UE20CS312-DATA ANALYTICS
PES UNIVERSITY,CSE DEPT.

HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

TEAM COCONUTWATER

RIA TREZA SERRAO(PES2UG20CS453)
POOJA H(PES2UG20CS448)
AYUSH DUDHE(PES2UG20CS420)

TABLE OF CONTENTS:

Abstract: Heart disease has been one of the ruling causes for death for quite some time now. About 31% of all deaths every year in the world take place as a result of cardiovascular diseases. A majority of the patients remain uninformed of their symptoms until quite late while others find it difficult to minimize the effects of risk factors that cause heart diseases. Machine Learning Algorithms have the potential in producing results with a high level of correctness thereby preventing the arrival of heart diseases in many patients and reducing the impact in the ones that are already affected by such diseases. It has helped medical researchers and doctors all over the world in recognising patterns in the patients resulting in early detections of heart diseases. **Keywords: Cardiovascular Diseases (CVDs); Support Vector Machine (SVM); K- Nearest Neighbor (KNN); Naive Bayes (NB); Random Forest (RF); Logistic Regression (LR); Machine Learning (ML); Prediction Model**

I. INTRODUCTION.

PAPER 1 AND 2

An expert system is defined as a software that attempts to reproduce the performance of one or more human experts, most commonly in a specific problem domain. It basically uses an inference engine connected to the knowledge base. A wide variety of methods can be used to simulate the performance of the expert however common to most or all are:

• the creation of a so-called “knowledgebase” which uses some knowledge representation formalism to capture the Subject Matter Expert’s (SME) knowledge; a process of gathering that knowledge from the SME and codifying it according to the formalism, which is called knowledge engineering. The paper is organized as following, in Section II, a brief overview on previous related works and in section III, introduction of Support Vector Machine and Feedforward Backpropagation neural network techniques described. Section IV, the proposed methodology and preparing Data for underlying neural network. Section V, Experimental analysis and how the coding is done with patients as well as medicine data is described. Section VI, Discussion and result of first five patients medicine given by the expert system and is compared. Finally, we concluded this paper in Section VII.

PAPER 3 AND 4

Cardiovascular diseases are very common these days, they describe a range of conditions that could affect your heart. World health organization estimates that 17.9 million global deaths from CVDS. It is the primary reason of deaths in adults. It recognizes who all are having any symptoms of heart disease such as chest pain or high blood pressure and can help in diagnosing disease with less medical tests and effective treatments, so that they can be cured accordingly. This project focuses on mainly three data mining techniques namely: (1) Logistic regression, (2) KNN and (3) Random Forest Classifier. The accuracy of our project is 87.5% for which is better than previous system where only one data mining technique is used. So, using more data mining techniques increased the HDPS accuracy and efficiency. Logistic regression falls under the category of supervised learning. The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases

PAPER 5 AND 6

Machine Learning being a subgroup of Artificial Intelligence has been one of the most important tools in recent times. It has helped further the progress of many healthcare industries. With such a tool, medical professionals and researchers have been able to diagnose and detect diseases with much accurate precision thus contributing to saving as many lives as possible . Majority of the heart related diseases can be prevented if people focus on their physical activity like walking,running,lifting weights and many more,have a balanced and nutritious diet,people should try to eat at least 5 portions of a 5 portions of a variety of fruits and vegetables every day. Then avoid consumption of products such as tobacco and alcohol that have a direct impact on the heart and thus deteriorating their health and prevent early deaths. Therefore, it is of prime importance that these diseases are detected as early as possible, so that its effects can be managed with medical advice,modern science and technologies and medicines. In this review paper, we are going to review various advancements and recent works that have been done using Machine learning in the prediction of heart diseases. Heart diseases are a result of a multitude of aspects that can influence the cardiovascular health of an individual such as age, blood sugar, blood pressure, cholesterol etc.

II. CHOOSING THE RIGHT DATASET.

PAPER 1 AND 2

About 300 patient's data were collected from the Sahara Hospital, Aurangabad. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not.

PAPER 3 AND 4

The UCI dataset's open-source dataset registry is used in analysis. It has numerous disease-related databases. For academic purposes, these freely accessible databases are used. In this model, the UCI dataset of heart disease is used. It is the dataset existence category data category. With 303 instances and 75 properties, it is a multivariate dataset.

