

A Feature Selection Approach To Determine Obesity Using Machine Learning Method

By

Shahidul Alam Antor

17101242

Jawad Ahmed

17101336

Zulker Nayen

17101273

Farisha Tabassum

17101154

Rasheda Mahbub

17301077

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
June 2021

© 2021. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis presented here is our original work during the degree at the Brac University.
2. No material accepted or submitted for any other degree or diploma at a university or other institution can be contained in the thesis.
3. The thesis does not include material previously published or written by third parties unless complete and accurate references adequately cite it.
4. We recognized all the primary sources of assistance.

Student's Full Name & Signature:

Jawad Ahmed

Jawad Ahmed
17101336

Shahidul

Shahidul Alam Antor
17101242

Rasheda Mahbub

Rasheda Mahbub
17301077

Zulker

Zulker Nayen
17101273

Farisha Tabassum

Farisha Tabassum
17101154

Approval

The thesis/project titled “A Feature Selection Approach To Determine Obesity Using Machine Learning Method” submitted by

1. Shahidul Alam Antor (17101242)
2. Jawad Ahmed (17101336)
3. Zulker Nayen (17101273)
4. Farisha Tabassum (17101154)
5. Rasheda Mahbub (17301077)

Of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 10, 2021.

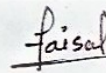
Examining Committee:

Supervisor:
(Member)



Rasif Ajwad
Lecturer
Department of Computer Science and Engineering
BRAC University

Co Supervisor:
(Member)



Faisal Bin Ashraf
Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)



Sadia Hamid Kazi
Associate Professor and Chairperson
Department of Computer Science and Engineering
BRAC University

Ethics Statement (Optional)

We have read different journals and papers and gathered information via Google form for our study purpose. We are committed to maintaining the anonymity of all the respondent's names. These data are only used for research purposes.

Abstract

The new health concern that is proliferating in developing and impoverished countries is obesity. It is recognized as a complex health issue caused by various factors such as genetics, behaviour, and other issues. Obesity is not just about physique or look; it is a persistent medical illness that opens the body to many diseases and shortens life. Obesity frequently results in a wide variety of other disorders, including cardiovascular disease, hypertension, diabetes, numerous malignancies, and more. The developed countries have already undertaken a few measures and are deeply concerned about their health issues. Thus, the people of low or mid-income countries are still unaware of this fact and will face significant health challenges in the future. Specifically, in Bangladesh, many people have diabetes, and recently, many people died due to heart disease and cancer, which could be prevented if they were health concerns. Recent studies say that the young generation is more prone to obesity as they are more influenced by western lifestyles, eating many junk foods, and spending the maximum of their time on the internet. Our research has collected more than 500 people's data from different groups of people around Bangladesh. We aim to predict the future outcome at which BMI value range people are more prone to diseases. To predict the outcome, we have analyzed our sample dataset using machine learning approaches such as Naive Bayes, Random Forest, decision tree, The k-nearest neighbours (KNN), Logistic Regression. Among these algorithms, Decision Tree has given the best accuracy of 96.67%. For selecting essential variables from the dataset, we used the BorutaShap wrapper feature selection method. This algorithm delivers a better subset of attributes from a high volume of data and trains the model faster. As the Boruta algorithm selects the best feature, reduces the model size, and identifies the key features, it became easy to train our data set, so we got a better accuracy level using this algorithm in our reach. This researcher will help the people of Bangladesh to understand obesity and its detrimental aspects. Moreover, it will assist them in being more conscious of their health conditions and predicting which BMI level is a risk for them.

Keywords: Obesity, BMI, Machine Learning, Naive Bayes, Random Forest, Decision tree, K-Nearest Neighbors (KNN), Logistic Regression

Dedication (Optional)

Each difficult task necessitates both self-effort and encouragement from elders, particularly those who were particularly dear to our hearts. We dedicate our humble efforts to our loving parents, whose affection, devotion, motivation, and prayer throughout the day and night make us deserving of such achievement and honor, as do all the dedicated and respected Teachers.

Acknowledgement

First, we express our gratitude to Allah for His favors, which have enabled us to continue our research without encountering serious obstacles. Additionally, we wanted to express our gratitude to all helpful faculty members, particularly our supervisor, for tolerating our errors and providing continual feedback to help us improve our research. Also, we want to express our gratitude to our parents and teammates for their unwavering support throughout the semester.

Index

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	v
Dedication	vi
Acknowledgment	vii
Table of Contents	viii
List of Figures	x
List of Tables	xi
Nomenclature	xiii
1 Introduction	1
1.1 Motivation	2
1.2 Problem Statement	3
1.3 Research Objective	4
1.4 Thesis Outline	5
2 Related Work	6
2.1 Literature Review	6
3 Data Collection and Feature Selection	21
3.1 Data Collection	21
3.1.1 Data Pre Processing	25
3.2 Feature Selection	27
3.2.1 Borutashap Feature Extraction	27
3.3 Feature Analysis	29
3.3.1 Heatmap of Data	34
3.3.2 Approach for Feature Selection	36

4	Model Selection and Result Analysis	39
4.1	Machine Learning	39
4.2	Supervised Learning	40
4.3	Naive Bayes Algorithm Implementation	41
4.4	Random Forest Model Implementation	42
4.5	KNN Algorithm Implementation	43
4.6	Logistic Regression Implementation	44
4.7	Decision Tree Implementation	45
4.8	Results and Analysis	46
5	Conclusion and Future Work	50
	Bibliography	53
	Appendix A Appendix	54
A.1	BorutaShap Code Sample	54
A.2	Comments from the panel members:	54

List of Figures

3.1	System module for ML-based approach	22
3.2	Variation of obesity due to age	23
3.3	Non-scaled Dataset	24
3.4	Scaled Dataset	24
3.5	Datatype	26
3.6	Datatype	27
3.7	Feature Extraction using BorutaShap	28
3.8	Number and Percentage of Weight	29
3.9	Meal Consumption	30
3.10	Smoking Percentage	30
3.11	Alcohol Consumption	31
3.12	Calorie Consumption	32
3.13	Water Consumption	32
3.14	Electronic device use	32
3.15	Comparison of Obesity vs Age, Height, Weight	33
3.16	Calorie consumption monitoring	33
3.17	Consumption of Meals	34
3.18	Correlation of Data	34
3.19	Heatmap of Dataset	35
3.20	Correlation of Comparing data	35
3.21	BorutaShap Feature Extraction	36
3.22	Feature Importance of our data	37
3.23	Feature Importance of the other dataset	38
4.1	Supervised learning	40
4.2	Random Forest Structure	42
4.3	Visualization of the formula	43
4.4	Logistic Regression	44
4.5	Decision tree	45
4.6	Decision tree accuracy prediction	46
4.7	Accuracy Chart	47
4.8	Confusion Matrix of Decision tree	48
4.9	Confusion Matrix of k-nearest neighbor	48
4.10	Confusion Matrix of Naive Bayes	49
4.11	Confusion Matrix of Random Forest	49

List of Tables

3.1 Feature Names	23
-----------------------------	----

Nomenclature

The following list describes several symbols & abbreviations that will later be used within the document body.

AAP American Academy of Pediatrics

ADA Americans with Disabilities Act

BMI Body Mass Index

CI confidence Interval

CKD Chronic Kidney Disease

CNN Convolutional Neural Network

CRA Colorectal Adenoma

CRC Colorectal Cancer

ED Erectile Dysfunction

FNPA The Family Nutrition and Physical Activity

HDL High-Density LipoProtein

KNN K-Nearest Neighbor

LDL Low-Density LipoProtein

LR Logistic Regression

LTPA Leisure Time Physical Activity

MSA Metropolitan Statistical Area

NHANES National Health and Nutrition Examination Survey

NSES National Science Education Standard

OR Odds Ratio

PA Physical Activity

PCA Principal Component Analysis

PCOD Polycystic ovary syndrome

SDS Standard Deviation Score

SMOTE Synthetic Minority Oversampling Technique

WHO World Health Organization

Chapter 1

Introduction

Although there are many different types of obesity, they are marked by a significant increase in body fat percentage. In addition, being overweight is defined as having a BMI over 25. Today, it is one-third of the world's population suffering from this severe health condition. The obesity rate has reached epidemic proportions and is no longer contained. A person who is obese will have a life that is not happy. There are numerous health risks tied to obesity, yet gaining weight is not the root of the problem. Many more it is possible to say that when we deal with obesity, our overall health will begin to decline. We must increase public understanding about the ways to prevent it. Nevertheless, on the other hand, many governments have already taken action to address this problem in countries at the beginning of their development and developing countries that still lag. For example, countries like Bangladesh, India, Pakistan, Afghanistan, etc., are yet to learn. Because of this, individuals in Bangladesh want to eat things that injure their health and have higher caloric values, such as fried foods or fast food. The increase in obesity is a result of this. It is necessary to get additional information and statistics to help educate the population about the impact of obesity on our health. People have the option of controlling or avoiding obesity if they can identify it. In our research, we examine our dataset using machine learning to find the root causes of obesity and determine how we might prevent future instances of it. In order to discover the primary cause, we correlated the data. BorutaShap wrapper feature selection approach, in addition to this, we also use the feature selection method to compare the data and identify differences. For each feature, various machine learning methods were employed to obtain accuracy. Using the decision tree algorithm, we found that it had the best accuracy. So our research may conclude that the major cause of obesity is known, and these are the characteristics of the cause. It will provide a simple explanation of the issue that caused this major problem.

1.1 Motivation

The main objective of our study is to trace whether an individual is obese or not via a machine learning approach. Obesity has become a matter of concern today. A connection exists between obesity and a range of chronic ailments, including heart disease, type 2 diabetes (high blood sugar), high blood pressure, many malignancies, and others. For our study purpose, we have gathered information through a google form. The dataset consists of 17 features from which the output is generated. Our questionnaires were based on height, weight, food habits, lifestyle, family history of obesity, food choosing, calorie intake, physical activities to identify a person as obsessed. If a person is diagnosed with obesity at the initial stage, it can be cured through proper medication rather than being a sufferer in future. Some diseases get worse for the obsessed patients. That is why it is essential to identify obesity to treat present disease or to avoid future sufferings by being concerned about It. Thus, we figured out the crucial features which can give a decision on obesity and create awareness.

1.2 Problem Statement

Obesity is a critical health condition that is a result of excess body fat. People who have higher BMI (body mass index) are called obese. Unhealthy lifestyles, eating habits, fewer body movements, genetics, food choices, less physical activity are the root cause of obesity. Obesity can negatively affect our health, increasing the chance of developing coronary heart disease, stroke, high blood pressure, type 2 diabetes, high LDL cholesterol, low HDL cholesterol, and osteoarthritis, among other medical issues. Nowadays, most obesity is vastly noticed in the young stars of our country as they are addicted to fast food and engage in less physical activities. On the other hand, in some diseases, such as particular cancer, pcod, and PCOS, hormonal issues worsen with increased BMI. If a person is warned about unhealthy body weight, it is easy to take the required steps to avoid this medical condition and avoid chronic diseases. Early detection of obesity can help patients who have cancer risks or are in the initial stage as they can get much to get through this critical condition. For the above reasons, we collected data from the people of our country to generate a dataset that determines the risks of obesity and thus works to better human life.

1.3 Research Objective

Our research has always had a singular objective. To ascertain the characteristics affecting our country's obese population. The main features that cause a person to become obese in his/her upcoming days. To accomplish this goal, we gathered data from individuals throughout our country, and, more precisely, we compared the characteristics to those of some other countries' data and research. As for our objectives:

1. One of our objectives is to compare our dataset with another dataset that carries the regional data of Colombia, Peru, and Mexico.
2. Using machine learning techniques to find the key features
3. Determining the accuracy of the exact features that affect the BMI of our region.
4. Evaluate the accuracy of the features that we extracted through the algorithm.
5. Comparing with other models to select the best approach for our study.
6. Describe the confusion matrix for each of the models that we used.

After completing all the shortcomings, we concluded with the most important and some ethical features to make our study more accurate. A set of questionnaires were made to collect a set of information among the people of our region. We set some parameters to set the validness of the data and to process that data. We wanted to know the exact features that affected the people to become obese. We generated a heatmap to see the correlation of the data and features for its validity. As for the other dataset, we also made a correlation to compare them both. So we used a feature extraction algorithm known as BorutaShap. This algorithm follows a Random Forest algorithm (RF) to make its decision more accurate and faster for the smaller datasets AS we had more than 500 people filling up our questionnaire. Then we were able to get the exact and the vital features that were the main troublemakers. We compared them with the other dataset by those features, and we made our assumption from our study. The features were then used under some machine learning algorithm to get the accuracy of the fact that those features were only responsible. We used the Logistic Regression model of machine learning and several others like Random Forest, Naive Bayes, Decision Tree, and KNN. All the algorithms showed some prospective study results, but our study Decision Tree algorithm happens to be the most suitable model for our study.

1.4 Thesis Outline

Our work is based on a prediction model to identify obesity which is alarming for human health. This paper aims to make people aware of their health condition by detecting obesity in the primary stage through analyzing the dataset from the context of Bangladesh. Firstly, the introduction part(chapter 1) reflects our motivation behind this paper. The reasons for choosing this topic, goals, methods implemented, and a brief discussion is included. In the literature review (chapter 2), paperwork related to obesity has been reviewed, selected from the computer science background. Reviewing this paper's strengths, shortcomings, and to-dos in the future were pointed out to make a picture of what has been done so far and our motivation for coming up with this burning issue. Thirdly, the data collection and feature selection(chapter 3) section state our features that are the basis of the result and why we have selected primary data instead of choosing a secondary one that is very available. Principal component analysis, feature selection, and extraction, which are essential to reduce time complexion by reducing features, are also discussed in this part. We also emphasized the reliability of our study in this sector. Model selection and result analysis (chapter 4) include our proposed model, comparing results with visual representations, which give different rates of obesity detection by implementing different algorithms. The findings are also analyzed in this section. Thus, we mentioned important features through section and extraction, which leads to obesity.

Chapter 2

Related Work

2.1 Literature Review

BMI (The body mass index) is utilized in almost every clinical setting to determine and classify body measurements. Although several alternatives have been suggested, the most accepted idea is that BMI is a person's level of obesity. Because it tended to cause numerous health problems, it is often viewed as a risk factor for certain conditions. It is also frequently utilized in public health policies formulation. The BMI has proven of great benefit in population-based studies because of its broad usage in determining certain categories of body mass as a health concern. A report by Keys et al. concluded that Metropolitan Life Insurance was fraudulent. A true statement is reported data, which includes both the tables of ideal weight for height at the time and the tables that classified persons who were either underweight or overweight. As a result, the word "obesity" was rarely used. The solution was that Keys et al. used better-reported weight for height data and distributed the concept of the Quetelet Index more widely by making an end-user population-based study available. To simplify things, they called it the BMI (body mass index) (BMI). Body mass index is defined as body weight (kilograms) divided by height squared (meters). This calculation yields the Quetelet Index. Reducing the height lowers the impact of height fluctuation on the weight-to-height relationship; that is, it helps to normalize the body mass distribution across various heights. This was assumed to be crucial, given that the trunk contained most of the body's fat. Even yet, Keys et al. note that the BMI does not provide an accurate estimate of body fat percentage. This classification technique is used in population-based research to estimate mortality in various BMI groups. It is also used to estimate the likelihood of a specific aetiology causing death. The easy way to measure body fat is by using the BMI Calculator. The relation of mortality rates with BMI commonly ignores various factors, including the incidence of several cancers in the family, metabolic syndrome, familial longevity, or diabetes in the family, among others.[23].

According to a new paper published in the International Journal of Obesity (2019), Mexican, Peruvian, and Colombian citizens have obesity levels according to their food habits and physical condition, according to Fabio Mendoza Palechor et al. This dataset has over 1700 variables and over 3000 documents. Obesity Level (Class variable N), in combination with each of these labels—Inadequate Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III—enables classification of the data, which determines the

dataset's classification value. 77% of the data was generated by Weka, while 23% was obtained directly from users through a web-based platform. This data is useful for calculating an individual's overall weight and helping to establish weight management programs that monitor weight trends. The source of this information involves analyzing various eating habits and physical conditions together with the age of the participants surveyed (14-61). The study's first source of knowledge was a survey whose participants took online to evaluate their food, exercise habits, and other physical aspects that affect their well-being. Once all the data had been gathered, it was processed such that various data mining algorithms could utilize it. Data was categorised using the Equation Body Mass Index= $Weight/(height * height)$. Following that, WHO and Mexican Normativity performed a comparison of results. The definitions of obesity after the labelling procedure were uneven. Using Weka and the filter SMOTE tool, a significant portion of the data was obtained, including up to 77 per cent. According to the filter, the data produced by the process is expected to be in the class of synthetic data production, to use a certain number of nearest neighbours, double the percentage, and utilize a randomly generated seed for random sampling. To see if there were cases of data that did not adhere to expectations, anomalous and incomplete data was furthermore investigated. Following this, 2111 records were retrieved. To the advantage of SMOTE, the balancing mechanism restricts biased learning in favour of a dominating class.[24].

