**EXPOSYS DATA LABS**

DATA SCIENCE INTERSHIP PROJECT

BUNETI DILIPKUMAR GOUDF

MALLA REDDY INSTITUTE OF TECHNOLOGY(MLTM)

# ABSTRACT

The world's data is growing more than 40% annually. Coupled with exponentially growing computing horsepower, this provides us withunprecedented basis for 'learning' useful things from the data throughstatistical induction without material human intervention and acting onthem. Philosophers have long debated the merits and demerits ofinduction as a scientific method, the latter being that conclusions arenot guaranteed to be certain and that multiple and numerous models canbe conjured to explain the observed data. I propose that 'big data'brings a new and important perspective to these problems in that itgreatly ameliorates historical concerns about induction, especially ifour primary objective is prediction as opposed to causal modelidentification. Equally significantly, it propels us into an era ofautomated decision making, where computers will make the bulk ofdecisions because it is infeasible or more costly for humans to do so.In this paper, I describe how scale, integration and most importantly,prediction will be distinguishing hallmarks in this coming era of DataScience.' In this brief monograph, I define this newly emerging fieldfrom business and research perspectives. Here in the given dataset, R&D Spend, Administration Cost and Marketing Spend of 50 Companies are given along with the profit earned.I targeted to prepare an ML model which can predict the profit value of a company if the value of its R&D Spend, Administration Cost and Marketing Spend,using different regression algorithms with spliting data in to train and test,regression metrics.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | **TABLE OF CONTENTS** | **PAGE NO** |
| **1** | **INTRODUCTION** | **1** |
| **2** | **EXISTING METHODS** | **2** |
| **3** | **PROPOSED METHODS WITH ARCHITECTURE** | **18** |
|  |  |  |
| **4** | **METHODOLOGY** | **30** |
|  |  |  |
| **5** | **IMPLEMENTATION** | **35** |
| **6** | **CONCLUSION** | **41** |

# 1.INTRODUCTION

The competition goal is to predict the profit of startup profit on the bases of data provided which are on the bases of Research and Development Spend(R&D Spend), Administration Spend, Marketing Spend. This model can help those people who want to invest in startup company by analysing profit of the company.Here 50 startups dataset containing 4 columns  like “R&D Spend”, “Administration”, “Marketing Spend”, “Profit”.In this dataset first 3 columns provides you spending on Research , Administration and Marketing respectively. Profit indicates how much profits earned by a startup.

# 2.EXISTING METHODS

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

# Data Science Life Cycle

Data science’s lifecycle consists of five distinct stages, each with its own tasks:

**1. Capture:** Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data

**2. Maintain:** Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.

**3. Process:** Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.

**4. Analyze:** Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.

**5.Communicate:** Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.

# Understanding Artificial Intelligence (AI)

Artificial Intelligence includes the simulation process of human intelligence by machines and special computer systems. The examples of artificial intelligence include learning, reasoning and self-correction. Applications of AI include speech recognition, expert systems, and image recognition and machine vision.

## Types of AI

**Reactive AI** The most basic type of artificial intelligence is reactive AI, which is programmed to provide a predictable output based on the input it receives. Reactive machines always respond to identical situations in the exact same way every time, and they are not able to learn actions or conceive of past or future.

**Example**

1. Deep Blue, the chess-playing IBM supercomputer that bested world champion Garry Kasparov
2. Spam filters for our email that keep promotions and phishing attempts out of our inboxes
3. The Netflix recommendation engine

**Limited Memory AI**

Limited memory AI learns from the past and builds experiential knowledge by observing actions or data. This type of AI uses historical, observational data in combination with pre-programmed information to make predictions and perform complex classification tasks. It is the most widely-used kind of AI today.

**Example**

Autonomous vehicles use limited memory AI to observe other cars’ speed and direction, helping them “read the road” and adjust as needed. This process for understanding and interpreting incoming data makes them safer on the roads

**Theory of Mind AI**

Want to hold a meaningful conversation with an emotionally intelligent robot that looks and sounds like a real human being? That’s on the horizon with theory of mind AI. With this type of AI, machines will acquire true decision-making capabilities that are similar to humans. Machines with theory of mind AI will be able to understand and remember emotions, then adjust behavior based on those emotions as they interact with people.

**Example**

1. The Kismet robot head, developed by Professor Cynthia Breazeal, could recognize emotional signals on human faces and replicate those emotions on its own face.
2. Humanoid robot Sophia, developed by Hanson Robotics in Hong Kong, can recognize faces and respond to interactions with her own facial expressions.

**Self-aware AI**

The most advanced type of artificial intelligence is self-aware AI. When machines can be aware of their own emotions, as well as the emotions of others around them, they will have a level of consciousness and intelligence similar to human beings. This type of AI will have desires, needs, and emotions as well.

# Data Wrangling.

Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging.

## **Importance Of Data Wrangling**

Data Wrangling is a very important step. The below example will explain its importance as :

Books selling Website want to show top-selling books of different domains, according to user preference. For example, a new user search for motivational books, then they want to show those motivational books which sell the most or having a high rating, etc.

But on their website, there are plenty of raw data from different users. Here the concept of Data Munging or Data Wrangling is used. As we know Data is not Wrangled by System. This process is done by Data Scientists. So, the data Scientist will wrangle data in such a way that they will sort that motivational books that are sold more or have high ratings or user buy this book with these package of Books, etc. On the basis of that, the new user will make choice. This will explain the importance of Data wrangling.

## **Data Wrangling in Python**

Data Wrangling is a crucial topic for Data Science and Data Analysis. Pandas Framework of Python is used for Data Wrangling. Pandas is an open-source library specifically developed for Data Analysis and Data Science. The process like data sorting or filtration, Data grouping, etc.

Data wrangling in python deals with the below functionalities:

**Data exploration:** In this process, the data is studied, analyzed and understood by visualizing representations of data.

**Dealing with missing values:** Most of the datasets having a vast amount of data contain missing values of NaN, they are needed to be taken care of by replacing them with mean, mode, the most frequent value of the column or simply by dropping the row having a NaN value.

**Reshaping data:** In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.

**Filtering data:** Some times datasets are comprised of unwanted rows or columns which are required to be removed or filtered

**Other:** After dealing with the raw dataset with the above functionalities we get an efficient dataset as per our requirements and then it can be used for a required purpose like data analyzing, machine learning, data visualization, model training etc.

## **Data exploration** here we assign the data, and then we visualize the data in a tabular format. Machine Learning

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

Once you have trained the model, you can use it to reason over data that it hasn't seen before, and make predictions about those data. For example, let's say you want to build an application that can recognize a user's emotions based on their facial expressions. You can train a model by providing it with images of faces that are each tagged with a certain emotion, and then you can use that model in an application that can recognize any user's emotion.

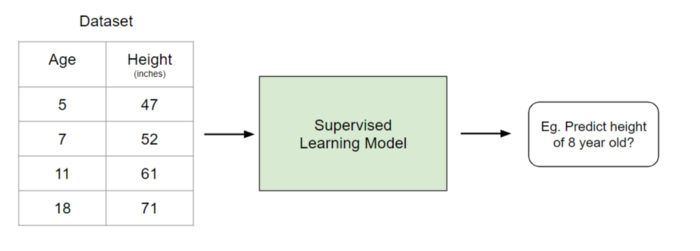
## Types of machine learning models

All machine learning models are categorized as either **supervised** or **unsupervised**. If the model is a supervised model, it’s then sub-categorized as either a **regression** or **classification** model. We’ll go over what these terms mean and the corresponding models that fall into each category below.

### Supervised Learning

Supervised learning involves learning a function that maps an input to an output based on example input-output pairs

For example, if I had a dataset with two variables, age (input) and height (output), I could implement a supervised learning model to predict the height of a person based on their age.



