Introduction to Machine Learning

Machine learning is a branch of artificial intelligence (AI) that empowers computers to learn from data without explicit programming. It allows systems to automatically improve their performance over time by recognizing patterns, making predictions, and adapting to new data.

Athira S Nair

Machine Learning Terminologies

Understanding the key terminologies is crucial for comprehending machine learning concepts and effectively applying them in practice.

Term	Definition
Features	Each feature, or column of a data set represents a measurable piece of data that can be used for analysis
Dataset	A set of data samples represented using features to solve problems
Model	A mathematical representation of the patterns learned from data.
Training	The process of feeding data to a model to learn its parameters.
Prediction	The output generated by a model based on input data.

What is Machine Learning?

Machine learning utilizes algorithms to analyze data and make predictions or decisions.
It involves training a model on a dataset to learn patterns and relationships, enabling it to perform tasks such as image recognition, natural language processing, and fraud detection.

1 **Learning from Data**

Machine learning systems learn from data, identifying patterns and relationships to make predictions or decisions.

2 **Building Predictive Models**

The goal of machine learning is to build models that can accurately predict future outcomes or classify data based on past experiences.

3 **Automation and Efficiency**

Machine learning automates complex tasks, freeing up human time and resources for more strategic activities.

4 **Continuous Improvement**

Machine learning models can adapt and improve over time as they encounter new data and feedback.

Applications of Machine Learning

Machine learning is transforming various industries by automating tasks, improving efficiency, and generating valuable insights. These applications demonstrate the wide-ranging impact of machine learning in modern society.

Healthcare

Medical diagnosis, drug discovery, personalized treatment plans.

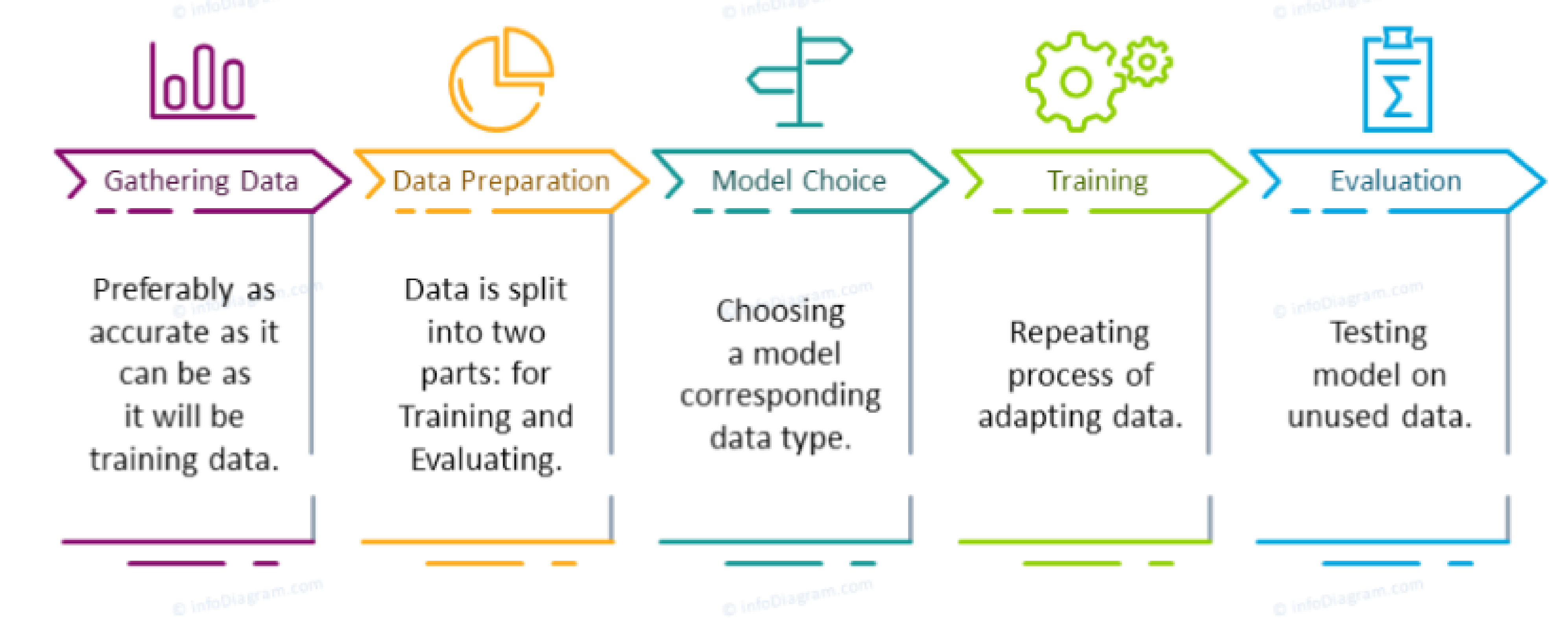
Finance

Fraud detection, risk assessment, algorithmic trading.

E-commerce

Personalized recommendations, targeted advertising, inventory management.

Machine Learning Process



Gathering Data

Preferably as accurate as it can be as it will be training data.

Data Preparation

Data is split into two parts: for Training and Evaluating.

Model Choice

Choosing a model corresponding data type.

Training

Repeating process of adapting data.

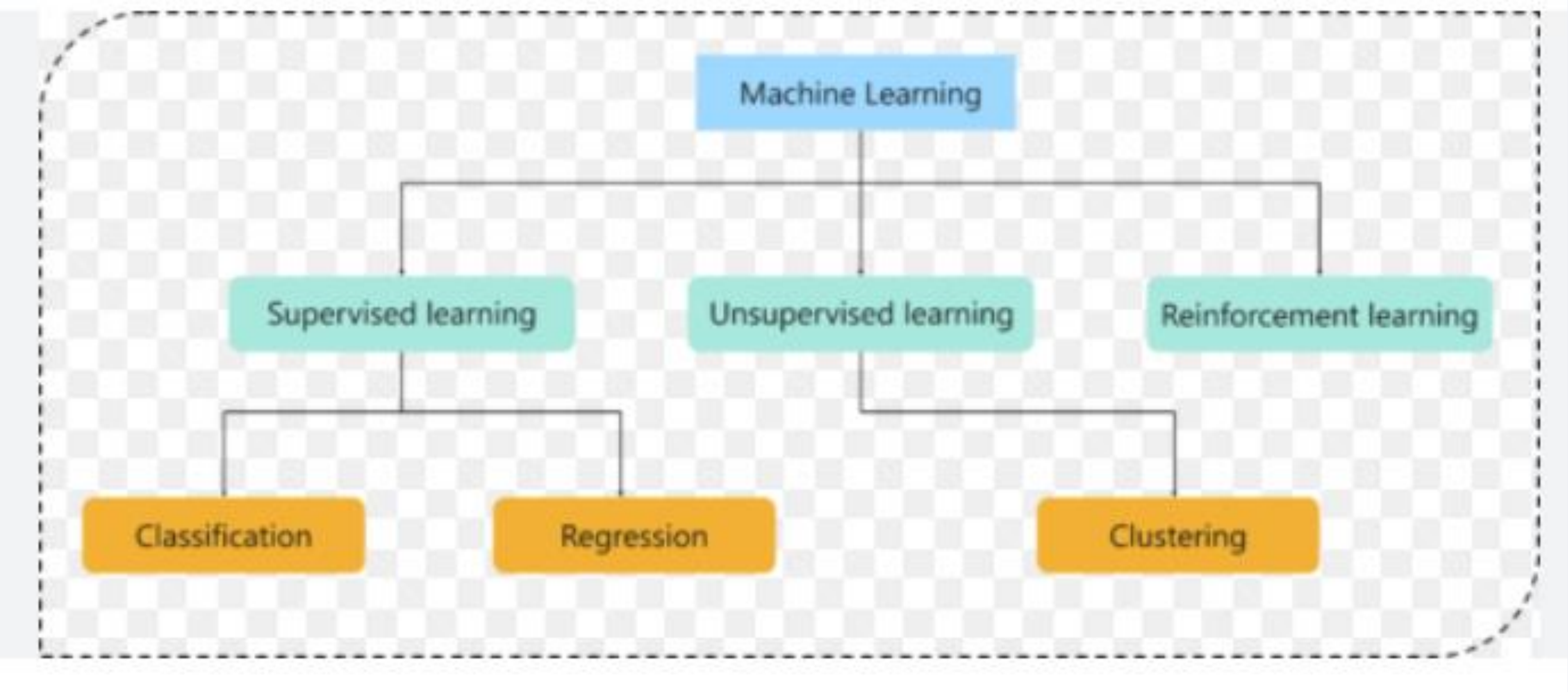
Evaluation

Testing model on unused data.

Machine Learning Approaches

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions.
There are four basic types of machine learning:

- Supervised learning,
- Unsupervised learning,
- Semi-supervised learning and
- Reinforcement learning



1. Supervised Learning Algorithms

- Supervised learning algorithms learn from labeled data, where the input features and corresponding output labels are provided.
- Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

  Y = f(X)

- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

- Learning stops when the algorithm achieves an acceptable level of performance.

- For example, the inputs could be the weather forecast, and the outputs would be the visitors to the beach.
- The goal in supervised learning would be to learn the mapping that describes the relationship between temperature and number of beach visitors.

- Being able to adapt to new inputs and make predictions is the crucial generalization part of machine learning.

- The output from a supervised Machine Learning model could be a category from a finite set e.g [low, medium, high].
- When this is the case, it’s is deciding how to classify the input, and so is known as classification.

  Input [temperature=20] -> Model -> Output = [visitors=high]

- Alternatively, the output could be a real-world scalar (output a number).When this is the case, it is known as regression

  Input [temperature=20] -> Model -> Output = [visitors=300]

Application areas

- Optical character recognition - recognizing character codes from their images
- Face recognition
- Medical diagnosis- the inputs are the relevant information we have about the patient and the classes are the illnesses
- speech recognition - the input is acoustic and the classes are words that can be uttered
- Natural language processing is used in sentimental analysis and spam filtering.
- Bio metric system uses bio-metric features for acceptance or rejection.
- Another use of machine learning is outlier detection, which is finding the instances that do not obey the rule and are exceptions.

Classification

- Classification is used to group the similar data points into different sections in order to classify them.
- Machine Learning is used to find the rules that explain how to separate the different data points.
- There are multiple ways to discover the rules.
- They all focus on using data and answers to discover rules that linearly separate data points.
- Linear separability is a key concept in machine learning.
- Linear separability means is ‘can the different data points be separated by a line?’.

- The lines drawn between classes are known as the decision boundaries.
- The entire area that is chosen to define a class is known as the decision surface.
- The decision surface defines that if a data point falls within its boundaries, it will be assigned a certain class.

