Regression and Analysis of Variance

Dr. Jyotismita Chaki

Regression

- 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.
- Regression is a supervised learning technique 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**.

Regression

- 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.

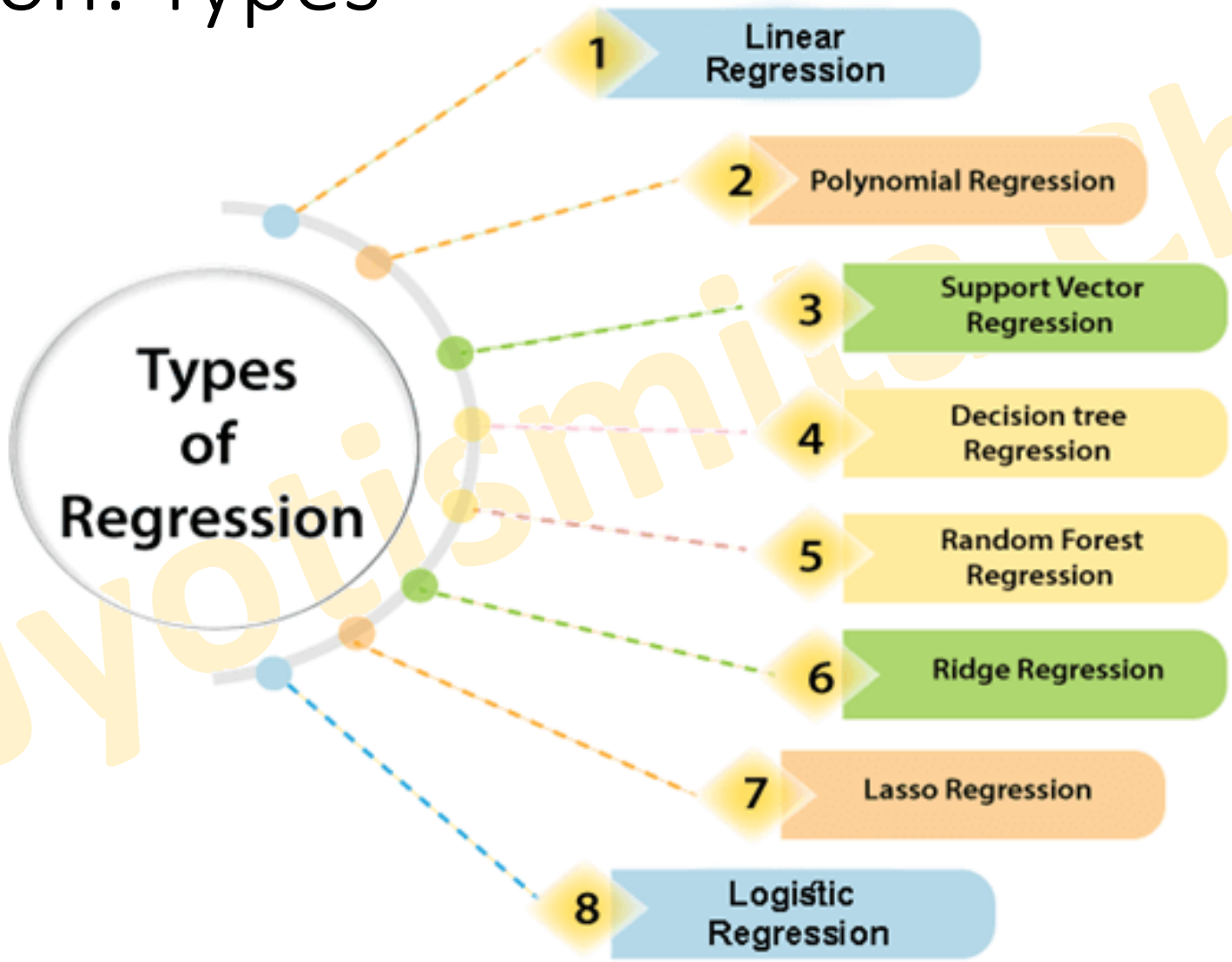
Regression: Terminologies

- **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**.

