WEB-BASED SYSTEM FOR THE PREDICTION OF HEART DISEASE

AARON RAJ A/L MAYA

SESSION 2018/2019

FACULTY COMPUTING AND INFORMATICS

MULTIMEDIA UNIVERSITY

FEBRUARY 2019

WEB-BASED SYSTEM FOR THE PREDICTION OF HEART DISEASE

BY

AARON RAJ A/L MAYA

SESSION
2018/2019

THE PROJECT REPORT IS PREPARED FOR

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
IN PARTIAL FULFILLMENT
FOR

BACHELOR OF COMPUTER SCIENCE
B.CS (HONS) DATA SCIENCE

FACULTY OF COMPUTING AND INFORMATICS

MULTIMEDIA UNIVERSITY

FEBRUARY 2019

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this thesis has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

Aaron Raj A/L Maya

Faculty of Computing & Informatics

Multimedia University

Date:

ACKNOWLEDGEMENT

I would like to take this opportunity to express my special thanks of gratitude to those who have guided and motivated me to complete this report, especially my supervisor who gave me the opportunity to do this project.

My highest gratitude goes to my supervisor of this project, Dr. Kannan Ramakrishnan who helped me a lot throughout the entire semester. He is a great mentor and have encouraged me when I was halfway through to complete this project. I would like to thank him for giving me such a good final year project experience.

Last but not least, I would also like to thank my family and friends who have supported me along the way. They have always helped me to release my stress and share their knowledge when I was in doubt. I also would like to thank everyone who helped me a lot in finalizing this project within the limited time frame.

ABSTRACT

The Human Heart is perhaps the most important major organ in the body. Recently, Medical research are showing more interest in their scientific research in implementing Artificial Intelligence in the health care industry. Various Data Mining techniques can be implemented for the prediction of heart disease. In this research, a web-based system has been implemented to predict the user heart disease by using Artificial Neural Network. The type of Neural Network being use is the Multilayer Perceptron with using backpropagation algorithm. The outcome of this project is to identify whether the person having heart disease and what type of heart disease the person having.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

1.0 INTRODUCTION

Heart disease is a term that is assigned to refer to a large number of medical conditions that is related to the heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts. Heart disease is a major health problem in today's time. In current literature, we find there are various ways to diagnose for heart disease by using various data mining methods and tools that we are about to discuss.

The background of this project is to implement a web-based system where the system will integrate with R programming. The reason for the integrate is to identify the accuracy of predicting the heart disease when user key in their credentials into the system. By using Artificial Neural Network, it can calculate and give the output to the user. The user can know what the user having heart disease or not having or what type of heart disease having.

Besides that, there are existing system but the disadvantage is the existing system can't predict the heart disease when the user key in their credentials. Most of the existing system are simple. Example, like book appointment, choose doctor, update doctor, what type of sick the user having and etc.

1.1 Project Overview

This project covers methods of prediction for heart disease by using a web based system. It reviews the health of a person's heart by asking questions that can gauge the person's awareness about the heart disease.

By using a web based system, the user need to key in their cholesterol level, high blood pressure, angina, diabetes and the all other requirements into

the system. After the user have key in all the information, the system will give the results that will show to the user what type of heart disease that the user have. After viewing the result that shown by the system, the user then can contact the doctor that is provided through the system.

The information that were gathered several patients will all be stored in the web based system for me to keep as a record of the users. After collecting the data of the patients, it can start to analyse the data whether there is any inconsistencies or missing data from the data set that has collected. From the data set, will be using the R programming language by applying data mining techniques to predict the accuracy of type of heart disease for the system.

The R Studio can determine the most accurate prediction of the type of heart disease by comparing all the different types of data mining algorithms. By comparing the different type of data mining techniques, R Studio can also use it to evaluate the best data mining techniques with the highest accuracy for predicting the type of heart disease for this system.

From the observations of the results of the data, it can analyse, compare and predict which gender have the most for heart disease, which type of heart disease that they suffer the most, at which age having the most heart disease and so on. After all these observation from the data collected, the R programming can find a way to prevent the problem of heart disease which keeps increasing rapidly in the number every year.

1.2 Problem Statement

Prediction of heart disease using data mining is one of the most interesting and challenging tasks. The high wrongly diagnosed cases need to develop a fast and efficient prediction of heart disease system. The main objective is to identify the features from the data set by using classification model. The attributes that are more relevant to prediction of heart disease can be observed. This will help to understand the root causes of disease in depth.

1.3 Project Objectives

This project aims to identify:

-The optimal Artificial Neural Network model for the prediction of heart disease.

Artificial Neural Network, Multilayer Perceptron by using Backpropagation algorithm is been used for the prediction of heart disease. The reason using Artificial Neural Network is to predict the accuracy of the user whether having heart disease or not having it and what type of heart disease the user having.

-Develop a web-based system for the prediction of heart disease.

The reason to develop a web based system is for user to key in their credentials such cholesterol level, blood pressure, sugar level and etc. When the user click analyse heart, the web-based system will predict by using Artificial Neural Network whether the user having heart disease or not having it and what type of heart disease the user having.

1.4 Scope

Here the scope of the project is to find the accuracy of heart disease of the user by using data mining technique. The main reason use data mining technique for the prediction is to see how the algorithm works when the user key in their details and what is the output of it to be.

1.5 Deliverable

Here the deliverable will be using R, PHP, PHPMYADMIN and HTML to implement a web-based system. The purpose of implement a web-based is to predict the heart disease of the particular user. When the user key in their credentials to the web-based system, the system will predict by using Neural Network whether the user having heart disease or not having or what type of heart disease the user having.

Data mining technique that be using for this semester is supervised learning, classification. Data mining tools will be using are Artificial Neural Network, Multilayer Perceptron by using Backpropagation algorithm. The first optimal model is to predict whether the user having heart disease or not having it. The second optimal model is to predict what type of heart disease the user having.

1.6 Overall Organization

In the chapter 2 will be explaining more details about review of different types of general papers where various researchers used various data mining techniques and review background systems to produce data for heart disease predictions.

In the chapter 2 will be explaining about the data set being used for this project, diagram for use case, sequence diagram, ERD, DFD and what type of data mining technique been applied for this project. In the chapter 4 will be explaining about the system that has been implemented and what type of pre-processing method been used. There are admin, user and doctor in the web-based system. Each role have its own function in the system.

Chapter 5 will explaining about the pseudocode by using R Studio, the screenshot of the system and development tools that has been used for this project. Chapter 6 will be explaining more about the diagram, the output by using R Studio, what algorithm and parameters been used in the code.

2.0 LITERATURE REVIEW

The literature review of this project presented is about the review of different types of general papers where various researchers used various data mining techniques and review background systems to produce data for heart disease predictions.

2.1 Review On Conference

2.1.1 Applying Advanced NN-based Decision Support Scheme for Heart Disease Diagnosis, 2012

Sameh Ghwanmeh has used Neural Network based decision support system for prediction of heart disease. The Artificial Neural Network decision support system is composed of two parts which are hardware and GUI software. The hardware is constructed to represent the pattern recognition sensor to collect heart sounds. Since the collected sound by stethoscope is at a low amplitude level, a microphone capsule is inserted between the stethoscope and the PCI sound card interface to amplify the sound waves fed by the stethoscope. The GUI software part is implemented using MATLAB and Neural Network Toolbox. He collected the data from two different sources. The first sources are provided by the Jordanian Royal Medical Services. While the second source is the recorded samples directly provided from the patient tested on the proposed system. The MATLAB FET function is used to extract the main features of the sound. For training of Neural Network, three Neural Network

are built which are mitral stenosis disease NN, ventricular septal defect disease NN and aortic stenosis disease NN. A series of experiments with different samples of data have been conducted to measure the effectiveness and accurateness of the proposed system. For the correctness of the knowledge base and outputs, NN are trained repeatedly and its parameters are modified where is needed. Results show that accurate outputs, with 98% classification accuracy, are resulted with the trained samples while the system produces acceptable results for non-trained samples, with 92% classification accuracy.

2.1.2 Decision Support in Heart Disease Prediction System using Naïve Bayes, 2011

Mrs.G. Subbalakshmi, Mr K. Ramesh, Mr.M. Chinna Rao has built the Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modelling technique namely Naïve Bayes to discover the relationship between variables in data for the healthcare industry. Form of web based is implemented in the questionnaire application.

2.1.3 Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, 2009

Shantakumar B.Patil, Y.S. Kumaraswamy proposed a methodology for the extraction of significant patterns from the heart disease data warehouse for the

prediction of heart attacks. Firstly, the data warehouse is pre-processed in order to make it suitable for the data mining process. After the pre-processing is over, the heart disease data warehouse is clustered with the aid of the K-means clustering algorithm, which will extract the data that is applicable to heart attacks data warehouse.

2.1.4 Prediction System For Heart Disease Using Naïve Bayes, 2012

Shadab Adam Pattekari and Asma Parveen proposed an Intelligent System that uses a data mining modelling technique namely Naïve Bayes. It is implemented as a web based application where the user answers several predefined questions. It retrieves hidden data from stored databases and compares the user values with the trained data set. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which the traditional decision support system cannot. By providing effective treatments, it also helps to reduce treatment costs.

2.1.5 A Survey of Data Mining Techniques on Risk Prediction: Heart Disease, 2013

K. Srinivas presented Applications of Data Mining Techniques in Healthcare and Prediction of Heart Disease. They presented how data mining techniques

such as Artificial Neural Network, Rule based, Naïve Bayes and Decision Tree can be used for classification of massive Volume of healthcare data. The data mining tool used for exploratory data analysis, machine learning and statistical learning algorithm is called Tanagra. The training data set consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers is evaluated and their results are analysed. The results of the comparison are based on 10 tenfold cross-validations. According to the attributes, the dataset is divided into two parts that are 70% of the data are used for training and 30% are used for testing. Among these classification algorithms, APRIORI algorithm is considered the best algorithm.

2.1.6 An Analysis of Heart Disease Prediction using Different Data Mining Techniques, 2012

Nidhi Bhatla and Kiran Jyoti used the model that combines the genetic algorithms for feature selection and fuzzy expert system for effective classification. Fuzzy set theory and fuzzy logic are highly suitable for developing knowledge-based systems and is conducted in MATLAB by using fuzzy tools. Mamdani model of the fuzzy system is used and through it, fuzzy rules are then generated. Initially, fuzzification is performed by collecting the set of input data. Thereafter, the fuzzy linguistic variable, fuzzy linguistic terms and membership functions are used to convert to a fuzzy set.

Lastly, the defuzzification step is formed. Nidhi and Kiran used their in-house dataset. They only used 6 attributes which they found out was sufficiently effective and necessary for the prediction of heart disease. The 6 selected attributes are then inputted in the system and the results of the output will either show a value of 0 or 1, which indicates the absence or presence of heart disease in patients respectively. The advantage is that the decision tree is highly accurate with the help of a genetic algorithm and feature subset selection.

2.1.7 Predicting Heart Attack Using Fuzzy C Means Clustering Algorithm, 2015

Dr.G. Rasitha Banu and J.H. Bousal Jamala proposed Fuzzy C-Means Clustering algorithm by using historical heart disease databases. The way it searches for the risk of heart disease in patients is by using the profiles collected from the patients. The FCM classifications can evolve an optimum number of clusters. It finds the abnormal and normal cases very efficiently. Pre-processing of the data is done to remove all the duplicate records and add missing data in the database. FCM classifications are used to classify the data in order to know whether the heart disease is present or not present. The dataset Banu and Jamala used are collected from the patients who suffer from heart disease. The data that was collected is from 270 patients where it contains 73 attributes and they used only 13 attributes to carry out the experiment. The input and output levels where two values are predicted

normal and abnormal are based on the 13 attributes. This method is used to diagnose heart disease in order to detect it early and to get the best accuracy for heart disease prediction. The efficiency of the classifier is tested using the records collected from 270 patients, which gives a classification accuracy of 92%. Based on the results, it shows that the proposed clustering algorithm can predict a patient's current heart disease in a more efficient and cost-effective way than the other well-known algorithms.

2.1.8 Heart Disease Prediction System using Associative Classification and Genetic Algorithm, 2012

Jabbar Akhil and Bulusu Deekshatulu proposed the APRIORI algorithm. They used it for discovering the rules for the algorithm and based on their research it requires multiple passes over the database in order to determine prediction of heart disease. The dataset they used is from in-house. The advantage of using this algorithm is a more efficient association classification heart disease prediction. Researcher used Gini Index which produces compact rule set and filter by applying Z-statistics and genetic algorithm to predict the accuracy of the heart disease. The main motivation for using a genetic algorithm in the discovery of high-level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interesting values. Based on the result, it shows that most of the classifier rules help in the best prediction of heart disease which even helps doctors in their diagnosis decision.

2.1.9 Review of Heart Disease Prediction System Using Data Mining and hybrid intelligent technique, 2013

R.Chitra and V.Seenivasagam proposed K-Means clustering with decision tree method to predict heart disease. In their research work, it suggested several centroid selection methods for k-means clustering to increase the efficiency. The sensitivity, specificity and accuracy are calculated with different initial centroids selection methods and different numbers of clusters. Comparing integrating k-means clustering and decision tree could enhance the accuracy of the decision tree in diagnosing heart disease. Integrating k-means clustering and the decision tree could achieve higher accuracy than the paging algorithm in the diagnosis of heart disease patients. The dataset that they used is from federal medical fields and the University of California. Th Prediction system for heart disease used is a system that contains a huge amount of data which is used to extract hidden information for making an intelligent medical diagnosis. To develop the system, medical terms such as sex type, blood pressure, cholesterol and 13 input attributes are used. To get more appropriate results, two more attributes which are obesity and smoking were considered as important attributes for heart disease. A Multi-layer Perceptron Neural Network that maps a set of input data that consists of 3 which are an input layer, a hidden layer and an output layer. There is a connection between each layer and weights are assigned to each connection. The primary function of neurons in the input layer is to divide the input into neurons in the hidden layer. The data set consists of a total of 573 records in heart disease database. The total

records are divided into two data sets, one is used for training consists of 303 records and another for testing consists of 270 records. Initially, the data set contains some fields in which some value in the records were missing. The advantage of using this algorithm is to ensure the k-means clustering will increase efficiency for predicting and by using this algorithm, it could enhance the accuracy of the decision tree in diagnosing the heart disease. They conclude that the Hybrid Intelligent Algorithm improves the accuracy of the prediction of heart disease system. The commonly used techniques for Heart Disease Prediction and their complexities are summarized in this area. By using this algorithm, it is easy to implement but difficult to predict the number of clusters in the data set.

