VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belgaum-590018

A PROJECT REPORT

ON

"OBESITY PREDICTION USING MACHINE LEARNING ALGORITHM"

Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

PAVAN KUMAR N	1KI19CS059
SHREYAS B R	1KI19CS100
VISHVESHWARA M N	1KI19CS119
VISMAYA K S	1KI19CS121

Under the Guidance of

Mr. Harish B M B.E., MTech.

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

KALPATARU INSTITUTE OF TECHNOLOGY

N H -206 , TIPTUR-572201

2022-2023

KALPATARU INSTITUTE OF TECHNOLOGY

N H -206, TIPTUR -572201

Department of Computer Science & Engineering

CERTIFICATE

Certified that the project work entitled **"OBESITY PREDICTION USING MACHINE LEARNING ALGORITHM"** is a bona fide work carried out by

PAVAN KUMAR N	**1KI19CS059**
SHREYAS B R	**1KI19CS100**
VISHVESHWARA M N	**1KI19CS119**
VISMAYA K S	**1KI19CS121**

in partial fulfillment for the award of Bachelor **of Engineering** in **Computer Science & Engineering** of the **Visvesvaraya Technological University, Belgaum** during the year 2022-2023. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Mr. Harish B M
Guide

Mr. Shashidara M S
Head of Department

Dr. G D GuruMurthy
Principal

External Viva

Name of Examiners — Signature with date

1.

2.

KALPATARU INSTITUTE OF TECHNOLOGY

N H -206, TIPTUR-572201

Department of Computer Science & Engineering

DECLARATION

We, the students of final semester of Computer Science & Engineering, Kalpataru Institute of Technology, N H -206 Tiptur -572201 , declare that the work entitled **"OBESITY PREDICTION USING MACHINE LEARNING ALGORITHM**" has been successfully completed under the guidance of Mr. Shashidara M S, Department of Computer Science & Engineering. This dissertation work is submitted to Visvesvaraya Technological University in partial fulfilment of the requirements for the award of Degree of Bachelor of Engineering in *Computer Science & Engineering* during the academic year 2022-2023. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to any university.

Place: Tiptur

Date:

Project Associates:

1. **PAVAN KUMAR N**
2. **SHREYAS B R**
3. **VISHVESHWARA M N**
4. **VISMAYA K S**

ACKNOWLEDGEMENT

At the outset I express my sincere thanks to the holy sanctum **"Kalpataru Institute of Technology"** the temple of learning, for giving us an opportunity to pursue the degree course in **Computer Science and Engineering** thus help shaping my career.

I consider it is privilege and honor to express my deep sincere gratitude to my Guide **Mr. Harish B M.** Assistant Professor, Department of CSE for his continuous support and guidance throughout the course of my technical seminar and also during the period of my stay in KIT.

I wish to extend my gratitude to my beloved HOD **Prof. Shashidhara M.S.,** Associate Professor and HOD, Department of CSE for his continuous encouragement during course my studies.

I wish to extend my gratitude to my beloved Principal **Dr.G.D. Gurumurthy** for his continuous encouragement during the course of my studies.

I would like to thank all teaching and non-teaching staff of the department of Computer Science and Engineering. I am grateful to all their cooperation and their guidance for completing my task well in time. I thank one and all who have helped me in one or the other way.

PAVAN KUMAR N (1KI19CS0059)
SHREYAS B R (1KI19CS100)
VISHVESHWARA M N (1K19CS119)
VISMAYA K S (1K19CS121)

ABSTRACT

Behavioral risk factors such as unhealthy habits, improper diet, and physical inactivity lead to physiological risks, and "obesity/overweight" is one of the consequences. "Obesity and overweight" are one of the major lifestyle diseases that leads to other health conditions, such as cardiovascular diseases (CVDs), chronic obstructive pulmonary disease (COPD), cancer, diabetes type II, hypertension, and depression. It is not restricted within the age and socio-economic background of human beings. The "World Health Organization" (WHO) has anticipated that 30% of global death will be caused by lifestyle diseases by 2030 and it can be prevented with the appropriate identification of associated risk factors and behavioral intervention plans. Health behavior change should be given priority to avoid life-threatening damages. The primary purpose of this study is not to present a risk prediction model but to provide a review of various machine learning (ML) methods and their execution using available sample health data in a public repository related to lifestyle diseases, such as obesity, CVDs, and diabetes type II. In this study, we targeted people, both male and female, in the age group of >20 and <60, excluding pregnancy and genetic factors. This paper qualifies as a tutorial article on how to use different ML methods to identify potential risk factors of obesity/overweight. Although institutions such as "Center for Disease Control and Prevention (CDC)" and "National Institute for Clinical Excellence (NICE)" guidelines work to understand the cause and consequences of overweight/obesity, we aimed to utilize the potential of data science to assess the correlated risk factors of obesity/overweight after analyzing the existing datasets available in "Kaggle" and "The University of California, Irvine (UCI) database", and to check how the potential risk factors are changing with the change in body-energy imbalance with data-visualization techniques and regression analysis. Analyzing existing obesity/overweight-related data using machine learning algorithms did not produce any brand-new risk factors, but it helped us to understand: (a) how are identified risk factors related to weight change and how do we visualize it? (b) what will be the nature of the data (potential monitorable risk factors) to be collected over time to develop our intended eCoach system for the promotion of a healthy lifestyle targeting "obesity and overweight" as a study case in the future? (c) why have we used the existing "Kaggle" and "UCI" datasets for our preliminary study? (d) which classification and regression models are performing better with a corresponding limited volume of the dataset following performance metrics?

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTERODUCTION

1.1 Introduction to Obesity

More than one-third of the adult population in the United States is obese and this is linked to certain factors, such as physical inactivity, improper diet, family history, and the environment [1]. As reported by "The GBD 2015 Obesity Collaborators" in 2015, a total of 107.7 million children and 603.7 million adults were obese [2]. After analyzing data from 68.5 million people from 195 countries between 1990 and 2015, the research team concluded that the burden of "obesity and overweight" is related to high body-mass index (BMI), age, and gender. With the number of obese people doubling in two decades (from 1.3 million people obese globally in 1980 to double in 2008), unhealthy habits (such as consumption of tobacco and alcoholic beverages), unhealthy diet (such as energy drinks, consumption of excess salt and sugar, intake of high saturated fat, and discretionary foods), and physical inactivity are the major pillars of "obesity and overweight". In 2016, more than 1.9 billion adults (39%) aged eighteen years and older were overweight, and of these, over 650 million (13%) were obese. In 2016, more than 340 million children and teenagers aged five to nineteen were overweight or obese, and in 2018, 40 million children under the age of five were overweight or obese. The universal predominance of "obesity and overweight" nearly tripled between 1975 and 2016. Juvenile obesity is linked to a higher chance of obesity, untimely death, and infirmity in adulthood [3–6]. The chronic conditions associated with "obesity and overweight" are considered as health care and social burdens. According to the latest study conducted by the "National Health and Nutrition Examination Survey (NHANES, 2007–2012)" on aggregated data (2007–2008, 2009–2010, and 2011–2012) collected from 15,208 adults with an age >=25, excluding pregnancy (n = 125) and incorrect noisy data (n = 827), a significant correlation was observed between sex, age, race, or ethnicity with "obesity and overweight" [1]. Potential risk factors related to obesity/overweight may vary in children under age five, adolescents, adults, older people, and pregnant women.

The consequences of "Obesity and overweight", which continues to be the foremost public health anxiety, increase the risk of the other four primary lifestyle diseases, such as cardiovascular diseases (CVD), cancers, diabetes (type II), and chronic lung diseases (chronic obstructive pulmonary disease (COPD), asthma). The burden of these diseases is extremely high among lower-income countries and populations. A total of 63% (36 million) of global death occurred

in 2008 due to lifestyle diseases or non-communicable diseases (NCDs). Additionally, 80% of the 36 million dead people belonged to low- and middle-income classes, 13% were from high-income classes, and 29% of the total NCD deaths occurred below the age of sixty years. An increase of 10 million deaths annually on average due to NCDs has been observed from selected literature study. In 2016, the number increased to
56.9 million (71%), and by 2030, it is predicted to achieve 75%, with 88.5% death in developed countries and 65% death in developing countries. The risk for the stated lifestyle diseases increases with "body mass index" (BMI) in direct proportion. BMI is a number calculated by "Weight/Height^2" and is used to assess body composition [3–7]. However, BMI is rather a bad indicator of the percent of body fat, as BMI does not capture information on the mass of fat in different body sites and is highly dependent on age [8]. In 2012, as identified by the Institute of Medicine, population-based obesity prevention strategies, such as physical activity, healthy diet, models of healthy social rules, and context-based and tailored recommendations by setting have the potential to combat "obesity and overweight" [9]. Thus, health behavior change should be given precedence to circumvent severe damages.

An electronic coaching ("eCoaching") system can empower people to manage a healthy lifestyle with early risk predictions and appropriate individualized recommendations. To develop an intelligent eCoach system for automated, personalized, contextual, and behavioral recommendations to achieve personal wellness goals, addressing obesity as a study case, we propose to (a) identify associated health risk factors, (b) perform data collection from identified controlled trials, (c) analyze the data, and (d) perform a predictive analysis with machine learning algorithms for future health risk predictions and behavioral interventions [10,11].

In this tutorial of ML models to identify the risk factors of overweight and obesity, we reviewed the performance of different machine learning algorithms (regression and classification) on existing datasets available in "Kaggle" and "UCI" so that we could create a list of risk factors associated with obesity/overweight with an appropriate quantitative analysis. The obtained result at the end of the study helped us to decide which risk factors health and wellness data would be collected on for our future research work— "eCoach behavioral interventions for obesity and overweight".

1.1.2 What is Obesity or Overweight?

Overweight and obesity are defined as abnormal or excessive fat accumulation that presents a risk to health. A body mass index (BMI) over 25 is considered overweight, and over 30 is obese. The issue has grown to epidemic proportions, with over 4 million people dying each year as a result of being overweight or obese in 2017 according to the global burden of disease.

Rates of overweight and obesity continue to grow in adults and children. From 1975 to 2016, the prevalence of overweight or obese children and adolescents aged 5–19 years increased more than four-fold from 4% to 18% globally.

