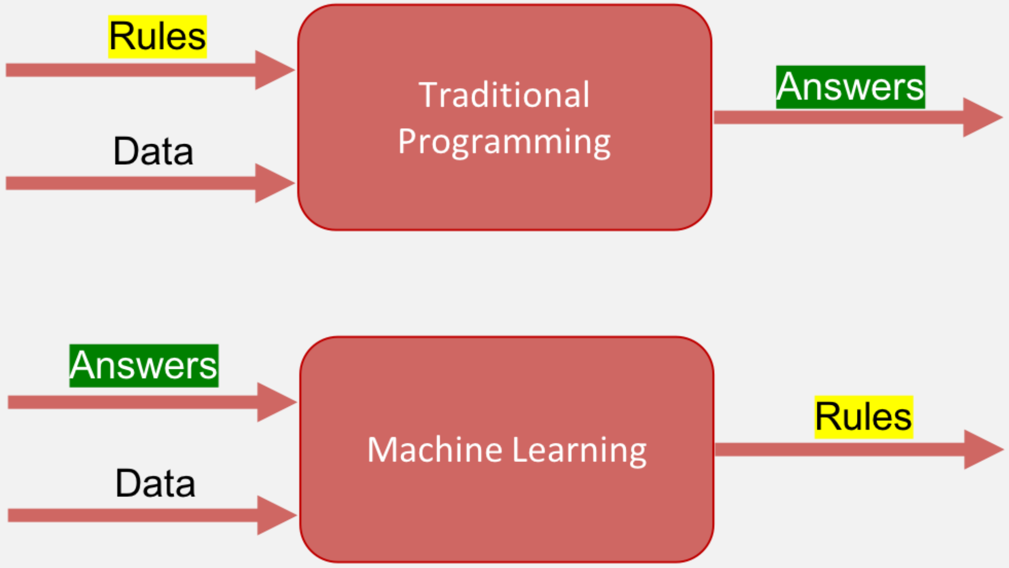
UNIT 4: MACHINE LEARNING

Dipesh Koirala

OUTLINE

- Introduction to Machine Learning, type of machine learning methods;
- Supervised vs Unsupervised learning
- Regression Techniques: Linear Regression
- Classification Techniques: Logistic Regression, KNN, Decision Tree, Naïve Bayes, SVMs
- Clustering Techniques: K Means, K Medoids; and their pros and cons.

MACHINE LEARNING

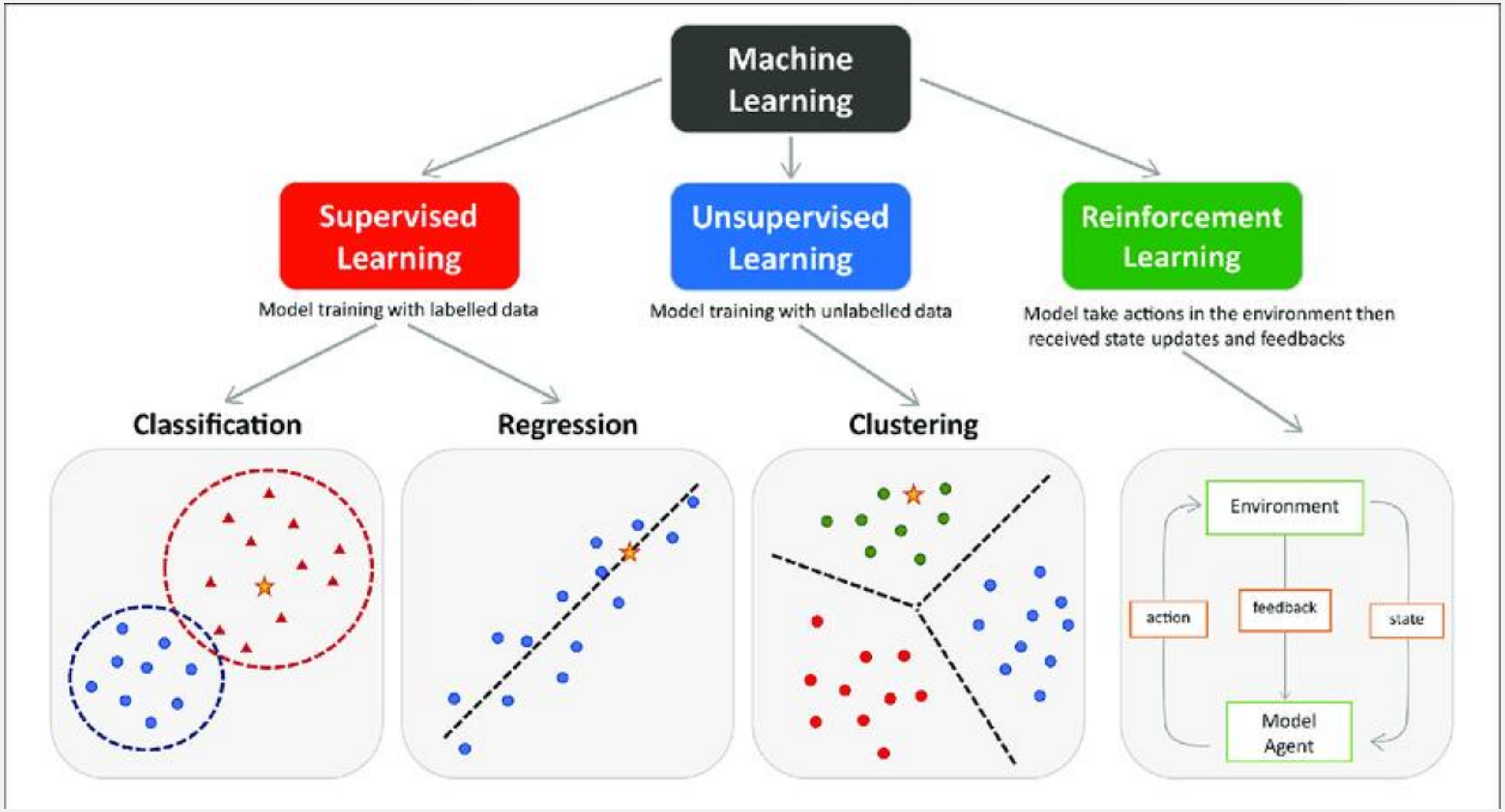


Machine learning is an application of AI that enables systems to learn and improve from experience without being explicitly programmed.

focuses on the use of data and algorithms to gradually improving its accuracy.

MACHINE LEARNING

- **Types:**

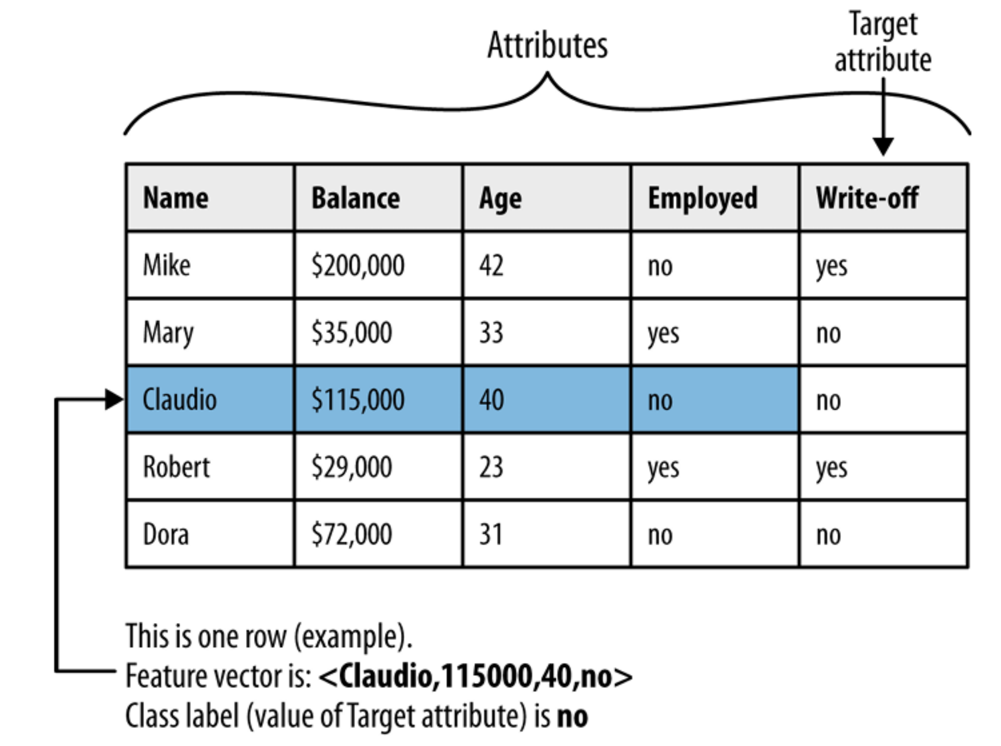


SUPERVISED LEARNING TECHNIQUE

- A type of learning that uses labelled data (or input with outputs) to train machine learning algorithms.

- The input are provided with corresponding outputs while training ML models.

- Since these algorithm needs external supervision on mapping input and expected output, they are called **supervised**.

