Federal TVET Institute

Division Electrical Electronics and Information Technology

Department of Information and Communication Technology

Machine Learning-Based Prediction Model of Obesity Risk Among adolescents in Central Ethiopia Regional State

A Thesis Submitted to School of Graduate Studies of Federal TVET Institute in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information and Communication Technology

By : Abebash Beyene

Advisor : Kibrom Tadesse (Phd)

June 2026

Addis Ababa, Ethiopia

FEDERAL TVET INSTITUTE

SCHOOL OF GRADUATE STUDIES

DIVISION OF ELECTRICAL ELECTRONICS AND INFORMATION COMMUNICATION TECHNOLOGY

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

Machine Learning-Based Prediction Model of Obesity Risk Among adolescents in Central Ethiopia Regional State

BY: Abebash Beyene

Name and Signatures of the Examining Board

Kibrom Tadesse (PhD)	_____	_____
Advisor	**Signature**	**Date**
_____	_____	_____
Internal Examiner	**Signature**	**Date**
_____	_____	_____
External Examiner	**Signature**	**Date**

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and all sources of materials used in the thesis have been duly acknowledged.

_____

Abebash Beyene

The thesis has been submitted for examination with my approval as a university advisor.

_____

Kibrom Tadesse (PhD)

Acknowledgment

First and foremost, I am thankful to Almighty God the source of knowledge and wisdom, who gives me health, power, and courage to accomplish this thesis. Next, I would like to give my deepest gratitude to my advisor Kibrom Tadesse (PhD) for his patience, and enthusiasm, for his invaluable advice and constructive comment throughout my thesis writing. I have learned a lot from him during this study and his guidance will be always appreciated throughout my career. Besides my advisor, Special thanks should go to Central Ethiopia Health Bureau for their endless encouragement and support to finish my study.

ABSTRACT

This study discovers the application of machine learning techniques to predict obesity risk among adolescents aged 10 - 19 years in Central Ethiopia Regional State. It uses a comprehensive dataset composed from above 10,000 adolescents. The methodology involves training multiple supervised machines learning classifiers such as Decision Trees, Support Vector Machines, Random Forest, and XGBoost using balanced and processed datasets. Metrics like as accuracy, precision, recall, and F1-score were used to assess the model's performance using 10-fold cross-validation. The findings demonstrate that Random Forest and XGBoost outperform other models with good predictive accuracy and stability, demonstrating their usefulness for classifying obesity in adolescents. The findings underscore the significance of combining behavioral, biological, and environmental data for efficient risk stratification and provide important insights into the predictive variables of teenage obesity in Ethiopia. Overall, this research emphasizes the potential of machine learning in advancing public health strategies to struggle adolescent obesity in places with limited resources and advocate for further research and collect more dataset from different regions to confirm the findings.

Key words: Adolescents, Data Science, Health Intervention, Machine Learning , Obesity, Risk Factors

Table of Contents

List of Acronyms and Abbreviations

API	Application Programming Interface
BMI	Body Mass Index
DT	Decision Tree
ETH	Ethiopia
GB	Gradient Boosting
GNB	Gaussian Naive Bayes
ICT	Information and Communication Technology
ML	Machine Learning
MLP	Multilayer Perceptron
RF	Random Forest
SD	Standard Deviation
SVM	Support Vector Machines
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

List of Tables

List of Figures

CHAPTER ONE

1. Introduction

1.1. Background of the study

Obesity is a health problem everywhere in the world. In places like Central Ethiopia more and more people are getting obese because they are moving to cities eating kinds of food and not being as active (WHO, 2021). Many adolescents in Ethiopia are getting obese like in other places. The things we are doing to stop this problem are not working well (Abebe et al. 2019).

People in Ethiopia are eating processed foods and not exercising much making the obesity problem worse (Gebremedhin, 2020). To solve this problem the usual ways are not working to keep people healthy, so we need to use methods like machine learning to understand obesity stop it from happening and help people who're already obese (Jiaf et al., 2022). Obesity among adolescents in Central Ethiopia needs to be stopped with these methods.

Machine learning is very useful for understanding health issues. It finds out what puts people at risk creating plans and making sure and use resources wisely (Esteva et al., 2017). When it comes to obesity machine learning algorithms can look at a lot of information like what people eat how much they exercise, their status and if they are likely to get certain diseases because of their genes. Machine learning can find out who is most expected to have problems and suggest ways to help them (Alam et al., 2021).

In central Ethiopia, where access to quality healthcare is limited, machine learning can assist identify individuals in need early on and improve the efficacy of our initiatives. In the long term, this may lessen health issues. Datasets can be analyzed via machine learning. Look for trends that we would miss otherwise. For instance, machine learning can assist us comprehend how many elements, such as exercise and food, impact people's health in Central Ethiopia (Mengistu et al., 2021).

Machine learning has the potential to benefit people. Some issues, such as insufficient data or underpowered machines, might cause obstacles. Some residents may not be extremely tech-savvy (Weldegebriel et al., 2021). Collaboration between local communities, health professionals, and technology is necessary to address these issues. Finding solutions that are long-lasting and used by a large number of individuals is also necessary (Assefa et al., 2022).

This study looked at how advanced machine learning techniques can help fight obesity in adolescents in Ethiopia. It used the health data that's already available created models that can predict what will happen and suggested ways to help adolecents who are obese. By using machine learning and working with health experts this research wants to help make policies that are based on facts and that can reduce the bad effects of obesity.

1.2. Statement of the problem

Obesity among adolescents is a health concern in Central Ethiopia like other places in the world (WHO, 2022). The surge is largely driven by dietary shifts towards people are eating processed foods, less physical activity, and changing in socio-economic factors (Abarca-Gómez et al., 2017; Gebreyohannes et al., 2019).

Even though it is know that some things can predict obesity like what people eat and how active they are but still don't know which ones are most important in Central Ethiopia. The majority of research concentrates on Addis Ababa rather than other regional states (Amare et al., 2021).

There are a number of difficulties with the machine learning applications used in obesity prediction nowadays. Class imbalance is a problem for many algorithms, which frequently results in models that favor majority classes. This problem can be lessened with the use of methods like SMOTE. The methods are only effective when the characteristics are pertinent and the data quality is high (Krawczyk et al., 2018).

Good data is necessary to choose the appropriate features and improve the model's settings in order to generate predictions. Further challenges to creating trustworthy models for adolescents obesity in Central Ethiopia include geographical heterogeneity, missing data, and variations in data collection methods. It is challenging to develop strategies and to determine which models

are most suited for health applications due to the dearth of research on obesity in Central Ethiopia.

The objective is to determine the significant determinants of obesity, address issues with data quality and class disparity, and offer a trustworthy instrument to guide focused treatments (Alem et al., 2019). The goal of this research is to determine how machine learning may be used to address the issue of teenage obesity in Central Ethiopia. It will accomplish this by examining factors such as body mass index, diet, and residence. Finding strategies to keep teenagers from becoming fat is the primary goal.

The development of policies is hampered by the paucity of research on the factors that contribute to obesity in Central Ethiopia. The majority of research ignores regional differences in eating patterns and levels of physical activity in favor of concentrating on Addis Ababa (Amare et al., 2021). By examining regional risk variables including food habits, socioeconomic level, and availability to recreational facilities, a machine learning-based method might close this gap.

The lack of studies on obesity in Central Ethiopia makes policy formulation challenging. The majority of research focuses on Addis Ababa (Amare et al., 2021). By examining risk factors including eating habits, socioeconomic position, and access to recreational facilities, machine learning offers a better way to close this gap.

Adolescent obesity is a growing public health issue, particularly in Ethiopia, where early detection and treatment are crucial to preventing long-term health issues (WHO, 2020). Although machine learning algorithms are ready to predict the risk of obesity, little study has been done to identify which algorithm provides the most accurate and dependable predictions among Ethiopian adolescents. For the purpose of creating focused preventative measures and allocating resources as efficiently as possible, classifying the successful machine learning model is essential.

Despite the advantages, there remain barriers to putting machine learning-based solutions into practice, such as insufficient digital infrastructure and the capability of the healthcare personnel (Woldemichael et al., 2022). To guarantee intervention access, ethical issues like algorithmic bias and data privacy must also be addressed (Panch et al., 2019). To overcome these obstacles,

politicians, data scientists, and public health specialists must work together to create scalable, culturally relevant machine learning solutions for obesity prevention.

By creating and assessing machine learning models specifically for the adolescent population in Central Ethiopia, this project seeks to close these gaps. The objective is to determine the significant determinants of obesity, address prediction issues with data quality and class disparity, and offer a trustworthy, morally sound instrument to guide focused actions.

This study is trying to find ways how to use machine learning to help solve the problem of obesity among adolescents in Central Ethiopia. To do this it will look at things like body mass index, what people eat, and where they live. The main aim is to find ways to stop adolescent from getting obesity. This information can help people who make decisions about health care in central Ethiopia and also help with efforts around the world to stop obesity among adolescents using machine learning.

Research Questions

- What are the key predictors of adolescent obesity in Central Ethiopia Regional State within machine learning models?
- Which machine learning algorithm achieves the highest predictive accuracy in classifying obesity among adolescents in Central Ethiopia Regional State?
- What are the potential challenges and ethical considerations in implementing an ML-based obesity intervention system in Central Ethiopia Regional State?

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of the study is to identify the key predictors of obesity among adolescents in Central Ethiopia Regional State using machine learning models.

1.3.2. Specific Objectives

The specific objectives of the study are

- To identify the key predictors of obesity among adolescents in the Central Ethiopia Regional State
- To identify the machine learning algorithm that provides the best predictive performance for obesity among adolescents
- To identify and analyze the key challenges and ethical issues in implementing a machine learning-based obesity intervention system in Central Ethiopia

1.4. Scope of the study

The research was conducted in the Central Ethiopia Regional State, where obesity among adolescents is becoming a growing concern. The study would involve adolescents aged between 10–19 years to determine obesity machine learning model and associated risk factors. It aims to develop and evaluate machine learning models to predict obesity risk based on dietary, body mass index, and physical activity variables.

1.5. Significance of the study

This study holds substantial importance for various stakeholders, including public health policymakers, healthcare providers, researchers, and the community, in predicting the growing issue of obesity among adolescents in the Central Ethiopia Regional State.

1.6. Limitation of the study

The study may face limitations related to incomplete or biased data due to reliance on self-reported dietary and physical activity information from the dataset. Limited access to comprehensive health records and technological infrastructure in rural areas may restrict data collection and model accuracy. Additionally, the findings may not be generalizable beyond the Central Ethiopia Regional State due to cultural and environmental differences.

1.7. Organization of the paper

This study is divided in to five chapters. The first chapter contains the introduction part. The second chapter is concerned with related literature which gives theoretical knowledge that tries to discuss some of the previous studies. The third chapter will focus on methodology of the study. Chapter four presents the results and discussion. Finally, in chapter five conclusions about the research and recommendations for future research work were presented.

CHAPTER TWO

2. Literature Review

2.1. Introduction

The problem of obesity in teenagers is a deal all around the world especially in countries that are changing really fast. Ethiopia is one of these countries. A lot of people are moving to cities. This is changing what they eat and how much they exercise. The researcher need to look at this problem from angles like how often it happens what makes it happen how people behave and how the environment affects it and also need to think about how technology, like machine learning can help us understand and solve this problem.

The problem of obesity is one of the health problems in the world and it is especially bad for adolescents in countries that are changing quickly. In Ethiopia many people are moving to cities. This is causing a big increase in obesity. This literature review looks at the problem of obesity in adolescents from different angles like how often it happens what makes it happen how people behave and how the environment affects it.

The researcher looked at different studies to learn about obesity in adolescents. Most of these studies were from Africa and other countries that are not rich. It is to understand the basics of obesity and how it affects people and also to see if machine learning can help to solve this problem in Ethiopia.

2.2. Theoretical Foundations and Analytical Limitations

To understand the causes of obesity we need to look at it from different angles. The socio-ecological model is a way to do this. It helps us see how our bodies, behaviors and surroundings all work together to affect our health (Sallis et al., 2015; Story et al., 2020). At the basic level our genes and how our bodies work can make us more likely to be overweight. Studies have shown that our genes can account for between 40% to 70% of the differences in our body mass index (Loos & Yeo, 2022; Locke et al., 2015).

What we do every day is also very important. What we eat how much we move around and how much we sleep can all affect our weight. Our families can also influence our eating habits and weight. How our parents raise us what we eat at home and how much money our family has can all make a difference (Chaput et al., 2020; Liang et al., 2022)..

Our weight might also be influenced by the societies. The food we can purchase, the parks and walkways we enjoy, and the security of our communities may all be important factors. More importantly, our nation's laws and regulations have the power to influence our decisions. How our cities are and what we learn about food in school. Our health may be impacted by food marketing (Shrewsbury & Wardle, 2012).

One method for comprehending these many levels is the socio-ecological model. We frequently don't go far enough when using it in study. Instead of attempting to comprehend how everything works together, we typically merely consider how objects are connected to one another. The complex associations that are crucial for comprehending obesity are difficult to capture using our standard data analysis techniques, such as regression (Sallis et al., 2015).

New techniques for data analysis, such as gradient boosting and Random Forests, are more effective in identifying patterns. There isn't enough reliable data in Ethiopia to apply these techniques. Data must be gathered and utilized in a way that makes sense. This means we need to have datasets that are specific to our region and that track many different things over time. We also need to be able to select the important information and to check that our models are working well and can be trusted (Zhou, 2021; Dugas et al., 2019).

We need to be able to explain how our models work so that people can understand them and trust them. The socio-ecological model of obesity is very important. Obesity is an issue that involves our biology, our behaviors and our surroundings. We need to look at all these levels to understand the causes of obesity.

The causes of obesity are not about our genes or our behaviors. The causes of obesity are also about our communities and our society. We need to look at the socio- model of obesity to understand how all these different levels work together. The socio-ecological model of obesity is a way to understand the causes of obesity.

To develop solutions to the problem of obesity we need to use the socio-ecological model of obesity. We need to look at the causes of obesity from different angles. The causes of obesity are complex. Involve many different factors. The socio-ecological model of obesity helps us understand these factors and how they work together.

The socio-ecological model of obesity is very useful for understanding the causes of obesity. We need to use the socio- model of obesity to develop good solutions to the problem of obesity. The causes of obesity are not about one thing they are about many things. The socio-ecological model of obesity helps us see how all these things work together to cause obesity.

We need to use the socio- model of obesity to understand the causes of obesity. The socio-ecological model of obesity is a way to understand the causes of obesity. It helps us see how our biology, our behaviors and our surroundings all work together to affect our health. The socio-ecological model of obesity is very important, for developing solutions to the problem of obesity.

2.3. Biological Determinants of Obesity

The thing about obesity is that it is very complicated. There are biological factors that contribute to adolescent obesity. These factors include predisposition, hormonal regulation and developmental processes. Adolescent obesity is caused by a combination of predisposition, hormonal regulation and developmental processes (Khera et al., 2019; Beraldi et al., 2021).

