Exploratory Data Analysis on the Diabetes Dataset

WRITTEN BY: Mogammad Hamit

Introduction

Diabetes is a growing global health concern, significantly impacting the quality of life of millions of individuals. Understanding the factors that influence the progression of diabetes is crucial for early detection, effective intervention, and improved patient outcomes. This exploratory analysis aims to identify key features that may contribute to, or correlate with, the progression of diabetes.The analysis is based on a standardised dataset, containing physiological and medical measurements, from individuals diagnosed with diabetes. The dataset comprises 442 records and 11 features, including:

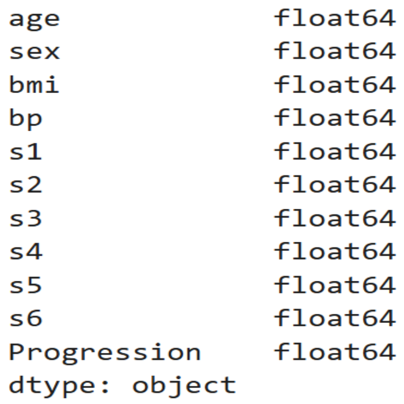
- Demographic and physical attributes: age, sex, bmi (body mass index), bp (blood pressure)
- Biochemical measurements: through s6 (various blood serum tests)

- Target variable: Progression (a quantitative measure of diabetes progression one year after baseline)

Through statistical summaries, visualizations, and correlation analysis, this study explores the relationships between these features and the target variable, offering insights into potential risk indicators for diabetes progression.

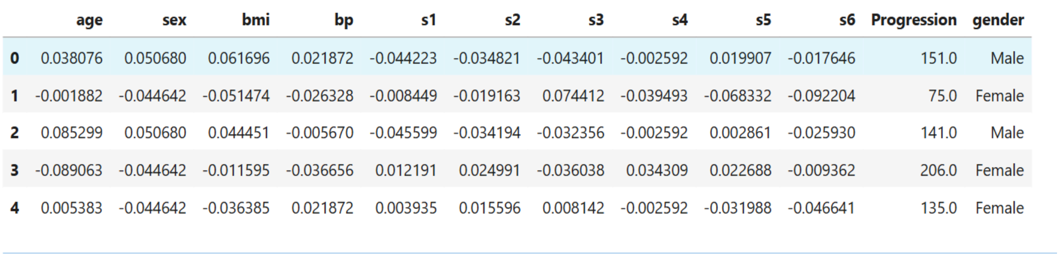
DATA CLEANING

All variables were checked for appropriate data types , each returned a type float which is expected from a standardised dataset. The “sex” feature was not easily interpretable in its original form . To improve this, a new categorical feature called “gender” was introduced , where the float type “sex” was converted to its gender and stored in the feature called “gender”



```
age            float64
sex            float64
bmi            float64
bp             float64
s1             float64
s2             float64
s3             float64
s4             float64
s5             float64
s6             float64
Progression    float64
dtype: object
```

Figure 1 : Table displaying data types of features.



	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	Progression	gender
0	0.038076	0.050680	0.061696	0.021872	-0.044223	-0.034821	-0.043401	-0.002592	0.019907	-0.017646	151.0	Male
1	-0.001882	-0.044642	-0.051474	-0.026328	-0.008449	-0.019163	0.074412	-0.039493	-0.068332	-0.092204	75.0	Female
2	0.085299	0.050680	0.044451	-0.005670	-0.045599	-0.034194	-0.032356	-0.002592	0.002861	-0.025930	141.0	Male
3	-0.089063	-0.044642	-0.011595	-0.036656	0.012191	0.024991	-0.036038	0.034309	0.022688	-0.009362	206.0	Female
4	0.005383	-0.044642	-0.036385	0.021872	0.003935	0.015596	0.008142	-0.002592	-0.031988	-0.046641	135.0	Female

Figure 2 : Table displaying the structure of the dataset after the gender feature was added.

MISSING DATA

The dataset was checked for missing values (NaN) and none were found . Duplicates were checked and no duplicates were detected.