Obesity is one side of the double burden of malnutrition, and today more people are obese than underweight in every region except sub-Saharan Africa and Asia. Once considered a problem only in high-income countries, overweight and obesity are now dramatically on the rise in low- and middle-income countries, particularly in urban settings. The vast majority of overweight or obese children live in developing countries, where the rate of increase has been more than 30% higher than that of developed countries

1.1.1.1. Risk factors

Risk factors for developing obesity include:

- **Lack of physical activity:** Lack of physical activity, combined with high amounts of TV, computer, video game, or other screen time has been associated with a high body mass index.
- **Unhealthy eating behavior:** Some unhealthy eating behaviors can increase your risk for overweight and obesity.
  - **Eating more calories than you use:** The number of calories you need will vary based on your sex, age, and physical activity level. Find daily calorie needs or goals for adults as part the dash eat planing. You can also find tipt sheets of parents for guidance on how many calories children need and ways to reduce screen time.
  - **Eating too much saturated fat:** According to the Dietary Guidelines for the American sexternal Link, the amount of saturated fat in your daily diet should be no more than 10% of your total calories. For a 2,000-calorie diet, that's about 200 calories or about 22 grams of saturated fat.
  - **Eating foods high in added sugar:** On a daily basis, try to limit the amount of added sugar in your diet to no more than 10% of your calories.
- **Not getting enough good-quality of sleep:** Research has shown a link between poor sleep — not getting enough sleep or not getting enough good-quality sleep — and a high BMI. Regularly getting less than 7 hours of sleep per night can affect the hormones that control hunger urges. In other words, not getting good-quality sleep can make us more likely to overeat or not recognize our body's signals that we are full.
- **High amount of stress :** Long-term and even short-term stress can affect the brain and trigger

your body to make hormones, such as cortisol, that control energy balances and hunger urges. These hormone changes can make you eat more and store more fat.

- **Health condition:** Some conditions, such as metabolic syndrome and polycystic ovary syndrome, cause people to gain weight. These medical conditions must be treated for a person's weight to come close to or into normal range.

- **Genetics:** Some people are predisposed to being heavier. Researchers have found at least 15 genes that influence obesity. Studies show that genetics may play a more important role in people with obesity than in people who are overweight. For people with a genetic high risk for obesity, making healthy lifestyle changes can help lower that risk.

- **Medicines:** Some medicines cause weight gain by disrupting the chemical signals that tell your brain you are hungry. These include:

  - Antidepressants
  - Antipsychotics
  - Beta-blockers, which are used to treat high blood pressure
  - Birth control
  - Glucocorticoids, which are often used to treat autoimmune disease
  - Insulin, which is a hormone taken to control blood sugar levels in people with diabetes

  Talk to your provider if you notice weight gain while you are using one of these medicines. Ask whether there are other forms of the same medicine or other medicines that can treat your medical condition but have less of an effect on your weight.

- **Your environment:** Your environment can contribute to unhealthy eating and a lack of physical activity. Your environment includes all of the parts where you live and work — your home, buildings in which you work or shop, streets, and open spaces. The types of restaurants and the amount of green space you have can contribute to overweight and obesity.

  Studies have shown that access to sidewalks and green spaces can help people be more physically active, and grocery stores and farmers markets can help people eat healthier. On the other hand, people living in neighborhoods with more fast food restaurants and inaccessible or no sidewalks or bath paths are more likely to be overweight or obese.

1.1.2. Features

Some of the attributes we used for Obesity Prediction and their correlation to BMI (Body mass index). This dataset consists of 12 features and a target variable. The detailed description of all the features are as follows:

- **Age:** Patients Age in years. (Numeric)
- **Sex:** Gender of patient. (Male - 1, Female - 0) (Nominal)
- **Height:** Persons Height. (Numeric)
- **Weight:** Person's weight. (Numerical)
- **Family_history_with_overweight:** Those with a family history of obesity had a higher BMI and were at increased risk of obesity. (Yes-0, No-1)
- **FAVC (frequent consumption of high caloric food):** The attribute related with eating habits. (Yes-1, No-0).
- **FCVC (Frequent consumption of vegetable):** The attribute related with eating habits. (Numeric).
- **NCP (Number of main meals):** The attribute related with eating habits.(Numeric).
- **CAEC (Consumption of food between meals):** The attribute related with eating habits.(No-0, Sometimes-1, Frequently-2, Always-3)
- **Smoke:** Persons Habit. (Yes-0, No-1)
- **CH_2O (Consumption of water daily):** Persons Habit (Numeric).
- **SCC (Squamous cell carcinoma):** It refers to High consumption of fat. (Yes-0, No-1)

Target Variable:

- **Target:** It is the target variable that we have to predict,(Insufficient Weight, Normal_Weight,Overweight_Level_I,Overweight_Level_II,Obesity_Type_I,Obesity_Type_II ,Obesity_Type_III)

1.2. PROBLEM STATEMENT

Obesity is quickly becoming one of the most prominent conditions affecting children and adolescents. Traditionally, treatment for obesity has been approached from a medical model. More recent research has found that in addition to changes in physical activity and caloric intake (medical health); emotional, social, and psychological (mental health) factors must be addressed in order to provide effective treatment to overweight and obese children and adolescents. In order to promote this holistic approach to obesity treatment, it is necessary to examine the existing evidence and compile it in a format that is easily accessed as well as informative for use of those in the medical health care professions.

1.2.1. Existing System

1.2.1.1 Description of existing system

In obesity researches, data mining techniques have been widely engaged to investigate the risk factors. Different advance machine learning algorithms have been used to that effect. Algorithms of K-nearest neighbor, Logistic Regression, Random Forest, Support Vector Machine, Multilayer Perceptron, Adaptive Boosting, Naïve Bayes, Gradient Boosting Classifier, and Decision Tree were used in designing models for prediction. [6 – 17].

1.2.1.2 Review of existing system:

In the early years, traditional and clinical procedures are being use in predicting obesity and this can be time-consuming, as it demands the employment of trained physicians in the processes needed to obtain diagnosed result. Most times, complications are due to late intervention as manual methods are mostly carried out when symptoms has manifested in patients. In recently, various researches have been carried out in the area of obesity prediction using machine learning techniques. These techniques being employ by different researchers produced varying results of different accuracy depending on its challenges as its related to the individual.

To diagnosed obesity, physical exams are performed on the patient, then tests are recommended by experts, these test and examinations generally include; taking the patient health history, generally examining the patient physically (checking height, blood pressure, heart rate and temperature), calculating the body mass index (BMI), measuring the waist circumference, lifestyle, hereditary background and checking for other problems. These processes require a high level expertise and lots of time.

1.2.1.3 Problem with the existing system:

The following are the problems facing the existing system

i. Inaccurate obesity diagnosis as patients are not diagnoses by an expert.

ii. Only one feature is used which is the body measure thus prediction might not be correct.

iii. Prone to low accuracy

1.2.2. Literature Review

Tries to examine the relationship between the weight status and the physical activities in human and also compare some machine learning and classical statistical models used in predicting obesity level. National Health and Nutrition Examination Survey Dataset was used in their model, and they made use eleven different algorithms which include the random sub space, logistic regression, decision table, Naïve Bayes, the Radial Basis Function, K-nearest neighbor, classification via regression, J48 and Multilayer perception are set of algorithms used for their implementation and evaluation. The evaluation metric used was the ROC and AUC and the algorithm that achieved the overall highest accuracy was the random subspace classifier algorithm.

[7] Use machine learning methods such as linear and logistic repressor, Artificial Neural Network, Deep Learning and Decision Tree analysis to predict and/or classify obesity level from large dataset gotten from sensors, smartphones, and electronic medical health records. They concluded that machine learning will provide a sophisticated tools to predict, classify and describe obesity related risks and its outcome.

[8] Tries to predict the percentage of obesity level in adult in the United States in the nearest future by using fitting multinomial regression to estimate the prevalence of 4 BMI categories (less than 25 is normal weight, 25 to less than 30 is overweight, 30 to less than 35 is moderate obesity and above 35 is severe obesity) on a self-reporting bias dataset from the behavioral risk factor surveillance system survey of 1993-1994 and 1999- 2016. They concluded from their analysis that obesity level in adult will always be on the increase in 2030 with 48.9% obese,24.2% severely obese and the occurrence will be higher than 25% in 25 states.

[9] Evaluated the BMI increase pattern in kids and also develop a system that predicts those children who have high risk of being obese before it gets to critical stage. From their evaluation, they reviled that the greatest increase in BMI happens between the age of 2 & 4 and their accurate prediction happens at 5 to 6 years of age.

[10] design a model that predict early childhood obesity using XGBoost, ID3 decision tree model and the Recurrent Neural Network(RNN) machine learning algorithm to model an electronic

health record data. From their area under curve the XGBoost has the best AUC-value of 0.81 (0.001) and it outperformed all other models.

[11], created a logistic regression model to evaluate the probability of BMI on children between the age of 2 and 17 in rural areas using publicly available dataset. The outcome shows that in small geographic areas, estimates are important to create effective involvements and to help planning of potential solutions to the problem as prevalence among census varies from 27 to 40%.

A model was presented by [12], which uses fuzzy signature to understand and manage intricacies on children obesity dataset and a solution that could handle the related risk with children motor development and early obesity.

[13] Had an initial method to predicting obesity which was developed with the information collected from primary sources: Parents, caretakers and the children themselves. The paper authors tries to identify the risk factors like the obesity, education level of the parents, habits, lifestyle of the children and the environment influence on children. The proposed framework uses a hybrid approach of the decision trees and the Naïve Bayes known as the NBTree.

[14] Predict children obesity using data mining. The aim of the suggested survey was to provide the required information on obesity problem. The NN, the Naïve Bayes and the Decision Trees models were used for their implementation.

[15] These authors paper showed the classification of obesity in grade 6 children, from two separate districts in Malaysia. A classification technique was used to model the information collected. The machine learning classification models used are Decision Trees, Support Vector Machines, Neural Networks and Bayesian Networks. An article published by.

[16] aims to predict the obesity in children over the age of 2 by using data collected before the second birthday of the children, they analyzed six different machine learning methods, the machine learning methods they used are the ID3, random tree, random forest, J48, Naïve Bayes and Bayes train on CHIKA data. The overall accuracy of 85% was gotten from the random forest .