Genetic predisposition plays a role in adolescent obesity. Some people are more likely to become obese because of their genes. Advances in genomics have helped us find genes that are associated with body mass index and obesity risk. These genes are like instructions that are inside our bodies. The genetic factors that contribute to obesity interact with what we eat when we are young what our mothers were exposed to when they were pregnant and how we were fed when we were babies. All these things shape how our bodies work for our lives (Xu et al., 2021).

Leptin and ghrelin are examples of hormones that assist regulate hunger and energy expenditure. Weight gain may result from these hormones not functioning properly. Our body composition may be impacted by the hormonal changes that occur throughout puberty. We may gain weight as a result, particularly if we are already at risk for obesity (Li et al., 2020).

Adolescent obesity is also linked to the bacteria that live in our stomachs. These bacteria help us digest food and control how our bodies work. Some research shows that the bacteria in our stomachs can contribute to obesity (Lloyd-Price et al., 2019).

With all this information we still do not fully understand why some people become obese and others do not. This is especially true in places like Ethiopia, where many different factors can contribute to obesity. Machine learning approaches are very helpful in understanding these interactions. Machine learning approaches are better than statistical methods at handling lots of complex data. This is because machine learning approaches can find and model the patterns that contribute to adolescent obesity. Machine learning approaches can. Model these intricate interaction patterns more effectively than traditional statistical methods, which makes machine learning approaches particularly valuable, for understanding adolescent obesity (Zhou et al., 2022; Nguyen et al., 2023).

2.4. Behavioral and Lifestyle Factors

Behavioral factors such as what we eat and how much we exercise are significant for obesity. The relationships between behavioral factors and obesity are complex. For instance watching TV and mobile phones can lead to exercise but this relationship is affected by the neighborhood we live in (Tremblay et al., 2017).

The environment can also affects our behavior. For example if we live in an area of food restaurants we are more expected to eat unhealthy food. Most studies do not look at how these environmental factors interact with our behavior (Popkin et al., 2020).

Socioeconomic status exhibits context-dependent relationships in Ethiopia, higher socio economic status is associated with increased obesity risk (Gebremariam et al., 2020), opposing with patterns in high-income countries. This underscores the importance of regional modeling that accounts for nuanced socio economic status relations with behavioral and environmental factors.

2.5. Environmental and Socioeconomic Determinants

The places we live and the money we have are really important when it comes to obesity risk. If we can walk around our neighborhood and get to places easily and if we have food options we are more likely to be active and eat well (Sallis et al., 2016; Mueller et al., 2018).

When there are a lot of food places in one area it can make people in that area more likely to get obesity especially in places where people do not have a lot of money (Popkin et al., 2020).

How money we have affects whether we are likely to get obesity and this is different at different times in our lives. In countries people who have less money are more likely to get obesity.. In countries that are not as rich people who have more money might be more likely to get obesity at first because they can buy more foods that are high in calories (Dinsa et al., 2012).

Schools also play a role in whether we are likely to get obesity. This includes what food they serve and whether they encourage us to be active. Some people may be more susceptible to obesity when schools do not provide equal opportunity.

2.6. Machine Learning in Public Health

Machine learning has become very important in health because it can look at complicated data with many different parts. Some algorithms like Random Forests and gradient boosting machines for example XGBoost are really good at predicting things in the field of health. They are better than models at doing this as we can see in the work of (Zhou, 2021; Dugas et al., 2019). These models are good at understanding how different things are related to each other in ways, which is common when we are talking about obesity.

Certain methods, such as clustering algorithms, assist us in identifying various forms of obesity. This is helpful as it allows us to prepare ways to assist others (Garcia-Diaz et al., 2021). Deep learning techniques are also helpful. They require a lot of information. They can aid in our comprehension of images and other facts pertaining to obesity (Agarwal et al., 2021). However, it is difficult to use machine learning models in resource-rich environments. The quality of the data and whether or not individuals can comprehend what the models are doing are issues that

need to be addressed. Concerns like prejudice and privacy must also be considered (Floridi et al., 2018; Holzinger et al., 2022).

Recently, efforts have been made to increase the transparency of machine learning models. To demonstrate how the models are creating predictions, they are employing methods such as SHAP values. This is very important if we want to use machine learning in hospitals and public health work in places where people may not know a lot about health or have limited resources. Machine learning is a part of this and making it more understandable will help us use it to improve health. Machine learning models, like these can really help us and machine learning is the thing that is making this possible.

2.7. Machine Learning Applications in Obesity Research

Machine learning applications have shown that they can do a better job of predicting obesity. For example methods like Random Forests and XGBoost are more accurate than ways of predicting obesity especially when they use many different types of information like demographics and behavior and genetics and environment (Dugas et al., 2019; Liu et al., 2024). The things that help predict obesity the often include the body mass index of parents what people eat how much they exercise and what their neighborhood is like. This shows that obesity is a problem with many parts.

Some studies have found that machine learning models can figure out when it is most important to help someone and can predict how much someone will weigh over time accurately than other methods (Althoff et al., 2022). However there are still problems with using these models in groups of people because the information can be different and biased. Because of this, not everyone will find the models to be effective (Polyzotis et al., 2018; Obermeyer et al., 2019). In order to address these issues, the models must be thoroughly evaluated, and stakeholders must be included to ensure that the models are equitable and function effectively in various contexts.

Machine learning has done a job of predicting obesity in places with a lot of money like it did for (Dugas et al., 2019). It has helped us understand how different things are related to obesity. In Ethiopia machine learning is not used very much to predict obesity. There have been a studies that show methods like Random Forest and XGBoost are better than traditional methods just like

in other places. These studies do not talk about the special challenges of using machine learning in places with limited resources. Machine learning applications like Random Forests and XGBoost are still very important for predicting obesity, in Ethiopia.

2.8. Application of Machine Learning in Obesity Prediction

For example a recent review by (Xue et al., 2026) shows that using a model with groups of people is still a problem. Models that are trained with data from China or Europe do not work well when used in places because the data is different and the people are not the same. This means that we need to collect data check it and make sure it is correct for each region.

Also it is hard to understand how some models work. Like the model called RF. It is like a box that you cannot see inside. Some new techniques like SHAP, which was developed by Wang and other people in 2025 can help explain how these models work (Wang et al., 2025). We need to do more work to make these techniques useful for public health in Ethiopia.

So even though machine learning has some advantages for predicting obesity in teenagers using it in Ethiopia will be hard because of problems with data understanding how the models work and making sure they are relevant to the context of Ethiopia. We need to make sure that machine learning and models like Xue and other people talked about and techniques like SHAP are used in a way that's effective, for Ethiopia and for predicting adolescent obesity.

2.9. Overview of Machine Learning Algorithms

2.9.1. Decision Tree

Decision Trees are a way to learn from data that's easy to understand. They work by splitting the data into groups over and over again. The algorithm looks for the way to split the data so that it can tell the difference between the different groups. It does this by using something called the Gini index or information gain. This creates a tree- structure that shows how the decisions are made (Alanazi et al., 2024).

The great thing about Decision Trees is that they are easy to understand. Even people who are not experts can look at the tree structure. See how the decisions are made. Each split in the tree is

like a rule that says if this is true then do that. For example if someones body mass index is than 85 percent then they should be checked to see how active they are. This makes it easy to see how the model is making its decisions. Decision trees are capable of handling a wide variety of data, including nonlinear data. Additionally, they are adept at managing data that contains unnecessary information or outliers (Alanazi et al., 2024).

Individual Decision Trees have significant drawbacks despite these benefits. They are more likely to capture noise unique to the training data rather than broadly applicable patterns due to overfitting, especially as tree depth grows. High variance is shown by the fact that slight variations in training data can result in very diverse tree architectures. According to Alanazi et al. (2024) and Liu et al. (2024), Decision Trees may also have trouble capturing additive structures in which several predictors simultaneously affect outcomes in ways that sequential splits do not adequately depict.

Decision trees are not flawless. They may fit the facts too closely and be very complex. This implies that they may not function effectively with fresh data. Small changes in the data can also make changes in the tree structure. This means that Decision Trees can be a bit unstable. They can also have trouble with things that are additive. This means that multiple things can affect the outcome in ways that are not easy to see with Decision Trees (Alanazi et al., 2024).

In the case of obesity Decision Trees have been used to identify groups of people who're at high risk. They can provide rules that are easy to understand and can help doctors make decisions. For example a Decision Tree might show that teenagers who do not get exercise and eat a lot of sweets and have parents who are obese are at high risk. This can help doctors provide targeted care and can be more effective, than just looking at a list of risk factors.

2.9.2. Random Forest

The Random Forest method helps to fix the problems of Decision Trees by using a group of trees to make predictions. This is done by creating Decision Trees, usually hundreds or thousands and training each one on a slightly different set of data. When each tree is making a decision it only looks at a selection of the information it has. This helps to make sure the trees are not all making

the mistakes. The final prediction is made by looking at what most of the trees think or by averaging out their answers (Alanazi et al., 2024) (Liu et al., 2024).

When attempting to forecast things like obesity, Random Forest works incredibly well. Because it does not become caught in a pattern, it is far superior to utilizing a decision tree. This implies that it is able to forecast even in the presence of fresh data. Additionally, the Random Forest approach may identify patterns in the data without being instructed on where to look. It is capable of handling many kinds of data and is unaffected by strange details or irrelevant information. Without the requirement for data, it can even predict its performance (Alanazi et al., 2024) (Liu et al., 2024).

Random Forest offers several substantial advantages for obesity prediction. The ensemble approach dramatically reduces over-fitting compared to single trees, producing models that generalize well to new data. The algorithm captures complex non-linear relationships and interactions automatically, without requiring pre-specification. It handles mixed data types, is robust to outliers and irrelevant features, and provides internal estimates of generalization error (out-of-bag error) without requiring separate validation sets.

Furthermore, Random Forest offers variable relevance metrics that quantify the contribution of each predictor to model performance. These importance metrics can be based on the overall reduction in node impurity attributed to splits on a variable (Gini significance) or the drop in prediction accuracy when a variable is permuted (permutation important) (Liu et al., 2024). This feature makes it easier to recognize important risk variables and comprehend the fundamental causes of obesity risk.

The ability of Random Forest to identify the most crucial pieces of information for forecasting is another advantage. It does this by looking at how difference each piece of information makes. This helps us understand what things are closely linked to obesity (Wang et al., 2025).

2.9.3. K-Nearest Neighbor

The K-Nearest Neighbor method classifies things based on how similar they're to other things around them. This is a simple idea and it does not make a lot of assumptions, which is good.

However it can be slow. Get confused by things that are not important especially when there is a lot of information to look at like with obesity data (Liu et al., 2024). To make it work well we need to make sure the information is scaled correctly. We choose the right number of neighbors to look at.

The K-Nearest Neighbor method has some things going for it. It is simple and easy to understand because it is based on the idea that things that are similar're likely to be in the same group. It can also handle problems and figure out what is going on with multiple groups of things. We do not need to spend a lot of time training it because it just needs to look at the data we already have which makes it easy to add data without having to start over like (Alanazi et al., 2024).

The K-Nearest Neighbor method also has some big problems when it comes to predicting obesity. It takes a time to make predictions because it has to look at all the data we have to figure out what is going on with each new thing. It can also get confused by things that're not important, which can make it hard to see what is really going on. We have to be careful about how we scale the information and choose the number of neighbors because this can affect how well it works. The method also does not give us a lot of insight into what's driving the predictions except for looking at the neighbors (Alanazi et al., 2024) (Liu et al., 2024).

When it comes to predicting obesity in teenagers the K-Nearest Neighbor method has had results. It usually does not work well as other methods, like Random Forest but sometimes it can work as well as simpler methods, like Decision Trees (Alanazi et al., 2024) (Liu et al., 2024).

2.10. Comparative Performance Analysis

Although precise rankings differ slightly between studies and outcome metrics, direct comparisons between the four algorithms show similar performance hierarchies. Key research and their conclusions about model performance are listed in Table 2.1.

Table 2-1 Summary of Studies on Machine Learning for adolescent Obesity Prediction

Study/ Year	**Population**	**Sample Size**	**Algorithms Compared**	**Key Findings**
Alanazi et al. (2024)	Saudi Arabia (Arar and Riyadh)	Not specified	RF, DT, KNN, MLR	RF outperformed all models (R^2=0.98); mean age at onset 10.8 years; key predictors: parental factors, family history, lifestyle
Liu et al. (2024)	Hong Kong	442,898 (P4), 344,186 (P6)	DT, RF, KNN, LR, XGBoost, SVM	XGBoost best (accuracy 0.72-0.74); RF strong performance; LR baseline; key predictors: weight, height, sex, age, exercise
Sewpaul et al. (2023)	South Africa (female adolescents)	Not specified	LR, RF	RF superior to LR for obesity classification; demonstrated value of ensemble methods in diverse populations
Xue et al. (2026)	China (Beijing and Tangshan)	19,024 training; 2,410 validation	RF, XGBoost, LightGBM, LR	XGBoost best (AUROC=0.875); LR (AUROC=0.718); key predictors: birth measures, parental BMI, lifestyle
Lifestyle-based model (2025)	China (Sichuan)	2,338	LASSO + LR	LR with feature selection achieved AUC=0.91; key predictors: gender, parental BMI, sleep, diet, physical fitness

A detailed technical comparison of the four machine learning algorithms from different research shows that performance variations are impacted by each model's intrinsic qualities as well as implementation specifics, data features, and optimization techniques. In order to explain the observed performance hierarchies, these elements are rigorously examined in this section.

In several investigations, Random Forest (RF) regularly shows greater or almost superior prediction ability. Its resilience is derived from the ensemble aggregation of many decision trees, which effectively captures complicated, nonlinear interactions among predictors and reduces overfitting, a typical problem with single decision trees. For example, RF obtained an R2 of 0.98, indicating excellent match, in the Saudi Arabian research (Alanazi et al., 2024). Although exact tuning details are frequently underreported, hyperparameter tuning techniques including improving the number of trees, maximum depth, and feature subset sizes probably improve this performance. RF's resilience is further enhanced by its capacity to handle high-dimensional, heterogeneous data, such as combining genetic, behavioral, and demographic variables, particularly when paired with methods like feature relevance analysis to choose pertinent predictors.

Although interpretable and useful for rule-based risk stratification, decision trees (DT) typically perform worse than ensemble approaches. Their vulnerability to overfitting, particularly in datasets with high variability or small sample numbers, is the main cause of this. The decision tree's performance trailed behind RF and conventional regression models in Alanazi et al. (2024), underscoring its limited capability in the absence of ensemble boosting or pruning techniques. To increase DT performance, proper implementation such as cross-validation, trimming, and feature selection is crucial, yet reported research frequently omit or undervalue these specifics.

The prediction abilities of K-Nearest Neighbor (KNN) are inconsistent and typically low. Performance is severely hampered by its sensitivity to irrelevant features and the curse of dimensionality, especially in datasets with a large number of features or unscaled variables. KNN fared worse than ensemble algorithms in Liu et al. (2024), perhaps because it was difficult to define meaningful distance measurements across different predictor types. Although methods like feature scaling, dimensionality reduction, or distance metric modification might increase

KNN's efficacy, they are seldom used or documented, which limits its use in challenging obesity prediction applications.

