“Review on Heart Disease Prediction System using Data Mining Techniques”

Harikrishnan R[1] MCA
Amal jyothi college of Engineering Kottayam, India

Abstract:

“In the context of the health care system, Health problems are enormous in this recent situation because of the prediction and the classification of health problems in different situations. The data mining area included the prediction and identification of abnormality and its risk rate in these domains. Today the health industry holds hidden information essential for decision-making.

This document is to analytic the cause of cardiac disease in peoples by looking through different data sets and by using mining tools we are predicting the disease with more accuracy. The dataset used in this study has 13 features, one target variable, and 303 instances in which 138 suffers from cardiovascular disease and 165 are healthy subjects. We are using this data sets to train a dataset and find out the accurate algorithms and its accuracy and design a suitable model for testing [5]”.

1.Introduction:

The Healthcare industry is generally “information rich”, but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making .Discovery of hidden patterns and relationships often goes unexploited. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting heart disease [6].

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects), among others. The symptoms may be different for men and women. For instance, men are more likely to have chest pain; women are more likely to have other symptoms along with chest discomfort, such as shortness of breath, nausea and extreme fatigue.

The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.

You might not be diagnosed with cardiovascular disease until you have a heart attack, angina, stroke or heart failure. It's important to watch for cardiovascular symptoms and discuss concerns with your doctor [1].

❖ **Key facts**

- Cardio Vascular Disease (CVDs) are the major cause of death globally: more people die annually from CVDs than from any other cause.
- An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke.
- Over three quarters of CVD deaths take place in low- and middle-income countries.
- Out of the 17 million premature deaths (under the age of 70) due to no communicable diseases in 2015, 82% are in

low- and middle-income countries, and 37% are caused by CVDs [2].

2.DATA MINING

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing, and computer processing.

Data mining processes are used to build machine learning models that power applications including search engine technology and website recommendation programs [3].

2.1 Data Mining Tools

- **Rapid Miner** : Rapid Miner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining and predictive analysis. The program is written entirely in Java programming language. The program provides an option to try around with a huge number of arbitrarily nestable operators which are detailed in XML files and are made with graphical user interference of rapid miner.

- **Oracle Data Mining** : The system works with a powerful data algorithm to target best customers. Also, it identifies both anomalies and cross-selling opportunities and enables users to apply a different predictive model based on their need [3].

- **IBM SPSS Modeler**: When it comes to large-scale projects IBM SPSS Modeler turns out to be the best fit. It helps to generate data mining algorithms with minimal or no programming. It used in anomaly detection, Bayesian networks, CARMA, Cox regression and basic neural networks that use multilayer perceptron with back-propagation learning [3].

- **KNIME:** Konstanz Information Miner is an open source data analysis platform. the data-driven innovation system helps uncover data potential. Also, it includes more than thousands of modules and ready-to-use examples and an array of integrated tools and algorithms [3].

- **Weka**: Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software developed at the University of Waikato, New Zealand. The program is written in Java. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling coupled with graphical user interface. Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection [3].

- **Orange**: It is an open source data visualization, machine learning and data mining toolkit. It features a visual programming front-end for exploratory data analysis and interactive data visualization. Orange is a component-based visual programming software package for data visualization, machine learning, data mining and data analysis [3].

- **MATLAB**: MATLAB is a high language and interactive environment for numerical computation, visualization and programming. Using MATLAB we can analyze data, develop algorithms and create models and applications. The language, tool and built-in math functions enable us to explore multiple approaches and reach a solution faster than with spreadsheets of traditional programming languages, such as C/C++ of JAVA [3]. .

- **Kaggle**: It is the world’s largest community of data scientist and machine learners. Kaggle kick-started by offering machine learning competitions but now extended towards public cloud-based data science platform. Kaggle is a platform that helps to solve difficult problems, recruit strong teams and accentuate the power of data science [3].

