

# **DATA MINING**

# **FINAL PROJECT**

**Full Name:** Francis Dcruz

**NJIT UCID:** fjd7

**Email:** [fjd7@njit.edu](mailto:fjd7@njit.edu)

**OPTION 1**

**Category 1:** Support Vector Machine

**Category 2:** Random Forest

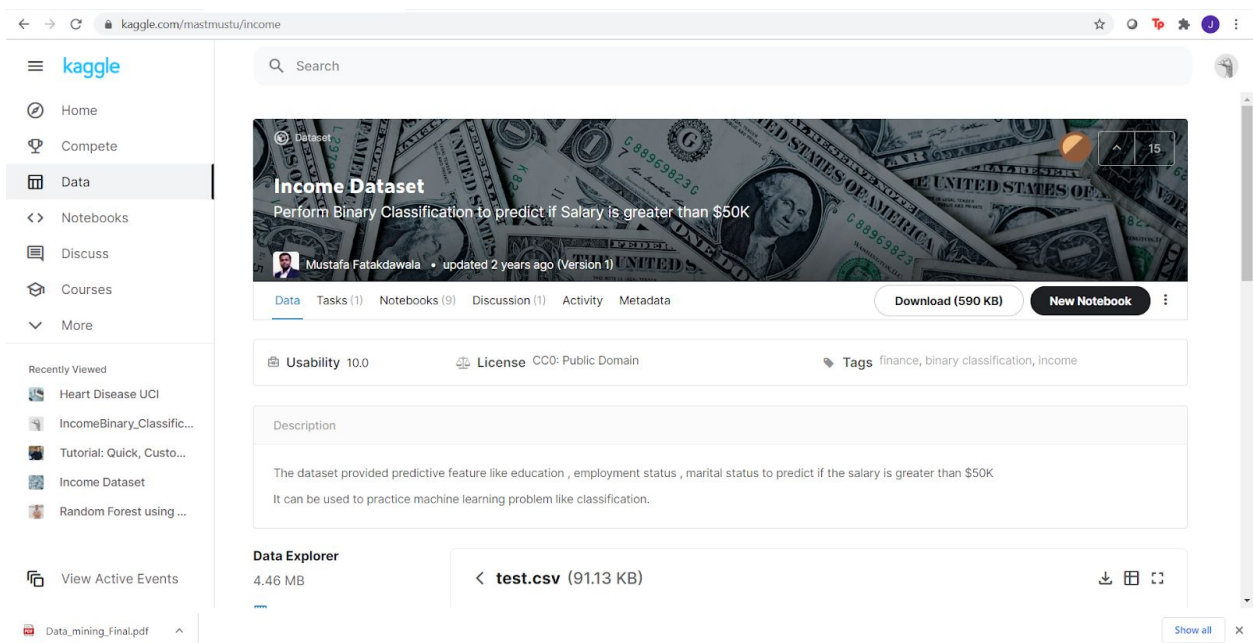
**Programming Language:** Python

**Tools:** Jupyter Notebook

**Dataset:** <https://www.kaggle.com/mastmustu/income>

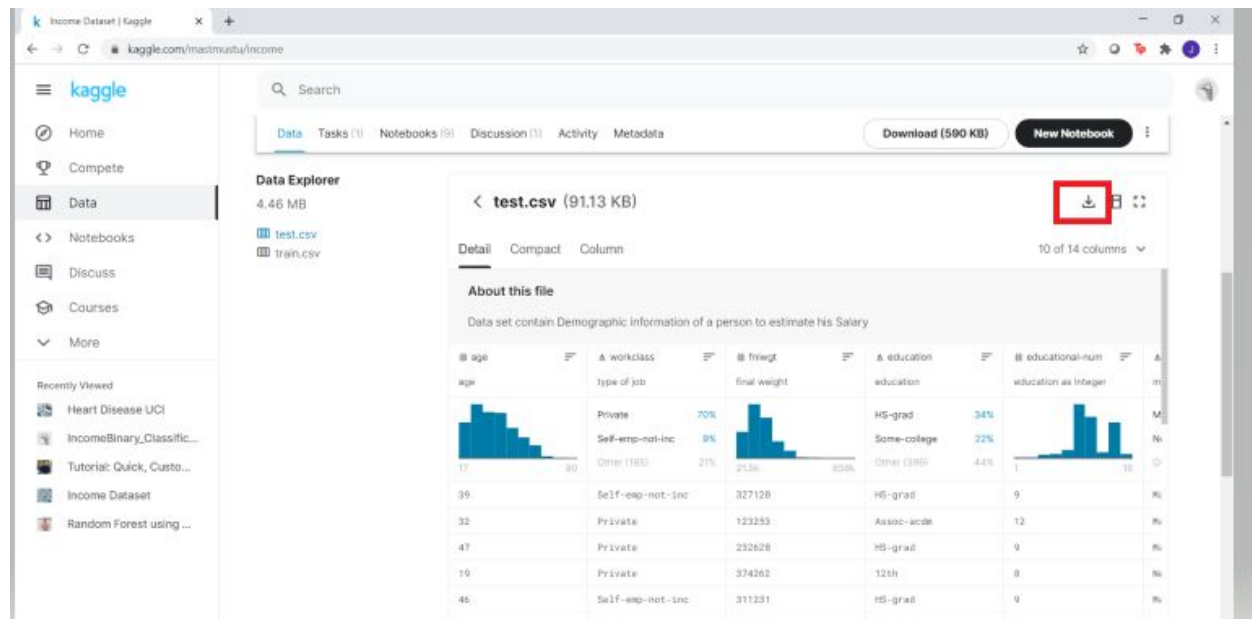
This dataset is a classification dataset about income based on various factors such as The dataset provided predictive feature like education , employment status , marital status to predict if the salary is greater than \$50K

To download the dataset, I clicked the URL mention above, and the following will be seen:

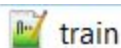


The screenshot shows the Kaggle website interface for the 'Income Dataset'. The left sidebar contains navigation links: Home, Compete, Data (selected), Notebooks, Discuss, Courses, and More. Below these are 'Recently Viewed' items and 'View Active Events'. The main content area displays the dataset title 'Income Dataset' with a subtitle 'Perform Binary Classification to predict if Salary is greater than \$50K'. It shows the creator 'Mustafa Fatakawala' and 'updated 2 years ago (Version 1)'. There are buttons for 'Download (590 KB)' and 'New Notebook'. Below this, the 'Usability' is 10.0, the 'License' is CC0: Public Domain, and 'Tags' include finance, binary classification, and income. The 'Description' section states: 'The dataset provided predictive feature like education , employment status , marital status to predict if the salary is greater than \$50K. It can be used to practice machine learning problem like classification.' At the bottom, the 'Data Explorer' shows a file named 'test.csv' (91.13 KB) with a size of 4.46 MB. A 'Show all' button is visible in the bottom right corner.

Then I scrolled down, clicked on the download option, and downloaded the dataset.



The downloaded data is stored in CSV format.



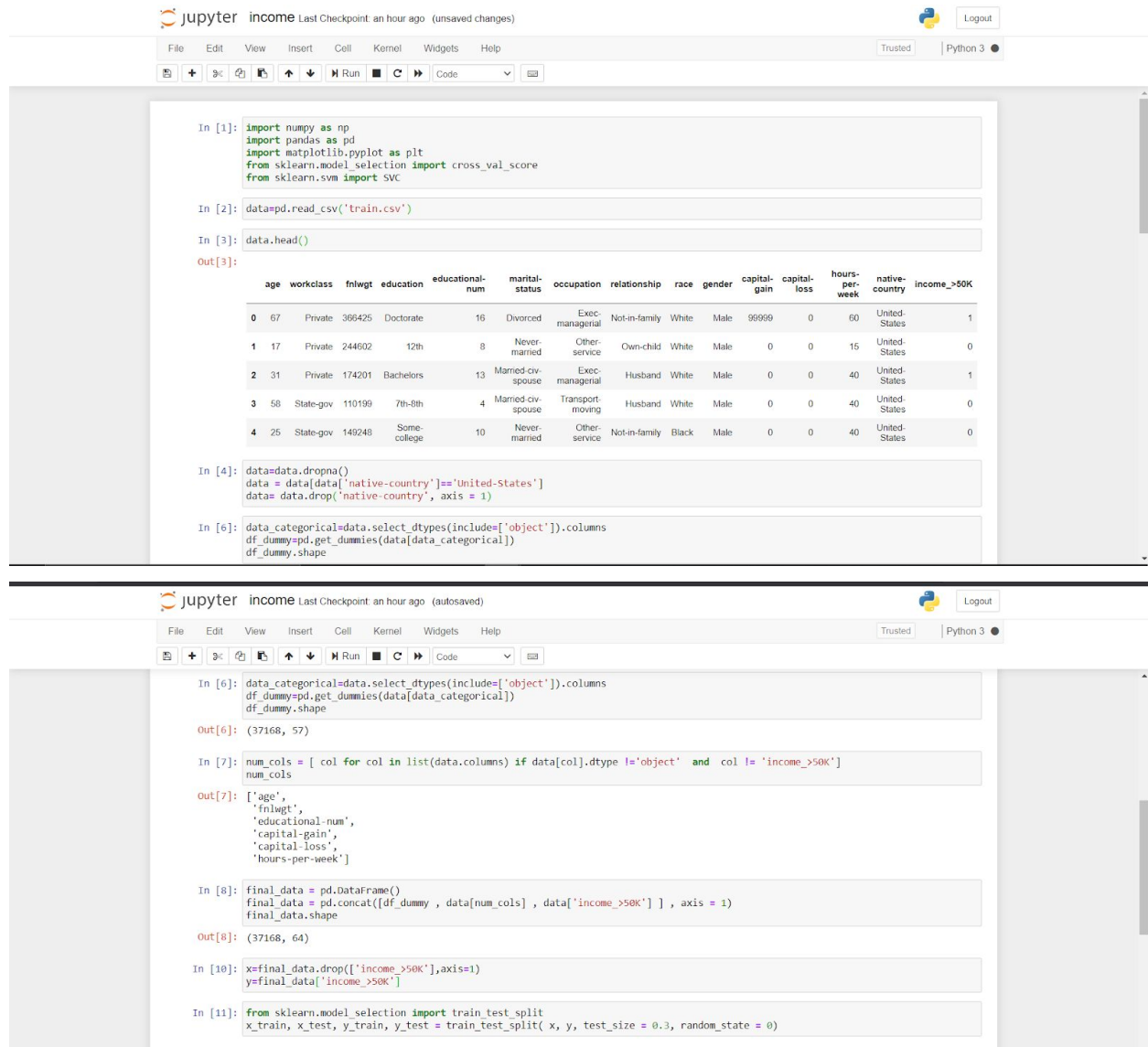
This is how the dataset “train.csv” looks. *Please note* this is not the full dataset, it’s just a snippet to show how the data looks.

```
age,workclass,fnlwgt,education,educational-num,marital-status,occupation,relationship,race,gender,capital-gain,capital-loss,hours-per-week,native-country,income_>50K
67,Private,366425,Doctorate,16,Divorced,Exec-managerial,Not-in-family,White,Male,99999,0,60,United-States,1
17,Private,244602,12th,8,Never-married,Other-service,Own-child,White,Male,0,0,15,United-States,0
31,Private,174201,Bachelors,13,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,0,40,United-States,1
58,State-gov,110199,7th-8th,4,Married-civ-spouse,Transport-moving,Husband,White,Male,0,0,40,United-States,0
25,State-gov,149248,Some-college,10,Never-married,Other-service,Not-in-family,Black,Male,0,0,40,United-States,0
59,State-gov,105363,HS-grad,9,Never-married,Adm-clerical,Own-child,White,Male,0,0,40,United-States,0
70,Private,216390,9th,5,Married-civ-spouse,Machine-op-inspct,Wife,White,Female,2653,0,40,United-States,0
35,Self-emp-not-inc,361888,Bachelors,13,Married-civ-spouse,Sales,Husband,White,Male,0,0,60,Japan,0
28,Private,74784,HS-grad,9,Never-married,Handlers-cleaners,Not-in-family,White,Male,0,0,50,United-States,0
28,Private,118089,HS-grad,9,Married-civ-spouse,Exec-managerial,Husband,White,Male,4386,0,45,United-States,1
21,Private,138513,Some-college,10,Never-married,Exec-managerial,Own-child,White,Male,0,0,25,United-States,0
30,Self-emp-not-inc,100252,HS-grad,9,Married-civ-spouse,Machine-op-inspct,Own-child,Asian-Pac-Islander,Male,0,0,60,South,0
59,Self-emp-not-inc,241297,Some-college,10,Widowed,Farming-fishing,Not-in-family,White,Female,6849,0,40,United-States,0
20,Private,39764,HS-grad,9,Never-married,Handlers-cleaners,Own-child,White,Male,0,0,40,United-States,0
45,Private,30690,7th-8th,4,Never-married,Other-service,Not-in-family,White,Male,0,0,10,United-States,0
76,Private,316185,7th-8th,4,Widowed,Protective-serv,Not-in-family,White,Female,0,0,12,United-States,0
30,Private,110239,10th,6,Married-civ-spouse,Transport-moving,Husband,White,Male,0,0,55,United-States,0
54,Federal-gov,278076,HS-grad,9,Married-civ-spouse,Exec-managerial,Husband,White,Male,5178,0,40,United-States,1
19,Local-gov,259169,Some-college,10,Never-married,Prof-specialty,Own-child,White,Female,0,0,30,United-States,0
44,Private,136986,Doctorate,16,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,0,50,United-States,1
31,Private,154227,Some-college,10,Married-civ-spouse,Transport-moving,Husband,White,Male,0,0,40,United-States,0
```

