IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Prediction of Heart Disease by Clustering and Classification Techniques

Reetu Singh[1], E. Rajesh[2]

[1] Department of Computing Science and Engineering, Galgotias University, Greater Noida, India
[2] Department of Computing Science and Engineering, Galgotias University, Greater Noida, India

DOI: https://doi.org/10.26438/ijcse/v7i5.861866 | Available online at: www.ijcseonline.org

Accepted: 16/May/2019, Published: 31/May/2019

***Abstract*—**Every year 19 million people approximately die from heart disease worldwide. A heart patient shows several symptoms and it is very tough to attribute them to the heart disease in so many steps of disease progression. Data mining, as an answer to extract a hidden pattern from the clinical dataset, are applied to a database in this analysis. All available algorithms in classification technique are compared to each other to achieve the highest accuracy. To further increase the correctness of the solution, the dataset is preprocessed by different unsupervised and supervised algorithms. The two important tasks which are needed for the development of classifier come under data mining and they are clustering and classification. In K-means clustering the initial point selection effects on the results of the algorithm, both in the number of clusters found and their centroids. Methods to enhance the k-means clustering algorithm are discussed. With the help of these methods efficiency, accuracy and performance are improved. So, to improve the performance of clusters the Normalization which is a pre-processing stage is used to enhance the Euclidean distance by calculating more nearer centers, which result in a reduced number of iterations which will reduce the computational time as compared to k-means clustering. Finally, the classifiers are developed with Logistic regression by using the data extracted by K-Means Clustering. The techniques adopted in the design of classifier perform relatively well in terms of classification results better compared to clustering techniques.

***Keywords*—** Data mining, Classification techniques, K-means clustering, Neural Networks, Logistic Regression

I. INTRODUCTION

Among all harmful disease, heart attacks diseases are considered as the most universal. Medical practitioners conduct so many surveys on heart diseases and collect information of heart patients, their disease progression and symptoms. Increasingly are describing patients with common diseases who have typical symptoms. Thus, there is valuable information hidden in their dataset to be taken out. Data mining is the technique of withdrawing hidden information from a large set of database. It helps researchers gain both profound insights of unprecedented understanding and novel of large medical datasets. The most important or we can say the main goals of data mining are prediction and description of diseases. It is attained through the processing of a set of variables (attributes) in the dataset and discovering the future states of remainder variables.

Extracting key information from a large amount of data is simply called data mining. The more appropriately this data mining is nothing but knowledge mining. Knowledge mining doesn't have the exact meaning of data mining, as it doesn't reflect the emphasis on the extraction of data from a large amount of data. In recent days, this data mining, that is "data" and "mining" has become very much popular among researchers. To carry out this data mining process, the following sequence of steps is very much important.

1. Data Web
2. Information Retrieval(Resource discovery)
3. Information Extraction(Selection/Preprocessor)
4. Generalization(Pattern Recognition)
5. Analyze(Validation)
6. Knowledge

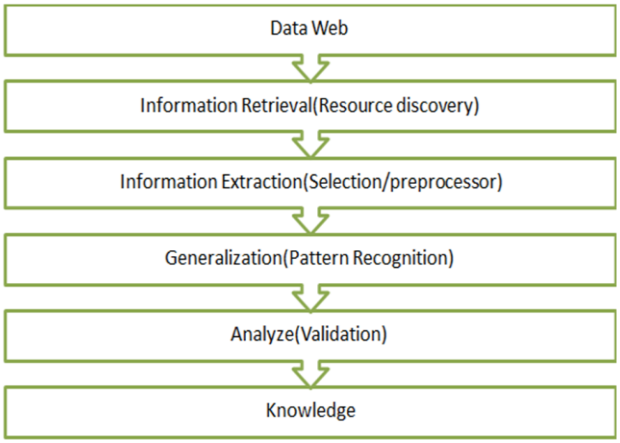
Data Web
Information Retrieval(Resource discovery)
Information Extraction(Selection/preprocessor)
Generalization(Pattern Recognition)
Analyze(Validation)
Knowledge

Figure 1:Data mining process

A. Clustering

Clustering analysis is a method used widely in the data mining community and beyond. This technique Summarizes a very large data set X with much smaller set $C= \{ci| i=1, 2,3 \cdots\cdots.k\}$ of the representative points called as centroids and a membership map $\gamma : X \rightarrow C$ relating every point in X to its representative in C. Various clustering algorithm like hierarchical, EM algorithm, Self organizing Maps, k-means etc. are used to make a set C of Representatives . In this paper, K-means clustering is used.

