

# **Data Mining Approach to Predict CVD Risk in the Sri Lankan Context**

**N.D.U. Gamage**

**2021**



# **Data Mining Approach to Predict CVD Risk in the Sri Lankan Context**

**A dissertation submitted for the Degree of Master of  
Business Analytics**

**N.D.U.Gamage**

**University of Colombo School of Computing**

**2021**





# DECLARATION

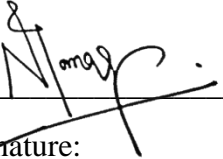
The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: N.D.U. Gamage

Registration Number: 2018/BA/013

Index number: 18880137

  
Signature: \_\_\_\_\_

12/09/2021

Date:

This is to certify that this thesis is based on the work of

Mr./Ms. \_\_\_\_\_ N. D. U . Gamage\_\_\_\_\_

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name: Dr. M.G. Noel A.S Fernando

\_\_\_\_\_  
Signature:

14/09/2021

Date:

I would like to dedicate this thesis to the Sri Lankan community to prevent major non-communicable diseases in Sri Lanka

## **ACKNOWLEDGEMENTS**

First and foremost, I would like to express my sincere gratitude to the project supervisor, Dr. M.G. Noel A.S Fernando, for his invaluable support, encouragement, and supervision. His kind guidance enabled me to conclude the project successfully.

I am also indebted to the medical students of Sri Jayawardhanapura University for their kind support and corporation in the process of collecting requirements.

Finally, I would like to thank all the people who have willingly helped me with their abilities.

## ABSTRACT

Cardiovascular diseases (CVDs) are considered the number one non-communicable disease that causes death and severe disabilities. CVDs associated with damage to the heart or blood vessels. Coronary heart disease, cerebrovascular disease, rheumatic heart disease, heart attack, heart failure are common cardiovascular diseases spread all around the world. The World Heart Federation predicts that there will be more than 23 millions CVD-related death per year by 2023. High blood pressure, smoking, high cholesterol, diabetes, physical inactivity, and family history of CVD are the main causes behind this tragedy. World Health Organization (WHO) and World Heart Foundation along with experts in the domain all around the world scale up the effort on CVD prediction and control through technical packages. Existing mobile and web-based packages are calculating the CVD risk using Framingham Risk Score. But existing solutions having age restrictions or are limited to major CVD risk factors. This study is introducing CVD classifier combining PCA to determine the CVD risk using literature-based classifiers and CVD risk factors. Results of the research indicate that the Random Forest classifier is most suitable to predict CVD risk prediction as the prediction accuracy is 85%. The data used for CVD risk prediction is not focused on Sri Lankan lifestyle, habits, and environment. Therefore, a questionnaire was prepared to identify CVD risk factors with the support of medical students and feedback from domain experts. Data collected from 1252 individuals all over Sri Lanka was fed into multiple classifiers to detect the best classifier, The experiment identified that the Random Forest classifier is suitable for CVD risk prediction in Sri Lankan oriented data due to it provides 100% accuracy for modifiable, contributing, and major CVD risk factors. As this is the first attempt to identify the CVD risk factors unique to the Sri Lanka context set of clustering and classifications was conducted to detect CVD risk groups and associations between factors. The association rule mining proved that neglecting age, age groups, or alcohol consumption is not suitable when predicting the CVD risk. Also, it proved that there is a high association between every non-communicable disease considered in this research (diabetics, blood pressure, and cholesterol). The cluster analysis identified that, if a person having any non-communicable diseases such as diabetics, cholesterol, or blood pressure and is not/less involved in physical exercises there is a high risk for getting into CVD.

# TABLE OF CONTENTS

DECLARATION.....	I
ACKNOWLEDGEMENTS .....	III
ABSTRACT .....	IV
TABLE OF CONTENTS .....	V
LIST OF FIGURES .....	VII
LIST OF TABLES .....	VIII
LIST OF ABBREVIATIONS.....	IX
CHAPTER 1 : INTRODUCTION.....	1
1.1 Motivation.....	2
1.2 Statement of the problem .....	4
1.3 Research Aims and Objectives .....	5
1.3.1 Aim .....	5
1.3.2 Objectives .....	5
1.4 Scope.....	6
1.5 Structure of the Thesis .....	7
CHAPTER 2 : LITERATURE REVIEW .....	8
2.1 Background of the Study .....	8
2.2 Comparison of Existing Literature-based Solutions .....	10
2.2.1 Feature Selection Approaches .....	10
2.2.2 CVD Risk Prediction Approaches .....	12
2.2.3 Comparison of Existing Literature-based Solutions.....	16
2.3 CVD Prediction Tools and Applications .....	17
2.3.1 Comparison of existing tools and applications.....	19
2.4 Related research and research gap .....	19
CHAPTER 3 : METHODOLOGY .....	21
3.1 Phase 1 .....	22
3.1.1 Data Accessibility.....	22
3.1.2 Data Pre-processing .....	22
3.1.3 Implementation of CVD Risk Prediction Model.....	22
3.1.4 Implementation of Principal Component Analysis .....	25
3.1.5 Identify the Influence of CVD Risk Factors for Betterment of the Life .....	28



3.2	Phase 2 .....	28
3.2.1	Data Gathering and Pre- processing .....	29
3.2.2	Implementation of CVD Risk Prediction Model with PCA .....	31
3.2.3	Identify the Influence of CVD Risk Factors for Betterment of Sri Lankans' Life	33
3.3	User Interaction with the Model .....	34
3.4	Tools and Technologies .....	35
3.5	Limitations .....	36
CHAPTER 4 : RESULTS AND EVALUATIONS.....		37
4.1	Experimental Results of Phase 1 .....	37
4.1.1	Results obtained in phase 1 to achieve objective 1 .....	37
4.1.2	Results obtained in phase 1 to achieve objective 2 .....	38
4.1.3	Results obtained in phase 1 to achieve objective 3 .....	40
4.2	Experimental Results of Phase 2 .....	43
4.2.1	Results obtained in phase 2 to achieve objective 1 and 2.....	44
4.2.2	Results obtained in phase 2 to achieve objective 3 .....	47
4.3	Experimental Results Evaluation.....	56
CHAPTER 5 : CONCLUSION AND FUTURE WORKS .....		61
5.1	Conclusions about Experiments.....	61
5.1.1	Detecting Critical Global CVD Risk Factors .....	61
5.1.2	Detecting Accurate CVD Prediction Classifier using Global CVD Risk Factors	62
5.1.3	Detecting Accurate CVD Prediction Classifier using CVD Risk Factors in Sri Lanka .....	62
5.1.4	Identification of CVD Risk Groups and Associations Between CVD Risk Factors	63
5.2	Conclusions about the Research Study .....	64
5.3	Future Work .....	65
6	REFERENCES .....	66
Appendix 1: Initial Questionnaire .....		69
Appendix 2: Data label assignment for each response for the google form to collect CVD factors in Sri Lankan Context.....		75

## LIST OF FIGURES

Figure 1: CVD risk prediction for SEAR B .....	3
Figure 2 : Overall system architecture of the research .....	21
Figure 3 : Code segment for build and evaluate model.....	24
Figure 4: Implementation of PCA .....	26
Figure 5: Explained variance of each attribute .....	26
Figure 6 : Mapping training and testing sets in to new feature set.....	27
Figure 7: Feedback obtained from domain experts for initial questionnaire.....	30
Figure 8 : User interface to display CVD risk prediction.....	34
Figure 9 : Accuracy classifiers with different PCA values. ....	37
Figure 10 : Accuracy scores returned by classification_report() function .....	39
Figure 11 : Clusters related to global CVD risk factors .....	40
Figure 12 : Best rules generated by Apriori association rule .....	42
Figure 13 : Accuracy scores returned by classification_report() function for all the CVD risk factors collected from Sri Lankan context.....	43
Figure 14 : Accuracy scores returned by classification_report() function for major risk factors .....	44
Figure 15: Accuracy scores returned by classification_report() function for modifiable risk factors .....	45
Figure 16: Accuracy scores returned by classification_report() function for contributing risk factors .....	45
Figure 17: Accuracy scores returned by classification_report() function for major and modifiable risk factors .....	46
Figure 18: Clusters related to Sri Lankan major CVD risk factors .....	47
Figure 19: Clusters related to Sri Lankan modifiable CVD risk factors .....	48
Figure 20 : Clusters related to Sri Lankan modifiable CVD risk factors .....	50
Figure 21 : Clusters related to Sri Lankan modifiable CVD risk factors .....	51
Figure 22 : Best rules generated by Apriori association rule – Major CVD risk factors .	52
Figure 23 : Best rules generated by Apriori association rule – Modifiable CVD risk factors .....	53
Figure 24 : Best rules generated by Apriori association rule – Contributing CVD risk factors .....	54
Figure 25 : Best rules generated by Apriori association rule – Major and modifiable CVD risk factors .....	55
Figure 26 : Comparison between before and after applying PCA for major CVD factors in Sri Lankan context.....	57
Figure 27 : Comparison between before and after applying PCA for modifiable CVD factors in Sri Lankan context.....	57
Figure 28: Comparison between before and after applying PCA for contributing CVD factors in Sri Lankan context.....	58
Figure 28: Comparison between before and after applying PCA for major and modifiable CVD factors in Sri Lankan context .....	58

## LIST OF TABLES

Table 1 : Comparison of CVD detection algorithms used during 2004-2016 time period (Banu and Swamy, 2016) .....	15
Table 2 : Comparison of literature-based state-of-the-art solutions .....	16
Table 3 : Comparison of existing state-of-the-art mobile solutions in the market .....	19
Table 4 : Comparison of classification accuracy before and after applying PCA.....	27
Table 5 : Accuracy comparison for CVD prediction datasets .....	32
Table 6: Software requirements for the proposed model .....	35
Table 7: Comparison of classification accuracy before and after applying PCA for dataset used in CVD prediction in Sri Lankan .....	56
Table 8: CVD risk clusters .....	59

## LIST OF ABBREVIATIONS

<b>Acronym</b>	<b>Meaning</b>
CVD	Cardiovascular Diseases
WHO	World Health Organization
NB	Naïve Bayes algorithm
DT	Decision Tree algorithm
API	Application programming interface
Sklearn	Sci-kit learn library
NCD	Non communicable diseases
IDE	Integrated development environment
WEKA	Weka Data Mining Software by university of Waikato

# CHAPTER 1

## INTRODUCTION

“Health is like money; we never have a true idea of its value until we lose it” – Josh Billings

Predicting future health risks immensely supports human beings with busy schedules and unhealthy lifestyles to enhance life expectancy. Therefore, this study is focused on predicting and preventing cardiovascular diseases in the Sri Lankan context.

Cardiovascular diseases (CVD) are defined as a set of irregularities of heart and blood vessels (World Health Organization, 2017) which includes;

- diseases in the vessels that supply the blood to heart muscles (coronary heart disease), brain (Cerebrovascular disease), arms and legs (peripheral arterial disease)
- damage to the heart muscles and heart valves due to rheumatic fever
- heart structure malfunction from birth (congenital heart disease)
- disability to supply blood to the head and lungs because of leg veins blocked by blood clots (Deep vein thrombosis and pulmonary embolism)

Above mentioned irregularities lead heart attacks and strokes due to it blocks the blood supply to the heart or brain. Irregular blood supply to the heart or brain leads to sudden deaths or long-term physical disability.

Cardiovascular diseases (CVD) are considered the main reason for global deaths according to World Health Organization (WHO). Estimations proved that nearly 17.9 individuals died from CVD in the year 2016 (World Health Organization, 2017). American Heart Association stated that nearly 18.6 million people die due to cardiovascular diseases globally. The statistics reflect that it is a 17.1% increase consider to the past decade. As it records more than 523.2 million CVD cases in the year 2019, the increase of CVD cases is 26.6% compared to 2010 (American Heart Association, 2021).

Even though rheumatic fever-based heart failures or congenital heart diseases are hard to pre-identified, it is possible to predict the presence of CVD in the future based on the lifestyle of any individual. The main reason for CVD-based deaths is the buildup of plaque in heart arteries

or blood vessels. The risk factors for build up a plaque are divided into 3 categories as major risk factors, modifiable risk factors, and contributing risk factors. The factors that are unable to change such as age, gender, race, and family history of CVD are major risk factors. The risk factors that can be controlled through changes in lifestyle and medication are considered modifiable risk factors. The use of tobacco, high blood cholesterol, high blood pressure, physical inactivity, diabetes, and obesity are examples of modifiable risk factors. Stress, alcohol diet, and nutrition are categorized as contributing risk factors where these factors can increase the risk of CVD but the significant impact for CVD is not yet determined (American Heart Association, 2016).

Therefore, this research is focusing on predicting CVD risk with high accuracy considering major, modifiable, and contributing risk factors based on online available global data. The research further extended to identify the CVD risk factors in the Sri Lankan context by considering local food habits, lifestyle, mental conditions, and environmental conditions. Therefore, there is a need to collect CVD related information from Sri Lankan context and predict the possibility of CVD occurrence. The research further extended to collect CVD related information from Sri Lankans via a questionnaire and predict the CVD risk using a classifier that provides the highest accuracy. The experiments conducted to determine the CVD risk with in global and Sri Lankan domains proves that the Random Forest classifier is more suitable for predicting CVD risk as the prediction accuracy was above 85% for global data and the prediction accuracy was 100% for Sri Lankan data. Identification of high and non-CVD risk clusters and association among other CVD factors also considered in this research work with the aim of reducing the CVD risk in Sri Lanka

## **1.1 Motivation**

Cardiovascular disease is considered an umbrella term that includes different types of irregularities in heart and blood vessels including heart attacks and strokes. According to the 2020 Heart disease & stroke statistical update fact sheet Asian/pacific islanders published by American Heart Association, 47.4% of males and 37.2% female Non-Hispanic Asians above 20 years old had CVD between 2013-2016 period (Benjamin et al., 2019). World heart federation along with WHO states that the annual global deaths caused by CVD are over 17 million (World Heart Federation, 2017). The article also highlighting that 31% of global deaths are caused because of CVD (World Heart Federation, 2017). The predicted deaths from CVD by 2030 are over 23 million.

Due to the high number of death and new CVD cases, World Health Organization kept a target to reduce premature deaths up to 25% - which occurs due to non-communicable diseases (NCDs), where the CVD make up the largest proportion of deaths by the year 2025. The organization planning to achieve the target by correcting the behavioral risk factors. To correct the behavioral risk factors, it is important to predict future CVD risks. Therefore, the WHO introduced the WHO/ISH risk prediction charts for 14 WHO epidemiological sub-regions (World Health Organization, 2014). According to the chart, Sri Lanka and other south Asian countries are belong to the WHO sub-region “Southeast Asia Region B (SEAR B)”. the CVD risk prediction for SEAR B countries is given in figure 1 below.

According to the above figure, smokers who are suffering from diabetics having a high risk of getting into CVD. The risk prediction is not limited to smoking or diabetics, charts introduced by WHO predict the CVD risk by considering gender, age, systolic blood pressure, total cholesterol, smoking status, and presence of diabetes mellitus. But the report itself stated that the CVD risk may increase with the presence of antihypertensive therapy, obesity, family history of premature coronary heart disease, triglyceride level. Therefore, it is important to identify the risk factors to provide an accurate CVD risk prediction.

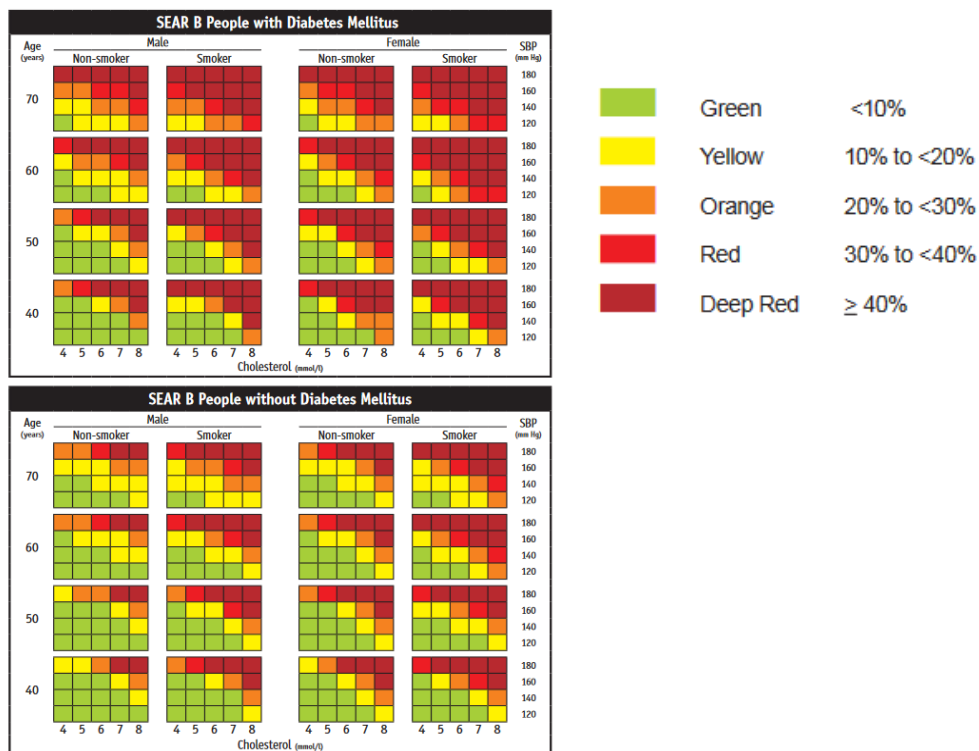


Figure 1: CVD risk prediction for SEAR B

The 9th edition of Kumar and Clark's clinical medicine (Kumar et al., 2017) categorized risk

factors for coronary disease into two groups as fixed risk factors, potential and changeable risk factors. Age, gender, CVD family history, and deletion polymorphism in ACE gene DD are considered as fixed risk factors for CVD while hyperlipidemia, smoking, hypertension, diabetes, lack of exercise, personality, obesity, gout, drug, and few more other factors are introduced as potentially changeable CVD risk factors. But commercially available web applications and mobile applications are poor in considering potentially changeable factors to predict CVD risk. Hence, it is important to research more on CVD risk factors and identify the most critical risk factors based on patient experience rather than depend on the WHO CVD prediction charts.

WageIndicator Network in Sri Lanka stated that there is no law to ensure all government and private-sector employees are secured under a medical insurance policy (WageIndicator Network, 2020). Therefore, the members of middle- and low-income families are reluctant to participate in regular medical checkups to ensure their current medical status. The busy lifestyle of modern society leads to consuming unhealthy and chaotic nutrition, insufficient sleep, and physical inactivity, this situation generates an unhealthy individual who is having the risk of getting into heart failure or stroke at any time.

According to the Final year medical students of Sri Jayawardhanapura Medical Faculty, currently, they are not practicing any CVD prediction mechanism in their clinics. Therefore, it is important to have a method to predict the CVD risk accurately and notify the people if they belong to a high CVD risk cluster, as it is always prevention is better than cure.

## **1.2 Statement of the problem**

Most of the Sri Lankans in mid and low-income families are reluctant to participate in frequent body checkups, as they cost a high amount of money. Unless they are faced with heart pain or stroke the consideration of personal health is low. Poor health facilities in government hospitals and medical coverage are limited, the situation getting worst. According to the features of existing tools and existing medical facilities in the Sri Lankan context, this research is focused on providing solutions for the following research question

**How to predict the CVD risk by detecting the most critical CVD risk factors in global and Sri Lankan context to prevent CVD before occurring while detecting high-risk CVD morbidity groups to issue health precautions?**



The above statement of the problem can be further decomposed as follows

1. What are the critical CVD risk factors among major, modifiable, and contributing risk factors in the global context?
2. How to identify the best data mining approach for CVD prediction?
3. How to predict CVD risk accurately using a selected data mining approach?
4. What are the CVD risk factors related to Sri Lanka domain?

## **1.3 Research Aims and Objectives**

### **1.3.1 Aim**

The main objective of the project is;

To develop a data mining approach to predict cardiovascular disease risk accurately by considering major, modifiable, and contributing CVD risk factors.

### **1.3.2 Objectives**

According to American Heart Association, CVD risk factors can be divided into 3 categories as major risk factors, modifiable risk factors, and contributing risk factors (American Heart Association, 2016).

- Major risk factors - Age, gender, race, and family history of CVD are major risk factors (The factors that are unable to change)
- Modifiable risk factors - Use of tobacco, high blood cholesterol, high blood pressure, physical inactivity, diabetes, and obesity (The risk factors that can be controlled through the changes in lifestyle and medication)
- Contributing risk factors - Stress, alcohol diet, and nutrition (Factors which can increase the risk of CVD but the significant impact for CVD is not yet determined)

The objectives are focused on performing datamining operations (testing and implementation) on the above CVD risk factors

1. To identify the critical CVD risk factors among major, modifiable, and contributing risk factors listed in existing research works.
2. To predict the CVD risk by using identified critical major and modifiable risk

factors after evaluating the accuracy of well-known data mining algorithms.

3. To detect major, modifiable, and contributing risk factors in the Sri Lankan domain
4. To evaluate the accuracy of CVD predicting algorithms by applying Sri Lankan data
5. To test the influence of CVD risk factors for the betterment of the life

1<sup>st</sup> and 2<sup>nd</sup> objectives were achieved using the global dataset which includes 12 attributes (11 features and 1 target variable). 3<sup>rd</sup>, 4<sup>th</sup>, and 5<sup>th</sup> objectives were achieved using collected Sri Lankan data which contains 1252 data records.

