Modelling Cohort Specific Metabolic Syndrome and Cardiovascular Disease Risk Using Supervised Machine Learning

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Paulina Genet Ngcayiya
1355485

Supervised by Dr Pravesh Ranchod

August 30, 2023

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science

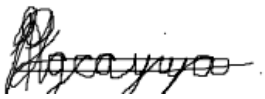
Abstract

Cardiovascular Disease (CVD) is the leading cause of death worldwide, with Coronary Heart Disease (CHD) being the most common type of CVD. The consequences of the presence of CVD risk factors often manifest as Metabolic Syndrome (MetS). In this study, a dataset from the Framingham Heart Study (FHS) was used to develop two different kinds of CHD risk prediction models. These models were developed using Random Forests (RF) and AutoPrognosis. Performance of the Framingham Risk Score model (AUC-ROC: 0.633) on the FHS dataset was used as the benchmark. The RF model with optimized hyperparameters (AUC-ROC: 0.728) produced the best results. This was by a very small margin to the AutoPrognosis model with an ensemble pipeline (AUC-ROC: 0.714). The performance of RF against AutoPrognosis when predicting the existence of MetS was evaluated using a dataset from the National Health and Nutrition Examination Survey (NHANES). The RF model with optimized hyperparameters (AUC-ROC: 0.851) produced the best results. This was by a small margin to the AutoPrognosis model with an ensemble pipeline (AUC-ROC: 0.851). Datasets, varying in size from 100 to 4900, were used to test the performance of RF against AutoPrognosis. The RF model with optimized hyperparameters had the best performance results.

Declaration

I, Paulina Genet Ngcayiya, hereby declare the contents of this dissertation to be my own work. This proposal is submitted for the degree of Bachelor of Master of Science (Dissertation) at the University of the Witwatersrand. This work has not been submitted to any other university, or for any other degree.

Portions of this work have appeared in the 23rd Annual International RAPDASA Conference joined by ROBMECH, PRASA and CoSAami which can be found in the MATEC Web of Conferences publication series [Ngcayiya and Ranchod 2022].



Signature

27 July 2023

Date

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr Pravesh Ranchod, for all the advice and guidance I received to help me complete my research and dissertation successfully. I am also grateful for the words of encouragement and motivation I was offered.

I would also like to thank my mother and all my friends for the support they've given me and for always believing in me.

Contents

List of Figures

List of Abbreviations and Acronyms

AUC-ROC Area Under the Curve of the Receiver Operating Characteristic.

BMI Body Mass Index.

BO Bayesian Optimization.

CART Classification and Regression Tree.

CHD Coronary Heart Diesease.

CVD Cardiovascular Disease.

FHS Framingham Heart Study.

FRS Framingham Risk Score.

GP Gaussian Process.

MetS Metabolic Syndrome.

NHANES National Health and Nutrition Examination Survey.

PSCP Pipeline Selection and Configuration Problem.

QDA Quadratic Discriminant Analysis.

RF Random Forests.

SCORE Systematic Coronary Risk Evaluation.

SMOTE Synthetic Minority Oversampling Technique.

THP Total Hip Osteoarthritis.

WHO World Health Organisation.

XGBoost eXtreme Gradient Boosting.

Chapter 1

Introduction

1.1 Problem Introduction

Cardiovascular Disease (CVD) is a medical term used to refer to a category of diseases that affect the heart and circulatory system (system of blood vessels). For most types of CVDs, there can be a minimal occurrence of symptoms related to the presence of the disease. Often, the experience of an acute event such as a heart attack, stroke or swelling in the arms and legs is the first warning sign of the presence of a CVD in the body. In some instances, such acute events can cause severe, long-term disability and even death [Car 2021].

Coronary Heart Disease (CHD) is the most frequently diagnosed type of CVD, killing approximately 9 million people each year [BHF 2022]. This disease occurs when blood vessels that supply blood and oxygen to the heart harden and become narrow due to a buildup of fatty substances known as cholesterol. This can eventually lead to the occurrence a heart attack or heart failure. The risk factors for CHD include the consumption of an unhealthy diet (high saturated fat intake), insufficient physical activity, alcohol abuse, and tobacco use [Car 2021].

The ramifications of the aforementioned risk factors manifest as Metabolic Syndrome (MetS). MetS is a medical term that refers to a group of conditions that can increase a person's risk of developing CVD and/or diabetes. MetS is characterized by hypertension, high blood glucose levels, unhealthy cholesterol levels, overweight, and obesity [Met 2021]. These factors are frequently included as core variables in models developed for CVD risk prediction. Other risk factors for CVD include having a family history of CVD, aging, poverty, and physiological factors such as stress. MetS affects an estimated 20–25% of the world's adult population. This population is three times more likely to suffer from a stroke or heart attack and twice as likely to die from the stroke or heart attack [Alberti *et al.* 2006].

Although the current statistics are alarming, clinicians reiterate that 80% of premature heart attacks and strokes can actually be prevented [WHO 2015]. Predicting future CHD risk and the presence of MetS at the individual, population, and subgroup levels will thus provide useful information to policymakers and healthcare authorities. This

can also motivate people to change their lifestyle choices, behaviours, and habits, as the majority of CHD/MetS risk factors are behavioural.

Over the past few years, multiple American and European countries have reported a reduction in CVD deaths [Sacco *et al.* 2016]. Unfortunately, this has not been the case for Africa. Figure 1.1 displays the changes in the number of CVD deaths, from 1990 to 2019, in the four Sub-Saharan African regions.

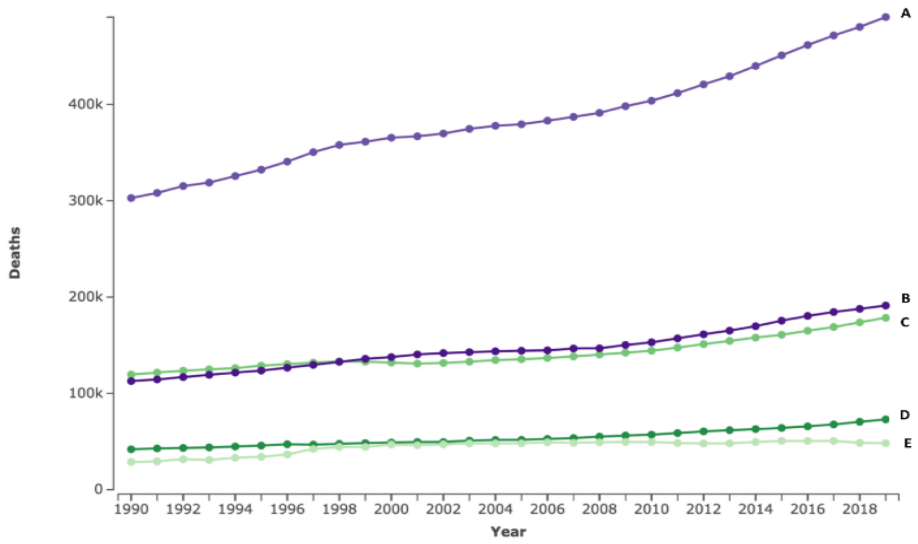


Figure 1.1: The number of CVD deaths, from 1990 to 2019. Line A represents all the four Sub-Saharan African regions combined. Lines B, C, D and E represent the Western, Eastern, Central and Southern regions, respectively [IHME 2020]

.

The majority of existing CVD risk prediction models, such as the Framingham Risk Score (FRS), were implemented and evaluated using data from American and European populations [Lloyd-Jones 2010]. As a result, these models may not accurately predict CVD risk for individuals from other regions, such as Asia or Africa, where lifestyle factors, socio-economics, and genetic predispositions may differ. In general, most conventionally developed models tend to either underestimate or overestimate CVD risk estimates.

In addition, many of these CVD risk prediction models rely on multivariate regression methods, which employ a small number of conventional risk factors and assume that the relationship between all of these factors and the CVD outcome is linear. These two characteristics limit the model's predictive performance for CVD, particularly for specific subgroups of the population, such as diabetics. Models based on ensemble machine learning algorithms often have the ability to realize more complex patterns within data of large repositories and can adequately model nonlinear relationships between the predictor variables and outcome. This is frequently advantageous, but can sometimes come at the expense of overfitting as well as time and resources.

However, in this research study, we consider the use of ensemble machine learning algorithms, namely Random Forest (RF) and Autoprognosis, with carefully regularized techniques to avoid overfitting with appropriate parameters and evaluate the effectiveness of the models using AUC-ROC, Accuracy, Confusion Matrices, F1 scores and Relative Entropy values.

1.2 Machine Learning Models

Expeditious advances in machine learning, increased computational power and digitized healthcare data have resulted in a growing number of applications for machine learning within healthcare. The development and implementation of machine learning models can be particularly useful for clinical prognosis (risk of future outcome) and diagnosis (prediction outcome of present condition). Supervised machine learning has previously been used to model multiple classification problems within the clinical space, with the most successful methods being RF and AutoPrognosis.

The RF algorithm [Breiman 2001] works by creating multiple decision trees which are trained independently on random subsets of the training dataset. The final output generated by the algorithm is the class, which is the mode of the predicted classes from the multiple decision trees. In the instance of regression, the final output is the mean of all the values predicted from the decision trees. The predictive performance of the RF on unseen data depends on the strength of each decision tree and on the correlation between any two trees in the forest.

AutoPrognosis is an automated machine learning framework which uses an advanced Bayesian optimization technique to automatically generate a prognostic model made up of a weighted ensemble of machine learning pipelines. Each pipeline is made up of a data imputation, feature pre-processing, classification, and calibration algorithm of its own. This method was proposed fairly recently by Alaa and Schaar [2018], whereby the practically and effectiveness of this method were established using nine major patient groups characterising various aspects of cardiovascular patient care.

The two aforementioned methods shall be discussed in further detail in Chapter 2 of this paper.

1.3 Results

The primary goals of this study were to identify the most significant risk factors associated with CHD and to conduct a performance evaluation study comparing the predictive performance of the FRS model to two different types of 10-year CHD risk prediction models based on RF and AutoPrognosis. This was carried out on a Framingham Heart Study (FHS) dataset of 4,240 people with 17 variables [Car 2020]. From the performance evaluation of the FRS model, an average AUC-ROC score of 0.633 on the FHS

dataset was achieved, which served as the benchmark. Our first research hypothesis was that ”both types of the CHD risk prediction models, based on RF and AutoPrognosis, would outperform the FRS model”. Results from the performance evaluation phase led to the acceptance this hypothesis. In fact, the RF model with optimized hyperparameters yielded the best performance results, with an AUC-ROC score of 0.728.

The performance of RF against AutoPrognosis for predicting the presence of Metabolic Syndrome was evaluated using a dataset from the National Health and Nutrition Examination Survey (NHANES) consisting of 7,821 records and 77 variables [Hoyt 2020b]. It was concluded that the RF model with optimized hyperparameters produced the best performance results of 0.851 for the AUC-ROC score. The results produced by the models for both 10-year CHD risk prediction and MetS detection, led to the rejection of the second research hypothesis that ”the AutoPrognosis models will always outperform the RF models”.

In some instances having an AutoPrognosis model with an ensemble pipeline can produce better results, as opposed to only having a single pipeline. However, it may not always be necessary to have an ensemble pipeline, especially if the model which uses a single pipeline produces satisfactory/comparable results. For CHD risk prediction, AutoPrognosis with a single pipeline (using the 7 conventional risk factors) had an AUC-ROC score of 0.703, which **slightly** outperformed AutoPrognosis with an ensemble pipeline (using the 7 conventional risk factors) with an AUC-ROC score of 0.696. The differences in performance can be almost negligible in some instances, therefore this led to the rejection of the third hypothesis that “The AutoPrognosis models with an ensemble pipeline will always outperform the AutoPrognosis model with a single pipeline, for both 10-year CHS risk prediction and MetS prediction respectively”.

In a similar manner, RF was evaluated against AutoPrognosis using various datasets ranging in size from 100 to 4900. The RF model with optimized hyperparameters produced the best performance results, followed by AutoPrognosis consisting of an ensemble pipeline, then AutoPrognosis consisting of a single pipeline, and finally the RF model with hyperparameters set to their default values. The results led to the acceptance of the fourth research hypothesis that ”AutoPrognosis will have satisfactory performance results on all the different sample sizes of data”, however the fifth research hypothesis, ” AutoPrognosis will always outperform RF on all the different sample sizes of data”, was rejected.

Once again, detailed information as to how these results were obtained and interpretations of the results are provided in Chapter 3 and Chapter 4, respectively.

Chapter 2

Background Information

2.1 Introduction

Chapter 1 briefly introduces an overview for the problem area of the research study presented in this paper. In chapter 2, information on the medical background and methods applicable to the research study are presented thoroughly. Sections 2.2, 2.3 and 2.4 provide a concise medical definition and understanding of CVD, CHD and MetS, respectively, and explores their impact on the global population. The Framingham Risk Score and the Systematic Coronary Risk Evaluation Charts are some of the most notable CVD risk prediction models used in the clinical space. Section 2.5 provides information on how these models were created and explores their usefulness. Two machine learning methods, based on Random Forests and AutoPrognosis, were used for the creation of the models produced by our research study. A description of the Random Forest algorithm and AutoPrognosis is presented in thorough detail in sections 2.6 and 2.7, respectively. Finally, section 2.8 concludes chapter 2 by providing a summary of the main points covered.

2.2 Cardiovascular Disease

At the start of the 20th century, less than 10% of global deaths were caused by CVDs. However, at present CVD is responsible for 32% of global deaths annually [Car 2021]. This accounts for an estimate of 17.9 million deaths per year, making CVD the current leading cause of death worldwide. CVD includes a wide range of illnesses, relating to diseases of the cardiac muscle and circulatory system (the body's network of blood vessels supplying the brain, heart and other vital organs).

More specifically, these diseases can be further classified into six categories [Car 2021]:

- Peripheral Arterial Disease – develops in the blood vessels supplying arms and legs.
- Congenital Heart Disease – an abnormality/defect in the structure of the heart which exists at birth.
- Coronary Heart Disease – develops in the blood vessels supplying the heart muscles.
- Deep Vein Thrombosis and Pulmonary Embolism – blood clots that develop in the leg veins, which can shift and travel up to the lungs and heart.
- Cerebrovascular Disease – develops in the blood vessels supplying the brain.
- Rheumatic Heart Disease – a deterioration of the heart valves and muscles as a result of rheumatic fever caused by streptococcal bacteria.

85% of CVD related deaths are caused by the onset of severe and sudden conditions; namely strokes and heart attacks [Car 2021]. A stroke or heart attack is usually the first warning sign of the presence of a CVD in the body. These conditions occur when plaque (fatty deposits) builds up in the blood vessels supplying the heart or brain, this causes the walls of these blood vessels to thicken, resulting in a blockage or constriction that reduces blood flow and prevents oxygen and other nutrients from reaching the brain or heart. Furthermore, a stroke can also be caused by a blood vessel rupturing and bleeding in the brain, or by the formation of a blood clot blocking off blood flow to the brain. The acuteness of these conditions can be fatal or cause chronic disabilities. These conditions are caused by the existence of a combination of risk factors in people's daily lives. The risk factors include consuming an unhealthy diet, lack of an adequate amount of physical activity, obesity, alcohol abuse, tobacco use, diabetes, hypertension and hyperlipidaemia [Car 2021].

The severity and number of behavioural risk factors can often be largely influenced by the social, cultural and economic changes in a society. It is for this reason that the prevalence of CVD differs in different regions of the world. In previous years, most CVD deaths occurred in developed countries. However, at present approximately 80% of these deaths now occur in developing countries. Many developing countries are experiencing increased urbanization and therefore changes in lifestyle choices. This has resulted in a direct relationship between increased urbanisation and an increased prevalence of CVD in these countries [Vorster 2002].

Since 2018, it has been reported that approximately 55% of the global population lives in urban areas. Furthermore, it is predicted that 68% of the global population will live in urban areas by 2050 [United Nations 2018]. The proliferation of restaurants, fast-food franchises and food vendors in urban areas means that people with in settings have increased and easy access to processed unhealthy foods which ultimately increases

the urban population's dietary consumption of saturated fats and animal protein, leading to higher "bad" cholesterol levels making their environment highly obesogenic. The urban population is more likely to have lower levels of physical activity due to their availability of motorised transportation, food and goods delivery services, jobs that rely on minimal physical activity and lifestyles which provide access to sedentary activities such as watching television, video gaming, etc. Some studies based on African populations have found that residents in urban areas are at a higher risk of tobacco use[Pampel 2005] and are more likely to start at a younger age [Townsend *et al.* 2006], due to the rise of psychological distresses as the cost of living in urban settings is higher.

It is important to note that, unlike in developed countries, people living in developing countries do not have the added benefit of integrated health care programmes for early detection and preventative treatment for high-risk individuals, which lessens the access these people can have to consistent and beneficial primary health care services which are sufficient for their needs. This results in late detection and premature deaths from CVD. CVD risk prediction will be particularly useful for prevention, early detection and management purposes to improve the overall health status of a nation and alleviate the economic burden caused by an increasing incidence of CVD.

2.3 Coronary Heart Disease

To ensure that the heart is kept pumping, blood vessels known as the coronary arteries need to supply the heart muscles with oxygen-rich blood. There are four main coronary arteries, namely the left coronary artery, right coronary artery, left circumflex artery and left anterior descending artery. These coronary arteries are located directly on top of and wrapped around the heart. Plaque made up of fatty substances, cholesterol, fibrin (the main component of blood clots), calcium, and waste products can build up in the inner lining of the body's arteries and cause what is known as Atherosclerosis. If the build-up of plaque continues this can clog up and damage the coronary arteries by making them harden and become narrow, which can limit or even stop the supply of oxygen-rich blood to the heart muscles. When Atherosclerosis specifically affects the coronary arteries it is called Coronary Heart Disease (CHD).

As the Atherosclerosis persists and the coronary arteries become even more clogged, the heart will have to do more work by pumping harder in order to still have the ability to push blood through the narrower and hardened arteries. It is important to note that CHD can go unnoticed for a very long time, because plaque build-up in the arteries can take years to decades before it becomes significant enough for someone to start developing symptoms. The most common symptoms include chest pain/discomfort, fatigue and shortness of breath after moderate physical activity and even at rest. Sometimes a person may not know that they have CHD until after they've had a heart attack. The occurrence of a heart attack is a medical emergency as it can cause permanent heart muscle damage and may even lead to death.

