

Analyzing Diabetes Dataset

[1] Introduction

Diabetes is a metabolic disorder which is prevalent in several millions across the globe thus leading to difficulties in managing diabetes. This is made even worse as there are very few patients who are able to get their diabetes diagnosed in its early stages. There is however light in that in terms of how the future might be as through technology and innovative expertise, diabetes management would be cost effective. Machine learning has turned into a dominant voice in predictive analysis pertaining to healthcare.

Dataset Source :

This report makes use of the very popular diabetes dataset modified from the **NIDDK database**. The dataset involves a total of 768 records of female patients who each have eight associated health attributes and the dependent variable indicates whether the patient is diabetic (1) or not (0). These attributes are the crucial medical indicators: plasma glucose concentration, skinfold thickness, insulin level, BMI, and diabetes pedigree function which have high clinical significance so detailed analysis and predictive modeling can be developed.

The dataset is of significance in helping construct reliable classifiers for the prediction of diabetes. This foreknowledge would give the healthcare workers the ability to make accurate diagnosis timely which would in turn facilitate appropriate decisions. Apart from that the dataset provides a great opportunity to investigate interdependencies and correlations between different health indicators.

Dataset Attributes:

Pregnancies : Number of times the patient has been pregnant.

Glucose: Plasma glucose concentration after a 2-hour oral glucose tolerance test.

BloodPressure: Diastolic blood pressure (mm Hg).

SkinThickness: Triceps skinfold thickness (mm).

Insulin: 2-hour serum insulin (mu U/ml).

BMI: Body mass index (weight in kg/(height in m)²).

DiabetesPedigreeFunction: A function that represents the patient's diabetes pedigree (i.e., likelihood of diabetes based on family history).

Age: Age of the patient (years).

Outcome: Binary outcome (0 or 1) where 1 indicates the presence of diabetes and 0 indicates the absence.

[2] Data Processing and Analysis Steps

Steps For Data Processing and Analysis

Data Collection

This dataset has been obtained from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and contains diagnostic health attributes for 768 females. They aimed to collect data that could serve as key medical indicators to understand the reason for the existence of diabetes in patients, and these include, glucose levels, body mass index (BMI), insulin levels, among others. The dataset has been arranged in such a way that there are seven features and one binary outcome, which denotes the diabetic status of the patient, that is, whether a patient is diabetic (1) or not (0).

Data cleaning :

Handling missing or erroneous data was a critical step to ensure the dataset's reliability for analysis.

The following steps were performed:

- Removing Anomalies:** Anomalous values (e.g., zeros in columns like Glucose, BMI, Insulin, which are biologically unrealistic) were treated as missing values.

- Replacing Zeros with NA:** Zeros in the following key columns were replaced with NA: Glucose, BloodPressure, SkinThickness, Insulin, and BMI.

This ensured that these missing values were appropriately flagged for further handling.

-**Counting NA Values:** The number of NA values in each column was calculated to decide on the appropriate treatment strategy:

-**Columns with a Small Number of Missing Values:** For columns where the count of NA values was small (e.g., Glucose, BMI), the rows containing these NA values were removed.

-**Columns with a Large Number of Missing Values:** For columns with a high proportion of missing values (e.g., SkinThickness, Insulin), the NA values were replaced with the mean of the respective column to retain as much data as possible.

-**Outcome:** After handling missing values, the dataset was complete and ready for analysis with minimal information loss.

Data preprocessing: To prepare the data for analysis, the following preprocessing steps were performed:

-**Feature Engineering:** A new categorical variable, GlucoseGroup, was created to classify glucose levels as "High Glucose" or "Low Glucose" based on the median value of Glucose.

-**Categorical Data Handling:** The target variable, Outcome, was treated as a binary categorical variable (0 = non-diabetic, 1 = diabetic) for hypothesis testing and visualizations.

-**Exploratory Data Analysis (EDA):**

- Statistical summaries (e.g., mean, median, standard deviation) were computed for all variables.
- Correlation matrices were generated to identify relationships among variables.
- Visualizations such as scatter plots, box plots, and heatmaps were created to identify trends, patterns, and anomalies.

