Data Mining Techniques to Construct a Model: Cardiac Diseases

Noreen Akhtar, Muhammad Ramzan Talib, Nosheen Kanwal
Department of Computer Science
Government College University
Faisalabad, Pakistan

***Abstract*—Using echocardiography flexible Transthoracic Echocardiography reported data set detecting heart disease by using mining techniques designed prediction model the data set can develop the reliability of analysis of cardiac diseases by echocardiography, using eight iterative and interactive steps consisting Knowledge Discovery in Database (KDD) methodology including from 209 patients with echocardiography to extracting the data important mode of action Transthoracic Echocardiography inspection report. This study used data from Faisalabad Institute of Cardiology study from 2012 to 2015. All models exposed the results of J48 decision tree, naïve bayes classifier and neural network that has extraordinary classification precision and predictive of heart disease cases are generally comparable. However, J48 model predictive classification accuracy shows of 80% based on the true positive rate ratio and performance slightly better. This study shows to predict heart disease cases and People can be used the results of our study to make more consistent diagnosis of cardiac disease and to help them as a support tool for cardiac disease specialists.**

Keywords*—Knowledge Discovery in Database (KDD); data mining; decision trees; neural networks; Bayesian classifier; heart disease

I. Introduction

Heart disease causes higher mortality rate in our Pakistan. In our country the male and female having the age 65-year-old they are facing the heart disease. Data mining technology technique is used to decrease cardiac disease in entirely over the world. In this study, researcher can easily identify heart diseases by skillfully doctor through extreme risk factors. To choose the best predictive method researcher use various data mining techniques to predict cardiac diseases at this end. The Manimekalai [1] says that different risky aspects in the manner that smoking, high blood pressure, diabetes, obesity did not increase heart diseases.

A. Discovering Knowledge for KDD process

Now a days, more information or data but lack of knowledge have health department. The researchers use the huge volume of information to the medical prediction of heart diseases treatment by Knowledge Discovery in Database (KDD).

In Fig. 1, researchers used a knowledge Discovery Database (KDD) approach to develop predictive models made by transthoracic echocardiography for predicting heart disease cases based on measurements. The data mining consists of nine steps that project's life cycle.

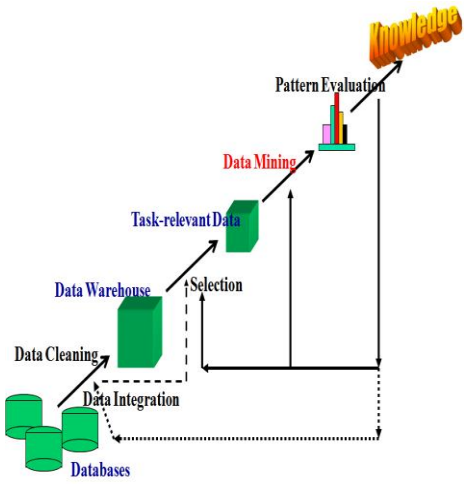

Fig. 1. Discovering knowledge for KDD process.

The researcher uses the various data mining techniques and compare with the same dataset to predict these techniques. Then I had to choose the best techniques to predict [2] Soni *et al.* says that Men and women almost equally affected low income countries out of the current CVD death of 82%. Low income society is unequally impressed. CVD is the most common disease about 2360 million folks died by CVD. Mostly heart disease and hit lead to social death. The major will be Eastern Mediterranean region percentage present. Southeast Asian lifestyle, work culture and eating habits change because the death clang increased the most. As a result of which changes social lifestyle reduces the disease.

The aim of this paper is extracting a data set bad or good key mode or feature. For heart disease diagnosis, we choose and identify the more relevant attributes. Decision tree, Neural Network and Bayesian classification compared to predict heart disease cases. With the help of domain experts, we chose the model to explain and analyze the results.