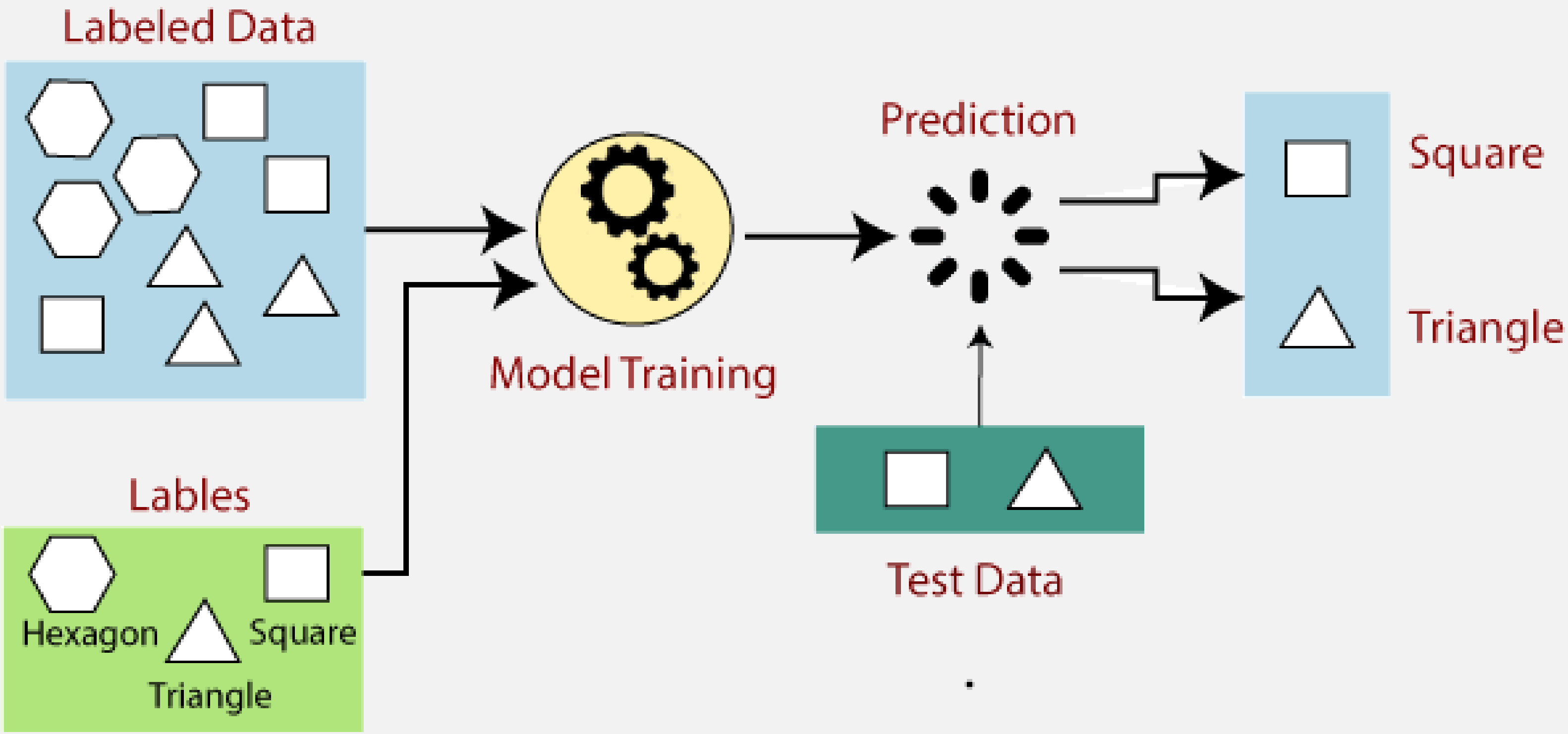


Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

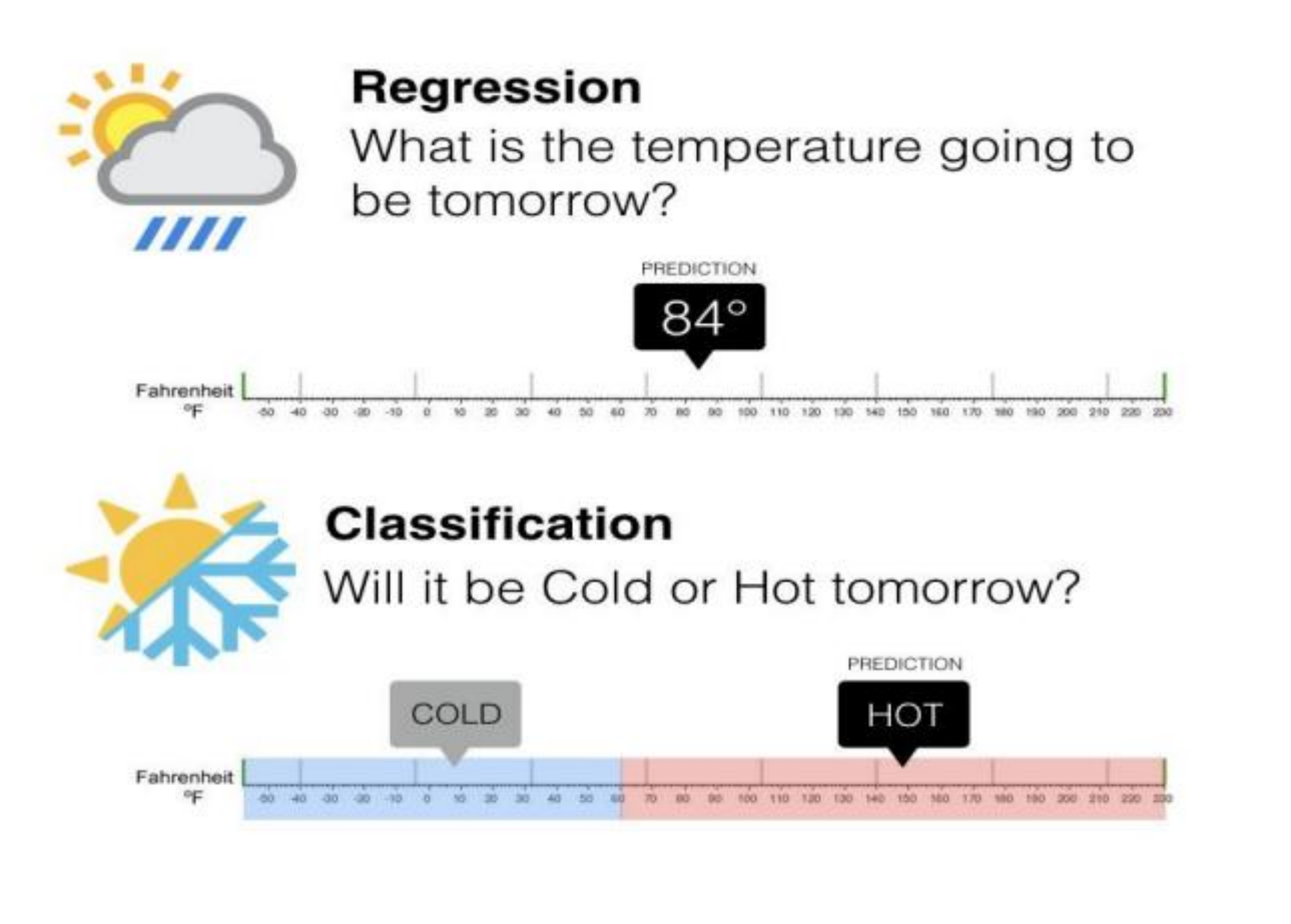
SUPERVISED LEARNING TECHNIQUE

- Understanding



SUPERVISED LEARNING TECHNIQUE

- Application



Regression

What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?

SUPERVISED LEARNING TECHNIQUE

Algorithms/ Techniques:

- Linear regression
- Logistic regression
- Naive Bayes
- Decision trees
- K-nearest neighbor algorithm
- Random forest algorithm
- Support Vector Machine etc.

UNSUPERVISED LEARNING

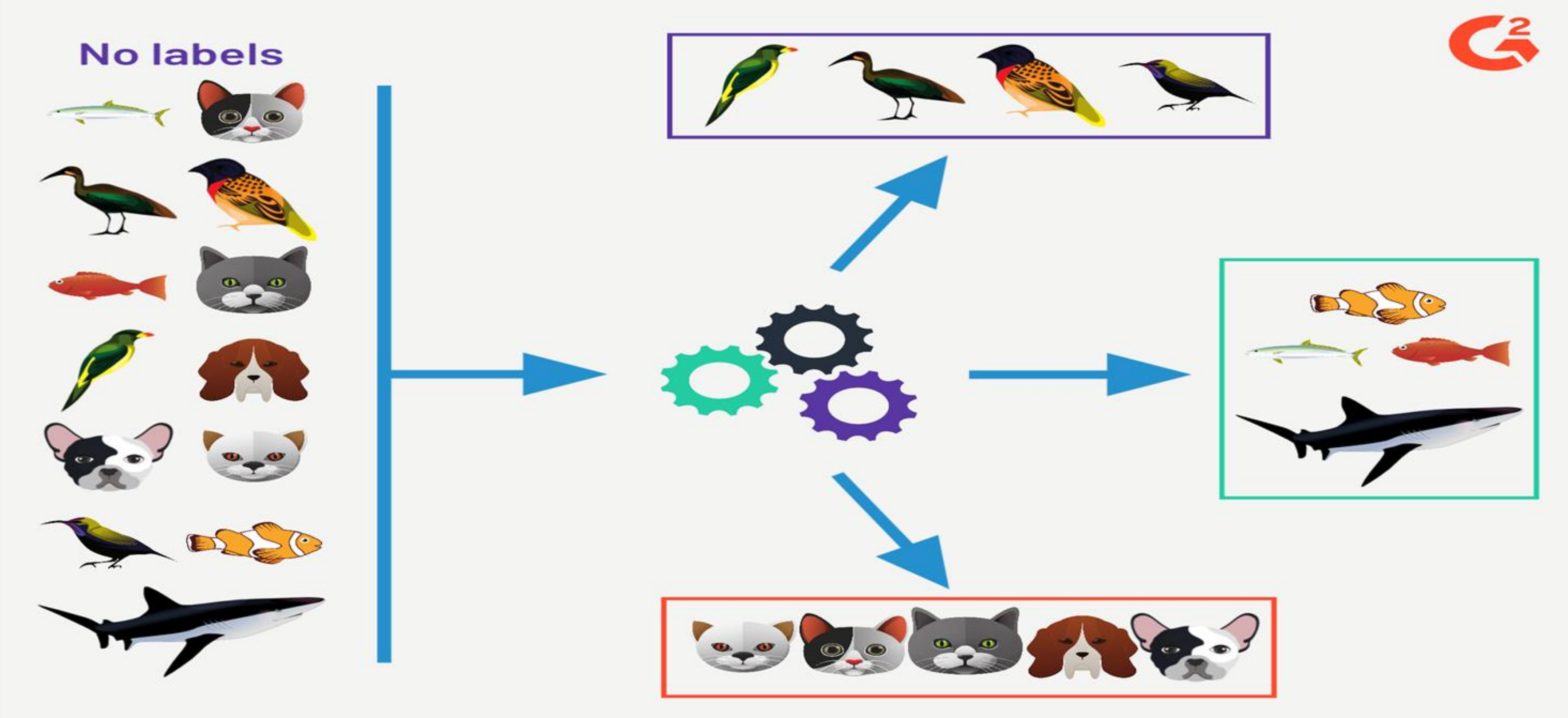
- A class of algorithm which has input data but *no corresponding output or label.*

- The goal for unsupervised learning is to model/ understand/ manipulate the underlying structure or pattern of the data in order to learn more about the data.

- There is *no supervision from expert through label or output*, thus, called unsupervised algorithm.

UNSUPERVISED LEARNING

- No Labels



UNSUPERVISED LEARNING

Algorithms/ Techniques:

- K-means clustering
- Hierarchical clustering
- Apriori algorithms
- Principal Component Analysis
- Autoencoders etc.

REGRESSION TECHNIQUES

- Regression analysis *is the process of curve fitting in which the relationship between the independent variables and dependent variables* which are modeled in the mth degree polynomial.

- Is done to *predict the value of dependent variable on the basis of independent variable.*

- Linear Regression

- **For example,** by looking at past customer purchase trends, regression analysis estimates future sales, so that more informed inventory purchases can be made.

