Prediction of Heart Disease using Classification Algorithms

Hlaudi Daniel Masethe, Mosima Anna Masethe

Abstract— The heart disease accounts to be the leading cause of death worldwide. It is difficult for medical practitioners to predict the heart attack as it is a complex task that requires experience and knowledge. The health sector today contains hidden information that can be important in making decisions. Data mining algorithms such as J48, Naïve Bayes, REPTREE, CART, and Bayes Net are applied in this research for predicting heart attacks. The research result shows prediction accuracy of 99%. Data mining enable the health sector to predict patterns in the dataset.

Index Terms—Algorithm, Classification, Diseases, Heart-Attack

I. Introduction

HEART attack diseases remains the main cause of death worldwide, including South Africa and possible detection at an earlier stage will prevent the attacks [1]. Medical practitioners generate data with a wealth of hidden information present, and it's not properly being used effectively for predictions [1]. For this purpose, the research converts the unused data into a dataset for modeling using different data mining techniques. People die having experienced symptoms that were not taken into considerations. There is a need for medical practitioners to predict heart disease before they occur in their patients [2]. The features that increase the possibility of heart attacks are smoking, lack of physical exercises, high blood pressure, high cholesterol, unhealthy diet, harmful use of alcohol, and high sugar levels [3][4]. Cardio Vascular Disease (CVD) incorporates coronary heart, cerebrovascular (Stroke), hypertensive heart, congenital heart, peripheral artery, rheumatic heart, inflammatory heart disease [3].

Data mining is a knowledge discovery technique to analyze data and encapsulate it into useful information [1]. The current research intends to predict the probability of getting heart disease given patient data set [5]. Predictions and descriptions are principal goals of data mining, in practice [6]. Prediction in data mining involves attributes or variables in the data set to find an unknown or future state values of other attributes [7]. Description emphasize on discovering patterns that explains the data to be interpreted by humans [6].

The purpose of predictions in data mining is to help discover trends in patient data in order to improve their health [1]. Due to change in life styles in developing countries, like South Africa, Cardio Vascular Disease (CVD) has become a leading cause of deaths [5]. CVD is projected to be a single largest killer worldwide accounting for all deaths [3]. An endeavor to exploit knowledge, experience and clinical screening of patients to diagnose or recognize heart attacks is regarded as a treasured opportunity [2]. In the health sectors data mining play an important role to predict diseases [7]. The predictive end of the research is a data mining model.

.

II. Related Work

The researchers [8] used pattern recognition and data mining methods in predicting models in the domain of cardiovascular diagnoses. The experiments were carried out using classification algorithms Naïve Bayes, Decision Tree, K-NN and Neural Network and results proves that Naïve Bayes technique outperformed other used techniques [8]. The researchers [9] uses K-means clustering algorithm on a heart disease warehouse to extract data relevant to heart disease, and applies MAFIA (Maximal Frequent Item set Algorithm) algorithm to calculate weightage of the frequent patterns significant to heart attack predictions.

The researchers [1] proposed a layered neuro-fuzzy approach to predict occurrences of coronary heart disease simulated in MATLAB tool. The implementation of the neuro-fuzzy integrated approach produced an error rate very low and a high work efficiency in performing analysis for coronary heart disease occurrences [1]. The researchers [5] also proposed a new approach for association rule mining based on sequence number and clustering transactional data set for heart disease predictions. The implementation of the proposed approach was implemented in C programming language and reduced main memory requirement by considering a small cluster at a time in order to be considered scalable and efficient [5].

The researchers [10] used the data mining algorithms decision trees, naïve bayes, neural networks, association classification and genetic algorithm for predicting and analyzing heart disease from the dataset. An experiment performed by [11] the researchers produced a model using neural networks and hybrid intelligent

Manuscript received July 17, 2014; revised August 15, 2014.
H.D. Masethe is with the Department of Software Engineering at Tshwane University of Technology, Pretoria 0001, South Africa (phone: +27 12-382-9714; fax: +27 866-214-011; (e-mail: masethehd@tut.ac.za).
M.A. Masethe is with the Department of Software Engineering at the Tshwane University of Technology eMalahleni , eMalahleni, 1035, South Africa (phone: +2784-888-6624; (e-mail: masethema@tut.ac.za).

algorithm, and the results shows that the hybrid intelligent technique improved accuracy of the prediction.

The research paper [12] describes the prototype using naïve bayes and weighted associative classifier (WAC) to predict the probability of patients receiving heart attacks. The researchers [13] developed a web based intelligent system using naïve bayes algorithm to answer complex queries for diagnosing heart disease and help medical practitioners with clinical decisions.