B. Logistic Regression

The logistic regression is a predictive analysis like all regression analysis; logistic regression is, for the most part, used to depict information and to clarify the connection between one dependent binary variable and at least one ordinal, nominal, interval or proportion level independent variable.

C. Heart Disease

The heart is a very important organ or part of our body. Life is itself dependent on proper working of heart. If operation of heart is not appropriate, it will affect the other body parts of human such as kidney, brain, etc. Heart is nothing more than a pump, which pumps blood through the body. If the circulation of blood in the body is inefficient the body parts like brain suffer and if the heart stops working altogether, death occurs within minutes. Life is utterly reliant on the proficient operation of the heart. The term Heart disease refers to the illness of the heart & blood vessel system in it. There are a number of factors that amplifies the risk of Heart disease [2] such as the Family history of heart disease, Poor diet, Cholesterol, Smoking, High blood pressure, High blood cholesterol, Hypertension, Physical inactivity, Obesity.

D. Symptoms of a Heart Attack

Symptoms of a heart attack are Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone. Anxiety burning at back, jaw, throat. indigestion, or arm Fullness or choking feeling (may feel like heartburn). Some of the common indications are Sweating, nausea, vomiting, or dizziness that also includes anxiety, extreme weakness, or shortness of breath, rapid or irregular heartbeats.

Rest of the paper is organized as follows, section I contains the introduction of the paper, section II contains the Background and Literature survey of the Paper, section III contains the Proposed research methodology of the paper in this section we tried to explain our proposed algorithm. Section IV contains all the experimental results of this research work, In Section V we try to compare the results of all the algorithms used in this paper for the research purpose and Last section, section VI concludes the research work.

II. BACKGROUND AND LITERATURE SURVEY

In any nation to have a progressive development in all the sectors, it is very important and particularly to have attention towards the health of their population. According to that, heart diseases and myocardial ischemic (Joe-Air Jiang et al. 2006) are the most common heart diseases which can lead to serious conditions and cause of death in most of the industrialized countries (Minami et al. 1999).These kinds of heart related diseases can be detected from the information taken from the so many heart attributes and they normally give the data related with health conditions of the patients. It is very much essential to enhance the patients living condition and treatment.

In this paper, authors proposed k-means clustering algorithm for predicting myocardial ischemic. Myocardial Ischemic is common disease in the world. This system extracts hidden information from historical datasets of heart disease. In this paper, number of input attributes is used for analyzing prediction systems for heart disease. Efficiency of output is increased by using k-means clustering. This method for predicting patients with heart disease is the most effective one. It is noticed various classifiers are frequently utilized in different studies to predict heart disease. Therefore, a comprehensive comparison of classification algorithms practically provides an insight into classifier performances. This comparison plays great importance to medical practitioners who desire to predict heart failure at a proper step of its progression. In any nation to have a progressive development in all the sectors, it is very important and particularly to have attention towards the health of their population. According to that, heart diseases and myocardial ischemic (Joe-Air Jiang et al. 2006) are the most common heart diseases which can lead to serious conditions and cause of death in most of the industrialized countries (Minami et al. 1999).These kinds of heart related diseases can be detected from the information taken from the so many heart attributes and they normally give the data related with health conditions of the patients. It is very much essential to enhance the patients living condition and treatment.

In this paper, authors proposed k-means clustering algorithm for predicting myocardial ischemic. Myocardial Ischemic is common disease in the world. This system extracts hidden information from historical datasets of heart disease. In this paper, number of input attributes is used for analyzing prediction systems for heart disease. Efficiency of output is increased by using k-means clustering. This method for predicting patients with heart disease is the most effective one. It is noticed various classifiers are frequently utilized in different studies to predict heart disease. Therefore, a comprehensive comparison of classification algorithms practically provides an insight into classifier performances. This comparison plays great importance to medical practitioners who desire to predict heart failure at a proper step of its progression.

III. PROPOSED RESEARCH METHODOLOGY

In k-mean clustering algorithm, the goal is to find groups of data (data is unlabeled) and after that functions are clustered based on feature similarity by using Euclidian distance formula. In this paper, the quality of clusters is increased by enhancing the Euclidian distance formula. The enhancement that has to do will be based on normalization. Normalization which is a pre-processing technique will enhance the accuracy and efficiency of clusters by calculating best distances from the dataset which will result in more accurate center points and as a result best clusters are formed, the feature which is added is for calculating normal distance metrics on the basis of normalization. The proposed technique is implemented in MATLAB.