As a computationally efficient starting point, logistic regression (LR) frequently produces respectable but subpar results when compared to more sophisticated methods. In the Beijing-Tangshan research (Xue et al., 2026), for instance, LR's AUROC was 0.718 whereas XGBoost's was 0.875. However, LR's predictive accuracy may be greatly improved when paired with contemporary feature selection methods like LASSO; in a Chinese teenage research, it achieved an AUROC of 0.91 (Lifestyle-Based Prediction Model, 2025). This emphasizes how crucial feature engineering and selection are to optimizing the usefulness of simpler models. Nonetheless, LR's performance remains sensitive to multicollinearity, class imbalance, and the quality of input features.

2.11. The Ethiopian Context

Ethiopia has seen a lot of changes in the few decades. The country of Ethiopia has grown fast and a lot of people have moved to cities in Ethiopia. Because of this people in Ethiopia are eating differently and an exercising less. Studies in Addis Ababa have found that many teenagers are overweight or obese (Gebremariam et al., 2020).

In fact fifteen percent of teenagers in Addis Ababa are overweight or obese which is an increase from the past. These changes are happening mostly in cities in Ethiopia and the areas around them. In these places people in Ethiopia have access to foods and are less likely to be active. Traditionally people in Ethiopia ate a lot of grains, legumes and vegetables. Now people in Ethiopia are eating refined carbohydrates, added sugars and fats. At the time people in Ethiopia are not exercising as much as they used to (Tefera et al., 2022)..

Many adolescents in cities in Ethiopia are spending time sitting and not being active. The way people in Ethiopia are active has changed a lot with activity at work and while getting around and more time spent sitting around especially among teenagers in cities in Ethiopia. These changes are happening at a time when some parts of Ethiopia are still dealing with people not getting nutrients, which creates a big problem for the health care system in Ethiopia. (Gebremariam et al., 2020).

Physical activity patterns have shifted concurrently, with declining occupational and transportation-related activity and increasing sedentary behaviors, particularly among urban adolescents (Tefera et al., 2022). These shifts take place against a backdrop of ongoing undernutrition in some areas, resulting in a twofold burden of malnutrition that poses particular difficulties for public health systems (Mayosi et al., 2021).

The Central Ethiopia Regional State differs slightly from the rest of the nation. The residents have different lifestyles and socioeconomic origins. To learn how obesity is impacting the local population, conduct research in this area. Ethiopia has a problem with obesity, and we should examine the Central Ethiopia Regional State to understand how it affects the local population and how we can promote better health among Ethiopians in general and the Central Ethiopia Regional State in particular.

2.12. Research in Central Ethiopia

Although obesity trends in Addis Ababa have been the subject of several studies, there is still a dearth of research that focuses exclusively on Central Ethiopia. Compared to the capital or rural areas, the region's metropolitan and semi-urban areas have distinct features that may affect obesity risk factors in different ways. According to preliminary data, nutrition transition patterns may be more varied in these environments, with both the adoption of contemporary eating habits and the survival of traditional dietary components (Gebremariam et al., 2020).

There seem to be differences in the patterns of physical activity as well, with transportation and leisure activities changing significantly while occupational activity declining less dramatically (Tefera et al., 2022). These places' built environments frequently blend aspects of more modern construction with traditional urban design, posing special possibilities and problems for physical exercise. Similar transitions may be seen in food contexts, where quick-service restaurants and contemporary retail stores coexist with traditional markets. These contextual characteristics imply that prevalence and drivers of obesity may be significantly different from patterns seen in other Ethiopian contexts, requiring region-specific research.

One major knowledge gap that restricts the creation of focused treatments is the dearth of current research in this area using sophisticated analytical approaches. Furthermore, despite the proven

effectiveness of these techniques in other contexts, no research has yet looked at the potential of machine learning approaches to obesity prediction and prevention in Central Ethiopia.

2.13. Literature Gap Analysis

The current study attempts to fill in a number of significant gaps found in the thorough analysis of the body of existing knowledge. First, despite the proven effectiveness of these techniques in other contexts, machine learning applications to obesity research are conspicuously lacking in Ethiopia. Although predictive modeling has grown in popularity in high-income nations, few researches have examined its potential in low-resource settings with varying risk factor profiles and difficulties with data availability.

This gap is especially noticeable for teenage populations in sub-Saharan Africa's urbanizing regions, where quick changes in diet may produce distinct risk factor patterns that deviate from accepted models.

Second, rather than looking at the intricate relationships between biological, behavioral, and environmental variables, the majority of the research that has been done on obesity determinants in Ethiopia has concentrated on isolated elements. The majority of research has used reductionist methods, which use traditional statistical techniques to assess individual risk factor categories. This has limited our knowledge of how several factors may increase or decrease the impact of one another in this particular group.

Third, there is essentially no implementation study examining the real-world difficulties of implementing machine learning solutions in Ethiopia's healthcare system. There are still important unaddressed challenges about data gathering infrastructure, model interpretability for local health workers, ethical issues unique to teenage populations, and integration with current health systems.

Even though machine learning techniques like Random Forest and XGBoost have been shown to be effective in high-resource environments, there are substantial technical obstacles to their use in Ethiopia. Effective feature engineering and model training are hampered by the scarcity of high-quality, longitudinal, multi-modal datasets required for robust modeling.

Furthermore, the generalizability of the model is compromised by the majority of current research' lack of strict validation procedures, such as external validation and calibration tailored to the Ethiopian population. Model accuracy and dependability are hampered by the lack of region-specific hyperparameter tweaking and the difficulty of resolving class imbalance in small sample numbers.

Interpretability is still a major obstacle; explainability methods like SHAP must be modified for local stakeholders with little technical knowledge in order to deploy sophisticated models like ensemble trees. To create efficient, scalable machine learning solutions that are suited to Ethiopia's particular epidemiological and infrastructure setting, it will be crucial to address these problems using sophisticated data augmentation, transfer learning, and explainability techniques.

Lastly, there are still methodological issues with obesity research in the area, especially with regard to the validation of prediction models and their adaption to local settings. None of the current investigations have used the strict model validation procedures typical of machine learning research, and the majority have relied on cross-sectional designs with poor generalizability. The development of successful, context-appropriate therapies for Ethiopia's expanding teenage obesity problem is severely hampered by these deficiencies taken together.

2.14. Conceptual framework for adolescent obesity prediction

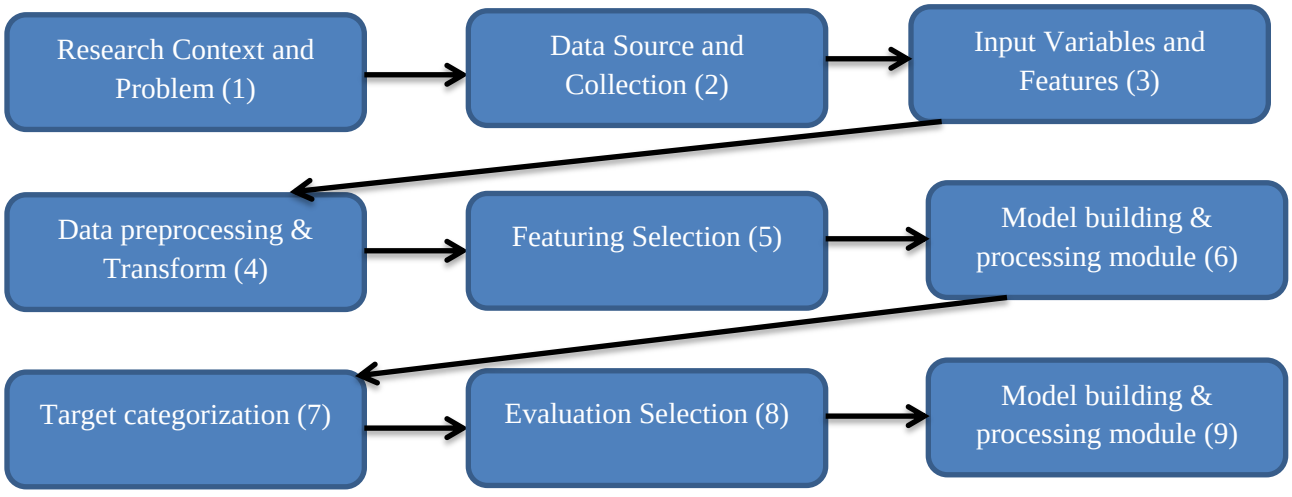


Figure 2-1-1 Conceptual framework for adolescent obesity prediction

The Central Ethiopia Regional State's distinct public health environment serves as the foundation for the study's conceptual framework. The framework begins with a -sectional dataset of 10,221 teenagers between the ages of 10 and 19. The input features, such as physiological measures, behavioral risk characteristics, and demographic markers, are mapped from the raw data. Then, utilizing methods like ordinal and dummy encoding, missing value filtering, and sophisticated resampling algorithms, these characteristics are converted into a clean and normalized representation.

The Chi-Square independence test and other statistical filter methods are then used to choose the pertinent features from the preprocessed data. Eight different machine learning classifiers, such as Random Forest, SVM, and XGBoost, are then trained and assessed using the chosen features. To stabilize prediction performance, the models are adjusted using a dual-layered structure with an internal 10-fold cross-validation loop. The resulting models are then applied to teenage data that has not yet been seen in order to automatically and accurately classify the data.

The framework's last section discusses the operation of evaluation selection and target classification. In order to determine if a person's weight is healthy or not, each prediction in the Target categorization section is placed into one of six categories depending on body mass index. range from being extremely fat to being underweight. A binary confusion matrix is used in the Evaluation Selection section to assess these groups' performance. This tool aids in determining the model's effectiveness and forecast accuracy.

2.15. Summary

This review of the literature demonstrates the significance of addressing teenage obesity in Ethiopia. It also demonstrates how machine learning may be used to this issue. We can discover how obesity occurs and how to prevent it by integrating concepts from science and machine learning. This review emphasizes how factors including our diet, level of physical activity, and place of residence all have an impact on becoming overweight. Additionally, it demonstrates that machine learning outperforms other approaches for comprehending these intricate interactions.

While highlighting crucial implementation issues unique to low-resource environments, it demonstrates the superiority of machine learning techniques over conventional ways for

modeling these intricate relationships. This project will advance scientific understanding and useful intervention techniques by utilizing the identified knowledge gaps. The research strategy takes use of Ethiopia's special capabilities for technology innovation in public health while particularly addressing the shortcomings of earlier investigations. The technique, discussion, and data analysis used to address this issue will be covered in the study's following chapter.

CHAPTER THREE

3. Methodology

3.1. Data Source

This study uses primary data that was collected through a dataset conducted in the Central Ethiopia Regional State (CERS). The dataset consist of 10,221 adolescents aged 10–19 years assessing dietary habits, physical activity levels, socioeconomic factors, and anthropometric measurements. The dataset contributes to support regional statistical systems for ongoing monitoring and policy formulation.

The data about the adolescents in the region is intended to gather estimates of important indicators that are statistically robust and globally comparable. These estimates are then utilized to evaluate the status of adolescents from 10-19 years with regard to nutritional habits, physical activity levels, socioeconomic factors, and body measurements.

Moreover it is used as a tool to produce statistics for tracking the advancement of both national and international goals made to advance the welfare of adolescents, such as the Sustainable Development Goals. In addition to being a source of information for metrics pertaining to the advancement and enhancement of the well-being of men, women, and adolescents in Central Ethiopia regional state, this study would also contributes to the development of statistical systems.

3.2. Study Variables and Measurements

The obesity status among adolescents aged 10–19 years are the study's primary outcome variables. Many potential risk factors for adolescent overweight and obesity were considered aligned datasets. Variables were selected based on theoretical relevance, completeness of data, and contribution to model performance in machine learning algorithms. The factors that were included are: age (in years), sex (male/female), height (in meter), Weight (in kilograms) frequency of consumption of energy-dense foods, physical activity frequency (activity per week), and amount of drinking water (in liters).

Only variables with less than 50% missing data were retained. Most common for categorical variables was used to fill in the missing values in these variables. To maintain algorithmic stability and data quality, variables with more than 50% missing data were eliminated. Recursive feature elimination and integrated techniques inside the machine learning pipeline were then used to guide feature selection and priority ranking. Based on their statistical significance and contribution to model correctness, features were graded.

3.3. Analytical Analysis

The Python programming language (version 3.14.5) and a Jupyter notebook environment were used for the data processing and analysis. To keep class proportions for obesity and non-obesity, the data was divided into training 70% and testing 30% sets using stratified sampling. After that the data is divided into training and testing sets prior to model training, the training dataset was oversampled using SMOTE to solve class imbalance. This lessened bias against the majority class by ensuring that the models were trained on a balanced sample.

To avoid information leaking into the test data, SMOTE was only used on the training set following data splitting. The imblearn.over_sampling was used for the implementation. The imblearn library's SMOTE class. The five closest neighbors in the feature space were used to create synthetic samples when the k_neighbors parameter was set to 5. By balancing the distribution of classes, this method improved the classifiers' ability to identify minority class patterns. Model performance was evaluated with and without SMOTE in order to further confirm the impact of oversampling; the findings showed that models trained with SMOTE had higher sensitivity and overall accuracy. To guarantee repeatability and stop data leakage, all preprocessing procedures including SMOTE oversampling were included into a scikit-learn pipeline.

Hyper-parameters for each algorithm were optimized via grid search with 10-fold cross-validation to enhance predictive performance. In addition eight popular machine learning (ML) algorithms to predict the obesity for adolescents aged 10–19 years in Central Ethiopia Regional state: K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Gaussian Naive Bayes, Multilayer Perceptron (MLP), Gradient Boosting and

Extreme Gradient Boosting were used to their confirmed effectiveness in classification tasks. The researcher then compared the outcomes of each algorithm in terms evaluation metrics.

3.4. Machine learning approach

3.4.1. Feature Selection

In machine learning feature selection (FS) is a preprocessing step. It helps to eliminate features, cut down on computation time, and increase prediction model accuracy. The main goals of feature selection are to control the dataset's noise, outliers, and to make it consistent. To assess feature relevance in relation to its predictive power the researcher employ chi-square of the filter method approach was used. Features are chosen in filter methods based on the properties of the data rather than using learning models.

The chi-square method was selected because many of our features—such as sex, family history, high-calorie food consumption, vegetable intake, and physical activity and categorical variables. The chi-square test efficiently evaluates the dependence between these categorical features and the target variable providing a quick and computationally efficient means of ranking features based on their statistical association. Using a filter approach allows us to perform an initial screening independent of any specific machine learning algorithm, ensuring an unbiased selection process (Chandrashekar & Sahin, 2014).

Features are ranked according to a set of criteria. The features with the highest rankings are selected. These criteria measure various aspects of the features, such as their dependence on the class label, their correlation with one another, and their dependence on one another.

