International Journal of Advanced Science and Technology
Vol. 29, No.8, (2021), pp. 6274-6282

Heart Disease Data Pre-Processing Using Enhanced Data Mining Techniques

[1] Ms. C. Keerthana, [2] Dr. B. Azhagusundari

Assistant professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, TamilNadu, India.

Abstract

Data mining is an iterative progress in which evolution is defined by detection, through normal or manual methods. Knowledge discovery and data mining have discovered different applications in scientific space Heart disease is a term for defining a huge measure of healthcare conditions that are identified with the heart. Diverse data mining techniques, for instance, affiliation rule mining, classification, clustering is utilized to predict the heart disease in medical services industry. The coronary illness database is preprocessed to make the mining cycle more successful. The preprocessed data is clustered utilizing bundling algorithms like K-intends to aggregate relevant data in database. Maximal Frequent Item Set Algorithm (MAFIA) is used for mining maximal frequent models in coronary illness database. Coronary illness is the principle wellspring of death in any place on the world over late years. Researchers have built up numerous cross variety data mining techniques for diagnosing coronary illness. This paper depicts a preprocessing strategy and dissects the accuracy for forecast subsequent to preprocessing the noisy data. It is also seen that the accuracy has been expanded to 91% subsequent to preprocessing and sum up the ebb and flow proof on the utilization of preprocessing techniques in coronary illness.

Keywords: Heart Disease Prediction, Data mining, Data Preprocessing, Data Cleaning, Data Integration.

I. INTRODUCTION

The heart is one of the fundamental pieces of the human body after the cerebrum. The essential capacity of the heart is to pumping blood to the whole body parts. Any disorder that can lead to upsetting the usefulness of the heart is called heart disease. Several types of heart disease are there on the planet; Coronary artery disease (CAD), and heart failure (HF) are the most common heart diseases that are present. The primary reason behind the coronary heart disease (CAD) is blockage or narrowing down of the coronary arteries Coronary arteries likewise responsible for providing blood for the heart. Computer aided design is the leading cause of death over 26 million people are suffering from coronary heart disease (CAD) around the globe, Heart disease is a critical issue, so there is a need for diagnosis or prediction of heart disease there are several methods to diagnose heart disease among them Angiography is the trending method which is used by the majority of the physicians across the world.

