Diabetes and Related Factors

Shekoofeh Ansari- 400305479

ansari.shekoofeh@stud.hs-fresenius.de

Contents

1 Abstract **3**

2 Introduction **4**

2.1 Data Science 4

2.2 Diabetes 5

2.2.1 Diabetes' Type 5

2.2.2 Type 1 Diabetes 5

2.2.3 Type 2 Diabetes 7

2.2.4 Gestational Diabetes 10

3 Data Processing **12**

3.1 Missing Values and Zero Value 13

4 Introduction-Research Question **14**

5 Structure **14**

5.1 Number of Pragnancies 15

5.2 Blood Sugar Regulator 16

5.2.1 Glucose 16

5.2.2 Insulin 17

5.3 Skin Thickness 18

5.4 Body Mass Index 19

5.5 Age 20

5.6 Diabetes Pedigree Function 21

6 Regresion **21**

6.1 Correlation Matrix 22

6.2 Logistic Regression 26

6.3 Cross Validation 28

7 Statistics **29**

7.1 Adding a Confounder 29

7.2 Plotting Histograms of Numeric Values 29

7.3 Decision Tree 33

8 Conclusion **36**

9 Challenges and Improvement Area **36**

Refrences **38**

Rendered at 24 July, 2023

Wordcount: 7424

1 Abstract

The data used in this project was obtained from the Pima Indians Diabetes Database, which was made available by the National Institute of Diabetes and Digestive and Kidney Diseases. The selection of instances for this study was subject to certain constraints. Specifically, all the patients included in the dataset are females aged 21 and above, and they belong to the Pima Indian heritage, which is a subgroup of Native Americans.

To analyze the data and build our classification model, we utilized RStudio, a popular environment for data science, along with various related packages. We began by collecting the data and then imported it into our project, reading it from a CSV file. We performed necessary data manipulations to prepare it for further analysis. Additionally, we used Knit to convert our data into a suitable format for inputting into our classification model.

Throughout the project, we employed several data visualization packages to generate informative visual representations of the data. These visualizations aided in gaining insights and understanding patterns in the dataset.

For the classification modeling, we imported the logistic regression algorithm and utilized cross-validation techniques to develop our classification model. This algorithm played a crucial role in building the model and making predictions based on the input features.

Finally, based on our analysis and modeling results, we drew conclusions regarding the relationship between diabetes and various factors under investigation.

2 Introduction

The modern world is rapidly transforming through the ever-evolving landscape of data science, a powerful interdisciplinary field that leverages data analysis, statistics, and machine learning to extract valuable insights and inform decision-making processes. Data science has emerged as a pivotal force across diverse industries, revolutionizing the way information is harnessed to address complex challenges and unlock hidden potentials(Provost 2013) .

Simultaneously, health conditions such as diabetes continue to pose significant public health concerns, affecting millions of lives worldwide. Diabetes, a chronic metabolic disorder characterized by abnormal blood glucose levels, presents an intricate interplay of genetic, environmental, and lifestyle factors. As the prevalence of diabetes escalates, understanding its complexities and discovering innovative solutions become imperative to mitigate its impact on global health (Babar 2022).

This paper seeks to embark on a dual exploration, beginning with a concise overview of data science and its fundamental principles. We will elucidate the key components of data science, including data collection, analysis, and interpretation, and its broad applications across diverse domains. Emphasizing the relevance of data-driven insights in healthcare, we aim to highlight how data science has catalyzed advancements in understanding and managing complex medical conditions like diabetes.

Following this, we delve into the domain of diabetes, providing an in-depth examination of its pathophysiology, risk factors, and prevalence. Our focus will encompass a detailed classification of diabetes into its various types, with particular attention to Type 1, Type 2 diabetes and Gestational Diabetes. Understanding the distinctions between these types is vital to comprehend the diverse challenges and interventions required for their management.

The seamless integration of data science and diabetes research has the potential to revolutionize diabetes care. By harnessing the power of data analytics, machine learning, and predictive modeling, researchers and healthcare practitioners can identify patterns, uncover hidden associations, and develop personalized treatment strategies for individuals living with diabetes.

The overarching goal of this paper is to establish a foundational understanding of both data science and diabetes, and subsequently explore how the symbiotic relationship between these fields can drive innovation in diabetes management. As we progress through this study, we hope to emphasize the significant role of data-driven methodologies in transforming the future of healthcare and contributing to improved outcomes for individuals affected by diabetes.

2.1 Data Science

Data science is an intriguing academic field that uses statistics, logical computing, logical strategies, forms, calculations, and frameworks to extricate or extrapolate information and experiences from noisy, organized, and unstructured data Data science moreover coordinates space information from the fundamental application space (e.g., characteristic sciences, data innovation, and medicine)(Provost 2013). Information science is multifaceted and can be depicted as a science, an inquiry about paradigm, an inquiry about strategy, a teaching, a workflow, and a profession. Data science could be a "concept to bind together measurements, data investigation, informatics, and their related strategies" in arrange to "get it and examine real phenomena" with data(T. Hastie and Friedman 2001). It employments strategies and hypotheses are drawn from numerous areas inside the setting of science, insights, computer science, data science, and space knowledge However, information science is diverse from computer science and data science(Donoho. 2017). The term "Science" denotes the acquisition of knowledge through organized investigation. One definition characterizes it as a methodical endeavor that constructs and arranges knowledge by providing explanations and predictions about the universe that can be tested. Data Science can be seen as emphasizing data and, consequently, Statistics, which involves the systematic examination of data organization, characteristics, analysis, and their role in drawing conclusions, including the level of confidence we have in those conclusions. So, why do we require a new term when Statistics has existed for centuries? The mere existence of massive amounts of data is not a sufficient reason to introduce a new term. (Provost 2013).

2.2 Diabetes

Diabetes is a chronic health condition that affects the body's ability to convert food into energy. When you consume food, your body breaks it down into sugar (glucose), which is then released into the bloodstream. In response to elevated blood sugar levels, the pancreas releases insulin. Insulin plays a crucial role in allowing the entry of blood sugar into the cells of your body, where it is used as energy. (Disease Control and Prevention 2022).

In the case of diabetes, either the body does not produce enough insulin or it is unable to effectively utilize the insulin it produces. Insufficient insulin or insulin resistance causes an accumulation of high levels of sugar (glucose) in the bloodstream. This prolonged elevation of blood sugar can lead to significant health problems over time, including heart disease, vision loss, and kidney disease. Insulin plays a vital role as a hormone in regulating blood glucose levels. Persistent high blood sugar, also known as hyperglycemia, is a common consequence of uncontrolled diabetes and can result in severe damage to various body systems, particularly the nerves and blood vessels.(Disease Control and Prevention 2022).

2.2.1 Diabetes' Type

There are three main types of diabetes: type 1, type 2, and gestational diabetes (diabetes while pregnant).

2.2.2 Type 1 Diabetes

Type 1 diabetes is believed to result from an immune system response where the body mistakenly attacks itself. This reaction impairs the production of insulin in the body. Approximately 5-10% of individuals with diabetes are affected by type 1. Symptoms of type 1 diabetes often manifest quickly, and it is typically diagnosed in children, adolescents, and young adults. If you have type 1 diabetes, daily insulin injections are necessary for survival. Currently, there is no known method to predict the development of type 1 diabetes.(Disease Control and Prevention 2022) Type 1 diabetes, also referred to as insulin-dependent, juvenile, or childhood-onset diabetes, is characterized by the inability of the body to produce insulin and necessitates daily administration of insulin. As of 2017, there were 9 million individuals living with type 1 diabetes, with the majority residing in high-income countries.The precise origins of type 1 diabetes and preventive measures are still uncertain. Indications of type 1 diabetes encompass frequent urination (polyuria), heightened thirst (polydipsia), continual hunger, weight loss, vision changes, and fatigue. These symptoms may manifest abruptly and without prior indication. (Health Organization 2022).

2.2.2.1 Symptoms Type 1 diabetes symptoms can appear suddenly and may include (Roglic et al. 2016) :

- Begin more thirsty than normal
- Urinating more than usual

- Enuresis In for children who never happened for them
- Starving
- Without any effort they lose weight
- Instability on moods
- Exhaustion
- Not having clear vision

2.2.2.2 Risk Factors Factors affecting the increase in the probability of having type 1 diabetes include:(Watson 2022):

- **Family history:** The presence of a parent or sibling with type 1 diabetes slightly elevates the risk of developing the condition.
- **Genetics:** Certain genes contribute to an increased susceptibility to type 1 diabetes.
- **Geography:** The prevalence of type 1 diabetes tends to be higher in regions farther away from the equator.
- **Age:** Although type 1 diabetes can manifest at any age, there are two distinct age peaks. The first peak occurs in children aged 4 to 7 years, while the second peak is observed in children aged 10 to 14 years.

2.2.2.3 Causes The exact cause of type 1 diabetes is not known. Usually, the body's immune system, which has the task of destroying harmful bacteria and viruses, destroys the insulin-producing cells(Islets of Langerhans)in the pancreas. Other possible causes include (Health Organization 2022):

- Hereditary factors
- Encounters with viruses and various elements in the environment

2.2.2.4 Complications Our research question is based on the need to explore a strong relationship between diabetes as explored factors and intelligible factors in arrange to foresee the likelihood of patients creating diabetes based on the analyzed information. Complications arising from diabetes have the potential to cause disabilities or pose a significant risk to one's life. The adverse effects of diabetes-related complications can result in impairments that affect daily functioning and overall well-being. In severe cases, these complications can even become life-threatening. It is essential to address and manage diabetes effectively to minimize the occurrence of complications and safeguard both quality of life and overall health. (clinic 2022).

Heart and cardiovascular disease: Diabetes raises the risk of various cardiovascular issues, including angina (chest pain), heart attacks, strokes, atherosclerosis (narrowing of arteries), and high blood pressure.

Nerve damage (neuropathy): Elevated blood sugar levels can cause damage to the small blood vessels that supply nerves, particularly in the legs. This can result in symptoms such as tingling, numbness, burning sensations, or pain. Typically, these sensations start in the toes or fingers and gradually progress upwards. If blood sugar is poorly controlled, it can eventually lead to loss of sensation in the affected limbs. Nerve damage can also affect the digestive system, leading to issues like nausea, vomiting, diarrhea, or constipation. Additionally, men may experience erectile dysfunction.

Kidney damage (nephropathy): Diabetes can harm the delicate blood vessels in the kidneys, impairing their function. Severe damage can lead to kidney failure or end-stage kidney disease, which may necessitate dialysis or a kidney transplant.

Eye damage: Diabetes can cause damage to the blood vessels in the retina, leading to a condition known as diabetic retinopathy. This can result in vision impairment or blindness. Diabetes also increases the risk of other serious eye conditions like cataracts and glaucoma.

Foot complications: Nerve damage in the feet, combined with poor blood circulation, increases the risk of foot problems. Without proper care, cuts and blisters can develop into severe infections, sometimes necessitating toe, foot, or leg amputations.

Skin and oral conditions: Diabetes can make individuals more susceptible to skin and oral infections, including bacterial and fungal infections. Gum disease and dry mouth are also more common.

