Comparative Analysis of ML Models

Obesity Level Classification

Lily Vogel

01 - Introduction

02 - Exploratory Methods

03 - Analysis

04 - Conclusions

Obesity

Classification

01 - Introduction

Objective

The objective of this project is to evaluate the accuracy of different modeling approaches in the context of classifying obesity level. The analysis will compare the performance of several methods, including multinomial logistic regression, boosted multinomial logistic regression, and random forests. Additionally, hyperparameter optimization will be used to try to enhance the models' predictive accuracy.

01 - Introduction

Hypothesis

I expect that the boosted logistic regression will outperform standard multinomial logistic regression and random forests in accurately classifying obesity levels in the dataset.

01 - Introduction

The Data

- 16 features - combination of continuous & categorical
- 1 multi-class response variable, Obesity Level

01 - Introduction

The Data

Eating Habit Features

- Do you eat high caloric food frequently?
- Do you usually eat vegetables in your meals?
- How many main meals do you have daily?
- Do you eat any food between meals?
- How much water do you drink daily?
- Do you monitor the calories you eat daily?
- How often do you drink alcohol?

01 - Introduction

The Data

Physical Condition Features

- Do you smoke?
- How often do you have physical activity?
- How much time do you use technological devices such as cell phone, videogames, television, computer, and others?
- Which transportation do you usually use?

01 - Introduction

The Data

Physical Condition Features

- Do you smoke?
- How often do you have physical activity?
- How much time do you use technological devices such as cell phone, videogames, television, computer, and others?
- Which transportation do you usually use?

01 - Introduction

The Data

Other Features

- Has a family member suffered or suffers from overweight?
- Gender (Female or Male)
- Age (years)
- Height (meters)
- Weight (kilograms)

