
# Arab American University

## Faculty of Graduate Studies

# Classification of Adults' Obesity and Overweight using Machine Learning Models: Case Study Palestine

By

**Ahmad Hakam Abd-Al hafiz Radwan**

Supervisor

**Prof. Dr. Mohammed Awad**

**This thesis was submitted in Partial Fulfillment of the Requirements for the Master's Degree in Data Science and Business Analytics.**

**Sep 2023**

# Classification of Adults' Obesity and Overweight using Machine Learning Models: Case Study Palestine

By

**Ahmad Hakam Abd-Al hafiz Radwan**

This Thesis was Defended Successfully on **09/09/2023** and approved by:

| Committee Members | Signature |
|---|---|
| **1.** Supervisor: **Prof. Mohammed Awad** |  |
| **2.** Internal Examiner: **Dr. Ahmad Ewais** |  |
| **3.** External Examiner: **Dr. Abdullatif Abuissa** |  |

## *Declaration*

This word provided in this thesis unless otherwise referenced, is the researcher own work, and has not been submitted elsewhere for any other degree or qualification.

Student Name: Ahmad Hakam Abd-al hafiz Radwan

Student Number: 20212228

Signature: 

Date: 31-10-2023

## *Dedication*

I dedicate this work to my father, mother, brothers, and sisters who helped me overcome the difficulties. I also dedicate this work to my soul mate my wife and my beloved children, who were the decisive factor in these successes

## *Acknowledgments*

There are no enough words to thank my supervisor Prof. Mohammed Awad for his patience, kindness, precise guidance, and support to complete this research with suitable work and advice.

## *Abstract*

Obesity and Overweight are considered a major cause of many illnesses on the global level. According to the World Health Organization (WHO), more than 30% of global population are suffering Obesity and Overweight. This percentage may increase in the coming years. This global phenomenon would lead to a series of dangers considering the emergence of other diseases. Obesity is often associated with other chronic diseases, such as arterial hypertension, and type 2 diabetes mellitus. The risk of suffering from these comorbidities is greater as the body mass index increases. However, there are infrequent studies investigating obesity in Palestine. Such lack of knowledge affects the prediction identification and prevention of obesity in Palestinian society.

As a result, this research aims to fill this gap in obesity research in Palestine, depending on real risk factors related to Palestinian society. The research dataset was collected from Palestine which included 902 participants. However, after classifying the participants into four categories, Underweight, Normal, Overweight, and Obesity, the first class was imbalanced which required using data balancing methods in this work we use Synthetic Minority Over-sampling Technique (SMOTE). The collected data was processed by ML models. The models include support vector machine (SVM), Random Forest (RF), decision tree (DT) multi-layer perceptron (MLPNNs), XGBoost, Adaboost, Extra tree classifier, and Gradient Boost (GB). The grid search was applied to models to obtain the appropriate set of parameters for each model. The performance of the models was evaluated on imbalance and balanced data, and the model that outperforms other models with the highest accuracy is the RF with (98.3%, and 98.6%) accuracy respectively


**Keywords:** *Adults Obesity, Obesity in Palestine, Classification Machine Learning Models, Synthetic Minority Over-sampling Technique (SMOTE)*

# *Table of Contents*

# *List of Figures*

## List of Tables



## List of Abbreviation

| Abbreviation | Explanation |
|---|---|
| AI | Artificial Intelligence |
| ANN | Artificial Neural Networks |
| AUC | Area Under Curve |
| BFP | Body Fat Percentage |
| BMI | Body Mass Index |
| CART | Classification and Regression Tree |
| DL | Deep Learning |
| DT | Decision Tree |
| EDA | Exploratory Data Analysis |
| ETC | Extra Tree Classifier |
| FN | False Negative |
| FP | False Positive |

| | |
|---|---|
| FPR | False Positive Rate |
| GB | Gradient Boosting |
| GM | Gut Microbiota |
| IG | Information Gain |
| IOT | Internet of Things |
| KNN | K-Nearest Neighbor |
| LR | Logistic Regression |
| MLPNNs | Multi-Layer Perceptron Neural Networks |
| NB | Naïve Bayes |
| NN | Neural Network |
| PC | Personal Computer |
| RF | Random Forest |
| RMSE | Root Mean Square Error |
| RNN | Recurrent Neural Network |
| SEMMA | Sample, Explore, Modify, Model, and Assess |
| SMOTE | Synthetic Minority Over-sampling Technique |
| SVM | Support Vector Machines |
| TN | True Negative |
| TP | True Positive |
| TPR | True Positive Rate |
| WHO | World Health Organization |

# Chapter 1. Introduction

## *1.1 Introduction*

According to the World Health Organization (WHO), Overweight and obesity are defined as irregular or exaggerated fat accretion that may impair health (*Obesity and Overweight*, n.d.). Environment and genetic factors are the main reasons for many people to be obese or overweight, also, physical activities affect health conditions, where the lack of activity is a factor for having obesity or overweight (*Obesity - Symptoms and Causes - Mayo Clinic*, n.d.-a)(*Causes of Obesity | Overweight & Obesity | CDC*, n.d.). More than 340 million children and youth between the ages of 5 and 19 were obese in 2016. The prevalence of overweight and obesity among children and adolescents aged 5 to 19 has substantially increased, rising from around 4 % in 1975 to just over 18 %. Male and female overweight rates also increased, rising from 19 % to 18 %, respectively (*Obesity*, n.d.). The frequency of obesity and overweight among adolescents and children has significantly increased during the past 40 years (Avgerinos et al., 2019).

People of all ages are affected by obesity(*Obesity - NHS*, n.d.), which is a major public health issue and has been connected to several illnesses like diabetes, heart disease, and chronic disease (Verma & Hussain, 2017). Some of the most common diseases associated with obesity include Type 2 Diabetes: Obesity is one of the most significant risk factors for developing type 2 diabetes. It can cause insulin resistance, high blood sugar levels, sleep disturbances, or androgen dysfunction, which can ultimately endanger the whole body(Piché et al., 2020). Cardiovascular Disease: Obesity can raise the risk of developing cardiac disease, stroke, hypertension, dyslipidemia, sleep disorders, or other cardiovascular conditions. The result of cardiac disease may lead to excessive cholesterol, and blood pressure, all of which can cause heart damage and blood vessels(Lopez-Jimenez et al., 2022). Certain Cancers: At least 13 anatomic parts of the body, including

endometrial, colon, esophageal, renal, pancreatic adenocarcinomas, hepatocellular carcinoma, gastric cancer, meningioma, multiple myeloma, colorectal, postmenopausal breast, ovarian, gallbladder, and thyroid cancers, are all associated with an increased risk of cancer when a person is overweight or obese (Avgerinos et al., 2019)(Lin et al., 2020). Covid-19: (COVID-19), caused by the severe acute respiratory distress coronavirus 2 (SARS-CoV2), is a vast and fast-evolving epidemic challenging humanity and posing unprecedented global health problems. Many datasets show that COVID-19 is connected with increased disease severity in people who are obese or overweight. Obesity is usually linked to dysregulated renin–angiotensin–aldosterone (RAAS) axis [11]. Depression: and obesity are closely relations to mental health illnesses, with a bidirectional relationship between them. Obesity could increase the risk of developing depression, while depression can lead to obtaining weight and obesity. Psychological factors, including low self-esteem and society problems (Blasco et al., 2020). Infertility: Obesity is strongly linked to infertility, affecting both males and females. In females, obesity can disrupt hormonal balance, leading to infrequent menstrual cycles and ovulatory dysfunction. In males, obesity can lower male hormones and raise female hormones, resulting in reduced sperm production and impaired sperm function(Dağ & Dilbaz, 2015).

Body mass index (BMI) is a measure of body fat (*Body Mass Index (BMI) | Healthy Weight, Nutrition, and Physical Activity | CDC*, n.d.), which is individual weight in kilograms divided by the square of height in meters. If BMI is 25 or above is considered to be overweight, and a BMI of 30 or more is considered obese (*Obesity - Symptoms and Causes - Mayo Clinic*, n.d.-b). figure 1.1 shows BMI classification, where each BMI range is related to a different weight category.


Figure 1.1 BMI classification according to WHO (*Free BMI with Age Calculator - Health by Science*, n.d.)

The development of artificial intelligence (AI) and machine learning (ML) opens a door to adequately carrying out the task of investigating the causes of the appearance and development of overweight and obesity. AI is understood as the development of systems endowed with intellectual processes typical of human beings (Ahuja, 2019). Among these processes, we find reasoning, generalization, improvement through past experiences, and discovery of meanings(Jassar et al., 2022).

Obesity is characterized by being multifactorial in its causes and associations, and the instruments to assess it tend to generate biased data. On the other hand, methods of statistics are useful in identifying factors of risk of some health problems, however, when modeling diseases related to lifestyles, where the causes are multifactorial, these methods are not that successful. ML can be considered an extension of more statistical models. It can be applied to problems in the health area, since they are more flexible to learn properties from the data, regardless of the human intervention for the selection of variables, and with additional benefits, such as incorporating more robust

strategies for handling missing or misplaced data. ML methods are characterized by considering complex relationships between variables, which allows them to dispense with subjective decisions or those that depend on the intervention, unlike statistical models. They are also capable of detecting patterns in large datasets (Beam & Kohane, 2018).

ML plays a great role in the medical sector, including Obesity and Overweight to predict a new value)Maciej Serda et al., 2013((*Top 10 Applications of Machine Learning in Healthcare - FWS*, n.d.). ML models emerge as tools adopted in the context of the study of obesity; and useful, since they have the potential to consider a large number of predictor variables, In the work, we proposed to apply ML methods to determine if some of the factors associated with or causing obesity and classify overweight or obesity in a certain way. Supervised Machine Learning algorithms like Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boost (GB), AdaBoost, Multi-layer perceptron (MLPNNs), Extra Tree Classifier (ETC), and XGBoost used to classify Obesity and overweight into four categories “Underweight, Normal, Overweight and Obesity” in Palestine. In Palestine, the absence of the use of ML models to classify Obesity in Adults of data has an impact on the Palestinian society's ability to forecast, identify, and prevent obesity. As a result, our ability, as Palestinians, to understand, predict or find answers is very limited.

## *1.2 Objectives*

The classification of Obesity based on Machine Learning (ML) has proven effective in getting highly accurate results. The main objective of this research is to propose ML models used to classify Obesity and overweight into four categories “Underweight, Normal, Overweight, and Obesity” in Palestine depending on ML models using a locally collected dataset based on the effect

of risk factors from adults. These models can discriminate the induvial weight for a specific class which can help nutritionists and physicians in dealing with each class in different ways to reach high-quality treatment.

***Specific Objectives***

- Exploring the Overweight and Obesity dataset available in Palestine. and statistical analysis of the selected data, to Know the study population and data density.
- Applying Machine learning techniques to feature selection and identification of the most relevant features or variables that contribute to obesity and Overweight.
- Implementing Machine Learning models for the classification and Prediction of Overweight and Obesity in adults in Palestine which leads to risk assessment of obesity development.
- By analyzing various factors such as demographics, lifestyle, and medical history, these models can provide personalized risk assessments and early intervention opportunities.
- Comparing the performance of the implemented ML models with tuning parameters that predict Overweight and Obesity.
- Reaching the classification model that can aid medical researchers in the future. Such knowledge would increase the likelihood of stopping the spread of Overweight and Obesity.

## *1.3 Contribution*

Obesity and Overweight research in Palestine play a significant role in addressing the growing epidemic and its impact on the population's health and well-being. Researchers in Palestine,

consider obesity and overweight using a statistical approach. The use of such approaches provides us with facts about the current situation or tries to build a model that can predict the trend in data.

However, in this research, an obesity and overweight dataset has been collected for adults in Palestine to classify Obesity and overweight into four categories (Underweight, Normal, Overweight, and Obesity). The researcher tried to deal with four main groups due to the data limitations. Synthetic Minority Over-sampling Technique (SMOTE) is used to deal with imbalanced classes, where the feature selection method is used along with ML classification models. This research's contribution is in treating a modern local, Palestinian, dataset by using classification ML approaches.

Based on such research, government intervention strategies can be modified or developed by Palestinian society. Researchers explore culturally appropriate interventions that promote healthy eating, physical activity, and eating habits. This may involve non-government-based programs, universities-based interventions, or government policy recommendations to construct supportive environments for healthy living.

## *1.4 Thesis Organization*

Chapter One presents the general background of Obesity and Overweight. It shows the statistical studies and the predicting class for people who could have one of four classes of obesity. Also, the relation between Obesity and Overweight with several chronic diseases is considered. Then, we introduce techniques used to deal with this issue. Finally, we explain how machine learning can be conducted to identify risk factors associated with Obesity and Overweight.

Chapter Two provides a literature review on Obesity and Overweight of existing research. Various approaches will be discussed for studying Obesity and overweight from a statistical approach perspective, and ML approach.

Chapter Three shows the process of developing, evaluating, and deploying ML models and involves the presentation of classification algorithms. The methodology presents the process, data pre-processing, feature engineering, feature selection, model deployment, and finally model evaluation.

Chapter Four includes the results which are presented by machine learning classification models. This chapter includes the presentation of results by all models and selecting the most appropriate result based on a high-quality comparison between models' results. Also, we will try to make the selected model generalized to make it approved.

Finally, chapter five present the research conclusion and future works.

# Chapter 2. Literature Reviews

## *2.1 Background*

Currently, there is a high rate of obesity and overweight in young people which causes chronic diseases at an early age in the different systems of the human body, On the other hand, it is known that technology is part of the development of the population. Today with the rise of technology and new information-gathering tools more and more volumes of data are collected per patient, which permits the analysis, and is a source of knowledge to find predictions that allow focus or

develop segmented strategies and have a deep knowledge of the patient to have support when making clinical decisions, and further facilitates the modeling of predictive techniques that identify anomalies and/or patient evolution automatically and with very high accuracy. In this research, different ML models will be applied with the aim of diagnosing Obesity and Overweight in adults using a dataset collected from Palestine. In this research, we seek to start developing a system for the classification of obesity in Palestine using ML techniques, in such a way as to facilitate the diagnosis of obesity to users. The application of ML models to classify and predict Obesity and Overweight aimed at a population of adults is a subject that has not been addressed in our country, to the best of our knowledge.

Statistical approach involves using methods and techniques to analyze and interpret data. This approach typically studies the past situation and creates a hypothesis for the study based on the research perspective, where researchers formulate specific research questions or hypotheses design studies and collect relevant data, applying statistical tests to analyze the data and draw conclusions based on statistical inference. Otherwise, in recent years, ML approaches have gained popularity in medical research. ML algorithms can automatically discover hidden patterns and complex relationships from large datasets without relying on predefined hypotheses from the researcher. These algorithms aim to build predictive models that can accurately classify obesity and overweight, predict outcomes, or identify the risk factors in medical data. AI offers tremendous potential to support physicians in the diagnostic process. By leveraging its capabilities in pattern recognition, medical decision support systems, risk prediction, differential diagnosis, real-time control, and data analysis, AI models can boost diagnostic accuracy, otherwise. It also can reduce failure, and improve efficiency in healthcare settings.

In this thesis, various classification machine learning models will be applied, like, SVM, RF, GBoost, MLPNNs, AdaBoost, Extra tree classifier, DT, and XGBoost, in our attempt to get the

most valuable classification results. Based on the literature review, we propose to create a classification dataset to classify Obesity and Overweight based on risk factors.

## *2.2 Literature Review*

Artificial Intelligence (AI) and analytical models are leveraging the medical decisions of patients and the data. In this work, we intend to give a clear idea of how we can identify risks associated with complications in the obese population through ML models in Palestine society. Several articles use statistical models to study Obesity and Overweight in Palestine, but few articles concentrate on using ML to study Obesity and Overweight, especially for adults.

Obesity studies have attracted many researchers around the world, looking into the epidemic among children as in (Al-Hazzaa et al., 2022)(Abdullah et al., 2017)(Massad et al., 2016)(Gannon et al., 2021)(Duchen et al., 2020)(Rashmi et al., 2021), or adults (Dinda Qatrunnada, n.d.). investigating Obesity can be purposive for treatment's sake in the future. In this part, the research tries to pin out other related literature that is related to obesity among adults.

Globally, several researchers look into the epidemic, its current situation, and future prevention, in (Jindal et al., 2018), The researchers predict obesity using a combination of machine learning techniques. They proposed an ensemble machine learning approach to the diagnosis of obesity and applied the ensemble prediction they utilized their projected value of obesity was 89.68% accurate. For its prediction model, the Python interface additionally makes use of the generalized linear model, random forest, and partial least squares. Other researchers use classification models as in (IEEE Computational Intelligence Society et al., n.d.-a), they used publicly available genetic profiles to conduct research studies on machine learning algorithms for obesity prediction. Algorithms predict sensitivity to chronic hepatitis using single nucleotide polymorphism data

(SNPs), they used the Support Vector Machine (SVM) algorithm, Decision Tree (DT), decision rule, and K-nearest neighbors (KNN) method. SVM provided the best outcome for their prediction model out of those methods. According to their simulation results, SVM produced the area under the curve with the greatest value of 90.5%.

As the previous research, researchers in (Cheng et al., 2021) attempt to investigate the connection between a person's weight status and physical activity. They also compared certain machine learning and conventional statistical models for predicting obesity levels. The National Health and Nutrition Survey Dataset was used in their model, the sample size was 7162 and they used eleven different algorithms for their implementation and evaluation, including the random subspace, logistic regression, decision tree, Nave Bayes, the Radial Basis Function, K-nearest neighbors, classification via regression, J48, and MLPNNs. The random subspace classifier method was the approach that attained the overall maximum accuracy according to the assessment metrics of the ROC and AUC. Also, the researchers in (Cervantes & Palacio, 2020) used the obesity dataset compiled from university students from Latin American nations, the researchers developed and contrasted the SVM model with the decision tree model, the decision tree achieved good precision and recall values. They also suggested a Decision Tree, a Simple K-Means model, which achieved 98.5% recall and 98.5% precision.

Overall, the potential of data mining techniques is tackling the issue of obesity. It can provide valuable insight into predicting obesity and overweight as in (Molina et al., 2021) The researchers utilized several classification methods, namely Logistic Model Tree (LMT), Random Forest (RF), Multi-Layer Perceptron (MLPNNs), and Support Vector Machines (SVM). Additionally, to assess the models' performance, the researchers found results revealed that LMT exhibited the highest precision, achieving an accuracy rate of 96.65%. In comparison, Random Forest achieved 95.62%, MLPNNs reached 94.41%, and SVM obtained 83.89%. Thus, the study implements that LMT is

a dependable method for analyzing obesity and similar datasets. While another researcher as in (Thamrin et al., 2021), applied ML models like LR, NB, and CART to predict obesity among adults based on risk factors on a public health dataset in Indonesia, CART obtained the highest accuracy 82%.

The related disease to Obesity such as Gut Microbiota (GM), the researcher in (Zeng et al., 2019) study GM markers for obesity in patients with various metabolic abnormalities and their relationships with clinical indicators, the researcher used the information from 1914 Chinese adults. The study identifies common biomarkers for obesity patients with high uric acid, high serum lipids, and high blood pressure, which hold the potential to predict obesity-related metabolic abnormalities. RF achieved a higher result with an AUC of 0.77. while another researcher, as in (Liu et al., 2022), used fecal samples for 2262 Chinese individuals besides any personal information. The researchers applied ML models, and SVM obtained the highest accuracy 0.716 and $R^2$ =0.485. Also, the researchers figured out that obesity is related to many diseases such as cardiovascular diseases (CVDs), chronic obstructive pulmonary disease (COPD), cancer, diabetes type II, hypertension, and depression. as (Powell-Wiley et al., 2021), the main goal is studying obesity based on other diseases like CVDs and others, they use public datasets for both genders from 20 – 60 years old. SVM obtained the best accuracy of 95% and $R^2$ = 0.96. other researchers investigate obesity Genetic Profiles as in (C. A. C. Montañez et al., 2017), this paper proposes using a machine learning approach for predicting obesity by genetic profiles from publicly available databases which contain 6622 different variables. The approach involves identifying genetic variants as risk factors and using them as inputs in various machine learning algorithms to classify individuals into normal or risk classes based on their body mass index status. SVM achieved higher results with AUC 90.5%.

In (Jeon et al., 2023) the authors used a 3-dimensional body measurements approach to investigate obesity. The researchers developed a new framework for the classification of obesity. They applied it on 160 Korean subjects where the sample included males and females in their 20s and 30s. The approach's accuracy was 80%. Other researchers study obesity for men and women separately, by using a dataset from National Health and Nutrition Examination Survey (NHANES) as in (DeGregory et al., 2018). The researchers use ML and Deep Learning for males' data, the highest accuracy percentage was for LR, NN, and DL models, with (87%, 87%, and 88%) respectively. For females' data, DL, NN and LR models achieved (88%, 89%, 89%) respectively. Other researchers study the changes in plasma lipidomes for obese as in (Gerl et al., 2019). The researchers investigate the prediction of different scales of obesity based on the plasma lipidome. the researchers targeted 1061 members, they conducted many ML models and the Lasso model predicted BFP with the highest $R^2 = 0.73$.

Another perspective for investigating obesity is to study the development of obesity overtime. As in (Xue et al., 2019) the researchers propose using RNN architecture in public clinical patient records. They pick out key details and look for unusual observations related to obesity; the accuracy obtained is 86%. In (Ferenci & Kovács, 2018), the research predicts BFP from main measurable data, such as age, gender, weight, height, waist circumference and other measures. The researchers applied many regression ML models, and SVM obtained a root mean square error (RMSE) of $0.0988 \pm 0.00288$. Some studies used risk factors, such as personal attributes like height and weight and some physical condition and eating attitudes, in studying obesity as in (De-La-Hoz-Correa et al., 2019). The researcher collected a dataset from university students in Mexico, Peru, and Colombia, and utilized the SEMMA data mining technique, the DT achieved 97.4%. However, the following researches use the same dataset. In (Kitis & Goker, 2023) uses the RF model

and achieved 95.78%. XGBoost model was utilized by (DANACI et al., 2023) and achieved (97%,97.16%, and 98.5%) respectively.

In Arab World, obesity was considered by many researchers as (Ibrahim et al., 2022)(Alzahrani et al., 2023)(Nasaif, 2022) for adults, and for children as in (Al-Hazzaa et al., 2022). All of the mentioned research adopted statistical methods in investigating obesity. However, new attempts to use ML and DL methods in investigating obesity have arisen. In (Hadi Albayati et al., n.d.), the researcher builds an ML model for the classification of obesity, the sample includes 137 youth students in Iraq between the ages of 18-25. The researcher applied many models, such as SVM and Linear discriminant analysis (LDA), and the highest accuracy assigned for SVM with 94.2%. Other researchers in Saudi Arabia study obesity among children, youth, and adults (Alsareii et al., 2022). They use a global dataset collected in Mexico, Peru, and Colombia, and propose using the Internet of Things (IoT) as a second step after using ML models such as RF, KNN, SVM, DT, LR, and NB. SVM obtained 96%.

Locally, although there are recent attempts to use other machine learning approaches to get insights from medical data, the traditional statistical approach is still used to deal with medical data. The usage of the statistical approach would provide us with information on the status quo of obesity, which aids medical researchers in understanding the problem without any further information on the problem's main source or possible treatment of obesity, as in (Abdeen et al., 2012a)(Abdeen et al., 2012b)(Ellulu et al., 2014)(Bentham et al., 2017)(El Bilbeisi et al., 2017)(Damiri et al., 2018)(Ghrayeb et al., n.d.)(El Kishawi et al., 2014). In the last 5 years, only one study of obesity using the ML approach could be traced, which is (Amro & Awad, 2021). The researchers study obesity by using data mining techniques and regression ML and DL to predict obesity by applying them to anthropometric measurements. The study included 8 variables in a global dataset collected in 1994 from the Human Performance Research Center at Brigham Young University and found that

artificial neural networks (ANNs) were the most effective method, with a correlation coefficient of $R^2$=0.77.

Based on the literature review of studies particularly in Palestine, scientific researchers consider obesity and overweight as a good topic to look into, but a statistical approach is mostly used. The use of such approaches provides us with facts about the current situation or tries to build a model that can predict the trend in data, in a regression case. However, this research uses ML for classification the of obesity and overweight for the first time on a dataset of adults collected from the Palestine population. In this research we collect data from Palestinian society in the year 2023 to achieve the best results of accuracy using different models of ML, considering obesity. then, the research will try to generalize the findings so they can be used on other data sources to help the medical community by contributing to the development of solutions.

# Chapter 3. Methodology

## *3.1 Proposed Method*

This section describes the methodology used in this research. The expected results are to get insights and analyze models to classify obesity into 4 categories (underweight, normal weight, overweight, and obese). The dataset includes attributes divided into three sections, which are: Demographic, eating habits, and physical condition. The methodology is going to be fulfilled by following three main steps. The first step is exploration and visualization of the collected dataset. The second step is Feature Engineering and Preprocessing, which can be considered the most challenging process during the classification. It contains the transformation and creation of new attributes from existing attributes in the dataset. Besides that, standardization (variance scaling) and encoding categorical variables. For dealing with imbalanced data, we proposed using SMOTE technique to make all categories approximately equal. The benefit of equalizing all categories is to prevent biasing between them and forbids weak performance for several models (*SMOTE for Imbalanced Classification with Python - MachineLearningMastery.Com*, n.d.).

Finally, the dataset will be prepared for modeling by selecting appropriate ML algorithms. Such as support vector machine (SVM), Random Forest (RF), Decision Tree (DT), Multi-Layer Perceptron Neural Networks (MLPNNs), XGBoost, Adaboost, Extra tree classifier, and Gradient

Boost (GB). Each one of these applied ML models has its parameters for accurate learning to make an intelligence decision support or classification. Then, the models will be applied to unseen datasets to obtain the most accurate results of testing. Estimated values of the parameters for each model will be empirically approached, Using Grid Search, Random Search, Hill Climbing, and other techniques (*Machine Learning Algorithms and The Art of Hyperparameter Selection | by Rosaria Silipo | Towards Data Science*, n.d.).

The next step after training and testing the ML models is evaluating the models' performance by running metrics such as accuracy, precision, recall, true positive rate, false positive rate, and F1 score. These metrics can determine the models' strengths or weaknesses. Furthermore, testing models will be applied to an unseen dataset to ensure it generalizes, which decides the model's capacity to adjust to and appropriately respond to the previously unknown dataset to be applicable in real cases (*What Is Generalization In Machine Learning? - Magnimind Academy*, n.d.).

Figure 3.1 presents the structure of the proposed model. Beginning with building an online questionnaire using the *Google Form* application to collect modern local datasets. Then, Exploratory data, to get more deeply and meaningful insights from datasets and distributions for all features, can help data scientists figure out the pattern of data, missing values, and correlations between features, and discover anomaly detection and outliers. Also, scientists use visualization libraries to obtain a powerful understanding of the collected data. After that, feature engineering and preprocessing help data scientists prepare data by solving the issues that may appear in the EDA process and dealing with imbalanced data. Next, the dataset will be divided into a training dataset of 80% and a testing dataset of 20% to ensure allow researcher's ability to evaluate the models' performances and avoid data leakage. After that, several ML models would be applied to classify obesity and overweight.

