Unit 1- Introduction to Machine Learning

by

Mr. S. S. Gangonda
Assistant Professor
E&TC Department
SKN Sinhgad COE, Pandharpur

Course Outcomes

Course Name	**:Machine Learning**	**Course Code:**	**:ET411**
Class	**:Final Year B. Tech**	**Semester**	**:I**
Academic Year	**:2021-22**	**Subject Teacher**	**:Gangonda S. S.**

CO No.	**Course Outcome Statements**	**Cognitive Level**
	At the end of course, the students will be able to-	
ET411.1	**Describe fundamental aspects of Machine Learning.**	**L1: Remember**
ET411.2	Illustrate different Machine Learning models.	L3: Apply
ET411.3	Discuss classification and regression algorithms	L2: Understand
ET411.4	Explain neural network for classification	L2: Understand
ET411.5	Distinguish between various characteristics of ML	L2: Understand
ET411.6	Interpret Machine learning techniques that enable to solve real world problems.	L2: Understand

How to get datasets for Machine Learning

- The key to success in the field of machine learning or to become a great data scientist is to practice with different types of datasets.
- But discovering a suitable dataset for each kind of machine learning project is a difficult task. So, in this topic, we will provide the detail of the sources from where you can easily get the dataset according to your project.

❑ **What is a dataset?**

- **A dataset** is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table. Below table shows an example of the dataset:

Country	Age	Salary	Purchased
India	38	48000	No
France	43	45000	Yes
Germany	30	54000	No
France	48	65000	No
Germany	40		Yes
India	35	58000	Yes

Cont..ed

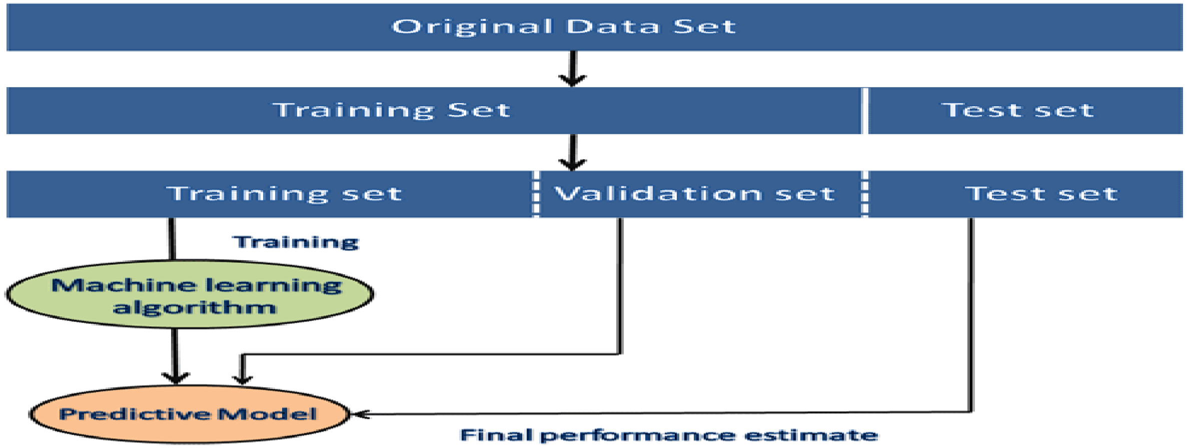
- A tabular dataset can be understood as a database table or matrix, where each column corresponds to a **particular variable,** and each row corresponds to the **fields of the dataset.** The most supported file type for a tabular dataset is **"Comma Separated File,"** or **CSV.** But to store a "tree-like data," we can use the JSON file more efficiently.

❑ **Types of data in datasets**

- **Numerical data:** Such as house price, temperature, etc.
- **Categorical data:** Such as Yes/No, True/False, Blue/green, etc.
- **Ordinal data:** These data are similar to categorical data but can be measured on the basis of comparison.
- *Note: A real-world dataset is of huge size, which is difficult to manage and process at the initial level. Therefore, to practice machine learning algorithms, we can use any dummy dataset.*

Need of Dataset

- To work with machine learning projects, we need a huge amount of data, because, without the data, one cannot train ML/AI models. Collecting and preparing the dataset is one of the most crucial parts while creating an ML/AI project.
- The technology applied behind any ML projects cannot work properly if the dataset is not well prepared and pre-processed.



