KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

PREDICTIVE ANALYSIS OF DIABETES USING LOGISTIC REGRESSION

By

Josephine Nyarko Peprah

Wirekoah Augustina Ama

Asare Atuahene Joseph

Kessey Opoku

A THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BSc. STATISTICS

September 11, 2023

Declaration

We hereby declare that this submission is my own work towards the award of the BSc. degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgment had been made in the text.

Josephine Nyarko Peprah	……	……
Student	Signature	Date

Wirekoah Augustina Ama	……	……
Student	Signature	Date

Asare Atuahene Joseph	……	……
Student	Signature	Date

Kessey Opoku	……	……
Student	Signature	Date

Certified by:

Mr. Isaac Akpor Ajei	……	……
Supervisor	Signature	Date

Certified by:

Prof. Atinuke Adebanji	……	……
Head of Department	Signature	Date

Dedication

We give thanks to God Almighty for providing us with the strength, knowledge, understanding, and wisdom required to complete this work. We most obviously would not have been able to finish this project without him. We also want to thank Mr. Issac Adjei Akpor, our supervisor for his constant support.

Acknowledgements

In order to help this project be completed successfully, We would want to sincerely thank everyone who helped.Thank you for your helpful advice and steadfast support along this journey, [Mr. Isaac Akpor Ajei]. This project has been greatly influenced by your advice and support. We would want to express our sincere gratitude to our family and friends for their unwavering support and compassion. Your confidence in us kept us inspired, and your support gave us the willpower to keep going.

Abstract

Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels, resulting from either insufficient insulin production or the body's inability to use insulin effectively. Diabetes comes in three different types: type 1, type 2, and gestational diabetes. The most prevalent form of diabetes is type 2. Diabetes can cause a number of complications, such as vision issues, neuropathy, and kidney damage. Numerous factors have been linked to diabetes in research, including age, gender, smoking status, blood pressure, hemoglobin, A1c, BMI, location, insurance, LDL and HDL levels. Utilizing statistical techniques that provide clinically significant data, the disease burden in Virginia will be assessed. The dataset for this study was derived from Sutham Jirapanakorn's physical examination in 2021. The specific goals of this study are to formulate the logistic regression model for the investigation, identify significant risk factors for diabetes, and predict diabetes risk. In this project, logistic regression model was used to examine the complex relationship between diabetes and various influencing factors. Logistic regression was used in the methodology, which enables us to analyze and quantify the effects of risk factors such as age, gender, smoking status, blood pressure, hemoglobin, A1c, BMI, location, insurance, LDL and HDL levels. With the help of this statistical method, the model was used to determine the likelihood of someone developing the diabetes and gain a better understanding of how these factors contribute to the development of diabetes.These elements have been carefully selected due to their potential impact on diabetes. It was concluded that the age, HDL, LDL were the factors contributing to an individual getting diabetes.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Background Of The Study

Diabetes Mellitus simply known as diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone produced in the pancreas by the islets of Langerhans, which regulates the amount of glucose in the blood. The lack of insulin causes a form of diabetes.

Diabetes Mellitus are in three types:

- The type 1 diabetes
- The type 2 diabetes
- The Gestational diabetes

The Type 1 diabetes once known as juvenile diabetes or insulin-dependent diabetes, is a chronic condition in which the pancreas produces little or no insulin by itself. The most common is Type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin. In the past three decades the prevalence of Type 2 diabetes has risen dramatically in countries of all income levels. Gestational diabetes develops in women during pregnancy. It generally occurs in the second half of pregnancy, when an increase in placental hormones leads to increased insulin resistance, which can cause high glucose levels in the blood. Approximately five to eight percent of pregnant women develop gestational

Hyperglycemia, also called raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over a time leads to serious damage to many of the body's system

such as hearts, kidney and eyes, especially the nerves and blood vessels. For people living with diabetes, access to affordable treatment, including insulin, is critical to their survival.
Referring to the World Health Organization (WHO), about 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.5 million deaths are directly attributed to diabetes a year. The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014.

There is a globally agreed target to halt the rise in diabetes and obesity by 2025. As reported by the Pan American Health Organization PAHO and World Health Organization WHO in 2021, it is estimated that 62 million people in the Americans live with Type 2 Diabetes Mellitus. This number has tripled in the Region since 1980 and is estimated to reach the 109 million mark by 2040 (Sun et al. 2021). Prevalence has been rising more rapidly in low- and middle-income countries than in high-income countries. Globally, between 2000 and 2016, there was a 5% increase in premature mortality from diabetes.

Diabetes is an epidemic in the United States. According to the Centers for Disease Control and Prevention (CDC) in 2021, over 34 million Americans have diabetes and face its devastating consequences. What's true nationwide is also true in Virginia. (Leon et al. 2021), Diabetes affects approximately over 10% of the U.S. population and costs $245 billion in medical costs/lost wages annually. It affects over 10% of the Virginia population as well, accounting for 733,302 Virginians. While diabetes is found in all populations and regions, certain demographics (ex. Black population) and areas (ex. Southwestern region) of the state are disproportionately affected.

As claimed by the America Diabetes Association (ADA) in 2021, Approximately 701,793 people in Virginia, or 10.4% of the adult population, have been diagnosed with diabetes. An additional 189,000 people in Virginia have diabetes but don't know it, greatly increasing their health risk. There are 2,208,000 people in Virginia, 33.3% of the adult population, who have prediabetes with blood glucose levels that are higher than normal but not yet high enough to be diagnosed

as diabetes. Every year an estimated 59,557 people in Virginia are diagnosed with diabetes. Diabetes alone can present major challenges and threats to health due to risks related to high blood glucose levels including diabetic ketoacidosis and hyperosmolar hyperglycemic nonketotic syndrome as well as susceptibility to other conditions such as depression, pneumonia, and influenza. In addition, activities of daily living may be negatively impacted.

Upon investigations by the American Diabetes Association -Virginia, Diabetes is expensive. People with diabetes have medical expenses approximately 2.3 times higher than those who do not have diabetes. Total direct medical expenses for diagnosed diabetes in Virginia were estimated at $6.1 billion in 2017. In addition, another $2.3 billion was spent on indirect costs from lost productivity due to diabetes. Solely, Diabetes was the seventh leading cause of death in the United States in 2019 based on the 87,647 death certificates in which diabetes was listed as the underlying cause of death. In 2019, diabetes was mentioned as a cause of death in a total of 282,801 certificates. It is the second leading Disability Adjusted Life Years (Pastors et al. 2002) reflecting the limiting complications that people with diabetes suffer throughout their lives.

Globally, between 2000 and 2016, there was a 5% increase in premature mortality from diabetes. In April 2021 WHO launched the Global Diabetes Compact, a global initiative aiming for sustained improvements in diabetes prevention and care, with a particular focus on supporting low- and middle-income countries. The Compact is bringing together national governments, UN organizations, non- governmental organizations, private sector entities, academic institutions, philanthropic foundations, people living with diabetes and international donors to work on a shared vision of reducing the risk of diabetes and ensuring that all people who are diagnosed with diabetes have access to equitable, comprehensive, affordable and quality treatment and care.

(Redondo et al. 2017), It is estimated that 537 million people are currently living with diabetes all over the world. By 2045, projections show this number rising to some 783 million

diabetics globally. In May 2021, the World Health Assembly agreed a Resolution on strengthening prevention and control of diabetes. It recommends action in areas including increasing access to insulin; promoting convergence and harmonization of regulatory requirements for insulin and other medicines and health products for the treatment of diabetes; and assessing the feasibility and potential value of establishing a web-based tool to share information relevant to the transparency of markets for diabetes medicines and health products.

1.2 Problem Statement

Diabetes is a chronic disease defined by high blood sugar levels caused by insufficient insulin production or the body's inadequate use of insulin. It is a major public health concern in Virginia, with an increasing prevalence that poses multiple challenges to individuals, healthcare systems, and the general well-being of the population.

- To identify the factors contributing to high risk of diabetes

1.3 Objective Of The Study

Diabetes presence in Virginia will be evaluated using statistical methods that provide clinically important data. Sutham Jirapanakorn's physical examination in 2021 provided the dataset used to build the model in this study. The cases were focused in the Virginia state of Buckingham and Louisa. .

The specific objectives of this study is to:

1. Formulate the logistic regression model for the study
2. To identify significant factors associated with diabetes
3. To predict the likelihood of diabetes.

1.4 Significance Of The Study

In many facets of healthcare, research and public health, the ability to predict diabetes using logistic regression is critically important. The following are some of the most profound impacts: early intervention, personalised medicine, disease management, lowering complications, etc. In outcome, using logistic regression to predict diabetes is a beneficial tool that can help patients have better outcomes, more effective healthcare systems, and a better understanding of the risk factors for the disease. It may have a favorable effect on people's personal health as well as the general public's health.

1.5 Research questions

1. How can logistic regression models determine the significance of the main demographic and lifestyle factors that affect the likelihood of developing diabetes?

2. How well does logistic regression perform when used in conjunction with a number of risk factors, such as age, BMI, family history, and dietary habits, to predict the onset of Type 2 diabetes?

3. Based on relevant variables like maternal age, weight gain, and glucose levels, can logistic regression models accurately distinguish between people at low, moderate, and high risk of developing gestational diabetes during pregnancy?

1.6 Organization Of The Study

The research will be divided into five chapters. We'll get into the introduction in Chapter one. This will include a review of the study's background, identification of the research problems, clarification of the study's objectives, consideration of research issues, clarification of the study's significance, establishment of its scope and limitations, definition of key terms and types, and an outline of the study's organization. We will begin the literature review framework in Chapter two after that. We will focus the early researchers' points of view in this chapter and

acknowledge their increases in output. The research methodology will be covered in chapter three. The logistics regression model will be use to analyse the data .The data was a secondary data obtained from Jirapanakron Sutham . The presentation, evaluation, and interpretation of the gathered data will all be covered in chapter four. We will carefully present, examine, and offer interpretations for the data gathered from the study's participants in this chapter. Finally, Chapter five will serve as the study's conclusion. offering a thorough conclusion and a summary of the study's main findings.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

This chapter focuses on a systematic review of the literary words of predictive analysis of diabetes in Virginia. This literature review aims to provide an overview of recent research findings and advancements in the understanding and management of diabetes.

