International Journal of Education and Science Research Review

E-ISSN 2348-6457

Volume-10, Issue-3 May-June-2023 P-ISSN 2349-1817

www.ijesrr.org Email- editor@ijesrr.org

DATA MINING TECHNIQUES IN PREDICTING HEART

Amandeep kaur

Assistant professor

Punjab Institute of technology, Nandgarh

Maharaja Ranjit Singh University, Bathinda

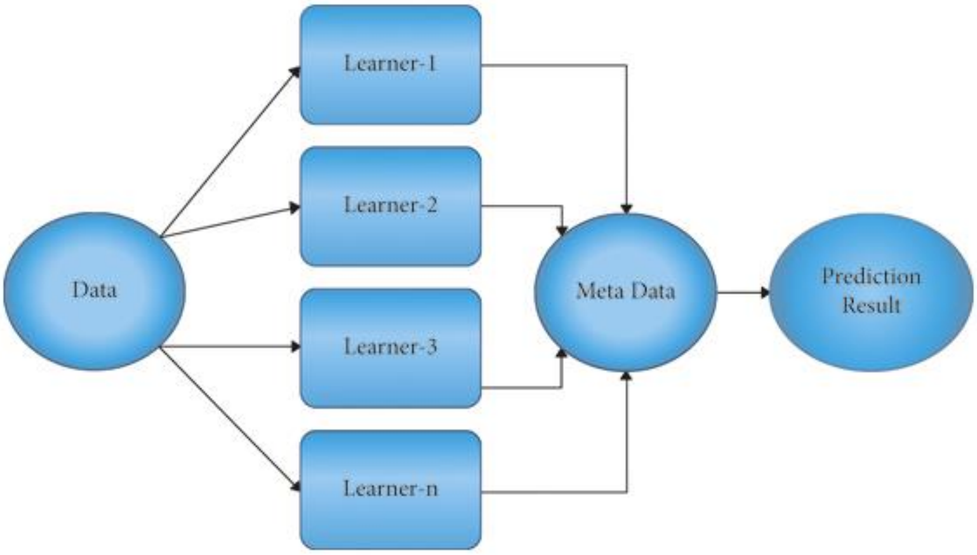
ABSTRACT

A regression model serves as the essential building block for every predictive analytics platform. The linear regression model investigates the relationship between one dependent (also known as a "response") variable and a number of independent (also known as a "predictor") variables. When it comes to medical treatment, there is a wealth of information available. The process of obtaining usable information from a large data set in order to aid informed decision making and prediction is referred to as "health data mining." In numerous of the currently ongoing studies, data mining strategies have been utilised for the purpose of predicting coronary disease. On the other hand, there hasn't been a lot of study done on the important components that go into determining whether or not someone would get cardiovascular disease. It is absolutely necessary to pick the appropriate group of important qualities that can improve the performance of the prediction models. This study aims to identify critical characteristics and data mining methodologies in order to improve the accuracy of cardiovascular disease forecasting. k-Nearest Neighbours, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network, and Vote (a hybrid technique combining Nave Bayes and Logistic Regression) were used to develop prediction models with varied feature combinations. Vote was a hybrid approach combining Nave Bayes and Logistic Regression. Data mining methods, such as Decision trees (J48), Bayesian classifiers, Multilayer preceptor, Simple logistic, and Ensemble approaches, are used to identify heart problems. In this investigation, we examine the usefulness of several data mining categorization strategies by comparing and contrasting them, with the end goal of constructing a database consisting of health-related data. The results of the classification will be shown visually utilising a wide variety of representation methods, such as two-dimensional diagrams, pie graphs, and others. The above described computations are assessed and studied in order to determine their correctness, time utilisation factor, region under ROC, and other relevant metrics.

