



[Click to Take the FREE Python Machine Learning Crash-Course](#)



# Your First Machine Learning Project in Python Step-By-Step

by Jason Brownlee on February 10, 2019 in [Python Machine Learning](#)

[Tweet](#)[Share](#)[Share](#)

Last Updated on September 3, 2019

Do you want to do machine learning using Python, but you're having trouble getting started?

In this post, you will complete your first machine learning project using Python.

In this step-by-step tutorial you will:

1. Download and install Python SciPy and get the most useful package for machine learning in Python.
2. Load a dataset and understand its structure using statistical summaries and data visualization.
3. Create 6 machine learning models, pick the best and build confidence that the accuracy is reliable.

If you are a machine learning beginner and looking to finally get started using Python, this tutorial was designed for you.

Discover how to prepare data with pandas, fit and evaluate models with scikit-learn, and more in my new book, with 16 step-by-step tutorials, 3 projects, and full python code.

Let's get started!

- **Update Jan/2017:** Updated to reflect changes to the scikit-learn API in version 0.18.
- **Update Mar/2017:** Added links to help setup your Python environment.
- **Update Apr/2018:** Added some helpful links about randomness and making predictions.
- **Update Sep/2018:** Added link to my own hosted version of the dataset as UCI has become unreliable.
- **Update Feb/2019:** Updated to address warnings with sklearn API version 0.20+ with SVM and Logistic Regression, also updated results and plots.

Your Start in Machine Learning



Your First Machine Learning Project  
Photo by cosmoflash,

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

## How Do You Start Machine Learning?

The best way to learn machine learning is by designing and building your own projects.

## Python Can Be Intimidating When Getting Started

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems.

There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming.

The best way to get started using Python for machine learning is to complete a project.

- It will force you to install and start the Python interpreter (at the very least).
- It will give you a bird's eye view of how to step through a small project.
- It will give you confidence, maybe to go on to your own small projects.

## Beginners Need A Small End-to-End Project

Books and courses are frustrating. They give you lots of recipes and snippets, but you never get to see how they all fit together.

When you are applying machine learning to your own datasets, you are working on a project.

A machine learning project may not be linear, but it has a number of well known steps:

1. Define Problem.

Your Start in Machine Learning

2. Prepare Data.
3. Evaluate Algorithms.
4. Improve Results.
5. Present Results.

The best way to really come to terms with a new platform or tool is to work through a machine learning project end-to-end and cover the key steps. Namely, from loading data, summarizing data, evaluating algorithms and making some predictions.

If you can do that, you have a template that you can use on dataset after dataset. You can fill in the gaps such as further data preparation and improving result tasks later, once you have more confidence.

## Hello World of Machine Learning

The best small project to start with on a new tool is the

This is a good project because it is so well understood

- Attributes are numeric so you have to figure out how to use them.
- It is a classification problem, allowing you to practice a simple learning algorithm.
- It is a multi-class classification problem (multi-node tree).
- It only has 4 attributes and 150 rows, meaning it is small enough to fit on an A4 page).
- All of the numeric attributes are in the same units so no scaling or transforms to get started.

Let's get started with your hello world machine learning project in Python.

## Machine Learning in Python: Step-By-Step Tutorial (start here)

In this section, we are going to work through a small machine learning project end-to-end.

Here is an overview of what we are going to cover:

1. Installing the Python and SciPy platform.
2. Loading the dataset.
3. Summarizing the dataset.
4. Visualizing the dataset.
5. Evaluating some algorithms.
6. Making some predictions.

Take your time. Work through each step.

Try to type in the commands yourself or copy-and-paste

### Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

If you have any questions at all, please leave a comment at the bottom of the post.

## Need help with Machine Learning in Python?

Take my free 2-week email course and discover data prep, algorithms and more (with code).

Click to sign-up now and also get a free PDF Ebook version of the course.

[Start Your FREE Mini-Course Now!](#)

### Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

[START MY EMAIL COURSE](#)

## 1. Downloading, Installing and Setting Up SciPy

Get the Python and SciPy platform installed on your system.

I do not want to cover this in great detail, because other parts of this tutorial are more straightforward, especially if you are a developer. If you are not a developer, then you will need to follow the steps in this section.

### 1.1 Install SciPy Libraries

This tutorial assumes Python version 2.7 or 3.5+.

There are 5 key libraries that you will need to install. Below is a list of the Python SciPy libraries required for this tutorial:

- scipy
- numpy
- matplotlib
- pandas
- sklearn

There are many ways to install these libraries. My best advice is to pick one method then be consistent in installing each library.

The [scipy installation page](#) provides excellent instructions for installing the above libraries on multiple different platforms, such as Linux, mac OS X and Windows. If you have any doubts or questions, refer to this guide, it has been followed by thousands of people.

- On Mac OS X, you can use macports to install Python 2.7 and these libraries. For more information on macports, [see the homepage](#).
- On Linux you can use your package manager, such as yum on Fedora to install RPMs.

Your Start in Machine Learning

If you are on Windows or you are not confident, I would recommend installing the free version of Anaconda that includes everything you need.

**Note:** This tutorial assumes you have scikit-learn version 0.18 or higher installed.

Need more help? See one of these tutorials:

- How to Setup a Python Environment for Machine Learning and Deep Learning with Anaconda
- How to Create a Linux Virtual Machine For Machine Learning Development With Python 3

## 1.2 Start Python and Check Versions

It is a good idea to make sure your Python environment has the right versions of libraries expected.

The script below will help you test out your environment. It imports several libraries and prints the version.

Open a command line and start the python interpreter.

```
1 python
```

I recommend working directly in the interpreter or writing scripts in a terminal window rather than big editors and IDEs. Keep things simple and focus on learning the toolchain.

Type or copy and paste the following script:

```
1 # Check the versions of libraries
2
3 # Python version
4 import sys
5 print('Python: {}'.format(sys.version))
6 # scipy
7 import scipy
8 print('scipy: {}'.format(scipy.__version__))
9 # numpy
10 import numpy
11 print('numpy: {}'.format(numpy.__version__))
12 # matplotlib
13 import matplotlib
14 print('matplotlib: {}'.format(matplotlib.__version__))
15 # pandas
16 import pandas
17 print('pandas: {}'.format(pandas.__version__))
18 # scikit-learn
19 import sklearn
20 print('sklearn: {}'.format(sklearn.__version__))
```

Here is the output I get on my OS X workstation:

```
1 Python: 3.6.8 (default, Dec 30 2018, 13:01:55)
2 [GCC 4.2.1 Compatible Apple LLVM 9.1.0 (clang-902.0.39.21)]
3 scipy: 1.1.0
4 numpy: 1.15.4
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

```

5 matplotlib: 3.0.2
6 pandas: 0.23.4
7 sklearn: 0.20.2

```

Compare the above output to your versions.

Ideally, your versions should match or be more recent. The APIs do not change quickly, so do not be too concerned if you are a few versions behind. Everything in this tutorial will very likely still work for you.

If you get an error, stop. Now is the time to fix it.

If you cannot run the above script cleanly you will not be able to complete this tutorial.

My best advice is to Google search for your error message or post a question on Stack Exchange.

## 2. Load The Data

We are going to use the iris flowers dataset. This dataset is one of the most famous datasets in machine learning and statistics by pretty much everyone.

The dataset contains 150 observations of iris flowers. Four features are given as continuous measurements in centimeters. The fifth column is the species name, which corresponds to one of three species.

You can learn more about this dataset on [Wikipedia](#).

In this step we are going to load the iris data from CSV file [IRIS.csv](#).

### 2.1 Import libraries

First, let's import all of the modules, functions and objects we are going to use in this tutorial.

```

1 # Load libraries
2 import pandas
3 from pandas.plotting import scatter_matrix
4 import matplotlib.pyplot as plt
5 from sklearn import model_selection
6 from sklearn.metrics import classification_report
7 from sklearn.metrics import confusion_matrix
8 from sklearn.metrics import accuracy_score
9 from sklearn.linear_model import LogisticRegression
10 from sklearn.tree import DecisionTreeClassifier
11 from sklearn.neighbors import KNeighborsClassifier
12 from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
13 from sklearn.naive_bayes import GaussianNB
14 from sklearn.svm import SVC

```

Everything should load without error. If you have an error, stop. You need a working SciPy environment before continuing. See the advice above about setting up your environment.

### 2.2 Load Dataset

Your Start in Machine Learning

We can load the data directly from the UCI Machine Learning repository.

We are using pandas to load the data. We will also use pandas next to explore the data both with descriptive statistics and data visualization.

Note that we are specifying the names of each column when loading the data. This will help later when we explore the data.

```
1 # Load dataset
2 url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv"
3 names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
4 dataset = pandas.read_csv(url, names=names)
```

The dataset should load without incident.

If you do have network problems, you can download the file locally and load it using the same method, changing URL to the local file path.

## 3. Summarize the Dataset

Now it is time to take a look at the data.

In this step we are going to take a look at the data a few different ways.

1. Dimensions of the dataset.
2. Peek at the data itself.
3. Statistical summary of all attributes.
4. Breakdown of the data by the class variable.

Don't worry, each look at the data is one command. These are useful commands that you can use again and again on future projects.

### 3.1 Dimensions of Dataset

We can get a quick idea of how many instances (rows) and how many attributes (columns) the data contains with the `shape` property.

```
1 # shape
2 print(dataset.shape)
```

You should see 150 instances and 5 attributes:

```
1 (150, 5)
```

### 3.2 Peek at the Data

It is also always a good idea to actually eyeball your data.

```
1 # head
2 print(dataset.head(20))
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

You should see the first 20 rows of the data:

	sepal-length	sepal-width	petal-length	petal-width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1		
14	5.8	4.0	1.2		
15	5.7	4.4	1.5		
16	5.4	3.9	1.3		
17	5.1	3.5	1.4		
18	5.7	3.8	1.7		
19	5.1	3.8	1.5		
20	5.7	3.8	1.7		
21	5.1	3.8	1.5		

### 3.3 Statistical Summary

Now we can take a look at a summary of each attribute.

This includes the count, mean, the min and max values.

```
1 # descriptions
2 print(dataset.describe())
```

We can see that all of the numerical values have the same scale (centimeters) and similar ranges between 0 and 8 centimeters.

	sepal-length	sepal-width	petal-length	petal-width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

### 3.4 Class Distribution

Let's now take a look at the number of instances (rows) that belong to each class. We can view this as an absolute count.

```
1 # class distribution
2 print(dataset.groupby('class').size())
```

We can see that each class has the same number of instances (50 or 33% of the dataset).

```
1 class
2 Iris-setosa      50
```

Your Start in Machine Learning

3 Iris-versicolor	50
4 Iris-virginica	50

## 4. Data Visualization

We now have a basic idea about the data. We need to extend that with some visualizations.

We are going to look at two types of plots:

1. Univariate plots to better understand each attribute.
2. Multivariate plots to better understand the relationships between attributes.

### 4.1 Univariate Plots

We start with some univariate plots, that is, plots of each attribute.

Given that the input variables are numeric, we can create

```
1 # box and whisker plots
2 dataset.plot(kind='box', subplots=True, layout=(4,2))
3 plt.show()
```

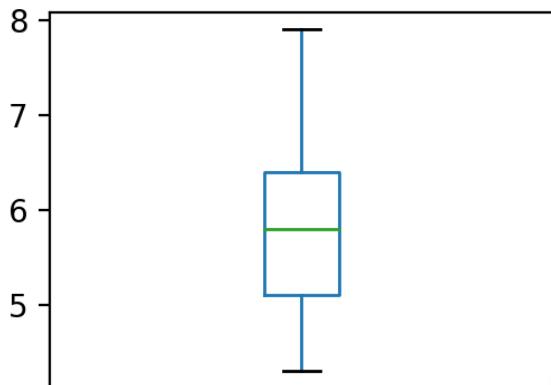
This gives us a much clearer idea of the distribution of

### Your Start in Machine Learning

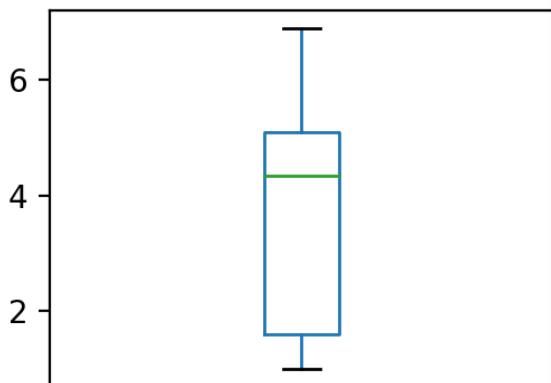
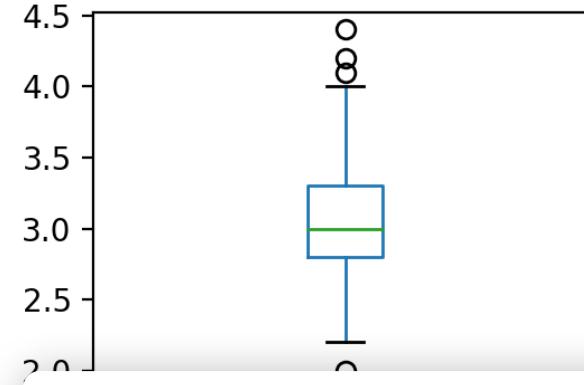
You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

Your Start in Machine Learning



sepal-length



petal-length

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

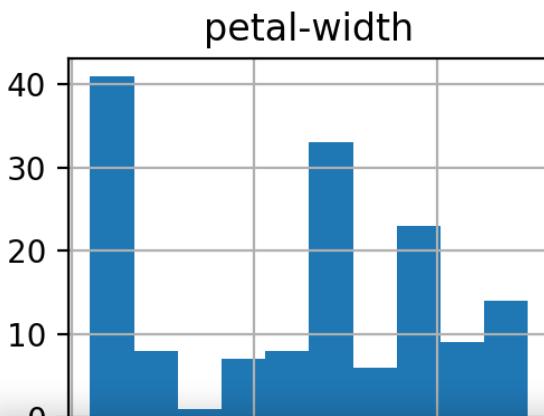
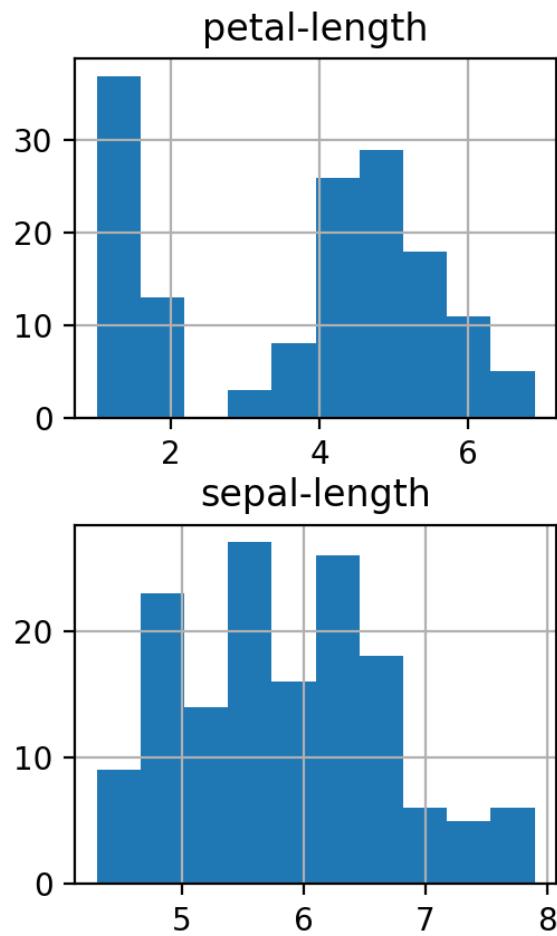
Box and Whisker Plots for Each Input Variable for the Iris Flowers Dataset

We can also create a histogram of each input variable to get an idea of the distribution.

```
1 # histograms
2 dataset.hist()
3 plt.show()
```

It looks like perhaps two of the input variables have a Gaussian distribution. This is useful to note as we can use algorithms that can exploit this assumption.

Your Start in Machine Learning



## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

Histogram Plots for Each Input Variable for the Iris Flowers Dataset

## 4.2 Multivariate Plots

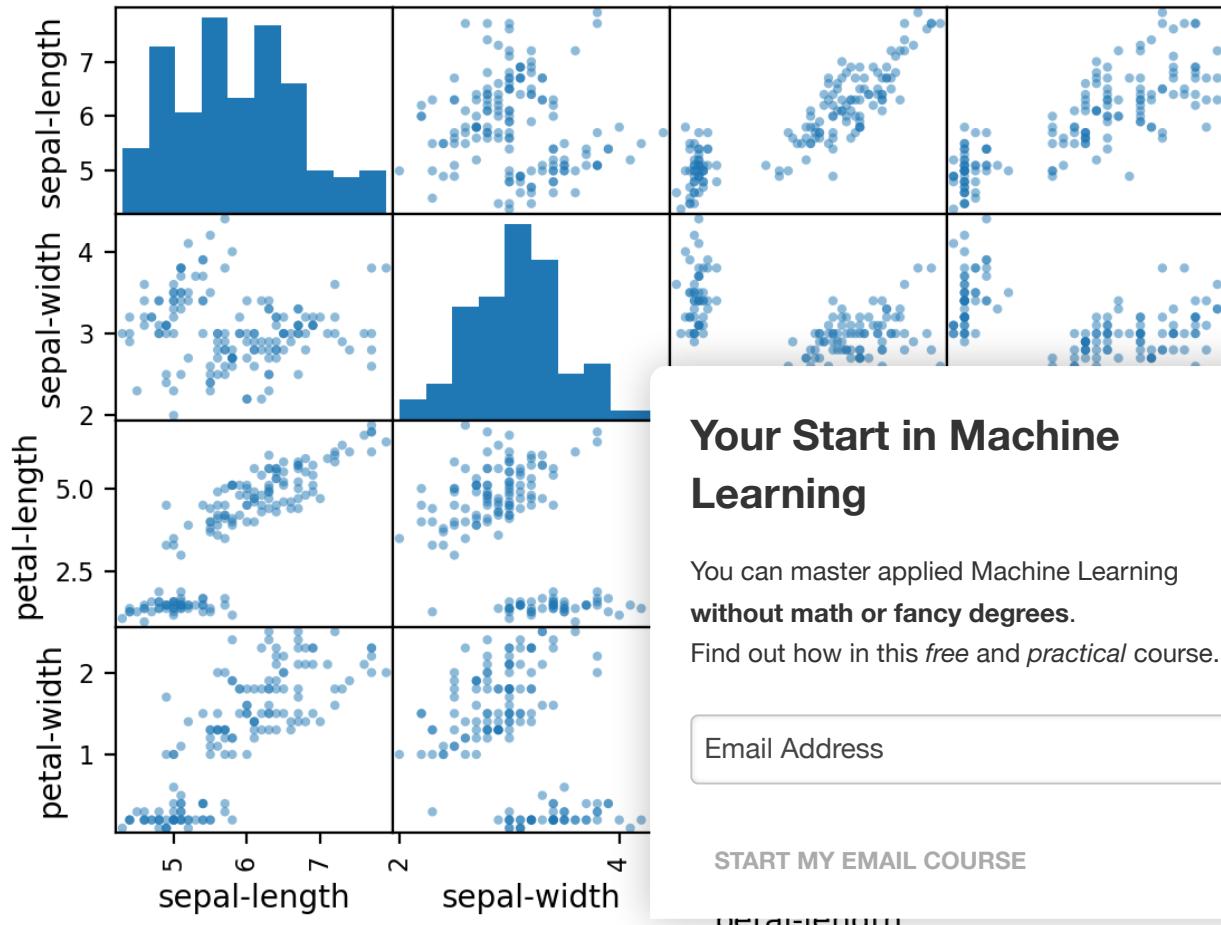
Now we can look at the interactions between the variables.

First, let's look at scatterplots of all pairs of attributes. This can be helpful to spot structured relationships between input variables.

```
1 # scatter plot matrix
2 scatter_matrix(dataset)
3 plt.show()
```

Note the diagonal grouping of some pairs of attributes. This suggests a high correlation and a predictable relationship.

Your Start in Machine Learning



## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)
[petal-length](#)

Scatter Matrix Plot for Each Input Variable for the Iris Flowers Dataset

## 5. Evaluate Some Algorithms

Now it is time to create some models of the data and estimate their accuracy on unseen data.

Here is what we are going to cover in this step:

1. Separate out a validation dataset.
2. Set-up the test harness to use 10-fold cross validation.
3. Build 5 different models to predict species from flower measurements
4. Select the best model.

### 5.1 Create a Validation Dataset

We need to know that the model we created is any good.

Later, we will use statistical methods to estimate the accuracy of the models that we create on unseen data. We also want a more concrete estimate of the accuracy of the best model on unseen data by evaluating it on actual unseen data.

[Your Start in Machine Learning](#)

That is, we are going to hold back some data that the algorithms will not get to see and we will use this data to get a second and independent idea of how accurate the best model might actually be.

We will split the loaded dataset into two, 80% of which we will use to train our models and 20% that we will hold back as a validation dataset.

```
1 # Split-out validation dataset
2 array = dataset.values
3 X = array[:,0:4]
4 Y = array[:,4]
5 validation_size = 0.20
6 seed = 7
7 X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=
```

You now have training data in the `X_train` and `Y_train` and validation data in the `X_validation` and `Y_validation` sets that we can use later.

Notice that we used a python slice to select the columns. If you are not familiar with slices, you might want to check-out this post:

- [How to Index, Slice and Reshape NumPy Arrays](#)

## 5.2 Test Harness

We will use 10-fold cross validation to estimate accuracy.

This will split our dataset into 10 parts, train on 9 and evaluate on the 10th part. We will repeat this process 10 times.

```
1 # Test options and evaluation metric
2 seed = 7
3 scoring = 'accuracy'
```

The specific random seed does not matter, learn more about pseudorandom number generators here:

- [Introduction to Random Number Generators for Machine Learning in Python](#)

We are using the metric of ‘accuracy’ to evaluate models. This is a ratio of the number of correctly predicted instances in divided by the total number of instances in the dataset multiplied by 100 to give a percentage (e.g. 95% accurate). We will be using the `scoring` variable when we run build and evaluate each model next.

## 5.3 Build Models

We don’t know which algorithms would be good on this problem or what configurations to use. We get an idea from the plots that some of the classes are partially linearly separable in some dimensions, so we are expecting generally good results.

Let’s evaluate 6 different algorithms:

- Logistic Regression (LR)

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

- Linear Discriminant Analysis (LDA)
- K-Nearest Neighbors (KNN).
- Classification and Regression Trees (CART).
- Gaussian Naive Bayes (NB).
- Support Vector Machines (SVM).

This is a good mixture of simple linear (LR and LDA), nonlinear (KNN, CART, NB and SVM) algorithms. We reset the random number seed before each run to ensure that the evaluation of each algorithm is performed using exactly the same data splits. It ensures the results are directly comparable.

Let's build and evaluate our models:

```

1 # Spot Check Algorithms
2 models = []
3 models.append(('LR', LogisticRegression(solver='liblinear', max_iter=500)))
4 models.append(('LDA', LinearDiscriminantAnalysis()))
5 models.append(('KNN', KNeighborsClassifier()))
6 models.append(('CART', DecisionTreeClassifier()))
7 models.append(('NB', GaussianNB()))
8 models.append(('SVM', SVC(gamma='auto'))))
9 # evaluate each model in turn
10 results = []
11 names = []
12 for name, model in models:
13     kfold = model_selection.KFold(n_splits=10)
14     cv_results = model_selection.cross_val_score(model, X, y, cv=kfold, scoring='accuracy')
15     results.append(cv_results)
16     names.append(name)
17     msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
18     print(msg)

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

## 5.4 Select Best Model

We now have 6 models and accuracy estimations for each. We need to compare the models to each other and select the most accurate.

Running the example above, we get the following raw results:

```

1 LR: 0.966667 (0.040825)
2 LDA: 0.975000 (0.038188)
3 KNN: 0.983333 (0.033333)
4 CART: 0.975000 (0.038188)
5 NB: 0.975000 (0.053359)
6 SVM: 0.991667 (0.025000)

```

Note, you're results may differ. For more on this see the post:

- Embrace Randomness in Machine Learning

In this case, we can see that it looks like Support Vector Machines (SVM) has the largest estimated accuracy score.

We can also create a plot of the model evaluation results and compare the spread and the mean accuracy of each model. There is a population of accuracy mea

Your Start in Machine Learning

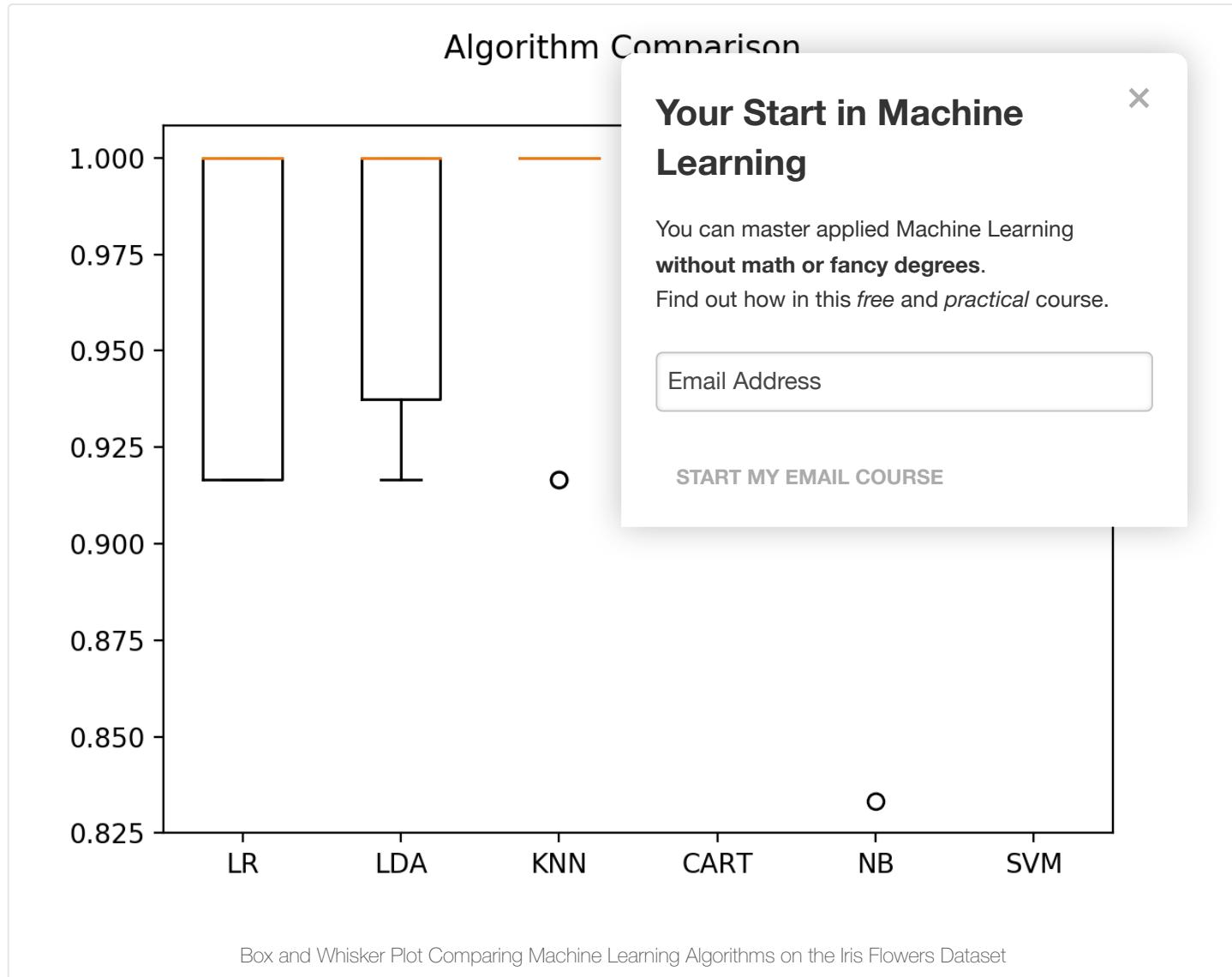
was evaluated 10 times (10 fold cross validation).

```

1 # Compare Algorithms
2 fig = plt.figure()
3 fig.suptitle('Algorithm Comparison')
4 ax = fig.add_subplot(111)
5 plt.boxplot(results)
6 ax.set_xticklabels(names)
7 plt.show()

```

You can see that the box and whisker plots are squashed at the top of the range, with many samples achieving 100% accuracy.



## 6. Make Predictions

The KNN algorithm is very simple and was an accurate model based on our tests. Now we want to get an idea of the accuracy of the model on our validation set.

This will give us an independent final check on the accuracy of the best model. It is valuable to keep a validation set just in case you made a slip during train

Your Start in Machine Learning

leak. Both will result in an overly optimistic result.

We can run the KNN model directly on the validation set and summarize the results as a final accuracy score, a confusion matrix and a classification report.

```
1 # Make predictions on validation dataset
2 knn = KNeighborsClassifier()
3 knn.fit(X_train, Y_train)
4 predictions = knn.predict(X_validation)
5 print(accuracy_score(Y_validation, predictions))
6 print(confusion_matrix(Y_validation, predictions))
7 print(classification_report(Y_validation, predictions))
```

We can see that the accuracy is 0.9 or 90%. The confusion matrix provides an indication of the three errors made. Finally, the classification report provides score and support showing excellent results (granted

1	0.9			
2	<code>[[ 7  0  0]</code>			
3	<code>[ 0 11  1]</code>			
4	<code>[ 0  2  9]]</code>			
5	precision			
6	recall			
7	f1-score			
8	Iris-setosa	1.00	1.00	1.00
9	Iris-versicolor	0.85	0.92	0.88
10	Iris-virginica	0.90	0.82	0.86
11	micro avg	0.90	0.90	0.90
12	macro avg	0.92	0.91	0.91
13	weighted avg	0.90	0.90	0.90

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

You can learn more about how to make predictions and predict probabilities [here](#).

- How to Make Predictions with scikit-learn

## You Can Do Machine Learning in Python

Work through the tutorial above. It will take you 5-to-10 minutes, max!

**You do not need to understand everything.** (at least not right now) Your goal is to run through the tutorial end-to-end and get a result. You do not need to understand everything on the first pass. List down your questions as you go. Make heavy use of the `help("FunctionName")` help syntax in Python to learn about all of the functions that you're using.

**You do not need to know how the algorithms work.** It is important to know about the limitations and how to configure machine learning algorithms. But learning about algorithms can come later. You need to build up this algorithm knowledge slowly over a long period of time. Today, start off by getting comfortable with the platform.

**You do not need to be a Python programmer.** The syntax of the Python language can be intuitive if you are new to it. Just like other languages, focus on function calls (e.g. `function()`) and assignments (e.g. `a =`

“b”). This will get you most of the way. You are a developer, you know how to pick up the basics of a language real fast. Just get started and dive into the details later.

**You do not need to be a machine learning expert.** You can learn about the benefits and limitations of various algorithms later, and there are plenty of posts that you can read later to brush up on the steps of a machine learning project and the importance of evaluating accuracy using cross validation.

**What about other steps in a machine learning project.** We did not cover all of the steps in a machine learning project because this is your first project and we need to focus on the key steps. Namely, loading data, looking at the data, evaluating some algorithms and making some predictions. In later tutorials we can look at other data preparation and result improvement tasks.

## Summary

In this post, you discovered step-by-step how to com

You discovered that completing a small end-to-end pr  
the best way to get familiar with a new platform.

## Your Next Step

Do you work through the tutorial?

1. Work through the above tutorial.
2. List any questions you have.
3. Search-for or research the answers.
4. Remember, you can use the `help("FunctionName")` in Python to get help on any function.

Do you have a question?

Post it in the comments below.

**Your Start in Machine Learning**

X

You can master applied Machine Learning without math or fancy degrees. Find out how in this *free* and *practical* course.

Email Address

[START MY EMAIL COURSE](#)

## Discover Fast Machine Learning in Python!

### Develop Your Own Models in Minutes

...with just a few lines of scikit-learn code

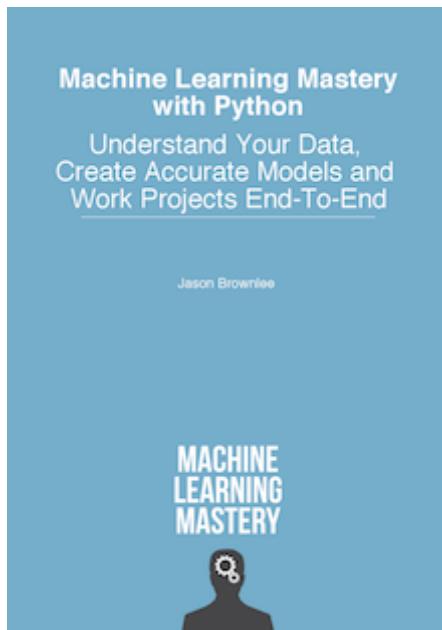
Learn how in my new Ebook:  
[Machine Learning Mastery With Python](#)

Covers **self-study tutorials** and **end-to-end projects** like:  
*Loading data, visualization, modeling, tuning, and much more...*

### Finally Bring Machine Learning To Your Own

[Your Start in Machine Learning](#)

Skip the Academics. Just Results.

[Tweet](#)[Share](#)[Share](#)

### About Jason Brownlee

Jason Brownlee, PhD is a machine learning researcher and author with modern machine learning methods via practical examples.

[View all posts by Jason Brownlee →](#)

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)[← How to Improve Performance With Transfer Learning for Deep Learning Neural Networks](#)[Framework for Better Deep Learning >](#)

1,488 Responses to *Your First Machine Learning Project in Python Step-By-Step*



**DR Venugopala Rao Manneni** June 11, 2016 at 5:58 pm #

[REPLY ↗](#)

Awesome... But in your Blog please introduce SOM ( Self Organizing maps) for unsupervised methods and also add printing parameters ( Coefficients )code.



**Jason Brownlee** June 14, 2016 at 8:17 am #

I generally don't cover unsupervised meth

[REPLY ↗](#)

[Your Start in Machine Learning](#)

This is because I mainly focus on and teach predictive modeling (e.g. classification and regression) and I just don't find unsupervised methods that useful.



**Rajesh** January 21, 2018 at 5:33 pm #

REPLY ↗

Jason,

Can you elaborate what you don't find unsupervised methods useful?



**Jason Brownlee** January 22, 2018 at 6:15 pm #

Because my focus is predictive m



**hamdy** November 19, 2018 at 8:04 pm #

DeprecationWarning: the imp module's documentation for alternatives what is the error?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 19, 2018 at 2:19 pm #

You can ignore this warning for now.



**Haider** June 16, 2019 at 7:23 pm #

Can you please help, where i'm doing mistake???

```
# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma='auto')))
# evaluate each model in turn
results = []
names = []
for name, model in models:
```

Your Start in Machine Learning

```
kfold = model_selection.KFold(n_splits=10, random_state=seed)
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold,
scoring=scoring)
results.append(cv_results)
names.append(name)
msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
print(msg)
```

ValueError Traceback (most recent call last)

in

13 for name, model in models:

14 kfold = model\_selection.KFold(n\_splits=10, random\_state=seed)

-> 15 cv\_results = model\_selection.cross\_val\_score(model, X\_train, Y\_train, cv=kfold,

scoring=scoring)

~\Anaconda3\lib\site-packages\sklear

cross\_val\_score(estimator, X, y, groups,

pre\_dispatch, error\_score)

400 fit\_params=fit\_params,

401 pre\_dispatch=pre\_dispatch,

-> 402 error\_score=error\_score)

403 return cv\_results['test\_score']

404

~\Anaconda3\lib\site-packages\sklear

cross\_validate(estimator, X, y, groups,

pre\_dispatch, return\_train\_score, return\_estimator, error\_score,

238 return\_times=True, return\_estimator=return\_estimator,

239 error\_score=error\_score)

-> 240 for train, test in cv.split(X, y, groups))

241

242 zipped\_scores = list(zip(\*scores))

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\parallel.py in \_\_call\_\_(self, iterable)

915 # remaining jobs.

916 self.\_iterating = False

-> 917 if self.dispatch\_one\_batch(iterator):

918 self.\_iterating = self.\_original\_iterator is not None

919

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\parallel.py in

dispatch\_one\_batch(self, iterator)

757 return False

758 else:

-> 759 self.\_dispatch(tasks)

760 return True

761

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\parallel.py in _dispatch(self, batch)
714     with self._lock:
715         job_idx = len(self._jobs)
-> 716         job = self._backend.apply_async(batch, callback=cb)
717     # A job can complete so quickly than its callback is
718     # called before we get here, causing self._jobs to

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\_parallel_backends.py in
apply_async(self, func, callback)
180     def apply_async(self, func, callback=None):
181         """Schedule a func to be run"""
-> 182         result = ImmediateResult(func)
183         if callback:
184             callback(result)

~\Anaconda3\lib\site-packages\sklear
__init__(self, batch)
547     # Don't delay the application, to a
548     # arguments in memory
-> 549     self.results = batch()
550

551     def get(self):
~\Anaconda3\lib\site-packages\sklear
223     with parallel_backend(self._backe
224     return [func(*args, **kwargs)
-> 225     for func, args, kwargs in self.it...]
226

227     def __len__(self):
~\Anaconda3\lib\site-packages\sklearn\externals\joblib\parallel.py in (.0)
223     with parallel_backend(self._backend, n_jobs=self._n_jobs):
224     return [func(*args, **kwargs)
-> 225     for func, args, kwargs in self.items]
226

227     def __len__(self):
~\Anaconda3\lib\site-packages\sklearn\model_selection\_validation.py in
_fit_and_score(estimator, X, y, scorer, train, test, verbose, parameters, fit_params,
return_train_score, return_parameters, return_n_test_samples, return_times,
return_estimator, error_score)
526     estimator.fit(X_train, **fit_params)
527     else:
-> 528     estimator.fit(X_train, y_train, **fit_params)
529

530     except Exception as e:

~\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py in fit(self, X, y,
sample_weight)

```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

Your Start in Machine Learning

```

1284 X, y = check_X_y(X, y, accept_sparse='csr', dtype=_dtype, order="C",
1285 accept_large_sparse=solver != 'liblinear')
-> 1286 check_classification_targets(y)
1287 self.classes_ = np.unique(y)
1288 n_samples, n_features = X.shape

~\Anaconda3\lib\site-packages\sklearn\utils\multiclass.py in check_classification_targets(y)
169 if y_type not in ['binary', 'multiclass', 'multiclass-multioutput',
170 'multilabel-indicator', 'multilabel-sequences']:
-> 171 raise ValueError("Unknown label type: %r" % y_type)
172
173

```

ValueError: Unknown label type: 'conti



**Jason Brownlee** June 17, 2019

I have some suggestions here  
<https://machinelearningmastery.com/f>  
code

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Aishwarya** April 11, 2018 at 1:49 pm #

I got quite different results though i used same seed and splits

Svm : 0.991667 (0.025) with highest accuracy

KNN : 0.9833

CART : 0.9833

Why ?

REPLY ↗



**Aishwarya** April 11, 2018 at 1:59 pm #

Im getting error saying

Cannot perform reduce with flexible type

While comparing algos using boxplots



**Jason Brownlee** April 11, 2018 at 4:26 pm #

Sorry, I have not seen this error before. Are you able to confirm that your environment is up to date?

Your Start in Machine Learning



**Ycyusa** August 5, 2018 at 9:31 am #

I followed your steps and I got the similar result as Aishwarya

SVM: 0.991667 (0.025000)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)



**Jason Brownlee** April 11, 2018 at 4:11 pm #

The API may have changed since  
small changes in predictions that are perhaps



**Aishwarya** April 11, 2018 at 10:54 pm #

Ive done this on kaggle.

Under ML kernal

<http://Www.kaggle.com/aishuvenkat09>

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Aishwarya** April 11, 2018 at 10:54 pm #

Sorry

<http://Www.kaggle.com/aishwarya09>



**Jason Brownlee** April 12, 2018 at 8:43 am #

Well done!



**manohar** April 23, 2018 at 6:49 pm #

Hi ,

I have same issues with above our friends discussed

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

Your Start in Machine Learning

CART: 0.983333 (0.033333)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

In that svm has more accuracy when compare to rest  
so i go ahead svm



**Jason Brownlee** April 24, 2018 at 6:26 am #

Yes.



**Ali** May 10, 2018 at 8:58 am #

Yes. I got the same. Dr. Jason had



**Sai Prasad** September 14, 2018 at 5:00 pm #

I also have the same result.

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.983333 (0.033333)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 15, 2018 at 6:01 am #

Nice.



**bharat** May 19, 2018 at 9:45 pm #

REPLY ↗

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold,
scoring=scoring)
```

sir i am getting error in this in of code.What should i do?

**Jason Brownlee** May 20, 2018 at 6:00 pm #

Your Start in Machine Learning



What error?

**AVNEESH UPADHYAY** June 25, 2018 at 5:00 am #

REPLY ↗

I think cv may be equal to the number of times you want to perform k-fold cross validation for e.g. 10,20etc. and in scoring parameter, you need to mention which type of scoring parameter you want to use for example 'accuracy'.

Hope this might help....

**Jason Brownlee** June 25, 2018 at 5:00 pm #

Correct.

More on how cross validation works here:  
<https://machinelearningmastery.com/k-fold-cross-validation/>

**Ved Anshu** September 21, 2018 at 4:22 pm #

Bro kindly use train\_test\_split() in

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

**Mohammed** March 25, 2019 at 2:54 pm #

REPLY ↗

thank you so much really its very useful

in the last step you are used KNN to make predictions why you are used KNN can we use SVM and can we make compare with all the models in predictions ?

**Jason Brownlee** March 26, 2019 at 7:58 am #

REPLY ↗

It is just an example, you can make predictions with any model you wish.

Often we prefer simpler models (like knn) over more complex models (like svm).

**Hasnain** July 8, 2017 at 8:55 pm #

REPLY ↗

I have installed all libraries that were in your How to Setup Python environment... blog. All went fine but when i run the starting imports code I get

[Your Start in Machine Learning](#)

named ‘pandas’”. But I did install it using “pip install pandas” command. I am working on a windows machine.



**Jason Brownlee** July 9, 2017 at 10:53 am #

REPLY ↗

Sorry to hear that. Consider rebooting your machine?



**Sheila Dawn** August 9, 2017 at 5:43 am #

REPLY ↗

I had the same problem initially, b  
libraries, and another for loading the iris da  
Then I decided to put the two commands



**Jason Brownlee** August 9, 2017 at 10:53 am #

Yes, all commands go in the c

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Dan Fiorino** July 16, 2017 at 2:37 am #

Hasnain, try setting the environment variable PYTHON\_PATH and PATH to include the path to the site packages of the version of python you have permission to alter

```
export PYTHONPATH="$PYTHONPATH:/path/to/Python/2.7/site-packages/"
export PATH="$PATH:/path/to/Python/2.7/site-packages/"
```

obviously replacing “/path/to” with the actual path. My system Python is in my /Users//Library folder but I’m on a Mac.

You can add the export lines to a script that runs when you open a terminal (“~/.bash\_profile” if you use BASH).

That might not be 100% right, but it should help you on your way.



**Jason Brownlee** July 16, 2017 at 8:00 am #

REPLY ↗

Thanks for posting the tip Dan, I hope it helps.

## Your Start in Machine Learning



**Jason Robinette** September 7, 2017 at 11:16 am #

got it to work have no idea how but it worked! I am like the kid at t-ball that closes his eyes and takes a swing!



**Jason Brownlee** September 7, 2017 at 12:58 pm #

I'm glad to hear that!



**Tanya** September 30, 2017 at 11:08 am #

I am starting at square 0, and after clearing the libraries at all... (as a newb), I didn't see what was what.  
# Load libraries  
import pandas  
from pandas.tools.plotting import scatter\_matrix  
import matplotlib.pyplot as plt  
from sklearn import model\_selection  
from sklearn.metrics import classification\_report  
from sklearn.metrics import confusion\_matrix  
from sklearn.metrics import accuracy\_score  
from sklearn.linear\_model import LogisticRegression  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.neighbors import KNeighborsClassifier  
from sklearn.discriminant\_analysis import LinearDiscriminantAnalysis  
from sklearn.naive\_bayes import GaussianNB  
from sklearn.svm import SVC

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** October 1, 2017 at 9:04 am #

REPLY ↗

Perhaps this step-by-step tutorial will help you set up your environment:  
<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**KASINATH PS** December 7, 2017 at 8:16 pm #

REPLY ↗

if u r using python 3

Your Start in Machine Learning

save all the commands as a py file

then in a python shell enter

```
exec(open("[path to file with name]").read())
```

if you open shell in the same path as the saved thing

then you only need to enter the filename alone

ex:

lets say i saved it as load.py

then

```
exec(open("load.py").read())
```

this will execute all commands in the current



**Rahul** December 7, 2017 at 10:28 pm #

Hi Tanya,

This tutorial is so intuitive that I went through it. To install PyCharm from JetBrains available here: <https://www.jetbrains.com/pycharm/download/#section=windows&code=PCC>

Install PIP (The de-facto python package manager) and then bring up the interactive DOS like terminal. Run the following commands:

```
pip install numpy
```

```
pip install scipy
```

```
pip install matplotlib
```

```
pip install pandas
```

```
pip install sklearn
```

All other steps in the tutorial are valid and do not need a single line of change apart from where its mentioned

```
from pandas.tools.plotting import scatter_matrix , change it to
```

```
from pandas.plotting import scatter_matrix
```



**Jason Brownlee** December 8, 2017 at 5:39 am #

Thanks for the tips Rahul.



**Murtaza** December 17, 2017 at 11:05 am #

Your Start in Machine Learning

For a beginner i believe Anacondas Jupyter notebooks would be the best option. As they can include markdown for future reference which is essential as beginner (backpropogation :p). But again varies person to person



**Jason Brownlee** December 18, 2017 at 5:19 am #

I find notebooks confuse beginners more than help.

Running a Python script on the command line is so much simpler.



**Jason** March 1, 2018 at 4:18 pm #

Except for me, on Debian Stretch  
from pandas.tools.plotting import scatter\_matrix



**Jason Brownlee** March 2, 2018 at 10:11 am #

You must update your version

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**avanish** March 25, 2018 at 7:11 pm #

REPLY ↩

use jupyter notebook ...there all the essential libraries are preinstalled



**Anmoldeep1509** October 31, 2018 at 6:50 am #

REPLY ↩

I also did a similar mistake, I am also a newbie to python, and wrote those import statements in the separate file, and imported the created file, without knowing how imports work...after your reply realized my mistake and now back on track thanks!



**Tushar** June 22, 2018 at 4:50 am #

REPLY ↩

I also had problems installing modules on windows. Although, there was no error of any kind if installed from PyCharm IDE.  
Also, use 32-bit python interpreter if you wanna use NLTK. It can be done even on 64-bit version, but was not worth the time it would it need.

Your Start in Machine Learning



**Karan sing** March 26, 2019 at 8:28 pm #

REPLY ↗

If you are working on virtual environment then you have to make script first and run it by activating the virtual environment,

If you are not working on virtual environment then run your scripts on time



**Yuvraj** July 13, 2018 at 1:56 am #

REPLY ↗

Could you please go into the mathematical concept behind KNN and why the accuracy resulted in the highest score? Thank you



**Mario** October 4, 2018 at 8:13 pm #

I like your tutorial for the machine learning where I am

```
# Compare Algorithms
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

This is the answer I am getting from it

```
TypeError Traceback (most recent call last)
```

```
in ()
```

```
3 fig.suptitle('Algorithm Comparison')
```

```
4 ax = fig.add_subplot(111)
```

```
--> 5 plt.boxplot(results)
```

```
6 ax.set_xticklabels(names)
```

```
7 plt.show()
```

```
~\Anaconda3\lib\site-packages\matplotlib\pyplot.py in boxplot(x, notch, sym, vert, whis, positions, widths, patch_artist, bootstrap, usermedians, conf_intervals, meanline, showmeans, showcaps, showbox, showfliers, boxprops, labels, flierprops, medianprops, meanprops, capprops, whiskerprops, manage_xticks, autorange, zorder, hold, data)
```

```
2846 whiskerprops=whiskerprops,
```

```
2847 manage_xticks=manage_xticks, autorange=autorange,
```

```
-> 2848 zorder=zorder, data=data)
```

```
2849 finally:
```

```
2850 ax._hold = washold
```

## Your Start in Machine Learning

X

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

```

~\Anaconda3\lib\site-packages\matplotlib\__init__.py in inner(ax, *args, **kwargs)
1853 "the Matplotlib list!)" % (label_namer, func.__name__),
1854 RuntimeWarning, stacklevel=2)
-> 1855 return func(ax, *args, **kwargs)
1856
1857 inner.__doc__ = _add_data_doc(inner.__doc__,

~\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py in boxplot(self, x, notch, sym, vert, whis,
positions, widths, patch_artist, bootstrap, usermedians, conf_intervals, meanline, showmeans,
showcaps, showbox, showfliers, boxprops, labels, flierprops, medianprops, meanprops, capprops,
whiskerprops, manage_xticks, autorange, zorder)
3555
3556 bxpstats = cbook.boxplot_stats(x, whis=whis, bootstrap=bootstrap)
-> 3557 labels=labels, autorange=autorange)
3558 if notch is None:
3559 notch = rcParams['boxplot.notch']

~\Anaconda3\lib\site-packages\matplotlib\cbook\__init__.py in autorange()
1839
1840 # arithmetic mean
-> 1841 stats['mean'] = np.mean(x)
1842
1843 # medians and quartiles

~\Anaconda3\lib\site-packages\numpy\core\fromnumeric.py in _mean(a, axis, dtype, out, keepdims)
2955
2956 return _methods._mean(a, axis=axis, dtype=dtype,
-> 2957 out=out, **kwargs)
2958
2959

~\Anaconda3\lib\site-packages\numpy\core\_methods.py in _mean(a, axis, dtype, out, keepdims)
68 is_float16_result = True
69
-> 70 ret = umr_sum(arr, axis, dtype, out, keepdims)
71 if isinstance(ret, mu.ndarray):
72 ret = um.true_divide()

TypeError: cannot perform reduce with flexible type

```

HOW CAN I FIX THIS?



**Jason Brownlee** October 5, 2018 at 5:33 am #

REPLY ↗

Perhaps post your code and error to stackoverflow.com?

Your Start in Machine Learning



**Swapna** December 7, 2018 at 8:42 pm #

REPLY ↗

Jason nice work.but I had some doubt about that Species column, in that we should predict t test for continuous and catagorical variable only 2 group..in this column there having 3 groups so how we predict t test.please give me answer



**Jason Brownlee** December 8, 2018 at 7:06 am #

The Student's t-test is for numerical data only, you can learn more here:

<https://machinelearningmastery.com/parametric-statistical-significance-tests-in-python/>

## Your Start in Machine Learning

X

You can master applied Machine Learning

**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Brandon** January 23, 2019 at 4:37 pm #

I also got a traceback on this section:  
TypeError: cannot perform reduce with flexible

Quick check on stackoverflow show's that plt.  
error in section 5.4 line 15.

Wrong code: results.append(results)

Coorect: reslts.append(cv\_results)

woohoo for tracebacks and wrong data-types. hope someone finds this helpful.



**Jason Brownlee** January 24, 2019 at 6:40 am #

REPLY ↗

Are you able to confirm that your python libraries are up to date?



**Ademola** November 27, 2018 at 7:49 am #

REPLY ↗

Well done



**Jan de Lange** June 20, 2016 at 10:43 pm #

REPLY ↗

Nice work Jason. Of course there is a lot more to tell about the code and the Models applied if this is intended for people starting out with ML (like me). Rather than telling which "button to press" to make work, it would be nice to know why also. I looked at a sample of you book (advanced) if you are covering the why also, but it looks like it's limited?

Your Start in Machine Learning

On this particular example, in my case SVM reached 99.2% and was thus the best Model. I gather this is because the test and training sets are drawn randomly from the data.



**Jason Brownlee** June 21, 2016 at 7:04 am #

REPLY ↗

This tutorial and the book are laser focused on how to use Python to complete machine learning projects.

They already assume you know how the algorithms work.

If you are looking for background on machine learning algorithms, take a look at this book:

<https://machinelearningmastery.com/master-machine-learning-algorithms/>



**Alan** July 26, 2017 at 10:50 pm #

Jan de Lange and Jason,

Before anything else, I truly like to thank Jason for this article. I have been trying to get the same accuracy result for SVM (0.991667 to be exact) as you have provided. I have installed the latest version of scikit-learn on my machine to version 0.18.1. The discrepancy could be because of its random selection of best model. It's a new 'can of worms' as some say, since the selection of best model is on the line.

In terms of the example you have provided, I can't seem to get the same accuracy result for SVM (0.991667 to be exact) as you have provided. I have installed the latest version of scikit-learn on my machine to version 0.18.1. The discrepancy could be because of its random selection of best model. It's a new 'can of worms' as some say, since the selection of best model is on the line.

Thank you again Jason for this practical article on ML.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 27, 2017 at 8:06 am #

REPLY ↗

Thanks Alan.

Absolutely. Machine learning algorithms are stochastic. This is a feature, not a bug. It helps us move through the landscape of possible models efficiently.

See this post:

<http://machinelearningmastery.com/randomness-in-machine-learning/>

And this post on finalizing a model:

<http://machinelearningmastery.com/train-final-machine-learning-model/>

Does that help?

**Per** December 15, 2017 at 7:36 pm #

Your Start in Machine Learning



Got it working too, changing the scatter\_matrix import like Rahul did.  
But I also had to install tkinter first (yum install tkinter).

Very nice tutorial, Jason!



**Jason Brownlee** December 16, 2017 at 5:24 am #

REPLY ↩

Glad to hear it!



**Nil** June 25, 2016 at 12:42 am #

Awesome, I have tested the code it is impressive.  
Iris-setosa or Iris-versicolor or Iris-virginica when I am  
width, petal-length and petal-width attributes?



**Jason Brownlee** June 25, 2016 at 5:09 am #

Great question. You can call model.predict()

For an example, see Part 6 in the above post.

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**JamieFox** March 28, 2017 at 6:38 am #

REPLY ↩

Dear Jason Brownlee, I was thinking about the same question of Nil. To be precise I was  
wondering how can I know, after having seen that my model has a good fit, which values of sepal-  
length, sepal-width, petal-length and petal-width corresponds to Iris-setosa ecc..  
For instance, if I have p predictors and two classes, how can I know which values of the predictors  
blend to one class or the other. Knowing the value of predictors allows me to use the model in the  
daily operativity. Thx



**Jason Brownlee** March 28, 2017 at 8:27 am #

REPLY ↩

Not knowing the statistical relationship between inputs and outputs is one of the down  
sides of using neural networks.



**JamieFox** March 29, 2017 at 7:03

Your Start in Machine Learning

Hi Mr Jason Brownlee, thks for your answer. So all algorithms, such as SVM, LDA, random forest.. have this drawbacks? Can you suggest me something else? Because logistic regression is not like this, or am I wrong?



**Jason Brownlee** March 29, 2017 at 9:14 am #

All algorithms have limitations and assumptions. For example, Logistic Regression makes assumptions about the distribution of variates (Gaussian) and more:  
[https://en.wikipedia.org/wiki/Logistic\\_regression](https://en.wikipedia.org/wiki/Logistic_regression)

Nevertheless, we can make useful models (skillful) even when breaking assumptions or pushing past limitations.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Sujon** September 6, 2016 at 8:19 am #

Dear Sir,

It seems I'm in the right place in right time! I'm doing my University. Could you give me some references for laughter audio conversion. Please send me anything on [sujon2100@gmail.com](mailto:sujon2100@gmail.com). Thanks a lot a



**Sujon** September 6, 2016 at 8:32 am #

Sorry I mean laughter audio to CSV conversion.

REPLY ↗



**Jason Brownlee** September 6, 2016 at 9:49 am #

REPLY ↗

Sorry, I have not seen any laughter audio to CSV conversion tools/techniques.



**Sujon** May 10, 2017 at 1:02 pm #

REPLY ↗

Hi again, do you have any publication of this article "Your First Machine Learning Project in Python Step-By-Step"? Or any citation if you know? Thanks.



**Jason Brownlee** May 11, 2017 at 8:28 am #

REPLY ↗

No, you can reference the blog post

Your Start in Machine Learning



**Roberto U** September 19, 2016 at 9:17 am #

REPLY ↗

Sweet way of condensing monstrous amount of information in a one-way street. Thanks!

Just a small thing, you are creating the Kfold inside the loop in the cross validation. Then, you use the same seed to keep the comparison across predictors constant.

That works, but I think it would be better to take it out of the loop. Not only is more efficient, but it is also much immediately clearer that all predictors are using the same Kfold.

You can still justify the use of the seeds in terms of replicability; readers getting the same results on their machines.

Thanks again!



**Jason Brownlee** September 20, 2016 at 8:27

Great suggestion, thanks Roberto.



**Francisco** September 20, 2016 at 2:02 am #

Hello Jaso.

Thank you so much for your help with Machine Learning and congratulations for your excellent website.

I am a beginner in ML and DeepLearning. Should I download Python 2 or Python 3?

Thank you very much.

Francisco

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 20, 2016 at 8:33 am #

REPLY ↗

I use Python 2 for all my work, but my students report that most of my examples work in Python 3 with little change.



**ShawnJ** October 11, 2016 at 5:24 am #

REPLY ↗

Jason,

Thank you so much for putting this together. I am been a software developer for almost two decades and am getting interested in machine learning. Found this

Your Start in Machine Learning



**Jason Brownlee** October 11, 2016 at 7:24 am #

REPLY ↩

Thanks ShawnJ, I'm glad you found it useful.



