"A MACHINE LEARNING MODEL TO IDENTIFY AND PREVENT THE OCCURRENCE OF ADAPTIVE THERMOGENESIS"

A thesis submitted to the
University of Petroleum & Energy Studies

For the award of
DOCTOR OF PHILOSOPHY

in

Computer Science

BY

Madam Chakradar

April 2023

SUPERVISOR

Dr. Alok Aggarwal

School of Computer Science
University of Petroleum & Energy Studies
DEHRADUN-248007: Uttarakhand

"A MACHINE LEARNING MODEL TO IDENTIFY AND PREVENT THE OCCURRENCE OF ADAPTIVE THERMOGENESIS"

A thesis submitted to the
University of Petroleum & Energy Studies

For the award of
DOCTOR OF PHILOSOPHY
in
Computer Science

BY

Madam Chakradar
(SAP ID- 500057009)

April 2023

Supervisor

Dr. Alok Aggarwal
Professor
School of Computer Science
University of Petroleum and Energy Studies

School of Computer Science and Engineering
University of Petroleum & Energy Studies
DEHRADUN-248007: Uttarakhand

April 2023

DECLARATION

I declare that the thesis entitles **"A MACHINE LEARNING MODEL TO IDENTIFY AND PREVENT THE OCCURRENCE OF ADAPTIVE THERMOGENESIS"** has been prepared by me under the guidance of **Dr. Alok Aggarwal**, Professor of school of computer science, university of petroleum and energy studies. No part of this thesis has formed the basis for the award of any degree or fellowship previously.

Madam Chakradar

Department of Computer Science,
School of Computer Science
Dehradun-248007: Uttarakhand

DATE: 25th April. 2023

CERTIFICATE

I certify that **Madam Chakradar** has prepared this thesis entitled **"A Machine Learning Model To Identify And Prevent The Occurrence Of Adaptive Thermogenesis"**, for award of PhD degree of the university of petroleum and energy studies, under my guidance. He has carried out the work at the department of School of computer science and engineering, University of petroleum and energy studies.

(Dr. Alok Aggarwal)
Professor
School of Computer Science,
School of Computer Science
Dehradun-248007: Uttarakhand

Date: 25th April. 2023

ABSTRACT

There are more people seeking to reduce extra weight because it would improve their health due to the rising percentage of overweight people in the world. So, the process of shredding excess on the human body isn't an easy task, because the process varies from body to body based on gender, age, height, weight, medical history, and sometimes even based on hereditary impacts. Hence, the solution of the problem would also not be the same for every possible combination of disparities earlier discussed. But based on the literature it is observed that a healthy diet and lifestyle contributes about 90% of the entire journey of successful sustainable weight loss and 10% to physical activity to boost the growth of fat free mass replacing fat mass in the human body. On top of following such rigorous diet and workouts, there are individuals who are still unable to get rid of weight and some regain the lost weight over a period of time. So, the projection of how much weight loss is permitted within a period of time is still a massive research gap. All these problems are being explained by a phenomenon called Adaptive Thermogenesis, wherein thermogenesis is the process of heat generation in living organism with the help of energy contained or consumed by the organism and adaptive thermogenesis is phenomenon where it affects this process of heat production in warm blooded beings. Poor body temperature means not providing better conditions for the body to perform at its best, over time degrading the unit health of organs and their functionality. And the energy difference of not producing more heat for the body is stored as fat adding more weight to the body over the time. Hence to understand such tricky problems like adaptive thermogenesis (AT) machine learning is used. In the current work it is being observed whether it is possible to identify whether the person is undergoing adaptive thermogenesis while losing weight and if it is happening how could one prevent this without performing any incisions on the human body.

To solve a problem like AT, one would need lots of tests and observations to extract conclusive evidence of what works and what doesn't for a particular human body. Due to the advancements in the fields of machine learning and artificial intelligence, researchers are now coming up with better and faster solutions irrespective of their areas of interest unless the researcher can break down the observations into machine understandable problems. A similar approach has been taken in the current study with the help of machine learning providing a probabilistic approach for the decision-making process. For an experiment to recreate, one would need some observations based on whether he/she could create the system. Over this system, numerous amounts of data are tested to see whether there is synergy between them or else there are disagreements.

In the current scenario the system is designed with the help of a real-world 2 yearlong clinical trial based on restricted calorie intake to lose weight called the CALERIE study. These observations from this study created a base for the models that are built in the current work.

The first phase of the work is to identify the occurrence of AT while losing excess body weight. As discussed earlier this study is purely focused on identifying the relationships between non-invasive parameters against the parameter of interest AT. Indirect calorimetry was used to measure the energy difference between the expected ideal resting metabolic rate and the actual resting metabolic rate in order to identify non-invasive parameters from the wide variety of tests. Negative value of AT supports the evidence of adaptive thermogenesis with tolerance levels of 5%. But there were few missing data in the observations which were filled building another model to refill the missing values. A bunch of models were using various machine learning techniques after converting the problem into a binary classification approach. Logistic regression model would generalize better even with lesser data, but Explainable boosting

machines (EBM) approach provided better results with greater amount of data for training achieving 78% accuracy. Therefore, the model built using EBM technique is used to observe the identification of adaptive thermogenesis in humans.

Now prevention of any situation is only possible with proper understanding of the problem. Overweight is the situation when the body is incapable of burning the energy and through literature review it is identified if the blood glucose levels if not consumed by the body will end up becoming unhealthy fat. As a metabolic failure sometimes the insulin whose primary job is to be carrier for blood glucose becomes weak and this condition is called insulin resistance. So, monitoring insulin resistance could help users track their metabolic activity helping them understand their body while undergoing weight loss. In the current study a model is proposed to identify whether the individual is experiencing insulin resistance. This model is built using machine learning based classification techniques. The data for this study is also used from the CALERIE study dataset. In the end logistic regression model provided a fruitful result providing an impressive accuracy score of 0.96 with nearly 0.6 AUC score. If both models, one for identification of Adaptive thermogenesis and two classification of individuals with insulin resistance are run simultaneously a collective goal of weight loss can be achieved with less consequences and pitfalls along the journey.

ACKNOWLEDGEMENT

I am highly indebted to my research supervisor Dr. Alok Aggarwal, who has guided me all through the completion of this Ph.D. research work. Their profound knowledge inspired my enthusiasm in this subject in this domain and elegant research styles. They have helped me with wise advice, valuable discussions, comments, and facilities. They never hesitated to spend precious time and effort to guide my work. My guides helped me overcome the various difficulties and challenges raised at different stages of this research work. They always motivated me with their positive inputs and always supported me whenever I got stuck. It is just ineffable to express my deep gratitude towards them.

My special thanks to my friends Dr. Kiran Kumar, School of Computer Science, for their enlightened guidance and encouragement during the preparation of the thesis.

I acknowledge my indebtedness to the Librarian Dr. Prem Prakash Sati, Mr. Jatendra Sharma Sr. Manager, and Dharmendra Chauhan, Server Administrator, Department of IT, University of Petroleum & Energy Studies, Dehradun, for providing me with the library and computing resources as and when required.

The idea of inspiration for this research topic is identified by Dr. K.D. Hall, PhD and Dr. Mcgrath's research work who inspired me to pursue the topic on adaptive

thermogenesis and mathematical understanding of human body functioning.

It wouldn't be possible without immense support and understanding I share with the love of my life, which is my wife Manusha.

I want to apologize to those whose names do not figure here but have helped me during my research tenure.

Lastly, I would like to dedicate this work to my parents Ram Narayana and Vanija for being present with me all the time with their blessings. As this work is written in focus to keep them healthy as long as I'm alive.

-MADAM CHAKRADAR

TABLE OF CONTENTS

ABBREVIATIONS

AT	:	Adaptive Thermogenesis
WHO	:	World health Organization
FM	:	Fat mass
FFM	:	Fat free mass
RMR	:	Resting metabolic rate
TDEE	:	Total Daily Energy Expenditure
BMR	:	Basal Metabolic Rate
nREE	:	non-Resting Energy Expenditure
REE	:	Resting Energy Expenditure
TEF	:	Thermic Effect of food
NEAT	:	Non-Exercise Activity Thermogenesis
EAT	:	Exercise Activity Thermogenesis
DIT	:	Diet Induced Thermogenesis
CIT	:	Cold Induced Thermogenesis
EE	:	Energy Expenditure
DXA	:	Dual-energy x-ray absorptiometry
RFE	:	Recursive Feature Elimination
RSME	:	Root square mean error
CV	:	cross validation
OLS	:	ordinary least squares
DT	:	Decision Tree
LR	:	logistic Regression
EBM	:	Explainable boosting machines
BMI	:	Body Mass Index
T2DM	:	Type 2 Diabetes Mellitus

ANN	:	Artificial Neural Networks
SVM	:	Support Vector Machine
ML	:	Machine Learning
TGs	:	Triglycerides
HDL-c	:	High-Density Lipoprotein-c
KNN	:	K Nearest neighbour
EMR	:	Electronic Medical Records
LDA	:	Linear Discriminant Analysis
AUC	:	Area under the curve
DNA	:	Deoxyribonucleic acid
WhtR	:	Waist to height ratio
HOMA-IR	:	Homeostasis model assessment-estimated insulin resistance
RFEcv	:	RFE with Cross-validation
PCA	:	Principal component analysis
GBM	:	Gradient boosted machine
WHR	:	Waist to Hip ratio
HDL-c	:	High-Density Lipoprotein-c
InGDR	:	Induced Glucose diffusion rate
RFC	:	Random Forest classifier
NBC	:	Naïve Bayes classifier
ROC	:	Receiver operating characteristic

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Prevention is better than cure but when prevention is not leading in the right direction then the problem is riddled, one such problem is called Adaptive Thermogenesis (AT). In 2007, researchers at the University of Ottawa proposed that “adaptive thermogenesis,” or the process by which body weight decreases when caloric intake is insufficient to meet energy demands, occurs in response to a decrease in energy intake that is not explained by changes to physical activity. The resultant of this is regaining the lost weight by over-expenditure of energy or under consumption of energy which ultimately leads to serious weight plateaus. Obesity and overweight rank sixth among causes of death worldwide. Every year, being overweight or obese contributes to at least 2.8 million adult deaths. As WHO suggested, many individuals identified maintaining a healthy weight as good for the body. Even those who are successful in losing weight temporarily frequently acquire back a significant amount of it. Nearly half of the weight lost during the first year following initial loss is gained back, suggesting a general propensity for weight cycling. The main goal for a successful weight loss is to reduce the loss of fat-free mass by reduction of fat mass alone. Now the disagreement for estimation of FFM after weight loss is referred to as AT (Adaptive Thermogenesis). And when FFM decreases it tends to impact RMR directly to reduce the overall energy expenditure. When the body is provided with lesser energy than its usual needs it intelligently moves to starvation mode allocating majority of energy to be stored as fat, dropping energy supplies to other non-essential needs of the body [56]. This has opened the gate for new problems where individuals ultimately get back to their old problems. As told before Prevention is better than cure but here prevention is a change in lifestyle for the human body, not just certain habits. Adaptive thermogenesis is an act of the brain to sustain lost body weight through many hormonal feedbacks in response to the current inputs and surroundings given to the body. So, in the current progressive world with growing demand for fitness trackers showed importance as these devices are more user centric with the algorithms mapped behind the inputs of these trackers.

The study of energy homeostasis is important for understanding obesity and developing new treatments. The potential of intelligible and individually predictive medically applicable models of energy balance is made possible by meticulous recording of each person's behaviour and novel imaging technologies. Due to the abundance of data coming from these sources, there is interest in using machine learning techniques to extract knowledge from these sizable, reasonably organised datasets.

For the current work CALERIE study's dataset is considered. The CALERIE trial was a 2-year randomized controlled trial with a 25% caloric restriction diet or an ad libitum control diet and the focus of the study was cardiometabolic risk factors. The trial involved 1,145 people in the calorie restriction group and 551 in the control group. It was conducted in three clinical facilities. The goal was to examine the physiological adaptations of calorie restriction, and if the results indicated the benefits and support of a calorie restricted diet. Also, to evaluate the effects of adaptations for cardiometabolic syndromes like blood glucose tolerance, blood pressure, few inflammatory markers, etc. One could argue there is a huge gap between good clinical trials focused on calorie restriction after Minnesota starvation study and CALERIE study has promising compilations. With such a wide variety of clinical parameters in the selection, this clinical trial has a good enough sample size to perform statistical and machine learning-based studies. The major outcomes of this study were to understand the body core temperature, resting metabolic rate in humans while regulated calorie restriction.

1.1 NEED AND SCOPE OF MACHINE LEARNING FOR ADAPTIVE THERMOGENESIS

Of the 2.8 million annual deaths worldwide due to overweight and obesity, 23 percent of ischaemic heart disease deaths, 7 to 41 percent of certain cancers, and 44 percent of type 2 diabetes are attributable to weight gain. Based on these facts the World Health Organization has ranked overweight and obesity as the fifth leading risk for global deaths, accounting for 5% deathrate in the world due to illness. According to the World Health Organization, between 1990 and 2016, there were 4 million more overweight or obese children in the WHO African area. That's a 125% growth in 16 years from the continent with the poorest economy-driven countries in the world. It means the world is heading towards a future with more obese and overweight-based health issues. So, enough attention needs to be given to these statistics to put forth a promising healthy future to tackle these problems.

So, the contribution of machine learning towards wellness technologies is through the algorithms created over the data derived from wearable fitness trackers like Fitbit, Garmin, Apple, Misfit, Polar, Samsung, Xiaomi, etc. Few other predefined calorie calculation applications using machine vision, questionnaires, etc. Examples of such applications are myfitnesspal, Healthifyme, etc providing a variety of diet regimes with the least information about their users. The success rate of such user-dependent applications is generally prone to huge errors. To counter these problems researchers have looked towards machine learning and big data for rescue. So, many clinical and behavioural research trials have shown some promise but weight loss though it is achieved once has to be sustained which can be sorted only when adaptive thermogenesis is kept under control. Hence, the problem for current research work is to identify the occurrence of adaptive thermogenesis and prevent it.

1.2 PROBLEM STATEMENT

To explore the causes behind weight regains after weight loss due to adaptive thermogenesis and model the behavioural pattern using the physiological inputs of an individual to avoid chances of weight regains in humans.

1.3 OBJECTIVE

"To explore the causes behind weight regains after weight loss due to adaptive thermogenesis and modelling the behavioural pattern using the physiological inputs of an individual to avoid chances of weight regains in humans"

Sub Objectives:

i. Identification of input parameters that impacts the physiological behavioral model of the user.
ii. Identification of Adaptive Thermogenesis by building a machine learning model with the data.
iii. Prevention of Adaptive Thermogenesis by building a machine learning model with the data.
iv. Validation of the model by fitting different profiles data.

1.4 ORGANIZATION OF THESIS

The thesis is organized chapter-wise so that the structure of the research work done can be read seamlessly. It is tried to keep the chapters of this thesis in a sequence so the readers can generate interest in reading and understanding. The research thesis starts with the introduction and government statistics of road accidents and road Infrastructure development. That makes the reader get a better experience of the requirements of this crucial research work. The remaining chapters mainly focus on the methodology, framework, and experiments conducted to validate the hypothesis. The organization of the thesis is shown as per figure 2 below.

Chapter 1: includes an elaborate discussion on thermogenesis, types of thermogenesis in humans, adaptive thermogenesis, different energy models, how it affects the humans over the period and how machine learning could help solve this problem. The chapter highlights the existence of adaptive thermogenesis, problem statement, and objectives of the research carried out during this novel research work.

Chapter 2: Deals with a literature survey of thermogenesis, adaptive thermogenesis, insulin resistance, machine learning techniques to understand clinical decision making. The chapter focuses on the research review findings that are necessary for the present research work.

Chapter 3: Presents the proposed methodology to identify the occurrence of adaptive thermogenesis using non-invasively trackable parameters from the human body. Feature identification, feature extraction, feature scaling and then machine learning models are proposed in this chapter. Findings are clearly explained with detailed graphs and references.

Chapter 4: Presents the possible explanation for early detection of adaptive thermogenesis in humans with the help of insulin resistance. Insulin resistance is extracted from various techniques with the available data in the CALERIE study dataset. A correlational analysis is done on insulin resistance with the rest of the non-invasive parameters identified. Finally, a bunch of machine learning techniques were applied to create a predictive model. This explanation and the observations are documented with detailed graphs and references.

Chapter 5: In this chapter synthetic data is generated by the model created by a neural network built by training on the existing CALERIE dataset. The models built in chapter 3 and chapter 4 are evaluated on their performance by providing the entire synthetic data for testing. All of this is explained in detail with graphs and references needed to understand.

Chapter 6: This chapter is my contribution towards diabetics who had other complications of covid-19 as this is formulated during one of the tough times in this generation. This chapter focuses on creating a fuzzy inference model which could generate data for various levels of covid-19 complications. Later the same

data is feed numerous machine learning models to spot the best possible model to identify which individual could face major health challenge due to covid-19.

Chapter 7: In this chapter the research work of this thesis is concluded with all the findings and observations.

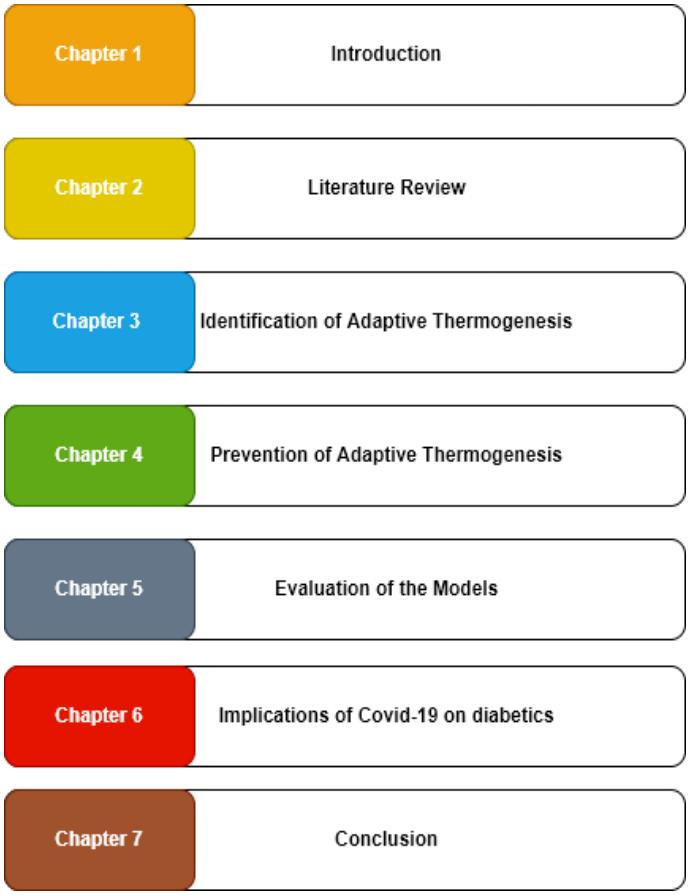


Figure 1.1 structure of the thesis

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

Before acquiring the knowledge of Adaptive thermogenesis, one needs to understand what is thermogenesis? Thermogenesis refers to the production of heat, especially in a human or animal body". As the human body needs to maintain a homeostatic state of body temperature 98.4 degrees F for its organs to work at their full potential. For that body needs to adapt to its current outer environmental changes as well as meet internal energy requirements for its organs. This energy is generated from food consumed by chewing, digesting, and absorbing the nutrients of the food, this entire process is called metabolism.

So, for a healthy human body to exist it needs to perform thermogenesis. Hence, thermogenesis is energy expenditure induced by different means of actions constituted daily called Total Daily Energy expenditure TDEE [2]. TDEE is the sum of Resting Energy Expenditure (REE) (energy spent while resting) and Non-Resting Energy Expenditure (nREE) (energy spent while performing some actions). Therefore, there are different types of thermogenesis as shown in the figure below. There are different types of thermogenesis. The biggest and the most impactful part of it is REE which contributes about 70% of the energy spending [3]. REE is the combination of both Resting Metabolic Rate (RMR) or Basal Metabolic Rate (BMR) and Adaptive Thermogenesis (AT).

There are two primary components of daily energy expenditure. The basal metabolic rate (BMR) refers to the energy used for normal body functions when a person is at rest. Approximately 60% of daily energy use is accounted for by the BMR. The amount of food eaten and other activities, such as exercise and digestion, determine the amount of calories used in a day. There are different ways of calculating RMR/BMR with formulae suggested by many researchers using the inputs like age, gender, weight, height, and then there is another one framed out of gender and fat-free mass (FFM).

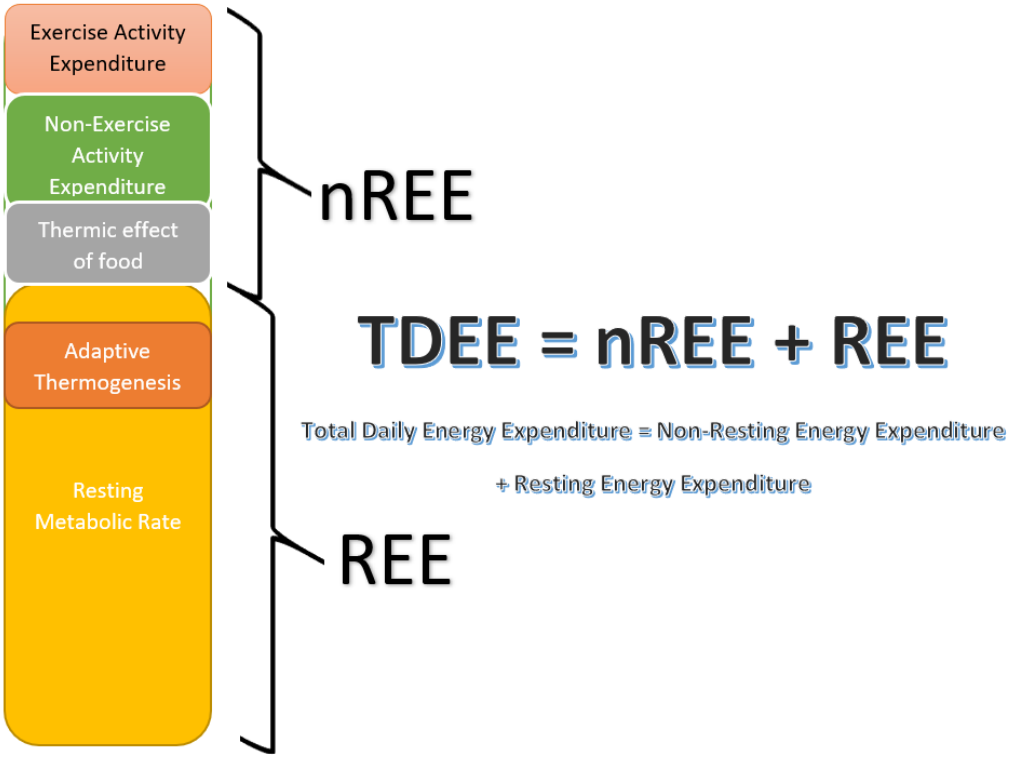


Figure 2.1 Total daily Energy Expenditure

2.2. DIFFERENT TYPES OF THERMOGENESIS

Thermic Effect of Food (TEF), Non-Exercise Activity Thermogenesis (NEAT), and Exercise Activity Thermogenesis (EAT) are then combined to form nREE. The energy used in digestion after eating is measured by the thermic effect of food (TEF), also known as diet-induced thermogenesis (DIT) or specific dynamic action (SDA). The TEF is affected by various factors, including food type, diet composition, and diet volume. TEF can be measured using additional calories as well as through the use of direct calorimetry, indirect calorimetry, and isotope tracers. The DIT response is composed primarily of fat (0–3%), carbs (5–10%), protein (20–30%), and alcohol (10–30%). Nearly 10% of daily energy use is made up of TEF. A recent study found that a high-protein diet led to a 100% increase in DIT activity and an associated increase in satiety.