LINEAR REGRESSION

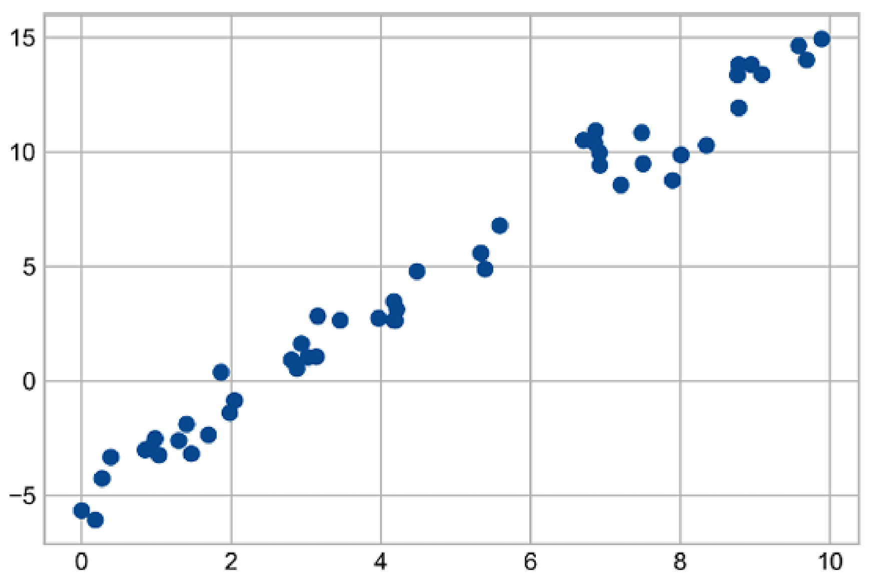
- **Linear Regression** is a machine learning technique used to predict a *dependent variable based on one or more independent variables,* assuming a linear relationship.

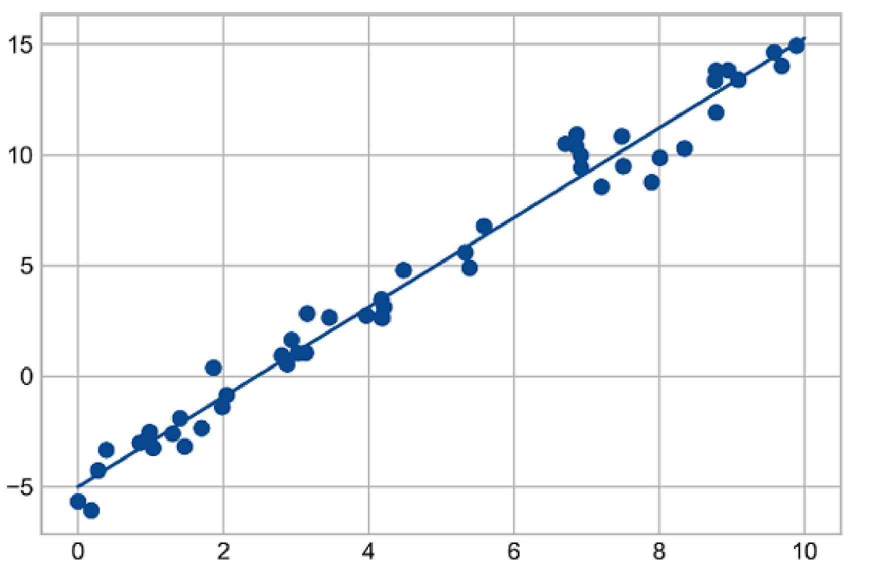
- In simple linear regression, is a statistical method used to model the relationship between two variables,
  - Independent Variable (X) : predictor or explanatory variable
  - Dependent Variable (Y) : the outcome or response variable
- fitting the data to a straight line, **often called as the regression line.**

$$Y = b_0 + b_1 * X + \varepsilon$$

LINEAR REGRESSION

- **Fits a regression line**





LINEAR REGRESSION

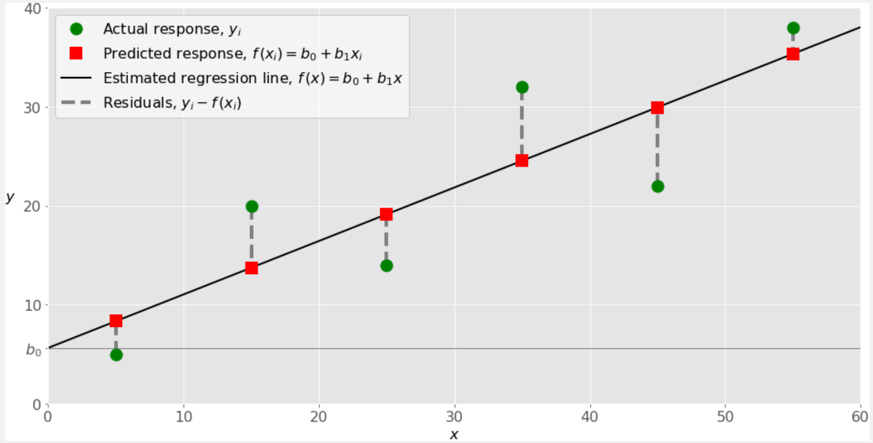
What is the Best Fit Line?

- is a line that minimizes the Sum of Squared Errors between observed and predicted values.

Random Error(Residuals)

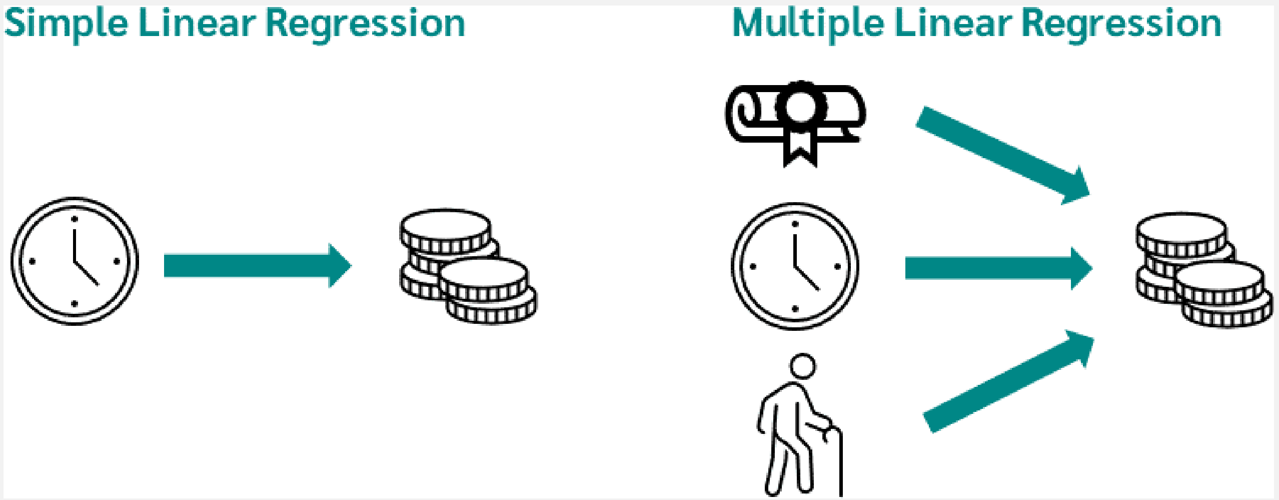
- The difference between the observed value of the dependent variable(**yi**) and the predicted value(**predicted**) is called the residuals.

$$\boldsymbol{\varepsilon\, i =}\ \ yi - ypred$$



LINEAR REGRESSION

Multiple linear regression



$Y = b_0 + b_1 * X_1 + b_2 * X_2 + \ldots\ldots.$

CLASSIFICATION TECHNIQUES

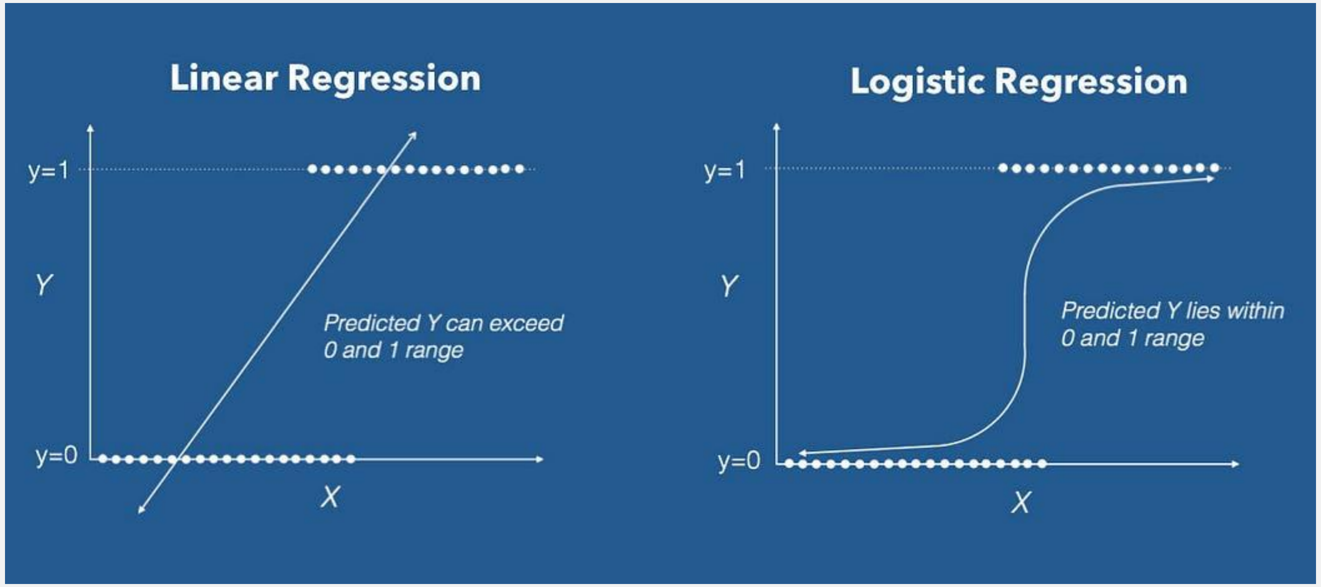
- In classification, *the goal is to predict a class label*, which is a choice from a predefined list of possibilities.