**Wendy G** October 14, 2016 at 5:37 am #

REPLY ↩

Jason,

Thanks for the great post! I am trying to follow this post by using my own dataset, but I keep getting this error "Unknown label type: array ([some numbers from possible solutions?]

Thanks,



**Jason Brownlee** October 14, 2016 at 9:08 am #

X

Hi Wendy,

Carefully check your data. Maybe print it on the screen so that you may need to convert to numbers using data

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**fara** October 20, 2016 at 7:15 am #

REPLY ↩

hi thanks for great tutorial, i'm also new to ML...this really helps but i was wondering what if we have non-numeric values? i have mixture of numeric and non-numeric data and obviously this only works for numeric. do you also have a tutorial for that or would you please send me a source for it? thank you



**Jason Brownlee** October 20, 2016 at 8:41 am #

REPLY ↩

Great question fara.

We need to convert everything to numeric. For categorical values, you can convert them to integers (label encoding) and then to new binary features (one hot encoding).



**fara** October 20, 2016 at 8:53 am #

REPLY ↩

after I post my comment here i saw this "DictVectorizer" i think i can use it for converting non-numeric to numeric, right?

Your Start in Machine Learning



**Jason Brownlee** October 20, 2016 at 11:15 am #

REPLY ↗

I would recommend the LabelEncoder class followed by the OneHotEncoder class in scikit-learn.

I believe I have tutorials on these here:

<http://machinelearningmastery.com/data-preparation-gradient-boosting-xgboost-python/>



**fara** October 21, 2016 at 3:53 am #

thank you it's great

X



**Mazhar Dootio** October 23, 2016 at 9:14 pm #

Hello Jason

Thank you for publishing this great machine learning tu  
It is really awesome awesome awesome.....!

I test your tutorial on python-3 and it works well but wh  
drive. I followed your give instructions but couldn't be  
My syntax is as under:

```
import unicodedata
url = open(r'C:\Users\mazhar\Anaconda3\Lib\site-packages\sindhi2.csv', encoding='utf-8').readlines()
names = ['class', 'sno', 'gender', 'morphology', 'stem', 'fword']
dataset = pandas.read_csv(url, names=names)
```

python-3 jupyter notebook does not loads this. Kindly help me in regard.

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** October 24, 2016 at 7:05 am #

REPLY ↗

Hi Mazhar, thanks.

Are you able to load the file on the command line away from the notebook?

Perhaps the notebook environment is causing trouble?



**Kenny** October 11, 2017 at 3:43 am #

REPLY ↗

Mazhar try this:

## Your Start in Machine Learning

```
import pandas as pd
.
.
.
file= \"namefile.csv\" #or c:/____/____/
df = pd.read_csv(file)

in Jupyter

https://www.anaconda.com/download/
https://anaconda.org/anaconda/python
```



**Mazhar Dootio** October 25, 2016 at 3:22 am #

Dear Jason

Thank you for response

I am using Python 3 with anaconda jupyter notebook so which python version you would like to suggest me file from local drive that how can I load utf-8 dataset fil

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)



**Jason Brownlee** October 25, 2016 at 8:32 am #

Hi Mazhar, I teach using Python 2.7 with examples from the command line.

Many of my students report that the code works in Python 3 and in notebooks with little or no changes.



**Kenny** October 11, 2017 at 3:50 am #

[REPLY](#) ↗

try with this command:

```
df = pd.read_csv(file, encoding='latin-1') #if you are working with csv "," or ";" put sep='|',
```



**Andy** October 27, 2016 at 11:59 pm #

[REPLY](#) ↗

Great tutorial but perhaps I'm missing something here. Let's assume I already know what model to use (perhaps because I know the data well... for example).

```
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
```

I then use the models to predict:

```
print(knn.predict(an array of variables of a record I war
```

[Your Start in Machine Learning](#)

Is this where the whole ML happens?

knn.fit(X\_train, Y\_train)

What's the difference between this and say a non ML model/algorithm? Is it that in a non ML model I have to find the coefficients/parameters myself by statistical methods?; and in the ML model the machine does that itself?

If this is the case then to me it seems that a researcher/coder did most of the work for me and wrap it in a nice function. Am I missing something? What is special here?



**Jason Brownlee** October 28, 2016 at 9:14 am #

REPLY ↗

Hi Andy,

Yes, your comment is generally true.

The work is in the library and choice of good library project can take you a very long way very quickly.

Stats is really about small data and understanding at least in common practice, is leaning towards automation (predictive modeling) at the expense of model interpretation. This trumps traditional goals of stats.

Because of the automation, the focus shifts more to engineering, automatic algorithm tuning and ensuring algorithms themselves taking more of a backseat to

Does that make sense?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Andy** November 3, 2016 at 10:36 pm #

REPLY ↗

It does make sense.

You mentioned 'data quality'. That's currently my field of work. I've been doing this statistically until now, and very keen to try a different approach. As a practical example how would you use ML to spot an error/outlier using ML instead of stats?

Let's say I have a large dataset containing trees: each tree record contains a species, height, location, crown size, age, etc... (ah! suspiciously similar to the iris flowers dataset 😊) Is ML a viable method for finding incorrect data and replace with an "estimated" value? The answer I guess is yes. For species I could use almost an identical method to what you presented here; BUT what about continuous values such as tree height?



**Jason Brownlee** November 4, 2016 at 9:08 am #

REPLY ↗

Hi Andy,

## Your Start in Machine Learning

Maybe “outliers” are instances that cannot be easily predicted or assigned ambiguous predicted probabilities.

Instance values can be “fixed” by estimating new values, but whole instance can also be pulled out if data is cheap.



**Shailendra Khadayat** October 30, 2016 at 2:23 pm #

REPLY ↩

Awesome work Jason. This was very helpful and expect more tutorials in the future.

Thanks.



**Jason Brownlee** October 31, 2016 at 5:26 am #

I'm glad you found it useful Shailendra.



**franklin** September 18, 2019 at 6:50 am #

Thank you for the good work you doing o  
i want to know how electricity appliance consumpti

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 18, 2019 at 2:05 pm #

REPLY ↩

Thanks, I'm glad it helped.

If you are referring to the time series examples, you can learn more about the dataset here:  
<https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption>



**Shuvam Ghosh** November 16, 2016 at 12:13 am #

REPLY ↩

Awesome work. Students need to know how the end results will look like. They need to get motivated to learn and one of the effective means of getting motivated is to be able to see and experience the wonderful end results. Honestly, if i were made to study algorithms and understand them i would get bored. But now since i know what amazing results they give, they will serve as driving forces in me to get into details of it and do more research on it. This is where i hate the orthodox college ways of teaching. First get the theory right then apply. No way. I need to see things first to get motivated.

## Your Start in Machine Learning



**Jason Brownlee** November 16, 2016 at 9:29 am #

REPLY ↗

Thanks Shuvam,

I'm glad my results-first approach gels with you. It's great to have you here.



**Puneet** November 17, 2016 at 12:08 am #

REPLY ↗

Thanks Jason,

while i am trying to complete this.

```
# Spot Check Algorithms
models = []
models.append('LR', LogisticRegression())
models.append('LDA', LinearDiscriminantAnalysis())
models.append('KNN', KNeighborsClassifier())
models.append('CART', DecisionTreeClassifier())
models.append('NB', GaussianNB())
models.append('SVM', SVC())
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

showing below error.-

```
kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
^
```

IndentationError: expected an indented block-

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 17, 2016 at 9:54 am #

REPLY ↗

Hi Puneet, looks like a copy-paste error.

Check for any extra new lines or white space around that line that is reporting the error.

## Your Start in Machine Learning



**Bram** March 10, 2018 at 7:51 am #

REPLY ↗

<https://stackoverflow.com/questions/4446366/why-am-i-getting-indentationerror-expected-an-indented-block>

This solved it for me. Copy code to notepad, replace all tabs with 4 spaces.



**Jason Brownlee** March 11, 2018 at 6:15 am #

REPLY ↗

Nice work.



**Puneet** November 17, 2016 at 12:30 am #

Thanks Json,

I am new to ML. need your help so i can run this.

as i have followed the steps but when trying to build a

```
-----  
# Spot Check Algorithms  
models = []  
models.append(('LR', LogisticRegression()))  
models.append(('LDA', LinearDiscriminantAnalysis()))  
models.append(('KNN', KNeighborsClassifier()))  
models.append(('CART', DecisionTreeClassifier()))  
models.append(('NB', GaussianNB()))  
models.append(('SVM', SVC()))  
# evaluate each model in turn  
results = []  
names = []  
for name, model in models:  
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)  
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)  
    results.append(cv_results)  
    names.append(name)  
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())  
    print(msg)
```

facing below mentioned issue.

File "", line 13

```
kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)  
^
```

IndentationError: expected an indented block

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Kindly help.



**Martin** November 18, 2016 at 5:18 am #

REPLY ↩

Puneet, you need to indent the block (tab or four spaces to the right). That is the way of building a block in Python



**Casey** December 2, 2018 at 3:58 am #

DEDI V ↩



I am also having this problem, I have executes. It seems to be waiting for more input happens. Is there something I am missing to e

```
>>> for name, model in models:  
... kfold = model_selection.KFold(n_splits=10,  
... cv_results = model_selection.cross_val_score  
... results.append(cv_results)  
... names.append(name)  
... msg = "%s: %f (%f)" % (name, cv_results.r  
... print(msg)  
...  
...
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 2, 2018 at 6:23 am #

REPLY ↩

Save the code to a file and run it from the command line. I show how here:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-run-a-script-from-the-command-line>



**george soilis** November 17, 2016 at 10:00 pm #

REPLY ↩

just another Python noob here, sending many regards and thanks to Jason :):)



**Jason Brownlee** November 18, 2016 at 8:22 am #

REPLY ↩

Thanks george, stick with it!

## Your Start in Machine Learning

**sergio** November 22, 2016 at 3:29 pm #

REPLY ↗

Does this tutorial work with other data sets? I'm trying to work on a small assignment and I want to use python

**Jason Brownlee** November 23, 2016 at 8:50 am #

REPLY ↗

It should provide a great template for new projects sergio.

**Brian** February 28, 2018 at 4:10 am #

I tried to use another dataset. I am no names, I still get the petal stuff as output. All then it gives me those old outputs.

**Albert** November 26, 2016 at 1:55 am #

Very Awesome step by step for me ! Even I am about Machine learning ~ supervised ML. Appreciate

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)**Jason Brownlee** November 26, 2016 at 10:38 am #

REPLY ↗

I'm glad to hear that Albert.

**Umar Yusuf** November 27, 2016 at 4:04 am #

REPLY ↗

Thank you for the step by step instructions. This will go along way for newbies like me getting started with machine learning.

**Jason Brownlee** November 27, 2016 at 10:21 am #

REPLY ↗

You're welcome, I'm glad you found the post useful Umar.

**Shiva Andure** March 18, 2019 at 3:08 pm**Your Start in Machine Learning**



Hello Jason,

```
from __future__ import division
models = []
models.append(("LR", LogisticRegression(solver='liblinear', multi_class='ovr')))
models.append(("LDA", LinearDiscriminantAnalysis()))
models.append(("KNN", KNeighborsClassifier()))
models.append(("CART", DecisionTreeClassifier()))
models.append(("NB", GaussianNB()))
models.append(("SVM", SVC(gamma='auto')))

# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=7)
    cv_results = model_selection.cross_val_score(model, X, y, cv=kfold)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

I am getting error of "ZeroDivisionError: float

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)


**Jason Brownlee** March 19, 2019 at 10:15 pm #

Sorry to hear that, I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Mike P** November 30, 2016 at 6:29 pm #

REPLY ↗

Hi Jason,

Really nice tutorial. I had one question which has had me confused. Once you chose your best model, (in this instance KNN) you then train a new model to be used to make predictions against the validation set. Should one not perform K-fold cross-validation on this model to ensure we don't overfit?

If this is correct how would you implement this, from my understanding `cross_val_score` will not allow one to generate a confusion matrix.

I think this is the only thing that I have struggled with in using scikit learn if you could help me it would be much appreciated?

Your Start in Machine Learning



**Jason Brownlee** December 1, 2016 at 7:26 am #

REPLY ↗

Hi Mike. No.

Cross-validation is just a method to estimate the skill of a model on new data. Once you have the estimate you can get on with things, like confirming you have not fooled yourself (hold out validation dataset) or make predictions on new data.

The skill you report is the cross val skill with the mean and stdev to give some idea of confidence or spread.

Does that make sense?



**Mike** December 2, 2016 at 1:30 am #

Hi Jason,

Thanks for the quick response. So to make sure I understand, you are estimating the skill of a model (mean of cross-validation) for a particular model.

Once you have this information you can just go ahead and train the model on the training set and test it against the validation set.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 2, 2016 at 8:17 am #

REPLY ↗

Hi Mike. Correct.

Additionally, if the validation result confirms your expectations, you can go ahead and train the model on all data you have including the validation dataset and then start using it in production.

This is a very important topic. I think I'll write a post about it.



**Sahana Venkatesh** November 30, 2016 at 8:15 pm #

REPLY ↗

This is amazing 😊 You boosted my morale



**Jason Brownlee** December 1, 2016 at 7:26 am #

REPLY ↗

I'm so glad to hear that Sahana.

## Your Start in Machine Learning



**Jhon** November 30, 2016 at 8:27 pm #

REPLY ↗

Hi

while doing data visualization and running commands dataset.plot(.....) i am having the following error.kindly tell me how to fix it

```
array([[],  
[],  
[],  
[]], dtype=object)
```



**Jason Brownlee** December 1, 2016 at 7:28 am #

Looks like no data Jhon. It also looks like

Are you running in a notebook or on the command  
command line).



**Brendon A. Kay** December 1, 2016 at 4:20 am #

Hi Jason,

Great tutorial. I am a developer with a computer science degree and mathematics, although I don't quite have the academic background for the latter except for what was required in college. So, this website has really sparked my interest as it has allowed me to learn the field in sort of the "opposite direction".

I did notice when executing your code that there was a deprecation warning for the `sklearn.cross_validation` module. They recommend switching to `sklearn.model_selection`.

When switching the modules I adjusted the following line...

```
kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
```

to...

```
kfold = model_selection.KFold(n_folds=num_folds, random_state=seed)
```

... and it appears to be working okay. Of course, I had switched all other instances of `cross_validation` as well, but it seemed to be that the `KFold()` method dropped the `n` (number of instances) parameter, which caused a runtime error. Also, I dropped the `num_instances` variable.

I could have missed something here, so please let me know if this is not a valid replacement, but thought I'd share!

Once again, great website!

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

Your Start in Machine Learning



**Jason Brownlee** December 1, 2016 at 7:33 am #

REPLY ↗

Thanks for the support and the kind words Brendon. I really appreciate it (you made my day)!

Yes, the API has changed/is changing and your updates to the tutorial look good to me, except I think `n_folds` has become `n_splits`.

I will update this example for the new API very soon.



**Brendon A. Kay** December 1, 2016 at 8:01 am #

REPLY ↗

😊 Now on to more tutorials for me!



**Jason Brownlee** December 2, 2016

You can access more here Brendon  
<http://machinelearningmastery.com/start-here/>



**Doug** March 9, 2018 at 5:56 am #

Jason, is everything on your website on that page? or is there another site map?  
 thanks!

P.S. your code ran flawlessly on my Jupyter Notebook fwiw. Although I did get a different result with SVM coming out on top with 99.1667. So I ran the validation set with SVM and came out with 94 93 93 30 fwiw.



**Jason Brownlee** March 9, 2018 at 6:29 am #

No, not everything, just a small and useful sample.

Yes, machine learning algorithms are stochastic, learn more here:  
<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Doug** March 9, 2018 at 6:46 am #

Thanks. I actually just read that article. Very helpful.

Your Start in Machine Learning



**Sergio** December 1, 2016 at 3:41 pm #

REPLY ↗

I'm still having a little trouble understanding step 5.1. I'm trying to apply this tutorial to a new data set but, when I try to evaluate the models from 5.3 I don't get a result.



**Jason Brownlee** December 2, 2016 at 8:13 am #

REPLY ↗

What is the problem exactly Sergio?

Step 5.1 should create a validation dataset. You can confirm the dataset by printing it out.

Step 5.3 should print the result of each algorithm a

Perhaps check for a copy-paste error or something



**sergio** December 2, 2016 at 9:13 am #

Does this tutorial work the exact same way for the  
Hello World dataset

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 3, 2016 at 8:23 am #

The project template is quite transferable.

You will need to adapt it for your data and for the types of algorithms you want to test.



**Jean-Baptiste Hubert** December 11, 2016 at 12:17 am #

REPLY ↗

Hi Sir,

Thank you for the information.

I am currently a student, in Engineering school in France.

I am working on date mining project, indeed, I have a many date ( 40Go ) about the price of the stocks of many companies in the CAC40.

My goal is to predict the evolution of the yields and I think that Neural Network could be useful.

My idea is : I take for X the yields from "t=0" to "t=n" and for Y the yields from "t=1 to t=n" and the program should find a relation between the data.

Is that possible ? Is it a good way in order to predict the evolution of the yield ?

Thank you for your time

Hubert

Jean-Baptiste

Your Start in Machine Learning



**Jason Brownlee** December 11, 2016 at 5:24 am #

REPLY ↗

Hi Jean-Baptiste, I'm not an expert in finance. I don't know if this is reasonable, sorry.

This post might help with phrasing your time series problem for supervised learning:

<http://machinelearningmastery.com/time-series-forecasting-supervised-learning/>



**Ernest Bonat** December 15, 2016 at 5:33 pm #

REPLY ↗

Hi Jason,

If I include an new item in the models array as:

```
models.append('LNR – Linear Regression', LinearReg
```

with the library:

```
from sklearn.linear_model import LinearRegression
```

I got an error in the \sklearn\utils\validation.py", line 52  
y = y.astype(np.float64)

as:

ValueError: could not convert string to float: 'Iris-setos

Let me know best to fix that! As you can see from my code, I would like to include the Linear Regression algorithms in my array model too!

Thank you for your help,

Ernest

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 16, 2016 at 5:39 am #

REPLY ↗

Hi Ernest, it is a classification problem. We cannot use LinearRegression.

Try adding another classification algorithm to the list.



**oumaima** December 9, 2017 at 11:29 am #

REPLY ↗

Hi Jason,

I am new to ML. need your help so i can run this.

```
>>> from matplotlib import pyplot
```

```
Traceback (most recent call last):
```

Your Start in Machine Learning

```
File "", line 1, in
File "c:\python27\lib\site-packages\matplotlib\pyplot.py", line 29, in
import matplotlib.colorbar
File "c:\python27\lib\site-packages\matplotlib\colorbar.py", line 32, in
import matplotlib.artist as artist
File "c:\python27\lib\site-packages\matplotlib\artist.py", line 16, in
from .path import Path
File "c:\python27\lib\site-packages\matplotlib\path.py", line 25, in
from . import _path, rcParams
'ImportError: DLL load failed: %1 n\x92est pas une application Win32 valide.\n'
```



**Jason Brownlee** December 10, 2017

Sorry, I have not seen that error before. It may be specific to your environment:  
<https://machinelearningmastery.com/setup-machine-learning-anaconda/>



**vanshika gupta** May 2, 2018 at 7:44am

hello oumaima,  
 i am also facing the same error? were you able to solve your error? now? please help.



## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Gokul Iyer** December 20, 2016 at 2:29 pm #

REPLY ↗

Great tutorial! Quick question, for the when we create the models, we do `models.append(name of algorithm, algorithm function)`, is `models` an array? Because it seems like a dictionary since we have a key-value mapping (algorithm name, and algorithm function). Thank you!



**Jason Brownlee** December 20, 2016 at 2:47 pm #

REPLY ↗

It is a list of tuples where each tuple contains a string name and a model object.



**Sasanka ghosh** December 21, 2016 at 4:55 am #

REPLY ↗

Hi Jason /any Gurus ,  
 Good post and will follow it but my question may be little off topic.  
 Asking this question as i am a data modeller /aspiring

Your Start in Machine Learning

I feel as Guru/Gurus you can clarify my doubt. The question is at the end .

In current Data management environment

1. Data architecture /Physical implementation and choosing appropriate tools,back end,storage,no sql,SQL, MPP, sharding, columnar ,scale up/out ,distributed processing etc .
2. In addition to DB based procedural languages proficiency at least one of the following i.e. Java/Python/Scala etc.
3. Then comes this AI,Machine learning ,neural Networks etc .

My question is regarding point 3 .

I believe those are algorithms which needs deep functional knowledge and years of experience to add any value to business .

Those are independent of data models and it's ,physical not data architecture domain .

If i take your above example say now 10k users trying be Data architects a domain and point 3 will be business between them to some extent .

Data Architect need not to be hands on/proficient in all Data architects job is not to invent business logic but i Business users/Analysts .

Am i correct in my assumption as i find the certain things expectations/benchmarks should be set right?

Regards

sasanka ghosh

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 21, 2016 at 8:46 am #

REPLY ↗

Hi Sasanka, sorry, I don't really follow.

Are you able to simplify your question?



**Sasanka ghosh** December 21, 2016 at 9:25 pm #

REPLY ↗

Hi Jason ,

Many thanks that u bothered to reply .

Tried to rephrase and concise but still it is verbose . apologies for that.

Is it expected from a data architect to be algorithm expert as well as data model/database expert?

Algorithms are business centric as well as specific times.

Your Start in Machine Learning

Giving u an example i.e. SHORTEST PATH ( take it as just an example in making my point)

An organization is providing an app to provide that service .

CAVEAT: Someone may say from comp science dept that it is the basic thing u learn but i feel it is still an algorithm not a data structure .

if we take the above scenario in simplistic term the requirement is as follows

1. there will be say million registered users
2. one can say at least 10 % are using the app same time
3. any time they can change their direction as per contingency like a military op so dumping the partial weighted graph to their device is not an option i.e. users will be connected to main server/server cluster.
4. the challenge is storing the spatial data in DP in correct data model  
scale out ,fault tolerance .
5. implement the shortest path algo and display dynamically.

My question is can a data architect work on the algorithm but have sufficient knowledge in other areas to help him/her to implement ?

I m asking this question as now a days people learning ,NLP,Data scientist etc. and the scenario is that they have minimal overlapping but continuous discussion.

I feel Algorithms are pure science that is a separate field. But to implement it in large scale Scientists/practitioners need to have minimal overlapping but continuous discussion.

Last but not the least if i make some sense what is the learning curve should i follow to try to be a data architect in unstructured data in general

regards  
sasanka ghosh

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 22, 2016 at 6:35 am #

REPLY ↗

Really this depends on the industry and the job. I cannot give you good advice for the general case.

You can get valuable results without being an expert, this applies to most fields.

Algorithms are a tool, use them as such. They can also be a science, but we practitioners don't have the time.

I hope that helps.

Your Start in Machine Learning



**Sasanka ghosh** December 22, 2016 at 7:00 pm #

Thanks Jsaon.

I appreciate your time and response .

I just wanted to validate from a real techie/guru like u as the confusion or no perfect answer are being exploited by management/HR to their own advantage and practice use and throw policy or make people sycophants/redundant without following the basic management principle,

The tech guys except " few geniuses" are always toiling and management is opening the cork, enjoying at the same time .

Regards  
sasanka ghosh



**Raveen Sachintha** December 21, 2016 at 8:51 pm

Hello Jason,

Thank you very much for these tutorials, i am new to ML and i am trying to work on my project to get started with rather than reading and reading

One question, when i tried this i got the highest accuracy for SVM

LR: 0.966667 (0.040825)  
LDA: 0.975000 (0.038188)  
KNN: 0.983333 (0.033333)  
CART: 0.983333 (0.033333)  
NB: 0.975000 (0.053359)  
SVM: 0.991667 (0.025000)

so i decided to try that out too,,

```
svm = SVC()
svm.fit(X_train, Y_train)
prediction = svm.predict(X_validation)
```

these were my results using SVM,

0.933333333333  
[[ 7 0 0]  
[ 0 10 2]  
[ 0 0 11]]  
precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7  
Iris-versicolor 1.00 0.83 0.91 12  
Iris-virginica 0.85 1.00 0.92 11

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

avg / total 0.94 0.93 0.93 30

I am still learning to read these results, but can you tell me why this happened? why did i get high accuracy for SVM instead of KNN?? have i done anything wrong? or is it possible?



**Jason Brownlee** December 22, 2016 at 6:33 am #

REPLY ↩

The results reported are a mean estimated score with some variance (spread).

It is an estimate on the performance on new data.

When you apply the method on new data, the performance may be in that range. It may be lower if the method has overfit the training data.

Overfitting is a challenge and developing a robust ourselves during model development is important

I hope that helps as a start.



**inzar** December 25, 2016 at 7:04 am #

i want to buy your book.  
i try this tutorial and the result is very awesome  
i want to learn from you  
thanks....

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 26, 2016 at 7:41 am #

REPLY ↩

Thanks inzar.

You can see all of my books and bundles here:  
<http://machinelearningmastery.com/products>



**lou** December 25, 2016 at 7:29 am #

REPLY ↩

Why the leading comma in X = array[:,0:4]?



**Jason Brownlee** December 26, 2016 at 7:42 am #

REPLY ↩

This is Python array notation for [rows,col]

Your Start in Machine Learning

Learn more about slicing arrays in Python here:  
<http://structure.usc.edu/numarray/node26.html>



**Thinh** December 26, 2016 at 5:05 am #

REPLY ↗

In 1.2 , should warn to install scikit-learn



**Jason Brownlee** December 26, 2016 at 7:49 am #

REPLY ↗

Thanks for the note.

Please see section 1.1 Install SciPy Libraries where

There are 5 key libraries that you will need to install



**Tijo L. Peter** December 28, 2016 at 10:34 pm #

Best ML tutorial for Python. Thank you, Jason

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 29, 2016 at 7:17 am #

REPLY ↗

Thanks!



**boso** December 29, 2016 at 12:38 am #

REPLY ↗

when i tried run, i have error message" TypeError: Empty 'DataFrame': no numeric data to plot"  
 help me



**Jason Brownlee** December 29, 2016 at 7:18 am #

REPLY ↗

Sorry to hear that.

Perhaps check that you have loaded the data as you expect and that the loaded values are numeric and not strings. Perhaps print the first few rows: `print(df.head(5))`

Your Start in Machine Learning



**baso** December 29, 2016 at 1:05 pm #

REPLY ↗

thanks very much Jason for your time

it worked. these tutorial very help for me. im new in Machine learning, but may you explain to me about your simple project above? because i did not see X\_test and target  
regard in advance



**Jason Brownlee** December 30, 2016 at 5:49 am #

REPLY ↗

Glad to hear it baso!



**Andrea** January 5, 2017 at 1:42 am #

Thank you for sharing this. I bumped into some issues. Eventually, yo get all dependencies installed on MacOs:

```
brew install python
pip install --user numpy scipy matplotlib ipython jupyter
export PATH=$PATH:~/Library/Python/2.7/bin
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** January 5, 2017 at 9:21 am #

REPLY ↗

Thanks for sharing Andrea.

I'm a macports guy myself, here's my recipe:

```
1 1. Install XCode and XCode Command Line Tools
2   Use the "Mac App Store" application
3   xcode-select --install
4 2. Install Macports
5   https://guide.macports.org/chunked/installing.macports.html
6 3. Install a SciPy Environment
7   sudo port install py27-numpy py27-scipy py27-matplotlib py27-ipython +notebook py27-sympy
8   sudo port select --set py-sympy py27-sympy
9   sudo port select --set cython cython27
10  sudo port select --set ipython py27-ipython
11  sudo port select --set ipython2 py27-ipython
12  sudo port select --set python python27
13  sudo port select --set python2 python27
14  sudo port select --set pip pip27
15 4. Install scikit-learn
16   sudo pip install -U scikit-learn
```

Your Start in Machine Learning



**Sohib** January 6, 2017 at 6:26 pm #

REPLY ↗

Hi Jason,

I am following this page as a beginner and have installed Anaconda as recommended. As I am on win 10, I installed Anaconda 4.2.0 For Windows Python 2.7 version (x64) and I am using Anaconda's Spyder (python 2.7) IDE.

I checked all the versions of libraries (as shown in 1.2 Start Python and Check Versions) and got results like below:

Python: 2.7.12 |Anaconda 4.2.0 (64-bit)| (default, Jun 29 2016, 11:07:13) [MSC v.1500 64 bit (AMD64)]

scipy: 0.18.1

numpy: 1.11.1

matplotlib: 1.5.3

pandas: 0.18.1

sklearn: 0.17.1

At the 2.1 Import libraries section, I imported all of them. In the 2.2 Load Dataset. But when I run it, it doesn't show anything.

Traceback (most recent call last):

```
File "C:\Users\gachon\spyder\temp.py", line 4, in 
from sklearn import model_selection
```

```
ImportError: cannot import name model_selection
```

Below is my code snippet:

```
import pandas
from pandas.tools.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)
print(dataset.shape)
```

When I delete "from sklearn import model\_selection" line I get expected results (150, 5).

Am I missing something here?

Thank you for your time and endurance!

## Your Start in Machine Learning



You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** January 7, 2017 at 8:23 am #

REPLY ↗

Hi Sohib,

You must have scikit-learn version 0.18 or higher installed.

Perhaps Anaconda has documentation on how to update sklearn?



**Sohib** January 10, 2017 at 12:15 pm #

REPLY ↗

Thank you for reply.

I updated scikit-learn version to 0.18.1 and it h  
The error disappeared, the result is shown, but

'import sitecustomize' failed; use -v for traceba  
is executed above the result.

I tried to find out why, but apparently I might n  
Is it going to be a problem in my further steps?  
How to solve this?

Thank you in advance!

## Your Start in Machine Learning



You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** January 11, 2017 at 9:25 am #

REPLY ↗

I'm glad to hear it fixed your problem.

Sorry, I don't know what "import sitecustomize" is or why you need it.



**Vishakha** January 7, 2017 at 10:10 pm #

REPLY ↗

Can i get the same tutorial with java



**Abhinav** January 8, 2017 at 8:27 pm #

REPLY ↗

Hi Jason,

Nice tutorial.

In univariate plots, you mentioned about gaussian distribution

Your Start in Machine Learning

According to the univariate plots, sepat-width had gaussian distribution. You said there are 2 variables having gaussain distribution. Please tell the other.

Thanks



**Jason Brownlee** January 9, 2017 at 7:49 am #

REPLY ↗

The distribution of the others may be multi-modal. Perhaps a double Gaussian.



**Thinh** January 13, 2017 at 5:07 am #

Hi, Jason. Could you please tell me the reaso



**Jason Brownlee** January 13, 2017 at 9:16 am #

Hi Thinh,

No reason other than it is an easy algorithm to run tutorial.



**Scott P** January 13, 2017 at 10:25 pm #

REPLY ↗

Hi Jason,

I'm trying to use this code with the KDD Cup '99 dataset, and I am having trouble with LabelEncoding my dataset in to numerical values.

```
#Modules
import pandas
import numpy
from pandas.tools.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn import cross_validation
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscrim
from sklearn.naive_bayes import GaussianNB
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

from sklearn.svm import SVC
from sklearn.preprocessing import LabelEncoder
#new
from collections import defaultdict
#
#Load KDD dataset
data_set = "NSL-KDD/KDDTrain+.txt"
names =
['duration', 'protocol_type', 'service', 'flag', 'src_bytes', 'dst_bytes', 'land', 'wrong_fragment', 'urgent', 'hot', 'num_failed_logins', 'logged_in', 'num_compromised', 'su_attempted', 'num_root', 'num_file_creations', 'num_shells', 'num_access_files', 'num_outbound_cmds', 'is_host_login', 'is_guest_login', 'count', 'srv_count', 'serror_rate', 'srv_serror_rate', 'rerror_rate', 'srv_rerror_rate', 'dst_host_count', 'dst_host_srv_count', 'dst_host_same_srv_rate', 'dst_host_diff_srv_rate', 'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate', 'dst_host_srv_rate', 'dst_host_srv_rerror_rate', 'class']

#Diabetes Dataset
#data_set = "Datasets/pima-indians-diabetes.data"
#names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
#data_set = "Datasets/iris.data"
#names = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class']
dataset = pandas.read_csv(data_set, names=names)

array = dataset.values
X = array[:,0:40]
Y = array[:,40]

label_encoder = LabelEncoder()
label_encoder = label_encoder.fit(Y)
label_encoded_y = label_encoder.transform(Y)

validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = cross_validation.train_test_split(X, label_encoded_y, test_size=validation_size, random_state=seed)

# Test options and evaluation metric
num_folds = 7
num_instances = len(X_train)
seed = 7
scoring = 'accuracy'

# Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))

```

Your Start in Machine Learning

```

models.append('NB', GaussianNB())
models.append('SVM', SVC())

# evaluate each model in turn
results = []
names = []

for name, model in models:
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)

msg = "%s: %f (%f)" % (name, cv_results.mean()*100, cv_results.std()*100)#multiplying by 100 to show
percentage
print(msg)

# Compare Algorithms
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(Y)
plt.show()

```

Am I doing something wrong with the LabelEncoder part?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)



**MegO\_Bonus** June 4, 2017 at 7:15 pm #

[REPLY](#) ↩

Hi. Change all symbols like “ to ” and ‘ to ’. LabelEncoder will work correctly but not all network. I try to create a neural network for NSL-KDD too. Have you any good examples?



**Jason Brownlee** June 5, 2017 at 7:40 am #

[REPLY](#) ↩

What is “NSL-KDD”?



**bugtime** December 10, 2017 at 8:22 pm #

[REPLY](#) ↩

Hello Jason,

Please see [https://github.com/defcom17/NSL\\_KDD](https://github.com/defcom17/NSL_KDD)

**Jason Brownlee** December 11

[Your Start in Machine Learning](#)



I'm not familiar with this, sorry.



**Rajnish** July 17, 2019 at 8:21 am #

REPLY ↩

How come it is concluded that KNN algorithm is accurate model when mean value for SVM algorithm is closer to 1 in comparison to KNN ?



**Jason Brownlee** July 17, 2019 at 8:32 am #

REPLY ↩

Either algorithm would be effective or



**Dan** January 14, 2017 at 4:56 am #

Hi, I'm running a bit of a different setup than you.

The modules and version of python I'm using are more:

Python: 3.5.2 |Anaconda 4.2.0 (32-bit)| (default, Jul 5 2016, 11:47:40)  
scipy: 0.18.1  
numpy: 1.11.3  
matplotlib: 1.5.3  
pandas: 0.19.2  
sklearn: 0.18.1

And I've gotten SVM as the best algorithm in terms of accuracy at 0.991667 (0.025000).

Would you happen to know why this is, considering more recent versions?

I also happened to get a rather different boxplot but I'll leave it at what I've said thus far.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** January 15, 2017 at 5:26 am #

REPLY ↩

Hi Dan,

You may get differing results for a variety of reasons. Small changes in the code will affect the result. This is why we often report mean and stdev algorithm performance rather than one number, to give a range of expected performance.

This post on randomness in ml algorithms might also help:

<http://machinelearningmastery.com/randomness-in-machine-learning/>

Your Start in Machine Learning



**Duncan Carr** January 17, 2017 at 1:44 am #

REPLY ↗

Hi Jason

I can't tell you how grateful I am ... I have been trawling through lots of ML stuff to try to get started with a "toy" example. Finally I have found the tutorial I was looking for. Anaconda had old sklearn: 0.17.1 for Windows – which caused an error "ImportError: cannot import name 'model\_selection'". That was fixed by running "pip install -U scikit-learn" from the Anaconda command-line prompt. Now upgraded to 0.18. Now everything in your imports was fine.

All other tutorials were either too simple or too complicated. Usually the latter!

Thank you again 😊



**Jason Brownlee** January 17, 2017 at 7:39 am

Glad to hear it Duncan.

Thanks for the tip for Anaconda uses.

I'm here to help if you have questions!



**Malathi** January 17, 2017 at 3:13 am #

Hi Jason,

Wonderful service. All of your tutorials are very helpful to me. Easy to understand.

Expecting more tutorials on deep neural networks.

Malathi



**Jason Brownlee** January 17, 2017 at 7:40 am #

REPLY ↗

You're very welcome Malathi, glad to hear it.



**Duncan Carr** January 17, 2017 at 7:32 pm #

REPLY ↗

Hi Jason

I managed to get it all working – I am chuffed to bits.

Your Start in Machine Learning

I get exactly the same numbers in the classification report as you do ... however, when I changed both seeds to 8 (from 7), then ALL of the numbers end up being 1. Is this good, or bad? I am a bit confused.

Thanks again.



**Jason Brownlee** January 18, 2017 at 10:14 am #

REPLY ↗

Well done Duncan!

What do you mean all the numbers end up being one?



**Duncan Carr** January 18, 2017 at 8:02 pm #

Hi Jason

I've output the "accuracy\_score", "confusion\_matrix" & getting a perfect score with seed=9? Many thanks.

(seed=7)

0.9

```
[[10 0 0]
 [ 0 8 1]
 [ 0 2 9]]
```

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 10

Iris-versicolor 0.80 0.89 0.84 9

Iris-virginica 0.90 0.82 0.86 11

avg / total 0.90 0.90 0.90 30

(seed=9)

1.0

```
[[13 0 0]
 [ 0 9 0]
 [ 0 0 8]]
```

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 13

Iris-versicolor 1.00 1.00 1.00 9

Iris-virginica 1.00 1.00 1.00 8

avg / total 1.00 1.00 1.00 30

(seed=10)

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

0.9666666666666666

```
[[10 0 0]
 [ 0 12 1]
 [ 0 0 7]]
```

precision recall f1-score support

	Iris-setosa	Iris-versicolor	Iris-virginica	avg / total
precision	1.00	1.00	1.00	10
recall	1.00	0.92	0.96	13
f1-score	0.88	1.00	0.93	7
support	50	50	50	150

avg / total 0.97 0.97 0.97 30



**Jason Brownlee** January 19, 2017 at 7:31 am #

Random chance. This is why it is a good idea to report mean and standard deviation scores.

More on randomness in machine learning here:  
<http://machinelearningmastery.com/randomness-in-machine-learning/>



**shivani** January 20, 2017 at 8:40 pm #

from sklearn import model\_selection  
 showing Import Error: can not import model\_selection



**Jason Brownlee** January 21, 2017 at 10:25 am #

REPLY ↗

You need to update your version of sklearn to 0.18 or higher.



**Jim** January 22, 2017 at 5:06 pm #

REPLY ↗

Jason

Excellent Tutorial. New to Python and set a New Years Resolution to try to understand ML. This tutorial was a great start.

I struck the issue of the sklearn version. I am using Ubuntu 16.04LTS which comes with python-sklearn version 0.17. To update to latest I used the site:

[http://neuro.debian.net/install\\_pkg.html?p=python-sklearn](http://neuro.debian.net/install_pkg.html?p=python-sklearn)

Which gives the commands to add the neuro repository and pull down the 0.18 version

## Your Start in Machine Learning

Also I would like to note there is an error in section 3.1 Dimensions of the Dataset. Your text states 120 Instances when in fact 150 are returned, which you have in the Printout box.

Keep up the good work.

Jim



**Jason Brownlee** January 23, 2017 at 8:37 am #

REPLY ↗

I'm glad to hear you worked around the version issue Jim, nice work!

Thanks for the note on the typo, fixed!



**Raphael** January 23, 2017 at 4:15 pm #

hi Jason.nice work here. I'm new to your blog



**Jason Brownlee** January 24, 2017 at 11:01 am #

Hi Raphael,

The y-axis in the box-and-whisker plots are the sc



**Kayode** January 23, 2017 at 8:42 pm #

REPLY ↗

Thank you for this wonderful tutorial.



**Jason Brownlee** January 24, 2017 at 11:03 am #

REPLY ↗

You're welcome Kayode.



**Raphael** January 26, 2017 at 2:28 am #

REPLY ↗

hi Jason,

In this line

```
dataset.groupby('class').size()
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

what other variable other than size could I use? I changed size with count and got something similar but not quite. I got key errors for the other stuffs I tried. Is size just a standard command?



**Jason Brownlee** January 26, 2017 at 4:48 am #

REPLY ↗

Great question Raphael.

You can learn more about Pandas groupby() here:

<http://pandas.pydata.org/pandas-docs/stable/groupby.html>



**Scott** January 26, 2017 at 10:35 pm #

Jason,

I'm trying to use a different data set (KDD CUP 99') which after modifying "names" and the array to account for the error of: "cannot convert string to a float".

In my data set, there are 3 columns that are text and they LabelEncoding but it gives me the same error, do you

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** January 27, 2017 at 12:08 pm #

REPLY ↗

Hi Scott,

If the values are indeed strings, perhaps you can use a method that supports strings instead of numbers, perhaps like a decision tree.

If there are only a few string values for the column, a label encoding as integers may be useful.

Alternatively, perhaps you could try removing those string features from the dataset.

I hope that helps, let me know how you go.



**Weston Gross** January 31, 2017 at 10:41 am #

REPLY ↗

I would like a chart to see the grand scope of everything for data science that python can do.

You list 6 basic steps. For example in the visualizing step, I would like to know what all the charts are, what they are used for, and what python library it comes from.

I am extremely new to all this, and understand that some steps have to happen for example

1. Get Data
2. Validate Data

Your Start in Machine Learning

- 3. Missing Data
- 4. Machine Learning
- 5. Display Findinds

So for missing data, there are techniques to restore the data, what are they and what libraries are used?



**Jason Brownlee** February 1, 2017 at 10:36 am #

REPLY ↗

You can handle missing data in a few ways such as:

1. Remove rows with missing data.
2. Impute missing data (e.g. use the Imputer class)
3. Use methods that support missing data (e.g. de

I hope that helps.



**Mohammed** February 1, 2017 at 1:11 am #

Hi Jason,

I am a Non Tech Data Analyst and use SPSS extensively.

I understand the above example very easily.

I want to work on Search – Language Translation and

Whats the best way forward ...

Do you also provide Skype Training / Project Mentoring..

Thanks in advance.

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 1, 2017 at 10:51 am #

REPLY ↗

Thanks Mohammed.

Sorry, I don't have good advice for language translation applications.



**Mohammed** February 1, 2017 at 1:14 am #

REPLY ↗

I dont have any Development / Coding Background.

However, following your guidelines I downloaded SciPy and tested the code.

Everything worked perfectly fine.

## Your Start in Machine Learning

Looking forward to go all in...



**Jason Brownlee** February 1, 2017 at 10:51 am #

REPLY ↗

I'm glad to hear that Mohammed



**Purvi** February 1, 2017 at 7:31 am #

REPLY ↗

Hi Jason,

I am new to Machine learning and am trying out the tu

```
>>> import sys
>>> print('Python: {}'.format(sys.version))
Python: 2.7.10 (default, Jul 13 2015, 12:05:58)
[GCC 4.2.1 Compatible Apple LLVM 6.1.0 (clang-602.0.53)]
>>> import scipy
>>> print('scipy: {}'.format(scipy.__version__))
scipy: 0.18.1
>>> import numpy
>>> print('numpy: {}'.format(numpy.__version__))
numpy: 1.12.0
>>> import matplotlib
>>> print('matplotlib: {}'.format(matplotlib.__version__))
matplotlib: 2.0.0
>>> import pandas
>>> print('pandas: {}'.format(pandas.__version__))
pandas: 0.19.2
>>> import sklearn
>>> print('sklearn: {}'.format(sklearn.__version__))
sklearn: 0.18.1
```

When I try to load the iris dataset, it loads up fine and prints dataset.shape but then my python interpreter hangs. I tried it out 3-4 times and everytime it hangs after I run couple of commands on dataset.

```
>>> url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
>>> names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
>>> dataset = pandas.read_csv(url, names=names)
>>> print(dataset.shape)
(150, 5)
>>> print(dataset.head(20))
sepal-length sepal-width petal-length petal-width class
0 5.1 3.5 1.4 0.2 Iris-setosa
1 4.9 3.0 1.4 0.2 Iris-setosa
2 4.7 3.2 1.3 0.2 Iris-setosa
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

3 4.6 3.1 1.5 0.2 Iris-setosa
4 5.0 3.6 1.4 0.2 Iris-setosa
5 5.4 3.9 1.7 0.4 Iris-setosa
6 4.6 3.4 1.4 0.3 Iris-setosa
7 5.0 3.4 1.5 0.2 Iris-setosa
8 4.4 2.9 1.4 0.2 Iris-setosa
9 4.9 3.1 1.5 0.1 Iris-setosa
10 5.4 3.7 1.5 0.2 Iris-setosa
11 4.8 3.4 1.6 0.2 Iris-setosa
12 4.8 3.0 1.4 0.1 Iris-setosa
13 4.3 3.0 1.1 0.1 Iris-setosa
14 5.8 4.0 1.2 0.2 Iris-setosa
15 5.7 4.4 1.5 0.4 Iris-setosa
16 5.4 3.9 1.3 0.4 Iris-setosa
17 5.1 3.5 1.4 0.3 Iris-setosa
18 5.7 3.8 1.7 0.3 Iris-setosa
19 5.1 3.8 1.5 0.3 Iris-setosa
>>> print(data)

```

It does not let me type anything further.  
I would appreciate your help.

Thanks,  
Purvi

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Jason Brownlee** February 1, 2017 at 10:55 am #

REPLY ↗

Hi Purvi, sorry to hear that.

Perhaps you're able to comment out the first parts of the tutorial and see if you can progress?



**sam** February 5, 2017 at 9:24 am #

REPLY ↗

Hi Jason

i am planning to use python to predict customer attrition.I have current list of attrited customers with their attributes.I would like to use them as test data and use them to predict any new customers.Can you please help to approach the problem in python ?

my test data :

customer1 attribute1 attribute2 attribute3 ... attrited

my new data

customer N, attribute 1,..... ?

Thanks for your help in advance.

Your Start in Machine Learning



**Jason Brownlee** February 6, 2017 at 9:42 am #

REPLY ↩

Hi Sam, as a start, this process will help you clearly define and work through your predictive modeling problem:

<http://machinelearningmastery.com/start-here/#process>

I'm happy to answer questions as you work through the process.



**Kiran Prajapati** February 7, 2017 at 6:31 pm #

X

Hello Sir, I want to check my data is how many Taluka , Total\_yield, Rain(mm) , types\_of\_soil

Nasik 12555 63.0 dark black

Igatpuri 1560 75.0 shallow

So on,

first, I have to check data is accurate or not, and next model.

Here is my model Total\_yield = Rain + types\_of\_soil

I use 0 and 1 binary variable for types\_of\_soil.

can you please help me, how to calculate data is accurate ? How many % ? and how to find predicted yield ?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 8, 2017 at 9:33 am #

REPLY ↩

I'm not sure I understand Kiran.

This process will help you describe and work through your predictive modeling project:

<http://machinelearningmastery.com/start-here/#process>



**Saby** February 15, 2017 at 9:11 am #

REPLY ↩

# Load dataset

```
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)
```

The dataset should load without incident.

## Your Start in Machine Learning

If you do have network problems, you can download the iris.data file into your working directory and load it using the same method, changing url to the local file name.

I am a very beginner python learner(trying to learn ML as well), I tried to load data from my local file but could not be successful. Will you help me out how exactly code should be written to open the data from local file.



**Jason Brownlee** February 15, 2017 at 11:39 am #

REPLY ↩

Sure.

Download the file as iris.data into your current working directory (where you are running the code from).

Then load it as:

```
1 dataset = pandas.read_csv('iris.data', n
```



**ant** February 15, 2017 at 9:54 pm #

Hi, Jason, first of all thank so much for this article.

Just for curiosity I have computed all the values obtained from the data set. I get 1.57500 instead of 1.60000. unsuccessfully. Is there an explanation? Tnx

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 16, 2017 at 11:07 am #

REPLY ↩

Not sure, perhaps you could look into the Pandas source code?



**ant** February 17, 2017 at 12:23 am #

REPLY ↩

OK, I will do.



**jacques** February 16, 2017 at 4:42 pm #

REPLY ↩

Hi Jason

I don't quite follow the KFOLD section ?

We started off with 150 data-entries(rows)

Your Start in Machine Learning

We then use a 80/20 split for validation/training that leaves us with 120

The split 10 boggles me ??

Does it take 10 items from each class and train with 9 ? what does the other 1 left do then ?



**Jason Brownlee** February 17, 2017 at 9:52 am #

REPLY ↗

Hi jacques,

The 120 records are split into 10 folds. The model is trained on the first 9 folds and evaluated on the records in the 10th. This is repeated so that each fold is given a chance to be the hold out set. 10 models are trained, 10 scores collected and we report the performance of the model on unseen data.

Does that help?



**Alhassan** February 17, 2017 at 4:02 pm #

I am trying to integrate machine learning into my project. Can you tell me how to do that using the guidelines you provided above?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 18, 2017 at 8:34 am #

REPLY ↗

I have not done this Alhassan.

Generally, I would advise developing a separate service that could be called using REST calls or similar.

If you are working on a prototype, you may be able to call out to a program or script from cgi-bin, but this would require careful engineering to be secure in a production environment.



**Simão Gonçalves** February 20, 2017 at 1:27 am #

REPLY ↗

Hi Jason! This tutorial was a great help, i'm truly grateful for this so thank you.

I have one question about the tutorial though, in the Scatterplot Matrix I can't understand how we can make the dots in the graphs whose variables have no relationship between them (like sepal-length with petal\_width).

Could you or someone explain that please? How do you make a dot that represents the relationship between a certain sepal\_length with a certain petal\_width?

Your Start in Machine Learning



**Jason Brownlee** February 20, 2017 at 9:30 am #

REPLY ↗

Hi Simão,

The x-axis is taken for the values of the first variable (e.g. sepal\_length) and the y-axis is taken for the second variable (e.g. petal\_width).

Does that help?



**Yopo** February 21, 2017 at 4:35 am #

REPLY ↗

you match each iris instance's length and number one is represented by a dot, and the dot's you take all these values and put them on a graph as you can see some in some of these plots the dot petal width – petal length graph it seems to be linearly related. hope this helped!



**Sébastien** February 20, 2017 at 9:34 pm #

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 21, 2017 at 9:34 am #

REPLY ↗

I'm glad you found it useful Sébastien.



**Raj** February 27, 2017 at 2:53 am #

REPLY ↗

Hi Jason,

I am new to ML & Python. Your post is encouraging and straight to the point of execution. Anyhow, I am facing below error when

```
>>> validataion_size = 0.20
>>> X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y,
test_size=validation_size, random_state = seed)
Traceback (most recent call last):
```

Your Start in Machine Learning

File "", line 1, in

NameError: name 'validation\_size' is not defined

What could be the miss out? I didn't get any errors in previous steps.

My Environment details:

OS: Windows 10

Python : 3.5.2

scipy : 0.18.1

numpy : 1.11.1

sklearn : 0.18.1

matplotlib : 0.18.1



**Jason Brownlee** February 27, 2017 at 5:54 am #

Hi Raj,

Double check you have the code from section "5.1" defined.

I hope that helps.



**Roy** March 2, 2017 at 7:38 am #

Hey Jason,

Can you please explain what precision, recall, f1-score, support actually refer to?

Also what the numbers in a confusion matrix refers to?

[ 7 0 0]

[ 0 1 1 1]

[ 0 2 9]]

Thanks.



**Jason Brownlee** March 2, 2017 at 8:24 am #

REPLY ↗

Hi Roy,

You can learn all about the confusion matrix in this post:

<http://machinelearningmastery.com/confusion-matrix-machine-learning/>

You can learn all about precision and recall in this article:

[https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Ahmed** December 25, 2017 at 2:21 am #

REPLY ↗

Hi Jason,

Thank you very much for your tutorial.

I am a little bit confused about the confusion matrix, because you are using a 3x3 matrix while it should be a 2x2 matrix.



**Jason Brownlee** December 25, 2017 at 5:25 am #

REPLY ↗

Learn more about the confusion matrix:  
<https://machinelearningmastery.com/confusion-matrix-machine-learning/>



**Ahmed** December 25, 2017 at 6:16 am #

X

Hi Jason,

Now I understand the meaning of your confusion matrix.  
Thank you and best regards.

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 26, 2017 at 5:11 am #

REPLY ↗

You're welcome.



**santosh** March 3, 2017 at 7:29 am #

REPLY ↗

what code should i use to load data from my working directory??



**Jason Brownlee** March 3, 2017 at 7:47 am #

REPLY ↗

This post will help you out Santosh:

<http://machinelearningmastery.com/load-machine-learning-data-python/>



**David** March 7, 2017 at 8:27 am #

REPLY ↗

Hi Jason,

Your Start in Machine Learning

I have a ValueError and i don't know how can i solve this problem

My problem like that,

ValueError: could not convert string to float: '2013-06-27 11:30:00.0000000'

Can u give some information about the fixing this problem?

Thank you



**Jason Brownlee** March 7, 2017 at 9:39 am #

REPLY ↗

It looks like you are trying to load a date-time column. Try to parse the date-time when loading or try removing it.



**Saugata De** March 8, 2017 at 6:11 am #

```
>>> for name, model in models:
... kfold=model_selection.Kfold(n_splits=10, random_state=seed)
... cv_results =model_selection.cross_val_score(model, X, y)
... results.append(cv_results)
... names.append(name)
... msg="%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
... print(msg)
...
...
```

After typing this piece of code, it is giving me this error. can you plz help me out Jason. Since I am new to ML, dont have so much idea about the error.

Traceback (most recent call last):

File "", line 2, in

AttributeError: module 'sklearn.model\_selection' has no attribute 'Kfold'

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Asad Ali** July 23, 2017 at 12:59 pm #

REPLY ↗

the KFold function is case-sensitive. It is " model\_selection.KFold(...)" not "model\_selection.Kfold(...)"  
update this line:  
kfold=model\_selection.KFold(n\_splits=10, random\_state=seed)



**ibtssam** February 12, 2018 at 9:17 pm #

REPLY ↗

THANK U

## Your Start in Machine Learning



**Ojas** March 10, 2017 at 10:58 am #

REPLY ↗

Hello Jason ,

Thanks for writing such a nice and explanatory article for beginners like me but i have one concern , i tried finding it out on other websites as well but could not come up with any solution.

Whatever i am writing inside the code editor (Jupyter Qtconsole in my case) , can this not be save as a .py file and shared with my other members over github maybe?. I found some hacks though but i have a thinking that there must be some proper way of sharing the codes written in the editor. , like without the outputs or plots in between.



**Jason Brownlee** March 11, 2017 at 7:55 am #

You can write Python code in a text editor or the command line as follows:

`1 python myfile.py`

Consider picking up a book on Python.



**manoj maracheea** March 11, 2017 at 9:37 pm #

Hello Jason,

Nice tutorials I done this today.

I didn't really understand everything, { I will follow your advice, will do it again, write all the question down, and use the help function.}

The tutorials just works, I take around 2 hours to do it typing every single line. install all the dependencies, run on each blocks types, to check.

Thanks, I be visiting your blogs, time to time.

Regards,



**Jason Brownlee** March 12, 2017 at 8:23 am #

REPLY ↗

Well done, and thanks for your support.

Post any questions you have as comments or email me using the “contact” page.

Your Start in Machine Learning



**manoj maracheea** March 11, 2017 at 9:38 pm #

REPLY ↗

Just I am a beginner too, I am using Visual studio code.

Look good.



**Vignesh R** March 13, 2017 at 9:59 pm #

REPLY ↗

What exactly is confusion matrix?



**Jason Brownlee** March 14, 2017 at 8:18 am #

Great question, see this post:

<http://machinelearningmastery.com/confusion-matrices-for-classification-with-python/>



**Dan R.** March 14, 2017 at 7:09 am #

Can I ask what is the reason of this problem?  
(In my code is just the section, where I Import all the n  
I have all libraries up to date, but it still gives me this e

File "C:\Users\64dri\Anaconda3\lib\site-packages\sklearn\model\_selection\\_search.py", line 32, in  
from ..utils.fixes import rankdata

ImportError: cannot import name 'rankdata'

```
(scipy: 0.18.1
numpy: 1.11.1
matplotlib: 1.5.3
pandas: 0.18.1
sklearn: 0.17.1)
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 14, 2017 at 8:31 am #

REPLY ↗

Sorry, I have not seen this issue Dan, consider searching or posting to StackOverflow.



**Cameron** March 15, 2017 at 5:28 am #

REPLY ↗

Jason,

Your Start in Machine Learning

You're a rockstar, thank you so much for this tutorial and for your books! It's been hugely helpful in getting me started on machine learning. I was curious, is it possible to add a non-number property column, or will the algorithms only accept numbers?

For example, if there were a "COLOR" column in the iris dataset, and all Iris-Setosa were blue. how could I get this program to accept and process that COLOR column? I've tried a few things and they all seem to fail.



**Jason Brownlee** March 15, 2017 at 8:16 am #

REPLY ↗

Great question Cameron!

sklearn requires all input data to be numbers.

You can encode labels like colors as integers and

Further, you can convert the integers to a binary encoding which is suitable if there is no ordinal relationship between



**Cameron** March 15, 2017 at 2:19 pm #

Jason, thanks so much for replying! To hot encoding I assume you mean (continuing the color (R,O,Y,G,B,V) and for each flower putting other colors?)

That's feasible for 6 colors (adding six columns) but how would I manage if I wanted to choose between 100 colors or 1000 colors? Are there other libraries that could help deal with that?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 16, 2017 at 7:58 am #

REPLY ↗

Yes you are correct.

Yes, sklearn offers LabelEncoder and OneHotEncoder classes.

Here is a tutorial to get you started:

<http://machinelearningmastery.com/data-preparation-gradient-boosting-xgboost-python/>



**Cameron** March 19, 2017 at 3:50 am #

Awesome! thanks so much Jason!

