Heart Diseases Prediction using Data Mining Techniques: A Literature Survey

Ekta M.Vyas[1], Asst.Prof. Saket J Swarndeep[2]

[1]Department of Computer Engineering, L.J Institute of Engineering & Technology (Gujarat Technology University), Ahmedabad, Gujarat, India
[2]Assistant Professor, L.J Institute of Engineering & Technology (Gujarat Technology University), Ahmedabad, Gujarat, India

---***---

Abstract: *Health care is very important in our life. Nowadays, heart diseases are common reason of death in the world. According to WHO, each year 17.9 million people die from heart diseases i.e., 31% deaths worldwide. Out of them 17 million deaths occurs under the age of 70 due to heart diseases. The reason of this can be address by behavioural factors like unhealthy diet, use of alcohol, tobacco etc. So we need to improve prediction system. For prediction many algorithm are used like SVM, Naïve Bayes classification, KNN, K-means etc. Here we can find different parameters to predict heart diseases. It can be helpful to reduce heart diseases. With use of data mining techniques we can get data from different source and then apply classification. By this technique we get better accuracy. The purpose of using this method is to get better performance of heart diseases prediction. For this prediction age, sex, blood pressure(BP), obesity, cholesterol, etc. are used.*

Key Words: Data mining, Heart diseases, Prediction, Classification, Naïve Bayes algorithm, Support Vector Machine(SVM).

1. INTRODUCTION

Data mining is the process which analysing the "knowledge discovery process in database(KDD)".Data mining is the process of collecting information and used it for workable pattern, which is applicable in many gadget. There are many usage of data mining techniques. Prediction is one of them. At the end, we get useful details from that data. The following diagram shows the knowledge discovery process of data.

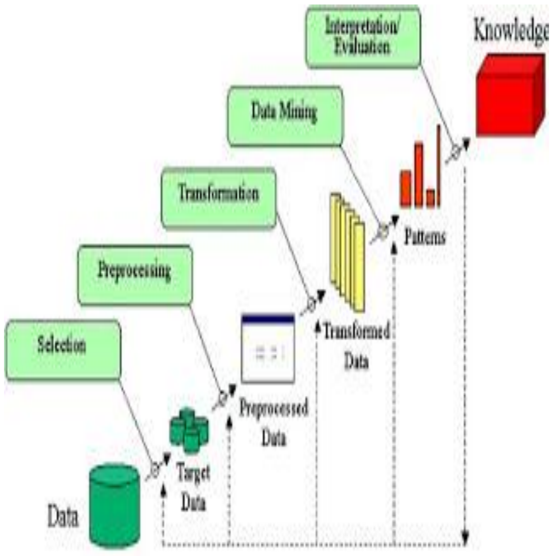

Fig-1: Knowledge discovery process using data mining [25].

Steps for KDD process.

First, Get the information about data. Select a dataset and create target dataset. And following steps are performed.

1. Data Cleaning: Remove noise or unusable data.

2. Data Integration: Combine multiple data source.

3. Data Selection: Relevent data are retrieved from database.

4. Data Transformation: Transform the data into appropriate pattern according to task.

5. Data mining: Select the process from classification, regression, clustering etc.to extract pattern.

6. Pattern Evaluation: This step evaluate the pattern.

7. Knowledge Representation: get the discovered knowledge.

Following are the types of data mining algorithms.

- Classification
- Regression
- Prediction
- Association
- Clustering
- Time series analysis
- Summarization

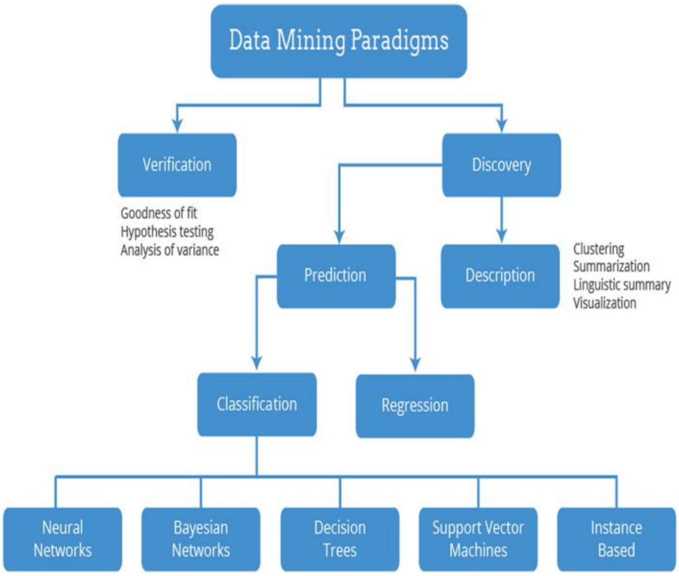

Fig-2: Basic schematic of data mining paradigm[14]

Classification: It is process in which models are finding to describes and differentiate data classes. It derived according to training dataset. It is used for prediction of unknown class label. To classify the data classifier is used. The goal of classification is to give prediction accurately. Decision tree(DT), Support vector machine (SVM), Naïve Bayes, Neural network, Genetic algorithm, Fuzzy algorithm are different classification techniques.