[4] In examining sedentary individuals, researchers have identified a decrease of 2% in the basal metabolic rate (BMR) for every 1 kg lost. Additionally, the thermic effect of food (TEF) has been shown to contribute slightly more than 4% to total daily energy expenditure (TDEE). The results of this study show that increased non-exercise activity thermogenesis significantly contributes to

the rise in the resting metabolic rate (RMR) and TDEE, while moderate weight loss also affects increases in NEAT through a reduction in BMR. Furthermore, because modest decreases in physical activity might be offset by tiny increases in that activity, changes in NEAT during an energy deficit are difficult to measure. However, maintaining NEAT might be necessary to guarantee good weight loss maintenance.

Exercise activity thermogenesis (EAT) is the last but significantly less stressed topic as it varies from age, gender. Hence practices for anaerobic activities are less observed rather than aerobic activities where a lot of oxygen is taken in to perform the activity like walking, jogging, staircase ascending and descending, etc., But aerobic activities comes under Non-Exercise Activity Thermogenesis (NEAT) though EAT has its importance of about 10% of TDEE the odds of practicing this activity is relatively least compared to other activities.

A state of energy deprivation is brought on by weight reduction, and this state decreases the metabolic rate to lower energy consumption. However, weight loss also induces various adaptive responses that hinder the ability to maintain weight loss (a.k.a Adaptive Thermogenesis). The metabolism of white and brown adipose tissue, skeletal muscle, and the digestive system can all be linked to these reactions. As white adipose tissue (WAT) is the result of storing excess calories in the form fat in the human body especially in the liver and brown adipose tissue (BAT) which is the heat generator by burning these fats and usually brown in color. Inorder to encourage thermogenesis or raise RMR one has to activate BAT cells. Hence cold induced thermogenesis is part of RMR as they help increase or maintain the body's optimal temperature for better functionality. Later, a study that examined the relationship between energy usage and body composition in overweight and obese people focused on energy usage based on physical activity [49].

2.3 ADAPTIVE THERMOGENESIS

In response to a reduction in energy intake, adaptive thermogenesis is the term for a decrease in energy expenditure that is more than that which might be predicted from body weight or its components (fat-free mass and fat mass) under conditions of standardised physical activity. It is a rebound algorithm run by the human brain to maintain the healthy body weight based on the hormonal inputs given by the organs with normal metabolic rate orchestrated by burning brown adipose tissue through Diet Induced Thermogenesis (DIT) and Cold Induced Thermogenesis (CIT) further followed by Shivering and Non-shivering Thermogenesis. Which do play a vital role in Non-Resting Energy Expenditure (nREE) also. There are different models proposed for energy homeostasis for the maintenance of reduced body weight.

[5] Models of changes in EE during maintenance of reduced body weight are:

A. Mechanical Model: As per mechanical model after weight loss, the body's energy expenditure decreases proportionally to the amount of energy stores lost, mostly fat.

B. Threshold Model: A threshold model of adaptive thermogenesis posits that when body fat falls below a certain level, metabolism slows, but further loss of weight does not trigger more slowing of metabolism.

C. Spring loading Model: The degree of adaptive thermogenesis increases as people lose weight, and the strength of adaptive thermogenesis depends on the amount of weight lost. In other words, someone who loses 50 kilograms will experience greater degrees of adaptive thermogenesis than someone who loses 10 kilograms. The relative strength of adaptive thermogenesis will decrease as energy stores decline. In the spring-loading model, the level of adaptive thermogenesis depends on how much of a decrease in available energy stores induces adaptive thermogenesis, analogous to Hooke's law, which describes the amount of force a spring exerts as a function of its length and tension. Which is

explained in equation no 1 where T is AT, x is the change in weight after weight loss and k is the factor which helps in understanding the stage of adaptive thermogenesis and usually varies from individual to individual.

$$T = k*x \Rightarrow (k = ?)$$ ----------Equation 1

Adaptive thermogenesis is a form of energy/glucose homeostasis that occurs in humans Veen [103] explains the logical control system which is proportional-integral (PI) based control model. A Dynamic model to handle glucose homeostasis is proposed specially for patients for virtually monitoring [102]. This model could benefit monitoring adaptive thermogenesis. This phenomenon being deeply inter-dependent [101] proposed a predictive distribution functions for scenarios which are dynamic decision dependent.

2.4 INSULIN RESISTANCE

Insulin resistance causes blood levels of insulin to increase and results in weight gain because it weakens insulin receptors. This cycle of fat loss increases before the insulin receptors start to respond to blood glucose and insulin levels. Contrarily, hyperglycaemia happens when this is not the case and blood glucose levels increase [3]. Due to a persistent imbalance between insulin demand (increase in blood glucose levels) and insulin production, glycemic levels develop to levels compatible with T2DM. Even if there are numerous life risk factors associated with insulin resistance, it can be difficult to self-test every day without a doctor's supervision [4]–[7]. Many types of research have clearly shown that insulin resistance is the primary sign of weight gain even in thyroid deficient individuals where the thyroid is affected due to inactivity of insulin towards blood sugar levels [52].

One could explain why insulin has a better correlation with an increase in body fat and overall body weight can be explained by leptin and ghrelin hormones as shown in figures 10-11. An eating disorder emerges as the individuals lose their ability to act upon fat reserves as leptin is related to the amount of fat deposits over the body maintaining satiety. Ghrelin helps in letting the individual know to stop consuming food since he/she has enough reserves of energy in the form

of fat. Figure 11 describes how hunger and files are controlled by leptin and ghrelin followed by insulin. But due to chronic poor lifestyle leptin tends to not react upon fat and this phenomenon is called leptin resistance. As shown in figure 11 there is a clear sign of the correlation between leptin and insulin which is why insulin resistance has opted as a disorder since leptin is a secondary hormone and availability is this data is very scarce whereas insulin tests are often performed in every clinical trial.

Numerous prediction models for insulin resistance and, consequently, T2DM, have been put out during the past decade using ML algorithms. These machine learning methods typically fall under the categories of regression and classification algorithms. Some of the most popular techniques are decision-making algorithms, which vary from the newest technologies to linear regression techniques and are aimed to generate a statistical prediction for a problem. Artificial Neural Networks (ANN) [8]-[21], [44]. Taking into account the growing number and complexity of the data, deep learning (DL) has also been applied. [22]-[26]. The clinical way of identifying insulin resistance in humans is carried out in 3 possible ways, them being the ratio between Triglycerides and HDL-c (TGs: HDL-c), Homa-IR, and serum c-peptide levels. Every technique has its range charts. [52]-[56]

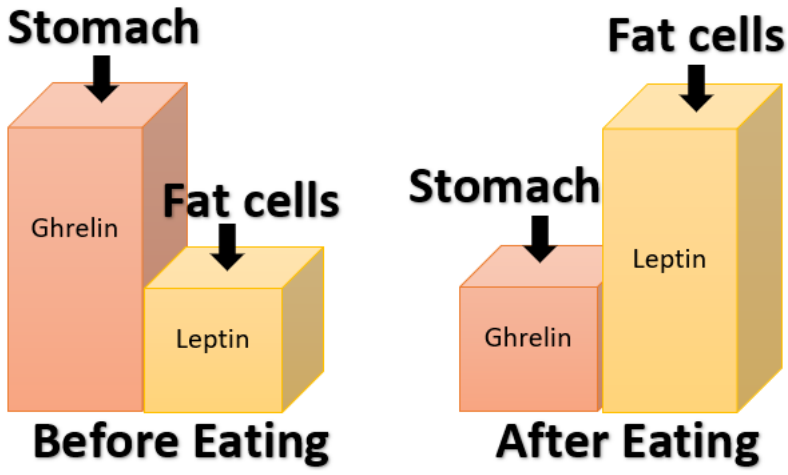


Figure 2.2 Leptin and ghrelin

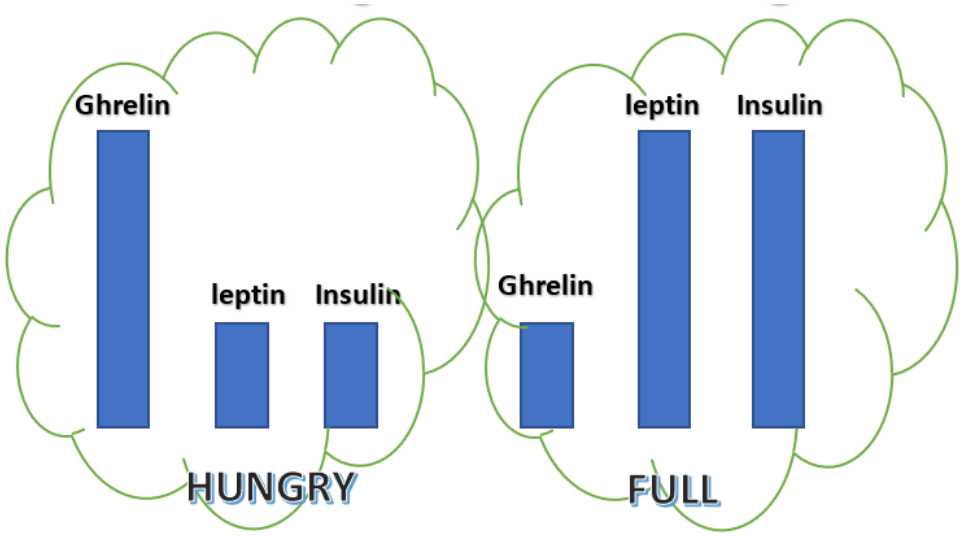


Figure 2.3 Imbalance in satiety

Numerous ML and DL-based combination techniques have also been put forth [27]-[31]. Over the PIMA Indian dataset, Kandhasamy et al. [8] applied a variety of classifiers, including KNN, SVM, etc. The dataset is validated using a 5-fold cross-validation (CV) procedure. The accuracy rate with KNN and Random Forest has been demonstrated to be 100% after data pre-processing, whereas the accuracy rate with J48 classifier without pre-processing is 74%. For the purpose of predicting diabetes, Tafa et al. [9] upgraded the Naive Bayes and SVM classifier model by integrating both of them. In the data with 402 individuals, 80 had type 2 diabetes, and eight factors were taken into consideration.

The proposed integrated method offers an accuracy of 97.6% when compared to the Naive Bayes accuracy and the SVM accuracy, according to the comparison. Six classifiers were used by Mercaldo et al. [10]: JRiP, BayesNet, RandomForest, J48, Hoeffding Tree, and Multilayer Perceptron. The PIMA Indian dataset is utilised. A 10-fold CV is applied to the dataset. The four characteristics have been determined to be age, diabetes pedigree function, BMI, and plasma glucose concentration. It is shown that the Hoeffding Tree method has produced the best results with precision scores of 0.757, F-measure of 0.759, and recall scores of 0.762.

TyG-er, an ML strategy, has been proposed by Michele Bernardini et al. [12] The dataset from the Italian Federation of General Practitioners is used.

Leukocytes, uricemia, and other uncommon clinical variables were identified. Patients with T2DM are excluded, while those in normal to high-risk situations are included. In [32]–[36], a similar study that made use of statistical analysis was also carried out. Konrad et al. [37] have published a machine learning-based technique for assessing insulin resistance in young people with type 1 diabetes. A total of 315 patients, ranging in age from 7.6 to 19.7, participated in the study. Byoung et al. [38] have suggested T2DM prediction models using machine learning techniques with an EMR dataset. The study included 8454 individuals who underwent treatment at a cardiology facility and had no history of diabetes after five years of follow-up. The linear regression model demonstrated the greatest prediction performance among several predictive models, including LR, LDA, KNN, etc. with an AUC of 0.78.

Most studies only used one data set, while a few studies, like [11], additionally employed two datasets for prediction. The Pima Indians dataset and Diabetes 130-US have been combined into one dataset. It is shown that using a pooled dataset can enhance diabetes prediction with an AUC of 0.72 after subjecting the data set to a 10-fold CV.

2.5 EXISTING TECHNIQUES TO MEASURE INSULIN RESISTANCE

The vast majority of methods used to identify insulin resistance are invasive. Very few solutions have emphasised simple, rapid, and inexpensive non-invasive techniques. The accuracy rate, however, is not especially alluring. For more precise type 2 diabetes and insulin resistance prediction, additional study into these non-invasive approaches is needed [39], [57].

In addition to the concept of non-invasively tracking the data, a continuous glucose monitor has long been a device that is accessible. It is an efficient method for keeping track of the rate of glucose absorption, which enables one to gauge the level of insulin resistance. The disadvantage is that it requires Bluetooth to transfer all of the data to a smartphone and that it will stay fastened to the patient's arm for around 10–4 days [45]. A growing body of research is being done on retinal microperimetry; a new technique that can only be used non-invasively [46]. DNA testing can reveal polygenic traits at the genome

level, which is a different non-invasively traceable method. Some studies have used these polygenic risk scores to effectively predict a person's likelihood of developing diabetes types I and II in the future [47]–[48]. Homa-IR level is a fantastic indicator to use because it persists in the bloodstream for a very long time, according to recent research. However, the examination necessitates a clinical strategy. [92] Because of advancements in the fields of genetic algorithms, machine learning, and artificial intelligence (AI), researchers are working to remove these clinical barriers and lessen the discomfort associated with taking a person's blood with a needle.

2.6 TRENDS OF HEALTH AND FITNESS BASED ALGORITHMS

A human body achieves homeostasis when the energy consumed is dissipated in one form or the other. Thanks to the advancements in the field of health and wellness with the help of wearable technology with the help of MEMS technology smaller and more accurate sensors are being designed. Hence began the 2nd wave of ICT revolution in south Asian countries [93]. Better algorithms are designed according to the sensors. Which helped researchers and developers to track the human activities which are used to calculate the calorie count precisely. Hence one component of TDEE is quantitatively instrumented. This helps researchers and developers bring pervasive tracking and enhanced user experience in the field of health and wellness. In the current discussion a literature is being reviewed for the advancement in the field of calorie shredding with the help of activity tracking based on the individual's physiological parameters. A similar study is being done in the next sub topic 2.7 where the literature review of energy intake is being focused.

Several techniques for measuring physical activity and energy expenditure [59] are shown in the table below:

Table 2.1 literature review on physical activity and energy expenditure measurement

s.no	Methods	Advantages	Disadvantages
1	DLW (doubly labelled water) [60]	● Gold-standard technique for measuring TEE that is extremely exact. ● Gives participants flexibility to engage in any activity.	● The approach is rather pricey (including the high price of DLW and expensive equipment for analysis). ● The personnel must possess expertise. ● The approach doesn't offer any detailed information on physical activity.
2	Direct calorimetry [61]	● It is the most precise way to measure metabolic rate.	● The method's high expense. ● 24-hour or longer subject confinement is necessary.
3	Indirect calorimetry [62]	● Accurate and non-invasive method. ● Offers details on the metabolic fuels being burned.	● Relatively high cost. ● Correct usage of the method requires trained people.
4	Accelerometry [63]	● Measurement of physical activity that is objective, non-invasive, and	● Predictive equations are inaccurate in converting activity counts into energy expenditure,

		less onerous for subjects. ● Applicable in both field and laboratory settings. ● Relatively inexpensive.	especially when applied to a wide variety of activities.
5	Heart rate monitor [64]	● Relatively inexpensive; Non-invasive; Versatile method; Can be used both in controlled settings and in free living conditions. Objective tool for the measurement of physical activity and energy expenditure.	● Measurements of sedentary and light activity are inaccurate, because ubiquitous electrical gadgets can cause electrical or magnetic interference.
6	Pedometry [65]	● Cost-effective and non-intrusive technique used to evaluate the most prevalent activity (walking). ● Can encourage people to continue engaging in physical activities.	● Is only accurate for tracking walking activity; Is inaccurate for calculating distance travelled and energy expenditure.
7	Self-report methods [66]	● Low burden on subjects; low expense, allowing application in	● Low dependability and accuracy, particularly related to their reliance on participant memory

		investigations with high sample sizes. ● Disclose the frequency of physical exercise.	

2.7 CURRENT RESEARCH ON DIETARY ASSESSMENT

Adaptive thermogenesis is a form of energy expenditure which can be tracked of energy intake as well as energy expenditure. Where energy expenditure can be done through physical activity monitoring at the same time BMR/RMR and energy intake are measured through dietary assessment as mentioned in the following literature review.

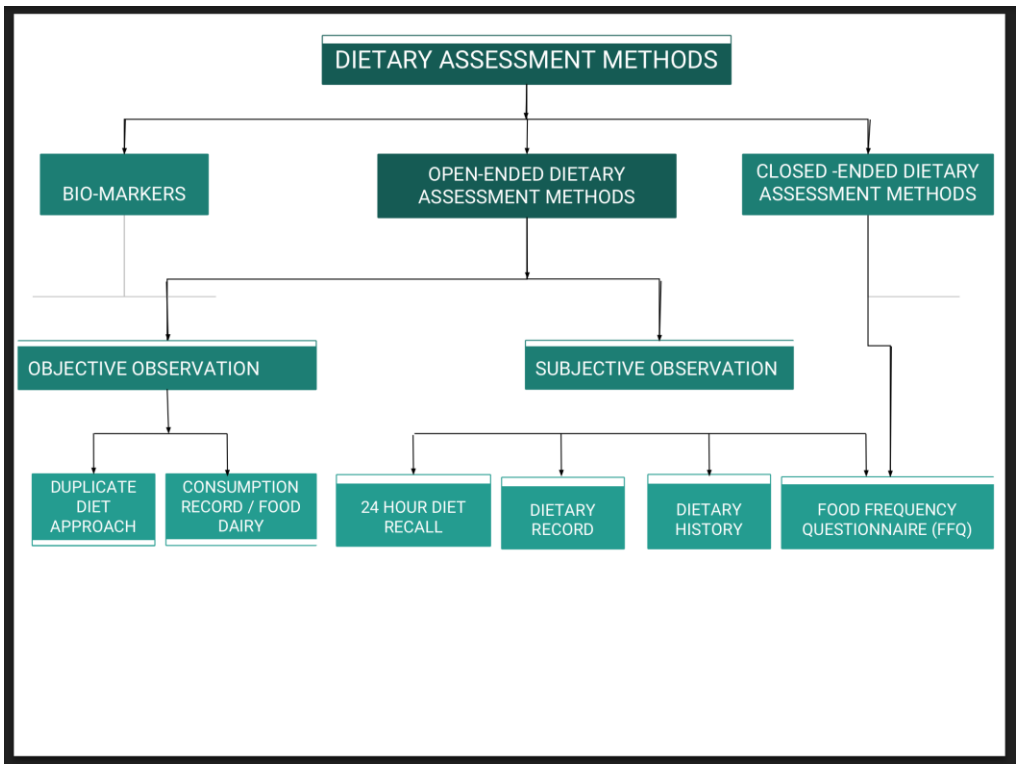


Figure 2.4 Dietary Assessment Methods

Table 2.2 literature review on dietary assessment methods

s.no	year	Device	sensors	process	pros	cons
1	2014	Digital photography of foods methods	Digital video cameras	1. Tracks the quantity and quality of food put on the plate and at the sametime, tracks the leftovers through images captured at public environments or self captured like RFPM (Remote Food Photography Method) 2. Their nutritional content is extracted from FNDDS [68,69]	1. This method is accurate in measuring the user's caloric intake when present in their natural living environment. 2. Very good adaptation technique for users undergoing various intervention and nursing programmes. 3. To avoid forgetting to capture images there are certain alerts based on the user's actions[70].	1. It would be difficult to calculate the calorie intake for a variety of foods with unique cooking styles and seasonings. 2. Lighting conditions for pictures which are captured from these cameras play an important role.
2	2014	AutoDietary [73]	Acoustic sensors around the throat like a necklace associated with a smartphone / tablet via bluetooth.	1. Identifies the sounds of chewing and swallowing. 2. [71] The amount of food is estimated using the CCS model. 3. Energy per meal is estimated by the amount of food chewed and swallowed	1. Good at classifying solid and liquid foods. 2. [72] Extracted 30 features over this input.	1. Accuracy is influenced by the disturbances of the collar (shirt,etc.) and low BMI users. 2. External sounds affect user accuracy of the device.
3	2014	Piezoelectric sensor based necklace	Piezoelectric sensor, accelerometer (added	1. Sports collar and [73] pendant based designs are proposed to	1. It's a low cost solution to monitor food intake.	1. Does not give a caloric quantity of food consumed. 2. Design flaws

			improvement), bluetooth, smartphone / tablet	track mechanical stress created at lower trachea. 2. Activity recognition is done with the help of an accelerometer.		and more false positives due to the lack of an accelerometer. 3. Fluctuations in signal due to bodily actions.
4	2014	E-button[74]	2 wide-angle cameras (stereo and depth), UV sensor, IMU, Wi-Fi, bluetooth, GPS, audio processor, proximity sensors.	1. It can track both physical activity and dietary intake. It captures images and sends them over to a smartphone for analysis. 2. It uses FNDDS as database for the food it is tracking.[74]	1. Compared to other techniques this method seems to be better in measurement of food intake. 2. Completely automatic as it can be helpful for blind people for support.	1. Depends on the lighting conditions and camera viewing angle. 2. [75]Tracking Complex shaped food items is still a challenge. 3. Significant amount of price to contain such technology into a button.
5	2016	Glasses wearable [77],[76]	Piezoelectric strain sensor, accelerometer	1. Temporalis muscle activity is recorded using a piezoelectric strain sensor and movements through an accelerometer then sent over to a smartphone using bluetooth. 2. Through single and multi-stage classification utilising linear SVM and decision trees, this	1. This is the most accurate measuring wearable device as the author states with almost 99% precision in detecting activity with multi stage classification. 2. Lowest power consuming device as the components are either passive sensors or MEMS based with BLE. 3. Few other advancements in the fields of carbon	1. User preferences in wearing glasses all the time. 2. It is challenging to maintain constant contact with the temporalis muscle at the back of the ear, where piezoelectric devices are supposed to be placed. 3. [90] IoT based deep learning could help the

				wearable can identify eating, walking, eating when resting, and eating while moving.	nanotubes could bridge for more pervasive tracking [88].	researchers bring in more valuable data.
6	2016	Bite-counter	Tri-axial accelero meter, gyroscop e worn on wrist.	1. Analysis of wrist torsion movement while the user picks up food from plate to his mouth. [78],[79],[81],[82] 2. An algorithm is proposed for Counts number of bites based on which an equation calculates calories consumed [79],[80]	1. Comfortable design for users. 2. Can identify other gestures that can also be mapped for other applications. 3. Survey [79] showed better results than human based estimation.	1. Can only predict precisely only when a user consumes food from his hands. 2. over estimates when eaten with fork, knife and spoon. 3. It allows an 8 seconds gap between 2 bites.
7	2017	AIM (Automatic Ingestion Monitor)[83]	Jaw motion sensor, hand gesture sensor, accelero meter,	1. This wearable is used to know the eating episode duration and food ingestion for micro structuring of meals.[84]	1. Accuracy of this wearable is almost similar to the actual self-reported duration or by push button and it is better than any food diaries.	1. There are a lot of seperate devices connected to make one system. Which makes the reliability of the system questionable.
8	2017	GlasSense [85]	Load cells, IMU	1. Load cell is placed at the hinge of the glass which detects the temporalis muscle contractions and relaxations. 2. Along with an IMU a pattern recognition	1. Almost 90% plus precise in detecting the actions promised. 2. One of the most economical ways to measure temporalis muscle activity. 3. A literature	1. There are bodily actions such as walking, sitting, running which are still not implemented. 2. It is being observed that there is a lot of noise from sensors based

				algorithm (SVM) helped researchers to classify head movements, chewing, talking and winking.	review on organic thin film transistor can make the tracking more and more invasive and accurate [89].	on the surroundings. 3. Not everyone is willing to wear glasses.
9	2017	EarBit [86]	IMU's, microphone, proximity sensor	1. This wearable is designed to detect the food-intake episodes in unconstrained environments. 2. And it is observed to be 90% accurate in detecting the events using a machine learning model.	1. One of the best devices to detect actions like talking, chewing, idle, eating, walking. 2. Deep learning with the data collected by wireless sensors on the body with the help of edge computing can help the sensors adapt the users [91].	1. Participants in the observation study didn't like the earphone styling as they felt them piercing and itchy. 2. Few didn't prefer to wear them while eating.