Sund et al. (2010) focused on family and neighbourhood variables as predictors of changes in BMI levels in the Norwegian adult population. There is solid evidence that the drivers of global obesity are numerous and occur in a wide variety of ways. Systems-based research provides a method to identify and quantify social influences that affect a person. In that study, researchers relied on two large-scale health surveys using longitudinal data gathered from 24,966 individuals who had been committed to staying in the same location for two years. HUNT 1 and 2 are two research studies in Norway that study the relationships between personal, family, and local variables and two outcomes: change in body mass index (BMI) at follow-up and change in BMI between baseline and follow-up. Individuals were all grouped into 17,500 households, 447 municipal wards, and 24 municipal entities called boroughs. Linear multilevel models were used in its design. To test these ideas, the researchers also analyzed many potential personal, family, and neighbourhood factors. These studies found a very strong association between human qualities and body mass index (BMI), with people who had a higher BMI often being male, unmarried, non-smokers, lacking experience, and not exercising. In general, the study found that individuals with a greater percentage of their body mass index (BMI) acquired more weight, and those who did not have employment income gained less. Other context-sensitive elements were important, with between 10% and 13% of the unexplained variation seen at the neighbourhood or municipality level, suggesting small environmental effects. [29]

Catherine S. Berkey et al. (2003) observed that among girls and boys, an increase in daily physical activity was linked to a reduction in the percentage of body fat, from 0.06 kg/m² per hour increase in activities (95% CI: 0.11, 0.01) in 1997 and 1998 to 0.07 kg/m² (95% CI: 0.12, 0.01) in 2003. There was a linear correlation between duration of inactivity and BMI growth in females (+0.05 kg/m² per hour increase in daily watching of TV/video/video games; CI: 0.02, 0.08). When weight increased, the effect was more pronounced. For certain girls, exercising for longer periods, such

as aerobics and dancing, was linked to reducing body mass index (BMI). They were trying to determine whether there was a relationship between one-year changes in BMI and one-year changes in the amount of time spent outside exercising or being inactive. A Cohort study was based on two surveys that were sent out one year apart. Between the ages of 10 and 15, 11,887 boys and girls from the Growing Up Today Study returned questionnaires. Changes in BMI that were associated with growth and development have also been considered. [2]

It was suggested in the investigation of (Myrskylä Chang, 2009) that Weight loss is directly associated with mortality increases. However, whether for benchmark well-being or weight growth, they are dependent on mortality decreases or increases. Weight gain does not just depend on an individual's baseline body mass index (BMI). Their study involved utilizing the Well-being and Retirement Study ($n = 13,104$) and Cox relapse study to see how life expectancy changes for 50- to 70-year-old adults based on a 2-year weight change for individuals with different BMIs. To highlight the point, they showed examples of the difference in weight reduction or growth of less than or equal to 1.2 BMI units and more than or equal to 3.5 BMI units. In the survey that they had previously conducted, two significant weight change mortality modifiers were discovered. First, they look at how a low initial BMI changes the consequences of weight gain. Since carrying extra weight indicates a significantly increased mortality risk, they propose that personal losses and gains resulting from increasing or decreasing weight are preferable to similar outcomes of losses and gains from more or less elevated weight levels. Next, they look at the overall impact of weight change, particularly concerning overall health. There are also confounding variables like well-being status, smoking, or actual work, which they consider—starting weight status and weight change each impact an individual, although most of their analyses have based on a particular modifier at a time. A comprehensive set of studies has assessed the long-term impact of the two modifiers in tandem. After analyzing all the data, they found an extremely broad outcome. In contrast to being outright malnourished, underweight is a medical condition that usually arises due to excessive body fat, such as being obese. Problems associated with increased BMIs include having an extremely elevated risk of death due to massive weight gain. Underlying BMI level has no relationship to small weight gains being accompanied by mortality. Disadvantages from lost weight such as health and well-being will likely show up in studies concerning emotional stability. While stating that it was the case that fat reduction was associated with mortality rates at the ends of papers for the average, overweight, and moderately obese persons, the authors promised that those older and who have lived a greater life would die at a slower rate. Thus, since the favourable effects of a lower BMI may outweigh the undesired effects, the probable benefits of a lower BMI must outweigh the disadvantages. Increasing mortality among important people with a BMI over 35 might be partially due to obesity. [22]

In this study, Wrotniak and colleagues(2004) found that a parent's standard deviation of body mass index (z -BMI) was associated with their child's standard deviation of body mass index (z -BMI). The design was a data analysis of three family-based, randomized, controlled weight increase trials, which evaluated weight improvements for both parents and children. 8- to 12-year-old obese children and their parents who had participated in family-based weight control treatments made up the participants. Parent z -BMI shift appeared to predict infant z -BMI change during the

time range of 0 to 6 months (P.001) and 0 to 24 months (P.009). The parent z-BMI shift was indeed a robust and progressive indication of the kid z-BMI change in hierarchical regression models. At six months, the extra r^2 range from 11.6% (P.001) to 3.8% (P =.02). When children of parents who were in the highest four of the z-BMI score shifted, their BMI changed significantly. However, parents in the lower quartiles had much more change than children of parents in the upper quartiles (P = 0.01). Researchers from the University of Buffalo looked into families enrolled in one of three family-based weight loss programs. If your child's BMI exceeds the 85th percentile and contributes to it with your own BMI, you may be a co-obesity offender. Of the people who were able to attend treatment sessions, 18 and older met the inclusion criterion. Diet and exercise habits and the temperature of the home were adjusted on request from parents. The parents were instructed to eat more fruits and vegetables and low-fat dairy products and increase their physical activity if the parents were determined to be not overweight. Therefore, regardless of their weight, both parents could make considerable health changes. BMI scores that placed parents at the 70th percentile of body weight were used to find the number of weight parents should allow their children to gain. Most parents who sought to lose weight but had BMIs below the 85th percentile tried to change their diet and activity habits to reduce weight. Families having complete parent and child data for all time points were included for the analysis. A stadiometer was used to measure a person's height in 0.32-cm increments, and a balance beam scale was used to measure a person's weight in 0.11-kg intervals. In order to compute BMI, the individuals' height and weight were used. Since both parents' BMIs and their children's BMIs were measured in two separate studies and calculated from two different references, standardized BMI indices (z-BMI) were calculated for both the mean (M) and standard deviation ($z\text{-BMI} = [\text{BMI} - M]/SD$) for each infant and partaking parent to examine the relationship between the change in the parents' BMI and changes in their children's BMIs during the first six months. P values, coefficients of determination, and 95% confidence intervals for confidence intervals, P values, or coefficients of determination were also applied to estimate the influence of weight reduction on weight gain. This research employed hierarchical models to study the association between parental weight and children's growth when other factors move into the first phase of the model. To adjust for the study group, dummy variables were added to the model. We researched the correlation between infant weight gain and the weight of the parents. An assessment of newborn weight shift by parental weight change was done at three different periods, with weights ranging from less than 0.5 kg to less than 24 months, and weights from 0.25 to 0.07 kg, and from 0.06 to 0.1. In weight change, the average decrease was 14 pounds, while the average increase was 5 pounds. In order to study these data, an aggregate two-type 2-factor analysis was done, in which the grouping variable (Grouping) was included as a grouping vector. When studying variations within groups, there was a comparison made using straight-line graphs. In addition to reporting child weight, sex, and age, the child's BMI was checked to see whether she was impersonating a younger or older child. Mean percentage and percentage of child weight vs height and mean percentage and percentage of child weight vs age ZBMI. The parent's weight was tracked as well as the 0- to 24-month z-BMI grades. They used a 2-way mixed variance analysis (ANOVA (also known as a two-factor analysis of variance or 2-way analysis of variance) that utilized grouping as the design vector and time

as the sample within, along with child sex and age as variables. Mean weight and BMI differed for each parent z-BMI shift. A survey of 84 children aged 10.2 and 60.9% overweight revealed that 58 boys and 84 girls were the ages 10.2 and 60.9% overweight, with a BMI of 27.9, a BMI percentile of 94.7, a z-BMI of 2.9, and a BMI of 59.0%. The average age of the parents was 40.8 (5.4) years, and they were 166.1 (8.1) cm tall. The mothers who participated (adjusted age = 40.3 years, BMI = 31.2, BMI percentile = 86.9, and z-BMI = 1.3) were in the 40.3-year-old age group (1.0). fathers who took part in the study were 41.9 years old, with a BMI of 31.2 and a BMI percentile of 88.0. (1.3). in 100 (88.7 per cent) of the households, the parents were superior to the 70th BMI percentile. Although 11.3% of families had a higher-weight contributing parent compared to a lower-weight contributing parent, in the remaining households, only the parent who contributed to the excess weight was above the 70th percentile. In terms of their socio-economic status, the households ranged from a low of 20 (the lower class) to a high of 66 (the upper-middle class) (upper class). There were 135 white households, six black families, and 1 Hispanic family. In line with expectations, a significant parent BMI z-score shift ($=.46$; 95% confidence interval, $=.31-=.61$; $r^2 = 0.21$) is expected to accompany infant BMI z-score change from 0 to 6 months (z-BMI $=.46$; 95% confidence interval, z-BMI $=.21-=.51$; $P=.001$), and from 0 to 24 months (z-BMI $=.29$; 95% confidence interval, z-BMI $=.07-=.51$; $P=.009$). Regression analysis found that changes in the parents' weight during the periods 0 to 6 months and from 6 to 24 months were important predictors of changes in their children's weight. Decreasing the number of people with z-BMI-quartile parents in the division (less than -0.5 z-BMI) showed a greater decrease in prevalence (as measured by z-BMI-quartile prevalence) over time compared to the reduced prevalence of z-BMI-zero parenthood in another three categories. In less than one year, the amount of Child Z-BMI that varies significantly from other Parent Z-BMI categories varies at the age of 0 to 6 months, at the age of 0 to 12 months, and at the age of 0 to 24 months ($P=.002$, $P_i.001$, and $P=.02$). This range is undergoing major changes. Each category has been significantly altered. Other interesting points to note are that the weight of all the students in the lowest z-BMI quartile (the one that the students were placed in after six months) increased relative to all the other grades (weight, 7.2 kg; height, 3.0 cm; and obese percentage, -27.6 per cent), while the weights and heights of students in the next two quartiles were no different (weight, 4.0 kg; height, 3.6 cm; and overweight percentage, 20.9 per cent). After two years, the bigger z-BMI shift quartile's parent children were at a disadvantage concerning height and weight, as their children in the smaller three-parent Z-BMI change quartiles were (weight, 7.1 kg, and height 11.2 cm). Without fluctuations in weight, they had a 16.8% weight shift. 10.8 kg; 11.7 cm; 10.5% Parent z-BMI variations also varied due to the quartiles of the parent z-BMI. Based on the study, it can be concluded that parents' weight gain directly influences their children's weight gain during and after recuperation. In order to better align both of our long-term plans, the partnership was shortened from 6 months to 2 to 1 month (0.2 months to 0.21) to 24 months (0.04). The research supports the increase in the period between treatment and follow-up. The largest shift occurs for parents in the third quintile, who have a child with a greater change in BMI. In order to better predict weight and height changes for those in the highest quartile of BMI, children in that quartile have greater weight and height growth for their height. An overweight child and his or her parents can impact each other's weight, which can

cause one or both parents to gain or lose weight. Good parents are more likely to be making healthier food choices, causing their children to be less sedentary at home. [31]

Berry et al. (2010) did research to study the influence of the built environment (subjective and objective body mass change) over time (BMI). Objectively, assessed community characteristics such as violence, traffic, physical activity, and schooling, and decreased BMI were investigated in this study. This study was designed to collect data on the influence of obesity on the built environment over six years (as measured by BMI). The main goal of this research was to see whether objectively assessed neighbourhood walkability and economic status, as well as the perception of neighbourhood characteristics (such as crime, traffic, and the type of scenery available), are linked to the expected change in BMI, along with other variables such as gender, age, educational attainment, physical activity, fruit and vegetable consumption, and smoking. In this study, the 2002 and 2008 Edmonton, Alberta region was the source of the data. An informal assessment on the general state of the city's population health was completed in 2003 by the Population Health and Research Department of the former Capital Health District of Edmonton (n=3174). The researchers' main goal was to determine the medical issues, causes, and needs in the Capital Region (including the City of Edmonton and outlying regions). Adults who resided in the former Capital Health Area were the targeted demographic. A total of 4,175 individuals took part in the poll. Regional layers were utilized to generate the walkability index, which only accounts for the City of Edmonton's citizens. Of these participants, there was a follow-up of 2362. In 2008, of the 822 applicants, 599 individuals had remained in the study cohort. At list-wise deletion of missed evidence, all 500 cases remained. Data were collected from October 28th, 2002 to December 15th, 2002. There was also a follow-up survey between November 2008 and January 2009. Canadian survey socio demographic questions were asked about age, race, schooling, work, and household income. Three query questions came from the Canadian Community Health Survey's Fruits and Vegetable module, asking how many days they consumed fruit and vegetables (in a week). Two classes were formed: those who consumed less than five servings of fruits and vegetables per week and those who ate more than five. Also, there was a term of smoking status. At both time points, BMI was estimated using self-reported height and weight. According to previous studies, they divided BMI changes into four categories: a decrease of at least 0.5 points (a decrease in BMI of -1.4 (standard deviation of 1.2) in men), little change (within 0.5 points; a change in BMI of 0.3 (standard deviation of 0.25) in men), moderate change (between 0.5 and 2 points; a change in BMI of 1.1 (standard deviation of 0.4) in men), and a substantial increase (greater than 2 points; a change in BMI of 1.1 (standard deviation of 0.4) in men) (mean BMI change 3.9). Taking into consideration that such variations in BMI have been shown to have a major impact on disease etiology, it's well worth investigating how constructed environment variables could contribute to the same BMI alterations. To verify null and tiny cells, they used cross-tabs. Three environmental issues were particularly difficult. Only a small percentage of individuals mentioned living in any type of residential building. about the questions dealing with transportation stops that are within 15 minutes of walking, 588 participants agreed. Therefore, Researchers excluded these three questions. From the remaining categories, they investigated if they were linked to BMI transition. Traffic was the only meaningful predictor of BMI and was the

only one used in the final regression. This metric was excluded from the analysis because the response rate was too low. An open-ended question yielded 35 answers (such as Aboriginal, Canadian, as well), with 30.7% describing themselves as being Canadian. Ordinal regression was the primary method. This research found that on average, the participants in the studies were wealthier and more prone to be married than the population used in the investigations. In addition, those who had not graduated from high school were less likely to participate in the study. Due to the prospective nature of the study and the inclusion of variables from 2002, predictors likely improved over time. Additionally, they discovered that 94 percent of participants' marital status remained stable, 82.8 percent's career status remained stable, and 91.2 percent's school status remained stable. Ninety-three percent of participants maintained their current smoking status. Smokers who have quit also have a greater body mass index (BMI) than smokers who have not quit. 28 According to the data, 71 per cent of fruit and vegetable intake remained constant, but most (55.4 per cent) increased their physical activity. At the same time, the average BMI (weight in kilograms divided by height in meters) increased by 0.4 (s.d. = 2.85) percentage points. A decrease of 2.21% in men's BMI and 1.74% in women's BMI occurred for 157 participants due to a fall in BMI of 0.5% for men and 0.4% for women. Participants' BMI stayed steady, while men's BMI climbed by 0.006, with a standard deviation of 0.19, while women's BMI declined by 0.002, with a standard deviation of 0.24. Of the 146 participants, an average of 1.17 points (a range of 0.5-2 points) increased in BMI for men, and 1.24 points (a range of 0.5-2 points) increased. Meanwhile, 93 participants saw a larger increase in BMI, averaging 3.88 points (a range of 2.17-4.41 points) for men and 4.41 points (a range of 3.21-4.97 points) for women. In keeping with prior studies, these improvements in BMI have remained stable. The most accurate ordinal regression model (aka complete ordinal regression model) was crucial, $\chi^2(18) = 46.32$, $P < 0.001$, and all of the goodness-of-fit checks were insignificant (i.e., none of the checks was significant by way of Pearson's χ^2 test ($\chi^2(1437) = 1449.70$, $P = 0.40$, or by the measure of deviance or by the difference of deviance from the null model). This statement claims that about 0.095 of the overall difference was corrected for (Nagelkerke=0.095). Also, they did not violate the parallel line test, $\chi^2(36) = 41.89$, $P = 0.231$, which means that the coefficient estimates were equal for all dependent variable ranges. BMI transition was predicted by factors such as local socioeconomic status and traffic. People in the lowest socioeconomic strata were much more likely to report a change in their body mass index (BMI) than those in the highest socioeconomic strata. Under the assumptions above, participants under 40 were also more likely to see their BMI rise than those in the 40 and up to age groups. [3]

Researchers from Norway and Sweden (Steinsbekk and Wichstrøm, 2015) investigated certain variables to predict BMI changes in preschool children between the ages of four and eight. They use several measurements such as physical activity levels, eating habits, and TV-watching time as predictors. More than two-thirds of the interviews for this study were done with Norwegian parents and children. Approximately 995 4-year-olds, 760 6-year-olds, and 687 8-year-olds participated in it. Body Mass Index Standard Deviation Score was questioned about, along with children's behaviour and activities, by the researchers (BMI SDS). Despite these findings, activity and time spent watching TV after the experiment did not correlate with the results. In this research, socioeconomic class and parental BMI are

included as covariates. The study's primary purpose was to identify patterns of physical activity and appetite on their own, as well as together. Higher levels of food reactivity, food satisfaction, and emotion overeating were associated with a rise in body mass index (BMI) faster. The correlation between satiety response and eating slowness and slower weight gain is opposite to that between slower weight gain and a better sensation of fullness. Parents who are overweight at 6 to 8 years old have greater BMI intercepts and higher BMI SDSs. Of the several parameters evaluated, food responsiveness parental BMI had no significant impact on a child's BMI SDS over 6 to 8. Food responsiveness (as indicated by a higher BMI SDS) is regarded to be a faster rise in BMI SDS. Even if a person has a high body mass index (BMI) SDS, having a higher food response but lower satiety responsiveness is possible. BMI SDS was not affected by physical activity or TV viewing time. Feeding babies exposed to and savouring the smells and sights of food increase the likelihood of weight gain. Excess weight weight growth are all risk factors for other health issues. [28]