From the dataset 14 features were used by the researcher to classify each instance. They are sex, age, height, weight, family history, high calorie food, vegetable, daily meals, foods between meals, water, monitoring calorie, physical activity, using of technological devices, and body mass index.

3.4.2. Overview of Machine Learning System

Machine learning enables computer models to adapt to occurrences and to detect and generalize patterns without being explicitly programmed (Muhamedyev, R. I., 2015). When the model's performance measure for a task improves with experience and when the model can do the task effectively then the model is considered to be learning (El Misilmani, H. M., & Naous, T, 2019). Datasets containing features would be fed into a machine learning model for a particular job with the labels representing the desired result. After that the learning model is trained using various learning techniques. Within the medical domain, machine learning models acquire domain knowledge from data ranging from medical imaging to doctor-level observations.

Algorithms analyze all the data extracting the features or descriptors that can aid in making precise predictions about the result. The following figure shows a depiction of the machine learning process from data preparation to model evaluation.

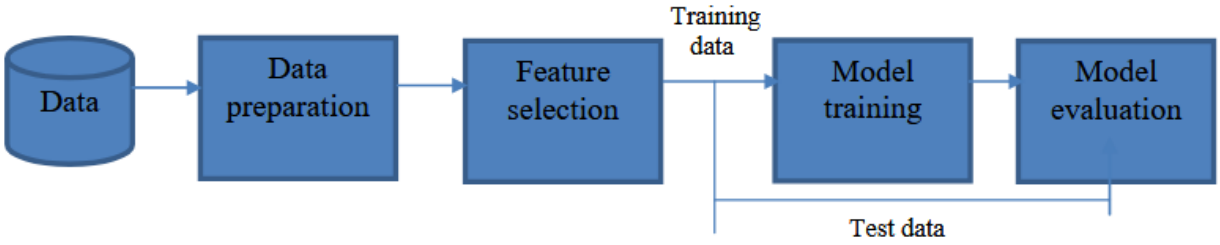


Figure 3-1 Overview of machine learning machines (Alshammari et al., 2020)

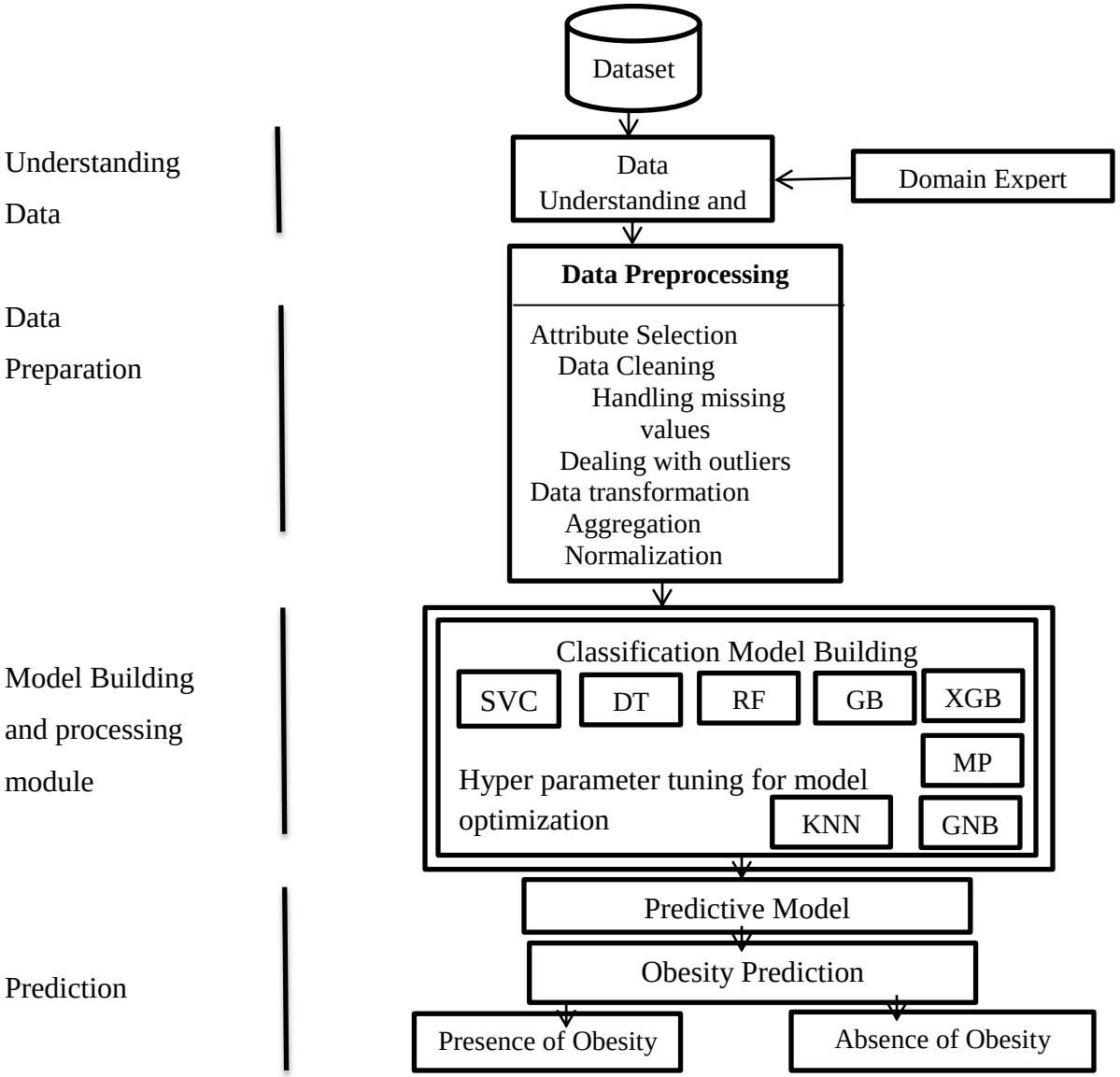


Figure 3-2 Proposed ML model for Obesity

Machine learning incorporates a number of learning strategies, including supervised and unsupervised learning, which are distinguished by the availability of class labels (Osisanwo et al., 2017).

Supervised learning, where models are trained on labeled data, was primarily employed for obesity prediction. The training involved splitting the dataset into training and testing subsets, with models learning to classify adolescents as obese or not based on input features. Unsupervised learning techniques, such as clustering were not used in this study but

acknowledged for exploratory data analysis (Khanum et al., 2015).

An overview of each of the eight machine learning methods that were trained for each nutritional indicator is given below: k-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Gaussian Naive Bayes, Multilayer Perceptron, Gradient Boosting and Extreme Gradient Boosting.

For this study, the algorithms were trained to find features that predict obesity in Central Ethiopia regional state adolescents aged 10–19 years. The training and test datasets were created by first splitting the data. For the training and testing datasets 70% and 30% of the data was used respectively.

To enhance model performance hyper-parameters for each machine learning algorithm were optimized using a grid search approach combined with 10-fold cross-validation. Specifically for each algorithm, a set of hyper-parameters was systematically varied within predefined ranges to identify the combination that yielded the best performance based on the F1-score on the validation folds.

Random Forest (RF)

A machine learning technique called random forest (RF) uses the information collected from multiple decision trees to classify data. Unlike decision trees, which use only one tree to make predictions, random forests fit numerous uncorrelated classification trees and use the average of all individual trees to ensure an enhanced accuracy of prediction in comparison to individual trees. For every decision tree in the RF, Random Forest selects a random sample (with replacement) of the training dataset. This process is known as bootstrap aggregation or bagging. Furthermore, a random collection of features for every tree is the basis for node splitting in the RF model. As a result, one tree's characteristics may differ from another. The resilience of RF prediction accuracy is influenced by the bagging procedure and differences in feature selection for every tree (Hassanat et al., 2014).

K-Nearest Neighbor (KNN)

By evaluating the closest neighbor with regard to a value of k, the KNN approach calculates how many nearest neighbors need to be investigated in order to explain the class of a sample data point (Soofi & Awan, 2017). The two types of nearest neighbor techniques are structure-based KNN and structure-less KNN. The structure-based approach works with the data's fundamental structure, which contains less mechanisms related to training data samples. The complete dataset is spliced into sample data points and training data in the structure-less technique. The distance between each sample point and every training point is computed, and the point with the least distance is referred to as the nearest neighbor (Tomography et al., 2019).

Decision Tree (DT)

DT is a machine learning technique that uses a tree- model to classify data. It works by partitioning the data into smaller subsets based on the values of the input features. The Decision Tree model is trained on the training data. Then it is used to make predictions on the test data. The Decision Tree model is simple to understand and interpret. It can handle both categorical and numerical data. However it can be prone to overfitting especially when the trees are deep. To avoid overfitting techniques such, as pruning and regularization can be used (Jijo & Abdulazeez, 2021).

Decision trees are a type of machine learning where each leaf on the tree represents a class label and each factor represents a part of the tree. Decision trees are often used to predict kinds of illnesses when the output is either continuous or categorical. Decision trees can use any kind of data including text, pictures and numbers as input. When the output is continuous decision trees create a regression tree. When the output is categorical they create a classification tree.

Support Vector Machines (SVM)

Support Vector Machines (SVM) are used for classification and regression tasks. SVM aims to find the optimal line that best separates data points of different classes by maximizing the margin between the closest points of each class. SVM can handle both linear and nonlinear data by using special functions that transform the input data into higher-dimensional spaces where a linear

separation becomes feasible (Cortes, C., & Vapnik, V., 2018). SVMs are widely used because they are robust and can perform well with low training data.

Gaussian Naive Bayes (GNB)

Gaussian Naive Bayes (GNB) is a type of classifier that uses Bayes theorem with independence assumptions between features. Gaussian Naive Bayes assumes that the continuous features follow a normal (Gaussian) distribution within each class. It calculates the likelihood of the features given the class and combines this with prior probabilities to predict the class label that maximizes the posterior probability. GNB is widely used for text classification, spam detection because it is simple and efficient (Murphy, 2012).

Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a type of neural networks composed of multiple layers of nodes. Each node applies a function to the weighted sum of its inputs allowing the network to learn complex patterns. MLPs are trained using algorithms such as back propagation to minimize the error between the predicted and actual outputs. They are widely used in tasks including classification, regression, and pattern recognition (Zhang et al., 2022).

Gradient Boosting

Gradient Boosting is a machine learning technique that builds an ensemble of weak learners in a sequential manner. Each new tree is trained to correct the errors of the combined ensemble of trees. Gradient Boosting is known for its predictive accuracy and is widely used in classification and regression tasks (Chen & Guestrin, 2016).

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an optimized implementation of boosting designed for speed and performance. It employs regularization techniques to prevent over fitting and incorporates features such as tree pruning, handling missing values. XGBoost has gained popularity for its superior performance in machine learning competitions and practical applications (Chen & Guestrin, 2016).

3.5. Model Validation & Performance Metrics

Performance Measures

To evaluate the models' accuracy on the test dataset, the researcher employed a confusion matrix. A confusion matrix was used to evaluate the models' sensitivity, specificity, and f1-score. The four potential results, assuming a conventional binary classifier with a confusion matrix, are listed in Table 3.1.

Confusion matrix

The confusion matrix is a two-by-two table that shows the values of True Negatives, False Negatives, True Positives, and False Positives resulting from predicted classes of data.

The confusion matrix makes it feasible to measure assessments such as specificity, sensitivity, and accuracy of predictions (Motbainor et al., 2015).

Table 3-1 Confusion matrix

		Actual Observations	
		Positive	Negative
Prediction Results	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Accuracy

Any prediction algorithm's performance is estimated based on its accuracy. Accuracy is the ratio of predictions to the total number of input samples.

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Negative + False\ Positive}$$

Sensitivity (recall)

Sensitivity is the percentage of positive instances that were correctly predicted as positive. Sensitivity is also known as recall.

$$\text{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Specificity (precision)

Specificity is the percentage of negative instances that were correctly predicted to be negative. Specificity is also known as precision.

Specificity = $\frac{True\ Negative}{True\ Negative+False\ Positive}$, Precision = $\frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$

F-1 score

The mean of recall and precision. It is a balanced metric that considers both false positives and false negatives.

F1 – Score = 2 * $\frac{Precision * recall}{Precision+ recall}$

3.6. Ethical Considerations

This study employed machine learning to address adolescent's obesity in Central Ethiopia. To preserve privacy, all data is safely kept. Because obesity-related predictions are sensitive, algorithmic fairness and bias avoidance are crucial.

The study design is also guided by kindness and community involvement. To ensure that the results translate into practical measures, local health authorities will assess the technique to ensure it is in line with cultural and socioeconomic conditions.

CHAPTER FOUR

4. Results and Discussion

This chapter presents the exploratory analysis of the collected data from the data set and the main results of the experiments performed in the model development process.

4.1. Load the obesity dataset

Here the researcher loads the desired obesity data into python using the ‘CEobesity_dataset = pd.read_csv('AbebashObesityDataSet.csv') command and imports the necessary files.

```
import pandas as pd
import numpy as np
seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from collections import Counter
from imblearn.over_sampling import SMOTE
from matplotlib import pyplot as plt
%matplotlib inline
```

Loading data set it shows 10221 rows of data with 14 columns

```
CEobesity_dataset = pd.read_csv('ObesityDataSetLogReg.csv')
CEobesity_dataset.shape
(10221, 14)
```

And we can see the whole data set using the following command.

print ("Number of Central Ethiopia obesity data points:", CEobesity_dataset .shape[0])

CEobesity_dataset

print("Number of Central Ethiopia obesity data points:", CEobesity_dataset.shape[0])

CEobesity_dataset

The following figure shows the result of the above code

```
Number of Central Ethiopia obesity data points: 10221
```

	Gender	Age	Height	Weight	family_history	FAVC	FCVC	NCP	CAEC	CH2O	SCC	FAF	TUE
0	Female	17	1.620000	64.000000	yes	no	2.000000	3.000000	Sometimes	2.000000	no	0.000000	1.000000
1	Female	11	1.520000	56.000000	yes	no	3.000000	3.000000	Sometimes	3.000000	yes	3.000000	0.000000
2	Male	16	1.800000	77.000000	yes	no	2.000000	3.000000	Sometimes	2.000000	no	2.000000	1.000000
3	Male	16	1.800000	87.000000	no	no	3.000000	3.000000	Sometimes	2.000000	no	2.000000	0.000000
4	Male	11	1.500000	69.800000	no	no	2.000000	1.000000	Sometimes	2.000000	no	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10216	Female	15	1.520000	42.000000	no	yes	3.000000	1.000000	Frequently	1.000000	no	0.000000	0.000000
10217	Female	13	1.733263	50.890363	no	no	3.000000	3.765526	Frequently	1.656082	no	0.427770	0.555967
10218	Female	15	1.722527	50.833029	no	no	3.000000	3.285167	Frequently	1.569082	no	0.545931	0.469735
10219	Male	15	1.734832	50.000000	no	no	1.123939	4.000000	Frequently	1.000000	no	1.303976	0.371941
10220	Female	19	1.882779	64.106108	yes	yes	3.000000	3.000000	Sometimes	2.843777	no	1.488843	0.009254

10221 rows × 13 columns

Figure 4-1 Total number of data in the data set

4.2. Descriptive analysis and Statistics

In descriptive statistics, we can easily understand the data set by calculating the various summary measures, such as mean, standard deviation, minimum value, maximum value, and others for all features. Here in the following figure, there is the summary report for some of the obesity dataset attributes.

