Amal Jyothi College of Engineering Kanjirappally, Kerala

"Review on Heart Disease Prediction System using Data Mining Techniques"

INTEGRATED MCA SEMINAR REPORT

Submitted in the partial fulfilment of the requirements for the Award of the Degree in

Integrated Master of Computer Applications

By

HARIKRISHNAN R
Reg No: AJC16MCA-I27

Under The Guidance Of

Ms. JETTY BENJAMIN

January 2021

DEPARTMENT OF COMPUTER APPLICATIONS

AMAL JYOTHI COLLEGE OF ENGINEERING

KANJIRAPPALLY

CERTIFICATE

This is to certify that the seminar report, ***"Review on Heart Disease Prediction System using Data Mining Techniques"*** *is the bonafide work of* ***HARIKRISHNAN R (Reg.No:AJC16MCA-I27)*** *in partial fulfilment of the requirements for the award of the Degree of Integrated Master of Computer Applications under APJ Abdul Kalam Technological University during the year 2020-21.*

Ms. Jetty Benjamin
Internal Guide

Fr. Rubin Thottupuram
Coordinator

Fr. Rubin Thottupuram
Head of the Department

ACKNOWLEDGEMENT

First and foremost, I thank God almighty for his eternal love and protection throughout the seminar. I take this opportunity to express my gratitude to all who helped me in completing this seminar successfully. It has been said that gratitude is the memory of the heart. I wish to express my sincere gratitude to our manager **Rev. Fr. Dr. Mathew Paikatt** and Principal **Dr. Z V Lakaparampil** for providing good faculty for guidance.

I owe a great depth of gratitude towards our Head of the Department **Fr. Rubin Thottupuram** for helping us. I extend my whole hearted thanks to the seminar coordinator **Fr. Rubin Thottupuram** for their valuable suggestions and for overwhelming concern and guidance from the beginning to the end of the seminar. I would also like to express sincere gratitude to my guide, **Ms. Jetty Benjamin** for her inspiration and helping hand.

I thank our beloved teachers for their cooperation and suggestions that helped me throughout the seminar. I express my thanks to all my friends and classmates for their interest, dedication, and encouragement shown towards the seminar. I convey my hearty thanks to my family for the moral support, suggestions, and encouragement to make this venture a success.

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]".

A large volume of data is downloaded from Kaggle repository. The data have been formed into two groups, one for training and another for testing. We use one of the Supervised Learning Algorithms for training. This classification model is saved and then used later to predict whether a given mushroom is edible or not. This is done with the help of weka tool (Waikato Environment for Knowledge Analysis).

CONTENTS

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 number 1 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 nest able 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 Modeller**: When it comes to large-scale projects IBM SPSS Modeller 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 modelling 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 analyse 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 pre-processing, 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 pre-processing, 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 pre-processing tools provided in WEKA helps to cleanse the data.
- Save the pre-processed 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.

2.3 working of weka tool

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, classification, regression, clustering, association rules; it also includes a visualization tool and I is made use of for the prediction. It is also well-suited for developing new machine learning schemes. 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 and 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. The type of algorithms that you apply is based largely on your domain knowledge. Even within the same type, for example classification, there are several algorithms available. You may like to test the different algorithms under the same class to build an efficient machine learning model. While doing so, you would prefer visualization of the processed data and thus you also require visualization tools. Each entry in a dataset is an instance of the java class: − weka.core.Instance. Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data. The various models can be applied on the same dataset. We can then compare the outputs of different models and select the best that meets our purpose. Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

2.4 Weka Data Formats

Weka uses the Attribute Relation File Format for data analysis, by default. But listed below are some formats that Weka supports, from where data can be imported:

- CSV
- ARFF
- Database using ODBC

Each instance consists of a number of attributes:

- Nominal: one of a predefined list of values − e.g. red, green, blue
- Numeric: A real or integer number
- String: Enclosed in “double quotes”
- Date
- Relational

The external representation of an Instances class consists of:

- A header: Describes the attribute types
- Data section: Comma separated list of data

Different WEKA Tool GUI Chooser the WEKA GUI Chooser application allows you to run five different types of applications as listed below:

- **Explorer** for data preparation, feature selection and evaluating algorithms.
- **Experiment Environment** for designing, running and analyzing the results from controlled experiments.

• **Knowledge Flow** Environment for graphically designing and executing machine learning pipelines.

• **Workbench** that incorporates all of the Weka tools into a single convenient interface.

• **Simple CLI** for using the Weka API from the command line.

Explorer

- The Weka Explorer is designed to investigate your machine learning dataset. It is an environment for exploring data. The explorer is where you play around with your data and think about what transforms to apply to your data, what algorithms you want to run in experiments. It is useful when you are thinking about different data transforms and modelling algorithms that you could investigate with a controlled experiment later. It is excellent for getting ideas and playing what-if scenarios. The interface is divided into 6 tabs, each with a specific function.

