CORONARY HEART DISEASE PREDICTION IN THE USA AND FACTORS THAT FAVOR ITS OCCURENCE

Jeremy Kibiru Gachanja, 101167

Submitted in partial fulfillment of the requirements for the Degree of Financial Economics at Strathmore University

SIMS
Strathmore University
Nairobi, Kenya

July 2020

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

Jeremy Kibiru Gachanja

.....

.....8/2/2021.....

This Research Project has been submitted for examination with my approval as the Supervisor.

.....Dr. Evans Otieno Omondi.....

.....

.....

Strathmore Institute of Mathematical Sciences

Strathmore University

Table of Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

- CHD: - Coronary Heart Disease
- CVD: - Cardiovascular Disease
- Exang: - exercise induced chest pain.
- Ca: - the number of blood vessels colored by fluoroscopy
- Thal: - Type of heart defect
- HDL- High density lipoproteins
- LDL- Low density lipoproteins
- Thalach: - maximum heart rate achieved.

ABSTRACT

Coronary Heart Disease (CHD) is the leading cause of deaths in adults in Europe and North America (WHO, 2017) . Early detection and treatment of this disease is thus a matter of life and death (Gonsalves, Thabtah, Mohammad, & Singh, 2019). This project has compared the predictive power of five machine learning algorithms namely: Support Vector Machine, Naïve Bayes, Logistic Regression, Decision Trees and Neural Networks, in predicting this disease. The objective of this study was to determine which of the five algorithms was best suited for CHD prediction and what level of the CHD risk factors favored the occurrence of CHD. This study had fourteen CHD risk factors that is gender, age, smoking habit, number of cigarettes smoked, use of blood pressure medication, prevalent stroke, prevalent hypertension, diabetes, total cholesterol, diastolic and systolic blood pressure, BMI, heart rate, and education. However, this study found that only age, systolic and diastolic blood pressure, prevalent hypertension, blood pressure medication and diabetes had a significant correlation with CHD occurrence. This study used these seven CHD risk factors to model CHD occurrence in the five algorithms. This study found that the logistic regression was best suited for predicting CHD, followed by Naïve Bayes then Decision Tree and lastly SVM and Neural Networks. This work found that CHD positive individuals had high cholesterol (235mm on average), high blood sugar (a maximum of 394mm), had a smoking habit (10.82 cigarettes per day on average), were obese (overweight BMI of 26.63 on average) and had high blood pressure (a maximum of 295/140 Mm Hg and 143/86 Mm Hg on average).

CHAPTER ONE: INTRODUCTION

1.0 Background to the study

Cardiovascular disease (CVD) is a compound term for all conditions affecting the heart and or blood vessels (NHS, Cardiovascular disease, 2018 and AmericanHeartAssosication, 2017). It is the leading causes of death globally (Ritchie, 2019).

In 2016, approximately 17.9 million deaths were reported (WHO, 2017). This is shown in Figure 1.1. In the United States CVDs have been reported as the leading cause of death since 1990 up to 2017 (Ritchie, 2019). In 2017, an estimated 974,655 people died from CVD in the USA (Ritchie, 2019) . This trend of death is shown in Figure 1.2.

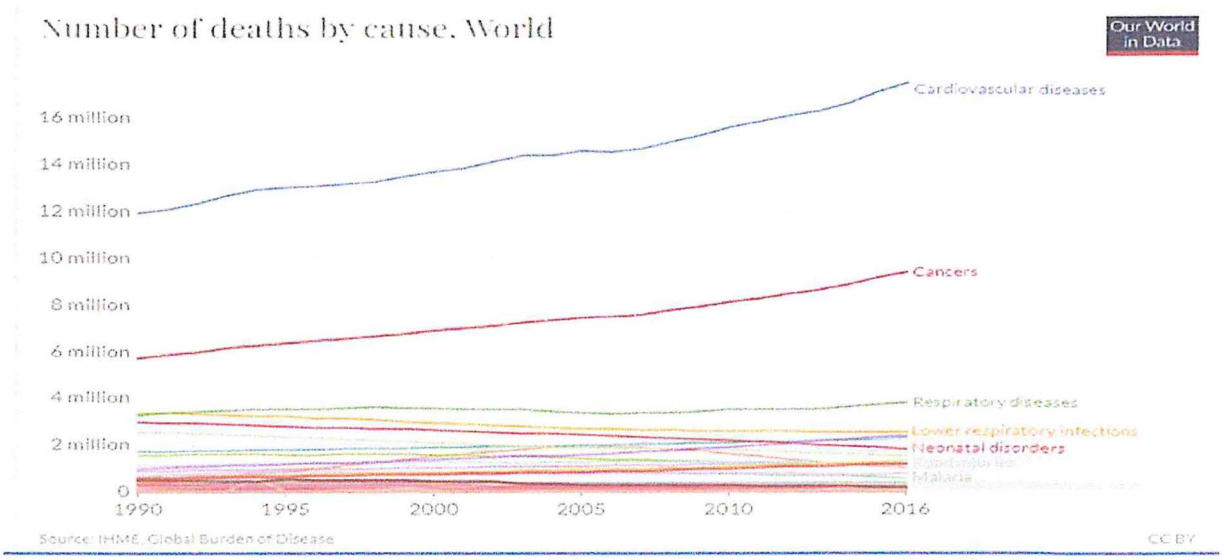

Figure 1.1: Global mortality due to CVDs. Source: (Ritchie, 2019)

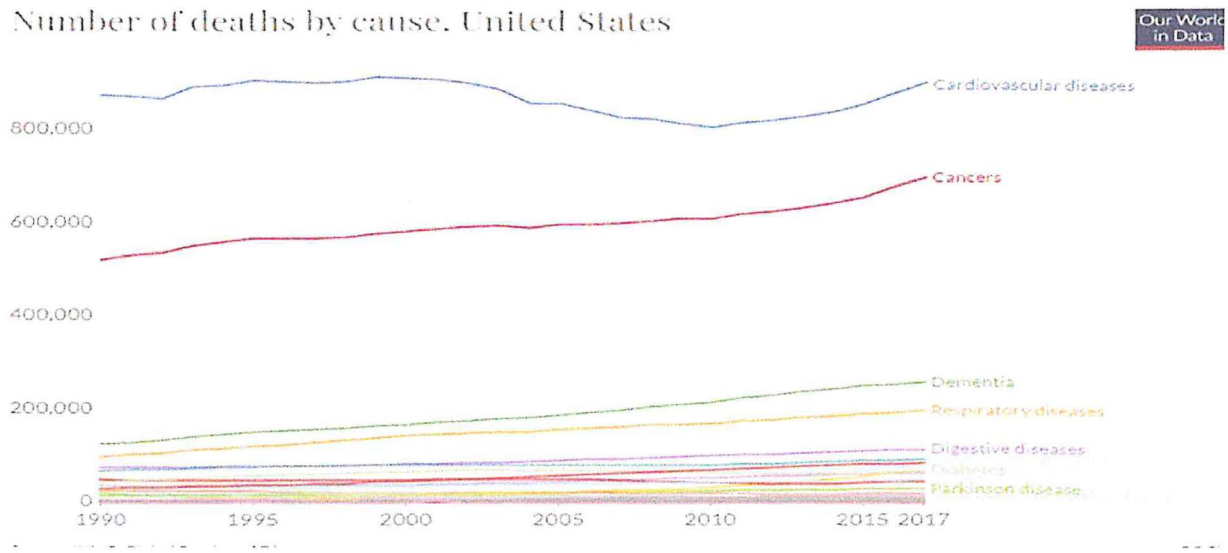

Figure 1.2: Deaths due to CVDs in the USA. Source: (Ritchie, 2019)

1.2 Causes of CVDs.

1.2.1 Behavioral factors

Unhealthy dieting: - Eating unhealthy foods rich in fat, cholesterol and salt is harmful to one`s circulatory system. Excess fat and cholesterol end up forming plaque on the linings of the blood vessels thus narrowing them leading to CVDs (NHS, Cardiovascular disease, 2018).

Physical inactivity: - This is when individuals do not exercise regularly at least 3 to 4 times a week (HealthNY, 1999). Exercising increases the heartbeat rate thus pumps more blood preventing thrombus formation and clearing them from the blood vessels, curbs obesity which causes atherosclerosis by narrowing blood vessels (HarvardPublishing, Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health, 2006).

Tobacco use: - Use of tobacco and or smoking leads to ingestion of harmful substances such as nicotine, carbon monoxide and tar which leads to plaque formation in the blood vessels (atherosclerosis), damaging blood vessels and the heart leading to CVD (NHS, Cardiovascular disease, 2018).

Harmful use of alcohol: - Excessive, long-term abuse of alcohol weakens the heart muscles, distorting its performance in pumping blood efficiently. Due to poor blood flow, one`s organs are deprived of blood disrupting their respective functions (Healthline, Alcoholic Cardiomyopathy: Causes, Symptoms, and Diagnosis, 2017) and (AmericanHeartAssociation, 2014).

Stress: - This may lead one to engaging in habits such as alcoholism, overeating, physical inactivity and smoking which are CVD risk factors (NHS, Cardiovascular disease, 2018). Such habits raise one`s blood pressure and damage their circulatory system leading to CVDs (NHS, Cardiovascular disease, 2018).

Poor hygiene: - The risk of suffering from CVDs significantly rises when one does not implement hygienic habits such as regularly washing one`s hands particularly if one is currently suffering from an underlying heart disease (MayoClinic, Heart disease - Symptoms and causes, 2018). Also, poor dental health may lead to bacterial infections in one`s blood stream, affecting their heart valves and this risk is significantly high for people with artificial valves (MayoClinic, Your teeth and your heart: What's the connection?, 2019).

1.2.2 Other risk factors that cause CVDs.

Age: - As a person ages their body organs become older and weaker and one`s circulatory system is no exception, the blood vessels become damaged and the heart becomes thicker and weaker (MayoClinic, Heart disease - Symptoms and causes, 2018) and (AmericanHeartAssociation, 2014).

Sex: - Women are at lower risk of suffering from CVDs because the estrogen hormones secreted by their bodies before menopause plays a protective role (HarvardPublishing, The heart attack gender gap - Harvard Health, 2016) . For men however they have low levels of estrogen in their bodies and are thus at higher risk of CVDs (MayoClinic, Heart disease - Symptoms and causes, 2018) . However, after menopause women and men have the same risk of suffering from CVD (HarvardPublishing, The heart attack gender gap - Harvard Health, 2016).

Family history: - If CVDs run in one`s family the risk of one suffering from it as well is high (genetics) especially if a parent developed it while young (before 55 years for male relatives, and 65 years for female relatives) (MayoClinic, Heart disease - Symptoms and causes, 2018).

High blood pressure: - Hypertension hardens and thickens one`s arteries thus narrowing the vessels through which blood flows (MayoClinic, Heart disease - Symptoms and causes, 2018).

High blood cholesterol levels: - High blood cholesterol leads to plaque formation in the blood vessels narrowing them (atherosclerosis) causing CVDs (MayoClinic, Heart disease - Symptoms and causes, 2018) and (BritishHeartFoundation, 2010).

Diabetes: - Diabetes raises one`s blood sugar levels to abnormally high levels, which could damage the nerves controlling the heart and blood vessels. Thus, the longer on has been a victim of diabetes the more likely they will develop heart diseases. Therefore, people with diabetes usually develop CVDs at an earlier age than those without diabetes (HealthInformation, 2017).

1.3 Types of CVDs

Coronary heart disease (CHD): - This disease affects the blood vessels that supply the heart and is usually because of accumulation of fatty deposits (atheroma) on the wall linings of the coronary artery. Chances of suffering from this illness are high if one: smokes, suffers from hypertension, has high cholesterol, is obese or overweight, is physically inactive, has a family history of CHD. (NHS, Coronary heart disease, 2020). This is the focus of this research study.

Peripheral arterial disease (PAD) or Peripheral Vascular Disease (PVD): - It is a disease where atheroma in the leg blood vessels in the legs limits the blood flow to the leg muscles. Atheroma narrows the leg arteries restricting blood flow to the legs. The risk of getting this disease increase if one: smokes, is diabetic, hypertensive, has high cholesterol and is of old age. (NHS, Peripheral arterial disease (PAD), 2019).

Cerebrovascular disease: - This is a group of diseases that impair the blood vessels thus causing abnormal blood supply in the brain. Some of these conditions include stroke, aneurysm, transient ischemic attack (TIA) and vascular malformation. These conditions can be caused by atherosclerosis (narrowing of arteries), thrombosis, arterial blood clots (blood clot in an artery of the brain) or by cerebral venous thrombosis (blood clot in a vein of the brain) (MedicalNewsToday, 2019).

Rheumatic heart disease: - This is a disease where the heart valves are permanently damaged. This is due to rheumatic fever which is an inflammatory infection that majorly affect the heart, joints, skin or brain and many other connective tissues. The inflamed and damaged heart valves lead to them narrowing thus impairing the heart functions (JohnHopkinsMedicine, 2009).

