Obesity Prediction using Ensemble Machine Learning

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Kapil Jindal
(Reg no: 801532025)

Under Supervision of
Dr. Niyati Baliyan
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004
July 2017

Declaration

I hereby certify that the work which is being presented in the thesis titled, "*Obesity Prediction using Ensemble Machine Learning*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering submitted in *Computer Science and Engineering Department* of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Dr. Niyati Baliyan and refers other researchers' work which are duly listed in the reference section. The matter presented in the thesis has not been submitted for the award of any other degree of this or any other University.

(Kapil Bindal)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Niyati Baliyan)

Assistant Professor

CSE Department

Abstract

At the present time, obesity is a serious health problem which causes many diseases such as - diabetes, cancer and heart ailments. Obesity, in turn, is caused by accumulation of excess fat. There are many determinants of obesity, namely, age, weight, height, and Body Mass Index. The value of obesity can be computed in numerous ways, however, they are not generic enough to be applied in every context (such as to a pregnant lady or to an old man) and yet provide accurate results. To this end, we employ the R ensemble prediction model and implement the same using Python interface. It is observed that on an average, the predicted values of obesity are 89.68% accurate, which is an improvement over previous such works. The Ensemble Machine Learning based prediction leverages Generalized Linear Model, Random Forest and Partial Least Squares. The current work can further be improvised to predict other health parameters and recommend corrective measures based on obesity values.

Keywords: Body Mass Index, Obesity, Bone Mass, Machine Learning Models.

Acknowledgement

First of all, I would like to thank the Almighty, who has always guided me to work on the right path of life. This work would not have been possible without the encouragement and able guidance of my supervisor **Dr. Niyati Baliyan**. I thank my supervisor for her time, patience, discussions and valuable comments. Her enthusiasm and optimism made this experience both rewarding and enjoyable.

I am equally grateful to **Dr. Maninder Singh**, Associate Professor, and Head, Computer Science and Engineering Department, a nice person, an excellent teacher and a well-credited researcher, who always encouraged me to keep working well and always advised me with his invaluable suggestions. I will be failing in my duty if I do not express my gratitude to **Dr. S.S. Bhatia**, Professor and Dean of Academic Affairs, Thapar University, for making provisions of infrastructure such as library facilities, computer labs equipped with Internet facilities, immensely useful for the learners to equip themselves with the latest in the field. I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love, and affection, which made my stay at Thapar University memorable.

Last but not least, I would like to thank my family whom I dearly love and without whose blessings none of this would have been possible. To my parents, I own thanks for their care and encouragement. I would also like to thank my brother since he insisted that I should do so. I would also like to thank my close friends for their constant support.

Date: July 2017

Place: Thapar University, Patiala

(Kapil Jindal)

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BMI	Body Mass Index
BMR	Basal Metabolic Rate
RMR	Resting Metabolic Rate
PLS	Partial Least Square
GLM	Generalized Linear Model
RF	Random Forest
RDA	Recommended Dietary Allowance
OLTP	Online Transaction Processing

Chapter 1

Introduction

Obesity is the condition of overweight or having excess weight. Obesity is related to Body Mass Index (BMI) however, the equation of BMI is same for every type of human, through obesity will be changed [1], for example, for a pregnant lady BMI will use same equation, however, the result of obesity is different because pregnant lady will gain an average weight of 25-35 pounds or 11-15 kg. For a normal person, it depends upon the BMI range. There are three classes of obesity - class 1, class 2 and class 3. The class depends upon the BMI range [2].

An excess of body fat may lead to obesity, which causes some diseases. If you know about your body fat you can prevent the obesity problem. We intend to use real data sets, either from existing websites or collect data from a sample population at/around Thapar University campus. The data would typically be that of various body circumference measures which will be less in number yet rich in information content. This data would then be utilized to gain knowledge about obesity levels. There are various risks which are caused by obesity [3]. These are:

- Heart disease
- Increase in blood pressure
- Diabetes
- Gall bladder disease
- Stroke
- High cholesterol
- Some cancers

There are a lot of ways to prevent the obesity [3] as follows:

a) **Exercise regularly:** Regular physical activity is imperative to stay active and maintain a healthy weight. Physical activity not only reduces weight, it also has other benefits. It decreases the risk of heart disease, blood pressure, diabetes,

stress, etc. However, in today's time, people stay busy in watching TV, Internet surfing and forget exercising. Without physical activity, we are not able to burn the energy that we get through food and later, this energy is converted to body fat.

b) **Eat less carbohydrate:** We should avoid eating much carbohydrates, because consuming more carbohydrates leads to obesity. Carbohydrates include ice-cream, fruit juices, bread, rice, potatoes, corn, sugary sweets etc. Later, carbohydrate via insulin is converted into fat. Instead of eating carbohydrates, we should prefer fiber-rich food to manage obesity. Fiber not only reduces weight, it also prevents heart diseases.

c) **Identify the right calories:** Healthy diet plan has safe calories, which helps in reducing weight. Human weight is determined by a balance between calorie intake and energy expenditure. The three main sources of calories are fat, carbohydrates and protein. Simple carbohydrates like rice, sugar, soda make us fat. Proteins are necessary for augmenting muscle mass. It is very important to identify bad or good calories in your diet plan. We should have a nutritious diet containing vitamins, minerals and essential fats.

d) **Eat more protein:** Protein is a very good nutrient for a healthy body with reduced weight. If we take more protein, it boosts metabolism, which helps in burning off more calories. Burning off calories helps in suppressing the appetite and changing the weight regulating hormones. Protein is of two kinds based on its source - plant based and animal based. Protein from vegetable sources offers fiber, minerals and vitamins, etc. To achieve and maintain a healthy weight it is imperative to combine good quality protein and low carbohydrates in every meal.

e) **Protein Recommended Dietary Allowance:** Also known as Protein RDA, they are used to calculate daily need of protein in grams. They are divided into two categories; first for a normal person and the other for athletes.

Protein RDA depends upon the body weight and work done by the body. To calculate the protein RDA for a normal person, use equation 1.1 [3] as

$$\text{Protein RDA} = \text{weight} \times 0.8 \quad (1.1)$$

To calculate a protein RDA for an athletic person, the equation depends on work done by him/her; which further has a range from 1.4 to 1.8, according to work done by the athlete. This is given by equation 1.2 [3] as

$$\text{Protein RDA} = \text{weight} \times \text{range} \quad (1.2)$$

Where,

weight is in kg

Protein RDA is in grams per day

f) Eat more fiber-rich food: Fiber is not digested by the body so it does not impart you more calories. It is imperative that you chew fiber food for a long period of time because prolonged chewing will give the signal to your brain that you have eaten enough. Hence, you can stop eating more food than you require. A diet high in fiber has low energy density, which means fewer calories in a particular food. Less number of calories will help in reducing weight faster.

g) Check your weight regularly: People who check their weight regularly are more successful in reducing weight, because with monitoring they can know whether their efforts to reduce weight are working or not.

h) Drink water: Drinking more water burns off more calories and hence reduces more weight. Various studies show that drinking water before each meal reduces 2 kg weight over a 12 week period. Water is 100% calorie free and will assist you to suppress your appetite. Water also detoxifies all impurities present in the body. Doctors suggest drinking 8-10 glasses of water in a day.

i) Get rid of fast food: Fast food is rich in calories and low in nutritional value so it augments your weight. In order to maintain weight, it is important that you eat less fast food and replace it with fruits, vegetables, nuts etc. Fast food contains a

high amount of fat and sugar but does not contain any minerals or vitamins that are important for health. Fast food reduces the quality of diet and increases the chances of the obesity. Fast foods are basically high in calories, sugar, carbohydrates, fat and salt.

j) **Stay active:** If we do regular physical activity, it not only reduces obesity, but also helps in decreasing the risk of heart disease, blood pressure, diabetes, stress etc. If we want to stay physically fit, both healthy diet plan and exercise is imperative. They will help to burn off calories by suppressing appetite.

k) **Eat slowly:** We should eat only when we need to have food. If we really want to reduce weight, eating slowly and mindfully can help you eat less and reduce weight. When we consume food quickly, our brain fails to register how much we have actually eaten and we end up eating too much.

1.1 Factors that affect the rate of obesity

In the following sections, we will discuss the determinants of obesity rate (section 1.1), causes of obesity (section 1.2) and data mining approach of classification to predict obesity levels (section 1.3).

1.1.1 Body Mass Index (BMI)

Many people suffer from overweightness, i.e., obesity, without even being aware of how to check obesity/overweightness. BMI his/her depends upon body mass (weight) and height of the body. Every person has different BMI according to his/her body weight and height. There are two ways to calculate BMI [3].

Metric: The body weight mass is in kilograms and body height is in meters, the metric formula for calculating BMI is given by equation 1.3 [4].

$$\text{BMI} = \text{Weight} / (\text{Height} \times \text{Height}) \tag{1.3}$$

Imperial: In this method, the body mass is in lb and body height is in inches, the imperial formula for calculating BMI is given by equation 1.4 [4].

$$\text{BMI} = (\text{Weight} / (\text{Height} \times \text{Height})) \times 703.0704 \quad (1.4)$$

Obesity levels categorizes people as follows:

- **Underweight:** If BMI of a person is less than 18.5 then the body is underweight but for a pregnant lady, they can be recommended weight gain 28-40 lbs (near about 13 to 18 kg).

- **Normal:** If a person has BMI between 18.5 and 24.9, then the body is normal but for a pregnant lady, they can be recommended weight gain 25-35 lbs (near about 11 to 16 kg).

- **Overweight**: This can further be divided into two categories:
  - **Overweight (At Risk):** If a person's BMI is between 25 and 27.9, then the body is overweight, but for a pregnant lady, they can be recommended weight gain 15-25 lbs (near about 7 to 11 kg).
  - **Overweight (Moderately Obese):** If a person's BMI is between 28 and 29.9, then the body is overweight, but for a pregnant lady, they can be recommended weight gain 12-23 lbs (near about 6 to 10 kg).

- **Obese:** If a person's BMI is more than 30, then the body is obese, but for a pregnant lady, they can be recommended weight gain 11-20 lbs (near about 5 to 9 kg).

1.1.2 Basal Metabolic Rate (BMR)

BMR depends upon age, weight, height, gender. It defines the rate of energy consumed by the human body [5]. According to Harris- Benedict the equation 1.5 and 1.6 [6,7] of BMR is divided into two parts, one for male another for female.

Metric Formula:

If the gender is male:

$$\text{BMR} = (\text{W} \times 13.7) + (\text{H} \times 5) - (\text{A} \times 6.8) + 66 \quad (1.5)$$

If the gender is female:

$$\text{BMR} = (\text{W} \times 9.6) + (\text{H} \times 1.8) - (\text{A} \times 4.7) + 66 \quad (1.6)$$

Where, W denotes weight (kg)

H denotes height (cm)

A denotes age (year)

Imperial Formula:

If the gender is male:

$$\text{BMR} = (\text{W} \times 6.23) + (\text{H} \times 12.7) - (\text{A} \times 6.8) + 66 \quad (1.7)$$

If the gender is female:

$$\text{BMR} = (\text{W} \times 4.35) + (\text{H} \times 4.7) - (\text{A} \times 4.7) + 655 \quad (1.8)$$

Where, W denotes weight (lb)

H denotes height (inches)

1.1.3 Resting Metabolic Rate (RMR)

RMR depends upon the age, weight, height, gender. RMR is also known as Resting Energy Expenditure and it is denoted by REE [8]. According to Mifflin-St Jeor, the equation 1.9 and 1.10 [3] of RMR is divided into two parts.