Overview of the Analytical Approach:

The project followed a structured approach to analyze the diabetes dataset. First, the data was cleaned by addressing missing values and filtering out incomplete rows to ensure accuracy. Exploratory Data Analysis (EDA) was conducted to identify trends and differences between diabetic and non-diabetic groups using descriptive statistics and visualizations like box plots and histograms. Hypothesis testing was performed to validate claims, such as differences in glucose levels, using t-tests. Relationships between key variables, such as BMI and glucose, were examined through correlation analysis and regression. Confidence interval simulations were conducted with varying sample sizes to observe precision and accuracy. Results were documented in a technical notebook and summarized in a report with visual insights. Finally, findings were presented with slides highlighting key patterns and actionable conclusions, ensuring clarity for diverse audiences.

Summary Statistics :The dataset was initially explored using summary statistics in order to appreciate the number of attributes that had been included in the total data set.

Summary findings :It is important to note that some attributes, for instance Glucose, BloodPressure, SkinThickness, Insulin, and BodyMass, contain zero values that are viewed as 'not applicable' if missing data.

[3] Challenges, Limitations, and Assumptions

- **Handling Missing Data:**

Challenge : Several attributes, such as **SkinThickness** (227 missing values) and **Insulin** (374 missing values), had significant missing data, represented as zeros.

Solution

- Replaced zeros with **NA** to accurately identify missing data.
- Imputed missing values using column means to retain dataset completeness while minimizing bias.

- **Potential Outliers**

- Outliers in attributes like BMI and Age were examined to ensure they did not unduly influence model performance

Limitations:

-Small Sample Sizes in Simulations: Confidence interval analysis with small sample sizes may introduce variability that does not reflect broader trends.

Assumptions:

1. Missing data imputation methods, such as mean replacement, were assumed to reasonably approximate actual values.
2. The relationships observed in the dataset (e.g., between glucose levels and diabetes) were assumed to hold true for the population represented by the dataset.
3. Statistical methods, such as t-tests and correlation analysis, were assumed to be appropriate given the data's characteristics.

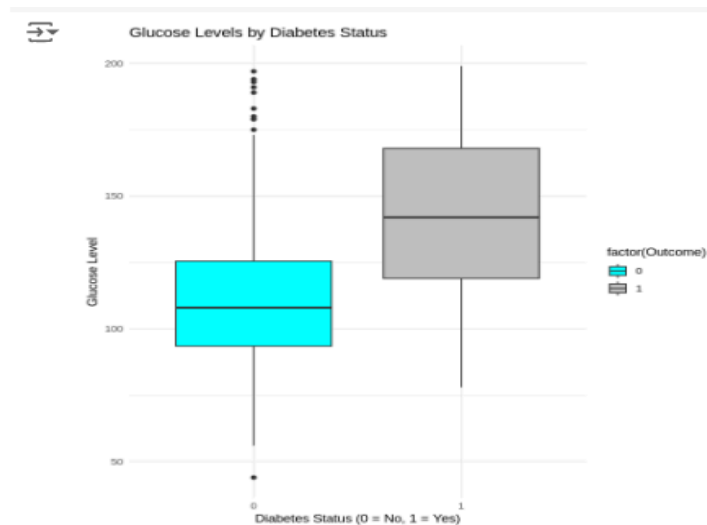
p-value

≥ 0.1	— No evidence against null hypothesis
0.09	
0.08	— Weak evidence against null hypothesis
0.07	
0.06	
0.05	
0.04	— Moderate evidence against null hypothesis
0.03	
0.02	
0.01	
0.009	
0.008	
0.007	
0.006	— Strong evidence against null hypothesis
0.005	
0.004	
0.003	
0.002	
≤ 0.001	— Very strong evidence against null hypothesis

[4] Results and Visualizations

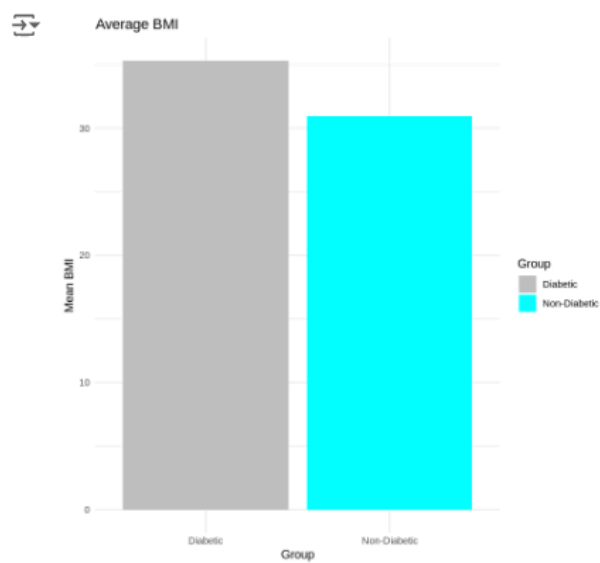
Visual [1]

The boxplot shows that diabetic patients (Outcome = 1) have higher glucose levels than non-diabetic patients (Outcome = 0), with clear group separation.