Congenital heart diseases (defects): - These are problems with the heart structure at birth. They are a common type of birth defects and happens during pregnancy as the fetal heart develops (Healthline, Congenital Heart Disease: Types, Symptoms, Causes, and Treatment, 2016). These defects can be on the walls and valves of the heart or on even the blood vessels near the heart. These defects could lead to slow blood flow, blood flow in the wrong direction, to the wrong place or blocked blood flow (Healthline, Congenital Heart Disease: Types, Symptoms, Causes, and Treatment, 2016).

Deep Vein Thrombosis (DVT) and Pulmonary Embolism: - DVT happens when a thrombus (blood clot) forms in the deep veins of the body, mostly in the legs causing leg pain or swelling and may also appear with no signs. It is caused by diseases that affect blood clotting, or if one is immobile for long periods of time such as when travelling for long distances or confined to a bed. These clots in the deep blood vessels are dangerous for they can dislodge and be carried into the lungs via the blood stream blocking blood flow to the lungs resulting in a pulmonary embolism (MayoClinic, Deep vein thrombosis - Symptoms and causes, 2019).

1.4 Treatment and prevention measures for the disease

1.4.1 Medical treatment and prevention

Medications: - lifestyle changes alone may not be enough, and one may be forced to take medication to control CVDs. The medication varies depending on the type of CVD one is suffering from (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

Medical procedures (surgery). The damage done to one`s circulatory system by CVDs may be too severe that even medicines and lifestyle changes cannot help. Heart surgeries such as bypasses may be carried out, but the procedure depends on the extent of damage and type of CVD one is suffering from (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

1.4.2 lifestyle changes

Lifestyle changes refers to actions an individual can take on their own to lead a healthier and longer life. These lifestyle changes are both a treatment and prevention measure against CVDs

(MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018) and (HealthAffairs, 2007). They are:

Quitting smoking: - Smoking may result in atherosclerosis, thus a major CVD risk factor. Quitting the habit is the best way to avoid suffering from CVDs and their complications (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). One can quit smoking by:

Going for Smoking addiction anonymous meetings. Here people suffering from addiction to smoking get to share their experiences with each other, they motivate each other to quit the habit and they watch over each other so that none of them relapses into the habit (Nichols, 2017).

Using Nicotine Replacement Therapy. This involves consuming substances that will lower one`s cravings for cigarettes and will reduce the withdrawal effects associated with quitting the habit. Some of these items are chewing gum and lozenges (Nichols, 2017).

Using non nicotine medication. These kinds of medications are aimed at reducing the pleasure one gets from smoking, lowering the cravings and withdrawal symptoms from quitting the habit by targeting the nicotine receptors in the brain. Examples of these drugs are Varenicline and Bupropion (Nichols, 2017).

Control your blood pressure: - Blood pressure may arise due to various factors such as stress and physical inactivity. It is recommended that one should go for hypertension tests annually or biannually (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). Luckily, hypertension can be regulated by exercising, avoiding, and managing stress, eating healthy and taking hypertension medication when prescribed by doctors (AmericanHeartAssosiation, 2017).

Check your cholesterol: - It is advised that one should go for regular cholesterol level tests and try as much as possible to avoid junk foods since they are high in cholesterol. Also, if one`s family has a history of high cholesterol one should go for cholesterol level tests early and avoid foods high cholesterol levels since they are already at a high risk (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

Keep diabetes under control: - When one has diabetes they can take steps to regulate their diabetes to ensure it does not lead to CVDs. A major way of diabetes control is the tight blood sugar control

method. This is a rigorous diabetes self-regulation where one ensures that their blood glucose levels mirror normal levels as possibly close without resulting in hypoglycemia (low blood sugar), with the goal of avoiding the complications from diabetes (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). An example of such a procedure is intensive insulin therapy meant to keep one`s blood glucose levels to match those of a non-diabetic as possibly close. The treatment necessitates high monitoring of the blood sugar levels and multiple insulin doses (MayoClinic, Intensive insulin therapy: Achieving tight blood sugar control, 2020).

Exercise: - Exercise allows one to maintain a healthy weight, avoid and control diabetes, lowers cholesterol levels, and regulates blood pressure, all of which are CVD risk factors. However, exercise intensity should be guided by one`s doctors in case one is suffering from heart defects since due to these defects one may not be able to do certain exercises. One should aim for 30 – 60 minutes of exercise at least thrice a week (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018) and (HarvardPublishing, Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health, 2006) .

Eat healthy foods: - Eating foods low in saturated fats, cholesterol, sodium and added sugars such as fruits, vegetables and whole grains allow one to regulate their weight, blood pressure and cholesterol (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

Maintain a healthy weight: - Being obese and overweight are major risk factors for CVDs. BMI (Body Mass Index) is used as a measure for weight. Thus, one should target having a BMI of below 25 to avoid CVDs (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018) and (HarvardPublishing, Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health, 2006) .

Manage stress: - One should reduce stress and manage it in healthy ways such as muscle relaxation and deep breathing and meditation (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). Avoid "managing" stress by engaging in activities such as alcoholism, overeating and physical inactivity which are major risk factors for CVDs occurrence (AmericanHeartAssociation, 2014).

Deal with depression: - Depression highly raises the chances of one suffering from CVDs. Depression may lead to unhealthy habits such as overeating, alcoholism and physical inactivity which are all risk factors for CVDs. (Holland, 2019). Going for counselling or seeking medical help from one`s doctor helps to avoid and overcome depression (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018) and (Holland, 2019).

Practice good hygiene: - One should avoid people with contagious diseases like colds, be vaccinated against the flu and take up hygienic practices such as brushing and flossing one`s teeth and washing hands (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

According to (HealthAffairs, 2007), in the Unites States they have created drugs that are managing the risk factors of heart disease. Examples of these drugs are statins (These are drugs that are meant to reduce one`s cholesterol level), antihypertensive agents (These are drugs that regulate blood pressure) and thrombolytic agents (These are drugs that dissolve clots in the veins and arteries).

1.5 Coronary heart disease

Coronary heart disease (CHD), or coronary artery disease, occurs when the coronary artery, which supplies oxygen to the heart, becomes too narrow thereby denying the heart enough oxygen which gives its muscle the energy to maintain a heartbeat (Felman, Coronary heart disease: Causes, symptoms, and treatment, 2019) and (familydoctor.orgstaff, 2019).

CHD is currently the leading cause of death worldwide. Approximately 3.8 million and 3.4 million men and women respectively die every year from CHD (Mackay J, 2004). In developed nations heart disease causes the highest number of deaths in men and women (NationalStatistics, 2006). In 2016, 15.5 million individuals who were 20 years and below were reported to have been suffering from CHD in the United States. As of 2017 the number of reported cases rose to about 18.2 million (CDC, 2019).

1.6 Causes of CHD.

1.6.1 Behavioral risk factors

Smoking: - Over time smoking leads to accumulation of plaque in the lining of the coronary artery making it narrower. Therefore, the amount of oxygenated blood which flows from the lungs into

the heart through this artery is reduced leading to deprivation of oxygen of the heart. Direct and indirect smokers are exposed to this risk from this habit (Sullivan & Felman, 2019) .

Overweight or obesity: - When one is obese or overweight, they have excess fat in their bodies and this excess fat ends up accumulating on the walls of the coronary artery thereby making it narrow. The oxygenated blood supply which flows from the lungs to the heart via the coronary artery reduces thus depriving the heart of enough oxygen (HarvardPublishing, Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health, 2006).

Physical inactivity: - This leads to one gaining weight thus becoming overweight and obese, having high cholesterol levels, suffering from diabetes and high blood pressure (HealthNY, 1999) and (HarvardPublishing, Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health, 2006). Therefore, physical inactivity worsens these risk factors for CHD which have been explained above.

High stress: - One may opt to relieve stress by engaging in negative actions such as alcohol abuse, overeating and not exercising which are all risk factors for CHD. Thus, high stress levels worsen the mentioned risk factors, and this leads to damaging of the coronary artery thus limiting oxygen supply to the heart (Healthline, Alcoholic Cardiomyopathy: Causes, Symptoms, and Diagnosis, 2017) and (Holland, 2019).

Unhealthy diet: - Consumption of unhealthy foods that are high in saturated fats, salt, sugar, and cholesterol raise the risk of suffering from CHD. These foods form plaque in the coronary artery thereby making it narrower limiting the supply of blood rich in oxygen to the heart leading to oxygen deprivation (Felman, Coronary heart disease: Causes, symptoms, and treatment, 2019) and (familydoctor.orgstaff, 2019).

Sleep apnea: - This is a condition where one repeatedly stops and start breathing while asleep. Sleep apnea leads to sudden drop in oxygen levels in the blood leading to hypertension and straining of the cardiovascular system possibly resulting in CHD (Felman, Coronary heart disease: Causes, symptoms, and treatment, 2019) and (MayoClinic, Coronary artery disease - Symptoms and causes, 2018).

Alcohol use: - Heavy alcohol consumption leads to hypertension making the coronary artery narrower and harder. The supply of oxygen rich blood to the heart from the lungs via the coronary artery is reduced leading to oxygen deprivation of the heart. Oxygen deprivation damages the heart muscle (Healthline, Alcoholic Cardiomyopathy: Causes, Symptoms, and Diagnosis, 2017) and (AmericanHeartAssociation, 2014).

1.6.2 Other risk factors

Age: - Over the years fatty deposits accumulate on the linings of the coronary artery and the hardening of blood vessels by becoming thicker making it narrower. The heart is therefore deprived of oxygen which it gets from the oxygenated blood which flows from the lungs and into the heart via the coronary artery. (NIH, 2018)

Sex: - Before menopause women have a lower risk of suffering from CHD due to the high levels of estrogen in their system which plays a protective role as compared to the levels in men (HarvardPublishing, The heart attack gender gap - Harvard Health, 2016). Thus, men develop CHD earlier in their lives than women (Fodor, 2004). However, after menopause the risk of CHD developing in both men and women is the same.

Family history: - An individual is at higher risk of suffering from CHD if their family has had a history of this illness. More particularly if an immediate relative got it while young (before 55 years for males and before 65 for females) (MayoClinic, Coronary artery disease - Symptoms and causes, 2018).

Hypertension: - Arteries are thick and narrow since blood flows through them under high pressure. When one is hypertensive their coronary artery thickens even more thus becoming narrower and harder. This makes it transport less oxygenated blood to the heart thereby resulting in oxygen deprivation of the heart (NHS, Coronary heart disease - Causes, 2020).

High blood cholesterol levels: - High cholesterol levels lead to accumulation of plaque on the lining of all blood vessels including the coronary artery, making it narrower and eventually clogging it (atherosclerosis). This limits the amount of oxygen rich blood flowing to the heart from

the lungs in the coronary artery resulting in oxygen deprivation of the heart which may then impair the heartbeat (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

Diabetes: - Diabetes leads to abnormally high levels of blood sugar which damages the nerves that facilitate the functions of the heart and the blood vessels that flow through it. Nerve damage to the coronary artery interferes with the flow of oxygen rich blood to the heart leading to fluctuations of oxygen levels within the heart muscle damaging the heart itself (NHS, Coronary heart disease - Causes, 2020).

Autoimmune diseases: - Inflammatory rheumatologic infections such as lupus and rheumatoid arthritis damage the heart and the blood vessels leading to atherosclerosis in the blood vessels limiting blood supply in the coronary artery thus causing CHD (JohnHopkinsMedicine, 2009).

1.6.3 Signs and symptoms

Chest pain (angina): - This pain is mostly felt on the left or middle side of one`s chest. It is due to oxygen deprivation of the heart and the pain can be stable or unstable (MayoClinic, Coronary artery disease - Symptoms and causes, 2018). Stable chest pains occur during physical activity (exercise or heavy manual work) while unstable chest pains occur randomly (MayoClinic, Coronary artery disease - Symptoms and causes, 2018).

Unstable angina is more dangerous and is characterized by chest pain, tightness, or heaviness and the pain can move to the arms, neck, stomach, back, or jaw (familydoctor.orgstaff, 2019). Stable angina subsides after a few minutes of stopping the physical or emotional stress (MayoClinic, Coronary artery disease - Symptoms and causes, 2018).

Shortness of breath: - All the muscles in the body need oxygen to contract and relax during movement. If your heart cannot supply enough oxygenated blood to them, they generate energy to contract and relax via anaerobic respiration which requires no oxygen. This leads to accumulation of lactic acid, causing a feeling of fatigue, in these muscles and therefore one will pant to breakdown the accumulated lactic acid (familydoctor.orgstaff, 2019).

Cramping: - This occurs when a muscle does not receive enough oxygen during use. If your heart is receiving inadequate oxygen then inadequate oxygen will be supplied to the body muscles

causing accumulation of lactic acid in them thereby causing pain (cramps) (Sullivan & Felman, 2019).