If the gender is male:

$$\text{RMR} = (\text{W} \times 10) + (\text{H} \times 6.25) - (\text{A} \times 5) + 5 \quad (1.9)$$

If the gender is female:

$$\text{RMR} = (\text{W} \times 10) + (\text{H} \times 6.25) - (\text{A} \times 5) - 161 \quad (1.10)$$

Where, W denotes weight (kg)

H denotes height (cm)

A denotes age (year)

According to World Health Organisation (WHO), Food and Agriculture Organisation (FAO) and United Nations Organisation (UNO) some more equations are used to calculate RMR [3]. By using age gap of males and females some equations are obtained as follows:

If the gender is male:

- Age is below 30 years:

$$\mathrm{RMR} = (\mathrm{W} \times 15.4) - (\mathrm{H} \times 27) + 717 \quad (1.11)$$

- Age is between 31 and 60 years:

$$\mathrm{RMR} = (\mathrm{W} \times 11.3) + (\mathrm{H} \times 16) + 901 \quad (1.12)$$

- Age is above 60 years:

$$\mathrm{RMR} = (\mathrm{W} \times 8.8) + (\mathrm{H} \times 1128) - 1071 \quad (1.13)$$

If the gender is female:

- Age is below 30 years:

$$\mathrm{RMR} = (\mathrm{W} \times 13.3) + (\mathrm{H} \times 334) + 35 \quad (1.14)$$

- Age is between 31 and 60 years:

$$\mathrm{RMR} = (\mathrm{W} \times 8.7) - (\mathrm{H} \times 25) + 865 \quad (1.15)$$

- Age is above 60 years:

$$\mathrm{RMR} = (\mathrm{W} \times 9.2) + (\mathrm{H} \times 637) - 302 \quad (1.16)$$

Where, W denotes weight (kg)

H denotes height (m)

1.2 Primary causes of obesity

1.2.1 Excess Calories

A calorie is a unit of energy provided by the food. Whatever you are eating, fats, carbohydrates, sugar, protein all are calories. A person's weight is determined by the stability between calorie intake and use of energy. To maintain the body weight, it is important to strike a balance between calorie intake and calorie used. Foods high in sugar or fat are the sources of obesity [3]. If a person eats more calories than he or she requires, then it increases the person's weight and if a person eats fewer calories than he or she requires, then it reduces the person's weight. If we do not burn off energy through exercise, much of this energy will be stored as fat in the body.

1.2.2 Excess Carbohydrates

Carbohydrates include milk, ice-cream, fruit juices, bread, rice, potatoes, corn, sugary sweets etc. Low carbohydrate diets are good for weight loss. There are a lot of studies which show that quality and not quantity of carbohydrates determines whether a person gains weight or not [3]. Besides causing obesity, carbohydrates also cause diabetes and cardiac diseases. A fiber-rich diet is important to manage obesity. The best way to maintain good health is to combine a balanced diet that includes fruit, vegetables, milk and cheese with exercise.

1.2.3 Sedentary lifestyle

Lack of physical activity is also a source of obesity. In this modern time, people do not like walking, but rely on their cars to go here and there. Technology has developed so much that people remain busy in watching TV, playing games and Internet surfing but forget to do regular exercise. There are a lot of people who do sedentary jobs, they also face obesity problems.

Today, instead of walking people uses their car to go here and there. Without physical activity, we are not able to burn off the energy we get through food and

later, this energy is converted to fat in the body. Getting too much sleep also causes obesity.

A sedentary lifestyle is one with little or no physical activity. If a person does not do much of physical activity, then the person becomes obese. Fewer calories will be burnt if we move less. A sedentary lifestyle can contribute to various risks such as depression, anxiety, breast cancer, diabetes, cardiovascular disease, high blood pressure, lipid disorder, obesity, and skin problems such as hair loss.

1.2.4 Medication

Medication is also responsible for augmenting the weight. There are a lot of medications, for example - diabetes drugs, antidepressants, blood pressure drugs, steroid hormones and medicine for schizophrenia that cause obesity. The effects are not same for everybody, one person might get 20 pounds on one drug, and other might not gain any weight. Body metabolism is affected by drugs, which slows down the burning of calories. If it is indispensable to take medication, then low carbohydrate diet and exercise can help to reduce weight. Some medicines can make you feel hungrier, others might slow the body's ability to burn calories. These drugs are: antipsychotic drugs, steroid hormone drugs, antidepressants and drugs for blood pressure.

1.2.5 Poor Diet

Obesity happens as a result of poor diet, such as

1. Drinking too much alcohol
2. Packed fruit juices
3. Eating more calories
4. Eating much sugar
5. Eating large amount of fast food

1.2.6 Hormones

Research shows that if we do not sleep as much as required then obesity doubles. If we do not sleep well, there is a hormone produced called Gerlin which stimulates appetite. Lack of sleep results in less Leptin, a hormone, which suppresses the appetite.

Leptin is a hormone that sends a signal to the brain that we have eaten full and now we need to stop eating. Then, if we really want to reduce our weight, it is important that we avoid eating much sugar and exercise properly.

1.2.7 Genetics

Obesity has a very high genetic impact. Children of the obese parents are more likely to be obese than the children of lean parents. These genetics are: Cohen, alstrom, bardet-biedl, prader-willi.

Chapter 2

Literature Survey

Machine learning is a type of Artificial Intelligence in which computers have the ability to learn without being programmed explicitly [9]. The aim of machine learning is to develop computer programs that can change themselves automatically when new data is exposed. The machine learning process is same as data mining. The goal of machine learning is to make a machine which can mimic human mind and to do that it needs learning abilities. Humans can learn from the past experience but computers follow only instructions. It changes solution as per the situation. For example, chess in which the machine decides its move based on the past experience. It works on the algorithms that can learn from data and can do prediction from data. Various applications of machine learning are, optical character recognition, spam filtering, search engine, speech recognition engine, google street view, climate modeling, detection of network intruders, and self-driving cars, etc.

Knowledge discovery is a popular research area among the scientific and industrial community and is basically used for finding patterns which give knowledge. An essential step in this whole process is modeling or data mining [10]. The techniques of data mining will help in making some intelligent decisions. The process of extracting information to make it useful, novel, understandable is basically the data mining or knowledge discovery in databases. Conventional database systems are often called as OLTP (Online Transaction Processing) systems which are designed for day to day running applications to obtain maximum throughput. On the other hand, a data warehouse is just the collection of historical data which may or may not contain the whole information. As in some applications, data mining requires complete information of the customers, but that may not be saved in the database of the data warehouse [11]. Therefore, it varies from application to application. Basically, a data warehouse is used for storing a summary of information and it consists of the OLTP systems in order to support the queries of customers.

The primary reasons for using data mining are:

- There is huge growth in the OLTP data. Nowadays, the data rate is increasing and large storage of data is required.
- Data from cards like online shopping sites and data from mobile phones are increasing day by day. There are lots of transactions which are being processed through debit and credit cards.
- Development in the data made available by websites has become the biggest source of data.
- Development in the area of banking transactions, immigration transaction, utility transactions.

Based on this improvement in these areas the data mining concept becomes popular.

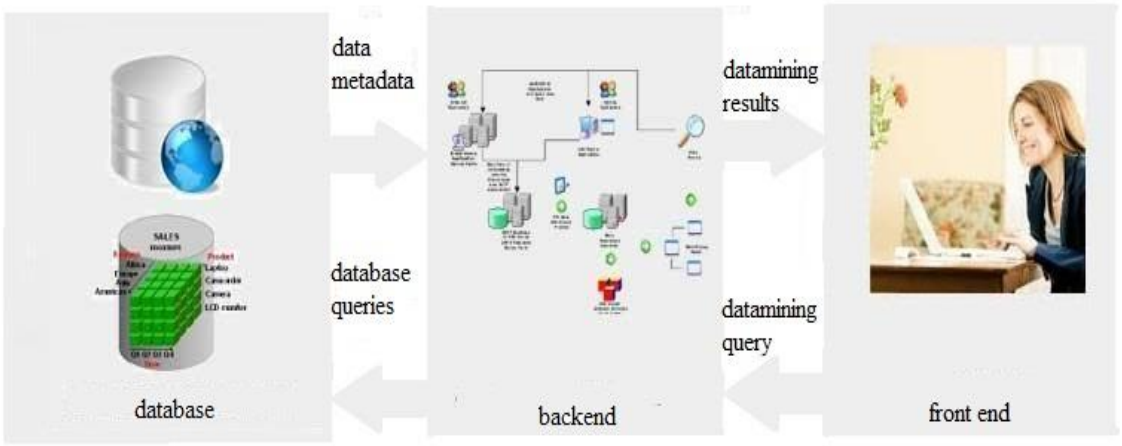


Figure 2.1: Working of data mining techniques

By applying various techniques, the new patterns will be obtained which give us new knowledge. The techniques can be predictive or descriptive this all depends upon the data mining task and application which is to be solved. If the technique which is used to make descriptive relationships is not only related to one specific characteristic of data but for all, then the descriptive model will be built.

2.1 Data Mining Techniques

Data mining algorithm consists of the following components [11,12]:

- **The model**: The model is defined by the functions, e.g. classification, clustering, association, regression etc.
- **The preference criterion**: Preference is dependent upon the data and the parameters taken for data. Too many degrees of freedom constrain the data space and smoothening is required to avoid over fitting of data.
- **The search algorithm**: It is used according to the specification of algorithms for finding parameters and the given model with data, precision, and models.

The intention of a predictive model is to construct models using support vectors, rule set, neural network and decision tree to predict the class of a recent data set as a future outcome. These models study recent and historical data, thereby allowing miners to make predictions about the future. All predictive data mining models are probabilistic in nature and can only forecast what might happen in the future. When we need to prescribe an action with this model, the business decision-maker may take this information and act further.

2.2 Function and Description of R Programming Language

R is a free and open source programming platform, mainly used for statistical computation. R also acts as the programming language (same as C or JAVA). Code written in R is very much similar to Python such as basic algebraic calculation [13]. Function and description of the machine learning models in R programming language are shown in Table 2.1.

Table 2.1: Machine learning models

Function	Description
Kernlab	Used for this model Kernlab
Glm	Used for implementing generalized linear model
Nnet	Used for implementing neural network
Fda	Used for implementing flexible discriminant analysis.
J48	Used for implementing J48.

LMT	Used for implementing the Logical model tree.
M5P	Used for implementing M5P.Package used for this model is Rweka
Mda	Used for implementing mixed discriminant analysis.
read.csv	Used for reading CSV file
set.seed	Used for setting the seed value
Predict	Used for calculating the prediction value
write.csv	Used for writing the output into CSV file
Library	Used for importing different library into R. for example library(ROCR)
Setwd	Used for setting the current working directory
Setdiff	Used for removing the desired attribute from the dataset
Rf	For implementing random forest. Package used for this model RandomForest
Ada	Used for implementing Ada boost model
Rpart	Used for implementing decision tree

R is a GNU project, it is rich in libraries. The number of packages present in R is approximately 5000. Thus, it avoids a great deal of line of code and saves time (unlike C++ and JAVA). Out of all the software tools present such as Weka, Java, C, etc., R is considered to be the best popular platform for machine learning because of its simplicity. R also displays graphical computation very impressively as compared to Mathematica and MATLAB. One important feature of R is that it is platform independent, hence R can be used in any operating system. R can also integrate with languages such as JAVA, C++ etc. Inside R, there are about 200 machine learning models present [13]. R is also used for finding missing values. Popular companies which use R are Bing (used for increasing the awareness in social search), Google (making online advertisements effectively), and Facebook (status analysis, prediction of friends or colleague interaction).

R was formed in 1993. It is open source software used for statistical and graphical computation. R is available on different interfaces which are:

- R studio: It is an Integrated Development Environment (IDE). For running R studio, first install R 3.2.1. This is the stable version which was released on 18 June 2015.
- Rattle: It is also a Graphical User Interface (GUI). It is basically used for data mining (feature selection. feature extraction etc.)
- R commander: It is also GUI. The stable version of R commander is 2.0.0,

released on 21 August 2013.

- RKWard: This is the extended GUI and IDE for R. It is basically written in C++ and ECMA script.

2.3 Machine Learning Techniques

It is a branch of Computer Science which deals with recognition and categorization. It is applicable to all areas of Artificial Intelligence. Machine learning was first developed in 1950. It is the study of how the computer realizes the behavior of the human being. An example of the application of machine learning is email filtration. The email can be filtered by two methods, one is machine learning another one is non-machine learning [9]. The categorization of email filtration describes the Heuristic approach, signature approach, hash-based, and traffic analysis while the machine learning approach is divided into two types - complete and complementary [9]. The complete approach is further divided into unified model and ensemble model. Schematic diagram of filtering techniques is shown in Figure 2.2.

