KIET Karachi Institute of Economics and Technology
CENTRE OF EXCELLENCE
25 Years & More...

DIABETES DATASET

EXPLORATORY DATA ANALYSIS, PREPROCESSING, AND LINEAR REGRESSION MODELING

NAME: MUHAMMAD WAQAS

DATE: 29-11-2025

Diabetes

Dataset: Exploratory Data Analysis, Preprocessing, and Linear Regression Modeling

INSTRUCTOR: SIR DR. MUHAMMAD ARIF HUSSAIN

NAME: MUHAMMAD WAQAS

ID: 65118

COURSE: DATA MINING

CLASS ID: 119243

DATE: 29-11-2025

3. Summary

This report details the Exploratory Data Analysis (EDA), preprocessing, and Linear Regression modeling performed on the Diabetes dataset. The project followed two distinct approaches: a comprehensive scikit-learn pipeline for the full dataset and a manual EDA for the first 10 rows. The full dataset EDA confirmed all features are numerical with no missing values. Histograms confirmed that most features are centrally distributed. Correlation analysis identified BMI (bmi), Blood Pressure (bp), and a blood serum measurement (s5) as having the strongest positive linear relationship with the target variable. A Linear Regression model was trained, achieving an R2 score of 0.453 and a Mean Squared Error (MSE) of 2900.19, indicating a moderate fit to the data. The manual EDA on the subset showed biased correlation results, confirming the necessity of a full and proper dataset analysis.

4. Table of Contents

5. Introduction

This project focuses on predicting a quantitative measure of diabetes progression after one year, based on ten baseline physiological variables (age, sex, body mass index, average blood pressure, and six blood serum measurements: s1-s6). The objective is to utilize the principles of Exploratory Data Analysis (EDA) to understand the dataset's structure, perform necessary preprocessing (scaling), and apply a foundational machine learning algorithm, Linear Regression, to model the relationship between the features and the target variable.

6. Methods

The project used two distinct approaches:

Method 1: Full Dataset Analysis using Scikit-learn

1. **Data Preparation:** The load_diabetes () dataset was converted into a pandas Data Frame. Initial checks confirmed no missing values and data types are suitable for numerical analysis.
2. **Preprocessing & Modeling:** The features were scaled using StandardScaler. The data was split into 80% training and 20% testing sets. A LinearRegression model was fitted to the scaled training data.

Method 2: Manual Exploratory Data Analysis (10 Rows)

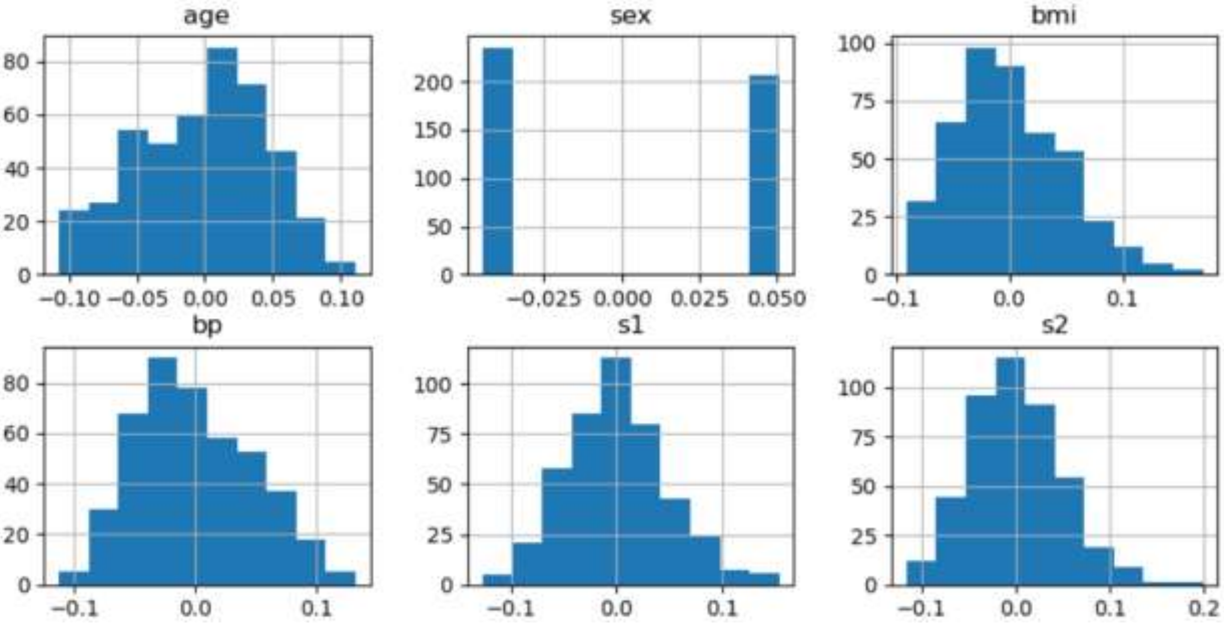
1. **Data Subset:** The first 10 rows of the dataset were isolated for manual inspection and visualization to demonstrate basic EDA concepts and check for data representation issues.
2. **Analysis:** Descriptive statistics, unique value counts, correlation matrix, and histograms were manually calculated or plotted for this small subset.

