EPIDEMIOLOGICAL SURVEILLANCE OF CARDIO VASCULAR DISEASE

A Project Report submitted in partial fulfillment of the requirements

for the Award of the Degree of

MASTER OF SCIENCE
IN
STATISTICS

Submitted by

RAMAMOORTHY D

Reg. No. 21STAC14

Under the Guidance of

Dr.K.M.SAKTHIVEL

PROFESSOR

DEPARTMENT OF STATISTICS
BHARATHIAR UNIVERSITY
COIMBATORE – 641 046

APRIL 2023

CERTIFICATE

This is to certify that the project work entitled **"EPIDEMIOLOGICAL SURVEILLANCE OF CARDIO VASCULAR DISEASE"** submitted to the Bharathiar University in partial fulfillment of the Award of Degree of Master of Science in Statistics is a record of original work done by **RAMAMOORTHY D (21STAC14)** during the period of his study in the Department of Statistics, Bharathiar University, Coimbatore. This work has not formed the basis for the award of any Degree / Diploma / Associateship / Fellowship or similar titled to any candidate of any University.

Head of the Department
(Dr. R.VIJAYARAGHAVAN)

Signature of the Guide
(Dr.K.M.SAKTHIVEL)

DECLARATION

I, hereby declare that the project work entitled **"EPIDEMIOLOGICAL SURVEILLANCE OF CARDIO VASCULAR DISEASE"** submitted to the Bharathiar University in partial fulfillment of the Award of Degree of Master of Science is Statistics is a record of original work done by me under the guidance of **Dr.K.M.SAKTHIVEL**, Professor, Department of Statistics, Bharathiar University, Coimbatore. This work has not formed the basis for the award of any Degree / Diploma / Associateship / Fellowship or similar titled to any candidate of any University.

Signature of the Guide
(Dr. K.M.SAKTHIVEL)

Signature of the Candidate
(RAMAMOORTHY D)

ACKNOWLEDGEMENT

In preparation of my project, I had to take the help and guidance of some respected people, who deserve my gratitude. As the completion of this project gave me much pleasure, I'm eternally grateful and itis my great privilege to express my deepest sense of gratitude to my supervising guide **Dr.K.M.Sakthivel,** Professor, Department of Statistics, Bharathiar University for guiding me throughout my project work.

I extend my profound thanks to **Dr. R.Vijayaraghavan,** Senior Professor and Head, Department of Statistics for his encouragement during the period of my study.

I render my sincere gratitude to the faculties of Department of Statistics, **Dr.R.Jaisankar, Dr.R.Muthukrishnan, Dr.V.Kaviyarasu, Dr.K.PradeepaVeerakumari, Dr. S.Jayalakshmi, Dr.S.GandhiyaVendhan and Dr. Sivasankari** for their constant support and encouragement while undergoing the project work.

It is an immense pleasure and my duty to express my sincere thanks to the Guest Lectures of the Department of Statistics for their support during my entire project work.

I would also extend my thanks to the non – teaching staffs of the Department of Statistics of Bharathiar University for arranging the Lab and Library Facility.

Words are insufficient to express my gratitude to the Research Scholar **Ms. Vidhya G, Ms. Nandhini S, Ms. Alicia Mathew** for her support and insightful suggestions throughout my project work.

I place my thanks to the God, my family members, and my friends for their great support and prayers for my project work.

I feel delighted to convey my thanks to my fellow project mates of Bharathiar University for rendering their support throughout my project work.

Last but not the least; a special thanks to my friends and all the students of the Department of Statistics, Bharathiar University for their support throughout my project work.

RAMAMOORTHY D

CONTENTS

Chapter	Title	Page
1	INTRODUCTION	1
2	METHODOLOGY	21
3	DATA PRESENTATION	40
4	ANALYSIS AND INTERPRETATION	48
5	SUMMARY AND CONCLUSION	60
	REFERENCE	63

Chapter 1

INTRODUCTION

1.1. HEART

The heart is a muscular organ about the size of a fist, located just behind and slightly left of the breastbone. The heart pumps blood through the network of arteries and veins called the cardiovascular system. The heart has four chambers: The right atrium receives blood from the veins and pumps it to the right ventricle. The right ventricle receives blood from the right atrium and pumps it to the lungs, where it is loaded with oxygen. The left atrium receives oxygenated blood from the lungs and pumps it to the left ventricle. The left ventricle (the strongest chamber) pumps oxygen-rich blood to the rest of the body. The left ventricle's vigorous contractions create our blood pressure. The coronary arteries run along the surface of the heart and provide oxygen-rich blood to the heart muscle. A web of nerve tissue also runs through the heart, conducting the complex signals that govern contraction and relaxation. Surrounding the heart is a sac called the pericardium. The heart's main function is to move blood throughout the body. The heart also controls the rhythm and speed of the heart rate and maintains blood pressure. The heart works with other body systems to control the heart rate and other body functions.

The primary systems are the nervous system and endocrine system. The nervous system helps control the heart rate. It sends signals that tell the heart to beat slower during rest and faster during stress. The endocrine system sends out hormones. These hormones tell the blood vessels to constrict or relax, which affects blood pressure. Hormones from the thyroid gland can also tell the heart to beat faster or slower. The parts of the heart are Walls, Chambers, Valves, Blood vessels, and the Electrical conduction system.

1.1.1. Heart walls

The heart walls are the muscles that contract and relax to send blood throughout the body. A layer of muscular tissue called the septum divides the heart walls into the left and right sides. The heart walls have three layers endocardium (Inner layer), myocardium (muscular middle layer), and epicardium (protective outer layer). The epicardium is one layer

of the pericardium. The pericardium is a protective sac that covers the entire heart. It produces fluid to lubricate the heart and keep it from rubbing against other organs.

1.1.2. Heart chambers

The heart is divided into four chambers. It has two chambers on the top (atrium, plural atria) and two on the bottom (ventricles), one on each side of the heart.

- *Right atrium:* Two large veins deliver oxygen-poor blood to your right atrium. The superior vena cava carries blood from your upper body. The inferior vena cava brings blood from the lower body. Then the right atrium pumps the blood to your right ventricle.
- *Right ventricle:* The lower right chamber pumps oxygen-poor blood to the lungs through the pulmonary artery. The lungs reload blood with oxygen.
- *Left atrium:* After the lungs fill the blood with oxygen, the pulmonary veins carry the blood to the left atrium. This upper chamber pumps the blood to your left ventricle.
- *Left ventricle:* The left ventricle is slightly larger than the right. It pumps oxygen-rich blood to the rest of the body.

1.1.3. Heart valves

The heart valves are like doors between the heart chambers. They open and close to allow blood to flow through.

- *The atrioventricular (AV) valves* open between the upper and lower heart chambers. They include:
- *Tricuspid valve:* Door between the right atrium and right ventricle.
- *Mitral valve:* Door between the left atrium and left ventricle.
- *Semilunar (SL) valves* open when blood flows out of the ventricles. They include:
- *Aortic valve:* Opens when blood flows out of the left ventricle to the aorta (the artery that carries oxygen-rich blood to the body).
- *Pulmonary valve:* Opens when blood flows from the right ventricle to the pulmonary arteries (the only arteries that carry oxygen-poor blood to the lungs).

1.1.4. Blood vessels

The heart pumps blood through three types of blood vessels:

- Arteries carry oxygen-rich blood from the heart to the body's tissues. The exception is the pulmonary arteries, which go to the lungs.
- Veins carry oxygen-poor blood back to the heart.
- Capillaries are small blood vessels where the body exchanges oxygen-rich and oxygen-poor blood.
- The heart receives nutrients through a network of coronary arteries. These arteries run along the heart's surface. They serve the heart itself.
- Left coronary artery: Divides into two branches (the circumflex artery and the left anterior descending artery).
- Circumflex artery: Supplies blood to the left atrium and the side and back of the left ventricle.
- Left anterior descending artery (LAD): Supplies blood to the front and bottom of the left ventricle and the front of the septum.
- Right coronary artery (RCA): Supplies blood to the right atrium, right ventricle, bottom portion of the left ventricle, and back of the septum.

1.1.5. Electrical conduction system

The heart's conduction system is like the electrical wiring of a house. It controls the rhythm and pace of your heartbeat. It includes:

- Sinoatrial (SA) node: Sends the signals that make the heartbeat.
- Atrioventricular (AV) node: Carries electrical signals from the heart's upper chambers to its lower ones.
- The heart also has a network of electrical bundles and fibers. This network includes:
- Left bundle branch: Sends electric impulses to the left ventricle.
- Right bundle branch: Sends electric impulses to the right ventricle.
- Bundle of His: Sends impulses from the AV node to the Purkinje fibers.
- Purkinje fibers: Make the heart ventricles contract and pump out blood.

1.1.6. Heart Rate

Heart rate (or pulse rate) is the frequency of the heartbeat measured by the number of contractions of the heart per minute (beats per minute, or bpm). The heart rate can vary according to the body's physical needs, including the need to absorb oxygen and excrete carbon dioxide, but is also modulated by numerous factors, including (but not limited to) genetics, physical fitness, stress or psychological status, diet, drugs, hormonal status, environment, and disease/illness as well as the interaction between and among these factors. It is usually equal or close to the pulse measured at any peripheral point.

The American Heart Association states the normal resting adult human heart rate is 60-100 bpm. Tachycardia is a high heart rate, defined as above 100 bpm at rest. Bradycardia is a low heart rate, defined as below 60 bpm at rest. When a human sleeps, a heartbeat with rates around 40–50 bpm is common and is considered normal. When the heart is not beating in a regular pattern, this is referred to as an arrhythmia. Abnormalities of heart rate sometimes indicate disease.

A heart rate below 60 beats per minute or above 100 beats per minute is often considered dangerous when a person is at rest (and awake) but can be normal in other states. During sleep, for instance, a person's heart rate may often drop below 60 beats per minute. Meanwhile, heart palpitations exceeding 100 beats per minute can be caused by non-dangerous conditions as well, such as a rigorous run or temporary stress. An abnormally high or low heart rate is considered worrisome when it occurs for multiple hours or days. It may also be considered dangerous when it's accompanied by other symptoms.

If the heart rate consistently measures above 100 beats per minute, it's classified as tachycardia, a condition marked by a rapid heart rate. There are several main types of tachycardia:

- Sinus tachycardia (ST) is a fast heart rhythm that results from faster-than-normal electrical impulses generated from the sinus node. A rare subset of sinus tachycardia is inappropriate sinus tachycardia when the heart beats quickly for no clear reason.
- Supraventricular tachycardia (SVT) is a fast heartbeat that originates from the heart's upper chambers. There are a wide variety of supraventricular tachycardias, the most

notable of which is atrial fibrillation, as it is the most common sustained arrhythmia worldwide.

- Junctional tachycardia (JT) originates from the junction, an area between the upper and lower chambers of the heart.
- Ventricular tachycardia (VT) occurs when the lower chambers of the heart beat very quickly. It's often more serious than SVT.

If the heart rate consistently sits under 60 beats per minute which, as mentioned above, can be considered normal in healthy, athletic people. It's unlikely to lead to complications unless the heart rate is regularly less than 40 beats per minute and/or causes other symptoms. There are several main types of bradycardia:

- Sinus bradycardia simply occurs when the heart rate is less than 60 beats per minute, and Sick sinus syndrome occurs when the sinus node (the heart's natural pacemaker) fails to function properly. Sinus pauses, a type of bradycardia in which the heart slows down because the sinus node isn't activating the electrical system throughout the rest of the heart correctly may occur in this setting. Various arrhythmias or combinations of arrhythmias can present in people with sick sinus syndrome.
- Tachycardia-bradycardia syndrome, also called a tachy-brady syndrome, often occurs in people with atrial fibrillation who also have sinus node dysfunction. Sometimes the heart beats too quickly while other times it beats too slowly.
- Heart block is an abnormality in the way electricity passes through the heart, "blocking" the electrical impulse on its normal path and resulting in a slower heart rate. There are multiple forms of heart block, with each differing in severity and corrective treatment.
- Ectopic bradycardia occurs when a focus other than the sinus node functions as the natural pacemaker for the heart. It can occur in the atria, termed ectopic atrial bradycardia, in the junction, called junctional bradycardia, or in the ventricles, dubbed idioventricular rhythm.

1.2. HEART DISEASE

Heart disease is a general term that includes many types of heart problems. It's also called cardiovascular disease, which means heart and blood vessel disease. Heart disease is

caused by a variety of factors, including genetics, lifestyle, and environmental factors. Heart disease describes a range of conditions that affect the heart. Heart diseases include Blood vessel diseases, such as coronary artery disease, Irregular heartbeats (arrhythmias), Born with heart problems (congenital heart defects), Disease of the heart muscle, and Heart valve disease. Coronary artery disease is the most common form of heart disease, and it occurs when the arteries that supply blood to the heart become narrowed or blocked due to the buildup of plaque in the artery walls. This can lead to a heart attack or myocardial infarction.

Congestive heart failure is a condition in which the heart is unable to pump enough blood to meet the body's needs. This can be caused by several factors including high blood pressure, heart attack, or abnormal heart rhythms. Arrhythmia is an abnormal heart rhythm, which can cause the heart to beat too fast, too slow, or erratically. This can lead to reduced blood flow to the body and can increase the risk of stroke.

Heart valve disease occurs when the valves that control the flow of blood in and out of the heart become damaged or narrowed. This can lead to poor blood flow throughout the body and can even cause the heart to stop beating. Heart disease is a serious condition that can have a significant impact on a person's life. Fortunately, there are several treatments available to help manage the symptoms and reduce the risk of further complications. These include lifestyle changes, such as eating a healthy diet, exercising regularly, and quitting smoking; medications; and, in some cases, surgery.

1.3. SYMPTOMS OF HEART DISEASE

Heart disease can manifest in a variety of ways, and symptoms can vary depending on the type of heart disease and the severity of the condition. Here are some common symptoms of heart disease:

1. *Chest pain or discomfort:* This is the most common symptom of heart disease. It may feel like pressure, tightness, or a squeezing sensation in the chest.
2. *Shortness of breath:* If experience any difficulty in breathing or feel like can't catch a normal breath.
3. *Fatigue:* Feeling tired or weak, especially with physical activity.