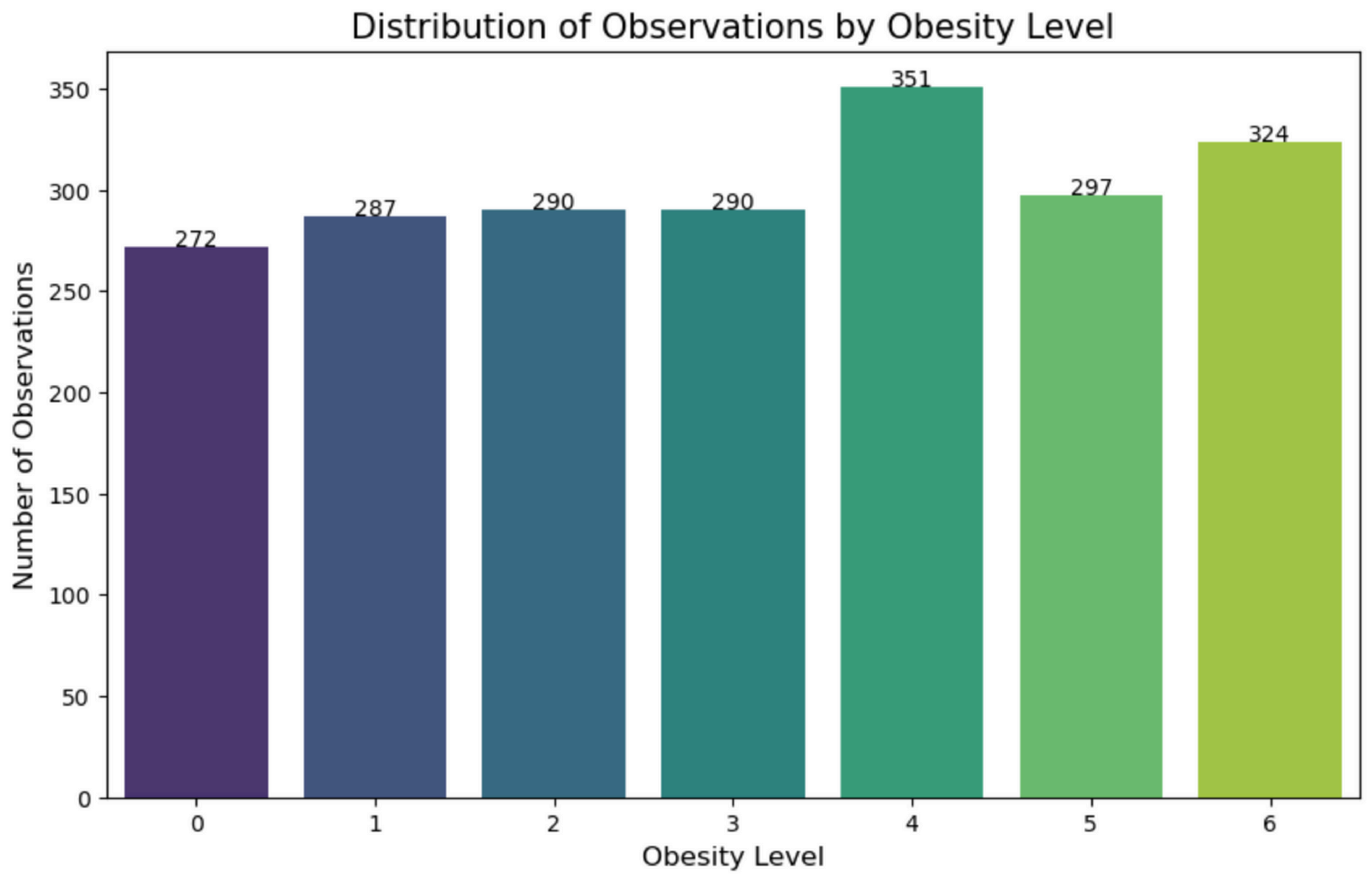
01 - Introduction

The Data

Response variable: Obesity Level (in BMI ascending order)

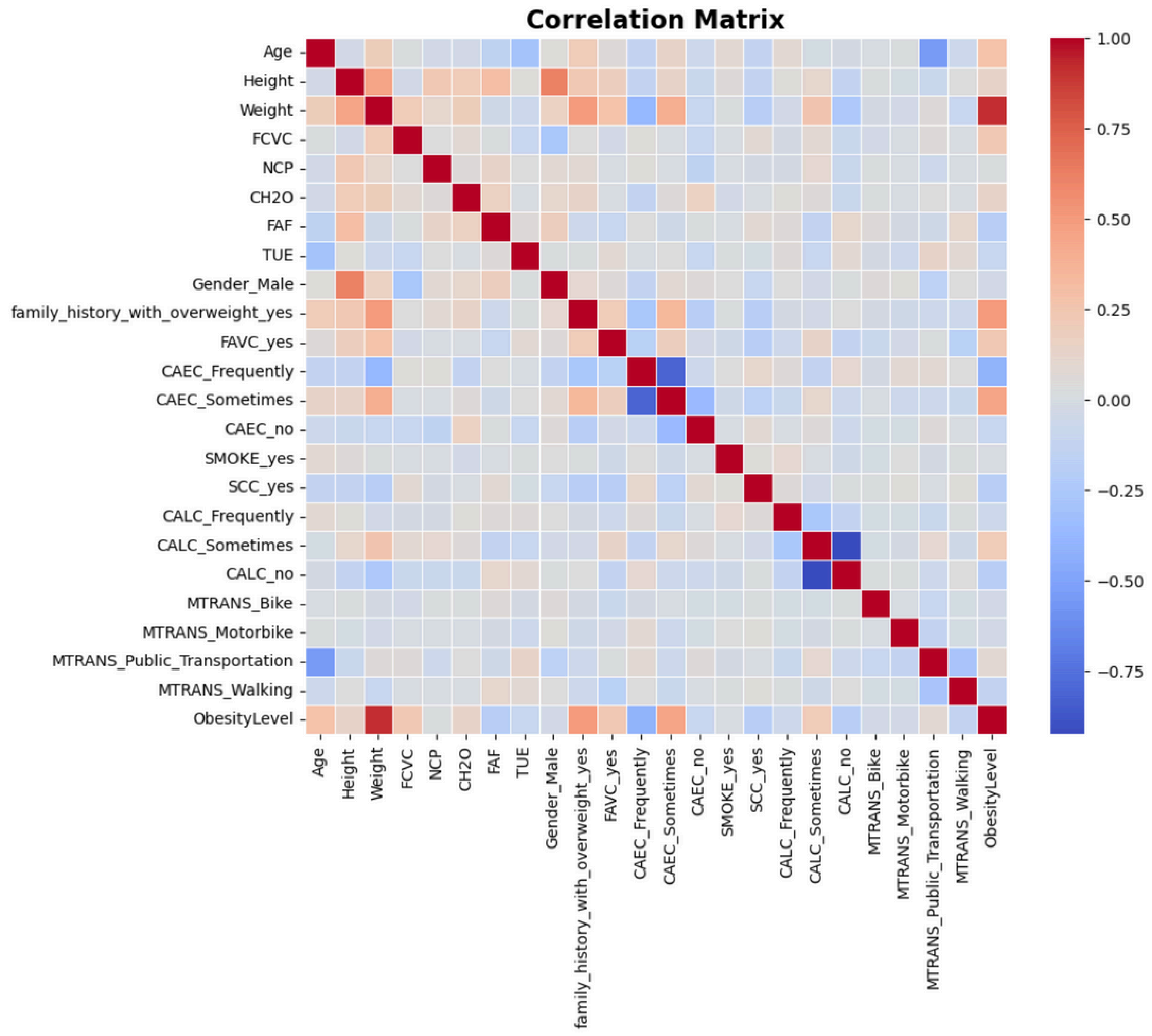
- insufficient weight
- normal weight
- overweight level 1
- overweight level 2
- obesity type 1
- obesity type 2
- obesity type 3