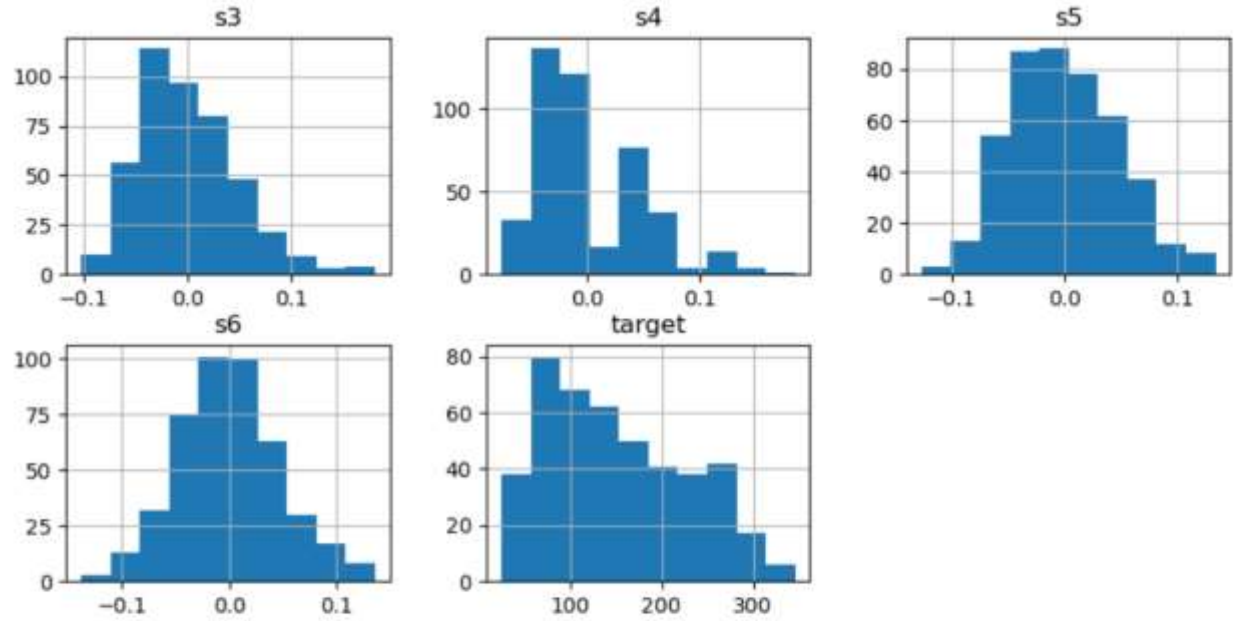
7. Discussion

7.1. EDA and Preprocessing (Scikit-learn)

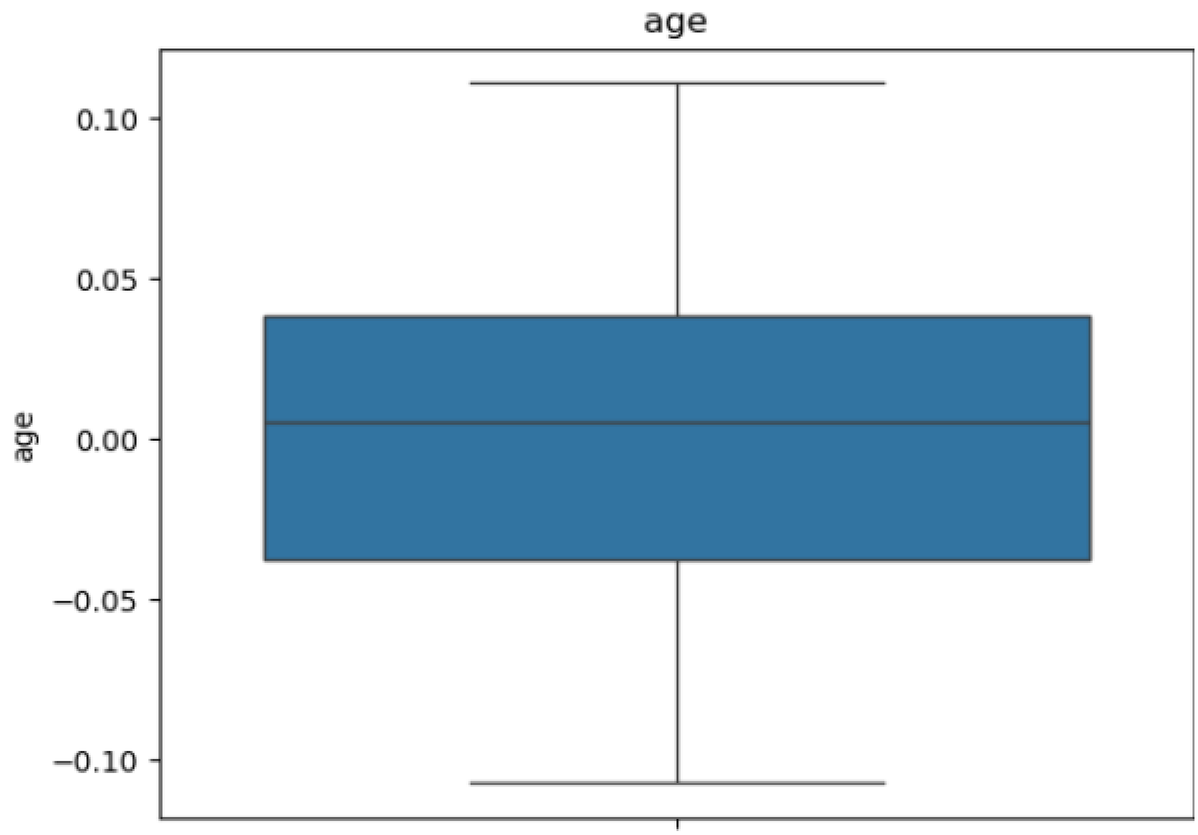
The initial exploration of the full dataset (442 entries, 11 columns) revealed the following:

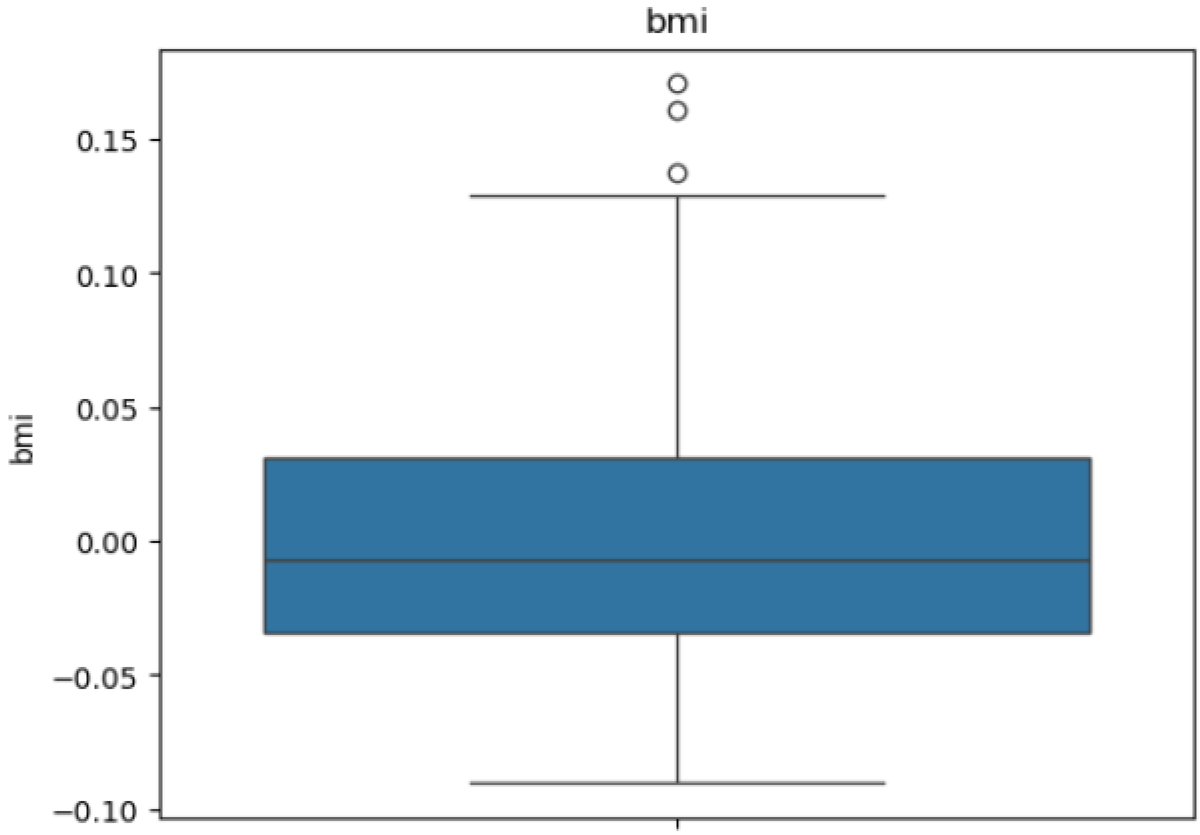
- Feature Distribution (Histograms):