Pregnancy complications: levated blood sugar levels during pregnancy can present dangers for both the mother and the baby. Inadequately managed diabetes heightens the likelihood of miscarriage, stillbirth, and congenital abnormalities. It can also result in maternal complications such as diabetic ketoacidosis, diabetic retinopathy, gestational hypertension, and preeclampsia

2.2.3 Type 2 Diabetes

Type 2 diabetes occurs when your body becomes resistant to insulin and is unable to maintain normal blood sugar levels. It accounts for approximately 90-95% of all diabetes cases. It develops gradually over time and is typically diagnosed in adults, although there is a rising prevalence among children, teenagers, and young adults. Unlike type 1 diabetes, you may not experience noticeable symptoms, highlighting the importance of getting your blood sugar tested if you are at risk. Making healthy lifestyle changes, such as losing weight, adopting a nutritious diet, and staying physically active, can help prevent or delay the onset of type 2 diabetes. The symptoms may resemble those of type 1 diabetes but are often less pronounced. As a result, the condition may be diagnosed years after its onset, when complications have already emerged. While type 2 diabetes used to be primarily observed in adults, it is now increasingly affecting children as well. (Health Organization 2022).

2.2.3.1 Symptoms The signs and symptoms of type 2 diabetes are not clear and it does not develop quickly.The truth is that you have been living for years, but your body is suffering from type 2 diabetes and you have not noticed it. But if there are signs and symptoms, it includes the following. (Roglic et al. 2016):

- Excessive thirst
- Frequent urination
- Increased hunger
- Unintended weight loss
- Fatigue
- Blurred vision
- Slow-healing sores
- Frequent infections
- Numbness or tingling in the hands or feet
- Darkened skin patches, typically found in the armpits and neck.

2.2.3.2 Causes Type 2 diabetes is primarily caused by two interconnected issues:

Insulin resistance: Cells in the muscles, fat tissues, and liver become resistant to the action of insulin. As a result, these cells do not effectively respond to insulin, leading to insufficient uptake of sugar (glucose) from the bloodstream.

Insufficient insulin production: The pancreas, which is responsible for producing insulin, fails to generate enough of it to adequately regulate blood sugar levels.

The exact reasons behind the development of insulin resistance and inadequate insulin production are not fully understood. However, being overweight and leading a sedentary lifestyle are significant contributing factors to the onset of type 2 diabetes. (Health Organization 2022).

2.2.3.3 Risk Factors Factors that increase the probability of getting type 2 diabetes include: (Watson 2022):

- **Weight:** One of the most important risks is overweight or obese

- **Fat distribution:** Having a higher propensity to accumulate fat in the abdominal area, as opposed to the hips and thighs, signifies an increased level of vulnerability. If you are a man with a waist circumference exceeding 40 inches (101.6 centimeters) or a woman with a measurement exceeding 35 inches (88.9 centimeters), your chances of developing type 2 diabetes elevate
- **Inactivity:** Your risk of certain health issues increases as your level of physical activity decreases. Engaging in physical activity assists in managing weight, utilizing glucose for energy, and enhancing insulin sensitivity in your cells.
- **Family history:** Having a parent or sibling with type 2 diabetes raises the likelihood of developing the condition.
- **Race and ethnicity:** The reasons remain unclear, but individuals belonging to specific races and ethnicities, such as Black, Hispanic, Native American, Asian, and Pacific Islander populations, have a higher susceptibility to developing type 2 diabetes compared to white individuals.
- **Blood lipid level:** Elevated risk is linked to decreased levels of high-density lipoprotein (HDL) cholesterol, commonly known as “good” cholesterol, as well as elevated levels of triglycerides.
- **Age:** As you age, particularly after reaching 35 years old, the risk of developing type 2 diabetes progressively rises..
- **Prediabetes:** Prediabetes is a state where your blood sugar level is higher than normal but not at the threshold for diabetes diagnosis. If left untreated, prediabetes frequently advances into type 2 diabetes.
- **Pregnancy-related risks:** If you experienced gestational diabetes during pregnancy or gave birth to a baby weighing over 9 pounds (4 kilograms), your risk of developing type 2 diabetes is heightened..
- **Polycystic ovary syndrome:**The presence of polycystic ovary syndrome (PCOS), a prevalent condition characterized by irregular menstrual periods, excessive hair growth, and obesity, amplifies the risk of developing diabetes.
- **Areas of darkened skin:** Typically occurring in the armpits and neck, this condition frequently signifies insulin resistance.

2.2.3.4 Complications Type 2 diabetes has wide-ranging effects on various major organs, including the heart, blood vessels, nerves, eyes, kidneys, and more. Additionally, many risk factors associated with diabetes also increase the likelihood of developing other serious chronic diseases. By effectively managing diabetes and controlling blood sugar levels, the risk of these complications or coexisting conditions (comorbidities) can be reduced.(clinic 2022).

The potential complications of diabetes and frequently occurring comorbidities include:

Heart and blood vessel disease Diabetes is linked to an elevated risk of heart disease, stroke, high blood pressure, and the narrowing of blood vessels (atherosclerosis).

Nerve damage (neuropathy) in limbs Prolonged high blood sugar levels can lead to nerve damage, resulting in symptoms such as tingling, numbness, burning, pain, or eventual loss of sensation. This typically starts at the extremities and progresses upward.

Other nerve damage Nerve damage associated with diabetes can contribute to irregular heart rhythms and digestive issues like nausea, vomiting, diarrhea, or constipation. In men, it may cause erectile dysfunction.

Kidney disease Diabetes can lead to chronic kidney disease or irreversible end-stage kidney disease, which may necessitate dialysis or a kidney transplant.

Eye damage Diabetes increases the risk of serious eye diseases, such as cataracts and glaucoma. It can also damage the blood vessels in the retina, potentially resulting in blindness.

Skin conditions People with diabetes are more susceptible to skin problems, including bacterial and fungal infections. Slow healing.Untreated cuts and blisters can develop into severe infections, healing poorly and potentially requiring toe, foot, or leg amputations.

Hearing impairment Hearing problems are more common in individuals with diabetes.

Sleep apnea Obstructive sleep apnea is prevalent among people with type 2 diabetes. Obesity frequently plays a role in the development of both conditions, although it remains uncertain whether addressing sleep apnea leads to enhanced blood sugar regulation.

Dementia Type 2 diabetes appears to increase the risk of Alzheimer's disease and other forms of dementia. Poor control of blood sugar levels is associated with a more rapid decline in memory and other cognitive abilities.

2.2.4 Gestational Diabetes

Gestational diabetes is a condition that occurs during pregnancy in women who have not previously had diabetes. If you are diagnosed with gestational diabetes, there is an increased risk of health issues for your baby. However, gestational diabetes typically disappears after giving birth. Nonetheless, it does raise the risk of developing type 2 diabetes later in life. Additionally, there is a higher likelihood that your child may have a higher birth weight and may be at an increased risk of developing type 2 diabetes in the future.(Health Organization 2022). Gestational diabetes refers to a condition where blood glucose levels during pregnancy are higher than normal but not as high as in diagnosed diabetes. It occurs in women during pregnancy. Women diagnosed with gestational diabetes are at a higher risk of experiencing complications throughout their pregnancy and during the delivery process. Both the mothers and their children are also at an elevated risk of developing type 2 diabetes later in life. Gestational diabetes is typically diagnosed through prenatal screening rather than based on specific symptoms. (Sarwar N 2010).

2.2.4.1 Causes The exact reasons why some women develop gestational diabetes while others do not are still unknown to researchers. However, pre-pregnancy overweight or obesity often contributes to its occurrence.

Typically, a balance of different hormones regulates blood sugar levels. However, during pregnancy, hormone levels undergo changes that make it more challenging for the body to process blood sugar effectively. As a result, blood sugar levels tend to increase. (Health Organization 2022).

2.2.4.2 Complications If gestational diabetes is not adequately controlled, it can result in elevated blood sugar levels. High blood sugar poses potential risks for both the mother and the baby, including an increased likelihood of requiring a cesarean section (C-section) for delivery.(clinic 2022).

Complications that may affect your baby

If you have gestational diabetes, your baby may have an elevated risk of the following complications:

Excessive birth weight Elevated blood sugar levels beyond the normal range can lead to excessive growth of the baby. This can result in the birth of a very large baby, weighing 9 pounds or more. Babies of this size are more prone to becoming stuck in the birth canal, experiencing birth injuries, or requiring a cesarean section (C-section) for delivery. It is crucial to manage blood sugar levels during pregnancy to minimize the risk of these complications and ensure a safer delivery for both the mother and the baby.

Early (preterm) birth High blood sugar levels can elevate the risk of preterm labor, resulting in the baby being delivered before the anticipated due date. Alternatively, early delivery may be recommended if the baby is excessively large due to the impact of high blood sugar.

Serious breathing difficulties Prematurely born infants can encounter respiratory distress syndrome, which is characterized by difficulties in breathing.

Low blood sugar (hypoglycemia) In certain cases, newborns may experience low blood sugar (hypoglycemia) shortly after birth. Severe instances of hypoglycemia can lead to seizures in the baby. Timely feedings and, occasionally, the administration of intravenous glucose solution can restore the baby's blood sugar levels to normal.

Obesity and type 2 diabetes later in life Babies are at an increased risk of developing obesity and type 2 diabetes in their later years.

Stillbirth If gestational diabetes is left untreated, it can lead to the unfortunate outcome of fetal demise, which refers to the death of the baby either before or shortly after birth. Proper management and monitoring of gestational diabetes are crucial to reduce such risks and ensure the well-being of both the mother and the baby.

Complications that may affect you Gestational diabetes can increase the likelihood of experiencing the following risks:(Watson 2022):

High blood pressure and preeclampsia Gestational diabetes raises the risk of developing high blood pressure and preeclampsia, a serious pregnancy complication characterized by high blood pressure and other symptoms that can endanger both the mother and the baby's well-being.

Having a surgical delivery (C-section) The chances of having a C-section are higher for women with gestational diabetes.

Subsequent diabetes If you have had gestational diabetes, there is an increased likelihood of developing it again in future pregnancies. Additionally, you have a higher risk of developing type 2 diabetes as you age.

2.2.4.3 Risk factors Factors that increase the risk of gestational diabetes comprise:(Sarwar N 2010) :

- Excessive weight or obesity
- Insufficient physical activity
- Preexisting prediabetes
- Prior history of gestational diabetes in a previous pregnancy
- Presence of polycystic ovary syndrome
- Family history of diabetes
- Previous delivery of a baby weighing over 9 pounds (4.1 kilograms)
- Belonging to certain racial or ethnic groups, such as Black, Hispanic, American Indian, and Asian American individuals.

3 Data Processing

Data processing is the systematic and organized manipulation of raw data to derive meaningful information, insights, or knowledge. Data processing is a crucial component of data management and plays a central role in various fields, including business, research, healthcare, finance, and more.

Frist of all, we use Head function. head is a function which returns the first 6 observations of the dataset.

```r
df <- read.csv("C:/Users/ansar/Desktop/final/diabetes.csv")
head(df)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

It has come to the point where we need to extract patients who do not possess data in any of the factors.

```r
df <- read.csv("C:/Users/ansar/Desktop/final/diabetes.csv")
missing_data <- df[,setdiff(names(df), c('Outcome'))]
features_miss_num <- apply(missing_data, 2, function(x) sum(x <= 0))
features_miss <- names(missing_data)[ features_miss_num > 0]

rows_miss <- apply(missing_data, 1, function(x) sum(x <= 0) >= 1)
summary(rows_miss)
```

```
##    Mode   FALSE    TRUE
## logical     336     432
```

Here we are computing the minimum,1st quartile, median, mean,3rd quartile and the maximum for all numeric variables of a dataset at once using summary() function.