## 1.4 Scope

Intending to provide accurate CVD risk prediction, this research work provides a mechanism to calculate CVD risk by considering major, modifiable, and contributing risk factors and any other social and environmental factors such as lifestyle, BMI, workplace stress, and living environment.

In order to calculate the CVD, risk this research work is divide into 2 phases.

- Phase 1 / First phase – Identify CVD risk factors and predict CVD risk with high accuracy using online available CVD data
- Phase 2 / Second Phase - Identify CVD risk factors and predict CVD risk with high accuracy using CVD data collected from Sri Lankans (1252 instances)

In the first phase the CVD data available online (in Kaggle.com (Ulianova, 2019) was be pre-processed and input to multiple data mining-based feature selection algorithms to determine the most relevant features among major, modifiable, and contributing risk factors. Majority of the CVD prediction applications available in the market not considering age groups below 40 years old when predicting CVD risk. According to WHO and the American heart association (American Heart Association, 2016) age consider as a major CVD risk factor. The use of a data mining approach and determining the patterns in CVD risk will avoid missing age group issues available in the majority of CVD calculating solutions in the future.

After identification of critical features in the CVD prediction online dataset, it input to

commonly used classification algorithms to evaluate the most accurate classification algorithm.

In the second phase, CVD data was collected via a questionnaire from Sri Lankan context to determine considerable CVD risk factors in Sri Lanka. Then the highest accurate algorithms were used to predict the CVD risk in Sri Lanka based on the unique features in the country

After analyzing the accuracy of commonly used data mining algorithms to predict CVD risk the most accurate algorithm use to predict the CVD risk of any intended individual. Users who are willing to track their CVD health conditions allows to observe the CVD risk based on their lifestyle and retrieve customized health insights using a data mining-based approach through the research. If the users having minor risks, they will be able to reduce the risk by following health insights and correct their lifestyles accordingly.

## **1.5 Structure of the Thesis**

The introduction chapter includes the motivation behind the study, research objective, summarized background study, and scope of the research to provide an insight into the dissertation. Background and related work chapter include critical analysis of most and recent similar work to justify the research work. The systematic approach followed to solve the identified research problem including data gathering, analyzing, and processing is presented in the methodology chapter. The chapter contains identified risks, and limitations as well. The architectural decisions taken for the solution is stated under the results and solution design chapter. The final chapter, evaluation describes the validity of the obtained results and the contribution to the domain

## **CHAPTER 2**

### **LITERATURE REVIEW**

Cardiovascular disease is defined as the major reason of death globally as it is the cause 80% of deaths all over the world. Pre identification of CVD might eventually reduce the number of deaths in considerable manner every year. In order to achieve the target researchers in multiple domains introduced smart solutions, applications, equations and algorithms. This section identifies state-of-the-art solutions designed for other similar studies from the problem domain that are currently in use.

The main objective of this project is to determine the critical factors for the CVD by considering major risk factors, modifiable risk factors and contributing risk factors defined by American Heart Association (American Heart Association, 2016) for both global and Sri Lankan. After identification of risk factors, CVD risk could be predicted for each user of the system. Health recommendations for betterment of life generated according to the current health conditions at the end of prediction. Following research works were studied to propose the solution for the identified research problems.

#### **2.1 Background of the Study**

The “Predictive analysis of rapid spread of heart disease with data mining” paper published by Radhanath Patra and Bonomali Khuntia (Patra and Khuntia, 2019) focused on identifying the most appropriate and best attributes and the classifier for heart disease identification. They have used UCI online machine learning repository as the dataset and feed the preprocessed data into the WEKA tool and a Python tool developed using Spider IDE. KNN, decision tree, support vector machine (SVM), RBF, Naïve Bayes, J48 algorithms were tested using the WEKA tool and Python tool to find the most accurate classifier. Based on the experiment the researchers found that the use of python-based Decision tree algorithm provides the highest accuracy than the WEKA tool-based J48, KNN, RBF, and naïve Bayes algorithms. The dataset used for the research contained 76 attributes that affect heart disease prediction, but the researchers selected 14 factors among them as the best attributes for heart disease prediction. The select attribute feature in the WEKA tool is used in the research for the feature selection process based on the information gain and entropy.

“Identification of Cardiovascular disease risk factors among diabetes patients using ontological data mining techniques” is a paper published by M.I.Qrenawi et.al. (Qrenawi and Al Sarraj, 2018). The research is more focused on predicting CVD risk among diabetes patients. The research builds an ontology for cardiovascular disease from domain terms defined in John Hopkins health systems: Cardiovascular disease – Glossary ([www.hopkinsmedicine.org](http://www.hopkinsmedicine.org), n.d.), conducting interviews with domain specialists and other reused ontologies such as CardioOWL and DBNO. The ontology construction minimized the number of attributes that are needed for the data mining process. After the ontology-based pre-mining process, the research team applied the FP-growth algorithm and generated association rules to predict the CVD risk. The results proved that the use of ontology with data mining algorithms exceeds the learning accuracy up to 90%.

Narain R, Saxena S, Goyal A in their paper of “Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network-based approach” (Narain et al., 2016) proved that the accuracy of CVD risk prediction using QNN is 98.57% while the accuracy of Framingham risk score was 19.22%. Even the accuracy is high the research is not focused on providing health insights.

M.G.Ruano et. el in their paper of “Reliability of Medical Databases for the use of Real Word Data and Data Mining Techniques for Cardiovascular Diseases Progression in Diabetic Patients” (Ruano et al., 2018) used real-world data (RWD) and big data mining (BDM) to analyze CVD disease progression of diabetes patients. The study was capable of revealing missing and misleading fields, inconsistencies in instrumental data inputs, user dependencies, and temporal fields. As a result, they suggested that the accessing of the same database at the same time by multiple individuals. The information may lose from one person to another if all the input records going to be updated by the hospital staff. To avoid such errors companies are requested to implement a strong method for filling their database for future reference.

The above research papers show the usage of data mining techniques supports predicting CVD with higher accuracy, but most of the research is focused on a few attributes in CVD risk prediction. Except for the research of M.I.Qrenawi et.al., other researchers are not much focusing on selecting the attributes for the data mining models. The paper of M.G.Ruano focuses on the importance of collecting data without missing or misleading data.

The CVD risk calculator EBMcalc developed by Foundation Internet Services, LLC stated that risk calculator may overestimate or underestimate CVD risk. By having a mechanism to calculate the CVD risk with a critical, considerable number of risk factors it is possible to provide accurate CVD risk estimation.

## **2.2 Comparison of Existing Literature-based Solutions**

### **2.2.1 Feature Selection Approaches**

Feature selections consider as a process of reducing dimensions from high dimensional data to reduce time complexity and increase the prediction precision from any classifier. Classification of high dimensional data using filtration attribute evaluation feature selection method of data mining research conducted by Ammisetty Veeraswamy et.al (Veeraswamy and Babu, 2019) introduced a model where numerous elements positioning methods combined together to produce solitary positioning rundown. In the first phase they identified different features from UCI dataset and performed feature ranking on top of 5 UCI datasets. Features were re-ranked and input to multiple classifiers such as Naïve Bayes, J48, SMO, Decision table, JRIP, Random Forest and K-Star in the second phase. The results of the experiment proved that use of ranking methods like relief attribute eval, gain ration eval etc. before feed the data to the feature selection algorithms gives better over-all performance.

Feature selection and extraction in data mining research conducted by Aparna.U.R et.al (Aparna U.R. and Paul, 2016) proposed a data mining based mechanism to select features effectively. The research proposed to use perceptron algorithm in and assign weights based on the input requirements. Data above the weight assigned suggested to excited and moves to the next mode as selected data while others were rejected. Features selected were proposed to truncate to the nearest possible value and apply distance measure evaluation. As researchers assumed that the features nearby could have slight difference among the distance calculate, Logistic regression was suggested to use to enhance the feature selection mechanism.

Maryam Zaffar et.al in their research of performance analysis of feature selection algorithm for education data mining (Zaffar et al., 2017) evaluated the performance of different feature selection algorithms. The research used WEKA (Waikato Environment

for Knowledge Analysis) tool to analyze the performance of CfsSubsetEval, ChiSquaredAttributeEval, FilteredAttributeEval, GainRatioAttributeEval, Principle Components, and ReliefAttributeEval feature selection algorithms regards to 15 feature selection algorithms named as BayesNet, Naïve Bayes, NaiveBayesUpdateable(NBU), MLP, Simple Logistic(SL), SMO, Decision Table(DT), Jrip, OneR, OneR, DecsionStump(DS), J48, Random Forest(RF), RandomTree(RT), REPtree(RepT) classification algorithms. The results were proved using precision, recall, F-measure and prediction accuracy. The results proved that among all the available feature selection algorithms resides in WEKA tool principal component feature selection algorithm showed better results with Random Forest classifier. Researchers stated that the parameter tuning need to be performed on top of feature selection algorithms to achieve better performance.

A literature review of feature selection techniques and applications done by S.Visalakshi and V. Radha (Visalakshi and Radha, 2014) investigated the performance of existing feature selection methods and applications. The critical literature survey supported researchers to conclude that instead of using filter and wrapper methods separately it is better to embed both methods for feature selection as it will increase the accuracy, speed and reduce the error rate.

All the above-mentioned research approaches were taken place to evaluate existing feature selection algorithms in different applications. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining paper published by R. Kavitha et.al (Kavitha and Kannan, 2016) can be considered as most relevant research work to this research work. Principle component analysis algorithm used in the first step of the research to determine similarities and differences between each attribute and to compute covariance matrix. Eigenvector with the highest eigen values chosen as principal component of the dataset and eigen values sorted in ascending order and data with highest significant values were taken while discarding other values. age, gender, cp, tresbps, chol,fbs ,slope, ca, thal, num, Ab pre, restecg, thalach, exang and oldpeak were filtered as most relevant 15 attributes. The steps followed to obtain the output is as follows. Mean, variance, covariance correlation, eigen values and eigen vectors calculated initially to make the above decision. Then the covariance matrix diagonalize by considering variable with highest variance is the highly relevant attribute where  $P =$  ordered set of principle components. The

diagonalization supported to minimize the redundancy and maximize the covariance. Variance values associated with P ranked accordingly. Then the calculate eigen values were ranked from lowest to highest and lesser significant values were ignored. Researchers created a matrix with eigen vectors as columns and selected the maximum significant values for extract relevant features.

Next the feature subset was selected with the support of wrapper filter. Wrapper filter ranked the attributes with the support of 5 cross fold validation. Heuristic searching algorithm used to conduct searching operations. With the support of feature subset search and classifier the wrapper method was capable to generate relevant feature subset. Finally, selected feature subset was ranked using information gain ratio. The researched concluded that the performance of the solution proposed improved compared to the well-known scoring functions such as Euclidean distance and Pearson correlation as it used wrapper filter for the feature selection. The framework could predict heart disease by using reduced test attributes successfully.

Therefore, the features will be reduced and input to the classifier algorithms in this research to determine the CVD risk as well.

### **2.2.2 CVD Risk Prediction Approaches**

Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques is a research conducted by K. Sirinivas et.al (Srinivas et al., 2010) to test the percentage of cardiovascular risk in coal mining area in Andhra Pradesh, India. They gather information over the phone from the workers in coal mining areas with the support of Andhra Pradesh health department to make the predictions. Their initial data mining model was developed by using UCI repository data to determine the accuracy of the data mining algorithms in CVD prediction. After preprocessing with the support of encoding, the data was input to different classification-based data mining techniques. Rule set classifier, decision tree algorithms (C4.5), CART and ID3, Neural Network Architecture with 20 input nodes, 10 hidden nodes and 10 output nodes, Neuro-Fuzzy stochastic back propagation algorithm, Bayesian network structure discoveries, and Support vector machine were tested in the research to predict the heart disease risk. The used algorithms supported to extract



pattens in generating heart risks by calculated weightages. Patterns which demonstrate values larges than the predefined threshold was chosen for the accurate prediction of heart attacks. The results proved that the decision tree is more efficient in predicting the heart attacks for the patients without heart morbidity when considering the model accuracy and sensitivity. For the patients with heart morbidity Neural network and SVM algorithms provide highest accuracy and sensitivity against other algorithms used. Researchers concluded that the model prediction should further enhanced with patient's financial status, stress, pollution, and medical history. Therefore, in this research, above mentioned contributing risk factors will be tested for accurate CVD risk prediction

Shan Xu et.al. in their paper of cardiovascular risk prediction method based on test analysis and data mining ensemble system (Xu et al., 2016) introduced to predict CVD risk in Chinese context. The model consists with 3 layers such as data source interface to store medical testing, imaging and diagnostic results of CVD patients, Text analysis took place in second step to understand and extract useful information from medical images and prescriptions in Chinese. Word segmentation and synonym identification was used by the researchers to convert raw data into numerical machine editable format to use in data preprocessing stage. Expert's marking, replace missing values, standardize, reducing dimensions are the preprocessing techniques used in the research. After the pre-processing the over 50 data mining techniques were evaluated by using cross validation with 10 folds. Researchers selected 6 most suitable classifier to generate ensemble system with the aim of reducing bias and improving accuracy. Naïve bayes, SWO function, IBI, regression, decision table rule-based method, and C4.5 decision tree algorithm provided highest accuracy in predicting CVD risk. Adjust voting method is the approach taken place to make the final decision. Prediction confidence from above mentioned sub classifiers used in the adjusted voting mechanism/ensembled system. Researchers proved that when classifier results input to the ensemble system it achieved precision of 79.3% which were better than traditional systems.

Efficient heart disease prediction system using decision tree is a research presented by Purushottam et.al. (Purushottam et al., 2015) with the aim of developing rules to predict CVD risk level accurately. The research considered age, gender, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca and thal parameters available in UCI repository to perform the evaluation. The WEKA tool used in the experiment for dataset analysis while KEEL (Knowledge Extraction based on Evolutionary Learning) tool used

to determine the classification decision rules. The experiment showed that the rules inferred from decision trees having 0.8675 (86%) of success in predicting CVD risk.

Heart disease prediction using lazy associative classification is a research presented by M.Akhil Jabbar (Jabbar et al., 2013). The researchers applied PCA and ranked all the attributes available in the dataset to identify the most crucial factors for CVD. Attributes with high rankings were used to develop class association rules. After identification of principle attributes subsets were created and calculated the probabilities. Subsets of attributes with largest posterior probabilities were assigned to each instance and calculated the accuracy. The accuracy of the lazy associative model was evaluated using 7 medical and non-medical UCI repositories. According to the results the researchers proved that the lazy associative datamining model perform well in predicting CVD risk in Andhra Pradesh, India

F. Mendonca et.al in their paper of intelligent cardiovascular disease risk estimation prediction system (Mendonca et al., 2019) introduced a CVD risk prediction approach with the support of K-nearest neighbors algorithm. The research used UCI repository and preprocessed the data before reduce number of attributes for better prediction. Apriori algorithm which follows association rule mining principle was used in the experiment to reduce number of attributes for 14 to 11. WEKA tool was used in the research and sensitivity, specialty, accuracy and Matthew's correlation coefficient used to evaluate the model accuracy. ANN, KNN, SVM and NB classifiers were evaluated against above mentioned methods and identified that the KNN shows highest values in every measurement. ANN classifier was the next highest in the list.

“Prediction of heart disease at an early stage using data mining and big data analytics: A survey” conducted by N.K.Salma Banu et.al (Banu and Swamy, 2016) compared different approaches used by researchers all over the world from 2004 – 2016. According to the survey, they were capable of identifying 18 different approaches. The below table 1 contains the comparison results that the research team obtained

Table 1 : Comparison of CVD detection algorithms used during 2004-2016 time period (*Banu and Swamy, 2016*)

Year	Author	Data mining technique/ algorithm	Accuracy
2004	Carlos Ordonez	Association Rule	-
2006	Hongmei Yan et al.	Multilayer perceptron	90%
2007	Yanwei X et al.	SVM, ANN, DT	92.1%, 91.0%, 89.6%
2007	Heon Gyu Lee et al.	Bayesian classification, CMAR, C4.5, SVM	81%, 80%, 78%, 85%
2008	Sellappan et al.	NN, Naïve Bayes, DT	85.68%, 86.12%, 80.4%
2009	Minas A Karolis	C4.5 (DT) Algorithm	82%
2010	K. Srinivas et al.	SVM, DT (C4.5), MLP	82.5%, 82.5%, 89.7%
2011	Leandro Pecchia et al.	Classification & Regression tree Techniques	96.39%, 79.31%
2011	Fahim Sufi et al.	Clustering Technique Apriori Algorithm	97%
2012	T John Peter et al	NB, DT, k-NN, NN	83.70%, 76.66%, 75.18%, 78.148%
2012	Chaitali S et al.	NN, DT, NB	100%, 99.62%, 90.74%
2012	Maishowman et al.	SVM with Bagging Algorithm	84.1%
2013	Syed Umar Amin et al.	NN Backpropagation Algorithm	96.2%
2013	Shamsher Bahadur et al.	DT, NB, Classification Clustering	99.2%, 96.5%, 88.3%
2013	I.S. Jenzi	NB	80.7%
2014	Hlaudi Daniel et al.	J48, REP TREE, NB, CART	99.0741%, 99.222%, 98.148%, 99.0741%
2014	B. Venkatalakshmi et al.	NB, DT	85.034%, 84.0136%
2015	Lokanath Sarangi et al	GA Technique	90%

The above comparison table indicates that the majority of researchers used Neural Network (NN), Decision Tree (DT), Naïve Bayes(NB), and Support Vector Machine (SVM) for the heart disease prediction. Among the algorithms highest accuracy is shown by DT and NN algorithms.

The above research papers show the usage of data mining techniques supports predicting the CVD with higher accuracy, but most of the research is focused on a few attributes in CVD risk prediction. Majority of the research works used ANN, KNN, SVM, NB, random forest classifiers in CVD risk prediction and highlighted the importance of selecting principal attributes in CVD risk prediction.

### 2.2.3 Comparison of Existing Literature-based Solutions

The below 2 table contains summarization about the existing research approaches studies for this research work

Table 2 : Comparison of literature-based state-of-the-art solutions

Paper	Classifiers / algorithms/methods used	Tools used	Selected best classifier/s	Novelty
<b>Feature selection</b>				
Ammisetty Veeraswamy et.al (Veeraswamy and Babu, 2019)	Naïve Bayes, J48, SMO, Decision Table, JRIP, Random forest, K-Star	N/A	N/A	Filter Attribute Evaluation method
Aparna.U.R et.al (Aparna U.R. and Paul, 2016)	Perceptron algorithm, Truncate distance measure evaluation, Logistic regression	N/A	N/A	Use of logistic regression
Maryam Zaffar et.al (Zaffar et al., 2017)	CfsSubsetEval, ChiSquaredAttributeEval, FilteredAttributeEval, GainRatioAttributeEval, Principle Components, and ReliefAttributeEval feature selection algorithms BayesNet, Naïve Bayes, NaiveBayesUpdateable(NBU), MLP, Simple Logistic(SL), SMO, Decision Table(DT), Jrip, OneR, OneR, DecsionStump(DS), J48, Random Forest(RF), RandomTree(RT), REPTree(RepT) classification algorithms	WEKA	principle component feature selection algorithm with Random Forest classifier.	parameter tuning performed on top of feature selection algorithms to achieve better performance
S.Visalakshi and V. Radha (Visalakshi and Radha, 2014)	Embed wrapper and filter methods.	N/A	N/A	N/A
R. Kavitha et.al (Kavitha and Kannan, 2016)	Principle component analysis algorithm, Eigenvector and eigen values calculation, Heuristic searching algorithm. Information gain ratio.	N/A	N/A	used wrapper filter
<b>CVD risk prediction</b>				
K. Sirinivas et.al (Srinivas et al., 2010)	Rule set classifier, decision tree algorithms (C4.5), CART and ID3, Neural Network , Neuro-	N/A	Neural network and SVM algorithms	N/A

	Fuzzy, stochastic back propagation algorithm, Bayesian network structure discoveries, and Support vector machine			
Shan Xu et.al. system (Xu et al., 2016)	50 data mining techniques	N/A	Naïve bayes, SWO function, IBI, regression, decision table rule based method, C4.5 decision tree algorithm	Ensemble system
Purushottam et.al. (Purushottam et al., 2015)	Decision tree	WEKA, KEEL	Decision tree	
M.Akhil Jabbar (Jabbar et al., 2013)	PCA, association rule	N/A	N/A	Combination of PCA and association rule
F. Mendonca (Mendonca et al., 2019)	Apriori algorithm ANN, KNN, SVM and NB	WEKA	KNN, ANN	Remove number of features using apriori algorithm
Salma Banu (Banu and Swamy, 2016)	SVM, NB, Decision tree (C4.5), random forest, ANN and KNN, association rule, NN, J48	N/A	SVM, NB, Decision tree (C4.5), random forest, ANN and KNN	N/A (A survey)

### 2.3 CVD Prediction Tools and Applications

- **ESC CVD Risk calculation app** (“ESC CVD Risk Calculation App,” n.d.)

This app can predict the risk of CVD for the next 10 years by considering current heart related medical treatments. But it is not considering the CVD risk factors such as alcohol consumption, BMI value, and physical inactivity. The app also not providing customized health tips based on the current lifestyle.

- **Absolute CVD risk calculator** (heartfoundation.org.au, n.d.)

The CVD risk calculator introduced by the Australian heart foundation is capable of calculating the CVD risk based on gender, age, systolic blood pressure, smoking status, total cholesterol, HDL cholesterol, and presence of diabetes. The individuals should be between age 35 – 74 to calculate the CVD risk. The results of the tool are limited to risk prediction but not focusing on providing health insights to control the risks

- **Reynolds risk score** (Cleveland Clinic medical professional, n.d.)

