Algorithms Used

In this project different ML algorithms are used and several techniques of data mining are also used to check the dataset that weather it is a balanced dataset or not and check for the data is structured or not for the disease prediction. The various ML algorithms used in this project are:

- Logistic Regression: The logistic regression is used the find the probability of a event that ~~weather a event is go~~ing to occur or not. The logistic regression is used in statistics to find the probability of the occurrence of a event like the probability the school will open or not is either 1 or 0 where 1 means that the school will open and 0 means that theschool will remain closed. For determining the logistic regression the sigmoid function is used in this algorithm i.e.

$$\mathit{Sigma}(t) = \frac{1}{1 + e^{-t}} \tag{3.1}$$

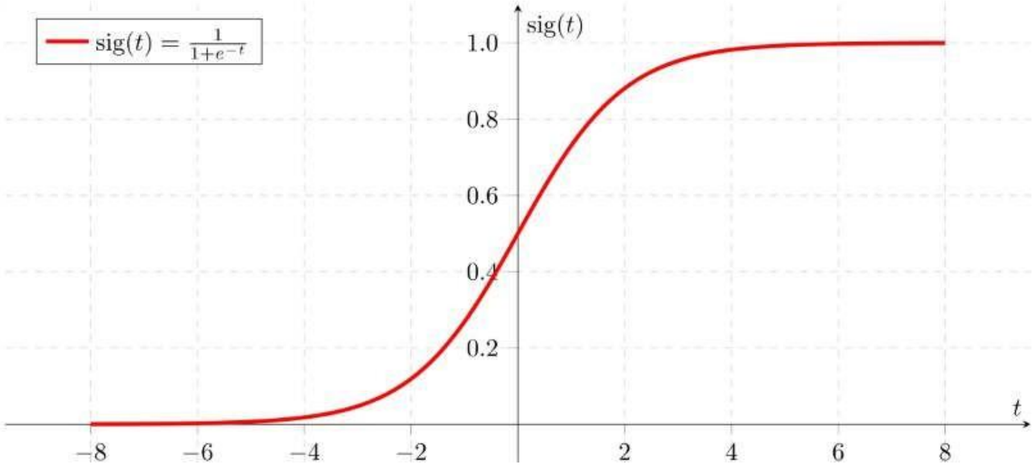

Figure 3.1: Sigmoid Activation Function

There are different types of logistic regression present in machine learning which have their specific uses:

  - Binary Logistic Regression: In Binary Logistic Regression there are only two cases i.e. only a event can happen or not.
  - Multinomial Logistic Regression: In Multinomial Logistic Regression there can be three number of possibilities like a person can purely vegetarian, purely non-vegetarian and both can consume both also in the third case. So this type of situation comes under the category of the Multinomial Logistic Regression.

- Ordinal Logistic Regression:Ordinal Logistic Regression works when there are 3 or more categories on which the logistic regression is to be applied like Rating the food of a hotel from 1-10 where 1 is for best and 10 is for bad.

- Naive Bayes Classifier: The Naive Bayes is used as a classification technique in the ML which is used to classify the things and give the answer on the basis of the classification. The main working of the Naive Bayes Classifier is based on the bayes theorm of the statistics and the bayes theorm states that:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{3.2}$$

By making the use of bayes theorm we can find the probability of X to occur when the probability of Y occurrence is given to us. For exmaple:

Table 3.1: Prediction Using Naive Bayes

S.No.	Temperature	Weather	Rain	Humid	Play Tennis
1	Hot	Sunny	No	High	Yes
2	Mild	Overcast	Yes	Low	Yes
3	Mild	Rainy	Yes	High	NO
4	Cold	Rainy	Yes	High	NO
5	Hot	Sunny	No	Low	Yes

This is kind of input which is given to the Naive Bayes Classifier for the prediction. On this kind of table input the bayes theorm is used the ML algorithm for the prediction, with this table the prediction can be done as probably of playing tennis when the temperature is mild, weather is rainy, rain is yes, humid is high like this the prediction is done in the ML model also.There are three different types of naive bayes classifier:

- Multinomial Naive Bayes: This is mainly used for the classification of big files and documents by dividing them into different categories as weather a file or a document is of sports category, politics, technology or something else. This method of classification is very widely used in Machine Learning algorithms.
- Bernoulli Naive Bayes: This method of classification is very similar to the multinomial naive bayes but the result of this type of naive bayes is only in either yes or no.
- Gaussian Naive Bayes: This type of naive bayes is used for the prediction of the continuous values while the other two types were used for the discrete value prediction.

In this way the Naive Bayes Classifier is used in different ways in the ML models for the process of prediction and analysis.