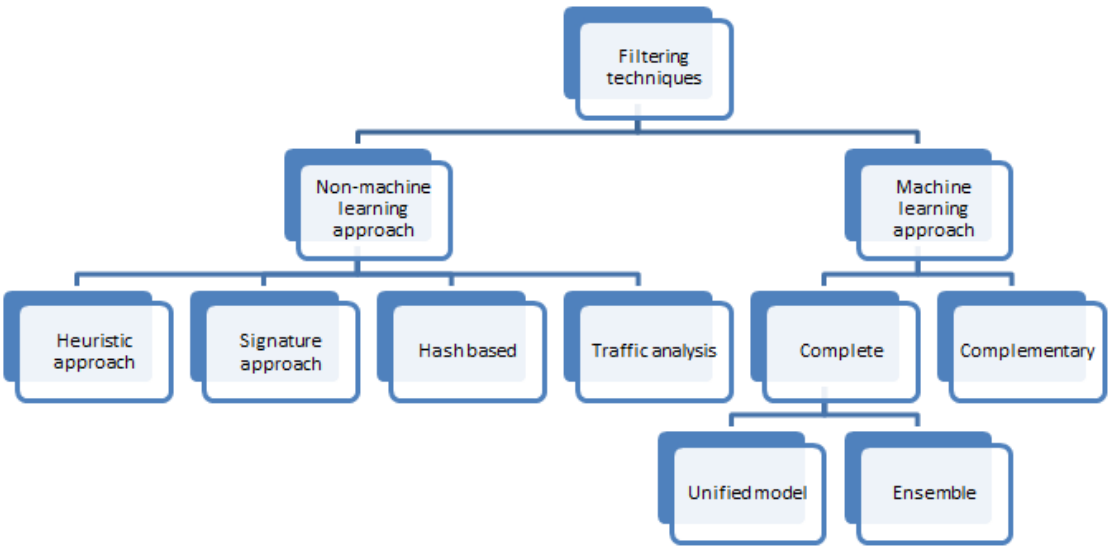


Figure 2.2: Filtering Techniques

While algorithms under ensemble model include - stacked generalization, boosting tree, and hidden Markov model, the algorithms under complementary approach are the adaptive and trust network.

Machine Learning is a relatively new field. Broadly, machine learning can be classified into three main categories [13]:

i. Supervised Machine Learning

ii. Unsupervised Machine Learning

iii. Ensemble Learning

2.3.1 Supervised Machine Learning

Let us consider an example to understand the underlying concept of supervised learning. Suppose that you are a child and your father tells you that this particular animal is a dog and tells you a few things about the dog. Later, if you see a new type of dog you will be able to identify it as a dog, this is supervised learning [3]. As another example, in a school there is a teacher to help you to learn concepts in such a way that if any unseen problem comes then, you might still be able to identify that problem. If we are training our machine for every input with the corresponding target, then it is called supervised learning [3]. After sufficient training, the system will be able to provide a target for new input. If there is continuous target space it is called regression problem.

- **Classification**: A data mining approach that allows items in a collection to be mapped to target categories or classes is known as classification. In general terms, it can be divided into two steps. First one is learning step which consists of the predetermined set of classes [3]. These concepts are built by the previous records and training database instances. The second step involves testing in which data sets are being tested the verification of learned model. e.g., a classification model could be useful for measuring the performance of the students to be high, medium or low.
- **Regression:** Statistical regression is one of the predictive data mining models that analyzes the dependency among attribute values. This is

the main difference between regression and classification. In other words, target attribute containing continuous values requires a regression technique. Continuous attributes are different from discrete ones which take finite or countably infinite values, i.e., those which can be mapped to natural numbers [11]. The most commonly used regression is linear regression. In this, the line that minimizes the average distance among all the points from the line, is sought.

2.3.2 Unsupervised Machine Learning

Unsupervised learning, as the name suggests is related to machine learning by self, with some algorithms without the interference of the user giving the input of data. The most popular unsupervised machine learning technique is clustering. It is the technique in which the distance formula plays the main role in deciding the mean of the points falling under one region [9]. Different regions form when the points come in contact with each other in reference to the distance calculated.

In unsupervised learning, there is no teacher to guide you and you have to find a solution to the problem on your own. If we train our machine only with the set of inputs, then it is called unsupervised learning. The unsupervised learning techniques are:

- Clustering
- Anomaly detection
- Heian learning

2.3.3 Ensemble methods

Ensembling of models means running two or more Machine Learning/Classification models, but getting the results in a single parameter, so that we can have better results. By ensembling, different model analysts can remove some drawbacks which are caused when the models are run individually on the dataset [9]. Ensembling is a supervised learning technique to provide high

degree of robustness. It performs best when the correlation value is low, i.e., models that are to be ensembled have very low linking value. It is not a good idea to ensemble models with high correlation value. Schematic diagram of ensemble model is shown in Figure 2.3.

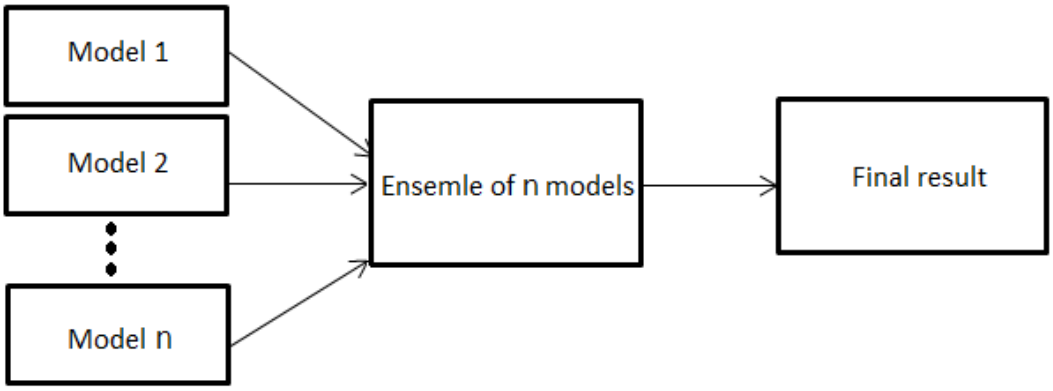


Figure 2.3: Ensemble Models

Ensemble models give better prediction values and are more stable with reduced number of drawbacks, hence providing a better decision making criterion. A good example of ensembling is Random forest models, these are a combination of various layers of decision trees. Ensemble modeling is the combination of two or more machine learning and data mining algorithms which work as a combined unit to give out the best result possible from either algorithm.

This technique is widely used these days in research in different fields. Usually, this technique increases the overall efficiency of the results because of the unity factor of the two models coming into play. However, this technique might also decrease the efficiency in cases where the data is vast and one model is better than another with a great margin. A wide variety of data sets might be thrown at the model and this will still give a decent result where one single model might not perform well at all.

The Kernel factory method of ensembling is basically an ensemble method on

kernel machine used for pattern recognition. In kernel method, a similarity function which is used for determining how similar the two machine learning algorithms are. For example, suppose there is a task for text classification. This can be done in two ways, first one is to take textual data as training data after that, calculate feature and put those features into machine learning algorithms and compute the performance of individual text and after that compute the similarity between each textual data. The second method is kernel method. In this case, define kernel function and with respect to that kernel function compute the similarity between two textual data.

Ensemble model not only combines algorithms, but also increases the performance of the algorithm [14]. Suppose that the naïve Bayes algorithm is implemented on some dataset and accuracy for a particular dataset is very less, so accuracy can be increased by two methods. First one is to use another algorithm and the second one, combine this algorithm with other algorithms for improving the performance. In this work, the second method is used for improving the performance of algorithms [11]. This application is also very useful to increase the efficiency and the accuracy of the algorithms.

2.4 Regression

For practical implementation, we use machine learning approaches. Every approach works at a dataset. The dataset is divided into two parts, first is for regression and another is for classification. For predicting obesity, we use regression type of data.

Regression analysis is used to check the relation between two or more variables, for e.g., if we have the data set (data of previous time from the current date) of sales of any company, by using this data we can predict the sale of current year, but if data is huge, then first we check the correlation by cleaning the data set (like some data is missing then it is replaced with the average value of that column). Cleaning or filtering of data set is a must for a large amount of data set otherwise accuracy reduces. There are many types of regression as explained in the following sections.

2.4.1 Linear Regression

This regression is mostly used for predicting, it works on dependent and independent variables. Dependent variables are continuous but the independent variables may have both (discrete and continuous) values. Schematic diagram [3] of linear regression is shown in Figure 2.4.

Figure 2.4: Linear Regression

They create a relation between these two or more than two variables by using regression Line, this is also known as a best fit straight line. This works on acceptable error value (error value is denoted as e) but the value of e should be less for better accuracy.

First, create a regression line then check the acceptable error value then also take those values which occur in that region [3]. After taking those values, it is ready for predicting the value or result also view the result via a graph, for example, a graph of total sales and amount of linear regression.

2.4.2 Logistic Regression

This depends on the dependent variable, which could be one of these types:

i. Binary Number (either 0 or 1)

ii. Boolean (either True or False)

iii. Binary Outcome (either Yes or No)

Mostly logistic regression is used for classification type of problems. They avoid both fitting (under and over fitting). For using regression, the data set should be large for better accuracy [3]. If we have a small size of data set, then the amount of data of success rate and failure rate should be same, otherwise, the accuracy of predicted result is very less. Schematic diagram of logistic regression is shown in Figure 2.5 [3].

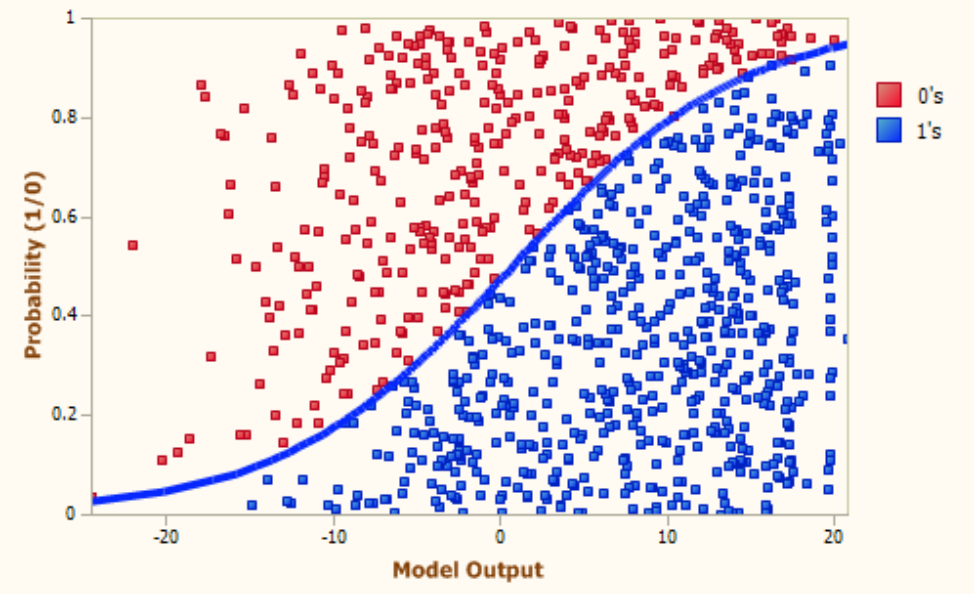


Figure 2.5: Logistic Regression

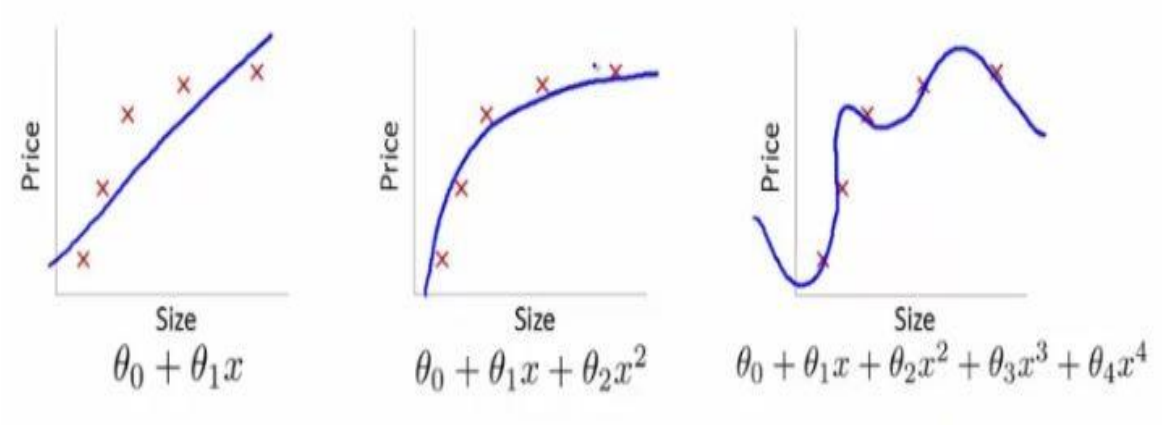
2.4.3 Polynomial Regression

This depends upon the exponential value of the mathematical expression or polynomial. If the maximum degree of the equation is greater than one, then this equation is a polynomial regression. It shows the result like best fit line, which is not a straight line. For getting a low error rate or better accuracy, the exponent of the equation should be high. The result of this technique shows data in these three ways as shown in Figure 2.6 (a, b, c) [3] respectively:

a) Under fitting / High Bias

b) Just right

c) Over fitting / High variance



(a) High bias (b) Just right (c) High variance

Figure 2.6: Polynomial Regression

2.4.4 Stepwise Regression

This regression deals with more than one independent variable. This technique is completed by with the help of an automation process. Basically, this is used to predict results with maximum accuracy by using of minimum number of variables. Schematic diagram [3] of stepwise regression is shown in Figure 2.7.