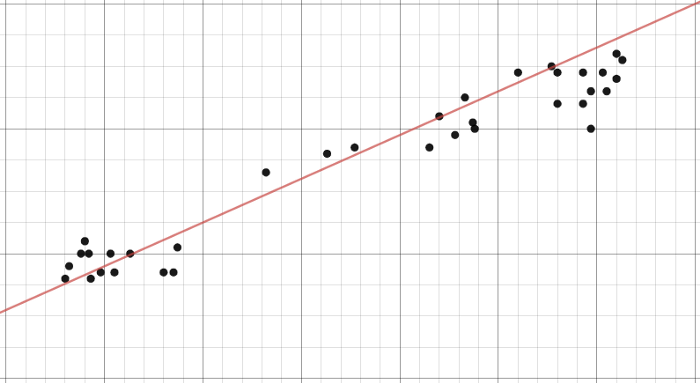
To re-iterate, within supervised learning, there are two sub-categories: regression and classification.

### Regression

In regression models, the output is continuous. Below are some of the most common types of regression models.

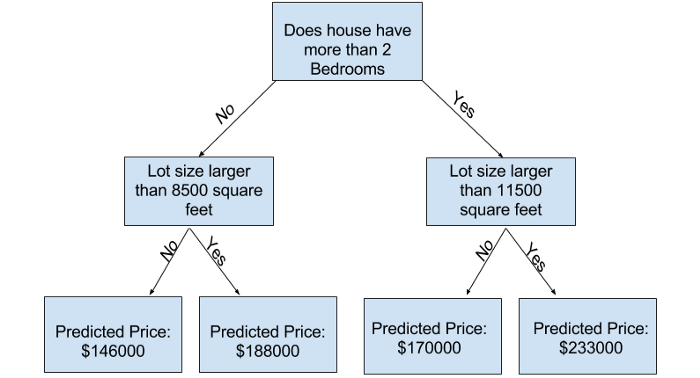
### Linear Regression

The idea of linear regression is simply finding a line that best fits the data. Extensions of linear regression include multiple linear regression (eg. finding a plane of best fit) and polynomial regression (eg. finding a curve of best fit).



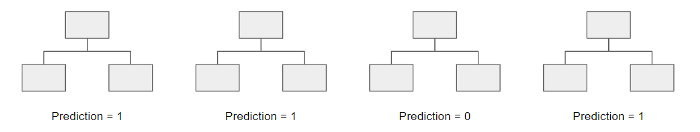
### Decision Tree

Decision trees are a popular model, used in operations research, strategic planning, and machine learning. Each square above is called a node, and the more nodes you have, the more accurate your decision tree will be (generally). The last nodes of the decision tree, where a decision is made, are called the leaves of the tree. Decision trees are intuitive and easy to build but fall short when it comes to accuracy.



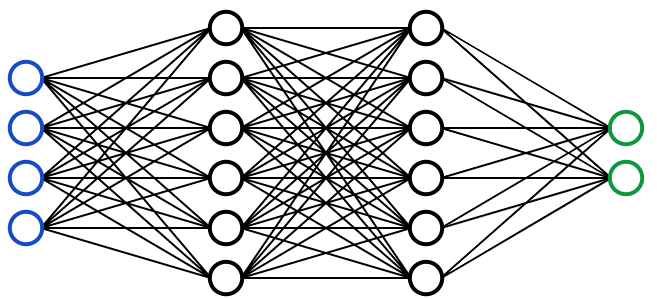
### Random Forest

Random forests are an ensemble learning technique that builds off of decision trees. Random forests involve creating multiple decision trees using bootstrapped datasets of the original data and randomly selecting a subset of variables at each step of the decision tree. The model then selects the mode of all of the predictions of each decision tree. What’s the point of this? By relying on a “majority wins” model, it reduces the risk of error from an individual tree.



For example, if we created one decision tree, the third one, it would predict 0. But if we relied on the mode of all 4 decision trees, the predicted value would be 1. This is the power of random forests.

### Neural Network



A Neural Network is essentially a network of mathematical equations. It takes one or more input variables, and by going through a network of equations, results in one or more output variables. You can also say that a neural network takes in a vector of inputs and returns a vector of outputs, but I won’t get into matrices in this article.

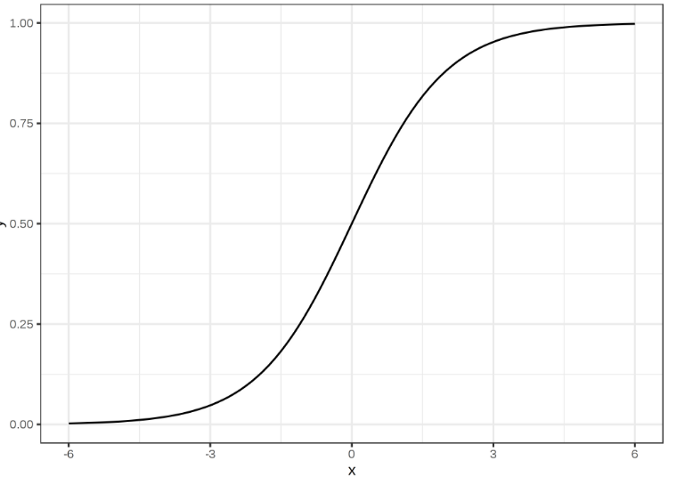
The blue circles represent the input layer, the black circles represent the hidden layers, and the green circles represent the output layer. Each node in the hidden layers represents both a linear function and an activation function that the nodes in the previous layer go through, ultimately leading to an output in the green circles.

### Classification

In classification models, the output is discrete. Below are some of the most common types of classification models.

### Logistic Regression

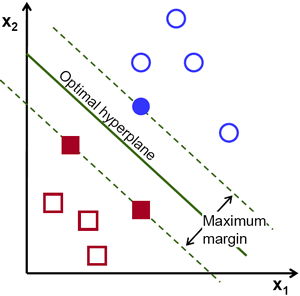
Logistic regression is similar to linear regression but is used to model the probability of a finite number of outcomes, typically two. There are a number of reasons why logistic regression is used over linear regression when modeling probabilities of outcomes. In essence, a logistic equation is created in such a way that the output values can only be between 0 and 1 (see below).



### Support Vector Machine

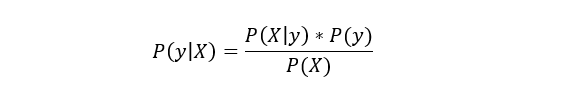
A Support Vector Machine is a supervised classification technique that can actually get pretty complicated but is pretty intuitive at the most fundamental level.

Let’s assume that there are two classes of data. A support vector machine will find a hyperplane or a boundary between the two classes of data that maximizes the margin between the two classes (see below). There are many planes that can separate the two classes, but only one plane can maximize the margin or distance between the classes.

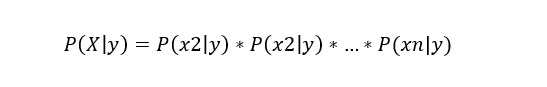


### Naive Bayes

Naive Bayes is another popular classifier used in Data Science. The idea behind it is driven by Bayes Theorem:



In plain English, this equation is used to answer the following question. “What is the probability of y (my output variable) given X? And because of the naive assumption that variables are independent given the class, you can say that:



As well, by removing the denominator, we can then say that P(y|X) is proportional to the right-hand side.

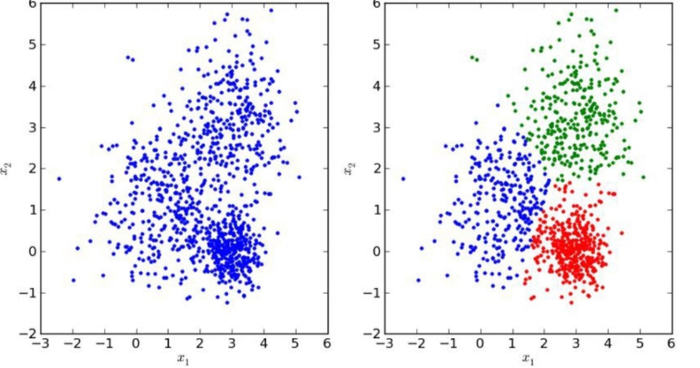


Therefore, the goal is to find the class y with the maximum proportional probability.