02 - Exploratory Methods



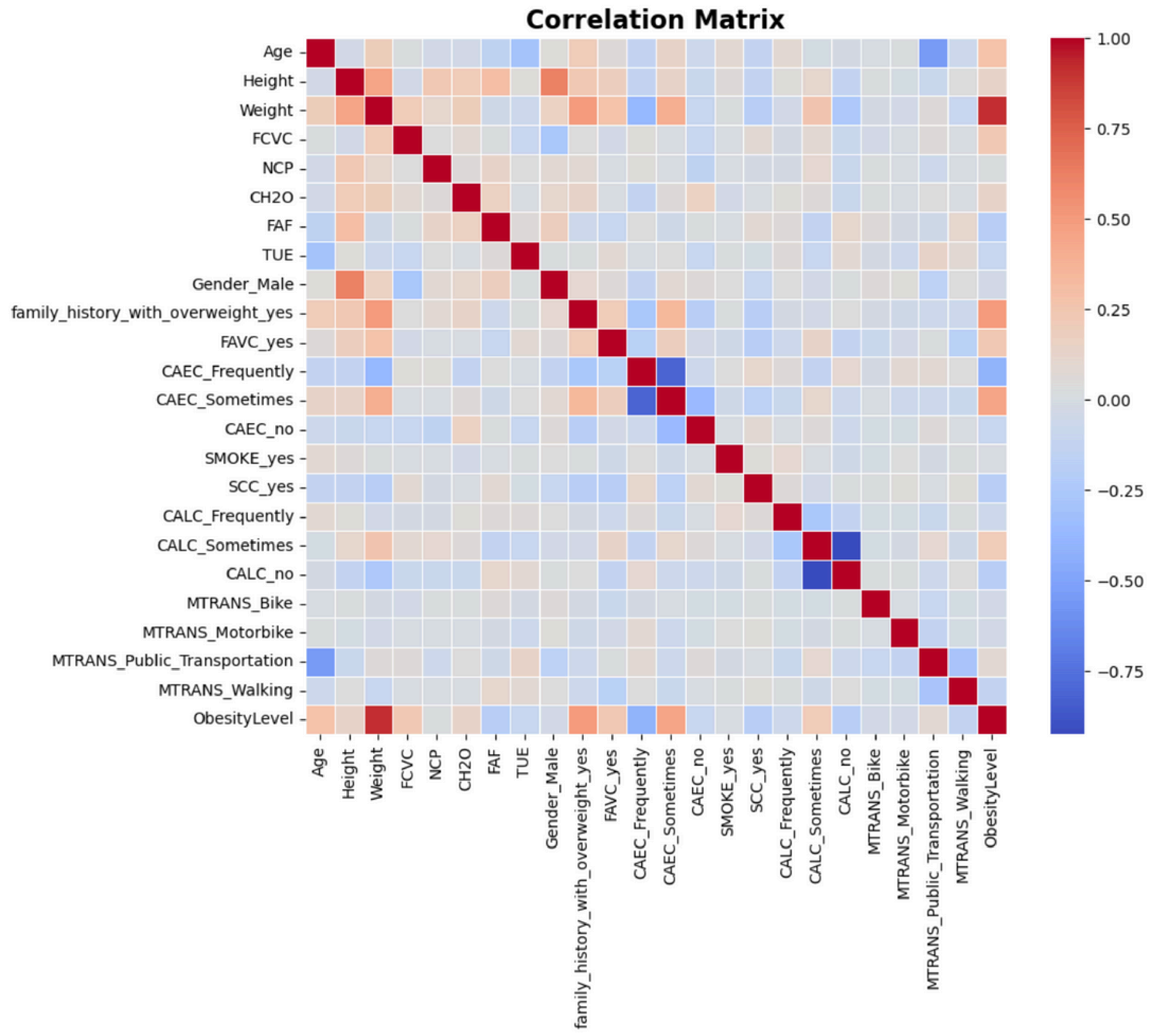
It appears that the observations are well distributed. Each class has at least 270 observations. This is a good distribution of observations since the class sizes are balanced.