In (Auerbach et al.,2017), a study was done to investigate the association between children's consumption of 100% fruit juice and their BMI. They conducted analyses with and without scaling factors used to regression coefficients to study the effect of standardizing one serving of 100% fruit juice to 6-8 oz. The number of 34470 youngsters involved in this survey is considerable. In the six-year-old children, they observed that 100% fruit juice was linked to a small weight increase. However, this was not evident in older children. In addition, kids between the ages of one to two years who drank one daily dose increment of 100% fruit juice were more likely to put on weight. Child under six years of age rise in weight is insignificant to the individual and unclear significance to the population. A study assessing the effect of 100% fruit juice consumption on metabolism and health outcomes, including such weight gain and obesity, is necessary on children between the ages of one and six. Until further research is conducted, the experts suggest the AAP's existing advice that children younger than six years old and older than eighteen consume no more than four to six ounces of 100 per cent fruit juice per day is sound and should be followed. In the study's view, long-term longitudinal studies with the best quality research represented in them and a literature review conducted by two authors throughout the investigation are a couple of the study's strengths. [1]

The findings of a study by Roberts et al. (2012) describe the interplay between biological and psychological variables and the effects of stress on body weight. This study had 71 nurses from a university-based nurse practitioner school who volunteered to be a part of it. At the beginning of the academic semester, the measurements were obtained. They were then taken again 12 weeks later when the participants had finished the evaluation. These factors have been proposed as predictors of weight loss, such as cortisol secretion, dietary inhibition, bingeing, mastery, temperament, and eating attitudes. Variations in individual behaviour established themselves as a risk factor for weight gain when faced with stressful life situations (dietary restraint and mastery). They were said to be within a standard deviation of the normal weight (BMI 25.2) for their age, with the documentation noting that the women were of average weight (BMI 25.2) for their age. A combination of dietary restrictions, weight, physical appearance, eating concerns, and physical well-being all increased, while depression, anxiety, and stress all reduced. The participants' average weight gain was 5.5 pounds in the trial, while their average weight loss was

2.5 pounds, and their average weight maintenance was 1.5 pounds. Changes in daily cortisol secretion were highly predictive of changes in BMI, and this association was much stronger for those who had mastered their task. This finding was supported by multiple regression analysis, which showed that an inverse association existed between dietary constraint and this connection. Loss of food control has also been predicted by loss of weight and self-mastery concerns. Changes in cortisol secretion were also associated with changes in dietary restraint, and these relationships were modified by dietary restraint when the academic semester began. This strategy is a significant advancement over previous studies that lack ecological validity and any longer-term measurement of whether improvements in eating behaviour are temporary and compensated for with the next meal. [25]

A behaviorally dependent screening method for identifying families that could predisposition children to be overweight was evaluated for predictive validity (Ihmels et al., 2009). This age group is connected to considerable annual BMI variations; hence it was decided to research the first graders. Before the start of the school year, parents completed the FNPA technique, and BMI data the following year was used to measure BMI change. For the sake of BMI data, 2,058 kids from all 37 of the district's campuses had their weight and height recorded. The investigation results into the information from the 1,030 participants revealed that 176 surveys had missing crucial data, making it impossible to use them for any study that involved data from the 854 participants. A strength of the study is that it utilized a diversified, ethnically varied population and used a longitudinal approach. An additional contributing factor may have been the domains that the ADA Evidence Analysis has found to be connected to childhood obesity. A follow-up test verified that the child's unique variation in BMI was captured by the FNPA total score ($p = 0.049$). In this study, 50% of the individuals had a considerable improvement in BMI per cent from the beginning of the study, with an average increase of 0.51 11.5%. This demonstrates that overweight habits can be reversed with sustained effort. As the survey's data sample of children with a large gain in BMI within a year was rather limited, the survey's capacity to identify changes in BMI that occur as a result of a family's home setting could have been compromised. This idea, plus the fact that each variable has to be weighted to get a risk score that reflects the total dangers, is something else you should know. These analytical tools, such as receiver operator curves (ROC) or other methodologies, as well as threshold scores, would be particularly valuable for this upcoming study because they enable risk categorization to be empirically computed. According to the experts, the future implications of childhood obesity trends will make this a highly regarded subject for potential investigation. [19]

A study was conducted by (Calitri et al., 2010) found that changes in BMI can be anticipated from the value of cognitive biases. A high BMI is related to a wide scope of medical issues, and it normally mirrors an inactive way of life and undesirable dietary patterns. These incorporate eating style, stress, and affectability to compensate. Their exploration was to consider psychological inclinations' intellectual predispositions for food-related data to anticipate changes in BMI. Their investigation investigated the prescient worth of intellectual inclinations to food signs (evaluated by enthusiastic Stroop and speck test assignments) on weight change over one year. This was a longitudinal report with college understudies ($N = 102$) living in shared understudy convenience. After controlling for the impacts of factors related

to weight (e.g., actual work, stress, limited eating, outer eating, and enthusiastic eating), no impacts of intellectual inclination were found with the spot test. A related key inquiry, and the focal point of the current work, is whether food-related psychological predispositions can anticipate changes in weight change. This issue has not been tended to in eating research, even though there have been promising equal discoveries for different psychopathology. Such reassuring outcomes have driven a few analysts to seek after cognitive-style intercessions, that is, meditations for a substance misuse issue that works by endeavoring to decrease comparing intellectual inclinations. There have been cross-sectional examinations in their hunger research showing how the size of attentional predispositions to food-related upgrades contrast between people who embrace diverse eating styles. Be that as it may, attentional predispositions for food-related data have been helpless indicators of current BMI. Investigation into attentional or intellectual predispositions to food-related data has regularly utilized either the spot test or the enthusiastic Stroop. For the passionate Stroop, intellectual inclination to unfortunate food sources anticipated an expansion in BMI though psychological predisposition to quality food sources was related to a decline in BMI. Further, Stroop and spot tests will probably gauge various parts of intellectual inclination; it gets hard to settle on the most proper apparatus to utilize. The enthusiastic Stroop has been the solitary measure to show prescient capacity concerning conduct change in excessive drinking. Besides, its inescapable appropriation, especially in the fixation writing, has prompted a decent comprehension of the variables which may influence members' presentation. Similarly, the speck test has been a normally utilized choice to survey more explicit parts of attentional inclinations identified with substance misuse or related practices. The important examination does not suggest reasonable other options. They received an exploratory methodology and tried members with food renditions of both the enthusiastic Stoop and the speck test along these lines. Their outcome was that intellectual predispositions seem to anticipate conduct change. In like manner, future exploration ought to think about techniques for attentional retraining, urging people to reorient consideration away from undesirable eating signals.[4]

A study (Drøyvold et al., 2004) showed that their main objective was to find the relation between body mass index(BMI) and leisure-time physical activity (LTPA). The body-weight firmly connects with dreariness and mortality, and the new weight pestilence may have sensational ramifications for human wellbeing and medical care frameworks. The unfriendly well-being-being outcomes of stoutness get impacted by body weight and muscle to fat ratio, the greatness of weight acquired during adulthood, and an inactive lifestyle. Weight acquires a result of higher energy admission than energy consumption after some time. Recreation time busy work (LTPA) is the most crucial factor in humans' absolute energy use. Nonetheless, the information about LTPA as an avoidance factor of the weight acquired is scanty at the populace level. Bodyweight and LTPA are various wonders among ladies and men. Men are by and considerably more genuinely dynamic than women. However, weight control is certainly not a solid rousing element for being dynamic among men. Numerous men do not perceive themselves as overweight. The weight record (BMI) level related to grimness and mortality is lower for men than women. To acquire better information about the weight acquiring pestilence and the relationship to LTPA, we feel that examinations ought to be performed separated by sex due to the sexual orientation heterogeneity and zero in explicitly on counteraction. It shows that

overweight and hefty individuals over-report physical activity and see bodyweight as a hindrance to being genuinely active. Dependent on dreariness and mortality connected to body weight, The World Wellbeing Association (WHO) has characterized bodyweight classes (underweight, typical weight, and overweight) for grown-ups utilizing BMI ranges. So they selected men characterized as a standard load gauge by utilizing WHO has normalized BMI characterization framework. To diminish heterogeneity and the capability of misclassification, they needed to examine the relationship between LTPA and change in BMI in a sound male populace with average body weight at benchmark with a potential for being dynamic in a forthcoming 11-y follow-up examination. Moreover, they needed to utilize the backward relationship between LTPA level and pulse as a determination measure since pulse and wellness are related. They utilized the pulse to help the legitimacy of the LTPA surveys, and solid men chose the age range 20–69 years by utilizing the backward relationship between pulse and LTPA level as a choice model. As a result, this examination has exhibited a moderate BMI impact of LTPA at the populace level. Despite this, an undeniable degree of LTPA did not forestall weight acquired during the 11-year follow-up period. [7]

This study (BMI, Body Build, Body Fatness, and Health Risks, 2012) found that the weight record (BMI) is frequently reprimanding because it is an off-base muscle-to-fat ratio list. In any case, verifiable in this analysis is that a superior muscle to fat ratio pointer would be adequate to isolate unfortunate bulky individuals from fragile sound individuals. Muscle to fat ratio measures are promptly accessible and are not better than BMI as pointers of wellbeing hazards. The BMI was created as a file of large body size by eliminating body stature's impact on body weight. They Utilized around one thousand significant class baseball players' BMI information. The BMI's values were outlined in two dissipated plots. Tallness and weight were connected, while stature and BMI were random. Epidemiological investigations evaluated that have decided BMI and percent muscle versus fat concerning demise rates and danger factors. On every occasion percent muscle to fat ratio is better than BMI in foreseeing danger, and BMI is more firmly identified with hazard factors than is body heftiness. In this way, abundance bulk or huge body size might be a critical indicator of infection that BMI catches. Critically, even though weight reduction can decrease muscle versus fat, it cannot change body size or strength. [10]

Another way to look at it: Dubowitz et al. (2012) looked at data from 60,775 women (i.e., $n = 60,775$ women). They discovered this data from the WHI CT - a public inquiry of postmenopausal women aged fifty to seventy-nine years of age. In studying the association between various food outlets in the 1.5 miles surrounding a woman's home, registration location, neighbourhood financial status (NSES), BMI, and circulatory strain, they observed a cross-sectional association (BP). Two writers used the NSES model and food outlets to demonstrate both straight and strategic relapse, responding to various sociodemographic parameters, the number of people, and the influence of irregular occurrences at the parcel and MSA levels. According to the study, it was found that the correlation between obesity, increased BMI, and high blood pressure (commonly known as hypertension) is much stronger when accounting for the study subjects' ability to obtain food and beverages. From the tenth to the ninety percentile, where BMI was higher by 0.28 kg/m², their research on the accessibility of fast food outlets expanded. However, the BMI result was lower by 0.30 kg/m² for the supermarket subjects as well. While expanding from the tenth

to the 90th percentile of the school's budget, NSES found that it cut 1.26 kg/m² off the population's BMI. The percentage of above the 90th percentile scores increased from the tenth to the 90th percentile, resulting in a 1.11 mm Hg decrease in systolic blood pressure and a 0.40 mm Hg decrease in diastolic blood pressure. Accessibility increased from the 90th percentile in the superstore, but 0.31 mm Hg reduced this room's diastolic blood pressure. By identifying a relationship between food and finances, postmenopausal women in this example demonstrated how fundamental, independent relationships exist between food, nutrition, and health. Because of these studies, it's suggested that local climate variations could also assist in efforts to regulate weight and hypertension.[8]

In the study of Goldschmidt et al. (2008), They disarranged the eating perspectives and practices which give off an impression of being very regular in youth, and overweight youth has been distinguished as a subset of the populace at exceptionally great danger for underwriting such indications. Overweight and dietary problem (ED) symptomatology autonomously give critical dangers to physical and psychosocial wellbeing. It showed solid connections with body weight gain and danger for ED improvement. At the point when simultaneous, the danger for negative wellbeing results might compound. This article aims to audit the writing's present status as it concerns scattered eating and its corresponding overweight youngsters and teenagers. Surviving writing on the control, distribution, corresponds, and etiology of disarranged eating aspects and studies (i.e., negative mentalities toward shape and weight, undesirable weight control practices, and overeating) in overweight youth are investigated and combined to make appraisal and treatment suggestions for medical services suppliers. The current writing proposes that early discovery of disordered eating in overweight youth ought to be a need to give the proper intercession, along these lines easing back the direction of weight acquisition and forestalling or lessening the drawn-out negative results related with the two conditions. Future examinations should zero in on elucidating formative pathways and creating novel counteraction and treatment intercessions for overweight youth displaying disarranged eating designs. [14]

With the help of BMI trajectories the PLOC i.e Prostate, Lung, Colorectal, and Ovarian Cancer can be analyzed. Through the screening process, this method uses the BMI data to conduct the research. A questionnaire from the authors was given to fulfill the necessary data and it was categorized with the criteria that WHO had given underweight ($< 18.5 \text{ kg m}^2$); normal ($18.5 \text{ to } 24.9 \text{ kg m}^2$); overweight ($25 \text{ to } 29.9 \text{ kg m}^2$); and obese ($> 30 \text{ kg m}^2$). The changes in weight during their adulthood were classified as a loss ($\leq 2 \text{ kg}$), stable ($> 2 \text{ to } < 5 \text{ kg}$), gain ($\geq 5 \text{ to } < 15 \text{ kg}$), and notable gain ($\geq 15 \text{ kg}$). To demonstrate the relationship between age-specific BMI, average BMI, and weight change, along with CRC risk, the authors employed Cox proportional hazard regression models. Four trajectories were so constructed as a result of this study. In their investigation, scientists discovered that being fat greatly increases the risk of having colorectal cancer in maturity. [32]

In the paper of Dar-Nimrod et al. (2014), the authors noticed that the public talk on hereditary inclinations for stoutness has prospered in many years. Their three investigations researched typically important associates and outcomes of an apparent hereditary etiology for heftiness. For their situation study 1, there was a survey of convictions about etiological clarifications for weight. More grounded underwriting

of genetic etiology was prescient of a conviction that fat individuals do not influence their weight. For their situation study 2, convictions about weight and its causes were evaluated following control of the apparent hidden reason. Contrasted and a hereditary attribution, a non-hereditary physiological attribution prompted an expanded impression of authority over one's weight. For their situation study 3, members read an anecdotal media report introducing either a hereditary clarification, a psychosocial clarification, or clarification (control) for weight. Results showed that members who read the hereditary clarification gobbled fundamentally more on a subsequent errand. Taken together, those three investigations of the creators exhibit the likely impacts of hereditary attributions for weight. [6]

It is estimated that colorectal or colon cancer is the third most common malignancy in men and women. The number of new cancer cases and fatalities projected to have occurred in 2008 is predicted to be over 1.2 million and 608,700. The 5-year survival rate of colorectal cancer is 63% if it's diagnosed properly at its early stage, colorectal cancer may be cured. Early detection of cancer can save the lives of thousands of individuals. In addition to this, proper treatment is dependent on how accurately symptoms are detected by appropriate medical exams. Haque et al. (2005) used an automated dataset to establish an algorithm especially to differentiate screening as well as diagnostic endoscopy medical tests of colorectal cancer. Using diagnostic codes as well as automated data, the system was able to detect which endoscopies were indications of the endoscopies as well as classify them as diagnostic or screening. Screening and endoscopy exams differ in cost, symptoms type, treatment type, methodologies, and time-bound. Besides, in colonoscopy, the whole colon is brought under observation, while a sigmoidoscopy examines only the lower part of the colon which is also known as the rectum and sigmoid colon. For this reason, a patient needs to be given proper treatment by identifying the symptoms. That's why the goal of this study was to build up a technique that will be helpful both for the patients and doctors. For the research purposes, two trained abstractors cooperated by reviewing medical records from 1997 to 2002. They also assessed the medical exams comparing those having gastrointestinal conditions or indications. Moreover, abstractors named the endoscopies as diagnostic if the unclear or specific symptoms were found which need proper treatment. On the other hand, all other endoscopies were listed for screening. It was mentioned in the paper that cross-tabulations were conducted between the approached technique and reports for the determination of the efficiency by considering , specificity, and sensitivity. After medical report analysis, the identification was named "Gold Standard". Here, sensitivity demonstrates similarities such that an analytic test for endoscopy was delegated such by clinical records audit. The possibility that the nondiagnostic endoscopy was labeled as screening is increased by specificity. It is expressed by . The methodology determines a temporary list of gastrointestinal phases and indications for testing and improvement, as suggested by the study. The algorithm's capability whether it can properly characterize the indication or can't, diversified by endoscopy exams. According to the procedure, the sensitivities for finding out diagnostic colonoscopy and sigmoidoscopy were 23.8% and 48.1%. Most of the diagnostic endoscopies were not able to be covered. On the other hand, the sensitivities for detecting screening colonoscopy and sigmoidoscopy rates were good enough (84.4% and 87.9%) but were found with low specificities. The was 76.3% for sigmoidoscopy but only 44.2% for colonoscopy. In assessing the algorithm's efficiency, medical records were used as

the "Gold Standard." Compared to medical data, the algorithm could detect some, but not all, screening endoscopies. Researchers have said that their approaches have certain drawbacks since diagnostic colonoscopies were missed in most of them. During the research time, the performance was only carried out on less than 25% of the members. By adding this initial list and checking to see if specific preventive health risks show different signs based on gender, Researchers have created an automated dataset with the intent of helping to find an effective method for screening and diagnostic distinction. But the number of gastrointestinal conditions, signs, and symptoms are limited for the algo to mark the endoscopies as diagnostic. Besides, it doesn't take the BMI index as an input to generate optimum results. [17]

