

# DATA SCIENCE EAST AFRICA

## INTRODUCTION TO MACHINE LEARNING & SCIKIT LEARN

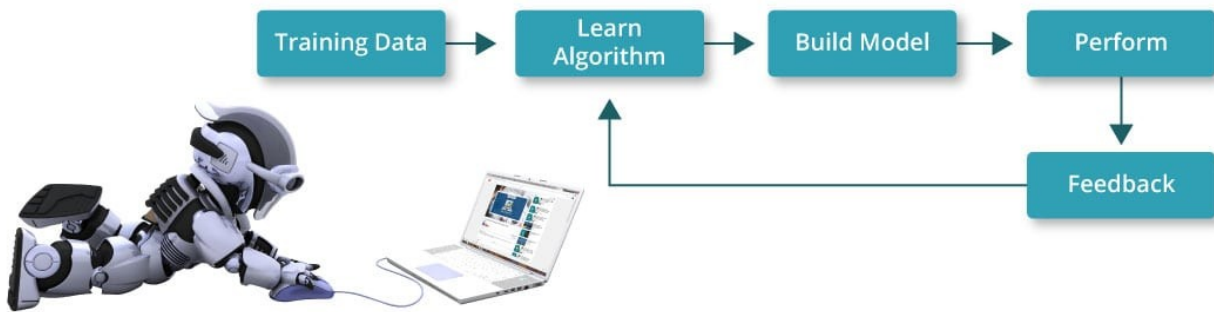
Day 14 /20

### What is Machine learning?

Machine learning is a type of artificial intelligence that allows software applications to learn from the data and become more accurate in predicting outcomes without human intervention.

But how does that happen? For that, the machine needs to be trained on some data and based on that, it will detect a pattern to create a model. This process of gaining knowledge from the data and providing powerful insights is all about machine learning.

Refer the below image to get a better understanding of its working:



Using the data, the system learns an algorithm and then uses it to build a predictive model. Later on, we adjust the model or we enhance the accuracy of the model using the feedback data. Using this feedback data, we tune the model and predict action on the new data set.

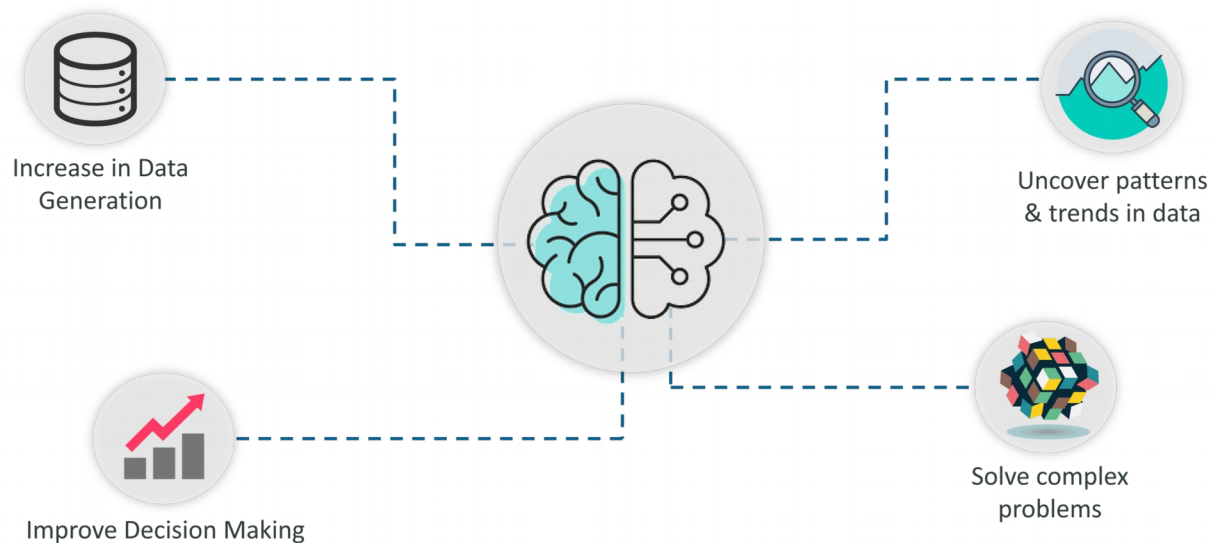
Reasons why Machine Learning is so important:

- Improve Decision Making:** By making use of various algorithms, Machine Learning can be used to make better business decisions. For example, Machine Learning is used to forecast sales, predict downfalls in the stock market, identify risks and anomalies, etc.
- Increase in Data Generation:** Due to excessive production of data, we need a method that can be used to structure, analyze and draw useful insights from data.

This is where Machine Learning comes in. It uses data to solve problems and find solutions to the most complex tasks faced by organizations.

- *Uncover patterns & trends in data:* Finding hidden patterns and extracting key insights from data is the most essential part of Machine Learning. By building predictive models and using statistical techniques, Machine Learning allows you to dig beneath the surface and explore the data at a minute scale. Understanding data and extracting patterns manually will take days, whereas Machine Learning algorithms can perform such computations in less than a second.

- *Solve complex problems:* From detecting the genes linked to the deadly ALS disease to building self-driving cars, Machine Learning can be used to solve the most complex problems.



## Machine Learning Definitions

**Algorithm:** A Machine Learning algorithm is a set of rules and statistical techniques used to learn patterns from data and draw significant information from it. It is the logic behind a Machine Learning model. An example of a Machine Learning algorithm is the Linear Regression algorithm.

**Model:** A model is the main component of Machine Learning. A model is trained by using a Machine Learning Algorithm. An algorithm maps all the decisions that a model is supposed to take based on the given input, in order to get the correct output.

**Predictor Variable:** It is a feature(s) of the data that can be used to predict the output.

**Response Variable:** It is the feature or the output variable that needs to be predicted by using the predictor variable(s).

**Training Data:** The Machine Learning model is built using the training data. The training data helps the model to identify key trends and patterns essential to predict the output.

**Testing Data:** After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.

### Machine Learning Process

The Machine Learning process involves building a Predictive model that can be used to find a solution for a Problem Statement. To understand the Machine Learning process let's assume that you have been given a problem that needs to be solved by using Machine Learning. The problem is to predict the occurrence of rain in your local area by using Machine Learning.

The below steps are followed in a Machine Learning process:

#### Step 1: Define the objective of the Problem Statement.

At this step, we must understand what exactly needs to be predicted. In our case, the objective is to predict the possibility of rain by studying weather conditions.

At this stage, it is also essential to take mental notes on what kind of data can be used to solve this problem or the type of approach you must follow to get to the solution.

## Step 2: Data Gathering

At this stage, you must be asking questions such as,

- What kind of data is needed to solve this problem?
- Is the data available?
- How can I get the data?

Once you know the types of data that is required, you must understand how you can derive this data. Data collection can be done manually or by web scraping. However, if you're a beginner and you're just looking to learn Machine Learning you don't have to worry about getting the data. There are 1000s of data resources on the web, you can just download the data set and get going.

Coming back to the problem at hand, the data needed for weather forecasting includes measures such as humidity level, temperature, pressure, locality, whether or not you live in a hill station, etc. Such data must be collected and stored for analysis.

### Step 3: Data Preparation

The data you collected is almost never in the right format. You will encounter a lot of inconsistencies in the data set such as missing values, redundant variables, duplicate values, etc. Removing such inconsistencies is very essential because they might lead to wrongful computations and predictions. Therefore, at this stage, you scan the data set for any inconsistencies and you fix them then and there.

### Step 4: Exploratory Data Analysis.

Grab your detective glasses because this stage is all about diving deep into data and finding all the hidden data mysteries. EDA or Exploratory Data Analysis is the brainstorming stage of Machine Learning. Data Exploration involves understanding the patterns and trends in the data. At this stage, all the useful insights are drawn and correlations between the variables are understood.

