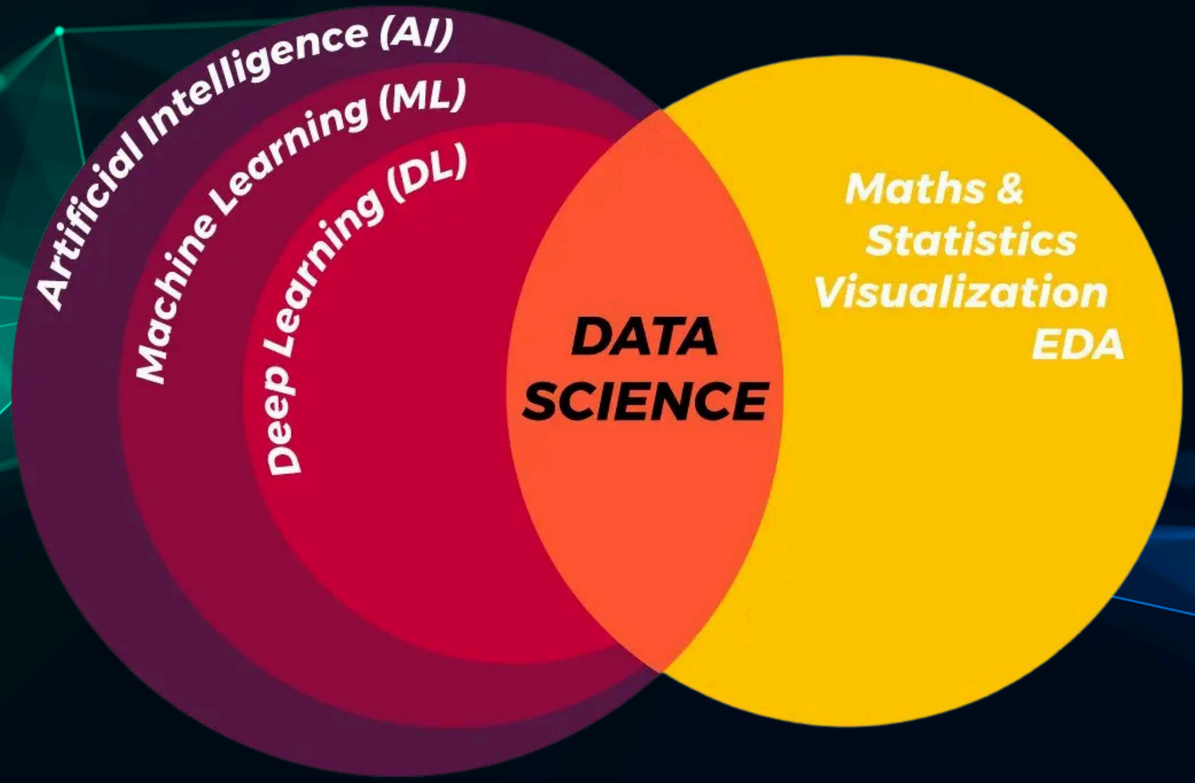
WINTER PROJECT

DATA ANALYTICS

LECTURE 2 :

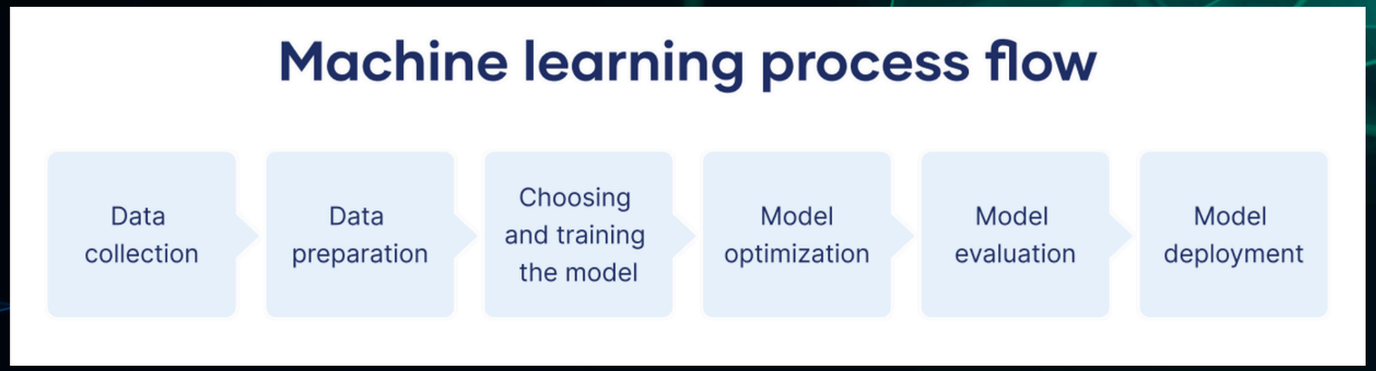
Intro to ML : Regression and Classification


Artificial Intelligence (AI)
Machine Learning (ML)
Deep Learning (DL)
DATA SCIENCE
Maths &
Statistics
Visualization
EDA

INTRODUCTION TO MACHINE LEARNING

What is this cool term?

