A PREDICTIVE STUDY ON OBESITY BASED ON HABITS AND PHYSICAL CONDITION

Project Report submitted to the

SDM COLLEGE(Autonomous)

in partial fulfillment of the degree of

MASTER OF SCIENCE

IN

STATISTICS

by

Dhanashree D

Under the guidence of

Ms. Supriya S.P.

Department of Postgraduate Studies in Statistics

SRI DHARMASTHALA MANJUNATHESHWARA

COLLEGE (Autonomous)

UJIRE - 574240

Karnataka, INDIA

SEPTEMBER 2022

SRI DHARMASTHALA MANJUNATHESHWARA COLLEGE

(AUTONOMOUS)

UJIRE - 574240

DEPARTMENT OF PG STUDIES IN STATISTICS

CERTIFICATE

Certified that this is the bonafide record of the project work done by Ms. Dhanashree D during the year 2022 as a part of her M.Sc (Statistics) fourth semester course work.

Reg. No.	**2**	**0**	**1**	**9**	**3**	**5**

Project Guide **Head of the Department**

Examiner

1.

2.

Place: Ujire
Date:

DECLARATION

I, Dhanashree D, here by declare the project report entitled **‘A PREDICTIVE STUDY ON OBESITY BASED ON HABITS AND PHYSICAL CONDITION’** is a bonafide record of project work carried out by me under the guidance and supervision of **Asst. Prof. Ms. Supriya S P**, Department of Statistics, SDM College, Ujire - 574240 Karnataka, India.I further declare that this project is the outcome of my own effort, that it has not been submitted to any other university for the award of any Degree.

Date: (Dhanashree .D)
Place: Ujire Email: dhanashree.d55gmail.com

CERTIFICATE

This is to certify that the project report entitled **‘A predictive study on obesity based on habits and physical condition’** is a bonafide record of an authentic work carried out by **Dhanashree D**, under my guidance and supervision in the Department of Post Graduate Studies and Research in Statistics, SDM College, Ujire, in partial fulfilment of the requirements for the award of the degree of Master of Science in Statistics, under Mangalore University, Mangalagangothri. I further certify that this report or part thereof has not previously been presented or submitted elsewhere for the award of any Degree, Diploma, Associateship, Fellowhsip or any other similar title of any other institution or university.

Date:

(Supriya S P)

Place: Ujire

E-mail: supriyasp@sdmcujire.in

ACKNOWLEDGEMENTS

Firstly, I would like to thank our Principal **Dr.P.N. Udayachandra** for providing the opportunity and the necessary facilities for the completion of this project work in our college.

I would also thank our Dean **Dr. Vishwanatha P** for his support.

It is my privilege to thank our HOD **Prof. Shanthiprakash** for his suggestions and support.

My deepest thanks to my project guide **Ms. Supriya S P**, Assistant Professor, Department of Statistics, SDM College, Ujire, for guiding me throughout my project work.

I would like to express my special thanks of gratitude to **Asst. Prof. Mr. Pradeep k** and **Asst. Prof. Ms. Shwetha Kumari**, Department of Statistics, SDM College, Ujire, for their support during the development of the project.

I would also like to thank my friends and family for their constant encouragement and support throughout the project.

Finally to all who helped me directly or indirectly in many ways, **'Thank You!**'

(Dhanashree. D)

Contents

1 Chapter 1 1

Introduction 1

1.1 Motivation . . . 2
1.2 Literature Review . . . 3
1.3 Objectives . . . 5
1.4 Scope of the Study . . . 5

2 Chapter 2 6

Methodology 6

2.1 Introduction . . . 6
2.2 About the data . . . 6
2.3 Statistical Techniques and Machine learning Model used for Data Analysis . . . 7
2.3.1 Chi-Square Test of Independence . . . 7
2.3.2 Cramer's V test . . . 8
2.3.3 Fisher Exact test . . . 8
2.3.4 Logistic Regression . . . 8
2.3.5 Hopkins statistic . . . 9
2.3.6 K-Means Clustering . . . 10
2.3.7 Elbow method . . . 10
2.3.8 K-Nearest Neighbors Algorithm (KNN) . . . 10
2.3.9 Random Forest . . . 11
2.3.10 Support Vector Machine (SVM) . . . 11
2.3.11 Box plot . . . 11
2.3.12 Confusion Matrix . . . 12

3 Chapter 3 14

Results and Discussion : Univariate Analysis 14

3.1 Characteristics of Sample Respondents . . . 14
3.1.1 Gender . . . 14
3.1.2 Age . . . 15
3.1.3 BMI . . . 16
3.1.4 Transportation . . . 18
3.1.5 Family History with Overweight . . . 19
3.1.6 Caloric Food Consumption . . . 20
3.1.7 Eating Vegetables . . . 21
3.1.8 Consumption of Alcohol . . . 22
3.2 Conclusion . . . 23

4 Chapter 4 **24**

Results and Discussion : Bivariate Analysis **24**

4.1 Relationship between BMI and various factors 24
4.2 Determining the association between physical factors and obesity . . 25
4.2.1 Age 25
4.2.2 Family history of being overweight 26
4.2.3 Physical activity 27
4.3 Determining affect of means of Transportation on obesity 28
4.4 Conclusion 30

5 Chapter 5 **31**

Results and Discussion : Multivariate models **31**

5.1 Analysis of Factors Influencing Obesity 31
5.1.1 Physical Factors 31
5.1.2 Individual habits 33
5.2 Data Prepartion 35
5.3 Unsupervised model-Clustering 35
5.3.1 K-Means Clustering 35
5.4 Supervised model 37
5.4.1 K Nearest Neighbor(K-NN) 37
5.4.2 Random Forest 38
5.4.3 Support Vector Machine(SVM) 40
5.5 Conclusion 43

6 Chapter 6 **44**

Conclusion **44**

7 Bibliography **45**

8 Appendix **46**

List of Tables

1 Frequency distribution of gender 14
2 Age distribution of respondents 15
3 Weight status distribution of respondents 17
4 Transportation distribution of respondents 18
5 Frequency of respondent with family history of overweight 19
6 Frequency of caloric food consumption by respondent 20
7 Frequency of eating vegetable 21
8 Frequency of alcohol consumption 22
9 Table showing obesity status of different age groups. 25
10 Table showing obesity status and family history of overweight 26
11 Table showing obesity status and physical activity frequency 27
12 Table showing obesity status and prefered means of transportation . . 28
13 Table showing summary of logistic regression model 31
14 Table showing summary of logistic regression model 33
15 Profile table of cluster and their characteristics 36
16 Model best parameter with their scores 41
17 Average performance evaluation of the techniques used 42

List of Figures

1 Distribution of Gender 14
2 Distribution of Age 15
3 Boxplot of BMI based on gender 16
4 Distribution of Weight Status 17
5 Distribution of mode of transportation used 18
6 Distribution of respondent with family history of overweight 19
7 Distribution of caloric food consumption by respondent 20
8 Distribution of frequency of eating vegetable 21
9 Distribution of frequency of alcohol consumption 22
10 Boxplot of BMI based various factors 24
11 Boxplot of BMI based Transportation 29
12 Elbow plot of clustering 35
13 2D Visualization of cluster 37
14 Confusion matrix of K-NN model 38
15 Visual representation random forest 39
16 Confusion matrix of random forest model 40
17 Confusion matrix of SVM model 41
18 Variable Importance of machine learning model 42

1 Chapter 1

Obesity is a complex disease involving an excessive amount of body fat. Obesity isn't just a cosmetic concern. Obesity is the condition of being obese, i.e., excess amount of body fat with a BMI of over 30. It's a medical problem that increases the risk of other diseases and health problems, such as heart disease, diabetes, high blood pressure and certain cancers.Obesity is now so common within the world's population that it is beginning to replace undernutrition and infectious diseases as the most significant contributor to ill health. It is a pathological condition in which excess body fat accumulated, leading adverse effects on health and life expectancy. A chronic disorder with complex interaction between genetic and environmental factors. It characterized by high cholesterol, fatty acid levels; imbalance in metabolic energy, insulin desensitization,lethargy, gallstones, high blood pressure, shortness of breath, emotional and social problems, and excessive adipose mass accumulation with hyperplasia and hypertrophy. There are many reasons why some people have difficulty losing weight. Usually, obesity results from inherited, physiological and environmental factors, combined with diet, physical activity and exercise choices.

Many low- and middle-income countries are now facing a 'double burden' of malnutrition.

- While these countries continue to deal with the problems of infectious diseases and undernutrition, they are also experiencing a rapid upsurge in noncommunicable disease risk factors such as obesity and overweight, particularly in urban settings.
- It is not uncommon to find undernutrition and obesity co-existing within the same country, the same community and the same household

Children in low- and middle-income countries are more vulnerable to inadequate pre-natal, infant, and young child nutrition. At the same time, these children are exposed to high-fat, high-sugar, high-salt, energy-dense, and micronutrient-poor foods, which tend to be lower in cost but also lower in nutrient quality. These dietary patterns, in conjunction with lower levels of physical activity, result in sharp increases in childhood obesity while undernutrition issues remain unsolved.

Common specific causes of obesity include:

- Genetics, which can affect how your body processes food into energy and how fat is stored.
- Growing older, which can lead to less muscle mass and a slower metabolic rate, making it easier to gain weight.

- Not sleeping enough, which can lead to hormonal changes that make you feel hungrier and crave certain high-calorie foods.

- Pregnancy, as weight gained during pregnancy may be difficult to lose and might eventually lead to obesity.

Obesity has been linked to a number of health complications, some of which can be life threatening if not treated.

Raised BMI is a major risk factor for noncommunicable diseases such as:

- Cardiovascular diseases (mainly heart disease and stroke), which were the leading cause of death in 2012.

- Diabetes.

- Musculoskeletal disorders (especially osteoarthritis – a highly disabling degenerative disease of the joints).

- Some cancers (including endometrial, breast, ovarian, prostate, liver, gallbladder, kidney, and colon). The risk for these noncommunicable diseases increases, with increases in BMI.

The good news is that even modest weight loss can improve or prevent the health problems associated with obesity. A healthier diet, increased physical activity and behavior changes can helps to lose weight. Prescription medications and weight-loss procedures are additional options for treating obesity.

