Estimation of Obesity Levels Based on Eating Habits and Physical Condition.

Realized by

Rajaa LEBNAITI

Date Last Edited: October 27, 2024

National School of Applied Sciences of Khouribga
5th Year: IT & Data Engineering

Abstract

Obesity has become a global public health concern, with an alarming increase in prevalence over recent decades. According to the World Health Organization (WHO), it is estimated that by 2030, more than 40% of the world's population will be overweight, and over a fifth will be classified as obese.

This project aims to estimate obesity levels based on individuals' eating habits and physical condition using machine learning techniques. A diverse dataset containing lifestyle and dietary attributes was preprocessed to handle both categorical and numerical variables.

Several models, including Logistic Regression, Random Forest, Gradient Boosting, and Multi-Layer Perceptron (MLP), were implemented to classify individuals into distinct obesity categories.

The performance of these models was evaluated using key metrics such as accuracy, precision, recall, and F1-score.

The Gradient Boosting model achieved the best performance with an accuracy of 96.89%.

These results demonstrate the potential of machine learning in predicting obesity levels and underline the significance of lifestyle factors in assessing health risks. The findings of this project could be instrumental in developing early detection and prevention tools to combat obesity.

Acknowledgements

I would like to extend my sincere thanks to HUBBLEMIND for providing me with the opportunity to work on this project. Their guidance and support have been instrumental in the completion of this work.

Lastly, I would like to acknowledge the availability of open datasets and machine learning resources that made this research feasible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Obesity is recognized as one of the most significant public health challenges of the 21st century. It has experienced a dramatic rise in prevalence across the globe, affecting both developed and developing countries. The World Health Organization (WHO) has projected that by 2030, more than 40% of the global population will be overweight, and over 20% will be classified as obese. Obesity is associated with an increased risk of numerous chronic conditions, including heart disease, type 2 diabetes, and certain cancers.

Given the growing burden of obesity on healthcare systems and its profound impact on individual well-being, early detection and intervention are crucial.

Advances in machine learning have opened new possibilities for predicting health conditions based on lifestyle and physiological data, offering promising tools for mitigating the risk of obesity and its related complications.

This project seeks to contribute to the growing body of research on the use of machine learning techniques to estimate obesity levels. By analyzing individuals' eating habits and physical conditions, we aim to develop predictive models that can assist in the early detection of obesity. This could pave the way for more personalized and timely interventions in combating the global obesity epidemic.

1.2 Objectives

The primary objective of this project is to develop a predictive model capable of estimating an individual's obesity level based on lifestyle and physiological data. Specifically, the project aims to:

- Analyze a dataset containing various attributes related to eating habits and physical condition to identify patterns linked to obesity levels.
- Preprocess the data by handling missing values, encoding categorical variables, and normalizing numerical features.
- Implement and compare the performance of multiple machine learning algorithms, including Logistic Regression, Random Forest, Gradient Boosting, and Multi-Layer Perceptron (MLP), in predicting obesity levels.
- Evaluate the performance of these models using accuracy, precision, recall, and F1-score to identify the most effective model for this task.
- Provide insights into how different lifestyle factors contribute to obesity

levels, which could inform future strategies for obesity prevention and management.

1.3 Scope

This project focuses on the use of machine learning techniques for predicting obesity levels based on a specific dataset that contains lifestyle and dietary attributes. The scope of this work includes the following:

- Data preprocessing, including handling of both categorical and numerical variables, to prepare the dataset for model training.
- Selection and implementation of four machine learning models: Logistic Regression, Random Forest, Gradient Boosting, and Multi-Layer Perceptron (MLP).
- Performance evaluation and comparison of these models using appropriate classification metrics.
- Interpretation of the models' results to provide actionable insights on the relationship between lifestyle habits and obesity risk.

The project is limited to the analysis of a single dataset, and its conclusions are based on the data at hand. As such, the findings may need further validation with additional datasets or in different populations. Additionally, the models are designed to assist in predicting obesity levels, and they do not replace clinical judgment or personalized medical advice.

Chapter 2

Literature Review

2.1 Previous Studies on Obesity

Obesity has been the focus of extensive research due to its growing impact on public health worldwide. Numerous studies have examined the causes, consequences, and prevention strategies for obesity. According to the World Health Organization (WHO), the global prevalence of obesity nearly tripled between 1975 and 2016. Factors contributing to this rise include changes in dietary patterns, reduced physical activity, and socioeconomic factors. The adverse effects of obesity extend beyond individual health, impacting healthcare systems through increased costs related to obesity-related diseases such as cardiovascular disease, diabetes, and certain cancers.

Research has shown that obesity is a complex condition influenced by genetic, behavioral, and environmental factors. For example, studies have identified strong associations between high-calorie diets, sedentary lifestyles, and the development of obesity. Behavioral interventions focusing on dietary

changes and increased physical activity have proven effective in reducing obesity rates, but their success depends on sustained long-term lifestyle changes, which are often challenging to maintain.