02 - Exploratory Methods



Based on this correlation matrix, weight and obesity level are highly correlated as we would expect. Other features that have notable correlation with obesity level are family with overweight history, 'CAEC_Sometimes' (sometimes has food between meals), and 'CAEC_Frequently' (frequently has food between meals). Some features with some correlation but not strong are age, FAVC_yes (frequently eats high-calorie foods), and `CALC_Sometimes` (sometimes consumes alcohol).

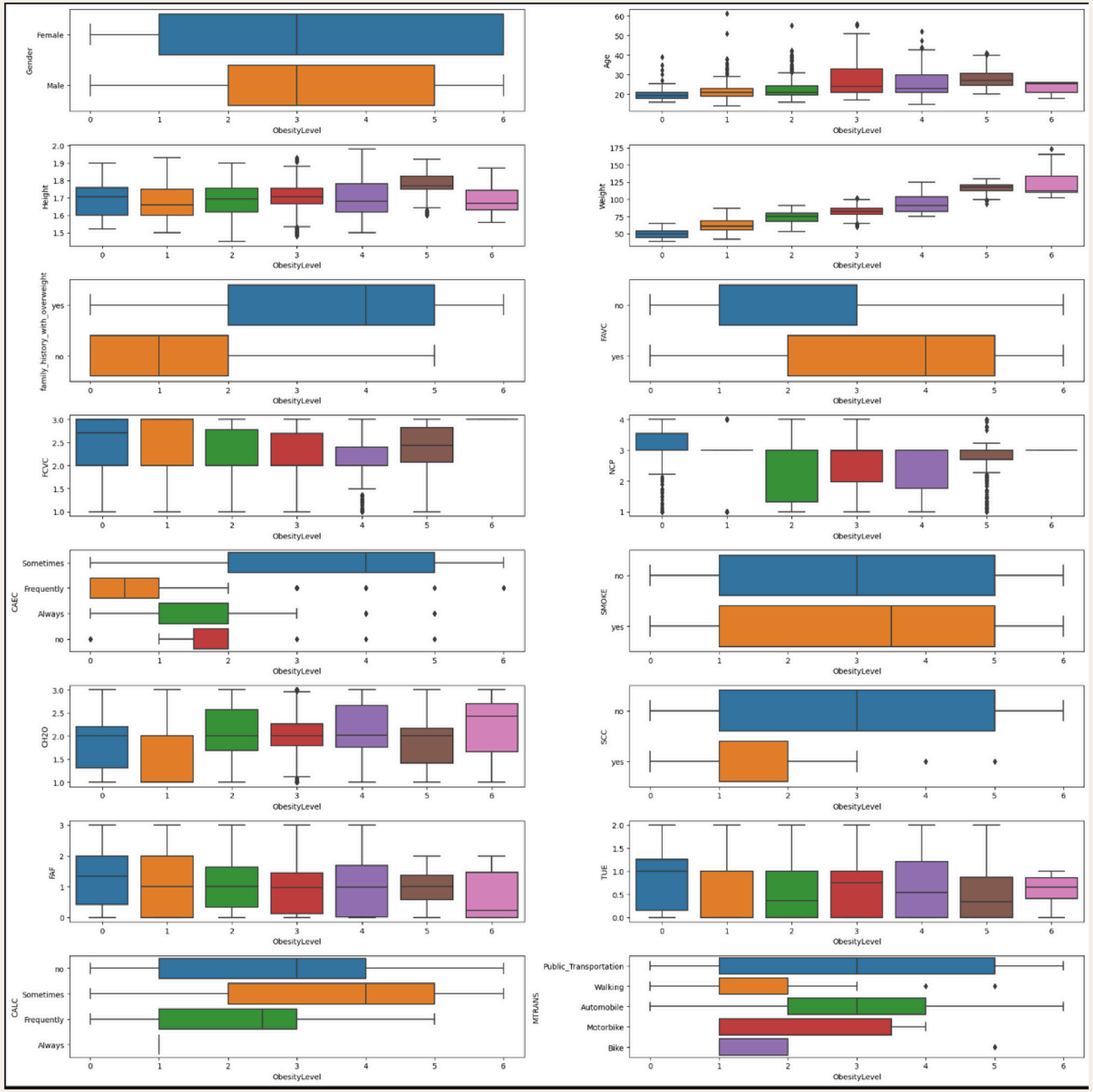
02 - Exploratory Methods



Based on this correlation matrix, weight and obesity level are highly correlated as we would expect. Other features that have notable correlation with obesity level are family with overweight history, ‘CAEC_Sometimes’ (sometimes has food between meals), and ‘CAEC_Frequently’ (frequently has food between meals). Some features with some correlation but not strong are age, FAVC_yes (frequently eats high-calorie foods), and `CALC_Sometimes` (sometimes consumes alcohol).

02 - Exploratory Methods

This visualization shows boxplots of each feature against the response variable, obesity level.



