**HANDLING DATASETS**

**File Types Used**

* **CSV** file contains data in **tabular** form.
* **JSON** files keep data in **tree** format.

**Types of Data in Datasets**

* **Numerical data:** **Values** of various entities.
* **Categorical data:** Can have **two possible values** like Booleans.
* **Ordinal data:** Similar to categorical data but **degree** of value may vary.

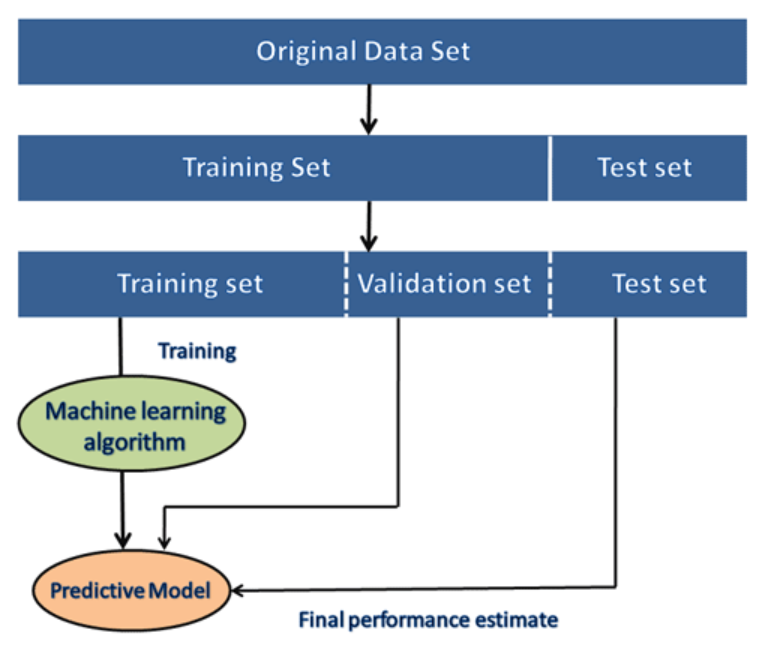
**Types of Datasets**

* **Image datasets:** Collection of **images**.
* **Text datasets:** Used in **NLP** techniques.
* **Time series datasets:** Generally used in **pattern examination**.
* **Tabular datasets**

**Data Pre-processing**

* Basically converting **raw data** into **useful information**.
* It also includes **cleansing data** & **removing irregularities** in it.

**Training Dataset & Test Dataset**



**Pre-processing Steps**

* **Step 1:** Getting the dataset
* **Step 2:** Importing libraries
* **Step 3:** Importing datasets
* **Step 4:** Finding missing data
* **Step 5:** Encoding categorical data
* **Step 6:** Splitting data into training set & test set
* **Step 7:** Feature scaling

**Getting The Dataset**

* It is preferred to store data in **spreadsheet file** and import it as ***.csv*** file.
* Because spreadsheet can hold **large** amount of data.

**Handling Missing Data**

* **Method 1:** Removing the **complete row** containing the **null/missing** value.
* **Method 2:** We **fill** that cell with **mean** of the entire column instead.

**Encoding Categorical Data**

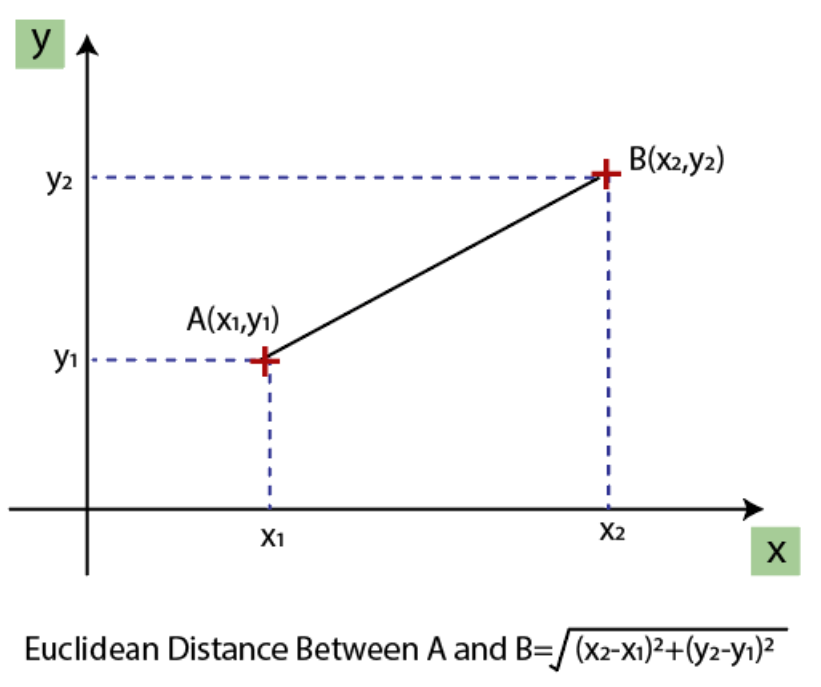
* Our model **doesn’t** know if any attribute represents **categorical data**.
* So, we **encode** our categorical data as **0s** and **1s**.
* **Dummy variable:** Variable that can be either **0** or **1**.

**Splitting Dataset Into Training Set & Test Set**

* We **separate** the **training set** & **test set** so that the training set **doesn’t** decrease performance of an already accurate model.
* Model gives unwanted output due to being **confused** between two different datasets.

**Feature Scaling**

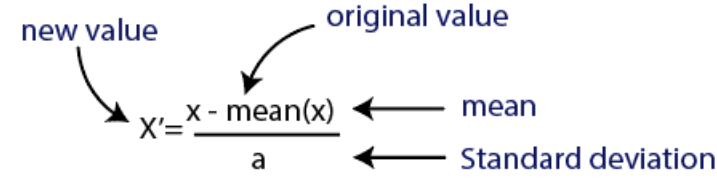
* Removing records with **too large** or **small** values.
* This affects the **consistency** of our data.
* These inconsistent data are known as **outlier**.
* So, we **squeeze** all the data into **confined space**, rescaling it.
* Feature scaling is based on **Euclidean distance**.



**Ways to Perform Feature Scaling**

Standardization:-

* **Less sensitive** to outliers.
* Used for **distance based** algorithms.



Normalization:-

* **More sensitive** to outliers.
* **Non-distance based** algorithms like neural networks.