II. Review of Literature

Shafique, et al. [3] studied the data mining is the region that reviews which implies that data and knowledge are helpful from past information. There are various strategies for information mining. Data mining can be utilized as a part of various regions including medical utilize. Heart or cardiovascular disease is a hot topic in the global health care industry. Chandna [4], explained the health of the professional data mining is effective in forecasting disease. The number of

test numbers must be from the patient to detect the necessary conditions for the diseases in any case, utilizing certainties mining innovation can lessen the quantity of exams required. Lakshmi, et al. [5] conducted heart diagnosis dirty disease is an important medical and annoying task. Healthcare department is commonly believed that "information rich" and "knowledge-poor". Alizadehsani, et al. [6] conducted experiment Cardiovascular disease is often very rare and is the important reason of decease. The fundamental sort of these sicknesses as Coronary Artery Disease (CAD) and the determination is essential. It has many side effects that are expensive and angiography is more accurate CAD Coronary Artery Disease diagnostic method. The existing research from the patient to collect data using several characteristics and the use of different data mining algorithms to achieve high precision side effects cost of the method. Gamberger et al. [7] studied that with the aid of information mining model Intelligent Heart Disease Forecasting System (IHDPS) innovation workmanship for example, decision trees, naïve Bayes and neural system. Chaurasia and Pal [8] explained that the death of the history of the largest study shows that heart disease has gradually become the world's number one killer. Death age group occurs from 25 to 69 years old about 25% due to heart disease. In the event that all age bunches are incorporated coronary illness represents around 19% of all passings. Muthukaruppan and Er [9] explained that Particle Swarm Optimization (PSO) founded fuzzy master framework aimed at the finding of Coronary Artery Disease (CAD). The framework is outlined in light of informational indexes of Cleveland and Hungarian coronary illness. Yeh et al. [10] studied that acquired 493 legitimate examples from expectation and conduct programs that cerebro vascular ailment and embraced three order calculations, decision trees, Bayesian classifier and BP neural system, to construct an arrangement demonstrate individually. Hand et al. [11] explained that artificial neural network is a highly parametric statistical model has attracted considerable attention in recent years. In the artificial neural network is a highly parameterized fact that they are actual springy so that they are correctly functioning with irregular model insignificant. Pham et al. [12] conducted an experiment to decision tree algorithms have been utilized as a part of numerous applications arrangement for example, comfort medication assembling and creation money related investigation, stargazing and sub-atomic science.

Khemphila and Boonjing [13] xplained that given meaning tree "which can be used to divide a large number of structures through over-application of simple sequence records gathered to decrease continuously record set decision-making rules. The KDD procedure demonstrate embraced in this examination along these lines as indicated by [14] Han and Kamber, sub-class is to locate a work of art (or process reason) depict and recognize information projects or thoughts keep in mind that end goal to foresee motivation behind the question class of the model can be utilized Its class tag is unidentified. Weka table is the first country into a set of data preprocessing algorithms and machine learning tools. It includes almost all popular algorithms. Its design allows you to quickly try new methods in a flexible way the existing method [15] Frank, et al. The data mining goal standard data collection strategy play no role. This is a lot of data mining statistical data where data is frequently used effective strategies to answer specific questions and collect different types of the method. Data mining is frequently called "secondary" information and for this reason investigation [16] Hand, et al.KDD focuses on data from including how data is stored and access, how the algorithm is extended to large data sets known to the whole process of knowledge discovery still operate effectively, how to solve interpretation and visualization of results and how to effectively support and overall modeling Robot Interaction [17] Sences Applied, says the attention of this paper by using data mining tools and techniques, particularly development of analytical models which can be identified in the situation of general predictive cardiac diseases classification technology. The experiments have been conducted, on the data which was collected from the Faisalabad Institute of cardiology hospital from 2012 to 2015.

III. Material and Method

The purpose of this study by applying classification techniques to detect heart disease and attempting to build up a forecast displays in view of decision tree, neural system and Bayes classifier. In this paper, researcher has done citation valuable information from the heart hospital Faisalabad Institute for the collection of Faisalabad institute of cardiology data including 4 years of validity [2012-2015] data cleaning, data selection, data conversion and data mining. Where in the presence of this paper was realized, and different prediction methods were used for the age of disease data in each step the value of chest pain, resting blood pressure, blood sugar and different steps resting electrocardiogram result, maximum speed of the heart rate, exercise angina, diseases and display the capability of data mining technology to predict the values.

A. Data Pre-Processing Steps

1) Data Cleaning

At this stage, we have to recover the missing data from the large amount of the datasets. Researches clean the data remove the data redundancy and recovered the missing values of the data. We had prepared the data according to appropriate format for data mining.

2) Data Selection

In this step, the applicable analytical data is determined from the data set to be retrieved. The second data compression technique applied to the data set is the attribute selection.

3) Normalization

The data is scale within small range for example 1 to 0 or 0 to1 and fall in only small range.

4) Attribute Construction

The new attribute is built in the dataset and add the new attribute in the given set that is used for mining data.

B. Knowledge Discovery

There are many data mining techniques that are used for statistical data mining and techniques for example outlier analysis, clustering, prediction and classification and association rule.