A. Working

1. Firstly we have loaded or generated the user-defined dataset in MATLAB.
2. Dataset is scattered and plotted in nature.
3. Now, we can apply k-means clustering on the generated dataset in which we can use the Euclidean distance formula for calculating centroids.
4. After applying k-means clustering accuracy of centers is calculated.
5. For classifying the dataset Logistic Regression is applied.
6. Now, after applying Normalization on the data in which iterations process started, and more nearest and accurate centers are calculated.
7. Normalization: It is a scaling technique or a pre-processing stage. Where we can discover a new range from a current one range. It can be useful for prediction or forecasting reason. Min-Max Normalization: Min-Max Normalization changes an esteem A to B which fits in the range [C, D]. It is given by the formula below: B = ((A − minimum value of A) / maximum value of A- minimum value of A)*(D-C) + C A=Original data point B=Normalized data point [C, D]= specified range.
8. Again accuracy is calculated in which we got better accuracy of clusters than k-means clustering.

IV. EXPERIMENTAL RESULTS

In this paper Heart disease, dataset is used for the research process and prediction analysis. The dataset is plotted between two attributes of Heart disease that is: Patient age and His/her Cholesterol. Cholesterol enables your body to construct new cells, protect nerves, and produce hormones. Regularly, the liver makes all the cholesterol the body needs. Yet, cholesterol additionally enters your body from sustenance, for example, animal-based nourishment's like milk, eggs, and meat. A lot of cholesterol in your body is a danger factor for heart disease.

Some suggest that everybody over age 20 ought to get their cholesterol levels estimated at least once every 5 years. The test that is performed is a blood test called a lipoprotein profile. That incorporates:

- Total cholesterol level
- LDL (the "bad" cholesterol)
- HDL (the "good" cholesterol)
- Triglycerides

Here's how to interpret your cholesterol numbers:

Table 1: Cholesterol numbers

Total Cholesterol	Category
Less than 200 to 240	Borderline High
240 and Above	High

With the help of this above table I can categorized the data and predict the result.

Step 1: Dataset is plotted

A scatter heart disease dataset is first loaded in MATLAB then plotted. I have added grid for easy viewing.

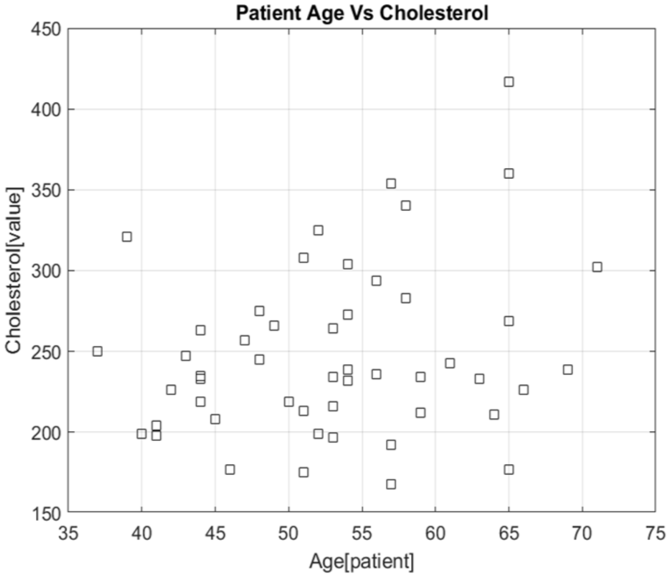

Figure 2: Data is Plotted

Step 2: K-means clustering is applied to the dataset.

Now the K-Means clustering is applied on the given dataset with k=2. The dataset is divided into 2 clusters, In k-means we randomly initialize centroids from the dataset and then Euclidean distance is calculated of each data point from centroids and depending upon the minimum distance between the centroids and data points, that data point is assigned to that centroids, and repeat again these steps till we get the same centroids that is no change in the centroids. In this way, two clusters are formed shown in figure.

3.

