International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

Classification Using Hybrid Machine Learning Techniques

Azhar M.A[1], Princy Ann Thomas [2]

P.G. Student, Department of Computer Engineering, Government Engineering College, Idukki, Kerala, India[1]

Associate Professor, Department of Computer Engineering, Government Engineering College, Idukki, Kerala, India [2]

ABSTRACT: Prediction of heart disease can be useful for the healthcare industry. Some Machine learning methods can help us remedy this circumstance. This helps in getting a better insight into the concept of a classification problem. However, there was an absence of successful analysis methods to find connections and patterns in health care data. Some Machine learning methods can help us remedy this circumstance. I have used the hybrid Classification method The work includes the classification of heart disease using Support Vector Machine and Multi-Layer Perceptron with a hybrid. A comparison is carried out among different methods based on performance measures.

KEYWORDS: Data mining, Hybrid, Classification model, Information gain

I. INTRODUCTION

Heart disease is one of the most critical human diseases in the world and over the last decade heart disease is the main reason for death in the world. To lower the number of deaths from heart diseases, there has to be a fast and efficient detection technique[1]. The hybrid method is one of the effective data mining methods. A major challenge facing healthcare organizations is the provision of good quality services at affordable costs. Quality service implies diagnosing patients correctly and administering effective treatments. one of the major challenges in heart disease is correct detection and finding the presence of it inside a human. Early techniques have not been so much efficient in finding it even a medical professor is not so efficient enough in predicting heart disease. There are various medical instruments available in the market for predicting heart disease there are two major problems in them, the first one is that they are very much expensive and the second one is that they are not efficiently able to calculate the chance of the art disease in human. With the headway in software engineering has gotten tremendous open doors in various zones, clinical science is one of the fields where the instrument of computer science can be used.Machine Learning is one such tool that is widely utilized in different domains because it doesn't require a different algorithm for different datasets. Reprogrammable capacities of machine learning bring a lot of strength and open new doors of opportunities for the area like medical science. In medical science heart disease is one of the major challenges; because a lot of parameters and technicality are involved for accurately predicting this disease. Machine learning could be a better choice for achieving high accuracy for predicting not only heart disease but also, another disease because this vary tool utilizes feature vector and it is various data types under various condition for predicting the heart disease, algorithms such as support vector machine and multilayer perceptron, are used to predicate risk of heart diseases. All these techniques are using old patient records for getting predication about new patients. This prediction system for heart disease helps specialists to predict heart disease in the early stage of disease resulting in saving a large number of lives. We included a hybrid feature selection model that produces an enhanced performance level with high accuracy. This model included a support vector machine and a multilayer perceptron. After the pre-processing Constructed a Decision Tree on a given dataset and construct a subset of features. Train the feature subset with any classification model and find a subset with the highest accuracy. There exist no feature selection algorithm which provides an optimal subset of features for any criterion function without doing exhaustive search where Some criterion function satisfies some properties then we can use those properties to obtain a feature selection algorithm that will give us optimal feature subset without doing an exhaustive search. So here we create a decision tree and select the one feature subset that shows better performance.

II. RELATED WORK

Comparison of different techniques in Data mining to find the best technique for creating a prediction model of heart disease at minimum effort. In machine learning, different methods used to find risk prediction of heart disease. To Study the performance measures between the classifications of the given dataset to understand their influence on prediction accuracy of different machine learning algorithms. There are several Comparision is performed for analysis of data.The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not mined to

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes, and Neural Network. Smartphone-Based Ischemic Heart Disease (Heart Attack) Risk Prediction: An Android-based prototype software has been developed by integrating clinical data obtained from patients admitted with IHD (Ischemic Heart Disease). The clinical data from 787 patients have been analyzed and correlated with risk factors like Hypertension, Diabetes, Dyslipidemia (Abnormal cholesterol), Smoking, Family History, Obesity, Stress, and existing clinical symptom which may suggest underlying non-detected IHD. The data is mined with data mining technology and a score is generated. Risks are classified into low, medium, and high for IHD.Analysis of Data Mining Techniques for Heart Disease Prediction: Heart disease is considered as one of the major causes of death throughout the world. It is hard to predict for the medical practitioners as it is a difficult task that demands expertise and higher knowledge for prediction. This paper addresses the issue of prediction of heart disease according to input attributes based on data mining techniques. Shouman M et al[1], proposed k-means clustering with the decision tree method to predict heart disease. In their work, they suggested several centroid selection methods for k-means clustering to increase efficiency. The thirteen input attributes were collected from the Cleveland Clinic Foundation heart disease data set. The sensitivity, specificity, and accuracy are calculated with different initial centroids selection methods and different numbers of clusters. This paper investigated integrating k-means clustering with decision trees in the diagnosis of heart disease patients. Initial centroid selection is one of the critical issues that strongly affect k-means clustering results. Their research systematically investigated applying different methods of initial centroid selection such as range, inlier, outlier, random attribute values, and random row methods for the k-means clustering technique in the diagnosis of heart disease patients. The results show that integrating Kmeans clustering and decision tree can enhance decision tree accuracy in the diagnosis of heart disease patients. The results also show that the best accuracy achieved is 83.9% by the inlier method with two clusters. Finally, some limitations on this work are noted that there is no feature selection method is used.Jabbar MA, al. [2] In this paper, M. A. Jabbar proposed a new algorithm to mine association rules from medical data based on digit sequence and clustering. The entire database is divided into partitions of equal size, each partition will be called a cluster. The dataset with fourteen attributes was used in that work and each cluster is considered one at a time by loading the first cluster into memory and calculating frequent itemsets. Then the second cluster is considered in the same way and calculating frequent itemsets. This approach reduces the main memory requirement since it considers only a small cluster at a time and it is scalable and efficient.This work provides a quick and easy understanding of various prediction models in data mining and helps to find the best model for further work. This is a unique approach because various techniques listed and expressed in table format to understand the accuracy level of each. These techniques are chosen based on their efficiency in the literature. In previous studies of different researchers expressed their effort on finding the best approach for the risk prediction model and here we found the best model by comparing those researcher's findings as a survey. This survey helps to understand the recent techniques involved in the risk prediction of heart disease at classification in data mining. Survey of relevant data mining techniques that are involved in risk prediction of heart disease provides the best prediction model as a hybrid approach comparing with the single model approach. The first one is applying a single model to various heart data and another one is applying a combined model to the data. The combined model is also known as a hybrid model Durga Devi et al [6].In this paper, a heart disease prediction system with better accuracy has been presented. The genetic algorithm has been used for attribute reduction and RBF networks for classification. Classification accuracy has been enhanced by reducing the number of attributes. From the experimental results, they revealed this prediction system predicts heart disease risk with enhanced accuracy. The enhanced accuracy is obtained by the attribute reduction. The data mining technique RBF Network is used to create the classifier. Genetic Algorithm is used for attribute selection or reduction. The result shows that the RBF Network performs better than the other data mining techniques Naïve Bayes and J48.