The use of predictive models in identifying individuals at risk of obesity is a growing area of interest. Early studies primarily relied on traditional statistical methods to analyze health data and predict obesity risks. However, with the emergence of more advanced computational techniques, researchers have begun to explore the potential of machine learning algorithms in improving the accuracy of obesity predictions.

2.2 Machine Learning in Health Predictions

Machine learning has gained increasing attention in healthcare, particularly in the areas of predictive analytics and personalized medicine. The ability to analyze large amounts of data and uncover patterns not easily discernible by traditional methods makes machine learning an ideal tool for health-related predictions. In recent years, machine learning has been applied to a range of health issues, from diagnosing diseases to predicting treatment outcomes. One of the key advantages of machine learning is its ability to handle complex, multi-dimensional datasets, which are typical in health research.

In the context of obesity prediction, machine learning algorithms have shown promise in accurately predicting obesity risks based on a combination of behavioral, dietary, and physiological variables. Models such as Decision Trees, Random Forests, and Neural Networks have been particularly successful in classification tasks related to health outcomes. The integration of

machine learning into public health initiatives offers the potential to identify at-risk individuals earlier and facilitate targeted interventions that could prevent the onset of obesity.

Several studies have successfully applied machine learning techniques to predict obesity based on lifestyle data. For instance, researchers have used Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), and Deep Learning methods to analyze data on dietary habits, physical activity, and metabolic factors to estimate obesity levels. These methods have demonstrated higher accuracy compared to traditional statistical techniques, highlighting their potential for clinical applications.

2.3 Related Work

Several studies have explored the application of machine learning in obesity prediction and health risk analysis.

The authors in Ref. [1] presented an intelligent method combining both supervised and unsupervised data mining techniques. Their approach integrated Decision Trees (DT), Support Vector Machines (SVM), and Simple K-Means clustering to classify obesity levels and provide recommendations for healthier lifestyles. Using a dataset of 178 students with 18 variables, they grouped the data into four categories. However, the distribution was imbalanced, with 69% of the participants classified as not prone to obesity and 31% classified as prone. The authors did not mention the use of any data balancing techniques. Their hybrid approach of DT and Simple K-Means achieved an accuracy of 98.5% and an ROC area of 99.5%, demonstrating

the potential of machine learning in detecting obesity levels [1].

In another study, Ref. [2] focused on predicting obesity risk in the Bangladeshi population using machine learning techniques. The dataset consisted of surveys from 1,100 individuals, with 28 factors identified as risk indicators for obesity. The data was divided into three groups: high risk (48%), medium risk (30%), and low risk (22%). Despite an evident class imbalance, the authors did not report using any balancing techniques. The models were trained using 80% of the dataset and evaluated using various supervised machine learning techniques, including k-Nearest Neighbors (k-NN), Logistic Regression, SVM, CART, Random Forest, Multi-Layer Perceptron (MLP), AdaBoost, and Gradient Boosting Machines (GBM). Logistic Regression performed best, achieving 97.09% accuracy, while GBM was the least effective with 64.08% accuracy and an F1-score of 57% [2].

The work in Ref. [3] introduced obesity estimation software based on the SEMMA data mining methodology, using Decision Trees, Bayesian Networks, and Logistic Regression. The dataset was drawn from a study involving 712 university students from Colombia, Mexico, and Peru, aged between 18 and 25. A survey was conducted to capture participants' physical characteristics, social habits, and other relevant factors. The models were validated using metrics such as Recall, true positive rate, and false positive rate. The Decision Tree model outperformed the others with an accuracy of 97.4%. The model was integrated into a desktop application, built with Java and supported by the Weka toolkit, allowing users to input their details to receive obesity predictions [3].

In a study focused on childhood obesity, Ref. [4] developed a predictive

model using Gradient Boosting Trees to estimate obesity risk in Israeli children. The dataset comprised 132,262 electronic health records, including demographic data, medical diagnoses, lab tests, and medication history, covering the years 2002 to 2018. The model was trained with data from the first two years of life to predict obesity risk at ages five and six. The Gradient Boosting Trees model achieved an area under the ROC curve (auROC) of 0.803 and an area under the Precision-Recall (auPR) curve of 0.312, showcasing its potential for early detection of childhood obesity [4].

Chapter 3

Methodology

3.1 Data Collection

The dataset used in this study for estimating obesity levels is publicly available and was obtained from the UCI Machine Learning Repository. The dataset comprises 2111 records of individuals, with each record containing 17 attributes related to personal characteristics, eating habits, and physical activity levels. The attributes include both categorical and continuous variables, such as 'Gender', 'Age', 'Height', 'Weight', 'Food consumption frequency', 'Physical activity frequency', and others. The target variable, 'NObeyesdad', classifies individuals into seven different obesity categories.