47,Private,117774,HS-grad,9,Married-civ-spouse,Sales,Husband,White,Male,0,0,40,Portugal,0  
 37,Self-emp-not-inc,174308,HS-grad,9,Married-civ-spouse,Craft-repair,Husband,White,Male,0,0,40,United-States,0  
 59,Private,189721,Bachelors,13,Married-civ-spouse,Handlers-cleaners,Husband,White,Male,0,0,40,Italy,1  
 27,Self-emp-not-inc,229125,11th,7,Married-civ-spouse,Craft-repair,Husband,White,Male,0,0,40,United-States,0  
 24,Private,341294,HS-grad,9,Married-civ-spouse,Machine-op-inspct,Husband,White,Male,0,0,40,United-States,0  
 49,Private,102583,Some-college,10,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,1848,44,United-States,1  
 36,Private,64874,Masters,14,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,0,60,United-States,1  
 54,Self-emp-inc,129432,Bachelors,13,Married-civ-spouse,Exec-managerial,Husband,White,Male,15024,0,40,United-States,1  
 44,Private,163331,Some-college,10,Widowed,Adm-clerical,Unmarried,White,Female,0,0,32,United-States,0  
 35,Private,38245,Some-college,10,Married-civ-spouse,Craft-repair,Husband,White,Male,0,0,60,United-States,0  
 19,Local-gov,167816,HS-grad,9,Divorced,Exec-managerial,Not-in-family,White,Female,0,0,35,United-States,0  
 71,Private,152307,HS-grad,9,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,2377,45,United-States,1  
 20,,249087,Some-college,10,Never-married,,Own-child,White,Female,0,0,40,United-States,0  
 62,Self-emp-not-inc,136684,HS-grad,9,Widowed,Adm-clerical,Other-relative,White,Female,0,0,30,United-States,0  
 35,Self-emp-not-inc,241126,HS-grad,9,Never-married,Craft-repair,Other-relative,White,Male,0,0,60,United-States,0  
 64,Local-gov,287277,9th,5,Married-civ-spouse,Other-service,Husband,White,Male,0,0,40,United-States,0  
 53,Private,188644,Preschool,1,Married-civ-spouse,Other-service,Husband,White,Male,0,0,40,Mexico,0  
 30,Private,182833,Some-college,10,Never-married,Exec-managerial,Own-child,Black,Female,0,0,40,United-States,0  
 34,Private,434463,Bachelors,13,Never-married,Machine-op-inspct,Not-in-family,White,Female,0,0,39,United-States,0  
 62,Private,312818,Bachelors,13,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,1902,1,United-States,1  
 21,State-gov,173534,Some-college,10,Never-married,Prof-specialty,Own-child,White,Female,0,0,40,Ecuador,0  
 42,Private,204235,HS-grad,9,Married-civ-spouse,Sales,Husband,White,Male,0,0,40,United-States,0  
 26,Private,193165,Some-college,10,Married-civ-spouse,Transport-moving,Husband,White,Male,0,0,52,United-States,1  
 43,Private,175133,Some-college,10,Divorced,Tech-support,Unmarried,Black,Female,0,0,35,United-States,0  
 28,Private,161538,Bachelors,13,Never-married,Tech-support,Not-in-family,White,Female,0,0,35,United-States,0  
 23,Private,177787,Bachelors,13,Never-married,Sales,Own-child,White,Female,0,0,30,England,0  
 56,Self-emp-not-inc,176280,HS-grad,9,Divorced,Other-service,Not-in-family,White,Male,0,0,50,United-States,0  
 19,,174233,Some-college,10,Never-married,,Own-child,Black,Male,0,0,24,United-States,0  
 21,Private,157893,HS-grad,9,Never-married,Transport-moving,Own-child,White,Female,0,0,40,United-States,0  
 30,Private,296462,Masters,14,Never-married,Prof-specialty,Not-in-family,Black,Male,0,0,30,United-States,0  
 58,Private,136951,Bachelors,13,Married-civ-spouse,Adm-clerical,Husband,Asian-Pac-Islander,Male,0,0,40,Philippines,0  
 40,Private,87771,HS-grad,9,Never-married,Adm-clerical,Not-in-family,White,Female,0,0,40,United-States,0  
 21,Private,206008,Some-college,10,Never-married,Adm-clerical,Own-child,Black,Male,0,0,50,United-States,0  
 25,Private,340288,9th,5,Married-civ-spouse,Handlers-cleaners,Husband,White,Male,0,0,40,United-States,0  
 63,Self-emp-not-inc,124015,Masters,14,Separated,Prof-specialty,Not-in-family,White,Male,0,0,40,United-States,1  
 50,Self-emp-inc,262777,Masters,14,Separated,Exec-managerial,Unmarried,Asian-Pac-Islander,Male,0,0,45,China,0  
 29,Local-gov,82393,HS-grad,9,Never-married,Handlers-cleaners,Unmarried,Asian-Pac-Islander,Male,0,1590,45,United-States,0  
 50,Local-gov,141875,HS-grad,9,Married-civ-spouse,Transport-moving,Husband,White,Male,0,0,40,United-States,0  
 27,Private,60288,Masters,14,Never-married,Sales,Own-child,White,Female,0,0,40,United-States,0  
 19,Private,366088,9th,5,Never-married,Machine-op-inspct,Own-child,White,Male,0,0,40,United-States,0  
 52,State-gov,254285,Doctorate,16,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,1887,70,Germany,1  
 33,Local-gov,100734,HS-grad,9,Divorced,Tech-support,Not-in-family,White,Female,0,0,55,United-States,0  
 38,State-gov,534775,Some-college,10,Never-married,Tech-support,Unmarried,Black,Female,0,0,50,United-States,0

Project:

This project runs train.csv with SVM and Random forest. I am using numpy, pandas, and cross\_val\_score from sklearn.model\_selection. The whole project is shown below.



The image displays two screenshots of a Jupyter Notebook interface, showing the initial steps of a machine learning project. The notebook is titled "income" and shows the first six cells of code.

**Cell 1:** Imports necessary libraries: `import numpy as np`, `import pandas as pd`, `import matplotlib.pyplot as plt`, `from sklearn.model_selection import cross_val_score`, and `from sklearn.svm import SVC`.

**Cell 2:** Loads the training data: `data=pd.read_csv('train.csv')`.

**Cell 3:** Displays the first five rows of the data: `data.head()`. The output shows a table with columns: age, workclass, fnlwgt, education, educational-num, marital-status, occupation, relationship, race, gender, capital-gain, capital-loss, hours-per-week, native-country, and income\_>50K. The first five rows are:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income_>50K
0	67	Private	366425	Doctorate	16	Divorced	Exec-managerial	Not-in-family	White	Male	99999	0	60	United-States	1
1	17	Private	244602	12th	8	Never-married	Other-service	Own-child	White	Male	0	0	15	United-States	0
2	31	Private	174201	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	1
3	58	State-gov	110199	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	40	United-States	0
4	25	State-gov	149248	Some-college	10	Never-married	Other-service	Not-in-family	Black	Male	0	0	40	United-States	0

**Cell 4:** Drops missing values: `data=data.dropna()`. Then, it filters the data to only include rows where the native-country is 'United-States': `data = data[data['native-country']=='United-States']`. Finally, it drops the 'native-country' column: `data=data.drop('native-country', axis = 1)`.

**Cell 5:** Selects categorical features: `data_categorical=data.select_dtypes(include=['object']).columns`. It then creates dummy variables for these features: `df_dummy=pd.get_dummies(data[data_categorical])`. The output shows the shape of the dummy variables: `df_dummy.shape`.

**Cell 6:** Displays the shape of the dummy variables: `df_dummy.shape`. The output is: `(37168, 57)`.

**Cell 7:** Defines the numerical columns: `num_cols = [ col for col in list(data.columns) if data[col].dtype != 'object' and col != 'income_>50K' ]`. The output shows the list of numerical columns: `num_cols`.

**Cell 8:** Concatenates the dummy variables and numerical features: `final_data = pd.DataFrame()`, `final_data = pd.concat([df_dummy , data[num_cols] , data['income_>50K'] ] , axis = 1)`. The output shows the shape of the final data: `final_data.shape`.

**Cell 9:** Displays the shape of the final data: `final_data.shape`. The output is: `(37168, 64)`.

**Cell 10:** Drops the 'income\_>50K' column: `x=final_data.drop(['income_>50K'],axis=1)`. The output shows the shape of the final data: `y=final_data['income_>50K']`.