PAPER 5 AND 6

For paper 5, First step for prediction system is data collection, data collect from net [12] after that used data wrangling for data cleaning and then determining about the training and testing dataset. In this project we have used 80% training dataset and 20% dataset used as testing dataset the system. After Cleaning, In this data set There are 14 columns and 383 rows created by code

For the 6th paper, This dataset contains 13 medical attributes of 304 patients that helps us detect if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 14 columns, where each row corresponds to a single record

III. EDA AND VISUALIZATION

PAPER 1 AND 2

The data is collected from daily OPD session while doctor examining the patients. The symptoms and information about patients details like Previous History(p1), Present History(p2), Personnel History(p3), Physical Examination(p4), Cardio Vascular System(CVS), Respiratory Rate(RS), Per Abdomen(PA), Central Nervous system(CVS), ECG and Blood Investigation(BI).

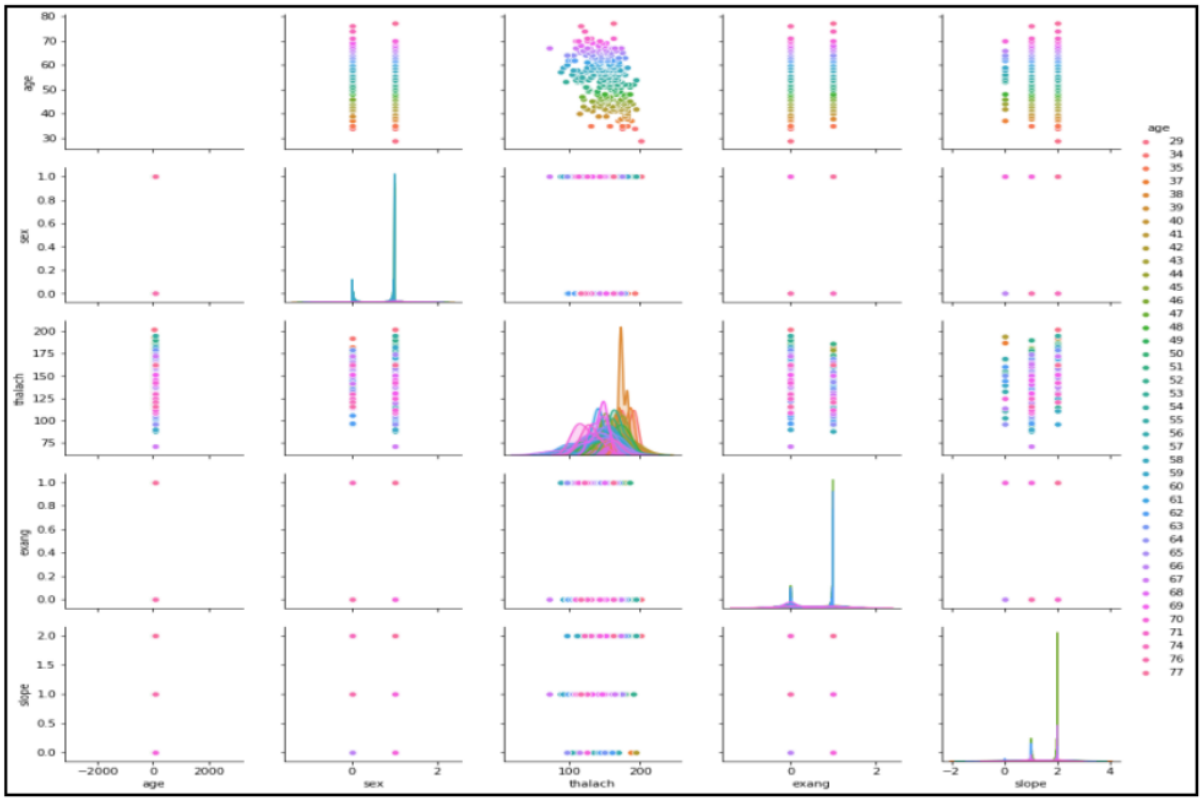
The main point is ECG from which the patient can easily diagnose whether the patient is having heart problem or not. The code is given to each symptoms, physical examination parameter or diseases in each sub-sheet for experimental work. On this data some pre-processing i.e. normalization, coding and decoding methods are applied for the expected output.