## Your Start in Machine Learning



**Jason Brownlee** March 19, 2017 at 9:11 am #

You're welcome, let me know how you go.



**James** March 19, 2017 at 6:54 am #

REPLY ↗

for name, model in models:

```
... kfold = cross_validation.KFold(n=num_instances,n_folds=num_folds,random_state=seed)
... cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

File "", line 3

```
cv_results = model_selection.cross_val_score(model, )
^
```

SyntaxError: invalid syntax

```
>>> cv_results = model_selection.cross_val_score(mo
```

Traceback (most recent call last):

File "", line 1, in

NameError: name 'model' is not defined

```
>>> cv_results = model_selection.cross_val_score(mo
```

Traceback (most recent call last):

File "", line 1, in

NameError: name 'kfold' is not defined

```
>>> cv_results = model_selection.cross_val_score(mo
```

kfold, scoring = scoring)

Traceback (most recent call last):

File "", line 1, in

NameError: name 'kfold' is not defined

```
>>> names.append(name)
```

Traceback (most recent call last):

File "", line 1, in

NameError: name 'name' is not defined

I am new to python and getting these errors after running 5.3 models. Please help me.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 19, 2017 at 9:12 am #

REPLY ↗

It looks like you might not have copied all of the code required for the example.



**Mier** March 20, 2017 at 10:26 am #

REPLY ↗

Hi, I went through your tutorial. It is super great! I wonder whether you can recommend a data set that

Your Start in Machine Learning



**Jason Brownlee** March 21, 2017 at 8:36 am #

REPLY ↩

Thanks Mier,

I recommend some datasets here:

<http://machinelearningmastery.com/practice-machine-learning-with-small-in-memory-datasets-from-the-uci-machine-learning-repository/>



**Medine H.** March 23, 2017 at 2:56 am #

REPLY ↩

Hi Jason,

That's an amazing tutorial, quite clear and useful.

Thanks a bunch!



**Jason Brownlee** March 23, 2017 at 8:50 am #

REPLY ↩

Thanks Medine.



**Sean** March 23, 2017 at 9:54 am #

REPLY ↩

Hi Jason,

Can you let me know how can I start with Fraud Detection algorithms for a retail website ?

Thanks,

Sean



**Jason Brownlee** March 24, 2017 at 7:51 am #

REPLY ↩

Hi Sean, this process will help you work through your problem:

<http://machinelearningmastery.com/start-here/#process>



**Raja** March 24, 2017 at 11:08 am #

REPLY ↩

You are doing great with your work.

Your Start in Machine Learning

I need your suggestion, i am working on my thesis here i need to work on machine learning.

Training : positive ,negative, others

Test : unknown data

Want to train machine with training and test with unknown data using SVM,Naive,KNN

How can i make the format of training and test data ?

And how to use those algorithms in it

Using which i can get the TPTN,FP,FN

Thanking you..



**Jason Brownlee** March 25, 2017 at 7:31 am #

REPLY ↗

This article might help:

[https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)



**Sey** March 26, 2017 at 12:38 am #

I m new in Machine learning and this was a re wanted to plot the predictions and the validation value like I really understood how I can plot it.

Can you please send me the piece of code with some thank you very much

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 26, 2017 at 6:13 am #

REPLY ↗

You can use matplotlib, for example:

```
1 yhat = model.predict(X)
2 from matplotlib import pyplot
3 pyplot.plot(y, yhat)
4 pyplot.show()
```



**Kamol Roy** March 26, 2017 at 7:25 am #

REPLY ↗

Thanks a lot. It was very helpful.



**Jason Brownlee** March 27, 2017 at 7:51 am #

REPLY ↗

You're welcome Kamol, I'm glad to hear it

Your Start in Machine Learning



**Rajneesh** March 29, 2017 at 11:31 pm #

REPLY ↗

Hi

Sorry for a dumb question.

Can you briefly describe, what the end result means (i.e.. what the program has predicted)



**Jason Brownlee** March 30, 2017 at 8:53 am #

REPLY ↗

Given an input description of flower meas

We are predicting the iris flower species as one of



**Anusha Vidapanakal** March 30, 2017 at 3:58 am #

LR: 0.966667 (0.040825)  
LDA: 0.975000 (0.038188)  
KNN: 0.983333 (0.033333)  
CART: 0.975000 (0.038188)  
NB: 0.975000 (0.053359)  
SVM: 0.991667 (0.025000)

Why am I getting the highest accuracy for SVM?

I'm a beginner, there was a similar query above but I couldn't quite understand your reply.

Could you please help me out? Have I done any mistake?

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 30, 2017 at 8:56 am #

REPLY ↗

Why is a very hard question to answer.

Our role is to find what works, ensure the results are robust, then figure out how we can use the model operationally.



**Anusha Vidapanakal** March 30, 2017 at 11:33 pm #

REPLY ↗

Okay. Thanks a lot for the prompt response!

The tutorial was very helpful.

Your Start in Machine Learning



**Jason Brownlee** March 31, 2017 at 5:54 am #

REPLY ↗

Glad to hear it Anusha.



**Vinay** March 31, 2017 at 11:10 pm #

REPLY ↗

Great tutorial Jason!

My question is, if I want some new data from a user, how do I do that? If in future I develop my own machine learning algorithm, how do I use it to get some new data?

What steps are taken to develop it?

And thanks for this tutorial.



**Jason Brownlee** April 1, 2017 at 5:56 am #

Not sure I understand. Collect new data first to collect it.



**walid barakeh** April 2, 2017 at 6:31 pm #

Hi Jason,

I have a question regards the step after trained the data and know the better algorithm for our case, how we could know the rules formula that the algorithm produced for future uses ?

and thanks for the tutorial, its really helpful



**Jason Brownlee** April 4, 2017 at 9:06 am #

REPLY ↗

You can extract the weights if you like. Not sure I understand why you want the formula for the network. It would be complex and generally unreadable.

You can finalize the mode, save the weights and topology for later use if you like.



**walid barakeh** April 5, 2017 at 7:40 pm #

REPLY ↗

the best algorithm results for my use case was the “Classification and Regression Trees (CART)”, so how could I know the rules that the algorithm created on my use case how I could extract the weights and use them

Your Start in Machine Learning

Thanks for your prompt response



**Jason Brownlee** April 9, 2017 at 2:34 pm #

REPLY ↩

See this post on how to finalize your model:

<http://machinelearningmastery.com/train-final-machine-learning-model/>



**Divya** April 4, 2017 at 4:58 pm #

REPLY ↩

Thank you so much...this document really helped me since a long time...this document gave the actual view of machine learning in python....Books and courses are really difficult to understand. I am currently working on such a vast concept... books n videos gave me a clear idea of how they all fit together.



**Jason Brownlee** April 9, 2017 at 2:30 pm #

I'm glad to hear that.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Divya** April 4, 2017 at 5:00 pm #

REPLY ↩

can i get such more tutorials for more detailed understanding?.....It will be really helpfull.



**Jason Brownlee** April 9, 2017 at 2:30 pm #

REPLY ↩

Sure, see here:

<http://machinelearningmastery.com/start-here/#python>



**Gav** April 11, 2017 at 5:17 pm #

REPLY ↩

Can't load the iris dataset either through the url or copied to working folder without the NameError:  
name 'pandas' is not defined

**Jason Brownlee** April 12, 2017 at 7:51 am #

Your Start in Machine Learning



You need to install the Pandas library.

See this tutorial:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Gavin** April 12, 2017 at 9:53 pm #

REPLY ↗

I've already installed Anaconda with Python 3.6 and the panda libraries are listed when I run versions.py. Everything has been fine up till trying to load the iris library. Do I need to use a different terminal within Anaconda?



**Jason Brownlee** April 13, 2017 at 1

You may need to close and re-open after installation.



**Sunil** June 4, 2017 at 2:31 am #

add a line  
import pandas  
at the top

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Ursula** April 13, 2017 at 7:33 pm #

REPLY ↗

Hi Jason,

Your tutorial is fantastic!  
I'm trying to follow it but gets stuck on 5.3 Build Models  
  
When I copy your code for this section I get a few Errors  
IndentationError: expected an indented block  
NameError: name 'model' is not defined  
NameError: name 'cv\_results' is not defined  
NameError: name 'name' is not defined

## Your Start in Machine Learning

Could you please help me find what I'm doing wrong?  
Thanks!

see the code and my “results” below:

```
>>> # Spot Check Algorithms
... models = []
>>> models.append('LR', LogisticRegression())
>>> models.append('LDA', LinearDiscriminantAnalysis())
>>> models.append('KNN', KNeighborsClassifier())
>>> models.append('CART', DecisionTreeClassifier())
>>> models.append('NB', GaussianNB())
>>> models.append('SVM', SVC())
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
... kfold = model_selection.KFold(n_splits=10, random_
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_st
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(mo
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> results.append(cv_results)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'cv_results' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> print(msg)
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Jason Brownlee** April 14, 2017 at 8:43 am #

REPLY ↗

Make sure you have the same tab indenting as in the example. Maybe re-add the tabs yourself after you copy-paste the code.

Your Start in Machine Learning



**Nathan Wilson** March 26, 2018 at 11:16 am #

[REPLY ↗](#)

I'm having this same problem. How would I add the Indentations after I paste the code?  
Whenever I paste the code, it automatically executes the code.



**Jason Brownlee** March 26, 2018 at 2:27 pm #

[REPLY ↗](#)

How to copy code from the tutorial:

1. Click the copy button on the code example (top right of code box, second from the end). This will select all code in the box.
2. Copy the code to the clipboard (control-click copy).
3. Open your text editor.
4. Paste the code from the clip board.

This will preserve all white space.

Does that help?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)



**Davy** April 14, 2017 at 10:14 pm #

Hi, one beginner question. What do we get after training is completed in supervised learning, for classification problem ? Do we get weights? How do i use the trained model after that in field, for real classification application lets say? I didn't get the concept what happens if training is completed. I tried this example: [https://github.com/fchollet/keras/blob/master/examples/mnist\\_mlp.py](https://github.com/fchollet/keras/blob/master/examples/mnist_mlp.py) and it printed me accuracy and loss of test data. Then what now?



**Jason Brownlee** April 15, 2017 at 9:35 am #

[REPLY ↗](#)

See this post on how to train a final model:

<http://machinelearningmastery.com/train-final-machine-learning-model/>



**Manikandan** April 14, 2017 at 11:36 pm #

[REPLY ↗](#)

Wow... It's really great stuff man.... Thanks you....



**Jason Brownlee** April 15, 2017 at 9:36 am #

Your Start in Machine Learning

I'm glad to hear that.



**Wes** April 15, 2017 at 3:16 am #

REPLY ↗

As a complete beginner, it sounds so cool to predict the future. Then I saw all these model and complicated stuff, how do I even begin. Thank you for this. It is really great!



**Jason Brownlee** April 15, 2017 at 9:40 am #

REPLY ↗

You're very welcome.



**Manjushree Aithal** April 16, 2017 at 7:41 am #

Hello Jason,

I just started following your step by step tutorial for machine learning. In each and every steps you specified, install all libraries

Traceback (most recent call last):

```
File "C:/Users/dell/PycharmProjects/machine-learning.py", line 1, in <module>
  from sklearn.linear_model import LogisticRegression
File "C:\Users\dell\Anaconda2\lib\site-packages\sklearn\linear_model\__init__.py", line 15, in <module>
  from .least_angle import (Lars, LassoLars, lars_path, LarsCV, LassoLarsCV,
File "C:\Users\dell\Anaconda2\lib\site-packages\sklearn\linear_model\least_angle.py", line 24, in <module>
  from ..utils import arrayfuncs, as_float_array, check_X_y
ImportError: DLL load failed: Access is denied.
```

Can you please help me with this?

Thank You!

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 16, 2017 at 9:33 am #

REPLY ↗

I have not seen this error and I don't know about windows sorry.

It looks like you might not have admin permissions on your workstation.



**Olah Data Semarang** April 17, 2017 at 3:03 pm #

REPLY ↗

## Your Start in Machine Learning

## Tutorial DEAP Version 2.1

<https://www.youtube.com/watch?v=drd11htJJCO>

A Data Envelopment Analysis (Computer) Program. This page describes the computer program Tutorial DEAP Version 2.1 which was written by Tim Coelli.



**Jason Brownlee** April 18, 2017 at 8:30 am #

REPLY ↗

Thanks for sharing the link.



**Federico Carmona** April 18, 2017 at 4:41 am #

X

Good afternoon Dr. Jason could help me with an algorithm to detect the most relevant variables?



**Jason Brownlee** April 18, 2017 at 8:34 am #

REPLY ↗

You can use feature importance scores from

Consider using sklearn to calculate and plot feature

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Bharath** April 18, 2017 at 10:09 pm #

Thank u...



**Jason Brownlee** April 19, 2017 at 7:52 am #

REPLY ↗

I'm glad the post helped.



**Amal** April 26, 2017 at 6:14 pm #

REPLY ↗

Hi Jason

Thanx for the great tutorial you provided.

I'm also new to MC and python. I tried to use my csv file as you used iris data set. Though it successfully loaded the dataset gives following error.

could not convert string to float: LipCornerDepressor

## Your Start in Machine Learning

LipCornerDepressor is normal value such as 0.32145 in excel sheet taken from sql server

Here is the code without library files.

```
# Load dataset
url = "F:\FINAL YEAR PROJECT\Amila\FTdata.csv"
names = ['JawLower', 'BrowLower', 'BrowRaiser', 'LipCornerDepressor',
'LipRaiser', 'LipStretcher', 'Emotion_Id']
dataset = pandas.read_csv(url, names=names)

# shape
print(dataset.shape)

# class distribution
print(dataset.groupby('Emotion_Id').size())

# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(
array, Y, test_size=validation_size, random_state=seed)

# Test options and evaluation metric
seed = 7
scoring = 'accuracy'

# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))

# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

Your Start in Machine Learning



**Jason Brownlee** April 27, 2017 at 8:37 am #

REPLY ↤

This error might be specific to your data.

Consider double checking that your data is loaded as you expect. Maybe print some raw data or plots to confirm.



**Chanaka** April 27, 2017 at 6:31 am #

REPLY ↤

Thank you very much for the easy to follow tutorial.



**Jason Brownlee** April 27, 2017 at 8:48 am #

X

I'm glad you found it useful.



**Sonali Deshmukh** April 27, 2017 at 7:07 pm #

REPLY ↤

Hi, Jason

Your posts are really good.....

I'm very naive to Python and Machine Learning.

Can you please suggest good reads to get basic clear for machine learning.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 28, 2017 at 7:38 am #

REPLY ↤

Thanks.

A good place to start for python machine learning is here:

<http://machinelearningmastery.com/start-here/#python>

I hope that helps.



**Ianndo** April 28, 2017 at 2:26 am #

REPLY ↤

Outstanding work on this. I am curious how to port out results that show which records were matched to what in the predictor, when I print(predictions) it does not show what records they are paired with. Thanks!

## Your Start in Machine Learning



**Jason Brownlee** April 28, 2017 at 7:51 am #

REPLY ↗

Thanks!

The index can be used to align predictions with inputs. For example, the first prediction is for the first input, and so on.



**NAVKIRAN KAUR** April 29, 2017 at 4:28 pm #

REPLY ↗

when I am applying all the models and printing message it shows me the error that it cannot convert string to float. how to resolve this error. my da



**Jason Brownlee** April 30, 2017 at 5:27 am #

Ensure you have converted your text data



**Shravan** May 1, 2017 at 6:29 am #

Awesome tutorial on basics of machine learni

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 2, 2017 at 5:51 am #

REPLY ↗

Thanks Shravan.



**Shravan** May 1, 2017 at 6:36 am #

REPLY ↗

Am using Anaconda Python and I was writing all the commands/ program in the 'python' command line, am trying to find a way to save this program to a file? I have tried '%save', but it errored out, any thoughts?



**Jason Brownlee** May 2, 2017 at 5:51 am #

REPLY ↗

You can write your programs in a text file then run them on the command line as follows:

```
1 python file.py
```

Your Start in Machine Learning



**Jason** May 1, 2017 at 2:05 pm #

REPLY ↗

Thank you for the help and insight you provide. When I run the actual validation data through the algorithms, I get a different feel for which one may be the best fit.

Validation Test Accuracy:

LR.....0.80

LDA.....0.97

KNN....0.90

CART..0.87

NB.....0.83

SVM....0.93

My question is, should this influence my choice of algos?

Thank you again for providing such a wealth of information.



**Jason Brownlee** May 2, 2017 at 5:56 am #

Tes it should.

ML algorithms are stochastic and you need to evaluate them multiple times.

This post might clarify what I mean:

<http://machinelearningmastery.com/randomness-in-machine-learning/>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**rahman** May 3, 2017 at 11:09 pm #

REPLY ↗

# Split-out validation dataset

array = dataset.values

X = array[:,0:4]

Y = array[:,4]

from my dataset , When i give Y=array[:,1] Its working , but if give 2 or 3 or 4 instead of 1 it gives following error !!

But all columns have similar kind of data .

Traceback (most recent call last):

File “/alok/c-analyze/analyze.py”, line 390, in

cv\_results = model\_selection.cross\_val\_score(model, X\_train, Y\_train, cv=kfold, scoring=scoring)

File “/usr/lib64/python2.7/site-packages/sklearn/model\_selection/\_validation.py”, line 140, in

cross\_val\_score

for train, test in cv\_iter)

File “/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py”, line 758, in \_\_call\_\_

while self.dispatch\_one\_batch(iterator):

File “/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py”, line 758, in \_\_call\_\_

Your Start in Machine Learning

```

dispatch_one_batch
self._dispatch(tasks)
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 571, in _dispatch
job = self._backend.apply_async(batch, callback=cb)
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 109, in
apply_async
result = ImmediateResult(func)
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 326, in
__init__
self.results = batch()
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 131, in __call__
return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "/usr/lib64/python2.7/site-packages/sklearn/model_selection/_validation.py", line 100, in
estimator.fit(X_train, y_train, **fit_params)
File "/usr/lib64/python2.7/site-packages/sklearn/discriminant_analysis.py", line 100, in
self._solve_svd(X, y)
File "/usr/lib64/python2.7/site-packages/sklearn/discriminant_analysis.py", line 100, in
fac = 1. / (n_samples - n_classes)

ZeroDivisionError: float division by zero

```



**Jason Brownlee** May 4, 2017 at 8:08 am #

Perhaps take a closer look at your data.



**rahman** May 4, 2017 at 4:29 pm #

But the very similar in all the columns .



**rahman** May 4, 2017 at 4:37 pm #

I meant there is no much difference in data from each columns ! but still its working only for first column !! It gives the above error for any other column i choose .



**rahman** May 4, 2017 at 4:46 pm #

Have a look at the data :

index,1column,2 column,3column,....,8column  
0,238,240,1103,409,1038,4,67,0

Your Start in Machine Learning

1,41,359,995,467,1317,8,71,0  
 2,102,616,1168,480,1206,7,59,0  
 3,0,34,994,181,1115,4,68,0  
 4,88,1419,1175,413,1060,8,71,0  
 5,826,10886,1316,6885,2086,263,119,0  
 6,88,472,1200,652,1047,7,64,0  
 7,0,322,957,533,1062,11,73,0  
 8,0,200,1170,421,1038,5,63,0  
 9,103,1439,1085,1638,1151,29,66,0  
 10,0,1422,1074,4832,1084,27,74,0  
 11,1828,754,11030,263845,1209,10,79,0  
 12,340,1644,11181,175099,4127,13,136,0  
 13,71,1018,1029,2480,1276,18,66,1  
 14,0,3077,1116,1696,1129,6,62,0

.....  
 .....

Total 105 data records

But the above error does not occur for  
 But the above same error happens wh



**hairo** May 3, 2017 at 11:13 pm #

How to plot the graph for actual value against the predicted value here ?

How to save this plotted graphs and again view them back when required from terminal itself ?



**Jason Brownlee** May 4, 2017 at 8:08 am #

REPLY ↗

It would make for a dull graph as this is a classification problem.

You might be better off reviewing the confusion matrix of a set of predictions.



**Sudarshan** May 5, 2017 at 12:18 pm #

REPLY ↗

How this can be applied to predict the value if statistical dataset is given  
 Say I have given with past 10 years house price now I want to predict the value for house in next one year,  
 two years

Can you help me out in this

I'm amateur in ML

Your Start in Machine Learning

Thank for this tutorial  
It gives me a good kickstart to ML  
I m waiting for your reply



**Jason Brownlee** May 6, 2017 at 7:30 am #

REPLY ↩

This is called a time series forecasting problem.

You can learn more about how to work through time series forecasting problems here:  
<http://machinelearningmastery.com/start-here/#timeseries>



**Sudarshan** May 6, 2017 at 3:15 pm #

I getting trouble in doing that please help me.  
Example I have a dataset containing plumber data. The attributes are experience\_level , date, rating, price/hour. I want to predict the price/hour for the next day. Can you please help me regarding this.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

X

REPLY ↩



**Jason Brownlee** May 7, 2017 at 5:34 am #

Sorry, I cannot write an example for you.



**Bane** May 8, 2017 at 4:30 am #

REPLY ↩

Great job with the tutorial, it was really helpful.

I want to ask, how can I use the techics above with a dataset that is not just one line with a few values, but a matrix NX3 with multiple values (measurements from an accelerometer). Is there a tutorial? How can I look up to it?



**Jason Brownlee** May 8, 2017 at 7:46 am #

REPLY ↩

Each feature would be a different input variable as in the example above.

## Your Start in Machine Learning



**Shud** May 9, 2017 at 12:04 am #

REPLY ↗

Hey Jason,

I have built a linear regression model. y intercept is abnormally high (0.3 million) and adjusted r<sup>2</sup> = 0.94. I would like to know what does high intercept mean?



**Jason Brownlee** May 9, 2017 at 7:45 am #

REPLY ↗

Think of the intercept as the bias term.

Many books have been written on linear regression models effectively. I would recommend diving into



**MK** May 11, 2017 at 12:19 am #

Excellent tutorial, i am moving from PHP to Py (http://thonny.org/) which is also very useful for python



**Jason Brownlee** May 11, 2017 at 8:33 am #

Thanks for sharing.



**Tmoe** May 14, 2017 at 4:31 am #

REPLY ↗

Thank you so much, Jason! I'm new to machine learning and python but found your tutorial extremely helpful and easy to follow – thank you for posting!



**Jason Brownlee** May 14, 2017 at 7:32 am #

REPLY ↗

Thanks Tmoe, I'm really glad to hear that!



**melody12ab** May 15, 2017 at 6:07 pm #

REPLY ↗

Thanks for all,now I am starting use ML!!!

Your Start in Machine Learning



**Jason Brownlee** May 16, 2017 at 8:39 am #

REPLY ↗

I'm glad to hear that!



**smith** May 15, 2017 at 9:36 pm #

REPLY ↗

# Spot Check Algorithms

```
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
```

When i print models , this is the output :

```
[('LR', LogisticRegression(C=1.0, class_weight=None,
intercept_scaling=1, max_iter=100, multi_class='ovr',
penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False)), ('LDA', LinearDiscriminatir
shrinkage=None,
solver='svd', store_covariance=False, tol=0.0001)), ('K
leaf_size=30, metric='minkowski',
metric_params=None, n_jobs=1, n_neighbors=5, p=2,
weights='uniform')]
```

What are these extra values inside LogisticRegression

How did they get appended ?

## Your Start in Machine Learning



You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** May 16, 2017 at 8:43 am #

REPLY ↗

You can learn about them in the sklearn API:

<http://scikit-learn.org/stable/modules/classes.html>



**pasha** May 15, 2017 at 9:45 pm #

REPLY ↗

When i print kfold :

```
KFold(n_splits=7, random_state=7, shuffle=False)
```

What is shuffle ? How did this value get added , as we had only done this :

```
kfold = model_selection.KFold(n_splits=10, random_state=seed)
```

## Your Start in Machine Learning



**Jason Brownlee** May 16, 2017 at 8:44 am #

REPLY ↗

Whether or not to shuffle the dataset prior to splitting into folds.



**pasha** May 16, 2017 at 3:17 pm #

REPLY ↗

Now i understand , jason thanks for amazing tutorials . Just one suggestion along with the codes give a link for reference in detail about this topics !



**Jason Brownlee** May 17, 2017 at 8:

Great suggestion, thanks pasha.



**sita** May 15, 2017 at 9:48 pm #

Hello jason

This is an amazing blog , Thank you for all the posts .

```
cv_results = model_selection.cross_val_score(model, )
```

Whats scoring here ? can you explain in detail " `model_selection.cross_val_score` " this line please .

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 16, 2017 at 8:45 am #

REPLY ↗

Thanks sita.

Learn more here:

[http://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.cross\\_val\\_score.html#sklearn.model\\_selection.cross\\_val\\_score](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html#sklearn.model_selection.cross_val_score)



**rahman** May 15, 2017 at 10:27 pm #

REPLY ↗

Please help me with this error Jason ,

ERROR :

Traceback (most recent call last):

File “/rahman/c-analyze/analyze.py”, line 390, in

Your Start in Machine Learning

```

cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
File "/usr/lib64/python2.7/site-packages/sklearn/model_selection/_validation.py", line 140, in
cross_val_score
for train, test in cv_iter)
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 758, in __call__
while self.dispatch_one_batch(iterator):
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 608, in
dispatch_one_batch
self._dispatch(tasks)
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 571, in _dispatch
job = self._backend.apply_async(batch, callback=cb)
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 109, in
apply_async
result = ImmediateResult(func)
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 109, in
__init__
self.results = batch()
File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 109, in
return [func(*args, **kwargs) for func, args, kwargs in s]
File "/usr/lib64/python2.7/site-packages/sklearn/model_selection/_validation.py", line 571, in
estimator.fit(X_train, y_train, **fit_params)
File "/usr/lib64/python2.7/site-packages/sklearn/discriminant_analysis.py", line 109, in
self._solve_svd(X, y)
File "/usr/lib64/python2.7/site-packages/sklearn/discriminant_analysis.py", line 109, in
fac = 1. / (n_samples - n_classes)

```

ZeroDivisionError: float division by zero

# Split-out validation dataset

My code :

```

array = dataset.values
X = array[:,0:4]

if field == "rh": #No error if i select this col
Y = array[:,0]

elif field == "rm": #gives the above error
Y = array[:,1]

elif field == "wh": #gives the above error
Y = array[:,2]

elif field == "wm": #gives the above error
Y = array[:,3]

```

Have a look at the data :

```

index,1column,2 column,3column,....,8column
0,238,240,1103,409,1038,4,67,0

```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

1,41,359,995,467,1317,8,71,0  
 2,102,616,1168,480,1206,7,59,0  
 3,0,34,994,181,1115,4,68,0  
 4,88,1419,1175,413,1060,8,71,0  
 5,826,10886,1316,6885,2086,263,119,0  
 6,88,472,1200,652,1047,7,64,0  
 7,0,322,957,533,1062,11,73,0  
 8,0,200,1170,421,1038,5,63,0  
 9,103,1439,1085,1638,1151,29,66,0  
 10,0,1422,1074,4832,1084,27,74,0  
 11,1828,754,11030,263845,1209,10,79,0  
 12,340,1644,11181,175099,4127,13,136,0  
 13,71,1018,1029,2480,1276,18,66,1  
 14,0,3077,1116,1696,1129,6,62,0

.....  
 .....

Total 105 data records

But the above error does not occur for 1 column , that

But the above same error happens when i choose any



**Jason Brownlee** May 16, 2017 at 8:45 am #

Perhaps try scaling your data?

Perhaps try another algorithm?



**suma** May 16, 2017 at 12:05 am #

REPLY ↗

fac = 1. / (n\_samples – n\_classes)

ZeroDivisionError: float division by zero

What is this error : fac = 1. / (n\_samples – n\_classes) ?

Where is n\_samples and n\_classes used ?

What may be the possible reason for this error ?



**bob** May 22, 2017 at 6:46 pm #

REPLY ↗

thank you Dr Jason it is really very helpfully. 😊

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 23, 2017 at 7:50 am #

REPLY ↗

You're welcome bob, I'm glad to hear that!

REPLY ↗



**Krithika** May 24, 2017 at 12:24 am #

Hi Jason

Great starting tutorial to get the whole picture. Thank you:)

I am a newbie to machine learning. Could you please tell why you have specifically chosen these 6 models?



**Jason Brownlee** May 24, 2017 at 4:57 am #

No specific reason, just a demonstration.

X

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Ram Gour** May 25, 2017 at 8:24 pm #

Hi Jason, I am new to Python, but found this tutorial very useful. I followed all the steps and return all the result as mention above by you, except for the scatter matrix graph. I am not able to generate the scatter matrix graph and the evaluation on 6 algorithms. I checked all the version and its higher or same as you mentioned in blog. Can you help if this issue can be resolved on my machine?

REPLY ↗



**Jason Brownlee** June 2, 2017 at 11:44 am #

Perhaps check the configuration of matplotlib and ensure you can create simple graphs on your machine?

REPLY ↗



**sridhar** May 25, 2017 at 8:50 pm #

Great tutorial.

How do I approach when the data set is not of any classification type and the number of attributes or just 2 – 1 is input and the other is output

say I have number of processes as input and cpu usage as output..  
data set looks like [10, 5] [15, 7] etc...

## Your Start in Machine Learning



**Jason Brownlee** June 2, 2017 at 11:45 am #

[REPLY ↗](#)

If the output is real-valued, it would be a regression problem. You would need to use a loss function like MSE.



**pierre** May 27, 2017 at 9:45 pm #

[REPLY ↗](#)

Many thanks for this — I already got a lot out of this. I feel like a monkey though because I was neither familiar enough with python nor had any clue of ML back alleys yesterday. Today I can see plots on my screen and even if I have no clue what I'm looking at, this is where I wanted to be, so thanks!

A few minor suggestions to make this perhaps even more useful:

- I'm on Mac and I used python3 because python2 is very old now. I understand you link, right? This stuff works in python3 if you needed further testing.
- when drawing plots, I started freaking out because the code was so different from R. I made an (unessential) suggestion to run plt.ion() first, I think. I don't know if this is good advice or not. I give up too easily. (BTW I find your use command line interface for plotting great one indeed!)
- There seems to be some 'hack' involved when defining the target variable. I mean, how do you get to load your dataset with an insight? Just a hint of clarification would help here. I'm feeling we can trust that we do the right thing in this case because the data is well understood (I mean, this is not really a big deal eh it's all par for the course but if I didn't have similar experience in R I'd feel completely lost I think).

I was a bit puzzled by the following sentence in 3.3:

"We can see that all of the numerical values have the same scale (centimeters) and similar ranges between 0 and 8 centimeters."

Well, just looking at the table, I actually can't see any of this. There is in fact really nothing telling this to us in the snippet, right? The sentence is a comment based on prior understanding of the dataset. Maybe this could be clarified so clueless readers don't agonise over whether they are missing some magical power of insight.

– Overall, I could run this and to some extent adapt it quickly to a different dataset until it became relevant what the data was like. I'm stumbling on the data manipulation for 5.1. I suppose it is both because I don't know python structures and also because I have no clue what is being done in the selection step.

I think in answer to a previous comment you link to doc for the relevant selection function, perhaps it would still be useful to have an extra, 'for dummies', detailed explanation of

```
X = array[:,0:4]
Y = array[:,4]
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

in the context of the iris dataset. This is what I have to figure out, I think, in order to apply it to say, a 11 column dataset and it would be useful to known what I'm trying to do.

The rest of the difficulties I have are with regards to interpretation of the output and it is fair to say this is outside of the scope of your tutorial which puts dummies like me in a very good position to try to understand while being able to fiddle with a bit of code. All the above comments are extremely minor and really about polishing the readability for ultimate noobs, they are not really important and your tutorial is a great and efficient resource.

Thanks again!

Pierre



**Jason Brownlee** June 2, 2017 at 12:04 pm #

Wonderful feedback pierre, thank you so much!



**Shaksham Kapoor** June 6, 2017 at 4:18 am #

I'm not able to figure out , what errors does the column(precision, recall, f1-score, support) in the class

And last but not the least thanks a lot Sir for this easy enough to express my gratitude, you have made a da

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 6, 2017 at 10:07 am #

REPLY ↗

You can learn more about the confusion matrix here:

<http://machinelearningmastery.com/confusion-matrix-machine-learning/>



**Shaksham Kapoor** June 7, 2017 at 3:39 am #

REPLY ↗

Thanks a lot Sir. Please suggest some data-sets from UCL repository on which I can practice some small projects...



**Jason Brownlee** June 7, 2017 at 7:26 am #

REPLY ↗

See here:

<http://machinelearningmastery.com/practice-machine-learning-with-small-in-memory-datasets-from-the-uci-machine-learning-repository/>

Your Start in Machine Learning



**Shaksham Kapoor** June 7, 2017 at 6:48 pm #

How do you classify problem into different categories example : Iris dataset was a classification problem and pima-indian-diabetes ,a binary problem. How can we figure out which problem belong to which category and which model to apply on that problem?



**Jason Brownlee** June 8, 2017 at 7:40 am #

By careful evaluation of the or



**Brian** June 6, 2017 at 11:11 pm #

Is this machine learning? what does the mac  
Statistics, used in a weird way...



**Jason Brownlee** June 7, 2017 at 7:14 am #

Yes, it is.



## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Raj** June 9, 2017 at 2:22 am #

REPLY ↗

your question can be answered like this...

consider the formula for area of triangle  $1/2 \times \text{base} \times \text{height}$ . When you learn this formula, you understand it and apply it many times for different triangles. BUT you did not learn anything ABOUT the formula itself. . for instance, how many people care that the formula has 2 variables(base and height) and that there is no CONSTANT(like PI) in the formula and many such things about the formula itself? Applying the formula does not teach anything about the nature of the formula itself

A lot of program execution in computers happen much the same way...data is a thing to be modified, applied or used, but not necessarily understood. When you introduce some techniques to understand data, then necessarily the computer or the 'Machine' 'learns' that there are characteristics about that data, and that at the least, there exists some relationship amongst data in their dataset. This learning is not explicitly programmed rather inferred, although confusingly, the algorithms themselves are explicitly programmed to infer the meaning of the dataset. The learning is then transferred to the end cycle of making prediction based on the gained ur

Your Start in Machine Learning

but like you pointed out, it is still statistics and all it's domain techniques, but as a statistician do you not 'learn' more about data than merely use it, unlike your counterparts who see data more as a commodity to be consumed? Because most computer systems do the latter(consumption) rather than the former(data understanding), a system that understands data(with prediction used as a proof of learning) can be called 'Machine Learning'.



**Alex** June 7, 2017 at 6:04 am #

REPLY ↗

Thanks for good tutorial Jason.

Only issue I encountered is following error while cross validation score calculation for model KNeighborsClassifier() :

AttributeError: 'NoneType' object has no attribute 'issupp...'

Is somebody got same error? How it can be solved?

I have installed following versions of tools:

Python: 2.7.13 |Anaconda custom (64-bit)| (default, Dec 2 2016, 16:41:25) [MSC v.1900 64 bit (AMD64)]  
scipy: 0.19.0

numpy: 1.12.1

matplotlib: 2.0.0

pandas: 0.19.2

sklearn: 0.18.1

Thanks,

Alex

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 7, 2017 at 7:27 am #

REPLY ↗

Ouch, sorry I have not seen this issue. Perhaps search on stackoverflow?



**thanda** June 8, 2017 at 6:31 pm #

REPLY ↗

Hi, Jason!

How can i get the xgboost algorithm in pseudo code or in code?



**Jason Brownlee** June 9, 2017 at 6:21 am #

REPLY ↗

You can read the code here:

<https://github.com/dmlc/xgboost>

I expect it is deeply confusing to read.

## Your Start in Machine Learning

For an overview of gradient boosting, see this post:

<http://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/>



**Shaksham Kapoor** June 9, 2017 at 1:14 am #

REPLY ↗

Sir,I've been working on bank\_note authentication dataset and after applying the above procedure carefully the results were 100% accuracy(both on trained and validation dataset) using SVM and KNN models. Is 100% accuracy possible or have I done something wrong ?



**Jason Brownlee** June 9, 2017 at 6:27 am #

That sounds great.

If I were to get surprising results, I would be skeptical.

Work hard to ensure your system is not fooling you.



**Shaksham Kapoor** June 9, 2017 at 3:10 pm #

Sir, I've considered various other aspects and the result is same 100%. How can I make sure what procedure can I apply to check the accuracy of my dataset ?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Rejeesh R** June 9, 2017 at 7:27 pm #

REPLY ↗

Hi, Jason!

I am new to python as well ML. so I am getting the below error while running your code, please help me to code bring-up

File "sample1.py", line 73, in

predictions = knn.predict(X\_validation)

File "/usr/local/lib/python2.7/dist-packages/sklearn/neighbors/classification.py", line 143, in predict

X = check\_array(X, accept\_sparse='csr')

File "/usr/local/lib/python2.7/dist-packages/sklearn/utils/validation.py", line 407, in check\_array

\_assert\_all\_finite(array)

Your Start in Machine Learning

File “/usr/local/lib/python2.7/dist-packages/scikit-learn/utils/validation.py”, line 58, in \_assert\_all\_finite  
” or a value too large for %r.” % X.dtype)

ValueError: Input contains NaN, infinity or a value too large for dtype('float64').

and my config

Python: 2.7.6 (default, Oct 26 2016, 20:30:19)

[GCC 4.8.4]

scipy: 0.13.3

numpy: 1.8.2

matplotlib: 1.3.1

pandas: 0.13.1

sklearn: 0.18.1

running in Ubuntu Terminal.



**Jason Brownlee** June 10, 2017 at 8:20 am #

You may have a NaN value in your database.



**Sats S** June 10, 2017 at 5:27 am #

Hello. This is really an amazing tutorial. I got caught up. I hit a snag. Can you help out?

Traceback (most recent call last):

```
File “/Users/sahityasehgal/Desktop/py/machinetest.py”, line 77, in
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
File “/Users/sahityasehgal/Library/Python/2.7/lib/python/site-
packages/scikit-learn/model_selection/_validation.py”, line 140, in cross_val_score
for train, test in cv_iter)
File “/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/scikit-learn/externals/joblib/parallel.py”, line 758, in __call__
while self.dispatch_one_batch(iterator):
File “/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/scikit-learn/externals/joblib/parallel.py”, line 608, in dispatch_one_batch
self._dispatch(tasks)
File “/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/scikit-learn/externals/joblib/parallel.py”, line 571, in _dispatch
job = self._backend.apply_async(batch, callback=cb)
File “/Users/sahityasehgal/Library/Python/2.7/lib/python/site-
packages/scikit-learn/externals/joblib/_parallel_backends.py”, line 109, in apply_async
result = ImmediateResult(func)
File “/Users/sahityasehgal/Library/Python/2.7/lib/python/site-
packages/scikit-learn/externals/joblib/_parallel_backends.
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

self.results = batch()
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/externals/joblib/parallel.py",
line 131, in __call__
return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-
packages/sklearn/model_selection/_validation.py", line 238, in _fit_and_score
estimator.fit(X_train, y_train, **fit_params)
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/linear_model/logistic.py",
line 1173, in fit
order="C")
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/utils/validation.py", line 526,
in check_X_y
y = column_or_1d(y, warn=True)
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/
in column_or_1d
raise ValueError("bad input shape {0}".format(shape))
ValueError: bad input shape (94, 4)

```



**Jason Brownlee** June 10, 2017 at 8:28 am #

Ouch. Are you able to confirm that you co

Also, are you able to confirm that your sklearn is u

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Sats S** June 10, 2017 at 11:10 am #

REPLY ↩

Yes i copied the code exactly as on the site. sklearn: 0.18.1  
 thoughts?



**Jason Brownlee** June 11, 2017 at 8:20 am #

REPLY ↩

I'm not sure but I expect it has something to do with your environment.

This tutorial may help with your environment:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Rene** June 11, 2017 at 1:25 am #

REPLY ↩

Very insightful Jason, thank you for the post!

Your Start in Machine Learning

I was wondering if the models can be saved to/loaded from file, to avoid re-training a model each time we wish to make a prediction.

Thanks,

Rene



**Jason Brownlee** June 11, 2017 at 8:26 am #

REPLY ↩

Yes, see this post:

<http://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/>



**Richard Bruning** June 12, 2017 at 11:42 am #

Mr. Brownlee,

This is, by far, is the most effective applied technology

You get right to the point and still have readers actually and of course machine learning. I am an electromechanic now, I have been bogged down trying to traipse through verbose machine learning theory knowing there exists

Thank you for showing me the way!

Rich

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 13, 2017 at 8:13 am #

REPLY ↩

Thanks Rich, you made my day! I'm glad it helped.



**Praver Vats** June 13, 2017 at 7:21 pm #

REPLY ↩

This was very informative....Thank You !

Actually I was working on a project on twitter analysis using python where I am extracting user interests through their tweets. I was thinking of using naive bayes classifier in textblob python library for training classifier with different type of pre-labeled tweets or different categories like politics,sports etc. My only concern is that will it be accurate as I tried passing like 10 tweets in training set and based on that I tried classifying my test set. I am getting some false cases and accuracy is around 85.

Your Start in Machine Learning



**Jason Brownlee** June 14, 2017 at 8:44 am #

REPLY ↗

Good question, I'd suggest try it and see.



**Kush Singh Kushwaha** June 14, 2017 at 4:14 am #

REPLY ↗

Hi Jason,

This was great example. I was looking for something similar on internet all this time, glad I found this link. I wanted to compile a ML code end-to-end and see my basic infra is ready to start with the actual course work. As you said, from here we can learn more about start a Youtube channel and upload some easy to learn Neural Networks.

Regards,  
Kush Singh



**Jason Brownlee** June 14, 2017 at 8:51 am #

X

Thanks.

Take a look at the rest of my blog and my books. I

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Shaksham Kapoor** June 14, 2017 at 4:34 am #

REPLY ↗

I've been working on a dataset which contains [Male,Female,Infant] as entries in first column rest all columns are integers. How can I replace [Male,Female,Infant] with a similar notation like [0,1,2] or something like that ? What is the most efficient way to do it ?



**Jason Brownlee** June 14, 2017 at 8:51 am #

REPLY ↗

Excellent question.

Use a LabelEncoder:

<http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>

I'm sure I have tutorials on this on my blog, try the blog search.

**Dev** June 14, 2017 at 12:52 pm #

Your Start in Machine Learning



Sir, while loading dataset we have given the URI but what if we already have one and wants to load it ?



**Jason Brownlee** June 15, 2017 at 8:42 am #

REPLY ↩

Change the URL to a filename and path.



**Vincent** June 18, 2017 at 2:26 am #

REPLY ↩

Hi,

Nice tutorial, thanks!

Just a little precision if someone encounter the same issue, if you get the error “This application failed to start because ‘windows’”

in “.” when you are trying to see your data visualization using PySide rather than PyQt.

In that case, add these lines before the “import matplotlib”

```
import matplotlib
matplotlib.use('Qt4Agg')
matplotlib.rcParams['backend.qt4']='PySide'
```

Hope this will help

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 18, 2017 at 6:33 am #

REPLY ↩

Thanks for the tip Vincent.



**Danielle** June 25, 2017 at 5:43 pm #

REPLY ↩

Fantastic tutorial! Running today I noticed two changes from the tutorial above (undoubtedly because time has passed since it was created). New users might find the following observations useful:

#1 – Future Warning

Ran on OS X, Python 3.6.1, in a jupyter notebook, anaconda 4.4.0 installed:

scipy: 0.19.0

numpy: 1.12.1

matplotlib: 2.0.2

pandas: 0.20.1

sklearn: 0.18.1

## Your Start in Machine Learning

I replaced this line in the #Load libraries code block:

```
from pandas.tools.plotting import scatter_matrix
```

With this:

```
from pandas.plotting import scatter_matrix
```

...because a FutureWarning popped up:

```
/Users/xxx/anaconda/lib/python3.6/site-packages/ipykernel_launcher.py:2: FutureWarning:  
'pandas.tools.plotting.scatter_matrix' is deprecated, import 'pandas.plotting.scatter_matrix' instead.
```

Note: it does run perfectly even without this fix, this may be more of an issue in the future

#2 – SVM wins!

In the build models section, the results were:

```
LR: 0.966667 (0.040825)  
LDA: 0.975000 (0.038188)  
KNN: 0.983333 (0.033333)  
CART: 0.966667 (0.040825)  
NB: 0.975000 (0.053359)  
SVM: 0.991667 (0.025000)
```

... which means SVM was better here. I added the following code:

```
# Make predictions on validation dataset  
svm = SVC()  
svm.fit(X_train, Y_train)  
predictions = svm.predict(X_validation)  
print(accuracy_score(Y_validation, predictions))  
print(confusion_matrix(Y_validation, predictions))  
print(classification_report(Y_validation, predictions))
```

which gets these results:

```
0.933333333333  
[[ 7 0 0]  
[ 0 10 2]  
[ 0 0 11]]  
precision recall f1-score support  
Iris-setosa 1.00 1.00 1.00 7  
Iris-versicolor 1.00 0.83 0.91 12  
Iris-virginica 0.85 1.00 0.92 11  
avg / total 0.94 0.93 0.93 30
```

I did also run the unmodified KNN block – # Make predictions on validation dataset – and got the exact results that were in the tutorial.

Excellent tutorial, very clear, and easy to modify 😊

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



Thanks for sharing Danielle.



**mr. disappointed** June 26, 2017 at 10:06 pm #

REPLY ↗

So this intro shows how to set everything up but not the actual interesting bit how to use it?



**Jason Brownlee** June 27, 2017 at 8:29 am #

REPLY ↗

What do you mean exactly? Putting the model into production? See here:

<http://machinelearningmastery.com/deploy-machine-learning/>



**Aditya** June 28, 2017 at 4:48 pm #

Excellent tutorial sir, I love your tutorials and I would love if you could provide a tutorial for sequence dataset.

Also I would be obliged if you could point me in some seq2seq

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 29, 2017 at 6:29 am #

REPLY ↗

I have one here:

<http://machinelearningmastery.com/learn-add-numbers-seq2seq-recurrent-neural-networks/>



**RATNA** June 30, 2017 at 4:19 am #

REPLY ↗

Hi Jason,

Awesome tutorial. I am working on PIMA dataset and while using the following command  
`# head  
print(dataset.head(20))`

I am getting NAN. HEPL ME.



**Jason Brownlee** June 30, 2017 at 8:18 am #

REPLY ↗

## Your Start in Machine Learning

Confirm you downloaded the dataset and that the file contains CSV data with nothing extra or corrupted.



**RATNA** June 30, 2017 at 4:14 pm #

REPLY ↩

Hi Jason,

I downloaded the dataset from UCI which is a CSV file but still I get NAN.

```
# Load dataset url = "https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data"
```

Thanks..



**Jason Brownlee** July 1, 2017 at 6:2

Sorry, I do not see how this could



**Deepak** July 2, 2017 at 1:50 am #

Hello Jason,

Thank you for a great tutorial.

I have noticed something , which I would like to share with you.

I have tried with random\_state = 4

```
"X_train,X_validation,Y_train,Y_validation = model_selection.train_test_split(X,Y, test_size = 0.2, random_state = 4)"
```

and surprisingly now “LDA” has the best accuracy.

LR: 0.966667 (0.040825)

LDA: 0.991667 (0.025000)

KNN: 0.975000 (0.038188)

CART: 0.958333 (0.055902)

NB: 0.950000 (0.055277)

SVM: 0.983333 (0.033333)

Any thoughts on this?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 2, 2017 at 6:33 am #

REPLY ↩

Machine learning algorithms are stochastic  
<http://machinelearningmastery.com/randomness-in-machine-learning/>

Your Start in Machine Learning



**Rui** July 3, 2017 at 12:31 pm #

REPLY ↗

Hi Jason,

Thanks for your great example, this is really helpful, this end-to-end project is the best way to learn ML, much better than text-book which they only focus on the separate concepts, not the whole forest, will you please do more example like this and explain in detail next time?

Thanks,

Rui



**Jason Brownlee** July 6, 2017 at 9:57 am #

Thanks.



**Vaibhav** July 4, 2017 at 4:33 pm #

`__init__()` got an unexpected keyword argument

I am getting this error while running the code upto “pri  
Can you please help me removing it.

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 6, 2017 at 10:12 am #

REPLY ↗

Update your version of sklearn to 0.18 or higher.



**Fahad Ahmed** July 5, 2017 at 12:31 am #

REPLY ↗

This is beautiful tutorial for the starters..

I am a lover of machine learning and want to do some projects and research on it.  
I would really need your help and guideline time to time.

Regards,  
Fahad



**Jason Brownlee** July 6, 2017 at 10:19 am #

Your Start in Machine Learning

Thanks.



**Neal Valiant** July 12, 2017 at 9:08 am #

REPLY ↗

Hi Jason,

Love the article. gave me a good start of understanding machine learning. One thing i would like to ask is what is the predicted outcome? Is it which type or “class” of flower that will happen next? i assume switching things up I could use this same outline as a way of getting a prediction on the other columns involved?



**Jason Brownlee** July 12, 2017 at 9:55 am #

Yes, the prediction is a number that maps

Correct, from the class and other measures you can



**Neal** July 13, 2017 at 3:50 am #

Hi again Jason,

Diving deeper into this tutorial and analyzing more data, you can see that the KNN algorithm has a higher accuracy percentage than the LDA algorithm. This means that the KNN algorithm is better at predicting the class of a new flower based on its features. The LDA algorithm has a lower accuracy percentage, which means that it is less accurate at predicting the class of a new flower based on its features.



**Jason Brownlee** July 13, 2017 at 9:59 am #

REPLY ↗

Machine learning algorithms are stochastic.

It is important to develop a robust estimate of the performance of machine learning models on unseen data using repeats. See this post:

<http://machinelearningmastery.com/evaluate-skill-deep-learning-models/>



**Neal** July 13, 2017 at 11:22 am #

Another great read Jason. This whole site is full of great pieces and it gives me a good answer on my question. I want to thank you for your time and effort into making such a great place for all this knowledge.

Your Start in Machine Learning



**Jason Brownlee** July 13, 2017 at 4:54 pm #

Thanks, I'm glad it helps Neal. Stick with it!

REPLY ↗



**Thomas** July 14, 2017 at 8:10 pm #

Hello Jason,

At the beginning of your tutorial you write: "If you are a machine learning beginner and looking to finally get started using Python, this tutorial was designed for you."

No offense but in this regards, your tutorial is not doing

You don't really go in detail so that we can understand rather weak.

Wrong expectations set i believe.

Cheers,

Thomas

## Your Start in Machine Learning

X

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 15, 2017 at 9:43 am #

It is a starting point, not a panacea.

Sorry that it's not a good fit for you.

REPLY ↗



**Mariah** July 15, 2017 at 7:11 am #

Hi Jason! I am trying to adapt this for a purely binary dataset, however I'm running into this problem:

```
# evaluate each model in turn
results = []
name = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

I get the error:

```
raise ValueError("Unknown label type: %r" % y_type)
```

Your Start in Machine Learning

ValueError: Unknown label type: 'unknown'

Am I missing something, any help would be great!



**Mariah** July 15, 2017 at 7:12 am #

REPLY ↗

All necessary indentations are correct, it just pasted incorrectly



**Jason Brownlee** July 15, 2017 at 9:46 am #

REPLY ↗

You can wrap pasted code in pre tags



**Jason Brownlee** July 15, 2017 at 9:46 am #

Sorry, the fault is not obvious to me.



**Daniel** September 12, 2017 at 1:14 am #

Hello Mariah,

Did you ever get a solution to this problem?

Jason..great guide here..THANKS!



**Sreeram** July 16, 2017 at 10:09 pm #

REPLY ↗

Hi. What should i do to make predictions based on my own test set.? Say i need to predict category of flower with data [5.2, 1.8, 1.6, 0.2]. ie i want to change my X\_test to that array. And the prediction should be like "setosa".

What changes should i do.? I tried giving that value directly to predict(). But it crashes.



**Jason Brownlee** July 17, 2017 at 8:47 am #

REPLY ↗

Correct.

Fit the model on all available data. This is called creating a final model:

<http://machinelearningmastery.com/train-final-model/>

Your Start in Machine Learning

Then make your prediction on new data where you do not know the answer/outcome.

Does that help?



**Sreeram** July 18, 2017 at 2:35 am #

REPLY ↗

Yes it helped. Can u show an example code for the same.?



**Jason Brownlee** July 18, 2017 at 8:46 am #

REPLY ↗

Sure:

```
1 # train on all data
2 model = ...
3 # make prediction on new 1D instance
4 result = model.predict(newX)
```



**Joe** July 18, 2017 at 7:49 am #

Hi Jason, i'm perú and i have to script write in python  
#Configurar para la red neural  
fechantedinicio = '1970-01-01'  
fechantedfinal = '1974-12-31'  
capasinicio = TodasEstaciones.ix[fechantedinicio:fechantedfinal].as\_matrix()[:,0,2,5]  
capasalida = TodasEstaciones.ix[fechantedinicio:fechantedfinal].as\_matrix()[:,1]  
#Construimos la Red Neural

```
from sknn.mlp import Regressor, Layer

neuronas = 8
tasaaprendizaje = 0.0001
numiteraciones = 7000

#Definition of the training for the neural network
redneural = Regressor()
layers=[Layer("ExpLin", units=neuronas),
Layer("ExpLin", units=neuronas), Layer("Linear")],
learning_rate=tasaaprendizaje,
n_iter=numiteraciones)
redneural.fit(capasinicio, capasalida)
```

#Get the prediction for the train set  
valortest = ([])

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

```
for i in range(capasinicio.shape[0]):
    prediccion = redneural.predict(np.array([capasinicio[i,:].tolist()]))
    valortest.append(prediccion[0][0])
```

and then run...

```
ModuleNotFoundError Traceback (most recent call last)
in ()
1 #Construimos la Red Neural
2
--> 3 from sknn.mlp import Regressor, Layer
4
5
```

ModuleNotFoundError: No module named 'sknn'  
i have install python in window 7 and i changed the sc

```
#construimos la red neural
import numpy as np
from sklearn.neural_network import MLPRegressor
#definicion del entrenamiento para el trabajo de la red
redneural = MLPRegressor(
hidden_layer_sizes=(100,), activation='relu', solver='ad
learning_rate='constant', learning_rate_init=0.01, power
random_state=0, tol=0.0001, verbose=False, warm_st
nesterovs_momentum=True,
early_stopping=False, validation_fraction=0.1, beta_1=0.1, beta_2=0.999, epsilon=1e-05,
```

redneural.fit(capasinicio,capasalida) and then shift + enter the run never end.

Thanks for your time.

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** July 18, 2017 at 8:49 am #

REPLY ↗

Consider posting to stackoverflow.

REPLY ↗



**Angel** July 18, 2017 at 6:06 pm #

Hello Jason, this is a fantastic tutorial! I am using this as a template to experiment with a dataset that has 0 or 1 as a value for each attribute and keep running into this error:

```
# Load libraries
import numpy
from matplotlib import pyplot
from pandas import read_csv
from pandas import set_option
```

Your Start in Machine Learning

```

from pandas.tools.plotting import scatter_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import ExtraTreesClassifier

# Load Dataset
filename = 'ML.csv'
names = ['Cities', 'Entertainment', 'RegionalFood', 'WesternFood', 'SuperBowl', 'Manufacturing']
data = read_csv(filename, names=names)
print(data.shape)
# types
set_option('display.max_rows', 500)
print(data.dtypes)
# head
set_option('display.width', 100)
print(data.head(20))
# descriptions, change precision to 3 places
set_option('precision', 3)
print(data.describe())
# class distribution
print(data.groupby('Cities').size())
# histograms
data.hist(sharex=False, sharey=False, xlabelsize=1, ylabelsize=1)
pyplot.show()
# correlation matrix
fig = pyplot.figure()
ax = fig.add_subplot(111)
cax = ax.matshow(data.corr(), vmin=-1, vmax=1, interpolation='none')
fig.colorbar(cax)
pyplot.show()

```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

Your Start in Machine Learning

```
# Split-out validation dataset
array = data.values
X = array[:,1:8]
Y = array[:,8]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = train_test_split(X, Y,
test_size=validation_size, random_state=seed)
# Test options and evaluation metric
num_folds = 3
seed = 7
scoring = 'accuracy'
# Spot-Check Algorithms
models = []
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
results = []
names = []
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = KFold(n_splits=3, random_state=seed)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

I get the following error:

```
File "C:\Users\Giselle\Anaconda3\lib\site-packages\sklearn\utils\multiclass.py", line 172, in
check_classification_targets
raise ValueError("Unknown label type: %r" % y_type)

ValueError: Unknown label type: 'unknown'

runfile('C:/Users/Giselle/.spyder-py3/temp.py', wdir='C:/Users/Giselle/.spyder-py3')
```



**Jason Brownlee** July 19, 2017 at 8:22 am #

REPLY ↗

Check that you are loading your data correctly.

**machine learning guy** July 18, 2017 at 9:15 pm

Your Start in Machine Learning



hey jason.

awesome detailed blog man.....i always love your method for explanation ..so clean and easy.  
Great ... i start machine learning with r but now doing with python too.

Regards

Kuldeep



**Jason Brownlee** July 19, 2017 at 8:23 am #

REPLY ↗

Thanks.



**Aayush A** July 18, 2017 at 9:17 pm #

Hey Jason,

Your sample code is amazing to get started with ML.

When I tried to run the code myself I get an

Can you please help me rectify this?

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Jason Brownlee** July 19, 2017 at 8:23 am #

What is the problem?



**Marco Roque** July 19, 2017 at 7:01 am #

REPLY ↗

Jason

Thanks for your help !!!! The Blog is super useful ... do you have another place that you recommend to learn more about the topic .... Thanks !!!!