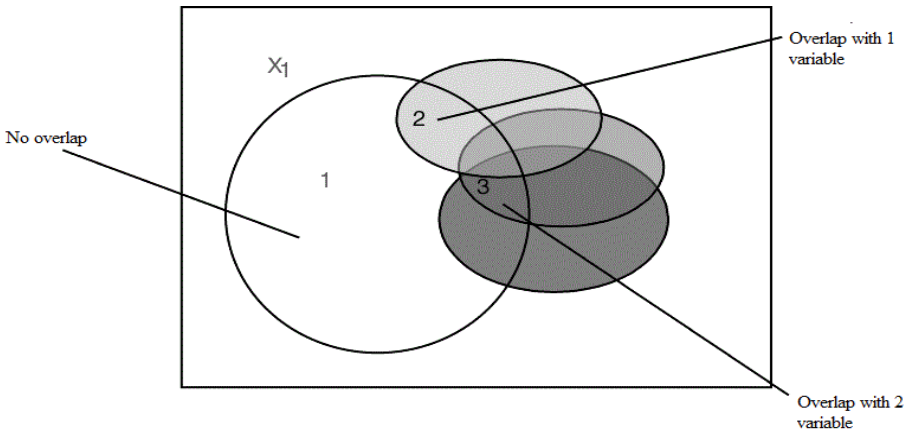


Figure 2.7: Stepwise Regression

2.4.5 Ridge Regression

When a variable is highly correlated, then we need this technique. In this regression, if the equation is linear then the result of prediction (if any error occurs) is divided into two parts: biased and variance. By using shrinkage parameter ridge regression, solve the problem of multicollinearity. Schematic diagram [3] of ridge regression is shown in Figure 2.8.

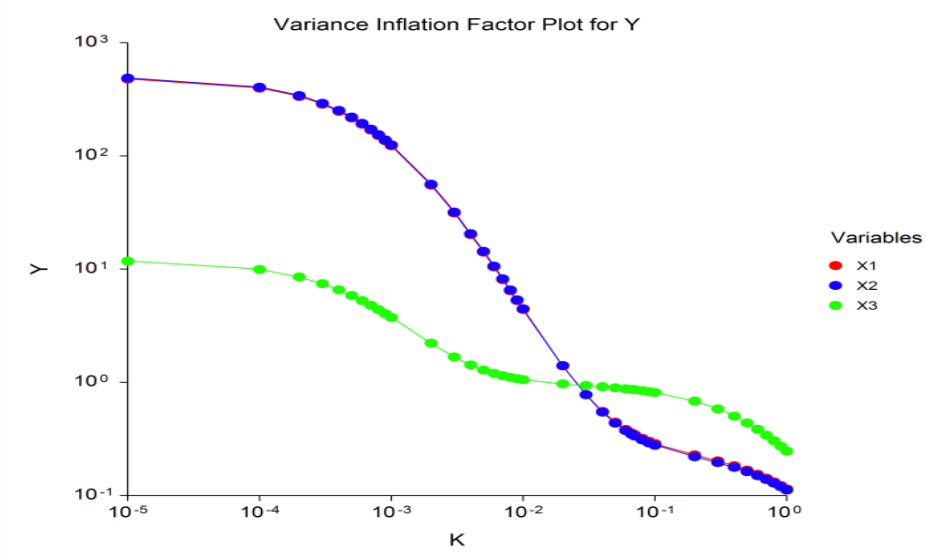


Figure 2.8: Ridge Regression

2.4.6 Lasso Regression

Lasso is defined as least absolute shrinkage and selection operator. This is very similar to ridge regression. It is used for reducing the variability and improves the result of prediction or accuracy of linear regression models. Schematic diagram [16] of lasso regression is shown in Figure 2.9.

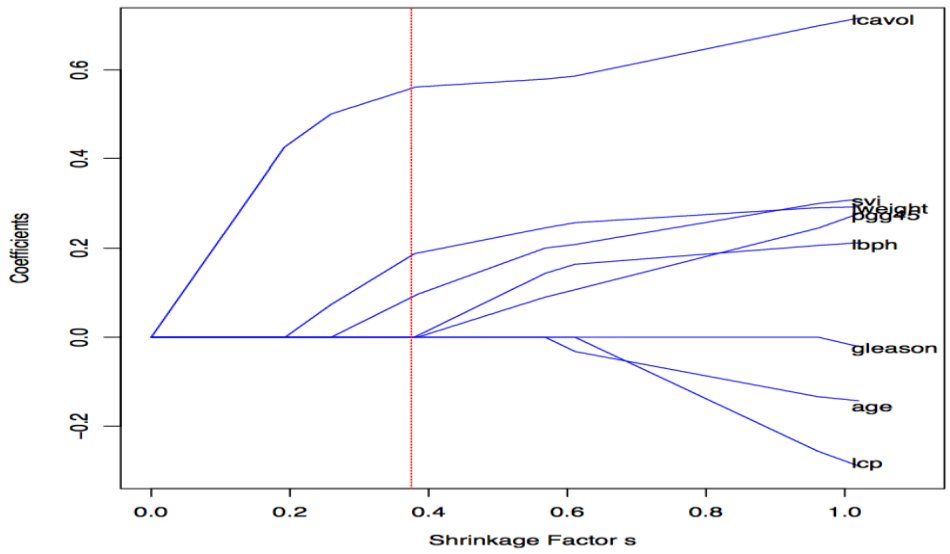


Figure 2.9: Lasso Regression

2.4.7 Elastic Net Regression

This is the hybrid of lasso regression technique and ridge regression technique. If we have multiple features which are correlated in the data set, then we use elastic net regression technique. It can suffer from multiple shrinkages. Schematic diagram [3] of elastic net regression is shown in Figure 2.10.

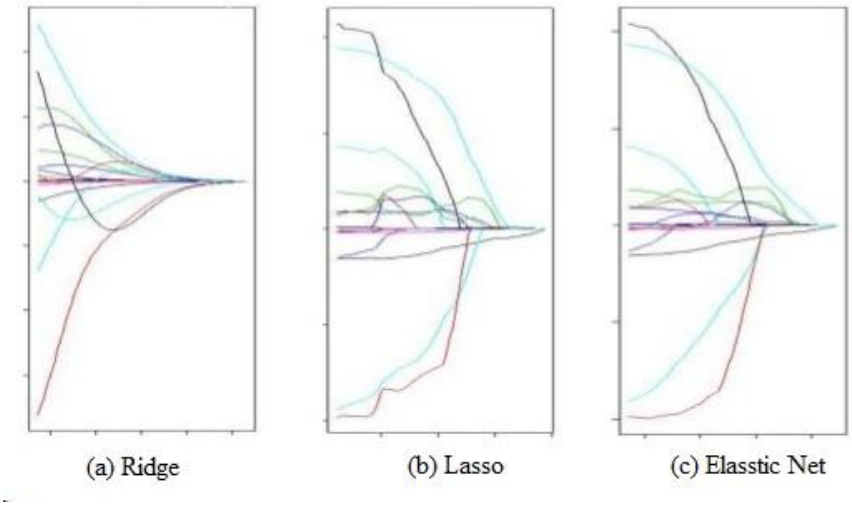


Figure 2.10: Elastic Net Regression

Chapter 3

Problem Statement

Excess of body fat may lead to obesity which causes some diseases. If you know about your body fats you can prevent the obesity problem. We can check the value of obesity of human body by using BMI. According to the range of BMI, we can describe the level of obesity. However, by using this formula we can not calculate the accurate value of obesity for these specific people.

- A pregnant lady
- An athlete person
- A physically disabled person

Let us examine how obesity affects these people. It is dependent on BMI, which further depends on the height and weight only. For females during pregnancy the average weight increase is 12 kg. According to pregnancy month, the weight of body increases. However, there is not a specific formula for calculating the obesity for a pregnant lady.

3.1 Research Gaps

For athlete people, the weight of bones is different from that of normal people, because the diet of athlete people is different and obesity is also affected according to the age of the person. At the age of 20 to 40, bone mass increases, however, after 40 years bone mass decreases. Therefore, the result of BMI range is not accurate for athlete person. For example, if a person has weight 70 kg and height 5’2” and the same weight and height of another person, however, the age is different like if the age of the first person is 70 years and the age of the second person is 25 years then the obesity level changes (minor change) because at the age of 25 to 30 years, the strength and bone size is maximum [3] but after the age of 40 years the bone mass reduces slowly. This is the major factor of age that affects obesity prediction.

Additionally, the bone mass depends upon gender [27]. If the gender is male then the bone mass of calcium of body at the age of 20-30 years is 1000-1500 grams, but if the gender is female then the bone mass of calcium of body is 600-1000 grams.

For physically disabled people, if any person is physically disabled by arm or leg, then it also affects on its body weight and it also affects the result of BMI. These are the major problems for calculating obesity for these people.

3.2 Research Objective

As a result of aforementioned open research problems, we are motivated to work in this area. Hence, we can formulate the problem statement as 'Obesity Prediction using Ensemble Machine Learning' which can further be broken into following modules.

- Train the dataset on ensemble of classifiers
- Check accuracy of ensemble classifier
- Use validated ensemble classifier to predict unseen data class (obesity level)

Chapter 4

Proposed Work

As discussed in the introductions chapters, obesity mainly depends on the weight and height, however, it also depends on age, because the human bone mass increases at the age of 20 years to 40 years, whereas, after 40 years it reduces. Therefore, if we check obesity without age like if user inputs weight as 77 kg and height as 170 cm then the model knows about the bone bass that is the reason for age to be a mandatory input. Gender and bone mass are also mandatory for a pregnant female. In males, at the age of 20 years to 40 years, bone mass ranges from 1500 to 2000 grams and in females it ranges from 1000 to 1200 grams.

4.1 Data Pre Processing

Every model is trained by using dataset and also tested with the same or different dataset. Generally speaking, if the accuracy of testing is above 95 then the model works well otherwise we need to clean the dataset [16] or change the machine learning approaches.

For cleaning the dataset, if any column is empty, then take the average value of all columns and place it as empty column's value. If nearly half of dataset is empty, then change the dataset for better accuracy because if we change the empty field with an average of all columns then maximum columns have same values and the prediction result is not good.

In Figure 4.1, we show how dataset works for training and testing the model. If we have a single dataset for training and testing, then select 75% of data for training and remaining 25% data for testing but do not forget to shuffle the data. If we do not shuffle the data then sometimes accuracy problem occurs, for example, if we have data from hospitals for patients for 4 diseases and it is sorted by disease then every disease takes 25% data of the dataset.

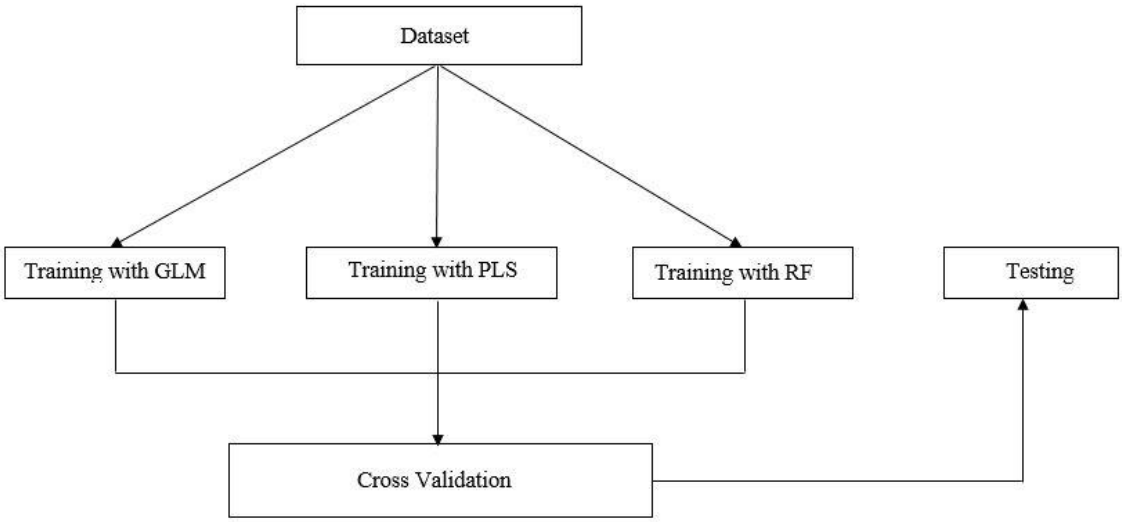


Figure 4.1: Training and Testing of dataset

If we train model without shuffling the data of dataset, then it takes first 75% of data of dataset and it is trained only for first three diseases and the last disease is excluded from for training of the model. If we test using remaining dataset, then they predict the wrong answer because they have no data on this disease. That is why shuffling is mandatory for the training of model if you have a single dataset for training and testing. If we have two different datasets for training and testing, then shuffling of data is not required. After training the model, we check on different dataset if accuracy is good then the model is valid, otherwise, change the model or clean the data.