1.1 Motivation

As per the World Health Organisation (WHO), the worldwide prevalence of obesity nearly tripled between 1975 and 2016. Obesity, once considered a problem in wealthy nations, is shifting in low and middle-income countries. According to the Global Burden of Disease (GBD) study, 5.02 million people died prematurely in 2019 due to obesity, nearly six times the number that died from HIV/AIDS that year.

According to the Centre for Disease Control and Prevention, having obesity increases the risk of severe illness from COVID-19. People who are overweight may also be at increased risk.

Obesity is on the rise in India, with one out of every four people being overweight. The prevalence of obesity among Indians increased in 2019-21 compared to 2015-16, as per the latest National Family Health Survey (NFHS-5) data. Nearly one in every four persons is overweight compared to one in every five earlier.

1.2 Literature Review

In 2021, Alanazi Talal Abdulrahman et al. carried out a data analysis and computational methods for assessing knowledge of obesity risk factors among saudi citizens. The findings showed that there is widespread knowledge of risk factors for obesity among citizens in Saudi Arabia. In addition, a significant association was found between demographics (gender, age, and level of education) and knowledge of risk factors for obesity, in assessing variables for knowledge of the risk factors for obesity in relation to the demographics of gender and level of education.

In 2020, Abduelmula R. AbduelkaremID et al. assessed obesity commonness and its associated risk factors among school-aged children in Sharjah, UAE. The result of the research where as follows: The percentage of overweight and obesity is high in both genders and across all ages of the studied population. The highest percentage of overweight is among children at the age of 11 years and this may be a consequence of their sedentary lifestyle, consumption of unhealthy food and family history.

In 2019,Tran B.X et al. carried out global evolution of obesity research in children and youths:setting priorities for interventions and policies. This study presents the global research trends and developments in the field of childhood obesity. In addition, it highlights the needs for improving international and local research capacities and collaborations to accelerate knowledge production and translation into contextualized and effective childhood obesity prevention.

In 2017, Mark O Goodarzi carried out research on genetics of obesity. Genome-wide association studies (GWAS) for BMI, waist-to-hip ratio, and other adiposity traits have identified more than 300 single-nucleotide polymorphisms (SNPs). Obesity arises from the interactions between an at-risk genetic profile and environmental risk factors, such as physical inactivity, excessive caloric intake, the intrauterine environment, medications, socioeconomic status, and possibly novel factors such as insufficient sleep, endocrine disruptors, and the gastrointestinal microbiome. The heritability of BMI has been estimated to be 40–70%.

In 2015, Sharon M. Fruh carried out a study on obesity: risk factors, complications and strategies for sustainable long-term weight management. The aim of the study was to review the effects of obesity on health and well-being and the evidence indicating they can be ameliorated by weight loss, and consider weight-managment strategies that may help patients achieve and maintain weight loss. It is found that over one third of U.S. adults have obesity and it is associated with range of comorbidities, including diabeted, cardiovascular disease, obstructive sleep apnea, and cancer; however, modest weight loss in the 5%-10% range and above can significantly improve health-related outcomes.

In 2014, Ahmed Salih Sahib et al. conducted study to investigate the eating habits in a sample of overweight and obese Iraqi subjects who attend Al-Kindy obesity therapy and research unit. And the study showed that The prevalence of obesity was more common among females compared to males. Eating habits of the subjects showed that the majority reported taking three meals per day. Also there is statistically significant positive association between specific eating behaviors, namely: number of meals per day, eating fatty food and fast food with obesity and overweight in a sample of Iraqi subjects.

In 2012, Alessandro Bonanno and Stephan J. Goetz conducted research to examine the simultaneous impact of different food outlet's density and expenditures in nutrition education programs on the incidence of healthy eating among the U.S adult population and, indirectly, on adult obesity incidence. The result shows that the presence of Wal-Mart Supercenters, in contrast, across specifications, is associated with lower percentages of individuals consuming fruit and vegetables regularly and, as a consequence, with higher levels of obesity. The effect of SNAP-Ed expenditures is consistently that of more healthy eating and, consequently, reduced obesity.

In 2008, Ayoub Cherd Kafyulilo carried out a study in Kinondoni and Njombe Districts in Tanzania to determine the prevalence and implications of overweight and obesity in children's health and learning behaviour. The results of this study were as follows: An average of 13.5% children, were overweight and obese. Economy status, household occupations, nutrition and inactivity were significant causes of overweight and obesity. Hypertension, excessive sweating, teasing and peer rejection were common to obese children. In addition, overweight and obese children were reported to underperform in academic and physical activities.

In 2004, George A. Bray carried out a study on medical consequences of obesity. The study was to review the effects of obesity resulting from two factors: the increased mass of adipose tissue and the increased secretion of pathogenetic products from enlarged fat cells. And it has showed that increased cytokine release may play a role in other forms of proliferative growth. The combined effect of these pathogenetic consequences of increased fat stores is an increased risk of shortened life expectancy.

In 2004, Srinivas Nammi et al. conducted an overview on obesity its current perspectives and treatment options. This article highlights various preventive aspects and treatment procedures of obesity with special emphasis on the latest research manifolds. Also, the researcher has highlighted that the obesity is not a social condition but is a rampant disease. It cannot be overviewed as just a matter of overeating.

1.3 Objectives

1. To determine the association between physical factors and obesity.
2. To study whether of mode of transportation prefered by individual has any impact on their obesity levels.
3. To determine the influence of various factors on obesity in individuals .
4. Segregate groups based on physical condition and their habits.
5. To determine best classification model to predict the weight status of individuals based on physical condition and their habits.

1.4 Scope of the Study

The results of this research is to help individuals to realise healthy eating habits and having healthy daily life routine are important for their health. Also this result could support public health management, evaluating groups at high risk of obesity and developing or providing effective interventions in the early stages.

2 Chapter 2

Methodology

2.1 Introduction

This chapter includes all information about the dataset considered for the study. Details of the various statistical techniques used for the exploratory data analysis are also provided.

2.2 About the data

A secondary data has been collected from the website of ‘UCI Machine Learning Repository’. The data presents information from different locations such as Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition. The data contains 15 attributes and 2111 records. The description about the variables considered in the analysis are given as follows:

- **Age** : Age of respondent in years
- **Gender** : Gender of respondent
- **Height** : Height of respondent in meters
- **Weight** : Weight of respondent in kilograms
- **Family history**: Has a family member suffered or suffers from overweight
- **FAVC** : Eat High Caloric Food Frequently (Yes, No)
- **FCVC** : Frequency of eating vegetables (Never, Sometimes, Always)
- **NCP** : Number of main meals (1,2,3)
- **CAEC** : Consumption of food between meals (No, Sometimes, Frequently, Always)
- **Smoke** : Smoking(Yes, No)
- **CH20** : Consumption of water daily in liter
- **SCC** : Caloric consumption monitoring (Yes,No)
- **FAF** : Physical activity frequency (0-4 days)
- **TUE** : Time using technology devices (0-2 , 3-5,more than 5 (in hours))
- **CALC** : Consumption of alcohol(No, Sometimes,Frequently,Always)

- **MTRANS**: Transportation used(Automobile, Bike, Public Transportation, Walking)
- **Weight status**: Weight status based on BMI (Underweight, Normal weight, Overweight I,Overweight II, Obesity I, Obesity II, Obesity III)

2.3 Statistical Techniques and Machine learning Model used for Data Analysis

'Python' and 'R' language has been used to carry out the analysis and the interpretation of the data. The statistical methods and models used in order to carry out the analysis are given as follows :

2.3.1 Chi-Square Test of Independence

The chi-square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. The frequency of each category for one nominal variable is compared across the categories of the second nominal variable. The data can be displayed in a contingency table. The hypothesis under consideration are given as follow :

H_0 : The two categorical variables are independent.
H_1 : The two categorical variable are dependent.

The test statistic for the chi-square test of independence is denoted χ^2, and is computed as:

$$\chi^2 = \frac{\sum (O-E)^2}{E}$$

where O represents the observed frequency, E is the expected frequency under the null hypothesis and is computed as follows :

$$E = \frac{row\ total \times column\ total}{sample\ size}$$

The expected frequency count for each cell of the table must be atleast 5.
The test procedure is to reject the null hypothesis H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$, where $\chi^2_{\alpha,(r-1)(c-1)}$ is the upper α^{th} percentile value of the central chi-square distribution with $(r-1)(c-1)$ degrees of freedom, r is the number of rows and c is the number of columns.
Also using the $p-value$ we can draw conclusion about the test. If the $p-value$ is less than 0.05, then reject the null hypothesis H_0.

2.3.2 Cramer's V test

Cramer's V statistic is a commonly used measure of association between two categorical variables. It takes on values between 0 and 1 (inclusive), with 0 corresponding to no association between the variables and 1 corresponding to one variable being completely determined by the other.
The assumptions for Cramer's V include:

- Categorical variables

If we have two categorical variables: X taking values from $\{a_1, \ldots, a_I\}$ and Y taking values from $\{b_1, \ldots, b_J\}$ with n obervations then Cramér's V statistic is given by:

$$V = \sqrt{\frac{\chi^2/n}{\min(I, J) - 1}}$$

2.3.3 Fisher Exact test

The Fisher Exact test is a test of significance that is used in the place of chi square test in 2 by 2 tables, especially in cases of small samples. It is used to determine if there are nonrandom associations between two categorical variables. The Fisher Exact test is computed in addition to the chi square test for a 2X2 table when the table consists of a cell where the expected number of frequencies is fewer than 5. The test uses the following formula:

$$p = \frac{((a + b)!(c + d)!(a + c)!(b + d)!)}{a!\ b!\ c!\ d!\ N!}$$

Where the 'a,' 'b,' 'c' and 'd' are the individual frequencies of the 2X2 contingency table, and 'N' is the total frequency. The Fisher Exact test uses this formula to obtain the probability of the combination of the frequencies that are actually obtained.