PAPER 3 AND 4

EDA : As at this preprocessing level, this is the important step; meaningful data is derived from the dataset of heart disease. This phase is compulsory because the raw data is not reliable and unfinished, so pre-processing is performed for more steps to render ready raw data. In this approach, the UCI Heart Disease Dataset, the data contains 75 attributes, and during pre-processing, 14 [6] attributes are extracted to understand the nature of patients' health better. The extricated 14 attributes include BP, sex, heart rate, chest, and others. The attribute's values are normalized and converted into numerical form. The quality of data plays an essential role, and the most carefully depicted thing to be. For this research, data cleaning has improved the quality of our dataset. Data cleaning is necessary as it removes unnecessary or irrelevant attributes of data from the dataset. This step of the model will make the dataset more precise and exact. In this part of approach, the Null (NaN) values are removed from the dataset to make it more useful as these values decrease the productivity of the algorithm [7]. At the data cleaning stage, the dataset is also normalized to not have any ambiguity after cleaning

VISUALIZATION: The dataset is in tabular form and it is hard to observe and understand the data in this or any other form. So the data is visualized Graphically below. It helps in knowing the trend of the data. Data visualization in this approach is a graphical representation of the data. In this analysis, using bar charts and scatter plots, the cleaned data acquired by pre-processing is visualized. It illustrates the actions of data attributes. It makes it easy to grasp the attribute's complicated relationship by graphical representation.

As mentioned above, this visualization plays a crucial role in data exploration. The various parameters of the dataset plotted dependent on the age of patients as shown below Using this Pairplot method from the Seaborn library of Python, it can be inferred the cardiac disorder, aka. Cardiovascular disorders rely primarily on the patient's age.