The researcher [14] uses association rules representing a technique in data mining to improve disease prediction with great potentials. An algorithm with search constraints was also introduced to reduce the number of association rules and validated using train and test approach [14]. Three popular data mining algorithms (support vector machine, artificial neural network and decision tree) were employed by the researchers [15] to develop a prediction model using 502 cases. SVM became the best prediction model followed by artificial neural networks [15].

The researchers [16] uses decision trees, naïve bayes, and neural network to predict heart disease with 15 popular attributes as risk factors listed in the medical literature.

Two kinds of data mining algorithms named evolutionary termed GA-KM and MPSO-KM cluster the cardiac disease data set and predict model accuracy [17]. This is a hybrid method that combines momentum-type particle swarm optimization (MPSO) and K- means technique. The comparison was made in the research conducted using C5, Naïve Bayes, K-means, Ga-KM and MPSO-KM for evaluating the accuracy of the techniques. The experimental results showed that accuracy improved when using GA-KM and MPSO-KM [17].

The researchers [18] created class association rules using feature subset selection to predict a model for heart disease. Association rule determines relations amongst attributes values and classification predicts the class in the patient dataset [18]. Feature selection measures such as genetic search determines attributes which contribute towards the prediction of heart diseases. The researchers [19] implemented a hybrid system that uses global optimization benefit of genetic algorithm for initialization of neural network weights. The prediction of the heart disease is based on risk factors such as age, family history, diabetes, hypertension, high cholesterol, smoking, alcohol intake and obesity [19].

III. Research methodology

The goal of the prediction methodology is to design a model that can infer characteristic of predicted class from combination of other data [20]. The task of data mining in this research is to build models for prediction of the class based on selected attributes. The research applies the following algorithms: J48, Bayes Net, and Naive Bayes, Simple Cart, and REPTREE algorithm to classify and develop a model to diagnose heart attacks in the patient data set from medical practitioners.

The objective of the research is to predict possible heart attacks from the patient dataset using data mining techniques and determines which model gives the highest percentage of correct predictions for the diagnoses.

IV. Patient Dataset

The patient data set is compiled from data collected from medical practitioners in South Africa. Only 11 attributes from the database are considered for the predictions required for the heart disease. The following attributes with nominal values are considered: Patient Identification Number (replaced with dummy values), Gender, Cardiogram, Age, Chest Pain, Blood Pressure Level, Heart Rate, Cholesterol, Smoking, Alcohol consumption and Blood Sugar Level.

Waikato Environment for Knowledge Analysis (WEKA) has been used for prediction due to its proficiency in discovering, analysis and predicting patterns [20].

V. RESEARCH RESULTS

The algorithms are applied on the data set using stratified 10-fold validation in order to assess the performance of classification techniques for predicting a class.

Confusion Matrix of J48 Algorithm

```
=== Confusion Matrix ===
  a  b   <-- classified as
 89  1 |  a = TRUE
  0 18 |  b = FALSE
```

Confusion Matrix of REPTREE Algorithm

```
=== Confusion Matrix ===
  a  b   <-- classified as
 89  1 |  a = TRUE
  0 18 |  b = FALSE
```

Confusion Matrix of NAÏVE BAYES Algorithm

```
=== Confusion Matrix ===

  a  b   <-- classified as
 88  2 |  a = TRUE
  1 17 |  b = FALSE
```

Confusion Matrix of BAYES NET Algorithm

```
=== Confusion Matrix ===

  a  b   <-- classified as
 88  2 |  a = TRUE
  0 18 |  b = FALSE
```

Confusion Matrix of SIMPLE CART Algorithm

```
=== Confusion Matrix ===

  a  b   <-- classified as
 89  1 |  a = TRUE
  0 18 |  b = FALSE
```

Table I Description of the Data Set

Attributes	Description	Possible Values
PatientId	Dummy Identification of the patient	
Gender		Male, Female

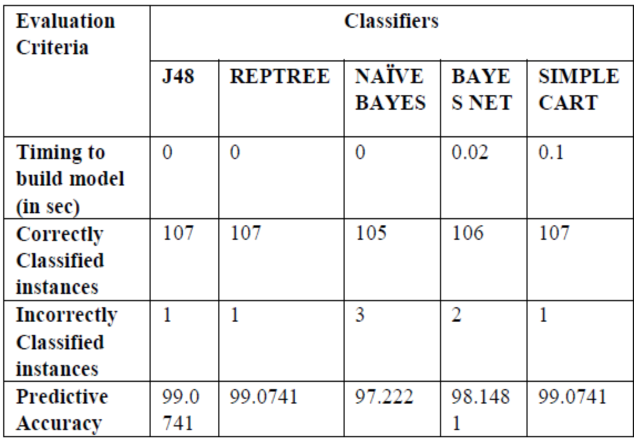
Age	Youth = 30-39,
	Young Adult =40-49
	Adult =50-59
	Old People =60-69
		Youth
		Young Adult
		Adult
		Old