The results also demonstrate that the reduced feature subset can have better prediction performance compared to the original set of attributes. Also, we can observe that the hybrid methods were more stable than the other individual technique. In conclusion, as identified through the literature survey, believe only a marginal success is achieved in the creation of a predictive model for heart disease patients, and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease. With the more amount of data being fed into the database, the system will be very intelligent.

III. PROPOSED METHOD

In this study (see Fig. 1) decision tree-based information gain was used to select important features, and on these selected features, the performance of the classifiers (hybrid SVM and MLP) was tested and compares the performance measures. After the Comparision of performance, we can understand that the information gain gets the best features,

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

and it shows the best classification result. Classifiers MLP, SVM and its hybrid combination were used in the system.. The methodology of the proposed system structured into five stages including(fig 1) (A) pre-processing of the dataset, (B) feature selection, (C) classification,(D)hybrid approaches and (E)classifier's performance evaluation F)Experiments and Results. In this work, the medical data related to Heart diseases are considered. This benchmark dataset was obtained from the Cleveland UCI repository. This is a publicly available dataset. Cleveland dataset concerns the classification of a person into a no cardiac and cardiac person regarding heart diseases. The dataset is divided into 2 sets of training (70%) and testing (30%). Python is used for the experiment. The data consists of 13 attributes (inputs) and 2 classes (outputs). Tables 1 illustrate the representation of the attributes.

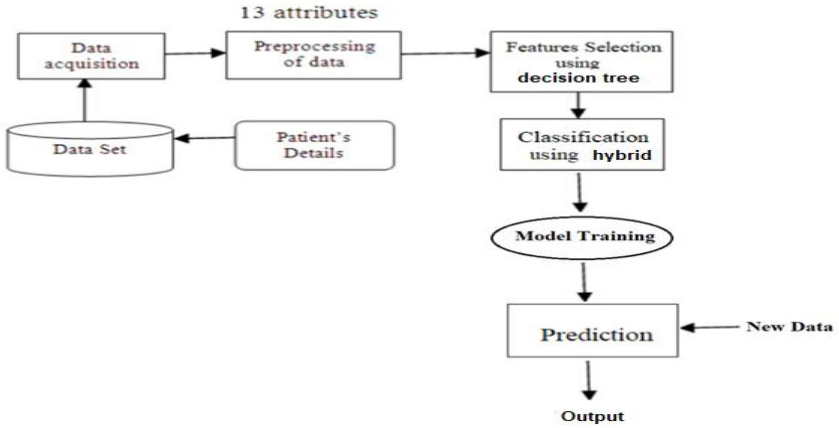


Fig 1: Block diagram of the proposed model

This paper also does a deep analysis of the utilization of learning in the field of predicting heart disease. We included a hybrid feature selection model that produces an enhanced performance level with high accuracy. This model included a support vector machine and a multilayer perceptron. After the pre-processing Constructed a Decision Tree on a given dataset and construct a subset of features. Train the feature subset with any classification model and find a subset with the highest accuracy. Finally, Learn and classify the Hybrid Model with selected features. the steps are followed.
Step-1: Pre-process the data set to remove the duplicate, missing, and unknown data.
Step-2: Construct a Decision Tree on a given Dataset.
Step-3: Traverse through the branches of the tree to construct a subset of features.
Step-4: Train the model with each subset of features.
Our result reveals that we can agree with feature selection but there is no consistent method over the feature selection because we are not in a position to go through all possible subsets. we cannot search all these subsets to get the optimal one. The other side is supposed we do not search for all possibilities then we will not get the optimal subset.

TABLE 1. UCI dataset attributes detailed information.

S.No	Description	Range	Type
1	Age	Continuous	Numeric
2	Sex	0,1	Nominal
3	CP	1,2,3,4	Nominal
4	trestBP	Continuous	Numeric
5	Chol -	Continuous	Numeric
6	FBS	0,1	Nominal
7	restECG	0,1,2	Nominal
8	Thalach	Continuous	Numeric
9	Exang	0,1	Nominal
10	OldPeak	Continuous	Numeric
11	Slope	0, 1, 2	Nominal
12	CA:	Continuous	Numeric
13	Thal	3,6,7	Nominal

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

A. Data set pre-processing

Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format data Preprocessing is necessary for efficient representation of data and machine learning classifier which should be trained and tested effectively.

Data Set Information: Heart disease data were collected from the UCI machine learning repository. There are four databases (i.e. Cleveland, Hungary, Switzerland, and the VA Long Beach). The Cleveland database was elected for this research because it is a commonly used database for ML researchers with comprehensive and complete records. The dataset contains 303 records. Although the Cleveland dataset has 76 attributes, the data set provided in the repository furnishes information for a subset of only 14 attributes.