- Decision Tree Algorithm: The decision tree algorithm is a type of a supervised learning ~~algorithm which can be~~ used in the case of classification as well as regression it has capability to solve both the kind of problems. In decision tree for predicting the value we start from the root of tree and then form a sub-tree from that root and finally cometo a conclusion which is nothing but the predicted value. The detailed overview of the decision tree can be given as:

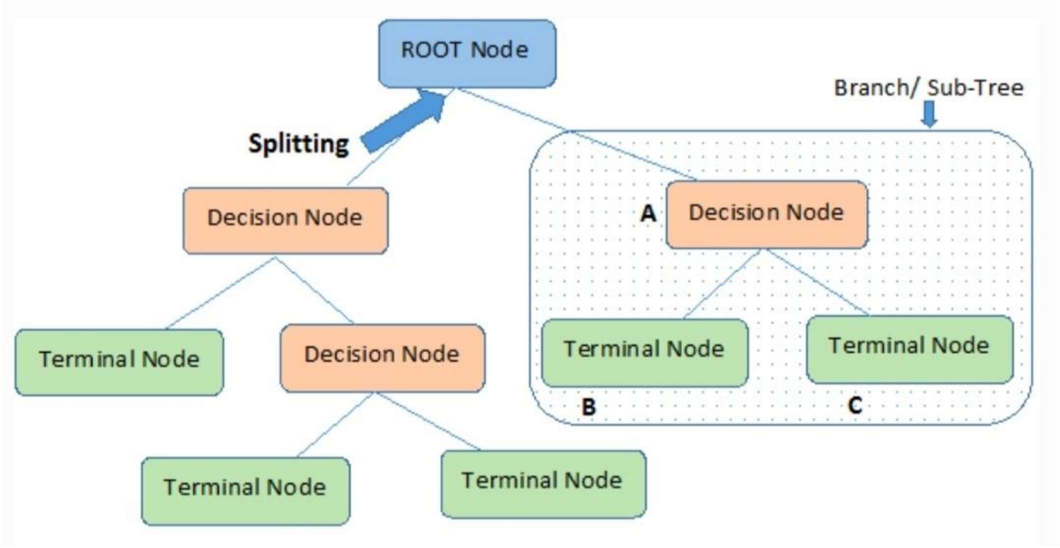

Figure 3.2: Decision Tree Overview

Decision tree helps in the method of classification as it sorts the values while traversing down the tree and predicts the right value.The decision tree starts from the root node on each step of moving downward in the tree the Information Gain and entropy are calculated, after that the branch with lowest entropy or the highest information gain is selected and the information gain and the entropy are calculated again.

$$\sum$$

$$Entropy =$$

$$Informationgain = entropy(parent) - [averageentropy(children)] \qquad (3.4)$$

When the entropy is high then it is concluded that the randomness of the dataset is high and then it is hard to predict the answer whereas the information gain tells us about that how well structured the dataset is and if the information gain is high then it is easy to predict the answer. All these steps are repeated until the decision tree reaches the leaf node i.e. the final predicted value by the decision tree. Different types of Decision tree present are:

Categorical Variable Decision Tree:

These types of decision tree works on the whole value i.e. suppose we have to choose between 0 and 1 or something which istaken as a whole.

– Continuous Variable Decision tree: These types of the decision tree works on the continuous values that’s why they are known as the continuous variable decision tree.

In this way the decision tree algorithm is used in classification and regression problems.

Random Forest Algorithm

- The random forest algorithm is a type of supervised learn- ing algorithm which is used for both classification as well as regression as the basic ideabehind the random forest algorithm is the decision tree algorithm, this algorithm createsmultiple decision trees and then predicts the values as shown in the figure 3.3

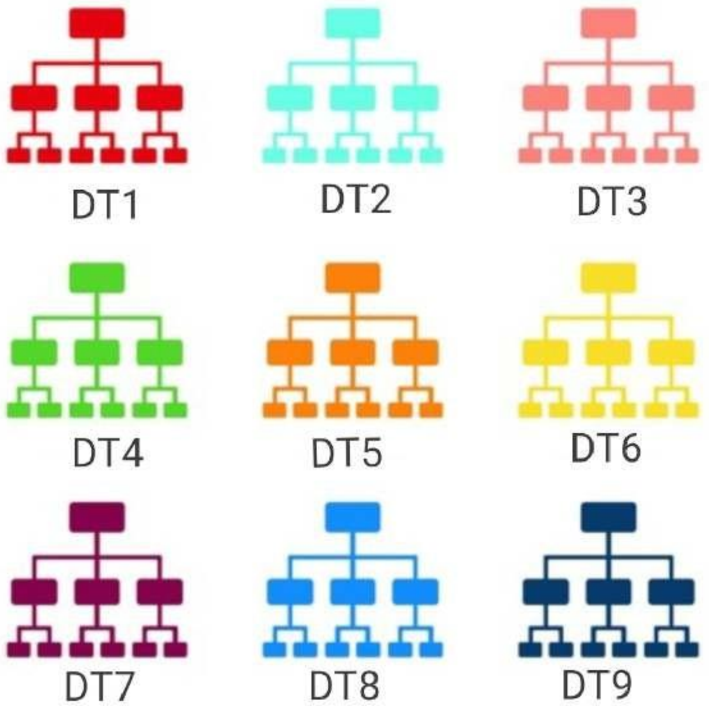

Figure 3.3: Random Forest Overview

As shown in the figure there as multiple decision trees in a random forest algorithm. Many of these decision trees will not be performing good enough and the best performingdecision trees are selected for prediction.