Project Report

Ameera Attiah S21107316 - Jana Abu Hantash S21107114 - Ahmad ElMaamoun S21207525

Fall 2023

Contents

1 Introduction

In the evolving landscape of health informatics, the focus of this project is centered on leveraging data science methodologies to construct a predictive model aimed at understanding and forecasting obesity trends. We delve into a detailed examination of a dataset, encompassing a wide array of physical and behavioral characteristics relevant to obesity, to uncover underlying patterns and insights.

The main goal of this project is to use the latest data analysis methods to build a model that predicts obesity. We want to create a model that shows how complex obesity is and helps us start tackling it early and plan better health approaches. We're using this dataset as our base, showing how different physical and lifestyle aspects come together to affect obesity.

The report will methodically cover our investigation of the obesity dataset. Subsequent sections will discuss in detail the data preprocessing steps, exploratory data analysis, feature selection, model building, and validation of the predictive model. The final part of this report will summarize our results, share the lessons we learned from our model, and discuss what our research could mean and how it might be used.

2 Problem Statement / Aim of our Analysis

The problem statement is to predict obesity levels in individuals based on various predictors like physical measurements and lifestyle choices. This analysis aims to understand the factors contributing to obesity and develop a predictive model that can accurately determine an individual's obesity level.

3 Literature Review

3.1 Predictive Modeling in Obesity Research

The challenge of obesity has been addressed through various predictive models, aiming to utilize data to forecast obesity levels based on lifestyle and physical conditions. The dataset provided by Mendoza & Montas (2019) serves as a foundational resource in such efforts, offering a rich compilation of variables from eating habits to physical activity, which are instrumental in predicting obesity levels with statistical models such as Random Forests and Linear Discriminant Analysis.

3.2 Obesity's Role in Health and Disease

Understanding the health implications of obesity is pivotal in predictive modeling. Research by Aziz et al. (2023) and Lam et al. (2023) investigates how obesity can increase the severity of diseases, especially viral respiratory infections, and its wider effects on health. These findings underscore the importance of incorporating medical data into predictive models to better estimate the risks associated with obesity.

3.3 Global Dietary Habits and Obesity Trends

Further research by Lafia et al. (2022) emphasizes the influence of global dietary habits on obesity, highlighting the diversity in nutritional patterns and their implications on health. This global perspective enriches predictive models by accounting for regional variations in obesity determinants.

4 Data

4.1 Data Set Variables

The unit of observation in this dataset is individual participants, with each row representing a unique individual's data.

4.1.1 The Outcome Variable

- The outcome variable is the obesity level, categorized into distinct classes 'Insufficient Weight', 'Normal Weight', 'Overweight Level I', 'Overweight Level II', 'Obesity Type I', 'Obesity Type II', and 'Obesity Type III'.
- The variable is derived from participants' physical and lifestyle data.
- The distribution of the outcome variable is illustrated in the graph and the frequency table below:
  - Insufficient Weight: 272
  - Normal Weight: 287
  - Overweight Level I: 290
  - Overweight Level II: 290
  - Obesity Type I: 351
  - Obesity Type II: 297
  - Obesity Type III: 324