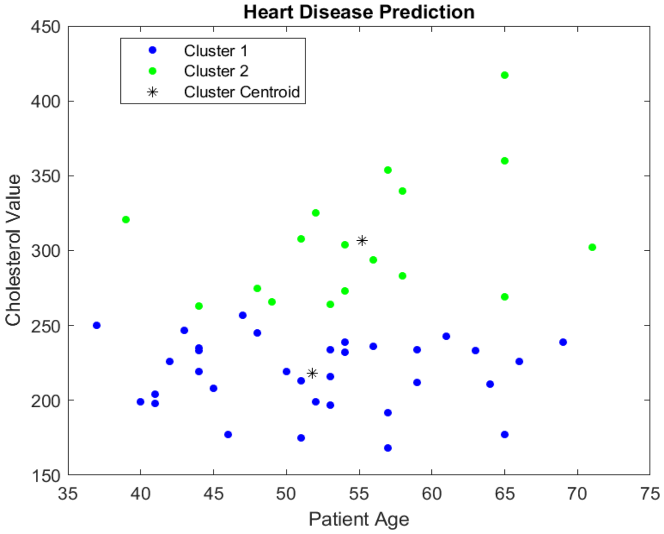

Figure 3: K-means clustering on data

Step 3: Classifier is applied on dataset

The Logistic Regression is applied on the dataset in a 2-d plot. The Logistic Regression is better because when I applied it to my dataset then I can easily get the values in the form of Borderline High and High, Borderline High means cholesterol in between less than 240 and High means cholesterol is more than 240.

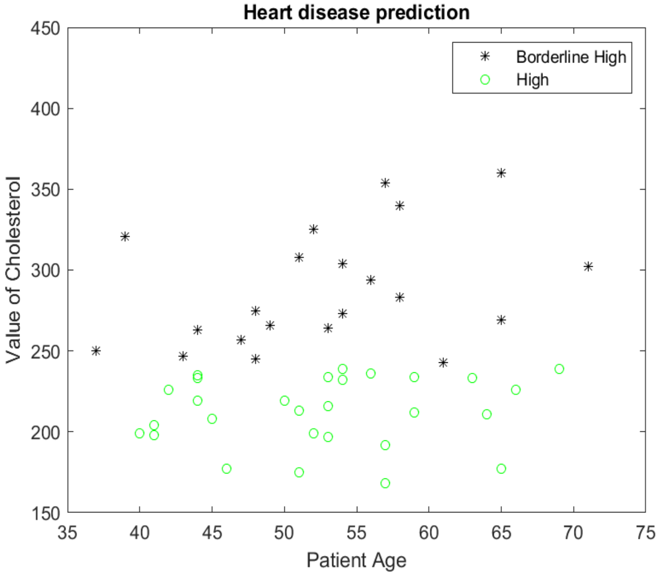

Figure 4: Dataset is classified

Step 4: Normalization is applied on the dataset

Normalization is applied to the dataset. While performing K-means clustering Euclidean distance is calculated for forming similar groups or clusters. But the clusters which we get are not accurate, here the Euclidian distance lacks, so Normalization(Min-Max) in which some range is specified is applied on the dataset for calculating the best distance for more nearer centers and best clusters are calculated according to the number of iterations. We have taken here 2 clusters, which are dataset is divided into two groups. It depends upon us how many clusters we want in our output.

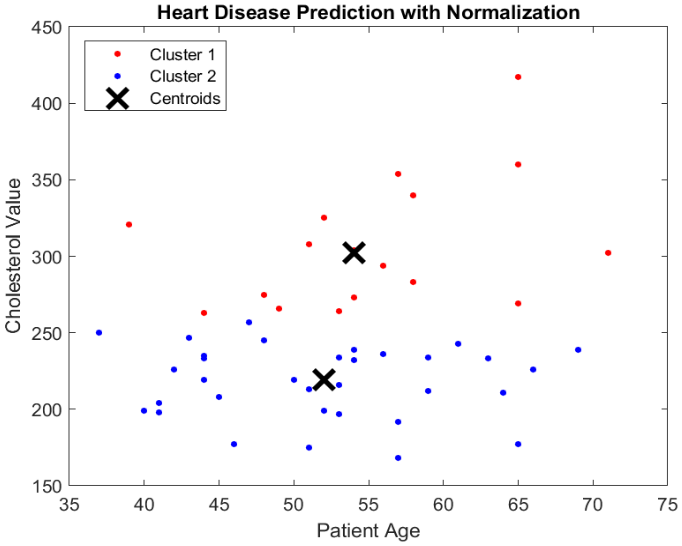

Figure 5: Normalized dataset

V. COMPARISON OF THE RESULTS

Table 2: Comparison of Clustering without Normalization, with Normalization and Classification