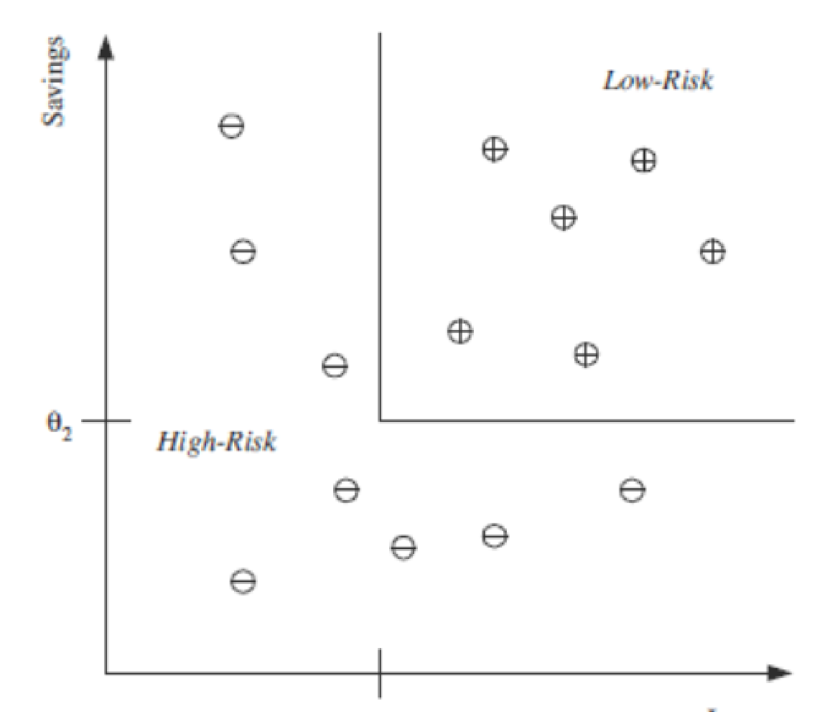
- After training with the past data, a classification rule learned may be of the form

IF income> ϑ1 AND savings> ϑ2 THEN low-risk ELSE high-risk

for suitable values of θ1 and θ2 .

This is an example of a discriminant;
 it is a function that separates the examples of different classes.

In some cases, instead of making a 0/1 (low-risk/high-risk) type decision, we may want to calculate a probability, namely, P(Y|X),



Regression

- Regression is another form of supervised learning.
- The difference between classification and regression is that regression outputs a number rather than a class.
- Regression is useful when predicting number based problems like stock market prices, the temperature for a given day, or the probability of an event.

- For eg: a system that can predict the price of a used car.
- Inputs are the car attributes—brand, year, engine capacity, mileage, etc
- The output is the price of the car.
- Such problems where the output is a number are regression problems.
- Regression is used in financial trading to find the patterns in stocks and other assets to decide when to buy/sell and make a profit.

The approach in machine learning is that we assume a model defined up to a set of parameters:

$$y=g(x|\theta)$$

where $g(\cdot)$ is the model and θ are its parameters.
Y is a number in regression
$g(\cdot)$ is the regression function

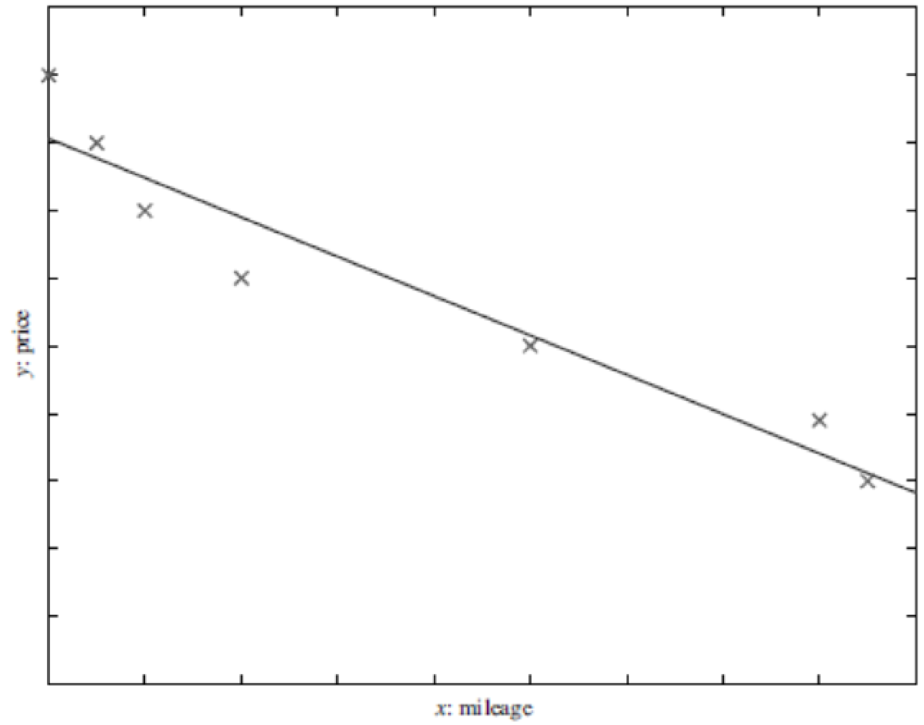
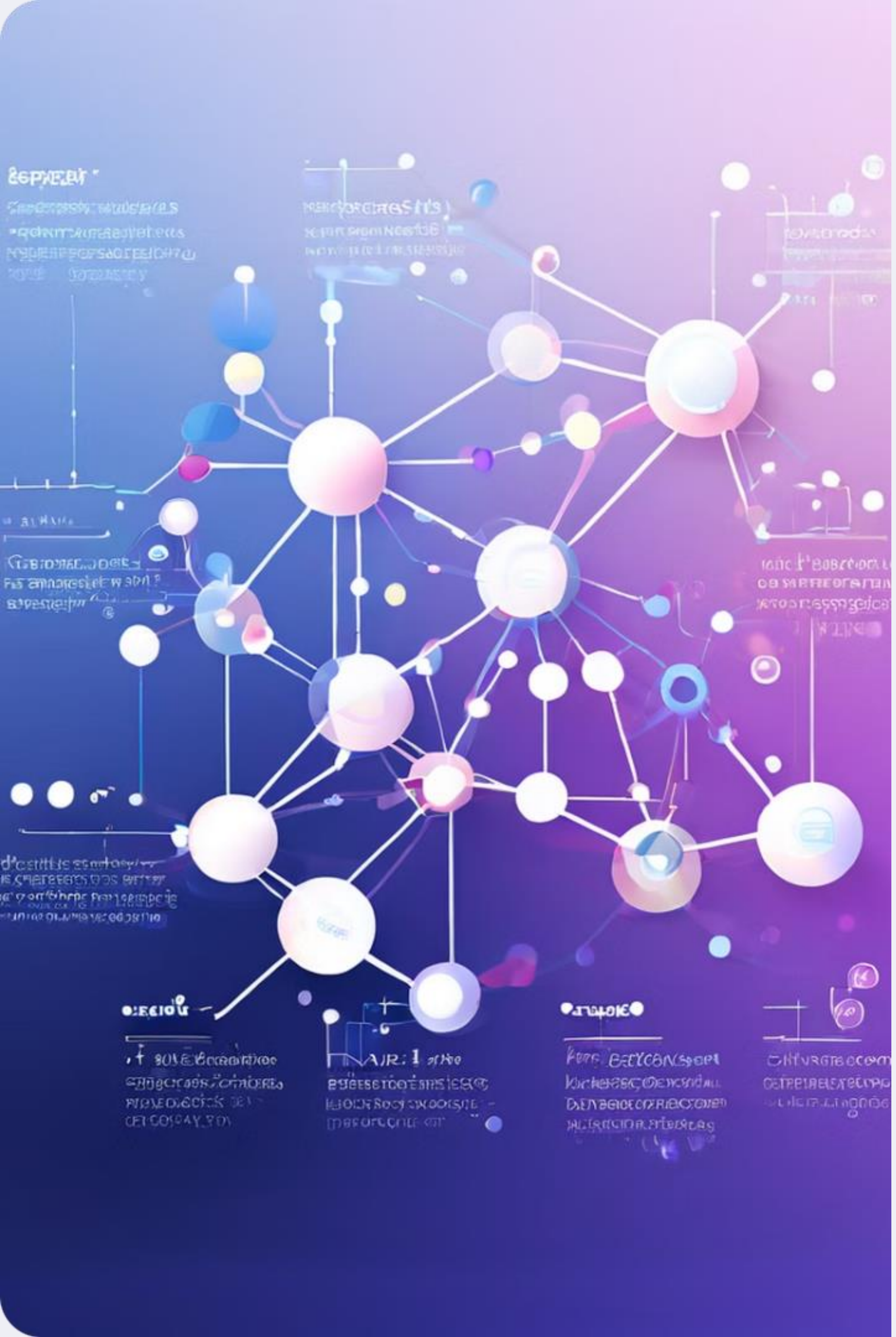


Figure A training dataset of used cars and the function fitted. For simplicity, mileage is taken as the only input attribute and a linear model is used.



2. Unsupervised Learning Algorithms

Unsupervised learning algorithms learn from unlabeled data, discovering patterns and structures without explicit guidance.

Clustering

Grouping similar data points together, uncovering hidden patterns in data.

Dimensionality Reduction

Simplifying complex datasets by reducing the number of features while preserving important information.

Association

Identifying relationships and patterns between different variables in data.

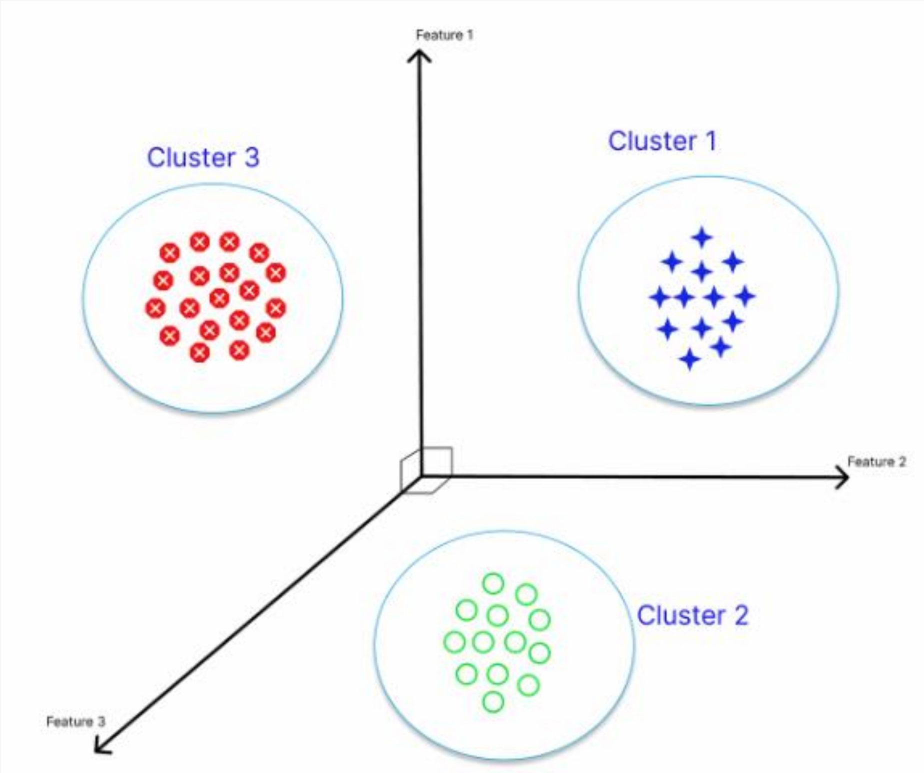
Anomaly Detection

Identifying outliers or unusual data points that deviate from normal behavior.