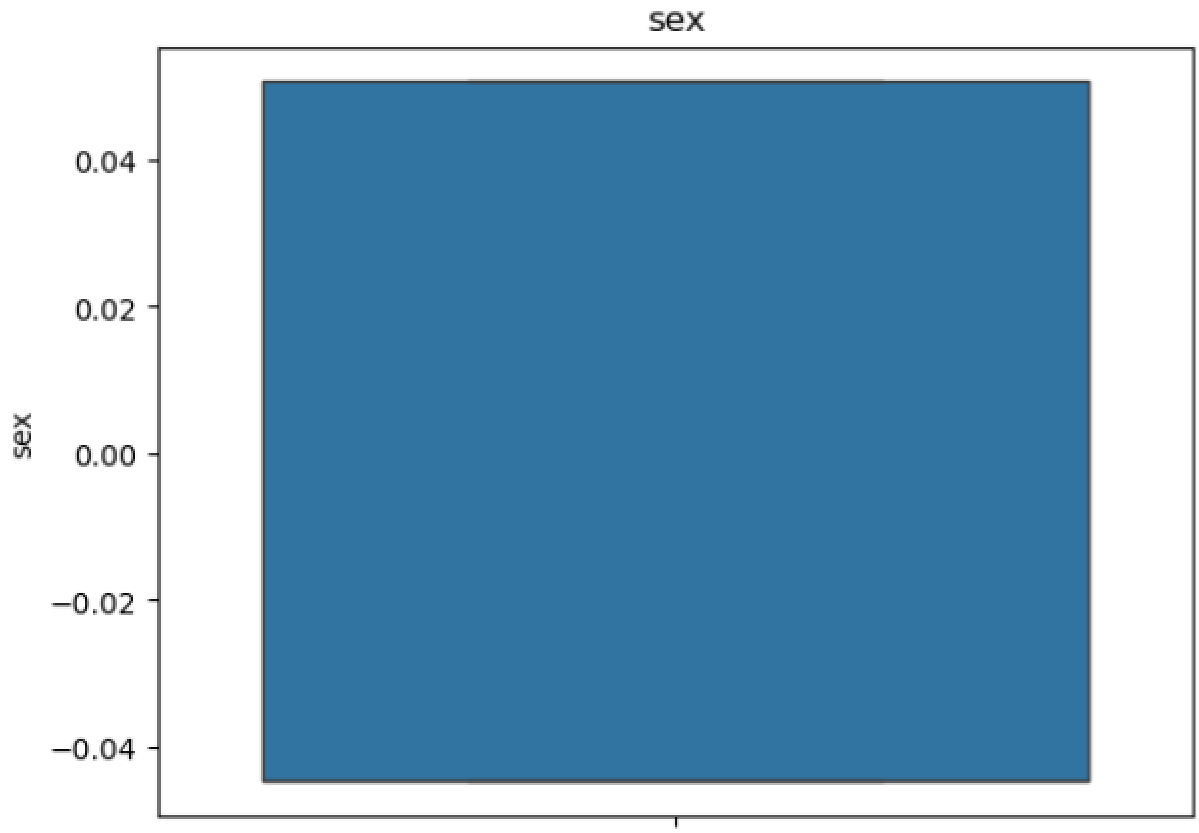


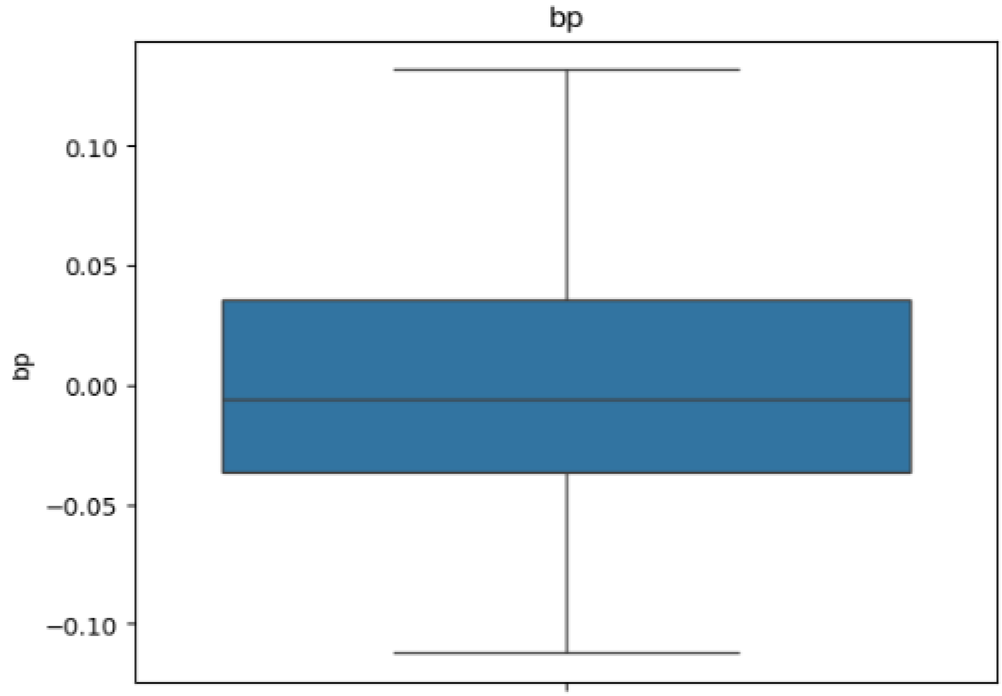


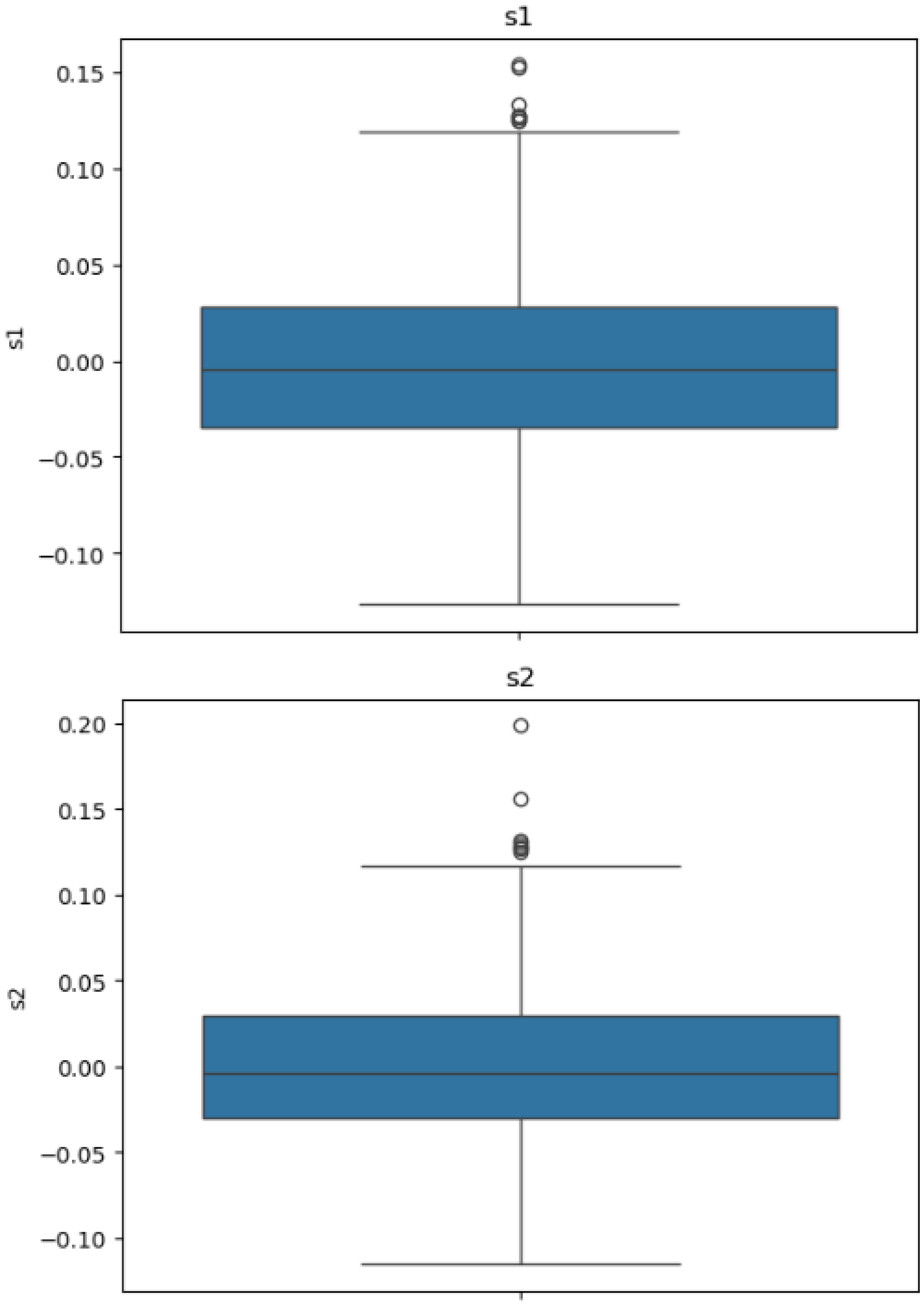
- **Explanation of Graph:** This visualization, comprising multiple histograms, shows the **distribution (spread)** of each individual feature in the dataset. Since the features are pre-normalized, they are centered around zero. The shape of most distributions is somewhat **bell-shaped (Gaussian/Normal)**, which is generally desirable for linear models, indicating that most data points are concentrated near the mean. Observing these distributions is crucial as it helps confirm the data's suitability for linear modeling after scaling, and identifies potential outliers or skewness that might impact the model's performance.
- Outlier Detection (Box Plots):