**Cell 11:** Imports the train\_test\_split function: `from sklearn.model_selection import train_test_split`. The output shows the shape of the final data: `x_train, x_test, y_train, y_test = train_test_split( X, y, test_size = 0.3, random_state = 0)`.

jupyter income Last Checkpoint: an hour ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Random Forest

```
In [12]: from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=200, criterion='gini', max_depth=15, max_features='auto', random_state=0)
classifier.fit(x_train, y_train)
y_pred_test = classifier.predict(x_test)
```

```
In [14]: from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred_test))

[[7882  434]
 [1223 1612]]
```

```
In [15]: from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred_test)
print("Accuracy for this model {} %".format(acc*100))

Accuracy for this model 85.14034615729531 %
```

```
In [21]: scores = cross_val_score(classifier, x_train, y_train, cv=10)
print(" Accuracy score for Random Forest using 10-fold cross validation: %.2f (+/- %.2f)" % (scores.mean(), scores.std() * 2))

Accuracy score for Random Forest using 10-fold cross validation: 0.86 (+/- 0.01)
```

```
In [16]: from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x = sc.fit_transform(x_train)
```

jupyter income Last Checkpoint: an hour ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```
x = sc.fit_transform(x_train)
```

SVM

```
In [17]: from sklearn.svm import SVC
classifier1 = SVC(kernel='rbf')
classifier1.fit(x_train, y_train)
y_pred_test1 = classifier1.predict(x_test)
```

```
In [19]: from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred_test1))

[[8307    9]
 [2416  419]]
```

```
In [20]: from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred_test1)
print("Accuracy for this model {} %".format(acc*100))

Accuracy for this model 78.25307147341046 %
```

```
In [18]: scores = cross_val_score(classifier1, x_train, y_train, cv=10)
print(" Accuracy score for SVM using 10-fold cross validation: %.2f (+/- %.2f)" % (scores.mean(), scores.std() * 2))

Accuracy score for SVM using 10-fold cross validation: 0.78 (+/- 0.01)
```

```
In [ ]:
```

Since I used Jupyter notebook to run the project, keep into consideration to place your CSV file and project file into the same folder.

jupyter Quit Logout

Files Running Clusters

Select items to perform actions on them. Upload New

<input type="checkbox"/>	0		data mining	Name	Last Modified	File size
<input type="checkbox"/>			..		seconds ago	
<input type="checkbox"/>			income.ipynb		Running a minute ago	18.5 kB
<input type="checkbox"/>			train.csv		an hour ago	4.58 MB

1. I first imported the required libraries, loaded the data using `pandas.read_csv`, and then stored the CSV data into a data frame.

```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import cross_val_score
from sklearn.svm import SVC
```

```
In [2]: data=pd.read_csv('train.csv')
```

2. Then I displayed the first 5 rows of the dataset using `head()` function. Also I dropped na from the dataset `dropna()`.

```
In [6]: data.head()
Out[6]:
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income_>50K
0	67	Private	369425	Doctorate	16	Divorced	Exec-managerial	Not-in-family	White	Male	99999	0	60	United-States	1
1	17	Private	244602	12th	8	Never-married	Other-service	Own-child	White	Male	0	0	15	United-States	0
2	31	Private	174201	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	1
3	58	State-gov	110199	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	40	United-States	0
4	25	State-gov	149248	Some-college	10	Never-married	Other-service	Not-in-family	Black	Male	0	0	40	United-States	0

```
In [7]: data=data.dropna()
data = data[data['native-country']!='United-States']
data=data.drop('native-country', axis = 1)
```

3. Stored the categorical data and then encoded it into numerical data using `get_dummies()`.

```
In [9]: data_categorical=data.select_dtypes(include=['object']).columns
df_dummy=pd.get_dummies(data[data_categorical])
df_dummy.shape
Out[9]: (37168, 57)
```

4. Fetching all the numerical column names from the dataset.

```
In [10]: num_cols = [ col for col in list(data.columns) if data[col].dtype !='object' and col != 'income_>50K']
num_cols
Out[10]: ['age',
'fnlwgt',
'educational-num',
'capital-gain',
'capital-loss',
'hours-per-week']
```

5. Creating another dataframe 'final\_data' using `pd.DataFrame()`. Then concatenating the categorical data, numerical data, and income prediction into 'final\_data'.

```
In [11]: final_data = pd.DataFrame()
final_data = pd.concat([df_dummy , data[num_cols] , data['income_>50K'] ] , axis = 1)
final_data.shape
Out[11]: (37168, 64)
```

6. Separating and storing the dependent and independent variables in x and y.

```
In [13]: x=final_data.drop(['income_>50K'],axis=1)
y=final_data['income_>50K']
```

7. Using `train_test_split`, the dataset is split into a training set and test set.

```
In [14]: from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)
```

8. Feature scaling on the split data using `StandardScaler` from `sklearn.preprocessing`.

```
In [9]: from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

## SVM

9. From `sklearn.svm` I used `SVC`. I used the kernel type as `rbf` to train the `SVC` model on the training set. After training the model, I predicted the Test set results using `predict()`.

```
SVM

In [17]: from sklearn.svm import SVC
classifier1 = SVC(kernel='rbf')
classifier1.fit(x_train, y_train)
y_pred_test1 = classifier1.predict(x_test)

In [19]: from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred_test1))

[[8307  9]
 [2416 419]]

In [20]: from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred_test1)
print("Accuracy for this model {} %".format(acc*100))

Accuracy for this model 78.25307147341046 %

In [18]: scores = cross_val_score(classifier1, x_train, y_train, cv=10)
print(" Accuracy score for SVM using 10-fold cross validation: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

Accuracy score for SVM using 10-fold cross validation: 0.78 (+/- 0.01)
```

10. Then I Evaluated the Model Performance and got an accuracy of 78.25%

```
SVM

In [17]: from sklearn.svm import SVC
classifier1 = SVC(kernel='rbf')
classifier1.fit(x_train, y_train)
y_pred_test1 = classifier1.predict(x_test)

In [19]: from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred_test1))

[[8307  9]
 [2416 419]]

In [20]: from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred_test1)
print("Accuracy for this model {} %".format(acc*100))

Accuracy for this model 78.25307147341046 %

In [18]: scores = cross_val_score(classifier1, x_train, y_train, cv=10)
print(" Accuracy score for SVM using 10-fold cross validation: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

Accuracy score for SVM using 10-fold cross validation: 0.78 (+/- 0.01)
```



11. In this step, I performed 10-fold cross-validation, and the accuracy score as 0.78 with a standard deviation of (+/- 0.01).

```
SVM

In [17]: from sklearn.svm import SVC
         classifier1 = SVC(kernel='rbf')
         classifier1.fit(x_train,y_train)
         y_pred_test1 = classifier1.predict(x_test)

In [19]: from sklearn.metrics import classification_report, confusion_matrix
         print(confusion_matrix(y_test, y_pred_test1))

[[8307  0]
 [2416 419]]

In [20]: from sklearn.metrics import accuracy_score
         acc = accuracy_score(y_test,y_pred_test1)
         print("Accuracy for this model {} %".format(acc*100))

Accuracy for this model 78.25307147341046 %

In [18]: scores = cross_val_score(classifier1, x_train, y_train, cv=10)
         print(" Accuracy score for SVM using 10-fold cross validation: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

Accuracy score for SVM using 10-fold cross validation: 0.78 (+/- 0.01)
```

## Random Forest

12. From sklearn.ensemble I used RandomForestClassifier. Trained the RandomForestClassifier model on the training set. After training the model, I predicted the Test set results using predict().

```
Random Forest

In [12]: from sklearn.ensemble import RandomForestClassifier
         classifier = RandomForestClassifier(n_estimators=200, criterion='gini', max_depth=15, max_features='auto', random_state=0)
         classifier.fit(x_train, y_train)
         y_pred_test = classifier.predict(x_test)

In [14]: from sklearn.metrics import classification_report, confusion_matrix
         print(confusion_matrix(y_test, y_pred_test))

[[7882  434]
 [1223 1612]]

In [15]: from sklearn.metrics import accuracy_score
         acc = accuracy_score(y_test,y_pred_test)
         print("Accuracy for this model {} %".format(acc*100))

Accuracy for this model 85.14034615729531 %

In [21]: scores = cross_val_score(classifier, x_train, y_train, cv=10)
         print(" Accuracy score for Random Forest using 10-fold cross validation: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

Accuracy score for Random Forest using 10-fold cross validation: 0.86 (+/- 0.01)
```

13. Then I Evaluated the Model Performance and got an accuracy of 85.14%

```
Random Forest

In [12]: from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=200, criterion='gini', max_depth=15, max_features='auto', random_state=0)
classifier.fit(x_train, y_train)
y_pred_test = classifier.predict(x_test)

In [14]: from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred_test))

[[7882  434]
 [1223 1612]]

In [15]: from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred_test)
print("Accuracy for this model {} %".format(acc*100))

Accuracy for this model 85.14034615729531 %

In [21]: scores = cross_val_score(classifier, x_train, y_train, cv=10)
print(" Accuracy score for Random Forest using 10-fold cross validation: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

Accuracy score for Random Forest using 10-fold cross validation: 0.86 (+/- 0.01)
```

14. In this step, I performed 10-fold cross-validation, and the accuracy score as 0.86 with a standard deviation of (+/- 0.01).

```
Random Forest

In [12]: from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=200, criterion='gini', max_depth=15, max_features='auto', random_state=0)
classifier.fit(x_train, y_train)
y_pred_test = classifier.predict(x_test)

In [14]: from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred_test))

[[7882  434]
 [1223 1612]]

In [15]: from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred_test)
print("Accuracy for this model {} %".format(acc*100))

Accuracy for this model 85.14034615729531 %

In [21]: scores = cross_val_score(classifier, x_train, y_train, cv=10)
print(" Accuracy score for Random Forest using 10-fold cross validation: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

Accuracy score for Random Forest using 10-fold cross validation: 0.86 (+/- 0.01)
```

**Source Code:** The code I implemented on the dataset I chose.

```
import numpy as np
import pandas as pd
from sklearn.model_selection import cross_val_score

data=pd.read_csv('train.csv')

data.head()

data=data.dropna()
data = data[data['native-country']=='United-States']
data= data.drop('native-country', axis = 1)

data_categorical=data.select_dtypes(include=['object']).columns
df_dummy=pd.get_dummies(data[data_categorical])
df_dummy.shape

num_cols = [ col for col in list(data.columns) if data[col].dtype !='object' and col !=
'income_>50K']
num_cols

final_data = pd.DataFrame()
final_data = pd.concat([df_dummy , data[num_cols] , data['income_>50K'] ] , axis = 1)
final_data.shape

x=final_data.drop(['income_>50K'],axis=1)
y=final_data['income_>50K']

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split( x, y, test_size = 0.3, random_state = 0)

from sklearn.preprocessing import StandardScaler
sc= StandardScaler()
x= sc.fit_transform(x_train)

Random Forest

from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators =200, criterion = 'gini',
max_depth=15, max_features='auto',random_state = 0)
classifier.fit(x_train, y_train)
```

```
y_pred_test = classifier.predict(x_test)

from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred_test))

from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred_test)
print("Accuracy for this model {} {}".format(acc*100))

scores = cross_val_score(classifier, x_train, y_train, cv=10)
print(" Accuracy score for Random Forest using 10-fold cross validation: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

SVM
from sklearn.svm import SVC
classifier1 = SVC(kernel='rbf')
classifier1.fit(x_train, y_train)
y_pred_test1 = classifier1.predict(x_test)

from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred_test1))

from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred_test1)
print("Accuracy for this model {} {}".format(acc*100))

scores = cross_val_score(classifier1, x_train, y_train, cv=10)
print(" Accuracy score for SVM using 10-fold cross validation: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```

**Related Source Code:** Some related source code I used in this project are posted below.

import pandas as pd

pd.read():

[LINK](#)

```
pandas.read_csv(filepath_or_buffer, sep=',', delimiter=None, header='infer', names=None, index_col=None, usecols=None, squeeze=False, prefix=None, mangle_dupe_cols=True, dtype=None, engine=None, converters=None, true_values=None, false_values=None, skipinitialspace=False, skiprows=None, skipfooter=0, nrows=None, na_values=None, keep_default_na=True, na_filter=True, verbose=False, skip_blank_lines=True, parse_dates=False, infer_datetime_format=False, keep_date_col=False, date_parser=None, dayfirst=False, cache_dates=True, iterator=False, chunksize=None, compression='infer', thousands=None, decimal='.', lineterminator=None, quotechar='"', quoting=0, doublequote=True, escapechar=None, comment=None, encoding=None, dialect=None, error_bad_lines=True, warn_bad_lines=True, delim_whitespace=False, low_memory=True, memory_map=False, float_precision=None)[source]
```

Read a comma-separated values (csv) file into DataFrame.