DATA STORIES AND VISUALISATIONS

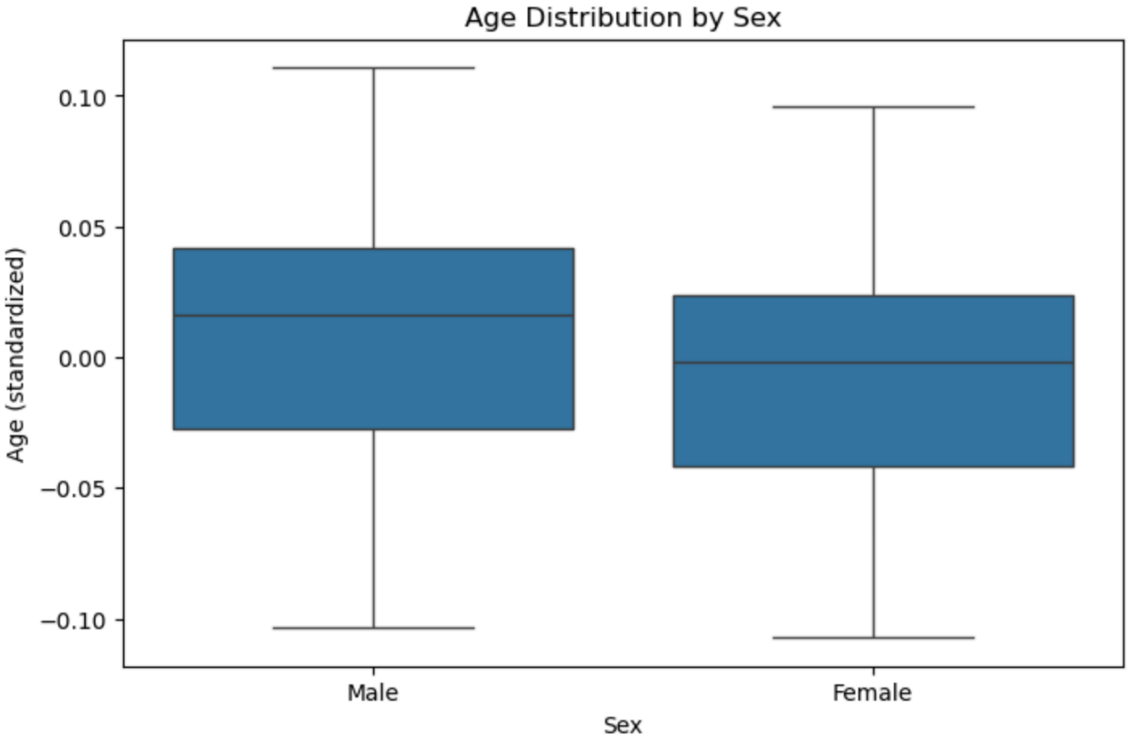


Figure 3 : Boxplot showing the age distribution of patients with diabetes, grouped by Sex.

From Figure 3, female patients tend to have standardised age values below the mean, which is centered at zero, while male patients tend to have values above the mean. This suggests that, on average, male patients in the dataset were older and female patients were younger.

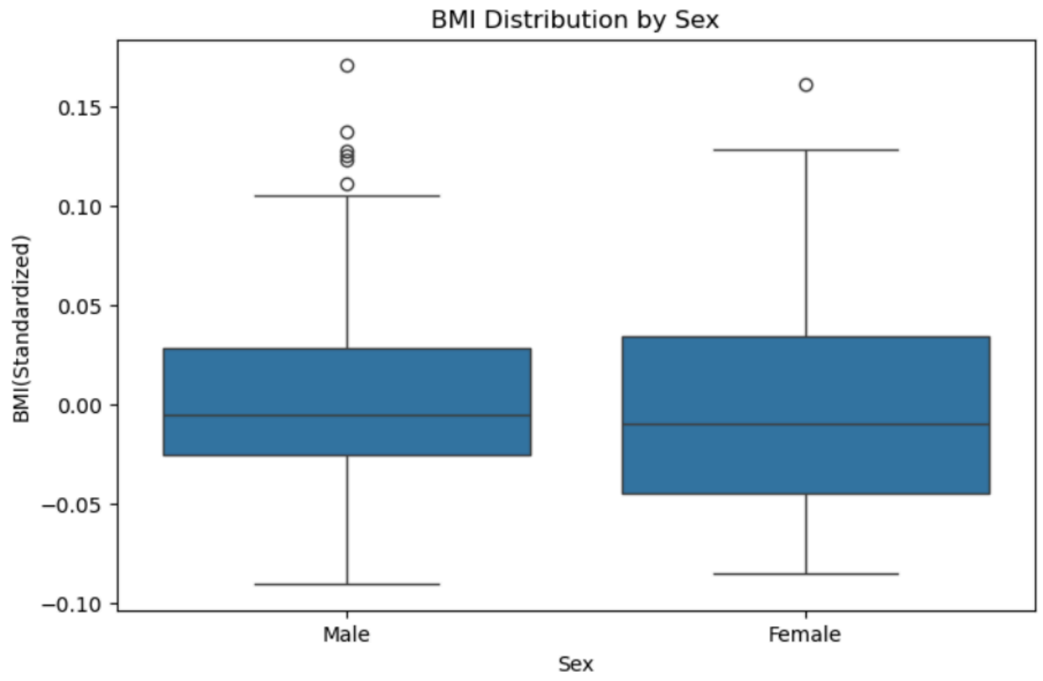


Figure 4 : Boxplot displaying the BMI distribution by sex.