CHAPTER 3: IDENTIFICATION OF ADAPTIVE THERMOGENESIS

3.1 INTRODUCTION TO RMR

Adaptive Thermogenesis is one of the most complex biological systems known. A complex network with multiple degrees of freedom makes the problem hard, but what makes it even harder is that adaptive metabolism depends not only on external temperature but also on the food eaten. Adaptive metabolism is not just inference within a single model, but simultaneous inference in spatially distributed models (multiple regions of the body). But as discussed earlier in chapter 1 that most of the energy expenditure in humans happens as RMR or BMR. Co-incidentally AT is the derived element of these parameters RMR/BMR as shown in equation 2. Though there is not more than 10-15% difference between RMR and BMR hence that additional factor is added later for RMR data.

AT = Predicted RMR – Measured RMR---------(2)

Here, measured RMR is extracted from CALERIE study dataset whereas measured RMR is calculated using BMR equations based on the following equations.

- `Men = 66.47 + (13.75 * weight [kg]) + (5.003 * size [cm]) – (6.755 * age [years])`
- `Women = 655.1 + (9.563 * weight [kg]) + (1.85 * size [cm]) – (4.676 * age [years])`

And generally, RMR is considered as 1.2 times of BMR which is averaged over a sample of a wide range of ethnic groups, sexes as well as ages. The current understanding of adaptive thermogenesis was acquired from Alexandra R. Martin's research helped decode the clinical side of the work [87].

3.2 CHALLENGES IN IDENTIFICATION OF AT

- AT had to be calculated since it has not been focused much in this clinical trial.
- There were a lot of missing fat mass details in the dataset.
- There are a lot of human errors following the continuation while data acquisition especially in body composition analysis documents.

3.2.1 DATA PRE-PROCESSING

- Identification of Categorical parameters
- Converting few categorical parameters into continuous parameters
- Fill missing parameters for DXA scan report for fat mass and fat-free mass data using another regression model.

3.2.2 HANDLING MISSING FAT MASS DATA

Fat mass data can’t be imputed with averages or mean of earlier data since it is purely individual and must maintain a relationship with other bodily parameters. Therefore, a model is supposed to be built since fat is a continuous parameter so a regression model is used to impute the rest of the fat mass and fat-free mass data.

3.3 METHODOLOGY FAT MASS IMPUTER MODEL

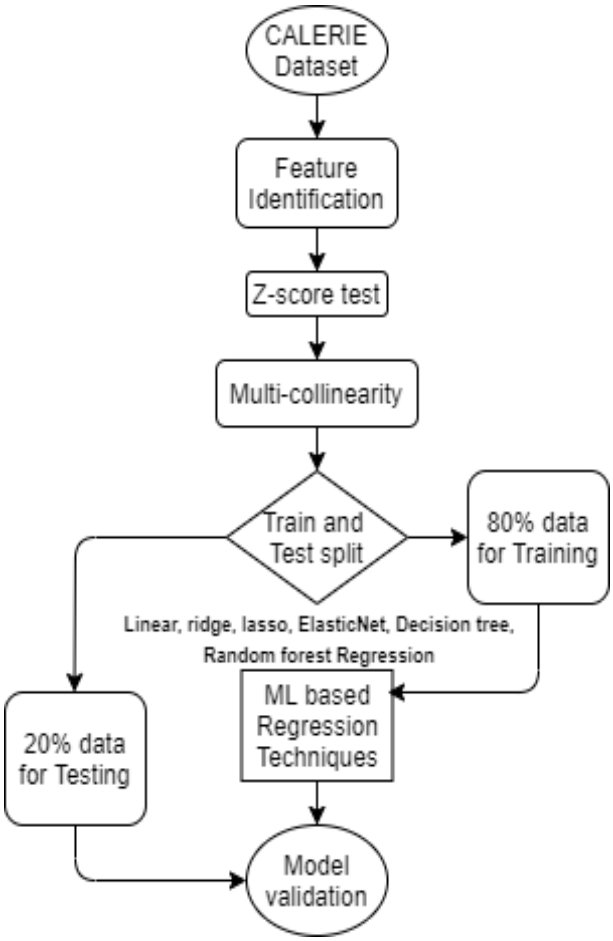


Figure 3.1 Methodology

Feature identification was done using the Recursive Feature Elimination (RFE) technique. RFE allows tries all the possible inputs individually and collectively to find a recursive solution for the best possible features as shown in figure 5 below and if observed a total of 18 features were selected out of which after 6 features there isn't any significant difference in the cross-validation scores of the model.

Above figure 3.2 shows the interrelationships of the parameters from table no 3.1 for the use case of the fat mass estimation model. Similar to this, a multi-collinearity test can reduce the unnecessary computation time and increase the model's efficacy. As a consequence, Pearson's correlation and z-score values are computed for each parameter and shown in figure 3.2 as a heatmap. Where the strongest association is believed to exist between the whitest portion and the lightest portion. Each cell in the heatmap will display its correlation score if it

is thoroughly inspected. The element that is most important for understanding the multi-collinearity problem is the output correlation, though. By picking the adjacent column or row with its matching scores next to it, you may determine the relationship between each parameter and the outcome, which is fat mass. Since the correlation values vary from 0 to 1, a score of 0.7 is desirable and is thought to be highly connected. But nothing came close to 0.7, with the exception of BMI (0.66). As a result, all input parameters are given consideration in order to comprehend their behaviour during the machine learning model's training. One of the finest approaches to comprehend feature importance and choose the proper characteristics for the experiment is using this method.

Table 3.1 Input parameters identified for the fat mass estimation model

Terms	Parameter
WhtR	Waist to height ratio
meanumb	Mean waist size over umbilical cord
BMI	Body Mass Index
clinwt	Body weight
GENDER	sex
height	Body Height
TEE	Total Energy Expenditure
RMR	Resting metabolic rate
PA	Physical activity level
AT	Residual RMR (Adaptive Thermogenesis)
EI	Energy Index
tgrams	Food consumption in grams
tfat	Fat consumption in grams
tcarb	Carbohydrates in grams
tprot	Protein in grams
aniprot	Animal protein in grams
vegprot	Vegetable protein in grams
alcohol	Alcohol quantity
Output	Fat mass based on DXA

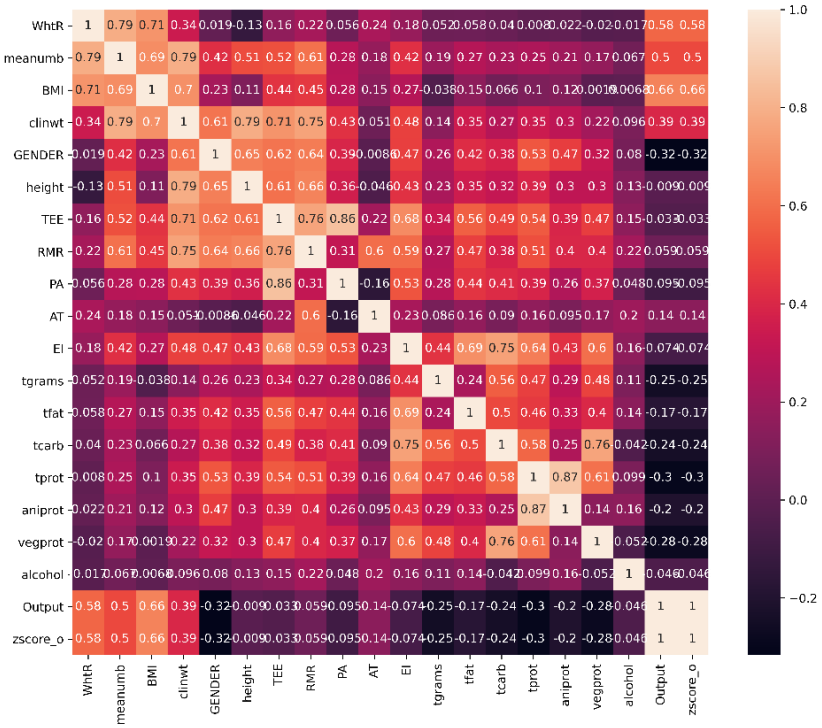


Figure 3.2 Correlational heatmap of the dataset

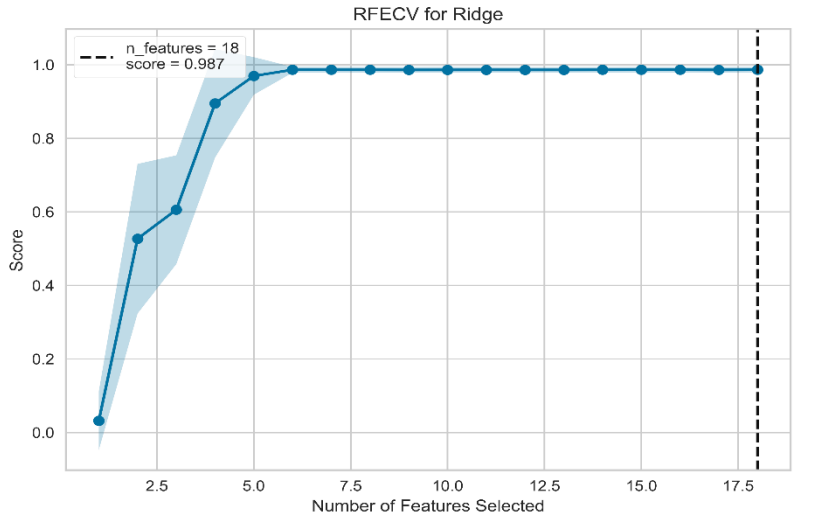


Figure 3.3 Recursive Feature Elimination (RFE)

As discussed in figure 3.3 RFE 6 features provided the best possible result out of 18 features. To find the 6 features, a feature importance plot is drawn as

shown in figure 3.4 which reveals the variable importance of the respective features on the y-axis. In which body weight, female gender, AT, RMR, protein consumption has shown importance whereas body weight and female gender showed significant importance. And 6th feature is picked either animal or veg protein.

An Outlier test is done on the overall data to see if the data has any underlying anomalies which prevent it from following a pattern. In this case, there isn't any such anomaly concerning the fat mass parameter which is done using z-test as shown in figure 3.5. Hence there are no outliers for this dataset therefore RSME is a good fit as a predictor and validator of the model. Also, the multi-collinearity test resulted negative which means the data is good to build a model.

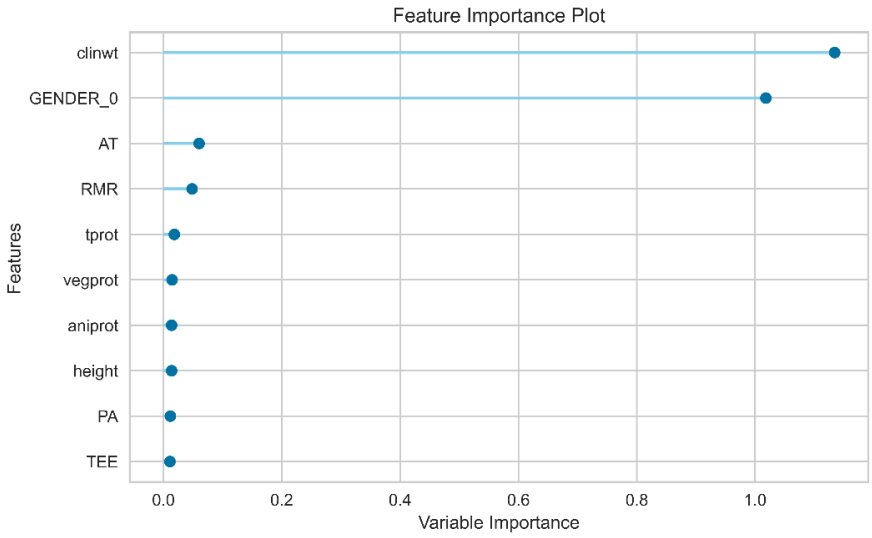


Figure 3.4 Feature importance plot

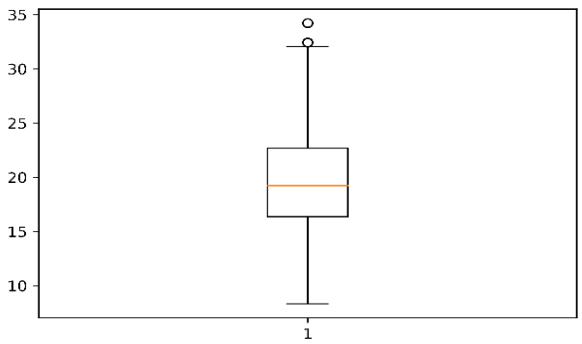


Figure 3.5 Z-score Test

Since the output is continuous, regression analysis should be used rather than categorization, which would produce discrete results. Since it is a regression problem, the basic linear regression model is the first approach to try, followed by variations like a lasso, ridge, and elastic net. In an effort to create better models, a decision tree regressor and an ensemble model akin to a random forest regressor were also tested. Table 3.4 lists the outcomes for each of the models mentioned above. Since using a linear regression model as a starting point is a no-brainer, this model performed admirably, as seen by its r-square score of 0.91 and cross-validated score of 0.988. In order to enhance this model, hyper-parameters for the linear regression model can be helpful. Although there are many hyper-parameters in linear regression models, some models developed from linear regression, including the ridge, lasso, and elastic net, use grid search CV as an optimization method. Elastic net regression model improves L1 and L2, ridge regression model improves L1 penalty, while lasso regression model improves L2 penalty. Each technique's parameter modifies a unique hyper-parameter. To validate those methods for the data, additional ensemble regressor and decision tree approaches were used. Table 3.2's results make it very evident that the Lasso regression modelling strategy has an advantage over other methods.

Table 3.2 Results of the regression analysis

Model	R-square	Cross-validation(mean)	Hyper-parameters	Optimization Technique
Linear Regression	0.91	0.988	-	-
Lasso Regression	0.97	0.962	L1 penalty	Gridsearch CV
Ridge Regression	0.986	0.978	L2 penalty	Gridsearch CV
ElasticNet Regression	0.867	0.79	L1 and L2 penalty	Gridsearch CV
Decision Tree Regressor	0.48	0.52	-	-

Random Forest Regressor	0.80	0.71	-	-

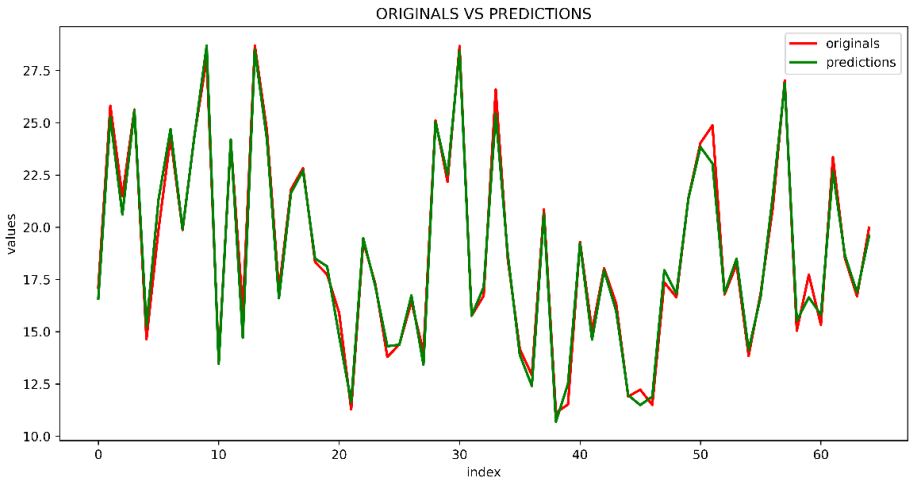


Figure 3.6 Ridge Regression

After the models have been developed, they are tested against the available data. Graphs are created to compare the original and predicted values of fat mass in order to visually understand the difference. This continuous graph displays the number of samples in the same order as those used for the model's prediction on the x-axis and the estimated fat mass on the y-axis. The no free lunch theorem states that no single algorithm can fit all issues while yet offering superior models. Because this problem contains more than two aspects, multivariate regression analysis must be used, the observer must start with the basic regression analysis. Starting with the linear regression model or OLS (ordinary least squares) method, the r-square value is calculated, which serves as a gauge of accuracy for regression-related difficulties. It does not follow that the model accurately predicts the outcome just because the r-squared value is higher. As a result, the dataset is split into train and test data before the model is developed, and the model's cross-validation test is then run on the model's unobserved data. The model is constructed using train data, and the output is cross-validated using test data. Plotting an original vs. forecast comparison chart is the greatest technique to cross-verify that cross-validation brought the exceptions or outliers

in the model. Numerous types of linear regression approaches are also utilised, in addition to decision trees and its ensemble counterpart, random forests, which were previously described.

3.4 MODEL FOR IDENTIFICATION OF ADAPTIVE THERMOGENESIS

In figure 3.7 is the correlational heatmap drawn against all the features along with a possible outlier test parameter called score. With the correlational heatmap, it should be obvious that BMI is highly correlated with adaptive thermogenesis outcome, along with waist to height ratio (WthR). The next parameter is mean waist size when measured over the umbilical cord and lastly, the body weight and sex of the individual showed some dependency compared to the rest other parameters. But total protein consumption showed an inverse relationship which should also be a dependable parameter according to Pearson's correlation.

Having identified these parameters bunch of regression-based machine learning algorithms were implemented to fit a model for adaptive thermogenesis. The results of this implementation can be seen in table 4 which are nowhere nearer to usable study, therefore a feature scaling would be suggestable based on the nature of the AT which is a continuous variable. This feature scaling method is explained in the coming section of feature scaling of AT just after the results table for AT and its highlights.

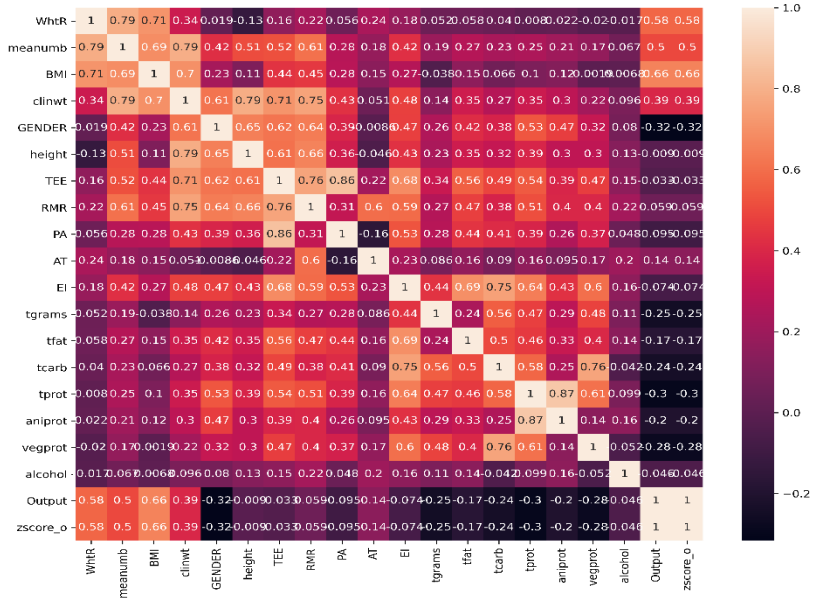


Figure 3.7 Correlational heatmap of features for Identification of AT

Table 3.3 Results of the regression analysis for AT

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Bayesian Ridge	88.1712	12699.8	110.6832	0.134	1.4561	2.8351
Elastic Net	87.5445	12716.17	111.2128	0.1136	1.4243	3.0603
Lasso Least Angle Regression	90.5831	13313.11	113.37	0.0995	1.63	1.8465
Lasso Regression	88.2838	12923.29	112.2009	0.0965	1.4249	3.0592
CatBoost Regressor	89.0506	13139.01	112.9782	0.0962	1.6327	3.0472
Extra Trees Regressor	90.1231	12914.26	112.3331	0.0962	1.6209	2.1601
Random Forest Regressor	89.7273	12913.15	112.4162	0.0956	1.5109	2.5539
AdaBoost Regressor	90.2995	13422.48	113.9426	0.0852	1.3387	2.575
Ridge Regression	89.1727	13113.72	113.055	0.0804	1.4539	3.0411
Linear Regression	89.8385	13336.11	113.9859	0.0635	1.4769	3.0427
Huber Regressor	94.4714	13819.4	116.1046	0.0239	1.416	3.2058
Least Angle Regression	91.962	13734.05	116.0432	0.0161	1.4799	3.4825
Orthogonal Matching Pursuit	96.274	14945.84	120.5176	-0.0268	1.5384	2.2719
Gradient Boosting Regressor	97.6204	14832.91	120.5589	-0.0425	1.472	3.1256
Light Gradient Boosting Machine	95.8494	15088.7	121.5515	-0.058	1.3896	3.4977
K Neighbors Regressor	99.6361	15017.46	120.7594	-0.059	1.6036	3.3874

- Massive underfitting is being observed due to low variance and high bias of the data while testing.
- The model is too simple to learn from too few parameters which contributes to fitting the dataset into a mathematical model.

3.5 FEATURE SCALING FOR AT

AT = Predicted RMR – Measured RMR

- IF AT <= -5% of RMR

  THEN AT == 1

  ELSE AT == 0

- Now, this becomes a Binary classification problem either 0 or 1.
- Therefore, a classifier is proposed to predict whether the individual is undergoing AT or not.

A lot of machine learning-based classification algorithms were run on the newly scaled variable of Adaptive Thermogenesis parameter which is the parameter of interest. Logistic regression is the first and foremost recommended algorithm because it is a binary classification problem (LR). The next best solution for classification problems is the decision tree (DT) classifier method. The latest state of the art of machine learning algorithm is called Explainable Boosting Machines (EBM). EBM produced a better result with more data provided for learning purposes whereas a linear regression algorithm could generalize the solution with fewer data provided for learning. As shown in figures 3.8, 3.11, and 3.13 below in figure 3.8 displays metrics like accuracy and F1-score for models built using the algorithms specified earlier with 80% of the data used for the training model and 20% used for testing. The results show EBM producing significantly better results compared to others with 78.4% accuracy and 73.9% F1 score.

Figure 23 follows the same steps, but the train and test size of the dataset varied from 70% training set to 30% testing set and a serious degradation can be observed with the EBM model as it is highly dependent on more data. Whereas LR improved as it was able to generalize better with lesser data whereas DT also improved significantly. Lastly, figure 3.13 shows a similar sequence of modelling but the train and test sizes varied to 60% and 40% of data subsequently. In this case, all the models degraded but LR yet provided a decent enough F1-score and accuracy.