For example, in the case of predicting rainfall, we know that there is a strong possibility of rain if the temperature has fallen low. Such correlations must be understood and mapped at this stage.

### Step 5: Building a Machine Learning Model.

All the insights and patterns derived during Data Exploration are used to build the Machine Learning Model. This stage always begins by splitting the data set into two parts, training data, and testing data. The training data will be used to build and analyze the model. The logic of the model is based on the Machine Learning Algorithm that is being implemented.

In the case of predicting rainfall, since the output will be in the form of True (if it will rain tomorrow) or False (no rain tomorrow), we can use a classification Algorithm such as Logistic Regression.

Choosing the right algorithm depends on the type of problem you're trying to solve, the data set and the level of complexity of the problem. In the upcoming sections, we will discuss the different types of problems that can be solved by using Machine Learning.

### Step 6: Model Evaluation & Optimization.

After building a model by using the training data set, it is finally time to put the model to a test. The testing data set is used to check the efficiency of the model and how accurately it can predict the outcome.



Once the accuracy is calculated, any further improvements in the model can be implemented at this stage. Methods like parameter tuning and cross-validation can be used to improve the performance of the model.

### Step 7: Predictions.

Once the model is evaluated and improved, it is finally used to make predictions. The final output can be a Categorical variable (eg. True or False) or it can be a Continuous Quantity (eg. the predicted value of a stock).

In our case, for predicting the occurrence of rainfall, the output will be a categorical variable.

## Machine Learning Types

A machine can learn to solve a problem by following any one of the following three approaches. These are the ways in which a machine can learn:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

### Supervised Learning

Supervised learning is a technique in which we teach or train the machine using data which is well labeled.

## Unsupervised Learning

Unsupervised learning involves training by using unlabeled data and allowing the model to act on that information without guidance.

## Reinforcement Learning

Reinforcement Learning is a part of Machine learning where an agent is put in an environment and he learns to behave in this environment by performing certain actions and observing the rewards which it gets from those actions.

There are three main types of problems that can be solved in Machine Learning:

1. **Classification:** In this type, the output is a categorical value. Classifying emails into two classes, spam and non-spam is a classification problem that can be solved by using Supervised Learning classification algorithms such as Support Vector Machines, Naive Bayes, Logistic Regression, K Nearest Neighbor, etc.
2. **Regression:** In this type of problem the output is a continuous quantity. So, for example, if you want to predict the speed of a car given the distance, it is a Regression problem. Regression problems can be solved by using Supervised Learning algorithms like Linear Regression.

3. **Clustering**: This type of problem involves assigning the input into two or more clusters based on feature similarity. For example, clustering viewers into similar groups based on their interests, age, geography, etc can be done by using Unsupervised Learning algorithms like K-Means Clustering.

### Introduction to Scikit Learn

Scikit learn is a library used to perform machine learning in Python.

Scikit learn is an open source library which is licensed under BSD and is reusable in various contexts, encouraging academic and commercial use.

It provides a range of supervised and unsupervised learning algorithms in Python. Scikit learn consists popular algorithms and libraries. Apart from that, it also contains the following packages:

- NumPy
- Matplotlib
- SciPy (Scientific Python)

To implement Scikit learn, we first need to import the above packages.

You can download these two packages using the command line or if you are using PyCharm, you can directly install it by going to your setting in the same way you do it for other packages (They come pre installed in the conda environment).

