Regression Analysis

Regression Analysis

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.

It predicts continuous/real values such as **temperature, age, salary, price,** etc.

It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

- ***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.***

- 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.

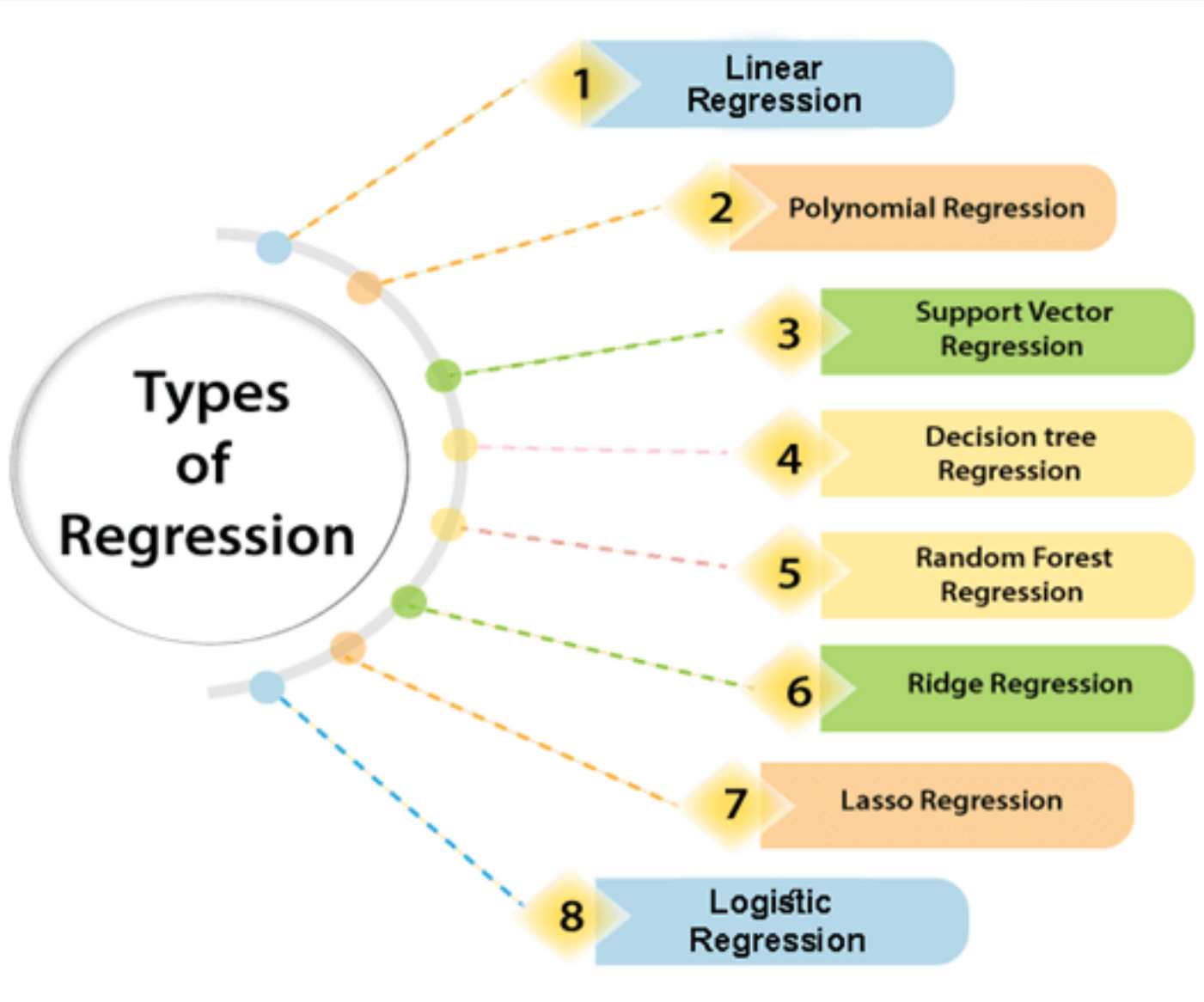
Terminologies

- **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?