Experimenter

- The Weka Experiment Environment is for designing controlled experiments, running them, then analyzing the results collected. It is an environment for performing experiments and conducting statistical tests between learning schemes. This interface is for designing experiments with your selection of algorithms and datasets, running experiments and analyzing the results. The tools for analyzing results are very powerful, allowing you to consider and compare results that are statistically significant over multiple runs.

KnowledgeFlow

- It is an environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning. Applied machine learning is a process and the Knowledge Flow interface allows you to graphically design that process and run the designs that you create. This includes the loading and transforming of input data, running of algorithms and the presentation of results.
- The Weka KnowledgeFlow Environment is a graphical workflow tool for designing a machine learning pipeline from data source to results summary, and much more. Once designed, the pipeline can be executed and evaluated within the tool. The KnowledgeFlow Environment is a powerful tool.

Workbench

- It is an environment that combines all of the GUI interfaces into a single interface. It provides three main ways to work on your problem: The Explorer for playing around

and trying things out, the Experimenter for controlled experiments, and the KnowledgeFlow for graphically designing a pipeline for your problem. The Weka Workbench environment that combines all of the GUI interfaces into a single interface. It is useful if you find yourself jumping a lot between two or more different interfaces, such as between the Explorer and the Experiment Environment.

Simple CLI

- Weka can be used from a simple Command Line Interface (CLI).This is powerful because you can write shell scripts to use the full API from command line calls with parameters, allowing you to build models, run experiments and make predictions without a graphical user interface.
- The Simple CLI provides an environment where you can quickly and easily experiment with the Weka command line interface commands. It provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

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 symptoms or behaviour of human body with a known target class. This data set is used as training data for the classifier.

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-angina pain Value 4: asymptomatic

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

5. chol: serum cholesterol 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

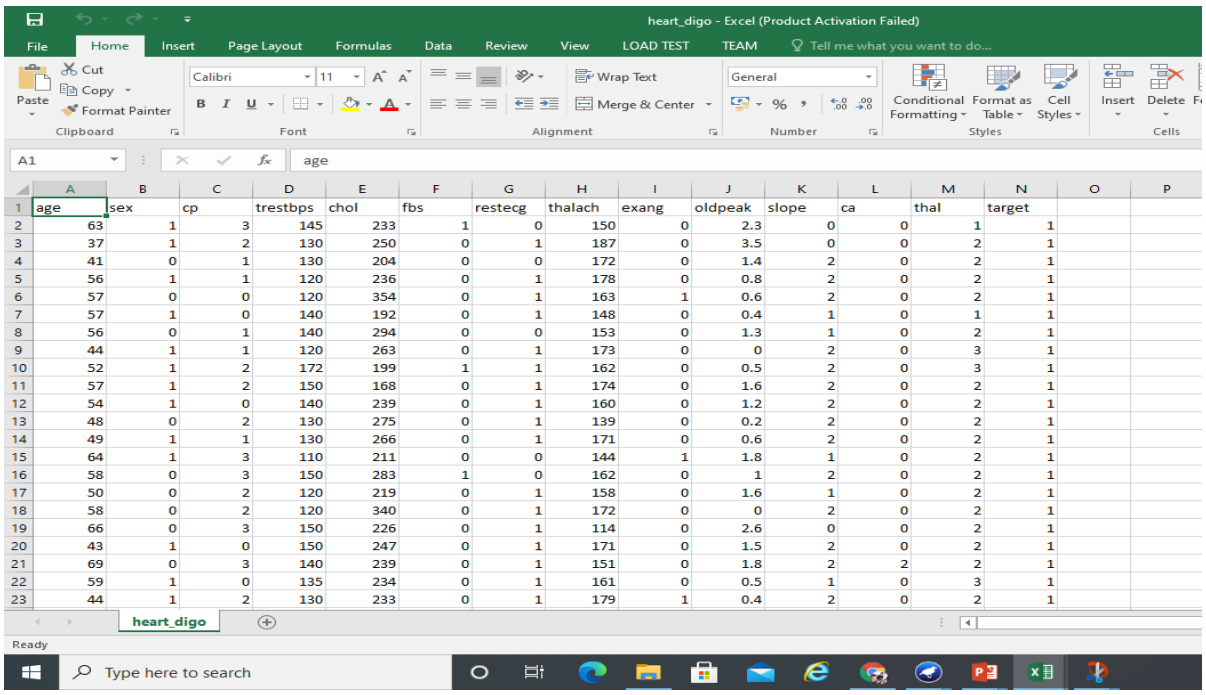
12. ca: number of major vessels (0-3) coloured by fluoroscopy