Globally, an estimate of 200 million people (110 million men and 60 million women)

live with CHD. This makes it the most commonly diagnosed type of CVD. Each year, CHD accounts for approximately 9 million deaths. Furthermore, for at least 3 decades now, CHD has been the leading cause of mortality worldwide [BHF 2022]. Individuals who are at an increased risk of developing CHD are those who have abnormal cholesterol levels, high blood pressure, overweight/obesity, diabetes, have a family history of CVD and are smokers. As a person ages their risk of CHD increases. According to the [Cleveland Clinic 2022] the risk for men increases at age 45, and for women at age 55. In some ethnic groups, the risk for CHD is higher, as people with similar cultural backgrounds tend to have similar traditional lifestyle practices (e.g. in diet choices) and share certain types of genes that make them predisposed to CHD. In South Africa, for instance, research performed by [Vorster 2002] concluded that the death rates from CVD asserted that stroke is a large public health predicament among black South Africans.

Several kinds of diagnostic tests can be conducted to determine if a person has CHD. Common tests include (this is not an exclusive list) [CHD 2021]:

- Coronary Artery Calcium Scan – a CT scan of the coronary arteries to check for calcium deposits and plaque build-up.
- Electrocardiogram Test– a test that uses electrodes placed on the chest and limbs to measure and record the heart's electrical activity to detect ischemia, an abnormal heartbeat rhythm and/or heart attack.
- Exercise Stress Test – a test to measure a person's heart rate while walking on a treadmill, to evaluate how well the heart functions during physical activity when it has to pump more blood.
- Echocardiogram Test– a test that uses ultrasound to generate an image of the heart to check its structure and function.
- Cardiac Catheterization Test – a test that involves placing small, flexible tubes through an artery in the neck, arm or groin to reach all the way to the heart. Using this, the intensity of blood flow through the heart's chambers and the heart's blood pressure can be measured. This method can also be used to collect blood samples from the heart.
- Chest X-ray Test – a test that involves the use of x-rays to generate an image of the heart, lungs, and organs in the chest.
- Blood Tests – tests that involve the collection of blood samples to check for levels of glucose, cholesterol, triglycerides, lipoprotein, c-reactive protein, etc, that may affect the arteries.

Once someone has been diagnosed with CHD, a health care professional will administer a treatment plan suited for the individual to reduce the risk of complications associated with CHD from arising. The first step involves making lifestyle changes, to decrease the presence of risk factors. Lifestyle changes include making the decision to quit smoking, limiting alcohol consumption, increasing levels of physical activity and diet re-adjustments. Medication can be prescribed to treat conditions that contribute to

CHD and to reduce the risk of developing blood clots.

Additionally, different procedures and surgery can be offered as treatment options. Balloon Angioplasty and Stenting are some very common procedures. These procedures involve making a small incision and inserting a catheter in an artery of the leg or wrist and guiding it to the clogged and narrow section of the coronary arteries. A balloon is then inflated to increase the diameter of the affected coronary artery to restore the flow of blood to the heart. A small metallic spring-like scaffold, called a stent, is then left in place to keep the artery open. Coronary Artery Bypass Graft is a another common surgery performed to restore blood flow to the heart by using a blood vessel that has been removed from other parts of the body to create a new detour pathway around the clogged arteries, to the heart. Occasionally, when medicine and traditional approaches are not successful, a heart transplant may be needed.

The World Health Organisation (WHO) has proposed a number of cost-effective and evidence-based strategies for preventing and controlling Non-Communicable Diseases such as CVDs [WHO 2007]. Examples include putting policies in place to reduce the salt and sugar content in processed food, increasing taxation on tobacco and alcoholic products, legislation to restrict public drinking and smoking zones and raising public awareness of CVD risks, dangers and solutions. From a clinical perspective, it has become mandatory to strengthen the responsiveness of the health care system to provide more access to essential preventive therapies and treatments such as statin therapy, to improve and build more health care facilities and infrastructure in remote areas, to increase the health workforce, and even incorporate the use of technology and machine learning approaches for CVD prognosis. Multiple clinical models exist and are currently being used to predict the risk of an individual developing a CVD in the future, to aid with prevention.

2.4 Metabolic Syndrome

Firstly, it is important to note that, Metabolic Syndrome (MetS), also commonly known as Syndrome X, is not a disease. It is a medical term used to refer to a collection of conditions. MetS includes high blood pressure, high blood glucose levels, unhealthy cholesterol levels and overweight/obesity [Met 2021]. Certainly, having any one of these conditions isn't ideal however, having a combination of the conditions is even more problematic as it can lead to some serious health problems. As discussed in Section 2.2 and 2.3, the occurrence of these conditions can result in the development of Atherosclerosis. The presence of MetS can increase an individual's chance of developing a CVD. It is estimated that almost 20-25% of the world's adult population live with MetS. Compared to people living without MetS, this population is three times as likely to suffer from a stroke or heart attack, and is twice as likely to die from a stroke or heart attack [Alberti *et al.* 2006]. MetS also increases an individual's risk of developing type 2 diabetes by five times. Type 2 diabetes accounts for 90% of all diabetes diagnoses worldwide [Alberti *et al.* 2006].

Symptoms associated with MetS develop over time, as such the appearance of symptoms is delayed. MetS may be caused by ageing, physical inactivity, hormonal changes, proinflammatory state and genetics, where the role of these factors may differ across ethnic groups. Central obesity and insulin resistance are considered the most significant risk factors.

According to the American Heart Association and the National Heart, Lung and Blood Institute, a person is diagnosed with MetS if they have a combination of any three or more of the following conditions [WebMD 2021]:

High Blood Pressure	Having a blood pressure of 130/85 or higher **OR** taking anti-hypertensive medicine
High Fasting Blood Sugar Level	Having a sugar level of 100 mg/dl or higher **OR** taking glucose-lowering medicine
Cholesterol (High Triglycerides)	150 mg/dL or higher **OR** taking cholesterol medicine
Cholesterol (Low High-Density Lipoprotein)	Less than 40mg/dL (for men) and less than 50 mg/dL (for women) **OR** taking cholesterol medicine
Large Waistline	40 inches or more (for men) and 35 inches or more (for women)

Table 2.1: Diagnostic Criteria for MetS

Due to the fact that MetS drives the two global epidemics of CVD and diabetes, it has become imperative to make early detections of MetS, so that lifestyle changes and treatments can be administered to prevent the onset of diabetes and/or CVD. MetS is prevented in the same way that it is treated through incorporating regular exercise, consuming a healthy diet and making the decision to quit smoking.

2.5 Some notable clinical models

2.5.1 The Framingham Risk Score

In 1998, the Framingham Risk Score (FRS) model was initially created to determine the risk of an individual developing CHD [Wilson *et al.* 1998]. The dataset used was from the Framingham Heart Study (FHS). This study is based on a largely white American population from Framingham, Massachusetts. The study began in 1948 and is still ongoing to date. It is important to note that FRS is based on multivariate regression

methods. Multiple adaptions and revisions of the FRS model have been created since 1998. Some noteworthy FRS versions include:

- The *2002 adaption of the Third Report of the National Heart, Lung and Blood Institute produced by the [NCEP Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults 2002] (FRS-ATP-III)*. Within this adaptation the diabetes status and family history risk factors were removed and replaced with the impact of hypertension treatment. Only the hard coronary heart disease endpoints were used in the calculations.

- The *2006 Lifetime Framingham CVD Risk Score Model* [Lloyd-Jones *et al.* 2006]. This adaptation estimates CHD risk for individuals aged 50 and above and is based on only four risk factors: systolic blood pressure, smoking status, diabetes status and total cholesterol levels.

- The *2008 Ten-Year Framingham CVD Risk Score Model* [D'agostino *et al.* 2008]. This adaptation was inclusive of additional CVD events (i.e. cerebrovascular events, transient ischemic attack and peripheral artery disease) which were previously not part of the model's derivation.

The *2008 Ten-Year Framingham CVD Risk Score (FRS-CVD) Model* [D'agostino *et al.* 2008] was the model chosen to be used in the study outlined by this paper. Over the years, researchers have formulated models to predict the risk of developing specific CVD events, like a stroke or heart attack, as opposed to general CVD risk. Disease-specific models exist because the influence of CVD risk factors may differ from one type of CVD to another. However, our chosen model (FRS-CVD Model) was created based on the proven fact that there exists enough commonality in the influence of risk factors to justify the development of a single general CVD risk prediction model, that can predict both global CVD risk and a specific CVD type risk with satisfactory accuracy results. The FHS defines CVD as a disease category composed of cerebrovascular events (transient ischemic attack, haemorrhagic stroke, ischemic stroke), peripheral artery disease (intermittent claudication), CHD (angina, myocardial infarction, coronary death, coronary insufficiency) and heart failure [Cupples 1987].

In the development of the FRS-CVD Model, Cox proportional hazards regression methods [Cox 1972] were used to model the relationship between CVD risk factors and the probability of a first CVD event occurring during a twelve-year follow-up period, over which the data used for the model was collected. The data came from 8491 (4522 women; 3969 men) participants from the FHS, who were between the ages of 30 and 74, free of CVD at the beginning of the follow-up period and were present for routine examination procedures. The summary statistics for the risk factors used in the FRS-CVD Model are shown in Figure 2.1.

Characteristics	Women (n=4522, 28% FOC)	Men (n=3969, 22% FOC)
Age, mean (SD), y	49.1 (11.1)	48.5 (10.8)
Total-C, mean (SD), mg/dL	215.1 (44.1)	212.5 (39.3)
HDL-C, mean (SD), mg/dL	57.6 (15.3)	44.9 (12.2)
Systolic BP, mean (SD), mm Hg	125.8 (20.0)	129.7 (17.6)
BP treatment, n (%)	532 (11.76)	402 (10.13)
Smoking, n (%)	1548 (34.23)	1398 (35.22)
Diabetes, n (%)	170 (3.76)	258 (6.50)
Incident CVD events, n (%)	456 (10.08)	718 (18.09)
FOC indicates Framingham original cohort; Total-C, total cholesterol; HDL-C, HDL cholesterol; and BP, blood pressure.		

Figure 2.1: Summary statistics for the risk factors used in the FRS-CVD Model [D'agostino *et al.* 2008]

From the Cox models, the mathematical CVD ten-year risk prediction functions were produced. The risk factors included in the Cox models were diabetes status, current smoking status, high blood pressure medication usage, systolic blood pressure, HDL cholesterol, total cholesterol and age. The CVD risk prediction functions are sex-specific. Other risk factors such as triglycerides, body mass index and diastolic blood pressure can be useful to consider however they were excluded from the main models as they were not statistically significant enough. To minimise the influence of outlier data points, and improve the discrimination and calibration of the prediction models, a natural logarithm transformation was applied to the covariates for risk factors which have continuous values, e.g. total cholesterol.

The general equation of the Cox model looks like this:

$$\hat{p} = 1 - S_0(t)^{exp(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \bar{X}_i)} \tag{2.1}$$

where $S_0(t)$ is the baseline survival value at the follow-up time t=10 years, β_i is the regression coefficient estimate, X_i is the value of the i-th risk factor (the value is log-transformed if it is continuous) and $\bar{X}_i$ is the corresponding mean value, p is the total number of risk factors.

The Regression Coefficient used in the model are shown in Figure 2.2

Variable	β*	P	Hazard Ratio	95% CI
Women [So(10)=0.95012]				
Log of age	2.32888	<0.0001	10.27	(5.65–18.64)
Log of total cholesterol	1.20904	<0.0001	3.35	(2.00–5.62)
Log of HDL cholesterol	−0.70833	<0.0001	0.49	(0.35–0.69)
Log of SBP if not treated	2.76157	<0.0001	15.82	(7.86–31.87)
Log of SBP if treated	2.82263	<0.0001	16.82	(8.46–33.46)
Smoking	0.52873	<0.0001	1.70	(1.40–2.06)
Diabetes	0.69154	<0.0001	2.00	(1.49–2.67)
Men [So(10)=0.88936]				
Log of age	3.06117	<0.0001	21.35	(14.03–32.48)
Log of total cholesterol	1.12370	<0.0001	3.08	(2.05–4.62)
Log of HDL cholesterol	−0.93263	<0.0001	0.39	(0.30–0.52)
Log of SBP if not treated	1.93303	<0.0001	6.91	(3.91–12.20)
Log of SBP if treated	1.99881	<0.0001	7.38	(4.22–12.92)
Smoking	0.65451	<0.0001	1.92	(1.65–2.24)
Diabetes	0.57367	<0.0001	1.78	(1.43–2.20)
So(10) indicates 10-year baseline survival; SBP, systolic blood pressure.				
*Estimated regression coefficient				

Figure 2.2: The Regression Coefficient and Hazard Ratios of the FRS-CVD Model [D'agostino *et al.* 2008]

As an alternative to using a direct application of the mathematical CVD risk prediction functions (based on the Cox model), CVD Risk Prediction Score Sheets [D'agostino *et al.* 2008] were derived for women and men, respectively. These score sheets are simpler to use and produce results just as accurate as directly applying the Cox models.

In terms of both discrimination and calibration, both versions of the male and female specific CVD risk prediction functions produced satisfactory results. Discrimination was measured by calculating the C-statistic [D'Agostino and Nam 2003] [Pencina and D'Agostino 2004], which ranged from 0.763 in men to 0.793 in women. Calibration was measured by calculating the Hosmer-Lemeshow (X^2) statistics [D'Agostino and Nam 2003], which were 13.48 (for the lack of fit, P=0.14) for men and 7.79 (for the lack of fit, P=0.56) for women.

Additionally, as an alternative to the main CVD risk prediction models and score sheets, simplified prediction models were developed using simple non-laboratory predictors that are obtained through routine office-based primary health care and do not need to go through laboratory processing and testing. For these models, the total cholesterol and HDL cholesterol predictors were replaced with body mass index. The rest of the variables remained unchanged and the same modelling procedures and assessment techniques were utilized for these simpler models.

2.5.2 The Systematic Coronary Risk Evaluation Charts

The Systematic Coronary Risk Evaluation (SCORE) Charts are another common and widely used method for CVD risk prediction, which we will only be discussed briefly as it was not included in our research. However, it is still worth presenting in order to provide us with more context to the background of the research. The SCORE project [Conroy *et al.* 2003] was established to create a risk scoring system to aid in the clinical management of CVD risk, subject to the clinical practice in Europe. The system predicts the likelihood of a fatal atherosclerotic CVD event occurring over a ten-year duration. The model was created using datasets from 12 European countries with varied cohorts being included to represent different levels of CVD risk across Europe. The SCORE risk charts include age, gender, total cholesterol, systolic blood pressure and smoking status for its risk factors.

The Weibull proportional hazards model [Barrett 2014] was used to calculate the underlying risk functions of the SCORE risk charts. The Weibull model is a parametric regression model which has two parts. One part of the model estimates the baseline survival function, and the other part estimates the relative risks correlated to the risk factors. A threshold value of 5% is used to determine if a person is at a high risk of experiencing a fatal cardiovascular event. [Conroy *et al.* 2003] evaluated the consequences of regression dilution bias (a bias in model coefficients cause by noise or measurement errors) on the risk estimates produced by SCORE and it was found that the effect on individuals whose risk lies between 2% and 5% were negligible, while individuals with very high or very low-risk estimates were affected significantly. Additionally, the SCORE risk charts only consider primary risk factors leaving out other factors such as a family history of CVD prevalence and blood glucose levels among others, which are generally known to also be important factors considered for CVD risk assessment in the clinical practice.

2.6 Random Forest Algorithm

Research in clinical prognosis is aimed at developing useful predictive models that can be used to advise clinicians about possible future outcomes of a patient's clinical condition. This information can be used to guide screening and treatment decisions. Expeditious advances in machine learning and the recent abundance of digitized healthcare data has resulted in a growing number of applications for machine learning within healthcare. Medical repositories often have vast quantities of data which has not been mined effectively. Machine Learning techniques, such as those based on the Random Forests algorithm, are effective when working with large volumes of data to efficiently establish variable importance, relationships among the variables and to make predictions (within both classification and regression problems). The Random Forests algorithm is a supervised machine learning algorithm that was developed by Leo Breiman [Breiman 2001].

The formal definition of a Random Forest is as follows:

"A random forest is a classifier consisting of a collection of tree-structured classifiers $h(x, \Theta_k), k = 1, ...$ where the Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x."

To put it simply, a random forest is an ensemble of multiple decision trees, where each decision tree is trained independently on a random subset of the training dataset. A decision tree can be graphically represented as a tree-like model of decisions. Each tree consists of nodes and edges ("branches") which connect the nodes. Each node represents a chosen feature/variable, while the edges represent a range of values for a particular feature. The range of values contains the partition points for each chosen feature, respectively. When a decision is made, the data is grouped according to the values of a chosen feature characteristic of the data thereby forming subsets of data. This step is applied to each subset of the data recursively. The recursive process continues until all the data items within the current subset reside in the same class. It is important to note that a decision tree is constructed from data that has already been pre-classified. There are different metrics used to decide which feature will be chosen to form the best split at each iteration of the recursion.

For a random forest, the input features are different from one tree to another. Each decision tree is constructed from a random subset of the primary feature set.

Within the random forest, each decision tree will produce a class prediction for a given input. The final output generated by the random forest will be the class which is the mode of the predicted classes from the multiple decision trees. However, in the instance of regression, the final output will be the mean of all the values predicted from the decision trees. The performance of a random forest on unseen data depends on two things: the strength of each decision tree in the random forest and the correlation between the trees.