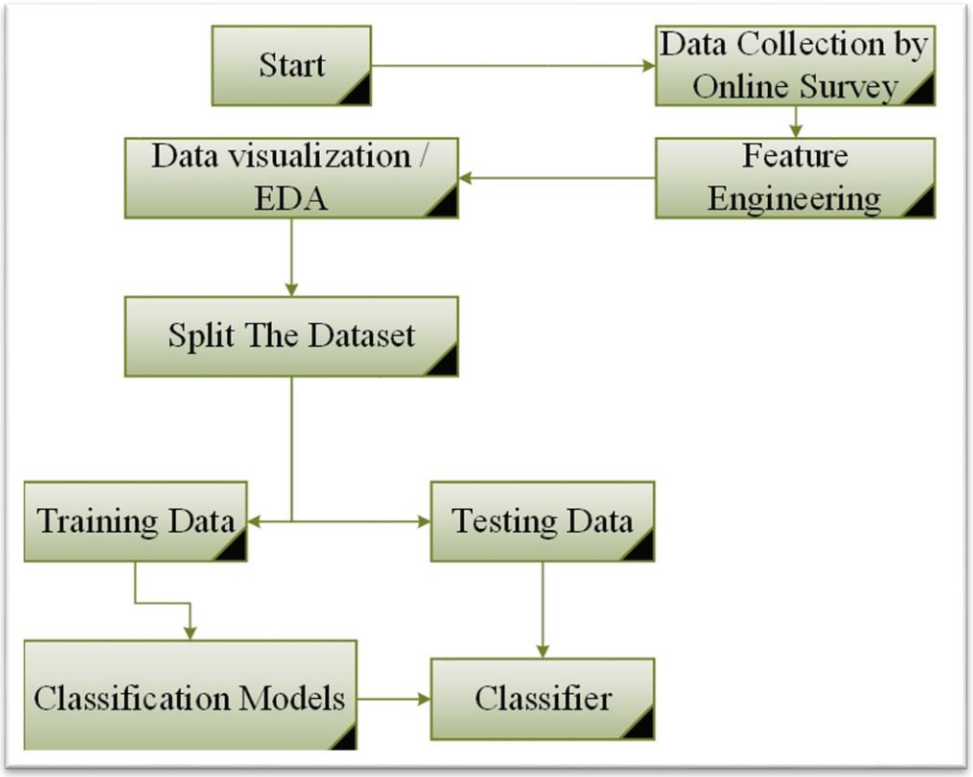


Figure 3.1 The proposed model for obesity and overweight classification

## *3.2 Data Description*

The researcher designed an online questionnaire, using a *Google Form* application that includes a variety of obesity-related questions depending on the literature, after that, the questionnaire was distributed online to collect the answers voluntarily. The questionnaire was modified after consulting an expert in Obesity and nutrition. The questionnaire is divided into three sections. the first one is general information, the second one consists of attributes related to eating habits, and the last one is related to physical condition. the questionnaire adopts (De-La-Hoz-Correa et al., 2019), which was distributed in Colombia, Peru, and Mexico in 2019. Moreover, the questionnaire expresses to the public community that it is made for the scientific research process.

Ethically, the researcher respects participants' confidentiality by not asking for any personal information and that the data will be collected for scientific research's sake only.

The questionnaire targets adults in Palestine, including different backgrounds, places of residence, or any other variables, it was distributed online starting from March,23 to April,23.

In the beginning 920 records where collected, and after the preprocessing stage, 902 records remained and constituted a well- prepared data.

## *3.3 Data Preprocessing*

Preprocessing the data before training in machine learning can be considered as an important step(*Python for Data Analysis, 3E - 7 Data Cleaning and Preparation*, n.d.). Preprocessing data includes many main steps: data cleansing or data scrubbing, transformation, aggregation, normalization, or feature scaling. The preprocessing stage in this work doesn't consist of any missing values in any records, but it contains two duplicated records.

Feature engineering is using current features to create a new feature or target (*Create New Features From Existing Features - Train a Supervised Machine Learning Model - OpenClassrooms*, n.d.). In this case, we use two features (height and weight) to create BMI variable based on the equation weight in kg divided by square height in m (*Body Mass Index (BMI) Calculator - Diabetes Canada*, n.d.).

Where weight is measured by kilogram, height is measured by meters.

$$BMI = \frac{weight(KG)}{height^2(M)} \quad (1)$$

For numeric data, standardization is usually used, which is critical for regularization techniques to avoid overfitting many algorithms (*Don't Overfit! II — How to Avoid Overfitting in Your Machine Learning and Deep Learning Models | by Md. Mubasir | Towards Data Science*, n.d.). In dealing with categorical data, we proposed using many techniques from many libraries such as sci-kit-learn, and we proposed to use a label encoder.

Figure 3.2 presents the obesity categories and counts for each category based on gender there is an imbalance in the dataset classes. For imbalance dataset cases there is many approaches to deal with imbalance data. Such as, oversampling techniques aim to amend class imbalance by increasing the number of instances in the minority class until it reaches a balance with the majority class. These methods encompass approaches such as random oversampling, which involves duplicating random instances from the minority class, synthetic oversampling utilizing techniques like Synthetic Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN). By oversampling, the model gains more data related to the minority class, enhancing its ability to make accurate predictions for that class. However, it is crucial to be cautious of potential overfitting, where the model becomes excessively biased towards the minority class due to the increased number of its instances.

On other hand, undersampling techniques work towards achieving class balance by reducing the number of instances in the majority class. Common methods include random undersampling, which entails the random removal of instances from the majority class, and Tomek connections, which eliminate instances that form Tomek connections with closest neighbors of different classes. Undersampling goal to diminish the dominance of the majority class, prompting the model to allocate greater attention to the minority class. Nevertheless, it comes with the drawback of potential information loss, as a significant portion of the majority class data is discarded. The choice between oversampling and undersampling depends on the specific dataset characteristics

and problem at hand, and careful experimentation is often necessary to determine the most effective strategy for addressing class imbalance.

This research proposes using SMOTE is a method used in machine learning to address class imbalance in datasets. It aims to balance the distribution of minority and majority classes by generating synthetic samples for the minority class. The process involves many steps, randomly selecting a minority instance, identifying its k-nearest neighbors, after that creating synthetic samples by interpolating between the chosen instance and its neighbors. This interpolation involves selecting random values between 0 and 1 for each feature to create new data points. These steps are repeated until the desired class balance is achieved. The imbalanced data can impure the classification process. SMOTE in many cases could improve the performance of models which can lead to high accurate classification results.

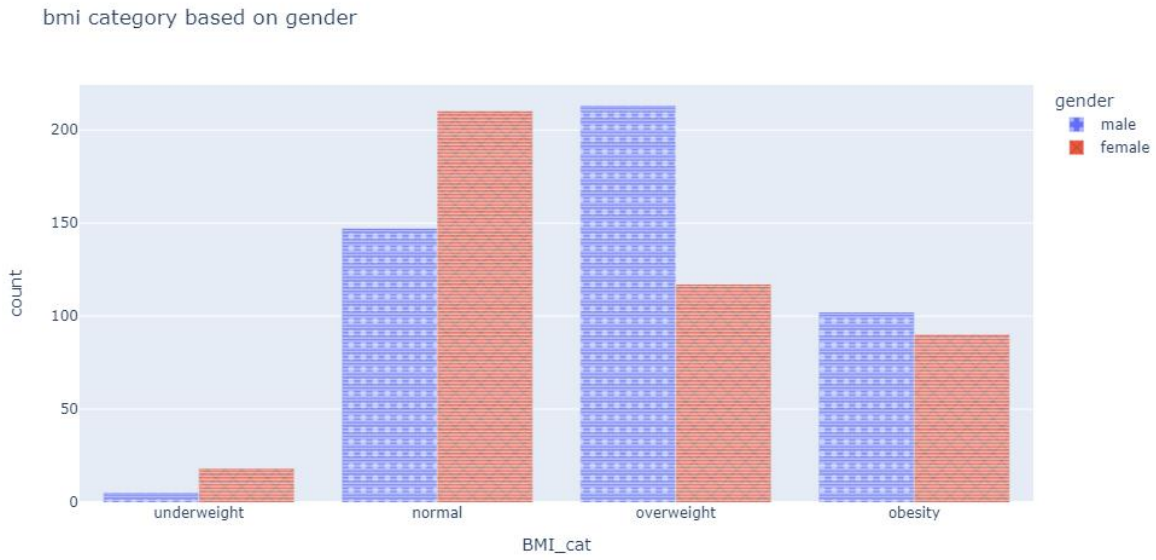


Figure 3.2 obesity categories based on gender

Figure 3.3 presents the obesity categories after using the SMOTE technique, the balance between all classes is attained.

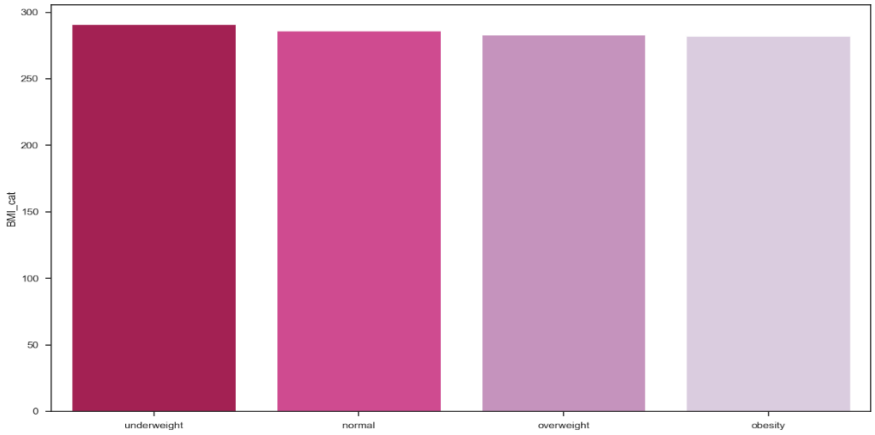


Figure 3.3 obesity categories after SMOTE technique

## *3.4 Exploratory Data Analysis (EDA)*

EDA is a significant stage in the data analysis. It includes visualization, exploration, descriptive statistics for numeric values, pattern recognition, anomaly detection, data types, correlation between features, etc. A better understanding of the data can lead to initial insights to pick up the appropriate ML. Figure 3.4 presents overall details for all features, such as data types, number of records, missing values, and duplicates.

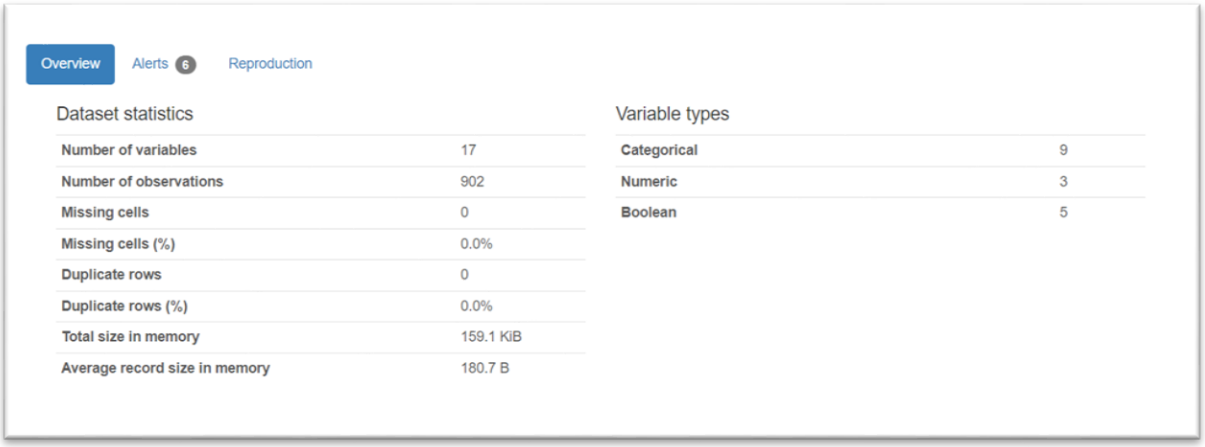


Figure 3.4 Overview of the Dataset

Figure 3.5 shows that males occupied (51.8%) of the data, while females occupied (48.2%), in this sense we can claim that there is an accepted balance between both classes which signifies gender non-bias in the data.

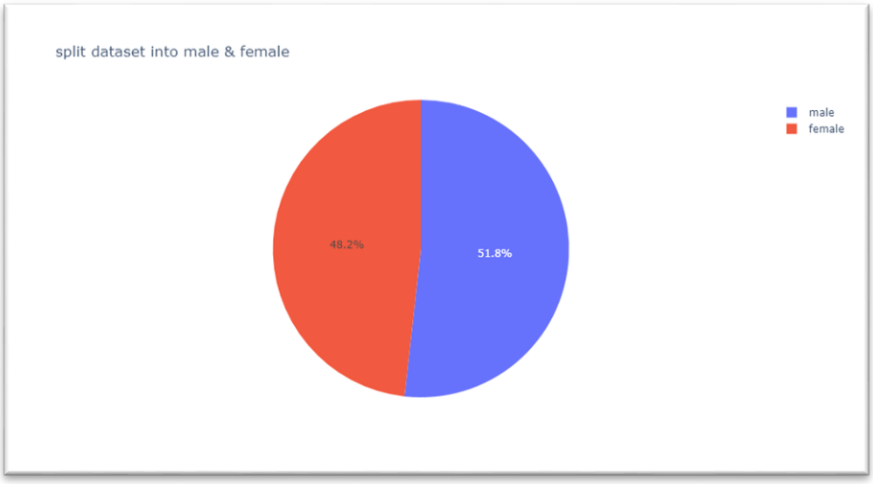


Figure 3.5 percentage of males & females

After exploring numeric values by using statistical figure boxplot. The measures contain a minimum value, Q1 (25%), Q2 (50% which is Median), Q3 (75%), maximum value and outliers. Figure 3.6 shows the distribution of weight based on gender. Males' 49 kg Min and 175 kg Max,

while females' 39 kg Min and 161 kg Max. Figure 3.7 shows us the distribution of height based on gender. Males' 158 cm Min and 200 cm Max, while females' 144 cm Min and 191 cm Max.

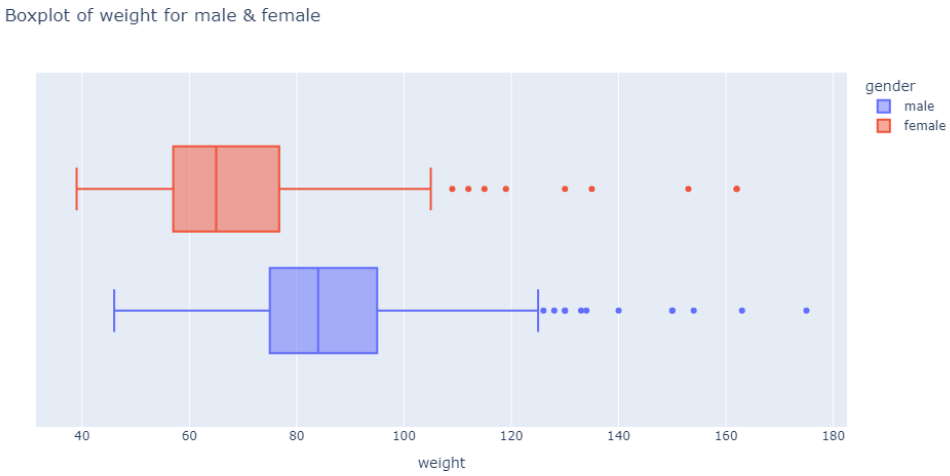


Figure 3.6 Boxplot for weight based on gender

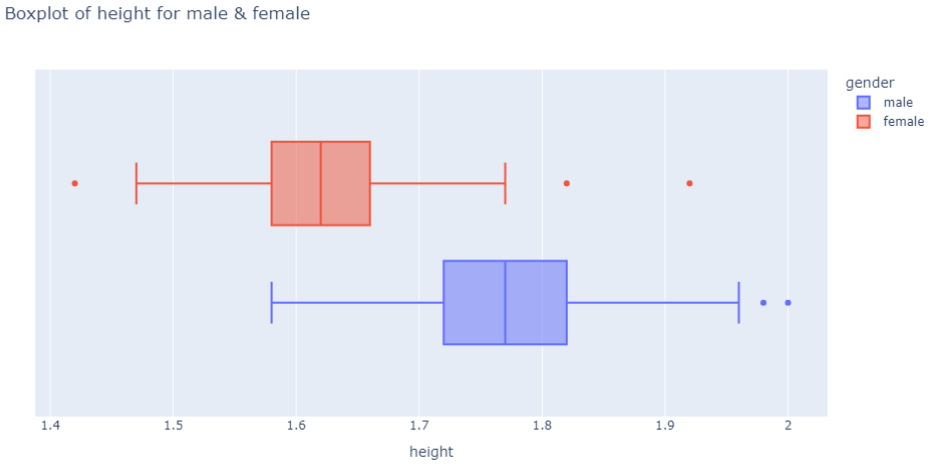


Figure 3.7 Boxplot for height based on gender

Figure 3.8, presents the family member who suffers from obesity or overweight based on gender. Approximately, the number of participants who have a family member obesity or overweight is

400. The number of participants who don’t have a family member suffering from obesity or overweight is 500.

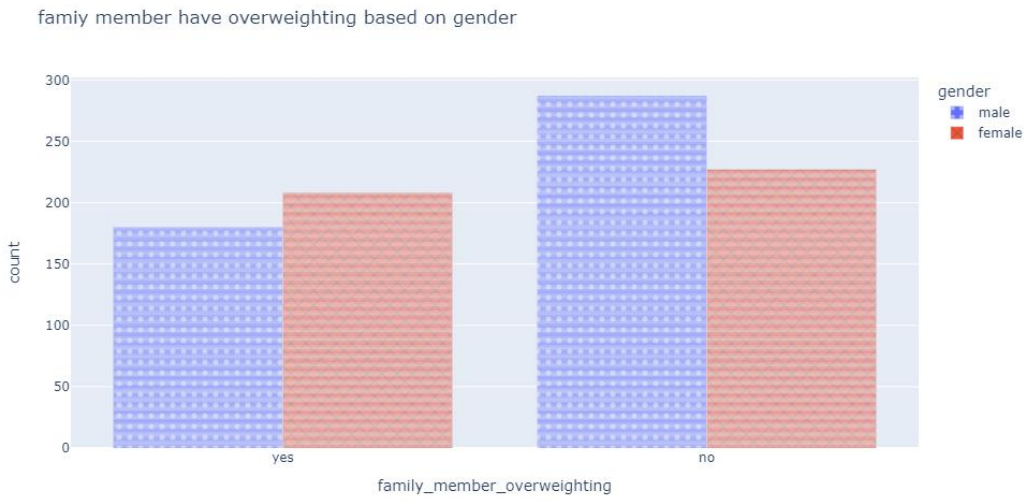


Figure 3.8 family member have overweight based on gender

Next, figure 3.9 presents the members of males and females who eat high caloric food, and the members who don’t eat high caloric food. Males are the high caloric food eaters, considering their daily meals, which may lead to obesity increased in this gender.

Figure 3.10, Presents the vegetables that are eaten in daily meals, it’s clear that there is balance in both genders. On the other hand, the vegetables are not the most eaten in the daily meals for the two genders.

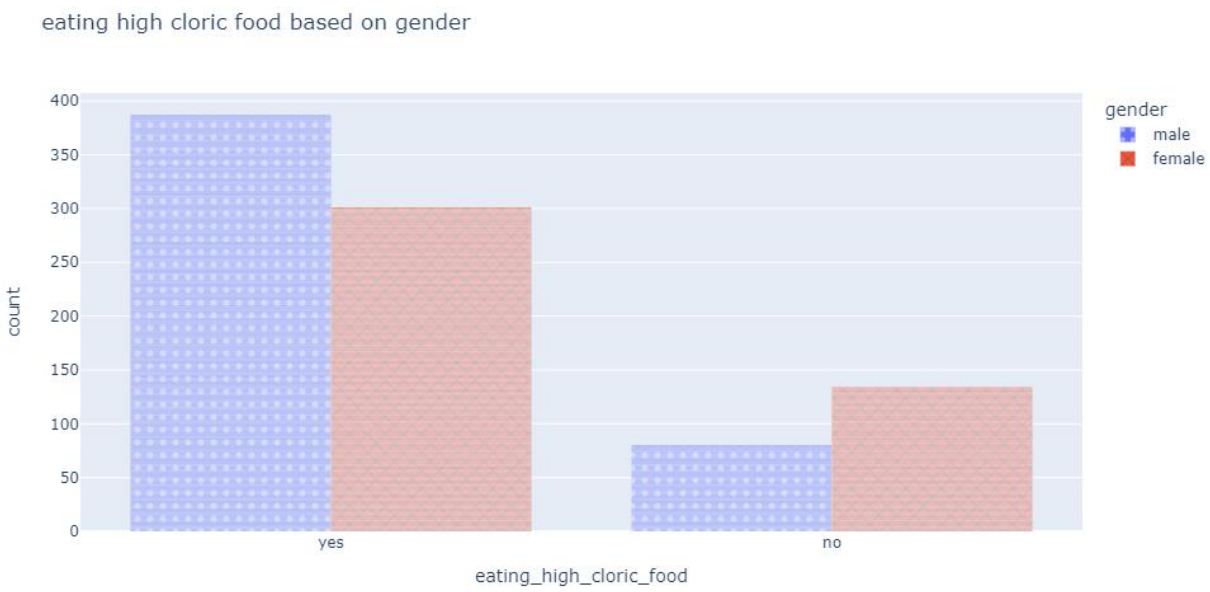


Figure 3.9 members eat high caloric food based on gender

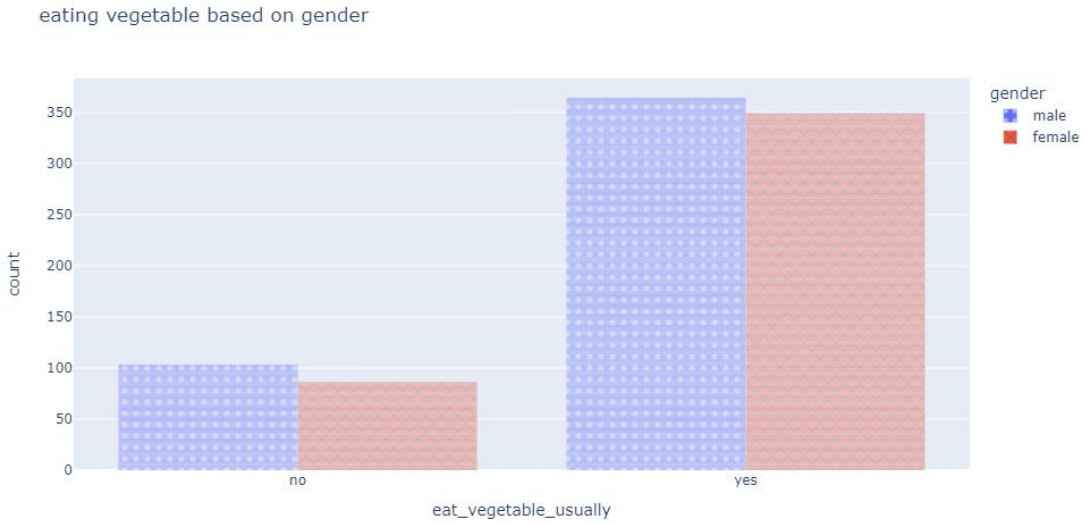


Figure 3.10 members eat vegetables based on gender

Figure 3.11, shows that non-smokers are more than smokers. The non-smoking females' number is higher than non-smoking males. However, smoking males' number is height than smoking females'.

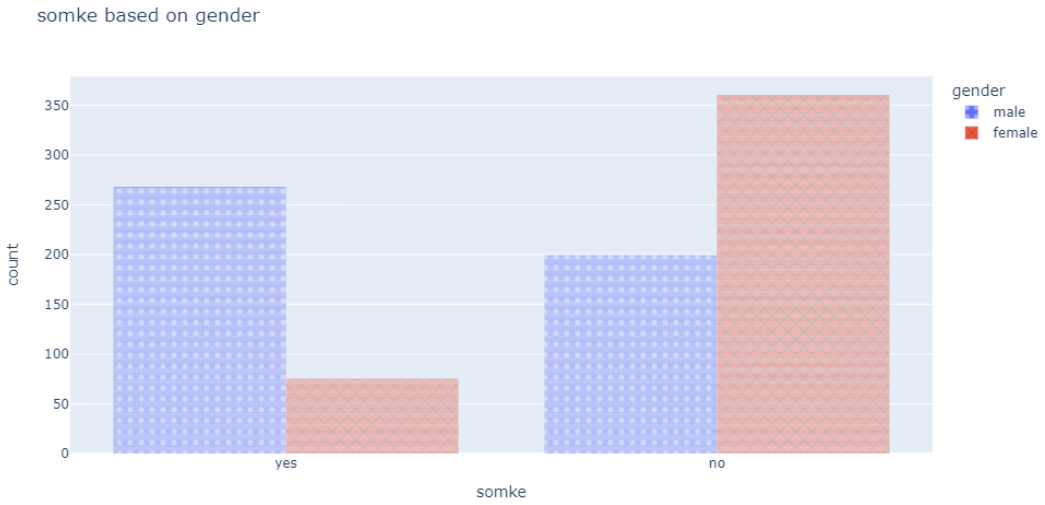


Figure 3.11 smokers or not based on gender

In figure 3.12, generally, there is no interest in monitoring calorie tracking on the meals. And there is a negative balance among males and females. While, for those who are monitoring calories in their meals, females' number is more than males' number.