CRC refers to the abnormal growth in the cells within the colon or rectal area. Different reasons can work as the culprit for CRC. Obesity is one of them. The more one is obese, the higher the BMI as well as the risk of colorectal cancer. (Sinn et al.,2017) showed how metabolically-healthy obesity can contribute to increasing the chances of having colorectal adenoma(CRA) which is an antecedent of CRC. They analyzed the connection between BMI chart, metabolic condition as well as CRC of adults who had cancers but no symptoms initially and had undergone colonoscopy Health Promotion Center of Samsung Medical Center. For the research purpose, having an equilibrium model assessment of insulin resistance was taken less than 2.5, and metabolically fit was referred to as not having any kind of abnormalities. For the study purpose, 9,182 metabolically-healthy and 17,407 unhealthy participants were chosen. The Possibilities of having CRA was categorized as any, multiple, and high risk. According to their finding, people with higher BMI have more risk in these types of sections than healthy ones.[27]

BMI (body mass index) is the ideal and standard process for controlling obesity among people. According to the House of Commons Select committee and two expert groups of USA, the BMI (body mass index) of school-going children should be observed, and parents should know about the outcome. Further, the Select Committee asked for an explanation from the parents whose children were overweight, and BMI cut-offs were chosen to give instructions to parents. BMI is related enough to the measurement of fatness used as a public health instrument for controlling obesity. As there is much separate inconsistency among BMI (body mass index) and body fat, heart disease, and other permanent health effects, BMI is a deficient agent for obesity. Childhood obesity is a threat to public health as it causes harmful effects on health. It causes unhappiness because a person who is obese feels unattractive.UK government made a target of reducing obesity by 2010. [15]

A prevalent cause of cancer-related death in women nowadays is breast cancer. A study by Wambui et al. (2005) found a significant association between age, weight, and the likelihood of getting breast cancer by the time women are 25 and have an initial breast cancer diagnosis. The shifting effects of demographic variables were also accounted for in this strategy. This information from the National Health and Nutrition Examination Survey was used. Those ladies responded to the poll (172 with breast cancer and 2723 with no breast cancer diagnosis). Though all three of these women were 50 or older, none were pregnant, and none of them had any non-breast cancer. The researchers applied multivariate logistic regression to calculate the odds ratio and confidence intervals at the 95% level. When the study was performed, women who had a higher body mass index (BMI) had a 2.1-fold increased risk (OR 2.1; 95% CI 1.1-3.8) of getting breast cancer compared to women with an

optimum BMI (risk of getting breast cancer). A meaningful link was established among general BMI and increased BMI in non-Hispanic black women (odds ratio 6.6, 95 per cent confidence interval 1.68-25.86). (odds ratio 4.2; 95 percent confidence interval 1.02 -17.75). To put it another way, women who are obese when they are over 25 are at greater risk of breast cancer. Even for non-Hispanic black women aged 50 and up, the risk is substantial. [12]

Child obesity is a huge problem because of the damage it does to health for life. Reducing obesity in kids over the long term will need having basic information. The effectiveness of physical activity in treating pediatric obesity is unmeasured, yet this method effectively treats obesity. The analysis of physical activity and waist circumference on overweight children for the sake of helping enlighten with intervention methods that assist with permanent weight loss in obese children could benefit from this type of study. A study conducted by Andrew et al. explored a relation between persistent PA-BMI in overweight children. Three thousand nine hundred fifty-eight children were observed, and 258 were enrolled, and in 3 years follow-up of 5 or 10 obese children in Melbourne, 182 were evaluated (76 eliminated for missing data). Some significant gains were observed when people wore the accelerometer for seven days. These include BMI z score (coefficient of determination was 0.02, with 95 per cent confidence interval from 0.04 to 0.09), as well as a difference in BMI category (odds ratio was 0.70; with 95 per cent confidence interval from 0.43 to 1.13) and changes in BMI values up to the point where a person's BMI was 0.5 were observed. Throughout the study, daily activity did not affect BMI change. BMI z score will be reduced by 0.11 for every 100 counts per minute increase in the use of exchange instead of a money donation (95 per cent CI 0.03 to 0.20 and p 0.006). Because it's been shown that maintaining increased physical activity levels for three years in obese youngsters reduces the BMI score. [30]

Body mass index was significantly associated with both mortality risk at the time of diagnosis and mortality risk following the diagnosis of colorectal cancer. Whilst preliminary study reveals that exercising can be suggested for people with cancer, outcomes show that suggestions are not warranted for those who are overweight in the immediate period. Despite medical research indicates that being physically fit in cancer patients is beneficial, data show that advocating for weight loss in symptomatic overweight people who experience the symptoms of their disease right after diagnosis is ill-advised. The paper employed covariance to evaluate linear variables by BMI categories and age at diagnosis in the context of people with a BMI at or above the median. [20]

Chapter 3

Data Collection and Feature Selection

3.1 Data Collection

In recent years, BMI has been treated as an essential topic in a variety of study domains. Our goal is to bring machine learning principles into this field of study. Machine learning technologies are being utilized to tackle social health problems in the new era of artificial intelligence. They are regarded as well adapted to handle specific problems since they produce improved precision. In the previous papers, the researchers have pointed out several problems and techniques to tackle the situation using some advanced tools, and they researched at a very high level. So we tried to conduct the study at a primary level and decide based on this topic. Since we were conducting the research, the world went through a big blow of COVID-19, and the world came under a massive threat, and we were bound to take the study securely. So as the country was under a complete lockdown and we could not research an online process. We collected our primary data from the google form with some questionnaires on it from different people across the country. Collecting this dataset is to make research more lenient and more accurate for this region. Our research also compared the attributes of other countries with our dataset and concluded with it. In order to make our dataset, we concluded that we needed 17 different columns to make our study more relevant to come up with a solution. As for our study, we collected more than five hundred data from our google questionnaire. And then, we collected the dataset based on behaviors and the characteristics of our daily life. We took the data about the person's family history of being obese or not. Because some studies have shown that a person is obese or not can be found out from their family and their surroundings where they live or spent most of their life. Genes can influence appetite management, food consumption, metabolism, body-fat distribution, body mass index (BMI), food preferences, exercise response, and an individual's eveningness or morningness, all of which can lead to obesity. With the conclusion of the Human Genome Project in 2003 and the accompanying advancement in technology, our understanding of the genetic basis of obesity has grown dramatically—though there is still more to learn. Obesity genetic associations are classified as polygenic, monogenic, or epigenetic.[13] In addition, body mass also varies greatly between men and women, even when they are the same age and gender. It appears that girls who start puberty early are more likely to be obese as

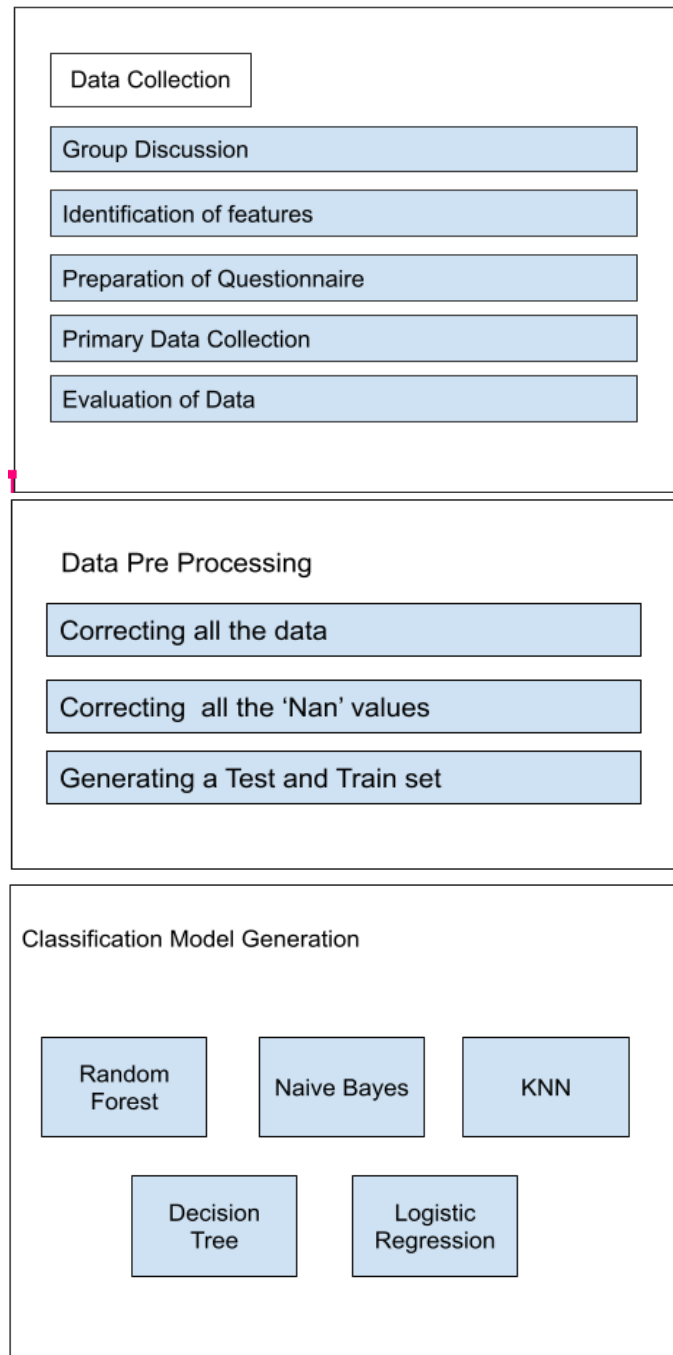


Figure 3.1: System module for ML-based approach

adults. In fig. 3.2 we can see that, childhood (pre-menarcheal) weight and height are strongly associated with earlier menarche, obesity in older adolescent females is more prevalent because obesity that started during childhood has continued into adulthood. The role of early menarche on adult obesity has been over-emphasized since further research is required. Most of the apparent effects of menarcheal age on adult obesity can be explained by the link between menarcheal age and obesity in childhood. [11]

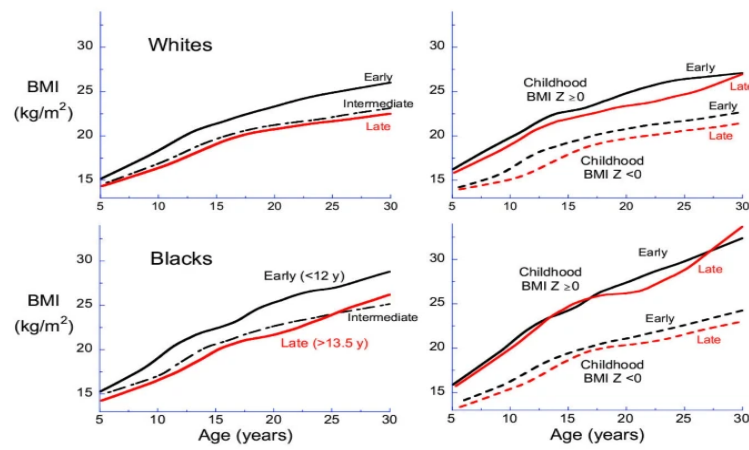


Figure 3.2: Variation of obesity due to age

Firstly, these were the features that we wanted to conduct our study, so the features shown in table 3.1 were included in our questionnaire. Nevertheless, these were indicated in our dataset as short-form using some keywords. For example, Frequent consumption of high caloric food (FAVC), Frequency of Consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC) Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS). The dataset was organized, and the class variable NObesity was defined with the values of Insufficient Weight, Normal Weight, Level I Overweight, Level II Overweight, Type I Obesity, Type II Obesity, and Type III Obesity based on the information obtained from WHO. In fig. 3.3 and in fig. 3.4 the scaled and non-scaled datasets are shown respectively.

Column	Feature Name
1	Gender
2	Age
3	Height
4	Weight
5	Family History
6	Water Consumption
7	Alcohol Consumption
8	Calorie Consumption
9	No. Of Meals
10	Smoking
11	Transportation
12	Physical Activity
13	Consumption Of Vegetables
14	Electronic Device Usage

Table 3.1: Feature Names

Gender	Age	Height	weight	BMI	Family_history	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC
Male	24	1.803	85	26.14733748	No	No	No	2	Always	No	1	No
Female	21	1.6	80	31.25	Yes	Yes	No	3	Sometimes	No	2	No
Male	24	1.72212	72	24.2779574	Yes	No	Yes	3	Sometimes	No	1	No
Male	18	1.778	84	26.57148171	Yes	No	Yes	4	Always	No	3	Yes
Male	24	1.68	64	22.67573996	No	Yes	Yes	3	Sometimes	Yes	4	No
Male	23	1.72212	78	26.30072872	Yes	Yes	Yes	2	Sometimes	Yes	2	No
Male	22	1.8034	52	15.98892307	No	Yes	Yes	4	Sometimes	Yes	1	No
Male	23	1.75	65	21.2244898	No	Yes	No	3	Frequently	Yes	1	No
Male	23	1.8288	91	27.20877355	Yes	Yes	No	2	Sometimes	No	1	No
Male	23	1.75	65	21.2244898	No	Yes	No	3	Frequently	Yes	1	No
Male	23	1.8288	91	27.20877355	Yes	Yes	No	2	Sometimes	No	1	No
Male	23	1.73736	68	22.53234438	No	Yes	Yes	3	Sometimes	No	2	No
Male	23	1.6764	72	25.61880395	No	Yes	Yes	1	Sometimes	Yes	3	No
Male	20	1.78	77	24.30246706	No	Yes	Yes	3	Sometimes	No	2	No
Male	18	1.73736	71	23.52224191	Yes	No	Yes	3	Sometimes	No	2	No
Male	17	1.79832	54	16.68782139	No	No	Yes	2	Sometimes	No	3	No
Male	23	1.72212	75	25.28919223	Yes	Yes	Yes	2	Frequently	No	1	No
Male	18	1.825752	92	27.59899338	Yes	Yes	Yes	3	Always	No	4	No
Male	29	1.6764	76	27.04321295	No	Yes	No	3	Sometimes	No	1	No
Female	24	1.7	78	26.88919338	No	Yes	No	4	No	Yes	4	Yes
Male	23	1.8	82	25.30864198	No	Yes	No	3	Sometimes	Yes	1	Yes
Male	18	1.72	88	29.74580855	No	Yes	Yes	1	Sometimes	Yes	2	No
Male	23	1.72	80	20.2912331	No	Yes	Yes	3	Sometimes	Yes	2	No
Male	17	1.67	45	16.13539388	No	Yes	Yes	3	Sometimes	No	4	No
Male	23	1.778	54	17.08166882	Yes	No	No	4	Sometimes	No	3	No
Male	21	1.8	46	14.19753869	No	No	No	2	Sometimes	No	2	No
Female	22	1.5748	70	28.2258629	Yes	No	Yes	3	Frequently	No	1	No
Male	18	1.7	64	22.14532872	No	No	Yes	1	Sometimes	No	1	No
Male	18	1.778	86	27.28413804	Yes	No	Yes	3	Sometimes	No	2	No
Female	24	1.6256	65	24.59721716	Yes	Yes	Yes	3	Sometimes	No	5	No
Male	20	1.8	87	26.85185185	No	Yes	Yes	3	Frequently	No	5	No
Male	24	1.778	68	21.5102471	No	Yes	Yes	1	Always	Yes	2	No

Figure 3.3: Non-scaled Dataset

Gender	Age	Height	weight	BMI	Family_hi	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	Nobeyes	RISK	
0	24	1.803	85	26.14734	1	1	1	2	3	1	1	1	1	0	1	0	0	2	0
1	21	1.6	80	31.25	0	0	1	3	1	1	2	1	0	2	0	1	3	1	1
0	24	1.72212	72	24.2776	0	1	0	3	1	1	1	1	0	2	0	0	1	0	0
0	18	1.778	84	26.57148	0	1	0	4	3	1	3	0	3	2	0	0	2	0	0
0	24	1.68	64	22.67574	1	0	0	3	1	0	4	1	0	1	0	0	1	0	0
0	23	1.72212	78	26.30073	0	0	0	2	1	0	2	1	2	2	0	0	2	0	0
0	22	1.8034	52	15.98892	1	0	0	4	1	0	1	1	0	2	0	0	0	0	0
0	23	1.75	65	21.22449	1	0	1	3	2	0	1	1	2	2	0	2	1	0	0
0	23	1.8288	91	27.20877	0	0	1	2	1	1	1	1	1	2	0	2	2	0	0
0	23	1.75	65	21.22449	1	0	1	3	2	0	1	1	2	2	0	2	1	0	0
0	23	1.8288	91	27.20877	0	0	1	2	1	1	1	1	1	2	0	2	2	0	0
0	23	1.73736	68	22.52834	1	0	0	3	1	1	2	1	1	2	0	2	1	0	0
0	23	1.6764	72	25.61989	1	0	0	1	1	0	3	1	0	2	2	2	3	2	0
0	20	1.78	77	24.30249	1	0	0	3	1	1	2	1	2	1	2	1	1	1	0
0	18	1.73736	71	23.52224	0	1	0	3	1	1	2	1	0	2	0	0	1	0	0
0	17	1.79832	54	16.69782	1	1	0	2	1	1	3	1	0	2	0	2	0	0	0
0	23	1.72212	75	25.28916	0	0	0	2	2	1	1	1	0	2	0	2	2	0	0
0	18	1.825752	92	27.59899	0	0	0	3	3	1	4	1	0	1	0	2	2	0	0
0	29	1.6764	76	27.04321	1	0	1	3	1	1	1	1	0	2	0	2	2	0	0
1	24	1.7	78	26.88962	1	0	1	4	0	0	4	0	3	1	3	1	2	0	0
0	23	1.8	82	25.30864	1	0	1	3	1	0	1	0	2	1	0	2	2	0	0
0	18	1.72	88	29.74581	1	0	0	3	1	0	2	1	0	1	2	1	1	0	0
0	23	1.72	60	20.28123	1	0	0	3	1	0	2	1	0	1	2	1	1	0	0
0	17	1.67	45	16.13539	1	0	0	3	1	1	4	1	0	2	0	2	0	0	0
0	23	1.778	54	17.08167	0	1	1	4	1	1	3	1	0	2	0	1	0	0	0
0	21	1.8	46	14.19753	1	1	1	2	1	1	2	1	0	1	0	2	0	0	0
1	22	1.5748	70	28.22586	0	1	0	3	2	1	1	1	0	2	0	1	2	0	0
0	18	1.7	64	22.14533	1	1	0	1	1	1	1	1	1	1	2	0	0	1	0
0	18	1.778	86	27.20414	0	1	0	3	1	1	2	1	2	2	0	1	2	0	0
1	24	1.6256	65	24.59722	0	0	0	3	1	1	5	1	0	2	2	2	1	0	0
0	20	1.8	87	26.85185	1	0	0	3	2	1	5	1	0	2	0	2	2	0	0
0	24	1.778	68	21.51025	1	0	0	1	3	0	2	1	2	1	0	2	1	0	0