Best

Marco



**Jason Brownlee** July 19, 2017 at 8:31 am #

REPLY ↗

Thanks.

Yes, search “resources” on the blog.

Your Start in Machine Learning



**Yug** July 20, 2017 at 2:59 am #

REPLY ↗

Hi Jason,  
Great tutorial!! very helpful!

I am getting an error executing below piece of code, can you help?

```
# evaluate each model in turn
```

```
results = []
```

```
names = []
```

```
for name, model in models:
```

```
kfold = ms.KFold(n_splits=10, random_state=seed)
```

```
cv_results = ms.cross_val_score(model, X_train, Y_train)
```

```
results.append(cv_results)
```

```
names.append(name)
```

```
msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_
```

```
print(msg)
```

Error that I am getting:

TypeError: get\_params() missing 1 required positional argument: 'estimator'

## Your Start in Machine Learning

X

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** July 20, 2017 at 6:22 am #

REPLY ↗

Sorry, I have not seen that error before. Perhaps confirm that your environment is installed correctly?

Also confirm that you have all of the code without extra spaces?



**Yug** July 20, 2017 at 8:02 am #

REPLY ↗

Yeah, environment is installed correctly. I made sure that there are no extra spaces in the code. It is still erroring out.



**Jason Brownlee** July 21, 2017 at 9:23 am #

REPLY ↗

Sorry, I'm running out of ideas.



**Sal** August 2, 2018 at 1:07 am #

REPLY ↗

## Your Start in Machine Learning

For anyone with this issue, the problem is a missing parenthesis in the line `models.append(('LR', LogisticRegression()))`



**Jason Brownlee** August 2, 2018 at 6:02 am #

REPLY ↗

Are you sure?



**Aawesh** July 21, 2017 at 8:40 am #

REPLY ↗

Great tutorial. Loved it. What's next?



**Jason Brownlee** July 21, 2017 at 9:37 am #

X

See here:

<http://machinelearningmastery.com/start-here/#py>

And for the higher-level goals (e.g. build a portfolio)

<http://machinelearningmastery.com/start-here/#ge>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Chandana** July 21, 2017 at 8:54 am #

REPLY ↗

I get the following results when the test is run against each model.

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.966667 (0.040825)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

Looks like SVN is the best and not KNN, what is the reason for this?



**Jason Brownlee** July 21, 2017 at 9:37 am #

REPLY ↗

Machine learning algorithms are stochastic:

<http://machinelearningmastery.com/randomness-in-machine-learning/>

**samkelo jiyane** July 21, 2017 at 4:24 pm #

## Your Start in Machine Learning



Hi Jason, have started to learn Machine learning basics using Keras (with TF/Theano as backend). I am going through examples on this site and other resources with the ultimate goal of implementing Document reading/interpretation on constrained data set, e.g bank statements, proof of residence, standard supporting document etc.

Any pointers ?



**Jason Brownlee** July 22, 2017 at 8:30 am #

REPLY ↩

Great!

Yes, start here:

<http://machinelearningmastery.com/start-here/#get-started>



**Asad Ali** July 23, 2017 at 1:04 pm #

Thank you Jason for this simple tutorial for beginers.

I just want to know that what is the effect of n-folds (in cross validation). If we change n-fold, the performance of algorithm varies, how?

kfold=model\_selection.Kfold(n\_splits=10, random\_state=42)

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 24, 2017 at 6:48 am #

REPLY ↩

The number of folds, and the specifics of the algorithm and data, will impact the stability of the estimated skill of the model on the problem.

Given a lot of data, often there is diminishing returns going beyond 10.

If in doubt, test the stability of the score (e.g. variance) by estimating model performance using a suite of different k values in k cross validation.



**Nelson D'souza** July 25, 2017 at 11:08 pm #

REPLY ↩

Hi! Jason,

Thanks for this amazing article/tutorial it is really very helpful.

I was working on a predictive model of my own

I seem to be occurring a problem nobody on the forum got 😊 xD

I am sorry but could you help me out or point me in a direction

## Your Start in Machine Learning

```
#####
#####
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.ensemble import RandomForestClassifier
from sklearn import svm
from sklearn.svm import SVR

from sklearn import linear_model

import csv

from numpy import genfromtxt

import time
import datetime

from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
```

```
date = []
usage = []

date = genfromtxt('date.csv')
usage = genfromtxt('usage.csv')
test = genfromtxt('test.csv')

print (len(date))

print (len(usage))

dataframe = pd.DataFrame({
'Date': (date),
'Usage': (usage)
})

#drop NaN data's
dataframe = dataframe.dropna()
print (dataframe)

df = dataframe.drop(dataframe.index[[-1,-4]])

array = df.values
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

Your Start in Machine Learning

```

X = array[:,0:1]
Y = array[:,1]

validation_size = 0.20
seed = 7

X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)

seed = 7
scoring = 'accuracy'

# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))

# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_st
    cv_results = model_selection.cross_val_score(model, X
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

# Compare Algorithms
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()

#####
OutPut :

Date length : 366
Usage Length: 366

the data frame :

Date Usage
1 1.451587e+09 47139.0
2 1.451673e+09 85312.0

```

Your Start in Machine Learning

```
3 1.451759e+09 14301.0
4 1.451846e+09 20510.0
5 1.451932e+09 24225.0
6 1.452019e+09 30051.0
7 1.452105e+09 42228.0
8 1.452191e+09 27256.0
9 1.452278e+09 33746.0
10 1.452364e+09 30035.0
11 1.452451e+09 85844.0
12 1.452537e+09 28814.0
13 1.452623e+09 31082.0
14 1.452710e+09 21565.0
15 1.452796e+09 19095.0
16 1.452883e+09 15995.0
17 1.452969e+09 6578.0
18 1.453055e+09 96143.0
19 1.453142e+09 20503.0
20 1.453228e+09 31373.0
21 1.453315e+09 30776.0
22 1.453401e+09 39357.0
23 1.453487e+09 45955.0
24 1.453574e+09 21379.0
25 1.453660e+09 43682.0
26 1.453747e+09 51304.0
27 1.453833e+09 47333.0
28 1.453919e+09 33629.0
29 1.454006e+09 24185.0
30 1.454092e+09 47052.0
...
336 1.480531e+09 74882.0
337 1.480617e+09 100712.0
338 1.480703e+09 45929.0
339 1.480790e+09 84837.0
340 1.480876e+09 85755.0
341 1.480963e+09 47184.0
342 1.481049e+09 62122.0
343 1.481135e+09 38140.0
344 1.481222e+09 46333.0
345 1.481308e+09 99399.0
346 1.481395e+09 101814.0
347 1.481481e+09 34078.0
348 1.481567e+09 45800.0
349 1.481654e+09 63657.0
350 1.481740e+09 33371.0
351 1.481827e+09 34921.0
352 1.481913e+09 33162.0
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

Your Start in Machine Learning

```
353 1.481999e+09 96179.0
354 1.482086e+09 27527.0
355 1.482172e+09 42291.0
356 1.482259e+09 112647.0
357 1.482345e+09 19299.0
358 1.482431e+09 52011.0
359 1.482518e+09 37571.0
360 1.482604e+09 78809.0
361 1.482691e+09 31469.0
362 1.482777e+09 69469.0
363 1.482863e+09 42879.0
364 1.482950e+09 31009.0
365 1.483036e+09 130637.0
```

[365 rows x 2 columns]

LR: 0.000000 (0.000000)

/Users/nelsonsouza/anaconda/lib/python2.7/site-pac

UserWarning: The priors do not sum to 1. Renormalizing UserWarning)

Traceback (most recent call last):

File "data\_0.py", line 111, in

cv\_results = model\_selection.cross\_val\_score(model, X,

File "/Users/nelsonsouza/anaconda/lib/python2.7/site-pac  
line 140, in cross\_val\_score  
for train, test in cv\_iter)

File "/Users/nelsonsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line  
758, in \_\_call\_\_

while self.dispatch\_one\_batch(iterator):

File "/Users/nelsonsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line  
608, in dispatch\_one\_batch

self.\_dispatch(tasks)

File "/Users/nelsonsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line  
571, in \_dispatch

job = self.\_backend.apply\_async(batch, callback=cb)

File "/Users/nelsonsouza/anaconda/lib/python2.7/site-

packages/sklearn/externals/joblib/\_parallel\_backends.py", line 109, in apply\_async

result = ImmediateResult(func)

File "/Users/nelsonsouza/anaconda/lib/python2.7/site-

packages/sklearn/externals/joblib/\_parallel\_backends.py", line 326, in \_\_init\_\_

self.results = batch()

File "/Users/nelsonsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line  
131, in \_\_call\_\_

return [func(\*args, \*\*kwargs) for func, args, kwargs in self.items]

File "/Users/nelsonsouza/anaconda/lib/python2.7/site-

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```
line 238, in _fit_and_score
estimator.fit(X_train, y_train, **fit_params)
File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/discriminant_analysis.py", line
468, in fit
self._solve_svd(X, y)
File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/discriminant_analysis.py", line
378, in _solve_svd
fac = 1. / (n_samples - n_classes)

ZeroDivisionError: float division by zero
```



**Jason Brownlee** July 26, 2017 at 7:55 am #

Sorry, I cannot debug your code. Consider



**Nelson D'souza** July 26, 2017 at 3:40 pm #

ok, Thanks 😊 Have a nice day!



**Nelson D'souza** July 26, 2017 at 6:49 pm #

I just thought I would let you know

my data set has 365 rows and only 2 columns is that a problem ?

Also I had a question, if you could lead me in a correct direction,  
If my dataset has a column 'Dates' .datetime object how should I go about handling it ?

thanks in advance 😊



## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Jason Brownlee** July 27, 2017 at 7:58 am #

Sounds like a time series forecasting problem. You should treat it differently.

Start here with time series forecasting:

<http://machinelearningmastery.com/start-here/#timeseries>

REPLY ↗



**Soumya** July 27, 2017 at 8:08 pm #

Your Start in Machine Learning

Awesome tutorial.. The program ran so smoothly without any errors. And it was easy to understand. Graphs looked fantastic. Although I could not understand each and every functionality. Do you have any reference to understand the very basics of machine learning in Python?

Thanks for your help.



**Jason Brownlee** July 28, 2017 at 8:31 am #

REPLY ↗

Yes, start right here:

<http://machinelearningmastery.com/start-here/#python>



**Razack** July 29, 2017 at 3:46 pm #

Hi Jason,

Very nice tutorial. This helped me a lot.

Is there a way to append the train set with new data so the train model. What I could see creating new train sets.

Please help

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Jason Brownlee** July 30, 2017 at 7:39 am #

Not sure I follow.

Once you choose a model, you can fit a final model on all available data and start using it to make predictions on new data.

You may want to update your model in the future, in which case you can use the same process above with new data.

Does that help?

REPLY ↗



**Dexter D'Silva** August 2, 2017 at 11:34 pm #

Thank you Jason!!!

Having done the Coursera ML course by Andrew Ng I wasn't sure where to go next.

Your clear and well explained example showed me the way!!! Looking forward to reading your other material and spending many many more hours learning and having fun. (And my first foray into Python wasn't as daunting as I expected thanks to you).

## Your Start in Machine Learning



**Jason Brownlee** August 3, 2017 at 6:51 am #

REPLY ↗

Thanks Dexter, well done on working through the tutorial!



**Gerry** August 3, 2017 at 5:51 am #

REPLY ↗

Hi Jason, I am using your tutorial for my own ML model and it's fantastic! I'm trying to predict make prediction on new data and am using

NB=GaussianNB()

new\_prediction = predict.nb(new data)

print(new\_prediction)

I am able to successfully get one prediction, how can I get 15 possible classifications and I'd like the predict function to return all predictions

Any help would be greatly appreciated, thank you so much!



**Jason Brownlee** August 3, 2017 at 6:57 am #

X

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



It sounds like your problem is a multi-class classification problem.

If so, you can predict probabilities and select the top N classes.

For example:

```
1 probabilities = model.predict_proba(X)
```



**Gerry** August 3, 2017 at 8:54 am #

REPLY ↗

Thanks, how can I match the probabilities to the class, or is there a way to have it return the class name?



**Gerry** August 3, 2017 at 9:08 am #

REPLY ↗

Here is the code:

```
ACN_prediction = NB.predict_proba([[ 0.80, 0.20, 0.70, 0.30, 0.99, 0.01, 0.98, 0.02, 0.95, 0.05,
0.95, 0.05, 1.00, 0]])
```

```
print (ACN_prediction)
```

And the result only displays:

```
[[ 0. 0. 0. ..., 0. 1. 0.]]
```

Your Start in Machine Learning

Is it just giving me the probabilities I have typed in?



**Jason Brownlee** August 4, 2017 at 6:44 am #

REPLY ↩

Each class is assigned an integer which is an index in the output array. This is done when you one hot encode the output variable.



**Gerry** August 3, 2017 at 9:30 am #

REPLY ↩

Using just the `NB.predict([[list of new data]])` I would get the class 'Flower'

-Sorry for the long winded question, I have been stuck



**Jason Brownlee** August 4, 2017 at 6:45 am #

If you just want one class label, then you `predict()` instead.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Gerry** August 4, 2017 at 10:20 am #

REPLY ↩

If I want it to predict n best class labels I need to use `predict_proba` and manually match the n best probabilities to their class label correct? There is no other way to to yield the top 5 class labels?



**Jason Brownlee** August 4, 2017 at 3:41 pm #

REPLY ↩

Yes. Correct.



**Gerry** August 5, 2017 at 6:10 am #

REPLY ↩

Thank you!



**Jason Brownlee** August 6, 2017 at 7:27 am #

REPLY ↩

I'm glad it helped.

## Your Start in Machine Learning



Fernando D Mera August 10, 2017 at 1:54 am #

REPLY ↗

Hello, Jason,

I am using python3 on my mac, and I am also using Jupyter notebooks in order to complete the assignment on this webpage. Unfortunately, when I save the Iris dataset in my Desktop folder, and then run the command `# shape print(dataset.shape)`, the output is `(193, 5)`

As you know, the output should be `(150,5)` and I am not sure why the dimensions of the dataset are wrong. Also, I tried to use the archive: <https://archive.ics.uci.edu/ml/machine-learning-databases/iris/> the Jupyter output was the following

```
-----
SSLError Traceback (most recent call last)
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in _opener(self, *args, **kwargs)
    1317     h.request(req.get_method(), req.selector, req.data, 
-> 1318         encode_chunked=req.has_header('Transfer-encoding'))
1319     except OSError as err: # timeout error
1320
1321     /Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in _send_request(self, 
url, body, headers, encode_chunked)
1322     1238     """Send a complete request to the server."""
-> 1323     self._send_request(method, url, body, headers, encode_chunked)
1324
1325     /Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in _send_request(self, 
method, url, body, headers, encode_chunked)
1326     1284     body = _encode(body, 'body')
-> 1325     self.endheaders(body, encode_chunked=encode_chunked)
1326
1327     /Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in endheaders(self, 
message_body, encode_chunked)
1328     1233     raise CannotSendHeader()
-> 1329     self._send_output(message_body, encode_chunked=encode_chunked)
1330
1331     /Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in _send_output(self, 
message_body, encode_chunked)
1332     1025     del self._buffer[:]
-> 1333     self.send(msg)
1334
1335     /Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in send(self, data)
963     if self.auto_open:
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

-> 964 self.connect()

965 else:

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in connect(self)

1399 self.sock = self.\_context.wrap\_socket(self.sock,

-> 1400 server\_hostname=server\_hostname)

1401 if not self.\_context.check\_hostname and self.\_check\_hostname:

How can I get the correct dimensions of the Iris dataset?



**Jason Brownlee** August 10, 2017 at 6:59 am #

REPLY ↩

Perhaps confirm that you downloaded the

Also, try running from the command line instead of challenging faults.



**Andrew Revoy** August 14, 2017 at 7:39 am #

I've been eyeballing this tutorial for a while and clear intro into machine learning! This has been the one evaluating the data / different models right off that bat.

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** August 15, 2017 at 6:26 am #

REPLY ↩

Thanks Andrew, and well done on working through it!



**Abi Yusuf** August 14, 2017 at 10:02 pm #

REPLY ↩

Hi Jason,

My sincere gratitude for this work you do to help us all out with ML. I have also been working away at this very wonderful field over the last 3 years now ( PhD research – studying gaze patterns and trying to build predictive models of gaze patterns which represent some sort of behavior). In any case, I was reviewing the code you built here and I was just thinking that I don't tend to declare the test\_size explicitly or the random\_state either – I just put it directly into the algorithm

so, your code goes:

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed) – totally spot on by the way,
```

My small addition/improvement – if you can call it that

Your Start in Machine Learning

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size= 0.2,
random_state= 7)
```

# test\_size keyword argument surely invokes the split method of the train\_test\_split module (I think) – meaning that the algorithm automatically assigns 80% to the training set and 20% to the test set

would you agree with this method? My python 3.x installation accepts this method just fine –

Also , I don't know if anyone else might have suggested this, but it is also worth pointing out that for cross\_val (cv) – the fold size can be quite resource intensive and also there are underfitting/overfitting issues to be aware of, when doing cross validation –

Can you sense check these thoughts please?

Many Thanks.

Cheers



**Jason Brownlee** August 15, 2017 at 6:36 am #

Evaluating algorithms is an important topic

Indeed the number of folds is important and we must consider the number of folds in relation to the broader problem.

As for specifying the test size a different way, that's a good question. The key is developing unbiased estimates of model performance.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Paul Wilson** January 11, 2019 at 2:32 am #

REPLY ↗

This is the bit where I'm currently stuck – when I type in the command:

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y,
test_size=validation_size, random_state=seed)
```

the shell hangs – or at least it isn't completing within 20 minutes or so. I'm guessing that shouldn't be the case on this small dataset?



**Jason Brownlee** January 11, 2019 at 7:53 am #

REPLY ↗

Are you running from the command line?

More help here:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-run-a-script-from-the-command-line>

Your Start in Machine Learning



**Sarbani** August 15, 2017 at 5:08 am #

REPLY ↗

Thank you, Jason Brownlee, the post is very helpful. I was really lost in so many articles, blogs, open source tools. I was not able to understand how to start ML. Your post really helped me to start at least. I installed ANACONDA, ran the classification model successfully.

Next Step – Understand the concept and apply on some real use cases.



**Jason Brownlee** August 15, 2017 at 6:44 am #

REPLY ↗

Well done Sarbani!



**Ryan Stoddard** August 15, 2017 at 3:39 pm #

Thanks for this extremely helpful example. I just was a little confused. It seems to me that you withhold cross-validation on only the 80% training data, then train with 20% validation data. Is this correct, and if so is it to get statistics about the best model is to simply use Why do you only perform cross-validation on 80% of the data with a single validation set?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** August 15, 2017 at 4:57 pm #

REPLY ↗

Great question Ryan!

We hold back a test set so that if we over fit the model via repeated cross validation (e.g. parameter tuning), we still have a final way of checking to see if we have fooled ourselves.

More here:

<http://machinelearningmastery.com/difference-test-validation-datasets/>



**vishnu** August 15, 2017 at 7:51 pm #

REPLY ↗

you above mention that scipy. it didn't available in pycharm (windows)..can u suggest another package for machine learning...?



**Jason Brownlee** August 16, 2017 at 6:33 am #

REPLY ↗

## Your Start in Machine Learning

This tutorial will help you set up your environment:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Adam Drake** August 17, 2017 at 11:23 pm #

REPLY ↗

The link to download the “iris.dat” file appears to be broken!



**Jason Brownlee** August 18, 2017 at 6:20 am #

REPLY ↗

Here is the direct link:

<https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>



**Ravindra Singh** August 17, 2017 at 11:32 pm #

REPLY ↗

Thanks. Loved your result-first approach... No problem. Hoping i would succeed !

A question

Given i will not have all the time to master writing new code, am an average developer from the past,(and new to Python but find it easy). I am thinking i should rather master how to prepare, present and interpret data – i understand domain very well – , and understand which algorithm (and libraries) to use for best results. I am guessing that, even to master applied ML, it will take many real projects !

I am keen in using ML in predicting data quality problems such as outliers that may need correction. any pointers ?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** August 18, 2017 at 6:22 am #

REPLY ↗

Thanks Ravindra!

No, I recommend using a library, here's more on the topic:

<http://machinelearningmastery.com/dont-implement-machine-learning-algorithms/>

My best advice is to first collect a lot of data.



**Brendan** August 17, 2017 at 11:34 pm #

REPLY ↗

## Your Start in Machine Learning

I am getting an error on the line starting with predictions?

```
# Make predictions on validation dataset
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
predictions = knn.predict(X_validation)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

I am using Python 3, is there something else I need to install



**Jason Brownlee** August 18, 2017 at 6:24 am #

What error?

This tutorial will show you how to setup your environment:  
<http://machinelearningmastery.com/setup-python-anaconda/>



**Ankith** August 18, 2017 at 4:56 am #

Hey Jason!!!!...Thanks for this!!!!...Also I appreciate your help! I guess an year!!! . I wish you good luck 😊



**Jason Brownlee** August 18, 2017 at 6:28 am #

REPLY ↩

Thanks Ankith, I'm glad the tutorial helped you.



**fb** August 18, 2017 at 9:54 am #

REPLY ↩

Thx a lot! Very helpful!



**Jason Brownlee** August 18, 2017 at 4:38 pm #

REPLY ↩

You're welcome.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**beginner** August 18, 2017 at 10:32 pm #

REPLY ↗

thank you this was really helpful >> too many indices for array  
 so I give him the data in 2 dimension instead of 1-D and use this >>> numpy.loadtxt( dataset ,  
 delimiter=None , ndmin=2) but he give me this error>>> could not convert string to float ,maybe because  
 there are float and string in the iris file  
 what's the solution please I have to split them 😞  
 i'm really sorry for the bad english and thank you again <3



**Jason Brownlee** August 19, 2017 at 6:20 am

REPLY ↗

Check your data file to makes sure it is a



**beginner** August 19, 2017 at 6:48 pm #

X

can you show me what do mean  
 my data file is the url you post it here, not an u  
 how can I do insure of this?( CSV file with no e

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Jason Brownlee** August 20, 2017 at 6:05 am #

REPLY ↗

Use the filename or URL to load a file. It is that simple.

REPLY ↗



**beginner** August 18, 2017 at 10:44 pm #

Sorry I don't know where the rest of the previous comment disappeared>>so i a got a question  
 how could I separate the data such like this  
`features = dataset[:,0:4]`  
`classification = dataset[:,4]`  
 which is mean in other words when I write `print (dataset.shape)` I want him to give me :  
`(150,4)` instead of `(150,5)` I told you that first I try to do this but he told me >> too many indices for array...  
 continue reading at the beginning in the comment above

REPLY ↗



**Xav** August 19, 2017 at 3:03 am #

I'd like to thank you for this concise but very helpful tutorial. I'm new to python and all the the code  
 is clear apart the following part:

Your Start in Machine Learning

```
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

It's not clear to me how this 'for' cycle works. Specifically what is name and model?



**Jason Brownlee** August 19, 2017 at 6:23 am #

It is evaluating the model using 10 fold cross-validation. This means that the data is split into 10 equal parts. Each part is used as a test set once, and the other 9 parts are used as training sets. The process is repeated 10 times, and the average score is calculated.

Does that help?



**beginner** August 19, 2017 at 7:19 am #

did you mean to write this command?

```
dataset = pandas.read_csv(url, names = parameters)
```

I did like you do in this lecture and imported the data file from the link ,But still can not separate the data

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Jason Brownlee** August 20, 2017 at 6:03 am #

REPLY ↗

What is the problem exactly?



**Cole** August 27, 2017 at 6:28 am #

REPLY ↗

I think what he is trying to say is: he followed the tutorial as required, but once he got to the part where he had to load the iris dataset, he received a traceback from the line "dataset = pandas.read\_csv(url, names = parameters)" in the python code provided. The traceback I received from this line was "NameError: name 'pandas' is not defined. Currently trying to fix, If I solve it before you get a chance to reply I will make sure to comment back on this thread what the problem was and how I fixed it.

Your Start in Machine Learning



**Cole** August 27, 2017 at 7:01 am #

REPLY ↗

for section 2.2 to fix this error, imported panda along with the script. hopefully this did the trick. I do not understand why pandas needed to be imported again, but, i did it.

```
# Load dataset
import pandas
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)
print("its goin")
```



**Jason Brownlee** August 28, 2017 a

Glad to hear it.



**Jason Brownlee** August 28, 2017 at 6:42

It sounds like pandas is not installed.

This tutorial will help you install pandas and get started with machine learning in Python:  
<http://machinelearningmastery.com/setup-python-machine-learning-anaconda/>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Ernst** August 20, 2017 at 8:29 am #

REPLY ↗

Wow. Great easy to use and understand example. It worked 100% for me. Thanks



**Jason Brownlee** August 21, 2017 at 6:04 am #

REPLY ↗

Thanks Ernst, I'm glad to hear that. Well done!



**Dharik** August 20, 2017 at 8:40 pm #

REPLY ↗

Hi Jason,

I found an error like this pls help me out.

## Your Start in Machine Learning

```
# Compare Algorithms
... fig = plt.figure()
>>> fig.suptitle('Algorithm Comparison')
```



**Jason Brownlee** August 21, 2017 at 6:05 am #

REPLY ↩

Looks like a typo, change it to fig.subtitle()



**Dharik** August 22, 2017 at 5:01 pm #

DEDIV ↩

But I copied it from your blog post.



**Jason Brownlee** August 23, 2017 a

Oh, my mistake.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Seaturtle** February 19, 2019 at 9:

Actually, it appears that `_sup_title` is correct; 'subtitle' is not recognized. (For me, it didn't work with 'subtitle', but worked like a charm with 'suptitle' which must stand for something like "supratitle"....)



**Dharik** August 22, 2017 at 7:21 pm #

REPLY ↩

And I would like to create dataset, which is precisely focused on handwritten language recognition using RNN. Would you please share some of your ideas, thoughts and resources.



**Jason Brownlee** August 23, 2017 at 6:45 am #

REPLY ↩

Perhaps start here:

<https://machinelearningmastery.com/handwritten-digit-recognition-using-convolutional-neural-networks-python-keras/>

**Dharik** August 24, 2017 at 3:50 pm

Your Start in Machine Learning



Thank you Jason.



**Jeremy** August 25, 2017 at 1:16 am #

REPLY ↩

Awesome tutorial! Thanks Jason



**Jason Brownlee** August 25, 2017 at 6:44 am #

REPLY ↩

Thanks Jeremy.



**Andrew** August 25, 2017 at 2:50 am #

Hi Jason, in your post 5.1 Create a Validation I

What is seed and why did you choose #7?

Why not seed 10 or seed 5?

Andrew from Seattle

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** August 25, 2017 at 6:45 am #

REPLY ↩

Great question.

It does not matter what the value is as long as it is consistent.

See this post for a good explanation:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



**ram** August 30, 2017 at 7:48 pm #

REPLY ↩

Hi , this article is really nice.. I am executing statements..and those are also working fine..But still i am not getting what i am doing..I mean where is the logic? And what is this validation set means.What actually we are doing here? What is the intention?



**Jason Brownlee** August 31, 2017 at 6:17 am #

REPLY ↩

More on validation sets here:

<https://machinelearningmastery.com/difference-validation-test-train-sets/>

Your Start in Machine Learning

More on the process of developing a predictive model end to end here:

<https://machinelearningmastery.com/start-here/#process>

Does that help?



**KK SINGH** September 1, 2017 at 4:08 am #

REPLY ↗

Hi jason,

Getting error in implementing

dataset.plot(kind='box', subplots=True, layout=(2, 2), sharex=False, sharey=False)

as:

```
super(FigureCanvasQT, self).__init__(figure=figure)
```

TypeError: 'figure' is an unknown keyword argument

Please help me.



**Jason Brownlee** September 1, 2017 at 6:51 am #

Might be an error in the way your environment is set up.

See this tutorial to setup your environment:

<http://machinelearningmastery.com/setup-python-anaconda/>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Ellie** September 5, 2017 at 12:33 am #

REPLY ↗

Hi Jason!

When plotting the multivariate and univariate plots in Jupyter, I found them rather small. Is there a way to increase their size?

I've tried using figsize, matplotlib.rcParams nothing seems to be working. Please help me out

Thanks!



**Jason Brownlee** September 7, 2017 at 12:36 pm #

REPLY ↗

Sorry, I don't use notebooks. I find them slow, hide errors and cause a lot of problems for beginners.

**Kay** September 6, 2017 at 11:11 pm #

Your Start in Machine Learning



Thank you, Jason.

Where in the model do you specify that you are predicting "class"? Did I miss that somewhere?



**Jason Brownlee** September 7, 2017 at 12:54 pm #

REPLY ↗

You can call `model.predict()`



**Langue cedric** September 8, 2017 at 2:12 am #

REPLY ↗

Very interesting.

That is my first tutorial on Machine learning.



**Jason Brownlee** September 9, 2017 at 11:46

Thanks!



**Sirish** September 8, 2017 at 4:54 pm #

Dear Jason,

Firstly thank you very much for this wonderful blog.

i was trying this code on my project on a 8 lac rows data set

when tried

```
array = dataset.values
```

```
X = dataset.iloc[:, [0, 18]].values
```

```
y = dataset.iloc[:, 19].values
```

```
validation_size = 0.20
```

```
seed = 7
```

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)
```

My Terminal gave me an error " positional indexers are out-of-bounds "

Summary of y data set is mentioned below

```
> print(dataset.shape)
```

```
> (787353, 18)
```

Could you pl help me in resolving this error

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** September 9, 2017 at 11:53 am #

REPLY ↗

Check your array slicing!



**Garima Shrivastava** September 8, 2017 at 11:21 pm #

REPLY ↗

Hi Jason

Grt work done by u.

I just completed this tutorial on python 2.7.1.but not able to predict the new class label using some new values



**Jason Brownlee** September 9, 2017 at 11:55

Why not?



**Albert** September 11, 2017 at 3:22 am #

When doing the

```
# Load dataset
```

```
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)
```

section, terminal says

NameError: name 'pandas' is not defined

Is it that I don't have pandas installed correctly?

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** September 11, 2017 at 12:09 pm #

REPLY ↗

You need to install pandas.

See this tutorial:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>

**Prashant** September 12, 2017 at 2:34 am #

Your Start in Machine Learning



hi Jason....first of all thank for such a good tutorial.

my question is: while execution my python interpreter stuck at the following line:

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y,
test_size=validation_size, random_state=seed)
```

and it neither produce any error nor correct output.

plz short it out...Thanks in advance.

I am using python 2.7.13



**Jason Brownlee** September 13, 2017 at 12:26 pm <#>

REPLY ↗

Perhaps wait a few minutes?



**cesar** September 13, 2017 at 5:14 pm <#>

Thank you so much Mr Joson, this tutorial is very good.  
I also got this to ask, can we get the training time for each iteration?  
The training vs testing error graph as well?

thank you again for the helping

## Your Start in Machine Learning



You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 15, 2017 at 12:00 pm <#>

REPLY ↗

I'm glad it helped.

Yes, you can develop these learning graphs, learn more here:

[http://scikit-learn.org/stable/auto\\_examples/model\\_selection/plot\\_learning\\_curve.html](http://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html)



**Trung Tiep** September 13, 2017 at 6:37 pm <#>

REPLY ↗

Hi Jason,

seem this line of code doesn't work

```
dataset.plot(kind = 'box', subplots = True, layout = (2,2), sharex = False, sharey = False)
plt.show()
```

It doesn't show anything. Could you help me?

Thanks you and best regard

**Jason Brownlee** September 15, 2017 at 12:00 pm <#>

Your Start in Machine Learning



Are you able to confirm your environment is installed and working correctly:  
<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>

Are you running the example as a Python script from the command line?



**Gary** April 27, 2019 at 6:18 am #

REPLY ↩

do you have import libraries piece at the top?

for this line –

```
import matplotlib.pyplot as plt
```



**Jason Brownlee** April 27, 2019 at 6:37 am #

Yes.



**Dr. Pulak Mishra** September 14, 2017 at 5:46 pm #

Traceback (most recent call last):  
File "machinelearning1.py", line 63, in  
kfold = model\_selection.Kfold(n\_splits=10,random\_state=seed)  
AttributeError: 'module' object has no attribute 'Kfold'

I have no idea about machine learning. just blindly following the tutorial example to just get an idea what is ML.

can you tell me how am I supposed to correct this error.

I also wish you will be explaining all codes and functions in details step by step in future lessons

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 15, 2017 at 12:12 pm #

REPLY ↩

Looks like you might need to update your version of sklearn.

See this tutorial:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Chad** September 15, 2017 at 2:47 am #

REPLY ↩

Hello Jason,

## Your Start in Machine Learning

Thank you for your tutorial, it is amazing. Could you possibly do a follow up to this where you show how to package this, and use it? For instance I am not sure how to feed in new values, either manually or dynamically and then how could I store this data in a csv?



**Jason Brownlee** September 15, 2017 at 12:16 pm #

REPLY ↩

Great question.

I have some ideas about putting models into production here that might help as a start:

<http://machinelearningmastery.com/deploy-machine-learning-model-to-production/>



**Silvio Abela** September 16, 2017 at 1:29 am #

This is a superbly put tutorial for someone starting people to actually understand and gain knowledge. That's made.



**Jason Brownlee** September 16, 2017 at 8:42 pm #

Thanks Silvio. Well done for working through



**Niklas Wilke** September 18, 2017 at 9:19 pm #

REPLY ↩

dataset.hist()

plt.show()

the 5&6 bar shows a different height on sepal-length ... did they change the dataset or anything? I'm not concerned, but just curious what could cause such a difference in display/result.

I imported everything properly, except the fact that I did not install theano because I'm planning to use TF. Can that have an issue on how it deals with data? Should I install it anyway?

<https://imgur.com/a/fC1TD>



**Niklas Wilke** September 18, 2017 at 10:20 pm #

REPLY ↩

Also I get different results when running my models... for me SVM is the best.

Could that be related to the visualization displaying something else before?

Your Start in Machine Learning

-Original-

LR: 0.966667 (0.040825)  
 LDA: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.975000 (0.038188)  
 NB: 0.975000 (0.053359)  
 SVM: 0.981667 (0.025000)

-Original-

-Result-

LR: 0.966667 (0.040825)  
 LDA: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.975000 (0.038188)  
 NB: 0.975000 (0.053359)  
 SVM: 0.991667 (0.025000)

-Result-



**Jason Brownlee** September 19, 2017 at 10:30 pm #

No, machine learning algorithms are smart.

Learn more here:

<https://machinelearningmastery.com/random-forest-machine-learning-python/>



## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Xinrui Li** September 20, 2017 at 2:10 pm #

I also got SVM as the best model.

LR: 0.966667 (0.040825)  
 LDA: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.966667 (0.040825)  
 NB: 0.975000 (0.053359)  
 SVM: 0.991667 (0.025000)



**Jason Brownlee** September 19, 2017 at 7:39 am #

REPLY ↗

That is odd, I don't have any ideas.



Could there be any changes to a newer version of the installed libraries ?  
NumPy now working differently after they adjusted an algorythm or something like that ?

Maybe all who use the updated versions of all the included tools get this result ;/



**Jason Brownlee** September 23, 2017 at 5:36 am #

REPLY ↗

Machine learning algorithms are stochastic and generally give different results each time they are run:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Dan Harris** September 23, 2017 at 4:27 pm #

Same here using python 3.6 (anaconda)

```
LR: 0.966667 (0.040825)
LDA: 0.975000 (0.038188)
KNN: 0.983333 (0.033333)
CART: 0.966667 (0.040825)
NB: 0.975000 (0.053359)
SVM: 0.991667 (0.025000)
```

Followed up with:

```
# Make predictions on validation dataset
svm = SVC()
svm.fit(X_train, Y_train)
predictions = svm.predict(X_validation)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

Resulting in:

```
0.933333333333
[[ 7  0  0]
 [ 0 10  2]
 [ 0  0 11]]
precision recall f1-score support
Iris-setosa 1.00 1.00 1.00 7
Iris-versicolor 1.00 0.83 0.91 12
Iris-virginica 0.85 1.00 0.92 11
avg / total 0.94 0.93 0.93 30
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning



**Jason Brownlee** September 24, 2017 at 5:14 am #

REPLY ↗

Nice work Dan!



**Niklas Wilke** September 27, 2017 at 6:38 pm #

REPLY ↗

you say they give out different results everytime , but it seems like everyone who is going through the tutorial right now is getting the “new” results.



**Jason Brownlee** September 28, 2017 at 10:30 am #

I tried to fix the random seed to make the results reproducible within the set of libraries and make a difference.



**Jean Nunes** September 26, 2017 at 6:06 am #

Hi, I'm new to machine learning. I started studying it on my own and I found your tutorial very helpful and I was able to make it work with different changes. For example, I had to change the metric from euclidean to manhattan, and also change the number of neighbors according to the parameter (in this case, sepal length and width and petal length and width). Thanks in advance!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 26, 2017 at 2:58 pm #

REPLY ↗

You can make predictions for new observations by calling `model.predict(X)`

Does that answer your question?



**delson** September 28, 2017 at 4:05 pm #

REPLY ↗

hi sir ,can you help to make an artificial neural network on how i import my train data(weight ,biases)in python programming to classify its category in class 1 to 4 manually and input the sample as the program execute or run sir ,i have 5 neuron to test my Ai.

thanks.

## Your Start in Machine Learning



**Jason Brownlee** September 28, 2017 at 4:47 pm #

REPLY ↗

I have an example of coding a network from scratch here that you could use as a template:  
<https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/>



**Suresh Kmar** September 29, 2017 at 12:28 am #

REPLY ↗

Great tutorial sir 😊

I'm facing a problem in logistic regression with python +numpy +sklearn

How to convert all feature into float or numerical format

Thanks



**Jason Brownlee** September 29, 2017 at 5:06 pm #

You can use an integer encoding and a one-hot encoder to convert categorical variables to numerical values. I have a blog post showing how to do this (use the search).



**Keshav** October 2, 2017 at 1:43 pm #

for me the result comes different:

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

SVM is more accurate than KNN

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**vaibhav** October 6, 2017 at 7:54 am #

REPLY ↗

same results. SVM is more accurate



**Soumendra Kumar Dash** October 3, 2017 at 1:56 am #

REPLY ↗

Hey

Your Start in Machine Learning

Nice guide. I did understand everything you have done but I had a small confusion regarding the seed variable being assigned to 7. I didn't understand its significance. Can you please tell me why we have considered the variable seed and why has it been assigned to 7 and not some other random number?



**Jason Brownlee** October 3, 2017 at 5:42 am #

REPLY ↗

It is to make the example reproducible.

You can learn more about the stochastic nature of machine learning algorithms here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Abhijeet Singh** October 3, 2017 at 5:40 pm #

In section 4.2 -> Note the diagonal grouping of correlation and a predictable relationship.

If u could explain how??



**Jason Brownlee** October 4, 2017 at 5:44 am #

Because the variables change together they plotted in 2D.



**Nas** October 3, 2017 at 11:15 pm #

REPLY ↗

File "ns.py", line 42

```
cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

^

IndentationError: unexpected indent

using my dataset I found this problem. How I can solve this type of problem please advice.



**Jason Brownlee** October 4, 2017 at 5:46 am #

REPLY ↗

Make sure you copy the code exactly.



**Nas** October 4, 2017 at 12:14 pm #

Your Start in Machine Learning

```

import pandas
from pandas.tools.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import KFold
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

dataset = pandas.read_csv("/home/nasrin/nslkdd/NSL-KDD/kddcup.data_10percent")

array = dataset.values
X = array[:,0:41]
Y = array[:,41]

validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = cross_validation.train_test_split(X, Y, test_size=validation_size, random_state=seed)

num_folds = 7
num_instances = len(X_train)
seed = 7
scoring = 'accuracy'

models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))

results = []
names = []
for name, model in models:
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring= Scoring)
    results.append(cv_results)
    names.append(name)

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

```
msg = "%s: %f (%f)" % (name, cv_results.mean()*100, cv_results.std()*100)
print(msg)
```

.....  
error is

Traceback (most recent call last):

File "ns.py", line 26, in

X\_train, X\_validation, Y\_train, Y\_validation = cross\_validation.train\_test\_split(X, Y, test\_size=validation\_size, random\_state=seed)

NameError: name 'cross\_validation' is not defined



**Jason Brownlee** October 4, 2017 at 3:37 pm #

It looks like you might not have the most



**Yusuf** October 5, 2017 at 10:52 am #

It's definitely the best site I've searched for m

I wish you success in your business..

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** October 5, 2017 at 5:22 pm #

REPLY ↗

Thank you so much.



**vaibhav** October 6, 2017 at 7:52 am #

REPLY ↗

Hey, i am getting better results with the SVM algorithm, Why is it so? although we use the same data set.



**Jason Brownlee** October 6, 2017 at 11:03 am #

REPLY ↗

It is the stochastic nature of machine learning algorithms:

<https://machinelearningmastery.com/randomness-in-machine-learning/>

Also, there may have been changes to the library.

## Your Start in Machine Learning



**Amit** October 6, 2017 at 5:03 pm #

REPLY ↗

Thanks Jason! its really beautiful to learn about ML . Thanks for your effort to make it effortless.



**Jason Brownlee** October 7, 2017 at 5:49 am #

REPLY ↗

Thanks Amit.



**Davis** October 8, 2017 at 12:26 am #

REPLY ↗

Thanks Jason its real great to do this project in python.Just have one questions i long does it take to learn and

its advisable to learn python libraries for machine learning before start learn different algorithms?



**Jason Brownlee** October 8, 2017 at 8:38 am #

REPLY ↗

You can make great progress in just a few weeks

Yes, I recommend starting with Python, you can address a lot of practical problems. Get started here:  
<https://machinelearningmastery.com/start-here/#python>



**Kevin** October 8, 2017 at 4:48 am #

REPLY ↗

Does anyone offer Machine Learning tutoring? I need help and am having a hard time finding anyone willing to actually speak and talk through examples.



**Jason Brownlee** October 8, 2017 at 8:42 am #

REPLY ↗

I do my best on the blog 😊

Perhaps you can hire someone on upwork?



**Praveen Kumar** October 9, 2017 at 10:23 pm #

REPLY ↗

Your Start in Machine Learning

Hey Its really nice bu i have a question that for other kind of data sets is that procedure remains same..?



**Jason Brownlee** October 10, 2017 at 7:45 am #

REPLY ↗

It is a good start. Also see this more general procedure:

<https://machinelearningmastery.com/start-here/#process>



**vinaya** October 9, 2017 at 10:46 pm #

REPLY ↗

can you explain

X = array[:,0:4]

Y = array[:,4]



**Jason Brownlee** October 10, 2017 at 7:46 am #

We are selecting columns using array slicing.

X is comprised of columns 0, 1, 2 and 3.

Y is comprised of column 4.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**sukanya** October 11, 2017 at 3:50 pm #

REPLY ↗

I am not clear with the seed value and its importance.can you explain this



**Jason Brownlee** October 11, 2017 at 4:41 pm #

REPLY ↗

It initializes the random number generator so that you get the same results as I do in the tutorial.

Generally, I recommend learning more about the stochastic nature of machine learning algorithms here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Ibrahim** October 13, 2017 at 1:11 am #

REPLY ↗

Thanks Jason! its really beautiful to learn about ML using Python . Thanks for your effort to make it effortless. would you please recommend me unsupervised learning

Your Start in Machine Learning

Thank you



**Jason Brownlee** October 13, 2017 at 5:49 am #

REPLY ↗

Thanks. Sorry, I cannot help you with HMMs. I hope to cover the topic in the future.



**Johnny** October 13, 2017 at 8:02 am #

REPLY ↗

Why do you split the data into train and validation sets at the very beginning using “train\_test\_split”? I thought the K-Fold cross validation

```
cv_results = model_selection.cross_val_score(model, X, y)
```

I would assume we want to use the most data possible of the data from this step?



**Jason Brownlee** October 13, 2017 at 2:53 pm #

We do this to double check the final model

<https://machinelearningmastery.com/difference-between-train-test-split-validation/>

Learn more about fitting a final model here:

<https://machinelearningmastery.com/train-final-machine-learning-model/>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

X

REPLY ↗



**Weizhi Song** October 13, 2017 at 3:24 pm #

Hi Jason,

Thanks for your tutorial, it is really awesome! I want to use machine learning approach for biology problems. I have a question below and hope you could me give me some suggestions. Thanks in advance.

I have eight DNA sequences which are labeled as either “TSS” or “NTSS”. If I want to use your code here to predict whether a DNA sequence is TSS or not, do I need to transfer these sequences into numbers? If yes, do you have any suggestions of how to do that?

ATATATAG TSS

ACATTAG TSS

ACATATAG TSS

ACTTATAG TSS

CCGTGTGG NTSS

CCGAGTG TSS

CCGTGCGG NTSS

CCGTCTGG NTSS

Your Start in Machine Learning

Thanks,  
Weizhi



**Jason Brownlee** October 14, 2017 at 5:38 am #

REPLY ↩

Yes, you will need to encode each char or each block as an integer, and then perhaps as a binary vector.

See this post:

<https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>



**Girmay** October 13, 2017 at 10:51 pm #

This step by step tutorial is very interesting.  
But I need yellow fever data set CSV file .. to predict you  
Please any one can help me...@ [teklegimay@gmail.com](mailto:teklegimay@gmail.com)



**Jason Brownlee** October 14, 2017 at 5:46 am #

Perhaps you can use google to find a suitable dataset.



**Gaurav** March 4, 2018 at 10:08 am #

REPLY ↩

go to CHEMBL dataset



**Rash** October 15, 2017 at 9:22 am #

REPLY ↩

Thanks for your help. This is awesome.  
I have one issue : How can I rescale the axis ?  
I have an error : ValueError: x and y must be the same size.  
I have 3 features and 1 class for more than 245 000 data points.  
please help.



**Jason Brownlee** October 16, 2017 at 5:40 am #

REPLY ↩

The error suggests that you must have the

Your Start in Machine Learning



**Manish Sogi** October 18, 2017 at 4:43 pm #

REPLY ↗

Hi Jason,

You might not aware that your tutorial is arousing motivation to learn ML in engineers who are far away from this domain too. Thanks a ton !



**Jason Brownlee** October 19, 2017 at 5:33 am #

REPLY ↗

I'm glad to hear it!



**Biswajith** October 20, 2017 at 7:53 pm #

X

Hi Jason,

Nice and precise explanation. But can you please elaborate step by step approach, still missing the actual problem

Below mentioned the basic stupid question.

What result we are expecting from this problem solution?

Biswa

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** October 21, 2017 at 5:33 am #

REPLY ↗

We are trying to predict the species given measurements of iris flowers.



**shivaprasad** October 24, 2017 at 4:46 am #

REPLY ↗

sir i am not getting what the classification report is ?, what is the meaning of precision, recall, f1 score and the support , what it actually tells us, what the table is for? , and what we understand with the help of the table



**Jason Brownlee** October 24, 2017 at 5:38 am #

REPLY ↗

Perhaps this article will help:

[https://en.wikipedia.org/wiki/Precision\\_and\\_recall#](https://en.wikipedia.org/wiki/Precision_and_recall#)

Your Start in Machine Learning



**shivaprasad** October 24, 2017 at 2:46 pm #

REPLY ↗

thank you sir



**Micah** October 25, 2017 at 3:58 am #

REPLY ↗

Great article. It's been a lot of help. I've been applying this to other free datasets to practice (e.g. the titanic dataset). One thing I haven't been able to figure out is how to show which columns are the most predictive. Do you know how to do that?

Thanks,  
Micah



**Jason Brownlee** October 25, 2017 at 6:53 am #

Feature selection methods can give you a  
<http://machinelearningmastery.com/an-introduction-to-feature-selection/>



**Daniel Bermudez** October 26, 2017 at 8:48 am #

Hi Dr Jason,

I can't say thank you enough. This step by step tutorial is awesome. I'm so interested to try ML in a real project and this is a good way. I agree with you, academic is a little slow even though we can see more details.

Regards!!

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** October 26, 2017 at 4:15 pm #

REPLY ↗

I'm glad to hear it helped Daniel, well done for making it through the tutorial!



**Aditya** October 26, 2017 at 6:12 pm #

REPLY ↗

Sir,

I really appreciate your post and very thankful to you.  
This post is very important for ML beginner like me.

Your Start in Machine Learning

I really loved the content and the way you make complex things simpler.

But I have one doubt, It would be very helpful to me if you help me building my understanding.

Question :

From the section “5.3 Build Models” line number 12

for name, model in models:

Please explain what is ” name, model ” here, its purpose and how it is working, (because I hadn’t seen any FOR loop like this. I had learn python from YouTube videos and have very basic understanding)

P.S. I ran your code and its perfectly working fine.



**Jason Brownlee** October 27, 2017 at 5:18 am #

In that loop, a model is an item from the li

I recommend taking some more time to learn basic

<https://wiki.python.org/moin/ForLoop>



**Aditya** October 27, 2017 at 4:28 pm #

Thank you, you are awsome



**Raj** October 29, 2017 at 4:12 pm #

REPLY ↩

Hello Jason, I am curious about ai and ml.Tons of thanks for your hard work and commitment.I have done installation of Anaconda and checked all the libraries successfully.My ignorance of programming is compelling me to ask this ridiculous question. But i cant understand that where to upload dataset ? To be more clear i mean i dont understand even that where to write those url and given command to upload dataset ? on Jupiter notebook, or on conda prompt window ??? Please reply for kind of stupid question. Thanking you in anticipation.



**Jason Brownlee** October 30, 2017 at 5:36 am #

REPLY ↩

The function call `pandas.load_csv()` will load a CSV data file, either as a filename on your computer or a CSV file on a URL.

Does that help?

Your Start in Machine Learning



**Kevin** November 3, 2017 at 1:43 pm #

REPLY ↗

Thanks Jason! It's such a great article! However, i come across problems when applying your code here to my own dataset.

```
import sys
import scipy
import numpy
import pandas
import sklearn

from sklearn import model_selection

dataset = pandas.read_csv('D:\CMPE333\Project\Spe
array = dataset.values
X = array[:,0:12]
Y = array[:,12]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_se
random_state=seed)
```

I got the error:

```
runfile('D:/CMPE333/Project/project.py', wdir='D:/CMF
Traceback (most recent call last):
```

File "", line 1, in

```
runfile("D:/CMPE333/Project/project.py", wdir='D:/CMPE333/Project')
```

File "C:\ProgramData\Anaconda3\lib\site-packages\spyder\utils\site\sitecustomize.py", line 710, in runfile
execfile(filename, namespace)

File "C:\ProgramData\Anaconda3\lib\site-packages\spyder\utils\site\sitecustomize.py", line 101, in execfile
exec(compile(f.read(), filename, 'exec'), namespace)

File "D:/CMPE333/Project/project.py", line 33, in

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_spilt(X, Y, test_size=validation_size,
random_state=seed)
```

AttributeError: module 'sklearn.model\_selection' has no attribute 'train\_test\_spilt'

The dataset is stored as comma delimited csv file and has been loaded into a dataframe.

Can you tell me where is wrong? Thank you!!!

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** November 3, 2017 at 2:18 pm #

REPLY ↗

You might need to update your version of sklearn to 0.18 or higher.

## Your Start in Machine Learning



**Kevin** November 4, 2017 at 6:35 am #

REPLY

Thanks for replying!

My sklearn version is 0.18.1

It works well when i use your data.

Is there something wrong when i load the data?



**Anil** November 3, 2017 at 6:11 pm #

REPLY

Hello Json, Thank you. But one thing didn't clearly ~~Can you tell me in above example output what we predict? What we find? We are getting summarized whos?~~



**Jason Brownlee** November 4, 2017 at 5:27 am #

We are predicting the iris flower species given



**Meghal** November 5, 2017 at 7:10 am #

Getting error in Class Distribution. If I give sum() instead of size() it works fine. Please suggest resolution.

```
=====
# class distribution
print(dataset.groupby('class').size())
=====
```

Output

Traceback (most recent call last):

File "C:\\Python\\ML\\ImportLibs.py", line 30, in

print(dataset.groupby('class').size())

File "C:\\Users\\Meghal\\AppData\\Roaming\\Python\\Python35\\site-packages\\pandas\\core\\base.py", line 59, in \_\_str\_\_

return self.\_\_unicode\_\_()

File "C:\\Users\\Meghal\\AppData\\Roaming\\Python\\Python35\\site-packages\\pandas\\core\\series.py", line 1060, in \_\_unicode\_\_

width, height = get\_terminal\_size()

File "C:\\Users\\Meghal\\AppData\\Roaming\\Python\\Python35\\site-packages\\pandas\\io\\formats\\terminal.py", line 33, in get\_terminal\_size

return shutil.get\_terminal\_size()

File "C:\\Users\\Meghal\\AppData\\Local\\Programs\\Python\\Python35-32\\lib\\shutil.py", line 1071, in get\_terminal\_size

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```
size = os.get_terminal_size(sys.__stdout__.fileno())
AttributeError: 'NoneType' object has no attribute 'fileno'
=====
```



**Jason Brownlee** November 6, 2017 at 4:44 am #

REPLY ↩

Perhaps double check you have the latest version of the libraries installed?

Confirm the data was loaded correctly?



**Jeff Guo** November 5, 2017 at 9:07 am #

Not sure why, but for me, SVM is giving me a score, but it ultimately has the same support score as



**Jason Brownlee** November 6, 2017 at 4:47 am #

Might be the stochastic nature of ML algos  
<https://machinelearningmastery.com/randomness-in-machine-learning/>



**xylo** November 6, 2017 at 2:21 am #

REPLY ↩

1.can someone explain compare algorithm graph? 2.why knn is best algorithm 3. why & when use which algorithm?? thnx in advance



**Jason Brownlee** November 6, 2017 at 4:53 am #

REPLY ↩

Generally, we cannot know what algorithm will be “best” for a given problem. Our job is to use careful experiment to discover what works best for a given prediction problem.

See this post:

<http://machinelearningmastery.com/a-data-driven-approach-to-machine-learning/>



**Georgios Koumakis** November 7, 2017 at 4:48 am #

REPLY ↩

Jason, you are the best!!  
 Thanks for putting together all that material in a meani

Your Start in Machine Learning

environment.

There are not enough words to say how thankful I am.



**Jason Brownlee** November 7, 2017 at 9:53 am #

REPLY ↗

Thanks, I'm glad it helped Georgios.



**Austin** November 8, 2017 at 12:08 pm #

REPLY ↗

Hey Jason, fantastic tutorial. I have one question though. Is it possible to inputting a flower and the computer identifying it? That would be great!



**Jason Brownlee** November 9, 2017 at 9:52 am #

Yes, you could input the measurements of a flower and the computer would identify it.



**Abhishek Jain** November 9, 2017 at 1:36 am #

REPLY ↗

Hi Jason, Thanks a lot for the excellent step by step material to give a quick run-through of the methodology.

I am a tenured analytics practitioner and somehow found some time off to learn Python and was looking through the IRIS project itself. I had hypothesised that by adding more ratio variables to the dataset, we should get a better result on the prediction. Your excellent article gives me a ready code to test my hypothesis. I will share my results once I have them. 😊

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 9, 2017 at 10:02 am #

REPLY ↗

Please do!



**Abhishek Jain** November 12, 2017 at 3:26 am #

REPLY ↗

Here are the k-Fold results: I used additional variables simply as all ratios of the original length variables respectively with no separate effort on dimensionality reduction.

LR: 0.950000 (0.040825)

LDA: 0.991667 (0.025000)

Your Start in Machine Learning

KNN: 0.958333 (0.055902)

CART: 0.950000 (0.066667)

NB: 0.966667 (0.055277)

SVM: 0.966667 (0.040825)

Drill down to the independent validation results for each technique:

Results for LR : 1.0

Results for LDA : 0.933333333333

Results for KNN : 1.0

Results for CART : 0.9

Results for NB : 0.9666666666667

Results for SVM : 1.0

Although validation results are better across the K-fold method because other models may require reduction effort.

I would be glad to hear more from you on this.



**Jason Brownlee** November 12, 2017 at 11:27 am #

Great work, thanks for sharing!



**narendra** November 11, 2017 at 11:27 am #

REPLY ↗

Hi Jason,

Thank you for the great tutorial. Once we run test and validate the model. How can we deploy the model. Also, how can we make the model predict on new data-set and still continuously learn from the new data.

Thank you,



**Jason Brownlee** November 12, 2017 at 9:00 am #

REPLY ↗

Great question.

This post has ideas on developing a final model:

<https://machinelearningmastery.com/train-final-machine-learning-model/>

This post has ideas on deploying a model:

<http://machinelearningmastery.com/deploy-machine-learning-model-to-production/>

**chaitanya** November 12, 2017 at 1:33 am #

Your Start in Machine Learning



Nice article to start with.  
Although I really do not understand what each of model does?  
So what should be the next step?



**Jason Brownlee** November 12, 2017 at 9:05 am #

REPLY ↩

You could learn more about how each model works:

<https://machinelearningmastery.com/start-here/#algorithms>



**Anh** November 13, 2017 at 9:15 pm #

Thanks a lot for your tutorial Jason. How should I handle dataset is text, not number?



**Jason Brownlee** November 14, 2017 at 10:11 pm #

Working with text is called natural language processing.

<https://machinelearningmastery.com/start-here/#nlp>

## Your Start in Machine Learning

X

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**sanjay** November 17, 2017 at 2:25 am #

REPLY ↩

“AxesSubplot” object has no attribute ‘set\_xticklables’



**Jason Brownlee** November 17, 2017 at 9:28 am #

REPLY ↩

Sorry to hear that, please confirm that you have setup your environment correctly:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Prateek Gupta** November 17, 2017 at 11:20 pm #

REPLY ↩

Thanks Jason for this well explained post!

I am an aspiring data scientist and currently working on Walmart’s sales forecasting dataset from kaggle. If it is possible can you please also share a post about predicting the sales for this dataset? It will be very helpful because I am not finding such a *step by step tutorial in Python*.

## Your Start in Machine Learning



**Jason Brownlee** November 18, 2017 at 10:18 am #

REPLY ↗

Thanks for the suggestion.

Perhaps this process will help you work through the problem systematically:

<https://machinelearningmastery.com/start-here/#process>



**ali** November 20, 2017 at 3:58 pm #

REPLY ↗

Thanks for the amazing guide

can i know how to get the sensitivity and specificity and  
you had a good Example Confusion Matrix in R with code  
but in the same page i could get the confusion for python  
sensitivity and specificity and recall

thank again



**Jason Brownlee** November 22, 2017 at 10:37 am #

REPLY ↗

Perhaps this will help:

<http://scikit-learn.org/stable/modules/classes.html>

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Nicola** November 22, 2017 at 6:09 am #

REPLY ↗

Thankyou very much for the great tutorial.

I analyzed every step but one thing it is not clear for me, and maybe it is the most important part of the tutorial 😊

At the end of all our steps I would expect a function or something else to answer Python questions like these:

1. I have a flower with sepal-length=5, sepal width=3.5, petal-length=1.3 and petal-width=0.3, which class is it?
2. I have an Iris-setosa with sepal-length=5, sepal width=3.5, petal-length=1.3. What could be the petal-width?

Isn't this one of the the main objectives of the ML?



**Nicola** November 22, 2017 at 6:49 am #

REPLY ↗

OK, I answer by myself, for question one '

## Your Start in Machine Learning

```
print(knn.predict([[5.0, 3.5, 1.3, 0.3]]))
```

to get “[‘Iris-setosa’]”

For question 2 I think that I need to rebuilt the whole model.



**Jason Brownlee** November 22, 2017 at 11:16 am #

REPLY ↗

Well done!



**Jason Brownlee** November 22, 2017 at 11:15 am #

Yes, you can train a final model on all data:

Here's more about that:

<https://machinelearningmastery.com/train-final-machine-learning-model/>

Here's how to save a model in Python:

<https://machinelearningmastery.com/save-load-machine-learning-models/>

You can predict on new data using:

```
1 X = ...
2 yhat = model.predict(X)
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Tash** November 22, 2017 at 11:26 am #

REPLY ↗

This is a brilliant tutorial, thank you. I have a few questions – you split the data in to training and validation, but in this case would it not be classed as training and test?