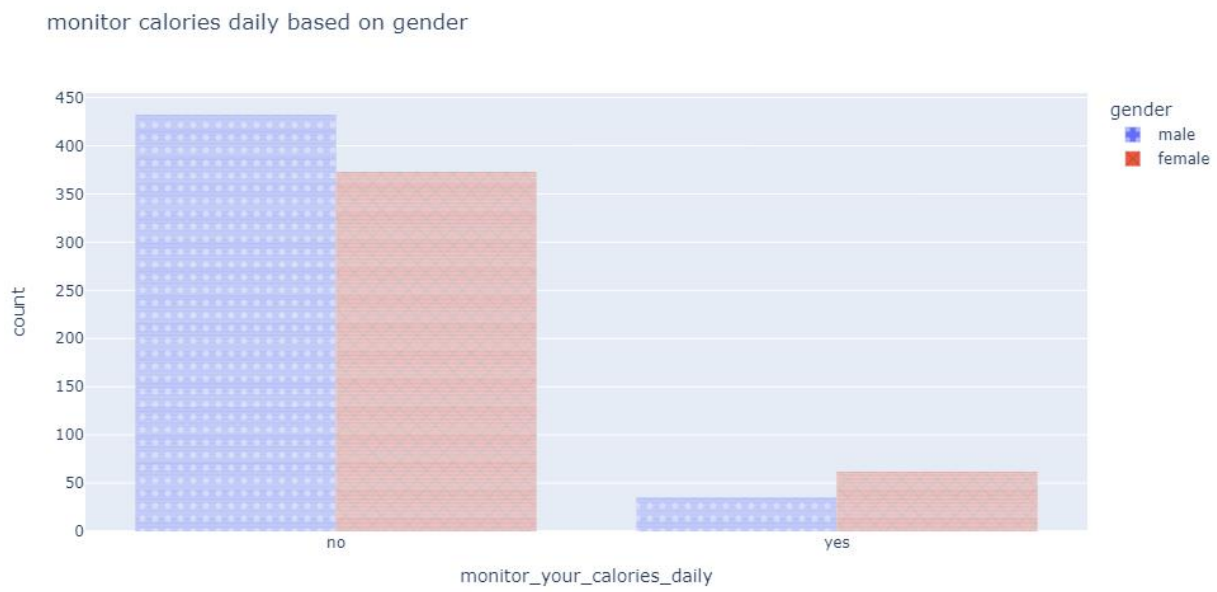


Figure 3.12 monitor calories or not based on gender

Then, in figure 3.13, most members eat one to two meals during their normal day, including males and females. Then, for those who eat three meals throughout a normal day, there is a balance between males and females. the least number of participants eat more than three meals a day, which includes males and females.

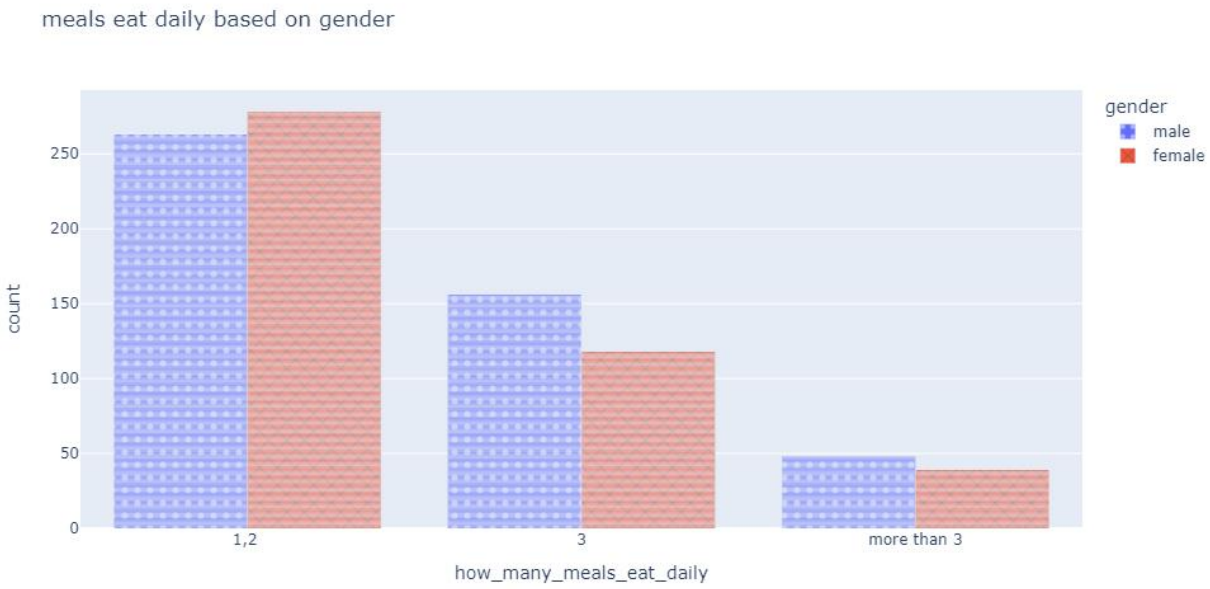


Figure 3.13 how many meals eat daily based on gender

Next, in figure 3.14, members who eat *sometime* between main meals have the highest count, and it applies to males and females. Then, the number of males' count who *never* eat between main meals is more than the number of females. Next, the number of members who *usually* eat between main meals is balanced between males and females. Finally, for those who *always* eat between males, females are more than males.

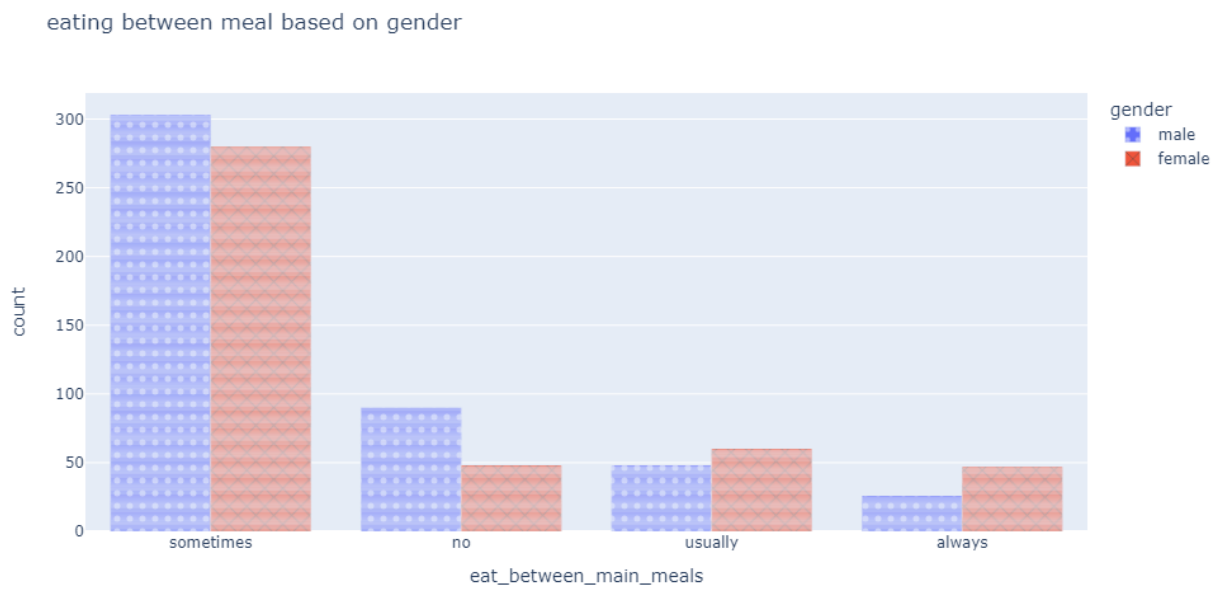


Figure 3.14 eat between meals based on gender

Figure 3.15 shows the result for drinking soft drinks. The highest count of those who *sometimes* drink soft drinks are males. While, for those who *do not drink* soft drinks, females are more than males. Then, for those who *mostly* drink soft drinks, the count is close to those who do not drink, and the males are more than females. Finally, there is a balance between males and females who *always* drink soft drinks.

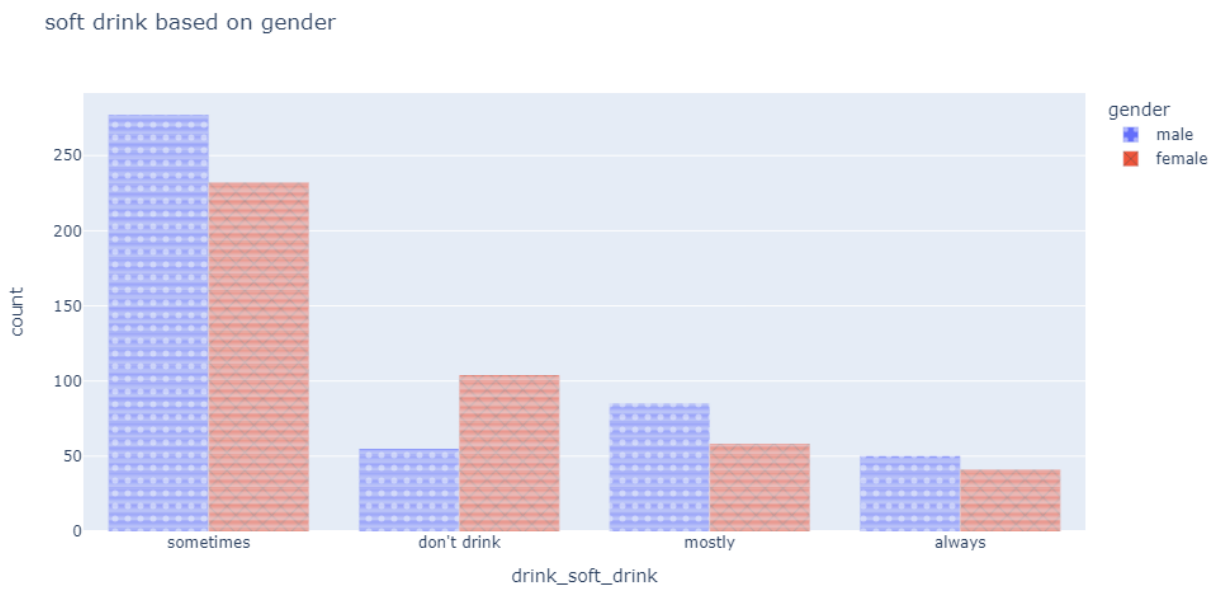


Figure 3.15 drink soft drink based on gender

Figure 3.16 shows the result for drinking water. The highest count for those who drink *one to two* liters daily are males, which is more than females. Then, females' count who drink *less* than one liter a day are more than males. Finally, for those who drink *more* than two liters daily, the males are more than females.

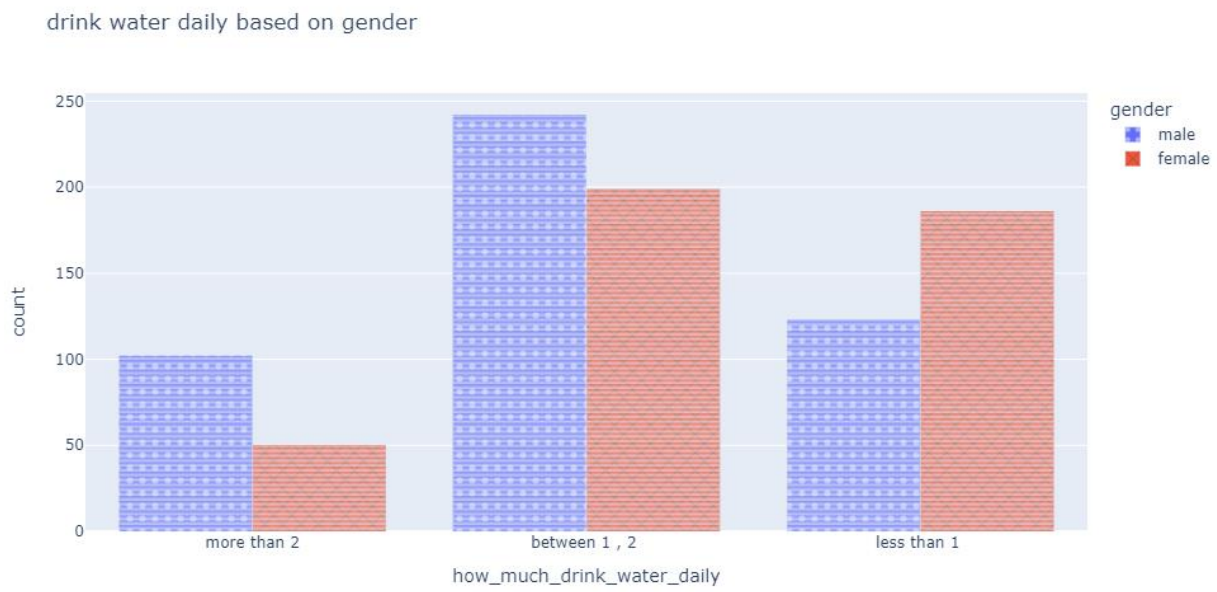


Figure 3.16 how much drink water daily based on gender

Then, in figure 3.17, the results of doing physical exercises during the week are shown. the count of members who never practice any physical exercises is almost equal for females and males. While, for those doing physical exercises for one or two days, weekly, the females are less than males. Then, for those doing physical exercises for three or four days a week, the balance between males and females appears. Finally, for those doing physical exercises five or six days weekly, there is a balance between males and females.

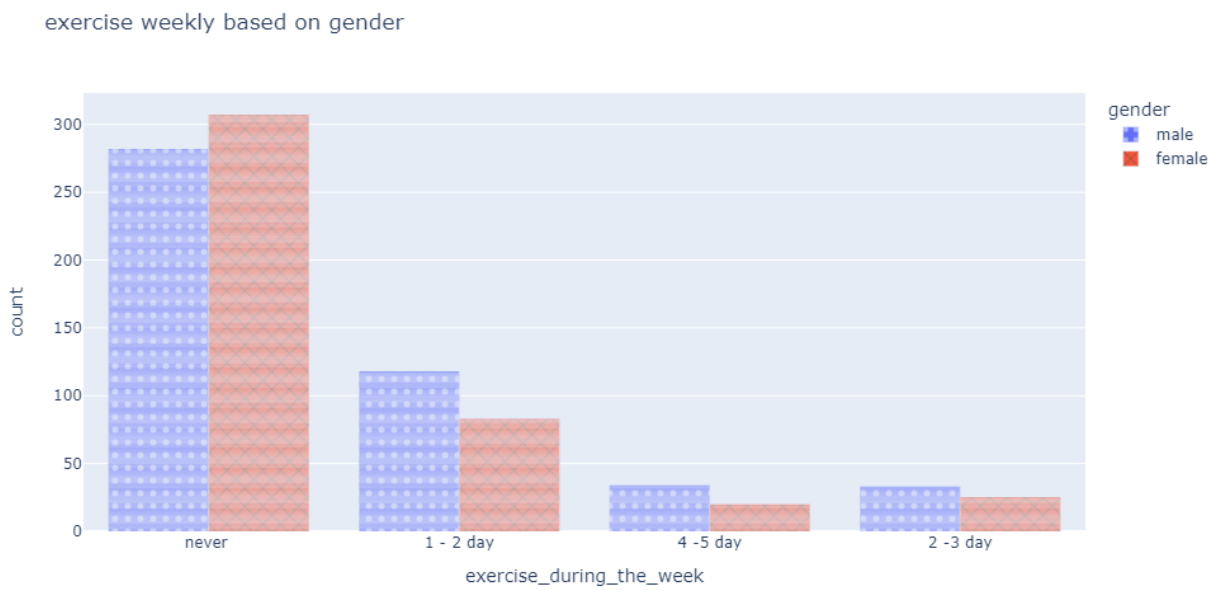


Figure 3.17 doing exercises weekly based on gender

Figure 3.18 shows the result of time-consuming on the mobile phone daily. The highest count of those who are using mobile phones for more than five hours daily are males. Then, for those using mobile phones for three to five hours, there is a balance between males and females. Finally, for those using mobile phones for less than two hours, the females are more than males.

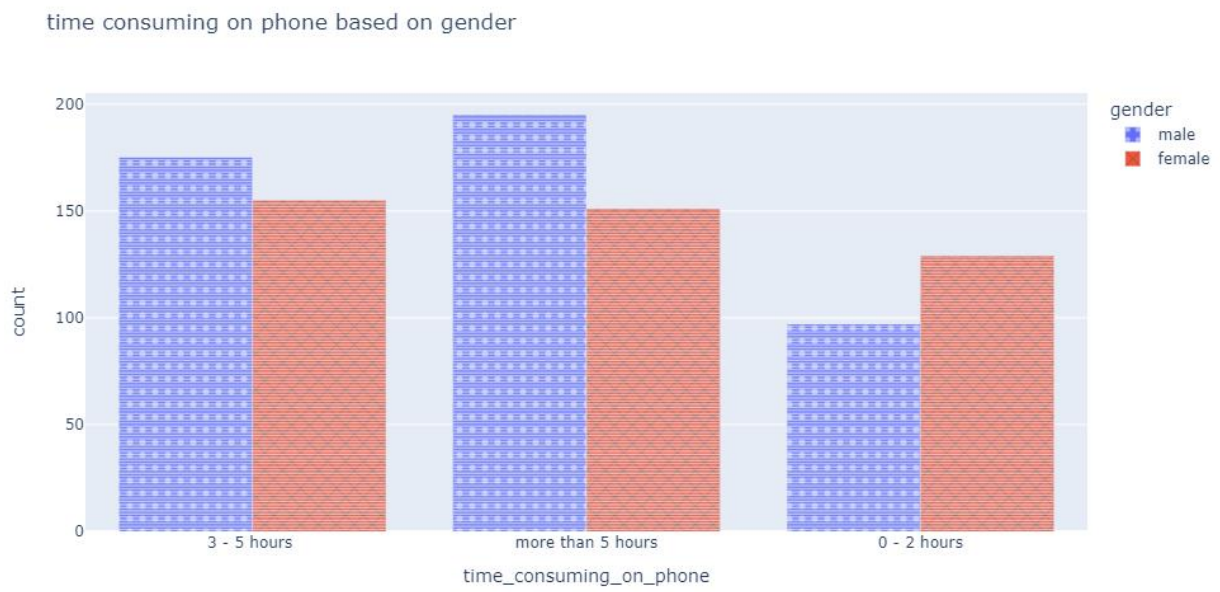


Figure 3.18 time consuming on phone based on gender

Figure 3.19 shows the result of transportation use. The highest count for private cars shows a balance between males and females. Also, using public transportation is balanced between males and females. Moreover, using walking as a way of transportation is balanced between males and females. There is less count for using motorcycles or bicycles for males.

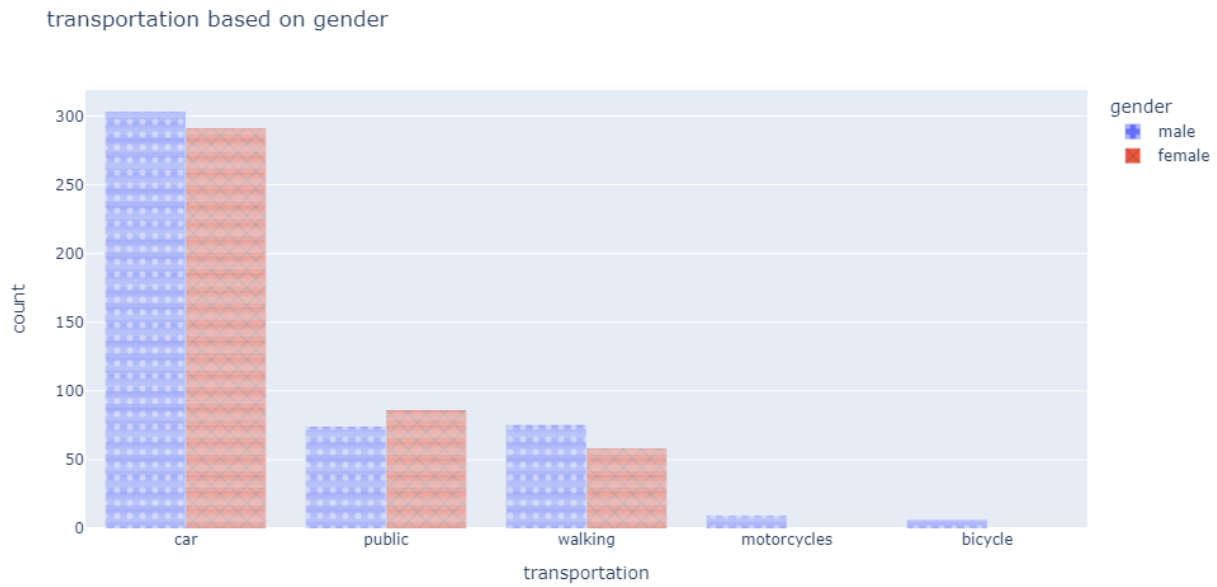


Figure 3.19 the way of transportation based on gender

Finally, figure 3.20 shows the relationship between an individual's height and weight, when analyzed in the context of gender, declares an interesting trend worth studying. It becomes evident that there exists a noticeable and obvious pattern between height and weight. When we study the height and weight on a graph split based on gender, a distinct positive slope emerges. This positive slope signifies that as height increases, weight tends to increase as well, and vice versa.

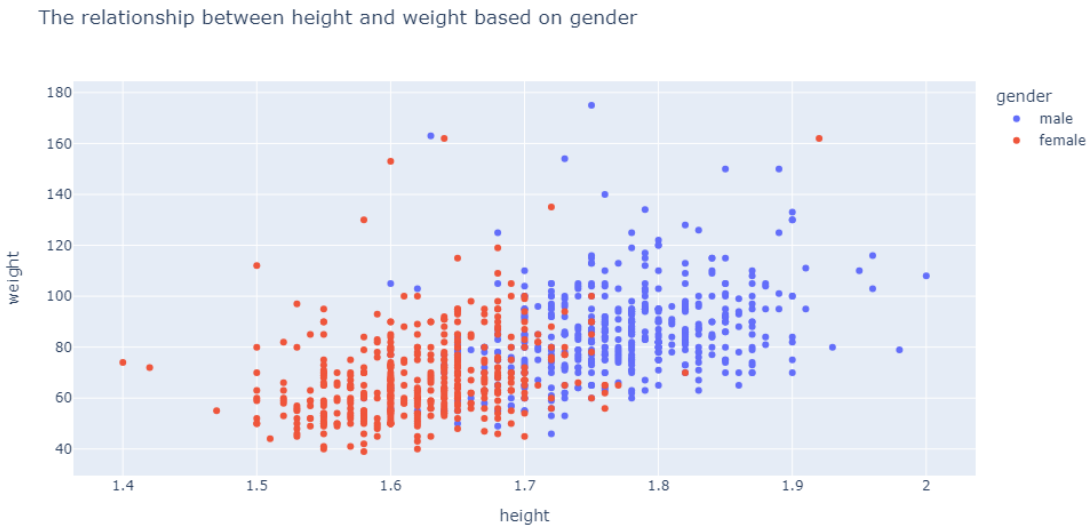


Figure 3.20 the relation between height and weight based on gender

## 3.5 Proposed Models

In this thesis, different ML models are used to classify obesity categories based on risk factors. The applied ML models will used to predict one of the fourth categories for obesity (underweight, normal, overweight, obese). Such as; SVM, RF, DT, MLPNNS, GB, XGBoost, AdaBoost, and Extra Tree Classifier models. To develop the preferment of these models in classification a *fine-tuning* step for the models via using techniques such as grid search to achieve more accurate results.

### 3.5.1 Support Vector Machines

Support Vector Machine (SVM) is one of the supervised machine learning algorithms used for linear or nonlinear classification and regression duties and even outlier detections. SVM can easily be conducted on a high dimensional dataset effectively (*1.4. Support Vector Machines — Scikit-Learn 1.3.0*

*Documentation*, n.d.)(Farnham et al., n.d.). In classification, the SVM can deal with binary or multiclass. SVM intends to find the best hyperplane in features space to distinguish data points of multiple classes. The hyperplane strives for a vast suitable distance between the nearest points of various classes. The number of features plays an important role in determining the dimension of hyperplane.

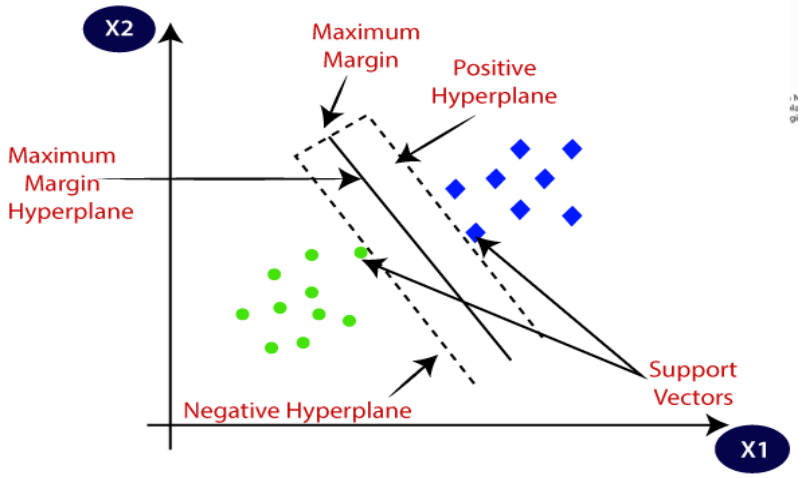


Figure 3.21 The basics of SVM (*Support Vector Machine(SVM): A Complete Guide for Beginners*, n.d.)

Figure 3.21 represents the basics of SVM, and how it works to separate data points into two classes and easily predict new data points based on the maximum marine hyperplane between support vectors that identify the best hyperplane. The types of hyperplane maximum margin classifier (hard margin), support vector classifier (soft margin), and support vector machines (non-linear).

The SVM equation is a linear equation to make a classification for data points and try to find the optimal hyperplane and the equation is (Farnham et al., n.d.)

$$W * X + b = 0 \tag{2}$$

Where x is the input vector, W is the weight vector and b is bias.

### *3.5.2 Decision Trees*

Decision Trees (DTs) are used in data mining methods to extract information from datasets, DTs are considered a non-parametric supervised machine learning utilized in classification or regression tree. The goal of DTs is to create a model that can predict the target by learning rules inferred from data attributes, it has a hierarchical shape, and the structure of the tree includes a root node, branches, internal nodes, and leaf nodes (terminal node), the DTs model assign the optimal solution or decision based on probability of target (Farnham et al., n.d.)(*1.10. Decision Trees — Scikit-Learn 1.3.0 Documentation*, n.d.).

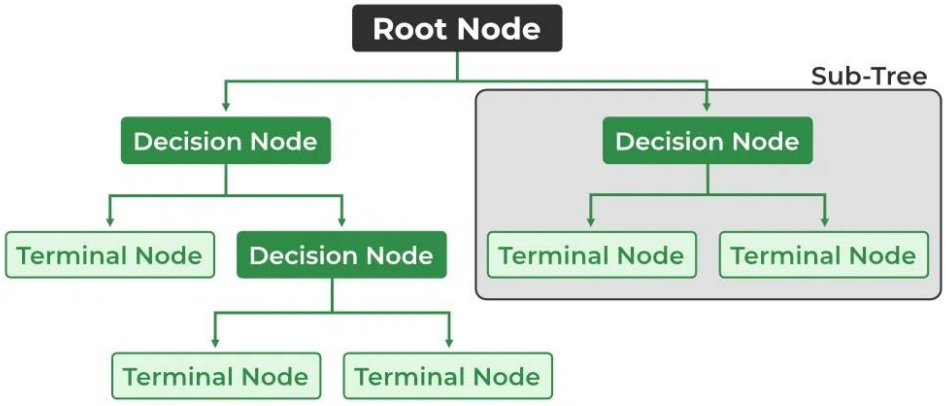


Figure 3.22 The basics of DTs (*Decision Tree - GeeksforGeeks*, n.d.)