Classifiers are like age, BP, Cholestrol, etc for heart diseases prediction. If classifier needs to predict object it has many classes. Classification is part of data mining and machine learning techniques.

2. LITERATURE REVIEW

To manage the risk of heart diseases affecting millions of people around the world. This paper tried to analyse heart disease dataset using important types of data mining techniques in order to create a 100% accurate model based on datamining algorithm. The results obtained can be a key for gaining insights from the dataset, forecasting the heart diseases status of the new patients and get good techniques for improving the accuracy, efficiency, and quality of the care processes for heart disease.

This paper has presented a naïve byes with SVM and implemented for the prediction of heart disease. This paper provides a systematic scheme for the heart diseases, and the relevant healthcare data is created by the use of UCI Repository dataset. Supervised learning is used for prediction. In this learning they used hybrid approach containing genetic algorithm and decision tree called as ensemble classifier, for better result. They give age, sex, BP, etc. risk factor as input of first stage. This data is preprocessed. Then the out of pre-processed data is given to ensemble classifier. Here, the features are initialized through decision tree and fitness is evaluated via genetic algorithm[1].Output of this classifier is give as input to find the type of heart dieases. Disicion tree contains training and testing phases. In training phase, the classifier gives the classification rate using number of DT. Classifier uses, the random optimized algorithm to give best tree model as output, in bagging it selects the random data replacing training data. Then CART(classification and regression) uses to get output. It gives type of heart attack as output. If there is any type of attack possibility is predicted then it will show the prediction by percentage value by the utilization of the regression method[1].

In this paper, DSS(Decision Support System) using Naïve Bayes algorithm. In this proposed system, first they collect the data such as age, sex, smoking details, blood sugar, type of chest pain etc. Which are given by users. Then they used Naïve Bayes classifier for supervised learning. It gives an independent variable as input. It reduced the time complexity and give better accuracy as compare to other techniques(Advance Encryption Standard) algorithm is used to secure the patient's data. Its revealed that in regard to accuracy, the prevailing technique surpasses the Naive Bayes by yielding an accuracy of 89.77%in spite of reducing the attributes[2].

In this article, divide process into 3stages.In first stage pre-processing is applied to raw data. Raw data are age, sex, Cp(related to chest pain), cholesterol, etc. Data transformation is done in this stage. Data transformation is done to make this problem a binary class problem[3].Then this pre-processed data used as input of second stage and feature selection is applied to get output of relevant feature. In feature selection irrelevant data are removed. Gain ration is used to get score of given attributes. At last stage classification algorithm is applied. Here they applied Naïve Bayes algorithm and Random Forest Algorithm to predict heart diseases. They used database that is publicly accessible. Confusion matrix ROC curve and area under curve are evaluated.

Here, three phases are applied to get prediction of heart diseases. Pre-processing is applied to first phase. In which data are filtered. In second phase different classification techniques are applied on output of phase one. Classification accuracy, precision, recall and f-measure will be used to evaluate the efficiency of the used techniques[4].Then they choose the highly efficient algorithms from the applied algorithms and then by applying hybridization, result is combine of choose algorithm. Phase three is Diagnose. In this, if the history of patient is available then compare it with result and then it predict heart disease. The main goal in this paper is to investigate available data mining techniques to predict heart disease and compare them, then combine the result from all of them to get most accurate result[4].

In this article, KNN, SVM and ANN used for the prediction of heart diseases. Also compare the result of this three algorithm and also used ensemble classifier. They used multiclass and binary classification. Two types of evolution are used percentage split and cross validation. Binary classification is higher then multiclass classification. The result of percentage split is higher then cross validation. In this method, data get from dataset and then by applying feature selection data is selected that data is used as input of model. The data is split onto two parts: training and testing dataset. At last cross validation is applied.

Here in this paper, different SVM models. The first stage uses a linear and L1 regularized SVM while the second stage uses L2 regularized SVM with different kernels including linear and RBF[17].In first step regularized linear model eliminate the irrelevant data. In second step is for prediction. At last they merged both the hyperparameter to get result in hybrid grid. For optimal search they used hybrid grid search algorithm. This proposed model improves the strength by 3.3%.It also give better result by using less features. It reduce the time so time complexity of proposed model is less.

3. PROPOSED SYSTEM

Following diagrams shows the proposed system of above survey.