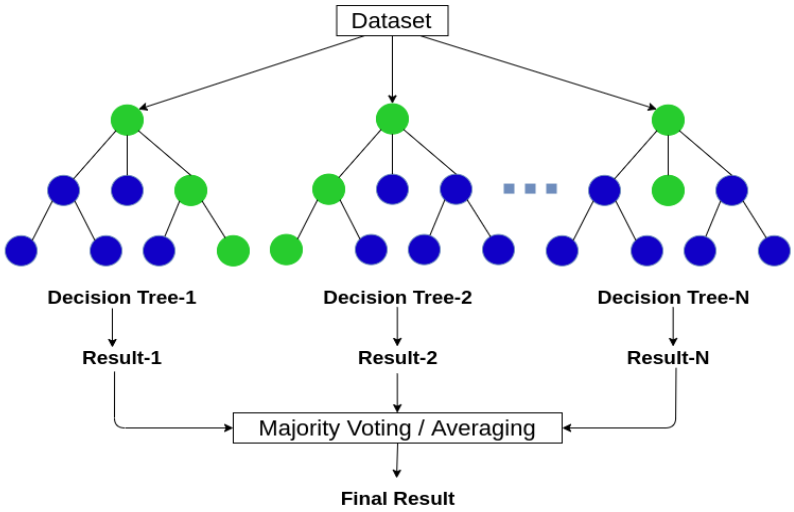


Figure 2.3: A Random Forest [Dec 2020]

The pseudocode for the random forest algorithm is displayed in Figure 2.4. This algorithm provides a more detailed explanation of the description above about how a random forest is created. In the algorithm, we can see that for each tree in the forest, a bootstrap sample is selected from the training dataset **S**, where $\mathbf{S}^i$ is the i-th bootstrap sample (line 4). For each bootstrap sample, a modified version of the decision tree learning algorithm is used to learn a decision tree (line 5). The modification in the decision tree learning algorithm is that at each tree node, instead of finding all the possible feature splits, a subset of all the features ($\mathbf{f} \subseteq \mathbf{F}$; **F** is the feature set) is randomly selected (line 12). Then at the node, a split is performed on the best feature in f (line 13). Making a decision on which feature to split can be computationally expensive however, it is important to note that **f** is much smaller in size than **F**. Therefore, by reducing the size of the feature set, the time taken to learn a tree is reduced.