Keywords: *Data mining; prediction model; classification methods; feature selection; heart disease prediction*

INTRODUCTION

One of the causes of death that is on the rise at a pace that is being estimated to be the quickest worldwide is cardiovascular disease, which is also one of the reasons of death that is on the rise. It is estimated that 17.9 million individuals lost their lives as a consequence of its effects in the year 2017, which accounts for around 15% of all deaths that can be attributed to natural causes. This information comes from projections made by the World Health Organisation. To put it another way, it is responsible for around 15 percent of all deaths that occur across the world. If vital indicators such blood pressure, cholesterol levels, heart rate, and glucose levels are monitored and evaluated on a regular basis, it is feasible to detect cardiovascular illness, also known as cardiovascular sickness, at an earlier stage. This disorder, which manifests itself in a variety of ways and manifests its effects on the cardiovascular system, is also referred to by a number of other names. Not only does cardiovascular disease have a significant impact on a country's economy, but it also has a significant impact on the structures that are in place to provide medical treatment in that country.[1] The effects of this are going to be felt by a very large audience. Researchers and developers working in the field of research and development are currently devoting a significant amount of attention and effort to the process of developing techniques for sickness prognosis that make use of data mining and machine learning. In a similar line, efforts are being made to create and assess algorithms that can assist in diagnosing and forecasting the early stages of cardiac sickness. These efforts are comparable to those that have already been mentioned. These attempts are very much like the ones that were detailed in the phrase before this one. These efforts are quite comparable to the ones that were discussed in the sentence that came before this one. The development of these algorithms is now underway, and it is being accomplished by the use of a variety of approaches, including data mining, machine learning, and hybrid techniques.

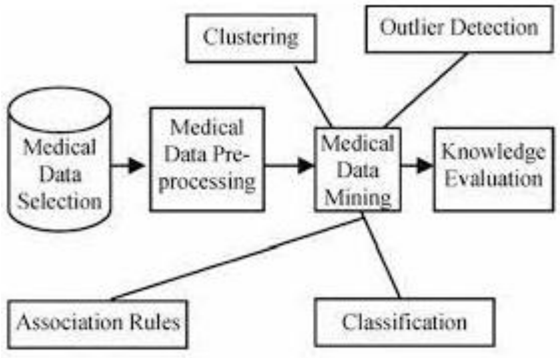


This examination makes use of artificial neural networks (ANN), K closest neighbours (KNN), genetic algorithms (GA), Naive bayes (NB), and decision trees (DT) in order to carry out an analysis of prior research on the categorization of the dataset. This analysis is carried out with the help of decision trees (DT). When ANN was compared to other classifiers that were trained on the same dataset, it was found that ANN required a smaller number of parameters for feature extraction and achieved a higher level of accuracy for heart prediction.[2] This was a result of the fact that ANN was capable of learning from the data in a manner that was more natural. This was the realisation that resulted from evaluating ANN in conjunction with many other classifiers.

After making the comparison, I came to the conclusion that this was one of the most important things that I had gleaned from my research. Because of the work that is scheduled to take place in the not-too-distant future, the ANN database may be upgraded or given new features. It is likely that some of these modifications or upgrades might include those that make it feasible to create more accurate estimates. This is something that is a possibility. Classifiers that are extremely similar to one another have the potential to be important because they can be used to develop prediction algorithms that are much more accurate and, as a consequence, have the potential to save a considerable number of lives. This is because similar classifiers can be used to design prediction algorithms that are significantly more accurate than those now in use. Classifiers that are very analogous to one another have the ability to cut down on the amount of needless deaths.

The process that is used to diagnose cardiac sickness requires both the existence of certain symptoms and the ability to draw conclusions about the patient's health based on the patient's medical history. Both of these are necessary components of the diagnostic process. It is likely that the attending medical experts will need extra time to come at a conclusive diagnosis for the patient in question if the patient is suffering from many disorders. The authors of the study came to this conclusion as a result of their investigation.[3] They arrived at this conclusion as a result of conducting an analysis of the prior research that had been done on the topic of the prediction of cardiovascular illness. The authors of the study came to the realisation that data mining strategies need to be considered as the "gold standard." Classification algorithms are gaining popularity in the field of healthcare, where they were previously uncommon because of their inability to assess massive volumes of data. This is a direct outcome of the capacity of classification algorithms to review large amounts of data. The field of medicine makes use of a wide variety of computer programmers, including but not limited to evolutionary algorithms, naive Bayesian, support vector machine, nearest neighbours, decision tree, fuzzy logic, fuzzy based neural network, artificial neural network, and all of the other varieties of neural networks that have been mentioned up to this point in this discussion. Since the beginning of the 20th century, diseases and disorders of the cardiovascular system have consistently ranked among the highest as leading causes of mortality across the entirety of the world. This pattern has been consistent over the whole of this century. In the year 2015, cardiovascular disease was the cause of death for 17.17 million individuals all over the world, according to the data that was published by the World Health Organisation (WHO).