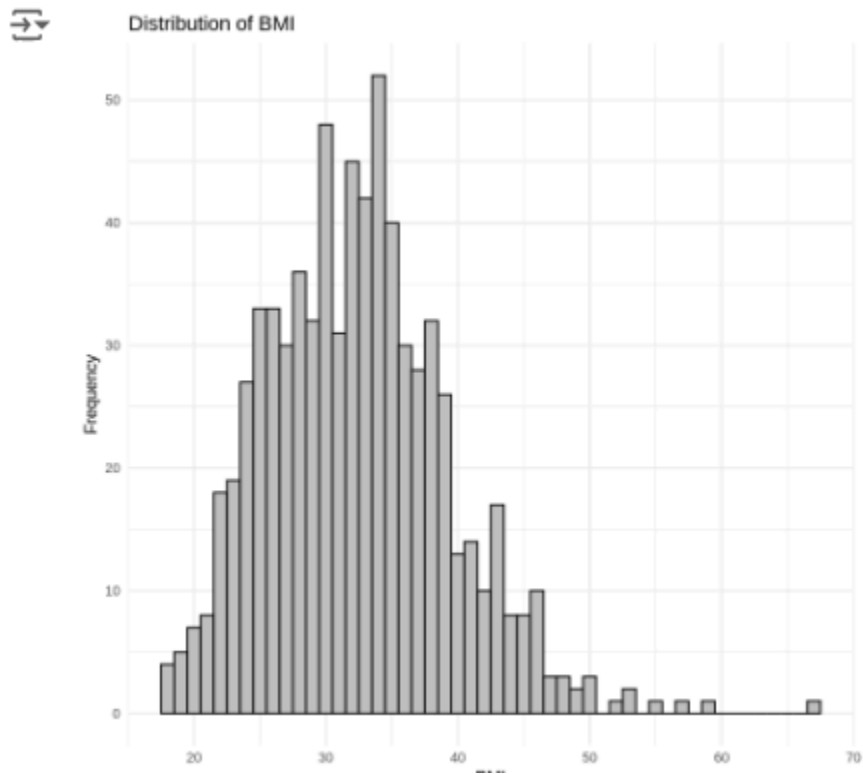
[2] Visual 2

The bar plot indicates that diabetic patients have a higher average BMI compared to non-diabetic patients, highlighting BMI as a potential risk factor for diabetes.



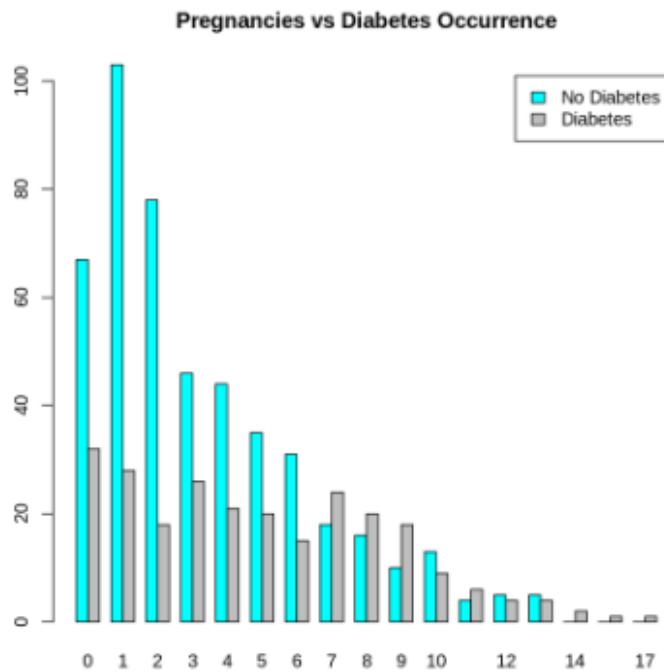
[3] Visual 3

The histogram illustrates the distribution of BMI values across all patients. The data shows a right-skewed pattern, with most BMI values concentrated between 20 and 40, indicating a higher prevalence of patients with normal to overweight BMI ranges.