In DTs, the nodes illustrate decision points or features, and branches represent the potential outcomes or decisions based on those attributes. The structure of the tree is built using training data and is used to make classifications on unknown or unseen data. DTs used Gini which measures

impurity utilized to build a decision tree or Entropy which measure of impurity or randomness in the data to select the root node and branches, internal nodes, and leaf nodes.

*The Gini formula is:*

$$1 - P(Class\ 1)^2 - P(Class\ 2)^2 \quad (3)$$

*The entropy formula is:*

$$- P(class\ 1)x\ log_2\big(P(class\ 1)\big) - P(class\ 2)x\ log_2(P(class\ 2)) \quad (4)$$

Where P is the probability and probability of the class by logarithm base= 2.

*Information Gain (IG)*

IG or reduction of Entropy which evaluates the uncertainty reduction for a given feature and determines which attribute should be chosen as a decision node or root node(Farnham et al., n.d.).

$$IG = E(Y) - E(Y/X) \quad (5)$$

Where E(Y) is the entropy for the dataset before any change. While E(Y/X) the conditional entropy is given by variable X.

### *3.5.3 Ensemble Methods*

Ensemble methods build thousands of trees instead of using one tree as in DT and collect these weak learners to find the most optimal answer for a strong learner. Voting classifiers use two approaches to obtain the optimal outcome, Bagging and Boosting. Bagging is creating independent samples from the training dataset; each sample takes a different subset of the training dataset with replacement. The bagging goal is to decrease the variance and overfitting by collecting all voters

and picking up the majority. While boosting takes another approach without replacement and the process is carried on through sequence samples, every tree solves the issues from the previous tree. Boosting goals reduces the training error at each iteration (Farnham et al., n.d.)(Kunapuli, n.d.).

### *3.5.3.1 Random Forests*

Random Forests are supervised machine learning algorithms, where hundreds or thousands of decision trees are built, and they gather all the results to make predictions or classifications. RF used Bagging techniques to find the optimal solution which is a strong tool to avoid overfitting and extract complex interaction between features and provide the feature importance (Farnham et al., n.d.)(Kunapuli, n.d.).

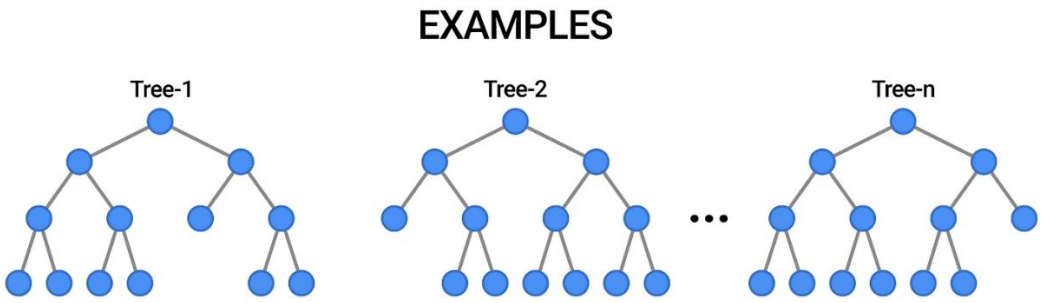


Figure 3.23 The basics of RF (How Random Forests & Decision Trees Decide: Simply Explained With An Example In Python | by Serafeim Loukas, PhD | Towards Data Science, n.d.)

The equation of RF is (Farnham et al., n.d.):

$$F(x) = 1/N \sum_{i} ((fi(x)) \tag{6}$$

Where F(x) is the output of the RF algorithm, N = the number of decision trees in the forest.

### *3.5.3.2 Extra Tree Classifier*

Extra Tree Classifiers like RF belong to Bagging techniques in Ensemble methods while introducing additional randomness during the tree construction process (Farnham et al., n.d.)(Kunapuli, n.d.).

Table 3.1 RF vs ETC (Farnham et al., n.d.)(Geurts et al., 2006)

| | RF | ETC |
|---|---|---|
| **Randomness** | Random feature subsampling at each split. | Random feature and threshold selection at each split |
| **Handling of Randomness** | Majority voting or averaging of predictions | Majority voting of predictions |
| **Computational Efficiency** | Evaluate multiple thresholds for each feature | Selects random thresholds without evaluating multiple candidates |
| **Variance** | Lower variance due to reduced randomness | Slightly higher variance due to increased randomness |
| **Hyperparameter Tuning** | Common hyperparameters (n_estimators, max_depth, etc.) | Different default values for some hyperparameters (e.g., larger max_features) |
| **Training Speed** | Slightly slower due to evaluating multiple thresholds | Faster due to not evaluating multiple candidates |

### *3.5.3.3 AdaBoost*

AdaBoost, which is an abbreviation for Adaptive Boosting, is a supervised machine-learning algorithm that merges multiple weak classifiers to make a strong classifier. It uses Boosting approach in ensemble learning techniques and AdaBoost aims to improve the accuracy of predictions by iteratively adjusting the weights assigned to training examples. The main idea

beyond AdaBoost is to set higher weights to the misclassified case in each iteration, thereby emphasizing the difficult instances and forcing subsequent weak classifiers to focus on them (Farnham et al., n.d.)(Kunapuli, n.d.).

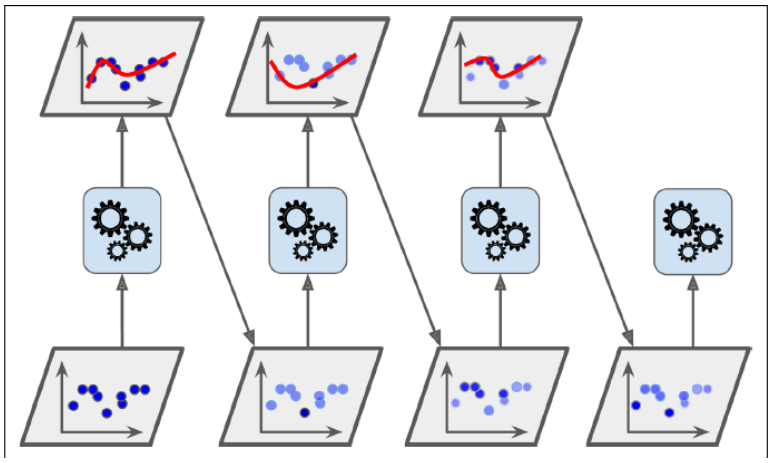

Figure 3.24 the basic of AdaBoost (Farnham et al., n.d.)

Adaboost equation is (Farnham et al., n.d.):

$$\hat{y}(x) = \underset{k}{\operatorname{argmax}} \sum_{\substack{j=1 \\ \hat{y}_j(x)=k}}^{N} a_j \tag{7}$$

Where N is the number of predictors.

### *3.5.3.4 Gradient Boosting*

Gradient Boosting (GB) is also an ensemble machine learning algorithm, like AdaBoost, using a sequential approach. GB aims to try to lessen errors from previous predictors. The "gradient" in Gradient Boosting refers to the technique's optimization process, which minimizes the loss function by repeatedly adding weak models to the ensemble. The privilege of using GB is to prevent overfitting, and also like other algorithms that could be used for classification or

regression, GB can handle complex datasets and obtain appropriate accuracy. Also, GB can handle missing in the data. Whatever kind of data, numerical or categorical, GB can handle it easily. Finally, GB could be applied to non-linear relationships and extract the pattern in the data (Farnham et al., n.d.)(Kunapuli, n.d.).

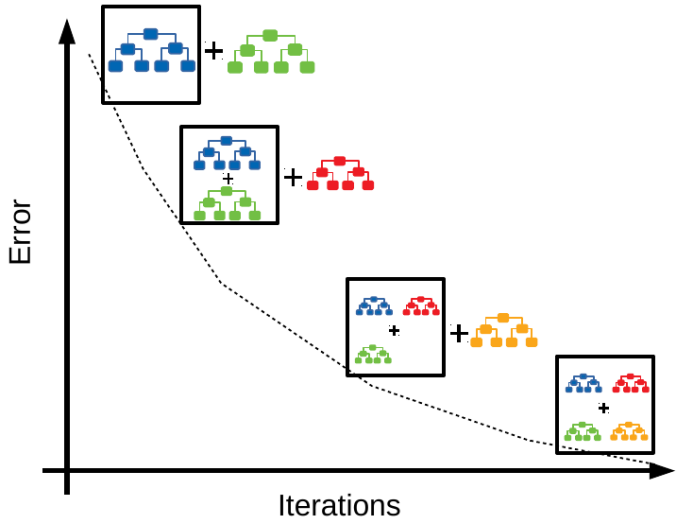


Figure 3.25 The Basics of Gradient Boosting (*Top 10 Interview Questions on Gradient Boosting Algorithms* -, n.d.)

### *3.5.3.5 XGBoost*

XGBoost is the development outcome of GB, which refers to Extreme Gradient Boosting. The goal of using XGBoost is to have a better version GB algorithm's ability. The advantage of XGBoost like GB is to prevent overfitting in datasets it could be used for classification or regression, and its ability to handle complex datasets and obtain appropriate accuracy. Also, XGBoost can handle missing in the data. Whatever kind of data, numerical or categorical, XGBoost can handle it efficiently. Finally, XGBoost could be applied to non-linear relationships and extract the pattern in the data (Farnham et al., n.d.)(Kunapuli, n.d.).

Table 3.2 AdaBoost vs GB vs XGBoost (Kunapuli, n.d.)

| | AdaBoost | Gradient Boosting | XGBoost |
|---|---|---|---|
| **Type** | Ensemble Learning | Ensemble Learning | Ensemble Learning |
| **Optimization** | An iterative process, adjusting weights | An iterative process, adjusting residuals | An iterative process, adjusting gradients |
| **Regularization** | No | Regularization techniques to control overfitting | Regularization techniques to control overfitting |
| **Handling Missing Values** | No | No | Yes, learns the best direction for missing values |
| **Parallel Computing** | No | No | Yes, column block parallelization |
| **Hyperparameter Tuning** | Limited | Moderate | Extensive |
| **Pruning** | No | No | Tree pruning |
| **Performance** | Good, but can be sensitive to noisy data | High performance, handles complex data | High performance, handles complex data |
| **Popular Uses** | Face detection, text classification | Various domains such as finance, healthcare, NLP | Various domains such as finance, healthcare, NLP |
| **Implementation** | Simple and easy to implement | More complex implementation | More optimized and efficient implementation |

## *3.6 Multilayer Perceptron Neural Networks*

Artificial Neural Networks (ANNs) is a supervised machine learning model. ANNs are computational models inspired by the functionality of biological neural networks in our brains. ANNs can deal with huge datasets to recognize hidden patterns and solve complex problems. An ANN contains interconnected nodes called artificial neurons or units. These units settle in layers, typically including an input layer as a first layer in the architecture, one or more hidden layers, and an output layer. The connections between these neurons are represented by weights and biases, which are modified weights and biases to minimize error during the learning process in every

iteration (Farnham et al., n.d.). Multi-Layer Perceptron is one type of Neural Network that consists of the input layer, at least one hidden layer, and an output layer

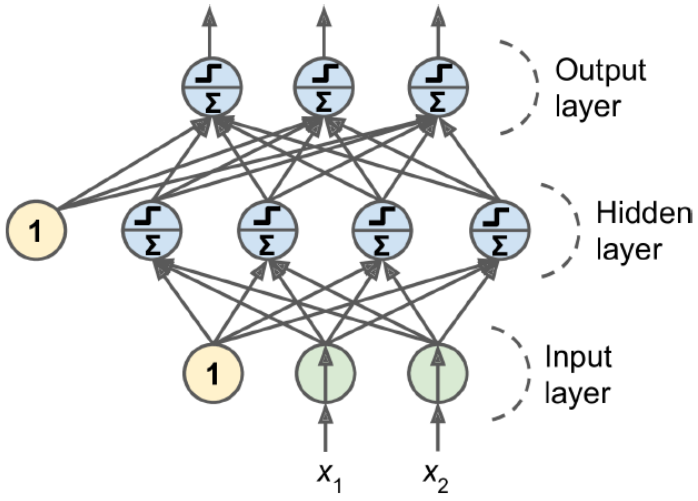


Figure 3.26 Architecture of Multilayer Perceptron NNs (Farnham et al., n.d.)

NNs have two stages: the first stage is forward. In this stage, the output is predicted and it calculates then sends the error back to the backward prorogation stage. The second stage is backward propagation, through this stage; the error is propagated back through the network to modify the values of the weights and to reduce the error value in the output layer. We can explain how NN works clearly through these main steps:

1. Input Data: In the first step we enter the training data for which we already know the correct output, after which it is processed through the hidden layers of the neural network.

2. Data processing by neurons (weights): At first, the input values are multiplied by the weights, according to this equation:

$$Y_{ij} = f\left(\sum_{i=1}^{n} Xi * w_{ij}\right) \quad (8)$$

Where $w_{ij}$ : is the connection weight between the *i*th node in the input layer and the *j*th node in the hidden layer, and $x_i$: is the *i*th input, and *f* is the activation function. It is worth noting that in each cell the layer is connected to all the cells in the next layer. The activation function *f* is calculated as in the following equation:

$$Y_j = \frac{1}{1 + e^{-y}} \tag{9}$$

Where $X_k$ is the input to the next layer's node.

3. Repeat the process with the other hidden layers: These operations are repeated with the other layers in the NNs until the last layer, which is the output layer.

4. Determine the output of the NN: The output values are ultimately determined by the type and function of the NNs, so we get the results and compare them with the real output.

5. Calculate the error: To break the process of training, there is a certain threshold θ is set depending on the error of the NN which represents the difference between the desired and actual output. The error is calculated using the following equation:

$$E = \frac{1}{2}\sum_{i}^{n}(y_d - y_j)^2 \tag{10}$$

Where $y_d$ is the desired output, $y_j$ is the actual output. So, by comparing the output with the results we already know.

6. Modify the weights to obtain more accurate results using the following equation:

$$w_{jk} = w_{jk} + \Delta w_{jk} \tag{11}$$

$$\Delta w_{j+1} = \alpha . E . x_i \tag{12}$$

$\alpha$ is the learning rate, which is a constant value between 0.01 to 1. The back-propagation algorithm repeats until the error on the output node is minimized

## *3.7 Tuning Parameter*

Tuning parameters in ML indicate the most appropriate set of parameters to boost the performance of models. These parameters should be determined in the training dataset, or hyperparameters which is the parameter tuned to achieve the most accurate results. Hyperparameter tuning can be chosen manually by an expert or can use automatic techniques to determine the hyperparameter such as grid search (evaluating all possible combinations), random search (sampling random combinations), or more advanced optimization techniques like Bayesian optimization or genetic algorithms (Farnham et al., n.d.). In this research, we propose using the Grid search technique to determine the best hyperparameter to achieve an appropriate result for our model.

SVM contains a few parameters that demand to be tuned for optimal performance for the model. There are parameters that, when tuned, could achieve appropriate results, such as the kernel function types that permit SVMs to deal with nonlinearity relationships between features and target variables, the regularization parameter (C) and gamma, (C) try to make the balance between training error and decision boundary complexity. Otherwise, gamma decides the influence of individual training examples. Using the grid search technique provides the most appropriate choice to determine the best parameter value to obtain appropriate results. Grid search technique will be used for all upcoming models to achieve the best values and results.

DT tuning parameters is an important step to enhancing the effectiveness of the model. There are many parameters that, when tuned, could achieve appropriate results, such as the maximum depth controls the depth of the tree, preventing overfitting when set too high. otherwise, A low value may result in underfitting. The minimum number of samples for splitting and leaf nodes control the stopping criteria for further partitioning in the tree. the criterion parameter specifies the quality measure used for splitting the nodes. The two commonly used criteria are Gini impurity and entropy to figure out the IG to determine the root of the tree and all nodes.

For tuning the parameters of bagging algorithms like Random Forest and Extra Trees classifiers, there are diverse parameters that can be tuned. Such as the number of trees (n_estimators), which controls the depth of trees (max_depth), minimum samples necessary to split an internal node (min_samples_split), minimum samples required at a leaf node (min_samples_leaf), the number of features considered for each split (max_features), and the use of bootstrap sampling (bootstrap).

For tuning the parameters of boosting algorithms like Adaboost, Gradient Boosting, and XGBoost, there are diverse parameters that can be tuned. Such as the learning rate (step size), the number of estimators (iterations), and weak learner parameters (e.g., maximum depth). In GB and XGBoost the sampling.

the tuning of the parameters of a Multilayer Perceptron (MLPNN), many parameters can be tuned. Such as the hidden layer of the architecture (the total number of layers and neurons in each layer), activation functions, learning rate, regularization techniques, batch size, number of epochs, and the optimization algorithm.

ML algorithms can be fine-tuned to obtain the most appropriate performance, highest result accuracy, and agile convergence, and to avoid overfitting or underfitting in our model.

## 3.8 Performance Evaluation

Evaluating the performance in ML is the technique to test the model results. It is considered an important stage in developing machine learning models. Also, the evaluation model's performance can help the data scientist to assess the model's behavior in with dealing the unseen dataset. To evaluate the model's performance there are many metrics to help the data scientist. Such as Confusing Matrix, precision, recall, F-1 score, and AUC and ROC curve. The cross-validation techniques can shuffle the dataset to avoid the pattern in the dataset which can help the models to achieve appropriate results for each model. After that, experts will compare the results between the model's performance by experts to determine which model could be the best one.

### 3.8.1 Confusion Matrix

The best way to evaluate classification algorithms is by applying the Confusion Matrix, which many researchers call the heart of the classification matrix. It is a tabular matrix including the values of actual results vs. predicted results and these values are (TP, TN, FP, FN) (Farnham et al., n.d.)(Bruce et al., n.d.)(Dangeti, n.d.).

| | Predicted: Yes | Predicted: No |
|---|---|---|
| **Actual: Yes** | TP | FN |
| **Actual: No** | FP | TN |

Figure 3.27 Confusion Matrix (Dangeti, n.d.)

Some concepts are used in a confusion matrix. TP is the number of true outcomes and the model classifies it as true. While FP is the number of false outcomes and the model classifies it as true. On the other hand. TN is the number of false outcomes and the model classifies it as false. FN is the number of true outcomes and the model classifies it as false. The most important formulas based on the Confusion Matrix are Accuracy, Precision, recall, and F-1 score (Farnham et al., n.d.)(Bruce et al., n.d.)(Dangeti, n.d.).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{13}$$

The accuracy formula can provide the percentage of correct predictions on all expectations.

$$Precision = \frac{TP}{(TP + FP)} \tag{14}$$

The precision can provide the percentage of correct values that are correct.

$$Recall = \frac{TP}{(TP + FN)} \tag{15}$$

The recall can provide the percentage of correct values that are predicted correctly.

$$F - 1\ score = \frac{2(precision * recall)}{(precision + recall)} \tag{16}$$

The f-1 score is a metric that seeks to measure both precision and recall.

### *3.8.2 ROC & AUC Curve*

The Receiver Operating Characteristic (ROC) is the curve that plots sensitivity on the y-axis against specificity on the x-axis. In other words, the relationship between True Positive Rate (TPR) and False Positive Rate (FPR). Aera Under Curve (AUC) points out the model's performance, more cover area under the curve is more accurate (Farnham et al., n.d.)(Bruce et al., n.d.)(Dangeti, n.d.).

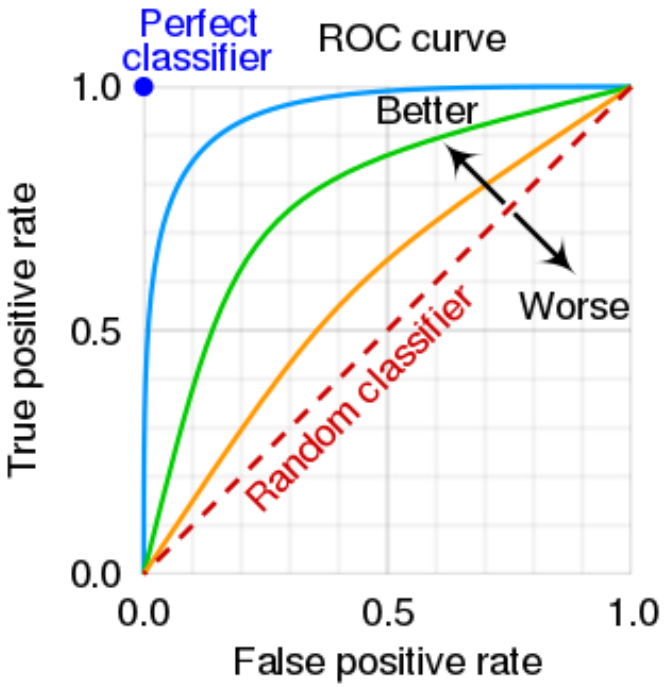


Figure 3.28 ROC & AUC Curve (Provost & Fawcett, 2001)

# Chapter 4. Result and Discussion

## *Experimental Results*

In this chapter, we represent the results of all ML models we propose to use in this research, the results will appear in two phases using the imbalance obesity dataset, phase two includes the using of SMOTE technique to deal with imbalance data. The results will include accuracy, precision, recall, F1 score, and AUC-ROC. Next, we will discuss the results for each model and compare the strengths and weaknesses spots for each model separately.

The Computing Environment in this research HP ZBook Firefly 15-inch G8 Mobile Workstation PC 11$^{th}$ Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz (8 CPUs), 16 GB RAM, 500 GB with Windows 11 Pro. For modeling Python, Anaconda Navigator, and Jupyter Notebook were used.

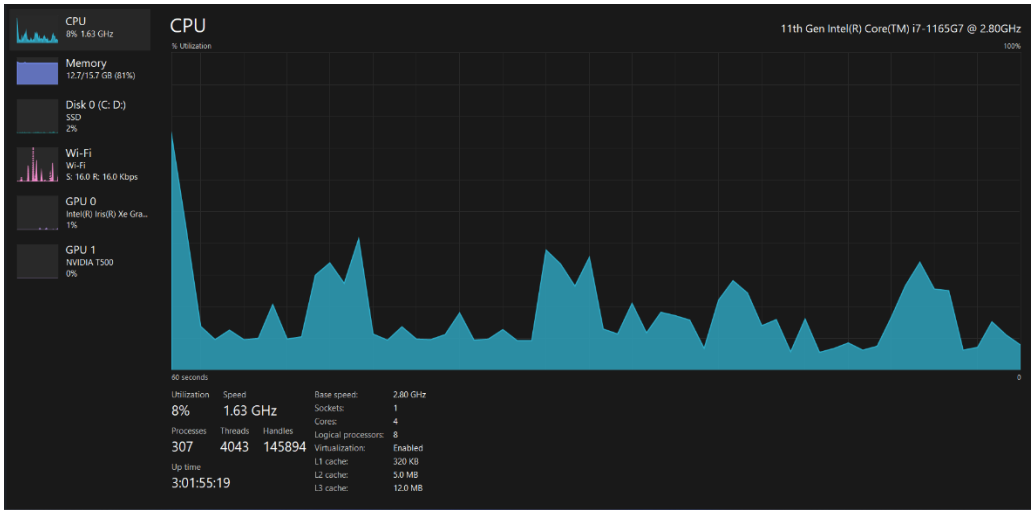


Figure 4.1 The performance of HP laptop and Jupyter Notebook

## 4.1 The Results of Classifications

The results of the ML classification models can be different based on the model and the training volume. The classification dataset was split into 80% for training and 20% for testing for each model using cross-validation.