4.2 Obesity prediction using Machine Learning

At present time, being overweight is a big problem for humans. There are many diseases like blood pressure, diabetes, cancer which are caused by fatness. There are so many reasons of fatness like oversleep, over diet, junk food etc. and on an average, fatness in youngsters is increasing day by day, however, people do not know about obesity, which is a major outcome of fatness [15].

Obesity has some stages like level 1, level 2 and level 3. If any person has obesity at level 1 then they can prevent it by exercise and diet control but after level 1 of obesity, medical treatment is also required so awareness of obesity is a must.

If obesity causes any serious disease to any person then it is imperative to control the level of obesity. Obesity largely depends on BMI, however, if we add some factors like gender, and BMR and RMR then we get better results. There is no formula to calculate obesity using BMR and RMR [6,8] also, therefore, we predict the value of obesity using BMR and RMR. For predicting the value of obesity we use machine learning technique.

For using machine learning models, first, create the dataset of obesity. Next, we take BMI, age, height, weight then we use them for predicting the result as obesity level. A sample of dataset looks as in Table 4.1.

Table 4.1: Sample dataset of BMI

BMI	**Age**	**Weight**	**Height**	**Obesity**
13.40	41	38	170	0
38.62	45	97	158.49	2.724
26.77	51	78	170.68	0.356
34.52	57	97	167.74	1.901
41.40	64	104	158.49	3.28
25.62	68	72	167.74	0.124
31.04	75	75	155.44	1.208

This type of dataset is assigned to machine learning model for training the model [2]. After training, we start testing the model, however, if we test for athletic person or pregnant female, they predict wrong output so we realize these factors are not sufficient to predict the obesity, without taking gender input it is not possible to predict obesity for a pregnant female. If we take one more input like gender, then if male then go forward, but if it is female then ask about pregnancy status. If the female is not pregnant then go forward otherwise change the current weight of the body by removing additional weight allowed during pregnancy.

According to the month of pregnancy, remove the weight from current weight of the body. If any female after pregnancy is also normal (not obese) then it is in underweight level. Obesity also depends upon the bone mass, if the age of person either male or female is 20-40 then the bone mass is maximum but after the age of 40 bones mass is

reduced. Therefore, it also affects the result of obesity because at the age of 20 – 40 the bone bass increases day by day. For males, range of bone mass is 1500 to 2000 grams, while in females, the range of bone mass is 1000 to 1200 grams.

4.3 Training the dataset

For training the model we create the dataset, but when the user enters details, they are stored in another Comma separated values(CSV) file. The model takes this file as input and gives output also in CSV file, and then we read the data from CSV file. There are two ways to solve the problem of a pregnant female for obesity. The first way is to take the dataset by adding the one more feature gender, 0 or 1, which can be for male and female, respectively. Then dataset look likes table 4.2.

Table 4.2: Sample dataset with additional feature gender

BMI	**Age**	**Weight**	**Height**	**Gender**	**Obesity**
13.40	41	38	170	0	0
38.62	45	97	158.49	1	2.724
26.77	51	78	170.68	1	0.356
34.52	57	97	167.74	1	1.901
41.40	64	104	158.49	0	3.28
25.62	68	72	167.74	0	0.124
31.04	75	75	155.44	1	1.208

This type of dataset trains the model for gender, i.e., male or female but not for a pregnant female. However, if the user selects the gender as female then interface of project asks about the pregnancy, if user select ‘yes’ then they ask about the month of pregnancy. According to user input, our interface automatically reduces the average weight according to pregnancy month from its current weight and gets the correct result of obesity.

4.4 Testing and Prediction

Now, if we want to predict the actual result with a model trained for the pregnant female, then if the user enters the input for a female it also reduces the weight according to pregnancy month of the female. If the dataset of obesity of a pregnant female is small, we use automatic method otherwise we train the model and predict the result then we take one more value for gender like 0, 1 and 2 for male, female and pregnant female, respectively.

For using this method, we need to have a huge dataset otherwise automatic method is better. The main benefit of model training is that if the user enters wrong input they never show error, they predict the result which is not accurate, whereas, if we use the automatic method they show an error for wrong input.

Table 4.3: Describe gender in details

BMI	Age	Weight	Height	Gender	Obesity
13.40	41	38	170	0	0
38.62	45	97	158.49	2	1.79
26.77	51	78	170.68	1	0.356
34.52	57	97	167.74	1	1.901
41.40	64	104	158.49	0	3.28
25.62	68	72	167.74	0	0.124
31.04	75	75	155.44	1	1.208

Compare the dataset in Table 4.3 with the previous dataset in Table 4.2 then you can see that whatever the gender is 2 or 1 with same weight and height then the prediction result is changed. This is a major difference between these dataset and model, training method is much accurate if we have a large sized dataset. However, dataset should be clean and correct, otherwise first clean the dataset and replace the empty columns with the arithmetic mean of all columns.

The regression-type dataset contains some attributes like age, gender, weight, height and BMI. After creating the dataset we apply machine learning approaches. These approaches are:

- Generalized Linear Model
- Partial Least Squares
- Random Forest

Chapter 5

Validation

5.1 Analytical Validation

For training the model, firstly check the model type. Every model works with the similar type of dataset, if model type is classification then it works with classification type of dataset. Thus, check the type of data of dataset and also the type of model. There are three types of models as discussed below.

Classification: In this type, the data should be of the same type as that of the class. Sometimes, it works with different types of data because if the dataset is small and little bit changes of data then it works with this type of data otherwise they show error at training time. Every model needs some library for predicting the result, training the data etc. Firstly install the libraries of particular model otherwise it shows error like this library missing.

Regression: In this type, the data of dataset have the type of regression. For using these types of models dataset should be large because it predicts the value in much detail. If the dataset is small then they cannot predict the class in full details.

Dual: In this type, we can use either type of dataset, i.e., either classification or regression. If we use continuous type of data then we use a dual type of model because in a continuous type of data, we are not sure about data of dataset, whether it is regression or classification.

Every model has some limitations like some models take so much time for predicting the result, while some models take less than one minute but they have less accuracy of the result. To overcome this problem we can use the hybrid technique. This is the combination of two or more than two models.

This technique works only for a large amount of dataset because here we divide the dataset according to a number of models and into an extra part for testing if we have a

single dataset for training and testing. If we have more dataset, then no need for the extra division for testing part.

In this technique, we take different models and assign the same amount of data to every model. Then take the result of every model and take the average. After taking average, this is the final result of dataset, now consider testing part. After testing, you will see the accuracy of the hybrid technique is better than the single model but it is applicable only to a large amount of dataset because if you have small dataset then it is difficult to divide it into training and testing data, so it is applicable only for a large amount of dataset.

The main problem in ensemble technique is this it takes much time than single model because here, the model executes serially and not in parallel. It takes two or more models so it takes much time. For using this technique we use only those types of models which take less time otherwise the execution of ensembling takes a lot of time, even though it is more accurate for predicting the result.

The second problem in this technique is that we can use only same type of models for example: - if we use regression type of model then every model's type will be same or dual type otherwise they can not be trained properly and can not predict the result. For overcoming this situation use only dual type models [2] because they work for both regression and classification type of data and are also used for the continuous or live data.

For better prediction result of obesity, we add some more factors like age, gender and athlete, for example:- if a person has weight 70 kg and height 5'2'' and the same weight and height of another person but the age [17] is different like first person age is 70 years and for second person age is 25 years then the obesity level has minor change because at the age of 25 to 30 years strength and bone size is maximum but after the age of 40 years the bone mass less slowly. This is the major factor of age that affects obesity prediction.

Additionally, the bone mass [27] depends upon gender, if the gender is male then the bone mass of calcium of body is 1000-1500 grams, whereas, if the gender is female then the bone mass of calcium of body is 600-1000 grams.

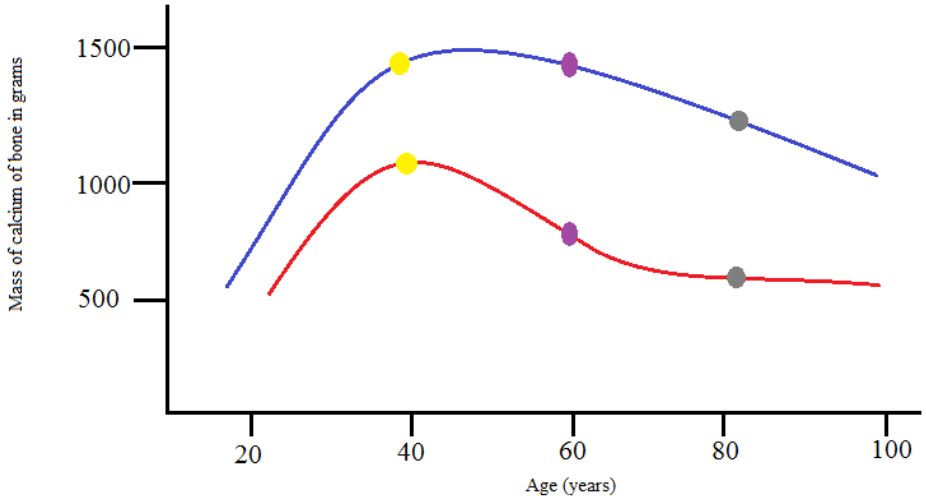


Figure 5.1: Bone Mass according to age

In Figure 5.1, the blue line is used for males and the red line is used for females. There are three stages in this graph which are denoted by circles.

1. Yellow Circle: In this stage, the bone mass increases in both male and females but the main difference is increase of calcium in grams. For males in this stage, the calcium of bone mass increases to 1000-1500 grams, however, in females, it is increased to 600-1000 grams.
2. Dark Pink Circle: In this stage, the body mass is maximum and the strength of bone is also maximum. For males, the calcium of bone mass is 1200-1500 grams and for females, the calcium of bone mass is 800-1000 grams.
3. Grey Circle: In this stage, the bone mass begins to decrease and strength of bone also decreases. For male calcium of bone mass decreases at about 900 grams and for females, it decreases at near about 400 grams.

Therefore, obesity also depends upon age [13] and gender but gender has some more categories to improve the result of obesity like an athlete (both male and female) and if gender is female then for a normal woman and for a pregnant woman the result of obesity is different. For athlete persons, the bone mass is greater than normal persons and a total number of calories intake per day is also above than normal person. This is valid for both male and female athlete as compared to normal male and female.

However, obesity additionally depends on age and gender. For instance, if the ages for two persons are 88 years and 22 years, while, the weight is same for both ages, then the BMI is same for both persons. However, the obesity level is different for both persons. In our knowledge, there is no mathematical formulation that explains and/or calculates this gap in obesity levels [28]. Therefore, we were motivated to employ machine learning techniques in order to achieve satisfactory results of obesity values in a wide variety of situations. The range of body fat as shown in Table 5.1.

Table 5.1: Body fat range according to gender

	BF % Range	**Optimal BW % Range**
Man	4 to 14 %	70 to 63 %
	22 to 24 %	57 to 55 %
	4 to 20 %	70 to 58 %
	30 to 32%	52 to 49 %
Woman	15 to 21%	63 to 57 %
	25 and over	55 to 37 %
	21 to 29%	58 to 52 %
	33 and over	49 to 37 %

In Table 5.1, athlete person takes more calories than a normal person and bone mass is also more than that of a normal person. This is a major reason that obesity also depends upon gender. Even though the normal male and female (normal and pregnant), and athlete male and female have the same weight and height but the result of obesity is different.

If we use gender factor, accuracy is improved, but if we want to improve further, then we will add some more factors like BMR. A total number of calories required per day also depends upon the body fitness. At the age of 20 years, requirement of calories is much higher as compared to the requirement of calories at the age of 60 years [21]. BMR affects the result of obesity at a minor level. If we take factor BMR then result will be improved, but if we check the obesity of same age people then BMR is not affected.

