INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

# AN ANALYSIS OF CARDIOVASCULAR DISEASE PREDICTION SYSTEMS USING DIFFERENT DATA MINING TECHNIQUES

**A.Shobana[1], P.Elango[2]**

[1]*M.Phil Research Scholar, PG & Research Dept. of Computer Science*
*Gobi Arts & Science College (Autonomous), Gobichettipalayam - 638 453*
shobanasa@gmail.com

[2]*Associate Professor of Comp. Science, PG & Research Dept. of Computer Science*
*Gobi Arts & Science College (Autonomous), Gobichettipalayam - 638 453*
pitchaielango@gmail.com

## Abstract:

Heart disease could be a term that assigns to an oversized variety of medical conditions associated with heart. These medical conditions describe the abnormal health conditions that directly influence the center and everyone its components. Cardiopathy could be a major pathological state in today's time. This paper aims at analyzing the varied data processing techniques introduced in recent years for cardiopathy prediction. The observations reveal that Neural networks with fifteen attributes have outperformed over all different data processing techniques. Another conclusion from the analysis is that call tree has conjointly shown smart accuracy with the assistance of genetic algorithmic program and has set choice.

**Keywords:** Heart disease, Fuzzy logic, Decision tree, Naive Bayes, Genetic algorithm.

## 1. Introduction

Data mining is that the method of finding antecedently unknown patterns and trends in databases and victimization that info to create prognosticative models. In tending, data processing is turning into more and more fashionable, if not more and more essential. tending trade nowadays generates great deal of complicated information concerning patients, hospitals resources, malady designation, electronic patient records, medical devices, etc. the massive quantity information could be a key resource to be processed and analyzed for knowledge extraction that permits support for cost-savings and higher cognitive process. Data processing provides a collection of tools and techniques that may be applied to the present processed information to find hidden patterns and additionally provides tending professionals a further supply of data for creating choices. The fundamental process model is shown in Figure 1.

The World Health Statistics 2012 report enlightens the actual fact that one in 3 adults worldwide have raised pressure – a condition that causes around half all deaths from stroke and cardiovascular disease. Cardiovascular disease, additionally called upset (CVD), encloses variety of conditions that influence the center –

not simply heart attacks. Cardiovascular disease additionally includes practical issues of the center like heart-valve abnormalities or irregular heart rhythms. These issues will result in coronary failure, arrhythmias and a bunch of alternative issues.

Effective and economical machine-controlled cardiopathy prediction systems will be useful in care sector for cardiopathy prediction. Our work tries to gift the careful study regarding the various data processing techniques which may be deployed in these machine-controlled systems. This automation will scale back the amount of tests to be taken by a patient. Hence, it will not solely save price however additionally the time of each, analysts and patients.

The rest of this paper is organized as follows. Section 2 describes the methodology of this research. In Section 3 explains the research observation on describes the detail of proposed system. In Section 4 Report the results of this research. In Section 5 describes the conclusion of this paper.

## 2. Methodology

This paper exhibits the analysis of varied data processing techniques which may be useful for medical analysts or practitioners for correct cardiopathy designation. the most methodology used for our work was by examining the publications, journals and reviews within the field of engineering science and engineering, data processing and upset in recent times [5].

## 3. Research Observations

### *3.1 Data Mining and Neural Networks*

An artificial neural network (ANN), typically simply referred to as a "neural network" (NN), could be a mathematical model or procedure model supported biological neural network. In alternative words, it's associate emulation of biological neural system. During this work, cardiovascular disease prediction system has been developed victimization 15 attributes [4]. Earlier 13 attributes were used for prediction however this analysis work incorporated a pair of additional attributes, i.e. blubber and smoking for economical designation of cardiovascular disease.

The data mining tool WEKA 3.6.6 is employed for experiment. Initially, missing values were known within the dataset and that they were replaced with applicable values victimization Replace Missing Values filter from 3.6.6 [4]. Further, varied data processing techniques are analyzed on cardiovascular disease information. Confusion matrix is obtained for every classifier.

Table 1 depicts the outcomes of this research work and it shows that neural networks has outplayed over other data mining techniques.