4. *Dizziness or light-headedness:* You may feel faint or dizzy, or you may faint.
5. *Heart palpitations:* This is the sensation of an irregular heartbeat or a feeling that your heart is racing or fluttering.
6. *Swelling in the legs, ankles, or feet:* This can be a sign of fluid build-up due to a weakened heart.
7. *Nausea or vomiting:* Some people with heart disease may experience these symptoms, especially if they have a heart attack.

If experience any of these symptoms, it's important to see a doctor as soon as possible, especially if you have risk factors for heart disease such as high blood pressure, high cholesterol, smoking, diabetes, or a family history of heart disease.

1.4. RISK FACTORS FOR HEART DISEASE

Heart disease, also known as cardiovascular disease, is a condition that affects the heart and blood vessels, including arteries, veins, and capillaries. It is a leading cause of death worldwide, responsible for nearly one-third of all deaths globally. The following are the major risk factors associated with heart disease:

1. *High blood pressure:* High blood pressure, also known as hypertension, is a major risk factor for heart disease. High blood pressure makes it harder for the heart to pump blood, which can lead to damage to the heart and blood vessels over time.
2. *High cholesterol:* High levels of cholesterol in the blood can lead to the build-up of plaque in the arteries, which can narrow the blood vessels and increase the risk of heart disease.
3. *Smoking:* Smoking damages the lining of the blood vessels and increases the risk of plaque buildup in the arteries. It also increases the risk of blood clots, which can cause a heart attack or stroke.
4. *Diabetes:* People with diabetes are at higher risk for heart disease, as high blood sugar levels can damage the blood vessels and increase the risk of plaque buildup.
5. *Family history:* If a close family member, such as a parent or sibling, has had heart disease, then the risk of developing heart disease is higher.

6. *Age:* The risk of heart disease increases with age, particularly for men over the age of 45 and women over the age of 55.
7. *Obesity:* Being overweight or obese can increase the risk of heart disease, as it can lead to high blood pressure, high cholesterol, and diabetes.
8. *Physical inactivity:* Lack of physical activity can increase the risk of heart disease, as it can lead to obesity, high blood pressure, and high cholesterol.
9. *Stress:* Chronic stress can increase the risk of heart disease, as it can raise blood pressure, increase inflammation, and lead to unhealthy coping mechanisms like smoking and overeating.
10. *Unhealthy diet:* Eating a diet high in saturated and trans fats, salt, and sugar can increase the risk of heart disease, as it can lead to high blood pressure, high cholesterol, and obesity.

Overall, taking steps to address these risk factors, such as maintaining a healthy weight, exercising regularly, quitting smoking, and eating a healthy diet, can help reduce the risk of heart disease. Additionally, regular check-ups with a healthcare provider can help identify and manage any underlying risk factors.

1.5. TYPES OF HEART DISEASE

Heart disease is a broad term that encompasses several different conditions that affect the heart and blood vessels. These conditions can range from relatively mild, such as high blood pressure, to severe, such as heart failure or heart attack. Here are some of the most common types of heart disease:

1. *Coronary artery disease (CAD):* This is the most common type of heart disease and occurs when the blood vessels that supply the heart with oxygen and nutrients become narrow or blocked. CAD can lead to chest pain (angina), heart attack, and heart failure.
2. *Heart failure:* Heart failure occurs when the heart is unable to pump enough blood to meet the body's needs. This can happen due to a variety of reasons, including CAD, high blood pressure, and damage to the heart muscle (such as from a heart attack).
3. *Arrhythmia:* An arrhythmia is an abnormal heart rhythm that can cause the heart to beat too fast, too slow, or irregularly. This can lead to symptoms such as dizziness, light-headedness, and fainting.

4. *Valvular heart disease:* This is a condition that affects the heart valves, which regulate blood flow through the heart. Valvular heart disease can cause the valves to become leaky or narrow, which can lead to symptoms such as shortness of breath, fatigue, and chest pain.
5. *Congenital heart disease:* Congenital heart disease is a type of heart disease that is present at birth. It can range from relatively mild, such as a small hole in the heart, to severe, such as a missing or improperly formed heart valve.
6. *Cardiomyopathy:* Cardiomyopathy is a disease that affects the heart muscle, making it harder for the heart to pump blood effectively. This can lead to symptoms such as fatigue, shortness of breath, and swelling in the legs.
7. *Peripheral artery disease (PAD):* PAD occurs when there is a build-up of plaque in the arteries that supply blood to the legs and other parts of the body. This can lead to symptoms such as leg pain and cramping and can increase the risk of heart attack and stroke.

These are just a few of the many types of heart disease that can affect people. It's important to talk to the doctor if experience any symptoms or risk factors for heart disease, as early detection and treatment can help prevent serious complications.

1.6. TREATMENTS FOR HEART DISEASES

The treatment options will vary depending on the type of heart disease a person has, but some common strategies include making lifestyle changes, taking medications, and undergoing surgery. Various medications and surgical methods are mentioned below:

1.6.1. Medications

- *Anticoagulants:* Also known as blood thinners, these medications can prevent clots. They include warfarin (Coumadin) and the direct oral anticoagulants dabigatran, rivaroxaban, and apixaban.
- *Antiplatelet therapies:* These include aspirin, and they can also prevent clots.
- *Angiotensin-converting enzyme inhibitors:* These can help treat heart failure and high blood pressure by causing the blood vessels to expand. Lisinopril is one example.
- *Angiotensin II receptor blockers:* These can also control blood pressure. Losartan is one example.

- *Angiotensin receptor neprilysin inhibitors:* These can help unload the heart and interrupt the chemical pathways that weaken it.
- *Beta-blockers:* Metoprolol and other medications in this class can reduce the heart rate and lower blood pressure. They can also treat arrhythmias and angina.
- *Calcium channel blockers:* These can lower blood pressure and prevent arrhythmias by reducing the pumping strength of the heart and relaxing the blood vessels. One example is diltiazem (Cardizem).
- *Cholesterol-lowering medications:* Statins, such as atorvastatin (Lipitor), and other types of drugs can help reduce levels of low-density lipoprotein cholesterol in the body.
- *Digitalis:* Preparations such as digoxin (Lanoxin) can increase the strength of the heart's pumping action. They can also help treat heart failure and arrhythmias.
- *Diuretics:* These medications can reduce the heart's workload, lower blood pressure, and remove excess water from the body. Furosemide (Lasix) is one example.
- *Vasodilators:* These are medications to lower blood pressure. They do this by relaxing the blood vessels. Nitroglycerin (Nitrostat) is one example. These medications can also help ease chest pain.

1.6.2. Surgery Methods

Undergoing heart surgery can help treat blockages and heart problems when medications are not effective. Some common types of surgery include:

- *Coronary artery bypass surgery:* This allows blood flow to reach a part of the heart when an artery is blocked. Coronary artery bypass grafting is the most common surgery. A surgeon can use a healthy blood vessel from another part of the body to repair a blocked one.
- *Coronary angiography:* This is a procedure that widens narrow or blocked coronary arteries. It is often combined with the insertion of a stent, which is a wire-mesh tube that allows easier blood flow.
- *Valve replacement or repair:* A surgeon can replace or repair a valve that is not functioning correctly.

- *Repair surgery:* A surgeon can repair congenital heart defects, aneurysms, and other problems.
- *Device implantation:* Pacemakers, balloon catheters, and other devices can help regulate the heartbeat and support blood flow.
- *Laser treatment:* Transmyocardial laser revascularization can help treat angina.
- *Maze surgery:* A surgeon can create new paths for electrical signals to pass through. This can help treat atrial fibrillation.

1.6.Preventive Measures

Some lifestyle measures can help reduce the risk of heart disease. Some preventive measures are as follows:

- *Eating a balanced diet:* Opt for a heart-healthy diet that is rich in fiber and favors whole grains and fresh fruits and vegetables. The Mediterranean diet and the DASH diet may be good for heart health. Also, it may help to limit the intake of processed foods and add fat, salt, and sugar.
- *Exercising regularly:* This can help strengthen the heart and circulatory system, reduce cholesterol, and maintain blood pressure. A person may wish to aim for 150 minutes of exercise per week.
- *Maintaining a moderate body weight:* A healthy body mass index (BMI) is typically between 20 and 25.
- *Quitting or avoiding smoking:* Smoking is a major risk factor for heart and cardiovascular conditions.
- *Limiting alcohol intake:* Women should consume no more than one standard drink per day, and men should consume no more than two standard drinks per day.
- *Managing underlying conditions:* Seek treatment for conditions that affect heart health, such as high blood pressure, obesity, and diabetes.

1.7. FACTORS INVOLVED IN HEART ATI'ACK

The given below are the factors involved in the data taken, which are used to predict the dependent variable.

1.7.1. Chest Pain Type

Angina is chest pain or discomfort that occurs when our heart doesn't get as much blood and oxygen as it needs. In angina, the need for increased blood flow isn't met for a short time. When increased demand for blood goes away, angina symptoms go away too. Angina and heart attack have the same root cause: atherosclerosis. This is the buildup of fatty substances (plaque) in the coronary arteries. If one or more arteries are partly clogged, not enough blood can flow through, and one can feel chest pain or discomfort. While the pain of angina may come and go, it's a sign of heart disease and can be treated. Lifestyle changes, medications, medical procedures, and surgery can help reduce angina. Angina is of four types. They are as follows:

- Typical Angina
- Atypical Angina
- Non-Anginal Pain
- Asymptomatic

1.7.2. Resting Blood Pressure

Blood pressure is the force of blood pushing against the walls of arteries as the heart pumps blood. When a health care professional measures our blood pressure, they use a blood pressure cuff around our arm that gradually tightens. The results are given in Bvo numbers. The first number, called systolic blood pressure, is the pressure caused by our heart contracting and pushing out blood. The second number, called diastolic blood pressure, is the pressure when our heart relaxes and fills with blood. A blood pressure reading is given as the systolic blood pressure number over the diastolic blood pressure number. Blood pressure levels are classified based on those two numbers.

Low Blood Pressure, or hypotension, is the systolic blood pressure lower than 90 or diastolic blood pressure lower than 60. If you have low blood pressure, you may feel lightheaded, weak, dizzy, or even faint. It can be caused by not getting enough fluids, blood loss, some medical conditions, or medications, including those prescribed for high blood pressure.

Normal blood pressure for most adults is defined as a systolic pressure of less than 120 and a diastolic pressure of less than 80.

Elevated blood pressure is defined as a systolic pressure between 120 and 129 with a diastolic pressure of less than 80.

High blood pressure is defined as 130 or higher for the first number, or 80 or higher for the second number.

1.7.3. Cholesterol

Cholesterol is a waxy substance. It's not inherently "bad." Our body needs it to build cells and make vitamins and other hormones. But too much cholesterol can pose a problem cholesterol comes from two sources. Our liver makes all the cholesterol we need. The remainder of the cholesterol in our body comes from foods and animals. For example, meat, poultry, and dairy products all contain dietary cholesterol. Cholesterol circulates in the blood. As the amount of cholesterol in our blood increases, so does the risk to our health. High cholesterol contributes to a higher risk of cardiovascular diseases, such as heart disease and stroke.

The two types of cholesterol are LDL cholesterol, which is bad, and HDL cholesterol which is good. Too much of the bad kind, or not enough of the good kind, increases the risk cholesterol will slowly build up in the inner walls of the arteries that feed the heart and brain. High cholesterol is one of the major controllable risk factors for coronary heart disease, heart attack, and stroke. If one has other risk factors such as smoking, high blood pressure, or diabetes, his risk increases, even more.

1.7.4. Fasting Blood Sugar

A test to determine how much sugar is in a blood sample after an overnight fast. The fasting blood sugar test is commonly used to detect diabetes mellitus. A blood sample is taken in a lab, physician's office, or hospital. The test is done in the morning before the person has eaten. The normal range for blood glucose is 70 to 100 mg /dl. Levels between 100 and 126 mg/dl are referred to as impaired fasting glucose or pre-diabetes. Diabetes is typically diagnosed when fasting blood glucose levels are 126 mg/dl or higher.

1.7.5. Resting Electrocardiographic Results

Resting Electrocardiography (ECG) is a non-invasive test that can detect abnormalities including arrhythmias, evidence of coronary heart disease, left ventricular hypertrophy, and bundle branch blocks. In the preoperative setting, resting ECG is used to assess known cardiovascular diseases, detect previously undiagnosed cardiovascular diseases, and provide a baseline standard against which to measure changes in the postoperative period.

When conducted by a suitably trained individual, the resting ECG is simple to perform and interpret, with the only described complication being a minor allergy to the ECG electrodes resulting in self-limiting skin reddening. However, there is uncertainty regarding the prognostic significance of different resting ECG abnormalities in the perioperative setting, especially in asymptomatic patients.

- *Thalach:* The person's maximum heart rate achieved. It is measured as 60-100 is normal, and >100 is Tachycardia
- *Exang:* Exang means exercise-induced angina. Angina is a pain in the chest that comes on with exercise, stress, or other things that make the heart work harder. It is an extremely common symptom of coronary artery disease, which is caused by cholesterol-clogged coronary arteries. This is the network of arteries that nourish the heart muscle,
- *Oldpeak:* Old peak means that the exercise is relative to rest. ST depression induced by exercise relative to rest. ST relates to positions on the ECG plot. The old peak ranges between 0-6.
- *Slope:* The ST segment shift relative to exercise-induced increments in heart rate, the ST/heart rate slope, has been proposed as a more accurate ECG criterion for diagnosing significant coronary artery disease. It is defined as, 0 = Downsloping, 1 = Flat, 2 = Upsloping.

1.7.6. Calcium Heart Score (CA)

A calcium heart score test is performed to identify the amount of calcified plaque in the heart's arteries. An elevated calcium heart score may be an indication that we need to make certain lifestyle adjustments to reduce our risk of a heart attack. Calcium heart scores

are also referred to as Angaston scores. It ranges from 0 and above. Higher scores indicate greater evidence of plaque deposits inside the lining of our heart arteries.