[17], used ensemble machine learning approach to predict obesity in human using attribute like the age and BMI (weight and height) of individuals. They use machine learning algorithms such as the linear model, the random forest and partial least square. They obtained an accuracy of 89.68% from random forest.

[8] carried out a research on the prediction of the percentage of obesity level in adult in the nearest future using Fitting Multinomial Regression to estimate the prevalence of four (4) BMI categories (less than 25 is normal weight, 25 to less than 30 is overweight, 30 to less than 35 is moderate obesity and above 35 is severe obesity) on a self-reported bias dataset. The drawback of this research is that only one technique was used to generalize the multiclass problems and the dataset is self-report biased with Body measure index (Height and weight). Only Body Mass Index (weight and height) is

not a perfect measure to determine an accurate obese person as it can sometimes give " gives a false positive" to an athletic people whose high BMIs are not due to excess body fat but to excess muscle.

Hence, it is essential to use more features to build advanced prediction model using machine learning approaches rather than using simpler methods like the manual or statistics approaches. This research will focus on improving an already existing model to obtain a more efficient and less human interference in predicting obesity status using a native dataset with standard components.

Base on this point, this paper considered using a machine learning techniques to build a prediction model of obesity in individuals using a locally available dataset from a health care diagnostic center with principle features. These form the research gap.

1.2.3. Proposed Solution

The proposed model will be build using the Python programming language, applying machine learning algorithm on the dataset collected from a healthcare center to model a system that will be capable of predicting obesity in patients with dataset that has been technically automated to suit the purpose of this research.

The advancement of machine learning which a branch of artificial intelligence is has led to different researches in different domains to help with lots of process that has been carried out manually. This model when built will help medical personnel in the prediction of obesity in patients with less time involvement. It will also help to handle the prediction of the class or level of obesity which will aid decision for early intervention. The system will have the following advantages.

i. Increase the reliability of result diagnosed
ii. Reduce the trouble of diagnosis of obesity to the barest minimal
iii. Provide real time system for classification of obesity
iv. Saves time and cut cost required for various medical test

CHAPTER 2

REQUIREMENT ANALYSIS, TOOLS &TECHNOLOGIES

2.1. Hardware & Software Requirements

2.1.1. Hardware Requirements

- **Processors:** Intel® Core™i3 or i5 processor, 4/8 GB of RAM.
- **Operating systems:** Windows* 7 or later, macOS, and Linux.

2.1.2. Software Requirements

- **Python* versions:** 2.7.X, 3.6.X, 3.9.X
- **Included development tools:** conda*, conda-env, Jupyter Notebook* (IPython),Google Colaboratory.
- **Compatible tools:** Microsoft Visual Studio*, PyCharm*.
- **Included Python packages:** NumPy, SciPy, scikit-learn*, pandas, Matplotlib, Seaborn, Numba*, Intel® Threading Building Blocks, pyDAAL, Jupyter, mpi4py, PIP*, and others.

2.2. Tools/ Languages/ Platform

- **Tools:** Flask, conda*, conda-env
- **Languages:** Python
- **Platform:** Microsoft Visual Studio, Jupyter Notebook