Clustering

Clustering is a method used for density estimation where the aim is to find clusters or groupings of input.

- For eg: a clustering model allocates customers similar in their attributes to the same group, providing the company with natural groupings of its customers called customer segmentation.
- Once such groups are found, the company may decide strategies, for example, services and products, specific to different groups; this is known as customer relationship management
- An interesting application of clustering is in image compression.
- In document clustering, the aim is to group similar documents
- Machine learning methods are also used in bioinformatics
- Clustering has successfully been used in marketing, it is regularly used to cluster customers into similar groups based on their behaviors and characteristics.



Association

- **D**iscover rules that describe large portions of your data, such as people that buy X also tend to buy Y.
- For example, if a person watches video A they will likely watch video B.
- Association rules are perfect for examples such as this where you want to find related items.
- Association learning is used for recommending or finding related items.
- A common example is market basket analysis.

Anomaly Detection

- The identification of rare or unusual items that differ from the majority of data.
- For example, your bank will use this to detect fraudulent activity on your card.
- Anomaly detection uses unsupervised learning to separate and detect these strange occurrences.

3. **Semi-supervised learning**

- Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems.
- These problems sit in between both supervised and unsupervised learning.
- A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled.
- Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.
- You can use unsupervised learning techniques to discover and learn the structure in the input variables.
- You can also use supervised learning techniques to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.
- A perfect example is in medical scans. A trained expert is needed to label these which is time consuming and very expensive. Instead, an expert can label just a small set of cancer scans, and the semi-supervised algorithm would be able to leverage this small subset and apply it to a larger set of scans.

Generative Adversarial Networks

- Generative Adversarial Networks (GANs) have been a recent breakthrough with incredible results.
- GANs use two neural networks, a generator and discriminator.
- The generator generates output and the discriminator critiques it.
- By battling against each other they both become increasingly skilled.

- By using a network to both generate input and another one to generate outputs there is no need for us to provide explicit labels every single time and so it can be classed as semi-supervised

4. Reinforcement Learning Algorithms

Reinforcement learning algorithms learn through trial and error, interacting with an environment to maximize rewards.

1 **Agent**

The learning system that interacts with the environment.

2 **Environment**

The context in which the agent operates and receives feedback.

3 **Reward**

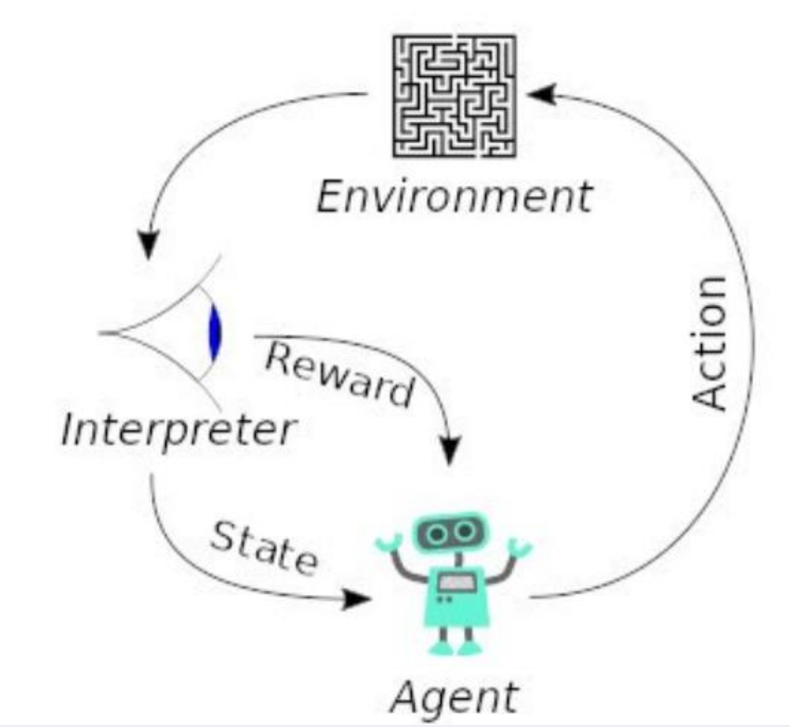
Positive feedback received by the agent for performing desirable actions.

4 **State**

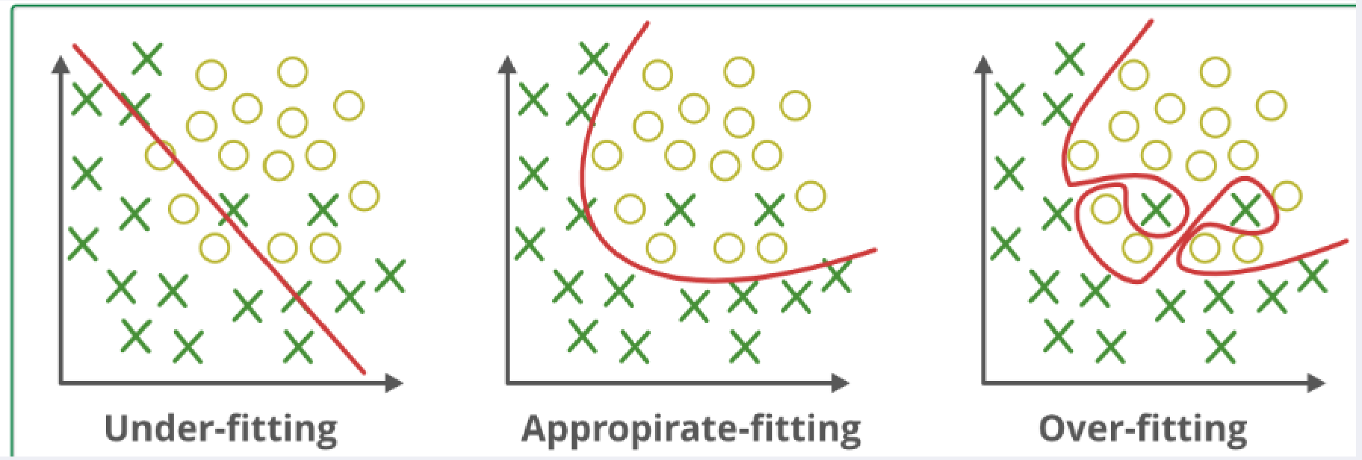
The current condition of the environment, providing information to the agent.

5 **Action**

The decision made by the agent to interact with the environment.



Overfitting and Underfitting



Overfitting and underfitting are common challenges in machine learning, affecting model performance and generalizability.

Overfitting

The model learns the training data too well, resulting in poor performance on unseen data.

Underfitting

The model fails to capture the underlying patterns in the data, resulting in poor performance on both training and unseen data.

Hyperparameters and Tuning

Hyperparameters are settings that control the learning process and influence the model's performance. Tuning these parameters is crucial for optimizing the model's accuracy and generalizability.

1 **Learning Rate**

The step size taken during the optimization process.

2 **Regularization**

Techniques used to prevent overfitting by adding a penalty to the model's complexity.

3 **Number of Hidden Layers**

The depth of the neural network architecture, determining the model's complexity.

4 **Activation Function**

The function used to introduce non-linearity into the model, enabling it to learn complex patterns.

Introduction to Validation Set

- The validation set is a crucial part of model development and evaluation in machine learning.
- It provides an unbiased assessment of your model's performance on unseen data.
- Specifically, we split the training data into two disjoint subsets. One of these subsets is used to learn the parameters.
- The other subset is our validation set, used to estimate the generalization error during or after training, allowing for the hyperparameters to be updated accordingly.
- The subset of data used to guide the selection of hyperparameters is called the validation set.
- Typically, one uses about 80% of the training data for training and 20% for validation.

k-fold cross-validation

- A partition of the dataset is formed by splitting it into k non-overlapping subsets.
- The test error may then be estimated by taking the average test error across k trials.
- On trial i, the i -th subset of the data is used as the test set and the rest of the data is used as the training set.

Importance of Validation Set

Unbiased Evaluation

The validation set provides an unbiased estimate of model performance on unseen data, which is essential for generalization.

Hyperparameter Tuning

The validation set helps to find the optimal hyperparameters for the model, leading to better performance.

Model Selection

The validation set can be used to compare different models and select the best one based on their performance.

Point estimation

In statistics, point estimation involves the use of sample data to calculate a single value which is to serve as a "best guess" or "best estimate" of an unknown population parameter

Point Estimators

A point estimator is a function that is used to find an approximate value of a population parameter from random samples

Bias

Bias refers to the error caused by the model's assumptions. High bias models tend to be too simple and may underfit the data.

Bias measures the expected deviation from the true value of the function or parameter.

Variance

Variance refers to the sensitivity of a model to the training data. High variance models tend to be too complex and may overfit the data.

Variance provides a measure of the deviation from the expected estimator value.

The square root of variance is called Standard Error SE

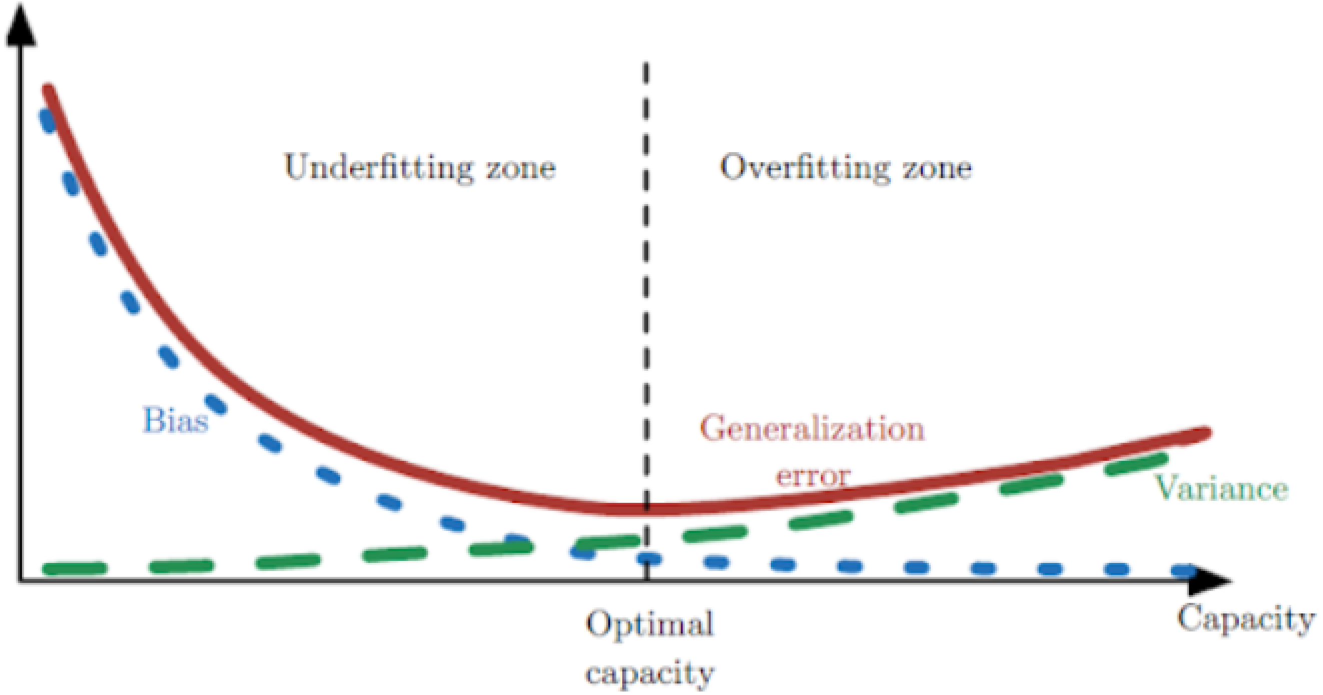
Trading off Bias and Variance

- **B**ias and variance measure two different sources of error in an estimator.
- The most common way to negotiate the trade-off is to use cross-validation.
- Empirically, cross-validation is highly successful on many real-world tasks.
- Alternatively, we can also compare the mean squared error (MSE) of the estimates
- The MSE measures the **overall expected deviation**
- Desirable estimators are those with small MSE
- These are estimators that manage to keep both their bias and variance somewhat in check.