[4] visual 4

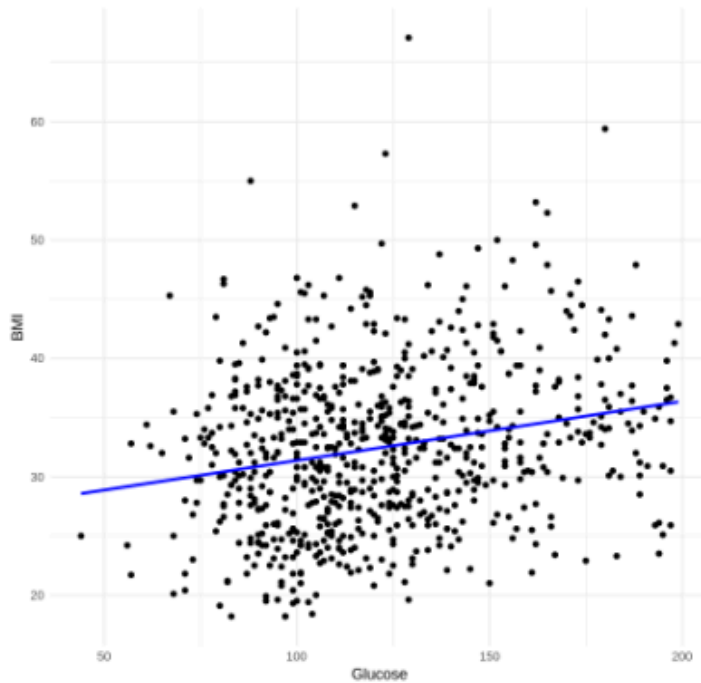
The bar plot reveals a trend where the occurrence of diabetes (gray bars) increases with the number of pregnancies, especially for higher pregnancy counts. This suggests a potential correlation between the number of pregnancies and diabetes prevalence.



[5] visual 5

The scatter plot with a trendline shows a positive correlation between glucose levels and BMI. This suggests that as glucose levels increase, BMI tends to increase slightly, indicating a potential relationship between these two factors.

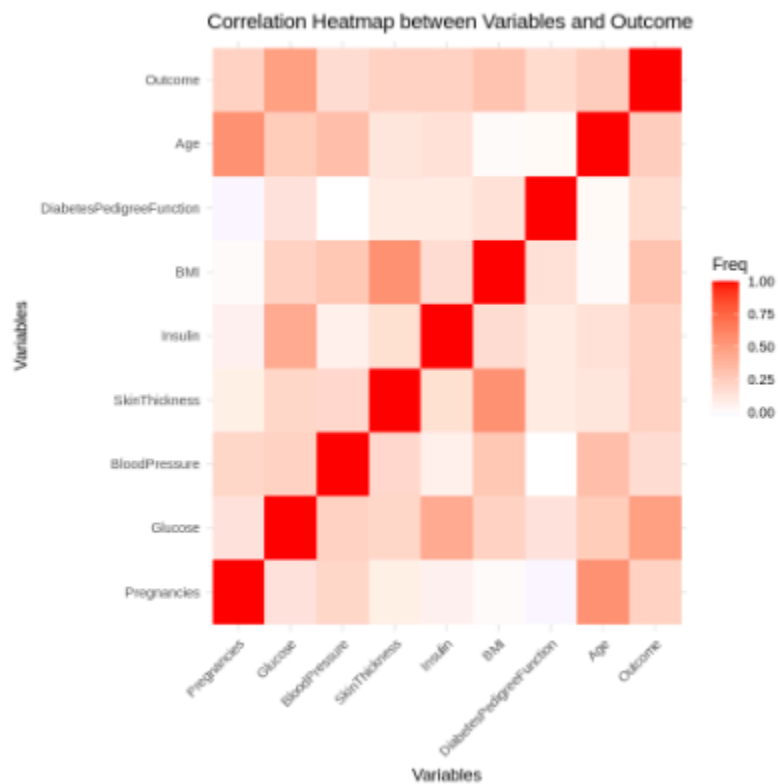
```
 `geom_smooth()` using formula = 'y ~ x'  
Correlation Between Glucose and BMI with Trendline
```