- 0 = (0) no plaque detected (risk of coronary artery disease is very low — less than 5%)
- 1 = (1-100) calcium detected in extremely minimal levels (risk of coronary diseases is still low — less than 10%)
- 2 = (11-100) mild levels of plaque detected with certainty (minimal narrowing of heart arteries is likely)
- 3 = (101-300) moderate levels of plaque detected (relatively high risk of a heart attack within 3-5 years)
- 4 = (300-400) extensive levels of plaque detected (very high risk of heart attack, high levels of vascular disease are present)

1.7.7. Thalassemia (Thal)

A blood disorder called thalassemia. Thalassemia is an inherited blood disorder that causes our body to have less hemoglobin than normal. Hemoglobin enables red blood cells to carry oxygen. It is valued as follows:

- 1 = Normal (Normal blood flow)
- 2 = Fixed Defect (no blood flow in some parts of the heart)
- 3 = Reversible Defect (a blood flow is observed but it is not normal)

1.8. BIOSTATISTICS

Biostatistics is a branch of science that deals with the application of statistics in the field of medicine and health care. It is a statistical science that focuses on the collection, analysis, interpretation, and presentation of biological data. Biostatistics is used to develop and test hypotheses, design experiments and analyze data. It is used to identify and quantify relationships between different factors and to identify patterns in data. It also helps to make predictions about the possible outcomes of health interventions and to determine the effectiveness of treatments.

Biostatistics is used to evaluate the safety and efficacy of new drugs and treatments, to measure the effectiveness of health interventions, to assess the impact of environmental

factors on health, and to identify risk factors for diseases. It is also used to analyze the impact of health policies and programs. Statistical methods allow for the organization and interpretation of large amounts of data, helping to make sense of what might otherwise seem to be chaotic or random.

Bioinformatics has its roots in biostatistics, which dates back to the 19th century. The earliest forms of biostatistics involved the collection, tabulation, and analysis of data from large groups of people. This data was used to inform public health policies and medical decision-making. In the early 20th century, biostatisticians developed methods of statistical inference, such as the chi-square test, t-test, and linear regression. These methods were used to analyze data and draw conclusions about the effectiveness of medical treatments and public health policies.

The advent of computers in the 1950s allowed biostatisticians to process data more quickly and efficiently, leading to the development of more sophisticated methods of data analysis. In the 1970s, bioinformatics emerged as an interdisciplinary field that combined the use of computers with biostatistics and other areas of biology. Bioinformatics uses algorithms and computer programs to analyze biological data, such as DNA sequences, gene expression data, and protein structures. Today, bioinformatics is used in a variety of fields, including genomics, proteomics, drug discovery, and medical research.

Some areas of application of biostatistics are:

1. ***Clinical Trials:*** Biostatistics is widely used in clinical trials to measure the safety and efficacy of treatments or interventions. Biostatisticians analyze data from clinical trials to determine whether a treatment has a statistically significant effect, and to measure its potential benefits and risks.
2. ***Epidemiology:*** Biostatistics is used in epidemiology to assess the risk of diseases and their possible causes. It is used to measure the incidence of diseases, analyze the data from case-control studies, and identify potential risk factors for diseases.
3. ***Public Health:*** Biostatistics is used in public health to monitor the health of populations and to identify and address health disparities. It is used to analyze data from surveys,

measure the effectiveness of public health interventions, and assess the impact of environmental factors on health.

4. ***Genetics:*** Biostatistics is used in genetics to identify gene-disease associations and to analyze the data from genome-wide association studies. It is also used to assess the heritability of diseases and to identify genetic markers for disease susceptibility.
5. ***Agriculture:*** Biostatistics is used in agriculture to analyze the data from field trials, measure the yield of crops, and identify the factors that affect crop production. It is also used to analyze the genetic data from livestock, assess the performance of farm animals, and identify traits that are associated with improved animal health.

1.9. PRELIMINARIES OF BIOSTATISTICS:

Basic concepts and terminology of biostatistics, including probability distributions, hypothesis testing, and sample size determination. Descriptive statistics is a set of techniques used to summarize and present data. It is used to describe the characteristics of a given data set, such as the mean, median, mode, and standard deviation. Descriptive statistics can also be used to present the data visually, such as in the form of charts and graphs. Descriptive statistics can help identify patterns, relationships, and trends in data. It can be used to compare the results of different studies and to analyze the effectiveness of different strategies. Descriptive statistics can be used to make predictions and decisions based on the data. Exploratory data analysis, including graphical and numerical methods for exploring data.

Inferential statistics is a type of statistical analysis that is used to make inferences and predictions about a population based on a sample of data. It is used to draw conclusions and make generalizations from a sample to a larger population. It is used to test hypotheses and make predictions about the population. By using inferential statistics, one can make statements about the population based on the sample.

Design of experiments (DOE) is a systematic approach to designing and conducting experiments to test hypotheses and gain insights into the behavior of a system. It involves selecting the most appropriate experimental design, analyzing the results, and drawing valid conclusions from the data. DOE is commonly used in the fields of engineering, manufacturing, and the sciences to improve products, processes, and services. DOE can be used to identify the most effective factors in a process, the optimal levels of those factors, and

the interactions between them. It is also used to optimize the design of experiments, reduce costs, and improve the efficiency of experiments.

Regression is a form of predictive analytics that looks at the relationship between a dependent variable (the one that you want to predict) and one or more independent variables (the ones that you use to predict the dependent variable). It is used to identify the underlying trends in data and can be used to make predictions about future values of the dependent variable. Regression is used in a variety of fields including economics, finance, marketing, and medicine. Classification, including supervised and unsupervised machine learning methods.

Survival analysis is a type of statistical analysis used to study the time until an event occurs. It is commonly used to study the time until death or other medical events, such as cancer recurrence, but can also be used to study the time until other events, such as job loss or marriage. Survival analysis is used to estimate the probability of an event occurring at a certain point in time and to compare different groups or treatments in terms of their effect on the time until the event occurs. It is used to identify risk factors that may be associated with an increased or decreased risk of an event occurring.

Bayesian statistics is a type of statistical analysis that relies on probability theory to make inferences about a given data set. It is based on the Bayes Theorem and uses prior knowledge and data to conclude the current data. It is particularly useful for making predictions based on past data. It allows for more accurate and reliable predictions than traditional statistics because it takes into account the uncertainty of the data. Additionally, it can be used to update beliefs and make decisions when new information is available.

1.10. ADVANTAGE OF BIOSTATISTICS

1. Biostatistics is a vital tool for healthcare research. It helps researchers to design, analyze, and interpret data from medical studies and surveys. This type of statistical analysis can be used to determine the effectiveness of certain treatments and medications, to study the prevalence of certain diseases, and to assess the impact of lifestyle and environmental factors on health. Additionally, biostatistics is used to create predictive models, which can

help healthcare professionals anticipate potential problems and make decisions about patient care.

2. Biostatistics can also provide insights that can help healthcare professionals to develop better screening tests, treatments, and preventive measures. Biostatistics is an essential tool used in healthcare research. It helps in analyzing diseases, treatments, and outcomes. It helps in the design of clinical trials and observational studies, data analysis, and interpretation of results.
3. Biostatistics is a branch of statistics that focuses on the collection, analysis, and interpretation of biological data. It is used to support decision-making in a variety of fields, including medicine, public health, and epidemiology. By providing an understanding of the underlying data and evidence for a given situation, biostatistical analysis helps inform decisions and improves outcomes.
4. Biostatistics can be used to design and evaluate studies, analyze data, create models to predict risk, identify trends and correlations, measure the effectiveness of treatments, and provide guidance on the best course of action. This data-driven approach to decision-making helps to reduce uncertainty and maximize the potential for successful outcomes.
5. Biostatistics is a field of study that involves the use of statistical methods, theories, and principles to gain insights into biological problems. It can be used to analyze data from clinical trials, laboratory experiments, and epidemiology studies. In doing so, biostatistics helps to determine the effectiveness of treatments, uncover the risk factors associated with a certain disease, and measure the impact of various treatments and interventions on health outcomes. By providing an accurate view of health and medical data, biostatistics can help to ensure that accurate results are obtained from clinical trials and other research studies.
6. Biostatistics helps to improve the quality of life by helping to identify and analyze patterns in health data. This includes identifying areas of high risk and identifying potential health interventions that can be used to address these areas.
7. Biostatistics can also be used to evaluate the effectiveness of interventions, identify trends in health outcomes, and create predictive models that can be used to make decisions about future health initiatives. By doing this, biostatistics can help to identify potential health

risks and provide evidence-based solutions that can lead to improved health outcomes and quality of life.

1.11. DISADVANTAGES OF BIOSTATISTICS

1. Biostatistics relies heavily on assumptions and statistical models that may not always accurately reflect the true nature of the data. This can lead to invalid conclusions and incorrect assumptions.
2. Biostatistics is heavily reliant on computerized algorithms, which can be difficult to interpret and understand. As such, it can be difficult to trust the results of biostatistical analysis without a thorough understanding of the algorithms used.
3. Biostatistics is heavily reliant on the data available, and the quality of that data can affect the accuracy of the analysis. If the data is incomplete, inaccurate, or biased, it can lead to incorrect conclusions.
4. A biostatistical analysis is also subject to measurement error, which can lead to incorrect interpretations of the data. If a study is not designed properly, or if the data is collected improperly, it can lead to problems in the analysis.
5. Finally, biostatistics is still a relatively new field and is constantly evolving. As such, it can be difficult to keep up with the latest developments and to know what methods are the most reliable and accurate.

OBJECTIVE OF THE STUDY

1. To display the data in the form of tables and diagrams.
2. To build the Regression model for the data using Logistic Regression.
3. To test whether the cholesterol level is under control level or not by using a t-test.
4. To classify the heart disease data by using the decision tree model.
5. To construct the Machine Learning model using a decision tree algorithm for the heart disease classification data.

Chapter 2

METHODOLOGY

The methodology is a set of principles and practices used to guide research and solve problems. It includes the tools, techniques, and processes used to systematically investigate a problem, collect data, analyze information, and draw conclusions. The methodology can be used in a variety of areas, including psychology, sociology, education, economics, and business.

2.1. DATA COLLECTION

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The data collection process involves the identification of data sources, the collection of data from those sources, and the organization and storage of the data. Data collection can be divided into two categories that are primary and secondary. Primary data collection is the process of collecting data directly from sources, such as surveys, interviews, experiments, and field observation. Secondary data collection involves collecting data from existing sources such as published books, newspapers, and online databases. The choice of data collection method depends on the type of data needed. For instance, surveys are best for collecting quantitative data, while interviews and field observations are better suited for collecting qualitative data. Additionally, different data collection methods can be used in combination to provide a more comprehensive analysis. Before beginning the data collection process, it is important to establish clear objectives and develop a data collection plan. The plan should outline the research design, the data sources, the data collection methods to be used, the timeline for data collection, and any logistical considerations, such as budget and staffing requirements. Furthermore, it is important to consider ethical considerations when collecting data,

2.2. DESCRIPTIVE STATISTICS

Descriptive statistics are used to summarize, organize, and interpret data. It includes methods such as measures of central tendency (mean, median, and mode) and measures of variability (range, quartiles, and standard deviation). Descriptive statistics are used to describe the basic features of a data set in quantitative terms. Descriptive statistics

can be used to provide a concise summary of a large data set. It can also be used to detect patterns and trends in the data and to make predictions based on the data. Descriptive statistics can be used to compare different data sets and to identify relationships between variables. Descriptive statistics can also be used to develop and test hypotheses. For example, a researcher may use descriptive statistics to determine if there is a correlation between two variables. Once a hypothesis is tested, descriptive statistics can be used to interpret the results and draw conclusions. Descriptive statistics are essential tools for data analysis and interpretation. They are used to make sense of complex data sets and to answer important research questions. Descriptive statistics provide a way to organize and interpret data, making it easier for researchers to draw meaningful conclusions about their research.

2.3. MEASURES OF CENTRAL TENDENCY

Central tendency is a way of describing the "center" of a population or sample of data. It is a measure of what is typical or average in the data. The three most common measures of central tendency are the mean, median, and mode. The mean is the average of all of the values in the data set. It is calculated by adding all of the values in a data set and dividing by the number of values. The mean is the most commonly used measure of central tendency and is often simply referred to as the "average". The median is the middle value in a data set when the values are arranged in numerical order. It is the value that is exactly in the middle of the data set.

If there are an even number of values, the median is the average of the two middle values. The median is less affected by extreme values (outliers) than the mean. The mode is the most frequently occurring value in a data set. It is the value that appears most often in the data set. A data set can have more than one mode if multiple values appear frequently. These three measures of central tendency can be used to describe the center of the distribution of a population or sample of data.

2.4. MEASURES OF DISPERSION

Measures of dispersion are statistical tools used to describe the spread of a set of data points. They provide a quantitative measure of how much the data points vary from each other and the mean. Measures of dispersion are useful in helping to identify trends in the data and to compare different data sets. Common measures of dispersion include the range, interquartile range (IQR), standard deviation, and variance.

The range is the simplest measure of dispersion. It is calculated by subtracting the smallest value from the largest value in the data set. It is sensitive to outliers, meaning that one particularly high or low value can drastically increase the range.

The interquartile range (IQR) is another measure of dispersion. It is calculated by subtracting the 25th percentile from the 75th percentile of the data set. The IQR is less sensitive to outliers than the range, as it ignores the extremes at the top and bottom of the data set.

The standard deviation is a measure of dispersion that is based on the distance between each data point and the mean. It is calculated by taking the square root of the variance. The standard deviation is a measure of how much the data points vary from the mean. It is less sensitive to outliers than the range and IQR, as it takes all data points into account.

The variance is the square of the standard deviation. It is calculated by taking the average of the squared distances between each data point and the mean. The variance is a measure of how much the data points vary from each other and the mean. It is less sensitive to outliers than the range and IQR, as it takes all data points into account.

In summary, measures of dispersion provide a quantitative measure of how much the data points vary from each other and the mean. The range, interquartile range (IQR), standard deviation, and variance are all common measures of dispersion. The range is the simplest measure of dispersion, but it is sensitive to outliers. The IQR is less sensitive to outliers than the range, as it ignores the extremes at the top and bottom of the data set. The standard deviation and variance are less sensitive to outliers than the range and IQR, as they take all data points into account.

2.5. GRAPHICAL / PICTORIAL METHODS

Graphical is an adjective that refers to something related to graphical representation or displaying of data or objects. Graphical representation is a visual representation of information or data using lines, bars, pie charts, etc. Graphical representations are used to make data easier to understand and interpret.

2.5.1. HISTOGRAMS

A histogram is a graphical representation of the distribution of a dataset. It is a way to visualize the frequency of observations or measurements that fall within specified intervals, or "bins," of a continuous variable.

In a histogram, the x-axis represents the range of values for the variable being measured, and the y-axis represents the frequency or count of observations that fall within each bin. Each bin is typically of equal width, although this may vary depending on the data and the purpose of the histogram.