Also supports optionally iterating or breaking of the file into chunks.

Additional help can be found in the online docs for [IO Tools](#).

#### Parameters

##### **filepath\_or\_buffer**, path object or file-like object

Any valid string path is acceptable. The string could be a URL. Valid URL schemes include http, ftp, s3, gs, and file. For file URLs, a host is expected. A local file could be: `file://localhost/path/to/table.csv`.

If you want to pass in a path object, pandas accepts any `os.PathLike`.

By file-like object, we refer to objects with a `read()` method, such as a file handler (e.g. via builtin `open` function) or `StringIO`.

##### **sep**, default `,`

Delimiter to use. If sep is None, the C engine cannot automatically detect the separator, but the Python parsing engine can, meaning the latter will be used and automatically detect the separator by Python's builtin sniffer tool, `csv.Sniffer`. In addition, separators longer than 1 character and different from `"s+"` will be interpreted as regular expressions and will also force the use of the Python parsing engine. Note that regex delimiters are prone to ignoring quoted data. Regex example: `"r\\t"`.

##### **delimiter**, default `None`

Alias for sep.

##### **header**, list of int, default `'infer'`

Row number(s) to use as the column names, and the start of the data. Default behavior is to infer the column names: if no names are passed the behavior is identical to `header=0` and column names are inferred from the first line of the file, if column names are passed explicitly then the behavior is identical to `header=None`. Explicitly pass `header=0` to be able to replace existing names. The header can be a list of integers that specify row locations for a multi-index on the columns e.g. `[0,1,3]`. Intervening rows that are not specified will be skipped (e.g. 2 in this example is skipped). Note that this parameter ignores commented lines and empty lines if `skip_blank_lines=True`, so `header=0` denotes the first line of data rather than the first line of the file.

#### **namesarray-like, optional**

List of column names to use. If the file contains a header row, then you should explicitly pass `header=0` to override the column names. Duplicates in this list are not allowed.

#### **index\_colint, str, sequence of int / str, or False, default None**

Column(s) to use as the row labels of the `DataFrame`, either given as string name or column index. If a sequence of int / str is given, a `MultiIndex` is used.

Note: `index_col=False` can be used to force pandas to *not* use the first column as the index, e.g. when you have a malformed file with delimiters at the end of each line.

#### **usecolslist-like or callable, optional**

Return a subset of the columns. If list-like, all elements must either be positional (i.e. integer indices into the document columns) or strings that correspond to column names provided either by the user in names or inferred from the document header row(s). For example, a valid list-like `usecols` parameter would be `[0, 1, 2]` or `['foo', 'bar', 'baz']`. Element order is ignored, so `usecols=[0, 1]` is the same as `[1, 0]`. To instantiate a `DataFrame` from `data` with element order preserved use `pd.read_csv(data, usecols=['foo', 'bar'])[['foo', 'bar']]` for columns in `['foo', 'bar']` order or `pd.read_csv(data, usecols=['foo', 'bar'])[['bar', 'foo']]` for `['bar', 'foo']` order.

If callable, the callable function will be evaluated against the column names, returning names where the callable function evaluates to `True`. An example of a valid callable argument would be `lambda x: x.upper() in ['AAA', 'BBB', 'DDD']`. Using this parameter results in much faster parsing time and lower memory usage.

#### **squeezebool, default False**

If the parsed data only contains one column then return a `Series`.

#### **prefixstr, optional**

Prefix to add to column numbers when no header, e.g. 'X' for X0, X1, ...

#### **mangle\_dupescolsbool, default True**

Duplicate columns will be specified as 'X', 'X.1', ... 'X.N', rather than 'X'...'X'. Passing in `False` will cause data to be overwritten if there are duplicate names in the columns.

#### **dtypeType name or dict of column -> type, optional**

Data type for data or columns. E.g. `{ 'a': np.float64, 'b': np.int32, 'c': 'Int64' }` Use str or object together with suitable `na_values` settings to preserve and not interpret dtype. If converters are specified, they will be applied INSTEAD of dtype conversion.

#### **engine{'c', 'python'}, optional**

Parser engine to use. The C engine is faster while the python engine is currently more feature-complete.

**convertersdict, optional**

Dict of functions for converting values in certain columns. Keys can either be integers or column labels.

**true\_valueslist, optional**

Values to consider as True.

**false\_valueslist, optional**

Values to consider as False.

**skipinitialspacebool, default False**

Skip spaces after delimiter.

**skiprowslist-like, int or callable, optional**

Line numbers to skip (0-indexed) or number of lines to skip (int) at the start of the file.

If callable, the callable function will be evaluated against the row indices, returning True if the row should be skipped and False otherwise. An example of a valid callable argument would be `lambda x: x in [0, 2]`.

**skipfooterint, default 0**

Number of lines at bottom of file to skip (Unsupported with engine='c').

**nrowsint, optional**

Number of rows of file to read. Useful for reading pieces of large files.

**na\_valuesscalar, str, list-like, or dict, optional**

Additional strings to recognize as NA/NaN. If dict passed, specific per-column NA values. By default the following values are interpreted as NaN: ' ', '#N/A', '#N/A N/A', '#NA', '-1.#IND', '-1.#QNAN', '-NaN', '-nan', '1.#IND', '1.#QNAN', '<NA>', 'N/A', 'NA', 'NULL', 'NaN', 'n/a', 'nan', 'null'.

**keep\_default\_nabool, default True**

Whether or not to include the default NaN values when parsing the data. Depending on whether na\_values is passed in, the behavior is as follows:

- If keep\_default\_na is True, and na\_values are specified, na\_values is appended to the default NaN values used for parsing.
- If keep\_default\_na is True, and na\_values are not specified, only the default NaN values are used for parsing.
- If keep\_default\_na is False, and na\_values are specified, only the NaN values specified na\_values are used for parsing.
- If keep\_default\_na is False, and na\_values are not specified, no strings will be parsed as NaN.

Note that if `na_filter` is passed in as `False`, the `keep_default_na` and `na_values` parameters will be ignored.

**`na_filterbool`, default `True`**

Detect missing value markers (empty strings and the value of `na_values`). In data without any NAs, passing `na_filter=False` can improve the performance of reading a large file.

**`verbosebool`, default `False`**

Indicate number of NA values placed in non-numeric columns.

**`skip_blank_linesbool`, default `True`**

If `True`, skip over blank lines rather than interpreting as NaN values.

**`parse_datesbool` or list of int or names or list of lists or dict, default `False`**

The behavior is as follows:

- boolean. If `True` -> try parsing the index.
- list of int or names. e.g. If `[1, 2, 3]` -> try parsing columns 1, 2, 3 each as a separate date column.
- list of lists. e.g. If `[[1, 3]]` -> combine columns 1 and 3 and parse as a single date column.
- dict, e.g. `{'foo' : [1, 3]}` -> parse columns 1, 3 as date and call result 'foo'

If a column or index cannot be represented as an array of datetimes, say because of an unparseable value or a mixture of timezones, the column or index will be returned unaltered as an object data type. For non-standard datetime parsing, use `pd.to_datetime` after `pd.read_csv`. To parse an index or column with a mixture of timezones, specify `date_parser` to be a partially-applied `pandas.to_datetime()` with `utc=True`. See [Parsing a CSV with mixed timezones](#) for more.

Note: A fast-path exists for iso8601-formatted dates.

**`infer_datetime_formatbool`, default `False`**

If `True` and `parse_dates` is enabled, pandas will attempt to infer the format of the datetime strings in the columns, and if it can be inferred, switch to a faster method of parsing them. In some cases this can increase the parsing speed by 5-10x.

**`keep_date_colbool`, default `False`**

If `True` and `parse_dates` specifies combining multiple columns then keep the original columns.

**`date_parserfunction`, optional**

Function to use for converting a sequence of string columns to an array of datetime instances. The default uses `dateutil.parser.parser` to do the conversion. Pandas will try to call `date_parser` in three different ways, advancing to the next if an exception occurs: 1) Pass one or more arrays (as defined by `parse_dates`) as arguments; 2) concatenate (row-wise) the string values from the columns defined by `parse_dates` into a single array and pass that; and 3) call `date_parser` once for each row using one or more strings (corresponding to the columns defined by `parse_dates`) as arguments.

**`dayfirstbool`, default `False`**



DD/MM format dates, international and European format.

**cache\_datesbool, default True**

If True, use a cache of unique, converted dates to apply the datetime conversion. May produce significant speed-up when parsing duplicate date strings, especially ones with timezone offsets.

*New in version 0.25.0.*

**iteratorbool, default False**

Return TextFileReader object for iteration or getting chunks with [get\\_chunk\(\)](#).

**chunksizaint, optional**

Return TextFileReader object for iteration. See the [IO Tools docs](#) for more information on [iterator](#) and [chunksize](#).

**compression{'infer', 'gzip', 'bz2', 'zip', 'xz', None}, default 'infer'**

For on-the-fly decompression of on-disk data. If 'infer' and filepath\_or\_buffer is path-like, then detect compression from the following extensions: '.gz', '.bz2', '.zip', or '.xz' (otherwise no decompression). If using 'zip', the ZIP file must contain only one data file to be read in. Set to None for no decompression.

**thousandsstr, optional**

Thousands separator.

**decimalstr, default '.'**

Character to recognize as decimal point (e.g. use ',' for European data).

**lineterminatorstr (length 1), optional**

Character to break file into lines. Only valid with C parser.

**quotecharstr (length 1), optional**

The character used to denote the start and end of a quoted item. Quoted items can include the delimiter and it will be ignored.

**quotingint or csv.QUOTE\_\* instance, default 0**

Control field quoting behavior per [csv.QUOTE\\_\\*](#) constants. Use one of QUOTE\_MINIMAL (0), QUOTE\_ALL (1), QUOTE\_NONNUMERIC (2) or QUOTE\_NONE (3).

**doublequotebool, default True**

When quotechar is specified and quoting is not [QUOTE\\_NONE](#), indicate whether or not to interpret two consecutive quotechar elements INSIDE a field as a single [quotechar](#) element.

**escapecharstr (length 1), optional**

One-character string used to escape other characters.

**commentstr, optional**

Indicates remainder of line should not be parsed. If found at the beginning of a line, the line will be ignored altogether. This parameter must be a single character. Like empty lines (as long as `skip_blank_lines=True`), fully commented lines are ignored by the parameter header but not by skiprows. For example, if `comment='#'`, parsing `#empty\na,b,c\n1,2,3` with `header=0` will result in 'a,b,c' being treated as the header.

**encodingstr, optional**

Encoding to use for UTF when reading/writing (ex. 'utf-8'). [List of Python standard encodings](#) .

**dialectstr or csv.Dialect, optional**

If provided, this parameter will override values (default or not) for the following parameters: delimiter, doublequote, escapechar, skipinitialspace, quotechar, and quoting. If it is necessary to override values, a ParserWarning will be issued. See csv.Dialect documentation for more details.

**error\_bad\_linesbool, default True**

Lines with too many fields (e.g. a csv line with too many commas) will by default cause an exception to be raised, and no DataFrame will be returned. If False, then these "bad lines" will be dropped from the DataFrame that is returned.

**warn\_bad\_linesbool, default True**

If error\_bad\_lines is False, and warn\_bad\_lines is True, a warning for each "bad line" will be output.

**delim\_whitespacebool, default False**

Specifies whether or not whitespace (e.g. ' ' or '\t') will be used as the sep. Equivalent to setting `sep='\s+'`. If this option is set to True, nothing should be passed in for the `delimiter` parameter.