2.2 Diabetes

Diabetes is the fourth major public health problem in the world, according to the National Institutes of Health in 2019. The majority of the food you eat is converted into sugar (glucose) by your body and then released into your bloodstream. When your blood sugar rises, it signals your pancreas to release insulin. Insulin functions as a key to allow blood sugar into cells for use as energy. Diabetes occurs when the body does not produce enough insulin or does not use insulin effectively, resulting in high blood sugar levels. This can cause serious problems over time. Numerous studies have investigated the epidemiology of diabetes, including its prevalence, incidence, and associated risk factors. These studies have identified obesity, sedentary lifestyle, unhealthy diet, family history, and genetic predisposition as key risk factors for the development of diabetes (Chan et al. 2020).

People with diabetes need to manage their blood sugar levels through diet,exercises and medications to prevent these complications. Diabetes can also cause problems with wound healing and can increase the risk factors of infections. Again, they need to monitor their blood sugar levels regularly and work with health care team to manage these condition. Diabetes is a chronic condition that requires ongoing care and attention

2.2.1 Types

There are primarily three types of diabetes: Type 1 diabetes, Type 2 diabetes, and gestational diabetes. Each type has distinct characteristics, causes, and management approaches. Here is an overview of each type

- Type 1

  According to the National Institute of Diabetes and Digestive and Kidney Diseases in 2022, Type 1 diabetes, also known as juvenile diabetes or insulin-dependent diabetes, is an autoimmune disease in which the immune system attack the cells in the pancreas that produce insulin. This lead to lack of insulin production which means that glucose can not enter the cells where it is needed for energy. According to National Health Organization, around 10 percent of people are all type 1 diabetes. In type 1 diabetes, the pancreas (a small gland behind the stomach) progressively reduces the amount of insulin (the hormone that regulates blood glucose levels) it produces until it stops producing any at all. If the amount of glucose in the blood is too high, it can seriously damage the body's organ over time.

  Knowledge of type 1 diabetes has rapidly increased over the past 25 years, resulting in a broad understanding about many aspects of the disease, including its genetics, epidemiology, immune and -cell phenotypes, and disease burden (Foster et al. 2019). People with type 1 diabetes need to take in insulin injection or use insulin pump to manage their blood sugar level. Type 1 diabetes usually develops in children between 4 and 6 or adolescence, but can occur at any age (Maahs et al. 2010). The exact cause of type1 diabetes is not known, but it is thought to be related to both genetic and environmental factor's.

- Type 2

  According to the International Diabetes Federation (IDF) in 2019, approximately 463 million adults (20-79 years) were living with diabetes worldwide. Type 2 diabetes accounts for around 90-95 percent of all diabetes cases.The prevalence of type 2 diabetes is increasing rapidly, primarily due to lifestyle changes, urbanization, and obesity. Type 2

diabetes is a condition in which the body becomes resistant to insulin or does not produce enough insulin. This lead to buildup of glucose in the bloodstream leading to variety of health problems.

According to the National Health Service in 2021, Type 2 diabetes can be managed through lifestyle changes such as a healthy diet, regular physical activity, weight management, and smoking cessation. Individual needs may dictate the use of medications such as oral anti-diabetic drugs and injectable insulin. Regular blood glucose monitoring and check-ups are required for effective management and prevention of complications.

- Gestational diabetes Gestational diabetes (GDM) is defined as a glucose intolerance resulting in hyperglycaemia of variable severity with onset during pregnancy (Baz et al. 2016) It is a type of diabetes that occur during pregnancy. It occurS by hormonal changes that can make the body less sensitive to insulin. Gestational diabetes usually develops in the second trimester or third trimester of pregnancy and causes high blood sugar levels in the mother and baby.
  (Baz et al. 2016) Women who are overweight, have a family history of diabetes or have had gestational diabetes. Treatment of gestational diabetes may include changing diet and regular exercise and in some case medication or insulin therapy. Most women with gestational diabetes are able to manage their blood sugar levels and have health pregnancies and babies.

2.2.2 Diabetes Worldwide

Diabetes is the leading global health concern affecting people of all ages and backgrounds according to the World Health Organization (WHO) in 2019. According to the International Diabetes Federation in 2019, an estimated 463 million adults (20-79) worldwide suffer from diabetes, accounting for 9.3 percent of the global adult population. If current trends continue, the number of people with diabetes is expected to rise to 578 million by 2030 and 700 million by 2045.The majority of diabetics live in low and middle-income countries with limited access to health care and diabetes management resources. Estimates from the International Diabetes

Federation (IDF) in 2019 show that approximately 463 million adults have diabetes.

Globally, The most recent meta-analysis (Munir et al. 2022) the global prevalence of Gestational Diabetes was 14.7 percent based on the International Association of Diabetes and Pregnancy Study Groups (IADPSG) criteria is affected by gestational diabetes . However, the prevalence can be significantly higher in certain populations. For example, in the United States, the Centers for Disease Control and Prevention (CDC) reported that the prevalence of gestational diabetes was approximately 6-9 percent of pregnancies.

The World Health Organization has identified diabetes as a priority health issue and has called for increased efforts to prevent and manage the disease,including promoting healthy lifestyle habits and improving access to diabetes care and resources.

2.2.3 Diabetes in Virginia

Diabetes which was the seventh leading cause of death in Virginia with 1,638 death attributed to the disease by Department of health, Virginia in 2017. According to Department, an estimated 11.3 percent of adults in Virginia had diagnosed diabetes in 2018. This represents an increase from 9.2 percent in 2011.

(Khan et al. 2021) the prevalence of diabetes in Virginia using data from the Behavioral Risk Factor Surveillance System (BRFSS), an overall prevalence of diabetes in Virginia at 9.2 percent in the year 2011. The prevalence varied by age group, with older adults having higher rates. The Department of Health,Virginia published an annual report called "Virginia Chronic Disease and Risk Factor Surveillance" that provides information on the prevalence of various chronic diseases, including diabetes. It indicated that approximately 10.3 percent of adults in Virginia had diagnosed diabetes. Another study published in the Journal of Community Health in 2019 focused on the prevalence of diabetes in rural Virginia. The study found that the prevalence of diabetes was higher in rural areas compared to urban areas of the state. African Americans and His panic American in Virginia are at higher risk of developing diabetes

than non-His panic whites.

The Virginia Diabetes Council, a nonprofit organization dedicated to diabetes education and advocacy in Virginia, provides resources and information related to diabetes prevalence in the state. They indicate that diabetes affects over 700,000 individuals in Virginia, accounting for approximately 10 percent the population. Diabetes alone has its own negative side effects, but is also a risk factor for heart disease, stroke, amputations, blindness, and kidney disease. This along with the financial implications, are why prevention is so important. People with prediabetes can prevent the onset of the disease with certain lifestyle changes and people already diagnosed with it can better learn how to manage their symptoms and prevent them from getting worse.

The CDC funds two grants in 1815 and 1817 to assist states, including Virginia, with diabetes prevention and education. Diabetes Prevention Program (DPP) and Diabetes Self-Management and Education (DSME) are both used. These programs, which have been shown to be effective for diabetes prevention and management, involve meeting with a trained or certified professional and a small group of people who also have prediabetes/diabetes. In 2019, Virginia had 44 DPP programs that reached 18,610 diabetics, which was approximately 5% more than in 2018. While this is encouraging, the Department of Health,Virginia continues to collaborate with external partners to expand both programs and target those people who are most at risk of developing diabetes or experiencing serious complications.

2.2.4 Incidence and Trends of Diabetes

- Incidence and Prevalence of Diabetes: Diabetes is a chronic metabolic disorder characterized by elevated blood sugar levels. It can be categorized into two main types: type 1 diabetes, which is typically diagnosed in childhood or adolescence and is characterized by the destruction of insulin-producing cells in the pancreas, and type 2 diabetes, which

is more common and usually develops in adulthood due to a combination of insulin resistance and impaired insulin production.

According to the International Diabetes Federation (IDF) in 2021, approximately 463 million adults (20-79 years) were living with diabetes worldwide. The global prevalence of diabetes is estimated to be 9.3% in this age group. It is projected that by 2045, the number of adults with diabetes will rise to 700 million.

- Incidence and Trends of Type 1 Diabetes

  Type 1 diabetes is considered an autoimmune disease in which the immune system attacks and destroys the insulin-producing cells in the pancreas. The incidence of type 1 diabetes varies among different populations and is influenced by genetic and environmental factors. The incidence of type 1 diabetes has been increasing globally over the past few decades. According to the International Diabetes Federation, between 2011 and 2021, the incidence of type 1 diabetes increased by 1.8% annually worldwide. It is particularly common in countries with a higher standard of living and is more frequently diagnosed in children and adolescents.

- Incidence and Trends of Type 2 Diabetes

  Type 2 diabetes accounts for the majority of diabetes cases worldwide and is closely linked to obesity, sedentary lifestyles, and poor dietary habits. The incidence and prevalence of type 2 diabetes have been rising rapidly, mainly due to increasing rates of overweight and obesity and an aging population.

  The global incidence of type 2 diabetes has been increasing at an alarming rate. According to the IDF, between 2011 and 2021, the incidence of type 2 diabetes increased by 5.5% annually worldwide. This increase is observed in both high-income countries and low- and middle-income countries. Type 2 diabetes is no longer limited to older adults but is also increasingly affecting younger individuals, including children and adolescents.

- Regional and Ethnic Disparities

  Diabetes prevalence and incidence rates vary across different regions and ethnic groups. Some populations have a higher susceptibility to diabetes due to genetic and environmental factors. For example, certain ethnic groups such as South Asians, Hispanics, and

Native Americans have a higher risk of developing Type 2 diabetes compared to other populations. Additionally, disparities in access to healthcare, socio-economic factors, and lifestyle behaviors contribute to variations in diabetes rates among different regions and ethnicities.