In Figure 4, the median BMI for males appears slightly above zero, while the median for females is slightly below zero. Since the data is standardised, this suggests that, on average, male patients have a marginally higher BMI than female patients. However, the overlapping interquartile ranges indicate considerable similarity between the sexes, suggesting that male and female diabetic patients (in this dataset) have broadly comparable BMI profiles.

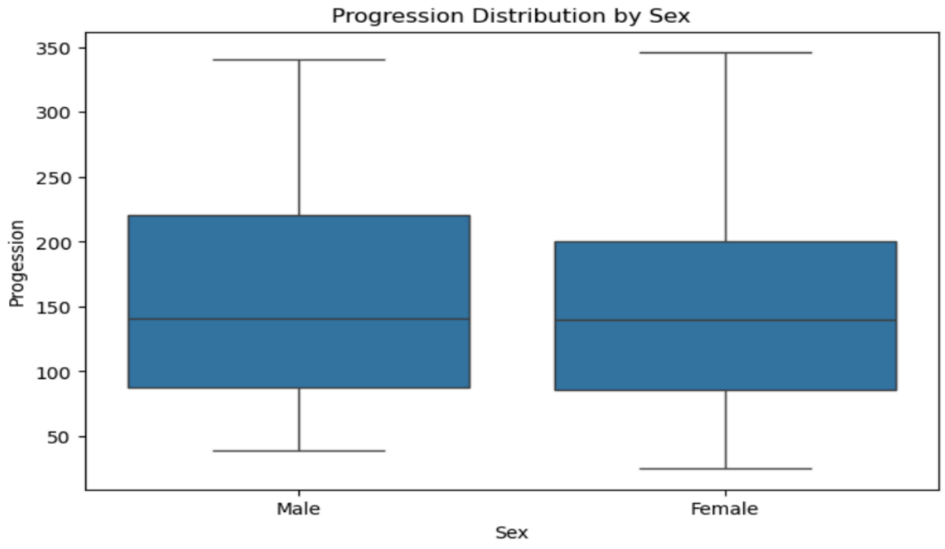


Figure 5 : Boxplot representing the Progression of diabetes by Sex.

Figure 5 boxplot shows that male and female patients have similar progression values. Both groups have similar medians, meaning the average progression is about the same. The spread of the data is also similar, although males show a slightly wider range. Overall, there doesn't appear to be a major difference in disease progression between the sexes in this dataset.