### 4.1.1 Classification Result of SVM

SVM Applied on the dataset without dealing with imbalance data and using without Grid search technique for tuning the parameters. Figure 4.2 presents the performance metrics which include: Accuracy, Precision, Recall, and F1 Score. Figure 4.3 presents the AUC-ROC curve. As shown in figure 4.2, the results of accuracy are 98.3%, precision for all classes (96%, 100%, 100%, 100%) respectively, recall for all classes (100%, 100%, 97%, 83%) respectively, and f-1 score (98%, 100%, 99%, 91%) respectively.

```
The accuracy is
 0.9834254143646409
------------------
The Confusion marix
 [[71  0  0  0]
 [ 0 32  0  0]
 [ 2  0 70  0]
 [ 1  0  0  5]]
------------------
The Classification Report
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        71
           1       1.00      1.00      1.00        32
           2       1.00      0.97      0.99        72
           3       1.00      0.83      0.91         6

    accuracy                           0.98       181
   macro avg       0.99      0.95      0.97       181
weighted avg       0.98      0.98      0.98       181
```

Figure 4.2 The results of SVM with imbalance data

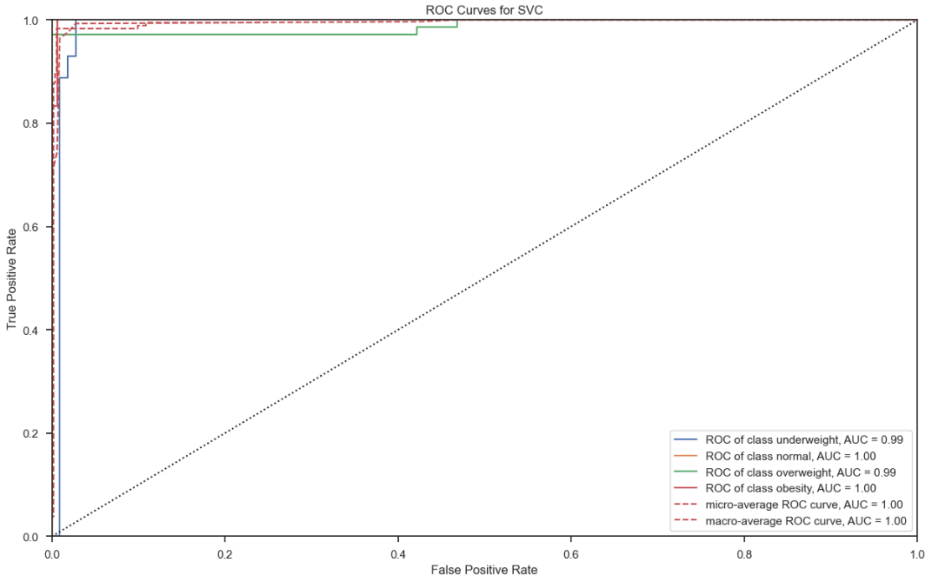


Figure 4.3 The AUC-ROC of SVM with imbalance data

On the other hand, when applying SVM to imbalance data by using the SMOTE technique and Grid search. Figure 4.4 presents the performance metrics which include: Accuracy, Precision, Recall, and F1 Score. Where figure 36 presents the AUC-ROC curve. As shown in figure 4.5, the produced results of accuracy are 95%, precision for all classes (92%, 97%, 94%, 99%)

respectively, recall for all classes (98%, 94%, 90%, 99%) respectively, and f-1 score (95%, 96%, 92%, 99%) respectively.

```
The accuracy is
 0.9545454545454546
------------------
The Confusion marix
 [[65  0  0  1]
 [ 0 68  4  0]
 [ 5  2 66  0]
 [ 1  0  0 74]]
------------------
The Classification Report
              precision    recall  f1-score   support

           0       0.92      0.98      0.95        66
           1       0.97      0.94      0.96        72
           2       0.94      0.90      0.92        73
           3       0.99      0.99      0.99        75

    accuracy                           0.95       286
   macro avg       0.95      0.96      0.95       286
weighted avg       0.96      0.95      0.95       286
```

Figure 4.4 The results of SVM with balanced data

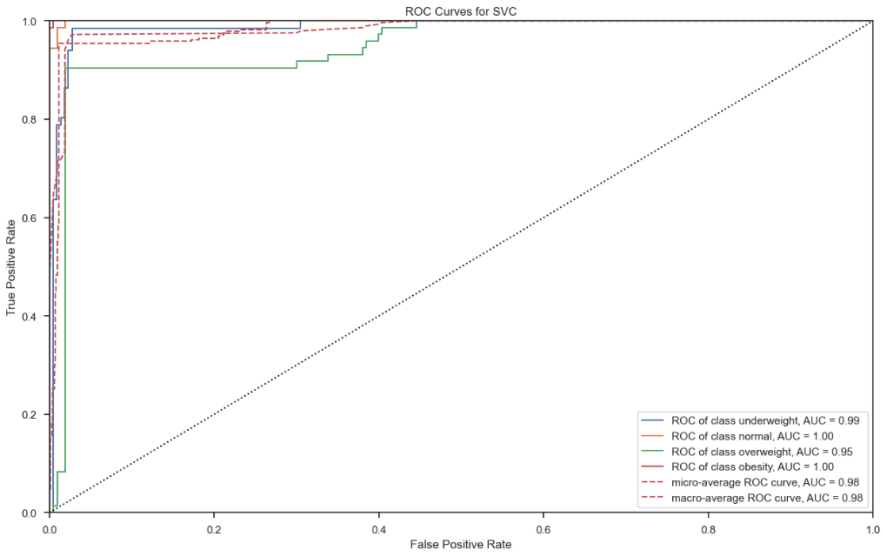


Figure 4.5 The AUC-ROC of SVM with balanced data

### *4.1.2 Decision Trees*

Phase 1: DT without dealing with imbalance data and using grid search technique to determine the best set of parameters, figure 4.6 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm DT, which include: Accuracy, Precision, Recall, and F1 Score. Figure 4.7 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in figure 4.6, the results we got was the accuracy of 90%, precision for all class (82%, 100%, 97%, 67%) respectively, recall for all classes (96%, 100%, 85%, 33%) respectively, and f-1 score (88%, 100%, 90%, 44%) respectively.

```
The accuracy is
 0.9005524861878453
------------------
The Confusion marix
 [[68  0  2  1]
 [ 0 32  0  0]
 [11  0 61  0]
 [ 4  0  0  2]]
------------------
The Classification Report
              precision    recall  f1-score   support

           0       0.82      0.96      0.88        71
           1       1.00      1.00      1.00        32
           2       0.97      0.85      0.90        72
           3       0.67      0.33      0.44         6

    accuracy                           0.90       181
   macro avg       0.86      0.78      0.81       181
weighted avg       0.91      0.90      0.90       181
```

Figure 4.6 The results of DT with imbalance data

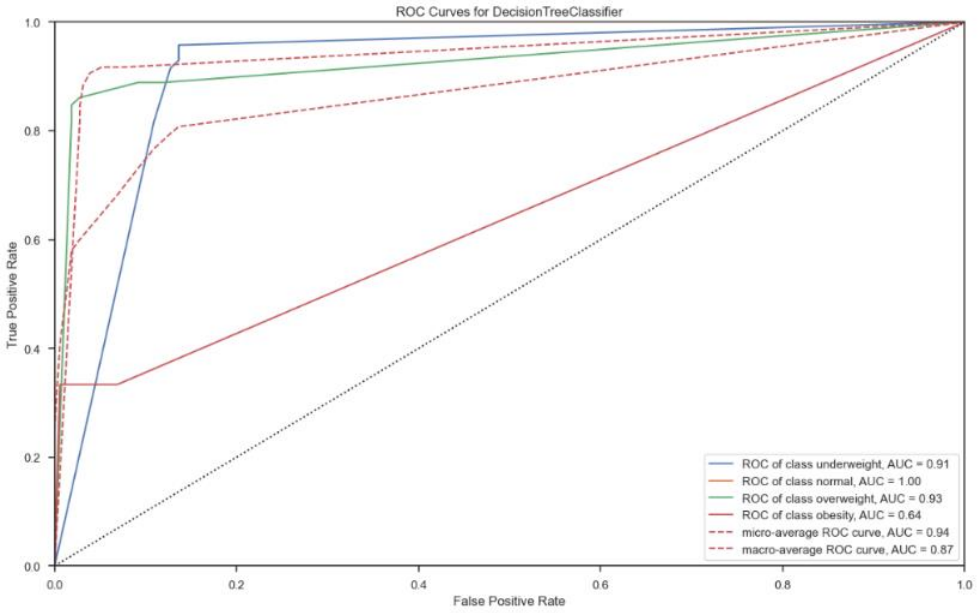


Figure 4.7 The AUC-ROC of DT with imbalance data

Phase 2: DT with dealing with imbalance data by using the SMOTE technique and using grid search technique to determine the best set of parameters, figure 4.8 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm DT, which include: Accuracy, Precision, Recall, F1 Score. The figure 4.9 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in figure 4.8, the results we got was the accuracy of 98%, precision for all class (97%, 99%, 100%, 97%) respectively, recall for all classes (97%, 99%, 100%, 97%) respectively, and f-1 score (97%, 99%, 99%, 97%) respectively.

```
The accuracy is
 0.9825174825174825
------------------
The Confusion marix
 [[64  0  0  2]
 [ 0 71  1  0]
 [ 0  0 73  0]
 [ 2  0  0 73]]
------------------
The Classification Report
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        66
           1       1.00      0.99      0.99        72
           2       0.99      1.00      0.99        73
           3       0.97      0.97      0.97        75

    accuracy                           0.98       286
   macro avg       0.98      0.98      0.98       286
weighted avg       0.98      0.98      0.98       286
```

Figure 4.8 The results of DT with balance data

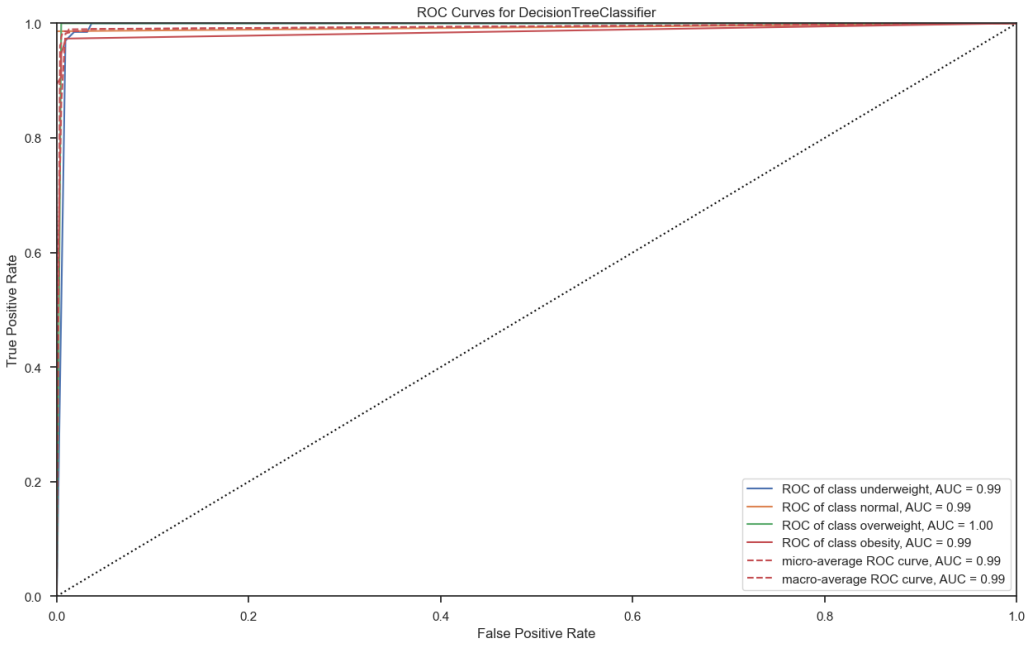


Figure 4.9 The AUC-ROC of DT with balanced data

### *4.1.3 Random Forests*

Phase 1: RF without dealing with imbalance data and using grid search technique to determine the best set of parameters, figure 4.10 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm RF, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.11 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.10, the results we got was the accuracy of 98%, precision for all class (96%, 100%, 100%, 100%) respectively, recall for all classes (100%, 100%, 100%, 50%) respectively, and f-1 score (98%, 100%, 100%, 67%) respectively.

```
The accuracy is
 0.9834254143646409
------------------
The Confusion marix
 [[71  0  0  0]
 [ 0 32  0  0]
 [ 0  0 72  0]
 [ 3  0  0  3]]
------------------
The Classification Report
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        71
           1       1.00      1.00      1.00        32
           2       1.00      1.00      1.00        72
           3       1.00      0.50      0.67         6

    accuracy                           0.98       181
   macro avg       0.99      0.88      0.91       181
weighted avg       0.98      0.98      0.98       181
```

Figure 4.10 The results of RF with imbalance data

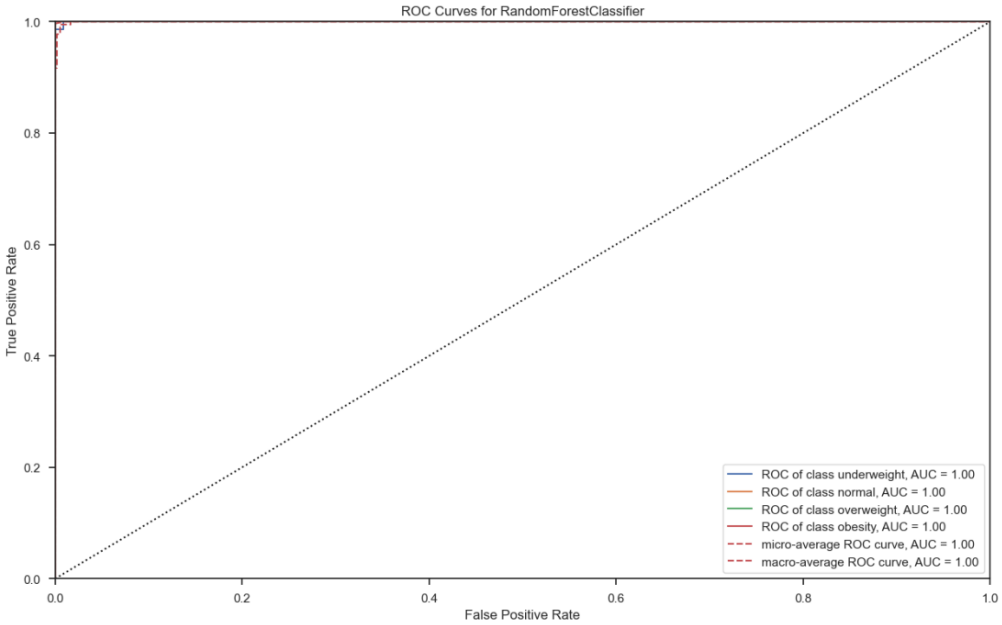


Figure 4.11 The AUC-ROC of RF with imbalance data

Phase 2: RF deals with imbalance data by using SMOTE technique and using grid search technique to determine the best set of parameters, figure 4.12 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm RF, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.13 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in figure 4.12, the results we got was an accuracy of 98.6%, precision for all class (97%, 100%, 97%, 100%) respectively, recall for all classes (100%, 97%, 100%, 97%) respectively, and f-1 score (99%, 99%, 99%, 99%) respectively.

```
The accuracy is
 0.986013986013986
------------------
The Confusion marix
 [[66  0  0  0]
 [ 0 70  2  0]
 [ 0  0 73  0]
 [ 2  0  0 73]]
------------------
The Classification Report
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        66
           1       1.00      0.97      0.99        72
           2       0.97      1.00      0.99        73
           3       1.00      0.97      0.99        75

    accuracy                           0.99       286
   macro avg       0.99      0.99      0.99       286
weighted avg       0.99      0.99      0.99       286
```

Figure 4.12 The results of RF with balanced data

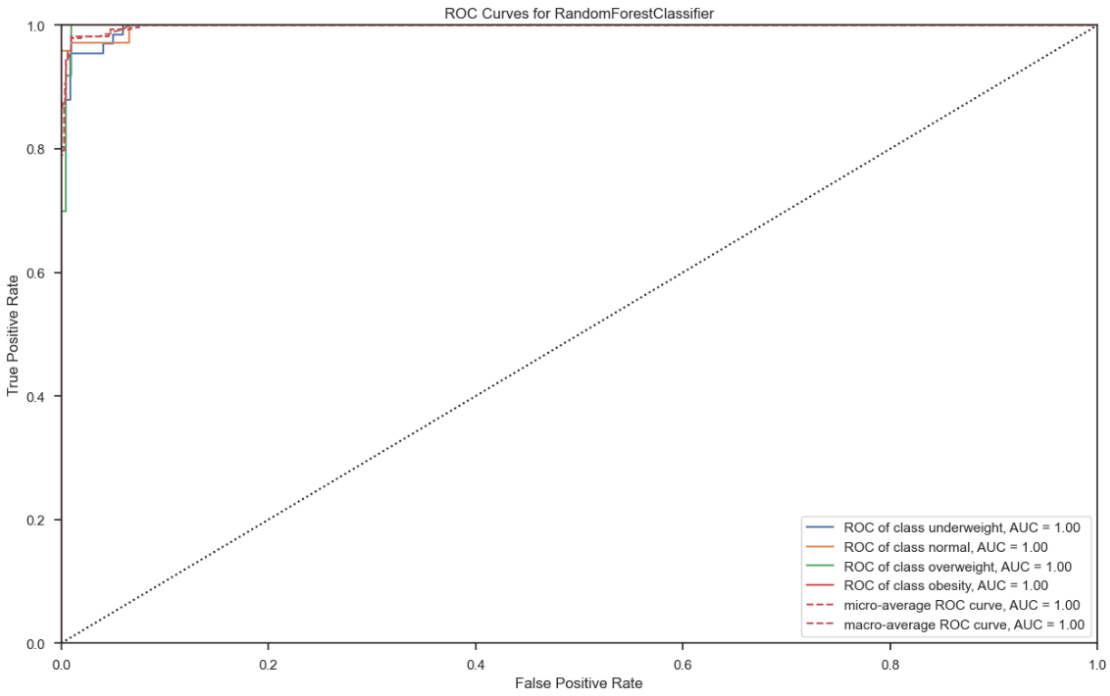


Figure 4.13 The AUC-ROC of RF with balanced data

### *4.1.4 Extra Tree Classifier*

Phase 1: ETC without dealing with imbalance data and using grid search technique to determine the best set of parameters, figure 4.14 represents the performance metrics used to evaluate the effectiveness of applying supervised machine learning algorithm ETC, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.15 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.14, the results we got was the accuracy of 84%, precision for all class (77%, 100%, 87%, 00%) respectively, recall for all classes (92%, 91%, 82%, 00%) respectively, and f-1 score (84%, 95%, 84%, 00%) respectively.

```
The accuracy is
 0.8453038674033149
------------------
The Confusion marix
 [[65  0  6  0]
 [ 0 29  3  0]
 [13  0 59  0]
 [ 6  0  0  0]]
------------------
              precision    recall  f1-score   support

           0       0.77      0.92      0.84        71
           1       1.00      0.91      0.95        32
           2       0.87      0.82      0.84        72
           3       0.00      0.00      0.00         6

    accuracy                           0.85       181
   macro avg       0.66      0.66      0.66       181
weighted avg       0.83      0.85      0.83       181
```

Figure 4.14 The results of ETC with imbalanced data

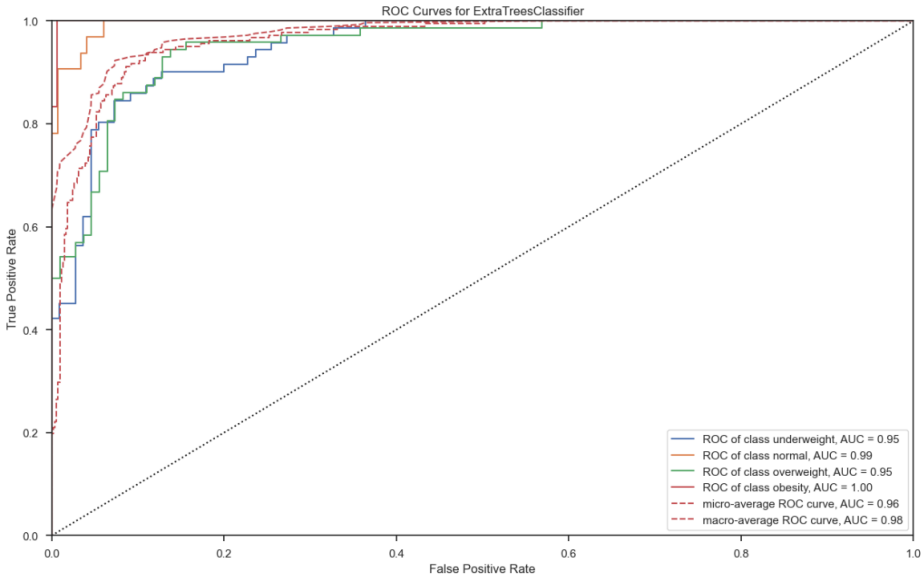


Figure 4.15 The AUC-ROC of ETC with imbalanced data

Phase 2: ETC deals with imbalance data by using the SMOTE technique and using grid search technique to determine the best set of parameters, figure 4.16 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm ETC, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.17 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.16, the results we got was the accuracy of 78%, precision for all class (74%, 74%, 79%, 86%) respectively, recall for all classes (76%, 83%, 56%, 97%) respectively, and f-1 score (75%, 78%, 66%, 91%) respectively.

```
The accuracy is
 0.7832167832167832
------------------
The Confusion marix
 [[50  5  2  9]
 [ 3 60  9  0]
 [13 16 41  3]
 [ 2  0  0 73]]
------------------
              precision    recall  f1-score   support

           0       0.74      0.76      0.75        66
           1       0.74      0.83      0.78        72
           2       0.79      0.56      0.66        73
           3       0.86      0.97      0.91        75

    accuracy                           0.78       286
   macro avg       0.78      0.78      0.77       286
weighted avg       0.78      0.78      0.78       286
```

Figure 4.16 The results of ETC with balanced data

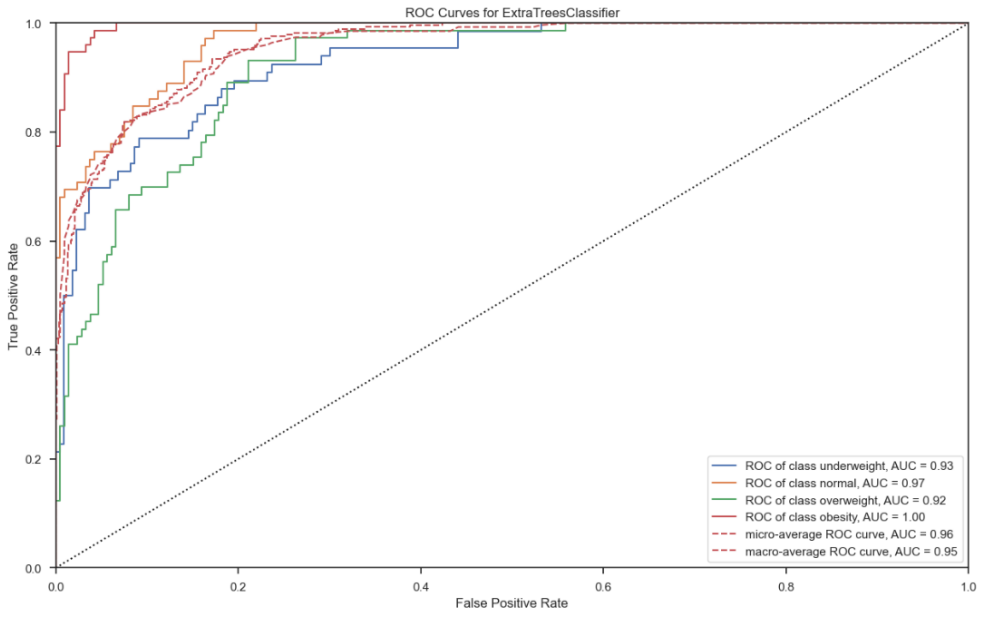


Figure 4.17 The AUC-ROC of ETC with balance data

### *4.1.5 AdaBoost*

Phase 1: AdaBoost without dealing with imbalance data and using grid search technique to determine the best set of parameters, Figure 4.18 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm AdaBoost, which include: Accuracy, Precision, Recall, F1 Score. And the figure 4.19 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.18, the results we got was the accuracy of 96.6%, precision for all class (92%, 100%, 100%, 00%) respectively, recall for all classes (100%, 100%, 100%, 00%) respectively, and f-1 score (96%, 100%, 100%, 00%) respectively.

```
The accuracy is
 0.9668508287292817
------------------
The Confusion marix
 [[71  0  0  0]
 [ 0 32  0  0]
 [ 0  0 72  0]
 [ 6  0  0  0]]
------------------
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        71
           1       1.00      1.00      1.00        32
           2       1.00      1.00      1.00        72
           3       0.00      0.00      0.00         6

    accuracy                           0.97       181
   macro avg       0.73      0.75      0.74       181
weighted avg       0.94      0.97      0.95       181
```

Figure 4.18 The results of AdaBoost with imbalanced data

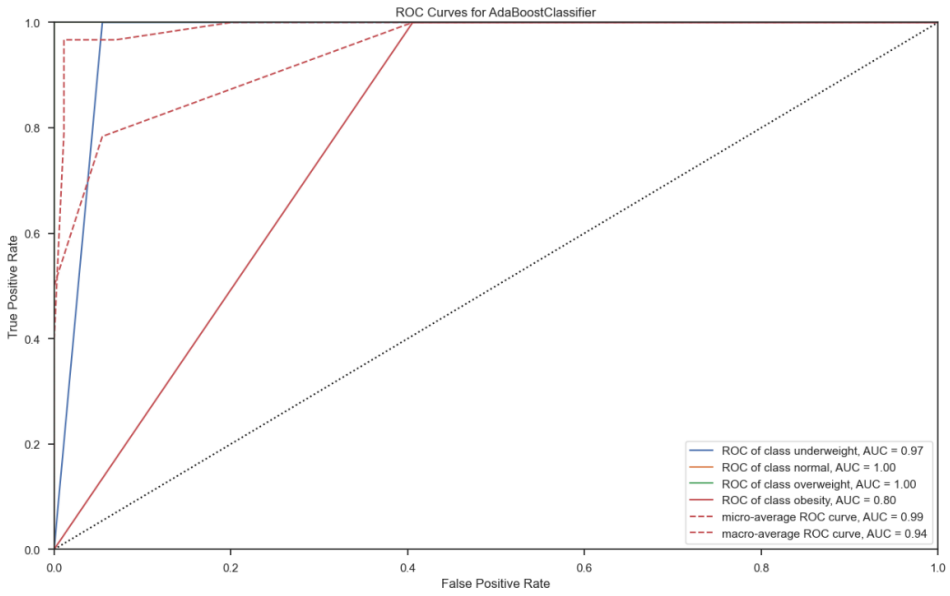


Figure 4.19 The AUC-ROC of AdaBoost with imbalanced data

Phase 2: AdaBoost deals with imbalance data by using the SMOTE technique and using grid search technique to determine the best set of parameters, Figure 4.20 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm AdaBoost, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.21 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.20, the results we got was the accuracy of 99.6%, precision for all class (100%, 100%, 99%, 100%) respectively, recall for all classes (100%, 99%, 100%, 100%) respectively, and f-1 score (100%, 99%, 99%, 100%) respectively.

```
The accuracy is
 0.9965034965034965
------------------
The Confusion marix
 [[66  0  0  0]
 [ 0 71  1  0]
 [ 0  0 73  0]
 [ 0  0  0 75]]
------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        66
           1       1.00      0.99      0.99        72
           2       0.99      1.00      0.99        73
           3       1.00      1.00      1.00        75

    accuracy                           1.00       286
   macro avg       1.00      1.00      1.00       286
weighted avg       1.00      1.00      1.00       286
```

Figure 4.20 The results of AdaBoost with balanced data

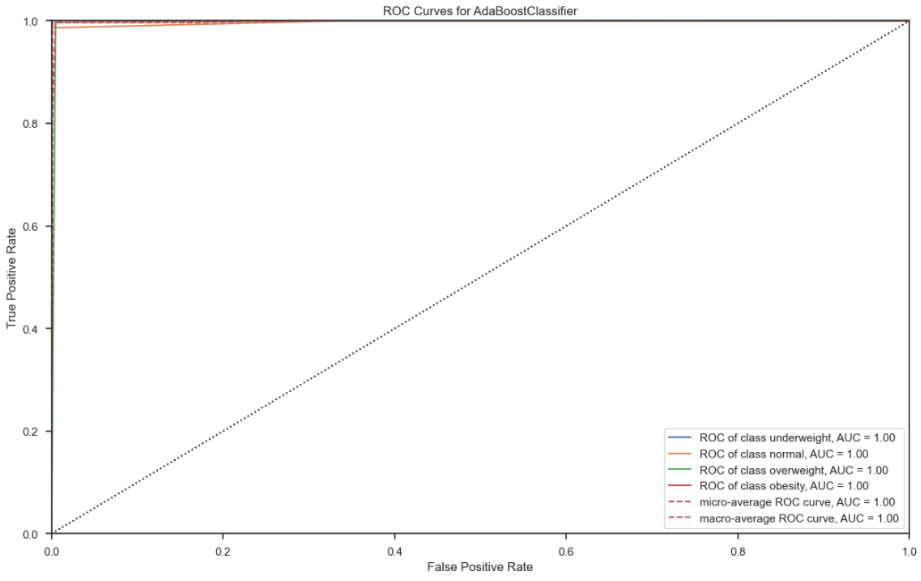


Figure 4.21 The AUC-ROC of AdaBoost with balanced data

### *4.1.6 Gradient Boosting*

Phase 1: GB without dealing with imbalanced data and using grid search technique to determine the best set of parameters, Figure 4.22 represents the performance metrics used to evaluate the effectiveness of applying supervised machine learning algorithm GB, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.23 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.22, the results we got was the accuracy of 100%, precision for all class (100%, 100%, 100%, 100%) respectively, recall for all classes (100%, 100%, 100%, 100%) respectively, and f-1 score (100%, 100%, 100%, 100%) respectively.

```
The accuracy is
 1.0
------------------
The Confusion marix
 [[71  0  0  0]
 [ 0 32  0  0]
 [ 0  0 72  0]
 [ 0  0  0  6]]
------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        71
           1       1.00      1.00      1.00        32
           2       1.00      1.00      1.00        72
           3       1.00      1.00      1.00         6

    accuracy                           1.00       181
   macro avg       1.00      1.00      1.00       181
weighted avg       1.00      1.00      1.00       181
```