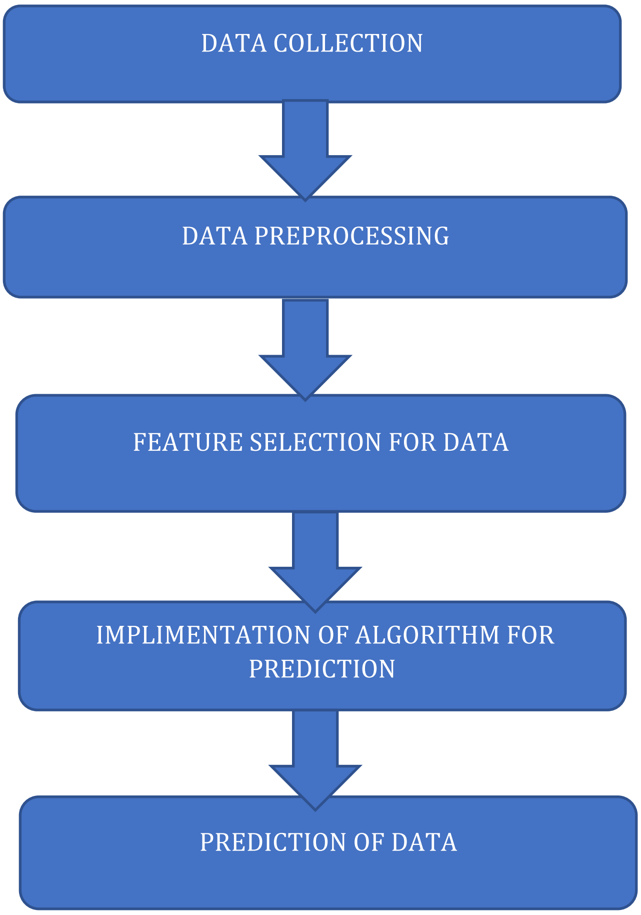
DATA COLLECTION
DATA PREPROCESSING
FEATURE SELECTION FOR DATA
IMPLIMENTATION OF ALGORITHM FOR PREDICTION
PREDICTION OF DATA

Fig-3: Block diagram of proposed method

Above Fig.3shows the block diagram of proposed method which shows the basic flow of our system.

Step 1-Data Collection: There are many sources available to collect the data. Using that data we are collecting the dataset.

Step 2-Data pre-processing: In this step we remove noise or irrelevant data and fill the data for missing values.

Step 3-Feature selection: It is process in which we find relevant data for input. This is used to identify and remove unrelated data that is not useful for the model.

Step 4-Implementation of algorithm: In this step, we take a data set of feature selection and predict heart diseases using ensemble classifier with SVM and naïve bayes algorithm. SVM uses mathematical function, known as kernel function. This function matches the new data from training data for prediction. Then that data is given as input to Naïve Bayes and it assumes the presence of a particular data in class is not related to presence of any other data.

Following fig.4 shows the block diagram of proposed method. First the data converted into small size using naïve bayes. It gives the presence or absence of data. Then using this data SVM is applied. SVM classify the data and give prediction. So here the combination of Naïve Bayes and SVM gives better result for heart diseases prediction.

Step 5-Prediction: In this step our proposed system gives the predicted value. This value is also useful for future prediction.

So this is the scenario of our model.

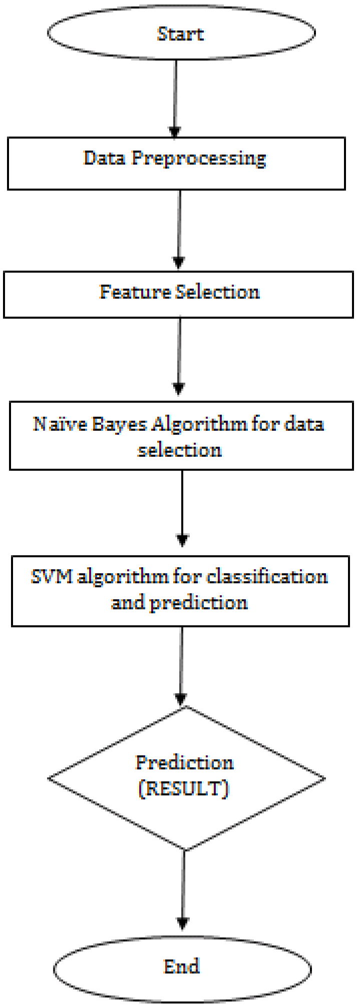
Start
Data Preprocessing
Feature Selection
Naïve Bayes Algorithm for data selection
SVM algorithm for classification and prediction
Prediction (RESULT)
End

Fig.4: Flow chart of proposed method

4. CONCLUSIONS

Many datamining techniques are used for prediction of heart diseases. By Comparing different algorithms like SVM, Random forest, decision tree, are applied on heart diseases dataset are obtained. Main parameters of this comparison give the sensitivity, accuracy, time complexity. By using the proposed system, it will help to get better prediction accuracy for heart diseases prediction.