Figure 3.4: Scaled Dataset

3.1.1 Data Pre Processing

As for data processing, we found it a great challenge to make it scaled from the non-scaled dataset. The reason behind this is that many persons who filled up the form only gave us the context type data, and from there, we had to convert it to numerical data to make the best use of the dataset for our study to progress further. We received around more than 500 hundred person data to work with them and process them to approach it in a machine learning manner. Nevertheless, as we know, contextual data can be a lot harder to process, and it becomes much more complicated when there are dummy data as some of the people gave it. In our form, we wanted to know around 17 questions from the users. In general, people gave their height and weight in feet and kilogram respectively. Nevertheless, to calculate the BMI and process the data, we needed to convert it into centimeters, and it was a challenge as more than 500 hundred data was needed to edit. For the approach towards our dataset, We converted all the contextual data such as the transportation they have been using (which is denoted as MTRANS), Consumption of food between meals (denoted as CAEC), Gender, *family_history*, frequent Consumption of high caloric food (denoted by FAVC), frequent Consumption of vegetables(denoted by FCVC), do the smoke or not (Smoke), do they drink alcohol or not (CALC), do they monitor their calorie intake (SCC) and the types of obese they are. We followed the data cleaning procedure to convert them into numerical data to make it more organized and more accurate for the feature extraction process. As per our dataset, we considered;

"Gender": "Male" as 0, "Female" as 1,

"*Family_history*": "Yes" as 0, "No" as 1,

"Nobeyesdad" : "Insuficient Weight" as 0 ,

"Normal Weight" as 1, "Overweight" as 2, "Obesity I" as 3, "Obesity II" as 4, "Obesity III" as 5,

"RISK": "Risk" as 1, "No risk" as 0,

"CAEC" : "Always" as 3, "Frequently" as 2, "Sometimes" as 1, "No" as 0,

"FAF": "I do not have" as 0, "1 or 2 days" as 1, "2 or 4 days" as 2, "4 or 5 days" as 3,

"FAVC": "Yes" as 0, "No" as 1,

"FCVC": "Yes" as 0, "No" as 1,

"SMOKE": "Yes" as 0, "No" as 1,

"SCC": "Yes" as 0, "No" as 1,

"TUE": "0-2 hours" as 1, "3-5 hours" as 2, "more than 5 hours" as 3,

"CALC": "Always" as 3, "sometimes" as 2, "Frequently" as 1, "I do not drink" as 0,

"MTRANS": "Walking" as 0, "Automobile" as 1, "Public Transportation" as 2, "Motorbike" as 3, "Bicycle" as 4.

And then, the dataset was all in numerical form as we expected as per shown in fig 3.5.

Gender	int64
Age	float64
Height	float64
weigth	float64
BMI	float64
Family_history	int64
FAVC	int64
FCVC	int64
NCP	int64
CAEC	int64
SMOKE	int64
CH2O	int64
SCC	int64
FAF	int64
TUE	int64
CALC	int64
MTRANS	int64
Nobeyesdad	int64
RISK	int64

Figure 3.5: Datatype

As we compare our dataset to the Dataset of Colombia, Peru, and Mexico, we had to process that dataset too. Furthermore, that dataset also carried the contextual forms. So we considered;

"Gender": "Male" as 0, "Female" as 1,

"family_history_with_overweight": "yes" as 0, "no" as 1,

"NObesyesdad": "Insufficient_Weight" as 0, "Normal_Weight" as 1, "Overweight_Level_I" as 2,

"Overweight_Level_II" as 3, "Obesity_Type_I" as 4, "Obesity_Type_II" as 5,

"Obesity_Type_III" as 6 ,

"RISK": "Risk" as 1, "No risk" as 0,

"CAEC" : "Always" as 3, "Frequently" as 2, "Sometimes" as 1, "no" as 0,

"FAVC": "yes" as 0, "no" as 1,

"SMOKE": "yes" as 0, "no" as 1,

"SCC": "yes" as 0, "no" as 1,

"CALC": "Always" as 3, "Sometimes" as 2, "Frequently" as 1, "no" as 0 ,

"MTRANS": "Walking" as 0, "Automobile" as 1, "Public_Transportation" as 2, "Bike" as 3, "Motorbike" as 4.

And then, the datatype was sorted just in the numerical form as shown in fig. 3.6.

Gender	int64
Age	float64
Height	float64
Weight	float64
family_history_with_overweight	int64
FAVC	int64
FCVC	float64
NCP	float64
CAEC	int64
SMOKE	int64
CH20	float64
SCC	int64
FAF	float64
TUE	float64
CALC	int64
MTRANS	int64
NObesidad	int64
BMI	float64
RISK	int64
..	..

Figure 3.6: Datatype

3.2 Feature Selection

3.2.1 Borutashap Feature Extraction

BorutaShap is a features extraction wrapper technique that combines the Boruta feature selection algorithm and Shapley values. This combination outperforms the original Permutation Importance technique in terms of both the performance and accuracy of the feature subset produced. This technique provides a superior subset of features and provides the most reliable and timely global feature rankings, which may be utilized for model inference. Unlike the original R package, which only enables the user to pick a Random Forest model as the base model in the feature selection process, BorutaShap allows the user to pick any Tree-Based learner as the basic model. As BorutaShap’s improvements in performance, the SHAP Tree-Explainer scales proportionally to the number of observations, making it difficult to employ with massive datasets. BorutaShap features a sampling technique that employs the lowest possible subsample of the data available at each step of the algorithm to fight this. Some steps that this extraction process follows are:

1. Create new copies of all of the other features in the data set, label them *shadow + feature* name, and then shuffle them to eliminate their correlations with the dependent variables
2. Attributes with a significance below the limit are termed ‘unimportant,’ and they are deleted from the process. Consider the traits that are much more essential than the criterion to be ‘important.’
3. Remove all shadowed features and continue the operation till every feature has been assigned a significance or the algorithm has reached the number of runs specified before.

It is calculated independently for each tree in the forest that uses a specific property to classify itself. The accuracy loss' average and standard deviation are then calculated. The Z score, which is calculated by dividing the average loss by its standard deviation, may also be used as a significance metric. As the general formula of the z-score:

$$z = \frac{(x-\mu)}{\sigma}$$

Nonetheless, in Boruta, we choose the Z score as the key metric since it accounts for the changes in mean accuracy loss across forest trees as we used a Random Forest classifier for this extraction process. The method determines if the feature Z-scores were improved by comparing the shuffled versions of the features to the original features at each iteration. In this case, the algorithm classifies the attribute as being required. To summarize, shown in fig. 3.7 the method measures the feature's importance against random samples shuffled, helping the software maintain robustness. To identify the features that outperformed the shadows, a binomial distribution is used.

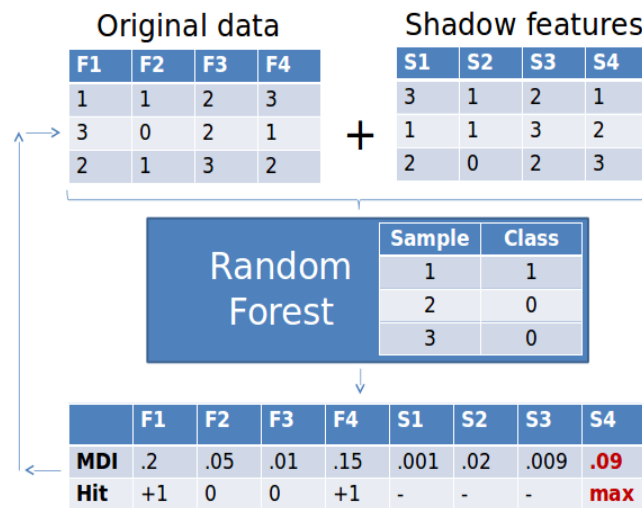


Figure 3.7: Feature Extraction using BorutaShap

3.3 Feature Analysis

In our research, the purpose is to identify the vulnerable person to become obese shortly. Also, to build a model which will enable the existing systems to identify obese people. We make a questionnaire of Over 17 attributes where the participants have to answer in a binary form. Some essential features have been identified later by running some algorithms like Logistic regression, Random Forest, Linear Regression, and K-nearest neighbor (KNN) put high value in identifying the prospective drug addicted person. We get the histogram as hown in fig.3.8, the Weight count from our dataset that shows the percentage of people weighted below or above the line.

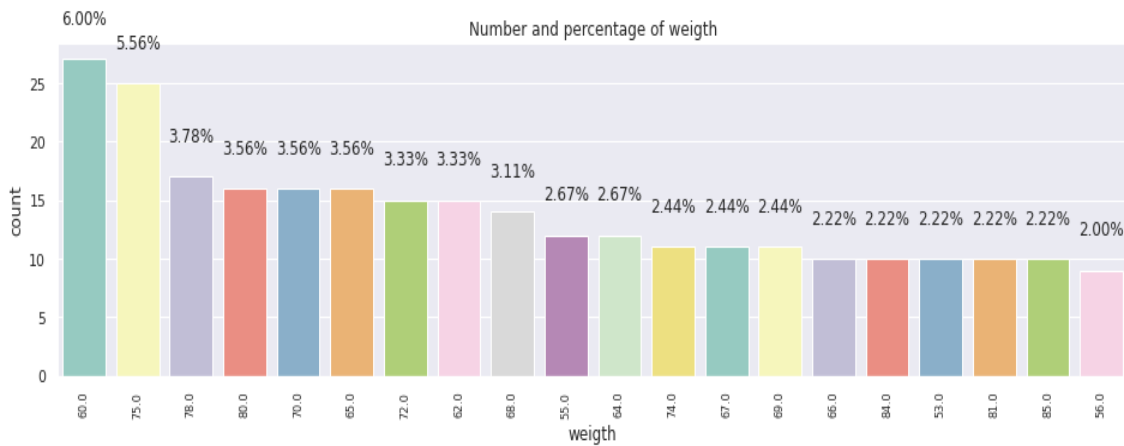


Figure 3.8: Number and Percentage of Weight

As we also collected the data according to their drinking habits and their meat consumption in fig.3.9. According to the data, fast food intake appears to be linked to a rising incidence of severe obesity. Given that morbid and super morbid obesity rates are increasing faster than moderate obesity, it is essential to investigate the behavioural factors linked to these trends.[16] From our Dataset, we get that meal consumption is 3 in most cases, but most people often get a meal between 1 and 2. Then we decided to collect the destructive factors that also affect the BMI of the

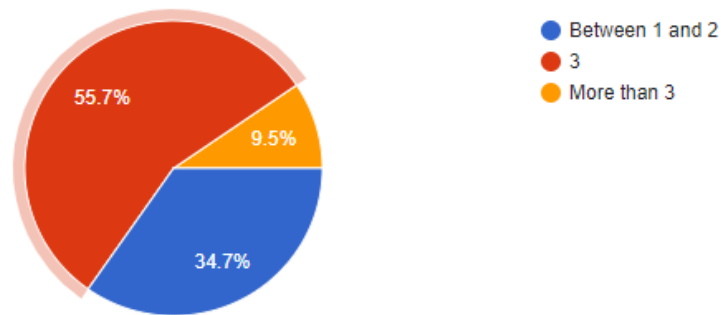


Figure 3.9: Meal Consumption

human body. In fig.3.10 smoking percentage is shown. Obesity is more common in males, and some studies have found that smoking is only a risk factor for males. Obesity, smoking, and physical inactivity were studied concerning chronic kidney disease (CKD) in the general population and if CKD risk was limited to males.[16] Most men of our country are usually involved in smoking, and also few are also involved in drinking.

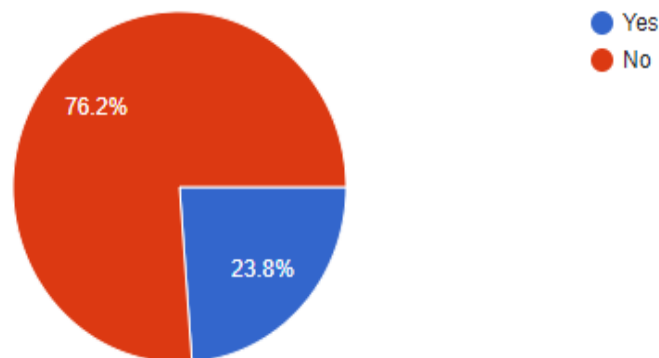


Figure 3.10: Smoking Percentage

As we can see that very few numbers of people in our country are also involved in drinking shown in fig.3.11, but a small group of people is always involved in it. In both genders, the fraction of energy underreporting dropped as the amount of alcohol consumed increased. Multiple logistic regression studies that accounted for energy underreporting, smoking, leisure-time physical activity, energy, and food quality found that drinking more than three glasses of alcohol was substantially related to the risk of abdominal obesity.[26]

According to our Dataset, we found out that a lot of the population does not count

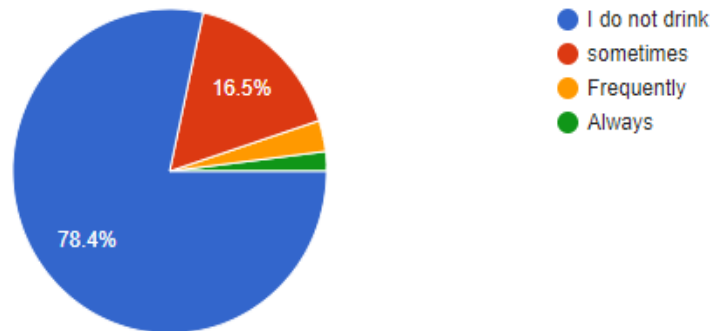


Figure 3.11: Alcohol Consumption

the amount of protein or calorie consumption in their daily life shown in fig.3.12 and the water consumption part too in fig.3.13. Because in our country people eat a lot of different things which makes them more affected to obese. Calorie measurement is necessary. Lifestyles, especially food patterns, have altered dramatically over the world in the last 20–30 years, accompanied by a global obesity pandemic. While physical activity has dropped in many places, particularly in low-income nations, the best available data shows that total physical activity has stayed consistent or even grown in high-income nations like the United States. This suggests that in the United States, a worsening diet is the primary cause of obesity. However, in most low-income nations, a combination of decreasing physical activity and a worsening diet is more likely. As nutrition is a significant risk factor for obesity, it is critical to understand how it causes weight gain. Obesity, according to popular belief, is caused by a calorie and energy imbalance, in which people consume too many calories while moving insufficiently. This viewpoint has influenced both scientific research and government policy. For decades, low-fat diets were advocated for obesity prevention merely because fat has a more extensive energy content per gram (9 kcal/g) than carbs and proteins (4 kcal/g). [21]

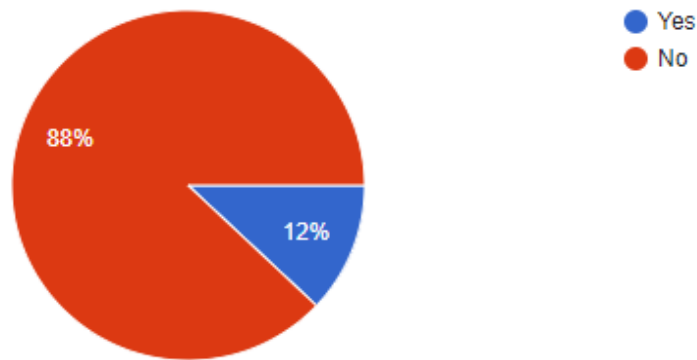


Figure 3.12: Calorie Consumption

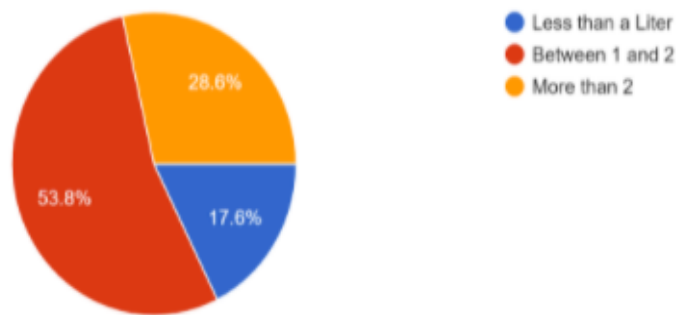


Figure 3.13: Water Consumption

In our present generation, the young generation is always involved in their mobile phones or their electronic devices such as gaming components and their laptop or their PC and recently a new evolution of individual components such as Tablets and many more shown in fig.3.14. People usually spend most of the time using them and making themselves idler. From our dataset, we can see that around 67.2% of people use their phone more than 5 hours which may be a considerable lead to obese and only a very few percentages of people are using it for less than 2 hours.