Also, do you have any posts on tuning hyperparameters such as the learning rate in Logistic Regression? It was my understanding that a validation set would be used for something like this, while holding back the test set until the models been fine-tuned...but now I'm not sure if I'm confused!

Thanks so much.



**Jason Brownlee** November 23, 2017 at 10:23 am #

REPLY ↗

Yes, it would be training and test, here's more on the topic:

<https://machinelearningmastery.com/difference-test-validation-datasets/>

## Your Start in Machine Learning



**Túlio Campos** November 24, 2017 at 11:56 am #

REPLY ↗

Why on

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)

you use only the training part instead of the full set since it's a cross-validation?
```



**Jason Brownlee** November 24, 2017 at 3:05 pm #

REPLY ↗

In this case I wanted to hold back a test set to evaluate the final chosen model.



**Túlio Campos** November 24, 2017 at 1:09 pm #

Also, in case I want to use X, Y by themselves don't have totally random results because my classes are not balanced. Thank you.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 24, 2017 at 3:08 pm #

Sorry, I don't follow. Do you have an example of what you mean?



**Túlio Campos** December 5, 2017 at 3:29 am #

REPLY ↗

If you directly use

```
cv_results = model_selection.cross_val_score(model, X, Y, cv=kfold, scoring=scoring)
```

With kfold = 3 for example. You will get 3 different groups, each with one type of iris flower because sklearn doesn't shuffle it by its own and the dataset is arranged by flower-type.

You would have to use something like ShuffleSplit

[http://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.ShuffleSplit.html](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ShuffleSplit.html)

Before doing so.



**Jason Brownlee** December 5, 2017 at 5:46 am #

REPLY ↗

Did you try this change, does it in-

Your Start in Machine Learning



**Túlio Campos** December 8, 2017 at 7:04 am #

Yes it does. In 3 fold I was getting under 70% accuracy. Shuffling makes it more evenly distributed (not 3 totally different groups). And I could get 90%\_ acc

Also, I figured that I could simply use the parameter “Shuffle=True” in .KFold



**Jason Brownlee** December 8, 2017 at 2:26 pm #

Nice!



**Goldi** November 25, 2017 at 12:30 pm #

Hi Jason,

Excellent way of explaining the basics of machine learning.

I assume that in almost all machine learning program if we apply algorithms we can understand much better a classification is the key in supervised and clustering is a good model.

Thanks a Lot.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 26, 2017 at 7:30 am #

REPLY ↩

I'm glad you found it useful.



**Meenakshi** November 26, 2017 at 9:42 am #

REPLY ↩

Thanks for the tutorial, it is very helpful!



**Jason Brownlee** November 27, 2017 at 5:42 am #

REPLY ↩

You're welcome, I'm glad to hear that.

## Your Start in Machine Learning



**BENNAMA** November 29, 2017 at 9:10 am #

REPLY ↗

I am working on windows 8.1

I am trying to apply the example by using python 2.7.14 anaconda

when arrived on section 4.1:

# box and whisker plots

```
dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
```

```
plt.show()
```

My cmd console shows an error “nameerror : plt name not defined”

To solve this problem i have added the line:

```
import matplotlib.pyplot as plt
```

it works

Thank's



**Jason Brownlee** November 30, 2017 at 8:04 am #

×

Glad to hear you fixed your issue.



**Deepak Gautam** December 2, 2017 at 5:23 am #

REPLY ↗

Hey! this is wonderful tutorial.

I goes through all the steps and it's great.

One thing I want to know that which is best model:-

- \* Linear Discriminant Analysis (LDA)

with 0.96

- \* K-Nearest Neighbors (KNN).

with 0.9



**Jason Brownlee** December 2, 2017 at 9:06 am #

REPLY ↗

It is up to the practitioner to choose the right model based on the complexity of the model and on mean and standard deviation of model skill results.



**John Wolter** December 4, 2017 at 10:04 am #

REPLY ↗

Here's a really nit-picky observation: You have

Your Start in Machine Learning

Nit-picking aside, this is an excellent starter for ML in Python. I am currently taking the Coursera / Stanford University / Dr. Andrew Ng Machine Learning course and being able to see some of these algorithms that we have been learning about in action is very satisfying. Thank you!



**Jason Brownlee** December 4, 2017 at 4:57 pm #

REPLY ↗

Thanks John, fixed section numbering.



**Ezra Axel** December 5, 2017 at 4:50 pm #

REPLY ↗

How do you respond to all the comments?



**Jason Brownlee** December 6, 2017 at 8:59 am #

X

It takes time every single day!

But I created this blog to hang out with people just

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**BukuBapi** December 8, 2017 at 3:17 pm #

You Mentioned that

[ We will use 10-fold cross validation to estimate accuracy.

This will split our dataset into 10 parts, train on 9 and test on 1 and repeat for all combinations of train-test splits. ]

In your code, I understand that you split it in 10 parts, but where is the 9:1 ratio mentioned. Unable to get that



**Jason Brownlee** December 9, 2017 at 5:36 am #

REPLY ↗

This is how cross-validation works, learn more here:

[https://en.wikipedia.org/wiki/Cross-validation\\_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics))



**Nil** December 11, 2017 at 12:39 am #

REPLY ↗

Hi Dr. Jason,

Your Start in Machine Learning

When evaluating we found that KNN presented the best accuracy, KNN: 0.983333 (0.033333). But when the validation set was used in KNN to have the idea of the accuracy, I see that the accuracy now is 0.9 so it decreased, while I was expecting the same accuracy. Can I consider this as over fitting? I can consider that KNN over fitted the train data? Is this difference of accuracy in the same model while training and validating acceptable?



**Jason Brownlee** December 11, 2017 at 5:26 am #

REPLY ↗

No, this is the stochastic variance of the algorithm. Learn more about this here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Nil** December 11, 2017 at 9:37 pm #

Thank you.

I will learn more in the recommended site.

Best Regards.



**bugtime** December 11, 2017 at 5:21 am #

Jason,

AWESOME ARTICLE, THANK YOU!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 11, 2017 at 5:34 am #

REPLY ↗

I'm glad it helped!



**Gulshan Bhatia** December 14, 2017 at 8:02 pm #

REPLY ↗

```
File "ml.py", line 73, in
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
File "/usr/local/lib/python2.7/dist-packages/sklearn/model_selection/_validation.py", line 342, in
cross_val_score
pre_dispatch=pre_dispatch)
File "/usr/local/lib/python2.7/dist-packages/sklearn/model_selection/_validation.py", line 206, in
cross_validate
for train, test in cv.split(X, y, groups))
```

Your Start in Machine Learning

```

File "/usr/local/lib/python2.7/dist-packages/scikit-learn/externals/joblib/parallel.py", line 779, in __call__
while self.dispatch_one_batch(iterator):
File "/usr/local/lib/python2.7/dist-packages/scikit-learn/externals/joblib/parallel.py", line 625, in
dispatch_one_batch
self._dispatch(tasks)
File "/usr/local/lib/python2.7/dist-packages/scikit-learn/externals/joblib/parallel.py", line 588, in _dispatch
job = self._backend.apply_async(batch, callback=cb)
File "/usr/local/lib/python2.7/dist-packages/scikit-learn/externals/joblib/_parallel_backends.py", line 111, in
apply_async
result = ImmediateResult(func)
File "/usr/local/lib/python2.7/dist-packages/scikit-learn/externals/joblib/_parallel_backends.py", line 332, in
__init__
self.results = batch()
File "/usr/local/lib/python2.7/dist-packages/scikit-learn/ex
return [func(*args, **kwargs) for func, args, kwargs in s
File "/usr/local/lib/python2.7/dist-packages/scikit-learn/m
_fit_and_score
estimator.fit(X_train, y_train, **fit_params)
File "/usr/local/lib/python2.7/dist-packages/scikit-learn/lin
check_classification_targets(y)
File "/usr/local/lib/python2.7/dist-packages/scikit-learn/ut
check_classification_targets
raise ValueError("Unknown label type: %r" % y_type)
ValueError: Unknown label type: 'unknown'

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)



**Gulshan Bhatia** December 14, 2017 at 8:08 pm #

[REPLY](#) ↗

urgent help required



**Jason Brownlee** December 15, 2017 at 5:31 am #

[REPLY](#) ↗

Confirm that you have copied all of the code and that your scipy/numpy/scikit-learn are all up to date.



**Justin** December 17, 2017 at 6:39 am #

[REPLY](#) ↗

Not sure if it's been mentioned, but this line: "pandas.read\_csv(url, names=names)" did not work for me until I replaced https with http after looking up docs for read\_csv

## Your Start in Machine Learning



**Jason Brownlee** December 17, 2017 at 8:55 am #

REPLY ↗

Thanks, Justin.

REPLY ↗



**Nawaz** December 19, 2017 at 7:59 pm #

hey Jason Brownlee,

Thanks for the tutorial

I got an error after I build five models

"urllib.error.URLError: "

Thanks



**Jason Brownlee** December 20, 2017 at 5:43 am #

Sorry to hear that. Perhaps ensure that yo



**Zeinab** December 20, 2017 at 4:42 pm #

Hello, Jason,

I am a beginner in python.

Unfortunately, when I load my dataset (it contains 4 features & 1 class “each with string datatype”), and then run the command

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring),
```

I found the following error:

ValueError: could not convert string to float:

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 21, 2017 at 5:23 am #

REPLY ↗

Perhaps confirm that your data is all numerical?

Perhaps try converting it to float before using sklearn?



**Steve H** December 22, 2017 at 3:53 am #

REPLY ↗

## Your Start in Machine Learning

- 1) I realize that this is just an example, but in general, is this the process that you personally use when you are building production models?
- 2) What would the next steps be in terms of taking this to the next level? Would you choose the model that you think performs best, and then attempt to tune it to get even better results?



**Jason Brownlee** December 22, 2017 at 5:36 am #

REPLY ↗

Mostly, this is the process in more detail:

<https://machinelearningmastery.com/start-here/#process>



**raymond doctor** December 23, 2017 at 11:51 pm #

Hello,

The tutorial worked like a charm and I had no problem with it. However, I am a linguist and my needs are different. The number of linguists is different.

As a linguist [and there are many like me throughout the world] we often need to generate rules for either a source language or between a source and a target language. At present I use an automata approach which states something like  $a \rightarrow b$  in environment  $x$ .

This however implies that rules have to be manually written by hand. If the number of rules grows, as does the amount of data this becomes a huge problem.

I have searched and not located a simple tool which does this job using RNN. The existing tools are extremely complex and adapting them to suit a simple requirement of the type outlined above is practically impossible.

What I need is:

- a. A tool which installs itself deploying Python and all accompanying libraries.
- b. Asks for input of parallel data
- c. generates out rules in the back ground
- d. Provides an interface for testing by entering new data and seeing if the output works.
- e. It should work on Windows. A large number of such prediction tools are Linux based depriving both Windows and Mac users the facility to deploy them. My Windows10 is hopefully Linux Compatible but I have never tested the shell.
- f. Above all ease of use. A large number if not all Linguists are not very familiar with coding.

Do you know of any such tool ? And can such a tool be made available in Open Source. You would have the blessings of a large number of linguists who at present have to do the tedious task of generating out rules by hand and once again generating out new rules every time a sample not considered pops up.

I know the Wishlist above is quite voluminous.Hoping to get some good news

Best regards and thanks,

R. Doctor

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 24, 2017 at 4:54 am #

REPLY ↗

Sounds like an interesting problem. I'm not aware of a tool.

Do you have some more information on this problem, e.g. some links to papers or blog posts?



**Prakash** December 26, 2017 at 1:45 am #

REPLY ↗

Thanks for awesome tutorial....

I am facing issue in 4.1 section, while installing

```
dataset.plot(kind='box', subplots=True, layout=(2,2), s
```

I am getting this error.

Traceback (most recent call last):

```
File "", line 1, in 
File "/usr/local/lib/python2.7/dist-packages/pandas/pl
sort_columns=sort_columns, **kwds)
File "/usr/local/lib/python2.7/dist-packages/pandas/pl
**kwds)
File "/usr/local/lib/python2.7/dist-packages/pandas/pl
plot_obj.generate()
File "/usr/local/lib/python2.7/dist-packages/pandas/pl
self._setup_subplots()
File "/usr/local/lib/python2.7/dist-packages/pandas/plotting/_core.py", line 299, in _setup_subplots
layout_type=self._layout_type)
File "/usr/local/lib/python2.7/dist-packages/pandas/plotting/_tools.py", line 197, in _subplots
fig = plt.figure(**fig_kw)
File "/usr/local/lib/python2.7/dist-packages/matplotlib/pyplot.py", line 539, in figure
**kwargs)
File "/usr/local/lib/python2.7/dist-packages/matplotlib/backend_bases.py", line 171, in
new_figure_manager
return cls.new_figure_manager_given_figure(num, fig)
File "/usr/local/lib/python2.7/dist-packages/matplotlib/backends/backend_tkagg.py", line 1049, in
new_figure_manager_given_figure
window = Tk.Tk(className="matplotlib")
File "/usr/lib/python2.7/lib-tk/Tkinter.py", line 1818, in __init__
self.tk = _tkinter.create(screenName, baseName, className, interactive, wantobjects, useTk, sync, use)
_tkinter.TclError: no display name and no $DISPLAY environment variable
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** December 26, 2017 at 5:18 am #

REPLY ↗

Sorry to hear that, looks like your Python

Your Start in Machine Learning

Perhaps this tutorial will sort things out:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Rizwan Mian** December 26, 2017 at 11:40 am #

REPLY ↗

Jason, I am learning so much from your work (thanks 😊)

– my model scores are different to ones reported in the post (Section 5.4)? what could be the possible reasons?

('algorithm', 'accuracy', 'mean', 'std')

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.983333 (0.033333)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

– What do the bars represent in Algorithm Comparison? accuracy and standard deviation are 0.98 and 0.04. The whisker at about 0.92. Take knn for another example, 1 and 0.03. However, the bar finishes at 1 and the whisker. Is y-axis accuracy?

– how to read the confusion matrix without labels? My guess is row and column (missing) labels represent actual and predicted classes, respectively. However, I am unsure about the order of classes. is there a way to switch on the labels?

I collected and annotated the code in a python script (iris.py), and placed it on the github:

<https://github.com/dr-riz/iris>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 26, 2017 at 3:01 pm #

REPLY ↗

The differences may be related to the stochastic nature of the algorithms:

<https://machinelearningmastery.com/randomness-in-machine-learning/>

You can learn more about box and whisker plots here:

[https://en.wikipedia.org/wiki/Box\\_plot](https://en.wikipedia.org/wiki/Box_plot)

You can learn more about the confusion matrix here:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>

Great annotations, please reference the URL of this blog post and the name of the blog as source.

## Your Start in Machine Learning



**Rizwan Mian** December 27, 2017 at 7:16 am #

REPLY ↗

Thanks for your reply and reminder. Credits and Source, URLs are now noted in README.



Re LDA example: the stated accuracy and standard deviation are 0.98 and 0.04. Yes, the box plot renders metrics such as minimum, first quartile, median, third quartile, and maximum but \*not\* necessarily mean. Hence, we don't see mean and std in the box plot in Section 5.4.

I reproduce this with a simple example.

```
lda_model = LinearDiscriminantAnalysis()
lda_results = model_selection.cross_val_score(lda_model, X_train, Y_train, cv=10,
scoring='accuracy')

np.size(lda_results) => 10 elements, 1 for each
investigation.

lda_results.max() # => 1
numpy.median(lda_results) # > 1
numpy.percentile(lda_results, 75) # => 1 — 3rd quartile
numpy.percentile(lda_results, 25) # => 0.9423
lda_results.min() # => 0.9091 — this is value we expect
lda_results.mean() # => 0.9749 — DONT expect
lda_results.std() # => 0.03849 — DONT expect

fig = plt.figure()
ax = fig.add_subplot(111)
plt.boxplot(lda_results)
ax.set_xticklabels(['LDA'])
plt.show()
```

As expected, we don't see mean and std in the box plot.

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 28, 2017 at 5:18 am #

REPLY ↗

Thanks.

Cross validation is creating 10 models and evaluating each on 10 different and unique samples of your dataset.



**Daniel** December 28, 2017 at 9:12 am #

REPLY ↗

Nice. Took me a little longer than 10 mins, but works as advertised. (I did everything under python3, no big difference I think.)

Your Start in Machine Learning

What would be really cool here would be a “what is going on here” section at the end. But it’s real nice to have something that actually runs, and be able to poke about with it a bit.

Thanks Jason. Good stuff.



**Jason Brownlee** December 28, 2017 at 2:10 pm #

REPLY ↩

Well done. Nice suggestion, thanks.



**MG5** December 29, 2017 at 3:26 am #

Hello Jason, I wanted to ask you if the seed dataset arrived at 97% accuracy, do you think it can still improve? <http://archive.ics.uci.edu/ml/datasets/seeds>.



**Jason Brownlee** December 29, 2017 at 5:25 am #

Perhaps, though that is an impressive result.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

X

REPLY ↩



**Sammy Lee** December 29, 2017 at 12:38 pm #

So how would we obtain individual new predictions using our own input data after going through this exercise?



**Jason Brownlee** December 29, 2017 at 2:37 pm #

REPLY ↩

Train a final model:

<https://machinelearningmastery.com/train-final-machine-learning-model/>

Then call:

```
1 X = ...
2 yhat = model.predict(X)
```

REPLY ↩



**Gage Russell** December 29, 2017 at 3:35 pm #

I am getting the syntax error pasted below at line 2. I have made sure that I am copying and pasting it directly.

## Your Start in Machine Learning

why this is occurring would be great! Thanks in advance!

for name, model in models:

File "", line 1

for name, model in models:

^

SyntaxError: unexpected EOF while parsing



**Jason Brownlee** December 30, 2017 at 5:17 am #

REPLY ↗

Ensure that you copy all of the code with the same formatting. White space has meaning in Python.



**Joe** January 1, 2018 at 10:00 am #

I put the requirements for this tutorial in a Dockerfile:  
<https://github.com/UnitasBrooks/docker-machine-learning>



**Jason Brownlee** January 2, 2018 at 5:31 am #

Thanks Joe.

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Rizwan Mian** January 1, 2018 at 2:21 pm #

REPLY ↗

The algorithms are instantiated with their default parameters. Is this a standard practise for spot checking algorithms?



**Jason Brownlee** January 2, 2018 at 5:33 am #

REPLY ↗

You can specify some standard or common configurations as part of the checking.



**abidh** January 1, 2018 at 6:32 pm #

REPLY ↗

I tried the above tutorial. But i got accuracies differ from the given above for the same dataset. why? also the boxplot for the same is changing each time

## Your Start in Machine Learning



**Jason Brownlee** January 2, 2018 at 5:34 am #

REPLY ↗

Yes, this is a feature not a bug, learn more here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Ben Hart** January 6, 2018 at 5:01 pm #

REPLY ↗

Hi Jason,

I think I downloaded the same dataset as you have here but the sepal-length data seems to have changed a bit. Not to worry though as you can easily follow the predictions using SVC ()

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.966667 (0.040825)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

0.933333333333

[[ 7 0 0]

[ 0 10 2]

[ 0 0 11]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 1.00 0.83 0.91 12

Iris-virginica 0.85 1.00 0.92 11

avg / total 0.94 0.93 0.93 30

It does give a better result which is nice.

Also I was wondering if you explain the confusion matrix anywhere on your website, I find it somewhat confusing 😊

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** January 7, 2018 at 5:04 am #

REPLY ↗

Yes, here is more on the confusion matrix:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>



**dj** January 6, 2018 at 7:01 pm #

REPLY ↗

What we predicted in the output with help of i

Your Start in Machine Learning



**Jason Brownlee** January 7, 2018 at 5:04 am #

REPLY ↗

The model predicts the species based on flower measurements.



**Praveen Chakravarthy** January 7, 2018 at 10:53 pm #

REPLY ↗

Hi Jason, watched your videos and you are awesome, can you tell me how to train our own image data database and split into train and test sets, labels...thank you for listening to me...



**Jason Brownlee** January 8, 2018 at 5:43 am #

I don't have any videos.



**prageeth** January 8, 2018 at 10:57 pm #

Thank you so much..

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** January 9, 2018 at 5:30 am #

REPLY ↗

You're welcome.



**Jackson** January 10, 2018 at 3:34 am #

REPLY ↗

Hi Jason,

Thanks for this great tutorial. It really helps.

Everything works fine except:

- In Section 4.1 – Histogram – the distribution in Sepal Length is quite different from yours. May be that's due to the random nature of Machine Learning ?
- In section 5.4 – Box and whisker plot: the plots for LR , LDA and CART are similar but for KNN, SVM; I could only get a "+" sign at around 0.92 (no box and no whisker shown). For NB, I could only get 1 "+" sign at 0.92 and 1 "+" sign at around "0.83".

Grateful if you could advise. Thanks.

Your Start in Machine Learning

I am using :

window 10, python 3.5.2 – Anaconda custom (64 bit)

scipy: 1.0.0

numpy: 1.13.3

matplotlib: 1.5.3

pandas: 0.18.1

statsmodels: 0.6.1

sklearn: 0.19.1

theano: 0.9.0.dev-unknown-git

Using TensorFlow backend.

keras: 2.1.2



**Jason Brownlee** January 10, 2018 at 5:30 am #

Well done!



**Jackson** January 11, 2018 at 2:45 am #

Thanks, but something goes “wrong”

In section 5.4 – Box and whisker plot: the plots your web page

but for KNN, SVM; I could only get a “+” sign at around 0.92 (no box and no whisker shown). For NB, I could only get 1 “+” sign at 0.92 and 1 “+” sign at around “0.83”.



**Jason Brownlee** January 11, 2018 at 5:53 am #

REPLY ↗

Interesting.



**NAVALUTI SHIVAKUMAR** January 13, 2018 at 6:02 am #

REPLY ↗

thank you so much for valuable blog.

I'm new to Python and ML. your blog is helped me a lot in learning.

in this I've not understand how data will train ( X\_train , Y\_train and )

thanks

Your Start in Machine Learning



**Jason Brownlee** January 13, 2018 at 7:49 am #

REPLY ↗

Thanks.



**Chandi** January 15, 2018 at 9:29 pm #

REPLY ↗

Hello Jason,

This is amazing tutorial and it's really helps me to understand well!!!.. Please I want to know, do you have this type of tutorials for "pyspark" ? Can you suggest me any links, books, pdf or any tutorials? Thank you



**Jason Brownlee** January 16, 2018 at 7:33 am #

Not at this stage, sorry.



**Nilotpal** January 16, 2018 at 2:19 pm #

It has a dependency with pillow library, but it

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** January 17, 2018 at 9:55 am #

REPLY ↗

Does it?

Perhaps this is contingent on how you setup your environment?



**EDUARDO DURAN** January 23, 2018 at 4:00 pm #

REPLY ↗

Dear ,

Maybe you have the .py file of the tutorial? could you send it to me please



**Jason Brownlee** January 24, 2018 at 9:51 am #

REPLY ↗

It is a part of this book:

<https://machinelearningmastery.com/machine-learning-with-python/>

## Your Start in Machine Learning



**Jude** January 26, 2018 at 12:08 am #

REPLY ↗

Thank you, Jason Brownlee. I did run the entire scripts. It worked simply well on my MacBookPro.  
You are the best!



**Jason Brownlee** January 26, 2018 at 5:43 am #

REPLY ↗

I'm glad to hear it, well done Jude!



**Sunil** January 27, 2018 at 4:55 am #

Hi Jason,

Very nice tutorial.

I am getting error while running models. It is complainin

Following is the stacktrace

Traceback (most recent call last):

File "C:\eclipse\_workspace\MachineLearning\Iris\_Project\trainData()

File "C:\eclipse\_workspace\MachineLearning\Iris\_Project\run\_algorithms(X\_train, Y\_train, seed, scoring)

File "C:\eclipse\_workspace\MachineLearning\Iris\_Project\src\IrisLoadData.py", line 79, in run\_algorithms  
cv\_results = model\_selection.cross\_val\_score(model, X\_train, Y\_train, cv=kfold, scoring=scoring)

File "C:\Python27\lib\site-packages\sklearn\model\_selection\\_validation.py", line 342, in cross\_val\_score  
pre\_dispatch=pre\_dispatch)

File "C:\Python27\lib\site-packages\sklearn\model\_selection\\_validation.py", line 206, in cross\_validate  
for train, test in cv.split(X, y, groups))

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\parallel.py", line 779, in \_\_call\_\_  
while self.dispatch\_one\_batch(iterator):

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\parallel.py", line 625, in dispatch\_one\_batch  
self.\_dispatch(tasks)

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\parallel.py", line 588, in \_dispatch  
job = self.\_backend.apply\_async(batch, callback=cb)

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\\_parallel\_backends.py", line 111, in  
apply\_async

result = ImmediateResult(func)

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\\_parallel\_backends.py", line 332, in \_\_init\_\_  
self.results = batch()

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\parallel.py", line 131, in \_\_call\_\_  
return [func(\*args, \*\*kwargs) for func, args, kwargs in self.items]

File "C:\Python27\lib\site-packages\sklearn\model\_selection\\_validation.py", line 159, in fit\_and\_score  
estimator.fit(X\_train, y\_train, \*\*fit\_params)

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

File "C:\Python27\lib\site-packages\sklearn\linear\_model\logistic.py", line 1216, in fit  
order="C")

File "C:\Python27\lib\site-packages\sklearn\utils\validation.py", line 573, in check\_X\_y  
ensure\_min\_features, warn\_on\_dtype, estimator)

File "C:\Python27\lib\site-packages\sklearn\utils\validation.py", line 441, in check\_array  
"if it contains a single sample.".format(array))

ValueError: Expected 2D array, got 1D array instead:

```
array=[2.8 3. 3. 3.3 3.1 2.2 2.7 3.2 3.1 3.4 3.8 3. 3.3 2.4 2. 2.8 3.4 2.9  
3.5 3.1 2.9 2.6 2.7 4.4 3.2 3.4 4. 2.6 2.5 3. 3. 3.2 2.9 3. 3. 3.8  
3.2 3.2 3. 2.6 2.4 3.1 4.2 3. 3.2 3.5 3.8 2.8 2.9 3.7 2.5 3.4 2.8 3.  
3.2 3.7 3.3 2.8 2.5 2.8 2.3 3.4 3.9 2.8 3. 3.7 2.7 3.2 3.4 2.8 2.3 3.1  
3.1 3.6 3. 2.9 2.8 2.8 3.1 2.9 3. 2.7 3. 2.3 2.8 3.4 3.3 2.5 3.8 3.8  
3.4 2.8 3. 3.5 3. 3. 2.2 3.4 3.2 3.2 2.5 2.5 3.3 2.7 2.6 2.
```

Reshape your data either using array.reshape(-1, 1) if you  
it contains a single sample.

Could you please take a look and help me out?



**Jason Brownlee** January 27, 2018 at 5:59 am #

Perhaps double check your loaded data r



**Sunil** January 28, 2018 at 5:16 am #

Hi Jason,

Yeah I made some mistake while loading the data. I corrected it.

I have some questions.

What is confusion matrix and support in final result? Can you please tell about these things? For logistic regression/ classification algorithms, we need to calculate weights and we need to provide learning rate for cost function and we need to minimize it right? Is it taken care in python libraries?

Thank you,

Sunil



**Jason Brownlee** January 28, 2018 at 8:27 am #

REPLY ↤

See this post on the confusion matrix:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>

Your Start in Machine Learning

**xd** January 25, 2019 at 10:28 am #

REPLY ↗

I also got the same error about reshaping the data. I double checked the loading of my data and it's loading fine. Not sure what the problem is. Any help will be appreciated. Great tutorial Jason!

**Jason Brownlee** January 25, 2019 at 12:03 pm #

I believe it's a warning that you can safely ignore.

**Pythor** January 27, 2018 at 2:16 pm #

This was fun for my first Machine learning project in Python

**Jason Brownlee** January 28, 2018 at 8:21 am #

Well done!

**Gopal Venugopal** January 28, 2018 at 9:58 am #

REPLY ↗

Hello,

I have a technical problem please! I have downloaded Anaconda 3.6 for windows in my desktop. However, I am unable to see Terminal window or Anaconda Prompt although I have the anaconda navigator installed. Is there something wrong?

Thank you very much for your advise,

Gopal.

**Jason Brownlee** January 29, 2018 at 8:14 am #

REPLY ↗

Perhaps this post will help:

<https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/>

**Jenny** January 29, 2018 at 5:33 pm #

<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>

Your Start in Machine Learning



I just want to say thank you this is very helpful!



**Jason Brownlee** January 30, 2018 at 9:47 am #

REPLY ↗

You're welcome, glad to hear that.



**kotrappa SIRBI** January 30, 2018 at 12:39 pm #

REPLY ↗

Very nice Machine Learning getting started like [HelloWorld](#). Thanks



**Jason Brownlee** January 31, 2018 at 9:36 am

I'm glad it helped.



**Blessy** January 30, 2018 at 3:57 pm #

i get this error after the line

```
" cv_results = model_selection.cross_val_score(model
```

Traceback (most recent call last):

File "", line 1, in

```
File "C:\Users\HP\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\model_selection\_validation.py", line 335, in cross_val_score
scorer = check_scoring(estimator, scoring=scoring)
```

```
File "C:\Users\HP\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\metrics\scorer.py", line 274, in check_scoring
```

"“fit” method, %r was passed" % estimator)

```
TypeError: estimator should be an estimator implementing ‘fit’ method, [(‘LR’, LogisticRegression(C=1.0,
class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
verbose=0, warm_start=False)), ('LDA', LinearDiscriminantAnalysis(n_components=None, priors=None,
shrinkage=None,
```

```
solver='svd', store_covariance=False, tol=0.0001)), ('KNN', KNeighborsClassifier(algorithm='auto',
leaf_size=30, metric='minkowski',
```

```
metric_params=None, n_jobs=1, n_neighbors=5, p=2,
weights='uniform')), ('CART', DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
```

```
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

Your Start in Machine Learning

```
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')), ('NB', GaussianNB(priors=None)), ('SVM', SVC(C=1.0, cache_size=200,
class_weight=None, coef0=0.0,
decision_function_shape='ovr', degree=3, gamma='auto', kernel='rbf',
max_iter=-1, probability=False, random_state=None, shrinking=True,
tol=0.001, verbose=False))] was passed
```



**Jason Brownlee** January 31, 2018 at 9:37 am #

REPLY ↗

Sorry to hear that, I have not seen this error. Perhaps try updating your libraries?



**Rahul** January 31, 2018 at 5:52 pm #

Sorry, If its a very basic question. I am a newb explanation.

I have a question at below code block, where we are s is the use of the output set ? What is its significance ?

```
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 1, 2018 at 7:16 am #

REPLY ↗

The output is the thing being predicted.

This post might help you understand how algorithms work:

<http://machinelearningmastery.com/how-machine-learning-algorithms-work/>



**Rahul** February 1, 2018 at 6:19 pm #

REPLY ↗

Jason, one more more clarification needed on the “output values” . In many articles , I have seen that ML works only on numeric values (even its of different type we need to convert it to numeric). Doesn’t it apply to the “output values” we are using ? Don’t we need to convert them to numeric ?

## Your Start in Machine Learning



**Jason Brownlee** February 2, 2018 at 8:09 am #

REPLY ↗

Generally, yes we do.



**Bipin Singh** January 31, 2018 at 8:43 pm #

REPLY ↗

Great article for beginners. Thanks you very much. Jason do you have any more articles for more in depth knowledge?



**Jason Brownlee** February 1, 2018 at 7:19 am

Yes, start here:

<https://machinelearningmastery.com/start-here/>



**Ityav Luke** February 1, 2018 at 1:20 pm #

Sir,

Through your article i have successfully installed python... ... unsuccess and every stage get success... Now as i tried to delve into this tutorial i am problems.

I first run a check on versions of libraries as you said and the result is okay:

Python: 2.7.14 |Anaconda custom (64-bit)| (default, Oct 15 2017, 03:34:40) [MSC v.1500 64 bit (AMD64)]

scipy: 0.19.1

numpy: 1.13.3

matplotlib: 2.1.0

pandas: 0.20.3

sklearn: 0.19.1

The next step which is to import libraries and i did by copy and pasting into a script file running with this command: python script.py and not error shown.

Where i had problem is to load the dataset csv from ML repo.

As i execute the command to load dataset from a script file

i have the following error

-----  
Traceback (most recent call last):

File "script.py", line 4, in

dataset = pandas.read\_csv(url, names=names)

NameError: name 'pandas' is not defined

## Your Start in Machine Learning

X

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Please what is the issue here?  
thanks



**Jason Brownlee** February 2, 2018 at 8:04 am #

REPLY ↗

Perhaps you have two versions of Python installed accidentally?



**Rahul** February 1, 2018 at 6:11 pm #

REPLY ↗

Got it now.

If i am correct, the initially supplied output values gives inputs, this would be the output ? And finally, based on entirely new inputs provided to the system ?



**Jason Brownlee** February 2, 2018 at 8:08 am #

Sorry, I don't follow.



**Bipin Singh** February 1, 2018 at 7:45 pm #

REPLY ↗

Just a minor suggestion which i encountered, pandas.tools.plotting is deprecated, use pandas.plotting instead.

Thanks 😊



**Jason Brownlee** February 2, 2018 at 8:16 am #

REPLY ↗

Thanks, fixed.



**chanid** February 1, 2018 at 8:44 pm #

REPLY ↗

Hello Jason,

I'm always fan of your tutorials. Please, have done any tutorials like this for explaining every algorithm in depth including mathematics behind them, how and what exactly happening in side the algorithm.

Thank you

Your Start in Machine Learning



**Jason Brownlee** February 2, 2018 at 8:18 am #

REPLY ↗

I have two books that explain how algorithms work:

<https://machinelearningmastery.com/products>



**Martine** February 2, 2018 at 8:25 pm #

REPLY ↗

Hello,

I get this error:

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/
170 if y_type not in ['binary', 'multiclass', 'multiclass-m
171 'multilabel-indicator', 'multilabel-sequences']:
-> 172 raise ValueError("Unknown label type: %r" % y
173
174
```

ValueError: Unknown label type: 'continuous'

I am using my own dataset. What is wrong here?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 3, 2018 at 8:35 am ..

REPLY ↗

Perhaps your dataset is the problem?



**Hugues Laliberte** February 4, 2018 at 7:12 am #

Hi Jason,

i'm also using my own dataset, and i get the same error as Martine above:

```
File "/Users/Hugues/anaconda3/lib/python3.6/site-packages/sklearn/utils/multiclass.py", line 172,
in check_classification_targets
raise ValueError("Unknown label type: %r" % y_type)
ValueError: Unknown label type: 'continuous'
```

I can check my dataset, but what should we be looking for ? I have used that dataset with the LSTM model without any error messages.

thanks

## Your Start in Machine Learning



The multiclass.py code that is giving the error is:

```
if y_type not in ['binary', 'multiclass', 'multiclass-multioutput',
    'multilabel-indicator', 'multilabel-sequences']:
    raise ValueError("Unknown label type: %r" % y_type)
```

line 172 is the last line

looks like 'continuous' is not expected. Where is 'continuous' coming from ?



**Hugues Laliberte** February 4, 2018 at 7:19 am #

REPLY ↗

my last column is binary, 0 or 1



**Hugues Laliberte** February 4, 2018 at 7:

googling this error code i find the following:  
"You are passing floats to a classifier which expects integers"

I thought my last column is categorical because there a way out ?



**Hugues Laliberte** February 4, 2018 at 7:37 am #

REPLY ↗

i changed my last column from 0 and 1 to 'zero' and 'one'  
now the error message changes to:  
ValueError: Unknown label type: 'unknown'  
I'm getting closer....



**Jason Brownlee** February 5, 2018 at 7:40 am #

REPLY ↗

Sorry, I have not seen this error before. Perhaps try posting to stackoverflow?



**Hugues** February 6, 2018 at 1:20 am #

REPLY ↗

i found the problem now. This part of your code above has to be changed according to the number of columns of our data set:

```
1 # Split-out validation dataset
2 array = dataset.values
3 X = array[:,0:4]
```

Your Start in Machine Learning

```

4 Y = array[:,4]
5 validation_size = 0.20
6 seed = 7
7 X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, )

```

So the 4 in X and Y needs to be changed. This seems obvious now but i'm new to Python and this is a rather dense language.

thanks a lot, the best output fo rmy data set is KNN with 85%. I will now try to improve on this by cleaning my data.



**Jason Brownlee** February 6, 2018 at 9:19 am #

REPLY ↗

Why does it need to be changed?



**jcridge** February 7, 2018 at 4:17 am #

Please change Section 2.1 out of date referen

#### CURRENT TEXT

from pandas.plotting import scatter\_matrix

#### TO REVISED TEXT

from pandas.tools.plotting import scatter\_matrix

as per comments already submitted

thanks

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 7, 2018 at 9:28 am #

REPLY ↗

The “pandas.tools.plotting” is outdated.

The latest version of Pandas uses “pandas.plotting”.

Consider updating your version of Pandas to v0.22.0 or higher.

Learn more here:

<https://pandas.pydata.org/pandas-docs/stable/visualization.html#scatter-matrix-plot>



**Phil** February 7, 2018 at 5:01 am #

REPLY ↗

Hi Jason

Apologies if this has already been asked.

## Your Start in Machine Learning

What would be the next step, therefore, if I wanted to apply this prediction to new data? I.e. if we got a new data set with just the measurements, how do we program the use of the predictions we've found to estimate the species?

P.s. great blog, really useful!



**Phil** February 7, 2018 at 5:06 am #

REPLY ↗

ah don't worry, you can just apply knn.predict() to a new array of the sizes right? That's easy



**Jason Brownlee** February 7, 2018 at 9:3

Correct.

Also see this post on creating a final model:

<https://machinelearningmastery.com/train-final-machine-learning-model/>



**jcridge** February 8, 2018 at 1:50 am #

RE: is the validation dataset nugatory given th

Whilst the idea of separating out a “final independent test data set (50 samples) away from the k-fold cross validation process seems nice, is it not actually wasting the opportunity to develop and compare the N model types using the larger and therefore more useful data set within the k-fold process ?

In short, the k-fold process seems to already be doing everything that the hold-out sample is purporting to do.

Out another way, surely the hold out data is no more independent than the i(th) hold out data partitioned within i(th) k-fold execution ?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 8, 2018 at 8:31 am #

REPLY ↗

There are many approaches at estimating out of sample model skill. I recommend finding an approach that is robust for your specific problem.



**Pallavee** February 9, 2018 at 6:28 pm #

REPLY ↗

Hello Jason,

## Your Start in Machine Learning

This post is a great starting point – I am new to coding (with only basics at hand), python with lot of interest in ML. The post has got me started with it... I was able to run most of the tutorial successfully with few experiments by changing the graphs, seed values, kfolds etc. Few questions though –

1. In one of the answers you have explained how kfolds works on February 17, 2017 –

Now in the for loop, where you define kfolds for a model at hand, that split is done only once right? I mean e.g. for LR, being first model to evaluate, we split the data of 120 in 10 folds with 12 items in each. Then as explained in the above post – The model is trained on the first 9 folds and evaluated on the records in the 10th. When we go for next set of 9, we are NOT resplitting the 120 items in new 10 sets right?

2. Also, when you say model is trained on first 9 folds – It means that we are looking at the relationships of the 4 numeric values and the class (out of 3 – Iris-setosa, Iris-versicolor, Iris-virginica) which they belong to, right?

3. When the dataset is split between X and Y values (Y values in X), where in the code are we actually mentioning that X are the independent variables and Y is our data?

Thanks a lot!

Pallavee



**Jason Brownlee** February 10, 2018 at 8:55 am

No, the same split into folds is reused with each time, systematically.

Yes, a fit model really means a learned mapping from inputs to outputs:

<http://machinelearningmastery.com/how-machine-learning-algorithms-work/>

We specify the inputs and outputs to the model as separate parameters in sklearn.



**Raghavendra** February 9, 2018 at 9:03 pm #

REPLY ↗

Hi Jason,

I am getting below errors.

Statement: from pandas.plotting import scatter\_matrix  
throws error as "No module named plotting"

Statement: from sklearn import model\_selection  
throws error as "cannot import name model\_selection"

Regards

Raghavendra

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 10, 2018 at 8:55 am #

REPLY ↗

You will need to update your version of pandas and sklearn to the latest versions.



**Bipin** February 9, 2018 at 9:34 pm #

REPLY ↗

Hi Jason on my dataset I used kfold but couldn't find any significant difference. Can you explain why this may happen. Also, does using kfold cross\_validation lead to overfitting?

P.S:

with cross\_validation without cross\_validation

LogisticRegression 0.816 0.816

LinearDiscriminantAnalysis 0.806 0.806

KNeighborsClassifier 0.79 0.79

DecisionTreeClassifier 0.810 0.816

GaussianNB 0.803 0.803

SVC 0.833 0.833

LinearSVC 0.806 0.806

SGDClassifier 0.7525 0.620

RandomForestClassifier 0.833 0.803

## Your Start in Machine Learning

X

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 10, 2018 at 8:56 am #

REPLY ↗

Both do the same job of performing k-fold cross validation.

You can overfit when evaluating models with cross validation, although it is less likely on average than using other evaluation methods.



**Akheel** February 10, 2018 at 6:36 pm #

REPLY ↗

Excellent tutorial Jason, and thanks very much for it.

One noob question here though –

Where do 'dataset' and 'plt' get associated in the code above? I ask this coz I don't see any code where we are associating 'dataset' and 'plt'; and yet when we call 'plt.show()', the plot that gets drawn has data from the 'dataset'.



**Jason Brownlee** February 11, 2018 at 7:53 am #

REPLY ↗

The dataset is loaded:

Your Start in Machine Learning

```
1 dataset = pandas.read_csv(url, names=names)
```

plt is the pyplot library

```
1 import matplotlib.pyplot as plt
```

A search on the page (control-f) would have helped you discover this for yourself.



**Akheel** February 13, 2018 at 12:51 am #

REPLY ↗

Thanks Jason, but that i know.

Let me try to make my question clearer –

From the examples I studied to understand py

1. set the range to be plotted along the x-axis
2. provide the corresponding values to be plot
3. Steps 1 and 2 are accomplished by the call
4. After the call to ‘plot’, the call to ‘show’ is m

ex:

```
e = np.arange(0.0, 2.0, 0.01)
f = 1 + np.sin(2*np.pi*t)
plt.plot(e, f)
plt.show()
```

As you can see, the call to ‘plot’ provides the values to plt and the call to show will cause the plotting and display of the same from ‘plt’.

However, in your example, I don’t see any line which is equivalent to the ‘plot’ call.

So my question is – When and where does ‘plt’ get the values from ‘dataset’ that it uses to draw the plot?

I hope it’s clearer now.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** February 13, 2018 at 8:04 am #

REPLY ↗

Here, I use pandas to make the calls to matplotlib via the pandas DataFrame (called dataset), then call plt.show().



**Mr D** February 11, 2018 at 7:58 am #

REPLY ↗

I installed Anaconda according to your instructions (<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>) but as I go to run python and check the versions of libraries I get this:

Your Start in Machine Learning

```
... import numpy
Traceback (most recent call last):
File "", line 2, in
ImportError: No module named numpy

How can I get passed this.
```



**Jason Brownlee** February 12, 2018 at 8:25 am #

REPLY ↩

It looks like numpy is not installed or you are trying to run code in a different version of Python from anaconda.



**Najmath** February 13, 2018 at 3:45 pm #

Hello Jason,  
I have a project in which it should predict the disease I  
and can you please help me with the attributes of sym



**Jason Brownlee** February 14, 2018 at 8:13 am #

REPLY ↩

I recommend this process:  
<https://machinelearningmastery.com/start-here/#process>



**pradnya** February 13, 2018 at 4:33 pm #

REPLY ↩

Thank you very much jason... for the great tutorial.  
its really great aratical...its help so much to our project..thanks...



**Jason Brownlee** February 14, 2018 at 8:14 am #

REPLY ↩

I'm glad it helped.



**Cor Colijn** February 16, 2018 at 10:10 am #

REPLY ↩

Hi Jason,  
Well I got the example running but only after I deleted

Your Start in Machine Learning

```

for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())

```

With "scoring=scoring" I received error message something like "scoring not defined". Then when I added "scoring=scoring" back I did not receive the error and the program runs fine.

What could this be?

Anyhow, great tutorial.

Regards,

Cor



**Jason Brownlee** February 16, 2018 at 2:57 pm #

Glad to hear you overcame your issue.

you might have missed a snippet from earlier in the



**Akshata** February 16, 2018 at 4:49 pm #

Hi Jason,

cv\_results = model\_selection.cross\_val\_score(model, X\_train, Y\_train, cv=kfold, scoring=scoring)

After typing that line in my command prompt, it shows this error:

Traceback (most recent call last):

File "", line 1, in

NameError: name 'model' is not defined

I tried copy pasting that line directly off the tutorial, I still faced the same error. What should I do??



**Jason Brownlee** February 17, 2018 at 8:40 am #

REPLY ↗

I think you may have missed some lines of code from the tutorial.



**Cor Colijn** February 16, 2018 at 11:52 pm #

REPLY ↗

Your Start in Machine Learning

I did get this exact error also. Then when I removed “scoring=scoring”, thinking ‘well, maybe the compiler or whatever is smart enough to deal with this’ , the code worked as expected. Then when I reinserted “scoring=scoring”, I did not get the error message and the code continued to run as expected.



**feedack** February 17, 2018 at 2:49 am #

REPLY ↗

When I run this code

```
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

i get this error

TypeError: cannot perform reduce with flexible type

and i get a blank graph where x\_axis and y\_axis both a

How do I fix it?

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 17, 2018 at 8:49 am #

REPLY ↗

Sorry, I have not seen this fault, perhaps post to stackoverflow?



**mufassal** February 19, 2018 at 3:37 am #

what algorithm should i use for weather prediction



**Jason Brownlee** February 19, 2018 at 9:09 am #

REPLY ↗

As far as I know, modern weather forecasting uses physical models, not machine learning methods.

That being said, if you do want to explore ML methods for weather forecasting, I would recommend this process:

<https://machinelearningmastery.com/start-here/#process>

**John Bagiliko** February 21, 2018 at 9:52 pm #

## Your Start in Machine Learning



```
from pandas.plotting import scatter_matrix
```

That did not work until I used

```
from pandas import scatter_matrix
```

Maybe this can help someone also.



**Jason Brownlee** February 22, 2018 at 11:17 am #

REPLY ↗

Interesting, perhaps you need to update your version of Pandas?

Here is the API for “pandas.plotting.scatter\_matrix”

<https://pandas.pydata.org/pandas-docs/stable/viz.html#scatter-matrix>



**Bob Fujita** February 22, 2018 at 11:56 am #

Just started your tutorial. Looks like the best I following error while trying to load the iris dataset. Would problem. Thanks.

```
=====
RESTART: /Users/TinkersHome/Doc
>>> dataset = pandas.read_csv(url, names=names)
Traceback (most recent call last):
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/urllib/request.py", line 1318, in
do_open
encode_chunked=req.has_header('Transfer-encoding')
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1239, in
request
self._send_request(method, url, body, headers, encode_chunked)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1285, in
_send_request
self.endheaders(body, encode_chunked=encode_chunked)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1234, in
endheaders
self._send_output(message_body, encode_chunked=encode_chunked)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1026, in
_send_output
self.send(msg)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 964, in send
self.connect()
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1400, in
connect
server_hostname=server_hostname)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 964, in send
self.connect()
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

```
_context=self, _session=session)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/ssl.py", line 814, in __init__
self.do_handshake()
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/ssl.py", line 1068, in
do_handshake
self._sslobj.do_handshake()
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/ssl.py", line 689, in do_handshake
self._sslobj.do_handshake()
ssl.SSLError: [SSL: CERTIFICATE_VERIFY_FAILED] certificate verify failed (_ssl.c:777)
```



**Jason Brownlee** February 23, 2018 at 11:51 am

Sorry, I have not seen this error. Perhaps you can post the code that caused the message?



**Angela** February 22, 2018 at 8:56 pm #

Hello experts,

When practise 5.Algorithm, I encountered this error message. I have checked all my packages, which are all up-to-date.  
Kindly please help me to fix it, thanks very much.

```
>>> # Spot Check Algorithms
... models = []
>>> models.append('LR', LogisticRegression())
>>> models.append('LDA', LinearDiscriminantAnalysis())
>>> models.append('KNN', KNeighborsClassifier())
>>> models.append('CART', DecisionTreeClassifier())
>>> models.append('NB', GaussianNB())
>>> models.append('SVM', SVC())
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
... kfold = model_selection.KFold(n_splits=10, random_state=seed)
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_state=seed)
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

```
>>> results.append(cv_results)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'cv_results' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'msg' is not defined
```



**Jason Brownlee** February 23, 2018 at 11:56 am #

Ensure that you copy all of the code for the example in the tutorial.



**Angela** February 23, 2018 at 8:38 pm #

REPLY ↗

I will retry. Thank you very much Jason. Cheers!



**Jason Brownlee** February 24, 2018 at 9:11 am #

REPLY ↗

Hang in there!



**Alan** February 22, 2018 at 11:32 pm #

REPLY ↗

Hi Jason,

Great tutorial, thanks!

I got an unique error that no one had posted here – special...

The error is at this line:

```
cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring='accuracy')
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

And it says: ValueError: This solver needs samples of at least 2 classes in the data, but the data contains only one class: 0.0

But my X\_train.shape shows (52480L, 25L) and my y\_train.shape is (52480L,). Any ideas please?

Thanks,  
Alan



**Jason Brownlee** February 23, 2018 at 11:58 am #

REPLY ↗

Hi Alan, it means that your data does not have enough samples from each class.

The dataset may be highly imbalanced.

If so, this post might give you some ideas:

<https://machinelearningmastery.com/tactics-to-convert-imbalanced-classes-machine-learning/>



**Bob Fujita** February 23, 2018 at 5:31 am #

Added the following lines to my load dataset function:

```
import ssl
ssl._create_default_https_context = ssl._create_unverified_context
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 23, 2018 at 12:03 pm #

REPLY ↗

Nice one!



**isaias** February 26, 2018 at 1:14 am #

REPLY ↗

Hello, Mr Jason!

I'm learning ML and PLN and i have a lot of doubts:

you can recommend some article, blog (and so on) to learn more about this? I have to implement a model switching different classifiers for predict/discriminate a class. The model is described below:

- I have a set S of words;
- Each word W of S is a class for prediction;

Two different of vector of features are used:

Your Start in Machine Learning

- 1 – The first is a vector which use PMI score between W and n-gram occurring before W and PMI between W and n-gram placed after W. Then, the vector length is twice length of S (set of words);
- 2 – Other is a vector of 500 most words (vocabulary) occurring in a context (variable size) surrounding all words of S. If the word (feature) exists in a sentence for training, the vector puts ‘1’ or ‘0’, otherwise. Frequency of word on document (context/sentence) don’t matter here.

I know that i have to vectorize features and create a array of counts, but i can't understand even a little about what way i've to follow after that steps (roughly explained).

Basically, above informations are the most important.

Finally, i wanna use the different classifiers in a “pluggable” way. Its possible?

Thanks in advance.



**Jason Brownlee** February 26, 2018 at 6:05 am #

My best advice for getting started with NLP:

<https://machinelearningmastery.com/start-here/#nlp>



**Phillip C.** February 26, 2018 at 11:43 pm #

Great tutorial!

In my case, I am POSTing the IRIS data to a Flask web service, but I don't see how to get that data into a pandas dataframe using any of the “read\_csv” or other methods available. I tried to use `io.String(csv_variable)`, then using `read_csv` on that, but it still doesn't work.

Suggestions?

Thanks,

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 27, 2018 at 6:32 am #

REPLY ↗

Perhaps try posting the question to stackoverflow?



**Griffin** February 27, 2018 at 2:14 am #

REPLY ↗

Hi Jason!

First of all, great introduction to cross validation! Your tutorial is comprehensive and I appreciate that you went through everything step-by-step as much as possible.

Just a question regarding section 5.3 Build Models. Th

Your Start in Machine Learning

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

As I have looked at other websites on cross validation as well, I am confused on the X and y inputs. Should it be X\_train and Y\_train or X and Y (original target and data)? Because I looked at sklearn documentation ([http://scikit-learn.org/stable/modules/cross\\_validation.html#cross-validation](http://scikit-learn.org/stable/modules/cross_validation.html#cross-validation)), it seems that the original target and data were used instead, and they did not perform a train\_test\_split to obtain X\_train and Y\_train.

Please clarify. Thank you!



**Jason Brownlee** February 27, 2018 at 6:36 am #

REPLY ↗

The goal in this part is to evaluate the skill sample of data from your domain.

Perhaps this post would clear things up for you:

<https://machinelearningmastery.com/difference-between-train-test-split-cross-validation/>



**Ron** February 28, 2018 at 1:19 pm #

What is the main objective of this project?



**Jason Brownlee** March 1, 2018 at 6:06 am #

REPLY ↗

To teach you something.

The model will learn the relationship between flower measurements and iris flower species. Once fit, it can be used to predict the flower species for new flower measurements.



**anushri** February 28, 2018 at 7:51 pm #

REPLY ↗

I believe there are many more pleasurable opportunities ahead for individuals that looked at your site.



**Jason Brownlee** March 1, 2018 at 6:12 am #

REPLY ↗

Thanks.

**Attharuddin** March 6, 2018 at 6:14 am #

Your Start in Machine Learning

for name, model in models:  
 kfolds = model\_selection.KFold(n\_splits=10, random\_state=seed)  
 cv\_results = model\_selection.cross\_val\_score(model, X\_train, Y\_train, cv=kfolds, scoring=scoring)  
 results.append(cv\_results)  
 names.append(name)  
 msg = "%s: %f (%f)" % (name, cv\_results.mean(), cv\_results.std())  
 print(msg)

I could not run this code, please help me out



**Jason Brownlee** March 6, 2018 at 6:21 am #

Why not? What was the problem?

REPLY ↗



**Christian Post** March 6, 2018 at 10:43 pm #

Great example to see what you can and can't do.  
 I ran this with my own sample and well, did not get over good 😊

I just had to do some small adjustment since this line is

```
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)
```

I had to change it because my dataset has only 3 independent variables:

```
# Split-out validation dataset
array = dataset.values
n = dataset.shape[1]-1
X = array[:,0:n]
Y = array[:,n]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)
```

I think this should work regardless of the number of attributes in any given dataset(?)

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 7, 2018 at 6:14 am #

REPLY ↗

Nice.



**mahima kapoor** March 7, 2018 at 1:43 am #

REPLY ↗

i need to build a taxi passenger seeking system using machine learning, i am a beginner. how should i go about it? please suggest some relevant source codes for reference



**Jason Brownlee** March 7, 2018 at 6:15 am #

X

Perhaps this process will help:

<https://machinelearningmastery.com/start-here/#p>



**Pauli Isoaho** March 10, 2018 at 8:49 am #

REPLY ↗

Excelnt guide, thank you

What enviroment you need to plot?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 11, 2018 at 6:16 am #

REPLY ↗

Thanks.

What do you mean by environment?