The data was collected through surveys designed to capture a comprehensive overview of participants' daily habits and physical conditions. These surveys were administered across various demographics to ensure a diverse sample population that could provide a broader insight into obesity patterns. The collected data was anonymized to protect participant privacy and

followed ethical guidelines for data usage in research.

To better understand the dataset and its distribution, Table 3.1 provides a summary of the attributes, including their types and descriptions. Additionally, Figure **??** shows the distribution of the target variable, indicating the number of individuals in each obesity category.

Attribute	**Type**	**Description**
Gender	Categorical	Gender of the individual (Male/Female)
Age	Continuous	Age of the individual (in years)
Height	Continuous	Height of the individual (in meters)
Weight	Continuous	Weight of the individual (in kilograms)
FAVC	Binary	Whether the individual frequently consumes high-calorie food (Yes/No)
FCVC	Continuous	Frequency of vegetable consumption in meals (Scale from 0 to 1)
NCP	Continuous	Number of main meals consumed daily
CAEC	Categorical	Food consumption between meals (No, Sometimes, Frequently, Always)
SMOKE	Binary	Whether the individual smokes (Yes/No)
CH2O	Continuous	Daily water consumption (in liters)
SCC	Binary	Whether the individual monitors their calorie intake (Yes/No)
FAF	Continuous	Frequency of physical activity per week
TUE	Continuous	Average daily time spent using technology devices (in hours)
CALC	Categorical	Alcohol consumption frequency (No, Sometimes, Frequently, Always)
MTRANS	Categorical	Mode of transportation (Automobile, Bike, Motorbike, Public Transportation, Walking)
NObeyesdad	Categorical	Obesity level categorized into 7 classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III

Table 3.1: Overview of Dataset Attributes

3.1.1 Source of Data

The dataset is sourced from a well-known repository (UCI Machine Learning Repository), which houses data collected from diverse populations. The original surveys aimed to gather information on individuals' dietary habits, physical activities, and health-related metrics. The use of publicly available datasets allows for reproducibility of the results and ensures the transparency of the research process.

Dataset: Download — Data Source: CC BY 4.0 — UCI Archive

3.1.2 Data Description

Each record in the dataset corresponds to a single individual and includes both input variables (such as height, weight, and food habits) and the target variable 'NObeyesdad', which categorizes obesity into seven classes ranging from "Insufficient Weight" to "Obesity Type III." The dataset is balanced to some extent across various categories, although some class imbalance is noticeable, as will be detailed in the following sections.

The next section will cover data preprocessing steps, including data cleaning, normalization, and handling of imbalanced data.

3.2 Data Preprocessing

In this section, we describe the steps taken to clean and prepare the data for modeling. Data preprocessing is essential to ensure the data is in the right format for machine learning algorithms and to improve the quality and performance of the models.

3.2.1 Handling Missing Values

One of the critical tasks in data preprocessing is managing missing data. Missing values can lead to biased estimates and inaccurate predictions. After inspecting the dataset, we found that there were no missing values, ensuring that the dataset was complete. In cases where datasets include missing data, common techniques such as imputation with mean/median values, or using advanced techniques like k-nearest neighbors (KNN) imputation, could be applied. However, since our dataset was complete, no imputation techniques were required.

3.2.2 Handling Duplicated Values

Duplicate entries in a dataset can distort the results of machine learning models by giving more weight to repeated observations. Upon examining the dataset for duplicated records, we identified 24 rows that were exact duplicates. These duplicate rows were likely due to data entry errors or redundancy during data collection.

To resolve this issue, we removed the 24 duplicated rows from the dataset to ensure that each observation is unique and representative. This step helped

improve the integrity of the data, ensuring that the model training process was not biased by repeated entries.

Duplicated rows:

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	\
98	Female	21.0	1.52	42.0	no	no	3.0	
106	Female	25.0	1.57	55.0	no	yes	2.0	
174	Male	21.0	1.62	70.0	no	yes	2.0	
179	Male	21.0	1.62	70.0	no	yes	2.0	
184	Male	21.0	1.62	70.0	no	yes	2.0	
209	Female	22.0	1.69	65.0	yes	yes	2.0	
309	Female	16.0	1.66	58.0	no	no	2.0	
460	Female	18.0	1.62	55.0	yes	yes	2.0	
467	Male	22.0	1.74	75.0	yes	yes	3.0	
496	Male	18.0	1.72	53.0	yes	yes	2.0	
527	Female	21.0	1.52	42.0	no	yes	3.0	
659	Female	21.0	1.52	42.0	no	yes	3.0	
663	Female	21.0	1.52	42.0	no	yes	3.0	
763	Male	21.0	1.62	70.0	no	yes	2.0	
764	Male	21.0	1.62	70.0	no	yes	2.0	
824	Male	21.0	1.62	70.0	no	yes	2.0	
830	Male	21.0	1.62	70.0	no	yes	2.0	
831	Male	21.0	1.62	70.0	no	yes	2.0	
832	Male	21.0	1.62	70.0	no	yes	2.0	
833	Male	21.0	1.62	70.0	no	yes	2.0	
834	Male	21.0	1.62	70.0	no	yes	2.0	
921	Male	21.0	1.62	70.0	no	yes	2.0	
922	Male	21.0	1.62	70.0	no	yes	2.0	
923	Male	21.0	1.62	70.0	no	yes	2.0	

Figure 3.1: Duplicated rows found in the dataset before removal.

By removing these duplicates, we ensured that the dataset's quality and accuracy were preserved, leading to more reliable model outcomes.

3.2.3 Encoding Categorical Variables

The dataset contains several categorical features, including variables like Gender, CAEC (food consumption between meals), CALC (alcohol consump-

tion frequency), and NObeyesdad (obesity levels). Machine learning models require numerical input, so these categorical variables were encoded into numerical representations. The following encoding techniques were used:

Label Encoding was applied to binary variables like Gender, family_history_with_overweight, FAVC, SMOKE, and SCC, where each category was replaced with a binary value (0 or 1).

One-Hot Encoding was applied to non-binary categorical variables such as CAEC, CALC, and MTRANS, where each category was converted into a separate binary column. This prevents the model from assuming any ordinal relationship between the categories.

Original Value	**Automobile**	**Bike**	**Motorbike**	**Public Transport**	**Walk**
Automobile	1	0	0	0	0
Bike	0	1	0	0	0
Motorbike	0	0	1	0	0
Public Transport	0	0	0	1	0
Walk	0	0	0	0	1

Table 3.2: Example of One-Hot Encoding for the 'MTRANS' (Mode of Transportation) Attribute

3.2.4 Outlier Detection and Treatment

Outliers are extreme values that can distort model training by affecting the performance of some machine learning algorithms. For continuous features like Weight and Height, we visually inspected the distribution using box plots to detect potential outliers.

In this study, outliers were capped using the IQR (Interquartile Range) method:

$$IQR = Q_3 - Q_1$$

$$\text{Lower Bound} = Q_1 - 1.5 \times IQR, \quad \text{Upper Bound} = Q_3 + 1.5 \times IQR$$

Values outside these bounds were treated as outliers.

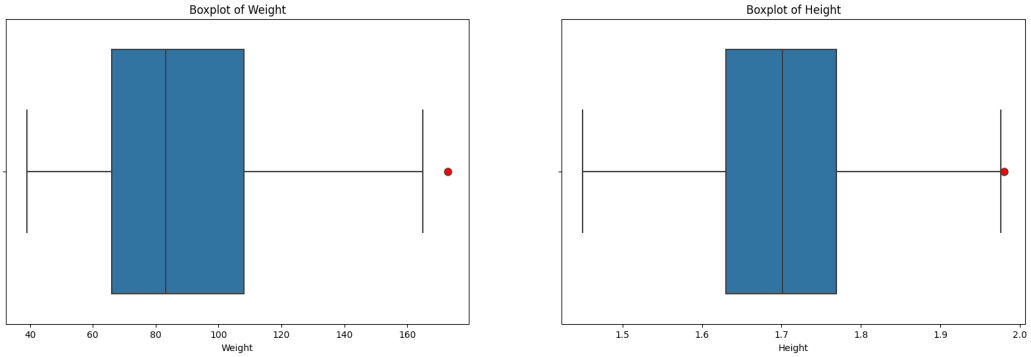

Figure 3.2: Box Plot for Continuous Features Highlighting Outliers

As shown in the figure above we found 2 outliers :

Weight Outliers Before Capping: 173.0

Height Outliers Before Capping: 1.98

Now, we will apply the capping method that enables us to replace the values of outliers with the value of Q3 (third quartil) or Q1 (first quartil)

Therefor, the values will be :

Weight Outliers After Capping: 171.039767

Height Outliers After Capping: 1.978461

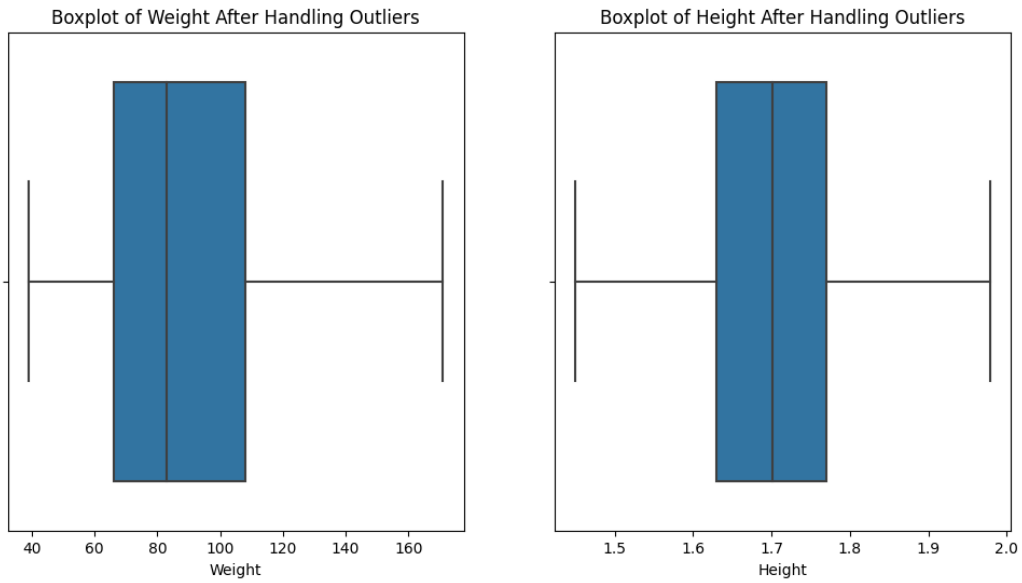

Figure 3.3: Box Plot for Continuous Features after Capping Outliers

3.2.5 Feature Scaling

In machine learning, especially for algorithms that rely on distance metrics, it is crucial to scale numerical variables.

The dataset contains several continuous variables, such as Age, Height and Weight, which need to be normalized to ensure equal weight during model training.

For this purpose, the MinMaxScaler was used, scaling the data to a range between 0 and 1:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

This transformation ensures that all features contribute proportionally to

the model training.

	Age	Weight	Height
count	2087.000000	2087.000000	2087.000000
mean	0.220279	0.362450	0.478131
std	0.135506	0.198333	0.176330
min	0.000000	0.000000	0.000000
25%	0.125871	0.204484	0.340949
50%	0.188247	0.333999	0.476069
75%	0.255319	0.522690	0.604570
max	1.000000	1.000000	1.000000

Table 3.3: Summary Statistics for Age, Weight, and Height after Scaling

3.2.6 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the patterns, relationships, and distributions in the dataset.

In this section, we visualize various features to gain insights and identify potential data preprocessing steps, such as handling outliers or scaling variables.

Visualizing the Distribution of Continuous Variables

To understand the distribution of continuous variables like Age, Weight, and Height, histograms and box plots were used.

These plots highlight any skewness or outliers that need attention.

For the Age distribution we found that:

- Skewness (1.51): This shows a right-skewed distribution, meaning the majority of data points are concentrated on the left, with a few higher values pulling the tail to the right.

- Kurtosis (2.77): A kurtosis greater than 0 indicates that the distribution has heavier tails and a sharper peak than a normal distribution, implying more outliers.

This distribution is not Normal, that's why we applied Square Root Transformation to reduce right skewness

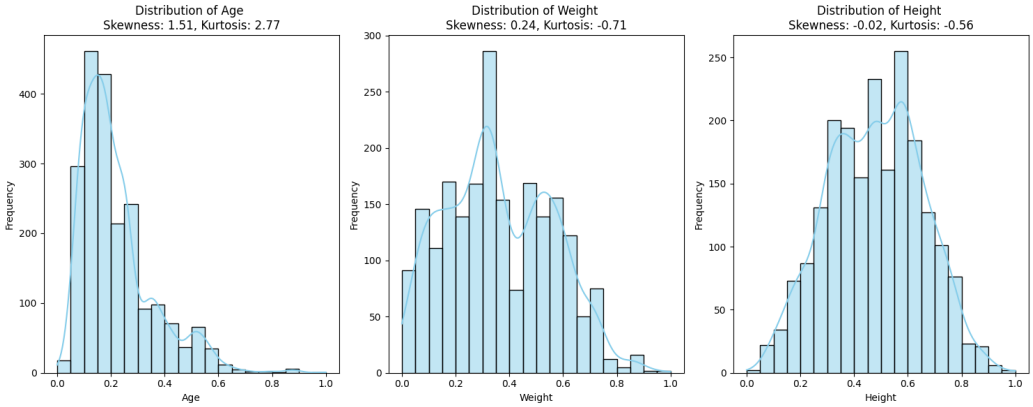

Figure 3.4: Histogram of Age, Weight and Height Distribution

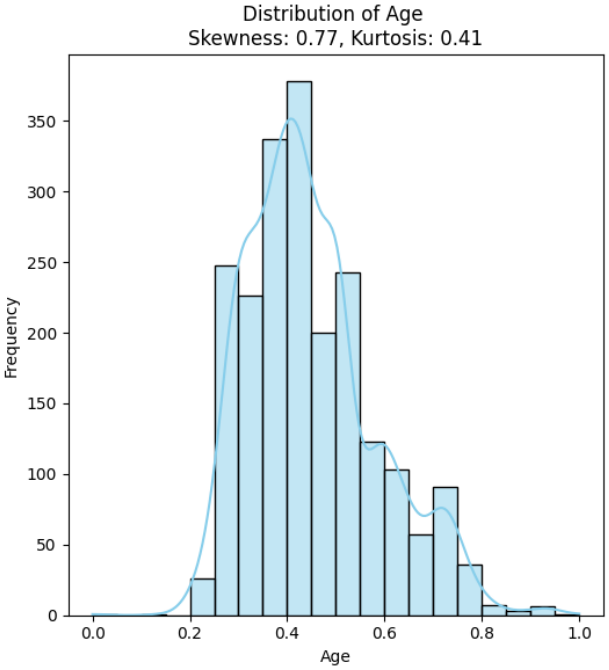

Figure 3.5: Histogram of Age Distribution after applying Square Root Transformation

Correlation Analysis

A correlation heatmap was created to examine the relationships between the continuous variables. Features such as Weight, Height, and Age were analyzed to identify strong correlations that might influence model performance.

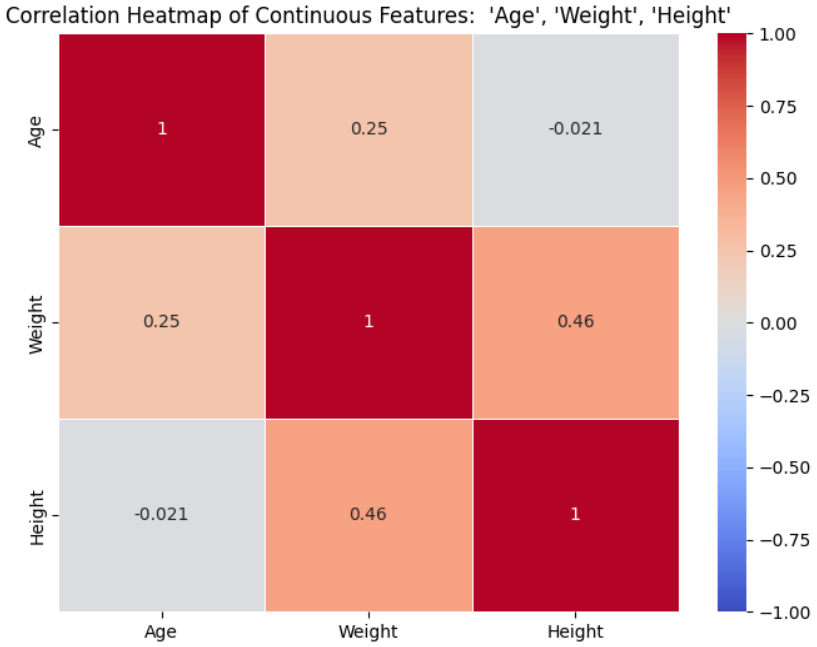

Figure 3.6: Correlation Heatmap of Continuous Variables

As seen in Figure 3.6, The strongest relationship here is between Weight and Height(0.46), which makes sense biologically. However, all other relationships are weak, indicating little to no linear relationship between Age and either Weight or Height.

BoxPlot of Weight and FAF vs NObeyesdad

In this part of the analysis, we visualized the relationships between Weight and Physical Activity Frequency (FAF) against the target variable NObeyesdad (representing obesity levels) using box plots.

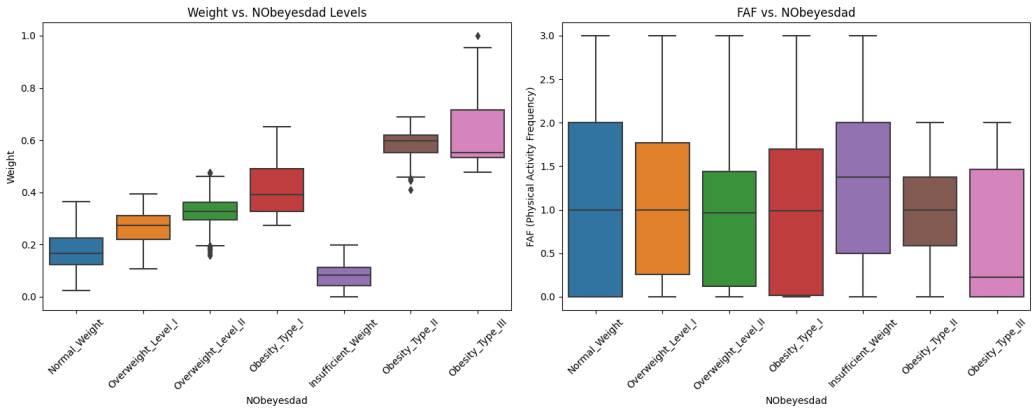

Figure 3.7: BoxPlot of Weight and FAF vs NObeyesdad

The first box plot illustrates the distribution of Weight across different obesity levels. As expected, the plot shows that individuals with higher obesity levels tend to have higher weights, with a clear upward trend in weight as obesity severity increases. The spread of values within each category also highlights the variability in weight across different classes.

The second box plot examines the relationship between FAF and NObeyesdad levels. It shows that individuals with higher physical activity frequency generally fall into lower obesity categories, while those with lower physical activity frequency are more likely to be in higher obesity classes. This confirms the expected inverse relationship between physical activity and obesity levels.

These visualizations provide insights into the correlation between weight, physical activity, and obesity levels, which will be valuable for feature selection and model training.

3.3 Model Selection

In this section, we prepared the data for model training by encoding binary categorical variables and splitting the dataset into training and testing sets. We then implemented four machine learning models: Logistic Regression, Random Forest, Gradient Boosting, and MLP Classifier.

3.3.1 Encoding Binary Features

To ensure our models can work with binary categorical variables, we applied label encoding to the following features:

family_history_with_overweight, *FAVC*, *CAEC*, *SMOKE*, *SCC*, and *CALC*.

Each binary feature was encoded into numerical values (0 and 1) using `LabelEncoder`.

3.3.2 Train-Test Split

The dataset was then split into training and testing sets using an 80-20 ratio, with 80% of the data used for model training and 20% reserved for testing. This ensures that the models are evaluated on unseen data to test their generalization performance.

3.3.3 Implemented Models

We implemented four models for predicting obesity levels:

- **Logistic Regression:** This is a simple yet effective classification model that estimates the probability of a categorical outcome based on the

logistic function. It is commonly used for binary and multi-class classification problems.

- **Random Forest:** A robust ensemble model that constructs multiple decision trees and combines their outputs to improve predictive performance and reduce overfitting. Random Forests are known for their flexibility and accuracy, especially in dealing with complex datasets.
- **Gradient Boosting:** An ensemble learning technique that builds models sequentially, with each new model correcting the errors of the previous ones. Gradient Boosting is effective at handling imbalanced datasets and capturing complex patterns in the data.
- **MLP Classifier (Multi-Layer Perceptron):** A type of artificial neural network that can learn complex non-linear relationships between inputs and outputs. The MLP consists of multiple layers of neurons and is trained using backpropagation to minimize error.

Each model was trained on the processed dataset, and performance metrics were calculated to assess their predictive capabilities. The results from these models will be discussed in the next section.

3.4 Model Evaluation and Results

To evaluate the performance of the models, we used four key metrics: Accuracy, Macro Average Precision, Macro Average Recall, and Macro Average F1 Score. These metrics give us a comprehensive understanding of the models' effectiveness across multiple classes.

3.4.1 Evaluation Metrics

Accuracy: Accuracy is the proportion of correct predictions made by the model out of all predictions. It is calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Precision: Precision measures the ability of the classifier to not label a negative sample as positive. Macro Average Precision is the unweighted mean of precision across all classes. The formula for precision is:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall: Recall, also known as sensitivity or true positive rate, measures the ability of the classifier to find all the positive samples. Macro Average Recall is the unweighted mean of recall across all classes. The formula for recall is:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure that takes both false positives and false negatives into account. Macro Average F1 Score is the unweighted mean of F1 scores across all classes. It is calculated as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.4.2 Results

The table below shows the results of our four models (Logistic Regression, Random Forest, Gradient Boosting, and MLP Classifier) based on the evaluation metrics discussed above.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.7368	0.7430	0.7368	0.7316
Random Forest	0.9450	0.9506	0.9427	0.9455
Gradient Boosting	0.9689	0.9726	0.9689	0.9689
MLP Classifier	0.9545	0.9529	0.9545	0.9535

Table 3.4: Model Performance Comparison

3.4.3 Discussion

From the results, we observe that Gradient Boosting achieved the highest accuracy of 96.89%, followed closely by the MLP Classifier with 95.45%. Random Forest also performed well, with an accuracy of 94.50%. Logistic Regression, while a simpler model, had a lower accuracy of 73.68%.

In terms of the macro-average metrics (precision, recall, and F1 score), the models followed a similar pattern, with Gradient Boosting consistently outperforming the other models. This suggests that Gradient Boosting was the most effective at distinguishing between the different obesity levels in the dataset.

Precision and Recall: High precision and recall values across the models indicate that they were good at correctly identifying true positives while minimizing false positives and false negatives.

F1 Score: The F1 score provides a balance between precision and recall.

A higher F1 score, as seen in the Gradient Boosting and MLP Classifier, indicates that these models maintain a good trade-off between precision and recall, making them reliable for multi-class classification in this dataset.

Chapter 4

Limitations and Perspectives

4.1 Limitations

Despite the overall success of the machine learning models in predicting obesity levels, several limitations should be noted:

- **Data Imbalance:** Some classes in the dataset had fewer instances compared to others, which might have affected the performance of the models. Specifically, the underrepresented classes could have led to lower recall rates for those specific obesity levels.
- **Limited Feature Set:** The dataset contains only a few key features, such as eating habits and physical activity. Other factors that may contribute to obesity, such as genetic predisposition or socio-economic factors, were not considered in the model.
- **Overfitting:** Some models, especially Random Forest and Gradient Boosting, could have suffered from overfitting due to the small size of

the dataset. This may lead to high accuracy on the training set but lower generalizability on unseen data.

- **Model Interpretability:** While complex models such as Gradient Boosting and MLP Classifier provided high accuracy, they are considered black-box models, making it challenging to interpret the decision-making process compared to more interpretable models like Logistic Regression.

- **Computational Resources:** The training of models like Gradient Boosting and MLP Classifier required significant computational resources, which could be a limiting factor in practical applications, especially for larger datasets or real-time use cases.

4.2 Perspectives

To address the limitations and improve future iterations of this work, several perspectives can be considered:

- **Data Augmentation:** In future work, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) could be applied to balance the dataset and improve the model's performance on underrepresented classes.

- **Feature Engineering:** Adding more features, such as demographic, genetic, and socio-economic factors, could enhance the model's ability to predict obesity levels more accurately. Additionally, leveraging do-

main knowledge for feature engineering may uncover relationships that current models do not capture.

- **Model Explainability:** To improve the transparency of the models, methods like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) could be applied to explain the predictions of more complex models.

- **Hybrid Models:** Exploring hybrid approaches that combine interpretable models (such as decision trees) with more powerful, black-box models (such as neural networks) could offer a balance between accuracy and interpretability.

- **Transfer Learning:** If a larger dataset on obesity is available, transfer learning techniques could be employed to leverage the patterns learned from that dataset, thus improving model performance and reducing overfitting.

- **Real-World Applications:** The models could be further adapted for integration into mobile health applications or public health platforms, providing personalized recommendations based on an individual's lifestyle data to prevent obesity.

Chapter 5

Conclusion

This project aimed to estimate obesity levels based on eating habits and physical conditions using several machine learning techniques. After conducting a comprehensive Exploratory Data Analysis (EDA), feature scaling, and data preprocessing, four machine learning models were trained: Logistic Regression, Random Forest, Gradient Boosting, and MLP Classifier.

The best-performing model, Gradient Boosting, achieved an accuracy of 96.89%, demonstrating its effectiveness in predicting obesity levels. The MLP Classifier and Random Forest also provided strong results with accuracies above 94%, while Logistic Regression performed reasonably well with an accuracy of 73.68%.

Despite the positive results, several limitations such as data imbalance, limited feature sets, and potential overfitting were identified. These challenges highlight areas for improvement and future research.

In terms of future perspectives, the application of techniques like data augmentation, transfer learning, and model explainability methods were pro-

posed. Furthermore, adding more diverse features and exploring hybrid model approaches could significantly improve the model's predictive capabilities.

In conclusion, this study demonstrated that machine learning models, particularly Gradient Boosting and MLP Classifier, hold significant potential in predicting obesity levels. The insights gained from this project could contribute to the development of data-driven solutions for obesity management and prevention, benefiting both individuals and public health systems.

You can find the GitHub project here: HEBBLEMIND - Estimation of Obesity Levels Based on Eating Habits and Physical Condition.

Bibliography

[1] R. Cañas Cervantes, U. Martinez Palacio, Estimation of obesity levels based on computational intelligence, *Informatics in Medicine Unlocked*, 21 (100472) (2020).

[2] F. Ferdowsy, K. Samsul, I. Jabiull, A machine learning approach for obesity risk prediction, *Current Research in Behavioral Sciences*, 2 (100053) (2021).

[3] E. De la Hoz Correa, R. Morales Ortega, F. Mendoza Palechor, A. De la Hoz Manotas, B. Sánchez Hernandez, Obesity level estimation software based on decision trees, *J Comput Sci*, 15 (10) (2019), pp. 67-77.

[4] H. Rossman, S. Shilo, S. Barbash-Hazan, N. Shalom Artzi, E. Hadar, R.D. Balicer, B. Feldman, A. Wiznitzer, E. Segal, Prediction of childhood obesity from nationwide health records, *J Pediatr*, 233 (2021), pp. 132-140.