CHAPTER 3

DESIGN AND IMPLEMENTATION

3.1. Architecture Model

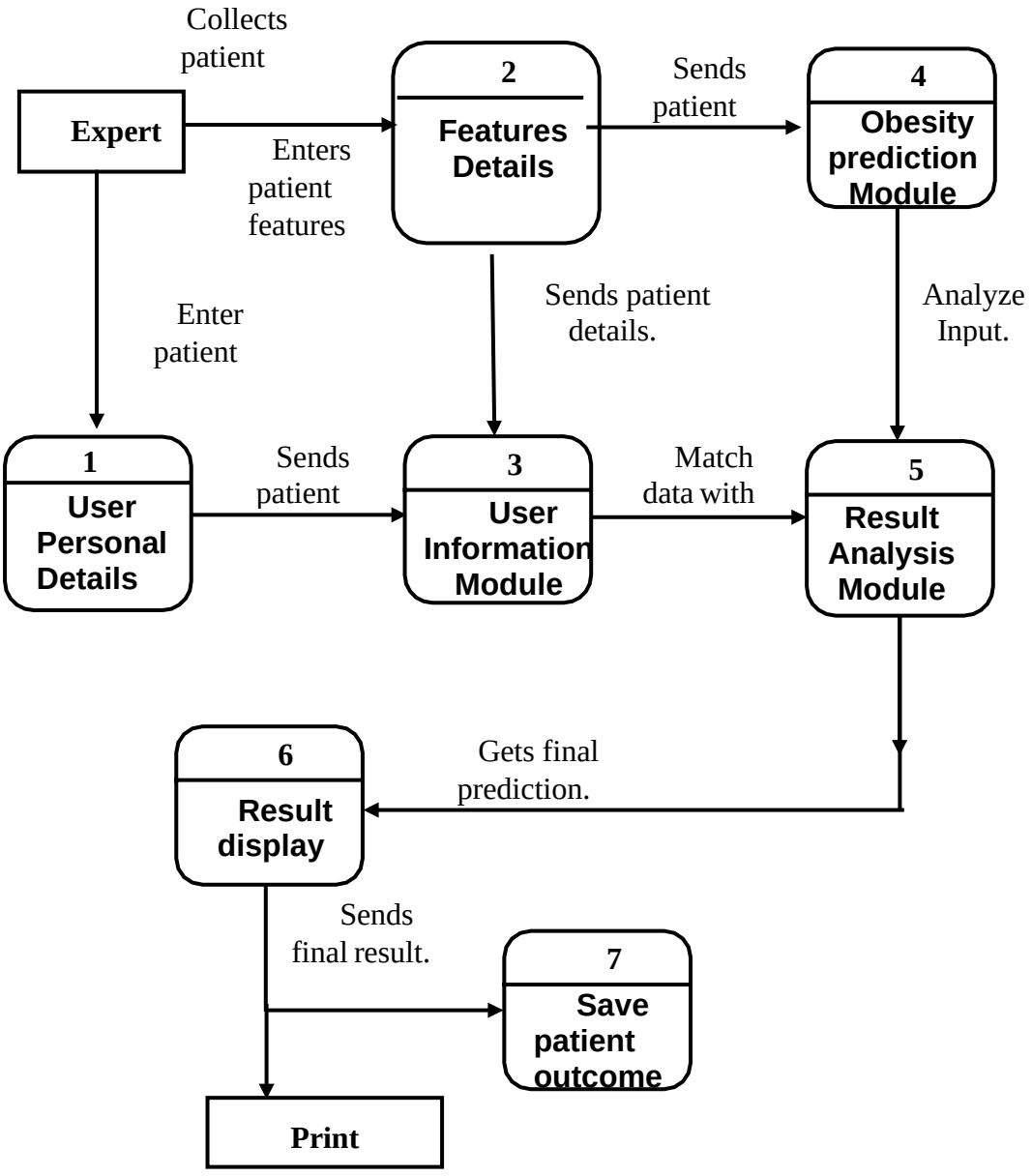


Figure: 3.1 Proposal model

3.2. Flowchart

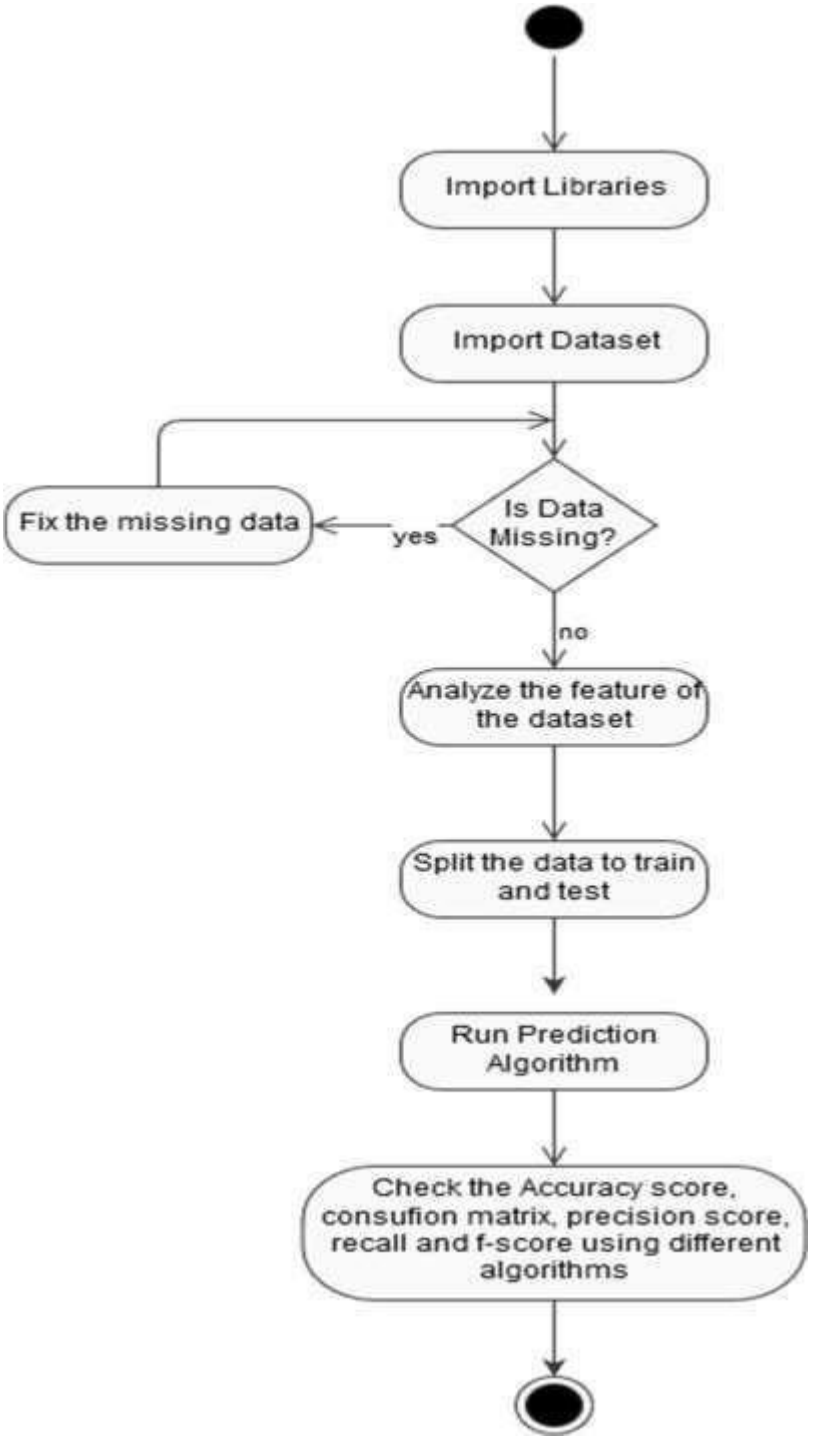


Figure: 3.2 Flow Chart

3.3. Sequence diagram

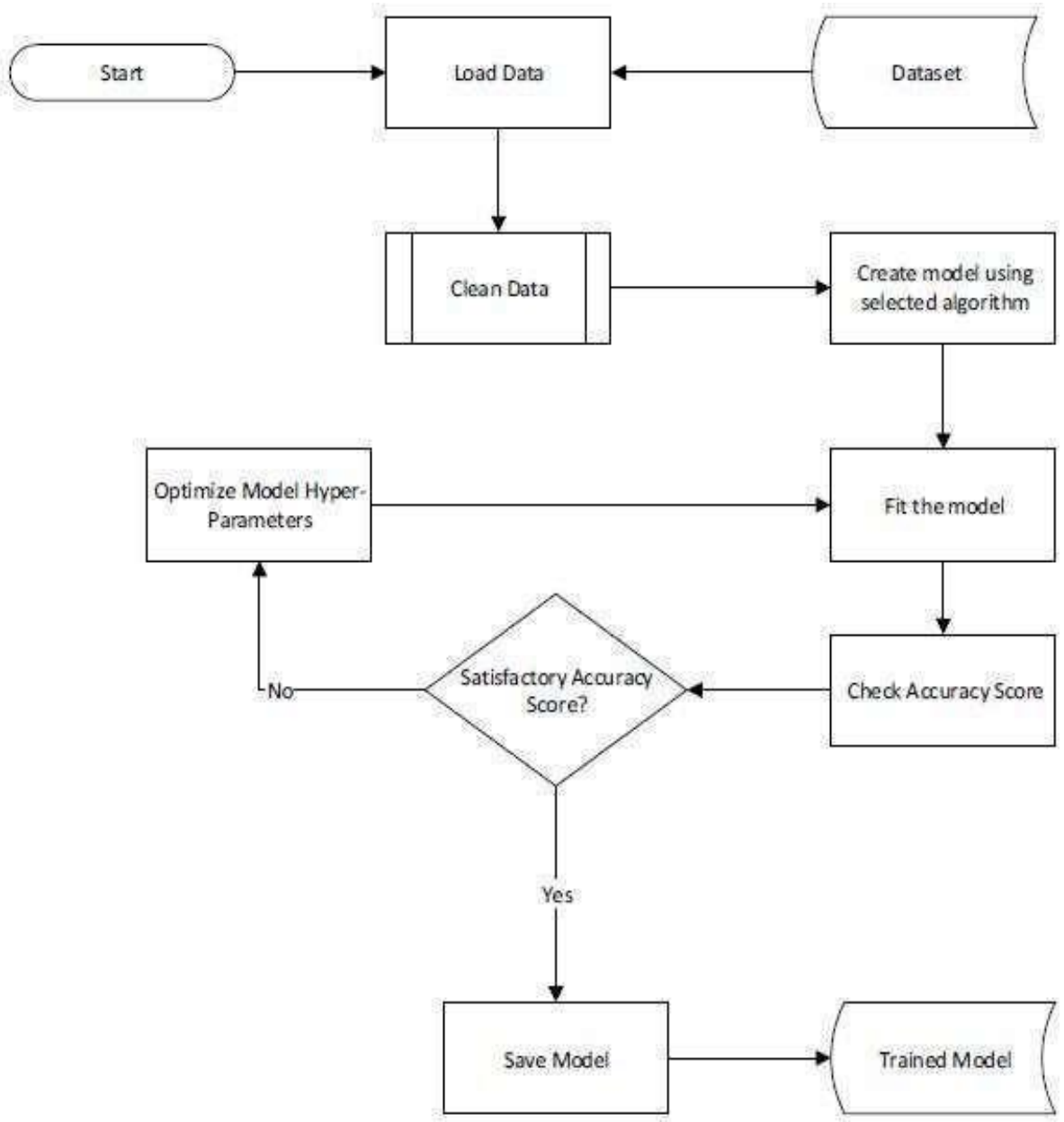


Figure: 3.3 Training Model Process

3.4. Algorithm & Code Segment

3.4.1. Random Forest

Random Forest is a supervised learning algorithm. Random forest can be used for both classification and regression problems, by using random forest regressor we can use random forest on regression problems. But we have used random forest on classification in this project so we will only consider the classification part.

3.4.1.1. Random Forest pseudocode

- Randomly select **"k"** features from total **"m"** features. Where **k << m**
- Among the **"k"** features, calculate the node **"d"** using the best split point.
- Split the node into **daughter nodes** using the **best split**.
- Repeat **1 to 3** steps until the "l" number of nodes has been reached.
- Build forest by repeating steps **1 to 4** for "n" number times to create **"n" numberof trees**.

3.4.1.2. Random Forest prediction pseudocode

- Takes the **test features** and use the rules of each randomly created decision treeto predict the outcome and stores the predicted outcome (target).
- Calculate the **votes** for each predicted target.
- Consider the **highly voted** predicted target as the **final prediction** from therandom forest algorithm.

Code:

```
max_accuracy = 0
for x in range(500):
rf_classifier = RandomForestClassifier(random_state=x)
rf_classifier.fit(X_train,Y_train)
Y_pred_rf = rf_classifier.predict(X_test)current_accuracy =
round(accuracy_score(Y_pred_rf,Y_test)*100,2)
if(current_accuracy>max_accuracy):
max_accuracy = current_accuracybest_x = x
print(max_accuracy)print(best_x) rf_classifier =
RandomForestClassifier(random_state=best_x)
rf_classifier.fit(X_train,Y_train) Y_pred_rf =
rf_classifier.predict(X_test)
```

```
Y_pred_rf.shape
score_rf = round(accuracy_score(Y_pred_rf,Y_test)*100,2)score_rf
```

3.4.2. K-Nearest Neighbors

We can implement a KNN model by following the below steps:

- Load the data
- Initialize the value of k
- For getting the predicted class, iterate from 1 to total number of training data points
  - Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
  - Sort the calculated distances in ascending order based on distance values
  - Get top k rows from the sorted array
  - Get the most frequent class of these rows
  - Return the predicted class

Code:

```
knn_classifier=
KNeighborsClassifier(n_neighbors=31,leaf_size=30)
knn_classifier.fit(X_train,Y_train)
Y_pred_knn = knn_classifier.predict(X_test)
score_knn = round(accuracy_score(Y_pred_knn,Y_test)*100,2)
score_knn
```

3.4.3. Decision Tree

Pseudocode:

- Place the best attribute of the dataset at the **root** of the tree.
- Split the training set into **subsets**. Subsets should be made in such a way that eachsubset contains data with the same value for an attribute.
- Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branchesof the tree.

Assumptions while creating a Decision Tree- At the beginning, the whole training set is considered as the root. Feature values are preferred to be categorical. If the values arecontinuous then they are discretized prior to building the model. Records are distributedrecursively on the basis of attribute values. Order to place attributes as root or internal

node of the tree is done by using some statistical approach.

The popular attribute selection measures:

- Information gain
- Gini index

Attribute selection method- A dataset consists of **"n"** attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy. Tosolve this attribute selection problem, researchers worked and devised some solutions. They suggested using some criterion like **information gain, Gini index,** etc. These criterions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e., the attribute with a high value (in case of information gain) is placed at the root. While using information Gain as a criterion, we assume attributes to be categorical, and for Gini index, attributes are assumed to be continuous. [11]

Gini Index - Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with a lower Gini index should be preferred.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Code:

```
dt_classifier = DecisionTreeClassifier(max_depth=20,
min_samples_split=2, min_samples_leaf=1,
min_weight_fraction_leaf=0.00001,max_features='auto',
random_state=46)
dt_classifier.fit(X_train, Y_train)
Y_pred_dt=dt_classifier.predict(X_test)
score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)score_dt
```

3.4.4. SVM (Support Vector Machine)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

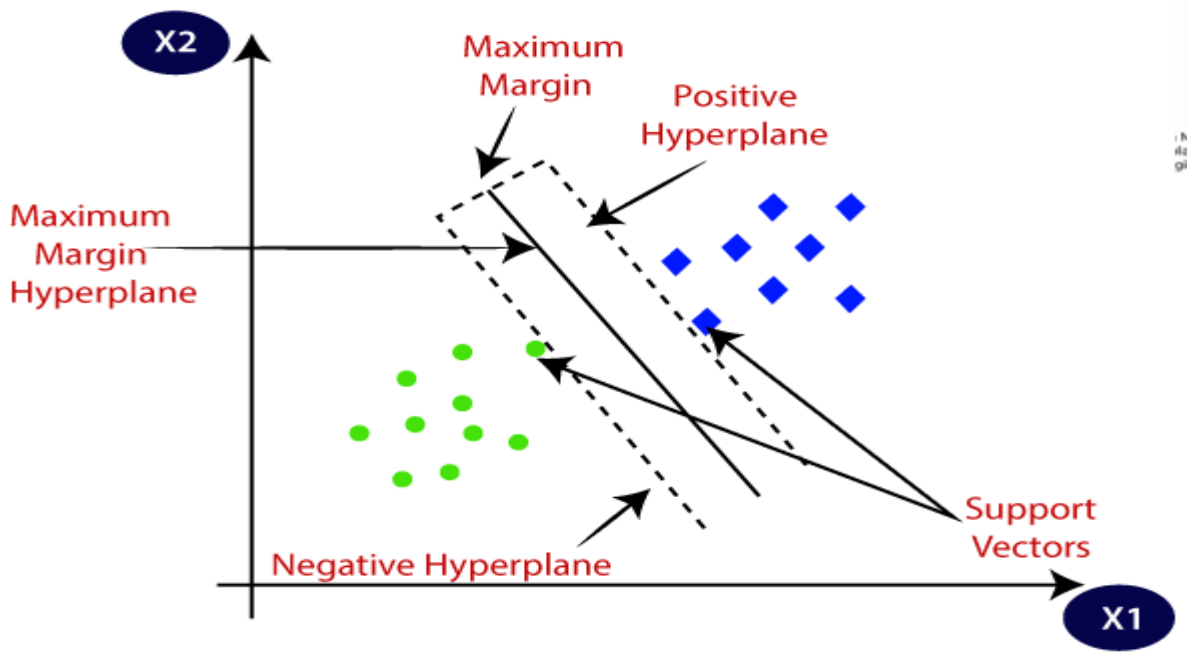


Fig 3.4.4.SVM

Types of SVM

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Code :

```
svm_clf = SVC(kernel='rbf',probability=True)
svm_clf.fit(X_train, Y_train)

# Predicting the Test set results
Y_pred_svm = svm_clf.predict(X_test)

from sklearn import metrics

print("Accuracy:", metrics.accuracy_score(Y_test, Y_pred_svm))

from sklearn.model_selection import cross_val_score

cv_svm = cross_val_score(svm_clf, X, Y, scoring='accuracy', cv=10)
cv_svm = pd.Series(cv_svm)
cv_svm.mean()
```