B. Feature selection and Reduction

Dimensionality Reduction is an important factor in predictive modeling. Features contain information about the target. The classification function is defined in terms of features, more features mean more information, better discrimination power, and better classification power. But the problem is that a large number of features leads to the "curse of dimensionality"- that means higher-dimensional data sets increase the time complexity and also the space required will be more. Some features are irrelevant to the output of the model In learning algorithms these irrelevant features introduce noise which degrades the learning algorithm's performance and makes the prediction wrong. After Pre-process the data remove the duplicate, missing, and unknown data. Then construct a Decision Tree on a given dataset and traverse through the branches of the tree to construct a subset of features. After the subset creation train any classification algorithm with each subset of features and find a subset with the highest accuracy

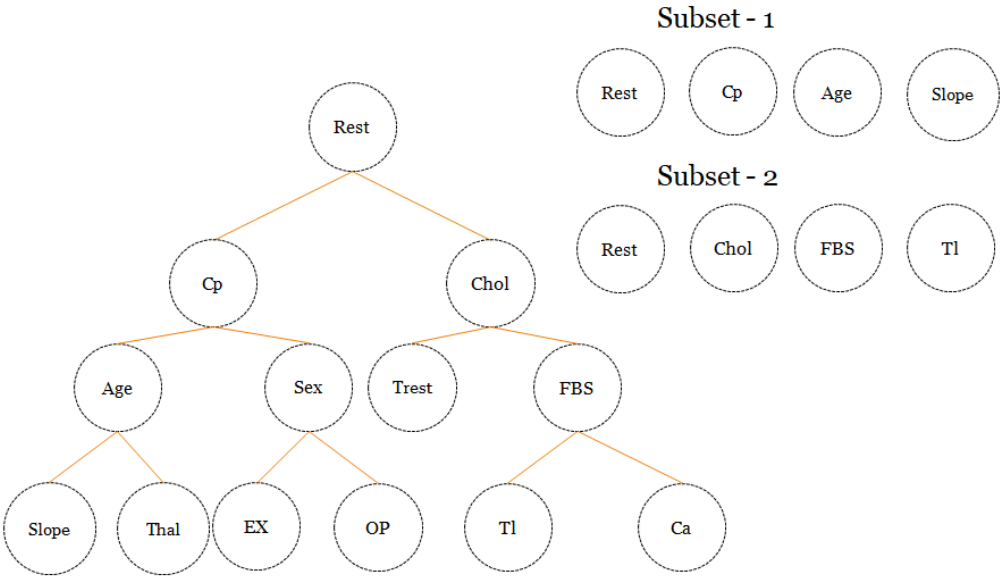


Fig 2: Features subset creation based on information gain

There are no single feature selection techniques that give consistent results for all types of data. The performance of feature selection techniques depends on the type of dataset that we have taken for experimenting. So, we can use a hybrid feature selection technique. The results also demonstrate that the reduced feature subset can have better prediction performance compared to the original set of attributes. Also, we can observe that the hybrid methods were more stable than the other individual technique.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

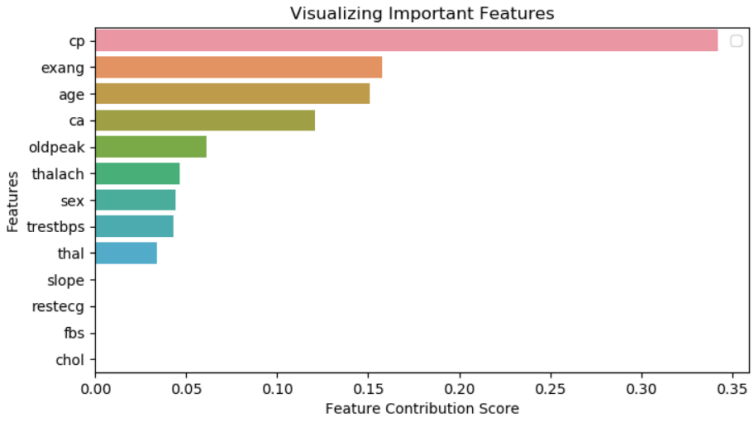


Fig 3: Features contribution based on information gain

C. Classification Modelling

In this section we have discussed the major classification techniques used in data mining for prediction of heart disease is discussed and with the help of accuracy analysis. we have shown that SVM and MLP are better than other methods.

1)Support Vector Machine (SVM)

SVM is one of the most effective classifiers among those which are sought of linear. It has a very good mathematical intuition behind the SVM and we can handle certain cases where there is on linearity by using nonlinear basis function is called kernel function.SVM has a clever way to prevent overfitting and we can work with a relatively larger number of features without requiring computation. It is a byproduct of Statistical learning theory.

Statistical Learning Theory:

$$(x_i, \theta_i): \forall i = 1,2,3,\dots n$$
$$(x_i, \theta_i): \forall i = 1,2,3,\dots n$$
$$x_i \in \mathbb{R}^N$$
$$\theta_i \in \{-1,1\} label$$

$P(x_i, \theta)$is a probability distribution on the data and p is unknown. We have a machine whose task is to learn the mapping. $x_i \rightarrow \theta_i$. In classification the problem is, we have given 'n' points and have corresponding class labels and we need to find the function from $x_i \rightarrow \theta_i$.ℶ={f(a)} the set of functions under consideration.'n' points can be labeled in 2^n ways. ℶ is said to shatter $x_1,x_2,\dots,x_n$. If for every labeling of n points we can get a function fϵ ℶ, which provides that labeling.VC dimension for a set of functions ℶ is defined as the number of points that can be shattered by ℶ. For example take two points $x_1$and x_2 , ℶ =All possible straight lines and we are in $\mathbb{R}^2$ (two-dimensional space).then 4 labeling are possible ie x1 and x_2 in +1 class, $x_1$and x_2 in -1 class, x_1 in -1 class and x_2 in +1 class and x_1 in +1 class and x_2 in -1 class(see figure 4).So two points they can be shattered by the set of straight lines. In anywhere we take every set of two points it can be shattered by a set of straight lines

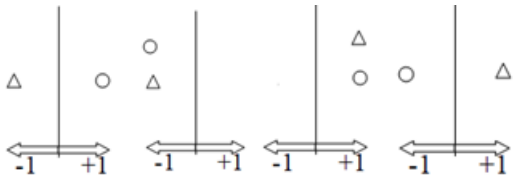


Fig 4: possible labeling of two points in a 2D space

Fig 5: The possible labeling of three points in a straight line

Similarly set of three points can be shattered in 23 ways. So we have eight sets of three points .this is shattered by a set of possible lines. But every set of three points cannot be shattered. Consider three points on a single line (fig 5).