The Reynolds risk score is another CVD calculator that predicts CVD risk by considering age, gender, smoking state, systolic blood pressure, total cholesterol HDL cholesterol, high sensitivity C-Reactive Protein, CVD history. The limitation of the application is, the prediction is recommended only for people between 45-80 years old. The alcohol consumption, BMI, or physical inactivity is also not considered in this application.

- **ASCVD Risk Estimator** (American College of Cardiology, n.d.)

American college of cardiology foundation introduced a tool named ASCVD by considering demographic factors, laboratory report values such as total cholesterol, HDL cholesterol, and systolic blood pressure, and personal history. They have considered the treatment for hypertension also as a CVD risk calculator factor.

### 2.3.1 Comparison of existing tools and applications

The below table 3 contains comparison between available CVD risk prediction tool in the market with the proposed approach.

Table 3 : Comparison of existing state-of-the-art mobile solutions in the market

Product Features	ESC CVD Risk calculation app	Absolute CVD risk calculator	Reynolds risk score	ASCVD Risk Estimator	New approach
Calculate CVD risk probability	✓	✓	✓	✓	✓
Consider modifiable risk factors	Partially	X	X	Partially	✓
No age restriction for risk calculation	X	X	X	X	✓
Deliver health insights	Partially	X	X	X	✓
Customized CVD prediction based on Sri Lankan lifestyle	X	X	X	X	✓
Identify the influence of CVD factors in CVD occurrence and betterment of life	X	X	X	X	✓

### 2.4 Related research and research gap

The existing research studies provide methods of calculating CVD risk, but most of them are not focusing on determining the most critical factors for CVD risk prediction. Hence, the applications developed by different organizations are using different factors for CVD risk calculations. Most of the traditional applications directly depend on the Framingham algorithm. According to the research of Narain R et.al. the accuracy of the Framingham algorithm in CVD risk calculation is low. Hence, it is better to move to a data mining-based approach. Identification of CVD is not enough to enhance the life expectancy of human beings, it is also important to identify CVD risk groups to have awareness programs. But above-mentioned research papers or applications are not focusing on identify user groups with CVD risk

Commercially available web and mobile products for CVD risk calculation having age limitations as they use mathematical calculations for the prediction. Absolute CVD risk calculator ([heartfoundation.org.au](http://heartfoundation.org.au), n.d.), Reynold risk score (Cleveland Clinic medical professional, n.d.), ASCVD risk estimator (American College of Cardiology, n.d.) are few applications which having age limitations in calculating CVD risk. As data mining-based approaches can make decisions by examining patterns in the available data “age” factor can be considered in CVD risk calculation in this research. According to Kumar and Clark (Kumar et al., 2017) age is one of the main CVD risk factors. Neglecting or restricting the “age” factor reduces the prediction accuracy according to European Heart Journal (Leening et al., 2016).

Applying and evaluating feature selection algorithms and classification algorithms could be able to eliminate above mentioned gaps while predicting CVD with high accuracy by considering most relevant risk factors.

The CVD risk factors could be vary based on the Sri Lankans food preference, lifestyle, and conditions such as stress. The literature survey highlighted that there is no proper mechanism to CVD risk prediction in the local context. Collect CVD data from Sri Lankan context, pre-process and apply them into pre-identified CVD prediction algorithms could eliminate the above gap and reduce the CVD death count in Sri Lanka.



# CHAPTER 3

## METHODOLOGY

The research conducted in two phases. The first phase dedicated to determining critical CVD factors and predict the CVD risk with high accuracy. The study used globally available dataset and for accurate CVD prediction. The second phase dedicated for CVD risk calculation in local context which used data collected from via a questionnaire from CVD effected and non-effected Sri Lankans all over the country. Below sections describes the methodology which followed to get the outputs. The figure 2 below illustrates the overall system diagram of the project.

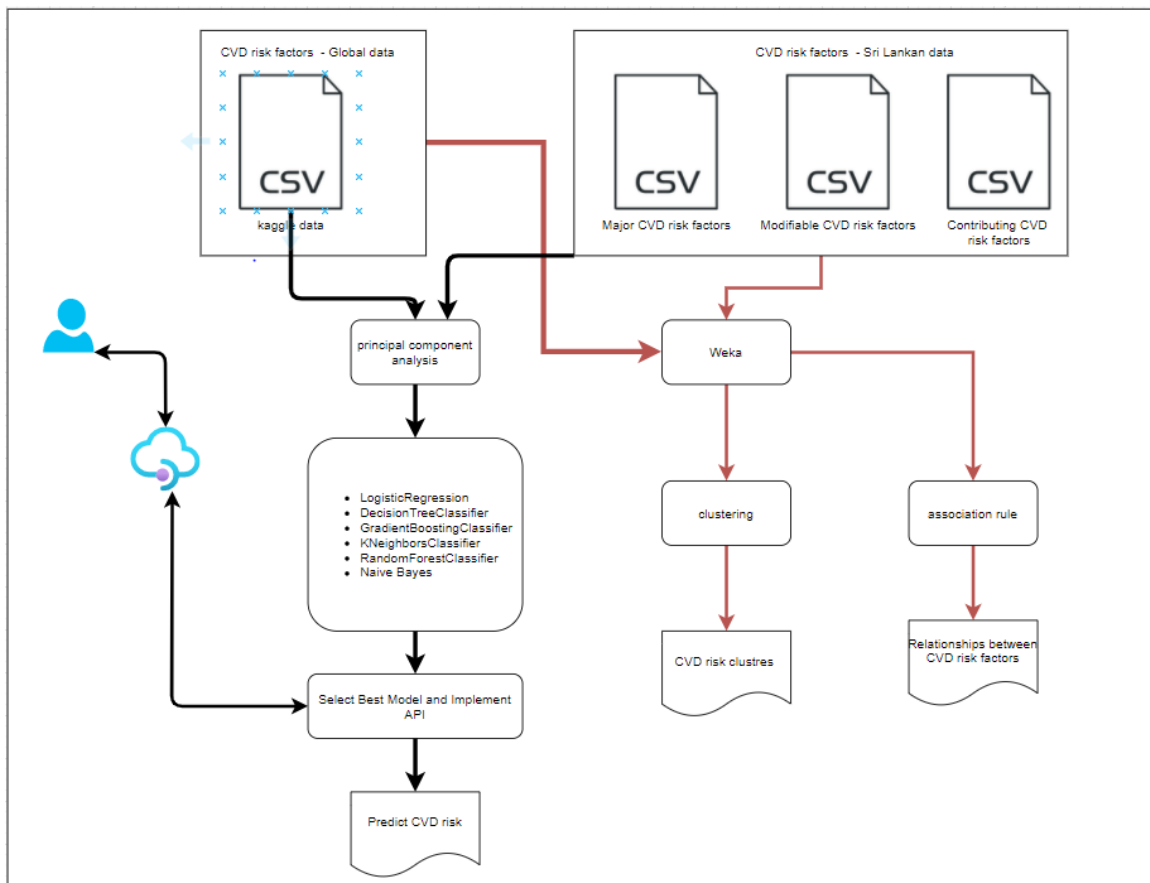


Figure 2 : Overall system architecture of the research

## **3.1 Phase 1**

### **3.1.1 Data Accessibility**

In the initial stage openly available dataset in Kaggle.com is using to identify most critical CVD risk factors and calculate the CVD risk. The dataset consists with 99 000 patient data records with 12 features including the class label (Ulianova, 2019).

### **3.1.2 Data Pre-processing**

The downloaded dataset contained outliers for systolic blood pressure, diastolic blood pressure and the diabetic values. The WEKA tool used to eliminate the outliers from the dataset. First the unsupervised learning attribute filter “InterquartileRange” applied to the dataset to identify the outliers and extreme values. When the filter applied two new attributes added to the attribute list in WEKA as outliers (attribute 13) and extreme points (attribute 14). Outlier attribute identified number of outliers available in the dataset. The Extreme point attribute displayed number of extreme points available in dataset. The “RemoveWithValues” instance filter used next to remove the outliers and the extreme points. Before applying the “RemoveWithValues” filter the generic object editor of the filter used to change the “attributeIndex” attribute need to be remove. The attribute needs to be remove changed to 13 (outlier attribute) and “nominalIndices” value changed to last to indicate to remove only the instances with outliers. Then the processed dataset saved as CSV file to use in the data models.

### **3.1.3 Implementation of CVD Risk Prediction Model**

#### **Attribute identification**

The focus of implementing the CVD risk prediction model is to identify the critical CVD risk factors and predict the occurrence of CVD risk. The pre-processed dataset consists of 12 attributes excluding the ID. The attribute list is given below,

- Age (in days)
- Gender (1 - women, 2 – men)
- Height (in cm)

- Weight (in kg)
- ap\_hi (Systolic blood pressure)
- ap\_lo (Diastolic blood pressure)
- Cholesterol (1: normal, 2: above normal, 3: well above normal)
- Gluc (Glucose level - 1: normal, 2: above normal, 3: well above normal)
- Smoke (whether the patient smokes or not)
- Alco (whether the patient consume alcohol or not)
- Active (whether the patient physically active or not)
- Cardio (whether the patient suffering from CVD or not / Target variable)

The CVD prediction model was implemented using python. The Scikit-learn library (Sklearn) which includes classification, regression, clustering and dimensionality reduction related machine learning and statistical modeling was used in along with python to do the implementation. The pre-processed dataset was loaded to Spyder IDE, saved as a dataframe and dropped the index column. `isnull()` function used to check whether is there any null values available in the dataset and it indicated that null values not existing in the dataset.

### **Feature selection, splitting data and scaling**

The columns in the data frame divided into two types of variables dependent variable (target variable) and independent variables (feature variables). The target variable is cardio (presence of CVD or not) and feature variables are the rest of the variables. Then to get understand about model performance, the data frame divided into training and testing set. The `train_test_split()` function available in the sklearn library.

Then used scikit learn preprocessing function called `StandardScaler()` to scale each feature to given range on the training set The scaled features were passed through `fit_transform()` method to center the data by calculating zero mean and unit standard error ( $x' = (x - \mu) / \sigma$ ) and apply the calculated transformation to particular set of features. The features which went through `fit_transform()` function saved to two numpy arrays `X_train` and `X_test`

## Build models

This step used predefined classifiers by the Sklearn library to build the models. Therefore, LogisticRegression, DecisionTreeClassifier, GradientBoostingClassifier, KNeighborsClassifier, RandomForestClassifier and GaussianNB classifier was imported.

To satisfy the need of determine best model for CVD prediction, within a for() loop creation of classifier object, train the classifier and predicting the response for test dataset was implemented. Above classifiers was selected for model building as they shows highest accuracy in CVD prediction during the literature survey.

## Evaluate model

To evaluate the accuracy of each model KFold cross validation which split data frame into 10 consecutive folds and return array of scores related to the model such as precision, recall, f1-score, support and accuracy. cross\_val\_score() function used with in the for() loop to get an array of score of the estimator for each run of the cross validation. The codes segment used for building and evaluate model is given using figure 3.

```
81
82 results = []
83 names = []
84 for name, model in models:
85     kfold = KFold(n_splits=10, random_state=None)
86     cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring='accuracy')
87     results.append(cv_results)
88     names.append(name)
89     msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
90     print(msg)
91
92     modell=model
93     modell.fit(X_train,y_train)
94     pred_y=modell.predict(X_test)
95
96     cm = metrics.confusion_matrix(y_test, pred_y)
97     plt.figure(figsize=(9,9))
98     sns.heatmap(cm, annot=True, fmt=".3f", linewidths=.5, square = True, cmap = 'Blues_r');
99     plt.ylabel('Actual label');
100    plt.xlabel('Predicted label');
101    all_sample_title = str(model)+' Confusion Matrix - score:'+str(metrics.accuracy_score(y_test,pred_y))
102    plt.title(all_sample_title, size = 15);
103    plt.show()
104    print(metrics.classification_report(y_test,pred_y))
105
106
107
```

Figure 3 : Code segment for build and evaluate model

The accuracy scores obtained from each model is given below. The output indicates that the accuracy of the models less than 85%

Logistic regression - 72%

Decision Tree Classifier – 81%

Gradient Boosting Classifier – 72%

Kneighbors Classifier – 71%

Random Forest Classifier - 83%

GaussianNB classifier - 70%

Principal Component Analysis (PCA) is a statistical approach which is capable of summarize substantial data tables and convert them into smaller set of “summary indicators” which support to easily visualize and analyze data. Therefore, it was decided to use PCA in this research work to represent multivariate CVD dataset as a smaller set of variables (education of the dimensionality) to observe the trends in CVD occurrence with high accuracy.

### **3.1.4 Implementation of Principal Component Analysis**

Principle Component Analysis (PCA) use widely for dimensionality reduction and make large datasets easily visualize and explore. Another advantage is PCA allows machine learning or datamining algorithms run faster by reducing the computational complexity. As literature says that the reduction of dimensionality might reduce the accuracy of the model below implementation took place to check the effect of PCA in datamining model building (Tripathi, 2020).

#### **Standardization**

Standardize the data with mean =0 and variance = 1 is the first step of implementing PCA. As the standardization already took place in the previous section the standardized data directly used here. Next it is required to obtain the eigenvectors and eigenvalues through the covariance matrix, sort eigenvalues in descending order and chose the top k eigenvectors, develop projection matrix from the selected k eigenvectors and obtain the new feature subsets. All the above steps were implemented using PCA object created by PCA class provide by sklearn. The sklearn PCA object capable of fit and convert/ transform the data frame into principle components.

## Apply PCA

The PCA object created from PCA() class called the fit\_transform() function to calculate the weights for principal components and transform the weighted components into a new set of principal components. The transformed principal components were saved in a data frames called “X\_train” – the training data frame and “X\_test” – the testing data frame.

The explain\_variance\_ratio\_ parameter in sklearn returned vector which explained by each dimension. The returned vector was saved in to a variable named as “explained\_variance. The implementation of PCA is given in below figure 4

```
from sklearn.decomposition import PCA

pca = PCA()
X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)
sssl=X_train

explained_variance = pca.explained_variance_ratio_
```

Figure 4: Implementation of PCA

The next step is to determine number of principle components for new feature subset. As the explained\_variance contains amount of information or the variance that could be attributed for each of selected principal components. The returned explained\_variance is given in figure 5.



The variance explained by the selected components in the above figure in in decreasing order. According to the figure it indicates that the cumulative variance of all the 11 features of the dataset contribute 94% of total variance (cumulative variance = 0.94). Thus, it is possible to imply that every feature in the dataset need to use for when projecting the new feature space. It can conclude that the all the 11 features are required to predict the CVD risk.

Figure 5: Explained variance of each attribute

Therefore, the 11 features of the dataset were applied for both training set and testing sets for the transform. The PCA transformed “X\_train” and “X\_test” values were used on top of the classification models to check the model accuracy. The mapping of new feature space to the training and testing variable given in the figure 6 below

```

Find_Most_Suitable=abs( pca.components_ )

from sklearn.decomposition import PCA

pca = PCA(n_components=11)
X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)

```

Figure 6 : Mapping training and testing sets in to new feature set

The classification accuracy of each model before and after apply PCA is given below in the table 4

Table 4 : Comparison of classification accuracy before and after applying PCA

		Accuracy
Logistic regression	Before PCA	71%
	After PCA	72%
Decision tree Classifier	Before PCA	80%
	After PCA	81%
Gradient Boosting Classifier	Before PCA	70%
	After PCA	72%
Kneighbors Classifier	Before PCA	71%
	After PCA	71%
Random Forest Classifier	Before PCA	83%
	After PCA	85%
GaussianNB classifier	Before PCA	70%
	After PCA	71%

The comparison illustrates that the accuracy values before and after applying PCA are almost equal. As stated by (Tripathi, 2020), due to the features of the dataset and number of rows, PCA dimension reduction method is not performing well in this situation. According to Laurae (Laurae, 2017) if PCA is applied on top of correlation robust algorithms such as Random Forest, applying PCA might not enhance the accuracy of the classifiers significantly.

But as PCA allows datamining algorithms to run faster by reducing the computational complexity, for rest of the implementations the PCA applied sub feature sets was used.

According to the comparison above it could conclude that the best classification algorithm for CVD prediction is Random Forest algorithm. There for to determine the CVD risk for any individual it can be recommended to use Random Forest algorithm.

### **3.1.5 Identify the Influence of CVD Risk Factors for Betterment of the Life**

The previously used globally available dataset was used to determine possible associations among CVD risk factors and the clusters that are high risk to get into the CVD risk. The CVD risk clusters, and association identification might lead to give awareness to the people who is not careering about their lifestyle and their habits

#### **Cluster analysis**

The WEKA tool was used to identify the CVD risk and non-risk clusters. The CVD dataset with 11 features and 1 target variable loaded to the WEKA tool. To preprocess the data for use it in clustering, all the attributes were converted into nominal variables using “NumericToNominal” unsupervised, attribute filter. In the cluster tab, the clustering algorithm selected as simpleKmeans algorithm and apply the algorithm on top of the dataset to determine the clusters related to CVD risk.

#### **Association rule mining**

To identify the association between CVD risk factors, the research used the association rule mining in WEKA tool. Apriori algorithm in the WEKA associator used to identify possible associations between the CVD risk factors.

## **3.2 Phase 2**

The second phase is focusing on predicting CVD risk in Sri Lankan domain. The CVD risk factors in Sri Lanka collected via a questionnaire which developed with the support of literature survey and domain experts (ResearchGate feedbacks). In the second phase data collected from



individuals who are currently suffering from CVD, not suffering from CVD and had slight pain in heart (moderate risk for getting into CVD). Due to current travel restrictions and the prevailing conditions, data collected from 1252 individuals with their consent for the research.

The collected data refined and loaded to the classifier model implemented in first phase to predict the CVD risk. And the same data used to perform cluster analysis to detect the user groups with high CVD risk and to determine the correlation with the CVD risk factors to prevent new CVD cases in the future.

### **3.2.1 Data Gathering and Pre- processing**

Prepared a questionnaire with the support of domain experts and online available valid questionnaires (“CVDQuestionnaire.pdf,” n.d.) to detect CVD risk factors in Sri Lankan context. The prominent CVD risk factors was identified via literature survey to develop the initial questionnaire. The initial questionnaire is attached as Appendix 1.

The initial questionnaire was published in the ResearchGate to gather the feedback of domain experts. Few feedbacks obtained from domain experts attached below using figure 7. The domain experts suggested to add following factors also to the questionnaire to get an efficient output.

High lipid profile, hyperglycemia, atherothrombotic disease

Chowdhury Meshkat Ahmed, a cardiologist from India suggested to include coconut milk and coconut oil consumption in Sri Lanka as high cholesterol content in coconut may related to the CVD risk in Sri Lanka.

The updated questionnaire based on the feedbacks converted into google form and circulated among the Sri Lankan individuals for the data collection.

The google form consists of 51 features related to major modifiable and contributing risk factors.

- Major risk factors : Age, gender, family history
  
- Modifiable risk factors : Diagnosed for high blood pressure, cholesterol, diabetics, HBCA<sub>1</sub> value, duration of diagnosis for each disease, under medication for each disease

- Contributing risk factors : Physical inactivity, smoking, alcohol consumption and amount consume, living environment, consume betel with tobacco, sleeping hours and sleeping difficulties, Occurrence of different life events such as death of spouse, marriage, working from home etc. , involvement in relaxation activities, presence of anxiety of depression feelings, bowel toxicity, consume oral contraceptive pill or antibiotics, food consumption habits such as use of fried foods, salt, sweet, soft drinks, coconut milk etc. , presence of inflammation/ pain and the occurring frequency.
- Target variable is the presence of cardiovascular disease, atherosclerosis, previous heart attack, stroke or PVD.

**William Feeman** added an answer

April 19

Dear Dr Awuchi,

Smoking is the clearly dominant ATD risk factor, followed closely by dyslipidemia (as I defined it) and more distantly by hypertension and high blood sugar levels. In the absence of smoking and dyslipidemia, early onset ATD is uncommon.

Mark as best answer
Recommended
Share

2 Recommendations

---

**Weiguo Zhang** added an answer

April 20

In a nutshell, established CV risk factors are:

1. Increasing Age (non-modifiable)
2. Male gender (non-modifiable)

... [Read more](#)

Mark as best answer
Recommend
Share
▼

2 Recommendations

---

**Chowdhury/Chaudhury Meshkat Ahmed** added an answer

April 20

Dear William Feemen, Sir and Dear Weiguo Zhang I had been through the questionnaire of Narmada Gamage. She has covered all most all the established risk factors for atherosclerosis. I have also suggested to add some emerging minor risk factors. Her project is about to form an IT base in order to establish a genetic link through nutrition to atherosclerosis in Sri Lanka.

Williams Feemen, Sir, it is truly useful to know, the order of he importance of individual risk factors from you.

I have little observation as this regards. Diabetes has differential Cardiovascular rate in different ethnic group. So from our part of world Diabetes is one of the top listed risk factor.

Dear Narmada Gamage, pl give us your feedback, if you need any more clarification.

---

**Chowdhury/Chaudhury Meshkat Ahmed** added an answer

April 27

Dear Narmada Gamage,

Much can be learnt about the behavioural risk factors for atherosclerosis from the link given Frank T. Edelmann. I want to draw your attention about a particular food with much consumption in Sri Lanka , that is coconut. Coconut is thought to high in cholesterol content and its high consumption may be related to the cardiovascular events. You can consider this separately as a risk factor. Interestingly it also increase the HDL C which is a protective cholesterol. So you may get interesting data from coconut consumption.