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST

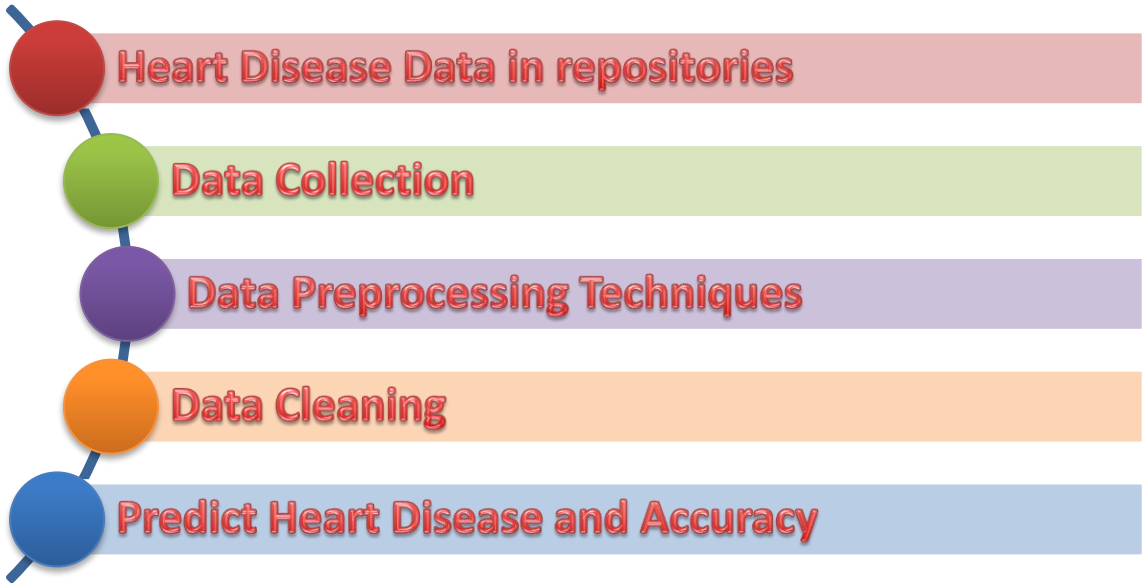


Figure 1: Heart Disease Analysis

Heart disease can be predicted based on different symptoms, for example, age, gender, pulse rate etc. Data examination in healthcare helps with foreseeing diseases, improving diagnosis, investigating symptoms, giving proper medicines, improving the idea of care, restricting expense, expanding the existence territory and reduces the death pace of heart patients. ECG (Electro Cardio Gram) helps in screening sporadic heart beat and stroke with the installed sensors by resting it on a chest to follow the patient's heart beat.

Heart disease forecast is being finished with the nitty gritty clinical data that could assist experts with making decision. Human life is profoundly dependent on proper working of blood vessels in the heart. The improper blood course causes heart inactiveness, kidney failure, imbalanced state of cerebrum, and even immediate death moreover. Some of the risk factors that can cause heart diseases are obesity, smoking, diabetes, blood pressure, cholesterol, lack of proactive tasks and unhealthy diet. Notwithstanding, there are some drawbacks associated with angiography technique. It is an expensive procedure and physicians have to analyze endless components to diagnose a patient hence this process makes doctor work very troublesome, so these impediments motivate to develop a non-invasive method for prediction of heart disease.

Data Mining involves few steps from crude data collection to some type of new knowledge. The heart disease pre-processing comprises of the going with steps like Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Representation. Medical Data Mining is a space of challenge which involves an enormous heap of imprecision and vulnerability. Arrangement of value services at affordable expense is the significant challenge faced in the health care organization. The poor clinical decision may lead to awful consequences. Health care data is massive.

Clinical decisions are frequently made based on the specialist's experience as opposed to on the knowledge rich data covered up in the database. This in some cases results in errors, excessive medical cost which affects the nature of service to the patients. Medical history data comprise of various tests essentials to diagnose a specific disease. It is possible to pick up the advantage of Data mining in health care by employing it as an intelligent demonstrative apparatus. Accuracy of the prediction can likewise be improved by utilizing Data Clustering Algorithms in heart disease investigation. Classification as respects:

a. The DP tasks and techniques most frequently utilized,
b. The impact of DP tasks and techniques on the presentation of classification in cardiology,
c. The general exhibition of classifiers when utilizing DP techniques, and

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST

d. Examinations of different blends classifier-preprocessing in terms of accuracy rate.

2. Data Mining in Heart Disease

Data mining is characterized as a strategy for extricating necessary and concealed predictive data from gigantic databases. Data mining classifiers find wide applications in medical area for diverse diagnosis. Different data mining techniques are used for prediction and decision making for different kinds of diseases like cancer, heart disease, diabetes etc. There are numerous types of data mining classifiers which can be applied to predict diseases. In this paper we are zeroing in on heart disease prediction. Data mining instruments predict future trends by permitting knowledge-driven decisions. The prediction of heart disease requires an immense size of data which is excessively perplexing and massive to process and examine by traditional techniques. Diverse data mining techniques are being used by experts. Our goal is to discover the suitable data mining technique that is computationally effective just as accurate for the prediction of heart disease.

Hence, data preprocessing techniques are being created and used in the undertaking to improve the exhibition of DM techniques in cardiology. Data preprocessing (DP) has been asserted by numerous researchers as a significant and basic step in the knowledge data discovery (KDD) process. KDD comprises of three rule steps: data preprocessing, data displaying and data post-processing. Several KDD studies and found that about 70% of exertion and time was spent on data preprocessing, while the DM step required interestingly about 25% of the total exertion.