The results of the worldwide poll that was carried out in 2015 were used to compile this information. When compared to other causes of mortality, either individually or collectively, cardiovascular illnesses take the lives of more people each year than any other cause of death. This is true whether you look at the numbers for individual causes of mortality or for all causes of mortality together.[4] This is the case regardless of whether one looks at the figures for individual causes of death or at the total number of deaths caused by all causes combined. If it were possible to anticipate deaths brought on by cardiovascular disease and offer early warnings for such deaths, then the number of people who passed away as a result of the condition would be far fewer than it is at the present time.



Heart Disease Prediction and Diagnosis System

The utilisation of computer-based methods and algorithms that arrive at conclusions throughout the pertinent phases of the diagnostic process in order to form judgements may make it possible to accomplish breakthroughs in medical diagnosis. These kinds of technological advancements are sometimes referred to by their abbreviation, DSSs, which stands for decision support systems. This is a common practise. Intelligence is another component that plays a role in this scenario and must be taken into consideration because of its importance. These systems provide a contribution to the prediction and diagnosis of the disease by integrating knowledge of the subject matter with information about the patient.

This information helps the system better understand the patient. The clinical data collection process will use this information. The Diagnosis Support System (DSS) aids to the overall enhancement of the quality of medical treatment that is being delivered by providing a diagnosis that can be relied on and is correct. The DSS has the potential to reduce the overall cost of treatment by providing a diagnosis that is both more accurate and more rapid than the methods that are now used, while at the same time reducing the amount of time that is necessary for the surgery itself.[5]

One way in which this may be achieved is by shortening the amount of time needed to complete the procedure. once being made available to the general public and being uploaded to the cloud, these services can be used by any health institution that so chooses once they have been made available to the general public.

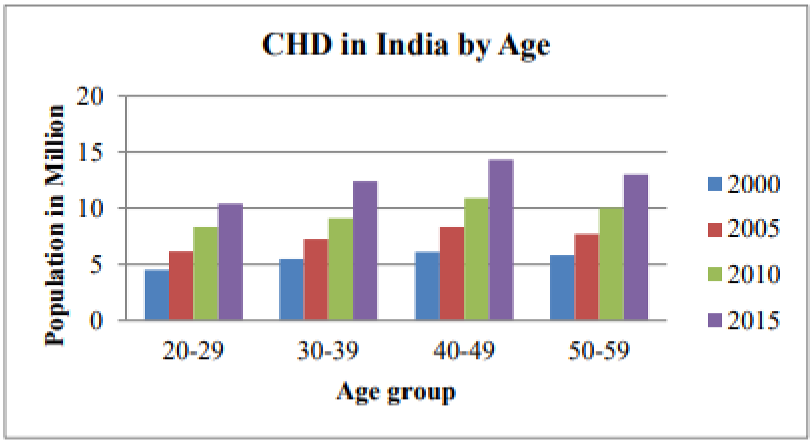


Figure 1. Age Wise Coronary Heart Disease In India

Surviving Hybrid Intelligent Heart Disease Prediction Techniques

Genetic algorithms and fuzzy logic are employed in the process of predicting heart illness, with the former being used for feature selection and the latter being used for both classification and prediction. Fuzzy logic is utilised in the process of predicting heart disease. The authors examine the fuzzy entropy-based method known as the NNTS and contrast it with the recommended strategy, which is known as the GAFL system. The writers of this article compare and contrast the effectiveness of various approaches. Accuracy, specificity, and sensitivity are three of the factors that are taken into consideration throughout the examination. By implementing the recommended technique, accuracy may be increased to 86 percent while simultaneously decreasing the total number of features from 13 to seven. In order to accomplish the objective of developing a method for the forecasting of cardiovascular illness, CANFIS and a genetic algorithm were both employed as tools in the course of the research. his model makes use of a variety of computer approaches, some of which include neural networks, fuzzy logic, and evolutionary algorithms, amongst others.[6]

The strategy that has been proposed not only shortens the amount of time required to finish training, but it also improves the accuracy with which classifications are accomplished. The authors introduced a unique classification approach for cardiovascular disorders that made use of an Artificial Neural Network (ANN) in conjunction with the selection of feature subsets as two of the primary components of the methodology. By picking only a small fraction of the attributes that are available, it is feasible to achieve the aim of streamlining the data collecting process. During the pre-processing phase of the analysis, the Principal Component Analysis (PCA) method is applied.

