SHIVA'S INTERNSHIP

MACHINE LEARNING

SEPTEMBER 10, 2022

SVCET
CHITTOOR

1. Introduction to Machine Learning

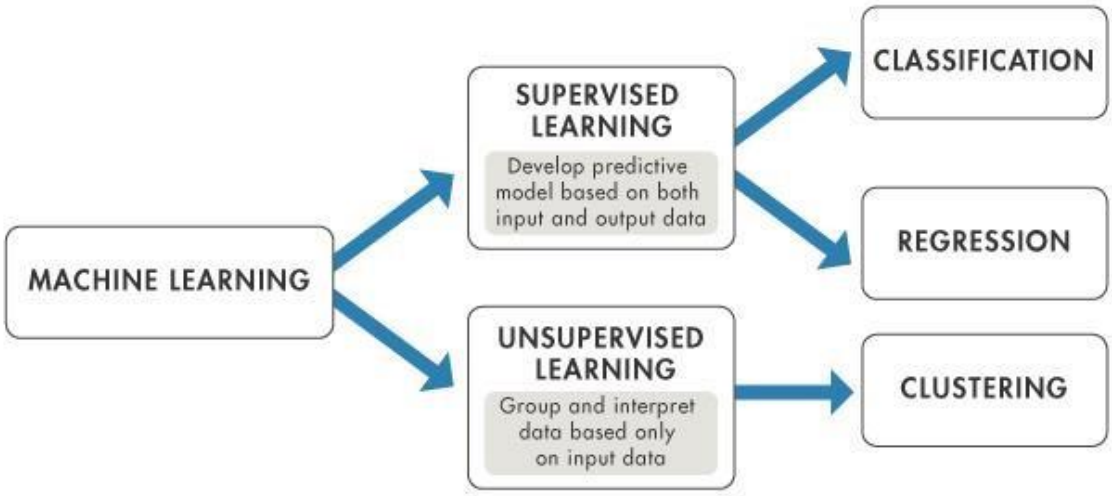
1.1. What is Machine Learning

- Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959.
- Over the past two decades Machine Learning has become one of the mainstays of information technology.
- With the ever-increasing amounts of data becoming available there is good reason to believe that smart data analysis will become even more pervasive as a necessary ingredient for technological progress.
- Machine learning is a subset of artificial intelligence (AI). It is focused on teaching computers to learn from data and to improve with experience – instead of being explicitly programmed to do so. In machine learning, algorithms are trained to find patterns and correlations in large data sets and to make the best decisions and predictions based on that analysis.

1.2. Types of Machine Learning

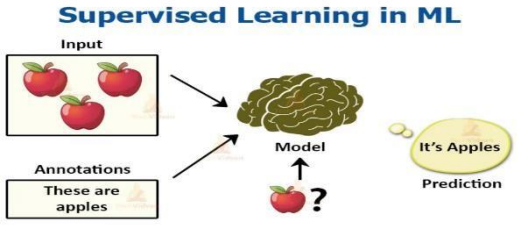
There are two types of Machine Learning:

- 1.Supervised Learning
- 2.Unsupervised Learning



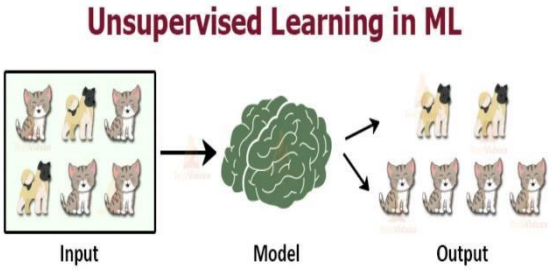
Supervised Learning :

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x)with the output variable(y)**.



Unsupervised Machine Learning :

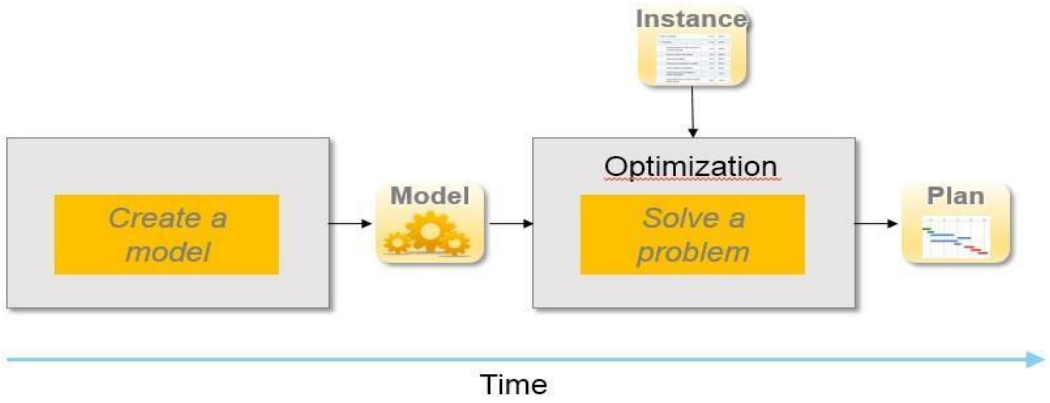
Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.



1.3 How Machine Learning works

- Machine Learning is, undoubtedly, one of the most exciting subsets of Artificial Intelligence. It completes the task of learning from data with specific inputs to the machine. It's important to understand what makes Machine Learning work and, thus, how it can be used in the future.
- The Machine Learning process starts with inputting training data into the selected algorithm. Training data being known or unknown data to develop the final Machine Learning algorithm. The type of training data input does impact the algorithm, and that concept will be covered further momentarily.
- New input data is fed into the machine learning algorithm to test whether the algorithm works correctly. The prediction and results are then checked against each other.
- If the prediction and results don't match, the algorithm is re-trained multiple times until the data scientist gets the desired outcome. This enables the machine learning algorithm to continually learn on its own and produce the optimal answer, gradually increasing in accuracy over time.