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

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

3.2 Data Sets

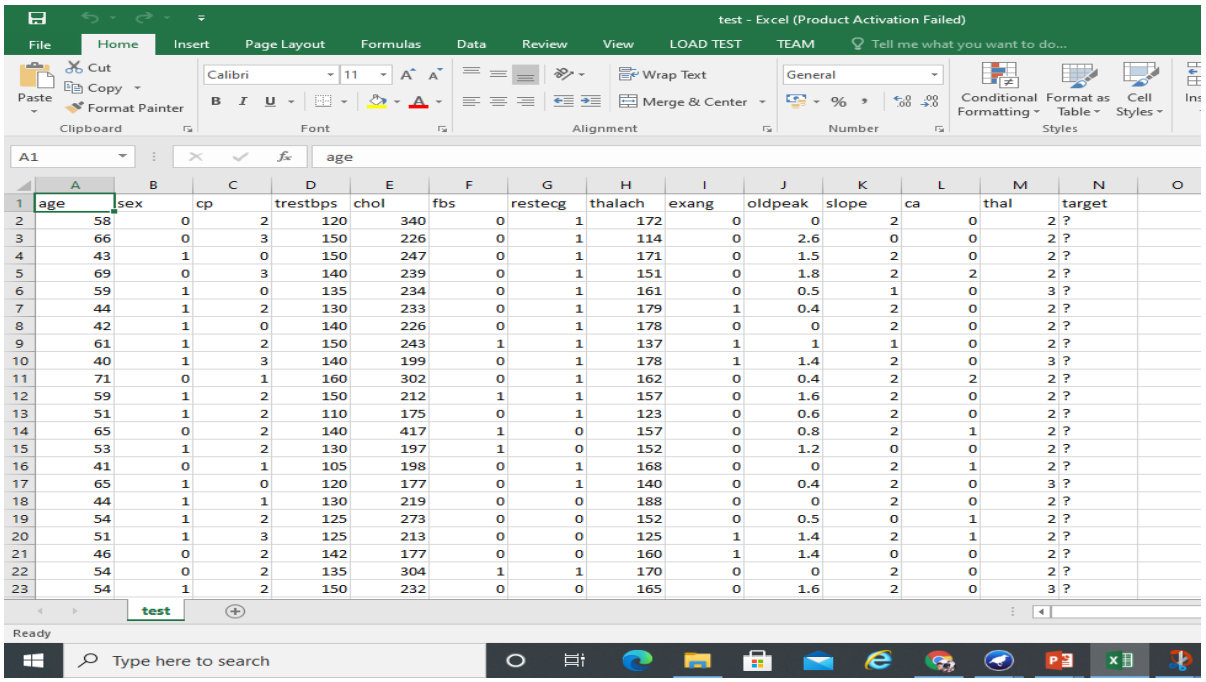
Training Data Set:



heart_digo - Excel (Product Activation Failed)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1

Test Data set:



test - Excel (Product Activation Failed)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
58	0	2	120	340	0	1	172	0	0	2	0	2	?
66	0	3	150	226	0	1	114	0	2.6	0	0	2	?
43	1	0	150	247	0	1	171	0	1.5	2	0	2	?
69	0	3	140	239	0	1	151	0	1.8	2	2	2	?
59	1	0	135	234	0	1	161	0	0.5	1	0	3	?
44	1	2	130	233	0	1	179	1	0.4	2	0	2	?
42	1	0	140	226	0	1	178	0	0	2	0	2	?
61	1	2	150	243	1	1	137	1	1	1	0	2	?
40	1	3	140	199	0	1	178	1	1.4	2	0	3	?
71	0	1	160	302	0	1	162	0	0.4	2	2	2	?
59	1	2	150	212	1	1	157	0	1.6	2	0	2	?
51	1	2	110	175	0	1	123	0	0.6	2	0	2	?
65	0	2	140	417	1	0	157	0	0.8	2	1	2	?
53	1	2	130	197	1	0	152	0	1.2	0	0	2	?
41	0	1	105	198	0	1	168	0	0	2	1	2	?
65	1	0	120	177	0	1	140	0	0.4	2	0	3	?
44	1	1	130	219	0	0	188	0	0	2	0	2	?
54	1	2	125	273	0	0	152	0	0.5	0	1	2	?
51	1	3	125	213	0	0	125	1	1.4	2	1	2	?
46	0	2	142	177	0	0	160	1	1.4	0	0	2	?
54	0	2	135	304	1	1	170	0	0	2	0	2	?
54	1	2	150	232	0	0	165	0	1.6	2	0	3	?