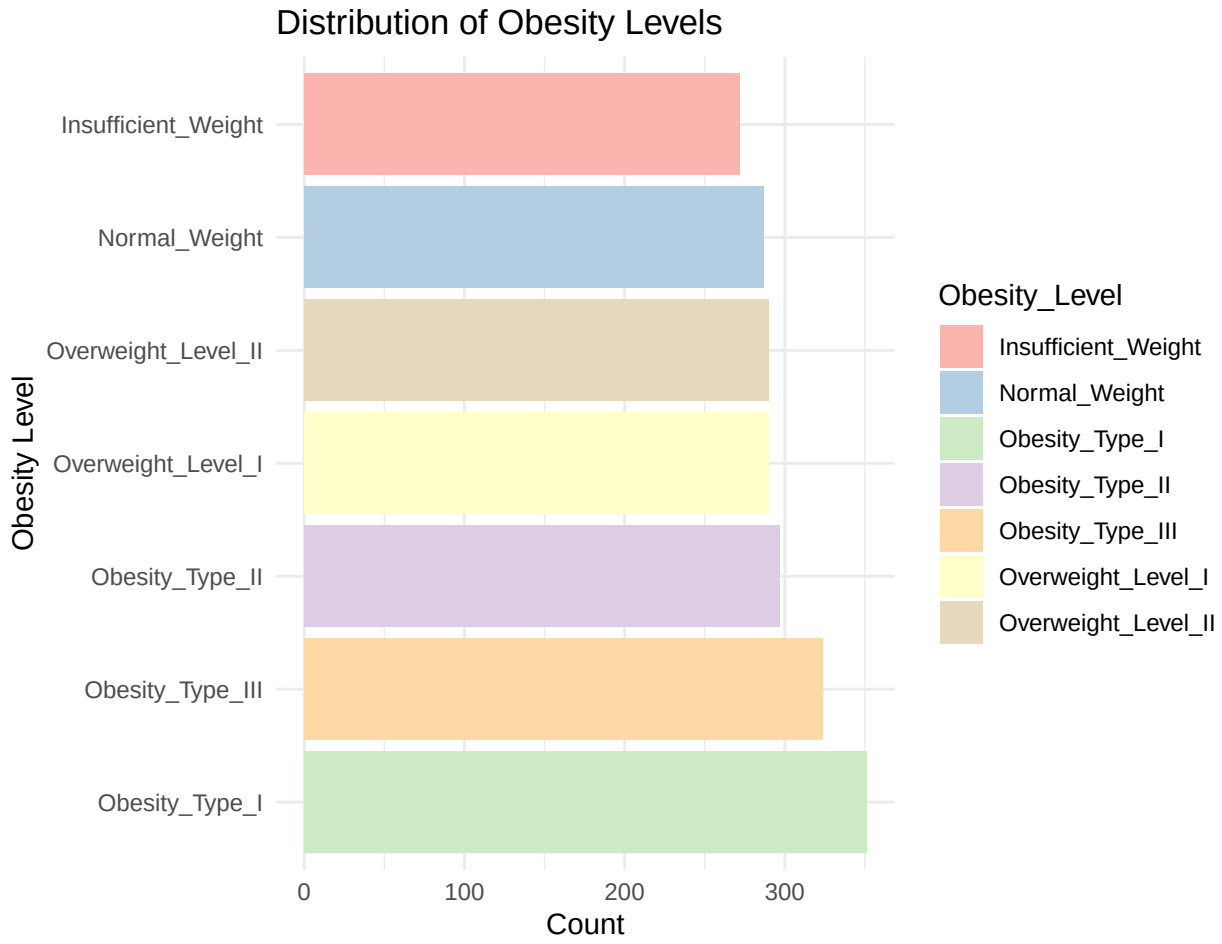

Figure 1. Distribution of Obesity Levels

4.1.2 Predictor Variables

- The predictor variables include age, gender, height, weight, family history of overweight, eating habits, physical activity, etc.
- These variables are measured through surveys or collected data.
- The distribution of each predictor will be presented using descriptive statistics and visualizations.

Descriptive Statistics of Key Variables:

1. **Age**
   - Mean: 24.31 years
   - Standard Deviation: 6.35 years
   - Range: 14 to 61 years
2. **Height**
   - Mean: 1.70 meters
   - Standard Deviation: 0.09 meters
   - Range: 1.45 to 1.98 meters
3. **Weight**
   - Mean: 86.59 kg
   - Standard Deviation: 26.19 kg
   - Range: 39 to 173 kg
4. **Frequency of Consumption of Vegetables (FCVC)**
   - Mean: 2.42
   - Standard Deviation: 0.53
   - Range: 1 to 3
5. **Number of Main Meals (NCP)**
   - Mean: 2.69
   - Standard Deviation: 0.78
   - Range: 1 to 4
6. **Water Consumption (CH2O)**
   - Mean: 2.01
   - Standard Deviation: 0.61
   - Range: 1 to 3
7. **Physical Activity Frequency (FAF)**
   - Mean: 1.01
   - Standard Deviation: 0.85
   - Range: 0 to 3
8. **Time Using Technology Devices (TUE)**
   - Mean: 0.66
   - Standard Deviation: 0.61
   - Range: 0 to 2

The histograms for each of these variables depict their distributions as shown in figure 2.Most variables show a diverse range of values, suggesting a good variety in the dataset for these predictors. For instance, 'Age' shows a right-skewed distribution, indicating a higher concentration of younger individuals in the dataset, while 'Weight' shows a more normal distribution. These visualizations help in understanding the spread and central tendencies of the predictor variables.