### Decision Tree, Random Forest, Neural Network

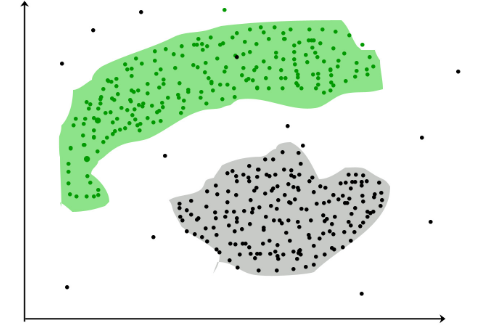
These models follow the same logic as previously explained. The only difference is that that output is discrete rather than continuous.

### Unsupervised Learning



Unlike supervised learning, unsupervised learning is used to draw inferences and find patterns from input data without references to labeled outcomes. Two main methods used in unsupervised learning include clustering and dimensionality reduction.

### Clustering



Clustering is an unsupervised technique that involves the grouping, or clustering, of data points. It’s frequently used for customer segmentation, fraud detection, and document classification.

Common clustering techniques include k-means clustering, hierarchical clustering, mean shift clustering, and density-based clustering. While each technique has a different method in finding clusters, they all aim to achieve the same thing.

### Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables [2]. In simpler terms, its the process of reducing the dimension of your feature set (in even simpler terms, reducing the number of features). Most dimensionality reduction techniques can be categorized as either feature elimination or feature extraction.

A popular method of dimensionality reduction is called principal component analysis.

### Principal Component Analysis (PCA)

In the simplest sense, PCA involves project higher dimensional data (eg. 3 dimensions) to a smaller space (eg. 2 dimensions). This results in a lower dimension of data, (2 dimensions instead of 3 dimensions) while keeping all original variables in the model.

## Choosing the Right Model

Well, there is no straightforward and sure-shot answer to this question. The answer depends on many factors like the problem statement and the kind of output you want, type and size of the data, the available computational time, number of features, and observations in the data, to name a few.

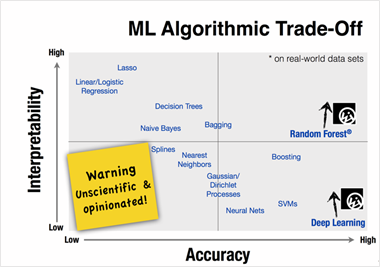
### 1. Size of the training data

It is usually recommended to gather a good amount of data to get reliable predictions. However, many a time, the availability of data is a constraint. So, if the training data is smaller or if the dataset has a fewer number of observations and a higher number of features like genetics or textual data, choose algorithms with high bias/low variance like Linear regression, Naïve Bayes, or Linear SVM.

If the training data is sufficiently large and the number of observations is higher as compared to the number of features, one can go for low bias/high variance algorithms like KNN, Decision trees, or kernel SVM.

### 2. Accuracy and/or Interpretability of the output

Accuracy of a model means that the function predicts a response value for a given observation, which is close to the true response value for that observation. A highly interpretable algorithm (restrictive models like Linear Regression) means that one can easily understand how any individual predictor is associated with the response while the flexible models give higher accuracy at the cost of low interpretability.



### 3. Speed or Training time

Higher accuracy typically means higher training time. Also, algorithms require more time to train on large training data. In real-world applications, the choice of algorithm is driven by these two factors predominantly.

Algorithms like Naïve Bayes and Linear and Logistic regression are easy to implement and quick to run. Algorithms like SVM, which involve tuning of parameters, Neural networks with high convergence time, and random forests, need a lot of time to train the data.

### 4. Linearity

Many algorithms work on the assumption that classes can be separated by a straight line (or its higher-dimensional analog). Examples include logistic regression and support vector machines. Linear regression algorithms assume that data trends follow a straight line. If the data is linear, then these algorithms perform quite good.

However, not always is the data is linear, so we require other algorithms which can handle high dimensional and complex data structures. Examples include kernel SVM, random forest, neural nets.

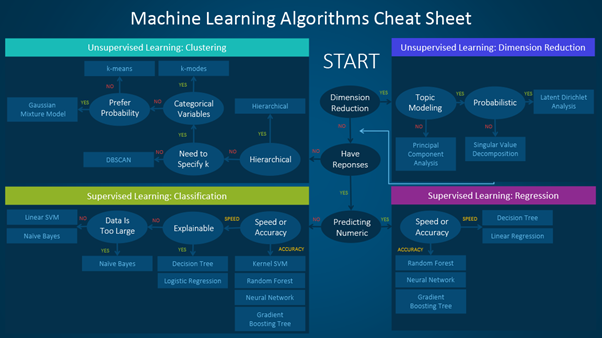
The best way to find out the linearity is to either fit a linear line or run a logistic regression or SVM and check for residual errors. A higher error means the data is not linear and would need complex algorithms to fit.

### 5. Number of features

The dataset may have a large number of features that may not all be relevant and significant. For a certain type of data, such as genetics or textual, the number of features can be very large compared to the number of data points.

A large number of features can bog down some learning algorithms, making training time unfeasibly long. SVM is better suited in case of data with large feature space and lesser observations. PCA and feature selection techniques should be used to reduce dimensionality and select important features.

Here is a handy cheat sheet that details the algorithms you can use for different types of machine learning problems.



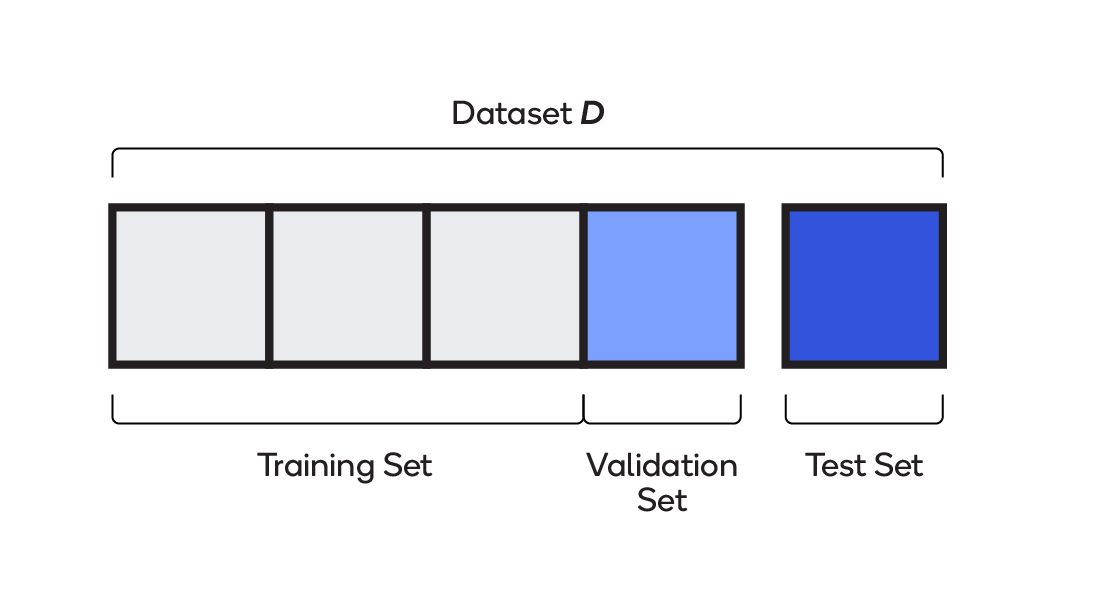
## Training, Testing and Evaluating Machine Learning Models

### Model training

Model training for Deep learning/Machine Learning includes splitting the dataset, tuning hyperparameters and performing batch normalization.

### Splitting the dataset

The data collected for training needs to be split into three different sets: training, validation and test.



Training — Up to 75 percent of the total dataset is used for training. The model learns on the training set; in other words, the set is used to assign the weights and biases that go into the model.

Validation — Between 15 and 20 percent of the data is used while the model is being trained, for evaluating initial accuracy, seeing how the model learns and fine-tuning hyperparameters. The model sees validation data but does not use it to learn weights and biases.

