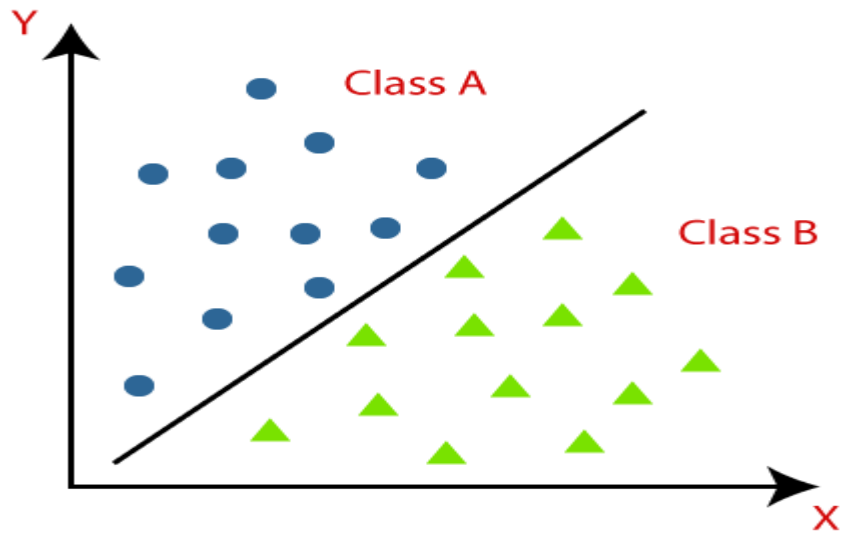
Classification Algorithms

- The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.
- Unlike regression, the output variable of Classification is categorical, not a value, such as Yes-No, Male-Female, True-false etc.



Classification Algorithms

- There are two types of Classifications:

➢ **Binary Classifier:** If the classification problem has only two possible outcomes, then it is called as Binary Classifier. Examples: YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.

➢ **Multi-class Classifier:** If a classification problem has more than two outcomes, then it is called as Multi-class Classifier. Example: Classifications of types of fruits, Classification of types of music.

- **Types of ML Classification Algorithms:**

  Classification Algorithms can be further divided into the Mainly two category:

➢ **Linear Models**
  - Logistic Regression
  - Support Vector Machines

➢ **Non-linear Models**
  - K-Nearest Neighbours
  - Kernel SVM
  - Naïve Baye's
  - Decision Tree Classification
  - Random Forest Classification

Evaluation of Classification Model

- **Confusion Matrix:**

	Actual Positive	Actual negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

Suppose a ML Algorithm identifies 8 dogs in a picture containing 10 cats and 12 dogs. Of the 8 identified as dogs, 5 actually are dogs (true positives), while the other 3 are cats (false positives). 7 dogs were missed (false negatives), and 7 cats were correctly excluded (true negatives).

- **Accuracy:** In classification, the commonly used metric is accuracy which is defined as: (TP + TN) / (TP + FP + FN +TN)

The algorithm's Accuracy is: (5 + 7) / (5 + 3 + 7 + 7) = 12 / 22 = .5454

Evaluation of Classification Model

- **Precision:** Precision is defined as: TP / (TP + FP).

  This fraction shows the ratio of the true positive prediction among all positive predictions.

  The algorithm's Precision is: 5 / (5 + 3) = 5 / 8 = .625

- **Recall/Sensitivity/True Positive Rate:** In classification, recall or true positive rate shows how many of the positives returned by the ML Algorithm. It is defined as:

  $$\text{Recall(TPR)} = TP / (TP + FN)$$

  The algorithm's Recall is: 5 / (5 + 7) = 5 / 12 = .4166

- **F1-Score:** Precision and recall are often combined into a single measure using their harmonic mean, known as the F1-score.

  $$\text{F1-Score} = 2.\text{recall}.\text{precision} / (\text{recall} + \text{precision})$$

  $$= 2TP / 2TP + FP + FN$$

  The algorithm's f1-Score is: 2*5 / 2*5 + 3 + 7 = 10 / 20 = .5

Evaluation of Classification Model

➢ **AUC-ROC Curve:**

- ROC curve stands for **Receiver Operating Characteristics Curve** and AUC stands for **Area Under the Curve**.
- It is a graph that shows the performance of the classification model at different thresholds.
- The ROC curve is plotted with TPR and FPR, where TPR (True Positive Rate) on Y-axis and FPR(False Positive Rate) on X-axis.
- **Specificity/True Negative Rate:** It tells us what proportion of the negative class got correctly classified.