Heart attack: - If the heart muscle is denied oxygen due to a narrow or completely blocked coronary artery the heart is denied oxygen which gives it the energy to beat. Once oxygen deprivation occurs in the heart the heart stops beating, and one ends up experiencing a heart attack. During a heart attack one may experience a crushing pressure in their chest also arm and shoulder pain, difficulty in breathing and sweating (NHS, Coronary heart disease - Symptoms, 2020) and (MayoClinic, Coronary artery disease - Symptoms and causes, 2018).

1.7 Treatment and prevention

CHD has no cure, however there are various treatments available for relieving the symptoms of CHD and there are actions one can take to avoid the disease (HealthDirect, 2020), (Felman, Coronary heart disease: Causes, symptoms, and treatment, 2019) and (NHS, Coronary heart disease - Treatment, 2020).

1.7.1 Medical treatment and prevention

Medicines

The medicines here are aimed at treating and controlling CHD and its risk factors such as hypertension and high cholesterol levels. They are:

low-dose aspirin: - Aspirin is an anticoagulant meaning it makes one`s blood thinner and prevents it from clotting in the blood vessels. By preventing thrombus formation in the blood vessels, it curbs heart attacks and chest pains since blood flows freely into the heart from the lungs. Due to this fact it is not recommended that you take this medicine if you are suffering from bleeding disorders such as hemophilia (HealthDirect, 2020) and (NHS, Coronary heart disease - Treatment, 2020).

Vasodilators (nitrates): - These kinds of medicines dilate (widen) one`s blood vessels facilitating free blood flow and preventing the formation of blockages and dislodges already present blockages in the blood vessels vessels. They come in forms of tablets, sprays, skin patches and ointments. These types of medications are also helpful in treating and management of chest pains

(HealthDirect, 2020) and (Felman, Coronary heart disease: Causes, symptoms, and treatment, 2019).

Statins: - These are drugs that are meant to reduce one`s cholesterol level. By reducing the blood cholesterol levels they curb the formation of plaque in the blood vessels thus allowing free flow of blood (HealthDirect, 2020).

Antihypertensive agents: - These are drugs that regulate blood pressure. Controlled blood pressure prevents the hardening and narrowing of blood vessels which inhibits blood flow to and from the heart (HealthDirect, 2020) and (Felman, Coronary heart disease: Causes, symptoms, and treatment, 2019).

Thrombolytic agents: - These are drugs that dissolve clots in the veins and arteries. In the case of clot formation (thrombosis) drugs such as warfarin can dissolve these clots thereby creating clear blood vessels for blood to flow through (NHS, Coronary heart disease - Treatment, 2020).

Antiplatelet medicines: - Platelets are responsible for the formation of clots which prevent external and internal blood loss through bleeding. However, platelets may end up causing formation of clots in the blood vessels thereby restricting blood flow. Medicines such as clopidogrel prevent platelets from causing blood clots in the blood vessels (Felman, Coronary heart disease: Causes, symptoms, and treatment, 2019).

Surgical procedures

Angioplasty and Stent Implantation: - Angioplasty is a procedure where a tube is inserted into an artery in one`s groin area, and is then moved up into the heart then a balloon is inflated in a narrowed part of the artery improving blood circulation to the heart (NHS, Coronary heart disease - Treatment, 2020) and (HealthDirect, 2020).

After angioplasty is performed, an expandable metal tube (stent) is placed in the artery, expanded, and left in the artery to keep it open (HealthDirect, 2020).

Bypass surgery: - This is a procedure where a blood vessel from another part of the body is grafted onto the coronary artery giving blood a detour route past the narrowed section of the coronary artery. A blood vessel can be drawn from the chest, leg, or forearm. This improves blood

circulation to the heart and reduces the risk of one experiencing a heart attack or chest pains (NHS, Coronary heart disease - Treatment, 2020).

Thrombolytic therapy: - This is where blood thinners are given through a drip to remove a blood clot clogging an artery (Felman, Coronary heart disease: Causes, symptoms, and treatment, 2019).

Implantable Cardiac Defibrator (ICD): - This is a device connected to one`s heart in their chest which monitors and corrects the heartbeat. It stops irregular heartbeats or "restart" the heart by giving it a controlled shock when it is beating either slowly or does not beat at all thus acts as a pacemaker which collects and stores information about one`s heart for the doctor to follow up. (HealthDirect, 2020)

1.7.2 Lifestyle changes

Lifestyle changes refers to actions an individual can take on their own to lead a healthier and longer life. These lifestyle changes are both a treatment and prevention measure against CHD (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). They are:

Quitting smoking: - Smoking may result in atherosclerosis, which is a major CHD risk factor. Quitting the habit is the best way to avoid suffering from CHD and complications associated with it (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). One can quit smoking by:

Going for Smoking addiction anonymous meetings. Here people suffering from addiction to smoking get to share their experiences with each other, they motivate each other to quit the habit and they watch over each other so that none of them relapses into the habit (Nichols, 2017).

Using Nicotine Replacement Therapy. This involves consuming substances that will lower one`s cravings for cigarettes and will reduce the withdrawal effects associated with quitting the habit. Some of these items are chewing gum and lozenges (Nichols, 2017).

Using non nicotine medication. These kinds of medications are aimed at reducing the pleasure one gets from smoking, lowering the cravings and withdrawal symptoms from quitting the habit by targeting the nicotine receptors in the brain. Examples of these drugs are Varenicline and Bupropion (Nichols, 2017).

Control your blood pressure: - Blood pressure may arise due to various factors such as stress and physical inactivity. It is recommended that one should for hypertension tests annually or biannually (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). Luckily, hypertension can be regulated by exercising, avoiding, and managing stress, eating healthy and taking hypertension medication when prescribed by doctors (AmericanHeartAssosiation, 2017).

Check your cholesterol: - It is advised that one should go for regular cholesterol level tests and try as much as possible to avoid junk foods since they are high in cholesterol. Also, if one`s family has a history of high cholesterol one should go for cholesterol level tests early and avoid foods high cholesterol levels since they are already at high risk (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

Keep diabetes under control: - When one has diabetes they can take steps to regulate their diabetes to ensure it does not lead to CHD. A major way of diabetes control is the tight blood sugar control method. This is a rigorous diabetes self- regulation where one ensures that their blood glucose levels mirror normal levels as possibly close without resulting in hypoglycemia (low blood sugar), with the goal of avoiding the complications from diabetes (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). An example of such a procedure is intensive insulin therapy meant to ensure that a diabetic`s blood glucose levels match those of a non-diabetic as possibly close. The treatment necessitates high monitoring of the blood sugar levels and multiple insulin doses (MayoClinic, Intensive insulin therapy: Achieving tight blood sugar control, 2020).

Exercise: - Exercise allows one to maintain a healthy weight, avoid and control diabetes, lowers cholesterol levels, and regulates blood pressure, all of which are CHD risk factors. However, exercise intensity should be guided by one`s doctors in case one is suffering from heart defects since due to these defects one may not be able to do certain exercises. One should aim for 30 – 60 minutes of exercise at least thrice a week (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018) and (HarvardPublishing, Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health, 2006) .

Eat healthy foods: - Eating foods low in saturated fats, cholesterol, sodium and added sugars such as fruits, vegetables and whole grains allow one to regulate their weight, blood pressure and cholesterol (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

Maintain a healthy weight: - Being obese and overweight are major risk factors for CHD. BMI (Body Mass Index) is used as a measure for weight. Thus, one should target having a BMI of below 25 to avoid CHD (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018) and (HarvardPublishing, Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health, 2006) .

Manage stress: - One should reduce stress and manage it in healthy ways such as muscle relaxation and deep breathing and meditation (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018). Avoid "managing" stress by engaging in activities such as alcoholism, overeating and physical inactivity which are major risk factors for CHD occurrence (AmericanHeartAssociation, 2014).

Deal with depression: - Depression highly raises the chances of one suffering from CHD. Depression may lead to unhealthy habits such as overeating, alcoholism and physical inactivity which are all risk factors for CHD (Holland, 2019). Going for counselling or seeking medical help from one`s doctor is advised to avoid or get through depression (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018) and (Holland, 2019).

Practice good hygiene: - One should avoid people with contagious diseases like colds, be vaccinated against the flu and take up hygienic practices such as brushing and flossing one`s teeth and washing hands (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018).

1.8 Statement of the problem

Coronary heart disease has been a plague in the United States for a long time. Researchers such as (Shirley, et al., 1990) and (Tavia , Castelli, Hjortland, Kannel, & Dawber, 1977) have come up with logistic algorithms that predict coronary heart disease. Others like (Wilson, Kannel, & Agostino, 1998) have proved that managing CHD risk factors aids in regressing the devastating effects of CHD on the human body and even avoiding it. Accurately predicting and testing for this disease is paramount since early detection facilitates the early creation of a treatment regimen to either regress the effects of CHD if one has it or transforming one`s current lifestyle to a healthy one if you are at risk of getting it.

Various algorithms have been proposed when it comes to prediction of CHD. These algorithms are: Decision Tree – C4.5 (J48),Naïve Bayes Algorithm and Support Vector Machine (SVM), Logistic regression and Neural Networks as proposed in the works by (Dangare & Sulabha , 2012), (Tavia , Castelli, Hjortland, Kannel, & Dawber, Predicting Coronary Heart Disease in Middle-Aged and Older Persons, 1977) and (Gonsalves, Thabtah, Mohammad, & Singh, 2019).

On measuring the accuracy sensitivity and specificity of these three models, (Gonsalves, Thabtah, Mohammad, & Singh, 2019) concluded that the Naïve Bayes algorithm was a good fit when it came to modelling CHD predictions and (Dangare & Sulabha , 2012) in their study concluded that Neural Networks outperformed Decision Trees and Naïve Bayes.

However, these algorithms have not been modelled together and then compared to each other when it comes to their ability to predict CHD occurrence. This study has added the Logistic regression as its fifth algorithm to compare to the four already mentioned.

This study compares the ability of these five models in estimating CHD occurrence in the United States and determine which amongst them is the best.

1.9 Objectives of the study

1.9.0 General Objective

The general aim of this study is to determine how a better and healthy lifestyle can help one avoid suffering from CHD.

1.9.1 Specific Objectives

1. To establish the risky lifestyles that lead to CHD in the United States of America
2. To determine the best machine learning algorithm for predicting occurrence of CHD in individuals.

1.10 Scope of the study

This study attempts to examine the influence of one`s current lifestyle on the risk of them contracting CHD in the future. This is because CHD is the leading cause of death in the world and more specifically in the United States. This study will impact the health agenda of increasing quality and years of healthy life and to eliminate health disparities in the USA by providing the best predictive tool that will allow individuals to take actions towards leading a healthier lifestyle to avoid contracting CHD. This study has applied a quantitative methodology by using cardiovascular health data from Kaggle to compare the prediction power of five different machine

learning algorithms in determining the occurrence of CHD for an individual based on their current lifestyle.

1.11 Significance of the study

The results generated by this study will have applications in the field of insurance. Insurance companies issue out health policies for all manner of diseases. The best model will allow insurers who are offering health covers for CHD to be able to set premiums for insured individuals based on the risk of them suffering from the disease in future and be prepared for the number of claims they will be faced with in the future when this risk occurs.

The best model here will also allow doctors to come up with lifestyle changes regiments for patients who are at risk of contracting CHD as a way of allowing the patients to take control of their health before even suffering from the disease.

The study here will allow governments to plan and approximate the expenditure for CHD diagnosis and treatment since with the model generated here, they will be able to know how many they expect to contract the disease and the various demographic and geographic aspects of the disease.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Timely CHD prediction has always been a matter of utter importance. Early detection is essential since it gives individuals a chance to start CHD regression treatments or lead healthier lives to avoid suffering from CHD. This chapter will focus on the research works previously done in prediction and prevention of CHD. It entails the methodology used in these studies, the results obtained and the gaps arising from the studies and how this present work intends to bridge these gaps.

2.2 Past Empirical studies CHD prediction

Various researchers such as (Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977), (Wilson, Kannel, & Agostino, 1998), (Dangare & Sulabha , 2012), (Gonsalves, Thabtah, Mohammad, & Singh, 2019), (Palaniappan & Awang, 2008), (Shirley, et al., 1990) and (X, J, Z, & Y, 2007) have conducted studies on prediction of CHD in individuals based on the lifestyle decisions

Traditionally CHD prediction was done by use of the logistic regressions as per the works of (Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977), (Wilson, Kannel, & Agostino, 1998) and (Shirley, et al., 1990). However, the logistic regression algorithm is not the only machine learning algorithm that can be used in the prediction of CHD and the emergence of data mining techniques have led to higher accuracy and timeliness in the prediction of CHD. Researchers such as (Dangare & Sulabha , 2012), (Palaniappan & Awang, 2008) and (Gonsalves, Thabtah, Mohammad, & Singh, 2019) have done works on CHD prediction by use of other algorithms such as: Naïve Bayes, Neural Networks, Decision Trees, and the Support Vector Machine algorithms.

The works by the above researchers were aimed at picking the best model for CHD prediction given the various available algorithms prior mentioned and based on the model that best fit their data. For (Palaniappan & Awang, 2008) and (Dangare & Sulabha , 2012), they were comparing the predictive power for CHD of the Naïve Bayes, Neural Networks and Decision Trees. However, for (Gonsalves, Thabtah, Mohammad, & Singh, 2019), they were comparing the predictive power for CHD of the Naïve Bayes, Decision Trees and the Support Vector Machine algorithms.