age
age
0.10
0.05
0.00
−0.05
−0.10

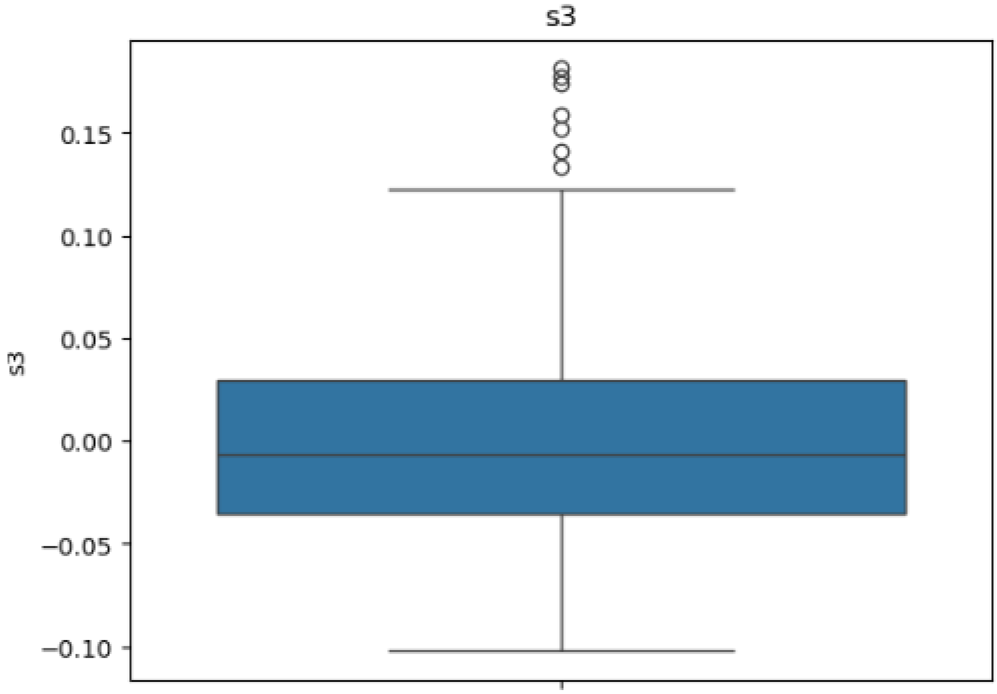

bmi
bmi
0.15
0.10
0.05
0.00
−0.05
−0.10

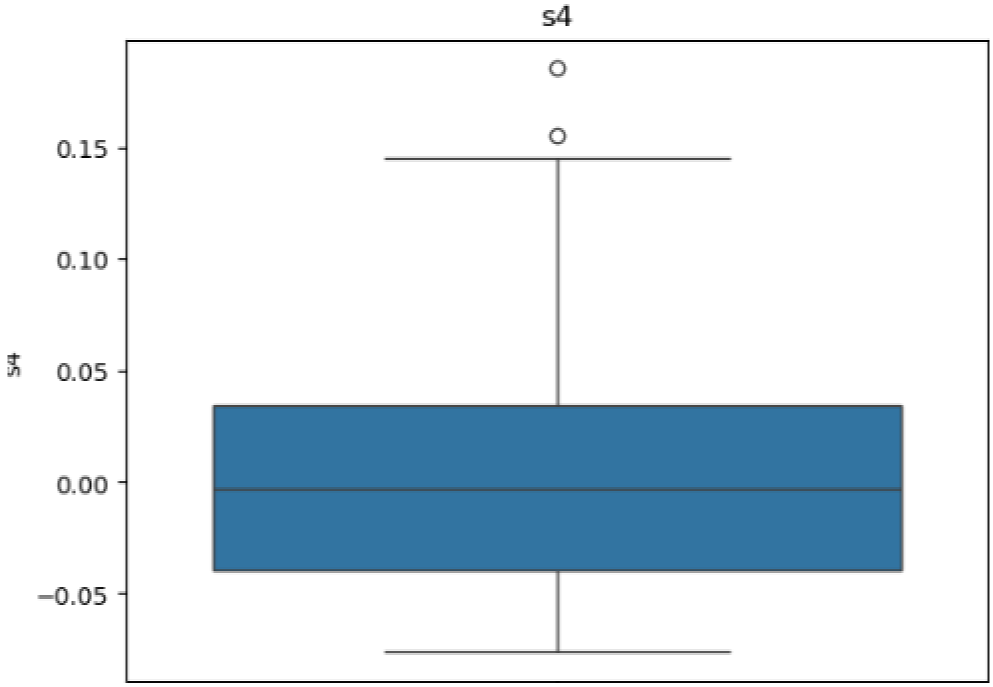

sex
0.04
0.02
0.00
−0.02
−0.04
sex

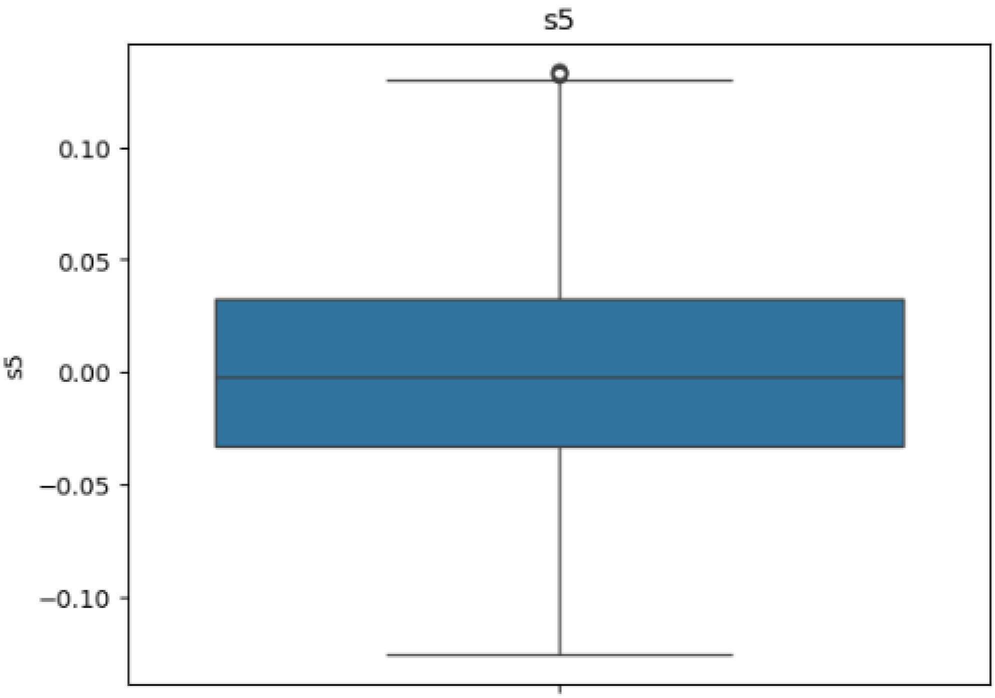

bp
0.10
0.05
0.00
−0.05
−0.10
bp

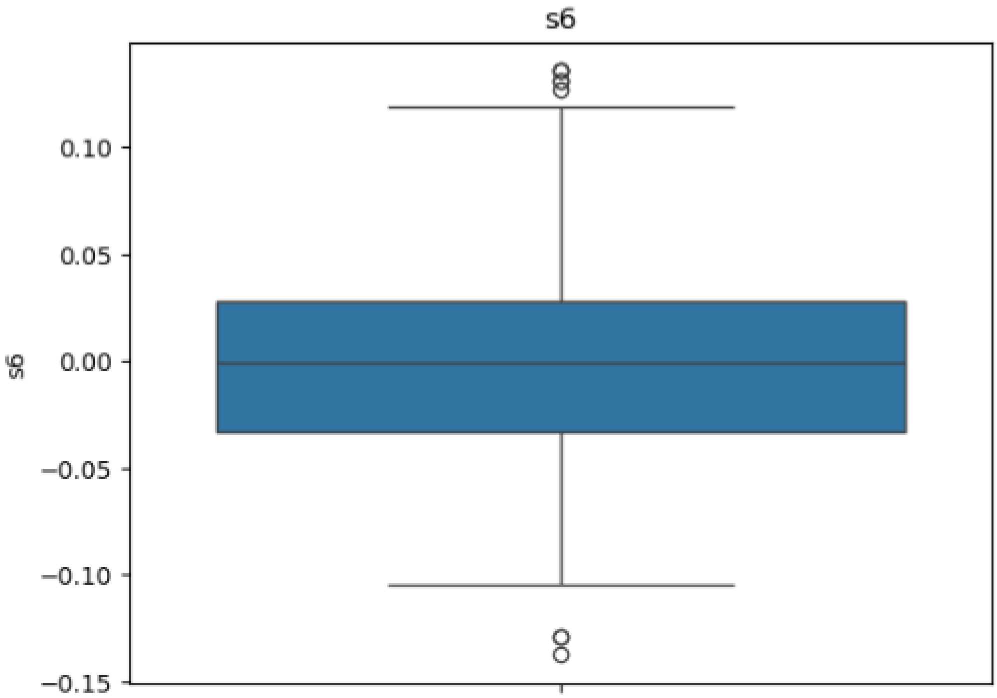


s1
0.15
0.10
0.05
0.00
−0.05
−0.10
s1
s2
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
s2