**low\_memorybool, default True**

Internally process the file in chunks, resulting in lower memory use while parsing, but possibly mixed type inference. To ensure no mixed types either set False, or specify the type with the dtype parameter. Note that the entire file is read into a single DataFrame regardless, use the chunksize or iterator parameter to return the data in chunks. (Only valid with C parser).

**memory\_mapbool, default False**

If a filepath is provided for filepath\_or\_buffer, map the file object directly onto memory and access the data directly from there. Using this option can improve performance because there is no longer any I/O overhead.

**float\_precisionstr, optional**

Specifies which converter the C engine should use for floating-point values. The options are None for the ordinary converter, high for the high-precision converter, and round\_trip for the round-trip converter.

**Returns**

**Dataframe or TextParser**

A comma-separated values (csv) file is returned as two-dimensional data structure with labeled axes.

```
from sklearn.model_selection import train_test_split
```

`train_test_split()`:

[LINK](#)

`sklearn.model_selection.train_test_split(*arrays, **options)`[\[source\]](#)

Split arrays or matrices into random train and test subsets

Quick utility that wraps input validation and `next(ShuffleSplit().split(X, y))` and application to input data into a single call for splitting (and optionally subsampling) data in a oneliner.

Read more in the [User Guide](#).

#### Parameters

***\*arrays****sequence of indexables with same length / shape[0]*

Allowed inputs are lists, numpy arrays, scipy-sparse matrices or pandas dataframes.

***test\_size****float or int, default=None*

If float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the test split. If int, represents the absolute number of test samples. If None, the value is set to the complement of the train size. If `train_size` is also None, it will be set to 0.25.

***train\_size****float or int, default=None*

If float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the train split. If int, represents the absolute number of train samples. If None, the value is automatically set to the complement of the test size.

***random\_state****int or RandomState instance, default=None*

Controls the shuffling applied to the data before applying the split. Pass an int for reproducible output across multiple function calls. See [Glossary](#).

***shuffle****bool, default=True*

Whether or not to shuffle the data before splitting. If `shuffle=False` then `stratify` must be None.

***stratify****array-like, default=None*

If not None, data is split in a stratified fashion, using this as the class labels.

#### Returns

***splittinglist***, *length=2 \* len(arrays)*

List containing train-test split of inputs.

*New in version 0.16:* If the input is sparse, the output will be a `scipy.sparse.csr_matrix`. Else, output type is the same as the input type.

#### Examples

```
>>>
>>> import numpy as np
>>> from sklearn.model_selection import train_test_split
>>> X, y = np.arange(10).reshape((5, 2)), range(5)
>>> X
array([[0, 1],
       [2, 3],
       [4, 5],
```

```

    [6, 7],
    [8, 9]])
>>> list(y)
[0, 1, 2, 3, 4]

>>>
>>> X_train, X_test, y_train, y_test = train_test_split(
...   X, y, test_size=0.33, random_state=42)
...
>>> X_train
array([[4, 5],
       [0, 1],
       [6, 7]])
>>> y_train
[2, 0, 3]
>>> X_test
array([[2, 3],
       [8, 9]])
>>> y_test
[1, 4]

>>>
>>> train_test_split(y, shuffle=False)
[[0, 1, 2], [3, 4]]

```

Import pandas as pd

pd.get\_dummies():

[LINK](#)

**pandas.get\_dummies**(data, prefix=None, prefix\_sep='\_', dummy\_na=False, columns=None, sparse=False, drop\_first=False, dtype=None)[\[source\]](#)

Convert categorical variable into dummy/indicator variables.

#### Parameters

**data**array-like, Series, or DataFrame

Data of which to get dummy indicators.

**prefix**str, list of str, or dict of str, default None

String to append DataFrame column names. Pass a list with length equal to the number of columns when calling get\_dummies on a DataFrame. Alternatively, prefix can be a dictionary mapping column names to prefixes.

**prefix\_sep**str, default '\_'

If appending prefix, separator/delimiter to use. Or pass a list or dictionary as with prefix.

**dummy\_na**bool, default False

Add a column to indicate NaNs, if False NaNs are ignored.

**columnslst-like, default None**

Column names in the DataFrame to be encoded. If columns is None then all the columns with object or category dtype will be converted.

**sparsebool, default False**

Whether the dummy-encoded columns should be backed by a **SparseArray** (True) or a regular NumPy array (False).

**drop\_firstbool, default False**

Whether to get k-1 dummies out of k categorical levels by removing the first level.

**dtype, default np.uint8**

Data type for new columns. Only a single dtype is allowed.

*New in version 0.23.0.*

**Returns**

**DataFrame**

Dummy-coded data.

import pandas as pd

pd.dataframe():

[LINK](#)

**class pandas.DataFrame(data=None, index=None, columns=None, dtype=None, copy=False)**[\[source\]](#)

Two-dimensional, size-mutable, potentially heterogeneous tabular data.

Data structure also contains labeled axes (rows and columns). Arithmetic operations align on both row and column labels. Can be thought of as a dict-like container for Series objects. The primary pandas data structure.

**Parameters**

**datandarray (structured or homogeneous), Iterable, dict, or DataFrame**

Dict can contain Series, arrays, constants, or list-like objects.

*Changed in version 0.23.0:* If data is a dict, column order follows insertion-order for Python 3.6 and later.

*Changed in version 0.25.0:* If data is a list of dicts, column order follows insertion-order for Python 3.6 and later.

**indexIndex or array-like**

Index to use for resulting frame. Will default to RangeIndex if no indexing information part of input

data and no index provided.

**columnsIndex or array-like**

Column labels to use for resulting frame. Will default to RangeIndex (0, 1, 2, ..., n) if no column labels are provided.

**dtype, default None**

Data type to force. Only a single dtype is allowed. If None, infer.

**copybool, default False**

Copy data from inputs. Only affects DataFrame / 2d ndarray input.

import pandas as pd

pd.concat

[LINK](#)

**pandas.concat**(objs: Union[Iterable['DataFrame'], Mapping[Label, 'DataFrame']], axis='0', join: str = "outer", ignore\_index: bool = 'False', keys='None', levels='None', names='None', verify\_integrity: bool = 'False', sort: bool = 'False', copy: bool = 'True') → 'DataFrame'[\[source\]](#)

**pandas.concat**(objs: Union[Iterable[FrameOrSeries], Mapping[Label, FrameOrSeries]], axis='0', join: str = "outer", ignore\_index: bool = 'False', keys='None', levels='None', names='None', verify\_integrity: bool = 'False', sort: bool = 'False', copy: bool = 'True') → FrameOrSeriesUnion

Concatenate pandas objects along a particular axis with optional set logic along the other axes.

Can also add a layer of hierarchical indexing on the concatenation axis, which may be useful if the labels are the same (or overlapping) on the passed axis number.

**Parameters**

**objs** a sequence or mapping of Series or DataFrame objects

If a mapping is passed, the sorted keys will be used as the keys argument, unless it is passed, in which case the values will be selected (see below). Any None objects will be dropped silently unless they are all None in which case a ValueError will be raised.

**axis**{0/'index', 1/'columns'}, default 0

The axis to concatenate along.

**join**{'inner', 'outer'}, default 'outer'

How to handle indexes on other axis (or axes).

**ignore\_index**bool, default False

If True, do not use the index values along the concatenation axis. The resulting axis will be labeled 0, ..., n - 1. This is useful if you are concatenating objects where the concatenation axis does not have meaningful indexing information. Note the index values on the other axes are still respected in the

join.

**keysequence, default None**

If multiple levels passed, should contain tuples. Construct hierarchical index using the passed keys as the outermost level.

**levelsof sequences, default None**

Specific levels (unique values) to use for constructing a MultiIndex. Otherwise they will be inferred from the keys.

**nameslist, default None**

Names for the levels in the resulting hierarchical index.

**verify\_integritybool, default False**

Check whether the new concatenated axis contains duplicates. This can be very expensive relative to the actual data concatenation.

**sortbool, default False**

Sort non-concatenation axis if it is not already aligned when join is 'outer'. This has no effect when `join='inner'`, which already preserves the order of the non-concatenation axis.

*New in version 0.23.0.*

*Changed in version 1.0.0:* Changed to not sort by default.

**copybool, default True**

If False, do not copy data unnecessarily.

**Returns**

**object, type of objs**

When concatenating all `Series` along the index (`axis=0`), a `Series` is returned. When `objs` contains at least one `DataFrame`, a `DataFrame` is returned. When concatenating along the columns (`axis=1`), a `DataFrame` is returned.

Import pandas as pd

pd.drop

[LINK](#)

**DataFrame.drop(labels=None, axis=0, index=None, columns=None, level=None, inplace=False, errors='raise')[source]**

Drop specified labels from rows or columns.

Remove rows or columns by specifying label names and corresponding axis, or by specifying directly index or column names. When using a multi-index, labels on different levels can be removed by specifying the level.

## Parameters

### labelssingle label or list-like

Index or column labels to drop.

### axis{0 or 'index', 1 or 'columns'}, default 0

Whether to drop labels from the index (0 or 'index') or columns (1 or 'columns').

### indexsingle label or list-like

Alternative to specifying axis (**labels**, **axis=0** is equivalent to **index=labels**).

### columnssingle label or list-like

Alternative to specifying axis (**labels**, **axis=1** is equivalent to **columns=labels**).

### levelint or level name, optional

For MultiIndex, level from which the labels will be removed.

### inplacebool, default False

If False, return a copy. Otherwise, do operation inplace and return None.

### errors{'ignore', 'raise'}, default 'raise'

If 'ignore', suppress error and only existing labels are dropped.

## Returns

### DataFrame

DataFrame without the removed index or column labels.

## Raises

### KeyError

If any of the labels is not found in the selected axis.

```
from sklearn.preprocessing import StandardScaler
```

StandardScaler():

[LINK](#)

```
class sklearn.preprocessing.StandardScaler(*, copy=True, with_mean=True, with_std=True)[source]
```

Standardize features by removing the mean and scaling to unit variance

The standard score of a sample  $x$  is calculated as:



$$z = (x - u) / s$$

where  $u$  is the mean of the training samples or zero if `with_mean=False`, and  $s$  is the standard deviation of the training samples or one if `with_std=False`.

Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using [transform](#).

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

For instance many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

This scaler can also be applied to sparse CSR or CSC matrices by passing `with_mean=False` to avoid breaking the sparsity structure of the data.

Read more in the [User Guide](#).

### Parameters

**`copy`***boolean, optional, default True*

If False, try to avoid a copy and do inplace scaling instead. This is not guaranteed to always work inplace; e.g. if the data is not a NumPy array or scipy.sparse CSR matrix, a copy may still be returned.

**`with_mean`***boolean, True by default*

If True, center the data before scaling. This does not work (and will raise an exception) when attempted on sparse matrices, because centering them entails building a dense matrix which in common use cases is likely to be too large to fit in memory.

**`with_std`***boolean, True by default*

If True, scale the data to unit variance (or equivalently, unit standard deviation).

### Attributes

**`scale`***ndarray or None, shape (n\_features,)*

Per feature relative scaling of the data. This is calculated using `np.sqrt(var_)`. Equal to `None` when `with_std=False`.

*New in version 0.17: `scale_`*

**`mean`***ndarray or None, shape (n\_features,)*

The mean value for each feature in the training set. Equal to `None` when `with_mean=False`.

**`var`***ndarray or None, shape (n\_features,)*

The variance for each feature in the training set. Used to compute `scale_`. Equal to `None` when `with_std=False`.