(Palaniappan & Awang, 2008) tested and compared the predictive power of the Naïve Bayes, Neural Networks and Decision Trees by fitting them onto the thirteen independent variables

(predictors) that were found to be CHD risk factors. These independent variables were: Gender (1 for male, 0 for female), chest pain type, Fasting Blood Sugar, resting electrographic results, exercise induced angina, slope of the peak exercise ST segment, the number of major vessels colored by fluoroscopy, Trest Blood Pressure, Serum Cholesterol, maximum heart rate achieved, ST depression induced by exercise relative to rest, defect type and Age in years. (Dangare & Sulabha , 2012) just added two more predictive variables to his study which were not covered by (Palaniappan & Awang, 2008) which were: obesity and smoking.

(Palaniappan & Awang, 2008) fitted the models on their dataset which they got from the Cleveland Heart Disease Database having 909 records containing 15 variables and tested the effectiveness of the three models by use of a confusion matrix and a lift chart to determine which of the three models was best for predicting individuals with CHD. From the confusion or classification matrix he found that the Naïve Bayes algorithm had the highest correct predictions for patients with CHD (86.53%) followed by neural networks with (85.80%) and finally decision trees. But the decision tree was best for predicting patients without CHD (89%) as compared to Naïve Bayes and Neural Networks.

In the life chart approach, one was done with a predictable value while the other was done without. The life chart with the predictable value tests a model's prediction accuracy by comparing it to predictions made by an ideal (perfect) model and random guess. By random guess this means a model that randomly guesses if one has CHD or not, statistically it should get at least half of these predictions right. By ideal this means a model that makes 100% accurate predictions. Therefore, for a model to be effective its predictions must lie between the predictions made by the ideal model and random guesses.

For the life chart without predictable value lacks the random guesses given that there is nothing to predict. In this case the ideal model will make 50% correct predictions if 50% of the data is used to test the model. In this case neural networks would make the highest correct predictions followed by Naïve Bayes and finally decision trees. Overall Naïve Bayes and Neural Networks do better in predicting CHD than Decision Trees.

Palaniappan & Awang, 2008 also measured their model's predictive powers by setting five major goals that these models would have meet for them to be deemed effective. They were: could the models predict patients most likely to suffer from CHD given their medical history? Could the models identify the relationship between the predictors and the dependent variable? Could they identify the effects the predictors had on the dependent variables? Could the models identify the traits of patients with CHD? Finally, if the models could distinguish factors favoring or disfavoring CHD patients and non-CHD patients?

As per the first goal all the models could accurately predict CHD with Naïve Bayes being the most accurate (95%) followed by Decision Trees (94.93%) and finally Neural Networks (93.54%). For the second goal Naïve Bayes and Decision Trees were able to show which were the most significant heart disease predictors such as chest pain type, thal, CA and exhang to those that were less significant such as fasting blood sugar as per Naïve Bayes and Trest blood pressure.
For the third goal only the decision Tree was able to capture the impact the predictors had on the target variable that is it gave an approximate probability of suffering from CHD of 99% if an individual had chest pain type 4, CA and exhang of 0 and a Trest blood pressure ranging between 146 and 159.

For the fourth goal only, Naïve Bayes could identify the characteristics of CHD patients such as 80% of them were males where about 40% of them were aged between 56 and 63 years. Finally, for the fifth goal both Neural Networks and Naïve Bayes could capture the predictor values that either favored or disfavored patients who were positive for CHD and those who were not. Neural Networks found that the factor favoring prediction of CHD present was old peak =3.81, CA=2 while those favoring CHD absent were cholesterol levels greater than or equal to 382.37 and CA=0. Naïve Bayes found that the factor favoring prediction of CHD present was chest pain type 4, thal = 7, exhang =1 and slope =2, while thal = 3, exhang = 0, CA=0 favored prediction of CHD absent in the individuals

The work by (Palaniappan & Awang, 2008) concluded that the three models were able to identify patterns between the predictors and the dependent variable and that the best model for heart disease prediction was Naïve Bayes followed by Neural Networks and finally Decision Tree. The

limitation of this study was that more variables could be introduced into the model and this was addressed by the work of (Dangare & Sulabha , 2012) who added two more predictors to the models which were smoking and obesity.

For (Dangare & Sulabha , 2012) and (X, J, Z, & Y, 2007) noted that the risk of suffering from CHD rose due to risk factors such as: Family history, Smoking, Poor diet, High blood pressure, High blood cholesterol, Obesity, Physical inactivity, Hypertension.
(Dangare & Sulabha , 2012) compared the predictive powers of Naïve Bayes, Neural Networks and Decision Trees via the same methodology by (Palaniappan & Awang, 2008) but the difference between the two studies was that (Dangare & Sulabha , 2012) addressed the limitation in the work of (Palaniappan & Awang, 2008) by adding two more predictors which were smoking and obesity.

However, (Dangare & Sulabha , 2012) encountered a challenge of missing values in his data and addressed them by replacing them with the mean mode method. Another difference between the two studies was that (Dangare & Sulabha , 2012) did not set any goals for his model to achieve to be deemed effective. A confusion matrix was used to test the three model's effectiveness and accuracy in predicting CHD. The confusion matrix compared the three model's accuracy and efficiency when thirteen predictors were used and when 15 predictors were used.

(Dangare & Sulabha , 2012) found that Neural Networks did a better job in predicting CHD occurrence when using both 13 and 15 predictors followed by Decision Trees and finally Naïve Bayes.

(Gonsalves, Thabtah, Mohammad, & Singh, 2019) first compared the predictive capability of three models which are: The Naïve Bayes Approach (NB), Support Vector Machine (SVM) and the Decision Tree (DT). Their research was aimed at finding the model that had the best accuracy, specificity, and sensitivity with respect to their CHD dataset sample which the obtained from CHD patients in South Africa.

By model classification accuracy we mean the ratio of the number of correct predictions to the total number of input samples. For example, if the model training sample contains 1000 CHD observations and the model predicts 800 of them correctly then the model accuracy is 0.8. now

applying the same model on the test sample data in which CHD observations are now 600 and non-CHD cases are 400, then test accuracy drops down to 60% (Mishra, 2018).

Therefore, high classification accuracy gives a false sense of achieving high accuracy in your predictions. Assuming the CHD observations were now 400 in the sample the rate of misclassification of a CHD occurrence to a non-CHD observation would be very high. Since CHD is a fatal disease the cost of misdiagnosing a CHD individual as a healthy individual would be more compared to sending a healthy individual to take more CHD tests (Mishra, 2018).

When it comes to prediction of CHD there are four categories of results one gets. There are the false positives, these are individuals the model predicts have CHD, but they do not. false negatives, these are individuals who have CHD, but the model deems them CHD free. True positives, these are individuals who the model predicts that they have CHD and they have it (Martin, 2020).

Finally, we have the true negatives, these are individuals who the model predict they do not have CHD and they do not have it. Now model specificity is the proportion of observed positives that the model predicted to be positive, out of the total number of CHD positive patients in our sample, how many of them did the model predict them to be positive. Model sensitivity is now the proportion of observed negatives that the model predicted to be negative, out of the CHD free individuals in the sample data how many of them did the model predict them to be CHD free (Martin, 2020).

The dependent variables used in the prediction of CHD were systolic blood pressure, volumes of tobacco consumed by their patients in Kgs, bad cholesterol levels (LDL), adiposity (percentage of one`s body fat), obesity (BMI), family history of CHD, type A behavior (competitive, impatient, and angry character individuals), alcohol consumption and age (Gonsalves, Thabtah, Mohammad, & Singh, 2019) and (Shirley, et al., 1990)

They fragmented their sample dataset into test and training samples, the training sample meant to derive and train the model on CHD prediction given the dependent factors and the test data for testing if the predictions made by the model were accurate with respect to the data sample provided.

(Gonsalves, Thabtah, Mohammad, & Singh, 2019) concluded that the Naïve Bayes algorithm had superior accuracy in predicting CHD as compared to DT and SVM. DT and SVM had similar

levels of sensitivity which was about 49% which was quite low since it generated many false negatives (people the model estimated they had no CHD, yet they did), the false negative rate was 50 percent from the study conducted. It was however noted that SVM and DT were the best when it came to predicting individuals who did not have CHD (Gonsalves, Thabtah, Mohammad, & Singh, 2019)

When it came to the sensitivity of the NB approach, it exhibited low levels of false negative predictions (individuals who the model predicted they did not have CHD, yet they did). It therefore had the highest sensitivity of 63 percent compared to SVM and DT. The model was deemed best for predicting the presence of CHD, but it also came with a high false positive level of prediction (estimated individuals who were CHD free as those who had CHD) (Gonsalves, Thabtah, Mohammad, & Singh, 2019).

The findings of (Gonsalves, Thabtah, Mohammad, & Singh, 2019) found that even though DT and SVM had the best specificity of 82 percent when it came to modelling CHD predictions, they had very low sensitivity levels of less than 50 percent. It was therefore recommended that the NB approach was best for predicting CHD due to its high accuracy and above average specificity and sensitivity measures.

Before other types of regression techniques were available, predicting CHD was only done by use of the logistic regression technique (Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977). Therefore, the study done by (Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977) was hinged on using the logistic regression algorithm in predicting CHD.

(Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977) got their data sample for their study from the town of Framingham of individuals aged between 49 to 82 years at the time of their eleventh biennial examination for CHD by the National Heart, Lung, and Blood Institute. The CHD risk factors used in this study were: systolic blood pressure, left ventricular hypertrophy (LVH), body mass index and if an individual was suffering from diabetes at the time.

Cigarette smoking was removed from this list of variables since it is unrelated to CHD after the age of 65 years. Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977 defined CHD as occurrence of myocardial infraction, angina pectoris, coronary insufficiency, or death due to CHD.

(Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977) concluded that the CHD risk factors used in this study for predicting CHD were the same for both the middle aged and younger generations. However, the effect of the risk factor's contribution to occurrence of CHD was not the same in the older generations as compared to the younger generations. A clear illustration of this was that risk factors such as total cholesterol was not a major CHD risk factor in the older generations, but it was a major CHD risk factor for younger generation. Also smoking in younger generations increased the chances of one suffering from CHD at a higher rate for the youths as compared to the middle aged and older generations (Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977).
It was observed that diabetes as a CHD risk factor was sex specific in the middle aged and older generations. This was because diabetes was a higher CHD risk factor in women but had no relation to CHD presence in men.

(Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977) concluded that with appropriate CHD risk factor variables CHD prediction could be accurately done to distinguish individuals who are and are not at risk of suffering from CHD.

A similarity in the prediction of CHD by (Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977) and (Wilson, et al., 1998) is that both used the logistic regression technique in predicting CHD and also they drew their research data samples from the data collected in the Framingham study of 1971 to 1974.

For (Wilson, et al., 1998) he created sex specific regressions for CHD prediction based on age, diabetes, smoking, total cholesterol (TC), LDL cholesterol, HDL cholesterol and obesity which was measured by use of BMI. One was classified as a smoker if they had been smoking 12 months prior to the study.

(Wilson, et al., 1998) categorized blood pressure (BP) observations into categories: optimal BP (systolic < 120mm Hg, diastolic < 80mm Hg), normal BP (systolic 120-129mm Hg, diastolic 80-84mm Hg), High normal BP (systolic 130-139mm Hg, diastolic 85-89mm Hg), Hypertension stage 1 (systolic 140-159mm Hg, diastolic 90-99mm Hg), Hypertension stage 2 – 4 (systolic ≥ 160 mm Hg, diastolic ≥ 100 mm Hg). This was done without considering if an individual was on hypertension medication or not. People who had BP at levels 2,3,4 were so few that they had to cluster them together into one category.

TC and LDL cholesterol (LDL-C) were also categorized into levels: <200 mm, (200-239) mm, (240-279) mm and ≥280 mm for total cholesterol while for LDL: <130mm, (130-159) mm, ≥280mm. the test subjects were then followed up for 12 years for signs of CHD as in the studies by (Wilson, et al., 1998)

The findings from (Wilson, et al., 1998) concluded that relative risk but not attributable risk of TC on CHD occurrence declined with age. HDL-C levels were associated with CHD occurrences in older generations. It was also observed that approximately half of the sample population had normal or optimal BP. Individuals who suffered from hypertension, had diabetes, a high BMI and TC levels were observed in the category of high BP individuals. The association between smoking and high blood pressure levels was more prevalent in men than in women.

In women it was observed that there was more association among HDL-C, LDL-C and BP levels but this was not so for men. Due to the above associations between the variable CHD occurrence in the sample data and certain BP, TC, HDL-C and LDL-C categories. Many CHD occurrences were observed in individuals classified under the stage 1 hypertension (Wilson, et al., 1998).