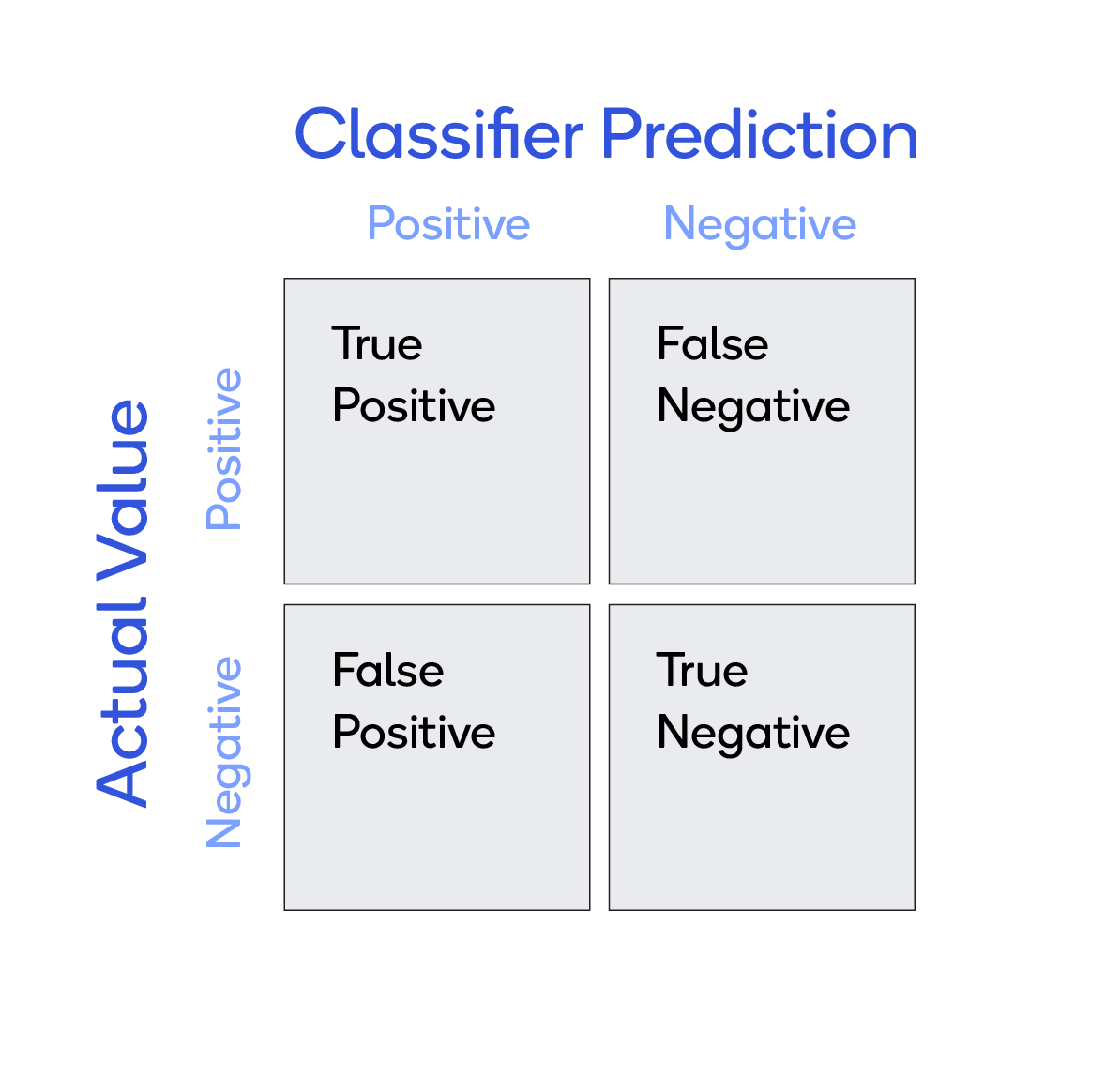
Test — Between 10 and 25 percent of the data is used for final evaluation. Having never seen this dataset, the model is free of any of its bias.

### Model evaluation and testing

Once a model has been trained, performance is gauged according to a confusion matrix and precision/accuracy metrics.

### Confusion matrix

A confusion matrix describes the performance of a classifier model, as in the 2x2 matrix depicted below.



Consider a simple classifier that predicts whether a patient has cancer or not. There are four possible results:

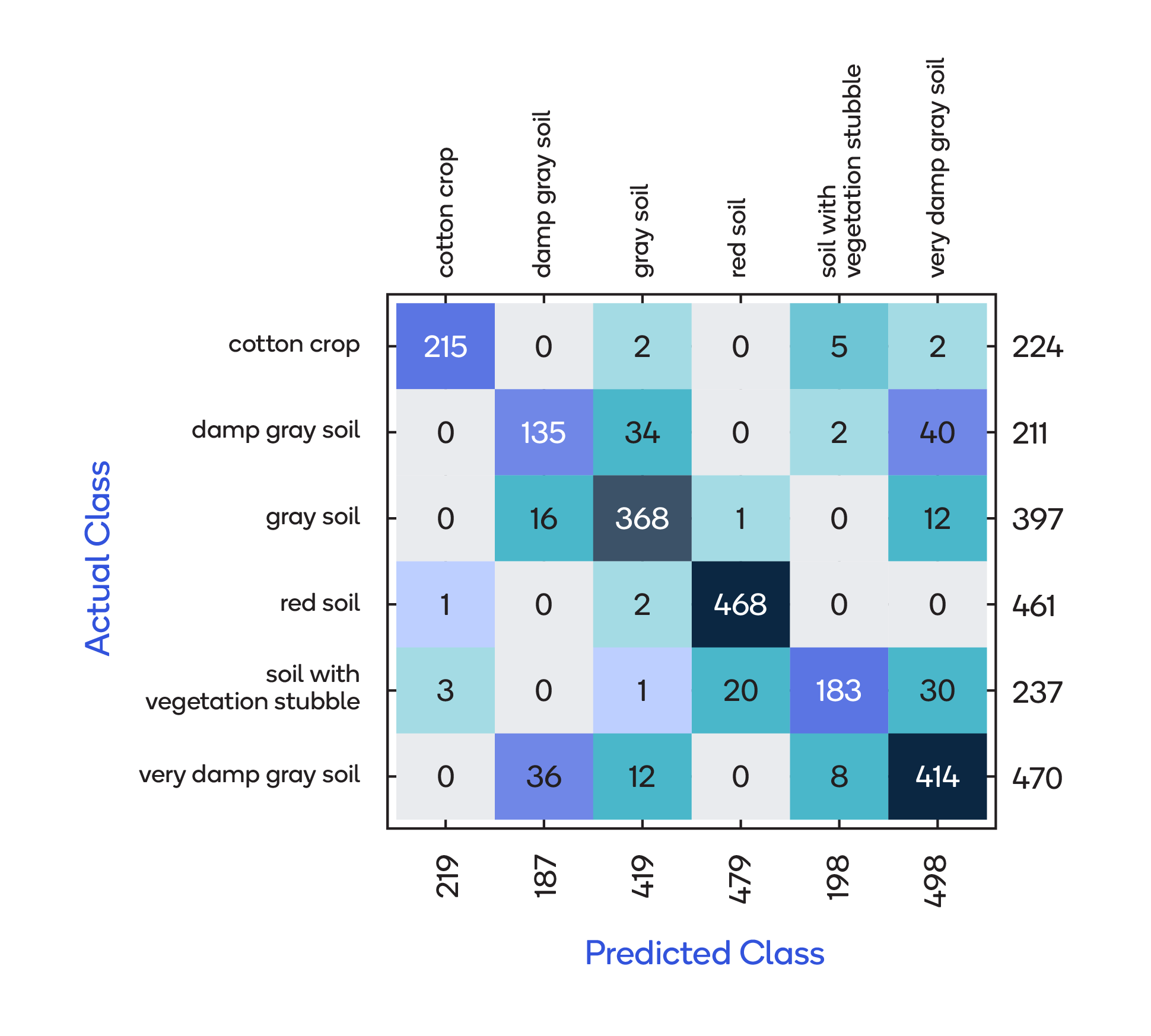
True positives (TP) — Prediction was yes and the patient does have cancer.

True negatives (TN) — Prediction was no and the patient does not have cancer.

False positives (FP) — Prediction was yes, but the patient does not have cancer (also known as a "Type I error").

