What is Artificial Intelligence (AI)

Most of us think AI is Robotics

- AI is not robotics
- AI is process
- AI is a study of how human brain think, learn, decide and work, when it tries to solve problem
- Robotics uses AI for bringing human intelligence to robots

Areas of AI

Two broad categories.

- ### Narrow Artificial Intelligence

  Its focus is on performing specific tasks.

  - Chat GPT, Amazon Alexa, YouTube Recommendation, Self Driving Cars.

- ### Artificial General Intelligence

  Human level intelligence

What is Machine Learning

Machine Learning (ML) is a discipline of artificial intelligence that allows machines to automatically learn from data and past experiences while identifying patterns to make predictions with minimal human intervention.

Machine Learning process

1. Collection of data from various source
2. Data cleaning and feature engineering
    a. Handle missing values
    b. Handle outliers
    c. Encoding
    d. Normalization and scaling
3. Model building using machine learning algorithm
    a. Regression
    b. Classification
    c. Clustering
4. Model testing using performance matrices
5. Model development

Types of Machine Learning

- Supervised Learning
    - Classification
        - Support vector machine
        - Decision Tree
        - Random forest
    - Regression
        - Linear regression
        - Natural network regression
        - Support vector regression

- Unsupervised Learning
    - Clustering
        - K – means clustering
        - Mean shit clustering
- Reinforcement Learning
    - Decision Making
        - Q – Learning
        - R – Learning

Supervised Learning

Algorithm learns from labeled training data to make predictions

Ex:

Classification – Predicting categories (e.g., spam detection)

Regression – Predicting continuous values (e.g., house prices)

Overfitting and Underfitting

Overfitting – refers to a model that models the training data too well.

1+1 =2

1+1=2

Underfitting – refers to a model that can neither model the training data nor generalize to new data.

1+1=2

1+1=2

1+1=3

Unsupervised Learning

The model is trained on unlabeled data, meaning the data does not have predefined labels or outcomes.

Example Applications

Customer segmentation

Market basket analysis

Anomaly detection in fraud detection

Image and video compression

Reinforcement Learning

A technique that trains software to make decisions to archive the most optimal results.

Applications

Game playing
Robotics

Data in Machine Learning

There are two main types of data which are

Quantitative

Qualitative

Data cleaning

Process of fixing or removing
- Incorrect
- Corrupted

How to clean data

1. Remove duplicate or irrelevant observations
2. Fix structural errors
3. Filter unwanted outliers
4. Handle missing data
5. Validate and quality assurance (QA)

Application of Machine Learning

Healthcare - Disease diagnosis, personalized treatment
Finance - Fraud detection, Algorithmic trading
Retail - Recommendation systems, Inventory management
Transportation - Autonomous vehicles, route optimization
Entertainment - Content recommendation, Audience analysis

Challenges in Machine Learning

Data quality - Garbage in, garbage out
Overfitting - The model performs well on training data but poorly on new data
Interpretability - Understanding how and why models make decisions
Scalability - Handling large volumes of data
Bias and Fairness - Ensuring ethical use and avoiding discrimination

Regression

Used to predict continuous values

Types of ML regression algorithms

- Linear Regression
- Regression trees
- Non-linear regression
- Bayesian linear regression
- Polynomial regression

Linear Regression

Simple linear regression : X=1 Y=1

$$Y=\beta_0+\beta_1 x$$

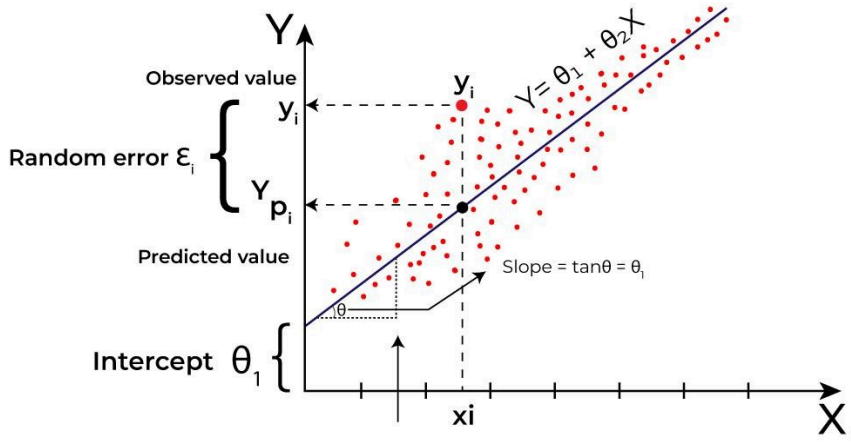
Multiple linear regression : plenty of X Y=1