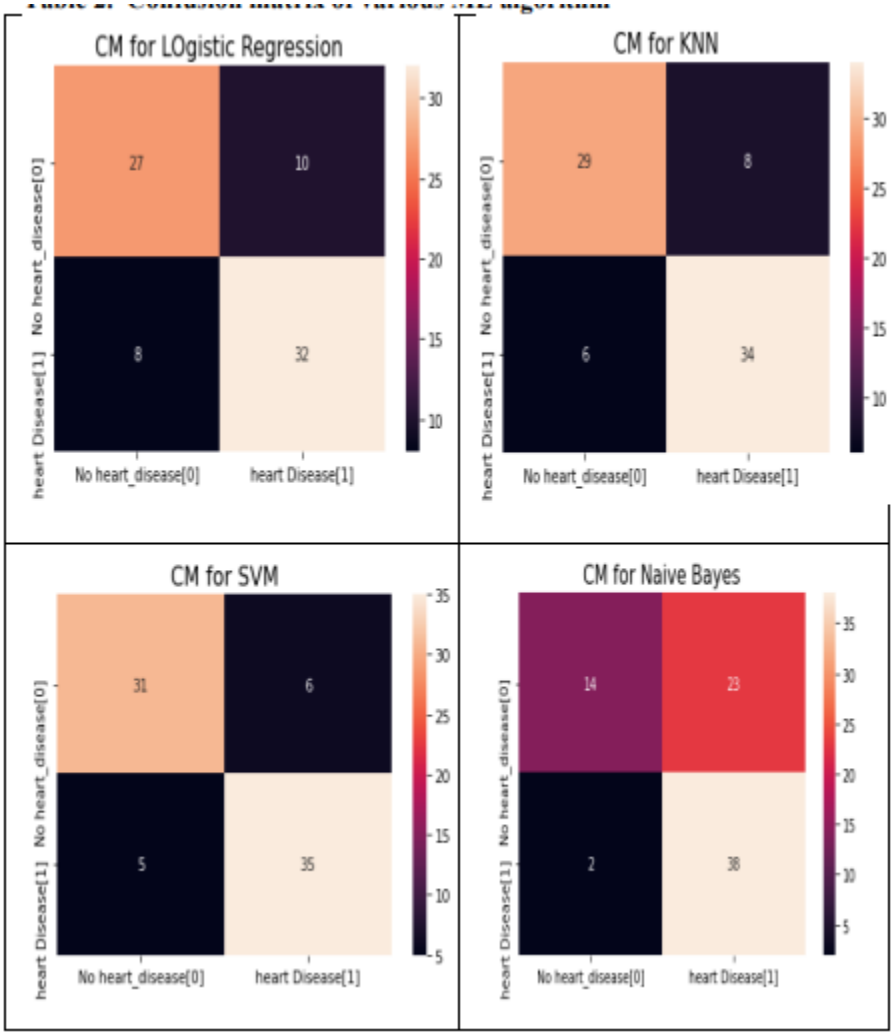


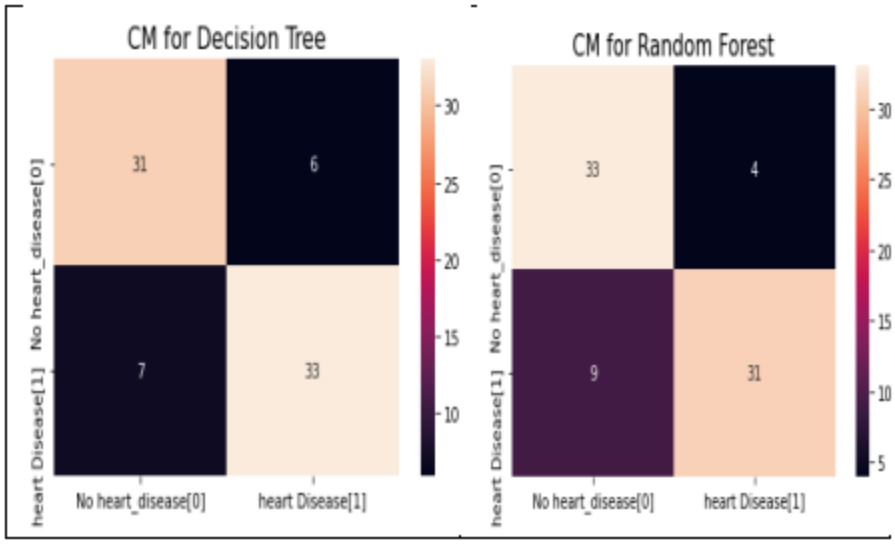
PAPER 5 AND 6

The confusion matrix is created by various classifier and it includes expected and real classifications information. The confusion matrix provides analysis to judge the effectiveness of proposed methodology [15]. Where, The number of actual negative cases in the data = Condition Negative (N) Condition Negative (N) = Total number of negative cases Condition Positive (P) = Total number of positive cases True Positive (TP) = number of correct positive prediction True Negative (TN) = number of correct negative prediction False Positive (FP) = Type I Error, No. of incorrect positive prediction False Negative (FN) = Type II Error, No. of incorrect negative prediction Accuracy: The accuracy of classification process is based on correct and incorrect predictions

Accuracy (ACC)

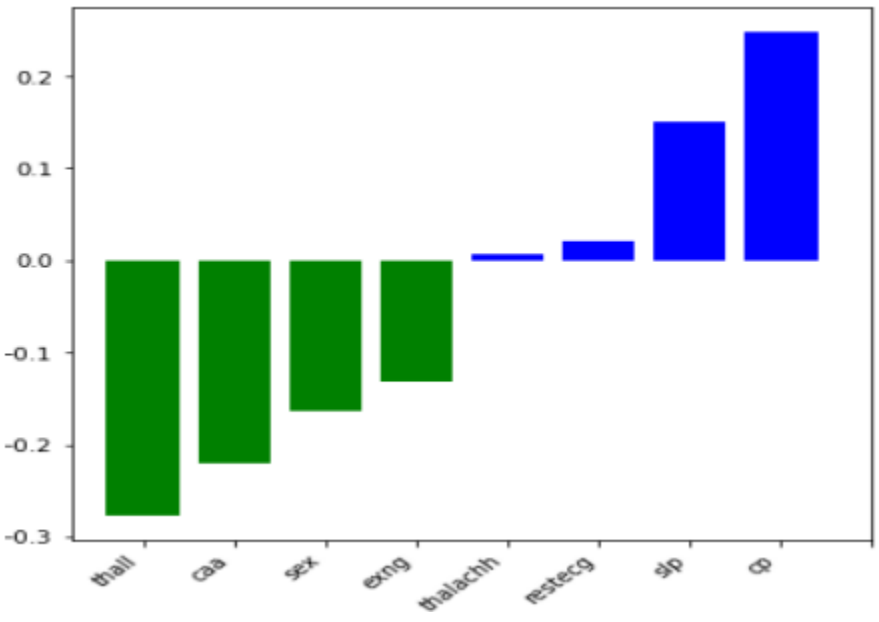
$$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN}$$





Features Selection

Features selection acts a indeed significant role particularly when functioning with large data set in machine learning. It can lead to better classification. The main features are shown in Fig.using SVM. Green features shows the negative factors and blue corresponding to positive ones.



