Common Machine Learning Algorithms

by

Abimbola Ayeni

(Trainee)

Tech Talent Academy (TTA)

55, Colmore Row, Birmingham, B3 2AA or

Origin Workspace, 40 Berkeley Square, Bristol, BS8 1HP

**Supervised by**

**Georgina Stanley**

**(Technical Trainer)**

January 17th, 2022

**Common Machine Learning Algorithms**

The term Machine Learning (ML) was coined by Arthur Samuel in 1959, an American pioneer in the field of computer gaming and artificial intelligence and stated that “it gives computers the ability to learn without being explicitly programmed” (Geeks, 2021). Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people (Lisa, 2017). Machine Learning Algorithms are defined as the algorithms that are used for training the models (Priya, 2020). Machine Learning Algorithms are now involved in more and more aspects of everyday life from what one can read and watch, to how one can shop, to who one can meet and how one can travel (Syed & Dharmendra, 2019). In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed (Lisa, 2017).

Machine Learning Algorithm can be divided into (Deepak et al, 2021):

(1) Supervised ML algorithms is a type of ML technique that can be applied according to what was previously learned to get new data using labelled data and to predict future events or labels. In this type of learning, supervisor (labels) is present to guide or correct.

(2) Unsupervised ML algorithms: In this type, there is no supervisor to guide or correct. This type of learning algorithm is used when unlabelled or unclassified information is present to train the system. The system does not define the correct output, but it explores the data in such a way that it can draw inferences (rules) from datasets and can describe hidden structures from unlabelled data.

(3) Semi-supervised ML algorithms are algorithms that are between the category of supervised and unsupervised learning. Thus, this type of learning algorithm uses both unlabelled and labelled data for training purposes, generally a small amount of labelled data and a large amount of unlabelled data. This type of method is used to improve the accuracy of learning.

(4) Reinforcement ML algorithms is a type of learning method that gives rewards or punishment based on the work performed by the system. If we train the system to perform a certain task and it fails to do that, the system might be punished; if it performs perfectly, it will be rewarded. It typically works on 0 and 1, in which 0 indicates a punishment and 1 indicates a reward.

Below are some of the Machine Learning algorithms (Priya, 2020)

#### **1. Linear Regression**

As the name suggests, this algorithm could be used in cases where the target variable, which is continuous in nature, is linearly dependent on the dependent variables. The goal of Linear Regression is to find the best fit line which would minimize the difference between the actual and the predicted data points.

**It falls under supervised learning.**

**Regression is used for predicting continuous values. It has various applications such as finance, economics, epidemiology, Sales prognostication, Risk analysis, Housing applications.**

#### **2. Logistic Regression**

In terms of maintaining a linear relationship, it is the same as Linear Regression. However, unlike in Linear Regression, the target variable in Logistic Regression is categorical, i.e., binary, multinomial or ordinal in nature. Moreover, the choice of the activation function is important in Logistic Regression as for binary classification problems, the log of odds in favour, i.e., the sigmoid function, is used.

**It falls under supervised learning.**

#### **3. K-Nearest Neighbours**

The idea behind the KNN method is that it predicts the value of a new data point based on its K Nearest Neighbours. K is generally preferred as an odd number to avoid any conflict. While classifying any new data point, the class with the highest mode within the Neighbours is taken into consideration. While for the regression problem, the mean is considered as the value.

**It falls under supervised learning.**

**KNN algorithms have been used since 1970 in many**[**applications like pattern recognition**](https://www.educba.com/pattern-recognition-applications/)**, data mining, statistical estimation, and intrusion detection, and many more.**

#### **4. Support Vector Machines (SVM)**

A classification algorithm where a hyperplane separates the two classes. In a binary classification problem, two vectors from two distinct classes are considered known as the support vectors, and the hyperplane is drawn at a maximum distance from the support vectors.

**It falls under supervised learning.**

**Common applications of the SVM algorithm are Intrusion Detection System, Handwriting Recognition, Protein Structure Prediction, Detecting Steganography in digital images, etc.**

#### **5. Naive Bayes**

It works on the principle of [Bayes Theorem](https://www.educba.com/bayes-theorem/), which finds the probability of an event considering some true conditions. Bayes Theorem is represented as –

The algorithm is called Naive because it believes all variables are independent, and the presence of one variable doesn’t have any relation to the other variables, which is never the case in real life.