Histograms are commonly used in statistics, data analysis, and data visualization to understand the distribution of data, identify patterns or outliers, and explore relationships between variables.

2.5.2. BAR GRAPH

A bar graph is a type of graph that displays data using rectangular bars. It is typically used to visualize the relative frequency of different categories or groups of data. The length of each bar is proportional to the value it represents. Bar graphs are useful for comparing different values, illustrating trends, and showing relationships between different data sets.

2.5.3. MULTIPLE BAR GRAPH

A multiple-bar graph is a type of graph that is used to compare the values of more than one group or data set. It is a useful tool for displaying the relationships between different data sets and for highlighting the differences and similarities between them. Multiple bar graphs are often used in scientific research, business analysis, and other data-driven scenarios.

2.5.4. PIE CHART

A pie chart is a type of graph that is used to display data in a circular format, where each category is represented by a portion of a circle. The size of each portion is determined by the relative size of each category compared to the total. Pie charts are often used to represent percentages or proportions visually. They are also used to compare different data sets or to show the relationship between different parts of a whole. Pie charts

can be used to easily identify patterns in data, as they allow viewers to quickly compare different categories of data at a glance.

2.6. CORRELATION

Correlation is a statistical measure that represents the degree of association or relationship between two variables. It tells us how two variables are related and whether changes in one variable are associated with changes in the other variable. Correlation can be positive, negative, or zero.

A positive correlation means that as one variable increases, the other variable also tends to increase. For example, there may be a positive correlation between a person's age and income, as older people tend to have more work experience and earn higher salaries.

A negative correlation means that as one variable increases, the other variable tends to decrease. For example, there may be a negative correlation between the amount of exercise a person gets and their body weight, as people who exercise more tend to weigh less.

A zero correlation means that there is no apparent relationship between the two variables. For example, there may be no correlation between a person's hair color and their intelligence.

Correlation is often expressed as a correlation coefficient, which is a number between -1 and +1 that represents the strength and direction of the relationship between two variables. A correlation coefficient of +1 represents a perfect positive correlation, while a correlation coefficient of -1 represents a perfect negative correlation. A correlation coefficient of zero represents no correlation.

It can help researchers understand the relationship between different variables and make predictions based on that relationship. However, it is important to note that correlation does not necessarily imply causation and further research is often needed to establish a causal relationship between variables.

Correlation heat maps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship. A correlation plot typically contains several numerical variables, with each variable represented by a column. The

rows represent the relationship between each pair of variables. The values in the cells indicate the strength of the relationship, with positive values indicating a positive relationship and negative values indicating a negative relationship.

Correlation heat maps can be used to find potential relationships between variables and to understand the strength of these relationships. In addition, correlation plots can be used to identify outliers and detect linear and nonlinear relationships. The color-coding of the cells makes it easy to identify relationships between variables at a glance. Correlation heat maps can be used to find both linear and nonlinear relationships between variables.

2.7. ONE SAMPLE T-TEST

The One Sample t Test (also known as t-test) is a parametric statistical test used to compare a sample mean to a known population mean. It is most commonly used to determine whether a sample of observations is representative of a larger population. The One Sample t-Test is used when a researcher is interested in evaluating the mean of a population but only has access to a single sample of data from that population.

The One Sample t-Test is a hypothesis testing procedure used to determine if a sample mean differs significantly from a known population means. The test is based on the Student's t-distribution, which is a continuous probability distribution that is symmetric around the mean. The t-distribution is used when the sample size is small ($n<30$), or when the population variance is unknown.

To conduct a One-Sample t-Test, the researcher first needs to state a null and alternative hypothesis. The null hypothesis states that the population mean is equal to the known value, while the alternative hypothesis states that the population mean is different from the known value. Next, the researcher calculates the sample mean and the standard deviation. The sample means and standard deviation are then used to calculate the t-statistic, which is the difference between the sample mean and the population means, divided by the standard error of the mean.

Finally, the researcher compares the calculated t-statistic to the critical value of the t-distribution to determine if the difference between the sample mean and the population mean is statistically significant. If the calculated t-statistic is larger than the critical value, then the researcher can reject the null hypothesis and conclude that the population means is different from the known value. If the calculated t-statistic is smaller than the critical

value, then the researcher cannot reject the null hypothesis and must conclude that the population means is equal to the known value.

2.8. LOGISTIC REGRESSION

Logistic Regression is a supervised machine learning algorithm used for classification problems. It works by taking a set of input features and estimating the probability of an output class. The estimated probability is then compared to a threshold value, and if the probability is greater than the threshold value, the output class is predicted. It is one of the most widely used algorithms in the field of machine learning.

Logistic Regression is a linear model, which means that it uses linear combinations of the input features to predict the probability of an output class. It is based on the assumption that the relationship between the input features and the output class can be represented by a linear function. The function is then fit to the data using a maximum likelihood estimation. The main goal of logistic regression is to find the best-fitting model for the data. To do this, the model is first initialized with a set of weights. These weights are then adjusted based on the data using an optimization algorithm. The optimized weights are then used to make predictions on new data.

Logistic regression can be used for a variety of tasks, such as predicting the probability of an event occurring (e.g. whether a customer will buy a product), classifying data into different categories (e.g. whether a customer is male or female), and estimating the probability of a dependent variable given a set of independent variables (e.g. predicting the probability of a customer buying a product given their age and income).

Logistic regression is a powerful and flexible algorithm that can be used for a variety of tasks. It is relatively simple to implement and can be used with both large and small datasets. It is also relatively fast to train, so it can be used in real-time applications.

2.8.1. ASSUMPTION OF LOGISTIC MODEL

Logistic regression is a powerful statistical model used for prediction and classification. It is widely used in fields such as medicine, finance, and marketing. The assumptions of logistic regression are important for accurate and reliable results.

1. A linear relationship between the independent and the dependent variable: The relationship between the independent variable and the dependent variable should be linear.
2. Multivariate normality: The data should be distributed normally.
3. No multicollinearity: There should be no correlation between the independent variables.
4. No auto-correlation: The observations should be independent of one another.
5. No omitted variable bias: All important independent variables should be included in the model.
6. Independence of errors: The errors should be independent of one another.
7. Homoscedasticity: The variance of the errors should be constant across all values of the independent variables. These assumptions are necessary for logistic regression to be accurate and reliable. If these assumptions are violated, the results of the logistic regression may be biased and unreliable.

2.9. MACHINE LEARNING MODEL

Machine learning is a branch of artificial intelligence that uses algorithms and statistical models to enable computers to learn from data and improve their performance on a given task. The main goal of machine learning is to develop models that can learn and make predictions or decisions based on data, similar to how humans learn from experience. The development of machine learning algorithms is based on statistical methods, which are used to build models that can learn from data and make predictions or decisions. There are three main types of machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning involves training a model on a labeled dataset, where the desired output is already known. The algorithm learns to map inputs to outputs by finding patterns in the data. Unsupervised learning, on the other hand, involves training a model on an unlabeled dataset, where the desired output is not known. The algorithm must find patterns or structures in the data without any prior knowledge of what it is looking for. Reinforcement learning involves training a model to interact with an environment and maximize a reward signal.

Machine learning has numerous applications in statistics, including image and speech recognition, natural language processing, predictive modeling, and anomaly detection. It is also used in industries such as finance, healthcare, and manufacturing to automate processes and make more accurate predictions.

Some popular machine learning algorithms used in statistics include decision trees, random forests, Naive Bayes, support vector machines (SVMs), and neural networks. These algorithms are used to build models that can make predictions or decisions based on the data they are trained on.

To use machine learning effectively in statistics, it is important to have a good understanding of the data, as well as the algorithms and tools. Also need to have a clear understanding of the problem you are trying to solve and what outcomes you are looking to achieve. With the right data, algorithms, and tools, machine learning can be a powerful tool for solving complex problems and making better decisions.

2.9.1. SUPERVISED LEARNING

Supervised learning is a type of machine learning algorithm that involves training a model on a labeled dataset, where the desired output is already known. The goal of supervised learning is to develop a model that can accurately predict the output for new, unseen data. Supervised learning involves two main types of problems: classification and regression.

Classification: In classification problems, the goal is to predict a categorical output variable based on one or more input variables. The model is trained on a labeled dataset, where each input is associated with a categorical output. The goal is to learn a mapping between the input variables and the output categories so that the model can accurately classify new, unseen data. Examples of classification problems include spam detection, image classification, and sentiment analysis.

Regression: In regression problems, the goal is to predict a continuous output variable based on one or more input variables. The model is trained on a labeled dataset, where each input is associated with a continuous output. The goal is to learn a mapping between the input variables and the output values so that the model can accurately predict the output for new, unseen data. Examples of regression problems include stock price

prediction, housing price prediction, and demand forecasting. The process of supervised learning typically involves several steps:

Data Preparation: The first step is to prepare the dataset by cleaning, preprocessing, and transforming the data into a format that can be used by the machine learning algorithm.

Model Selection: The next step is to select an appropriate model for the problem at hand. Many different types of models can be used for supervised learning, including decision trees, logistic regression, support vector machines, and neural networks.

Training the Model: The model is then trained on the labeled dataset, using an optimization algorithm to find the best set of parameters that minimize the error between the predicted output and the actual output.

Model Evaluation: The final step is to evaluate the performance of the model on a separate, unseen dataset. The performance is typically measured using metrics such as accuracy, precision, recall, and F1 score.

Supervised learning is widely used in many industries, including healthcare, finance, marketing, and engineering. It has many applications, such as fraud detection, customer segmentation, predictive maintenance, and medical diagnosis. With the right data and a well-designed model, supervised learning can be a powerful tool for solving complex problems and making better decisions.

2.9.2. UNSUPERVISED LEARNING

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information makes it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction. Below we'll define each learning method and highlight common algorithms and approaches to conduct them effectively.

Clustering: Clustering is a data mining technique that groups unlabeled data based on their similarities or differences. Clustering algorithms are used to process raw, unclassified

data objects into groups represented by structures or patterns in the information. Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic.

Exclusive and Overlapping Clustering: Exclusive clustering is a form of grouping that stipulates a data point can exist only in one cluster. This can also be referred to as "hard" clustering. The K-means clustering algorithm is an example of exclusive clustering. K-means clustering is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group's centroid. The data points closest to a given centroid will be clustered under the same category. A larger K value will be indicative of smaller groupings with more granularities whereas a smaller K value will have larger groupings and fewer granularities. K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression. Overlapping clusters differ from exclusive clustering in that it allows data points to belong to multiple clusters with separate degrees of membership. "Soft" or fuzzy k-means clustering is an example of overlapping clustering.

Hierarchical clustering: Hierarchical clustering, also known as hierarchical cluster analysis (HCA), is an unsupervised clustering algorithm that can be categorized in two ways; they can be agglomerative or divisive. Agglomerative clustering is considered a "bottom-up approach". Its data points are isolated as separate groupings initially, and then they are merged iteratively based on similarity until one cluster has been achieved. Four different methods are commonly used to measure similarity:

- *Ward's linkage:* This method states that the distance between two clusters is defined by the increase in the sum of squared after the clusters are merged.
- *Average linkage:* This method is defined by the mean distance between two points in each cluster
- *Complete (or maximum) linkage:* This method is defined by the maximum distance between two points in each cluster
- *Single (or minimum) linkage:* This method is defined by the minimum distance between two points in each cluster

Euclidean distance is the most common metric used to calculate these distances; however, other metrics, such as Manhattan distance, are also cited in clustering literature.

Divisive clustering can be defined as the opposite of agglomerative clustering; instead, it takes a "top-down" approach. In this case, a single data cluster is divided based on the differences between data points. Divisive clustering is not commonly used, but it is still worth noting in the context of hierarchical clustering. These clustering processes are usually visualized using a dendrogram, a tree-like diagram that documents the merging or splitting of data points at each iteration.

Probabilistic clustering: A probabilistic model is an unsupervised technique that helps us solve density estimation or "soft" clustering problems. In probabilistic clustering, data points are clustered based on the likelihood that they belong to a particular distribution. The Gaussian Mixture Model (GMM) is one of the most commonly used probabilistic clustering methods. Gaussian Mixture Models are classified as mixture models, which means that they are made up of an unspecified number of probability distribution functions. GMMs are primarily leveraged to determine which Gaussian, or normal, probability distribution a given data point belongs to. If the mean or variance is known, then we can determine which distribution a given data point belongs to. However, in GMMs, these variables are not known, so we assume that a latent, or hidden, variable exists to cluster data points appropriately. While it is not required to use the Expectation-Maximization (EM) algorithm, it is commonly used to estimate the assignment probabilities for a given data point to a particular data cluster.

Association Rules: An association rule is a rule-based method for finding relationships between variables in a given dataset. These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products. Understanding the consumption habits of customers enables businesses to develop better cross-selling strategies and recommendation engines. While there are a few different algorithms used to generate association rules, such as Apriori, Eclat, and FP-Growth, the Apriori algorithm is the most widely used.

Apriori algorithms: Apriori algorithms have been popularized through market basket analyses, leading to different recommendation engines for music platforms and online retailers. They are used within transactional datasets to identify frequent item sets, or collections of items, to identify the likelihood of consuming a product given the consumption of another product. Apriori algorithms use a hash tree to count item sets, navigating through the dataset in a breadth-first manner.

Dimensionality reduction: While more data generally yields more accurate results, it can also impact the performance of machine learning algorithms (e.g. over fitting) and it can also make it difficult to visualize datasets. Dimensionality reduction is a technique used when the number of features, or dimensions, in a given dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the integrity of the dataset as much as possible. It is commonly used in the preprocessing data stage, and there are a few different dimensionality reduction methods that can be used, such as:

Principal component analysis: Principal component analysis (PCA) is a type of dimensionality reduction algorithm which is used to reduce redundancies and compress datasets through feature extraction. This method uses a linear transformation to create a new data representation, yielding a set of "principal components." The first principal component is the direction which maximizes the variance of the dataset. While the second principal component also finds the maximum variance in the data, it is completely uncorrelated to the first principal component, yielding a direction that is perpendicular, or orthogonal, to the first component. This process repeats based on the number of dimensions, where the next principal component is the direction orthogonal to the prior components with the most variance.

Singular value decomposition: Singular value decomposition (SVD) is another dimensionality reduction approach that factorizes a matrix, A, into three, low-rank matrices. SVD is denoted by the formula, A = USVT, where U and V are orthogonal matrices. S is a diagonal matrix, and S values are considered singular values of matrix A. Similar to PCA, it is commonly used to reduce noise and compress data, such as image files.