Figure 7: Feedback obtained from domain experts for initial questionnaire

The collected data was pre-processed by removing null values and assigning labels for each selection. The data collected via google form and the label assigned for each feature is given in the appendix 2. Example label assignment is given below.

Ex :

<b>How many serves of bread, pasta, rice, potatoes or other starchy foods do you have a day? *</b>	0 – 1 serves daily	2 serves daily	3 serves daily	4 or more serves daily
<b>Label assigned</b>	0	1	2	3

The outliers of the collected dataset was removed using WEKA tunsupervised learning attribute filter “InterquartileRange” and the “RemoveWithValues” instance filter as explained under section 3.1.1

Then the pre-processed dataset saved as a separate CSV file to use it in the CVD risk prediction.

### 3.2.2 Implementation of CVD Risk Prediction Model with PCA

The literature survey highlighted that there is no mechanism to predict CVD risk in Sri Lanka which consider local food habits, lifestyle and mental conditions. Therefore, the dataset split into three separate datasets,

Dataset 1 - major CVD risk factors along with the target variable

Dataset 2 - modifiable risk factors along with the target variable,

Dataset 3 - contributing risk factors along with the target variable to provide efficient classification results. As there are 3 datasets 3 different models implemented to load the 3 datasets.

The steps followed for the implementation of the classification models were similar to the model implementation with PCA described under 3.1.3 and 3.1.4. The feature selection and splitting data with train\_test\_split() function, scaling with StandardScaler() and fit\_transform(), apply PCA with PCA(), build and evaluate model with sklearn classifiers are the major steps followed in the implementation.

Logistic regression, Decision Tree Classifier, Gradient Boosting Classifier, Kneighbors Classifier, Random Forest Classifier and GaussianNB classifier are the classifiers used to perform CVD risk prediction. Table 5 contains accuracies obtained for each experiment. Accuracy values lead to identify the most suitable, accurate CVD prediction among wide used classification algorithms.

Major risk factors only contained 3 variables as gender, age and CVD family history. Due to less number of features, model displayed less accuracy in each model. To overcome the situation the major and modifiable risk factors combined to create totally 4 datasets.

Dataset 4 - major and modifiable risk factors along with the target variable

Table 5 : Accuracy comparison for CVD prediction datasets

	Accuracy of CVD prediction for major risk factors	Accuracy of CVD prediction for modifiable risk factors	Accuracy of CVD prediction for contributing risk factors	Accuracy of CVD prediction for major + modifiable risk factors
Logistic regression	59%	62%	69%	62%
Decision Tree Classifier	69%	100%	100%	100%
Gradient Boosting Classifier	69%	98%	100%	97%
Kneighbors Classifier	65%	77%	81%	80%
Random Forest Classifier	69%	100%	100%	100%
GaussianNB classifier	60%	60%	57%	60%

The table indicate that decision tree and random forest classifiers show better performance when predicting the CVD in Sri Lankan context. Therefore, it is possible to recommend these algorithms to use when predicting CVD risk.

### 3.2.3 Identify the Influence of CVD Risk Factors for Betterment of Sri Lankans' Life

To identify the influence of CVD risk factors for betterment of Sri Lankan's life the datasets introduced in the section 3.2.2 were used in cluster analysis and association rule mining. Following section describe the methodology used to perform cluster analysis and association rule mining.

#### Cluster analysis

The WEKA tool was used to identify the CVD risk and non-risk clusters. The 4 datasets created in the previous section 3.2.3. Features and target variable for each dataset is given below. The target variable in each dataset is given in Bold font style. The target variable contains 3 class labels as high CVD risk, moderate CVD risk and no CVD risk.

- Dataset 1 (Major CVD risk factors) :
- 3 features and 1 target variable (Age, Sex, Family, **Cardio**)

Dataset 2 (Modifiable CVD risk factors) :

- 17 features and 1 target variable
- (BMI, Diagnosed for Diabetes, Diabetic Level, HbA1c, Diagnose Period for diabetics, Medication for diabetics, Digesting issues, Diagnose for blood pressure, Systolic blood pressure, Diastolic blood pressure, Diagnose Duration, Medication for blood pressure, Diagnose for Cholesterol, Cholesterol Level, Diagnose for cholesterol, Medication for Cholesterol, Exercise, **Cardio**)

Dataset 3 (Contributing risk factors)

- 29 features and 1 target variable
- Smoking, Passive Smoking, Alcohol consumption, Amount of Alcohol Consumption, Living environment, Betel Consumption, Sleep Hours, Sleep Issues, special events, relaxation activities, Stress, Bowel Toxicity, consume Contraceptive Pills, Consume Antibiotic, Antibiotic Duration, Food habits – Fried Food, Carbohydrate, Sweet Food Sugar, Salt, Fish, Fruit, Vegetable, Coffee, Soft-Drink, Water, Coconut Milk, Anxiety, Anxiety Frequency, **Cardio**)

Dataset 4 (Major and modifiable CVD risk factors) :

- 20 features and 1 target variable

Above mentioned CVD datasets loaded to the WEKA tool separately. To preprocess the data for use it in clustering, all the attributes were converted into nominal variables using “NumericToNominal” unsupervised, attribute filter. In the cluster tab, the clustering algorithm selected as simpleKmeans algorithm and apply the algorithm on top of the datasets to determine the clusters related to CVD risk.

### Association rule mining

To identify the association between CVD risk factors, the research used the association rule mining in WEKA tool. Apriori algorithm in the WEKA associator used to identify possible associations between the CVD risk factors in Sri Lankan context.

### 3.3 User Interaction with the Model

Python based API created to connect with the best CVD predictions classifier obtained in section 3.1.4 and 3.2.2. Using a simple GUI (created using visual studio) users can insert there major, modifiable and contributing risk factors into 3 forms. The best classifier takes the user inputs and returns the CVD risk related to the given user data.

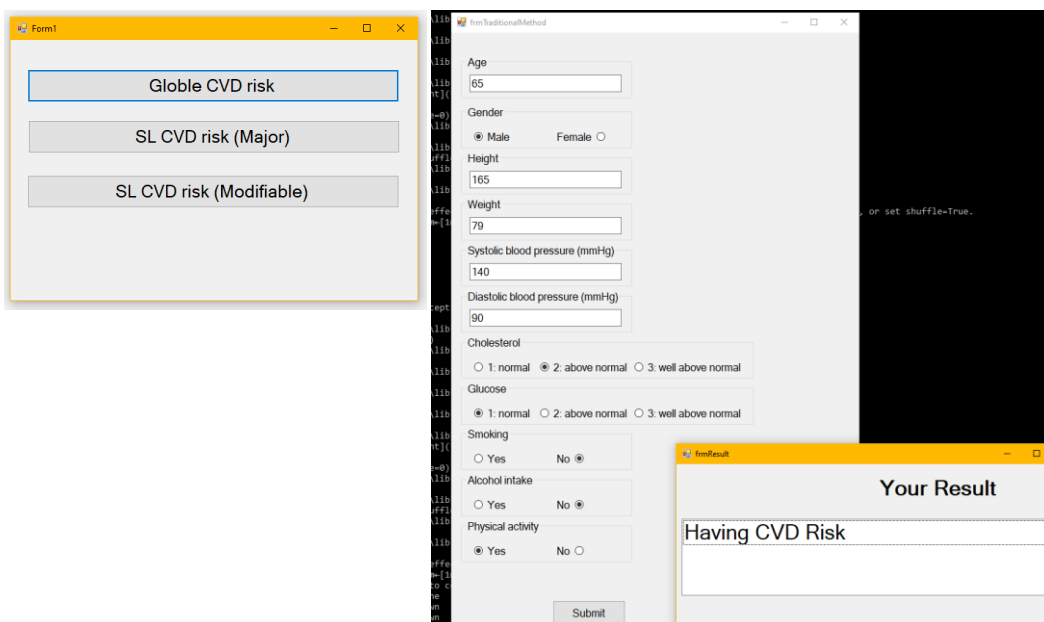


Figure 8 : User interface to display CVD risk prediction

The first interface allows the users to select the CVD risk calculator, CVD risk prediction based on Sri Lankan dataset or global dataset. After that user redirect to an interface to provide their personal information, medical history and lifestyle information related to CVD prediction. By using the best classifier the CVD risk ( whether high, moderate or no CVD risk) displayed to the users in via a pop-up window. Figure 8 contains the CVD prediction user interface when user selected global data-based CVD prediction along with the prediction CVD occurrence prediction

### 3.4 Tools and Technologies

Below table 6 contains the software and hardware requirements related to the research work

Table 6: Software requirements for the proposed model

<b>Technology</b>	<b>Technology Software tools</b>
<b>Python:</b> High-level, general-purpose programming language. Python contains inbuilt data mining and machine learning-based modules	Anaconda Navigator with Spyder IDE
<b>Asp.net (C#):</b> Self-contained, server-side scripting language,	Visual studio
	WEKA

### **3.5 Limitations**

- As suggested by the supervisor and the proposal presentation evaluation panel project was divided in to two phases to considering CVD risk in Sri Lankan context. Research was planned to collect clinical data from CVD risk individuals. Due to COVID 19 pandemic it was not permitted to visit hospitals to collect information. Therefore, by considering the recommendations of supervisor and the cardiologist from Peradeniya teaching hospital Dr. Dhaminda CVD related data collected online using a online questionnaire after gaining user consent.
- When the Google sheet populated to gather CVD risk factors in Sri Lankan context via social media platforms majority of individuals between 20 – 35 age categories intended to fill the google form. The presence of CVD was to capture in the collected data. Therefore, google form shared among personal contacts and asked to gather the information from their parents, or any relative with CVD risk. At the same time printed datasheets distributed among CVD patients and gathered the data. Therefore, it took longer time than expected for the data collections.



# CHAPTER 4

## RESULTS AND EVALUATIONS

This chapter specifies in detail, the results from the experiment conducted to obtain CVD prediction with high accuracy. Further, the experimental results are evaluated based on multiple criteria's and against the goals and objectives of the study.

### 4.1 Experimental Results of Phase 1

#### 4.1.1 Results obtained in phase 1 to achieve objective 1

This research focused to achieve below 3 objectives during the phase 1 implementation.

(Section 1.3.2 list down 5 objectives for entire research. This section derives objective 1,2 and 5 from the section 1.3.2 and renumbered as 1, 2 and 3 for the explanation purpose)

1. To identify the critical CVD risk factors among major, modifiable and contributing risk factors listed in existing research works.
2. To predict the CVD risk by using identified critical major and modifiable risk factors after evaluating the accuracy of well-known data mining algorithms.
3. To test the influence of CVD risk factors for the betterment of the life

The dataset used in the phase I had 11 global CVD risk factors along with 1 target variable. To determine the critical risk factors by reducing the dimensionality PCA implemented on top of classifiers. But according to the figure 4, explained variance vector, it shows cumulative variance of 11 features account for 94% variance in the dataset.

Number of components	Logistic regression	Decision tree	Gradient boosting	K neighbors	Random Forest	Naïve bayes
1	0.66	0.79	0.66	0.66	0.79	0.68
2	0.71	0.8	0.71	0.7	0.83	0.73
3	0.71	0.81	0.71	0.71	0.84	0.7
4	0.71	0.81	0.71	0.71	0.84	0.7
5	0.71	0.81	0.71	0.71	0.85	0.7
6	0.71	0.81	0.72	0.71	0.85	0.7
7	0.71	0.81	0.72	0.71	0.85	0.7
8	0.73	0.81	0.72	0.71	0.85	0.69
9	0.72	0.81	0.72	0.71	0.85	0.72
10	0.72	0.81	0.72	0.71	0.85	0.72
11	0.72	0.81	0.73	0.72	0.85	0.69

Figure 9 : Accuracy classifiers with different PCA values.

Reduction of features/ dimensions leads to reduce the accuracy of the classifiers. The model run multiple iterations by incrementing components one by one to observe the accuracy of the classifiers. The result obtained is given by figure 9.

Chapter 3 depicts that the data should be pre-processed and fed into multiple classification algorithms. As the initial stage missing values were removed and the dataset was fed into logistic regression, decision tree classifier, gradient boosting classifier, k-neighbors classifier and random forest classifiers. Figure 4 depicts the accuracy values obtained using python IDE

- Majority of the classifiers except Gaussian NB all the other classifiers showed highest accuracy when the number of components in the PCA is equals to 11. Therefore, I decided Continues the CVD prediction model implementation with 11 PCA components.
- According to table 3 in chapter 3, the accuracy of the classifiers not increased when using the PCA. But literature stated that PCA have the capacity to allows datamining algorithms to run faster by reducing the computational complexity (Tripathi, 2020). Therefore, it decided to use PCA along with classifiers to predict the CVD risk
- Based on the PCA experiment it is possible to decide that each feature is important to predict the CVD risk.

#### **4.1.2 Results obtained in phase 1 to achieve objective 2**

The precision, recall, F1, support and accuracy values obtained for each classifier used in the implementation given in below figure 10.

The sklearn function “`metrics.classification_report(y_test,pred_y)`” is used in the implementation to print all the above values. According to the accuracy scores in figure 10 it shows that the highest accuracy in predicting CVD is with the Random Forest. Therefore, it can conclude that the Random Forest is best classifier in predicting CVD. The decision tree classifier shows the next best performance while logistic regression and KNeighbors classifiers and next in line. Lowest performance showed by gradient booting classifier and Naive Bayes classifiers.

- By considering the above scenario it is decided to use Random Forest algorithm in CVD risk prediction when any individual willing to check the CVD risk.

```

-----LogisticRegression-----
precision    recall  f1-score   support

     1       0.70      0.77      0.73     10070
     2       0.74      0.67      0.70     9930

 accuracy          0.72     20000
 macro avg         0.72      0.72      0.72     20000
 weighted avg      0.72      0.72      0.72     20000

-----KNeighborsClassifier-----
precision    recall  f1-score   support

     1       0.71      0.72      0.72     10070
     2       0.71      0.70      0.71     9930

 accuracy          0.71     20000
 macro avg         0.71      0.71      0.71     20000
 weighted avg      0.71      0.71      0.71     20000

-----DecisionTreeClassifier-----
precision    recall  f1-score   support

     1       0.81      0.82      0.82     10070
     2       0.81      0.81      0.81     9930

 accuracy          0.81     20000
 macro avg         0.81      0.81      0.81     20000
 weighted avg      0.81      0.81      0.81     20000

-----RandomForestClassifier-----
precision    recall  f1-score   support

     1       0.85      0.85      0.85     10070
     2       0.85      0.85      0.85     9930

 accuracy          0.85     20000
 macro avg         0.85      0.85      0.85     20000
 weighted avg      0.85      0.85      0.85     20000

-----GradientBoostingClassifier-----
precision    recall  f1-score   support

     1       0.71      0.76      0.73     10070
     2       0.74      0.69      0.71     9930

 accuracy          0.72     20000
 macro avg         0.72      0.72      0.72     20000
 weighted avg      0.72      0.72      0.72     20000

-----Naive Bayes-----|
precision    recall  f1-score   support

     1       0.71      0.72      0.72     10070
     2       0.71      0.70      0.71     9930

 accuracy          0.71     20000
 macro avg         0.71      0.71      0.71     20000
 weighted avg      0.71      0.71      0.71     20000

```

Figure 10 : Accuracy scores returned by classification\_report() function

### 4.1.3 Results obtained in phase 1 to achieve objective 3

#### Cluster analysis

Identification of CVD risk user groups and the CVD non risk user group might support the medical sector to reach and treat the correct individuals. The simple Kmeans algorithm used in WEKA clustering resulted 10 clusters with and without CVD risk. The resulted output given in figure 11.

```

Final cluster centroids:
Attribute      Full Data      Cluster#
                (99999.0)    (6910.0) (27879.0) (7266.0) (4772.0) (8907.0) (7828.0) (6049.0) (12149.0) (10521.0) (7718.0)
-----
age            56.02         56.05         50.04         47.93         57.64         52.04         58.08         57.92         56.08         54.22         60.07
gender         1             1             1             2             1             1             2             1             2             2             1
height        165          158          170          168          156          160          170          168          160          168          165
weight        65           65           75           74           60           65           80           80           78           70           65
ap_hi         120          120          120          110          120          110          140          130          140          120          120
ap_lo         80           80           80           70           80           70           80           80           90           80           80
cholesterol   1            2            1            1            1            1            2            3            1            1            1
gluc          1            2            1            1            1            1            1            3            1            1            1
smoke         1            1            1            1            1            1            1            1            1            1            1
alco          1            1            1            1            1            1            1            1            1            1            1
active        2            2            2            2            2            2            2            2            2            2            1
cardio        No           No           No           No           No           No           Yes          Yes          Yes          No           Yes

Time taken to build model (full training data) : 0.73 seconds

=== Model and evaluation on training set ===

Clustered Instances

0          6910 ( 7%)
1          27879 ( 28%)
2          7266 ( 7%)
3          4772 ( 5%)
4          8907 ( 9%)
5          7828 ( 8%)
6          6049 ( 6%)
7          12149 ( 12%)
8          10521 ( 11%)
9          7718 ( 8%)

```

Figure 11 : Clusters related to global CVD risk factors

#### Glossary for cluster analysis

- Gender (1 - women, 2 – men), Height (in cm), Weight (in kg), Systolic blood pressure, Diastolic blood pressure, Cholesterol (1: normal, 2: above normal, 3: well above normal), Gluc (Glucose level - 1: normal, 2: above normal, 3: well above normal) , Smoke (Not smoking -1 , Smoking – 2) , Alco (consume alcohol -1 , not consume alcohol – 2) , Active (Not active -1 , Active – 2), Cardio (whether the patient suffering from CVD – Yes)

According to the percentage of instances in each cluster majority of the people belongs to cluster 1, cluster 7 and cluster 8. The definitions for those clusters are given below

Cluster 1: Females in in mid-50's whose BMI is marginal overweight range (BMI = 23.6 when height = 165cm and weight = 65kg), having normal blood pressure value 120/80 mmHg, if they are not smoking and not consuming alcohol even though they are having cholesterol and diabetics above normal value when they are physically active there is a less risk to getting into CVD

Cluster 7: Males in mid-50's (56 years old) whose BMI is in overweight range (BMI = 28.7 when height = 165cm and weight = 78kg), having high blood pressure value 140/90 mmHg, if they are not smoking and not consuming alcohol even though they are having cholesterol and diabetics above normal value when they are physically active there is a less risk to getting into CVD

Cluster 8: Males in mid-50's (54 years old) whose BMI is in normal range (BMI = 24.8 when height = 168cm and weight = 70kg), having normal blood pressure value 120/80 mmHg, and don't have cholesterol, diabetics, or not smoking and not consuming alcohol when they are physically active, they are not having CVD risk.

But according to the cluster, if a 60 year old female whose BMI is in normal range (BMI = 23.6 when height = 165cm and weight = 65kg), having normal blood pressure value 120/80 mmHg, if they are not smoking and not consuming alcohol or not having cholesterol and diabetics still if they are physically inactive there is a risk to getting into CVD

- By considering the clusters generated, in WEKA it is possible to identify individuals with high or low CVD risk. The majority of the clusters highlights that even though the blood pressure values is slightly above normal (normal = 120/80 mmHg), having diabetics above normal (normal fasting blood sugar = 70–99 mg/dl (3.9–5.5 mmol/L) or having cholesterol above normal (normal total cholesterol = 125 to 200mg/dL) still if it is possible to be physically active, it will reduce the CVD risk.
- All the clusters prove that if any individual does not smoke or not consume alcohol it will reduce the CVD risk

## Association rule mining

Identification of relationships between CVD risk factors might support the community to fix their bad habits.

The results of initial association rules analysis 11 features represented in figure 12.

```
Apriori
=====

Minimum support: 0.7 (69999 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5
Size of set of large itemsets L(2): 6
Size of set of large itemsets L(3): 2

Best rules found:

1. gluc=1 smoke=1 77605 ==> alco=1 75479 <conf:(0.97)> lift:(1.03) lev:(0.02) [1994] conv:(1.94)
2. smoke=1 91295 ==> alco=1 88593 <conf:(0.97)> lift:(1.02) lev:(0.02) [2145] conv:(1.79)
3. smoke=1 active=2 73148 ==> alco=1 70876 <conf:(0.97)> lift:(1.02) lev:(0.02) [1612] conv:(1.71)
4. cholesterol=1 74945 ==> alco=1 71384 <conf:(0.95)> lift:(1.01) lev:(0) [418] conv:(1.12)
5. gluc=1 85021 ==> alco=1 80650 <conf:(0.95)> lift:(1) lev:(0) [143] conv:(1.03)
6. active=2 80415 ==> alco=1 75928 <conf:(0.94)> lift:(1) lev:(-0) [-216] conv:(0.95)
7. gluc=1 alco=1 80650 ==> smoke=1 75479 <conf:(0.94)> lift:(1.03) lev:(0.02) [1848] conv:(1.36)
8. alco=1 94689 ==> smoke=1 88593 <conf:(0.94)> lift:(1.02) lev:(0.02) [2145] conv:(1.35)
9. alco=1 active=2 75928 ==> smoke=1 70876 <conf:(0.93)> lift:(1.02) lev:(0.02) [1556] conv:(1.31)
10. gluc=1 85021 ==> smoke=1 77605 <conf:(0.91)> lift:(1) lev:(-0) [-15] conv:(1)
```

Figure 12 : Best rules generated by Apriori association rule

First 5 rules out of 10 best rules found using WEKA – apriori algorithms shows 95% confidence. The interpretation for first 5 rules is given below,

- If the glucose level is normal and the person is not smoking, with 97% confidence it is possible to state that the person does not consume alcohol
- If a person is nonsmoker with 97% confidence it is possible to state that the person is nonalcoholic.
- If person is a nonsmoker and physically active with 97% confidence it is possible to state that the person is nonalcoholic.
- If a person maintains normal cholesterol level with 95% confidence it is possible to state that the person is a non-alcoholic
- If the glucose level is normal with 95% confidence it is possible to state that the person does not consume alcohol

The rules having confidence greater or equal to 95% sums up that if a person maintain better glucose, cholesterol levels and not smoking, then the alcohol consumption of the person is rare condition. It leads to the assumption, if it is possible to avoid alcohol consumption, smoking habits the cholesterol and glucose levels will be normal. This will eventually reduce the CVD risk.

