Classification Algorithm in Machine Learning

Supervised Machine Learning algorithm

- Supervised Machine Learning algorithm can be broadly classified into Regression and Classification Algorithms. In Regression algorithms, we have predicted the output for continuous values, but to predict the categorical values, we need Classification algorithms.

What is the Classification Algorithm

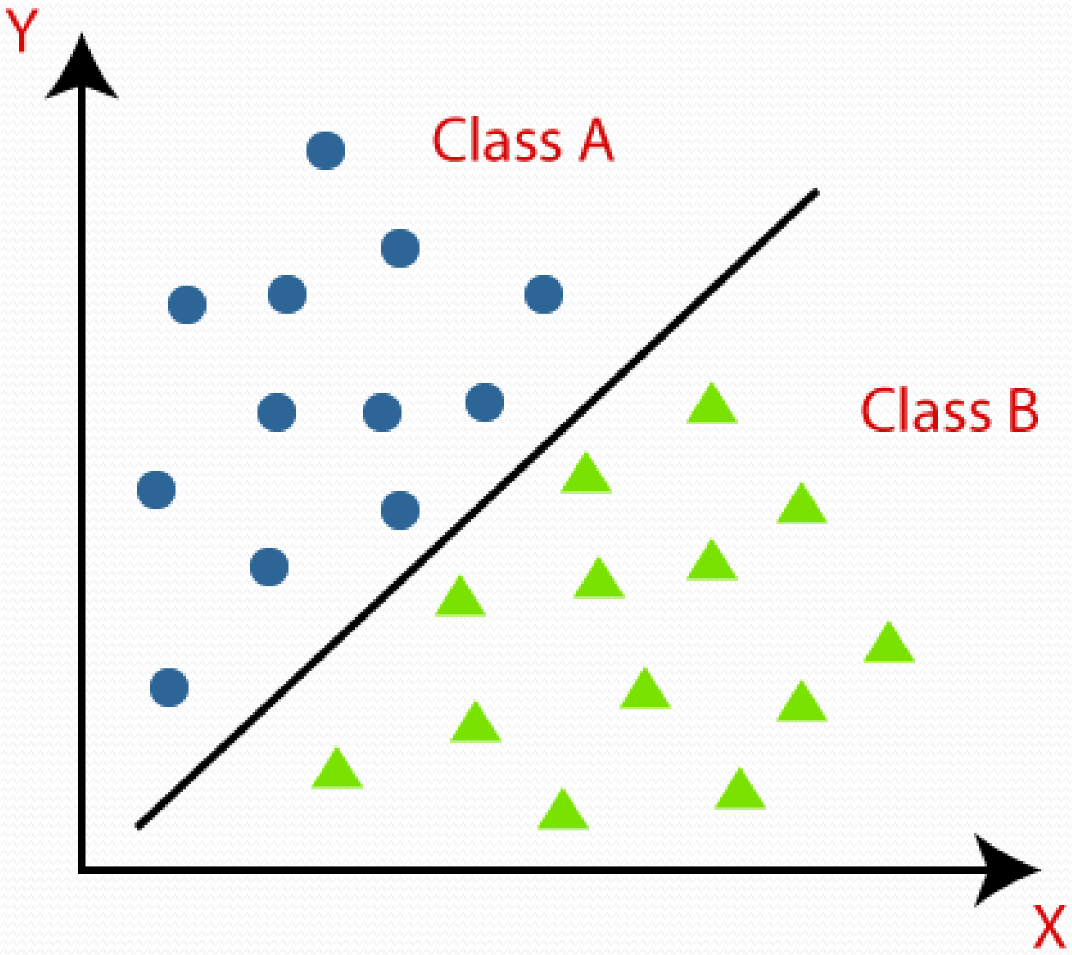
- The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.
- In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups.
- Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

Cont…

- Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc.
- Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.
- $y=f(x)$, where y = categorical output
- The best example of an ML classification algorithm is **Email Spam Detector**.

Classification

- The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.
- Classification algorithms can be better understood using the below diagram. In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.


Y
Class A
Class B
X

There are two types of Classifications:

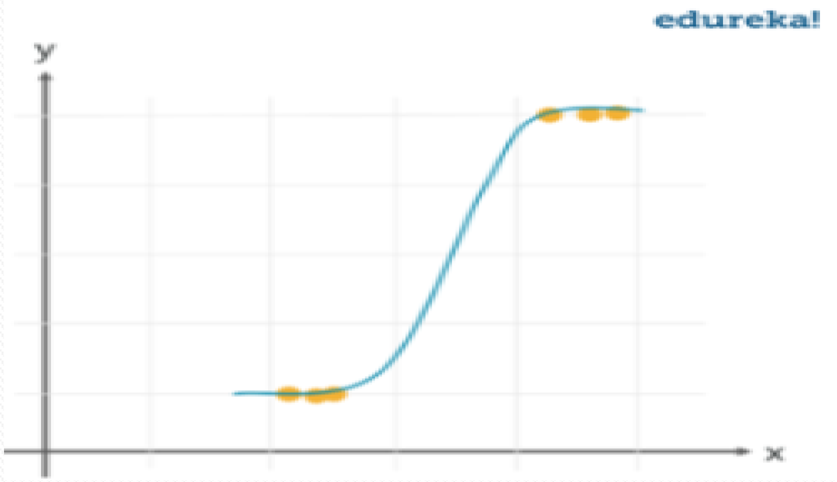
- **Binary Classifier:** If the classification problem has only two possible outcomes, then it is called as Binary Classifier.
  Examples: YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.
- **Multi-class Classifier:** If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.
  Example: Classifications of types of crops, 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 Bayes
  - Decision Tree Classification
  - Random Forest Classification

Logistic Regression

- It is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome. The outcome is measured with **have only two possible outcomes**.
- The goal of logistic regression is to find a best-fitting relationship between the dependent variable and a set of independent variables.



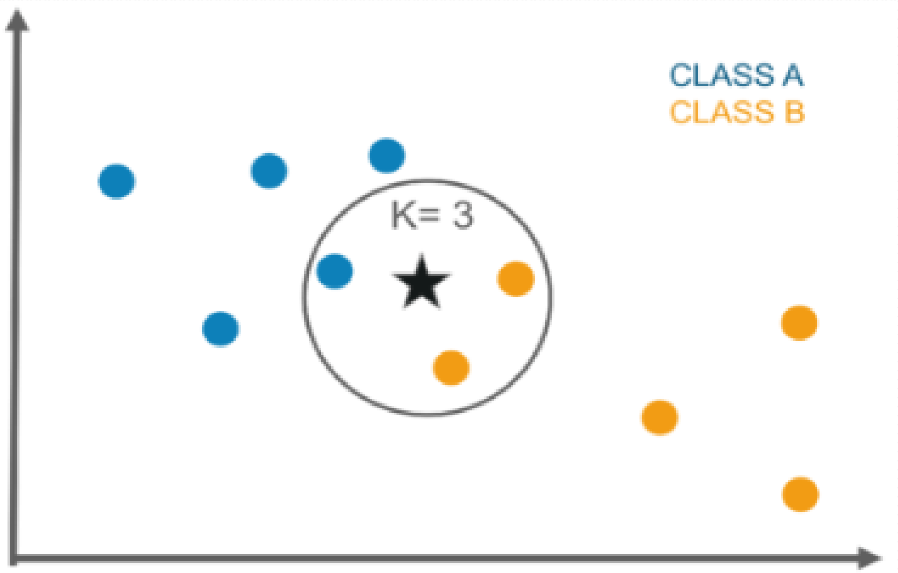
Naive Bayes Classifier

- It is a classification algorithm based on **Bayes's theorem** which gives an assumption of independence among predictors.
- Naive Bayes model is easy to make and is particularly useful for comparatively large data sets.
- The Naive Bayes classifier requires a small amount of training data to estimate the necessary parameters to get the results. They are extremely fast in nature

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \frac{P(x_1, x_2 \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2 \ldots, x_n)}\ for\ 1 < i < k$$

K-Nearest Neighbor

- It is a algorithm that **stores all instances corresponding to training data in n-dimensional space**. It is a **lazy learning algorithm** as it does not focus on constructing a general internal model, instead, it works on storing instances of training data.