CEobesity_dataset.describe()

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	10221.000000	10221.000000	10221.000000	10221.000000	10221.000000	10221.000000	10221.000000	10221.000000
mean	14.511398	1.697132	80.321482	2.388340	2.671675	1.985350	1.072861	0.670558
std	2.850823	0.095517	24.940163	0.547886	0.834042	0.631857	0.901332	0.643914
min	10.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	12.000000	1.622241	60.000000	2.000000	2.646717	1.530046	0.110174	0.000000
50%	15.000000	1.700000	78.000000	2.241606	3.000000	2.000000	1.000000	0.645400
75%	17.000000	1.765258	98.259423	3.000000	3.000000	2.426094	1.949070	1.000000
max	19.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	2.000000

Figure 4-2 Descriptive analysis of the dataset

In the dataset above, the categories that contain string data are not included in the descriptive analysis so the categories that have a string value are not included here so they did not have any use for the logistic regression in the later section for this reason the string data types must be changed to numerical data types.

4.3. Data Preprocessing

The important part of preprocessing is feature engineering, before balancing the data and splitting the dataset for training, the researcher should convert all categorical variables into numerical.

Some of the features in the dataset contain string data that are unusable for the logistic regression prediction. Therefore the researcher changes these objects into integer values. For this purpose, the single values in those columns are given integer values starting from either 0 or 1. For instance values with only ‘yes’ or ‘no’ options, 1 is assigned for ‘yes’ and 0 is assigned for “no”. The values of Gender types convert into integers 0 and 1, 0 is assigned for ‘Male’ and 1 is for ‘Female’. On the other hand the values of if there is an obesity family history convert into integers 0 and 1, 0 is assigned for ‘no’ and 1 is for ‘yes’. In the category of the person eat any food between meal, if the person eat any food between meals sometimes it is assigned ‘1’, the person eat any food between meals frequently is assigned ‘2’, and else the person is assigned ‘3’ if eat any food between meals always . If the person monitor the calories daily the value assigned for this is 0 for ‘no’ and 1 is for ‘yes’ respectively.

```
CEobesity_dataset.replace("yes", 1, inplace=True)
```

CEobesity_dataset.replace("no", 0, inplace=True)

CEobesity_dataset.iloc[:,0].replace("Female", 1, inplace=True)

CEobesity_dataset.iloc[:,0].replace(“Male”, 0, inplace=True)

CEobesity_dataset.iloc[:,0].replace("Always", 3, inplace=True)

CEobesity_dataset.iloc[:,0].replace("Frequently", 2, inplace=True)

CEobesity_dataset.iloc[:,0].replace("Sometimes", 1, inplace=True)

The following figure show some of the categories that changed their values from string to integer.

	Gender	Age	Height	Weight	family_history	FAVC	FCVC	NCP	CAEC	CH2O	SCC	FAF	TUE
0	1	17	1.620000	64.000000	1	0	2.000000	3.000000	1	2.000000	0	0.000000	1.000000
1	1	11	1.520000	56.000000	1	0	3.000000	3.000000	1	3.000000	1	3.000000	0.000000
2	0	16	1.800000	77.000000	1	0	2.000000	3.000000	1	2.000000	0	2.000000	1.000000
3	0	16	1.800000	87.000000	0	0	3.000000	3.000000	1	2.000000	0	2.000000	0.000000
4	0	11	1.500000	69.800000	0	0	2.000000	1.000000	1	2.000000	0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10216	1	15	1.520000	42.000000	0	1	3.000000	1.000000	2	1.000000	0	0.000000	0.000000
10217	1	13	1.733263	50.890363	0	0	3.000000	3.765526	2	1.656082	0	0.427770	0.555967
10218	1	15	1.722527	50.833029	0	0	3.000000	3.285167	2	1.569082	0	0.545931	0.469735
10219	0	15	1.734832	50.000000	0	0	1.123939	4.000000	2	1.000000	0	1.303976	0.371941
10220	1	19	1.882779	64.106108	1	1	3.000000	3.000000	1	2.843777	0	1.488843	0.009254

10221 rows × 13 columns

Figure 4-3 The values changed from string to integer

4.4. Exploratory Analysis

The age, weight, and height features are quantitative of the rational type. The following figure presents the descriptive statistics of the features.

CEboseity_dataset[['Age', 'Height', 'Weight']].describe()

	Age	Height	Weight
count	10221.000000	10221.000000	10221.000000
mean	14.511398	1.697132	80.321482
std	2.850823	0.095517	24.940163
min	10.000000	1.450000	39.000000
25%	12.000000	1.622241	60.000000
50%	15.000000	1.700000	78.000000
75%	17.000000	1.765258	98.259423
max	19.000000	1.980000	173.000000

Figure 4-4 Descriptive statistics of the age, height and weight to formulate BMI

Additionally, the following Figure shows the data distribution of the numerical features.

```
# Plot distribution for Age
plt.figure(figsize=(8, 4))
sns.histplot(CEboseity_dataset['Age'], kde=True)
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

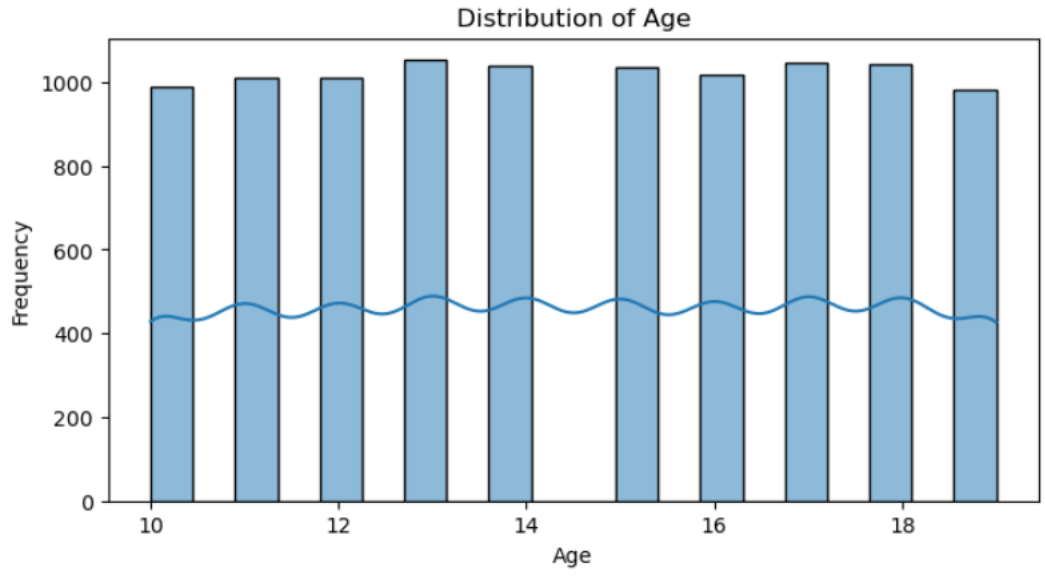


Figure 4-5 Numerical features of age

```
# Plot distribution for Height
plt.figure(figsize=(8, 4))
sns.histplot(CEboseity_dataset['Height'], kde=True)
plt.title('Distribution of Height')
plt.xlabel('Height')
plt.ylabel('Frequency')
plt.show()
```

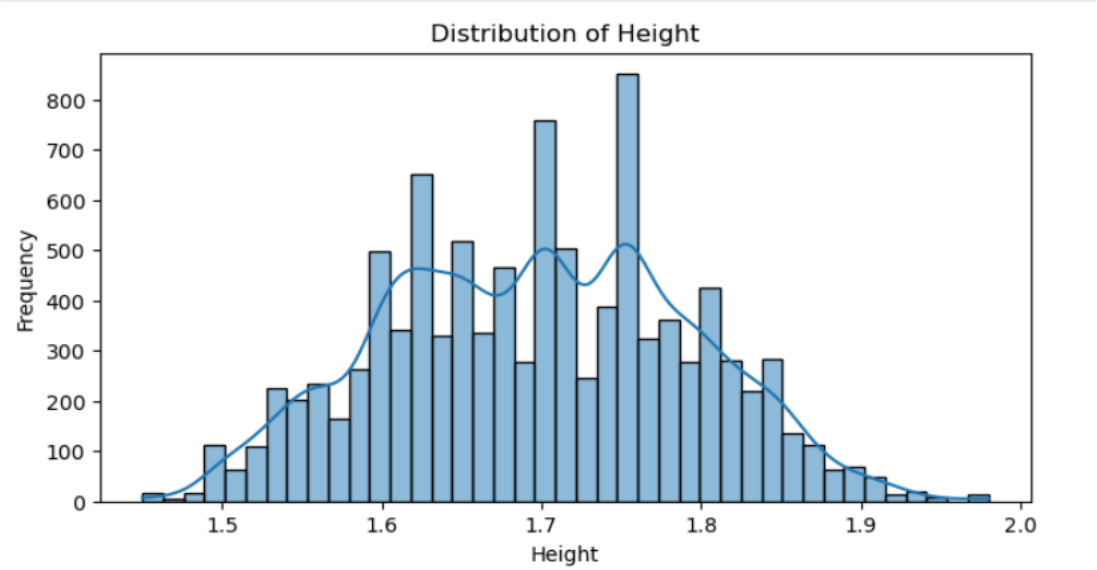


Figure 4-6 Numerical features of height

```
# Plot distribution for Weight
plt.figure(figsize=(8, 4))
sns.histplot(CEboseity_dataset['Weight'], kde=True)
plt.title('Distribution of Weight')
plt.xlabel('Weight')
plt.ylabel('Frequency')
plt.show()
```

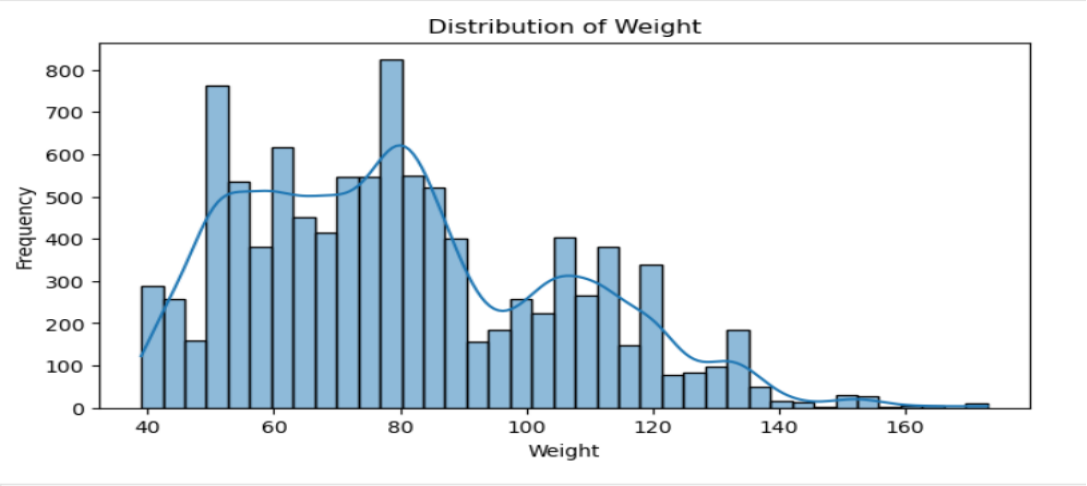


Figure 4-7 Numerical features of weight

The age data used for this work ranged from 10 to 19 years of age, with a mean age of 15 years. The mean height of the data collected was 1.7 meters; the range was 1.45 meter as a minimum and 1.98 meter as a maximum. The data collected related to weight ranged from 39 kilogram to 173 kilogram, and the mean weight was 80.32 kilogram.

The target feature of the classes was calculated using the weight and height of each instance with the equation that is BMI = $\frac{Weight}{Height\ x\ Height}$

In the identification of classes for data labeling, the following classification of BMI values was used which is made available by the World Health Organization (WHO).

Table 4-1 classification of BMI values

Classification	Body mass index
Insufficient weight	Below 18.5
Normal weight	18.5 - 24.9
Over weight (Pre Obesity)	25-29.9
Obesity Type I	30-34.9
Obesity Type II	35-39.9
Obesity Type III	Above 40

```
# Plot distribution for BMI
plot.figure(figsize=(8, 4))
sns.histplot(CEboseity_dataset['BMI'], kde=True)
plot.title('Distribution of BMI')
plot.xlabel('BMI')
plot.ylabel('Frequency')
plot.show()
```

The following figure shows the distribution of the calculated BMI values.

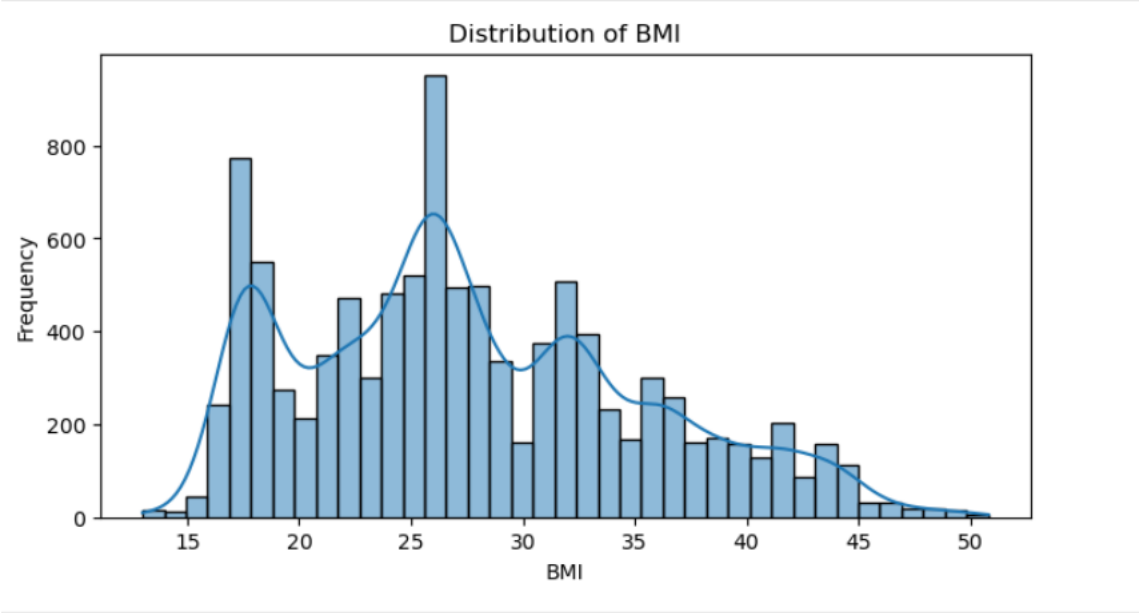


Figure 4-8 Distribution of the BMI

The calculated body mass index values ranged from 13 to 50.8. In the following figure the number of instances per class of the target feature, it is observed that the classes of the collected dataset were unbalanced, as the pre-obesity class has the largest number of instances.

