APPLIED DATA SCIENCE – 1

ASSIGNMENT 1: STATISTICS AND TRENDS

TITLE: Data Analysis of Health Indicators for Diabetes Prediction

ABSTRACT

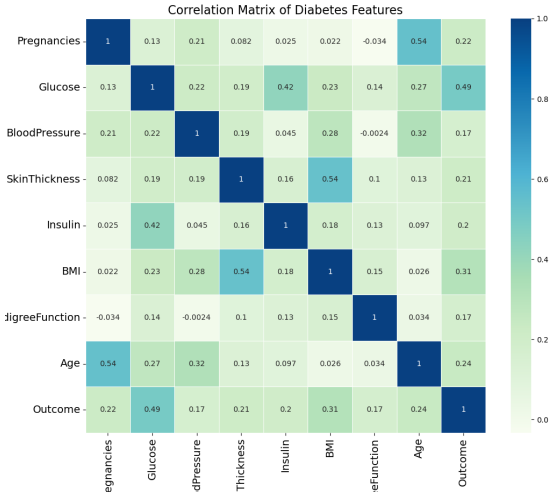
In this report, we analyse a diabetes dataset to uncover patterns and relationships that may be associated with diabetes risk. By examining how factors such as age, glucose levels, blood pressure, BMI, and insulin levels relate to diabetes diagnoses, we aim to understand which features are most linked to the condition. Through various visualizations, we bring to light the correlations and distributions in the data, with the goal of identifying key risk indicators for diabetes. The insights derived from this analysis could help inform early detection strategies and support the development of preventative measures for individuals at risk of diabetes.

NAME: *PRANEET SIVAKUMAR*
STUDENT ID: *23095964*
GITHUB LINK: https://github.com/Prani8/diabetes_assignment-1

Introduction

Diabetes is a chronic health condition impacting millions globally, often associated with factors like high blood pressure, elevated glucose levels, increased body weight, and age. Understanding these interconnected risk factors is vital for early detection and prevention. In this report, we analyse a diabetes dataset to explore how variables such as glucose levels, BMI, age, blood pressure, and insulin levels relate to diabetes diagnoses. Through visualizations, we uncover patterns and relationships that shed light on the complex nature of diabetes risk, aiming to support more informed approaches to diabetes management and preventive healthcare.

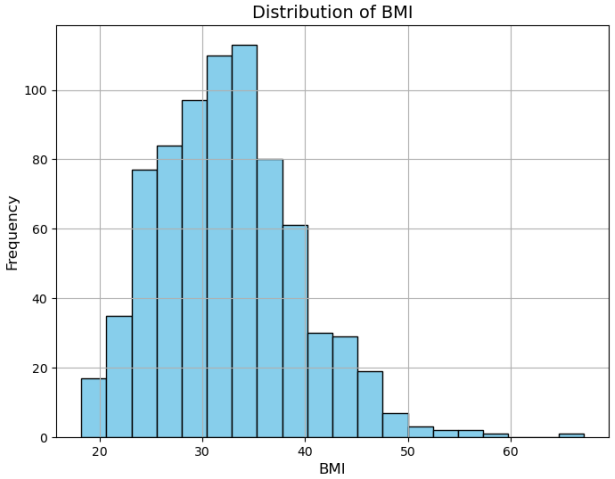
1. Correlation Matrix of Diabetes Features



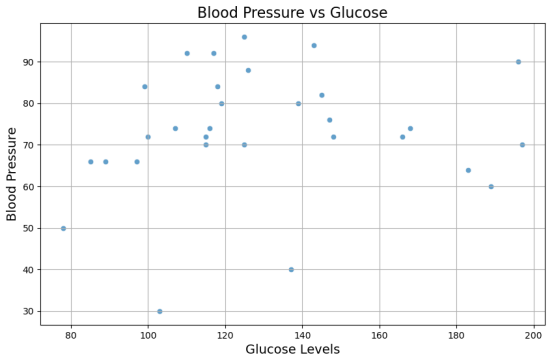
This heatmap visualizes how various factors in the dataset, such as glucose levels, blood pressure, BMI, and others, relate to each other. Each cell in the heatmap represents the strength of the relationship between two variables, with darker shades indicating stronger correlations. A notable observation is the relatively strong positive correlation between glucose levels and diabetes outcome, which aligns with what we know about high glucose being a primary indicator of diabetes. Additionally, we see moderate correlations between age and BMI, as well as between BMI and skin thickness, offering insight into the interconnected nature of these health metrics.