BMR is affected only if we check the obesity of different age persons. When we add the factor gender for a pregnant lady then we have two options. We can assign the data of pregnant women in the dataset or we can directly pass the previous weight of pregnant lady directly to the model. However, if we add the factor BMR then we have only one option. We can add the data of BMR directly in the dataset. We can not pass the value of BMR automatically. After adding the BMR data, dataset looks as in Table 5.2

Table 5.2: Sample of dataset with additional feature BMR

BMI	**Age**	**Weight**	**Height**	**BMR**	**Obesity**
13.40	41	38	170	1797.24	0
38.62	45	97	158.49	2098.85	2.724
26.77	51	78	170.68	2305.04	0.356
34.52	57	97	167.74	2411.43	1.901
41.40	64	104	158.49	2066.66	3.28
25.62	68	72	167.74	1993.32	0.124
31.04	75	75	155.44	1364.95	1.208

If we want to remove some factor, then we can remove factor BMR because its effect on obesity result is insignificant. Other factors such as age, weight, height, and BMI are mandatory for predicting obesity, however, for better prediction gender is an important factor. We have one more factor, i.e., RMR, similar to BMR. This also affects the result of obesity slightly, so we can neglect this factor.

After preparing the complete dataset, we decide about the machine learning model. First, we take any random model of regression or dual type and check the accuracy and total time taken by model. After taking the first model, divide the dataset into two parts, first part of the dataset for training of model and second part of the dataset for testing of the model. Divide the dataset in the ratio of 75:25 for training and testing [22]. For checking the dataset, execute the model two or three times for the same dataset because it takes random data for training. After three times check the accuracy of these three results, if the result is similar then dataset is correct otherwise clean the dataset first. Noisy dataset looks like the one in Table 5.3.

Table 5.3: Sample of dataset having noise

BMI	Age	Weight	Height	BMR	Obesity
13.40	41	38	170	1797.24	0
38.62	45	97	158.49	2098.85	
26.77	51	78	170.68	2305.04	0.356
34.52	57	97		2411.43	
41.40	64	104	158.49		3.28
25.62	68	72		1993.32	
31.04	75	75	155.44	1364.95	1.208

In this dataset, some values are missing then we change these values by taking the average of the entire column. The cleaning of dataset is only possible if some values of data are missing, however, if maximum data is empty then neglect this dataset [10]. Cleaning of dataset improves the accuracy, however, if we have correct value of this missing value then model predicts the accurate answer. An example of neglected dataset owing to a large number of missing values looks as in Table 5.4.

Table 5.4: Sample of neglected dataset

BMI	Age	Weight	Height	BMR	Obesity
13.40	41		170	1797.24	0
	45				
26.77					0.356
	57			2411.43	
41.40	64	104	158.49		3.28
	68			1993.32	
31.04	75	75	155.44	1364.95	1.208

In Table 5.4, majority of the cells are empty, and it represents a sparse matrix, so we cannot replace these values by taking the average of the entire column, hence we neglect this dataset or improve the dataset for prediction of the result of obesity.

If the time taken by the model is quite high, then change the model because this is good for forecasting or live dataset, however, if we have a static dataset then select only those

types of models which take less time. After applying the first model, try with another model which are shown in Appendix A and compare the result of this model with the result of the previous model. For better accuracy try some more models out of those given in Appendix A and select those models which have high accuracy. Finally, create a hybrid model of these models.

The combination of two or more models is known as a hybrid model. The main requirement of the hybrid model is that the dataset should be large [14]. If we take small dataset then we cannot divide it some parts for training the models, so we need large dataset. After cleaning the dataset, divide the dataset according to models and one more part for testing. A sample of the dataset for predicting obesity looks as in Table 5.5.

Table 5.5: Sample of dataset for ensembling

Obesity	**Age**	**Weight**	**Height**	**Gender**	**BMI**
0	41	38	170.68	1	13.04
0	41	47	170.68	0	16.13
0	41	56	170.68	1	19.22
0	41	64	170.68	0	21.97
0.342	41	78	170.68	0	26.71
0.766	41	84	170.68	0	28.83
1.66	41	97	170.68	1	33.3
2.28	41	105	170.68	1	36.04
3.914	41	124	170.68	0	42.57
3.46	41	137	170.68	0	47.03
0	41	38	158.49	1	15.13
0	41	45	158.49	0	17.91
0	41	53	158.49	1	21.1
0.134	41	67	158.49	1	26.67
0.892	41	74	158.49	1	29.46
0.766	41	84	170.68	0	28.83
0	41	38	170.68	1	13.04

In equation (5.1), m is used for training of m number of models, 1 is used to denote use of at least one model for testing. If you have the dataset of n number of rows and hybrid model also takes m number of models then a total division of dataset is:

$$\text{Total Division} = n / (m + 1) \quad (5.1)$$

For checking the result of obesity, we consider the hybrid model as a combination of the following three models.

- Generalized Linear Model
- Partial Least Squares
- Random Forest

5.2 Experimental Validation

For prediction of obesity, firstly we divide the dataset into 4 parts because we use three models. Three parts of the dataset for training and one part for model testing. Dataset looks as in Table 5.6.

Table 5.6: Sample of dataset assigned to model GLM

Obesity	**Age**	**Weight**	**Height**	**Gender**	**BMI**
0	41	47	170.68	0	16.13
0	41	56	170.68	1	19.22
0	41	64	170.68	0	21.97
0.342	41	78	170.68	0	26.71
0.766	41	84	170.68	0	28.83

If we take the sample of dataset, then the dataset shown in Table 5.6 is assigned to Generalized Linear model (GLM) [28]. Obesity is the target value which the model predicts by using other columns like age, weight, height, gender, and BMI [31]. This dataset is used only for training the GLM model. Some models require some packages to predict the value of dataset according to target value but the GLM does not require any package for prediction.

The sample of the dataset for Partial Least Square (PLS) model is as in Table 5.7.

Table 5.7: Sample of dataset assigned to model PLS

Obesity	Age	Weight	Height	Gender	BMI
0.342	41	78	170.68	0	26.71
0.766	41	84	170.68	0	28.83
1.66	41	97	170.68	1	33.3
2.28	41	105	170.68	1	36.04
3.914	41	124	170.68	0	42.57
3.46	41	137	170.68	0	47.03
0	41	38	158.49	1	15.13

This dataset is assigned to PLS model, obesity is the target for this model. PLS requires one package pls, for using this model. Firstly, install this package then we can use this model for prediction of obesity. It takes more time as compared to GLM [28]. Every model takes an equal amount of data from dataset, however, they take data randomly so dataset should be large for hybrid technique otherwise some model predicts biased output which affects the result of the model. The tuning parameters of this model are Ncomp.

For hybrid technique, the third model is Random Forest (RF). The sample dataset of this model as shown in Table 5.8.

Table 5.8: Sample of dataset assigned to model RF

Obesity	Age	Weight	Height	Gender	BMI
0	41	38	158.49	1	15.13
0	41	45	158.49	0	17.91
0	41	53	158.49	1	21.1
0.134	41	67	158.49	1	26.67
0.892	41	74	158.49	1	29.46
0.766	41	84	170.68	0	28.83
0	41	38	170.68	1	13.04

This sample of dataset is assigned to RF model, obesity is the target value for every model. Random forest also needs some packages like randomForest and one tuning parameter, i.e., try. After installing this package we can use this model for prediction. The main overview of the working of the model is shown in Figure. 5.2.

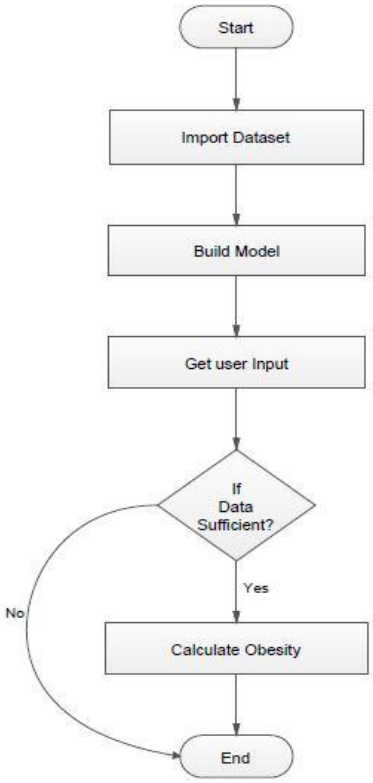


Figure. 5.2: Flow of proposed prediction process

All the models are of dual type. They can work with both regression and classification type of data. If we have a specific type of data, then we can use only these types of models that can work only at regression type of data.

5.2.1 Interface

Python is used for the interface of R. They work at some library files for different functions.

For CSV file:

```
Import csv
```

For opening CSV file:

```
Open ('file_name.csv', 'wb')

Open ('file_name.csv', 'rb')
```

w is used for writing mode

b is writing in binary mode

r is used for the reading mode.

For executing R script:

```
Import os

os.system ("Rscript File_name.R")
```

Every machine learning model needs packages and parameters. Some models have inbuilt packages and some models require to install those packages. Firstly, set path of the environment variable of your system to the bin folder of R. Then package will be installed via command prompt or R scripts.

The dataset is divided into two partitions, one for training data another for testing. If we use more than one model then training data is partitioned. After building the model, we check the accuracy of the model by using testing data. If the accuracy is high, then use this model otherwise execute the code again because the model is trained at random sample of data.

After building the model, create an interface of the model via Python. Python takes input from the user and stores it into a file with an extension of the file as CSV. R script executes the model and imports CSV file as input, the model returns the output via another CSV file. By using Python, we read this file. This is the final result of the model.

Every model has method argument value, type, packages, and tuning parameters. Argument value leads to calling the function. The type defines the type of models such as classification, regression, and dual use. If any model has dual use type, then it executes both classification and regression type of data. There are a number of packages needed to execute model. These packages and tuning parameters are shown in Table 5.9.

Table 5.9: Features of machine learning models

Model	**Argument**	**Type**	**Packages**	**Tuning Parameters**
Generalized Linear Model	Glm	dual use	None	None
Partial Least Squares	Kernelpls	dual use	Pls	Ncomp
Random Forest	Rf	dual use	randomForest	Mary

Frontend and backend both work in parallel. Firstly user enters inputs like age, weight, height, gender, and athlete. Attribute gender has sub-parts like male or female if user selects female then two more options occur, first for a pregnant lady and another for a normal lady. Athlete attribute works for both male and female.

After passing the input we store these values in CSV file because the model works with CSV file and model also returns the output in CSV file. Then we fetch the data from CSV file and return to the user. Obesity is predicted by machine learning model but BMI, BMR, RMR, Fat Percentage, Protein Recommended Dietary Allowance are calculated by formulae.

There are two ways to calculate BMI.

Metric: In this method, the body weight (kg) and body height (m × m) and the metric formula for calculating the BMI is shown in equation (1.3).

Imperial: In this method, the body mass (lb) and body height (in × in), the imperial formula for calculating the BMI is shown in equation (1.4).

Basal metabolic rate is denoted by BMR and depends upon age, weight, height, gender. It defines the rate of energy consumed by the human body. According to Harris- Benedict the equation [6] of basal metabolic rate is divided into two parts, one for male another for

female. Its metric formulae are given by equations (1.5) and (1.6), and imperial formulae are given by equations (1.7) and (1.8).

Resting Metabolic Rate is denoted by RMR [8] and depends upon the age, weight, height, gender. Resting metabolic rate is also known as Resting Energy Expenditure and it's denoted by REE [3]. According to the Mifflin-St Jeor, the equation [4] of resting metabolic rate is divided into two parts, one for male another for female, as given in equations (1.5) and (1.6)

Fat Percentage: It was researched by Deurenberg [15]. Body fat percentage is calculated by using BMI, age, and gender. The equation [5] of body fat percentage is divided into two categories, first equation (5.2) for child body fat percentage and the second equation (5.3) for an adult body fat percentage.

Fat percentage for child:

$$\text{Fat}\,\% = (\text{bmi} \times 1.5) - (\text{A} \times 0.70) - (\text{G} \times 3.6) + 1.4 \qquad (5.2)$$

Fat percentage for Adults:

$$\text{Fat}\,\% = (\text{bmi} \times 1.2) - (\text{A} \times 0.23) - (\text{G} \times 10.8) - 5.4 \qquad (5.3)$$

Where, BMI is the value of body mass index.

Unit of age is year.

If gender is male, then value of gender in 1.

If gender is female, then the value of gender is 0.

Protein Recommended Dietary Allowance (Protein RDA) is used to calculate daily need of protein in grams. It is divided into two categories, first for a normal person and another for athletes. It depends on the body weight and work done by the body. To calculate protein RDA for normal person, the equation (5.4) is

$$\text{Protein RDA} = \text{weight (kg)} \times 0.8 \qquad (5.4)$$

For calculating protein RDA for an athlete person, the equation depends on the work done by him and ranges from 1.4 to 1.8, according to athlete work done.

$$\text{Protein RDA} = \text{weight (kg)} \times \text{range} \quad (5.5)$$

The value of range = 1.4 to 1.8.