C. Outlier Analysis

Information libraries can contain general conduct or occasionally utilized information model of the information question. These information objects are special case ranges. It is first applied to the early removal of outliers to avoid its impact on other mining methods.

D. Clustering

In this research paper we have used K-means clustering. *K*-means clustering is come in unsupervised learning. The k-means clustering is used to grouping the data on the base of similarity.

E. Classification

Sub class is used to describe and differentiate the data to find the class/concept is to use the model for predicting the object class and its class label is unknown process models. The classification models are IF-THEN rules, J48 decision tree, Neural Network and Naïve Bayes and can be expressed in these forms.

F. Prediction

Forecast has been complicated in quite a lot of attention given the success of forecasting business setting. However, predictions of the time related data missing, or increasing/decreasing trends are more frequently mentioned. The main purpose is to use past values of larger numbers to consider future possible values.

IV. Results and Discussion

Four experiments managed for this paper and we done all observations in both cases is considered that contains all 8 and containing other attribute sand 4 one of the selected attributes. With four experiments and eight different scenes a total of eight models of development work.

A. Experiment

1) Performance Measure for J48 Experiment

The first purpose of the experiment was to evaluate a J48 performance class unpruned tree to predict heart disease and investigate the properties of selected effect. In this Table I, first of all we select 8 attributes after completion of the all attribute experiment and then start the selected 4 attributes.

Algorithm In a first aspect of containing 209 instance of the training set has a complete run 8 attributes spent 0.45 seconds to build the model and model size of tree generated by the tree 50 times 30 leaves

TABLE. I. Confusion Matrix

Model	Confusion-Matrixes		
J48:unpruned with attributes	Positive Predicted	Negative Predicted	Actual results
	66	26	Positive
	24	93	Negative
J48:unpruned with Selected attributes	Yes Predicted	No Predicted	Actual results
	77	15	Positive
	28	89	Negative

As shown in Table II, the model correctly identified 66 patients who were enrolled in 92 patients with heart disease and the remaining 26 were identified by errors that were disease free and in fact these had a disease. This result gives the model of 0.756 Precision rate. The better model is to determine the negative cases as a model of TN rate is 0.78 correctly identify 92 patients were 117 patient who had no heart disease and the remaining 25 were identified have the disease but he had not actually.

TABLE. II. Detail Performance of J48 Experiments

Models	**Accuracy**	**TP-Rate**	**FP-Rate**	**Precisions**	**F-Measure**	**ROC-Area**
J48 Unpruned with all attributes	76%	0.756	0.252	0.756	0.756	0.828
J48 Unpruned with selected attribute	79%	0.794	0.194	0.804	0.795	0.771

For precision score model labeled as belonging to class positive patients 79% (a) determining a real belong to the class affirmative (YES) and marked as belonging to the class-negative patients 76% (no) is not really a real negative part of the class (no). With 80.4% of the average precision it is in a very successful pattern for each class to retrieve the relevant values. With the 0.795 F-measured values it can be concluded that the accuracy and model recall rates are significantly balanced.

The results of this experiment show that a J48 decision unpruned tree algorithm is highly capable of when a prediction of heart disease. In addition the results show the impact of attributes select the classification accuracy, the size of the decision tree and the complexity of the model.

2) Naïve Bayes Classifiers

In this Table III, we predict the heart disease through Naïve Bayesian classifier and assess the performance of the experiment. In the third experiment, two scenarios are considered first we take all attributes 8 and the other we take selected 4 attributes.

In the first embodiment of the algorithm for solving the 209 instance of the complete set of training run 8 points and the attributes of the model execution time of 0.04 seconds. In a second embodiment the algorithm contained 209 one instance selected 4 attributes and a complete run on a training set of the model execution time of 0.00 seconds.

TABLE. III. Confusion Matrix of Naïve Bayes

Models	**Confusion_Matrixes**		
Naïve-Bayes with attributes	Positive Predicted	Negative Predicted	Actual results
	67	25	Positive
	18	99	Negatives
Naïve Bayes with Selected attributes	Yes Predicted	No Predicted	Actual results
	75	17	Positive
	29	88	Negatives

As shown in Table IV, the overall classification accuracy of the model than all similar experiments performed better properties but it is still more than the success of the more. The model correctly identified 163 (77%) patients for the 209

embodiment who heart disease and the remaining 46 (22%) is determined to be error from the disease-free charges but they actually had the disease. This result gives the model TP rate of 0.78. This model is better in the case of determining the negative because the model of TN rate is 0.74 pass through correctly identified 76 patients performed 92 Li who had no heart disease and the remaining 16 were identified have the disease but he had not actually.