1. Distribution of BMI Levels

This histogram displays the distribution of BMI values among the dataset's individuals. On the x-axis, we have BMI levels, while the y-axis represents the frequency of each BMI range. Most individuals fall within a BMI range of 20-40, with a peak around 30. This right-skewed distribution suggests that some individuals have significantly high BMI values, which could potentially increase their risk of developing diabetes. This visualization underscores the importance of weight management as a factor in diabetes prevention, as higher BMI levels are often associated with increased diabetes risk.



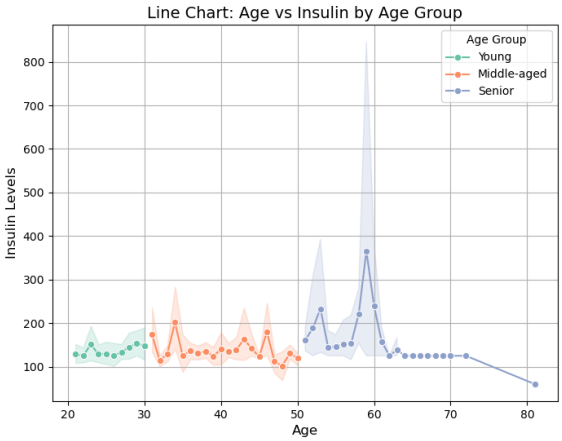
2. Blood Pressure vs Glucose Levels



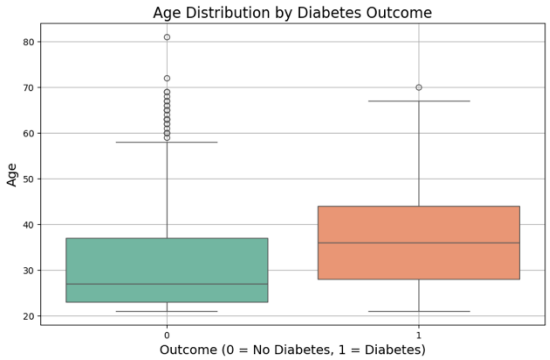
The scatter plot illustrates the relationship between blood pressure and glucose levels. The x-axis represents glucose levels, and the y-axis shows blood pressure readings. A positive trend can be seen, with individuals who have higher glucose levels often exhibiting higher blood pressure. This relationship highlights a known link between elevated glucose and blood pressure, both of which are recognized risk factors for diabetes. This scatter plot supports the idea that individuals with higher glucose and blood pressure levels may be at greater risk of developing diabetes.

3. Age vs Insulin by Age Group

The line chart provides a breakdown of insulin levels across different age groups: young (under 30), middle-aged (30-50), and senior (50+). Here, the x-axis represents age, and the y-axis shows insulin levels. We observe that insulin levels vary more widely in older age groups, particularly among seniors, suggesting that insulin resistance may increase with age. This trend aligns with clinical knowledge that aging is often accompanied by changes in insulin sensitivity, making older adults more vulnerable to conditions like diabetes.



4. Age Distribution by Diabetes



The box plot illustrates the distribution of age across individuals with and without diabetes. Here, the x-axis categorizes people based on their diabetes status: “0” represents those without diabetes, and “1” represents those with diabetes. The y-axis shows their age. We observe that individuals with diabetes tend to be older on average. However, there are also a few older individuals in the non-diabetic group, indicating that age alone may not be the sole predictor of diabetes. This visualization suggests that, while age plays a role, it is likely part of a broader combination of risk factors.

Conclusion

In summary, our analysis highlights several interconnected factors associated with diabetes risk, with glucose levels emerging as a primary indicator. Higher BMI and older age also show a connection to diabetes, although these factors alone are not definitive. Instead, the data suggests that diabetes risk is influenced by a combination of age, body composition, blood pressure, and insulin levels. These insights emphasize the importance of a holistic approach to diabetes prevention and management, where multiple health metrics are monitored together rather than in isolation. By understanding these relationships, healthcare providers and individuals can make more informed decisions, potentially leading to early detection and better management of diabetes.