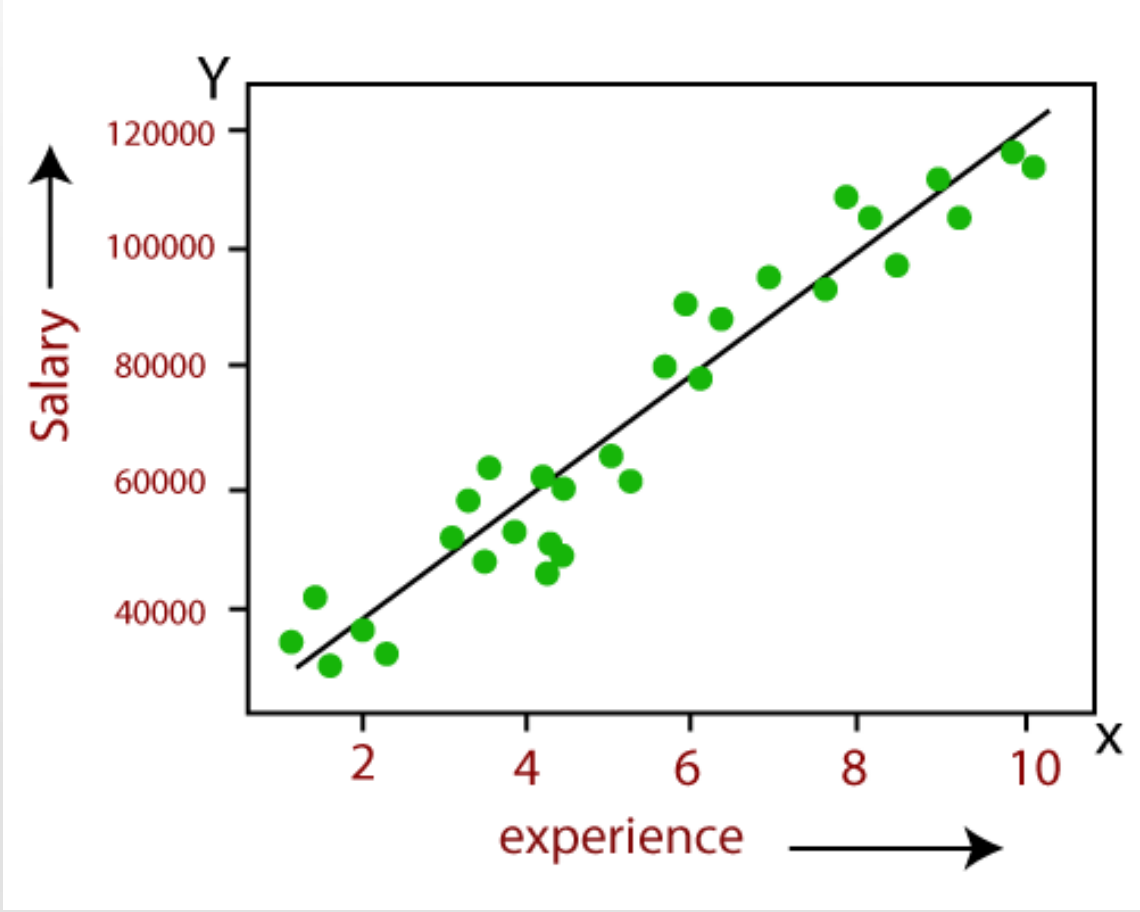
- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the **most important factor, the least important factor, and how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

Types of Regression



Linear Regression

- 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**.



➢ Mathematical equation for Linear regression:

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:

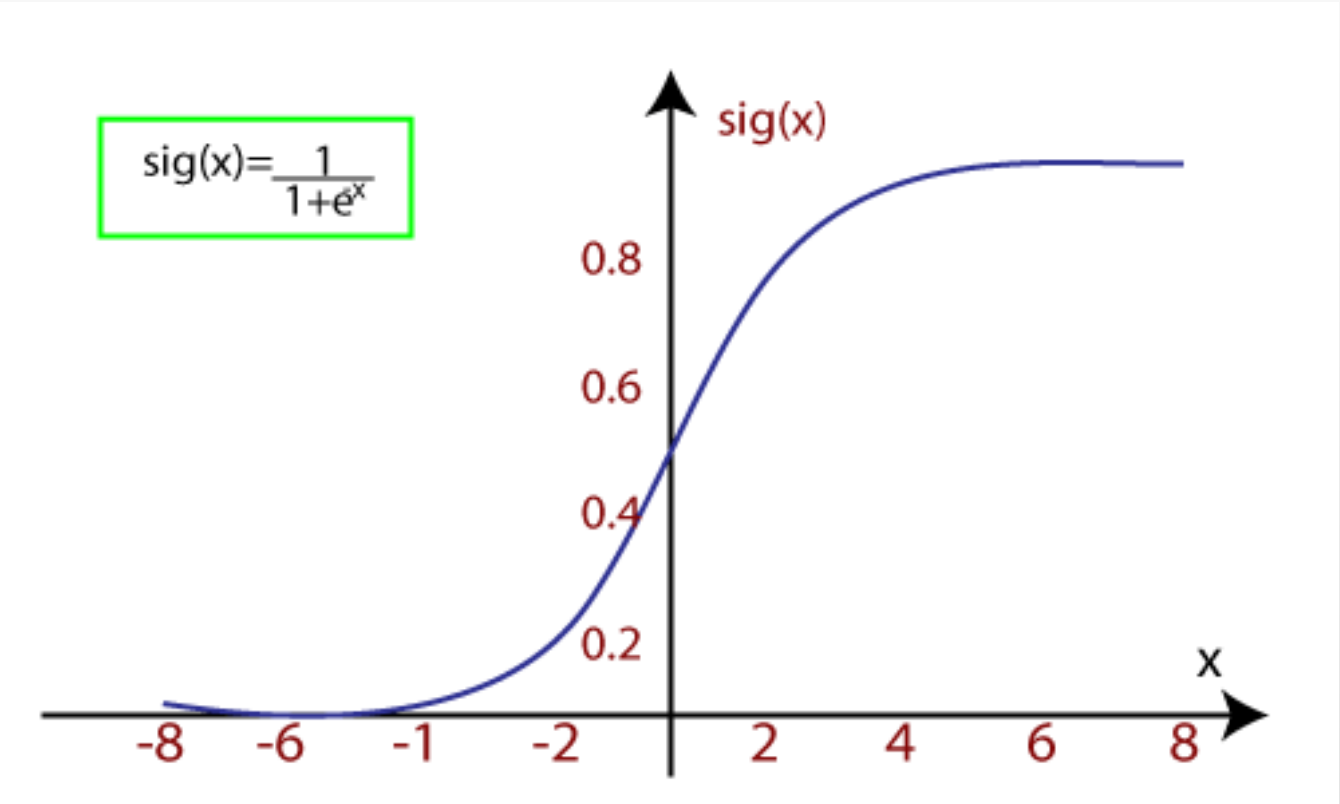
- **Analyzing trends and sales estimates**
- **Salary forecasting**
- **Real estate prediction**
- **Arriving at ETAs in traffic.**

Logistic Regression

- Logistic regression 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 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:

$$f(x)= \frac{1}{1+e^{-x}}$$

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

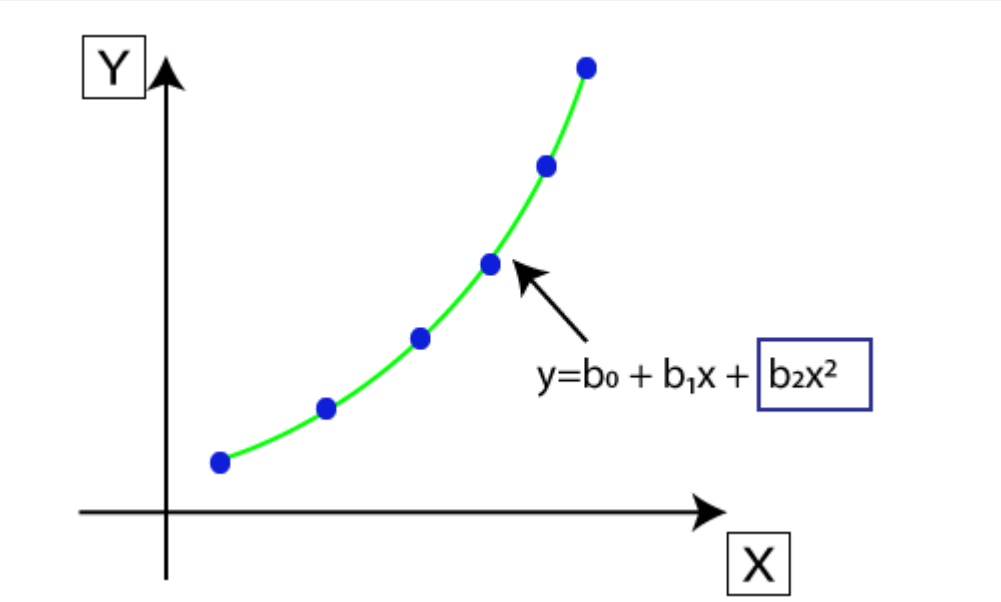


It uses the concept of threshold levels, values above the threshold level are rounded up to 1, and values below the threshold level are rounded up to 0.

- There are three types of logistic regression:
- **Binary(0/1, pass/fail)**
- **Multi(cats, dogs, lions)**
- **Ordinal(low, medium, high)**

Polynomial Regression

- 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.

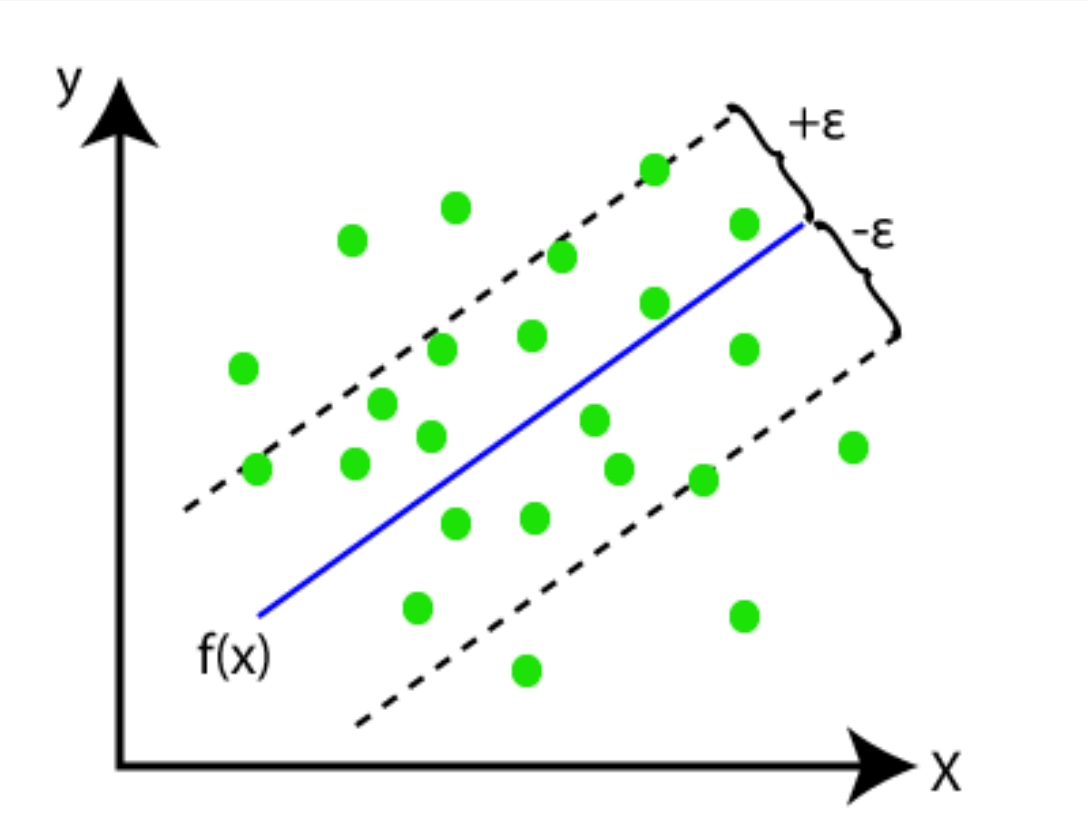


- The equation for polynomial regression also derived from linear regression equation that means Linear regression equation Y= b_0+ b_1x, is transformed into Polynomial regression equation Y= b_0+b_1x+ b_2x^2+ b_3x^3+.....+ b_nx^n.
- Here Y is the **predicted/target output, b_0, b_1,... b_n are the regression coefficients**. x is our **independent/input variable**.