TABLE. IV. DETAIL PERFORMANCE OF NAÏVE BAYES

Models	Accuracy	TP-Rate	FP-Rate	Precisions	F-Measures	ROC-Area
Naïve-Bayes with attribute	80%	0.798	0.217	0.798	0.799	0.872
Naïve-Bayes with selected attribute	77%	0.780	0.210	0.788	0.781	0.827

For precision score model, labeled as belonging to class positive patients 78% (a) determining a real belong to the class affirmative (YES) and marked as belonging to the class-negative patients 74%(no) is not really a real negative part of the class (no). With 74% of the average precision it is in a very successful pattern for each class to retrieve the relevant values. 0.78 F-measured values can be concluded that the accuracy and model recall rates are significantly balanced. In here, the better naive Bayes model selected property.

3) Neural Network

This experiment was designed to explore the ability of the neural network to predict the disease. Neural carried out by a multi-layer perception network algorithm is selected experiments.

In Table V, a first embodiment of the algorithm 209 run instances of complete training set 8 points and the attributes of the algorithm taken 0.56 seconds to build the model and super over 3 of 5 bell to produce confusion matrix. In the second embodiment of the algorithm for solving the 209 instance selected 4 complete attributes operation training set and the 0.17 seconds to build the models and super over 2 of 5 minute to produces confusion matrix.

TABLE. V. CONFUSION MATRIX OF NETWORK EXPERIMENT

Models	Confusion_Matrixes		
Neural-network with attribute	Positive predicted	Negative predicted	Actual results
	66	26	Positives
	22	95	Negatives
Neural-network with Selected attribute	Yes predicted	No predicted	Actual results
	77	15	Positives
	29	88	Negatives

In Table VI, all 8 of the first attributes of neural network model correctly classified 160 (76.55%) of the instance while Example 49(23.45%) class. The overall accuracy of the velocity model is successful models discussed so far. The model correctly identified 64-patients performed 92 who heart disease and the remaining 28 were identified errors are free from the disease and they actually had the disease. This result gives the model of 0.766 Purpose price rate. The model was determined in the negative case the better TN rate model was 0.820 correctly identified 96 patients who were enrolled in 117 patients who had no heart disease and the remaining 21 had identified the disease while they did not actually.

Models	Accuracy	TP-Rate	FP-Rate	Precisions	F_Measure	ROC-Area
Neural-network with attribute	75%	0.767	0.248	0.766	0.765	0.845
Neural-network with selected attribute	78%	0.789	0.200	0.798	0.790	0.797

For precision score model labeled as belonging to class positive patients 78% (a) determining a real belong to the class affirmative (YES) and marked as belonging to the class-negative patients 75%(no) is not really a real negative part of the class (no). With 79.8% of the average precision it is in a very successful pattern for each class to retrieve the relevant values. With the 0.790 F-measured values it can be concluded that the accuracy and the recall rate of the model are significantly balanced. The result shows that the neural networks model of the selected properties better than the whole property. The classification accuracy rate increased from 75% to 79.8%. Moreover the execution time decreased significantly from 0.56 to 0.16 seconds.

TABLE. VI. DETAIL PERFORMANCE OF ALL IMPLEMENTED ALGORITHMS

Algorithms	Accuracy (%)	TP-rate	FP-rate	Precisio	F_measu	ROC-Cur	Time: (sec)
J48-Decision Tree-pruned with all	77.04%	0.771	0.240	0.771	0.771	0.818	0.05
J48-Decision Tree-pruned with selected attribute	79.5%	0.795	0.193	0.805	0.794	0.772	0.04
J48-Decision Tree-un-pruned with all attribute	75.61%	0.757	0.251	0.757	0.757	0.827	0.46
J48-Decision Tree un-pruned with selected	79.43%	0.795	0.193	0.805	0.794	0.772	0.01
Multilayer-perceptron with all	76.56%	0.767	0.248	0.766	0.766	0.844	0.57
Multilayer-perceptron with selected	78.96%	0.788	0.201	0.799	0.791	0.798	0.17
Naïve-bayes with all attribute	79.91%	0.798	0.217	0.798	0.797	0.872	0.03
Naïve-bayes with selected attribute	77.98%	0.781	0.211	0.787	0.780	0.828	0.01

In Table VII, all sub-class algorithms have almost as high as 80% of the remarkable accuracy and precision of a minimum score of 76%. Naïve Bayes classifier to achieve the highest accuracy in the all property (80%) while a Naïve Bayes classifier to achieve a selected attribute it is a 78% of the sub-class accuracy followed On the other hand simple two implement a decision tree classifier score and the entire group selected attribute properties lowest sub class accuracy which were 75% and 77%.