2.2.5 Risk Factors On Diabetes

Certainly! Here is further explanation and information on the risk factors of diabetes, including cholesterol, age, gender, weight, height, and high-density lipoprotein (HDL), along with examples for each factor

Cholesterol: Abnormal cholesterol levels, specifically high levels of low-density lipoprotein (LDL) cholesterol and triglycerides, and low levels of high-density lipoprotein (HDL) cholesterol, are associated with an increased risk of diabetes. High LDL cholesterol and triglycerides contribute to insulin resistance and impair glucose metabolism, increasing the likelihood of developing type 2 diabetes American Heart Association in 2021. For example, (Peiris et al. 2018) found that individuals with high LDL cholesterol and low HDL cholesterol had a higher risk of developing type 2 diabetes compared to those with healthier lipid profiles.

Age: Advancing age is a significant risk factor for diabetes, particularly type 2 diabetes. The incidence of diabetes increases with age due to factors such as reduced physical activity, increased insulin resistance, and gradual deterioration of pancreatic function according to the American Diabetes Association in 2021. For instance, (Sun et al. 2022) demonstrated a clear age-related increase in the incidence of Type 2 diabetes, with the highest rates observed in older individuals.

Gender: Gender plays a role in diabetes risk, although the impact differs between type 1 and type 2 diabetes. Historically, men had a higher risk of developing type 2 diabetes than women. However, this gender difference has become less prominent in recent years due to changes in lifestyle and obesity prevalence (Zheng et al. 2018). In contrast, type 1 diabetes tends to be

slightly more prevalent in males, particularly during childhood and adolescence (Harjutsalo et al. 2008). For example, (Wang et al. 2017) found that among individuals diagnosed with type 1 diabetes before the age of 30, the incidence was slightly higher in males compared to females.

Weight: Excess weight, especially abdominal obesity, is a significant risk factor for developing type 2 diabetes. Adipose tissue releases pro-inflammatory substances and hormones that interfere with insulin signaling and glucose regulation (American Diabetes Association, 2021). For instance,()drouin2019changes revealed that individuals with a higher body mass index (BMI) had an increased risk of developing type 2 diabetes compared to those with a normal BMI.

Height: Research suggests that taller stature is associated with a higher risk of developing diabetes, particularly type 2 diabetes. Multiple studies have demonstrated a positive correlation between height and diabetes risk, although the underlying mechanisms remain unclear (Shin & Song 2015). For example, (Karamanos et al. 2002) reported that taller individuals had an elevated risk of developing type 2 diabetes compared to shorter individuals, even after accounting for other risk factors.

High-Density Lipoprotein (HDL): Low levels of HDL cholesterol, often referred to as "good cholesterol," are associated with an increased risk of diabetes. HDL cholesterol plays a crucial role in lipid metabolism and helps remove excess cholesterol from cells, contributing to improved insulin sensitivity (American Heart Association, 2021). (Lee et al. 2020) found that individuals with low HDL cholesterol levels had a higher risk of developing type 2 diabetes compared to those with higher levels of HDL cholesterol.

2.2.6 Prognostic Factors Of Survival

Prognostic factors are the measurements available at the time of diagnosis that are associated with diabetes or overall survival. The various measurements are explained below:

Glycemic Control: Maintaining optimal glycemic control is crucial for improving survival rates in individuals with diabetes. High blood glucose levels have been associated with an increased risk of complications and mortality. Regular monitoring of blood glucose levels and adherence to appropriate diabetes management strategies, such as medication, diet, and exercise, are essential in achieving good glycemic control.

Cardiovascular Risk Factors: Diabetes is closely linked to an elevated risk of cardiovascular diseases (CVDs). Prognostic factors that influence survival in diabetes include hypertension, dyslipidemia, obesity, and smoking. Effective management of these risk factors through lifestyle modifications and appropriate medications can reduce the risk of CVD-related mortality in individuals with diabetes.

Diabetic Kidney Disease: Diabetic kidney disease (DKD) is a common and severe complication of diabetes. It significantly impacts survival rates in affected individuals. Prognostic factors for survival in DKD include the degree of albuminuria, estimated glomerular filtration rate (eGFR), blood pressure control, and the presence of other comorbidities. Early detection and management of DKD, including strict blood pressure control and the use of renin-angiotensin-aldosterone system inhibitors, are vital in improving outcomes.

Diabetes-related Complications: The presence of diabetes-related complications, such as diabetic retinopathy, neuropathy, and foot ulcers, can negatively affect survival rates. Timely identification and appropriate management of these complications, including regular eye examinations, foot care, and neuropathy screening, are crucial in preventing further deterioration and improving prognosis.

2.2.7 Clinical Treatment On Diabetes

Certainly! Here is further explanation and information on the clinical treatment of diabetes, including cholesterol, age, gender, weight, height, and high-density lipoprotein (HDL),

Cholesterol :In Statin therapy, Medications such as atorvastatin or simvastatin are commonly prescribed to manage high cholesterol levels. Also with Lifestyle modifications, Dietary changes, regular exercise, and weight management can help reduce cholesterol levels.

Age : Individualized management: As age increases, personalized treatment plans should be developed, taking into account the patient's overall health, comorbidities, and goals of care.Regular health screenings: Age-specific screenings for diabetes and other related conditions should be conducted regularly to monitor health status

Gender: Treatment for diabetes risk factors is not gender-dependent. However, women with gestational diabetes may require specific interventions during pregnancy. Concerning specific considerations, Healthcare providers may consider hormonal influences, reproductive health, and other gender-related factors when managing diabetes risk in women.

Weight loss interventions: Lifestyle changes including a balanced diet, increased physical activity, and behavioral interventions can aid in weight reduction and diabetes prevention. Bariatric surgery: In cases of severe obesity, bariatric surgery may be considered as a treatment

High-density lipoprotein (HDL) levels: Regarding Lifestyle modifications, regular exercise, a healthy diet (including sources of monounsaturated fats), moderate alcohol consumption, and smoking cessation can help increase HDL levels. Also with medications,in certain cases,medicines such as fibrates may be prescribed to raise HDL levels.

Chapter 3

METHODOLOGY

3.1 Introduction

The aim of this study is to investigate the predictive analysis of diabetes using logistic regression analysis based on secondary data. The research question centers around examining the relationship between various demographic, lifestyle, and clinical variables and the presence or absence of diabetes. By analyzing existing data, we seek to gain insights into the prevalence and predictors of diabetes at Virginia

3.2 Specification Of The Study Area

The secondary data for this study is obtained from Jirapanakorn Sutham conducted by the Public Health at Imperial College London. Imperial College London is a prestigious institution known for its excellence in education and research with good records of data on diabetes .The dataset used is the most recent available , covering years 2021 to 2023 to ensure a sufficiently large sample size.

3.3 Variable Selection And Data Cleaning and Preparation

Variables of interest include the dependent variable, "diabetes status" (presence or absence), as well as a set of independent variables such as patient's age, gender, smoking status, blood pressure, hemoglobin A1c, BMI, location, insurance, LDL and HDL level as predictors. These variables were chosen based on their known associations with diabetes prevalence in the literature. The secondary data obtained from the Jirapanakorn Sutham will undergo thorough

cleaning and preparation. This includes identifying and handling missing values, outliers, and inconsistencies. Missing data will be addressed using appropriate techniques such as backwards eliminations method based on the extent of missingness. Variable recoding and transformations will be performed to ensure consistency and compatibility for logistic regression analysis.

3.4 Experimental Design And Subject

Logistic regression analysis will be conducted to estimate the odds ratios and associated confidence intervals for the selected independent variables in relation to the prevalence of diabetes. The logistic regression model will be fitted, with "presence or absence of diabetes" as the dependent variable and the selected independent variables as predictors. Adjustments for potential confounding variables, such as age and BMI, will be made as necessary.The goodness-of-fit of the logistic regression model will be evaluated using appropriate statistical measures such as the Likelihood ratio test and McFadden's R-Squared test. These assessments will gauge how well the model fits the observed data and whether there is evidence of model misfit. Furthermore, odds ratios, p-values, and hypothesis testing will be utilized to interpret the strength and significance of the associations between the independent variables and diabetes prevalence.

3.5 Data analysis and model

3.5.1 Logistic Regression

Logistic regression is a statistical modeling technique used to analyze the relationship between one or more independent variables and a binary or categorical dependent variable. It is particularly useful when the outcome variable is dichotomous, meaning it has two possible outcomes (e.g., yes/no, success/failure, presence/absence). The objective of logistic regression is to estimate the probability of the occurrence of an event or the presence of a specific outcome based on the values of the independent variables. Unlike linear regression, which models continuous dependent variables, logistic regression models the log-odds or logit transformation of the probability. The logistic regression model can be expressed as follows;

$$\ln\left(\frac{p(y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta p X p \tag{3.1}$$

Logistics regression model is the logit transformation of the probability of the event occurring (p) divided by the probability of the event not occurring (1-p).

β_0, β_1,β_2, ..., β_p are the coefficients or log-odds associated with the independent variables (X_1, X_2, ..., X_p) in the model.

P is the probability of the event occurring given the values of the independent variables.

The logistic regression model assumes that the relationship between the independent variables and the log-odds of the dependent variable is linear. However, the relationship between the independent variables and the probability of the event occurring is nonlinear due to the logistic transformation.

The coefficients (β_0,β_1,β_2, ...,β_p) in the logistic regression model represent the effects of the independent variables on the log-odds of the dependent variable.

These coefficients can be exponentiated to obtain the odds ratios associated with the independent variables. The odds ratio represents the change in odds (probability of success divided by the probability of failure) for a one-unit change in an independent variable while holding other variables constant.

The logistic regression model can be estimated using maximum likelihood estimation (MLE) techniques, which find the parameter values that maximize the likelihood of the observed data. Once the model is estimated, the coefficients can be tested for statistical significance, and the model's fit can be assessed using various measures such as the Hosmer-Lemeshow test, deviance, AIC, and BIC.