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**:

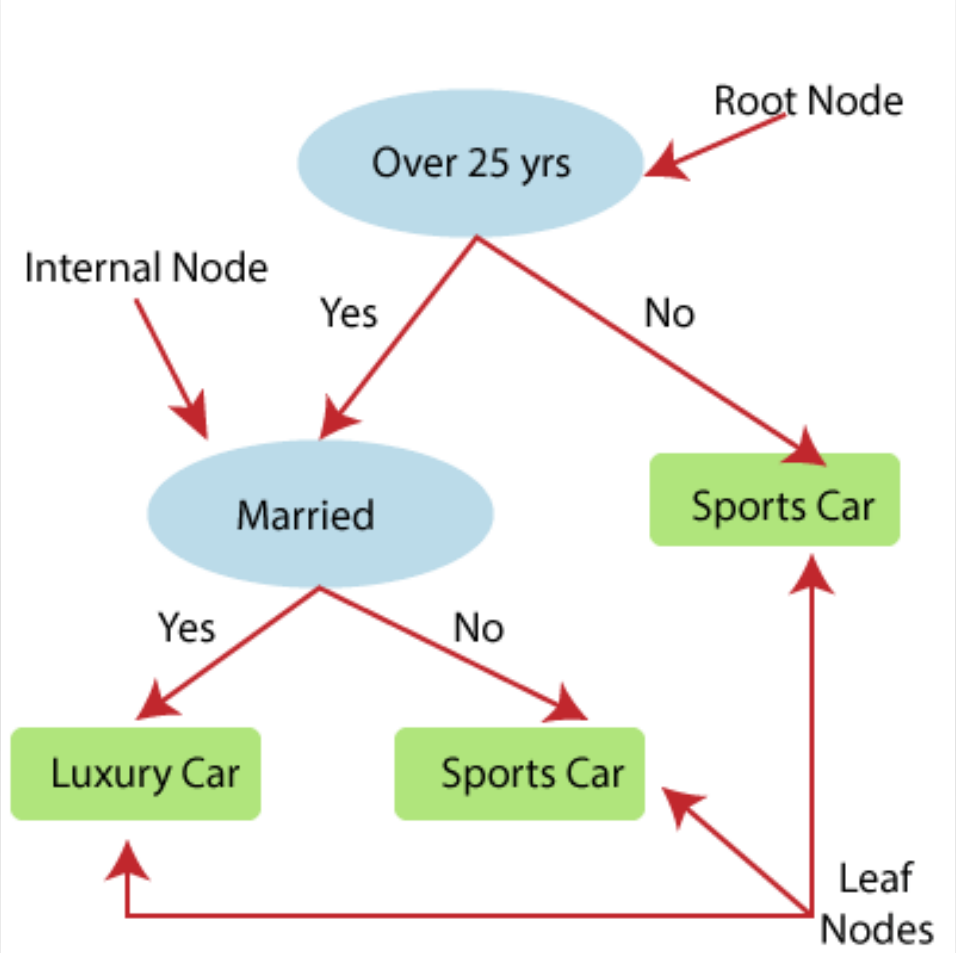
- **Kernel:** It is a function used to map a lower-dimensional data into higher dimensional data.
- **Hyperplane:** In general SVM, it is a separation line between two classes, but in SVR, it is a line which helps to predict the continuous variables and cover most of the datapoints.
- **Boundary line:** 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.