When it came to LDL-C and TC levels, CHD occurrences were high in individuals who had LDL-C levels ≥ 160mm and TC levels ≥240. Relative risk of CHD rose with respect to higher levels of BP and TC categories. Low levels of HDL were associated with high risk of CHD occurrence while high levels of HDL-C were associated with low chances of CHD occurrence. On average the attributable risk of CHD with respect to BP levels above normal was 28 percent for men and 29 percent for women (Wilson, et al., 1998).

Attributable risk of CHD with respect to TC levels > 200 on average was 27% in men and 34% in women and when it came to high LDL-C levels the occurrence of CHD rose with it. Higher levels of BP coupled with abnormal cholesterol levels and elevated BMI (obesity levels) was associated with high occurrence of diabetes.

(Wilson, et al., 1998) found that BP, TC, LDL-C and HDL-C are major CHD prediction risk factors and therefore risk factor prevention mechanisms such as healthy living should be incorporated.

This study differed from (Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977) in that left ventricular hypertrophy risk factor was not included since it lacked a universally accepted ECG

(electrocardiogram) criteria at the time this study was done and exercise would have been appropriate in their study, but it also could not be included in this study since data about it was not present. Therefore, high levels of BMI, TC, BP, low HDL-C levels, and diabetes increases one`s chances of suffering from CHD.

A related study by (Shirley, et al., 1990) dove into the possible ways of managing and avoiding CHD. Shirley, et al., 1990 empirically showed that with comprehensive lifestyle changes for the better can lead to the regression of even severe CHD in patients after one year of implementation without the use of lipid lowering drugs. They did this by splitting their volunteer CHD patients into two groups that is the treatment (22 participants) and control groups (19 participants). For the treatment group they were advised to make comprehensive lifestyle changes such as eating healthy, exercising moderately, managing their stress and quitting smoking and alcohol consumption. For the control group the comprehensive lifestyle changes were optional for them. Coronary angiography tests were administered to both groups before the research experiment began to establish baseline characteristics. The tests were then done after six months and one year after the research experiment began (Shirley, et al., 1990).

These baseline characteristics were total cholesterol, HDL cholesterol, apolipoproteins (A1 and B) and triglyceride concentration. The selection of patients for their study was done on the San Francisco population on patients who had to meet conditions such as: had no life-threatening diseases, had not suffered a myocardial infraction in the preceding six weeks, had no or scheduled heart bypass surgery and was not on lipid lowering medication.

In the treatment group adherence to the comprehensive lifestyle change was graded according to how one observed the lifestyle changes advised. A score of zero was awarded to poor adherers, one was awarded to good adherers and those who did more than what was set were awarded adherence score higher than one (Shirley, et al., 1990).

On comparing the baseline characteristics of the control and treatment groups there was no significant difference between them. After one year however the deviations between them manifested (Shirley, et al., 1990).

After one year it was observed that the members of the control group showed regression of diameter stenosis while for the control group diameter stenosis progressed. Diameter stenosis is

regarded as the narrowing of blood vessels (reduction in luminal diameter), narrowing of blood vessels by more than 50 percent is of high concern since it limits the volume of blood flowing in the respective blood vessel (Harris, et al., 1980).

In the treatment group CHD regression was observed in eighteen of its members, three showed slight CHD progression while one showed significant regression due to poor adherence. CHD progression was however observed in the treatment group by ten of its members, slight regression was observed in eight of its members four of whom were women, and one showed no change. It was therefore concluded by (Shirley, et al., 1990) that regression of CHD was dependent on an individual's adherence to living a healthier lifestyle.

2.3 Past empirical studies on CHD risk lifestyles

Even though the methodologies used by (Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977), (Wilson, Kannel, & Agostino, 1998), (Dangare & Sulabha , 2012), (Gonsalves, Thabtah, Mohammad, & Singh, 2019), (Palaniappan & Awang, 2008), (Shirley, et al., 1990) and (X, J, Z, & Y, 2007) were different, they were all aimed at predicting CHD while also pointing out the risk factor levels that favored the occurrence of CHD in their respective samples.

Key risk factors such as high cholesterol, high blood sugar, smoking habit, obesity, hypertension, heart defects, heart rate and chest pain suffered by a patient were both risk factors and early signs for the onset of CHD.

Shirley, et al., 1990 recommended strict adherance to a comprehensive lifestyle change to combat CHD backed by the results of their study which showed adoption of better lifestyle choices such as exercising, proper dieting, quiting smoking and drinking and managing stress led to regression of CHD. If these lifestyle choices could regress the damaging effects of CHD, then they should be adequate to avoid CHD occurrence.

Dangare & Sulabha , 2012, Palaniappan & Awang, 2008 and Wilson, Kannel, & Agostino, 1998 in their works sited that high cholesterol levels, high blood pressure, being obese and diabetes significantly rose the chances of an individual of suffering from CHD.

Wilson, Kannel, & Agostino, 1998 found that there was a higher association between high blood pressure and smoking habits in men. Given this high relationship, men are adviced not to smoke since it would lead to hypertension leading to CHD.

Tavia, Castelli, Hjortland, Kannel, & Dawber, 1977 found that middle aged and older women were at a higher risk of diabetes which would then raise the risk of them suffering from CHD. Women in their middle ages and above should therefore go for regular diabetes tests and diet properly to avoid it or detect early enough for fast treatment. Delayed diabetes treatment would result in CHD occurrence.

For (Dangare & Sulabha , 2012) and (Palaniappan & Awang, 2008) they were able to distinguish the risk factors that significantly favored the occurrence of CHD. They were cholesterol levels of between 318 to 383, level 4 chest pains and individuals with reversible heart defects. Therefore, one should aim for cholesterol levels below 318, seek medical attention when suffering from severe chest pains and take medication or go for surgical procedures to correct reversible heart defects CHD being one of them also.

In conclusion, to avoid CHD, one should lead a better and healthier lifestyle. This necessitates exercising, eating better, managing one`s blood sugar, avoiding and managing stress, and quitting harmful habits such as smoking and drinking alcohol.

2.4 Conceptual Framework

This study will predict CHD occurrence in individuals by use of 14 predictors which are classified into four groups that is: demographic variables, behavioral variables, historical and current health variables. The variables contained in the Figure 2.1 are further discussed in the methodology section.

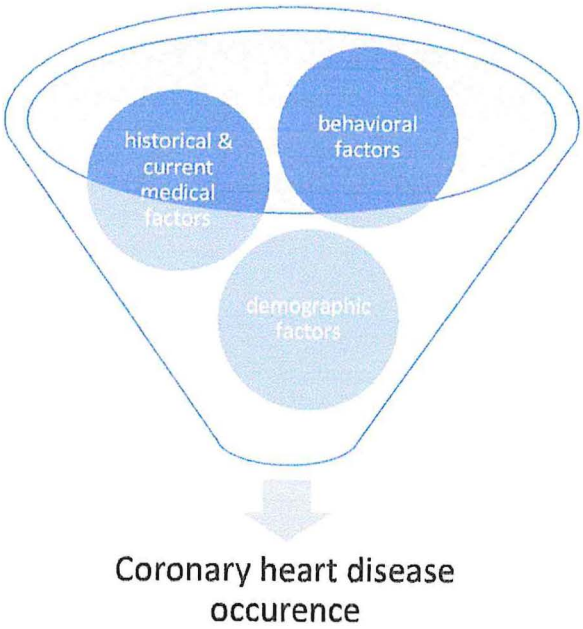

Figure 2.1: Conceptual Framework

CHAPTER THREE: METHODOLOGY

3.1 Introduction

This study has compared the predictive power of SVM, Naïve Bayes, Neural Networks, Decision Trees and Logistic Regression in predicting the chance that an individual may end up suffering from CHD in 10 years' time given their current lifestyle choices.

3.2 Research design

The research questions this study has answered is which risky lifestyle decisions increase one's risk of contracting CHD also the chance of that individual of suffering from CHD given their current lifestyle. The five algorithms have been used to classify individuals as CHD positive or negative. The summary statistics of CHD positive individuals have been used to determine what factors promote the occurrence of CHD.

3.3 Population and sampling techniques

3.3.1 Population

The Framingham heart study is the longest running CHD group study in USA currently 70 years old which is still being conducted in the town of Framinghton in Massachusetts (Levy, Benjamin, Johnson, Andersson, & Ramachandran , 2019). The population captured in this dataset over the 7 decades consists of approximately 15,000 observations (Levy, Benjamin, Johnson, Andersson, & Ramachandran , 2019). This data has been collected from the 1940s, 1970s, 1990s and currently in the 21st century (2000s) (NHBL & NIH, 2019).

3.3.2 Sample size

The data sample observations are the medical records of more than 4000 individuals drawn from the 15000-population sized dataset in the Framingham area in Massachusetts. The sample size for any population is calculated in 2 main steps. The first one is to get the sample size for an infinite population and secondly the sample size is adjusted to the required population. To get the sample size for this study this formula is used ($s=[z^2*p(1-p)]/m^2$) as proposed by (Cochran, 1977). Where s is for the sample size, z^2 is for the z value for the confidence interval (CI) used, p is for the population proportion (assumed to be 50%), and m^2 is for the margin of error. For an infinite population at a 95 % CI (recommended level of CI) the sample size would be 384.16 observations.

Thus, adjusting for the population of approximately 15000 the desired sample size as per the formula s=s/(1+(s-1)/population would be 375 observations. Also, using the Yamane sampling equation formula of ($n= N/1+Ne^2$) where n is the sample size, e the margin of error and N the known population sample, the desired sample size would be 390 observations (Yamane, 1967).

Thus, this sample of 4239 observations is convenient to use since the estimated models have plenty of data to learn from. Also, with the many observations available the models were able to explore and establish stronger relationships between the variables thus raising the model`s predictive power.

3.3.3 Sampling techniques

The observations in the data sample are randomly sampled from the Framingham population dataset. It is random since the probability of an individual of being observed with CHD or without is equal. It is therefore random since participants randomly went for the test. No bias was introduced by sampling a particular age, gender, or any other category. The sampling method was thus random. In the choice of sample size, the recommendations of (Cochran, 1977) and (Yamane, 1967) were to be applied but the sample size of 4239 observations has met their recommendations.

3.4 Data collection

The data for the study is available and was got from Kaggle. The data however is drawn from the Framingham population dataset which was collected in the small town of Framingham in Massachusetts.

3.5 Data analysis

The data for this study was sourced from the Kaggle data website and has 4239 observations. However, the data has missing values and to deal with them, this study dropped them to avoid introducing bias into the data.

Table 3.1: Operationalization of variables

(I)Independent variables	
(A)Demographic variables	
Sex	As one ages, all their organs age and the blood vessels and the heart are no exception. With older age the heart becomes weaker and blood vessels become thick and stiff due to lifetime plaque accumulation on the walls of the vessels. This raises blood pressure (CHD risk factor in itself), which may lead to a heart attack (NIH, 2018) Before menopause women`s estrogen levels are higher than those of men. Estrogen plays a protective role against CHD and thus women are less likely to develop CHD at early age as compared to men. However, after menopause the levels of estrogen in women reduces and thus the risk of a woman or a man suffering from CHD is the same. (HarvardPublishing, The heart attack gender gap - Harvard Health, 2016)
Age	
(B)Behavioral variables	
Current Smoker	Smoking lines one`s arteries with plaque narrowing the blood vessels to and from the heart, making the heart work even harder. This raises one`s blood pressure, causes an irregular heart rhythm and restricts oxygen flow to the heart via the coronary artery thus causing CHD and strokes (Sullivan & Felman, 2019).
Cigarettes Per Day	
(C)Medical variables (historical)	
Blood pressure medication	Hypertension medication are one of the best management measures for high blood pressure. They also reduce the risk of heart attacks, strokes, and CHD by preventing the rapture of blood vessels (NHS, Coronary heart disease - Causes, 2020).

Prevalent Stroke	If an individual has suffered from a stroke in the past chances are high that they will suffer from coronary heart disease and future stroke incidents
Prevalent Hypertension	Hypertension damages the blood vessels leading to impaired blood flow to the heart resulting in various CVDs such as CHD, stroke, heart failure, atrial fibrillation, and peripheral vascular disease. (NHS, Coronary heart disease - Causes, 2020)
Diabetes	Diabetes leads to abnormally high levels of blood sugar which damages the nerves that facilitate the functions of the heart and the blood vessels that flow through it. Nerve damage to the coronary artery interferes with the flow of oxygen rich blood to the heart leading to fluctuations of oxygen levels within the heart muscle damaging the heart itself (NHS, Coronary heart disease - Causes, 2020).
(D)Medical variables (current)	
Total Cholesterol	The body needs cholesterol to build healthy cells, but high levels of cholesterol can increase the risk of heart disease. High cholesterol, forms plaque in the blood vessels, narrowing them and limiting blood circulation and deprivation of oxygen and nutrients of vital organs such as the heart leading to CHD (MayoClinic, Heart disease - Symptoms and causes, 2018) and (BritishHeartFoundation, 2010)
Systolic Blood Pressure	This is the amount of pressure in the arteries during the contraction of the heart. CHD and all other CVDs occur in individuals with high systolic blood pressure and may result in death (NHS, Coronary heart disease - Causes, 2020).
Diastolic Blood Pressure	This is quantity of pressure in the arteries during the relaxion of the heart (resting between heart beats). High diastolic and high systolic blood pressure damages the heart and the blood vessels around it leading to occurrence of various CVDs such as CHD, stroke, and heart attacks.