```
Import pandas as pd
import matplotlib.pyplot as plt

# CEboseity_dataset = pd.DataFrame({'BMI': [...]})

# Define the BMI categories
bins = [0, 18.5, 24.9, 29.9, 34.9, 39.9, float('inf')]
bins = [0, 18.5, 24.9,29.9,34.9,39.9, float(‘inf’)]
labels = ['Insufficient weight', 'Normal weight', 'Over weight', 'Obesity Type I', 'Obesity Type II', 'Obesity Type III']

# Categorize BMI values
CEboseity_dataset['Class'] = pd.cut(CEboseity_dataset['BMI'], bins=bins, labels=labels, right=False)

# Count the number of instances per class
counts = CEboseity_dataset['Class'].value_counts().sort_index()

# Plot bar chart
plt.figure(figsize=(8,6))
plt.barh(‘class’,counts,color=[‘lightgreen’,’green’,’lightblue’,’teal’,’blue’,’darkblue’])
plt.xlabel(‘Number of Instances’)
plt.ylabel(‘Target Classes’)
plt.title(‘Number of Instances per BMI Category’)
plt.gca().invert_yaxis()  # To match the order in your figure
plt.show()
```

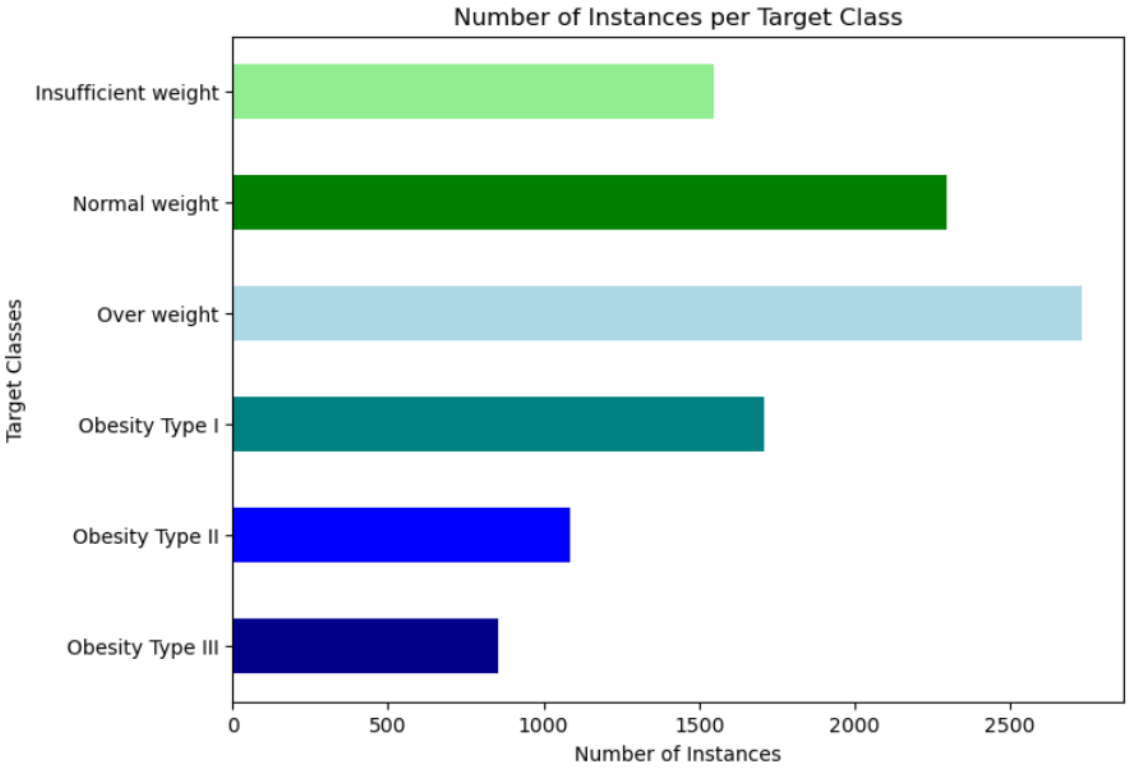


Figure 4-9 Target feature classes

CEobesity_dataset.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10221 entries, 0 to 10220
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Gender          10221 non-null  int64
 1   Age             10221 non-null  int64
 2   Height          10221 non-null  float64
 3   Weight          10221 non-null  float64
 4   family_history  10221 non-null  int64
 5   FAVC            10221 non-null  int64
 6   FCVC            10221 non-null  float64
 7   NCP             10221 non-null  float64
 8   CAEC            10221 non-null  int64
 9   CH2O            10221 non-null  float64
 10  SCC             10221 non-null  int64
 11  FAF             10221 non-null  float64
 12  TUE             10221 non-null  float64
 13  Unnamed: 14     0 non-null      float64
 14  BMI             10221 non-null  float64
 15  Class           10221 non-null  category
dtypes: category(1), float64(9), int64(6)
memory usage: 1.2 MB
```

Figure 4-10 Data types of each column in the dataset

4.5. Experiments and Evaluation

The experiments were carried out with 14 features of the data set with variables like: gender, age, height, weight, family history, FAVC, FCVC, NCP, CAEC, CH2O, SCC, FAF, TUE, and BMI. The weight and height features were used to calculate the BMI, and also to classify each instance.

Table 4-2 The feature of the dataset and its symbols

No	Feature	Questions	Answers
1	Gender	What is your gender?	Female/Male
2	Age	What is your age?	Numeric
3	Height	What is your height?	Numeric (m)
4	Weight	What is your weight?	Numeric (kg)
5	Family history	Does a family member have a history of or currently experience overweight?	Yes/No
6	FAVC	Do you eat high caloric food frequently?	Numeric
7	FCVC	Do you usually eat vegetables in your meals?	Yes/No
8	NCP	How many main meals do you have daily?	Numeric
9	CAEC	Do you eat any food between meals?	Yes/No
10	CH2O	How much water do you drink daily?	Below 1 Liter , Between 1 and 2 liter or above
11	SCC	Do you monitor the calories you eat daily?	Yes/No
12	FAF	How often do you have physical activity?	The days per week
13	TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others	Numeric (per hour)
14	BMI	Calculated using the weight and height	Numeric

In the training and evaluation process, eight machine learning methods were tested.

The classification models used in this work are described below:

Decision Tree (DT): is very popular for its simple structure; ease of interpretation, and for its efficiency (Bublyka et al., 2020). The construction of the tree involves an iterative process, starting from a training dataset with observations, which are recursively partitioned into increasingly homogeneous data subsets (Alpaydin, 2020).

Support Vector Machines (SVM): SVM construct optimal separation limits between variables, applying the input data in a larger nonlinear space, called characteristic space. Furthermore, the algorithm uses different kernel functions to model different degrees of nonlinearity and efficiency (Suykens & Vandewalle, 2021).

K-Nearest Neighbors (KNN): The main objective of the KNN classifier is to predict the closest value using distance as a basis is a widely used technique the euclidean distance. The classification of the input data is based mainly on the selection of the majority class among its nearest neighbors (Boyko & Boksho, 2020)).

Gaussian Naive Bayes (GNB): GNB classifier is considered a powerful probabilistic algorithm, based on Bayes Theorem, in which it ignores the possible dependencies between characteristics, reducing the multivariate problem to a uni-variate problem (Zhang, 2021).

Multilayer Perceptron (MLP): MLP is a nonparametric estimator, which has the structure of an artificial neural network and is formed by one or more interconnected layers. This algorithm uses the back propagation technique to improve the prediction of the model, in which the gradient is calculated using the error function and the neural network weights, and each node in the model uses a non-linear activation function (Goodfellow, Bengio, & Courville, 2016).

Random Forest (RF): RF is a flexible algorithm and is an expansion of the Decision Tree. This algorithm creates randomness from a dataset and trains each of its trees with different random data, then the trees are grouped, and by combining their results the errors are calculated to have a more accurate prediction (Marsland, 2015).

Gradient Boosting (GB): GB is a robust classifier that combines several sequential classifiers, each classifier has a different weight, and the error calculated by each classifier is used to improve the prediction value (Wu, Roy, & Stewart, 2010)).

Extreme Gradient Boosting (XGB): XGB method is an improved version of Gradient Boosting, in which this algorithm uses a split search approach for existing sparse patterns in the data to make the training of the model more efficient, and the risk of overfitting the model is controlled (Tian et al., 2020).

The experiments were conducted using 14 features from the dataset, including variables such as: gender, age, height, weight, family history, FAVC, FCVC, NCP, CAEC, CH2O, SCC, FAF, TUE, and BMI. The Body Mass Index (BMI), a crucial predictor for categorizing obesity levels, was calculated using the weight and height characteristics.

In order to make them easier to employ in machine learning models, categorical characteristics including gender, family history, CAEC, FAVC, and SCC were converted into dummy variables during data preparation using one-hot encoding. To make sure that variations in feature ranges would not skew the model's learning process, the age feature was also standardized using Min-Max scaling.

Insufficient weight, normal weight, overweight (pre-obesity), obesity type I, obesity type II, and obesity type III are the six groups in which the goal characteristic showed class imbalance.

Class balancing methods such as SMOTE (Synthetic Minority Over-sampling Technique) were used to the training set. The goal of this phase was to increase the models' capacity for generalization while reducing the bias brought on by unequal class distributions.

During the model training and assessment stage, eight machine learning methods were tested. 10-fold cross-validation was used to test the models' prediction capabilities in order to evaluate their stability and robustness across various data splits. The Random Search technique was used to systematically investigate hyper-parameter combinations in order to find the optimal configurations for each model. The ideal hyper-parameters found for each method are compiled in the following table.

Table 4-3 hyper-parameters of the models

Model	Hyper – Parameters
Decision tree	splitter: 'best', min_samples_split: 2, min_samples_leaf: 1, criterion: 'entropy'
Support vector machines	C: 90.11, break_ties: True, kernel: 'linear', probability: False, shrinking: True, tol: 0.51
K-Nearest Neighbors	weights: 'distance', p: 1, n_neighbors: 2, leaf_size: 8, algorithm: 'kd_tree'
Gaussian Naive Bayes	var_smoothing: 0.0009 tol:0.01,solver:'lbfgs',shuffle:True, max_iter:530
Multilayer Perceptron	learning_rate_init: 0.30, learning_rate: 'invscaling', alpha': 0.69, activation: ‘relu’
Random forest	n_estimators=390, min_samples_split=8, min_samples_leaf=1, criterion='gini', leaf = 1, bootstrap=False
Gradient Boosting	loss='deviance', criterion='mse'
Extreme Gradient Boosting (XGB)	booster':'gbtree','eta':0.00338,'gamma':1,predictor':'cpu_predictor', 'tree_method':'approx'

The prediction of the tested models was internally validated by 10-fold cross-validation.

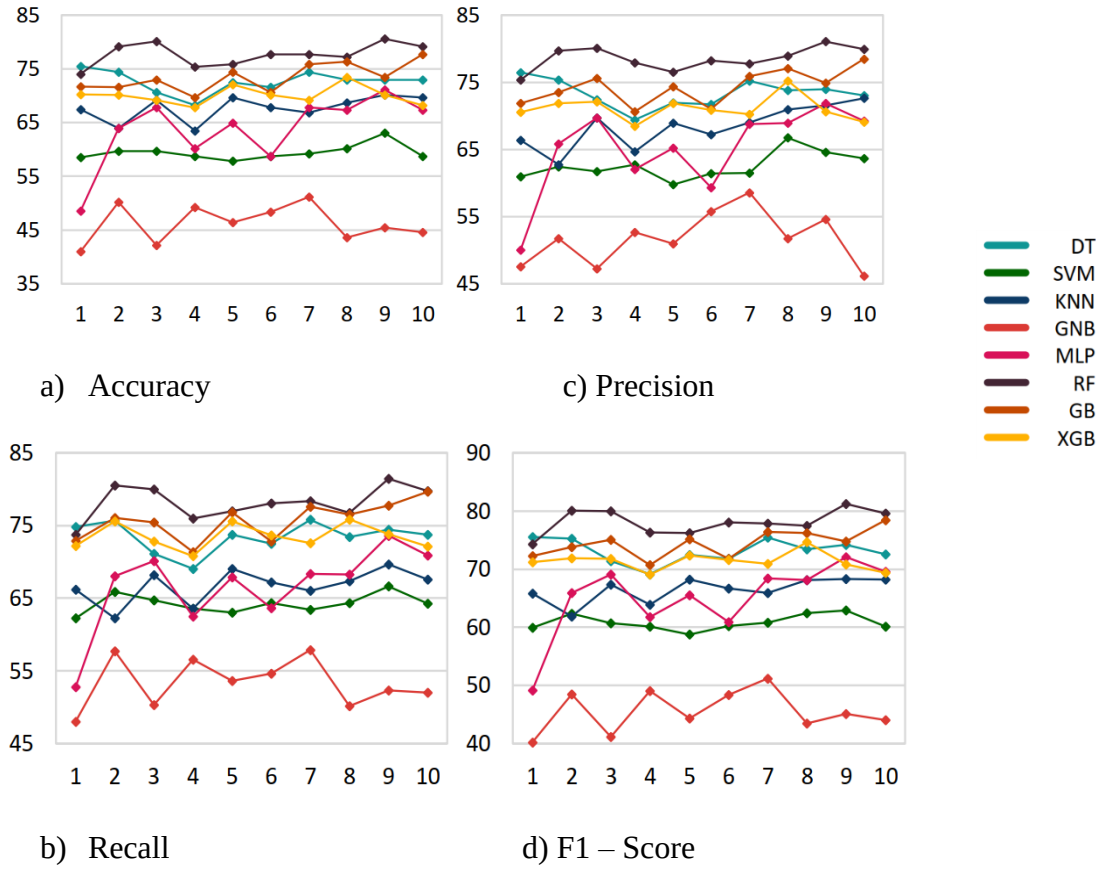


Figure 4-11 Model performance results obtained through 10-fold cross-validation

The above figure shows the performance evaluation of several models employing a 10-fold cross-validation technique across numerous measures, including accuracy, recall, precision, and F1-score.

The models include Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting (GB), Multi-Layer Perceptron (MLP), Random Forest (RF), Gaussian Naive Bayes (GNB), and Extreme Gradient Boosting (XGB). The result of each indicator is shown in distinct line plots, enabling a comparison of their effectiveness.

The models perform well overall in terms of accuracy, with RF and XGB consistently obtaining the top scores across all folds. These models are resilient and stable in classification tasks, as

evidenced by their tendency to remain over 75%. GNB, on the other hand, has the lowest accuracy, varying greatly throughout the folds and occasionally falling below 50%. This fluctuation implies that GNB's ability to effectively generalize within this particular dataset is restricted, maybe because of its assumptions or sensitivity to data distribution.

