# 10-Year Cardiovascular Disease (CVD) Risk Prediction of Sri Lankans: A Longitudinal Cohort Study

M.B Solangaarachchige
(Reg. No.: MS19805306)
M.Sc. in IT

Supervisor: Mr Prasanna S. Haddela

December 2021

Word Count 21,496

**Faculty of Graduate Studies & Research**
**Sri Lanka Institute of Information Technology**

# Acknowledgments

I'd like to express my gratitude to everyone whose encouragement and support helped me complete my thesis.

First and foremost, I want to thank my supervisor, Mr Prasanna S. Haddela, for his guidance, support, and encouragement in supporting the completion of this thesis. He has provided me with an abundance of invaluable advice and knowledge that will benefit me well in both my career and personal life. He was indeed the most patient and most understanding lecturer a student could ask for. Mr. Prasanna's guidance and teaching helped me understand and use machine learning techniques in this research. Thank you very much, Sir, for your unwavering support throughout this study.

Second, I'd like to express my heartfelt gratitude to Prof K.Chamila D.Mettananda for her invaluable guidance and unwavering support and assistance throughout this study. Prof Chamila's medical expertise was fundamental and extremely beneficial. Without her assistance, this research would not have been a success. Her guidance, support, and encouragement were priceless from the start of the project, when gathering the data, until the finish by validating the model.

I would also like to express my sincere gratitude to Vidyajyothi Professor H Janaka De Silva, Professor A R Wickremasinghe and Professor Anuradhani Kasthuriratne for believing in me and agreeing to provide a valuable dataset to make this project a success. Prof Janaka was extremely humble to discuss the feasibility of this study with me and my supervisor and ultimately decided to grant permission to use the dataset. Thank you, Sir, for making this possible.

I would like to thank our MSc in IT coordinator Mr. Samantha Rajapaksha for his support and valuable time and input throughout this research. His counsel to issues that developed over the two years of MSc degree and this thesis study is remarkable.

Lastly, I would like to express my sincere appreciation to my family for encouraging and supporting me throughout the study.

# Abstract

Cardiovascular diseases are one of the leading causes of mortality in the world. A cornerstone of preventive cardiology is identifying individuals at risk of cardiovascular diseases (CVD) at the earliest. Clinical guideline primarily recommends risk prediction models that are based on a limited number of predictors that perform poorly across all patient groups. Predicting cardiovascular risk is crucial for making treatment decisions, especially in the primary prevention of CVDs using a total risk approach. Despite the fact that several cardiovascular risk prediction models exist, only a handful are specifically designed for Asians, and none are generated from South Asians, including Sri Lankans. Machine learning (ML) and neural networks appear to be increasingly promising in supporting decision-making and forecasting from the huge amounts of data generated by the healthcare industry. This led us to develop a CVD model using Machine Learning to predict 10-year risk of developing a CVD in Sri Lankans. We investigated whether we could adopt ML to develop a model and whether there is an improvement in including non-traditional variables for the accuracy of CVD risk estimates and how to validate the ML model with existing WHO risk charts.

Using data on 2596 participants without CVD at baseline data collection of Ragama Medical Officer of Health (MOH) area in Sri Lanka, we developed a ML-based model for predicting CVD risk based on 75 available variables. However, the ratio of developing a CVD vs no CVD in 10 years was 7:93, which is extremely unbalanced. Therefore, at first, we derived a balanced dataset from the main dataset and build a ML model and it recorded an 80.56% accuracy. Secondly, to alleviate the dataset's imbalance, we adopted two techniques, which are 10-fold cross validation and stratified 10-fold cross validation (SKF) and trained six ML classification algorithms. They are Random Forest (RF), Decision Tree, AdaBoost, Gradient Boosting, K-Nearest Neighbor and 2D Neural Network. Out of these six algorithms RF model with SKF showed the highest accuracy in predicting a CVD event with an accuracy of 93.11%. Our ML model included predictors that are not usually considered in existing risk prediction models. Systolic blood pressure was the most important variable in this model. There were six non-traditional variables in the most ten important variable list and three of them were non-laboratory variables. To validate the model with existing WHO risk charts, we explored an experimental approach by developing a simple logistic regression function using the same techniques as the best selected model, with the seven traditional risk factors used in WHO risk charts and our

Random Forest model indicated the highest accuracy compared to the WHO model, with a difference of 26.20 %.

Our ML model improves the accuracy of CVD risk prediction in the Sri Lankan population. This approach justifies that the CVD prediction models also can be derived using ML for each subregion individually. Additionally, our research discovered novel CVD disease factors that may now be investigated in prospective studies.

**Keywords**: cardiovascular disease, risk assessment, models, machine learning, classification

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| IT | Information Technology |
| UN | United Nation |
| WHO | World Health Organization |
| CVD | Cardiovascular Disease |
| CVDs | Cardiovascular Diseases |
| ML | Machine Learning |
| SKF | Stratified K-Fold |
| RF | Random Forest |
| 2D | 2 Dimensional |
| EDA | Exploratory Data Analysis |
| RHS | Ragama Health Study |
| CV | Cardiovascular |
| CVEs | Cardiovascular Events |
| kNN | k-Nearest Neighbor |
| KNN | K-Nearest Neighbor |
| ROC | Receiver Operating Characteristic |
| AUC | Area Under the Curve |
| PR | Precision-Recall |
| IHD | Ischemic Heart Disease |
| NCDs | Noncommunicable Diseases |
| NCD | Noncommunicable Disease |
| MOH | Medical Officer of Health |
| GN | Grama Niladhari |