3.5. Libraries used

Python has a vast reserve of inbuilt standard libraries which includes areas like web services tools, string operation, data analysis, and machine learning, etc. The complex programming tasks can be dealt with ease using these inbuilt libraries as it reduces the sizeof code with many inbuilt functions that do the job pretty well for its user.

3.5.1. Data Visualization

- **Matplotlib:**

Matplotlib is a cross-platform, data visualization and graphical plotting library forPython and its numerical mathematics extension NumPy, a big data numericalhandling resource.

  - pyplot
  - rcParams
  - rainbow

- **Seaborn:**

Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with dataframes and the Pandas library. The graphs created can also be customized easily.

3.5.2. Data Manipulation

- **NumPy:** The NumPy library in python is used for scientific computing and array manipulation. It can perform different operations such as indexing of an array, sequencing, and slicing, etc.

- **Pandas:** The Pandas library in python is used for structuring, manipulating, and organizing data in a tabular structure called the data frame which is further used for data analysis.

- **Scikit-learn:**

  - sklearn.model_selection
    - train_test_split
  - sklearn.preprocessing
    - StandardScaler
    - LabelEncoder

3.5.3. Data Modeling

- **Scikit-learn:**

Scikit-learn is one of the most useful libraries that python offers. It has various statistical learning algorithms such as regression models (linear regression, logistic regression), SVM's, random forest for classification tasks and k-means for clustering, etc.

  - sklearn.ensemble.RandomForestClassifier
  - sklearn.neighbors.KNeighborsClassifier
  - sklearn.tree.DecisionTreeClassifier
  - sklearn.naive_bayes.GaussianNB

3.5.4. Data Validation

- **Scikit-learn-metrics:**

The sklearn.metrics module implements several loss, score, and utility functions to measure classification performance.

sklearn.metrics -

log_loss, roc_auc_score, precision_score, f1_score, recall_score, roc_curve, auc, plot_roc_curve, classification_report, confusion_matrix, accuracy_score, fbeta_score, matthews_corrcoef

- **Mlxtend:**

Mlxtend (machine learning extensions) is a Python library of useful tools for day-to-day data science tasks.

mlxtend.plotting -

plot_confusion_matrix

CHAPTER 4

OBSERVATIONS AND RESULTS

4.1. Testing

Testing is the process used to help identify the correctness, completeness, security, and quality of the developed computer software. Testing is the process of technical investigation and includes the process of executing a program or application with the intent of finding errors.

In the training process, our model learns to associate a particular input (i.e. features) to the corresponding output (tag) based on the test samples used for training. Input features and tags are fed into the machine learning algorithm to generate a model.

A comparative analysis of different classifiers was performed for the classification of the Obesity dataset in order to correctly classify and predict Obesity cases with minimal attributes.

Input	**Expected Output**	**Actual Output**
Data Visualization	Various visual representations of the data to understand more about the relationship between various features.	Pass
Data Processing	Convert some categorical variables into dummy variables and scale all the values before training the Machine Learning models.	Pass

Dataset	Split the dataset into trainingand testing datasets.	Pass
Training dataset	Train the model using thetraining dataset.	Pass
Testing dataset	Tests if the model is accurate based on the output of the testing dataset.	Pass

Table 4.1 Training and subsequent testing

Input	**Expected Output**	**Actual Output**
Obesity Type 1	Should be labeled as 1 (Obesity type 1) and should show output as "The person is not likely to have obesity".	Pass
Overweight 1	Should be labeled as 2 (overweight 1) and should show output as "The patient is likely to have Obesity".	Pass
Overweight 2	Should be labeled as 3 (Overweight 2) and should show output as "The patient is likely to have Obesity".	Pass

Table 4.2 Obesity Test

4.1.1. Model Evaluation

The most important evaluation metrics for this problem domain are Accuracy, Sensitivity, Specificity, Precision, F1-measure,Log Loss, ROC and Mathew correlation coefficient.

- **Accuracy:** which refers to how close a measurement is to the true value and can be calculated using the following formula:

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

- **Precision:** which is how consistent results are when measurements are repeated and can be calculated using the following formula:

Precision = True Positive / (True Positive + False Positive)

- **Sensitivity:**

Sensitivity is a measure of the proportion of actual positive cases that gotpredicted as positive (or true positive). Sensitivity is also termed as Recall.

Sensitivity = True Positive / (True Positive + False Negative)

- **Specificity:**

Specificity is defined as the proportion of actual negatives, which got predicted asthe negative (or true negative).

Specificity = True Negative / (True Negative + False Positive)

- **Mathew Correlation coefficient (MCC):**

The Matthews correlation coefficient (MCC), instead, is a more reliable statisticalrate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}}$$

(worst value: −1; best value: +1)

- **Log Loss:**

Logarithmic loss measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value. A perfect model would have a log loss of
0. Log loss increases as the predicted probability diverges from the actual label. So predicting a probability of .012 when the actual observation label is 1 would be bad and result in a high log loss.

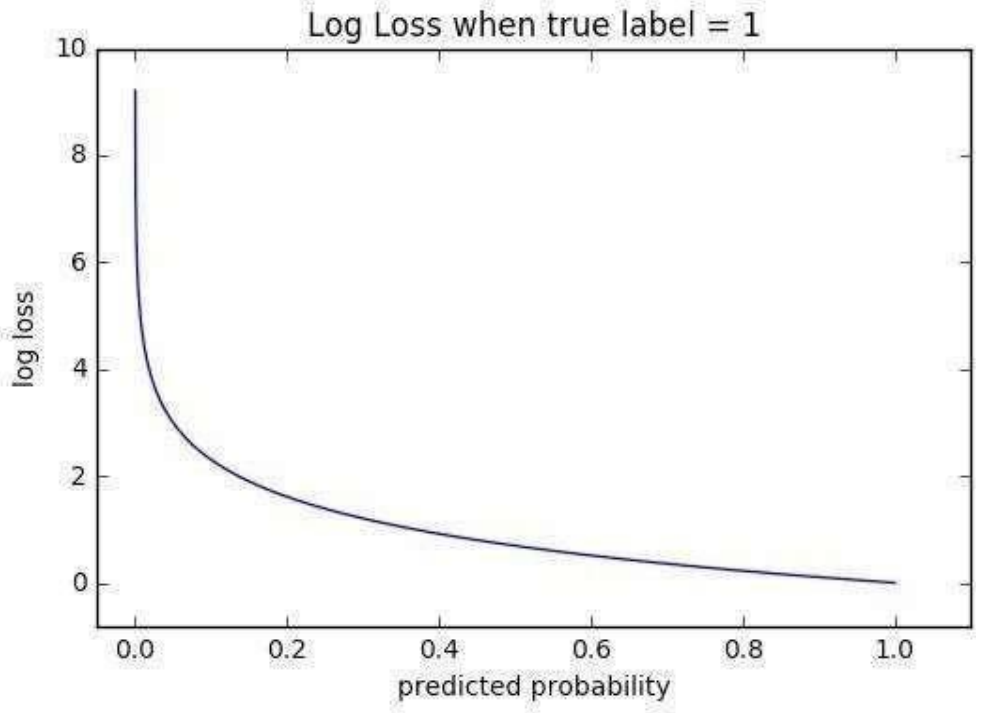


Figure: 4.1.1 Log Loss Graph

- **F1 Score:**

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 score is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost.

F1 Score = 2(Recall Precision) / (Recall + Precision)

- **ROC Curve:**

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate & False Positive Rate.