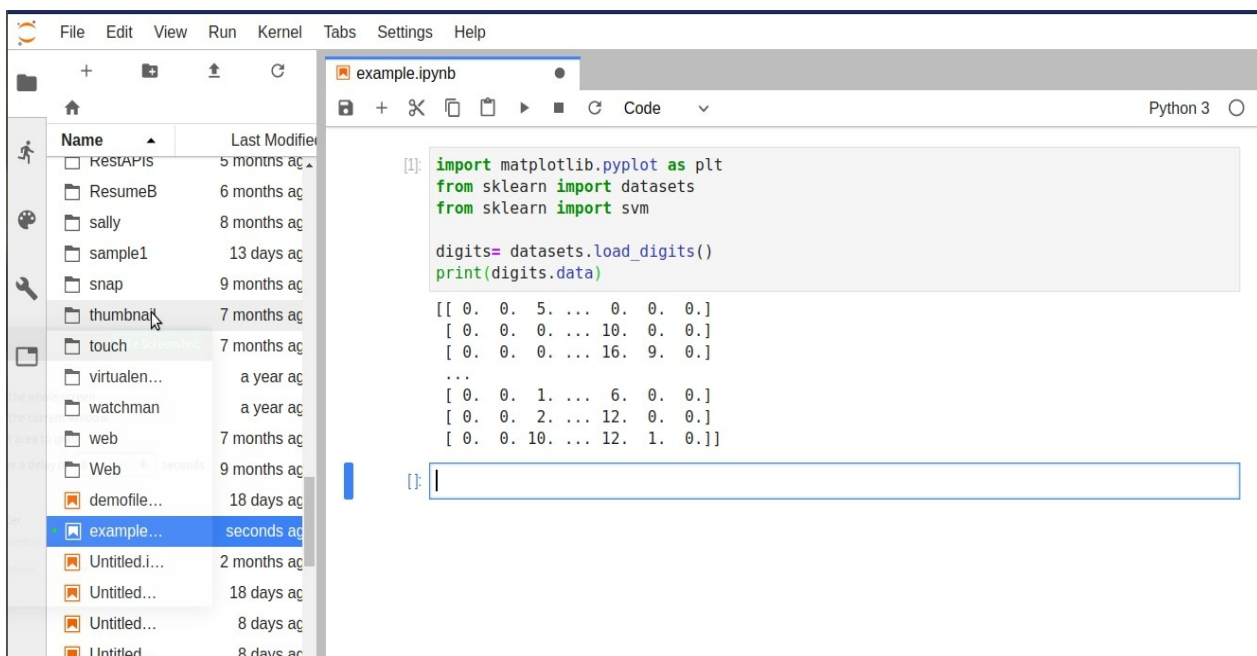
Next, in a similar manner, you have to import Sklearn. Scikit learn is built upon the SciPy (Scientific Python) that must be installed before you can use Scikit-learn. You can refer to this [website](#) to download the same. Also, install Scipy and wheel package if it's not present, you can type in the below command:

```
pip install scipy
```

And after downloading and importing the required libraries, let's dig deeper and understand how exactly Scikit learn is used.

Scikit learn comes with sample datasets, such as iris and digits. You can import the datasets and play around with them. After that, you have to import SVM which stands for Support Vector Machine. SVM is a form of machine learning which is used to analyze data.

Let us take an example where we will take *digits* dataset and it will categorize the numbers for us, for example- 0 1 2 3 4 5 6 7 8 9. Refer to the code below:



The screenshot shows a Jupyter Notebook interface with a file explorer on the left and a code editor on the right. The file explorer lists various files and folders, including 'example.ipynb' which is selected. The code editor shows the following code:

```
[1]: import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn import svm

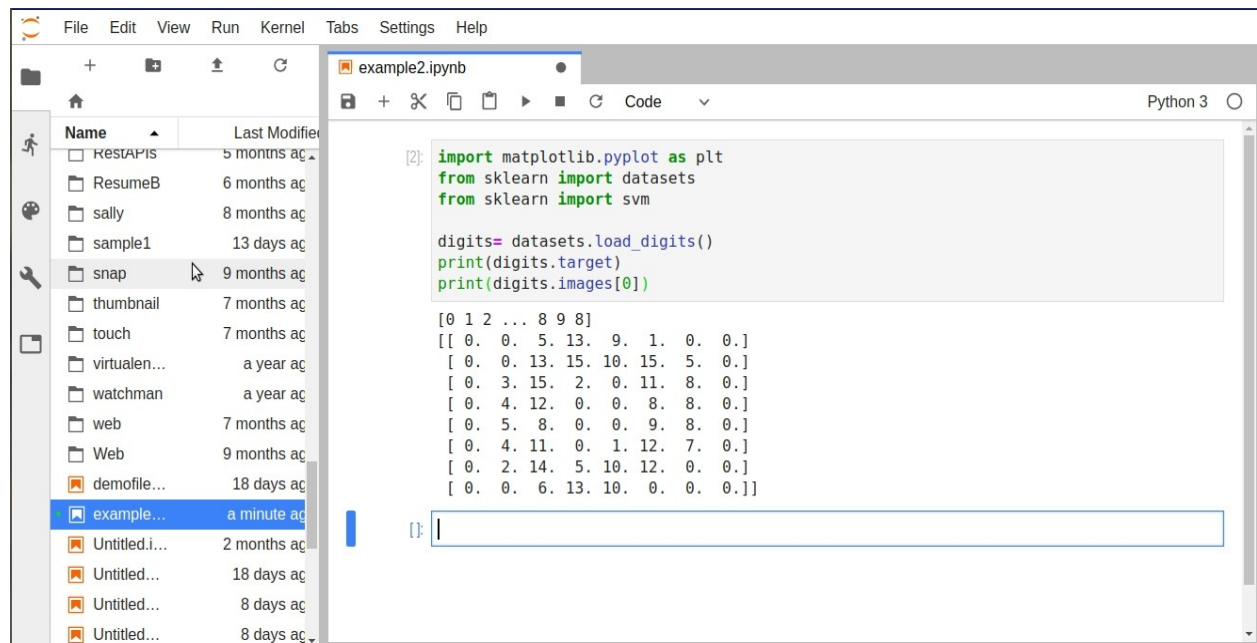
digits= datasets.load_digits()
print(digits.data)
```

The output of the code is a 2D array of digit data, represented as a list of lists. Each inner list represents a single digit sample, where each element is a value from 0 to 16, representing the intensity of a pixel in a 10x10 grid. The output shows the first few samples of the digits dataset.

```
[[ 0.  0.  5. ...  0.  0.  0.]
 [ 0.  0.  0. ... 10.  0.  0.]
 [ 0.  0.  0. ... 16.  9.  0.]
 ...
 [ 0.  0.  1. ...  6.  0.  0.]
 [ 0.  0.  2. ... 12.  0.  0.]
 [ 0.  0. 10. ... 12.  1.  0.]]
```

Here we have just imported the libraries, SVM, datasets and printed the data. It's a long array of digits data where the data is stored. It gives the access to the features that can be used to classify the *digits* samples.

Next, you can also try some other operations such as target, images etc. Consider the example below:



```
File Edit View Run Kernel Tabs Settings Help
example2.ipynb Python 3
[2]: import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn import svm

digits= datasets.load_digits()
print(digits.target)
print(digits.images[0])

[0 1 2 ... 8 9 8]
[[ 0.  0.  5. 13.  9.  1.  0.  0.]
 [ 0.  0. 13. 15. 10. 15.  5.  0.]
 [ 0.  3. 15.  2.  0. 11.  8.  0.]
 [ 0.  4. 12.  0.  0.  8.  8.  0.]
 [ 0.  5.  8.  0.  0.  9.  8.  0.]
 [ 0.  4. 11.  0.  1. 12.  7.  0.]
 [ 0.  2. 14.  5. 10. 12.  0.  0.]
 [ 0.  0.  6. 13. 10.  0.  0.  0.]]

[]
```

As you can see above, the target digits and the image of the digits are printed. `digits.target` gives the ground truth for the digit dataset, that is the number corresponding to each digit image. Next, data is always a 2D array which has a shape `(n_samples, n_features)`, although the original data may have had a different shape. But in the case of the digits, each original sample is an image of shape `(8,8)` and can be accessed using `digits.image`.

Prepare to learn more on sk-learn on future tutorial.