Range depends upon work done by the athlete.

Unit of Protein RDA is grams per day.

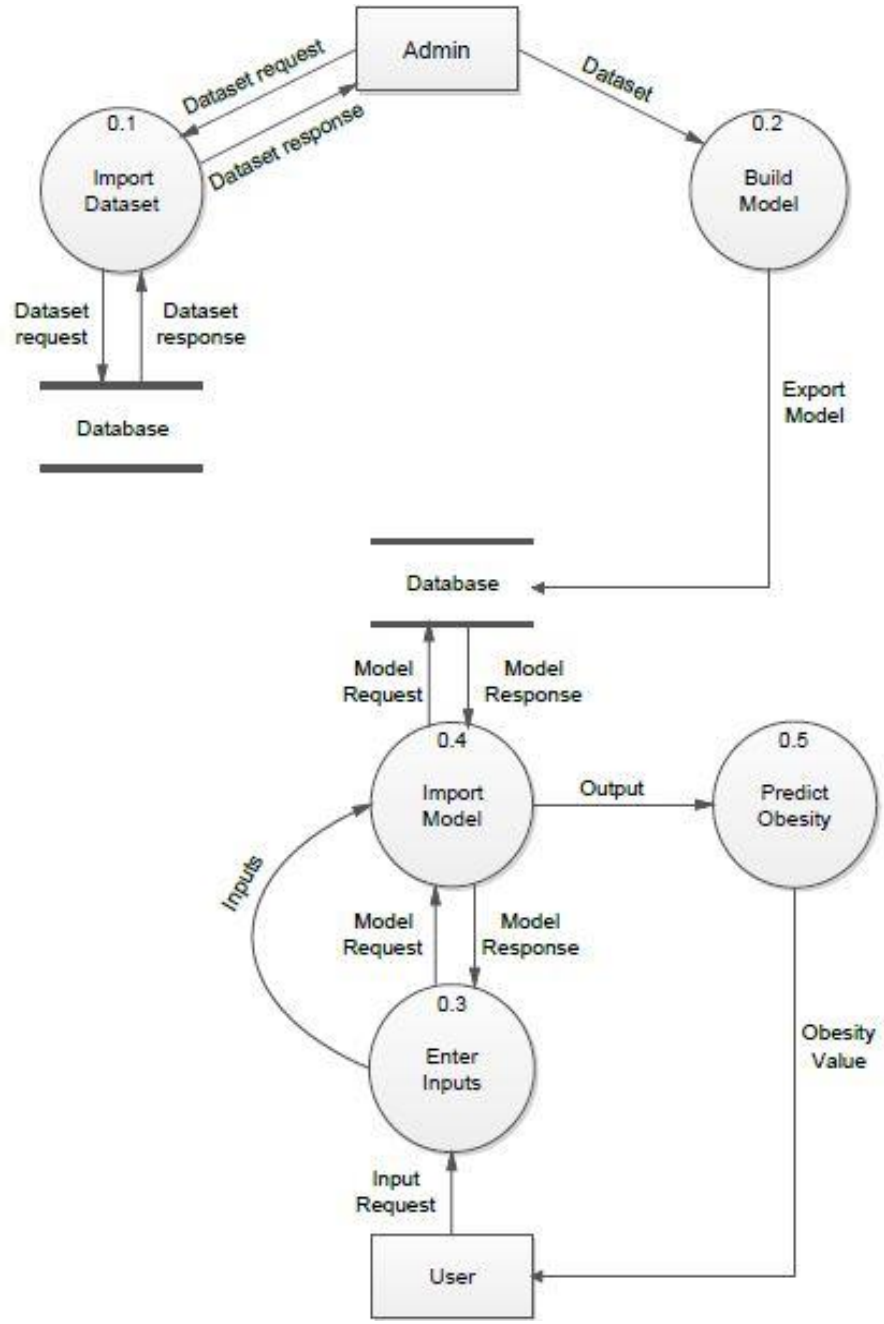


Figure 5.3: Detailed working of proposed model

Working of entire process is shown in Figure 5.3.

- 0.1 – 0.5 refers the control of the entire model.
- User input stores at CSV file.
- Model output store at CSV file.
- Python is used for interface of R

5.3 Result Analysis

For obesity prediction, we use machine learning techniques. The dataset is divided into four equal parts, the first part is assigned to the GLM, and the second part assigned to PLS, the third part assigned to the RF and the fourth part is testing data. Take the arithmetic mean of the output of every machine learning model, then test with fourth part of data. Execute this code 50 times for checking better accuracy. Choose the model which has best accuracy among all models[4]. Description of dataset is shown in Table 5.10.

Table 5.10: Description of dataset

Data Count (No. of rows)	**Model Name**
200 rows	Generalized Linear Model
200 rows	Partial Least Squares
200 rows	Random Forest
200 rows	Testing

Firstly, we input the dataset to RF model. After 10 iterations we note that the minimum value of accuracy is 79 and the maximum value of accuracy is 98 [5]. The iteration vs. accuracy graph in the RF model is shown in Figure 5.3.

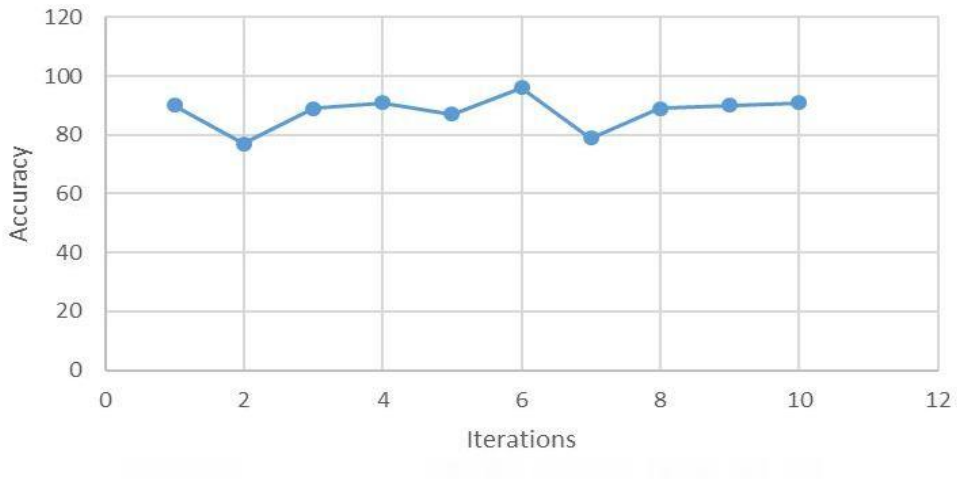


Figure 5.3: Accuracy result of RF

Fluctuation of accuracy is so high, then we apply PLS model on this dataset. After 10 iterations, we analyze the minimum value of accuracy is 81 and the maximum value of accuracy is 97. The iteration vs. accuracy graph of PLS model is shown in Figure 5.4.

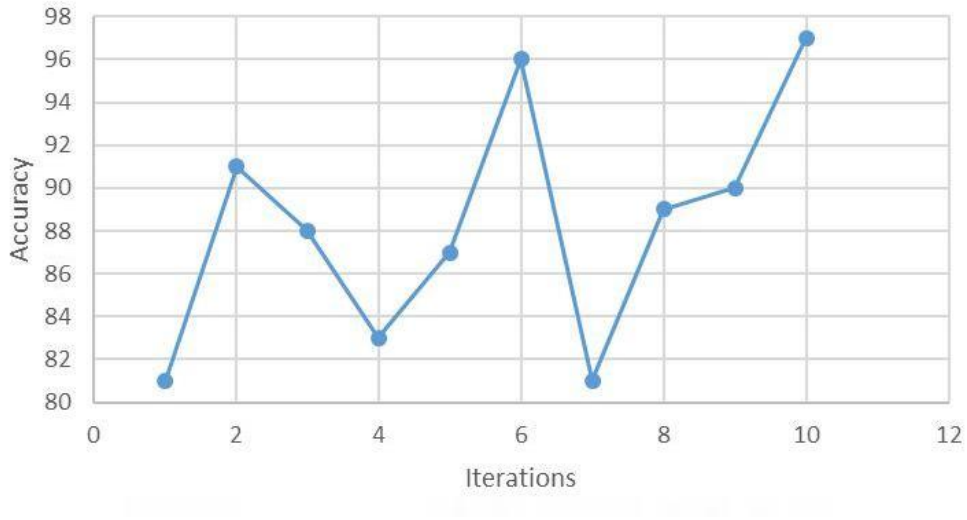


Figure 5.4: Accuracy result of PLS

Fluctuation of accuracy is so high, then we apply GLM on data set[28]. After 10 iterations we analyze the minimum value of accuracy is 78 and the maximum value of accuracy is 92. The iteration vs. accuracy graph of the GLM is shown in Figure 5.5.

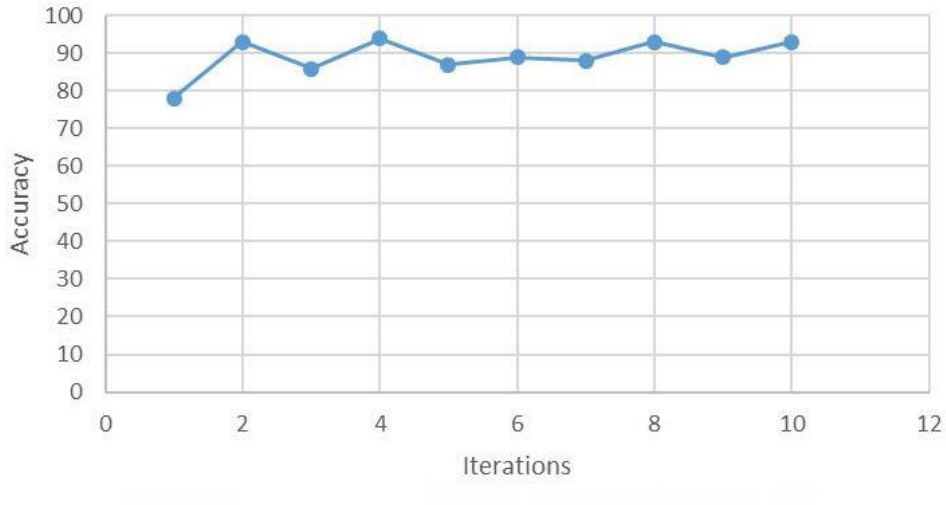


Figure 5.5: Accuracy result of GLM

Fluctuation of accuracy is still high, therefore, after using the RF, PLS and GLM, we apply ensemble model on the dataset by taking three models, these models are RF, PLS and GLM. After 10 iterations we analyze that the minimum value of accuracy is 96.5 and the maximum value of accuracy is 99. Fluctuation of accuracy is now lesser as compared to other models, so we finalize this ensemble model to predict obesity [14]. The iteration vs. accuracy graph of the ensemble model [14] is shown in Figure 5.6.

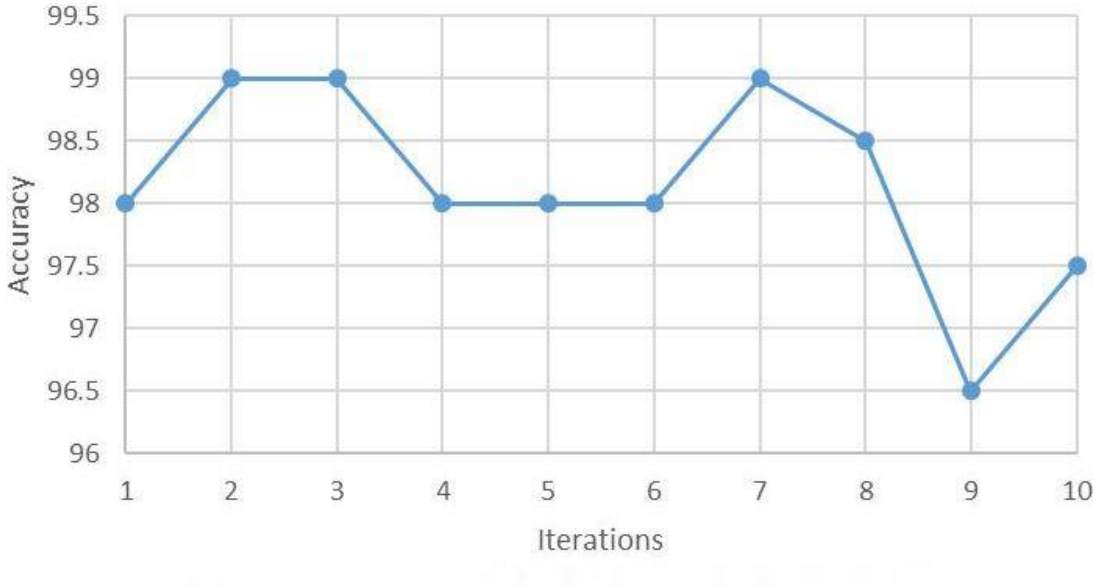


Figure 5.6: Accuracy of Ensemble model

Chapter 6

Concluding Remarks

6.1 Conclusion

To solve the problem of accuracy in the predicted value of obesity for specific people, we use machine learning technique. Every machine learning model needs packages and parameters, some models have inbuilt packages, while some models require the packages to be installed. The dataset is divided into two partitions, one for training data the other for testing. Training data is used to build a model, further, if we use more than one model then training data is further partitioned. After building the model, we check the accuracy of the model by using testing data. If the accuracy is high, then use this model otherwise execute the code again because model is trained on random sample of data. The average result of multiple runs is considered to retain robustness of the machine learning model.

