**Assignment Cover Page**

|  |  |
| --- | --- |
| **Subject Code** | COSC2999 |
| **Subject Name** | Practical Data Science with Python |
| **Location** | SGS |
| **Title of Assignment** | Assignment 2: Data Modelling |
| **Student Name** | Nguyễn Đình Đăng Nguyên |
| **Student Number** | S3759957 |
| **Lecturer Name** | DR. Thuy Nguyen |
| **Assignment due date** | December 2nd, 2024 |
| **Assignment submission date** | December 2nd, 2024 |
| **Numbers of pages including this one** | 15 |
| **Word count** |  |

*I declared that in submitting all work for this assessment I have read, understood and agree to the content and expectations of the Assessment Declaration*

Contents

[Abstract 2](#_Toc185633204)

[Introduction 2](#_Toc185633205)

[Methodology 2](#_Toc185633206)

[Data Preparation 2](#_Toc185633207)

[Exploratory data analysis (EDA) 3](#_Toc185633208)

[Feature analysis and selection 6](#_Toc185633209)

[Machine learning algorithms 7](#_Toc185633210)

[Grid search optimization and hyper-parameter tuning 7](#_Toc185633211)

[Evaluation Metrics 9](#_Toc185633212)

[Results 10](#_Toc185633213)

[Discussion 12](#_Toc185633214)

[Conclusion 12](#_Toc185633215)

[Reference 13](#_Toc185633216)

# Abstract

Obesity is a medical condition characterized by an excessive accumulation of body fat, which can have significant adverse effects on an individual's health [ref]. This study aimed to classify obesity levels using two distinct machine learning approaches—classification and regression—while focusing on physical activity, dietary habits, and genetic factors. The research adopted an observational design, gathering data from the UCI repository [ref] via a web-based survey to assess participants' eating patterns and physical activity levels. The dataset comprised variables such as gender, age, height, weight, family history of obesity, dietary behaviors, and frequency of physical activity. For the modeling phase, three classification algorithms were used to predict obesity levels: Random Forest (applied to both classification and regression tasks), Extreme Gradient Boosting (XGBoost, for classification), and k-nearest neighbors (KNN), with grid search optimization for hyperparameter tuning. Model performance was evaluated using a range of metrics: accuracy, recall, precision, F1-score, area under the curve (AUC), and precision-recall curve for classification tasks, and Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) for regression tasks. The Random Forest model demonstrated the most robust performance overall, and feature selection was found to enhance model efficiency. These results underscore the importance of physical activity and nutritional habits in mitigating the growing obesity epidemic.

# Introduction

The rise in obesity has emerged as a significant global public health challenge. According to the World Health Organization (WHO), obesity rates have steadily increased over the past few decades, driven by various factors, including genetics, environment, diet, physical activity levels, and behavioral patterns [ref]. Poor dietary habits, characterized by high-calorie consumption, lack of physical activity, genetic predispositions, and certain medical conditions or medications, all contribute to the growing prevalence of obesity.

Addressing obesity is critical not only for improving individual health but also for reducing the strain on healthcare systems and enhancing overall societal health outcomes. Immediate and sustained efforts are necessary to prevent obesity, ensuring healthier futures for individuals and communities worldwide. Effective prevention strategies should emphasize promoting balanced eating habits, increasing physical activity, and creating environments that encourage healthier lifestyle choices.

This report explores the effectiveness of using Machine Learning techniques to identify and address obesity-related variables, such as gender, age, height, weight, family history of obesity, dietary patterns, and physical activity frequency, utilizing the UCI dataset titled "Estimation of Obesity Levels Based On Eating Habits and Physical Condition" [ref].

# Methodology

## Data Preparation

**Data Collection**

The dataset used in this study, sourced from UCI [ref], contains data on obesity levels among individuals from Mexico, Peru, and Colombia. Participants, aged between 14 and 61, are evenly distributed across two genders and represent a diverse range of dietary habits and physical conditions. The dataset comprises 2,111 records, with 16 variables serving as input features and one variable as the output target. The following is a description of each feature included in the dataset:

|  |  |
| --- | --- |
| **Features** | **Description** |
| Gender | Gender |
| Age | Age |
| Height | Height |
| Weight | Weight |
| family\_history\_with\_overweight | Has a family member suffered or suffers from overweight? |
| FAVC | Do you eat high caloric food frequently? |
| FCVC | Do you usually eat vegetables in your meals? |
| NCP | How many main meals do you have daily? |
| CAEC | Do you eat any food between meals? |
| SMOKE | Do you smoke? |
| CH2O | How much water do you drink daily? |
| SCC | Do you monitor the calories you eat daily? |
| FAF | How often do you have physical activity? |
| TUE | How much time do you use technological devices such as cell phone, videogames, television, computer and others? |
| CALC | How often do you drink alcohol? |
| MTRANS | Which transportation do you usually use? |

The target of the study, obesity levels (NObeyesdad), is commonly calculated using **Body Mass Index (BMI)**, a commonly used measure that is related to weight and height. BMI is calculated using the following formula [ref]:

Where:

* Weight is in kilograms (kg)
* Height is in meters (m)

After calculating the Body Mass Index (BMI), the results were classified into the following categories based on established BMI thresholds [ref]:

* Underweight: Less than 18.5
* Normal: 18.5 to 24.9
* Overweight: 25.0 to 29.9
* Obesity I: 30.0 to 34.9
* Obesity II: 35.0 to 39.9
* Obesity III: Greater than 40