2.3.4 Logistic Regression

Logistic regression estimates the probability of an event occurring, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. Logistic regression analysis is used to examine the association of (categorical or continuous) independent variable(s) with one dichotomous dependent variable. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$Logit(pi) = \frac{1}{(1 + e^{(-pi)})}$$

$$ln(\frac{pi}{1-pi}) = \beta_0 + \beta_1 * X_1 + \ldots + \beta_k * X_k$$

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). The exponentiated regression coefficient represents the strength of the association of the independent variable with the outcome. More specifically, it represents the increase (or decrease) in risk of the outcome that is associated with the independent variable.
Statistical inference of β can be constructed by performing an approximate Wald test, the so-called Z-test. Given the standard error estimate, the null and the alternative hypotheses, written as $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$, can be statistically tested by using the following Wald statistic:

$$Z = \frac{\hat{\beta}_i}{\sqrt{\hat{V}(\hat{\beta}_i)}}$$

where the Z asymptotically follows a standard normal distribution.

2.3.5 Hopkins statistic

The Hopkins statistic is used to assess the clustering tendency of a data set by measuring the probability that a given data set is generated by uniform data distribution. In other words, it tests the spatial randomness of the data.

Under the method we define hypothesis as:

- Null hypothesis: The data set D is uniformly distributed (i.e., no meaningful clusters)

- Alternative hypothesis: The data set D is not uniformly distributed (i.e., contains meaningful clusters)

Hopkins statistic,

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d}$$

If the value is around 0.5 or lesser, the data is uniformly distributed and hence it is unlikely to have statistically significant clusters. If the value is between 0.7, ..., 0.99, it has a high tendency to cluster and therefore likely to have statistically significant clusters.

2.3.6 K-Means Clustering

Clustering is an unsupervised machine learning technique, which partitions the entire dataset into a fixed number of clusters or groups in such a way that observations in each group exhibit similar behavior or characteristics and are homogeneous. The observations in the same group are more similar than the observations in other groups. In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem.The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. The number of clusters found from data by the method is denoted by the letter 'K' in K-means.

2.3.7 Elbow method

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k. In this method the cost function (variance) is plotted against different values of k. As the value of k increases, the number of clusters decreases and the average distortion also decreases. The optimal value of k is where we see the sharp change in the rate of change of error.

2.3.8 K-Nearest Neighbors Algorithm (KNN)

The k-nearest neighbors is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another. For classification problems, a class label is assigned on the basis of a majority vote—i.e. the label that is most frequently represented around a given data point is used.This algorithm uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. The k-nearest neighbor classifier fundamentally relies on a distance metric. A common distance metric to identify the k nearest neighbors is the Euclidean distance measure,

$$d(X, Y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

which is a pairwise distance metric that computes the distance between two data points x and y over the m input features.

2.3.9 Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches. This sequence continues until a leaf node is attained. The leaf node cannot be segregated further. The nodes in the decision tree represent attributes that are used for predicting the outcome. Decision nodes provide a link to the leaves. Entropy is a metric for calculating uncertainty. Information gain is a measure of how uncertainty in the target variable is reduced, given a set of independent variables. Min samples split represents the minimum number of samples required to split an internal node and n estimators is the number of trees to be build before taking the maximum voting or averages of predictions.

2.3.10 Support Vector Machine (SVM)

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. 'C' parameter in SVM is penalty parameter of the error term. It controls the trade off between smooth decision boundary and classifying the training points correctly. The SVM kernel is a function that takes low dimensional input space and transforms it into higher-dimensional space, i.e., it converts not separable problem to separable problem.

There are mainly four different types of kernels (Linear, Polynomial, RBF, and Sigmoid) that are popular in SVM classifier. Linear Kernel is used when the data is Linearly separable, that is, using a single Line. Some methods that can make support vector machines capable of dealing with more than two classes. We call these methods heuristic methods. One of these method is One-vs-Rest (OvR). This approach mainly splits the multiclass data as binary classification data so that the binary classification algorithms can be applied to convert binary classification data.

2.3.11 Box plot

A box plot is a chart that shows data from a five-number summary including one of the measures of central tendency. It does not show the distribution in particular as much as a stem and leaf plot or histogram does. But it is primarily used to indicate a distribution is skewed or not and if there are potential unusual observations (also called outliers) present in the data set. Boxplots are also very beneficial when large numbers of data sets are involved or compared.

2.3.12 Confusion Matrix

Confusion Matrix is the visual representation of the Actual VS Predicted values. It is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. It represents the different combinations of Actual VS Predicted values.

TP: True Positive: The values which were actually positive and were predicted positive.

FP: False Positive: The values which were actually negative but falsely predicted as positive. Also known as Type I Error.

FN: False Negative: The values which were actually positive but falsely predicted as negative. Also known as Type II Error.

TN: True Negative: The values which were actually negative and were predicted negative.

Evaluation Metrics

1. **Accuracy:** Accuracy is defined as the percentage of the total sample that was correctly classied. It can be measured by the following equation:

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

2. **Recall / Sensitivity:** The recall is the measure to check correctly positive predicted outcomes out of the total number of positive outcomes.

$$Recall = \frac{TP}{TP + FN}$$

3. **Precision:** Precision checks how many outcomes are actually positive outcomes out of the total positively predicted outcomes.

$$Precision = \frac{TP}{TP + FP}$$

4. **F-1 score:** F-1 score is the harmonic mean of Precision and Recall and it captures the contribution of both of them.

$$F - 1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

5. **Matthews correlation coefficient (MCC) :** The Matthews correlation coefficient is used in machine learning as a measure of the quality of binary and multiclass classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction.It is calculated as:

$$MCC = \frac{(TP * TN – FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3 Chapter 3

Results and Discussion : Univariate Analysis

3.1 Characteristics of Sample Respondents

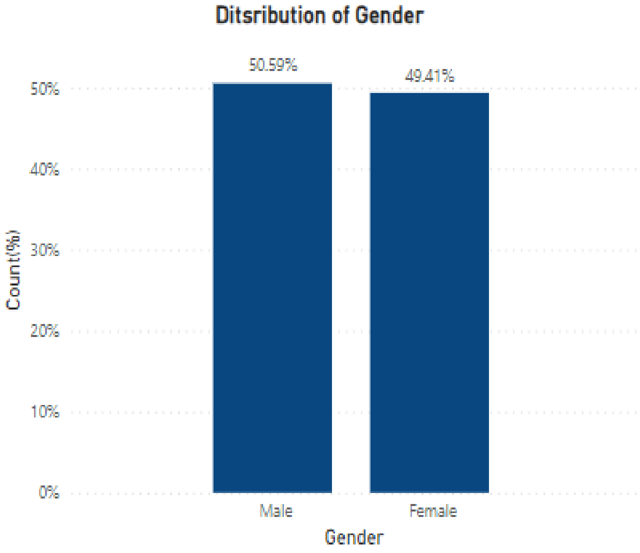
3.1.1 Gender

The below table shows the frequency of the gender of the respondents considered in this study.

Table 1: Frequency distribution of gender

Gender	Frequency	Percentage
Male	1068	50.5
Female	1043	49.4

Figure 1: Distribution of Gender



This dataset contains 2111 records out which 50.5% are male respondents and 49.4% are female respondents.

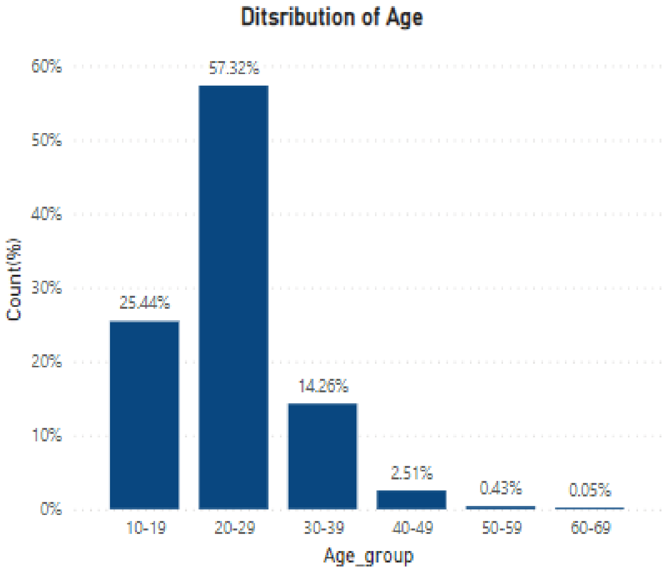
3.1.2 Age

The following table shows the frequency of the age(in years) of the respondents considered in this study.

Table 2: Age distribution of respondents

Age group	Frequency	Percentage
10-19	537	25.44
20-29	1210	57.32
30-39	301	14.26
40-49	53	2.51
50-59	9	0.43
60-69	1	0.05

Figure 2: Distribution of Age



Here, we observe that 57.32% of respondent considered in our study belongs to age category of 20-29.

3.1.3 BMI

Each weight status was labelled by body mass index calculated for each individual using the equation shown below:

$$BMI = \frac{Weight}{Height^2}$$

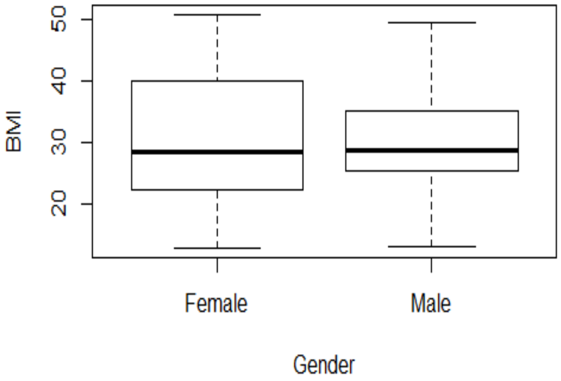
where weight is in kilogram and height is in metre.

The weight status was categoriesed as follows:

- Underweight : BMI $<$ 18.5
- Normal weight : 18.5 $\leq$ BMI $<$ 25
- Over weight : 25 $\leq$ BMI $<$ 30
- Obesity I : 30 $\leq$ BMI $<$ 35
- Obesity II : 35 $\leq$ BMI $<$ 40
- Obesity III : 40 $\leq$ BMI

The following figure shows the BMI of male and female.