The combination of two or more models is known as an ensemble model, whose primary requirement is that the dataset should be large, otherwise, if we take small dataset then we can not divide its training tuples any further. After cleaning the dataset, divide the dataset according to models and hold out one part for testing. The supervised learning models used in our ensemble model are - generalized linear model, random forest, and partial least square.

After building the model, we created an interface of the model via Python which takes input from user and stores into a file with an extension of the file being CSV. R script executes the model and imports CSV file as input, the model returns the output via another CSV file read via Python. This is the final result of the model.

Frontend and backend both work in parallel. Firstly, user enters inputs like age, weight, height, gender, and athlete. The attribute gender has sub-parts like male or female, if user selects female then two more options occur, first for a pregnant lady and another for a normal lady. Athlete attribute works for both male and female. After passing the input,

we store these values in CSV file, and then we fetch the data from CSV file and return it to the user.

Obesity is predicted by machine learning model but BMI, BMR, RMR, Fat Percentage, Protein RDA are calculated by formulae. If we check the obesity for pregnant lady or athlete person by ensemble model, and by the simple formula of BMI, then the values are different. Ensemble model is much more accurate as compared to the simple formula, for example - if we take the sample of two ladies with same height and weight but one lady is pregnant and another one is not, then simple formula shows the same output for both ladies, while the ensemble model analyzes the data first and then predicts the much accurate value of obesity, this is where we can see the difference between both techniques.

6.2 Threats to validity

Few limitations of our current work can be listed as follows:

- Our ensemble machine learning model requires more memory and takes so much time for prediction the result, at the cost of being accurate.
- It is difficult to clean the data and analyze the accuracy of every model because some models work at regression type of data, some models work at classification type of data and some models have type as dual, which deals with both regression and classification.
- If we work with dynamic data or live data, then it is much difficult to analyze the model for this dataset. If live data have some changes then model gives an error because it works with the specific data type.
- Some general weaknesses of machine learning techniques are:
    - Each application needs to be specially trained and requires huge amounts of hand-crafted data and structured training data.
    - Training data or dataset must be tagged.
    - Never learn incrementally, in real time.
    - Never learn interactively, in real time.

- Poor transfer learning ability.
- Poor reusability of modules and poor integration.
- Debugging is much difficult.
- Performance is not consistent for a long time.

6.3 Future Research Directions

By using the ensemble technique, we improve the accuracy of predicting obesity for pregnant lady and athlete persons but the increase of weight during pregnancy is not same for every lady and increase of bone mass of every athlete is not same. For improving the accuracy, we need some more facts and some more data. Then we can add some more models and get a more accurate result.

For physically disabled people, every person doesn't have the same weight of body parts like if we analyze the data of 10-20 people, and then the weight of arms or legs of every person is different. For analyzing these differences, we need some more time. This could be a possible extension to the current work. Another direction of work in the future would be to reduce the training as well as running time of models, and further take fewer inputs. Additionally, we may also improve the front end look, because we use Python for front-end look and it is executed in Command Prompt, we will try for front-end in PHP then we will improve the Graphical User Interface which looks very attractive.

We will try some platform and/or machine independent new techniques to fetch the data of obesity for every person because the coding of machine learning by using R requires some more RAM and memory of the computer and if we execute on another computer then all packages need to be installed again. If any other techniques require less memory and RAM, and also predict the accurate result of obesity then we will use those techniques.

References

[1] Jindal, K., Baliyan, N., and Rana, P.S., "Obesity Prediction using Ensemble Machine Learning Approaches," presented at Springer 5th International Conference on Advanced Computing, Networking, and Informatics, NIT Goa, to be published as book chapter in Advances in Intelligent Systems and Computing, 2017

[2] P. Deurenberg, J. A. Weststrate, and J. C. Seidell, "Body mass index as a measure of body fatness: age- and sex-specific prediction formulas," British Journal of Nutrition, vol. 65, no. 02, p. 105, 1991.

[3] "The Free Encyclopedia," Wikipedia. [Online]. Available: http://www.wikipedia.com/. [Accessed: 21-Jan-2017].

[4] D. Parkin, "BMI - Body Mass Index Tool | Calculate Imperial and Metric | Weight & Height Chart," BMI - Body Mass Index Tool | Calculate Imperial and Metric | Weight & Height Chart. [Online]. Available: http://www.mehndiskinart.com/bmi_calculator.htm. [Accessed: 17-Mar-2017].

[5] C. R. White and R. S. Seymour, "Mammalian basal metabolic rate is proportional to body mass2/3," Proceedings of the National Academy of Sciences, vol. 100, no. 7, pp. 4046–4049, 2003.

[6] "BMR Calculator For Weight Loss," The Calculator Site. [Online]. Available: http://www.thecalculatorsite.com/health/bmr-calculator.php. [Accessed: 7-Dec-2016].

[7] A. E. Black, "Critical evaluation of energy intake using the Goldberg cut-off for energy intake: basal metabolic rate. A practical guide to its calculation, use and limitations," International Journal of Obesity, vol. 24, no. 9, pp. 1119–1130, Jan. 2000.

[8] J. Daley, "Resting Metabolic Rate (RMR) Calculator," Resting Metabolic Rate (RMR) Calculator | SHAPESENSE.COM. [Online]. Available: http://www.shapesense.com/fitness-exercise/calculators/resting-metabolic-rate-rmr-calculator.shtml. [Accessed: 19-Feb-2017].

[9] E. Alpaydin, Introduction to machine learning. Cambridge: The MIT Press, 2014.

[10] L. Ling, Encyclopedia of database systems. New York: Springer, 2009.

[11] I. Kononenko and Kukar Matjaž, Machine learning and data mining introduction to principles and algorithms. Oxford: Woodhead Publ., 2013.

[12] G.-Y. Shi and S. Liu, "Model selection of RBF kernel for C-SVM based on genetic algorithm and multithreading," 2012 International Conference on Machine Learning and

Cybernetics, 2012.

[13] R. Gentleman, R programming for bioinformatics. Boca Raton FL: CRC Press, 2009.

[14] S. Banik, A. F. M. K. Khan, and M. Anwer, "Hybrid Machine Learning Technique for Forecasting Dhaka Stock Market Timing Decisions," Computational Intelligence and Neuroscience, vol. 2014, pp. 1–6, 2014.

[15] D. Halls, "Body fat percentage formula from body mass index," Moose and Doc, 26-Jun-2017. [Online]. Available: http://halls.md/body-fat-percentage-formula/. [Accessed: 24-Apr-2017].

[16] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," 2014 IEEE International Conference on Image Processing (ICIP), 2014.

[17] A. S. Singh, Marijke J. M. Chin A Paw, J. Brug, and W. V. Mechelen, "Dutch Obesity Intervention in Teenagers," Archives of Pediatrics & Adolescent Medicine, vol. 163, no. 4, p. 309, Jun. 2009.

[18] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," Statistics Surveys, vol. 4, pp. 40–79, 2010.

[19] T. D. Müller, A. Hinney, A. Scherag, T. T. Nguyen, F. Schreiner, H. Schäfer, J. Hebebrand, C. L. Roth, and T. Reinehr, "Fat mass and obesity associated gene (FTO): No significant association of variant rs9939609 with weight loss in a lifestyle intervention and lipid metabolism markers in German obese children and adolescents," BMC Medical Genetics, vol. 9, no. 1, 2008.

[20] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, p. 1, Jan. 2004.

[21] E. Mcgarvey, A. Keller, M. Forrester, E. Williams, D. Seward, and D. E. Suttle, "Feasibility and Benefits of a Parent-Focused Preschool Child Obesity Intervention," American Journal of Public Health, vol. 94, no. 9, pp. 1490–1495, 2004.

[22] N. Cristianini, "Cross-Validation (K-Fold Cross-Validation, Leave-One-Out, Jackknife, Bootstrap)," Dictionary of Bioinformatics and Computational Biology, 2004.

[23] D. Frankenfield, L. Roth-Yousey, and C. Compher, "Comparison of Predictive Equations for Resting Metabolic Rate in Healthy Nonobese and Obese Adults: A Systematic Review," Journal of the American Dietetic Association, vol. 105, no. 5, pp. 775–789, 2005.

[24] M. Saiedullah, M. Sha, M. Siddique, Z. Tamannaa, and Z. Hassan, "Healthy Bangladeshi individuals having lower high-density lipoprotein cholesterol level compared to age-, gender-, and body mass index-matched Japanese individuals: A pilot study," Journal of

Molecular Pathophysiology, vol. 6, no. 1, p. 1, 2017.

[25] K. M. Flegal, M. D. Carroll, B. K. Kit, and C. L. Ogden, “Prevalence of Obesity and Trends in the Distribution of Body Mass Index Among US Adults, 1999-2010,” Jama, vol. 307, no. 5, p. 491, Jan. 2012.

[26] T. Reinehr, M. Temmesfeld, M. Kersting, G. D. Sousa, and A. M. Toschke, “Four-year follow-up of children and adolescents participating in an obesity intervention program,” International Journal of Obesity, vol. 31, no. 7, pp. 1074–1077, Jan. 2007.

[27] Y. M. Henry, D. Fatayerji, and R. Eastell, “Attainment of peak bone mass at the lumbar spine, femoral neck and radius in men and women: relative contributions of bone size and volumetric bone mineral density,” Osteoporosis International, vol. 15, no. 4, pp. 263–273, Jan. 2004.

[28] “Generalized Linear Models (GLMs),” Generalized, Linear, and Mixed Models Wiley Series in Probability and Statistics, pp. 135–155, 2005.

[29] D. Frankenfield, L. Roth-Yousey, and C. Compher, “Comparison of Predictive Equations for Resting Metabolic Rate in Healthy Nonobese and Obese Adults: A Systematic Review,” Journal of the American Dietetic Association, vol. 105, no. 5, pp. 775–789, 2005.

[30] D. T. Larose, Discovering knowledge in data: an introduction to data mining. Hoboken, NJ: Wiley, 2014.

[31] “Body Mass Index,” Encyclopedia of Obesity.

List of Publications

[1] Jindal, K., Baliyan, N., and Rana, P.S., "Obesity Prediction using Ensemble Machine Learning Approaches," presented at Springer 5th International Conference on Advanced Computing, Networking, and Informatics, NIT Goa, to be published as book chapter in Advances in Intelligent Systems and Computing, 2017

[2] Jindal, K. and Baliyan, N., "An Ensemble Machine Learning Based System for Prediction of Obesity" under preparation for communication to Journal of Computer Information Systems, Taylor & Francis, 2017

Appendix A

List of Models

Table A1: List of dual type models

Model Name	Method	Type
Boosted Linear Model	BstLm	Dual
Boosted Tree	blackboost	Dual
Gaussian Process	gaussprLinear	Dual
Penalized Ordinal Regression	ordinalNet	Dual
Regularized Random Forest	RRF	Dual
Self-Organizing Map	Bdk	Dual
Neural Network	Nnt	Dual

Video URL

A Video has been uploaded on YouTube to describe the working of the system named “Ensemble R code with Python interface”. The URL of the video is as follows:

https://www.youtube.com/watch?v=KBnUul4Upgg&t=35s

Thesis_17_07_17

ORIGINALITY REPORT

%3	%2	%0	%1
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.gdeepak.com Internet Source	%2
2	www.webmd.com Internet Source	<%1
3	Submitted to Manipal University Student Paper	<%1
4	www.tex.ac.uk Internet Source	<%1
5	www.ctibooks.com.cn Internet Source	<%1
6	Submitted to Leeds Beckett University Student Paper	<%1
7	Submitted to Miami Palmetto Senior High School Student Paper	<%1
8	Submitted to Harrisburg University of Science and Technology Student Paper	<%1