Test size = 20%			
	LR	DT	EBM
Accuracy	0.7076	0.6153	0.7846
F1 Score	0.6771	0.5966	0.7397
C.M	0 T 1 | 33 | 9 | 10 | 13 | 0 P 1	0 T 1 | 27 | 15 | 10 | 13 | 0 P 1	0 T 1 | 39 | 3 | 11 | 12 | P 0 1

Figure 3.8 metrics for 20% test size

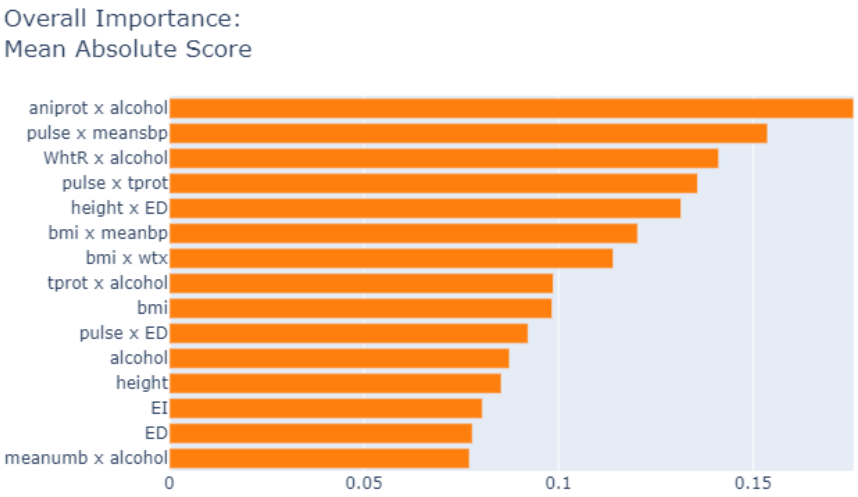


Figure 3.9 feature importance for EBM model

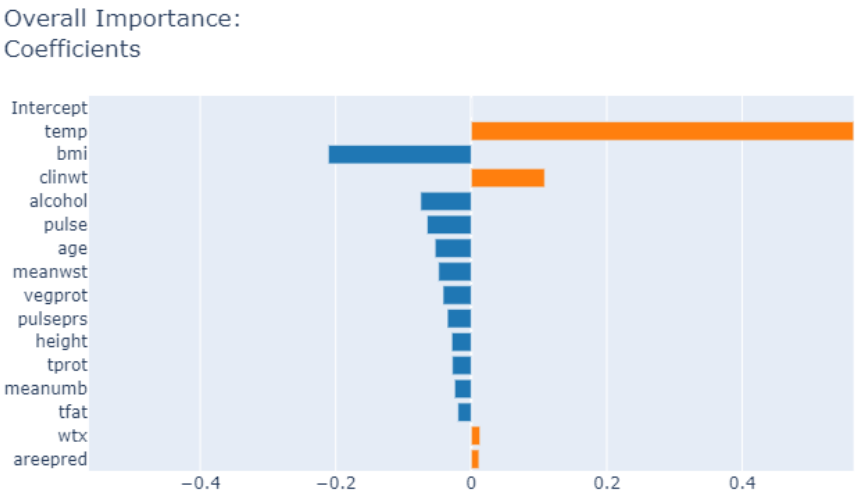


Figure 3.10 Feature importance for LR model

These results reveal feature temperature of the body has a significant dependency over other features extracted from the correlation heatmap from figure 3.7. Following this feature BMI, bodyweight which could be a cause for multicollinearity since BMI has a relationship with body weight. Alcohol provided another inverse relationship following pulse and age.

Test size= 30%			
	LR	DT	EBM
Accuracy	0.7319	0.690	0.680
F1 Score	0.709	0.623	0.629
C.M	0 T 1: 49, 14 / 12, 22 — 0 P 1	0 T 1: 54, 9 / 21, 13 — 0 P 1	0 T 1: 51, 12 / 19, 15 — 0 P 1

Figure 3.11 metrics for 30% test size

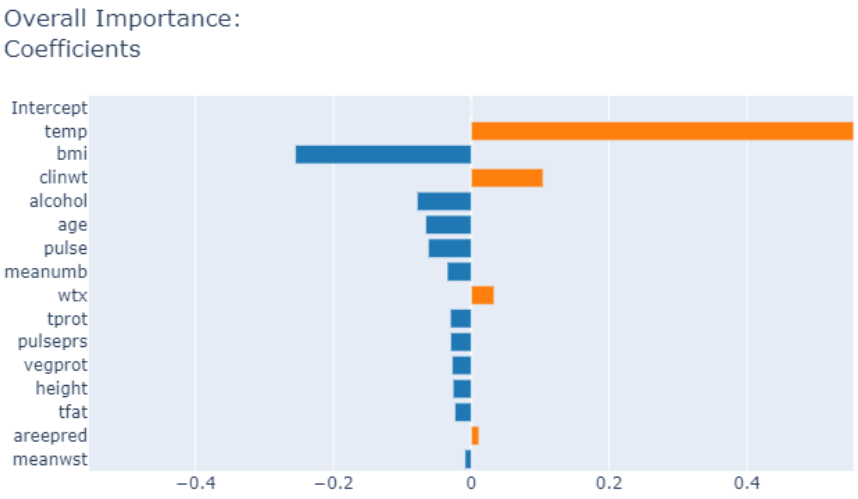


Figure 3.12 Feature importance for LR model

Test size = 40%			
	LR	**DT**	**EBM**
Accuracy	0.7131	0.6589	0.6976
F1 Score	0.6859	0.6116	0.6578
C.M	T 0: 65, 21 / 1: 16, 27 — P 0, 1	T 0: 65, 21 / 1: 23, 20 — P 0, 1	T 0: 67, 19 / 1: 20, 23 — P 0, 1

Figure 3.13 metrics for 40% test size

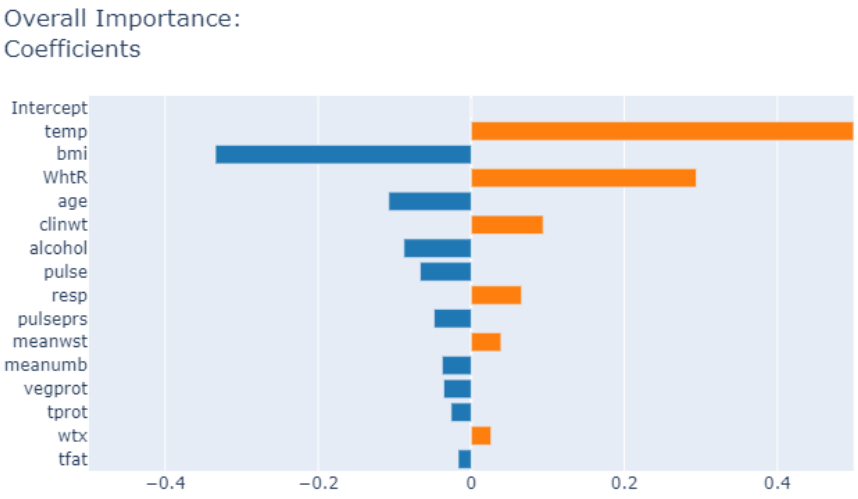


Figure 3.14 Feature importance for LR model

Since both the experiments with test sizes of 30% and 40% resulted in Logistic Regression (LR) as the most probable choice, the feature importance charts for only LR models for both the experiments. If observed in all the cases body temperature, BMI, Alcohol, pulse, and waist to height ratio (WhtR) showed up at the forefront of the importance levels.

And for model validation, a synthetic dataset is created using the original dataset to fit different models. And 30% test data model is opted to validate the model which can be observed in section 5.4 objective 4. Results provided from these models were satisfactory as the model for identification of Adaptive Thermogenesis (AT) in humans.

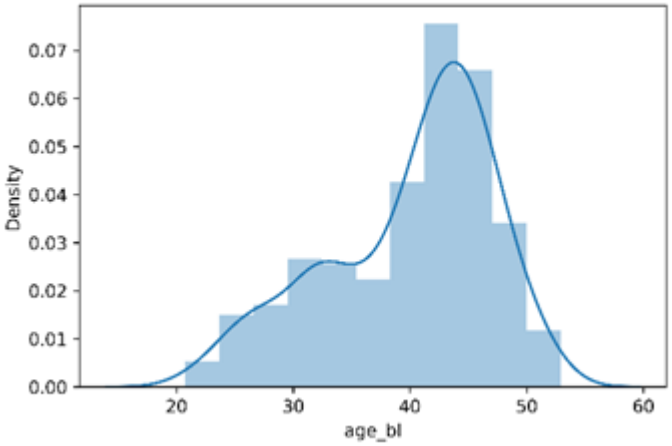
CHAPTER 4: PREVENTION OF ADAPTIVE THERMOGENESIS

As discussed earlier Adaptive thermogenesis is the phenomenon that affects the body weight to grow along with its composition analysis. It has been identified that excess mass in the body after weight loss is generally stored in fat adipose tissues in the liver and around the abdominal region. The storage of fat on the body is generally affected by inactive insulin which avoids using up available blood sugars. This syndrome is termed insulin resistance in medical science terminology. Insulin resistance slows down the body's ability to process glucose, which raises the amount of insulin produced and causes hyperinsulinemia. In 2014, 8.5% of people worldwide had Type 2 diabetes (T2DM), and by 2016, 1.6 million people had died as a result of the disease, causing tremendous losses in terms of human life as well as societal and economic costs. Diabetes prevalence is 9.3% globally as of 2019, and is projected to increase to 10.2% by 2030 and 10.9% by 2045 [1] [2].

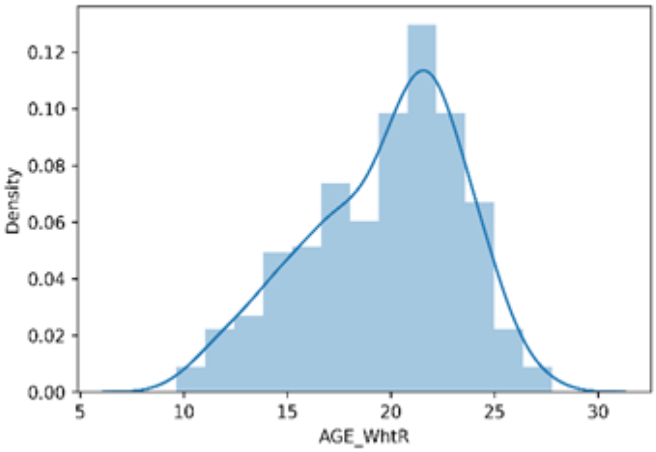
All the parameters shown in the table below can be non-invasively achievable. In this table, there are direct parameters like gender, age, etc as well as combinational parameters like BMI, WhtR, AGE/BMI, etc. All these parameters provide diversity to the model to understand the wide range of relationships that can be observed using correlational graphs. There are many categorical variables here like gender and a few others are converted based on their respective ranges set according to WHO for BMI, WhtR which decides whether the person is underweight, overweight, obese, or normal.

Table 4.1 Feature description

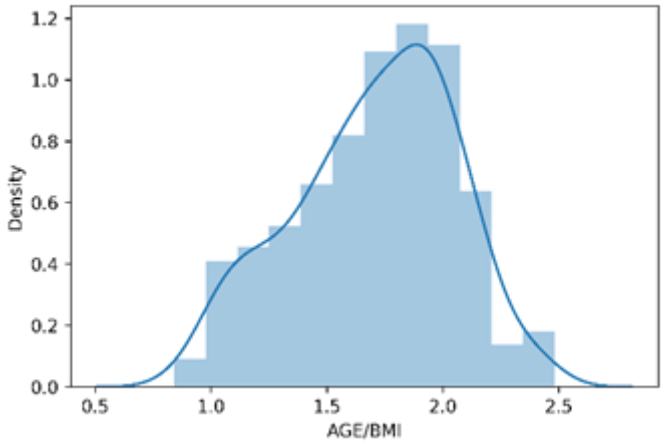
S.no	Feature	Feature Description
1	GENDER	Gender
2	age_bl	Age
3	fma	Fat mass
4	ffma	Fat free mass
5	1 clinwt	Clinical weight of the body
6	bmi	Body mass index
7	B meanwst	Mean waist circumference size
8	B meanumb	Mean waist circumference size around umbilical
9	B pulse	Pulse rate
10	B temp	Body temperature
11	B resp	Respiration rate
12	B meansbp	Mean systolic blood pressure
13	B meandbp	Mean diastolic blood pressure
14	B meanbp	Mean blood pressure
15	B pulseprs	Pulse pressure per second
16	WhtR	Waist to height ratio
17	AGE/BMI	age_bl / bmi
18	AGE_WhtR	Age * WhtR
19	D1	Output variable (TGs:HDL-c)
20	D1	Output variable (c-peptide)
21	D1	Output variable (Homa-IR) = (fasting insulin*glucose)/22.5


Density
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
age_bl

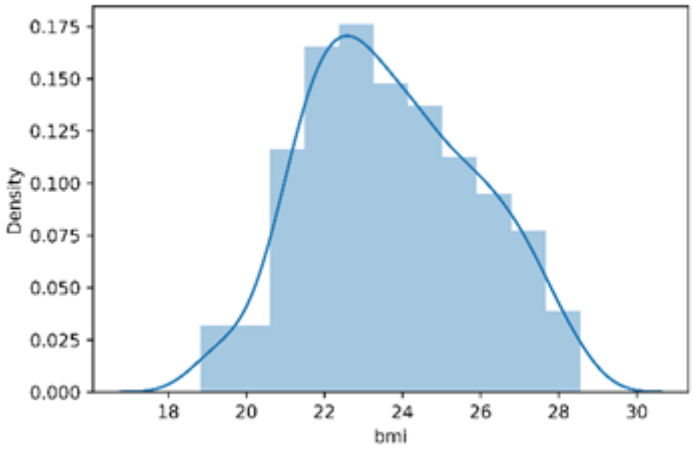
(a)


Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
5
10
15
20
25
30
AGE_WhtR

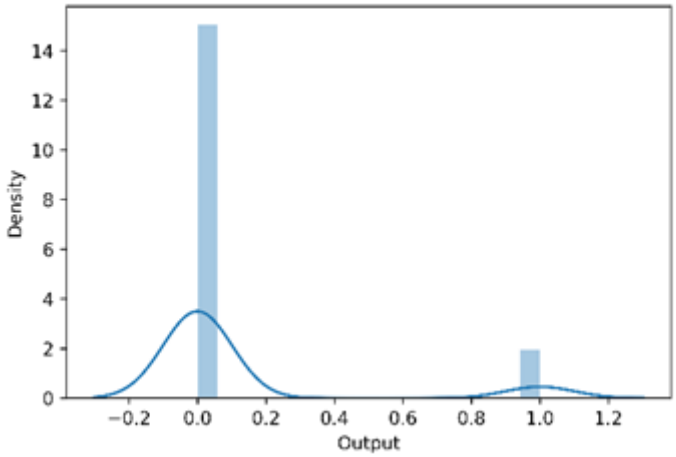
(b)


Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
AGE/BMI

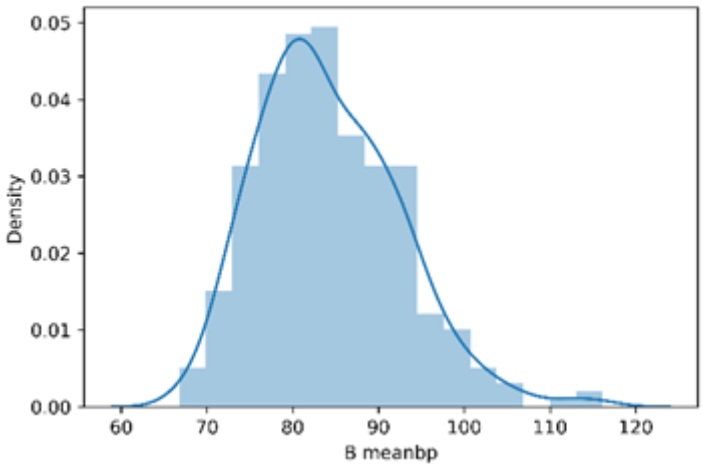
(c)


0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Density
18
20
22
24
26
28
30
bmi

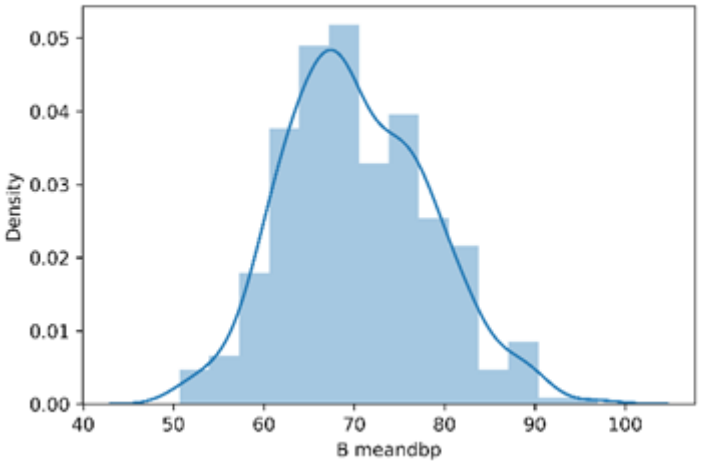
(d)


14
12
10
8
6
4
2
0
Density
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Output

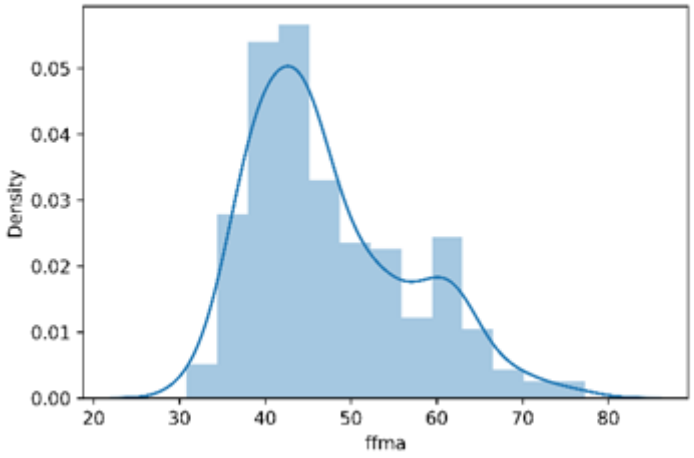
(e)


0.05
0.04
0.03
0.02
0.01
0.00
Density
60
70
80
90
100
110
120
B meanbp

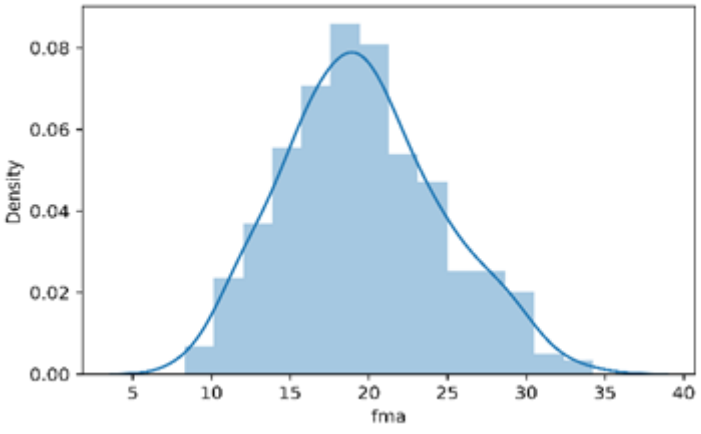
(f)


Density
0.05
0.04
0.03
0.02
0.01
0.00
40
50
60
70
80
90
100
B meandbp

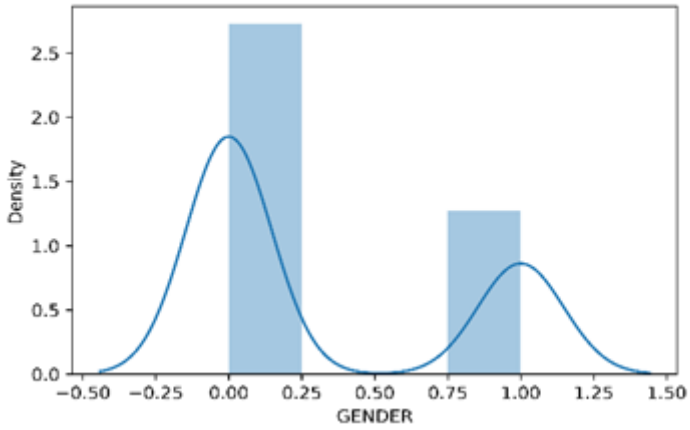
(g)


Density
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
70
80
ffma

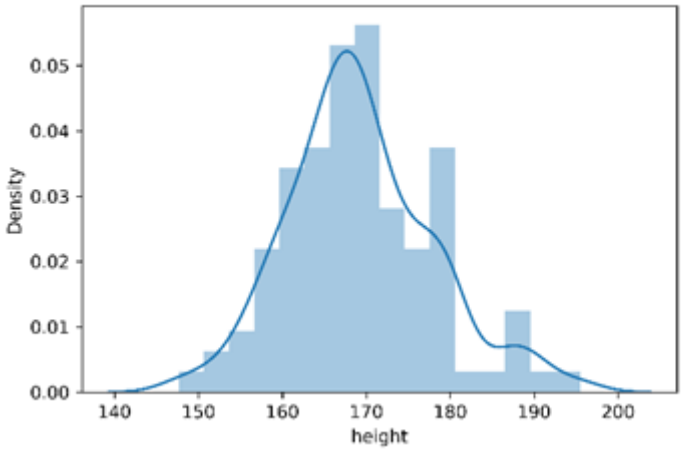
(h)


Density
0.08
0.06
0.04
0.02
0.00
5
10
15
20
25
30
35
40
fma

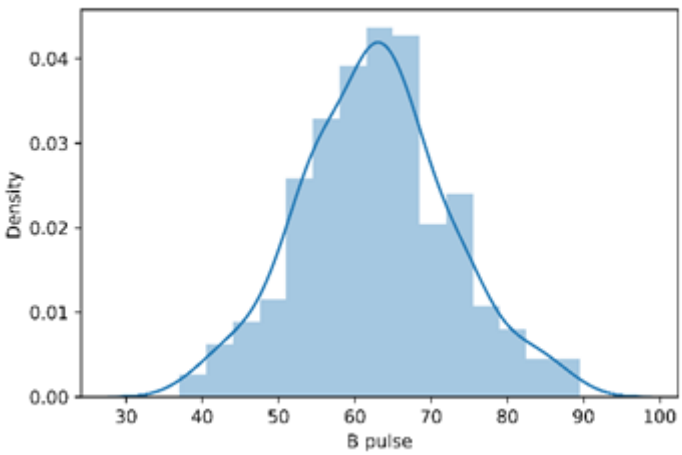
(i)


Density
2.5
2.0
1.5
1.0
0.5
0.0
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
GENDER

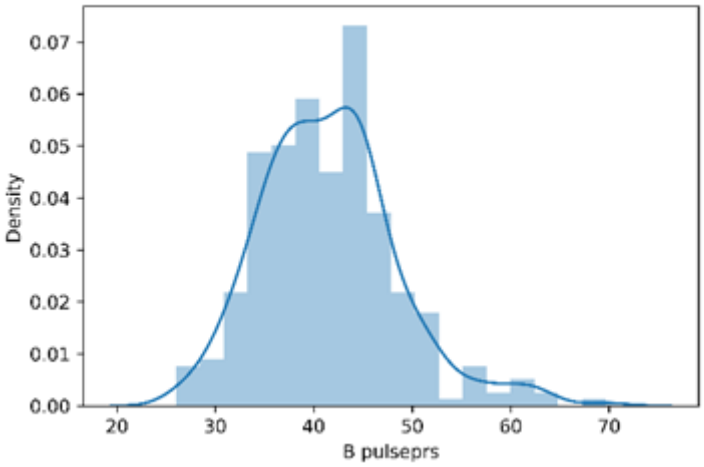
(j)