Autoencoders: Autoencoders leverage neural networks to compress data and then recreate a new representation of the original data's input. The hidden layer specifically acts as a bottleneck to compress the input layer before reconstructing it within the output layer. The stage from the input layer to the hidden layer is referred to as "encoding" while the stage from the hidden layer to the output layer is known as "decoding."

Applications of unsupervised learning: Machine learning techniques have become a common method to improve a product user experience and to test systems for quality assurance. Unsupervised learning provides an exploratory path to view data, allowing businesses to identify patterns in large volumes of data more quickly when compared to

manual observation. Some of the most common real-world applications of unsupervised learning are:

- *News Sections:* Google News uses unsupervised learning to categorize articles on the same story from various online news outlets.
- *Computer vision:* Unsupervised learning algorithms are used for visual perception tasks, such as object recognition.
- *Medical imaging:* Unsupervised machine learning provides essential features to medical imaging devices, such as image detection, classification, and segmentation, used in radiology and pathology to diagnose patients quickly and accurately.
- *Anomaly detection:* Unsupervised learning models can comb through large amounts of data and discover atypical data points within a dataset. These anomalies can raise awareness around faulty equipment, human error, or breaches in security.
- *Customer personas:* Defining customer personas makes it easier to understand common traits and business clients' purchasing habits. Unsupervised learning allows businesses to build better buyer persona profiles, enabling organizations to align their product messaging more appropriately.
- *Recommendation Engines:* Using past purchase behavior data, unsupervised learning can help to discover data trends that can be used to develop more effective cross-selling strategies. This is used to make relevant add-on recommendations to customers during the checkout process for online retailers.

Unsupervised vs. supervised vs. semi-supervised learning

Unsupervised learning and supervised learning are frequently discussed together. Unlike unsupervised learning algorithms, supervised learning algorithms use labeled data. From that data, it either predicts future outcomes or assigns data to specific categories based on the regression or classification problem that it is trying to solve. While supervised learning algorithms tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately.

However, these labeled datasets allow supervised learning algorithms to avoid computational complexity as they don’t need a large training set to produce intended

outcomes. Common regression and classification techniques are linear and logistic regression, naïve Bayes, KNN algorithm, and random forest.

Semi-supervised learning occurs when only part of the given input data has been labeled. Unsupervised and semi-supervised learning can be more appealing alternatives as it can be time-consuming and costly to rely on domain expertise to label data appropriately for supervised learning.

2.10. DECISION TREE

A decision tree is a tree-like model used to visualize and analyze decision-making processes and outcomes. It is a type of supervised learning algorithm used in machine learning and data mining. The decision tree consists of nodes and branches. The nodes represent decision points, and the branches represent the possible outcomes or decisions that can be made. The decision tree algorithm iteratively selects the best attribute to split the data into different classes. The splitting of the data is done in a way that the resulting subsets are as homogeneous as possible concerning the target variable. There are two types of nodes in a decision tree:

Decision Nodes: A decision node represents a decision or a test. It has two or more branches representing the possible outcomes of the decision. Each branch represents a possible value for the attribute being tested.

Leaf Nodes: A leaf node represents a classification or a decision. It represents the outcome of a decision path through the tree. A leaf node contains the class label or target variable value.

The process of building a decision tree involves selecting the best attribute to split the data at each decision node. The selection criterion for choosing the best attribute can vary, but the most commonly used methods are:

- Information gain
- Gini index
- Chi-squared test

The decision tree algorithm continues to split the data until a stopping criterion is met, such as when all instances belong to the same class or when the maximum depth of the tree is reached.

Once the decision tree is built, it can be used to classify new instances by following the decision path through the tree until a leaf node is reached. The class label or target variable value associated with the leaf node is then assigned to the new instance. Decision trees have several advantages over other classification models, including their interpretability and ease of use. They can also handle both categorical and numerical data, and they can be used for both classification and regression problems. However, decision trees also have some limitations, such as their tendency to overfit the data, their sensitivity to small changes in the data, and their inability to capture complex relationships between variables. In summary, a decision tree is a useful tool for visualizing and analyzing decision-making processes and outcomes. It is a popular algorithm used in machine learning and data mining for classification and regression problems.

2.10.1. ADVANTAGE OF DECISION TREE

Decision trees are a popular machine learning algorithm that can be used for both classification and regression tasks. Here are some advantages of using decision trees:

1. *Easy to understand:* Decision trees are easy to understand and interpret, making them a great tool for both experts and non-experts.
2. *Suitable for both categorical and numerical data:* Decision trees can handle both categorical and numerical data, which means they can be used for a wide variety of tasks.
3. *Efficient for large datasets:* Decision trees can be trained efficiently on large datasets, which is important for tasks that involve a lot of data.
4. *Robust to noise:* Decision trees are robust to noise in the data, which means they can still perform well even if the data contains errors or outliers.
5. *Non-parametric method:* Decision trees are a non-parametric method, which means they do not make any assumptions about the distribution of the data. This makes them flexible and able to capture complex relationships in the data.
6. *Feature selection:* Decision trees can be used for feature selection, which means they can help identify the most important features in the data for a particular task.

Overall, decision trees are a powerful and flexible machine learning algorithm that can be used for a wide range of tasks.

2.10.2. DISADVANTAGE OF DECISION TREE

One disadvantage of decision trees is that they can easily become complex and difficult to interpret, particularly when dealing with large datasets or when the tree has many levels or branches. This can make it difficult for users to understand the reasoning behind the decision-making process and may also result in over-fitting, where the model performs well on the training data but poorly on new, unseen data.

Another potential disadvantage of decision trees is that they may be sensitive to small changes in the input data, which can result in significant changes to the final decision tree structure and predictions. This can make the model less robust and reliable, particularly when dealing with noisy or inconsistent data.

Additionally, decision trees can also suffer from bias if the training data is not representative of the broader population or if there are imbalances in the distribution of the target variable. This can result in the model making incorrect predictions or being less accurate for certain groups or subsets of data.

2.10.3. USES OF DECISION TREE

Decision trees are a popular machine-learning technique that is commonly used for solving classification and regression problems. They are versatile tools that can be used for a wide range of applications across various industries. Here are some of the common uses of decision trees:

1. *Predictive Analytics:* Decision trees are often used for predictive analytics, where they are used to predict future outcomes based on past data. For example, they can be used to predict whether a customer is likely to churn or not, based on their past behavior.

2. *Medical Diagnosis:* Decision trees are widely used in medical diagnosis to help doctors diagnose diseases and conditions. They can help identify the most likely diagnosis based on symptoms, lab results, and other patient information.

3. *Credit Scoring:* Decision trees are used in the finance industry to predict the creditworthiness of a borrower. Based on factors like income, credit score, and loan history, a decision tree can be used to determine whether a borrower is a good credit risk or not.

4. *Fraud Detection:* Decision trees are also used in fraud detection, where they help identify patterns of fraudulent behavior. For example, a decision tree can be used to flag credit card transactions that are suspicious based on factors like location, amount, and frequency.
5. *Marketing:* Decision trees are used in marketing to help identify the most effective marketing campaigns. They can be used to segment customers based on factors like demographics, buying behavior, and interests, and then target them with personalized marketing campaigns.
6. *Customer Service:* Decision trees are used in customer service to help automate the resolution of common customer issues. For example, a decision tree can be used to guide customers through troubleshooting steps for a product or service.
7. *Supply Chain Management:* Decision trees are used in supply chain management to help optimize supply chain operations. They can be used to identify the most efficient routes for transportation, the best suppliers to work with, and the optimal inventory levels to maintain.

2.10.4. APPLICATION OF DECISION TREE

Decision trees are a popular machine learning algorithm used for both classification and regression tasks. They are used in a wide range of applications, including:

1. *Healthcare:* Decision trees are used to predict diseases based on symptoms, and to identify the most effective treatments based on patient characteristics.
2. *Finance:* Decision trees are used to predict stock prices, credit risk, and fraud detection.
3. *Marketing:* Decision trees are used to identify customer segments based on demographic and behavioral data, and to personalize marketing messages.
4. *Engineering:* Decision trees are used to identify defects in manufacturing processes, and to optimize the design of products and systems.
5. *Customer Service:* Decision trees are used to provide customer support by providing step-by-step instructions for resolving issues and answering common questions.

6. *Agriculture:* Decision trees are used to predict crop yields based on weather conditions and soil properties.
7. *Education:* Decision trees are used to predict student performance based on demographic and academic data, and to identify factors that contribute to success or failure.

Overall, decision trees are a powerful and versatile tool for solving a wide range of problems in various industries.

Chapter 3

DATA DESCRIPTION

The dataset for our study on heart disease is discussed in this chapter. The heart disease dataset is a collection of information about heart disease from a variety of sources. This dataset contains information such as patient demographics, clinical measurements, and lifestyle factors. It is useful for research on cardiovascular disease and the risk factors associated with it. It can be used to build predictive models to identify patients at risk of developing heart disease and to develop strategies for early intervention and prevention.

The dataset is available in Kaggle. The dataset has been taken from the source of UCI Machine Learning Repository; the website link is given as: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

About 303 samples were collected. It is used to predict whether a patient will have less chance of heart attack/ more chance of heart attack based on the factors like age, gender, chest pain type and also by various factors. The dataset consists of 303 observations with 14 factors. The details of the various factors with reference to heart attack are taken up and given below:

ATTRIBUTES OF DATASET

- Age: Age of the patient
- Sex: Gender of the patient
  (1=male, 0=female)
- cp: Chest Pain type
  value 0: typical angina,
  value 1: atypical angina,
  value 2: non-anginal pain, and
  value 3: asymptomatic.
- trtbps: resting blood pressure (in mm Hg)
- Chol: cholesterol in mg/dl fetched via BMI sensor
- fbs: (fasting blood sugar > 120 mg/dl)
  (1 = true; 0 = false)
- rest_ecg: resting electrocardiographic results with

value 0: normal,

value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV),

value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria.

- exang: exercise-induced angina
  (1 = yes; 0 = no)
- thalach: maximum heart rate achieved
- oldpeak = ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment
  value 1: upsloping,
  value 2: flat, and
  value 3: downsloping.
- ca: number of major vessels (0-3) coloured by fluroscopy
- thall : A blood disorder called thalassemia
  value 1: normal (blood flow),
  value 2: fixed defect (no blood flow in some part of heart),
  value 3 for reversible defect (a blood flow is observed but it is not normal).
- heart attack: Target variable
  value 0 indicates less chance of heart attack and
  value 1 indicates more chance of heart attack.

HEART DISEASE DATASET

age	sex	cp	trtbps	chol	fbs	restecg	thalach	exng	oldpeak	slp	ca	thall	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1
40	1	3	140	199	0	1	178	1	1.4	2	0	3	1

71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
51	1	2	110	175	0	1	123	0	0.6	2	0	2	1
65	0	2	140	417	1	0	157	0	0.8	2	1	2	1
53	1	2	130	197	1	0	152	0	1.2	0	0	2	1
41	0	1	105	198	0	1	168	0	0	2	1	2	1
65	1	0	120	177	0	1	140	0	0.4	2	0	3	1
44	1	1	130	219	0	0	188	0	0	2	0	2	1
54	1	2	125	273	0	0	152	0	0.5	0	1	2	1
51	1	3	125	213	0	0	125	1	1.4	2	1	2	1
46	0	2	142	177	0	0	160	1	1.4	0	0	2	1
54	0	2	135	304	1	1	170	0	0	2	0	2	1
54	1	2	150	232	0	0	165	0	1.6	2	0	3	1
65	0	2	155	269	0	1	148	0	0.8	2	0	2	1
65	0	2	160	360	0	0	151	0	0.8	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
48	1	1	130	245	0	0	180	0	0.2	1	0	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
53	0	0	130	264	0	0	143	0	0.4	1	0	2	1
39	1	2	140	321	0	0	182	0	0	2	0	2	1
52	1	1	120	325	0	1	172	0	0.2	2	0	2	1
44	1	2	140	235	0	0	180	0	0	2	0	2	1
47	1	2	138	257	0	0	156	0	0	2	0	2	1
53	0	2	128	216	0	0	115	0	0	2	0	0	1
53	0	0	138	234	0	0	160	0	0	2	0	2	1
51	0	2	130	256	0	0	149	0	0.5	2	0	2	1
66	1	0	120	302	0	0	151	0	0.4	1	0	2	1
62	1	2	130	231	0	1	146	0	1.8	1	3	3	1

44	0	2	108	141	0	1	175	0	0.6	1	0	2	1
63	0	2	135	252	0	0	172	0	0	2	0	2	1
52	1	1	134	201	0	1	158	0	0.8	2	1	2	1
48	1	0	122	222	0	0	186	0	0	2	0	2	1
45	1	0	115	260	0	0	185	0	0	2	0	2	1
34	1	3	118	182	0	0	174	0	0	2	0	2	1
57	0	0	128	303	0	0	159	0	0	2	1	2	1
71	0	2	110	265	1	0	130	0	0	2	1	2	1
54	1	1	108	309	0	1	156	0	0	2	0	3	1
52	1	3	118	186	0	0	190	0	0	1	0	1	1
41	1	1	135	203	0	1	132	0	0	1	0	1	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
35	0	0	138	183	0	1	182	0	1.4	2	0	2	1
51	1	2	100	222	0	1	143	1	1.2	1	0	2	1
45	0	1	130	234	0	0	175	0	0.6	1	0	2	1
44	1	1	120	220	0	1	170	0	0	2	0	2	1
62	0	0	124	209	0	1	163	0	0	2	0	2	1
54	1	2	120	258	0	0	147	0	0.4	1	0	3	1
51	1	2	94	227	0	1	154	1	0	2	1	3	1
29	1	1	130	204	0	0	202	0	0	2	0	2	1
51	1	0	140	261	0	0	186	1	0	2	0	2	1
43	0	2	122	213	0	1	165	0	0.2	1	0	2	1
55	0	1	135	250	0	0	161	0	1.4	1	0	2	1
51	1	2	125	245	1	0	166	0	2.4	1	0	2	1
59	1	1	140	221	0	1	164	1	0	2	0	2	1
52	1	1	128	205	1	1	184	0	0	2	0	2	1
58	1	2	105	240	0	0	154	1	0.6	1	0	3	1
41	1	2	112	250	0	1	179	0	0	2	0	2	1