Similarly, no set of four points can be by straight-line fig 6.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

Fig 6:The problem of labeling four points

Vapnik–Chervonenkis theory, the Vapnik–Chervonenkis (*VC*) *dimension* is a measure of the capacity of a space of functions that can be learned by a statistical classification algorithm.VC dimension of the straight line is ≥3.It can be proved that the VC dimension of hyper-planes is (N+1). There are infinitely many such vectors. How do we choose one optimal classifier? The solution is the maximum margin classifier or SVM classifier. We shall set the margin (min distance of the hyperplane to the +ve points= min distance of hyperplane to the –ve points =1)to one and achieve it with minimum weight.

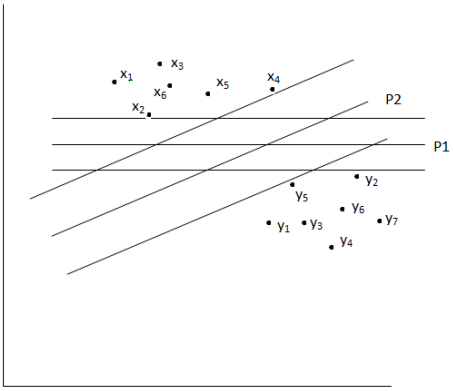


Fig 7: The classification of points using different planes

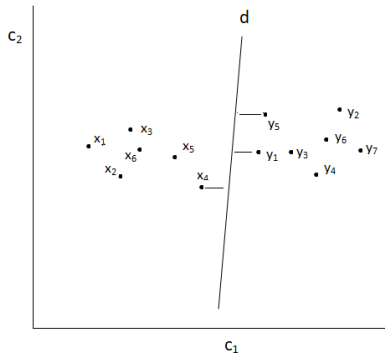


Fig 8: The classification of points using the suitable plane

Here we have two classes x_1, x_2,….x_6 and y_1,y_2,….y_7 points. There is a hyperplane for the plane 'P_1' we take the distance of the hyperplane with every one of the point x_1,x_2,…x_6. Find out the one that has minimum distance then the minimum distance is x_2. Again for the hyperplane P_1, find out the distance of its hyperplane with every one of the points. Find out the one that has minimum distance then we can get the point y_2.we can choose the hyperplane P_1 in such a way that the shift of (x_2,p_1) is one and shift of (y_2,p_1)is also one. So it(x_2,y_2) becomes two, and distance is 2/Normal of the vector.ie $\frac{2}{\|w\|}$. Then we look at the hyperplane P_2. We take the distance of the hyperplane with every one of the point x_1, x_2…x_6, and y_1,y_2,…y_7. The minimum distance is x4 and y5. Then we can find that the distance between the hyperplanes P_2 is more than the distance between the hyperplanes P_1. We would like to choose the hyperplane in such a way that for that hyperplane with the same shift we get a +ve point at the same shift we get a –ve point. So that the distance between x_2,y_2 is, $\frac{2}{||w||}$w gives the equation of the hyperplane. we aim to maximize $\frac{2}{\|w\|}$ or minimize $\bar{w}.w$, where $||w||= \bar{w}.w$ So it is also called maximum margin classifier because the finally selected margin it is a maximum of all the possible margin we can have. Here the points x_2,y_2, and x_4,y_5 is called the support vectors or the points which are close to the decision surface these are called the support vectors. Support vectors are which actually determine the equation of the line and typically the support vector number will be very small.

Consider the figure 8: then c_1 and c_2 are the x and y coordinates respectively and we have different points suppose x_1,x_1,…x_6 are positive points and y_1,y_2,…y_7 are negative points' is the decision surface. Some points are closes to the decision surface have less confidence. In SVM we need a classifier so that it maximizes the separation between the points and decision surface. Among all possible decision surface, we want to choose that one for which the margin width is highest. So margin is denoted by the maximum distance of a training instance from the decision surface and we want to choose that decision surface for which the margin width is highest. If we consider a decision surface 'd' and we also consider points near to the decision surface x_4,y_1,y_4. These points are closer to the decision surface and almost equidistance from the decision surface. These points are called the support vectors. Here we aim to maximize the margin width, the distance of the closest –ve point to the line, and the closest +ve points to the line will be the same and there is the minimum number of support vectors. There can be more support vectors but typically the number of

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

support vectors is extremely small. These support vectors determine the equation of the line. This particular decision surface decides which point given a test point and can do the test with the side of the decision surface, It is based on that we can classify by the points as +ve or –ve.

Functional margin: Consider a point (x_i,y_i) for a decision surface. Then the equation of the decision surface is $wx+b=0$,w and x are two vectors, suppose $x_1$and x_2 are two features then $w_1x_1+w_2x_2+b=0$.The functional margin is measured by the distance of the point$(x_i,y_i)$from the decision boundary. Suppose **fig** is a straight line whose equation is wx+b=0. The vector w1 and w2 are normal to this straight line .So we can define the distance (γ^i) of (x_i,y_i) from (w,b).$\gamma^i=y_i(w^Tx_i+b)$,which is the normal distance from the point(x_i,y_i) to the decision surface. Consider another point (x_j,y_j). It will have another functional margin (γ_j). So γ^j is greater than γ^i.That means if any point away from the decision surface we have higher confidence in the classification of the points.So here (x_j,y_j) have higher confidence than (x_i,y_i). And a larger functional margin means more confidence in predicting the class of that point. The problem of defining the margin is as follows w and b can be arbitrarily scaled, suppose the equation of the line is 3x+2y+1=0, where $w_1=3$,$w_2=2$, and b=1. We can write the same equation as 6x+4y+2=0 another equation as 30x+20y+10=0. Here if we scaled the equation, the equation of the line is the same but the functional margin becomes larger. It means that the functional margin does not depend on the actual line, it also depends on the coefficient that we are choosing. So we need to use some normalization. Consider a set of points $S=\{(x_1,y_1),(x_2,y_2)\ldots..(x_m,y_m)\}$,the functional margin is defined as $\gamma = \min_{i=1}^{m} \gamma^i$. Among all functional margins, we choose the smallest functional margin that is the smallest functional margin of the different points is refined to be the functional margin of the set of points.