Regression: Need

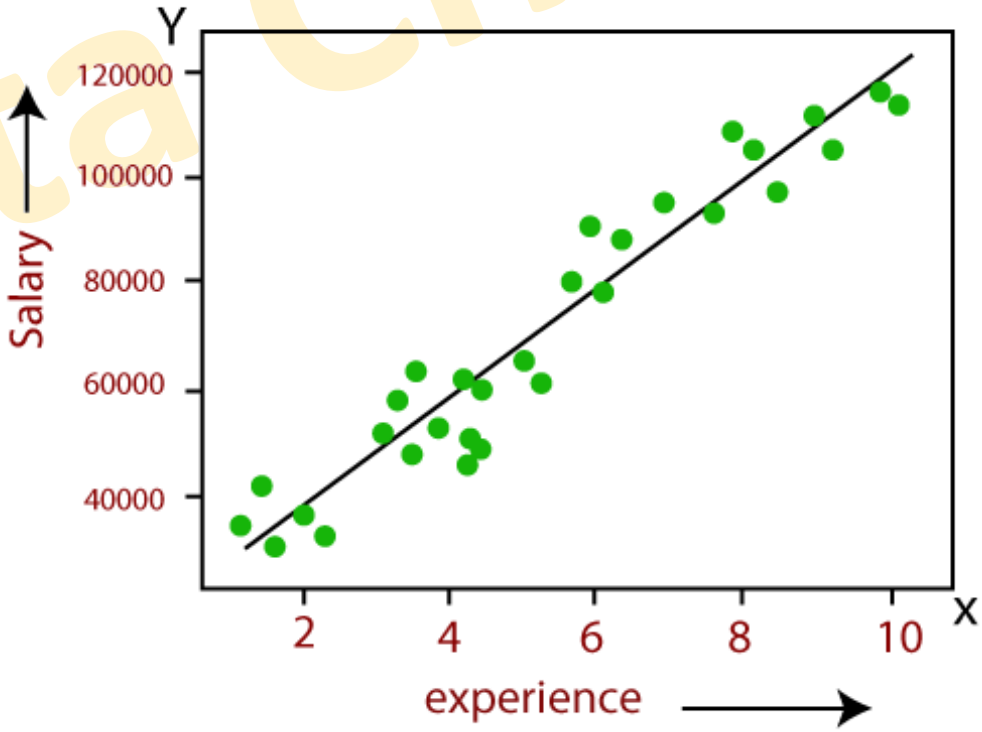
- Regression analysis helps in the prediction of a continuous variable.
- There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately.
- 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**.

Regression: Types



Regression: Types: 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**.



Regression: Types: Linear Regression

- The 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.**

Regression: Types: Logistic Regression

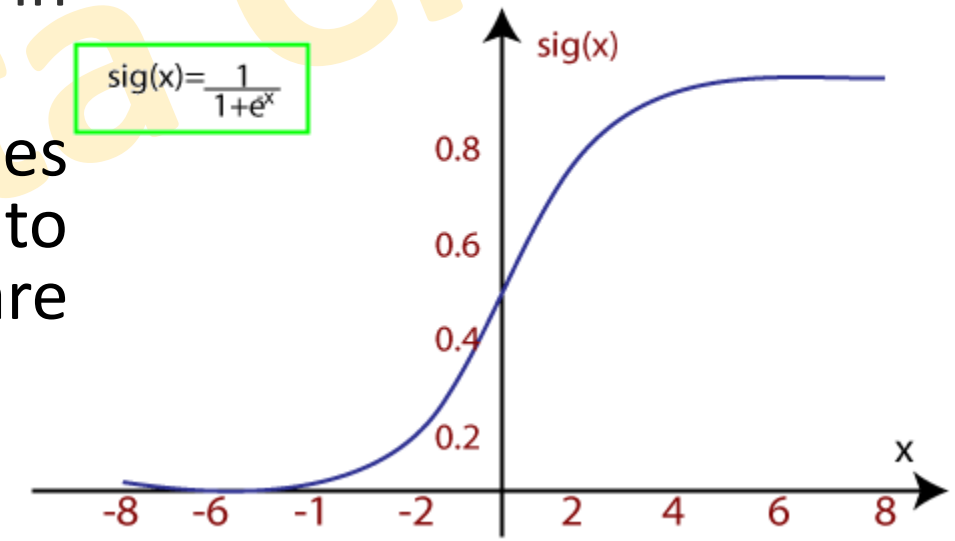
- Logistic regression is another supervised learning algorithm which is used to solve the classification problems. In **classification problems**, we have dependent variables in a binary or discrete format such as 0 or 1.
- Logistic regression algorithm works with the categorical variable such as 0 or 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:

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

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

Regression: Types: Logistic Regression

- When we provide the input values (data) to the function, it gives the S-curve as shown in figure.
- 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)**