2.2.0 Study of Heart Disease Prediction using Data Mining, 2014

K.Sudhakar and Dr.M.Manimekalai proposed J48 algorithm using the pruning method to build a tree. Pruning is a technique that reduces the size of the tree by removing over fitting data which leads to poor accuracy in predictions. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides a balance of flexibility and accuracy. The data set they used is from the health care industry. The data set contains records of 13 attributes in each record. The supervised networks which are the Neural Network with back propagation algorithm are used for training and testing of the data. The result of the

dataset contains records of patients having heart disease. Three constraints were introduced to decrease the number of patterns. The three constraints are first that the attributes have to appear on only one side of the rule. Secondly, the attributes are separated into groups, for example, uninteresting groups. Thirdly, in a rule, there should be a limited number of attributes. The data set consist of 670 people, distributed into two groups, namely normal people and patients with heart disease were employed to carry out the experiment for the associative classifier. The three different data mining classification technique which are Neural Network, Decision Tree and Naïve Bayes are used to analyse the data set. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers which are an input layer, a hidden layer and an output layer. There is a connection between each layer and weights are assigned to each connection. The primary function of neurons in the input layer is to divide the input into neurons in the hidden layer. The neuron of the hidden layer adds input signals with weights of respective connections from the input layer. The advantages of using the decision tree approach are more powerful for classification problems by using the J48 algorithm for predicting the accuracy of the heart disease. These large amounts of data are very important in the filed of Data Mining to extract useful information and generate relationships among the attributes. Heart disease diagnosis is a complex task which requires much experience and knowledge. Heart disease is a single largest cause of death in developing countries and predicting the heart disease by using data mining tools could save their life.

2.2 Comparison Table

Title Number	Description of Methods	Data set	Result	Advantage	Limitation	Comments
2.1.6	Combine genetic to specify the classification	6 attributes are effective for heart disease prediction	Decision tree produce the highest accuracy 99.2% with 0.09s for the model construction time	The decision tree is shown very good accuracy	Mutation rate should be low	Decision tree find good accuracy for predicting the heart disease
2.1.7	The Fuzzy C-Means Clustering method can evolve number of clusters	The database contains 73 attributes but uses 13 attributes	Fuzzy C-Means accuracy has achieved as expected	Fuzzy C Means Clustering Algorithm get the best accuracy for the prediction	Long computational time	The algorithm ignores noise sensitivity deficiency of Fuzzy C-Means
2.1.8	APRIORI association produce huge number of association rule	Gini Index used as filter to reduce number of candidate item sets	APRIORI produce best accuracy 88.9% for the prediction	Applying Z-statistics and genetic algorithm predict the best accuracy	Use a uniform minimum support threshold	Very easy to implement and use large itemset

2.1.9	Increase efficiently when using centroid	15 attributes are used to get more appropriate result	Naive Bayes produce the best accuracy 52.33% with 609 miliseconds time taken for the model construction	K-means clustering increase the efficiency	K-means can only handle numeric data	Easy to implement but difficult to predict number of cluster in the dataset
2.2.0	Pruning reduce size of tree by removing overfitting data	The dataset contains 13 attributes	Neural Network produce the highest accuracy among other algorithms	Decision tree powerful for classification	Difficulty in representing functions such as parity or exponential size	The decision tree is robust to outliers

2.3 Limitations

Sameh Ghwanmeh, Nidhi Bhatla and Kiran Jyoti are those researcher who used MATLAB for the heart disease prediction. MATLAB is great for prototyping and investigating data but it is just an awful language for building a complete application. The language was designed around small scripts to do two-dimensional matrix math and everything else is a bolt-on. It leads to an astounding number of gotchas that can lead to real bugs and limitations. The limitation is when it takes much CPU time for computation. It makes real time application very complicated and confuse the programmer.

Dr.G. Rasitha Banu and J.H. Bousal Jamala are those researchers who used VISUAL BASIC.NET for the heart disease prediction. Visual basic is a proprietary programming language written by Microsoft. The programs written in Visual Basic cannot easily be transferred to other operating systems. Besides, there are some fairly minor disadvantages compared with C. C has a better declaration of arrays and it is possible to initialise an array of structures in C at declaration time. This is impossible to do in Visual Basic.

Nidhi Bhatla and Kiran Jyoti are those researchers who used WEKA for the heart disease prediction. The limitation of using WEKA is that it only can handle small datasets. If the datasets are too big and there only a few megabytes of memory available, an Out Of Memory error occurs. Another limitation is that one is to copy the entire dataset to another file and use the filename as the entry in that history table. What will happen is that it will be

very slow because it has to copy the entire dataset every time when executing an algorithm and it would be very space consuming for secondary storage.

2.4 Survey Similar Application/Website

2.4.1 : Cardio Smart Web-based System

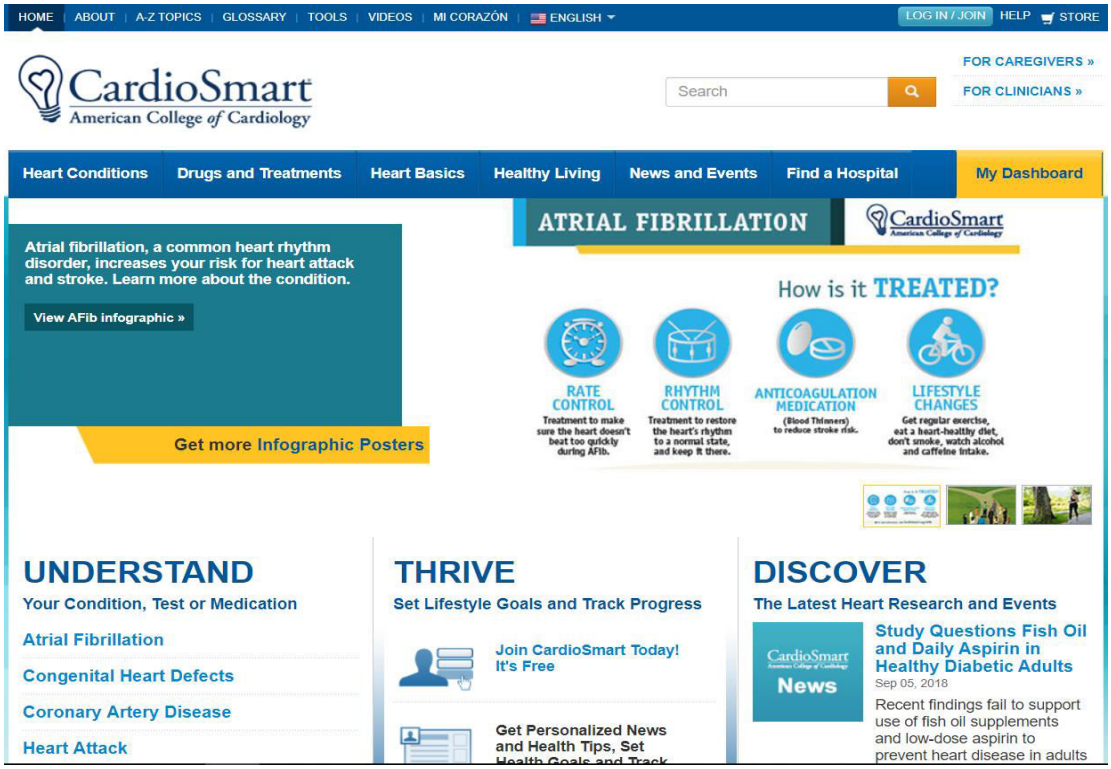


Figure 2.4.1 : Cardio Smart Web-based system.

As shown in the figure 2.4.1 above, the Cardio Smart web-base system. Basically, this system just provide general knowledge of heart disease, how to prevent the heart disease, what causes of having the heart disease and so on.

2.4.2 : Mayo Clinic Web-based System

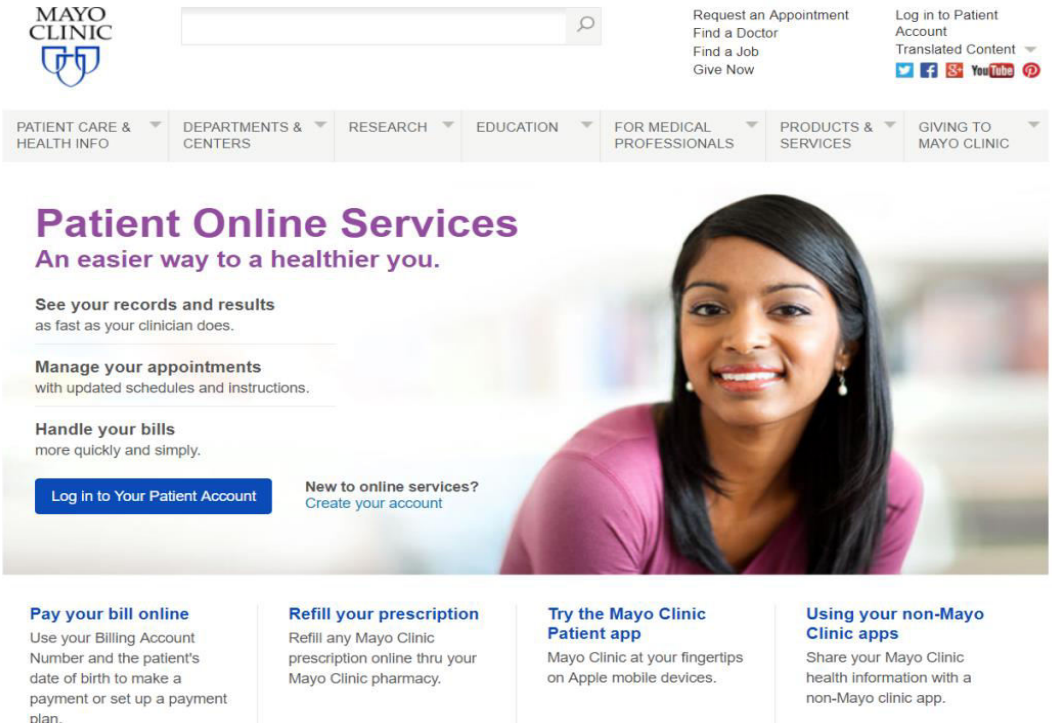


Figure 2.4.2 : Mayo Clinic Web-based System.

As shown in the figure 2.4.2 above, the Mayo Clinic Web-based System. If the user are new to the system, the user need to create an account in order to access to the system. The user then need to login as a patient account as shown above. Once the user access to the system, the user can find a doctor which is near to the user location. Besides that, the user can request an appointment to the system where the user can set the date and time to make appointment with the doctor.

2.4.3 : American Heart Association Web-based System

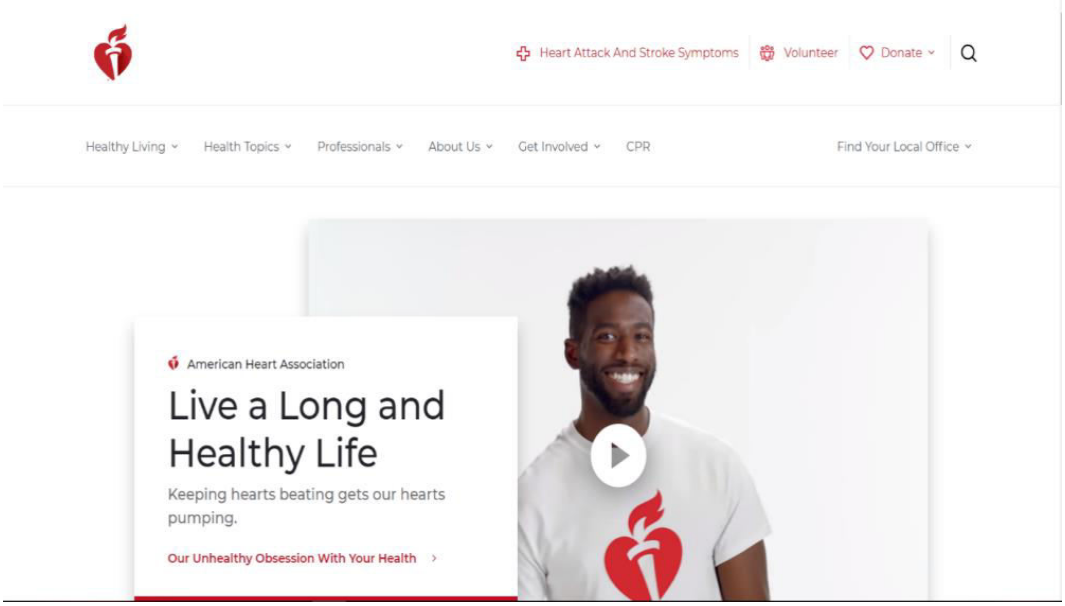


Figure 2.4.3 : American Heart Association Web-based system.

As shown in the figure 2.4.3 above, the American Heart Association web-based system. Basically, this system just provide general knowledge about heart disease. This system also have Volunteer where the user can be volunteer to spread about how to prevent the heart disease and so on. Besides that, the user can make donation. The purpose of the donation is to help save lives from heart disease.

2.5 Summary

The researcher has done various data mining technique in order to get the best accuracy of the prediction of heart disease. Some researchers use an additional algorithm to combine the data mining algorithm that they used to get the exact accuracy. Besides, some of the researchers reduce the number of attributes in the dataset. By reducing the number of attributes, it will increase the efficiency and also less time is taken for it to produce good accuracy for the prediction. Most of the researcher didn't develop web-based systems for heart disease prediction. Developing web-based systems will be useful for the people who suffer from heart disease because it helps to save lives.

3.0 THEORETICAL FRAMEWORK

3.1 Dataset Description

The dataset that I am using for this project is based on the Cleveland.csv dataset which I got it from the ethernet. In the Cleveland.csv dataset, there are 15 attributes which are age, gender, cp, trestbps, cholesterol, fbs, restecg, thatach, exang, oldpeak, slope, ca, thal, diag, and grp. The age range is between 29 years old to 77 years old to 77 years old with a mixture of male and female. The cp in the third column of the dataset means types of chest pain. For example, typical angina, atypical angina and etc. The next column is trestbps which means resting blood pressure.

The following column is cholesterol. Total cholesterol level less than 200 milligrams per decilitre (mg/dL) is considered desirable and optimal for a normal person. A reading between 200 and 239 mg/dL is considered borderline high and a reading of 240 mg/dL and above is considered high. The six column which is fbs means fasting blood sugar. There are 2 types of classes which if true means the person's fasting blood sugar level is less than 100 mg/dL which is considered normal and if it is false, it means the person's fasting blood sugar level is higher than 126 mg/dL which is considered as diabetic which could lead to heart disease.

The next attribute is restecg which is resting electrocardiographic results. The resting electrocardiographic results show that the patient if their heart is normal or hypertrophy. The eight attribute is thatach which means maximum heart rate achieved. For example, for a 50 years old person, the estimated maximum age-related heart rate would be 170 beats per minute (bpm) for a moderate intensity physical activity. The next attribute is exang which means exercise-induced angina where there are two classes in that attribute which if true means that the person has exercise-induced angina and if it is false means that the person doesn't have exercise induce angina. The 10^{th} attribute is old peak which means depression induced by exercise relative to rest. The normal old peak is around 2.5. If the normal old peak is around 3.0 and above, then it is considered not normal because it makes the heart pump harder to circulate blood throughout the body. The next attribute is slope which means the slope of the peak exercise segment.