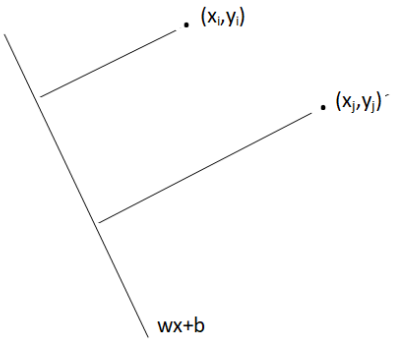


Fig 9: The solution of SVM classification problems

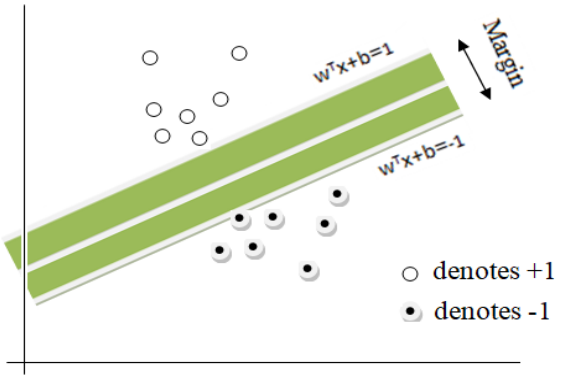


Fig 10: The SVM classification of points

Geometrical Margin: It is invariant of the scale of the equation. Consider the decision surface given by wx+b=0,w is normal to the decision surface and the unit vector normal to the decision surface is given by $\frac{w}{||w||}$ .Let us looked at the point 'P'**(fig)**has the value (a_1,a_2) and draw a normal to the decision surface meets at the point 'Q'(b_1,b_2). If we want to find the distance in the direction of the normal vector. So we can write $P=Q+\gamma\frac{w}{||w||}$, γ is the distance from Q to the decision surface. Coordinates of the point P is given by

$(a_1,a_2)=(b_1,b_2)+\gamma\frac{w}{||w||}$ ——————————(1)

$$\gamma = \frac{(w^T(a_1,a_2)+b)}{|w|}$$

$$=\frac{w^T}{||w||}.(a_1,a_2)+\frac{b}{||w||}$$

$$\gamma = y.\left(\frac{w^T}{|w|}.(a_1,a_2)+\frac{b}{|w|}\right)$$

Geometric mean ||w||=1.Geometric margin of (w,b) with respect to $S=\{(x_1,y_1),(x_2,y_2)\ldots..(xm,ym)\}$.So $\gamma = \min_{i=1}^{m} \gamma^i$ is the smallest of the geometric margin of individual points. We assume that training examples are linearly separable. The classifier with the maximum margin width is robust to outliers and thus has the strongest generalization ability.maximizing $\frac{1}{||w||}$ or minimizing w.w subject to the constraints, for all (x_i,y_i) i=1,2,…m: If (w,b) Characterizes

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

the decision surface then $\frac{\gamma}{||w||}$ is the geometric margin and we want to learn the values of (w,b) so that the geometric margin is largest subject to the constraints.

$wx_i + b \geq \gamma$ for +points

$wx_i + b \leq \gamma$ for - points

We can scale it by $\gamma = 1$ and we can maximize it $\frac{1}{||w||}$ or minimize w.w

$w^T x_i + b \geq 1$,if. $y_i = +1$

$w^T x_i + b \leq -1$,if. $y_i = -1$ for all ines

In the proposed system we provide two methods of classification, Support vector machine and multilayer perceptron. It is found that it performs better with high accuracy when the data is preprocessed and given as input. From the below analysis, Sequential minimal optimization in the Support vector machine is more effective. The support vectors are the data coordinates that are on the boundary of the margin. Mathematical functions are involved in SVM design which is frequently used to model real-world problems when we have large data set of entries The system with SVM will help for early diagnosis of heart disease for any given patient. This system when deployed can complement traditional heart disease detection systems and can help not only the doctors but also the patients. This system can act as a boon for the medical industries for the detection of heart disease with better accuracy. The dataset with 13 attributes is fed to the SVM classifier and the output is mapped into two classes cardiac and noncardiac. SVM aims to identify the best classification function to distinguish between members of the two classes from the training data. The classification performance of the heart disease dataset is discussed in the following section. This system when deployed can complement traditional heart disease detection systems and can help not only the doctors but also the patients.

In the proposed model, we used a dataset having 297 total number of patient records. A large part of records in the dataset is used for training and the rest of them are used for testing. The main difference between the dataset given as input to training and testing is that, the input we are giving to training is the data with the correct diagnosis and whereas the information on testing doesn't have the right diagnosis purposely. The Diagnosis field refers to the presence or absence of heart disease of that respective patient. It is an integer-valued field, having value 0(absence of disease) or 1(presence of disease). So that at the end of the testing process we can check the result in the output file created after testing and verify the efficiency of the proposed model in terms of accuracy. In the proposed framework two techniques for classification are given, Support vector machine and Artificial neural system. The performance of both the classification techniques is compared in terms of the time needed for the classification and accuracy of the system.

2)Artificial neural network(ANN)

Artificial neural network (ANN) is a popular machine learning algorithm that attempts to mimic how the human brain processes information It provides a flexible way to handle regression and classification problems. Artificial neural network models are a first-order mathematical approximation to the human nervous system that has been widely used to solve various nonlinear problems. It is a “connectionist” computational system.ANNs consist of single processing units called neurons. Each of the neurons in the input layer receives one of the variables on which the variables that we wish to forecast depend. The neurons of the output layer return the values of the variables that we wish to forecast There can also be a series of intermediate layers, called hidden layers. A true neural network does not follow a linear path or it exploits the non –linearity. Every layer of the NN computes a nonlinear function of the input feature vector. The input to the particular layer of neurons and the output of that layer of the neuron is an intermediate feature vector of the same or different dimension depending upon how many neurons have in that particular layer. One of the key elements of a neural network is its ability to learn and it can change its internal structure based on the information flowing through it. A neural network is not just a complex system, but a complex adaptive system, meaning it typically, this is achieved through the adjusting of weights.

Classification with perception model:The perceptron was one of the first neurally-motivated classification models. It is the easiest neural network system conceivable: a computational model of a single neuron: a computational model of a single neuron. A perceptron comprises of at least one or more inputs, a processor, and a single output. The number of outputs is normally determined by the number of classes in the data. These processing elements (PEs) include the individual scaled commitments and react to the whole input space.we have a Perceptron convergence algorithm.