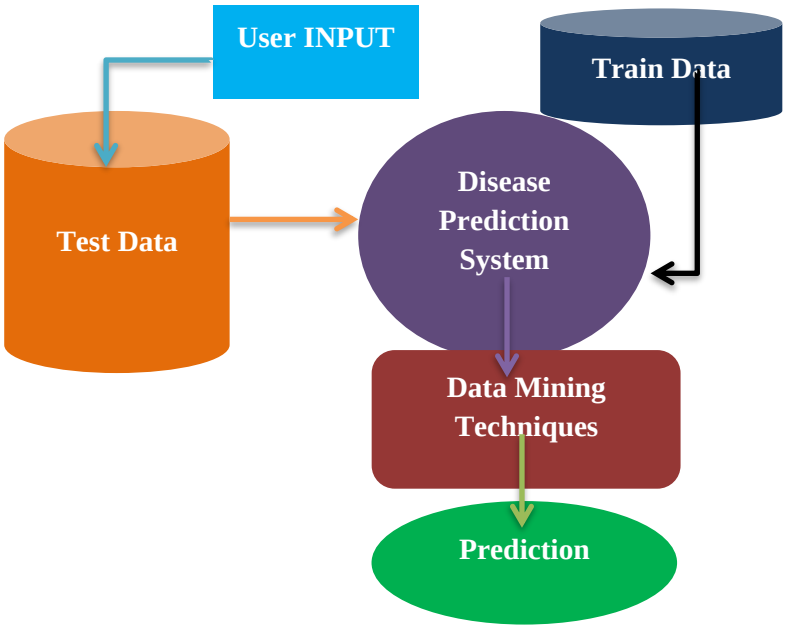


Figure 2: Data mining Techniques in Heart Disease Prediction

The purpose of data preprocessing is to manage crude data imperfections and challenges, for instance, instability, noise, not presented data, outliers, high largeness and imbalanced data.

Data Cleaning

Data cleaning is the process of planning data for investigation by eliminating or changing data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is generally not necessary or helpful when it comes to dissecting data because it might hinder the process or provide inaccurate results. There are several methods for cleaning data depending on how it is stored alongside the answers being looked

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST

for. Data cleaning isn't just about erasing data to make space for new data, but instead figuring out how to maximize a data set's accuracy without necessarily deleting data. For one, data cleaning includes more activities than removing data, for example, fixing spelling and language structure errors, normalizing data sets, and correcting mistakes, for instance, empty fields, missing codes, and distinguishing copy data points.

In this context, the need for data cleaning to improve data quality is becoming essential. Duplicate records elimination is a challenging data cleansing task. In this paper, we present a duplicate records elimination way to deal with improve the nature of data. Data quality has become a significant issue, due to its tangible effect on the nature of the provided decisions. This issue becomes more and more significant in the medicine area, where the need for effective decision making is high, and where helpless data quality really affects people's lives. Ensuring data quality is principally about applying the data cleaning process. The purpose of data cleaning (likewise called data cleansing) is to remove the anomalies that affect the data.

3. Heart Disease Preprocessing

This section presents a system overview and the description of the existing arrangement Heart Care. The arrangement is divided into three phases.

- Data Pre-processing.
- Model Training Phase
- Live Prediction.

The brief description of the relative multitude of phases is in the accompanying subsections. In this section, we describe our existing system. Figure 3 illustrates the general architecture of the system. We follow two steps before making quality decisions. The initial step comprises of cleaning the messy data from duplicate records. This step comprises of duplicate records detection and duplicate records correction. The second step permits predicting the presence or absence of heart disease based on cleaned data.

De Duplication of records

The duplicate records elimination process is based on two processes; duplicate records detection and duplicate records correction. Duplicate records detection process permits identifying coordinating records and duplicate records correction process permits generating new records by merging the coordinating ones.

Duplicate records detection

The duplicate records detection approach; Figure 3 illustrates the general architecture of our methodology and follows three steps to detect duplicate records. The initial step comprises of indexing the record sets to be compared. The second step generates embedding's vectors of records and computes likeness vectors between record sets. The last step predicts if a record pair matches or not matches.