The following attribute which is ca means a number of major vessels which range from 0 to 3. The next attribute is thal which means Thalassemia where it is an inherited blood disorder characterized by abnormal haemoglobin production. The person who has Thalassemia is either normal or fixed defect or reversible defect. The 14th attribute is diag which means diagnosis where there are two variable whether the person is buff or sick. The last attribute is a group where it will be the output. There are five variables under the group attribute which are healthy, S1, S2, S3, and S4. S1 means the person has typical angina disease, S2 means the person has atypical angina disease followed by S3 where the person has non-anginal pain disease and lastly S4 where the person has asymptomatic disease.

3.2 Data Pre-processing Techniques for Data Mining

3.2.1 Data Integration

In this project will be using Data Integration for the Cleveland data set where it combine the data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes or flat files. There are a number of issues to consider during data integration. For example, how can the data analyst or the computer be sure that cholesterol attribute in one database and chol in another refer to the same entity. Database and data warehouse typically have metadata which means data about the data. The metadata can be used to help avoid errors in schema integration. Redundancy is another important issue. An attribute may be redundant if it can be derived from another table such as annual revenue.

Inconsistencies in attribute or dimension naming also may cause redundancies in the resulting data set.

3.2.2 Data Reduction

In this project will be using Data Reduction for the Cleveland data set where the data will be reducing the number of attributes. Complex data analysis and mining on huge amounts of data may take a very long time, making such analysis impractical or in feasible. Data reduction techniques have been helpful in analyzing the reduced representation of the data set without compromising the integrity of the original data and yet producing the quality knowledge. The concept of data reduction is commonly understood as either reducing the volume or reducing the dimensions. There are a number of methods that have facilitated in analyzing a reduced volume or dimension of data and yet yield useful knowledge. Certain partition based methods work on the partition of data tuples, that is, mining on the reduced data set should be more efficient yet produce the same analytical results.

3.3 Data Mining Techniques

3.3.1 Classification

The Classification method been used under the Supervised Learning in the Machine Learning. The meaning of Supervised Learning is where the program is trained on a predefined set of training examples which then facilitate its ability to reach an accurate conclusion when given new data. A classification problem is when the output variable is a category, such as "having heart disease" or "not having heart disease". Example of classification algorithms are

Decision Tree, Neural Network, ID3, Naïve Bayes, Support Vector Machines and etc.

3.3.2 Neural Network

The Neural Network algorithm to apply it for the Cleveland data set. The reason why choosing Neural Network algorithm is that it has the ability to learn and model non-linear and complex relationships, which is really important because in real life, many of the relationships between inputs and outputs are non-linear as well as complex. Unlike many other prediction techniques, Neural Network does not impose any restrictions on the input variables. Additionally, many studies have shown that NN can be better model for heteroskedasticity. For example, data with high volatility and non-constant variance, given its ability to learn hidden relationships in the data without imposing any fixed relationships in the data.

3.4 Diagram for the Use-case

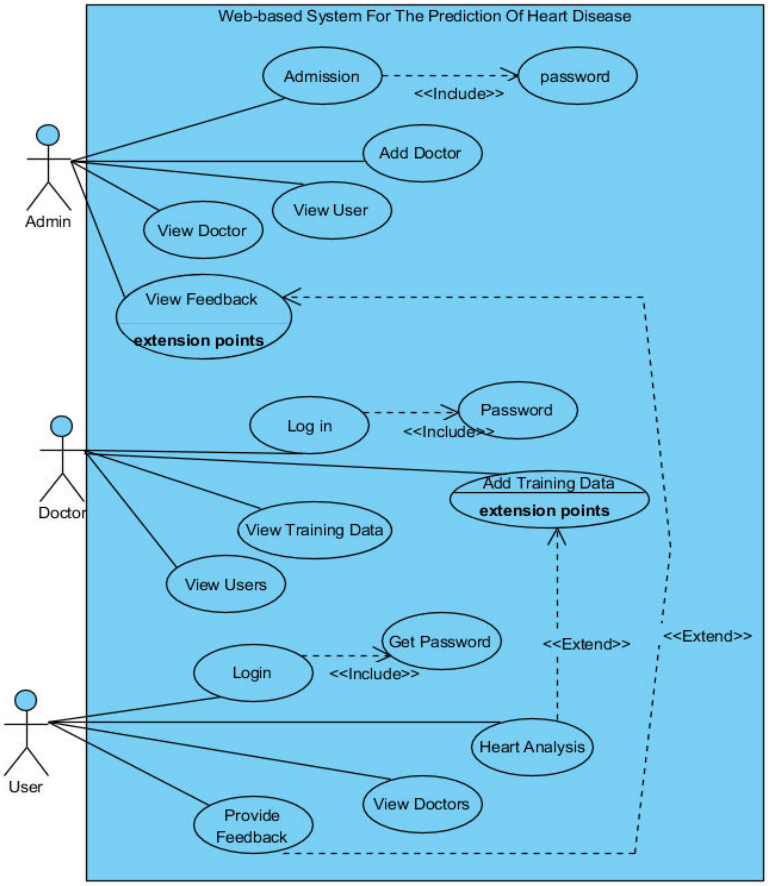


Figure 3.4 : Use-case Diagram for Web-based System For Heart Disease Prediction

The above Figure 3.4 shows the use case diagram of this system. The use cases in this project are based on the functional requirements after analyzing the user requirements and comparing the existing system. There are three actor in this system, which is Admin, Doctor and User. The use cases are linked to the actors to show their functions in the system.

3.5 Sequence Diagram

3.5.1 Admin

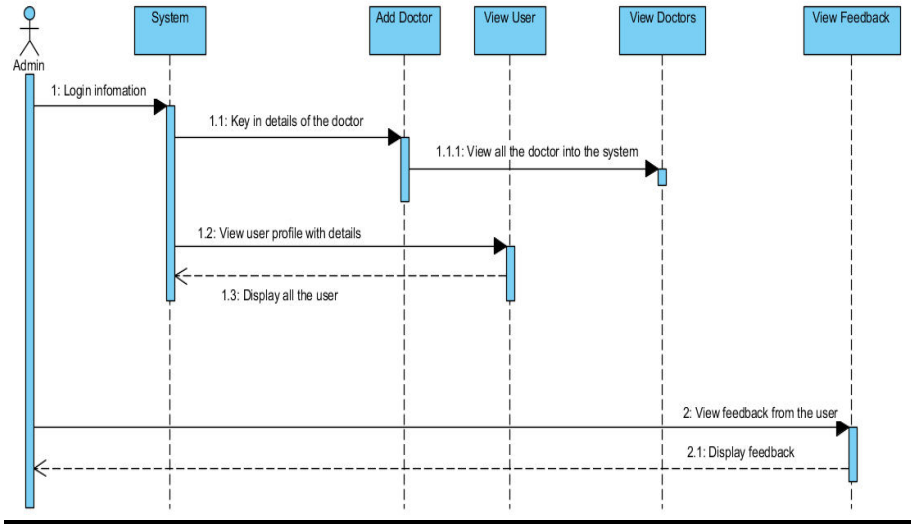


Figure 3.5.1: Sequence Diagram for Admin

As shown in the Figure 3.5.1 above, the admin need to login information to the system. In the system, the admin can key in details of the doctor. After adding the doctors into the system, Admin can view all the doctors details. Admin can view all the user profile with details in the system and can view feedback from the user.

3.5.2 Doctor

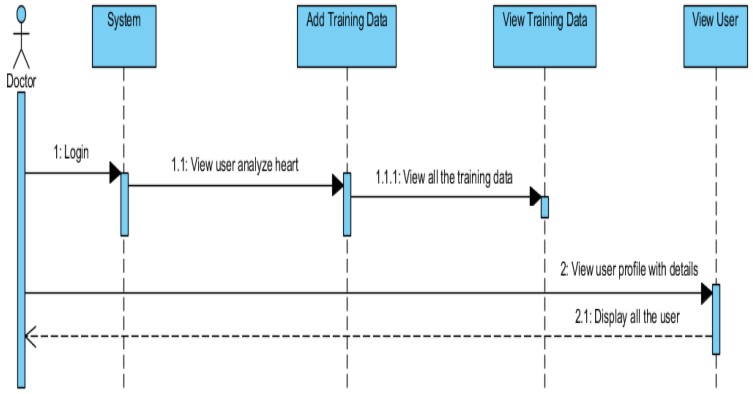


Figure 3.5.2: Sequence Diagram for Doctor

As shown in the Figure 3.5.2 above, the doctor need to provide credentials to the system. The doctor can add training data that provided by the user where it display the result of the user after key in the details. The doctor can view training data where doctor can view all the details of analyse heart details and result of disease of the user. The doctor can view all the user with details.

3.5.3 User

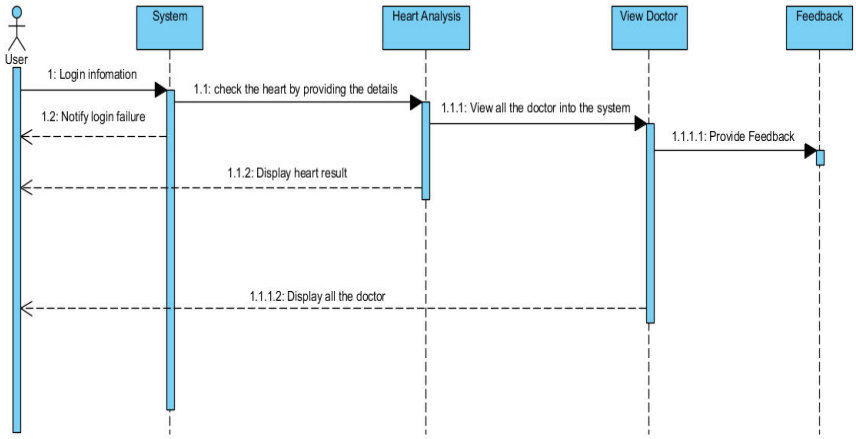


Figure 3.5.3: Sequence Diagram for User

As shown in the Figure 3.5.3 above, the registered user has to insert the login details in order to get access to the system. If invalid details are inserted, the system will prompt a notification where require the user to login the details. The unregistered user has to register for a new account in order to access to the system. The registered user can key in the details for heart analysis where the heart analysis will provide the result for the user which heart disease type he or she having. Then user can view all the doctor with details. The user can choose the doctor where the heart disease type result is matching with the doctor that specialize in that area of the heart. The user can provide feedback whether they satisfy with the doctor or they need better doctor.

3.6 Entity Relationship Diagram

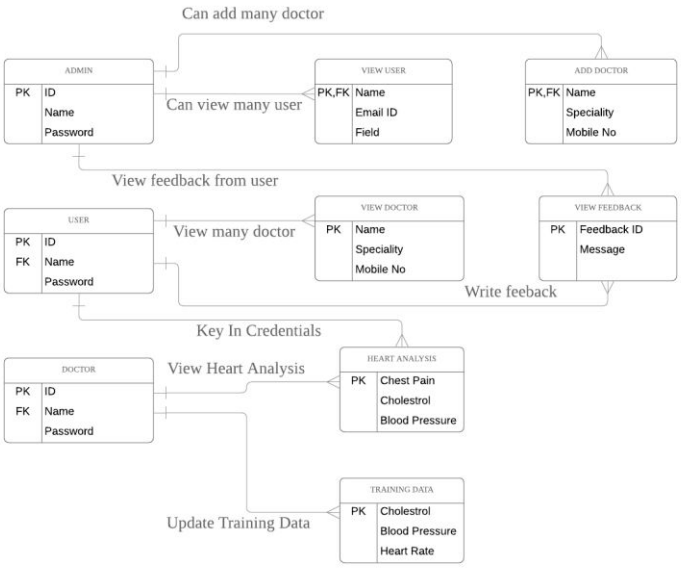


Figure 3.6 : Entity Relationship Diagram For Web-based System For The Prediction Of Heart Disease

As shown in the Figure 3.6, there are 9 tables in this diagram including 'Admin','User','Doctor','View_User','Add_Doctor','Heart_Analysis','View_Docto r', 'Training_Data', 'View_Feedback'. The Admin to view user, add doctor, view doctor and view feedback are one-to-many. The User to heart analysis, view doctor and feedback are one to one. The Doctor to view heart analysis (add training data), view training data and view user details are one to one.

3.7 Data Flow Diagram

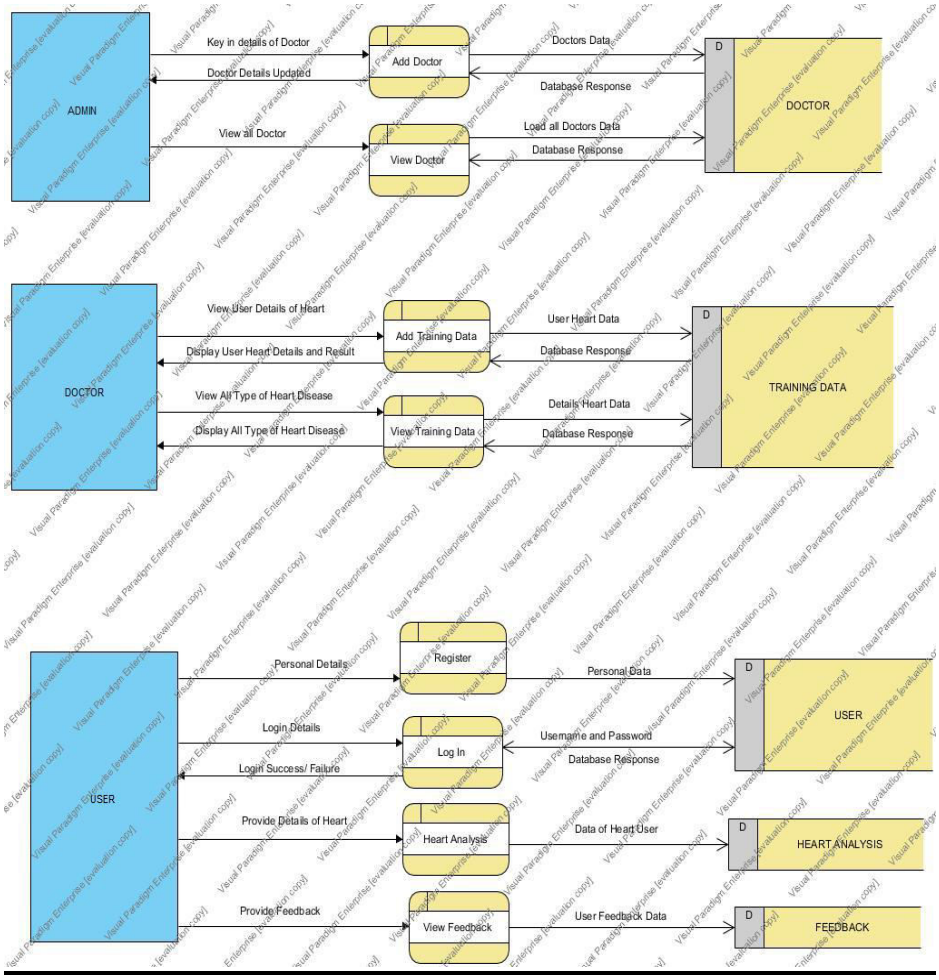


Figure 3.7 : Data Flow Diagram For Web-Based System Heart Disease Prediction

Above Figure 3.7, there are 3 external entities, 5 data stores and 8 processes.

