**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 3](#_Toc185633206)

[Data Preparation 3](#_Toc185633207)

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

[Feature analysis and selection 3](#_Toc185633209)

[Machine learning algorithms 3](#_Toc185633210)

[Grid search optimization and hyper-parameter tuning 4](#_Toc185633211)

[Evaluation Metrics 5](#_Toc185633212)

[Results 7](#_Toc185633213)

[Discussion 8](#_Toc185633214)

[Conclusion 9](#_Toc185633215)

[Reference 9](#_Toc185633216)

# Abstract

Obesity is a medical condition characterized by an excessive accumulation of body fat, which can negatively impact an individual’s health. This study aimed to classify obesity levels using two machine learning approaches (Classification and Regression), with a focus on physical activity, nutritional habits, and genetic factors. The study utilized an observational design, collecting data from the UCI repository [ref] through a web-based survey to assess participants' eating habits and physical activity levels. The dataset included variables such as gender, age, height, weight, family history of obesity, dietary patterns, and physical activity frequency. For the modeling process, three classification algorithms were employed to predict obesity levels, including Random Forest (for both classification and regression tasks), Extreme Gradient Boosting (for classification), and k-nearest neighbors with grid search optimization for hyperparameter tuning. Model performance was evaluated using various 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 exhibited the most robust performance, and feature selection was shown to improve model efficiency. These findings emphasize the critical role of physical activity and nutritional habits in addressing the growing obesity epidemic.

# Introduction

The rise in obesity has become a significant public health concern worldwide. According to WHO, over the past few decades, obesity rates have steadily increased, driven by factors such as genetics, environment, diet, physical activity levels, and behaviors [ref]. Poor diet (high in calories), lack of physical activity, genetics, and certain medical conditions or medications can all contribute to obesity.

Addressing obesity is not only essential for individual well-being but also for reducing the overall burden on healthcare systems and improving societal health outcomes. Immediate and sustained efforts to prevent obesity are necessary to ensure healthier futures for individuals and communities around the world. Preventing obesity is crucial to reversing these trends and improving public health. Prevention strategies should focus on promoting healthier eating habits, increasing physical activity, and fostering environments that support healthier choices.

This report will outline the efficiency of using Machine Learning methodologies in detecting and preventing obesity based variables such as gender, age, height, weight, family history of obesity, dietary patterns, and physical activity frequency provided by UCI’s “Estimation of Obesity Levels Based On Eating Habits and Physical Condition” dataset [ref]

# Methodology

## Data Preparation

The dataset used in this study, obtained from UCI [ref], contains information on obesity levels among individuals from Mexico, Peru, and Colombia. The participants, aged between 14 and 61, balanced between 2 genders, represent a wide range of dietary habits and physical conditions. The dataset includes 2,111 records, consisting of 16 variables serving as inputted features and 1 as the outputted target. Each of the feature descriptions as listed as follows:

|  |  |
| --- | --- |
| **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 simple and widely used measure that helps categorize individuals based on their weight relative to their height. BMI is calculated using the following formula [ref]:

Where:

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

After all calculation was made to obtain the mass body index for each individual, the results were classified as [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 Higher than 40

## Exploratory data analysis (EDA)

## Feature analysis and selection

The dataset was divided into training and testing sets using an 8:2 ratio, while preserving the class distribution in both subsets.

## Machine learning algorithms

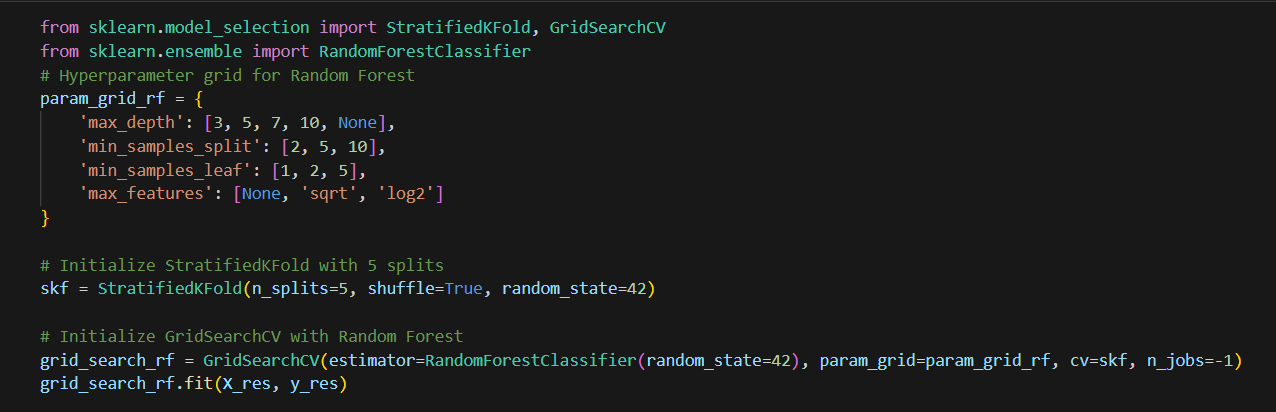
In this study, state-of-the-art Random Forest, K-nearest neighbors (KNN) and Extreme gradient boosting (XGBoost) algorithms were employed, to predict obesity levels based on various demographic, lifestyle, and health-related features.