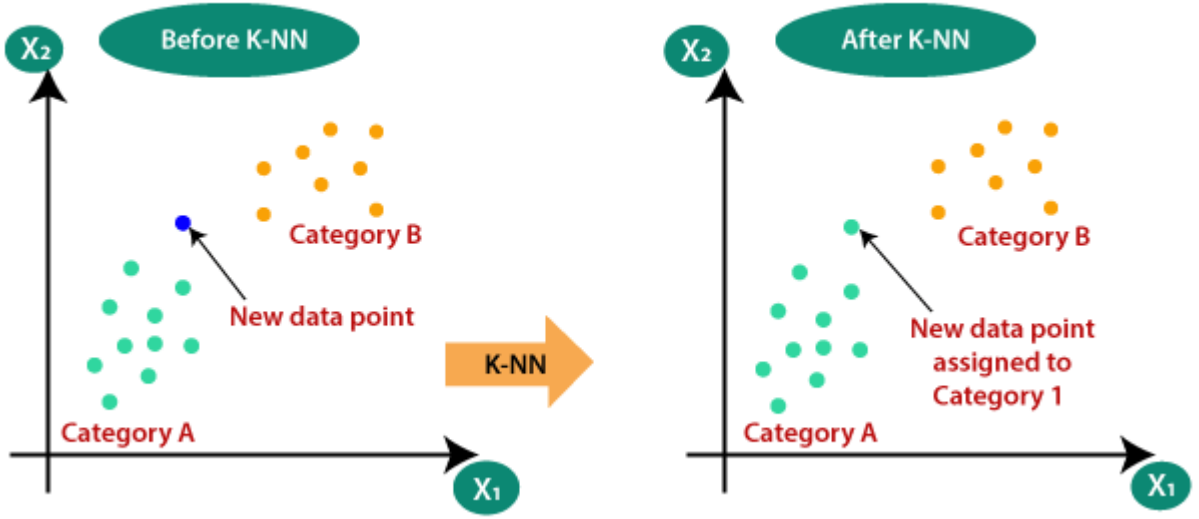
  Specificity = TN / (TN + FP)

- **False Positive Rate:** It tells us what proportion of the negative class got incorrectly classified by the classifier.

  FPR = FP / (TN + FP) = 1- Specificity

K-Nearest Neighbor Algorithm

- This algorithm is used to solve the classification problems.
- K-nearest neighbor or K-NN algorithm basically creates an imaginary boundary to classify the data. When new data points come in, the algorithm will try to predict that to the nearest of the boundary line.
- Suppose there are two categories, i.e., Category A and Category B, and we have a new data point(in blue colour), so K-NN algorithm can be used to identify the category in which this data point will lie.



K-Nearest Neighbor Algorithm

- Select the number ‘k’ of the neighbors.(k=5)
- Calculate the Euclidian Distances of the new data point with all the existing data points
- Take the k nearest neighbors and count the number of data points in each category.
- Assign the new data points to that category for which the number of the neighbour is maximum.

- **Standardization: When independent variables in training data are measured in different units, it is important to standardize variables before calculating distance.**
- **Xstd = (X – mean) / standard deviation**
- **Xstd = (X-mean)/(max-min)**

K-Nearest Neighbor Example

- Suppose we have height, weight and T-shirt size of some customers.
- We have to predict the size of the T-shirt of a new customer whose height and weight is given.
- Let us consider the following available information:
- Find T-shirt size of the new customer whose
  - Height: 161 cm
  - Weight: 61 kg
  - T-Shirt Size: ?

Height (cm)	Weight (kg)	T-shirt Size
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M
163	60	M
163	61	M
160	64	L
163	64	L
165	61	L
165	62	L
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L