IV. REVIEWING THE LITERATURE REPORTS

PAPER 1 AND 2

1. Decision Tree

A Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences including chance event outcomes and utility. It is one of the ways to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis to help and identify a strategy that will most likely reach the goal. It is also a popular tool in machine learning. A Decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one. Finally by following these rules, appropriate conclusions can be reached

2. K-means Algorithm

K-means creates k groups from a set of given objects so that the members of a group are more similar. Other than specifying the number of clusters, k-means also “learns” the clusters on its own without any information about which cluster a particular observation should belong to. That’s why k-means can be called as semi-supervised learning method. K-means is specially effective over large datasets.

3. ID3 Algorithm

The ID3 algorithm (Quinlan86) is a Decision tree building algorithm which determines the classification of objects by testing the values of the properties. It builds the tree in a top down fashion, starting from a set of objects and the specification of properties. At each node of the tree, a property is tested and the results used to partition the object at that point are set. This process is recursively continued till the set in a given sub tree is homogeneous with respect to the classification criteria. Then it becomes a leaf node. At each node, information gain is maximized and entropy is minimized. In simpler words, that property is tested which divides the candidate set in the most homogeneous subsets.

4. Support Vector Machine(SVM)

It is a supervised learning method which classifies data into two classes over a hyper plane. Support vector machine performs a similar task like C4.5 except that it doesn't use Decision trees at all. Support vector machine attempts to maximize the margin (distance between the hyper plane and the two closest data points from each respective class) to decrease any chance of misclassification. Some popular implementations of support vector machine are scikit-learn, MATLAB and of LIBSVM.

5. Naive Bayes(NB)

It is a simple technique for constructing classifiers. It is a probabilistic classifier based on Bayes' theorem. All Naive Bayes classifiers assume that the value of any particular feature is independent of the value of any other feature, given the class variable. Bayes theorem is given as follows: $P(C|X) = P(X|C) * P(C)/P(X)$, where X is the data tuple and C is the class such that P(X) is constant for all classes. Though it assumes an unrealistic condition that attribute values are conditionally independent, it performs surprisingly well on large datasets where this condition is assumed and holds.

PAPER 3 AND 4

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

Recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. (Beyene & Kamat, 2018) suggested Heart Disease Prediction System using Data Mining Techniques. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Chala Beyene recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time.
The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computed using WEKA software. (Soni, Ansari, & Sharma, 2011) proposed to use non- linear classification algorithm for heart disease prediction.

It is proposed to use big data tools such as Hadoop Distributed File System (HDFS), Map reduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM. (Science & Faculty, 2009) suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques.

By using data mining techniques, the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease. (Sai & Reddy, 2017) proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc.

(A & Naik, 2016) recommended to develop the prediction system which will diagnosis the heart disease from patient's medical data set. 13 risk factors of input attributes have considered to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed. He used k-means and naïve Bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of

300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

PAPER 5 AND 6:

Naïve bayes[NB]

NB is a supervise classification algorithm. It is a simple technique using Bayes theorem. To get the probability, mathematical concept is used with the support of bayes theorem. The correlation is neither related to each other nor predictor to one another. All parameters work autonomously for getting the maximum probability.

$$P(x/y) = \frac{p(y/x) \times p(x)}{p(y)}$$

P(x/y) = (1) Where p(x)=Class predictor probability, p(y)= Predictor Probability, P(x/y)= Posterior probability, P(y/x)=possibility, probability of predictor

Decision Tree[DT]

DT is an algorithm that classifies parameters in categorical form in spite of arithmetic data. Tree like structure is created by DT. Many large data set related to medical have analyzed by DT due to its simple nature. It works on tree node for analysis. Leaf Node: Signify the solution of every Test Interior Node: Handle numerous element Main Node[Root Node]: Other nodes work based on main node Data is to be divided into two or more parallel set by applying this algorithm. Then entropy of each parameter is calculated. After that divide the data with predictor having extreme information gain that means minimum entropy

$$\text{Entropy} = -\sum_{i,j=0}^{Ng-1} p(i,j)\log(p(i,j))$$

Random forest [RF]