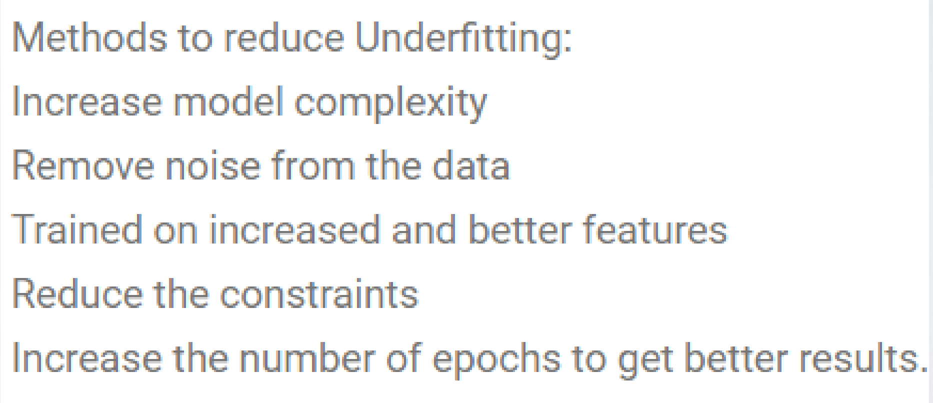
$$Bias^2(\hat{\theta}_m) + Var(\hat{\theta}_m) = MSE$$

- As the number of data points m in our dataset increases, our point estimates converge to the true value of the corresponding parameters
- The condition described by equation is known as **consistency**.
- The symbol $plim$ indicates convergence in probability

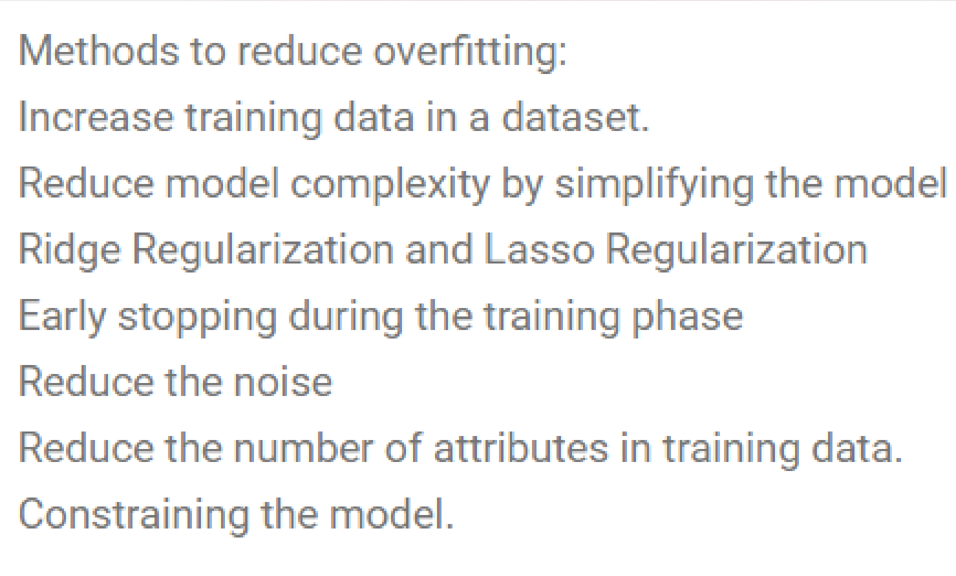
$$plim_{m \to \infty} = \hat{\theta}_m = \theta$$



Challenges in Machine Learning



Methods to reduce Underfitting:
Increase model complexity
Remove noise from the data
Trained on increased and better features
Reduce the constraints
Increase the number of epochs to get better results.



Methods to reduce overfitting:
Increase training data in a dataset.
Reduce model complexity by simplifying the model
Ridge Regularization and Lasso Regularization
Early stopping during the training phase
Reduce the noise
Reduce the number of attributes in training data.
Constraining the model.

1

Data Quality

Poor data quality, such as missing values, inconsistencies, or outliers, can severely impact model performance.

2

Overfitting & Underfitting

When a model learns the training data too well, it may fail to generalize to new data. This is known as overfitting.

3

Feature Engineering

Selecting and transforming the right features is crucial for building effective models. It requires domain expertise and careful analysis.

Linear Regression

- Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
- When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.
- Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there are more than one dependent variables, it is known as Multivariate Regression.

Simple Linear Regression

- This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.
- Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available.
- When we have a single input, we can use statistics to estimate the coefficients.
- The dependent variable must be a continuous/real value.
- The objectives are:
- **Model the relationship between the two variables**. Such as the relationship between Income and expenditure, experience and Salary, etc.
- **Forecasting new observations. Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.**

Simple Linear Regression

1

Data Preparation

Clean and prepare your data. Ensure it is in a suitable format and handle missing values or outliers.

2

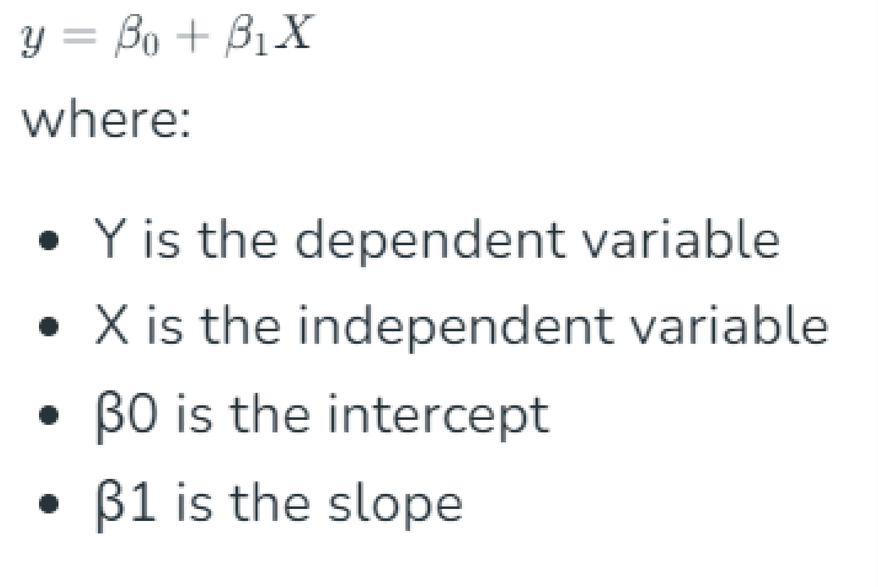
Model Training

Train a linear regression model on the prepared data. The model will learn a linear relationship between the input and output variables.

3

Model Evaluation

Evaluate the model's performance on the validation set to assess its ability to generalize to unseen data.



$y = \beta_0 + \beta_1 X$

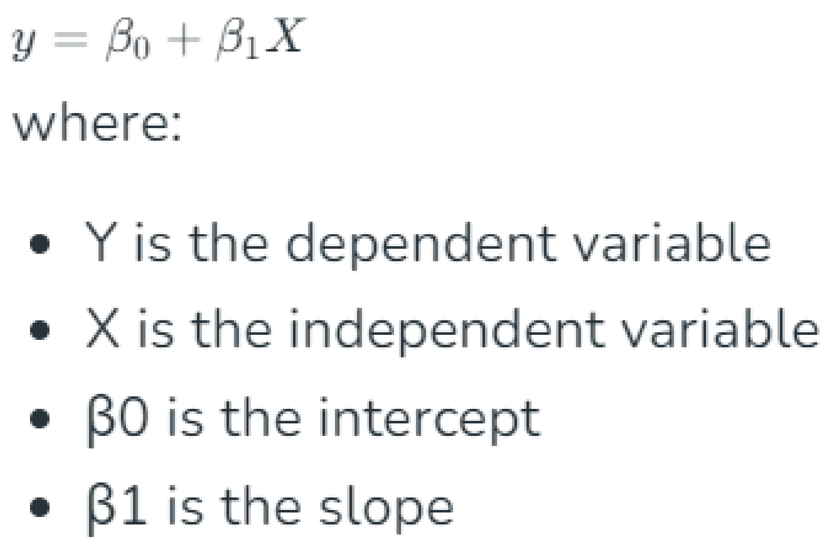
where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Best Fit Line

- Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum.
- There will be the least error in the best-fit line.
- The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables.
- The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

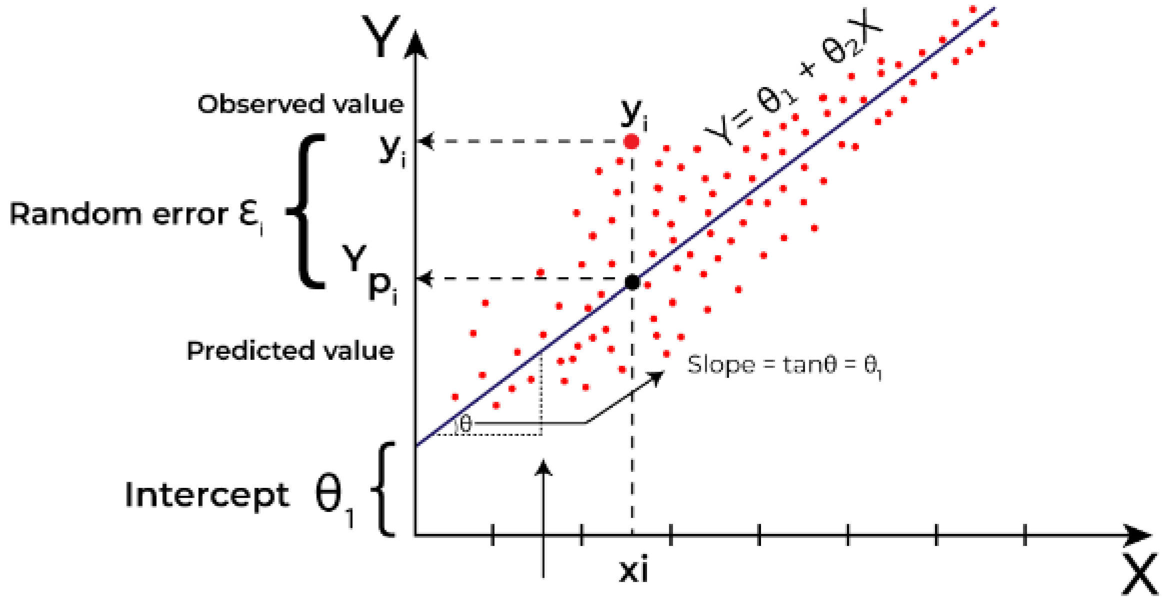
- Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y.
- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)).
- The regression line is the best-fit line for our model.
- We utilize the cost function to compute the best values in order to get the best fit line since different values for weights or the coefficient of lines result in different regression lines.