Density
0.05
0.04
0.03
0.02
0.01
0.00
140
150
160
170
180
190
200
height

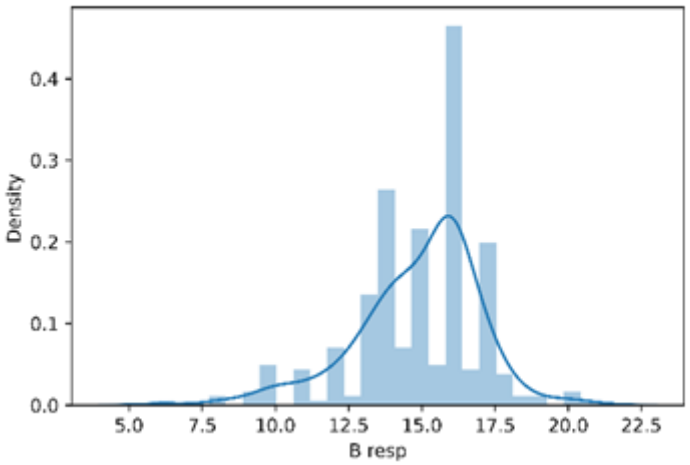
(k)


Density
0.04
0.03
0.02
0.01
0.00
30
40
50
60
70
80
90
100
B pulse

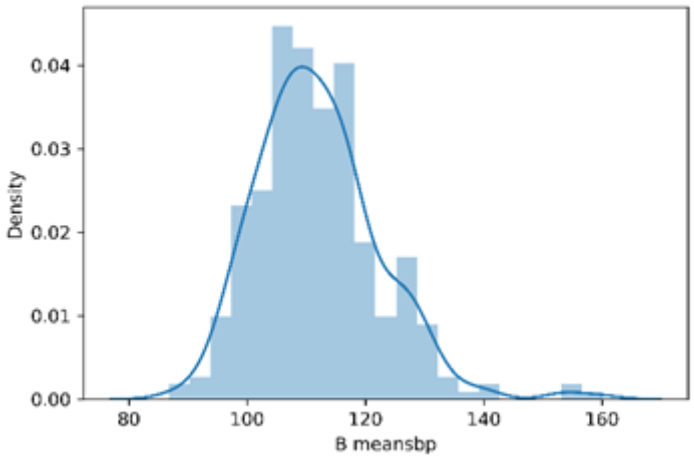
(l)


0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
20
30
40
50
60
70
B pulseprs

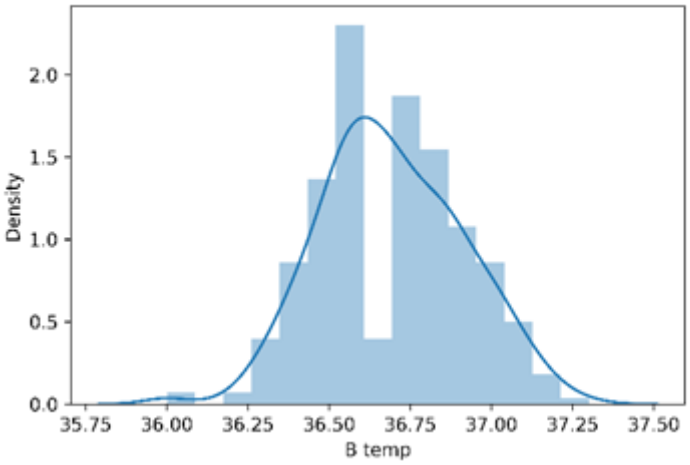
(m)


0.4
0.3
0.2
0.1
0.0
Density
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
B resp

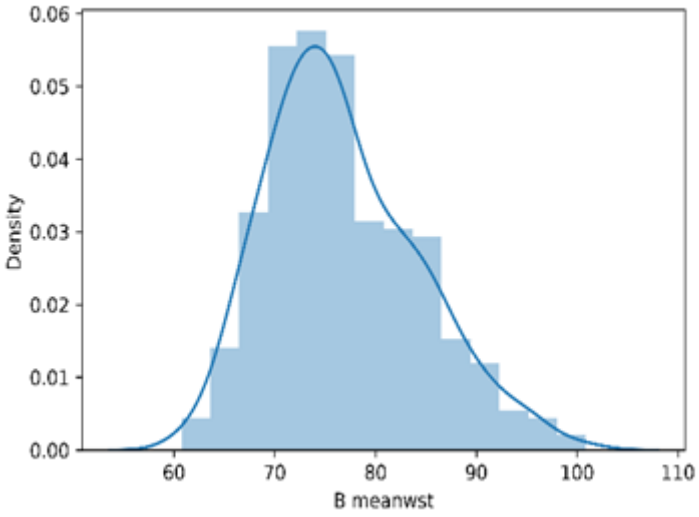
(n)


0.04
0.03
0.02
0.01
0.00
Density
80
100
120
140
160
B meansbp

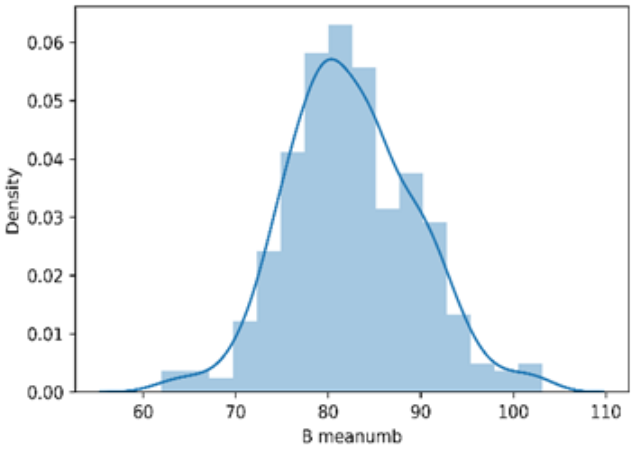
(o)


Density
2.0
1.5
1.0
0.5
0.0
35.75
36.00
36.25
36.50
36.75
37.00
37.25
37.50
B temp

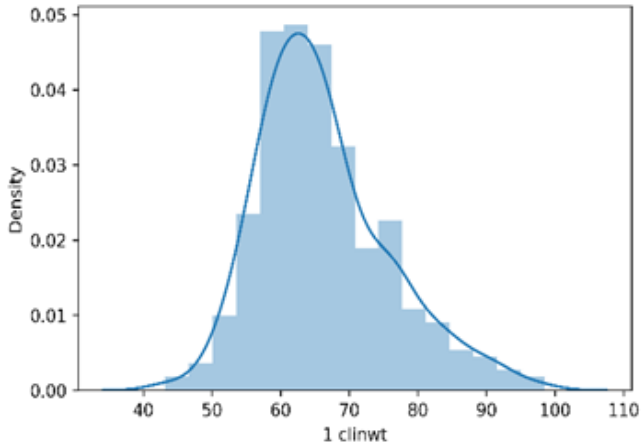
(p)


Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
60
70
80
90
100
110
B meanwst

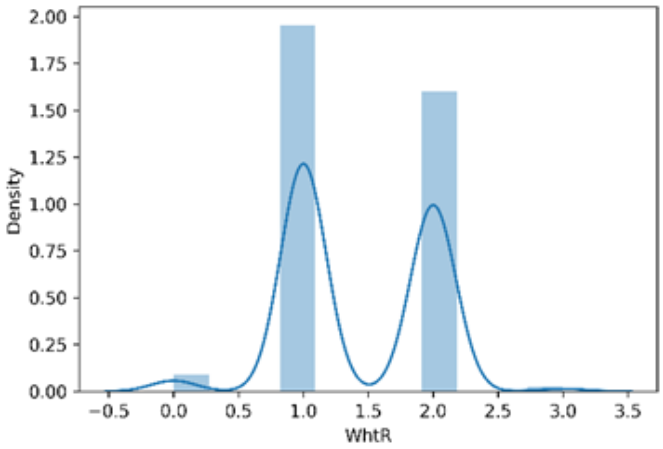
(q)


Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
60
70
80
90
100
110
B meanumb

(r)



(s)



(t)

Figures 4.1 (a-t) Dataset distribution

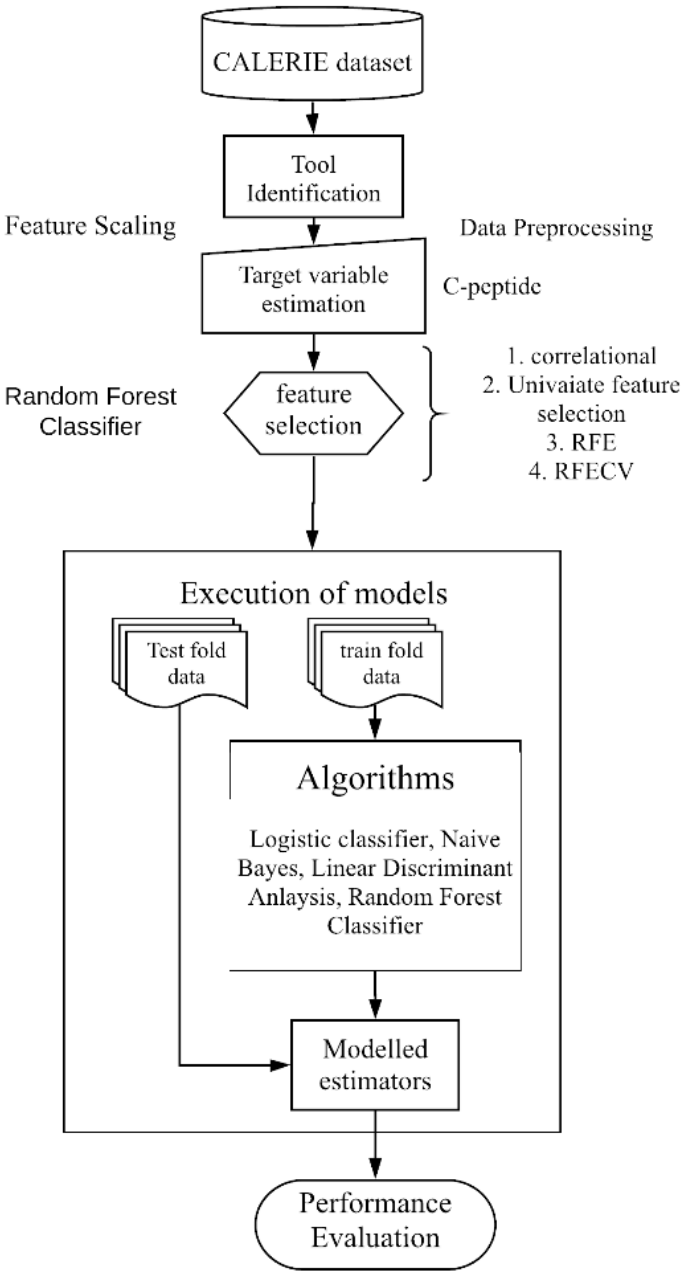


Figure 4.2 block diagram

4.1 FEATURE EXTRACTION, SCALING AND IDENTIFICATION

Based on these scales the target variable is divided well into their respective output labels. These target variables are continuous parameters that are now converted into categorical parameters. Therefore, this process turned the machine learning approach easier by transitioning into a classification problem. As there are either 2 or 3 classes in each target variable.

One of the workouts for handling any classification or regression task is feature selection. The pre-processing method used for the CALERIE study's data is known as attribute removal, and it eliminates information gathered by intrusive methods including blood sample reports and urological reports. The database, which has been retrieved in a variety of ways as described in [96], is fundamentally reduced as a result of this. After that, another adjustment is needed before using feature selection approaches. For feature selection, target variables are formulated [41]–[42]. These numbers are scaled to fall between 0 and 1, with 0 representing normal and 1 representing insulin resistance. HOMA-IR, c-peptide, and the ratio of triglycerides to HDL-c are the three indicators of insulin resistance. Although these measurements are made using insulin and fasting blood sugar levels [52]–[56]. However, they are scaled between 0, 1, and occasionally 2 classes in order to transform this problem into a classification problem as detailed in [94] [95]. According to the parameter of interest, tables 4.2–4.4 provide content verification for this strategy, which is also supported by literature. The target variable for the feature selection approach is now being satisfied by evaluating the aforementioned amounts. Now that the data has the necessary features, it must be trained. In order to reduce biases in the training data, stratified K fold cross-validation is performed.

Table 4.2 Triglycerides and HDL-c ratio scales

Sex	Range	condition	Outcome label
Male	<2.75	Normal	0
Male	>2.75	Insulin resistant	1
Female	<1.65	Normal	0
Female	>1.65	Insulin resistant	1

Table 4.3 C-peptide scales

C-peptide Range	Condition	Outcome Label
<=0.5	Type 1 Diabetic	2
0.5-1.8	Healthy	0
>1.8	Insulin resistance	1

Table 4.4 Homa-IR scales

Homa-IR Range	Condition	Outcome label
0.5-1.4	Healthy	0
>1.4	Early insulin resistance	1
>2.9	Significant insulin resistance	2

4.1.1 CORRELATIONAL ANALYSIS

The statistical evaluation method known as correlation analysis is used to assess the degree of relationships between two components that are quantitatively quantified. Finding any potential connections between factors or variables is made easier by this study. Extreme red indicates a high correlation, extreme blue indicates a negative correlation, and white indicates a neutral correlation in the Pearson's correlational heat map. It demonstrates the heat maps for the Spearman's (P), Kendall's (), and Phi k (k) correlations, respectively. Phi k (k) correlation is coloured from vivid red to deep blue and scaled between 0 and 1. The dataset's data points have very little, if any, relationship to one another. There are very few correlational data points for the IR (ratio of Triglycerides and HDL-c aim variable) in various plots. Since the output signal is boolean, logistic regression analysis is frequently used.

4.1.2 EXTRA TREE CLASSIFIER

Correlational mapping can be used in conjunction with feature identification. Accuracy might be used as a benchmark. Based on the features picked, a decision tree shows the correctness of the result. The extra trees feature classifier is an example of a classifier of this type. The initial training sample is used to construct each decision tree in the extra trees forest. Table 4.5–4.7 shows the findings, with the clinical body weight, mean waist size, mean waist size over umbilical cord, and fat mass showing greater relevance for the model. To create the model, all characteristics with scores larger than 0.05 were nevertheless picked.

4.1.3 UNIVARIATE FEATURE DIAGNOSIS

To establish the statistical significance of an attribute in relation to the target element, one feature at a time is subjected to a statistical Chi-squared test against the target variable or feature. The Chi-square test is run after every feature has been introduced. Four univariate feature analyses are used to rank the importance of different features: Select K-best, RFE, RFEcv, and PCA.

4.1.3.1 Select K-best In this type of univariate feature evaluation, a model is created by selecting the k-best features from among those that have the highest accuracy across numerous versions constructed over a range of capacities. Figure 7 displays selected k-best features together with Chi-squared values.

4.1.3.2 Recursive Feature Elimination (RFE) By examining the precision and removing one feature at a time, this method produces a variant. From the accuracy's design, characteristic importance is obtained by adding and eliminating the same feature. As a result, a ranking system is created in which the lowest position denotes the most worth.

4.1.3.3 RFE with Cross-validation (RFEcv) is a method of feature selection and deletion that is backward-compatible. This strategy starts by building a model and then checking its accuracy before deleting one feature at a time. The accuracy of the model's deviance when adding and removing the same feature is used to determine a feature's relevance. With the lowest rank indicating the highest significance, a ranking system is created.

4.2 FEATURE IDENTIFICATION

All conceivable physical and clinical lab findings are included in the CALERIE dataset. As was covered in the section above, the relative significance of each characteristic to the target variable is ascertained using a variety of methodologies. These strategies are split into thirteen categories, from feature 1 to feature 13, each of which has a specific technique. Based on their respective methodologies, these procedures produce various traits that vary in both quality and quantity.

4.3 STRATIFIED K FOLD CROSS-VALIDATION

To execute layouts with the proper precision and variance, the suggested model needs to be validated. The validation procedure puts proposed relationships between variables into numbers. In order to investigate the under-/ over-/ and well-generalized states of the specified model, an appraisal on hidden data must be made from the current situation evaluation of performance for multiple machine learning models. Cross-validation is used to verify the effectiveness of the machine learning model. In cases where there is a lack of data, it is also a resampling process used to test a variant. Stratified K fold CV is used in this situation since this function binary classifies the target variable based on Homa-IR levels.

Table 4.5 Features based on feature selection techniques for Target variable Ratio of Triglycerides and HDL-c.

S.No.	Features category	Feature Selection type	Features
1.	Features 1	i) Pearson correlation ii) Spearman's (P) correlation iii) Kendall's (τ) correlation iv) Phi k (φk) correlation	All parameters are taken
2.	Features 2	Chi-squared Select K-best	All parameters are taken
3.	Features 3	Extra Trees Classifier	Ffma, bmi, body temperature, mean waist (umb), pulse pressure, respiration rate, age, body weight, AGE_WhtR, AGE/BMI, WhtR, fma, sex.

4.	Features 4	Feature importance based on RFE	All parameters are taken
5.	Features 5	RFE with cross-validation	All parameters are taken
6.	Features 6	Embedded Random forests	Ffma, bmi, mean waist size (umb), mean systolic BP, height, fat mass, mean waist size, pulse, mean diastolic BP.
7.	Features 7	Lasso with Logistic Regression	age, fat free mass, bmi, body weight, mean waistsize (umb), WhtR, AGE_WhtR, fat mass, mean waist size, pulse.
8.	Features 8	Light GBM classifier	All parameters are taken

Table 4.6 Features based on feature selection techniques for Target variable C-peptide.

S.No.	**Features category**	**Feature Selection type**	**Features**
1.	Features 1	Pearson correlation Spearman's (P) correlation Kendall's (τ) correlation Phi k (φk) correlation	Fma, meanwst, meanumb, Age_WhtR, clinwt, bmi, WhtR, meanbp
2.	Features 2	Chi-squared Select K-best	Fma, meanwst, pulseprs, meanumb, clinwt, bmi, WhtR, resp, meandbp
3.	Features 3	Extra Trees Classifier	Fma, meanwst, pulseprs, pulse, meansbp, meanumb, clinwt, height, meanbp, meandbp
4.	Features 4	Recursive Feature Elimination (RFE)	Fma, ffma, pulseprs, clinwt, bmi, WhtR, gender, temp, resp
5.	Features 5	RFE with cross-validation	fma, meanwst, ffma, pulseprs, clinwt, bmi, WhtR, gender, temp, resp
6.	Features 6	Embedded Random forests	fma, meanwst, ffma, pulse, meanumb, meansbp, Age_WhtR, clinwt
7.	Features 7	Lasso with Logistic Regression	fma, meanwst, age, pulseprs
8.	Features 8	Light GBM classifier	All parameters are taken

Table 4.7 Features based on feature selection techniques for Target variable HOMA-IR.

S. No.	Features category	Feature Selection type	Features
1.	Features 1	Pearson correlation Spearman's (P) correlation Kendall's (τ) correlation Phi k (φk) correlation	All parameters are taken
2.	Features 2	Chi squared Select K-best	All parameters are taken
3.	Features 3	Extra Trees Classifier	Fat mass, mean waist size, pulse pressure, pulse, mean waistsize (umb), body weight, height, mean bp, mean diastolic bp, WhtR, age, AGE_WhtR, sex, body temperature.
4.	Features 4	Recursive Feature Elimination (RFE)	age, fat free mass, bmi, body weight, sex, mean waistsize (umb), Body temperature, respiration rate, mean systolic bp, WhtR, AGE_WhtR, height, fat mass, mean waistsize, pulse, mean diastolic bp, mean bp, pulse pressure.
5.	Features 5	RFE with cross-validation	Bmi
6.	Features 6	Embedded Random forests	age, fat free mass, bmi, mean waistsize (umb), AGE_WhtR, height, fat mass, mean waist size, pulse, mean diastolic bp, mean bp.
7.	Features 7	Lasso with Logistic Regression	age, fat free mass, bmi, body weight, mean waistsize (umb), WhtR, AGE_WhtR, fat mass, mean waist size, pulse.
8.	Features 8	Light GBM classifier	All parameters are taken

4.4 RESULTS AND COMPARISON

Table 4.8 Comparison of insulin resistance (IR) identification

Authors	**Population**	**Parameters Status**	**Parameters**	**Algorithms and Models**	**Results**
Zheng et al. [35]	36	Invasive	HbA1c, Diastolic BP, WHR	$\ln GDR = 4.964 - 0.121 \times HbAlc\ (\%) - 0.012 \times$ diastolic blood pressure (mmHg) $- 1.409 \times WHR$	R2=0.616, p<0.01
Stawiski et al. [11] NIRCa	315	invasive	Waist size, Triglycerides, HbA1c	1. MARSplines 2. ANN	• R2=0.44, p<0.0001, median error=3.6 % • R2=0.66, median error= 0.6%
Bernardini et al. [5] TyG-er	968	invasive	uricemia, leukocytes, gamma-glutamyltransferase and protein profile	Ensemble Random Forests	R2=0.666, p<0.05
Farran et al. [32]	1837	Non-invasive (with history reports)	Age, BMI, family history of diabetes, hypertensive status, family history of hypertensive, sex	1. KNN 2. LR 3. SVM (AUC scores)	A. 3 years data 1. 0.83 2. 0.74 3. 0.73 B. 5 years data 1. 0.83 2. 0.72 3. 0.68 C. 7 years data 1. 0.79 2. 0.72 3. 0.71
Proposed work (Triglycerides and HDL-c ratio)	321	Non-invasive (without history)	BMI	1. Logistic Regression 2. CatBoost 3. LDA 4. XGBoost	• **KNN:** Accuracy =0.74, AUC = 0.72. • **Catboost:** Accuracy

					=0.73, AUC = 0.73
Proposed work (c-peptide)	321	Non-invasive (without history)	fma, Gender, WhtR, fma, meanwst, age, pulseprs	1. SVM 2. Logistic Regression 3. KNN 4. ensemble classifier 5. LDA 6. NaïveBayes 7. Random forests 8. Adaboost 9. XGBoost CARTs	• **Naïve Bayes:** Accuracy = 0.8437, AUC= 0.69.
Proposed work (Homa-IR)	321	Non-invasive (without history)	BMI	1. Logistic Regression 2. SVM 3. LDA 4. Xgboost	• **logistic Regression** Accuracy =0.96, AUC=0.58

If observed, c-peptide showed a better overall performance with Dependable AUC scores of 0.69 for the accuracy of 0.8437. This model is described with the help of Gaussian Naïve Bayes though other algorithms in the system performed better than their AUC scores and other metrics were not supporting for recommending the model ahead for validation.

As illustrated in figure 4.3, the block diagram outlining the process, the model was initially built using a random forests classifier by dividing the data into train and test sets. Random forest is typically chosen, but in cases of decision-making issues, when the categorization between 0 and 1 is the current difficulty and the dataset for the CALERIE study is described in the article so far. The accuracy of this model's performance was 0.855, and Figures 4.4 and 4.5 display the confusion matrix for this experiment. This outcome applies to the entire dataset of 19 supposedly important parameters. As a result, gathering 19 different parameters is a laborious operation, and the dataset may contain less important

or less impactful parameters that could degrade the model. Feature selection is the process of removing these attributes from the dataset, and the approaches that will be applied are covered.