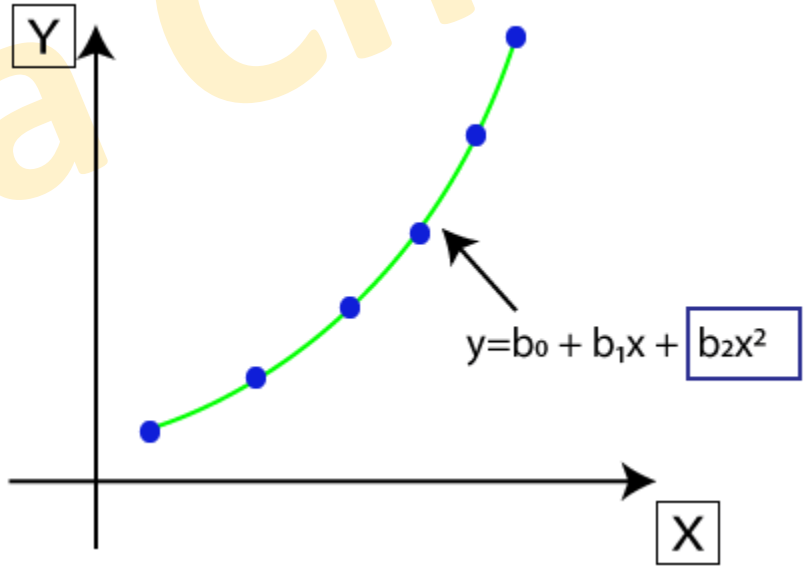


Regression: Types: Polynomial Regression

- 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.

Regression: Types: Polynomial Regression

- 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**.
- The model is still linear as the coefficients are still linear with quadratic

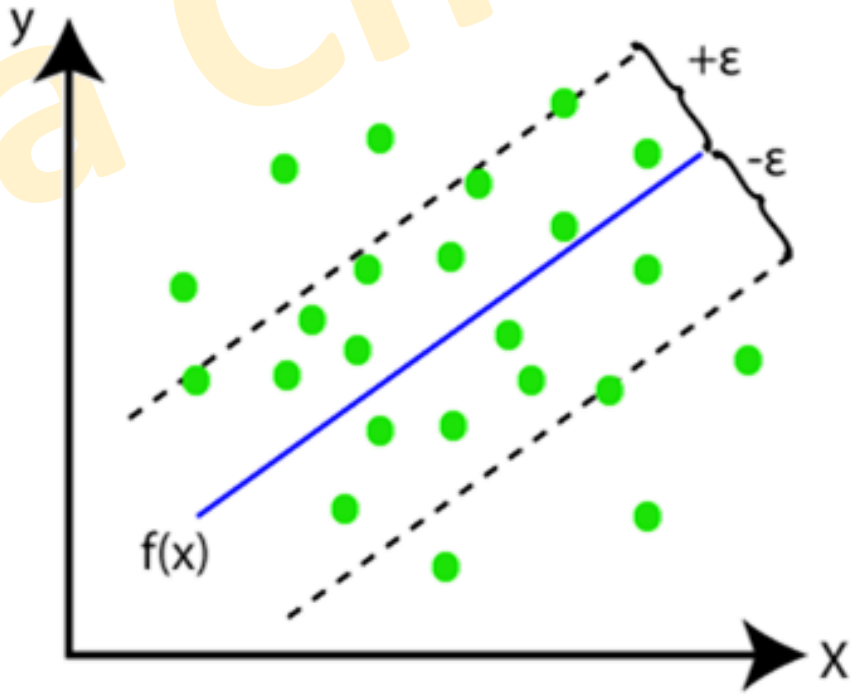


Regression: Types: Support Vector Regression

- 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.