## 4.2 Experimental Results of Phase 2

This research focused to achieve below 3 objectives during the phase 2 implementation. (Section 1.3.2 list down 5 objectives for entire research. This section derives objective 3,4 and 5 from the section 1.3.2 and renumbered as 1, 2 and 3 for the explanation purpose)

1. To detect major, modifiable and contributing risk factors in Sri Lankan domain
2. To evaluate the accuracy of CVD predicting algorithms by applying Sri Lankan data
3. To test the influence of CVD risk factors for the betterment of the life

The figure 13 depicts the accuracy of CVD prediction when there are 51 features together

5 rows x 51 columns]					KNeighborsClassifier: 0.596000 (0.125793)					
LogisticRegression: 0.644000 (0.076837)					precision	recall	f1-score	support		
		precision	recall	f1-score	support	0	1.00	0.83	0.91	18
	0	0.87	0.72	0.79	18	5	0.63	0.59	0.61	29
	5	0.68	0.66	0.67	29	10	0.33	0.44	0.38	16
	10	0.40	0.50	0.44	16					
	accuracy			0.63	63				0.62	63
	macro avg	0.65	0.63	0.63	63	accuracy	0.65	0.62	0.63	63
	weighted avg	0.66	0.63	0.64	63	macro avg	0.66	0.62	0.64	63
						weighted avg				
DecisionTreeClassifier: 0.564000 (0.060531)					RandomForestClassifier: 0.608000 (0.111427)					
		precision	recall	f1-score	support	precision	recall	f1-score	support	
	0	0.68	0.83	0.75	18	0	0.93	0.78	0.85	18
	5	0.70	0.66	0.68	29	5	0.61	0.66	0.63	29
	10	0.50	0.44	0.47	16	10	0.29	0.31	0.30	16
	accuracy			0.65	63	accuracy			0.60	63
	macro avg	0.63	0.64	0.63	63	macro avg	0.61	0.58	0.59	63
	weighted avg	0.65	0.65	0.65	63	weighted avg	0.62	0.60	0.61	63
GradientBoostingClassifier: 0.608000 (0.100876)					Naive Bayes: 0.588000 (0.111786)					
		precision	recall	f1-score	support	precision	recall	f1-score	support	
	0	0.93	0.78	0.85	18	0	0.88	0.83	0.86	18
	5	0.62	0.62	0.62	29	5	0.63	0.83	0.72	29
	10	0.32	0.38	0.34	16	10	0.25	0.12	0.17	16
	accuracy			0.60	63	accuracy			0.65	63
	macro avg	0.62	0.59	0.60	63	macro avg	0.59	0.60	0.58	63
	weighted avg	0.63	0.60	0.62	63	weighted avg	0.61	0.65	0.62	63

Figure 13 : Accuracy scores returned by classification\_report() function for all the CVD risk factors collected from Sri Lankan context

## 4.2.1 Results obtained in phase 2 to achieve objective 1 and 2

According to the figure 13, all the classifiers illustrate accuracy less than 70%. Therefore, it was decided to split the dataset in to three separate sets referring to the CVD risk factors given by American heart association (Benjamin et al., 2019). The attributes each dataset consists with is given in section 3.2 above.

- From the collected 51 features, separate datasets were prepared by assuming major risk factors are the CVD factors which unable eliminate or reduce by changing the lifestyle, modifiable risk factors are the factors which can eliminate or reduced by changing the life style, contributing risk factors are the factors that might enhance the CVD risk but not yet confirmed by medical research. (Kumar et al., 2017),(Benjamin et al., 2019)

The datasets were separately input to the classifier with PCA implementation build in phase 1 to identify the CVD prediction accuracy. The accuracy scores return by `classification_report()` function is given below for each dataset. Figure 14 is the illustration of accuracy scores for major risk factors, figure 15 illustrates accuracy scores for modifiable risk factors, while figure 16 illustrates the accuracy values for contributing risk factors.

LogisticRegression: 0.607475 (0.050003)					KNeighborsClassifier: 0.665337 (0.040282)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.73	0.70	0.71	66	0	0.89	0.82	0.85	66
5	0.53	0.83	0.64	98	5	0.57	0.83	0.67	98
10	0.65	0.25	0.36	87	10	0.68	0.37	0.48	87
accuracy			0.59	251	accuracy			0.67	251
macro avg	0.63	0.59	0.57	251	macro avg	0.71	0.67	0.67	251
weighted avg	0.62	0.59	0.56	251	weighted avg	0.69	0.67	0.65	251
DecisionTreeClassifier: 0.725287 (0.020424)					RandomForestClassifier: 0.727287 (0.023494)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.87	0.82	0.84	66	0	0.87	0.82	0.84	66
5	0.59	0.85	0.69	98	5	0.59	0.84	0.69	98
10	0.73	0.40	0.52	87	10	0.72	0.41	0.53	87
accuracy			0.69	251	accuracy			0.69	251
macro avg	0.73	0.69	0.69	251	macro avg	0.73	0.69	0.69	251
weighted avg	0.71	0.69	0.67	251	weighted avg	0.71	0.69	0.67	251
GradientBoostingClassifier: 0.723317 (0.028542)					Naive Bayes: 0.607475 (0.052348)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.87	0.82	0.84	66	0	0.73	0.70	0.71	66
5	0.59	0.84	0.69	98	5	0.52	0.86	0.65	98
10	0.73	0.43	0.54	87	10	0.78	0.24	0.37	87
accuracy			0.69	251	accuracy			0.60	251
macro avg	0.73	0.69	0.69	251	macro avg	0.68	0.60	0.58	251
weighted avg	0.71	0.69	0.68	251	weighted avg	0.67	0.60	0.57	251

Figure 14 : Accuracy scores returned by `classification_report()` function for major risk factors

- According to the figure 14, decision tree, random forest and gradient booting classifier shows the highest accuracy. But the accuracy is less than 75% in each classifier. Therefore, it is decided to combine major risk factors and modifiable risk factors and prepare a separate dataset to get high accuracy in CVD prediction.



LogisticRegression: 0.693297 (0.040816)					KNeighborsClassifier: 0.750267 (0.033014)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.71	0.68	0.70	66	0	0.86	0.76	0.81	66
5	0.56	0.69	0.62	98	5	0.72	0.78	0.75	98
10	0.64	0.49	0.56	87	10	0.76	0.77	0.77	87
accuracy			0.62	251	accuracy			0.77	251
macro avg	0.64	0.62	0.63	251	macro avg	0.78	0.77	0.77	251
weighted avg	0.63	0.62	0.62	251	weighted avg	0.77	0.77	0.77	251
DecisionTreeClassifier: 0.987010 (0.011877)					RandomForestClassifier: 0.993010 (0.006399)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	1.00	1.00	66	0	1.00	1.00	1.00	66
5	1.00	1.00	1.00	98	5	1.00	1.00	1.00	98
10	1.00	1.00	1.00	87	10	1.00	1.00	1.00	87
accuracy			1.00	251	accuracy			1.00	251
macro avg	1.00	1.00	1.00	251	macro avg	1.00	1.00	1.00	251
weighted avg	1.00	1.00	1.00	251	weighted avg	1.00	1.00	1.00	251
GradientBoostingClassifier: 0.987020 (0.014168)					Naive Bayes: 0.669257 (0.052973)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	0.95	0.98	66	0	0.73	0.65	0.69	66
5	0.94	1.00	0.97	98	5	0.55	0.62	0.58	98
10	1.00	0.97	0.98	87	10	0.57	0.53	0.55	87
accuracy			0.98	251	accuracy			0.60	251
macro avg	0.98	0.97	0.98	251	macro avg	0.62	0.60	0.61	251
weighted avg	0.98	0.98	0.98	251	weighted avg	0.60	0.60	0.60	251

Figure 15: Accuracy scores returned by classification\_report() function for modifiable risk factors

- According to the figure 15, decision tree and random forest both can predict the CVD risk with 100% accuracy when the modifiable risk factors are known.
- The results obtained leads to a decision Random Forest classifier and decision tree classifiers can be used to predict CVD risk with high accuracy for modifiable CVD risk factors.
- But logistic regression, KNeighbors and Naïve bayes classifiers are not suitable in this condition as the accuracy of prediction is less than 80%

LogisticRegression: 0.648386 (0.031113)					KNeighborsClassifier: 0.726356 (0.058684)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.85	0.92	0.88	66	0	0.97	0.97	0.97	66
5	0.62	0.66	0.64	98	5	0.83	0.87	0.85	98
10	0.59	0.51	0.55	87	10	0.84	0.80	0.82	87
accuracy			0.68	251	accuracy			0.87	251
macro avg	0.69	0.70	0.69	251	macro avg	0.88	0.88	0.88	251
weighted avg	0.67	0.68	0.67	251	weighted avg	0.87	0.87	0.87	251
DecisionTreeClassifier: 0.992010 (0.009796)					RandomForestClassifier: 0.995010 (0.009214)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	1.00	1.00	66	0	1.00	1.00	1.00	66
5	1.00	1.00	1.00	98	5	1.00	1.00	1.00	98
10	1.00	1.00	1.00	87	10	1.00	1.00	1.00	87
accuracy			1.00	251	accuracy			1.00	251
macro avg	1.00	1.00	1.00	251	macro avg	1.00	1.00	1.00	251
weighted avg	1.00	1.00	1.00	251	weighted avg	1.00	1.00	1.00	251
GradientBoostingClassifier: 0.986010 (0.014969)					Naive Bayes: 0.608455 (0.040938)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	1.00	1.00	66	0	0.71	0.74	0.73	66
5	1.00	0.97	0.98	98	5	0.50	0.80	0.62	98
10	0.97	1.00	0.98	87	10	0.59	0.18	0.28	87
accuracy			0.99	251	accuracy			0.57	251
macro avg	0.99	0.99	0.99	251	macro avg	0.60	0.57	0.54	251
weighted avg	0.99	0.99	0.99	251	weighted avg	0.59	0.57	0.53	251

Figure 16: Accuracy scores returned by classification\_report() function for contributing risk factors

- According to the figure 16, decision tree, random forest both can predict the CVD risk with 100% accuracy and gradient boost classifier can predict the risk with 99% accuracy when use contributing risk factors only.
- The results obtained leads to a decision Random Forest and decision tree classifiers can be used to predict CVD risk with high accuracy for contributing CVD risk factors.
- But logistic regression and Naïve bayes classifiers are not suitable in this condition as the accuracy of prediction is less than 80%

As it decided to prepare a dataset by combining major and modifiable risk factors by considering figure 14, the below figure 17 illustrates the accuracy score when major and modifiable risk factors are well-known.

LogisticRegression: 0.668327 (0.049176)					KNeighborsClassifier: 0.772257 (0.037418)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.78	0.65	0.71	66	0	0.93	0.83	0.88	66
5	0.56	0.67	0.61	98	5	0.71	0.84	0.77	98
10	0.58	0.53	0.55	87	10	0.84	0.74	0.79	87
accuracy			0.62	251	accuracy			0.80	251
macro avg	0.64	0.62	0.63	251	macro avg	0.83	0.80	0.81	251
weighted avg	0.63	0.62	0.62	251	weighted avg	0.81	0.80	0.80	251
DecisionTreeClassifier: 0.990010 (0.014832)					RandomForestClassifier: 0.992010 (0.011660)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	1.00	1.00	66	0	1.00	1.00	1.00	66
5	1.00	1.00	1.00	98	5	1.00	1.00	1.00	98
10	1.00	1.00	1.00	87	10	1.00	1.00	1.00	87
accuracy			1.00	251	accuracy			1.00	251
macro avg	1.00	1.00	1.00	251	macro avg	1.00	1.00	1.00	251
weighted avg	1.00	1.00	1.00	251	weighted avg	1.00	1.00	1.00	251
GradientBoostingClassifier: 0.979030 (0.019196)					Naive Bayes: 0.671347 (0.046643)				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	0.89	0.94	66	0	0.70	0.64	0.67	66
5	0.93	1.00	0.97	98	5	0.54	0.68	0.61	98
10	1.00	1.00	1.00	87	10	0.62	0.48	0.54	87
accuracy			0.97	251	accuracy			0.60	251
macro avg	0.98	0.96	0.97	251	macro avg	0.62	0.60	0.60	251
weighted avg	0.97	0.97	0.97	251	weighted avg	0.61	0.60	0.60	251

Figure 17: Accuracy scores returned by classification\_report() function for major and modifiable risk factors

- According to the figure 17, the combination of CVD major and modifiable risk factors enhance the accuracy of CVD prediction significantly.
- Decision tree, random forest both can predict the CVD risk with 100% accuracy and gradient boost classifier can predict the risk with 97% accuracy when use major and modifiable risk factors for CVD risk prediction.
- The results obtained leads to a decision Random Forest and decision tree classifiers can be used to predict CVD risk with high accuracy for combination of major and modifiable risk factors.
- But logistic regression and Naïve bayes classifiers are not suitable in this condition as the accuracy of prediction is less than 80%

By considering the above results and discussions it is decided to use Random Forest algorithm in CVD risk prediction for Sri Lankan context as it satisfies all the above scenarios.

## 4.2.2 Results obtained in phase 2 to achieve objective 3

### Cluster analysis

Identification of CVD risk user groups and the CVD non risk user group might support the medical sector to reach and treat the correct individuals. The simple Kmeans algorithm used in WEKA clustering resulted 10 clusters with and without CVD risk for each scenario described above. The results and the decisions made give below.

#### 1. Scenario 1 : Cluster analysis for major CVD factors.

The simple Kmeans algorithm used in WEKA clustering resulted 5 clusters with and without CVD risk. The resulted output given in figure 18.

```
Final cluster centroids:
Attribute      Full Data      Cluster#
                (1252.0)      0      1      2      3      4
                (336.0)      (432.0)  (180.0)  (216.0)  (88.0)
=====
Age             10             9       10       8        0        10
Sex             2              2       1        2        1        2
Family         5              5       5        5        2.5      7.5
Cardio         5              0       5        10       0        10

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances
0      336 ( 27%)
1      432 ( 35%)
2      180 ( 14%)
3      216 ( 17%)
4       88 (  7%)
```

Figure 18: Clusters related to Sri Lankan major CVD risk factors

According to the percentage of instances in each cluster majority of the people belongs to cluster 1 to cluster 4. The definitions for those clusters are given below

Cluster 1: For a female in age group 10 (above 75 years old) if the family history of CVD is moderate, the CVD occurrence risk is moderate

Cluster 2: For a male between age group 65-69 with moderate family history of CVD have high risk of getting into CVD

Cluster 3: A female under 30 with who have low family history with CVD occurrence there is no CVD risk at the moment

Cluster 4: For a male in age group 10 (above 75 years old) with moderately high CVD family history there is a high risk of getting into CVD

As gender, CVD family history and age are major CVD risk factors, each individual should be careful about their health condition when getting old

## 2. Scenario 2: Cluster analysis for modifiable CVD factors.

The simple Kmeans algorithm used in WEKA clustering resulted 10 clusters with and without CVD risk. The resulted output given in figure 19.

```

Final cluster centroids:
Attribute          Cluster#
                   Full Data  0      1      2      3      4      5      6      7      8      9
                   (1252.0) (204.0) (128.0) (80.0) (220.0) (108.0) (296.0) (52.0) (64.0) (36.0) (64.0)
-----
BMI                23.9    21    24.1    28.5    21    22.7    23.9    25.6    20.7    27.7    28.4
DiagnosedDibetics  2        2        2        2        1        2        1        2        2        2        2
DibeticLevel       0       190     0       108     0       270     0       151     154     150     140
HbA1c              0        8        0       7.3     0       7.8     0       7.2     7.3     9        7.4
DiagnosePeriodDB  0        3        2        2        0        5        0        1        1        4        2
MedicationDB       1        1        1        1        0        1        0        1        1        1        1
DiegestIssues      0        0        0        0        1        0        0        0        0        0        0
DiagnoseBP         2        2        1        2        2        2        1        2        2        2        2
SystolicBP         0       159     0       160     160     146     0       145     159     160     149
DiastolicBP        0       90      0       110     105     96     0       91     95     95     97
DiagnoseDuration  0        3        0        2        4        3        0        3        4        5        5
MedicationBP       1        1        0        1        1        1        0        1        1        1        1
DiagnoseCholesterol 2        2        2        2        2        2        1        2        2        2        2
ChoLevel           0       260     210     206     244     222     0       209     200     202     0
DiagnoseChol       3        3        4        2        3        3        0        1        5        2        3
MedicatioCholesterol 1        1        1        1        1        1        0        1        1        1        1
Exercise           5        0       10      5        5        10     5        5        5        0        5
Cardio             5        5        5       10      5        10     0       10     10     10     10
  
```

```

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      204 ( 16%)
1      128 ( 10%)
2       80 (  6%)
3      220 ( 18%)
4      108 (  9%)
5      296 ( 24%)
6       52 (  4%)
7       64 (  5%)
8       36 (  3%)
9       64 (  5%)
  
```

Figure 19: Clusters related to Sri Lankan modifiable CVD risk factors

According to the percentage of instances in each cluster majority of the people belongs to cluster 5, 3 and 0. The definitions for those clusters are given below

Cluster 5: For a person with normal BMI (BMI =23.9) not diagnosed for diabetic, high blood pressure or cholesterol and involve in moderate exercise few times a week there is no risk to getting into CVD

Cluster 3: For a person with normal BMI diagnosed for high blood pressure and cholesterol even though they are under medication and engaged in moderate exercises there is a moderate CVD risk

Cluster 0: For a person with normal BMI diagnosed and under medication for cholesterol, blood pressure and sugar and engaged in exercises, there is a moderate CVD risk.

- When observing clusters, it was identified that the not considering gender with when generating the cluster might decrease the quality of the cluster.
- By observing all the clusters, it is possible to decide that, even the BMI is within the normal range, if a person diagnosed with NCDs like high blood pressure, diabetics or cholesterol, it is important to engage with frequency exercises. It is not good to be physical inactive. When a person having NCDs even though they are under medication for longer period of time still there is a risk of getting into CVD risk

### **3. Scenario 3: Cluster analysis for contributing CVD factors.**

The simple Kmeans algorithm used in WEKA clustering resulted 5 clusters with and without CVD risk. The resulted output given in figure 20.

According to the percentage of instances in each cluster majority of the people belongs to cluster 4. The definition for the cluster is given below

Cluster 4 : A person who is not smoking or not expose to passive smoking, if he consume about 50ml of alcohol only during a special occasion, lived in a busy city near to a main road, not consuming betel, having enough sleep and having moderate sleeping issues to different types of life events with in a short period of time and having enough mind relaxation activities and have high stress level, anxiety and the frequency of occurring anxiety there is a moderate risk for getting in to CVD risk. According to the values in 4<sup>th</sup> cluster it can further explain having a balanced diet might protecting the person from getting into high CVD risk.

By observing all the clusters generated it is possible to summaries that for a person who maintains balanced social life, having healthy diet, not consuming alcohol or nonsmoker having less risk for getting into CVD

```

Missing values globally replaced with mean/mode

Final cluster centroids:
Attribute          Full Data          Cluster#
                   (1252.0)          0          1          2          3          4
                   (1252.0)          (272.0)          (384.0)          (64.0)          (124.0)          (408.0)
=====
Smoking            0          0          6          1          0          0
PassiveSmoking    2          2          2          2          2          1
Alcohol            1          1          2          2          1          1
AlcoholConsumption 0          0          2          8          0          0
Environment        10         10         10         10         1          10
BetaConsumption   0          0          0          2.5        0          0
SleepHours         0          0          0          2.5        2.5        0
SleepIssues        5          0          5          5          5          5
Events             20         20         2          10         33         54
ExtraActivities    0          10         0          0          2.5        0
Stress             3          3          3          5          0          4
BowelToxic         2          1          2          2          1          2
ConcetraptivePills 1          1          1          1          1          1
AntibioticDuration 0          0          0          6          0          0
FriedFood          1          5          1          1          0          1
Corbohydrate       2          0          2          0          2          2
SweetFood          2          2          2          2          2          2
Sugar              0          1          0          0          0          0
Salt               2          0          2          0          0          2
Fish               5          5          5          5          10         5
Fruit              2          2          2          2          2          2
Vegetable          3          5          3          5          3          3
Coffee             0          0          0          2          0          0
Soft-Drink         2          0          2          0          0          2
Water              0          5          0          10         5          0
CoconutMilk        2          2          2          2          2          2
Anxity             5          0          5          5          5          5
AnxityFrequency    6          0          6          6          6          6
Cardio             5          0          5          10         0          5

Time taken to build model (full tr
=== Model and evaluation on traini
Clustered Instances
0          272 ( 22%)
1          384 ( 31%)
2          64 ( 5%)
3          124 ( 10%)
4          408 ( 33%)

```

Figure 20 : Clusters related to Sri Lankan modifiable CVD risk factors

#### 4. Scenario 4 : Cluster analysis for major and modifiable CVD factors.

The simple Kmeans algorithm used in WEKA clustering resulted 5 clusters with and without CVD risk. The resulted output given in figure 21. When combine major and modifiable CVD risk factors for cluster analysis cluster 0 shows the cluster instances percentage. According to the cluster 0 a female between 70-74 years old having moderate CVD family history, who diagnosed for diabetics, blood pressure and cholesterol for long period of time, but involve in exercises, there is a moderate risk in getting in to CVD. If this cluster compare with a cluster 4 it is easily can identify the for a male above 75 years old having high CVD risk if they are diagnosed under medication for blood pressure, diabetics and cholesterol for a longer period of time, if he not involved in any exercise the CVD risk is high.