4.1.1.1. Random Forest Classifier

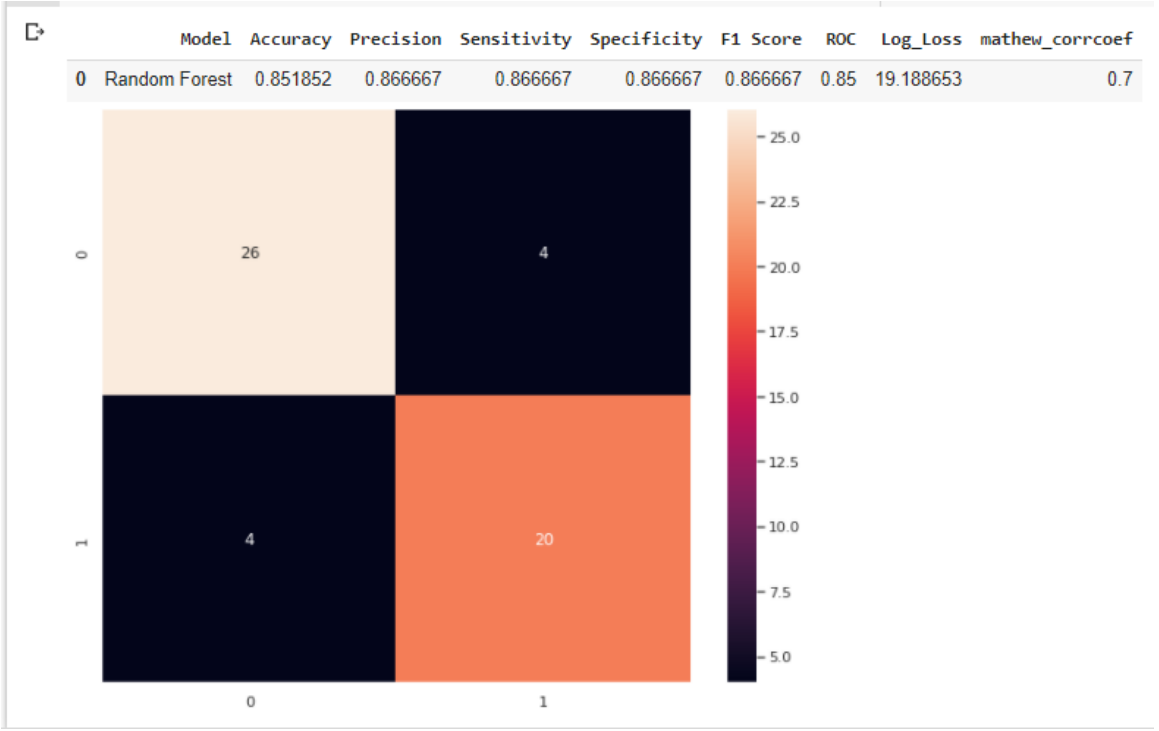
```
y_pred_rfe = rf_classifier.predict(X_test)

plt.Figure(Figuresize=(10, 8))
CM=confusion_matrix(Y_test,y_pred_rfe)sns.heatmap(CM,
annot=True)

TN = CM[0][0]
FN = CM[1][0]
TP = CM[1][1]
FP = CM[0][1]
specificity = TN/(TN+FP)
loss_log = log_loss(Y_test, y_pred_rfe) acc=
accuracy_score(Y_test, y_pred_rfe) roc=roc_auc_score(Y_test,
y_pred_rfe)
prec = precision_score(Y_test, y_pred_rfe)rec =
recall_score(Y_test, y_pred_rfe)
f1 = f1_score(Y_test, y_pred_rfe)
mathew = matthews_corrcoef(Y_test, y_pred_rfe)

model_results =pd.DataFrame([['Random Forest',acc,
prec,rec,specificity, f1,roc, loss_log,mathew]],
                columns = ['Model', 'Accuracy','Precision',
'Sensitivity','Specificity', 'F1
Score','ROC','Log_Loss','mathew_corrcoef'])

model_results
```



	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Random Forest	0.851852	0.866667	0.866667	0.866667	0.866667	0.85	19.188653	0.7

Figure: 4.1.1.1 Random Forest Confusion Matrix

```
Y_pred_rf = np.around(Y_pred_rf)
print(metrics.classification_report(Y_test,Y_pred_rf))


plot_roc_curve(rf_classifier,X_test,Y_test)plt.xlabel('False
Positive Rate') plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic Curve')
```

4.1.1.2. K-Nearest Neighbors Classifier

```
y_pred_knne = knn_classifier.predict(X_test)

plt.Figure(Figuresize=(10, 8))
CM=confusion_matrix(Y_test,y_pred_knne)sns.heatmap(CM,
annot=True)

TN = CM[0][0]
FN = CM[1][0]
TP = CM[1][1]
FP = CM[0][1]
specificity = TN/(TN+FP)
loss_log      =      log_loss(Y_test,      y_pred_knne)      acc=
accuracy_score(Y_test,   y_pred_knne)   roc=roc_auc_score(Y_test,
y_pred_knne)
prec = precision_score(Y_test, y_pred_knne)rec =
recall_score(Y_test, y_pred_knne)
f1 = f1_score(Y_test, y_pred_knne)
mathew = matthews_corrcoef(Y_test, y_pred_knne)

model_results =pd.DataFrame([['K-Nearest Neighbors ',acc,
prec,rec,specificity, f1,roc, loss_log,mathew]],
               columns = ['Model', 'Accuracy','Precision',
'Sensitivity','Specificity', 'F1
Score','ROC','Log_Loss','mathew_corrcoef'])

model_results
```

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	K-Nearest Neighbors	0.796296	0.787879	0.866667	0.866667	0.825397	0.7875	19.188653	0.586094

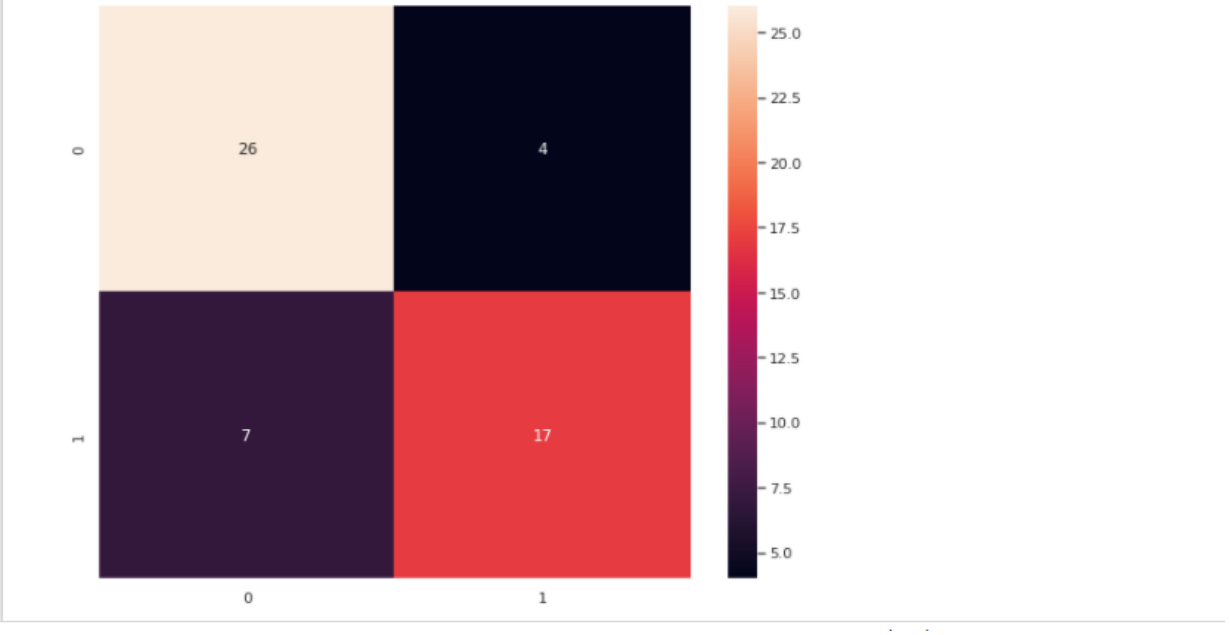


Figure: 4.1.1.2 K-Nearest Neighbors Confusion Matrix

```
Y_pred_knn = np.around(Y_pred_knn)
print(metrics.classification_report(Y_test,Y_pred_knn))
```

```
    plot_roc_curve(knn_classifier,X_test,Y_test)plt.xlabel('False
Positive Rate') plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic Curve')
```

4.1.1.3. Decision Tree Classifier

```
y_pred_dte = dt_classifier.predict(X_test)

plt.Figure(Figuresize=(10, 8))
CM=confusion_matrix(Y_test,y_pred_dte)sns.heatmap(CM,
annot=True)

TN = CM[0][0]
FN = CM[1][0]
TP = CM[1][1]
FP = CM[0][1]
specificity = TN/(TN+FP)
loss_log      =      log_loss(Y_test,      y_pred_dte)      acc=
accuracy_score(Y_test,    y_pred_dte)    roc=roc_auc_score(Y_test,
y_pred_dte)
prec = precision_score(Y_test, y_pred_dte)rec =
recall_score(Y_test, y_pred_dte)
f1 = f1_score(Y_test, y_pred_dte)
mathew = matthews_corrcoef(Y_test, y_pred_dte)

model_results =pd.DataFrame([['Decision Tree',acc,
prec,rec,specificity, f1,roc, loss_log,mathew]],
               columns = ['Model', 'Accuracy','Precision',
'Sensitivity','Specificity', 'F1
Score','ROC','Log_Loss','mathew_corrcoef'])

model_results
```

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Decision Tree	0.703704	0.769231	0.666667	0.666667	0.714286	0.708333	19.188653	0.414371