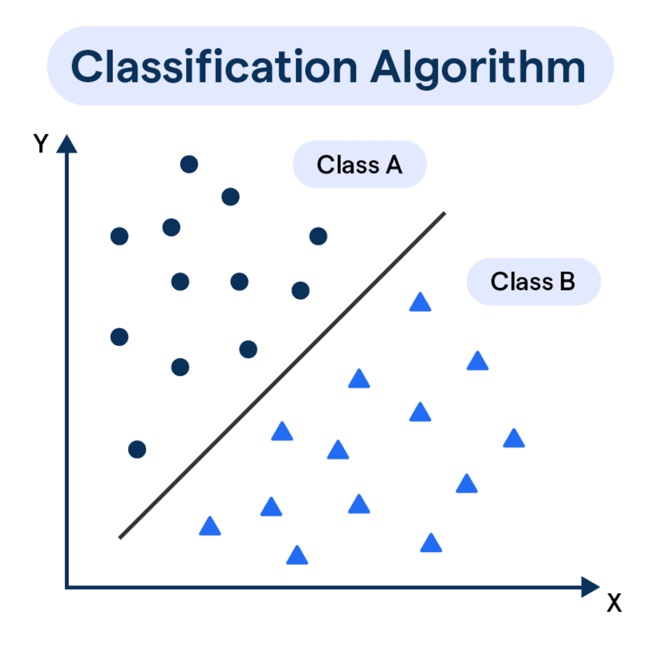
Algorithms:

- Logistic Regression
- K – Nearest Neighbors
- Decision Trees
- Naïve Bayes

LOGISTIC REGRESSION

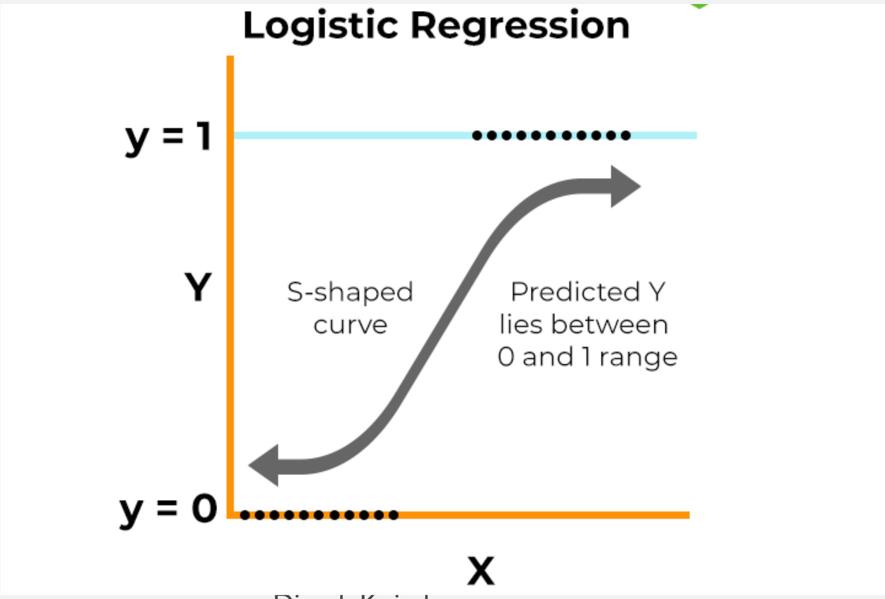
- Logistic Regression is a Machine Learning algorithm *which is used for the classification problems,* it is a predictive analysis algorithm and based on the concept of probability.

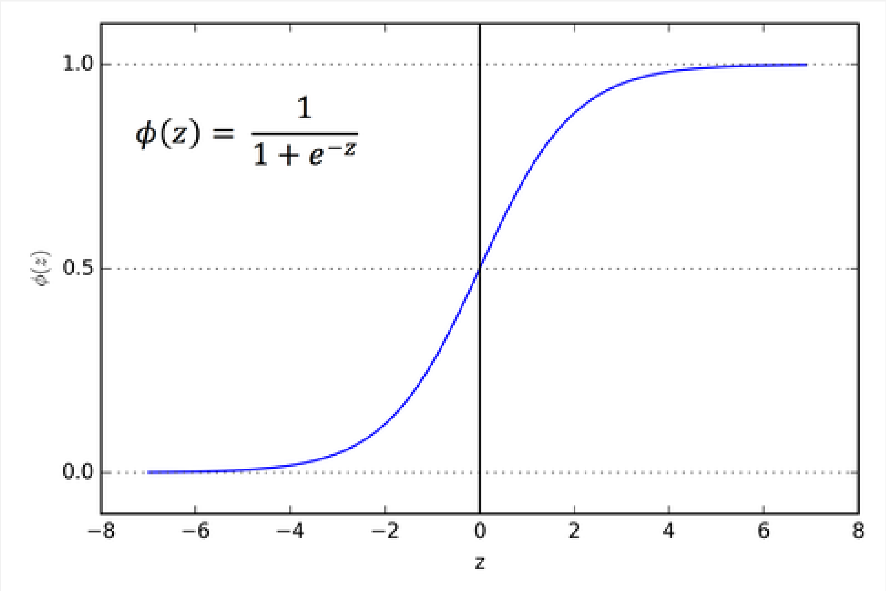




LOGISTIC REGRESSION

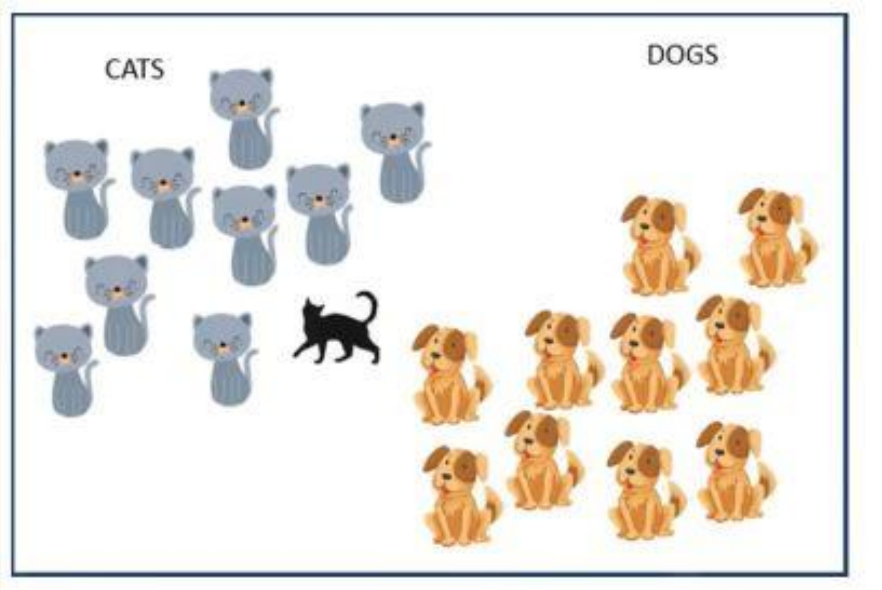
- Logistic regression is a supervised machine learning algorithm that performs binary classification tasks by predicting the probability of an outcome.
- The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

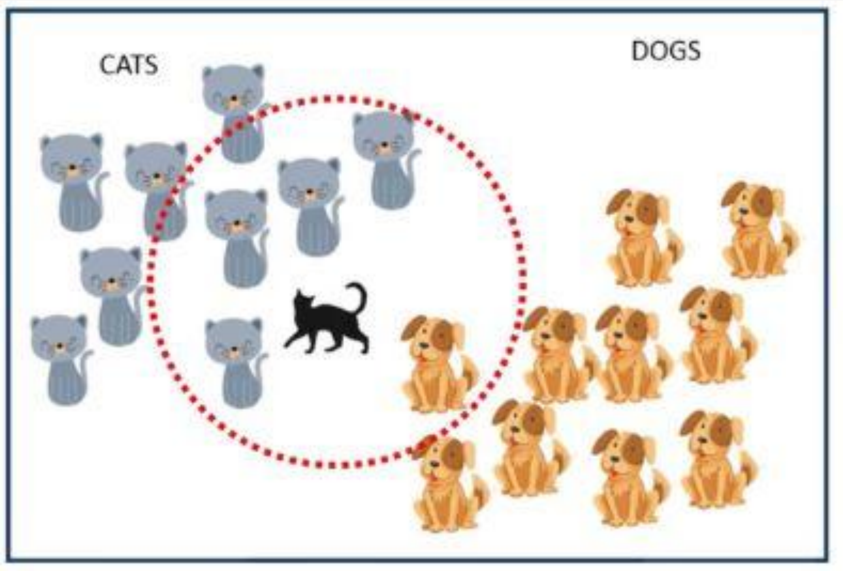




K-NEAREST NEIGHBORS

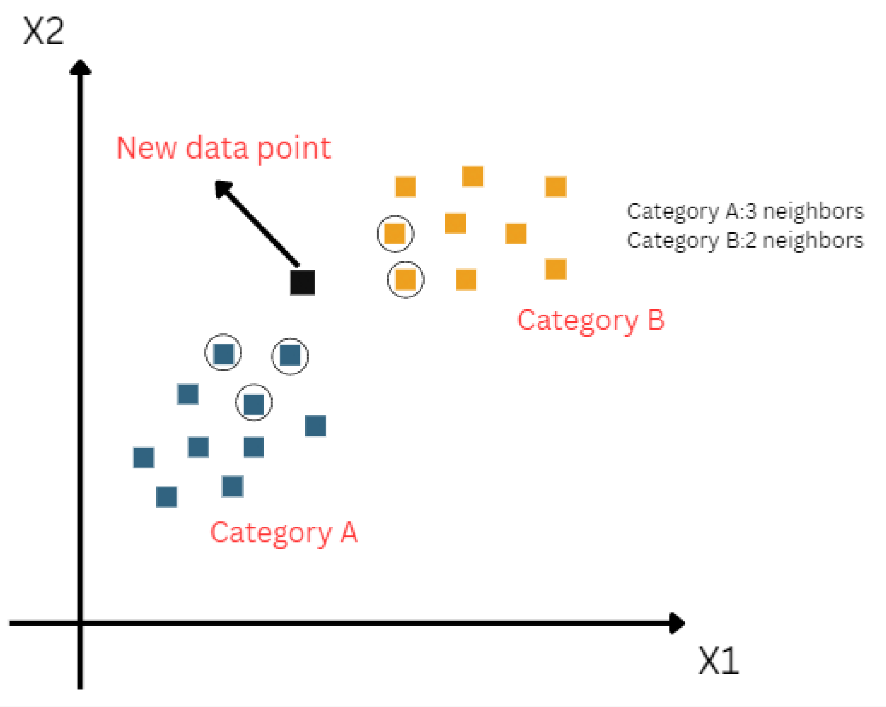
- K-Nearest Neighbors is one of the simplest supervised machine learning algorithms used for classification. It classifies a data point *based on its neighbors' classifications.*
- Can be used for both regression and classification, but it is more commonly utilized for classification tasks.

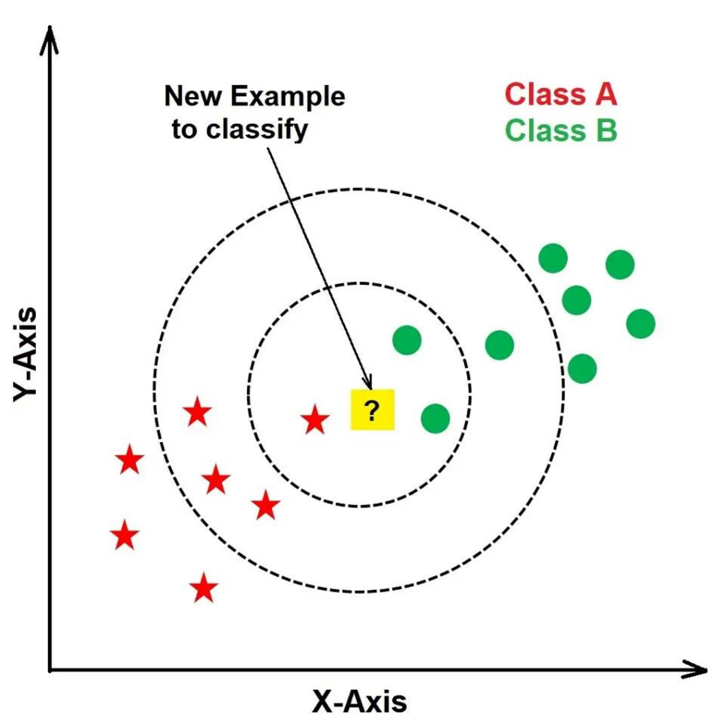




K-NEAREST NEIGHBORS

- Assumes that the *new case/data and existing cases are similar and places the new case in the category that is most similar to the existing categories.*
- For classification problems, a class label is assigned on the basis of a majority vote





K-NEAREST NEIGHBORS

Algorithm:

1. Choose the number of neighbors 'k'.
2. *Calculate the distance between the new data point* and all training data points.
3. Identify the 'k' nearest neighbors based on the calculated distances.
4. Count the number of data points in each category (for classification) or average the values (for regression).
5. Assign the new data point to the category with the most neighbors (for classification) or predict the value (for regression).