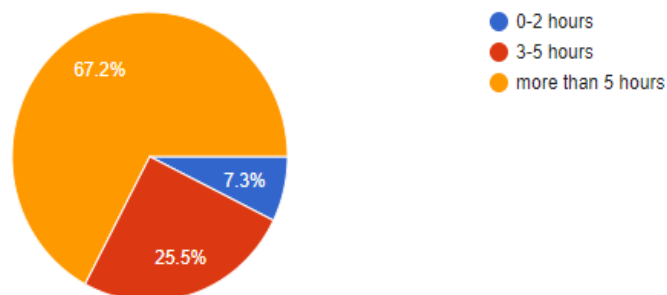


Figure 3.14: Electronic device use

The fig.3.15 shows the histogram of the Obesity level vs age of our data and the Obesity vs Height and Obesity vs Weight of our dataset.

The below situation in fig.3.16 describes the calorie intake monitoring of a person

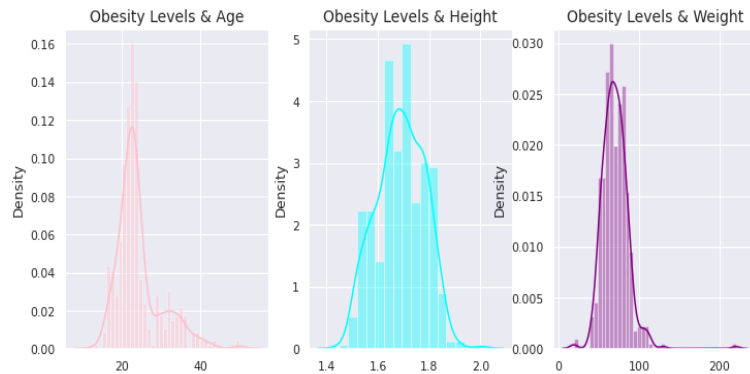


Figure 3.15: Comparison of Obesity vs Age, Height, Weight

and in fig.3.17 shows consumption of meal, as calorie intake is highly involved with the obese these days. Nutrition was calorie-dense but micronutrient-deficient before nutritional management. In BMC, blood concentrations of 25-hydroxyvitamin-D, vitamin C, selenium, iron, and β -carotene, vitamin C, and lycopene were found to be deficient at the foundation. Even more, participants reported lower micronutrient intake of vitamin C (serum, BMC), zinc, and lycopene after a three-month formula diet. Lipophilic blood vitamin contents and iron and C-reactive protein had a strong negative connection with body fat. [5]

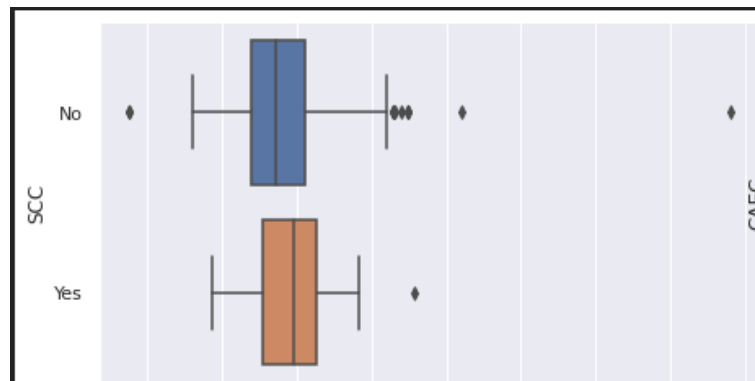


Figure 3.16: Calorie consumption monitoring

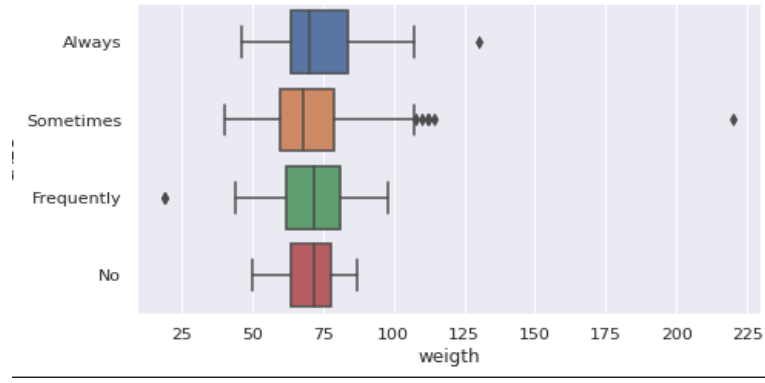


Figure 3.17: Consumption of Meals

3.3.1 Heatmap of Data

A heat map (or heatmap) is a two-dimensional data visualization tool that displays the magnitude of phenomena as colour. The colour fluctuation might be via hue or intensity, providing the reader clear visual indications about how the occurrence is clustered or evolves. As it depends solely on the correlation of the data. The correlation of our data is shown in fig.3.18

	Gender	Age	Height	...	MTRANS	Nobeyesdad	RISK
Gender	1.000000	0.091026	-0.557838	...	-0.026112	-0.070708	0.006774
Age	0.091026	1.000000	0.026637	...	0.198294	0.102513	0.064041
Height	-0.557838	0.026637	1.000000	...	-0.013431	-0.112859	-0.120208
weight	-0.310069	0.107087	0.394335	...	0.043036	0.774069	0.529254
BMI	-0.054576	0.113303	-0.101372	...	0.070238	0.919938	0.655638
Family_history	-0.042392	-0.042272	-0.032546	...	0.068947	-0.143937	-0.096679
FAVC	0.013538	-0.038864	-0.094969	...	-0.075663	-0.142035	-0.158074
FCVC	0.023743	0.099462	0.047625	...	0.053945	0.030671	0.016751
NCP	-0.056887	0.019064	0.105068	...	-0.053445	0.090511	0.051864
CAEC	-0.055878	-0.020968	0.063550	...	-0.012152	0.036032	0.064676
SMOKE	0.202208	0.006753	-0.124930	...	-0.130229	-0.072845	-0.004829
CH2O	-0.123937	0.107965	0.136009	...	0.061546	0.043455	0.049001
SCC	-0.088611	-0.069833	0.006359	...	0.121038	-0.061108	0.017523
FAF	-0.033626	0.156958	0.092176	...	-0.030998	-0.044973	-0.121957
TUE	0.001406	-0.050790	0.043503	...	-0.075126	0.135338	0.098148
CALC	0.044225	0.295471	0.113091	...	0.165677	0.170700	0.051970
MTRANS	-0.026112	0.198294	-0.013431	...	1.000000	0.057092	0.054217
Nobeyesdad	-0.070708	0.102513	-0.112859	...	0.057092	1.000000	0.736990
RISK	0.006774	0.064041	-0.120208	...	0.054217	0.736990	1.000000

Figure 3.18: Correlation of Data

Two essentially distinct forms of heat maps: the cluster heat map and the spatial heat map. Magnitudes are placed in a fixed cell size matrix whose rows and columns represent discrete phenomena and categories. The ordering of rows and columns is purposeful and somewhat random, intending to identify clusters or show them as determined by statistical research as shown in fig.3.19.

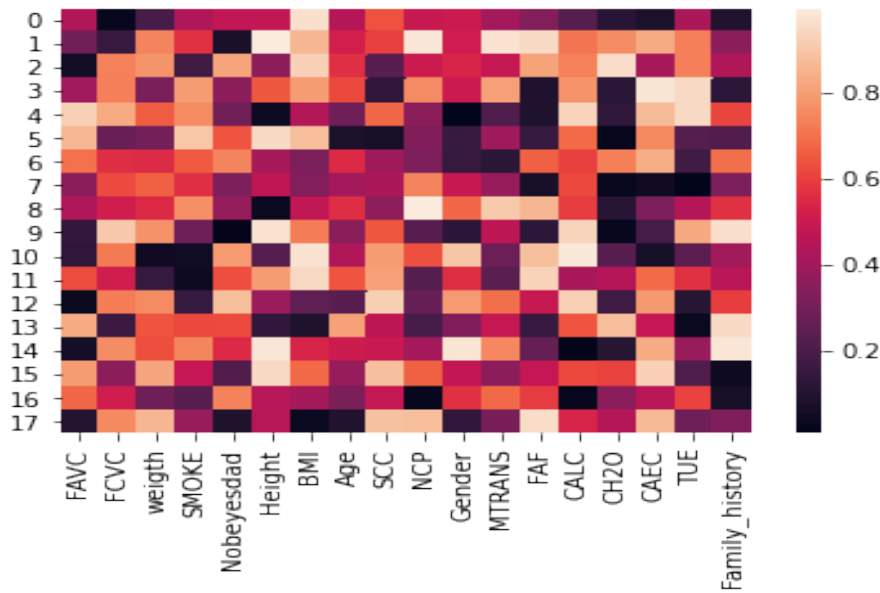


Figure 3.19: Heatmap of Dataset

The cell size is variable, but it must be large enough to be seen. The position of a magnitude on a spatial heat map, on the other hand, is determined by its location in that space, and there is no concept of cells; the phenomena are assumed to change constantly. We have generated the heat Map of our data. A heat map is one of the best visual representations for a large dataset. A different shade explains the difference in the correlation in our scattered dataset. We have around 17 columns and their questionnaires along with them. As we are also comparing with the dataset of other regions, the correlation of that dataset is also shown in the heatmap and the correlation in fig.3.20 to make it more straightforward and visually understandable, as the heatmap is one of the best approaches.

	Gender	Age	...	BMI	RISK
Gender	1.000000	-0.048394	...	0.053035	-0.000442
Age	-0.048394	1.000000	...	0.244163	0.202689
Height	-0.618466	-0.025958	...	0.131785	0.135505
Weight	-0.161668	0.202560	...	0.934806	0.734576
family_history_with_overweight	0.102512	-0.205725	...	-0.483508	-0.390373
FAVC	0.064934	-0.063902	...	-0.246097	-0.249334
FCVC	0.274505	0.016291	...	0.263651	0.170898
NCP	-0.067600	-0.043944	...	0.039969	0.049479
CAEC	0.091543	-0.083739	...	-0.313278	-0.217127
SMOKE	0.044698	-0.091987	...	0.000819	0.035262
CH20	-0.107930	-0.045304	...	0.144200	0.090598
SCC	-0.102633	0.116283	...	0.184286	0.188339
FAF	-0.189607	-0.144938	...	-0.177537	-0.133609
TUE	-0.017269	-0.296931	...	-0.099720	-0.077609
CALC	0.030571	0.002455	...	0.229584	0.141751
MTRANS	0.112189	-0.425414	...	0.088297	0.068353
NOBeyesdad	0.031464	0.282913	...	0.977826	0.797062
BMI	0.053035	0.244163	...	1.000000	0.773249
RISK	-0.000442	0.202689	...	0.773249	1.000000

Figure 3.20: Correlation of Comparing data

3.3.2 Approach for Feature Selection

From our extensive research, we have been able to identify several universal characteristics. Most of the papers listed many additional features, but to produce the optimum outcome, we had to select only a small number of them. Many data dimensionality reduction methods are routinely utilized, such as Principal Component Analysis, Singular Value Decomposition, etc. Given this information, you must wonder why you would ever need to use any other features extraction approaches. Unsupervised feature selection approaches have problems with these techniques: As an example, we may look at PCA, which detects the component using the difference in data variance. These solutions disregard the relationship between extracted characteristics and the target class or values. These processes also require adjustments before being employed because certain assumptions, such as normality, are connected. Restrictions on data do not apply to all forms of data.

The use of filters in the preprocessing process is rather widespread. Feature selection has no bearing on machine learning algorithms of any kind. Instead, attributes are selected based on their connection to the outcome variable, which is demonstrated through various statistical tests. Additional typical filter approaches include Pearson, Spearman, Distance, Chi-Squared test, and Anova.

Here, Forward Selection and Backward Elimination are but a few of the many wrapper approaches that can be used.

As, we used the BorutaShap Model of extracting our data which is also a wrapper method of training a dataset shown in fig.3.21. As from our selected features, the Borutashap algorithm extracted 8 of the essential features from 17 features of the dataset and the remaining 10 of them as non-important features as the pseudocode describes the nature of this process.

```
from BorutaShap import BorutaShap
from sklearn.ensemble import RandomForestClassifier

df2 = pd.read_csv('dataset.csv')
X = df2.drop(['RISK'],axis=1)
y = df2['RISK']
model = RandomForestClassifier()

# no model selected default is Random Forest, if classification is
False it is a Regression problem
Feature_Selector = BorutaShap(model=model, importance_measure='shap',
classification=True)

Feature_Selector.fit(X=X, y=y, n_trials=100, random_state=0)
```

Figure 3.21: BorutaShap Feature Extraction

After this process, our desired result is achieved, and we get all the essential features that we needed for this study to make sure the research is conducted in an exemplary manner. The importance of the features are shown in fig.3.22. We also conducted the feature importance test to prove the features that we used to make our final decision.

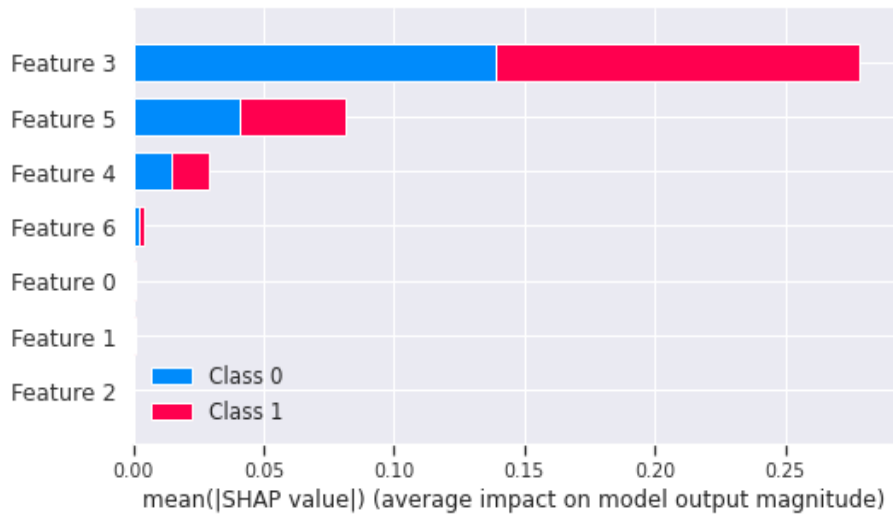


Figure 3.22: Feature Importance of our data

As of now, for the filtering process, we faced several challenges. Because many people gave many dummy inputs and did not give any input, we had the best members to cover them up and correct them. As the persons who filled up the form given their Height in Feets, but for our study to conduct, we needed them in Centimeters, so our members converted it using machine techniques. Furthermore, to convert our work in the best way, we needed the value for BMI and the Level of obesity. We also calculated them and also by levelling the highest from the lowest. Moreover, the risk factors were also denoted by their side, and they were levelled as “Risk” and “no Risk”. As from the feature extraction process, we extracted the target features of the other dataset as per fig.3.23, too, with which we are about to compare our work.

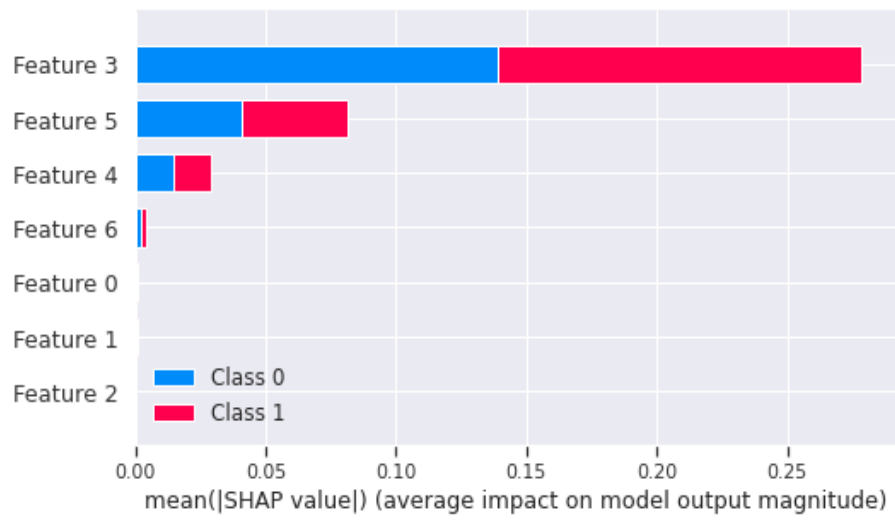


Figure 3.23: Feature Importance of the other dataset

Chapter 4

Model Selection and Result Analysis

4.1 Machine Learning

Machine learning is a rapidly emerging area of computing where computers are trained to do things that are thought to be intelligent by human beings. Larger quantities of data have become workhorses in the present era of so-called big data. A variety of industries are using machine learning to improve efficiency and save money, including computer vision, pattern recognition, spacecraft engineering, banking, and entertainment.[9] Machine learning algorithms are employed in numerous different applications. Artificial Intelligence is applicable in a wide range of domains, including image analysis and computer vision. Explicitly describing how something should be done is seldom a good idea. Machine Learning is almost always recommended. Advances in technology have changed the current state of machine learning. The researchers conducted experiments to investigate if the computer could learn without being taught to carry out a specific activity. While algorithms are trained using labelled instances, such as an intake where the outcome is known, supervised learning or unsupervised learning may be employed. Failure and operational data items may be labelled "F" (Failed) or "R" (passed). An algorithm learns by comparing its output to the correct outputs to identify faults. Once it has done that, it will apply the changes to the model. The classification, regression, prediction, and gradient boosting methods are used to anticipate label values on data that has not yet been tagged by predicting what the values will be. Under these conditions, supervised learning is usually used. In conclusion, based on our investigation, we found that we do supervise the learning of anticipating the effect of obesity within our region.