```r
summary(missing_data)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
```

3.1 Missing Values and Zero Value

Now we want to determine the missing values in our database and then zero value in our table with following codes:

```r
cat("Number of missing value:", sum(is.na(df)), "\n")
```

```
## Number of missing value: 0
```

```r
df <- read.csv("C:/Users/ansar/Desktop/final/diabetes.csv")

row_sub = apply(df, 1, function(row) all(row !=0 ))
```

4 Introduction-Research Question

Diabetes is one of the foremost expensive diseases for national health systems. The combination of the two variables i.e. the expansive rate of world populace that is affected by diabetes, and the high costs of diabetes for national well-being frameworks make the urgency for a political, and organizational reaction that's either arranged to advance personal and social well-being either able to diminish the fetched of diabetes medicines for patients. The alter in person behaviors, the understanding of the co-cause of diabetes, and indeed the investigation and improvement in the pharmacological divisions are able to advance a more profound understanding of diabetes and the methodology to advance way better well-being at the personal and social levels and a decrease in open costs for the care of diabetes patients (Alessandro Massaro 2022).

Our contribution tries to examine the nearness of the cause of diabetes attempting to get the presence of a significant affiliation between diabetes on one side and other person's well-being conditions on the other side understanding of the co-cause of diabetes, and indeed the inquiry about an improvement in the pharmacological divisions are able to advance a more profound understanding of diabetes and the methodology to advance way better wellbeing at the personal and social levels and a diminishment in open costs for the care of diabetes patients.

Our research question is based on the need to explore a strong relationship between diabetes as explored factors and intelligible factors in arrange to foresee the likelihood of patients creating diabetes based on the analyzed information.

5 Structure

First of all we should recognize which factors can have impact in our body, for this purpose, the following features have been provided to help us to predict whether a person is diabetic or not:

1. Number of times pregnant
2. Glucose (ml/DL): Plasma glucose concentration over 2 hours in an oral glucose tolerance test
3. Blood Pressure (mmHg)
4. Skin Thickness : Triceps skin fold thickness

5. Insulin : 2-Hour serum insulin (mu U/ml)
6. BMI : (Kg/ m*m)
7. Diabetes pedigree function: a function which scores likelihood of diabetes based on family history
8. Age: (years)

Number of Instances: 768

Number of Attributes: 8 plus class

For Each Attribute: (all numeric-valued)

Outcome: Class variable (0 if non-diabetic, 1 if diabetic)

Class Distribution: (class value 1 is interpreted as “tested positive for diabetes”)

we will discuss each factor in the following content

5.1 Number of Pragnancies

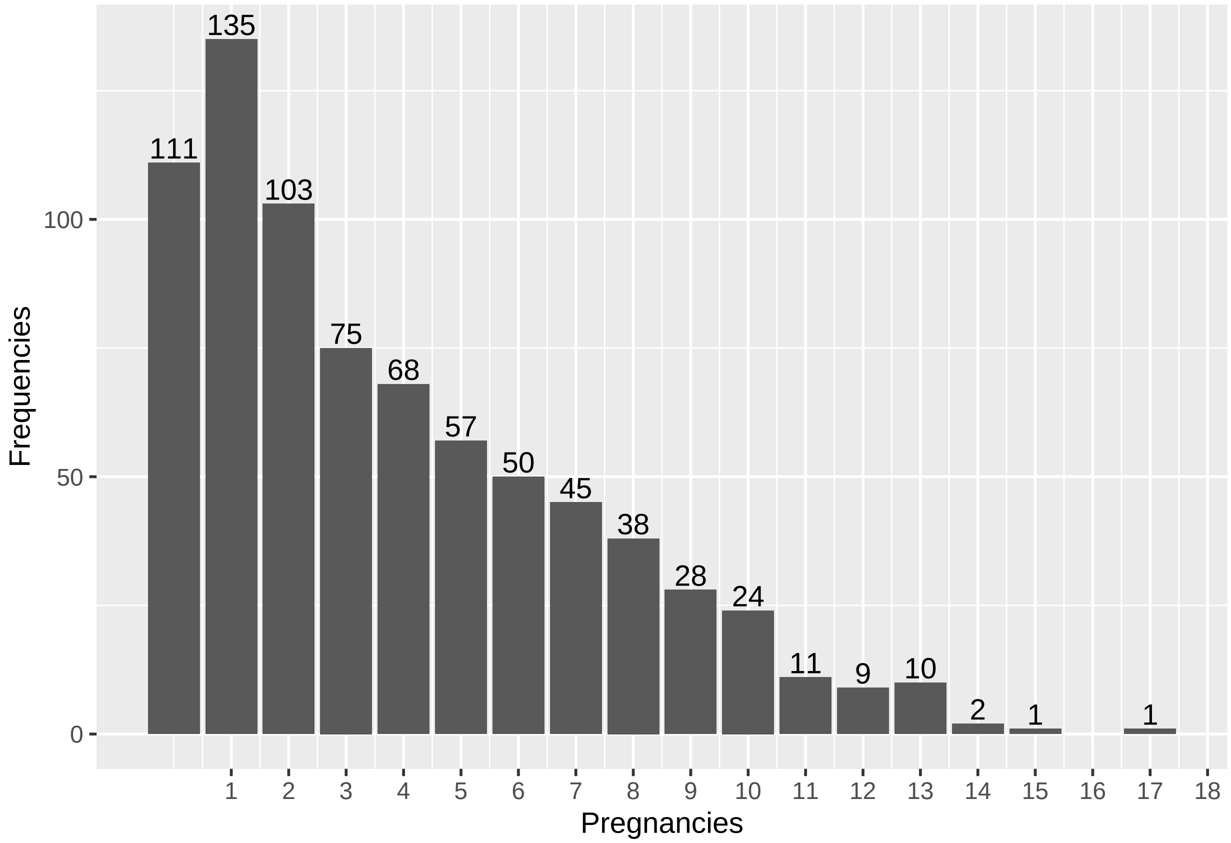
However, it is still vague whether the repeated presentation of these changes will impact the improvement of gestational diabetes mellitus (GDM). Within the present consideration, we examined the affiliation between the number of pregnancies and GDM.

```
ID <- 1:20

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df_grp <-  df %>% group_by(Pregnancies)  %>%
  summarise(Frequencies = length(Pregnancies),
            .groups = 'drop')
```