4.0 RESEARCH METHODOLOGY

4.1 Model for Prediction

4.1.1 Collecting Data

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes. Data collection methods can be divided into two categories. The two categories are secondary methods of data collection and primary methods of data collection.

Secondary data is a type of data that has already been published in journals, books, online portals etc. The secondary data is based on the research papers that have been summarized. It is used in the study and plays an important role in terms of increasing the levels of research validity and reliability.

Primary data collection methods can be divided into two groups which are quantitative and qualitative.

Quantitative data collection methods are based on mathematical calculations in various formats. The Cleveland dataset is considered a Quantitative data collection methods. From there, R programming can calculate the methods of correlation and regression, mode, mean and median. Qualitative research methods do not involve numbers or mathematical calculations. Qualitative

research is closely associated with words, colours and other elements that are non-quantifiable. Qualitative studies aim to ensure a greater level of depth of understanding and qualitative data collection methods include observation, case studies etc.

4.1.2 Pre-Processing

Data Integration and Data Reduction pre-processing are applied to the data set. The reason for using the Data Integration is because the Cleveland data set have fewer data and it's not big enough for the data to fulfil the requirement. The current data set has less than 300 data. In order to fulfil the requirement, Data Integration pre-processing is needed to be used for the Cleveland data set. The Cleveland data set is combined with other data set which have the same attributes as the Cleveland data set to make data larger in the data set. In Data Integration, it also eliminates redundant tasks such as reconciliation and the same data entry into multiple systems. By using Data Integration, it leads to more accuracy where the bigger the data in the data set, the more accurate the result produced for the output. Besides that, it also expedites data cleaning and the reconciliation process and reduces the risk of data entry errors. Data Reduction pre-processing is also applied to the data set. The reason being, to reduce the number of the attribute or reduce the number of the data set to ensure that it produce the same or almost the same results.

4.1.3 Applying Different Techniques

Artificial Neural Networks are one of the main tools used in machine learning. The Neural Networks itself consists of many small units called Neurons. These Neurons are grouped into several layers. Neurons of one layer connect to the next layer through weighted connections. Neural Networks consist of input and output layers, as well as a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are too complex or numerous for a human programmer to extract and the machine to recognize.

Information flows through a neural network in two ways. When it's learning or operating normally, patterns of information are fed into the network via the input units, which trigger the layers of hidden units, and these in turn arrive at the output units. This common design is called a feed-forward network. Each unit receives inputs from the units to its left, and the inputs are multiplied by the weights of the connections they travel along. Every unit adds up all the inputs it receives in this way and if the sum is more than a certain threshold value, the units it’s connected to are triggered.

Neural networks learn things in exactly the same way, typically by a feedback process called back-propagation. This involves comparing the output network it produces with the output it was meant to produce, and by using the difference between them to modify the weights of the connections between

the units in the network and working form output units through the hidden units to the input units which means going backward. In time, back-propagation causes the network to learn, reducing the difference between actual and intended output to the point where the two exactly coincide, so the network figures things out exactly as it's should.

4.2 Developing Web-based System

4.2.1 Description of the modules

Admin

The admin needs to provide the credentials to access the system. There are four modules which are accessible by the admin. The first module is "add doctor" where admin can add a new doctor in the system. The admin need to key in the details of the doctor such as name, gender, address, age, phone number, email address and speciality of the doctor. The details of the doctor are being added to the system. The ID of the doctor is automatically generated for the next doctor. The second module is "view user" where admin can view user profile with the details. The third module is "view doctor" where admin can view all the doctor that set into the system. The fourth module is "view feedback" where the admin can view feedback from the user.

User

If the user is new to the system, the user needs to register first by providing the user details. The ID of the user is automatically generated. From that ID, the user can log in to the system after the user provides the credentials. There are three modules that are accessible by the user. The first module is "heart

analysis" where the user can check his or her heart by providing the details. After providing all the details, the user can click analyse heart to predict the disease of the heart. The result has been described when the user key in all the details. The result will show what type of heart disease the user had. The second module is "view doctor" where the user can view all the doctor into the system and contact the doctor accordingly. The third module is "feedback" where the user can provide or submit to the administrator.

Doctor

The doctor needs to provide the credentials to access the system. There are three modules which are accessible by the doctor. The first module is "add training data". Training data means the data which is specified in the system to analyse further reason. Based on the training data, the user analyses his or her heart and get an accurate result from the system. The second module is "view training data" where the doctor can view all the training data that add to the system. The training data contains a lot of details of the heart such as chest pain type, resting blood sugar and etc. Based on the training data, the user analyses the heart and get the accurate based on patients details. The third module is "view user" where the doctor can view user profile with details.

4.2.2 Block Definition Diagram

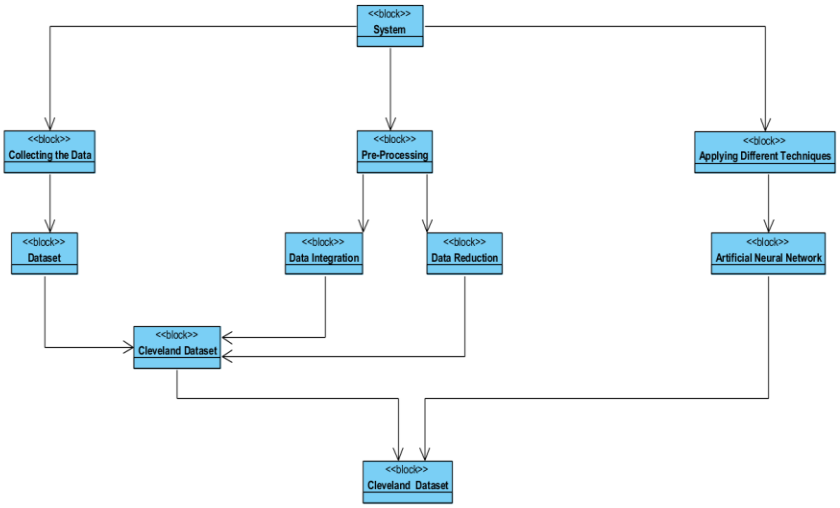


Figure 4.2.2 : Block Definition Diagram For Model Prediction

As shown in the figure above, the Block Definition Diagram for Model Prediction. There are three model prediction collecting the data, pre-processing and applying different techniques. The collecting the data is where get the dataset from the source. The pre-processing that be using are Data Integration and Data Reduction and from there the pre-processing method apply to the Cleveland dataset. From applying the different techniques, Artificial Neural Network is the chosen classification from data mining technique. After done the pre-processing, the Artificial Neural Network will be apply to the Cleveland Dataset by using R-Studio tool to run it.

5.0 IMPLEMENTATION

This chapter describes what implementation and tools that be doing for this semester.

5.1 Development Tools

5.1.1 R

R is an high level programming language and powerful programming language. R programming language easy to read, easy to learn and easy to code. Besides that, R has library for many things, use quickly build, the lower the performance and often less powerful prototype. The R is also great for validating ideas for so many different projects. R can integrate with Web-based System which will be implement for this semester. Besides that, R can be use to apply data mining techniques. For example, Artificial Neural Network will be use in R programming language.

5.1.2 HTML

HTML is a standard mark-up language for creating web application. Cascading Style Sheets (CSS) and JavaScript it forms a triad of cornerstone technologies for the World Wide Web. The system is using Cascading Style Sheets (CSS) to create desired appearance of the Graphical User Interface.

5.2 User Interface

5.2.1.1 Overall Modules User Interface

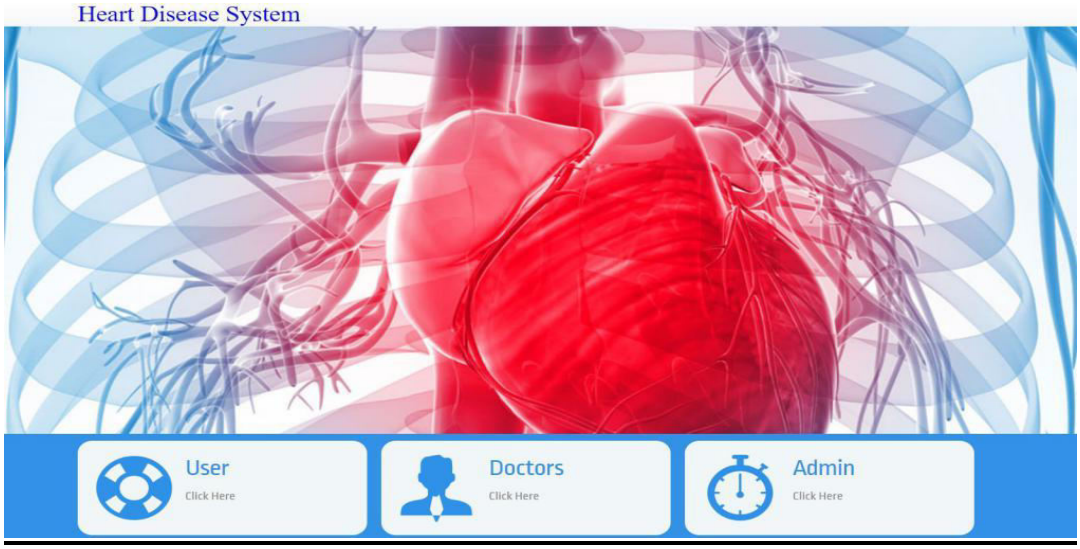


Figure 5.2.1.1 : Overall Modules Interface

As shown in the figure above, the overall modules for the system. There are three modules which are Admin, Doctor and User.

5.2.1.2 Admin

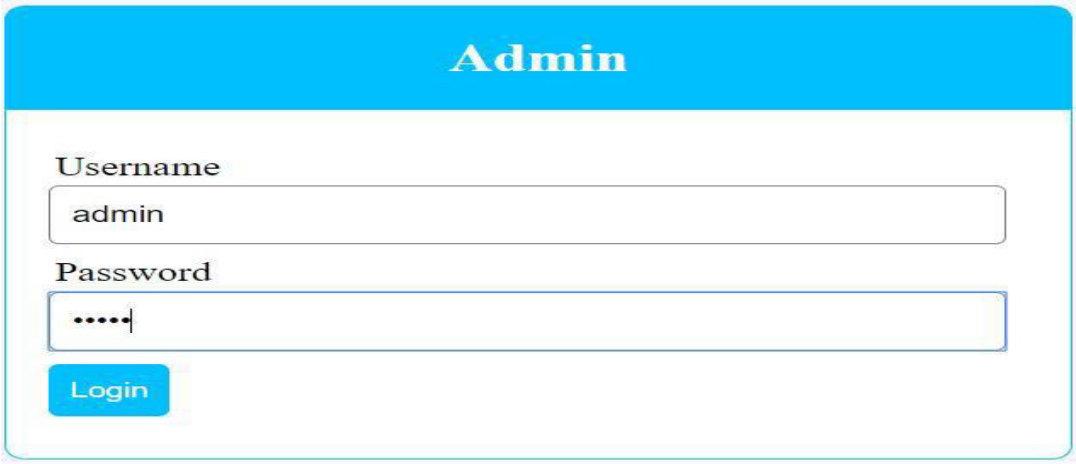


Figure 5.2.1.2 : Admin Login Interface

As shown in the figure above, the admin need to provide credentials to access the system.

5.2.1.3 Add Doctor Interface

Figure 5.2.1.3 : Add Doctor Interface

As shown in the figure above, the admin need to key in the details of the doctor to add into the system. The doctor ID automatically generated when submit the doctor details into the system.

5.2.1.4 View User Interface

Figure 5.2.1.4 : View User Interface

As shown in the figure above, the admin can view all the user details.

5.2.1.5 View Doctor Interface

Figure 5.2.1.5 : View Doctor Interface

As shown in the figure above, the admin can view all the doctor after submit the doctor details into the system.

5.2.1.6 Feedback User Interface

Figure 5.2.1.6 : Feedback User Interface

As shown in the figure above, the admin can view all the feedback from user where user provide feedback.

5.2.1.7 Doctor

Doctor Login Interface

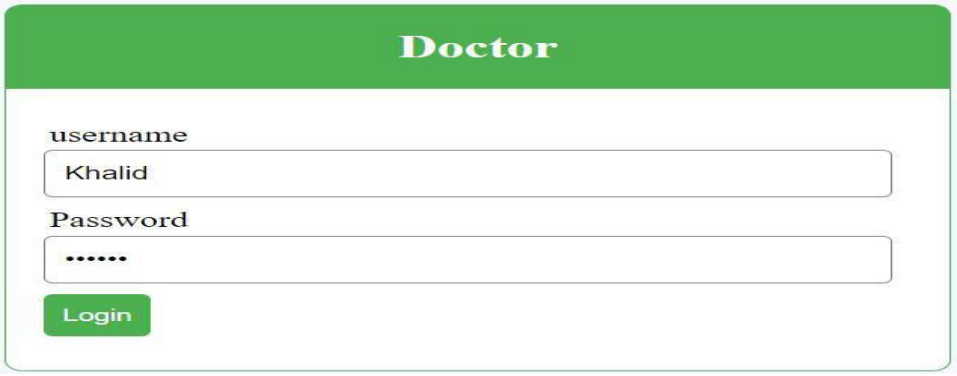


Figure 5.2.1.7 : Doctor Login Interface

As shown in the figure above, the doctor need to key in their ID and password in order to access to the system.

5.2.1.8 Doctor Add Training Data User Interface

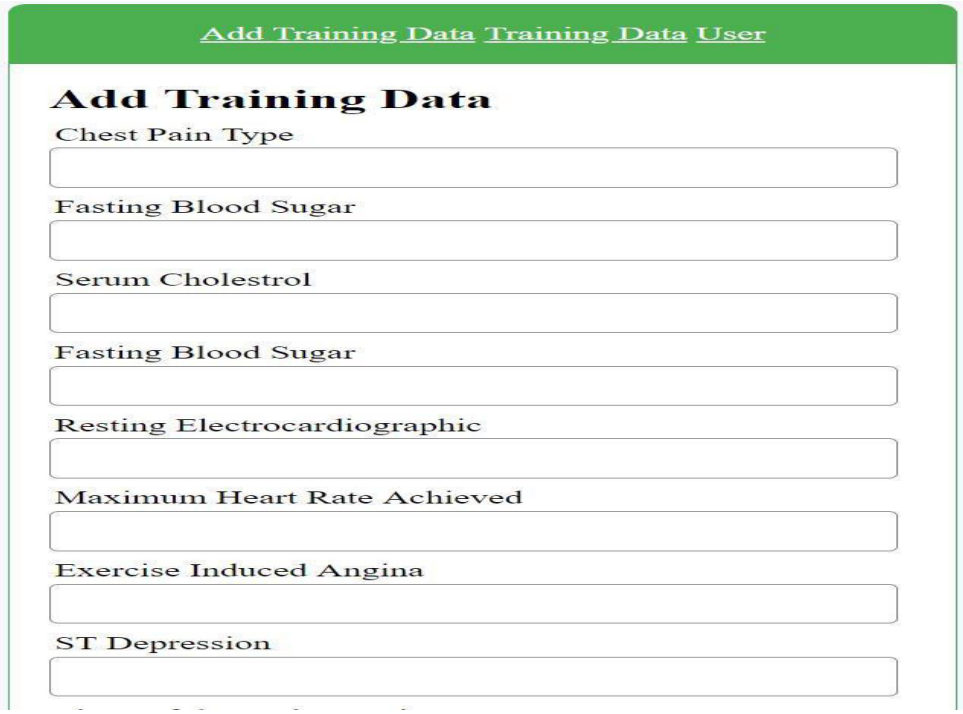


Figure 5.2.1.8 : Doctor Add Training Data Interface

As shown in the figure above, the doctor can add training data where the user provide their details of the heart.

