

Technical Report

Introduction

The diabetes dataset is central to understanding the relationships between various health metrics and diabetes risk. This dataset includes key health indicators such as glucose, BMI, age, insulin, blood pressure, and pregnancy history. It helps evaluate how these factors contribute to diabetes occurrence.

The primary objective of this analysis is to explore how these health metrics correlate with diabetes outcomes, identify risk profiles, and understand the factors that could influence diabetes risk.

Data Processing and Analysis Steps

The dataset was sourced from a public database and contains key health metrics of patients, labeled as diabetic or non-diabetic. Data cleaning included handling missing values, removing rows with zero values where necessary, and ensuring the integrity of the variables for analysis. The dataset was then explored using statistical methods and visualizations to identify trends and patterns.

The analysis focused on relationships between health metrics (e.g., glucose, BMI) and the presence of diabetes. We used visual tools like bar plots, scatter plots, and box plots to uncover correlations and patterns.

Challenges, Limitations, and Assumptions

Challenges included handling missing values and rows with zero values, which could distort analysis. These were addressed through imputation or removal. One limitation is that the dataset only includes a limited number of health metrics, which may not fully represent all factors influencing diabetes. Assumptions were made regarding the linearity of relationships between variables and the exclusion of potential confounders not included in the dataset.

Results and Visualizations

Diabetes Rate

- A bar plot shows the rate of diabetes among patients, with a noticeable skew toward non-diabetic individuals.
- The analysis indicates that diabetes is more prevalent in individuals with higher glucose levels, higher BMI, and older age.

Average Glucose, BMI, and Age

- The average glucose level is higher among diabetic patients, as shown in bar plots comparing diabetic and non-diabetic groups. This supports the well-known association between high blood glucose and diabetes.

- BMI shows higher values in diabetic patients, reinforcing the link between obesity and diabetes risk.
- Older patients tend to have higher glucose and insulin levels, as indicated by scatter plots and line graphs.

Pregnancies and Diabetes

- Patients with a higher number of pregnancies tend to have a higher likelihood of being diabetic, as visualized through boxplots.

Blood Pressure and Age

- Blood pressure increases with age for both diabetic and non-diabetic individuals, with a stronger trend observed in diabetic patients.

Correlations and Risk Profiles

- Diabetic patients generally show higher glucose, insulin, and BMI levels compared to non-diabetic patients. This suggests that managing these factors could help mitigate diabetes risk.
- Risk profiles for diabetic patients highlight the role of high glucose, BMI, and insulin as key risk factors.

Conclusion

The findings suggest that elevated glucose levels, higher BMI, and older age are significant factors contributing to diabetes risk. The analysis highlights the importance of managing these health metrics to prevent or manage diabetes. Future research could explore other potential risk factors, improve data quality, and examine causal relationships using more sophisticated models.

Future Directions

Future analysis could include more extensive datasets that account for additional factors such as lifestyle, diet, and genetic predisposition. Further studies might also incorporate machine learning models to predict diabetes outcomes more accurately.