Relation to Optimization

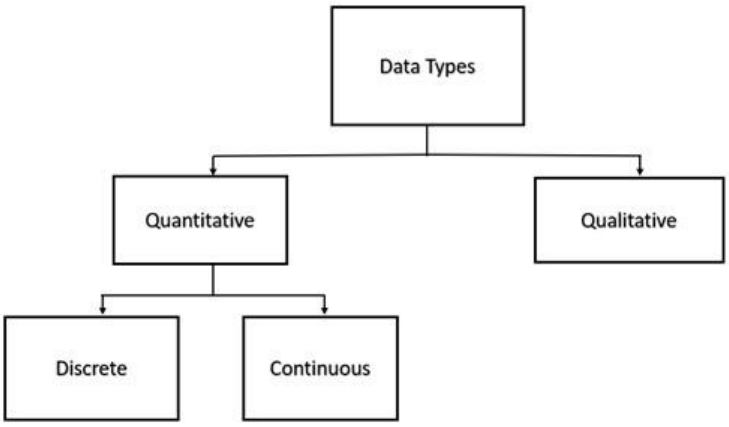


2.DATA

Different Types of Data Types

The Data Type Is Broadly Classified Into

1. Quantitative data
2. Qualitative data



1.Quantitative Data Type: –

This Type of Data Type Consists Of Numerical Values. Anything Which Is Measured By Numbers.

E.G., Profit, Quantity Sold, Height, Weight, Temperature, Etc.

This can be divided into:

A.)Discrete Data Type: –
The Numeric Data Which Have Discrete Values Or Whole Numbers. This Type Of Variable Value If Expressed In Decimal Format Will Have No Proper Meaning. Their Values Can Be Counted. E.G.: – No. Of Cars You Have, No. Of Marbles In Containers, Students In a classroom ,etc..

B.)Continuous Data Type: –

The Numerical Measures Which Can Take The Value Within A Certain Range. This Type Of Variable Value If Expressed In Decimal Format Has True Meaning. Their Values Can Not Be Counted But Measured. The
Value Can Be Infinite
E.G.: – Height, Weight, Time, Area, Distance, Measurement Of Rainfall, Etc.

2. Qualitative Data Type: –
These Are The Data Types That Cannot Be Expressed In Numbers. This Describes Categories Or Groups And Is Hence Known As The Categorical Data Type.

This Can Be Divided Into:-

A. Structured Data:
This Type Of Data Is Either Number Or Words. This Can Take Numerical Values But Mathematical Operations Cannot Be Performed On It. This Type Of Data Is Expressed In Tabular Format. E.G.) Sunny=1, Cloudy=2, Windy=3 Or Binary Form Data Like 0 Or1, Good Or Bad, Etc.

B. Unstructured Data:

This Type Of Data Does Not Have The Proper Format And Therefore Known As Unstructured Data .This Comprises Textual Data, Sounds, Images, Videos, Etc.

Besides This, There Are Also Other Types Refer As Data Types Preliminaries Or Data Measures:-

1. Nominal data
2. Ordinal data

These Can Also Be Refer Different Scales Of Measurements.

Nominal Data Type:
This Is In Use To Express Names Or Labels Which Are Not Order Or Measurable.

E.G., Male Or Female (Gender), Race, Country, Etc.

Ordinal Data Type:
This Is Also A Categorical Data Type Like Nominal Data But Has Some Natural Ordering Associated With
It.

E.G., Likert Rating Scale, Shirt Sizes, Ranks, Grades, Etc.

DATA ANALYTICS:

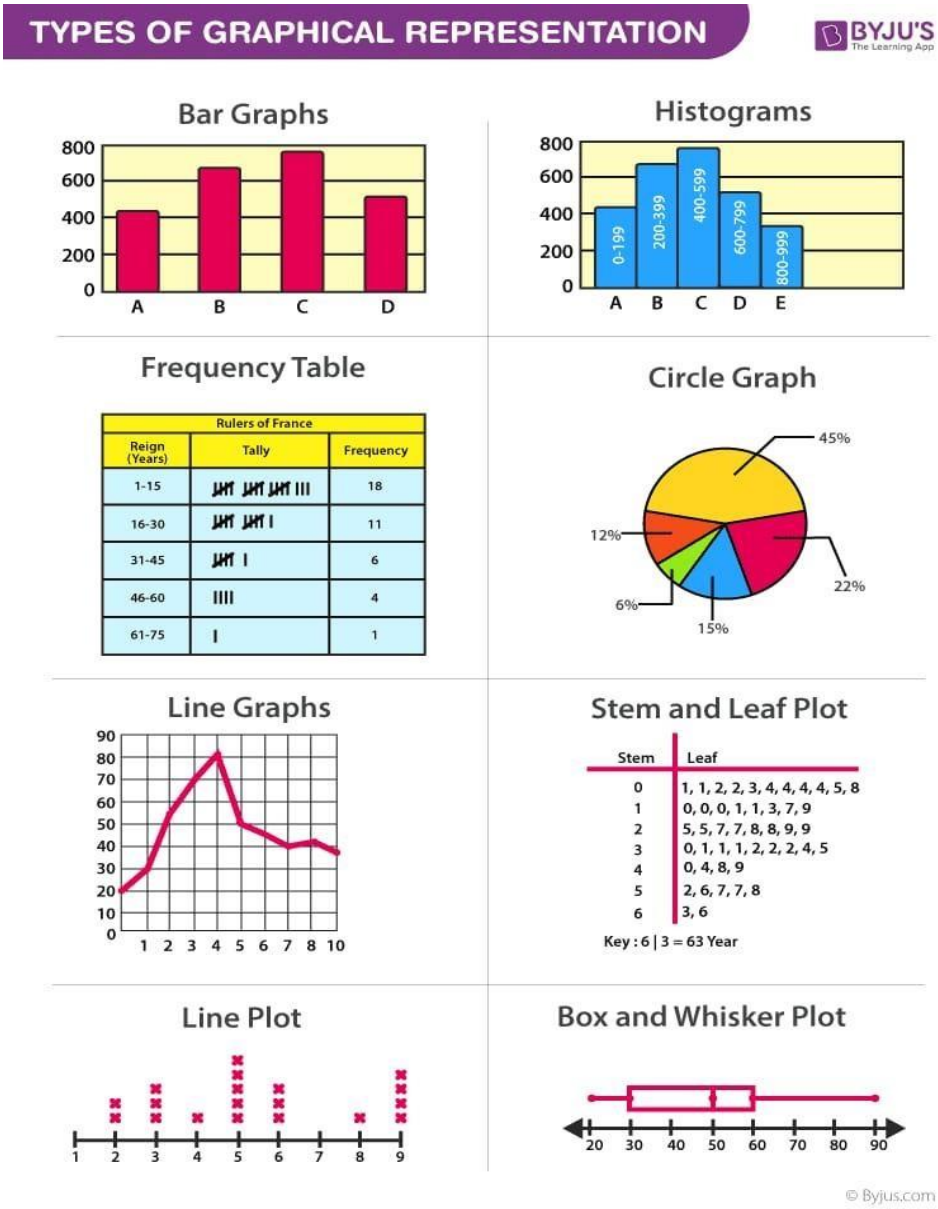
- Is the process of studying in the available data and drawing valuable insights or information from it. With the help of software.
- Is being used every day & every where to enable the business to take smart and accurate decisions.