Algorithm 1 Random Forest

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
1  function RandomForest(S, F)
2      H ← ∅
3      for i ∈ 1, ..., B do
4          S^(i) ← A bootstrap sample from S
5          h_i ← RandomizedTreeLearn(S^(i), F)
6          H ← H ∪ {h_i}
7      end for
8      return H
9  end function
10 function RandomizedTreeLearn(S, F)
11     At each node:
12         f ← very small subset of F
13         Split on best feature in f
14     return The learned tree
15 end function
```

Figure 2.4: Pseudocode for Random Forest Algorithm

Lowering the error rate of a decision tree makes it a stronger classifier, therefore increasing the strength of the individual trees will ultimately improve the overall performance of the random forest. Decreasing the correlation between trees increases the discriminative power of the random forest, therefore, reducing its error rate. Breiman [2001] introduces the correlation and strength attributes through an upper bound for the generalization error given by:

$$PE^* \leq \frac{\overline{\rho}(1 - s^2)}{s^2} \tag{2.2}$$

where s is the strength and ρ is the correlation

[Breiman 2001] presented empirical results on strength and correlation and concluded that the lower this ratio the higher the accuracy of the random forest.

This method combines the bagging technique [Breiman 1996] and the concepts of random split selection proposed by [Dietterich 2000], the random subspace method by [Ho 1998], and geometric feature selection by [Amit and Geman 1997]. In the traditional bagging algorithm, the same features are likely to be repeatedly used for splitting the bootstrap samples. Therefore, the individual decision trees may end up being highly correlated. In the random forest algorithm, each split test is restricted to a smaller, random sample of the feature set. This reduces the correlation between the trees in the ensemble. Additionally, since the features considered at each node are restricted to a subset of the feature set, each decision tree can be learned faster, hence more decision trees can be learnt in a given time. The more decision trees are used the better the performance of the random forest. However, the gained improvement in performance decreases as the quantity of trees increases. Thus eventually, the benefit of improved performance will be lower than the cost of computational time and resources for adding more trees.

Overfitting is a term used in machine learning to refer to the occurrence of a modelling error that happens when a model is trained to fit the training dataset too closely. Therefore, the model not only captures information about the existing data patterns in the training dataset but also some specific irrelevant details such as noise and outliers. This affects the model's generalisation ability, making it have a poor predictive performance on unseen data. However, the bagging technique is known to decrease variance. As mentioned previously, the random forest algorithm incorporates bagging making it robust to noise, outliers and overfitting.

As published in the original paper on random forests, other advantages include:

- It is a simple method that can easily be parallelised.
- Its performance accuracy is just as good, and sometimes even better than Adaboost.
- Performance is faster than boosting or bagging.
- Useful estimates of variable importance, strength, correlation, and error rates are automatically generated.

2.7 AutoPrognosis

2.7.1 Description of AutoPrognosis' components

It has been establish that there is a progressive number of applications for machine learning models and techniques in prognostic research. However, there often exists a rift between the potential and actual practicality of these machine learning approaches. This is because clinicians without any adequate knowledge and experience in data science are challenged with having to manually design and tune machine learning modelling pipelines before they can put them into use. AutoPrognosis was developed to circumvent this challenge, as it is an automated machine learning framework specifically designed for clinical prognosis. To put it very briefly, AutoPrognosis takes a dataset for a cohort of patients as its input and uses the data to automatically configure machine learning pipelines to ultimately produce a prognostic model which predicts the patients' risks accompanied with explanations of the predicted outcomes, as its output. Figure 2.5 provides a diagrammatic depiction of how AutoPrognosis works, which we will now discuss in detail in reference to this figure.

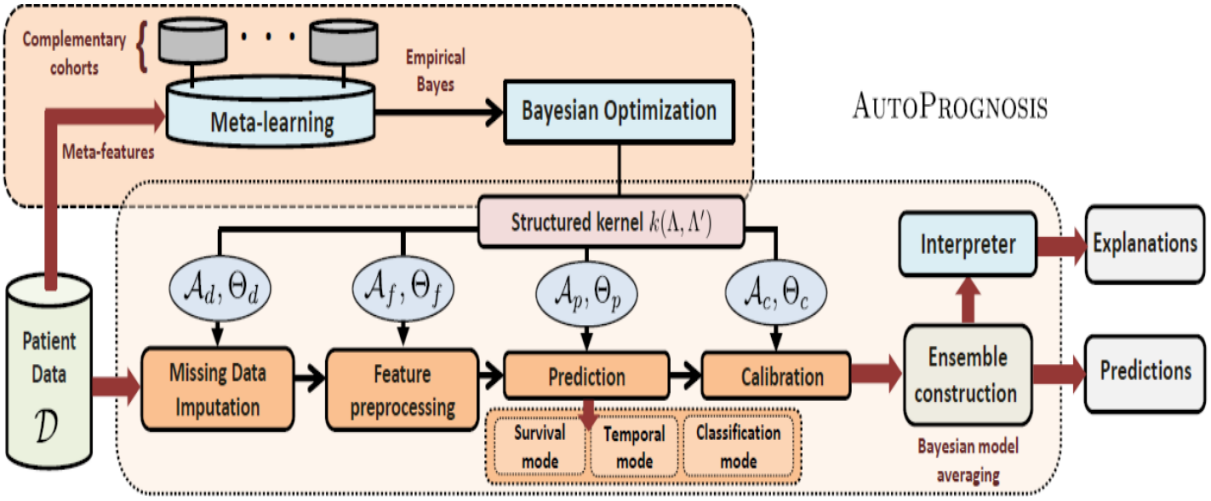


Figure 2.5: AutoPrognosis' components [Alaa and Schaar 2018]

AutoPrognosis accepts an input which is a dataset $\mathcal{D} = (x_i, y_i)_{i=1}^n$ for patients of a cohort of size n, where x_i is the feature values for patient i and y_i is the target value (patient's clinical endpoint). The algorithm which uses this input data to automatically configure the machine learning pipelines is the most important component of AutoPrognosis. Each machine learning pipeline configuration constitutes an algorithm for all phases of the prognostic modelling process, i.e. every pipeline has an algorithm for Missing Data Imputation, Feature Preprocessing, Prediction and Calibration. Figure 2.6 displays all the algorithms considered at each phase of the pipeline configuration. A pipeline can be described in the form of $P = \{missForest, Linear\,SVM, Random\,Forest, Isotonic\}$. There are a total of 4 800 possible pipeline configurations in AutoPrognosis. In order to obtain a complete description of the configuration for a single pipeline, the hyperparameters of the specific algorithms that constitute the pipeline need to be resolved.

Determining the optimal configuration of a pipeline is the fundamental objective of AutoPrognosis. Thus, for a specified dataset of patients $\mathcal{D}$, the optimal configuration of a pipeline $P^*_{\theta^*} \in \mathcal{P}_\Theta$ is determined by using J-fold cross-validation as follows:

$$P^*_{\theta^*} \in argmax_{P_\theta \in \mathcal{P}_\Theta} \frac{1}{J} \sum_{i=1}^{J} \mathcal{L}(P_\theta; \mathcal{D}^{(i)}_{train}, \mathcal{D}^{(i)}_{valid}) \tag{2.3}$$

$\mathcal{L}$ represents an accuracy metric such AUC-ROC, c-index, etc, $\mathcal{D}^{(i)}_{train}$ and $\mathcal{D}^{(i)}_{valid}$ are training and validation datasets obtained through performing splits on $\mathcal{D}$ in the i-th fold, $P = A_d \times A_f \times A_p \times A_c$ where $(\mathcal{A}_d, \mathcal{A}_f, \mathcal{A}_p, \mathcal{A}_c)$ is the sets of all the algorithms considered by AutoPrognosis at each respective phase of prognostic modelling, $\Theta = \Theta_d \times \Theta_f \times \Theta_p \times \Theta_c$ is the space of hyperparameter configurations.

Pipeline Stage	Algorithms				
□ Data Imputation	□ missForest (2)	□ Median (0)	□ Most-frequent (0)	□ Mean (0)	□ EM (1)
	□ Matrix completion (2)	□ MICE (1)	□ None (0)		
♣ Feature process.	♣ Feature agglo. (4)	♣ Kernel PCA (5)	♣ Polynomial (3)	♣ Fast ICA (4)	♣ PCA (2)
	♣ R. kitchen sinks (2)	♣ Nystroem (5)	♣ Linear SVM (3)	♣ Select Rates (3)	♣ None (0)
• Prediction	• Bernoulli NB (2)	• AdaBoost (4)	• Decision Tree (4)	• Grad. Boost. (6)	• LDA (4)
	• Gaussian NB (0)	• XGBoost (5)	• Extr. R. Trees (5)	• Light GBM (5)	• L. SVM (4)
	• Multinomial NB (2)	• R. Forest (5)	• Neural Net. (5)	• Log. Reg. (0)	• GP (3)
	• Ridge Class. (1)	• Bagging (4)	• k-NN (1)	• Surv. Forest (5)	• Cox Reg. (0)
★ Calibration	★ Sigmoid (0)	★ Isotonic (0)	★ None (0)		

Figure 2.6: Algorithms used in the machine learning pipeline. The number of parameters required by each algorithm is specified in the parenthesis [Alaa and Schaar 2018]

This is an optimisation problem which is referred to as the Pipeline Selection and Configuration Problem (PSCP). To solve this problem the algorithm responsible for constructing a pipeline configuration uses Bayesian Optimization (BO) [Snoek *et al.* 2012] to approximate the performance of various pipeline configurations as a black-box function with a Gaussian Process (GP) prior and through learning a structured kernel decomposition encoding the correlation among the performances of the various pipeline configurations.

The PSCP is treated specifically as a black-box optimisation problem because equation 2.3 is an expression that does not have an analytic form. Therefore,
$\frac{1}{J} \sum_{i=1}^{J} \mathcal{L}(P_\theta; \mathcal{D}^{(i)}_{train}, \mathcal{D}^{(i)}_{valid})$ becomes a black-box function of the form $f : \mathbf{\Lambda} \to \mathbb{R}$, where $\mathbf{\Lambda} = \Theta \times \mathcal{P}$.

BO is used to seek for $P^*_{\theta^*}$, which is the pipeline configuration that maximises f (the black-box function). A Gaussian Process (GP) prior on f is specified by the BO algo-

rithm in the form:

$$f \sim GP(\mu(\Lambda), k(\Lambda, \Lambda')) \tag{2.4}$$

$\mu(\Lambda)$ is the mean function specifying the expected performance for various pipelines, $k(\Lambda, \Lambda')$ is the covariance kernel specifying the correlation among the performances of the various pipeline configurations.

The function f is defined over $\mathbf{\Lambda}$, which is a D-dimensional space.
Therefore, $D = dim(\mathbf{\Lambda})$ and is defined as:

$$D = dim(P) + \Sigma_{v \in \{d,f,p,c\}} \Sigma_{a \in A_v} dim(\Theta_v^a) \tag{2.5}$$

In AutoPrognosis, $\mathbf{\Lambda}$ is the pipeline configuration space and has a high dimensionality $(D = 106)$. Equation (2.5) specifies that D can be calculated by adding the total number of pipeline stages and the total number of hyperparameters of the algorithms included in every pipeline stage (see Figure 2.6), i.e. $4 + 102 = 106$. This high-dimensionality would generally make the standard GP-based BO impractical for usage because the computational complexity when optimizing the acquisition function and the sample complexity of finding the nonparametric estimation become exponential in D [Kandasamy *et al.* 2015]. Recent studies have found that the standard GP-based BO is only suitable for usage when $D \leq 10$ [Wang *et al.* 2013], i.e. it is only feasible for hyperparameter optimisation of a single machine learning model at a time. However, AutoPrognosis uses the knowledge that for a given dataset, one machine learning algorithm's performance may not be indicative of another machine learning algorithm's performance therefore GP-based BO can still be used if the algorithms are modelled independently to create sub-problems. Based on the data, the BO used by AutoPrognosis learns such a decomposition which will reduce the high-dimensional problem into a collection of sub-problems of lower dimensions. The kernel mentioned earlier is what models the decomposition of the algorithms. The correlation between the expected performances of the various pipelines are encoded in the kernel $k(\Lambda, \Lambda')$. Therefore, the following sparse additive kernel decomposition can be used to represent the primary structure that relates the various hyperparameters:

$$k(\Lambda, \Lambda') = \Sigma_{m=1}^{M} k_m(\Lambda^{(m)}, \Lambda'^{(m)}) \tag{2.6}$$

$\Lambda^m \in \mathbf{\Lambda}^m, \forall m \in 1, ..., M$ where $\{\mathbf{\Lambda}^{(m)}\}_m$ is a set of disjoint subspaces of $\mathbf{\Lambda}$. This means that, $\Lambda^{(m)} \cap \Lambda^{(m')} = \emptyset$ and $\cup_m \Lambda^{(m)} = \mathbf{\Lambda}$. Furthermore, $\Sigma_m dim(\Lambda^m) = D$.

The sparse additive decomposition decomposes the function f as follows:

$$f(\Lambda) = \Sigma_{m=1}^{M} f_m(\Lambda^{(m)}) \tag{2.7}$$

As a result, the sample complexity of finding the nonparametric estimation is reduced from $O(n^{\frac{-\gamma}{2\gamma+D}})$ to $O(n^{\frac{-\gamma}{2\gamma+D_m}})$. D_m specifies the highest number of dimensions in any subspace. The hyperparameters for the algorithms with uncorrelated performances are found in different subspaces. An example of a subspace decomposition for the hyperparameters of algorithms in the prediction, feature processing and imputation stages of the pipeline is illustrated in Figure 2.7.

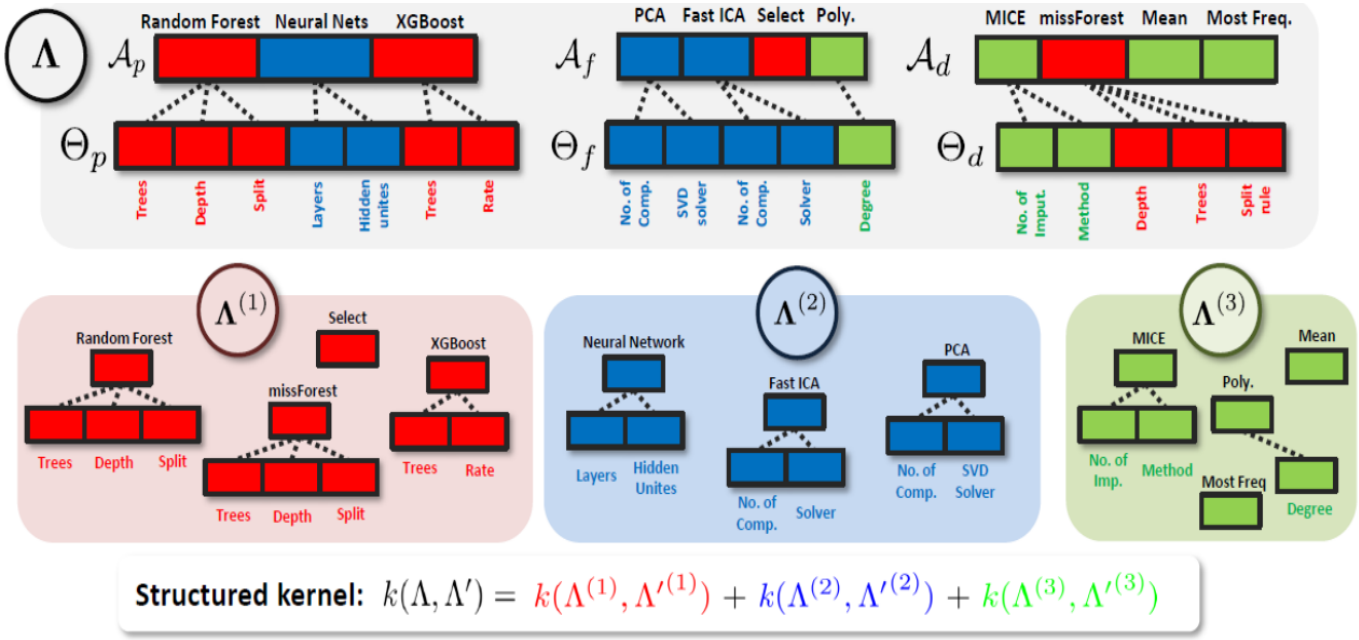


Figure 2.7: Example of subspace decomposition [Alaa and Schaar 2018]

It is important to note that the structured kernel is learnt from the given dataset (the input) and is not known prior. In order to learn the subspace decomposition, AutoPrognosis makes use of a Bayesian approach simultaneously with the BO technique. During this process, a Dirichlet-Multinomial prior is placed over the kernel

$$\alpha \sim Dirichlet(M, \gamma), z_{v,a} \sim Multi(\alpha), \tag{2.8}$$

$\forall a \in A_v, v \in \{d, f, p, c\}$, whereas $\gamma = \{\gamma_m\}_m$ is one of the parameters of the Dirichlet prior, $z_{v,a}$ determines to which subspace an algorithm a in A_v belongs. An update to the posterior distribution of the subspace decomposition $\{\Lambda^{(m}\}_m$ is made at each iteration of the BO procedure, and that is how the kernel decomposition is learned.

Furthermore, to solve the PSCP, AutoPrognosis uses a parallelised adaptation of the BO technique where it selects B configured pipelines for evaluation at every iteration t; where each pipeline has a specific algorithm from a distinct subspace. The whole process takes place in the form of a batched exploration scheme consisting of the following two steps:

Step 1: Identify and select the frequentist kernel decomposition $\{\hat{\mathbf{\Lambda}}^{(m)}\}_m$. This particular kernel decomposition should maximise the posterior $\mathbb{P}(z|H_t)$, whereas $z = \{z_{v,a} : \forall a \in A_v, \forall v \in \{d, f, p, c\}\}$and H_t specifies the history of all the different evaluations, from iteration 0 to t, of the black-box function.

Step 2: Identify and select the B configured pipelines $\{P^b_\Theta\}^B_{b=1}$ which have the largest acquisition function $\{A(P^b_\Theta; H_t)\}^B_{b=1}$ values. Every pipeline P^b_Θ , $b \in \{1, ..., B\}$should have a different prediction algorithm from a different subspace in $\{\hat{\mathbf{\Lambda}}^{(m)}\}$.

The Ensemble Construction module (see Figure 2.5) uses Bayesian model averaging to construct an ensemble of weighted pipelines, as opposed to using the frequentist approach which simply selects the best performing pipeline from $\{P^1_{\theta^1}, ..., P^t_{\theta^t}\}$ and throws away the rest of the evaluated pipelines. When a weighted ensemble is used the uncertainty in the performances of the pipelines can also be captured. Furthermore, it is important to note the Meta-learning module. This module integrates preceding information procured from previous executions of AutoPrognosis on a repo of complementary cohorts $\{\mathcal{D}_1, ..., \mathcal{D}_K\}$ to warmstart the BO procedure.

Finally, we discuss the interpreter module. This module takes the learned model for a given dataset and a set of risk strata R as its input, and outputs a set of logical association rules which attempts to explain predictions made by the learned model. These association rules are generated through a Bayesian associative classifier [Kruschke 2008].

$$C_1 \wedge C_2 \wedge ... \wedge C_{l(r)} \longrightarrow r, \ \forall r \in \mathbb{R} \tag{2.9}$$

$\{C_1, ..., C_{l(r)}\}$ is a set of Boolean conditions associated with r

More specifically, an example of an "explanation" produced by the interpreter module will look like this: $Diabetic \wedge LipidLowering \wedge (Age \geq 40) \Longrightarrow HighCVDrisk$.

It is worth mentioning that the AutoPrognosis pipelines have three different modes of operation: (i) survival mode, (ii) temporal mode, and (iii) classification mode. This means that AutoPrognosis can handle diverse types of clinical data and their outcomes. The survival mode deals with time-to-event data. This data provides information about the time taken until a specific event occurs. The survival mode is comprised of the classification algorithm mentioned above, along with survival models (e.g. the Cox proportional hazards model) and models for multiple competing risks. The temporal

mode deals with time series and longitudinal data. The classification algorithms are applied to data found in the sliding window of the time series. The classification mode deals with datasets where the outcomes have binary values. All the classification algorithms in the scikit-learn library and other useful models, such as XGBoost are included in this mode.

2.8 Conclusion

This chapter provides information on the medical background and methods applicable to the problem area relevant to the research study presented in this paper. The chapter presents a medical definition and understanding of CVD, CHD and MetS. Some notable clinical models currently used for CVD prognosis are briefly discussed. Detailed descriptions of the Random Forest and AutoPrognosis algorithms are provided.

Chapter 3

Related Work

3.1 Introduction

In the previous chapter, information on the medical background and methods applicable to our research study was presented. Chapter 3 discusses some of the major contributions to the field, relevant to the themes related to the problem area of our research study. Essentially, sections 3.2 and 3.3 present an outline of some of the previous applications of Random Forests and AutoPrognosis in terms of how it has informed the direction of this research thesis. Section 3.4 concludes chapter 3 by providing a summary of the main points covered.

3.2 Applications of Random Forests

Existing CVD risk prediction models discussed in Section 2.5, make use of regression methods. Generally, regression models use a limited number of conventional risk factors and assume that the correlation between all such factors and the CVD outcomes is linear. These two characteristics limit the model's predictive performance for CVD, especially for certain subgroups of the population such as individuals living with HIV/AIDS, diabetes, etc (this is because for people whose health is already compromised, risk factors tend to have more complex interactions with their pre-existing conditions, therefore such individuals may be at an increased of developing CVD and often require their own models). Models based on ensemble machine learning algorithms, such the Random Forest algorithm, often have the ability to realize more complex patterns within data of large repositories and can adequately model nonlinear relations for interactions between the predictor variables and outcome. Multiple comparative studies have been conducted to evaluate the differences in performances for CVD risk prediction between Random Forests and traditionally used CVD risk prediction models, as well as other standard ML models.

One such study was performed by Yang *et al.* [2020] on a cohort of 29 930 CVD high-risk individuals from the Chinese population. From 2014-2016, consistent follow-ups were conducted to update information related to the occurrence of a CVD event for each individual over the specified time period. The data was used to develop a 3-

year CVD risk prediction model based on Random Forests. The performance of the model was compared to that of models based on Multivariate Regression, Naïve Bayes, AdaBoost, Classification and Regression Tree (CART) and Bagged Trees. For the performance evaluation, the Multivariate Regression model (AUC-ROC = 0.7143) was used as the benchmark model. As displayed in Figure 3.1, results show that the Random Forest model was superior to the other models with an AUC-ROC value of 0.787. Essentially, [Yang *et al.* 2020] successfully demonstrated that the RF model outperforms the FRS model due to its ability to model more complex nonlinear relationships between the predictor variables and the CVD risk outcome. Modelling complex nonlinear relationships and including more variables beyond the conventional risk factors to improve CHD risk prediction is one of the central themes of the research study presented by this paper, therefore contributions from studies such as that presented by [Yang *et al.* 2020] influenced the decision of the model choice for our research study.

Model	AUC	AUC Change
Multivariate Regression	0.7143	Benchmark
CART	0.7025	−1.18%
Naïve Bayes	0.7074	−0.69%
Bagged Trees	0.7448	3.05%
Ada Boost	0.7862	7.19%
Random Forest	0.7872	7.29%