V. CONCLUSION

Known information mining and Knowledge disclosure (KDD) expressions is utilized to extract the learning (mode) from an extensive number of information acquired is helpful for a given application or information data. From the generated knowledge of the user can determine and meet our requirements. For detecting the heart disease classification and prediction techniques developed in this study. The main aim of this paper is to diagnose heart disease and prevent attacks on people. To this end we use three different monitoring machine learning algorithm to build the model to facilitate the people. Different oversight algorithm is a decision tree classification algorithm using a Bayesian classifier and neural networks 3.8.1 of Weka machine learning software. For predicting heart disease, we have collect the heart patient data from Faisalabad Institute of cardiology contain 209patients, from 2012 to 2015. We use three constracting supervised machine learning algorithms, for example naïve Bayes is plain on j48 and Multilayer Perceptron Weka 3.8.1 machine learning to run the learning software. We established model tests or diagnosed heart disease by pretreatment of chest echocardiographic data sets. All sub-class algorithms have almost as high as 80% of the remarkable accuracy and precision of a minimum score of 76%. Naïve Bayes classifier to achieve the highest accuracy in the all property (80%) while a naïve Bayes classifier to achieve a selected attribute it is a 78% of the sub-class accuracy followed on the other hand simple two implement a decision tree classifier score and the entire group selected attribute properties lowest sub class accuracy which were 75% and 77%.

VI. FUTURE WORK

With respect to future work the specialists intend to lead progressively extra exploratory informational data and algorithms to enhance sub class precision and have the capacity to assemble the model sort of particular expectation of heart illness. To improve the model further research should be carried out using a classification accuracy of different sub class algorithms such as Support Vector Machines (SVM) and rule induction. Most of the experiments carried out this study the default parameters used to implement the algorithm further studies should use a different set of parameters to carry out, in order to increase strength and ability to predict model to expand.

REFERENCES

[1] Manimekalai, K. (2016). Prediction of Heart Diseases using Data Mining Techniques.International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol, 4.

[2] Soni, J., Ansari, U., Sharma, D., &Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[3] Shafique, U., Majeed, F., Qaiser, H., & Mustafa, I. U. (2015). Data mining in healthcare for heart diseases. International Journal of Innovation and Applied Studies, 10(4), 1312.

[4] Chandna, D. (2014). Diagnosis of heart disease using data mining algorithm. (IJCSIT) International Journal of Computer Science and Information Technologies, 5(2), 1678-1680.

[5] Lakshmi, K., Krishna, M. V., & Kumar, S. P. (2013). Performance comparison of data mining techniques for predicting of heart disease survivability. International Journal of Scientific and Research Publications, 3(6), 1-10.

[6] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., . . . Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 111(1), 52-61.

[7] Gamberger, D., Lavrač, N., &Krstačić, G. (2003). Active subgroup mining: a case study in coronary heart disease risk group detection. Artificial Intelligence in Medicine, 28(1), 27-57.

[8] Chaurasia, V., & Pal, S. (2013). Early prediction of heart diseases using data mining techniques. Caribbean Journal of Science and Technology, 1, 208-217.

[9] Muthukaruppan, S., &Er, M. J. (2012). A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Systems with Applications, 39(14), 11657-11665.

[10] Yeh, D.-Y., Cheng, C.-H., & Chen, Y.-W. (2011). A predictive model for cerebrovascular disease using data mining. Expert Systems with Applications, 38(7), 8970-8977

[11] Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining: MIT press.

[12] Pham, B. T., Bui, D., Prakash, I., &Dholakia, M. (2016). Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. J Geomat, 10, 71-79.

[13] Khemphila, A., &Boonjing, V. (2010). Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. Paper presented at the Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on.

[14] Jaiwei, H., &Kamber, M. (2006). Data mining: concepts and techniques. ed: Morgan Kaufmann San Francisco.

[15] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2005). Weka. Data mining and knowledge discovery handbook, 1305-1314.

[16] Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining: MIT press.

[17] Sen, A. K., Patel, S. B., & Shukla, 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, 2(9).