The findings of the research indicate that the methodology that was proposed fares notably better in terms of accuracy than the standard classification methods do. This study's objective is to research and assess the relative effectiveness of two separate types of learning algorithms, namely feed forward neural networks and back propagation neural networks, in comparison to one another. The weights of the networks will be determined with the assistance of a genetic algorithm, which will make this task significantly simpler.

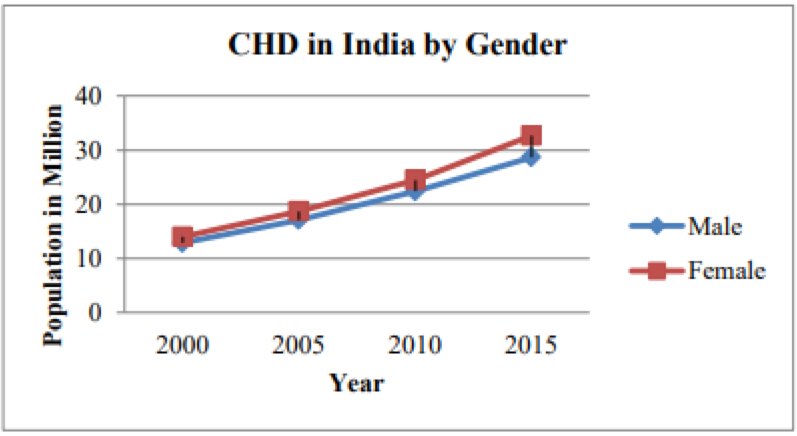


Figure 2. CHD In India Based On Gender

The acronym NN refers to a certain kind of algorithm used in data mining.

A neural network is a parallel and distributed information processing structure that is made up of a large number of processing modules that are linked together by one-way channels of communication that are called connections. These processing modules are connected to one another in a neural network.[7] Nodes are a term used to describe these processing modules that are part of the structure. Every processing node has a single output connection that splits off into many other ones and sends out the identical signal in each and every one of them. This link is severed once it reaches the final node in the processing chain. There are primarily two categories of NN that may be distinguished by the way in which individuals choose to take in information. Both supervised and unsupervised learning can be considered valid methods of acquiring knowledge. During the process of supervised learning, the network will first calculate an output for each input, and then it will evaluate how well that result satisfies the constraints of the job at hand. The weights of the network are changed in accordance with a learning rule if it is discovered that the response that was computed is not the same as the value that was intended. The single-layer perceptron and the multi-layer perceptron are both types of examples that demonstrate supervised learning in action.

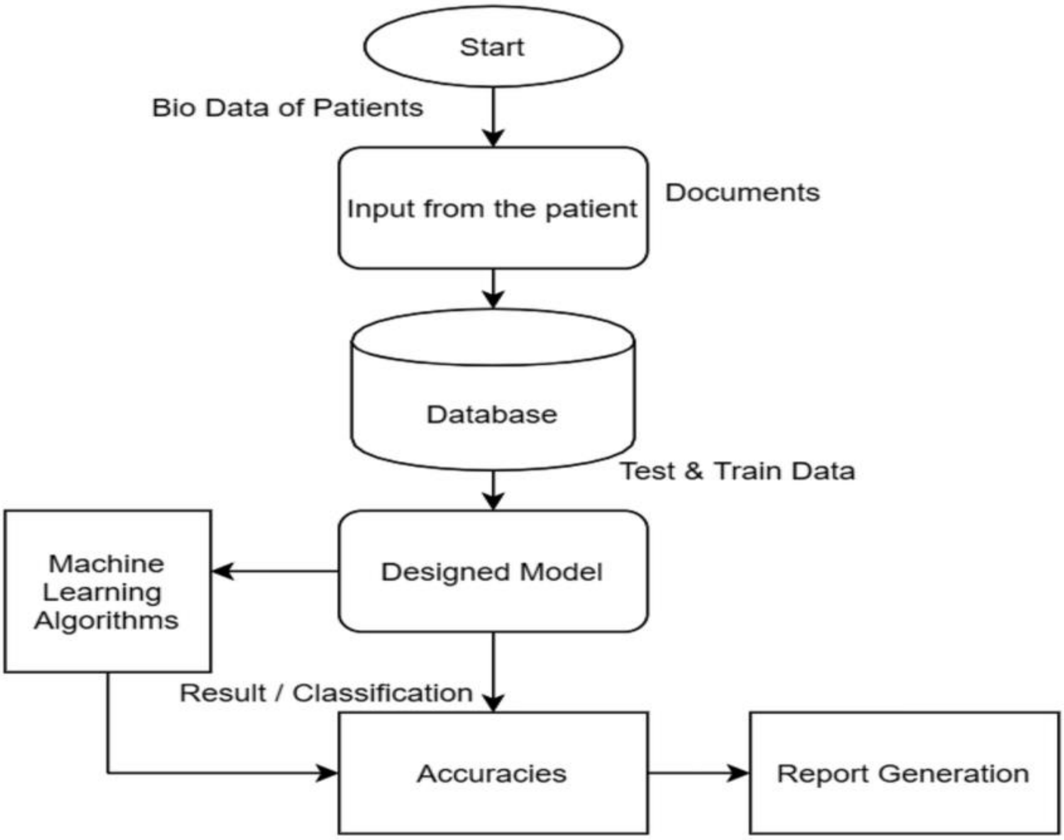
Learning in which the networks learn on their own depending on the particular elements of the situations they come across is referred to as unsupervised learning. One type of learning is known as unsupervised learning, and one example of this type is known as self-organizing feature maps.