Framingham Score	0.7596	4.53%

Figure 3.1: Performance of prediction models from [Yang *et al.* 2020]

Furthermore, multiple other studies have been conducted with the aim of applying different machine learning methods to discover the optimal model for MetS prediction. One such study was performed by Kim *et al.* [2022] on a middle-aged population in Korea. A dataset of 20 variables for 1991 participants, aged 30-55 years old, was collected from a cohort study known as the Korean Medicine Daejeon Citizen Cohort and used for this study. The variables included age, gender, anthropometric measurements, blood indicators and data related to lifestyle choices/habits. Participants who satisfied two of the five NECP-ATP III (refer to table 2.1) were considered as pre-MetS individuals, while those with three or more of the five NECP-ATP III criteria were considered to have MetS. The study aimed to evaluate and compare the performances of nine different ML models based on Random Forest, Support Vector Machine, Gaussian Naïve Bayes, Decision Tree, Logistic Regression, K-nearest neighbour, Multi-layer Perceptron, eXtreme Gradient Boosting (XGBoost) and 1D Convolutional Neural Network, respectively.

Before training the models, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the data to resolve the issue of data imbalance. Three different models were trained for each machine learning method used, i.e. for the first step the model was trained only using the anthropometric and demographic features, the second step involved adding the lifestyle-related features to the feature set, and the third step involved adding the biochemical measurements. The experiment results in Figure 3.2 showed that the XGBoost and random forest models produced the best performances, with AUC-ROC values of 0.851 and 0.844, respectively. When the SMOTE technique was applied to the data, the ROC-AUC value of the models' performance increased by up to 0.091. MetS was identified in 33.85% of the participants. The waist-to-hip ratio and BMI features were established to be the most significant features in the prediction models for MetS in the cohort. The paper concluded that tree-based machine learning methods are useful for detecting MetS in the Korean population. Early detection of MetS provides great benefits for disease control and prevention and requires an approach which is multidimensional. Yet again, contributions from studies such as that presented by [Kim *et al.* 2022] influenced the decision of the model choice for our research study.

	F1-score		Accuracy		Sensitivity		Specificity		AUC	
	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE	Original	SMOTE
4 Features (Demographic and anthropometric Features)										
Decision Tree	0.711 (0.66–0.76)	0.758 (0.71–0.80)	0.711 (0.66–0.76)	0.758 (0.71–0.80)	0.573 (0.52–0.63)	0.758 (0.71–0.80)	0.782 (0.74–0.83)	0.758 (0.71–0.80)	0.677 (0.63–0.73)	0.758 (0.71–0.80)
Gaussian NB	0.789 (0.75–0.83)	0.780 (0.74–0.82)	0.790 (0.75–0.83)	0.780 (0.74–0.82)	0.684 (0.63–0.73)	0.790 (0.75–0.83)	0.844 (0.80–0.88)	0.769 (0.72–0.81)	0.764 (0.72–0.81)	0.780 (0.74–0.82)
KNN	0.774 (0.73–0.82)	0.783 (0.74–0.83)	0.777 (0.73–0.82)	0.783 (0.74–0.83)	0.619 (0.57–0.67)	0.826 (0.79–0.87)	0.859 (0.82–0.90)	0.740 (0.69–0.79)	0.739 (0.69–0.79)	0.783 (0.74–0.83)
XGBoost	0.771 (0.73–0.82)	0.802 (0.76–0.84)	0.773 (0.73–0.82)	0.802 (0.76–0.85)	0.626 (0.57–0.68)	0.812 (0.77–0.85)	0.848 (0.81–0.89)	0.792 (0.75–0.84)	0.737 (0.69–0.78)	0.802 (0.76–0.85)
RF	0.772 (0.73–0.82)	0.813 (0.77–0.86)	0.774 (0.73–0.82)	0.814 (0.77–0.86)	0.628 (0.58–0.68)	0.832 (0.79–0.87)	0.850 (0.81–0.89)	0.795 (0.75–0.84)	0.739 (0.69–0.79)	0.814 (0.77–0.86)
Logistic R	0.777 (0.73–0.82)	0.783 (0.74–0.83)	0.787 (0.74–0.83)	0.784 (0.74–0.83)	0.558 (0.50–0.61)	0.799 (0.76–0.84)	0.904 (0.87–0.94)	0.768 (0.72–0.81)	0.731 (0.68–0.78)	0.784 (0.74–0.83)
SVM	0.787 (0.74–0.83)	0.785 (0.74–0.83)	0.795 (0.75–0.84)	0.785 (0.74–0.83)	0.585 (0.53–0.64)	0.809 (0.77–0.85)	0.903 (0.87–0.93)	0.762 (0.72–0.81)	0.744 (0.70–0.79)	0.786 (0.74–0.83)
MLP	0.785 (0.74–0.83)	0.770 (0.72–0.82)	0.792 (0.75–0.84)	0.772 (0.73–0.82)	0.607 (0.55–0.66)	0.735 (0.69–0.78)	0.887 (0.85–0.92)	0.809 (0.77–0.85)	0.747 (0.70–0.79)	0.772 (0.73–0.82)
1D-CNN	0.779 (0.73–0.82)	0.783 (0.74–0.83)	0.782 (0.74–0.83)	0.784 (0.74–0.83)	0.657 (0.61–0.71)	0.784 (0.74–0.83)	0.846 (0.81–0.88)	0.784 (0.74–0.83)	0.752 (0.71–0.80)	0.784 (0.74–0.83)
12 Features (Lifestyle-related features added)										
Decision Tree	0.722 (0.67–0.77)	0.765 (0.72–0.81)	0.724 (0.68–0.77)	0.765 (0.72–0.81)	0.570 (0.52–0.62)	0.776 (0.73–0.82)	0.803 (0.76–0.85)	0.755 (0.71–0.80)	0.686 (0.64–0.74)	0.765 (0.72–0.81)
Gaussian NB	0.775 (0.73–0.82)	0.766 (0.72–0.81)	0.774 (0.73–0.82)	0.766 (0.72–0.81)	0.685 (0.64–0.74)	0.773 (0.73–0.82)	0.820 (0.78–0.86)	0.759 (0.71–0.81)	0.753 (0.71–0.80)	0.766 (0.72–0.81)
KNN	0.738 (0.69–0.78)	0.780 (0.73–0.82)	0.743 (0.70–0.79)	0.782 (0.74–0.83)	0.551 (0.50–0.60)	0.879 (0.84–0.91)	0.842 (0.80–0.88)	0.685 (0.63–0.73)	0.696 (0.65–0.75)	0.782 (0.74–0.83)
XGBoost	0.778 (0.73–0.82)	0.834 (0.79–0.87)	0.782 (0.74–0.83)	0.834 (0.79–0.87)	0.622 (0.57–0.67)	0.837 (0.8–0.88)	0.863 (0.83–0.90)	0.832 (0.79–0.87)	0.743 (0.70–0.79)	0.834 (0.79–0.87)
RF	0.791 (0.75–0.83)	0.838 (0.80–0.88)	0.795 (0.75–0.84)	0.838 (0.80–0.88)	0.635 (0.58–0.69)	0.850 (0.81–0.89)	0.876 (0.84–0.91)	0.826 (0.79–0.87)	0.756 (0.71–0.80)	0.838 (0.80–0.88)
Logistic R	0.785 (0.74–0.83)	0.779 (0.73–0.82)	0.792 (0.75–0.84)	0.779 (0.73–0.82)	0.595 (0.54–0.65)	0.791 (0.75–0.83)	0.893 (0.86–0.93)	0.767 (0.72–0.81)	0.744 (0.70–0.79)	0.779 (0.73–0.82)
SVM	0.790 (0.75–0.83)	0.783 (0.74–0.83)	0.797 (0.75–0.84)	0.783 (0.74–0.83)	0.605 (0.55–0.66)	0.796 (0.75–0.84)	0.894 (0.86–0.93)	0.770 (0.72–0.82)	0.750 (0.70–0.80)	0.783 (0.74–0.83)
MLP	0.772 (0.73–0.82)	0.797 (0.75–0.84)	0.778 (0.73–0.82)	0.798 (0.75–0.84)	0.619 (0.57–0.67)	0.790 (0.75–0.83)	0.859 (0.82–0.90)	0.806 (0.76–0.85)	0.739 (0.69–0.79)	0.798 (0.75–0.84)
1D-CNN	0.771 (0.73–0.82)	0.770 (0.72–0.82)	0.776 (0.73–0.82)	0.774 (0.73–0.82)	0.635 (0.58–0.69)	0.861 (0.82–0.90)	0.848 (0.81–0.89)	0.688 (0.64–0.74)	0.742 (0.69–0.79)	0.775 (0.73–0.82)
20 Features (Biochemical measurements added)										
Decision Tree	0.743 (0.70–0.79)	0.777 (0.73–0.82)	0.743 (0.70–0.79)	0.778 (0.73–0.82)	0.631 (0.58–0.68)	0.797 (0.75–0.84)	0.801 (0.76–0.84)	0.758 (0.71–0.80)	0.716 (0.67–0.76)	0.778 (0.73–0.82)
Gaussian NB	0.786 (0.74–0.83)	0.759 (0.71–0.81)	0.795 (0.75–0.84)	0.762 (0.72–0.81)	0.577 (0.52–0.63)	0.646 (0.59–0.70)	0.906 (0.87–0.94)	0.878 (0.84–0.91)	0.741 (0.69–0.79)	0.762 (0.72–0.81)
KNN	0.748 (0.70–0.79)	0.787 (0.74–0.83)	0.756 (0.71–0.80)	0.788 (0.74–0.83)	0.540 (0.49–0.59)	0.871 (0.83–0.91)	0.866 (0.83–0.90)	0.705 (0.66–0.75)	0.703 (0.65–0.75)	0.788 (0.74–0.83)
XGBoost	0.801 (0.76–0.84)	0.851 (0.81–0.89)	0.804 (0.76–0.85)	0.851 (0.81–0.89)	0.662 (0.61–0.71)	0.859 (0.82–0.9)	0.877 (0.84–0.91)	0.843 (0.8–0.88)	0.769 (0.72–0.81)	0.851 (0.81–0.89)
RF	0.815 (0.77–0.86)	0.843 (0.80–0.88)	0.818 (0.78–0.86)	0.844 (0.80–0.88)	0.690 (0.64–0.74)	0.857 (0.82–0.89)	0.883 (0.85–0.92)	0.831 (0.79–0.87)	0.786 (0.74–0.83)	0.844 (0.80–0.88)
Logistic R	0.812 (0.77–0.85)	0.804 (0.76–0.85)	0.818 (0.78–0.86)	0.804 (0.76–0.85)	0.638 (0.59–0.69)	0.812 (0.77–0.85)	0.910 (0.88–0.94)	0.796 (0.75–0.84)	0.774 (0.73–0.82)	0.804 (0.76–0.85)
SVM	0.811 (0.77–0.85)	0.810 (0.77–0.85)	0.817 (0.78–0.86)	0.810 (0.77–0.85)	0.636 (0.58–0.69)	0.831 (0.79–0.87)	0.909 (0.88–0.94)	0.790 (0.75–0.83)	0.773 (0.73–0.82)	0.810 (0.77–0.85)
MLP	0.807 (0.76–0.85)	0.811 (0.77–0.85)	0.812 (0.77–0.85)	0.812 (0.77–0.85)	0.638 (0.59–0.69)	0.836 (0.80–0.88)	0.901 (0.87–0.93)	0.787 (0.74–0.83)	0.770 (0.72–0.81)	0.812 (0.77–0.85)
1D-CNN	0.799 (0.76–0.84)	0.814 (0.77–0.86)	0.803 (0.76–0.85)	0.815 (0.77–0.86)	0.662 (0.61–0.71)	0.807 (0.76–0.85)	0.875 (0.84–0.91)	0.822 (0.78–0.86)	0.768 (0.72–0.81)	0.815 (0.77–0.86)

Presented are the results before (Original) and after (SMOTE) applying the synthetic minority oversampling technique

AUC Area under the receiver operating characteristic curve, *Gaussian NB* Gaussian naïve bayes classifier, *KNN* K-nearest neighbor, *XGBoost* Extreme gradient boosting, *Logistic R* Logistic regression, *RF* Random forest, *SVM* Support vector machine, *MLP* Multilayer perceptron, *1D-CNN* 1-dimensional convolutional neural network

Figure 3.2: Performance of prediction models from [Kim *et al.* 2022]

3.3 Applications of AutoPrognosis

[Alaa *et al.* 2019] used a dataset from the UK Biobank for a cohort of 423 604 participants, with 473 available variables, to test whether AutoPrognosis is a more suitable option for CVD risk prediction in contrast to traditional approaches. A machine learning model developed using AutoPrognosis and all 473 available variables on the entire dataset was compared to the Framingham Risk Score model (an approved risk prediction model based on conventional CVD risk factors), a Cox Proportional Hazards model based on several well-established CVD risk factors (i.e. BMI, smoking status, recipience of hypertension treatment, systolic blood pressure, age, gender and patient's history of diabetes), a Cox Proportional Hazards model based on all 473 variables, and some standard machine learning models. The results of these comparisons are shown in Figure [2.4], where it is depicted that, for the population from the UK Biobank, AutoPrognosis improves CVD risk prediction results.

Model	AUC-ROC	Absolute AUC-ROC Change
Framingham Score	0.724 ± 0.004	Baseline model
Cox PH Model (7 core variables)	0.734 ± 0.005	+ 1.0%
Cox PH Model (all variables)	0.758 ± 0.005	+ 3.4%
Support Vector Machines	0.709 ± 0.061	- 1.5%
Random Forest	0.730 ± 0.004	+ 0.6%
Neural Networks	0.755 ± 0.005	+ 3.1%
AdaBoost	0.759 ± 0.004	+ 3.5%
Gradient Boosting	0.769 ± 0.005	+ 4.5%
AutoPrognosis (7 core variables)	0.744 ± 0.005	+ 2.0%
AutoPrognosis (369 non-lab. variables)	0.761 ± 0.005	+ 3.7%
AutoPrognosis (104 lab. variables)	0.735 ± 0.008	+ 1.1%
AutoPrognosis (all variables)	0.774 ± 0.005	+ 5.0%

The Framingham score is provided as the reference model for comparative purposes.

https://doi.org/10.1371/journal.pone.0213653.t002

Figure 3.3: Performances of prediction models from [Alaa *et al.* 2019]

Since the dataset from the UK Biobank contains a large number of significant variables (473 variables), an information gain is guaranteed because more risk factors beyond the conventional ones can get incorporated into a model. Any data-driven model, such as a basic Cox PH model, can benefit from this information gain and have improved performance results, in contrast to the Framingham Risk Score Model which only uses a set number of conventional risk factors. Based on the table shown in Figure 3.3, when all 473 variables are used, AutoPrognosis had better a performance than the standard Cox PH model, despite the information gain that was also attained by the standard Cox

PH model. This occurred because AutoPrognosis is a fairly complex model and therefore not only does it benefit from the information gain but also from a modelling gain.

AutoPrognosis achieves a modelling gain because of its ability to select the best machine learning model (and tune its hyperparameters) among several models that have different degrees of complexity and robustness for datasets that have a high dimensionality. Furthermore, results which can be found in the paper written by [Alaa *et al.* 2019] which showed that AutoPrognosis performs significantly better for CVD risk prediction in the sub-population of individuals who have a history of diabetes. As a result of both the information and modelling gain, AutoPrognosis was able to learn risk factors specific to diabetic patients which do not get captured by existing prediction models. [Alaa *et al.* 2019] states that this was the first comprehensive study of the performance of machine learning models for CVD risk prediction in a substantially large group of participants with such an immense number of variables.

Although [Alaa and Schaar 2018] and [Alaa *et al.* 2019] only produced studies which were conducted with the aim of improving CVD risk prediction, just like other machine learning methods AutoPrognosis can be used to solve other classification problems. [Shah *et al.* 2021] demonstrates how AutoPrognosis can be used to model and predict the risk of a patient developing major complications after undergoing a procedure known as Total Hip Arthroplasty (THP) to treat hip osteoarthritis. The paper presents results from a retrospective study of a cohort of 89 986 people who had THP performed between 2015 and 2017 and were at risk of developing major complications such as a pulmonary/cardiac complication, venous thromboembolism, or an infection. The discrimination and calibration ability of their AutoPrognosis model was compared to that of logistic regression and other standard machine learning models. AutoPrognosis demonstrated a greater risk prediction performance in contrast to the other models for predicting the risk of THP post-op complications.

Lastly, [Alaa and van der Schaar 2018] performed a study demonstrating the use of AutoPrognosis as a prognostic model for Cystic Fibrosis. This was a retrospective study on a cohort that contained 99% of the UK's Cystic Fibrosis population. The study aimed to establish the optimal time needed to determine when a patient suffering from terminal respiratory failure can be referred for a lung transplant. Current clinical practices recommend referring patients when the value of their Forced Expiratory Volume (FEV) decreases to below 30% of the predicted nominal value. FEV has been established to be a strong predictor for this particular classification problem. Experimental results showed that the performance of AutoPrognosis surpassed that of other competing models, such as Bagging, Gradient Boosting, Support Vector Machines, etc. and existing clinical guidelines.

Essentially, [Shah *et al.* 2021] and [Alaa and van der Schaar 2018] demonstrated how the applications of AutoProgonosis can be useful for classification problems other than CVD risk prediction. In the research study presented by this paper, one of the other central themes was related to demonstrating how the modelling gain achieved by AutoPrognosis can also be useful for improving MetS prediction results. Furthermore, the

performance of the algorithm that a model is based on can also be affected by the size, quality and dimensionality of the dataset used to train the model. In papers such as those presented by Alaa and Schaar[2018] and Alaa et al. [2019] the performance of AutoPrognosis was evaluated using datasets with at least 30 000 and 423 604 patient records, respectively. Evaluating the performance of AutoPrognosis on significantly smaller datasets becomes essential as some ML models, such as Neural Networks, are data hungry and require large quantities of data during training in order to produce satisfactory performance results. Once again, to the best of our knowledge, our study is unique in that it is the first to conduct a comparative analysis for a performance evaluation between models based on RF against AutoPrognosis on different sample sizes of data.

3.4 Conclusion

This chapter provides information about the related work applicable to the problem area relevant to the research study presented by this paper. Contributions from studies such as those presented by [Yang *et al.* 2020] and [Kim *et al.* 2022] influenced the decision to make use of the RF algorithm for our models, as modelling complex nonlinear relationships to improve CHD risk prediction and MetS detection is one of the central themes of our research study. Results from [Shah *et al.* 2021] and [Alaa and van der Schaar 2018] showed how applications of AutoProgonosis can be useful for classification problems other than CVD risk prediction. Therefore, in our research study, we aimed to demonstrate how AutoPrognosis can also be useful specifically for improving MetS prediction results. Furthermore, our study is unique in that it is the first to conduct a comparative performance analysis for RF against AutoPrognosis on significantly small sample sizes of data.

Chapter 4

Research Methodology

4.1 Introduction

In chapter 3, all the necessary information about the related work applicable to this research study was presented. Previously, the performance of AutoPrognosis was evaluated using significantly large datasets. In our research study, we evaluate the performance of RF against AutoPrognosis on significantly smaller sample sizes of data. Furthermore, we determined the most significant risk factors related to MetS and as a novel study, we assessed the predictive performance of RF vs AutoPrognosis specifically for determining the presence of MetS. Within this chapter, the objectives and research hypotheses are formally stated in section 4.2 and section 4.3, respectively. Thereafter, section 4.4 describes the data and methodology followed which led to an acceptance or rejection of the proposed research hypotheses. The limitations of this research are presented in section 4.5 . Finally, section 4.6 provides a short synopsis of the main points covered in this chapter.

4.2 Research Objectives

The main objectives of this research were:

- Using the RF algorithm, we developed a 10-year CHD risk prediction model and a MetS prediction model.
- Using AutoPrognosis, we developed a 10-year CHD risk prediction model and a MetS prediction model.
- We evaluated and compared the predictive performances between FRS, RF without optimised hyperparameters, RF with optimised hyperparameters, AutoPrognosis with a single pipeline and AutoPrognosis with an ensemble pipeline, for 10-year CHD risk prediction and MetS prediction, respectively
- We used an RF model to perform variable ranking, in order to determine the most significant risk factors related to CHD and MetS, respectively.

- We evaluated and compared the predictive performances of RF against AutoPrognosis on different samples of the dataset. The size of the datasets ranged from 100 to 4900.

4.3 Research Hypotheses

The research hypotheses were the following:

1. Both types of CHD risk prediction models, based on RF and AutoPrognosis, will outperform the FRS model. (The FRS model was used as a benchmark model because it is one of the oldest and most popularly used CVD risk prediction models.)

2. The AutoPrognosis models will always outperform the RF models, for both 10-year CHD risk prediction and MetS detection, respectively.

3. The AutoPrognosis models with an ensemble pipeline will always outperform the AutoPrognosis models with a single pipeline, for both 10-year CHD risk prediction and MetS detection, respectively.

4. AutoPrognosis will have satisfactory performance results on all the different sample sizes of data.

5. AutoPrognosis will always outperform RF on all the different sample sizes of data.

4.4 Methodology

In this section, a description of the data and methodology followed to achieve the objectives of this research is presented in the following sections.

4.4.1 Modelling CHD risk prediction

Sample Design and Population

FHS [NHLBI 2016] is an epidemiologic study based on CVD. The study began in 1948 in the town of Framingham in Massachusetts, USA. When the study began, 5209 males and females aged between 30 and 60 were recruited by FHS researchers. They were used for the first course of wide-ranging physical examinations and interviews that were carried out. Thereafter, the subjects continued to return every two to six years to participate in further physical exams and laboratory tests, while also providing detailed information on their medical history, as a means of updating information for the study. As the years progressed, the FHS transformed into a multigenerational study by collecting data from the two generations (children and grandchildren) of the original cohort. Initially, FHS was based on a largely white American population. However, since then the study has extended to become more inclusive of more ethnically diverse populations by enrolling participants of Hispanic, Indian, African American, Asian, Native American and Pacific Islander descent.

Data Collection and Preparation