**Table 1:** Comparison of various data mining techniques

| Classification Techniques | Accuracy |
|---|---|
| Naive Bayes | 90.74% |
| Decision Trees | 99.62% |
| Neural Networks | 100% |

### *3.2 Fuzzy Logic and Genetic Algorithm*

The projected technique during this analysis work is associate extended version of the model that mixes the genetic algorithms for feature choice and fuzzy knowledgeable system for effective classification. Fuzzy pure mathematics and mathematical logic are extremely appropriate for developing data based mostly systems in health care for

designation of diseases [2]. Experiments are conducted in Matlab victimization fuzzy tool. For this, Mamdani model of fuzzy system is employed. The fuzzy rules are generated supported experts' data during this domain. The dataset from UCI machine learning repository is employed, and solely half-dozen attributes are found to be effective and necessary for cardiovascular disease prediction. Within the projected system, the input is that the set of all the chosen options and also the output of the system are to realize a price zero or one that indicates the absence or presence of cardiovascular disease in patients.

In mathematical logic method, at the start fuzzification is performed by collection the crisp set of computer file and changing it to a fuzzy set victimization fuzzy linguistic variables, fuzzy linguistic terms and membership functions. After that, associate logical thinking is created supported a collection of rules and last, defuzzification step is performed [2]. This technique generates the fuzzy rules supported the support sets obtained. Table a pair of shows this support set.

**Table 2:** Values of the features in the support set

| S.No. | Attributes | Support Set | |
|---|---|---|---|
| | | Heart Patients | Non – Heart Patients |
| 1. | Chest Pain Type | 4 | 1,2, 3 |
| 2. | Rbps | 134-153 | 142-154 |
| 3. | Exang | Yes | No |
| 4. | Oldpeak | 2.06-6.2 | <2.06 |
| 5. | Thalach | 71-136 | 136-168 |
| 6. | Ca | 1,2,3 | 0 |

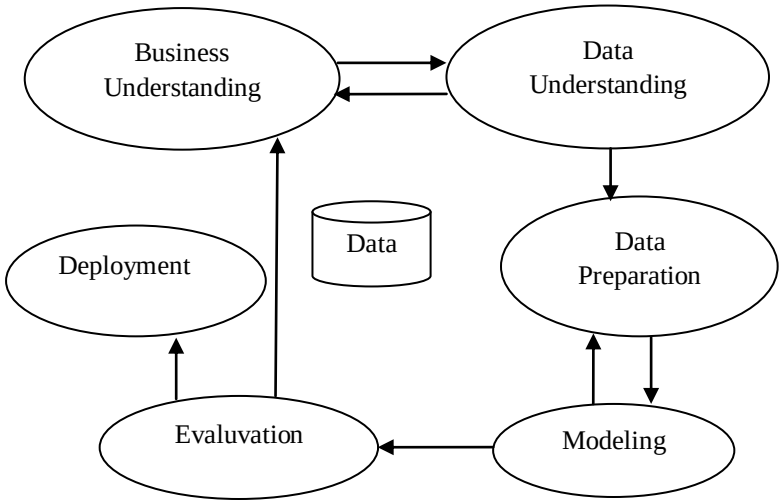


**Figure 1:** Data Mining Process Model

### *3.3 Data Mining and Supervised Machine Learning Algorithms*

This analysis work has conferred the information classification supported numerous supervised machine learning algorithms, namely, Naive Bayes, call List and KNN. TANAGRA tool is employed to classify information and also the data is evaluated victimization 10- fold cross validation. TANAGRA [20] could be a data processing tool for tutorial and analysis functions. It proposes many data processing ways from exploratory information analysis, applied mathematics learning, machine learning and databases space. It provides associate degree easy-to-use interface by permitting the users to research either real or artificial information. This tool conjointly planned design to the users permitting them to simply add their own data processing ways, to match their performances. It's a large set of knowledge sources, direct access to information warehouses and databases, information cleansing, interactive utilization. Experiments are conducted by victimization the coaching information set of 3000 instances with fourteen completely different attributes.