4. IMPLEMENTATION RESULTS

Here we are performing three different algorithm on given data set in order to create or to made out some useful 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]

a. 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]

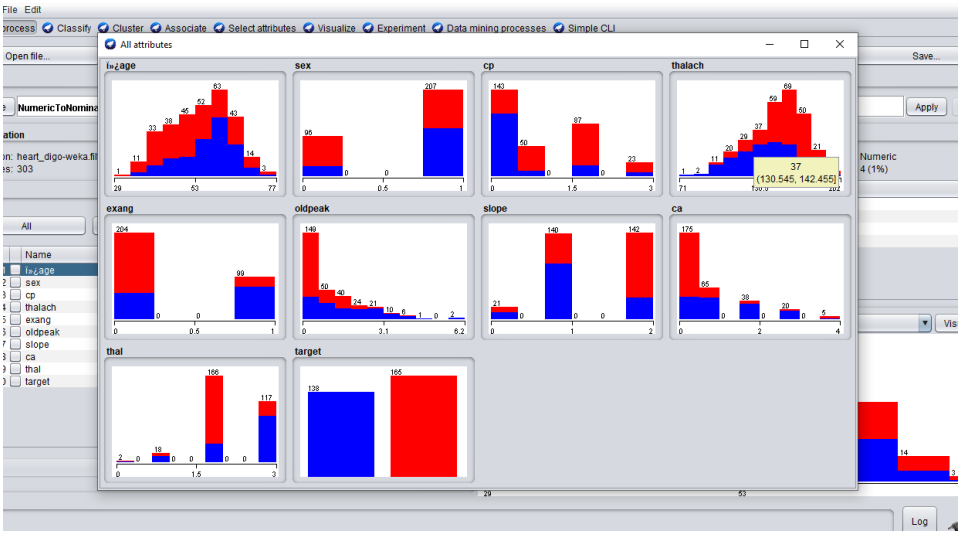
4.2 Steps:

1. Open Workbench in weka Explorer.

2. Under "pre-processing" tab add new file

* Here you can filter the attributes and its instance using the filter option then you can increase the data set reliability

*At the bottom of the window, you see the visual representation of the class values by clicking on the Visualize All button, you will be able to see all features in one single window as shown below.



To predict nominal or numeric quantities, we have classifiers in Weka. Available learning schemes are decision-trees and lists, support vector machines, instance- based classifiers, logistic regression and Bayes' nets. Once the data has been loaded, all the tabs are enabled. Based on the requirements and by trial and error, we can find out the most suitable algorithm to produce an easily understandable representation of data. To classify the data set based on the characteristics of attributes, Weka uses classifiers.

3. Under "classify" tab specify the Algorithm for training the data set

4. Choose "use training set" from "test option"

5. Then start the operation.

Perform this step by changing the selected attribute and algorithm such that you came with higher accuracy number then opt that algorithm modal then save that modal by giving right click and save it.

Once the model have been found out or created. Then you can predict the value or targeted value of the test data set

6. for Predicting:

1. Load the saved model by pressing right mouse button in "result list" and choosing load option and load respective model
2. Choose "Supplied test set" from "test option"
    a. Click "set". (A test instance dialog box appears).
    b. Open file—choose the test data set.

3. Then right click on loaded model in “result list”—click “re-evaluate using the current test set”

7. Visuals Representation of tress are possible by right clicking the modal and click visualize.

So a tree will be displayed according the attributes and the targeted values will be predicted

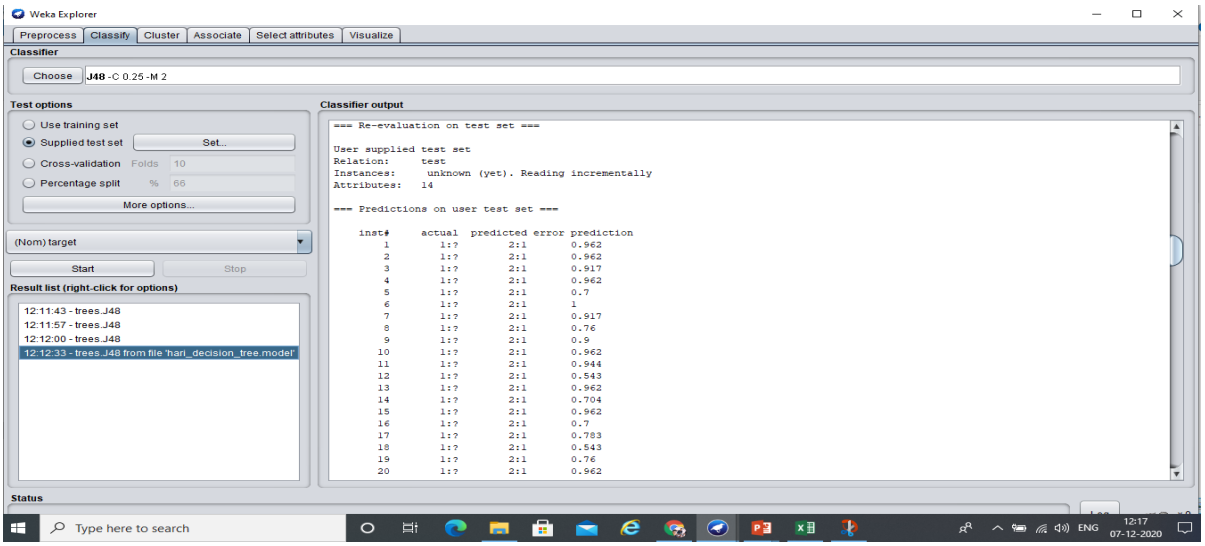
NOTE: For seeing predicted value we must choose “plaintext” by

