**VARIOUS REGRESSION MODELS**

**Introduction**

* ***Regression*** prediction can involve **multiple** variables.
* The **less** the distance between **datapoints** & **regression curve**, the **more** the chances for accurate prediction are.

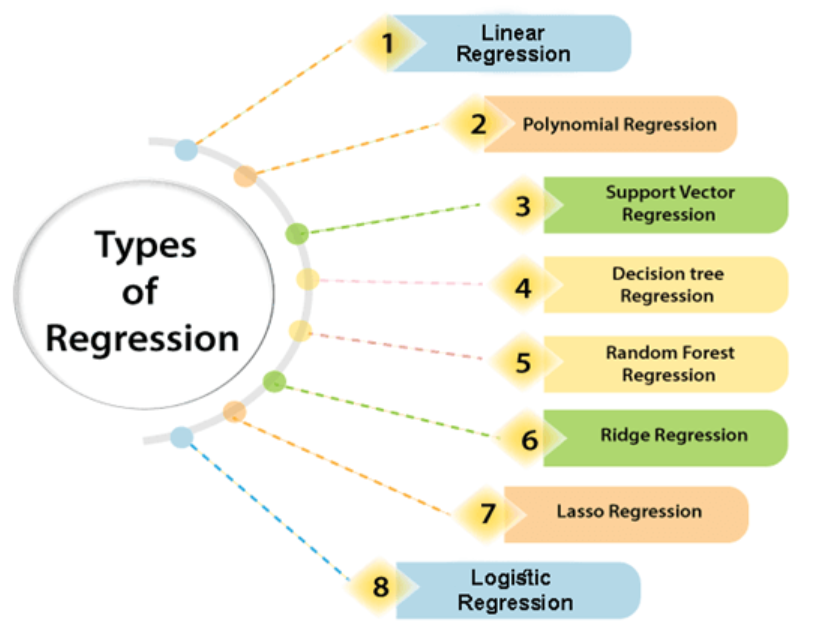
**Related Terminologies**

* **Independent variable:** Predictor (input).
* **Dependent variable:** Target variable (output).
* **Multicollinearity:** It is when multiple independent variables are **highly correlated** with each other.
* **Overfitting:** It is when our algorithm works **well** with **training dataset**; but **not** with **test dataset**.
* **Underfitting:** It is when our algorithm **doesn’t** work well with **training dataset**.

**Need for Regression Analysis**

* It helps us determine the **most** & **least important** factors in a dataset.
* And also gives insight into how they affect each other.

**Types of Regression**



**Linear Regression**

* **Simple linear regression:** One input variable.
* **Multiple linear regression:** Multiple input variable.

**Y = aX + b**

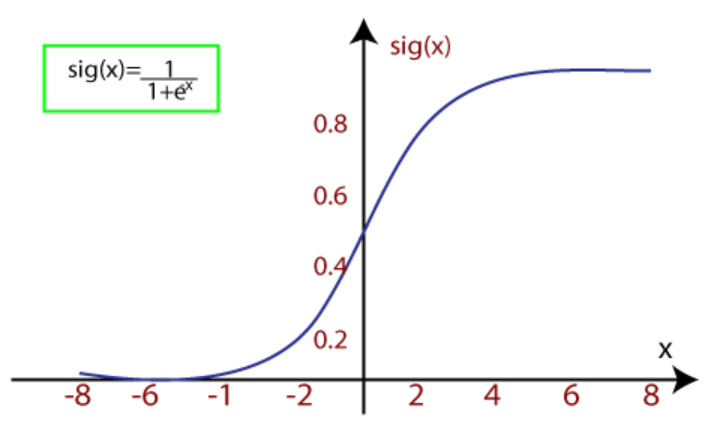
**Y = Dependant variable**

**X = Independent variable**

**a = Linear regression coefficient, b = Line’s intercept**

**Logistic Regression**

* It is used for solving ***classification*** problems.
* **Dependent variables** here are in **binary** format.
* It works on the concept of **probability**.
* Uses **sigmoid/logistic function**, also known as **complex cost** **function**.



**f(x) = 1 / (1 + e-x)**

**f(x) = Output in range [0,1]**

**x = Input variable**

**e = Base of natural logarithm**

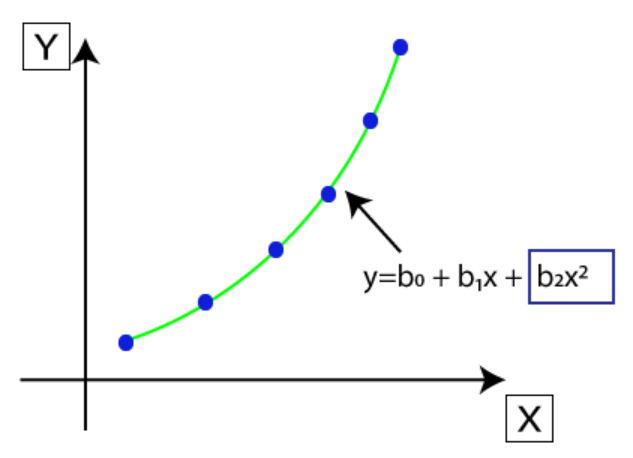
* It works on **threshold levels**, as per which the output is rounded off to **0** or **1**.
* Value **below** threshold means output will be rounded off to **0**, else **1** for **above** it.

Types of logistic regression:-

* Binary
* Multi (cats, mouses, dogs)
* Ordinal (low, medium, high)

**Polynomial Regression**

* It’s a **non-linear** regression method.
* In some situations, **linear regression** can be used but **non-linear regression** seems **more feasible** option there.