Depending upon the attributes, the dataset is classed into 2 components, i.e. 70% of the information is employed for coaching and rest half-hour is employed for testing. Performance of every algorithmic program is set and comparison is created supported the accuracy and analysis time of calculation for every algorithmic program [12]. It's been determined that Naive Bayes algorithmic program performed higher compared to different 2 algorithms. Table three illustrates the performance study of varied algorithms.

**Table 3:** Performance analysis of various Algorithms

| Algorithm Used | Accuracy | Time Taken |
|---|---|---|
| Naive Bayes | 52.33% | 609ms |
| Decision List | 52% | 719ms |
| KNN | 45.67% | 1000ms |

### *3.4 Data Mining and Genetic Algorithm*

The objective of this work was to cut back the quantity of attributes that were used for cardiovascular disease diagnosing. Earlier, 13 attributes were used for this prediction however this analysis work reduced the quantity of attributes to 6 solely victimization Genetic rule and have set choice.

Genetic rule [6] incorporates natural evolution methodology. The genetic search started with zero attributes, and an initial population with randomly generated rules. Supported the thought of survival of the fittest, new population was created to match with fittest rules within the current population, yet as offspring of those rules. Offspring were generated by applying genetic operators; cross over and mutation. The method of generation continued till it evolved a population P wherever each ruling P satisfies the fitness threshold. With initial population of twenty instances, generation continued until the 20th generation with cross over likelihood of 0.6 and mutation likelihood of 0.033. The genetic search resulted in 6 attributes out of 13 attributes.

CFS judge is additionally employed in addition to the genetic rule. Observations area unit conducted victimization weka 3.6.0 tool. Initially, information set of 909 records with thirteen attributes was used. All attributes were created categorical and inconsistencies were resolved for simplicity. When reduction of thirteen attributes to 6 attributes, numerous classifiers area unit used on the dataset similar to these half dozen attributes for cardiovascular disease prediction. Performance analysis of those classifiers is shown in Table 4. It is perceived from the table that call Tree has outperformed with highest accuracy and least mean absolute error.

**Table 4:** Comparison Table for three Classifiers

| DM Techniques | Accuracy | Model Construction Time | Mean Absolute Error |
|---|---|---|---|
| Naive Bayes | 96.5% | 0.02s | 0.044 |
| Decision Tree | 99.2% | 0.09s | 0.00016 |
| Classification via Clustering | 88.3% | 0.06s | 0.117 |

### *3.5 IHDPS and Data Mining Techniques*

This analysis has developed a paradigm Intelligent Heart Disease Prediction System (IHDPS) mistreatment data processing technique; namely, call Trees, Naive Thomas Bayes and Neural Networks. IHDPS is web-based, easy, scalable, reliable and expandable system that is enforced on the .NET platform [15].

IHDPS will discover and extract hidden information related to cardiovascular disease from historical cardiovascular disease info. It will answer complicated queries for designation cardiovascular disease and so facilitate aid analysts and practitioners to form intelligent clinical choices that ancient call support systems cannot. It conjointly helps in reducing treatment prices by providing effective treatments. Moreover, it displays the results each in tabular and graphical forms. This IHDPS relies on fifteen attributes.

A total of 909 records were obtained from the Cleveland cardiovascular disease info. The records were equally divided into 2 datasets, i.e. coaching dataset (455 records) and testing dataset (454 records). it's been discovered throughout the analysis that Naive Thomas Bayes seems to be handiest because it has the very best proportion of correct predictions (86.53%) for patients with cardiovascular disease, followed by Neural Network (85.53%) and call Trees seems to be handiest just in case of predicting patients with no cardiovascular disease, i.e. (89%) as compared to alternative 2 models.

**Table 5:** Performance analysis of IHDPS

| DM Techniques | Accuracy |
|---|---|
| Naïve Bayes | 86.53% |
| Decision Trees | 89% |
| Neural Network | 85.53% |

## 4. Results

For higher understanding, results for every data processing techniques are shown severally in several tables. Varied classifiers square measure used together with totally different data processing techniques for cardiovascular disease prediction. It will be perceived from the observations that in some cases, an equivalent classifier produces totally different accuracy for various data processing techniques.