**It falls under supervised learning.**

**The Naive Bayes algorithm performs better or has a higher success rate in text classification; therefore, it is used in Sentiment Analysis and Spam filtering. As a result, naive Bayes could be used in Email Spam classification and in text classification.**

#### **6. Decision Tree**

Used for classification and regression problems, the Decision Tree algorithm is one of the simplest and easily interpretable Machine Learning algorithms. Moreover, it is not affected by outliers or missing values in the data and could capture the non-linear relationships between the dependent and the independent variables.

**It falls under supervised learning.**

**As one of the most important and supervised algorithms, Decision Tree plays a vital role in decision analysis in real life**

#### **7. Random Forest**

To reduce overfitting in the Decision Tree, it is required to reduce the variance of the model, and thus the concept of bagging came into place. Bagging is a technique where the output of several classifiers is taken to form the final output. Random Forest is one such bagging method where the dataset is sampled into multiple datasets, and the features are selected at random for each set. Then on each sampled data, the [Decision Tree algorithm](https://www.educba.com/decision-tree-algorithm/) is applied to get the output from each mode. Random Forest is not influenced by outliers, missing values in the data, and it also helps in dimensionality reduction as well. However, it is not interpretable, which is a drawback for Random Forest.

**It falls under supervised learning.**

**In banking, it is used to determine loyal customers and fraud customers. It is used to detect which customer will be able to pay the loan back. Because in banking it is very important to issue loans only to those customers who will be able to pay it in time. Also, a random forest is used to predict if a customer is fraudulent or not. Bank’s growth depends on such type of prediction.**

**In the medicinal field, the random forest is used to diagnose the disease based on the patient’s past medical records.**

**In the stock market, the random forest is used to identify the market and stock behaviour.**

**In the field of e-commerce, this algorithm is used to predict the customer’s preference based on past behaviour.**

#### **8. K-means Clustering**

Here the data is unlabelled and needs to be clustered into specific groups. There are several clustering techniques available. However, the most common of them is the K-means clustering. Ink-means, k refers to the number of clusters that need to be set in prior to maintaining maximum variance in the dataset. Once the k is set, the centroids are initialized. The centroids are then adjusted repeatedly so that the distance between the data points within a centroid is maximum and the distance between two separate is maximum. Euclidean distance, Manhattan distance, etc., are some of the distance formula used for this purpose.

**It falls under unsupervised learning.**

**It is used in Social network analysis and in identifying crime-prone areas.**

**References**

[1.] Deepak Painuli, Divya Mishra, Suyash Bhardwaj, Mayank Aggarwal (2021): “Forecast and prediction of COVID-19 using machine learning”, Academic Press, Data Science for COVID-19, Pages 381-397, Editor(s): Utku Kose, Deepak Gupta, Victor Hugo C. de Albuquerque, Ashish Khanna, ISBN 9780128245361, https://doi.org/10.1016/B978-0-12-824536-1.00027-7.

(https://www.sciencedirect.com/science/article/pii/B9780128245361000277)

<https://doi.org/10.1016/B978-0-12-824536-1.00027-7>[Get rights and content](https://s100.copyright.com/AppDispatchServlet?publisherName=ELS&contentID=B9780128245361000277&orderBeanReset=true)

[2.] GeeksforGeeks (2021): “An introduction to Machine Learning”

<https://www.geeksforgeeks.org/introduction-machine-learning/>

[3,] Prya Pedamkar (2020): “Introduction to Machine Learning Algorithms”

<https://www.educba.com/machine-learning-algorithms/>

[4.] Syed Muzamil Basha and Dharmendra Singh Rajput (2019): “[Survey on Evaluating the Performance of Machine Learning Algorithms: Past Contributions and Future Roadmap](https://www.sciencedirect.com/science/article/pii/B9780128167182000166)”, Academic Press, Editor(s): Arun Kumar Sangaiah, Pages 153-164, ISBN 9780128167182,

https://doi.org/10.1016/B978-0-12-816718-2.00016-6.

(https://www.sciencedirect.com/science/article/pii/B9780128167182000166)