**`n_samples_seen`***int or array, shape (n\_features,)*

The number of samples processed by the estimator for each feature. If there are not missing samples, the `n_samples_seen` will be an integer, otherwise it will be an array. Will be reset on new calls to `fit`, but increments across `partial_fit` calls.

From sklearn.svm import SVC

SVC():

[LINK](#)

```
class sklearn.svm.SVC(*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True,
probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1,
decision_function_shape='ovr', break_ties=False, random_state=None)[source]
```

C-Support Vector Classification.

The implementation is based on libsvm. The fit time scales at least quadratically with the number of samples and may be impractical beyond tens of thousands of samples. For large datasets consider using [sklearn.svm.LinearSVC](#) or [sklearn.linear\\_model.SGDClassifier](#) instead, possibly after a [sklearn.kernel\\_approximation.Nystroem](#) transformer.

The multiclass support is handled according to a one-vs-one scheme.

For details on the precise mathematical formulation of the provided kernel functions and how `gamma`, `coef0` and `degree` affect each other, see the corresponding section in the narrative documentation: [Kernel functions](#).

Read more in the [User Guide](#).

#### Parameters

**`Cfloat, default=1.0`**

Regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty.

**`kernel{'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'}, default='rbf'`**

Specifies the kernel type to be used in the algorithm. It must be one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' or a callable. If none is given, 'rbf' will be used. If a callable is given it is used to pre-compute the kernel matrix from data matrices; that matrix should be an array of shape (n\_samples, n\_samples).

**`degreeint, default=3`**

Degree of the polynomial kernel function ('poly'). Ignored by all other kernels.

**`gamma{'scale', 'auto'} or float, default='scale'`**

Kernel coefficient for 'rbf', 'poly' and 'sigmoid'.

- if `gamma='scale'` (default) is passed then it uses  $1 / (n\_features * X.var())$  as value of gamma,
- if 'auto', uses  $1 / n\_features$ .

*Changed in version 0.22:* The default value of `gamma` changed from 'auto' to 'scale'.

**`coef0float, default=0.0`**

Independent term in kernel function. It is only significant in 'poly' and 'sigmoid'.

**`shrinkingbool, default=True`**

Whether to use the shrinking heuristic. See the [User Guide](#).

**probabilitybool, default=False**

Whether to enable probability estimates. This must be enabled prior to calling `fit`, will slow down that method as it internally uses 5-fold cross-validation, and `predict_proba` may be inconsistent with `predict`. Read more in the [User Guide](#).

**tolfloat, default=1e-3**

Tolerance for stopping criterion.

**cache\_sizefloat, default=200**

Specify the size of the kernel cache (in MB).

**class\_weightdict or 'balanced', default=None**

Set the parameter C of class i to `class_weight[i]*C` for SVC. If not given, all classes are supposed to have weight one. The “balanced” mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as  $n_{\text{samples}} / (n_{\text{classes}} * \text{np.bincount}(y))$

**verbosebool, default=False**

Enable verbose output. Note that this setting takes advantage of a per-process runtime setting in libsvm that, if enabled, may not work properly in a multithreaded context.

**max\_iterint, default=-1**

Hard limit on iterations within solver, or -1 for no limit.

**decision\_function\_shape{'ovo', 'ovr'}, default='ovr'**

Whether to return a one-vs-rest (‘ovr’) decision function of shape (n\_samples, n\_classes) as all other classifiers, or the original one-vs-one (‘ovo’) decision function of libsvm which has shape (n\_samples, n\_classes \* (n\_classes - 1) / 2). However, one-vs-one (‘ovo’) is always used as multi-class strategy. The parameter is ignored for binary classification.

*Changed in version 0.19:* `decision_function_shape` is ‘ovr’ by default.

*New in version 0.17:* `decision_function_shape='ovr'` is recommended.

*Changed in version 0.17:* Deprecated `decision_function_shape='ovo'` and `None`.

**break\_tiesbool, default=False**

If true, `decision_function_shape='ovr'`, and number of classes > 2, `predict` will break ties according to the confidence values of [decision\\_function](#); otherwise the first class among the tied classes is returned. Please note that breaking ties comes at a relatively high computational cost compared to a simple `predict`.

*New in version 0.22.*

**random\_stateint or RandomState instance, default=None**

Controls the pseudo random number generation for shuffling the data for probability estimates. Ignored when `probability` is False. Pass an int for reproducible output across multiple function calls. See [Glossary](#).

**Attributes**

**support\_ndarray of shape (n\_SV,)**

Indices of support vectors.

**support\_vectors\_ndarray of shape (n\_SV, n\_features)**

Support vectors.

**n\_support\_ndarray of shape (n\_class,), dtype=int32**

Number of support vectors for each class.

**dual\_coef\_ ndarray of shape (n\_class-1, n\_SV)**

Dual coefficients of the support vector in the decision function (see [Mathematical formulation](#)), multiplied by their targets. For multiclass, coefficient for all 1-vs-1 classifiers. The layout of the coefficients in the multiclass case is somewhat non-trivial. See the [multi-class section of the User Guide](#) for details.

**coef\_ ndarray of shape (n\_class \* (n\_class-1) / 2, n\_features)**

Weights assigned to the features (coefficients in the primal problem). This is only available in the case of a linear kernel.

`coef_` is a readonly property derived from `dual_coef_` and `support_vectors_`.

**intercept\_ ndarray of shape (n\_class \* (n\_class-1) / 2,)**

Constants in decision function.

**fit\_status\_int**

0 if correctly fitted, 1 otherwise (will raise warning)

**classes\_ ndarray of shape (n\_classes,)**

The classes labels.

**probA\_ ndarray of shape (n\_class \* (n\_class-1) / 2)**

**probB\_ ndarray of shape (n\_class \* (n\_class-1) / 2)**

If `probability=True`, it corresponds to the parameters learned in Platt scaling to produce probability estimates from decision values. If `probability=False`, it's an empty array. Platt scaling uses the logistic function  $1 / (1 + \exp(\text{decision\_value} * \text{probA\_} + \text{probB\_}))$  where `probA_` and `probB_` are learned from the dataset [2]. For more information on the multiclass case and training procedure see section 8 of [1].

**class\_weight\_ ndarray of shape (n\_class,)**

Multipliers of parameter C for each class. Computed based on the `class_weight` parameter.

**shape\_fit\_tuple of int of shape (n\_dimensions\_of\_X,)**

Array dimensions of training vector X.

```
from sklearn.metrics import confusion_matrix, accuracy_score
```

`confusion_matrix()`:

[LINK](#)

```
sklearn.metrics.confusion_matrix(y_true, y_pred, *, labels=None, sample_weight=None,
normalize=None)[source]
```

Compute confusion matrix to evaluate the accuracy of a classification.

By definition a confusion matrix

C

is such that

$C_{i,j}$

is equal to the number of observations known to be in group

$i$

and predicted to be in group

$j$

.

Thus in binary classification, the count of true negatives is

$C_{0,0}$

, false negatives is

$C_{1,0}$

, true positives is

$C_{1,1}$

and false positives is

$C_{0,1}$

.

Read more in the [User Guide](#).

### Parameters

***y\_true***array-like of shape *(n\_samples,)*

Ground truth (correct) target values.

***y\_pred***array-like of shape *(n\_samples,)*

Estimated targets as returned by a classifier.

***labels***array-like of shape *(n\_classes,)*, *default=None*

List of labels to index the matrix. This may be used to reorder or select a subset of labels. If `None` is given, those that appear at least once in `y_true` or `y_pred` are used in sorted order.

***sample\_weight***array-like of shape *(n\_samples,)*, *default=None*

Sample weights.

*New in version 0.18.*

***normalize***{*'true'*, *'pred'*, *'all'*}, *default=None*

Normalizes confusion matrix over the true (rows), predicted (columns) conditions or all the population. If `None`, confusion matrix will not be normalized.

### Returns

***C***ndarray of shape *(n\_classes, n\_classes)*

Confusion matrix whose i-th row and j-th column entry indicates the number of samples with true label being i-th class and predicted label being j-th class.

`accuracy_score()`:

[LINK](#)

`sklearn.metrics.accuracy_score(y_true, y_pred, *, normalize=True, sample_weight=None)`[\[source\]](#)

Accuracy classification score.

In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must *exactly* match the corresponding set of labels in `y_true`.

Read more in the [User Guide](#).

#### Parameters

**`y_true`***Id array-like, or label indicator array / sparse matrix*

Ground truth (correct) labels.

**`y_pred`***Id array-like, or label indicator array / sparse matrix*

Predicted labels, as returned by a classifier.

**`normalize`***bool, optional (default=True)*

If `False`, return the number of correctly classified samples. Otherwise, return the fraction of correctly classified samples.

**`sample_weight`***array-like of shape (n\_samples,), default=None*

Sample weights.

#### Returns

**`score`***float*

If `normalize == True`, return the fraction of correctly classified samples (float), else returns the number of correctly classified samples (int).

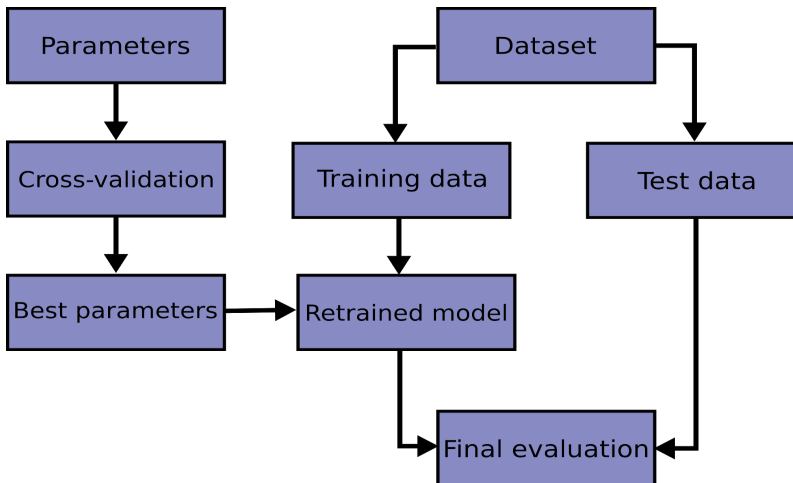
The best performance is 1 with `normalize == True` and the number of samples with `normalize == False`.

```
from sklearn.model_selection import cross_val_score
```

`cross_val_score()`:

[LINK](#)

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called **overfitting**. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a **test set**  $X_{\text{test}}, y_{\text{test}}$ . Note that the word “experiment” is not intended to denote academic use only, because even in commercial settings machine learning usually starts out experimentally. Here is a flowchart of typical cross validation workflow in model training. The best parameters can be determined by [grid search](#) techniques.