Clicking “more” from “test options” a dialogue box appears

Then choose “plaintext” from output predictions

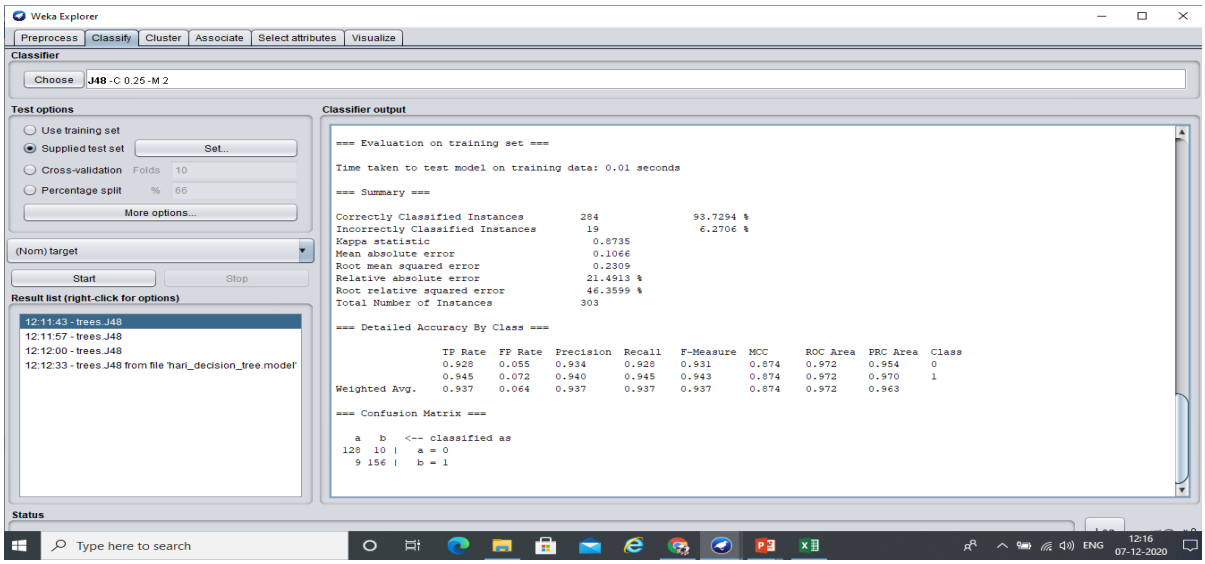
4.3 Results:

a. Test data set evaluated using j48 trained data set

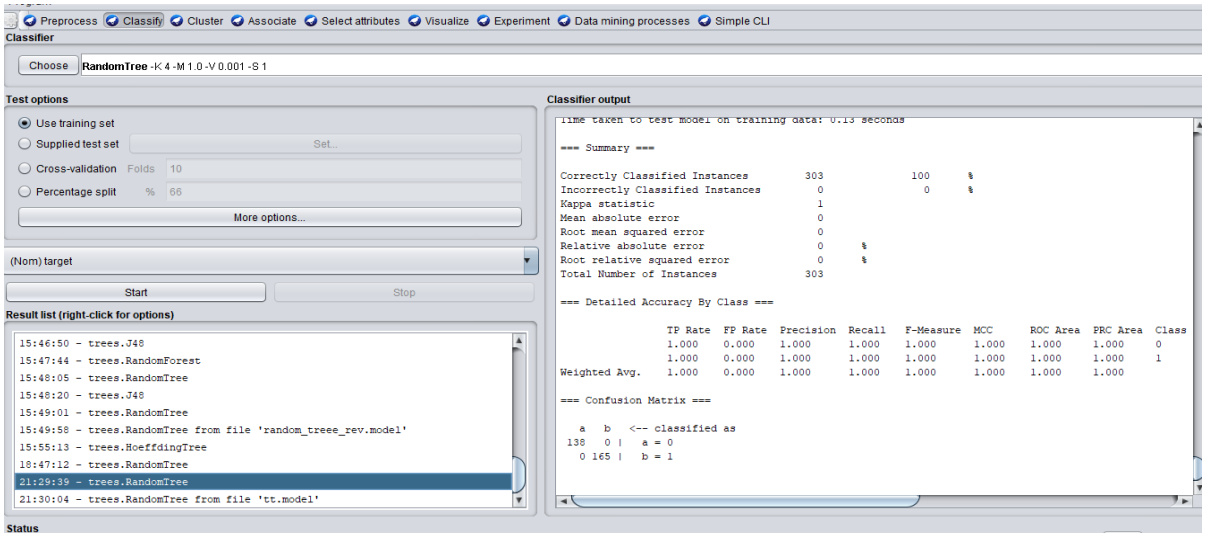


Predicted value

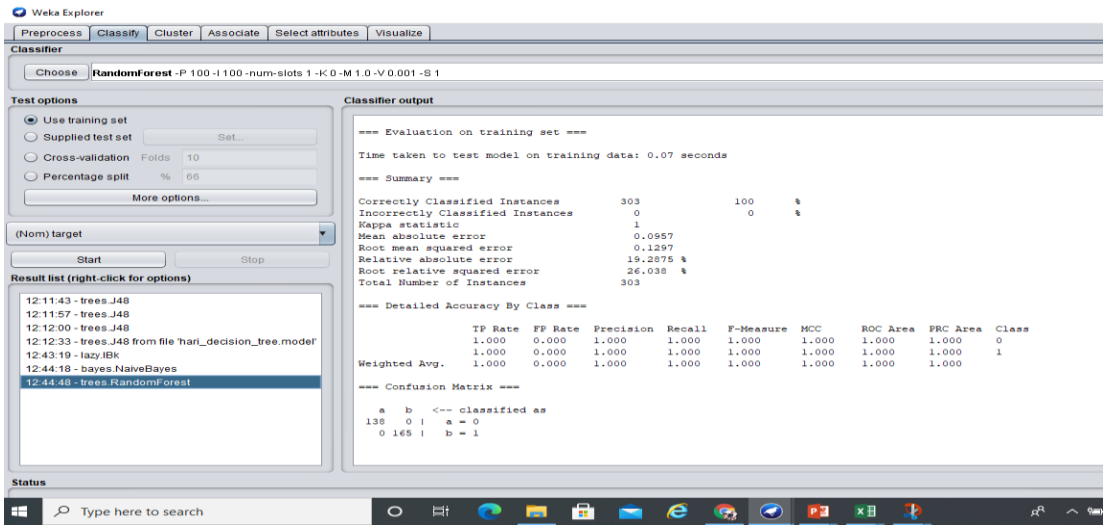
:



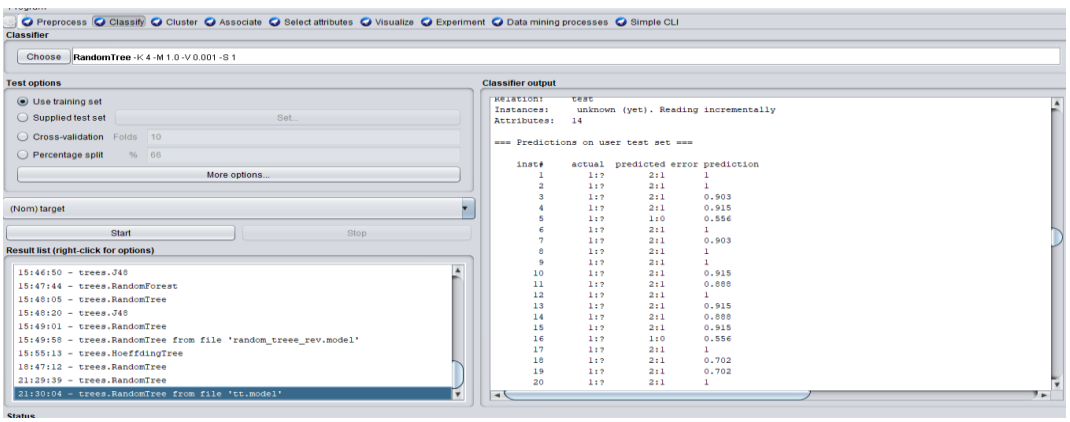
b. Test data set evaluated using Random Tree trained data set



