Predicting Bodyfat Using Variable Selection Techniques

A Data-Driven Approach

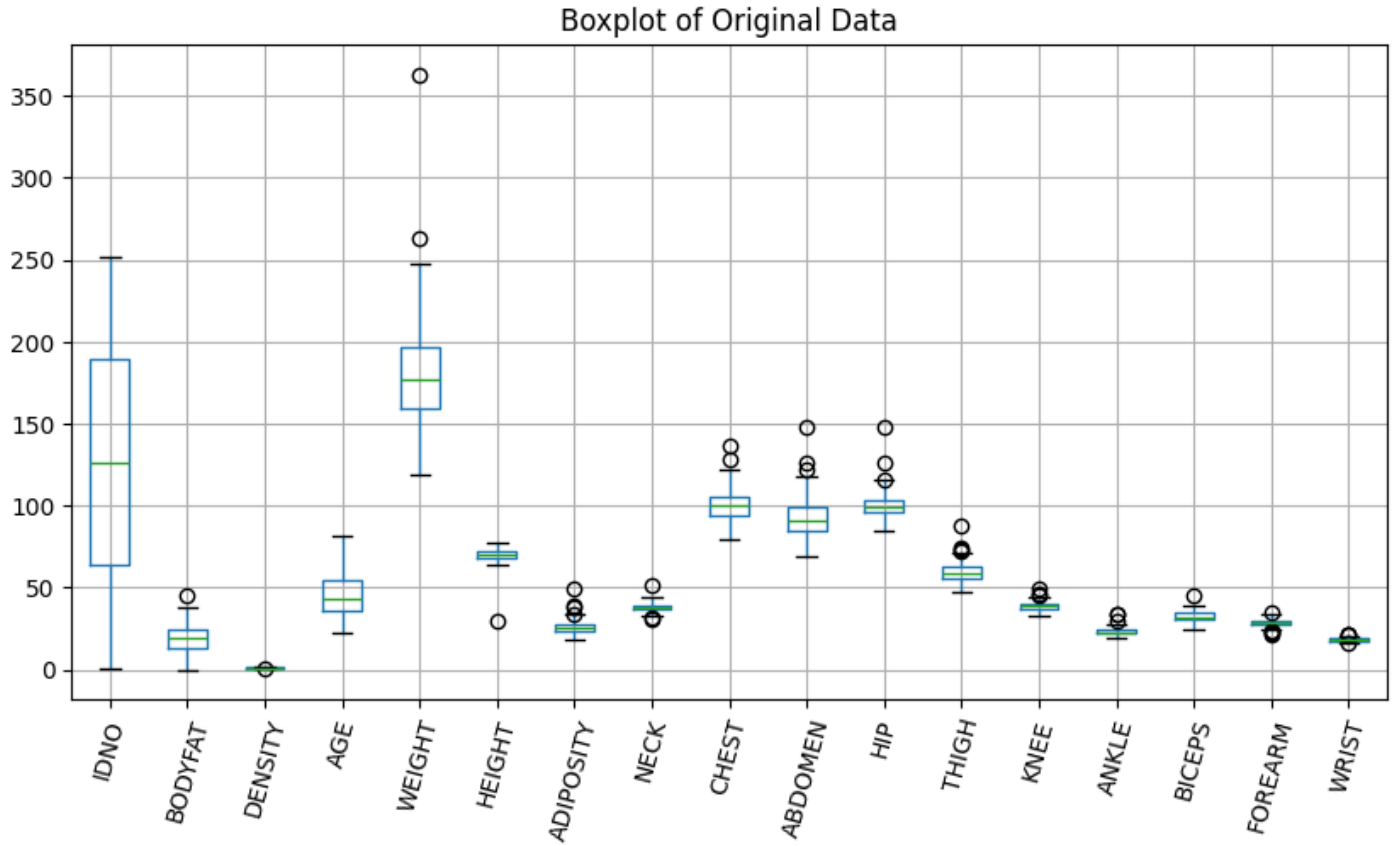
Shuo Li, Xinrui Zhong, Yunze Wang

Data Science Project
University of Wisconsin-Madison
October 2024

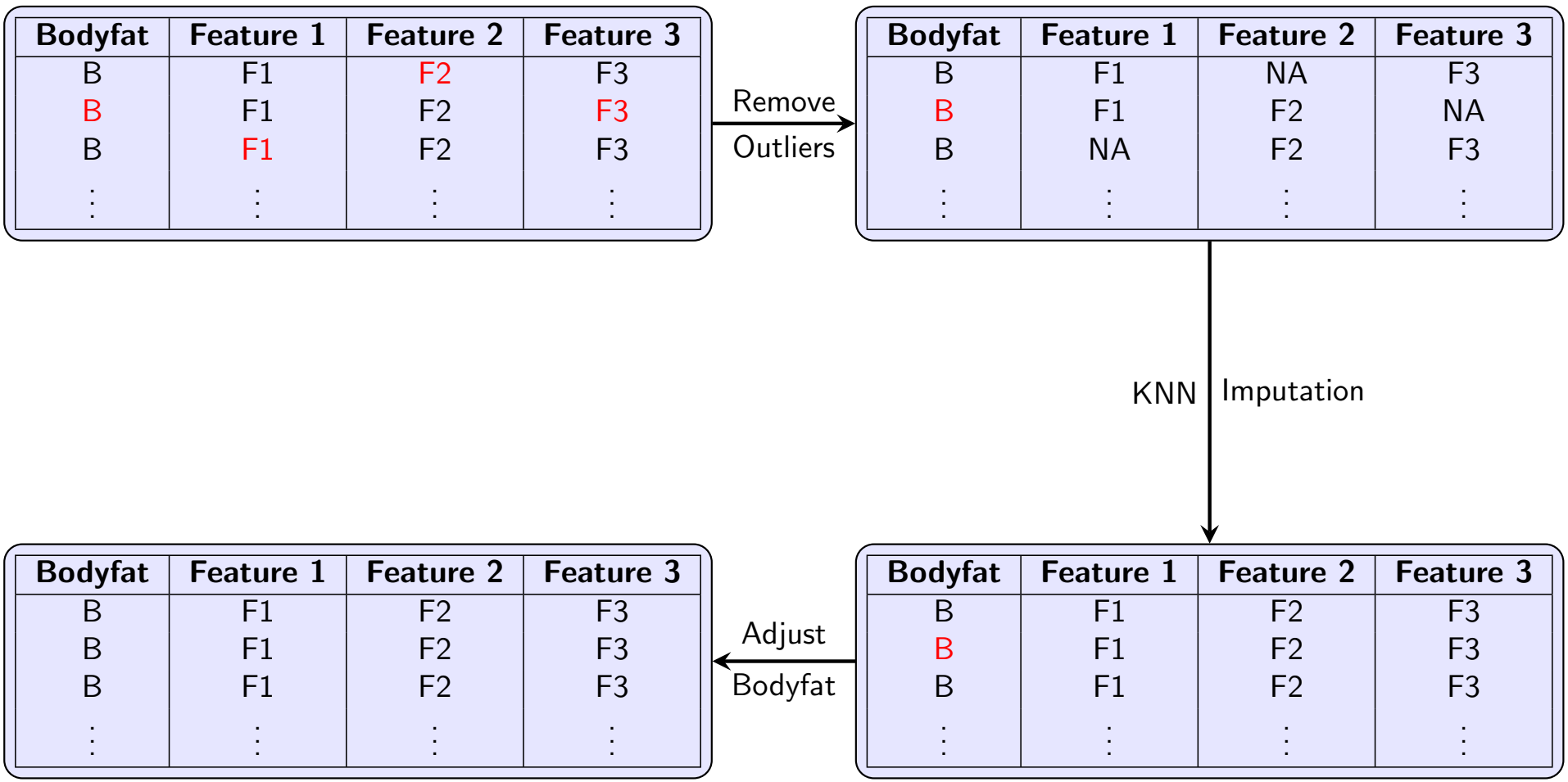
Outline

Data Preprocessing

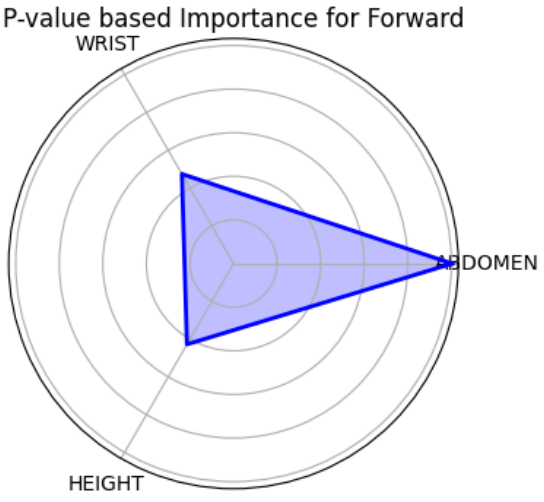