In scikit-learn a random split into training and test sets can be quickly computed with the [train\\_test\\_split](#) helper function. Let's load the iris data set to fit a linear support vector machine on it:

```
>>>
>>> import numpy as np
>>> from sklearn.model_selection import train_test_split
>>> from sklearn import datasets
>>> from sklearn import svm

>>> X, y = datasets.load_iris(return_X_y=True)
>>> X.shape, y.shape
((150, 4), (150,))
```

We can now quickly sample a training set while holding out 40% of the data for testing (evaluating) our classifier:

```
>>>
>>> X_train, X_test, y_train, y_test = train_test_split(
...     X, y, test_size=0.4, random_state=0)

>>> X_train.shape, y_train.shape
((90, 4), (90,))
>>> X_test.shape, y_test.shape
((60, 4), (60,))

>>> clf = svm.SVC(kernel='linear', C=1).fit(X_train, y_train)
>>> clf.score(X_test, y_test)
0.96...
```

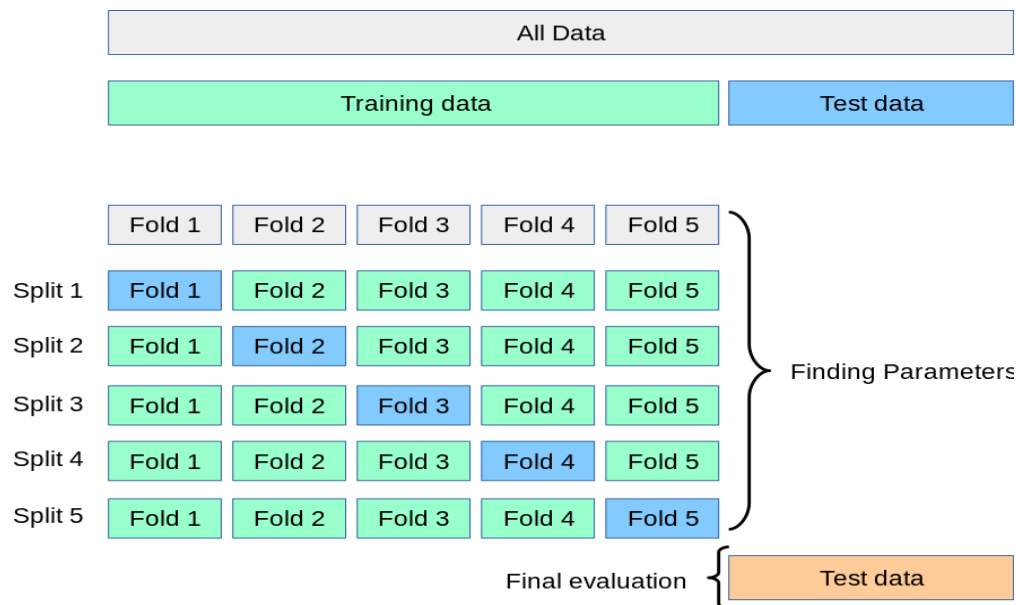
When evaluating different settings (“hyperparameters”) for estimators, such as the `C` setting that must be manually set for an SVM, there is still a risk of overfitting *on the test set* because the parameters can be tweaked until the estimator performs optimally. This way, knowledge about the test set can “leak” into the model and evaluation metrics no longer report on generalization performance. To solve this problem, yet another part of the dataset can be held out as a so-called “validation set”: training proceeds on the training set, after which evaluation is done on the validation set, and when the experiment seems to be successful, final evaluation can be done on the test set.

However, by partitioning the available data into three sets, we drastically reduce the number of samples which can be used for learning the model, and the results can depend on a particular random choice for the pair of (train, validation) sets.

A solution to this problem is a procedure called **cross-validation** (CV for short). A test set should still be held out for final evaluation, but the validation set is no longer needed when doing CV. In the basic approach, called *k*-fold CV, the training set is split into *k* smaller sets (other approaches are described below, but generally follow the same principles). The following procedure is followed for each of the *k* “folds”:

- A model is trained using
- $k-1$
- of the folds as training data;
- the resulting model is validated on the remaining part of the data (i.e., it is used as a test set to compute a performance measure such as accuracy).

The performance measure reported by *k*-fold cross-validation is then the average of the values computed in the loop. This approach can be computationally expensive, but does not waste too much data (as is the case when fixing an arbitrary validation set), which is a major advantage in problems such as inverse inference where the number of samples is very small.



### 3.1.1. Computing cross-validated metrics

The simplest way to use cross-validation is to call the `cross_val_score` helper function on the estimator and the dataset.



The following example demonstrates how to estimate the accuracy of a linear kernel support vector machine on the iris dataset by splitting the data, fitting a model and computing the score 5 consecutive times (with different splits each time):

```
>>>
>>> from sklearn.model_selection import cross_val_score
>>> clf = svm.SVC(kernel='linear', C=1)
>>> scores = cross_val_score(clf, X, y, cv=5)
>>> scores
array([0.96..., 1. ..., 0.96..., 0.96..., 1. ...])
```

The mean score and the 95% confidence interval of the score estimate are hence given by:

```
>>>
>>> print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
Accuracy: 0.98 (+/- 0.03)
```

By default, the score computed at each CV iteration is the `score` method of the estimator. It is possible to change this by using the `scoring` parameter:

```
>>>
>>> from sklearn import metrics
>>> scores = cross_val_score(
...     clf, X, y, cv=5, scoring='f1_macro')
>>> scores
array([0.96..., 1. ..., 0.96..., 0.96..., 1. ...])
```

See [The scoring parameter: defining model evaluation rules](#) for details. In the case of the Iris dataset, the samples are balanced across target classes hence the accuracy and the F1-score are almost equal.

When the `cv` argument is an integer, `cross_val_score` uses the **KFold** or **StratifiedKFold** strategies by default, the latter being used if the estimator derives from **ClassifierMixin**.

It is also possible to use other cross validation strategies by passing a cross validation iterator instead, for instance:

```
>>>
>>> from sklearn.model_selection import ShuffleSplit
>>> n_samples = X.shape[0]
>>> cv = ShuffleSplit(n_splits=5, test_size=0.3, random_state=0)
>>> cross_val_score(clf, X, y, cv=cv)
array([0.977..., 0.977..., 1. ..., 0.955..., 1. ...])
```

Another option is to use an iterable yielding (train, test) splits as arrays of indices, for example:

```
>>>
>>> def custom_cv_2folds(X):
...     n = X.shape[0]
...     i = 1
...     while i <= 2:
...         idx = np.arange(n * (i - 1) / 2, n * i / 2, dtype=int)
...         yield idx, idx
...         i += 1
...
>>> custom_cv = custom_cv_2folds(X)
```

```
>>> cross_val_score(clf, X, y, cv=custom_cv)
array([1.         , 0.973...])
```

### Data transformation with held out data

Just as it is important to test a predictor on data held-out from training, preprocessing (such as standardization, feature selection, etc.) and similar [data transformations](#) similarly should be learnt from a training set and applied to held-out data for prediction:

```
>>>
>>> from sklearn import preprocessing
>>> X_train, X_test, y_train, y_test = train_test_split(
...   X, y, test_size=0.4, random_state=0)
>>> scaler = preprocessing.StandardScaler().fit(X_train)
>>> X_train_transformed = scaler.transform(X_train)
>>> clf = svm.SVC(C=1).fit(X_train_transformed, y_train)
>>> X_test_transformed = scaler.transform(X_test)
>>> clf.score(X_test_transformed, y_test)
0.9333...
```

A [Pipeline](#) makes it easier to compose estimators, providing this behavior under cross-validation:

```
>>>
>>> from sklearn.pipeline import make_pipeline
>>> clf = make_pipeline(preprocessing.StandardScaler(), svm.SVC(C=1))
>>> cross_val_score(clf, X, y, cv=cv)
array([0.977..., 0.933..., 0.955..., 0.933..., 0.977...])
```

See [Pipelines and composite estimators](#).

#### 3.1.1.1. The cross\_validate function and multiple metric evaluation

The [cross\\_validate](#) function differs from [cross\\_val\\_score](#) in two ways:

- It allows specifying multiple metrics for evaluation.
- It returns a dict containing fit-times, score-times (and optionally training scores as well as fitted estimators) in addition to the test score.

For single metric evaluation, where the scoring parameter is a string, callable or None, the keys will be - ['test\_score', 'fit\_time', 'score\_time']

And for multiple metric evaluation, the return value is a dict with the following keys - ['test\_<scorer1\_name>', 'test\_<scorer2\_name>', 'test\_<scorer...>', 'fit\_time', 'score\_time']

`return_train_score` is set to `False` by default to save computation time. To evaluate the scores on the training set as well you need to be set to `True`.

You may also retain the estimator fitted on each training set by setting `return_estimator=True`.

The multiple metrics can be specified either as a list, tuple or set of predefined scorer names:

```
>>>
>>> from sklearn.model_selection import cross_validate
>>> from sklearn.metrics import recall_score
>>> scoring = ['precision_macro', 'recall_macro']
>>> clf = svm.SVC(kernel='linear', C=1, random_state=0)
>>> scores = cross_validate(clf, X, y, scoring=scoring)
```

```
>>> sorted(scores.keys())
['fit_time', 'score_time', 'test_precision_macro', 'test_recall_macro']
>>> scores['test_recall_macro']
array([0.96..., 1. ..., 0.96..., 0.96..., 1. ...])
```

Or as a dict mapping scorer name to a predefined or custom scoring function:

```
>>>
>>> from sklearn.metrics import make_scorer
>>> scoring = {'prec_macro': 'precision_macro',
...           'rec_macro': make_scorer(recall_score, average='macro')}
>>> scores = cross_validate(clf, X, y, scoring=scoring,
...                          cv=5, return_train_score=True)
>>> sorted(scores.keys())
['fit_time', 'score_time', 'test_prec_macro', 'test_rec_macro',
 'train_prec_macro', 'train_rec_macro']
>>> scores['train_rec_macro']
array([0.97..., 0.97..., 0.99..., 0.98..., 0.98...])
```

Here is an example of `cross_validate` using a single metric:

```
>>>
>>> scores = cross_validate(clf, X, y,
...                          scoring='precision_macro', cv=5,
...                          return_estimator=True)
>>> sorted(scores.keys())
['estimator', 'fit_time', 'score_time', 'test_score']
```

### 3.1.1.2. Obtaining predictions by cross-validation

The function `cross_val_predict` has a similar interface to `cross_val_score`, but returns, for each element in the input, the prediction that was obtained for that element when it was in the test set. Only cross-validation strategies that assign all elements to a test set exactly once can be used (otherwise, an exception is raised).

#### **Warning** Note on inappropriate usage of `cross_val_predict`

The result of `cross_val_predict` may be different from those obtained using `cross_val_score` as the elements are grouped in different ways. The function `cross_val_score` takes an average over cross-validation folds, whereas `cross_val_predict` simply returns the labels (or probabilities) from several distinct models undistinguished. Thus, `cross_val_predict` is not an appropriate measure of generalisation error.

#### **The function `cross_val_predict` is appropriate for:**

- Visualization of predictions obtained from different models.
- Model blending: When predictions of one supervised estimator are used to train another estimator in ensemble methods.

The available cross validation iterators are introduced in the following section.

### **Examples**

- [Receiver Operating Characteristic \(ROC\) with cross validation](#),
- [Recursive feature elimination with cross-validation](#),
- [Parameter estimation using grid search with cross-validation](#),
- [Sample pipeline for text feature extraction and evaluation](#),
- [Plotting Cross-Validated Predictions](#),
- [Nested versus non-nested cross-validation](#).

### 3.1.2. Cross validation iterators

The following sections list utilities to generate indices that can be used to generate dataset splits according to different cross validation strategies.

#### 3.1.2.1. Cross-validation iterators for i.i.d. data

Assuming that some data is Independent and Identically Distributed (i.i.d.) is making the assumption that all samples stem from the same generative process and that the generative process is assumed to have no memory of past generated samples.

The following cross-validators can be used in such cases.

#### NOTE

While i.i.d. data is a common assumption in machine learning theory, it rarely holds in practice. If one knows that the samples have been generated using a time-dependent process, it is safer to use a [time-series aware cross-validation scheme](#). Similarly, if we know that the generative process has a group structure (samples collected from different subjects, experiments, measurement devices), it is safer to use [group-wise cross-validation](#).