Inexperienced A categorization scheme based on Bayesian logic The Bayes theorem is a tool that may be utilised to demonstrate that naïve The Bayesian method is a method of categorization that calculates a probability by looking at the frequency with which particular values and combinations of values have occurred in the past. Thomas Bayes, a mathematician and statistician from the United Kingdom, is the inspiration for the name Bayes. The likelihood of a third event may be estimated using Bayes' theorem, which does this by computing the likelihood of the third event based on the probability of an earlier occurrence. If A is known, then the likelihood of B, also known as the probability of B given A, is equal to the probability of A given B. If A is unknown, then the likelihood of B is not equal to the probability of B given A.

The capacity of this method to estimate the necessary classification parameters while only using a limited amount of training data is one of its most notable strengths.[8] This is the most important factor in determining the algorithm's performance.

Diagram of the Decision

According to the definition of a decision tree provided by Berry and Linoff, a decision tree is "a structure that can be used to divide up a large collection of records into successive smaller sets of records by applying a sequence of simple decision rules." The term "decision tree" refers to "a structure that can be used to divide up a large collection of records into successive smaller sets of records." This is one definition of what a decision tree is. This is only one point of view. This specification was provided so that the decision tree would be able to function properly. A decision tree is another name for "a structure that can be used to divide up a large collection of records into successive smaller sets of records." Another name for "a structure that can be used to divide up a large collection of records," a decision tree is another word for "a structure that can be used to divide up a large collection of records into successive smaller sets of records."



Another definition of a decision tree describes it as "a structure that can be used to divide up a large collection of records." Another explanation of what a decision tree actually is has been provided here. To restate this definition, one could say that a decision tree is "a structure that can be used to divide up a large collection of records." This would be an accurate description. The process that is carried out in order to carry out the divisions has a direct influence on the degree to which the individual components of the sets that are created as a consequence of the divisions become progressively comparable to one another.[9] This is because the technique that is followed in order to carry out the divisions.

This similarity may be traced back to the process that was followed in order to carry out the divisions, which had a direct impact on the phenomenon. The fact that this occurs as a direct result of the procedure that is now being carried out is unavoidable. Iterative dichotomiser 3, often known as ID3 for short, is a sort of decision tree model that builds a decision tree by applying a sequence of training samples that have been selected in advance. This method is known as iterative dichotomiser 3. This model is also known as the Iterative dichotomiser 3 by a large number of people. The iterative dichotomiser 3 approach is the name given to this particular methodology. The iterative dichotomiser 3 approach is what the vast majority of people refer to when they talk about this methodology. When referring to models of this particular ilk, the term "ID3" is frequently applied as a general description because of its widespread use. It's conceivable that you've heard people refer to ID3 by its full name, which is iterative dichotomizer 3, but this use isn't very frequent. It's possible that you've heard people refer to ID3 by its entire name. However, it is probable that you are already familiar with its usage. If so, you're lucky.[10]