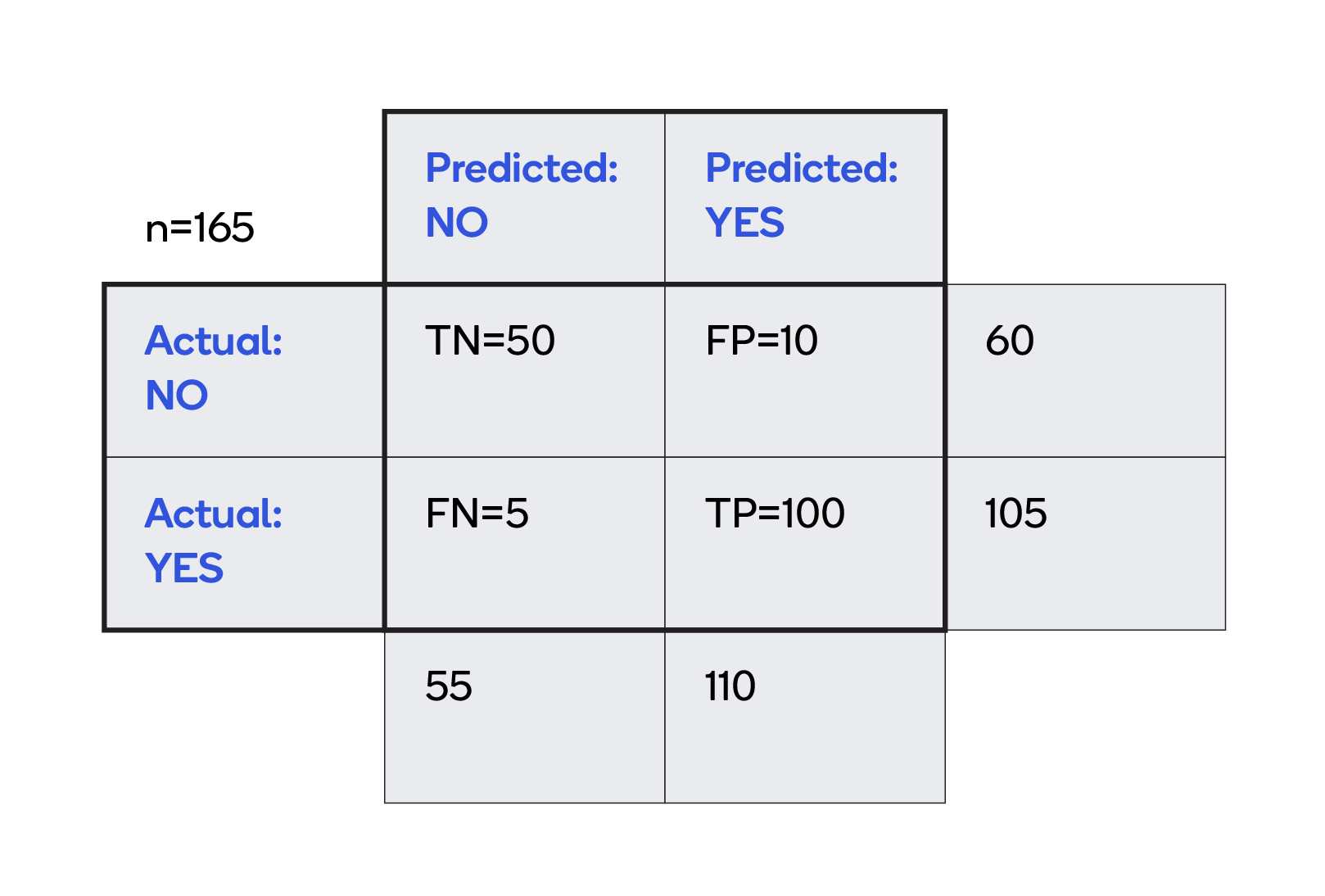
False negatives (FN) — Prediction was no, but the patient does have cancer (also known as a "Type II error")

A confusion matrix can hold more than 2 classes per axis, as shown here:



### Precision / Accuracy

It is also useful to calculate the precision and accuracy based on classifier prediction and actual value.



Accuracy is a measure of how often, over all observations, the classifier is correct. The calculation, based on the grid above, is (TP+TN)/total = (100+50)/(60+105) = 0.91.

Precision is a measure of how often the actual value is Yes when the prediction is Yes. In this case, that calculation is TP/predicted yes = 100/(100+10) = 0.91.

# 3.PROPOSED MENTHOD WITH ARCHITECTURE

# Regression Analysis in Machine learning

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

We can understand the concept of regression analysis using the below example:

**Example:** Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:



Now, the company wants to do the advertisement of $200 in the year 2019 **and wants to know the prediction about the sales for this year**. So to solve such type of prediction problems in machine learning, we need regression analysis.

Regression is a [supervised learning technique](https://www.javatpoint.com/supervised-machine-learning) which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, **"Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum."** The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

* Prediction of rain using temperature and other factors
* Determining Market trends
* Prediction of road accidents due to rash driving.

## **Terminologies Related to the Regression Analysis:**

* **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
* **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
* **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
* **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
* **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

## **Why do we use Regression Analysis?**

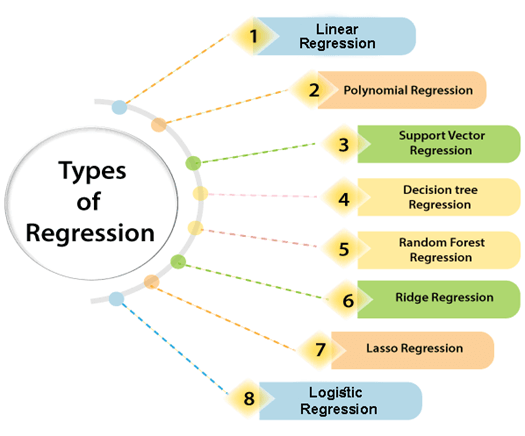
As mentioned above, Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately. So for such case we need Regression analysis which is a statistical method and used in machine learning and data science. Below are some other reasons for using Regression analysis:

* Regression estimates the relationship between the target and the independent variable.
* It is used to find the trends in data.
* It helps to predict real/continuous values.
* By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.

## **Types of Regression**

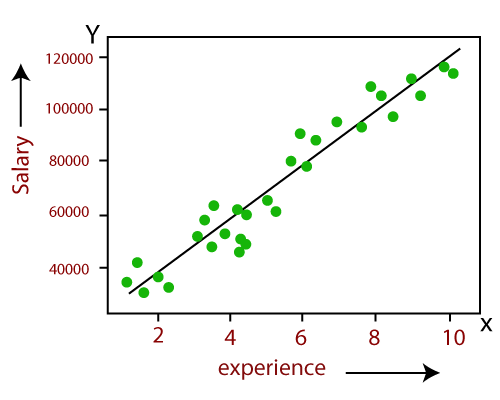
There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables. Here we are discussing some important types of regression which are given below:

* **Linear Regression**
* **Logistic Regression**
* **Polynomial Regression**
* **Support Vector Regression**
* **Decision Tree Regression**
* **Random Forest Regression**
* **Ridge Regression**
* **Lasso Regression:**



### **Linear Regression:**

* Linear regression is a statistical regression method which is used for predictive analysis.
* It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
* It is used for solving the regression problem in machine learning.
* Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
* If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
* The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



* Below is the mathematical equation for Linear regression:

1. Y= aX+b

**Here, Y = dependent variables (target variables),**  
**X= Independent variables (predictor variables),**  
**a and b are the linear coefficients**

Some popular applications of linear regression are:

* **Analyzing trends and sales estimates**
* **Salary forecasting**
* **Real estate prediction**
* **Arriving at ETAs in traffic.**

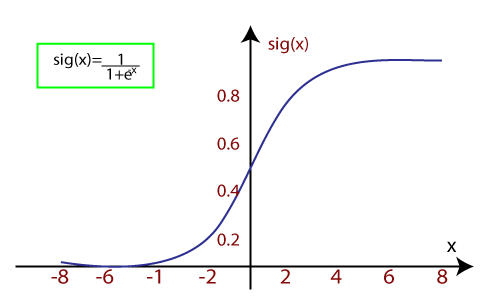
### **Logistic Regression:**

* Logistic regression is another supervised learning algorithm which is used to solve the classification problems. In **classification problems**, we have dependent variables in a binary or discrete format such as 0 or 1.
* Logistic regression algorithm works with the categorical variable such as 0 or 1, Yes or No, True or False, Spam or not spam, etc.
* It is a predictive analysis algorithm which works on the concept of probability.
* Logistic regression is a type of regression, but it is different from the linear regression algorithm in the term how they are used.
* Logistic regression uses **sigmoid function** or logistic function which is a complex cost function. This sigmoid function is used to model the data in logistic regression. The function can be represented as:

Regression Analysis in Machine learning

* f(x)= Output between the 0 and 1 value.
* x= input to the function
* e= base of natural logarithm.