**Data Validation**

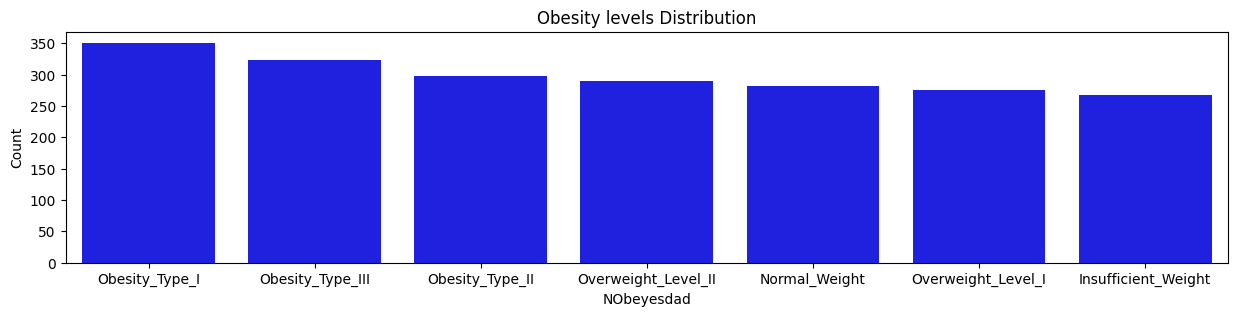
Overall, the Obesity Level dataset appears to be valid across the following categories:

* The dataset contains no missing values (NaN or null).
* All records fall within the valid range of responses for each survey question, with no incorrect values identified.
* A total of 24 duplicate records were found and removed during the data preprocessing phase.
* While some outliers are present in certain variables, they have been retained in the dataset, as their values still fall within the acceptable range.
* The dataset consists of 9 categorical variables ('Gender', 'family\_history\_with\_overweight', 'FAVC', 'CAEC', 'SMOKE', 'SCC', 'CALC', 'MTRANS', 'NObeyesdad') and 8 numerical variables ('Age', 'Height', 'Weight', 'FCVC', 'NCP', 'CH2O', 'FAF', 'TUE').

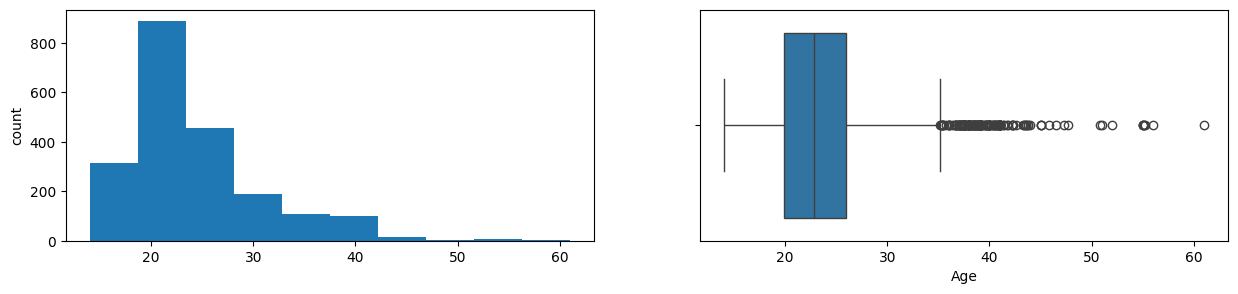
## Exploratory data analysis (EDA)

In this section, three findings of univariate and bivariate analysis are presented to show relationships of different variables.

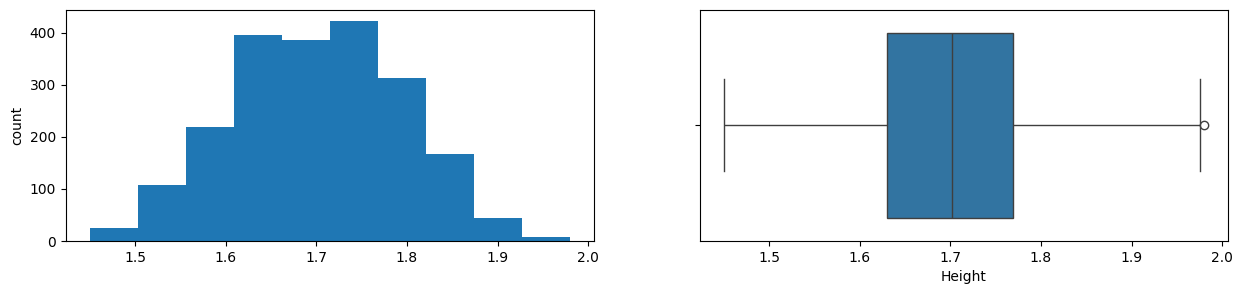
**Univariate analysis**



The obesity level (NObeyesdad) was selected as the target variable for analysis, as it is the primary focus of the dataset, which was designed to identify factors influencing obesity. Based on the graphical representation and descriptive statistics of the "NObeyesdad" variable, it can be concluded that there are 7 distinct classes of obesity types. Furthermore, the distribution of observations across these classes is relatively uniform, suggesting that the survey responses were likely influenced by participants' self-perceptions or appearance, rather than being randomly distributed.