$$y = \beta_0 + \beta_1 X$$

where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope



Y
Observed value
y_i
y_i
Random error ε_i
Y_{p_i}
Predicted value
$Y = \theta_1 + \theta_2 X$
Slope = tanθ = θ_1
θ
Intercept θ_1
xi
X

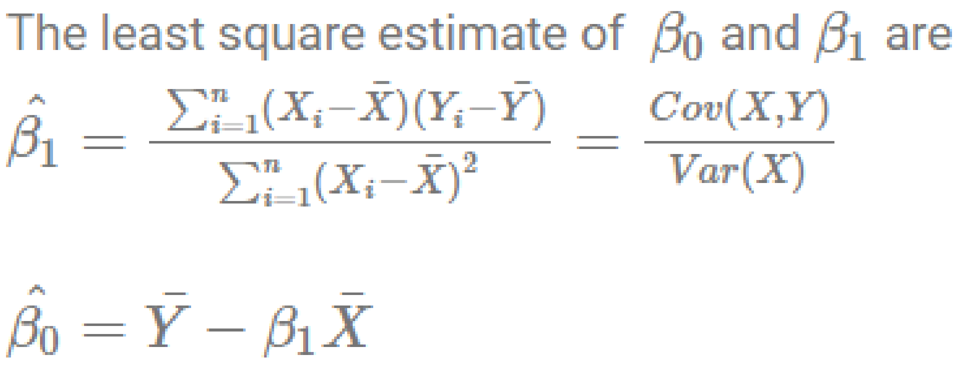
Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called **Ordinary Least Squares**.
It is common to therefore refer to a model prepared this way as **Ordinary Least Squares Linear Regression** or just **Least Squares Regression.**

In a simple regression problem (a single x and a single y), the form of the model would be:

$$y = \beta_0 + \beta_1 * x$$

The ordinary least square estimate of β0 and β1 try to minimize the sum of squared error

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \text{ where } \hat{Y} \text{ is the estimated value of } Y$$



The least square estimate of β_0 and β_1 are

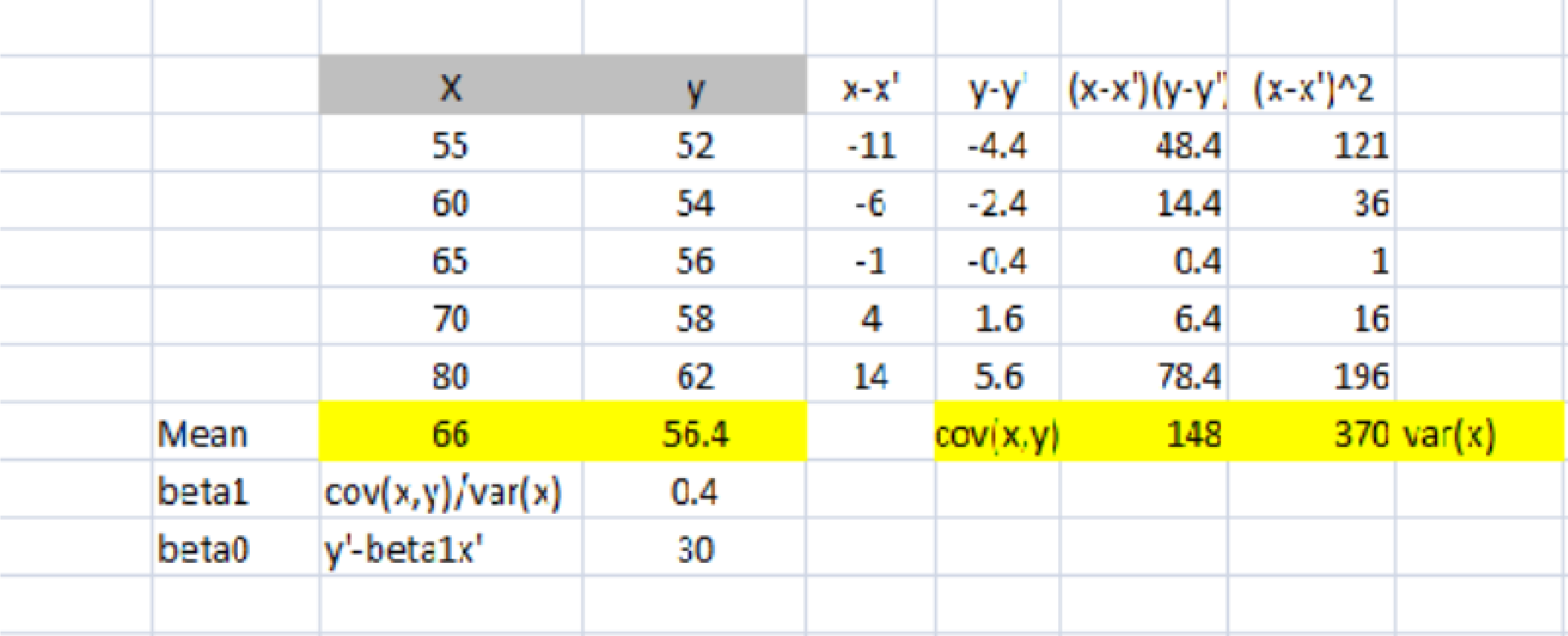
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

Example: (University question)

Determine the regression equation by finding the regression slope coefficient and the intercept value using the following data.

$$\begin{vmatrix} x & 55 & 60 & 65 & 70 & 80 \\ y & 52 & 54 & 56 & 58 & 62 \end{vmatrix}$$

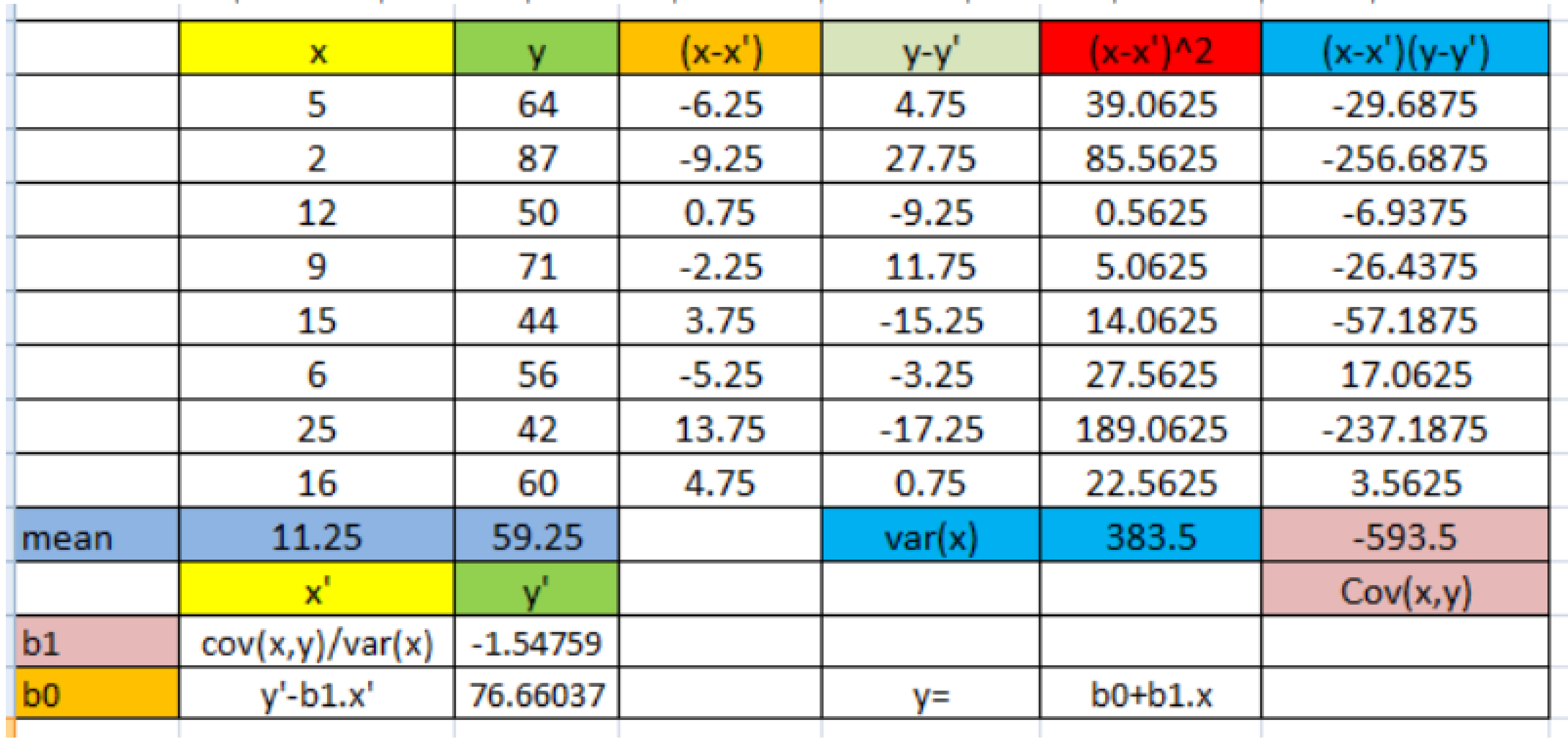


	x	y	x-x'	y-y'	(x-x')(y-y'	(x-x')^2	
	55	52	-11	-4.4	48.4	121	
	60	54	-6	-2.4	14.4	36	
	65	56	-1	-0.4	0.4	1	
	70	58	4	1.6	6.4	16	
	80	62	14	5.6	78.4	196	
Mean	66	56.4		cov(x,y)	148	370	var(x)
beta1	cov(x,y)/var(x)	0.4					
beta0	y'-beta1x'	30					

so the regression equation is $y = 30 + 0.4x$

Use the following data to construct a linear regression model for the auto insurance premium as a function of driving experience. (university question)

DrivingExp	5	2	12	9	15	6	25	16
MonthlyPremium	64	87	50	71	44	56	42	60



	x	y	(x-x')	y-y'	(x-x')^2	(x-x')(y-y')
	5	64	-6.25	4.75	39.0625	-29.6875
	2	87	-9.25	27.75	85.5625	-256.6875
	12	50	0.75	-9.25	0.5625	-6.9375
	9	71	-2.25	11.75	5.0625	-26.4375
	15	44	3.75	-15.25	14.0625	-57.1875
	6	56	-5.25	-3.25	27.5625	17.0625
	25	42	13.75	-17.25	189.0625	-237.1875
	16	60	4.75	0.75	22.5625	3.5625
mean	11.25	59.25		var(x)	383.5	-593.5
	x'	y'				Cov(x,y)
b1	cov(x,y)/var(x)	-1.54759				
b0	y'-b1.x'	76.66037		y=	b0+b1.x	

y=b0+b1.x

y=76.66-1.55x

Logistic Regression

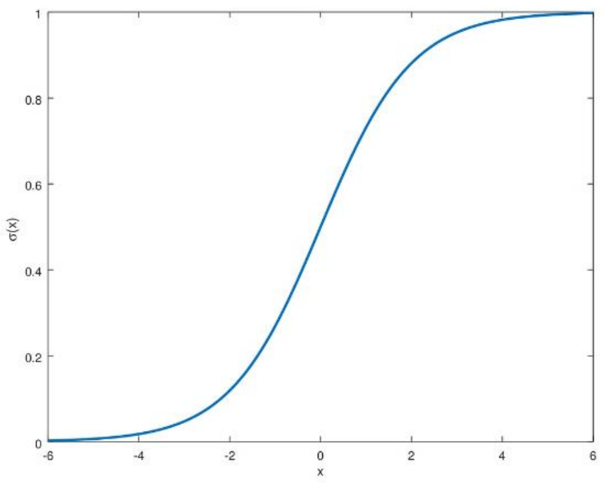
- Logistic Regression is a **supervised machine learning algorithm** used for **classification tasks**
- The goal is to predict the probability that an instance belongs to a given class or not.
- Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.