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

Decision Tree Regression

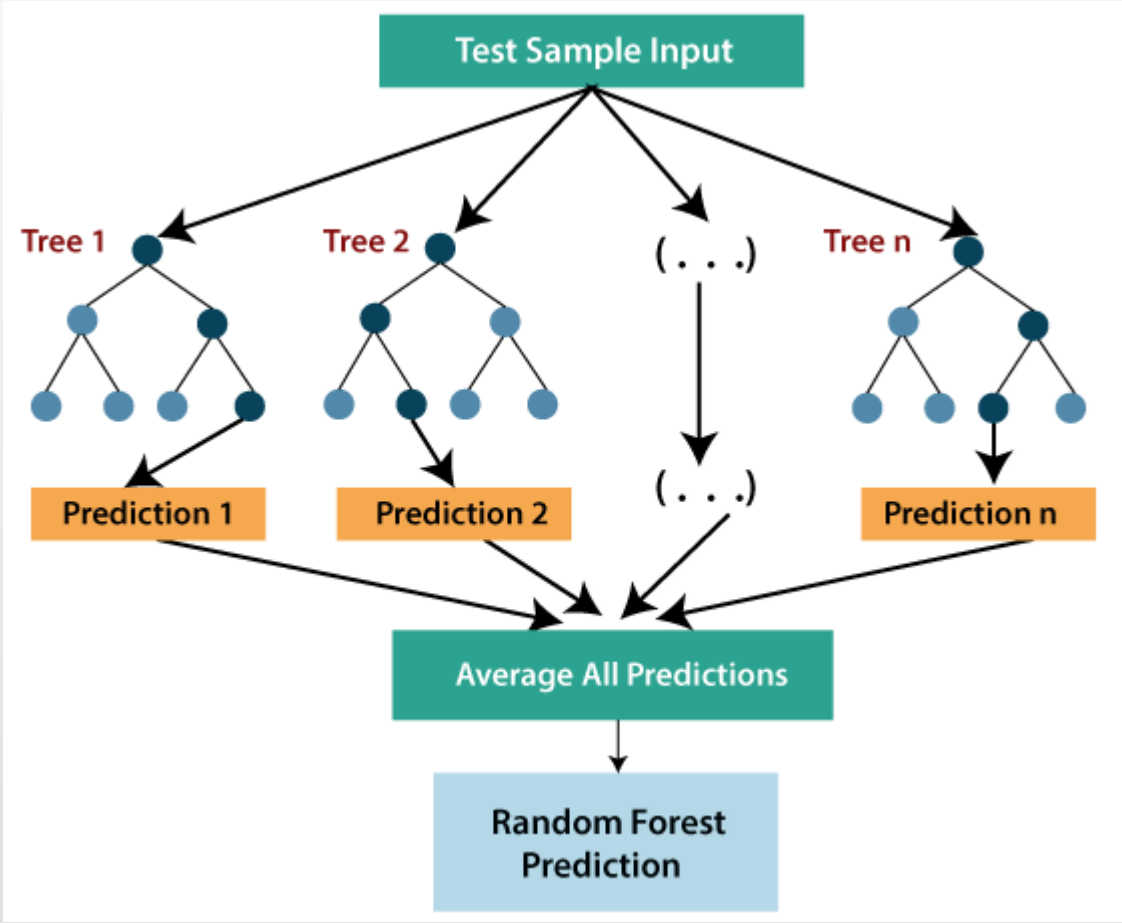
- Decision Tree regression builds a tree-like structure in which each internal node represents the "test" for an attribute, each branch represent the result of the test, and each leaf node represents the final decision or result.



here, the model is trying to predict the choice of a person between Sports cars or Luxury car.

- 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:

$$g(x)= f_0(x)+ f_1(x)+ f_2(x)+....$$



- 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.

$$L(x, y) = \text{Min}\left(\sum_{i=1}^{n} (y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n} (w_i)^2 \right)$$

- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.

Lasso Regression

- It is similar to the Ridge Regression except that penalty term contains only the absolute weights instead of a square of weights.

- It is also called as **L1 regularization**. The equation for Lasso regression will be:

$$L(x, y) = \text{Min}\left(\sum_{i=1}^{n} (y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n} |\mathbf{w}_i| \right)$$