45	1	1	128	308	0	0	170	0	0	2	0	2	1
60	0	2	102	318	0	1	160	0	0	2	1	2	1
52	1	3	152	298	1	1	178	0	1.2	1	0	3	1
42	0	0	102	265	0	0	122	0	0.6	1	0	2	1
67	0	2	115	564	0	0	160	0	1.6	1	0	3	1
68	1	2	118	277	0	1	151	0	1	2	1	3	1
46	1	1	101	197	1	1	156	0	0	2	0	3	1
54	0	2	110	214	0	1	158	0	1.6	1	0	2	1
58	0	0	100	248	0	0	122	0	1	1	0	2	1
48	1	2	124	255	1	1	175	0	0	2	2	2	1
57	1	0	132	207	0	1	168	1	0	2	0	3	1
52	1	2	138	223	0	1	169	0	0	2	4	2	1
54	0	1	132	288	1	0	159	1	0	2	1	2	1
45	0	1	112	160	0	1	138	0	0	1	0	2	1
53	1	0	142	226	0	0	111	1	0	2	0	3	1
62	0	0	140	394	0	0	157	0	1.2	1	0	2	1
52	1	0	108	233	1	1	147	0	0.1	2	3	3	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
57	1	0	152	274	0	1	88	1	1.2	1	1	3	0
56	1	0	132	184	0	0	105	1	2.1	1	1	1	0
48	1	0	124	274	0	0	166	0	0.5	1	0	3	0
56	0	0	134	409	0	0	150	1	1.9	1	2	3	0
66	1	1	160	246	0	1	120	1	0	1	3	1	0
54	1	1	192	283	0	0	195	0	0	2	1	3	0

69	1	2	140	254	0	0	146	0	2	1	3	3	0
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
43	1	0	132	247	1	0	143	1	0.1	1	4	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
67	1	0	100	299	0	0	125	1	0.9	1	2	2	0
59	1	3	160	273	0	0	125	0	0	2	0	2	0
45	1	0	142	309	0	0	147	1	0	1	3	3	0
58	1	0	128	259	0	0	130	1	3	1	2	3	0
50	1	0	144	200	0	0	126	1	0.9	1	0	3	0
62	0	0	150	244	0	1	154	1	1.4	1	0	2	0
38	1	3	120	231	0	1	182	1	3.8	1	0	3	0
66	0	0	178	228	1	1	165	1	1	1	2	3	0
52	1	0	112	230	0	1	160	0	0	2	1	2	0
53	1	0	123	282	0	1	95	1	2	1	2	3	0
63	0	0	108	269	0	1	169	1	1.8	1	2	2	0
54	1	0	110	206	0	0	108	1	0	1	1	2	0
66	1	0	112	212	0	0	132	1	0.1	2	1	2	0
55	0	0	180	327	0	2	117	1	3.4	1	0	2	0
49	1	2	118	149	0	0	126	0	0.8	2	3	2	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
56	1	0	130	283	1	0	103	1	1.6	0	0	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
61	1	3	134	234	0	1	145	0	2.6	1	2	2	0
67	1	0	120	237	0	1	71	0	1	1	0	2	0
58	1	0	100	234	0	1	156	0	0.1	2	1	3	0
47	1	0	110	275	0	0	118	1	1	1	1	2	0
52	1	0	125	212	0	1	168	0	1	2	2	3	0
58	1	0	146	218	0	1	105	0	2	1	1	3	0

57	1	1	124	261	0	1	141	0	0.3	2	0	3	0
58	0	1	136	319	1	0	152	0	0	2	2	2	0
61	1	0	138	166	0	0	125	1	3.6	1	1	2	0
42	1	0	136	315	0	1	125	1	1.8	1	0	1	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
59	1	2	126	218	1	1	134	0	2.2	1	1	1	0
40	1	0	152	223	0	1	181	0	0	2	0	3	0
61	1	0	140	207	0	0	138	1	1.9	2	1	3	0
46	1	0	140	311	0	1	120	1	1.8	1	2	3	0
59	1	3	134	204	0	1	162	0	0.8	2	2	2	0
57	1	1	154	232	0	0	164	0	0	2	1	2	0
57	1	0	110	335	0	1	143	1	3	1	1	3	0
55	0	0	128	205	0	2	130	1	2	1	1	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
58	0	0	170	225	1	0	146	1	2.8	1	2	1	0
67	1	2	152	212	0	0	150	0	0.8	1	0	3	0
44	1	0	120	169	0	1	144	1	2.8	0	0	1	0
63	1	0	140	187	0	0	144	1	4	2	2	3	0
63	0	0	124	197	0	1	136	1	0	1	0	2	0
59	1	0	164	176	1	0	90	0	1	1	2	1	0
57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
57	0	1	130	236	0	0	174	0	0	1	1	2	0

Chapter 4

ANALYSIS AND INTERPRETATION

In this chapter, we discuss the analysis and interpretation of the heart disease patients dataset. The variables which are considered for this study are chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, old peak, slope type, exercise-induced angina, maximum heart rate achieved, thalassemia, and the target variable. The bar diagram, pie diagram, one sample t-test, correlation, logistic regression, and decision trees are the statistical models to analyze the dataset.

4.1. DESCRIPTIVE STATISTICS

The Descriptive statistics are displayed using SPSS software for obtaining the mean, median, standard deviation, variance, skewness, Standard Error of Skewness, kurtosis, Standard Error of kurtosis, Minimum, Maximum, Range, etc.

Table 4.1: Descriptive Statistics

Descriptive Statistics	Trestbps	Chol	Thalachh	Oldpeak
Mean	129.17	245.26	160.20	0.6449
Median	130.00	236.00	162.00	0.4000
Std. Deviation	14.499	55.476	18.188	0.8191
Variance	272.22	3077.53	330.78	0.6710
Skewness	0.330	1.910	-0.573	1.5560
Std.Error of Skewness	0.213	0.213	0.213	0.2130
Kurtosis	0.388	8.181	0.248	2.8340
Std.Error of Kurtosis	0.423	0.423	0.423	0.4230
Minimum	94	126	111	0.0000
Maximum	180	564	202	4.2000
Range	86	438	91	4.2000

The table above 4.1 shows the descriptive statistics for the quantitative variables of the dataset such as Trestbps which is resting blood pressure in mm Hg, cholesterol in mg/dl, Thalachh which is the maximum heart rate achieved, and old peak which is ST depression induced by exercise relative to rest.

4.2. BAR GRAPH

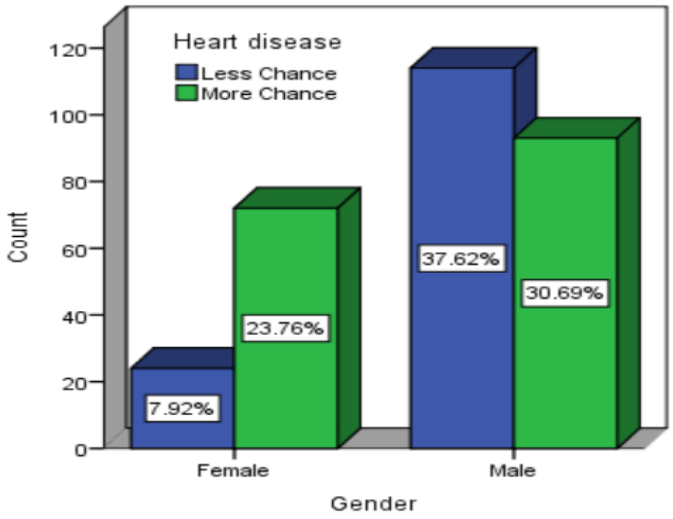


Fig 4.1: Bar chart of Heart Disease Chances by Gender

Interpretation

Figure 4.1 describes that happening of heart disease in females has less chance for 24 (7.92%) and more chance for 72 (23.76%) patients. Of the male patients 114 (37.62%) for a lesser chance and 93(30.69%) having a higher chance of happening heart disease.

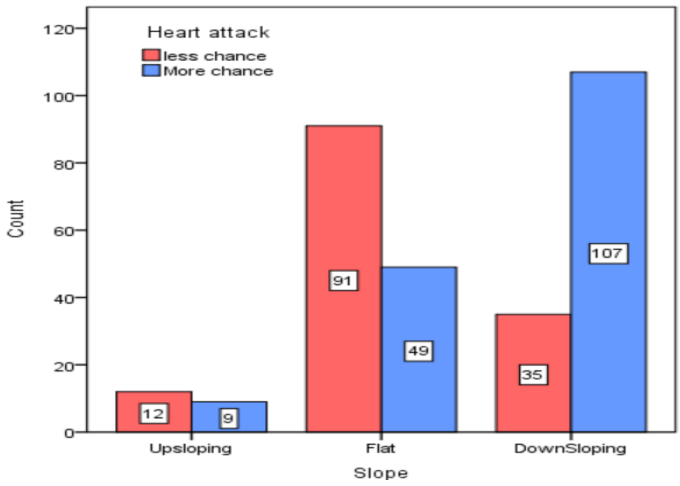


Fig 4.2: Bar chart for Slope Vs Target

Interpretation

Figure 4.2 displayed above shows the chance of heart disease associated with the sloping values. The high chance of heart attack in downsloping is 107, in flat is 49, and in upsloping is 9 among the whole observations. In downsloping many patients have more chance of having heart disease.

4.3.PIE CHART AND STACKED BAR GRAPH

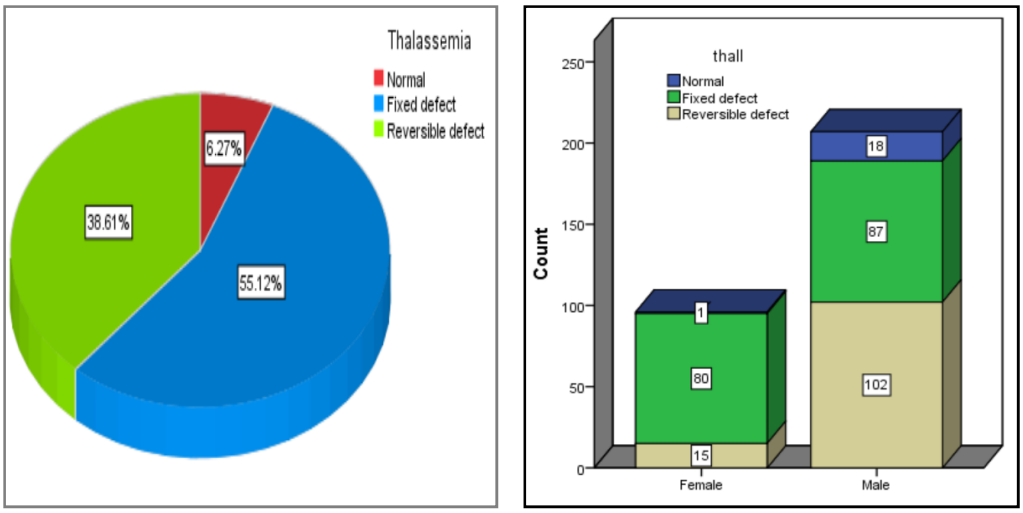


Fig: 4.3: Pie chart and Stacked bar graph for Thalassemia level

Interpretation

The above-left figure 4.3 depicts the category of thalassemia called a blood disorder in heart patients. The patients with normal blood flow in the heart are 6.27 percent, for the fixed defect is no blood flow in some part of the heart is 38.61 percent, and for the reversible defect which is blood flow is observed but it is not normal 55.12 percent. The more patients have reversible defect category of blood disorder. The right side of the above figure describes the blood disorder of female patients with 1 from normal, 80 from fixed defect, and 15 from reversible category, and for male patients 18 is from normal, 87 from fixed defect, and 102 from reversible category.

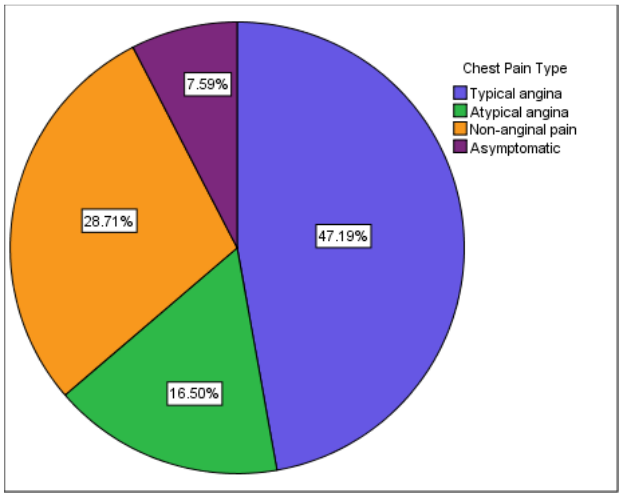


Fig 4.4: Pie chart of Chest Pain stages

Interpretation

From the above Figure 4.4, most of the patients that are 47.19 percent have typical angina which is described as squeezing, pressure, heaviness, tightness, or pain in the chest, 28.71 percent have non-angina pain, 16.50 percent have atypical angina usually feels like a stabbing or burning pain in the chest and may sometimes have characteristics similar to indigestion, and 7.59 percent have asymptomatic chest pain can be induced by physical or mental stress but may occur without any obvious trigger.

4.4. SCATTER DIAGRAM

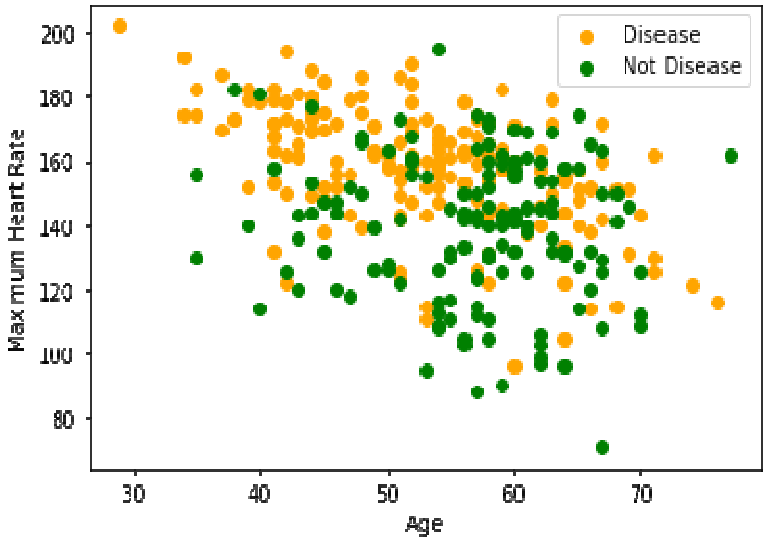


Fig4.5: Scatter diagram for heart disease

Interpretation

The above scatter diagram 4.5 shows the spread of the target variable depending on the age group of the heart patients in this study.