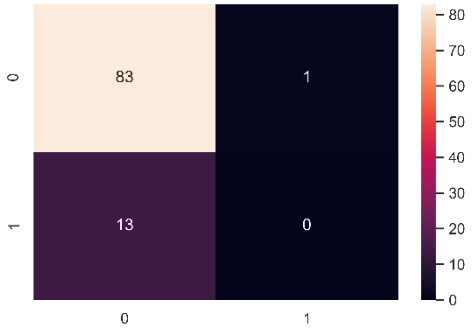


Figure 4.3 confusion matrix of random forest classifier

When the RFE technique, as explained in the previous section, is used to perform an experiment, the technique generates a limited number of characteristics based on the contribution of the model. Although they were provided all the characteristics, cross validating these features led to RFEcv in order to determine the right number of features that may produce superior outcomes.

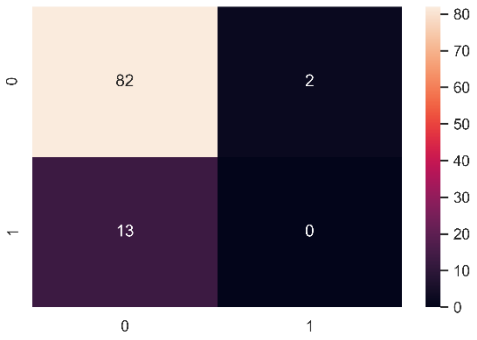


figure 4.4 confusion matrix of RFC with select K-best

As can be seen in figure 4.6, the cross-validation scores of the models against all features are plotted on the x-axis. This may help to explain the next effective method for determining the right number of features to include in the model to improve performance.

According to the graph in figure 4.6, it will be effective to select 6 features over 19 characteristics and acknowledge the feature importance charts made based on their feature important ratings. With the features on the x-axis and the associated feature importance scores on the y-axis, Figure 4.7 demonstrates this feature significant score. The characteristics presented in figure 4.7 include fat mass, mean waist size, mean waist size over an umbilical chord, age*WhtR, body weight, and body age. To achieve better results, machine learning techniques were used to these parameters.

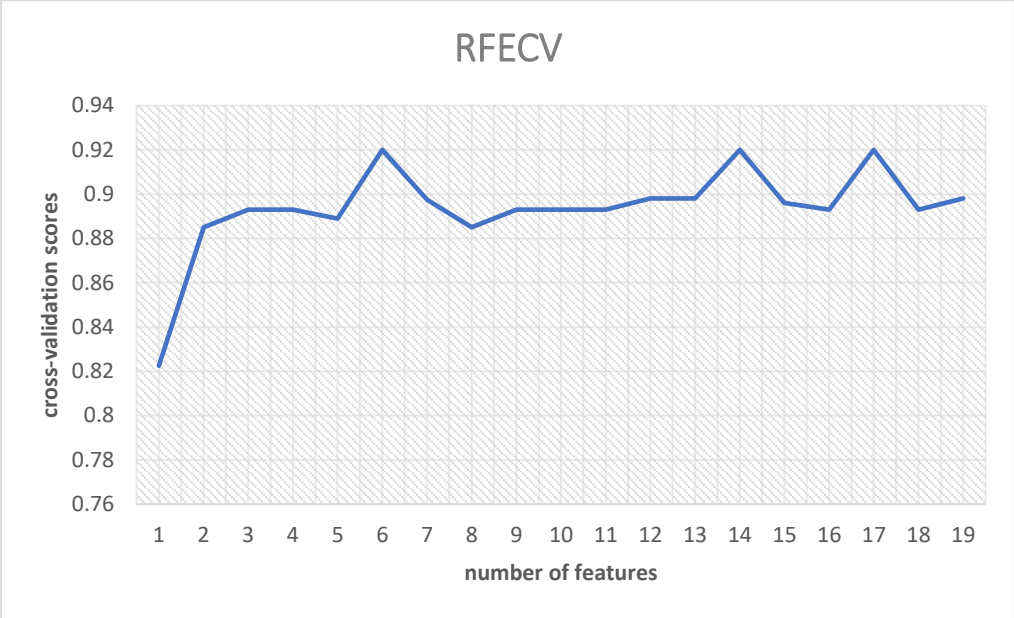


Figure 4.5 RFEcv number of features versus cross-validation scores

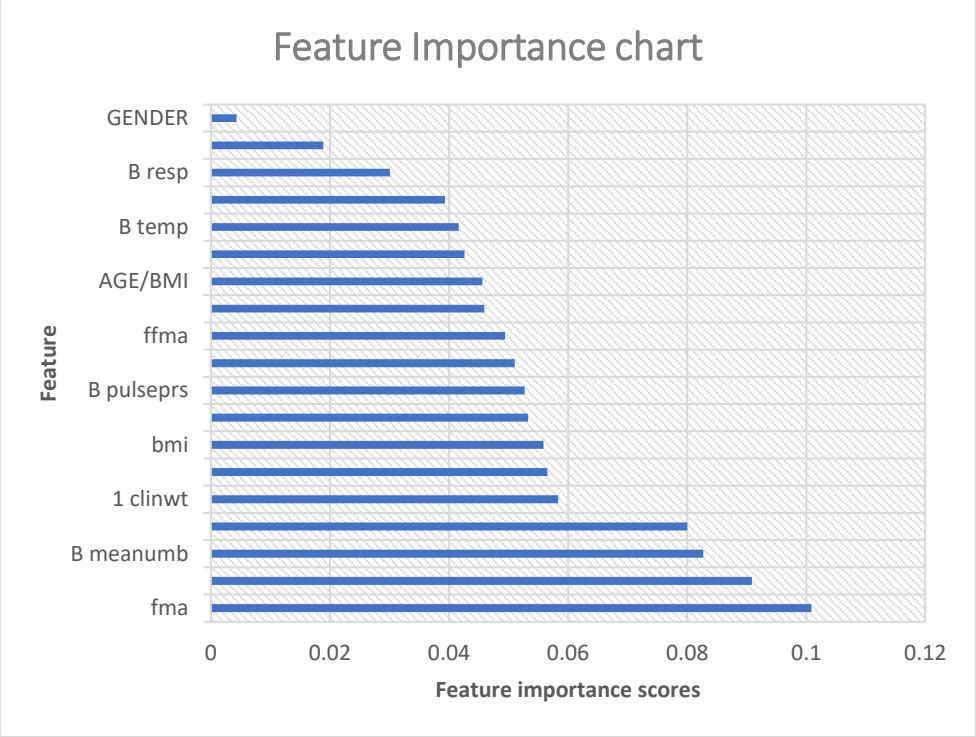


Figure 4.6 Feature importance chart

Table no 4.9 performance characteristics of the models after feature selection

Model	**Accuracy**	**AUC**	**Recall**	**Precision**	**F1-score**
Logistic Classifier	0.8749	0.6636	0.0333	0.1000	0.0500
Naïve Bayes	0.8437	0.6984	0.2333	0.3500	0.2700
Linear Discriminant Analysis	0.8751	0.6685	0.0333	0.1000	0.0500
Random Forest Classifier	0.8530	0.5157	0.0333	0.0333	0.0333

[50]-[51] The literature and other relevant data are used to select a few machine learning techniques, such as the logistic classifier, naive bayes classifier, linear discriminant analysis, and random forest classifier. Table 4.8 shows the outcomes for these strategies. It describes the machine learning approaches that should be applied as well as their performance traits, such as accuracy, recall, precision, and F1-Score. These criteria aid in determining which deployment strategies are most suited. Based on the table, NBC achieved better findings and showed a significant improvement in identifying a person with insulin

resistance. As seen in the confusion matrix figure 4.8, NBC was able to successfully identify the true positive much more effectively than earlier algorithms.

Despite having higher accuracy, logistic classifiers and linear discriminant analysis fell short in terms of other performance metrics like AUC, making them less reliable models to rely on for deployment. The performance of a random forest classifier without feature selection improves when a confusion matrix is generated for naive Bayes classifiers. While making some poor judgement calls, the prediction gradient of true positives for naive Bayes classifiers increased by over 500% from 1 to 6. Following feature selection, it is the most practical model created for the dataset, with a precision of 0.55 as shown in figure 4.9 and a very positive AUC score of 0.84 in figure 4.10.

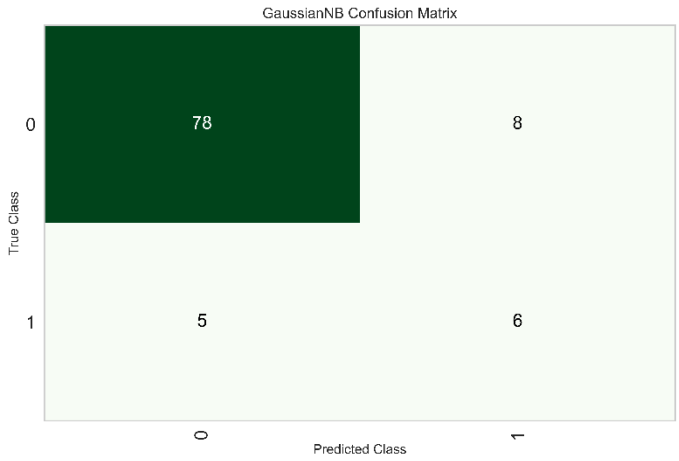


Figure 4.7 Confusion matrix of Naïve Bayes Classifier after feature selection

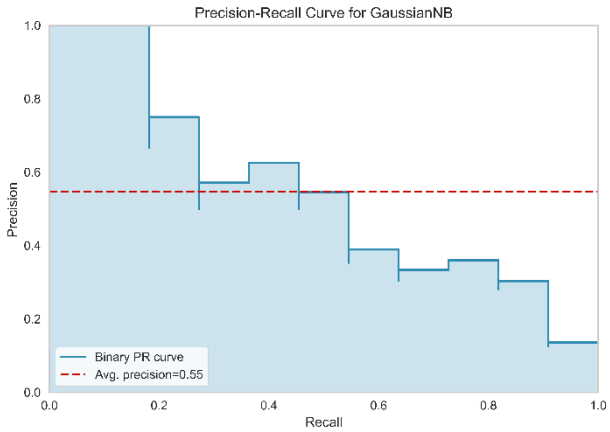


Figure 4.8 precision-recall curve for naïve Bayes

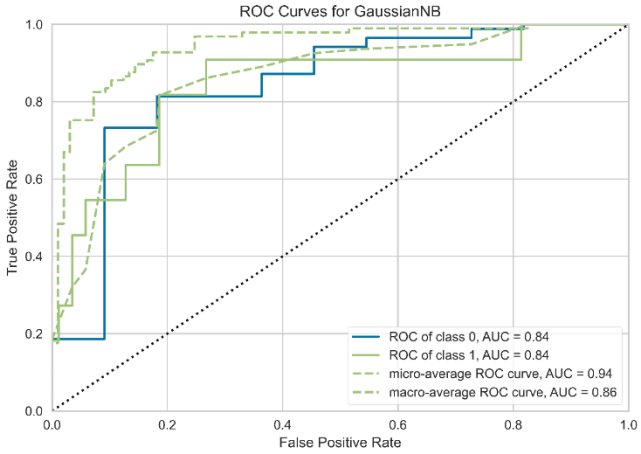


Figure 4.9 AUC and ROC curve

CHAPTER 5: EVALUATION OF THE MODELS

5.1 GENERATION OF SYNTHETIC DATASET AND RESULTS

Though the current CALERIE study data is limited to 235 individuals after the data filtering and pre-processing based on availability and reproducibility of the parameters. But to validate the models built for the identification of AT and prevention of AT one needs diversified data over different ages, sexes, heights, weights, foods habits, etc. Therefore, synthetic data created with 100000 individuals who are diversified over various parameters is generated using the same pattern as followed by the CALERIE study dataset features as shown in Tables 2 and 7. This creates an opportunity to test over a wide variety of audiences. This synthetic data when tested to know the similarity pattern it received over 87.3% accuracy of CALERIE study dataset. It is always better to not have highly accurate synthetic data to its source.

With the objective 2 model which is the identification of AT where one would need the parameters body temperature, body weight, alcohol, and protein consumption, waist to height ratio, pulse, and blood pressure levels. Similar parameters were identified and AT is derived as explained in objective 2 into categorical values either 0 or 1. The discussed parameters from synthetic data and huge test data of 100000 individuals are ready to be tested over the model. The model test accuracy score for AT identification over the synthetic data modelled after the CALERIE data is 0.66372 which is a satisfactory result.

Similarly, objective 3 models are tested over the synthetic data. Here the input parameters can be seen from the results and comparison table 12. In this case, BMI was the only appropriate parameter needed and it is tested against c-peptide component of blood serum which is scaled as shown in table 9. This data is used to test the model generated in objective 3 and test results achieved an accuracy of 0.7827.

Both the models performed 66.3% and 78.2% accurately respective to the objective 2 AT identification model and objective 3 AT prevention model. This test dataset has still maintained the same age groups but has improved the

permutations and combinations of the other parameters which should be a good fit as test data.

5.2 LIMITATIONS OF PROPOSED WORK

- Due to the limitation of a lesser volume of data hence the results are satisfactory.
- While scaling the continuous target variable into discrete variable sometimes the data associated with it gets generalized and becomes noise for the model.
- Though this is tested on generated synthetic data and a good advantage would be when tested on lab reports.

But these limitations can be overturned with more and more advanced monitoring and tracking systems from social media [100], enterprise tracking for energy/food consumption in the real world through surveillance [98]. Majority of the solutions can be sorted by predefining the possible outcomes through simulation modelling and crisis management [97], followed by e-commerce-based food delivery system which can be scheduled according to the user's bodily requirements through cloud scheduling [99].

CHAPTER 6: IMPLICATIONS OF COVID-19 ON DIABETICS

6.1 SUMMARY

Individuals with pre-existing diabetes tend to be more vulnerable to COVID-19 due to fluctuations in blood sugar levels and complications from diabetes. Around 20–50% of those who contracted the coronavirus had diabetes, as was seen globally. There is no new evidence, though, indicating people with diabetes are more likely than people without diabetes to get COVID-19. Recent research, however, suggests that diabetic complications may be as least twice as likely to cause death. Given the multifold mortality rate of COVID-19 in diabetic patients, this work proposes a COVID-19 risk prediction model for diabetic patients using a fuzzy inference system and machine learning approaches. This study sought to determine the risk level of COVID-19 in diabetic patients without seeking medical help in order to take immediate action and prevent the multifold death rate of COVID-19 in diabetic patients. Eight input parameters, which were discovered to represent the symptoms that diabetic patients perceived to be most influential, are used in the suggested model. 15 models were constructed over the rule base using various cutting-edge machine learning approaches. The CatBoost classifier has the highest accuracy, recall, precision, F1 score, and kappa score for the current data created using fuzzy inference technique. After adjusting for hyperparameters, the CatBoost classifier outperformed logistic regression and XGBoost, achieving accuracy of 76% and gains in recall, precision, F1 score, and kappa score. Stratified k-fold cross-validation is used for validation.

6.2 INTRODUCTION

The emergence of the SARS-CoV-2 has presented a hitherto unheard-of challenge to the global healthcare industry. The coronavirus disease 19 (COVID-19) has spread quickly over the world due to its high infectivity and relatively moderate severity. COVID-19 first appeared on December 8, 2019, in the Chinese province of Hubei. Since then, it has spread to many other nations, with a total of 21,294,845 cases and 7,61,779 deaths documented globally [105]. COVID-19 symptoms, which include fever (98.6%), cough

(59.4%), and sore throat (5%), frequently appear 2 to 14 days after infection. To counteract COVID-19 and minimise human involvement, many cutting-edge solutions are being investigated [106–109]. The most recent examinations have also revealed that COVID-19's case was significantly influenced by age, sex, recent travel history, and pre-existing medical issues. Serious issues [110], like pneumonia or death, could result from it. Diabetes patients are more likely to experience severe side effects such adult respiratory distress [111–113]. Although no recent research has shown that people with diabetes are more likely than those without diabetes to get COVID-19, some recent research has shown that people with diabetes may be at least twice as likely to die from diabetes-related complications. In a study of 52 trauma victims, diabetes was shown to be a disease in 22% of the 32 non-survivors [114], in 16.2% of 173 patients with acute disease, and in 12% of 140 hospitalised patients [115, 116]. In terms of COVID19, there has been a two-fold increase in the prevalence of diabetic patients in intensive care as compared to non-diabetic patients. Diabetes patients appear to have a roughly threefold higher mortality rate [105, 114–117]. Patients with diabetes have been found to have a higher risk of COVID-19. A few recent studies have shown that people with diabetes have an up to 50% higher chance of having fatal findings from the COVID-19 [118].

A number of classification techniques, including support vector machine (SVM), decision tree (DT), and K-nearest neighbour (KNN), are used on COVID-19 data in [119] under the machine learning (ML) methodology. The disease risk levels dataset is generated using an adaptive neuro-fuzzy inference system (ANFIS). In terms of these data, SVM offered 100% accuracy, however when evaluated against the test data, it produced a risk prediction of 80%. Similar to this, a variety of patient data is collected, including the patients' treatment modalities as categorical values, their place of origin, and the number of survivors [120]. The random forest method was used, and this algorithm was then strengthened using the AdaBoost algorithm, producing a remarkable 0.86 F1 score. Additionally, it provided a survival probability rate based on travel history, citizenship, gender, and age group. By using fuzzy membership functions and fuzzy rule descriptions, Kerk et al. [121] developed the parametric

requirements for the Takagi-Sugeno-Kang fuzzy inference system model to act as an n-ary aggregation function.

This work suggests a COVID-19 risk prediction model for diabetic patients using a fuzzy inference system and machine learning methodologies in light of the multifold death rate of COVID-19 in diabetic patients. The suggested model requires eight input variables, which were discovered to be the most significant symptoms in diabetic patients who contracted COVID-19, including age, sex, and travel history for the previous three weeks. These parameters are: fever, cough, sore throat, cardiovascular disease, high blood pressure, and age. Chest cough, dry tickling cough, bronchitis, post-viral cough, and whooping cough have all been thought of as possible levels of cough. Similar to this, all potential fever phases are taken into account. Another important element in COVID-19 has also been sexual orientation. When all significant suffering states were averaged out, men made up 61.8% of the cases, while women made up 38.2%. According to research, the case fatality rate for COVID-19 increases by about 1.4% with age. Numerous applications [128-132] use computational intelligence approaches, such as fuzzy logic [122-127]. The language and imprecise data, which are not meant to describe the process, are altered by a fuzzy logic controller. In this instance, specialised information is gleaned from a few doctors' expertise treating illnesses like COVID-19. It makes it possible to use actual rules, like the way people think, and it can mimic human intelligence. Iwendi et al. [120] employed an adaptive neuro-fuzzy inference system (ANFIS), which is used to describe and operate ill-defined and unpredictable systems, to forecast the risk variables for COVID-19.

Support vector machines were used to classify the COVID-19 dataset, and they outperformed all other classifiers in accuracy with a score of 100%. Therefore, for COVID-19 patients, a risk prediction of 80% has been reached. Additionally, to estimate the severity of COVID-19, the authors of [120] examined a variety of data about COVID-19 patients, including travel, health, and age. This forecast, which is enhanced by the AdaBoost algorithm, was made using the random forest model. The end result is an accuracy of 94% and an F1 score of 0.86 has been achieved. The authors of [121] outlined the parametric

requirements for the Takagi-Sugeno-Kang fuzzy inference system model to work as an n-ary aggregation function as well as fuzzy membership functions and fuzzy rules.

In a specific area of information, a fuzzy controller can be used to rank imprecision and uncertainty. Domain-specific knowledge and experience in treating various diseases, such as COVID-19, are crucial for building fuzzy traffic controllers and designing the language protocols that create the control input to the control system. Eight input factors, one output, and a total of 3,888 criteria (3*3*3*3*3*4*2*2) are used to create the COVID-19 risk level, which gives the risk level of COVID-19, five in number, to diabetic patients. The lowest degree of danger is level 1, and the highest level is level 5. The following fifteen models were created using cutting-edge machine learning methods: logistic regression, AdaBoost, CatBoost, gradient boosting, random forest, extreme gradient boosting, extra trees, light gradient boosting machine, decision tree, linear discriminant analysis, K-neighbors, SVM-linear kernel, ridge, naive Bayes, and quadratic discriminant analysis. It has been determined how well these fifteen models performed in terms of accuracy, recall, precision, F1 score, and kappa score. The best performing model is further optimised using the hyperparameter method. The block diagram of the proposed inference pipeline is shown in Figure 6.1.

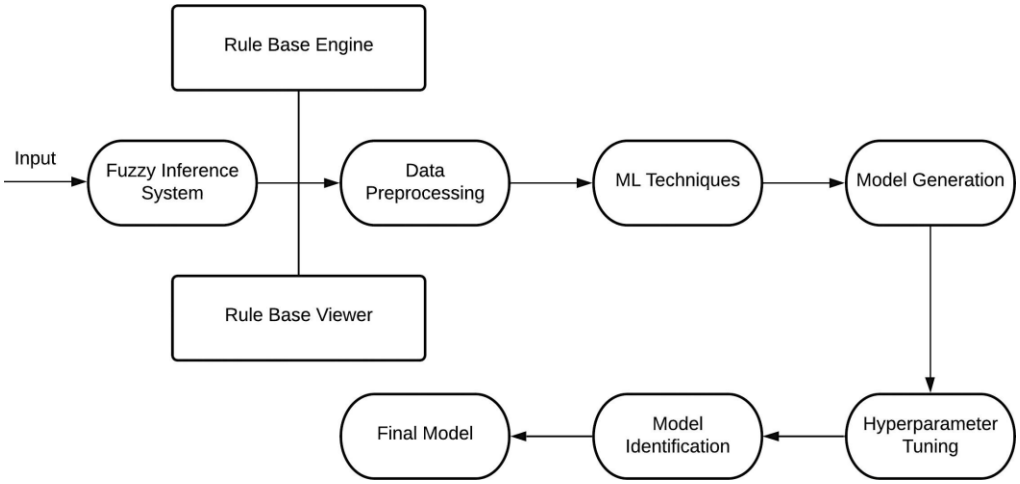


Figure 6.1 Proposed Inference pipeline

The following are the work's main contributions:

(i) Using a fuzzy inference system (FIS) and machine learning (ML) approaches, the effects of COVID-19 on diabetic patients are determined.

(ii) Using a variety of ML techniques, different machine learning models are trained, and the effectiveness is checked using stratified K-fold cross-validation. Bias and variance issues are eliminated from consideration.

(iii) The output ML model can be directly used as a learning metric to improve the accuracy of the current model as well as to validate the actual data.

The remainder of the investigation is conducted in the manner listed below. The proposed work's rule base, fuzzy set membership functions for input and output variables, and the fuzzy model for the eight-input fuzzy traffic controller are all described in Section 6.3. The simulation's specifics are provided in Section 6.4. The machine learning models are presented in Section 6.5, along with the outcomes of several machine learning models.

6.3 FUZZY INFERENCE SYSTEM

A fuzzy logic-based controller is presented with eight input parameters—fever, cough, sore throat, cardiovascular disease, high blood pressure, age, sex, and travel history for the previous three weeks—and one output that gives the COVID-19 risk level, which is five in number, to the diabetes patients. The lowest degree of danger is level 1, and the highest level is level 5. Low, medium, and high degrees of coughing have been assigned to fuzzy sets of levels. Cough intensity is classified as low, medium, or high depending on how mild or intense it is. All five types of cough have been taken into account, including whooping cough, dry tickling cough, bronchitis, post-viral cough, and chest cough [131]. Low, medium, and high fever levels have been denoted by fuzzy sets. A fever level of 98.0°F to 99.0°F is considered low, 98.0°F to 101°F is considered medium, and 100.0°F and higher is considered high. It covers all phases of fever, including those with no symptoms, prodromal, second-stage chills, third-stage flushes, and defervescence [132]. The sore throat has been categorised

into fuzzy sets of low, medium, and high levels No coughing to stage 1 of coughing is considered low, stage 1 to stage 2 is considered medium, and stage 2 to stage 3 is considered high. The sore throat has been taken into account in all three stages. In stage 1, the patient may experience exhaustion, tiredness, and runny or clogged nose. At this stage of the cold in stage 2, the patient may experience a runny nose, minor aches, sneezing, tiredness, weariness, or cough. The third stage of a cold is the most severe and is marked by congestion, a sore throat, and other symptoms [131].