- **This is a branch of AI that enables computers to learn from data without need for explicit logic-feeding or programming.**
- **Just like humans learn from experience, machines can be trained using DATA.**
- **This is how the process works:**



Used for :

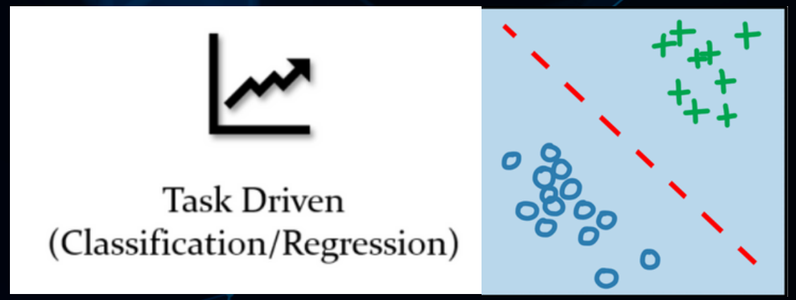
1. **Detect patterns**/structures/themes/trends etc. in the data.
2. Make **predictions** about future data and make decisions.

- **Having already received brief idea on the topics of data collection and data preparation (basically what you did in EDA), we will cover the models you can select for your data in today's session.**

TYPES OF ML:

Supervised learning

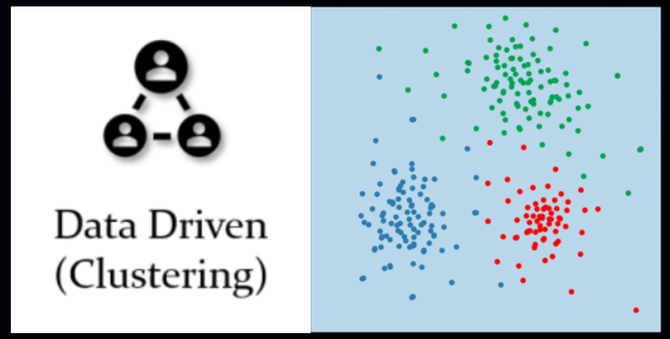
Here, the model is trained on data where the **input features** are mapped with the corresponding **output targets.**



DATA WITH LABELS

Unsupervised learning

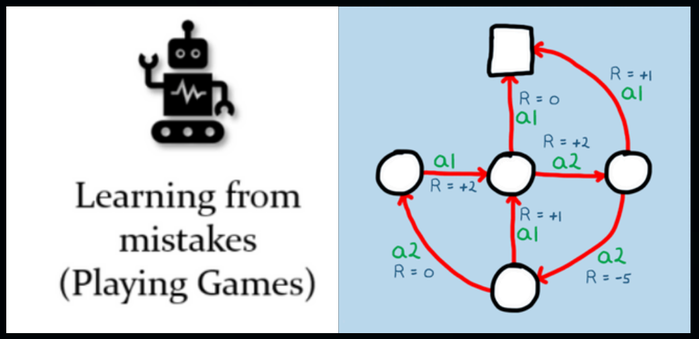
Here the model recognises **pattern in data** without being provided explicit output labels.