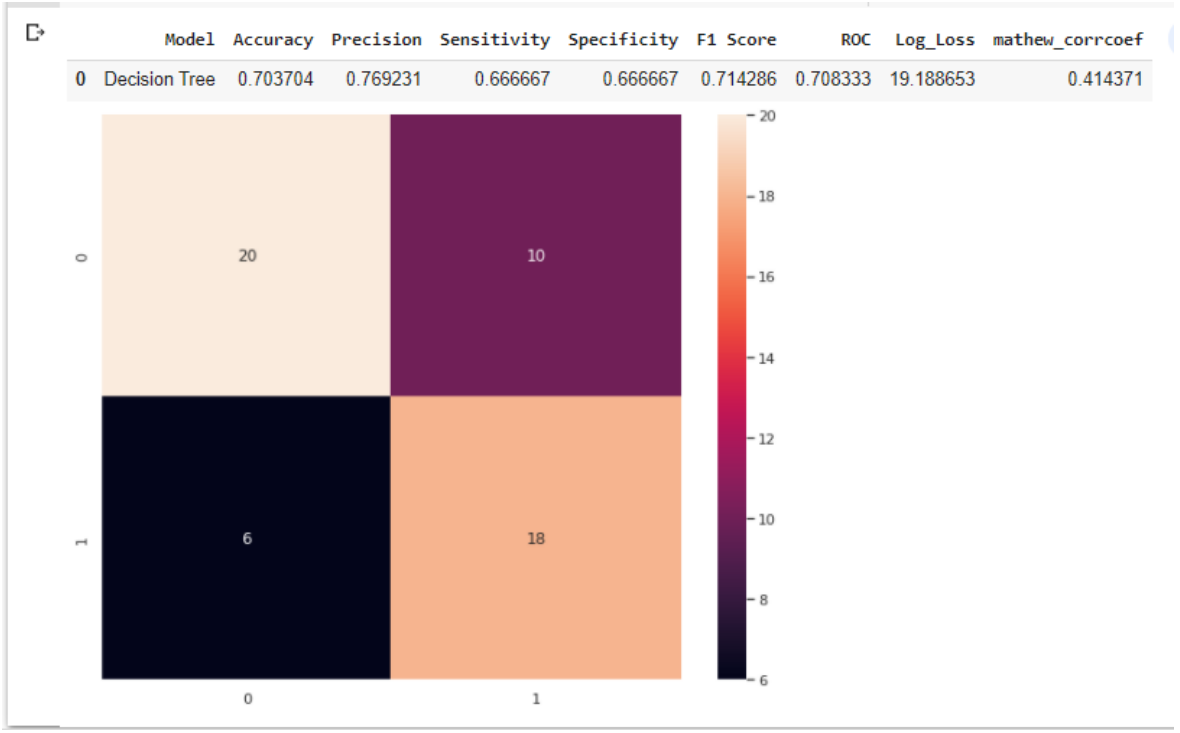


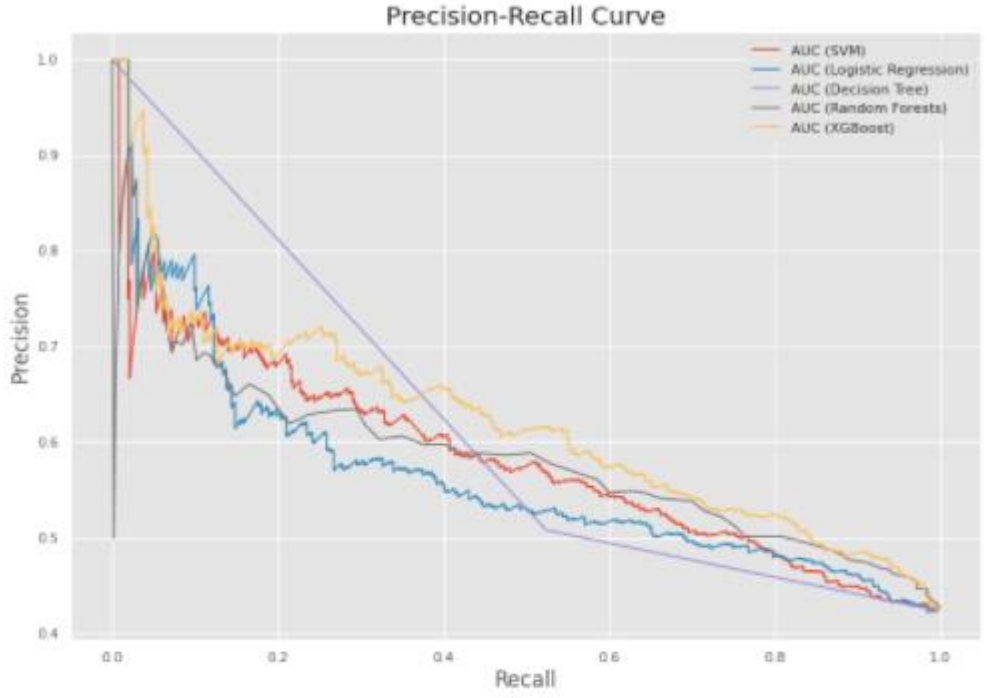
Figure: 4..1.1.3 Decision Tree Confusion Matrix

4.2. Results

Algorithms	Accuracy	Precision	Recall	F1-score
SVM	63.49%	64%	32%	43%
Logistic Regression	61.33%	57%	35%	43%
Decision Tree	58.36%	51%	51%	51%
Random Forest	63.40%	59%	45%	51%
XGBoost	64.93%	61%	47%	53%

Fig 4.2 Final Result

4.3. Graphs




```
AUC of Logistic Regression: 0.56
AUC of SVM: 0.58
AUC of Random Forest: 0.58
AUC of Decision Tree: 0.61
AUC of XGBoost: 0.62
```