5.2.1.9 Doctor View Training Data Interface

Figure 5.1.2.9 : Doctor View Training Data Interface

As shown in the figure above, the doctor can view all the training data where the doctor can view how many got stroke disease, at which age having the most heart disease, which gender have heart disease the most and etc.

5.2.2.0 Doctor View User Interface

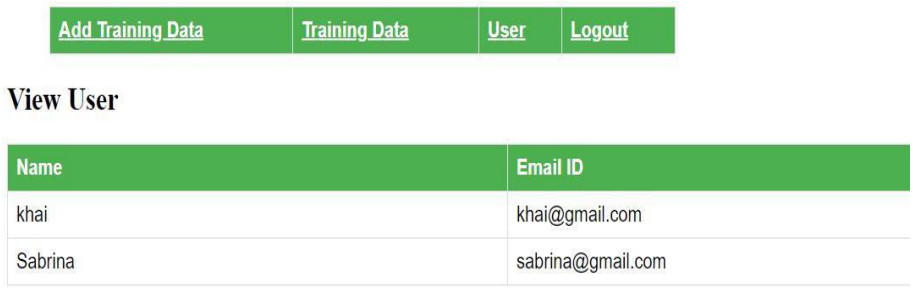


Figure 4.3.3.3 : Doctor View User Interface

As shown in the figure above, the doctor can view the user details where the user choose the specific doctor for their heart disease.

5.2.2.1 User

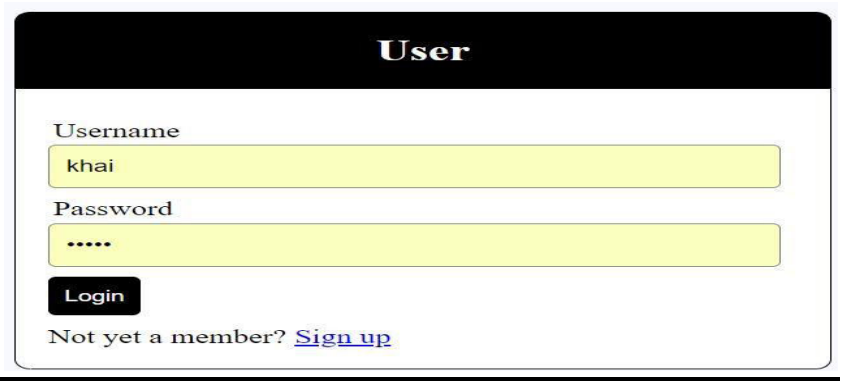


Figure 5.2.2.1 : User Interface Login

As shown in the figure above, User need to login the username and password in order to access to the system. If the user are new to the system, the user can sign up before access to the system.

5.2.2.2 Heart Analysis User Interface

Figure 5.2.2.2 : Heart Analysis User Interface

As shown in the figure above, the user need to key in all the details of the user heart. Once the user key in all the details, then user click the analyse heart button and the system will show the result of what type of heart disease the user having.

5.2.2.3 View Doctor Interface

Figure 5.2.2.3 : View Doctor Interface

As shown in the figure above, the user can view all the doctors after the user know what type of heart disease he or she had. User can view all the doctors with details. The user can choose any doctor with specialty which related to their heart disease type.

5.2.2.4 Provide Feedback

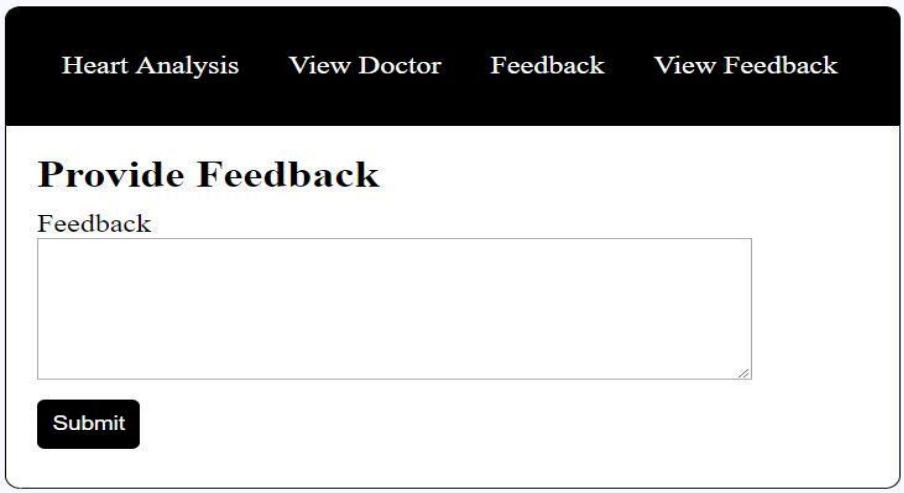


Figure 5.2.2.4 : Provide Feedback

As shown in the figure above, the user can provide feedback to the system. The user can provide feedback whether he or she satisfy with the doctor or need more quality doctor.

5.2.2.5 View Feedback

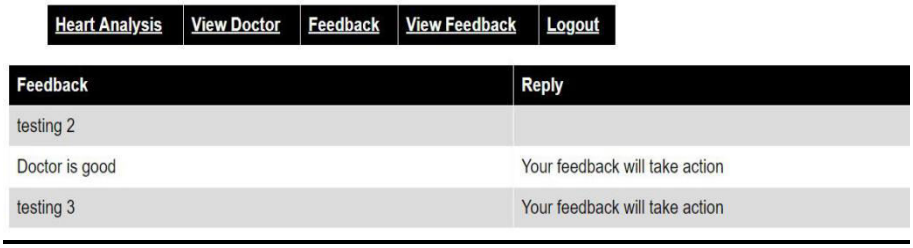


Figure 5.2.2.5 : View Feedback

As shown in the figure above, the user can view feedback to the system. The user can view reply feedback form from the admin whether is there any action be taken or not.

5.3 Pseudo-code

In this Pseudocode, I be explaining how the code works and function of using that code in R Studio. There are 2 models that being implemented using that code.

5.3.1.0 Calling the file

```
#Calling the CSV File
setwd("C:/Users/AARON/Documents/FYP Dataset")
health <- read.csv("Health1234.csv")
attach(health)
```

Figure 5.3.1.0 : Calling the file

The figure above shows the data set for the first model being called. First be set the working directory where it shows the setwd where the file of the data is located. Then it will read the file to call the data. Finally the data is attach in the R Studio.

5.3.1.1 Training and Testing Data

```
#Training and Testing Data for Validation
library(caret)
set.seed(10)
inTrainRows <- createDataPartition(health$grp,p=0.9,list=FALSE)
trainData <- health[inTrainRows,]
testData <-  health[-inTrainRows,]
nrow(trainData)/(nrow(testData)+nrow(trainData)) #checking whether really 90% -> OK
```

Figure 5.3.1.1 : Training and Testing Data

The figure above shows Training and Testing Data for the Validation. The library caret being used in this programming. The library function of caret is whether to classify or to do regression. The set seed function is to generate any random number to produce it. The create data partition is to divide into two one for train and one for testing by using the probability of 0.9. The train data will have more data compare to the test data. The last line of the code to check whether the data is already divided by using the probability of 0.9 to ensure the train Data will have more data.

5.3.1.2 Neural Network for the model 1

```
#NeuralNetwork for health dataset
library(neuralnet)
nn <-
  neuralnet(
    grp ~ age + trestbps + chol + thatach + oldpeak + ca + gender +
      cp + fbs + restecg + exang + slope + thal + diag,
    data = trainData,
    hidden = c(8, 2),
    linear.output = FALSE,
    threshold = 0.01
  )
plot(nn)
nn$result.matrix
nn$weights
```

Figure 5.3.1.2 : Neural Network for the model 1

The figure above shows the generate the Artificial Neural Network by using the code above. In order to run the Neural Network, the library neuralnet need to be use or to run it. The library neuralnet have its package whether type of Neural Network is Multilayer Perceptron and the algorithm inside the neuralnet is backpropagation. The parameters have been use above to generate it. The group having been classify and other fourteen variables being the input. The train data is being used to implement it. The number of hidden means how many hidden layers to generate. Set 8 and 2 hidden layers to classify the output. The threshold parameters being used where it can run when the data is in numeric or vector. After all the parameters is set and run the code where it will set it. Plot the neural network where it will generate Artificial Neural Network. Then it will show the result matrix and the weights when run the code.

5.3.1.3 Testing Result Output

```
#Testing Result Output
temp_test <-
  subset(
    trainData,
    select = c(
      "age",
      "trestbps",
      "chol",
      "thatach",
      "oldpeak",
      "ca",
      "gender",
      "cp",
      "fbs",
      "restecg",
      "exang",
      "slope",
      "thal",
      "diag"
    )
  )
head(temp_test)
nn.results <- compute(nn, temp_test)
```

```
results <- data.frame(actual = trainData$grp, prediction = nn.results$net.result)
results
```

Figure 5.3.1.3 : Testing Result Output

The above codes shows that testing the result output where set the subset. It will generate the actual and prediction value where it shows the results. The first for the actual is set for the type of heart disease is whether 0 or 1. The second for the prediction is set for the neural network result where it generate the Artificial Neural Network.

5.3.1.4 Round off the Prediction

```
#Roundoff the prediction to 0 or 1
results$prediction <- round(results$prediction)
results <- data.frame(actual = results$actual, prediction = results$actual)
results
```

Figure 5.3.1.4 : Round off the Prediction

The above codes shows where it will round-off the prediction value. The prediction value before round-off was almost the same with the actual. In order to round-off the prediction it will get equal value with the actual.

5.3.1.5 Confusion Matrix

```
#Confusion Matrix
roundedresults<-sapply(results,round,digits=0)
roundedresultsdf=data.frame(roundedresults)
attach(roundedresultsdf)
conf_matrix <-table(results$actual,results$prediction)
conf_matrix
```

Figure 5.3.1.5 : Confusion Matrix for model 1

The above codes shows the confusion matrix. The roundedresults means that the prediction is equal with the actual set the digits with zero so that it will generate the confusion matrix starting from zero. The function attach being use where it will use the roundedresult of the actual and prediction value. In order to run the confusion matrix have to use conf matrix where it will generate the matrix of 0 and 1.

5.3.1.6 Accuracy

```
#ACCURACY
n = sum(conf_matrix) # number of instances
nc = nrow(conf_matrix) # number of classes
diag = diag(conf_matrix) # number of correctly classified instances per class
rowsums = apply(conf_matrix, 1, sum) # number of instances per class
colsums = apply(conf_matrix, 2, sum) # number of predictions per class
p = rowsums / n # distribution of instances over the actual classes
q = colsums / n # distribution of instances over the predicted classes
accuracy = sum(diag) / n
accuracy
```

Figure 5.3.1.6 : Accuracy

The above codes shows how to run the accuracy in order to get the best accuracy. Set the function for the accuracy and once set it will give the accuracy value.

5.3.1.7 Code for plotting

```
ggplot(trainData, aes(x = grp, y = prediction)) +
  geom_line(color="red")+
  geom_point(color = "blue")
```

Figure 5.3.1.7 : Code for plotting

The above codes shows how to run the plot to generate the graph. The 2nd code shows the line that going plot for the straight line and set the color. The 3rd code shows the point in the graph and can set the color. The function of this to generate complex graph.

5.3.1.8 Sensitivity and Specificity

```
#Call library Caret
library(caret)

#Sensitivity/True Positive Rate
sensitivity(conf_matrix)

#Specificity/True Negative Rate
specificity(conf_matrix)
```

Figure 5.3.1.8 : Sensitivity and Specificity

The above codes shows how to run the sensitivity and specificity. In order to run the code, call the library where it can use to classify or to regression. Once run it, code for sensitivity and specificity can get the results.

5.3.1.9 Algorithm for Fast and Frugal Trees

```
#Algorithm Fast & Frugal Trees
#Decision Tree
#Testing Train&Test
library(FFTrees)
FFTrees.guide()

health<- FFTrees(formula = grp ~ .,                                  # Criterion and (all) predictors
                 data = trainData,                                   # Training data
                 data.test = testData,                               # Testing data
                 main = "Heart Disease",                             # General label
                 decision.labels = c("Healthy", "Have Heart Disease"))  # Labels for decisions
health
summary(health)
```

Figure 5.3.1.9 : Algorithm for Fast and Frugal Trees

The above codes shows the algorithm Fast and Frugal trees. Run the library of Fast and Frugal trees in order to visualize and evaluate. The data being use is from train data and the main and decision labels function is to label it for the title.

5.3.2.0 Fast and Frugal Trees Plotting

```
# Plot the best FFT when applied to the test data
plot(health, data= "train")

# Plot only the tree without accuracy statistics
plot(health, stats = FALSE)

plot(health, what = "cues")
```

Figure 5.3.2.0 : Fast and Frugal Trees Plotting

The above codes shows the plot of Fast and Frugal.

5.3.2.1 Calling the CSV File for 2nd Model

```
#Calling the CSV File
setwd("C:/Users/AARON/Documents/FYP Dataset")
nhealth <- read.csv("NotHealthy1234.csv")
attach(nhealth)
```

Figure 5.3.2.1 : Calling the CSV File for 2nd Model

The figure above shows the data set for the second model being called. First be set the working directory where it shows the setwd where the file of the data is located. Then it will read the file to call the data. Finally the data is attach in the R Studio.

5.3.2.2 Splitting Training and Testing Data

```
#Training and Testing Data for Validation
library(caret)
set.seed(10)
inTrainRows <- createDataPartition(nhealth$grp,p=0.7,list=FALSE)
trainData <- nhealth[inTrainRows,]
testData <-  nhealth[-inTrainRows,]
nrow(trainData)/(nrow(testData)+nrow(trainData)) #checking whether really 70% -> OK
```

Figure 5.3.2.2 : Splitting Training and Testing Data

The figure above shows Training and Testing Data for the Validation. The library caret being used in this programming. The library function of caret is whether to classify or to do regression. The set seed function is to generate any random number to produce it. The create data partition is to divide into two one for train and one for testing by using the probability of 0.7. The train data will have more data compare to the test data. The last line of the code to check whether the data is already divided by using the probability of 0.7 to ensure the train Data will have more data.