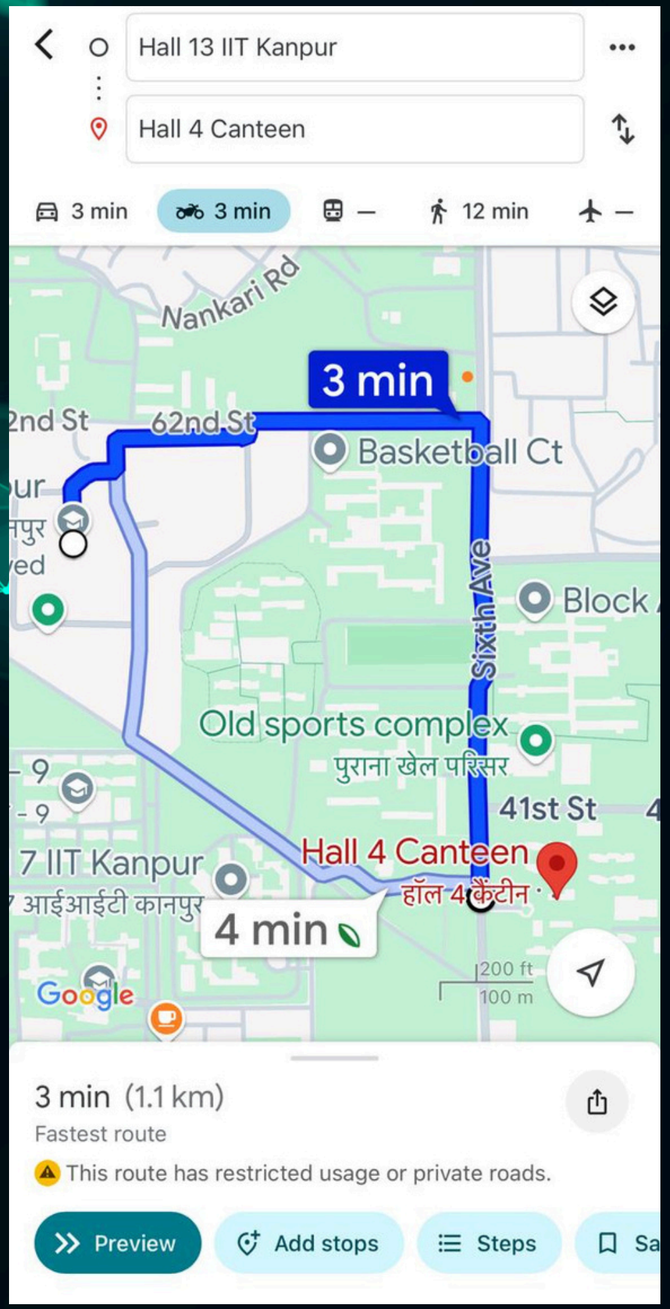
DATA WITHOUT LABELS

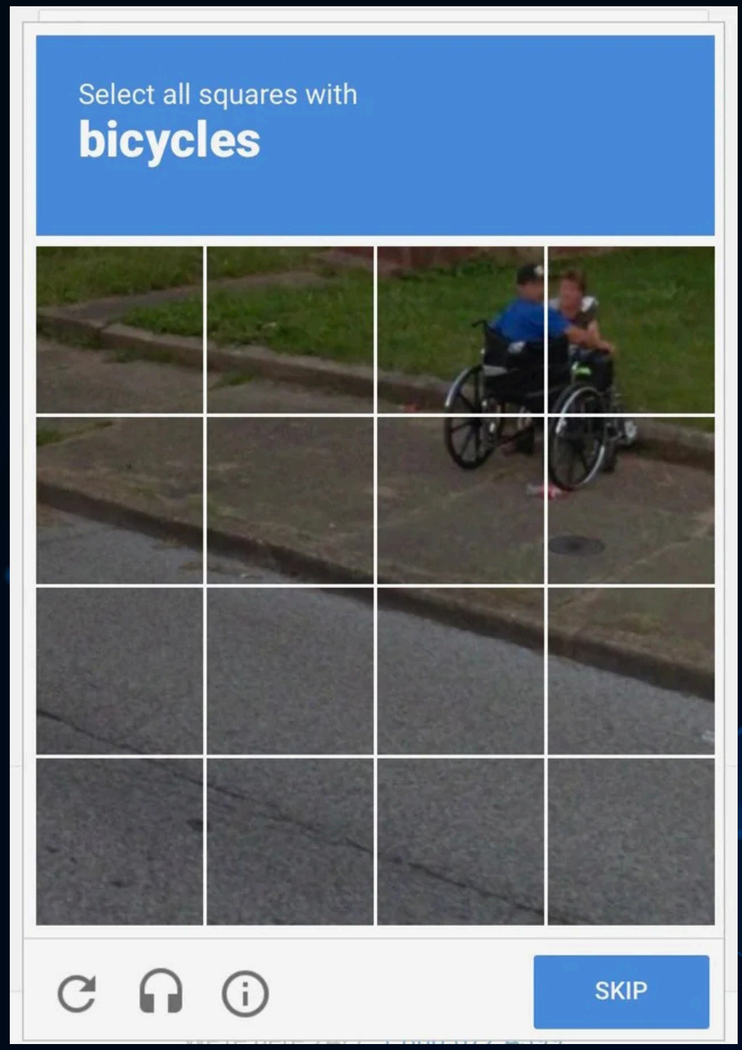
Reinforcement learning

Here the goal is to train the model by **rewarding or punishing** in accordance with its actions. The objective for it is to maximise the rewards.