```r
library(ggplot2)
ggplot(df_grp, aes(x=Pregnancies, y=Frequencies)) +
  geom_bar(stat = "identity")  +
  scale_x_continuous("Pregnancies", labels = as.character(ID), breaks = ID)+
  geom_text(aes(label = Frequencies), vjust = -0.2)
```



During the process of creating this chart, we encountered difficulty in appropriately labeling the columns to ensure visibility of the numerical values associated with each column.However, with the aid of geom_text(aes(label =) formula successfully depicted in the chart.

5.2 Blood Sugar Regulator

The two hormones counterbalance each other to stabilize blood glucose

5.2.1 Glucose

The term “glucose” originates from the Greek word meaning “sweet.” It refers to a type of sugar that is obtained from the foods we consume, and our body utilizes it as a source of energy. When glucose circulates through our bloodstream and reaches our cells, it is referred to as blood glucose or blood sugar.

In the case of individuals with diabetes, their blood glucose levels are higher than normal. This can occur due to either insufficient insulin production or the cells' reduced responsiveness to insulin.

Prolonged elevation of blood glucose levels can lead to damage in various organs, including the kidneys, eyes, and other vital organs. The exact mechanisms by which high blood glucose causes these complications are complex and multifactorial.

Managing blood glucose levels is crucial for individuals with diabetes to prevent or minimize the risk of complications. This typically involves interventions such as medication, insulin therapy, dietary adjustments, regular physical activity, and monitoring blood glucose levels. By effectively controlling blood glucose levels, the risk of long-term organ damage can be reduced. (Watson 2022).

5.2.2 Insulin

Insulin plays a vital role as a hormone in the body. Its main function is to facilitate the movement of glucose from the bloodstream into the cells, where it is utilized for energy or stored for future use. Essentially, insulin helps convert food into energy and helps regulate blood sugar levels. However, in the case of diabetes, the body either does not produce enough insulin or cannot effectively utilize it.

To address this issue, healthcare providers may prescribe synthetic insulin that can be administered through injections (shots), injectable pens, or pumps. Insulin pumps are small, computerized devices, roughly the size of a small cell phone. They deliver controlled doses of insulin according to a pre-programmed schedule.

Insulin pumps offer several advantages. They provide a continuous supply of insulin, reducing the need for frequent injections. This can be particularly beneficial for individuals, including children, who may have difficulty remembering or administering insulin injections. Moreover, some people find insulin pumps more convenient since they remain attached to the body. Additionally, there are also inhalable forms of insulin available.

Overall, insulin plays a crucial role in managing blood sugar levels for individuals with diabetes, and various insulin delivery methods, including insulin pumps, are available to meet individual needs and preferences. Working closely with a healthcare provider is important to determine the most suitable insulin regimen for each person's specific requirements. (United States Renal Data System. National Institutes of Health, Digestive, and Kidney Diseases 2021).

5.3 Skin Thickness

Collagen content largely determines the thickness of the skin, and in individuals with insulin-dependent diabetes mellitus (IDDM), the skin thickness is increased. Studies have observed that thickened skin on the back of the hands and fingers is a common characteristic in diabetes mellitus, which is associated with a higher prevalence of diabetic retinal microvascular disease. (Huntley AC 1990).

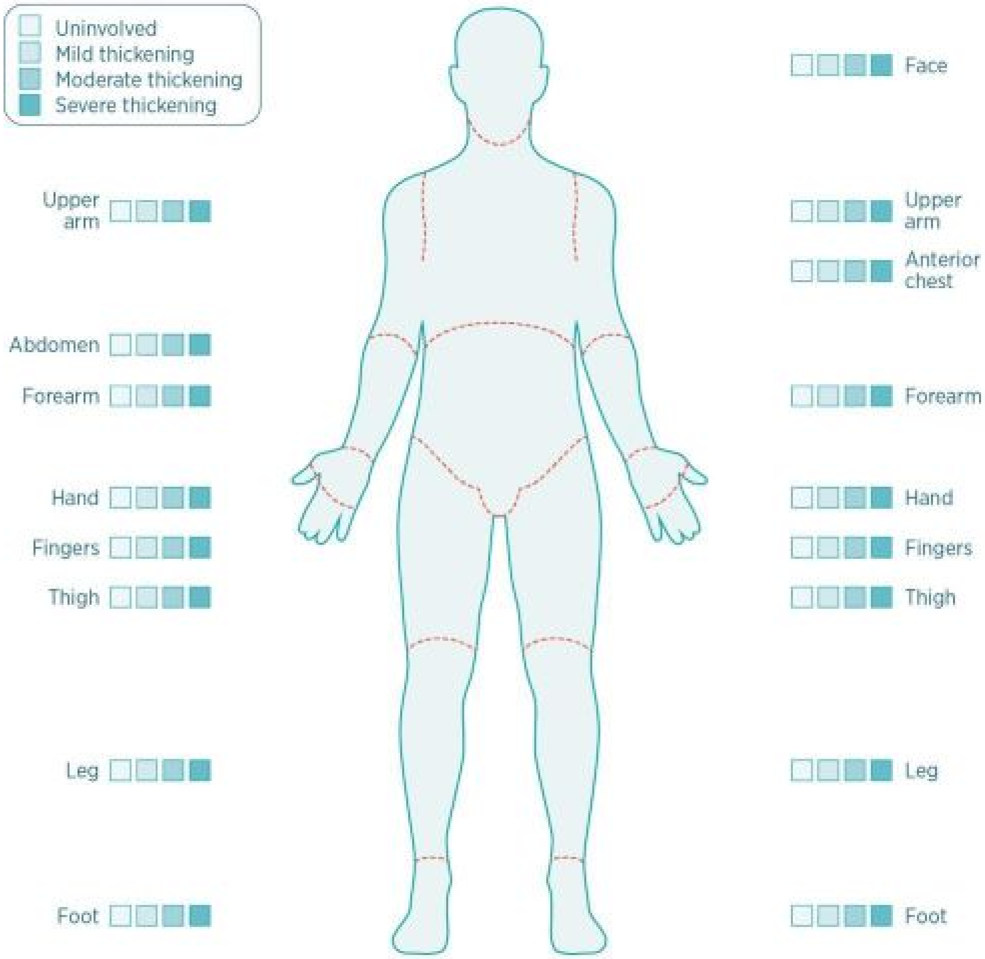


Figure 1: skin thickness (P J Clements 1 1993)

Regrettably, the image possessed a considerable size, and I endeavored to diminish its dimensions exclusively utilizing the Rstudio program, without resorting to additional tools solely for resizing.Fortunately, I managed to shrink the photo by resize formula which allowing me to effectively reduce the image size as desired.

5.4 Body Mass Index

Body Mass Index (BMI) is a measure of body weight relative to height. It is calculated by dividing a person's weight in kilograms (or pounds) by the square of their height in meters (or feet). A higher BMI value can indicate a higher level of body fatness. While BMI serves as a screening tool for weight categories that may pose health risks, it does not provide a definitive diagnosis of an individual's body fatness or overall health status.

$$BMI = \frac{Weight_{(kg)}}{Height^2_{(m)}}$$

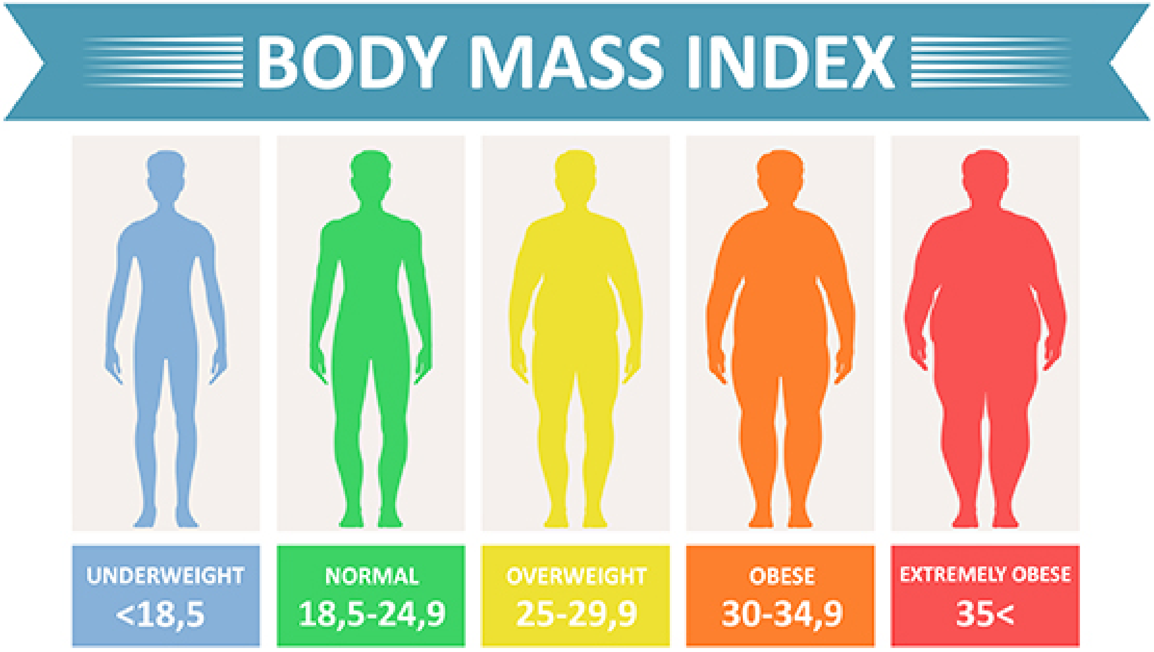
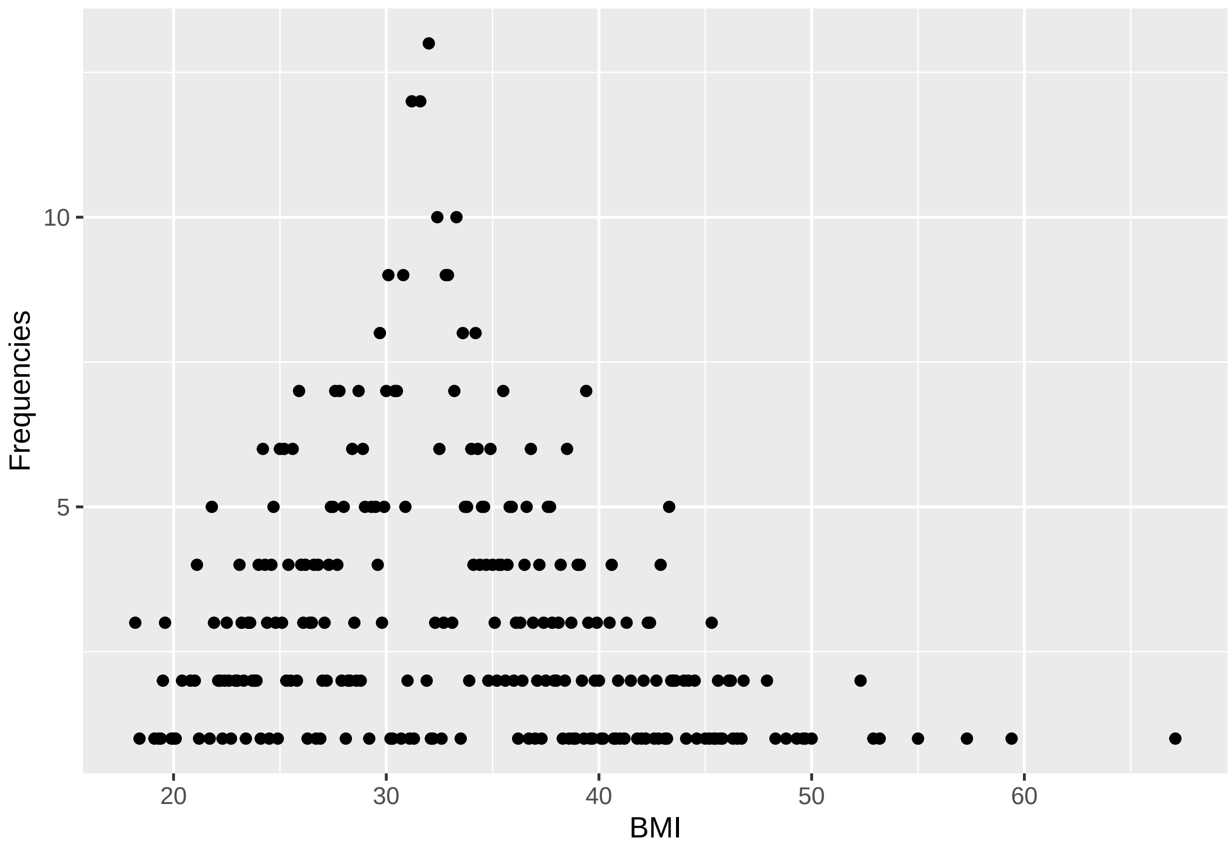


Figure 2: BMI (Lawlor, n.d.)

Our study offers epidemiological evidence indicating that obesity plays a significant role in the development of diabetes among the elderly population in rural areas.

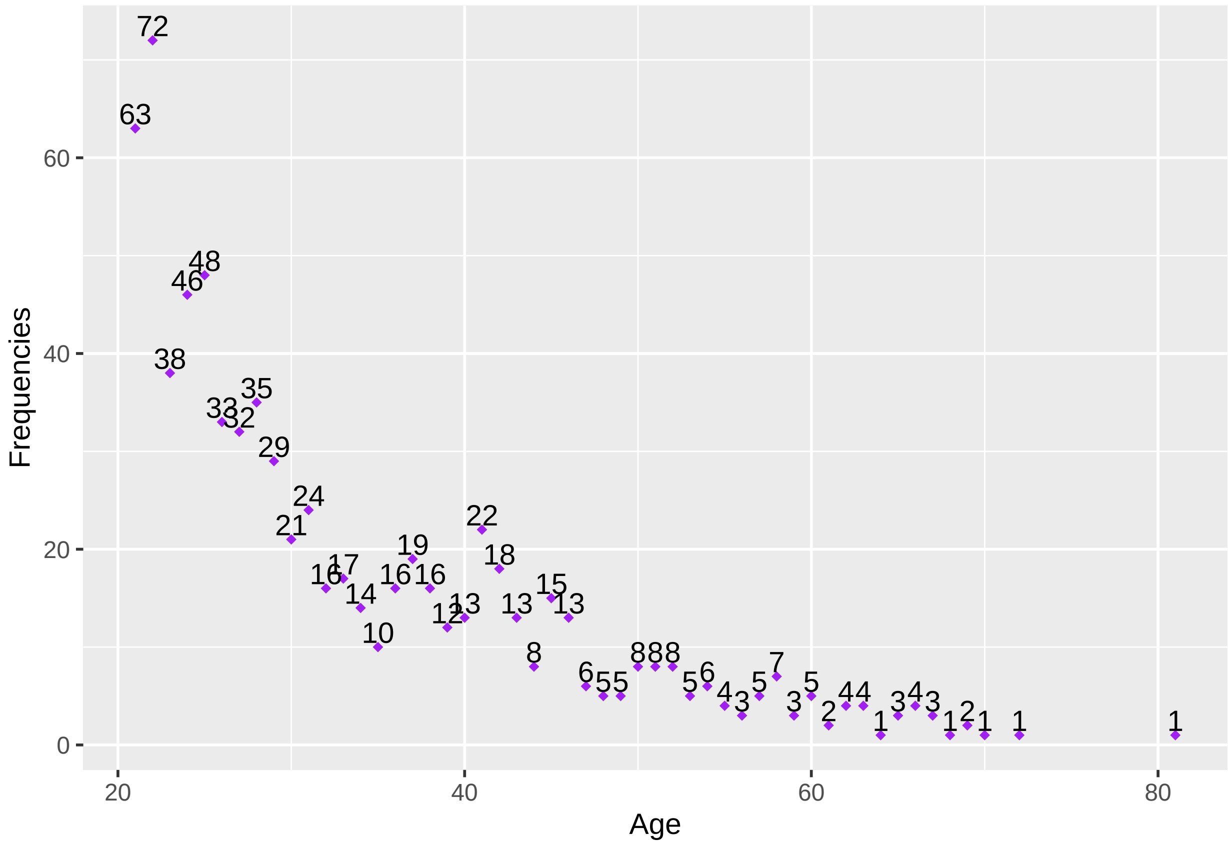
```
library(dplyr)
library(ggplot2)
df_bmi <-  missing_data %>% group_by(BMI)  %>%
  summarise(Frequencies = length(BMI),
            .groups = 'drop')
P15 <- filter(df_bmi, BMI >= 18 & BMI <= 70)
ggplot(data = P15 ,mapping = aes(x = BMI, y = Frequencies ))  +
  geom_point()
```



5.5 Age

The age at which diabetes is diagnosed is inversely related to the potential harm it can cause. In other words, the younger a person is when they are diagnosed with diabetes, the greater the potential harm they may experience.