s3
0.15
0.10
0.05
0.00
−0.05
−0.10
s3

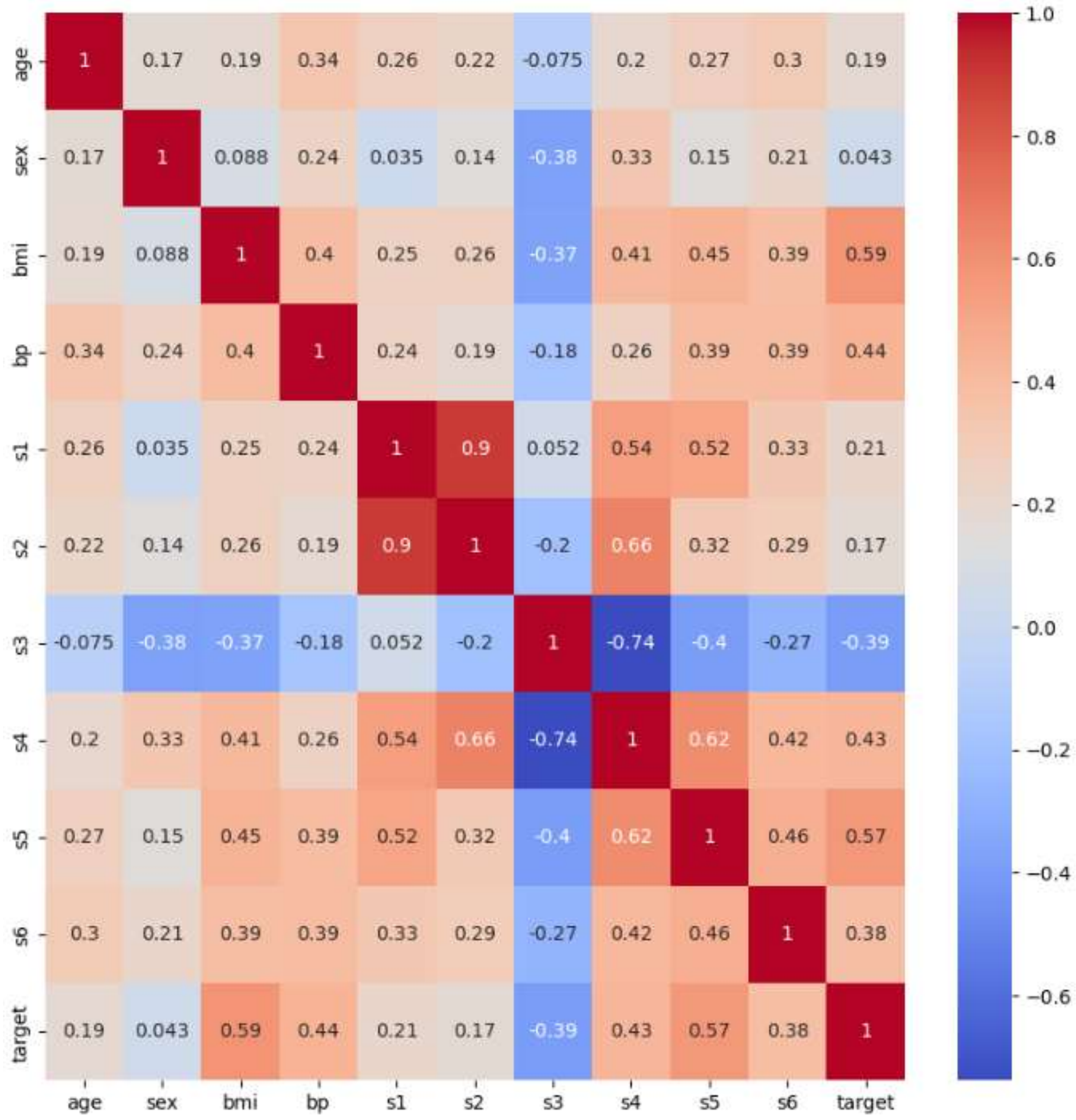

s4
0.15
0.10
0.05
0.00
−0.05
s4





7.2. Feature Selection and Visualization

A Correlation Heatmap was used to find the linear relationships between all features and the target variable.

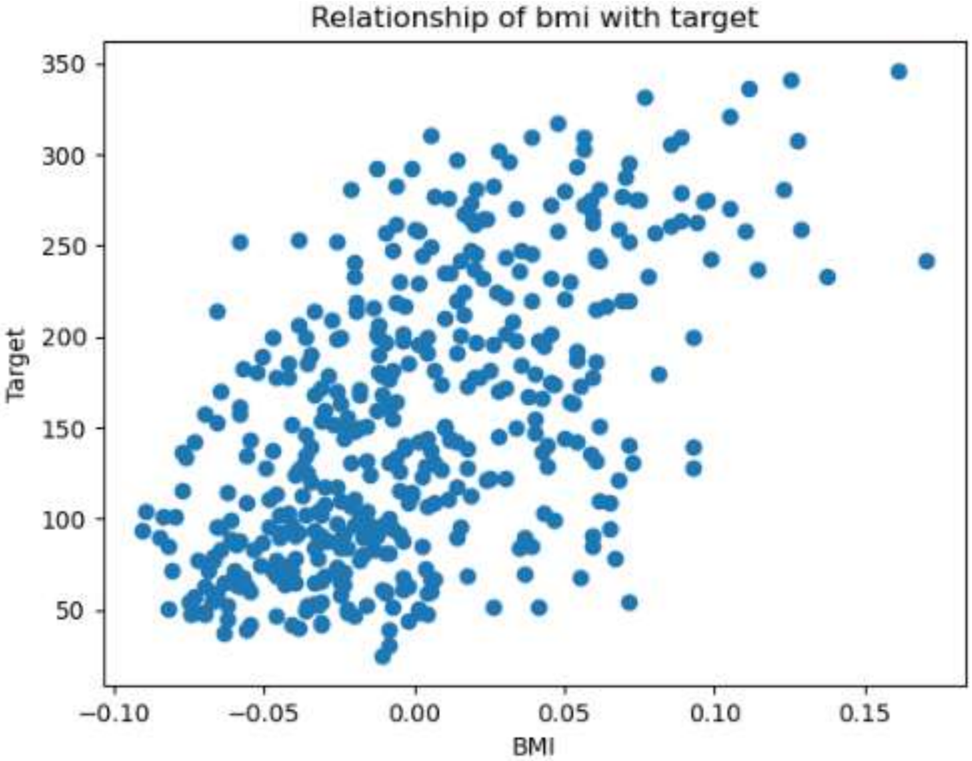


	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	target
age	1	0.17	0.19	0.34	0.26	0.22	-0.075	0.2	0.27	0.3	0.19
sex	0.17	1	0.088	0.24	0.035	0.14	-0.38	0.33	0.15	0.21	0.043
bmi	0.19	0.088	1	0.4	0.25	0.26	-0.37	0.41	0.45	0.39	0.59
bp	0.34	0.24	0.4	1	0.24	0.19	-0.18	0.26	0.39	0.39	0.44
s1	0.26	0.035	0.25	0.24	1	0.9	0.052	0.54	0.52	0.33	0.21
s2	0.22	0.14	0.26	0.19	0.9	1	-0.2	0.66	0.32	0.29	0.17
s3	-0.075	-0.38	-0.37	-0.18	0.052	-0.2	1	-0.74	-0.4	-0.27	-0.39
s4	0.2	0.33	0.41	0.26	0.54	0.66	-0.74	1	0.62	0.42	0.43
s5	0.27	0.15	0.45	0.39	0.52	0.32	-0.4	0.62	1	0.46	0.57
s6	0.3	0.21	0.39	0.39	0.33	0.29	-0.27	0.42	0.46	1	0.38
target	0.19	0.043	0.59	0.44	0.21	0.17	-0.39	0.43	0.57	0.38	1

- **Explanation of Graph:** This heatmap uses color intensity to visually represent the **Pearson correlation coefficient** between every pair of variables. The color and numerical values range from -1 to +1. Values close to +1 (darker shade) indicate a strong positive linear relationship, while values close to -1 indicate a strong negative linear relationship. By specifically focusing on the bottom row (or last column) labeled **target**, we can identify the features that have the strongest linear association with the diabetes progression score. The graph clearly shows strong **positive correlations** for **bmi (0.59)**, **s5 (0.56)**, and **bp (0.44)**. This visualization is the primary tool used for **feature selection**, confirming that these three are the most promising predictors for the linear model.