- During the development of the ML project, the developers completely rely on the datasets. In building ML applications, datasets are divided into two parts:

Training dataset:

Test Dataset

Popular sources for Machine Learning datasets

Below is the list of datasets which are freely available for the public to work on it:

1. Kaggle Datasets

- Kaggle is one of the best sources for providing datasets for Data Scientists and Machine Learners. It allows users to find, download, and publish datasets in an easy way. It also provides the opportunity to work with other machine learning engineers and solve difficult Data Science related tasks.
- Kaggle provides a high-quality dataset in different formats that we can easily find and download.
- The link for the Kaggle dataset is https://www.kaggle.com/datasets

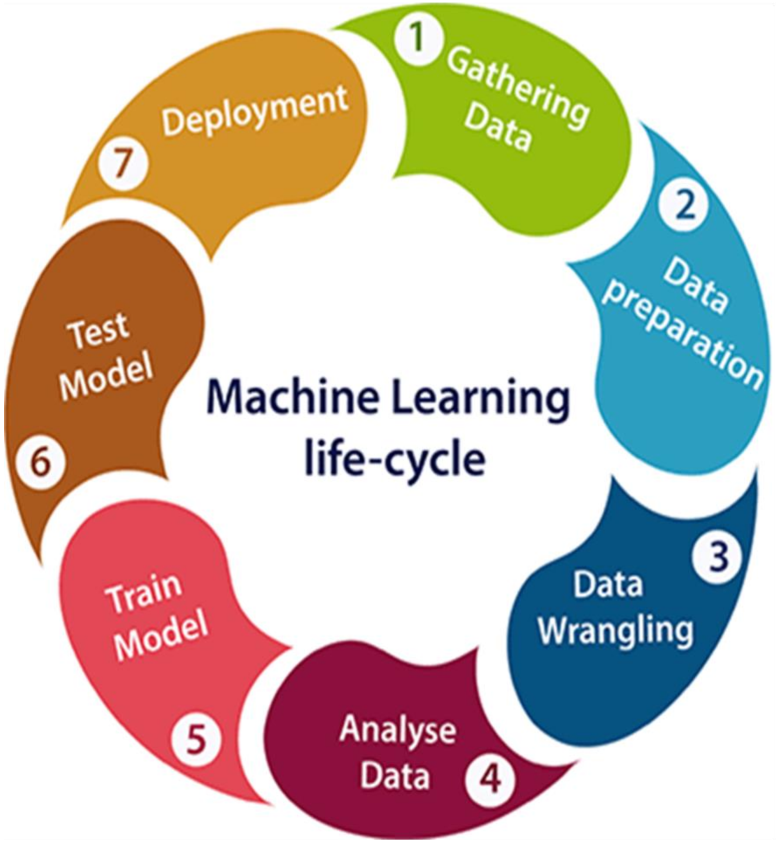
2. UCI Machine Learning Repository

- It classifies the datasets as per the problems and tasks of machine learning such as **Regression, Classification, Clustering, etc.** It also contains some of the popular datasets such as the **Iris dataset, Car Evaluation dataset, Poker Hand dataset, etc.**
- The link for the UCI machine learning repository is https://archive.ics.uci.edu/ml/index.php

3. Datasets via AWS

We can search, download, access, and share the datasets that are publicly available via AWS resources.

Machine Learning Life Cycle



1. **Gathering Data:** The goal of this step is to identify and obtain all data-related problems.
2. **Data preparation:** Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
3. **Data Wrangling:** Data wrangling is the process of cleaning and converting raw data into a useable format.
4. **Data Analysis:** The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome.
5. **Train Model:** Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.
6. **Test Model:** Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.
7. **Deployment:** If the prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system.

Thank You!