Figure 4.22 The results of GB with imbalanced data

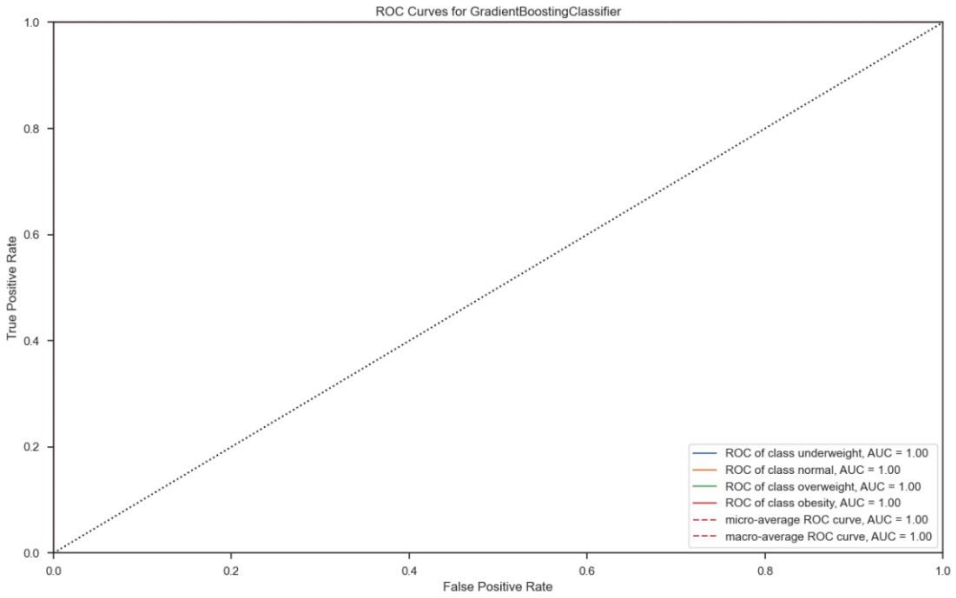


Figure 4.23 The AUC-ROC of GB with imbalanced data

Phase 2: GB deals with imbalance data by using the SMOTE technique and using grid search technique to determine the best set of parameters, Figure 4.24 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm GB, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.25 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.24, the results we got was the accuracy of 99.6%, precision for all class (100%, 100%, 99%, 100%) respectively, recall for all classes (100%, 99%, 100%, 100%) respectively, and f-1 score (100%, 99%, 99%, 100%) respectively.

```
The accuracy is
 0.9965034965034965
------------------
The Confusion marix
 [[66  0  0  0]
 [ 0 71  1  0]
 [ 0  0 73  0]
 [ 0  0  0 75]]
------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        66
           1       1.00      0.99      0.99        72
           2       0.99      1.00      0.99        73
           3       1.00      1.00      1.00        75

    accuracy                           1.00       286
   macro avg       1.00      1.00      1.00       286
weighted avg       1.00      1.00      1.00       286
```

Figure 4.24 The results of GB with balance data

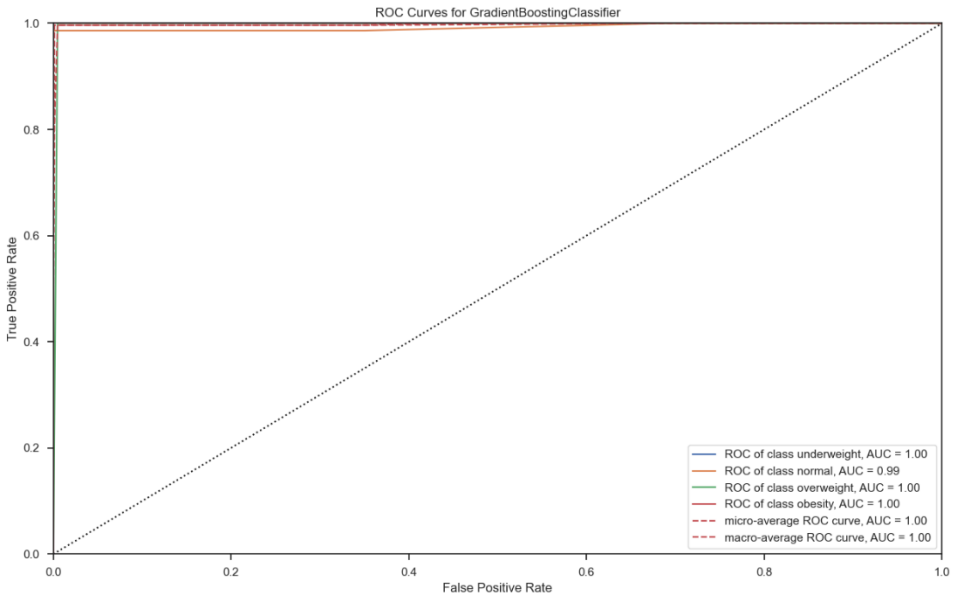


Figure 4.25 The AUC-ROC of GB with balance data

### *4.1.7 XGBoost*

Phase 1: XGB without dealing with imbalance data and using grid search technique to determine the best set of parameters, Figure 4.26 represents the performance metrics used to evaluate the effectiveness of applying supervised machine learning algorithm XGB, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.27 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.26, the results we got was the accuracy of 100%, precision for all class (100%, 100%, 100%, 100%) respectively, recall for all classes (100%, 100%, 100%, 100%) respectively, and f-1 score (100%, 100%, 100%, 100%) respectively.

```
The accuracy is
 1.0
------------------
The Confusion marix
 [[71  0  0  0]
 [ 0 32  0  0]
 [ 0  0 72  0]
 [ 0  0  0  6]]
------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        71
           1       1.00      1.00      1.00        32
           2       1.00      1.00      1.00        72
           3       1.00      1.00      1.00         6

    accuracy                           1.00       181
   macro avg       1.00      1.00      1.00       181
weighted avg       1.00      1.00      1.00       181
```

Figure 4.26 The results of XGB with imbalanced data

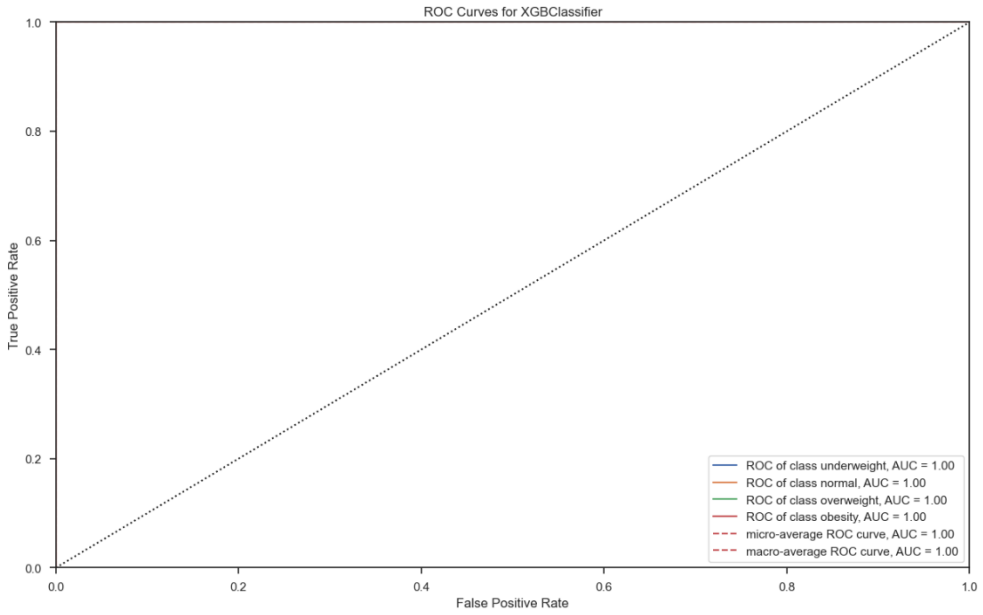


Figure 4.27 The AUC-ROC of XGB with imbalanced data

Phase 2: XGB deals with imbalance data by using the SMOTE technique and using grid search technique to determine the best set of parameters, Figure 4.28 represents the performance metrics used to evaluate the effectiveness of applying the supervised machine learning algorithm XGB, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.29 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.28, the results we got was the accuracy of 99.6%, precision for all class (100%, 100%, 99%, 100%) respectively, recall for all classes (100%, 99%, 100%, 100%) respectively, and f-1 score (100%, 99%, 99%, 100%) respectively.

```
The accuracy is
 0.9965034965034965
------------------
The Confusion marix
 [[66  0  0  0]
 [ 0 71  1  0]
 [ 0  0 73  0]
 [ 0  0  0 75]]
------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        66
           1       1.00      0.99      0.99        72
           2       0.99      1.00      0.99        73
           3       1.00      1.00      1.00        75

    accuracy                           1.00       286
   macro avg       1.00      1.00      1.00       286
weighted avg       1.00      1.00      1.00       286
```

Figure 4.28 The results of XGB with balanced data

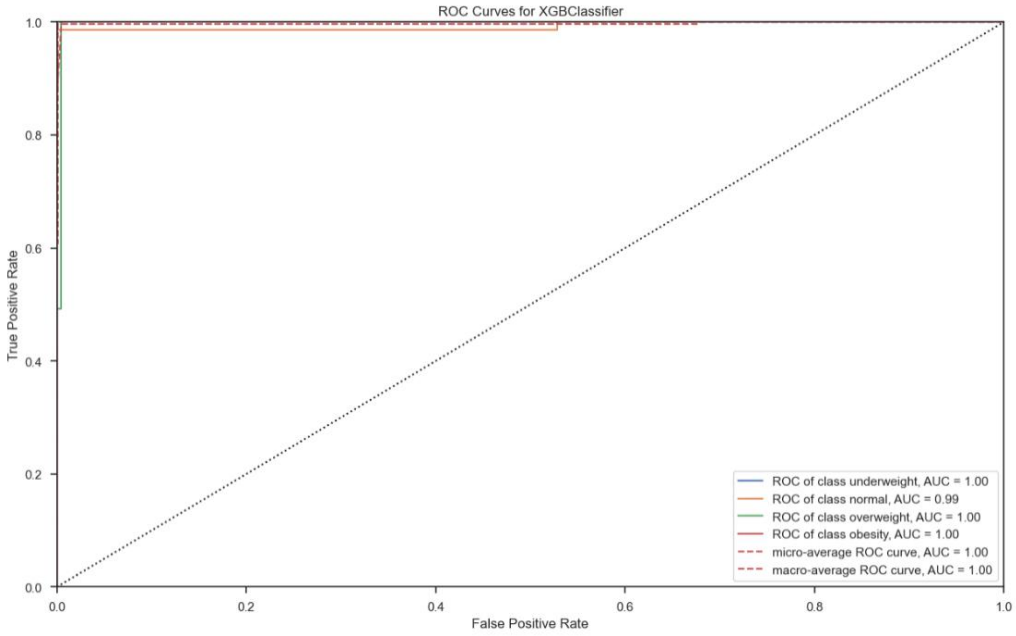


Figure 4.29 The AUC-ROC of XGB with balance data

### *4.1.8 Multilayer Perceptron Neural Networks*

Phase 1: MLPNNs without dealing with imbalance data and using grid search technique to determine the best set of parameters, Figure 4.30 represents the performance metrics used to evaluate the effectiveness of applying supervised machine learning algorithm MLPNNs, which include: Accuracy, Precision, Recall, F1 Score. The figure 4.31 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.30, the results we got was the accuracy of 94.4%, precision for all class (88%, 100%, 100%, 100%) respectively, recall for all classes (100%, 100%, 89%, 67%) respectively, and f-1 score (93%, 100%, 94%, 80%) respectively.

```
The accuracy is
 0.9447513812154696
------------------
The Confusion marix
 [[71  0  0  0]
 [ 0 32  0  0]
 [ 8  0 64  0]
 [ 2  0  0  4]]
------------------
              precision    recall  f1-score   support

           0       0.88      1.00      0.93        71
           1       1.00      1.00      1.00        32
           2       1.00      0.89      0.94        72
           3       1.00      0.67      0.80         6

    accuracy                           0.94       181
   macro avg       0.97      0.89      0.92       181
weighted avg       0.95      0.94      0.94       181
```

Figure 4.30 The results of MLPNNs with imbalanced data

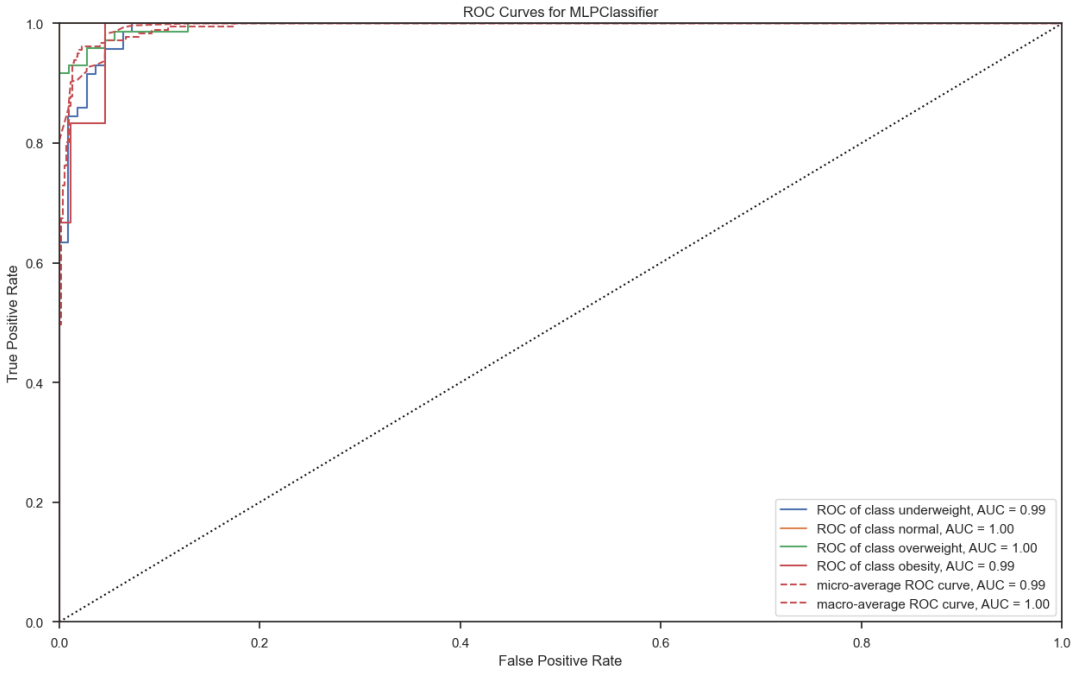


Figure 4.31 The AUC-ROC of MLPNNs with imbalanced data

Phase 2: MLPNNs deal with imbalance data by using SMOTE technique and using grid search technique to determine the best set of parameters, figure 4.32 represents the performance metrics used to evaluate the effectiveness of applying supervised machine learning algorithm MLPNNs, which include: Accuracy, Precision, Recall, F1 Score. Figure 4.33 represents the AUC-ROC curve. These metrics were utilized to conduct a comprehensive comparison of the model performances.

As we can see in Figure 4.32, the results we got was the accuracy of 95.1%, precision for all class (91%, 97%, 94%, 97%) respectively, recall for all classes (94%, 97%, 89%, 100%) respectively, and f-1 score (93%, 97%, 92%, 99%) respectively.

```
The accuracy is
 0.951048951048951
------------------
The Confusion marix
 [[62  0  2  2]
 [ 0 70  2  0]
 [ 6  2 65  0]
 [ 0  0  0 75]]
------------------
              precision    recall  f1-score   support

           0       0.91      0.94      0.93        66
           1       0.97      0.97      0.97        72
           2       0.94      0.89      0.92        73
           3       0.97      1.00      0.99        75

    accuracy                           0.95       286
   macro avg       0.95      0.95      0.95       286
weighted avg       0.95      0.95      0.95       286
```

Figure 4.32 The results of MLPNNs with balance data

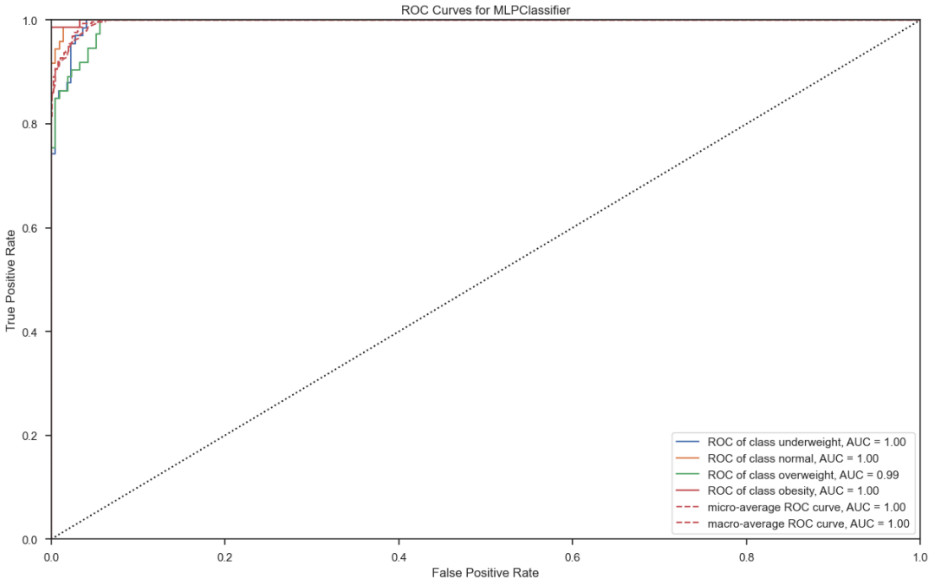


Figure 4.33 The AUC-ROC of MLPNNs with balance data

## *4.2 Discussion of Results and Comparison*

It's important to comprehensively investigate the reasons behind any unexpected behavior for any algorithm and should consider using appropriate evaluation metrics and techniques designed for imbalanced and balanced datasets (Stefanowski, 2015)(Jeni et al., 2013a). Ultimately, the choice of model will depend on the specific characteristics of your dataset and the problem you are trying to solve.

Figures (4.2, and 4.4) show the results of SVM in two phases (imbalance, and balance) and the results including Precision, Recall, and F-1 scores represent in Table 4.1 and Table 4.2.

Table 4.1 The result of SVM with imbalanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 96% | 100% | 100% | 100% |
| Recall | 100% | 100% | 97% | 83% |
| F-1 score | 98% | 100% | 99% | 91% |

Table 4.2 The result of SVM with balanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 92% | 97% | 94% | 99% |
| Recall | 98% | 94% | 90% | 99% |
| F-1 score | 95% | 96% | 92% | 99% |

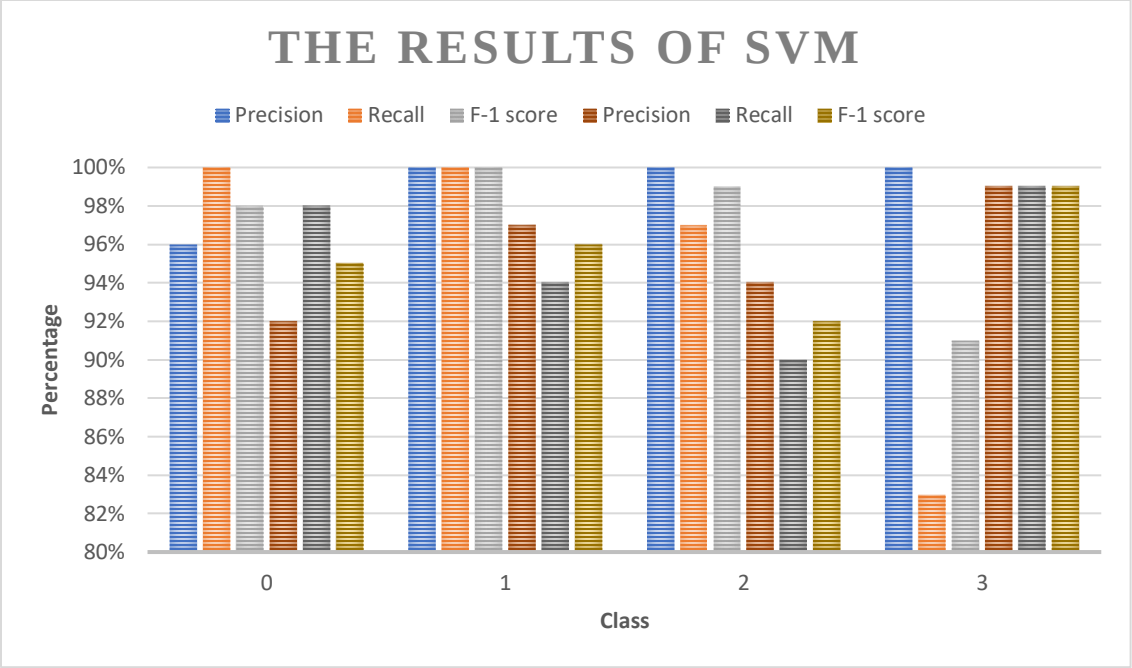


Figure 4.34 Visualize the result of SVM

Comparing both results as shown in figure 4.34, the SVM applied to imbalanced data consistently achieves higher precision, recall, and F-1 scores for most classes. This is perhaps due to the algorithm's emphasis on correctly classifying the minority class instances, which could lead to lower FN and elevated TP. However, it's important to interpret these results correctly, as the apparent performance gains on imbalanced data might not necessarily translate to better generalization on new data. Balancing the data attempts to more equitable distribution of the model's attention, probably leading to better overall performance in cases where a balanced consideration of all classes is crucial(Jeni et al., 2013b).

When applying DT with imbalanced data, it is generally expected that decision trees may exhibit different behaviors depending on the class distribution. Figures (4.6, and 4.8) show the results of DT in two phases (imbalance, and balance) and the results including Precision, Recall, and F-1 scores represent in Table 4.3 and Table 4.4.

Table 4.3 The result of DT with imbalanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 82% | 100% | 97% | 67% |
| Recall | 96% | 100% | 85% | 33% |
| F-1 score | 88% | 100% | 90% | 44% |

Table 4.4 The result of DT with balance data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 97% | 100% | 99% | 97% |
| Recall | 97% | 99% | 100% | 97% |
| F-1 score | 97% | 99% | 99% | 97% |

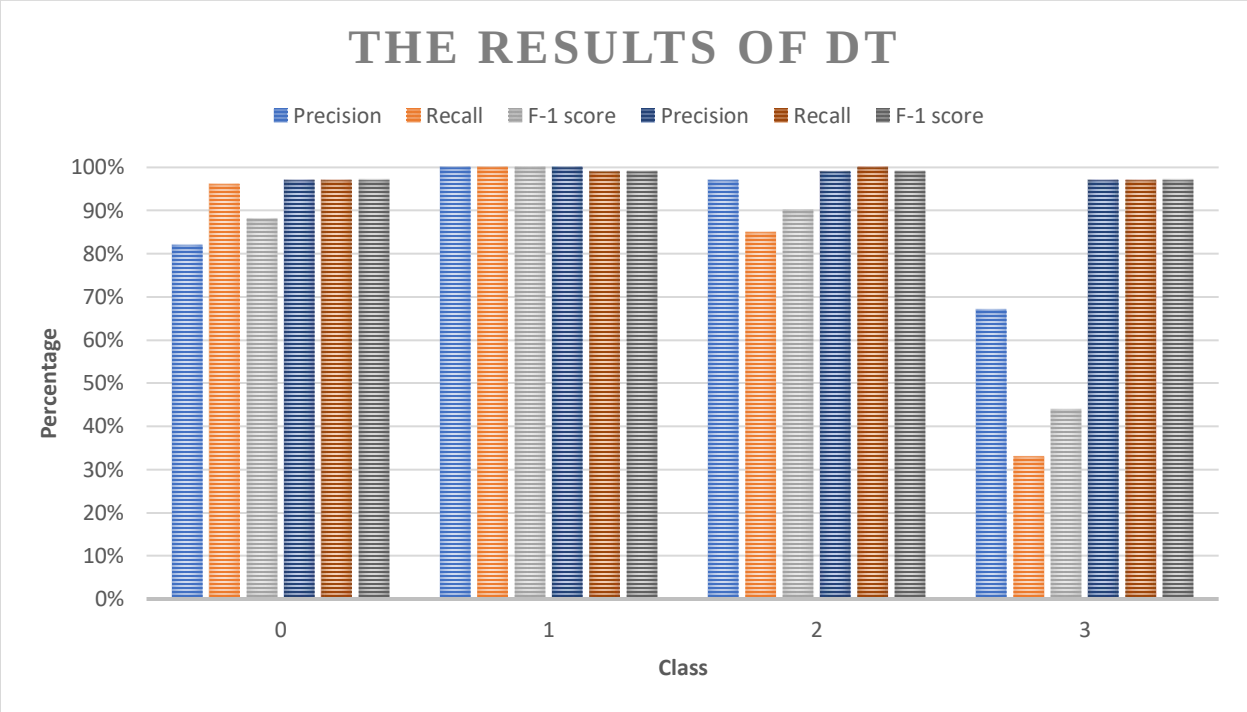


Figure 4.35 Visualize the result of DT

Comparing both results as shown in figure 4.35, the DT applied to balanced data consistently yields higher precision, recall, and F-1 scores across all classes. This is likely attributed to the balanced data distribution, enabling the model to equally consider all classes during training, leading to improved generalization and performance. In contrast, the imbalanced data presents challenges for the DT, particularly in capturing TP and achieving balanced precision-recall trade-offs, especially

evident in class 3. The balanced data ensures that the DT's attention is evenly distributed among classes, resulting in more accurate positive predictions and comprehensive class capture, ultimately leading to enhanced overall performance and better generalization abilities(Jeni et al., 2013b).

The performance of Random Forest can be influenced by the distribution of classes in the dataset, especially when dealing with imbalanced data. The figures (4.10, 4.12) show the results of RF in two phases (imbalance, and balance) and the results including Precision, Recall, and F-1 score represent in Table 4.5 and Table 4.6.

Table 4.5 The result of RF with imbalanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 96% | 100% | 100% | 100% |
| Recall | 100% | 100% | 100% | 50% |
| F-1 score | 98% | 100% | 100% | 67% |

Table 4.6 The result of RF with balanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 97% | 100% | 97% | 100% |
| Recall | 100% | 97% | 100% | 97% |
| F-1 score | 99% | 99% | 99% | 99% |


Figure 4.36 Visualize the result of the RF

Comparing both results as shown in figure 4.36, RF applied to balanced data consistently outperforms its imbalanced data in terms of precision, recall, and F-1 scores. The balanced data approach enables RF to train with a more even consideration of all classes, leading to improved generalization and performance. While the imbalanced data shows impressive precision values, the drop in recall and F-1 scores for class 3 highlights the challenge of capturing TP when class distribution is skewed. On the other hand, the balanced data allows RF to effectively balance precision and recall for all classes, resulting in enhanced overall performance and better generalization abilities(Jeni et al., 2013b).