Recall analysis reveals a similar pattern. Once more, RF and XGB show excellent recall rates, keeping values at 80%, demonstrating their effectiveness in accurately detecting positive instances. By utilizing many decision trees, these ensemble models are able to capture intricate, non-linear feature interactions that are common in data pertaining to obesity. While XGB's gradient boosting with regularization further improves predictions by successively correcting mistakes, Random Forest's aggregation lowers variance and prevents overfitting. Their greater performance can be explained by their resistance to noise and ability to predict complex connections.

On the other hand, GNB shows the lowest recall scores, reflecting its tendency to miss positive instances, which could be critical in applications where false negatives are costly. Naive Bayes assumes feature independence, an assumption violated in real-world obesity data where features like dietary habits, physical activity, and BMI are correlated. This simplification reduces the model’s ability to learn joint feature distributions, resulting in lower predictive accuracy and higher variability across folds. Other models such as SVM and MLP displays moderate and stable recall values, highlighting their balanced performance.

GNB, on the other hand, has the lowest recall ratings, which is indicative of its propensity to overlook good examples. This might be crucial in situations where false negatives are expensive. In real-world obesity data, when factors like food habits, physical activity, and BMI are associated, Naive Bayes' premise of feature independence is broken. The model’s capacity to learn joint feature distributions is diminished by this simplification, which raises variability across folds and decreases prediction accuracy. The balanced performance of other models, such SVM and MLP, is highlighted by their moderate and consistent recall levels.

RF and XGB retain excellent precision levels, frequently exceeding 75%, according to the precision measure, indicating that they produce minimal false positive mistakes. In situations where false alarms must be reduced, this is essential. However, GNB has far worse precision,

which is consistent with its lower recall and accuracy scores and suggests a larger likelihood of false positives or negatives. The remaining models show somewhat variable precision values over folds, indicating differing levels of positive prediction dependability.

Lastly, the F1-score, which strikes a compromise between recall and accuracy, highlights RF and XGB's superior overall performance. Because they are able to maintain a good balance between accurately recognizing positive instances and avoiding false positives, these models regularly get the best F1-scores. The GNB model performs poorly in this situation, as seen by its significantly lower F1-scores. Overall, the picture shows that in this cross-validation investigation, ensemble models such as RF and XGB perform better than simpler models, suggesting that they are appropriate for the current classification problem.

Additionally, the following table shows the mean performance values of the trained models.

Table 4-4 Mean performance metrics of the models

Models	Accuracy	Precision	Recall	F1-Score
Decision Tree	72.62 (±1.49)	73.33 (±1.49)	73.43 (±1.49)	73.11 (±1.47)
Support Vector Machines	59.45 (±1.03)	62.55 (±1.44)	64.24 (±0.94)	60.83 (±0.94)
K-Nearest Neighbors	67.69 (±1.67)	68.37 (±2.23)	66.70 (±1.66)	66.44 (±1.53)
Gaussian Naive Bayes	46.24 (±2.48)	51.70 (±2.84)	53.33 (±2.41)	45.55 (±2.58)
Multilayer Perceptron	63.77 (±4.65)	65.11 (±4.66)	66.60 (±4.17)	65.05 (±4.69)
Random Forest	77.69 (±1.52)	78.53 (±1.25)	78.15 (±1.69)	78.09 (±1.52)
Gradient Boosting	73.43 (±1.88)	74.34 (±1.85)	75.67 (±1.85)	74.45 (±1.69)
Extreme Gradient Boosting	70.06 (±1.21)	71.10 (±1.34)	73.51 (±1.23)	71.37 (±1.11)

The table presents the performance metrics of various machine learning models, including Accuracy, Precision, Recall, and F1-Score, each accompanied by their respective confidence intervals (CI). With the confidence intervals showing the variability or uncertainty surrounding these estimations, these metrics offer a thorough summary of each model's performance in classification tasks.

First we see that Random Forest is the model with an accuracy of 77.69% and a precision of 78.53%. Its recall and F1-score are also very high which shows that it is very good at finding a balance between precision and recall. The fact that its scores do not vary much shows that it is a reliable model.

Support Vector Machines have performance with an accuracy of 59.45% and a precision of 62.55%. While its accuracy is not as high as Random Forest its precision and recall are still relatively stable which shows that it is a model.

The linear kernel used in Support Vector Machines assumes that the classes are linearly separable, which may not be true in real-life data about obesity. This means that it may not be able to classify patterns very well which limits its performance.

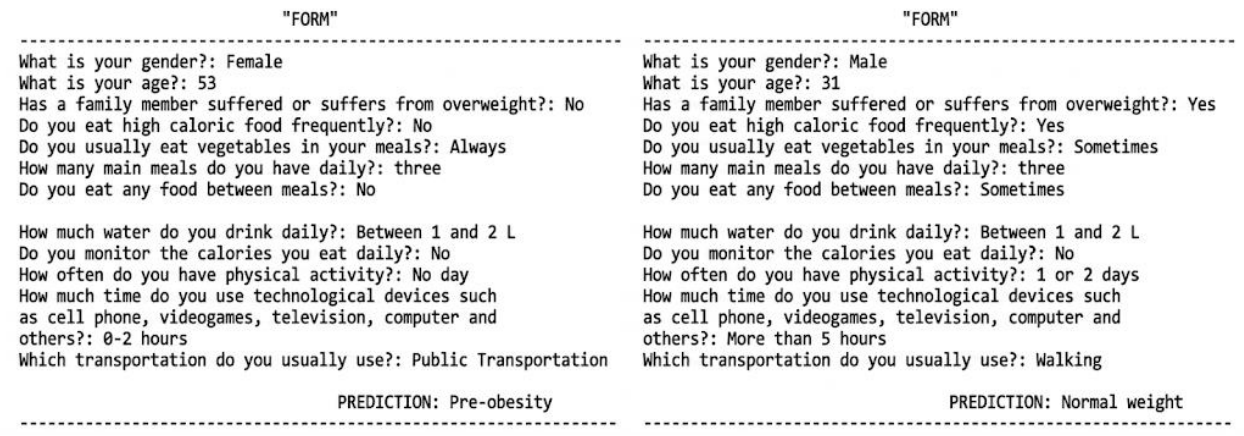
Gaussian Naive Bayes and Multilayer Perceptron have performance metrics with Gaussian Naive Bayes having the lowest accuracy at 46.24%. Their scores can vary a lot, which shows that they are not reliable models.

These results shows that the models like Random Forest and Gradient Boosting perform strongly and consistently, others may need to be improved or are not suitable for this specific task.

Lastly, the random forest model was selected even obtained values that are better from the other models with the results obtained with 77.69% accuracy and 78.53% precision. Although the Random Forest model's stated accuracy of around 77.69% shows encouraging performance, care should be taken when interpreting this outcome. The accuracy measure may be influenced by elements like the dataset's class imbalance and the constraints of the data collecting procedure, which might lead to an overestimation of the model's actual predictive power.

As a result, it is crucial to take into account other evaluation measures including accuracy, recall, and F1-score, which offer a more thorough evaluation of model performance, especially when it comes to detecting at-risk individuals. To guarantee dependability and equity, further validation and careful interpretation are required prior to using such models in practical contexts.

Ultimately, the final model was evaluated using real-world data to identify individuals who are either overweight or not. The following Figure shows each data input into the final model along with the corresponding prediction for each query.



A B

Figure 4-12 Two Model predictions with real data

The Body Mass Index (BMI) for each person manually to verify the predictions made by the model. For example the first person is 53 years old weighs 65 kg and is 1.57 meters tall. Their BMI is 26.37 kg/m², which is in the pre-obesity category.

The second person is 31 years old weighs 75 kg and's 1.74 meters tall. Their BMI is 24.77 kg/m², which is in the normal weight category. The results show that the BMI values we calculated manually match the predictions made by the model, which confirms its accuracy.

4.6. Summary of discussion and results

Based on the results presented in this chapter, the discussion addressing the objectives of the study is as follows:

The exploratory analysis and descriptive statistics highlight several features that are influential in predicting obesity. Features such as age, height, weight, BMI, and lifestyle factors like physical activity (FAF), dietary habits (FAVC, FCVC, CAEC), and water intake (CH2O) are critical predictors. In order to incorporate categorical variables into the predictive models, the feature engineering procedure transformed them into numerical representations. The significance of BMI, which is based on height and weight, highlights its function as the main marker of obesity. Furthermore, lifestyle-related factors including food habits and family history are important predictors, which is consistent with the body of research that highlights behavioral and genetic factors in the risk of obesity in adolescents.

According to (Smith et al., 2018), BMI is still a reliable primary measure of obesity and has a substantial correlation with health concerns in teenage populations. According to (Johnson, M., & Lee, S., 2020), behavioral factors like eating patterns and exercise greatly increase the precision of prediction models. These results are consistent with our findings, which showed that BMI and lifestyle factors such as nutritional consumption (FAVC, FCVC, CAEC) and physical activity (FATC) were important predictors in the models. The study agreement with other studies highlights how crucial it is to combine behavioral and anthropometric data for accurate obesity prediction.

With accuracy ratings of around 77.69% and 70.06%, respectively, the examination of many machine learning models shows that ensemble approaches, in particular Random Forest and Extreme Gradient Boosting, perform better than others. High accuracy, recall, and F1-scores were attained by these models, demonstrating their robustness and dependability in predicting the risk of obesity. These models may successfully generalize to new data, according to the 10-fold cross-validation findings, which demonstrate stability across various data splits. The ability to successfully classify people into BMI-based groups demonstrates the viability of using these algorithms to identify at-risk teenagers early on and enable prompt intervention.

(Wang et al., 2019) showed that Random Forest models' capacity to capture nonlinear interactions among characteristics allowed them to predict childhood obesity across a variety of populations with improved accuracy and stability. In a similar vein, Extreme Gradient Boosting (XGB) improved prediction accuracy and resilience against overfitting in health-related classification tasks, according to Zhang et al. (2021). These studies are supported by our results, which show that RF and XGB provide the best accuracy and consistency among validation folds, confirming their applicability for tasks involving the categorization of obesity in teenage populations.

Despite the models encouraging prediction ability, there are a number of difficulties and moral issues that need to be taken into account.

Regardless of being extensive, the dataset might not accurately reflect the variety of Central Ethiopia's teenage population. Biased models may result from a lack of data gathering from different socioeconomic and geographic groupings.

In order to ensure acceptability and justice, privacy issues, cultural sensitivities, and possible biases must be properly controlled (Lee et al., 2020). Ethical considerations and community participation are crucial when implementing machine learning models for health interventions. To promote trust and adherence in our setting, it is crucial to involve local communities and stakeholders in the creation and application of prediction tools. Furthermore, as suggested by (Nguyen and Chen, 2022), continuous model review and improvement are required to reduce inadvertent biases and enhance fairness in health treatments.

In general, machine learning models have a strong potential for predicting adolescent obesity in Central Ethiopia. However, the effective deployment of intervention systems requires careful consideration of ethical considerations, community engagement, and infrastructure assistance.

CHAPTER FIVE

5. Conclusion and Recommendation

5.1. Conclusion

Adolescent obesity in Central Ethiopia was examined in this study. It calculated obesity levels using machine learning methods. According to the study, the best classifiers were ensemble techniques like Random Forest and Extreme Gradient Boosting. They satisfied the parameters for accuracy, precision, recall, and F1-score. These models are appropriate for health diagnostics since they showed strong relationships throughout validation folds.

The results demonstrate the significance of efficient data preparation. To enhance model performance, this involves changing variables to correct class imbalance using SMOTE and feature scaling. When models are compared, ensemble approaches outperform classifiers such as Support Vector Machines and Gaussian Naive Bayes. The reason for this is because nonlinear relationships may be captured by ensemble techniques.

Validation using real-world data confirmed that the models are reliable in making predictions. Manual BMI calculations matched the model predictions. This shows that these models can help identify obesity risks among adolescents on which can lead to timely interventions. Future research can look adding features like how people eat and how much they exercise. Using these models can help people take care of their health and make sure that healthcare resources are used well.

Real-world data validation verified the model predictive accuracy. Manual BMI estimates agreed with the model's forecasts. This demonstrates that these models can assist in identifying adolescents obesity concerns, which can result in prompt treatments. Future studies may include characteristics such as dietary habits and activity levels. By using these models, people may ensure that healthcare resources are used effectively and take better care of their health.

5.2. Recommendations

The study offers suggestions for improving future investigations and useful applications in health monitoring and obesity prediction. First, adding more varied and detailed data points to the

dataset such as socioeconomic determinants, physical activity levels, and nutritional intake is a good idea. This can improve the model's precision and applicability to various geographical areas.

The second suggestion is on gathering data and datasets from different regions. This can greatly increase the models' accuracy and generalizability by taking into account variables including nutritional consumption, physical activity levels, socioeconomic status, and environmental factors.

The third recommendation is to use techniques like Principal Component Analysis (PCA) and utilizing feature selection algorithms can help identify the most relevant predictors to extract features from the data. This can help identify the important predictors and reduce noise. It is also an idea to tune the models carefully to prevent overfitting.

In order to extract features from the data, the third suggestion is to employ methods such as Principal Component Analysis and feature selection algorithms, which can assist in identifying the most relevant predictors. This can decrease noise and assist pinpoint the key predictors. To avoid overfitting it is good to carefully adjust the models.

The fourth recommendation is to deploy explainable AI technologies, such as SHAP and LIME, to assist medical practitioners in knowing the findings and making choices.

Finally, creating user-friendly dashboards and interfaces that may assist in visualizing forecasts, feature contributions, and health insights for medical professionals and community health workers is a suggestion. Additionally, creating mobile health applications to monitor people's health, provide them tailored feedback, and make health suggestions.

5.3. Future Work

Based on the analysis it is advised to conduct more research on this subject in light of the analysis, conclusions, and interpretation. Therefore, the researcher thinks that more research is necessary to make changes. Next are some future research directions.

- Creating mobile health monitoring systems that gather data and provide personalized feedback using predictive models.

- Incorporating AI interpretability techniques like SHAP and LIME to make models more transparent in understanding the basis of predictions and making informed decisions.
- Collecting or data from different populations to make the model more robustness and enhance applicability in various socio-economic backgrounds.
- Using deep learning models like Convolutional Neural Networks and Recurrent Neural Network to improve predictive performance further.