The FHS is still ongoing to date. The dataset used in the study presented by this paper is from the FHS and is publicly available on the Kaggle website [Car 2020]. The dataset has 4240 records and 16 variables for each record. After a series of data cleaning procedures, the dataset included 3658 records and 15 variables for each record. Data cleaning included the standard procedures of removing duplicated records, removing/imputing missing values, removing invalid points and outliers, deleting obvious error messages and data standardisation. Each variable is a possible CHD risk factor. The variables can be categorised into three categories of risk factors: demographic, medical and behavioural risk factors. The target variable was the outcome for the 10-year risk of CHD. The list of variables and definitions are provided in Appendix A. Prior to implementing and training the models used in the study outlined in this paper, an oversampling technique known as SMOTE was applied to the datatset in order to resolve the issue of data imbalance and reduce the possibility of having highly biased models.

Models Developed and Tested

Framingham Risk Score – As mentioned in chapter 2, multiple adaptions and revisions of the original FRS model have been made over the years for assessing the risk of an individual developing a specific type of CVD. However, there also exists an FRS model that can be used to predict the general risk of an individual developing any type of CVD or CVD event. This model has been proven to perform just as well as the separate disease-specific variations of FRS, to predict an individual component of CVD, such as CHD. Furthermore, it is the model which was chosen to be used in the study outlined by this paper.

The standard version of general CVD risk prediction model includes HDL cholesterol as one of the variables in its algorithm. However, this is a variable which was not provided in the dataset used for this study. Fortunately, FRS provides a variation of the standard model based on non-laboratory variables, which substitutes the lipid variables such as HDL cholesterol with Body Mass Index (BMI). Therefore, in this study, we used the predicting equations of the BMI-based FRS model for general CVD risk prediction which was published by D'agostino *et al.* [2008]. These equations were composed of beta-coefficients and survival functions. The model has seven core variables: sex, age, diabetes status, smoking status, treatment for hypertension, systolic blood pressure and BMI. All these variables were present in our dataset.

Random Forest - For this study, the RF model was implemented using the Scikit-learn library in the Python programming language. Our dataset was split into training and test datasets which are randomly selected to form a 70-30% split, respectively. The model was trained by fitting the Random Forest Classifier with the training data and tuning the parameters (maximum tree depth and maximum features) for optimization via grid search. The hyperparameter (the number of decision trees in the forest) was determined by assessing the out-of-bag error rate. The test data set was used to procure the AUC-ROC value to indicate the performance of the model on unseen data. A total

of 10 trials were conducted to obtain the mean AUC-ROC value of the predictive model.

AutoPrognosis – AutoPrognosis uses an advanced Bayesian Optimisation technique to automatically generate a prognostic model made up of a weighted ensemble of machine learning pipelines. Each pipeline is made up of a data imputation, feature processing, classification, and calibration algorithm of its own. For training the model, AutoPrognosis is set to conduct several iterations of the Bayesian Optimization procedure. In each iteration, a new machine learning is explored, and its hyper-parameters are tuned. In this study, the number of iterations was set to 10. In every iteration, 5-fold cross-validation was used to assess the performance of the pipeline being evaluated. In this study, two different versions of the AutoPrognosis model were created: one based on the seven core variables used in the FRS model, and the other based on all the variables provided in our dataset. (The code for the state-of-the-art AutoPrognosis model used in this study has been made available by Alaa and Schaar [2019])

Performance Evaluation

The FRS model was used as the baseline model for a comparative performance evaluation of the proposed models used in this study. The mean AUC-ROC, confusion matrix, Precision, Recall, and F1 score values for the FRS, RF and AutoPrognosis models were calculated. These values were attained after testing was done on the test datasets. Once all the models were developed and tested, a comparative analysis of their performances based on their performance metric values was done to see model had better predictive abilities for CHD risk prediction. (In some instances, where model improvements were necessary, interpretation of the results led to the need to return to the previous phases of training and testing to adjust the models and their parameters.)

Variable Ranking

Each variable used to create a classification model contributes differently to the predictions made by the model. To determine the relative significance of each variable to the CHD risk prediction outcome, a random forest model was fitted to the data with the patients' variables as inputs and the predictions made by the model (the model which performed better between the AutoPrognosis or Random Forest model) as the outputs. Based on that, variable importance scores based on the Mean Decrease in Impurity were allocated for each variable based on feature permutation.

4.4.2 Modelling Metabolic Syndrome outcome prediction

Sample Design and Population

The NHANES [NCHS 2017] is an initiative of studies created to assess the health and nutritional status of both people in their youth and adulthood in the USA. NHANES began in 1959 and is still ongoing. Data is collected through a survey which combines physical examinations and interviews. The survey is carried out on a sample of approximately 5000 people for every iteration of data collection once a year. The people

of the sample reside in different counties of the USA and are chosen to be nationally representative of all age and ethnicity groups in the USA population. Interviews are conducted in participants' homes and include socioeconomic, demographic, dietary and other health-related questions about lifestyle choices. The physical examinations are conducted in mobile health centres and include dental, medical, and physiological measurements and laboratory tests. Workers of the study consist of multiple physicians, health and medical technicians, as well as health and dietary interviewers.

Data Collection and Preparation

The dataset used in the latter part of the study presented by this paper is from the NHANES and is publicly available on the Data World [Hoyt 2020b] website. The dataset has 7821 records and 77 variables for each record. After a series of data cleaning procedures, the dataset included 6781 records and 24 variables for each record. Data cleaning included the standard procedures of removing duplicated records, removing/imputing missing values, removing invalid points and outliers, deleting obvious error messages and data standardisation. A person is diagnosed with MetS if they have three or more of the following conditions: hypertension, diabetes, dyslipidemia, and overweight/obesity. Although, our dataset did not come with variables for the dyslipidemia and overweight/obesity conditions we were still able to produce them using the HDL and triglyceride cholesterol and waist circumference variables, respectively. The variables can be categorised into three categories of risk factors: demographic, medical and behavioural risk factors. The target variable, indicative of the presence/absence of MetS, was produced by checking if each patient was diagnosable with MetS based on the criteria stated above and assigned them with a binary value. The list of variables and definitions are provided in Appendix B. Prior to implementing and training the models used in the study outlined in this paper, an oversampling technique known as SMOTE was applied to the datatset in order to resolve the issue of data imbalance and reduce the possibility of having highly biased models.

Models Developed and Tested, Performance Evaluation, Variable Ranking

Two types of models were built for predicting the presence of MetS. The first model was based on Random Forests, while the other was based on AutoPrognosis. Information for the descriptions of the models and the procedures followed during the training and testing phases can be found under the Models Tested sub-section in the Modelling CHD risk prediction section further above, as the exact same information applies to this half of the study presented by this paper. Likewise, information on how performance evaluation and variable ranking was performed can be found in the Performance Evaluation and Variable Ranking sub-sections above.

4.4.3 Models comparing the performance of Random Forests against AutoPrognosis on different sample sizes of data

Alaa and Schaar [2018] evaluated the performance of AutoPrognosis using datasets with at least 30 000 patient records. Sometimes attaining data that large and training

models on it can be challenging and computationally expensive. Therefore, for the final phase of the study, using the MetS dataset, we trained, tested and compared the performance of Random Forests against AutoPrognosis on 13 different sample sizes of data ranging from 100 to 4900 (at intervals of 400). Models were trained and tested as described in sections further above.

4.5 Limitations

Initially, the main objectives of our research study were to determine the most significant CVD risk factors and develop a 10-year CVD risk prediction model specific to the **South African** population. However, due to time constraints, obtaining a local dataset could not be achieved, therefore limiting the scope of our research study. This led to the alteration of our original research question, as we had to rely on using publicly available datasets which were not collected from the desired geographic location. Another major limitation to using this kind of secondary data is that we did not know exactly how the data collection process was done. We had no information about any potential biases which may have existed, and the extent to which the data may have been affected by issues such as low response rates or survey respondents misinterpreting specific survey questions. Futhermore, we had no control over elements such as data quality. For future works, the procedure and methods used in the research study can still be replicated and extended for use on a South African dataset.

4.6 Conclusion

Chapter 3 formally presents the research objectives and hypotheses of the study. The methodology followed, which will produce results eventually leading to either the acceptance or rejection of the research hypotheses, was discussed. Finally, limitations of the research were presented. The next chapter will provide the experiment results and conclusions that can be drawn from them.

Chapter 5

Results

5.1 Introduction

In chapter 3, a full description of the data and research methodology followed was presented. In this chapter, section 4.2 presents the experiment results for the performances of the different CHD risk prediction models. The results are presented in tabular form. A list of the variables ranked according to their contribution to our best-performing CHD risk prediction model is presented in the form of a graph. Section 4.3 resembles section 4.2, except that it presents the MetS prediction results. Section 4.4 presents a graph displaying the results for the performance of RF against AutoPrognosis on different sample of data varying in size. Finally, section 4.5 summaries the main points presented in this chapter.

5.2 CHD Risk Prediction Results

5.2.1 Characteristics of the study population

The dataset used in the study presented by this paper is from the Framingham Heart Study and is publicly available on the Kaggle website [Car 2020]. A total number of 3 658 participants had enough data for inclusion in our study. The mean (std) age of participants was 49.55 (8.56) years, while 1 623 (44.37%) were male and 2 035 (55.63%) were female. 557 participants (15.23%) were predicted to have a 10-year risk of developed CHD, with the mean (std) age being 54.27 (7.99) where 307 (55.12%) are male and 250 (44.88%) are female.

5.2.2 Comparison of prediction models

The AUC-ROC values for the different models under evaluation for 10-year CHD risk prediction are shown in Table 5.1. The Framingham Risk Score model (AUC-ROC: 0.633) was used as the baseline model for performance evaluation. There were two RF models under evaluation. For one of the RF models, hyperparameters were ignored during training (meaning that the hyperparameters were set to default values). For the other RF model, the hyperparameters were optimized using the out-of-bag error value.

Both models outperformed the baseline model however, the RF model with optimized hyperparameters (AUC-ROC: 0.728) performed significantly better than the RF model with default values for the hyperparameters (AUC-ROC: 0.653). The AutoPrognosis model can be created with a single pipeline or an ensemble pipeline. In some instances, the ensemble helps (however it was not always necessary). AutoPrognosis (single pipeline, using the 7 core variables) (AUC-ROC: 0.703) slightly outperformed AutoPrognosis (ensemble pipeline, using the 7 core variables) (AUC-ROC: 0.696). However, AutoPrognosis (ensemble pipeline, using all variables) (AUC-ROC: 0.714) outperformed AutoPrognosis (ensemble pipeline, using all variables) (AUC-ROC: 0.704). All these different versions of AutoPrognosis significantly outperformed the baseline model. The RF model with optimized hyperparameters had the best overall performance.

Model	**AUC-ROC**	**AUC Change**
Framingham Risk Score	0.633	Baseline Model
Random Forest (default hyperparameters)	0.653	+2.0%
Random Forest (optimised hyperparameters)	0.728	+9.5%
AutoPrognosis (single pipeline) - using the 7 core variables	0.703	+7.0%
AutoPrognosis (ensemble pipeline) - using the 7 core variables	0.696	+6.3%
AutoPrognosis (single pipeline) - using the all variables	0.704	+7.1%
AutoPrognosis (ensemble pipeline) - using the all variables	0.703	+8.1%

Table 5.1: Performance of 10-year CHD risk prediction models

The Confusion Matrix, Precision, Recall and F-1 Score values for the CHD risk prediction models with the lowest and best performances are displayed in Table 5.2 and Table 5.3, respectively.

		True Value		
		No CHD	**CHD**	Total
Predicted Value	**No CHD**	772	148	920
	CHD	263	120	383
	Total	1035	268	1303

	No CHD	**CHD**
Precision	0.84	0.31
Recall	0.75	0.45
F1-Score	0.79	0.37

Table 5.2: Confusion Matrix, Precision, Recall and F1-score values of the FRS model

		True Value		
		No CHD	**CHD**	Total
Predicted Value	**No CHD**	756	164	920
	CHD	155	228	383
	Total	911	392	1303

	No CHD	**CHD**
Precision	0.82	0.60
Recall	0.83	0.58
F1-Score	0.83	0.60

Table 5.3: Confusion Matrix, Precision, Recall and F1-score values of the best performing CHD risk prediction model

5.2.3 AutoPrognosis' selected algorithms

Table 5.4 presents AutoPrognosis' selected algorithms for the Data Imputation, Feature Preprocessing and Classification pipeline stages for each 10-year CHD risk prediction AutoPrognosis model of this study. The values of the weights for each pipeline in the ensemble is also included. For the AutoPrognosis models, where the seven conventional risk factors were used, the Logistic Regression algorithm was chosen as the most suitable for classification. While for the AutoPrognosis models, where the whole feature set was used, the Neural Networks and Quadratic Discriminant Analysis (QDA) algorithms which produce more complex models were chosen as the most suitable.

	Data Imputation	**Feature Preprocessing**	**Classification**	**Ensemble Pipeline Weights**
AutoPrognosis (single pipeline) - using the 7 core variables	missForest	PCA	Logistic Regression	1
AutoPrognosis (ensemble pipeline) - using the 7 core variables	mean	Gaussian Random Projections	Logistic Regression	0.021
	missForest	PCA	Logistic Regression	0.173
	median	PCA	Logistic Regression	0.806
AutoPrognosis (single pipeline) - using the all variables	most-frequent	Scaler	NeuralNet.	1
AutoPrognosis (ensemble pipeline) - using the all variables	most-frequent	Gaussian Transformer	Neural Net.	0.494
	mean	Gaussian Random Projections	QDA	0.452
	missForest	Uniform Transformer	QDA	0.054

Table 5.4: AutoPrognosis' selected algorithms for each pipeline stage for the various 10-year CHD risk prediction AutoPrognosis models

5.2.4 Variable Importance

Figure 5.1 displays a list of all variables ranked based on their contribution to the RF model with optimized hyperparameters (best performing model). The importance scores are based on their mean decrease in impurity. Along with the conventional CHD risk factors, 'heartRate' was among the top ranking. Although the 'currentSmoker' and 'diabetes' variables were ranked the lowest, 'glucose' and 'cigsPerDay' were in the top ranking.

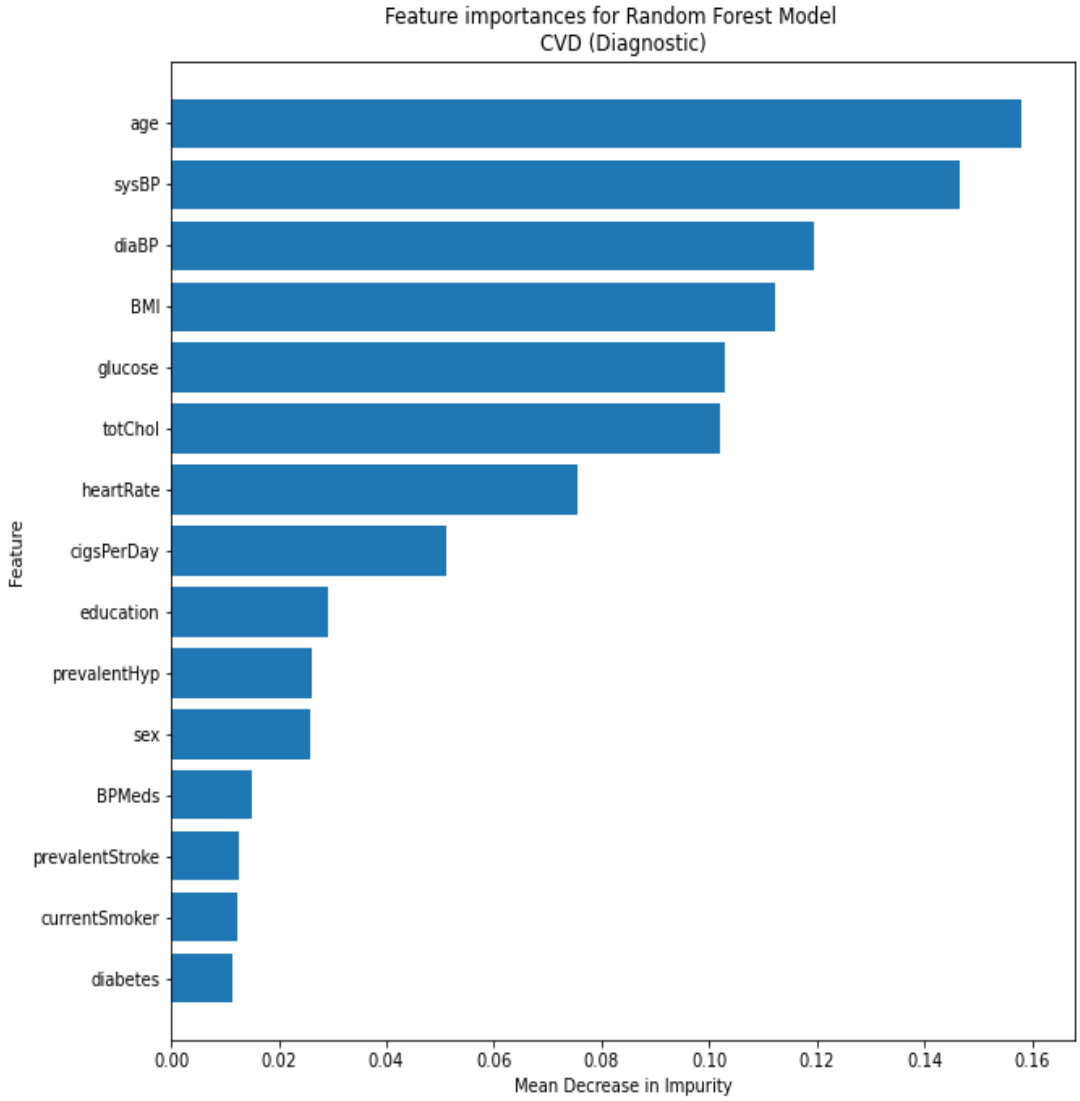


Figure 5.1: Variable Ranking of RF (optimised hyperparameters) model for 10-year CHD Risk Prediction

5.3 Metabolic Syndrome Prediction Results

5.3.1 Characteristics of the study population

The dataset used in the latter part of the study presented by this paper is from the NHANES and is publicly available on the Data World website [NCHS 2017]. A total number of 6 781 participants had enough data for inclusion in our study. The mean (std) age of participants was 39.10 (21.82) years, while 3 412 (50.31%) were male and 3 369 (49.69%) were female. 1061 participants (15.65%) had MetS, with the mean (std) age being 50.24 (19.97) where 531 (50.05%) are male and 530 (49.95%) are female.

5.3.2 Comparison of prediction models

The AUC-ROC values for the different models under evaluation for MetS prediction are shown in Table 5.5. The Random Forest model where hyperparameters were left to their values during training (AUC-ROC: 0.753) was used as the baseline model for performance evaluation. For the other RF model (AUC-ROC: 0.858), the hyperparameters were optimized using the out-of-bag error value. This model significantly outperformed the baseline model. AutoPrognosis (ensemble pipeline, using all variables) (AUC-ROC: 0.851) significantly outperformed AutoPrognosis (single pipeline, using all core variables) (AUC-ROC: 0.773). Both AutoPrognosis models outperformed the baseline model. The RF model with optimized hyperparameters had the best overall performance, but only by a small difference (0.007) to the AutoPrognosis (ensemble pipeline) - using all variables.

Model	AUC-ROC	AUC Change
Random Forest (default hyperparameters)	0.753	Baseline model%
Random Forest (optimised hyperparameters)	0.858	+10.5%
AutoPrognosis (single pipeline) - using all the variables	0.773	+2.0%
AutoPrognosis (ensemble pipeline) - using all the variables	0.851	+9.8%

Table 5.5: Performance of the best performing MetS prediction models

The Confusion Matrix, Precision, Recall and F-1 Score values for the MetS prediction models with the lowest and best performances are displayed in Table 5.6 and Table 5.7, respectively.

		True Value		
		No MetS	**MetS**	Total
Predicted Value	**No MetS**	1511	212	1723
	MetS	113	199	312
	Total	1624	411	2035

	No MetS	**MetS**
Precision	0.88	0.64
Recall	0.93	0.48
F1-Score	0.90	0.55

Table 5.6: Confusion Matrix, Precision, Recall and F1-score values of the baseline MetS prediction model

		True Value		
		No MetS	**MetS**	Total
Predicted Value	**No MetS**	1521	192	1713
	MetS	93	229	322
	Total	1614	421	2035

	No MetS	**MetS**
Precision	0.89	0.71
Recall	0.94	0.56
F1-Score	0.91	0.62

Table 5.7: Confusion Matrix, Precision, Recall and F1-score values of the best performing MetS prediction model

5.3.3 AutoPrognosis' selected algorithms