5. REFERENCES

[1] K. Chandra Shekar, Priti Chandra and K. Venugopala Rao," An Ensemble Classifier Characterized by Genetic Algorithm with Decision Tree for the Prophecy of Heart Disease"Springer(2019).

[2] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin," Design And Implementing Heart Disease Prediction Using Naives Bayesian"IEEE(2019).

[3] Kanika Pahwa, Ravinder Kumar" Prediction of Heart Disease Using Hybrid Technique For Selecting Features"IEEE(2017).

[4] Monther Tarawneh and Ossama Embarak" Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques"Springer(2019).

[5] Ritika Chadha, Shubhankar Mayank" Prediction of heart disease using data mining techniques"Springer(2016).

[6] M.A. Jabbar, B.L. Deekshatulu and Priti Chandra" Prediction of Heart Disease Using Random Forest and Feature Subset Selection"Springer(2016).

[7] G. Thippa Reddy,M. Praveen Kumar Reddy, Kuruva Lakshmanna,Dharmendra Singh Rajput, Rajesh Kaluri,Gautam Srivastava" Hybrid genetic algorithm and a fuzzy logic classifier for heart disease Diagnosis"Springer (2019).

[8] S. Ramasamy and K. Nirmala" Disease prediction in data mining using association rule mining and keyword based clustering algorithms" International Journal of Computers and Applications, (2017).

[9] P. Ramprakash1, R. Sarumathi2, R. Mowriya2, S. Nithyavishnupriya" Heart Disease Prediction Using Deep Neural Network"IEEE(2020)

[10] VirenViraj Shankar,Varun Kumar, Umesh Devagade, Vinay Karanth1 · K. Rohitaksha" Heart Disease Prediction Using CNN Algorithm"Springer (2020)

[11] Harshali Dube1, Shweta Madge2, Prajakta Jagtap3, Pooja Potdar4, and Mr.Nilesh Bhandare" Review on Heart Disease Classification"IEEE(2020)

[12] C. Sowmiya1 · P. Sumitra" A hybrid approach for mortality prediction for heart patients using ACO-HKNN"Springer(2020)

[13] Theresa Princy. R, J. Thomas" Human Heart Disease Prediction System using Data Mining Techniques"

International Conference on Circuit, Power and Computing Technologies ICCPCT](2016).

[14] K. Mathan,Priyan Malarvizhi Kumar, Parthasarathy Panchatcharam,Gunasekaran Manogaran, R.Varadharajan" A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease"Springer(2018)

[15] Meenal Saini, Niyati Baliyan*, Vineeta Bassi" Prediction of Heart Disease Severity with Hybrid Data Mining" 2nd International Conference on Telecommunication and Networks (TEL-NET 2017).

[16] Dželila Mehanović, Zerina Mašetić, and Dino Kečo,"Prediction of Heart Diseases Using Majority Voting Ensemble Method"Springer(2020).

[17] LIAQAT ALI,AWAIS NIAMAT,JAVED ALI KHAN, NOORBAKHSH AMIRI GOLILARZ4, AND XIONG XINGZHONG" An Expert System Based on Optimized Stacked Support Vector Machines for Effective Diagnosis of Heart Disease"IEEE(2016).

[18] Abhishek Rairikar, Vedant Kulkarni, Vikas Sabale, Harshavardhan Kale" HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES" International Conference on Intelligent Computing and Control (I2C2 2017).

[19] Mehrbakhsh Nilashi• Hossein Ahmadi,• Azizah Abdul Manaf•Tarik A. Rashid• Sarminah Samad• Leila Shahmoradi• Nahla Aljojo•Elnaz Akbari," Coronary Heart Disease Diagnosis Through Self-Organizing Map and Fuzzy Support Vector Machine with Incremental Updates International Journal of Fuzzy Systems(2020).

[20] Deepali Chandna" Diagnosis of Heart Disease Using Data Mining Algorithm" (IJCSIT) International Journal of Computer Science and Information Technologies(2014).

[21] H. Benjamin Fredrick David and S. Antony Belcy" HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES" ICTACT JOURNAL ON SOFT COMPUTING(2018).

[22] Ilias Tougui & Abdelilah Jilbab & Jamal E Mhamdi" Heart disease classification using data mining tools and machine learning techniques"Springer(2020).

[23] Montu Saw, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, Nidhi Lal" Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning"IEEE(2020).

[24] KDD process steps accessed on 20 April, "https://data-flair.training/blogs/data-mining-and-knowledge-discovery/".