When we provide the input values (data) to the function, it gives the S-curve as follows:



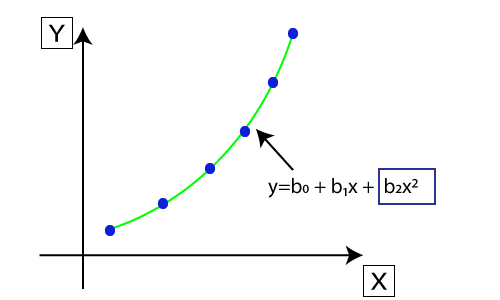
* It uses the concept of threshold levels, values above the threshold level are rounded up to 1, and values below the threshold level are rounded up to 0.

There are three types of logistic regression:

* **Binary(0/1, pass/fail)**
* **Multi(cats, dogs, lions)**
* **Ordinal(low, medium, high)**

### **Polynomial Regression:**

* Polynomial Regression is a type of regression which models the **non-linear dataset** using a linear model.
* It is similar to multiple linear regression, but it fits a non-linear curve between the value of x and corresponding conditional values of y.
* Suppose there is a dataset which consists of datapoints which are present in a non-linear fashion, so for such case, linear regression will not best fit to those datapoints. To cover such datapoints, we need Polynomial regression.
* I**n Polynomial regression, the original features are transformed into polynomial features of given degree and then modeled using a linear model.** Which means the datapoints are best fitted using a polynomial line.



* The equation for polynomial regression also derived from linear regression equation that means Linear regression equation Y= b0+ b1x, is transformed into Polynomial regression equation Y= b0+b1x+ b2x2+ b3x3+.....+ bnxn.
* Here Y is the **predicted/target output, b0, b1,... bn are the regression coefficients**. x is our **independent/input variable**.
* The model is still linear as the coefficients are still linear with quadratic

#### Note:**This is different from Multiple Linear regression in such a way that in Polynomial regression, a single element has different degrees instead of multiple variables with the same degree.**

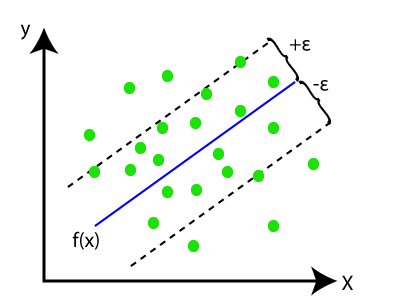
### **Support Vector Regression:**

Support Vector Machine is a supervised learning algorithm which can be used for regression as well as classification problems. So if we use it for regression problems, then it is termed as Support Vector Regression.

Support Vector Regression is a regression algorithm which works for continuous variables. Below are some keywords which are used in **Support Vector Regression**:

* **Kernel:** It is a function used to map a lower-dimensional data into higher dimensional data.
* **Hyperplane:** In general SVM, it is a separation line between two classes, but in SVR, it is a line which helps to predict the continuous variables and cover most of the datapoints.
* **Boundary line:** Boundary lines are the two lines apart from hyperplane, which creates a margin for datapoints.
* **Support vectors:** Support vectors are the datapoints which are nearest to the hyperplane and opposite class.

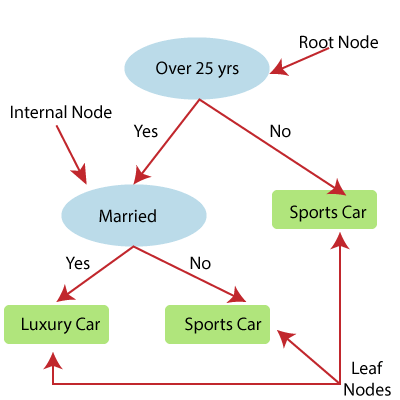
In SVR, we always try to determine a hyperplane with a maximum margin, so that maximum number of datapoints are covered in that margin. **The main goal of SVR is to consider the maximum datapoints within the boundary lines and the hyperplane (best-fit line) must contain a maximum number of datapoints**. Consider the below image:



Here, the blue line is called hyperplane, and the other two lines are known as boundary lines.

### **Decision Tree Regression:**

* Decision Tree is a supervised learning algorithm which can be used for solving both classification and regression problems.
* It can solve problems for both categorical and numerical data
* Decision Tree regression builds a tree-like structure in which each internal node represents the "test" for an attribute, each branch represent the result of the test, and each leaf node represents the final decision or result.
* A decision tree is constructed starting from the root node/parent node (dataset), which splits into left and right child nodes (subsets of dataset). These child nodes are further divided into their children node, and themselves become the parent node of those nodes. Consider the below image:

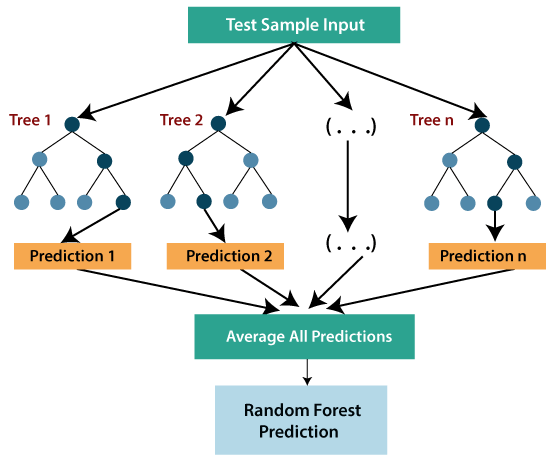


Above image showing the example of Decision Tee regression, here, the model is trying to predict the choice of a person between Sports cars or Luxury car.

* Random forest is one of the most powerful supervised learning algorithms which is capable of performing regression as well as classification tasks.
* The Random Forest regression is an ensemble learning method which combines multiple decision trees and predicts the final output based on the average of each tree output. The combined decision trees are called as base models, and it can be represented more formally as:

g(x)= f0(x)+ f1(x)+ f2(x)+....

* Random forest uses **Bagging or Bootstrap Aggregation** technique of ensemble learning in which aggregated decision tree runs in parallel and do not interact with each other.
* With the help of Random Forest regression, we can prevent Overfitting in the model by creating random subsets of the dataset.



### **Ridge Regression:**

* Ridge regression is one of the most robust versions of linear regression in which a small amount of bias is introduced so that we can get better long term predictions.
* The amount of bias added to the model is known as **Ridge Regression penalty**. We can compute this penalty term by multiplying with the lambda to the squared weight of each individual features.
* The equation for ridge regression will be:

Regression Analysis in Machine learning

* A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.
* Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.
* It helps to solve the problems if we have more parameters than samples.

### **Lasso Regression:**

* Lasso regression is another regularization technique to reduce the complexity of the model.
* It is similar to the Ridge Regression except that penalty term contains only the absolute weights instead of a square of weights.
* Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
* It is also called as **L1 regularization**. The equation for Lasso regression will be:

Regression Analysis in Machine learning

# 4.METHODLOGY

# Linear Regression in Machine Learning

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 algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

y= a0+a1x+ ε

**Here,**

Y= Dependent Variable (Target Variable)  
X= Independent Variable (predictor Variable)  
a0= intercept of the line (Gives an additional degree of freedom)  
a1 = Linear regression coefficient (scale factor to each input value).  
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

## **Types of Linear Regression**

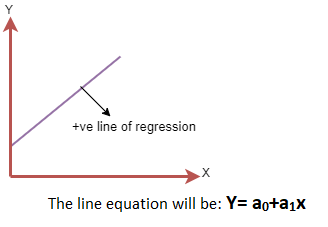
Linear regression can be further divided into two types of the algorithm:

* **Simple\_Linear\_Regression:**  
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
* **Multiple\_Linear\_regression:**  
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

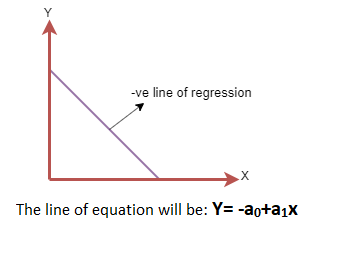
## **Linear Regression Line**

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

* **Positive Linear Relationship:**  
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



* **Negative Linear Relationship:**  
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



## **Finding the best fit line:**

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.