Figure 3: Boxplot of BMI based on gender



Here, we can observe that there are no outliers.
The average BMI of the male is 29.28 and of female is 30.13, which comes under the category 'overweight'. Also there is no much variability.

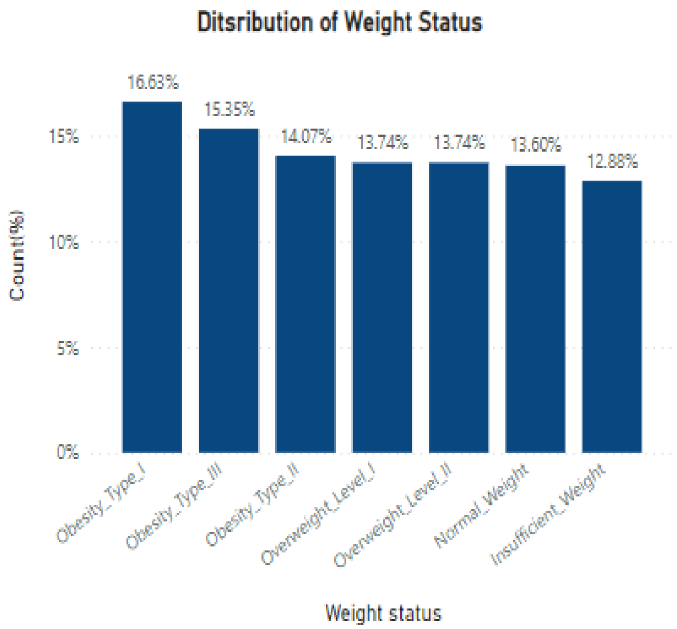
- **Weight Status**

The following table shows the frequency of the weight status of the respondents considered in this study.

Table 3: Weight status distribution of respondents

Weight Status	Frequency	Percentage
Insufficient weight	272	12.88
Normal weight	287	13.60
Overweight I	290	13.74
Overweight II	290	13.74
Obesity I	297	16.68
Obesity II	324	14.07
Obesity III	351	15.35

Figure 4: Distribution of Weight Status



Here, we observe that most of the respondents considered in this study belongs to 'Obesity I' category .i.e, 16.62% of our respondents are of type 'obesed I'.

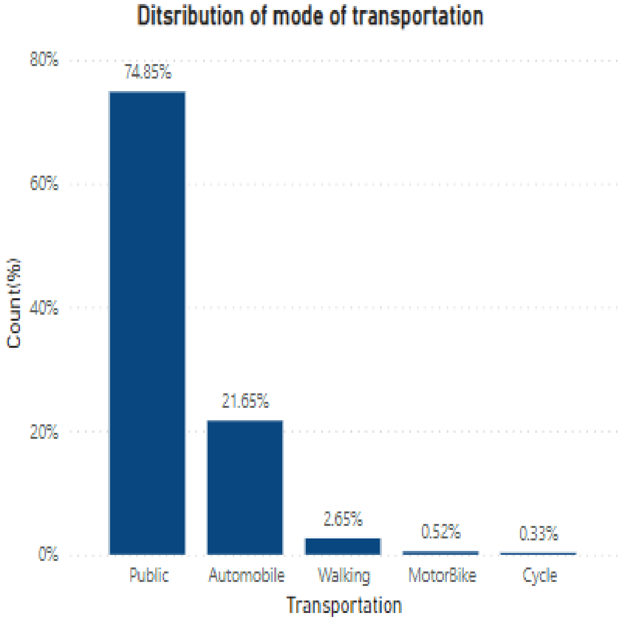
3.1.4 Transportation

The following table shows the frequency of transportation used by respondent.

Table 4: Transportation distribution of respondents

Transportation	Frequency	Percentage
Cycle	7	0.33
Motorbike	11	0.52
Walking	56	2.65
Automobile	457	21.64
Public Transportation	1580	74.84

Figure 5: Distribution of mode of transportation used



We observe that most of our respondent in our study use public transportation and less used means of transportation is cycle.

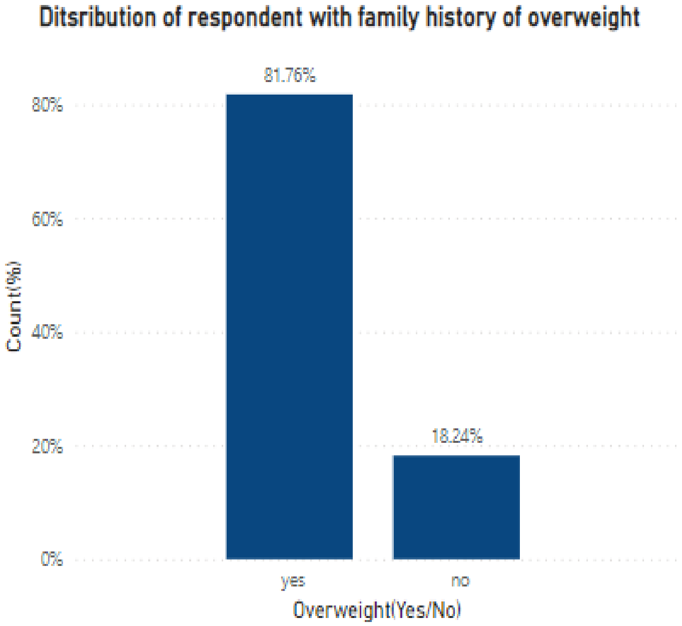
3.1.5 Family History with Overweight

The following table shows the frequency of respondent with family history of overweight.

Table 5: Frequency of respondent with family history of overweight

Family history with overweight	Frequency	Percentage
No	385	18.23
Yes	1726	81.76

Figure 6: Distribution of respondent with family history of overweight



Here, we observe that most of respondent in our study has family history of being overweight.

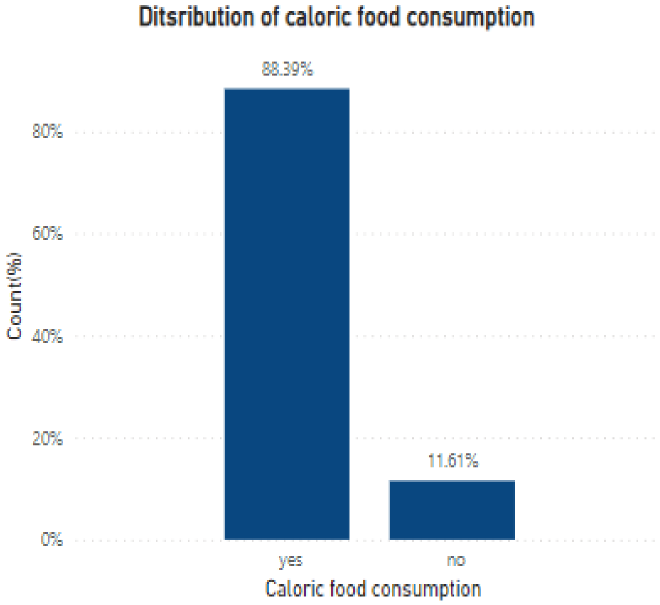
3.1.6 Caloric Food Consumption

The following table shows the frequency of caloric food consumption.

Table 6: Frequency of caloric food consumption by respondent

Caloric food consumption	Frequency	Percentage
No	245	11.61
Yes	1866	88.39

Figure 7: Distribution of caloric food consumption by respondent



Most of respondent in our study has high caloric food consumption habit.

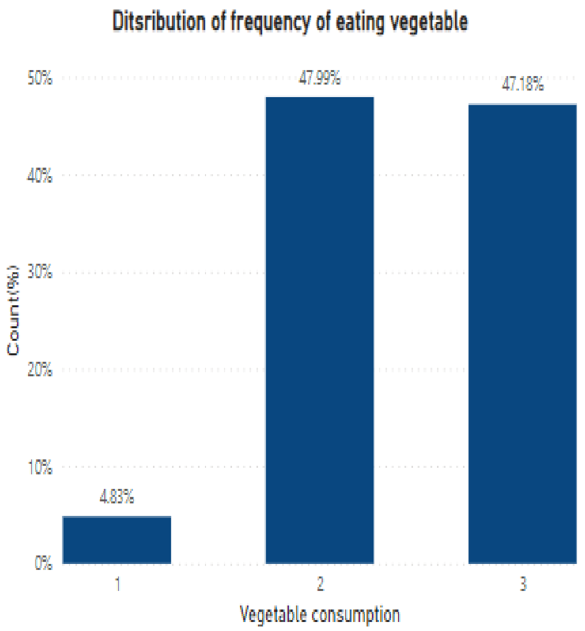
3.1.7 Eating Vegetables

The following table shows the frequency of eating vegetable.

Table 7: Frequency of eating vegetable

Vegetable food consumption	Frequency	Percentage
Never(1)	102	4.83
Sometimes(2)	1013	47.99
Always(3)	996	47.18

Figure 8: Distribution of frequency of eating vegetable



Here, we observe that most of respondent in our study has habit of eating vegetables.

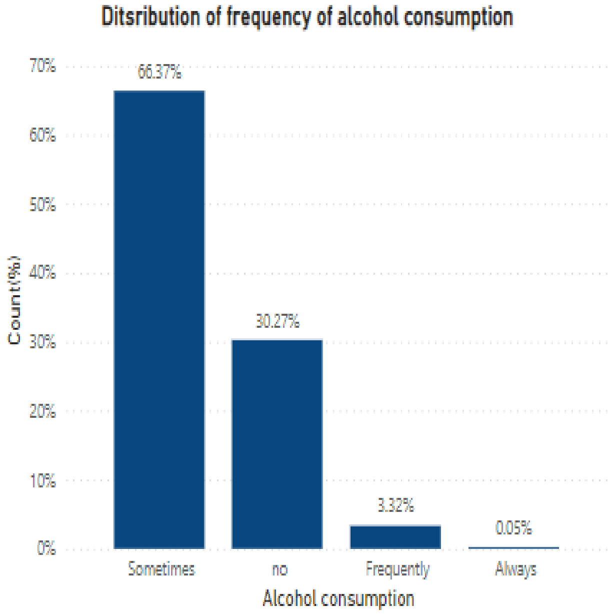
3.1.8 Consumption of Alcohol

The following table shows the frequency of alcohol consumption.