**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, where each tree is trained on a random subset of features. When making predictions, the Random Forest model aggregates the outputs of all the individual trees by majority voting (for classification) or averaging (for regression). This method helps in mitigating the overfitting problem commonly seen with individual decision trees. Random Forest is particularly effective in handling large datasets with numerous features (particularly in this dataset, 2111 records with 17 features), and it can manage missing values and maintain accuracy even when a significant proportion of data is absent. Its ability to generalize well and its robustness to outliers and noise make it a popular choice for many classification and regression tasks. This model is employed in classification and regression tasks.

**XGBoost (eXtreme Gradient Boosting)** is a highly efficient and scalable machine learning algorithm, renowned for its exceptional performance in classification problems [ref]. Unlike Random Forest, which builds multiple trees independently, XGBoost builds trees sequentially, with each new tree designed to correct the errors made by the previous ones [ref. This boosting technique enhances the predictive power of the model by focusing on difficult-to-classify data points. XGBoost employs gradient boosting, where the model minimizes a loss function by updating the weights of the trees in a direction that reduces prediction errors. Additionally, it incorporates regularization techniques to prevent overfitting, making it well-suited for tasks involving complex, high-dimensional data. XGBoost’s speed, accuracy, and flexibility in handling different data types and distributions have made it a popular choice in machine learning competitions and real-world applications. This model is employed in classification tasks.

**K-Nearest Neighbors (KNN)** is a simple and intuitive classification algorithm that assigns a class label to a data point based on the majority class of its nearest neighbors [ref]. The model works by calculating the distance (typically Euclidean distance) between the test point and all training samples. It then identifies the 'k' nearest neighbors and predicts the class label that appears most frequently among these neighbors. The value of 'k' is a user-defined parameter, and its selection can significantly influence the model's performance, as a small 'k' can lead to overfitting, while a large 'k' can result in underfitting. KNN is a non-parametric method, meaning it makes no assumptions about the underlying data distribution, making it versatile for various types of classification tasks. However, it can be computationally expensive, especially with large datasets, as it requires calculating distances for each query point. Despite this, KNN is widely used for its simplicity, effectiveness, and ability to handle multi-class classification problems. This model is employed in regression tasks.

## 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 robust evaluation of the model, the code uses **StratifiedKFold** with 5 splits. This technique divides the data into 5 subsets (folds), where each fold is used once as a test set while the remaining folds are used as training sets. The stratified sampling ensures that each fold maintains the same proportion of class labels as the original dataset, which is especially useful in the case of imbalanced datasets. This method provides a more reliable estimate of model performance across different subsets of the data.

**GridSearchCV** is initialized with the model and the hyperparameter grid, and it is set to run using 5-fold cross-validation, parallelizing the computation with n\_jobs=-1. This means that the process will utilize all available CPU cores to speed up the search. After fitting the model to the dataset **X\_res** (features) and **y\_res** (target labels), **GridSearchCV** will evaluate all hyperparameter combinations and identify the set that maximizes the model’s performance based on cross-validation results. This process ensures that the final model is tuned to perform optimally on unseen data.

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

**For classification tasks**

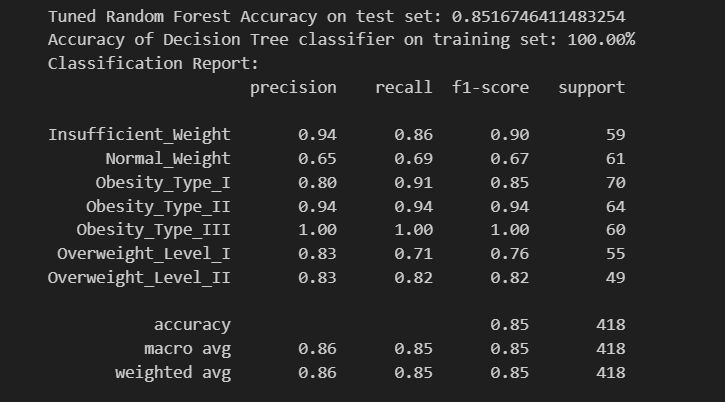
|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyper-parameter** | **Values** | **Best group** |
| Random Forest | 'max\_depth' | 3, 5, 7, 10, None | None |
| 'min\_samples\_split' | 2, 5, 10 | 2 |
| 'min\_samples\_leaf' | 1, 2, 5 | 1 |
| 'max\_features' | None, 'sqrt', 'log2' | 'sqrt' |
| XGBoost | '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**

