1. **System Architecture**
2. **Implementation Phase**

In this chapter, we have used three models:-

* Random Forest
* K-Nearest Neighbors
* Logistic Regression

## Random Forest: -

The Random Forest algorithm is an algorithm used extensively in the emerging field of Machine Learning. This is mainly a supervised classification algorithm, which means that it is used for segregating or assigning the object to a particular class of instances where it might belong. An example of classification is to classify a cancer as benign or malignant. There are myriad such algorithms used for classification, like Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN) and numerous others [15]. This algorithm in addition to being used for classification can also be seldom used for regression.

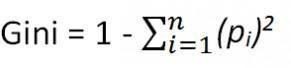
The Random Forest algorithms is a huge collection and cluster of numerous decision trees.

DECISION TREE:

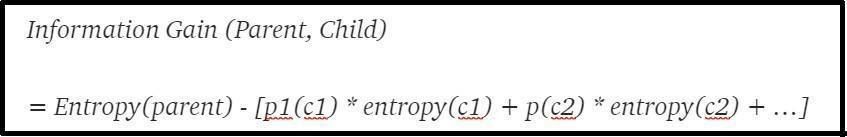
The decision trees have a representation similar to the tree structure. A tree-like flowchart is drawn, where each node resembles an attribute, the branches denote a decision rule and the leaf nodes are the final outcomes. These trees, as suggested by the name help in decision making process by the numerous recursive branches of the tree. CART (Classification and Regression techniques) are applied in deriving and deciding the final outcome of any event.

The results are governed by various factors like probability and various indices. Parameters like Gini Index and Entropy are some of the factors which are used to determine the root node and construct the final decision tree. The outcomes are then drawn by traversing the branches according to the conditions and values posit in the data tuple. [16]

To construct the decision tree, to predict the outcome of a single tree, we can use the Gini Impurity or the Entropy.



Using this above formula and multiplying this with the probability of the instances of the tuple, we obtain the Gini index. The attribute with the least value becomes the root node and the same procedure is now carried out again, once the tuples are classified according to the new root node.

This repeats till we get definite leaf nodes which signify the final outcome, when the tree is traversed using a given tuple.

Entropy is calculated in a similar fashion, but the one with the largest value is made the root node in case of decision tree formation using entropy calculation. [17]

This algorithm falls under the category of Ensemble machine learning algorithm, which are meta- algorithms which combine several machine learning techniques into one single model which decreases variance, bias and improves predictions. [18]

The random forest being a collection of numerous decision trees depend on the outputs of these trees. The final output of the random forest algorithm is decided by choosing the outcome with maximum votes as decided by all the individual trees. Thus, this ensemble approach provides a highly efficient output, since in this manner the individual errors are compensated for by the other trees. [19]

The correlation between the trees present in the forest is very low and this plays a very paramount role in the decision of outcome.

A pictorial representation of the Random forest algorithm can be depicted by the image below. [20]

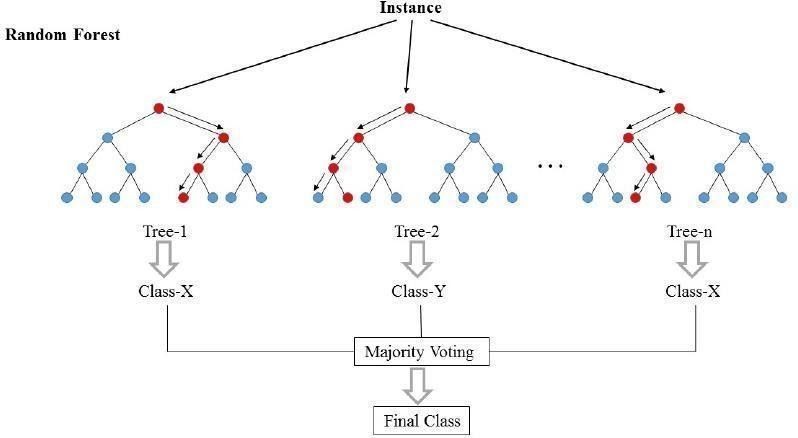


Fig.1 [21]

## K-Nearest Neighbours: -

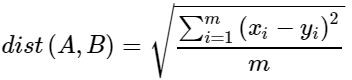
The kNN Algorithm of Supervised Machine learning stands for K-Nearest Neighbours. This is a classification algorithm, used for many applications in classifying the data tuples. To explicate this, we can consider a data tuple N, to be classified, among the retrieved k nearest neighbours to tuple N. The classification of the tuple N is dependent on the weight based factors (distance of the tuple from the neighbours) and is classified into the neighbourhood with minimum associated weight [22].