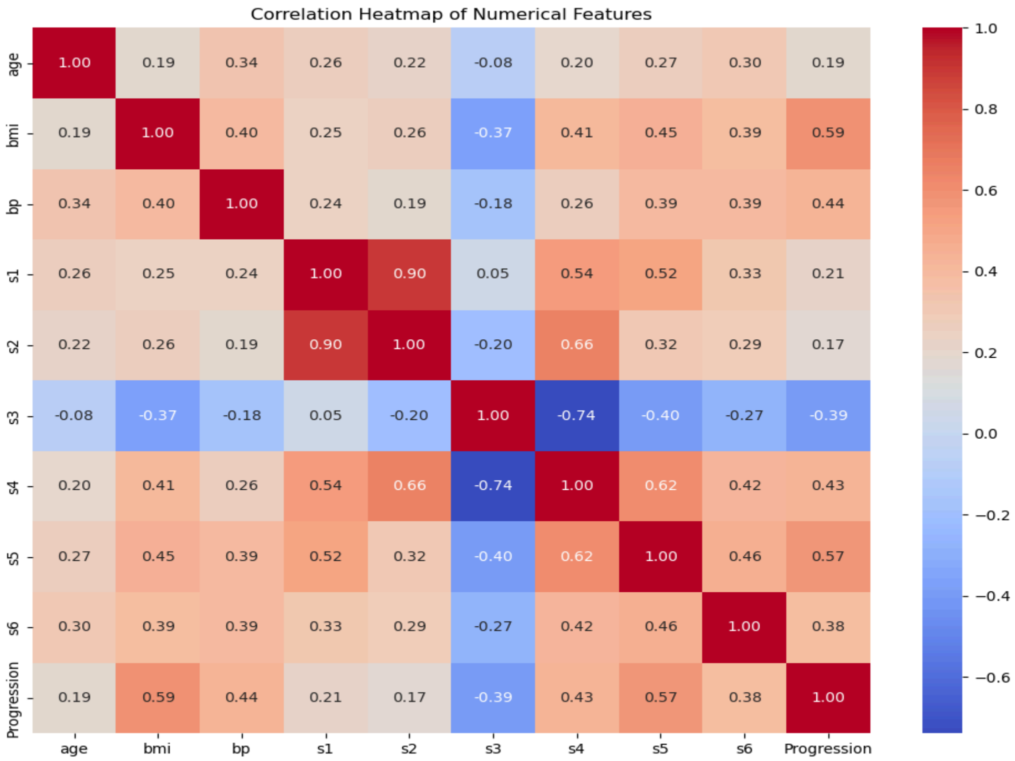


Figure 6 : Correlation heatmap of numerical values.

The correlation heatmap in figure 6 , displays the relationships between numerical features and their association with disease progression. Among the features, BMI shows the strongest positive correlation with progression (0.59), indicating that patients with higher BMI tend to experience more severe disease. Similarly, the s5 feature also shows a strong positive correlation (0.57), suggesting it may be a key biochemical marker of disease severity. Blood pressure (0.44) and s6 (0.38) also show moderate positive correlations. In contrast, s3 exhibits a moderate negative correlation (–0.39), implying that higher values may be linked to less severe progression. Additionally, some features are highly correlated with each other, such as s1 and s2 (0.90), and s3 and s4 (–0.74), which may indicate redundancy. These patterns can guide feature selection by highlighting which variables are most informative and which may contribute to multicollinearity in predictive models.

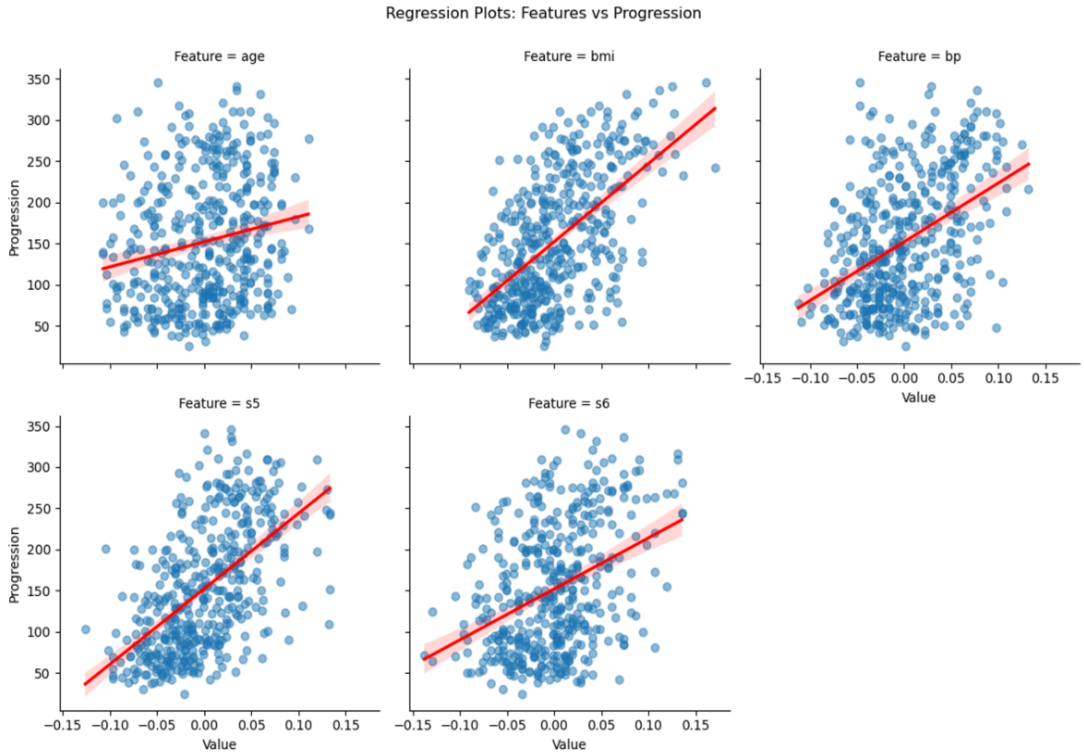


Figure 8 : Multi-panel regression plot, displaying the relationship between Features and diabetes progression