Table 5.8 presents AutoPrognosis' selected algorithms for the Data Imputation, Feature Preprocessing and Classification pipeline stages for each MetS prediction AutoPrognosis model of this study. The values of the weights for each pipeline in the ensemble is also included. The Gradient Boosting and XGBoost algorithms which produce tree-based models were chosen as the most suitable.

	Data Imputation	**Feature Preprocessing**	**Classification**	**Ensemble Pipeline Weights**
AutoPrognosis (single pipeline) - using the all variables	most-frequent	None	Gradient Boosting	1
AutoPrognosis (ensemble pipeline) - using the all variables	mean	Gaussian Transformer	Gradient Boosting	0.5
	missForest	Uniform Transformer	Gradient Boosting	0.0
	mean	Scaler	XGBoost	0.5

Table 5.8: AutoPrognosis' selected algorithms for each pipeline stage for the different MetS prediction AutoPrognosis models

5.3.4 Variable Importance

Figure 5.2 displays a list of all variables ranked based on their contribution to the RF model with optimized hyperparameters. The importance scores are based on their mean decrease in impurity.

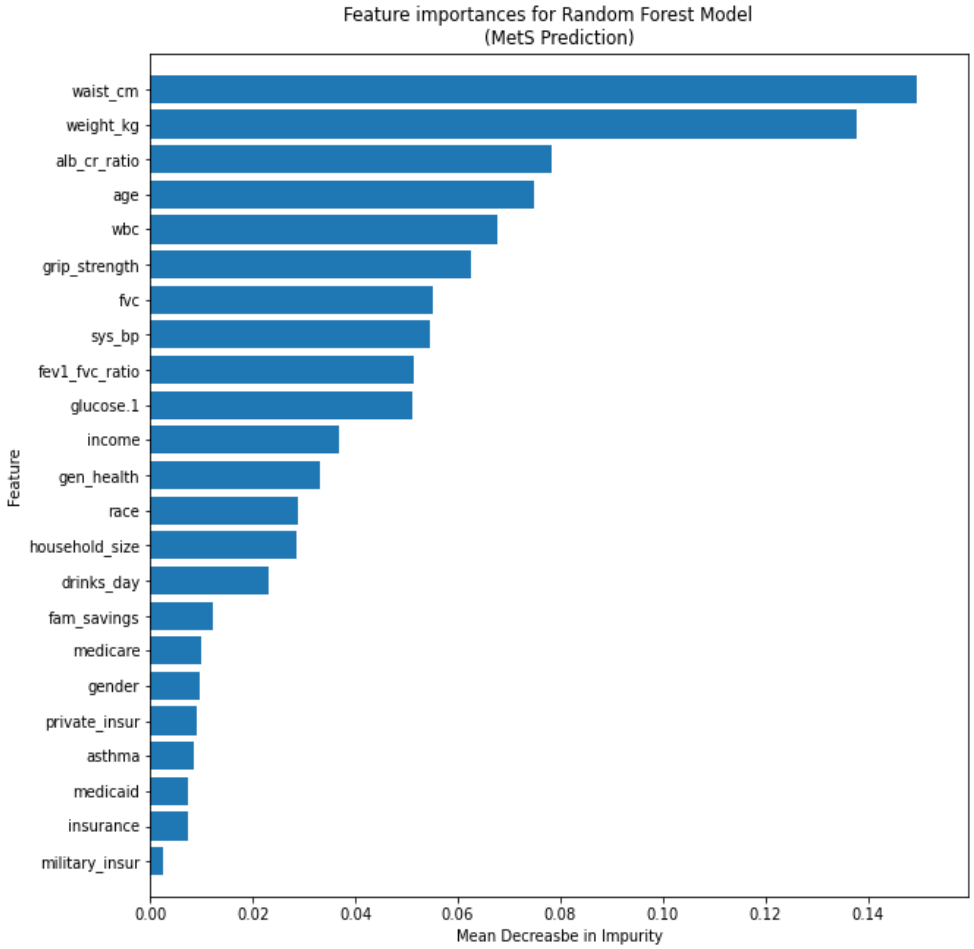


Figure 5.2: Variable Ranking of RF (optimised hyperparameters) model for MetS Prediction

5.4 Performance results of RF against AutoPrognosis on different sample sizes of data

Figure 5.3 displays the performances, indicated by the AUC-ROC value, of four different models which were each built using the NHANES dataset used in the previous section for predicting the presence of MetS. Once again, the four models are based on AutoPrognosis with a single pipeline, AutoPrognosis with an ensemble pipeline, Random Forest with default hyperparameter values and Random Forest with optimized hyperparameters values, respectively. All of these models had satisfactory performance results, on all the different sample sizes of data as shown in Figure 5.3. The Random Forest model with optimized hyperparameters values had the best overall performance, followed by AutoPrognosis with an ensemble pipeline, then AutoPrognosis with a single pipeline and finally the Random Forest model with default hyperparameter values.

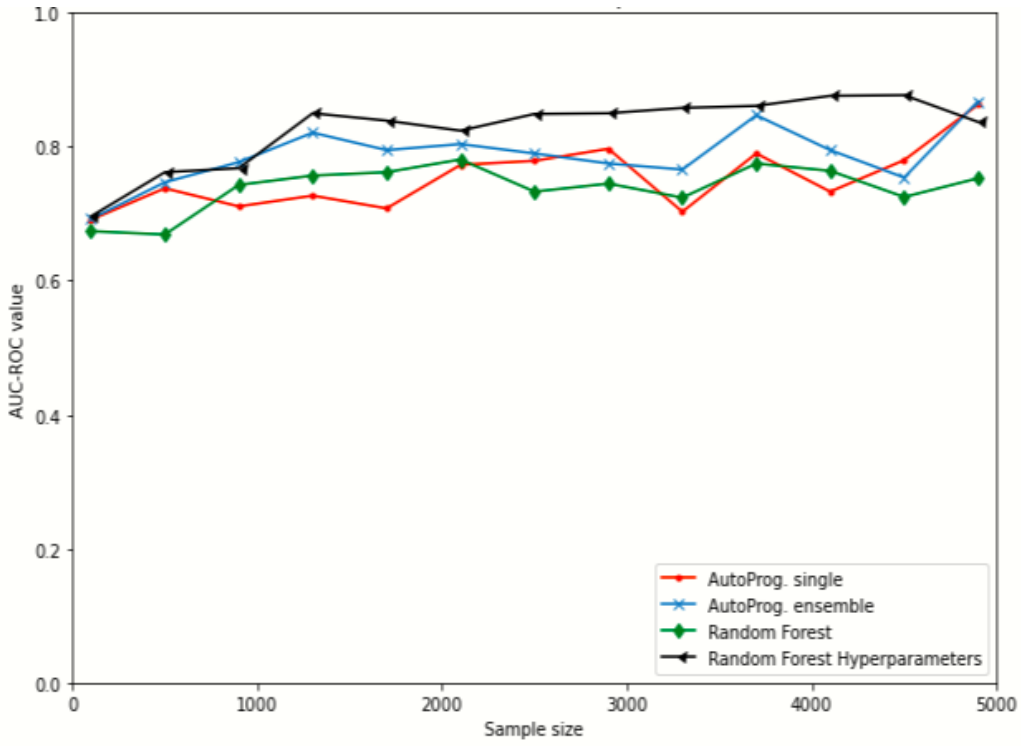


Figure 5.3: MetS prediction (AUC-ROC value vs Sample Size)

5.5 Conclusion

The results show that the RF model with optimized hyperparameters had the best performance for both CHD risk prediction and MetS outcome. The performance of RF against AutoPrognosis on different samples of data, varying in size from 100 to 4900, was also presented. The Random Forest model with optimized hyperparameter values produced the best overall performance, proceeded by AutoPrognosis consisting of an

ensemble pipeline, then AutoPrognosis consisting of a single pipeline and lastly the Random Forest model with hyperparameters set to its default values. In the next chapter, a discussion of the results and how they can be interpreted is provided.

Chapter 6

Discussion of Results

6.1 Introduction

In the previous chapter, a concise presentation of the experiment results was provided. In this chapter the results and their interpretations are discussed in detail. In Section 5.2, each subsection presents a discussion of the results for CHD Risk Prediction, MetS Prediction and the performance of RF against AutoPrognosis on different sample sizes of data, respectively. We also review how the experiment results lead to the acceptance/rejection of the research hypotheses stated in chapter 3 (section 3.3). Section 5.3 highlights some future work that can still be explored. Finally, section 5.4 provides a summary of the important points presented in this chapter.

6.2 Discussion

CVD accounts for approximately 32% of all deaths worldwide, making it the leading cause of mortality globally, with CHD being the most common type of CVD [Car 2021]. However, clinicians reiterate that 80% of heart attacks and CVD related events can be prevented [WHO 2015]. The individual components of MetS are recognized as CHD risk factors and are often included as some of the core variables used in models developed for CHD risk prediction. Predicting future CHD risk and MetS occurrence at an individual level, population level and in specific subgroups of the population will provide useful information for policymakers and healthcare authorities about these risks. CHD risk prediction and MetS detection can also motivate individuals to adjust their lifestyle choices, behaviours, and habits.

6.2.1 CHD Risk Prediction

Table 5.1 shows that for 10-year CHD risk prediction, both types of our machine learning models significantly outperformed the FRS model (AUC-ROC: 0.633) which was developed using the seven conventional CVD risk factors discussed in section 2.5. Based on these results, our first hypothesis ("Both types of CHD risk prediction models, based on RF and AutoPrognosis, will outperform the FRS model.") is accepted. Therefore, using machine learning methods for clinical prognostic applications will be beneficial.

However, it is important to remember that RF and AutoPrognosis are both ensemble machine learning methods. This means that they are fairly complex methods which often have the ability to realise more complex data relationships in large datasets compared to other methods that only rely on using a single type of machine learning classifier as opposed to an ensemble. Therefore, not every machine learning method will always have significant improvements over the FRS model. Some may not even lead to any improvements. The results in section 2.7 (Figure 3.1), from the study performed by Yang *et al.* [2020], show how the Naïve Bayes classifier actually underperformed the FRS model. From this, we learn that selecting the most suitable machine learning method for application is necessary.

From the results, it is also important to notice that even when we reduce the number of features used and only include the seven conventional CVD risk factors used in the FRS model to train our AutoPrognosis models, the FRS model still significantly underperformed AutoPrognosis despite using datasets which are exactly the same with an equal amount of information gain. This is because of the modelling gain AutoPrognosis achieves over the FRS model (which is based on standard Cox proportional hazards regression models). Further details related to the information/modelling gain achieved by AutoPrognosis were covered in **??**. AutoPrognosis was able to select the most suitable model in terms of numerical robustness and complexity to adequately handle the dimensionality of the data due to the complex interactions that may exist between the features. Each of our AutoPrognosis models, using the seven conventional risk factors, had Logistic Regression as the most suitable algorithm for classification.

In addition to choosing the most suitable machine learning model, tuning the hyperparameters of the model is important. Every machine learning model has a set of different parameters whose values are learnt from the training dataset. Hyperparameters are simply a particular type of parameters whose values define more advanced machine learning concepts, such as the model's learning rate, complexity, penalties, number of estimators, etc. Hyperparameters need to be predefined, either manually or by use of an optimisation algorithm, as they cannot be learnt directly from the data during the model training. Being able to tune and find the optimal hyperparameter values is important as it gives us the ability to adjust the models to suit specific use cases and can be used to prevent under/overfitting to yield even better accuracy results on unseen data. Furthermore, it improves the convergence rate and efficiency of an algorithm. This explains why our 10-year CHD risk prediction RF model with optimised hyperparameters (AUC-ROC: 0.728) outperformed our RF model where the hyperparameters (AUC-ROC: 0.653) were left set to their default values.

Based on the results shown in Table 5.1, when using all the variables, AutoPrognosis with an ensemble pipeline (AUC-ROC: 0.714) outperforms AutoPrognosis with a single pipeline (AUC-ROC: 0.704). However, when using the 7 conventional risk factors, AutoPrognosis with a single pipeline (AUC-ROC: 0.703) had a very similar performance to AutoPrognosis with an ensemble pipeline (AUC-ROC: 0.696). Even though in both cases one model outperforms the other, the performance results of the models are very similar. This contrast suggests that the ensemble pipeline may not always be necessary,

especially if the model which uses a single classifier produces satisfactory/comparable results. Generally, ensembles are meant to reduce the variance in a model. However, the performance of the average of n models in an ensemble is significantly improved only if the individual models are significantly independent of each other. In other words, if the individual classifiers have very similar performances, the average performance will be also similar to that of an individual classifier. Although not common, it is also possible for an ensemble to significantly underperform an individual classifier model. This can happen when you have a dataset with very few strong predictors leading to the ensemble having only a few moderate/good classifiers with many bad classifiers. Taking the average performance of a few true models and many bad models produces a fairly bad ensemble. Meaning that our third hypothesis ("The AutoPrognosis models with an ensemble pipeline will always outperform the AutoPrognosis models with a single pipeline, for both 10-year CHS risk prediction and MetS prediction respectively.") is rejected.

The RF model with optimized hyperparameters (AUC-ROC: 0.728) outperformed the AutoPrognosis model with an ensemble pipeline (AUC-ROC: 0.714) but only by a very small margin. Therefore, our second hypothesis ("The AutoPrognosis models will always outperform the RF models, for both 10-year CHD risk prediction and MetS detection respectively.") is rejected. For the data used in our study, the performances of the two models are comparable. Hence, both models can still be used to improve CHD risk prediction results. Along with the conventional CHD risk factors, '*heartRate*' was among the top ranking. Although the 'diabetes' and '*currentSmoker*' variables were ranked the lowest due to having a considerable number of missing values, '*glucose*' and '*cigsPerDay*' were in the top rankings.

6.2.2 Metabolic Syndrome Prediction

To the best of our knowledge, our study is unique in that it is the first to conduct a comparative analysis for a performance evaluation between models based on RF vs AutoPrognosis for MetS prediction. Although Alaa and Schaar [2018] and Alaa *et al.* [2019] presented studies which were conducted with the aim of applying AutoPrognosis to improve CVD risk prediction, just like other machine learning methods AutoPrognosis can be used to solve other classification problems as demonstrated by Alaa and van der Schaar [2018] and Shah *et al.* [2021]. Hence, we used AutoPrognosis to develop MetS prediction models.

Our study discloses several significant points about the use of machine learning for MetS prediction. Firstly, both types of our machine learning models based on RF and AutoPrognosis had satisfactory performance results as displayed in Table 5.5. Therefore, using machine learning models for applications in MetS prediction will be beneficial. Furthermore, predicting MetS for the purpose of getting it treated can also help prevent the development of a CVD like CHD, as the different conditions associated with MetS are some of the major CVD risk factors. Among the conventional risk factors for MetS, at the top of our variable ranking list (produced by a random forest model fitted to the dataset) are: "*alb_cr_ratio*" (ratio of albumin to creatine in urine; this measure

is used to identify kidney disease that occurs due to complications of diabetes), "*wbc*" (white blood cell count), "*fvc*" (forced vital capacity; this is the maximum amount of air that can be exhaled forcibly after inhaling fully) and '*fev1 fvc ratio*' (ratio of the amount of air expelled on one second to forced vitality capacity). Including these variables, along with the conventional risk factors, as part of the diagnostic criteria may lead to improved MetS prediction results in the clinical space.

Our RF model with optimized hyperparameters (AUC-ROC: 0.858) outperformed the AutoPrognosis model with an ensemble pipeline (AUC-ROC: 0.851) but only by a very small margin. However, even though the two models' performances are comparable, AutoPrognosis may be the preferred method applicable for classification problems. This is because clinicians without the adequate knowledge and expertise in data science are challenged to manually design and tune ML modelling pipelines before they can put them to use. Therefore, even though there has been an increased number of applications for machine learning models and techniques in clinical prognostic research, often there exists a gap between the potential and actual practicality of these machine learning approaches. However, AutoPrognosis was developed to circumvent this challenge, as it is an automated ML framework specifically designed for clinical prognosis. As shown in Table 5.8, AutoPrognosis selected the Gradient Boosting and XGBoost as the most suitable classification algorithms for our dataset. It is important to take note of the applications of the Data Imputation and Feature Preprocessing, as the algorithms chosen during these steps affect the dataset and can therefore also have an effect on the overall performance of the models. Excluding these steps may change the performance results as factors such as missing data and outliers make the data biased.

Once again, the RF model with optimised hyperparameters (AUC-ROC: 0.858) significantly outperformed the RF model with default hyperparameter values (AUC-ROC: 0.753). The value of optimising hyperparameters is discussed in **??** above.

6.2.3 Performance of RF against AutoPrognosis on different sample sizes of data

The size, quality and dimensionality of the dataset used to train a model will often affect the performance of the algorithm the model is based on. Therefore, the choice of the best machine learning model and the tuning of its hyperparameters are important for ensuring the possible advantages of machine learning applications. AutoPrognosis automates these processes, making it more easily useful for machine learning applications in the clinical space. However, in papers such as those presented by Alaa and Schaar [2018] and Alaa *et al.* [2019] the performance of AutoPrognosis was evaluated using datasets with at least 30 000 and 423 604 patient records, respectively. Sometimes obtaining data that large and training models on it can be challenging and computationally expensive. Furthermore, there can be many specific challenges associated with obtaining large quantities of medical data such as insufficient data for rare diseases, getting ethics clearance due to doctor-patient privacy concerns or the lack of laboratory equipment for the collection and testing of human samples to generate specific data (this the case in many developing countries).

Evaluating the performance of AutoPrognosis on significantly smaller datasets becomes essential as some ML models, such as Neural Networks, are data hungry and require large quantities of data during training in order to produce satisfactory performance results. Once again, to the best of our knowledge, our study is unique in that it is the first to conduct a comparative analysis for a performance evaluation between models based on RF against AutoPrognosis on different sample sizes of data. Based on Figure 5.3, our results concluded that AutoPrognosis with a single pipeline, AutoPrognosis with an ensemble pipeline, Random Forest with default hyperparameter values and Random Forest with optimized hyperparameters values all had satisfactory performance results, on all the different sample sizes of data. As a result, our fourth hypothesis ("AutoPrognosis will have satisfactory performance results on all the different sample sizes of data ranging from 100 to 4900.") is accepted. The best performance results were produced by the RF model with optimized hyperparameters, proceeded by AutoPrognosis consisting of an ensemble pipeline, then AutoPrognosis consisting of a single pipeline and lastly the Random Forest model with hyperparameters set to their default values. Finally, our fifth hypothesis ("AutoPrognosis will always outperform RF on all the different sample sizes of data ranging from 100 to 4900.") is rejected.

6.3 Future Works

Initially, the main objectives of our research study were to determine the most significant CVD risk factors and develop a 10-year CVD risk prediction model specific to the **South African** population. However, due to time constraints, obtaining a local dataset could not be achieved, therefore limiting the scope of our research study. This led to the alteration of our original research question, as we had to rely on using publicly available datasets which were not collected from the desired geographic location. Another major limitation to using this kind of secondary data is that we did not know exactly how the data collection process was done. We had no information about any potential biases which may have existed, and the extent to which the data may have been affected by issues such as low response rates or survey respondents misinterpreting specific survey questions. Futhermore, we had no control over elements such as data quality. For future works, the procedure and methods used in the research study can still be replicated and extended for use on a South African dataset.

6.4 Conclusion

Within this chapter, the results which were presented in chapter 5 are thoroughly discussed. Additionally, the research hypotheses which were formerly presented in chapter 5 are reviewed in relation to the experiment results obtained. Finally, any potential future works which pertain to this research are presented in section 6.3. In chapter 7 we conclude this document with a summary of all the important points.

Chapter 7

Conclusion

CVD is a medical term used to refer to a category of diseases that affect the heart and blood vessels. Globally, CHD is the most commonly diagnosed CVD, with approximately 200 million people living with the disease. Behavioural risk factors such as consuming an unhealthy diet, high alcohol intake, smoking and inadequate physical activity are some of the major determinants of developing CHD. The consequences of these behavioural risks often manifest themselves as MetS. MetS refers to a collection of conditions, namely high blood pressure, high blood glucose levels, abnormal cholesterol level, overweight and obesity. These conditions are known to increase a person's chance of developing a CVD and diabetes. It is estimated that 20-25% of the world's adult population lives with MetS.

CVD is responsible for 32% of global deaths, making it the leading cause of death worldwide. 85% of CVD related deaths are caused by the onset of severe and sudden conditions, namely strokes and heart attacks. Even though the present statistics are distressing, medical professionals have reiterated that 80% of premature heart attacks and strokes can be prevented as the major risk factors are largely influenced by lifestyle choices that people can change. Predicting the presence of MetS and future CHD risk will provide information to policymakers and healthcare authorities which can be used to help decide on regulations to reduce the presence of risk factors, e.g. putting policies in place to reduce the salt and sugar content allowed in processed food. Risk prediction can also motivate individuals to adjust their lifestyle choices, behaviours, and habits that lead to CHD/MetS.

FRS and SCORE are examples of some of the most commonly used CVD risk prediction models. These models have been helpful however they were created and validated on data from North American and European populations (populations of developed countries). Therefore, when applied to populations from developing countries, where lifestyle choices differ, these models can underestimate/overestimate CVD risk. Additionally, many of these CVD risk prediction models rely on standard regression methods. Generally, a limited number of conventional risk factors are used and the regression models tend to assume that the correlation between the variables and the CVD outcome is linear. This affects the performance metrics used in evaluating such a model. Models based on ensemble machine learning methods, such as Random Forests