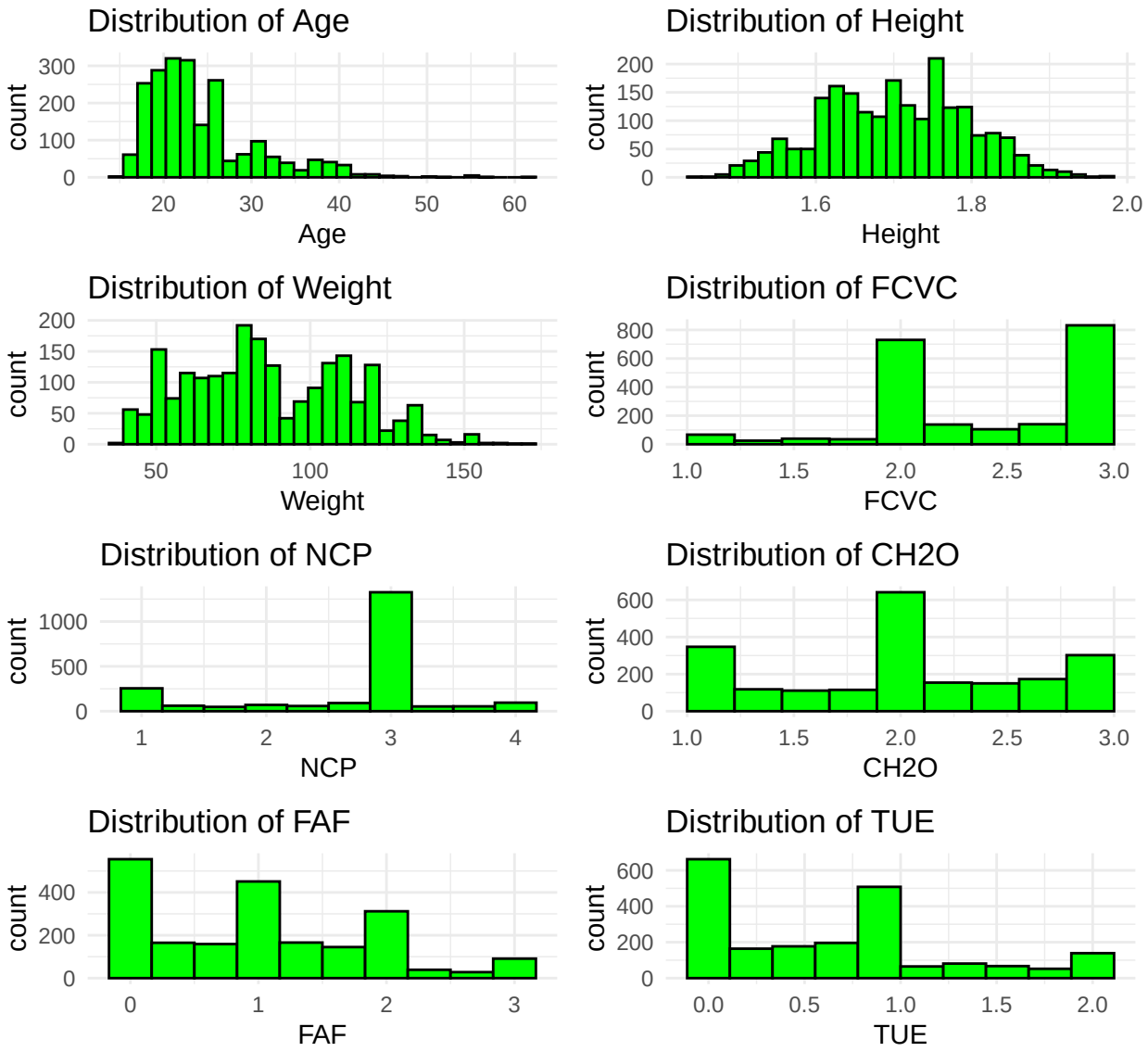

Figure 2. Histograms of the Variable Distributions

4.2 Addressing Data Issues and Solutions

4.2.1 Potential issues with the data

- Some variables might have limited variation, affecting the model's ability to learn from them.
- Bias could be introduced due to self-reported data or sampling methods.