GRAPHICAL REPRESENTATION OF DATA:

- It is one of the simple techniques for drawing insights from the data.
- It helps us to study relationship between the variables.
- Helps us to identify the trend & patterns across the variables.

MAJOR TYPES :

1. Line graph
2. Bar graph
3. Histogram
4. Pie chart
5. Scatter plot



3.INTRODUCTION TO PYTHON

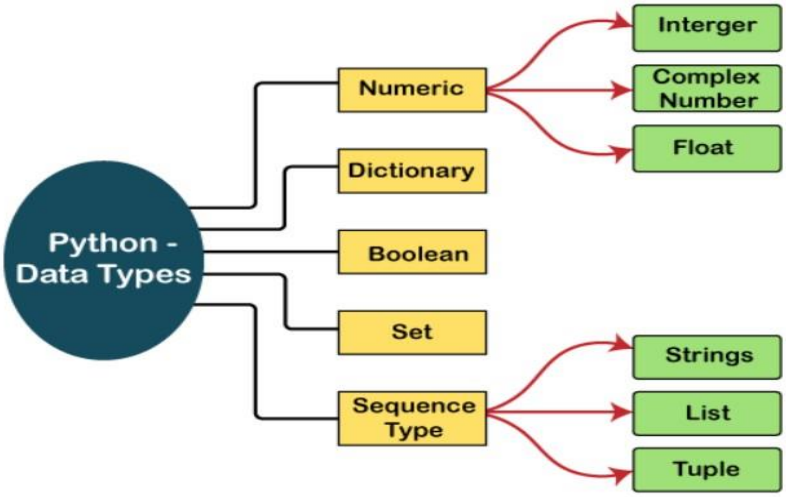
Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.

It is used for:

- web development (server-side),
- software development, mathematics,

- system scripting.

3.1 DATA TYPES IN PYTHON:



Int ->Integer value can be any length such as integers 10, 2, 29, -20, -150 etc. Python has no restriction on the length of an integer. Its value belongs to int

Float ->Float is used to store floating-point numbers like 1.9, 9.902, 15.2, etc. It is accurate up to 15 decimal points.

String->The string can be defined as the sequence of characters represented in the quotation marks. In Python, we can use single, double, or triple quotes to define a string.

```
a = 5
print("The type of a", type(a))

b = 40.5
print("The type of b", type(b))
```

OUTPUT:

```
The type of a <class 'int'>
The type of b <class 'float'>
```

3.2 CONDITIONAL STATEMENT

IF:

These conditions can be used in several ways, most commonly in "if statements" and loops.

An "if statement" is written by using the if keyword.

```
If statement:
```

```
a = 33
b = 200
if b > a:
  print("b is greater than a")
```

ELIF:

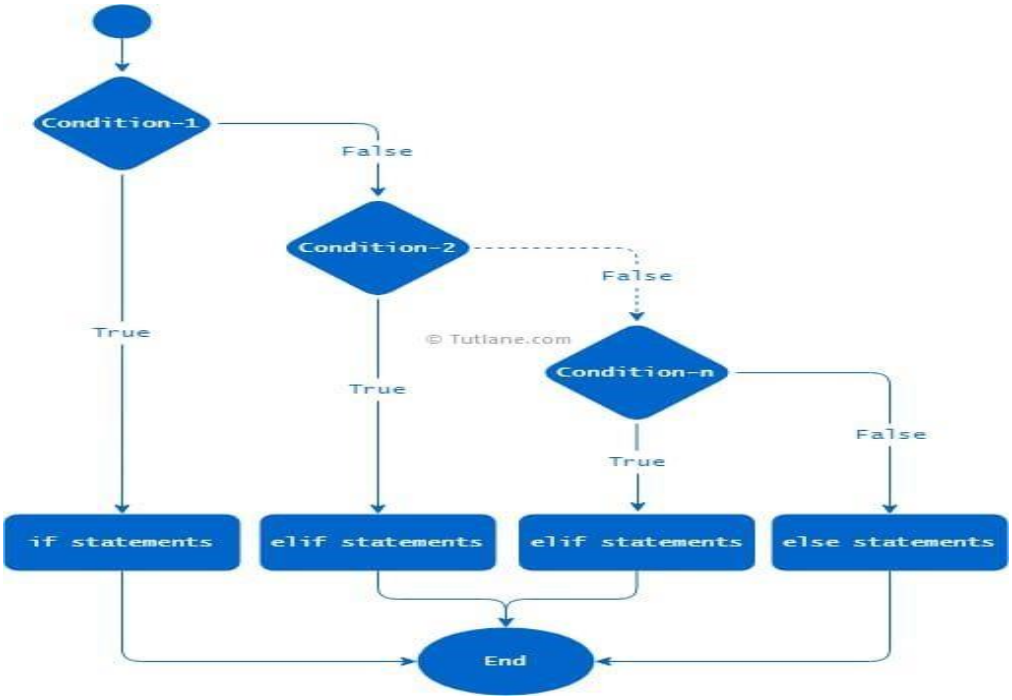
The Elif keyword is pythons way of saying "if the previous conditions were not true, then try this condition".

ELSE:

The else keyword catches anything which isn't caught by the preceding conditions.

NESTED IF:

You can have `if` statements inside `if` statements, this is called *nested* `if` statements.



3.3 ITERATIVE STATEMENTS

Iteration statements or loop statements allow us to execute a block of statements as long as the condition is true.