Low, medium, and high levels of cardiovascular disease have been categorised as fuzzy sets of different levels. No evidence of cardiovascular disease to stage B is considered low, stage A to stage C is considered medium, and stage B to stage E is considered high. Stage A is seen as a condition before a heart attack. Stage B is also known as systolic left ventricular dysfunction, or pre-heart failure when there have been no symptoms of the condition. A stage B echocardiography is one that shows an ejection fraction (EF) of 40% or less and reduced EF (HFrEF) due to particular reasons. Stage C patients are those who have been diagnosed with cardiac disease and who currently exhibit or previously displayed signs and markers of the ailment. Patients in stage E who do not improve with treatment are taken into consideration [129]. Low, medium, and high blood pressure levels have been included in fuzzy sets. Low blood pressure is defined as 110 to 120, medium blood pressure as 115 to 135 and high blood pressure as 130 to 140+ [132]. With increasing age, the case fatality rate for COVID-19 is reported to be almost 1.4%. According to the Health Ministry of the Government of India, 63% of coronavirus deaths in India have been reported in those 60 and older, which is consistent with statistics on COVID-19 mortality rates from other countries. Fuzzy sets of age have been classified as low, medium, high, and extremely high in light of this. Ages 0 to 20 are considered low, 15 to 35 are considered medium, 35 to 55 are considered high, and 45 and above are considered very high. Additionally, sex has been crucial to COVID-19. According to data, men are substantially more likely than women to experience acute symptoms and pass away. Data gathered by Global Health 50/50 [135] were taken into account. In Italy, there were 71% more male

deaths than female deaths, while in Spain, there was 35% more female deaths than male deaths. A mean of all significant nations has been determined, and it reveals that men were responsible for 61.8% of case deaths while women were responsible for 38.2% of those deaths. In light of these factors, fuzzy sets of sex have been rated as low and high. Gender male denotes high, while gender female denotes low. The last three weeks' worth of hazy travel history has been seen as both low and high. No travel history during the past three weeks is considered low; however, if any, it is considered high (Table 6.2). The three subprocesses in the proposed version are fuzzification, fuzzy inference, and defuzzification. Sharp values are transformed into fuzzy sets serving membership needs during fuzzification. These fuzzy sets are subsequently supplied into the if-then sentences that make up the rule base. Finally, Table 6.1 shows fuzzy sets of the input and output components.

The final stage of this paradigm, defuzzification, involves using the fuzzy rule basis to provide precise output signal values. It is the fuzzification process's opposite. Mamdani created an inference method that uses centroid defuzzification to turn ambiguously described locations into exact values. Figure 2 displays the membership features. Table 6.2 displays the fuzzy set's rule base. Eight input factors, one output, and a total of 3,888 (3∗3∗3∗3∗3∗4∗2∗2) criteria are used to create the COVID-19 risk level, which is five for diabetes individuals. The lowest danger is level 1, while the worst risk is level 5.

6.4 FUZZY SIMULATION

The fuzzy logic toolkit and MATLAB 8.1 were used to complete the simulation. The fuzzy logic toolkit was used for two reasons. This toolkit can be used to establish a rule base fast and efficiently in the beginning, and adjustments can be made as necessary. Second, it cuts down on the time needed to build the rule foundation. The rule framework for various input variables and outputs is shown in Table 6.1. A sample of eight outcomes is displayed in Table 6.2.

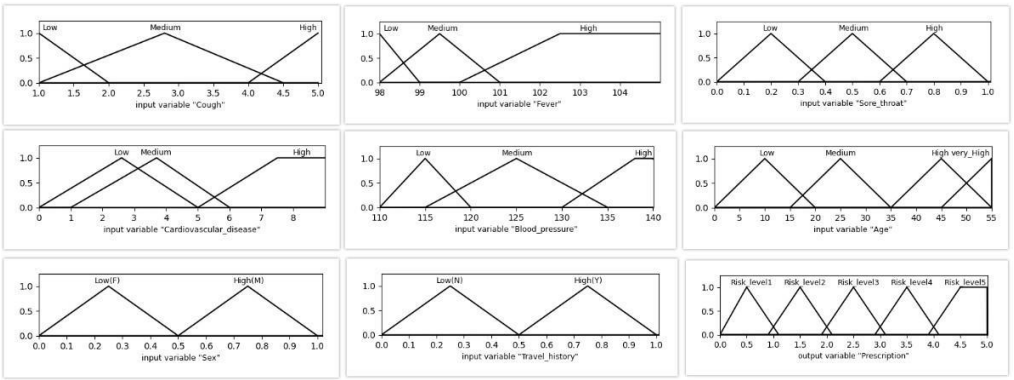


Figure 6.2 Fuzzy set membership diagrams

TABLE 6.1 Input/Output Variables And Their Fuzzy Sets

Input/Output Variables	**Fuzzy sets**
Cough (Input1)	{low, medium, high}
Fever (Input2)	{low, medium, high}
Sore throat (Input3)	{low, medium, high}
Cardiovascular disease (Input4)	{low, medium, high}
High Blood pressure (Input5)	{low, medium, high}
Age (Input6)	{low, medium, high, very high}
Sex (Input7)	{low, high}
Travel History during the last 3 weeks (Input8)	{low, high}
Prescription (Output)	{risk level 1, 2, 3, 4, 5}

Table 6.2 Rule Base Of The Fuzzy Inference

Sl. No.									**Output Parameter**
	Cough	**Fever**	**Sore Throat**	**Cardio. Disease**	**B.P.**	**Age**	**Sex**	**Travel History**	**Risk Level**
1	Low	Low	Low	Low	Low	Low	Low	Low	Risk Level 1
2	Medium	Low	Low	Low	Low	Low	Low	Low	Risk Level 1
3	High	Low	Low	Low	Low	Low	Low	Low	Risk Level 1
4	Low	Medium	Low	Low	Low	Low	Low	Low	Risk Level 1
5	Medium	Medium	Low	Low	Low	Low	Low	Low	Risk Level 1
6	High	Medium	Low	Low	Low	Low	Low	Low	Risk Level 1
7	Low	High	Low	Low	Low	Low	Low	Low	Risk Level 1
8	Medium	High	Low	Low	Low	Low	Low	Low	Risk Level 1
9	High	High	Low	Low	Low	Low	Low	Low	Risk Level 1
10	Low	Low	Medium	Low	Low	Low	Low	Low	Risk Level 1
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
3888	High	High	High	High	High	Very High	High	High	Risk Level 5

Table 6.3 Sample Of Eight Outputs

Cough	**Fever**	**Sore Throat**	**Cardiovascular Disease**	**BP**	**Age**	**Sex**	**Travel History**	**Prescription**
3	101.5	0.5	4.5	120	27	0.5	0.5	Risk level2
2	105	0.6	5	110	37.5	0.8	0.8	Risk level3
5	103	0.3	4	140	47	0.6	0.8	Risk level4
5	105	0.8	5	150	32	0.4	0.7	Risk level5
2	99	0.3	4	110	48	0.5	0.4	Risk level1
6	99	0.6	3	120	25	0.4	0.4	Risk level2
6	103	0.2	6	125	20	0.5	0.8	Risk level4
2	98	0.6	5	135	55	0.3	0.3	Risk level3

6.5 MACHINE LEARNING

It is classified as a multiclass classification problem since there are eight input parameters and only one output parameter—the goal variable. During the data pre-processing stage, the input and output parameters' crisp values are converted into numeric values. Once the dataset is prepared, several machine learning models are used to train on it and test it. This is done by applying fifteen machine learning models, including quadratic discriminant analysis, K-neighbours, SVM-linear kernel, ridge, Naive Bayes, logistic regression, AdaBoost, CatBoost, gradient boosting, random forest, extreme gradient boosting, additional trees, and light gradient boosting machine. The performance of these fifteen models was also evaluated using the accuracy, AUC, recall, precision, F1, and kappa scores criteria. Based on the values of these parameters, the best model is selected from among these fifteen models. Finally, hyper-parameter adjustment is done depending on the dataset and various patterns to optimise performance. The performance features of ML techniques on COVID-19 symptoms are shown in Table 6.4. Five performance metrics—accuracy, recall, precision, F1 score, and kappa score—are employed. With the exception of the F1 score, which is generated using recall and precision, all parameters are independently determined. All parameters follow the same trend, with the exception of the AUC score. It has been found that the AdaBoost and CatBoost classifiers perform second and third best to the logistic regression model. The hyper-parameter tuning procedure can help to further enhance these qualities. Accuracy, recall, precision, F1 score, kappa, confusion matrix, ROC, and AUC curves were employed as performance indicators in this study. In addition, graphs showing learning rates were also created in conjunction with the quantity of training instances. Accuracy is the possibility of being able to tell the correct class apart from all other classes, to put it simply. Precision is the proportion of accurately classified positive classifications among all positive classifications. Less probability of misclassifying a class as another class results from more precision. Recall or sensitivity can be used to locate and review the observations that were correctly identified among all other possible true observations in the experiment. The weighted average of recall

and precision, known as the F1 score, can be used to detect the uneven class distribution. The classes are distributed equally in this instance. As a result, accuracy is preferred instead. Therefore, only accuracy scores are frequently addressed. Kappa values are used to describe the distribution of the class variable and data collection. This statistic does not provide any additional value over accuracy because the present dataset was produced using a fuzzy rule base.

Each machine learning model is shown in Table 6.4 along with its corresponding performance metrics, including accuracy, recall, precision, F1 score, and kappa score. Based on these rankings, one would choose the appropriate models, ideally testing and predicting further use cases. Even though these models are imperfect, you can still use them by optimising the algorithm's hyperparameters as you train the model. The creation of bar graphs for each model and performance metric comes next (Figures 6.3–6.7). Based on these results, CatBoost, logistic regression, and XGBoost all outperformed the control groups. Hyper-parameter optimization was followed by a significant improvement in every model produced by these techniques. CatBoost's accuracy increased by almost 3 percent. Over 1.1% and 3% improvements were made to XGBoost and logistic regression, respectively. Now, each model is tested for accuracy, recall, precision, F1 score, and kappa score, among other performance metrics. Optimization of hyperparameters is utilised to boost performance even more. The accuracy, recall, precision, F1 score, and kappa score after hyperparameter optimization are shown in Figures 6.3-6.7 respectively.

Table 6.4 PERFORMANCE CHARACTERISTICS OF ML TECHNIQUES ON COVID-19 SYMPTOMS

S.no	Model	Accuracy	Recall	Precision	F1-Score	Kappa
1	Logistic Regression	0.7391	0.503	0.7536	0.7195	0.5995
2	Ada Boost Classifier	0.7324	0.549	0.7433	0.7093	0.5908
3	CatBoost Classifier	0.7166	0.601	0.7159	0.7136	0.5817
4	Light Gradient Boosting Machine	0.7041	0.557	0.7031	0.6997	0.561
5	Gradient Boosting Classifier	0.6968	0.483	0.7052	0.6816	0.537
6	Extreme Gradient Boosting	0.6935	0.473	0.7037	0.6757	0.5303
7	Extra Trees Classifier	0.6928	0.562	0.6929	0.6908	0.5494
8	Decision Tree Classifier	0.6909	0.59	0.697	0.6922	0.5501
9	Random Forest Classifier	0.6909	0.558	0.6898	0.6884	0.5459
10	SVM - Linear Kernel	0.6733	0.449	0.703	0.639	0.4971
11	K Neighbours Classifier	0.6534	0.495	0.6474	0.6461	0.485
12	Ridge Classifier	0.6487	0.345	0.4885	0.5572	0.4365
13	Quadratic Discriminant Analysis	0.5182	0.426	0.5352	0.5067	0.3164
14	Naive Bayes	0.4943	0.493	0.6474	0.5279	0.3152

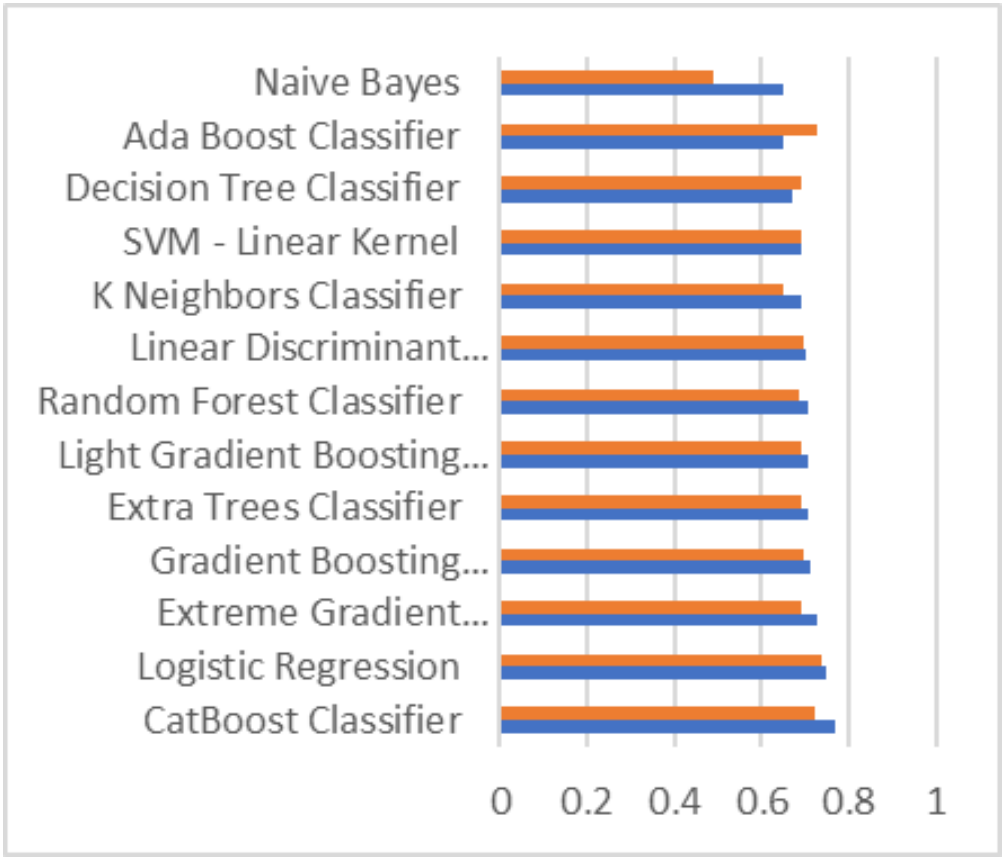


Figure 6.3 Comparison of accuracy after hyper-parameter optimization.

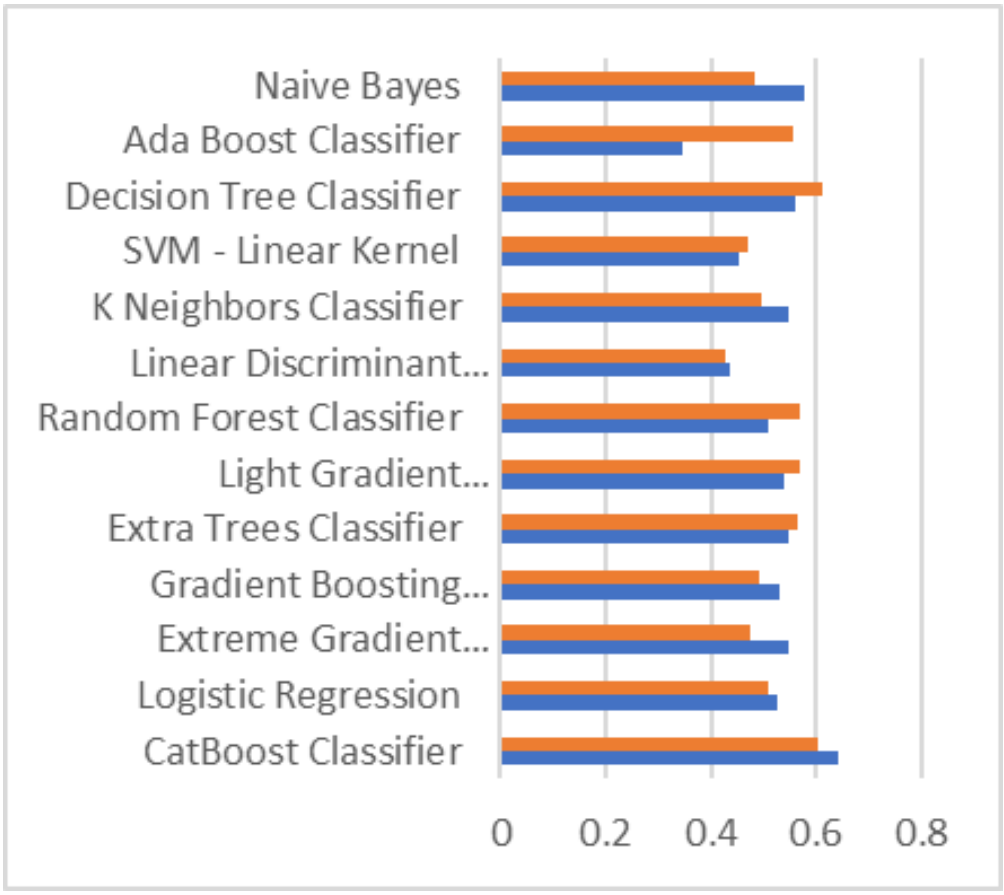


Figure 6.4 Comparison of recall after hyper-parameter optimization.

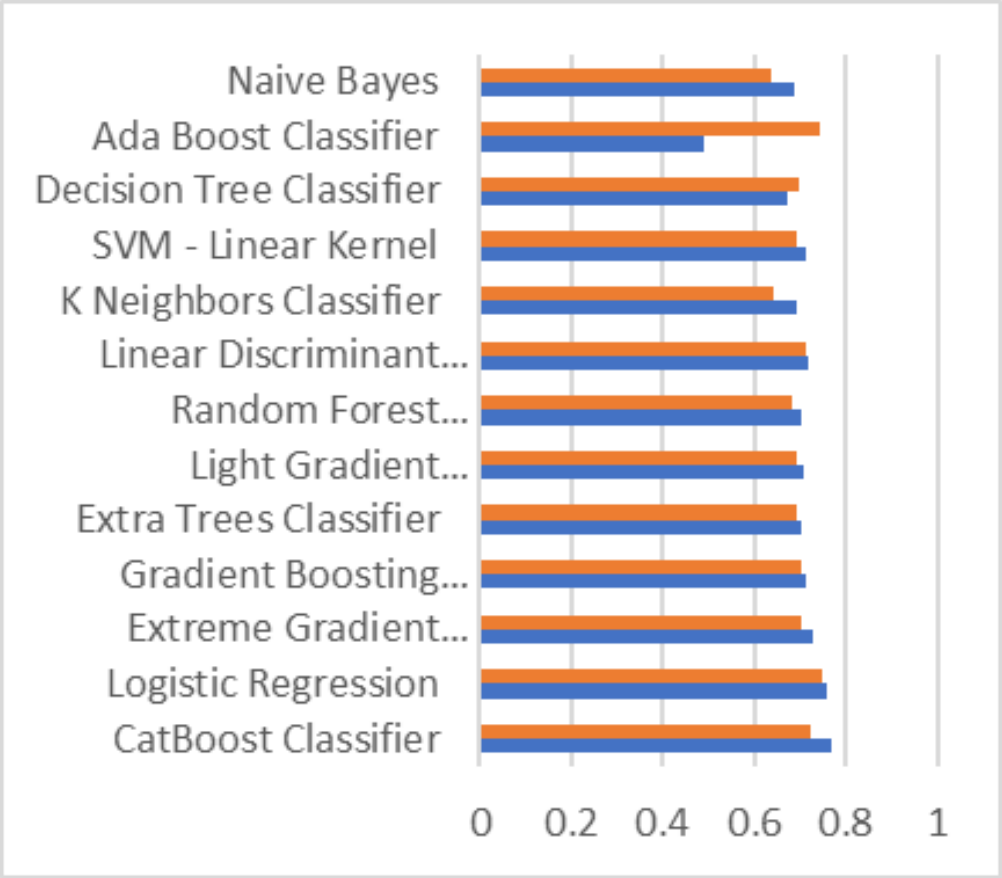


Figure 6.5 Comparison of precision after hyper-parameter optimization.

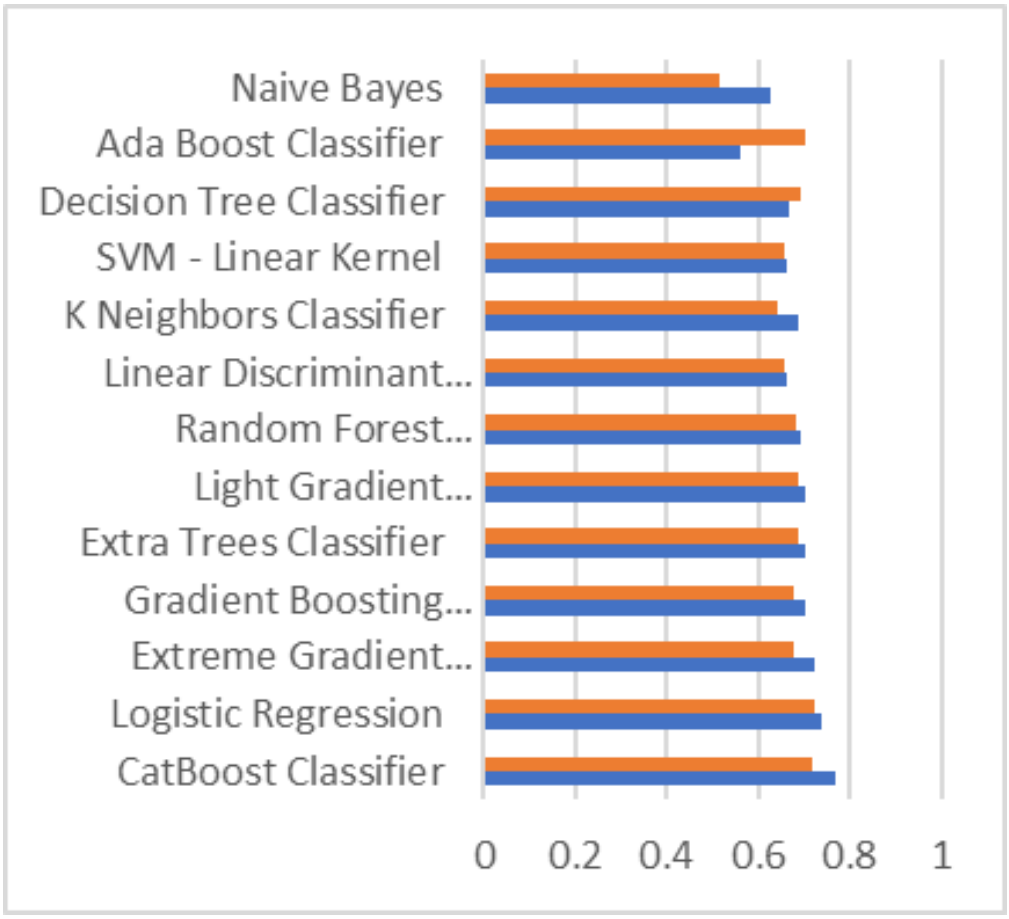


Figure 6.6 Comparison of kappa after hyper-parameter optimization.