4.2.2 Solutions to the issues

- The issues will be addressed through data cleaning, handling missing values, ensuring diversity in the dataset, and applying statistical techniques to mitigate bias.

5 Analysis

5.1 Methods/Tools Explored

For our project, a comprehensive set of tools and methodologies were employed to address the problem statement and to analyze the provided dataset. The primary software used was R, a powerful tool for statistical computing and graphics. This choice was made due to R's versatility in handling various types of data and its extensive range of packages for data manipulation, visualization, and machine learning.

Key packages utilized in R included:

- `dplyr` and `tidyr` for data manipulation.
- `ggplot2` for data visualization.
- `caret` and `randomForest` for machine learning and predictive modeling.

5.2 Feature Selection

Feature selection was performed using both visual analysis of density plots and Recursive Feature Elimination (RFE). The visual analysis helped us identify patterns and behaviors across different classes within the features, while RFE provided a systematic approach for selecting the most significant features for predictive modeling.

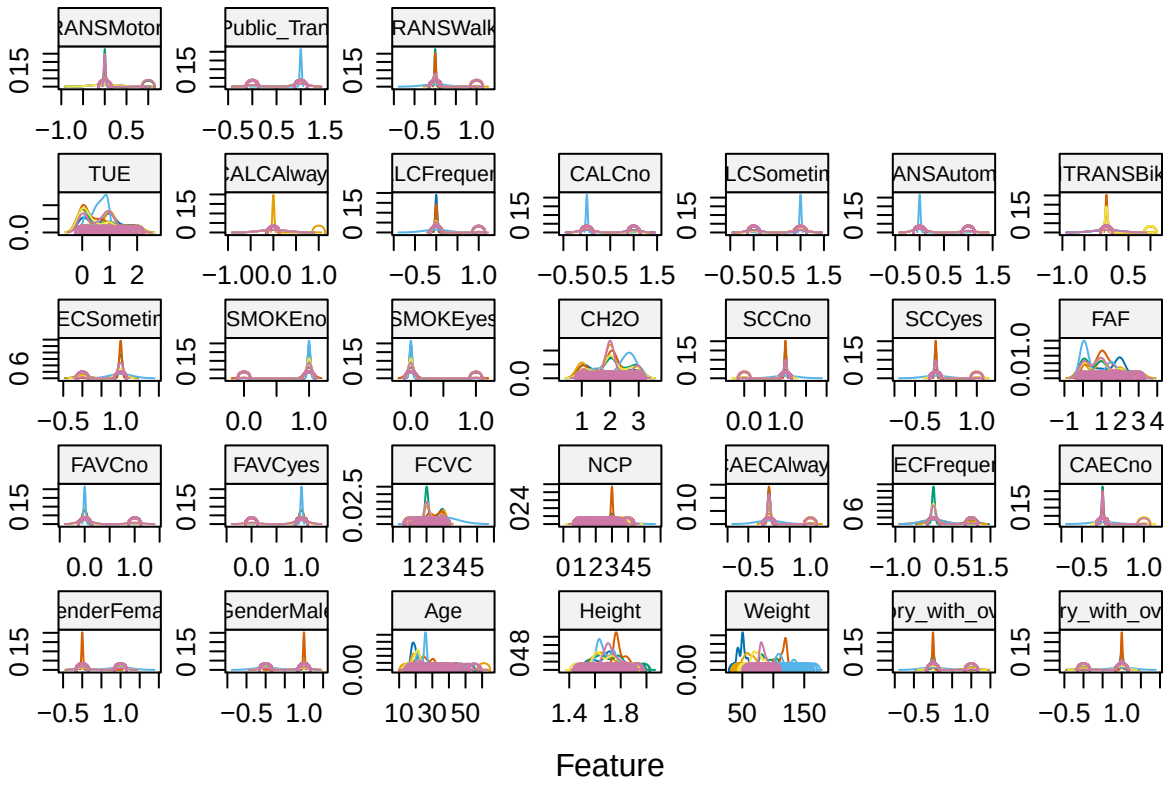

Figure 3. Feature Selection Graph

5.2.1 Recursive Feature Elimination (RFE) Results

Based on the RFE method, we identified the top five variables that contribute the most to the predictive model's performance. These were:

1. Weight
2. Age
3. Height
4. Frequency of Consumption of Vegetables (FCVC)
5. Number of Main Meals (NCP)

These variables showed the highest significance and were therefore chosen for the final model training.