[6] Visual 6

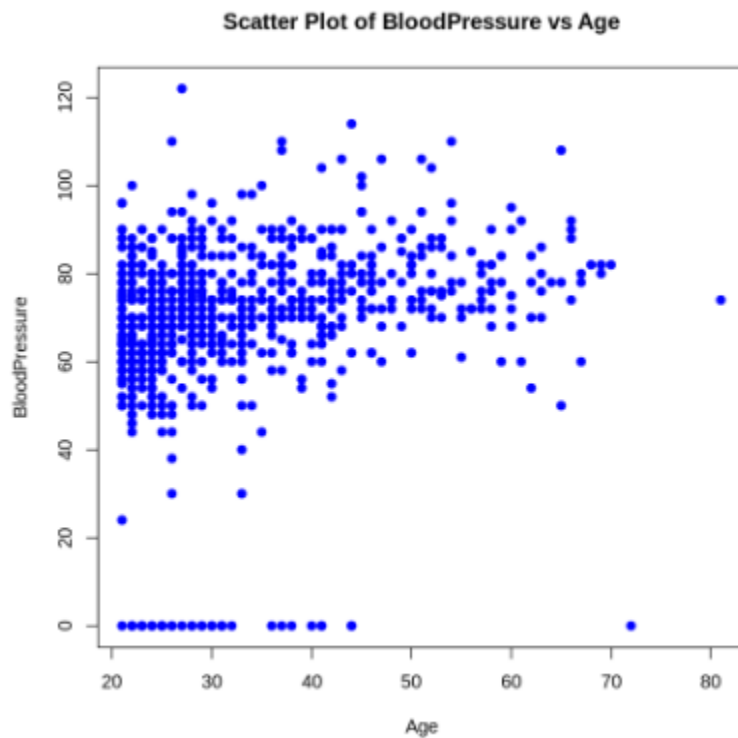
"Glucose" and "BMI" are the strongest predictors of diabetes (moderate correlation).

"BloodPressure" and "SkinThickness" show little to no correlation with diabetes. The heatmap effectively summarizes the relationships, making it easy to identify key predictors visually.



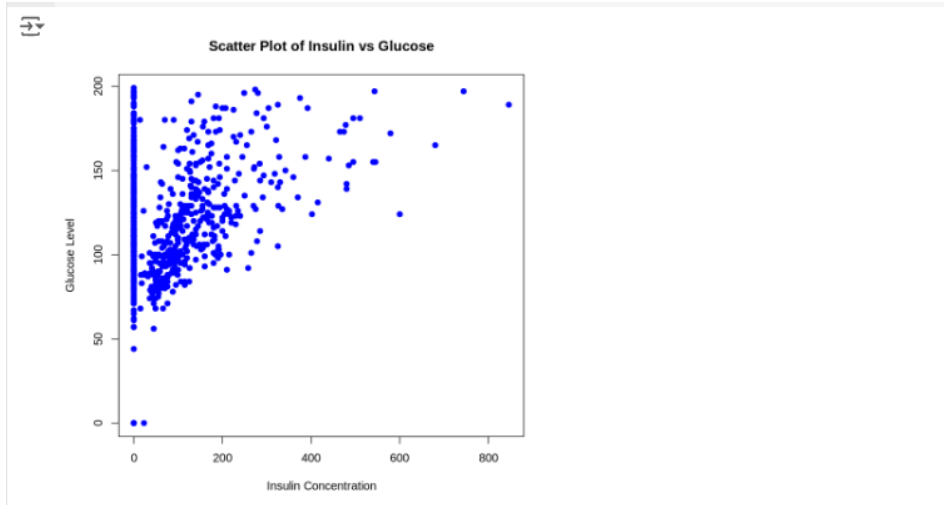
[7] visual 7

The Pearson correlation between Age and BloodPressure is 0.2395, indicating a weak positive relationship



[8] visual 8

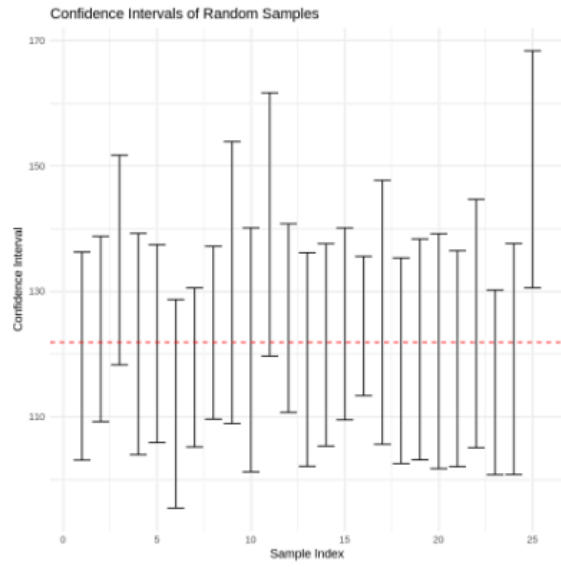
The Pearson correlation between Insulin and Glucose is 0.3314, indicating a moderate positive relationship.