Regression: Types: Support Vector Regression

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

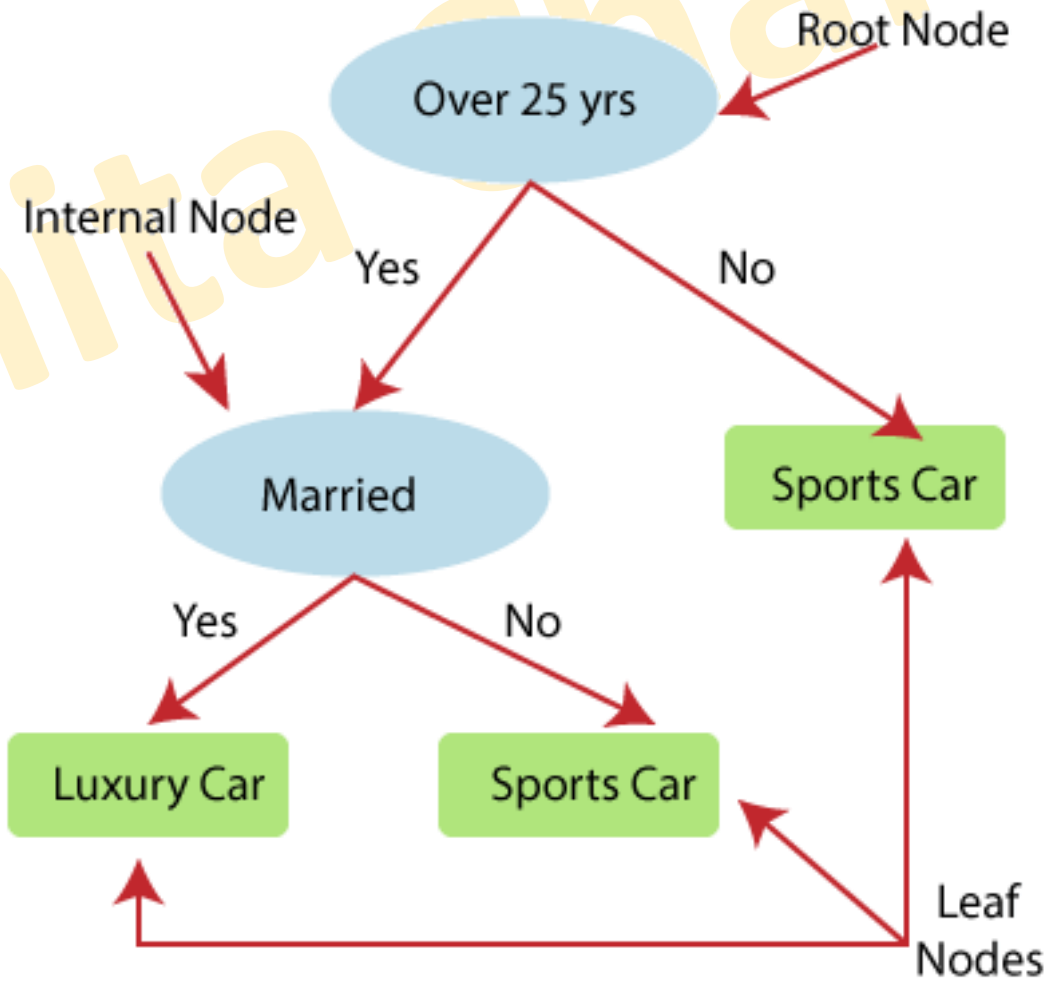


Regression: Types: Decision Tree Regression

- 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.

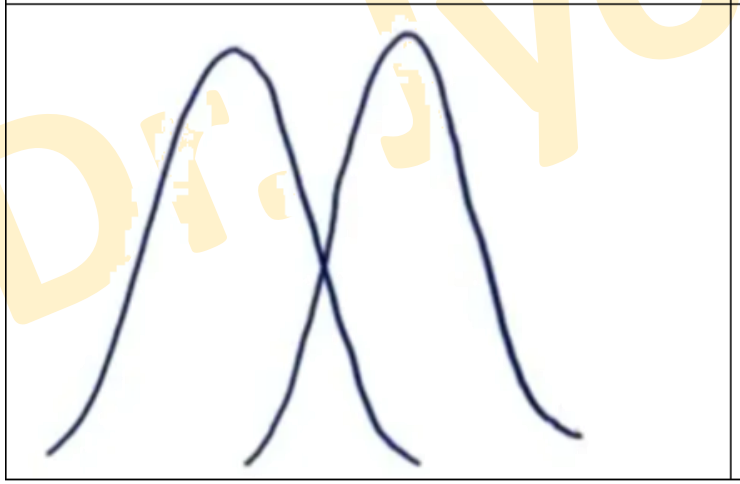
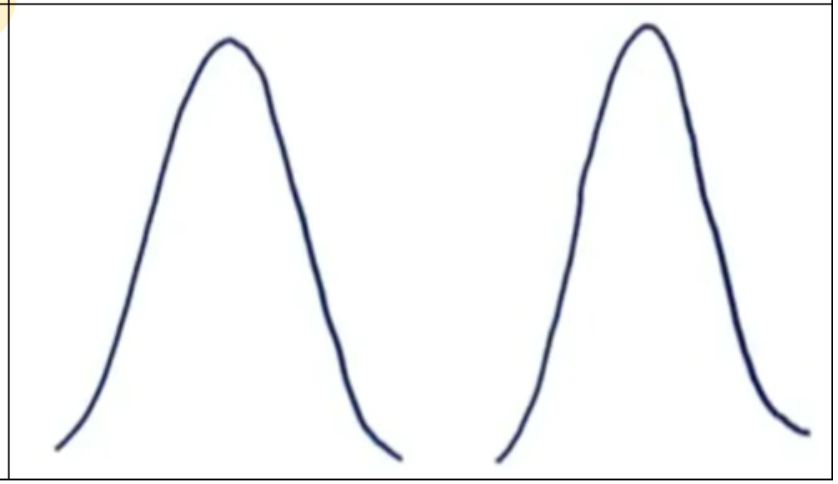
Regression: Types: Decision Tree Regression