Final cluster centroids:

Attribute	Cluster#					
	Full Data (1252.0)	0 (440.0)	1 (132.0)	2 (248.0)	3 (292.0)	4 (140.0)
Age	10	9	10	10	0	10
Sex	2	1	2	2	1	2
Family	5	5	5	5	2.5	5
BMI	23.9	21	24.1	28.5	23.9	25
DiagnosedDibetics	2	2	2	2	1	2
DibeticLevel	0	0	0	0	0	184
HbA1c	0	0	0	0	0	7.2
DiagnosePeriodDB	0	0	2	0	0	5
MedicationDB	1	1	1	1	0	1
DiegestIssues	0	0	0	0	0	0
DiagnoseBP	2	2	1	2	1	2
SystolicBP	0	159	0	160	0	146
DiastolicBP	0	95	0	110	0	97
DiagnoseDuration	0	4	0	4	0	3
MedicationBP	1	1	0	1	0	1
DiagnoseCholestorol	2	2	2	2	1	2
ChoLevel	0	244	210	0	0	209
DiagnoseChol	3	3	4	2	0	3
MedicatioCholestorol	1	1	1	1	0	1
Exercise	5	0	5	5	5	10
Cardio	5	5	5	10	0	10

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

```

0      440 ( 35%)
1      132 ( 11%)
2      248 ( 20%)
3      292 ( 23%)
4      140 ( 11%)

```

Figure 21 : Clusters related to Sri Lankan modifiable CVD risk factors

## Association rule mining

Association rule mining allows to identify the relationships between CVD risk factors. The Association rules were generated for below 4 scenarios to identify the association of risk factors.

### 1. Scenario 1: Association rule mining for major CVD factors.

Figure 22 depicts the association rule generated when there is a less number of feature as the experiment used only the major CVD factors.

```
Apriori
=====

Minimum support: 0.1 (125 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 20
Size of set of large itemsets L(3): 6

Best rules found:

1. Age=0 180 ==> Cardio=0 180    <conf:(1)> lift:(3.6) lev:(0.1) [129] conv:(129.97)
```

Figure 22 : Best rules generated by Apriori association rule – Major CVD risk factors

According to the rule generated using only the major CVD risk factors it is possible to say that if the age of an individual is less than 30 there is a no CVD risk associated with the person



## Scenario 2: Association rule mining for modifiable CVD factors.

Figure 23 contains the best rules generated for modifiable risk factors.

```
Apriori
=====

Minimum support: 0.6 (751 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 8

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6
Size of set of large itemsets L(2): 7
Size of set of large itemsets L(3): 4
Size of set of large itemsets L(4): 1

Best rules found:

1. MedicatioCholesterol=1 920 ==> DiagnoseCholesterol=2 920 <conf:(1)> lift:(1.34) lev:(0.19) [232] conv:(232.2)
2. DiagnoseBP=2 MedicatioCholesterol=1 784 ==> DiagnoseCholesterol=2 784 <conf:(1)> lift:(1.34) lev:(0.16) [197] conv:(197.88)
3. MedicationBP=1 MedicatioCholesterol=1 768 ==> DiagnoseCholesterol=2 768 <conf:(1)> lift:(1.34) lev:(0.15) [193] conv:(193.84)
4. DiagnoseBP=2 MedicationBP=1 MedicatioCholesterol=1 764 ==> DiagnoseCholesterol=2 764 <conf:(1)> lift:(1.34) lev:(0.15) [192] conv:(192.83)
5. DiagnosedDibetics=2 752 ==> DiegestIssues=0 752 <conf:(1)> lift:(1.4) lev:(0.17) [216] conv:(216.23)
6. MedicationBP=1 828 ==> DiagnoseBP=2 824 <conf:(1)> lift:(1.47) lev:(0.21) [263] conv:(53.44)
7. MedicationBP=1 DiagnoseCholesterol=2 780 ==> DiagnoseBP=2 776 <conf:(0.99)> lift:(1.47) lev:(0.2) [247] conv:(50.34)
8. MedicationBP=1 MedicatioCholesterol=1 768 ==> DiagnoseBP=2 764 <conf:(0.99)> lift:(1.47) lev:(0.19) [243] conv:(49.56)
9. MedicationBP=1 DiagnoseCholesterol=2 MedicatioCholesterol=1 768 ==> DiagnoseBP=2 764 <conf:(0.99)> lift:(1.47) lev:(0.19) [243] conv:(49.56)
10. MedicationBP=1 MedicatioCholesterol=1 768 ==> DiagnoseBP=2 DiagnoseCholesterol=2 764 <conf:(0.99)> lift:(1.56) lev:(0.22) [275] conv:(55.94)
```

Figure 23 : Best rules generated by Apriori association rule – Modifiable CVD risk factors

The output provides few straight forward rules with 100% confidence such as,

- If a person on medication for cholesterol that person already diagnosed for cholesterol
- If a person diagnosed for blood pressure and medication for cholesterol, that person already diagnosed for cholesterol
- If a person medication for blood pressure and medication for cholesterol, that person already diagnosed for cholesterol
- If a person on medication for blood pressure that person already diagnosed for blood pressure
- If a person medication for blood pressure and medication for cholesterol, that person already diagnosed for blood pressure

Above set of rules obtained from WEKA associator implied that high blood pressure, diabetics and cholesterol are non – communicable diseases which associated with each other. Therefore, it can recommend that if somebody diagnosed with any of above diseases, then extra care should taken place to by having health diet, exercising and expose to less stress. Otherwise, there is a high possibility for getting into other diseases which will lead to CVD risk

### Scenario 3: Association rule mining for contributing CVD factors.

Figure 24 contains the best rules generated for contributing risk factors.

```
Apriori
=====

Minimum support: 0.45 (563 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 24

Size of set of large itemsets L(2): 75

Size of set of large itemsets L(3): 21

Best rules found:

1. SleepHours=0 CoconutMilk=2 660 ==> BetalConsumption=0 620 <conf:(0.94)> lift:(1.09) lev:(0.04) [52] conv:(2.26)
2. Salt=2 CoconutMilk=2 640 ==> BetalConsumption=0 588 <conf:(0.92)> lift:(1.07) lev:(0.03) [37] conv:(1.7)
3. Vegetable=3 CoconutMilk=2 640 ==> BetalConsumption=0 588 <conf:(0.92)> lift:(1.07) lev:(0.03) [37] conv:(1.7)
4. Environment=10 672 ==> BetalConsumption=0 616 <conf:(0.92)> lift:(1.07) lev:(0.03) [38] conv:(1.66)
5. SleepHours=0 Fruit=2 668 ==> BetalConsumption=0 612 <conf:(0.92)> lift:(1.07) lev:(0.03) [37] conv:(1.65)
6. SleepHours=0 Vegetable=3 616 ==> BetalConsumption=0 564 <conf:(0.92)> lift:(1.07) lev:(0.03) [34] conv:(1.63)
7. Corbohydrate=2 Fruit=2 644 ==> BetalConsumption=0 588 <conf:(0.91)> lift:(1.06) lev:(0.03) [34] conv:(1.59)
8. Smoking=0 716 ==> BetalConsumption=0 652 <conf:(0.91)> lift:(1.06) lev:(0.03) [36] conv:(1.55)
9. Corbohydrate=2 760 ==> BetalConsumption=0 692 <conf:(0.91)> lift:(1.06) lev:(0.03) [38] conv:(1.55)
10. Fruit=2 Vegetable=3 664 ==> BetalConsumption=0 604 <conf:(0.91)> lift:(1.06) lev:(0.03) [33] conv:(1.53)
```

Figure 24 : Best rules generated by Apriori association rule – Contributing CVD risk factors

The output provide arbitrary set of associations as the dataset contains 29 feature such as,

- If a person takes a proper sleep, and consume 2-3 cups of coconut milk per day that person is not consuming betel
- If a person takes 3 spoons of salt per day, and consume 2-3 cups of coconut milk per day that person is not consuming betel

Therefore, it can conclude that the variability of the features in the modifiable risk factors unable to produce proper meaningful association rules related to CVD risk predictions.

#### Scenario 4: Association rule mining for major and modifiable CVD factors.

With the aim of identification associations between major and modifiable CVD risk factors the dataset was pruned by removing the diagnosed period for NCDs and whether the person is under medication or not as the association between NCDs and medication already identified in the scenario 2. The best rule found when used the pruned dataset is given below using figure 25.

Best rules found:

```
1. DiagnoseBP=2 Cardio=5 420 ==> DiagnoseCholestoral=2 404 <conf:(0.96)> lift:(1.29) lev:(0.07) [90] conv:(6.24)
2. Family=5 DiagnoseBP=2 564 ==> DiagnoseCholestoral=2 540 <conf:(0.96)> lift:(1.28) lev:(0.09) [118] conv:(5.69)
3. Sex=2 DiagnoseBP=2 444 ==> DiagnoseCholestoral=2 424 <conf:(0.95)> lift:(1.28) lev:(0.07) [92] conv:(5.34)
4. Family=5 DiagnosedDibetics=2 DiagnoseBP=2 416 ==> DiagnoseCholestoral=2 396 <conf:(0.95)> lift:(1.27) lev:(0.07) [84] conv:(5)
5. Family=5 DiagnosedDibetics=2 464 ==> DiagnoseCholestoral=2 440 <conf:(0.95)> lift:(1.27) lev:(0.07) [93] conv:(4.68)
6. Cardio=5 492 ==> DiagnoseCholestoral=2 464 <conf:(0.94)> lift:(1.26) lev:(0.08) [96] conv:(4.28)
7. Age=10 412 ==> DiagnoseCholestoral=2 388 <conf:(0.94)> lift:(1.26) lev:(0.06) [79] conv:(4.16)
8. DiagnosedDibetics=2 DiagnoseBP=2 612 ==> DiagnoseCholestoral=2 576 <conf:(0.94)> lift:(1.26) lev:(0.09) [118] conv:(4.17)
9. DiagnoseBP=2 848 ==> DiagnoseCholestoral=2 796 <conf:(0.94)> lift:(1.26) lev:(0.13) [162] conv:(4.04)
10. Family=5 DiagnoseCholestoral=2 600 ==> DiagnoseBP=2 540 <conf:(0.9)> lift:(1.33) lev:(0.11) [133] conv:(3.17)
```

Figure 25 : Best rules generated by Apriori association rule – Major and modifiable CVD risk factors

According to the figure 25 following associations were observed with 95% confidence.

- If a person diagnoses for blood pressure and had moderate CVD risk, then the person have to diagnose for cholesterol as well (95% confidence)
- If there is a moderated CVD family history, and diagnose for blood pressure, then person have to diagnose for cholesterol as well (95% confidence)
- If a gender of a person is male and diagnose for blood pressure with 95% confidence, stated that the person diagnoses for cholesterol as well
- If there is a moderated CVD family history, diagnose for diabetic and blood pressure, then person have to diagnose for cholesterol as well (95% confidence)
- If there is a moderated CVD family history, and diagnose for diabetics then person have to diagnose for cholesterol as well (95% confidence)

This experiment also proves the relationship between non communicable diseases. Therefore, as a summary it can conclude that male patients, who have moderate CVD family history need to be careful if they are suffering from any NCD. If a person has at least one NCD, person might be affected with other NCDs which leads to CVD in future.

### 4.3 Experimental Results Evaluation

#### CVD risk prediction

This study used Logistic regression, Decision Tree Classifier, Gradient Boosting Classifier, Kneighbors Classifier, Random Forest Classifier and Gaussian Naïve Bayes classifier for predicting CVD risk. The accuracy values obtained in each experiment proved that the Random Forest algorithm among others shows better performance in predicting CVD risk. Table 7 given below compares the accuracy of classifiers before and after applying PCA for different datasets described in section 3.2.2

Table 7: Comparison of classification accuracy before and after applying PCA for dataset used in CVD prediction in Sri Lankan

		Accuracy For major risk factors	Accuracy For modifiable risk factors	Accuracy For contributing risk factors	Accuracy For major and modifiable risk factors
Logistic regression	Before PCA	56%	64%	63%	63%
	After PCA	59%	62%	68%	62%
Decision tree Classifier	Before PCA	69%	100%	100%	100%
	After PCA	69%	100%	100%	100
Gradient Boosting Classifier	Before PCA	66%	95%	95%	96%
	After PCA	69%	98%	99%	97%
Kneighbors Classifier	Before PCA	64%	77%	73%	79%
	After PCA	67%	77%	87%	80%
Random Forest Classifier	Before PCA	69%	100%	100%	100%
	After PCA	69%	100%	100%	100%
GaussianNB classifier	Before PCA	56%	55%	58%	56%
	After PCA	60%	60%	57%	60%

Below figures 26 to 29 clearly illustrates that there is a slight difference between the prediction accuracy before and after applying PCA.

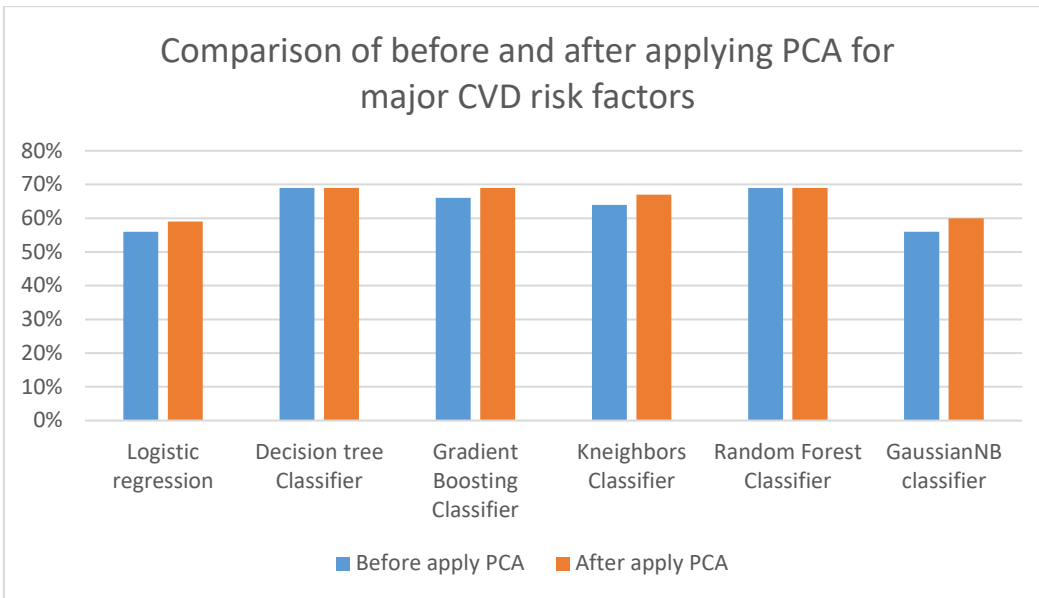


Figure 26 : Comparison between before and after applying PCA for major CVD factors in Sri Lankan context

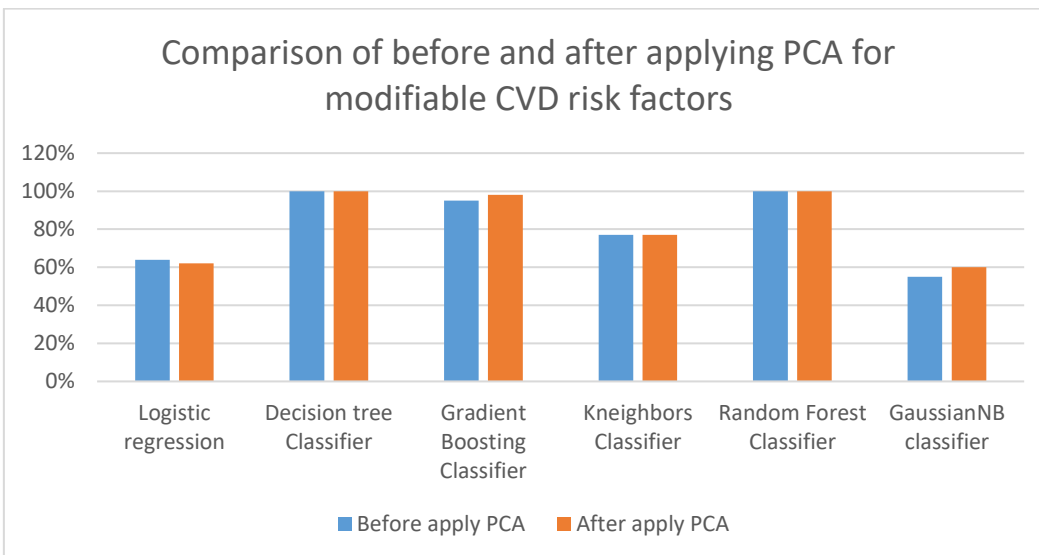


Figure 27 : Comparison between before and after applying PCA for modifiable CVD factors in Sri Lankan context

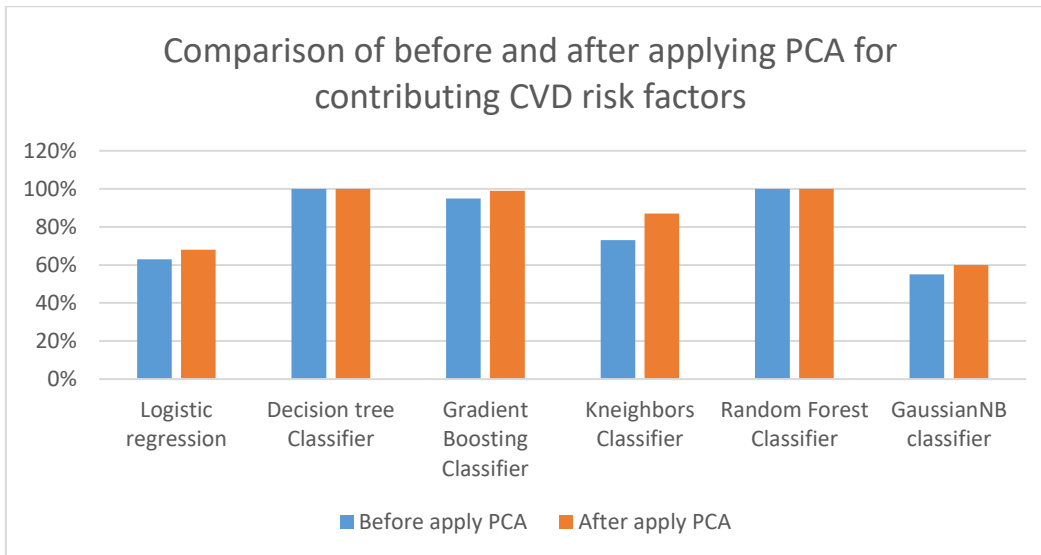


Figure 28: Comparison between before and after applying PCA for contributing CVD factors in Sri Lankan context

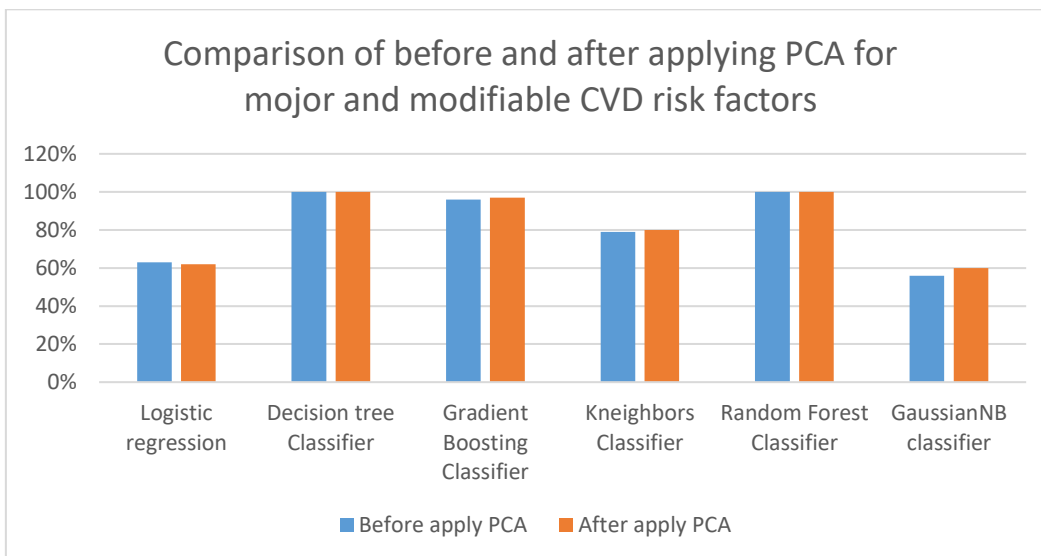


Figure 29: Comparison between before and after applying PCA for major and modifiable CVD factors in Sri Lankan context

Above figures clearly illustrates that in every condition Random Forest and decision tree classifiers predicts the CVD with highest accuracy in Sri Lankan context. Therefore, it is possible to use decision tree or random forest classifier when predicting the CVD risk for new individual. Also, above figure indicates that when predicting CVD risk there is no significant improvement in using PCA. As data mining algorithms used in this research work are correlation robust algorithms it is possible to use random forest or decision tree classifiers in predicting CVD risk without dimensionality reduction function like PCA. But the experiment proves that when using PCA, it is possible to obtain the classification results with in a short

period of time as PCA reducing the classification complexity. Therefore, in terms of enhance the efficiency of the classifiers it is good to use PCA

### Cluster analysis

According to the cluster analysis conducted following clusters identified as high-risk CVD clusters in Sri Lanka. The summarization of clusters generated in above experiments is given below table 8. According to the table, health officers in Sri Lanka can have their focus on high CVD risk groups when providing CVD awareness sessions or in the treatments.