```
library(ggplot2)
library(dplyr)
df_rt <-  df %>% group_by(Age)  %>%
  summarise(Frequencies = length(Age),
            .groups = 'drop')
df_rt %>%
    group_by(Age) %>%
    mutate(mean_age = mean(Age)) %>%
    ggplot(aes(x = Age, y = Frequencies)) +
    geom_point(color = "purple", shape = 18) +
  geom_text(aes(label = Frequencies), vjust = -0.2)
```



While composing the code for this graph, I aimed to modify the shape and color of the data points. With the aid of various websites, specifically referring to GGPLOT2 ESSENTIALS FOR GREAT DATA VISUALIZATION IN R , I was able to identify the most suitable shape for the points, resulting in an enhanced visualization.(Wickham 2016)

5.6 Diabetes Pedigree Function

The Diabetes Pedigree Function is a measure that assesses the probability of having diabetes based on family history. In simpler terms, it is a function that calculates the risk of type 2 diabetes by taking into account the individual's family background. A higher value of the function indicates a greater risk of developing type 2 diabetes.(Alessandro Massaro 2022).

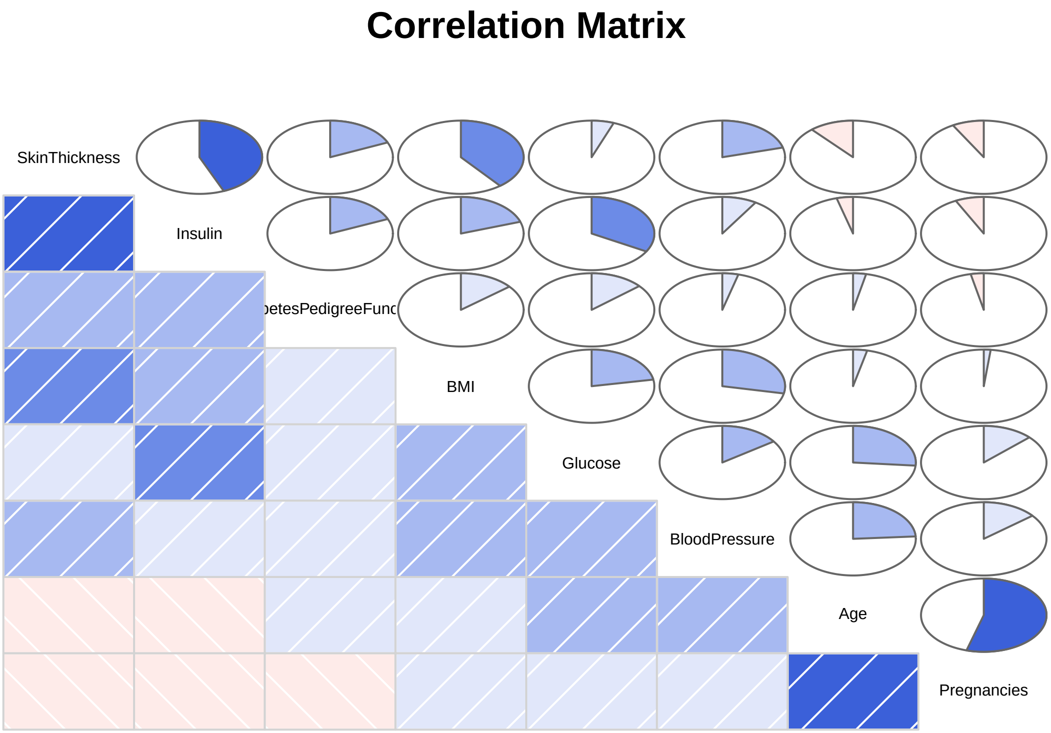
6 Regresion

Since all the variables exhibit reasonably broad distributions, they will be retained for the regression analysis. This indicates that the data for each variable spans a wide range and captures diverse values, making them suitable for inclusion in the analysis.

6.1 Correlation Matrix

Regarding the correlation between numeric variables, regression analysis can help determine the relationship and strength of association between these variables. By analyzing the correlation coefficients, which quantify the degree of linear relationship between variables, we can assess the extent to which the variables are correlated with each other.

```r
library(corrgram)
corrgram(missing_data, order=TRUE, lower.panel= panel.shade ,
  upper.panel= panel.pie, text.panel=panel.txt,
  main="Correlation Matrix")
```



Insulin is a hormone that controls blood glucose.Hyperglycemia, also known as elevated blood glucose or high blood sugar, is a common consequence of uncontrolled diabetes and can lead to severe damage to various systems in the body, particularly the nerves and blood vessels. In 2014, approximately 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was directly responsible for 1.5 million deaths, and 48% of those deaths occurred before the age of 70. An additional 460,000 deaths from kidney disease were attributed to diabetes, and approximately 20% of cardiovascular deaths were caused by elevated blood glucose. From 2000 to 2019, age-standardized mortality rates from diabetes increased by 3%, with a 13% increase in lower-middle-income countries.

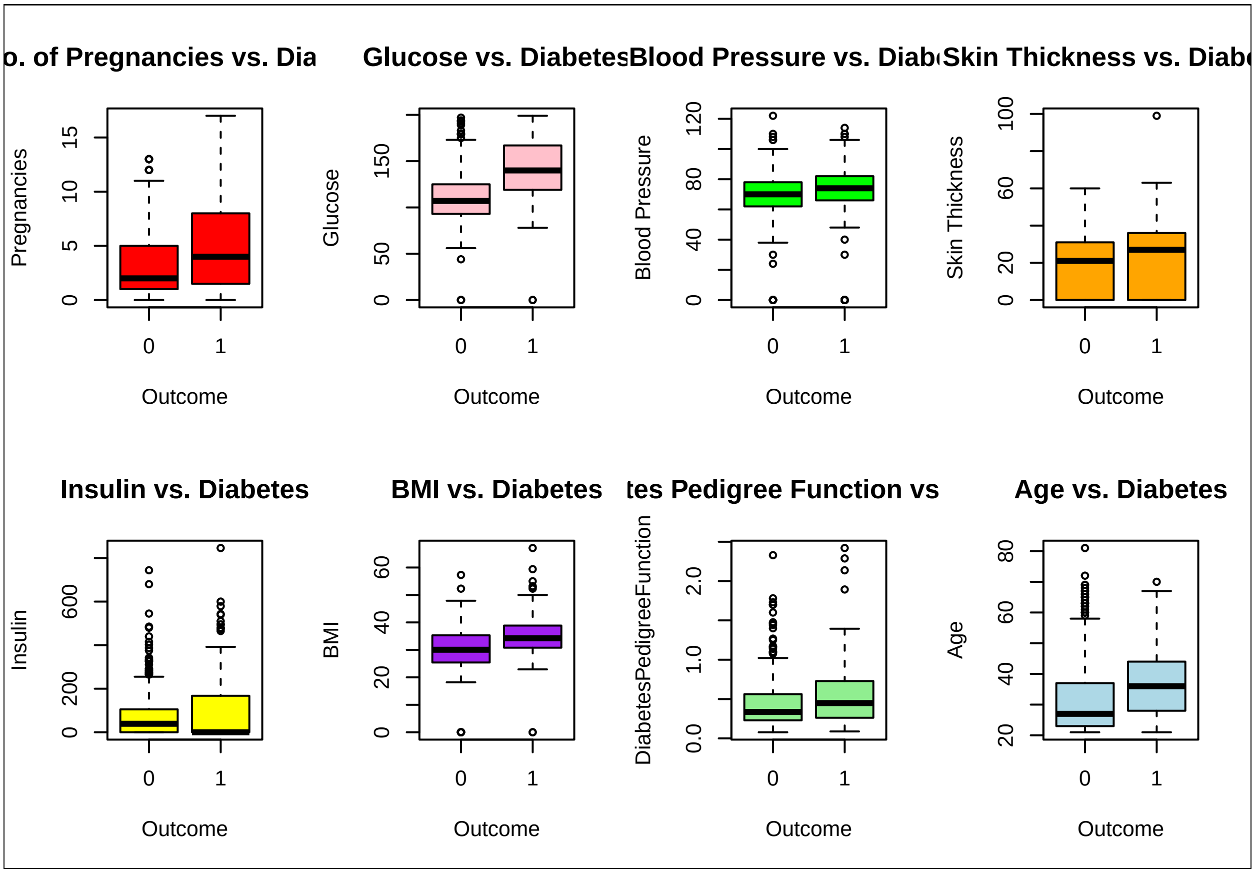
In contrast, the likelihood of dying from any of the four major non-communicable diseases (cardiovascular diseases, cancer, chronic respiratory diseases, or diabetes) between the ages of 30 and 70 decreased by 22% globally between 2000 and 2019.

```
df$BloodPressure <- NULL
df$SkinThickness <- NULL
train <- df[1:540,]
test <- df[541:768,]
model <-glm(Outcome ~.,family=binomial(link='logit'),data=train)
summary(model)
```

```
##
## Call:
## glm(formula = Outcome ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4366  -0.7741  -0.4312   0.8021   2.7310
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.3461752  0.8157916 -10.231  < 2e-16 ***
## Pregnancies               0.1246856  0.0373214   3.341 0.000835 ***
## Glucose                   0.0315778  0.0042497   7.431 1.08e-13 ***
## Insulin                  -0.0013400  0.0009441  -1.419 0.155781
## BMI                       0.0881521  0.0164090   5.372 7.78e-08 ***
## DiabetesPedigreeFunction  0.9642132  0.3430094   2.811 0.004938 **
## Age                       0.0018904  0.0107225   0.176 0.860053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 700.47  on 539  degrees of freedom
## Residual deviance: 526.56  on 533  degrees of freedom
## AIC: 540.56
##
## Number of Fisher Scoring iterations: 5
```

Based on the low p-values, the top three most relevant features in the analysis are “Glucose,” “BMI,” and “Number of times pregnant.” These variables have demonstrated significant statistical associations with diabetes. On the other hand, the variables “Insulin” and “Age” do not appear to be statistically significant in relation to diabetes in this context.