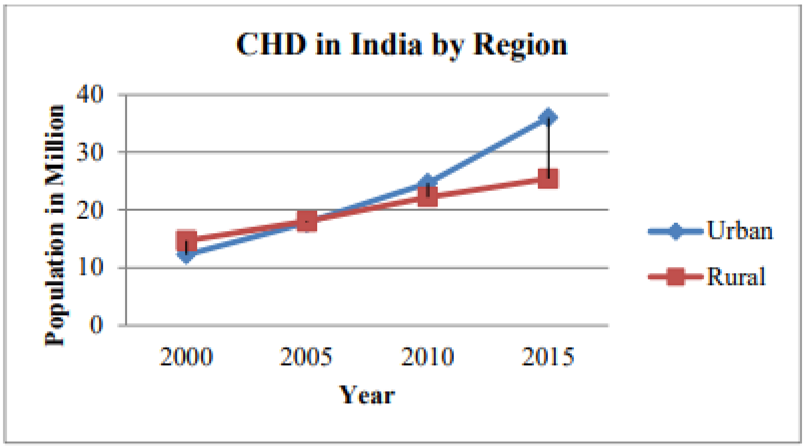


Figure 3. CHD In Urban And Rural India

The newly established ID3 induction mechanism went through a number of iterations before reaching its current state, which is the C4.5 version of the system. These iterations took place over the course of the past several years. The decision tree technique, which was presented for the very first time in C4.5 and is now documented in C5.0, has received a significant boost as a direct consequence of this upgrade, which has directly led to a significant rise in accuracy. This enhancement has resulted in a considerable increase in the number of correct predictions. A more in-depth discussion of this method was previously had and can be found in the C4.5 paper.

The J48 decision tree has been updated to incorporate the ID3 procedure as a candidate for one of the available options.

OBJECTIVES OF THE STUDY

1. To the study of the Data Mining Techniques.
2. To the study of the Heart Disease Prediction using Hybrid Intelligent Techniques

RESEARCH METHODOLOGY

Because of its capabilities, the testing that was included in this research was carried out with the assistance of RapidMiner Studio. It provides an environment for the building of predictive analytic algorithms that is both sophisticated and user-friendly from a visual design standpoint. A pictorial explanation of the procedure that can be comprehended quickly and easily is an example of an educational tool that may be helpful to novices.[11] The use of this kind of educational tool can be beneficial. In addition to this, it encourages the creation of open-source software, which increases not only the availability of the programme but also its capability of carrying out the functions for which it was designed. Figure 2 provides a visual representation of the steps that need to be taken in order to successfully carry out the experiment. RapidMiner was utilised in order to do data analysis on the dataset concerning heart disease that was produced by UCI Cleveland. The dataset was compiled by UCI Cleveland. The initial phase in data mining is known as pre-processing, and it is followed by feature engineering (during which one-of-a-kind combinations of characteristics are selected), followed by classification modelling (during which prediction models are developed), and finally, feature mining (during which unique combinations of characteristics are picked).

The procedure of collecting features and producing models was repeated with each and every conceivable combination of the characteristics that were used. The results of this approach were then analysed. A minimum of three out of a total of thirteen attributes are selected at the beginning of each iteration of the loop, and the model is applied to those specific qualities.

After the process has been finished, an output will be provided that provides information about how well each model performed based on the characteristics and data mining technique that was applied at each stage of the operation. This information will be produced after the procedure has been finished. After the operation has been carried out to its conclusion, you will be given this information.[12]

This section provides a comprehensive breakdown of the processes involved in preparing the data, selecting the features to use in the modelling of the classification, and evaluating the outcomes of the modelling. The results of our assessments of your performance are detailed below in the table that we've provided for your convenience.

DATA ANALYSIS

In this section, the outcomes of the tests are utilised to illustrate how key characteristics and approaches to data mining were chosen. These were decided upon by referring to the findings of the part that came before this one. After we have finished our study and examined its results, we will figure out which aspects of data mining and techniques of data mining are the most relevant when it comes to the building of prediction models for cardiovascular sickness. Specifically, we will determine which are the most significant elements of data mining and methods of data mining.[13]

Display an analysis of the most important components of the data mining methods that were employed for this particular research endeavour, and then proceed to clarify what those tactics were. This will help the reader have a better understanding of the study that was conducted. It is hardly possible to place enough emphasis on how significant this truth is.