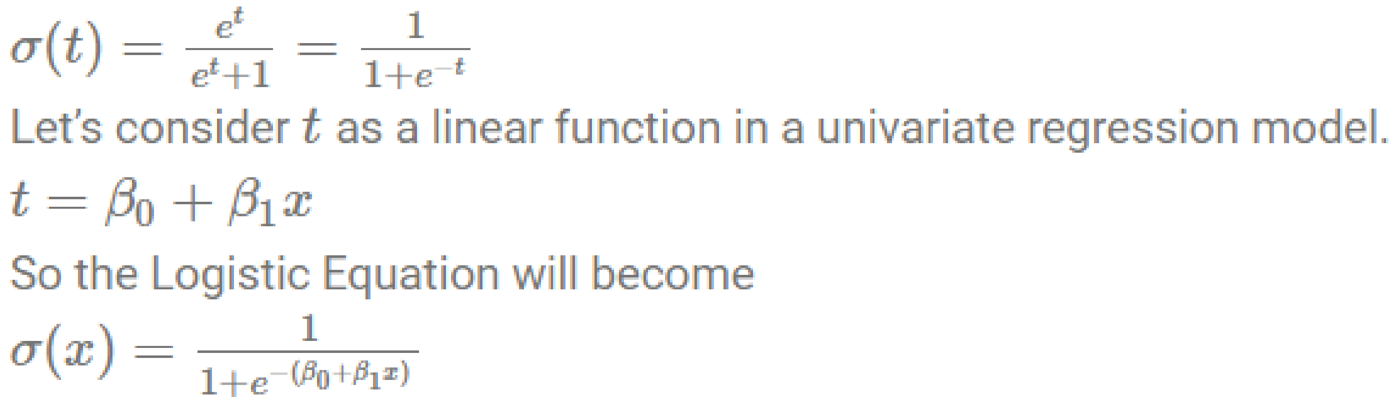
In Logistic regression, instead of fitting a regression line, we fit an “S” shaped logistic function, which predicts two maximum values (0 or 1).

Algorithm	Simple linear regression	Logistic regression
Output Variable	Continuous	Binary (0 or 1)
Function	Linear	Sigmoid
Applications	Predicting numerical values	Classifying data into two categories

Logistic Function – Sigmoid Function

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the “S” form.
- The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.





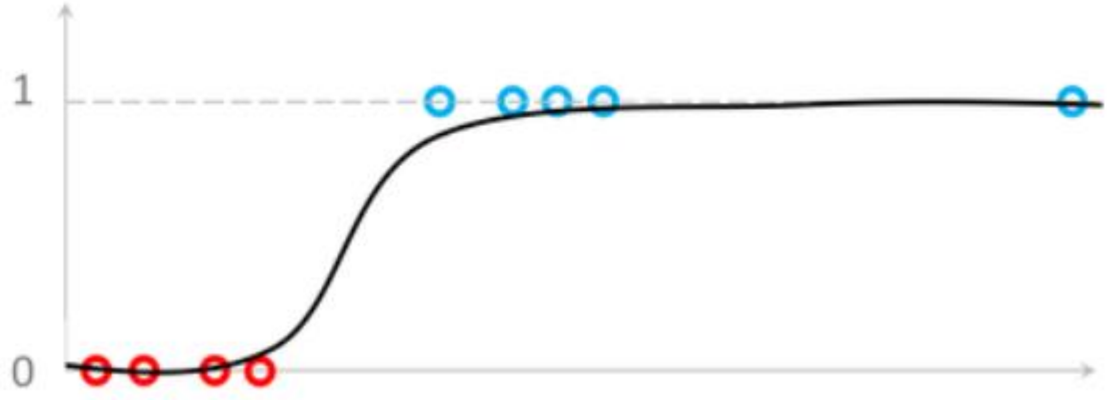
$$\sigma(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}}$$

Let’s consider t as a linear function in a univariate regression model.

$$t = \beta_0 + \beta_1 x$$

So the Logistic Equation will become

$$\sigma(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$



Types of Logistic Regression

On the basis of the categories, Logistic Regression can be classified into three types:

1. **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
3. **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”.

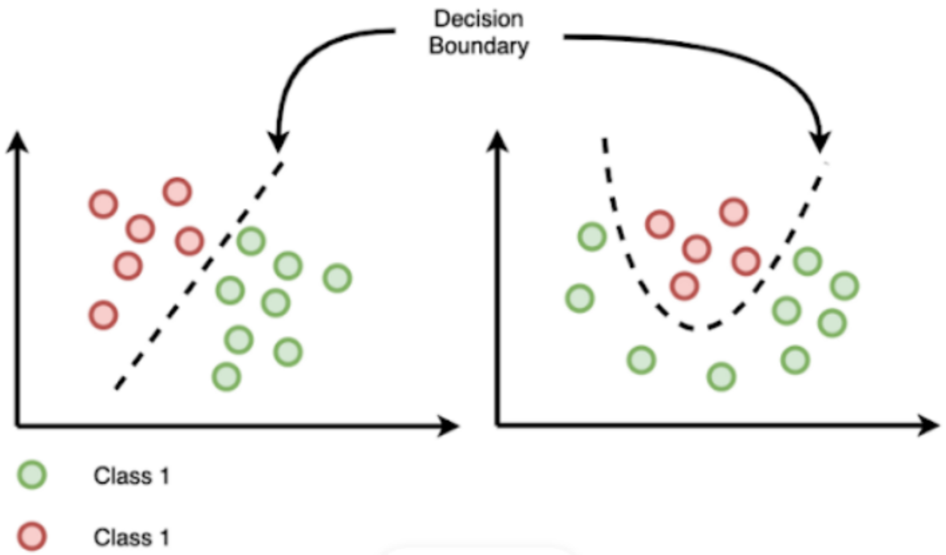
Decision Boundary

To predict which class a data belongs, a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes.

Say, if $predicted\text{–}value \geq 0.5$, then classify email as spam else as not spam.

Decision boundary can be linear or non-linear.
Polynomial order can be increased to get complex decision boundary.
For the binary classification problem , the decision boundary is always linear.

The fundamental application of logistic regression is to determine a decision boundary for a binary classification problem.
The approach can be very well applied for scenarios with multiple classification classes or multi-class classification.



Multinomial Classification problem

- The probability distribution that defines multi-class probabilities is called a multinomial probability distribution.
- A logistic regression model that is adapted to learn and predict a multinomial probability distribution is referred to as **Multinomial Logistic Regression**.
- Similarly, we might refer to default or standard logistic regression as Binomial Logistic Regression.

- **Binomial Logistic Regression**: Standard logistic regression that predicts a binomial probability (i.e. for two classes) for each input example.

- **Multinomial Logistic Regression**: Modified version of logistic regression that predicts a multinomial probability (i.e. more than two classes) for each input example.

Advantages of Logistic Regression

The advantages of the logistic regression are as follows:

1. Logistic Regression is very easy to understand.
2. It requires less training.
3. It performs well for simple datasets as well as when the data set is linearly separable.
4. It doesn’t make any assumptions about the distributions of classes in feature space.
5. A Logistic Regression model is less likely to be over-fitted but it can overfit in high dimensional datasets. To avoid over-fitting these scenarios, One may consider regularization.

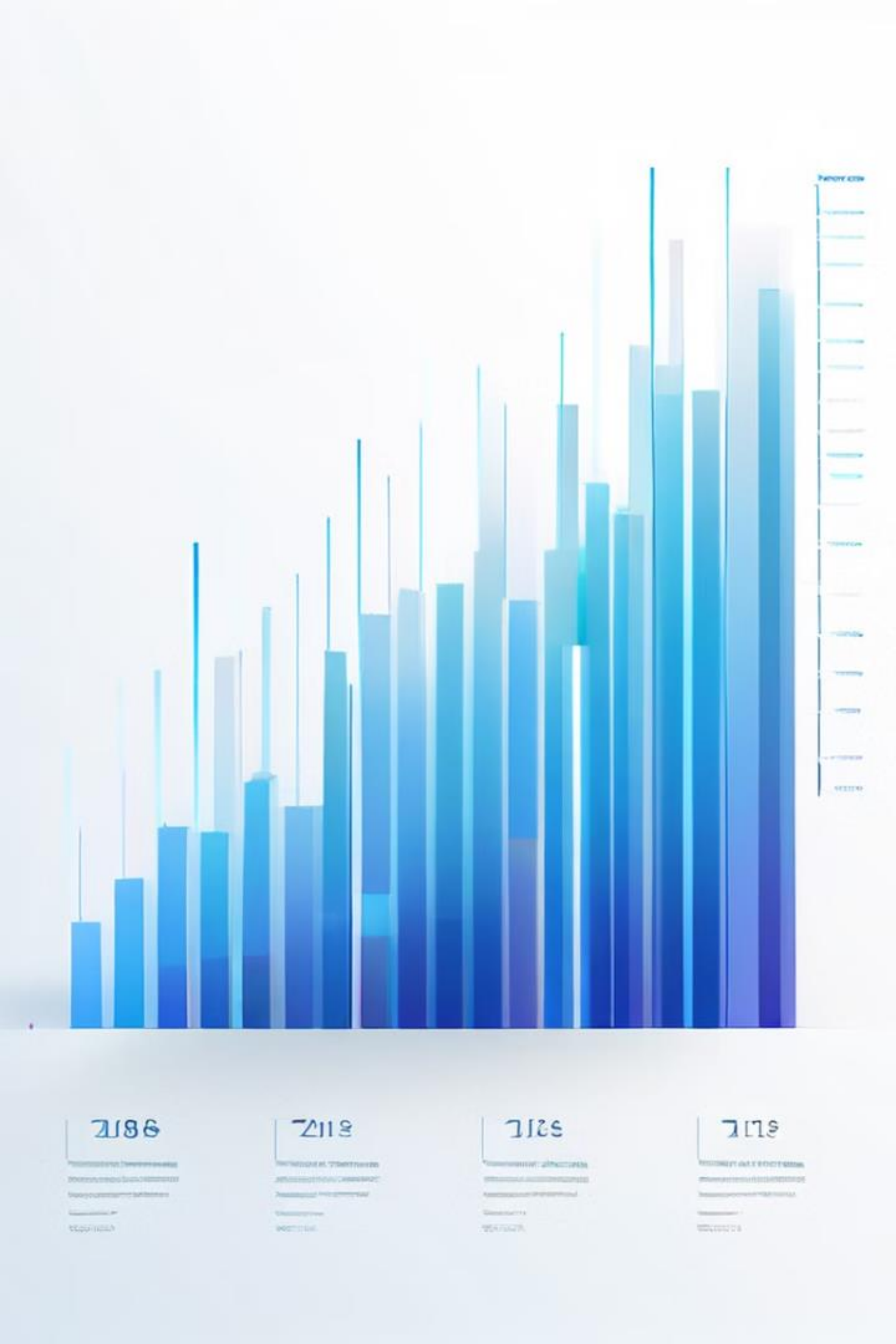
Disadvantages of Logistic Regression

The disadvantages of the logistic regression are as follows:

1. Sometimes a lot of Feature Engineering is required.
2. If the independent features are correlated with each other it may affect the performance of the classifier.
3. It is quite sensitive to noise and overfitting.
4. Logistic Regression should not be used if the number of observations is lesser than the number of features, otherwise, it may lead to overfitting.
5. By using Logistic Regression, non-linear problems can’t be solved because it has a linear decision surface.

Introduction to Performance Measures Measures

Performance measures are essential for evaluating the effectiveness of machine learning models. They provide insights into the model's ability to accurately predict outcomes. Numerous metrics are used in the evaluation of a machine learning model. Selection of the most suitable metrics is important to fine-tune a model based on its performance.



Classification Accuracy and its Limitations

Classification accuracy is the ratio of correct predictions to total predictions made.

$$Classificationaccuracy = correctpredictions / totalpredictions$$

It is often presented as a percentage by multiplying the result by 100.

$$classificationaccuracy=(correctpredictions/totalpredictions)\ast 100$$

Classification accuracy can also easily be turned into a misclassification rate or error rate by inverting the value, such as:

$$errorrate=(1-(correctpredictions/totalpredictions))\ast 100$$

The main problem with classification accuracy is that it hides the detail you need to better understand the performance of your classification model.
Classification accuracy can hide the detail you need to diagnose the performance of your model. But thankfully we can tease apart this detail by using a confusion matrix.

Confusion Matrix

The confusion matrix is a foundational tool for understanding model performance.
It classifies predictions into four categories: true positives, true negatives, false positives, and false negatives.
A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier.
It is used to measure the performance of a classification model.
It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score.

True Positives (TP): The number of points that the classifier correctly predicts as positive:

False Positives (FP)Type 1 error: The number of points the classifier predicts to be positive, which in fact belong to the negative class:

False Negatives (FN)Type 2 error: The number of points the classifier predicts to be in the negative class, which in fact belong to the positive class:

True Negatives (TN) : The number of points that the classifier correctly predicts as negative:

Accuracy

Accuracy measures the overall proportion of correct predictions.
It is calculated by dividing the sum of true positives and true negatives by the total number of predictions.

$$Accuracy = \frac{TP+TN}{n}$$

Error Rate

The error rate for the binary classification case is given as the fraction of mistakes (or false predictions):

$$ErrorRate = \frac{FP+FN}{n}$$

Precision

Precision indicates the proportion of correctly predicted positive cases among all cases among all predicted positive cases.
It is calculated by dividing the number of true positives by the sum of true true positives and false positives.

$$prec_P = \frac{TP}{TP+FP}$$

$$prec_N = \frac{TN}{TN+FN}$$

Recall

Recall measures the proportion of correctly predicted positive cases among all actual positive cases.
It is calculated by dividing the number of true positives by the sum of true positives and false negatives.

$$recall_P = \frac{TP}{TP+FN}$$

$$recall_N = \frac{TN}{FP+TN}$$

F-measure (F1 Score)

F-measure tries to balance the precision and recall values, by computing their harmonic mean
For a perfect classifier, the maximum value of the F-measure is 1.
The higher the Fi value the better the classifier.
The overall F-measure for the classifier M is the mean of the class-specific values:

$$F_i = \frac{2}{\frac{1}{Prec_i} + \frac{1}{Recall_i}} = 2.\frac{Prec_i * Recall_i}{Prec_i + Recall_i}$$

Sensitivity

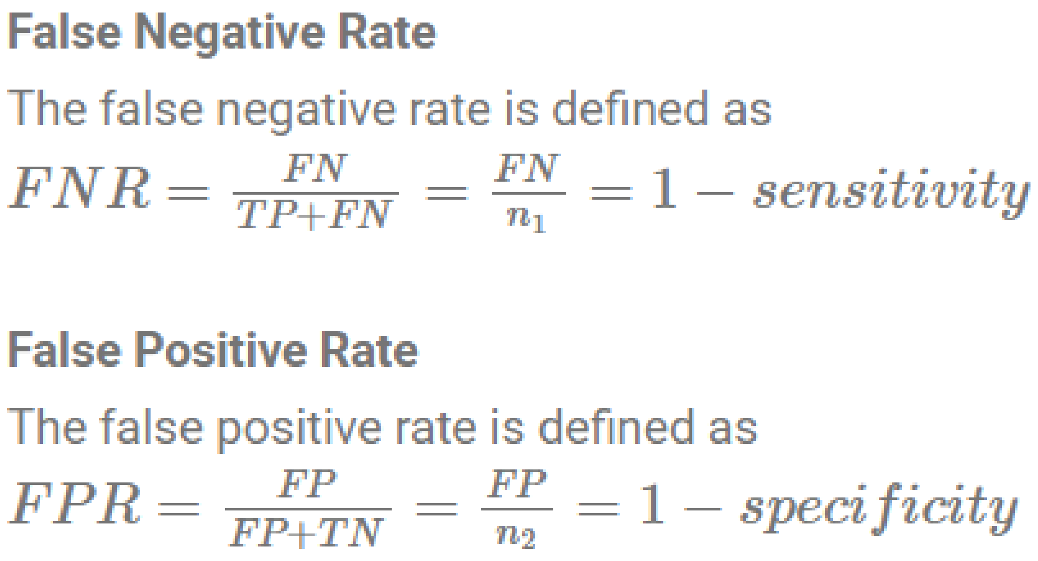
Sensitivity, also known as true positive rate, indicates the model's ability to identify actual positive cases. It is equivalent to recall.

$$TPR = sensitivity = recall_P = \frac{TP}{TP+FN}$$

Specificity

- Specificity, also known as true negative rate, is simply the recall for the negative classes.
- It indicates the model's ability to correctly identify actual negative cases.
- It is calculated by dividing the number of true negatives by the sum of true negatives and false positives.

$$TNR = specificity = recall_N = \frac{TN}{FP+TN}$$



False Negative Rate

The false negative rate is defined as

$$FNR = \frac{FN}{TP+FN} = \frac{FN}{n_1} = 1 - sensitivity$$

False Positive Rate

The false positive rate is defined as

$$FPR = \frac{FP}{FP+TN} = \frac{FP}{n_2} = 1 - specificity$$

When to use Accuracy / Precision / Recall / F1-Score?

a. Accuracy is used when the True Positives and True Negatives are more important. Accuracy is a better metric for Balanced Data.

b. Whenever False Positive is much more important use Precision.

c. Whenever False Negative is much more important use Recall.

d. F1-Score is used when the False Negatives and False Positives are important. F1-Score is a better metric for Imbalanced Data.

In the detection of spam mail, it is okay if any spam mail remains undetected (false negative), but what if we miss any critical mail because it is classified as spam (false positive).
In this situation, False Positive should be as low as possible. Here, precision is more vital as compared to recall.
Similarly, in the medical application, we don’t want to miss any patient. Therefore we focus on having a high recall.

Example (University Question)

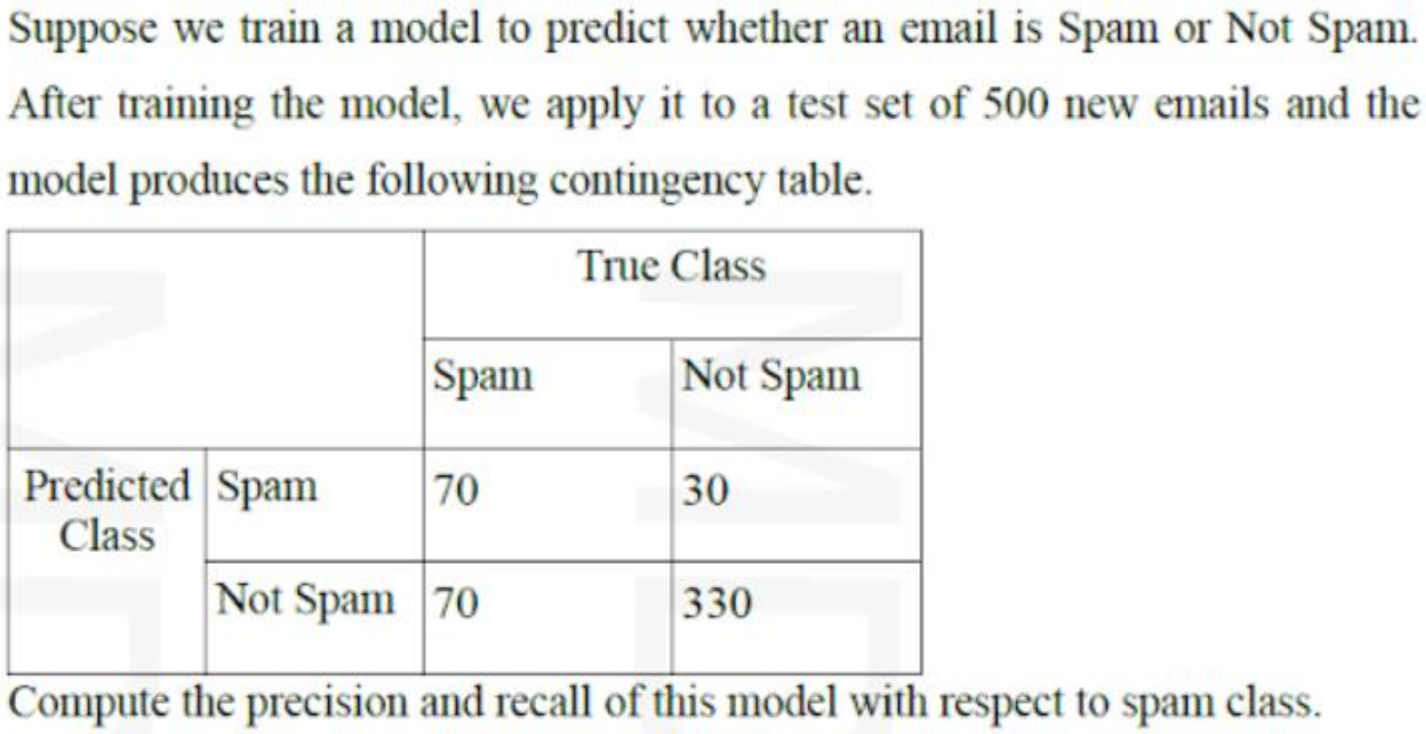
Imagine an application that is developed to recognize cats and dogs. This identifies eight dogs in a picture containing 12 dogs and some cats. Of the eight dogs identified, five actually are dogs while the rest are cats. Compute the True Positive, False Positive and False Negative values.

