A REVIEW ON HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

Meenu Shukla
Assistant Professor
Computer Science Department
Krishna Engineering College
Ghaziabad, India

Sandeep Kumar
UG Student
Krishna Engineering College
Ghaziabad, India

Rishabh Sharma
UG Student
Krishna Engineering College
Ghaziabad, India

Saurabh Sharma
UG Student
Krishna Engineering College
Ghaziabad, India

Rishabh Tyagi
UG Student
Krishna Engineering College
Ghaziabad, India.

ABSTRACT: Heart is the second most significant feature relative to the human body's nucleus. It filters the blood, providing all of the body's organs. Data processing allows monitoring further knowledge and lets the patient identify various diseases. Significant numbers of patient-related data are held regularly. The accumulated data can be useful as a tool for forecasting possible disease happens. Such as Decision Tree, Naive Bayes, Support Vector Machine as well as K-Nearest Neighbor some of the Data mining and Machine learning methods are being used to forecast cardiovascular disease. The paper presents a nearby of the successful algorithm and discusses the current work in general. There is also an approx. estimation of data which is compounded using all the datasets.

***KEYWORDS*:** *Data Mining, forecasting heart disease, pre-processing model, prediction classifiers.*

1. INTRODUCTION

Heart condition is among the big diseases which nowadays may cut the standard of human existence. 17.5 million People suffer per year from heart disease. Life demands that the heart act as a variable, since the heart is a crucial part of our bodies. Heart disorder is a syndrome that influences cardiac activity. Estimating the likelihood of cardiovascular disease in one individual is important for a lot of areas of health endorsement and medical drug. Restoration of a danger evaluation possibly has done using a multivariate regression examination with a statistical sample [12]. Due to the rapid growth of new technologies, health care institutions keep vast amounts of data of their archives, so it becomes really complex well as challenging to analyse.

ML techniques play an essential role in medical centre assessment of the various results. These methods as well as strategies can be used uncompromisingly on even a database to construct these models or else to make essential conclusions as well as inferences as of the set of data. Sex, fasting blood pressure, age, category of heart problems, latent ECG, amount of large fluoroscopic coloured veins, test blood pressure, serum cholesterol, thalassic, ST depression [1].

Medical research cardiac damage from numerous serious diseases has gained a great deal of notice. Heart disease detection is a challenging task which must require statistical examination of the cardiac attack of the victim so that further identification will be rendered quickly. The diagnosed with heart failure usually is focused mostly on outcome diagnosis of the patient signs and symptoms. There are a number of factors that improve the hazard of cardiac damage, for example use of alcohol, systemic cholesterol rates, family record of obesity, heart disease, lack of work out, high blood pressure as well as high blood pressure. There are several medical records systems built to assist patient accounting, inventory control and the simple generation of statistics [2][13]. A few other healthcare facilities use the decision systems, however they are usually limited. They the address basic questions like "and what's the average age of diagnosis with cardiac disease?" How many operations have resulted in medical stays longer than 10 days? "Identify single female women, aged 30 years old, who've been diagnosed for cancer." However, these cannot address specific questions such as 'provided medical history, estimate the risk of patients dying from cardiac failure.' Clinical evaluations are frequently prepared on the origin of physicians' intuition as well as knowledge, than the hidden facts- rich information in their record [11].

This practice results in unwanted beliefs, inaccuracies and increased hospital expenses affecting the quality of patient service. The suggested approach is that combining clinical decision making through computer-based health records will minimize physician mistakes, increase patient protection, reduce unnecessary treatment variability, and enhance patient results. This advice is hopeful as data analysis as well as simulation technologies, e.g. records processing, are capable of creating a fact-rich surroundings that could theoretically considerably improve productivity of treatment practice. Professional treatment requires effective medical identification and the delivery of suitable therapies. Terrible healthcare professionals will cause catastrophic consequences [10], even undesired.