Random Error(Residuals)

ε = Predicted value - Actual Value

$$\varepsilon = \hat{y}_{(i)} - y_{(i)}$$

Best Fit Line



Simple Linear Regression(Least Square Method)

This is a method to find out the slope and the intercept of the best fir line of a given data set

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

Evaluation Metrics

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE)

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

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R-Squared (R^2) Score

R square value should be 0 to 1,if R squared value is closer to one model is good

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total}}$$

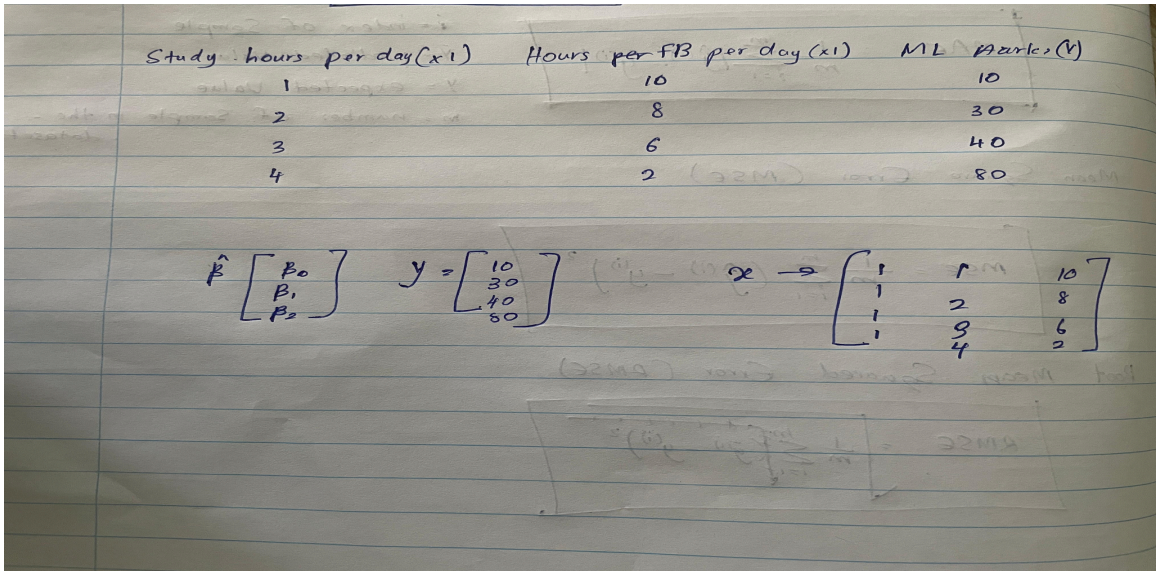
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Multiple linear regression

Multiple independent variables and a single dependent variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$



CLASSIFICATION TYPES

- Binary classification
- Multi-label classification
- Multi-class classification
- Imbalance classification