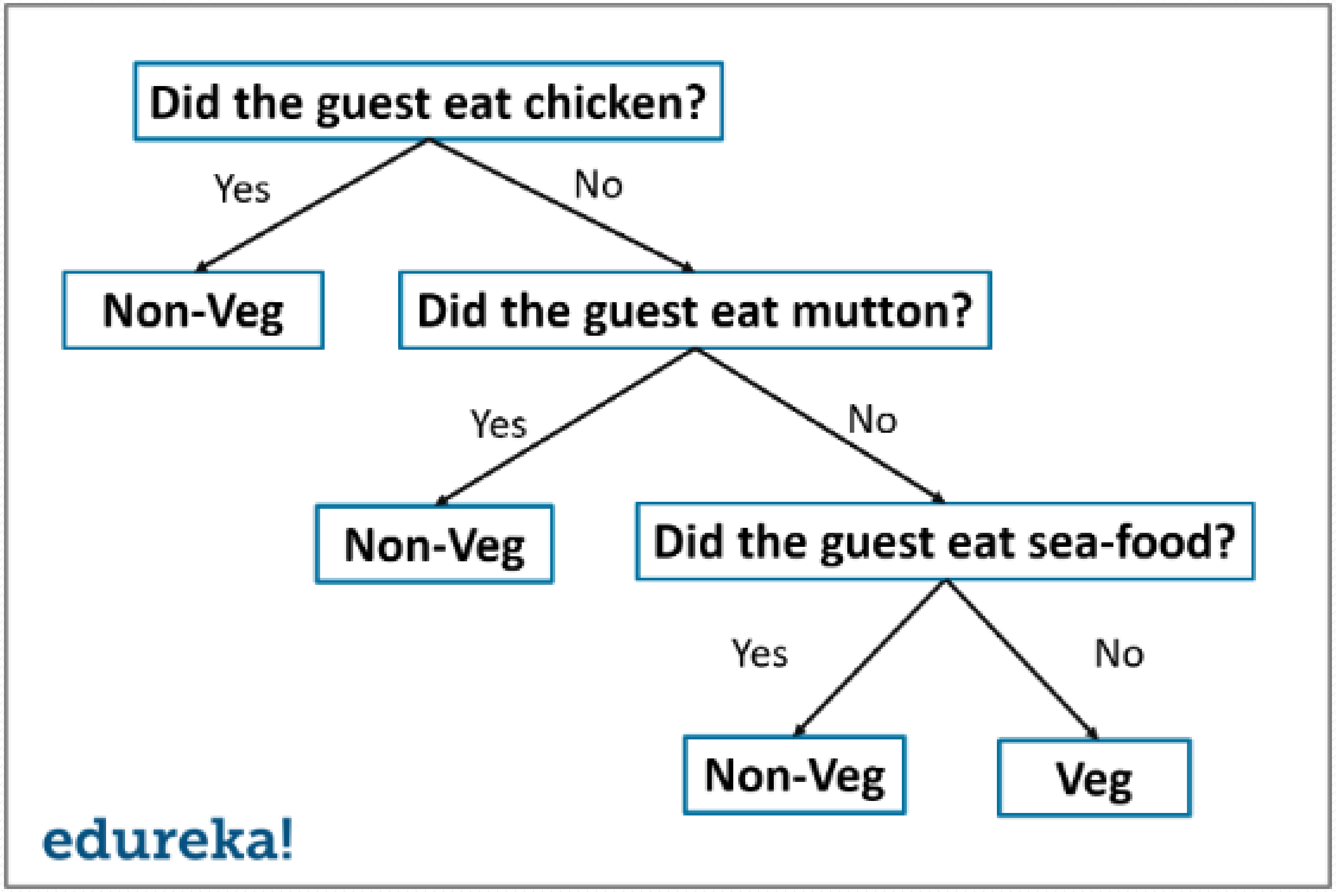
Cont…

- Classification is computed from a simple majority vote of the k nearest neighbors of each point.
- It is supervised and takes a bunch of labeled points and uses them to label other points.
- To label a new point, it looks at the labeled points closest to that new point also known as its nearest neighbors.

Decision Tree

- The decision tree algorithm builds the classification model in the form of a **tree structure**.
- It utilizes the if-then rules which are equally exhaustive and mutually exclusive in classification. The process goes on with breaking down the data into smaller structures and eventually associating it with an incremental decision tree
- The final structure looks like a tree with nodes and leaves. The **rules are learned sequentially** using the training data one at a time

Decision Tree



Did the guest eat chicken?

Yes

No

Non-Veg

Did the guest eat mutton?

Yes

No

Non-Veg

Did the guest eat sea-food?

Yes

No

Non-Veg

Veg

edureka!

Evaluating a Classification model:

- Once our model is completed, it is necessary to evaluate its performance; either it is a Classification or Regression model.
- So for evaluating a Classification model, we have the following ways:
- **1. Log Loss or Cross-Entropy Loss**
- **2.Confusion Matrix**
- **3. AUC-ROC curve:**

Log Loss or Cross-Entropy Loss

- It is used for evaluating the performance of a classifier, whose output is a probability value between the 0 and 1.
- For a good binary Classification model, the value of log loss should be near to 0.
- The value of log loss increases if the predicted value deviates from the actual value.
- The lower log loss represents the higher accuracy of the model.
- For Binary classification, cross-entropy can be calculated as:

Confusion Matrix

- The confusion matrix provides us a matrix/table as output and describes the performance of the model.
- It is also known as the error matrix. The matrix consists of predictions result in a summarized form, which has a total number of correct predictions and incorrect predictions.

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

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total Population}}$$

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.
- To visualize the performance of the multi-class classification model, we use the AUC-ROC Curve.
- 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.

Use cases of Classification Algorithms

- Classification algorithms can be used in different places. Below are some popular use cases of Classification Algorithms:
- Email Spam Detection
- Speech Recognition
- Identifications of Cancer tumor cells.
- Drugs Classification
- Biometric Identification, etc.