ChestPainType	Stable Angina – Predictable Chest Pain Unstable Angina –Chest pain that signal impending heart attack Prinzmetal's Angina –have coronary artery disease	Stable angina Non-angina Unstable angina Prinzmetal’s angina Asymptomatic
HeartRate		
Cholesterol	Low-density lipoproteins (LDL) (Bad Cholesterol), High-density lipoproteins (HDL) (Good Cholesterol)	LDL HDL
Smoking		Yes, No
BloodSugar	If Blood Sugar level is > 120 mg/dl -Increase the risk	True, False
BloodPressure	Normal-(systolic<139mm Hg), Prehypertension-(systolic >140 mmHg), High – (systolic > 160 mmHg)	Normal Prehypertension High
Electrocardio graphicR (ECG)	Normal - ST_T wave Abnormality, Left Ventricular Hypertrophy (LVH) {Electrocardiographic results }	Normal Abnormal
Diet		Healthy, Unhealthy
Alcohol		True, False

The confusion matrix obtained calculate the accuracy, sensitivity and specificity measures [15]. The matrix denotes samples classified as true, others as false and others misclassified. Evaluation of the confusion matrix shows that J48, REPTREE and SIMPLE CART show a prediction model of 89 cases with a risk factor positive for heart attacks. The techniques strongly suggest that data mining algorithms are able to predict a class for diagnoses. The confusion matrix clearly categorizes the accuracy of the mode. The matrix validates the effectiveness of the model.

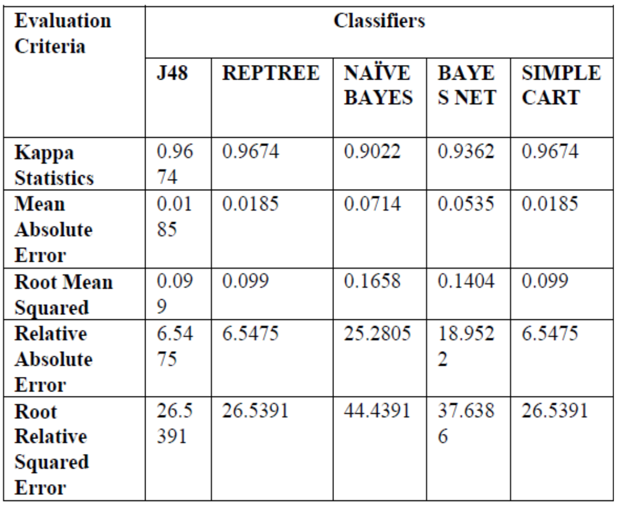
Table II and Table III shows classification accuracy based on different techniques applied, which proves the best classification technique to be J48, REPTREE and SIMPLE CART algorithm perform similar in this data set, while Bayes Net algorithm out-performed the Naïve Bayes algorithm.

Table II Predictive performance of the classifiers



Evaluation Criteria	Classifiers				
	J48	REPTREE	NAÏVE BAYES	BAYES NET	SIMPLE CART
Timing to build model (in sec)	0	0	0	0.02	0.1
Correctly Classified instances	107	107	105	106	107
Incorrectly Classified instances	1	1	3	2	1
Predictive Accuracy	99.0741	99.0741	97.222	98.1481	99.0741

Table III Comparison of estimates



Evaluation Criteria	Classifiers				
	J48	REPTREE	NAÏVE BAYES	BAYES NET	SIMPLE CART
Kappa Statistics	0.9674	0.9674	0.9022	0.9362	0.9674
Mean Absolute Error	0.0185	0.0185	0.0714	0.0535	0.0185
Root Mean Squared	0.099	0.099	0.1658	0.1404	0.099
Relative Absolute Error	6.5475	6.5475	25.2805	18.9522	6.5475
Root Relative Squared Error	26.5391	26.5391	44.4391	37.6386	26.5391

VI. Decision tree model

The J48 algorithm grows an initial tree using the divide and conquers technique. Fig 1 shows the visualization of the tree from modeling the dataset using the J48 algorithm. The tree is pruned to evade over fitting. The tree-construction in J48 differs with the tree-construction in several respects from REPTREE in Fig 2. These two trees show a graphical representation of the relations that exist in the dataset. Knowledge is represented mainly from the classification and prediction model in a tree structure.