The logistic regression model can be used for prediction by plugging in values of the independent variables into the model equation to obtain the estimated log-odds or probabilities. Inference involves interpreting the coefficients, odds ratios, and their confidence intervals to understand the associations between the independent variables and the probability of the event occurring. Overall, the logistic regression model is a powerful tool for analyzing binary or categorical outcomes and understanding the relationships between variables in various fields of study.

3.5.2 Assumptions

Logistic regression is a statistical method used for binary classification problems, where the outcome variable (dependent variable) is binary or dichotomous (e.g., 0 or 1, true or false). Here are the key assumptions of logistic regression:

1. Binary dependent variable:

   There should only be two possible outcomes or categories for the dependent variable, which is known as a dichotomous variable.

2. Linearity of the logit

   The logit (log-odds) of the dependent variable and the independent variables should be linearly related. The logistic regression model assumes that the connection between the independent variables and the logit is linear because the logit is the natural logarithm of the odds ratio.

3. Independence of observations

   Observations ought to be made separately from one another. To put it another way, there shouldn't be any dependence or correlation between the data values. A lack of independence may result in skewed parameter estimates.

4. No multicollinearity

   The independent variables should not have any multicollinearity. When there is a strong correlation between two or more independent variables, multicollinearity occurs. It may be challenging to isolate the specific influences of each independent variable on the dependent variable.

5. Large sample size

   With a big sample size, logistic regression often performs well. The estimates and standard errors for the model parameters are more precise when the sample size is higher.

6. Linearity of the independent variables and logit

   Each independent variable should have a roughly linear connection with the dependent variable's logit. The model must adhere to this premise in order to successfully generalize to new data.

3.5.3 The Chi-Square Test

In logistic regression, the deviation between observed values and predicted values is utilized as the statistic for overall fit of the linear regression model rather than R^2. Where P_i is the observed dependent variable for the ith subject and P_i is the corresponding prediction from the model, residuals in linear regression are defined as $(P_i - \hat{P}_i)$. In logistic regression, when P_i is equal to either 0 or 1, the same idea is applicable, and the equivalent prediction from the model is given as ;

$$\hat{P}_i = \pi = \frac{e^{\beta_o + \beta_1 X_1 + ... + \beta_k X_k}}{1 + e^{\beta_o + \beta_1 X_1 + ... + \beta_k X_k}} \tag{3.2}$$

The residuals, $(P_i - \hat{P}_i)$, can be used as the basis for a chi-square test. When the standard deviation of the residuals is $\hat{P}_i(1 - \hat{P}_i)$, the residual is said to be standardized. The resulting χ^2 statistic may then be formed as follows:

$$\chi^2 = \sum_{i=1}^{n} r_i^2 \tag{3.3}$$

In order to determine p values, this statistic follows an χ^2 distribution with $n(k+1)$ degrees of freedom.

3.5.4 Maximum Likelihood Estimation

Given $\text{logit}(y) = \alpha + \beta x$ resembles a straightforward linear regression model, the parameters and cannot be estimated with OLS since the underlying distribution is binomial. Instead, the maximum likelihood of witnessing the sample values is typically used to estimate the parame-

ters. The likelihood of getting the dataset is maximized by using values of α and β that come from maximum likelihood. Assume that each person in the population we are sample has the same likelihood that an event will occur. For each individual in our sample of size n, $P_i = 1$ indicates that an event occurs for the ith subject, otherwise, $P_i = 0$. The observed data are $Y_1, ..., Y_n$ and $X_1, ..., X_n$. The joint probability of the data (the likelihood) is given by

$$L = \prod_{i=1}^{n} p(y/x)^{P_i}(1 - p(y/x))^{1-P_i} = p(y/x)^{\sum_{i=1}^{n} Y_i}(1 - p(y/x))^{n-\sum_{i=1}^{n} Y_i} \tag{3.4}$$

Natural logarithm of the likelihood is

$$l = \log(L) = \sum_{i=1}^{n} Y_i \log[p(y/x)] + \left(n - \sum_{i=1}^{n} Y_i\right) \log[1 - p(y/x)] \tag{3.5}$$

In which

$$p(y/x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{3.6}$$

The result is the maximum likelihood estimates of α and β. (Park, Hyeoun-Ae, 2013)

3.5.5 Odds And Odds Ratio

Odds: The likelihood or proportion that an event will occur divided by the probability that it will not occur is known as the odds of an occurrence. If an event's probability of occurrence is π, then its probability of not happening is $(1 - \pi)$. The chances that coincide are then determined by

$$odds\ of\ an\ event = \frac{\pi}{1 - \pi} \tag{3.7}$$

The influence of independent variables is typically described in terms of odds since logistic regression estimates the chance of an event occurring over the probability of an event not occurring. In logistic regression, the result variable's mean is expressed in terms of an explanatory

variable x using the equation

$$p = \beta_o + \beta_1 X_1 \tag{3.8}$$

Unfortunately, this is a flawed model since extreme x values will result in $\beta_o + \beta_1 X_1$ values outside of the range between 0 and 1. In the logistic regression solution to this problem, the probabilities are transformed using the natural logarithm. As a linear function of the explanatory factor, we use logistic regression to describe the natural log chances:

$$logit(\pi) = \ln(odds) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_o + \beta_1 X_1 + ... + \beta X_k \tag{3.9}$$

where π is the probability of interested outcome and x is the explanatory variable. The parameters of the logistic regression are β_0 and β_i . This is the simple logistic model. Taking the antilog on both sides, one can derive an equation for the prediction of the probability of the occurrence of interested outcome;

$$\pi = \frac{e^{\beta_o+\beta_1 X_1+.....+\beta_k X_k}}{1+e^{\beta_o+\beta_1 X_1+.....+\beta_k X_k}} = \frac{1}{1+e^{-(\beta_o+\beta_1 X_1+.....+\beta_k X_k)}} \tag{3.10}$$

where x is the explanatory variable and π is the probability of the result of interest.

Odds Ratio: A tool for contrasting two odds in relation to different outcomes is the odds ratio. The likelihood that two events A and B will occur in close proximity to one another is

$$Odds\ ratio\ A\ vs\ B = \frac{odds\ A}{odds\ B} = \frac{\frac{\pi_A}{1-\pi_A}}{\frac{\pi_B}{1-\pi_B}} \tag{3.11}$$

A measure of the relationship between an exposure and an outcome is an odds ratio. The odds ratio compares the likelihood of an outcome (such as a disease or illness) occurring in the presence of that exposure. (such as a health behavior or medical history) to the likelihood of the outcome occurring in the absence of a specific exposure. In other words, The odds ratio compares how likely it is that an event (such a sickness or illness) will happen in the presence of that exposure compared to the likelihood of the result happening in the absence of a certain exposure. When a logistic regression is calculated, the regression coefficient (β_1) is the expected

rise in the logged odds of the result for each unit increment in the value of the independent variable. In other words, the exponential function of the regression coefficient ($e^{(\beta_1)}$) is the odds ratio that corresponds to a one unit increase in the independent variable. The odds ratio can be used to assess the relative significance of different risk factors for an event and to establish whether a specific exposure is a risk factor for a specific outcome. Thus;

1. The odds ratio is the ratio of the probability of a result occurring with a therapy versus the odds of the outcome occurring without the treatment.

2. Odds are computed by dividing the likelihood of an event happening by the likelihood of an event not happening.

3. Despite widespread assumption, probability and odds are not synonymous. Probabilities range from 0 to 1 (0% to 100%), whereas odds can be any number.

4. An odds ratio of one (1) implies that the exposure has no effect on the outcome odds; odds ratios less than one (1) indicate that the exposure is associated with lower outcome odds; and odds ratios greater than one (1) indicate that the exposure is associated with higher outcome odds.

The odds ratio can be used to establish whether a specific treatment is a "risk factor" for a particular outcome. To compute the odds ratio, the frequencies of two dichotomous variables must be known.

3.5.6 Hypothesis Testing

Hypothesis testing will be used in this study to determine the most significant variables when predicting the diabetes status in an individual

H_o : The variable coefficients are not significant.

H_A : The variable coefficients are significant.

If the p-value associated with each independent variable is less than the significant level, we

reject the null hypothesis (H_o) and conclude that the variable has a significant impact in predicting diabetes in this study.

3.5.7 Model Diagnostics

Model diagnostics typically refer to the evaluation and assessment of machine learning models to understand their performance and potential issues. It involves various techniques and tools to analyze how well a model is doing and identify areas for improvement. Common model diagnostic techniques include:

- A confusion matrix is a fundamental tool in classification problems, especially in machine learning. It's a table that is used to evaluate the performance of a classification model by comparing its predicted classifications against the actual known ground truth. The confusion matrix has four main components:

Table 3.1: Confusion Matrix

	Test positive	Test negative
No.of Disease	True positive	False negative
No. without Disease	False positive	True negative

- True Positives (TP): These are cases where the model correctly predicted the positive class (i.e., it predicted "yes" when the actual answer was "yes").

- True Negatives (TN): These are cases where the model correctly predicted the negative class (i.e., it predicted "no" when the actual answer was "no").

- False Positives (FP): These are cases where the model incorrectly predicted the positive class (i.e., it predicted "yes" when the actual answer was "no"). Also known as Type I errors.

- False Negatives (FN): These are cases where the model incorrectly predicted the negative class (i.e., it predicted "no" when the actual answer was "yes"). Also known as Type II errors.

Checking Assumptions:

Ensure that logistic regression assumptions, such as linearity of the log-odds, are met. Plotting residuals against predictors can help identify potential issues.

Multicollinearity:

Check for multicollinearity among predictor variables, as high correlations can affect the model's stability and interpretability. You can use variance inflation factors (VIFs) to assess multicollinearity.

Variable Selection:

Consider the significance of predictor variables using p-values or other criteria. Remove variables that are not statistically significant if they do not contribute to the model's performance.