Table 8: CVD risk clusters

High CVD risk	<ul style="list-style-type: none"> <li>• Male individuals in between age 65-69 with moderate CVD family history</li> <li>• Male individuals above 75 years old with moderate CVD family history</li> <li>• Male above 75 years old with moderate CVD family history with marginally high or high BMI diagnosed for diabetics, blood pressure and cholesterol, and under medication for all the NCDs if not following regular exercise the CVD risk is high</li> <li>• Individuals involve in passive smoking, consume alcohol, lived in toxic environment near to a main road, having 5-6 hours sleeping, having sleeping issues and not having enough water while having a balanced diet and facing to anxiety in high frequency having high CVD risk</li> </ul>
Moderate CVD risk	<ul style="list-style-type: none"> <li>• Female above 75 years old with moderate CVD family history</li> <li>• Female between 70-74 years old with moderate CVD family history with normal BMI diagnosed for diabetics, blood pressure and cholesterol, and under medication for all the NCDs if following regular exercise, the CVD risk is moderate</li> </ul>
Low CVD risk	<ul style="list-style-type: none"> <li>• Female under 30 with low CVD family history</li> <li>• Female under 30 maintaining normal BMI having low CVD family history not diagnosed for diabetics, cholesterol, blood</li> </ul>

	<p>pressure and involve in regular physical exercises the CVD risk is low.</p> <ul style="list-style-type: none"><li>• Not smoking, not consume alcohol, lives near to a main road, but having proper sleep without any issues, having low stress and no anxiety levels when consume a balanced diet with enough water no risk of getting into CVD</li></ul>
--	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

As there is no research conducted in Sri Lanka to detect CVD risk groups and associations between CVD risk factors the output of this research can be used to identify the needy user groups in Sri Lankan domain.



## **CHAPTER 5**

### **CONCLUSION AND FUTURE WORKS**

The research is focusing on develop an accurate CVD prediction by evaluating the literature-based classifiers and critical risk factors. After identification of accurate CVD prediction classifier, CVD risk factors were collected from Sri Lankan context with the aim of developing customized CVD prediction for Sri Lanka. Globally validated CVD risk factors and CVD risk factors identified in Sri Lankan context further analyzed to identify high risk and low risk CVD groups and associations among risk factors. This chapter summarizes the results obtained and the contributions to the domain

#### **5.1 Conclusions about Experiments**

##### **5.1.1 Detecting Critical Global CVD Risk Factors**

Literature survey highlighted PCA can be used to reduce the features of a given dataset. Therefore, the research was planned to have PCA on top of the CVD prediction classifiers. But the obtained results after applying PCA indicated that there is no significant difference between before and after applying PCA. Also, it indicated that all the 11 principal components are needed to get the highest accuracy of majority of existing classifiers. Literature also highlighted that PCA enhance the performance of data mining algorithms by reducing the computational complexity via reducing the dimensionality of the dataset. Therefore, with the aim of reducing the classification complexity and enhance accuracy in small proportion (according to the table 4, chapter 3)the research adopted PCA on top of CVD classifiers.

According to the Explained variance vector return after applying PCA it was decided that all the 11 features of the dataset contribute 94% of total variance. Therefore, if any principal component dropped to reduce dimensionality it the contribution for the feature to cumulative variance getting reduce. Remaining components insufficient to contribute to the total variance. Therefore, even though the PCA applied on top of the classifiers all the 11 principal components used in the research. This mechanism was unable to use for identifying most critical risk factors in the given feature set, but it enhanced the efficiency of the classifier.

### **5.1.2 Detecting Accurate CVD Prediction Classifier using Global CVD Risk Factors**

Majority of CVD risk prediction applications available online use Framingham algorithm which have multiple limitations in predicting CVD risk. Therefore, identification of CVD risk prediction classifier using data mining will enhance capabilities in predicting CVD risk. This research used 6 classifiers and compare the accuracy scores obtained from each classifier. As the classifier incorporate with PCA classifiers provided accuracy scores within shorter period of time and with better results than classifier without PCA. By considering the accuracy values obtained it was identified that Random Forest classifier with PCA predicts the CVD risk with 85% accuracy. Therefore, it was decided to use Random Forest for CVD risk prediction applications. As an output of the research a basic UI created to input CVD risk factors of any user to check the CVD risk. For the application Random Forest classifier was used as it produced the highest accuracy. Majority of online available applications and CVD risk prediction applications neglect age factor, or some age groups or habits like smoking and alcohol consumption when predicting the CVD risk. As the datamining-based approach allowed to input all the literature based well-known CVD risk factors to predict the CVD risk, it could provide reliable CVD prediction.

### **5.1.3 Detecting Accurate CVD Prediction Classifier using CVD Risk Factors in Sri Lanka**

When the Sri Lankan CVD risk factors dataset with 51 features loaded to the previously build classifier (with PCA) returned lower accuracy values. Therefore, the dataset split 3 groups considering major, modifiable contributing risk factors. Application of PCA along with classifiers returned the accuracy scores within less time and higher accuracy. To predict the CVD risk factors for all the 3 scenarios the best accuracy results showed by random forest classifier and the decision tree classifier.

As major risk factors dataset consists only with 3 features, the accuracy was 69% with Random Forest classifier. Therefore, the major risk factors combined with modifiable risk factors dataset to enhance the accuracy of classifier. This decision leads to generate meaningful association and clustering results as well. The advantage of this split might be is any individual unaware

about his blood pressure value, diabetic level or cholesterol level, still he will be able to now the future CVD risk considering current known values.

#### **5.1.4 Identification of CVD Risk Groups and Associations Between CVD Risk Factors**

Cluster analysis and association rule mining available in WEKA used to achieve this objective. Based on the cluster analysis it was identified that user groups who suffering from blood pressure, diabetics and cholesterol but if the cluster is physically active, the risk of getting in CVD is low. But if a cluster of females aged 60 years having normal BMI, not diagnosed with any NCD and not consuming smoke or alcohol, if there are physically inactive there the cluster have the risk of getting into CVD risk.

All the clusters prove that if any individual does not smoke or not consume alcohol and physically active it will reduce the CVD risk

When it comes to the Sri Lankan context clustering proves that the age group greater than 60 years having high risk in getting into CVD.

Also, a cluster with normal BMI, if a person diagnosed with NCDs like high blood pressure, diabetics or cholesterol, it is important to engage with frequency exercises. When a person having NCDs even though they are under medication for longer period of time still there is a risk of getting into CVD. But if the people in this cluster can be physically active, then they will be able to move from the cluster to a CVD risk less cluster

When considering contributing CVD risk factors in Sri Lankan context, by observing all the clusters generated it is possible to summarize that for a person who maintains balanced social life, with mind relaxing activities, not involve in life changing events frequently, having healthy diet, not consuming alcohol or nonsmoker having less risk for getting into CVD

As the contributing risk factors contributed to multiple clusters can assume that all the collected possible factors affect the CVD in different manners.

The association rules generated proves that there is a high association between high blood pressure, cholesterol and diabetics. If anyone having one of these there is a high possibility of diagnosis for other NCDs as well.

It also highlighted that there is a high association between NCDs, alcohol consumption and smoking habit. It leads to the assumption, if it is possible to avoid alcohol consumption, smoking habits the cholesterol and glucose levels will be normal. This will eventually reduce the CVD risk.

When it comes to the Sri Lankan context it highlighted that there is a 100% confidence in not having CVD risk for the age group below 30. Therefore, neglecting the age group below 30 in CVD risk prediction will be ok. But as the dataset does not contain any data record it is not recommended to remove the age group entirely.

When there are multiple features in the contributing risk factors data set the variability between the factors unable to produce proper associations rule. But with the support of a medical research if the dataset is pruned it will show the associations with different food habits, lifestyles and mental health conditions. As this is the first approach of identification CVD risk factors in Sri Lankan context this research was unable to track the exact features to be pruned to get better associations.

## **5.2 Conclusions about the Research Study**

This research work introduced CVD classifier combining PCA to determine the CVD risk using literature-based classifiers and CVD risk factors. Results of the research supported to identify that the Random Forest classifier is the accurate classifier in CVD risk prediction. The section 4.3 proves that application PCA on top of the classifiers do not improve the performance of the classifiers in significant amount as well. Therefore, to predict the CVD risk in both global and Sri Lankan context use the Random Forest classifier without PCA is acceptable. But still as it reduce the classification complexity and provide the results within a short period of time the PCA can be adopted in CVD prediction classifiers.

The data used for CVD risk prediction not focused Sri Lankan lifestyle, habits and environment. Therefore, CVD possible risk factors identified to prepare the questionnaire with the support of medical students and feedbacks from domain experts. The collected data feed to multiple classifiers to detect the best classifier and it was identified that Random Forest classifier is suitable for CVD risk prediction in Sri Lankan oriented data as well. As this is the first attempt to identify the CVD risk factors unique to Sri Lanka exact set of clustering and classifications conducted to detect CVD risk groups and associations between factors.

### **5.3 Future Work**

The dataset used for CVD prediction in Sri Lankan domain contains 51 features. Due to the high variance in the data collected it was unable to identify most critical risk factors in Sri Lankan risk prediction. The research can be further improved by expanding the size of the dataset. As the current experiment highlighted there is an association with the age group below 30 and not getting into CVD, it might lead to decide it is possible to remove the age group below 30 when predicting the CVD. But the Sri Lankan dataset used for the experiment doesn't contain any user data from CVD patients below 30 years old. In future with the support of a cardiologist it is important to identify the CVD patients below 30 years old and collect the data from them to enhance the results of the research.

The clusters generated using dataset 3 highlighted that use of contraceptive pills or use of coconut milk does not directly affect the CVD. The reason behind the result is more than 80% of instances in the dataset use 2-3 cups of coconut milk daily and not consume contraceptive pills. During the methodology, experts of the domain suggested that there might be a relationship between CVD and those factors as well. Therefore, the experiments can be further expanded by collecting data from identified specific user groups to check the influence of coconut milk consumption and use of contraceptive pills in CVD occurrence.

## REFERENCES

- American College of Cardiology, n.d. ASCVD Risk Estimator [WWW Document]. URL [https://tools.acc.org/ldl/ascvd\\_risk\\_estimator/index.html#!/calculate/estimator/](https://tools.acc.org/ldl/ascvd_risk_estimator/index.html#!/calculate/estimator/)
- American Heart Association, 2021. Heart disease #1 cause of death rank likely to be impacted by COVID-19 for years to come [WWW Document]. EurekAlert! URL <https://www.eurekalert.org/news-releases/658325> (accessed 5.10.21).
- American Heart Association, 2016. Understand Your Risks to Prevent a Heart Attack [WWW Document]. [www.heart.org](http://www.heart.org). URL <https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack> (accessed 8.19.20).
- Aparna U.R., Paul, S., 2016. Feature selection and extraction in data mining, in: 2016 Online International Conference on Green Engineering and Technologies (IC-GET). Presented at the 2016 Online International Conference on Green Engineering and Technologies (IC-GET), IEEE, Coimbatore, India, pp. 1–3. <https://doi.org/10.1109/GET.2016.7916845>
- Banu, N.K.S., Swamy, S., 2016. Prediction of heart disease at early stage using data mining and big data analytics: A survey, in: 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECOT). Presented at the 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECOT), IEEE, Mysuru, India, pp. 256–261. <https://doi.org/10.1109/ICEECOT.2016.7955226>
- Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R., Delling, F.N., Djousse, L., Elkind, M.S.V., Ferguson, J.F., Fornage, M., Jordan, L.C., Khan, S.S., Kissela, B.M., Knutson, K.L., Kwan, T.W., Lackland, D.T., Lewis, T.T., Lichtman, J.H., Longenecker, C.T., Loop, M.S., Lutsey, P.L., Martin, S.S., Matsushita, K., Moran, A.E., Mussolino, M.E., O’Flaherty, M., Pandey, A., Perak, A.M., Rosamond, W.D., Roth, G.A., Sampson, U.K.A., Satou, G.M., Schroeder, E.B., Shah, S.H., Spartano, N.L., Stokes, A., Tirschwell, D.L., Tsao, C.W., Turakhia, M.P., VanWagner, L.B., Wilkins, J.T., Wong, S.S., Virani, S.S., On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee, 2019. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. *Circulation* 139. <https://doi.org/10.1161/CIR.0000000000000659>
- Cleveland Clinic medical professional, n.d. Reynolds Risk Score [WWW Document]. URL <http://www.reynoldsriskscore.org/Default.aspx>
- CVDQuestionnaire.pdf, n.d.
- ESC CVD Risk Calculation App [WWW Document], n.d. URL <https://www.escardio.org/Education/ESC-Prevention-of-CVD-Programme/Risk-assessment/esc-cvd-risk-calculation-app>, <https://www.escardio.org/Education/ESC-Prevention-of-CVD-Programme/Risk-assessment/esc-cvd-risk-calculation-app> (accessed 5.12.21).
- heartfoundation.org.au, n.d. Cardiovascular Disease Risk Calculator | Heart Foundation [WWW Document]. URL <https://www.heartfoundation.org.au/health-professional-tools/cvd-risk-calculator>
- Jabbar, M.A., Deekshatulu, B.L., Chandra, P., 2013. Heart disease prediction using lazy associative classification, in: 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (IMac4s). Presented at the 2013 International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), IEEE, Kottayam, pp. 40–46. <https://doi.org/10.1109/iMac4s.2013.6526381>

- Kavitha, R., Kannan, E., 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining, in: 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS). Presented at the 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), IEEE, Pudukkottai, India, pp. 1–5. <https://doi.org/10.1109/ICETETS.2016.7603000>
- Kumar, P.J., Clark, M.L., Feather, A. (Eds.), 2017. Kumar and clark's clinical medicine, 9th edition. ed. Elsevier, Edinburgh London New York Oxford Philadelphia St Louis Sydney Toronto.
- Laurae, 2017. How to not be dumb at applying Principal Component Analysis (PCA)? Data Science & Design. URL <https://medium.com/data-design/how-to-not-be-dumb-at-applying-principal-component-analysis-pca-6c14de5b3c9d> (accessed 9.20.21).
- Leening, M.J.G., Cook, N.R., Ridker, P.M., 2016. Should we reconsider the role of age in treatment allocation for primary prevention of cardiovascular disease? *Eur Heart J* 1542–1547. <https://doi.org/10.1093/eurheartj/ehw287>
- Mendonca, F., Manihar, R., Pal, A., Prabhu, S.U., 2019. Intelligent Cardiovascular Disease Risk Estimation Prediction System, in: 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). Presented at the 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), IEEE, Mumbai, India, pp. 1–6. <https://doi.org/10.1109/ICAC347590.2019.9036738>
- Narain, R., Saxena, S., Goyal, A., 2016. Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network based approach. *PPA Volume 10*, 1259–1270. <https://doi.org/10.2147/PPA.S108203>
- Patra, R., Khuntia, B., 2019. Predictive Analysis of Rapid Spread of Heart Disease with Data Mining, in: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). Presented at the 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), IEEE, Coimbatore, India, pp. 1–4. <https://doi.org/10.1109/ICECCT.2019.8869194>
- Purushottam, Saxena, K., Sharma, R., 2015. Efficient heart disease prediction system using decision tree, in: International Conference on Computing, Communication & Automation. Presented at the 2015 International Conference on Computing, Communication & Automation (ICCCA), IEEE, Greater Noida, India, pp. 72–77. <https://doi.org/10.1109/CCAA.2015.7148346>
- Qrenawi, M.I., Al Sarraj, W., 2018. Identification of Cardiovascular Diseases Risk Factors among Diabetes Patients Using Ontological Data Mining Techniques, in: 2018 International Conference on Promising Electronic Technologies (ICPET). Presented at the 2018 International Conference on Promising Electronic Technologies (ICPET), IEEE, Deir El-Balah, pp. 129–134. <https://doi.org/10.1109/ICPET.2018.00030>
- Ruano, M.G., Almeida, G.P., Palma, F., Raposo, J.F., Ribeiro, R.T., 2018. Reliability of medical databases for the use of real word data and data mining techniques for cardiovascular diseases progression in diabetic patients, in: 2018 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE). Presented at the 2018 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE), IEEE, Porto, pp. 1–6. <https://doi.org/10.1109/GMEPE-PAHCE.2018.8400769>
- Srinivas, K., Rao, G.R., Govardhan, A., 2010. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques, in: 2010 5th International Conference on Computer Science & Education. Presented at the Education (ICCSE 2010), IEEE, Hefei, China, pp. 1344–1349. <https://doi.org/10.1109/ICCSE.2010.5593711>
- Tripathi, A., 2020. A Complete Guide to Principal Component Analysis — PCA in Machine Learning [WWW Document]. Medium. URL <https://towardsdatascience.com/a->

- complete-guide-to-principal-component-analysis-pca-in-machine-learning-664f34fc3e5a (accessed 9.12.21).
- Ulianova, S., 2019. Cardiovascular Disease dataset [WWW Document]. URL <https://kaggle.com/sulianova/cardiovascular-disease-dataset> (accessed 5.11.21).
- Veeraswamy, A., Babu, A.M., 2019. Classification of High Dimensional Data Using Filtration Attribute Evaluation Feature Selection Method of Data mining, in: 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT). Presented at the 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), IEEE, Mysuru, India, pp. 8–12. <https://doi.org/10.1109/ICEECCOT46775.2019.9114804>
- Visalakshi, S., Radha, V., 2014. A literature review of feature selection techniques and applications: Review of feature selection in data mining, in: 2014 IEEE International Conference on Computational Intelligence and Computing Research. Presented at the 2014 IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–6. <https://doi.org/10.1109/ICCIC.2014.7238499>
- WageIndicator Network, 2020. Medical Insurance in Sri Lanka [WWW Document]. WageIndicator subsite collection. URL <https://salary.lk/labour-law/work-and-sickness/medical-insurance> (accessed 8.19.20).
- World Health Organization, 2017. Cardiovascular diseases (CVDs) [WWW Document]. URL [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)) (accessed 8.17.20).
- World Health Organization, 2014. WHO/ISH Risk prediction charts for 14 WHO epidemiological sub-regions.
- World Heart Federation, 2017. Cardiovascular diseases - Global facts and figures. World Heart Federation. URL <https://world-heart-federation.org/resource/cardiovascular-diseases-cvds-global-facts-figures/> (accessed 8.19.20).
- [www.hopkinsmedicine.org](http://www.hopkinsmedicine.org), n.d. Heart and Vascular [WWW Document]. URL <https://www.hopkinsmedicine.org/health/heart-and-vascular> (accessed 8.20.20).
- Xu, S., Shi, H., Duan, X., Zhu, T., Wu, P., Liu, D., 2016. Cardiovascular risk prediction method based on test analysis and data mining ensemble system, in: 2016 IEEE International Conference on Big Data Analysis (ICBDA). Presented at the 2016 IEEE International Conference on Big Data Analysis (ICBDA), IEEE, Hangzhou, China, pp. 1–5. <https://doi.org/10.1109/ICBDA.2016.7509809>
- Zaffar, M., Hashmani, M.A., Savita, K.S., 2017. Performance analysis of feature selection algorithm for educational data mining, in: 2017 IEEE Conference on Big Data and Analytics (ICBDA). Presented at the 2017 IEEE Conference on Big Data and Analytics (ICBDA), IEEE, Kuching, pp. 7–12. <https://doi.org/10.1109/ICBDAA.2017.8284099>



## Appendix 1: Initial Questionnaire



### Purpose of the Study

The study hopes to identify the most critical cardiovascular diseases (referred to as “CVD” hereon) risk factors to calculate the CVD risk and to provide accurate customized health tips using a data mining based approach. Further, the study expects to enable CVD risk patients to track their health condition over time by themselves without the interaction of a medical officer.