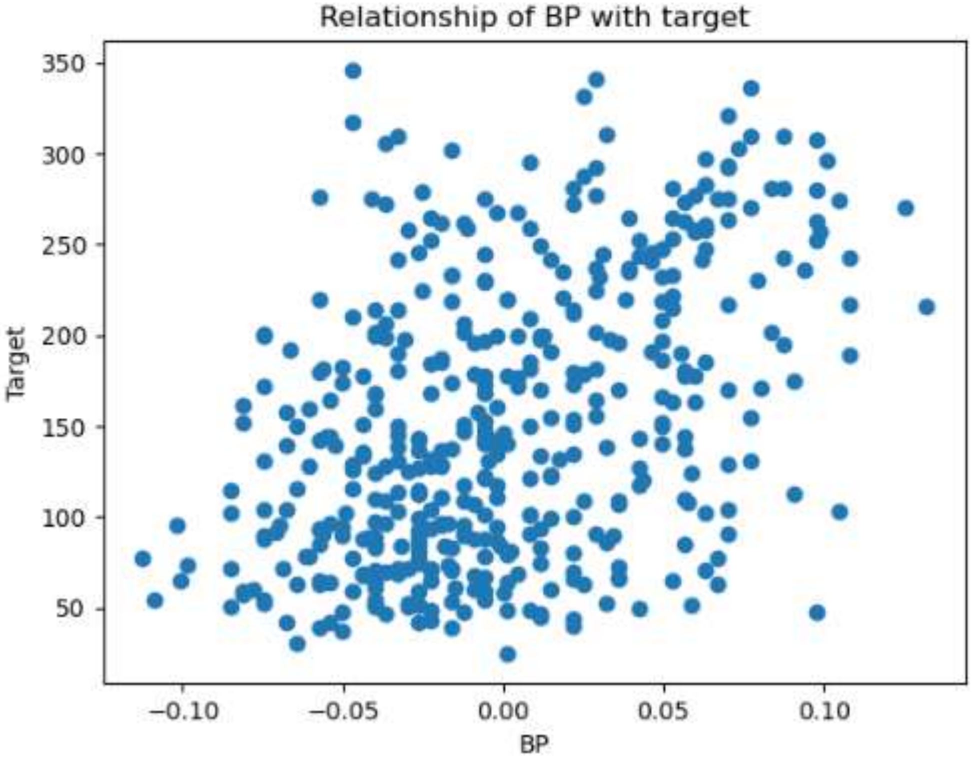
The linear relationships between the top correlated features and the target are shown below:

- Scatter Plot (BMI - Target):



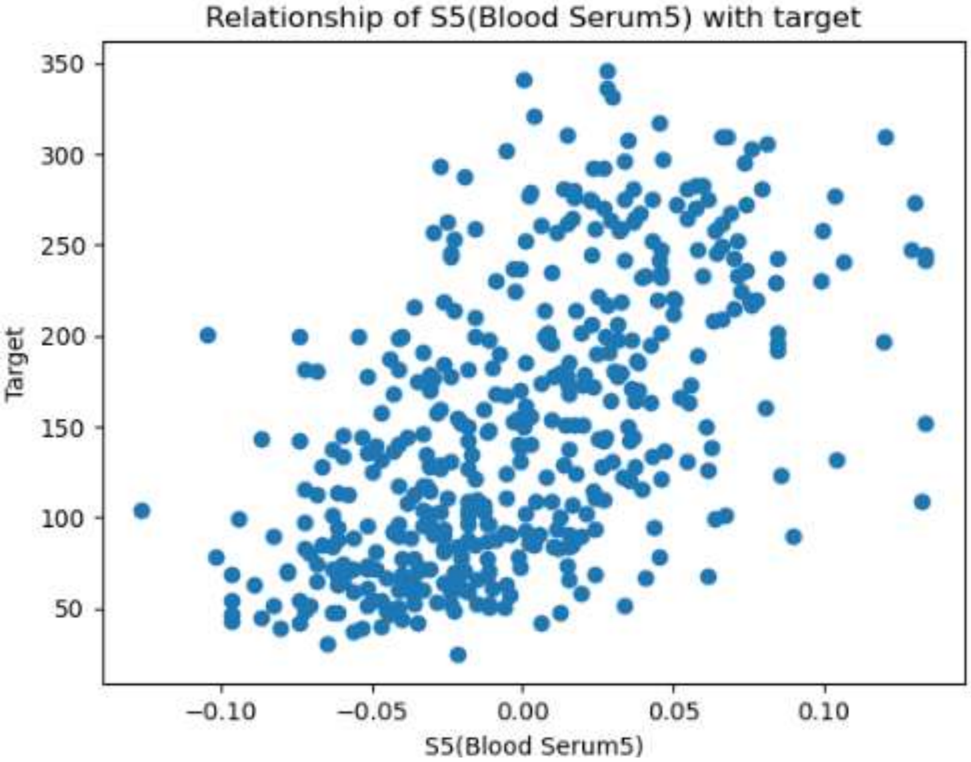
Explanation of Graph: This scatter plot compares the **Body Mass Index (BMI)** on the X-axis with the **Target (diabetes progression score)** on the Y-axis. The plot exhibits a clear, moderately strong **positive linear trend**. As the BMI value increases (moving right), the data points generally move upwards along a conceptual line. This confirms a significant positive relationship, suggesting that individuals with higher BMI tend to have a higher measure of diabetes progression.

- Scatter Plot (BP - Target):



Explanation of Graph: This plot visualizes the relationship between **Blood Pressure (BP)** and the **Target**. It suggests a **positive linear trend**, but the data points are noticeably more **scattered** and less concentrated around a line compared to the BMI plot. This higher dispersion of points visually confirms the lower correlation coefficient (0.44). The graph indicates that while higher BP is generally associated with higher progression, the relationship is weaker and less reliable as a single linear predictor.

- Scatter Plot (S5 - Target):



Explanation of Graph: This scatter plot displays the relationship between the **S5 blood serum measurement** and the **Target**. It reveals the **strongest and most distinct positive linear pattern** among the three plots. The data points cluster most closely around a potential straight line, indicating the highest degree of linear predictability, confirming its high correlation coefficient (0.56).

7.3. Linear Regression Model

A linear regression model was trained using the scaled data for the selected features (bmi, bp, s5).

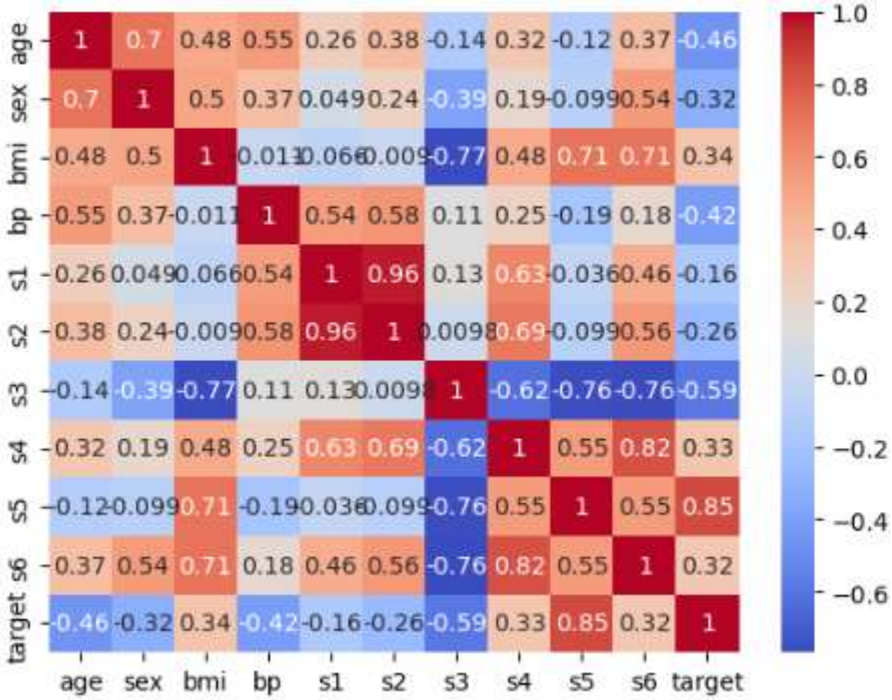
- **Model Performance:**
  - Mean Squared Error (MSE): 2900.19
  - R2 Score: 0.453