Like other tree-based algorithms, tends to be influenced by the class distribution in the dataset. The figures (4.14, 4.16) show the results of ETC in two phases (imbalance, and balance) and the results including Precision, Recall, and F-1 score represent in Table 4.7 and Table 4.8.

Table 4.7 The result of ETC with imbalanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 77% | 100% | 87% | 0% |
| Recall | 92% | 91% | 82% | 0% |
| F-1 score | 84% | 95% | 84% | 0% |

Table 4.8 The result of ETC with balanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 74% | 74% | 79% | 86% |
| Recall | 76% | 83% | 56% | 97% |
| F-1 score | 75% | 78% | 66% | 91% |

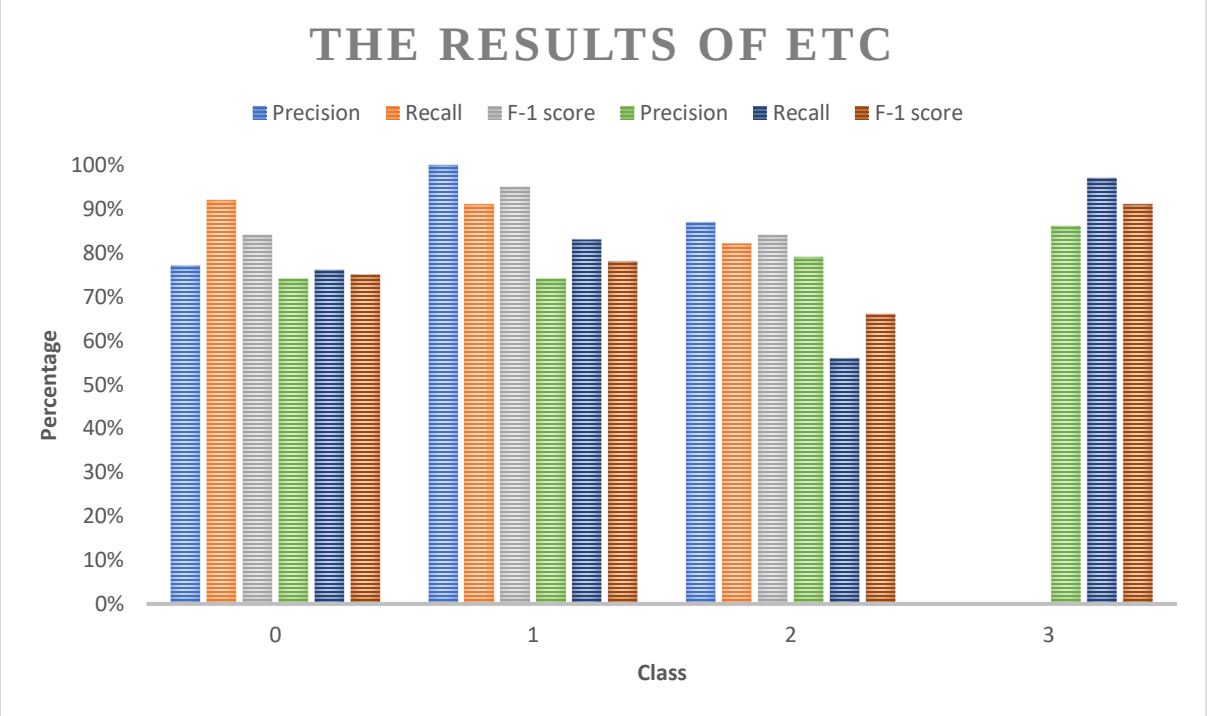


Figure 4.37 Visualize the result of ETC

Comparing both results as shown in figure 4.37, ETC applied to balanced data generally produces lower precision, recall, and F-1 scores compared to its imbalanced data. In the imbalanced data, ETC achieves higher precision and recall values for certain classes, especially class 1. However, the complete absence of both precision and recall values for class 3 highlights the model's inability to handle extremely imbalanced data. On the other hand, the balanced data yields more balanced precision and recall values across classes, despite the slightly lower scores. While ETC's performance on imbalanced data might seem better for some classes, the model's limitations become evident in scenarios with significant class imbalances, emphasizing the importance of balanced data for more robust generalization and improved overall performance.

Like other tree-based algorithms, tends to be influenced by the class distribution in the dataset. The figures (4.18, 4.20) show the results of AdaBoost in two phases (imbalance, and balance) and the results including Precision, Recall, and F-1 score represent in Table 4.9 and Table 4.10.

Table 4.9 The result of AdaBoost with imbalanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 92% | 100% | 100% | 0% |
| Recall | 100% | 100% | 100% | 0% |
| F-1 score | 96% | 100% | 100% | 0% |

Table 4.10 The result of AdaBoost with balanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 100% | 100% | 99% | 100% |
| Recall | 100% | 99% | 100% | 100% |
| F-1 score | 100% | 99% | 99% | 100% |

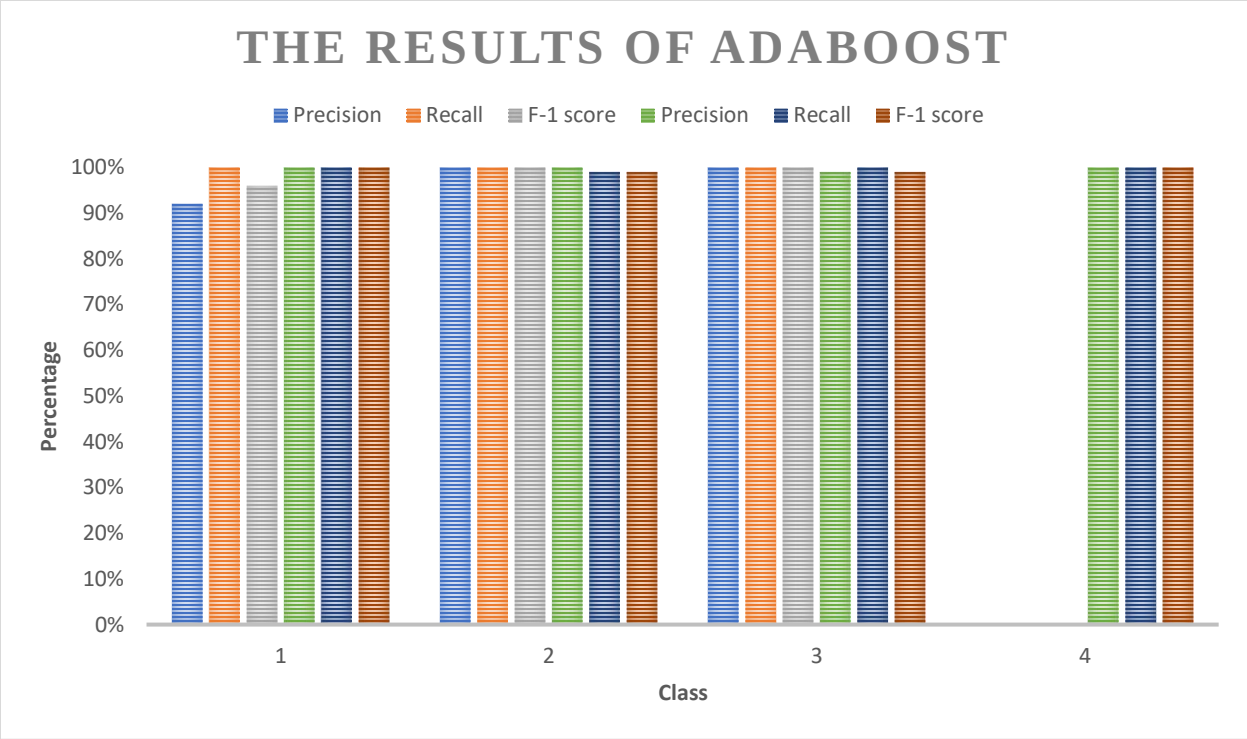


Figure 4.38 Visualize the result of AdaBoost

Comparing both results as shown in figure 4.38, AdaBoost applied to balanced data demonstrates admirable performance in terms of precision, recall, and F-1 scores across all classes compared to its imbalanced data. In the imbalanced data, AdaBoost achieves high precision and recall values for some classes but fails in capturing any instances for class 3. This highlights AdaBoost's limitations when dealing with extreme class imbalances. Conversely, the balanced data enables AdaBoost to obtain consistently high precision, recall, and F-1 scores across all classes, underlining the significance of balanced data for acceptable model robustness, generalization, and overall performance(Jeni et al., 2013b).

As tree-based algorithms. The figures (4.22, 4.24) show the results of GB in two phases (imbalance, and balance) and the results including Precision, Recall, and F-1 score represent in Table 4.11 and Table 4.12.

Table 4.11 The result of GB with imbalanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 100% | 100% | 99% | 100% |
| Recall | 100% | 100% | 100% | 100% |
| F-1 score | 100% | 100% | 100% | 100% |

Table 4.12 The result of GB with balance data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 100% | 100% | 99% | 100% |
| Recall | 100% | 99% | 100% | 100% |
| F-1 score | 100% | 99% | 99% | 100% |

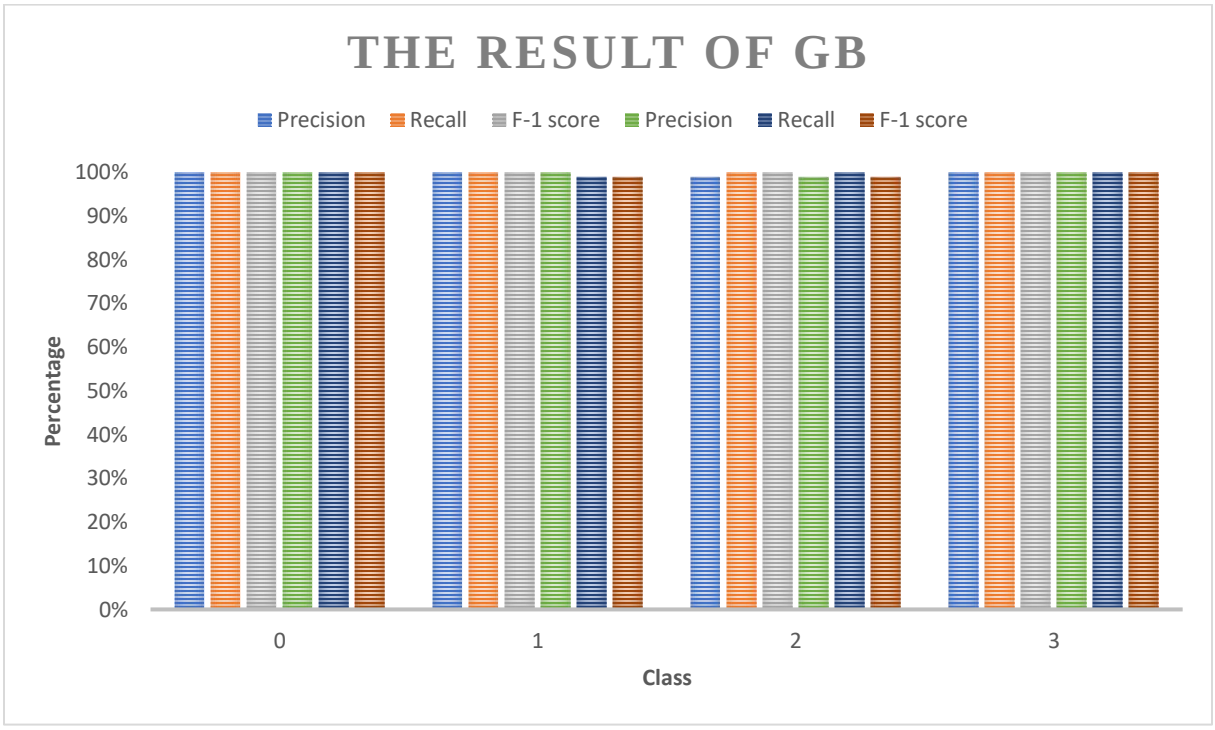


Figure 4.39 Visualize the result of GB

Comparing both results as shown in figure 4.39, GB applied to both imbalanced and balanced data consistently achieves exceptionally high precision, recall, and F-1 scores across all classes. This demonstrates GB's robustness and stability in handling both data cases. The balanced data showcases slightly reduced recall and F-1 scores for class 1 but still maintains excellent overall performance. particularly, in both cases, GB maintains perfect precision, suggesting a model that excels in accurate positive predictions. The similarity in performance between the two tables indicates that GB is highly effective in maintaining its performance even in the presence of class imbalances, making it a strong candidate for diverse classification tasks.

The figures (4.26, 4.28) show the results of XGB in two phases (imbalance, balance) and the results including Precision, Recall, and F-1 score represent in Table 4.13 and Table 4.14.

Table 4.13 The result of XGB with imbalanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 100% | 100% | 99% | 100% |
| Recall | 100% | 100% | 100% | 100% |
| F-1 score | 100% | 100% | 100% | 100% |

Table 4.14 The result of XGB with balance data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 100% | 100% | 99% | 100% |
| Recall | 100% | 99% | 100% | 100% |
| F-1 score | 100% | 99% | 99% | 100% |


Figure 4.40 Visualize the result of XGB

Comparing both results as shown in figure 4.40, XGB applied to both imbalanced and balanced data consistently demonstrates appropriate precision, recall, and F-1 scores across all classes. XGB's performance remains virtually unchanged between the two cases, highlighting its robustness and stability even in the presence of class imbalances. The balanced data showcases minimal variations in recall and F-1 scores for class 1 while maintaining appropriate overall performance. This consistency underscores XGB's powerful and reliable classifier, with the capability of delivering outstanding results regardless of the data distribution, making it a compelling choice for diverse classification tasks.

Multilayer Perceptron’s Neural Networks (MLPNNs), which are a type of artificial neural network, can exhibit different behavior when trained on imbalanced data compared to balanced

data. The figures (4.30, 4.32) show the results of MLPNNs in two phases (imbalance, balance) and the results including Precision, Recall, and F-1 score represent in Table 4.15 and Table 4.16.

Table 4.15 The result of MLPNNs with imbalanced data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 88% | 100% | 100% | 100% |
| Recall | 100% | 100% | 89% | 67% |
| F-1 score | 93% | 100% | 94% | 80% |

Table 4.16 The result of MLPNNs with balance data

| Class | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Precision | 91% | 97% | 94% | 97% |
| Recall | 94% | 97% | 89% | 100% |
| F-1 score | 93% | 97% | 92% | 99% |

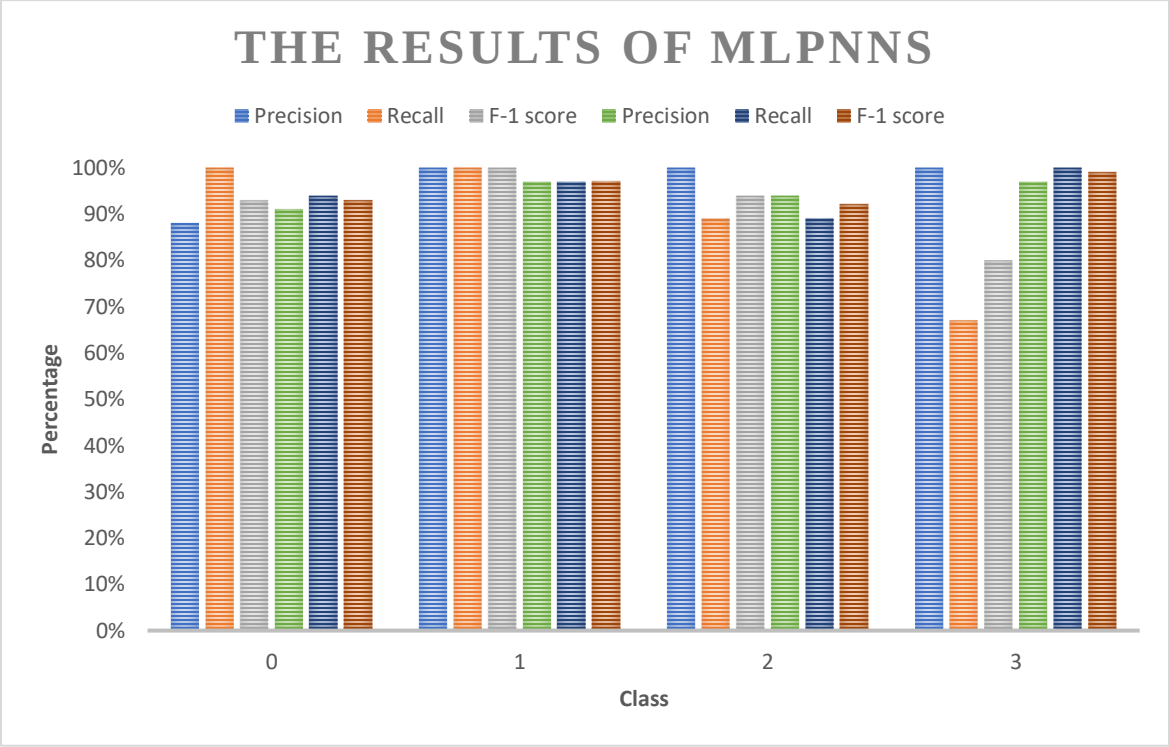


Figure 4.41 Visualize the result of MLPNNs

Comparing both results as shown in figure 4.41, MLPNNs applied to balanced data generally demonstrate better performance in terms of recall and F-1 scores across all classes compared to imbalanced data. While precision values are slightly lower in the balanced data, the model's balanced consideration of all classes during training leads to more consistent trade-offs between precision and recall. In other case, the imbalanced data earnings high precision values but exhibits variation in recall and F-1 scores, especially in class 3. The balanced data ensures a more equitable distribution of the model's attention, resulting in improved generalization and overall performance across all classes(Jeni et al., 2013b).

Table 4.17 represents the comparison of accuracy between all models before dealing with imbalance data and after dealing with balance data.

Table 4.17 The comparison of accuracy between all models

| Model | Accuracy with imbalance | Accuracy with balance |
|---|---|---|
| SVM | 98.3% | 95.4% |
| DT | 90.0% | 98.2% |
| RF | 98.3% | 98.6% |
| ETC | 84.0% | 87.0% |
| AdaBoost | 96.6% | 99.6% |
| GB | 100.0% | 99.6% |
| XGB | 100.0% | 99.6% |
| MLPNNs | 94.4% | 95.1% |

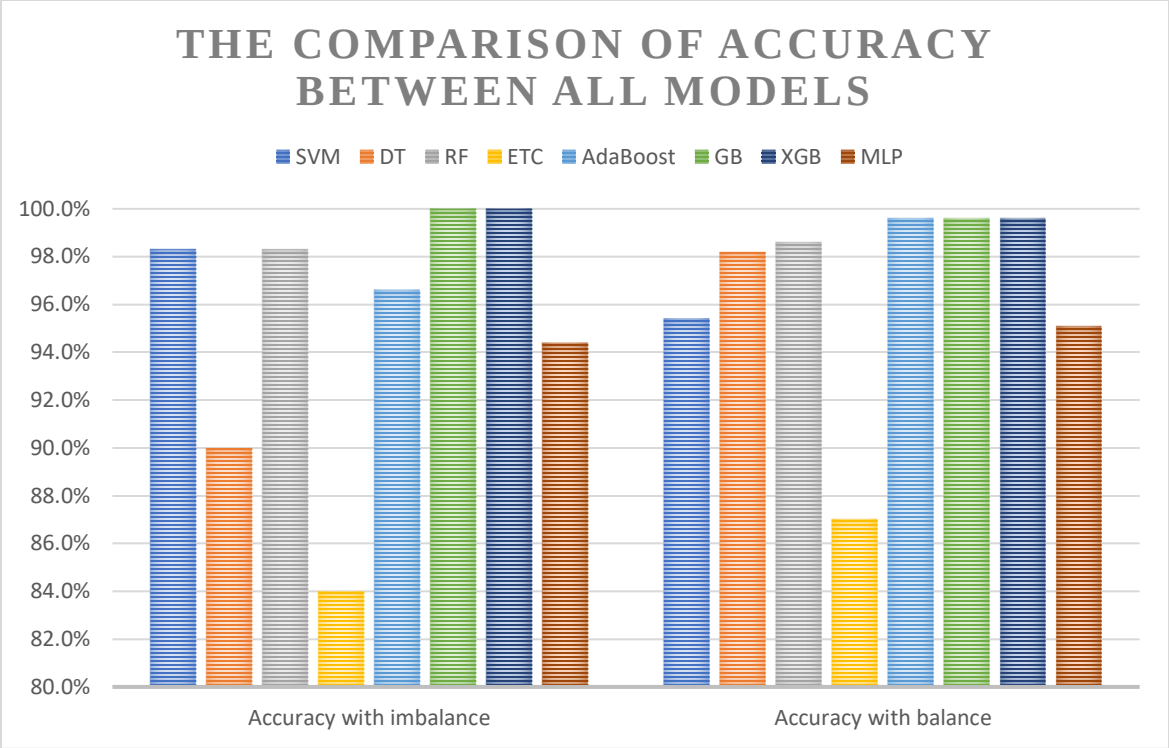


Figure 4.42 The comparison of accuracy between all models

The figure 4.42 shows the difference in accuracy between these two cases can be attributed to how the SVM operates and how it responds to class imbalance. In the imbalanced dataset case, the SVM tends to heavily favor the majority classes as it strives to minimize classification errors overall. Consequently, it achieves high accuracy on the majority classes but often struggles to maintain similar accuracy on the minority class. In contrast, in the balanced dataset case, the SVM is compelled to give equal consideration to all classes, making it more challenging to achieve equally high accuracy on both fronts, thus resulting in a slightly lower overall accuracy.

While, the difference in accuracy between these two cases can be attributed to key factors. In the imbalanced dataset case, the DT classifier demonstrated its ability to create splits that effectively addressed both the majority and minority classes, leading to relatively strong performance. However, in the balanced dataset scenario, the classifier encountered a less demanding task as there was no significant class imbalance to navigate. This allowed it to craft a tree structure that

was finely tuned to the characteristics of all classes, resulting in a significantly higher accuracy rate.

On other hand, the RF classifier achieved an impressive accuracy of 98.3% when dealing with an imbalanced dataset. This signifies its exceptional ability to accurately predict all the majority and minority classes, even when one class significantly outnumbers the others. The strength of RF lies in their robustness when handling imbalanced datasets, achieved through the construction of multiple decision trees and the amalgamation of their predictions, which effectively mitigates the impact of class imbalance. In the second case, the RF classifier surpassed its prior performance, attaining a remarkable accuracy of 98.6% on a balanced dataset, where instances for each class were roughly equal in number. When faced with a balanced dataset, RF can shine even brighter. Free from the need to contend with class imbalance issues, they can create a diverse ensemble of trees, collectively delivering highly accurate and reliable predictions, which led to a slightly higher overall accuracy. When ETC achieved an accuracy of 84% when dealing with an imbalanced dataset. This implies that when one class significantly outnumbers the others, the classifier demonstrated the ability to predict all of the majority and minority classes with moderate accuracy. ETC, like RF, are ensemble ML methods that construct multiple DT and combine their predictions. However, they introduce a higher degree of randomness in the tree-building process. In the second case, the ETC achieved a slightly higher accuracy of 87% on a balanced dataset, where there was an approximately equal number of instances for each class. When confronted with a balanced dataset, ETC still benefitted from the ensemble approach, but with the added advantage of working with a more evenly distributed training dataset. This often led to an improvement in accuracy compared to cases with class imbalance.

While, AdaBoost classifier achieved an accuracy of 96.6% when run on an imbalanced dataset. This indicates the classifier's remarkable ability to provide strong performance when number

classes significantly outnumber other classes. AdaBoost, popular for its ensemble sequential approach, combines multiple weak classifiers to form a strong classifier, and excels at dealing with class imbalance by ensuring accurate predictions for both majority and minority classes. In the second case, the AdaBoost classifier showed higher accuracy, reaching 99.6%, when working with a balanced dataset, where the instances of each class were approximately equal. In situations with balanced data sets, AdaBoost really shines. Freed from the complexities of class imbalance, he can focus on raising overall classification performance without fear of favoring any particular class.

GB classifier demonstrated exceptional performance by achieving a flawless accuracy of 100% when dealing with an imbalanced dataset. This remarkable result signifies that even in situations where one class significantly dominates the others, the GB model excels at accurately predicting both the majority and minority classes. GB, which constructs a potent classifier by progressively adding decision trees, stands out in such cases due to its ability to focus on challenging samples, including those from the minority class. In the second case, the GB classifier maintained impressive accuracy, reaching 99.6%, while operating on a balanced dataset with roughly equal instances for each class. In balanced datasets, Gradient Boosting still delivers remarkable results, but a minor reduction in accuracy compared to the imbalanced scenario may occur due to the absence of class imbalance challenges. The results XGB were also similar to the results of GB.

Finally, MLPNNs classifier achieved an accuracy of 94.4% when dealing with an imbalanced dataset. This signifies the model's robust performance even when some classes significantly outweigh the others. MLPNNs, as part of the broader family of neural networks, possess the ability to get deep insights of data relationships, making them valuable for handling imbalanced datasets. Nevertheless, it's important to note that their performance can still be influenced by the distribution of classes within the dataset. In the second case, the MLPNNs classifier attained a slightly higher accuracy of 95.1% when working with a balanced dataset, where each class is represented in

approximately equal proportions. When the dataset exhibits balance, MLPNNs continue to demonstrate proficiency, and the marginal uptick in accuracy compared to the imbalanced case suggests that balancing the dataset might have offered a subtle enhancement.

## *4.3 Challenges and Limitations*

The research faces some challenges and limitations considering the data collection process. The first challenge is attributed to the questionnaire distribution online, which may affect the participants' reliability. Also, there may be another personal factor related to the feeling of embarrassment and refusal of sharing personal information about height, etc. on the other hand, the number of participants in the data collection process was less than expected.

The imbalance in classes is one of the challenges in this research, moreover the lack of data regarding obesity and overweight in official sources, a lack of statistical data, and in the dataset of obesity for Palestinian society.