- Dataset: `BodyFat.csv`
- Removed outliers using the IQR method.
- Imputed missing values using KNN.
- Handling Bodyfat column.

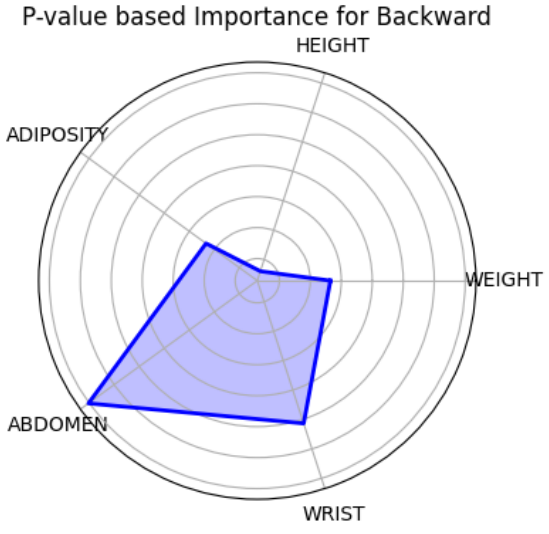


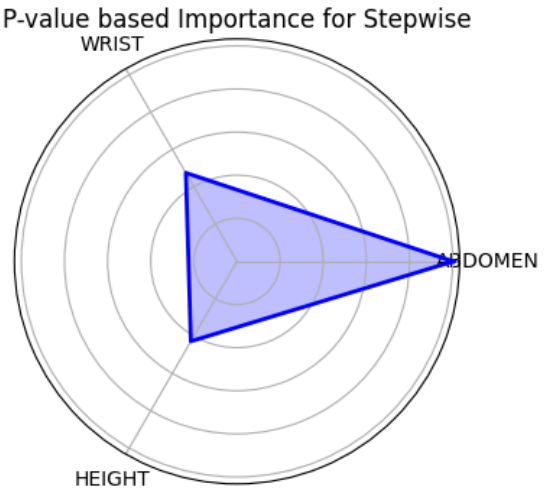
Flowchart with Dataset Illustrations

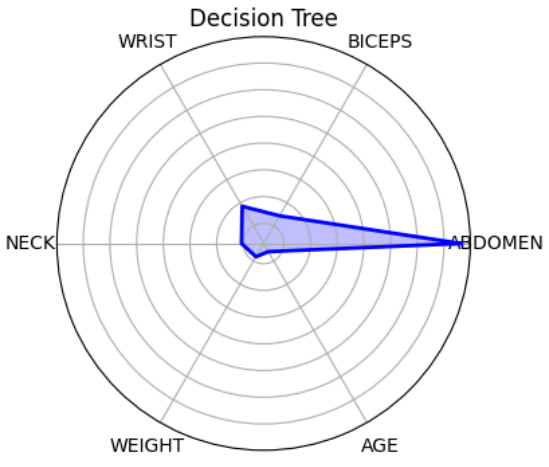


Feature Selection









Model Comparison

- Best model: Multiple Linear Regression with Forward Selection
- Comparison of model performance:

	MLR_Forward	MLR_Backward	Decision Tree
Feature numbers	3	5	6
R^2	0.731	0.732	0.828
Adjusted R^2	0.727	0.726	0.823
MSE	13.25	13.19	8.47
Cross-validation MSE	13.84	14.00	28.45
F-test (p-value)	2.58e-70	3.56e-68	NA
Jarque-Bera test (p-value)	0.112	0.110	NA