5.3 Training the Data

In this section, we outline the process and methodologies involved in training our predictive models. Training a model is a crucial step in the machine learning workflow where the model learns the patterns from the provided dataset to make predictions on new, unseen data.

5.3.1 Data Preprocessing

Before training, the data underwent several preprocessing steps to ensure its quality and suitability for modeling. These steps included:

- **Cleaning Data**: Removing or imputing missing values and outliers.
- **Feature Selection**: Selecting the most relevant features to reduce dimensionality and improve model performance
- **Data Transformation**: Normalizing or scaling features to ensure that they contribute equally to the model's performance.

5.3.2 Model Selection

We evaluated several machine learning algorithms to find the best performer for our specific problem. The models considered included:

- Random Forest (RF): An ensemble learning method for classification and regression.
- K-Nearest Neighbors (KNN): A simple, instance-based learning algorithm.
- Support Vector Machines (SVM): Effective in high-dimensional spaces.
- Linear Discriminant Analysis (LDA): A generalization of Fisher's linear discriminant.

5.3.3 Model Training

The models were trained using a subset of the data known as the training set. The following steps were involved:

- **Splitting Data**: The data was divided into training and testing sets, using a 80:20 ratio.
- **Cross-Validation**: We used k-fold cross-validation to assess how the results of a statistical analysis will generalize to an independent dataset.
- **Training Models**: Each model was trained using the `caret` package, which provides a fast and efficient way to create predictive models.

5.4 Model Comparison and Evaluation

Our evaluation process involved comparing several machine learning models using accuracy and Kappa statistics. Box plot was generated to visualize the performance metrics for Random Forest (RF), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), and Support Vector Machine with a linear kernel (SVMLinear).

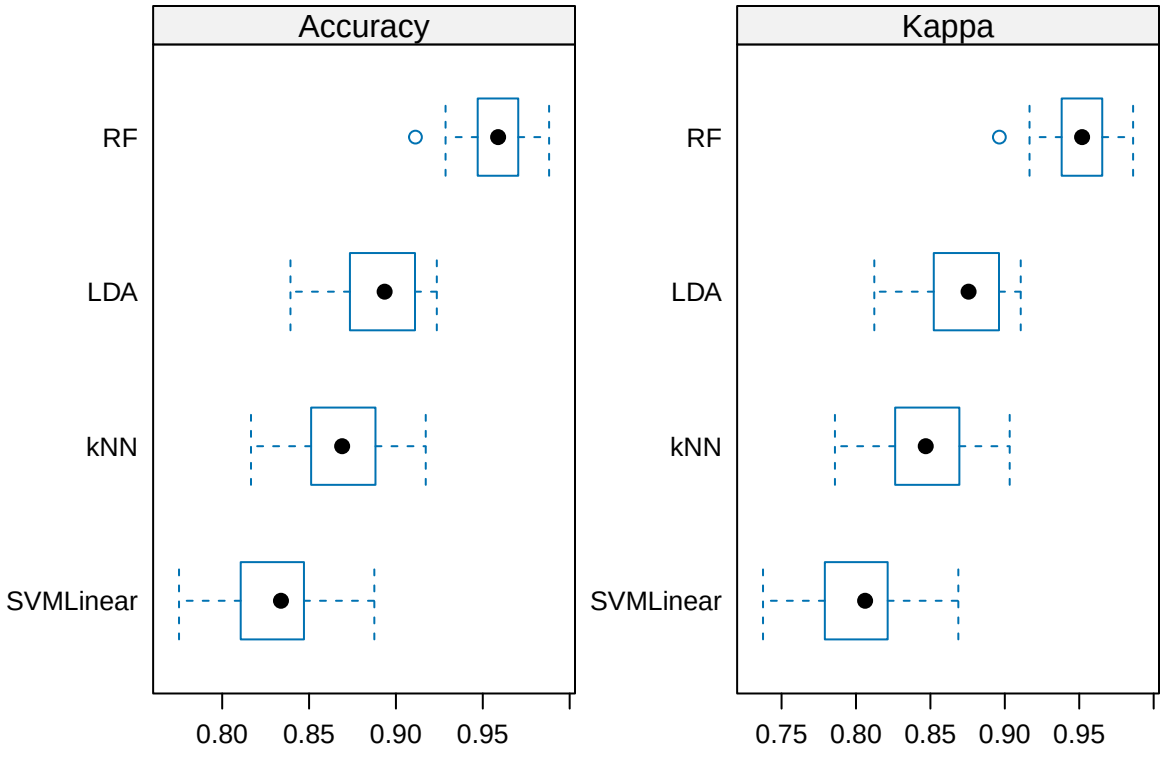

Figure 4. Comparing Machine Learning Models

5.4.1 Performance Metrics Analysis

The analysis revealed that:

- **Random Forest (RF)** showed the highest potential for accuracy and Kappa, but also the greatest variability across different tests, which could be indicative of overfitting or sensitivity to the dataset's nuances.
- **Linear Discriminant Analysis (LDA)** offered a balance between high performance and consistency, with less variability than RF, making it a potentially reliable choice for the model.
- **k-Nearest Neighbors (kNN)** had lower median values for accuracy and Kappa compared to RF and LDA, with moderate variability, suggesting that while it is less complex, it may not capture the complexities of the dataset as effectively.
- **Support Vector Machine (SVMLinear)** had the lowest median accuracy and Kappa values, suggesting that it may not be as suitable for this particular dataset, despite its computational efficiency and lower variance.

5.4.2 Model Selection Decision

The decision on which predictive model to select was made with careful consideration of the observed trade-offs in performance. The Random Forest model emerged as our choice, driven by its superior predictive accuracy and Kappa scores. While it demonstrated some variability in performance, we determined that this could be managed with appropriate measures to prevent overfitting, such as regular cross-validation

and potential use of a validation set. The model's ability to handle complex interactions and nonlinear relationships within the data made it especially appealing for our objectives.

Linear Discriminant Analysis was also considered for its consistency and lower variability, but the need for the highest predictive performance led us to favor Random Forest. With the implementation of Random Forest, we anticipate better generalization to new data, which is essential for the practical application of our findings.

6 Results

Following the comprehensive data analysis and machine learning process, we have arrived at significant findings. The Recursive Feature Elimination (RFE) identified Weight, Age, Height, Frequency of Consumption of Vegetables (FCVC), and Number of Main Meals (NCP) as the top five predictors. Model comparisons demonstrated that the Random Forest algorithm achieved the highest accuracy and Kappa scores, suggesting its superior predictive ability, although it also exhibited the most considerable variability. Linear Discriminant Analysis (LDA) followed closely, offering more consistent results across evaluations. Conversely, k-Nearest Neighbors (kNN) and Support Vector Machine with a linear kernel (SVMLinear) lagged in performance, indicating a lesser fit for our dataset.

7 Conclusion

The project's analysis underscores the importance of thorough feature selection and model evaluation in predictive modeling. The identified key predictors contribute significantly to the understanding and prediction of the outcome variable. The chosen Random Forest model, despite its variability, is recommended for deployment due to its high performance, with the consideration that additional validation techniques should be implemented to control for overfitting.

8 Limitations

Our study acknowledges several limitations. The variability of the Random Forest model suggests a possible overfitting to the training data. Also, our analysis is limited to the variables provided in the dataset; there may be other unmeasured factors contributing to the outcomes. Moreover, the dataset size and potential biases within it can limit the generalizability of our findings.

9 Future Expansion & Recommendations

For future work, we recommend:

1. **Data Enrichment**: Including additional relevant variables to provide a more holistic analysis.
2. **Model Optimization**: Further hyperparameter tuning and exploration of ensemble methods to stabilize the Random Forest's performance.
3. **Exploring Alternative Models**: Investigating newer or more complex algorithms that might yield better performance or interpretability.

With these recommendations, we aim to refine the predictive models further and enhance their applicability to real-world scenarios.

10 References

- Aziz R, Sherwani AY, Al Mahri S, Malik SS, Mohammad S. Why Are Obese People Predisposed to Severe Disease in Viral Respiratory Infections? Obesities. 2023; 3(1):46-58.
- Fabio Mendoza, & Alexis Montas. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru, and Mexico. Data in Brief.
- Lam, B. C. C., et al. (2023). The impact of obesity: a narrative review. Singapore medical journal, 64(3), 163–171.
- Lafia, A., et al. (2022). Dietary habits, prevalence of obesity and overweight in developed and developing countries. Research, Society and Development. 11. e249111032769.