4.5.ONE SAMPLE T-TEST

The hypothesis for one sample test is given as

Null Hypothesis H_0: $\mu = 200$, There is no significant difference between the two means, that is the mean cholesterol level is equal to 200 mg/dL.

Alternative Hypothesis H_1: $\mu \neq 200$, There is a significant difference between the two means, that is the mean cholesterol level is not equal to 200 mg/dL.

Table 4.2: One-Sample Test: Test Value = 200 mg/dL

	t	df	Significance (2-tailed)	Mean Difference	95% CI of the Difference	
					Lower	Upper
chol	15.537	302	.000	46.264	40.40	52.12

The significant p-value is less than 0.05, so there is no evidence to accept the null hypothesis. We accept the alternative hypothesis that there is a significant difference between the two means. The mean cholesterol level of the heart disease victim is greater than the normal cholesterol of 200 mg/dL.

4.6.CHI-SQUARE TEST

Table 4.3: Gender * Output Crosstabulation

Gender	Output		Total
	less chance	More chance	
Female	24	72	96
Male	114	93	207
Total	138	165	303

From Table 4.3, the proportion of females with less chance of heart attack is 24, and more chance of happening a heart attack is 72 among 96 female patients, the proportion of male patients with less chance is 114, and more chance of having a heart attack is 93 among 207 males. The total number of patients having less chance is 138 and more chance is 165 of 303 cases.

Table 4.4: Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	23.914[a]	1	.000		
Continuity Correction[b]	22.717	1	.000		
Likelihood Ratio	24.841	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	23.835	1	.000		
N of Valid Cases	303				
a). 0 cells (0.0%) have an expected count of less than 5. The minimum expected count is 43.72. b). Computed only for a 2x2 table					

The value of the test statistic is 23.914 with 1 degree of freedom. Since the p-value is less than our chosen significance level of 0.05, we can reject the null hypothesis and conclude that there is an association between gender and the chances of happening heart attack (less or more).

4.7. CORRELATION

The coefficients of correlation between variables are shown in a correlation table. The relationship between cholesterol by age is shown below.

Table 4.5: Correlation Table

		age	trtbps	chol
age	Pearson Correlation	1	.279	.214
	Sig. (2-tailed)		.000	.000
	N	303	303	303
trtbps	Pearson Correlation	.279	1	.123
	Sig. (2-tailed)	.000		.032
	N	303	303	303
chol	Pearson Correlation	.214	.123	1
	Sig. (2-tailed)	.000	.032	
	N	303	303	303

Null Hypothesis: There is no correlation between age, resting blood pressure, and cholesterol level.

Alternative Hypothesis: There is a significant correlation between the age, resting blood pressure, and cholesterol level of the heart patient.

The significant p-value is greater than 0.05, so there exists a correlation that is there is relationship between the age, resting blood pressure, and the cholesterol level of the heart patient.

4.8. HEAT MAP

A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. It displays the correlation between multiple variables as a color-coded matrix. It displays the correlation between multiple variables as a color-coded matrix. It’s like a color chart that shows us how closely related different variables are. In a correlation heatmap, each variable is represented by a row and a column, and the cells show the correlation between them.

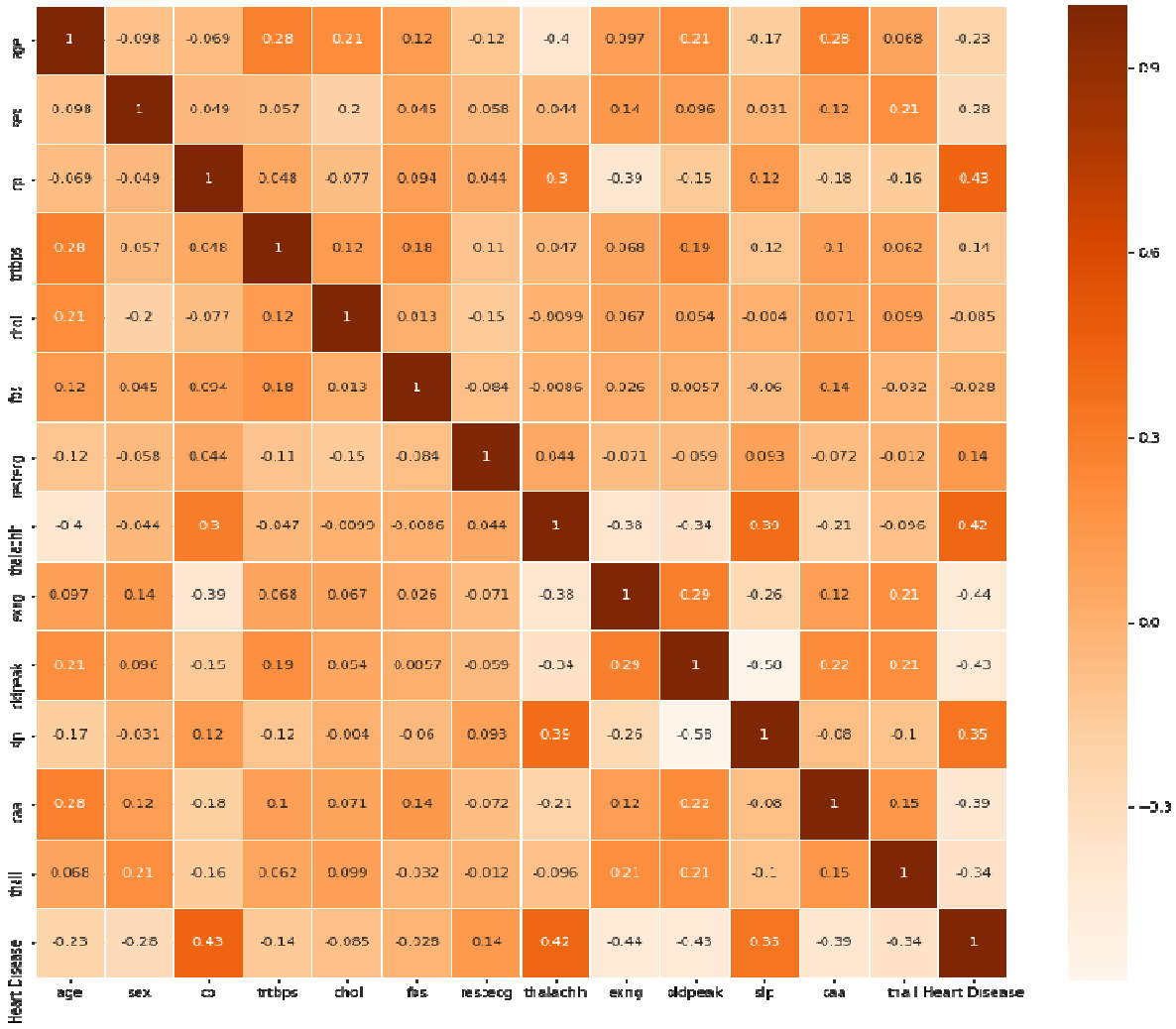
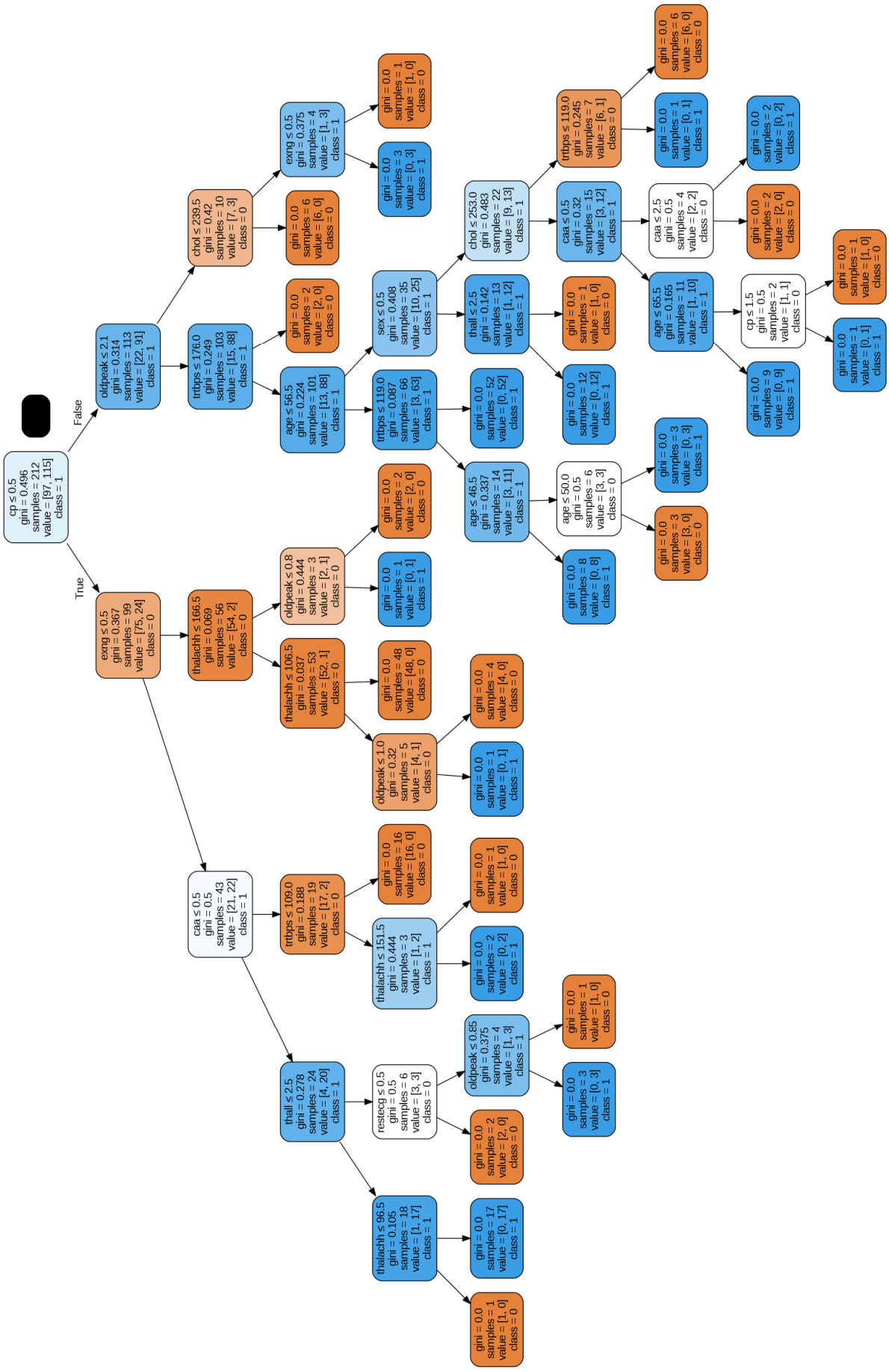


Fig 4.6: Heat Map for Heart disease data

From the above matrix figure 4.6, the color of each cell represents the strength and direction of the correlation, with darker colors indicating stronger correlations. Darker colors indicate stronger correlations, while lighter colors indicate weaker correlations. Positive correlations (when one variable increases, the other variable tends to increase) are usually represented by dark warm colors. Negative correlations (when one variable increases, the other variable tends to decrease) are usually represented by light cool colors in the matrix.

4.9. DECISION TREE

The decision tree has been constructed using Python in Google Colab, The Accuracy of the Decision tree model is 0.7362637362637363


cp ≤ 0.5
gini = 0.496
samples = 212
value = [97, 115]
class = 1
True
False
exng ≤ 0.5
gini = 0.367
samples = 99
value = [75, 24]
class = 0
oldpeak ≤ 2.1
gini = 0.314
samples = 113
value = [22, 91]
class = 1
caa ≤ 0.5
gini = 0.5
samples = 43
value = [21, 22]
class = 1
thalachh ≤ 166.5
gini = 0.069
samples = 56
value = [54, 2]
class = 0
trtbps ≤ 176.0
gini = 0.249
samples = 103
value = [15, 88]
class = 1
thall ≤ 2.5
gini = 0.278
samples = 24
value = [4, 20]
class = 1
trtbps ≤ 109.0
gini = 0.188
samples = 19
value = [17, 2]
class = 0
oldpeak ≤ 0.8
gini = 0.444
samples = 3
value = [2, 1]
class = 0
thalachh ≤ 106.5
gini = 0.037
samples = 53
value = [52, 1]
class = 0
chol ≤ 239.5
gini = 0.42
samples = 10
value = [7, 3]
class = 0
age ≤ 56.5
gini = 0.224
samples = 101
value = [13, 88]
class = 1
thalachh ≤ 96.5
gini = 0.105
samples = 18
value = [1, 17]
class = 1
restecg ≤ 0.5
gini = 0.5
samples = 6
value = [3, 3]
class = 0
thalachh ≤ 151.5
gini = 0.444
samples = 3
value = [1, 2]
class = 1
oldpeak ≤ 1.0
gini = 0.32
samples = 5
value = [4, 1]
class = 0
exng ≤ 0.5
gini = 0.375
samples = 4
value = [1, 3]
class = 1
sex ≤ 0.5
gini = 0.408
samples = 35
value = [10, 25]
class = 1
trtbps ≤ 119.0
gini = 0.087
samples = 66
value = [3, 63]
class = 1
oldpeak ≤ 0.85
gini = 0.375
samples = 4
value = [1, 3]
class = 1
age ≤ 46.5
gini = 0.337
samples = 14
value = [3, 11]
class = 1
chol ≤ 253.0
gini = 0.483
samples = 22
value = [9, 13]
class = 1
thall ≤ 2.5
gini = 0.142
samples = 13
value = [1, 12]
class = 1
trtbps ≤ 119.0
gini = 0.245
samples = 7
value = [6, 1]
class = 0
caa ≤ 0.5
gini = 0.32
samples = 15
value = [3, 12]
class = 1
caa ≤ 2.5
gini = 0.5
samples = 4
value = [2, 2]
class = 0
age ≤ 65.5
gini = 0.165
samples = 11
value = [1, 10]
class = 1
cp ≤ 1.5
gini = 0.5
samples = 2
value = [1, 1]
class = 0
age ≤ 50.0
gini = 0.5
samples = 6
value = [3, 3]
class = 0
gini = 0.0
samples = 16
value = [16, 0]
class = 0
gini = 0.0
samples = 1
value = [1, 0]
class = 0
gini = 0.0
samples = 2
value = [0, 2]
class = 1
gini = 0.0
samples = 48
value = [48, 0]
class = 0
gini = 0.0
samples = 4
value = [4, 0]
class = 0
gini = 0.0
samples = 1
value = [0, 1]
class = 1
gini = 0.0
samples = 2
value = [2, 0]
class = 0
gini = 0.0
samples = 1
value = [0, 1]
class = 1
gini = 0.0
samples = 6
value = [6, 0]
class = 0
gini = 0.0
samples = 1
value = [1, 0]
class = 0
gini = 0.0
samples = 3
value = [0, 3]
class = 1
gini = 0.0
samples = 2
value = [2, 0]
class = 0
gini = 0.0
samples = 3
value = [0, 3]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 0
gini = 0.0
samples = 17
value = [0, 17]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 0
gini = 0.0
samples = 52
value = [0, 52]
class = 1
gini = 0.0
samples = 12
value = [0, 12]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 0
gini = 0.0
samples = 8
value = [0, 8]
class = 1
gini = 0.0
samples = 3
value = [3, 0]
class = 0
gini = 0.0
samples = 3
value = [0, 3]
class = 1
gini = 0.0
samples = 6
value = [6, 0]
class = 0
gini = 0.0
samples = 1
value = [0, 1]
class = 1
gini = 0.0
samples = 2
value = [0, 2]
class = 1
gini = 0.0
samples = 2
value = [2, 0]
class = 0
gini = 0.0
samples = 9
value = [0, 9]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 0
gini = 0.0
samples = 1
value = [0, 1]
class = 1

Interpretation

The decision tree has been constructed and classified according to nominal output. It has several iterations. For each iteration, Gini Index, Samples, Value, and class have been calculated which increases the accuracy of the model. Above we can see the tree built after training.