Table 8: Frequency of alcohol consumption

Alcohol consumption	Frequency	Percentage
No	639	30.27
Sometimes	1401	66.37
Frequently	70	3.32
Always	1	0.05

Figure 9: Distribution of frequency of alcohol consumption



Here, 66.3% of our respondent in our study has habit of alcohol consumption and 30.2% do not consume alcohol.

3.2 Conclusion

1. The number of respondent considered in this data are 2111 out of which 50.5% are male and 49.4% are female.

2. Most of respondent belongs to age group of 20-29.

3. The average BMI of the male is 29.28 and of female is 30.13, which comes under the category ‘overweight’ and ‘Obesity I’ respectively.

4. About 16.62% of the respondents considered in this study belongs to ’Obesity I’ category.

5. Considering the transportation preference, cycle is less prefered comparing with other mode of transportation.

6. About 81.76% respondent has family history of being overweight.

7. Respondents has high caloric food consumption habit i.e., 88.3% of respondents in our study consume high caloric food.

8. About 4.8% of respondents do not have the habit of eating eating vegetables.

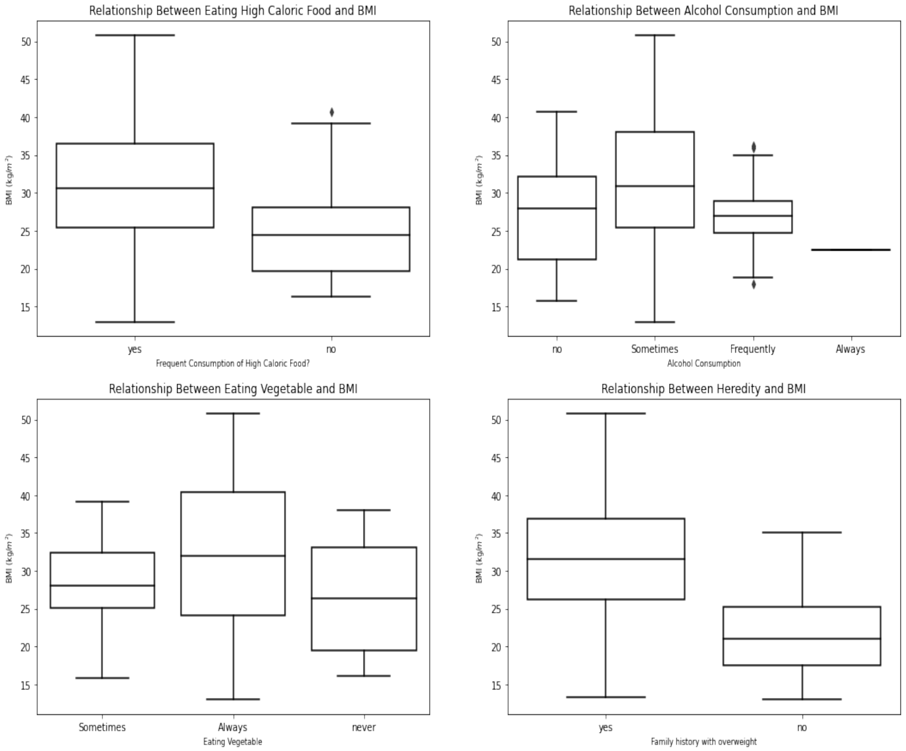
9. About 66.37% of respondent has habit of alcohol consumption sometimes and 30.2% do not consume alcohol.

4 Chapter 4

Results and Discussion : Bivariate Analysis

4.1 Relationship between BMI and various factors

Figure 10: Boxplot of BMI based various factors



(a) The first subplot shows that that frequently consumption of high-calorie food have a median BMI higher than the median BMI of those who don't. This indicates that calorie count is likely a contributing factor to increased body fat.

(b) The second subplot shows that there is little relationship between alcohol consumption and BMI, as those who frequently drink alcohol had the same median BMI as those who do not drink alcohol at all.

(c) The third subplot shows that there is little relationship between vegetable consumption and BMI, as those who always eat vegetables has the more median BMI as those who do consume less.

(d) The fourth subplot suggests that a family history of obesity is also a contributing factor, as the median BMI of those with a family history of obesity is around 31 kg/m^2 higher.

4.2 Determining the association between physical factors and obesity

4.2.1 Age

Affect of respondent age on obesity using Chi-square test of Independence

The following table shows the classification of age(in yrs) based on their obesity status.

Table 9: Table showing obesity status of different age groups.

Obesed	Age group			
	10-20	20-30	30-40	40 & above
No	417	556	139	25
Yes	120	654	162	38

The hypothesis are as follows:

H_0 : There is no significant association between the variable ‘age’ and ‘obesity’.

H_1 : There is significant association between the variable ‘age’ and ‘obesity’.
The values obtained are as follows:

- Chi-square test statistic: 177.192
- p-value : 2.2e-16
- Degrees of freedom : 3

Here we can observe that the obtained p-value is less than 0.05. Hence we reject H_0. Thus, the variable ‘Age’ and ‘obesity’ are dependent. Also, Cramer’s V value turns out to be 0.2796, which indicates a fairly weak association between the two variables.

Ageing is associated with an increase in obesity. Adults who have a normal BMI often start to gain weight in young adulthood and continue to gain weight until they are ages 60 to 65. In addition, children who have obesity are more likely to have obesity as adults.

4.2.2 Family history of being overweight

Affect of heredity on the obesity using Chi-square test of Independence

The following table shows the classification of respondent obesity status based on their family history of being overweight.

Table 10: Table showing obesity status and family history of overweight

Obesed	Family history	
	No	Yes
No	377	760
Yes	8	966

The hypothesis are as follows:

H_0 : There is no significant association between respondent being obesed and their family history of being overweight.

H_1 : There is significant association between respondent being obesed and their family history of being overweight.

The values obtained are as follows:

- Chi-square test statistic: 365.6931
- p-value : 1.621e-81
- Degrees of freedom : 1

Here we can observe that the obtained p-value is less than 0.05. Hence we reject H_0. Thus, the variable 'Family history' and 'Obesity' are dependent. Also, Cramer's V value turns out to be 0.4162, which indicates there is a fairly weak association between the variables 'Family history' and 'Obesity'.

Overweight and obesity tend to run in families, suggesting that genes may play a role. The chances of being obesed are greater if one or both of our parents are overweight or have obesity.

4.2.3 Physical activity

Affect of physical activity on the obesity using Chi-square test of Independence

The following table shows the classification of respondent obesity status and their frequency of physical activity(in days).

Table 11: Table showing obesity status and physical activity frequency

Obesed	Physical activity(in days)			
	0-1	1-2	2-3	3-4
No	335	418	290	94
Yes	385	358	206	25

The hypothesis are as follows:

H_0 : There is no significant association between respondent being obesed and physical activity frequency.

H_1 : There is significant association between respondent being obesed and physical activity frequency.

The values obtained are as follows:

- Chi-square test statistic: 50.058
- p-value : 7.764e-11
- Degrees of freedom : 3

Here we can observe that the obtained p-value is less than 0.05. Hence we reject H_0. Therefore, there is significant association between respondent being obesed and their physical activity frequency. Also, Cramer's V value turns out to be 0.1539, which indicates a fairly weak association between the two variables.

Obesity results from energy imbalance. The World Health Organization, the U.S. Dept. of Health and Human Services, and other authorities recommend that for good health, one should get the equivalent of two and a half hours of moderate-to-vigorous physical activity each week.

4.3 Determining affect of means of Transportation on obesity

Affect of means of transportation on obesity using Fisher Exact Test

The following table shows the classifiaction of respondents based on their obesity status and means of transportation .

Table 12: Table showing obesity status and prefered means of transportation

Obesed	Transportation				
	Public Transportation	Walking	Automobile	Motorbike	Cycle
No	820	53	250	8	6
Yes	760	3	207	3	1

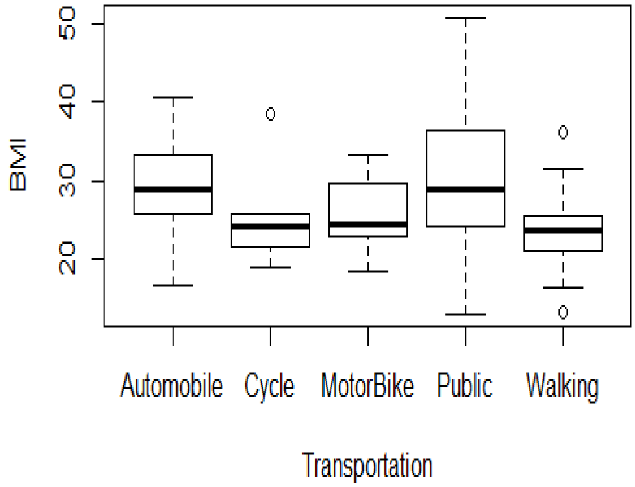
The hypothesis are as follows:

H_0 : There is no significant association between the means of transportation and obesity .

H_1 : There is significant association between the means of transportation and obesity.

From Fisher's exact test it is found that p-value is 3.126e-11 which is less than 0.05. Hence we reject H_0. Thus, means of transportation are significantly associated with obesity. Also, Cramer's V value turns out to be 0.1451, which indicates a fairly weak association between the two variables.

Figure 11: Boxplot of BMI based Transportation



The figure shows us that there is a trend between more strenuous activity, such as walking or cycling, with a lower BMI.

4.4 Conclusion

(a) Frequently consumption of high-calorie food have a median BMI higher than the median BMI of those who don't.

(b) There is little relationship between alcohol consumption and BMI, as those who frequently drink alcohol had the same median BMI as those who do not drink alcohol at all.

(c) There is relationship between vegetable consumption and BMI, as those who always eat vegetables has the more median BMI as those who do consume less.

(d) Family history of obesity is also a contributing factor, as the median BMI of those with a family history of obesity is around 11 kg/m^2 higher.

(e) Ageing is associated with an increase in obesity. Adults who have a normal BMI often start to gain weight in young adulthood and continue to gain weight until they are ages 60 to 65, which is true in our study also.