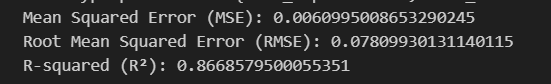


*Figure: Evaluation metrics for classification tasks*

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 all instances in the dataset [ref]. It is calculated as:
* **Precision:** Precision (also known as Positive Predictive Value) measures the proportion of true positive predictions among all positive predictions made by the model [ref]. It tells you how many of the predicted positive instances are actually positive. The formula for precision is:
* **Recall:** Recall (also known as Sensitivity or True Positive Rate) measures the proportion of actual positive instances that were correctly identified by the model [ref]. It answers the question: Of all the actual positives, how many did the model correctly identify? 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 a measure of how much error there is 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, meaning larger errors have a more significant 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**

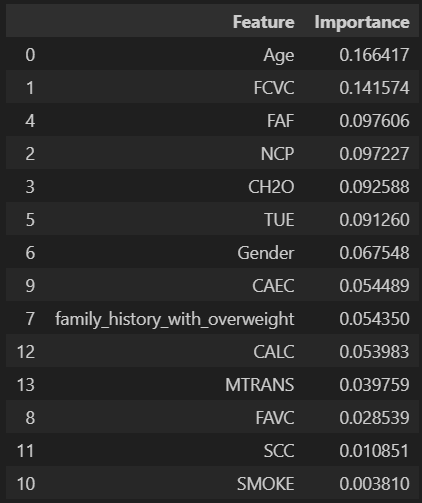
The Random Forest model's superior performance in predicting obesity levels, with an accuracy of 85.2%, highlights its ability to handle complex, high-dimensional data. Accuracy, in this case, indicates that over 85% of the predictions made by the model were correct. This level of performance is slightly better than XGBoost, which achieved an accuracy of 84.4%. The slight difference in accuracy suggests that both models are competitive, but Random Forest may be slightly better at generalizing to unseen data, likely due to its ensemble learning approach that reduces overfitting by combining multiple decision trees.

In addition to accuracy, Random Forest also demonstrated strong performance across other important classification metrics: precision, recall, and F1-score. Precision, at 86.0%, means that 86% of the instances predicted as belonging to a certain obesity category were correctly classified. This is crucial for applications where false positives (misclassifying individuals into the wrong obesity level) can lead to inappropriate recommendations or interventions. Meanwhile, the recall of 85.0% indicates that the model was able to identify 85% of all true instances for each obesity category, ensuring that most individuals were correctly classified within the appropriate group.

The F1-score of 85.0% further supports the effectiveness of Random Forest, as it is the harmonic mean of precision and recall, providing a balanced measure of the model’s ability to correctly identify both positive and negative instances.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 revealed that weight, height, and gender were the most influential factors in predicting obesity levels. This finding aligns with the established understanding that body mass index (BMI), which is calculated based on weight and height, is a primary indicator of obesity. The importance of gender in obesity prediction can be attributed to the differences in body composition and hormonal factors between males and females.

****

**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, which indicates that the model was able to explain approximately 86.7% of the variance in the target variable. This performance was slightly better than KNN, which achieved an R² score of 0.821, reflecting its lower ability to capture the variance in the data compared to Random Forest. The higher R² score of the Random Forest regressor suggests that it provided more accurate predictions by effectively capturing complex relationships in the dataset.

In addition to the R² score, the Random Forest model demonstrated a strong consistency in its performance, with notably low error rates. The model achieved an RMSE (Root Mean Squared Error) of 0.078, indicating that the average difference between the predicted and actual values was very small. Similarly, the MSE (Mean Squared Error) was 0.006, which further supports the model's reliability and its ability to produce predictions that are close to the actual values. These error metrics are particularly impressive considering that the target values were normalized to a range of 0-1, making them more sensitive to small variations. Overall, the Random Forest regressor's performance was superior to KNN, both in terms of its ability to explain the variance in obesity levels and in minimizing prediction errors.

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



# Discussion

In the task of predicting obesity levels, the 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’s ensemble learning approach, which aggregates 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.