1. If the output unit is correct to leave its weights alone.
2. If the output unit is incorrect outputs a zero, add the input vector to the weight vector
3. If the output unit is incorrect outputs a one, subtract the input vector to the weight

Vectorization of perceptron model

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

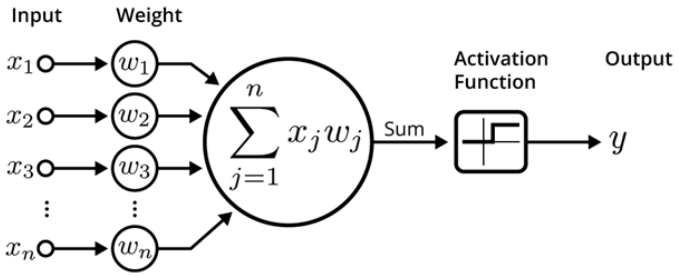


Fig 11: An illustration of an artificial neuron

$$\begin{bmatrix} x_0 \\ x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \text{X} \begin{bmatrix} w_0 & w_1 \ldots w_n \end{bmatrix} = x_0 w_0 + x_1 w_1 + .., + x_n w_n$$

$$y = \sum_{i=0}^{n} x_i w_i + b$$

$$= w^T.x$$

$$b = w_0 x0, \text{where } x_0 = 1$$

The perceptron learning rule loops and within each loop, it iterates over the data one at a time. We are going to use {−1, +1} for the labels. As each datum is examined, there are two possible cases: either $y_n \cdot w^T x_n \geq 0$ or $y_n \cdot w^T x_n < 0$. In in the former case, we're classifying datum n correctly, and in the latter case, we're getting it wrong. Note that we're multiplying the inner product by the true label to take advantage of the fact that being correct means having the same sign. If we are getting this example right, then there's no reason to update w. There are two ways to get it wrong, however. What should we do in each case? If $\text{sgn}(w^T x_n) = +1$ when $y_n = -1$, then we want the inner product to go down. If $\text{sgn}(w^T x_n) = -1$ when $y_n = +1$, then we want the inner product to go up. Of course, to make this inner product go up or down, we need to know about the sign of each entry in x_n. For example, if we want $w^T x_n$ to get bigger, and $x_n, d < 0$, then we need wd to get smaller. If our problem is a linear problem we can solve it using a single layer neuron but it is not in the linear problem or in the case of a nonlinear problem we need multiple neural networks.

Multi-Layer Perceptron: Every layer of neuron computes a nonlinear function and we have a cascade of the nonlinear function that is in f (i+1) th layer it computes the nonlinear function of $f^{(i)}$ on the feature vector which are computed by previous layer $f^{(i-1)th}$ similarly we can cascade all of them together.

The above diagram is known as a multi-layered perceptron, a network of many neurons. In the system for heart disease prediction, the multilayer perceptron architecture of the neural network is used. The system consists of two steps, in the first step 13 clinical attributes are accepted as input and then the training of the network is done with training data by the back-propagation learning algorithm. Some are input neurons and receive the inputs, some are part of what's called a "hidden" layer (as they are connected to neither the inputs nor the outputs of the network directly), and then there are the output neurons, from which we read the results. Training these networks is much more complicated. With the simple perceptron, we could easily evaluate how to change the weights according to the error. But here there are so many different connections, each in a different layer of the network. The solution to optimizing the weights of a multi-layered network is known as backpropagation. Multilayer is feed-forward neural networks trained with the standard back-propagation algorithm. It is supervised networks so they require a desired response to be trained. It learns how to transform input data into the desired response, so they are widely used for pattern classification. With one or two hidden layers, they can approximate virtually any input-output map. It has been shown to approximate the performance of optimal statistical classifiers in difficult problems.

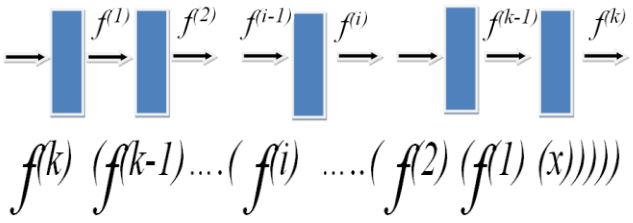


Fig 13: Mathematical illustration of MLP

The output of the network is generated in the same manner as a perceptron. The inputs multiplied by the weights are summed and fed forward through the network. The difference here is that they pass through additional layers of

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

neurons before reaching the output. Training the network (i.e. adjusting the weights) also involves taking the error (desired result - guess). The error, however, must be fed back through the network. The final error ultimately adjusts the weights of all the connections. The most popular static network in the multilayer. The multilayer is trained with error-correction learning, which is appropriate here because the desired multilayer response is the arteriographic result and as such known. Error correction learning works in the following way from the system response at neuron j at iteration t, yj (t), and the desired response dj (t)for given input pattern an instantaneous error ej (t)is defined by $ej(t) = dj(t) - yj(t)$ (1) Using the theory of gradient descent learning, each weight in the network can be adapted by correcting the present valueof the weight with a term that is proportional to the present input and error at the weight, i.e. $wjk(t + 1) = wjk(t) + \eta\delta j(t)xk(t)$ (2) The $\eta(t)$ is the learning rate parameter.The $wjk(t)$ is the weight connecting the output of neuron k to the input neuron j at iteration t. The local error$\delta j(t)$ can be computed as a weighted sum of errors at the internal neurons. The BP algorithm has served as a useful methodology to train multilayer perceptron for a wide range of applications. The BP network calculates the difference between real and predicted values, which is circulated from output nodes backward to nodes in the previous layer.

D. Hybrid Approaches

In this study, we include a hybrid feature selection model that produces an enhanced performance level with high accuracy.This model includes a Support vector machine with a multi layer perceptron.The data may require some pre-processing the attribute values may be missing, the data contain outliers, the data is not in a suitable format or the values appear inconsistent and so on. After the pre-processing Construct a Decision Tree on a given dataset and construct a subset of features. Train the feature subset with Support vectormachine and find a subset with highest accuracy. Finally, Learn and classify the Hybrid Model with selected features. The steps are followed. Table-3 shows a Comparison of various models with the hybrid model.