$TP = 5$

$FP = 8 - 5 = 3$

$FN = 12 - 5 = 7$

Example: University question



Suppose we train a model to predict whether an email is Spam or Not Spam. After training the model, we apply it to a test set of 500 new emails and the model produces the following contingency table.

		True Class	
		Spam	Not Spam
Predicted Class	Spam	70	30
	Not Spam	70	330

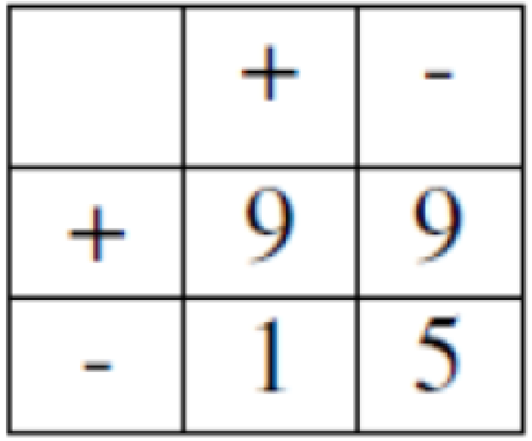
Compute the precision and recall of this model with respect to spam class.

$Precision = TP/(TP + FP) = 70/(70 + 30) = 0.70$

$Recall = TP/(TP + FN) = 70/(70 + 70) = 0.5$

Example:(University Question)

For a classifier, the confusion matrix is given
What is the precision, recall and accuracy of that classifier?

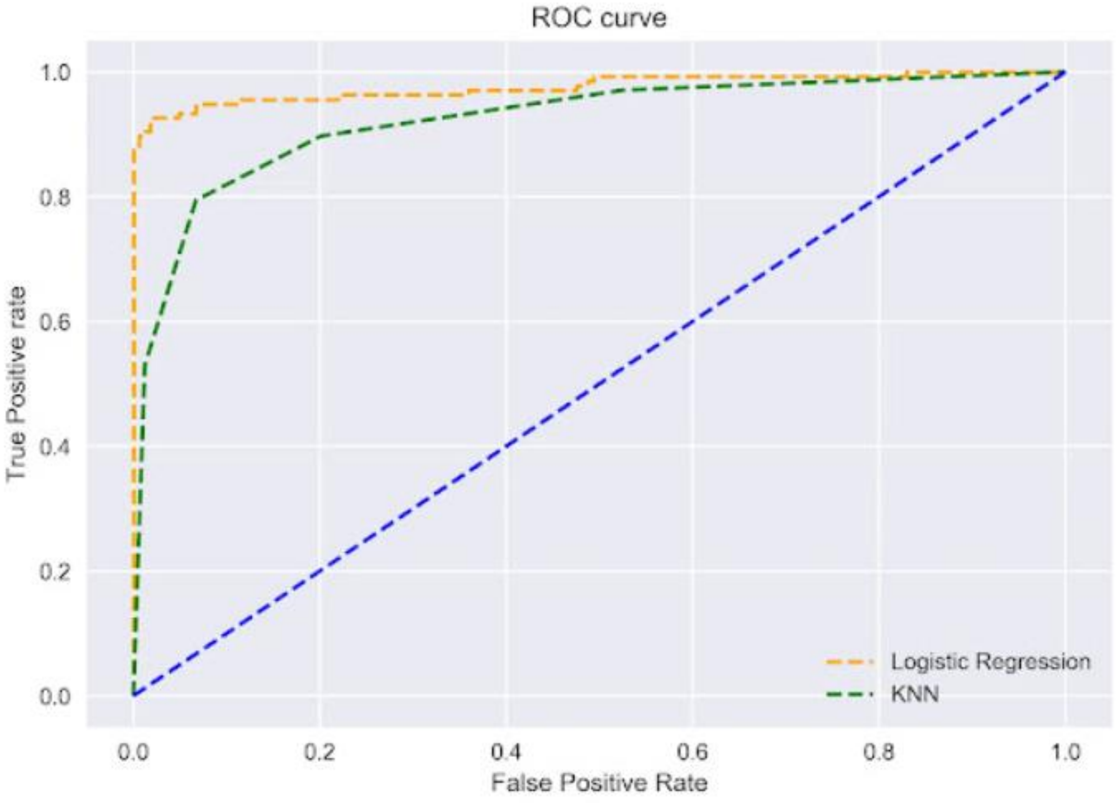


	+	-
+	9	9
-	1	5

$$prec_P = \frac{TP}{TP+FP} = 9/(9+9) = 9/18 = 0.5$$
$$prec_N = \frac{TN}{TN+FN} = 5/6 = 0.83$$

$$recall_P = sensitivity = TPR = \frac{TP}{TP+FN} = 9/(9+1) = 9/10 = 0.9$$
$$recall_N = specificity = TNR = \frac{TN}{TN+FP} = 5/(5+9) = 5/14 = 0.357$$

$$Accuracy = \frac{TP+TN}{n} = 14/24 = 0.583$$



Receiver Operating Characteristic (ROC) Curve

The ROC curve plots the true positive rate against the false positive rate at various threshold values.

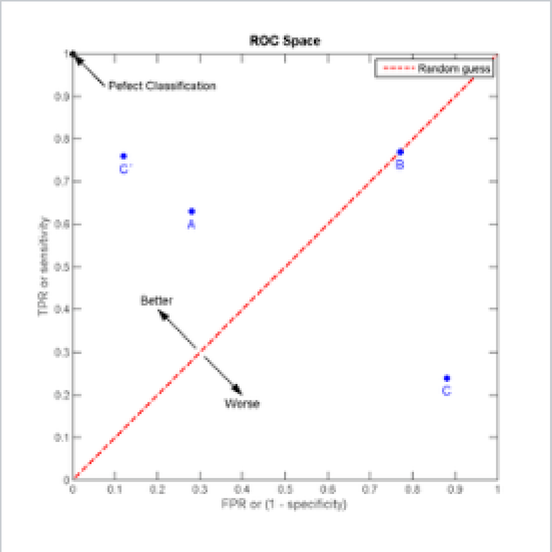
It helps visualize the trade-off between sensitivity and specificity.

ROC is a probability curve

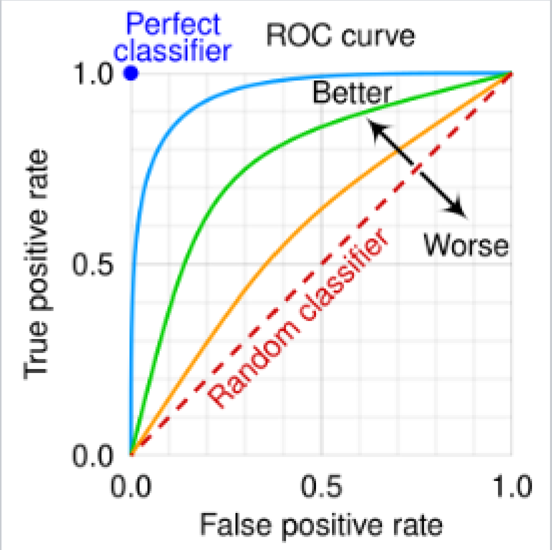
ROC is drawn by taking false positive rate in the x-axis and true positive rate in the y-axis.

ROC space

- A ROC space is defined by FPR and TPR as *x* and *y* axes, respectively, which depicts relative trade-offs between true positive and false positive.
- The diagonal divides the ROC space.
- Points above the diagonal represent good classification results (better than random); points below the line represent bad results (worse than random).



The ROC space and plots of the four prediction examples.



The ROC space for a "better" and "worse" classifier.

Area Under Curve (AUC)

The area under the ROC curve (AUC) provides a single, comprehensive measure of model performance.

A higher AUC indicates a better ability to distinguish between positive and negative cases.

AUC represents the degree or measure of separability

If we use a random model to classify, it has a 50% probability of classifying the positive and negative classes correctly. Here, the AUC = 0.5.

A perfect model has a 100% probability of classifying the positive and negative classes correctly.

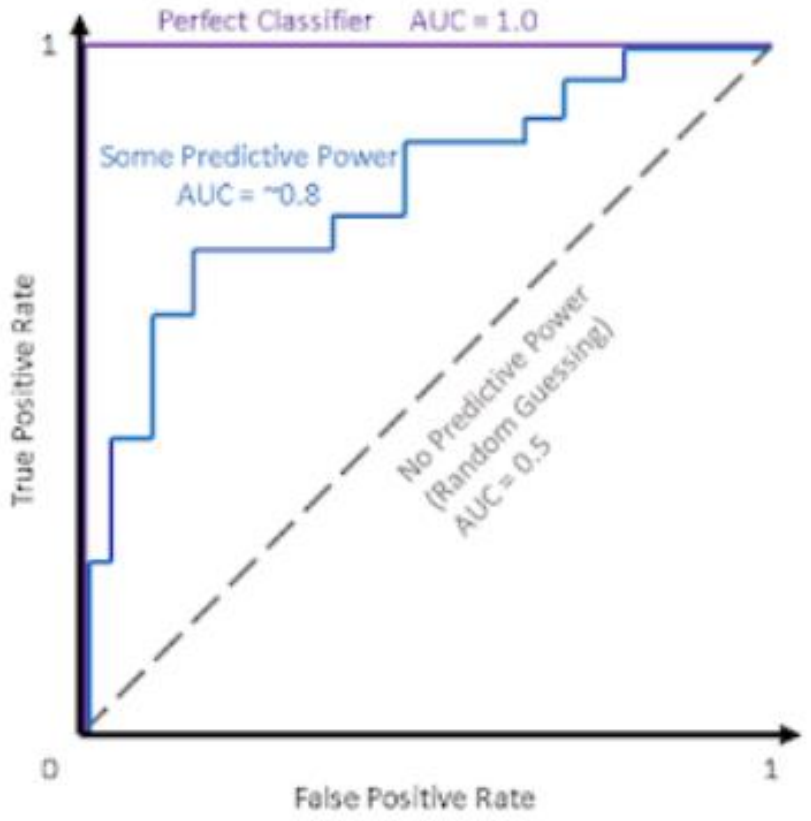
Here, the AUC = 1.

So when we want to select the best model, we want a model that is closest to the perfect model.

In other words, a model with AUC close to 1.

When we say a model has a high AUC score, it means the model's ability to separate the classes is very high (high separability).

This is a very important metric that should be checked while selecting a classification model.


Perfect Classifier AUC = 1.0
Some Predictive Power
AUC = ~0.8
True Positive Rate
No Predictive Power
(Random Guessing)
AUC = 0.5
False Positive Rate