### How old are you?

Under 30	<input type="checkbox"/>
30 – 34	<input type="checkbox"/>
35 – 39	<input type="checkbox"/>
40 – 44	<input type="checkbox"/>
45 – 49	<input type="checkbox"/>
50 – 54	<input type="checkbox"/>
55 – 59	<input type="checkbox"/>
60 – 64	<input type="checkbox"/>
65 – 69	<input type="checkbox"/>
70 – 74	<input type="checkbox"/>
75 and over	<input type="checkbox"/>

### Cardiovascular history

Do you have diagnosed cardiovascular disease, atherosclerosis, previous heart attack, and/or previous stroke	<input type="checkbox"/>
Have you experienced angina (heart pain) within the last 3 months	<input type="checkbox"/>

### Family History

Mother with Cardiovascular Disease at less than 65 years (high blood pressure, heart attack, angina, stroke, hardening of the arteries)	<input type="checkbox"/>
Father with Cardiovascular Disease at less than 55 years (high blood pressure, heart attack, angina, stroke, hardening of the arteries)	<input type="checkbox"/>
Parent with Type II Diabetes (adult-onset diabetes)	<input type="checkbox"/>

## Lifestyle

### Exercise

Moderate exercise is brisk walking, jogging, cycling, swimming, playing sports or any exercise that increases breathing and heart rate continuously for at least 20 minutes

Sedentary – moderate exercise less than once a week	<input type="checkbox"/>
Moderate exercise (average once per week)	<input type="checkbox"/>
Moderate exercise (average 2 – 3 times per week)	<input type="checkbox"/>
Moderate exercise (average 4 – 5 times per week)	<input type="checkbox"/>
Moderate exercise (average 5 or more times per week)	<input type="checkbox"/>

### Smoking

Never smoked	<input type="checkbox"/>
Ex-smoker	<input type="checkbox"/>
Smoker (If “yes” select below options	<input type="checkbox"/>
Current smoker less than 20 cigarettes/day	<input type="checkbox"/>
Current smoker more than 20 cigarettes/day	<input type="checkbox"/>

### Passive smoking

(a non-smoker exposed to smoke most days at home or work)

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

### Alcohol

Do you consume: Male: 5 or more drinks (50 ml per drink) Female: 3 or more drinks (50 ml per drink) in one sitting on a fortnightly or more frequent basis? <i>(If you are not consuming alcohol keep the cage blank)</i>	<input type="checkbox"/>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------

### Environment

Do you live on a main road?	<input type="checkbox"/>
Do you live in a city?	<input type="checkbox"/>
Do you live in an industrial area with gas emissions?	<input type="checkbox"/>
Do you work with any chemicals, cleaners, pesticides, petrochemicals, paints, exhausts?	<input type="checkbox"/>

### Sleep

How many hours of sleep do you have on average per night?

0-4	<input type="checkbox"/>
5-6	<input type="checkbox"/>
7-8	<input type="checkbox"/>
More than 8 hours	<input type="checkbox"/>

**Do you experience?**

Snoring	<input type="checkbox"/>
Insomnia, difficulty falling asleep or interrupted sleep	<input type="checkbox"/>

**Stress**

**Have you experienced any of the following events in the past 6 months?**

Death of spouse	<input type="checkbox"/>
Death of family member	<input type="checkbox"/>
Divorce/separation	<input type="checkbox"/>
Marital reconciliation	<input type="checkbox"/>
Jail term	<input type="checkbox"/>
Major illness/injury/surgery	<input type="checkbox"/>
Marriage	<input type="checkbox"/>
Dismissal from work	<input type="checkbox"/>
Retirement	<input type="checkbox"/>
Death of a friend	<input type="checkbox"/>
Illness in the family	<input type="checkbox"/>
Sexual difficulties	<input type="checkbox"/>
Pregnancy	<input type="checkbox"/>
Moving to a new town/city/country	<input type="checkbox"/>
Family/relationship disputes	<input type="checkbox"/>
Change in financial status	<input type="checkbox"/>
Change in occupation	<input type="checkbox"/>
Change in the work responsibilities	<input type="checkbox"/>
Mortgage	<input type="checkbox"/>
Major family events – Wedding, births in the immediate family	<input type="checkbox"/>
Look after son or daughter	<input type="checkbox"/>
Personal difficulties at work	<input type="checkbox"/>
Outstanding personal achievement	<input type="checkbox"/>
Change in residence	<input type="checkbox"/>
Change in schools	<input type="checkbox"/>
Change in social habits	<input type="checkbox"/>
Change in routine	<input type="checkbox"/>
Holidays	<input type="checkbox"/>
Christmas	<input type="checkbox"/>
Minor violations of the law	<input type="checkbox"/>

**Do you participate in any of the following activities for more than an hour a week?**

Medication / prayer	<input type="checkbox"/>
Yoga /stretching / relaxation exercises	<input type="checkbox"/>
Community events/ social activities / sports	<input type="checkbox"/>
Play with pets	<input type="checkbox"/>

**Select if applicable**

Do you feel anxiety, worry, fear, sudden feelings of panic, inability to control breathing and accelerated heart rate when upset, or recurrent feelings of unease?	Weekly or more	<input type="checkbox"/>
	Monthly or more	<input type="checkbox"/>
Do you have feelings of sadness, depression, hopelessness, apathy, gloom, helplessness, isolation, loneliness, or lack of interest in social interaction?	Weekly or more	<input type="checkbox"/>
	Monthly or more	<input type="checkbox"/>
Are you easily angered or frustrated, feel resentment or hostility towards others or frequently irritable?	Weekly or more	<input type="checkbox"/>
	Monthly or more	<input type="checkbox"/>

**Bowel Toxicity**

**Do you regularly experience lower abdominal pain, gas, bloating, diarrhoea, constipation, straining when passing bowel motions, excessively smelly stools and/or a feeling that your bowels do not completely empty?**

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

**Have you taken the oral contraceptive pill for more than 6 months in the last year?**

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

**For what length of time have you been on antibiotics in the last year?**

Less than 2 weeks	<input type="checkbox"/>
2 weeks – 2 months	<input type="checkbox"/>
2 – months	<input type="checkbox"/>
Longer than 6 months	<input type="checkbox"/>

**Blood Sugar**

**Select if applicable**

Do you feel your energy levels drop within an hour of eating? and /or Do you experience cravings for sweets or chocolate? and /or Do you have headaches or an inability to concentrate which is relieved by eating?	<input type="checkbox"/>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------

### Are you diabetic?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

### Diet

How often do you usually eat fried foods?	Less than once a week <input type="checkbox"/>	1 – 2 times a week <input type="checkbox"/>	3 – 6 times a week <input type="checkbox"/>	Every day <input type="checkbox"/>
How many serves of bread, pasta, rice, potatoes or other starchy foods do you have a day?	0 – 1 serves daily <input type="checkbox"/>	2 serves daily <input type="checkbox"/>	3 serves daily <input type="checkbox"/>	4 or more serves daily <input type="checkbox"/>
How many servings of sweet foods like cakes, biscuits, lollies and/or chocolate do you consume a day?	Usually none <input type="checkbox"/>	1 – 2 serves daily <input type="checkbox"/>	More than 2 serves daily <input type="checkbox"/>	
How many teaspoons of sugar do you consume daily in hot drinks, added to foods, etc.?	0 – 3 <input type="checkbox"/>	4– 6 <input type="checkbox"/>	7 – 9 <input type="checkbox"/>	10 or more <input type="checkbox"/>
How often do you usually eat fish?	Rarely <input type="checkbox"/>	1 – 2 times a week <input type="checkbox"/>	3 – 6 times a week <input type="checkbox"/>	Every day <input type="checkbox"/>
How many pieces of fruit do you usually eat a day?	Usually none <input type="checkbox"/>	1 – 3 pieces daily <input type="checkbox"/>	4 or more pieces daily <input type="checkbox"/>	
How many serves of vegetables (excluding potatoes) do you usually eat a day? (1 serve = approximately 1 handful)	Usually none <input type="checkbox"/>	1 – 2 serves daily <input type="checkbox"/>	3 – 4 serves daily <input type="checkbox"/>	5 or more serves daily <input type="checkbox"/>
How many cups of coffee do you usually drink a day?	Usually none <input type="checkbox"/>	1 – 2 cups daily <input type="checkbox"/>	3 – 4 cups daily <input type="checkbox"/>	5 or more cups daily <input type="checkbox"/>
How much soft-drink do you consume on average?	Less than 500 ml per week <input type="checkbox"/>	1 – 2 litres per week <input type="checkbox"/>	3 – 4 litres per week <input type="checkbox"/>	5 or more litres per week <input type="checkbox"/>

How much water do you drink a day?	0 – 500 ml <input type="checkbox"/>	501 ml – 1.25 litres <input type="checkbox"/>	More than 1.25 litres <input type="checkbox"/>	
------------------------------------	----------------------------------------	--------------------------------------------------	---------------------------------------------------	--

### **Inflammation and Pain**

**Do you experience any of the following symptoms more than once a month?**

Wheezing, sneezing, a runny nose, sore throat, itchy or watery eyes, coughing and/or blocked nose	<input type="checkbox"/>
Heart palpitations or headaches after certain foods	<input type="checkbox"/>

**Do you experience recurrent pain?**

Daily	<input type="checkbox"/>
Weekly	<input type="checkbox"/>
Monthly or less	<input type="checkbox"/>
Never	<input type="checkbox"/>

## Appendix 2: Data label assignment for each response for the google form to collect CVD factors in Sri Lankan Context

Below assignments of label is based on Metagenics Genetic potential through nutrition project conducted in Australia and New Zealand (“CVDQuestionnaire.pdf,” n.d.)

1. How old are you?

Under 30	0
30 – 34	1
35 – 39	2
40 – 44	3
45 – 49	4
50 – 54	5
55 – 59	6
60 – 64	7
65 – 69	8
70 – 74	9
75 and over	10

2. What is your height in CM

3. What is your weight (weight in kg)?

4. What is your gender?

Male	2
Female	1

5. Are you diagnosed to have blood sugar/diabetics

Yes	2
No	1

6. Mention latest fasting blood sugar level (in mg/dL)

7. Mention latest HbA1c level in percentage (if known)

8. From how long you have been diagnose for diabetics

Less than 6 months	5
More than 6 months but less than a year	4
More than a year but less than 5 years	3
More than 5 years but less than 10 years	2
More than 10 years	1

9. Are you on medication for diabetes

Yes	2
No	1

10. Select if applicable

Do you often feel hungry	1
Do you often feel thirsty	1
Increase frequency of passing urine	1

11. Are you diagnose to have blood pressure

Yes	2
No	1

12. Mention latest systolic blood pressure (in mm Hg)

13. Mention latest diastolic blood pressure (in mm Hg)

14. From how long you have been diagnose for blood pressure

Less than 6 months	5
More than 6 months but less than a year	4
More than a year but less than 5 years	3
More than 5 years but less than 10 years	2
More than 10 years	1

15. Are you on medication for blood pressure

Yes	2
No	1

16. Are you diagnose to have cholesterol

Yes	2
No	1

17. Mention latest cholesterol level (in mg/dL)

18. From how long you have been diagnose for cholesterol

Less than 6 months	5
More than 6 months but less than a year	4
More than a year but less than 5 years	3
More than 5 years but less than 10 years	2
More than 10 years	1

19. Are you on medication for cholestorol

Yes	2
No	1

20. Exercise - Moderate exercise is brisk walking, jogging, cycling, swimming, playing sports or any exercise that increases breathing and heart rate continuously for at least 20 minutes

Sedentary – moderate exercise less than once a week	10
Moderate exercise (average 2 – 3 times per week)	5



Moderate exercise (average 4 – 5 times per week)	3
Moderate exercise (average 5 or more times per week)	0
None	10

21. Smoking

Never smoked	0
Ex-smoker	1
Current smoker 1 to 5 cigarettes per day	5
Current smoker 6 to 10 cigarettes per day	6
Current smoker 11 to 15 cigarettes per day	8
Current smoker 16 to 20 cigarettes per day	9
Current smoker more than 20 cigarettes per day	10

22. Passive smoking – and -smoker exposed to smoke most days at home or work

Yes	2
No	1

23. Alcohol consumption

Yes	2
No	1

24. Select the applicable section from below option [Consume alcohol daily ]

	1-2 drinks in one sitting(50 ml per drink)	2-5 drinks in one sitting(50 ml per drink)	More than 5 drinks in one sitting(50 ml per drink)
Consume alcohol daily	7	8	10
Consume alcohol weekly	4	5	6
Consume alcohol occasionally (in special events)	1	2	3

25. Environment

Do you live on a main road?	10
Do you live in a city?	5
Do you live in an industrial area with gas emissions?	7.5
Do you work with any chemicals, cleaners, pesticides, petrochemicals, paints, exhausts?	10
Village	0

26. Consume betel

Consume betel daily with Tabaco	10
Consume betel occasionally with Tabaco	7.5
Consume betel occasionally with Tabaco	5
Consume betel without Tabaco	2.5
Not consume Tabaco or betel	0

27. How many hours of sleep do you have on average per night?

0-4	10
5-6	5
7-8	2.5
More than 8 hours	0

28. Do you experience followings when sleeping?

Insomnia, difficulty falling asleep or interrupted sleep	10
Snoring	5
None of the above	0

29. Events

Death of spouse	10
<ul style="list-style-type: none"> <li>• Death of family member</li> <li>• Divorce/separation</li> <li>• Marital reconciliation</li> <li>• Jail term</li> <li>• Major illness/injury/surgery</li> </ul>	9
<ul style="list-style-type: none"> <li>• Marriage</li> <li>• Dismissal from work</li> <li>• Retirement</li> <li>• Isolation due to COVID</li> </ul>	8
<ul style="list-style-type: none"> <li>• Death of a friend</li> <li>• Illness in the family</li> </ul>	7
<ul style="list-style-type: none"> <li>• Pregnancy</li> <li>• Moving to a new town/city/country</li> <li>• Family/relationship disputes</li> <li>• Working from home due to COVID</li> </ul>	5
<ul style="list-style-type: none"> <li>• Change in financial status</li> <li>• Change in occupation</li> <li>• Change in the work responsibilities</li> <li>• Mortgage</li> <li>• Major family events – Wedding, births in the immediate family</li> <li>• Major conflict/ issue/argument with a child</li> <li>• Personal difficulties at work</li> </ul>	4
<ul style="list-style-type: none"> <li>• Outstanding personal achievement</li> </ul>	2

<ul style="list-style-type: none"> <li>• Change in residence</li> <li>• Change in your or child's education institutes</li> <li>• Change in social habits</li> <li>• Change in routine</li> <li>• Holidays</li> <li>• Seasonal events like new year celebrations, Christmas etc</li> <li>• Minor violations of the law</li> </ul>	
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--

30. Do you participate in any of the following activities for more than an hour a week?

Meditation / prayer	0
Yoga /stretching / relaxation exercises	2.5
Community events/ social activities / sports	5
Play with pets	7.5
Non of above	10

31. Stress

	Weekly or more	Monthly or more
Do you feel anxiety, worry, fear, sudden feelings of panic, inability to control breathing and accelerated heart rate when upset, or recurrent feelings of unease?	10	5
Do you have feelings of sadness, depression, hopelessness, apathy, gloom, helplessness, isolation, loneliness, or lack of interest in social interaction?	8	4
Are you easily angered or frustrated, feel resentment or hostility towards others or frequently irritable?	6	3

32. Do you regularly experience lower abdominal pain, gas, bloating, diarrhoea, constipation, excessively smelly stools and/or a feeling that your bowels do 1t completely empty?

Yes	2
No	1

33. Have you taken the oral contraceptive pill for more than 6 months in the last year?

Yes	2
No	1

34. For what length of time have you been on antibiotics in the last year?

Not taken recently	0
Less than 2 weeks	2
2 weeks – 2 months	4
2 – months	6
Longer than 6 months	10

35. How often do you usually eat fried foods?

Less than once a week	0
1 – 2 times a week	0
3 – 6 times a week	5
Every day	10

36. How many serves of bread, pasta, rice, potatoes or other starchy foods do you have a day?

0 – 1 serves daily	0
2 serves daily	0
3 serves daily	2
4 or more serves daily	4

37. How many servings of sweet foods like cakes, biscuits, and/or chocolate do you consume a day?

Usually none	0
1-2 serves daily	2
More than 3 serves daily	8

38. How many teaspoons of sugar do you consume daily in hot drinks, added to foods, etc.?

0-3	0
4-6	1
7-9	4
10 or more	7

39. How many teaspoons of salt do you consume daily in food and beverages

0-3	0
4-6	2
7-9	8
10 or more	10

40. How often do you usually eat fish?

Rarely	10
1 – 2 times a week	5
3-6 times a week	2
Every day	0

41. How many pieces of fruit do you usually eat per week?

Usually none	10
1-3 pieces	2

4 or more pieces	0
------------------	---

42. How many serves of vegetables (excluding potatoes) do you usually eat a day?

usually none	10
1-2 serves daily	5
3-4 serves daily	3
5 or more serves daily	0

43. How many cups of coffee do you usually drink a day?

usually none	10
1-2 cups daily	5
3-4 cups daily	2
5 or more cups daily	0

44. How much soft-drink do you consume on average?

Usually none	0
Less than 500 ml per week	2
1-2 liters per week	4
3-4 liters per week	8
5 or more liters per week	10

45. How much water do you consume on average?

0-500 ml	10
501 ml to 1.25 liters	5
More than 1.25 liters	0

46. How much coconut milk do you consume on average per day when preparing meal

0-1 cup	0
2-3 cups	2
3-5 cups	5
More than 5 cups	10

47. Do you experience any of the following symptoms more than once a month?

Wheezing, sneezing, a runny nose, sore throat, itchy or watery eyes, coughing and/or blocked nose	5
Heart palpitations or headaches after certain foods	5
Non of the above	0

48. How frequently above pains occurred

Daily	10
Weekly	8
Monthly or less	6
Never	0

**49. Presence of Cardiovascular disease – Target variable**

<b>High risk - Have diagnosed cardiovascular disease, atherosclerosis, previous heart attack, stroke or PVD</b>	<b>10</b>
<b>Moderate risk - Experienced angina (heart pain) within the last 3 months</b>	<b>5</b>
<b>Low risk – Not experience any of above</b>	<b>0</b>



## MBA 3007 – Individual Project in Master of Business Analytics Thesis Submission Form

Recommendation by the Main Supervisor \*

### 1. Personal Details

Name with Initials	N.D.U. Gamage		
Full Name	Narmada Dushmanthi Udumalagala Gamage		
E-Mail Address	narmadadg@gmail.com	Registration No.	2018/BA/013
Phone Number(s)	0717404036	Index No.	


### 2. Research Project Details

Tentative Title of the Research	Data mining based approach to predict CVD risk in the Sri Lankan context
Area(s) of Research	Data mining

### 3. Supervisor Details

Please state the name (s) of the selected supervisors here. At least one of the supervisors should be a UCSC permanent senior academic staff member.

Main Supervisor (UCSC)		Co - Supervisor(Optional)- External / UCSC	
Name	Dr. M.G. Noel A.S Fernando	Name	
Position		Position	
		Organization	
		Qualifications	
		Tel. No.	
		E-mail	

Supervisor's Comments	<p>The work carried out in the study is satisfactory. But the contents of the reports are mixed up.</p> <p>After modifying the content arrangement, I recommended submitting the thesis. Recommended to re-arrange to contents of the last 3 chapters (methodology, design, conclusion) as instructions given in the body of the thesis.</p> <p>Mentioned the size of the local data set.</p> <p>Pay your attention correct the grammar and typos throughout the report (I corrected few mistakes and highlighted them)</p>		
Supervisor Recommendation	<input checked="" type="checkbox"/>	Recommend submitting the report with suggested modification	Do not recommend to submit
Name	MGNAS Fernando		Signature
			Date
			
			14/09/2021

\*Main Supervisor must be a UCSC senior academic staff member

Recommendation: Co-Supervisor (UCSC / External)			
Co-Supervisor's Comments			
Co-Supervisor Recommendation	<input type="checkbox"/>	Recommend to submit	<input type="checkbox"/>
			Do not recommend to submit



Name		Signature	
		Date	

## Summary of changes

#	Change Requested by the supervisor	<b>How</b> you addressed the supervisor request
1	Recommended to update the project title. Remove word "based" from the title  Previous title : Data Mining Based Approach to Predict CVD Risk in the Sri Lankan Context	Update the title by removing word "based"  Updated title : Data Mining Approach to Predict CVD Risk in the Sri Lankan Context
2	Show accuracy percentage briefly in the abstract	Added accuracy values for CVD prediction in global context and Sri Lankan context (for major, modifiable and contributing risk factors)
3	Label the chapters as follows in the table of content Chapter 1: introduction Chapter 2: Literature review. Chapter 3 design or methodology. Etc.	Table of content updated as mentioned
4	using the global data set, how could you Predict CVD Risk in the Sri Lankan Context? please justify???	Justification provided
5	Objectives are ok, what is data used in the SL domain?	To test CVD risk in Sri Lanka data gathered from 1252 individuals all over the country via a google document. The Sri Lankan data collection mechanism added the "objectives" sub topic
6	Clearly mention 2 phases of the research in the section 1.4	Phases define at the beginning of section 1.4
7	Rename the highlevel diagram	Renamed the diagram accordingly
8	Justify why component analysis applied on top of the implementation	Justify the reason for applying PCA for the classifier
9	Application of PCA shows minor improvement, Justify why it is getting minor improvement	Justification provided. PCA might not give significant results when use along with correlation robust algorithms
10	no of data records are sufficient/ significant to apply the data mining algorithms – Explain under phase 2 – section 3.2.1	No of data collected given
11	Provide the meaning of dataset1, dataset 2 in the section 3.2.3	Datasets define separately
12	Comment given for Chapter 4 :  In chapter 3, under the methodology, you have already explained how to achieve the objectives through the design approach. Is this results presentation chapter ??? not clear, is this chapter also belongs to design/methodology.  This chapter belongs to the results presentations. Arrange the information in	Chapter 4 and 5 rearranged. New sub section introduced to explain the evaluation and conclusion

## Summary of changes

	the design(methodology) and results presentation clearly, accommodating the correct contents. Contents are mixed up. Recommended to highlight the important conclusions in the conclusion chapter	
--	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--