1.While Loop :

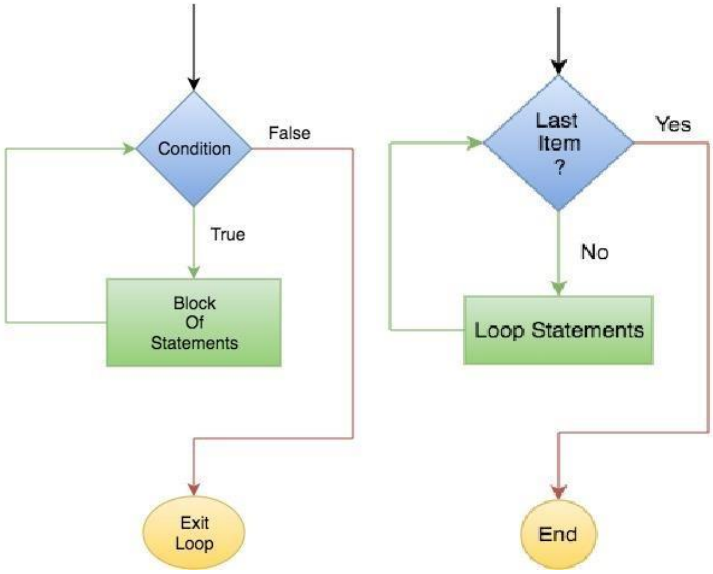
While Loop In Python is used to execute a block of statement as long as a given condition is true. And when the condition is false, the control will come out of the loop

2. For Loop :

For loop in Python is used to iterate over items of any sequence, such as a list or a string.

3.nested for loop

Loop with in a loop an inner loop within a body of an outer one



3.4 FUNCTIONS

A function is a block of code which only runs when it is called.

Creating a Function

In Python a function is defined using the **def** keyword.

```
def my_function():
  print("Hello from a function")
```

Calling a Function

To call a function, use the function name followed by parenthesis.

```
def my_function():
  print("Hello from a function")

my_function()
```

3.5 BASIC LIBRARIES IN PYTHON

LIBRARY:

has created several open-source libraries, each with its root source. A library is an initially merged Python collection of code scripts that can be used iteratively to save time. It's similar to a physical library in that it holds reusable resources, as the name implies.

Matplotlib:

The plotting of numerical data is the responsibility of this library. It's for this reason that it's used in analysis of data. It's an open-source library that plots high-definition figures such as pie charts, scatterplots, boxplots, and graphs, among other things.

NumPy:

NumPy is one of the most widely used open-source Python packages, focusing on mathematical and scientific computation. It has built-in mathematical functions for convenient computation and facilitates large matrices and multidimensional data. It can be used for various things, including linear algebra, as an N-dimensional container for all types of data. The NumPy Array Python object defines an N-dimensional array with rows and columns. A long with this, it can be used as a random number generator.

Pandas:

Pandas is an open source library licensed under the Berkeley Software Distribution (BSD). In the domain of data science, this well-known library is widely used. They're mostly used for analysis, manipulation, and cleaning of data, among other things. Pandas allows us to perform simple data modelling and analysis without having to swap to another language like R.

Scikit- learn:

Scikit-learn is also an open-source machine learning library based on Python. Both supervised and unsupervised learning processes can be used in this library. Popular algorithms and the SciPy, NumPy, and Matplotlib packages are all already pre-included in this library. The most well-known Scikit-most-learn application is for Spotify music recommendations.

4 . DATA EXPLORATION AND PRE-PROCESSING

4.1 Data Exploration – Target Variable :

The target variable is **the variable whose values are modeled and predicted by other variables**. A predictor variable is a variable whose values will be used to predict the value of the target variable.

Why are Target Variables Important?

.In the absence of a labeled target, supervised machine learning algorithms would not be able to map available data to outcomes.

- Understand data and make sure it is ready to be used in a model.
- A model would be as 9000 as the data it is built on.
- Take a structure and step by step approach in understand and preparing the data

$$x' = \frac{x - \bar{x}}{\sigma}$$

4.2 Data Exploration-independent variable

Overview

A complete tutorial on data exploration (EDA)
We cover several data exploration aspects, including missing value imputation, outlier removal and the art of feature engineering

Steps of Data Exploration and Preparation:

1.Variable Identification

First, identify **Predictor** (Input) and **Target** (output) variables. Next, identify the data type and category of the variables.

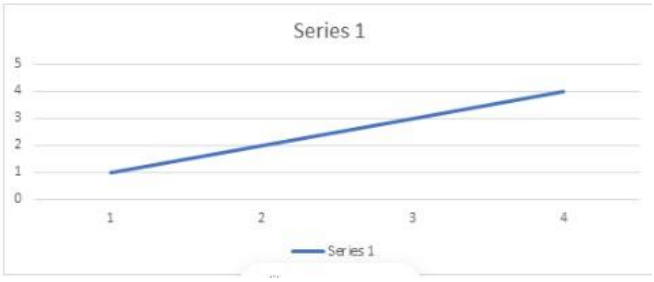
2. **Continuous Variables:-** In case of continuous variables, we need to understand the central tendency and spread of the variable.

3.Categorical Variables:- For categorical variables, we'll use frequency table to understand distribution of each category.

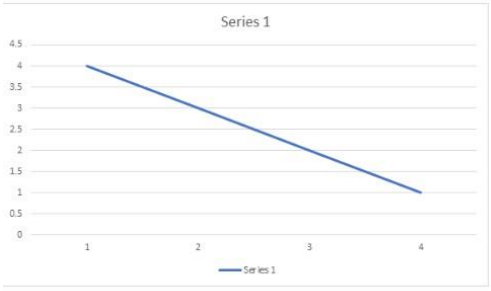
4.Continuous & Continuous: While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables.

4.3 correlation : Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

Positive Correlation: Two features (variables) can be positively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) also increases.

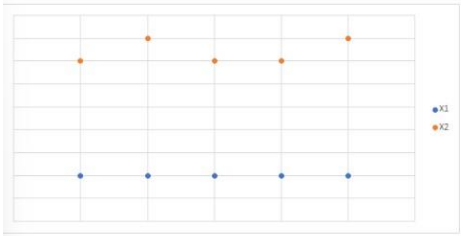


Negative Correlation: Two features (variables) can be negatively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) decreases.



Negative Correlation

No Correlation: Two features (variables) are not correlated with each other. It means that when the value of one variable increase or decrease then the value of the other variable(s) doesn't increase or decreases.