Note: It's also known as a lazy learner algorithm since it doesn't learn from the training set right away

K-NEAREST NEIGHBORS

- Similarity is calculated by using *Minkowski distance*, which is a generalization of both Euclidean distance and Manhattan distance. It is defined as:

$$d(x,y) = \left(\left|x_2 - x_1\right|^p + \left|y_2 - y_1\right|^p\right)^{1/p}$$

- It represents the Manhattan distance when p = 1 (i.e., L_1 norm) and Euclidean distance when p = 2 (i.e., L_2 norm).

- **E.g., :** Class A : (2, 2), (2, 3), (3, 2)

  Class B : (4, 5), (5, 3), (6, 2)

Check for: (4, 3)

K-NEAREST NEIGHBORS

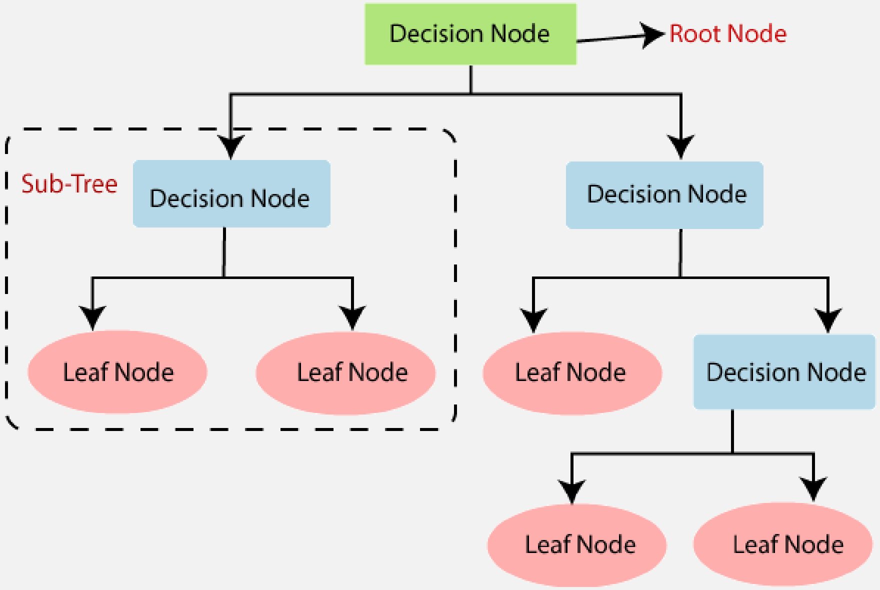
- Use k = 5 and classify

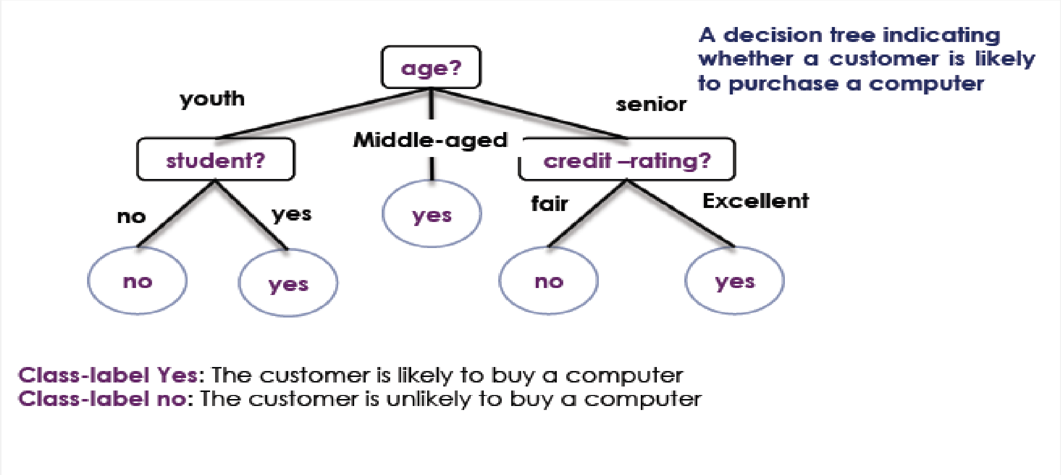
Sepal Length	Sepal Width	Species (Class)
5.3	3.7	Setosa
5.1	3.8	Setosa
5.4	3.4	Setosa
7.2	3.0	Verginica
7.4	2.8	Verginica
5.8	2.7	Verginica
6.1	2.8	Versicolor
6.3	2.3	Versicolor
5.5	2.4	Versicolor
5.2	3.1	?

Sepal Length	Sepal Width	Species (Class)	Distance
5.3	3.7	Setosa	0.61
5.1	3.8	Setosa	0.71
5.4	3.4	Setosa	0.36
7.2	3.0	Verginica	2.00
7.4	2.8	Verginica	2.22
5.8	2.7	Verginica	0.72
6.1	2.8	Versicolor	0.95
6.3	2.3	Versicolor	1.36
5.5	2.4	Versicolor	0.76

DECISION TREE

- Decision tree is *a flowchart-like tree structure for classification task.*
- internal nodes (non leaf node) denotes a test on an attribute or feature, branches represent outcomes of tests, and Leaf nodes (terminal nodes) hold class labels.



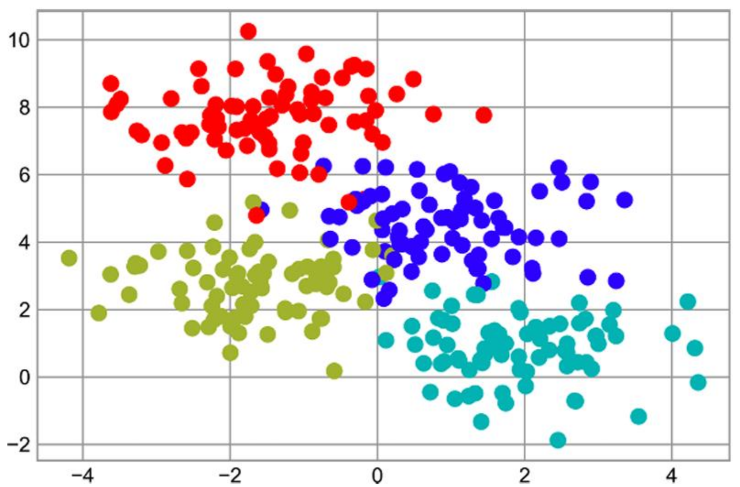


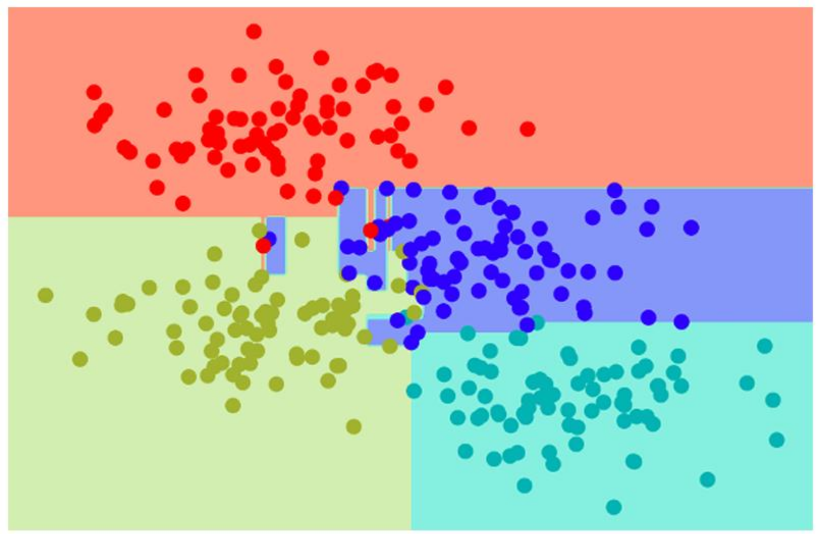
A decision tree indicating whether a customer is likely to purchase a computer

Class-label Yes: The customer is likely to buy a computer
Class-label no: The customer is unlikely to buy a computer

DECISION TREE

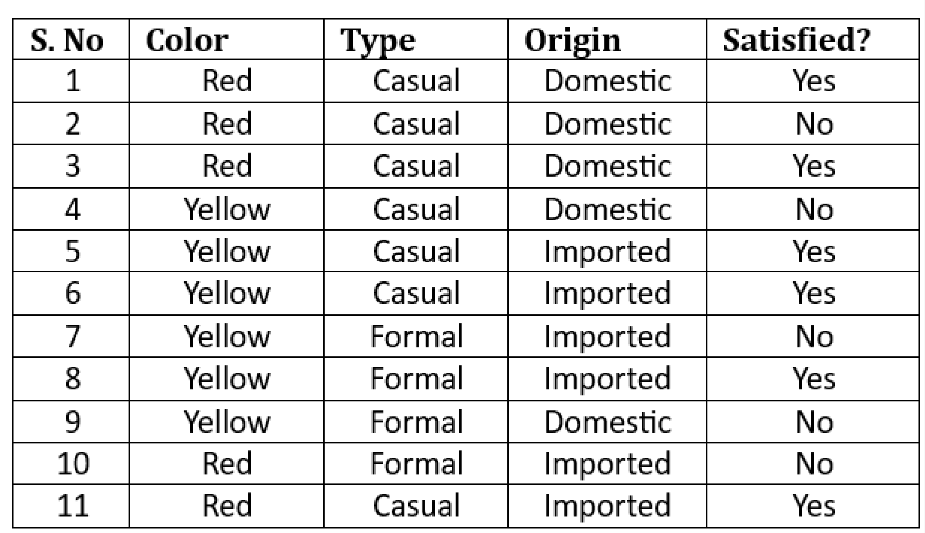
- Example





DECISION TREES

- The following table represents a dataset of 10 objects with attributes Color, Type, Origin and the “class”, whether the customer who bought was satisfied or not.