Residual Analysis:

Examine the residuals of the logistic regression model. You can use deviance residuals or Pearson residuals to check for any patterns or outliers. A well-fitted model should have residuals that are approximately normally distributed.

Log-Likelihood and Deviance:

Evaluate the log-likelihood and deviance of your logistic regression model. Lower deviance values suggest a better fit to the data.

3.5.8 Logit Plot

A logit plot is a graph that displays the logit-transformed probability (logit(p)) on the vertical axis and one or more predictor variables on the horizontal axis. Each data point on the plot corresponds to an observation in the dataset. In a logit plot, you can visualize how the log-odds of the event occurring change as the predictor variable(s) change. This helps you understand the nature and strength of the relationship between the predictors and the event. If the logit plot shows a clear linear relationship between the logit and the predictor(s), it suggests that

the logistic regression model is a good fit for the data. If the logit plot displays a non-linear pattern, it may indicate that the relationship is more complex than a simple logistic regression model can capture, and further model refinement or feature engineering may be necessary.

3.5.9 Likelihood Ratio

The likelihood ratio is a fundamental concept in statistics, particularly in the context of hypothesis testing and model comparison. It's used to assess the relative likelihood of two different hypotheses or models given observed data. The likelihood ratio is defined as the ratio of the likelihood of the data under one hypothesis or model compared to another.
Where

- Null Hypothesis (Ho): This is the default or status quo hypothesis. It represents a specific statement or assumption about the population or data, often suggesting no effect or no difference.
- Alternative Hypothesis H_A: This is the hypothesis that researchers are typically interested in proving or supporting. It represents an alternative explanation or effect different from the null hypothesis.

The likelihood ratio is calculated as follows:

$$likelihoodratio = \frac{\textbf{likelihood ratio of the data } H_o}{\textbf{likelihood ratio of the data } H_A} \tag{3.12}$$

Where

- Likelihood of the Data Under H_o : This is the probability of observing the given data if the null hypothesis is true.
- Likelihood of the Data Under H_A : This is the probability of observing the given data if the alternative hypothesis is true.

In hypothesis testing, the likelihood ratio is often used to compare these two hypotheses. There are different statistical tests based on the likelihood ratio, such as the likelihood ratio test

(LRT). The LRT compares the likelihood ratio to a distribution (usually chi-squared) to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

In the context of model comparison, the likelihood ratio can help assess which of the two or more models better explains the observed data. It quantifies how much more likely one model is compared to another, given the data.

3.5.10 McFadden's R Squared

McFadden's pseudo R-squared, is a statistic commonly used in the context of logistic regression and other models with likelihood-based goodness-of-fit measures. It was developed by Daniel McFadden as a way to assess the goodness of fit of a logistic regression model.

Unlike traditional R-squared values used in linear regression, which measure the proportion of variance explained by the model, McFadden's pseudo R-squared is based on the likelihood function. It compares the log-likelihood of the model you've fitted to the log-likelihood of a null model (a model with no predictors, only an intercept term). The formula for McFadden's pseudo R-squared is as follows:

$$R^2_{\text{McFadden}} = 1 - \frac{\textbf{Log-Likelihood of Model}}{\textbf{Log-Likelihood of Null Model}} \tag{3.13}$$

where

- Log-Likelihood of Model: The log-likelihood of your logistic regression model given the observed data.
- Log-Likelihood of Null Model: The log-likelihood of a null model with no predictors, essentially a model that predicts the outcome based solely on the overall probability of the event happening.

McFadden's pseudo R-squared typically ranges from 0 to 1, where 0 indicates that the model fits no better than the null model (no improvement in prediction), and 1 indicates a perfect fit where the model explains all of the variation in the data.

3.5.11 Homer-Lemeshow Test

The Hosmer-Lemeshow test (HL test) measures the goodness of fit of logistic regression models, particularly risk prediction models. A goodness of fit test determines how well your data matches the model. The HL test determines whether observed event rates match expected event rates in population subgroups. The test is only applicable to binary response variables (variables with only two outcomes, such as alive or dead, yes or no). The Hosmer-Lemeshow test is shown below:

$$X^2 = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - \frac{E_j}{n_j})}$$

where;

- X^2 = chi squared.
- n_j = number of observations in the jth group.
- O_j = number of observed cases in the jth group.
- O_j = number of expected cases in the jth group.

Chapter 4

Data Analysis

4.1 Introduction

In this study, diabetes was the main dependent variable of interest. Understanding the factors that could cause diabetes in an individual or patient is important. The independent variables were categorized using categorical and continuous groupings.To learn more about the dataset, a thorough Exploratory Data Analysis (EDA) has been conducted. Investigating the relationships between the variables involved looking at the distributions of the variables. Several statistical techniques and visualizations have been used with EDA to better understand the data.Since then, a multiple logistic regression analysis has been performed. The statistical method of logistic regression is used to model a categorical dependent variable (in this case, diabetes status, either having diabetes or not) and one or more independent variables.

4.2 Description of Variables

The variables used in the analysis are categorized into two groups namely;

- Categorical variables. The categorical variables used in the study were, smoking, diabetes status, location and insurance.

- Continuous Variables. The continuous variables included age, gender, systolic and diastolic blood pressure, cholesterol and body mass index.

Response Variable

The response variable is the diabetes status in the study. The table below explains the frequency of the disease among the study participants.

Table 4.1: Diabetes status

Description	Present	Absent
Total	30	100
Percentage (%)	23.08	76.92

The results shows that 30 people with 23.08% of the sample, were diagnosed with the condition out of the total number of participants. However, 100 people with 76.92% of the sample, did not have diabetes. This represents the majority.

Gender Piechart

The gender distribution of the study participants is shown in Figure 4.1.

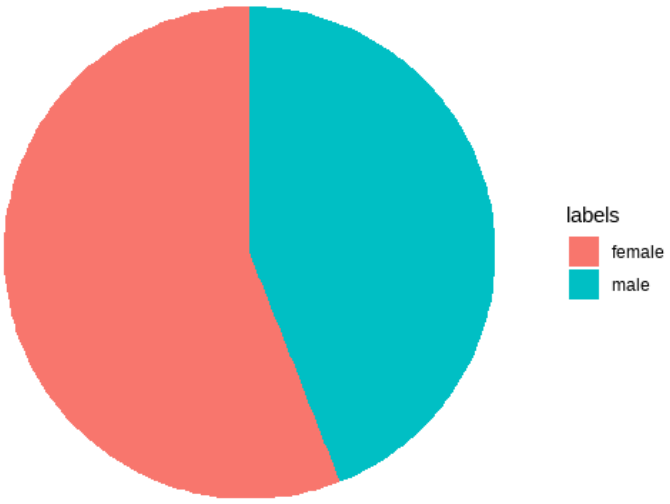

Figure 4.1: Age distribution pie chart

According to the data, 56% of the participants are female and 44% are male. The results show that there are more females in the study population than males.

Diabetes among the location in the study

Table 4.2 below provides an overview of the distribution of diabetes status based on the location of the study participants.

Table 4.2: Diabetes among the location

Location	Diabetes Mellitus	Total	Percentage(%)
Buckingham	Present	13	28.26
	Absent	33	71.74
Louisa	Present	17	20.24
	Absent	67	79.76

Specifically, the table shows the percentage of individuals with diabetes in two locations: Buckingham and Louisa. In Buckingham, 28.26% of the participants have been identified with diabetes, indicating a considerable proportion of the population in this area is affected by the disease. On the other hand, the majority, constituting 71.74% of the participants in Buckingham, are free from diabetes. In the location of Louisa, the prevalence of diabetes is slightly lower, with 20.24% of the participants being diagnosed with the disease. The remaining 79.76% of individuals in Louisa do not have diabetes.

Diabetes Status by Smokers

The table below reveals the participants with or the without the disease according to their smoking status.

Table 4.3: Diabetes Status By Smokers

Variable	Diabtes status	Total
smoking_1	Present	24
	Absent	9
smoking_2	Present	57
	Absent	16
smoking_3	Present	19
	Absent	5

Table 4.4 explains the three levels of smoking, smoking_1 (a never smoker), smoking_2 (a current smoker) and smoking_3 (an experienced smoker). The table also reveals

that participants who have never smoked were diabetic. The table below also reveals that participants who have never smoked were diabetic.

Descriptive Analysis

The below explains the central tendency and the variability of the continuous variables of the dataset used in the study.

Table 4.4: Descriptive Analysis of continuous variables.

Variable	Mean	25%	50%	75%	Min	Max
age	50.77	40.00	51.00	61.75	20.00	89.00
bp.1s	155.6	140.5	150.0	164.2	100.0	230.0
bp.1d	94.45	89.00	94.00	100.00	60.00	124.00
glyhp	5.973	4.393	4.975	6.487	2.850	16.110
bmi	30.21	25.35	29.72	34.01	17.22	51.37
Ldl	217.4	184.2	212.0	241.5	134.0	443.0
hdl	50.52	37.00	46.00	59.75	23.00	120.00

Table 4.4 presents the results of the descriptive analysis conducted on the numerical independent variables in the dataset. The descriptive statistics offer valuable insights into the central tendency and variability of the numerical variables. The average age is 50.77 years which suggests that this study involved older individuals.

Correlation Plot

The correlation plot shown by Figure 4.3 illustrates the linear relationship between the numerical variables in the study.