ACTIONS AND IMPROVEMENTS


Hall 13 IIT Kanpur
Hall 4 Canteen
3 min
3 min
12 min
Nankari Rd
3 min
62nd St
Basketball Ct
Sixth Ave
Block
Old sports complex
पुराना खेल परिसर
41st St
Hall 4 Canteen
हॉल 4 कैंटीन
7 IIT Kanpur
आईआईटी कानपुर
4 min
200 ft
100 m
Google
3 min (1.1 km)
Fastest route
This route has restricted usage or private roads.
Preview
Add stops
Steps


Select all squares with
bicycles
SKIP

TYPES OF SUPERVISED LEARNING

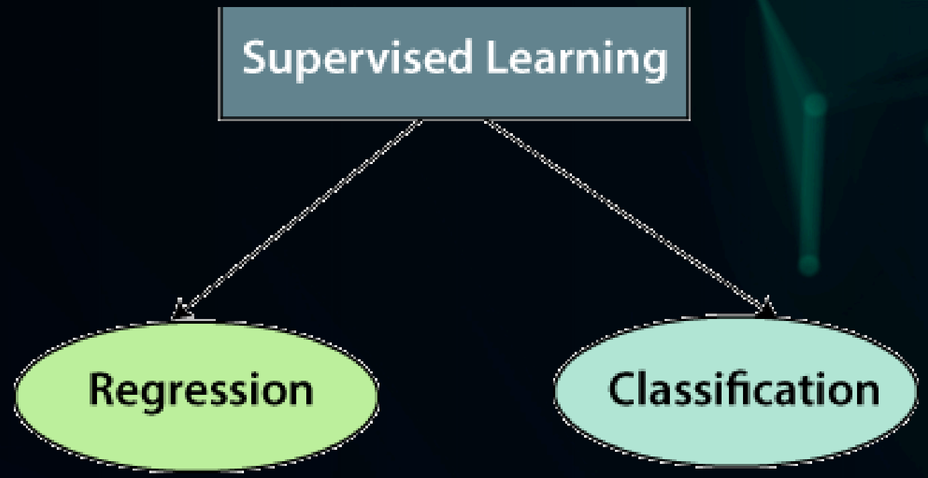
Regression

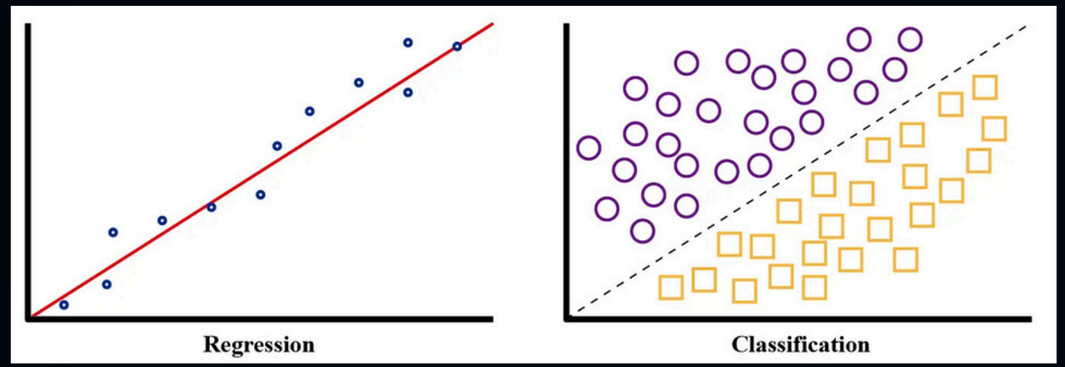
Here our objective is to establish a correspondance between the input and output.

Remember the **output should be a continuous** variable here.

Nature of problem: “**How much**”,”**How many.**”

Example: Finding stock prices or the price of houses based on given square footage, location etc.





Classification

Like the name suggests, this process is about classifying a given input to a **discrete** output label or category.

Nature of problem: “**Which class?**” or “**Which category**” are answered by this method.

Example: To classify types of buyers on shopping sites or to recognise a mail as spam or not.

ML IN FINANCE

Objective : To help us take Decisions

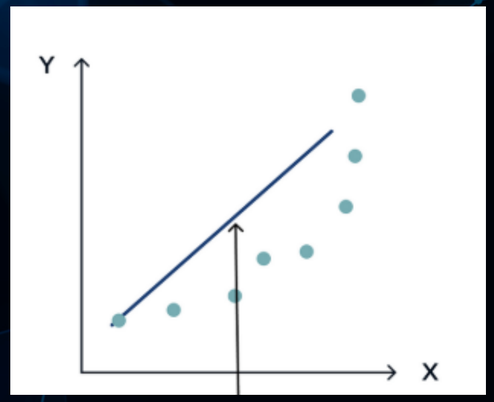
- **Fraud Detection**
  - Goes through millions of data points in real time every second and detect irregular transactions and unusual data patterns by itself for businesses, and financial institutions.
- **Credit Risk Assessment**
  - **B**anks need to make decisions about who to lend money to, check their transactions, their account behaviour, the borrower information to fix an interest rates or deny loans
- **Portfolio Optimization**
  - Companies or individuals make decisions about where to allocate capital by analyzing patterns about which invesments are likely to be more profitable to improve overall return performance.
- **AlgoTrading**
  - Supervised Learning: Predicting prices, trends, or signals using labeled data.
  - Unsupervised Learning: Clustering stocks, candlestick patterns and price-action.
  - Reinforcement Learning: Designing trading bots that learn optimal strategies.