**Nick F** March 10, 2018 at 8:43 pm #

REPLY ↗

Thanks for the tutorial. When I run the code, the Support Vector Machine got the best score (precision 0.94), while the knn got precision 0.90, as in your example. I am using Python 3. Is the different result caused by the global warming? 😊



**Jason Brownlee** March 11, 2018 at 6:24 am #

REPLY ↗

Nice work.

A difference in results is caused by the stochastic  
<https://machinelearningmastery.com/randomness-in-machine-learning/>

## Your Start in Machine Learning



**Frank984** March 10, 2018 at 9:55 pm #

REPLY ↗

I have Python: 2.7.10 (default, May 23 2015, 09:40:32) and the following versions of the libraries:  
 scipy: 0.15.1  
 numpy: 1.9.2  
 matplotlib: 1.4.3  
 pandas: 0.16.2  
 sklearn: 0.18.1

I have modified your example considering the following structure for the dataset:

```
Age Weight Height Metbio RH Tair Trad PMV TSV gender
0 61 61.4 175 2.14 31.98 21.35 20.58 -0.38 0 male
1 39 81.0 178 2.19 46.88 24.25 24.09 0.30 1 male
[...]
```

All works fine, except for the following part:

I have created a validation dataset considering:

```
# Split-out validation dataset
array = dataset.values
X = array[:,0:8]
#the line above is interpreted as "all rows for columns
Y = array[:,9]
#the line above is interpreted as "all rows for column 9
validation_size = 0.20
# 20% as a validation dataset
seed = 7
#what does this parameter means?
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)
```

Now when I try to built and evaluate the 6 models with this code:

```
# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_st
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```

cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
results.append(cv_results)
names.append(name)
msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
print(msg)

```

It appears this message:

```

>>> # Spot Check Algorithms
... models = []
>>> models.append(('LR', LogisticRegression()))
>>> models.append(('LDA', LinearDiscriminantAnalysis()))
>>> models.append(('KNN', KNeighborsClassifier()))
>>> models.append(('CART', DecisionTreeClassifier()))
>>> models.append(('NB', GaussianNB()))
>>> models.append(('SVM', SVC()))
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
...     kfold = model_selection.KFold(n_splits=10, random_state=7)
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_st
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(mo
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> results.append(cv_results)
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> print(msg)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'msg' is not defined
>>>

```

Could you explain how can I solve?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Frank984** March 10, 2018 at 10:24 pm #

REPLY ↗

I have tried also anaconda prompt and the following versions:

Python 3.6.1 |Anaconda 4.4.0 (64-bit)| (default, May 11 2017, 13:25:24)  
scipy: 0.19.0  
numpy: 1.12.1  
matplotlib: 2.0.2  
pandas: 0.20.1  
sklearn: 0.18.1

Same error when I try to build and evaluate the six models considering the script of paragraph 5.3



**Jason Brownlee** March 11, 2018 at 6:26 am #

Versions look ok. Ensure you have all



**Jason Brownlee** March 11, 2018 at 6:26 am #

Looks like a copy-paste error.

Ensure you copy all of the code and maintain the s

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Frank984** March 12, 2018 at 5:51 am #

REPLY ↗

Solved considering this post:

<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/#comment-431754>



**Kevin** March 13, 2018 at 10:47 am #

REPLY ↗

Hi Jason,

Your Instruction were great. I am new to coding and I would like to know if you have codes for fantasy sports. Will the process above work with fantasy sports.



**Jason Brownlee** March 13, 2018 at 3:05 pm #

REPLY ↗

Not at this stage. I have worked on sports datasets using rating systems and had great success:

Your Start in Machine Learning

[https://en.wikipedia.org/wiki/Elo\\_rating\\_system](https://en.wikipedia.org/wiki/Elo_rating_system)



**Qasem** March 13, 2018 at 9:57 pm #

REPLY ↗

how long will it take to run the program? i follow all instruction, and there is no errors, but still running and only get the first graph, and the dataset description? is it take to long to complete run ? note i use windows 7



**Jason Brownlee** March 14, 2018 at 6:20 am #

REPLY ↗

Seconds. No more than minutes.



**Qasem** March 14, 2018 at 12:08 pm #

so what do you think is the problem?



**Qasem** March 14, 2018 at 12:27 pm #

I have done like this and its just work till # histograms, there problem the pycharm 3 does not show any error.

```
# Load libraries
import pandas
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

# Load dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)

# shape
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```

print(dataset.shape)
# head
print(dataset.head(20))
# descriptions
print(dataset.describe())
# class distribution
print(dataset.groupby('class').size())
# box and whisker plots
dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
plt.show()
# histograms
dataset.hist()
plt.show()
# scatter plot matrix
scatter_matrix(dataset)
plt.show()
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation
test_size=validation_size, random_state=s
# Test options and evaluation metric
seed = 7
scoring = 'accuracy'
# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold,
    scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

Your Start in Machine Learning

```
# Compare Algorithms
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()

# Make predictions on validation dataset
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
predictions = knn.predict(X_validation)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```



**Jason Brownlee** March 14, 2018 at 9:11 pm #

Perhaps try and run from the command line? It's easier to hide output messages and error messages.



**Qasem** March 14, 2018 at 9:11 pm #

I have solved the problem, where I should close the figures and the results will be displayed, I have tried to change the dataset for example to Heart Dataset, where there are 14 attributes and only two classes, for sure there were errors. Sir, if I use the heart dataset in which part of the project should I do the modifications? thanks in advance I'm just started to learn Python in Machine learning. your help is really appreciated



**Jason Brownlee** March 15, 2018 at 6:30 am #

This process will help you work through your problem systematically:  
<https://machinelearningmastery.com/start-here/#process>



**Daniel** March 13, 2018 at 10:50 pm #

REPLY ↗

Jason,

Thanks a bunch for the awesome example. Like others I received 0.991667 for SVM. The problem, however, I am having relates to the last step — getting prediction values. Below you can find my stack trace.

Your Start in Machine Learning

NOTE: I am mac with python 2.7

Any clue?

--

```
ValueError Traceback (most recent call last)
in ()
3 knn.fit(X_train, Y_train)
4 predictions = knn.predict(X_validation)
--> 5 print(accuracy_score(Y_validation, predictions))
6 print(confusion_matrix(Y_validation, predictions))
7 print(classification_report(Y_validation, predictions))
```

```
/usr/local/lib/python2.7/site-packages/sklearn/metrics/classification.pyc in accuracy_score(y_true, y_pred,
normalize, sample_weight)
```

174

```
175 # Compute accuracy for each possible representa
-> 176 y_type, y_true, y_pred = _check_targets(y_true
177 if y_type.startswith('multilabel'):
```

```
178 differing_labels = count_nonzero(y_true - y_pred,
```

```
/usr/local/lib/python2.7/site-packages/sklearn/metrics
```

```
69 y_pred : array or indicator matrix
```

70 """

```
--> 71 check_consistent_length(y_true, y_pred)
```

```
72 type_true = type_of_target(y_true)
```

```
73 type_pred = type_of_target(y_pred)
```

```
/usr/local/lib/python2.7/site-packages/sklearn/utils/validation.pyc in check_consistent_length(*arrays)
```

202 if len(uniques) > 1:

203 raise ValueError("Found input variables with inconsistent numbers of"

```
--> 204 " samples: %r" % [int(l) for l in lengths]
```

205

206

ValueError: Found input variables with inconsistent numbers of samples: [4, 30]

--



**Jason Brownlee** March 14, 2018 at 6:23 am #

REPLY ↗

I have not seen this error sorry. Perhaps double check that you have copied all of the code?



**Daniel** March 16, 2018 at 2:06 am #

REPLY ↗

Found it!!!

Did try to make some changes in the code but

Your Start in Machine Learning

Thanks a lot. That is an awesome example!



**Jason Brownlee** March 16, 2018 at 6:20 am #

REPLY ↗

Glad to hear it Daniel.



**Frank984** March 14, 2018 at 7:46 pm #

REPLY ↗

Hi Jason,

I have a dataset structured as reported here:

<https://app.box.com/s/mi97crz44bz2r7f96wy2z6ztf68c>

(you can download it here: <https://app.box.com/s/c2b>)

It is composed by 9871 rows e 5 columns:

<https://app.box.com/s/xasyqbhtsmov9gqnvg7siop47>

When I try to describe it only the first and second colu

<https://app.box.com/s/9wez8izysrfwivns0sus6ql2ahko>

Also if I try to plot a scatter matrix, the data of the first

<https://app.box.com/s/41x56gxd5bil0c4e0tz000433ph>

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 15, 2018 at 6:27 am #

REPLY ↗

Nice work. Note none of your links work.



**Frank984** March 15, 2018 at 6:07 pm #

REPLY ↗

I have solved the issue and cancelled the folder.



**Jason Brownlee** March 16, 2018 at 6:11 am #

REPLY ↗

Great!



**Abhay Sapru** March 16, 2018 at 6:42 am #

REPLY ↗

till step 5.2 its fine for me but from point 5.3 a

Your Start in Machine Learning

```
# Spot Check Algorithms
... models = []
>>> models.append('LR', LogisticRegression())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'LogisticRegression' is not defined
>>> models.append('LDA', LinearDiscriminantAnalysis())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'LinearDiscriminantAnalysis' is not defined
>>> models.append('KNN', KNeighborsClassifier())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'KNeighborsClassifier' is not defined
>>> models.append('CART', DecisionTreeClassifier())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'DecisionTreeClassifier' is not defined
>>> models.append('NB', GaussianNB())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'GaussianNB' is not defined
>>> models.append('SVM', SVC())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'SVC' is not defined
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
... kfold = model_selection.KFold(n_splits=10, random_state=seed)
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_state=seed)
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model_selection' is not defined
>>> results.append(cv_results)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'cv_results' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

Your Start in Machine Learning

```
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> print(msg)
```



**Jason Brownlee** March 16, 2018 at 2:20 pm #

REPLY ↗

It looks like you are not preserving the indenting of the code. White space is important in python, the tabs and new lines must be preserved



**Abhay Sapru** March 17, 2018 at 8:02 pm #

ok i'll try it on ipython may be directly  
and one more thing do i have to define a logo r  
in results square brackets



**Abhay Sapru** March 17, 2018 at 9:56 pm #

Below is the code i am trying to run:-

```
1 # Load libraries
2 import pandas
3 from pandas.plotting import scatter_matrix
4 import matplotlib.pyplot as plt
5 from sklearn import model_selection
6 from sklearn.metrics import classification_report
7 from sklearn.metrics import confusion_matrix
8 from sklearn.metrics import accuracy_score
9 from sklearn.linear_model import LogisticRegression
10 from sklearn.tree import DecisionTreeClassifier
11 from sklearn.neighbors import KNeighborsClassifier
12 from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
13 from sklearn.naive_bayes import GaussianNB
14 from sklearn.svm import SVC
15 # Load dataset
16 url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
17 names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
18 dataset = pandas.read_csv(url, names=names)
19 # shape
20 print(dataset.shape)
21 # head
22 print(dataset.head(20))
23 # descriptions
24 print(dataset.describe())
25 # class distribution
26 print(dataset.groupby('class').size())
27 # box and whisker plots
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

28 dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
29 plt.show()
30 # histograms
31 dataset.hist()
32 plt.show()
33 # scatter plot matrix
34 scatter_matrix(dataset)
35 plt.show()
36 # Split-out validation dataset
37 array = dataset.values
38 X = array[:,0:4]
39 Y = array[:,4]
40 validation_size = 0.20
41 seed = 7
42 X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X,
43 # Test options and evaluation metric
44 seed = 7
45 scoring = 'accuracy'
46 # Spot Check Algorithms
47 models = []
48 models.append(('LR', LogisticRegression()))
49 models.append(('LDA', LinearDiscriminantAnalysis()))
50 models.append(('KNN', KNeighborsClassifier()))
51 models.append(('CART', DecisionTreeClassifier()))
52 models.append(('NB', GaussianNB()))
53 models.append(('SVM', SVC()))
54 # evaluate each model in turn
55 results = []
56 names = []
57 for name, model in models:
58     kfold = model_selection.KFold(n_splits=10)
59     cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
60     results.append(cv_results)
61     names.append(name)
62     msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
63     print(msg)

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)


**Katti** March 17, 2018 at 2:59 am #

REPLY ↗

Where can we see the visual representation of variate and univariate plots? I'm only seeing textual representation of the data. Please notify where to type dataset.plot(..) code



**Katti** March 17, 2018 at 3:08 am #

REPLY ↗

My bad, I never used the plt.show() function to visualize my data. I can see the plots very nicely.



**Jason Brownlee** March 17, 2018 at 8:44 am #

REPLY ↗

Perhaps it would help you to re-read section 4 of the above tutorial?

## Your Start in Machine Learning



**German Loiti Azcue** March 19, 2018 at 8:29 pm #

REPLY ↩

Hi Jason, I really found your guide useful and easy to follow. I am developing my Master Thesis and I am trying to apply ML to predict electricity prices (therefore numerical class). Which algorithm would you recommend me more (more than one if it is possible)?

As far as I know, classification algorithms are used in those cases where the class is binary like in this example. Why do we compare regression model with other classification models in this example then? Does that make sense? Can regression models be applied for classification purposes and vice versa?

Again thanks for your help and your time.



**Jason Brownlee** March 20, 2018 at 6:17 am #

If you are predicting a quantity, you will want to testing a suite of methods to see which works best.

Here is more info on the difference between regression and classification:  
<https://machinelearningmastery.com/classification-vs-regression/>



**Sirish** March 22, 2018 at 3:24 am #

Why is that same dataset gave two different kinds of tools, LDA with R and KNN with Python?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 22, 2018 at 6:26 am #

REPLY ↩

What do you mean exactly?



**Vaibhav V** March 26, 2018 at 8:56 pm #

REPLY ↩

Well explained concept. Kudos to you.



**Jason Brownlee** March 27, 2018 at 6:34 am #

REPLY ↩

Thanks!

## Your Start in Machine Learning



**Danish bhatia** March 26, 2018 at 9:18 pm #

REPLY ↗

What is “seed” ?



**Jason Brownlee** March 27, 2018 at 6:35 am #

REPLY ↗

Good question.

The random number generator used in the splitting of data and within some of the algorithms is actually a pseudorandom number generator. We can seed it so that it will generate the same sequence of random numbers each time the code is run. This is what I got.

Learn more about this here:

<https://machinelearningmastery.com/randomness-in-python/>



**Mathew** March 27, 2018 at 7:33 am #

Hi Jason,

Thank you for the explanation. please find the below query

1. I changed file name to iris22==> it gave error OK
2. I removed all data in iris.data ==> it gave the same output.
3. If any changes in the iris.data file does not change the output

Can you please explain.

Mathews

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 27, 2018 at 4:15 pm #

REPLY ↗

Perhaps confirm that your modified file is still being loaded and used in the code?



**Prachi** May 8, 2018 at 6:30 pm #

REPLY ↗

Then is that command not required to actually run the code? Only to run it in a specific manner?

## Your Start in Machine Learning



Here is information on how to run a script from the command line:  
<https://machinelearningmastery.com/faq/single-faq/how-do-i-run-a-script-from-the-command-line>



**Saumya Gupta** March 27, 2018 at 10:12 pm #

REPLY ↗

Hey Jason,

I trained my data on a linear regression model, now I want to predict the value of label based on the values of indicators that the user inputs. Can this be done?

I'm really not getting it anywhere.

Please help me out



**Jason Brownlee** March 28, 2018 at 6:27 am #

Linear regression is a model for predicting

This post might clear things up for you:

<https://machinelearningmastery.com/classification/>



**Jeffrey Foster** April 1, 2018 at 2:22 pm #

I just want to say that this was fantastic. Knew the basics of Python and had it installed already, and everything worked without a hitch.

In my case I just wanted to get a sense of what's involved on a step by step level in machine learning but I'm definitely not a data scientist and only somewhat a developer, so while some of the concepts that came up are not familiar (not yet anyway) the whole thing gave me a good feel for what it would be like. Well done.



**Jason Brownlee** April 2, 2018 at 5:18 am #

REPLY ↗

Thanks Jeffrey, well done!



**Jarrar** April 3, 2018 at 6:35 am #

REPLY ↗

```
cv_results=model_selection.cross_val_score(model,X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
```

Your Start in Machine Learning

plz help me Sir i'll b very thankful to you



**Jason Brownlee** April 3, 2018 at 6:43 am #

REPLY ↗

Ensure you copy the complete code example.



**mars** April 3, 2018 at 7:08 pm #

REPLY ↗

hey Jason,

I currently working on ML projects and I found Gaussian process Regression to be the best choice for my problem.

In the validation phase, I predicted Values with an error of 10%.

Is this a good model? or do I need to retrained the data?

Thanks in advance for your reply!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**mars** April 3, 2018 at 11:13 pm #

I REFORMULATE MY QUESTION ABOVE

I am currently working on a ML project. I found Gaussian process Regression to be the best choice for my problem.

The validation error is twice higher than the trained model error.

Is this ok? or do I need to retrained the data or maybe look for another algorithm?

Thanks in advance for your reply!



**Jason Brownlee** April 4, 2018 at 6:10 am #

REPLY ↗

A good model can only be defined by comparing it to simple baseline methods like the Zero Rule method.

Alternately, you can interpret the RMSE using domain expertise because the units are the same as the output variable.



**Shamir** April 4, 2018 at 11:24 pm #

REPLY ↗

## Your Start in Machine Learning

Thanks so much Jason. After finishing this tutorial, what do you think are good next steps and projects to try to work on?

Thanks again – love your site!



**Jason Brownlee** April 5, 2018 at 6:03 am #

REPLY ↩

Perhaps start working through a suite of standard problems:

<http://machinelearningmastery.com/practice-machine-learning-with-small-in-memory-datasets-from-the-uci-machine-learning-repository/>

Get good at the process of working through problems.



**Megan** April 5, 2018 at 10:06 am #

Excellent intro tutorial — thank you for sharing!



**Jason Brownlee** April 5, 2018 at 3:12 pm #

Thanks, I'm glad it helped.

X

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Mujtaba ASAD** April 5, 2018 at 8:37 pm #

REPLY ↩

Hi Jason can u provide a link which guides the syntax of all model for validation that u have to use in this..

As you have only use KNN for validation but i want to all the other models for learning. as i am a total beginner and little bit bit confused what parameters to use in SVM or Linear Regression etc..



**Jason Brownlee** April 6, 2018 at 6:30 am #

REPLY ↩

I'm not sure I follow.

Perhaps here would be a good place to start:

<https://machinelearningmastery.com/start-here/#process>



**Jed** April 7, 2018 at 1:40 am #

REPLY ↩

## Your Start in Machine Learning

Great Article!! I would like to know how one could improve the accuracy of an algorithm such as KNN or Logistic regression?



**Jason Brownlee** April 7, 2018 at 6:35 am #

REPLY ↩

There are many ways, see this post for some ideas:

<http://machinelearningmastery.com/machine-learning-performance-improvement-cheat-sheet/>



**Gaurav Keswani** April 7, 2018 at 4:04 am #

REPLY ↩

`plt.boxplot(results)`

Error is showing in this statement while working in jupyter

`TypeError : cannot perform reduce with flexible type`



**Jason Brownlee** April 7, 2018 at 6:36 am #

REPLY ↩

I recommend not using a notebook.

Also, ensure you have all of the code for the example

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Shobha** April 10, 2018 at 2:30 pm #

REPLY ↩

I loved the tutorial. great work!!

first I tried it on ubuntu 14.04 LTS, but because of version problems, I had to upgrade to ubuntu 16.04 LTS. I could run the tutorial successfully. Thanks 😊



**Jason Brownlee** April 11, 2018 at 6:32 am #

REPLY ↩

I'm glad you got there in the end, well done.



**HKumar** April 12, 2018 at 7:37 pm #

REPLY ↩

Excellent tutorial Json. I am new to python as well to ML. It worked a like charm. Pls keep up the good work.

## Your Start in Machine Learning



**Jason Brownlee** April 13, 2018 at 6:38 am #

REPLY ↗

Thanks, I'm glad it helped!



**Ahmed Khan** April 14, 2018 at 6:29 am #

REPLY ↗

Hello Jason,

It is really a great article, I learned a lot.

One question:

How it will be used in production env or for a new exa



**Jason Brownlee** April 14, 2018 at 6:51 am #

See this post:

<https://machinelearningmastery.com/make-predictions-machine-learning-scikit-learn/>



**Ahmed Khan** April 14, 2018 at 12:04 pm #

Thank you!

So if I want to update data file, should I use all 5 attributes or only 4?

Please give an example.

Thanks,

Ahmed

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 15, 2018 at 6:20 am #

REPLY ↗

What do you mean by update the data file?



**rich** April 15, 2018 at 3:11 pm #

REPLY ↗

Hello! Great learning thank you for taking the time to do this. Few questions if you don't mind answering them i'm very very new to all this including python forgive me.

In 5.1 what is Seed? why is it 7?

## Your Start in Machine Learning

Also for the K-fold say you have 5 sets of data [1,2,3,4,5] each with 10 data set size do you do [1(for testing),2,3,4,5] and 2-5 as training until every bin has cycled through as testing set? Like after that it would be [1,2(for testing),3,4,5] and 1,3,4,5 as training until it's complete?

Also why do you have validation\_size = 0.20 if your using K-fold? Isn't K-fold cross validation already solving it?

Also now that we have the model how can I extract it? So I can use it so i can plug in my own values for the attributes and have the model give me a classification?



**Jason Brownlee** April 16, 2018 at 6:08 am #

REPLY ↗

Great questions!

Seed is the initialization of the pseudorandom number generator used by the algorithm and evaluation of the algorithm. The seed controls the randomness in ml here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>

Correct re k-fold cross-validation (CV). We use CV to evaluate the algorithm and use the validation set to confirm that indeed the error is reduced in some major way.

You can make use of the final model to make predictions.

<https://machinelearningmastery.com/make-predictions-machine-learning-models-python-scikit-learn/>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Arjun** April 18, 2018 at 1:43 am #

REPLY ↗

Hello Sir...

I'm truly saying from the bottom of my heart your tutorial really helps me a lot especially beginners like me. if you could also provide some more projects like above step-by-step procedures on like Titanic Data Set, Loan Prediction Data Set, Bigmart Sales Data Set and Boston Housing Data Set that would be really really a great help to beginners like me.



**Jason Brownlee** April 18, 2018 at 8:12 am #

REPLY ↗

Thanks. Yes. I have a few in my book.



**Hazem** April 18, 2018 at 6:09 pm #

REPLY ↗

Thank you very much for your interesting explanation

But I have an important question as to how we transfo

Your Start in Machine Learning

enter data for this plant and the application predicts any type of plant

I would be very thankful for this (how to convert the project into an application that can be used)

The application is also rich with Python with Anconda



**Jason Brownlee** April 19, 2018 at 6:27 am #

REPLY ↗

Great question. I would recommend start by collecting a large dataset of plant details and their associated species.



**Sanej** April 19, 2018 at 7:29 am #

Hello Jason,

Excellent tutorial It was such a fun runing the code. Th

Just in case if somebody else will get an error. When I

from pandas.plotting import scatter\_matrix

I get -> ImportError: No module named 'pandas.plotting'

I tried to update the pandas library -> not working

Solution was:

from pandas.tools.plotting import scatter\_matrix

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 19, 2018 at 2:46 pm #

REPLY ↗

Thanks, well done!

I recommend updating to the latest version of Pandas, you can learn more about this here:

[https://machinelearningmastery.com/faq/single-faq/i-think-you-meant-pandas-tools-plotting-scatter\\_matrix](https://machinelearningmastery.com/faq/single-faq/i-think-you-meant-pandas-tools-plotting-scatter_matrix)



**Chathura** April 25, 2018 at 3:53 pm #

REPLY ↗

I'm new in python and machine learning  
when i run the code i face an error in this line

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

It makes many errors and the final error given by running is

File "C:\Users\Chathura Herath\PycharmProjects\MoreModels\venv\lib\site-packages\sklearn\utils\validation.py", line 433, in chec

## Your Start in Machine Learning

```
array = np.array(array, dtype=dtype, order=order, copy=copy)
```

ValueError: could not convert string to float: 'PentalWidth'

please healp me



**Jason Brownlee** April 26, 2018 at 6:21 am #

REPLY ↗

I'm sorry to hear that, try these steps:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Chathura** April 25, 2018 at 4:04 pm #

cycler 0.10.0 0.10.0

kiwisolver 1.0.1 1.0.1

matplotlib 2.2.2 2.2.2

numpy 1.14.2 1.14.2

pandas 0.22.0 0.22.0

pip 9.0.1 10.0.1

pyparsing 2.2.0 2.2.0

python-dateutil 2.7.2 2.7.2

pytz 2018.4 2018.4

scikit-learn 0.19.1 0.19.1

scipy 1.0.1 1.1.0rc1

setuptools 28.8.0 39.0.1

six 1.11.0 1.11.0

sklearn 0.0 0.0

these are the installed packages

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 26, 2018 at 6:21 am #

REPLY ↗

So far so good.



**Neha** April 25, 2018 at 8:33 pm #

REPLY ↗

I am getting the same output for different active user input using KNN algorithm can you suggest something?

## Your Start in Machine Learning



**Jason Brownlee** April 26, 2018 at 6:29 am #

REPLY ↗

Here are some ideas:

<http://machinelearningmastery.com/machine-learning-performance-improvement-cheat-sheet/>



**darren** April 27, 2018 at 4:24 am #

REPLY ↗

this is a great start. works a treat. thank you.

for what its worth to others i installed py using anaconda.

there is an development environment in this called Spy



**Jason Brownlee** April 27, 2018 at 6:09 am #

Yes, but I generally recommend beginners

<https://machinelearningmastery.com/faq/single-page/#q1>



**Kevin Burke** April 27, 2018 at 5:00 am #

Hi Jason, hope all is well and thank you for all your support and inspiration to me...

I hope this has not been asked! So the goal is predicting outcomes on unseen data, what I would like to be able to do is say something like this.

"I predict with 90% accuracy that this rowid in the dataframe will be Iris-virginica."

But the rowid is not part of the training or test set

How can I tie my prediction to the rowid of the unseen data so I know which rowid I am referring to?

Thanks Jason

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 27, 2018 at 6:15 am #

REPLY ↗

The predict() function will take a list of rows and return a list of predictions in the same order. The order links the two.

Learn more about how to make predictions here:

<https://machinelearningmastery.com/make-predictions-scikit-learn/>

## Your Start in Machine Learning



**Kevin Burke** April 27, 2018 at 8:51 pm #

REPLY ↗

Thank you Jason, awesome! (I'm looking for a DS mentor, interested??!!) 😊



**Jason Brownlee** April 28, 2018 at 5:29 am #

REPLY ↗

I answer this question here:

<https://machinelearningmastery.com/faq/single-faq/can-you-be-my-mentor-or-coach>



**Peter** May 8, 2018 at 7:44 am #

i'm not new to machine learning but new to p  
You skip certain parts to start it all..



**Jason Brownlee** May 8, 2018 at 2:49 pm #

I had to draw the line somewhere for a on

What are the most important topics do you think I



**ro** May 8, 2018 at 10:36 am #

REPLY ↗

hello

```
models.append(('LR',LogisticRegression()))
models.append(('LDA',LinearDiscriminantAnalysis()))
models.append(('KNN',KNeighborsClassifier()))
models.append(('CART',DecisionTreeClassifier()))
models.append(('NB',GaussianNB()))
models.append(('SVM',SVC()))
are there more for cosine similarity, euclidean distance, mahalanobis distance?
```



**Jason Brownlee** May 8, 2018 at 2:53 pm #

REPLY ↗

Do you mean as distance functions on the knn?

Here's advice on changing the distance function:

<http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

Your Start in Machine Learning



**Prachi** May 8, 2018 at 6:25 pm #

REPLY ↗

What is a confusion matrix and how do I read it?



**Jason Brownlee** May 9, 2018 at 6:18 am #

REPLY ↗

You can learn about the confusion matrix here:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>



**Ali** May 10, 2018 at 9:00 am #

Waw Dr. this is amazing. You made it very easy! Thank you so much! Greetings from the USA!



**Jason Brownlee** May 11, 2018 at 6:31 am #

Thanks.



**Ahmad Zaki** May 14, 2018 at 5:40 pm #

REPLY ↗

Hi Jason

Thanks for the work youve done im sure its been a great help for a lot of people.

So i wanted so make sure of something. in step number 5 and 6 which is evaluating an algorithm and making predictions. So step 5 basically dividing 80% of the data to become training data and the 20% to validate the trained model.

What i wanted to ask is when we use the 10-fold cross validation to estimate accuracy of the model, we split up the dataset to 10 part, 9 of which we use to train and 1 part of the dataset to test the model. Now is the dataset were dividing from the training part of the original dataset or in other words 80% of the original dataset?

Another thing is it says that the 10-fold cross validation to spilt the dataset into 10 parts then train and validate for all combinations of train and test splits. It means that for 1 combination of train and test data, lets say the first of the ten part of data becomes the test data while the rest becomes the train data, then on another combination of train test data, the second part of the ten part of data becomes the test data etc for all combinations?

Thanks a lot

Zaki

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 15, 2018 at 7:52 am #

REPLY ↗

It is a choice. It can be a good idea to hold back a portion of the dataset to validate the final model.

Learn more here:

<https://machinelearningmastery.com/difference-test-validation-datasets/>



**Hari M** May 16, 2018 at 10:07 pm #

REPLY ↗

Hi Jason....

Your efforts are really helpful for me.

I am learning the code line by line. What is meant by validation set .

seed = 7

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=0.3, random_state=seed)
```

Why do we use seed. Also why it is hardcoded as 7.

Can you please let me know

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 17, 2018 at 6:32 am #

REPLY ↗

This is a common question that I answer here:

<https://machinelearningmastery.com/faq/single-faq/what-value-should-i-set-for-the-random-number-seed>



**Felipe Fernandes** May 17, 2018 at 5:51 am #

REPLY ↗

Jason, thank you for your post. I am from Rio de Janeiro, Brazil and I am currently finishing my Computer Engineering course on College. We have learned the very basics of machine learning. It would be very useful if you go ahead and show us how to feed these algorithms with real images and show us the result.

I am using Sublime Text as the IDE and Python 2.7 with all the necessary environment. Your tutorial worked fine for me, without any error when building.

Your Start in Machine Learning



**Jason Brownlee** May 17, 2018 at 6:40 am #

REPLY ↗

Thanks for the suggestion.



**Abhijit** May 17, 2018 at 2:58 pm #

REPLY ↗

hey jason, thanks for post, i completed intro course of machine learning on udacity but didnt able to hand on code that much. without application and practising codes there is no way to learn. please suggest me the project based website for practise and anything new i should do as per your concern...



**Jason Brownlee** May 17, 2018 at 3:14 pm #

Here are some suggested projects:

<https://machinelearningmastery.com/faq/single-faq/>



**Noah Roberts** May 18, 2018 at 8:39 am #

X

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 18, 2018 at 9:49 am #

REPLY ↗

Perhaps your environment is not installed correctly?

This tutorial might help:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Jason** May 18, 2018 at 1:30 pm #

REPLY ↗

Hi,

What is the function of the instructions above, and how would we implement this into our own programs?



**Jason Brownlee** May 18, 2018 at 2:40 pm #

REPLY ↗

What do you mean exactly?

## Your Start in Machine Learning



**Jonathan** May 22, 2018 at 12:04 pm #

REPLY ↗

Hi, I just started out in ML and tried to run your code in the Anaconda command line and am getting the following error in the code below. Thanks

```
#Spot Check Algorithms
... models = []
>>> models.append(('LR',LogisticRegression()))
>>> models.append(('LDA',LinearDiscriminantAnalysis()))
>>> models.append(('KNN',KNeighborsClassifier()))
>>> models.append(('CART',DecisionTreeClassifier()))
>>> models.append(('NB',GaussianNB()))
>>> models.append(('SVM',SVC()))
>>> #evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
...     kfold = model_selection.cross_val_score(model, X_
File "", line 2
kfold = model_selection.cross_val_score(model, X_trai
^
IndentationError: expected an indented block
>>> kfold= model_selection.cross_val_score(model, X_
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> kfold = model_selection.KFold(n_splits=10,random_state=seed)
>>> cv_results= model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> results = []
>>> names = []
>>> for name, model in models:
...     kfold = model_selection.KFold(n_splits=10,random_state=seed)
File "", line 2
kfold = model_selection.KFold(n_splits=10,random_state=seed)
^
IndentationError: expected an indented block
>>> kfold= model_selection.KFold(n_splits=10,random_state=seed)
>>> cv_results= model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.** Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 22, 2018 at 2:57 pm #

REPLY ↗

It looks like you might not have copied the code with all of the indenting.

This might help:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-copy-code-from-a-tutorial>



**kotrappa sirbi** May 23, 2018 at 11:01 pm #

REPLY ↗

array = dataset.values

NameError: name 'dataset' is not defined



**Jason Brownlee** May 24, 2018 at 8:13 am #

Ensure you copy all of the code required.



**Sorina Chirilă** May 24, 2018 at 7:40 pm #

Hello, Jason, Great, great article. Thank You 😊

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 25, 2018 at 9:22 am #

REPLY ↗

Thanks!



**Jonathan** May 25, 2018 at 5:35 am #

REPLY ↗

I will try that, thanks very much!



**Sreenivasa Rao Gubba** May 25, 2018 at 9:14 pm #

REPLY ↗

Hi Jason

I started working on this project. I have encountered an issue with 5.1

array = dataset.values

Your Start in Machine Learning

it is saying ndarray object of numpy module. I am using latest Anaconda. I have check the installs as you mentioned. All modules are installed and are of higher version.

your help is much appreciated.

Sreenivasa



**Jason Brownlee** May 26, 2018 at 5:57 am #

REPLY ↗

Did you copy all of the code?



**Jaya** May 27, 2018 at 2:26 am #

hai jason

this is good publication

I know the ML algorithms theory wise but new to practice. But by following your tutorial I could install all the libraries. As I started to implement "your first machine learning script" code.

There is no >>> prompt in anaconda prompt.

Please help me its all new. Should I type every thing in python filename.py

or should i type the code separately

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 27, 2018 at 6:48 am #

REPLY ↗

The code goes into a script and is run from the command line.

More on running code from the command line here:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-run-a-script-from-the-command-line>



**Jaya** May 27, 2018 at 2:57 am #

REPLY ↗

Hai Jason

Finally I got it.

It was thrilling

Thank you

## Your Start in Machine Learning



**Jason Brownlee** May 27, 2018 at 6:50 am #

REPLY ↗

Well done!



**Bento Silva** May 29, 2018 at 5:12 am #

REPLY ↗

Great tutoria! Thanks!

My results:

LR: 0.966667 (0.040825)  
 LDA: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.966667 (0.040825)  
 NB: 0.975000 (0.053359)  
 SVM: 0.991667 (0.025000)

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**mike** June 2, 2018 at 1:33 am #

REPLY ↗

Great intro!! Really appreciated. The one part that didn't work for me was all the plt.show(). I have triple checked my versions. Any idea what I am doing wrong?



**Jason Brownlee** June 2, 2018 at 6:38 am #

REPLY ↗

Perhaps you are running inside an IDE or notebook instead of from the commandline?



**Amarnath** June 3, 2018 at 2:58 pm #

REPLY ↗

Hi Jason,  
 Thanks for the post.

i have tried your above approach on Iris data set with seed = 7, i got the same result as expected in this approach. when i tried the below approach with seed (or) random state=42 . aettina the 100 % accuravc. i

Your Start in Machine Learning

didn't understand why changing the seed (or) random\_state=42 increased the performance or there is any mistake in my code ?

Please find the belowcode

```
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 42
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)
```

```
# Test options and evaluation metric
seed = 42
scoring = 'accuracy'
```

```
# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

Result :

```
LR: 0.950000 (0.040825)
LDA: 0.975000 (0.038188)
KNN: 0.950000 (0.055277)
CART: 0.950000 (0.055277)
NB: 0.950000 (0.055277)
SVM: 0.958333 (0.041667)

# Make predictions on validation dataset
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
#predictions = []
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

```
#print(predictions)
predictions = knn.predict(X_validation)
#print(X_validation)
#print(predictions)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

Result :

```
1.0
[[10 0 0]
 [ 0 9 0]
 [ 0 0 11]]
precision recall f1-score support
Iris-setosa 1.00 1.00 1.00 10
Iris-versicolor 1.00 1.00 1.00 9
Iris-virginica 1.00 1.00 1.00 11
avg / total 1.00 1.00 1.00 30
```

```
# Make predictions on validation dataset
svc = SVC()
svc.fit(X_train, Y_train)
#predictions = []
#print(predictions)
predictions = svc.predict(X_validation)
#print(X_validation)
#print(predictions)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

Result :

```
1.0
[[10 0 0]
 [ 0 9 0]
 [ 0 0 11]]
precision recall f1-score support
Iris-setosa 1.00 1.00 1.00 10
Iris-versicolor 1.00 1.00 1.00 9
Iris-virginica 1.00 1.00 1.00 11
avg / total 1.00 1.00 1.00 30
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



You can learn more about the impact of randomness in machine learning here:  
<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Ahmed Yunus** June 3, 2018 at 6:30 pm #

REPLY ↗

Hello sir ,

In this tutorial you have showed a basic project which load pre-defined dataset.Can you please tell me how can I create my own dataset and load it here ? And also I have trained data and now how can I input new image so that machine can identify that and print it's name ?



**Jason Brownlee** June 4, 2018 at 6:23 am #

X

This post shows you how to load a new dataset  
<http://machinelearningmastery.com/load-machine-learning-dataset/>

This post shows you how to make a prediction with your own dataset  
<https://machinelearningmastery.com/make-predictions-machine-learning-dataset/>

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Karthik** June 3, 2018 at 8:06 pm #

What type of dataset can be used for the linear regression? (Can we use all types of datasets?)

REPLY ↗



**Jason Brownlee** June 4, 2018 at 6:24 am #

Numerical data input and numerical data output.

REPLY ↗



**Karthik** June 3, 2018 at 8:09 pm #

How to select a particular dataset for particular algorithm (knn, linear regression.....)?

REPLY ↗



**Jason Brownlee** June 4, 2018 at 6:24 am #

This is a common question that I answer here:

<https://machinelearningmastery.com/faq/single-faq/what-algorithm-config-should-i-use/>

## Your Start in Machine Learning



**Jorge** June 5, 2018 at 9:42 am #

REPLY ↗

Hi, in what part of the code can I put my new data for classification?



**Jason Brownlee** June 5, 2018 at 3:05 pm #

REPLY ↗

This post explains how to make a prediction:

<https://machinelearningmastery.com/make-predictions-scikit-learn/>



**ajay** June 9, 2018 at 12:22 am #

i am a high school passed out and wanted to these things



**Jason Brownlee** June 9, 2018 at 6:55 am #

Great!



**John David** June 9, 2018 at 8:43 pm #

REPLY ↗

I only came recently across this blog post. Very well written, congratulations. I have a question about the ‘brute force’s approach you used to define the best predictive ML approach. You tried all of them. But due to the very small dataset would you rely on such a small difference? That is within the variance of the model, so I could pick almost any of those. Do you have posted about a dataset ? eventually larger) where trends might be eventually different?



**Jason Brownlee** June 10, 2018 at 6:02 am #

REPLY ↗

Indeed, with overlapping skill scores, we might have to use statistical hypothesis tests to see if indeed there is a meaningful difference between the skill of the different methods. The student’s t-test would be a good starting point.



**Padmaja Shukla** June 11, 2018 at 1:34 pm #

REPLY ↗

## Your Start in Machine Learning

Very nice blog to start with. Thanks for the same. I am following most of your emails in my ML journey.

Started a week ago.

A small issue in this blog.

```
from sklearn.neighbors import KNeighborsClassifiers
```

Traceback (most recent call last):

File "", line 1, in

```
from sklearn.neighbors import KNeighborsClassifiers
```

ImportError: cannot import name 'KNeighborsClassifiers'

Please suggest .. Rest all I am able to understand



**Jason Brownlee** June 11, 2018 at 1:51 pm #

Perhaps ensure that you have the sklearn

This tutorial can help you to setup your environment

<https://machinelearningmastery.com/setup-python-for-machine-learning/>



**Luiz** June 12, 2018 at 2:59 am #

Awesome stuff! one thing, when you apply the new mapping function or it uses the one it created during the test phase?



**Jason Brownlee** June 12, 2018 at 6:47 am #

REPLY ↗

In knn, the training data is used to make a prediction on the test dataset.



**Maker Athian** June 12, 2018 at 7:44 pm #

REPLY ↗

Good afternoon sir,

I am having network problem, I downloaded the Iris dataset on my directory, kindly how do I load the dataset to my python IDE?

Thanks,

Maker

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 13, 2018 at 6:17 am #

REPLY ↗

I recommend using a text editor, not an IDE.

You can copy the .csv file into the same directory as your .py files.



**heybqqy** June 14, 2018 at 10:02 pm #

REPLY ↗

ty for this m8 😊 very good toot



**Jason Brownlee** June 15, 2018 at 6:44 am #

X

Thanks.



**Luke** June 15, 2018 at 12:14 am #

REPLY ↗

This was incredible, thank you so much. A ve



**Jason Brownlee** June 15, 2018 at 6:44 am #

REPLY ↗

I'm glad it helped.



**Dipanjan Moitra** June 17, 2018 at 5:29 am #

REPLY ↗

Hi,

I am getting this error when I am running the code with my own dataset:

ValueError: Unknown label type: 'continuous'

my dataset is having 161 instances and 54 attributes.

Please help!



**Jason Brownlee** June 17, 2018 at 5:42 am #

REPLY ↗

Looks like you need to change your output type from classification to regression.

Your Start in Machine Learning



**Robin** June 19, 2018 at 7:34 pm #

REPLY ↗

Having the following error

```
NameError: name 'msg' is not defined
>>> models = []
>>> models.append(('LR', LogisticRegression()))
>>> models.append(('LDA', LinearDiscriminantAnalysis()))
>>> models.append(('KNN', KNeighborsClassifier()))
>>> models.append(('CART', DecisionTreeClassifier()))
>>> models.append(('NB', GaussianNB()))
>>> models.append(('SVM', SVC()))
>>> results = []
>>> names = []
>>> for name, model in models:
... kfold = model_selection.KFold(n_splits=10, random_
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_st
^
```



**Jason Brownlee** June 20, 2018 at 6:25 am #

×

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



Ensure you copy all of the code in the example and ensure indenting matches the example.

Learn how to copy code from the tutorial here:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-copy-code-from-a-tutorial>

REPLY ↗



**Manoj T** June 22, 2018 at 9:14 pm #

Thank you Dr.Jason for writing wonderful simple Machine learning project for the beginners. I am getting exactly same results for the accuracy as given in your tutorial. I am finding bit difficulty in interpreting statistical results.



**Jason Brownlee** June 23, 2018 at 6:17 am #

REPLY ↗

Well done.

What results are you having trouble with?

Your Start in Machine Learning



**Alice** June 25, 2018 at 5:24 pm #

REPLY ↗

Hi,

I have been working on binary text classification, so, I used the above code but before predicting the output I converted it into numerical data using

```
df = handle_non_numerical_data(dataset)
```

now, Prediction on training,validation data all worked fine, but How to give a new set to predict the class, when I am trying to use the above function it classifies the new dataset differently as in there is no relation between training dataset and this dataset. How to solve this problem ?



**Jason Brownlee** June 26, 2018 at 6:34 am #

X

What is the function "handle\_non\_numerical\_data"?



**Kaushal Dave** June 28, 2018 at 4:27 pm #

REPLY ↗

Hello Jason,

I am a newbie, trying to learn Machine learning with little knowledge. Its awesome to learn it from here!!!

I want to know 1 thing here why we have separated data into X\_train and Y\_train? Can't we keep the data and classes in one single table say X\_train only so that the very first row contains the data and the second row contains the classes say

5.9,3,5.1,1.8,Iris-virginica

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 29, 2018 at 5:51 am #

REPLY ↗

The models learn a mapping from inputs to outputs.

The libraries expect the data to be separated. This is why we separate them.



**Maria Shoukat** June 28, 2018 at 8:10 pm #

REPLY ↗

Assalam-o alaikum!

Very nice tutorial.. Can you give me any idea about simplest implementation of any of Machine Learning algorithms for processing big data? I want the implementation to in Python like you have did above in your tutorial.

Regards

Your Start in Machine Learning



**Jason Brownlee** June 29, 2018 at 5:57 am #

REPLY ↗

I provide a suite of tutorials that you can use to get started here:

<https://machinelearningmastery.com/start-here/#python>



**Devin Crane** June 30, 2018 at 1:30 am #

REPLY ↗

I have a few questions:

- 1) How do I print out the confusion matrix of TP, FP, TN, FN, rather than just the precision, recall, etc?
- 2) How do I just train on one set of data and test on another?
- This would require the ability to save my model. How can I do this without the need to re-train?
- 3) Is there a best way to selectively scale discrete values?
- 4) Is the n\_spits always a good way to go? How do I know if it's not?

Thanks

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 30, 2018 at 6:13 am #

Good questions Devin .

I have more on the confusion matrix here, including how to print it:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>

You can call:

```
1 model.fit(trainX,trainy)
2 yhat = model.predict(testX)
```

More here:

<https://machinelearningmastery.com/make-predictions-scikit-learn/>

You will have to split the data up by column, scale it, then reassemble. Look into using slicing to select and hstack() to combine, more here:

<https://machinelearningmastery.com/gentle-introduction-n-dimensional-arrays-python-numpy/>

10 splits for CV has been found to be effective on a wide range of problems, more here:

<https://machinelearningmastery.com/k-fold-cross-validation/>



**NAVEEN KUMAR** July 5, 2018 at 5:38 am #

REPLY ↗

## Your Start in Machine Learning

hi jason  
how KNN is better  
can you explain on what basis we find the better one algorithm



**Jason Brownlee** July 5, 2018 at 8:03 am #

REPLY ↗

We can choose an algorithm based on its average expected performance when making predictions on unseen data.



**Sanjib** July 6, 2018 at 10:19 pm #

Hello Jason,

I am stuck at confusion matrix. looking at the output b

```
[[ 7 0 0]
 [ 0 11 1]
 [ 0 2 9]]
```

I was trying to follow below statements, but could not versicolor/Iris-virginica) looking at above output matrix

Expected down the side: Each row of the matrix corresponds to a predicted class.  
Predicted across the top: Each column of the matrix cor

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 7, 2018 at 6:16 am #

REPLY ↗

I explain more here:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>



**Sanjib** July 7, 2018 at 11:29 pm #

REPLY ↗

Thank you.



**fawaz** July 8, 2018 at 7:48 am #

REPLY ↗

Hello Doctor, First of all, thank you very much for this tutorial.  
I have implemented this code on my own dataset that I have created. It is one class to differentiate between two types of attacks. The dataset contain 267 features experimental, I created randomly a small database of 2

## Your Start in Machine Learning

output is as follows:

```
LR: 0.927639 (0.020943)
LDA: 0.964074 (0.008784)
KNN: 0.763901 (0.045070)
CART: 0.979401 (0.007253)
NB: 0.680964 (0.021898)
SVM: 0.560485 (0.022857)
```

---

```
-----SVM-----
```

```
0.5464135021097046
```

```
[[256 0]
```

```
[215 3]]
```

```
precision recall f1-score support
```

```
Benign 0.54 1.00 0.70 256
malicious 1.00 0.01 0.03 218
```

```
avg / total 0.75 0.55 0.39 474
```

---

```
-----Decision Tree Classifier (CART) -----
```

```
accuracy_score=:
```

```
0.9852320675105485
```

```
confusion_matrix=:
```

```
[[252 4]
```

```
[ 3 215]]
```

```
classification_report=:
```

```
precision recall f1-score support
```

```
Benign 0.99 0.98 0.99 256
```

```
malicious 0.98 0.99 0.98 218
```

```
avg / total 0.99 0.99 0.99 474
```

---

```
-----LinearDiscriminantAnalysis-----
```

```
Warning (from warnings module):
```

```
File "C:\python36\lib\site-packages\sklearn\discriminant_analysis.py", line 388
```

```
warnings.warn("Variables are collinear.")
```

```
UserWarning: Variables are collinear.
```

```
accuracy_score=:
```

```
0.9556962025316456
```

```
confusion_matrix=:
```

```
[[256 0]
```

```
[ 21 197]]
```

```
classification_report=:
```

```
precision recall f1-score support
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

Your Start in Machine Learning

Benign 0.92 1.00 0.96 256  
 malicious 1.00 0.90 0.95 218  
 avg / total 0.96 0.96 0.96 474

---

.Note that this is the first test of samples of dataset.

Does this look right? does makes sense

If the problem is not linear why the result is less in SVM? while in the CART (0.99)

Any suggestion would be appreciated

Thank you introduction



**Jason Brownlee** July 9, 2018 at 6:30 am #

It is always a good idea to test a suite of models on your data.  
 We cannot know a priori.



**Naveen** July 9, 2018 at 2:38 am #

hi jason  
 tell me after getting 90% accuracy how i predict the values  
 with practical

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 9, 2018 at 6:36 am #

REPLY ↗

This post explains how to make predictions:

<https://machinelearningmastery.com/how-to-make-classification-and-regression-predictions-for-deep-learning-models-in-keras/>



**Ahmed** July 12, 2018 at 6:00 pm #

REPLY ↗

Thanks, I like that you've mentioned in the end of the tutorial, that we don't have to know or understand everything in the tutorial.

I like that your lesson are so concise. long tutorial make me lost

my question is where should I go from here so I can understand and apply the machine learning to my goals

**Jason Brownlee** July 13, 2018 at 7:35 am #

Your Start in Machine Learning



Thanks.

A next step would be here:

<https://machinelearningmastery.com/start-here/#python>



**Ahmed** July 13, 2018 at 11:41 pm #

REPLY ↗

Man!, where are you before few months!  
you replay fast, and you are always following up with your students  
I lost so much time trying to read over the internet to get started  
I wish that I found your tutorials before few months ago  
please keep doing what you are doing now  
Thanks a lot



**Jason Brownlee** July 14, 2018 at 6:

Thanks!



**Shekhar** July 12, 2018 at 9:24 pm #

Installed sklearn still got ImportError: No module named discriminant\_analysis. any suggestion?



**Jason Brownlee** July 13, 2018 at 7:40 am #

REPLY ↗

Are you able to confirm that you have the latest version of sklearn installed?



**Rahul** July 13, 2018 at 1:44 pm #

REPLY ↗

Hi Jason

First of all thanks for helping newbie.

I want to know what are the prerequisite to learn this course as i have no understanding of python.



**Jason Brownlee** July 14, 2018 at 6:12 am #

REPLY ↗

Your Start in Machine Learning

Perhaps start with Weka instead:

<https://machinelearningmastery.com/how-to-run-your-first-classifier-in-weka/>



**Deepika** July 13, 2018 at 7:13 pm #

REPLY ↗

Hi jason!

i have more interested ML . I'm in a beginner stage now .

I have one doubt

ML is, that

“we giving past input and output data , based on that we are expecting machines to give same output as in the past data for our future input”????

Like the following

data set:

input output

AA 1

BB 2

CC 3

in future if i give AA it should return 1.

but tradition programming also doing the same right?

only one thing is different that is unsupervised learning

kindly clarify my doubt ..

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 14, 2018 at 6:16 am #

REPLY ↗

The model does not memorize, instead it generalizes.

More information here:

<https://machinelearningmastery.com/what-is-generalization-in-machine-learning/>



**Ganesh** July 13, 2018 at 10:08 pm #

REPLY ↗

Hi

there is no prediction algorithm here ?

how to make the prediction step

how many variable of test data will be used to prediction ?

where is x – and y axis column

you just build the model gives good accuracy but how to make use of prediction

## Your Start in Machine Learning

Regards,  
Ganesha



**Jason Brownlee** July 14, 2018 at 6:18 am #

REPLY ↗

You can learn more about how to make predictions with your final model here:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-make-predictions>

REPLY ↗



**Ally** July 15, 2018 at 7:46 am #

Thank you for this, this is amazing. Helped me a lot.

Thanks again.



**Jason Brownlee** July 16, 2018 at 6:09 am #

You're welcome, I'm glad to hear that.



**swati** July 17, 2018 at 9:54 pm #

I am using url = "https://raw.githubusercontent.com/uiuc-cse/data-fa14/gh-pages/data/iris.csv" since UCI is not working.

All the code is getting executed but plt.hist() is showing error

-----  
ValueError Traceback (most recent call last)

in ()

1 # histograms

—> 2 dataset.hist()

3 plt.show()

```
~\Anaconda3\lib\site-packages\pandas\plotting\_core.py in hist_frame(data, column, by, grid, xlabelsize,
```

```
xrot, ylabelsize, yrot, ax, sharex, sharey, figsize, layout, bins, **kwds)
```

```
2176 fig, axes = _subplots(naxes=naxes, ax=ax, squeeze=False,
```

```
2177 sharex=sharex, sharey=sharey, figsize=figsize,
```

```
-> 2178 layout=layout)
```

```
2179 _axes = _flatten(axes)
```

```
2180
```

```
~\Anaconda3\lib\site-packages\pandas\plotting\_tools.py in _subplots(naxes, sharex, sharey, squeeze,
```

```
subplot_kw, ax, layout, layout_type, **fig_kw)
```

```
235
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

236 # Create first subplot separately, so we can share it if requested
-> 237 ax0 = fig.add_subplot(nrows, ncols, 1, **subplot_kw)
238
239 if sharex:
    ~\Anaconda3\lib\site-packages\matplotlib\figure.py in add_subplot(self, *args, **kwargs)
1072 self._axstack.remove(ax)
1073
-> 1074 a = subplot_class_factory(projection_class)(self, *args, **kwargs)
1075
1076 self._axstack.add(key, a)

~\Anaconda3\lib\site-packages\matplotlib\axes\_subplots.py in __init__(self, fig, *args, **kwargs)
62 raise ValueError(
63 "num must be 1 <= num 64 maxn=rows*cols, num=")
65 self._subplotspec = GridSpec(rows, cols)[int(num) -]
66 # num - 1 for converting from MATLAB to python in

```

ValueError: num must be 1 <= num <= 0, not 1



**Jason Brownlee** July 18, 2018 at 6:34 am #

You can get the dataset here as well:

<https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv>



## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↩



**Amulya** July 18, 2018 at 12:27 am #

Can we access two .pb files in a single model?

Thanks in advance.

REPLY ↩



**Jason Brownlee** July 18, 2018 at 6:36 am #

What is a .pb file?

REPLY ↩



**AMIRUL** July 18, 2018 at 4:46 pm #

sir i got this error

File “C:\Users\Amirul\Anaconda3\lib\urllib\request.py”, line 1320, in do\_open  
raise URLError(err)

URLError:

please help me



**Jason Brownlee** July 19, 2018 at 7:47 am #

REPLY ↗

I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Kiran** July 18, 2018 at 11:15 pm #

REPLY ↗

I installed everything and am trying to print th



**Jason Brownlee** July 19, 2018 at 7:54 am #

X

I have some ideas here:

<https://machinelearningmastery.com/faq/single-fa>



**Rajat** July 20, 2018 at 1:50 pm #

REPLY ↗

Hi

my data set contains 143 colomns, so I change the X Y values for new array. Good.

But in the for loop

my code is breaking at cv\_results line. How do I overcome it?

Pls help, thanks!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 21, 2018 at 6:29 am #

REPLY ↗

I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Adoh** July 21, 2018 at 5:23 pm #

REPLY ↗

What an awesome! Really easy-to-follow tutorial!

Thanks for advices you gave along the tutorial!

## Your Start in Machine Learning



**Jason Brownlee** July 22, 2018 at 6:21 am #

REPLY ↗

Thanks, I'm glad it helped.



**H.G. Lison** July 23, 2018 at 12:54 am #

REPLY ↗

Dear Dr. Brownlee,

You are a true hero, someone who gives their time and energy to helping others.

Bravo!!!

H.G. Lison



**Jason Brownlee** July 23, 2018 at 6:13 am #

I'm glad it helped.



**Ken** July 23, 2018 at 1:42 am #

I really like that you solved the same problem future modeling of real-world problems because it shows me that I can easily compare results in my particular case to pick the best model. I understand that some of them may give dramatically better results depending on the problem and training/validation data. Thanks for sharing this! I'm looking forward to reading more of your posts.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 23, 2018 at 6:14 am #

REPLY ↗

Thanks Ken, I'm glad it helped.



**Navid Akbari** July 24, 2018 at 4:16 am #

REPLY ↗

Hi Jason,

thanks for your tutorial. Really helpful. I am a complete beginner. I am seeing two errors checking for the right models. First is an indentationError (couldn't fix it by deleting spaces). Second is NameError: name 'model' is not defined

Please assist. Thanks!

Your Start in Machine Learning

```

>>> # Spot Check Algorithms
... models = []
>>> models.append('LR', LogisticRegression())
>>> models.append('LDA', LinearDiscriminantAnalysis())
>>> models.append('KNN', KNeighborsClassifier())
>>> models.append('CART', DecisionTreeClassifier())
>>> models.append('NB', GaussianNB())
>>> models.append('SVM', SVC())
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
... kfold = model_selection.KFold(n_splits=10, random_
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_st
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(mo
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> results.append(cv_results)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'cv_results' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> print(msg)

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)


**Jason Brownlee** July 24, 2018 at 6:22 am #

REPLY ↗

Be sure to copy all of the code, here's some help on how:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-copy-code-from-a-tutorial>



**Oliver** July 25, 2018 at 12:47 am #

REPLY ↗

Hi Jason,

Your Start in Machine Learning

Very helpful introduction. Thanks for that!  
 I'm wondering how I could get the equation for example of the logistic regression.  
 Could you please guide me in the right direction?



**Jason Brownlee** July 25, 2018 at 6:21 am #

REPLY ↗

This might help:

<https://machinelearningmastery.com/implement-logistic-regression-stochastic-gradient-descent-scratch-python/>



**Purnima** July 28, 2018 at 4:03 pm #

Hi Jason

i have a question about algorithm comparison figure, v  
 also i used the same code but i not getting that dotted



**Jason Brownlee** July 29, 2018 at 6:08 am #

They are box and whisker plots, you can l

[https://en.wikipedia.org/wiki/Box\\_plot](https://en.wikipedia.org/wiki/Box_plot)

They may be solid lines in the latest version of matplotlib.

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Paul Burkart** July 30, 2018 at 6:02 am #

REPLY ↗

Support Vector Machines seems to be a better option for this particular problem. Sorry for any  
 formatting issues that may occur.