Algorithms	Clustering without Normalization	Clustering with Normalization	Classification Result
Accuracy	70.58%	84.84%	90%

VI. CONCLUSION

Heart disease is one of the biggest problems in now a day which leads to causality. Prediction of heart diseases is possible only by the consideration of attributes. This analyzing method of the attributes can be achieved by the inclusion of data mining techniques. Data mining methodologies embrace methods such as Neural networks, naïve Bayes, clustering mechanisms, classification, big data, etc. Further implementation has to be done in order to predict heart disease in a big data environment. The prediction analysis is the technique in which the user predicts the future on the basis of current situations. The prediction analysis consists of two steps. The first step is of clustering, which will cluster the similar and dissimilar type of data. The second step consists of classification which will classify the clustered data for the prediction analysis. In this work, K-mean clustering is used for the clustering. The Logistic Regression classifier is used to classify the dataset for predicting the complex data. The k-mean clustering consists of two steps. In the first step, the arithmetic mean of the loaded dataset is calculated which will be the centroid point. In the second step, Euclidean distance from the centroid point is calculated which defines the similarity between the data points. The accuracy of clustering and classification is reduced when some points remain un-clustered or wrongly clustered. In this work, the technique of Normalization is being applied which will reduce the complexity of large datasets, will calculate Euclidean distance in a dynamic manner and retain maximum accuracy as normalization works better when dataset is complex in nature. Normalized outcomes make the information reasonable for specific analysis and expectation to be performed. The proposed improvement leads to increase the accuracy of classification. The proposed improvement and existing technology are being implemented in MATLAB and it is being analysed that accuracy is increased.

References

[1] de Carvalho Junior, Helton Hugo, et al. "*A heart disease recognition embedded system with fuzzy cluster algorithm."* Computer methods and programs in biomedicine 110.3 (2013): 447-454.

[2] Mirmozaffari, Mirpouya, Alireza Alinezhad, and Azadeh Gilanpour. "*Data Mining Apriori Algorithm for Heart Disease Prediction.*" Int'l Journal of Computing, Communications & Instrumentation Engg (IJCCIE) 4.1 (2017).

[3] Khaing, Hnin Wint. *"Data mining based fragmentation and prediction of medical data."* Computer Research and Development (ICCRD), 2011 3rd International Conference on. Vol. 2. IEEE, 2011.

[4] Patel, Ajad, Sonali Gandhi, Swetha Shetty, and Bhanu Tekwani. *"Heart Disease Prediction Using Data Mining."* (2017).

[5] Wghmode, Mr Amol A., Mr Darpan Sawant, and Deven D. Ketkar. *"Heart Disease Prediction Using Data mining Techniques."* Heart Disease (2017).

[6] Vijayashree, J., and N. Ch SrimanNarayanaIyengar. *"Heart disease prediction system using data mining and hybrid intelligent techniques: A review."* Int. J. Bio-Sci. Biotechnol 8 (2016): 139-148.

[7] Singla, Meenu, and Kawaljeet Singh. *"Heart Disease Prediction System using Data Mining Clustering Techniques."*

[8] Cp, Prathibhamol, Anjana Suresh, and Gopika Suresh. *"Prediction of cardiac arrhythmia type using clustering and regression approach (P-CA-CRA)."* Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017.

[9] Banu, NK Salma, and Suma Swamy. *"Prediction of heart disease at early stage using data mining and big data analytics: A survey."* Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 016 International Conference on. IEEE, 2016.

[10] Mane, Tejaswini U. *"Smart heart disease prediction system using Improved K-means and ID3 on big data."* Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on. IEEE, 2017.

[11] Senthil Kumar, B., and Dr Gunavathi R. *"A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis."* IJARCCE 5 (2016): 463-467.

[12] Kumar, B. Senthil, and R. Gunavathi. *"Comparative and Analysis of Classification Problems."* Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org 7.8 (2017).

[13] G. Kesavaraj, Dr. S.Sukumaran*,"A Study on Classification Techniques in Data Mining"*,IEEE-31661.

[14] F.U.Siddiqui, N.A.Mat Isa, *"Optimized k-means clustering algorithm for image segmentation",* School of Electrical and electronic engineering, university Sains Malaysia, 14300, Nibong Tebel, Penang, Malaysia, 2012.

[15] Shital A. Raut and S. R. Sathe, *"A Modified Fastmap KMeans Clustering Algorithm for Large Scale Gene Expression Datasets",* International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 1, No. 4, page 120- 124, November 2011.