S. No	Color	Type	Origin	Satisfied?
1	Red	Casual	Domestic	Yes
2	Red	Casual	Domestic	No
3	Red	Casual	Domestic	Yes
4	Yellow	Casual	Domestic	No
5	Yellow	Casual	Imported	Yes
6	Yellow	Casual	Imported	Yes
7	Yellow	Formal	Imported	No
8	Yellow	Formal	Imported	Yes
9	Yellow	Formal	Domestic	No
10	Red	Formal	Imported	No
11	Red	Casual	Imported	Yes

Now classify a new object with the following properties:

- Color = Red, Origin = Domestic and Type = Formal

DECISION TREES

- ID3 stands for Iterative Dichotomiser 3. It uses top-down greedy approach to build decision tree model.

- This algorithm *computes information gain for each attribute* and then selects the attribute with the highest information gain.

- **Information gain** measures reduction in entropy after data transformation. *It is calculated by comparing entropy of the dataset before and after transformation.*

DECISION TREES

- Entropy is the measure of impurity or uncertainty of the dataset

$$E(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \qquad (1)$$

- Consider *attribute A having v distinct values*. The attribute A can be used to split D n partitions $\{D_1, D_2, .., D_n\}$. Now, the total entropy of data partitions while partitioning D around attribute A is calculated as:

$$E_A(D) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} \times E(D_i)$$

- **Information Gain:**

$$IG(A) = E(D) - E_A(D)$$

DECISION TREES

- ID3 - Entropy
- CART – Gini Index

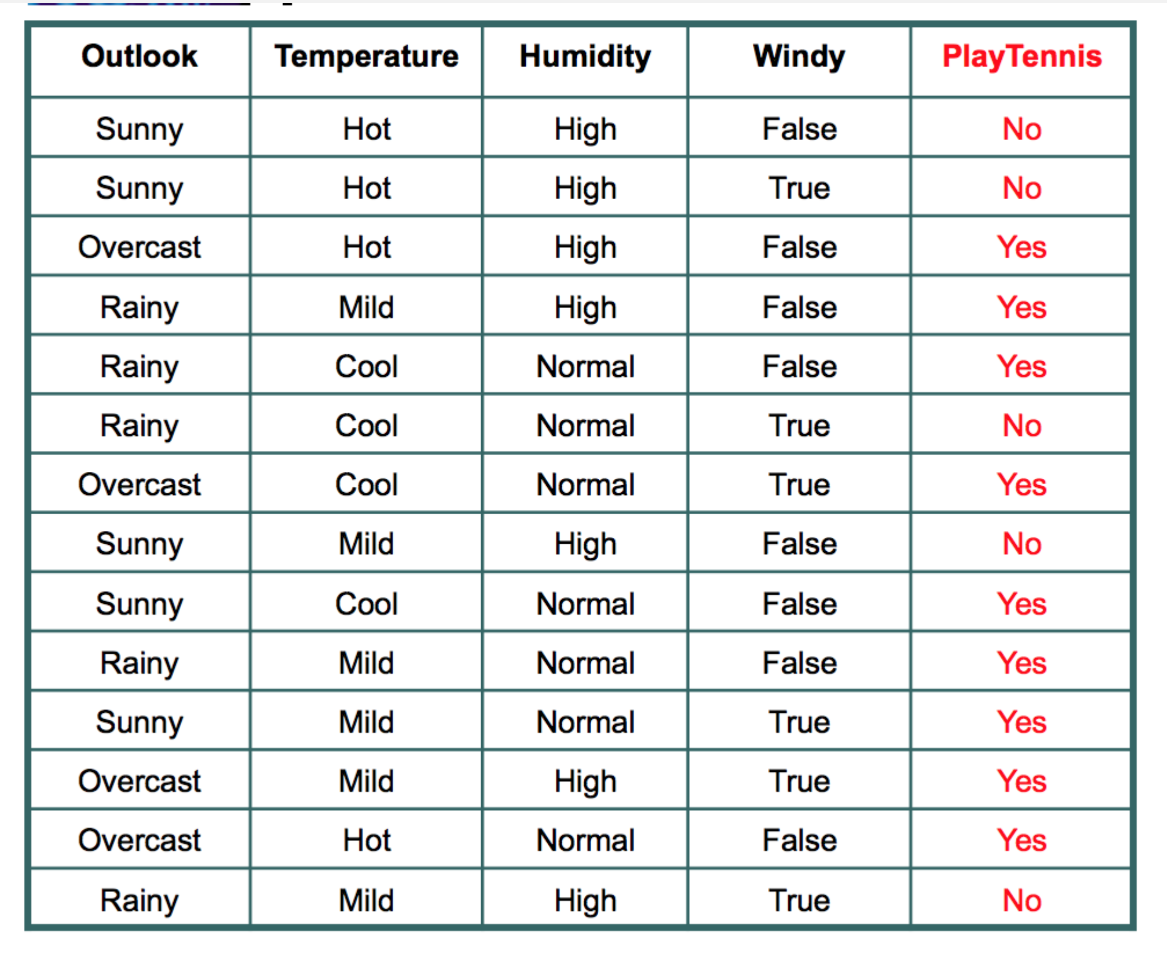
$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

- **IN Conclusion:**

1. Choose the best feature
2. Split the dataset
3. Repeat
4. Stop when the leaf nodes are pure

DECISION TREES

- Among Outlook, Temperature, Humidity and Windy, find one that has highest IG.



Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

NAIVE BAYES

- Nave Bayes classification is **based on Bayes' theorem**. Bayes' theorem is given by:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Let D be a database and $C_l, C_2\ldots\ldots C_m$ are m classes. Now above Bayes rule can be written as:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

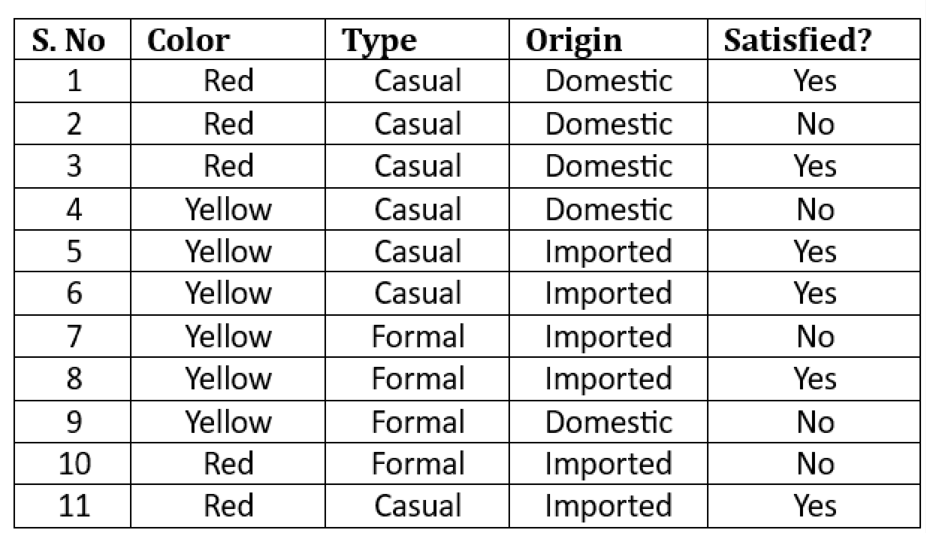
NAIVE BAYES

- Let X is the set of attributes $\{x_1, x_2, x_3 \ldots\ldots x_n\}$ where attributes are independent of one another. Now the probability P(X|C$_i$) is given by the equation given below:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \ldots\ldots \times P(x_n|C_i)$$

NAÏVE BAYES

- The following table represents a dataset of 10 objects with attributes Color, Type, Origin and the "class", whether the customer who bought was satisfied or not.



S. No	Color	Type	Origin	Satisfied?
1	Red	Casual	Domestic	Yes
2	Red	Casual	Domestic	No
3	Red	Casual	Domestic	Yes
4	Yellow	Casual	Domestic	No
5	Yellow	Casual	Imported	Yes
6	Yellow	Casual	Imported	Yes
7	Yellow	Formal	Imported	No
8	Yellow	Formal	Imported	Yes
9	Yellow	Formal	Domestic	No
10	Red	Formal	Imported	No
11	Red	Casual	Imported	Yes

Now classify a new object with the following properties:

- Color = Red, Origin = Domestic and Type = Formal

NAÏVE BAYES

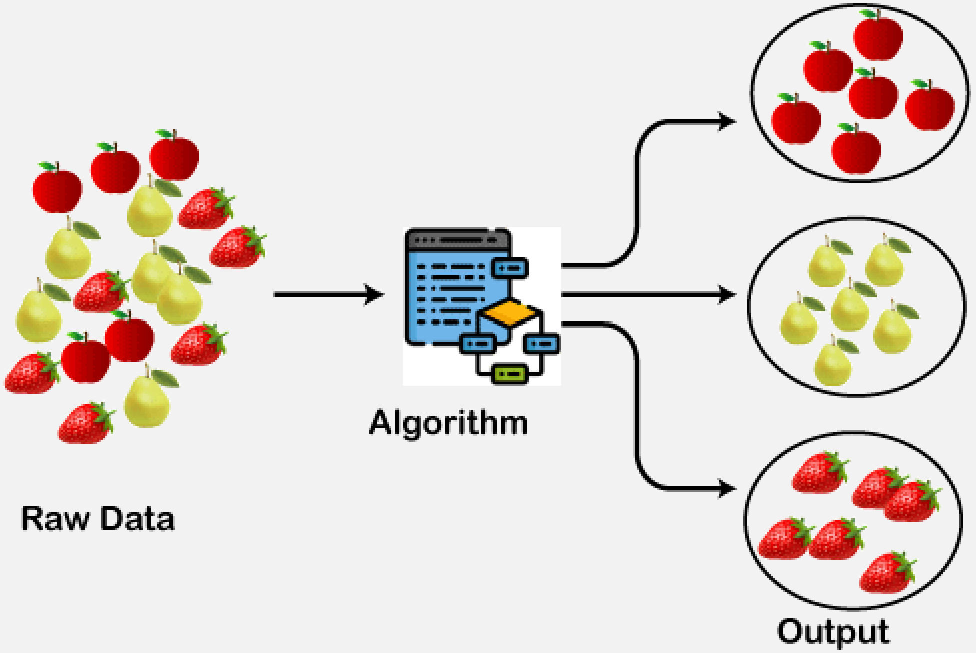
- Understanding

CLUSTERING TECHNIQUES

- The process of grouping *a set of objects into classes of similar objects is called clustering.*

- It is an unsupervised learning technique.

- A cluster is a collection of data objects that are *similar to one another* within the same cluster and are *dissimilar* to the objects in other clusters.