### **Cost function-**

* The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
* Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
* We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

Linear Regression in Machine Learning

**Where,**

N=Total number of observation  
Yi = Actual value  
(a1xi+a0)= Predicted value.

**Residuals:** The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

### **Gradient Descent:**

* Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
* A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
* It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

## **Model Performance:**

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

**1. R-squared method:**

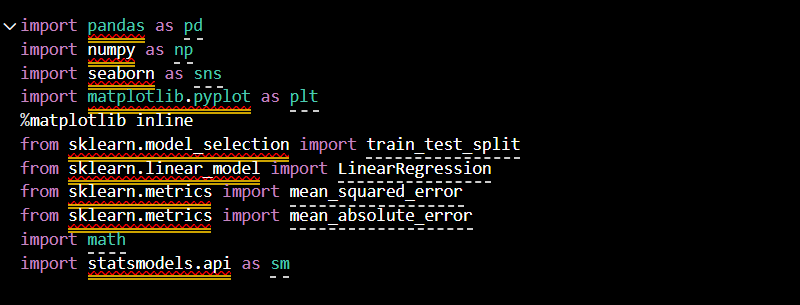
* R-squared is a statistical method that determines the goodness of fit.
* It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
* The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
* It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
* It can be calculated from the below formula:

Linear Regression in Machine Learning

# 5.IMPLEMENTATION

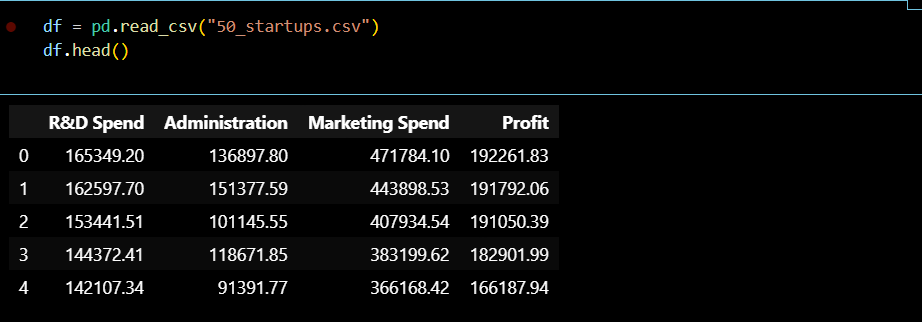
Here is the total process of implementation:

This about the complete libraries that are used in this Project.



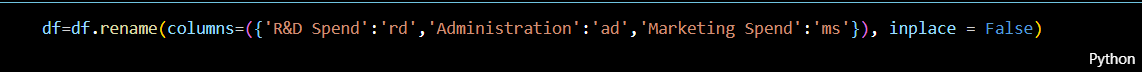
**Figure 5.1**

This is the step of reading csv file from the folder

****

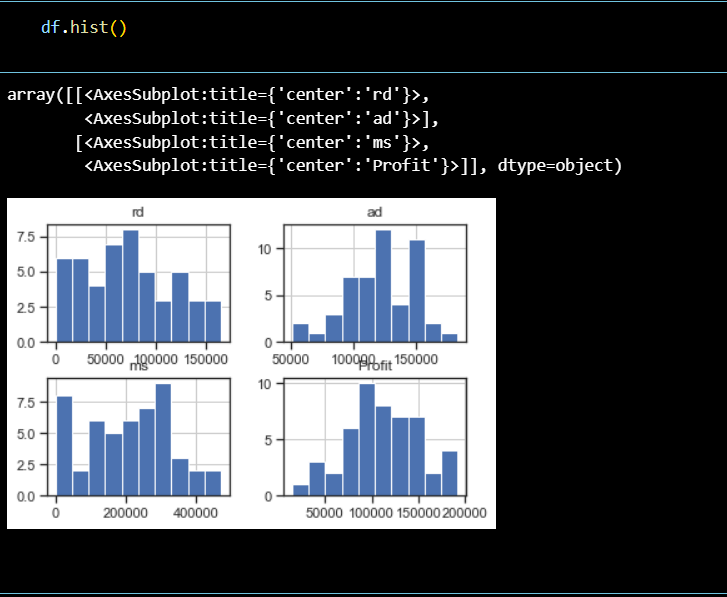
**Figure 5.2**

Herei made names into shorter for code convenience.

****

**Figure 5.3**

Here I have drawn a hist. This histogram is a chart that displays numeric data in ranges, where each bar represents how frequently numbers fall into a particular range.

****

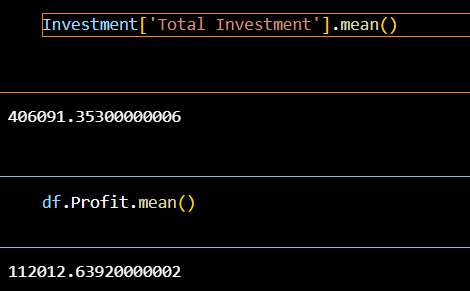
**Figure 5.4**

In next I have made whole attributes of investing capita’s as investment.

****

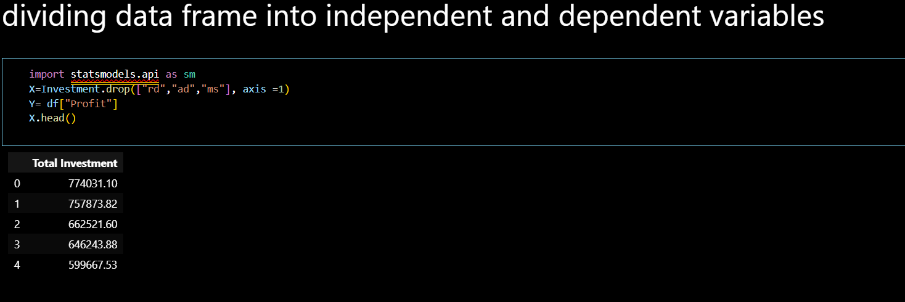
**Figure 5.5**

Here we have mean of both Total Investment and Profit.

****

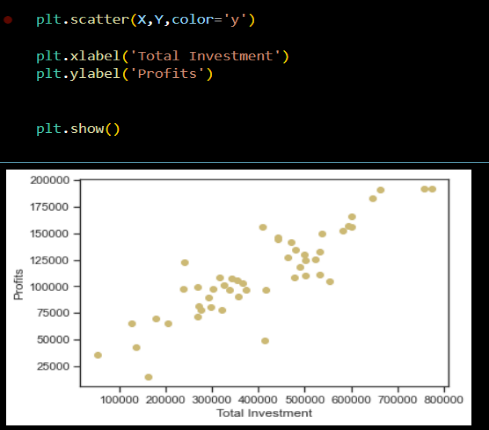
**Figure 5.6**

Here I have divided data frame into X,Yvariables.