##### 3.1.2.1.1. K-fold

**KFold** divides all the samples in

$k$

groups of samples, called folds (if

$k=n$

, this is equivalent to the *Leave One Out* strategy), of equal sizes (if possible). The prediction function is learned using

$k-1$

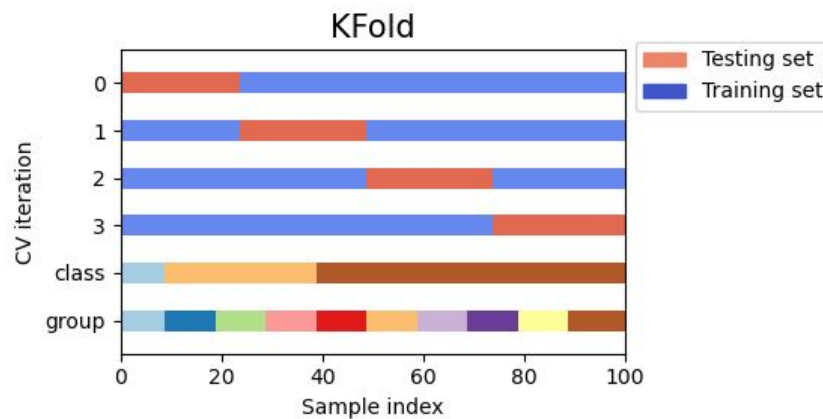
folds, and the fold left out is used for test.

Example of 2-fold cross-validation on a dataset with 4 samples:

```
>>>
>>> import numpy as np
>>> from sklearn.model_selection import KFold

>>> X = ["a", "b", "c", "d"]
>>> kf = KFold(n_splits=2)
>>> for train, test in kf.split(X):
...     print("%s %s" % (train, test))
[2 3] [0 1]
[0 1] [2 3]
```

Here is a visualization of the cross-validation behavior. Note that **KFold** is not affected by classes or groups.



Each fold is constituted by two arrays: the first one is related to the *training set*, and the second one to the *test set*. Thus, one can create the training/test sets using numpy indexing:

```
>>>
>>> X = np.array([[0., 0.], [1., 1.], [-1., -1.], [2., 2.]])
>>> y = np.array([0, 1, 0, 1])
>>> X_train, X_test, y_train, y_test = X[train], X[test], y[train], y[test]
```

#### 3.1.2.1.2. Repeated K-Fold

**RepeatedKFold** repeats K-Fold n times. It can be used when one requires to run **KFold** n times, producing different splits in each repetition.

Example of 2-fold K-Fold repeated 2 times:

```
>>>
>>> import numpy as np
>>> from sklearn.model_selection import RepeatedKFold
>>> X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])
>>> random_state = 12883823
>>> rkf = RepeatedKFold(n_splits=2, n_repeats=2, random_state=random_state)
>>> for train, test in rkf.split(X):
...     print("%s %s" % (train, test))
...
[2 3] [0 1]
[0 1] [2 3]
[0 2] [1 3]
[1 3] [0 2]
```

Similarly, **RepeatedStratifiedKFold** repeats Stratified K-Fold n times with different randomization in each repetition.

#### 3.1.2.1.3. Leave One Out (LOO)

**LeaveOneOut** (or LOO) is a simple cross-validation. Each learning set is created by taking all the samples except one, the test set being the sample left out. Thus, for

n

samples, we have

n

different training sets and

n

different tests set. This cross-validation procedure does not waste much data as only one sample is removed from the training set:

```
>>>
>>> from sklearn.model_selection import LeaveOneOut

>>> X = [1, 2, 3, 4]
>>> loo = LeaveOneOut()
>>> for train, test in loo.split(X):
...     print("%s %s" % (train, test))
[1 2 3] [0]
[0 2 3] [1]
[0 1 3] [2]
[0 1 2] [3]
```

Potential users of LOO for model selection should weigh a few known caveats. When compared with

k

-fold cross validation, one builds

n

models from

n

samples instead of

k

models, where

$n > k$

. Moreover, each is trained on

$n-1$

samples rather than

$(k-1)n/k$

. In both ways, assuming

k

is not too large and

$k < n$

, LOO is more computationally expensive than

$k$

-fold cross validation.

In terms of accuracy, LOO often results in high variance as an estimator for the test error. Intuitively, since

$n-1$

of the

$n$

samples are used to build each model, models constructed from folds are virtually identical to each other and to the model built from the entire training set.

However, if the learning curve is steep for the training size in question, then 5- or 10- fold cross validation can overestimate the generalization error.

As a general rule, most authors, and empirical evidence, suggest that 5- or 10- fold cross validation should be preferred to LOO.

from sklearn.ensemble import RandomForestClassifier

RandomForestClassifier():

[LINK](#)

```
class sklearn.ensemble.RandomForestClassifier(n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None)[source]
```

A random forest classifier.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the `max_samples` parameter if `bootstrap=True` (default), otherwise the whole dataset is used to build each tree.

Read more in the [User Guide](#).

#### Parameters

**n\_estimators***int, default=100*

The number of trees in the forest.

*Changed in version 0.22:* The default value of `n_estimators` changed from 10 to 100 in 0.22.

**criterion**{“gini”, “entropy”}, default=“gini”

The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. Note: this parameter is tree-specific.

**max\_depth**int, default=None

The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min\_samples\_split samples.

**min\_samples\_split**int or float, default=2

The minimum number of samples required to split an internal node:

- If int, then consider min\_samples\_split as the minimum number.
- If float, then min\_samples\_split is a fraction and  $\text{ceil}(\text{min\_samples\_split} * n\_samples)$  are the minimum number of samples for each split.

*Changed in version 0.18:* Added float values for fractions.

**min\_samples\_leaf**int or float, default=1

The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min\_samples\_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.

- If int, then consider min\_samples\_leaf as the minimum number.
- If float, then min\_samples\_leaf is a fraction and  $\text{ceil}(\text{min\_samples\_leaf} * n\_samples)$  are the minimum number of samples for each node.

*Changed in version 0.18:* Added float values for fractions.

**min\_weight\_fraction\_leaf**float, default=0.0

The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Samples have equal weight when sample\_weight is not provided.

**max\_features**{“auto”, “sqrt”, “log2”}, int or float, default=“auto”

The number of features to consider when looking for the best split:

- If int, then consider max\_features features at each split.
- If float, then max\_features is a fraction and  $\text{int}(\text{max\_features} * n\_features)$  features are considered at each split.
- If “auto”, then  $\text{max\_features} = \text{sqrt}(n\_features)$ .
- If “sqrt”, then  $\text{max\_features} = \text{sqrt}(n\_features)$  (same as “auto”).
- If “log2”, then  $\text{max\_features} = \text{log2}(n\_features)$ .
- If None, then  $\text{max\_features} = n\_features$ .

Note: the search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max\_features features.

**max\_leaf\_nodes**int, default=None

Grow trees with max\_leaf\_nodes in best-first fashion. Best nodes are defined as relative reduction in impurity. If None then unlimited number of leaf nodes.

**min\_impurity\_decrease**float, default=0.0

A node will be split if this split induces a decrease of the impurity greater than or equal to this value.

The weighted impurity decrease equation is the following:

$$N_t / N * (\text{impurity} - N_{t\_R} / N_t * \text{right\_impurity})$$



$- N_{t\_L} / N_t * \text{left\_impurity}$ )

where  $N$  is the total number of samples,  $N_t$  is the number of samples at the current node,  $N_{t\_L}$  is the number of samples in the left child, and  $N_{t\_R}$  is the number of samples in the right child.

$N$ ,  $N_t$ ,  $N_{t\_R}$  and  $N_{t\_L}$  all refer to the weighted sum, if `sample_weight` is passed.

*New in version 0.19.*

**min\_impurity\_split***float, default=None*

Threshold for early stopping in tree growth. A node will split if its impurity is above the threshold, otherwise it is a leaf.

*Deprecated since version 0.19: min\_impurity\_split has been deprecated in favor of min\_impurity\_decrease in 0.19. The default value of min\_impurity\_split has changed from 1e-7 to 0 in 0.23 and it will be removed in 0.25. Use min\_impurity\_decrease instead.*

**bootstrap***bool, default=True*

Whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree.

**oob\_score***bool, default=False*

Whether to use out-of-bag samples to estimate the generalization accuracy.

**n\_jobs***int, default=None*

The number of jobs to run in parallel. `fit`, `predict`, `decision_path` and `apply` are all parallelized over the trees. None means 1 unless in a `joblib.parallel_backend` context. -1 means using all processors. See [Glossary](#) for more details.

**random\_state***int or RandomState, default=None*

Controls both the randomness of the bootstrapping of the samples used when building trees (if `bootstrap=True`) and the sampling of the features to consider when looking for the best split at each node (if `max_features < n_features`). See [Glossary](#) for details.

**verbose***int, default=0*

Controls the verbosity when fitting and predicting.

**warm\_start***bool, default=False*

When set to True, reuse the solution of the previous call to fit and add more estimators to the ensemble, otherwise, just fit a whole new forest. See [the Glossary](#).

**class\_weight***{“balanced”, “balanced\_subsample”, dict or list of dicts, default=None*

Weights associated with classes in the form `{class_label: weight}`. If not given, all classes are supposed to have weight one. For multi-output problems, a list of dicts can be provided in the same order as the columns of `y`.

Note that for multioutput (including multilabel) weights should be defined for each class of every column in its own dict. For example, for four-class multilabel classification weights should be `[{0: 1, 1: 1}, {0: 1, 1: 5}, {0: 1, 1: 1}, {0: 1, 1: 1}]` instead of `[{1:1}, {2:5}, {3:1}, {4:1}]`.

The “balanced” mode uses the values of `y` to automatically adjust weights inversely proportional to class frequencies in the input data as `n_samples / (n_classes * np.bincount(y))`

The “balanced\_subsample” mode is the same as “balanced” except that weights are computed based on the bootstrap sample for every tree grown.

For multi-output, the weights of each column of `y` will be multiplied.

Note that these weights will be multiplied with `sample_weight` (passed through the fit method) if `sample_weight` is specified.

**`ccp_alpha` non-negative float, default=0.0**

Complexity parameter used for Minimal Cost-Complexity Pruning. The subtree with the largest cost complexity that is smaller than `ccp_alpha` will be chosen. By default, no pruning is performed. See [Minimal Cost-Complexity Pruning](#) for details.

*New in version 0.22.*

**`max_samples` int or float, default=None**

If bootstrap is True, the number of samples to draw from X to train each base estimator.

- If None (default), then draw `X.shape[0]` samples.
- If int, then draw `max_samples` samples.
- If float, then draw `max_samples * X.shape[0]` samples. Thus, `max_samples` should be in the interval (0, 1).

*New in version 0.22.*

**Attributes**

**`base_estimator` *DecisionTreeClassifier***

The child estimator template used to create the collection of fitted sub-estimators.

**`estimators_list` of *DecisionTreeClassifier***

The collection of fitted sub-estimators.

**`classes` ndarray of shape (n\_classes,) or a list of such arrays**

The classes labels (single output problem), or a list of arrays of class labels (multi-output problem).

**`n_classes` int or list**

The number of classes (single output problem), or a list containing the number of classes for each output (multi-output problem).

**`n_features` int**

The number of features when `fit` is performed.

**`n_outputs` int**

The number of outputs when `fit` is performed.

**`feature_importances` ndarray of shape (n\_features,)**

The impurity-based feature importances.

**`oob_score` float**

Score of the training dataset obtained using an out-of-bag estimate. This attribute exists only when `oob_score` is True.

**`oob_decision_function` ndarray of shape (n\_samples, n\_classes)**

Decision function computed with out-of-bag estimate on the training set. If `n_estimators` is small it might be possible that a data point was never left out during the bootstrap. In this case, `oob_decision_function` might contain NaN. This attribute exists only when `oob_score` is True.