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST

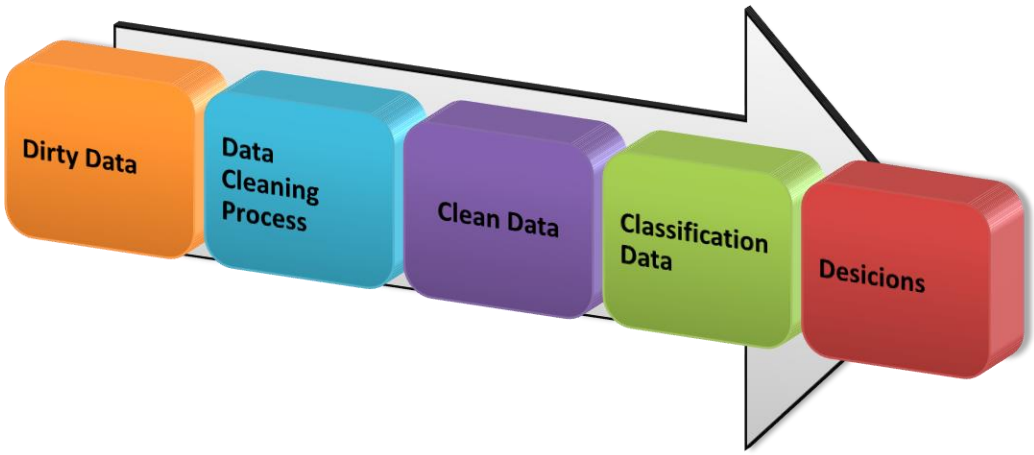


Figure 3: General Architecture for Existing Method

III. Data Pre-processing

This phase comprises of dataset retrieval, data cleaning and feature selection so that the obtained dataset can be used for extra steps to improve results. The steps of this phase are listed below:

- **Collecting Dataset:** The dataset is retrieved from a reliable source.
- **Addressing the missing values:** In this step the missing data is addressed utilizing the measurable method 'mode' that is replacing the missing values with the most frequent estimation of each element.
- **Feature Selection:** In this certain useful features are selected from the wide and used the correlation matrix to check if the attributes are correlated or not.

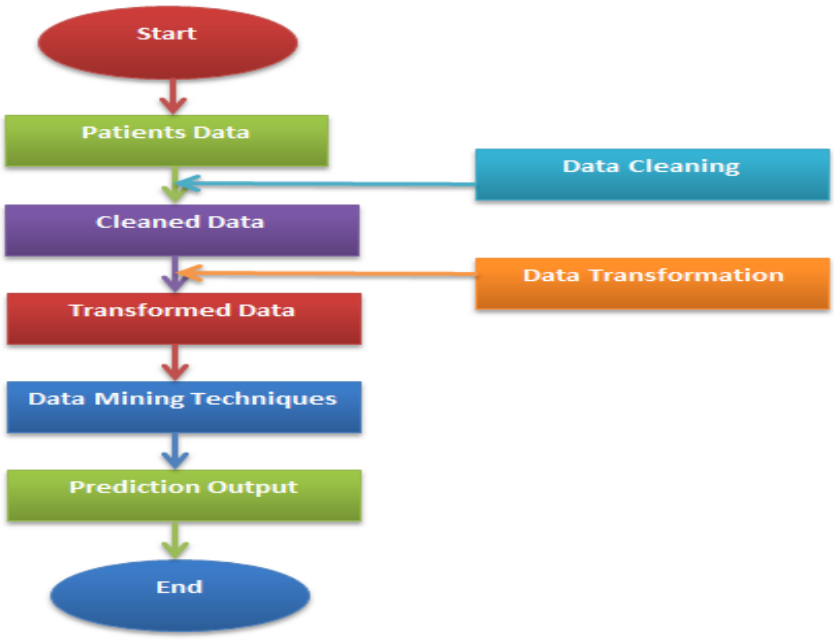


Figure 4: Flow Diagram of Preprocessing in Heart Disease Analysis

Model Training

The second existing system is included splitting the preprocessed dataset into training and testing dataset and afterward afterwards training the model utilizing the training dataset. Then the predictions are made utilizing the testing dataset on different trained models (trained utilizing different algorithms). From

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST

these tests, the best model is selected based on the performance metrics like accuracy, sensitivity and miss rate. The model which performs better in the entirety of the cases is selected as the last training model for further live predictions.

