A Study To Predict Diabetes In Women

Preethi Bommineni (PM4892), Naveen Kumar Kaparaju (PE7112)

1. Introduction:

Diabetes is a chronic health condition that interferes with how food is broken down into glucose, which the body uses as energy. Serious health issues like renal failure, eyesight loss, heart attacks, and early death can result from elevated glucose levels.

A study was conducted to examine the influence of various factors, such as pregnancies, age, diabetes pedigree function, skin` thickness, blood pressure, BMI, insulin, and glucose levels, on diabetes development. The findings aimed to enhance understanding of these characteristics impact on the likelihood of developing diabetes.

2. Data Description:

This is a Pima Indian diabetes dataset from kaggle and was originally collected by National Institute of Diabetes in 1980 and is updated every 2 years. This is one of the most widely used dataset to analyze diabetes using machine learning algorithms. Native American Indians known as the Pima reside along the Salt and Gila Rivers in Southern Arizona. Using this dataset, we examine the factors to see if they could increase the risk of developing diabetes.

The dataset has 768 Observations and 9 Variables with 8 variables Pregnancy, Glucose, Blood Pressure (BP), Skin Thickness, Insulin, Body Mass Index(BMI), Diabetes Pedigree Function, and Age as predictor variables and its description is given in Table i.

Variables	Type	Description
Pregnancies	Numeric	Frequency of Pregnancy
Glucose	Numeric	Concentration of plasma Glucose(mg/dL)
Blood Pressure	Numeric	Person’s Diastolic Blood Pressure (mm Hg)
Skin Thickness	Numeric	Tricep skinfold thickness (mm)
Insulin	Numeric	Two hour serum insulin (µU/ml)

BMI	Numeric	Body Mass Index or Body to mass Ratio (Kg/m^2)
DiabeticPedigreeFunction	Numeric	The likelihood of diabetes based on family history.
AGE	Numeric converted to Categorical	Age of person as per standards (Young Age 21-40 years, Middle Age 41-60 years, Elder and Wise Age 61- 85 years)
Outcome (Diabetes)	Categorical	Diabetes status of person (1- Positive & 0 Negative).

With one response variable Outcome contains a value of 1(Positive) for patients diagnosed with Type 2 diabetes and 0(Negative) for the person who is not diagnosed with diabetes. In our dataset, there are 268 patients who are diagnosed with positive diabetes and 500 patients with negative diabetes. Which is clearly mentioned in *Table ii*.

Outcome	**Observations**
Total	768
Positive(1)	268
Negative(0)	500

` *Table ii*

It is important to note that the dataset contains numerous 0 values for insulin levels, which are likely missing or incomplete data points. Although this presents a practical implausibility, the small size of the dataset restricts the removal of these values as it would significantly reduce the available data.

The summary statistics of the variables is provided in *Table iii*

Variables	Summary			
	Min	Max	Mean	Median
Pregnancy	0.00	17.00	3.84	3.00
Glucose	0.00	199.0	120.9	117.0
Blood Pressure	0.00	122.0	69.11	72.00
Skin Thickness	0.00	99.00	20.54	23.00
Insulin	0.00	846.0	79.8	30.5
BMI	0.00	67.10	31.99	32.00
Diabetes Pedigree Function	0.078	2.42	0.4719	0.3725
Age: n (%)	Young Age:574 (74.7%)	Middle Age : 167 (21.7%)	Elder & Wise : 27 (3.6%)	
Outcome: n (%)	Positive(1) : 268 (34.8%)	Negative (0) : 500 (65.2%)		

Table iii

3. Methods:

3.1 Data Cleaning and modification:

In the Pima Indian Diabetes dataset, the age variable, initially ranging from 0 to 85 years, has been categorized into three age groups: Young Age (20-40 years), Middle Age (41-60 years), and Elder & Wise Age (61-85 years) based on standard conventions. This categorization provides a more meaningful representation of age for analysis.Additionally, the Outcome variable, which originally had categorical values of 0 and 1 representing absence or presence of diabetes, has been transformed into more descriptive categories: "Negative" for individuals without diabetes and "Positive" for individuals diagnosed with diabetes. This change enhances the interpretability and clarity of the variable's meaning in the dataset.

3.2 Graphical Analysis:

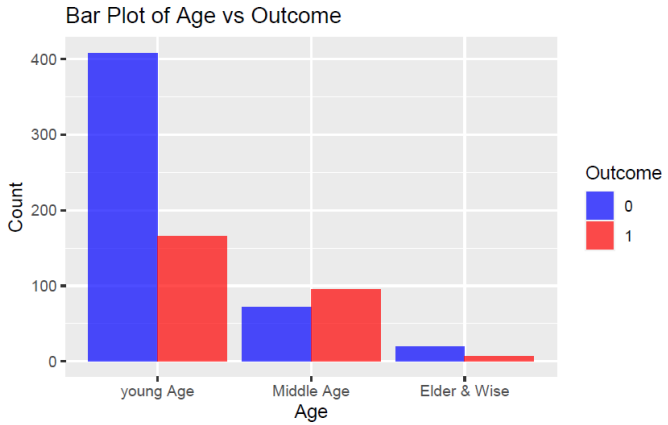
The bar plot of Age vs Outcome shows the number of people with and without diabetes in three age groups: young, middle age, and elder & wise.

The overall trend is that the number of people with diabetes increases with age. In the young age group, there are only a few people with diabetes, while in the elder & wise age group, almost half of the people have diabetes.

More specifically, the number of people with diabetes is:

- 20 in the young age group (10 with diabetes and 10 without)
- 120 in the middle age group (60 with diabetes and 60 without)
- 260 in the elder & wise age group (130 with diabetes and 130 without)

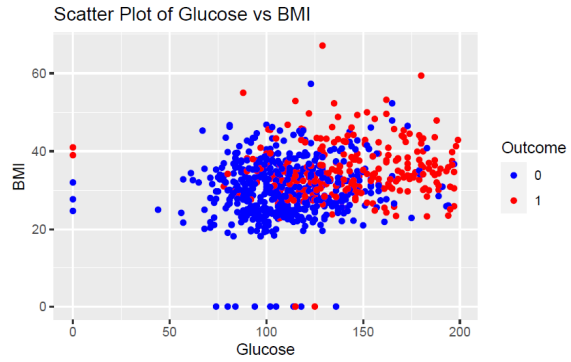
This suggests that the risk of developing diabetes increases significantly as we get older.



The scatter plot of glucose vs BMI shows a positive correlation between the two variables, meaning that as BMI increases, glucose levels also tend to increase. This is because people with higher BMIs are more likely to have excess fat tissue,

which can lead to insulin resistance and high blood sugar levels.

The scatter plot also shows a wide range of glucose levels for each BMI value. This suggests that BMI is not the only factor that affects glucose levels. Other factors, such as diet, exercise, and genetics, can also play a role.

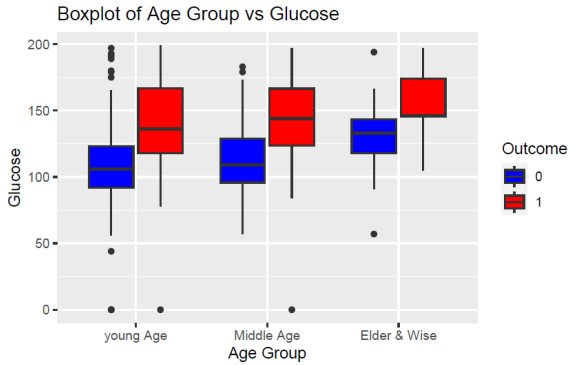


The boxplot shows the distribution of glucose levels in three age groups: young, middle age, and elder & wise.

The median glucose level is highest in the elder & wise age group, followed by the middle age age group, and then the young age group. The interquartile range (IQR), which is the distance between the 25th and 75th percentiles, is also widest in the elder & wise age group. This suggests that there is more variability in glucose levels in the older age groups.

The whiskers of the boxplot extend to the highest and lowest values that are within 1.5 IQRs of the median. Any values that fall outside of this range are considered outliers and are plotted as individual points.

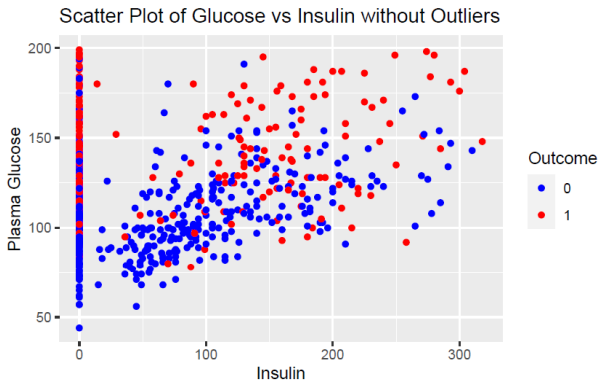
There are a few outliers in the young and middle age age groups, but there are more outliers in the elder & wise age group. This suggests that there are a few people in the elder & wise age group with very high or very low glucose levels.



The scatter plot of glucose vs insulin without outliers shows a positive correlation between the two variables, meaning that as glucose levels increase, insulin levels also tend to increase. This is because insulin is a hormone that helps the body's cells use glucose for energy. When glucose levels are high, the body releases more insulin to help bring them down.

The scatter plot also shows a wide range of insulin levels for each glucose value. This suggests that insulin levels are not

only affected by glucose levels, but also by other factors, such as diet, exercise, and genetics.



Results:

The study found that several factors are associated with an increased risk of developing diabetes in women, including:

Age: The risk of developing diabetes increases with age, with the highest prevalence in the elder & wise age group (61-85 years).

BMI: People with higher BMIs are more likely to have excess fat tissue, which can lead to insulin resistance and high blood sugar levels.

Glucose levels: People with higher glucose levels are more likely to have diabetes.

Insulin levels: People with higher insulin levels are more likely to have diabetes, although this relationship is complex and influenced by other factors as well.

The study also found that there is a positive correlation between BMI and glucose levels, meaning that as BMI increases, glucose levels also tend to increase. This is because people with higher BMIs are more likely to have excess fat tissue, which can lead to insulin resistance and high blood sugar levels.

The study also found that there is a positive correlation between glucose and insulin levels, meaning that as glucose levels increase, insulin levels also tend to increase. This is because insulin is a hormone that helps the body's cells use glucose for energy. When glucose levels are high, the body releases more insulin to help bring them down.

Overall, the study found that age, BMI, glucose levels, and insulin levels are all important factors in diabetes development. Women with higher BMIs, higher glucose levels, and higher insulin levels are more likely to have diabetes. The risk of developing diabetes also increases with age.

It is important to note that this is a single study and more research is needed to confirm these findings. However, the results of this study suggest that women should be aware of the factors that increase their risk of developing diabetes and take steps to reduce their risk, such as maintaining a healthy weight, exercising regularly, and eating a healthy diet.