REGRESSION : THE FITTING TECHNIQUE

- Statistical and machine learning technique used to model the relationship between a dependent variable (target) and one or more independent variables (Features).
- The fit between Features and Target is generalised to make predictions.
- Helps in decision making by identifying critical factors influencing an outcome.

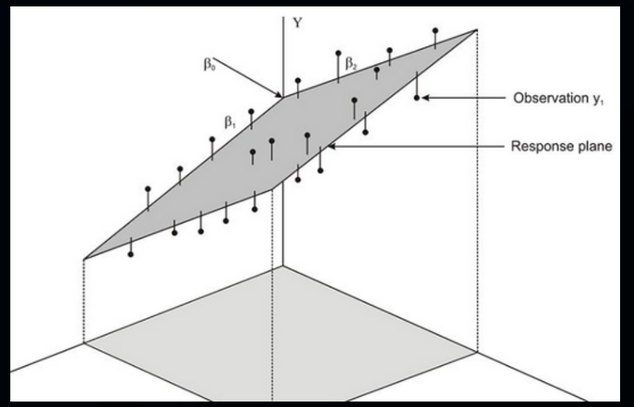
Types of Regression :

- **Linear Regression :**
  - Simple Linear Regression: Involves one independent variable.
  - Multiple Linear Regression: Involves two or more independent variables.



SLR

y is target
x is feature
The fit is a straight line



MLR

y is target
x1, x2 are features
The fit is a plane for 3D

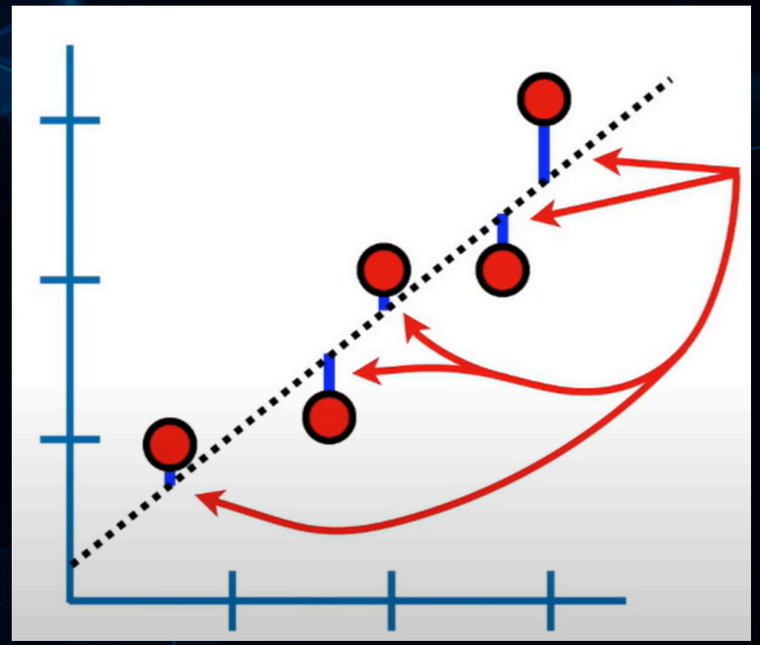
How do models perfects themselves?

-> By minimising ERROR.

-> Key metrics for evaluating error :