(f) There is significant association between respondent being obesed and their family history of being overweight. Overweight and obesity tend to run in families, suggesting that genes may play a role.

(g) Obesity results from energy imbalance. Thus, physical activity frequency plays an important role.

(h) Also, means of transportation are significantly associated with obesity. Considering the transportation preference, cycle is less prefered comparing with other means of transportation.

5 Chapter 5

Results and Discussion : Multivariate models

Logistic Regression

5.1 Analysis of Factors Influencing Obesity

In this section, logistic regression model is developed to find out the factors that influence to obesity in individual. Here, factors like physical factors and individual habits are considered.

5.1.1 Physical Factors

Under physical factors 'Gender, Age, Family history with overweight(Heredity) and physical activity frequency' are factors to be included in the logistic model. For this purpose, chi-square test is carried so that only the predictors which are significant are included.

Thus, these factors are considered as predictor(X_i) and 'obesity'(Y) as study variable. The results where as follows:

Table 13: Table showing summary of logistic regression model

Predictor	Estimate	Std Error	z value	p-value
Gender	-0.185	0.121	-1.527	0.127
Age	0.052	0.010	5.171	2.3 x 10^{-7}
Heredity	3.963	0.418	9.461	2 x 10^{-15}
Physical activity	-0.306	0.070	-4.363	1.2 x 10^{-5}

Also,

- Null deviance: 2040.4 on 1477 degrees of freedom
- Residual deviance: 1643.9 on 1473 degrees of freedom
- AIC: 1653.9

Here we can observe the p-value of 'Age, Heredity and Physical activity' is less than 0.05. Hence we conclude that in this study, the predictor Age, Heredity(Family history with overweight) and Physical activity frequency of the individual does play an important role in explaining obesity.

Testing for model significance

The hypothesis are as follows:

H_0 : The model is not significant in explaining the relationship between the predictor and response.

H_1 : The model is significant in explaining the relationship between the predictor and response.

Testing for model significance we found p-value to be 0.0011. Thus, model is significant in explaining the relationship between the predictor and response.

Wald test is used to check the significance of individual predictor.
The hypothesis are as follows:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0.$

The results where as follows:

Predictor	F-value	p-value
Age	18.70	0.16 x 10^{-4}
Heredity	89.70	0.02 x 10^{-13}
Physical activity	27.62	0.16 x 10^{-6}

It is found that ‘Age, Heredity and Physical activity’ are significant predictors with p value less than 0.05. The coefficient estimate of the variable ‘Age’ is 0.052, which is positive. Thus increase in ‘Age’ is associated with increase in the probability of being obesed. However the coefficient for the variable ‘Physical activity’ is -0.306, which is negative. Thus, increase in physical activity will be associated with a decreased probability of being obesed.

5.1.2 Individual habits

Under individual habits, eating habits like consumption of 'High caloric food, Vegetable, Water, number of main meals and habits including Smoking, Alcohol consumption frequency, time spent on technology devices' are the factors to be included in the logistic model.

For this purpose, chi-square test is carried so that only the predictors which are significant are included.

Thus, these factors are considered as predictor(X_i) and 'obesity'(Y) as study variable.

The results where as follows:

Table 14: Table showing summary of logistic regression model

Predictor	Estimate	Std Error	z value	p value
High caloric food	2.394	0.26662	8.982	2 x 10^{-15}
Eating Vegetable	-0.62147	0.09752	6.373	1.85 x 10^{-15}
Main meals	0.07775	0.06907	1.126	0.260
Smoking	0.15030	0.40865	0.368	0.713
Water	0.35541	0.09174	3.874	0.00010
Time on Technology devices	-0.25841	0.08326	-3.103	0.0019
Alcohol	0.13215	0.11203	1.180	0.2381

Also,

- Null deviance: 2040.4 on 1477 degrees of freedom
- Residual deviance: 1842.3 on 1470 degrees of freedom
- AIC: 1858.3

Here we can observe the p-value of 'High caloric food, Eating Vegetable, Water and Time spent on Technology devices' is less than 0.05. Hence we conclude that in this study, the predictor i.e., consumption of high caloric food, Vegetable, Water and Time spent on Technology devices by the individual does play an important role in explaining obesity.

Testing for model significance

The hypothesis are as follows:

H_0 : The model is not significant in explaining the relationship between the predictor and response.

H_1 : The model is significant in explaining the relationship between the predictor and response.

Testing for model significance we found p-value to be 0.09x10^{-8}. Thus, model is significant in explaining the relationship between the predictor and response.

Wald test is used to check the significance of individual predictor.
The hypothesis are as follows:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0.$

The results where as follows:

Predictor	F-value	p-value
High caloric food	39.30	0.04 x 10^{-4}
Eating Vegetable	78.8	0.02 x 10^{-13}
Water	7.43	0.006
Time on Technology Devices	6.609	0.010

It is found that 'High caloric food, water, Time on Technology devices and Eating Vegetable' are significant predictors. The coefficient estimate of the variable 'High caloric food and water' is positive. Thus increase in 'High caloric food and water' is associated with increase in the probability of being obesed. However the coefficient for the variable 'Time on Technology devices and Eating Vegetable' is negative.
Thus, more time spent on technology devices and eating Vegetable will be associated with a decreased probability of being obesed.

5.2 Data Prepartion

Each obesity level was labelled by mass body index calculated for each individual. The dataset contains categorical and ordinal attributes which are not ideal for applying model. Thus, all the categorical attributes where converted into numerical.

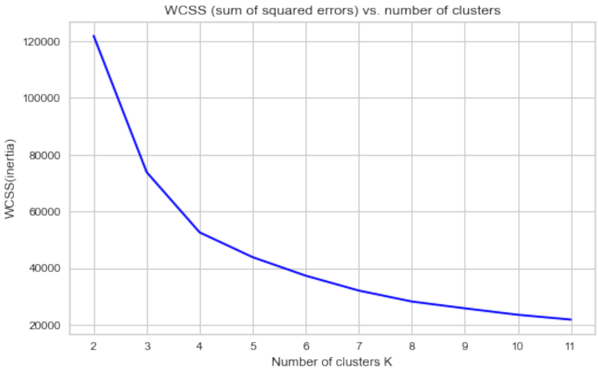
5.3 Unsupervised model-Clustering

5.3.1 K-Means Clustering

In this section, a K-means clustering model is developed. At first, a ’hopkins’ test was conducted to assess the clustering tendency of a data. From the result it is observed the data has clustering tendency. And then performed segmentation based on principal components scores.

Used elbow plot to determine the optimal k value

Figure 12: Elbow plot of clustering



The scree plot of the distortion metric reports an elbow at k=4 clusters with silhouette score 0.423.

The profile table lists the four clusters in its columns and displays their characteristics in its rows. For categorical variables: the mode; for numerical values the mean.

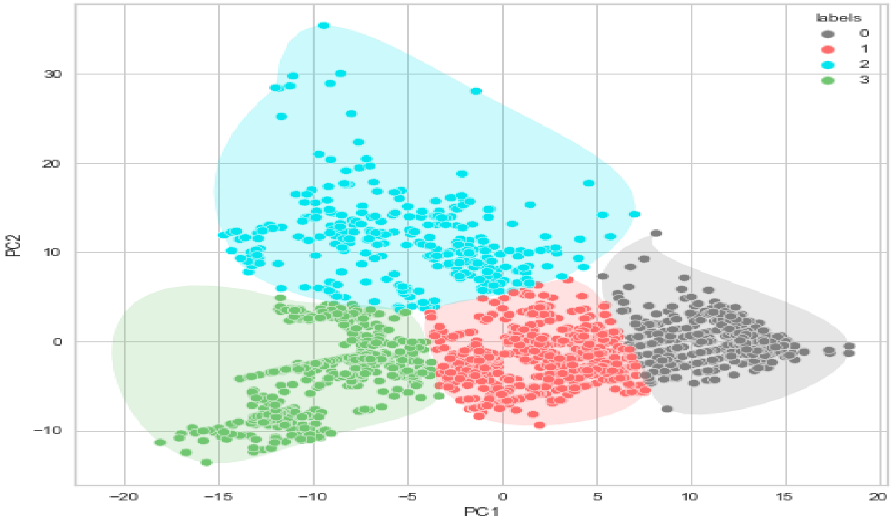
Table 15: Profile table of cluster and their characteristics

Variables	**Cluster 1(0)**	**Cluster 2(1)**	**Cluster 3(2)**	**Cluster 4(3)**
Age	19.876	21.497	36.543	24.970
BMI	19.089	28.449	29.865	39.653
Number of main meals	3	2	2	3
Water (ltr)	1.846	2.116	1.865	2.080
Gender	Female	Male	Male	Female
Family history with overweight	No	Yes	Yes	Yes
Physical activity (days)	2	1	0	0
Caloric food	Yes	Yes	Yes	Yes
Eating Vegetable	Always	Sometimes	Sometimes	Always
Smoking	No	No	No	No
Alcohol	Sometimes	Sometimes	Sometimes	Frequently
Transportation	Public_Transportation	Bike	Automobile	Public Transportation

By looking at the results table, we can derive information about the physical factor, preferences and habits of each of the clusters.

- Cluster 1 subjects has average BMI of 19.08 which is least and it is evident since physical activity frequency and consumption of vegetables is more compared to other segments even though respondents has habit of consumption of caloric food.
- Cluster 2 subjects has average BMI of 28.4 which is approximately similar to cluster 3. Also, there is quite similarities in subjects habits.
- Cluster 4 subjects has average BMI of 39.6 which is quite more and it is evident since subjects are not involved in physical activity but has habit of eating high caloric food and frequent consumption of alcohol.

Figure 13: 2D Visualization of cluster



5.4 Supervised model

In the section classification model is developed. Also data preprocessing, data transformation, model building, evaluation and interpretation of the results were performed.

5.4.1 K Nearest Neighbor(K-NN)

To build knn model, weight status is considered as predictor variable and all other variables are considered as features. In weight status there are 7 classes(Underweight, normalweight, overweight(I and II),Obesity(I,II,III). The dataset is partitioned into training (80%) and testing (20%) sets.