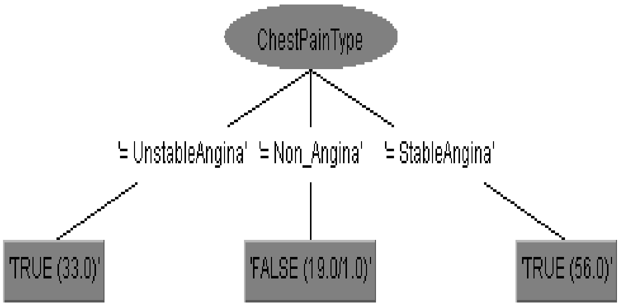


Fig. 1. J48 Pruned Tree

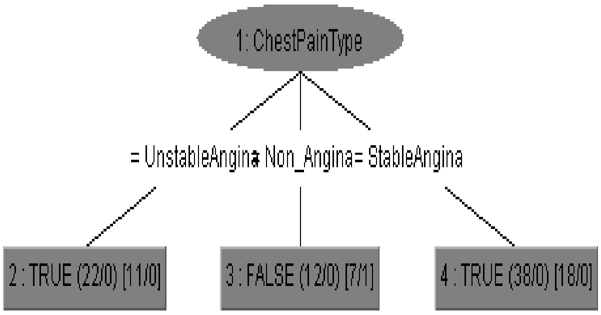


Fig .2. REPTREE Pruned Tree

VII. Conclusion

The research undertook an experiment on application of various data mining algorithms to predict the heart attacks and to compare the best method of prediction. The research results do not presents a dramatic difference in the prediction when using different classification algorithms in data mining. The experiment can serve as an important tool for physicians to predict risky cases in the practice and advise accordingly. The model from the classification will be able to answer more complex queries in the prediction of heart attack diseases. The predictive accuracy determined by J48, REPTREE and SIMPLE CART algorithms suggests that parameters used are reliable indicators to predict the presence of heart diseases.

.

References

[1] A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," *International Journal of Engineering and Computer Science*, vol. 2, no. 9, pp. 1663–1671, 2013.

[2] S. . Ishtake and S. . Sanap, "‘ Intelligent Heart Disease Prediction System Using Data Mining Techniques ’," *International Journal of healthcare & biomedical Research*, vol. 1, no. 3, pp. 94–101, 2013.

[3] V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208–217, 2013.

[4] D. S. Chaitrali and A. S. Sulabha, "A Data Mining Approach for Prediction of Heart Disease Using Neural Networks," *International Journal of Computer Engineering & Technology (IJCET)*, vol. 3, no. 3, pp. 30–40, 2012.

[5] M. Jabbar, P. Chandra, and B. Deekshatulu, "CLUSTER BASED ASSOCIATION RULE MINING FOR," *Journal of Theoretical & Applied Information Technology*, vol. 32, no. 2, pp. 196–201, 2011.

[6] R. Rao, "SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 1, no. 3, pp. 14–34, 2011.

[7] S. Vijiyarani and S. Sudha, "Disease Prediction in Data Mining Technique – A Survey," *International Journal of Computer Applications & Information Technology*, vol. II, no. I, pp. 17–21, 2013.

[8] T. J. Peter and K. Somasundaram, "AN EMPIRICAL STUDY ON PREDICTION OF HEART DISEASE USING CLASSIFICATION DATA MINING TECHNIQUES," 2012.

[9] [9] S. B. Patil and Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 9, no. 2, pp. 228–235, 2009.

[10] K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.

[11] [11] R. Chitra and V. Seenivasagam, "REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES," *Journal on Soft Computing (ICTACT)*, vol. 3, no. 4, pp. 605–609, 2013.

[12] N. A. Sundar, P. P. Latha, and M. R. Chandra, "PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE," *International Journal of Engineering Science & Advanced Technology*, vol. 2, no. 3, pp. 470–478, 2012.

[13] S. A. Pattekari and A. Parveen, "PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES," *International journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290–294, 2012.

[14] C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction.," *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 10, no. 2, pp. 334–43, Apr. 2006.

[15] Y. Xing, J. Wang, Z. Zhao, and A. Gao, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease," in *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, 2007, pp. 868–872.

[16] K. Srinivas, K. Raghavendra Kao, and A. Govardham, Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in *The 5th International Conference on Computer Science & Education*, 2010, pp. 1344–1349.

[17] J. Liu, Y.-T. HSU, and C.-L. Hung, "Development of Evolutionary Data Mining Algorithms and their Applications to Cardiac Disease Diagnosis," in *WCCI 2012 IEEE World Congress on Computational Intelligence*, 2012, pp. 10–15.

[18] P. Chandra, M. . Jabbar, and B. . Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in *12th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2012, pp. 628–634.

[19] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in *Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013)*, 2013, no. Ict, pp. 1227–1231.

[20] A. AZIZ, N. ISMAIL, and F. AHMAD, "MINING STUDENTS’ACADEMIC PERFORMANCE.," *Journal of Theoretical & Applied Information Technology*, vol. 53, no. 3, 2013.