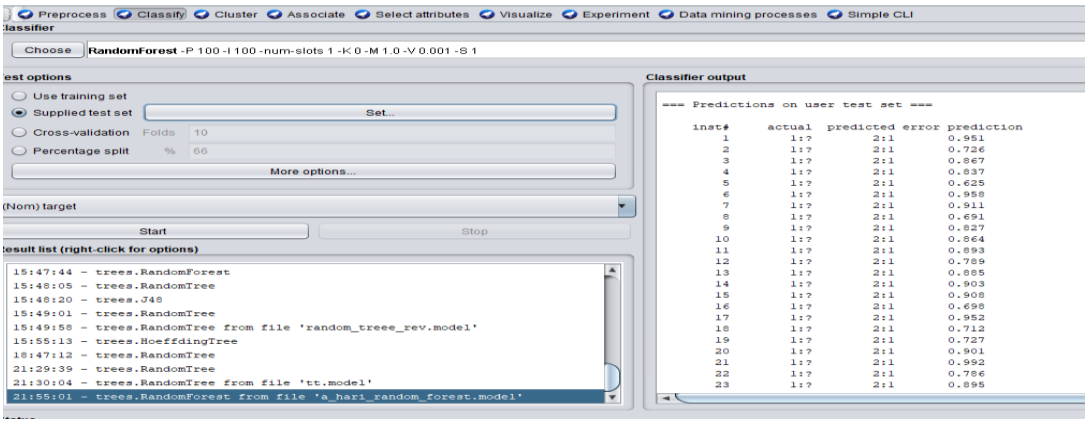
Predicted Value:



c. Test data set evaluated using Random Forest trained data set



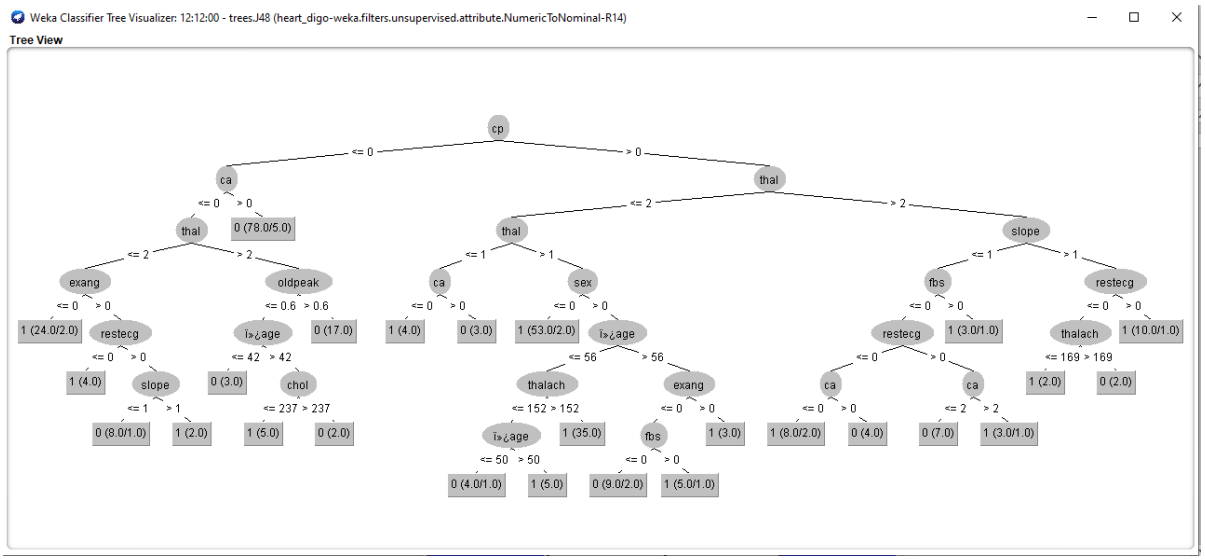
Predicted value



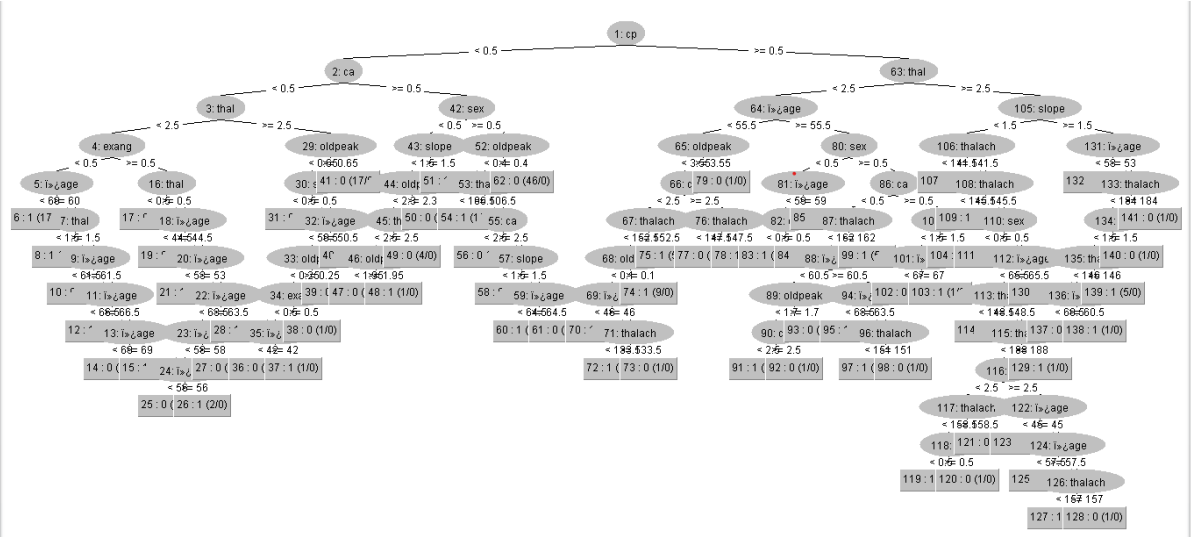
4.3.a Tree Visualization:

Visualize tree to get a visual representation of the traversal tree as seen in the screenshot below. Here cp attribute is the root node, because it have the highest information gain. Information gain measures the amount of information contained in a set of data. It gives the idea of importance of an attribute in a dataset.

Tree Visualization of the Predicted Value(**J48**):



Tree Visualization of the Predicted Value(**Random Forest**):



Summarized Results

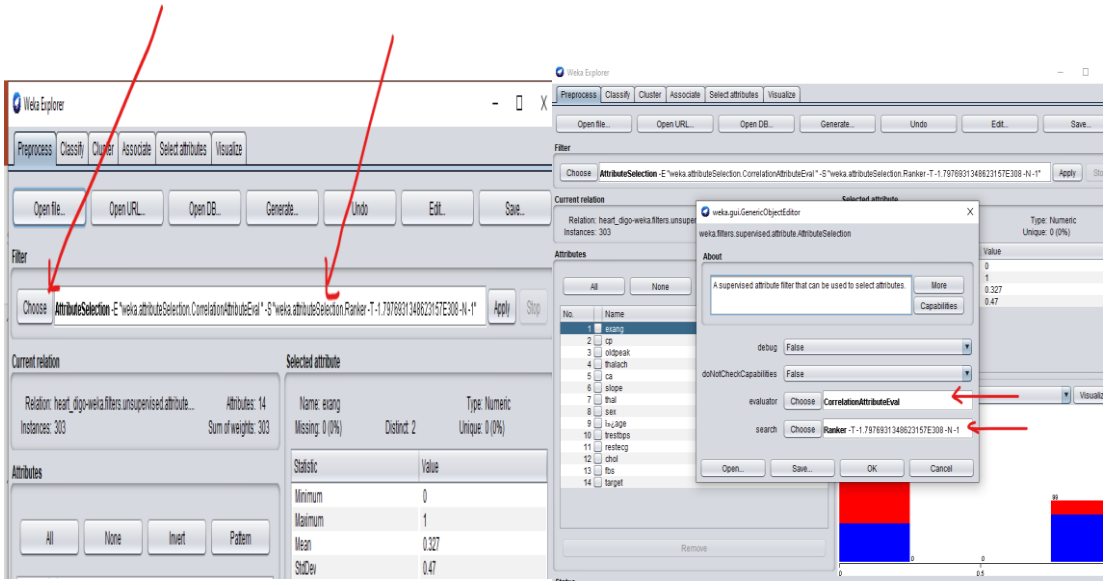
Techniques	Accuracy	Time
Decision Tree Algorithm (j48)	93.7294 %	0.01
Random Tree	100 %	0.08
Random Forest	100 %	0.07

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

Filteration Details:

1. Attribute Selection is Taken from chose >> Suppervised >> attribute selection
2. Evaluator >> correlationAttribute eval

   search >> Ranker



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 forest 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 analyse and predict the future behaviour 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.

WEKA offers a wide range of sample datasets to apply machine learning algorithms. The users can perform machine learning tasks such as classification, regression, attribute

selection, association on these sample datasets, and can also learn the tool using them. WEKA explorer/workbench is used for performing several functions, starting from preprocessing. Preprocessing takes input as an .arff file, processes the input, and gives an output that can be used by other computer programs. In WEKA the output of preprocessing gives the attributes present in the dataset which can be further used for statistical analysis and comparison with class labels.

NOTE: 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/geomaticahelp/concepts/focus_c/oa_classif_intro_rt.html

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