- **Explanation:** The R2 score of 0.453 means that the linear model explains approximately 45.3% of the variance in the target variable. The MSE of 2900.19 quantifies the average squared prediction error.

7.4. Manual EDA on First 10 Rows

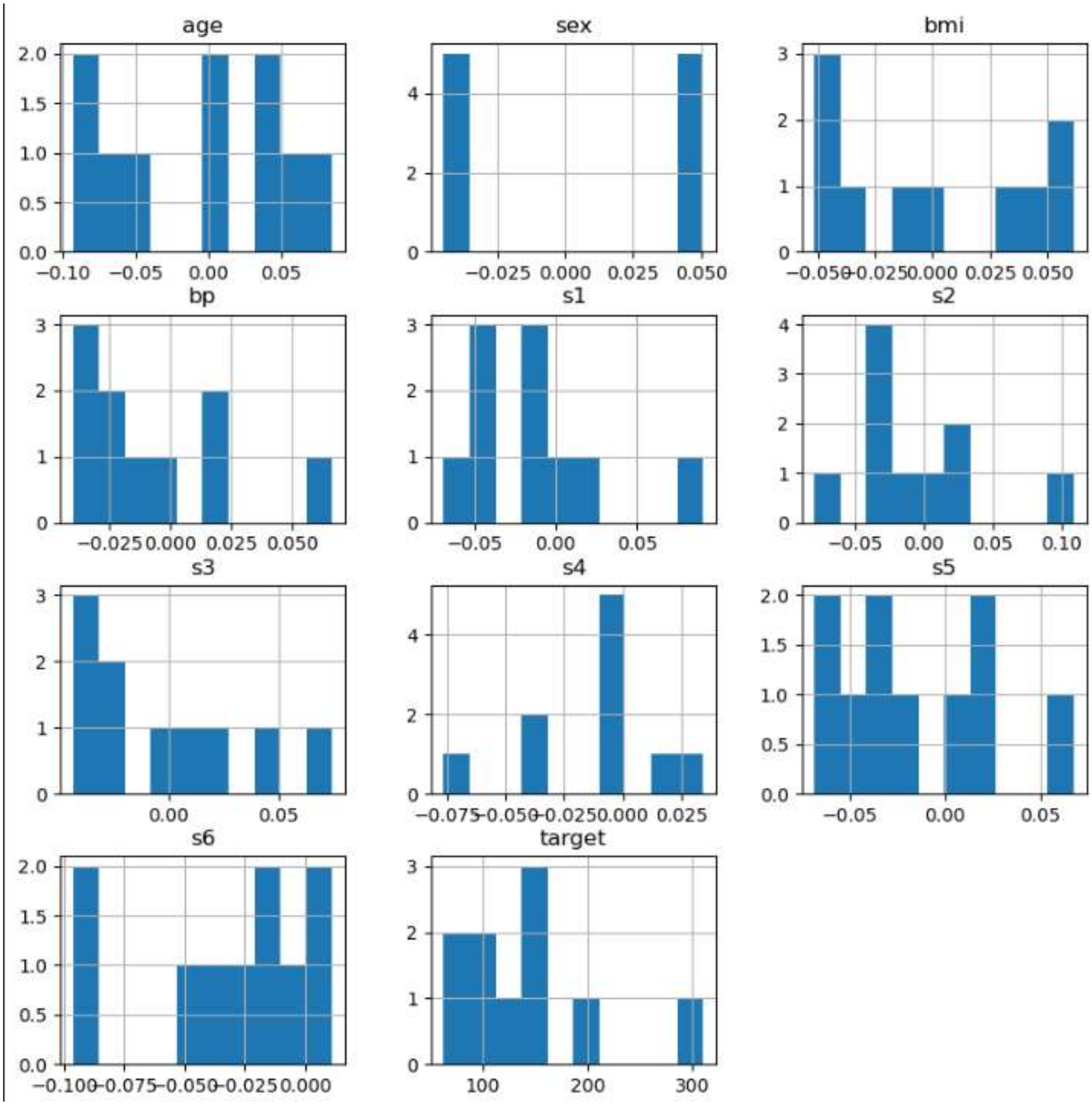
The manual EDA on a subset of 10 rows highlighted the importance of a sufficient sample size.

- Correlation Visualization (10-Row Subset):



Explanation of Graph: This graph (likely a bar plot or small heatmap) shows the correlation values calculated only from the first 10 rows. Critically, it displays a highly **inflated correlation** for S5 to the target (e.g., **0.848**), which is much higher than the full dataset's true correlation (0.56). This visualization clearly demonstrates the danger of **sampling bias**: a small, non-representative sample can lead to overly optimistic and misleading conclusions about the true strength of the relationship in the entire population.

- Feature Distribution (10-Row Subset):



- **Explanation of Graph:** These histograms for the 10-row subset are extremely sparse, often showing only a handful of bars (or single values) for each feature. They do not resemble any standard statistical distribution. This visualization confirms that a tiny sample is **insufficient** to reliably infer the true shape or distribution of the underlying data, making any statistical inference based on this subset invalid.

8. Conclusions

Data Condition: The Diabetes dataset is clean and ready for modeling with **no missing values**.

Predictive Features: Body Mass Index (bmi) and the blood serum measurement **s5** are the strongest linear predictors of diabetes progression.

Model Adequacy: The Linear Regression model provides a **moderate fit (R2 = 0.453)**. While it explains almost half of the target variance, more complex non-linear models may be necessary to achieve higher accuracy.

Sampling Warning: The disparity in results between the full dataset and the small 10-row subset proves that **EDA must be performed on a sufficiently large and representative sample** to draw valid conclusions.

9. Recommendations (Optional)

- **Explore Non-Linear Models:** Investigate algorithms like Random Forest, Support Vector Regression, or neural networks to capture non-linear relationships that the current model may be missing.
- **Feature Engineering:** Test the effect of interaction terms (e.g., bmi * s5) on model performance.

10. References (Optional)

EDA and Preprocessing using Scikit-learn: diabetes-data-analysis-sklearn.ipynb

Manual EDA: manually-eda.ipynb

12. Appendices (Optional)

- *Content omitted for brevity.*