	(NHS, Coronary heart disease - Causes, 2020)
BMI	Being obese and overweight are major risk factors for CVDs. BMI (Body Mass Index) is used as a measure for weight. Thus, one should target having a BMI of below 25 to avoid CVDs (MayoClinic, Heart disease - Diagnosis and treatment - Mayo Clinic, 2018) and (HarvardPublishing, Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health, 2006) .
Heart Rate	An elevated heart rate increases causes cardiovascular disease and sudden death in patients with known or suspected coronary heart disease.
Glucose	High levels of blood sugar damage the nerves that facilitate the functions of the heart and the blood vessels that flow through it. Nerve damage to the coronary artery interferes with the flow of oxygen rich blood to the heart leading to fluctuations of oxygen levels within the heart muscle damaging the heart itself (NHS, Coronary heart disease - Causes, 2020).
(II)Dependent variable	
10-year risk of coronary heart disease	CHD occurrence in an individual.

This study has visualized the data by use of graphs, scatter plots and correlation heatmaps to show the relationship amongst the predictors themselves and their relationship between them and the dependent variable.

This study has modelled CHD occurrence using the five algorithms based on the predictors deemed significant by the correlation heatmap. The summary statistics of CHD positive individuals has been used to determine the risky lifestyles favoring CHD occurrence.

3.6 Model building

This study has compared the predictive power of the Support Vector Machine, Neural Networks, Naïve Bayes, Decision Tree and Logistic regressions in predicting CHD occurrence in the USA. The model's performance was measured by accuracy, specificity, and sensitivity. The Specificity measure measures how many times the models predict and individual with CHD correctly. Sensitivity measures how many times the models predict an individual without CHD correctly. Accuracy gives an overall measure of how many times a model is correct in classifying an individual with or without CHD.

This study has estimated the five algorithms and compared their powers in predicting CHD occurrence. The variables used in these models were: age, diabetes, cigarette smoking, total cholesterol and blood pressure, prevalent hypertension, heart rate, glucose blood level, prevalent stroke, and blood pressure medication.

This study has replicated the methodology by (Wilson, Kannel, & Agostino, 1998) used in blood pressure and cholesterol categorization.

In this study, first we have categorized blood pressure (BP) observations into categories: optimal BP (systolic less than 120mm Hg, diastolic less than 80mm Hg), normal BP(systolic between 120 and 129mm Hg, diastolic between 80 to 84mm Hg), High normal BP (systolic ranging from 130 to139mm Hg, diastolic ranging between 85 to89mm Hg), Hypertension stage 1(systolic between 140 to159mm Hg, diastolic between 90 to99mm Hg), Hypertension stage 2 – 4 (systolic greater or equal to 160 mm Hg, diastolic greater or equal to 100 mm Hg). This has been done without considering if an individual is on hypertension medication or not. TC has been categorized into levels: <200 mm, (200-239) mm, (240-279) mm and ≥280 mm.

This study has then divided the data sample into a training and testing data sample. By training dataset this is the portion of our original data sample that has been used to estimate the five algorithms while the testing dataset is the portion of the original data sample that has been used to test for the accuracy of the five algorithms estimated by the training data portion.

Naïve Bayes

The Naïve Bayes algorithm estimation process consists of 5 steps. First, the data was separated by class. The classification variable in the sample dataset is the 10-year risk of coronary heart disease which takes the value of 1 if CHD is observed and 0 if not. Separating by class is having all the 10-year risk of coronary heart disease that have the value of one stacked together and those that have the value of 0 stacked together. By doing this it was easier to generate the class statistics of the data; the characteristics of the observations under the two classes.

The second step was to summarize the dataset. This was by getting the mean and variance of the independent variables. The third step was summarizing the data by class. This was accomplished by getting the mean and variance statistics in the independent variables in their respective classes, getting the mean and standard deviation of the independent variables in the CHD present class and then getting those same statistics for the CHD absent class.

The fourth step was estimating a Gaussian Probability Distribution function. Each of the input variables contain their individual observations and calculating the probability of observing a real value of any of these individual observations is hard. A solution to this roadblock was to assume that these observations of the independent variables were drawn from a certain distribution. We used a Gaussian distribution in this case since it is described by two values the mean and variance and its PDF is given as:

$$f(X) = \frac{1}{\sqrt{(2\Pi\sigma)}} \times e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

The fifth and final step was to get the class probabilities. In this step the probability of an observation composed of the independent variables was calculated to see if it belongs to the class of CHD present and the probability that it may belong to the class of CHD absent. The Gaussian probabilities of observing the various observations of the independent variables were used in this fifth step. The probability of an observation belonging to either of the classes was given by:

$$P(class \mid data) = P(X1 \mid class) \times P(X2 \mid class) \ldots \times P(X_n \mid class))$$

Given that we have 14 independent variables the estimated naïve Bayes equation is:

$$P(TenYearCHD = 1 \mid X1, X2 \ldots., X14) = P(X1 \mid TenYearCHD = 1) \times P(X2 \mid TenYearCHD = 1) \ldots \times P(X14 \mid TenYearCHD = 1))$$

To estimate CHD present and

$$P(TenYearCHD = 0 \mid X1, X2 \ldots., X14) = P(X1 \mid TenYearCHD = 0) \times P(X2 \mid TenYearCHD = 0) \ldots \times P(X14 \mid TenYearCHD = 0))$$

To estimate CHD absent.

Neural Networks

A neural network is a mathematical model that mimics the biological neural system. We used a three-layered neural network, with one input, one hidden and one output layer. Each layer is connected to each other and weights are assigned to these connections (Dangare & Sulabha , 2012).

The neurons in the input layer divide the CHD predictors into neurons in the hidden layer. The hidden layer then takes these inputs and weights of each connection to generate output, that is CHD positive or negative (Dangare & Sulabha , 2012).

SVM

SVM aims to find the maximum marginal hyperplane that best separates different classes. A hyperplane is simply a line that separates one class from another. In our case it is be separating those who have CHD and those who do not.

SVM, maps data to points in a multi-dimensional space and labels them to different classes using maximum margin hyper-plane. The side of the margin that the new data falls on predicts the future outcome (Gonsalves, Thabtah, Mohammad, & Singh, 2019).

Logistic Regression

This is a classification algorithm that assigns observations to discrete classes in our case CHD positive and negative. There are two types of logistic regressions: binary and multi linear. In our case we shall use the binary logistic regression since our target variable (Ten Year CHD) is binary that is 1 and 0 (Pant, 2019) and (Tavia , Castelli, Hjortland, Kannel, & Dawber, Predicting Coronary Heart Disease in Middle-Aged and Older Persons, 1977).

The algorithm uses probabilities to classify observations. It uses a sigmoid (logistic) cost function to map the predicted values to probabilities. A decision bound of 0.3 has been set such that if the probability of an observation exceeds the threshold, it is classified as CHD positive (1) otherwise it is classified as CHD negative (0) (Pant, 2019) and (Tavia , Castelli, Hjortland, Kannel, & Dawber, Predicting Coronary Heart Disease in Middle-Aged and Older Persons, 1977).

Decision Tree

A decision tree finds ways of splitting data based on different conditions, making it good for classification problems. The decision rules are usually if then else statements. These decision rules are implemented in the tree nodes since they test each attribute, and this process is recursive until the item has been appropriately classified (Hacker Earth, 2020).

3.7 Data presentation

This study has made use of tables, figures, and charts to convey the results of the study. The relationships between the variables are presented using a correlation heatmap diagram. A confusion matrix table has been presented to show the model accuracy in terms of false negatives, false positives, true negatives, and true positives as in the work by (Palaniappan & Awang, 2008) and (Dangare & Sulabha , 2012). The factor (predictor) levels that favor or disfavor the occurrence of CHD are presented in table of summary statistics.

CHAPTER FOUR: RESULTS

4.0 Introduction

This chapter gives the results for the study. It is broken down into descriptive statistics for individuals with and without CHD and the modelling results.

This study did not use the 4239 observations since the data was unevenly balanced (too many CHD negative individuals than CHD positive individuals). We thus had to resample the data and on randomly sampling without replacement, we arrived at a sample of 2000 observations. We thus had 292 CHD positive individuals and 1708 CHD negative individuals.

4.1 Descriptive statistics

The study results in Table 4.1 1 indicate that for CHD positive individuals their average glucose level is 85.71 mm with the highest glucose level is 394mm while the lowest is 45 mm. Their diastolic blood pressure is 86.977 mm Hg while their highest and lowest blood pressures were 140 mm Hg and 51 mm Hg, respectively. Their average BMI was 26.63, while the highest and the lowest BMI were 43.3 and 15.96, respectively. For systolic blood pressure, it averaged at 143.08 mm Hg while the highest and lowest systolic blood pressures were 295 mm Hg and 83.5 mm Hg. As for cholesterol, it had an average of 247.09 mm while the highest and lowest cholesterol levels were 464 mm and 124 mm, respectively. When it comes to smoking individuals with CHD smoke an average of 10.81 cigarettes per day, with some smoking a maximum of 43 cigarettes while other do not smoke at all.

Table 4.1: Descriptive statistics for the CHD present individuals

	count	mean	std	min	max
gender	292.0	0.571918	0.495650	0.00	1.0
age	292.0	54.143836	8.064593	35.00	69.0
education	292.0	1.804795	1.045155	1.00	4.0
currentSmoker	292.0	0.503425	0.500847	0.00	1.0
cigsPerDay	292.0	10.811644	13.097918	0.00	43.0
BPMeds	292.0	0.075342	0.264396	0.00	1.0
prevalentStroke	292.0	0.010274	0.101012	0.00	1.0
prevalentHyp	292.0	0.486301	0.500670	0.00	1.0
diabetes	292.0	0.051370	0.221130	0.00	1.0
totChol	292.0	247.092466	46.668256	124.00	464.0
sysBP	292.0	143.080479	26.887340	83.50	295.0
diaBP	292.0	86.977740	14.436388	51.00	140.0
BMI	292.0	26.633801	4.423150	15.96	43.3
heartRate	292.0	76.085616	12.254582	52.00	120.0
glucose	292.0	85.708904	32.824131	45.00	394.0
TenYearCHD	292.0	1.000000	0.000000	1.00	1.0
diaBPgrps	292.0	2.777397	1.531492	1.00	5.0
sysBPgrps	292.0	3.205479	1.423503	1.00	5.0
cholgrps	292.0	2.606164	0.966285	1.00	4.0

The study results in Table 4.1 2 for CHD negative individuals indicate that their average glucose level is 80.29 mm, their highest glucose level is 370 mm while their lowest is 44 mm. The average diastolic blood pressure is 82.27 Mm Hg while the highest and lowest blood pressures were 142.5 Mm Hg and 52 Mm Hg, respectively. The average BMI was 25.56, while the highest and the lowest BMI were 51.28 and 15.54, respectively. For systolic blood pressure, it averaged at 130.098 Mm Hg while the highest and lowest systolic blood pressures were 243 Mm Hg and 83.5 Mm Hg. As for cholesterol, it had an average of 235.87 mm while the highest and lowest cholesterol levels were 453 mm and 119 mm, respectively. When it comes to smoking individuals without CHD smoke an average of 8.60 cigarettes per day, with some smoking a maximum of 60 cigarettes while others do not smoke at all.

Table 4.2: Summary statistics for CHD absent individuals

	count	mean	std	min	max
gender	1708.0	0.419789	0.493669	0.00	1.00
age	1708.0	48.750585	8.304393	32.00	69.00
education	1708.0	2.005855	1.006115	1.00	4.00
currentSmoker	1708.0	0.485363	0.499932	0.00	1.00
cigsPerDay	1708.0	8.596019	11.421294	0.00	60.00
BPMeds	1708.0	0.017564	0.131400	0.00	1.00
prevalentStroke	1708.0	0.005855	0.076315	0.00	1.00
prevalentHyp	1708.0	0.281616	0.449919	0.00	1.00
diabetes	1708.0	0.018735	0.135629	0.00	1.00
totChol	1708.0	235.870609	43.826029	119.00	453.00
sysBP	1708.0	130.097775	20.707544	83.50	243.00
diaBP	1708.0	82.272834	11.386313	52.00	142.50
BMI	1708.0	25.563864	3.946701	15.54	51.28
heartRate	1708.0	76.015808	11.991028	44.00	143.00
glucose	1708.0	80.286885	17.296468	44.00	370.00
TenYearCHD	1708.0	0.000000	0.000000	0.00	0.00
diaBPgrps	1708.0	2.228337	1.359070	1.00	5.00
sysBPgrps	1708.0	2.442623	1.367327	1.00	5.00
cholgrps	1708.0	2.367096	0.986978	1.00	4.00

4.2 The modelling results

As seen in Figure 4.2 we first conducted a correlation test to check for the significance of the predictors in predicting CHD occurrence. However, using too many variables with low correlations was straining the model leading to poor model performance. A minimum correlation threshold of 0.1 between a CHD risk factor and Ten-Year CHD was set. Any variable that had a correlation below this was dropped.