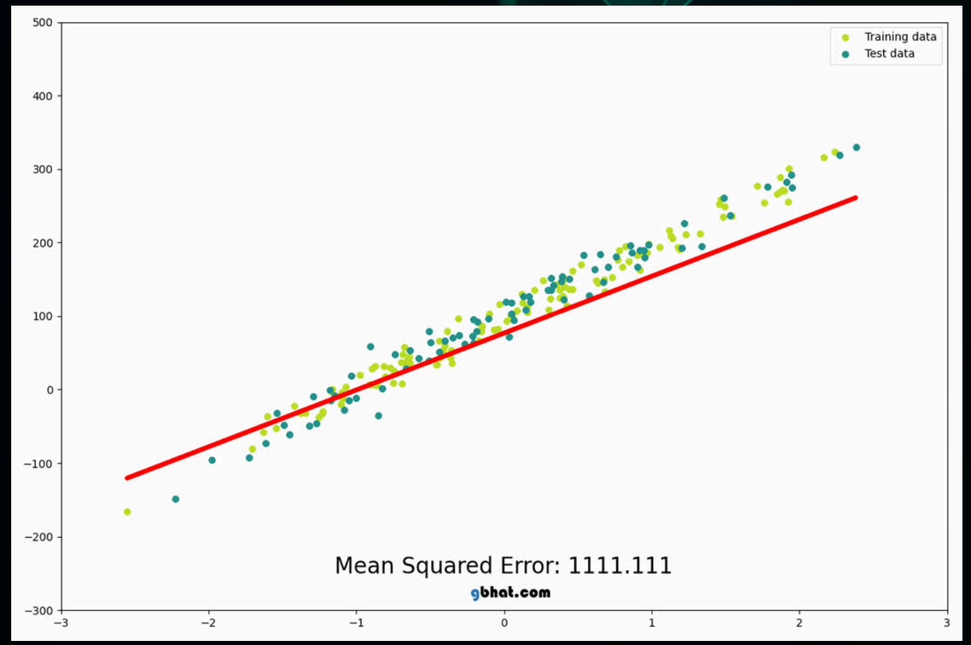
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Accuracy etc.

Ex : Linear Regression



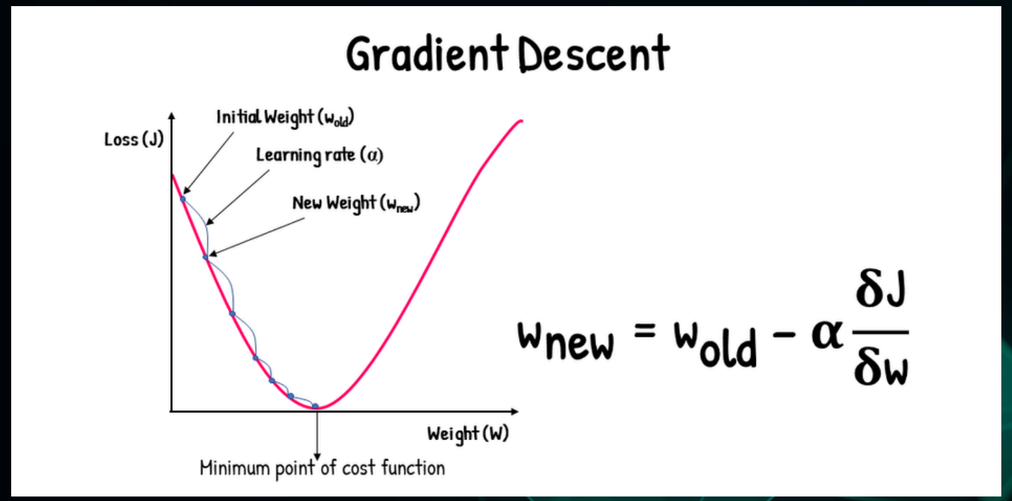
$$\hat{y}_i = b_0 + b_1 x_i$$

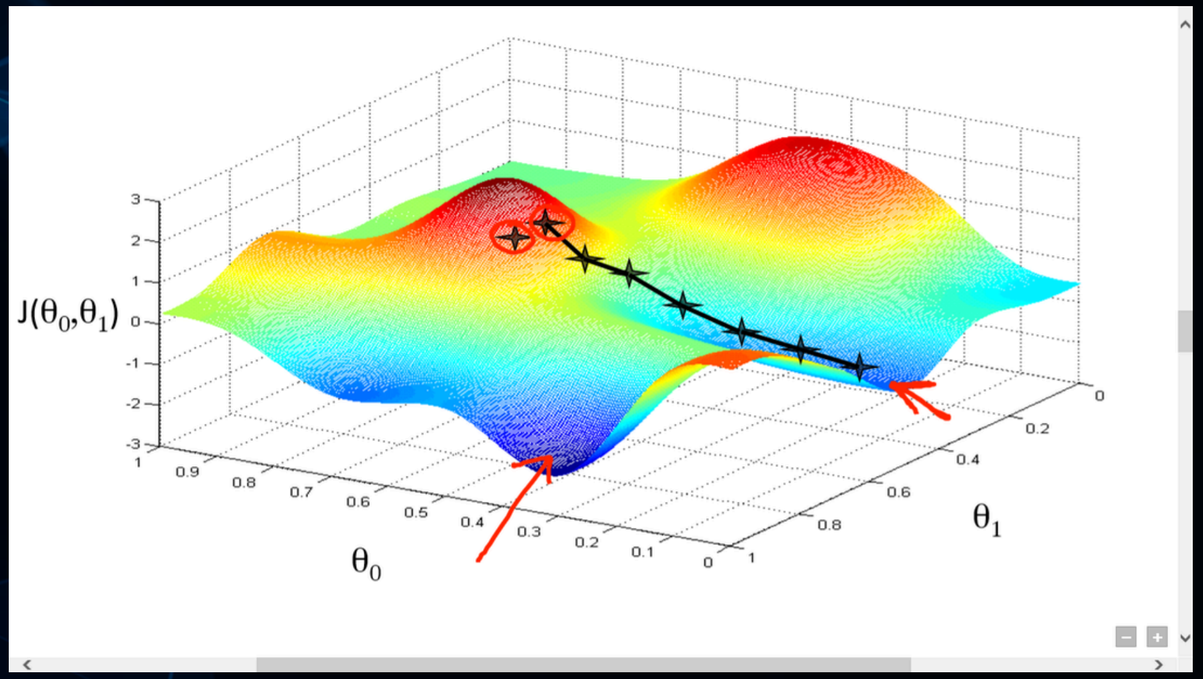
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