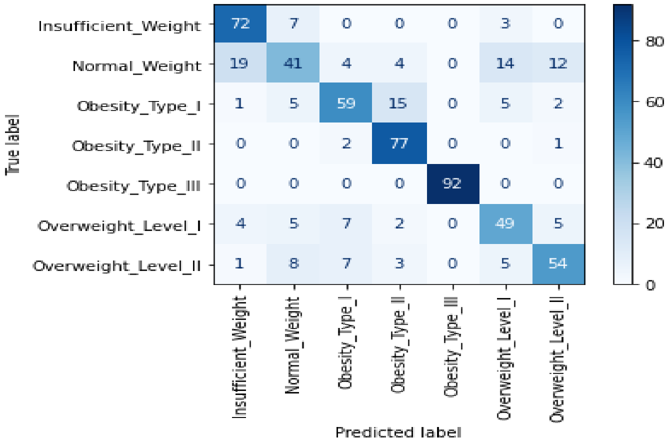
In order to determine best k values model was optimized via gridsearchcv and best k value was obtained. The model results are as follows:

- K : 5
- Train score : 84.4%
- Test score : 75.3%
- F1-Score : 81.1%
- Precision : 81.5%
- Recall : 82.14%

From above we can observe that the model is overfitted since accuracy of test and train data is not near.

Confusion Matrix

Figure 14: Confusion matrix of K-NN model



5.4.2 Random Forest

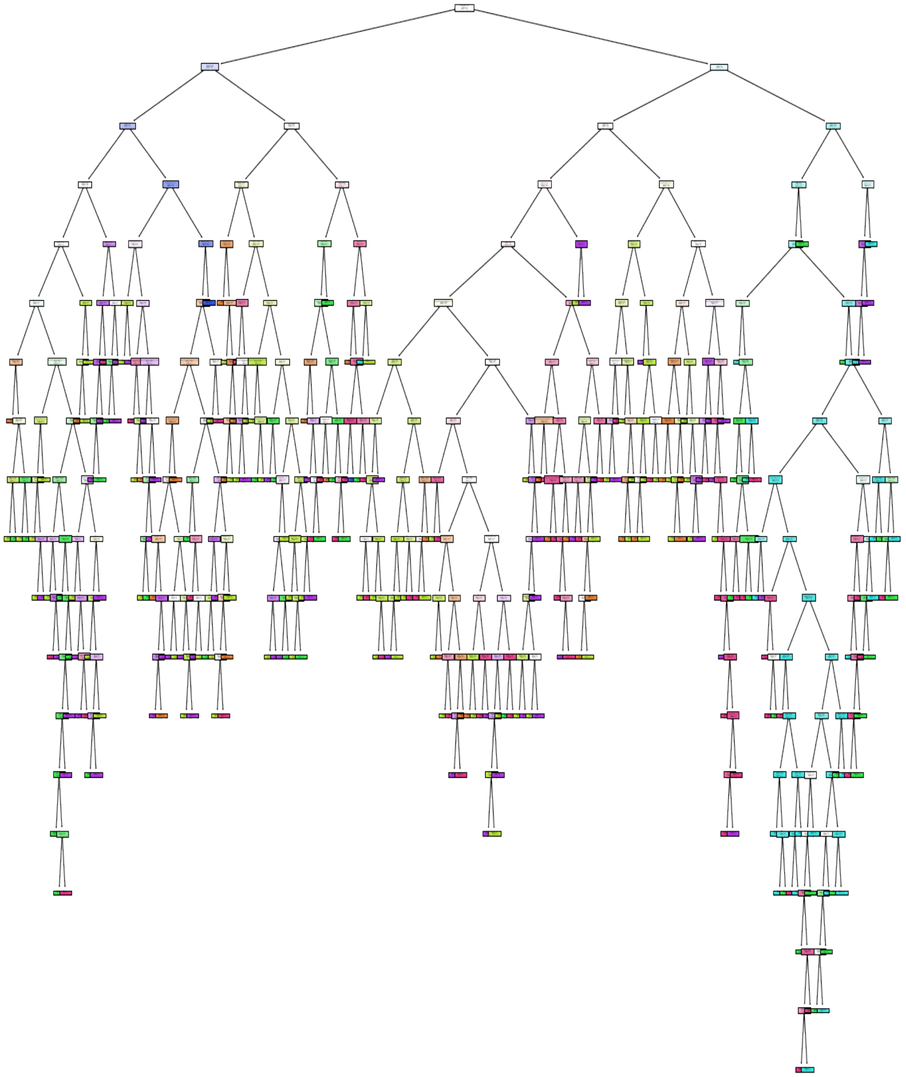
To build random forest model, weight status is considered as predictor variable and all other variables are considered as features. In weight status there are 7 classes(Underweight, normalweight, overweight(I and II),Obesity(I,II,III). The dataset is partitioned into training (80%) and testing (20%) sets.

In order to determine best parameter model was optimized via gridsearchcv and best parameters where obtained. The model results are as follows:

- Best parameters : ‘criterion’: ‘gini’, ‘min samples split’: 2, ‘n estimator:50
- Train score : 89.1%
- Test score : 81.2%
- F1-Score : 81.1%
- Precision : 81.5%
- Recall : 82.14%

From above we can observe that the model is pretty good since all the scores are good and model is not overfitted.

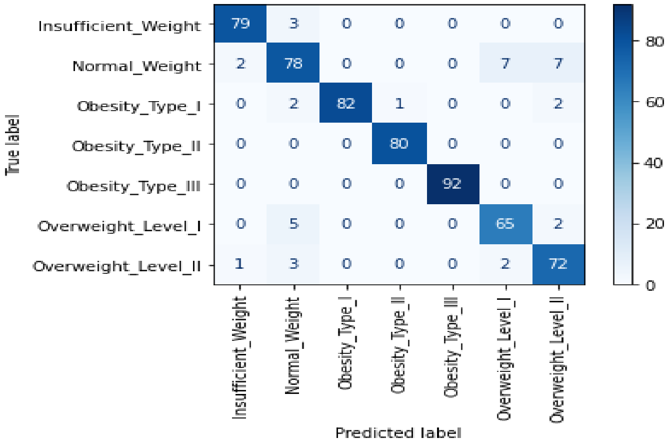
Figure 15: Visual representation random forest



The above plot is the visualization of randomforest consist of 50 decision tree.

Confusion Matrix

Figure 16: Confusion matrix of random forest model



5.4.3 Support Vector Machine(SVM)

To build support vector machine model, weight status is considered as predictor variable and all other variables are considered as features. In weight status there are 7 classes(Underweight, normalweight, overweight(I and II),Obesity(I,II,III). The dataset is partitioned into training (80%) and testing (20%) sets.

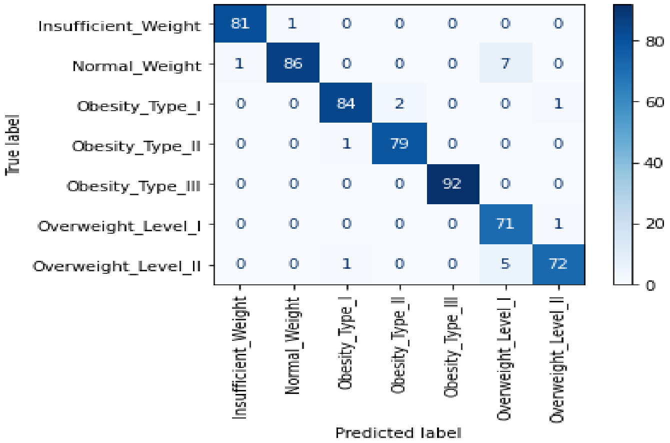
In order to determine best parameter model was optimized via gridsearchcv and best parameters where obtained. The model results are as follows:

- Best parameters : 'C': 20, 'kernel': 'linear'
- Train score : 96.1%
- Test score : 94.9%
- F1-Score : 97%
- Precision : 96.6%
- Recall : 96.7%

From above we can observe that the model is pretty good since all the scores are good and model is not overfitted.

Confusion Matrix

Figure 17: Confusion matrix of SVM model



The overall model result is given below:

Table 16: Model best parameter with their scores

model	**Train_score**	**Test_score**	**best_params**
SVM	96.1%	94.9%	‘C’: 20, ‘kernel’: ‘linear’
Random Forest	89.1%	81.2%	‘criterion’: ‘gini’, ‘min_samples_split’: 2, ‘n_estimator:50
KNN	84.4%	75.3%	‘leaf_size’: 10, ‘n_neighbors’: 5

Thus from above result, we can observe that Random forest and support vector machine(SVM) model has better score.

Classification report

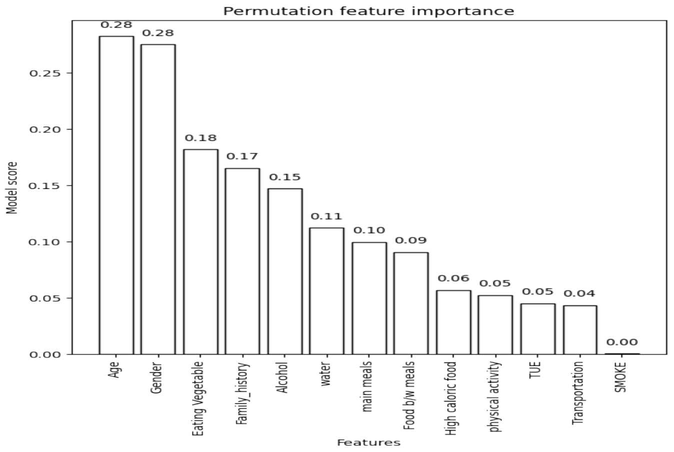
Table 17: Average performance evaluation of the techniques used

Techinque	Accuracy	MCC	F1-Score	Precision	recall
SVM	95.5%	0.97	97%	96.6%	96.7%
Random Forest	85.1%	0.82	81.1%	81.5%	82.14%
KNN	80.1%	0.61	81.1%	81.5%	82.14%

On comparing the classification report of KNN, random forest and support vector machine, support vector machine is doing better in performance. Since the precision, recall and f1-score of support vector machine is much better than random forest. Also Matthews correlation coefficient is good for support vector machine. Therefore, the support vector machine model is a good fit.