Hybrid algorithm

Step-1: Pre-process the data set to remove the duplicate, missing and unknown data.

Step-2: Construct a Decision Tree on a given Dataset.

Step-3: Traverse through the branches of the tree to construct a subset of features.

Step-4: Train SVM with each subset of features and find the subset with the highest accuracy.

Step-5: Learn and Classify the Hybrid Model (SVM +MLP) with selected features

IV. PERFOMANCE MEASURES

Several standard performance measures such as accuracy, precision, and error in classification have been considered for the computation of the performance efficiency of this model. To check the performance of the classifiers, various performance evaluation metrics were used in this project work. We used the confusion matrix and ROC for evaluation. In the confusion matrix, every observation in the testing set is predicted in exactly one box. With just two classes, there are four possible results with the classification. The upper left and lower right quadrants represent the correct actions. The remaining two quadrants are incorrect actions. The performance of classification could be determined by associating costs with each of the quadrants. Classification accuracy is normally determined by deciding the percentage of tuples placed in the correct class. This ignores the fact that there likewise also be an expense related to an inaccurate assignment to an inappropriate class. This perhaps should also be determined.

Confusion Matrix: is a table that is regularly used to portray the performance of a classification model on a lot of test information for which the genuine values are known.

TP (True Positive):Predict1 as 1.

TN (True Negative):predict 0 as 0.

FP (False Positive):Predict 0 as 1.

FN (False Negative) :Predict 1 as 0.

Accuracy: It is the number of samples predicted correctly out of total number of samples .ie (TN+TP) / (TN+TP+FN+FP).

Precision: Of the samples classified target value ,how many are actually target i.e. TP/TP+FN.

Sensitivity: of the samples that are actually target value, how many are classified as target ie TN/TN+FP.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

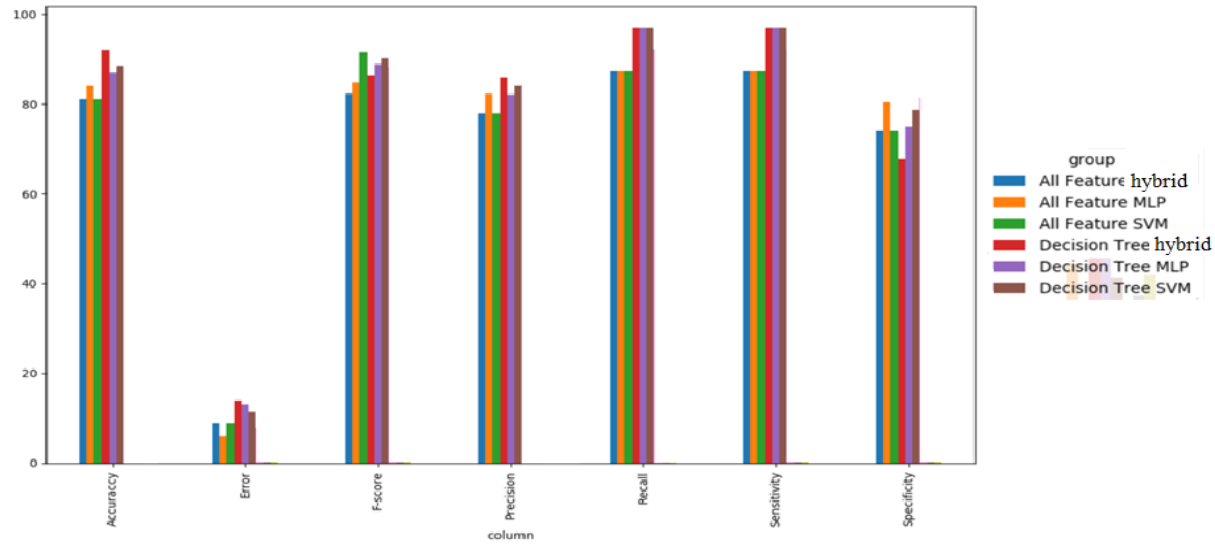
Receiver operating characteristic curve (ROC): ROC curve shows the relationship between false positives and true positives. A ROC curve was originally used in the communications area to examine false alarm rates. It has also been used in information retrieval to examine fallout (percentage of retrieved that are not relevant) versus recall (percentage of retrieved that are relevant). InIn the OC curve, the horizontal axis has the percentage of false positives and the vertical axis has the percentage of true positives for a database test. At the beginning of evaluating a sample, there is none of either category, while at the end there is 100 percent of each. When evaluating the results for a specific sample, the curve looks like a jagged stair-step, as seen in Figure, as each new tuple is either a false positive or true positive. A progressively smoothed variant of the OC curve can likewise be obtained. A confusion matrix outlines the accuracy of the answer to a classification problem.

V. EXPERIMENT AND RESULTS

This section of the paper involved the discussion on the classification models and outcomes from different perspectives. First, we checked the performance of the following machine learning algorithms SVM, MLP, and hybrid(SVM+MLP) on full features. In the second, we used a decision tree feature selection algorithm for important feature subset selection and apply these subsets of features to the above-mentioned machine learning algorithms.To check the performance of classifiers performance evaluation metrics were applied.The experimental results in Table 2 show that the selected set of attributes using information gain, hybrid , and the appropriate training set gives a better performance when compared to the existing method. Next, we get the performance of the machine learning algorithm SVM, MLP, and hybrid(SVM+MLP) with include feature subset selection using decision tree and get the presentation of the machine learning algorithm SVM, MLP, and hybrid(SVM+MLP) with all features.