```
diabetes <- read.csv("C:/Users/ansar/Desktop/final/diabetes.csv")
attach(diabetes)
par(mfrow=c(2,4))
boxplot(Pregnancies~Outcome, main="No. of Pregnancies vs. Diabetes",
        xlab="Outcome", ylab="Pregnancies",col="red")
boxplot(Glucose~Outcome, main="Glucose vs. Diabetes",
        xlab="Outcome", ylab="Glucose",col="pink")
boxplot(BloodPressure~Outcome, main="Blood Pressure vs. Diabetes",
        xlab="Outcome", ylab="Blood Pressure",col="green")
boxplot(SkinThickness~Outcome, main="Skin Thickness vs. Diabetes",
        xlab="Outcome", ylab="Skin Thickness",col="orange")
boxplot(Insulin~Outcome, main="Insulin vs. Diabetes",
        xlab="Outcome", ylab="Insulin",col="yellow")
boxplot(BMI~Outcome, main="BMI vs. Diabetes",
        xlab="Outcome", ylab="BMI",col="purple")
boxplot(DiabetesPedigreeFunction~Outcome, main="Diabetes Pedigree Function vs. Diabetes'
boxplot(Age~Outcome, main="Age vs. Diabetes",
        xlab="Outcome", ylab="Age",col="lightblue")
box(which = "outer", lty = "solid")
```



Based on the observation that blood pressure and skin thickness demonstrate minimal variation with diabetes, they will be excluded from the model. However, considering that the remaining variables exhibit varying degrees of correlation with diabetes, they will be retained for further analysis.

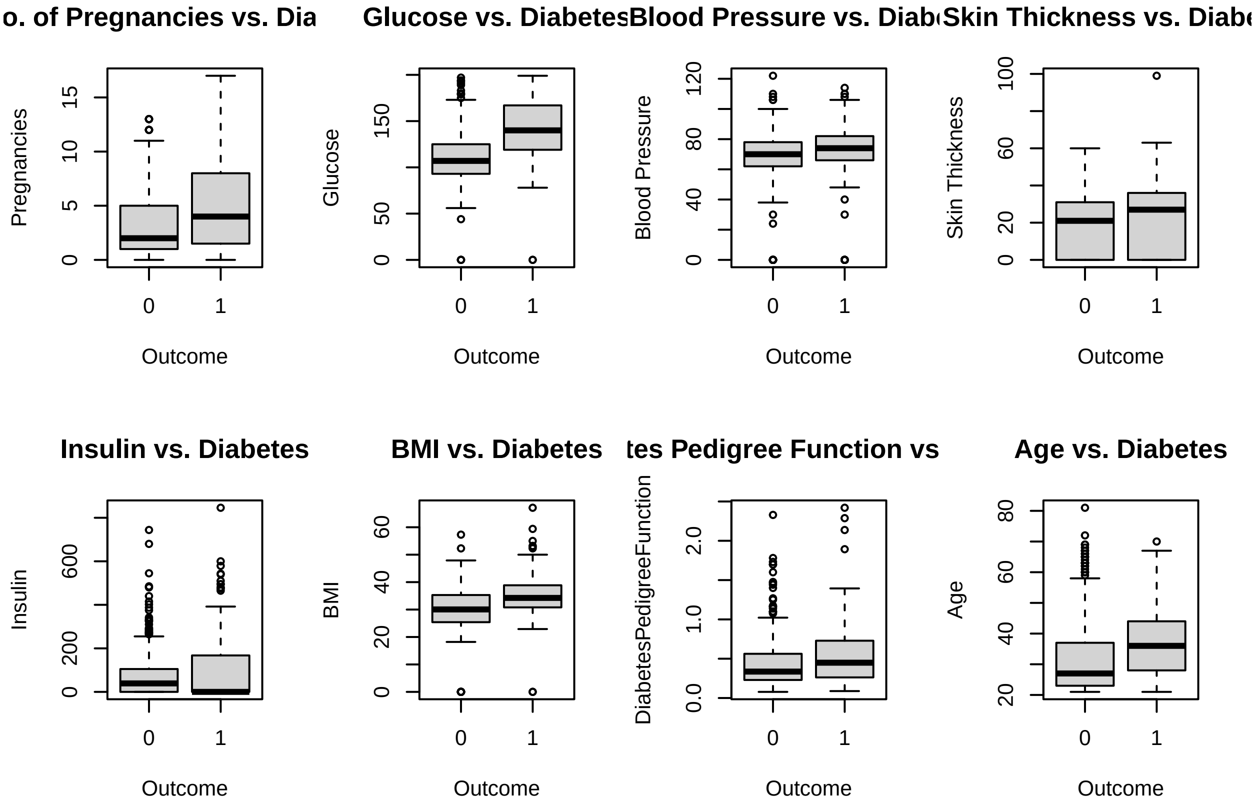
The numeric variabls are almost not correlated.

Correlation bewteen numeric variables and outcome.

```
attach(missing_data)
```

```
## The following objects are masked from diabetes:
##
##     Age, BloodPressure, BMI, DiabetesPedigreeFunction, Glucose,
##     Insulin, Pregnancies, SkinThickness
```

```
par(mfrow=c(2,4))
boxplot(Pregnancies~Outcome, main="No. of Pregnancies vs. Diabetes",
        xlab="Outcome", ylab="Pregnancies")
boxplot(Glucose~Outcome, main="Glucose vs. Diabetes",
        xlab="Outcome", ylab="Glucose")
boxplot(BloodPressure~Outcome, main="Blood Pressure vs. Diabetes",
        xlab="Outcome", ylab="Blood Pressure")
boxplot(SkinThickness~Outcome, main="Skin Thickness vs. Diabetes",
        xlab="Outcome", ylab="Skin Thickness")
boxplot(Insulin~Outcome, main="Insulin vs. Diabetes",
        xlab="Outcome", ylab="Insulin")
boxplot(BMI~Outcome, main="BMI vs. Diabetes",
        xlab="Outcome", ylab="BMI")
boxplot(DiabetesPedigreeFunction~Outcome, main="Diabetes Pedigree Function vs. Diabetes'
boxplot(Age~Outcome, main="Age vs. Diabetes",
        xlab="Outcome", ylab="Age")
```



Due to the limited variation observed in blood pressure and skin thickness in relation to diabetes, it is decided to exclude these variables from the model. On the other hand, since the remaining variables demonstrate varying degrees of correlation with diabetes, they will be retained in the analysis. This decision is based on the understanding that the retained variables provide more meaningful insights and have stronger associations with the occurrence of diabetes.

Taking into consideration the understanding of color blindness, I made a deliberate effort to limit the use of colors in the design. Instead, I opted for a color scheme that predominantly features black and white, ensuring that the information remains accessible for individuals with this condition. By providing a black and white version alongside the colored version, I aimed to accommodate the needs of both individuals with normal color vision and those affected by color blindness.

6.2 Logistic Regression

Logistic Regression is a statistical method used for binary classification tasks, where the goal is to predict the probability of an event occurring or not occurring. It is commonly used in machine learning and statistics to model the relationship between one or more independent variables (features) and a binary outcome variable (target).

```
df$BloodPressure <- NULL
df$SkinThickness <- NULL
train <- df[1:540,]
test <- df[541:768,]
model <-glm(Outcome ~.,family=binomial(link='logit'),data=train)
summary(model)
```

```
##
## Call:
## glm(formula = Outcome ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4366  -0.7741  -0.4312   0.8021   2.7310
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.3461752  0.8157916 -10.231  < 2e-16 ***
## Pregnancies               0.1246856  0.0373214   3.341 0.000835 ***
## Glucose                   0.0315778  0.0042497   7.431 1.08e-13 ***
## Insulin                  -0.0013400  0.0009441  -1.419 0.155781
## BMI                       0.0881521  0.0164090   5.372 7.78e-08 ***
## DiabetesPedigreeFunction  0.9642132  0.3430094   2.811 0.004938 **
## Age                       0.0018904  0.0107225   0.176 0.860053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 700.47  on 539  degrees of freedom
## Residual deviance: 526.56  on 533  degrees of freedom
## AIC: 540.56
##
## Number of Fisher Scoring iterations: 5
```

The top three most relevant features are “Glucose”, “BMI” and “Number of times pregnant” because of the low p-values.

“Insulin” and “Age” appear not statistically significant.

```
anova(model, test="Chisq")
```

```
## Analysis of Deviance Table
```

```
##
## Model: binomial, link: logit
##
## Response: Outcome
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                       539     700.47
## Pregnancies               1   26.314       538     674.16 2.901e-07 ***
## Glucose                   1  102.960       537     571.20 < 2.2e-16 ***
## Insulin                   1    0.062       536     571.14  0.803341
## BMI                       1   36.135       535     535.00 1.841e-09 ***
## DiabetesPedigreeFunction  1    8.414       534     526.59  0.003723 **
## Age                       1    0.031       533     526.56  0.860201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table of deviance, we can see that adding insulin and age have little effect on the residual deviance.

6.3 Cross Validation

Cross-validation is a technique used to assess the performance and generalizability of a machine learning model. It involves partitioning the available data into multiple subsets to train and evaluate the model multiple times, providing a more robust estimate of its performance.

```
fitted.results <- predict(model,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$Outcome)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.789473684210526"
```

In this project, a comparison was made between the performance of Logistic Regression and Decision Tree algorithms. It was observed that Logistic Regression outperformed the Decision Tree algorithm on the standard, unaltered dataset. However, to enhance the generalization performance in decision tree induction, techniques such as pruning can be implemented. These future posts will focus on applying pruning methods to improve the performance of the Decision Tree algorithm. By implementing pruning, it is expected to achieve better results in terms of the model's ability to generalize to unseen data.

7 Statistics

This paper delves into the critical aspects of handling confounder variables, the significance of plotting histograms for numeric values, and briefly explores their impact on predictive modeling. In particular, we employ decision tree-based techniques to predict outcomes based on the data attributes. The investigation presents a comprehensive analysis of confounder variables and their potential influence on data interpretation, followed by a detailed exposition on the usage of histogram plotting for visualizing numeric data distributions. Lastly, we implement decision tree models to predict the target variable, demonstrating the effectiveness of this approach in generating accurate predictions..

7.1 Adding a Confounder

To enhance our existing model, we will incorporate a confounding variable. We possess a variable named “Pregnancies,” which indicates the number of previous pregnancies. Our objective is to establish a new variable with a binary outcome: “History of Pregnancy” or “No History of Pregnancy.” To accomplish this, we must subset the data and establish specific criteria. Individuals with one or more pregnancies will be assigned a code of 1, indicating a history of pregnancy. Conversely, those without any previous pregnancies will be assigned a code of 0, denoting no history of pregnancy.

```r
df$pregnancy.history[df$Pregnancies == 0] = 0
df$pregnancy.history[df$Pregnancies > 0] = 1

table(df$pregnancy.history)
```

```
##
##   0   1
## 111 657
```

We observe that out of the total population, there are 657 women who have a record of past pregnancies, while 111 women do not have any history of pregnancy.

A DAG diagram illustrating the relationship between Pregnancy History as a confounder on the Age to Glucose relationship can be drawn.