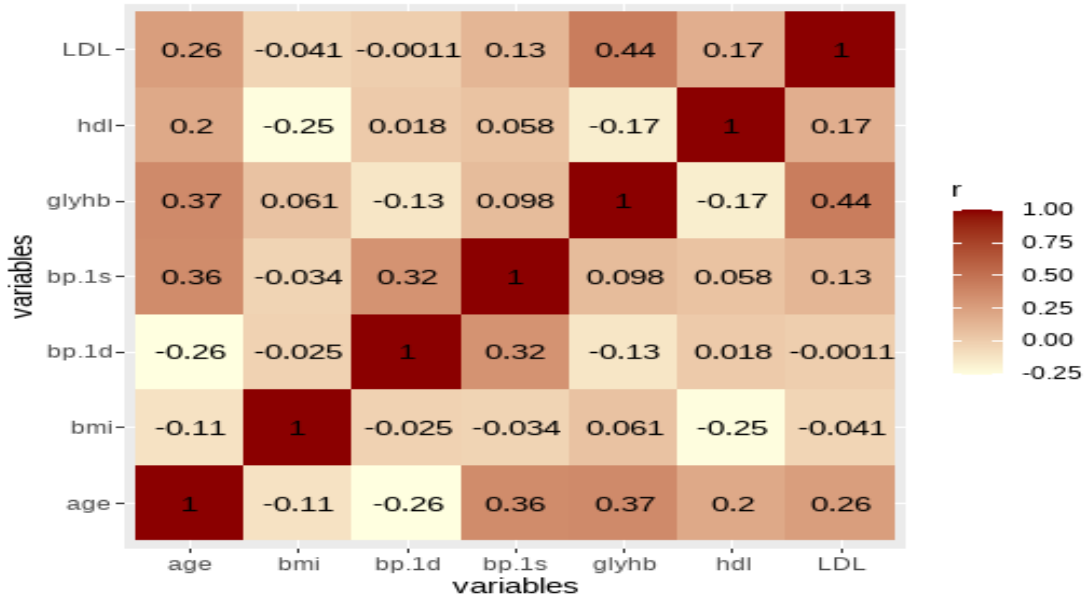

Figure 4.2: correlation plot

According to the findings shown in Figure 4.3, the Age variable has a negative relationship with the BMI variable. This negative correlation suggests that as people get older, their BMI decreases. In other words, older study participants have lower BMI values than younger participants. Furthermore, the cholesterol variable has a negative relationship with both diastolic blood pressure and BMI. This negative correlation implies that higher cholesterol levels in the study population are associated with lower diastolic blood pressure and lower BMI values. However, in this study, the correlations among the variables remain below the critical threshold of 0.7, indicating no significant concerns of high correlation between predictors. This is advantageous for the study's analysis, as it ensures that the relationships between the independent variables and the dependent variable (diabetes status) can be more reliably assessed without the effects of multicollinearity.

Age Histogram

Figure 4.3 below depicts the Age variable Histogram, which shows that the majority of the patients in the study were between the ages of 40 and 70.

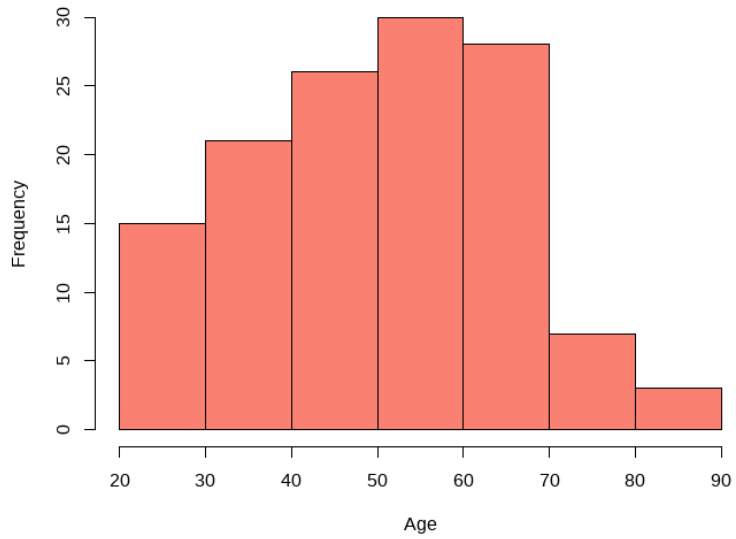

Figure 4.3: Patient's age Histogram

This observation above suggests that the study primarily included older people.

Cholesterol Histogram

The highest number of recorded cholesterol values falls within the range of 150mg/dl to 250mg/dl, as shown by Figure 4.4 below. The histogram's concentration of cholesterol values between 150 mg/dl and 250 mg/dl may represent a higher-than-ideal level of cholesterol in the study subjects.

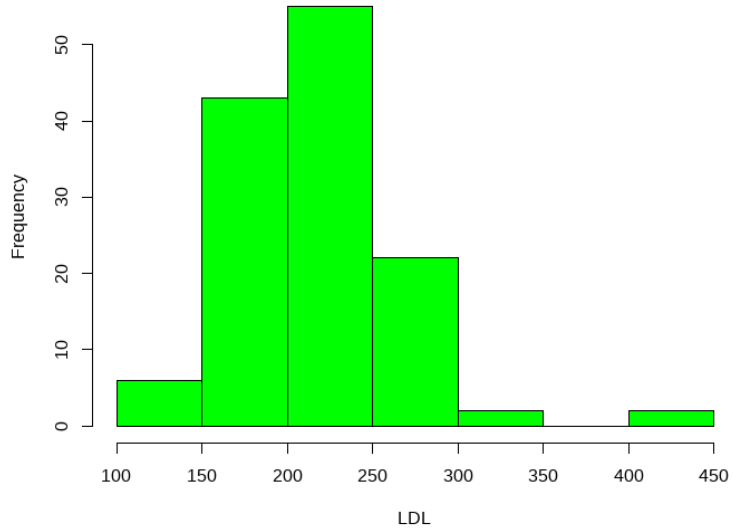

Figure 4.4: Cholestrol Histogram

This result suggests that a significant amount of the study subjects had cholesterol levels that were outside the normal range of cholesterol levels in the human body. Low-density lipoprotein (LDL) cholesterol levels should be less than 150 mg/dl, and high-density lipoprotein (HDL) cholesterol levels should be 40 to 60 mg/dl or more. For healthy adults, total cholesterol levels are frequently advised to be under 200 mg/dl.

4.3 Factors Influencing Diabetes Status Based on The Logistic Regression Output

The possibility of an individual being diabetic is dependent on a number of factors. These factors include age, High Density Lipoprotein (HDL),gender, Body Mass Index (BMI),LDL and smoking. The logistic regression model stated in chapter 3 was used to estimate the probability of an individual being a diabetic. The result of the estimation is shown in a table below.

Table 4.5: Logistic Estimation for Factors Associated with Diabetes (FULL MODEL)

Variables	Parameter	Estimate	Standard Error	P-value
(Intercept)	β_0	-6.440476	-1.820	0.06883
age	β_1	0.072319	0.024171	0.00277 **
bp.1s	β_2	-0.005451	0.014308	0.70325
bp.1d	β_3	-0.013098	0.026347	0.61908
gender	β_4	0.007261	0.597876	0.99031
insurance	β_5	-0.079181	0.332229	0.81162
location	β_6	-0.298817	0.526285	0.57018
Ldl	β_7	0.021439	0.006993	0.00217 **
smoking_2	β_8	0.052762	0.622465	0.93245
smoking_3	β_9	-0.232837	0.787393	0.76745
bmi	β_{10}	0.023999	0.042375	0.57116
hdl	β_{11}	-0.038824	0.016631	0.01957 *

Note: Significant levels are ***p <0.001, ** p <0.01, * p <0.05

According to the Table 4.7, age and Lower density lipoprotein(Ldl) have 0.00277 and 0.00217 p-values respectively. Since their p-values are less then the significance level 0.01, we conclude that age and Lower density lipoprotein(Ldl) are significant at a significance level of 0.01. Also, higher density lipoprotein(hdl), with a p-value of 0.01957, is significant in the logistic regression model at a significance level of 0.01. From the results from the logistic model in Table 4.7 above, the most significant independent variables in predicting diabetes are age, lower density lipoprotein (ldl), and higher density lipoprotein (hdl).

Table 4.6: Logistic Estimation for Factors Associated with Diabetes (REDUCED MODEL)

Variable	Parameter Estimate	Standard Erorr	P-Value	Odds Ratio
Intercept	-7.549110	1.730942	1.29e-05 ***	0.0005
age	0.071648	0.020727	0.000547 ***	1.0743
Ldl	0.020203	0.006458	0.001758 **	1.0204
hdl	-0.042952	0.015536	0.005698 **	0.9580

Table 4.6 as shown above represents the reduced logistic regression after conducting a hypothesis test and choosing the important variables. Age, LDL and HDL are included as independent variables in this simplified state to predict the dependent variable, diabetes status. The table also includes the odds ratio of the significant variables. It is depicted that, after eliminating some predictors, the model coefficient do not change much.

The Age variable's associated coefficient in the reduced logistic regression model is 0.071648. In order to interpret logistic regression coefficients, it is expected that the exponentiated value of the coefficient is estimated (also known as the odds ratio) in order to understand the effect of a one-unit increase in the independent variable on the odds of the dependent variable (diabetes status), while holding other variables constant. Thus it is estimated that the odds of an individual being diagnosed of diabetes is expected to increase by a factor of 1.0743 for a unit increase in age holding LDL and HDL constant.

The odds of an individual being diabetic is expected to increase by a factor of 1.0204 for a unit increase LDL holding age and hdl constant. Again, the exponentiated value of -0.042952 for the HDL is 0.9580. This means that, when age and LDL are held constant, the probability of developing diabetes should decrease by a factor of 0.9580 (approximately 0.95), for a one-unit increase in HDL (e.g., increasing HDL by 1 mg/dL).

4.4 Reduced Model

The reduced model is the logistic regression that contains all the significant variables with their respective coefficients and estimates.

$$p(y=1) = \frac{e^{-7.549110+0.071648*age+0.020203*ldl-0.042952*hdl}}{1+e^{-7.549110+0.071648*age+0.020203*ldl-0.042952*hdl}} \tag{4.1}$$

4.5 Model Diagnostics

In the logit plot shown in figure 4.5, the figure visualizes how the log-odds of the event occurring change as the predictor variable(s) changes. This helps to understand the nature and strength of the relationship between the predictor variables age, bp.1s, bp.1d, BMI, glyph, LDL and hdl and the event which shows a clear linear relationship between the logit and the predictor(s), it

suggests that the logistic regression model is a good fit.