K-MEANS ALGORITHM

- K-means is one of the simplest partitioning based clustering algorithm.
- The procedure follows a simple and easy way *to classify a given data set into a certain number of clusters (assume k clusters) fixed Apriori.*
- The main idea **is to define k centers**, one for each cluster.
- These centers should be selected cleverly because of different location causes different result.
- So, the better choice is to place them as much as possible far away from each other

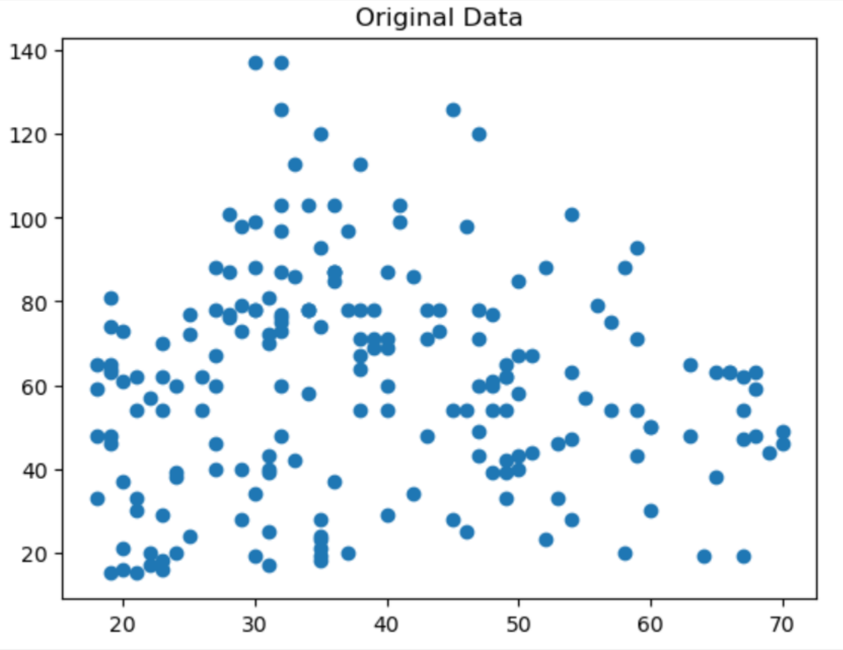
K-MEANS ALGORITHM

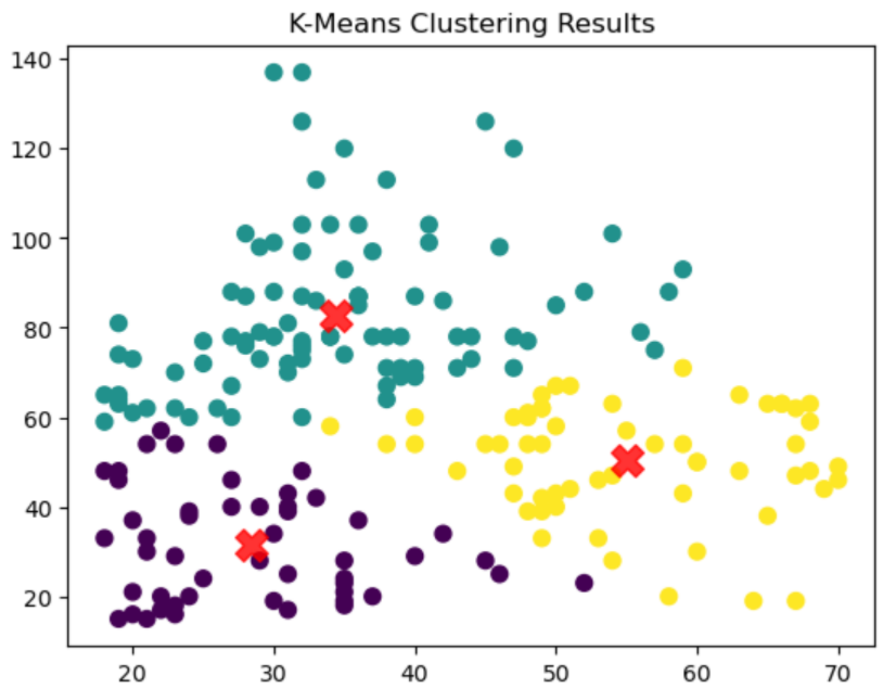
- Let $X = \{x_1, x_2, x_3, \ldots\ldots.., x_n\}$ be the set of data points.

1. Randomly select *k* cluster centers, $C = \{c_1, c_2, \ldots\ldots., c_k\}$
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. If
   - No data is reassigned then terminate
5. Else
   - Recalculate the new cluster center using centroid.
   - Recalculate the distance between each data point and new cluster centers.
   - Go to step 3

K-MEANS ALGORITHM

- K-Means Clustering





K-MEANS ALGORITHM

E.g.: Divide the data points {(2,10), ((2,5), (8,4), (5,8), (7,5), (6,4)} into two clusters.

Solution

Let p1=(2,10)   p2=(2,5)   p3=(8,4)   p4=(5,8)   p5=(7,5)   p6=(6,4)

Initial step

Choose Cluster centers randomly

Let c1=(2, 5) and c2=(6, 4) are two initial cluster centers.

K-MEANS ALGORITHM

Iteration 1: Calculate distance between clusters centers and each data points

d(c1, p1) = 5	d(c2, p1) = 7.21
d(c1, p2) = 0	d(c2, p2) = 4.12
d(c1, p3) = 6.08	d(c2, p3) = 2
d(c1, p4) = 4.24	d(c2, p4) = 4.12
d(c1, p5) = 5	d(c2, p5) = 1.41
d(c1, p6) = 4.12	d(c2, p6) = 0

Thus, Cluster1= {p1, p2} cluster2 = {p3, p4, p5, p6}

K-MEANS ALGORITHM

Iteration 2: *New Cluster centers: c1=(2, 7.5)   c2=(6.5, 5.25)*

Again, Calculate distance between clusters centers and each data points

d(c1, p1) = 2.5	d(c2, p1) = 6.54
d(c1, p2) = 2.5	d(c2, p2) = 4.51
d(c1, p3) = 6.95	d(c2, p3) = 1.95
d(c1, p4) = 3.04	d(c2, p4) = 3.13
d(c1, p5) = 4.59	d(c2, p5) = 0.56
d(c1, p6) = 5.32	d(c2, p6) = 1.35

Thus, Cluster1= {p1, p2, p4}   cluster2 = {p3, p5, p6}

MODEL EVALUATION

- **Model evaluation** is a method of *assessing the correctness of models on test data.*
- The test data consists of data points that have not been seen by the model before.

Regression Metrics

- Root Mean Squared Error, Mean Absolute Error, Mean Percentage Error

Classification Metrics

- Confusion Matrix, Accuracy, Precision, Recall

MODEL EVALUATION

Mean Squared Error (MSE)

- MSE is a metric that calculates the difference between the actual value and the predicted value (error), squares it and then *provides the mean of all the errors.*

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

where y is actual value and $\hat{y}_i$ is predicted value

if the model predicts a house to be 30,000 and it's 32,000, *the squared error is square of $(32{,}000-30{,}000)^2 = 2000^2$*

MODEL EVALUATION

- MSE is very sensitive to outliers and will show a very high error value even if a few outliers are present in the otherwise well-fitted model predictions.

Root Mean Squared Error (RMSE)

- RMSE is the root of MSE and is beneficial because it *helps to bring down the scale of the errors closer to the actual values*, making it more interpretable.

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

MODEL EVALUATION

Mean Absolute Error or MAE

- MAE is the mean of the absolute error values (actuals – predictions).

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

- If one *wants to ignore the outlier values to a certain degree*, MAE is the choice since it reduces the penalty of the outliers significantly with the removal of the square terms.

MODEL EVALUATION

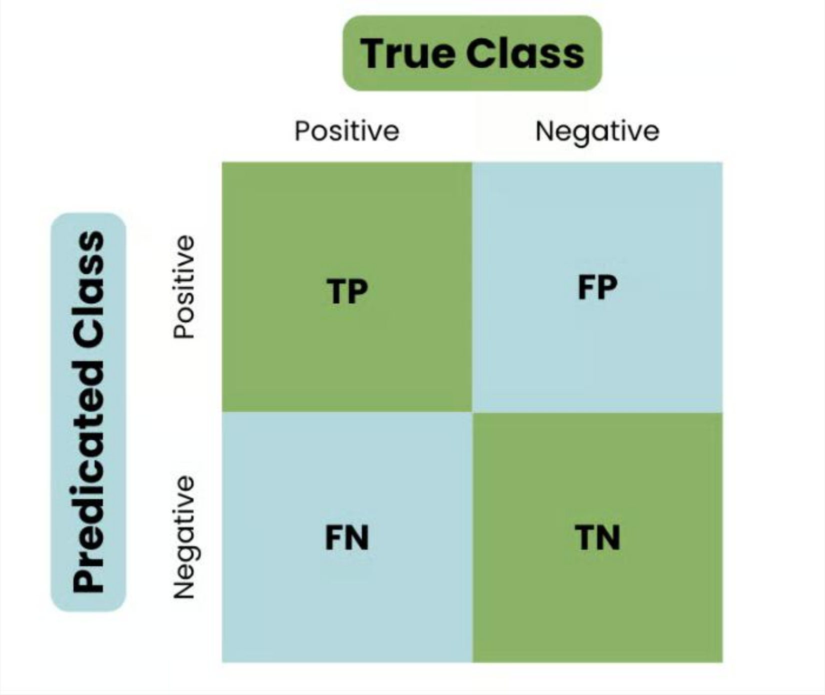
Mean Absolute Percentage Error

- *Expresses the error as a percentage of the actual values*, providing an easy-to-understand metric.
- **For e.g.,** if a house is worth 200,000 and predicted price is180,000, the error is 10%.
- This percentage-based approach makes MAPE very interpretable, especially when explaining model performance to stakeholders who might not be technical.

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

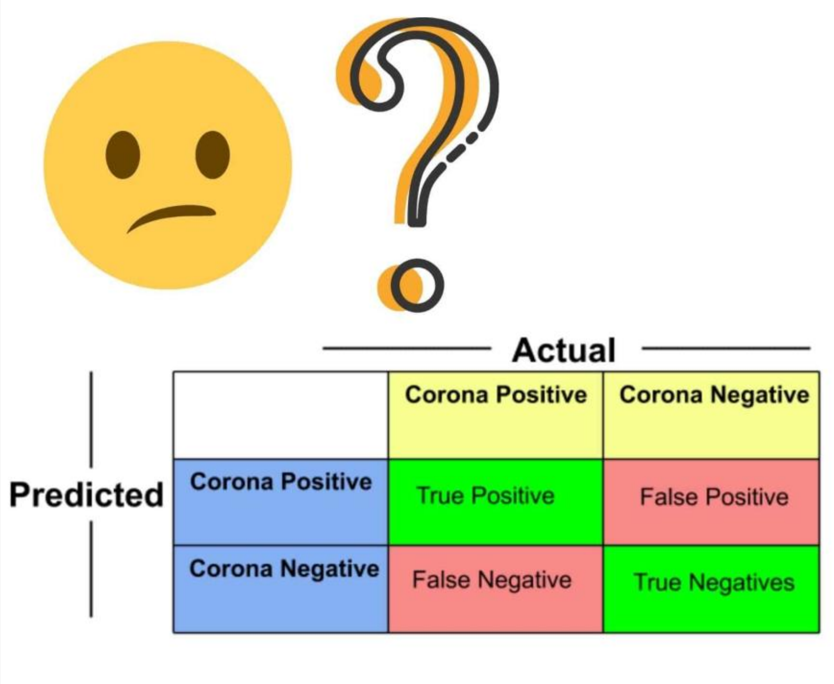
MODEL EVALUATION

- **Confusion matrix** is an N x N matrix used *for evaluating the performance of a classification model,* where N is the number of target classes.
- The matrix compares the actual target values with those predicted by the machine learning model.