Using Permutation Importance the feature importance was calculated to analyse which attribute leads to obesity.

Figure 18: Variable Importance of machine learning model



From bar graph we can observe that all attributes has contribution to the model. The feature importance calculations show us that age, gender,family history, consumption of vegatables and alcohol had the largest influence on the models above. Other factors that had a large influence on one of the above models were the consumption of water, number main meals, food before meals, physical activity frequency.

Factors that were not impactful across the board were smoking and calorie consumption monitoring.

5.5 Conclusion

(a) Physical factors like age, heredity and physical activity frequency of the individual does play an important role in explaining obesity of individual. Thus, these are the physical factors that influence to obesity in individual.

(b) Increase in ‘Age’ is associated with increase in the probability of being Obesed and increase in physical activity is associated with a decreased probability of being obesed.

(c) When it comes to individual habits factors like consumption of high caloric food, Vegetable, Water and time spent on technology devices by the individual does play an important role in explaining obesity.

(d) Thus increase in ‘High caloric food and water’ is associated with increase in the probability of being Obesed. Thus, these are the habits that influence to obesity in individual.

(e) Consumption of Vegetable is associated with a decreased probability of being obesed.

(f) The data has clustering tendency. Thus, we got 4 cluster with different character possesed by them.

(g) From cluster analysis it is found that subjects having less BMI has a bit healthy life routine and it is evident since physical activity frequency and consumption of vegetables is more compared to other segments even though respondents has habit of consumption of caloric food.

(h) Subjects having high BMI has no healthy life routine and it is evident since subjects are not involved in physical activity, has habit of eating high caloric food and frequent consumption of alcohol.

(i) Three classification model where built named ‘K Nearest Neighbour(Knn), Random Forest and Support Vector Machine’.

(j) The best parameters for the model where found via gridsearchcv and the best parameters obtained where used to fit the model.

(k) The accuracy of K Nearest Neighbour model is 75.3%, Random Forest is 85.1% and of Support vector machine(SVM) is 95.1%. On camparing performance metric, Support vector machine performed better than Knn and Random forest. Thus, Support vector machine model is best fit for this data.

(l) By the analysis, the important features which where contributing to model were determined, and it was confirmed that indiviual habits as well as physical factors has effect/influence on obesity.

6 Chapter 6

Conclusion

The following are the overall conclusion:

(a) Ageing is associated with an increase in obesity. Adults who have a normal BMI often start to gain weight in young adulthood and continue to gain weight until they are ages 60 to 65.

(b) The chances of being obesed are greater if one or both of our parents are overweight or have obesed.

(c) Obesity results from energy imbalance. Thus, for good health one should get the equivalent of moderate-to-vigorous physical activity each week.

(d) Age, Heredity(Family history with overweight) and Physical activity frequency of the individual does play an important role in explaining obesity. Increase in 'Age' is associated with increase in the probability of being Obesed. Whereas, increase in physical activity is associated with a decreased probability of being obesed.

(e) In medical point of view, increase in 'High caloric food' consumption is associated with increase in the probability of being Obesed. Which is true in our study also.

(f) Also, usage of technology is one of the risk factor that leads to increase in obesity. But in our study it is observed that it is associated with a decreased probability of being obesed.

(g) Consumption of vegetables which are in rich in fibre helps in loosing weights. Which is true in our study.

(h) In our study it is observed that, individuals having least BMI has better daily routine interms of their habits. Individuals having more BMI value are not involved in physical activity but has habit of eating high caloric food and frequent consumption of alcohol.

A limitation of present research is that the data set used in this research consists of 2,111 records, which was not balanced in the same proportion of data for the seven levels of obesity, had a smaller number of records of people with underweight and overweight I. This study is to help individuals to realise that healthy eating habits and having healthy daily life routine are important for their health. However, it is necessary to confirm if the factors are risk factors or reducers, so the governments and health organisations are able to publish a detailed guideline based on scientific facts.

7 Bibliography

(a) Bonanno, A. and Goetz, S.J. (2012). *Food Store Density, Nutrition Education, Eating Habits and Obesity*.vol.15.

(b) Gupta, P. (2011). *Obesity: An Introduction and Evaluation*. Journal of Advanced Pharmacy Education Research

(c) Abduelkarem, A.R. (2020). *Obesity and its associated risk factors among school-aged children in Sharjah.* UAE.

(d) Kopelman, P.G. (2000). *Obesity as a medical problem*.vol.404.

(e) Lindstrom.M (2008). *Means of transportation to work and overweight and obesity:A population-based study in southern Sweden*. Sweden.

(f) Crosnoe, R. (2007). Gender, Obesity, and Education. *University of Texas at Austin,* Vol. 80, 241–260.

(g) Goodarzi, M.O. (2017). Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. USA.

(h) Bray, G.A. (2004). Medical Consequences of Obesity. *Pennington Biomedical Research Center, Baton Rouge, Louisiana 70808*

(i) Sahib, A.S. (2016). Eating Behavior in a Sample of Overweight and Obese: A Cross Sectional Study.*Journal of Obesity and Weight-loss Medication*

(j) https://www.ibm.com/in-en/topics/logistic-regression

(k) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6088226/

(l) https://nutritionj.biomedcentral.com/articles/10.1186/1475-2891-3-3

(m) https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

(n) https://medium.com/analytics-vidhya/implementation-of-principal-component-analysis-pca-in-k-means-clustering-b4bc0aa79cb6

(o) https://medium.com/more-python-less-problems/principal-component-analysis-and-k-means-clustering-to-visualize-a-high-dimensional-dataset-577b2a7a5fe2

(p) https://www.datacamp.com/tutorial/random-forests-classifier-python

(q) https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

(r) https://www.sciencedirect.com/topics/mathematics/wald-test

8 Appendix

Codes of techinques and model used for the analysis in the project

- **Chi-square test of independence:**

```
tab1 = table(data1$Obesed_Y_N,data1$Age_Grp)
test1 = chisq.test(tab1)
```

- **Fisher exact test:**

```
tab2 = table(data1$Obesed_Y_N,data1$MTRANS_)
test2 = fisher.test(tab2)
```

- **Logistic regression**

```
library(caret)

'%ni%' = Negate('%in%')
options(scipen=999)

set.seed(100)
trainDataIndex = createDataPartition(phy_dta$Obesed_Y_N,
p=0.7,list=FALSE)

trainData = phy_dta[trainDataIndex, ]
testData = phy_dta[-trainDataIndex,]

set.seed(100)
up_train = upSample(x=trainData[, colnames(trainData)
  %ni% "Class"],y = trainData$Obesed_Y_N)


logitmod1 = glm(Obesed_Y_N~Gender + Age +Family_history+FAF,
family = "binomial",data=trainData)

trainData = as.data.frame(trainData)
summary(logitmod1)
```

- **Test the significance of model and regressors of logistic model**

```
pchisq(deviance(logitmod1), df=df.residual(logitmod1),
lower.tail = FALSE)

library(survey)
regTermTest(logitmod1,"Age")
regTermTest(logitmod1,"Family_history")
regTermTest(logitmod1,"FAF")
```

- **K-Means clustering**

```
# kmeans: looking for the elbow -
# compare number of clusters by their inertia scores
inertia_scores = [KMeans(
                   n_clusters=k,
                   init='k-means++',
                   n_init=10, max_iter=100, random_state=RNDN). \
                   fit(dfa3).inertia_ \
                   for k in range(2,nK)]


dict_inertia = dict(zip(range(2,nK), inertia_scores))
print("inertia scores (sum of squared errors) by number of clusters:")
_ = [print(k, ":", f'{v:,.0f}') for k,v in dict_inertia.items()]

# scree plot: look for elbow
plt.figure(figsize=[8,5])
plt.plot(range(2,nK), inertia_scores, color="blue")
plt.title("WCSS (sum of squared errors) vs. number of clusters")
plt.xticks(np.arange(2,nK,1.0))
plt.xlabel("Number of clusters K")
plt.ylabel("WCSS(inertia)");

#model

model = KMeans(n_clusters=4, random_state=RNDN)
clusters = model.fit_predict(dfa3)
labels = pd.DataFrame(clusters)
labeleddata = pd.concat((data2,labels),axis=1)
labeleddata = labeleddata.rename({0:'labels'},axis=1)
```

- **K-Nearest Neighbors Classifier model**

```
from sklearn.neighbors import KNeighborsClassifier

knn =  KNeighborsClassifier()
params = {
    'n_neighbors': [x for x in range(5,20)],
    'leaf_size':[10,20,30,40,50,60]
}
```

```
from sklearn.model_selection import GridSearchCV

clf2 =  GridSearchCV(knn, params, cv=10,
return_train_score=False,verbose=10,n_jobs=1)
clf2.fit(X_train, y_train)
clf2.best_params_

knn1 = KNeighborsClassifier(leaf_size= 10, n_neighbors= 5)
knn1.fit(X_train, y_train)

y_test_pred = knn1.predict(X_test)
y_train_pred = knn1.predict(X_train)
```

- **Random Forest Model**

```
rf =  RandomForestClassifier()
params = {
    'n_estimators': [5,10,20,50],
    'criterion':['gini','entropy'],
    'min_samples_split':[2,3,4]
}

clf3 =  GridSearchCV(rf, params, cv=10,
 return_train_score=False,verbose=10,n_jobs=1)
clf3.fit(X_train, y_train)
clf3.best_params_

rf = RandomForestClassifier(n_jobs=-1,n_estimators=50,
criterion= 'entropy',min_samples_split= 2)
rf.fit(X_train, y_train)

y_test_pred = rf.predict(X_test)
y_train_pred = rf.predict(X_train)
```

- **Support Vector Machine**

```
from sklearn.svm import SVC
svc = SVC()

params = {
    'C': [1,10,20],
    'kernel': ['rbf','linear']
}
clf1 =  GridSearchCV(svc, params, cv=10,
return_train_score=False,verbose=10,n_jobs=1)
```

```
clf1.fit(X_train, y_train)
clf1.best_params_
svc_bvt=SVC(C=20,kernel="linear")
svc_bvt.fit(X_train,y_train)
```