5.3.2.3 Neural Network for 2nd Model

```
#NeuralNetwork
#Algorithm-> Backpropagation
#Hidden Layers->8,2
library(neuralnet)
trainData.nn <-
  neuralnet(
    grp ~ age + trestbps + chol + thatach + oldpeak + ca + gender +
      cp + fbs + restecg + exang + slope + thal + diag,
    data = trainData,
    hidden = c(8, 2),
    linear.output = FALSE,
    threshold = 0.01
  )
plot(trainData.nn)
trainData.nn$result.matrix
```

Figure 5.3.2.3 : Neural Network for 2nd Model

The figure above shows the generate the Artificial Neural Network by using the code above. In order to run the Neural Network, the library neuralnet need to be use or to run it. The library neuralnet have its package whether type of Neural Network is Multilayer Perceptron and the algorithm inside the neuralnet is backpropagation. The parameters have been use above to generate it. The group having been classify and other fourteen variables being the input. The train data is being used to implement it. The number of hidden means how many hidden layers to generate. Set 8 and 2 hidden layers to classify the output. The threshold parameters being used where it can run when the data is in numeric or vector. After all the parameters is set and run the code where it will set it. Plot the neural network where it will generate Artificial Neural Network. Then it will show the result matrix and the weights when run the code.

5.3.2.4 Testing Result Output for 2nd Model

```
#Testing Result Output
temp_test <-
   subset(
      trainData,
      select = c(
         "age",
         "trestbps",
         "chol",
         "thatach",
         "oldpeak",
         "ca",
         "gender",
         "cp",
         "fbs",
         "restecg",
         "exang",
         "slope",
         "thal",
         "diag"
      )
   )
head(temp_test)
```

```
trainData.nn.results <- compute(trainData.nn, temp_test)
results <- data.frame(actual = trainData$grp, prediction = trainData.nn.results$net.result)
results
```

Figure 5.3.2.4 : Testing Result Output for 2nd Model

The above codes shows that testing the result output where set the subset. It will generate the actual and prediction value where it shows the results. The first for the actual is set for the type of heart disease is whether 1,2,3 or 4. The second for the prediction is set for the neural network result where it generate the Artificial Neural Network.

5.3.2.5 Round off Prediction Value

```
#Roundoff the Prediction to 1,2,3,4
#For Heart Disease Prediction
results$prediction <- round(results$prediction)
results <- data.frame(actual = results$actual, prediction = results$actual)
results
```

Figure 5.3.2.5 : Round off Prediction Value

The above codes shows where it will round-off the prediction value. The prediction value before round-off was almost the same with the actual. In order to round-off the prediction it will get equal value with the actual.

5.3.2.6 Confusion Matrix

```
#Confusion Matrix
roundedresults<-sapply(results,round,digits=0)
roundedresultsdf=data.frame(roundedresults)
attach(roundedresultsdf)
conf_matrix <-table(results$actual,results$prediction)
conf_matrix
```

Figure 5.3.2.6 : Confusion Matrix

The above codes shows the confusion matrix. The roundedresults means that the prediction is equal with the actual set the digits with zero so that it will generate the confusion matrix starting from zero. The function attach being use where it will use the roundedresult of the actual and prediction value. In order to run the confusion matrix have to use conf matrix where it will generate the matrix of 1,2,3 and 4.

5.3.2.7 Accuracy

```
#ACCURACY
n = sum(conf_matrix) # number of instances
nc = nrow(conf_matrix) # number of classes
diag = diag(conf_matrix) # number of correctly classified instances per class
rowsums = apply(conf_matrix, 1, sum) # number of instances per class
colsums = apply(conf_matrix, 2, sum) # number of predictions per class
p = rowsums / n # distribution of instances over the actual classes
q = colsums / n # distribution of instances over the predicted classes
accuracy = sum(diag) / n
accuracy
```

Figure 5.3.2.7 : Accuracy

The above codes shows how to run the accuracy in order to get the best accuracy. Set the function for the accuracy and once set it will give the accuracy value.

5.3.2.8 Plotting the graph

```
ggplot(trainData, aes(x = grp, y = prediction)) +
  geom_line(color="green")+
  geom_point(color = "blue")
```

Figure 5.3.2.8 : Plotting the graph

The above codes shows how to run the plot to generate the graph. The 2nd code shows the line that going plot for the straight line and set the color. The 3rd code shows the point in the graph and can set the color. The function of this to generate complex graph.

5.3.2.9 Sensitivity and Specificity

```
#Call library Caret
library(caret)

#Sensitivity/True Positive Rate
sensitivity(conf_matrix)

#Specificity/True Negative Rate
specificity(conf_matrix)
```

Figure 5.3.2.9 : Sensitivity and Specificity

The above codes shows how to run the sensitivity and specificity. In order to run the code, call the library where it can use to classify or to regression. Once run it, code for sensitivity and specificity can get the results.

5.3.3.0 Pie Charts for plotting

```
#PieCharts Based on Actual vs Prediction Table
slices <- c(42, 20, 25, 10)
lbls <-
  c(
    "Arrhythmia",
    "Stroke",
    "Coronory Artery Disease",
    "Angina"
  )
pie(slices, labels = lbls, main="Pie Chart of Heart Disease Type")
```

Figure 5.3.3.0 : Pie Charts for plotting

The above codes how to run the pie charts based on the confusion matrix. The slices mean to put the value in order to generate the pie charts. The second line code is where each description being label according to the value given. The last line of the code is for the title when plot the pie charts. So it is easy to visualize based on the pie charts.

6.0 RESEARCH CONTRIBUTION

There are 2 model which use for the simulation. There are 15 variables. The inputs use are 14 variables and the output is 1. The first model use to predict whether the person having heart disease or not. The second model have 4 type of heart disease and it predict what type of heart disease the user having. The prediction for both models using Artificial Neural Network. Type of Neural Network use is the Multilayer Perceptron.In R Studio there is a library called neuralnet where inside neuralnet package contain backpropagation. The backpropagation function is to track the weight backtracking.

6.1 FIRST MODEL

The first model will predict the person whether having heart disease or not.

6.1.2 Artificial Neural Network

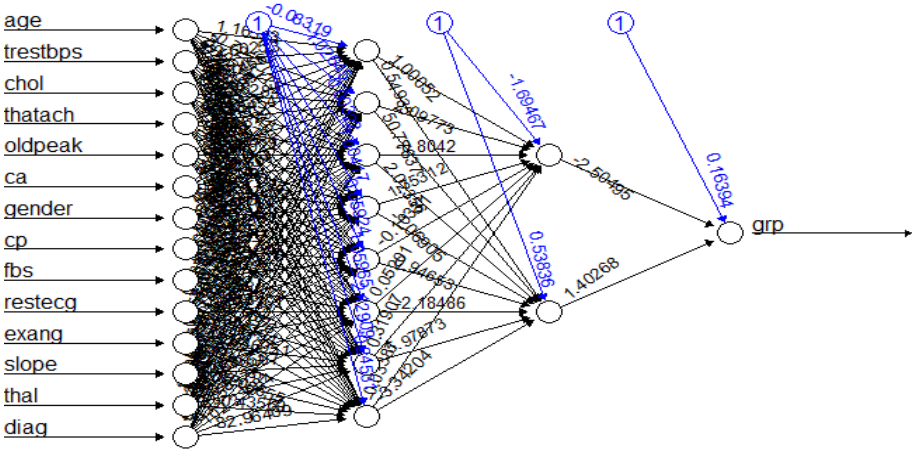


Figure 6.1.2 : Artificial Neural Network

The above diagram shows the Artificial Neural Network where there are 14 inputs with 8 and 2 for the hidden layers and 1 output either 0 or 1. The blue color represent the bias.

6.1.3 Actual vs Prediction

	actual	prediction
1	0	-0.00023985031659
3	1	0.99974962906097
4	0	-0.00023985031659
5	0	-0.00023985031659
6	0	-0.00023985031659
7	1	0.99975460301665
8	0	-0.00096650283480
9	1	0.99974962906104
10	1	0.99974963023189
11	0	-0.00023985030010
12	0	-0.00023985031659
13	1	0.99974966535594
14	0	0.00001089231167
15	0	-0.00023985031659
17	0	-0.00023985031659
18	1	0.94106530260306
19	1	0.99973822339653
20	1	0.99974962906097

	actual	prediction
1	0	0
2	1	1
3	0	0
4	0	0
5	0	0
6	1	1
7	0	0
8	1	1
9	1	1
10	0	0
11	0	0
12	1	1
13	0	0
14	0	0
15	0	0
16	1	1
17	1	1
18	1	1
19	0	0

Figure 6.1.3 : Actual vs Prediction

The above screenshot shows the actual and prediction value. The prediction value give almost the exact value compare to the actual. When round off the number of prediction it became exact value with the actual.

6.1.4 Confusion Matrix

	0	1
0	141	0
1	0	126

Figure 6.1.4 : Confusion Matrix

The above screenshot shows the confusion matrix. The zero represent the user having no heart disease and the one represent the user having heart disease.

6.1.5 Accuracy

```
[1] 1
```

Figure 6.1.5 : Accuracy

The above screenshot shows the accuracy for the 1st model where it shows the model have 100% accuracy.

6.1.6 Graph

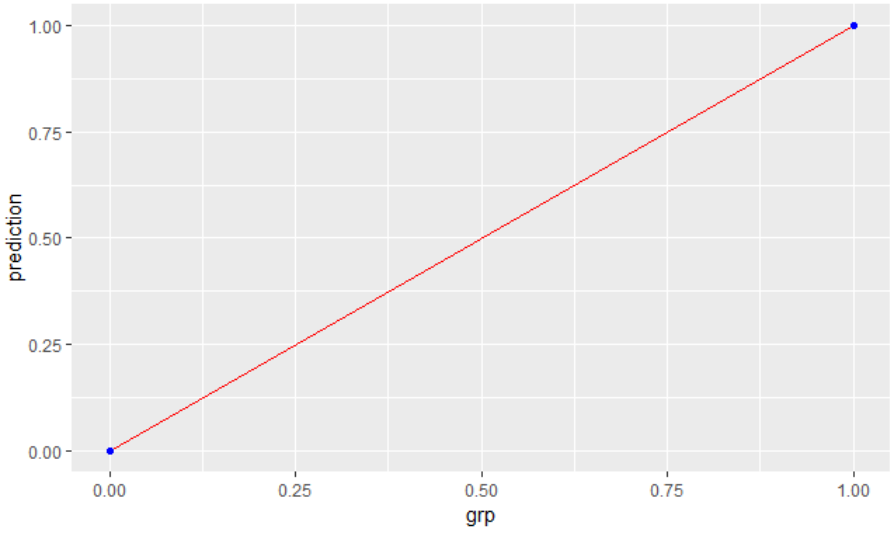


Figure 6.1.6 : Graph

The above screenshot shows the group and prediction graph for model 1 where it shows the group and the prediction get the exact value. When plotting the straight line it get nice straight line graph.

6.1.7 Sensitivity and Specificity

```
> #Sensitivity/True Positive Rate
> sensitivity(conf_matrix)
[1] 1
>
> #Specificity/True Negative Rate
> specificity(conf_matrix)
[1] 1
```

Figure 6.1.7 : Sensitivity and Specificity

The above screenshot shows the sensitivity and specificity of confusion matrix where both get the value of 1.

6.1.8 Fast and Frugal for Train

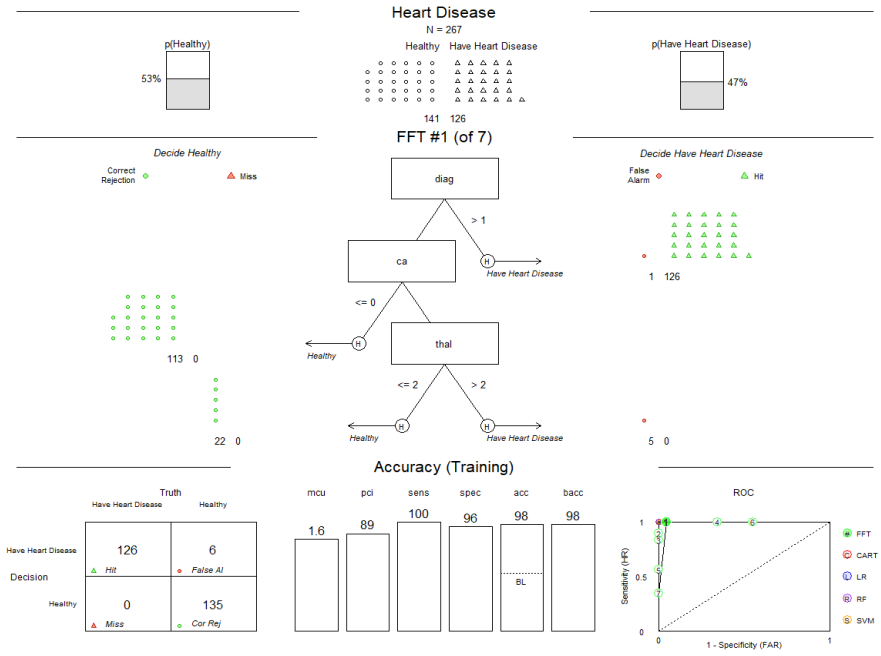


Figure 6.1.8 : Fast and Frugal for Train

The above screenshot shows the Fast and Frugal for Train where it shows the overall.

6.1.9 Fast and Frugal for Sensitivity and Specificity

Figure 6.1.9 : Fast and Frugal for Sensitivity and Specificity

The above graph shows the fast and frugal for the sensitivity and specificity where all the variables are the dots in the graph.

6.2 SECOND MODEL

The second model will predict what type of heart disease.

6.2.1 Artificial Neural Network

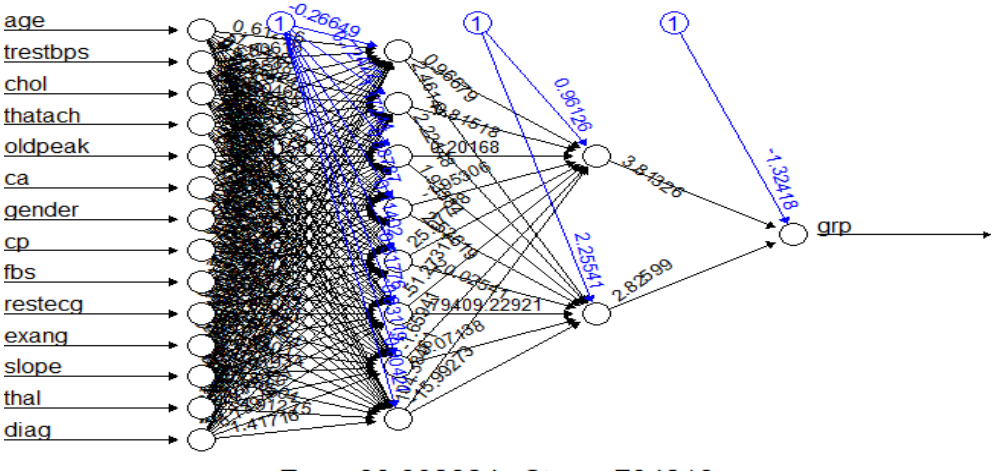


Figure 6.2.1 : Artificial Neural Network

The above diagram shows the Artificial Neural Network where there are 14 inputs with 8 and 2 for the hidden layers and 1 output either 1,2,3,4. The blue color represent the bias.