and AutoPrognosis, can combine multiple different algorithms in machine learning, like Decision Tress (DT) to find non-linear relationships between the dependent and independent variables at a specific threshold between precision and recall to ensure overfitting is minimized. These ensemble methods also provide a more robust and accurate prediction compared to single algorithm models, making them a popular choice in various fields such as healthcare and finance. The proposed model in this research study is implemented to model complex nonlinear relationships in large repositories.

The primary goals of this study were to identify the most important CHD risk factors and to conduct a performance evaluation study to compare the predictive performance of the FRS model with two different types of CHD risk prediction models based on ensemble machine learning methods, namely RF and AutoPrognosis. Furthermore, we identified the most significant MetS risk factors, and in a novel study, we compared the predictive performance of RF vs. Autoprognosis for determining the presence of MetS.

In summary, the experiment results demonstrated that both proposed methods, RF and AutoPrognosis, outperformed the FRS model for CHD risk prediction, which had an average AUC-ROC score of 0.633. With optimized hyperparameters, the RF model obtained an AUC-ROC score of 0.728. However, the results were comparable to the AutoPrognosis model, with an AUC-ROC of 0.714. Both models can still be used to improve the accuracy of CHD risk prediction. The results also showed that using machine learning models for MetS prediction will be beneficial, as both types of our machine learning models based on RF and AutoPrognosis performed well. The RF model with optimized hyperparameters performed similarly to the AutoPrognosis model, with an average AUC-ROC score of 0.858 and an average AUC-ROC score of 0.851, respectively. AutoPrognosis may be the preferred method for classification problems because it was designed to eliminate the difficulty of manually designing and tuning machine learning modeling pipelines by automating the entire process.

Sometimes attaining large datasets of medical data to train models can be challenging and computationally expensive. Hence, the performance of RF vs AutoPrognosis on different sample sizes of data, ranging from 100 to 4900 (at intervals of 400), was also tested for the final phase of this study. That was another objective achieved in this study. Acquiring large datasets of medical data to train models can be difficult and computationally expensive at times. As a result, the performance of RF vs. Autoprognosis on various data sample sizes ranging from 100 to 4900 (at 400 intervals) was also tested for the final phase of this study. This study also accomplished another goal. The results revealed that the RF model with optimized hyperparameters produced the best performance results, followed by AutoPrognosis, which consists of an ensemble pipeline, and finally the RF model with hyperparameters set to their default values.

In conclusion, machine learning applications in health care are very promising and have the potential to make a significant difference. Machine learning models have shown value in clinical prognosis and diagnosis, leading to better outcomes and the potential to save lives and medical costs. Additionally, these models can also help healthcare providers identify patients who are at risk of developing certain conditions, allowing

for early intervention and preventative measures to be taken. With continued research and development, machine learning has the potential to revolutionize the healthcare industry and improve patient outcomes on a global scale. In general, the application of machine learning could be beneficial to medical technologists, but it must also be regulated to ensure that the technologists are aligned and adhere to ethical issues.

Appendix A

List of CHD risk prediction variables

Variable	Definition
Sex	Male or Female
Age	Age in years
Current Smoking Status	Yes - the patient is currently a smoker No - the patient is not a smoker
Cigarettes per day	The average number of cigarettes the patient smokes per day
Blood Pressure Medication	Yes - the patient takes BP medication No - the patient does not take BP medication
Stroke History	Yes - the patient has had a stroke No - the patient has never had a stroke
High Blood Pressure Status	Yes - the patient is hypertensive No - the patient is not hypertensive
Diabetes Status	Yes - the patient has diabetes No - the patient does not have diabetes
Total Cholesterol	Total Cholesterol Level
Systolic Blood Pressure	Systolic Blood Pressure Level
Diastolic Blood Pressure	Diastolic Blood Pressure Level
BMI	Body Mass Index
Heart Rate	Number of heart beats per minute
Glucose Level	Blood Glucose Level
10-year CHD Risk	Yes - patient is at risk of developing CHD in 10 years No - patient is not at risk of developing CHD in 10 years

Table A.1: List of CHD risk prediction variables and definitions

Appendix B

List of MetS prediction variables

Variable	Definition
Sex	Male or Female
Age	Age in years
Race	Black, White, Mexican Hispanic, Other Hispanic or Other
Income	Household Income
Household Size	Number of people in household
Insurance	Yes - the person is covered by insurance No - the person is not covered by insurance
Private Insurance	Yes - the person has private insurance No - the person does not have private insurance
Medicare	Yes - the person has Medicare insurance No - the person does not have Medicare insurance
Medicaid	Yes - the person has Medicaid insurance No - the person does not have Medicaid insurance
Military Insurance	Yes - the person has military insurance No - the person does not have military insurance
General Health	Excellent, Moderate or Poor
Family Savings	Accumulated family wealth
Asthma	Yes - the person has asthma No - the person does not have asthma
Drinks per day	The average number of alcoholic drinks the person consumes a day
Weight	Weight in kg
Waist circumference	Measure of the waist circumference in cm
Systolic Blood Pressure	Systolic Blood Pressure Level
Albumin to Creatinine Ratio	Albumin to Creatinine Ratio in urine. A measure of protein in the urine.
White Blood Cell Count	White Blood Cell Count
Glucose	Blood Glucose Level

Grip Strength	Measure of muscular strength in the hand
Forced Vital Capacity	Maximum amount of air the person can exhale forcibly after fully inhaling
Forced Vital Capacity to Forced Expiratory Volume Ratio	Ratio of the maximum amount of air the person can exhale forcibly after fully inhaling to the amount of air the person can exhale forcibly in one second
Metabolic Syndrome Prediction	Yes - the person has metabolic syndrome No - the person does not have metabolic syndrome

Table B.1: List of MetS risk prediction variables and definitions

References

[Alaa and Schaar 2018] Ahmed Alaa and Mihaela Schaar. Autoprognosis: Automated clinical prognostic modeling via bayesian optimization with structured kernel learning. In *International conference on machine learning*, pages 139–148. PMLR, 2018.

[Alaa and Schaar 2019] Ahmed Alaa and Mihaela Schaar. *AutoPrognosis: Automated Clinical Prognostic Modelling via Bayesian Optimization*. https://github.com/ahmedmalaa/AutoPrognosis, December 2019. (Accessed on 15/07/2021).

[Alaa and van der Schaar 2018] Ahmed M Alaa and Mihaela van der Schaar. Prognostication and risk factors for cystic fibrosis via automated machine learning. *Scientific reports*, 8(1):1–19, 2018.

[Alaa *et al.* 2019] Ahmed M Alaa, Thomas Bolton, Emanuele Di Angelantonio, James HF Rudd, and Mihaela van der Schaar. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants. *PloS one*, 14(5):e0213653, 2019.

[Alberti *et al.* 2006] George Alberti, Paul Zimmet, Jonathan Shaw, Scoot M Grundy, et al. The idf consensus worldwide definition of the metabolic syndrome. *Brussels: International Diabetes Federation*, 23(5):469–80, 2006.

[Amaratunga *et al.* 2008] Dhammika Amaratunga, Javier Cabrera, and Yung-Seop Lee. Enriched random forests. *Bioinformatics*, 24(18):2010–2014, 2008.

[Amit and Geman 1997] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.

[Bader-El-Den and Gaber 2012] Mohamed Bader-El-Den and Mohamed Gaber. Garf: towards self-optimised random forests. In *International conference on neural information processing*, pages 506–515. Springer, 2012.

[Barrett 2014] James Barrett. Weibull-cox proportional harzard model. 2014.

[BHF 2022] British Heart Foundation BHF. Global heart circulatory diseases factsheet. 2022.

[Bloom *et al.* 2011] David E Bloom, Dan Chisholm, Eva Jané-Llopis, Klaus Prettner, Adam Stein, Andrea Feigl, et al. *From burden to" best buys": reducing the economic impact of non-communicable disease in low-and middle-income countries*. Technical report, Program on the Global Demography of Aging, 2011.

[Breiman 1996] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[Breiman 2001] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[Car 2020] *Logistic Regression To predict heart disease— Kaggle*. https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression, 2020. (Accessed on 05/11/2021).

[Car 2021] *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), June 2021. (Accessed on 01/05/2022).

[CHD 2021] *Coronary Artery Disease (CAD)*. https://www.cdc.gov/heartdisease/coronary_ad.htm, July 2021. (Accessed on 30/05/2022).

[Choe *et al.* 2018] Eun Kyung Choe, Hwanseok Rhee, Seungjae Lee, Eunsoon Shin, Seung-Won Oh, Jong-Eun Lee, and Seung Ho Choi. Metabolic syndrome prediction using machine learning models with genetic and clinical information from a nonobese healthy population. *Genomics & informatics*, 16(4), 2018.

[Cleveland Clinic 2022] Cleveland Clinic. *Coronary artery disease: Symptoms, causes & treatment*. https://my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease, Aug 2022. (Accessed on 13/09/2022).

[Conroy *et al.* 2003] Ronán Michael Conroy, K Pyörälä, AP el Fitzgerald, S Sans, A Menotti, Gui De Backer, Dirk De Bacquer, P Ducimetiere, P Jousilahti, U Keil, et al. Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project. *European heart journal*, 24(11):987–1003, 2003.

[Cox 1972] DR Cox. Regression models and life tables. *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

[Cupples 1987] LA Cupples. Section 34: some risk factors related to the annual incidence of cardiovascular disease and death in pooled repeated biennial measurements. *Framingham Heart Study: 30 Year Follow Up*, pages 1–22, 1987.

[D'Agostino and Nam 2003] Ralph B D'Agostino and Byung-Ho Nam. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, 23:1–25, 2003.

[Dec 2020] *Decision Tree vs. Random Forest - Which Algorithm Should you Use?* https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/, 2020. (Accessed on 05/11/2021).

[Dietterich 2000] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

[D'agostino *et al.* 2008] Ralph B D'agostino, Ramachandran S Vasan, Michael J Pencina, Philip A Wolf, Mark Cobain, Joseph M Massaro, and William B Kannel. General cardiovascular risk profile for use in primary care. *Circulation*, 117(6):743–753, 2008.

[Expert Panel on Detection and others 2001] Evaluation Expert Panel on Detection et al. Executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). *Jama*, 285(19):2486–2497, 2001.

[Goff *et al.* 2014] David C Goff, Donald M Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B D'agostino, Raymond Gibbons, Philip Greenland, Daniel T Lackland, Daniel Levy, Christopher J O'donnell, et al. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Part B):2935–2959, 2014.

[Goldstein *et al.* 2017] Benjamin A Goldstein, Ann Marie Navar, and Rickey E Carter. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal*, 38(23):1805–1814, 2017.

[Gouda *et al.* 2019] Hebe N Gouda, Fiona Charlson, Katherine Sorsdahl, Sanam Ahmadzada, Alize J Ferrari, Holly Erskine, Janni Leung, Damian Santamauro, Crick Lund, Leopold Ndemnge Aminde, et al. Burden of non-communicable diseases in sub-saharan africa, 1990–2017: results from the global burden of disease study 2017. *The Lancet Global Health*, 7(10):e1375–e1387, 2019.

[Gurney 1997] Kevin Gurney. *An introduction to neural networks*. CRC press, 1997.

[Gutiérrez-Esparza *et al.* 2020] Guadalupe Obdulia Gutiérrez-Esparza, Oscar Infante Vázquez, Maite Vallejo, and José Hernández-Torruco. Prediction of metabolic syndrome in a mexican population applying machine learning algorithms. *Symmetry*, 12(4):581, 2020.

[Ho 1998] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[Hoyt 2020a] Robert Hoyt. *Open Data Project with NHANES 2011-2012 Data: Codebook_NHANES_2011_2012.xlsx*. https://data.world/rhoyt/librehealth-educational-ehr/workspace/file?filename=Codebook_NHANES_2011_2012.xlsx, 2020. (Accessed on 17/10/2021).

[Hoyt 2020b] Robert Hoyt. *Open Data Project with NHANES 2011-2012 Data: Merged_Unique_Names_V2.csv*. https://data.world/rhoyt/librehealth-educational-ehr/workspace/file?filename=Merged_Unique_Names_V2.csv, 2020. (Accessed on 17/10/2021).

[IHME 2020] IHME. *GBD Compare Data Visualization*. https://vizhub.healthdata.org/gbd-compare/, 2020. (Accessed on 09/07/2023).

[Kandasamy *et al.* 2015] Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. High dimensional bayesian optimisation and bandits via additive models. In *International conference on machine learning*, pages 295–304. PMLR, 2015.

[Kim *et al.* 2022] Junho Kim, Sujeong Mun, Siwoo Lee, Kyoungsik Jeong, and Younghwa Baek. Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in korea. *BMC Public Health*, 22(1):1–10, 2022.

[Kruschke 2008] John K Kruschke. Bayesian approaches to associative learning: From passive to active learning. *Learning & behavior*, 36(3):210–226, 2008.

[Latinne *et al.* 2001] Patrice Latinne, Olivier Debeir, and Christine Decaestecker. Limiting the number of trees in random forests. In *International workshop on multiple classifier systems*, pages 178–187. Springer, 2001.

[Lloyd-Jones *et al.* 2006] Donald M Lloyd-Jones, Eric P Leip, Martin G Larson, Ralph B d'Agostino, Alexa Beiser, PW Wilson, Philip A Wolf, and Daniel Levy. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. *Circulation*, 113(6):791–798, 2006.

[Lloyd-Jones 2010] Donald M Lloyd-Jones. Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation*, 121(15):1768–1777, 2010.

[Met 2021] *Metabolic syndrome*. https://www.mayoclinic.org/diseases-conditions/metabolic-syndrome/symptoms-causes/syc-20351916, 2021. (Accessed on 15/12/2022).

[NCEP Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults 2002] NCEP Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults. *Third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III)*. Number 2. The Program, 2002.

[NCHS 2017] NCHS. *About the National Health and Nutrition Examination Survey*. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm, September 2017. (Accessed on 03/05/2022).

[Ngcayiya and Ranchod 2022] Paulina Ngcayiya and Pravesh Ranchod. Comparative performance analysis of random forests against autoprognosis for predicting coronary heart disease risk and metabolic syndrome: A retrospective cohort study. In *MATEC Web of Conferences*, volume 370, page 07005. EDP Sciences, 2022.

[NHLBI 2016] NHLBI. *Framingham Heart Study (FHS)*. https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs, April 2016. (Accessed on 15/03/2022).

[Pampel 2005] Fred C Pampel. Patterns of tobacco use in the early epidemic stages: Malawi and zambia, 2000–2002. *American journal of public health*, 95(6):1009–1015, 2005.

[Pencina and D'Agostino 2004] Michael J Pencina and Ralph B D'Agostino. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*, 23(13):2109–2123, 2004.

[Robnik-Šikonja 2004] Marko Robnik-Šikonja. Improving random forests. In *European conference on machine learning*, pages 359–370. Springer, 2004.

[Sacco *et al.* 2016] Ralph L Sacco, Gregory A Roth, K Srinath Reddy, Donna K Arnett, Ruth Bonita, Thomas A Gaziano, Paul A Heidenreich, Mark D Huffman, Bongani M Mayosi, Shanthi Mendis, et al. The heart of 25 by 25: achieving the goal of reducing global and regional premature deaths from cardiovascular diseases and stroke: a modeling study from the american heart association and world heart federation. *Circulation*, 133(23):e674–e690, 2016.

[Saffari *et al.* 2009] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. On-line random forests. In *2009 ieee 12th international conference on computer vision workshops, iccv workshops*, pages 1393–1400. IEEE, 2009.

[Shah *et al.* 2021] Akash A Shah, Sai K Devana, Changhee Lee, Reza Kianian, Mihaela van der Schaar, and Nelson F SooHoo. Development of a novel, potentially universal machine learning algorithm for prediction of complications after total hip arthroplasty. *The Journal of Arthroplasty*, 36(5):1655–1662, 2021.

[Snoek *et al.* 2012] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944*, 2012.

[Sou 2018] *South Africa's population — South Africa Gateway*. https://southafrica-info.com/people/south-africa-population/, 2018. (Accessed on 05/11/2021).

[Townsend *et al.* 2006] L Townsend, AJ Flisher, T Gilreath, and G King. A systematic review of tobacco use among sub-saharan african youth. *Journal of Substance Use*, 11(4):245–269, 2006.

[Tsymbal *et al.* 2006] Alexey Tsymbal, Mykola Pechenizkiy, and Pádraig Cunningham. Dynamic integration with random forests. In *European conference on machine learning*, pages 801–808. Springer, 2006.

[United Nations 2018] United Nations. *68% of the world population projected to live in urban areas by 2050, says un — UN Desa Department of Economic and Social Affairs*. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html, May 2018. (Accessed on 17/05/2022.

[Vorster 2002] HH Vorster. The emergence of cardiovascular disease during urbanisation of africans. *Public health nutrition*, 5(1a):239–243, 2002.

[Wang *et al.* 2013] Ziyu Wang, Masrour Zoghi, Frank Hutter, David Matheson, Nando De Freitas, et al. Bayesian optimization in high dimensions via random embeddings. In *IJCAI*, pages 1778–1784. Citeseer, 2013.

[WebMD 2021] WebMD. *Metabolic syndrome: Risk factors & causes*. https://www.webmd.com/heart/metabolic-syndrome/metabolic-syndrome-what-is-it, October 2021. (Accessed on 14/09/2022).

[Weng *et al.* 2017] Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.

[WHO 2007] WorldHealthOrganization WHO. *Prevention of cardiovascular disease: guidelines for assessment and management of total cardiovascular risk*. World Health Organization, 2007.

[WHO 2015] WHO. *Cardiovascular diseases: Avoiding heart attacks and strokes*. https://www.who.int/news-room/questions-and-answers/item/cardiovascular-diseases-avoiding-heart-attacks-and-strokes, September 2015. (Accessed on 15/03/2022).

[Wilson *et al.* 1998] Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

[Xu *et al.* 2012] Baoxun Xu, Joshua Zhexue Huang, Graham Williams, and Yunming Ye. Hybrid weighted random forests for classifying very high-dimensional data. *International Journal of Data Warehousing and Mining*, 8(2):44–63, 2012.

[Xu *et al.* 2017] Shan Xu, Zhen Zhang, Daoxian Wang, Junfeng Hu, Xiaohui Duan, and Tiangang Zhu. Cardiovascular risk prediction method based on cfs subset evaluation and random forest classification framework. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 228–232. IEEE, 2017.

[Yang *et al.* 2020] Li Yang, Haibin Wu, Xiaoqing Jin, Pinpin Zheng, Shiyun Hu, Xiaoling Xu, Wei Yu, and Jing Yan. Study of cardiovascular disease prediction model based on random forest in eastern china. *Scientific reports*, 10(1):1–8, 2020.

[Yu *et al.* 2020] Cheng-Sheng Yu, Yu-Jiun Lin, Chang-Hsien Lin, Sen-Te Wang, Shiyng-Yu Lin, Sanders H Lin, Jenny L Wu, Shy-Shin Chang, et al. Predicting metabolic syndrome with machine learning models using a decision tree algorithm: retrospective cohort study. *JMIR medical informatics*, 8(3):e17110, 2020.