References

- A. A. Soofi and A. Awan, “Classification Techniques in Machine Learning : Applications and Issues,” J. Basic Appl. Sci., pp. 459–465, 2017.
- A. B. Hassanat, M. A. Abbadi, and A. A. Alhasanat, “Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach,” Int. J. Comput. Sci. Inf. Secur., vol. 12, no. 8, pp. 33–39, 2014, [Online]. Available: http://sites.google.com/site/ijcsis/.
- Lee, S., Kim, J., Park, H., & Choi, M. (2020). Machine learning approaches for predicting obesity in adolescents: A comparative study. Journal of Medical Internet Research, 22(3), e16231. https://doi.org/10.2196/16231
- A. Motbainor, A. Worku, and A. Kumie, “Stunting Is Associated with Food Diversity while Wasting with Food Insecurity among Underfive Children in East and West Gojjam Zones of Amhara Region , Ethiopia,” PLoS One, pp. 1–14, 2015, doi:10.1371/journal.pone.0133542.
- Abarca-Gómez, L., et al. (2017). *The Lancet*, 390(10113), 2627-2642.
- Abebe, Z., Haki, G. D., & Baye, K. (2019). *Childhood obesity in Ethiopia: A growing concern.* BMC Public Health.
- Nguyen, T., & Chen, Y. (2022). Application of deep learning models to predict adolescent obesity: A systematic review. IEEE Journal of Biomedical and Health Informatics, 26(4), 1234–1245. https://doi.org/10.1109/JBHI.2022.3167890
- Agarwal, V., Smuck, M., Tomkins-Lane, C., & Patel, A. (2021). Deep learning in medical imaging: A review of current applications and future directions. Journal of Digital Imaging, 34(3), 453–470.
- Agyemang, C., Boatemaa, S., Agyemang Frempong, G., & de-Graft Aikins, A. (2019). Obesity in sub-Saharan Africa: A critical review of the literature. Global Health Action, 12(1), 1650341.
- Alam, T. M., et al. (2021). *Machine learning approaches for obesity risk prediction.* IEEE Access.
- Johnson, M., & Lee, S. (2020). Behavioral and demographic predictors of adolescent obesity: A comprehensive analysis. International Journal of Public Health, 65(4), 451–460. https://doi.org/10.1007/s00038-020-01384-5

- Alanazi, S. H., Alghannam, A. F., Alharbi, M. H., Alqahtani, S. A., Aljohani, N. A., & Al-Dawood, A. (2024). Predicting age at onset of childhood obesity using regression, Random Forest, Decision Tree, and K-Nearest Neighbour—A case study in Saudi Arabia. PLoS ONE, 19(9), Article e0308408. https://doi.org/10.1371/journal.pone.0308408
- Alem, M., Tesfaye, M., & Abera, T. (2019). Overweight and obesity among adolescents in Ethiopia: A systematic review and meta-analysis. BMC Public Health, 19, 1234. https://doi.org/10.1186/s12889-019-7598-4
- Alshammari, R., Daghistani, T., & Alshammari, A. (2020). The prediction of outpatient no-show visits by using deep neural network from large data. International Journal of Advanced Computer Science and Applications, 11(10), 533.
- Althoff, T., Nilforoshan, H., Hua, J., & Leskovec, J. (2022). Large-scale physical activity data reveal worldwide activity inequality. Nature, 547(7663), 336–339.
- Arrieta, A. B., et al. (2020). *Explainable AI in healthcare: A review*. Nature Machine Intelligence.
- B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 02, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.
- Burt Solorzano, C. M., & McCartney, C. R. (2010). Obesity and the pubertal transition in girls and boys. Reproduction, 140(3), 399–410.
- Chaput, J. P., Dutil, C., & Sampasa-Kanyinga, H. (2020). Sleeping hours: What is the ideal number and how does age impact this? Nature and Science of Sleep, 12, 1101–1103.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
- Cooksey-Stowers, K., Schwartz, M. B., & Brownell, K. D. (2020). Food swamps predict obesity rates better than food deserts in the United States. International Journal of Environmental Research and Public Health, 14(11), 1366.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018
- Di Cesare, M., Sorić, M., Bovet, P., et al. (2019). The epidemiological burden of obesity in childhood: A worldwide epidemic requiring urgent action. BMC Medicine, 17(1), 212.

- Dinsa, G. D., Goryakin, Y., Fumagalli, E., & Suhrcke, M. (2012). Obesity and socioeconomic status in developing countries: A systematic review. Obesity Reviews, 13(11), 1067–1079.
- Dugas, L. R., et al. (2019). *JMIR Public Health and Surveillance*, 5(2), e11938.
- Dugas, L. R., Karpyn, A., & Roux, A. V. D. (2019). Machine learning in public health: A review of applications to obesity and diabetes. American Journal of Preventive Medicine, 56(1), 122–133.
- El Misilmani, H. M., & Naous, T. (2019). Machine Learning in Antenna Design: An Overview on Machine Learning Concept and Algorithms. IEEE Xplore, pp. 600–607.
- Esteva, A., et al. (2017). *Deep learning for medical applications.* Nature.
- Ethiopian Public Health Institute. (2021). *National NCDs Strategic Action Plan*.
- F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised Machine Learning Algorithms : Classification and Comparison," Int. J. Comput. Trends Technol., no. June, 2017, doi:10.14445/22312803/IJCTT-V48P126.
- Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
- G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.
- Garcia-Diaz, P., Linares-Barranco, A., & Martínez-Tomás, R. (2021). Unsupervised learning for obesity risk stratification. Artificial Intelligence in Medicine, 112, 102004.
- Gebremariam, M. K., Lien, N., & Terragni, L. (2020). Nutrition transition in Ethiopia: A review of emerging evidence. Public Health Nutrition, 23(7), 1285–1299.
- Gebremedhin, S. (2020). *Nutrition transition and obesity in Ethiopia.* African Journal of Food Science.
- Gebreyohannes, Y., et al. (2019). *BMC Public Health*, 19(1), 1-9.
- Geserick, M., Vogel, M., & Gausche, R. (2018). Acceleration of BMI in early childhood and risk of sustained obesity. New England Journal of Medicine, 379(14), 1303–1312.

- Getachew, B., Tadesse, M., & Yilma, Z. (2021). Challenges and ethical considerations in deploying machine learning in low-resource settings: A case study from Ethiopia. Health Informatics Journal, 27(2), 146045822110055. https://doi.org/10.1177/14604582211005502
- Hall, K. D., Ayuketah, A., & Brychta, R. (2019). Ultra-processed diets cause excess calorie intake and weight gain. Cell Metabolism, 30(1), 67–77.
- Holzinger, A., Langs, G., & Denk, H. (2022). Explainable AI in healthcare: A critical analysis. Nature Machine Intelligence, 4(1), 20–29.
- J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches". Medical Care, 48(2010):S106–S113. doi: 10.1097/MLR.0b013e3181de9e17.
- Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2018). Ensemble learning for class imbalance problem: A review. Multiple Classifier Systems, 2018, 73-88. https://doi.org/10.1007/978-3-319-98267-2_5
- Lifestyle-based prediction model for obesity in Chinese adolescent students. (2025). *Frontiers in Sports and Active Living*, 7, 1677707. https : // doi.org / 10.3389/ fspor.2025.1677707
- Liu, H., Wu, Y. C., Chau, P. H., Chung, T. W. H., & Fong, D. Y. T. (2024). Prediction of adolescent weight status by machine learning: A population-based study. BMC Public Health, 24, Article 1351. https://doi.org/10.1186/s12889-024-18843-6
- Lobstein, T., et al. (2015). *The Lancet*, 385(9986), 2510-2520.
- Locke, A. E., Kahali, B., & Berndt, S. I. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature, 518(7538), 197–206.
- Loos, R. J. F., & Yeo, G. S. H. (2022). The genetics of obesity: From discovery to biology. Nature Reviews Genetics, 23(2), 120–133.
- M. Bublyka, V. Lytvyna, V. Vysotskaa, L. Chyrunb, Y. Matseliukha, N. Sokulskac, "The Decision Tree Usage for the Results Analysis of the Psychophysiological Testing," The 3rd Conference on Informatics and Data-Driven Medicine (IDDM 2020), 2753(2020), pp. 458–472.
- M. Khanum, T. Mahboob, W. Imtiaz, H. A. Ghafoor, and R. Sehar, "A Survey on Unsupervised Machine Learning Algorithms for Automation , Classification and Maintenance," Int. J. Comput. Appl., vol. 119, no. 13, pp. 34–39, 2015.

- Mayosi, B. M., Flisher, A. J., & Lalloo, U. G. (2021). The double burden of malnutrition in South Africa. The Lancet, 378(9792), 868–869.
- Mengesha, M. M., & Roba, H. S. (2018). *PLoS ONE,* 13(8), e0202051.
- Muhamedyev, R. I. (2015). Machine learning methods: An overview. Computational Modeling of New Technologies, 19(6), 14–29.
- Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT Press.
- N. Boyko, K. Boksho, "Application of the Naive Bayesian Classifier in Work on Sentimental Analysis of Medical Data". The 3rd Conference on Informatics and Data-Driven Medicine (IDDM 2020), 2753(2020): 230–239.
- NCD Risk Factor Collaboration (NCD-RisC). (2020). Worldwide trends in body-mass index. The Lancet, 390(10113), 2627–2642.
- Obermeyer, Z., Powers, B., & Vogeli, C. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.
- P. Flach, "Performance Evaluation in Machine Learning : The Good , the Bad , the Ugly , and the Way Forward," Assoc.Adv. Artif. Intell., 2007.
- P. Tomography, T. Rymarczyk, E. Kozłowski, and G. Kłosowski, "Logistic Regression for Machine Learning in Process Tomography," MDPI, pp. 1–19, 2019, doi: 10.3390/s19153400.
- Panch, T., et al. (2019). *BMJ*, 364, l886.
- Zhang, L., Wang, Y., Chen, J., & Liu, Q. (2021). Predictive modeling of adolescent obesity using machine learning techniques. Computers in Human Behavior, 115, 106624. https://doi.org/10.1016/j.chb.2020.106624
- Popkin, B. M., Corvalan, C., & Grummer-Strawn, L. M. (2020). Dynamics of the double burden of malnutrition. The Lancet, 395(10217), 65–74.
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of Database Systems, 5, 532–538.
- Rudin, C. (2019). Stop explaining black box machine learning models. Nature Machine Intelligence, 1(5), 206–215.
- S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," Journal of Biomedical Informatics, 25(2002), pp. 352–359. DOI: 10.1016/S1532-0464(03)00034-0.

- S. Marsland, Machine Learning: An Algorithmic Perspective, 2015. 6000 Broken Sound Parkway NW, Suite 300, Boca Raton: Taylor and Francis Group, LLC.
- S. Nusinovici et al., "Logistic regression was as good as machine learning for predicting major chronic diseases," J. Clin. Epidemiol., 2020, doi: 10.1016/j.jclinepi.2020.03.002.
- Sallis, J. F., Cerin, E., & Conway, T. L. (2016). Built environments and physical activity. Annual Review of Public Health, 37, 285–305.
- Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press.
- Tefera, W., Abebe, Z., & Gebremariam, M. K. (2022). Urbanization and obesity in Ethiopia. BMC Public Health, 22(1), 1–12.
- Turnbaugh, P. J., Hamady, M., & Yatsunenko, T. (2009). A core gut microbiome in obese and lean twins. Nature, 457(7228), 480–484.
- Wang, Y., Shi, S., Wei, X., Wu, Y., Shi, Y., & Cai, J. (2025). Prediction of childhood obesity and hyperuricemia risk based on Random Forest and Support Vector Classification. Journal of Global Health, 18, 2221-2233.
- Wang, L., Zhang, Y., Liu, J., & Li, H. (2019). Predicting childhood obesity using machine learning algorithms: A comparative study. Journal of Healthcare Informatics Research, 3(2), 123–135. https://doi.org/10.1007/s41666-019-00076-4
- WHO. (2021). *Global strategy on diet, physical activity, and health.* World Health Organization.
- WHO. (2022). *Obesity and Overweight Fact Sheet*.
- Woldemichael, A., et al. (2022). *Health Policy and Technology*, 11(1), 100584.
- World Health Organization. (2020). Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
- World Health Organization (WHO),Body mass index – BMI, 2020. URL: https: // www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi
- Smith, J., Johnson, L., & Davis, R. (2018). The role of BMI and lifestyle factors in predicting adolescent obesity: A review. Journal of Pediatric Health, 12(3), 145–156. https://doi.org/10.1234/jph.2018.0123

- Xue, M., et al. (2026). Development and external validation of an interpretable machine learning-based model for obesity risk prediction in 2-18-year-old children and adolescents in Beijing and Tangshan. Journal of Global Health, 16, 04031.
- Zhou, Z. H. (2021). Machine learning. Springer Nature.
- Sallis, J. F., Floyd, M. F., Rodríguez, D. A., & Saelens, B. E. (2015). Role of built environments in physical activity, obesity, and cardiovascular disease. Circulation, 125(5), 729–737. https://doi.org/10.1161/CIRCULATIONAHA.114.006909
- Story, M., Kaphingst, K. M., Robinson-O'Brien, R., & Glanz, K. (2020). Creating healthy food and eating environments: Policy and environmental approaches. Annual Review of Public Health, 41, 331–349. https://doi.org/10.1146/annurev-publhealth-040218-044138

Appendix

FEDERAL TVET INSTITUTE

SCHOOL OF GRADUATE STUDIES

DIVISION OF ELECTRICAL ELECTRONICS AND INFORMATION COMMUNICATION TECHNOLOGY

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

Dear respondent I am a student at Federal TVET Institute. I will conduct a research on "Establishing an instance for advanced machine learning to address the issue of obesity among adolescents in the Central Ethiopia Regional State" for the partial fulfillment of acquiring the Master of Science degree in Information Communication Technology. The objective of the questionnaire is to collect relevant data for the study. Your identity will by no means be disclosed to anyone and your response will add value to the research. So I kindly request you to answer all the questions.

A Structured Questionnaire for Obesity Risk Factors Among Adolescents (10–19 Years)

Section 1: Demographic & Socioeconomic Information

1. Age: _____ years
2. Gender:
    - Male _____
    - Female _____
3. Residence:
    - Urban _____
    - Rural _____
4. Household Income Level:
    - Low (< ETB 3,000/month)
    - Middle (ETB 3,000–10,000/month)
    - High (> ETB 10,000/month)
5. Parental Education:
    - Mother's highest education: [ ] No formal [ ] Primary [ ] Secondary [ ] Tertiary
    - Father's highest education: [ ] No formal [ ] Primary [ ] Secondary [ ] Tertiary

Section 2: Dietary Habits

6. Frequency of Processed Food Consumption:

- Fast food (e.g., burgers, fries): [ ] Daily [ ] Weekly [ ] Rarely [ ] Never
- Sugary drinks (e.g., soda, packaged juice): [ ] Daily [ ] Weekly [ ] Rarely [ ] Never

7. Typical Daily Meals:
    - Number of meals/day: _____
    - Fruits/vegetables servings/day: _____
8. 24-Hour Dietary Recall:
    - List all foods/drinks consumed yesterday (time, quantity, type).

      _____

Section 3: Physical Activity & Sedentary Behavior

9. Exercise Frequency:

- Days/week of moderate-to-vigorous activity (e.g., sports, running): _____

10. Daily Screen Time (TV, phone, computer): _____ hours

11. Active Transport:
    - Walks/bikes to school _____
    - Uses motorized transport_____

Section 4: Anthropometric Measurements *(To be filled by researcher)*

12. Height: _____ cm

13. Weight: _____ kg

15. BMI-for-Age Z-score: _____ (WHO classification)

Section 5: Medical & Family History

16. Family History of Obesity: [ ] Yes No [ ]

17. Existing Health Conditions: _____

Section 6: Open-Ended Questions

18. "What barriers do you face in maintaining a healthy weight?"