## 5. Conclusions

The objective of this work is to supply a study of various data processing techniques that may be used in machine-controlled cardiovascular disease prediction systems. Numerous techniques and data processing classifiers area unit outlined during this work that has emerged in recent years for economical and effective cardiovascular disease designation. The analysis shows that Neural Network with 15 attributes has shown the best accuracy i.e. 100% to date. On the opposite hand, call Tree has conjointly performed well with 99.62% accuracy by exploitation 15 attributes. Moreover, together with Genetic rule and half dozen attributes, call Tree has shown 99.2% potency.

## 6. REFERENCES

[1] P .K. Anooj, 2012, Journal of King Saud University, Computer and Information Sciences, Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules, 24, 27-40.
[2] E.P.Ephzibah, Dr. V. Sundarapandian, February 2012, International Journal of Fuzzy Logic Systems (IJFLS), Framing Fuzzy Rules using Support Sets for Effective Heart Disease Diagnosis‖; Vol.2, No.1.
[3] A.Sudha, P.Gayathri, N.Jaisankar, March 2012, International Journal of Computer Applications, Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability‖; (0975 – 8887) Volume 41– No.17.
[4] Chaitrali S. Dangare, Sulabha S. Apte, June 2012, International Journal of Computer Applications, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques‖; (0975 – 888) Volume 47– No.10.
[5] Jyoti Soni, Sunita Soni et al., March 2011, International Journal of Computer Applications, Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction‖; (0975 – 8887) Volume 17– No.8.
[6] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, 2010, International Journal of Engineering Science and Technology, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; Vol. 2(10).
[7] E.Sivasankar, Dr.R.S.Rajesh, 2010, IEEE, Knowledge Discovery in Medical Datasets Using a Fuzzy Logic rule based Classifier‖; 978-1-4244-7406-6/10/$26.00.
[8] M.A. Saleem Durai, et. al. 2102-2108, 2010, International Journal of Engineering Science and Technology, Effective analysis and diagnosis of lung cancer using fuzzy rules‖; Vol. 2(6).
[9] Mostafa Fathi Ganji, Mohammad Saniee Abadeh, 2010, IEEE, Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease‖; Proceedings of ICEE 2010, May 11-13, 2010, 978-1-4244-6760-0/10/$26.00©.
[10] Huang Hai, 2010, International Conference On Computer Design And Applications (ICCDA), Data Mining Based on a Compensative Fuzzy Neural Network‖.
[11] M.A.SaleemDurai,N.Ch.S.N.Iyengar, 2010, International Journal of Engineering Science and Technology, Effective Analysis and Diagnosis of Lung Cancer Using Fuzzy Rules‖; Vol. 2(6).
[12] Asha Rajkumar, G. Sophia Reena, September 2010, Global Journal of Computer Science and Technology, Diagnosis of Heart Disease Using Datamining Algorithm‖; Page | 38 Vol. 10 Issue 10 Ver. 1.0.
[13] Shantakumar B.Patil, Y.S.Kumaraswamy, 2009, European Journal of Scientific Research, ISSN 1450-216X , Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network‖; Vol.31 No.4.
[14] Rupa G. Mehta, Dipti P. Rana, Mukesh A. Zaveri, 2009, World Congress on Computer Science and Information Engineering, A Novel Fuzzy Based Classification for Data Mining using Fuzzy Discretization‖.
[15] Sellappan Palaniappan, Rafiah Awang, 2008, IEEE, Intelligent Heart Disease Prediction System Using Data Mining Techniques‖; 978-1-4244-1968-5/08/$25.00©.
[16] Cleveland database: http://archive.ics.uci.edu/ml/datasets/Heart+Disease
[17] Han, J., Kamber, M, 2006, Data Mining Concepts and Techniques‖; Morgan Kaufmann Publishers.
[18] American Heart Association. Heart Disease and Stroke Statistics — 2004 Update. Dallas, Tex.: American Heart Association; 2003.
[19] Statlog database: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart
[20] http://eric.univ-lyon2.fr/~ricco/tanagra/