True Class

Predicated Class		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN



Actual

		Corona Positive	Corona Negative
Predicted	Corona Positive	True Positive	False Positive
	Corona Negative	False Negative	True Negatives

MODEL EVALUATION

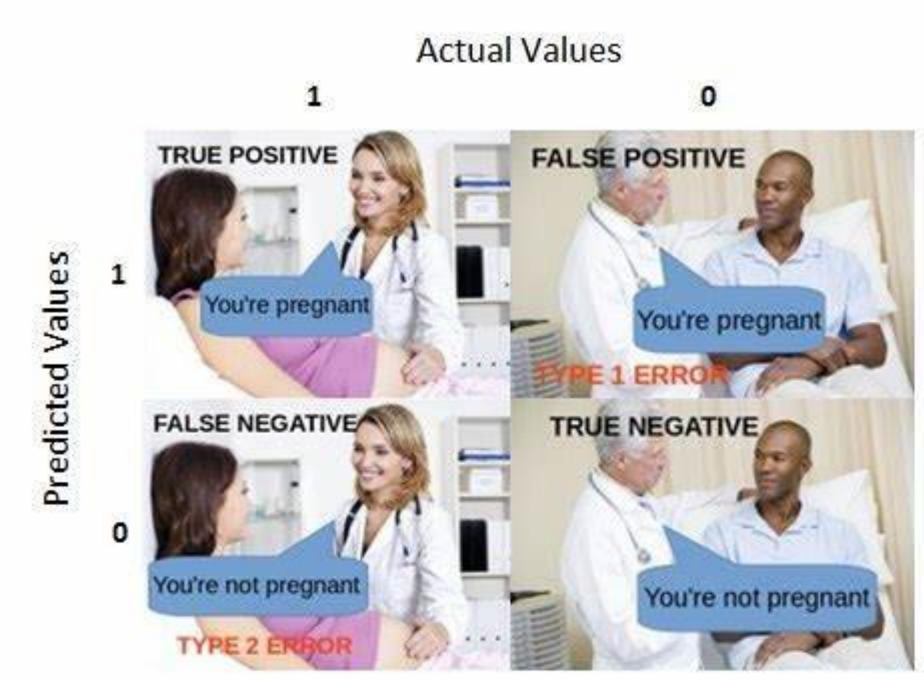
Confusion matrix

True Positive(TP): It represents correctly classified positive classes.

False Positive (FP): It represents incorrectly classified positive classes.

False Negative (FN): It represents incorrectly classified negative classes.

True Negative(TN): It represents correctly classified Negative classes.



MODEL EVALUATION

- **Accuracy:** It is the percentage of correct predictions made by the model and is given as below:

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

- **Precision:** measures the proportion of correctly predicted positive observations out of all the predictions made for the positive class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{Total Number of Predicted Positives}}$$

- **Recall:** measures the proportion of correctly predicted positive observations out of all the actual positive samples.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{Total Number of actual positives}}$$

MODEL EVALUATION

E.g.,

- Suppose that we have to classify 100 people as pregnant or not pregnant. This includes 40 pregnant women and the remaining 60 are not pregnant. Out of 40 pregnant women 30 pregnant women are classified correctly and the remaining 10 pregnant women are classified as not pregnant by the machine learning algorithm. On the other hand, out of 60 people in the not pregnant category, 55 are classified as not pregnant and the remaining 5 are classified as pregnant.
- Compute accuracy, precision, recall and F1-score for the above example.

END OF UNIT 4

Thank You

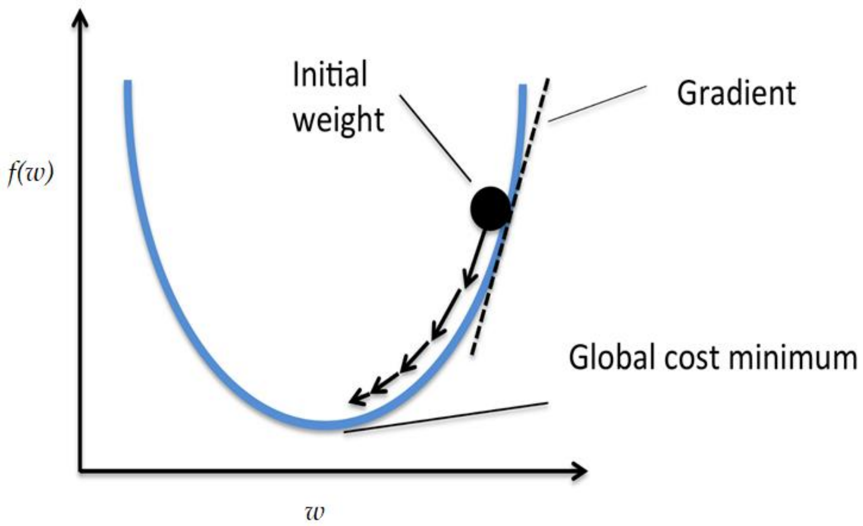
EXTRA MATERIALS

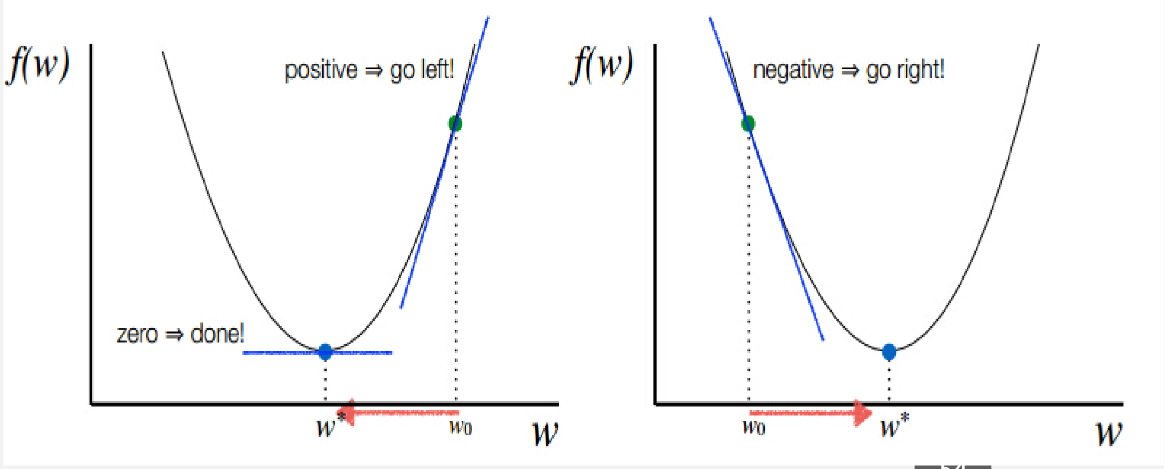
- Gradient Descent

GRADIENT DESCENT

- **Iterative approach for finding the minimum of a function**
- **Gradient descent** is an *optimization algorithm used to minimize some convex function* by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.

- In machine learning, gradient descent *is used to update the parameters or weights* of the model.





GRADIENT DESCENT

- Said it more mathematically, *a gradient is a partial derivative with respect to its inputs.*

- The higher the gradient, *the steeper the slope and the faster a model can learn*. But if the slope is zero, the model stops learning.

- How big the steps are that Gradient Descent takes into the direction of the local minimum are determined by the learning rate.

LINEAR REGRESSION

$y = w_0 + w_1 x$

- Let us suppose that $\{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$ are given data points. Loss function for the n data points is given by:

$$L = \frac{1}{2n} \sum_{i=1}^{n} e_i^2$$

$$L = \frac{1}{2n} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2$$

LINEAR REGRESSION

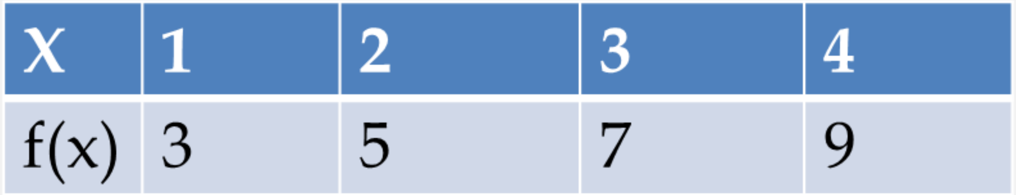
- Now, coefficients or weights can be determined or updated using gradient decent method as below:

$$\boldsymbol{w_0} = w_0 - \alpha \frac{\partial E}{\partial w_0} = w_0 + \alpha \frac{1}{n} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)$$

$$\boldsymbol{w_1} = w_1 - \alpha \frac{\partial E}{\partial w_1} = w_1 + \alpha \frac{1}{n} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) x_i$$

LINEAR REGRESSION

- ***E.g.,*** Fit a straight line through the following data using SGD. Show one epoch of training.



X	1	2	3	4
f(x)	3	5	7	9

Solution

General form of linear regression equation is: $y = w_0 + w_1 x$

Let us assume that initial values of parameters are:

$$w_0 = w_1 = 0$$

PRECISION AND RECALL

- Accuracy may not be a good measure if the dataset is not balanced.
- Consider the following scenario: There are 90 people who are healthy (positive) and 10 people who have some disease (negative).
- In this example, ***TP*** = 90, ***FP*** = 10, ***FN*** = 0 and ***TN*** = 0.

LINEAR REGRESSION

<u>Iteration 1</u>: *x*=1, y=f(*x*)=3 $\alpha = 0.01$

$$w_0 = w_0 + \alpha(y - w_0 - w_1 x) = 0 + 0.01 \times 3 = 0.03$$

$$w_1 = w_1 + \alpha(y - w_0 - w_1 x)x = 0 + 0.01 \times 3 = 0.03$$

<u>Iteration 2</u>: *x*=2, y=f(*x*)=5

$$w_0 = w_0 + \alpha(y - w_0 - w_1 x) = ?$$

$$w_1 = w_1 + \alpha(y - w_0 - w_1 x)x = ?$$

In the same way perform iteration 3 and 4.

PRECISION AND RECALL

- Precision answers: "*Of all the samples predicted as positive, how many are actually positive?*“
- High precision indicates that the classifier makes fewer false positive errors.
- Useful in scenarios where false positives are costly (e.g., spam email detection — we don’t want non-spam emails classified as spam).

- Recall answers: "*Of all the actual positive samples, how many did the classifier identify correctly?*“
- High recall indicates that the classifier captures most of the positive cases.
- Useful in scenarios *where false negatives are costly* (e.g., medical diagnosis — missing a disease is more serious than falsely diagnosing it).

PRECISION AND RECALL

- Machine Learning
- Feature Engineering
- Predictive Data Analysis (Time Series Data Analytics)
- Introduction to Deep Learning, backpropagation

THANK YOU