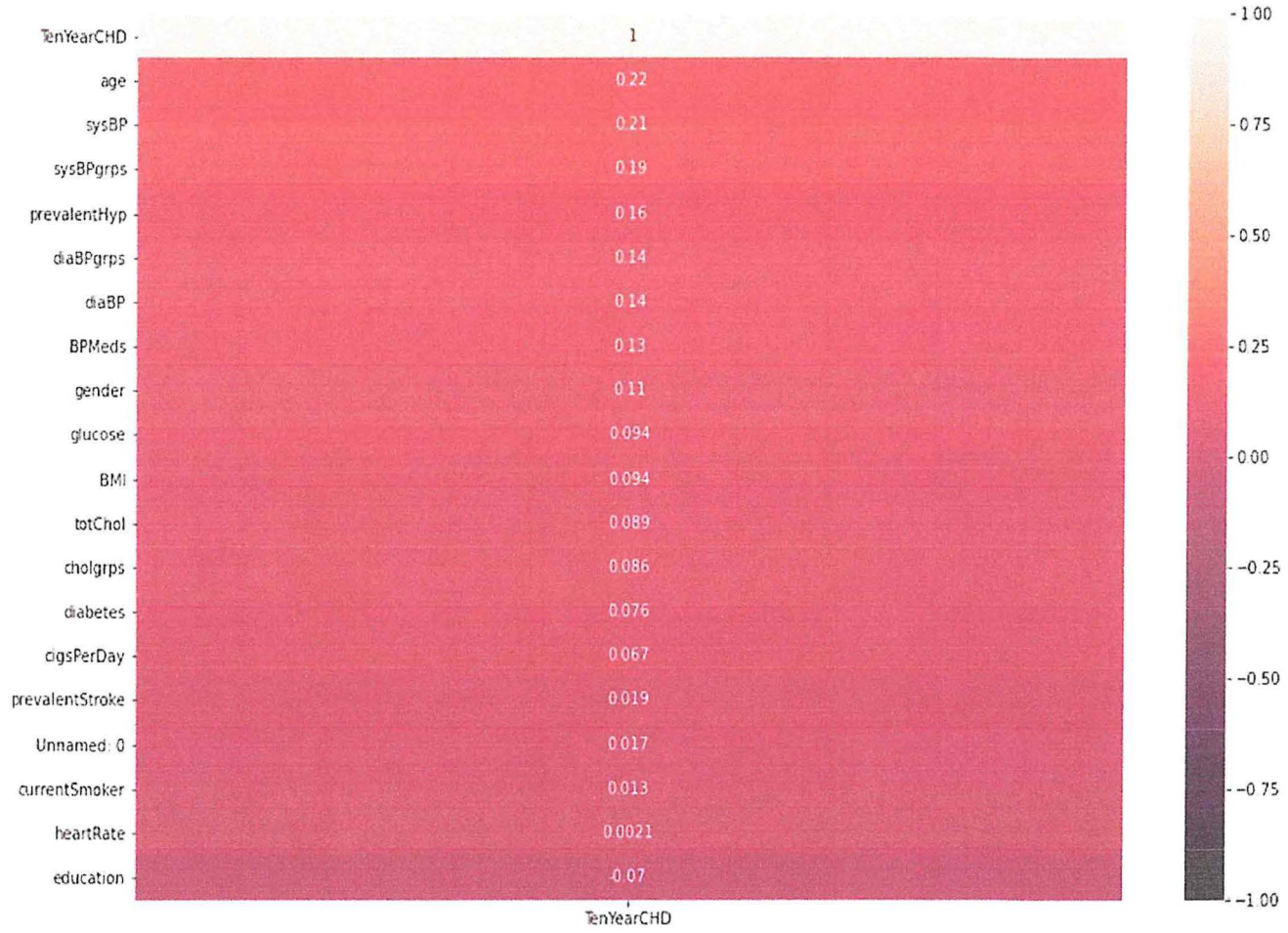

Figure 4.1: Correlation heatmap for all CHD risk factors.

We then re- estimated another correlation heatmap using the variables age, systolic blood pressure group, prevalent hypertension, blood pressure medication and diabetes as our predictors of CHD and got the correlation map shown in Figure 4.2.

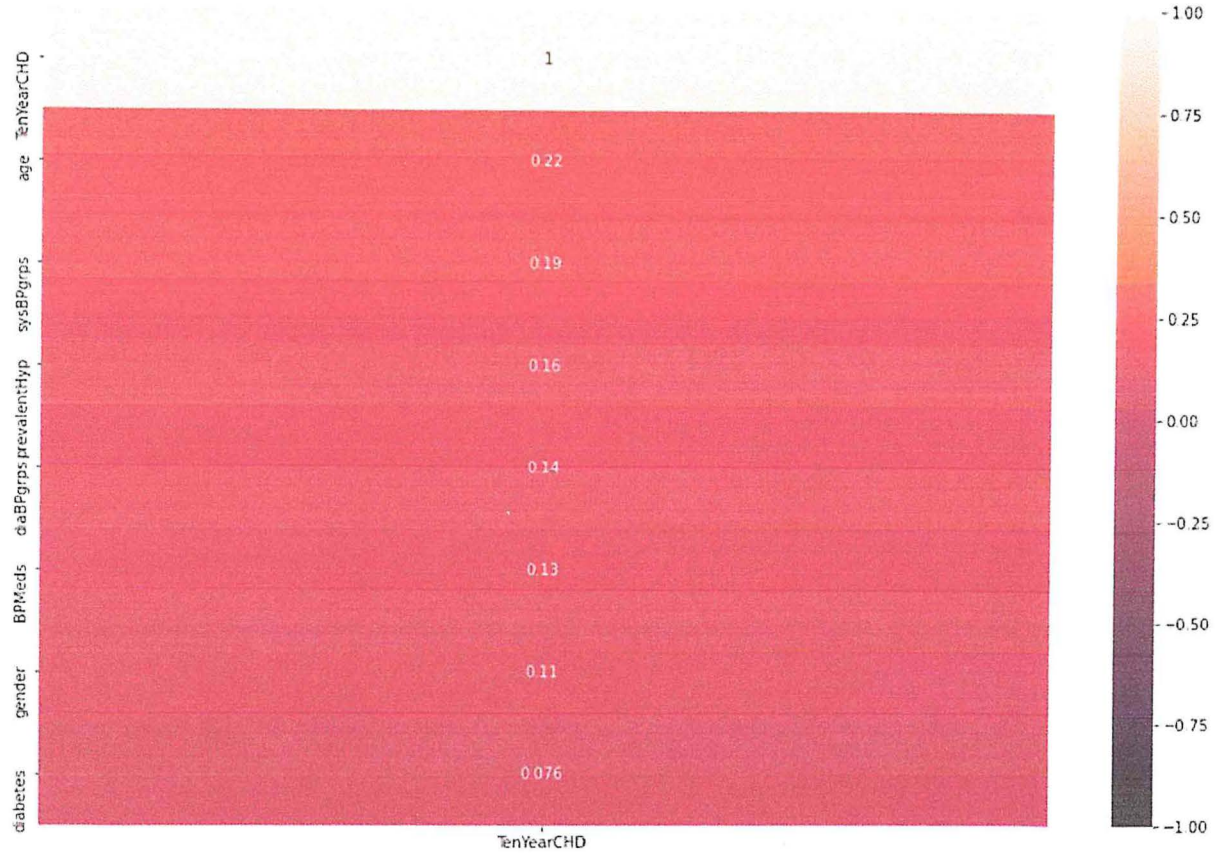

Figure 4.2: Correlation heatmap for variables used to model CHD.

The variables in Figure 4.2 were then used to estimate the 5 models in this study. The performance of these models was captured by 3 measures that is specificity, accuracy, and sensitivity.

Specificity measures the number of times our model correctly predicts CHD positive individuals. Sensitivity measures how many times our model correctly predicts individuals without CHD while accuracy measures how well our model can distinguish CHD positive and negative individuals generally.

Table 4.3 summarizes the performance of the estimated models.

Table 4.3: Model Performance

Test	Naïve bayes	Decision tree	Logistic regression	Neural network	Support Vector Machine
Accuracy	81.25%	80.5%	82.5%	86.75%	86.75%
specificity	26.42%	16.98%	30.19%	0%	0%
sensitivity	89.63%	90.2%	90.49%	100%	100%

SVM and neural networks had the highest accuracy of 86.75%, followed by 82.5% for the logistic regression, then 81.25% for Naïve Bayes and lastly by 80.5% for Decision Trees.

The model with the highest specificity of 30.19% was the logistic regression, followed by Naïve Bayes with 26.42%, then Decision Tree with 16.98%. The models with the lowest specificity of 0% were SVM and Neural networks.

SVM and neural networks had the highest sensitivity of 100%, followed by 90.49% from the logistic regression, 90.2% from the Decision Tree and the lowest sensitivity of 89.63% from Naïve Bayes.

A good model maintains a balance in accuracy, specificity, and sensitivity. The model that satisfies this constraint is the logistic regression, followed by the naïve bayes, then decision and finally by the SVM and neural networks. Thus, the best model for CHD prediction is the logistic regression.

Confusion matrices

The confusion matrices for each of the estimated models are shown in Table 4.4. The confusion matrix shows the number of individuals correctly and incorrectly classified as CHD positive or negative.

Table 4.4: Confusion Matrices

SVM & Neural Networks				
		Actual		
		Has CHD	No CHD	Total
Predicted	Has CHD	0	53	53
	No CHD	0	347	347
	total	0	400	
Naïve Bayes				
		Actual		
		Has CHD	No CHD	Total
Predicted	Has CHD	14	39	53
	No CHD	36	311	347
	total	50	350	
Logistic Regression				
		Actual		
		Has CHD	No CHD	Total
Predicted	Has CHD	16	37	53
	No CHD	33	314	347
	total	49	351	
Decision Tree				
		Actual		
		Has CHD	No CHD	Total
Predicted	Has CHD	9	44	53
	No CHD	34	313	347
	total	43	357	

It is clear from Table 4.2.2 that:

SVM and Neural Networks do a better job at classifying individuals without CHD since they do not misclassify any CHD negative individual. However, the Logistic regression misclassifies 33 as CHD negative and 37 CHD positive. Naïve Bayes misclassified 36 as CHD negative and 39 as CHD positive. Decision Tree misclassified 44 as CHD positive and 34 as CHD negative.

The model ranking based on highest power of predicting CHD negative individuals would be SVM and Neural Networks having the highest, followed by the logistic regression, then Decision Tree and finally Naïve Bayes.

The algorithm best suited to predict CHD positive individuals is the logistic regression since it misclassifies the least number of CHD positive individuals (37 individuals). It is then followed by the Naïve Bayes (39 misclassified), then Decision Trees (44 misclassified) and lastly SVM and Neural Networks (53 misclassified).

The model best for predicting CHD positive and negative individuals would be the logistic regression since it misclassifies the least, followed by Naïve Bayes, then Decision Trees and finally SVM and Neural Networks.

CHAPTER FIVE: DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.0 Introduction

This chapter gives the discussion, conclusion and recommendations for the study conducted.

5.1 Discussion

This study found that Neural networks and SVM were the best in predicting CHD negative individuals, since they both had a sensitivity of 100%. However, the works of (Palaniappan & Awang, 2008), found that the Decision Tree was best in predicting CHD negative individuals.

The Naïve Bayes algorithm was better than Neural Networks, SVM, and Decision Tree at predicting CHD positive individuals given it had the highest specificity of 26.42% this finding is consistent with the findings of (Palaniappan & Awang, 2008).

However, on adding the logistic regression, the logistic regression outperforms Naïve Bayes (specificity of 30.19%) in predicting CHD positive individuals. We thus found that the Logistic regression is the best in predicting CHD positive individuals.

The Naïve Bayes algorithm, having an accuracy of 81.25%, had a lower accuracy than SVM that had an accuracy of 86.75%. We found that Decision Trees (90.2% sensitivity), Neural Networks (100% sensitivity), and SVM (100% sensitivity) are best at predicting individuals without CHD. These findings are consistent with those of (Gonsalves, Thabtah, Mohammad, & Singh, 2019).

We find that SVM and Neural Networks are best for predicting CHD presence and absence when using model accuracy in assessing model performance (86.75% accuracy), followed by the logistic regression (82.5%), Naïve Bayes (81.25%) and finally Decision Tree (80.5%). If only comparing Naïve Bayes, Decision Trees and SVM, we note that SVM and Naïve Bayes outperform Decision Trees on accuracy as seen in the study by (Gonsalves, Thabtah, Mohammad, & Singh, 2019).