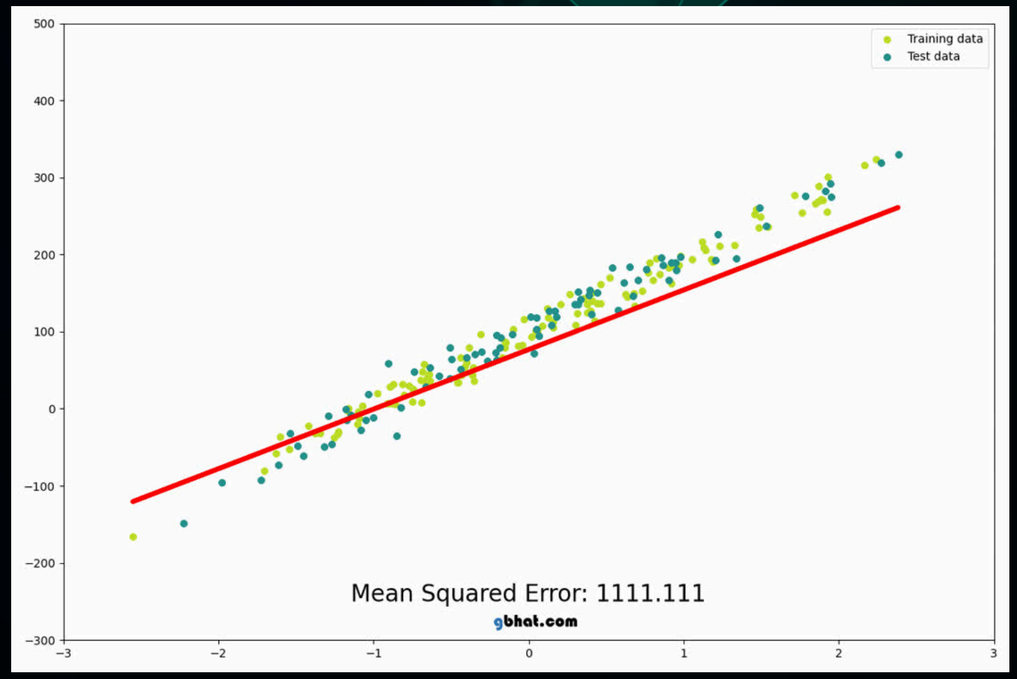


How do models perfects themselves?

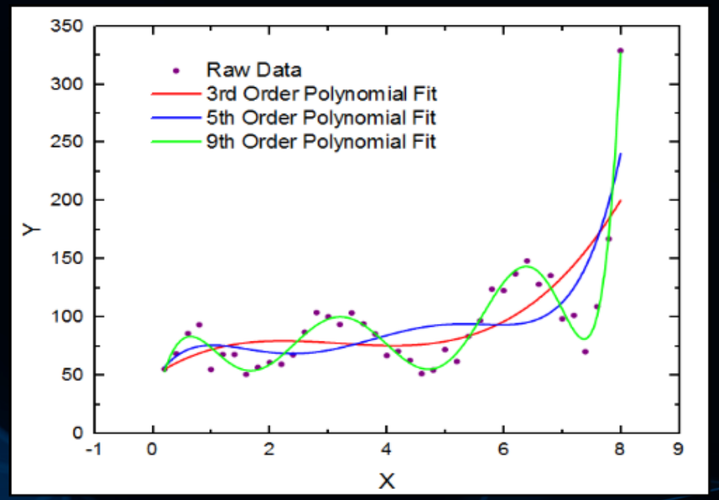
- Process ->
  - Assumption : This is Best Fit line equation
  - Squared Error is to be minimised.
  - (Gradient Descent)
  - Differentiated and equated to zero (Gradient =0).
  - Obtain b0, b1 .
  - Generalize the equation.
  - Predict ‘y’ for ‘x’.