The regression plots visualize the linear relationship between selected features and disease progression. Among the variables, BMI and s5 display the strongest positive linear trends, with data points closely aligned along the regression line, reinforcing earlier findings from the correlation heatmap that these are key indicators of disease severity. Blood pressure (bp) and s6 also show a positive association with progression, but slightly more dispersed, suggesting moderate predictive value. Age presents a weaker positive relationship, with a flatter slope and more scattered data, indicating that while older patients may tend to have higher progression values, the relationship is less pronounced. These visual trends align well with the correlation heatmap, where BMI and s5 had the highest correlation with progression (0.59 and 0.57, respectively), followed by bp and s6. This suggests that BMI, s5, bp, and s6 are the most relevant features for predicting diabetes progression in this dataset.

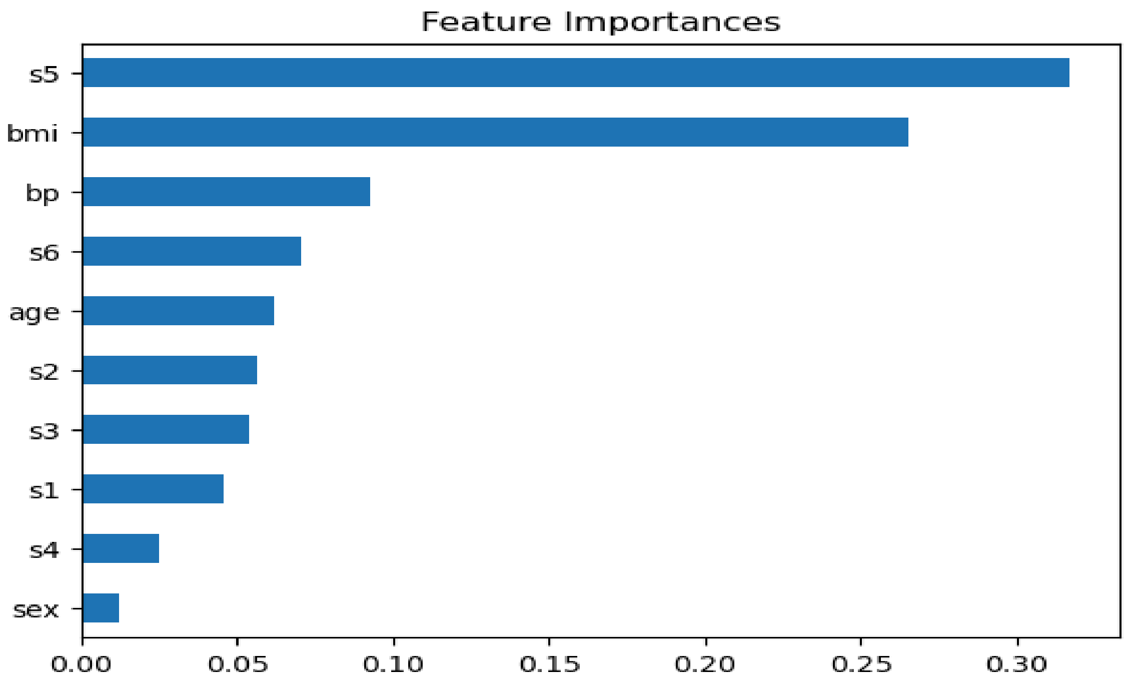


Figure 9 : Relative importance of features in predicting ProgessLine using Random Forest Regressor

In the feature importance analysis using a Random Forest Regressor, feature s5 emerged as the most influential predictor of diabetes progression, while sex showed the least importance, relative to the other variables. Notably, BMI was ranked second in importance, emphasising earlier findings from both the correlation heatmap and regression plots, where BMI demonstrated a strong positive association with disease progression (correlation coefficient r = 0.59). Although sex had minimal predictive value in the model, this is consistent with the boxplot analysis Figure 4, which revealed substantial overlap in diabetes progression scores between males and females, suggesting that sex alone is not a significant factor in disease progression. These findings collectively highlight that metabolic-related features such as s5, BMI, and blood pressure play a key role in predicting diabetes progression in this dataset, whereas demographic variables like sex and age are less predictive.

Although this dataset offers valuable insights into diabetes progression, there are a few limitations to consider. The biochemical variables labeled s1 through s6 lack clear definitions, which restricts the depth of biological interpretation. The absence of important health indicators ,such as body fat percentage reduces the ability to assess the full metabolic profile of patients. Including such variables could have improved the analysis by providing more context, and improved the predictive modeling of disease progression.