6.2.2 Actual vs Prediction

	actual	prediction
1	1	1.0035161827
4	1	2.0540866814
5	1	2.0540866814
6	1	2.0540866814
7	1	2.0540866814
8	1	2.0540866814
9	1	2.0540866814
10	1	0.9981272612
11	1	2.0540864949
12	1	0.9987785434
14	1	2.0540866753
15	1	2.0540866814
17	1	2.0540866814
19	1	2.0540866814
20	1	2.0540866814

35	1	1
36	1	1
37	1	1
38	1	1
39	1	1
40	1	1
41	1	1
42	1	1
43	2	2
44	2	2
45	2	2
46	2	2
47	2	2
48	2	2
49	2	2
50	2	2
51	2	2
52	2	2

Figure 6.2.2 : Actual vs Prediction

The above screenshot shows the actual and prediction value. The some prediction value give almost the exact value compare to the actual. When round off the number of prediction it became exact value with the actual.

6.2.3 Confusion Matrix

	1	2	3	4
1	42	0	0	0
2	0	20	0	0
3	0	0	25	0
4	0	0	0	10

Figure 6.2.3 : Confusion Matrix

The above screenshot shows the confusion matrix. The one represent the user having Arrhythmia and the two represent the user having Stroke. The three represent the user having Coronary Artery Disease. The four represent the user having Angina.

6.2.4 Accuracy

```
[1] 1
```

Figure 6.2.4 : Accuracy

The above screenshot shows the accuracy for the 2nd model where it shows the model have 100% accuracy.

6.2.5 Graph

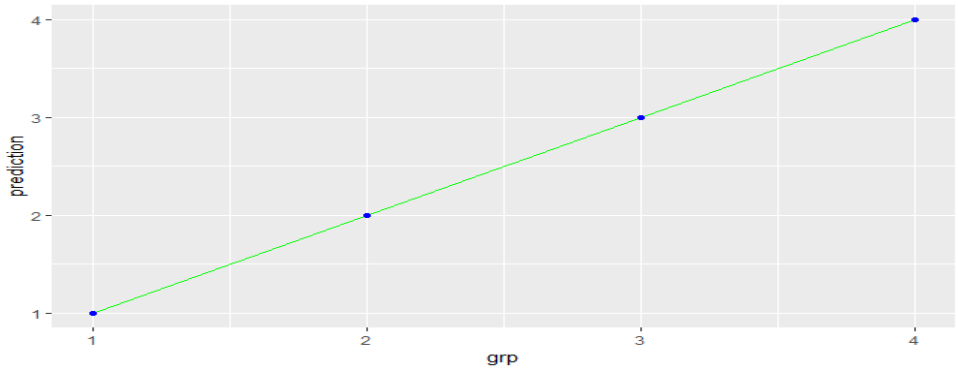


Figure 6.2.5 : Graph

The above screenshot shows the group and prediction graph for model 2 where it shows the group and the prediction get the exact value.

6.2.6 Sensitivity and Specificity

```
> #Sensitivity/True Positive Rate
> sensitivity(conf_matrix)
[1] NA
>
> #Specificity/True Negative Rate
> specificity(conf_matrix)
[1] NA
```

Figure 6.2.6 : Sensitivity and Specificity

The above screenshot shows the sensitivity and specificity of confusion matrix where both get the value of null.

6.2.7 Pie Charts

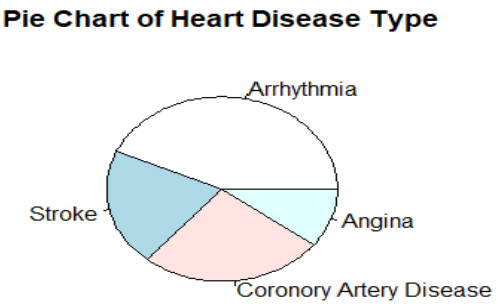


Figure 6.2.7 : Pie Charts

The above graphs shows the pie charts where there are four type of heart disease. Arrhythmia have the highest disease follow by Coronory Artery Disease follow by Stroke and lastly Angina.

7.0 CONCLUSION

This project is about prediction of heart disease by using Artificial Neural Network in order to achieve the highest accuracy for the prediction. Other than that, this project also main is the web-based system for prediction of heart disease to make the user to check their heart whether they have heart disease or not.

In order to achieve the goal, there are several activities that have been conducted. Firstly, the literature review has been carried out to better understand the prediction of heart disease by using various data mining techniques and also studied about the existing system. There are three existing system has been reviewed to identify functionalities, strength and weakness. In summaries, one of the existing system does not provide the details of the doctors, two existing system does not provide the login for the user or patient, and all the system have donation for the user or patient who having heart disease.

After the completion of literature reviews, the theoretical framework has been identified. The theoretical framework includes register, login, view doctors, view users, view feedback, add training data, view training data, heart analysis, add doctors and logout.

REFERENCES

Dr.G.Rasitha Banu & J.H. Bousal Jamala. (2015). Predicting Heart Attack using Fuzzy C Means Clustering Algorithm. Retrieved from International Journal of Latest Trend in Engineering and Technology.

G.Purusothaman & P.Krishnakumari. (2015). A Survey of Data Mining Techniques on Risk Prediction : Heart Disease. Retrieved from Indian Journal of Science and Technology.

K.Sudhakar & Dr.M.Manimekalai. (2014). Study of Heart Disease Prediction using Data Mining. Retrieved from International Journal of Advanced Research in Computer Science and Software Engineering.

M.Akhil Jabbar & Dr.B.L Deekshatulu. (2012). Heart Disease Prediction System using Associative Classification and Genetic Algorithm. Retrieved from International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies.

Mrs.G.Subbalakshmi, Mr.K.Ramesh, & Mr.M.Chinna Rao. (2011). Decision Support in Heart Disease Prediction System using Naïve Bayes. Retrieved from Indian Journal of Computer Science and Engineering.

Nidhi Bhatla & Kiran Jyoti. (2012). An Analysis of Heart Disease Prediction using Different Data Mining Techniques. Retrieved from International Journal of Engineering Research and Technology.

R.Chitra & V.Seenivasagam. (2013). Review Of Heart Disease Prediction System using Data Mining and Hybrid Intelligent Techniques. Retrieved from International Conference on Technology Journal on Soft Computing.

Sameh Ghwanmeh. (2012). Applying Advanced NN-based Decision Support Scheme for Heart Diseases Diagnosis. Retrieved from International Journal of Computer Application.

Shadab Adam Pattekari & Asma Parveen. (2012). Prediction System For Heart Disease using Naïve Bayes. Retrieved from International Journal of Advanced Computer and Mathematical Sciences.

APPENDICES

APPENDIX A:
PROPOSAL FOR DRAFT RESEARCH PAPER

Web-Based System For The Prediction Of Heart Disease

Aaron Raj A/L Maya
Faculty of Computing and Informatics
Multimedia University
63100 Cyberjaya, Selangor, Malaysia
aa.raj1434@gmail.com

Ramakrishnan Kannan
Faculty of Computing and Informatics
Multimedia University
63100 Cyberjaya, Selangor, Malaysia
kannan.ramakrishnan@mmu.edu.my

I. Introduction

Heart disease is a term that is assigned to refer to a large number of medical conditions that is related to the heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts. Heart disease is a major health problem in today's time. Prediction of heart disease using data mining is one of the most interesting and challenging tasks. The high wrongly diagnosed cases need to develop a fast and efficient prediction of heart disease system. The main objective is to identify the features from the data set by using classification model. The attributes that are more relevant to prediction of heart disease can be observed. This will help to understand the root causes of disease in depth.This project aims to identify:

-The optimal Artificial Neural Network model for the prediction of heart disease

-Develop a web-based system.

Here the scope of this research is to find the accuracy of heart disease of the user by using data mining technique. The main reason use data mining technique for the prediction is to see how the algorithm works when the user key in their details and what is the output of it to be.

II. Literature Review

The literature review of this project presented is about the review of different types of general papers where various researchers used various data mining techniques and review background systems to produce data for heart disease predictions.

Jabbar Akhil and Bulusu Deekshatulu proposed the APRIORI algorithm. They used it for discovering the rules for the algorithm and based on their research it requires multiple passes over the database in order to determine prediction of heart disease. The dataset they used is from in-house. The advantage of using this algorithm is a more efficient association classification heart disease prediction. Researcher used Gini Index which produces compact rule set and filter by applying Z-statistics and genetic algorithm to predict the accuracy of the heart disease. The main motivation for using a genetic algorithm in the discovery of high-level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interesting values. Based on the result, it shows that most of the classifier rules help in the best prediction of heart disease which even helps doctors in their diagnosis decision.

The researcher has done various data mining technique in order to get the best accuracy of the prediction of heart disease. Some researchers use an additional algorithm to combine the data mining algorithm that they used to get the exact accuracy. Besides, some of the researchers reduce the number of attributes in the dataset. By reducing the number of attributes, it will increase the efficiency and also less time is taken for it to produce good accuracy for the prediction. Most of the researcher didn't develop web-based systems for heart disease prediction. Developing web-based systems

will be useful for the people who suffer from heart disease because it helps to save lives.

III. Theoretical Framework

The data set that be using for this project is based on the Cleveland data set which got it from the Ethernet. In the Cleveland data set, there are 15 attributes which are age, gender, cp, trestbps, cholesterol, fbs, restecg, thatach, exang, oldpeak, slope, ca, thal, diag, and grp. 14 variables from the data set be use for the input and the 1 variable will be classify as the output.

Pre-processing method being use in the R Studio tool. The 15 variables are convert from nominal to numeric in order to generate the Artificial Neural Network Diagram. The data set split into two, one for train data and one for test data. The 90% is for the train data and 10% is for the test data. Train data being used to generate Artificial Neural Network with function of threshold, linear output and hidden layers.The reason choose Neural Network algorithm is that it has the ability to learn and model non-linear and complex relationships, which is really important because in real life, many of the relationships between inputs and outputs are non-linear as well as complex. Unlike many other prediction techniques, Neural Network does not impose any restrictions on the input variables. Additionally, many studies have shown that Neural Network can be better model for heteroskedasticity. For example, data with high volatility and non-constant variance, given its ability to learn hidden relationships in the data without imposing any fixed relationships in the data.

A. Limitation

- Dr.G. Rasitha Banu and J.H. Bousal Jamala are those researchers who used VISUAL BASIC.NET for the heart disease prediction. Visual basic is a proprietary programming language written by Microsoft. The programs written in Visual Basic cannot easily be transferred to other operating systems. Besides, there are some fairly minor disadvantages compared with C. C has a better declaration of arrays and it is possible to initialize an array of structures in C at declaration time. This is impossible to do in Visual Basic. Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Nidhi Bhatla and Kiran Jyoti are those researchers who used WEKA for the heart disease prediction. The limitation of using WEKA is that it only can handle small data sets. If the data sets are too big and there only a few megabytes of memory available, an Out Of Memory error occurs. Another limitation is that one is to copy the entire data set to another file and use the filename as the entry in that history table. What will happen is that it will be very slow because it has to copy the entire data set every time when executing an algorithm and it would be very space consuming for secondary storage.
- Sameh Ghwanmeh, Nidhi Bhatla and Kiran Jyoti are those researcher who used MATLAB for the heart disease prediction. MATLAB is great for prototyping and investigating data but it is just an awful language for building a complete application. The language was designed around small scripts to do two-dimensional matrix math and everything else is a bolt-on. It leads to an astounding number of gotchas that can lead to real bugs and limitations. The limitation is when it takes much CPU time for computation. It makes real time application very complicated and confuse the programmer.

IV. Research Methodology

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes. Data collection methods can be divided into two categories. The two categories are secondary methods of data collection and primary methods of data collection. Secondary data is a type of data that has already been published in journals, books, online portals etc. The secondary data is based on the research papers that I have summarized. It is used in the study and plays an important role in terms of increasing the levels of research validity and reliability. Primary data collection methods can be divided into two groups which are quantitative and qualitative.

A. Pre-Processing

Pre-Processing will be use into the data set by using R Studio programming language. All the data be changing from nominal to numeric. The data variables consist of factors, nominal and integer. In order to built the Artificial Neural Network, the variables must be in numeric in order to generate the graph of Artificial Neural Network. 300 data is being used and then split into two. One for training and another for testing. The splitting part is 90% 10% which means training data is 90% and test data is 10%. The reason why using 90% 10% is list as below:

1) To find whether the prediction value is equal to the actual value

2) Whether the accuracy will be more higher if

a) Round off the value

b) Didn't round off the value

B.Artificial Neural Network

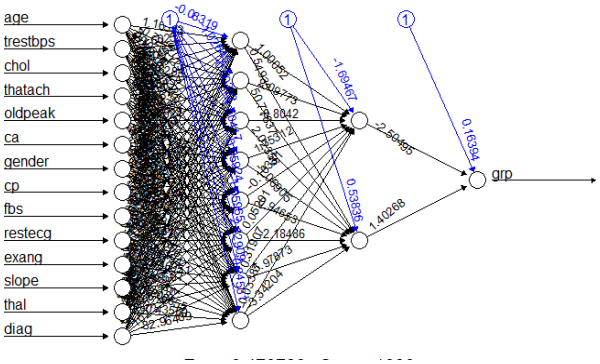
Artificial Neural Networks are one of the main tools used in machine learning. The Neural Networks itself consists of many small units called Neurons. These Neurons are grouped into several layers. Neurons of one layer connect to the next layer through weighted connections. Neural Networks consist of input and output layers, as well as a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are too complex or numerous for a human programmer to extract and the machine to recognize.

Information flows through a neural network in two ways. When it's learning or operating normally, patterns of information are fed into the network via the input units, which trigger the layers of hidden units, and these in turn arrive at the output units. This common design is called a feedforward network. Each unit receives inputs from the units to its left, and the inputs are multiplied by the weights of the connections they travel along. Every unit adds up all the inputs it receives in this way and if the sum is more than a certain threshold value, the units it's connected to are triggered.