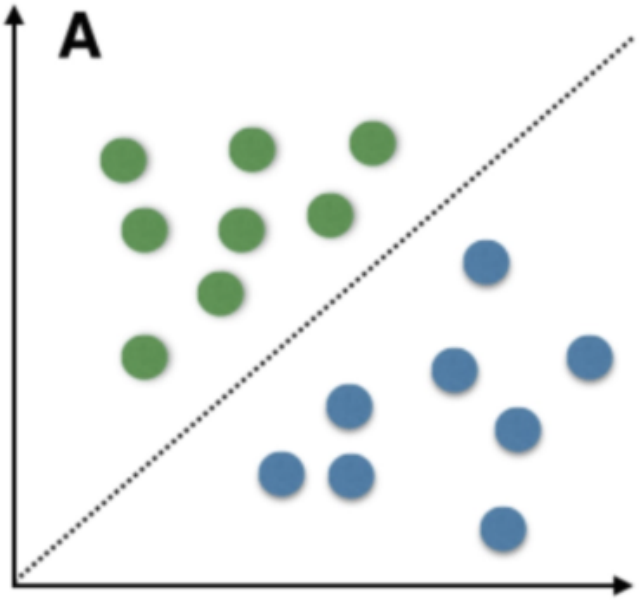
Binary classification

Two outputs are their …it can be

- Buy or not
- Spam or not
- Pass or fail

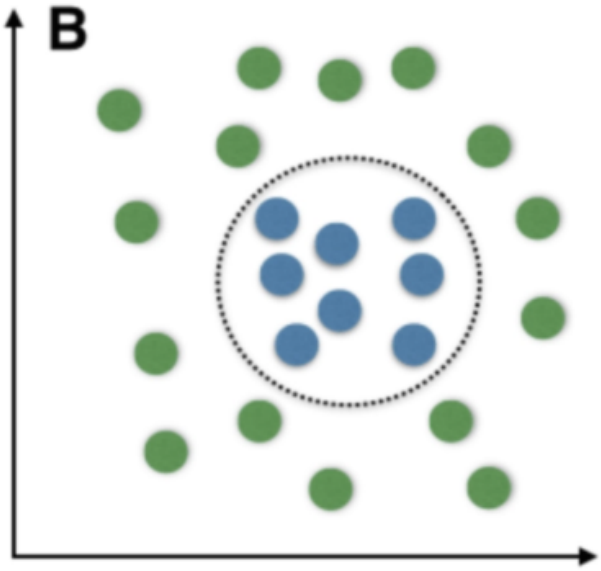
Linear classifiers

- Logistic Regression
- Support Vector Machines Having kernel = ‘linear’



Non Linear classifiers

- K-Nearest Neighbours
- Kernel SVM
- Decision Tree Classification
- Ensemble Learning Classifiers
- Random Forests
- AdaBoost
- Bagging Classifier



Logistic Regression

Here we have two outputs

Sigmoid Function

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

P >= Some Cutoff → Assign to class 01
P <= Some cutoff → Assign to class 02
Normally general cutoff values is =0.5

Model evaluation

Miss classification Error(MCE)
Accuracy=1-MCE
F1 Score

Classification (03/03/2025)

- True Positive(TP)
    - The predicted class is positive class. Predicted value and actual value are the same

- False Positive(FP)
    - The predicted class is positive class .but the predicted value and the actual value are not the same

- True Negative(TN)
    - The predicted class is negative class. Predicted value and actual value are the same

- False Negative(FN)
    - The predicted class is negative class. but the predicted value and the actual value are not the same

		Actual Class	
		Positive (**P**)	Negative (**N**)
Predicted Class	Positive (**P**)	True Positive (**TP**)	False Positive (**FP**)
	Negative (**N**)	False Negative (**FN**)	True Negative (**TN**)