[9] Visual 9

red dashed line ---> shows the true population mean.

The CIs vary in their width because each sample may have different variability and means.



[10] Visuals 10

Coverage proportion: Generally around 95% but may vary slightly due to randomness or non-normality in the data.

Interval width: Decreases with increasing sample size, reflecting greater precision.



Coverage Proportion:

n = 10: 1

n = 15: 0.96

n = 100: 0.88

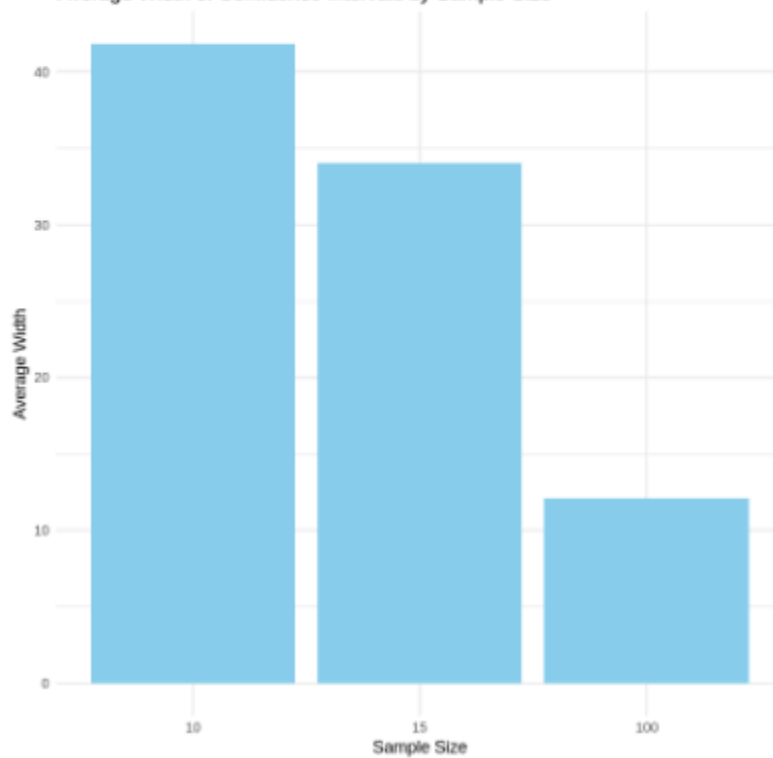
Average Width of Confidence Intervals:

n = 10: 41.82203

n = 15: 33.99821

n = 100: 12.09743

Average Width of Confidence Intervals by Sample Size



[5] Conclusion

This analysis provided valuable insights into the factors contributing to diabetes and the relationships among key health metrics. Through exploratory data analysis, significant trends were identified, such as the higher glucose levels and BMI observed in diabetic patients compared to non-diabetic ones. Hypothesis testing confirmed these differences, highlighting the critical role of glucose levels in predicting diabetes. Furthermore, the study revealed correlations between health metrics, such as the positive relationship between BMI and glucose levels, emphasizing the interplay of various factors in diabetes risk.

The simulation task demonstrated the impact of sample size on the precision and reliability of confidence intervals, underlining the importance of using adequate sample sizes for robust statistical inference. Despite challenges such as demographic limitations and handling missing data, the analysis remained thorough and systematic, offering a solid foundation for understanding the dataset.

Future research could address the dataset's limitations by incorporating more diverse demographics and exploring additional variables that may influence diabetes onset. Advanced machine learning models could also enhance predictive accuracy and uncover complex patterns within the data. This study serves as a stepping stone for more in-depth investigations, contributing to better understanding and management of diabetes risks.

Presentation Link :

https://www.canva.com/design/DAGahfqkXoM/iNfDFBuFVXyK8oww_SpqqA/edit