In our study we added the logistic regression into the mix, and it ends up outperforming the Naïve Bayes. We thus conclude that the Logistic regression is best in predicting CHD positive and negative individuals due to its higher well-balanced accuracy, sensitivity, and specificity measures.

Also, on observing individuals with CHD we realized that they suffered from high cholesterol (235mmg on average), high blood sugar (a maximum of 394mmg), had a smoking habit (10.82 cigarettes per day on average), were obese (overweight BMI of 26.63 on average) and had high

blood pressure (a maximum of 295/140 Hmg and 143/86 Hmg on average) as seen in the works of (Shirley, et al., Can lifestyle changes reverse coronary heart disease, 1990), (Palaniappan & Awang, 2008) and (Tavia, Castelli, Hjortland, Kannel, & Dawber, Predicting Coronary Heart Disease in Middle-Aged and Older Persons, 1977)

5.2 Conclusion

The best model for CHD prediction given the data sample used would be the Logistic regression, followed second by Naïve Bayes, then by Decision Tree and finally Neural Networks and SVM. Based on measures of accuracy, sensitivity and specificity, the logistic regression algorithm had the highest of these three (82.5%, 90.49% and 30.19%).

5.3 Recommendation

5.3.1Recommendations for further studies

Further studies need to be done using new datasets. Unfortunately, heart disease data is hard to come by. Future research should be done with more recent samples to capture the evolution of both the CHD risk factors and the human body.

Future studies can also focus on CHD occurrence in children since in the US the rate of obesity amongst children is on the rise 13.7 million adolescents and children currently (Centers for Disease Control and Prevention (CDC), 2020).

5.3.2 Recommendation for policy making.

CHD is a deadly disease which can be avoided by better healthy living. Governments should regularize free CHD, diabetes, and hypertension checkups across all public hospitals to ensure early detection of these diseases. The warnings on cigarettes packaging have been made clearer but more public awareness programs on the side effects of smoking should be enacted.

Physical education in schools is mandatory in schools thus ensuring all children are living healthy fit lives. For adults, gym memberships could be subsidized to encourage more adults to take up physical fitness affordably.

Healthy diets should be adopted by all schools to curb obesity and high cholesterol levels in schools. Also, foods rich in trans fats should be banned or heavily regulated given their high cholesterol content.

Subsidizing hypertension and diabetes medication will go a long way in combating these two diseases and combating CHD since the two diseases are CHD risk factors.

REFERENCES

Dangare, C. S., & Sulabha , A. S. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *International Journal of Computer Applications (0975 – 888) Volume 47– No.10*.

Levy, D., Benjamin, E. J., Johnson, A. D., Andersson, C., & Ramachandran , V. S. (2019). 70-year legacy of the Framingham Heart Study. *Nature Reviews Cardiology*, 200 - 2012.

AmericanHeartAssociation. (2014, Jun 17). *Stress and Heart Health*. Retrieved from www.heart.org: https://www.heart.org/en/healthy-living/healthy-lifestyle/stress-management/stress-and-heart-health

AmericanHeartAssosiation. (2017, Nov 30). *Five Simple Steps to Control Your Blood Pressure*. Retrieved from www.heart.org: https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/five-simple-steps-to-control-your-blood-pressure

AmericanHeartAssosication. (2017, May 31). *What is Cardiovascular Disease?* Retrieved from www.heart.org: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease

BritishHeartFoundation. (2010). *Focus on: Chemotherapy and the heart*. Retrieved from Bhf.org.uk: https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/chemotherapy

CDC. (2019, December 2). *Heart Disease Facts | cdc.gov*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/heartdisease/facts.htm

Centers for Disease Control and Prevention (CDC). (2020). *https://www.cdc.gov/*. Retrieved from Childhood Obesity Facts: https://www.cdc.gov/obesity/data/childhood.html#:~:text=Prevalence%20of%20Childhood%20Obesity%20in%20the%20United%20States&text=The%20prevalence%20of%20obesity%20was,to%2019%2Dyear%2Dolds.

Cochran, W. G. (1977). *Sampling Techniques (3rd Ediction).* New York: John Wiley & Sons.

DiabetesSelfManagement. (2006, JUNE 16). *Tight Control - Diabetes Resources & Information | Diabetes Self-Management*. Retrieved from Diabetes Self-Management: https://www.diabetesselfmanagement.com/diabetes-resources/definitions/tight-control/

familydoctor.orgstaff. (2019, March 27). *Coronary Heart Disease - Coronary Artery Disease | familydoctor.org*. Retrieved from familydoctor.org: https://familydoctor.org/condition/coronary-heart-disease-chd/

Felman, A. (2019, July 5). *Coronary heart disease: Causes, symptoms, and treatment*. Retrieved from Medicalnewstoday.com: https://www.medicalnewstoday.com/articles/184130

Felman, A. (2019, July 5). *Coronary heart disease: Causes, symptoms, and treatment*. Retrieved from Medicalnewstoday.com: https://www.medicalnewstoday.com/articles/184130

Fodor, G. (2004). Coronary heart disease: is gender important? *The Journal of Men's Health and Gender VOL. 1*, 32 - 37.

Gonsalves, A. H., Thabtah, F., Mohammad, R. M., & Singh, G. (2019). Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis.

Hacker Earth. (2020). *https://www.hackerearth.com*. Retrieved from Decision Tree: https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/

Harris, P. J., Behar, V. S., Conley, M. J., Harrel, F. E., Lee, K. L., Peter, R. H., . . . P. D. (1980). *The Prognostic Significance of 50% Coronary Stenosis in Medically Treated Patients with Coronary Artery Disease.* Circulation 62.

HarvardPublishing. (2006, June). *Atherosclerosis : Exercise essential in combating arterial disease - Harvard Health*. Retrieved from Harvard Health: https://www.health.harvard.edu/press_releases/atherosclerosis_arterial_diseas

HarvardPublishing. (2016, April). *The heart attack gender gap - Harvard Health*. Retrieved from Harvard Health: https://www.health.harvard.edu/heart-health/the-heart-attack-gender-gap

HealthAffairs. (2007, JANUARY/FEBRUARY). *Advances In The Prevention And Treatment Of Cardiovascular Disease | Health Affairs*. Retrieved from Healthaffairs.org: https://www.healthaffairs.org/doi/full/10.1377/hlthaff.26.1.25

HealthDirect. (2020, January). *Coronary heart disease treatment*. Retrieved from Healthdirect.gov.au: https://www.healthdirect.gov.au/coronary-heart-disease-treatments

HealthInformation. (2017, February). *Diabetes, Heart Disease, and Stroke | NIDDK*. Retrieved from National Institute of Diabetes and Digestive and Kidney Diseases: https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke

Healthline. (2016 , February 22). *Congenital Heart Disease: Types, Symptoms, Causes, and Treatment*. Retrieved from Healthline: https://www.healthline.com/health/congenital-heart-disease

Healthline. (2017, May 2). *Alcoholic Cardiomyopathy: Causes, Symptoms, and Diagnosis*. Retrieved from Healthline: https://www.healthline.com/health/alcoholism/cardiomyopathy

HealthNY. (1999, August). *Physical Inactivity and Cardiovascular Disease*. Retrieved from Health.ny.gov: https://www.health.ny.gov/diseases/chronic/cvd.htm

Holland, K. (2019, June 25). *The Alcohol-Depression Connection: Symptoms, Treatment & More*. Retrieved from Healthline: https://www.healthline.com/health/mental-health/alcohol-and-depression

JohnHopkinsMedicine. (2009). *Rheumatic Heart Disease*. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/rheumatic-heart-disease

Lee, H. G., Noh, K. Y., & Ryu, K. H. (May 2007). Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV. *Emerging Technologies in Knowledge Discovery and Data Mining*, 56-66.

Mackay J, M. G. (2004). *The Atlas of Heart Disease and Stroke.* Geneva: World Health Organisation. Retrieved from Health Knowledge: https://www.healthknowledge.org.uk/public-health-textbook/disease-causation-diagnostic/2b-epidemiology-diseases-phs/chronic-diseases/coronary-heart-disease#ref%201

Martin, K. G. (2020). *Measures of Predictive Models: Sensitivity and Specificity*. Retrieved from theanalysisfactor.com: https://www.theanalysisfactor.com/sensitivity-and-specificity/

MayoClinic. (2018, May 16). *Coronary artery disease - Symptoms and causes*. Retrieved from Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-20350613

MayoClinic. (2018, March 22). *Heart disease - Diagnosis and treatment - Mayo Clinic*. Retrieved from Mayoclinic.org: https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124

MayoClinic. (2018, March 22). *Heart disease - Symptoms and causes*. Retrieved from Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

MayoClinic. (2019, July 18). *Deep vein thrombosis - Symptoms and causes*. Retrieved from Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/deep-vein-thrombosis/symptoms-causes/syc-20352557

MayoClinic. (2019, Jan 25). *Your teeth and your heart: What's the connection?* Retrieved from Mayo Clinic: https://www.mayoclinic.org/healthy-lifestyle/adult-health/expert-answers/heart-disease-prevention/faq-20057986

MayoClinic. (2020, March 19). *Intensive insulin therapy: Achieving tight blood sugar control*. Retrieved from Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/diabetes/in-depth/intensive-insulin-therapy/art-20043866

MedicalNewsToday. (2019, August 2). *Cerebrovascular disease: Causes, symptoms, and treatment*. Retrieved from Medicalnewstoday.com: https://www.medicalnewstoday.com/articles/184601

Mishra, A. (2018 , Feb 24). *Metrics to Evaluate your Machine Learning Algorithm*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

NationalStatistics. (2006, Spring). Health Statistics Quarterly 30.

NHBL, & NIH. (2019). *Framingham Heart Study (FHS) | NHLBI, NIH.* Retrieved from Nhlbi.nih.gov: https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs#how-it's-conducted

NHS. (2019, October 31). *Peripheral arterial disease (PAD)*. Retrieved from nhs.uk: https://www.nhs.uk/conditions/peripheral-arterial-disease-pad/

NHS. (2018 , September 17). *Cardiovascular disease*. Retrieved from nhs.uk: https://www.nhs.uk/conditions/Cardiovascular-disease/

NHS. (2020, March 10). *Coronary heart disease*. Retrieved from nhs.uk: https://www.nhs.uk/conditions/coronary-heart-disease/

NHS. (2020, March 10). *Coronary heart disease - Causes*. Retrieved from nhs.uk: https://www.nhs.uk/conditions/coronary-heart-disease/causes/

NHS. (2020, March 10). *Coronary heart disease - Symptoms*. Retrieved from nhs.uk: https://www.nhs.uk/conditions/coronary-heart-disease/symptoms/

NHS. (2020, March 10). *Coronary heart disease - Treatment*. Retrieved from nhs.uk: https://www.nhs.uk/conditions/coronary-heart-disease/treatment/

Nichols, H. (2017, September 18). *Five ways to quit smoking*. Retrieved from Medicalnewstoday.com: https://www.medicalnewstoday.com/articles/319460#3.-Consider-non-nicotine-medications

NIH. (2018, June 01). *Heart Health and Aging*. Retrieved from National Institute on Aging: https://www.nia.nih.gov/health/heart-health-and-aging

Palaniappan, S., & Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. *IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8*.

Pant, A. (2019, Jan 22). *https://towardsdatascience.com*. Retrieved from Introduction to Logistic Regression: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

Ritchie, H. (2019, December). *Causes of Death*. Retrieved from Our World in Data: https://ourworldindata.org/causes-of-death

Shirley, E. B., William, T. A., Thomas, A. P., Dean, O., Brand, R. J., Gould, L. K., & Kirkeeide, R. L. (1990). Can lifestyle changes reverse coronary heart disease. *THE LANCET*, 129-133.

Shirley, E. B., William, T. A., Thomas, A. P., Dean, O., Brand, R. J., Gould, L. K., . . . McLanahan, S. M. (1990). Can lifestyle changes reverse coronary heart disease. *THE LANCET*, 129-133.

Sullivan, D., & Felman, A. (2019, July 5). *Coronary heart disease: Causes, symptoms, and treatment*. Retrieved from Medicalnewstoday.com: https://www.medicalnewstoday.com/articles/184130

Tavia , G., Castelli, W. P., Hjortland, M. C., Kannel, W. B., & Dawber, T. R. (1977). Predicting Coronary Heart Disease in Middle-Aged and Older Persons. *The Framington Study*.

Tavia, G., Castelli, W. P., Hjortland, M. C., Kannel, W. B., & Dawber, T. R. (1977). Predicting Coronary Heart Disease in Middle-Aged and Older Persons. *JAMA*, 497-499.

WHO. (2017). *Cardiovascular diseases*. Retrieved from Who.int.

Wilson, P. W., Kannel, W. B., & Agostino, R. B. (1998). *Prediction of Coronary Heart Disease Using Risk Factor Categories.*

X, Y., J, W., Z, Z., & Y, G. (2007). Combination data mining models with new medical data to predict outcome of coronary heart disease. *Proceedings International Conference on Convergence Information Technology* , 868 – 872.

Yamane, T. (1967). *Statistics, An Introductory Analysis.* New York: Harper and Row.