****

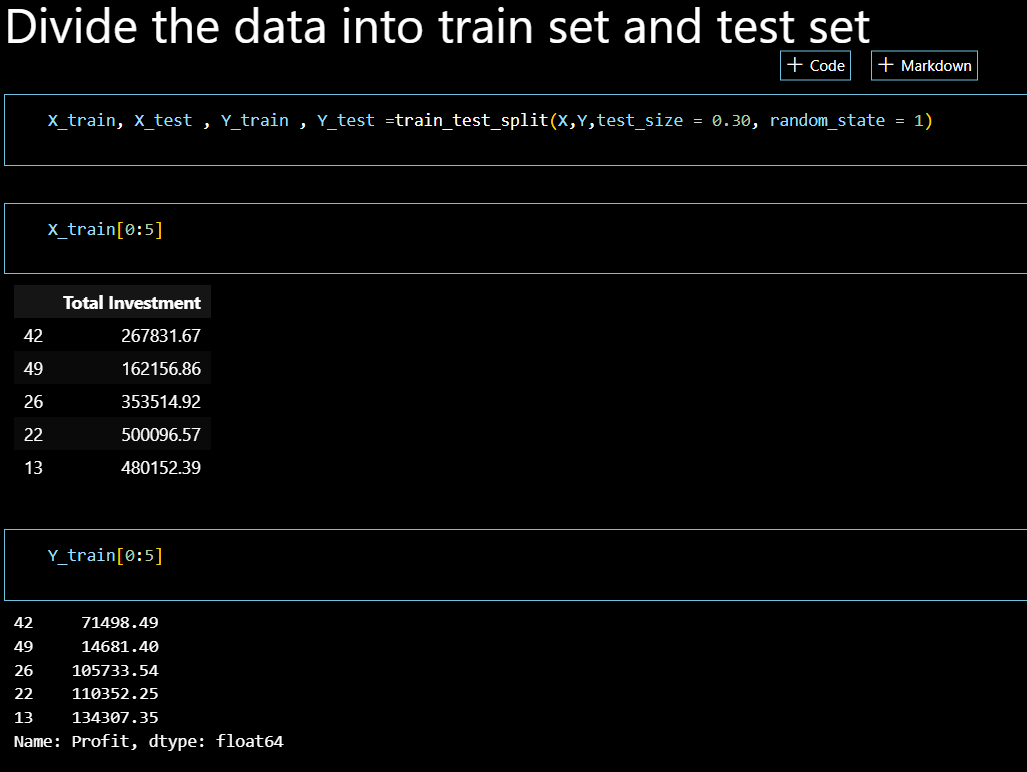
**Figure 5.7**

Here I plotted independent variables against dependent variable.

****

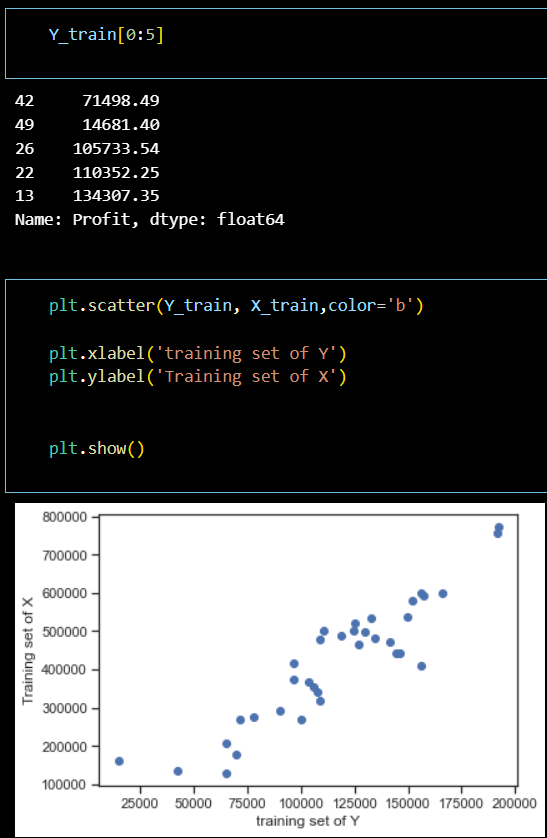
**Figure 5.8**

Dividing data into 70% Training data and 30% Test data.

****

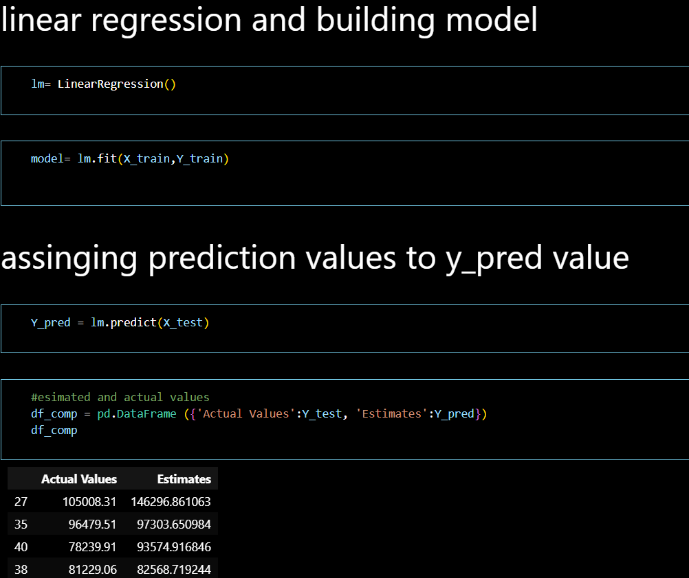
**Figure 5.9**

Plotting between Training sets.

****

**Figure 5.10**

Building the linear regression model.

****

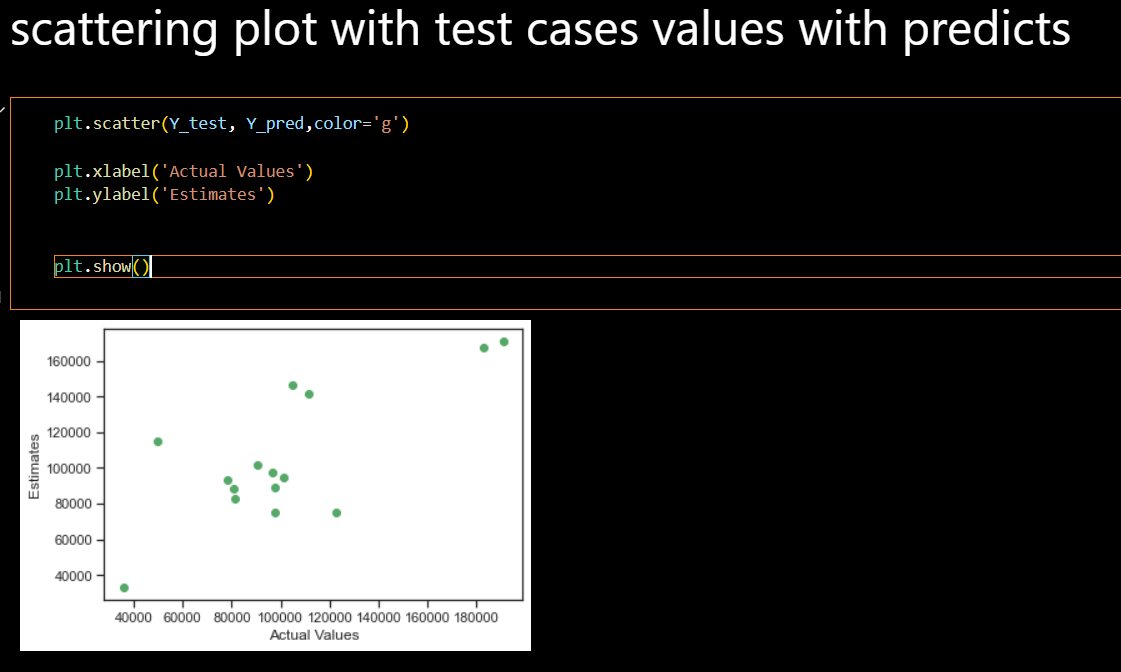
**Figure 5.11**

Regressionmetrics and model scores.

****

**Figure 5.12**

Scattering Plot with test cases values.

****

**Figure 5.13**

# 6.CONCLUSION

The competition goal is predicted the profit of startup profit on the bases of data provided which are on the bases of Research and Development Spend(R&D Spend), Administration Spend, Marketing Spend. I used Linear regression in this model because we have to predict profit(dependent variable) on bases of multiple field(independent variables) as a one enitity i.e total investment.This model can help those people who want to invest in startup company by analysing profit of the company.Here 50 startups dataset containing 4 columns  like “R&D Spend”, “Administration”, “Marketing Spend”, “Profit”.In this dataset first 3 columns provides you spending on Research , Administration and Marketing respectively. At last in Test cases I have predicted the Profits with those investment criterias.Which is 30% of data set (15) and a scattered plot is drawn for better visualisation.