Performance Metrics For Classification

- Accuracy

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Presition

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1 Score

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

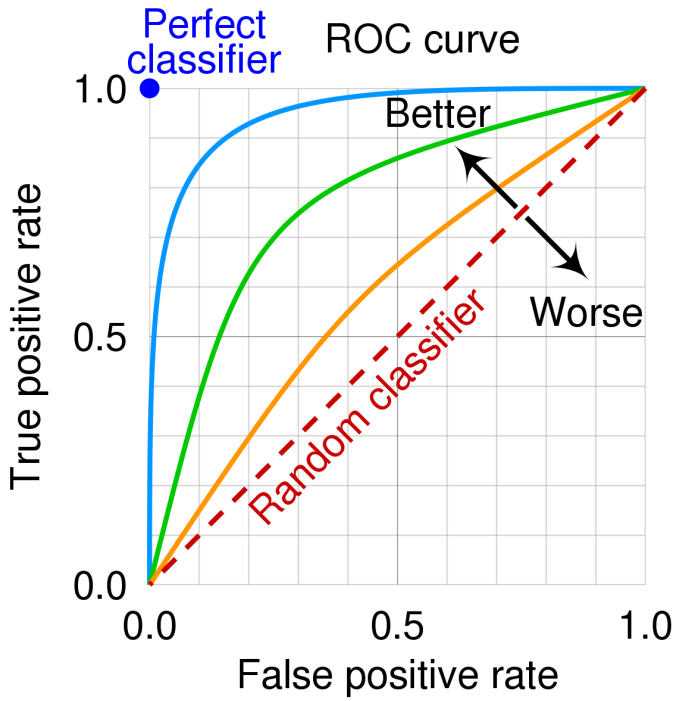
- True positive Rate

$$\text{TPR} = \frac{TP}{TP + FN}$$

- False Positive Rate

$$\text{FPR} = \frac{FP}{FP + TN}$$

AUC-ROC(Receiver Operating Characteristics)



AUC Should between 0.5 and 1

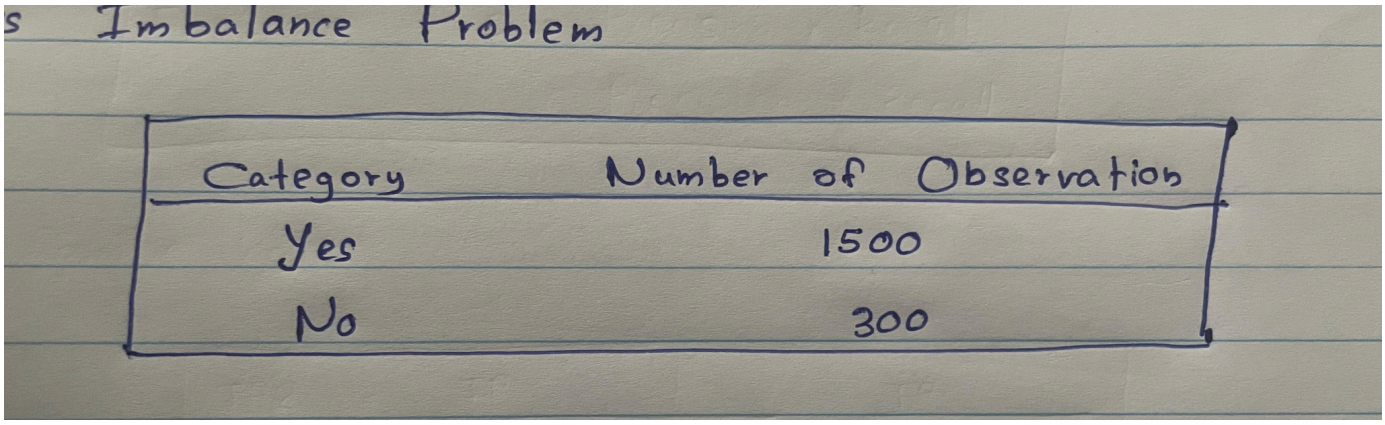
AUC-Area under the curve

AUC Range	Classification Level
0.90 - 1.00	Excellent
0.80 - 0.90	Good
0.70 - 0.80	Fair
0.60 - 0.70	Poor
0.50 - 0.60	Failure

Best range are excellent,good, fair

Class imbalance problem in classification

The observation in both classes are not balanced.this problem is called the class imbalance problem



Options

SMOTE
ADASYN
Hybridization: SMOTE+Tomek Links
Hybridization: SMOTE+ENN

K-Nearest Neighbors (KNN)

First define **k** value,usually **k** value is more than 5.then calculate the distance from the point where you need to predict to the other data points.then obtain the closest distance values according to the given k value (id k is 5 get first 5 closes distances)

If it is a classification classify according to the number of neighbours .if it is a regression obtain the values by considering the mean value of the closest neighbours

How to measure the distance in KNN

1. Euclidean distance
2. Manhattan distance
3. Minkowski distance
4. Hamming distance

Euclidean distance