2.2 WEKA

- WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms.
- WEKA is open source software issued under the GNU General Public License.
- It provides tools for data preprocessing, implementation of several machine learning algorithms, and visualization tools so that we can develop machine learning techniques and apply them to real-world data mining problems.
- It is a collection of machine learning algorithms for data mining tasks.
- The algorithms are applied directly to a dataset.
- WEKA implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tool and is made use of for the prediction.
- The new machine learning schemes can also be developed with this package.
- The raw data collected from the field may contain several null values and irrelevant fields.
- The data preprocessing tools provided in WEKA helps to cleanse the data.
- Save the preprocessed data in your local storage for applying ML algorithms.
- Next, depending on the kind of ML model that are trying to develop we can select one of the options such as Classify, Cluster, Associate etc.
- The Attributes Selection allows the automatic selection of features to create a reduced dataset.
- Under each category, WEKA provides the implementation of several algorithms. We can select an algorithm of our choice, set the desired parameters and run it on the dataset.
- Then, WEKA would give you the statistical output of the model processing.
- It provides you a visualization tool to inspect the data.[4]

3.Data Source

- Here we consider the mushroom dataset from Kaggle repository, which contains 303 instances and 13 attributes.
- The prediction of heart disease is based on several parameters.
- The data set from Kaggle repository contains the attributes of symtops or behaivour of human body with a known target class. This data set is used as training data for the classifier. [4]

3.1 Input Attributes:

1. age: age in years

2. sex: sex (1 = male; 0 = female)

3. cp: chest pain type Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic

4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)

5. chol: serum cholestoral in mg/dl

6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7. restecg: resting electrocardiographic results Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. thalach: maximum heart rate achieved 9. exang: exercise induced angina (1 = yes; 0 = no) 10. oldpeak = ST depression induced by exercise relative to rest

11. slope: the slope of the peak exercise ST segment Value 1: up-sloping Value 2: flat Value 3: down-sloping

12. ca: number of major vessels (0-3) colored by flourosopy

13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

14. num: diagnosis of heart disease (angiographic disease status) Value 0: < 50% diameter narrowing Value 1: > 50% diameter narrowing.[4]

4.Implementation Results

Here we are performing three different algorithm on given data set inorder to create or to made out some usefull model for predicting the scope of the heart disease for given person under certain condition.

4.1 Algorithms

a. **Decision Tree algorithm :**

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

b. **Random Tree**

The random trees classifier, unlike the Support Vector Machine (SVM), can handle a mix of categorical and numerical variable. The Random Trees is also less sensitive to data scaling while SVM often required data to be normalized prior to the training/classification. However, SVM is reported to perform better when the training set is small or unbalanced. The Random Trees classifier is computationally less intensive than SVM and works better and faster with large training sets.

Many versions of the Random Trees algorithm exist. Object Analyst uses the OpenCV implementation which use the Gini Impurity index to determine what is a good split point for a node on the classification tree and the minimum number of samples, the maximum tree depth and the accuracy of the trees as stopping criteria.[7]

c. **Random forest**

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.[8]

Techniques	Accuracy	Time
Decision Tree Algorithm (j48)	93.7294 %	0.01
Random Tree	100 %	0.08
Random Forest	100 %	0.07

After Applying some filtration to the attribute we are able to increase the accuracy of the decision tree about 0.38%.

Techniques	Accuracy	Time
Decision Tree Algorithm (j48)	94.0594 %	0.01
Random Tree	100 %	0.08

Random Forest	100 %	0.07

We see that the highest accuracy for the training set is achieved Random Tree and random foreset which is equal to 100%.

5.Conclusion:

Data mining techniques help in finding the hidden knowledge in a team of disease data that can remain used to analyze and predict the future behavior of diseases. Classification is one the records mining methods which assigned a class label to a set of unclassified cases.

In this paper the focus is on using different algorithms and combinations of several target attributes for effective heart attack prediction using data mining. Random Tree and Decision forest Tree has outperformed with 100% accuracy by using 14 attributes.

Moreover the algorithm Random Forest having accuracy 100% as well as the performance timing of .07 gives a chance over the Random Tree.

The algorithms are implemented with the default parameters only.

6.References:

[1] https://www.mayoclinic.org/diseases-conditions/heart-disease/

[2] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[3]https://www.investopedia.com/terms/d/datamining

[4] Aditya Methaila1 , Prince Kansal2 , Himanshu Arya3 , Pankaj Kumar4 1NetajiSubhas Institute of Technology,India and 2 Student, B.Tech (CSE), Maharaja Surajmal Institute of Technology New Delhi, India :“EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES”

[5] Basma Saleh1, Ahmed Saedi2, Ali al-Aqbi3, Lamees Salman4, ”Analysis of Weka Data Mining Techniques for Heart Disease Prediction System”.

[6] Aditya Methaila1, P. Kansal2, P. Kumar3, Computer Science, ”Early Heart Disease Prediction Using Data Mining Techniques”

[7] https://www.pcigeomatics.com/geomatica-help/concepts/focus_c/oa_classif_intro_rt.html

[8] https://builtin.com/data-science/random-forest-algorithm