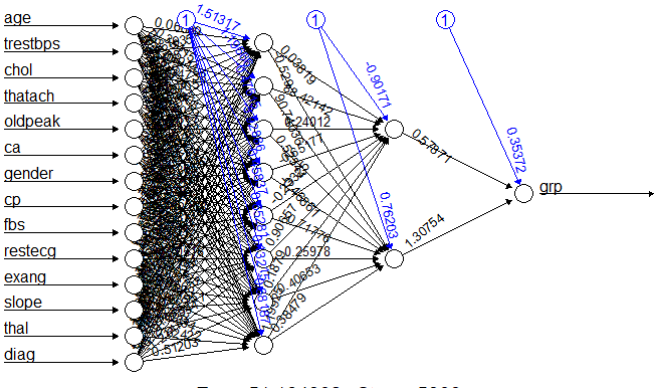
Neural networks learn things in exactly the same way, typically by a feedback process called backpropagation. This involves comparing the output network it produces with the output it was meant to produce, and by using the difference between them to modify the weights of the connections between the units in the network and working form output units through the hidden units to the input units which means going backward. In time, backpropagation causes the network to learn, reducing the difference between actual and intended output to the point where the two exactly coincide, so the network figures things out exactly as it's should.

C.R

a) R is an high level programming language and powerful programming language. R programming language easy to read, easy to learn and easy to code. Besides that, R has library for many things, use quickly build, the lower the performance and often less powerful prototype. The R is also great for validating ideas for so many different projects. R can integrate with Web-based System which will be implement for this semester. Besides that, R can be use to apply data mining techniques. For example, Artificial Neural Network will be use in R programming language. The figure below shows the diagram of ANN for model 1 which is having heart disease or not and the model 2 which is having 4 types of heart disease.



Model 1



MODEL 2

TABLE I. ACCURACY TABLE

	Model of Prediction of Heart Disease Accuracy		
	Type of models	***Model 1***	***Model 2***
1	Accuracy	100%	100%

Fig. 1. Shows the accuracy table

```
    0   1
0 141   0
1   0 126
```

```
   1  2  3  4
1 42  0  0  0
2  0 20  0  0
3  0  0 25  0
4  0  0  0 10
```

Fig. 2. Shows the confusion matrix for Model 1 and Model 2

	actual	prediction
1	0	0
2	1	1
3	0	0
4	0	0
5	0	0
6	1	1
7	0	0
8	1	1
9	1	1
10	0	0
11	0	0
12	1	1
13	0	0
14	0	0
15	0	0
16	1	1
17	1	1
18	1	1
19	0	0

35	1	1
36	1	1
37	1	1
38	1	1
39	1	1
40	1	1
41	1	1
42	1	1
43	2	2
44	2	2
45	2	2
46	2	2
47	2	2
48	2	2
49	2	2
50	2	2
51	2	2
52	2	2

Fig. 3. Shows the actual and prediction value for Model 1 and Model 2

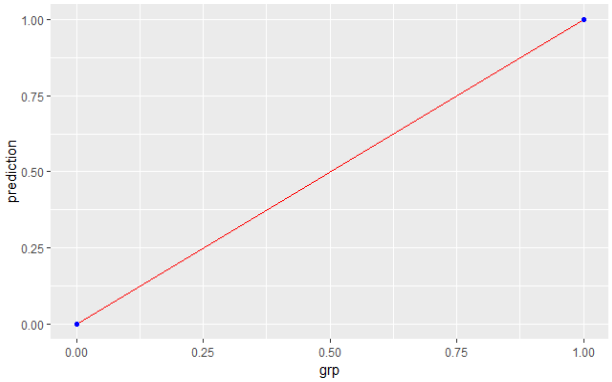


Fig. 4. Shows the graph between group and prediction in Model 1

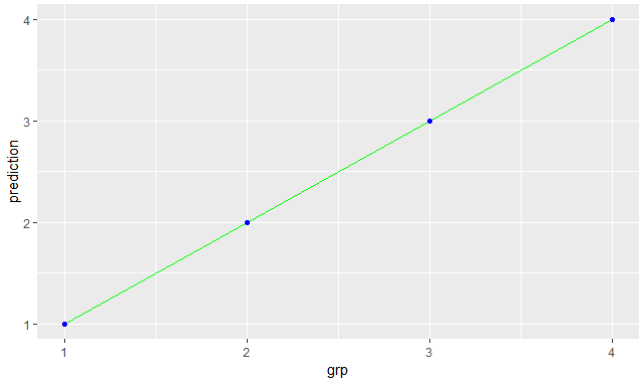


Fig. 5. Shows the graph between group and prediction in Model 2

Fig. 6. Shows the web-based system for the homepage

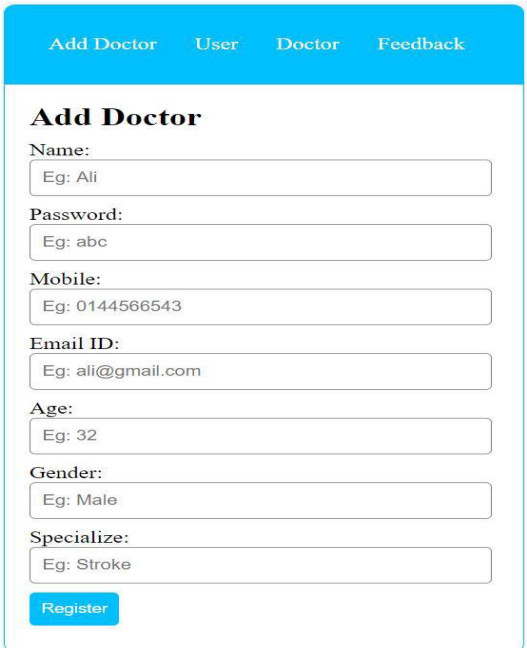


Fig. 7. Shows the admin can add the doctor in the system

ACKNOWLEDGMENT

I would like to take this opportunity to express my special thanks of gratitude to those who have guided and motivated me to complete this report, especially my supervisor who gave me the opportunity to do this project.

REFERENCES

[1] Shadab Adam Pattekari & Asma Parveen, "Prediction System For Heart Disease using Naïve Bayes," International Journal of Advanced Computer and Mathematical Sciences, vol. 3, pp. 290–294, June 2012.

[2] Sameh Ghwanmeh, Applying Advanced NN-based Decision Support Scheme for Heart Diseases Diagnosis, vol. 44. International Journal of Computer Application, April 2012.

[3] R.Chitra & V.Seenivasagam, "Review Of Heart Disease Prediction System using Data Mining and Hybrid Intelligent Techniques,"International Conference on Technology Journal on Soft Computing, July 2013.

[4] Nidhi Bhatla & Kiran Jyoti, “An Analysis of Heart Disease Prediction using Different Data Mining Techniques,” International Journal of Engineering Research and Technology, vol. 1, October 2012.

[5] Mrs.G.Subbalakshmi, Mr.K.Ramesh, & Mr.M.Chinna Rao, “Decision Support in Heart Disease Prediction System using Naïve Bayes,” Indian Journal of Computer Science and Engineering. 2011.

[6] M.Akhil Jabbar & Dr.B.L Deekshatulu, “Heart Disease Prediction System using Associative Classification and Genetic Algorithm,” International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies. 2012.

[7] K.Sudhakar & Dr.M.Manimekalai, “Study of Heart Disease Prediction using Data Mining,”International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, 2014.

APPENDIX B:
WEEKLY LOG, TURNITIN REPORT

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE: 13.12.2018	MEETING NO.: 1
PROJECT ID: 1173	
PROJECT TITLE : Web-based System for the Prediction of Heart Disease	
SESSION : 2018/2019	SUPERVISOR : Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

1. WORK DONE

[Please write the details of the work done after the last meeting.]

Correction of FYP1 report and modify the report according to the format given

2. WORK TO BE DONE

Study the implement ANN, what type of neural network use and algorithm use for the implementation.
Code Neural Network in RStudio

3. PROBLEMS ENCOUNTERED
4. COMMENTS

.....
Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan

.....
Co-Supervisor's Signature

.....
Student's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE: 20.12.2018	MEETING NO.: 2
PROJECT ID: 1173	
PROJECT TITLE : Web-based System for the Prediction of Heart Disease	
SESSION : 2018/2019	SUPERVISOR : Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

1. WORK DONE

[Please write the details of the work done after the last meeting.]

Study the implement ANN, what type of neural network use and algorithm use for the implementation.
Halfway progress doing code in RStudio for the implementation of Neural Network

2. WORK TO BE DONE

Implement Artificial Neural Network with Multilayer Perceptron using back-propagation algorithm with using some parameters
Generate the ANN diagram

3. PROBLEMS ENCOUNTERED
4. COMMENTS

.....
Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan

.....
Student's Signature

.....
Co-Supervisor's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE: 27.12.2018	MEETING NO.: 3
PROJECT ID: 1173	
PROJECT TITLE : Web-based System for the Prediction of Heart Disease	
SESSION : 2018/2019	SUPERVISOR : Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

1. WORK DONE

[Please write the details of the work done after the last meeting.]

Implement Artificial Neural Network with Multilayer Perceptron using back-propagation algorithm with using some parameters.

Generate the ANN diagram.

2. WORK TO BE DONE

Find the confusion matrix, actual and prediction and accuracy

Ensure the prediction value get almost the same with the actual

3. PROBLEMS ENCOUNTERED
4. COMMENTS

.....
Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan

.....
Student's Signature

.....
Co-Supervisor's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE: 3.1.2019	MEETING NO.: 4
PROJECT ID: 1173	
PROJECT TITLE : Web-based System for the Prediction of Heart Disease	
SESSION : 2018/2019	SUPERVISOR : Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

1. WORK DONE

[Please write the details of the work done after the last meeting.]

Find the confusion matrix, actual and prediction and accuracy
Ensure the prediction value get almost the same with the actual

2. WORK TO BE DONE

Correction for accuracy to ensure it will get 100% accuracy for the 1st model and for the 2nd model to get the specific value
Correction for actual and prediction to ensure the prediction value get the nearest to the actual value

3. PROBLEMS ENCOUNTERED
4. COMMENTS

.....
Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan

.....
Student's Signature

.....
Co-Supervisor's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE:10.1.2019	MEETING NO.: 5
PROJECT ID:1173	
PROJECT TITLE :Web-based System for the Prediction of Heart Disease	
SESSION :2018/2019	SUPERVISOR :Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

1. WORK DONE

[Please write the details of the work done after the last meeting.]

Correction for accuracy to ensure it will get 100% accuracy for the 1st model and for the 2nd model to get the specific value.

Correction for actual and prediction to ensure the prediction value get the nearest to the actual value for the first and second model.

2. WORK TO BE DONE

Plot some graph to compare the prediction with the group

Round off the prediction value to get the 100% for the 1st model

3. PROBLEMS ENCOUNTERED
4. COMMENTS

.....
Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan

.....
Student's Signature

.....
Co-Supervisor's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE:17.1.2019	MEETING NO.:6
PROJECT ID:1173	
PROJECT TITLE :Web-based System for the Prediction of Heart Disease	
SESSION :2018/2019	SUPERVISOR :Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

1. WORK DONE

[Please write the details of the work done after the last meeting.]

Plot some graph to compare the prediction with the group

Round off the prediction value to get the 100% for the 1st model

2. WORK TO BE DONE

Change some parameter to get the accurate result

3. PROBLEMS ENCOUNTERED
4. COMMENTS

.....
Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan

.....
Student's Signature

.....
Co-Supervisor's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE:24.1.2019	MEETING NO.:7
PROJECT ID:1173	
PROJECT TITLE :Web-based System for the Prediction of Heart Disease	
SESSION :2018/2019	SUPERVISOR :Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

1. WORK DONE
[Please write the details of the work done after the last meeting.]
Change some parameter to get the accurate result

2. WORK TO BE DONE
Implement Web-based system
Continue progress coding in R Studio

3. PROBLEMS ENCOUNTERED
4. COMMENTS

.....
Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan

.....
Student's Signature

.....
Co-Supervisor's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE:31.1.2019	MEETING NO.:8
PROJECT ID:1173	
PROJECT TITLE :Web-based System for the Prediction of Heart Disease	
SESSION :2018/2019	SUPERVISOR :Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

1. WORK DONE
[Please write the details of the work done after the last meeting.]
Implement Web-based system login, logout and other functions
Continue progress coding in R Studio

2. WORK TO BE DONE
Integrate R programming with PHP in web-based system
Complete the Web-based system

3. PROBLEMS ENCOUNTERED
Integrate R programming with PHP in web-based system Modify the system
4. COMMENTS

.....
Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan

.....
Co-Supervisor's Signature

.....
Student's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

MULTIMEDIA UNIVERSITY

Faculty of Computing & Informatics
Final Year Project Meeting Log

MEETING DATE: 7/2/2019	MEETING NO.: 9
PROJECT ID: 1173	
PROJECT TITLE : Web-based System for the Prediction of Heart Disease	
SESSION : 2018/2019	SUPERVISOR : Dr Kannan
STUDENT ID & Name:1131120615 Aaron Raj	CO- SUPERVISOR :

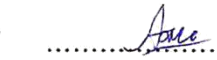
1. WORK DONE
[Please write the details of the work done after the last meeting.]
75% of the system complete
Report complete including correction and format just left testing and appendix

2. WORK TO BE DONE
Integrate R programming with PHP in web-based system
Report modification
Research paper modification

3. PROBLEMS ENCOUNTERED Integrate R Studio with PHP
4. COMMENTS

Supervisor's Signature

Dr. KANNAN RAMAKRISHNAN
Senior Lecturer
Faculty of Computing & Informatics
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya
Selangor Darul Ehsan



Student's Signature

.....
Co-Supervisor's Signature

NOTES:

1. Items 1 – 3 are to be completed by the students before coming for the meeting. Item 4 is to be completed by the supervisor.
2. Minimum six log sheets are to be submitted (at least one every other week).
3. Log sheets are compulsory assessment criteria for FYP. Student who fails to meet the requirements of log sheets will not be allowed to submit FYP report.

Final hardcover Report_Aaron Raj

ORIGINALITY REPORT

%18	%17	%8	%18
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Multimedia University Student Paper	%3
2	ictactjournals.in Internet Source	%2
3	Submitted to General Sir John Kotelawala Defence University Student Paper	%2
4	citeseerx.ist.psu.edu Internet Source	%1
5	Submitted to Prestige Academy (Pty) Ltd Student Paper	%1
6	Submitted to UT, Dallas Student Paper	%1
7	Submitted to National Institute Of Technology, Tiruchirappalli Student Paper	%1
8	research.ijcaonline.org Internet Source	%1

9	www.iosrjournals.org Internet Source	%1
10	Submitted to Sardar Patel Institute of Technology Student Paper	%1
11	W R SAM EMMANUEL, S JASMINE MINIJA. "Fuzzy clustering and Whale-based neural network to food recognition and calorie estimation for daily dietary assessment", Sādhanā, 2018 Publication	%1
12	www.ijert.org Internet Source	%1
13	Submitted to VIT University Student Paper	%1
14	www.slideshare.net Internet Source	%1
15	rep.bntu.by Internet Source	%1
16	pdfs.semanticscholar.org Internet Source	%1
17	www.idosi.org Internet Source	%1
18	www.viva-technology.org Internet Source	%1

19	Submitted to CSU, San Jose State University Student Paper	%1

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	< 1%
EXCLUDE BIBLIOGRAPHY	OFF		