4. Algorithmic Procedure for Preprocessing

In the proposed system early diagnosis of the heart disease is carried utilizing the data mining techniques. A huge measure of healthcare data, which tragically, are not mined to discover concealed data for compelling decision making. This research provides a prototype "Heart Disease Diagnosis Using Data Mining Technique. It is a searching that imitates the process of typical development. This inquisitive is routinely used to make useful answers for optimization and search problems. In our system the algorithm is used to extract attribute from a huge attribute set. The extracted attribute are as per the following,

III. Experiment Results on Heart Disease Preprocessing

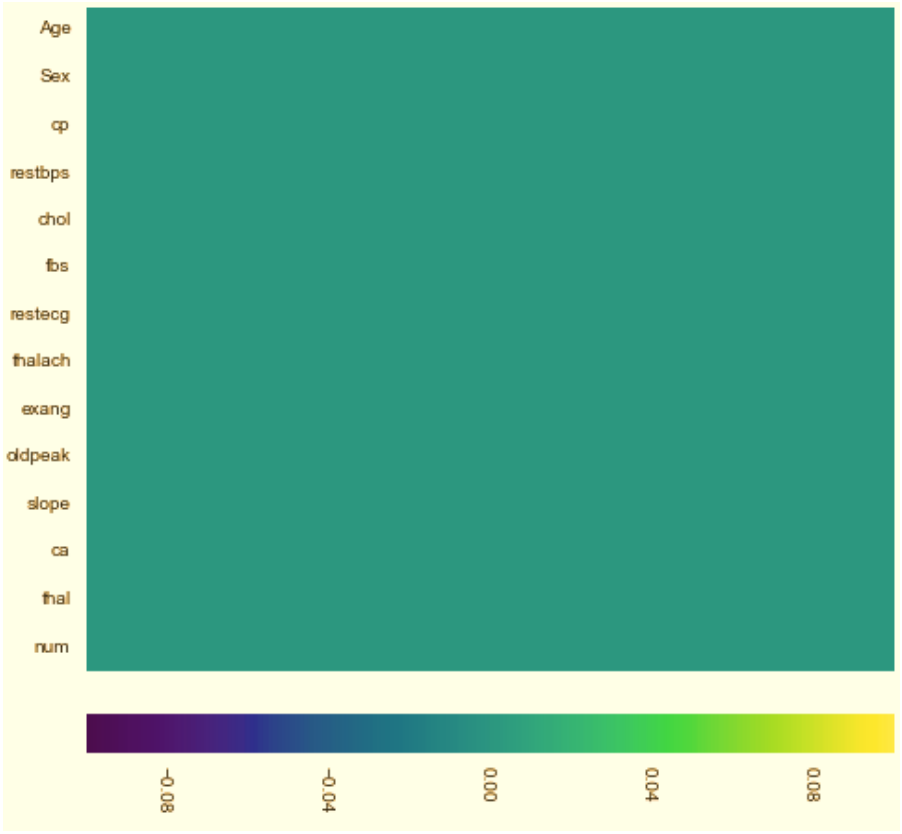


Figure 5: Chart for Heart Disease Cases by Age

The above Figure 5 explains the Number of Heart Patient Cases by Age in Heart Disease Analysis.