RF algorithm is supervised primarily based learning. It is used as classifier in numerous fields. By using this more trees makes a forest. If we have more number of trees then it create higher accuracy. It is also used for regression task. but it accomplish well when classify the task. And may overwhelmed misplaced values. There are three approach of RF: Forest RC(Random Blend) Forest RI(Random input) And combination of RC and RI

Logistic regression [LR]

LR is the supervised ML learning method. It is established on the association between dependent and independent variable as seen in Fig.A variable “a” and “b” are dependent

variable and independent variable and relation between them is shown by equation of line which is linear in nature that why this approach is called linear regression. b a Fig.A Relation between a and b It gives a relation equation to predict a dependent variable value “b” based on a independent variable value “a” as we can see in the Fig.A so it is concluded that linear regression technique give the linear relationship between a(input) and b(output).

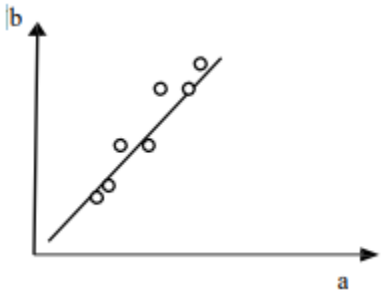


Fig.A:Relation between a and b

Support Vector Machine SVM

Support Vector Machine SVM is one type of ML method that work on the conception of hyper plan. It is used to find a hyper plan in n dimensional space, using this data point can be classified specifically [13]. (Xa, Ya) is training sample of data set where a=1,2,3,…….n and Ya is the target vector and Xa is the ith vector. Hyper plan quantity select the variety of support vector such as example if a line is used as hyper plan then method is called linear support vector. Y X Fig.B Linear Regression

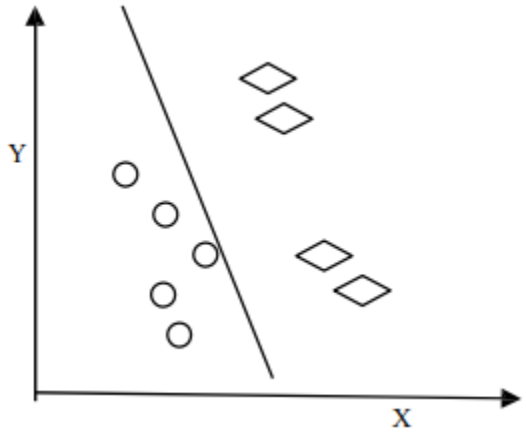


Fig.B:Linear Regession

K-nearest Neighbor [KNN]

KNN is a classification algorithm that belongs to supervised learning. It categorizes the entity that reliant on nearest neighbor. KNN could be a wide applied methodology used as a classifier and regression in numerous field like image process, data processing, pattern recognition and different applications. The output result of the algorithmic program depends on K nearest neighbor class that enforced by finding K- variety of coaching points nearest to the specified character and contemplate the votes among the K object. The algorithmic program is incredibly easy. However, is capable of learning highly-complex non-linear call boundaries and regression functions [14]. The intuition of KNN that similar instances ought to have similar category labels (in classification) or similar target values (regression). On the drawback, the algorithmic program is computationally high-priced, and is vulnerable to over fitting.

V. SUMMARIZING THE LITERATURE REPORTS

PAPER 1 AND 2

Medical prescription for heart disease patient SVM networks and FFBP NN both are examples of nonlinear layered feed-forward networks and they are universal approximations. The basic comparison of SVM and FFBP NN for the medical prescription for heart disease patient presented in table 14 and table 15.

1) In both the NN model 250 data samples has given as input.

2) The FFNN model takes 10,000 epochs to train it while SVM NN model takes only 250 epochs for the training of the model.

3) If the training performance error is compared FFBP NN and SVM are both gives less result.