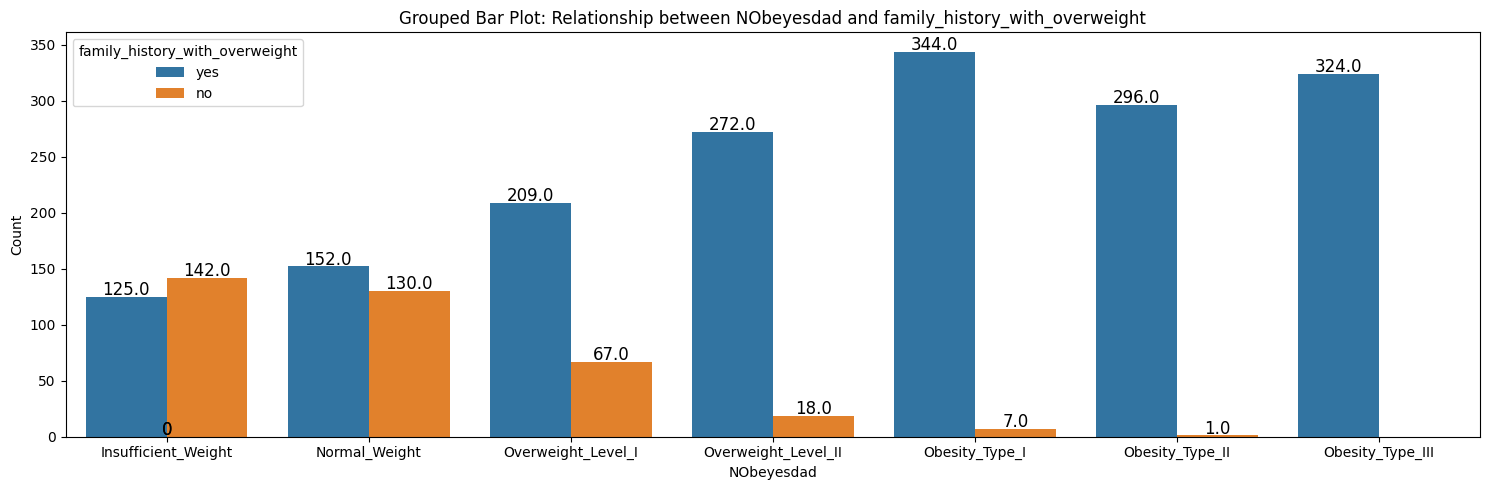


Based on the graph and descriptive statistics of the "Age" variable, several observations can be made that the survey population ranges from 14 to 61 years old. However, the data exhibits a relatively strong right skew, with a skewness index of 1.53, indicating that the survey predominantly targeted a younger audience, as reflected by the mean age of 24. The effective range of the data spans only up to an age of 35, with very few observations beyond this age, which can be considered outliers. Despite this, these outliers can be retained in the dataset, as their values still fall within the acceptable range.



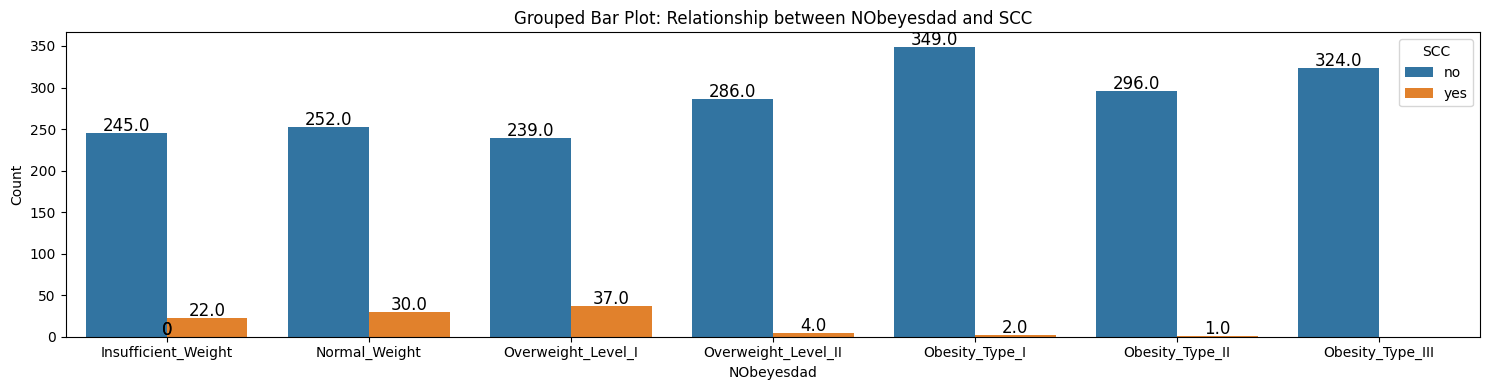
Based on the graph and descriptive statistics of the "Height" variable, the following observations can be made that the height of the survey population ranges from 1.45m to 1.98m. The distribution of height is approximately normal, with a skewness index of -0.01, indicating a near symmetrical distribution. The mean height is around 1.7m, with the only outlier being the value of 1.98m, which is significantly higher than the other data points.

**Bivariate analysis**

******

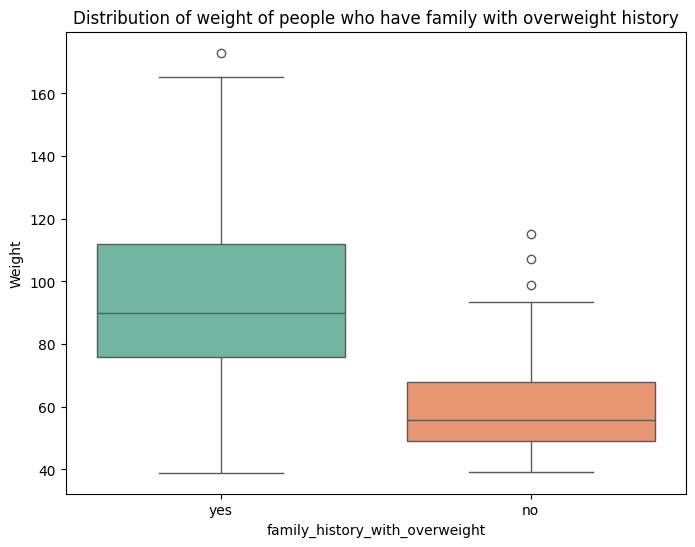
***Hypothesis: If you have no member in your family who has obesity, you have a high chance of not having obesity.***

Observations from the graph support this hypothesis, as the number of individuals with no family history of obesity gradually decreases from the first weight level to the highest (dropping from 146 to 0). Conversely, the number of people with at least one family member who has obesity increases steadily from the first to the last weight level (rising from 126 to 324). In the categories of insufficient and normal weight, the proportion of families with a history of obesity is notably lower compared to the overall population. This suggests that individuals with no family history of obesity are less likely to experience weight-related issues, supporting the idea that the absence of obesity in the family reduces the likelihood of an individual having obesity or being overweight.



***Hypothesis: People who monitor their health daily are less likely to have obesity.***

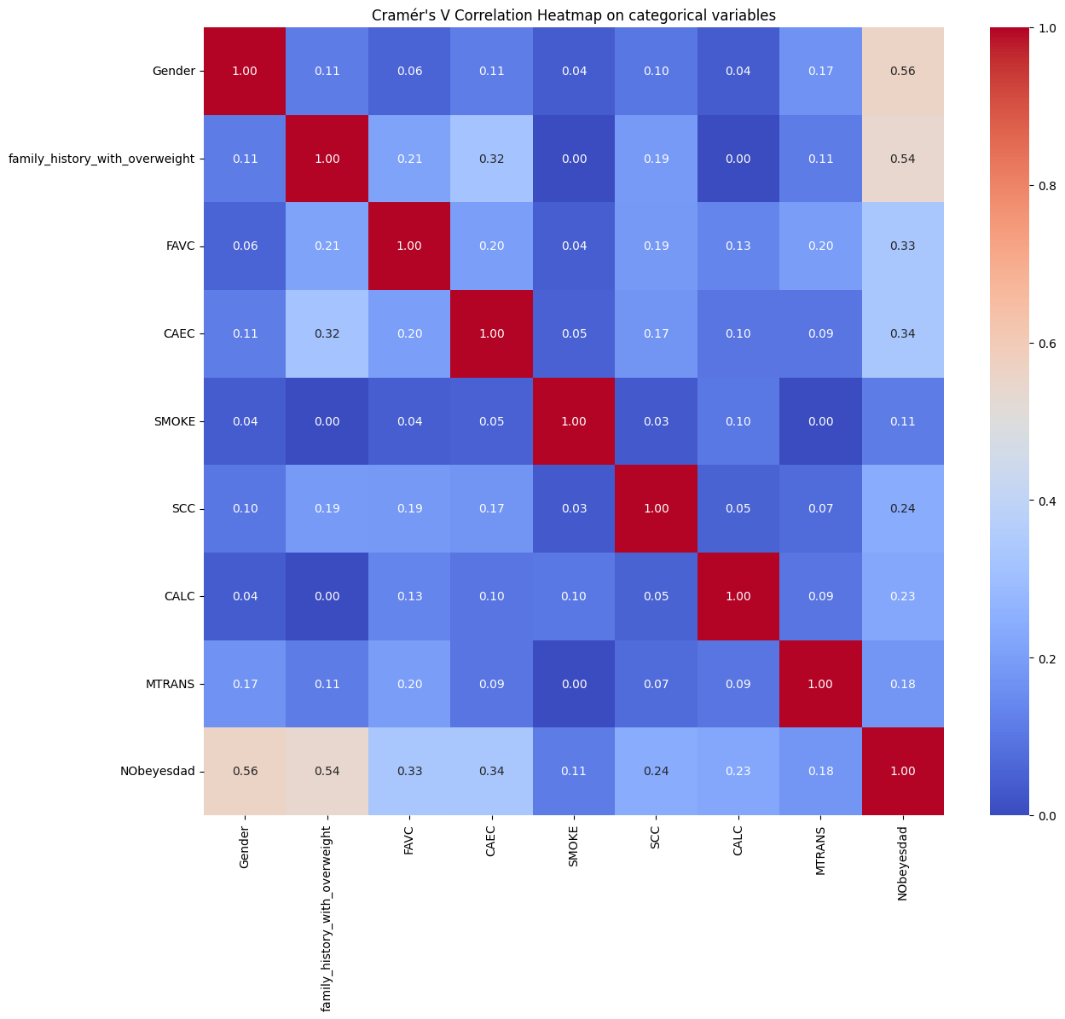
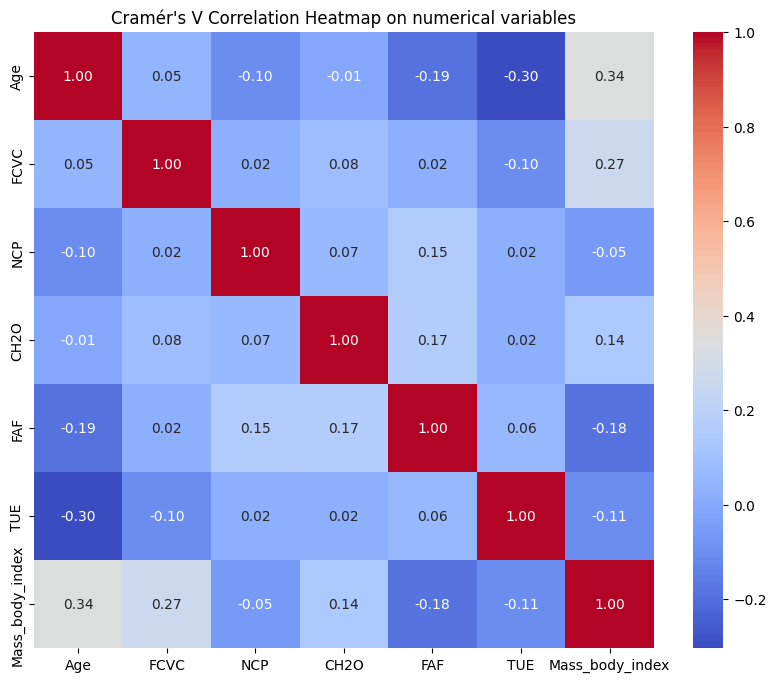
Observations from the graph support this hypothesis, as the number of individuals who monitor their health daily is more concentrated at lower weight levels. In contrast, the number of individuals who do not monitor their health daily gradually increases toward higher weight levels, although the rate of increase is relatively low. This suggests that daily health monitoring may contribute to a lower likelihood of obesity or overweight. However, it is important to note that the correlation between daily health monitoring and obesity is relatively weak, indicating that while there is some association, other factors likely play a more significant role in determining obesity levels.



*Hypothesis: People with family members who have an overweight history weigh more than those without.*

The data supports this hypothesis, as individuals with a family history of overweight predominantly fall within the weight range of 78 to 115 kg, with a small number of outliers above 160 kg. In contrast, those without a family history of overweight are mostly concentrated between 50 to 70 kg, with a few outliers near 110 kg. This clearly indicates that, on average, individuals with a family history of overweight tend to weigh more than those without such a history. Moreover, the weight distributions for both groups do not overlap significantly, with only a few data points falling within similar weight ranges, further reinforcing the idea that family history of overweight is a strong indicator of higher body weight.

## Feature analysis and selection



The following data transformation tasks were carried out prior to the training phase: First, the target variables, which were composite variables of Height and Weight calculated using the Body Mass Index (BMI) formula (BMI = Weight / Height²), were removed from the dataframe. Numerical data distributions were transformed using the PowerTransformer to make the data more Gaussian-like, enhancing model performance. The numerical values were then scaled using MinMaxScaler to standardize the range of the features. All categorical variables were encoded into numerical values using LabelEncoder. The dataset was split into training and testing sets using an 8:2 ratio, while ensuring that the class distribution was preserved in both subsets. Through several experiments, it was observed that including all features resulted in the best performance for the classification task, with NObeyesdad as the target variable. For the regression task, including only those features with a correlation score of greater than 0.1 with the target variable led to the highest model performance, with Mass\_body\_index (BMI) as the target variable.

## Machine learning algorithms

In this study, three state-of-the-art machine learning algorithms—Random Forest, K-nearest neighbors (KNN), and Extreme Gradient Boosting (XGBoost)—were employed to predict obesity levels based on eating habits and physical conditions.

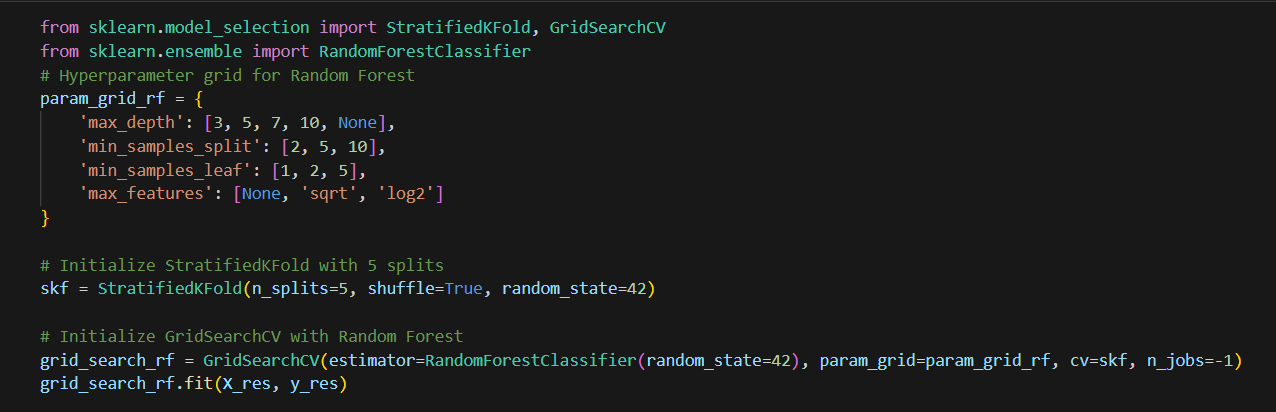
**Random Forest** is an ensemble learning algorithm that combines multiple decision trees to improve classification accuracy and reduce overfitting [ref]. It works by generating a collection of decision trees through bootstrapped sampling of the training data, with each tree trained on a random subset of features. When making predictions, the Random Forest model aggregates the outputs of all the individual trees, using majority voting for classification or averaging for regression. This method mitigates the overfitting problem commonly seen with individual decision trees. Random Forest is particularly effective for large datasets with numerous features (such as this dataset with 2,111 records and 17 features) and can handle missing values, maintaining accuracy even with significant data gaps. Its ability to generalize well, combined with robustness to outliers and noise, makes it a popular choice for many classification and regression tasks, and it was employed here for both.

**XGBoost (eXtreme Gradient Boosting)** is a highly efficient and scalable algorithm, renowned for its exceptional performance in classification tasks [ref]. Unlike Random Forest, which builds multiple trees independently, XGBoost constructs trees sequentially, with each new tree correcting the errors of the previous ones [ref]. This boosting technique enhances the model’s predictive power by focusing on difficult-to-classify data points. XGBoost employs gradient boosting, minimizing a loss function by updating the weights of trees to reduce prediction errors. It also includes regularization techniques to prevent overfitting, making it particularly well-suited for complex, high-dimensional data. The algorithm’s speed, accuracy, and flexibility in handling various data types and distributions have made it a preferred choice in machine learning competitions and real-world applications. For this study, XGBoost was utilized for classification tasks.

**K-Nearest Neighbors (KNN)** is a simple and intuitive classification algorithm that assigns a class label based on the majority class of its nearest neighbors [ref]. The model calculates the distance (typically Euclidean) between the test point and all training samples, identifying the 'k' nearest neighbors and predicting the class label that appears most frequently. The choice of 'k' is crucial, as a small value can lead to overfitting, while a large value may cause underfitting. KNN is a non-parametric method, making no assumptions about the underlying data distribution, and is versatile for a wide range of classification tasks. However, it can be computationally expensive, particularly with large datasets, since it requires calculating distances for each query point. Despite this limitation, KNN is widely used for its simplicity, effectiveness, and ability to handle multi-class classification problems. In this study, KNN was employed for regression tasks.

These algorithms were chosen for their robust performance in handling the complexities of obesity prediction, leveraging both classification and regression techniques based on the dataset's variables.

## Grid search optimization and hyper-parameter tuning



*Figure: Fitting Random Forest model with hyper-parameter tuning and grid search optimization*

The aforementioned sample of code is implemented for 2 tasks of classification and regression and varied in 4 models. This code employed hyperparameter tuning using **GridSearchCV** along with **StratifiedKFold** cross-validation to optimize each model. The goal is to find the best combination of hyperparameters for the models to improve its performance. Each of these parameters has a predefined range of values, and the **GridSearchCV** will exhaustively search over all combinations of these values to find the optimal set for the model.

To ensure a robust evaluation of the model, the code employs **StratifiedKFold** with 5 splits. This technique partitions the data into five subsets (or folds), where each fold is used once as the test set, while the remaining folds serve as the training sets. Stratified sampling ensures that each fold maintains the same proportion of class labels as the original dataset, making it particularly useful for imbalanced datasets [ref]. This approach provides a more reliable estimate of model performance across different data subsets.

Additionally, **GridSearchCV** is used in conjunction with the model and the hyperparameter grid. It is set to run with 5-fold cross-validation, with the computation process parallelized by setting n\_jobs=-1. This configuration allows the process to utilize all available CPU cores, accelerating the search. After fitting the model to the dataset (with features denoted as **X\_res** and target labels as **y\_res**), **GridSearchCV** evaluates all possible hyperparameter combinations and identifies the optimal set that maximizes the model’s performance based on cross-validation results [ref]. This procedure ensures that the final model is finely tuned and capable of performing optimally on unseen data.

With the given strategies, each of the model’s hyper-parameters are well-tuned with following groups:

**For classification tasks**

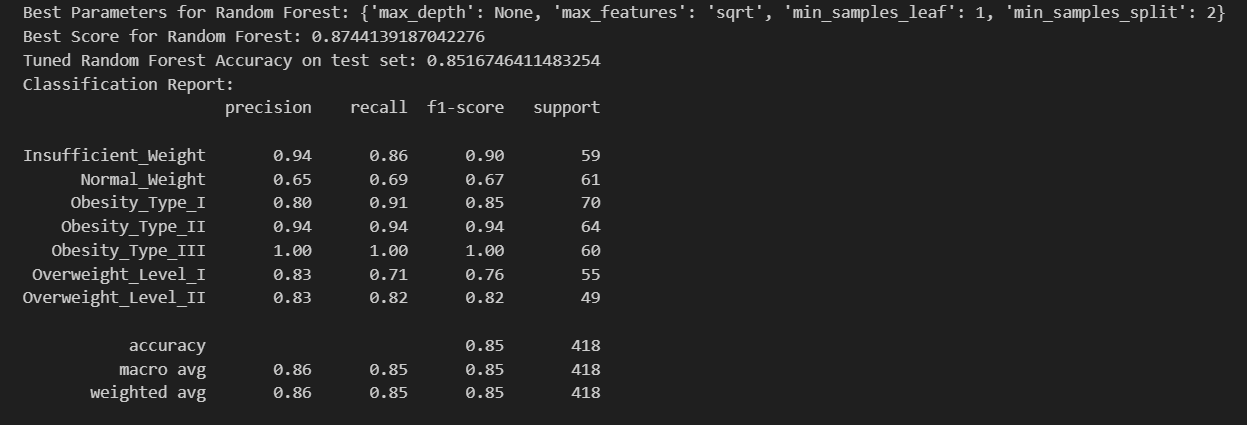
|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyper-parameter** | **Values** | **Best group** |
| Random Forest | 'max\_depth' | 3, 5, 7, 10, None | None |
| 'min\_samples\_split' | 2, 5, 10 | 2 |
| 'min\_samples\_leaf' | 1, 2, 5 | 1 |
| 'max\_features' | None, 'sqrt', 'log2' | 'sqrt' |
| XGBoost | 'n\_estimators' | 50, 100, 200 | 200 |
| 'max\_depth' | 3, 5, 7, 10 | 10 |
| 'learning\_rate' | 0.01, 0.1, 0.05 | 0.05 |
| 'subsample' | 0.5, 0.7, 0.8, 1.0 | 0.8 |

**For regression tasks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyper-parameter** | **Values** | **Best group** |
| KNN | 'n\_neighbors' | 3, 5, 7, 9, 11 | 5 |
| 'weights' | 'uniform', 'distance' | 'distance' |
| 'metric' | 'euclidean', 'manhattan' | 'manhattan' |
| Random Forest | 'n\_estimators' | 100, 200, 300 | 200 |
| 'max\_depth' | 10, 20, 30, None | None |
| 'min\_samples\_split' | 2, 5, 10 | 2 |
| 'min\_samples\_leaf' | 1, 2, 4 | 1 |
| 'max\_features' | None, 'sqrt', 'log2' | 'sqrt' |

## Evaluation Metrics

**For classification tasks**

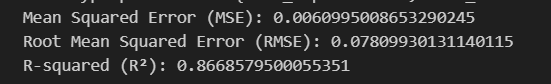


*Figure: Evaluation metrics for classification tasks*

In classification analysis, accuracy, precision, recall, and F1-score are utilized to evaluate the performance of Random Forest and XGBoost model [ref]:

* **Accuracy:** Accuracy measures the proportion of correctly classified instances out of the total number [ref]. It is calculated as:
* **Precision:** Precision (also known as Positive Predictive Value) measures the proportion of true positive predictions of the entire positive predictions[ref]. It presents how many of the predicted positive instances are positive. The formula for precision is:
* **Recall:** Recall (also known as Sensitivity or True Positive Rate) presents the proportion of actual positive instances that were correctly predicted [ref]. It answers the question: How many did the model correctly identify out of the actual positives? The formula for recall is:
* **F1-score:** F1-score is the harmonic mean of precision and recall, balancing both metrics in one number [ref]. It is especially useful when you need to balance the tradeoff between precision and recall. The formula for F1-score is:

**For regression tasks**



*Figure: Evaluation metrics for regression tasks*

In regression analysis, RMSE (Root Mean Squared Error), MSE (Mean Squared Error), and R² (Coefficient of Determination) are utilized to evaluate the performance of Random Forest and KNN model [ref]:

* **RMSE (Root Mean Squared Error):** RMSE provides the error rate between the predicted and actual values, with a lower RMSE indicating better model performance [ref]. The formula for RMSE is:
* **MSE (Mean Squared Error):** MSE is the average of the squared differences between the predicted values and the actual values [ref]. It is sensitive to outliers, indicating that larger errors have a more greater impact on the MSE. The formula for MSE is:
* **R² (Coefficient of Determination):** R² measures how well the regression model explains the variability in the data [ref]. R² values range from 0 to 1, with 1 indicating perfect prediction and 0 indicating that the model does not explain any variability. The formula for R² is:

# Results

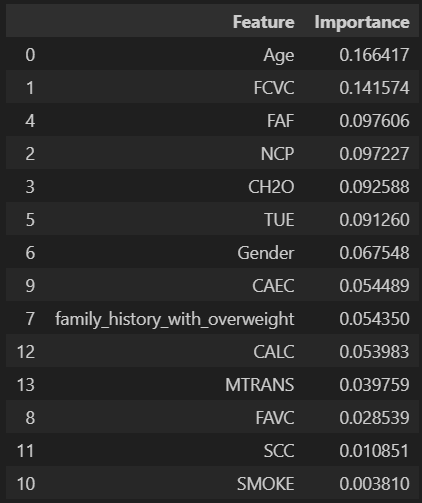
**For classification tasks**

The higher performance of the Random Forest model in predicting obesity levels, with an accuracy of 85.2%, underscores its ability to handle complex, high-dimensional data. This accuracy indicates that over 85% of the predictions made by the model were correct, reflecting a strong model performance. When compared to XGBoost, which achieved an accuracy of 84.4%, the difference is minimal, suggesting that both models are highly competitive. However, Random Forest’s slightly higher accuracy may stem from its ensemble learning approach, which helps reduce overfitting by combining the outputs of multiple decision trees, thus enhancing its ability to generalize to unseen data.

Beyond accuracy, Random Forest also demonstrated robust performance in other key classification metrics, including precision, recall, and F1-score. The precision of 86.0% indicates that 86% of the instances predicted as belonging to a specific obesity category were correctly classified. This is particularly important in scenarios where false positives could lead to incorrect recommendations or interventions. The recall rate of 85.0% reveals that the model was able to identify 85% of the true instances for each obesity category, ensuring that most individuals were accurately assigned to the correct group. The F1-score, also at 85.0%, further highlights the model's effectiveness, as it represents the harmonic mean of precision and recall, providing a balanced measure of performance. These metrics collectively affirm the Random Forest model's strong predictive capabilities in classifying obesity levels.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **F1 score** | **Recall** |
| Random Forest | 0.852 | 0.860 | 0.850 | 0.850 |
| XGBoost | 0.844 | 0.850 | 0.850 | 0.840 |

The feature importance analysis reveals that Age, FCVC (Frequency of Consumption of Vegetables), and FAF (Frequency of Physical Activity) are the most influential factors. These features reflect key lifestyle and physiological aspects that directly influence obesity levels.



**For regression tasks**

The Random Forest regressor outperformed the K-Nearest Neighbors (KNN) algorithm in predicting obesity levels, achieving an R² score of 0.867. This indicates that the model was able to explain approximately 86.7% of the variance in the target variable, a notable achievement in capturing the underlying patterns in the data. In comparison, KNN obtained an R² score of 0.821, demonstrating a lower ability to capture the variance in the data. The higher R² score for Random Forest suggests that it provided more accurate and reliable predictions by effectively modeling the complex relationships within the dataset.

In addition to the impressive R² score, the Random Forest model exhibited strong consistency in performance, with exceptionally low error rates. The model achieved an RMSE (Root Mean Squared Error) of 0.078, which indicates that the average difference between predicted and actual values was minimal. Furthermore, the MSE (Mean Squared Error) was 0.006, reinforcing the model's accuracy and reliability in producing predictions that closely matched the actual values. These error metrics are particularly significant given that the target values were normalized within the 0-1 range, making the model's accuracy even more impressive by being sensitive to small variations. Overall, the Random Forest regressor demonstrated superior performance to KNN in both explaining the variance in obesity levels and minimizing prediction errors, cementing its effectiveness for this regression task.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **MSE** | **R2** |
| Random Forest | 0.078 | 0.006 | 0.867 |
| KNN | 0.090 | 0.008 | 0.821 |

For regression task, most important features that affect people body mass index (BMI) values are Age, FCVC (Frequency of Consumption of Vegetables), and family\_history\_with\_overweight (value if people have family members with overweight history).



# Discussion

In the task of predicting obesity levels (or, BMI segmentation), Random Forest model demonstrated strong performance when compared to both K-Nearest Neighbors (KNN) and XGBoost, in both classification and regression tasks. In classification, Random Forest outperformed XGBoost slightly, achieving an accuracy of 85.2% compared to XGBoost's 84.4%. Additionally, Random Forest achieved robust precision, recall, and F1-scores (86.0%, 85.0%, and 85.0%, respectively), indicating its ability to balance false positives and false negatives effectively. These results highlight Random Forest’s effectiveness in accurately categorizing individuals into obesity level groups, making it a reliable choice for classification tasks. In regression, Random Forest also surpassed KNN, achieving an R² score of 0.867, which reflects its ability to explain a greater proportion of the variance in the obesity level data. Random Forest’s consistent low error rates, with an RMSE of 0.078 and MSE of 0.006, further demonstrate its reliability in predicting continuous outcomes. In contrast, KNN, while simple and intuitive, struggled in both classification and regression, particularly in handling larger, more complex datasets, as indicated by its lower performance in these metrics. Overall, Random Forest, which combines predictions from multiple decision trees, made it more robust and accurate than KNN and XGBoost across both types of tasks, providing a comprehensive solution for predicting obesity levels.

In the machine learning task of predicting obesity levels, features such as age, vegetable consumption, physical activity, and a family history of overweight significantly impact the model's performance. Age plays a critical role because, as individuals get older, their metabolism tends to slow down, and they may become less physically active, leading to a higher likelihood of weight gain. Vegetable consumption is another important factor, as it reflects a person's dietary habits; a higher intake of vegetables, which are nutrient-dense and low in calories, is associated with healthier eating patterns and a lower risk of obesity. Physical activity is equally influential, as regular exercise helps burn calories, boosts metabolism, and maintains muscle mass, all of which are essential for weight management. Additionally, having family members with a history of overweight or obesity suggests a genetic predisposition or shared lifestyle factors that increase the likelihood of obesity. These features together highlight the complex interplay of genetics, lifestyle choices, and physiological changes that contribute to obesity, underscoring the importance of considering these factors when building predictive models.

# Conclusion

In conclusion, this study highlights the exceptional performance of the Random Forest model in predicting obesity levels among individuals in Mexico, Peru, and Colombia, leveraging a comprehensive set of demographic, dietary, physical, and health-related variables. The Random Forest model outperformed other widely used algorithms, such as XGBoost and KNN, achieving an accuracy of 85.2%, along with robust precision, recall, and F1-scores of 86.0%, 85.0%, and 85.0%, respectively. Feature importance analysis revealed several key predictors of obesity, including BMI, age, physical activity level, daily calorie intake, and family history of obesity. These findings align with existing literature and offer valuable insights into the primary factors driving obesity within this specific population. Notably, the results underscore the critical influence of lifestyle factors, such as physical activity and diet, along with genetic predisposition. Additionally, the study highlights the potential of machine learning models, particularly Random Forest, to accurately pinpoint the factors contributing to obesity, providing useful tools for shaping public health strategies and personalized interventions to address the obesity crisis in Latin American countries.

***\*Declaration of AI Usage:*** *This report has been prepared with the assistance of artificial intelligence (AI) tools, specifically in the areas of code assistance, lexical and grammar correction, and ensuring the clarity and coherence of written sections. While AI played a supporting role, all interpretations, conclusions, and final content have been thoroughly reviewed and validated by the author (myself). The use of AI aligns with RMIT’s guidelines on academic integrity, and the report maintains the principles of originality and ethical conduct.*

# Reference