7.2 Plotting Histograms of Numeric Values

Over time, diabetes can cause damage to the heart, blood vessels, eyes, kidneys, and nerves. Adults with diabetes have a two- to three-fold expanded risk of heart attacks and strokes . Combined with reduced bloodstream, neuropathy (nerve harm) within the feet increases the chance of foot ulcers, disease, and the inevitable requirement for limb amputation. Diabetic retinopathy is a critical cause of the visual deficiency and happens as a result of long-term

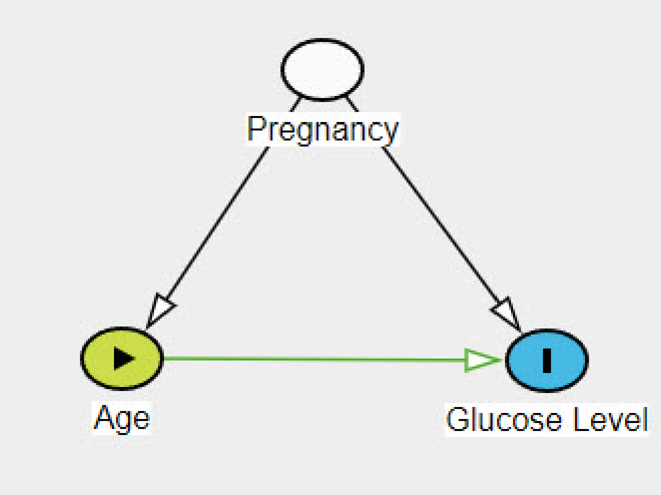


Figure 3: Confounder (Bounthavong 2022)

accumulated harm to the small blood vessels within the retina. Near to 1 million individuals are dazed due to diabetes . Diabetes is among the driving causes of kidney failure . People with diabetes are more likely to have poor results for a few irresistible illnesses, counting COVID-19 (Nair 2022)

```r
library(ggplot2)
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(corrplot)
```

```
## corrplot 0.92 loaded
```

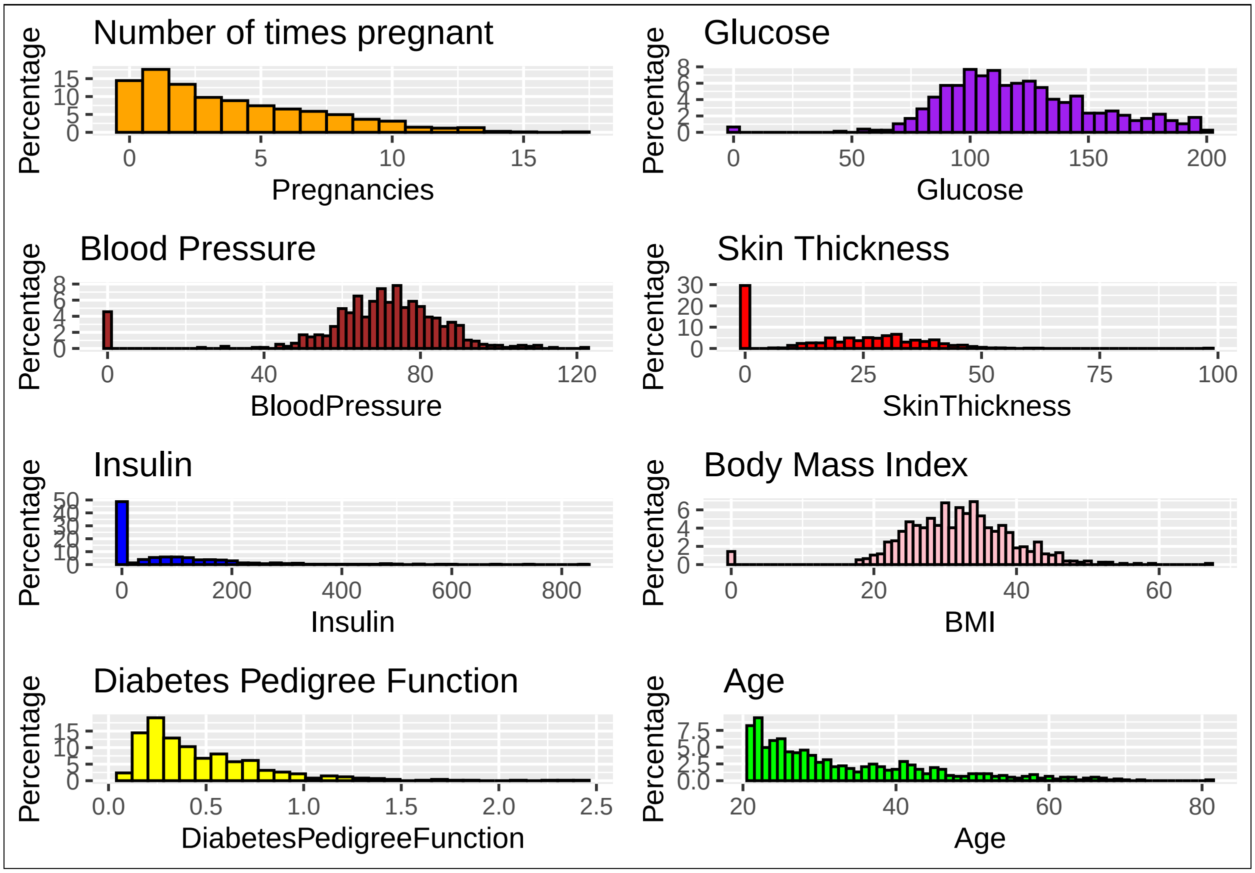
```r
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'
## The following object is masked from 'package:corrgram':
##
##     panel.fill
```

```
library(e1071)

p1 <- ggplot(missing_data, aes(x=Pregnancies)) + ggtitle("Number of times pregnant") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 1, colour="Black",
p2 <- ggplot(df, aes(x=Glucose)) + ggtitle("Glucose") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 5, colour="black",
p3 <- ggplot(df, aes(x=BloodPressure)) + ggtitle("Blood Pressure") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 2, colour="black",
p4 <- ggplot(df, aes(x=SkinThickness)) + ggtitle("Skin Thickness") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 2, colour="black",
p5 <- ggplot(df, aes(x=Insulin)) + ggtitle("Insulin") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 20, colour="black"
p6 <- ggplot(df, aes(x=BMI)) + ggtitle("Body Mass Index") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 1, colour="black",
p7 <- ggplot(df, aes(x=DiabetesPedigreeFunction)) + ggtitle("Diabetes Pedigree Function"
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), colour="black", fill="yellow"
p8 <- ggplot(df, aes(x=Age)) + ggtitle("Age") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth=1, colour="black", fi
grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, ncol=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
grid.rect(width = 1, height = 1, gp = gpar(lwd = 1, col = "black", fill = NA))
```

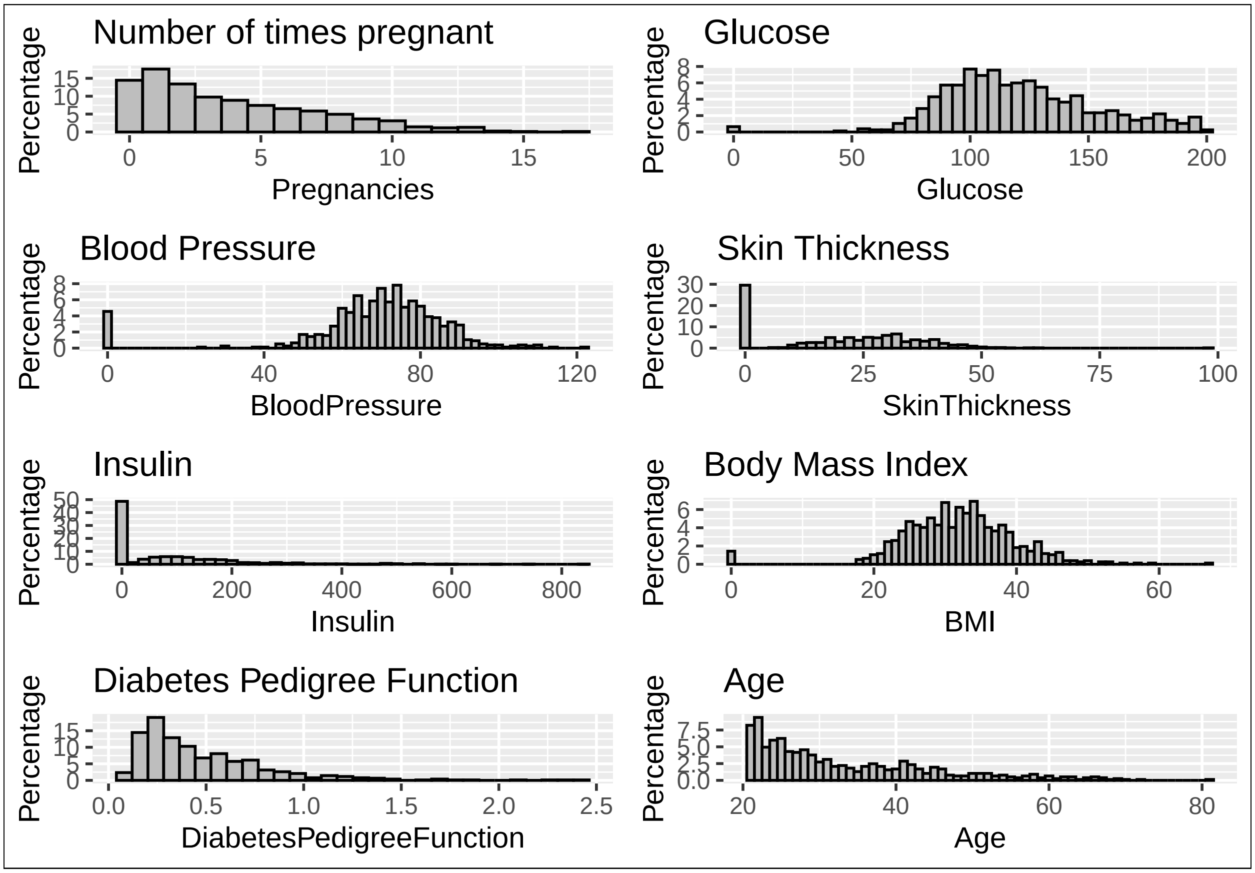


and black and white version

```
library(ggplot2)
library(grid)
library(gridExtra)
library(corrplot)
library(caret)
library(e1071)

df <- read.csv("C:/Users/ansar/Desktop/final/diabetes.csv")
p1 <- ggplot(df, aes(x=Pregnancies)) + ggtitle("Number of times pregnant") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 1, colour="black",
p2 <- ggplot(df, aes(x=Glucose)) + ggtitle("Glucose") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 5, colour="black",
p3 <- ggplot(df, aes(x=BloodPressure)) + ggtitle("Blood Pressure") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 2, colour="black",
p4 <- ggplot(df, aes(x=SkinThickness)) + ggtitle("Skin Thickness") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 2, colour="black",
p5 <- ggplot(df, aes(x=Insulin)) + ggtitle("Insulin") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 20, colour="black"
p6 <- ggplot(df, aes(x=BMI)) + ggtitle("Body Mass Index") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth = 1, colour="black",
p7 <- ggplot(df, aes(x=DiabetesPedigreeFunction)) + ggtitle("Diabetes Pedigree Function"
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), colour="black", fill="gray") +
p8 <- ggplot(df, aes(x=Age)) + ggtitle("Age") +
  geom_histogram(aes(y = 100*(..count..)/sum(..count..)), binwidth=1, colour="black", fi
grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, ncol=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
grid.rect(width = 1, height = 1, gp = gpar(lwd = 1, col = "black", fill = NA))
```



Given that all the variables exhibit reasonably broad distributions, they will be retained for the regression analysis.(Babar 2022)