No Correlation

data exploration categorical variables:

Exploring categorical variables is generally simpler than working with numeric variables because we have fewer options, or at least life is simpler if we only require basic summaries. We'll work with the year and type variables in storms to illustrate the key ideas.

feature scaling: Feature scaling is a method used to normalize the range of independent variables or features of data.

min-max normalization

Also known as min-max scaling or min-max normalization, rescaling is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization (Z-score Normalization)

In machine learning, we can handle various types of data, e.g. audio signals and pixel values for image data, and this data can include multiple dimensions. Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning algorithms

$$x' = \frac{x - \bar{x}}{\sigma}$$

.

5. Linear Regression

5.1 Regression model

A regression model is used to investigate the relation between two or more variables and estimate one variable based on the others.
In regression analysis variables can be independent which are used as the predictor or casual input and dependent, which are used as a response variables in experimental studies independent variable X is a variable that can be controlled and variable y is the variable that reflects the changes in independent variable X.

5.2 Model evaluation metrics

Predictive models have become a trusted advisor to many businesses and for a good reason. These models can "foresee the future", and there are many different methods available, meaning any industry can find one that fits their particular challenges.
When we talk about predictive models, we are talking either about a regression model (continuous output) or a classification model (nominal or binary output). In classification problems, we use two types of algorithms (dependent on the kind of output it creates):

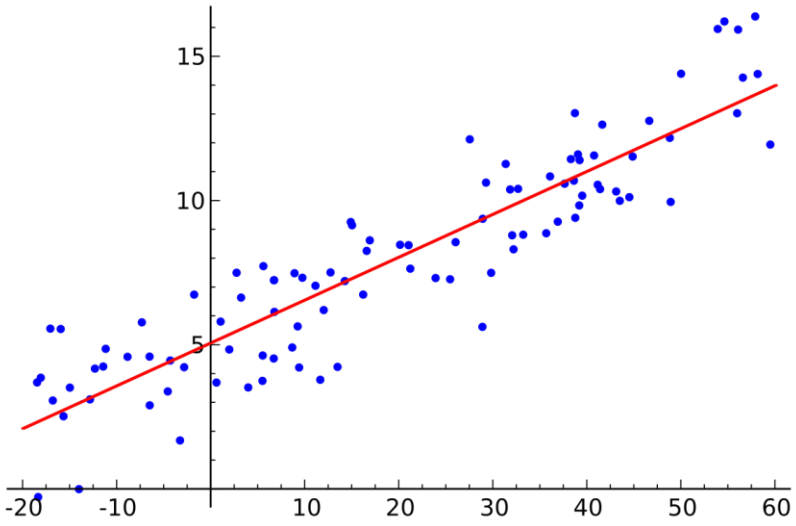
		Predicted Values: Negative	Predicted Values: Positive
Actual Values	Negative	TN True Negative	FP False positive
	Positive	FN False Negative	TP True Positive

Implementing Linear regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a **statistical method that is used for predictive analysis**. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression, make use of the following key principles:

1. Define the equation used for making predictions
2. Define the parameters to learn to make predictions
3. Define the cost function (or loss function) required to train the model
4. Train the model using gradient descent in order to minimize the cost function 5. Make predictions using the trained parameters



Linear regression is a linear approach for modelling the relationship between a scalar dependent variable y and an independent variable x.

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$$

where x, y, w are vectors of real numbers and w is a vector of weight parameters.

The equation is also written as:

y = wx + b

where b is the bias or the value of output for zero input

5.3 Gradient Descent

Gradient Descent is known as one of the most commonly used optimization algorithms to train machine learning models by means of minimizing errors between actual and expected results. Further, gradient descent is also used to train Neural Networks.

In mathematical terminology, Optimization algorithm refers to the task of minimizing/maximizing an objective function f(x) parameterized by x. Similarly, in machine learning, optimization is the task of minimizing the cost function parameterized by the model's parameters.

The main objective of gradient descent is to minimize the convex function using iteration of parameter updates. Once these machine learning models are optimized, these models can be used as powerful tools for Artificial Intelligence and various computer science applications.

5.4 Training model

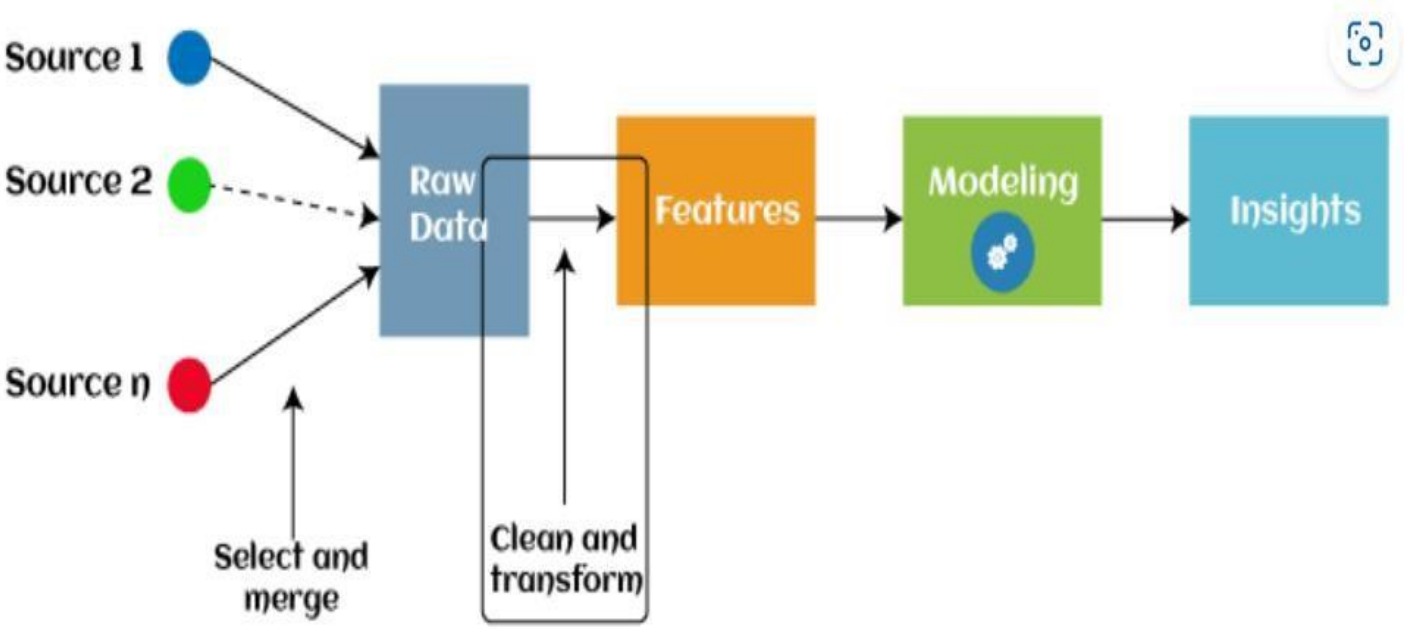
Training ML Models. The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process. The training data must contain the correct answer, which is known as a target or target attribute.