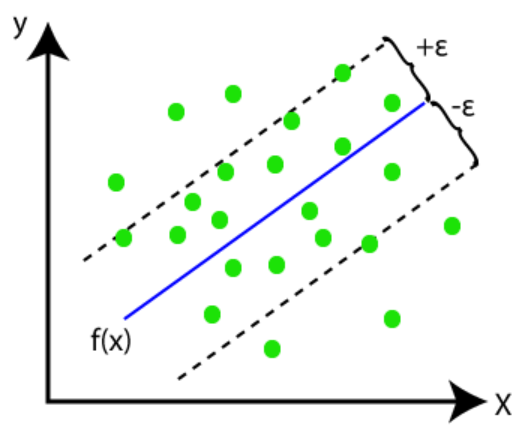


**Y = b0 + b1x + b2x2 + b3x3 + ..... + bnxn**

* It is different from ***multiple linear regression*** because coefficients here are of **different** **powers**.

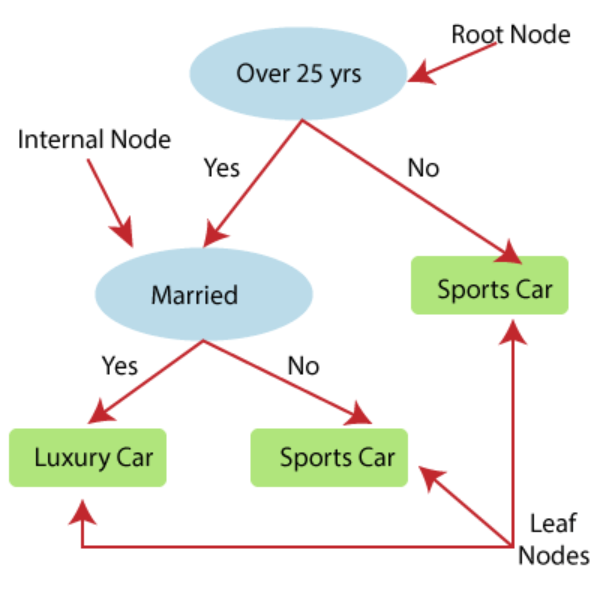
**Support Vector Regression**

* **Support vector machine (SVM)** is used for both ***regression*** as well as ***classification*** problems.
* When we use it for ***regression***, we call it **support vector regression**.
* **Kernel:** A function used for mapping **low-dimensional** data into **high-dimensional** data.
* **Hyperplane:** In SVM, it is generally a **line separating the datapoints** into different **classes**. Whereas for SVR, it is just a simple **regression line**.
* **Boundary line:** Creates **margin** for datapoints.
* **Support vectors:** Datapoints **nearest** to ***hyperplane***.
* Main goal of SVR is to get **maximum datapoints** between **boundary lines** & **hyperplane**, as shown below:



**Decision Tree Regression**

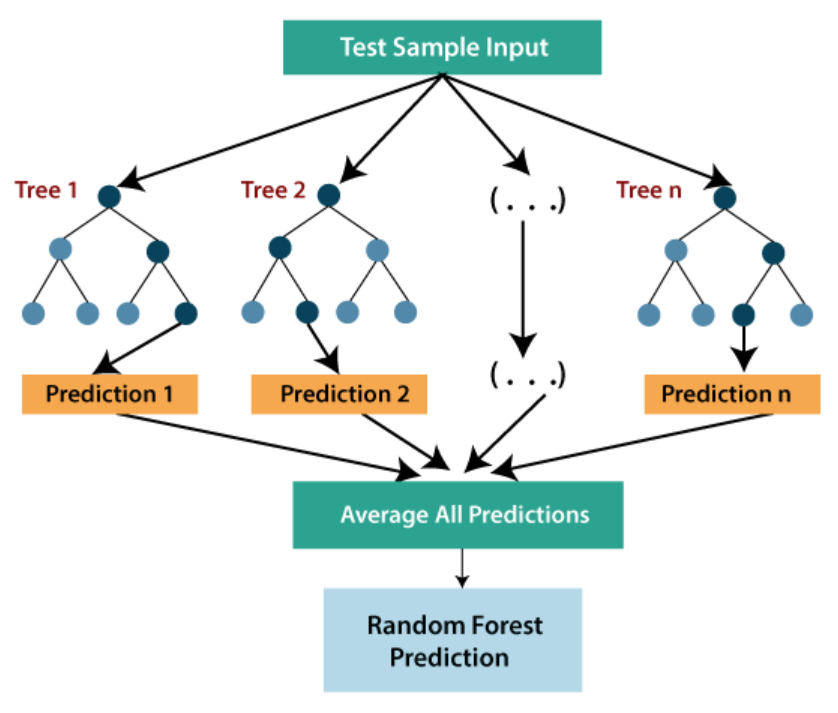
* Used for solving both ***classification*** & ***regression*** problems.
* Resembles **tree like** structure, where **branches** represent **test** & **leaf nodes** represent **final result**.



**Random Forest**

* A very powerful algorithm for performing both ***regression*** & ***classification***.
* It combines **multiple** decision trees.
* Output is based on **average** of each tree’s result.
* These decision trees when combined, are known as **base models**.
* It is also used sometimes for **preventing overfitting**.

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



**Ridge Regression**

* A type of regression where we can introduce some amount of **bias**.
* **Ridge regression penalty:** Amount of bias added.



* The term involving **lambda** is used for calculating the **penalty**.
* When there is presence of **high collinearity** among **independent variables**, ***ridge regression*** must be used instead ***linear*** or ***polynomial***.
* Helps when we have **less data** but a **lot of input** **parameters**.

**Lasso Regression**

* Used for **reducing complexity** of the model.
* Here, the ***penalty*** term contains ***absolute value*** instead ***square***.