During the training phase, the decision tree adds nodes, and split them into branches that lead to leaves. When the Decision Tree has to predict a target, heart disease, for the chance of heart attack belonging to the testing set, it travels down the tree from the root node until it reaches a leaf, deciding to go to the left or the right child node by testing the feature value of the heart disease being tested against the parent node condition cp for heart disease of two case labels 0 and 1.

A decision boundary is decided by testing all the possible decision boundaries splitting the dataset and choosing the one that minimizes the Gini impurity of the two splits. As the Gini impurity is 0 for heart disease, i.e. we cannot have a more homogeneous group, the algorithm will not try to split this part anymore.

The node's Gini attribute measures its impurity. A node is said to be pure when all training instances it applies belong to the same class. Here, we can see that in class heart disease of the leaf node has gini appears to be 0 which means it is "pure".

4.10. LOGISTIC REGRESSION

A logistic regression model has been obtained by using the MINITAB software. The following outcomes are used to indicate the logistic regression.

Logistic regression is mainly used for prediction and also for calculating the probability of success. Logistic regression models allow us to fit a regression model to categorical data. Here we focus on the heart disease of a patient. The Link function (Logit) method is used in this case.

Categorical predictor coding (1, 0) is more chance, less chance of heart attack. Here we have taken the more-chance of heart attack as the reference event.

The target variable of heart disease has 165 patients having more chance of heart attack event and 138 have less chance of heart attack of the total 303 patients.

Table 4.6: Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	9	183.169	20.3521	183.17	0.0000
trtbps	1	4.639	4.6386	4.64	0.0310
thalch	1	12.523	12.5229	12.52	0.0000
oldpeak	1	13.119	13.1193	13.12	0.0000
Gender	1	9.177	9.1769	9.18	0.0020
cp	3	35.596	11.8652	35.6	0.0000
thal	2	19.832	9.9158	19.83	0.0000
Error	293	234.469	0.8002		
Total	302	417.638			

From the above table 4.6, the p-values are less than 0.05 so we do reject the null hypothesis, and our model is statistically significant to the heart disease dataset. The influencing factors of more chance of heart attack are the variables displayed in the table through logistic regression analysis.

Table 4.7: Coefficients

Term	Coef	SE Coef	VIF
Constant	1.83	1.88	
Trtbps	-0.01978	0.00935	1.11
Thalch	0.02906	0.00865	1.14
Oldpeak	-0.626	0.182	1.17
Gender (male)	-1.247	0.424	1.29
cp (Atypical angina)	-0.795	0.712	2.44
cp (Non Anginal Pain)	-0.189	0.619	2.77
cp (Typical angina)	-2.255	0.609	3.42
thal (Normal)	-0.374	0.691	1.2
thal (Reversible defect)	-1.555	0.361	1.17

In coefficient table 4.7, we have the coefficients received by all independent variables. The variance influence function (VIF) greater than 1 indicates that there exists a correlation between the predictor variables.

Table 4.8: Odds Ratios for Continuous Predictors

Odds	Ratio	95% CI
Trtbps	0.9804	(0.9626, 0.9985)
Thalch	1.0295	(1.0122, 1.0471)
Oldpeak	0.5349	(0.3740, 0.7649)

Positive odds ratios indicate that the event is more likely to occur, whilst negative odd ratios indicate the event is less likely to occur.

Note that the coefficient is the log odds ratio. The 'log' part of the log-odds ratio is just the logarithm of the odds ratio, as a logistic regression uses a logarithmic function to solve the regression problem. It is much easier to just use the odds ratio, so we must take the exponential of the log-odds ratio to get the odds ratio.

For categorical features or predictors, the odds ratio compares the odds of the event occurring for each category of the predictor relative to the reference category, given that all other variables remain constant.

Table 4.9: Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Gender			
Male	Female	0.2875	(0.1253, 0.6595)
Chest pain type (cp)			
Atypical_angina	Asymptomatic	0.4514	(0.1118, 1.8226)
NonAnginalPain	Asymptomatic	0.8275	(0.2459, 2.7848)
Typical_angina	Asymptomatic	0.1049	(0.0318, 0.3458)
NonAnginalPain	Atypical_angina	1.8331	(0.6181, 5.4362)
Typical_angina	Atypical_angina	0.2324	(0.0875, 0.6176)
Typical_angina	NonAnginalPain	0.1268	(0.0564, 0.2853)
Thall			
Normal	Fixed defect	0.6882	(0.1778, 2.6642)
Reversible defect	Fixed defect	0.2113	(0.1041, 0.4289)
Reversible defect	Normal	0.307	(0.0807, 1.1677)

From above table 4.9, we observe that males are highly affected by heart disease compared to females. Odd ratios for level A relative to level B for different categories of the variables are considered here in the table.

The Regression Equation for the probability of getting more chance of heart attack is given as follows:

$$P(\text{More chance}) = \frac{\exp(Y')}{\left(1 + \exp(Y')\right)}$$

Y' = 1.83 - 0.01978 trtbps + 0.02906 thalch - 0.626 oldpeak + 0.0 Gender_Female
- 1.247 Gender_Male + 0.0 cp_Asymptomatic - 0.795 cp_Atypical_angina
- 0.189 cp_NonAnginalPain - 2.255 cp_Typical_angina + 0.0 thal_Fixed defect
- 0.374 thal_Normal - 1.555 thal_Reversible defect

Table 4.10: Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	293	234.47	0.995
Pearson	293	303.18	0.329
Hosmer-Lemeshow	8	4.97	0.761

The model fitness of the observed data can be assessed in several ways; we will use the Pearson; Deviance; Hosmer-Lemenshow tests. When the Deviance test is significant (p-value<0.05) it means the model does not describe the data well. This means that the null hypothesis of data fitting the model can be rejected. In other words, the goodness of fit is not that good.

H_0: model does fit the data

H_A: model does not fit the data.

Pearson and Deviance are both types of residuals. The larger the p-value the better is the fit of the model to the data. When the model does not fit (reject H0) it would be best to try alternative models and opt for the one that produces the largest p-values.

In our result from above table 4.10, the p-value of Deviance is 0.997, Pearson is 0.221, and Hosmer-Lemeshow is 0.866 which gives us significant evidence that our model is fitting our data. That is, the model adequately describes the relationship between the response and predictor variables in the data.

Chapter 5

SUMMARY AND CONCLUSION

In this chapter, we discuss a summary of the analysis and its inference to heart disease patients. The study data consist of 303 observation for analysis. The pie chart, and bar graph was used to visualize the heart disease chances, one sample t-test, correlation, logistics regression, and decision tree are the analysis used for the data. The data is collected from the UCI machine learning repository. The descriptive statistics for the quantitative variables of the dataset such as resting blood pressure in mm Hg, cholesterol in mg/dL, maximum heart rate achieved, and ST depression induced by exercise relative to rest were calculated.

The Bar graph displayed the chance of heart disease associated with the sloping values. In downsloping many patients have more chance of having heart disease. The Pie chart depicts the category of thalassemia called a blood disorder in heart patients. The patients with normal blood flow in the heart are 6.27 percent, for the fixed defect is no blood flow in some part of the heart is 38.61 percent, and for the reversible defect which is blood flow is observed but it is not normal can be 55.12 percent. The more patients were from the reversible defect category of blood disorder. The Pie chart for chest pain type shows that most of the patients that are 47.19 percent have typical angina which is described as squeezing, pressure, heaviness, tightness, or pain in the chest, 28.71 percent have non-angina pain, 16.50 percent have atypical angina usually feels like a stabbing or burning pain in the chest and may sometimes have characteristics similar to indigestion, and 7.59 percent have asymptomatic chest pain can be induced by physical or mental stress but may occur without any obvious trigger.

The correlation analysis shows a significant p-value which was greater than 0.05, so there exists a correlation that there was a relationship between the age, resting blood pressure, and the cholesterol level of the heart patient. The correlation heat map matrix figure displays, the color of each cell represents the strength and direction of the correlation. Darker colors indicate stronger correlations, while lighter colors indicate weaker correlations. Positive correlations when one variable increases, the other variable tends to increase are usually represented by dark warm colors. Negative correlations when one variable increases, and the other variable tends to decrease are usually represented by light cool colors in the matrix.

A logistic regression model has been obtained by using the MINITAB software. Positive odds ratios indicate that the event is more likely to occur, whilst negative odd ratios indicate the event is less likely to occur. In our result, Pearson and Deviance are both types of residuals, the p-value of Deviance is 0.997, Pearson is 0.221, and Hosmer-Lemeshow is 0.866 which gives us significant evidence that our model is fitting our data. That is, the model adequately describes the relationship between the response and predictor variables in the data.

The decision tree has been constructed using Python in Google Colab. The accuracy of the decision tree model is 0.736267. The decision tree has been constructed and classified according to nominal output. It has several iterations. For each iteration, Gini Index, Samples, Value, and class have been calculated which increases the accuracy of the model. During the training phase, the decision tree adds nodes, and split them into branches that lead to leaves. Once the model is trained, it can be tested on a separate set of data to evaluate its performance. This testing phase helps assess the model's ability to generalize and make accurate predictions on unseen data. If the model performs well, it can be used to make predictions on new, unseen data.

One of the applications of the decision tree algorithm is in the field of healthcare, where it can be used to predict the risk of certain medical conditions, such as heart attacks. For the heart attack dataset, the decision tree algorithm can classify patients into different risk categories based on their age, gender, cholesterol levels, blood pressure, and other relevant features. The output variable of the dataset, which contains the categorization information, can be used to validate the accuracy of the model's predictions.

A decision boundary is decided by testing all the possible decision boundaries splitting the dataset and choosing the one that minimizes the Gini impurity of the two splits. As the Gini impurity is 0 for heart disease, i.e. we cannot have a more homogeneous group, the algorithm will not try to split this part anymore. The node’s Gini attribute measures its impurity. A node is said to be pure when all training instances it applies belong to the same class. Here, we can see that in class heart disease of the leaf node has gini appears to be 0 which means it is “pure”.

Furthermore, the results of the decision tree algorithm can be visualized to gain insights into the decision-making process. The tree structure can be represented graphically, showing the sequence of questions and decisions made by the model to arrive

at its predictions. This visual representation provides a clear understanding of how the model is categorizing or predicting based on the input features.

In conclusion, a decision tree algorithm is a supervised learning technique used for categorization or prediction tasks. It is trained on labeled data and uses a tree-like structure to make decisions based on input features. The accuracy of the model's predictions can be validated using the output variable of the dataset. The visualization of the decision tree provides insights into the model's decision-making process, making it a valuable tool in various fields, including healthcare.

REFERENCE

1. Bingham, N. H., and John M. Fry. (2010), Regression: Linear Models in statistics *Springer*.
2. Daniel, W. W and Chad .L, Cross. (2018), Bio-Statistics : A Foundation for analysis in the Health Science, Eleventh Edition , John Willey and Sons.
3. David G. Kleinbaum., and Mitchel Klein. (2010), Logistic Regression A self learning text, Third edition, Science for Biology and Health, *Springer*.
4. Frank, E., and Harrell, Jr. (2015), Regression Modeling Strategies with Applications to Linear Models, Logistic and ordinal Regression and Survival analysis, *Springer*.
5. Gupta, S. C., and Kapoor, V. K. (2016)., Fundamental of Mathematical Statistics, Sultan Chand & Sons, New Delhi.
6. Miroslav Kubat. (2017), An Introduction to Machine Learning, Second edition, *Springer*.
7. Muller, A. C., and Guido, S. (2016), Introduction to Machine Learning with python, *O'Reilly*.
8. Norman Matloff (2017), statistical Regression and classification from Linear Models to Machine Learning, *CRC Press*.
9. Rohatgi, V. K., and Md Ehsanes Saleh A. K. (2001) An Introduction to probability and statistics, *John Wiley & Sons, Inc*.
10. UCI Repository: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
11. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.
12. https://towardsdatascience.com/logistic-regression-using-minitab-d58a80ec548a.
13. https://scikit-learn.org/stable/modules/tree.html.
14. https:// www.healthknowledge.org.uk.
15. https://pythonprogramminglanguage.com/what-are-the-advantages-of-using-a-decision-tree-for-classification/
16. https://developers.google.com/machine-learning/crash-course/logistic-regression/calculating-a-probability.

*****