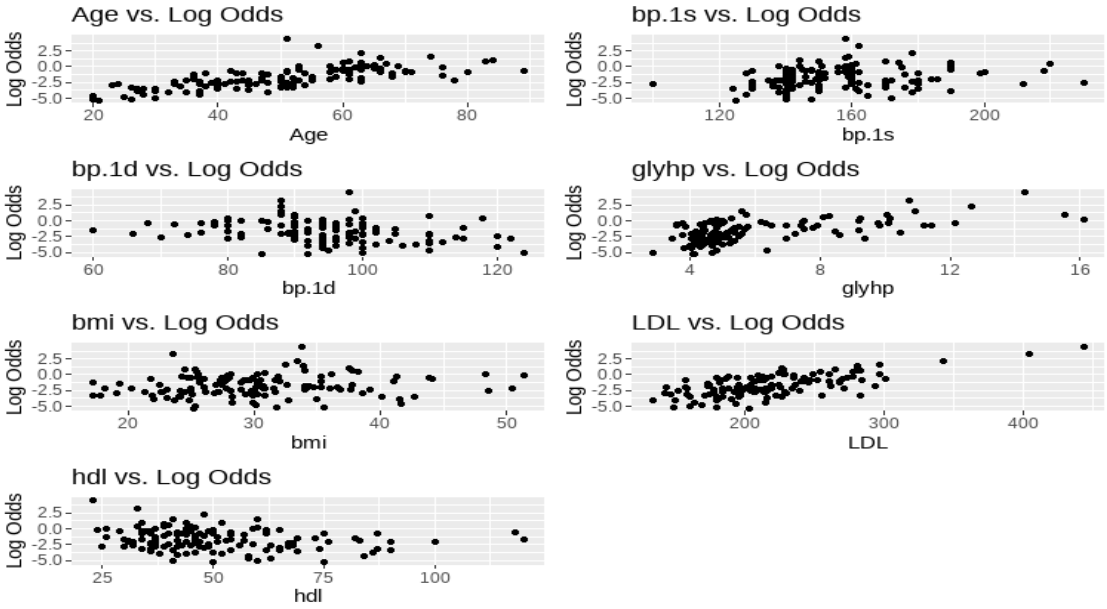

Figure 4.5: Logit Plot

Residual Plot

Figure 4.6 below reflect the differences between fitted and observed values in the study's dataset. Focusing solely on the information about the residual plot,Figure 4.6 it can be concluded that our analysis of the residual plot reveals a crucial insight regarding the appropriateness of the logistic model for our data. The fact that the points are randomly dispersed around zero in the residual plot indicates that the logistic model is a suitable and well-fitting model for the data. This randomness in the residuals suggests that the model adequately captures the underlying relationships within the dataset, without any systematic bias or trend in the model's errors.

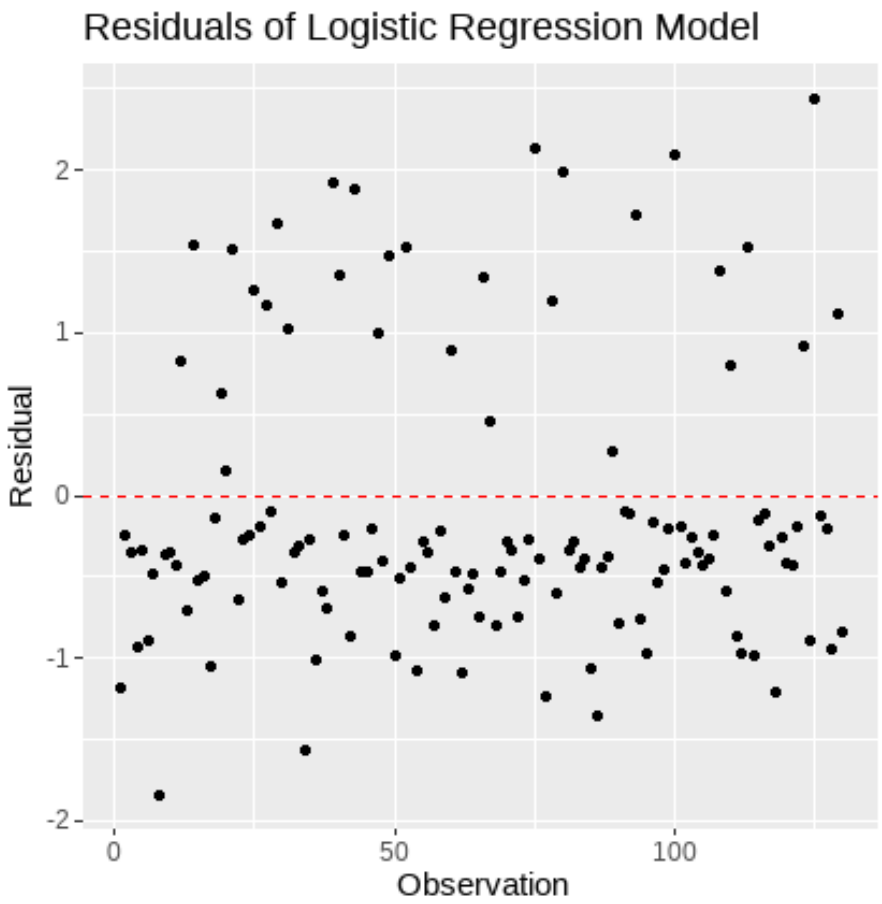

Figure 4.6: Residuals Plots

Therefore, based on the evidence from the residual plot, we can confidently conclude that the logistic model is appropriate for this dataset.

4.5.1 Likelihood Ratio Test

In the context of model comparison, the likelihood ratio can help assess which of the two or more models better explains the observed data.

Table 4.7: Likelihood Ratio Test Results For Model Comparison

Model	Resi.df	Resi.dev	DF	Dev.	P-value
Reduced	126	103.64			
Full	118	102.06	8	1.5785	0.9913

- Null Hypothesis (H_o): The reduced model is a better fit or equally good as the full model.

- Alternative Hypothesis (H_A): The full model is a better fit than the reduced model.

Table 4.7 shows a p-value of 0.9913, with which we can fail to reject the null hypothesis (H_o). This suggests that there is no strong statistical evidence that the full model is a significantly better fit than the reduced model. The data does not support the claim that the additional variables in the full model significantly improve its fit compared to the reduced model.

In conclusion, based on the p-value and the hypotheses, it is reasonable to conclude that the reduced model is as good or possibly better than the full model for explaining the data.

4.6 Test for Goodness of fit

A goodness of fit test determines how well your data matches the model

Mcfadden's R Squared

McFadden's pseudo R-squared, is a statistic commonly used in the context of logistic regression and other models with likelihood-based goodness-of-fit measures.

Table 4.8: McFadden's R squared

Test	Log. Likelihood
Reduced	0.2621366
Full	0.2733753

McFadden's pseudo R-squared values are both positive, indicating that both models explain some of the variation in the dependent variable. When compared to the reduced model, the full model has a slightly higher McFadden's pseudo R-squared value (0.2733753).

Hosmer and Lemeshow Test.

The Hosmer-Lemeshow test (HL test) measures the goodness of fit of logistic regression models, particularly risk prediction models.

Table 4.9: Hosmer and Lemeshaw Test For Goodness of Fit

Test	X^2	P-value
Full	3.8508	0.8703
Reduced	5.0395	0.7533

The p-values for the full and reduced model are 0.8703 and 0.7533 respectively. The model is a good fit if the null hypothesis is failed to be rejected. Since the p-values for both models are greater than the significance level ($\alpha = 0.05$), we conclude that the model is a good fit.

Confusion Matrix For The Full Model

The table below explains the performance metrics of the full model.

Table 4.10: Confusion Matrix For The Full Model

	Present	Absent
Positive Test	19	2
Negative Test	2	3

From Table 4.10, we realized that out of 21 patients that actually had diabetes,the model correctly predicted 19 of them and out of 5 patients that were free from disease, the model correctly predicted 3 of them.

4.7 Performance Metrics For Full Model

Performance Metrics for the full model explains the confusion matrix, precision, recall, and F1 score gives better intuition of prediction results as compared to accuracy.

4.7.1 Sensitivity OR Recall

Sensitivity measures the proportion of correctly predicted positive cases(Diabetic) out of the actual positive cases.From Table 4.10, the value is estimated below:

$$\textbf{Sensitivity} = \frac{19}{19+2} = 0.904762 \tag{4.2}$$

The recall value of 0.9048 means that, the model correctly identified approximately 90.5% of the actual positive cases.

4.7.2 Specificity

Specificity is a measure that tells us what proportion of patients that did not have diabetes, were predicted by the model as non-diabetic.From Table 4.10, it is shown below:

$$\textbf{Specificity} = \frac{3}{3+2} = 0.6 \tag{4.3}$$

The specificity value of 0.600 means that the model correctly identified approximately 60% of the actual negative cases.

4.7.3 Precision

Precision measures the proportion of correctly predicted positive cases (Diabetic) out of the cases predicted as positive(Diabetic). From Table 4.10, this has been depicted below:

$$\textbf{Precision} = \frac{19}{19+2} = 0.9048 \tag{4.4}$$

The Precision value of 0.9048 indicates that out of all the cases the model predicted as present,approximately 90.5% were actually diabetic cases.

4.7.4 F1 Score

The F1 Score is the harmonic mean precision and recall which provides a balanced measure of the model's performance. From Table 4.10, it is shown below:

$$\textbf{F1-Score} = \frac{20.90480.9048}{0.9048 + 0.9048} = 0.9048 \tag{4.5}$$

With an F1 Score of 0.9048, it indicates a very good balance between precision and recall.The value suggests that the model performs well in correctly identifying both positive (present) and negative (Absent).

4.7.5 Model Accuracy

Accuracy measures the overall correctness of the model's predictions.From Table 4.10, the full model's accuracy has been illustrated below:

$$\textbf{Model Accuracy} = \frac{19 + 3}{19 + 3 + 2 + 2} = 0.8462 \tag{4.6}$$

The accuracy value of 0.8462 means that the model predicted the outcome (whether present or absent of diabetes) for approximately 85% of the cases.