Healthcare facilities can increase the cost of clinical tests. These results can be achieved by the use of reliable, computer-base in sequence as well as judgment systems. The healthcare manufacturing is gathering huge amount of health records that are unfortunately not "harvested" for positive decision-making to discover proprietary information. Sometimes the unseen patterns and experiences go unexplored. Advanced data management techniques should help in remedying this situation. It is a potential medical diagnosis System (HDPS) [3][14] that essentially decides whether or not patient is prone to heart failure, the recommended approach will include questions gathered for the person to answer. The solution will be generated on the basis of the patient's responses.

A major problem for providers in healthcare services, such as hospitals and community centres, is providing quality services at affordable levels. Professional care calls for correct condition evaluation and effective diagnosis. The open heart disease database contains quantitative as well as categorical tests. On these records washing or retrieval [15] is performed to delete the out-dated information from the files before any further storage. The proposed approach would analyse reliable hidden knowledge from a previous cardiac attack context, i.e. patterns and relationships associated with heart disease. It can also answer the difficult heart disease patient queries; hence it can be helpful for health care professionals to take responsible medical choices. Results revealed that the future approach has its unique efficacy within attaining organizational objective of mining targets set.

2. LITERATURE REVIEW

Frequent research related to disease forecast model has been performed using numerous mining methods and learning algorithms in health centres.

Polaraju et al, suggested cardiovascular disease Forecasting using several linear regression, and K shows that multiple regression analysis is appropriate for forecasting risk of cardiac attack [4]. The analysis is done by means of planning data collection consisting of 3000 instances with 13 specific attributes previously specified. The collection of data is separated into 2 parts, which account for 70% of the data used for teaching also 30% for research. Depending resting on the outcomes, it's obvious to Regression method's arrangement accuracy is better associated with other methodologies.

Centred on results of different SMO and Bayes Net variables [8], output is optimal compared to Kstar, multilevel observation and j48 approaches utilizing k-fold cross validation.Marjia et al. recognized prediction of heart disease using weka tools, using Kstar, j48, smo, and Bayes net and Multilevel the precision performance achieves with these methodologies is at rest unsatisfactory. Consequently, efficiency of the precision is enhanced further to offer better diagnostic disease judgment.

S. Seema et al, focus to strategies which can forecast chronic illness with the help of Support Vector Machine, Artificial Neural Network, Decision Tree and naive Bayes, extracting the data found in historic health documents. To calculate the improved efficiency at an acceptable scale, a trained analysis is performed on classificatory. SVM yielded the best precision score from such a trial, while Naive Bayes gives the lowest precision for diabetes.

P. Chandra, M. Jabbar recycled the array of subsets of features to render class association [6] guidelines for cardiovascular disease identification. Instructions for the association control the partnership among attributes and organisation to forecast components in a series of patient information. Functional choice, like genetic search, governs the beneficial features of predicting cardiac disease.

Usha Rani proposed a system for forecasting cardiac disease with aid of ANN [5], incorporating feed forward and back propagation techniques. Studies were done using representations of single and multiphase neural networks. Parallel processing is performed to speed up the training process of all hidden embedding layer within each nerve cell.

Along with aid of recommendation laws Carlos Ordonez designed the diagnosis of heart failure. They had used a basic algorithm for visualization. This satisfaction algorithm assigns to be empirical or unambiguous. It is used for translation of medical documents into transaction formats. Using enhanced algorithms [7], established classification methods are minified. The filtering system is planned, and the quality of the component is assigned to the object. Decision trees are included in data processing because they separate numerical values robotically. Break points are seldom utilized as specified in the decision tree. Clustering is used to get a general overview of the findings.

Flowchart:

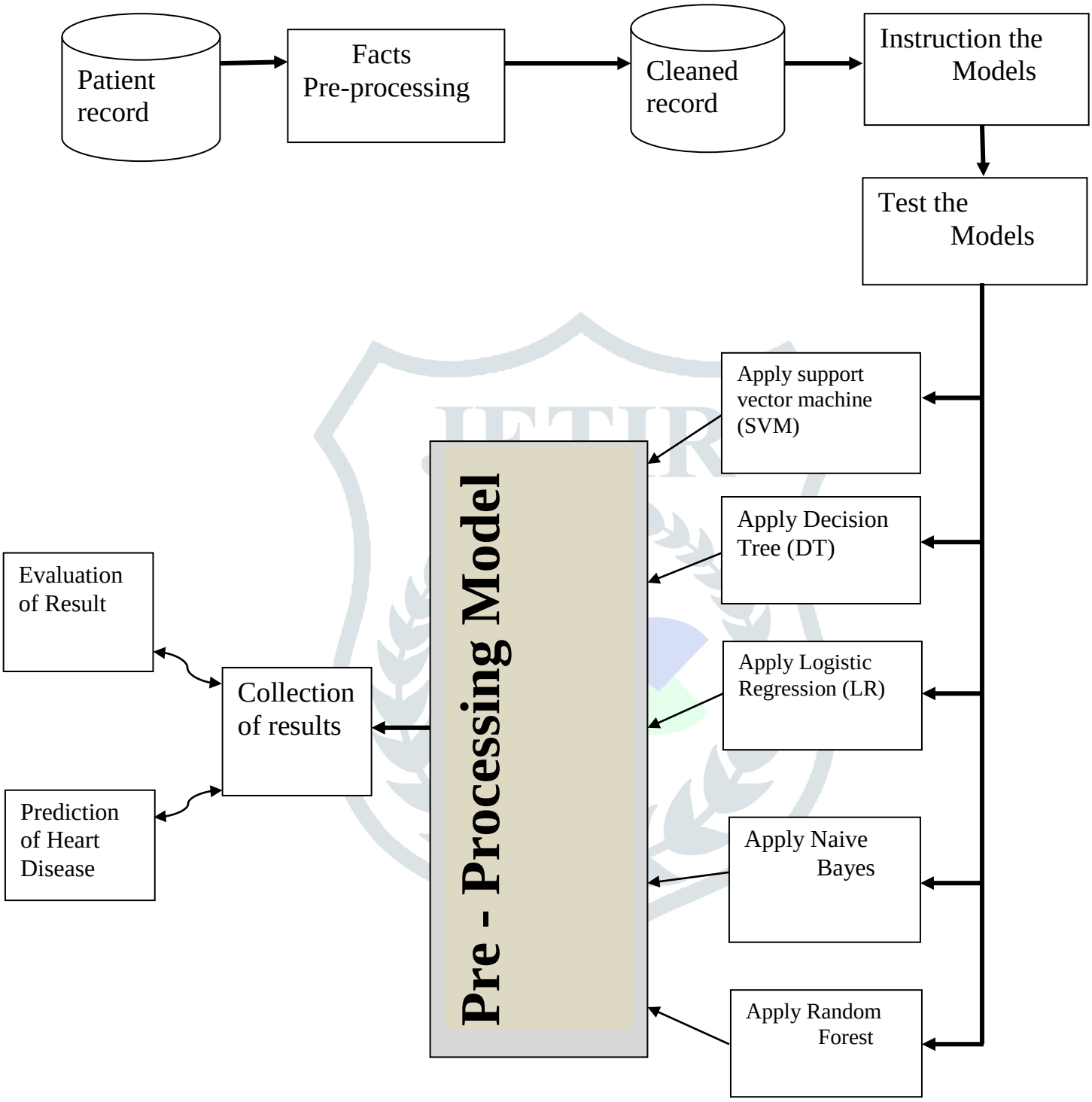

Fig.1: Flowchart for pre-processing model.

3. TECHNIQUES

In this project we have worked on SYPDER as software in python. and hardware devices like 8GB RAM and i5 processor are also being used for completion of this project.

Data Mining: Data mining means working deep interested in data which is in dissimilar forms towards grow up pattern, and to increase information resting on that model. In the method of data mining, huge records sets are first sorted, after that pattern are known and relationships are well-known to make solve problems and data analysis.

3.1 Decision Tree

- Decision tree set of the rules cataracts under the category of administer learning. And it is also used to crack both waning and understanding problems.
- Decision tree use the tree demonstration to crack the difficulty in which every leaf node correspond towards the class tag and aspects be represented on the inner node of the tree.