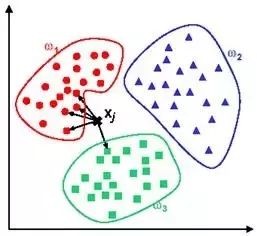
There is no predetermined value of k, which can certainly yield an optimum outcome. The value of k is changed each time, which reflects the change in accuracy of the model. The value of k, at which maximum accuracy is obtained, can be the right value for our purpose.

[23] However, research is being carried out to determine the value of k using a k-Tree Method to determine the optimal value of k, as proposed by the paper mentioned in [24].

In the diagram below, the distance of attribute/tuple xj, is calculated from each of the 3 neighbours (k=3). The neighbours - w1,w2,w3 represent the classes, which denotes the general characteristics of all the data points included in the respective class. The closest class to xj, will determine the class of xj, thus completing and yielding an outcome of the classification problem.

The Euclidean distance between points A(x1,x2….xm) and B(y1,y2...ym), *m* being the dimensionality of the feature space is given by the formula below [25].





## Logistic Regression: -

The Logistic Regression algorithm is an algorithm used extensively in the emerging field of Machine Learning. This is mainly a supervised classification algorithm, which means that it is used for segregating or assigning

the object to a particular class of instances where it might belong. y=f(x) is a method for fitting a regression curve in logistic regression as the function varies between zero and unity

i.e success and failure. It requires iterative methods to fit a logistic regression model and we use the maximum likelihood function to maximise the probability using the values of alpha and beta concerned in the data set. Logistic regression deals with the relationship of multiple independent variables and dependent variables to analyse the probability of the event concerned.[26]

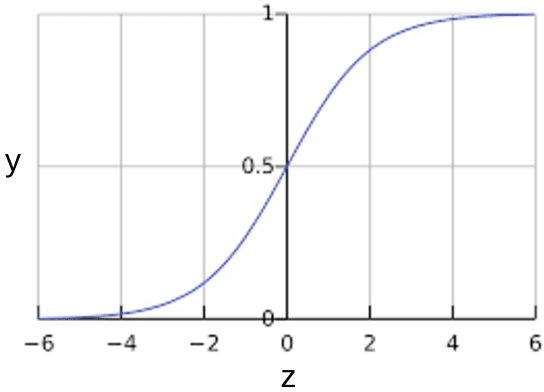
The model is used to predict a plethora of dependent variables distinguished from the pile of independent variables. Overall Evaluation of the model and goodness of fit-statistics are some important factors to be considered while fitting the logistic model efficiently. It's a unique type of multivariate which analyses variables used in high frequency.

We first find the regression coefficient b1 which we used to estimate the increase in the odds of the outcome for every increase in the value of the independent variable. The coefficients are the key representations in a logistic regression model. The odds play a vital role in the accuracy of the model as they signify the ratio of probability of success to failure which are mapped into the model. It tends to behave as a specialised case of the linear regression model but gives a better accuracy in most of the cases and fits the data efficiently.

The Sigmoid Function is a special characteristic function. It is a function which ranges between the values of zero and unity. It has a characteristic “S”-shaped curve which is extensively used in logistic regression [27]. A common example of sigmoid function is the logistic function shown in the figure.

The expression of sigmoid function is:

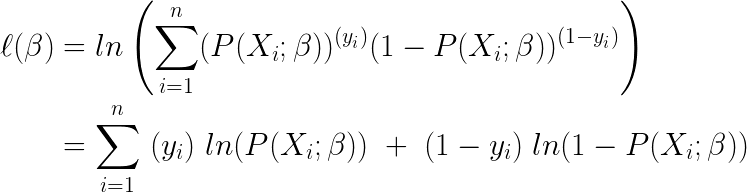


[28]

Maximum likelihood estimation

We will find the β parameters that reach the global maximum of the log likelihood function using maximum likelihood estimation.

Using the maximum likelihood estimation we can fit the model by the equation:



1. **Results and Analysis**
2. **Conclusion and Future Scope**
3. **Reference**