03 - Analysis

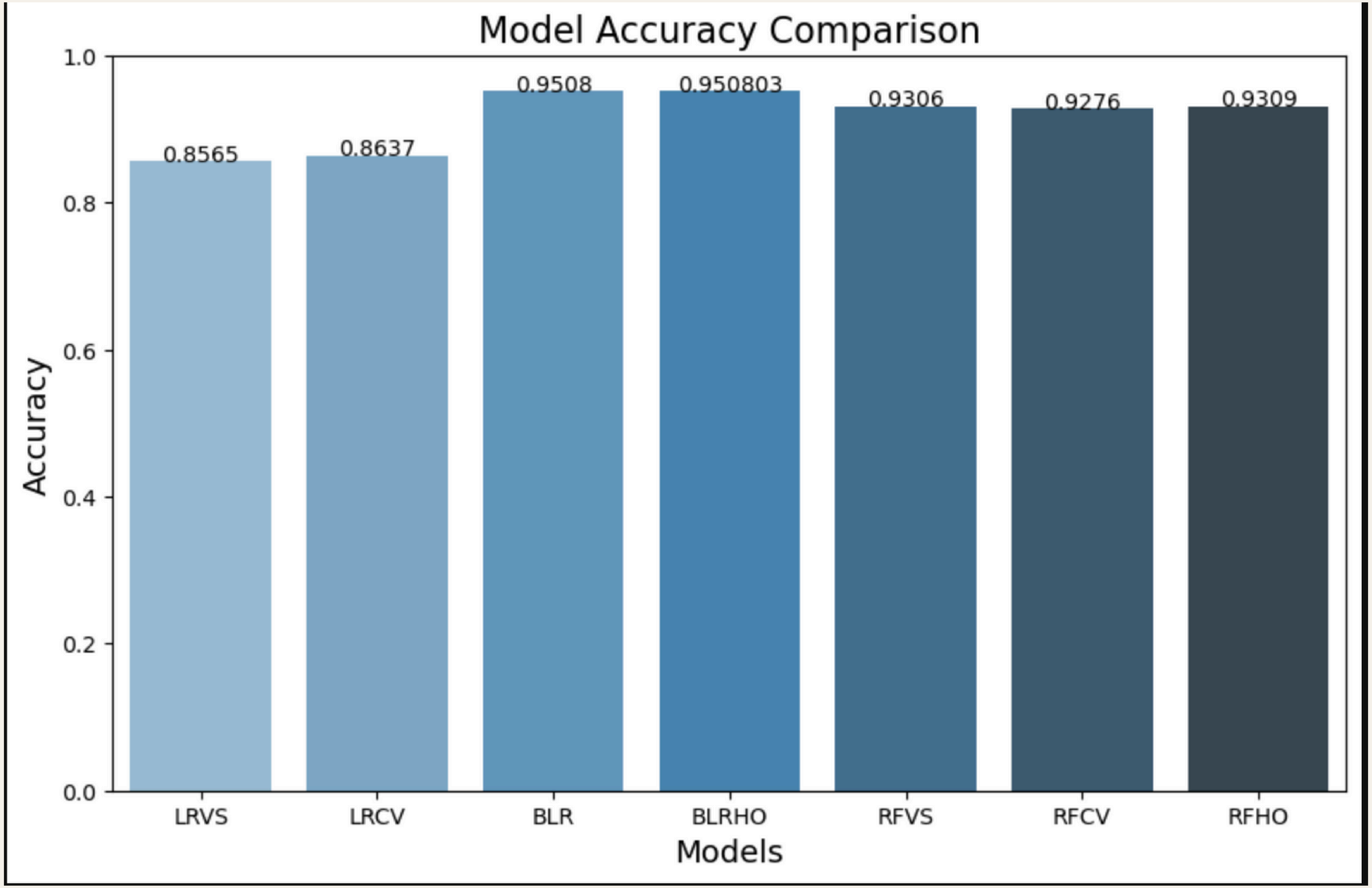
Models:

- Multinomial Logistic Regression using Validation Set
- Multinomial Logistic Regression using Cross-Validation
- Boosted Multinomial Logistic Regression
- Boosted Multinomial Logistic Regression with Hyperparameter Optimization
- Random Forests using Validation Set
- Random Forests using Cross-Validation
- Random Forests with Hyperparameter Optimization

03 - Analysis

Legend:

- LRVS = Log Reg using Validation Set
- LRCV = Log Reg using CV
- BLR = Boosting
- BLRHO = Boosting with Hyperparameter Optimization
- RFVS = Random Forests using Validation Set
- RFCV = Random Forests using CV
- RFHO = Random Forests with Hyperparameter Optimization



04 - Conclusions

The Boosted models outperform the other models, suggesting that boosting is highly effective for this dataset.

The Random Forest models perform well but not as well as the boosted models. The accuracy improvement with hyperparameter tuning (`RFHO`) is small.

The Logistic Regression models perform well but are outperformed by the more complex models, suggesting that while Logistic Regression works well for this problem but more complex models can lead to improvements in model accuracy.