The results of this study demonstrate the superior performance of the CatBoost model in predicting obesity levels compared to other ML algorithms. The high accuracy, precision, recall, and F1-score achieved by the CatBoost model highlight its effectiveness in classifying individuals into their respective obesity level categories. These findings are consistent with previous studies that have shown the advantages of gradient boosting algorithms, particularly CatBoost, in various classification tasks. The CatBoost model's ability to handle categorical variables effectively is a key factor contributing to its success in this study. By utilizing ordered target statistics for categorical features, CatBoost can capture the relationships between these features and the target variable more efficiently than other algorithms that require extensive preprocessing. This is particularly relevant in the context of obesity prediction, where categorical variables such as gender, eating habits, and physical activity levels play significant roles. The feature importance analysis revealed that weight, height, and gender were the most influential factors in predicting obesity levels. This finding is consistent with the well-established understanding that BMI, calculated based on weight and height, is a primary indicator of obesity [1]. The importance of gender in obesity prediction can be attributed to the physiological differences between males and females, such as body composition and hormonal factors, which affect the development and distribution of body fat . The identification of dietary habits (FCVC and CH2O) and sedentary behavior (TUE) as important features underscores the role of lifestyle factors in the development of obesity. Previous studies have shown that unhealthy eating patterns, characterized by high consumption of energy-dense foods and low intake of fruits and vegetables, are associated with an increased risk of obesity. Similarly, excessive screen time and sedentary behavior have been linked to weight gain and obesity. The inclusion of these features in the CatBoost model highlights the potential for targeting these modifiable risk factors in obesity prevention and management strategies. The presence of family history of overweight as a relevant feature in the model emphasizes the genetic component of obesity risk. Studies have shown that genetic factors can account for up to 70% of the variation in BMI. The CatBoost model's ability to incorporate this information in its predictions demonstrates its potential to identify individuals with a higher genetic predisposition to obesity, allowing for early intervention and personalized management approaches. To contextualize the findings of this study, it is important to compare the CatBoost model's performance with similar studies using different ML algorithms or feature sets for obesity prediction. Dugan et al. used decision trees, random forests, and support vector machines to predict obesity based on lifestyle factors and demographic variables. Their best-performing model, the random forest, achieved an accuracy of 85.2%, lower than the 95.98% accuracy obtained by the CatBoost model in our study. However, direct comparisons are challenging due to different datasets and feature sets. Yi et al. employed deep learning with convolutional neural networks (CNNs) for obesity prediction based on body images, achieving an accuracy of 91.7%. While innovative, their approach relies on visual data rather than the demographic, lifestyle, and health-related features used in our study Muse et al. used a combination of feature selection techniques and ML algorithms, including support vector machines and artificial neural networks, for obesity prediction. Their best-performing model achieved an accuracy of 93.2%, comparable to the CatBoost model's performance. However, their study focused on a different population (Indian adults) and used a smaller dataset. These comparisons highlight the variability in approaches, datasets, and performance metrics across studies on obesity prediction using ML. While the CatBoost model's performance is promising, further research is needed to establish its superiority over other algorithms in diverse settings and populations.

# Conclusion

In conclusion, this study demonstrates the superior performance of the CatBoost model in predicting obesity levels among Indonesian adults based on demographic, lifestyle, and health-related factors. The CatBoost model outperformed other commonly used algorithms, including logistic regression, KNN, random forest, and naive Bayes, achieving an accuracy of 95.98%, precision of 96.08%, recall of 95.98%, and F1-score of 96.00%. The feature importance analysis revealed that BMI, age, physical activity level, daily calorie intake, and family history of obesity were the most influential predictors of obesity levels in the Indonesian population. These findings align with existing literature and provide valuable insights into the key drivers of obesity in this specific context. The study's novelty lies in its application of the CatBoost algorithm, which has not been extensively explored in the domain of obesity prediction, particularly in the Indonesian setting. The CatBoost model's ability to handle categorical variables effectively, resist overfitting, and provide interpretable results makes it a promising tool for obesity risk assessment and classification. The practical implications of this study are significant. The high accuracy and interpretability of the CatBoost model can assist healthcare professionals and policymakers in identifying individuals at high risk of obesity and developing targeted prevention and intervention strategies. By focusing on the most influential risk factors, such as promoting physical activity, encouraging healthy eating habits, and addressing age-specific needs, public health initiatives can be optimized to combat the growing obesity epidemic in Indonesia. The CatBoost model's superior performance in predicting obesity levels, coupled with its ability to handle categorical variables and provide interpretable results, makes it a promising tool for obesity classification and risk assessment. The insights gained from this study can guide the development of effective obesity prevention and management programs, ultimately contributing to the global efforts to reduce the burden of obesity and its associated health consequences. Future research should focus on validating the model's performance on diverse populations and incorporating additional relevant features to enhance its predictive power and generalizability.

***\*Declaration of AI Support:*** *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