7.3 Decision Tree

A Decision Tree is a popular and interpretable machine learning algorithm used for both classification and regression tasks. It is a tree-like structure where each internal node represents a decision based on a feature, each branch represents an outcome of that decision, and each leaf node represents the final prediction or decision. In this paper we used this method for better prediction.

```
anova(model, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Outcome
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                       539     700.47
## Pregnancies               1   26.314       538     674.16 2.901e-07 ***
## Glucose                   1  102.960       537     571.20 < 2.2e-16 ***
## Insulin                   1    0.062       536     571.14  0.803341
## BMI                       1   36.135       535     535.00 1.841e-09 ***
## DiabetesPedigreeFunction  1    8.414       534     526.59  0.003723 **
## Age                       1    0.031       533     526.56  0.860201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

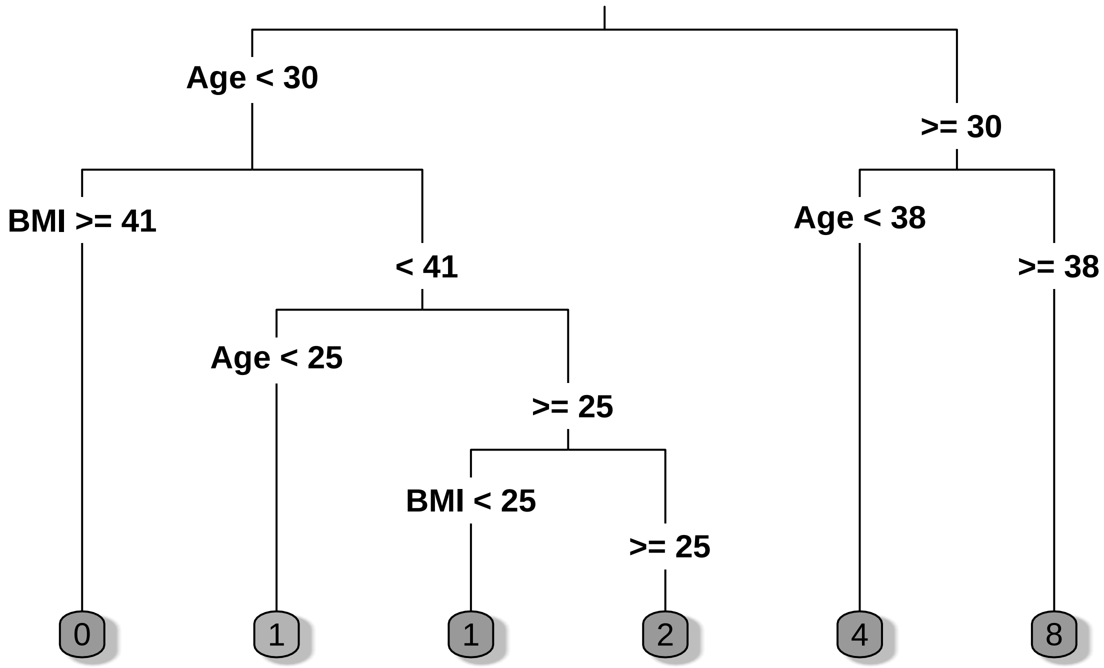
```
library(mlbench)
library(rpart)
library(rpart.plot)
library(caret)
library(Metrics)
```

```
##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```
data(diabetes)
dfn <- na.omit(df)
dplyr::glimpse(dfn)
```

```
## Rows: 768
## Columns: 9
## $ Pregnancies              <int> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, ~
## $ Glucose                  <int> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125~
## $ BloodPressure            <int> 72, 66, 64, 66, 40, 74, 50, 0, 70, 96, 92, 74~
## $ SkinThickness            <int> 35, 29, 0, 23, 35, 0, 32, 0, 45, 0, 0, 0, 0, ~
## $ Insulin                  <int> 0, 0, 0, 94, 168, 0, 88, 0, 543, 0, 0, 0, 0, ~
## $ BMI                      <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0, 35.~
## $ DiabetesPedigreeFunction <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.2~
## $ Age                      <int> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 3~
## $ Outcome                  <int> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, ~
```

```
set.seed(123)
index <- sample(2, nrow(df), prob = c(0.8, 0.2), replace = TRUE)
dfn_train <- df[index==1, ]
dfn_test <- df[index == 2, ]
dfn_model <- rpart(formula = dfn ,
                        method = "class")
rpart.plot(x = dfn_model, yesno = 2, type = 3, extra = 0,shadow.col = "gray", box.palett
```



According to the analysis, individuals with a BMI below 45.4 and a Diabetes Pedigree function lower than 0.8745 have an elevated probability of having diabetes. These specific thresholds for BMI and the Diabetes Pedigree function indicate a higher risk of diabetes based on the findings of this particular analysis.

8 Conclusion

The number of individuals with diabetes has significantly increased from 108 million in 1980 to 422 million in 2014. This rise has been more pronounced in low- and middle-income countries compared to high-income countries. Diabetes is a leading cause of visual impairment, kidney failure, heart attacks, strokes, and lower limb amputation. Between 2000 and 2019, there was a 3% increase in diabetes mortality rates by age. In 2019 alone, it was estimated that diabetes and kidney disease related to diabetes resulted in 2 million deaths.

Embracing a nutritious eating plan, participating in consistent physical activity, achieving and maintaining a healthy body weight, and refraining from tobacco consumption are successful approaches to stave off or postpone the development of type 2 diabetes. The management of diabetes and the prevention of associated complications can be achieved by implementing appropriate dietary practices, engaging in regular exercise, adhering to prescribed medications, undergoing routine screenings, and receiving treatment for any arising complications. Impaired glucose tolerance (IGT) and impaired fasting glycaemia (IFG) serve as transitional stages between normal blood sugar levels and diabetes.Individuals with IGT or IFG are at a high risk of progressing to type 2 diabetes, although this progression is not inevitable.(Forouhi and Wareham 2010).

In this project, we conducted a comparison of the performance between Linear Regression and Decision Tree algorithms. Our findings revealed that Linear Regression outperformed Decision Tree on the unaltered dataset, exhibiting higher accuracy. This indicates that Linear Regression is more suitable for the given task.

In the future, the developed system, utilizing machine learning classification algorithms, can be applied to predict or diagnose other diseases. There is potential for further extension and enhancement of the work, aiming to automate the analysis of diabetes and incorporate additional machine learning algorithms for improved results.

9 Challenges and Improvement Area

Throughout this paper, various challenges have been encountered, some of which have been mentioned in the text. Given more time, a concerted effort would have been made to enhance the efficiency and quality of the graphs presented. A particular area that requires improvement is the data processing phase, as the database contains inexplicable zero values, such as a BMI of 0, which is implausible. Although these discrepancies are evident in the graphs, addressing them adequately would necessitate additional time and resources.

Another concern observed in other research papers is the indiscriminate removal of data without justifiable reasons. Such data exclusion can lead to skewed results and introduce bias. It is crucial to identify outliers and handle them judiciously. While the removal of outliers is acceptable under certain circumstances, it must be accompanied by valid reasons. For instance, data entry errors could warrant correction or exclusion. In our case, the certainty of whether the outliers resulted from erroneous data entry remains uncertain. Consequently, the outright removal of these observations lacks sufficient justification, prompting the need for a more thoughtful approach to handling outliers.

Moreover, there exist numerous methods for calculating regression, each with its own advantages and drawbacks. Unfortunately, due to time constraints, a comprehensive exploration of all available regression techniques could not be undertaken in this study.

Additionally, a structural issue that was identified involves certain headers on pages overlapping with the project's title (e.g., page 36-38). This inconsistency should be addressed to ensure the document's visual coherence and professionalism.

In the final moments of this paper, it became apparent that the inclusion of numerical labels and a reference list for the graphs would have significantly improved their comprehensibility and verifiability. Regrettably, time limitations hindered the implementation of these enhancements.

Upon completing our report, we encountered the challenge of making contributions using Git and GitHub. In order to overcome this obstacle, I diligently studied a tutorial that thoughtfully combined step-by-step textual explanations with video demonstrations.

At the outset, we encountered difficulty in adding my name to the list of contributors. The intricacies of Git commands and GitHub procedures proved to be a hurdle. Thankfully, the generous support and guidance from Prof. Dr. Huber proved instrumental in resolving this issue.

Thanks to Prof. Dr. Huber's assistance, I was able to successfully add my name to the list of contributors. As a result, I now possess the knowledge and understanding necessary to make contributions using Git and GitHub confidently.

In conclusion, despite encountering various challenges throughout this paper, concerted efforts were made to present the research findings accurately. Nevertheless, there remain areas that warrant further attention, and given additional time, improvements could be made to refine the visual representation of data and address structural discrepancies. In future research, an exploration of alternative regression methods and a more thorough examination of data processing techniques could provide valuable insights and enrich the quality of the study's outcomes.

Refrences

Alessandro Massaro, Gabriele Cosoli, Nicola Magaletti. 2022. “The Prediction of Diabetes,” June.

Babar, Shraddha. 2022. “DIABETES PREDICTION.” https://rpubs.com/Shraddha20/Diabetes_Prediction_using_R.

Bounthavong, Mark. 2022. “R Tutorial on Linear Regression Model.” https://rpubs.com/mbounthavong/linear_regression_using_R.

clinic, Mayo. 2022. “Types of Diabetes.” https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/symptoms-causes/syc-20351193.

Disease Control, Centers for, and Prevention. 2022. “What Is Diabetes?” https://www.cdc.gov/diabetes/basics/diabetes.html#print.

Donoho., David L. 2017. “50 Years of Data Science.” *Journal of Computational and Graphical Statistics* 26: 745–66.

Forouhi, Nita Gandhi, and Nicholas J Wareham. 2010. “Epidemiology of Diabetes.” *Medicine* 38 (11): 602–6.

Health Organization, world. 2022. “Diabetes.” https://www.who.int/news-room/fact-sheets/detail/diabetes.

Huntley AC, Walter RM Jr. 1990. “Quantitative Determination of Skin Thickness in Diabetes Mellitus: Relationship to Disease Parameters.” *Journal of Medicine* 21: 257–264.

Lawlor, D. A. et al. n.d. “Body Mass Index (BMI).” https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html#print.

Nair, Chauhan, D. R. 2022. “Were US Asian Indian Decedents with Atherosclerosis More Likely to Have Concurrent Diabetes Mellitus? Analysis of National Multiple Cause of Mortality Data.” *Diabetology & Metabolic Syndrome* 14(1): 159.

P J Clements 1, J R Seibold, P A Lachenbruch. 1993. “Skin Thickness Score in Systemic Sclerosis: An Assessment of Interobserver Variability in 3 Independent Studies.” *The Journal of Rheumatology*, 1892–96.

Provost, & Fawcett, F. 2013. *Data Science for Business*. Sebastopol, Calif.

Roglic, Gojka et al. 2016. “WHO Global Report on Diabetes: A Summary.” *International Journal of Noncommunicable Diseases* 1 (1): 3.

Sarwar N, Seshasai SR, Gao P. 2010. “Diabetes Mellitus, Fasting Blood Glucose Concentration, and Risk of Vascular Disease: A Collaborative Meta-Analysis of 102 Prospective Studies. Emerging Risk Factors Collaboration.” 2215–22.

T. Hastie, R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics Springer New York Inc.

United States Renal Data System. National Institutes of Health, National Institute of Diabetes, Digestive, and MD Kidney Diseases Bethesda. 2021.

Watson, Stephanie. 2022. “What Is Glucose?” https://www.webmd.com/diabetes/glucose-diabetes.

Wickham, Hadley. 2016. *GGPLOT2 ESSENTIALS FOR GREAT DATA VISUALIZATION IN r*. Springer-Verlag New York. https://ggplot2.tidyverse.org.