TABLE 2. Result of various models with proposed model.

lassification Methods	Feature selection Methods	Accuracy	Precision	Sensitivity	Specificity	F-Score	Error
SVM	**All features**	81.5	77.84	87.27	73.93	**91.43**	8.85
MLP		84.1	82.29	87.27	80.36	84.71	**5.9**
Hybrid(SVM +MLP)		81.15	77.84	87.27	73.93	82,29	8.85
SVM	**Decision tree**	84.1	81.81	89.24	78.04	85.36	10.9
MLP		87.21	84.44	92.12	81.43	88.12	7.79
Hybrid(SVM +MLP)		**87.21**	**84.44**	**92.12**	**81.43**	88.12	7.79



V. CONCLUSION AND FUTURE WORK

This work will be useful in identifying possible patients who may suffer from heart disease in the coming years. The comparison of the accuracy rates of all classifiers can be seen in Table 7. Overall, there is little fluctuation in the accuracy rates of all classifiers. Inter-Related feature selection can outperform traditional feature selection methods. Hybrid classifiers shows the best performance measures, their accuracy rates are the highest compared to other techniques. Here we use decision tree based feature selection technique. The results also demonstrate that the reduced feature subset can have better prediction performance compared to the original set of attributes. Our proposed system tries to extract these interrelationships between features and select the best feature subset and thereby improve the performance of the Machine Learning algorithm. The performance of feature selection techniques depends on the type of dataset that we have taken for experimenting. Also, we can observe that the hybrid methods were more stable than the other individual techniques. The results indicated that the system can be useful and helpful for the doctors for timely diagnoses the chances of a heart attack in a patient.

There are some shortcomings to be a further study in the future. The focus of the latter research will be on the optoimal feature selection to verify its classification performance. At the same time, our training sample data is too small, it is necessary to increase more data sets to test better. As far as UCI dataset concerns, the dataset needs to be amplified.

REFERENCES

Bibliography and Citations

[1] Jabbar MA, Chandra P, Deekshatulu BL. Cluster-based association rule mining for heart attack prediction. Journal of Theoretical and Applied Information Technology. 2011; 32(2):197–201.

[2] Sudha A, Gayathiri P, Jaisankar N. Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications. 2012; 43(14):26–31.

[3] Amin SU, Agarwal K, Beg R. Genetic neural network-based data mining in the prediction of heart disease using risk factor. Proceeding of IEEE Conference on Information and Communication Technologies (ICT); 2013 Apr. p. 1227– 31.

[4] Deepika N, Chandrashekar K. Association rule for classification of Heart Attack Patients. International Journal of Advanced Engineering Science and Technologies. 2011; 11(2):253–57.

[5] Sellappan Palaniappan and Rafiah Awang (2008): Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244- 1968- 5/08/ IEEE.

[6] Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm M. ANBARASI, E. ANUPRIYA,N.CH.S.N.IYENGAR.-2012

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 7, July 2020||

[7] Feature Selection using Artificial Bee Colony for Cardiovascular Disease Classification B.Subanya, Dr.R.R.Rajalaxmi-2014
[8] "Analysis of data mining techniques for heart disease prediction," 2016-2017, M. Sultana, A. Haider, and M. S. Uddin
[9] D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," J.Theor. Appl. Inf. Technol., 2009
[10] N. Bhatia and C. Author, "Survey of Nearest Neighbor Techniques," IJCSIS) Int. J. Comput. Sci. Inf. Secur., vol. 8, no. 2, pp. 302–305, 2010.
[11] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P.Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, 2013.
[12] Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data Simon Fong, Raymond Wong, and Athanasios V. Vasilakos, Senior Member, IEEE
[13] T.Peter and K.Sonausundaram "An empirical study on prediction of heart disease using classification data minings techniques" in IEEE International Conference and Management-2012
[14] A. Khemphila and V.Boonjing "Comparing Performance of logistic regression decision trees and neural networks for classifying heart disease patients "in International Conference on Computer Information systems and Industrial Management Applications"
[15] Data Mining Techniques on Risk Prediction: Heart Disease, G. Purusothaman* and P. Krishnakumari, Indian Journal of Science and Technology-2015
[16] A Fuzzy Rule-based Approach to Predict Risk Level of Heart Disease, By Kantesh Kumar Oad & Xu Delhi-2014
[17] [18]The Best Two Independent Measurements Are Not the Two Best"-1974, THOMAS. M. COVER
[18] Enhanced Prediction of Heart Disease by Genetic Algorithm and RBF Network-2015, A. Durga Devi
[19] E. J. Benjamin, P. Muntner, and et al. Alonso, Alvaro, ―Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association,‖ Circulation, vol. 139, no. 10, 2019.
[20] M. Ramaraj and T. A. Selvadoss, ―A Comparative Study of CN2 Rule and SVM Algorithm and Prediction of Heart Disease Datasets Using Clustering Algorithms,‖ Netw. Complex Syst., vol. 3, no. 10, pp. 1–6, 2013.
[21] A. Gavhane, G. Kokkula, I. Pandya, and P. K. Devadkar, ―Prediction of Heart Disease Using Machine Learning,‖ in Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018, 2018, pp. 1275–1278.
[22] C.-S. Lee and M.-H. Wang, ―A fuzzy expert system for diabetes decision support application.,‖ IEEE Trans. Syst. MAN, Cybern. B Cybern., vol. 41, no. 1, pp. 139–153, 2011.
[23] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.
[24] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," IEEE transactions on pattern analysis and machine intelligence, vol. 24, pp. 301-312, 2002.
[25] H. M. Harb and M. A. Moustafa, "Selecting an optimal subset of features for student performance model," Int J Comput Sci, p. 5,2012.
[26] A. Figueira, "Predicting Grades by Principal Component Analysis: A Data Mining Approach to Learning Analytics," in Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on, 2016, pp. 465-467.
[27] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," International Journal of Database Theory and Application, vol. 9, pp. 119-136, 2016.
[28] K. Patel, J. Vala, and J. Pandya, "Comparison of various classification algorithms on iris datasets using WEKA," Int. J.Adv. Eng. Res. Dev. (IJAERD), vol. 1, 2014.
[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update,"ACM SIGKDD explorations newsletter, vol. 11, pp. 10-18,2009.

Webliography

[30] https://en.wikipedia.org/wiki/Feature_selection#/media/File: created on 7 march 2019 and accessed on 13 September 2019
[31] https://web.stanford.edu › class › handouts › CS276: Information Retrieval and Web Search Christopher Manning and Pandu Nayak accessed on 26 September 2019
[32] https://images.app.goo.gl/7SGi2Y6sHwwiFGNa9 created on 13 march 2029 and accessed on 10 April 2020