Figure: 4.3 Accuracy Score Precision-Recall Curve

4.4. Snapshots

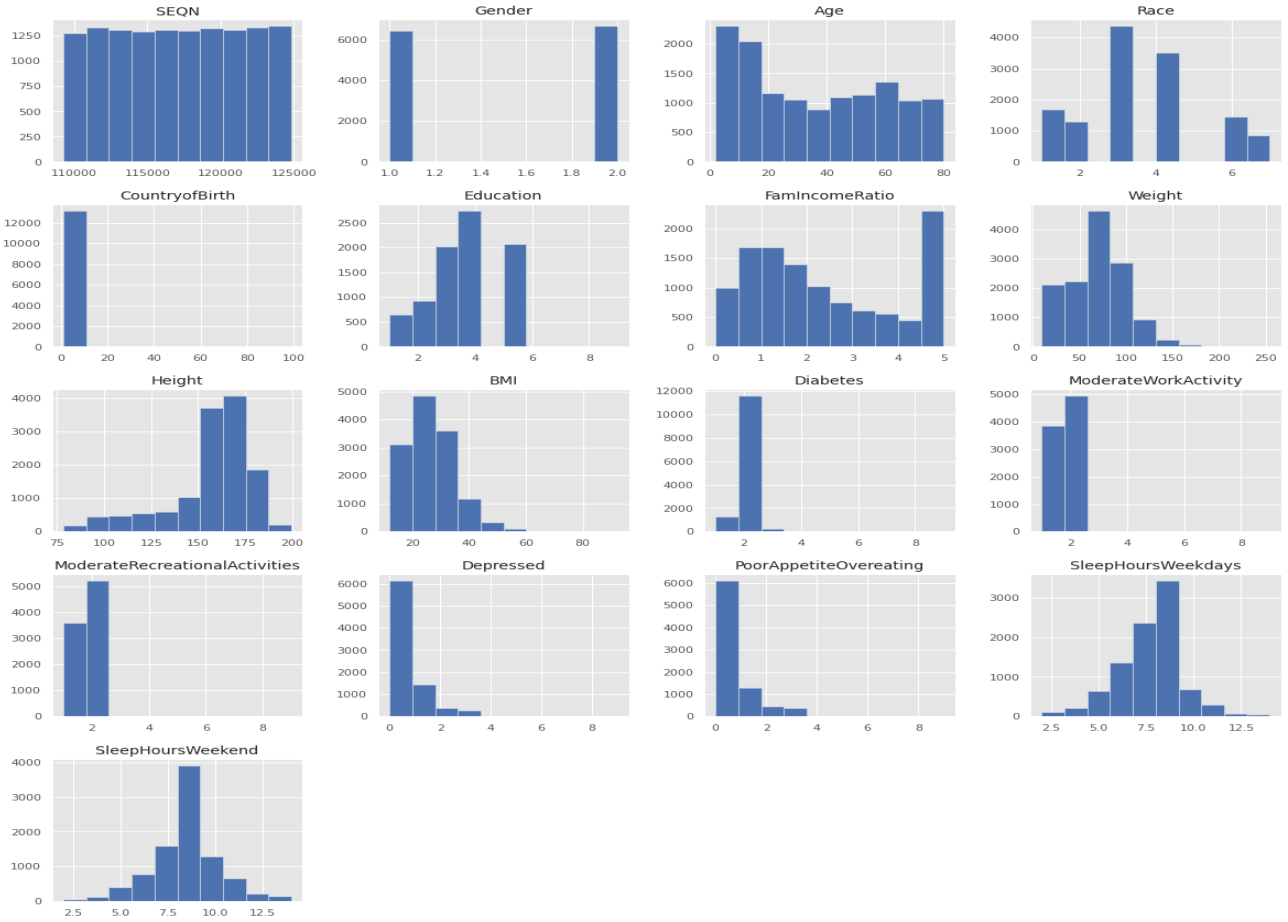


Fig 4.4.1 Histogram

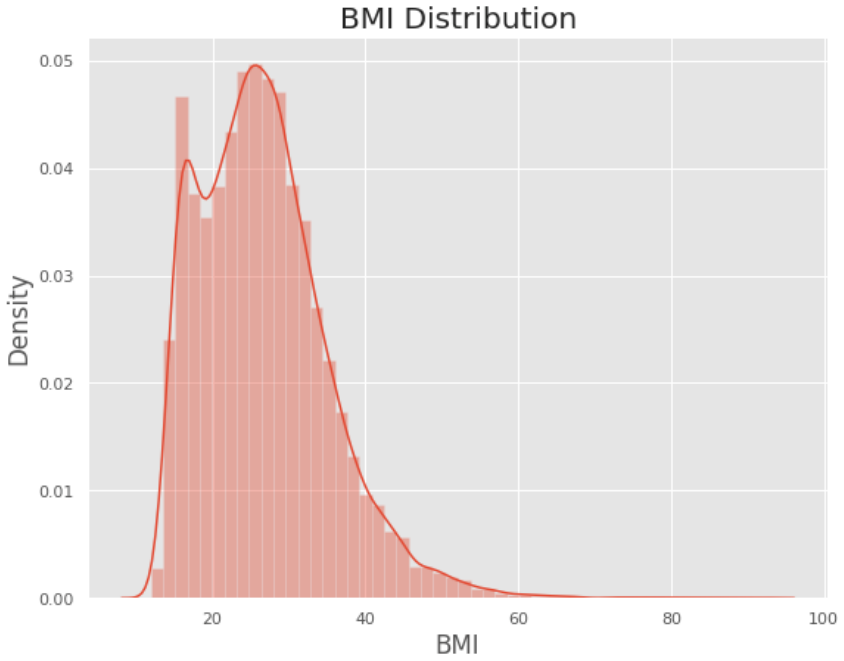


Fig 4.4.2 BMI Distribution

Fig 4.4.4 Heatmap

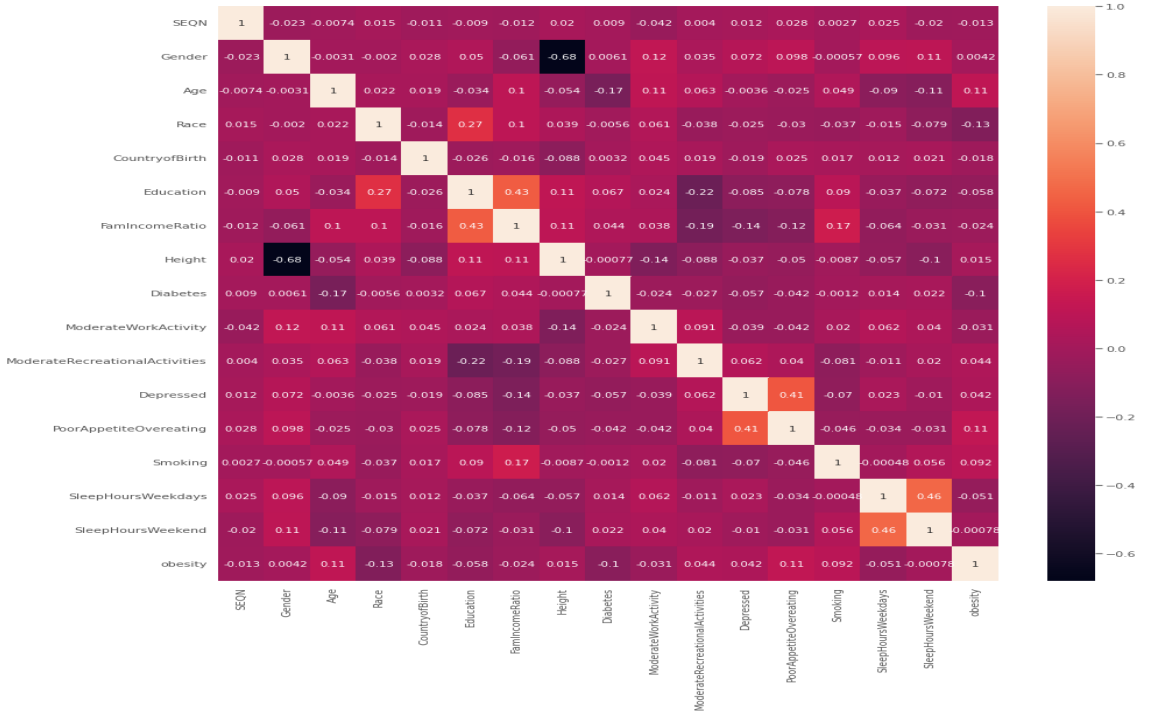


Fig 4.4.5 Download

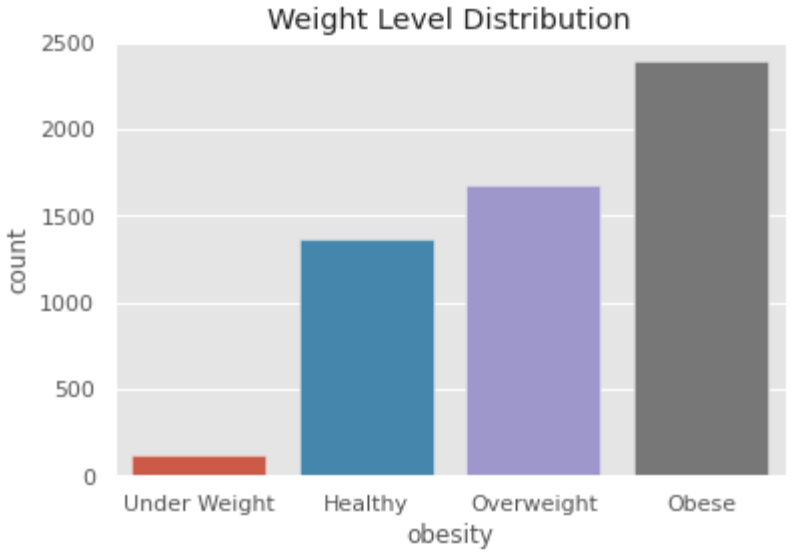


Fig 4.4.6 Weight Level Distibution

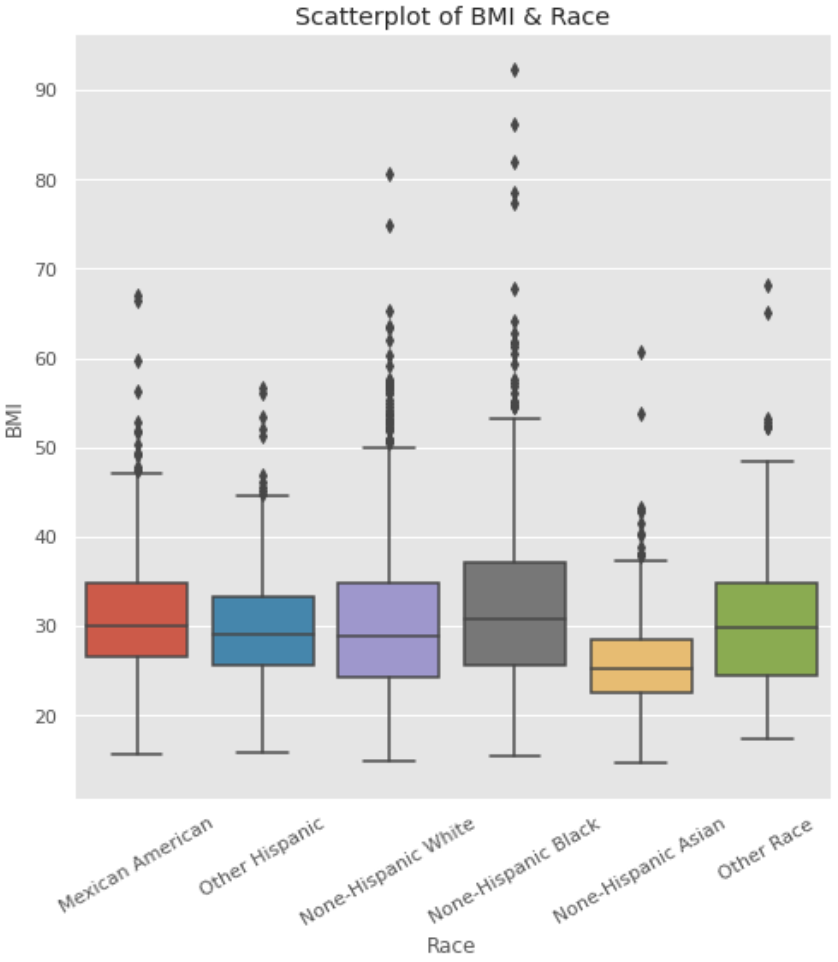


Fig 4.4.7 Scatterplot of BMI and Race

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1. Conclusion

Diagnosing obesity is difficult, as it is a complex disease that varies in nature. Improvement of the diagnosis of obesity is needed in the health sector to help reduce the risk/implications to the barest minimal. To determine the obesity status of a patient, a physician must carry out some physical assessment on the patient and examine the results of a patient's test to assess prior judgments because it is subject to the doctor's interpretation. This project with the development of the Obesity prediction model would solve these issues of validity of diagnosis, time-consuming factor and also provide a reliable diagnosis system that can be used for all gender. A reliable and time-saving obesity prediction model has been developed in this research. The machine learning model was designed using the python programming language. Algorithms that were used to achieve the aim of this research are random forest classifier, K-nearest neighbor, and Support vector machine. The Support vector machine algorithm outperformed the other algorithms compared, with an accuracy of 65.05%. This prediction model is highly recommended for hospitals, clinics, diagnostic centers and the health sector in general, as it will help them to accurately predict obesity status in patients before it gets to a complex stage

5.2. Limitations

The Algorithms used in our project do not give a 100% accuracy, so the prediction is not 100% feasible. Clinical diagnosis and diagnosis using our project may differ slightly because the prediction is not 100% accurate. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on the doctor's intuition and experience rather than on the knowledge-rich data collected from the dataset.

5.3. Future Work

In addition, the report predicted that the percentage of the global population with obesity will increase from 14% in 2020 to 24% in 2035. Obesity rates are also expected to grow significantly among children and adolescents. Among boys, obesity rates are expected to grow from 10% to 20% between 2020 and 2035, and obesity rate among girls are expected to grow from 8% to 18% during that same period.

In the United States, the annual increase in adult obesity is high at 2.1%, and the percentage of U.S. adults with obesity is predicted to reach 58% by 2035. Among children, the annual increase in obesity is estimated to be 2.4%, and the percentage of U.S. boys and girls with obesity is predicted to be over 40% and close to 35%, respectively, by 2035.

According to Louise Baur, a professor and pediatrician at the **University of Sydney** and president of WOF, no country has seen their obesity rates, including childhood obesity rates, decline since 1975.

"This means more adolescents now enter adulthood with established risk factors for chronic disease — they're more likely to develop type 2 diabetes, or have heart disease risk factors or orthopedic problems, sleep apnea or fatty liver disease," Baur said.

REFERENCES

[1] Halaudi Daniel M., "Prediction of heart disease using classification algorithms.," WCSECS, pp. 22–24, 2014.

[2] U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," in 2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016, 2017.

[3] T. J. Peter and K. Somasundaram, "Study and Development of Novel Feature Selection Framework for Heart Disease Prediction," Int. J. Sci. Res. Publ., 2012.

[4] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," IEEE Trans. Cybern., 2013.

[5] C. Ordonez, "Improving Heart Disease Prediction using Constrained Association Rules," Tech. Semin. Present. Univ. Tokyo, 2004.

[6] M. C. and P. M. Franck Le Duff, CristianMunteanb, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method," Stud. Health Technol. Inform., vol. Vol. 107, no. 2, p. No. 2, pp. 1256–1259, 2004.

[7] W. J. F. and G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview," AI Mag., vol. Vol. 13, N, no. 3, pp. 57–70, 1996.

[8] K. Y. N. and K. H. R. Heon Gyu Lee, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," Proc. Int. Conf. Emerg. Technol. Knowl. Discov. Data Min., p. pp. 56–66, 2007.

[9] L. P. and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFISand Genetic Algorithm," Int. J. Biol. Biomed. Med. Sci., vol. Vol. 3, no. No. 3, pp. 1-8,2008.

[10] UCI Dataset (https://archive.ics.uci.edu/ml/datasets/statlog+(heart)).

[11] M. Kukar, I. Kononenko, C. Grošelj, K. Kralj, and J. Fettich, "Analyzing and improving the diagnosis of ischaemic heart disease with machine learning," Artif. Intell. Med., vol. 16, no. 1, pp. 25–50, May 1999.