Confusion Matrix of The Reduced Model

The table below explains the performance metrics of the reduced model.

Table 4.11: Confusion Matrix For The Reduced Model

	Present	Absent
Positive Test	18	0
Negative Test	5	3

From Table 4.11, we realized that out of 21 patients that actually had diabetes,the

model correctly predicted 18 of them and a total of 3 patients that were free from disease, the model correctly predicted all of them.

4.8 Performance Metrics For Reduced Model

Performance Metrics for the reduced Model explains the confusion matrix, precision, recall, and F1 score gives better intuition of prediction results as compared to accuracy.

4.8.1 Sensitivity

Sensitivity measures the proportion of correctly predicted positive cases(Diabetic) out of the actual positive cases.From Table 4.11, the result is shown below:

$$\textbf{Sensitivity} = \frac{18}{18+5} = 0.7826 \tag{4.7}$$

The sensitivity value of 0. means that, the model correctly identified approximately 78.3% of the actual positive cases.

4.8.2 Specificity

The Specificity measures the actual number of patients that were indeed free from the disease. From Table 4.11, the specificity value is illustrated below:

$$\textbf{Specificity} = \frac{3}{3+0} = 1.00 \tag{4.8}$$

The specificity value of 1.00 means that the model correctly identified the 100% of all the actual

negative cases.

4.8.3 Precision

Precision measures the proportion of correctly predicted positive cases (Diabetic) out of the cases predicted as positive(Diabetic). From Table 4.11, the precision value is shown below:

$$\textbf{Precision} = \frac{18}{18+0} = 1.00 \tag{4.9}$$

The Precision value of 1.00 indicates that all the cases the model predicted as present,were all actually diabetic cases with a 100%.

4.8.4 F1 Score

The F1 Score of the reduced model is calculated below from Table 4.11

$$\textbf{F1-score} = \frac{210.7826}{1+0.7826} = 0.8780 \tag{4.10}$$

With an F1 Score of 0.8780 , it indicates a very good balance between precision and recall.The value suggests that the model performs well in correctly identifying both positive (present) and negative (Absent).

4.8.5 Model Accuracy

The Model accuracy for the reduced model is explained below from Table 4.11. The Model accuracy of the reduced is less as compared to the full model.

$$\textbf{Model Accuracy} = \frac{18+3}{18+0+5+3} = 0.80769 \tag{4.11}$$

The accuracy value of 0.80769 means that the model predicted the outcome (whether present or absent of diabetes) for approximately 80.8% of the cases.

Chapter 5

Discussion and Conclusion

5.1 Introduction

This final chapter summarizes the output of our study undertaken on the prevalence of diabetes in virginia. The chapter is subdivided into three sections. Section one has a summary of the study, Section two presents the conclusion and final section presents the recommendation made from the analysis of the available data.

5.2 Summary

Diabetes is one of the most chronic non-communicable diseases with increasing trend worldwide. The purpose of this study was to determine the prevalence of diabetes and it risk factors in virginia. The analysis in this study was based on the data from Sutham Jirapanakorn, a medical Doctor in Virginia for Public Health by Imperial College London course, on February,2021.

The number of respondents were 130 in the study. A multiple logistic regression was for the study. The dependent variable was the diabetes status of the patient. Thus, whether the patient is diabetic or not; and also the dependent variable was made continuous (with the multiple regression model).

The independent variables of the study were age, BMI, gender, HDL and smoking status of the participants. The statistical software, R (version ,4.3.1) was used for all estimations in the research study. The correlation matrix was used to find the association between the variables. There was no issue of multicollinearity.

5.3 Conclusion

Based on the analysis, it is observed that age, Low Density Lipoprotein (LDL), and High Density Lipoprotein (HDL) are the key factors influencing an individual's likelihood of developing diabetes. From the study's analysis and findings, the following conclusions can be drawn: In the full model, the probability of a person having diabetes, as per the logistic regression model, is influenced by an increase of 0.072319 in age, 0.021439 in LDL, and a decrease of 0.038824 in HDL. In the reduced model, this probability is affected by a rise of 0.071648 in age and 0.020203 in LDL.

Since the full model has an accuracy of 84.62% which is higher than the reduced model's accuracy of 80.8%, the full model is more effective for predicting diabetes in Virginia.

5.4 Recommendation

1. It is therefore recommended, people with low density lipoprotein (LDL) are advised to reduce smoking,reduce eating a lot of foods which have high saturated fats, and people with high density lipoprotein (HDL) should engage in physical activities,limit trans fats and medication to manage cholesterol levels.

2. Furthermore, as people age, it is advised that healthy improved diet is adopted , increased exercise, quitting smoking and maintaining a healthy weight.

REFERENCES

Baz, B., Riveline, J.-P. & Gautier, J.-F. (2016), 'Endocrinology of pregnancy: gestational diabetes mellitus: definition, aetiological and clinical aspects', *European journal of endocrinology* **174**(2), R43–R51.

Chan, J. C., Lim, L.-L., Wareham, N. J., Shaw, J. E., Orchard, T. J., Zhang, P., Lau, E. S., Eliasson, B., Kong, A. P., Ezzati, M. et al. (2020), 'The lancet commission on diabetes: using data to transform diabetes care and patient lives', *The Lancet* **396**(10267), 2019–2082.

Foster, N. C., Beck, R. W., Miller, K. M., Clements, M. A., Rickels, M. R., DiMeglio, L. A., Maahs, D. M., Tamborlane, W. V., Bergenstal, R., Smith, E. et al. (2019), 'State of type 1 diabetes management and outcomes from the t1d exchange in 2016–2018', *Diabetes technology & therapeutics* **21**(2), 66–72.

Harjutsalo, V., Sjöberg, L. & Tuomilehto, J. (2008), 'Time trends in the incidence of type 1 diabetes in finnish children: a cohort study', *The Lancet* **371**(9626), 1777–1782.

Karamanos, B., Thanopoulou, A., Angelico, F., Assaad-Khalil, S., Barbato, A., Del Ben, M., Dimitrijevic-Sreckovic, V., Djordjevic, P., Gallotti, C., Katsilambros, N. et al. (2002), 'Nutritional habits in the mediterranean basin. the macronutrient composition of diet and its relation with the traditional mediterranean diet. multi-centre study of the mediterranean group for the study of diabetes (mgsd)', *European Journal of Clinical Nutrition* **56**(10), 983–991.

Khan, R., Siddiqui, A. A., Alshammary, F., Shaikh, S., Amin, J. & Rathore, H. A. (2021), Diabetes in the arab world, *in* 'Handbook of Healthcare in the Arab World', Springer, pp. 1029–1051.

Lee, K. W., Ching, S. M., Hoo, F. K., Ramachandran, V., Chong, S. C., Tusimin, M., Nordin, N. M., Devaraj, N. K., Cheong, A. T. & Chia, Y. C. (2020), 'Neonatal outcomes and its association among gestational diabetes mellitus with and without depression, anxiety and stress symptoms in malaysia: A cross-sectional study', *Midwifery* **81**, 102586.

Leon, N., Namadingo, H., Bobrow, K., Cooper, S., Crampin, A., Pauly, B., Levitt, N. & Farmer, A. (2021), 'Intervention development of a brief messaging intervention for a randomised controlled trial to improve diabetes treatment adherence in sub-saharan africa', *BMC Public Health* **21**(1), 1–14.

Maahs, D. M., West, N. A., Lawrence, J. M. & Mayer-Davis, E. J. (2010), 'Epidemiology of type 1 diabetes', *Endocrinology and Metabolism Clinics* **39**(3), 481–497.

Munir, S., Rodriguez, B. S. Q., Waseem, M. & Haddad, L. M. (2022), Addison disease (nursing), *in* 'StatPearls [Internet]', StatPearls Publishing.

Pastors, J. G., Warshaw, H., Daly, A., Franz, M. & Kulkarni, K. (2002), 'The evidence for the effectiveness of medical nutrition therapy in diabetes management', *Diabetes care* **25**(3), 608.

Peiris, H., Park, S., Louis, S., Gu, X., Lam, J. Y., Asplund, O., Ippolito, G. C., Bottino, R., Groop, L., Tucker, H. et al. (2018), 'Discovering human diabetes-risk gene function with genetics and physiological assays', *Nature communications* **9**(1), 3855.

Redondo, M. J., Oram, R. A. & Steck, A. K. (2017), 'Genetic risk scores for type 1 diabetes prediction and diagnosis', *Current diabetes reports* **17**, 1–10.

Shin, D. & Song, W. O. (2015), 'Prepregnancy body mass index is an independent risk factor for gestational hypertension, gestational diabetes, preterm labor, and small-and large-for-gestational-age infants', *The Journal of Maternal-Fetal & Neonatal Medicine* **28**(14), 1679–1686.

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C., Mbanya, J. C. et al. (2022), 'Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045', *Diabetes research and clinical practice* **183**, 109119.

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B., Stein, C., Basit, A., Chan, J., Mbanya, J., Pavkov, M., Ramachandaran, A., Wild, S., James, S., Herman, W., Zhang, P., Bommer, C., Kuo, S., Boyko, E. & Magliano, D. (2021), 'Idf diabetes atlas:

Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045', *Diabetes Research and Clinical Practice* .

Wang, Z.-L., Zou, L., Lu, Z.-W., Xie, X.-Q., Jia, Z.-Z., Pan, C.-J., Zhang, G.-X. & Ge, X.-M. (2017), 'Abnormal spontaneous brain activity in type 2 diabetic retinopathy revealed by amplitude of low-frequency fluctuations: a resting-state fmri study', *Clinical Radiology* **72**(4), 340–e1.

Zheng, Y., Ley, S. H. & Hu, F. B. (2018), 'Global aetiology and epidemiology of type 2 diabetes mellitus and its complications', *Nature reviews endocrinology* **14**(2), 88–98.