Table 1: Comparing the Highest-Performing Attributes

Features Occurrence	Age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
Highest Precision	2	7	7	1	2	5	4	3	4	6		7	5
Highest F-measure	2	7	7	1	2	5	4	3	4	6	4	7	5
Highest Precision	0	6	4	2	1	2	2	2	4	2	4	5	4
Total Occurrence	4	20	18	4	5	12	10	8	12	14	12	19	14

The priority elements that we had picked for our proposed model required the addition of data mining as a crucial component. Based on an analysis that takes into account both the accuracy and the precision of the findings of various experiments, this study provides a ranking of the top three data mining methodologies.[14] The average accuracy and precision of each data mining method was used to choose the top three approaches, which were chosen on the basis of this information. We may draw the following conclusion from the data: these three approaches are among the most accurate and exact procedures that are accessible to us. Vote, Naive Bayes, and Support Vector Machines are going to be utilised in the construction of the heart disease prediction models.

Table 2: Cleveland-Statlog comparison

Comparison Category	Cleveland Dataset					Statlog Dataset	
No. of Attributes	13					13	
Attributes	age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal					age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal	
Class Attribute	num					num	
Different values for “num”	0,1,2,3,4					1,2	
Distribution of “num”	0	1	2	3	4	1	2
	164	55	36	35	13	150	120
Records with Missing Values	6					0	
Total number of instances	303					270	

The comprehensive method that was used when carrying out the investigation that was being questioned. It was necessary to do preprocessing on the Statlog dataset before it could be utilised for analysis. In order to ensure that the Statlog dataset is consistent with the Cleveland dataset, the value of the class variable "num" was altered from "1" to "0" and from "2" to "1." The result that was expected from the produced data set only had two potential values: zero if there was no evidence of heart disease, and one if there was. After the transformation has been implemented, the original 270 records are now divided between '0' and '1' (respectively 150 and 120), creating a balanced distribution. After that, the data was organised so that it could be utilised in a scenario including classification.[15]

CONCLUSION

Cardiovascular disease is responsible for the deaths of millions of people every single year. This situation is only going to get worse as time goes on and as the globe and its inhabitants grow. On the other hand, if the illness can be forecasted in advance, it will be possible to save a great number of lives. In order to diagnose diseases at an earlier stage, data mining has shown to be a very helpful tool. The process of data mining allows for the extraction of information from patient records, such as possible risk factors for cardiovascular disease and the best way to organise a prediction system.

REFERENCES

1) Anooj, P. K. 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University-Computer and Information Sciences, 24(1), 27-40.

2) Bhatla, N., & Jyoti, K. 2012. An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.

3) Chaurasia, V., Pal, S., 2013. Early prediction of heart diseases using data mining

4) techniques. Carib. J. SciTech. 1, 208-217.

5) Dey, A., Singh, J., Singh, N., 2016. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. Analysis. 140(2), 27-31.

6) Dua, D., Karra Taniskidou, E., 2017. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

7) El-Bialy, R., Salamay, M. A., Karam, O. H., Khalifa, M. E., 2015. Feature analysis of coronary artery heart disease data sets. Procedia Computer Science. 65, 459-468.

8) Ismaeel, S., Miri, A., Sadeghian, A., Chourishi, D., 2015. An Extreme Learning Machine (ELM) Predictor for Electric Arc Furnaces' vi Characteristics. IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), New York, pp. 329-334.

9) Kavitha, R., & Kannan, E., 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, pp. 1-5.

10) Khemphila, A., & Boonjing, V. 2011. Heart disease classification using neural network and feature selection. In 21st International Conference on Systems Engineering (ICSEng), Las Vegas, pp. 406-409. IEEE.

11) Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. 2017. A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Computational and mathematical methods in medicine, 2017.

12) Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P., 2013. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications. 40(1), 96-104.

13) Nahato, K. B., Harichandran, K. N., & Arputharaj, K. 2015. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and Mathematical Methods in Medicine, 2015, 1-13.

14) Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. A. H. 2016. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 145-150. IEEE.

15) Sen, A. K., Patel, S. B., & Shukla, D. D. 2013. A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level. International Journal Of Engineering And Computer Science (IJECS), 2(8), 2663-2671.