```
1 svm = SVC()
2 svm.fit(X_train, y_train)
3 predictions = svm.predict(X_validation)
4 print(accuracy_score(y_validation, predictions))
5 print(confusion_matrix(y_validation, predictions))
6 print(classification_report(y_validation, predictions))
```

Output:

```
1 Accuracy Score: 0.9333333333333333
2
3 Confusion Matrix:
4 [[ 7  0  0]
5 [ 0 10  2]
6 [ 0  0 11]]
7
8 Classification Report
```

Your Start in Machine Learning

		precision	recall	f1-score	support
12	Iris-setosa	1.00	1.00	1.00	7
13	Iris-versicolor	1.00	0.83	0.91	12
14	Iris-virginica	0.85	1.00	0.92	11
15					
16	avg / total	0.94	0.93	0.93	30



**Jason Brownlee** July 30, 2018 at 6:09 am #

REPLY ↗

Nice work!



**Renata** July 31, 2018 at 6:54 am #

Is there a way of printing the p-value  
msg = "%s: %f (%f)" % (name, cv\_results.me



**Jason Brownlee** July 31, 2018 at 2:14 pm #

I explain how to calculate p-value  
<https://machinelearningmastery.com/parametric-statistical-significance-tests-in-python/>



**vishal** August 1, 2018 at 4:39 am #

REPLY ↗

# Spot Check Algorithms

```
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
print(msg)
ValueError Traceback (most recent call last)
in ()
11 for name, model in models:
12 kfold = model_selection.KFold(n_splits=10, random_state=seed)
-> 13 cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
14 results.append(cv_results)
15 names.append(name)
ValueError: Unknown label type: 'unknown'

```



**Jason Brownlee** August 1, 2018 at 7:49 am #

Ensure you copy the code exactly and press **Run**.  
<https://machinelearningmastery.com/faq/single-faq/>



**Stepan** August 2, 2018 at 4:01 am #

Hello Jason, do you have any articles on your blog about machine learning? Could you share a link on it?

Kind regards!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Jason Brownlee** August 2, 2018 at 6:02 am #

REPLY ↗

Here's one:

<https://machinelearningmastery.com/avoid-overfitting-by-early-stopping-with-xgboost-in-python/>



**WallWall** August 8, 2018 at 9:43 pm #

REPLY ↗

Hello Jason,

I use LDA to predict and the result seems to better than SVC:

0.966666666667

[[ 7 0 0]

[ 0 1 1 1]

[ 0 0 1 1]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 1.00 0.92 0.96 12

Your Start in Machine Learning

Iris-virginica 0.92 1.00 0.96 11

avg / total 0.97 0.97 0.97 30

even the estimated accuracy score is worse than SVC:

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)



**Jason Brownlee** August 9, 2018 at 7:39 am #

Well done!



**George** August 12, 2018 at 6:30 pm #

Dear Jason

Big thanks for your great posts!! You are contributing so much knowledge!!

2 questions please for you or anyone in the community

I 've been using WEKA and now I am also entering in the world of Python scikit.

WEKA gives you the option to include the p-value in the results, but it seems there is nothing around (or I completely missed it) in Python scikit..

Question 1:

– How can we also include the Statistical Significance (with p-value=0.05, for paired t-test ) in the above command line that gave this results list:

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.981667 (0.025000)

It is helpful to know the p-value of the result in order to confidently claim the difference between the accuracy performance of the compared algorithms/models we are comparing.

In other words, what do we have to do to also display in the list of the above results the p-value?

Question 2:

– What if we wanted to calculate the AUC ROC instead of the accuracy?

Should we switch the following

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```
seed = 7
scoring = 'accuracy'

into just

seed = 7
scoring = 'auc' . ?
```

Many thanks in advance and apologies to you and the rest of the community for my ignorance.

Best regards,  
George



**Jason Brownlee** August 13, 2018 at 6:16 am

You can calculate p-values in Python using  
<https://machinelearningmastery.com/parametric-statistical-hypothesis-testing/>



**kestas** August 15, 2018 at 12:19 am #

Hi Jason,

Thanks for this, how quickly could i see the output of t

```
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
...     kfold = model_selection.KFold(n_splits=10, random_state=seed)
...     cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
...     results.append(cv_results)
...     names.append(name)
...     msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
...     print(msg)
...
...
```

For me it stops here, no errors showing in the entire code.



## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Jason Brownlee** August 15, 2018 at 6:04 am #

REPLY ↗

Are you running from the command line?

Notebooks and IDEs can introduce problems.

## Your Start in Machine Learning

**Taz** August 15, 2018 at 12:25 am #

REPLY ↩

LR: 0.908333 (0.078617)

LDA: 0.975000 (0.038188)

KNN: 0.966667 (0.040825)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.975000 (0.038188)

**Jason Brownlee** August 15, 2018 at 6:04 am #

REPLY ↩

Well done.

**qausain** August 17, 2018 at 12:08 am #

Hello, the code what you have given in this website i got the same outcome offline.:)

**Jason Brownlee** August 17, 2018 at 6:30 am #

REPLY ↩

I don't understand, can you elaborate?

**qausain** August 28, 2018 at 1:43 am #

REPLY ↩

I tried this code and i have also tired it in my own way by using excel file as data base instead of url.... Hope you understood me.... Thank you

**Jason Brownlee** August 28, 2018 at 6:02 am #

REPLY ↩

Sorry, I cannot help you connecting to an excel file.

I recommend saving your data into CSV format before working with it.

**SB** August 26, 2018 at 2:37 am #

REPLY ↩

Thanks so much for the wonderful website and

Your Start in Machine Learning

If I understand this correctly, we have built a model that will look at the data and predict the type of flower based on sepal/petal length/width.

Quick question:

After we have our final model for the dataset, how can we see what variables (sepal/petal length/width) are the most significant for prediction?

Thanks again!



**Jason Brownlee** August 26, 2018 at 6:30 am #

REPLY ↗

Correct.

We often give up this insight (from statistics) in fav



**Shashank** August 27, 2018 at 7:57 am #

The great post ...quickly building the confide



**Jason Brownlee** August 27, 2018 at 1:56 pm

Thanks!

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Tom** August 28, 2018 at 8:30 pm #

REPLY ↗

Hi. I'm trying to use this with a csv with two cols (date, price) but get the error: "could not convert string to float: '2014-12-31'".

Could anyone tell me what I'm doing wrong please?



**Jason Brownlee** August 29, 2018 at 8:11 am #

REPLY ↗

You can get started with time series problems here:

<https://machinelearningmastery.com/start-here/#timeseries>



**Sudeshna** August 30, 2018 at 10:49 pm #

REPLY ↗

`cv_results = model_selection.cross_val_score`

Your Start in Machine Learning

Here “model\_selection.cross\_val\_score” calculates the score based on the training data. But score/accuracy are calculated for the model with respect to validation data. This gives the performance of the model. But herein you have used this method prior to using the validation data. Could you please explain the logic behind. I am new to Machine learning and have gone through the algorithms also. So have come up with this question. Please help!



**Jason Brownlee** August 31, 2018 at 8:13 am #

REPLY ↗

You can learn more about validation sets here:

<https://machinelearningmastery.com/difference-test-validation-datasets/>



**Sudeshna** September 15, 2018 at 12:40 am #

Hello Jason,

I went through the link you shared. And also the link you provided in your previous post:  
<https://machinelearningmastery.com/evaluate-machine-learning-algorithms-cross-validation-resampling/>

Please confirm me if my understanding is correct.

Estimates of performance for our machine learning algorithm using “Cross Validation” is done by the following way :

First the original training data set is split into training data and test validation data. Then this derived training set is again split into n-number of folds using KFold(). Now with n-1 number of folds(sets of data), algorithm under consideration is trained. Then with the n-th fold(set) of data, algorithm is tested and the accuracy/ score is calculated between {the result obtained with this test data set} and the result obtained for each of {n-1 folds of training data set}. So we obtain n-1 counts of accuracy values for these n-1 folds of data. Finally the mean of this is calculated which gives the net accuracy of the algorithm used.

Please confirm me if my understanding is correct or not.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 15, 2018 at 6:11 am #

REPLY ↗

Sounds good. Except we get k accuracy scores, not k-1.



**Sudeshna** September 27, 2018 at 12:54 am #

Thanks a lot Jason!

## Your Start in Machine Learning



**Elizabeth Keleshian** September 4, 2018 at 11:23 am #

REPLY ↗

You may have answered this question before, so please excuse the possible repetitiveness:  
As you were exploring the relationships between the features, you noticed some correlations/patterns. Did that allow you to narrow down your choices of algorithms? If so, how?

My overall question: when do you know you can really leverage on the correlative relationships and/or gaussian representations when choosing a model? Is it true that sometimes it's too expensive (and hence not preferred in the workplace) to run and test six different algorithms when the data can get really big?



**Jason Brownlee** September 4, 2018 at 1:51 pm #

DEEPIKA ↗

Yes, if the data looks gaussian I think about correlation, I think about feature selection methods.

A good starting point is to test many methods and Often these intuitions breakdown in the face of rig



**Yadesh** September 5, 2018 at 1:55 am #

Why do we have included the LABEL column  
 $X = \text{array}[:,0:3]$  instead  $X = \text{array}[:,0:4]$

f ↗

X

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

↗



**Jason Brownlee** September 5, 2018 at 6:42 am #

REPLY ↗

No, the label is never included in the input.

↗



**Nickmachine** September 6, 2018 at 12:12 am #

REPLY ↗

Hello my friend.Nice tutorial.I am a little rookie in machine learning and i am struggling to complete the tutorial with this dataset: <http://archive.ics.uci.edu/ml/datasets/Wine>.

Can you please help me?It is important for me to understand how it works.

Thank you very much for your time and the tutorial.



**Jason Brownlee** September 6, 2018 at 5:39 am #

REPLY ↗

## Your Start in Machine Learning

This process will help you work through your dataset:

<https://machinelearningmastery.com/start-here/#process>

These tutorials will show you how to use the process with Python:

<https://machinelearningmastery.com/start-here/#python>



**Nick s** September 7, 2018 at 8:33 pm #

REPLY ↗

Very nice introduction to get some hands on experience, thanks!



**Jason Brownlee** September 8, 2018 at 6:04 am #

X

I happy you found it useful Nick!



**shamsah** September 8, 2018 at 6:06 am #

REPLY ↗

thanks for useful lessons

in my code the SVM achieved the best accuracy so I v  
when I am trying to change the code of prediction from  
can you help please

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 8, 2018 at 6:17 am #

REPLY ↗

What problem are you having exactly with this change?



**dhanadhawan** September 10, 2018 at 3:37 am #

REPLY ↗

how these datasets help to predict?



**Jason Brownlee** September 10, 2018 at 6:23 am #

REPLY ↗

What do you mean exactly?

## Your Start in Machine Learning



**Vipin Chauhan** September 11, 2018 at 5:09 pm #

REPLY ↗

A Very good course for beginners to get a feel of how thing really work in ML and how algo can be applied on data. I think this is the best way to start ML journey for anyone. Later on you can build deep understanding and expertise in python as well as ML Algos. Great work! Jason!.



**Jason Brownlee** September 12, 2018 at 8:10 am #

REPLY ↗

Thanks, I'm happy that it helped.



**Brittany** September 12, 2018 at 4:09 am #

This tutorial was superb – thank you!



**Jason Brownlee** September 12, 2018 at 8:15

Thanks, I'm happy that it helped.



**Dilip** September 12, 2018 at 10:23 pm #

REPLY ↗

Hi,

I'm getting this error when I execute the line  
`cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv = kfold, scoring = scoring_met)`

`ValueError: Found input variables with inconsistent numbers of samples: [120, 30]`

What am I doing wrong?



**Jason Brownlee** September 13, 2018 at 8:04 am #

REPLY ↗

I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Sai Prasad** September 14, 2018 at 6:47 pm #

REPLY ↗

Your Start in Machine Learning

LR: 0.966667 (0.040825)  
 LDA: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.983333 (0.033333)  
 NB: 0.975000 (0.053359)  
 SVM: 0.991667 (0.025000)

Above is what I ended up with. made minor modification to script before make prediction step on the validation set  
 knn = SVC()

Accuracy on the validation set was 90%.



**Jason Brownlee** September 15, 2018 at 6:03

Nice work!



**Saiprasad Josyula** September 14, 2018 at 6:49 pm

Thanks Jason. Great tutorials to get us on the wisdom. Hats off sir.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 15, 2018 at 6:03 am #

REPLY ↗

Thanks, I'm happy it helped!



**Dany** September 14, 2018 at 10:56 pm #

REPLY ↗

Hi Jason, great article you have there, it's simple and clear. Congrats.

I'm trying to use this concept to classify a data based on description (texts), but as I understood these functions that you use just accept numbers. DO you have any suggestions in how can I scalonate my texts?



**Jason Brownlee** September 15, 2018 at 6:09 am #

REPLY ↗

There are many ways to encode and represent text. This field is called natural language processing, you can get started here:

<https://machinelearningmastery.com/start-here/#nlp>

## Your Start in Machine Learning



**Yasmin Sajitha** September 15, 2018 at 12:53 am #

REPLY ↗

I am a newbie to ML and not a programmer. This tutorial explained to me all the steps in detail and was easy to understand. It gave me a new level of confidence which I didn't get after going through so many courses and theory. Thank you so much !



**Jason Brownlee** September 15, 2018 at 6:12 am #

REPLY ↗

Thanks, I'm happy to hear that!



**Matheus** September 15, 2018 at 2:54 am #

Good afternoon teacher, after you have finished what you said above, not all the steps of a machine learning project, but just the first few steps, having done all these tests and validated the model, how do we move forward with the remaining data?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** September 15, 2018 at 6:19 pm #

Perhaps this process will help:

<https://machinelearningmastery.com/start-here/#process>



**Chidi** September 17, 2018 at 12:04 pm #

REPLY ↗

I work through the project. I had to type most of the codes to help me understand the what each function and object meant and it was very intellectual. Thanks. Appreciate!



**Jason Brownlee** September 17, 2018 at 2:07 pm #

REPLY ↗

Well done!



**jens holm** September 17, 2018 at 9:36 pm #

REPLY ↗

i just found this and i am truly impressed. i was about to write something like this, but instead i will just link to yours! problem solved. well done on breaking

Your Start in Machine Learning

charm.



**Jason Brownlee** September 18, 2018 at 6:14 am #

REPLY ↗

Thanks, I'm happy it helped!



**Rajani** September 20, 2018 at 8:43 am #

REPLY ↗

Hi. I have a doubt regarding the seed value.

How to choose seed value? Is this value really affect the results?

Thank you in advance



**Jason Brownlee** September 20, 2018 at 2:27 pm #

This is a common question that I answer in my free course:  
<https://machinelearningmastery.com/faq/single-faq/how-to-set-the-random-seed/>

X

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Ali** September 28, 2018 at 6:11 pm #

Hi Jason,

Thank you for this tutorial, it's very useful and helped me a lot. I was only wondering if I can graphically display the models that come from the algorithms? So for example when making a decision tree, that I actually show it on the screen.

Thanks in advance



**Jason Brownlee** September 29, 2018 at 6:33 am #

REPLY ↗

You may be able to, I don't have a tutorial on that topic sorry.

X



**Parwaz** October 1, 2018 at 2:47 am #

REPLY ↗

Hii..

Tys given for good tutorial ...

Your Start in Machine Learning

Problem how the download dataset on his work.

And give any simple project templet such as example. .

New dataset download and its how to use in python



**Jason Brownlee** October 1, 2018 at 6:27 am #

REPLY ↗

You can download the dataset here:

<https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv>



**Jendiwiw** October 1, 2018 at 4:41 am #

Hi Jason,

I really appreciate this tutorial. It makes machine learning examined 1-by-1 every syntax you used, and then, the model is SVC. After that, I was curious about the other other models and I compared each other. LDA gave a happen? Does this case depend on the value of validation step 1 until step 5.

Here are the results:

```
LR: 0.966667 (0.040825)
LDA: 0.975000 (0.038188)
KNN: 0.983333 (0.033333)
CART: 0.966667 (0.040825)
NB: 0.975000 (0.053359)
SVM: 0.991667 (0.025000)
```

Logistic Regression

0.8

```
[[ 7 0 0]
 [ 0 7 5]
 [ 0 1 10]]
```

precision recall f1-score support

```
Iris-setosa 1.00 1.00 1.00 7
Iris-versicolor 0.88 0.58 0.70 12
Iris-virginica 0.67 0.91 0.77 11
```

avg / total 0.83 0.80 0.80 30

Linear Discriminant Analysis

0.9666666666666667

```
[[ 7 0 0]
 [ 0 1 1]]
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

[ 0 0 11]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 1.00 0.92 0.96 12

Iris-virginica 0.92 1.00 0.96 11

avg / total 0.97 0.97 0.97 30

K-Neighbors Classifier

0.9

[[ 7 0 0]

[ 0 11 1]

[ 0 2 9]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 0.85 0.92 0.88 12

Iris-virginica 0.90 0.82 0.86 11

avg / total 0.90 0.90 0.90 30

Decision Tree Classifier

0.9

[[ 7 0 0]

[ 0 11 1]

[ 0 2 9]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 0.85 0.92 0.88 12

Iris-virginica 0.90 0.82 0.86 11

avg / total 0.90 0.90 0.90 30

Gaussian Naive-Bayes

0.8333333333333334

[[7 0 0]

[0 9 3]

[0 2 9]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 0.82 0.75 0.78 12

Iris-virginica 0.75 0.82 0.78 11

avg / total 0.84 0.83 0.83 30

Support Vector Machines

0.9333333333333333

[[ 7 0 0]

[ 0 10 2]

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

[ 0 0 11]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 1.00 0.83 0.91 12

Iris-virginica 0.85 1.00 0.92 11

avg / total 0.94 0.93 0.93 30



**Jason Brownlee** October 1, 2018 at 6:33 am #

REPLY ↗

Nice work!

The difference in the result could be real or statistical.

In order to discover if the difference is real, statistical

<https://machinelearningmastery.com/start-here/#statistical-significance>



**Ahmad Nashreen** October 3, 2018 at 3:52 pm #

Hi,

I'm wondering, is it possible to make confusion matrix

If it is possible, how? I've search, and used the paramete

instead of  $4 \times 4$  (the attribute has 4 categories). I wonder how it ended like that, and whether I had code it wrongfully. Can you help give me some tips or explain how does this happen.

Thanks.

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** October 3, 2018 at 4:17 pm #

REPLY ↗

If you are trying to predict a class with 65 levels, that is challenging.

You can create a confusion matrix of  $65 \times 65$ , but it will be very difficult to read.

Nevertheless, here's some code you can use:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>



**Martin** October 8, 2018 at 5:12 pm #

REPLY ↗

Nice work! Very helpful

## Your Start in Machine Learning



**Jason Brownlee** October 9, 2018 at 8:33 am #

REPLY ↗

Thanks. I'm happy it helped.



**Zishan** October 10, 2018 at 5:31 am #

REPLY ↗

Hello Jason How are you, your tutorial is so much effective to learn machine learning from scratch for all beginner like me. i have run your code successfully, but i faced problem during working on various data set csv file, like : "<https://www.kaggle.com/new-york-city/nyc-baby-names>". which contains various New York City baby names, including (mother's) ethnicity information when i run your code with this data set i got this error "ValueError: could not convert string to float" in the csv file column number to your irish data set column error, Please give me a solution, thanks in advance



**Jason Brownlee** October 10, 2018 at 6:18 am #

I expect the code will require some modification

I recommend that you follow this process:

<https://machinelearningmastery.com/start-here/#python>

Perhaps some of these tutorials will help:

<https://machinelearningmastery.com/start-here/#python>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Sandra** October 14, 2018 at 10:28 pm #

REPLY ↗

Hello Jason, I got all the results right. But I also got three warnings while building the models:  
C:\Python27\lib\site-packages\sklearn\linear\_model\logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning. FutureWarning)

C:\Python27\lib\site-packages\sklearn\linear\_model\logistic.py:459: FutureWarning: Default multi\_class will be changed to 'auto' in 0.22. Specify the multi\_class option to silence this warning. "this warning.", FutureWarning)

C:\Python27\lib\site-packages\sklearn\svm\base.py:196: FutureWarning: The default value of gamma will change from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoid this warning. "avoid this warning.", FutureWarning)

I did not change anything in the code. Can you please tell me what is the error?

**Jason Brownlee** October 15, 2018 at 7:27 am #

Your Start in Machine Learning



You can ignore the warning for now.



**KC Cheung** November 23, 2018 at 7:18 am #

REPLY ↗

```
import warnings
warnings.filterwarnings("ignore", category=FutureWarning)
```

Put it in the beginning of code



**Jason Brownlee** November 23, 2018 at 11:15 pm #

X

Nice tip.



**Hannan** October 17, 2018 at 11:54 am #

Hi Jason,

Thanks for your efforts, undoubtedly it was a good start. But it'd be really nice if you can please add little more (e.g. how they're calculating the error and how they're providing such information) and the steps involved in the process.

And last but not the least, would you please let us know which other tutorials should we follow afterwards? Please provide the links with priorities, one must follow in terms of diving a bit more into it but not yet intelligent enough in prioritising the guidelines /learning process. 😊

Thanks.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** October 17, 2018 at 2:28 pm #

REPLY ↗

A good place to start for more tutorials and their ordering is right here:

<https://machinelearningmastery.com/start-here/#python>



**Hannan** October 17, 2018 at 8:45 pm #

REPLY ↗

Thanks, I'll check now. 😊



**tim** October 19, 2018 at 2:02 am #

REPLY ↗

## Your Start in Machine Learning

Absolutely fantastic page... I'm just starting out with ML (with only fairly basic Python skills.. but a lot of programming background) but this is a great way to get going

My only suggestion would be to add a bit more text at the top to explain what we are trying to achieve with the flower data (sorry if I've missed it).

I think it's 'given the data.. predict what type of Iris each row (or subsequent rows) is'.. but.. I'm not 100% sure



**Jason Brownlee** October 19, 2018 at 6:09 am #

REPLY ↗

Thanks Tim.



**Fath U Min Ullah** October 26, 2018 at 1:40 pm #

hey!

Can we use it for any other image classification ? like e.g. features in this training like hog, sift or surf features etc  
thank you.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** October 26, 2018 at 2:42 pm #

REPLY ↗

Sure.



**Whitt** October 28, 2018 at 5:24 am #

REPLY ↗

Thank you very much for your thorough & helpful tutorial!



**Jason Brownlee** October 28, 2018 at 6:15 am #

REPLY ↗

I'm glad it helped.



**Flavin** October 29, 2018 at 3:36 am #

REPLY ↗

Hi Jaison,

This tutorial was very useful for a beginner like me. I ha

Your Start in Machine Learning

1. How to save the trained model to some other file and use it for prediction, so that I need not run this entire code every time I want to do prediction for an input data?
2. How to visualize the training function on any plot of the data set after training? i.e., the curves separating the regions for the 3 classes we are having, on the data set plot.



**Jason Brownlee** October 29, 2018 at 6:01 am #

REPLY ↗

This post shows how to save a trained final model:

<https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/>

I think you mean: how to plot the decision surface?

Sorry, I don't have an example of this, it's more of



**Terefe Feyisa** November 2, 2018 at 11:05 pm #

I am very new to ML. I thought the field of ML oriented-step-by-step approach, I kind of like it. Many

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 3, 2018 at 7:07 am #

Thanks, I'm happy the tutorials are helping!

REPLY ↗



**sravanthi padavala** November 3, 2018 at 4:31 am #

iam getting an error saying pandas not defined in loading the data step.please help me out.



**Jason Brownlee** November 3, 2018 at 7:10 am #

REPLY ↗

Sounds like you need to install Pandas.

Perhaps this tutorial will help:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>

REPLY ↗



**sravanthi padavala** November 3, 2018 at 4:03 pm #

## Your Start in Machine Learning

Thank you. I had to write import statement in the code.

I got it now.

I am getting an error called name error that dataset is not defined in 5.1



**Jason Brownlee** November 4, 2018 at 6:24 am #

REPLY ↗

I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**oded** November 4, 2018 at 5:02 pm #

X

hi. thanks for the great tutorial!

one thing i don't understand though- in section 5.3 you

"We get an idea from the plots that some of the classes  
so we are expecting generally good results."

could you please elaborate a little bit about that? it seems  
parameters combinations are very heterogeneous in re

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 5, 2018 at 6:10 am #

REPLY ↗

I am suggesting that if the classes look linearly separable, that most models will find a way to separate them.



**Muhammad Zaka Ud Din** November 8, 2018 at 7:20 pm #

REPLY ↗

I am applying on my dataset that raises accuracy of about 92% in matlab apps, but here I am trying on both nn and on the examples above, my accuracy is not increasing then that of 40%...



**Jason Brownlee** November 9, 2018 at 5:20 am #

REPLY ↗

I have some suggestions for improving neural network performance here:

<http://machinelearningmastery.com/improve-deep-learning-performance/>



**Rabia** November 9, 2018 at 1:24 pm #

REPLY ↗

Hi Jason!

Your Start in Machine Learning

It's really helpful. Can you suggest me how to plot the classified samples to show visual classification to a lay man. That see how was the original data and how it is after classifying?

Thanks.



**Jason Brownlee** November 9, 2018 at 2:03 pm #

REPLY ↗

I don't understand, how would this plot look exactly?



**Cipher** November 10, 2018 at 11:24 pm #

Hi Jason,

Thank you so much for these perfect tutorials. I however have no experience with machine learning analysis, and as I am a beginner in this field I am not sure where to start here which makes the search for the answer relatively difficult. I have seen your post on how to answer the question on one of the page of the website.

I have a dataset made of objects belonging to either class A or B. I want to determine for each object its class. And this work perfectly with your code (using logistic regression and KNN algorithms). However, I 'overfed' on purpose my prediction models, while usually only a third of this N objects are correctly classified (when classifying these objects by hand, I get 100% accuracy).

I believe – but perhaps I am wrong here – that the ML models will weight each of the input parameters in term of relevance, and I would like now to access to these weights and I want to see if the classification is only made using the parameters known to be relevant or if another parameters left usually aside is also of importance for the classification.

So is there a way to extract the weight of each parameters as set by the prediction model?

Best regards,

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 11, 2018 at 6:07 am #

REPLY ↗

An algorithm may or may not make the "weight" of each input available to you.

Instead, you can use methods designed to report the relevance or importance of each input variable. Some of these methods are called feature selection methods and others are called feature importance methods. You can get started here:

<http://machinelearningmastery.com/an-introduction-to-feature-selection/>

**Noor** November 11, 2018 at 9:31 pm #

Your Start in Machine Learning



what about the audio dataset?



**Jason Brownlee** November 12, 2018 at 5:38 am #

REPLY ↗

I hope to cover audio data in the future.



**Li Yuan** November 12, 2018 at 2:35 am #

REPLY ↗

Here is another algorithm called Self-Organizing Maps apply on IRIS dataset and works very well. The source code and demo have been posted on GitHub, free to enjoy it.



**Jason Brownlee** November 12, 2018 at 5:39 am #

Thanks for sharing.



**john** November 12, 2018 at 5:52 am #

I have a question about how to find which algorithm is the best. Although it is a very basic question, I need it to know? In your example

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 12, 2018 at 6:03 am #

REPLY ↗

Good question, I answer it here:

<https://machinelearningmastery.com/faq/single-faq/what-algorithm-config-should-i-use>



**David Hull** November 15, 2018 at 10:51 am #

REPLY ↗

I simply have to say, the number of errors following your trail is truly frustrating.  
-Dave.



**Jason Brownlee** November 15, 2018 at 11:30 am #

REPLY ↗

What do you mean exactly Dave? Typos?

Your Start in Machine Learning



**jack** November 15, 2018 at 6:56 pm #

REPLY ↗

Hello Jason,

thank you very much for your input. The logistic regression is binary 1 and 0 .How can it determine 4 types of IRIS.Thank you very much



**Jason Brownlee** November 16, 2018 at 6:13 am #

REPLY ↗

Good question. It can be used in a one vs



**Ronakkumar Ashokbhai Modi** November 19,

Hii,

when i am going to install scipy library with python 3.4 registry”.

But i already install python 3.4.So,give me proper solu

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 20, 2018 at 6:32 am #

REPLY ↗

Perhaps use Python 3.5 or 3.6?



**Waseem Ahmed** November 20, 2018 at 12:36 am #

REPLY ↗

Thanks a lot, Jason. you'll easy-to-understand tutorial gave me a very very quick intro to ML using Python. And it also pointed me to the advanced use of ML algorithms. Speeded up my work considerably. Thanks a lot!!!



**Jason Brownlee** November 20, 2018 at 6:37 am #

REPLY ↗

Thanks, I'm glad it helped.



**Jimi** November 20, 2018 at 10:46 am #

REPLY ↗

Hi Jason

## Your Start in Machine Learning

I tried like what you said but none of them was more 40% accuracy! In addition how can I do regression to find misclassified?

Thanks



**Jason Brownlee** November 20, 2018 at 2:04 pm #

REPLY ↗

I don't follow sorry, how do you want to use regression for classification exactly?



**Roman Parajuli** November 24, 2018 at 4:05 am #

Great !! This was the first model I trained myself. Great idea of yours to create a walkthrough



**Jason Brownlee** November 24, 2018 at 6:35 am #

Thanks, well done!



**Anicetus Odo** November 24, 2018 at 8:50 pm #

Thanks Jason.

I followed your step-by-step implementation in the tutorial and got similar results and I found it very helpful.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)



**Jason Brownlee** November 25, 2018 at 6:52 am #

REPLY ↗

Well done!



**Ashish** November 25, 2018 at 2:22 am #

REPLY ↗

sir i just want to know after writing this code spyder where we have to run this code for see its working.



**Jason Brownlee** November 25, 2018 at 6:58 am #

REPLY ↗

Your Start in Machine Learning

I recommend saving the code to a text file and running from the command line.

I show how here:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-run-a-script-from-the-command-line>



**Sunil** November 25, 2018 at 4:12 pm #

REPLY ↗

Hi Jason,

Thanks for this tutorial, please see the results that i had that were similar to yours, but in my case, the boxplot for the Algorithm Comparison did not have the blue dotted lines that you had for KNN, NB and SVM. The code is the same as yours and hence i am p

LR: 0.966667 (0.040825)  
 LDA: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.975000 (0.038188)  
 NB: 0.975000 (0.053359)  
 SVM: 0.991667 (0.025000)

Thanks,  
 Sunil

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** November 26, 2018 at 6:15 am #

Well done!

Differences may be due to the stochastic nature of the algorithms:  
<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Abdallah Mohamed Hassan** December 5, 2018 at 8:56 am #

REPLY ↗

I just want to thank u for this efforts , iam new at the track and this tutorial took about 3 days from me to understand most things ;")

but it really helped me . it is a very good starting point . again thank u very much  
 God bless you



**Jason Brownlee** December 5, 2018 at 2:22 pm #

REPLY ↗

Well done for making it through!

Your Start in Machine Learning



**Tayyab** December 6, 2018 at 5:37 am #

REPLY ↗

Hi Jason Brownlee. I am following your tutorials from the last 2 months time to time and I am learning things quite in a nice manner. I have a question why is the result different for selecting the best model when I am printing the results in a separate for loop:

```
for count in range(len(names)):
    msg = "{0}: {1} ({2})".format(names[count], cv_results[count].mean(), cv_results[count].std())
    print(msg)
```

SVC: 1.0 (0.0)

LR: 0.9166666666666666 (0.0)

KNN: 1.0 (0.0)

CART: 0.8333333333333334 (0.0)

GNB: 1.0 (0.0)

LDA: 1.0 (0.0)

It seems like it rounds it but why not in the other ones?  
I would appreciate your response.

## Your Start in Machine Learning



You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Jason Brownlee** December 6, 2018 at 6:03 am #

Perhaps this will help:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Sruthissree R** December 11, 2018 at 2:04 pm #

REPLY ↗

## Your Start in Machine Learning

It has been specified that either theano or tensorflow will be required. pertaining to the fact that tensorflow is cumbersome to install in windows, I successfully installed theano. But installation and verification of keras requires tensorflow as it contains commands with tensorflow module. Trying to install tensorflow gave problems as told. How do I proceed with the setting up of the environment?



**Jason Brownlee** December 11, 2018 at 2:34 pm #

REPLY ↗

Keras can be configured to use Theano instead of TensorFlow:

<https://keras.io/backend/#switching-from-one-backend-to-another>



**mamina sahu** December 12, 2018 at 8:38 pm #

nice posts..



**Jason Brownlee** December 13, 2018 at 7:51 am #

Thanks.



**Arsalan** December 15, 2018 at 7:39 am #

REPLY ↗

I'm new in python.. What exactly we predict in this project with the help of different algorithms?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)



**Jason Brownlee** December 16, 2018 at 5:17 am #

REPLY ↗

You are learning how to predict the species of iris flower given measurements of the flowers.



**Cason Cherry** December 20, 2018 at 7:09 am #

REPLY ↗

Hey Jason – nice tutorial. I wanted to collect your thoughts (apologies if this was addressed earlier in the thread, but the thread is quite long). I've run this exercise in both Python and R, as I wanted to compare the algorithms in both languages, and I've noticed that the predictive power in R seems to be consistently higher on the test sets (see confusion matrix), even though overall accuracy is lower, with Linear Discriminant Analysis (LDA) consistently the most performant. In Python, the test sets seem to not be predicted as well (see confusion matrix) even though accuracy is generally higher and Support Vector Machines (SVM) consistently more performant in Python. What explains this difference? It surprised me

Your Start in Machine Learning

because I considered I might model something in R and then convert the code over to Python, but this somewhat alters those kinds of plans if the model would need to change in the process.

R

Accuracy

Min. 1st Q u. Median Mean 3rd Qu. Max. NA's

lda 0.9666667 0.9666667 0.9833333 0.9833333 1.0000000 1 0

cart 0.8666667 0.9416667 0.9666667 0.9533333 0.9666667 1 0

knn 0.9333333 0.9666667 0.9666667 0.9733333 0.9916667 1 0

svm 0.9333333 0.9666667 1.0000000 0.9833333 1.0000000 1 0

rf 0.9000000 0.9666667 0.9666667 0.9633333 0.9666667 1 0

Linear Discriminant Analysis

120 samples

4 predictor

3 classes: 'setosa', 'versicolor', 'virginica'

No pre-processing

Resampling: Repeated Train/Test Splits Estimated (10)

Summary of sample sizes: 90, 90, 90, 90, 90, 90, ...

Resampling results:

Accuracy Kappa

0.9833333 0.975

Confusion Matrix and Statistics

Reference

Prediction setosa versicolor virginica

setosa 10 0 0

versicolor 0 10 1

virginica 0 0 9

Overall Statistics

Accuracy : 0.9667

95% CI : (0.8278, 0.9992)

No Information Rate : 0.3333

P-Value [Acc > NIR] : 2.963e-13

Kappa : 0.95

Mcnemar's Test P-Value : NA

Python:

looping through each model and evaluating

LR: 0.983333 (0.033333)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.983333 (0.033333)

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

Support Vector Machine:

0.9333333333333333

[ [ 7 0 0 ]

[ 0 10 2 ]

[ 0 0 11 ] ]

precision recall f1-score support

setosa 1.00 1.00 1.00 7

versicolor 1.00 0.83 0.91 12

virginica 0.85 1.00 0.92 11

micro avg 0.93 0.93 0.93 30

macro avg 0.95 0.94 0.94 30

weighted avg 0.94 0.93 0.93 30



**Jason Brownlee**

December 20, 2018 at 1:56 pm #

Interesting.

It might be differences in a range of things, for example seeds, implementation details, etc.



**Brahim**

December 22, 2018 at 12:56 pm #

REPLY ↗

Hello,

results = []

names = []

for name, model in models:

kfold = model\_selection.KFold(n\_splits=10, random\_state=seed)

cv\_results = model\_selection.cross\_val\_score(model, X\_train, Y\_train, cv=kfold, scoring=scoring)

results.append(cv\_results)

names.append(name)

msg = "%s: %f (%f)" %(name, cv\_results.mean(), cv\_results().std())

print(msg)

I had this error, msg = "%s: %f (%f)" %(name, cv\_results.mean(), cv\_results().std())

TypeError: 'numpy.ndarray' object is not callable

what was it?

thanks

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** December 23, 2018 at 6:03 am #

REPLY ↗

Sorry to hear that, I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Anmol** December 28, 2018 at 5:47 pm #

REPLY ↗

sir can you help me to run the above code am getting confused to use any other application for it or in python IDLE it self



**Jason Brownlee** December 29, 2018 at 5:50 am #

I explain how to run code from the command line.

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Venkat** January 4, 2019 at 2:21 am #

I am getting a output showing the error message. Can you please clarify my doubt?

Traceback (most recent call last):

```
File "E:\Project\Implementation\sample.py", line 48, in 
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\model_selection\_validation.py", line 342, in cross_val_score
    pre_dispatch=pre_dispatch)
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\model_selection\_validation.py", line 206, in cross_validate
    for train, test in cv.split(X, y, groups))
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 779, in __call__
    while self.dispatch_one_batch(iterator):
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 625, in dispatch_one_batch
    self._dispatch(tasks)
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 588, in _dispatch
    job = self._backend.apply_async(batch, callback=cb)
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\externals\joblib\_parallel_backends.py", line 574, in apply_
    result = ImmediateResult(func)
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\externals\joblib\_parallel_backends.py", line 332, in __init__
    self.results = batch()
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 131, in __call__
    return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 131, in
    return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\model_selection\_validation.py", line 458, in _fit_and_score
    estimator.fit(X_train, y_train, **fit_params)
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\linear_model\logistic.py", line 1217,
    check_classification_targets(y)
File "C:\Users\user\AppData\Local\Programs\Python\Python36-32\lib\site-
packages\sklearn\utils\multiclass.py", line 172, in check_
    raise ValueError("Unknown label type: %r" % y_type)
ValueError: Unknown label type: 'unknown'
>>>

```



**Jason Brownlee** January 4, 2019 at 6:32 am #

Sorry to hear that, I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Cody Bradley** January 14, 2019 at 7:30 am #

REPLY ↗

After much failure, I was able to get this to work!

however I had to set the LR model as follows to prevent error due to getting a 'future warning error'

LogisticRegression(solver='lbfgs', multi\_class='auto', max\_iter=1000)

as well as:

SVC(gamma='auto')

my results were as follows:

LR: 0.983333 (0.033333)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.983333 (0.033333)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

-----

0.9333333333333333

[[ 7 0 0]

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

[ 0 10 2]

[ 0 0 11]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 1.00 0.83 0.91 12

Iris-virginica 0.85 1.00 0.92 11

micro avg 0.93 0.93 0.93 30

macro avg 0.95 0.94 0.94 30

weighted avg 0.94 0.93 0.93 30

I am a complete beginner with ML but this at least gave me a place to start. Do you think the changes I made to the parameters or the models could have changed the data to make it less accurate?

again thanks for this tutorial!



**Jason Brownlee** January 14, 2019 at 11:14 am #

Well done!

I believe they were just warnings, not errors. You can



**saint** January 16, 2019 at 7:52 pm #

very nice job done ,can you make on A.I



**Jason Brownlee** January 17, 2019 at 5:24 am #

REPLY ↗

AI is a large field of study and ML is a subfield of AI, more here:

<https://machinelearningmastery.com/faq/single-faq/how-are-ai-and-ml-related>



**AA** January 21, 2019 at 7:52 am #

REPLY ↗

Hey Guys – Need help.

import pandas errors out – raise ImportError('dateutil 2.5.0 is the minimum required version')

Forums talks about lowering version – are they referring to downgrade from version 2.7 of python?

then import sklearn fails – ImportError: No module named sklearn

I was able to install sklearn from this command sudo pip install -U scikit-learn scipy matplotlib  
my pip version is 9.0.1. Is that the problem?

Your Start in Machine Learning



**Jason Brownlee** January 21, 2019 at 12:01 pm #

REPLY ↗

I have not seen this error, perhaps try posting on stackoverflow?



**Ping Liu** January 25, 2019 at 12:53 pm #

REPLY ↗

Thank you for the instruction. I am learning how to use the method to do my project. I have a dataset with X and Y, X are all 5-min resolution data , Y has both 5-min and 30-min data. Now I need to forecast 30-min data and the probability, which way should I go?

1) aggregate all 5-min X data to 30-min X data by averaging the 5-min X data and 30-min Y data to do training and testing, in concern is I have some time sensitive X data. If I use 3 variability of X data as accurate as in 5-min resolution.  
2) use all 5-min X data and 5-min Y data to do training trained model, then average the 5-min Y data into 30-min probability for the 30-min Y data, the trained model can there any way to convert the probability from 5-min res

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** January 26, 2019 at 6:08 am #

The above tutorial won't be very useful if you

You can get started with time series forecasting here:

<https://machinelearningmastery.com/start-here/#timeseries>

I have advanced material here:

[https://machinelearningmastery.com/start-here/#deep\\_learning\\_time\\_series](https://machinelearningmastery.com/start-here/#deep_learning_time_series)



**Saddam** January 27, 2019 at 3:44 am #

REPLY ↗

Sir, you are too good. It took me just hours to learn the basics of machine learning on Python. Thank you so much.



**Jason Brownlee** January 27, 2019 at 7:41 am #

REPLY ↗

Well done!

**khalil** February 1, 2019 at 1:15 am #

Your Start in Machine Learning



Hello

Thanks for your good training.

I have a question from you.

I want to predict the probability value for every 0

That is, how much is it possible to convert from 0 to 1

what do I do

help me.

thanks a lot



**Jason Brownlee** February 1, 2019 at 5:40 am #

REPLY ↗

You can use `model.predict_proba()`

I explain more here:

<https://machinelearningmastery.com/make-predictions-machine-learning-model/>



**Susovan** February 1, 2019 at 1:33 am #

## Your Start in Machine Learning

X

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 1, 2019 at 5:41 am #

REPLY ↗

Yes, I have many examples, perhaps start here:

<https://machinelearningmastery.com/spot-check-regression-machine-learning-algorithms-python-scikit-learn/>



**Toufik** February 2, 2019 at 11:43 pm #

REPLY ↗

hello Jason i bought your book (deep learning with python ) it's very important. so my question is what's the best function activation used for multiclassification (Example IRIS) .



**Jason Brownlee** February 3, 2019 at 6:18 am #

REPLY ↗

The activation function in the output layer categorical cross entropy.

## Your Start in Machine Learning



**Toufik** February 4, 2019 at 1:31 am #

REPLY ↩

thank 's Jason



**pedro** February 3, 2019 at 2:59 am #

REPLY ↩

(base) C:\Users\pedro>python

Python 3.7.1 (default, Dec 10 2018, 22:54:23) [MSC v.1915 64 bit (AMD64)] :: Anaconda, Inc. on win32

Type "help", "copyright", "credits" or "license" for more information.

>>> #numpy

... import numpy

>>> print('numpy: %s' % numpy.\_\_version\_\_)

numpy: 1.15.4

>>> #matplotlib

... import matplotlib

>>> print('matplotlib: %s' % matplotlib.\_\_version\_\_)

matplotlib: 3.0.2

>>> #pandas

... import pandas

>>> print('pandas: %s' % pandas.\_\_version\_\_)

pandas: 0.23.4

>>> #statsmodels

... import statsmodels

>>> print('statsmodels: %s' % statsmodels.\_\_version\_\_)

statsmodels: 0.9.0

>>> #scikit\_learn

... import sklearn

>>> print('sklearn: %s' % sklearn.\_\_version\_\_)

sklearn: 0.20.1

>>>

## Your Start in Machine Learning

X

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 3, 2019 at 6:20 am #

REPLY ↩

Well done!



**Ayman Mikhail** February 4, 2019 at 2:22 pm #

REPLY ↩

No bugs. Got it to work in Ubuntu and Windows.

Your Start in Machine Learning



**Jason Brownlee** February 5, 2019 at 8:12 am #

REPLY ↗

Well done!



**JOSEPH WILLIAMS** February 5, 2019 at 8:23 am #

REPLY ↗

Great article.



**Jason Brownlee** February 5, 2019 at 8:30 am #

Thanks.



**Mamta** February 6, 2019 at 4:30 pm #

Thank you for the tutorial. Amazing work done. I followed the tutorial and got same cross validation score as you. I tried score for each of the models and got the result as follows:  
LR : 0.8

LDA : 0.9666666666666667

KNN : 0.9

CART : 0.9

NB : 0.8333333333333334

SVM : 0.9333333333333333

Based on the cross validation score if we select KNN but the prediction score of LDA is highest here. Why is that? Can you help me in drawing some conclusion here.

Thanks 😊

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 7, 2019 at 6:36 am #

REPLY ↗

You can expect some variability around the model evaluation, I explain more here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



**Fredrick Ughimi** February 10, 2019 at 10:32 am #

REPLY ↗

Hello Jason,

Your Start in Machine Learning

Thank you for the tutorials. Really amazing! It was really straight forward.

I didn't have to change a thing. What next after this.

My results are similar to yours.

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

Best regards.



**Jason Brownlee** February 11, 2019 at 7:53 am #

Well done!



**Luzuko** February 11, 2019 at 8:49 pm #

i am happy to say that i have used your some Algorithms to perfection.



## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 12, 2019 at 7:58 am #

REPLY ↩

Thanks, I'm glad it helped!



**red** February 13, 2019 at 7:10 pm #

REPLY ↩

how do you manage to fix the warning error? i also have that error in my different code.



**Jason Brownlee** February 14, 2019 at 8:42 am #

REPLY ↩

Perhaps ensure that your libraries are up to date?

What warnings?

**red** February 14, 2019 at 1:36 pm #

Your Start in Machine Learning



Multiple error like this

```
/home/user/.local/lib/python3.5/site-packages/sklearn/utils/validation.py:761:
DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please
change the shape of y to (n_samples, ), for example using ravel().
y = column_or_1d(y, warn=True)
/home/user/.local/lib/python3.5/site-packages/sklearn/utils/validation.py:761:
DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please
change the shape of y to (n_samples, ), for example using ravel().
y = column_or_1d(y, warn=True)
/home/user/.local/lib/python3.5/site-packages/sklearn/utils/validation.py:761:
DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please
change the shape of y to (n_samples, ), for exa...n's using ravel()
y = column_or_1d(y, warn=True)
/home/user/.local/lib/python3.5/site-packages/sklearn/utils/validation.py:761:
DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please
change the shape of y to (n_samples, ), for exa...n's using ravel()
y = column_or_1d(y, warn=True)
main.py:122: DataConversionWarning: A colur
expected. Please change the shape of y to (n_
knn.fit(X_train, Y_train)
KNN: 0.957953 (0.006179)
CART: 0.987552 (0.003800)
NB: 0.916668 (0.006903)
SVM: 0.658934 (0.055898)
LR: 1.000000 (0.000000)
LDA: 0.977768 (0.005342)
KNN: 0.957953 (0.006179)
CART: 0.988441 (0.004228)
NB: 0.916668 (0.006903)
SVM: 0.658934 (0.055898)
0.9649390243902439
[[973 35]
 [ 34 926]]
precision recall f1-score support
L 0.97 0.97 0.97 1008
W 0.96 0.96 0.96 960
micro avg 0.96 0.96 0.96 1968
macro avg 0.96 0.96 0.96 1968
weighted avg 0.96 0.96 0.96 1968
```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
 Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**red** February 14, 2019 at 1:42 pm #

although it print the desire output

REPLY ↗

Your Start in Machine Learning

libraries version

```
Python: 3.5.1 (default, Jul 5 2018, 13:06:10)
[GCC 5.4.0 20160609]
scipy: 1.2.1
numpy: 1.16.1
matplotlib: 3.0.2
pandas: 0.24.1
sklearn: 0.20.2
```



**Jason Brownlee** February 14, 2019 at 2:10 pm <#>

You can fix your errors by reading:

More on reshaping numpy arrays here:  
<https://machinelearningmastery.com/reshape-numpy-arrays-python/>



**Ziad** February 14, 2019 at 1:31 am <#>

Dear Jason,

Thanks for the useful and interesting materials.

I have a question please: you said in 5.4 Select Best Model: "In this case, we can see that KNN has the largest estimated accuracy score."

In fact LR has the lowest mean. do you mean low mean = high accuracy? but we could have high mean with high accuracy. Could you please make it clear? thank you.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

REPLY



**Jason Brownlee** February 14, 2019 at 8:48 am <#>

REPLY



**Ziad** February 14, 2019 at 7:18 pm <#>

Hi,

I guess SVN has the highest accuracy not KNN, or I am wrong.  
 please see the results:

```
LR: 0.966667 (0.040825)
LDA: 0.975000 (0.038188)
KNN: 0.983333 (0.033333)
CART: 0.975000 (0.038188)
```

Your Start in Machine Learning

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

Thanks



**Jason Brownlee** February 15, 2019 at 8:00 am #

REPLY ↩

Yes, I have updated the text accordingly. Thanks!



**rick** February 14, 2019 at 1:33 pm #

hello jason, how to you manage the warning errors? same error



**Jason Brownlee** February 14, 2019 at 2:17 pm #

I will have a post about how to fix warning errors.

Until then, I recommend reading the warning messages and looking up what they mean. It will help you learn how to fix the warnings.



**SK Pandey** February 14, 2019 at 8:15 pm #

REPLY ↩

How can we get the Model function which we have created in this section ? means structure of the model in the forms of variables



**Jason Brownlee** February 15, 2019 at 8:01 am #

REPLY ↩

We typically do not get the equation for machine learning models as it is often intractable.



**Naren** February 15, 2019 at 4:41 am #

REPLY ↩

Though you've mentioned my results may vary... from top till bottom, I got the exact same result as your screenshots... bang... Thanks for the article... though a longer path to go still, one step at a time. Thanks.

## Your Start in Machine Learning



**Jason Brownlee** February 15, 2019 at 8:16 am #

REPLY ↗

I'm glad to hear that!



**Renato** February 15, 2019 at 9:42 pm #

REPLY ↗

Hi Jason,

I got the same results, but I don't understand why you mention "K-Nearest Neighbors (KNN) has the largest estimated accuracy score." According to the list, SVM presents a higher score

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

why?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** February 16, 2019 at 6:18 am #

I describe SVM as getting the best result.



**Seaturtle** February 19, 2019 at 9:10 am #

REPLY ↗

Thank you, Jason. This is an excellent resource, as are your other posts.



**Jason Brownlee** February 19, 2019 at 1:58 pm #

REPLY ↗

Thanks, I'm glad it helped!



**Farru Khan** February 21, 2019 at 9:28 pm #

REPLY ↗

can we use two machine learning algorithm simultaneously like Clustering (K-means) with Naive Bayes?

## Your Start in Machine Learning



**Jason Brownlee** February 22, 2019 at 6:18 am #

REPLY ↗

Sure.



**Darek** February 23, 2019 at 1:16 am #

REPLY ↗

Can you please help me to understand. First you make standard test\_train\_spit and next you make cross validation. Shouldn't we do either this or that? You use cross validation only to select best model but you do predictions on initially created train,test datasets (80%,20%).



**Jason Brownlee** February 23, 2019 at 6:34 am #

We can overfit during cross validation more so help confirm the chosen model/models are skillful.

This is just a suggestion, you can model the problem



**Neha Kavatage** February 23, 2019 at 3:57 pm #

cannot import name 'cross\_validation' from 'sklearn.cross\_validation' (C:\...\ProgramData\Anaconda3\lib\site-packages\sklearn\\_\\_init\\_\\_.py)

I'm getting error for this line ...how can i fix this??



**Jason Brownlee** February 24, 2019 at 9:05 am #

REPLY ↗

You must ensure that your version of scikit-learn is up to date, e.g. 0.18 or higher.



**Wizytor** February 27, 2019 at 7:49 am #

REPLY ↗

Just to make sure. I was given a task: Use leave-one-out cross-validation to determine the correct model and report the results in terms of average performance across cross-validation samples.

First I split dataset to Train/Test samples.

Then I use leave one out cross val (on train sample) to determine best model.

After that I predict values using cross\_val\_score on test sample only or on whole dataset?

## Your Start in Machine Learning



**Jason Brownlee** February 27, 2019 at 2:36 pm #

REPLY ↗

That is one approach.

Instead, I would recommend split into train/test, use k-fold cv on train for model selection, then fit a final model on all train and evaluate on test to get an unbiased idea of how good the model might be. Then fit a new final model on all data and start using it to make predictions on real unseen data.

Does that help?



**Wizytor** February 27, 2019 at 5:38 pm #

REPLY ↗

X

Yes, thank you! It makes perfect sense. What (test, train, whole?)

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Larry** March 1, 2019 at 2:02 am #

REPLY ↗

Fantastic – thank you for the tutorial – got mine working first time – now reading back through it to understand more. Many Thanks Jason.



**Jason Brownlee** March 1, 2019 at 6:24 am #

REPLY ↗

Thanks, well done!



**zahida** March 2, 2019 at 12:17 pm #

REPLY ↗

X

Dear Jason,  
Thanks for the useful and interesting materials. But, how to handle the Outliers.  
Is there any best practices to do so? Should it be handle before we split the data?



**Jason Brownlee** March 3, 2019 at 7:56 am #

REPLY ↗

## Your Start in Machine Learning

Great question, a good place to start is here:

<https://machinelearningmastery.com/how-to-identify-outliers-in-your-data/>

Here is a very simple and effective method that you can use:

<https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/>



**Catherine** March 7, 2019 at 12:53 am #

REPLY ↗

Hello sir, I hope this meets you well. Thank you very much for this tutorial.

Right now, I'm trying to use this lesson to assist me in my own predictions.

I am using a lung cancer dataset that has attributes of lung cancer.

I've been getting some errors from the statistical summary.

Secondly, if I am able to successfully make predictions suggest, how do I implement this prediction in my web

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 7, 2019 at 6:54 am #

Perhaps some of these tutorials will help:

<https://machinelearningmastery.com/start-here/#predicting-lung-cancer>

I have some advice about developing a final model here:

<https://machinelearningmastery.com/train-final-machine-learning-model/>

And about putting it into production:

<http://machinelearningmastery.com/deploy-machine-learning-model-to-production/>



**Catherine** March 13, 2019 at 5:49 am #

REPLY ↗

Thank you, I tried it and It worked perfectly!

I read your post on how to save and load a model with sci-kit learn to make predictions but I don't quite get it.... After saving this model using pickle, how do I enter new inputs to get a prediction from this model??

Please I need clarification



**Olego** March 10, 2019 at 2:46 am #

REPLY ↗

## Your Start in Machine Learning

Hi! this is really awesome first project! and the blog as a whole is amazing and very useful!

Thanks a lot!

in the sklearn docs I found an option for ordinary KFold() function StratifiedKFold().

This is basically the same with only difference it returns stratified folds. The folds are made with preserving the percentage of samples for each class. I think this is especially useful with very unbalanced classes distribution



**Jason Brownlee** March 10, 2019 at 8:19 am #

REPLY ↗

Nice work, yes, it is a good idea to use the stratified version if the classes are imbalanced.



**yukti** March 11, 2019 at 4:57 pm #

hello the project is really helpful  
i wanted to know how to load the data from the stored  
and how to use something else rather than panda??



**yukti** March 11, 2019 at 5:00 pm #

as i am working with air quality data to ca  
predictions for the air quality please rply sir



**Jason Brownlee** March 12, 2019 at 6:46 am #

REPLY ↗

If you are working with time series, I recommend starting here:

<https://machinelearningmastery.com/start-here/#timeseries>



**Jason Brownlee** March 12, 2019 at 6:46 am #

REPLY ↗

This will help you:

<http://machinelearningmastery.com/load-machine-learning-data-python/>



**yukti** March 13, 2019 at 5:20 pm #

REPLY ↗

hey i tried doing things as you have suggested but the file that i have to fetch is something like this  
[https://github.com/yukti23/Data\\_Predictions/blob/master/airquality.csv](https://github.com/yukti23/Data_Predictions/blob/master/airquality.csv)

Your Start in Machine Learning

please help how to fetch this



**Jason Brownlee** March 14, 2019 at 9:18 am #

REPLY ↗

What problem are you having exactly?



**yukti** March 14, 2019 at 3:32 pm #

REPLY ↗

this is the error

```
File "", line 3
filename = 'test.csv' as csv file
 ^
SyntaxError: invalid syntax
```



**Jason Brownlee** March 15, 2019 at

I have some suggestions here:  
<https://machinelearningmastery.com/faq/solve-work-for-me>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
 Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**yukti** March 13, 2019 at 9:04 pm #

REPLY ↗

using yours dataset and implementing things the way you implemented that is working correctly  
 but further when i m implementing for my own dataset the error comes



**Jason Brownlee** March 14, 2019 at 9:22 am #

REPLY ↗

What errors?



**shrivaths** March 15, 2019 at 7:14 pm #

REPLY ↗

hi sir,

I am facing error in the step of “cv\_results = model\_selection.cross\_val\_score(model, X\_train, Y\_train, cv=kfold, scoring=scoring)”  
 will you please resolve.I am unable to understand this.

Your Start in Machine Learning

error named is :