- Let’s start with a crucial but sometimes overlooked step: Spending your data. Think of your data as a limited resource.

- You can spend some of it to train your model (feed it to the algorithm). You can spend some of it to evaluate (test) your model. But you can't reuse the same data for both!

5.6 Feature Engineering

Feature engineering is the pre-processing step of machine learning, which extracts features from raw data. It helps to represent an underlying problem to predictive models in a better way, which as a result, improve the accuracy of the model for unseen data. The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.



6. Introduction to Dimensionality Reduction

6.1 Common Dimensionality Reduction Technique:

In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play.

Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction**.**

6.2 Advanced Dimensionality Reduction Technique:

Here are two components of dimensionality reduction : Feature selection: In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
Filter,
Wrapper,
Embedded.

7. Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1)

.

EVALUATION METRIX:

Confusion Matrix

A confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix. Here are a few definitions, you need to remember for a confusion matrix :

Accuracy: the proportion of the total number of predictions that were correct.
Positive Predictive Value or Precision: the proportion of positive cases that were correctly identified.
Negative Predictive Value: the proportion of negative cases that were correctly identified.
Sensitivity or Recall : the proportion of actual positive cases which are correctly identified. **Specificity**: the proportion of actual negative cases which are correctly identified.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

F1 Score:

precision and recall for classification problems and also highlighted the importance of choosing precision/recall basis our use case. What if for a use case, we are trying to get the best precision and recall at the same time? F1-Score is the harmonic mean of precision and recall values for a classification problem. The formula for F1-Score is as follows:

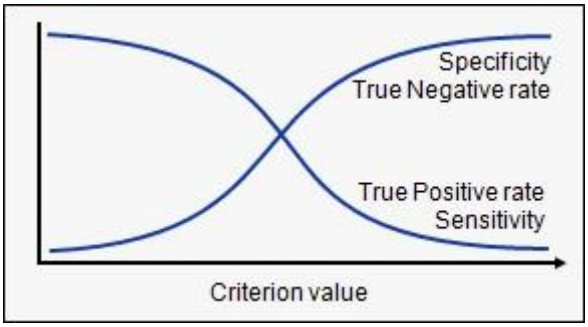
$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Area Under the ROC curve (AUC – ROC):

This is again one of the popular metrics used in the industry. The biggest advantage of using ROC curve is that it is independent of the change in proportion of responders. This statement will get clearer in the following sections.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Hence, for each sensitivity, we get a different specificity .The two vary as follows:



The ROC curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known as false positive rate and sensitivity is also known as True Positive rate. Following is the ROC curve for the case in hand.

8.DECISION TREE

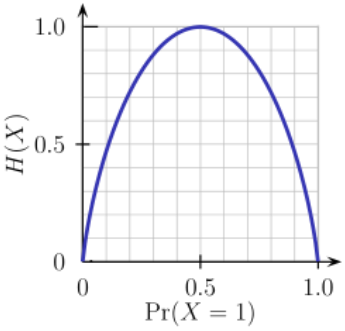
8.1 HOW DECISION TREE WORKS

The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria are different for classification and regression trees.

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

Entropy

Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information. Flipping a coin is an example of an action that provides information that is random.



8.2 implementing decision tree

Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**

In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

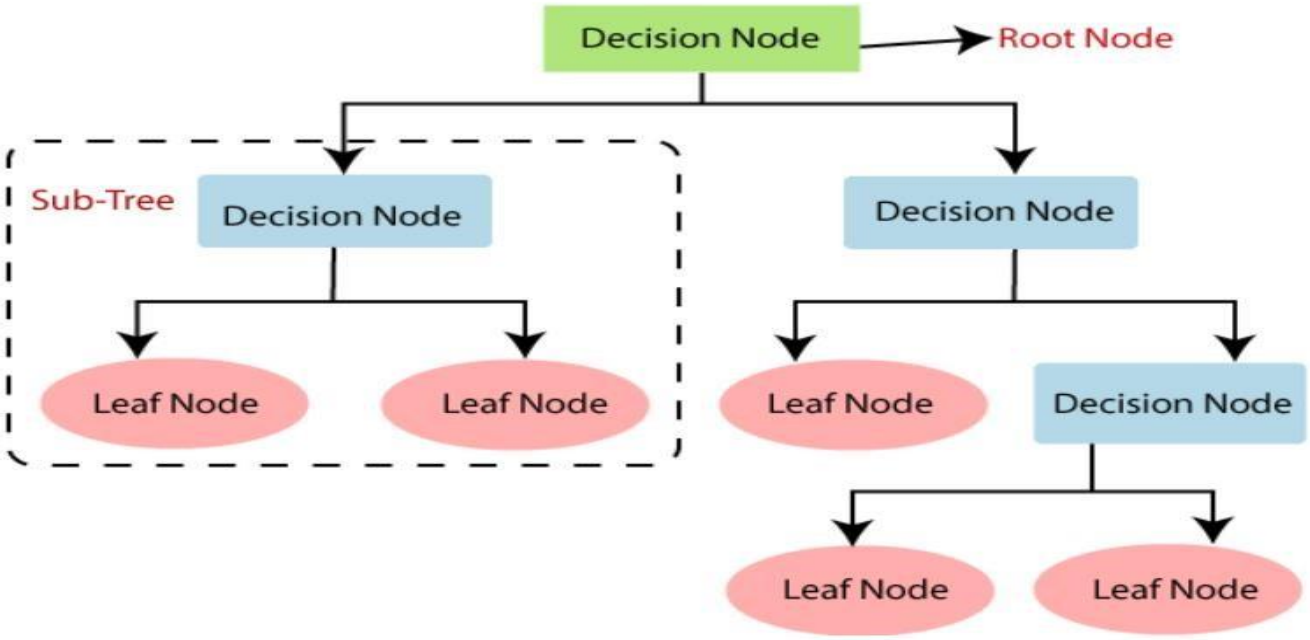
It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:



9.ENSEMBLE MODELS

9.1 BASIC ENSEMBLE TECHNIQUES

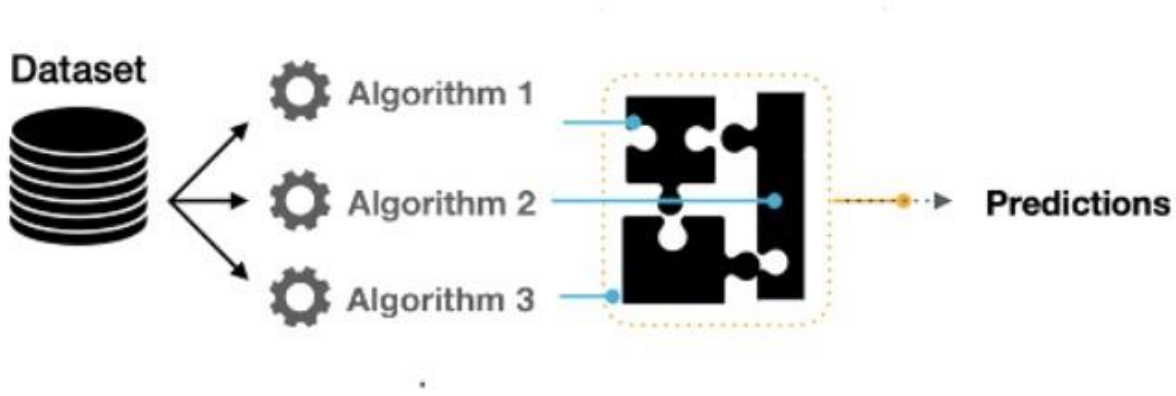
Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data. The motivation for using ensemble models is to reduce the generalization error of the prediction. As long as the base models are diverse and *independent*, the prediction error of the model decreases when the ensemble approach is used. The approach seeks the wisdom of crowds in making a prediction. Even though the ensemble model has multiple base models within the model, it acts and performs as a single model. Most of the practical data mining solutions utilize ensemble modeling techniques. Classification covers the approaches of different ensemble modeling techniques and their implementation in detail

BAGGING

The idea of bagging is based on making the training data available to an iterative learning process. Each model learns the error produced by the previous model using a slightly different subset of the training data set. Bagging reduces variance and minimizes overfitting. One example of such a technique is the random forest algorithm.

Ensemble Algorithm

A single algorithm may not make the perfect prediction for a given data set. Machine learning algorithms have their limitations and producing a model with high accuracy is challenging. If we build and combine multiple models, we have the chance to boost the overall accuracy. We then implement the combination of models by aggregating the output from each model with two objectives

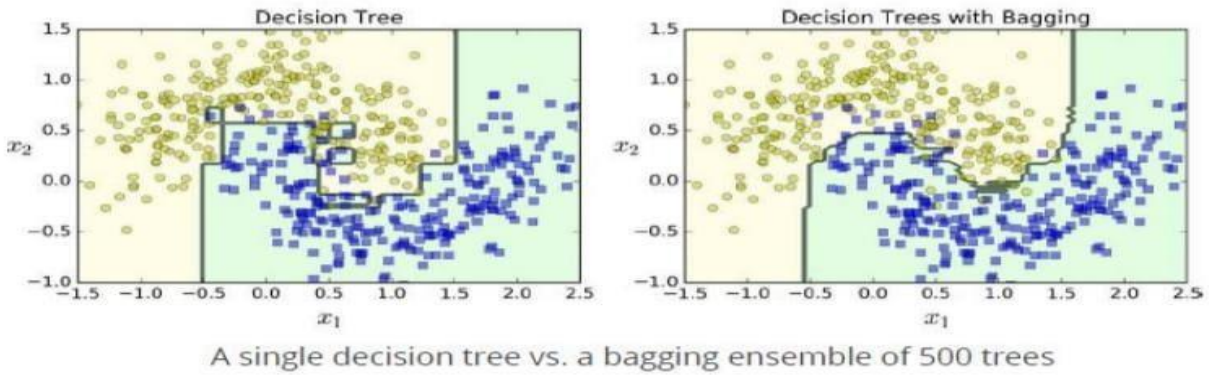


RANDOM FOREST

This technique uses a subset of training samples as well as a subset of features to build multiple split trees. Multiple decision trees are built to fit each training set. The distribution of samples/features is typically implemented in a random mode.

Ensemble Learning uses the same algorithm multiple times or a group of different

algorithms together to improve the prediction of a model.



A single decision tree vs. a bagging ensemble of 500 trees

10 . Clustering

10.1 Clustering:

- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

- Clustering or cluster analysis is a machine learning technique, which groups the un-labelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with anothergroup."

10.2 K-means:

- K-Means clustering is an **unsupervised learning algorithm**. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

- First, each data point is randomly assigned to one of the K clusters. Then, we compute the centroid (functionally the center) of each cluster, and reassign each data point to the cluster with the closest centroid. We repeat this process until the cluster assignments for each data point are no longer changing.

- K refers to the number of clusters that are required to be created for grouping data inputs. As this algorithm is a centroid-based clustering algorithm, each data cluster is attached to a centroid.

Project : Analyzing Raw Data To Predicting Sale Price

Problem statement:

Analysis on raw data set and visualization ,Which contains information about house selling price regarding to there Flat size ,Year ,Layout , location , No. of floors.

By using Python libraries like pandas, Matplotlib.

Here ,I took a data set which contains more number of data ,in this I taken Housing Data .

Pandas

It is useful in deploying data of existed files or user Entered provided data.

Pandas is an open source Python package that is most widely used for Data Science / Data analysis and **Machine Learning** tasks.