ACCURACY OF DECISION TREE ALGORITHM: 75.409

3.2 Logistic Regression:

Logistic Regression be baptized used for the role and it is used as the core of the way, the logistic gathering. The Logistic gathering is as well-known as the sigmoid role which was developed with maxing out at the carrying volume of the surroundings and geometers towards explain chattels of people growth in natural balance, rising quickly. It's an S-shaped camber that can take any real-valued number and map it into a value between 0 and 1, but under no circumstances exactly at those limits.

Logistic Regression = 1 / (1 + e^-value)

Here is the base of the ordinary logarithms (Euler's number or the EXP () utility in your record) and assessment is the actual arithmetical value that you desire to convert . Lower is a plot of the statistics between-5 and 5 converted into the assortment 0 and 1 use the logistic function.

ACCURACY OF LOGISTIC REGRESSION ALGORITHM: 80.327

3.3 Random Forest:

Random Forest is an algorithm for managing research that is used both for classification and regression. Yet it is mostly used for scandalous grouping. As we recognize a forest consists of trees and extra trees mean extra dynamic forest. Equally, random forest algorithms construct decision trees on records samples and after that get prophecy from every of them and eventually pick the top answer through arbitrary means. It is a method of enterprise that is superior to a single result tree since it reduces the over fitting with being around the effect.

ACCURACY OF RANDOM FOREST: 89.473

3.4 SVM (Support Vector Machine):

Support Vector Machines be analytical model in process wisdom, with attendant education systems that grab evidence that is used for sorting and regression examination. Identified a series of research instances, every labeled since heading to 1 or the other of 2 groups, an SVM retaining an acceptable algorithm forms a test order to allocate new samples to one class otherwise the other, forming a non-probabilistic dualistic linear classifier (although means such as Platt scaling happen to use SVM in a probabilistic ordering system).The optimal SVM is to explain the specimens as spatial data, intended to divide the specimens in the different groupings by a wide distance that is as varied as probable. So fresh conditions are built for the particular place and are forced to contribute to an organization centered on the side of the slit they fall upon.

An appropriate SVM is to show the examples as space points, designed to distinguish the independent grouping examples by a simple slit as large as possible. Then, new conditions are mapped in the same space and expected to belong to a classification centred on the side of the slit they collapse through.

ACCURACY OF SUPPORT VECTOR MACHINE: 88.524

3.5 Naive Bayes:

Software scholarship naïve Bayes classifiers are viewed as basic "probabilistic classifier" based on related Bayes inference by clear (naive) assumptions of outrageousness between the functions. They are among the unissued goals of the Bayesian network.

Since the 1960s, Naïve Bayes former regarded expansively. It was introduce in the information extraction group in the early 1960s and remains a ordinary (baseline) strategy for assessing papers as belonging to one or the former group (for example sports, politics or spam, etc.) with word frequencies since identifiers. It is economical in this area for more advanced techniques like SVM, for appropriate pre-processing. In predictive medical diagnosis this sometimes finds appeal.

ACCURACY OF NAÏVE BAYES: 81.196

4. ANALYSIS RESULTS

The accuracy of different classification algorithm are given below:

Classification error -

It applies to the data sets misidentified from the corresponding listed documents.

Real positive percentage: This refers to the amount of positive specimens which the identification system accurately expected.

False Positive percentage: It refers to the amount of false specimens so as to the recognition system incorrectly expected.

The responsiveness, specificity, precision and accuracy is measured for evaluation of the enactment of each combination:-

1. Specificity- It measures the percentage of denials which are correctly notorious.

$$\text{Specificity} = \frac{TN}{FP + TN}$$

2. Recall- Is the fraction of relevant illustrations that are retrieved.

$$\text{Recall} = \frac{TP}{TP + FN}.$$

3. Precision-It is the segment of retrieved instances that are relevant.

$$\text{Precision} = \frac{TP}{TP + FP}.$$

4. Accuracy- It refers to the total number of records that are correctly confidential by the classifier

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

Table.1: collective classification of all the techniques with their accuracy.