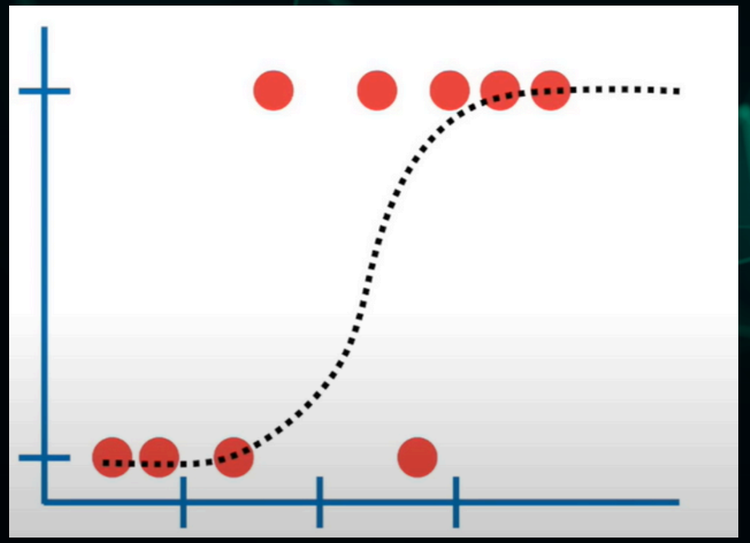






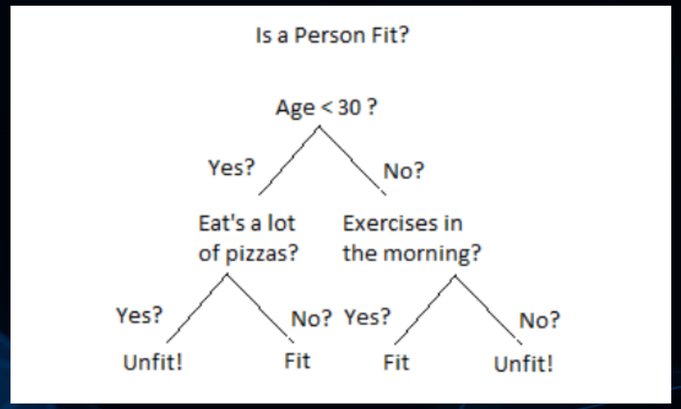
- **Polynomial Regression :**
  - The fit is in non-linear relationship form like a polynomial equation between features and targets
  - Possible fits can be of various orders, and the best ordered fit is the one most accurate with

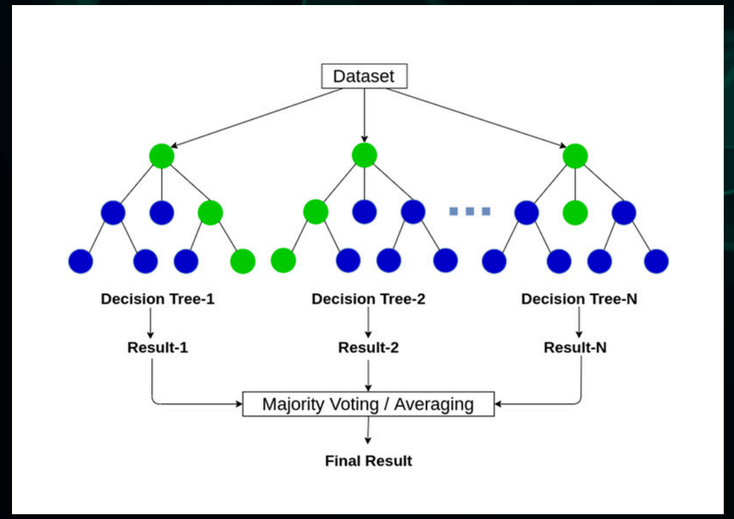




- **Logistic Regression :**
  - Used for classification problems, where the target variable is categorical (Ex : yes/no, spam/not spam).
  - Logistic regression is a classification algorithm even though it has regression in its name.
  - The fit is like S- Curve.
  - Very helpful in cases to find out which variables or features have a impact on outcome

- **Decision Trees :**
  - Splits the data into smaller subsets based on certain conditions and predict a continuous value
  - The tree splits data at each node based on a **condition** that minimizes a metric like Mean Squared Error (MSE).
  - Overfitting risk: A single decision tree can overfit, it performs well on training data but poorly on new data.





- **Random Forest Regression :**
  - Similar to above, but this solves the problem mentioned.
  - Is an ensemble method that combines multiple decision trees to improve predictions and reduce overfitting .
  - It takes the average of predictions from many decision trees to provide a final output.
  - Random Forest Technique is used in both Regression (final output is Average) and Classification (is result with highest Frequency)

THANK YOU!