Pandas comes with two Data structure for manipulating data :

- Series
- Data Frame

1.First import pandas library to use Housing Data.

```
In [2]: import pandas as pd
        Houseing_data=pd.read_csv("Transformed_Housing_Data2.csv")
        Houseing_data
```

Out[2]:

	Sale_Price	No of Bedrooms	No of Bathrooms	Flat Area (in Sqft)	Lot Area (in Sqft)	No of Floors	No of Times Visited	Overall Grade	Area of the House from Basement (in Sqft)	Basement Area (in Sqft)	...	Waterfront_View_Yes	Zipcode_Group_Zipcode_Gro
0	221900.0	3	1.00	1180.0	5650.0	1.0	0	7	1180.0	0	...	0	
1	538000.0	3	2.25	2570.0	7242.0	2.0	0	7	2170.0	400	...	0	
2	180000.0	2	1.00	770.0	10000.0	1.0	0	6	770.0	0	...	0	
3	604000.0	4	3.00	1960.0	5000.0	1.0	0	7	1050.0	910	...	0	
4	510000.0	3	2.00	1680.0	8080.0	1.0	0	8	1680.0	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
21604	360000.0	3	2.50	1530.0	1131.0	3.0	0	8	1530.0	0	...	0	
21605	400000.0	4	2.50	2310.0	5813.0	2.0	0	8	2310.0	0	...	0	
21606	402101.0	2	0.75	1020.0	1350.0	2.0	0	7	1020.0	0	...	0	

In the above section , observe that we initialized a CSV File to Houseing _ data variable by using pandas

library. It will show the data from beginning to end.

2. To get particular variable data on top list data use head() function , in-front of variable name

3. To get particular variable data of bottom list data use head() function , in-front of variable name

```
In [3]: Houseing_data['Sale_Price'].head()
Out[3]: 0    221900.0
        1    538000.0
        2    180000.0
        3    604000.0
        4    510000.0
        Name: Sale_Price, dtype: float64

In [3]: import pandas as pd
        Houseing_data=pd.read_csv("Transformed_Housing_Data2.csv")
        Houseing_data['Sale_Price'].tail()
Out[3]: 21604    360000.0
        21605    400000.0
        21606    402101.0
        21607    400000.0
        21608    325000.0
        Name: Sale_Price, dtype: float64
```

These functions() helps in getting quick summary of what there top & bottom data , which is easy for **Analysis :**

4.By using info() function we will get list of Datatypes in our Housing Data set.

```
In [5]: Houseing_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21609 entries, 0 to 21608
Data columns (total 31 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Sale_Price                             21609 non-null  float64
 1   No of Bedrooms                         21609 non-null  int64
 2   No of Bathrooms                        21609 non-null  float64
 3   Flat Area (in Sqft)                    21609 non-null  float64
 4   Lot Area (in Sqft)                     21609 non-null  float64
 5   No of Floors                           21609 non-null  float64
 6   No of Times Visited                    21609 non-null  int64
 7   Overall Grade                          21609 non-null  int64
 8   Area of the House from Basement (in Sqft)  21609 non-null  float64
 9   Basement Area (in Sqft)                21609 non-null  int64
 10  Age of House (in Years)                21609 non-null  int64
 11  Latitude                               21609 non-null  float64
 12  Longitude                              21609 non-null  float64
 13  Living Area after Renovation (in Sqft)  21609 non-null  float64
 14  Lot Area after Renovation (in Sqft)    21609 non-null  int64
 15  Years Since Renovation                 21609 non-null  int64
 16  Condition_of_the_House_Excellent       21609 non-null  int64
 17  Condition_of_the_House_Fair            21609 non-null  int64
 18  Condition_of_the_House_Good            21609 non-null  int64
```

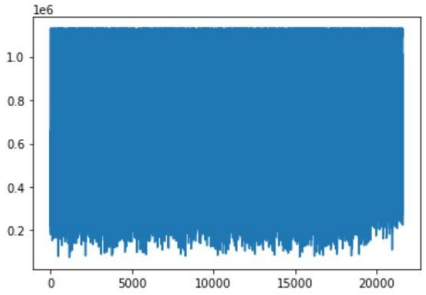
Now we are aware about that what kind of datatypes there in Dataset ,Now we jump into Matplotlib library

MATPLOTLIB LIBRARY IS USED FOR VISUAL REPRESENTATION OF DATA IN GRAPHICAL MANNER

5.Now we will plot a line graph for Sale price variable for visual understanding

```
In [7]: import matplotlib.pyplot as plt
        plt.plot(Houseing_data['Sale_Price'])

Out[7]: [<matplotlib.lines.Line2D at 0x264f9b5b3d0>]
```

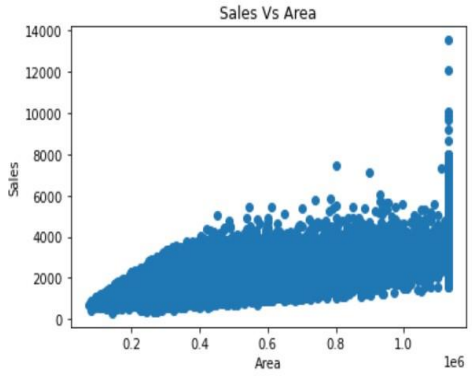


Here , you can see it plotted a line graph for Sales price.

In next step we will compare two variables difference/change in plotting graph

6.Finally we plotted a Scatter plot for Sales price and Flat Area[in sq feet]

```
In [3]: import pandas as pd
        Houseing_data=pd.read_csv("Transformed_Housing_Data2.csv")
        import matplotlib.pyplot as plt
        plt.scatter(x=Houseing_data['Sale_Price'],y=Houseing_data['Flat Area (in Sqft)'])
        plt.xlabel('Area')
        plt.ylabel('Sales')
        plt.title('Sales Vs Area')
        plt.show()
```



By,Using X,Y-Axis , named as Area and Sales Also compared both values.

CONCLUSION:

I believe the trial has shown conclusively that it is both possible and desirable to use Python as the principal teaching language. It is Free (as in both cost and source code). It is trivial to install on a Windows PC allowing students to take their interest further. For 56 many the hurdle of installing a Pascal or C compiler on a Windows machine is either too expensive or too complicated. It is a flexible tool that allows both the teaching of traditional procedural programming and modern OOP. It can be used to teach a large number of transferable skills. It is a real-world programming language that can be and is used in academia and the commercial world. It appears to be quicker to learn and, in combination with its many libraries, this offers the possibility of more rapid student development allowing the course to be made more challenging and varied and most importantly,

its clean syntax offers increased understanding and enjoyment for students. The training program having three destination was a lot more useful than staying at one place throughout the whole 6 weeks. In my opinion. I have gained lots of knowledge and experience needed to be successful in great engineering challenge as in my opinion, Engineering is after all a Challenge, and not a job.