- Consider the following image: The 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.



Analysis of Variance (ANOVA)

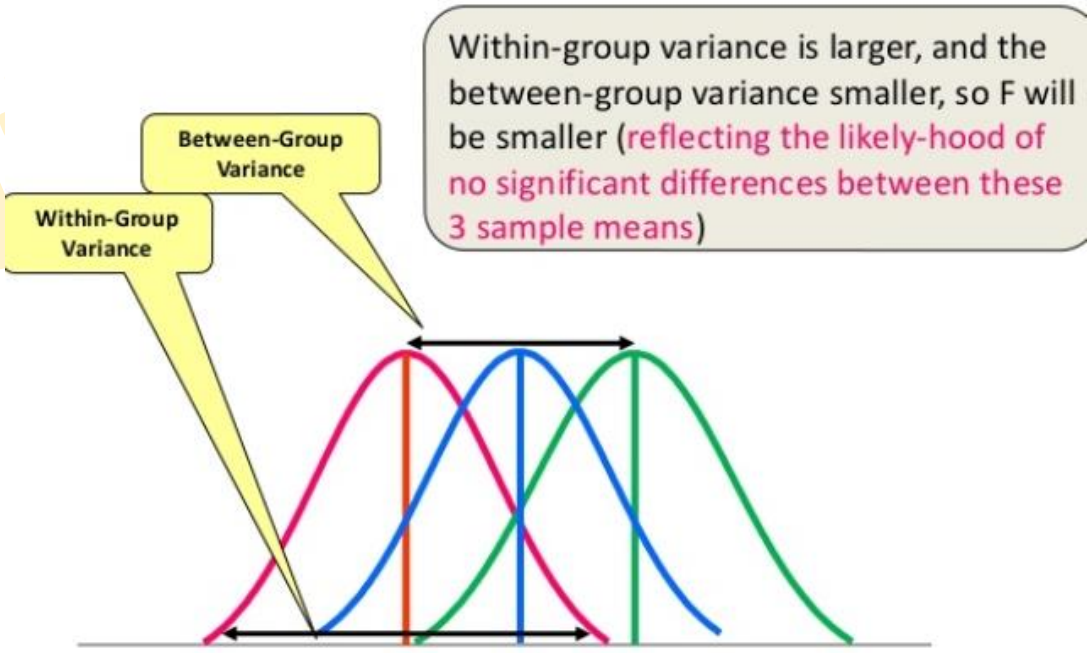
- Statistical method, used to check the means of two or more groups that are significantly different from each other.
- Consider two distributions and their behavior in below fig.

a. Two distributions Overlap each Other	b .Two distributions differ from each other
• Individual means don't differ by great margin. • Difference between individual means and grand mean would be less.	• Individual means and grand mean differ by larger distance. • Difference between individual means and grand mean would be large.
	

From the fig, we can say If the distributions overlap or close, the grand mean will be similar to individual means whereas if distributions are far, the grand mean and individual means differ by larger distance.

Analysis of Variance (ANOVA)

- *In ANOVA, we will compare Between-group variability to Within-group variability.*
- ANOVA uses F-test to check if there is any significant difference between the groups. [**F = Between group variability / Within group variability**]
- If there is no significant difference between the groups that all variances are equal, the result of ANOVA's F-ratio will be close to 1.



ANOVA: One way and Two way

- One way: Used to determine how one factor affects a response variable. Suppose a professor wants to know if three different studying techniques lead to different exam scores.

- Two way: Used to determine how two factors affect a response variable, and to determine whether or not there is an interaction between the two factors on the response variable.

Suppose a botanist wants to know whether or not plant growth is influenced by sunlight exposure and watering frequency.