[16] Tapas kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y.Wu*,"An Efficient k-Means Clustering Algorithm: Analysis and Implememtation*", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.24, No.7,2002.

[17] Kapil Joshi, Himanshu Gupta, Prashant Chaudhary, Punit Sharma*," Survey on Different enhanced K-means Clustering Algorithm"*, International Journal Of Engineering Trends And Technology, Vol. 27 ,No. 4-September 2015.

[18] Adil Fahad, Najlaa Alshatri, Zahir Tari, Addullah Alamri, Ibrahim Khalil, albert Y.Zomaya, Sebti Foufou, Abdelaziz Bouras, *"A survey of Clustering Algorithms for Big Data: Taxonomy And empirical Analysis",* IEEE TTANSACTION ON Emerging Topics in Computing.

[19] Jiawei Han, Micheline kamber, Jian pei, *"Data Mining Concepts and Techniques",* Third Edition, © 2012, Elsevier Inc.

[20] K.Rajalakshmi,Dr.S.S.Dhenakaran,N.Roobin*"Comparative Analysis of K-Means Algorithm in Disease Prediction",* International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, Issue 7, July 2015. [

[21] BV Sumana, T.Santhanam*,"Prediction of diseases by Cascading Clustering and Classification",*International Conference on Advances in Electronics , Computers and Communication(ICAECC),2014.

[22] K. R. Lakshmi, M. V. Krishna and S. P. Kumar*, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability,"* International Journal of Scientific and Research Publications, ISSN 2250-3153, Vol.3, Issue.6, June 2013.

[23] M. Kumari, R. Vohra and A. Arora, *"Prediction of Diabetes Using Bayesian Network,"* International Journal of Computer Science and Information Technologies, Vol. 5 (4), 5174-5178, 2014. M. A. Banu, B. Gomathy, "Disease forecasting system using data mining methods," in IEEE International Conference on Intelligent Computing Applications (ICICA'14), pp. 130-133, 2014. https://doi.org/10.1109/icica.2014.36

[24] B. Bahrami, and M. H. Shirvani, *"Prediction and Diagnosis of Heart Disease by Data Mining Techniques,"* Journal of Multidisciplinary Engineering Science and Technology (JMEST), ISSN: 3159-0040, Vol. 2, Issue 2, February 2015.

[25] I. H. Witten and E. Frank, *"Data Mining Practical Machine Learning Tools and Techniques,"* Morgan Kaufman Publishers, 2005

[26] R. Jyoti, G. Preeti, *"Analysis of Data Mining Techniques for Diagnosing Heart Disease,"* International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 5, ISSUE. 7, July 2015.

[27] K. Manimekalai, *"prediction of heart disease using data mining techniques,"* IJIRCCE, Vol.4, Issue 2, February 2016.

[28] A. Durgadevi and K. Saravanapriya, *"comparative study of data mining classification algorithm in heart disease prediction,"* international journal of recent research in mathematics computer science and information technology, Vol.2, Issue 2, March 2016.

[29] O. Y. Atkov, *"Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters,"* Elsevier, 2012.

[30] R. Alizadehsani, J. Habibi, M. Hosseini, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z. Alizadehsani, *"A Data Mining Approach for Diagnosis of Coronary Artery Disease, "* Elsevier, 2013.

[31] S. U. Amin, K. Agarwal, and R. Beg, *"Genetic neural network based data mining in prediction of heart disease using risk factors,"* presented at the IEEE Conference on Information & Communication Technologies, 2013.

[32] Priyanka, Sana Khan, Tulsi Kour "*Investigation on Smart Health Care Using Data Mining Methods"* International Journal Of Scientific Research in Computer Sciences and Engineering 2016.

Authors Profile

Reetu Singh ,pursuing M.Tech in computer Science And Engineering (Final Year) in Galgotias Universi -ty .I have done my B.Tech from United Institute Of Technology,Allahabad.

Dr.E.Rajesh, Assistant Professor at galgotias University". Head Of Department in School of Information Technology (2016-2018) SRM University, Sikkim,Gangtok, East Sikkim. Assistant Professor in Computer science and Engineering(2014-2016) Sri Jayram Institute Of Engineering and Technology, Cuddalore, Tamilnadu Lecturer in Computer Science and Engineering (2006- 2008) IFET college of Engineering ,Villupuram, Tamilnadu . 10 years Of teaching Experience.

.