Table: Comparison of different models

Model Diagnostics

- Normality of the error terms :Jarque-Bera test
- Homoskedasticity:Residual plot

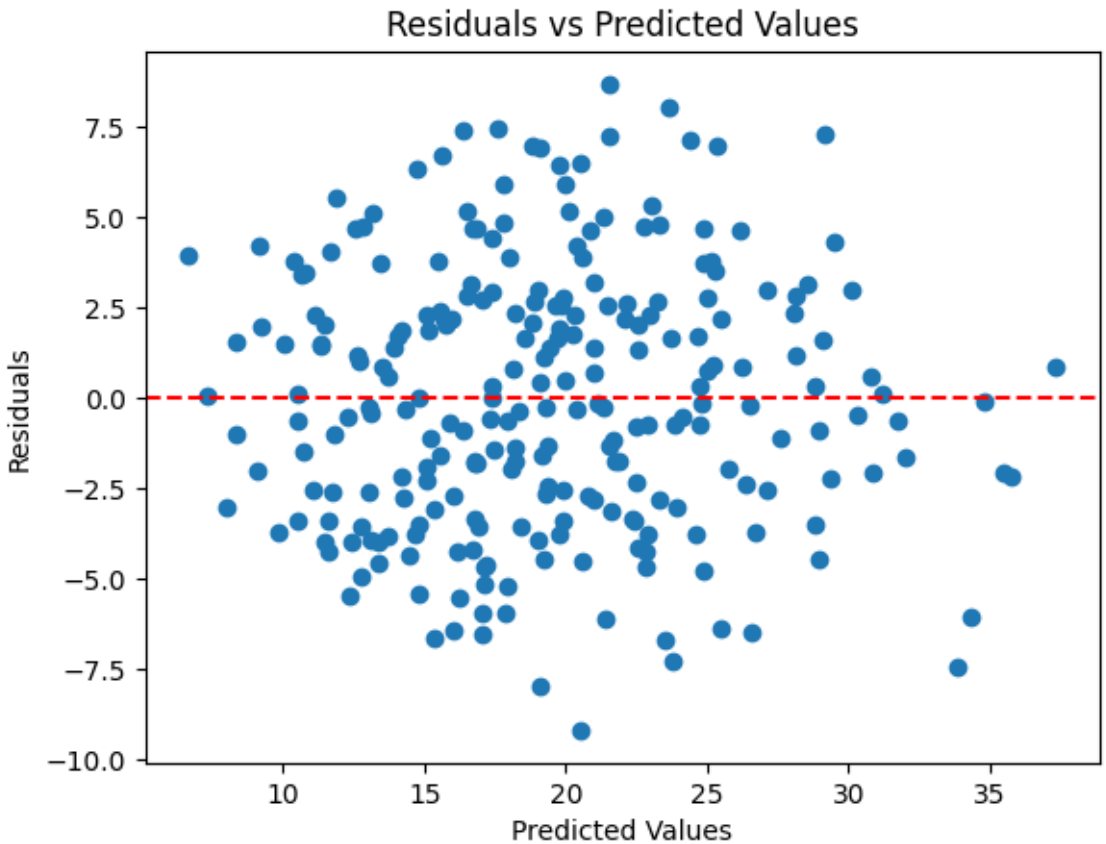


Figure: Resudual plot of MLR_Forward

Conclusion

$$BF = 9.43 + 32.97 \times \frac{Abd - 69.4}{118 - 69.4} - 6.55 \times \frac{Wrt - 16.1}{20.4 - 16.1} - 5.17 \times \frac{Height - 64}{77.75 - 64} \quad (1)$$

- Advantages
  - The multiple linear regression model is simple in structure, making it easy to implement and use.
  - MLR is easy to interprete, together with the statistical meaning of the parameters.
- Disadvantages
  - MLR relies heavily on the assumption that the residuals (errors) follow a normal distribution.
  - It's difficult to clearly explain the individual effect of each variable.

Shiny APP

Shiny App link: **click here**

Thank you!