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST

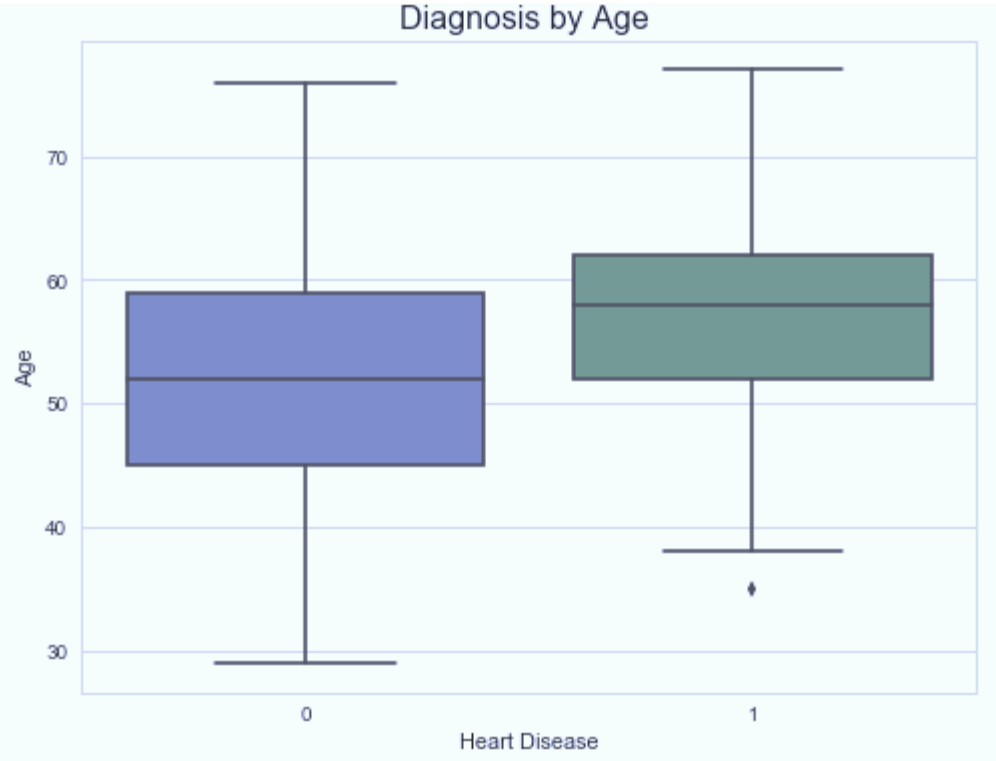


Figure 6: Heart Disease by Age

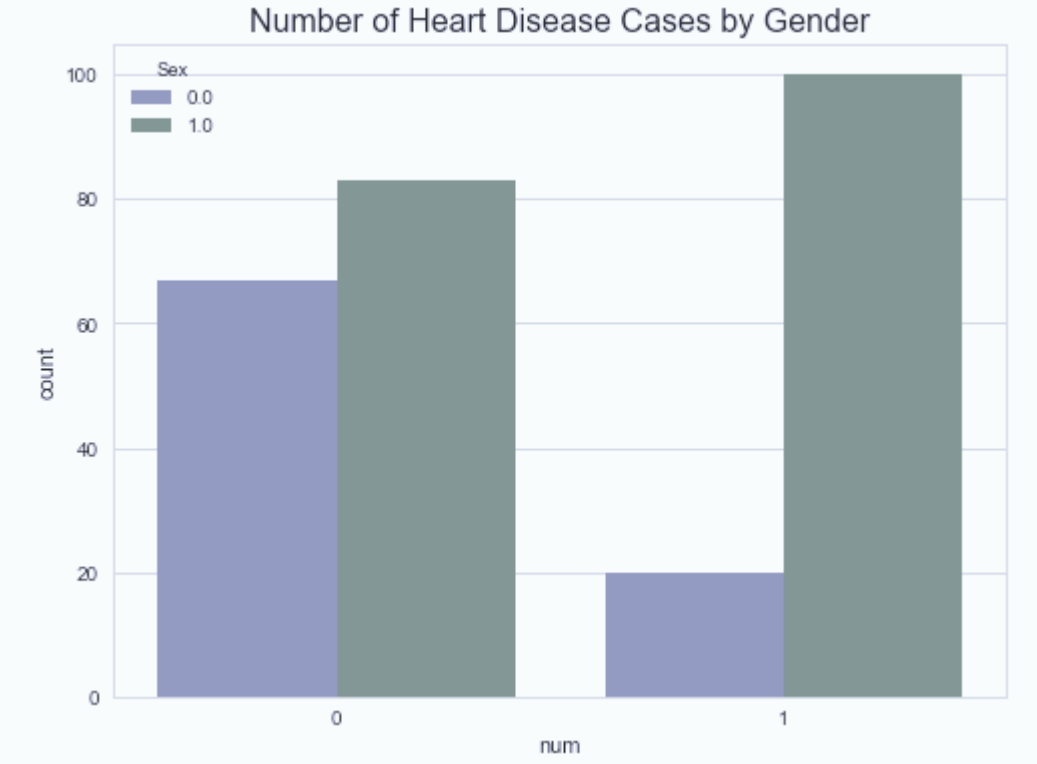


Figure 7: Number of Heart Disease Cases by Gender

The above Figure 7 explains the Number of Heart Patient Cases by Gender in Heart Disease Diagnosis.

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST

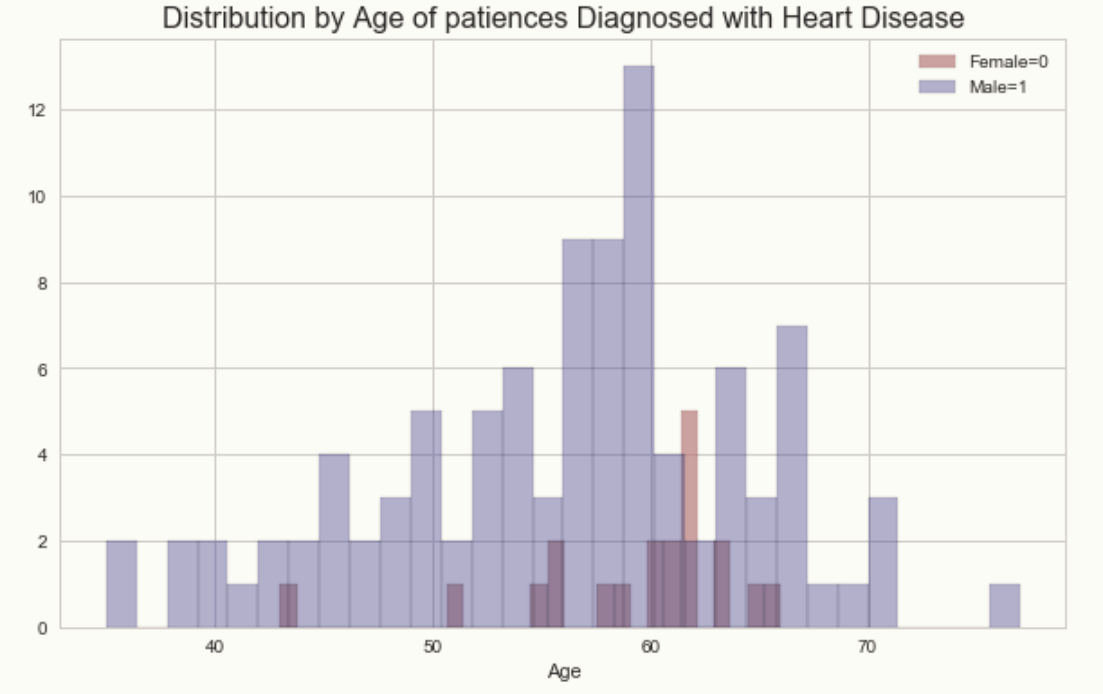


Figure 8: Heart Disease Diagnosis by Patient's Age

In this result the attention is on utilizing different algorithms in data preprocessing and sequence of several attributes for compelling heart disease prediction and its diagnosis. It has tremendous efficiency utilizing fourteen attributes, after applying proposed algorithm to reduce the real data size to get the ideal subset of attribute acceptable for heart disease prediction. In this paper the attention is on utilizing different algorithms in data mining and sequence of several attributes for effective heart disease prediction and its diagnosis. It has tremendous efficiency utilizing fourteen attributes, after applying this algorithm to reduce the genuine data size to get the ideal subset of attribute acceptable for the heart disease prediction.