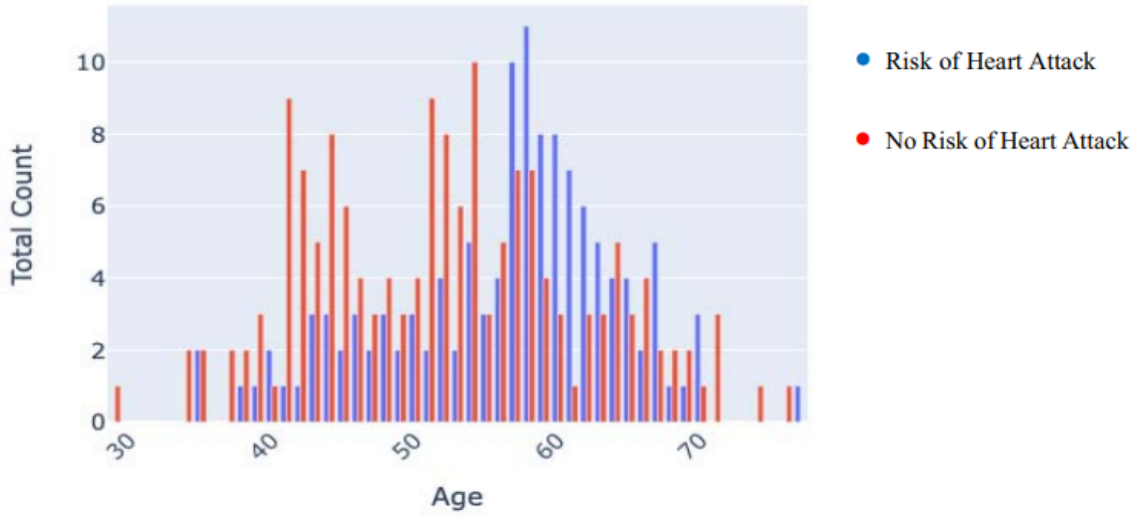
4) The time taken by FFBP NN to train the model is near about 40-45 minutes while SVM NN model takes only 5-10 seconds. So Medicines given by the expert system using FFBP and SVM is not producing the appropriate result.

PAPER 3 AND 4

From these results we can see that although most of the researchers are using different algorithms such as SVC, Decision tree for the detection of patients diagnosed with Heart disease, KNN, Random Forest Classifier and Logistic regression yield a better result to out rule them.

The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 94.5% which is greater from previous researches. So, we summarize that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease. The following figure' shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain

Showing the risk on the basis of their age



PAPER 5 AND 6

Majority of researchers have used the Cleveland Heart Disease Dataset available from the UCI repository containing 76 attributes and 303 instances, of which only 14 attributes are used due to missing values [11]. There are huge benefits to having feature selection methods so as to minimize the number of attributes that one has to use in order to build an accurate model by checking the correlation between various attributes and their impact on the accuracy of the models. It can be seen from various research papers in the field that KNN and Neural Network works quite accurately in most cases for the prediction of heart diseases.

VI. MAKING A FINAL REPORT

PAPER 1 AND 2

In this paper, around 300 patient's information is collected from Sahara Hospital, under supervision of Dr. Abdul Jabbar, (MD Medicine) Sahara Hospital, Roshan Gate, Aurangabad. The collected information is coded, normalized and entered into 13 different excel sub-sheets. All the patients data is trained by using SVM and FFBP. Around 50 samples were tested with these two techniques. If the more data set is used for the training the NN model gives more robust results. The analysis model by using SVM and FFBPof ANN gives less appropriate result for medical prescription for heart disease patient. However, there are several techniques that can improve the speed and performance of the back propagation algorithm, weight initialization, use of momentum and adaptive learning rate. It is found that the result of testing data by using SVM and FFBP is not satisfactory as per the result verified by the doctor. In future, this work may be extend using Radial Basis Function or Regression technique to improve the accuracy and to improve the performance of the expert system.

PAPER 3 AND 4

The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients.

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health's diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

PAPER 5 AND 6

Heart acts a major role in corporeal organism. The diseases of heart wants more perfection and exactness for diagnose and analyses.In real time heart diseases may not be detect in early stage. This need further analysis. In proposed work, an accurate and early heart diseases prediction is presented by using data set of heart diseases .The presented methodology requires various ML algorithms. The analysis is carried out based on Confusion matrix and comparing accuracy among them and get SVM is finest algorithm. Thus the efficacy of presented work has been verified. This technique may be used as an support for early and accurate prediction of heart disease. There are many more ML algorithms that can be used for finest exploration and for earlier prediction of heart diseases for the upcoming possibility. This needs further diagnosis.

VII. REFERENCES

1) https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.2431&rep=rep1&type=pdf
2)
3) A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–281. https://doi.org/10.15680/IJIRSET.2016.0505545
4) https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf