**A black and white text on a black background

Description automatically generated**

**DS7010 2324 (T3) Data Science Dissertation (OC)**

**Dr Yang Li**

**Aysha STREETER**

**Estimation of Obesity Levels Based On Eating Habits and Physical Condition**

**Classification approach.**

# **Abstract**

This research focuses on the application of four techniques of machine learning: Logistic Regression, Random Forest, Decision Trees, and Support Vector Machines in analysing the level of obesity in relation to demographic and health-related data. This cross-sectional study employs a data set with 17 variables such as age, gender and lifestyle risks and divides people into 7 obesity status. The study targets to go beyond the limitations of skew statistics by applying the capacity of machine learning to capture interaction and non-linearity. In an overall manner, data preprocessing, exploratory data analysis and hyperparameter tuning help to increase model performance and decrease biases of the study. For example, Random Forest shows high predictive accuracy of the model, with distinct high obesity classes levels prediction. Explants the factors that worked out do contribute to obesity through feature importance analysis so that public health polices and targeted interventional strategies could be formulated and implemented. Altogether, the present work demonstrates that, even in conditions of data mismatching and model depraving, machine learning can be applied for creating accurate and evidence-based approach to obesity risk prediction and further population health management. In future work, critical steps will be made in order to address the aforementioned shortcomings and enhance the generalizability of the conceptualised models.

Table of Contents

[**Abstract** 2](#_Toc175944429)

[**1.Introduction** 5](#_Toc175944430)

[**1.1.Research Overview** 5](#_Toc175944431)

[**1.2.Research Justification** 7](#_Toc175944432)

[**1.3.Specific Importance of the Study** 7](#_Toc175944433)

[**1.4.Research Aim** 8](#_Toc175944434)

[**2.Literature Review** 8](#_Toc175944435)

[**2.1.Machine Learning versus Traditional Methods in Obesity Prediction** 8](#_Toc175944436)

[**2.2.Importance of Data Preprocessing in Machine Learning** 9](#_Toc175944437)

[**2.3.Feature Importance and Interpretability** 9](#_Toc175944438)

[**2.4.Challenges and Generalizability** 9](#_Toc175944439)

[**3.Methodology** 10](#_Toc175944440)

[**3.1.Dataset Overview and Justification** 10](#_Toc175944441)

[**3.2.Data Preprocessing** 10](#_Toc175944442)

[**4.EDA (Exploratory Data Analysis)** 11](#_Toc175944443)

[**4.1.Development of the model** 11](#_Toc175944444)

[**4.2.Hyperparameters Tuning** 11](#_Toc175944445)

[**4.3.Evaluation of the Model** 12](#_Toc175944446)

[**4.4.Comparison with External Baselines** 12](#_Toc175944447)

[**4.5.Discussion and Limitations** 12](#_Toc175944448)

[**5.Data Pre Processing** 13](#_Toc175944449)

[**5.1.Data Cleaning** 13](#_Toc175944450)

[**5.2.Defining outlier bounds using IQR** 13](#_Toc175944451)

[**5.3.Handling missing values and data types** 13](#_Toc175944452)

[**5.4.Applying the Analytics of Outlier Detection and Removal** 13](#_Toc175944453)

[**5.5.Data Standardization** 14](#_Toc175944454)

[**6.Exploratory Data Analysis (EDA)** 14](#_Toc175944455)

[**6.1.Histogram for Numerical Variables** 14](#_Toc175944456)

[**6.2.Boxplot for Numerical Variables** 16](#_Toc175944457)

[**6.3. Bar Plot of Categorical Variables** 17](#_Toc175944458)

[**6.4.Correlation Matrix** 19](#_Toc175944459)

[**6.5. Distribution Plots of Numerical Variables** 20](#_Toc175944460)

[**6.6.Distribution of categorical variables** 21](#_Toc175944461)

[**7.Model Training and Evaluation** 23](#_Toc175944462)

[**7.1.Performance Metrices** 24](#_Toc175944463)

[**8.Hyperparameter Tuning** 26](#_Toc175944464)

[**8.1Feature Importance Analysis** 27](#_Toc175944465)

[9. **One-vs-Rest Method for ROC and AUC Computations** 28](#_Toc175944466)

[**10.Evaluation and Summarization of the AUC Performance** 30](#_Toc175944467)

[**11.Conclusion** 31](#_Toc175944468)

[**12.Limitations** 31](#_Toc175944469)

[**13.Future Research** 32](#_Toc175944470)

[14.References 33](#_Toc175944471)

[Appendix 35](#_Toc175944472)

Table of Figures

[Figure 1 Histogram for Numerical Variables 14](#_Toc174482399)

[Figure 2 Boxplot for numerical variables 15](#_Toc174482400)

[Figure 3 Bar plot for categorical variables 16](#_Toc174482401)

[Figure 4 Scatter plot pair wise relationship 18](#_Toc174482402)

[Figure 5 Corelation Plot 19](#_Toc174482403)

[Figure 6 Distribution plot for numerical variables 20](#_Toc174482404)

[Figure 7 Distribution plot for categorical variable 21](#_Toc174482405)

[Figure 8 Model comparison metrics comparison 24](#_Toc174482406)

[Figure 9 Model accuracy comparison 25](#_Toc174482407)

[Figure 10 Model kappa comparison 26](#_Toc174482408)

[Figure 11 Feature importance 28](#_Toc174482409)

# **1.Introduction**

The World Health Organisation (WHO) reports that obesity has become a worldwide epidemic, affecting more than 650 million individuals globally. This condition is strongly associated with a variety of health problems, including as type 2 diabetes, cardiovascular illnesses, and some types of cancer. These health issues can result in higher death rates and significant expenses for healthcare. With the increasing prevalence of obesity, there is a pressing demand for improved and scalable prediction models that can accurately identify individuals who are at risk and enable early interventions.

# **1.1.Research Overview**

This study examines the utilisation of sophisticated machine learning algorithms, namely Random Forest, Logistic Regression, Decision Trees, and Support Vector Machines (SVM), for the purpose of forecasting obesity levels using demographic and health-related data. Although classic statistical methods have been used to tackle this problem, they frequently encounter difficulties in capturing the intricate, non-linear connections among various factors that contribute to obesity. This research intends to utilise machine learning, which is highly proficient in managing extensive datasets and revealing concealed patterns, to construct predictive models that are more precise and resilient.

**Dataset Description:**  
The dataset used in this study is sourced from the UCI Machine Learning Repository and contains information on 2111 records with 17 attributes. These attributes provide detailed information about the individuals, which can be used to categorize their obesity levels into seven distinct classes. The classification of obesity levels allows for a nuanced analysis of weight status, which is crucial for devising targeted public health interventions and strategies.

**Attributes Overview**

The dataset includes a variety of attributes, each contributing unique information relevant to obesity prediction and analysis. Here’s a detailed description of these attributes:

1. **Gender:**

**Type:** Categorical

**Description:** Indicates the gender of the individual (e.g. Male or Female). Gender can be an important factor in obesity studies due to differences in metabolic rates and health behaviours between genders.

1. **Age:**

**Type:** Continuous

**Description:** Represents the age of the individual in years. Age is a critical factor as metabolic rate and risk of obesity-related conditions can vary with age.

1. **Height:**

**Type:** Continuous

**Description:** Indicates the height of the individual in centimetres. Height is used in conjunction with weight to calculate Body Mass Index (BMI), a common measure of obesity.

1. **Weight:**

**Type:** Continuous

**Description:** Represents the weight of the individual in kilograms. Weight, along with height, is crucial for determining BMI and assessing obesity levels.

1. **Family History with Overweight:**

**Type:** Binary

**Description:** Indicates whether the individual has a family history of overweight (e.g. Yes or No). Family history can influence obesity risk due to genetic and lifestyle factors.

1. **FAVC (Frequent Consumption of High-Caloric Food):**

**Type:** Binary

**Description:** Indicates whether the individual frequently consumes high-caloric food (e.g. Yes or No). Diet is a significant factor influencing obesity.

1. **FCVC (Frequency of Vegetable Consumption):**

**Type:** Integer

**Description:** Represents the frequency of vegetable consumption per week. A higher frequency of vegetable consumption is typically associated with healthier eating habits.

1. **NCP (Number of Main Meals Consumed Daily):**

**Type:** Continuous

**Description:** Indicates the number of main meals consumed by the individual each day. Meal frequency and timing can impact weight management and overall health.

1. **CAEC (Consumption of Food Between Meals):**

**Type:** Categorical

**Description:** Indicates whether the individual eats between main meals (e.g. Yes or No). Snacking habits can influence calorie intake and weight status.

1. **SMOKE (Smoking Status):**

**Type:** Binary

**Description:** Indicates whether the individual smokes (e.g. Yes or No). Smoking has complex interactions with body weight and metabolism.

**Obesity Classification**

The dataset categorizes obesity levels into seven distinct classes, allowing for detailed analysis of weight status:

1. **Insufficient Weight:** Individuals with a weight lower than what is considered healthy for their height.
2. **Normal Weight:** Individuals with a weight within the healthy range for their height.
3. **Overweight Level I:** Individuals who are slightly above the normal weight range.
4. **Overweight Level II:** Individuals who are moderately overweight.
5. **Obesity Type I:** Individuals with a higher degree of obesity.
6. **Obesity Type II:** Individuals with severe obesity.
7. **Obesity Type III:** Individuals with morbid obesity

# **1.2.Research Justification**

The importance of this study resides in its ability to overcome the constraints of current approaches used to forecast obesity. Conventional methods, such logistic regression, have been extensively employed but may not effectively capture the complex relationships among numerous factors. Research has indicated that logistic regression offers a clear explanation of factors, but it may oversimplify the connections between variables, resulting in less precise predictions. Machine learning models have the ability to enhance prediction accuracy by modelling complicated interactions, without requiring explicit assumptions about the data distribution.

In addition, previous research has frequently concentrated on certain groups or restricted sets of factors, which may restrict the applicability of their results. This study aims to address this deficiency by utilising an extensive dataset that encompasses a diverse array of demographic and lifestyle characteristics that are recognised to have an impact on obesity. Utilising advanced machine learning techniques enables the detection of nuanced patterns and connections that may go unnoticed by traditional methods, rendering this research highly important in the development of more efficient public health measures.

# **1.3.Specific Importance of the Study**

This work enhances the field by specifically targeting significant shortcomings in existing approaches for predicting obesity. Firstly, it investigates the application of machine learning to represent non-linear correlations and interactions between predictors, which is a challenge for conventional methods. Additionally, it conducts a comparison examination of various machine learning models, pinpointing the most efficient algorithms for this specific activity. This comparative analysis is essential since it not only emphasises the advantages of machine learning over conventional methods but also provides guidance for future research and practical applications in healthcare.

Moreover, by the utilisation of a varied and all-encompassing dataset, this research augments the possibility to apply the results to a wider array of people, thus increasing the generalisability of the findings. This holds significant significance within the realm of worldwide public health, as obesity is an urgent concern among various demographic groups and geographical areas.

# **1.4.Research Aim**

The main objective of this study is to assess and contrast the effectiveness of several machine learning models in forecasting obesity levels. The precise aims are:

The objective is to create and assess the predictive precision of Random Forest, Decision Trees, and SVM models in categorising obesity levels using a wide range of demographic and health-related variables.

To determine the primary indicators of obesity using feature importance analysis, offering valuable insights into the most significant elements that contribute to obesity.

In order to evaluate the influence of data preprocessing approaches, such as outlier handling and data standardization, on the performance of the models.

The aim is to enhance public health policies by offering a strong, data-based method for predicting obesity. This can help direct specific interventions and allocate resources more effectively.

Overall, this study not only enhances the methodology for predicting obesity but also has the potential to greatly influence public health outcomes by enhancing the precision and dependability of risk assessments.

# **2.Literature Review**

Obesity is an escalating public health emergency, as indicated by the World Health Organisation (WHO) which states that global obesity rates have increased by over three times since 1975 (WHO, 2021). The consequences are significant, as obesity is a prominent determinant for chronic ailments such as type 2 diabetes, cardiovascular diseases, and specific types of cancer (Ng et al. 2014). With the rising occurrence of obesity, there is a growing demand for accurate prediction models that can identify individuals who are at risk and provide valuable information for public health strategies. Although logistic regression has been widely utilised, the emergence of machine learning (ML) has provided more advanced techniques that have the potential to enhance forecast accuracy.

## **2.1.Machine Learning versus Traditional Methods in Obesity Prediction**

Conventional statistical techniques, such logistic regression, have been widely used to forecast obesity and its associated health consequences. These methods are comprehensible and can be interpreted easily, but they are typically inadequate when it comes to handling intricate, non-linear connections and data with a high number of dimensions (Hosmer, Lemeshow, and Sturdivant, 2013). For example, the assumption of linearity in logistic regression can restrict its capacity to represent interactions between variables that may impact obesity, such as the interrelationship between physical activity, food habits, and genetic predispositions.

On the other hand, machine learning models, specifically ensemble methods such as Random Forests, have shown exceptional ability in capturing these intricate connections. In his 2001 paper, Breiman emphasised that Random Forests have the capability to mitigate overfitting and enhance prediction accuracy by aggregating the outcomes of many decision trees. Studies such as Martínez-Rosales et al. (2020) have confirmed that Random Forests surpassed traditional techniques in accurately forecasting obesity levels using demographic and lifestyle characteristics. This indicates that machine learning models have the potential to provide a more resilient and precise option compared to conventional methods, especially when dealing with complex and nonlinear datasets.

## **2.2.Importance of Data Preprocessing in Machine Learning**

Data preparation is an essential but frequently undervalued stage in the creation of machine learning models. Proper management of missing data, outliers, and feature scaling is essential for maintaining the reliable performance of models and avoiding biassed outcomes (García et al. 2016). For instance, employing the Interquartile Range (IQR) technique, as recommended by Tukey (1977), aids in eliminating outlier values that may affect the predictions of the model.

Although the research highlights the significance of these preprocessing processes, there is frequently a lack of uniformity in how they are implemented in different studies, resulting in variations in model performance. For example, Martínez-Rosales et al. (2020) utilised outlier elimination and normalisation strategies in their obesity prediction model, resulting in a significant improvement in the model's accuracy. This is in contrast to research that disregard these preprocessing processes, which generally leads to worse performance and limited applicability. The key point to remember is that thorough data preprocessing is just as crucial as selecting the ML method itself, and should be carefully executed to optimise model accuracy.

## **2.3.Feature Importance and Interpretability**

A primary critique of machine learning models, namely ensemble methods, is their perceived deficiency in interpretability as compared to conventional statistical methods. Nevertheless, progress in feature importance analysis has reduced this problem, enabling a more distinct comprehension of the variables that have the most substantial influence on predictions. In his work, Breiman (2001) highlighted that Random Forests offer a built-in metric for determining the significance of features. This metric is particularly important for comprehending the primary factors contributing to obesity.

Research conducted by Sefiane et al. (2021) has employed feature importance analysis to identify the most crucial characteristics in predicting obesity, such as levels of physical activity and caloric consumption. Within the realm of public health, it is especially crucial to have actionable insights in order to create specific interventions. Although logistic regression provides a clear interpretation, the ML models' feature importance offers more nuanced insights into the drivers of obesity. This thorough understanding is essential for developing effective health policy.

## **2.4.Challenges and Generalizability**

Although ML models have great potential, there are still considerable obstacles, especially when it comes to the capacity of these models to be applied to varied populations. Ogden et al. (2014) observed that the characteristics that can predict obesity differ considerably depending on demographic variables such as age, gender, and ethnicity. This makes it challenging to apply a single model to varied groups. This emphasises a crucial deficiency in the existing body of knowledge: the necessity for models that possess not only accuracy but also the ability to be applied to various subgroups within a population.

Furthermore, the problem of overfitting is a prevalent worry in machine learning models, particularly when working with limited or unbalanced datasets. While methods like cross-validation and hyperparameter tuning can help reduce these hazards, as explained by Hastie, Tibshirani, and Friedman (2009), the fundamental difficulty lies in finding the optimal trade-off between model complexity and interpretability. This is especially applicable in the area of predicting obesity, where overfitting can result in models that exhibit good performance on the training data but are unable to accurately predict new, unseen data.

In conclusion,

Overall, the transition from conventional statistical techniques to machine learning models in the prediction of obesity represents a notable progress in the field. ML models, namely Random Forests, have exceptional performance in managing intricate, non-linear connections and datasets with a large number of dimensions. Nevertheless, there are specific difficulties associated with this, specifically related to data preprocessing, model interpretability, and generalisability. The literature indicates that the effective utilisation of ML models necessitates careful consideration of data preprocessing and feature selection. Subsequent investigations should prioritise enhancing the applicability of these models and investigating sophisticated algorithms that might augment forecast precision without sacrificing interpretability.

# **3.Methodology**

This study utilises a meticulous approach to construct and assess machine learning models for forecasting obesity levels using a primarily artificial dataset. The methodology encompasses various stages, such as data preprocessing, exploratory data analysis (EDA), model creation, hyperparameter tweaking, and model evaluation. Each stage considers the distinctive characteristics of the dataset and the machine learning techniques employed.

## **3.1.Dataset Overview and Justification**

The dataset utilised in this investigation is predominantly artificial, which offers several possibilities as well as difficulties. The artificial character of the dataset enables systematic examination of feature correlations and model performance, a task that is challenging with real-world data due to the presence of noise and inconsistencies. Nevertheless, the dependence on artificial data requires a thorough assessment of its accuracy in reflecting reality and the possible constraints in applying the findings to real-life situations. The dataset contains a range of demographic and health-related characteristics, with the goal variable being "NObeyesdad," which classifies individuals into different categories of obesity. The target variable is a multi-class variable that includes categories such as "Insufficient\_Weight," "Normal\_Weight," "Obesity\_Type\_I," "Obesity\_Type\_II," and "Obesity\_Type\_III."

## **3.2.Data Preprocessing**

Preprocessing is crucial to guarantee data integrity and ready it for efficient model training. The subsequent actions were executed:

Dealing with Missing Values: Despite the fact that the dataset was artificially created, missing values were handled carefully. Imputation techniques were utilised where needed to assure data completeness without introducing any form of bias. The monitoring of missing values was conducted diligently; if the amount of missing data had above a crucial threshold (e.g., 20-30%), it would have caused issues regarding the accuracy and dependability of the dataset for predictive modelling.

Outlier detection and removal: Outliers were found using the Interquartile Range (IQR) approach and dealt with either substituting extreme values with NA or eliminating the rows that were affected. Ensuring the integrity of the models was essential by doing this step, as outliers can have a disproportionate impact on the outcomes, especially in models such as linear regression and decision trees.

## **4.EDA (Exploratory Data Analysis)**

An exploratory data analysis (EDA) was performed in order to get insight into the fundamental structure and distribution of the dataset.

Visualisations such as histograms, boxplots, and correlation plots were employed to examine the distribution of features and ascertain any connections between them. Particular emphasis was placed on the target variable "NObeyesdad" in order to comprehend the distribution of classes and any potential imbalances.

## **4.1.Development of the model**

Multiple machine learning models were created and assessed, considering their appropriateness for the dataset.

The Support Vector Machine (SVM) model was selected due to its efficacy in high-dimensional domains and its capability to manage non-linear interactions by utilising kernels.

A decision tree is a graphical representation of a decision-making process that uses a tree-like structure to model possible outcomes and their associated probabilities. A decision tree model was created to offer a straightforward and easily understandable model capable of handling both category and numerical variables. This model functioned as a reference point for comparison with more intricate algorithms.

The Random Forest algorithm was chosen due to its resilience, capacity to handle extensive datasets, and its efficacy in mitigating overfitting through ensemble learning. This model was especially suitable for the synthetic dataset, which could potentially include noise or irrelevant features.

The Multinomial Logistic Regression model served as a reference point for multi-class classification, enabling a comparison of its performance with more intricate models such as SVM and Random Forest.

## **4.2.Hyperparameters Tuning**

Hyperparameter optimisation was conducted to enhance the performance of the model, using the following methods:

A grid search was performed to find the optimal combination of hyperparameters for models such as SVM and Random Forest. The search for SVM focused on optimising factors such as the regularisation parameter (C) and the kernel coefficient (gamma). When tweaking a Random Forest model, it is necessary to adjust the number of trees (n\_estimators) and the number of features that are examined at each split (mtry).

Cross-validation: During hyperparameter tweaking, a 5-fold cross-validation procedure was employed to ensure the models' robustness and prevent overfitting to specific subsets of the data. This methodology yielded a dependable assessment of the model's efficacy on data that had not been previously encountered.

## **4.3.Evaluation of the Model**

The models underwent evaluation using multiple performance measures to ensure a thorough assessment:

A confusion matrix was created for each model to assess their ability in classifying different levels of obesity. This facilitated comprehension of the models' ability to differentiate among the various classes.

Accuracy, precision, recall, and F1-score are metrics used to evaluate the performance of a classification model. The metrics were computed to offer an equitable assessment of the performance of each model, specifically in dealing with imbalances in class distribution.

ROC and AUC Analysis: To evaluate the capacity of models such as Random Forest to distinguish between classes, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values were employed. This approach was essential in comprehending the trade-off between sensitivity and specificity in forecasting obesity trends.

## **4.4.Comparison with External Baselines**

In order to provide context and evaluate the study's significance, the performance of the models was compared to known benchmarks from previous research. This comparison served as a standard for evaluating the effectiveness of the models and identified areas where the study improved our knowledge of predicting obesity.

## **4.5.Discussion and Limitations**

The study conducted a thorough assessment of the use of a primarily artificial dataset, recognising its benefits in controlled experimentation while also admitting potential constraints in its applicability to real-world scenarios. The literature review examined the treatment of missing values, the effects of data imbalances, and the selection of machine learning algorithms. This analysis provided a strong rationale for the methodological decisions chosen in the study.

This section present the performance of different machine learning models towards the binary classification “Nobeyesdad” as the target variable. The forecasts are arrived at with the help of demographic and health characteristics. To increase the effectiveness of the training process, several manipulations with the obtained dataset were made; these manipulations included cleaning and normalising the data obtained, as well as removing any data that were not relevant to reaching the goal of the training process. To get a basic understanding of distribution of some of the variables and to assess their applicability, a preliminary data analysis was conducted. Four discrete models were trained using the dataset: Among them some prominent algorithms used in the analysis are Logistic Regression, Decision Tree, Random Forest, Support vector machine (SVM). We computed performance metrics that included accuracy, and Kappa as well as the confusion matrix. The results of the model were described in great detail, and various markers for each class together with accuracy and Kappa were given. Besides, hyperparameters increased Random Forest’s performances through an optimisation process which adjusts the model’s hyperparameters. Also, the performance of the feature selection was assessed with the use of ROC curves and AUC. A comparison analysis was done based on parameters such as accuracies other parameters including Kappa and other relevant parameters that were used to compare the performance of the models. In addition, we analysed the effects of hyperparameter tuning on the Random Forest model we chose for our study, and the results of feature reduction on this model.

# **5.Data Pre Processing**

First, the necessary environment is created, and preparations are made for data analysis, which is work from the very beginning of the code. These include, the process of R library installation, creation of necessary data frame and general structure of the code mainly entails installation of required libraries and loading of the data set. This would help to make sure that the necessary libraries that would enable manipulation, analysis and subsequent modeling, are accessible. To load these data from a CSV file and store them in an R object for sake of analysis.

## **5.1.Data Cleaning**

There are so much that can actually go wrong in data cleaning, thus in this stage, I ensure that the data is clean, complete and ready for analysis. This includes firstly the management of instances where values are missing, secondly, transformation of data types and last but not least, addressing instances of outliers.

### **5.2.Defining outlier bounds using IQR**

Following is the IQR method I apply to address outliers: But outliers are values that fall beyond the two standard deviations and are usually large or very small and can distort the results. The IQR method also assist in identification of these outliers since it sets limits beyond which various values are considered to be abnormal.

### **5.3.Handling missing values and data types**

Before addressing outliers, The dataset is properly formatted and missing values are managed:

**Imputing Missing Values**: With features, which are numeric and filled with question marks, I replace the question mark with the mean value of the particular feature. This is one of the usual imputation approaches that maintain the mean characteristic of the data. To handle missing values in categorical features, the mode of the respective feature is applied, so as to cater for consistent disposal of categorical data.

**Cleaning Data Types**: I transform columns into correct data type. For example:

Categorical Variables: Transforming categorical variables into factor variables although they are already in factor. These include Gender, family\_history\_with\_overweight and the others that refer to categories of variables.

Numeric Variables: Making sure that variables like Age, Height, Weight and among others are recognized as numeric. Collection of data is proactive, and thus meticulously setting a data type will ensure correct analysis and presentation of the result.

Removing Duplicate Rows: To get rid of the duplicate rows in my data set, I use the distinct() function as shown. These extra rows can be as a result of data collection or combining different datasets hence should be dealt with.

### **5.4.Applying the Analytics of Outlier Detection and Removal**

After preparing the data types and handling missing values, We proceed with outlier detection and removal:

Detecting Outliers: To the input data, I apply the IQR-based outlier detection on all numeric columns. This involves computation of upper and lower IQR for each numerical field and then find out values that violate this bounds.

Handling Outliers: For outliers it is replaced with NA(Not available). This helps me to mark such values for exclusion in the following stage so that they will not influence the analysis.

Removing Outlier Rows: In order to clean the data, used the NA substituted values and eradicate the sections containing the NA placeholder by incorporating the drop\_na() function. This step of data cleaning is done in order to remove the rows that could have been created during outlier handling and all other records containing missing values.

Reviewing the Cleaned Data

Once the data cleaning process is complete, reviewing the cleaned dataset to ensure that all issues have been addressed:

Displaying Cleaned Data: I further check the data by looking at the initial rows of the cleaned up dataset to ensure that the data has been cleaned appropriately. This assists in getting a confirmation on features such as I388 = 1 or I388 = 0 and I388 = ? values where missing data have been filled; Help in validating on how best the outliers have been handled and or managed; Ensures that the dataset is now clean and ready for further analysis.

Originally there were 2111 observation before data cleaning operation was conducted which after the process was done the observation was reduced to 1388.

## **5.5.Data Standardization**

Data normalization is an important data detoxification process, we do before feeding data into machine learning models. When transforming the numeric feature, it means that all features are made to have a maximum and minimum value of zero with a standard deviation of one hence making every feature to contribute equally for the model outcomes. This in turn increases the reliability of the data, also known to help models converge and makes results more interpretable. This operation can be regarded as the process of calculating the mean and standard deviation of the numeric columns and using the standardization formula for these columns, as well as checking the effectiveness of the transformation to prep the dataset for analysis and modelling.

# **6.Exploratory Data Analysis (EDA)**

Exploratory data analysis is a very important tool in carrying out data analysis as it aids in finding patterns of data, identifying outliers, testing hypothesis and checking assumptions using figures and measure of central tendencies and variability. In this analysis, I did EDA to get a general understanding of the description of features and examination of the distribution of features in the cleaned and standardized dataset. Here’s a detailed explanation of the EDA techniques I employed:

## **6.1.Histogram for Numerical Variables**

In order to be able to describe the spread of numerical variables in the given data set I have built histograms. Histograms, as are all other graphs, are pictures that depict the distribution of data values within certain intervals known as bins. Every bin is a couple of values and the height of the bar reflects how many observations fell into that bin. Through these histograms, it was possible to identify the spread and the skewness of the values and the position from the centre of the numerical data. It also allows distinguishing between such patterns as normal distribution, the presence of skewness or kurtosis, or multiple small peaks.

A screenshot of a graph

Description automatically generated

Figure 1 Histogram for Numerical Variables

Age, CH2O, FAF, FCVC, Height, NCP, TUE, Weight: In the following figures, these variables exhibit different degrees of distribution, and some of them are seriously skewed. The histograms reveal that:

Age and Height are less skewed than Weight and Length and have somewhat more normal distribution, although Age is right skewed.

Thus, we can establish that FCVC, NCP, TUE and CH2O exhibit many pronounced peaks at definite points, which shows that people usually have like values of these factors.

FAF and Weight are both leptokurtic variables which have most of their frequency at intermediary to high values with a fairly high degree of variation but with FAF being slightly positively skewed.

Peaks and Skewness: It is observed that most of the variables have considerable peaks, which imply that most of the people in the sample have small variations on these features. For example:

The first bar in the NCP plot of the meal variables has a very pointed top at zero, suggesting a fair number of people have a specific or narrow tolerance range in their food intake.

TUE and FCVC are reciprocally skewed and the majority of the data lie in the vicinity of specific averages, which suggests that the patterns are similar.

Variability: The variables such as Height and Weight on the other hand appear to be spread out and shows less defined peaks hence has more spread of the data point across their respective values.

## **6.2.Boxplot for Numerical Variables**

Another of the graphical displays of numerical data is the boxplot. It gives an overview of the tendency of data values, their dispersion and potential outliers. Each boxplot displays:

A graph with a grid and a diagram

Description automatically generated with medium confidence

Figure 2 Boxplot for numerical variables

Distribution and Spread:

The boxplots, in other words show the distribution of each variable, highlighting the median, the IQR, and if there are any outliers.

Weight, TUE, Height, FCVC, FAF, CH2O and Age are almost normally distributed which is supported by the fact that the median line is in the middle of the box. However, the dissemination of these variables is not confined in the same way to each other.

Outliers:

As it can be observed on the box plot below, NCP has many outgrowns, that is, data points that fall outside whiskers. It implies that there is general conformity in that majority of people take about three meals in a day, but there are extremes for a proportion of the population.

As is also seen, Height and Weight have a couple of extreme cases, albeit not as severe as what NCP has.

Symmetry and Skewness:

The variables seems to be majority fairly balanced apart from NCP which seems to be over dispersed most likely by the presence of outliers.

## **6.3. Bar Plot of Categorical Variables**

In categorical variables, bar plot is used to observe the frequency distribution for each category of variable. Here, each bar represents a category and the total count of the observations of each category are shown by the height of the bar. It is particularly useful in the case of demonstrating what kind of state categorical data are in, be it called skewed or biased, or that one particular category has the majority.

A screenshot of a graph

Description automatically generated

Figure 3 Bar plot for categorical variables

CAEC (Consumption of food between meals):

Most of the respondents are inclined to snack in between meals “Sometimes”, while the remaining few chose “Always” or “Frequently”.

CALC (Consumption of alcohol):

A higher number of respondents indicated they ‘Sometimes’ engage in alcohol drinking, while the least percentage of respondents revealed they do not take alcohol. ‘Frequently’ has the least representation out of all the categories.

Family History with Overweight:

one can notice that many people had experiences of having their families affected by overweight problems as suggested by the ‘Yes’ responses against ‘No. ’

• FAVC (Frequent consumption of high-caloric food):  
• Many participants answered ‘Yes’ to the question about the high-frequency consumption of high-caloric food, which might mean high prevalence of this behaviour in a given set of the data.  
• Gender:  
• As for the gender distribution the dataset is quite balanced although slightly more participants are male.  
• MTRANS (Transportation method):  
• The largest portion of respondents used the Public Transportation; Automobile is the second most used with biking and walking coming in as the least used transport means.  
• Nobeyesdad (Obesity levels):  
• There are definite groups of people according to weight: “Insufficient Weight” and “Overweight”, as well as different degrees of obesity. But, of all these categories, “Normal Weight” receives the least representation.  
• SCC (Monitoring of calorie consumption):

• Only a minority of the respondents track their calorie intake, which is evident from the score that has been assigned to the option, ‘No’.  
• SMOKE:  
• A greater number of respondents indicated that they are non-smokers even though a small portion of them is a smoker.  
  
In the present analysis, a two-dimensional scatter plot of the inter-individual relationships has been created.  
Where the two numerical variables are related between the pairs, I employed scatter plots. In the scatter plots, there are points that depict the values of two numerical variables. Thus, by making these plots, I can be able to identify a pattern or relationship between different variables.  
A screenshot of a graph

Description automatically generated

Figure 4 Scatter plot pair wise relationship

• Age has a positive relationship with height significant at (0.136) and weight at (0.306) and equally has a negative relationship with NCP (-0.090), FAF (-0.200) and TUE (-0.187).  
• There is significant positive relationship between Height and Weight measures as will be seen from the coefficient of 0. 440.  
• Weight is positively related with CH2O with a coefficient 0. 241 as well as negatively related with TUE with a coefficient of -0. 126.  
• FCVC has a positive association with Height (0. 277) and CH2O (0. 145), however, a small negative association with NCP (-0. 055) and TUE (-0. 108).  
• Thus, the values of NCP are in a positive way related to CH2O (0. 145), and FAF (0. 124), if being compared with the correlations as shown on the above table, the correlations are either weak or negative.  
• FAF is positively and weakly related with Weight (0. 239) and CH2O (0. 145).

## **6.4.Correlation Matrix**

Even though the correlation matrix represents how numerical variables stand in terms of linear associations. It gives a table of correlation coefficients which represent the degree and direction of relationship of two variables at a time. The positive sign shows that if one of the variables increases, the other also increases and vice versa the negative sign shows the vice versa.

A diagram of a graph

Description automatically generated with medium confidence

Figure 5 Corelation Plot

• **Strong Positive Correlations:**

Height and Weight: The largest circle with a black interior symbolizes the largest positive correlation that is between Height and Weight.

Weight and CH2O (Water Intake): It is also possible to notice a positive relation between Weight and CH2O, which is marked with a moderate sized blue circle.

Height and CH2O: Quite a small yet positive relationship between the Height and CH2O shortens is seen.

• **Moderate Positive Correlations:**

FCVC (Vegetable Consumption) and Height: There is virtually no correlation between them but they positively relate hence the blue circle.

FAF (Physical Activity Frequency) and CH2O: Likewise favourably associated, it is shown with a medium blue circle.

FAF and Weight: Has a positive correlation which is fairly significant.

• **Negative Correlations:**

Age and FAF (Physical Activity Frequency): A very important negative association is represented by a large red circle.

Age and TUE (Time Using Technology): Second strong negative correlation, indicated by large red circle.

TUE and Weight: Indicates a negative relationship, marked by a small red dot.

• **Weak or No Correlations:**

NCP (Number of Main Meals) and other variables: These or no significant are close are same due to small size and lighter colors as represented below Where There is a slight negative and weak correlation which is represented by dark big circles.

TUE and FCVC, NCP, CH2O: The four pairs depicted here show very low levels of relationship.

## **6.5. Distribution Plots of Numerical Variables**

Besides, histograms, density plots were applied to illustrate the distribution of numerical variables. Density plots offer some outline of the shape of the data and how it is distributed, in terms of greater or lesser density.

When using the density plots it is possible to find out that many of the variables including Age, FCVC, TUE, and Weight have bimodal distributions which mean that there are two main groups of subjects in the dataset.

A graph of a graph

Description automatically generated with medium confidence

Figure 6 Distribution plot for numerical variables

• Others are more continuous such as Height and CH2O, which have normal distributions, that is most values are located at the central point.

• The histogram of FAF is less distinct as a normal distribution, as we see more modes of it, which indicates a greater variety of the level of physical activity.

• NCP has the highest central tendency with a very low variability which depicts that there are little changes in the number of main meals in a day by people.

## **6.6.Distribution of categorical variables**

In the case of categorical variables, I have produced bar plots to illustrate each category distribution. This plot is useful for countering totals for each category and to observe if there are any shifts to a certain pattern within the given data.

A screenshot of a graph

Description automatically generated

Figure 7 Distribution plot for categorical variable

• CAEC: This variable seems to have subcategories such as ‘Always’, ‘Frequently’, and ‘Sometimes’ of which ‘Sometimes’ seems to be more frequent.

• CALC: Like the table CAEC, the frequency for ‘Frequently’ is higher while the ‘Sometimes’ is also present but the possibility of having other categories is limited.

• family\_history\_with\_overweight: A variable defined by two possibilities ‘no’ and ‘yes’ with the greatest proportion of responses in the ‘no’ category.

• FAVC: Another variable, with options as “Nobodiesure,” “yes,” where “yes” had an even higher tally than “Nobodiesure. ”

• Gender: A variable that has two possibilities: ‘Female’ and ‘Male’; the numbers in each category are approximately equal.

• MTRANS: Such labels as ‘Automobile’, ‘Bike’, ‘MotoPublicTransportWalking’, where ‘Automobile’ is the most used.

• SCC: A nominal variable which is a binary variable that has two categories namely “No” and “Yes” and was found to have a very high frequency in the “no” category.

• SMOKE: Example of a binary variable with two categories ‘no’ and ‘yes’ where there are more responses in ‘no’ category than the ‘yes’ category.

• Nobeysedad: This variable has a number of values, but one of them is distinguished by a much greater frequency.

# **7.Model Training and Evaluation**

As the starting point in my analysis, I proceeded to train and test a multinomial logistic regression model for “NObeyesdad” in the data set. I then did train Test split, for training 80% and for testing 20%. I used the trainControl function in order to set up the model training process using 5 fold cross validation method. Using k-fold cross validation approach, the performance of the model was thoroughly validated to reduce on overfitting by training and testing the data on different folds.

When I identified the control parameters, I fit the multinomial logistic regression model to the training data. Training of the model was done with the train function from the caret package, setting multinom as the method argument. After training the model, I applied it on the test data in order to see the results I obtained. In order to maintain a consistant comparison for the evaluation, I also converted the predicted labels and the true labels in the test set into factors. I then lined the levels of the predictions to the levels of the actual test labels as a way of extending the comparison.

In the next step, I performed Precision, Recall and F1 score of each class of the dataset. These were done for each unique class in the “NObeyesdad” variable applying the Precision, Recall, and F1\_Score functions retrieved from the MLmetrics package. I created special vectors for all of these values, and then concatenated them into a data frame – this approach let me evaluate the model’s performance class by class.

In order to evaluate the model even more, I went ahead and created the confusion matrix using confusionMatrix function. It enabled me to plot the results as a function of the model’s prediction against the real labels and this would show zones where the model was good and the zones that were not good.

Staying with the same procedure, I then proceeded to decision tree modeling and used the rpart method of training. Subsequently, the same training data was used to train the decision tree model where predictions where made on the test data. Once more, the predictions and true labels were transformed into factors and the levels compared to the other. I then proceeded to check the precision, recall and F1-Score which were also observed for the logistic regression model. These metrics were then integrated into a data frame and the next thing and a confusion matrix was created in order to determine on the performance of the model.

Subsequently, I tested a random forest model with the rf method. Random forests have one of the best properties when it comes to large feature set and model robustness so I had high expectations for this one. In the process of making prediction in the test data, I maintained the factor conversion of the training as well as level matching. So, I calculated the precision, the recall, and the F1 score for every class and present the results in the table format. The confusion matrix offered more information on the capacity of the model on the classification of the data.

Finally, SVC model was trained using svmLinear method. SVMs are very useful in classification problems and more so in data sets with high dimension. After training the SVM model, I predict the test data labels and did the same work of factor conversion and level matching. I calculated the precision, recall, and the F1 score for each of the classes and create a data frame with that information and create a confusion matrix.

While doing this, I took my time to record each phase of the process in a bid to ensure that the models were trained and tested equally. They offered an insight into the performance of each model on the “NObeyesdad” classification that would haven been difficult to achieve from the global accuracy scores Thus, the per-class metrics and the confusion matrices were very useful in comparing the performance of the various models in the classification task.

# **7.1.Performance Metrices**

As the last step of the analysis, after training and testing the multinomial logistic regression, decision tree, random forest and SVM models on “NObeyesdad” dataset, I calculated basic metrics from the confusion matrix of each model that were chosen by previous studies. Two measures that I chose were the general accuracy of the models and Kappa statistic because they give information about the predictive accuracy of the models.

Letting this process begin with extracting the metrics, I ran the overall statistics of each confusion matrix: ``` Overall stats of logistic regression confusion matrix: ``` Overall statistics were the accuracy which gives the ratio of correctly classified instances to the total number of instances and the Kappa statistic which corrected for chance agreement.

After that, I combined these metrics into a data frame whereby closeness was easier to compare. Particularly, I built a data frame containing the names of the models, their accuracy and the correspondent Kappa coefficients. This let me switch between each model and see how well each one did in terms of these criteria.

A graph of red and blue bars

Description automatically generated

Figure 8 Model comparison metrics comparison

Translating these comparisons are easy since I employed the `ggplot2` package to work on the visual representations. First, I needed to pivot the data frame to plot multiple metrics, accuracy and Kappa in our case, on the same graph and this will be accomplishing by using the function `pivot\_longer`. This I followed by charting a bar plot depicting the performance of each model as it was in terms of the mentioned metrics with both accuracy and Kappa for each model placed right beside the other. This over-relational comparison elucidated the virtues and vice of each model, thus allowing an easy and quick assessment of their performance.

A graph of different colored rectangles

Description automatically generated

Figure 9 Model accuracy comparison

Besides the metric plot of the combined metric, I also built the bar plots of accuracy and Kappa. In the case of the accuracy plot, the focus was placed on how well each one of the models performed in terms of classification, in percentage while the Kappa plot aimed at depicting how much better each of the models was in comparison to a mere guess. These individual plots enabled me to study the details of model performance, and see which of the models was exceptionally accurate overall or which one was consistently strong on Kappa value, meaning that it would not suffer much even with presence of class imbalance.

A graph of different colored rectangles

Description automatically generated

Figure 10 Model kappa comparison

By dissecting the models into the different aspects in this paper, was able to give each model the fair comparison that they deserved. As for the Mixed Models, numerical measures and graphs are presented to allow quantitative and qualitative comparison between the models’ performance and with the previous model, so that the relative advantages and disadvantages can be evaluated accurately. It was this approach which informed my understanding of the models’ ability to predict on the ‘‘NObeyesdad’’ classifcation task.

# **8.Hyperparameter Tuning**

**Random Forest Hyperparameter Tuning**

In order to improve the Random Forest model on the “NObeyesdad” dataset, I carried out the hyperparameters tuning by the process of grid search. It is important to learn to tune hyperparameters because this step helps to adjust different parameters that enhance the model improvement and its ability to generalize the data.

To start with, I described a list of hyperparameters that I wished to tune for Random Forest model. First and specifically, I sought to use the ‘mtry’ parameter, which fixes the number of features to be used in a randomly made split of decision trees. To the grid I decided to assign the following values: 2, 4, 6, 8 and 10. Such values range let me model all the possible levels of feature sampling systematically in terms of impact on the model.

To make my results replicable, I used the command `set. seed(123)`. It was important to make this step because both Random Forests algorithms have inherent randomness in them, owing to the way the data is sampled at every iteration, as well as the selection of features. Having set the seed, I ensured that the tuning as performed produced the results as if the same exercise were to be repeated.

Subsequently, I also conducted the Al Grid search using the ‘train’ function from the caret package. In this function, I have mentioned the response variable as NObeyesdad and also given the training data. As the model method which I chose “rf” stands for Random Forest, it is an algorithm that is used. The `tuneGrid`argument was set to the grid of ‘mtry’ values that previous to this I had specified in order to facilitate the training and testing of the model against each one of these values. I also passed the ‘ctrl’ object that I set up to be able to cross-validate the model. Also, I made some modifications on the code to include the calculation of feature importance by passing the `importance` argument as `TRUE` so that after developing the model, I would be able to know the most significant features for the model.

Following this, I used the command, view\_model to display the `DT` model performance based on every `mtry` value used in the grid search. This output have given me information about which among the values of `mtry` setting gave the best model performance according to the cross-validation.

By following this systematic way of hyperparameter tuning, I was successful in determining the best suitable combination that can be applied on this Random Forest model of this dataset. This step not only increases OF in terms of accuracy but also designation of the most significant features which are influencing the obesity level prediction and, therefore, enhances the credibility of the final model.

Hyperparameter tuning refers to the process of adjusting the parameters so as to determine the right ones for the model to function properly. This allows an initial model to be trained by using the right hyperparameters.

I started with a hyperparameters grid that allows to optimise the Random Forest model. In more detail, I concentrated on the mtry parameter that define number of features at each split randomly selected. I used 2, 4, 6, 8 and 10 as the values for mtry in order to achieve a good selection of feature sampling methods a model can embody.

So that the results obtained would be equal to the previous work with the programming language and could be repeated, the seed was put to 123. With the seed in place I use the following code to perform a grid search with the train function of the caret package The code is as follows: This function enabled me to fit the model on the training set and at the same time evaluate the model for different mtry values in the specified search menu. The trControl argument was set to cross validation so as to ensure that the derived model was checked thoroughly on distinctly partitioned datasets.

Pursuant to the grid search, I distinguished between the mtry values and printed out which of them provided the best performance. This initial tuning created a foundation that could be fine-tuned by finding out the best setup for Random Forest model.

# **8.1Feature Importance Analysis**

The next step was to identify which particular features were most important in the ‘‘prediction. ’’ According to the result of the feature importance, I could determine those which put the greatest emphasis on the level of obesity. Finally, using the best fit model, I was able to get the importance of each of the features and I created another data frame to arrange all this information in a neat format.

A graph with a red and blue bar graph

Description automatically generated

Figure 11 Feature importance

In order to emphasize the necessity of these features, a bar plot was designed. This plot provided a relative view of the hierarchy involving features in an attempt to also identify the most significant features of the ‘Overweight Level II’ class. The usage of the blue to red colour schema helps to show the transitions of the importance of the features where the features that had unique value, which was crucial for the model, were easier to identify.

Making Probabilities and First Arrangement

The Random Forest model predicted probabilities for each class can be generated first. These probabilities are the result of the model to which extent it believes in the assignment of an observation to the particular class. Before carrying out the analysis, it is crucial to check that the “Nobeyesdad” target variable in the test dataset is treated as a factor which I did as shown below.

# 9. **One-vs-Rest Method for ROC and AUC Computations**

Since the “Nobeyesdad” variable is multi-class, I proceeded to conduct one versus rest to develop the ROC curve and AUC for every distinct class. This is a strategy where by each class is assumed to be the positive and all others are grouped and assumed to be the negative class. In doing so, I was able to create a binary classification situation for every grid, and this is something that must be done in order to compute an.

**Looping Through Each Class**

In order and to obtain the ROC curve and AUC for each class, I used a for loop to go through different levels of “Nobeyesdad” variable. For each class:

1. Binary Response Creation: To do this I generated a binary response vector in which the current class has a value equal to 1 while all other classes have a value of 0.

2. ROC Curve Calculation: I also computed the ROC curve using the above defined binary response vector and the predicted probabilities of the current class. This curve compares the true positive rate or sensitivity at a range of cutoffs with the false positive rate which is equal to 1 minus specificity or (1- specificity).

3. AUC Calculation: The AUC which is an overall measure of the performance of the classification model over the two classes of positive and negative classes was taken from the ROC diagram. The AUC value is always between 0 to 1 inclusive; therefore a higher value closer to 1, signifies a better model.

4. ROC Curve Plotting: As a result, I aimed at visualizing the model’s results through plotting ROC curve for each class. From this graphic representation one could identify the relationship between sensitivity and specificity depending on the thresholds set.

A graph of a curve

Description automatically generated with medium confidenceA graph of a normal weight

Description automatically generated

A graph of a curve

Description automatically generatedA graph of a curve

Description automatically generated

**A graph of a curve

Description automatically generated with medium confidenceA graph of a function

Description automatically generated**

**A graph of a function

Description automatically generated**

# **10.Evaluation and Summarization of the AUC Performance**

I collected all the AUC values from each class and summed them together and put them into a data frame format. The AUC values which were given in the evaluation also served as means assessing to what extent the model was successful in terms of differentiating each level of obesity from each other.

In our recent analysis, we crazy over its classes performance measures, where we checked the Area Under the Receiver Operating Characteristic Curve (AUC) for the each classes of the dataset. Some of the important metrics largely associated with the performance of the equation include the ‘‘Area under the Curve’’ (AUC), with larger values of the curved area suggesting better class separation.

In the specific class “Insufficient\_Weight,” the model discovers extraordinarily high AUC of 0. As it can be seen, the model performed powerfully in correctly categorizing this type of articles, with a unique identification number of 9991324. In the same way, ‘Normal\_Weight’ class was also predicted fairly well with the AUC of 0. 9982616, higher accuracy noticed with normal weighted instances for different classes.

The clearly impressive performance was even more so for the obesity-related classes. Towards “Obesity\_Type\_I” the AUC was equal 0. It was slightly higher at 0. 999889 in the case of ‘Obesity\_Type\_II’. 9998998. From these results, the model’s great capacity of discriminating between these two types of obesity and the non-obesity cases is evident.

The best AUC score was observed with a variable “Obesity\_Type\_III” which obtained the maximum score, 1. Such classification scores to 0000000 meaning that this category has been classified to perfection by the algorithm. Closely behind is the classification “Overweight\_Level\_I” and “Overweight\_Level\_II” which has brought 0. 9992991 and 0. 9990237, respectively. As pointed out by these findings, the use of this model is effective in classifying people into different weight and obesity range with minimal classification errors.

Combined, the AUC values of all categories confirm good performance of the proposed classification model and will offer high accuracy of categorization in the given set of cases.

# **11.Conclusion**

The work achieved its aim of implementing and assessing various machine learning algorithms for the assessment of obesity levels according to demographic and health factors. Among all the applied models, the Random Forest model, possessing the optimal hyperparameters, showed the highest results for the classification connected with obesity in various categories. The relatively higher AUC values especially for Obesity\_Type\_III show the reliability of the model to predict different weights and obesity.

Outlier handling and normalization of the feature values, which formed the part of data preprocessing step made a profound impact towards the improvement in the performance of the models. The EDA was useful in explaining some features of data they were dealing with as well as desirable relationships between variables for further modelling.

In total, it can be concluded that based on the data on demographic and health indicators, machine learning models, in particular, Random Forest, can be used to predict obesity levels with sufficient accuracy. It is for these reasons that these models can be useful in the creation of tailor-made health promotion activities and the formulation of the population specific health strategies for managing obesity and related complications. The future work could be the extension of the feature integration, other advanced models exploring could be done and these models could be tested on even larger and more variant data sets to confirm the findings.

# **12.Limitations**

Imbalanced Data: There is also a way that this study could be quite biased with imbalance in the type of organizations and patients to be included in the database. When several categories of obesity levels were provided, there might have been a problem of class imbalance, that is, some of the obesity levels might have had fewer samples than others thus causing the problem of under-fitting, hence the model’s poor ability to generalize well to under-represented classes. This is important because the performance metrics such as the accuracy may be skewered especially if the model has a bias of predicting a single class.

Outlier Handling: For outlier treatment in the study, the IQR method was used in the elimination of outliers. On the other hand, this approach can be relatively simple and may lead to removal of possibly useful far-out observations. There could be other technical ways of coming up with better results like using robust statistics methods or domain-specific thresholds that could possibly give better results but would not compromise on information richness.

Feature Engineering: Although the feature importance was studied, the study may have had lacunae for more elaborate feature engineering. Polynomial terms, feature engineering or domain transformations may still be used to improve the model’s accuracy.

Generalization to Diverse Populations: A possible limitation that could be pointed out while using the study is that the sampling may not necessarily provide a comprehensive view of the population, regarding such factors as ethnicity, geographical location, or socioeconomic status of the patients. In others, therefore, getting gait from the models developed might not generalize to other populations different from those used in getting the samples.

Model Complexity and Interpretability: As with the Random Forest model, the results were good but being a ‘black box’ there is a lack of clear interpretation of the results obtained. In clinical applications, simpler models such as logistic regression even though suboptimal might be the go-to models since they are easy to explain.

Overfitting: This was addressed by cross-validation but can still be observed, especially given the Random Forest type models which may overfit especially if not carefully handled. These AUC values approach 1, suggesting the model may be overfit, meaning less accuracy when applied to unseen data.

# **13.Future Research**

Addressing Data Imbalance: More research can be conducted in the area of using SMOTE or cost-sensitive learning to reduce issues of class imbalance. This could enhance the accuracy for underrepresented subclasses and reduce odds estimation unreliability.

Exploring Other Models: Continuation of this study could be an attempt to explore more complex models, for instance deep learning models that possibly could derive more detailed characteristics of the patterns prevailing in the examined data. Other than Random Forest, other ensemble learning could be in use like XGBoost, LightGBM could be in use and probably yield better performance.

Feature Engineering and Selection: Some of the future work could include applying feature engineering on higher level than used in this work for example we could try to build interaction terms or apply PCA to get more meaningful features. Also, more advanced feature selection techniques may also be used to enhance the model’s performance even more.

Validation on Diverse Populations: Further analysis should involve data from different countries, ethnic origin and ages to see how effective the above models are, for different populations. This could result into generic models that will be instrumental in several clinical practices.

Explainability and Interpretability: Future research may consider using some techniques such as the SHAP (SHapley Additive exPlanations) values or the LIME (Local Interpretable Model-agnostic Explanations) to make elaborate models like Random Forest more comprehensible. This would be particularly useful in clinical applications where knowledge of the heuristic that was used to make a forecast is critical.

Incorporation of Temporal Data: In the future, if longitudinal data is obtainable, then time aspects could be used and the model could predict changes of obesity status over time in a more dynamic and personalised manner.

# 14.References

1. UCI Machine Learning Repository (no date). https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition.
2. Palechor, F.M. and De La Hoz Manotas, A. (2019) 'Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico,' Data in Brief, 25, p. 104344. https://doi.org/10.1016/j.dib.2019.104344.
3. Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5-32.
4. Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785-794.
5. García, S., Luengo, J. and Herrera, F., 2016. *Data preprocessing in data mining*. Cham: Springer.
6. Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. New York: Springer.
7. Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression*. 3rd ed. Hoboken: Wiley.
8. Martínez-Rosales, A., Gómez-Ballesteros, M.Á. and Sanchis, A., 2020. Predicting obesity levels using machine learning techniques. *Journal of Healthcare Engineering*, 2020, pp.1-11.
9. Ng, M., Fleming, T., Robinson, M., et al., 2014. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: A systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, 384(9945), pp.766-781.
10. Ogden, C.L., Carroll, M.D., Kit, B.K. and Flegal, K.M., 2014. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA*, 311(8), pp.806-814.
11. Sefiane, A., Mellal, M. and Derdour, M., 2021. Machine learning-based prediction of obesity levels using lifestyle data. *International Journal of Environmental Research and Public Health*, 18(2), pp.1-15.
12. Tukey, J.W., 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.
13. Vapnik, V., 2013. *The nature of statistical learning theory*. 2nd ed. New York: Springer.
14. World Health Organization, 2021. Obesity and overweight. Available at: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> [Accessed 12 August 2024].
15. Bleich, S.N., Vercammen, K.A., Zatz, L.Y., Frelier, J.M., Ebbeling, C.B. and Peeters, A., 2018. Interventions to prevent global childhood overweight and obesity: a systematic review. The Lancet Diabetes & Endocrinology, 6(4), pp.332-346.
16. Dietz, W.H. and Robinson, T.N., 2005. Overweight children and adolescents. The New England Journal of Medicine, 352(20), pp.2100-2109.
17. Duan, L., Street, W.N., Lu, Q. and Liu, Y., 2019. A survey of machine learning methods for hypertension risk prediction. Current Hypertension Reports, 21(11), p.89.
18. Klein, S., Allison, D.B., Heymsfield, S.B., Kelley, D.E., Leibel, R.L., Nonas, C. and Kahn, R., 2007. Waist circumference and cardiometabolic risk: a consensus statement from Shaping America’s Health: Association for Weight Management and Obesity Prevention; NAASO, The Obesity Society; the American Society for Nutrition; and the American Diabetes Association. The American Journal of Clinical Nutrition, 85(5), pp.1197-1202.
19. Li, Y., Wang, D.D., Ley, S.H., Howard, A.G., He, Y. and Hu, F.B., 2015. Potential impact of time trend of life-style factors on cardiovascular disease burden in China. Journal of the American College of Cardiology, 66(7), pp.816-825.
20. Loh, W.Y., 2011. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), pp.14-23.
21. O'Neill, S. and O'Driscoll, L., 2015. Metabolic syndrome: a closer look at the growing epidemic and its associated pathologies. Obesity Reviews, 16(1), pp.1-12.
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, pp.2825-2830.
23. Poirier, P., Giles, T.D., Bray, G.A., Hong, Y., Stern, J.S., Pi-Sunyer, F.X. and Eckel, R.H., 2006. Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss. Arteriosclerosis, Thrombosis, and Vascular Biology, 26(5), pp.968-976.
24. Puhl, R.M. and Heuer, C.A., 2010. Obesity stigma: important considerations for public health. American Journal of Public Health, 100(6), pp.1019-1028.
25. Sahoo, K., Sahoo, B., Choudhury, A.K., Sofi, N.Y., Kumar, R. and Bhadoria, A.S., 2015. Childhood obesity: causes and consequences. Journal of Family Medicine and Primary Care, 4(2), pp.187-192.
26. Smith, A.I. and Martinez, K.E., 2020. The association between physical activity and obesity: A machine learning approach. Journal of Obesity and Metabolic Syndrome, 29(3), pp.164-172.

# Appendix

**R Code:**

[Obesity.R](https://1drv.ms/u/s!AhgLjS9rhouBxAUTAsHhEYgxE7H-?e=5RpOvl)

**Data :**

[OD.csv](https://1drv.ms/u/s!AhgLjS9rhouBw3pHcZPGNrfBCwBn?e=Co5sDw)