4.2 Supervised Learning

Supervised machine learning design focuses on creating algorithms capable of developing broad pattern and hypotheses by using previously given instances to forecast future examples. In fig.4.1, classification algorithms that use supervised learning attempt to use previously learned knowledge to categorize new data. In data science challenges, classification is used rather frequently. For tackling such issues, several practical strategies have been presented, including rule-based approaches, logic-based approaches, instance-based approaches, and stochastic strategies. The usefulness of supervised machine learning algorithms is discussed in terms of accuracy, learning speed, complexity, and the danger of overfitting metrics in this study.[8] There are various supervised learning algorithms available, each with its own set of advantages and disadvantages. There is no one-size-fits-all learning algorithm that can solve all supervised learning issues. If you have a set of \mathbf{N} training examples in the form $\{\{x_1, y_1\}, \dots, \{x_N, y_N\}\}$, such that x_i is the feature vector of the i -th example and y_i is its label (i.e., class), a learning algorithm seeks a function $g: X \rightarrow Y$, where X is the input space, and Y is the output space. The function g is a component of the hypothesis space, which is a collection of alternative functions \mathcal{G} . The risk $R(g)$ of function g is defined as the expected loss of g . It can be assumed from the training data as.

$$R_{emp}(g) = \frac{1}{N} \sum_i L(y_i, g(x_i))$$

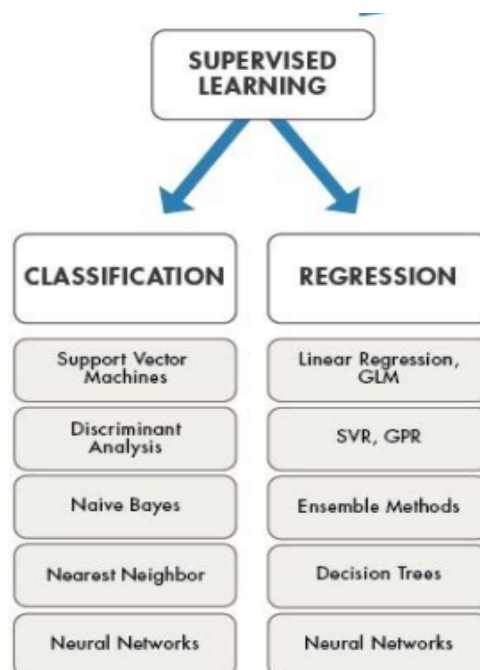


Figure 4.1: Supervised learning

The training algorithms in supervised learning require labelled inputs and outputs, whereas the training algorithms in unsupervised learning do not. This algorithm "learns" by repeatedly making a prediction on the training dataset and fine-tuning those predictions based on how accurate they are in supervised learning.

4.3 Naive Bayes Algorithm Implementation

The Bayes Theorem is utilized in a wide range of various classification problems, including the Naive Bayes method. The Naive Bayes algorithm demonstrates the basic approach of generating classifiers using Naive Bayes. Instead of building a single algorithm to build classifiers, a broad range of algorithms with similar principles form a family. In our research, we used the Gaussian Naive Bayes method in particular and discussed the results. We are all familiar with the Bayes Theorem, which goes like this:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

In this formula, y stands for class, and X stands for data. In this case, we have one observation to predict and two possible classes: RISK, no - RISK. As a result, we will compute two posteriors: one for Obese people and. When working with continuous data, it is common to assume that the continuous values associated with each class follow a standard (or Gaussian) distribution.

$$P(y|x_1, \dots, x_n) = \frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)P(y)}{P(x_1)P(x_2)\dots P(x_n)}$$

Here x_1, x_2, x_3, \dots are the features that we used here. Assume the training data includes a continuous attribute called x. The mean and variance of x are computed in each class after the data is segmented by class. As for our study, we selected 8 features like the essential features, and then we calculated the accuracy through this algorithm. A test_size of 30% was taken from our dataset, and the other was for the training of the data. We were able to get an accuracy of 90.5% by the selected features that we used. It yielded one of the best results that we achieved throughout our study. On the other dataset that concerns Colombia, Peru, and Mexico, we were able to get an accuracy of 88.64% by the newly selected data according to that dataset. The model gave a decent accuracy for our dataset, but for the other dataset, the accuracy was a bit lower than was expected.

4.4 Random Forest Model Implementation

The first time that Ho proposed using random decision forests was in 1995.[18] He discovered that forests of trees might have more accurate split with oblique hyperplanes if the dimensions of the forest were varied randomly, as long as those dimensions were only sensitive to a few features. A Random Forest classification/regression technique leverages the ensemble learning method. Instead of using a boosting strategy, random forest employs bagging. Random forests contain trees that grow side-by-side as shown in fig4.2. At the moment, the trees are being built, and they are not interacting with each other. It models using many decision trees, then determines the class that is the most commonly found among the groups (classification) or the individual trees' mean prediction (regression). A random forest is a "meta coordinate estimator" (i.e., it aggregates the findings of numerous estimators by bringing in individual trees).

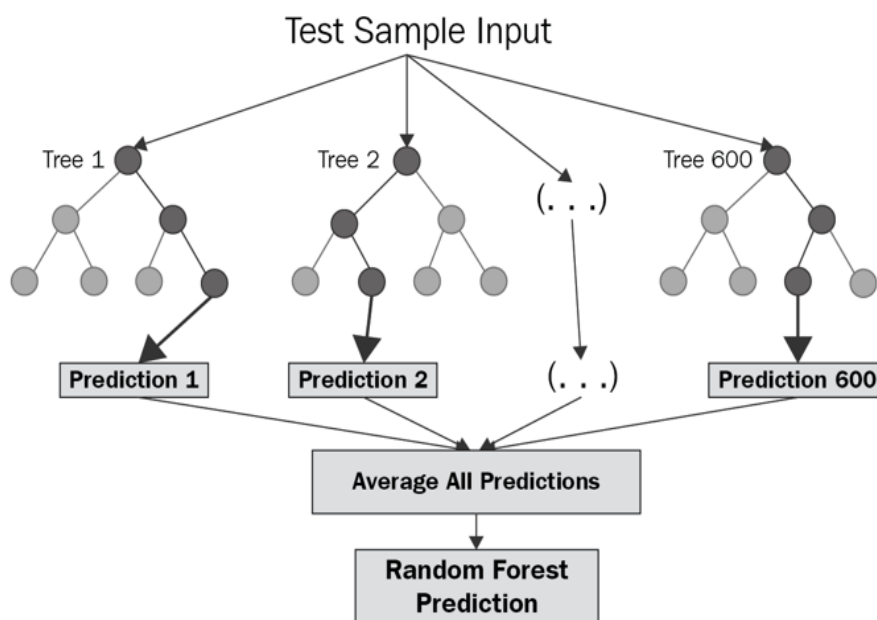


Figure 4.2: Random Forest Structure

When splitting features on a node, a particular percentage (the hyperparameter) of the total features is allowed to be used. Ensemble methods aim to avoid reliance on any single feature, and they need to assess all possibly predictive features equally. Using random samples adds an element of randomness that inhibits overfitting. Given a training set $X=x_1, x_2, x_3, \text{ etc.}, x_N$, and answers $Y=y_1, y_2, y_3, \text{ etc.}, y_N$. We select random samples from the training set.

$$\hat{f} = \frac{1}{N} \sum_{i=1}^N f_i(x')$$

and fit trees to them numerous times. The decision tree for the variables x_i and y_i is $f(x_i, y_i)$. Finally, we may average the outcomes of all the trees that correspond to x' to forecast the outcomes for x' (in continuous). According to this algorithm we

considered the `n_estimators = 70` and `min_sample_leaf = 30`. The accuracy that we yielded is 95.56%, which is the second-highest in this study that we conducted; on the other dataset, we yielded 94.321% accuracy.

4.5 KNN Algorithm Implementation

One of the most fundamental machine learning approaches is the KNN algorithm. Many people in the business love it because of how simple it is to use and how fast it calculates. The k-Nearest-Neighbors (KNN) approach is a basic yet successful classification approach. The two most significant disadvantages of KNN are

- (1) It is poor efficiency (being a lazy learning approach precludes it from being used in many applications like dynamic web mining for a large repository) and
- (2) Its reliance on the selection of a “good value” for k.

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

The fundamental mathematical notions that are embedded in KNN are analogous to many other things in life. KNN is created by first identifying the various characteristics in the data and then converting those values into vectors (or numbers). After these mathematical numbers are calculated, the algorithm estimates the distance between them. In the example shown below, the Euclidean distance is employed most often to calculate this distance. KNN employs this formula to calculate both measurement result and test data. This calculation results in determining which points in fig.4.3 given as X and Y, are the most comparable to the test data, and those points are classed accordingly.

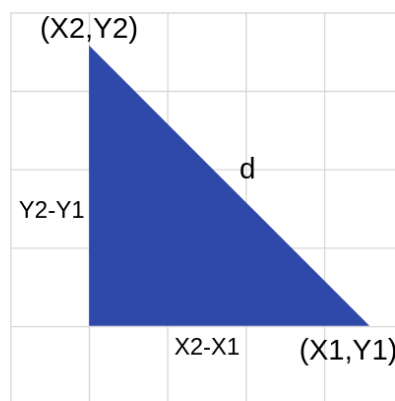


Figure 4.3: Visualization of the formula

Simple optimization techniques, image recognition technologies, and decision-making models frequently employ KNN. The naïve form is simple to construct since it computes the distances between the test example and all stored instances. However, it is computationally costly when dealing with large training sets. Even for enormous data sets, k-NN is computationally tractable because of an approximate closest neighbor search technique. Over the years, a slew of nearest neighbor search algorithms has been presented. We extracted our features according to the content for our study, and we achieved an accuracy of 91.11%. Furthermore, for the other dataset, we achieved an accuracy of 93.21%. We estimated the `n_neighbors` as 15.

4.6 Logistic Regression Implementation

A binary predictor variable is modelled by using a logistic function to approximate a logistic function. Nevertheless, this is not the only advanced variation. Logistic regression (or logit regression) is a technique used in regression analysis to estimate the parameters of a logistic model (a form of binary regression). Logistic regression (LR) finds a formula that predicts an output for a binary variable, Y , from one or with a dependent variable, X .

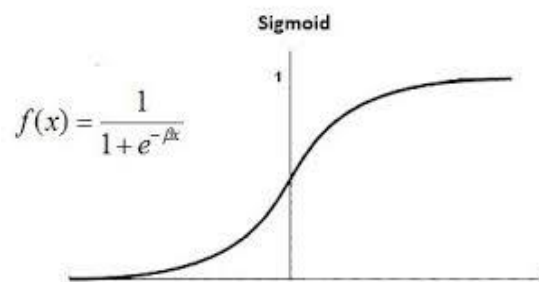


Figure 4.4: Logistic Regression

Logistic function, or sigmoid function shown in fig.4.4, was developed by mathematicians to describe how populations increase in ecology, for example, at a rapid rate and peak before the carrying capacity is reached. The curve can accurately and accurately transfer any real-valued integer between 0 and 1 but cannot precisely and accurately transfer between those two locations.

$$\frac{1}{1 + e^{-value}}$$

To calculate the logarithm of e to base e (Euler's number or the Excel `EXP()` function), you must start with the logarithm of e (the number or value you want to convert) and then perform the calculation. It would anticipate values for the default class and the other class near each other: this would yield the optimal parameter settings. The key feature of logistic regression is maximizing the likelihood of the coefficients to result in the projected probabilities fitting the actual data. We extracted our features according to the content for our study, and we achieved an accuracy of 91.11%. Moreover, for the other dataset, we achieved an accuracy of 93.84%.

4.7 Decision Tree Implementation

The Decision Tree method is included in the supervised learning algorithms family. A key benefit of the decision tree learning methodology is that it can also be applied to regression and classification problems. The induction of decision trees, also known as decision tree learning, is one of the statistical approaches used in data mining, machine learning, and statistics. After making observations about an object, one moves on to determining the worth of the item's goal using a decision tree (represented in the leaves) shown in fig.4.5. We will use the Decision Tree, which has two categories because there are two sorts of decision trees. A Categorical variable decision tree is termed that.

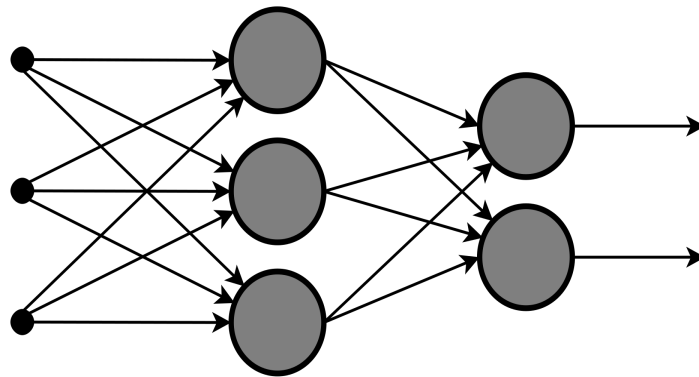


Figure 4.5: Decision tree

A decision tree is a network in which the root node, branches, and leaf nodes are linked together. The attribute tests are represented by internal nodes, the test results by branch nodes, and the class labels by leaf nodes. The root node is the topmost node in the tree. Decision Tree Classifier, like other classifiers, accepts two arrays as input: a sparse or dense array X of the form $(n_samples, n_features)$ containing the training samples, and an array Y of integer values of the form $(n_samples, n_labels)$ containing the class labels for the training samples.

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

The frequentist probability of an element/class I in our data is just ' P_i .' Let us pretend we have two classes, a positive class and a negative class, for simplicity. As a result, I might be either $+$ or $-$. ($-$). We extracted our features according to the content for our study, and we achieved an accuracy of 96.67%. Furthermore, for the other dataset, we achieved an accuracy of 94.4%.

4.8 Results and Analysis

Following the development of the model, its performance was assessed to see how well it might predict the Obesity Risk of our region compared to other regions, including Colombia, Peru, and Mexico. The performance evaluation metrics we employed were based on four factors of a matrix of perplexity. True positive, False positive, True negative and False-negative were the parameters. The number of accurately anticipated observations was expressed as a true negative. The ratio of accurately predicted samples to total samples was used to measure accuracy. The following equation was used to calculate the classifier's correct prediction rate:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

As per our algorithm, we got the best results in the decision tree algorithm shown in fig.4.6 as the entire training set is regarded as the root at first. Then by analyzing its features, we were able to get the most accuracy for this algorithm in our study.

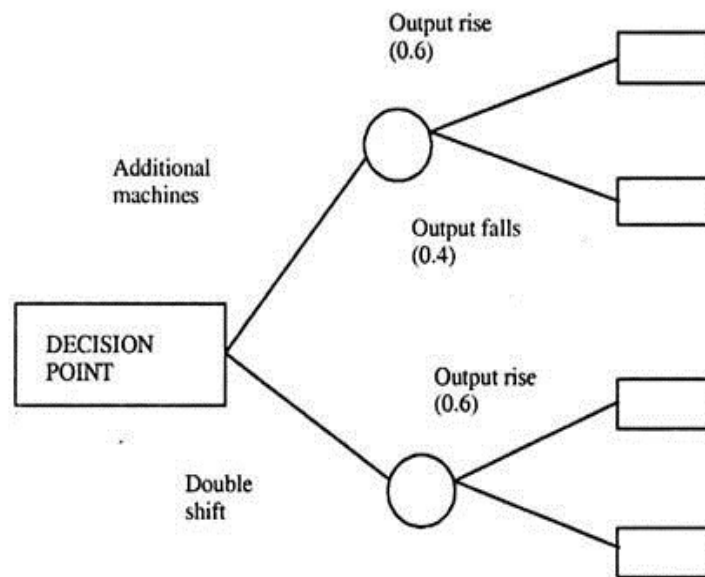


Figure 4.6: Decision tree accuracy prediction

As each leaf calculates its output accuracy, the other falls behind due to the lower one. Thus this process brings the best result of accuracy into our study. In the below fig.4.7, we showed the accuracy and the respective algorithms on which we have worked.

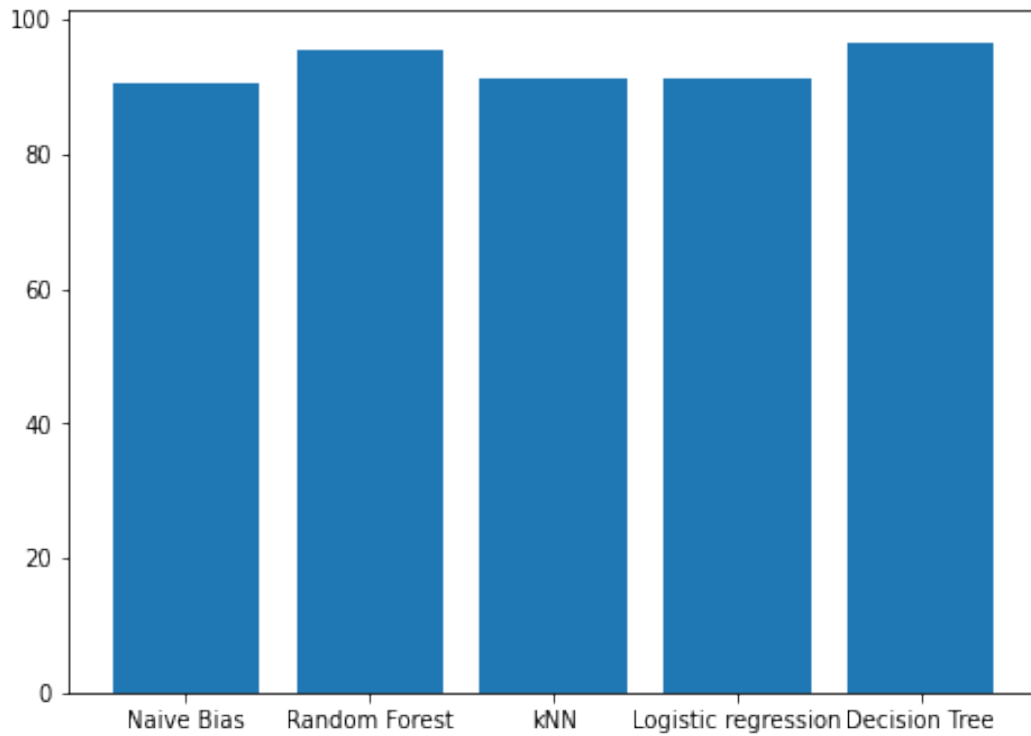


Figure 4.7: Accuracy Chart

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

Precision also denoted the proportion of accurately predicted positive observations to the total number of positive observations in the test set. A true positive is when the model predicts the positive class adequately. On the other hand, a real negative results in which the model correctly predicts the negative class. Furthermore, a false positive occurs when the model forecasts the positive class inaccurately. A false negative is an outcome in which the model forecasts the negative class inaccurately. As of now, we show the TN, FP, FN, and TP of our algorithms that we got in this study for each algorithms in fig.4.8, fig.4.9, fig.4.10, fig.4.11 respectively.