```
C:\Users\HPPC\Anaconda3\lib\site-packages\sklearn\model_selection\_validation.py:542: FutureWarning:  
From version 0.22, errors during fit will result in a cross validation score of NaN by default. Use  
error_score='raise' if you want an exception raised or error_score=np.nan to adopt the behavior from  
version 0.22.  
FutureWarning)
```

-----  
ValueError Traceback (most recent call last)

in

12 for name, model in models:

13 kfold = model\_selection.KFold(n\_splits=10, random\_state=seed)

-> 14 cv\_results = model\_selection.cross\_val\_score(model, X\_train, Y\_train, cv=kfold, scoring=scoring)

15 results.append(cv\_results)

16 names.append(name)

~\Anaconda3\lib\site-packages\sklearn\model\_selection\cross\_val.py:238: FutureWarning:

groups, scoring, cv, n\_jobs, verbose, fit\_params, pre\_dispatch=pre\_dispatch,

400 fit\_params=fit\_params,

401 pre\_dispatch=pre\_dispatch,

-> 402 error\_score=error\_score)

403 return cv\_results['test\_score']

404

~\Anaconda3\lib\site-packages\sklearn\model\_selection\cross\_val.py:238: FutureWarning:

groups, scoring, cv, n\_jobs, verbose, fit\_params, pre\_dispatch=pre\_dispatch,

error\_score)

238 return\_times=True, return\_estimator=return\_estimator,

239 error\_score=error\_score)

-> 240 for train, test in cv.split(X, y, groups))

241

242 zipped\_scores = list(zip(\*scores))

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\parallel.py in \_\_call\_\_(self, iterable)

915 # remaining jobs.

916 self.\_iterating = False

-> 917 if self.dispatch\_one\_batch(iterator):

918 self.\_iterating = self.\_original\_iterator is not None

919

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\parallel.py in dispatch\_one\_batch(self, iterator)

757 return False

758 else:

-> 759 self.\_dispatch(tasks)

760 return True

761

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\parallel.py in \_dispatch(self, batch)

714 with self.\_lock:

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

**START MY EMAIL COURSE**

```

715 job_idx = len(self._jobs)
-> 716 job = self._backend.apply_async(batch, callback=cb)
717 # A job can complete so quickly than its callback is
718 # called before we get here, causing self._jobs to

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\_parallel_backends.py in apply_async(self, func,
callback)
180 def apply_async(self, func, callback=None):
181     """Schedule a func to be run"""
-> 182 result = ImmediateResult(func)
183 if callback:
184     callback(result)

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\_parallel_backends.py in ImmediateResult(self, func)
547 # Don't delay the application, to avoid keeping the
548 # arguments in memory
-> 549 self.results = batch()
550
551 def get(self):

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\_parallel_backends.py in batch(self)
223 with parallel_backend(self._backend, n_jobs=self._n_jobs):
224     return [func(*args, **kwargs)
-> 225     for func, args, kwargs in self.items]
226
227 def __len__(self):

~\Anaconda3\lib\site-packages\sklearn\externals\joblib\parallel.py in __len__(self)
223 with parallel_backend(self._backend, n_jobs=self._n_jobs):
224     return [func(*args, **kwargs)
-> 225     for func, args, kwargs in self.items]
226
227 def __len__():

~\Anaconda3\lib\site-packages\sklearn\model_selection\_validation.py in _fit_and_score(estimator, X, y,
scorer, train, test, verbose, parameters, fit_params, return_train_score, return_parameters,
return_n_test_samples, return_times, return_estimator, error_score)
526 estimator.fit(X_train, **fit_params)
527 else:
-> 528 estimator.fit(X_train, y_train, **fit_params)
529
530 except Exception as e:

~\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py in fit(self, X, y, sample_weight)
1284 X, y = check_X_y(X, y, accept_sparse='csr', dtype=_dtype, order="C",
1285 accept_large_sparse=solver != 'liblinear')
-> 1286 check_classification_targets(y)
1287 self.classes_ = np.unique(y)
1288 n_samples, n_features = X.shape

```

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

Your Start in Machine Learning

```

~\Anaconda3\lib\site-packages\sklearn\utils\multiclass.py in check_classification_targets(y)
169 if y_type not in ['binary', 'multiclass', 'multiclass-multioutput',
170 'multilabel-indicator', 'multilabel-sequences']:
-> 171 raise ValueError("Unknown label type: %r" % y_type)
172
173

```

ValueError: Unknown label type: 'continuous'



**Jason Brownlee** March 16, 2019 at 7:50 am #

REPLY ↗

I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/>



**Sherri** March 17, 2019 at 6:45 am #

Hi,

Great tutorial, every thing works fine until I actually try  
I get an error

line 79, in

```
cv_results = model.selection.cross_val_score(model, X,
```

AttributeError: 'LogisticRegression' object has no attribute 'selection'

|

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** March 18, 2019 at 6:00 am #

REPLY ↗

I think there is a typo in your code, perhaps double check the tutorial. e.g. `model.selection`  
should be `model_selection`.



**ZAK** March 25, 2019 at 9:07 am #

REPLY ↗

Hi thank you for this tutorial. Do you have any links dealing with the problem of missing values



**Jason Brownlee** March 25, 2019 at 2:15 pm #

REPLY ↗

## Your Start in Machine Learning

Yes, you can get started with missing data here:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-handle-missing-data>



**max\_s** March 28, 2019 at 7:11 am #

REPLY ↗

very nicely done, Jason! I used Jupyter notebook and had no issues replicating your findings using similar package versions. All the errors I encountered were my own typos.

a few questions:

1. SVM seems to have performed better; is there a reason you chose to show validation for KNN instead? (my validation of SVM shows 93% accuracy.)
2. Is the reason you call knn.fit() on the training data after appending results to the list?



**Jason Brownlee** March 28, 2019 at 8:25 am #

X

Well done!

Not really, just an example.

Fit will create an efficient representation of the train

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Alex** April 5, 2019 at 7:19 am #

REPLY ↗

Thanks so much Jason! This (along with your “How to Setup a Python Environment) were incredibly straightforward and easy to follow. The only minor confusion was that you need to run all the code within one file, but I was able to figure that out from the comments (might be worth noting up top though). I’ve never done a coding tutorial that worked so cleanly 😊

I am very excited to have just completed my first ML project.

Thank you!



**Jason Brownlee** April 5, 2019 at 1:57 pm #

REPLY ↗

Thanks, great suggestion Alex!

More on running a script from the command line here:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-run-a-script-from-the-command-line>

**Enzo** April 6, 2019 at 6:17 am #

## Your Start in Machine Learning



Very good tutorial Jason, thank you very much!

I'm trying to apply ML to a project using what I learned here, currently in the phase of reshaping my model training data and could use some help with a problem.

Currently, all the values of my attributes are either a negative integer or "Not available" and I want the model to be trained to take into account when an attribute value is "Not available" because for a same Class I have rows with a value on that attribute and rows with "Not available" in that attribute. You have any tips on how to go about that?



**Jason Brownlee** April 6, 2019 at 6:55 am #

REPLY ↗

Not available sounds like missing data.

This will help:

<https://machinelearningmastery.com/faq/single-faq/>



**yannick masua** April 9, 2019 at 12:00 am #

## Your Start in Machine Learning

×

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 9, 2019 at 6:27 am #

REPLY ↗

Sorry to hear that, perhaps this will help:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**uzair mushtaq** April 10, 2019 at 4:43 pm #

REPLY ↗

How to increase accuracy of predictive model.



**Jason Brownlee** April 11, 2019 at 6:30 am #

REPLY ↗

Great question, I have some suggestions here:

<http://machinelearningmastery.com/machine-learning-performance-improvement-cheat-sheet/>

## Your Start in Machine Learning



**ayush** April 12, 2019 at 2:42 am #

REPLY ↗

Build an application / web-page / mobile app which will perform the following tasks:

The program will take the following input: Weather (for example sunny, rainy etc), Season (e.g., summer, winter), Geographic Scene (e.g., hilly terrain, open field, crowded market etc) and other inputs which can be thought of by the students themselves. Given the input the program will generate a virtual reality scene. The generated virtual scene can be used for training ML algorithms to detect objects in varying environmental conditions.

can you give me suggestion in above problem??



**Jason Brownlee** April 12, 2019 at 7:51 am #

X

Perhaps talk to your teacher if you having

I believe a GAN would be required.



**its** April 16, 2019 at 6:32 am #

First ever example which worked without erro

Just want to add my +1

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 16, 2019 at 6:55 am #

REPLY ↗

Well done!



**Joe Feverati** April 18, 2019 at 6:01 pm #

REPLY ↗

Hi Jason,

thanks for your tutorial.

I don't understand why the predictions are not made with the model previously constructed models[2] but with a new fit. Would it be possible to use the previous one?



**Jason Brownlee** April 19, 2019 at 6:05 am #

REPLY ↗

Yes, you can make a prediction.

## Your Start in Machine Learning

Here's an example:

<https://machinelearningmastery.com/how-to-make-classification-and-regression-predictions-for-deep-learning-models-in-keras/>



**punch** April 18, 2019 at 11:15 pm #

REPLY ↗

Hi Jason,

i went to the tutorial.It is very helpful beginner. But i have a query regarding target variable how we will select class if it is not given in the data set.



**Jason Brownlee** April 19, 2019 at 6:10 am #

If you don't have a class, perhaps you wa

More here:

<https://machinelearningmastery.com/faq/single-faq/regression>



**LB** April 20, 2019 at 10:54 am #

Hey, I'm having problems with step 2.1 Import libraries. I have checked and my environment should be correct. it is printing out this code so far:

Python: 3.7.3 (default, Mar 27 2019, 16:54:48)

[Clang 4.0.1 (tags/RELEASE\_401/final)]

scipy: 1.2.1

numpy: 1.16.2

matplotlib: 3.0.3

pandas: 0.24.2

statsmodels: 0.9.0

sklearn: 0.20.3

theano: 1.0.3

tensorflow: 1.13.1

Using TensorFlow backend.

keras: 2.2.4

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 21, 2019 at 8:17 am #

REPLY ↗

Looks great, problem are you having exactly?

## Your Start in Machine Learning



**LB** April 24, 2019 at 3:32 am #

REPLY ↗

When I run the code:

```
# Load libraries
import pandas

from pandas.tools.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

in pycharm it turns grey and wont run
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 24, 2019 at 8:10 am #

REPLY ↗

I recommend running code from the command line:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-run-a-script-from-the-command-line>



**LB** April 24, 2019 at 7:59 am #

REPLY ↗

I can run everything up to the:

```
dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
plt.show()
```

Then the error i get is:

This application failed to start because it could not find or load the Qt platform plugin "cocoa" in "".

Reinstalling the application may fix this problem.



**Jason Brownlee** April 24, 2019 at 8:10 am #

REPLY ↗

## Your Start in Machine Learning

Perhaps try following this tutorial to setup your workstation:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Sara Kunwar** April 24, 2019 at 7:17 pm #

REPLY ↗

Hello Sir

Your information was so important for me for my project but sir i want a classified image as an output. Please tell me the solution for this.



**Jason Brownlee** April 25, 2019 at 8:10 am #

You can get started here:

<https://machinelearningmastery.com/start-here/#dl>



**Sayan Saha** April 29, 2019 at 8:21 pm #

Hi,

I got the result of print(msg) as

```
LR: 0.966667 (0.040825)
LDA: 0.975000 (0.038188)
KNN: 0.983333 (0.033333)
CART: 0.983333 (0.033333)
NB: 0.975000 (0.053359)
SVM: 0.991667 (0.025000)
```

Where KNN and CART has the same result. I followed your project step by step. Why is my answer different?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** April 30, 2019 at 6:54 am #

REPLY ↗

Well done!

Good question, I answer it here:

<https://machinelearningmastery.com/faq/single-faq/why-do-i-get-different-results-each-time-i-run-the-code>

## Your Start in Machine Learning



Thanks. I ran it again and got

LR: 0.966667 (0.040825)  
 LDA: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.975000 (0.038188)  
 NB: 0.975000 (0.053359)  
 SVM: 0.991667 (0.025000)



**Jason Brownlee** May 1, 2019 at 6:59 am #

REPLY ↩

Nice work.



**Qi Qi** May 3, 2019 at 11:49 pm #

```
# Load dataset
url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'species']
dataset = pandas.read_csv(url, names=names)

/Users/qiqi/PycharmProjects/ml/venv/bin/python /Users/qiqi/PycharmProjects/ml/ml53.py
Traceback (most recent call last):
File "/Users/qiqi/PycharmProjects/ml/ml53.py", line 5, in <module>
    dataset = pandas.read_csv(url, names=names)
NameError: name 'pandas' is not defined
```

Process finished with exit code 1

Excuse me, I met the following error. And pandas are not in the last step. Thank you very much!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

X



**Jason Brownlee** May 4, 2019 at 7:09 am #

REPLY ↩

It suggests that pandas is not installed.

You can follow this tutorial to setup your development environment:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**Qi Qi** May 6, 2019 at 4:30 am #

REPLY ↩

Yes, I am following the instruction.  
 As a matter of fact, in the last step it showed t

Your Start in Machine Learning

matplotlib: 3.0.3 pandas: 0.24.2 sklearn: 0.20.3.

I am curious about what is the problem. I saw someone met this question either but the answer does work for me. And I installed it on mac and am using Pycharm CE version. I will check it. Even if I used import pandas, it didn't work. Thank you very much!



**Jason Brownlee** May 6, 2019 at 6:51 am #

REPLY ↗

I recommend running code from the command line:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-run-a-script-from-the-command-line>



**Anj** May 5, 2019 at 2:44 am #

Hello Dr.Jason,

I am using Pycharm IDE and in this particualr line :

```
cv_results= model_selection.cross_val_score(model, x
C:\Users\Lenovo\PycharmProjects\Sample_Project\ve
packages\sklearn\model_selection\_validation.py:542:
will result in a cross validation score of NaN by default
raised or error_score=np.nan to adopt the behavior fro
FutureWarning)
```

Traceback (most recent call last):

```
File "C:/Users/Lenovo/PycharmProjects/Sample_Project/readingdatasets/Irisdataset.py", line 63, in
cv_results= model_selection.cross_val_score(model, x_train, y_train, cv=kfold, scoring=scoring)
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\model_selection\_validation.py", line 402, in cross_val_score
error_score=error_score)
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\model_selection\_validation.py", line 240, in cross_validate
for train, test in cv.split(X, y, groups))
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 917, in __call__
if self.dispatch_one_batch(iterator):
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 759, in dispatch_one_batch
self._dispatch(tasks)
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 716, in _dispatch
job = self._backend.apply_async(batch, callback=cb)
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\externals\joblib\_parallel_backends.
```

## Your Start in Machine Learning



You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```

result = ImmediateResult(func)
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\externals\joblib\_parallel_backends.py", line 549, in __init__
self.results = batch()
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 225, in __call__
for func, args, kwargs in self.items]
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\externals\joblib\parallel.py", line 225, in
for func, args, kwargs in self.items]
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\model_selection\_validation.py", line 528, in _fit_and_score
estimator.fit(X_train, y_train, **fit_params)
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\linear_model\logistic.py", line 1289,
check_classification_targets(y)
File "C:\Users\Lenovo\PycharmProjects\Sample_Project\venv\lib\site-
packages\sklearn\linear_model\logistic.py", line 171, in check_classification_targets
raise ValueError("Unknown label type: %r" % y_type)
ValueError: Unknown label type: 'unknown'

```

Please help here



**Jason Brownlee** May 5, 2019 at 6:33 am #

I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**roberto lupo** May 5, 2019 at 12:25 pm #

REPLY ↗

Hello Dr.Jason,

i use anaconda terminal on a windows 8.1 64 bit, python 3.7.3 64 bit  
when import scipy i get this error :

```

(base) C:\Users\roberto>python
Python 3.7.3 (default, Mar 27 2019, 17:13:21) [MSC v.1915 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import scipy
Traceback (most recent call last):
File "", line 1, in
File "C:\Users\roberto\Anaconda3\lib\site-packages\scipy\__init__.py", line 62, in
from numpy import show_config as show_numpy_config
File "C:\Users\roberto\AppData\Roaming\Python\Python37-32\site-packages\scipy\__init__.py", line 140, in
from . import core

```

Your Start in Machine Learning

File "C:\Users\roberto\AppData\Roaming\Python\Python37\site-packages\numpy\core\\_\_init\_\_.py", line 23,  
in

```
WinDLL(os.path.abspath(filename))
```

File "C:\Users\roberto\Anaconda3\lib\ctypes\\_\_init\_\_.py", line 356, in \_\_init\_\_  
self.\_handle = \_dlopen(self.\_name, mode)

OSError: [WinError 193] %1 non è un'applicazione di Win32 valida

>>>

but if i use python 3.7.3 32bit it's all ok and i get all results as on your tutorial,  
what's happens? and what i have do to use anaconda terminal 64bit ?

Thank you very much!

```
(base) C:\Users\roberto>anaconda3  
3.7.3 (v3.7.3:ef4ec6ed12, Mar 25 2019, 21:26:53) [MSC v.1916 64 bit (AMD64)]
```

```
(base) C:\Users\roberto>python packVersXml.py  
scipy: 1.2.1
```

numpy: 1.16.2

matplotlib: 3.0.3

pandas: 0.24.2

statsmodels: 0.9.0

sklearn: 0.20.3

(150, 5)

sepal-length sepal-width petal-length petal-width class:

0 5.1 3.5 1.4 0.2 Iris-setosa

1 4.9 3.0 1.4 0.2 Iris-setosa

2 4.7 3.2 1.3 0.2 Iris-setosa

3 4.6 3.1 1.5 0.2 Iris-setosa

4 5.0 3.6 1.4 0.2 Iris-setosa

5 5.4 3.9 1.7 0.4 Iris-setosa

6 4.6 3.4 1.4 0.3 Iris-setosa

7 5.0 3.4 1.5 0.2 Iris-setosa

8 4.4 2.9 1.4 0.2 Iris-setosa

9 4.9 3.1 1.5 0.1 Iris-setosa

10 5.4 3.7 1.5 0.2 Iris-setosa

11 4.8 3.4 1.6 0.2 Iris-setosa

12 4.8 3.0 1.4 0.1 Iris-setosa

13 4.3 3.0 1.1 0.1 Iris-setosa

14 5.8 4.0 1.2 0.2 Iris-setosa

15 5.7 4.4 1.5 0.4 Iris-setosa

16 5.4 3.9 1.3 0.4 Iris-setosa

17 5.1 3.5 1.4 0.3 Iris-setosa

18 5.7 3.8 1.7 0.3 Iris-setosa

19 5.1 3.8 1.5 0.3 Iris-setosa

sepal-length sepal-width petal-length petal-width

count 150.000000 150.000000 150.000000 150.000000

mean 5.843333 3.054000 3.758667 1.198667

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**

Your Start in Machine Learning

```
std 0.828066 0.433594 1.764420 0.763161
min 4.300000 2.000000 1.000000 0.100000
25% 5.100000 2.800000 1.600000 0.300000
50% 5.800000 3.000000 4.350000 1.300000
75% 6.400000 3.300000 5.100000 1.800000
max 7.900000 4.400000 6.900000 2.500000
class
Iris-setosa 50
Iris-versicolor 50
Iris-virginica 50
```



**Jason Brownlee** May 6, 2019 at 6:44 am #

I recommend saving the script in a .py file

See this:

<https://machinelearningmastery.com/faq/single-faq/>



**Qi Qi** May 6, 2019 at 8:20 am #

Hi, Jason,

When I walked the step 4 of plt.show()

NameError: name 'plt' is not defined.

Should I install plt or what's the potential error?

Thank you so much!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 6, 2019 at 2:32 pm #

REPLY ↗

Perhaps you skipped some lines of code?



**Anjali Muralidharan** May 6, 2019 at 6:01 pm #

REPLY ↗

Thank you, Dr.Jason , my code worked and got my output,  
Thanks for the help .

I just added one line line to my code ie.  
y = y.astype('int') and my code worked perfectly fine after that

## Your Start in Machine Learning



**Jason Brownlee** May 7, 2019 at 6:14 am #

REPLY ↗

Glad to hear it.



**p** May 7, 2019 at 8:14 am #

REPLY ↗

I don't understand how to see the visualizations portion. I'm getting an output of the numeric values but can't see the graphs.



**Jason Brownlee** May 7, 2019 at 2:27 pm #

Try running the code from the command line:

<https://machinelearningmastery.com/faq/single-faq/>



**sbkr** May 14, 2019 at 9:29 pm #

Does the DecisionTreeClassifier() do pruning?  
view the output hypothesis?

## Your Start in Machine Learning

X

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 15, 2019 at 8:14 am #

REPLY ↗

Yes it does, learn more here:

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>



**puja** May 15, 2019 at 2:27 pm #

REPLY ↗

After executing the code of validation dataset we are not getting the graph of Box and Whisker Plot Comparing Machine Learning Algorithms on the Iris Flowers Dataset....We are getting nameError:  
name 'model\_selection' is not defined...please give solution...



**Jason Brownlee** May 15, 2019 at 2:46 pm #

REPLY ↗

The error suggests you need to update your version of the sklearn library.

## Your Start in Machine Learning



**iuri prado** May 17, 2019 at 11:59 pm #

REPLY ↗

hello!

thank you for the tutorial. it was great to follow it along.

yes, i got the results in the end, indeed, but how to i input data to get a prediction for the trained model?



**Jason Brownlee** May 18, 2019 at 7:38 am #

REPLY ↗

You can use `model.predict()`.

I explain more here:

<https://machinelearningmastery.com/make-predictions-machine-learning-scikit-learn/>



**Shravani** May 20, 2019 at 1:02 am #

Hi Jason. Great tutorial. I have a small question.  
Under section “6. Make Predictions” you say “KNN algorithm based on our tests”. How did you come to this conclusion?

Previously, we established that SVM is most accurate here?

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 20, 2019 at 6:33 am #

REPLY ↗

You can choose any model you wish, I chose knn because it did well and is not complex.



**NR** May 23, 2019 at 6:26 am #

REPLY ↗

Hi Jason,

Thank you for this post 😊

I have a question.

Every time I run the ‘for’ loop of section 5.3. the mean accuracy score and standard deviation for the Decision Tree Classifier changes.

This is not observed for any other model, but only for the Decision Tree model.

What could be the reason for this?

(I understand that the other models’ scores remain sam

<https://www.udemy.com/machine-learning-in-python-step-by-step/>

Your Start in Machine Learning

Best Regards.



**Jason Brownlee** May 23, 2019 at 2:29 pm #

REPLY ↗

Good question, this is common, I explain more here:

<https://machinelearningmastery.com/faq/single-faq/why-do-i-get-different-results-each-time-i-run-the-code>



**NR** May 26, 2019 at 4:09 am #

DEDIV ↗

Thanks for the link, Jason!

I have some questions –

Does the seed value to the parameter ‘random’ function and the ‘KFold()’ function.

You have used 7 here for both. Is that just a co

Am I correct in understanding that the ‘seed’ value shuffling and uses the same data splits which

Also, what is the life of this state (random\_state)? Does it persist in memory or is this restricted to

Best Regards.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**NR** May 26, 2019 at 4:47 am #

REPLY ↗

Also, are we evaluating the algorithms with both mean and standard deviation? I understand that it is standard practice to include both as it gives you a correct idea of the variation in the data values. But in this case, does variation really matter?

If we add a 3rd column, “Coefficient of Variation”, should we deduce that the model with the least varied scores is the best performer or should we stick to the mean accuracy?

Best Regards.



**Jason Brownlee** May 26, 2019 at 6:51 am #

Ideally we would pick a model that best serves a project goals/stakeholders. This might be a model that is more stable (lower variance).

## Your Start in Machine Learning



**Jason Brownlee** May 26, 2019 at 6:50 am #

REPLY ↗

The random state is just for the session, the run.

In modern tutorials, I don't recommend fixing the random seed:

<https://machinelearningmastery.com/faq/single-faq/why-do-i-get-different-results-each-time-i-run-the-code>



**Kaustubh** May 29, 2019 at 10:36 pm #

REPLY ↗

Thank you very much for such an amazing tutorial!



**Jason Brownlee** May 30, 2019 at 9:00 am #

You're welcome, I'm glad it helped.



**Jerome** May 30, 2019 at 1:57 am #

Hi Jason,

For improving my results using feature selection, I am those features which have a relatively strong positive correlation with the target variable 'quality'. Should the variables which show strong negative correlation be excluded or included in this case? Can you explain more on how to use the correlation matrix to arrive at decisions related to feature selection? Thanks for this helpful post BTW!

- Jerome

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** May 30, 2019 at 9:03 am #

REPLY ↗

A strong positive or negative correlation may be useful.

This might help:

<https://machinelearningmastery.com/how-to-calculate-nonparametric-rank-correlation-in-python/>



**Jerome** June 7, 2019 at 5:16 am #

REPLY ↗

Hi Jason,

Thanks for providing the reference to the corre some basic questions –

Your Start in Machine Learning

Q.1. – How do I use negative correlation?

If you can provide your comments on how negative correlation can be useful in this particular example (wine dataset), it will help me draw analogies and work out other problems using similar understanding.

Q.2. – Is the call on which features to include/exclude initially made by looking at the correlation matrix values? What is the process you personally follow when you have features negatively correlated with your target variable?

Do we only look at the magnitude of correlation when making these decisions?

Thanks in advance Jason.



**Jason Brownlee** June 7, 2019 at 8:00 pm #

Sign does not matter.

A strong negative or positive correlation between inputs and outputs may be a sign of prediction difficulty.



**mohsen** May 31, 2019 at 11:33 pm #

thanks Dr. Jason



**Jason Brownlee** June 1, 2019 at 6:15 am #

REPLY ↩

You're welcome.



**teimoor** June 2, 2019 at 11:41 pm #

REPLY ↩

hi have you ever worked with ecg classification system in physionet? i have trouble loading the dataset to work with. should i load them in csv file?



**Jason Brownlee** June 3, 2019 at 6:42 am #

REPLY ↩

Sorry, I have not heard of "physionet".



**reuben** June 4, 2019 at 6:58 pm #

Your Start in Machine Learning

NameError Traceback (most recent call last)  
in  
2 # box and whisker plots  
3 dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)  
--> 4 plt.show()

NameError: name 'plt' is not defined

i face this problem in line 4.1 4.1 Univariate Plots

i have directly copied the code but unfortunately it keep showing this code.

Please help me out



**Jason Brownlee** June 5, 2019 at 8:35 am #

Looks like you might have missed the ma



**Jeswin Augustine** June 5, 2019 at 9:31 pm #

Hi Jason,

This tutorial was really helpful to get started. But when classifier/estimator for a project?

In real world use cases, I assume that, there might be large amount of data . So training a classifier will take large amount of time. So, is it possible to train multiple estimators and pick-out the best one as we did here, considering time and space complexity?

Or how is it done in real use cases with millions of data?



**Jason Brownlee** June 6, 2019 at 6:28 am #

REPLY ↗

Yes, test a suite of methods and select one that meets the objectives of the project (performance, complexity, etc.).

Often we want the simplest model (reliable) that preforms the best (skill).



**ZAK** June 12, 2019 at 9:56 am #

REPLY ↗

I tried it for the first time, it worked but for the second time when i run this :

# Spot Check Algorithms

models = []

Your Start in Machine Learning

```

models.append('LR', LogisticRegression())
models.append('LDA', LinearDiscriminantAnalysis())
models.append('KNN', KNeighborsClassifier())
models.append('CART', DecisionTreeClassifier())
models.append('NB', GaussianNB())
models.append('SVM', SVC())
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

```

I have this error

NameError Traceback (most recent call last)  
in

11 names = []

12 for name, model in models:

→ 13 kfold = cross\_validation.KFold(n=num\_instances,

14 cv\_results = cross\_validation.cross\_val\_score(mode

15 results.append(cv\_results)

NameError: name 'cross\_validation' is not defined

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 12, 2019 at 2:23 pm #

REPLY ↗

Looks like you might have forgotten the import statements?



**ZAK** June 12, 2019 at 7:57 pm #

REPLY ↗

No in the beginning i put this and i run it

```

import pandas
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

```

Your Start in Machine Learning

```
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
```



**AMJAD IQBAL** June 13, 2019 at 12:51 pm #

REPLY ↩

hi sir!

it's great to see such kind of post from you. I have app  
of result. sir i have some other dataset and the code is  
Your help will be highly appreciated  
waiting for your kind response



**Jason Brownlee** June 13, 2019 at 2:36 pm #

Sorry, I don't have tutorials in matlab, I ca



**neer** June 13, 2019 at 4:54 pm #

REPLY ↩

hi jason,

i tried a lot to solve indented block error....but I am stuck at it..pls help!



**Jason Brownlee** June 14, 2019 at 6:37 am #

REPLY ↩

This will show you how to copy and paste the code and preserve the indenting:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-copy-code-from-a-tutorial>



**neer** June 13, 2019 at 6:02 pm #

REPLY ↩

hi jason,

```
>>> # Spot Check Algorithms
```

```
... models = []
```

```
>>> models.append(('LR', LogisticRegression(solver='liblinear', C=1000)))
```

```
>>> models.append(('LDA', LinearDiscriminantAnalysis()))
```

Your Start in Machine Learning

```

>>> models.append('KNN', KNeighborsClassifier())
>>> models.append('CART', DecisionTreeClassifier())
>>> models.append('NB', GaussianNB())
>>> models.append('SVM', SVC(gamma='auto'))
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
... kfold = model_selection.KFold(n_splits=10, random_state=seed)
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_state=seed)
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(mo
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> results.append(cv_results)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'cv_results' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'msg' is not defined
NameError: name 'msg' is not defined
tried a lot to solve this ....but I am stuck.

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

**START MY EMAIL COURSE**



**Jason Brownlee** June 14, 2019 at 6:38 am #

REPLY ↗

This will show you how to safely copy code and preserve the indenting:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-copy-code-from-a-tutorial>



**neer** June 13, 2019 at 8:40 pm #

REPLY ↗

Your Start in Machine Learning

thanks a lot....i did it....!!!

precision recall f1-score support

	Iris-setosa	Iris-versicolor	Iris-virginica	
precision	1.00	1.00	1.00	7
recall	0.85	0.92	0.88	12
f1-score	0.90	0.82	0.86	11
support	50	50	50	
micro avg	0.90	0.90	0.90	30
macro avg	0.92	0.91	0.91	30
weighted avg	0.90	0.90	0.90	30



**Jason Brownlee** June 14, 2019 at 6:41 am #

Well done!



**teimoor** June 17, 2019 at 4:49 pm #

hi i am trying detecting myocardial infarction on this dataset  
<https://blog.orkami.nl/diagnosing-myocardial-infarction-with-tensorflow/>  
<https://www.kaggle.com/mwaskom/sepsis-detection>  
<https://www.kaggle.com/c/heart-disease/leaderboard>

but after some records processed it gives me the following error:

Using TensorFlow backend.

```
0%| | 0/549 [00:00<?, ?it/s]
0%| | 2/549 [00:00<00:55, 9.86it/s]
1%| | 3/549 [00:00<01:06, 8.17it/s]
1%| | 4/549 [00:00<01:14, 7.29it/s]
1%| | 5/549 [00:00<01:27, 6.19it/s]
1%|1 | 6/549 [00:01<01:42, 5.32it/s]
1%|1 | 7/549 [00:01<01:39, 5.46it/s]
1%|1 | 8/549 [00:01<01:37, 5.57it/s]
2%|1 | 9/549 [00:01<01:45, 5.10it/s]
2%|1 | 10/549 [00:01<01:44, 5.17it/s]
2%|2 | 11/549 [00:02<01:53, 4.76it/s]
2%|2 | 12/549 [00:02<01:49, 4.92it/s]
2%|2 | 13/549 [00:02<02:03, 4.32it/s]
3%|2 | 14/549 [00:02<02:01, 4.40it/s]
3%|2 | 15/549 [00:02<01:59, 4.45it/s]
3%|2 | 16/549 [00:03<02:26, 3.65it/s]
3%|3 | 17/549 [00:03<02:29, 3.56it/s]
3%|3 | 18/549 [00:04<02:49, 3.14it/s]
3%|3 | 19/549 [00:04<02:35, 3.41it/s]
4%|3 | 20/549 [00:04<02:28, 3.57it/s]
```

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Your Start in Machine Learning

```

4%|3 | 21/549 [00:04<02:51, 3.07it/s]
4%|4 | 22/549 [00:05<02:44, 3.20it/s]
4%|4 | 23/549 [00:05<02:54, 3.02it/s]
4%|4 | 24/549 [00:06<03:15, 2.69it/s]
5%|4 | 25/549 [00:06<03:27, 2.52it/s]
5%|4 | 26/549 [00:07<04:07, 2.11it/s]
5%|4 | 27/549 [00:07<03:54, 2.23it/s]
5%|5 | 28/549 [00:08<04:04, 2.13it/s]
5%|5 | 29/549 [00:08<03:41, 2.35it/s]
5%|5 | 30/549 [00:08<03:16, 2.65it/s]
6%|5 | 31/549 [00:09<04:08, 2.08it/s]
6%|5 | 32/549 [00:09<03:58, 2.16it/s]
6%|6 | 33/549 [00:10<04:16, 2.01it/s]
6%|6 | 34/549 [00:10<03:56, 2.17it/s]
6%|6 | 35/549 [00:11<03:52, 2.21it/s]
7%|6 | 36/549 [00:11<04:42, 1.81it/s]
7%|6 | 37/549 [00:12<04:41, 1.82it/s]
7%|6 | 38/549 [00:13<05:06, 1.67it/s]
7%|7 | 39/549 [00:13<04:45, 1.78it/s]
7%|7 | 40/549 [00:14<04:47, 1.77it/s]Traceback (most
File "C:\Program Files\Python\Python37\diagnosingus
record = io.rerecord(record_name=os.path.join('ptbdb'
File "C:\Program Files\Python\Python37\lib\site-pacak
ignore_skew)
File "C:\Program Files\Python\Python37\lib\site-packa
smooth_frames)[:, r_w_channel[fn]]
File "C:\Program Files\Python\Python37\lib\site-packages\wfdb\io\_signal.py", line 992, in _rd_dat_signals
signal = sig_data.reshape(-1, n_sig)
ValueError: cannot reshape array of size 868190 into shape (12)

```



**Jason Brownlee** June 18, 2019 at 6:33 am #

REPLY ↗

I have some suggestions here that might help:

<https://machinelearningmastery.com/faq/single-faq/can-you-read-review-or-debug-my-code>



**teimoor** June 21, 2019 at 5:22 pm #

REPLY ↗

hi i investigate the problem and it seems that data i am using has varying length therefore it throws this exception. how can i fix it to get rid of this reshape error: ValueError: cannot reshape array of size 868190 into shape (12)?

Your Start in Machine Learning



**Jason Brownlee** June 22, 2019 at 6:35 am #

REPLY ↗

Perhaps work with less data as a first step?



**Khadeejah Saeed** June 17, 2019 at 7:36 pm #

REPLY ↗

Here is my Code it is giving some errors. Please help me to sort it out. I have tried same this code in my own dataset.

```
# Python version
import sys
print('Python: {}'.format(sys.version))
# scipy
import scipy
print('scipy: {}'.format(scipy.__version__))
# numpy
import numpy
print('numpy: {}'.format(numpy.__version__))
# matplotlib
import matplotlib
print('matplotlib: {}'.format(matplotlib.__version__))
# pandas
import pandas
print('pandas: {}'.format(pandas.__version__))
# scikit-learn
import sklearn
print('sklearn: {}'.format(sklearn.__version__))

# Load libraries
import pandas
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

# Load dataset
url = r"C:\Users\Khadeej\spyder-py3\DataScience\pc"
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```

names =
['age','sex','cp','trestbps','chol','fbs','restecg','thalach','exang','oldpeak','slope','ca','thal','heartpred']
dataset = pandas.read_csv(url, names=names)

# shape
print(dataset.shape)

# head
print(dataset.head(20))
# descriptions
print(dataset.describe())

# class distribution
print(dataset.groupby('class').size())

# box and whisker plots
dataset.plot(kind='box', subplots=True, layout=(2,2), s=
plt.show()

# histograms
dataset.hist()
plt.show()

# scatter plot matrix
scatter_matrix(dataset)
plt.show()

# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)

# Test options and evaluation metric
seed = 7
scoring = 'accuracy'

# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma='auto')))

# evaluate each model in turn
results = []
names = []

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

START MY EMAIL COURSE

Your Start in Machine Learning

```

for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

# Compare Algorithms
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()

# Make predictions on validation dataset
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
predictions = knn.predict(X_validation)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))

```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)


**Jason Brownlee** June 18, 2019 at 6:37 am #

REPLY ↗

This is a common question that I answer here:

<https://machinelearningmastery.com/faq/single-faq/can-you-read-review-or-debug-my-code>



**Peter** June 27, 2019 at 6:48 pm #

REPLY ↗

Hi, I have problem with this line:

```
import sklearn
```

It has output: „ImportError: No module named ‘sklearn’“

But I tried almost everything (reinstalling, installing version for Python 3 only, ...), but nothing helps.

Thank for your advice.



**Peter** June 27, 2019 at 6:50 pm #

REPLY ↗

Now it works. I work on Python 3.5, and i

Your Start in Machine Learning



**Jason Brownlee** June 28, 2019 at 5:59 am #

REPLY ↗

Well done Peter!



**Jason Brownlee** June 28, 2019 at 6:00 am #

REPLY ↗

I recommend this tutorial:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**pavani** June 29, 2019 at 7:59 pm #

hiii.....

the tutorial poin very useful...its pretty good

i have to project on ..IPL WINNER PREDICTION  
what data should I load?

## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** June 30, 2019 at 9:38 am #

REPLY ↗

Thanks.

Perhaps start here:

[https://machinelearningmastery.com/faq/single-faq/where-can-i-get-a-dataset-on-\\_\\_](https://machinelearningmastery.com/faq/single-faq/where-can-i-get-a-dataset-on-__)



**Eric** July 3, 2019 at 12:49 pm #

REPLY ↗

in section 3.1 im getting unable to initialize device PRN, and thoughts?

thanks!



**Jason Brownlee** July 4, 2019 at 7:37 am #

REPLY ↗

I have not seen that before, sorry.

Perhaps confirm that your libraries are installed correctly:

<http://machinelearningmastery.com/setup-python-anaconda/>

## Your Start in Machine Learning



**Ashish Pratap Singh** July 15, 2019 at 3:04 pm #

REPLY ↗

```
models = []
models.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma='auto')))

# evaluate each model in turn
results = []
names = []

for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=7)
    cv_results = model_selection.cross_val_score(model, X, cv=kfold)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

WHEN I RUN THIS, I GET

ValueError: Unknown label type: 'unknown'
```

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.** Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** July 16, 2019 at 8:12 am #

REPLY ↗

I'm sorry to hear that, I have some suggestions here that might help:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>



**Rob** July 16, 2019 at 8:32 pm #

REPLY ↗

to illustrate the structure of the data, I added color to the scatter matrix:

```
1 color_map={}
2 'Iris-setosa': 'r',
3 'Iris-versicolor': 'g',
4 'Iris-virginica': 'b'
5 }
6
7 dataset['color']=dataset['class'].map(color_map)
8
9 scatter_matrix(dataset,color=dataset['color'])
10 plt.show()
```

## Your Start in Machine Learning



**Jason Brownlee** July 17, 2019 at 8:23 am #

REPLY ↩

Well done Rob!

REPLY ↩



**ToanNguyen** July 17, 2019 at 1:56 am #

Thank you so much. it's my first time with Python.

LR: 0.966667 (0.040825)  
 LDR: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.975000 (0.038188)  
 NB: 0.975000 (0.053359)  
 SVM: 0.991667 (0.025000)



**Jason Brownlee** July 17, 2019 at 8:29 am #

Well done!



**RFI** July 17, 2019 at 6:04 am #

why tensorflow is not installing in python 3.7?



**Jason Brownlee** July 17, 2019 at 8:32 am #

REPLY ↩

Perhaps this will help:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



**aquaman** July 18, 2019 at 6:45 pm #

REPLY ↩

"The confusion matrix provides an indication of the three errors made."

Where are the three errors?



**Jason Brownlee** July 19, 2019 at 9:15 am #

REPLY ↩

Your Start in Machine Learning

Prediction errors.

The report does not indicate what specific instances these were, only the nature of the errors.

You could manually make a prediction for each example and inspect those that had an error to learn more about them.



**Tracy** July 21, 2019 at 12:40 pm #

REPLY ↗

Hello Jason,

```
models.append('LR', LogisticRegression(solver='liblinear', multi_class='ovr')
```

Can you explain what are solver and mult\_class for?



**Jason Brownlee** July 22, 2019 at 8:14 am #

They were set to overcome warnings after

<https://machinelearningmastery.com/how-to-fix-future-warning-messages-in-scikit-learn/>

More on their meaning here:

<https://scikit-learn.org/stable/modules/generated/sklearn.exceptions.FitTransformWarning.html>



**Tracy** July 21, 2019 at 1:50 pm #

REPLY ↗

Hello Jason,

Another question about StandardScaler? why does X\_train need fit and transform and X\_test only need transform?

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X_train_std=sc.fit_transform(X_train)
X_validation_std=sc.transform(X_test)
```



**Jason Brownlee** July 22, 2019 at 8:23 am #

REPLY ↗

The coefficients are calculated on the training set then applied to the train and test sets.



**Tracy** July 21, 2019 at 2:23 pm #

REPLY ↗

Hello Jason,

I guess that fit\_transform does fit and transform, the S

X\_test and X\_train actually are processed in the same way.



**Jason Brownlee** July 22, 2019 at 8:23 am #

REPLY ↗

Yes.



**Ghanshyam** July 28, 2019 at 4:44 pm #

REPLY ↗

Great tutorials



**Jason Brownlee** July 29, 2019 at 6:10 am #

X

Thanks!



**Akash** July 29, 2019 at 4:39 pm #

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.  
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

how do you get the visualizations to appear etc  

```
dataset.plot(kind='box', subplots=True, layout(2,2), sharex=False, sharey=False)
plt.show()
#histograms
dataset.hist()
plt.show()
```

and I get this error.

```
file "/Users/akashchandra/Desktop/Python and ML/python course/iris.py", line 32
dataset.plot(kind='box', subplots=True, layout(2,2), sharex=False, sharey=False)
^
```

SyntaxError: positional argument follows keyword argument  
[Finished in 1.6s with exit code 1]



**Jason Brownlee** July 30, 2019 at 6:02 am #

REPLY ↗

Sorry to hear that you are having trouble, I have some suggestions here:

<https://machinelearningmastery.com/faq/single-faq/why-does-the-code-in-the-tutorial-not-work-for-me>

## Your Start in Machine Learning



**Prafull S Vernekar** August 4, 2019 at 7:45 pm #

REPLY ↗

Dear Mr. Jason Brownlee,

First and foremost thanks for this wonderful, awesome post.

Just worked seamlessly in the very first attempt, being struggling with other tutorials which really never works in the first try.

Please do keep up your sincere efforts.

Thanks and Regards



**Jason Brownlee** August 5, 2019 at 6:51 am #

X

Thanks, I'm happy it worked for you!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

REPLY ↗



**Jason Brownlee** August 7, 2019 at 7:42 am #

Sorry, I don't have a tutorial on this topic, I hope to cover it in the future.

REPLY ↗



**anupam agarwal** August 11, 2019 at 11:38 pm #

sir i am a beginner and want to make robot on ml can you suggest some idea on it.

REPLY ↗



**Jason Brownlee** August 12, 2019 at 6:37 am #

Sorry, I don't know about robots.

REPLY ↗



**Jigyasa** August 15, 2019 at 4:48 pm #

Hi Jason,

Your Start in Machine Learning

I wanted to know one question regarding the training of the model. If my data is having the same trend can my model also predict the data on different offset? or I have to train my model for all the offset?

Best regards,



**Jason Brownlee** August 16, 2019 at 7:46 am #

REPLY ↗

Not sure I follow, do you mean time series and a trend in the series?



**Joseph** August 17, 2019 at 5:29 pm #

Hi Jason,

First, thanks very much for this tutorial. it is easy to follow how to interpret the Algorithm comparison chart? K classification\_report? Finally, based on the knn results

Many thanks



**Jason Brownlee** August 18, 2019 at 6:39 am

Perhaps focus just on accuracy, and start off by choosing a model that has the highest average accuracy.



**Chung Liang** August 26, 2019 at 6:02 pm #

REPLY ↗

Hi Dr. Brownlee,

This was my first ML tutorial in python. Thank you for writing such a simple and easy to follow tutorial. I followed every step and my results were as follows:

LR: 0.966667 (0.040825)  
 LDA: 0.975000 (0.038188)  
 KNN: 0.983333 (0.033333)  
 CART: 0.966667 (0.040825)  
 NB: 0.975000 (0.053359)  
 SVM: 0.991667 (0.025000)

If one wanted to use a different model, where can we find tutorials on the code or are the models already built into the sklearn? Which book would you recommend for beginners in ML without any Statistics background knowledge?

Thanks again for the excellent tutorial.

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** August 27, 2019 at 6:37 am #

REPLY ↗

Well done!

Yes, a good place to explore different models in sklearn us here:

<https://machinelearningmastery.com/start-here/#python>



**Nivitus** September 3, 2019 at 3:02 am #

REPLY ↗

Hai sir , how can i start the machine learning



**Jason Brownlee** September 3, 2019 at 6:19 am #

X

You can use an existing project as a temp

Also, this process will help:

<https://machinelearningmastery.com/start-here/#process>

## Your Start in Machine Learning

You can master applied Machine Learning without math or fancy degrees. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Febil** September 3, 2019 at 9:06 pm #

REPLY ↗

hi i want to do a mini project on weather forecasting. Can you help me to find out what all functions and models can be prepared out from it..



**Jason Brownlee** September 4, 2019 at 5:57 am #

REPLY ↗

Perhaps this process will help:

<https://machinelearningmastery.com/start-here/#process>



**maryam** September 4, 2019 at 2:56 am #

REPLY ↗

Hi Jason,

I have learned machine learning by your clear tutorials like this one.

tell you the truth I am trying to visualize a dataset's distribution, but I do not know how to plot the samples belongs to 2 different class sing two different colors as you did plot all the samples with one color, blue. U have tested some other links, but they do not work.

Your Start in Machine Learning

please let me know about it  
Best  
Maryam



**Jason Brownlee** September 4, 2019 at 6:02 am #

REPLY ↩

Perhaps this will help:

<https://machinelearningmastery.com/data-visualization-methods-in-python/>



**maryam** September 5, 2019 at 8:48 am #

Dear Jason,

I have read it, but all the illustrated figures in the article applied this command and it works for me.

```
import seaborn as sns
```

```
sns.pairplot(hepatit_pca2,
hue = 'Target', diag_kind = 'kde',
plot_kws = { 'edgecolor': 'k'},
size = 6);
```

## Your Start in Machine Learning

X

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

START MY EMAIL COURSE



**Jason Brownlee** September 5, 2019 at 1:47 pm #

REPLY ↩

Thanks for your note.



**Eran** September 5, 2019 at 10:50 pm #

REPLY ↩

Hello, can you please advise on an example with 2 input files :

1. training input file
2. test file

so have code of M learning that knows to predict result (like if transaction is a fraud) in missing result column at test file based on what it learned in the training file



**Jason Brownlee** September 6, 2019 at 5:01 am #

REPLY ↩

That sounds like a great project.

What problem are you having exactly?

Your Start in Machine Learning



**Eran** September 6, 2019 at 2:25 pm #

REPLY ↗

Need advice how to output on screen entire csv columns and rows (like if opened with Excel)



**Jason Brownlee** September 7, 2019 at 5:17 am #

REPLY ↗

What do you mean exactly?

You can output the data and predictions using the print() function, does that help?

## Your Start in Machine Learning

X

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Eran** September 6, 2019 at 3:35 pm #

For example how can I put on screen the vali

```
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```



**Jason Brownlee** September 7, 2019 at 5:19 am #

REPLY ↗

What do you mean put on screen?

Do you mean print to screen? If so, you can use the print() function.



**AI PASSIONATE** September 7, 2019 at 1:23 am #

REPLY ↗

Hello,

I'm following your tutorial but using different dataset that includes dates, entry id, temp, humid, moisture etc so when give this dataset to the model it gives me error that couldn't convert string to float and secondly, the graphs I'm trying to plot is not plotting idk why. Kindly help me.

Thanks in advance.

## Your Start in Machine Learning



**Jason Brownlee** September 7, 2019 at 5:37 am #

REPLY ↗

Perhaps some one or more of the columns contains strings.

If they categorical, they must be encoded to a number, such as with an integer encoding or a one hot encoding. More details here:

<https://machinelearningmastery.com/faq/single-faq/how-to-handle-categorical-data-with-string-values>



**Eran** September 7, 2019 at 3:28 pm #

REPLY ↗

Thanks Jason, I am trying find algorithm where I can load another csv and not slicing from train data (so simulate a real scenario). Can you please refer me to such?



**Jason Brownlee** September 8, 2019 at 5:13 am #

Perhaps this post will help you to understand how to load multiple CSV files into memory:  
<http://machinelearningmastery.com/load-machine-learning-data-multiple-csv-files/>

And this for slicing an array:

<https://machinelearningmastery.com/index-slice-arrays-numpy/>

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**poorvi** September 8, 2019 at 7:14 pm #

REPLY ↗

python code for a tv cable provider has 170 customers over 8km radius. the service provider wishes to restrict his service over 2 km radius w& retain maximum customers as possible .the remaining customers will be transferred to other service provider.i want idea about this problem plz can anybody help me plz.



**Jason Brownlee** September 9, 2019 at 5:13 am #

REPLY ↗

I recommend following this process:

<https://machinelearningmastery.com/start-here/#process>



**pleaseHelp** September 9, 2019 at 5:42 pm #

REPLY ↗

Hi

## Your Start in Machine Learning

I have Create a machine learning keras model and I want to deploy it to los application.  
how should I Convert keras model to coreml.

Thank you.



**Jason Brownlee** September 10, 2019 at 5:37 am #

REPLY ↗

That sounds like a great project.

Sorry, I don't know about iOS.



**Eran** September 13, 2019 at 4:01 pm #

Thanks to this example. Please advise for example how kind of improvement programmer can test



**Jason Brownlee** September 14, 2019 at 6:12 pm #

You can modify the algorithm by changing the learning algorithm.



**Sabrina** September 15, 2019 at 4:43 am #

REPLY ↗

Its actually helpful thank you very much!... I want to know how can the recall , precision and f1 score of each model can be represented in a bar diagram instead of box plots for comparison?



**Jason Brownlee** September 15, 2019 at 6:26 am #

REPLY ↗

You can use matplotlib and call bar()

[https://matplotlib.org/3.1.1/api/\\_as\\_gen/matplotlib.pyplot.bar.html](https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.bar.html)



**Greg Denson** September 15, 2019 at 5:19 am #

REPLY ↗

Dr. Jason, you have a unique website! Because...

- Your Python code examples work – that's my highest compliment to anyone because this scenario seems to have become a great rarity these days!
- Your information is vey useful, and is absolutely the best!

Your Start in Machine Learning

- You take the time to respond to all the emails.
- You know what it takes to teach this subject, and share it clearly.

You are so correct about this being the best way to teach ML. After wasting my money on a stack of ML books, I found your website. So, now, instead of trying to read and understand those books, they've just become a reference library that I seldom turn to – because I come to this website first! (And based on all learned from this site, I did just buy one more book – YOURS!)

Congratulations on a job extremely well done!!!



**Jason Brownlee** September 15, 2019 at 6:27 pm #[4](#)

REPLY ↗

Thanks for your support Greg, I really app



**peter morris** September 20, 2019 at 7:12 pm #[5](#)

thanks it worked first time using anaconda, ba  
get into ML



**Jason Brownlee** September 21, 2019 at 6:50 pm #[6](#)

REPLY ↗

Well done Peter!



**Ayobami** September 21, 2019 at 11:36 pm #[7](#)

REPLY ↗

Hello, please I'm a student. I have a project that I'm about to start on building a classification system for malware with machine learning using python but i don't know where to start. Please i need your counsel on this.



**Jason Brownlee** September 22, 2019 at 9:29 am #[8](#)

REPLY ↗

Perhaps start with this process:

<https://machinelearningmastery.com/start-here/#process>



**Vlad** September 25, 2019 at 3:29 am #[9](#)

REPLY ↗

Your Start in Machine Learning

Does it make sense, when evaluating models, to divide mean by sd, given that I (supposedly) want a high mean and a low std? These are the results:

LR: 0.966667 (0.040825) 23.678401  
 LDA: 0.975000 (0.038188) 25.531493  
 KNN: 0.983333 (0.033333) 29.500000  
 CART: 0.983333 (0.033333) 29.500000  
 NB: 0.975000 (0.053359) 18.272330  
 SVM: 0.991667 (0.025000) 39.666667

Which clearly shows SVM is superior.



**Jason Brownlee** September 25, 2019 at 6:02

Probably not, the samples are small and a



**Villanova** September 25, 2019 at 3:05 pm #

Hey Jason, first of all want to congratulate you! I don't have a programming background and I am almost new to machine learning. My objective in the mid term is to dive into image/pattern recognition (specifically human body behavior captured from pictures). Do you have any tips or resources that I can use? In a few words about what should be my "path" to master machine learning, deep learning, AI is very messy. Just want to hear from you. Thanks and greetings from Brazil!



**Jason Brownlee** September 26, 2019 at 6:29 am #

REPLY ↗

Thanks!

Great question, a great starting point is here:

<https://machinelearningmastery.com/start-here/#getstarted>

I have more on self-study here that I think will help:

<https://machinelearningmastery.com/faq/single-faq/how-do-i-self-study-machine-learning>



**MD Parwaz** September 28, 2019 at 1:06 am #

REPLY ↗

Thanks for these types of help of programmer ..can give me suggestion for object recognition project .....

## Your Start in Machine Learning



**Jason Brownlee** September 28, 2019 at 6:21 am #

REPLY ↗

Perhaps start with some of the tutorials here:

<https://machinelearningmastery.com/start-here/#dlfcv>



**kef** September 29, 2019 at 10:51 pm #

REPLY ↗

any help pls

ImportError: cannot import name 'RandomizedLogisticRegression' from 'sklearn.linear\_model'  
(C:\Users\Kefyalew\Anaconda2\envs\FakenewsEnv\lib\site-packages\sklearn\linear\_model\\_\_init\_\_.py)



**Jason Brownlee** September 30, 2019 at 6:10 pm #

It looks like you are using a different model.

I have some suggestions here that might help:

<https://machinelearningmastery.com/faq/single-page/>



**Sami Cordahi** October 6, 2019 at 7:55 am #

Hi,

I managed to go through the whole example but I found it easier to use Spyder! I got exactly the same output and numbers as in your findings.

Next step; going deeper and learning the syntax and the algos then moving into deep learning example...

Thank you Jason!

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**.

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE



**Jason Brownlee** October 6, 2019 at 8:18 am #

REPLY ↗

Well done!



**sultan** October 9, 2019 at 8:57 pm #

REPLY ↗

if i want to learn machine learning, what should i do, im beginner

**Jason Brownlee** October 10, 2019

Your Start in Machine Learning



Here:

<https://machinelearningmastery.com/start-here/#getstarted>

REPLY ↗

**sultannnnn** October 9, 2019 at 9:04 pm #

```
def add(x, y):
    return x + y

def do_twice(func, x, y):
    return func(func(x,y), func(x,y))

a = 5
b = 10

print(do_twice(add, a,b))

what the output of this code? if I use C# language
```

X

## Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees.**  
Find out how in this *free* and *practical* course.

[START MY EMAIL COURSE](#)

REPLY ↗

**Benjamin** October 13, 2019 at 8:32 am #

Hello, thank you so much sir for this beginner lesson its really been helpfull, however i found this an error "from pandas.plotting import scatter\_matrix" since pandas have been imported already 'from pandas import scatter\_matrix' should do .

REPLY ↗

**Jason Brownlee** October 13, 2019 at 8:36 am #

You must update your version of scikit-learn, see here for instructions:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>

REPLY ↗

**kamran** October 14, 2019 at 5:39 pm #

Great stuff.

Thank you.

A little suggestion (if I did not miss it :P), please if you

[Your Start in Machine Learning](#)

you that you think we should follow to move on.

## Leave a Reply

Name (required)

Email (will not be published) (requi

Website

[SUBMIT COMMENT](#)

### Your Start in Machine Learning

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

[START MY EMAIL COURSE](#)



Welcome! I'm **Jason Brownlee** PhD and I help developers get results with machine learning.  
[Read More](#)

### Picked for you:



[Your First Machine Learning Project in Python Step-By-Step](#)



[How to Setup Your Python Environment for Machine Learning with Anaconda](#)



[Feature Selection For Machine Learning in Python](#)

[Your Start in Machine Learning](#)



## Python Machine Learning Mini-Course



Save and Load Machine Learning Models in Python with scikit-learn

## Loving the Tutorials?

The Machine Learning Mastery with Python EBook is where I keep the **Really Good** stuff.

SEE WHAT'S INSIDE



## Your Start in Machine Learning

You can master applied Machine Learning  
**without math or fancy degrees.**

Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

© 2019 Machine Learning Mastery Pty. Ltd. All Rights Reserved.  
Address: PO Box 206, Vermont Victoria 3133, Australia. | ACN 624 333 008  
[RSS](#) | [Twitter](#) | [Facebook](#) | [LinkedIn](#)

[Privacy](#) | [Disclaimer](#) | [Terms](#) | [Contact](#) | [Sitemap](#) | [Search](#)

Your Start in Machine Learning