CLASSIFICATION OF TECHNIQUE	**ACCURACY**	
	TRAINING SET	TESTING SET
SUPPORT VECTOR MACHINE	91.322%	88.524%
LOGISTIC REGRESSION	82.644%	80.327%
RANDOM FOREST	89.473%	88.157%
NAÏVE BAYES	82.231%	81.967%
DECISION TREE	100%	75.409%

Table 1.1, describes how we can easily evaluate the classified model of different types of clustering algorithm which is used here. Some of the training dataset is provided to the machine and as a result the machine gives a certain output as in the form of testing dataset with a required amount of accuracy results.

5. FUTURE SCOPE

- Desktop Application: On Android and IOS devices, we will build Smartphone applications to make the program simple to use. We will also have an SMS service that will give the patient a letter about his heart disease as well as whether to receive appropriate method to minimize threat.
- Pacemaker**:** We could attach defibrillator to the device. A defibrillator is a compact tool for monitoring irregular heartbeats that is mounted in the chest or abdomen. This system utilizes electric pulses to quickly beat the heart at a regular rhythm. We can often use this form of prediction techniques to render many applications. Because only a second improvement is obtained in the development of an empirical model for patients with cardiac disease, combinational and more composite preparation are required to improve the precision of predicting the premature arrival of cardiovascular syndrome. The further records is fielded hooked on the servers machine would be really light. None of the less, suitable to time limitations the following examine has to be carried out for the upcoming.

6. CONCLUSION

By using dissimilar types of machine learning method to expect the rate of heart disease include concise. By decisive the prediction acts of every algorithm and apply the proposed organization for the area it necessary. Use extra significant element collection methods to get better the perfect act of algorithms. There are some action methods for long-suffering, if they once diagnosed with the particular form of heart disease.

There are frequent feasible improvements that could be explored to improve the scalability and accuracy of this forecast organization. Would like to use various discretization techniques, various classifiers and decision tree testing to improve the system's greater validity and stability.

7. REFERENCES

[1] AnimeshHazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Research Gate Publications, July 2017, pp.2137-2159.

[2] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review", International Journal of Computer Applications, February 2016.

[3] Guizhou Hu, Martin M. Root, "Building Prediction Models for Coronary Heart Disease by Synthesizing Multiple Longitudinal Research Findings", European Science of Cardiology, 10 May 2005.

[4] T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, "A Heart Disease Prediction Model using SVM- Decision Trees- Logistic Regression (SDL)", International Journal of Computer Applications, vol. 68, 16 April 2013.

[5] Shadab Adam Pattekari, Asma Parveen"Prediction System for Heart using Naive Bayes", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.

[6] http://www-igm.univ-mlv.fr/~lecroq/string/node14.html, Boyer Moore, accessed on 25th, April, 2014.

[7] Resul Das, Ibrahim Turkoglub and Abdul kadir Sengurb, "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, 2009, pp. 7675–7680.

[8] Robert Detrano, "Cleveland Heart Disease Database", V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

[9] MoloudAbdar, Sharareh R. NiakanKalhori,et.al, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases" ,InternationalJournal of Electrical and Computer Engineering (IJECE) ,Vol. 5, No. 6, December 2015, pp. 1569-1576.4

[10] Sumit B, Praveen P, G.N. Pillai. "*SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features".* WCECS Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, October 2008.

[11] Kyle E. Walker, Sean M. Crotty. "Classifying high-prevalence neighbourhoods for cardiovascular disease in Texas", *Applied Geography*, 2014.

[12] K.Rajeswari, V.Vaithiyanathan, T.R. Neelakantan. "Ischemic Heart Disease Identification using Feed Forward Neural Networks", International Symposium onRobotics and Intelligent Sensors,(IRIS 2012), Procedia Engineering, pp 1818-1823.

[13] A.V Senthil Kumar "Generating Rules for Advanced Fuzzy Resolution Mechanism to Diagnosis Heart Disease", *International Journal of Computer Applications*, 2013.

[14] A.Rafiah and P.Sellappan "Intelligent Heart Disease Prediction System Using Data Mining Techniques", 2008.

[15] I.H. Karlberg, and S.L. Elo "Validity and utilization of epidemiological data: A study of Ischaemic heart disease and coronary risk factors.