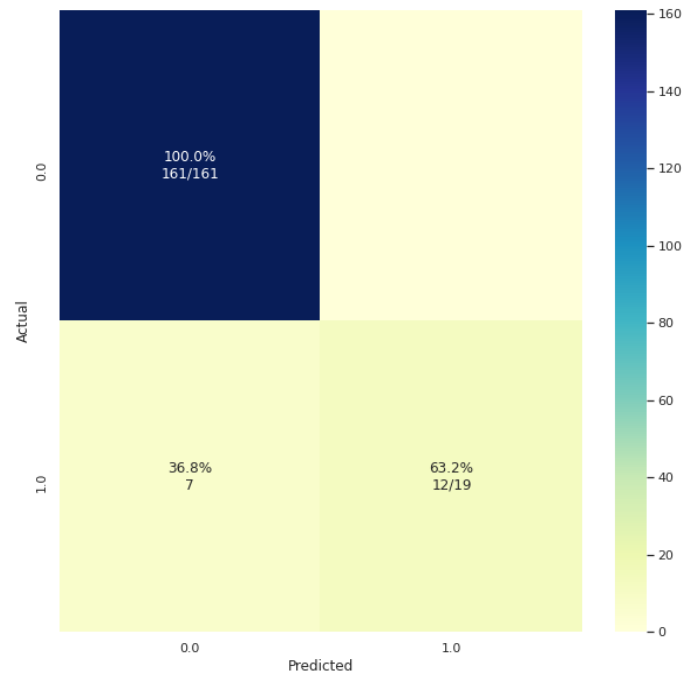


Figure 4.8: Confusion Matrix of Decision tree

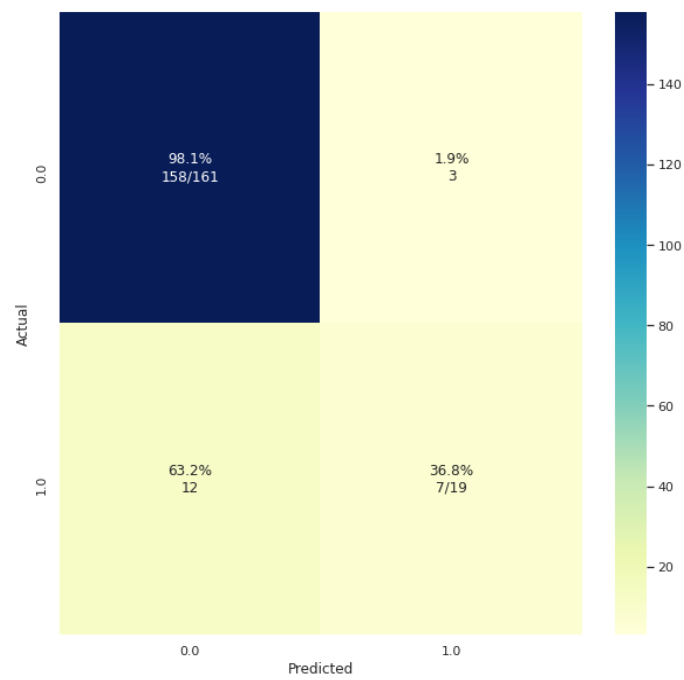


Figure 4.9: Confusion Matrix of k-nearest neighbor

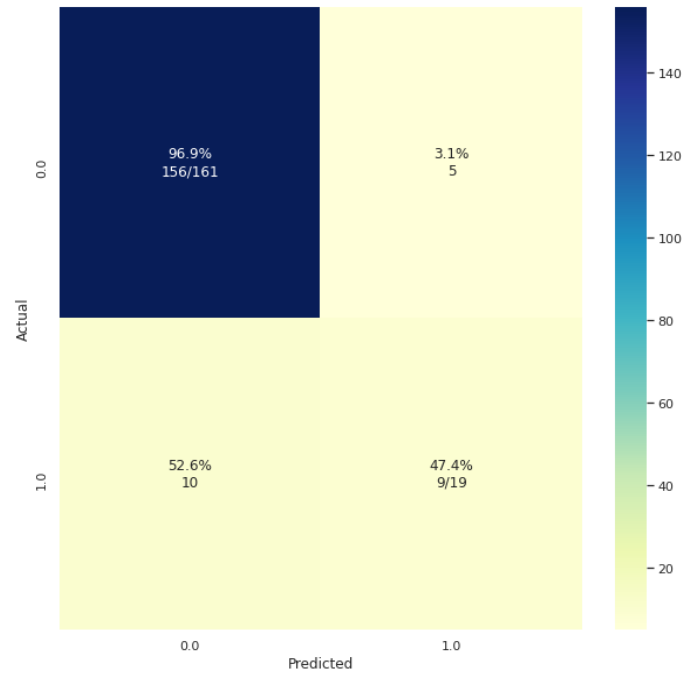


Figure 4.10: Confusion Matrix of Naive Bayes

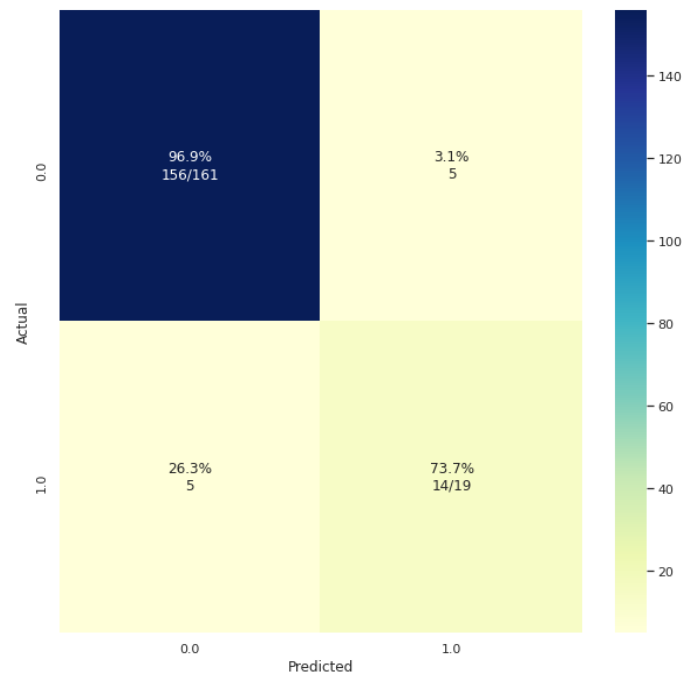


Figure 4.11: Confusion Matrix of Random Forest

Chapter 5

Conclusion and Future Work

A person who has a healthy BMI have a chance of healthy and improved life. From research, we know that if the BMI is higher, the risk of certain diseases like heart disease, high blood pressure, diabetes and cancer is also higher. BMI can help to determine the risk of diseases, and people can work on that to reduce the risk. In Bangladesh, the obesity rate is rising gradually. Among adults, the obesity rate became 2% to 4%. We can see many people suffering from heart disease, high blood pressure, cancer in Bangladesh nowadays. From our data, we can see that 6% of the population are nearly 60 kg. Which are the highest percentage among the count and 2.22% people are near 85 kg. We can also see the connection of fast food intake with the increase in obesity. Obesity is more common in men compared to female, many of them have a habit of smoking, and some are associated with drinking, which is also connected with high BMI. The calorie intake also has an impact on obesity.

In our study, we have used many methods. They are Logistic regression, Naive bayes, KNN, Decision tree, Random forest. We found decision tree algorithm is the best suitable model for our study as we achieved an accuracy of 96.67%, which is the highest. The remaining algorithm also showed prospective study results. For example, by the Random Forest algorithm, we were able to accomplish 95.56% accuracy. We collected primary data from a random person in our country through google form (more than 500), and many of them gave context type data; it was challenging for us to convert them. If we work on this study in future and collect data from hospitals, we can get more accurate data. Thus we will be able to accomplish the more accurate result.

Bibliography

- [1] B. J. Auerbach, F. M. Wolf, A. Hikida, P. Vallila-Buchman, A. Littman, D. Thompson, D. Loudon, D. R. Taber, and J. Krieger, “Fruit juice and change in bmi: A meta-analysis,” *Pediatrics*, vol. 139, no. 4, 2017.
- [2] C. S. Berkey, H. R. Rockett, M. W. Gillman, and G. A. Colditz, “One-year changes in activity and in inactivity among 10-to 15-year-old boys and girls: Relationship to change in body mass index,” *Pediatrics*, vol. 111, no. 4, pp. 836–843, 2003.
- [3] T. Berry, J. Spence, C. Blanchard, N. Cutumisu, J. Edwards, and C. Nyki-foruk, “Changes in bmi over 6 years: The role of demographic and neighborhood characteristics,” *International journal of obesity*, vol. 34, no. 8, pp. 1275–1283, 2010.
- [4] R. Calitri, E. M. Pothos, K. Tapper, J. M. Brunstrom, and P. J. Rogers, “Cognitive biases to healthy and unhealthy food words predict change in bmi,” *Obesity*, vol. 18, no. 12, pp. 2282–2287, 2010.
- [5] A. Damms-Machado, G. Weser, and S. C. Bischoff, “Micronutrient deficiency in obese subjects undergoing low calorie diet,” *Nutrition journal*, vol. 11, no. 1, pp. 1–10, 2012.
- [6] I. Dar-Nimrod, B. Cheung, M. Ruby, and S. Heine, “Can merely learning about obesity genes lead to weight gain,” *Appetite*, vol. 81, pp. 269–276, 2014.
- [7] W. Drøyvold, J. Holmen, K. Midthjell, and S. Lydersen, “Bmi change and leisure time physical activity (ltpa): An 11-y follow-up study in apparently healthy men aged 20–69 y with normal weight at baseline,” *International journal of obesity*, vol. 28, no. 3, pp. 410–417, 2004.
- [8] T. Dubowitz, M. Ghosh-Dastidar, C. Eibner, M. E. Slaughter, M. Fernandes, E. A. Whitsel, C. E. Bird, A. Jewell, K. L. Margolis, W. Li, *et al.*, “The women’s health initiative: The food environment, neighborhood socioeconomic status, bmi, and blood pressure,” *Obesity*, vol. 20, no. 4, pp. 862–871, 2012.
- [9] I. El Naqa and M. J. Murphy, “What is machine learning?” In *machine learning in radiation oncology*, Springer, 2015, pp. 3–11.
- [10] P. Ernsberger, “Bmi, body build, body fatness, and health risks,” *Fat Studies*, vol. 1, no. 1, pp. 6–12, 2012.
- [11] G. Garcia, T. S. Sunil, and P. Hinojosa, “The fast food and obesity link: Consumption patterns and severity of obesity,” *Obesity surgery*, vol. 22, no. 5, pp. 810–818, 2012.

- [12] W. G. Gathirua-Mwangi, T. W. Zollinger, M. J. Murage, K. R. Pradhan, and V. L. Champion, “Adult bmi change and risk of breast cancer: National health and nutrition examination survey (nhanes) 2005–2010,” *Breast cancer*, vol. 22, no. 6, pp. 648–656, 2015.
- [13] A. Golden and C. Kessler, “Obesity and genetics,” *Journal of the American Association of Nurse Practitioners*, vol. 32, no. 7, pp. 493–496, 2020.
- [14] A. B. Goldschmidt, V. P. Aspen, M. M. Sinton, M. Tanofsky-Kraff, and D. E. Wilfley, “Disordered eating attitudes and behaviors in overweight youth,” *Obesity*, vol. 16, no. 2, pp. 257–264, 2008.
- [15] D. M. Hall and T. J. Cole, “What use is the bmi?” *Archives of disease in childhood*, vol. 91, no. 4, pp. 283–286, 2006.
- [16] S. Hallan, R. de Mutsert, S. Carlsen, F. W. Dekker, K. Aasarød, and J. Holmen, “Obesity, smoking, and physical inactivity as risk factors for ckd: Are men more vulnerable?” *American journal of kidney diseases*, vol. 47, no. 3, pp. 396–405, 2006.
- [17] R. Haque, V. Chiu, K. R. Mehta, and A. M. Geiger, “An automated data algorithm to distinguish screening and diagnostic colorectal cancer endoscopy exams,” *JNCI Monographs*, vol. 2005, no. 35, pp. 116–118, 2005.
- [18] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, vol. 1, 1995, pp. 278–282.
- [19] M. A. Ihmels, G. J. Welk, J. C. Eisenmann, S. M. Nusser, and E. F. Myers, “Prediction of bmi change in young children with the family nutrition and physical activity (fnpa) screening tool,” *Annals of Behavioral Medicine*, vol. 38, no. 1, pp. 60–68, 2009.
- [20] C. H. Kroenke, R. Neugebauer, J. Meyerhardt, C. M. Prado, E. Weltzien, M. L. Kwan, J. Xiao, and B. J. Caan, “Analysis of body mass index and mortality in patients with colorectal cancer using causal diagrams,” *JAMA oncology*, vol. 2, no. 9, pp. 1137–1145, 2016.
- [21] D. Mozaffarian, “Foods, obesity, and diabetes—are all calories created equal?” *Nutrition reviews*, vol. 75, no. suppl_1, pp. 19–31, 2017.
- [22] M. Myrskylä and V. W. Chang, “Weight change, initial bmi, and mortality among middle-and older-aged adults,” *Epidemiology*, pp. 840–848, 2009.
- [23] F. Q. Nuttall, “Body mass index: Obesity, bmi, and health: A critical review,” *Nutrition today*, vol. 50, no. 3, p. 117, 2015.
- [24] F. M. Palechor and A. de la Hoz Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico,” *Data in brief*, vol. 25, p. 104344, 2019.
- [25] C. Roberts, N. Troop, F. Connan, J. Treasure, and I. C. Campbell, “The effects of stress on body weight: Biological and psychological predictors of change in bmi,” *Obesity*, vol. 15, no. 12, pp. 3045–3055, 2007.

- [26] H. Schröder, J. A. Morales-Molina, S. Bermejo, D. Barral, E. S. Mándoli, M. Grau, M. Guxens, E. de Jaime Gil, M. D. Álvarez, and J. Marrugat, “Relationship of abdominal obesity with alcohol consumption at population scale,” *European journal of nutrition*, vol. 46, no. 7, pp. 369–376, 2007.
- [27] D. H. Sinn, Y. W. Min, H. J. Son, P.-L. Rhee, S. W. Paik, S. N. Hong, and G.-Y. Gwak, “Metabolically-healthy obesity is associated with higher prevalence of colorectal adenoma,” *PloS one*, vol. 12, no. 6, e0179480, 2017.
- [28] S. Steinsbekk and L. Wichstrøm, “Predictors of change in bmi from the age of 4 to 8,” *Journal of pediatric psychology*, vol. 40, no. 10, pp. 1056–1064, 2015.
- [29] E. R. Sund, A. Jones, and K. Midthjell, “Individual, family, and area predictors of bmi and bmi change in an adult norwegian population: Findings from the hunt study,” *Social science & medicine*, vol. 70, no. 8, pp. 1194–1202, 2010.
- [30] A. Trinh, M. Campbell, O. C. Ukoumunne, B. Gerner, and M. Wake, “Physical activity and 3-year bmi change in overweight and obese children,” *Pediatrics*, vol. 131, no. 2, e470–e477, 2013.
- [31] B. H. Wrotniak, L. H. Epstein, R. A. Paluch, and J. N. Roemmich, “Parent weight change as a predictor of child weight change in family-based behavioral obesity treatment,” *Archives of pediatrics & adolescent medicine*, vol. 158, no. 4, pp. 342–347, 2004.
- [32] R. Zheng, M. Du, B. Zhang, J. Xin, H. Chu, M. Ni, Z. Zhang, D. Gu, and M. Wang, “Body mass index (bmi) trajectories and risk of colorectal cancer in the plco cohort,” *British journal of cancer*, vol. 119, no. 1, pp. 130–132, 2018.

Appendix A

Appendix

A.1 BorutaShap Code Sample

```
1. from BorutaShap import BorutaShap
   from sklearn.ensemble import RandomForestClassifier
   df2 = pd.read_csv('new - Copy.csv')
   X = df2.drop(['RISK'],axis=1)
   y = df2['RISK']
   model = RandomForestClassifier()
```

No model selected default is Random Forest, if classification is False it is a Regression problem

```
Feature_Selector = BorutaShap(model=model, importance_measure='shap',
classification=True)
```

```
Feature_Selector.fit(X=X, y=y, n_trials=100, random_state=0)
```

A.2 Comments from the panel members:

Mr. Rubayat Ahmed Khan:

- Rephrasing, Good work. Check if you have the best accuracy among the same criteria papers to publish.

Ms. Ipshita Bonhi Upoma:

- Rephrasing, Good work. You reviewed a lot of papers. There are some minor spelling mistakes/typos. Check them and you are good to go.

Ms. Ahanaf Hassan Rodoshi:

- Rephrasing, I think your presentation was good. Great job guys.