Diabetes Analysis Report

Introduction

For this project I have chosen the Diabetes dataset to perform analysis. This dataset is restricted to women of at least 21 years of age, from the Pima Indian heritage. The dataset contains information about the Glucose levels, Skin thickness, Pregnancies, Blood Pressure, Insulin, BMI, Diabetes Pedigree Function, Age, as well as the outcome stating whether the person has diabetes. To support the analysis, I have implemented two machine learning algorithms to confirm the influential factors for diabetes.

Data

The dataset used in this analysis is sourced from [1]. It contains health-related data from 768 Pima Indian women, all aged 21 or older. The dataset includes the following features:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

Analysis

Exploring the dataset

The dataset was explored to assess the distribution of the features and to identify any missing values that could impact the analysis. Additionally, the dataset was transformed to include a new "Age Group" column, categorizing individuals into age ranges to dwell deeper into how age-related factors influence diabetes.

Visualization and Analysis

The data was further visualized to uncover key insights. A heatmap of feature correlations revealed that glucose levels, BMI, diabetes pedigree function, Pregnancies and age had a stronger influence on the outcome compared to other variables. These findings guided the focus of the analysis.

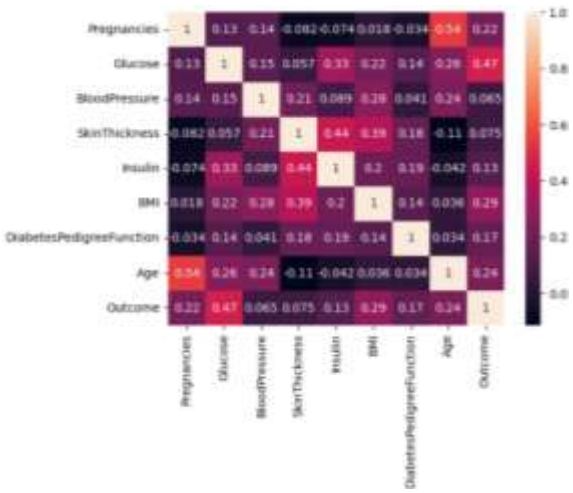

Fig 1: Correlation between the variables

The distribution of the most correlated variables—Glucose, BMI, Age, and Diabetes Pedigree Function was analyzed based on the diabetes outcome. Results showed that individuals with diabetes had higher averages in these variables compared to those without diabetes. This trend was further explored across different age groups, consistently confirming that the diabetes-positive cases exhibited higher values for Glucose, BMI, and Diabetes Pedigree Function in each age group.

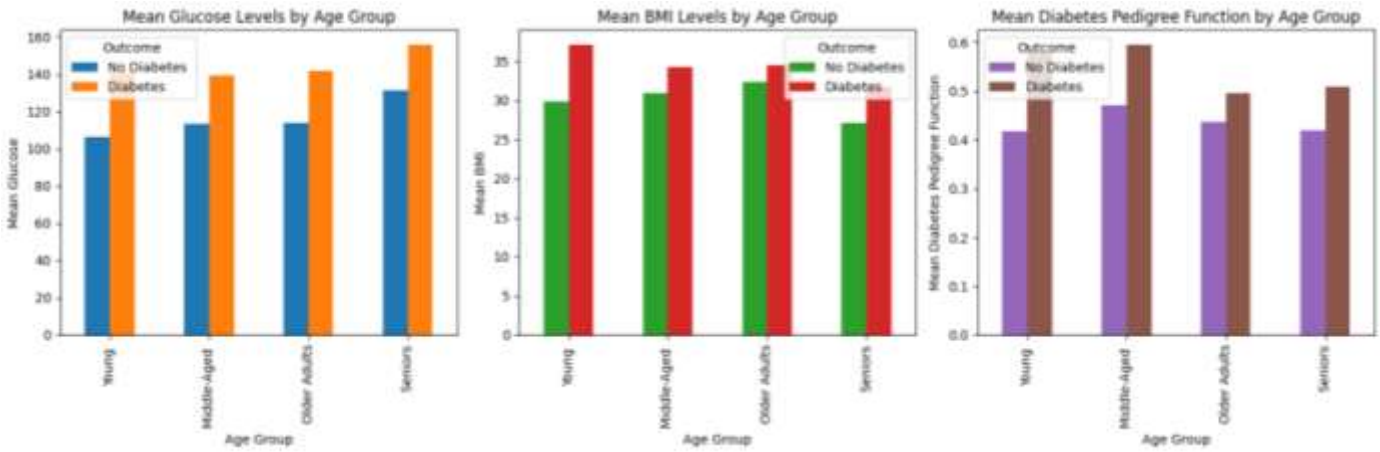

Fig 2: Mean Glucose, BMI, Diabetes Pedigree Function among different age groups

However, when analysing pregnancies within each age group, this variable did not follow the same trend observed with Glucose, BMI, and Diabetes Pedigree Function. But, when examined by the number of pregnancies rather than age, a clear pattern emerged, individuals with a higher number of pregnancies showed a greater likelihood of developing diabetes.

Validation Using Machine Learning

Two machine learning algorithms, Logistic Regression and AdaBoost, were applied to validate the findings. The independent variables used in both models were Glucose, BMI, Age, Pregnancies, and Diabetes Pedigree Function.

For the Logistic Regression model, the coefficients for these variables were [0.03287071, 0.07967181, 0.01228861, 0.11877822, 0.82152637], respectively. This suggests that the Diabetes Pedigree Function had the greatest influence on determining whether a person develops diabetes. The accuracy of the Logistic Regression model was 77.94%.

The AdaBoost algorithm, trained on the same variables, yielded a slightly higher accuracy of 79.81%.

Conclusion

In conclusion, the analysis of the key variables—Glucose, BMI, Age , and Diabetes Pedigree Function, revealed showed higher averages for those diagnosed with the condition. Pregnancies, however was the only variable that defied the trend further investigation by the number of pregnancies showed a higher likelihood of diabetes with increasing pregnancies.

The use of two machine learning models, Logistic Regression and AdaBoost, confirmed these findings. The Logistic Regression model indicated that Diabetes Pedigree Function had the greatest influence on predicting diabetes, with an accuracy of 77.94%. The AdaBoost algorithm achieved a slightly better performance, with an accuracy of 79.81%.

References

1. Diabetes Dataset: https://www.kaggle.com/datasets/mathchi/diabetes-data-set