IV. Conclusion

The data mining techniques are to predict heart diseases which are discussed here. Heart disease is a human disease by its disposition. This disease makes several problems, for example, heart attack and death. In the medical area, the significance of data mining is perceived. Different steps are taken to apply pertinent techniques in the disease prediction. In this way diagnosing patients effectively based on time is an urgent limit with regards to medical support. An invalid diagnosis done by the emergency clinic leads for losing reputation. The precise diagnosis of heart disease is the predominant biomedical issue. The inspiration of this paper is to develop an efficacious treatment utilizing data mining techniques that can help remedial circumstances. The research works with effective techniques that are done by different researchers were studied in this paper. From the comparative examination we can conclude that data mining technique is an efficient method for predicting heart disease. It gives great accuracy by observing different research papers.

References

1. Sarath Babu,Vivek EM,Famina KP,Fida K,Aswathi P,Shanid M,Hena M,"Heart disease Digagnosis using data mining Technique",IEEE,2018.
2. Cincy Raju,Philipsy E,siji Chacko,L.Padma Suresh,Deepa Rajan S,"A Survey on predicting heart disease using data mining Techniques",IEEE,2018.
3. Salha M. Alzahani, Afnan Althopity, Ashwag Alghamdi, -"An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction" Lecture Notes on Information Theory, Volume: 02, No. 04 [2]

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST

4. H K Shifali, Dr. B. Srinivasu, Rajashekar Shastry, B N Ranga Swamy-"Mining of Medical Data to Identify Risk Factors of Heart Disease Using Frequent Itemset", International Research Journal of Engineering and Technology (IRJET),Volume: 04 Issue: 01 [3]
5. Prof. Mamta Sharma 1, Farheen Khan2, Vishnupriya Ravichandran- "Comparing Data Mining Techniques Used For Heart Disease", International Research Journal of Engineering and Technology, Volume: 04 Issue: 06
6. T. A. Gaziano, "Reducing the growing burden of cardiovascular disease in the developing world," Health Affairs. 2007, doi: 10.1377/hlthaff.26.1.13.
7. N. Yilmaz, O. Inan, and M. S. Uzer, "A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases," J. Med. Syst., vol. 38, no. 5, 2014, doi: 10.1007/s10916-014-0048-7.
8. H. Benhar, A. Idri, and J. L. Fernández-Alemán, "A Systematic Mapping Study of Data Preparation in Heart Disease Knowledge Discovery," J. Med. Syst., vol. 43, no. 1, p. 17, Jan. 2019, doi: 10.1007/s10916-018-1134-z
9. Annoj P.K.," Clinical decision support system: Risk level prediction of heart disease using Data Mining Algorithms", Journal of King Saud University- Computer and Information Sciences, pp 27-40,2012.
10. AshaRajkumar and Mrs. Sophia Reena, "Diagnosis of Heart Disease using Data Mining Algorithms", Global Journal of Computer Science and Technology, vol. 10(10), pp 38-43, 2010. 4. BalaSundar V, "Development of Data Clustering Algorithm for predicting Heart", IJCA, Vol 48(7), pp 8-13, June 2012.
11. BhagyashreeAmbulkar and VaishaliBorkar "Data Mining in Cloud Computing",MPGINMC, Recent Trends in Computing, ISSN 0975-8887, pp 23-26, June 2012.
12. Bhuvaneswari. R, "Naïve Bayesian Classification Approach in Health Care Application", International Journal of Computer Science and Telecommunication, Volume 3(1), pp 106-112, Jan 2012.
13. Carlos Ordonez, Edward Omincenski and Levien de Braal "Mining Constraint Association Rules to Predict Heart Disease", Proceeding of 2001, IEEE International Conference of Data Mining, IEEE Computer Society, ISBN-0-7695-1119-8, pp: 433-440,2001.
14. Cengizcolak.M ,Cemizcolak and HasanKocatruk "Predictingcoronary artery disease using different artificial neural network models", CAD and Artificial neural network, pp 249-254, 2008.
15. Chaltrali S. Dangare and Sulabha, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", IJCA, Vol 47(10), pp 44-48, June 2012

Copyright ⓒ 2021SERSC
ISSN: 2005-4238 IJAST