# Chapter 5. Conclusions

## *5.1 Conclusions*

ML has become an essential part of technology development in the healthcare field. Machine learning models can deeply understand a huge and complex amount of data. After analyzing the data, we can get new insights, as a result, new extracted knowledge can assist the physicians and

doctors to figure out the medical issues quickly with high accuracy, which can lead the physicians to process a larger quantity and overcome errors. Classification machine learning models have proven to be functional mechanisms in classifying and diagnosing Obesity and Overweight issues. Using machine learning models can predict the level of obesity for participants to determine how nutrition can help them with special treatment for every case. For that, in this research, we apply numerous classification machine learning algorithms and compare the model results. The results for all models were promising in both phases of imbalance and balance data. As shown in Table 11, the models deal well with imbalanced data and fine-tuning as in SVM, RF, AdaBoost and MLPNNS (98.3%, 98.3%,96.6%, 94.4%) respectively. While GB and XGB showing It can be considered as Overfitting (100%, 100%). On the other hand, DT, ETC showed appropriate results with (90.0%, and 84.0%) respectively. After using the SMOTE technique and fine-tuning some algorithms have shown an improvement in performance as DT, RF, ETC, AdaBoost, and MLPNNs (98.2%, 98.6%, 99.6%, 87%, 95.1%) respectively. While, some algorithms showed a decrease in performance as SVM, GB, and XGB (95.4%, 99.6%, 99.6%) respectively. Based on that, the best performance of the model in imbalanced data is SVM, RF, and AdaBoost. On the other hand, the best performance of the model in balance data is DT and RF.

## *5.2 Future Work and Recommendations*

While published research on Obesity Level classification covers many topics, studies rarely apply classification machine learning algorithms in the researcher's country, Palestine. This research represents the classification of Obesity levels based on risk factors to build an intelligence decision support system and Web based App depending on the RF model, which can be related to nutrition and diet models with the help of the professional's es in this field.  In order to improve the results,

in future research it is necessary to count in advance with a larger amount of data, sufficient to demonstrate and observe changes in the training of models.

On the other hand, after studying Obesity and Overweight based on BMI with risk factors to classify Obesity levels. We believe that the classification of Obesity levels should be studied based on a Body Fat Percentage (BFP) on a local dataset from Palestine. And try to compare results between BMI and BFP to obtain an accurate result to treat Obesity and Overweight Problems.

# References

*1.4. Support Vector Machines — scikit-learn 1.3.0 documentation*. (n.d.). Retrieved July 2, 2023, from https://scikit-learn.org/stable/modules/svm.html

*1.10. Decision Trees — scikit-learn 1.3.0 documentation*. (n.d.). Retrieved July 5, 2023, from https://scikit-learn.org/stable/modules/tree.html

Abdeen, Z., Jildeh, C., Dkeideek, S., Qasrawi, R., Ghannam, I., & Al Sabbah, H. (2012a). Overweight and obesity among Palestinian adults: Analyses of the anthropometric data from the first national health and nutrition survey (1999-2000). *Journal of Obesity*, *2012*. https://doi.org/10.1155/2012/213547

Abdeen, Z., Jildeh, C., Dkeideek, S., Qasrawi, R., Ghannam, I., & Al Sabbah, H. (2012b). Overweight and obesity among Palestinian adults: Analyses of the anthropometric data from the first national health and nutrition survey (1999-2000). *Journal of Obesity*, *2012*. https://doi.org/10.1155/2012/213547

Abdullah, F. S., Manan, N. S. A., Ahmad, A., Wafa, S. W., Shahril, M. R., Zulaily, N., Amin, R. M., & Ahmed, A. (2017). Data mining techniques for classification of childhood obesity among year 6 school children. *Advances in Intelligent Systems and Computing*, *549 AISC*, 465–474. https://doi.org/10.1007/978-3-319-51281-5_47

Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, *2019*(10). https://doi.org/10.7717/peerj.7702

Al-Hazzaa, H. M., Alrasheedi, A. A., Alsulaimani, R. A., Jabri, L., Alhowikan, A. M., Alhussain, M. H., Bawaked, R. A., & Alqahtani, S. A. (2022). Prevalence of overweight and obesity among saudi children: A comparison of two widely used international standards and the national growth references. *Frontiers in Endocrinology*, *13*. https://doi.org/10.3389/fendo.2022.954755

Alsareii, S. A., Shaf, A., Ali, T., Zafar, M., Alamri, A. M., AlAsmari, M. Y., Irfan, M., & Awais, M. (2022). IoT Framework for a Decision-Making System of Obesity and Overweight Extrapolation among Children, Youths, and Adults. *Life*, *12*(9). https://doi.org/10.3390/life12091414

Alzahrani, A., Alshehri, Y., Alageel, N., Bakraa, R., Rawas, R., & Alwagdani, R. (2023). The Relation between Mean Platelets Volume and Obesity among patients in the Primary Health-care Centers at National Guard in Jeddah, Saudi Arabia. *Open Access Macedonian Journal of Medical Sciences*, *11*(B), 309–313. https://doi.org/10.3889/oamjms.2023.11308

Amro, H., & Awad, M. (2021). Prediction of Body Fat Percentage Based on Anthropometric Measurements Using Data Mining Approach. In *Journal of the Arab American University مجلة الجامعة العربية الامريكية للبحوث* (Vol. 7, Issue 2). Prof. Mohammed.

Avgerinos, K. I., Spyrou, N., Mantzoros, C. S., & Dalamaga, M. (2019). Obesity and cancer risk: Emerging biological mechanisms and perspectives. In *Metabolism: Clinical and Experimental* (Vol. 92, pp. 121–135). W.B. Saunders. https://doi.org/10.1016/j.metabol.2018.11.001

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. In *JAMA - Journal of the American Medical Association* (Vol. 319, Issue 13, pp. 1317–1318). American Medical Association. https://doi.org/10.1001/jama.2017.18391

Bentham, J., Di Cesare, M., Bilano, V., Bixby, H., Zhou, B., Stevens, G. A., Riley, L. M., Taddei, C., Hajifathalian, K., Lu, Y., Savin, S., Cowan, M. J., Paciorek, C. J., Chirita-Emandi, A., Hayes, A. J., Katz, J., Kelishadi, R., Kengne, A. P., Khang, Y. H., … Cisneros, J. Z. (2017). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *The Lancet*, *390*(10113), 2627–2642. https://doi.org/10.1016/S0140-6736(17)32129-3

Blasco, B. V., García-Jiménez, J., Bodoano, I., & Gutiérrez-Rojas, L. (2020). Obesity and depression: Its prevalence and influence as a prognostic factor: A systematic review. In *Psychiatry Investigation* (Vol. 17, Issue 8, pp. 715–724). Korean Neuropsychiatric Association. https://doi.org/10.30773/pi.2020.0099

*Body Mass Index (BMI) | Healthy Weight, Nutrition, and Physical Activity | CDC*. (n.d.). Retrieved June 2, 2023, from https://www.cdc.gov/healthyweight/assessing/bmi/index.html

*Body Mass Index (BMI) Calculator - Diabetes Canada*. (n.d.). Retrieved June 23, 2023, from https://www.diabetes.ca/managing-my-diabetes/tools---resources/body-mass-index-(bmi)-calculator

Bruce, P., Bruce, A., Gedeck, P., & Safari, an O. M. Company. (n.d.). *Practical Statistics for Data Scientists, 2nd Edition*.

*Causes of Obesity | Overweight & Obesity | CDC*. (n.d.). Retrieved June 2, 2023, from https://www.cdc.gov/obesity/basics/causes.html

Cervantes, R. C., & Palacio, U. M. (2020). Estimation of obesity levels based on computational intelligence. *Informatics in Medicine Unlocked*, *21*. https://doi.org/10.1016/j.imu.2020.100472

Cheng, X., Lin, S. Y., Liu, J., Liu, S., Zhang, J., Nie, P., Fuemmeler, B. F., Wang, Y., & Xue, H. (2021). Does physical activity predict obesity—a machine learning and statistical method-based analysis. *International Journal of Environmental Research and Public Health*, *18*(8). https://doi.org/10.3390/ijerph18083966

*Create New Features From Existing Features - Train a Supervised Machine Learning Model - OpenClassrooms*. (n.d.). Retrieved June 20, 2023, from https://openclassrooms.com/en/courses/6389626-train-a-supervised-machine-learning-model/6398776-create-new-features-from-existing-features

Dağ, Z. Ö., & Dilbaz, B. (2015). Impact of obesity on infertility in women. In *Journal of the Turkish German Gynecology Association* (Vol. 16, Issue 2, pp. 111–117). AVES Ibrahim Kara. https://doi.org/10.5152/jtgga.2015.15232

Damiri, B., Abualsoud, M. S., Samara, A. M., & Salameh, S. K. (2018). Metabolic syndrome among overweight and obese adults in Palestinian refugee camps. *Diabetology and Metabolic Syndrome*, *10*(1). https://doi.org/10.1186/s13098-018-0337-2

DANACI, Ç., AVCI, D., & ARSLAN TUNCER, S. (2023). Komşuluk Bileşen Analizi Tabanlı Makine Öğrenimi Yöntemleri ile Obezite Seviyelerinin Tahmini. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*. https://doi.org/10.35234/fumbd.1234161

Dangeti, P. (n.d.). *Statistics for machine learning : build supervised, unsupervised, and reinforcement learning models using both Python and R.*

*Decision Tree - GeeksforGeeks*. (n.d.). Retrieved July 5, 2023, from https://www.geeksforgeeks.org/decision-tree/

DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., Fisher, C. B., Harness, D., Viswanath, S., Heymsfield, S. B., Dungan, I., & Thomas, D. M. (2018). A review of machine learning in obesity. In *Obesity Reviews* (Vol. 19, Issue 5, pp. 668–685). Blackwell Publishing Ltd. https://doi.org/10.1111/obr.12667

De-La-Hoz-Correa, E., Mendoza-Palechor, F. E., De-La-Hoz-Manotas, A., Morales-Ortega, R. C., & Adriana, S. H. B. (2019). Obesity level estimation software based on decision trees. *Journal of Computer Science*, *15*(1), 67–77. https://doi.org/10.3844/jcssp.2019.67.77

Dinda Qatrunnada, R. (n.d.). *Faktor Penyebab Kejadian Kelebihan Berat Badan dan Obesitas pada Anak-anak dan Dewasa Factors that Cause Overweight and Obesity in School-Age Children and Adult*.

*Don't Overfit! II — How to avoid Overfitting in your Machine Learning and Deep Learning Models | by Md. Mubasir | Towards Data Science*. (n.d.). Retrieved June 18, 2023, from https://towardsdatascience.com/dont-overfit-ii-how-to-avoid-overfitting-in-your-machine-learning-and-deep-learning-models-2ff903f4b36a

Duchen, K., Jones, M., Faresjö, Å. O., Faresjö, T., & Ludvigsson, J. (2020). Predicting the development of overweight and obesity in children between 2.5 and 8 years of age: The prospective ABIS study. *Obesity Science and Practice*, *6*(4), 401–408. https://doi.org/10.1002/osp4.418

El Bilbeisi, A. H., Shab-Bidar, S., Jackson, D., & Djafarian, K. (2017). The Prevalence of Metabolic Syndrome and Its Related Factors among Adults in Palestine: A Meta-Analysis. In *Ethiopian journal of health sciences* (Vol. 27, Issue 1, pp. 77–84). https://doi.org/10.4314/ejhs.v27i1.10

El Kishawi, R. R., Soo, K. L., Abed, Y. A., & Muda, W. A. M. W. (2014). Obesity and overweight: Prevalence and associated socio demographic factors among mothers in three different areas in the Gaza Strip-Palestine: A cross-sectional study. *BMC Obesity*, *1*(1). https://doi.org/10.1186/2052-9538-1-7

Ellulu, M. S., Abed, Y., Al-Kahlout, M., Ellulu, M., & Rahmat, A. (2014). Updates of Overweight and Obesity Status and Their Consequences in Palestine Overweight and Obesit y among Palest inian Adult s: Analyses of t he Ant hropomet ric Dat a from t he Fir… Ziad Abdeen Nut rit ional Risk Fact ors among St roke Survivors in Gaza Governorat es: A case-cont rol st udy Updates of Overweight and Obesity Status and Their Consequences in Palestine. *Pakistan Journal of Nutrition*, *13*(2), 116–121.

Farnham, B., Tokyo, S., Boston, B., Sebastopol, F., & Beijing, T. (n.d.). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION*.

Ferenci, T., & Kovács, L. (2018). Predicting body fat percentage from anthropometric and laboratory measurements using artificial neural networks. *Applied Soft Computing Journal*, *67*, 834–839. https://doi.org/10.1016/j.asoc.2017.05.063

*Free BMI with Age Calculator - Health by Science*. (n.d.). Retrieved September 11, 2023, from https://healthbyscience.co.uk/age-well/is-there-a-bmi-with-age-calculator-for-adults/

Gannon, J., Pollock, A. J., Allen, D. B., & Kling, P. J. (2021). A Practical Screening Tool to Predict Early Childhood Obesity Risk: Examining a Birth Cohort. *Clinical Pediatrics*, *60*(3), 178–183. https://doi.org/10.1177/0009922820971006

Gerl, M. J., Klose, C., Surma, M. A., Fernandez, C., Melander, O., Männistö, S., Borodulin, K., Havulinna, A. S., Salomaa, V., Ikonen, E., Cannistraci, C. V., & Simons, K. (2019). Machine learning of human plasma lipidomes for obesity estimation in a large population cohort. *PLoS Biology*, *17*(10). https://doi.org/10.1371/journal.pbio.3000443

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghrayeb, F., Ghrayeb, F. A., Rusli, M., Al Rifai, A., & Ismail, M. (n.d.). *Prevalence of Overweight and Obesity Among Adolescents in Tarqumia, Palestine*. https://www.researchgate.net/publication/275624991

Hadi Albayati, A., Esmaiel Mahmood, S., Thannon, A. H., & Qasim Hasan, S. (n.d.). *Computer Scope*. https://www.researchgate.net/publication/362862467

*How Random Forests & Decision Trees Decide: Simply Explained With An Example In Python | by Serafeim Loukas, PhD | Towards Data Science*. (n.d.). Retrieved July 8, 2023, from https://towardsdatascience.com/how-do-random-forests-decision-trees-decide-simply-explained-with-an-example-in-python-6737eb183604

Ibrahim, M., Ibrahim, M., Alzaher, A., Alshumemri, W., Alfaleh, A., Alabdulkareem, K., & Alabdulkareem, K. (2022). Lifestyle and its association with obesity among adolescents in Kingdom of Saudi Arabia. *The Egyptian Journal of Community Medicine*, *0*(0), 0–0. https://doi.org/10.21608/ejcm.2022.124700.1209

C. A. C. Montañez et al., "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2743-2750, doi: 10.1109/IJCNN.2017.7966194.

IEEE Computational Intelligence Society, International Neural Network Society, & Institute of Electrical and Electronics Engineers. (n.d.-b). *IJCNN 2017 : the International Joint Conference on Neural Networks.*

Jassar, S., Adams, S. J., Zarzeczny, A., & Burbridge, B. E. (2022). The future of artificial intelligence in medicine: Medical-legal considerations for health leaders. *Healthcare Management Forum*, *35*(3), 185–189. https://doi.org/10.1177/08404704221082069

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013a). Facing imbalanced data - Recommendations for the use of performance metrics. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 245–251. https://doi.org/10.1109/ACII.2013.47

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013b). Facing imbalanced data - Recommendations for the use of performance metrics. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 245–251. https://doi.org/10.1109/ACII.2013.47

Jeon, S., Kim, M., Yoon, J., Lee, S., & Youm, S. (2023). Machine learning-based obesity classification considering 3D body scanner measurements. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-30434-0

Jindal, K., Baliyan, N., & Rana, P. S. (2018). Obesity prediction using ensemble machine learning approaches. *Advances in Intelligent Systems and Computing*, *708*, 355–362. https://doi.org/10.1007/978-981-10-8636-6_37

Kitis, S., & Goker, H. (2023). Detection of Obesity Stages Using Machine Learning Algorithms. *Anbar Journal of Engineering Sciences*, *14*(1), 80–88. https://doi.org/10.37649/aengs.2023.139350.1045

Kunapuli, G. (n.d.). *Ensemble Methods for Machine Learning*.

Lin, C. J., Chang, Y. C., Cheng, T. Y., Lo, K., Liu, S. J., & Yeh, T. L. (2020). The association between metabolically healthy obesity and risk of cancer: A systematic review and meta-analysis of prospective cohort studies. In *Obesity Reviews* (Vol. 21, Issue 10). Blackwell Publishing Ltd. https://doi.org/10.1111/obr.13049

Liu, W., Fang, X., Zhou, Y., Dou, L., & Dou, T. (2022). Machine learning-based investigation of the relationship between gut microbiome and obesity status. *Microbes and Infection*, *24*(2), 104892. https://doi.org/10.1016/J.MICINF.2021.104892

Lopez-Jimenez, F., Almahmeed, W., Bays, H., Cuevas, A., Di Angelantonio, E., le Roux, C. W., Sattar, N., Sun, M. C., Wittert, G., Pinto, F. J., & Wilding, J. P. H. (2022). Obesity and cardiovascular disease: mechanistic insights and management strategies. A joint position paper by the World Heart Federation and World Obesity Federation. *European Journal of Preventive Cardiology*, *29*(17), 2218–2237. https://doi.org/10.1093/eurjpc/zwac187

*Machine Learning Algorithms and The Art of Hyperparameter Selection | by Rosaria Silipo | Towards Data Science*. (n.d.). Retrieved June 16, 2023, from https://towardsdatascience.com/machine-learning-algorithms-and-the-art-of-hyperparameter-selection-279d3b04c281

Maciej Serda, Becker, F. G., Cleary, M., Team, R. M., Holtermann, H., The, D., Agenda, N., Science, P., Sk, S. K., Hinnebusch, R., Hinnebusch A, R., Rabinovich, I., Olmert, Y., Uld, D. Q. G. L. Q., Ri, W. K. H. U., Lq, V.,

Frxqwu, W. K. H., Zklfk, E., Edvhg, L. V, … فاطمی, ح. (2013). Synteza i aktywność biologiczna nowych analogów tiosemikarbazonowych chelatorów żelaza. *Uniwersytet Śląski*, *7*(1), 343–354. https://doi.org/10.2/JQUERY.MIN.JS

Massad, S., Deckelbaum, R. J., Gebre-Medhin, M., Holleran, S., Dary, O., Obeidi, M., Bordelois, P., & Khammash, U. (2016). Double Burden of Undernutrition and Obesity in Palestinian Schoolchildren. *Food and Nutrition Bulletin*, *37*(2), 144–152. https://doi.org/10.1177/0379572116637720

Molina, D., De-La-Hoz, A., & Mendoza, F. (2021). CLASSIFICATION AND FEATURES SELECTION METHOD FOR OBESITY LEVEL PREDICTION. *Journal of Theoretical and Applied Information Technology*, *15*(11). www.jatit.org

Nasaif, H. (2022). *Prevalence of Overweight and Obesity among Nurses in Bahrain: A Cross-Sectional Study*. https://doi.org/10.21203/rs.3.rs-1957139/v1

*Obesity*. (n.d.). Retrieved November 19, 2022, from https://www.who.int/health-topics/obesity#tab=tab_1

*Obesity - NHS*. (n.d.). Retrieved June 2, 2023, from https://www.nhs.uk/conditions/obesity/

*Obesity - Symptoms and causes - Mayo Clinic*. (n.d.-a). Retrieved June 2, 2023, from https://www.mayoclinic.org/diseases-conditions/obesity/symptoms-causes/syc-20375742

*Obesity - Symptoms and causes - Mayo Clinic*. (n.d.-b). Retrieved June 2, 2023, from https://www.mayoclinic.org/diseases-conditions/obesity/symptoms-causes/syc-20375742

*Obesity and overweight*. (n.d.). Retrieved February 12, 2023, from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Piché, M. E., Tchernof, A., & Després, J. P. (2020). Obesity Phenotypes, Diabetes, and Cardiovascular Diseases. *Circulation Research*, 1477–1500. https://doi.org/10.1161/CIRCRESAHA.120.316101

Powell-Wiley, T. M., Poirier, P., Burke, L. E., Després, J. P., Gordon-Larsen, P., Lavie, C. J., Lear, S. A., Ndumele, C. E., Neeland, I. J., Sanders, P., & St-Onge, M. P. (2021). Obesity and Cardiovascular Disease A Scientific Statement From the American Heart Association. In *Circulation* (Vol. 143, Issue 21, pp. E984–E1010). Lippincott Williams and Wilkins. https://doi.org/10.1161/CIR.0000000000000973

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*(3), 203–231. https://doi.org/10.1023/a:1007601015854

*Python for Data Analysis, 3E - 7  Data Cleaning and Preparation*. (n.d.). Retrieved June 18, 2023, from https://wesmckinney.com/book/data-cleaning.html

Rashmi, R., Umapathy, S., & Krishnan, P. T. (2021). Thermal imaging method to evaluate childhood obesity based on machine learning techniques. *International Journal of Imaging Systems and Technology*, *31*(3), 1752–1768. https://doi.org/10.1002/ima.22572

*SMOTE for Imbalanced Classification with Python - MachineLearningMastery.com*. (n.d.). Retrieved June 14, 2023, from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Stefanowski, J. (2015). Dealing with data difficulty factors while learning from imbalanced data. In *Challenges in Computational Statistics and Data Mining* (Vol. 605, pp. 333–363). Springer International Publishing. https://doi.org/10.1007/978-3-319-18781-5_17

*Support Vector Machine(SVM): A Complete guide for beginners*. (n.d.). Retrieved July 4, 2023, from https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Thamrin, S. A., Arsyad, D. S., Kuswanto, H., Lawi, A., & Nasir, S. (2021). Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018. *Frontiers in Nutrition*, *8*. https://doi.org/10.3389/fnut.2021.669155

*Top 10 Applications of Machine Learning in Healthcare - FWS*. (n.d.). Retrieved June 2, 2023, from https://www.flatworldsolutions.com/healthcare/articles/top-10-applications-of-machine-learning-in-healthcare.php

*Top 10 Interview Questions on Gradient Boosting Algorithms -*. (n.d.). Retrieved July 10, 2023, from https://www.analyticsvidhya.com/blog/2022/11/top-10-interview-questions-on-gradient-boosting/

Verma, S., & Hussain, M. E. (2017). Obesity and diabetes: An update. In *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* (Vol. 11, Issue 1, pp. 73–79). Elsevier Ltd. https://doi.org/10.1016/j.dsx.2016.06.017

*What Is Generalization In Machine Learning? - Magnimind Academy*. (n.d.). Retrieved June 14, 2023, from https://magnimindacademy.com/blog/what-is-generalization-in-machine-learning/

Xue, Q., Wang, X., Meehan, S., Kuang, J., Gao, J. A., & Chuah, M. C. (2019). Recurrent Neural Networks Based Obesity Status Prediction Using Activity Data. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 865–870. https://doi.org/10.1109/ICMLA.2018.00139

Zeng, Q., Li, D., He, Y., Li, Y., Yang, Z., Zhao, X., Liu, Y., Wang, Y., Sun, J., Feng, X., Wang, F., Chen, J., Zheng, Y., Yang, Y., Sun, X., Xu, X., Wang, D., Kenney, T., Jiang, Y., … Dai, W. (2019). Discrepant gut microbiota markers for the classification of obesity-related metabolic abnormalities. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-49462-w

# Appendix 1

**The questionnaire:**

*Researcher Ahmed Hakam Abd Al-hafiz Radwan is conducting a scientific study to " collect a dataset to estimate levels of obesity based on eating habits and physical condition of individuals in Palestine", to obtain a master's degree in data science and business analysis. This questionnaire was prepared as a study tool for data collection, which includes three parts: general information, eating habits, and physical condition*

*Your participation in this study is completely voluntary. Your answers will be strictly confidential, and the data will be treated in general terms only and will be used for scientific research purposes only.*

**General Information:**

**This axis includes a set of questions related to the personal aspect of the respondent.**

- What is your gender? (Male, Female).
- What is your age? (19-61).
- What is your height? number (CM).
- What is your weight? number (KG).
- Has a family member suffered or suffered from being overweight? (Yes, No).

**Attributes related to eating habits**

**This axis includes several questions related to personal habits related to food.**

- Have you often eaten high-caloric food? (Rice, Bread, Fast Food)? (Yes, No).
- Do you usually eat vegetables in your meals? (Yes, No).
- How many main meals do you have daily? (1 - 2, 3, more than 3).
- Do you eat any food between meals? (No, sometimes, Usually, always).
- how often do you drink soda? (Soda or energy drink) (I do not drink, sometimes, Mostly, always).
- How much water do you drink daily? (Less than a liter, Between 1 and 2 L, more than 2 L).

**Attributes related to the physical condition**

**This axis includes a set of questions related to the personal aspect of the respondent**

- Do you smoke? (Yes, No).
- Do you monitor the calories you eat daily? (Yes, No).
- How often do you have physical activity? (Never, 1 - 2 days, 3 - 4 days, 5 - 6 days).
- How much time do you use technological devices such as cell phones, video games, television, computer, and others? (0-2 hours, 3-5 hours, more than 5 hours).
- Which transportation do you usually use? (Car, Motorbike, Bike, Public Transportation, Walking).

## الملخص

تعتبر السمنة وزيادة الوزن من الأسباب الرئيسية للعديد من الأمراض على المستوى العالمي. وفقًا لمنظمة الصحة العالمية (WHO)، يعاني أكثر من 30٪ من سكان العالم من السمنة وزيادة الوزن. هذه النسبة قد تزيد في السنوات القادمة. هذه الظاهرة العالمية ستؤدي إلى خطر متسلسل بالنظر إلى ظهور أمراض أخرى. غالبًا ما ترتبط السمنة بأمراض مزمنة أخرى، مثل ارتفاع ضغط الدم الشرياني وداء السكري من النوع 2. يزداد خطر الإصابة بهذه الأمراض المصاحبة مع زيادة مؤشر كتلة الجسم. ومع ذلك، هناك دراسات غير متكررة حول السمنة في فلسطين. يؤثر هذا النقص في المعرفة على تحديد التنبؤ بالسمنة والوقاية منها في المجتمع الفلسطيني.

نتيجة لذلك، يهدف هذا البحث إلى سد هذه الفجوة في أبحاث السمنة في فلسطين. سيتم تحقيق هذا الهدف من خلال تصنيف السمنة، والذي يشمل عوامل الخطر الحقيقية المتعلقة بالمجتمع الفلسطيني. تم جمع بيانات البحث عن طريق استبيان عبر الإنترنت وشملت

902 مشاركًا. ومع ذلك، بعد تصنيف المشاركين إلى أربع فئات، نقص الوزن، والوزن الطبيعي، وزيادة الوزن والسمنة، كان الصف الأول غير متوازن وهو ما تطلب استخدام طرق موازنة البيانات .SMOTE تمت معالجة البيانات التي تم جمعها بواسطة نماذج .ML تشمل آلة متجه الدعم (SVM) و Random Forest (RF) وشجرة القرار (DT) و Perceptron متعدد الطبقات (MLPNNs) و XGBoostو Adaboost ومصنف الشجرة الإضافية و(GB) Gradient Boost. تم تطبيق البحث الشبكي على النماذج للحصول على مجموعة المعلمات المناسبة لكل نموذج. تم تقييم أداء النماذج على بيانات الخلل والتوازن، والنموذج الذي تفوق على النماذج الأخرى بأعلى دقة هو RF بنسبة (98.3٪، 98.6٪) من الدقة على التوالي.