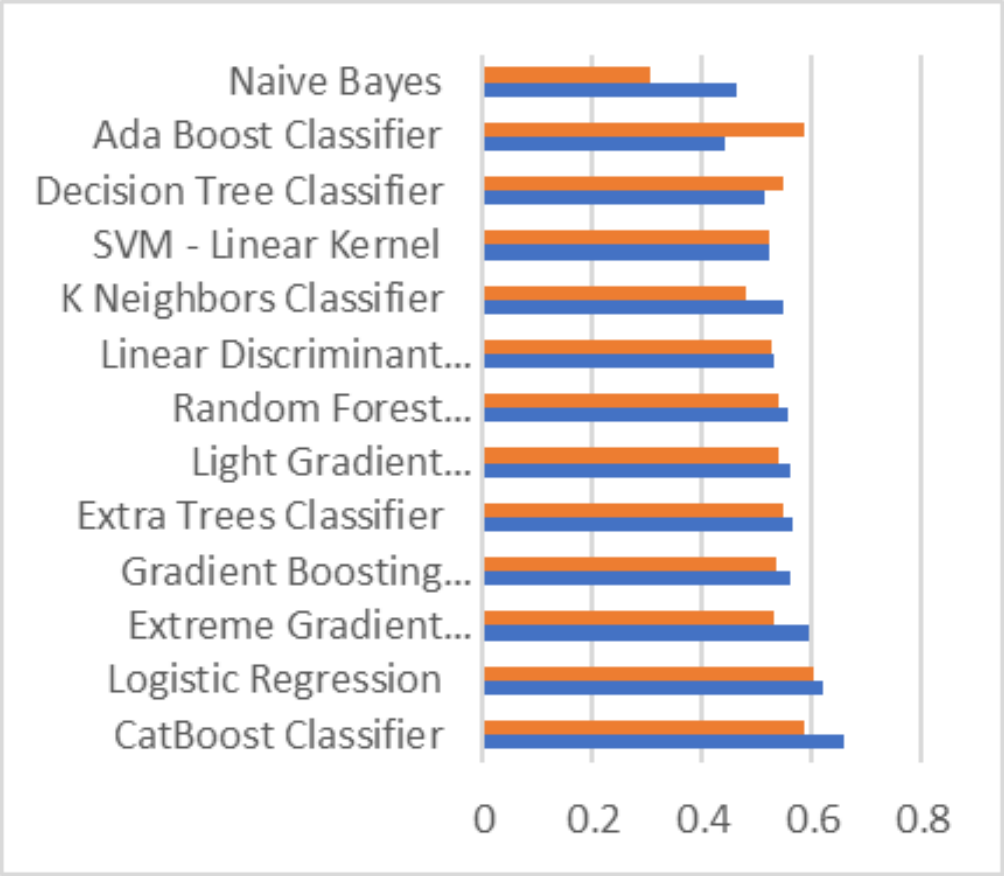


Figure 6.7 Comparison of F1-score after hyper-parameter optimization.

The best accuracy, recall, precision, F1 score, and kappa score are provided by the CatBoost classifier. We choose the CatBoost classification model to test against the confusion matrix. The confusion matrices of the CatBoost classifier are shown in Figures 6.8 and 6.9 both before and after hyperparameter adjustment. After hyperparameter optimization, a ROC curve is created against the CatBoost model's false-positive rate and true positive rate prediction scores, which indicate whether the model is operating incorrectly. AUC scores for each of the output parameter's five classes were seen to be nearly one and in the significantly good range. The CatBoost classifier's ROC curve with AUC values is displayed in Figure 6.10. This CatBoost classifier model is trained and tested once more for roughly 2500 examples following hyperparameter tweaking. Once more, stratified k-fold cross-validation is employed, with a 74% accuracy rate. The accuracy variance is indicated by the darkened region around the line. The cross-validation and training scores are validated in Figure 6.11. Since the current model is a

multinomial decision-making problem, Figure 6.8 combines these categorization, decision-making, and ensemble procedures. The model is validated using particular performance metrics like the AUC and confusion matrix. The performance of the model is typically verified using the area under the curve (AUC). Measures like the true-positive rate (TPR) and false-positive rate are used to determine AUC (FPR). Using TPR and FPR calculations, a confusion matrix is a useful tool for identifying true positives, true negatives, false positives, and false negatives. The graph typically tends to fall below the zone of operation (which is a diagonal line from the origin) if the FPR is higher, rendering the model useless. Generally speaking, the AUC score ought to be higher than 0.5, above the diagonal line. The TPR would be larger, though, if the model finds the best fit. Since the majority of the data in this situation are categorical, CatBoost is seen to have a higher TPR.

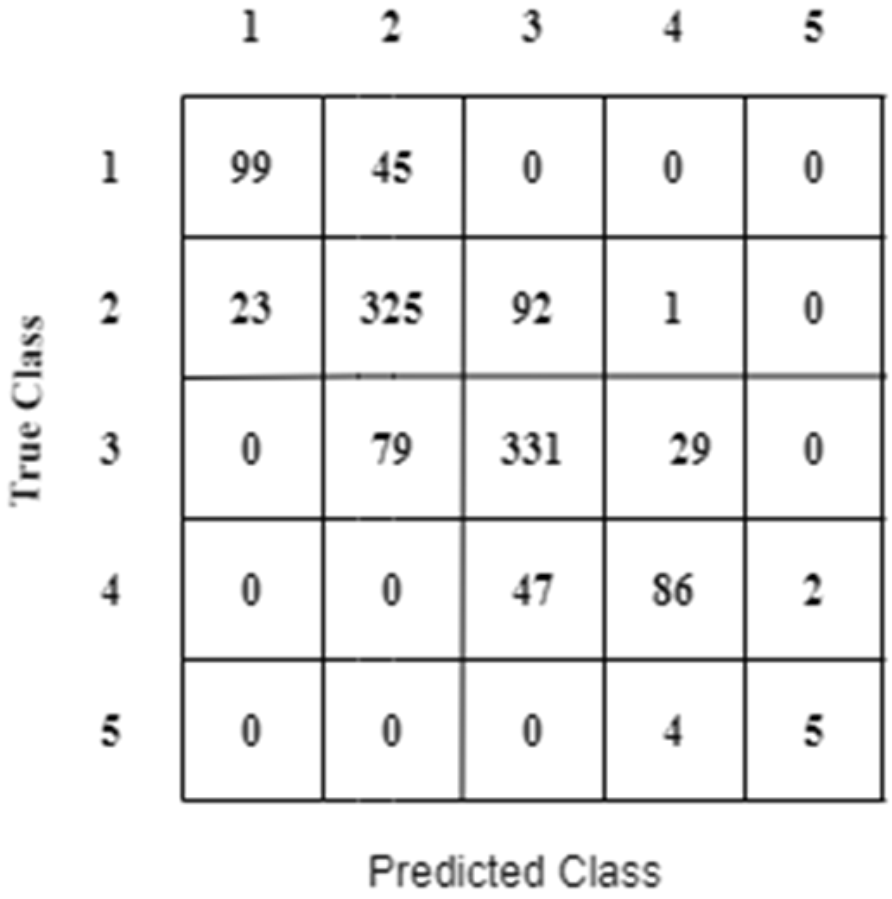


Figure 6.8 Confusion matrices of CatBoost classifier before Hyper-parameter tuning.

True Class \ Predicted Class	1	2	3	4	5
1	90	54	0	0	0
2	13	347	80	1	0
3	0	66	367	6	0
4	0	0	68	67	0
5	0	0	0	8	1

Figure 6.9 Confusion matrices of CatBoost classifier after Hyper-parameter tuning.

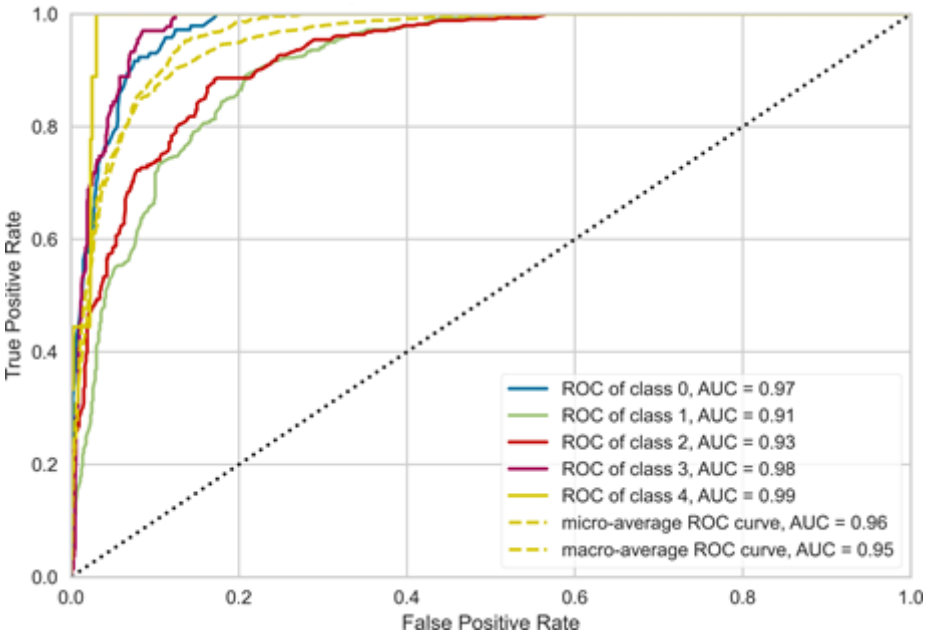


Figure 6.10 ROC curve for CatBoost classifier with AUC scores.

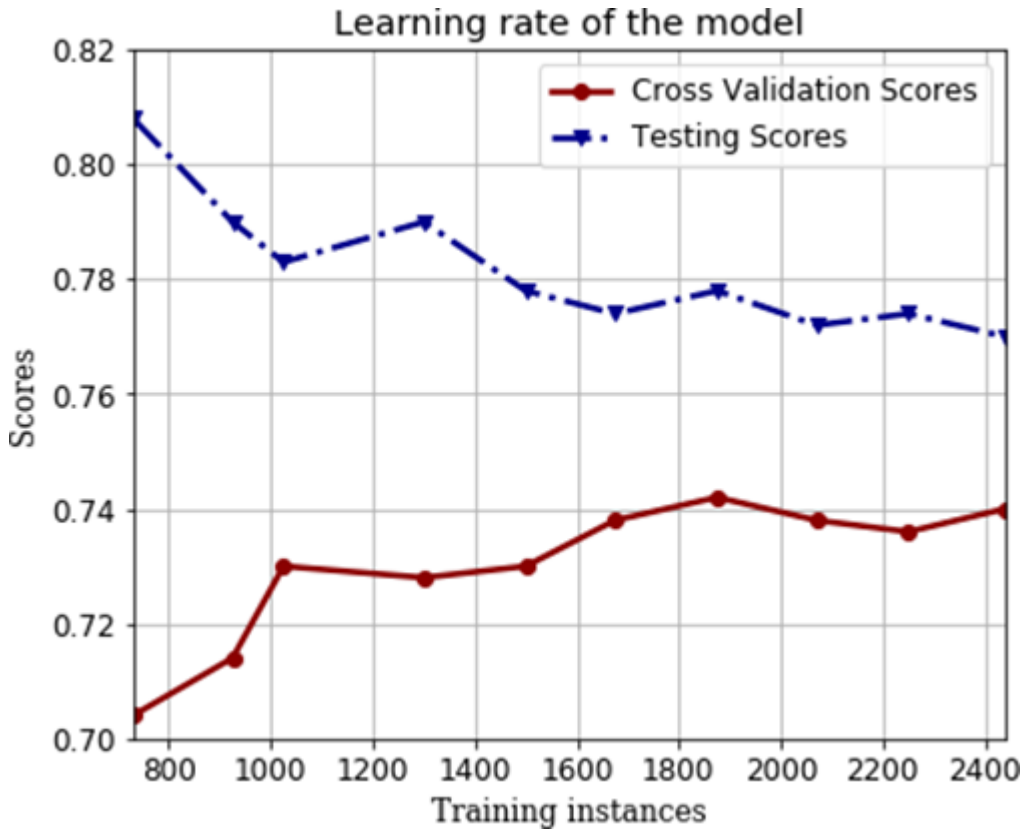


Figure 6.11 Validation of training and cross-validation scores.

Table 6.5 Top Three Best Performing Models After Tuning Hyper-Parameter

S.No.	Model	Accuracy	Recall	Precision	F1 score	Kappa score
1	CatBoost classifier	0.7582	0.64	0.772	0.767	0.663
2	Logistic Regression	0.751	0.57	0.753	0.731	0.631
3	XGboost classifier	0.7471	0.55	0.727	0.721	0.591

Table 6.5 lists the top three models in terms of accuracy, recall, precision, F1 score, and kappa score after modifying hyper-parameters. The phrase "hyper-parameters" may be used while modelling a machine learning model. The model is trained over an algorithm, but hyper-parameters are not altered throughout that process; instead, the model is verified based on its correctness. As a result, these hyper-parameters are happy to take advantage of new opportunities to improve the present machine learning model. Modeling just takes accuracy into account, despite the fact that several other variables, including as accuracy, recall, precision, F1 score, kappa score, and AUC, may also have an impact on how well a model performs. Hyper-parameters therefore help to improve

performance. The performance of the machine learning model is improved overall while maintaining the present accuracy through hyper-parameter adjustment. Applying an optimization problem on top of the current model will help find the optimal hyper-parameter. In order to investigate the suitable parameters and raise the AUC and accuracy scores for the given situation, a randomised grid search is chosen. The CatBoost classifier, which uses gradient boosting methods to make decisions, is now the best in its field. It is an ensemble technique like extra trees, AdaBoost, and XGBoost classifier, but because it sometimes needs categorical information, it provides substantially better parametric identification. Values from input and output category categories are combined across the full dataset in the current investigation. the problem is modified to the particular use case for CatBoost after grid searching the essential hyper-parameters like learning rate (0.05) and depth (6). Other hyper-parameters were left at their default settings since they were either machine- or data-specific. The performance of the model was superior to that of the other methods after the algorithm had been implemented with the revised hyper-parameters. For binary classification, the logistic regression approach is typically preferred. Solver, penalty, C, and max iterations are the hyper-parameters for logistic regression classifiers. The effectiveness of the model is greatly influenced by these parameters. Sklearn provides five alternatives for solver and saga for this job (stochastic average gradient descent with L1 regularisation). The penalty's severity, denoted by the letter C, is 7.0028. Since the data are not particularly large, the default setting for the maximum iterations was chosen. The following hyper-parameters for the XGBoost gradient boosting technique in the decision-making process are found after grid searching: min child weight = 7, max depth = 6, learning rate = 0.1, gamma = 0.4, and sample tree = 0.5. Although XGBoost has additional parameters, the AUC and accuracy were the most affected.

CHAPTER 7: CONCLUSION

While exploring CALERIE study's dataset to see whether the dataset can be used for tracking adaptive thermogenesis. A lot of unknown questions were raised like missing data problems, and a lot of continuous data provided unsatisfactory results. Above all, it has been observed that only non-invasively trackable parameters were preferred to study the relationships with adaptive thermogenesis. Meanwhile, the same data was used for fat mass prediction where few participants' data were missing. The model performed well to reach up to 0.986 R-square scores along with a cross-validation score of 0.978 with the Lasso Regression technique.

Then after EBM and LR analysis was performed to model the data for adaptive thermogenesis. Which scored insignificantly therefore a feature scaling technique is deployed to scale the AT value from continuous to discrete. Then the next segment of the research was to prevent the chances of adaptive thermogenesis. Through extensive literature review, it has been identified that insulin resistance was the primary source of body weight gain. A few techniques were identified that were used to measure insulin resistance in a human.

The parameters for these techniques were available in the CALERIE study's dataset, which was scaled between 0,1, and 2. Since the target variable is converted to classes, classification algorithms were employed on all the parameters but models built on c-peptide performed better. Naïve Bayes and logistic regression were appropriate techniques as they produced over 0.84 accuracies with AUC scores of 0.78.

To validate the models created using various resources and techniques associating statistical and machine learning methods a synthetic dataset is created based on the original CALERIE dataset with 87% efficiency of the model. This synthetic dataset is used as testing for models created earlier. This could prove whether the models created were at any use if provided on a scale of 100000 individuals.

Limitations such as 78.46% probability to identify an individual with Adaptive thermogenesis with all non-invasively trackable parameters. Where body

temperature, alcohol, and protein consumption played a major role. When this model is tested with synthetically generated data AT classification achieved an accuracy of 0.66372. The limitations of correctly classifying an individual with insulin resistance have a probability of 84.37% with an AUC score of 0.58. which can be further improved with more data. When the Insulin Resistance model is tested with synthetically generated data a model accuracy of 0.7827 is achieved.

Connecting both the models to enhance the overall promise of identification and prevention of adaptive thermogenesis with just a few non-invasively trackable parameters. The future scope of this research is to understand the degree of adaptive thermogenesis through continuous data monitoring. With the development of these models, it can be a substitute for invasive techniques but still has its limitations in being correct all the time.

No clinical report has shown that people with diabetes had a higher vulnerability to COVID-19, despite the fact that diabetes was linked to worse results in COVID-19 patients. Humans are twice as likely to die from diabetic complications, according to a few recent studies. Patients with diabetes experienced an approximately threefold increase in overall COVID-19 mortality in China from January to April 2020. This work proposes the COVID-19 risk prediction model for diabetic patients using a fuzzy inference system and machine learning approaches to estimate the risk level of COVID-19 in diabetic patients, making it possible for timely action. This is due to the multifold mortality rate of COVID-19 in diabetic patients. The suggested model lessens the need for medical guidance from specialists who treat COVID-19 patients and estimate the risk level of COVID19 for diabetic patients. The suggested model uses eight input variables, including age, sex, and recent travel history, which were discovered to be the most significant symptoms in diabetic patients who contracted COVID-19. These eight variables are fever, cough, sore throat, cardiovascular disease, high blood pressure, and age. Eight input factors, one output, and a total of 3,888 (3,3,3,3,3,4,2,2) rules were used to create the COVID-19 risk level for diabetes patients, which is five in number. The risk levels range from 1 to 5, with 5 being the highest. The following fifteen models

were created using cutting-edge machine learning methods: logistic regression, AdaBoost, CatBoost, gradient boosting, random forest, extreme gradient boosting, extra trees, light gradient boosting machine, decision tree, linear discriminant analysis, K-neighbors, SVM-linear kernel, ridge, naive Bayes, and quadratic discriminant analysis. The accuracy, recall, precision, F1 score, and kappa score are all highest for the CatBoost classifier. The CatBoost classifier demonstrated 76% accuracy following hyperparameter adjustment and improvements in the recall, precision, F1 score, and kappa score. Logistic regression and XGBoost came in second and third, respectively, with 75.1% and 74.7% accuracy. For validation, stratified k-fold cross-validation was employed. Although the knowledge base of the fuzzy inference system has some useful insights, accuracy and precision can be increased by using the actual medical record. Better synthetic data creation methods can reduce variance fluctuations and the slight bias of being entirely naive, avoiding the need for hyper-parameter optimization.

REFERENCES

[1] V. Ormazabal, S. Nair, O. Elfeky, C. Aguayo, C. Salomon, and F. A. Zuñiga, "Association between insulin resistance and the development of cardiovascular disease," *Cardiovasc. Diabetol.*, vol. 17, no. 1, pp. 1–14, 2018, doi: 10.1186/s12933-018-0762-4.

[2] P. Saeedi *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 157, p. 107843, 2019, doi: 10.1016/j.diabres.2019.107843.

[3] Freeman AM, Pennings N. Insulin Resistance. [Updated 2020 Jul 10]. "In: StatPearls [Internet]," *Treasure Island (FL): StatPearls Publishing*; 2020 Jan. [online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK507839/ [Accessed october 5, 2020]

[4] O. O. Woolcott and R. N. Bergman, "Relative Fat Mass as an estimator of whole-body fat percentage among children and adolescents: A cross-sectional study using NHANES," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, 2019, doi: 10.1038/s41598-019-51701-z.

[5] "Global report on diabetes," WHO, ISBN 978 92 4 156525 7, 2016. https://www.who.int/diabetes/global-report/en/ (accessed March 1, 2020)

[6] X. Ren *et al.*, "Association between triglyceride to HDL-C Ratio (TG/HDL-C) and insulin resistance in chinese patients with newly diagnosed type 2 diabetes mellitus," *PLoS One*, vol. 11, no. 4, pp. 1–13, 2016, doi: 10.1371/journal.pone.0154345.

[7] N. A. K. Z. Iwani *et al.*, "Triglyceride to HDL-C Ratio is Associated with Insulin Resistance in Overweight and Obese Children," *Sci. Rep.*, vol. 7, no. August 2016, pp. 1–7, 2017, doi: 10.1038/srep40055.

[8] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Comput. Sci.*, vol. 47, no. C, pp. 45–51, 2015, doi: 10.1016/j.procs.2015.03.182.

[9] Z. Tafa, N. Pervetica, and B. Karahoda, "An intelligent system for diabetes prediction," *Proc. - 2015 4th Mediterr. Conf. Embed. Comput. MECO 2015 - Incl. ECyPS 2015, BioEMIS 2015, BioICT 2015, MECO-Student Chall. 2015*, pp. 378–382, 2015, doi: 10.1109/MECO.2015.7181948.

[10] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques," *Procedia Comput. Sci.*, vol. 112, pp. 2519–2528, 2017, doi: 10.1016/j.procs.2017.08.193.

[11] A. Negi and V. Jaiswal, "A first attempt to develop a diabetes prediction method based on different global datasets," *2016 4th Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2016*, pp. 237–241, 2016, doi: 10.1109/PDGC.2016.7913152.

[12] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, and L. Burattini, "TyG-er: An ensemble Regression Forest approach for identification of clinical factors related to insulin resistance condition using Electronic Health Records," *Comput. Biol. Med.*, vol. 112, no. June, p. 103358, 2019, doi: 10.1016/j.compbiomed.2019.103358.

[13] N. Yuvaraj and K. R. SriPreethaa, “Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster,” *Cluster Comput.*, vol. 22, 2019, doi: 10.1007/s10586-017-1532-x.

[14] E.O. Olaniyi, K. Adnan, "Onset diabetes diagnosis using artificial neural network," *Int. J. Sci. Eng*, Res. 5 754-759, 2014.

[15] Z. Soltani and A. Jafarian, “A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 6, pp. 89–94, 2016, doi: 10.14569/ijacsa.2016.070611.

[16] A. Sarwar and V. Sharma, “Comparative analysis of machine learning techniques in prognosis of type II diabetes,” *AI Soc.*, vol. 29, no. 1, pp. 123–129, 2014, doi: 10.1007/s00146-013-0456-0.

[17] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records,” *Sci. Rep.*, vol. 6, no. January, pp. 1–10, 2016, doi: 10.1038/srep26094.

[18] T. Pham, T. Tran, D. Phung, and S. Venkatesh, “Predicting healthcare trajectories from medical records: A deep learning approach,” *J. Biomed. Inform.*, vol. 69, pp. 218–229, 2017, doi: 10.1016/j.jbi.2017.04.001.

[19] A. Askarzadeh and A. Rezazadeh, “Artificial neural network training using a new efficient optimization algorithm,” *Appl. Soft Comput. J.*, vol. 13, no. 2, pp. 1206–1213, 2013, doi: 10.1016/j.asoc.2012.10.023.

[20] N. MadhuSudana Rao, K. Kannan, X. zhi Gao, and D. S. Roy, “Novel classifiers for intelligent disease diagnosis with multi-objective parameter evolution,” *Comput. Electr. Eng.*, vol. 67, pp. 483–496, 2018, doi: 10.1016/j.compeleceng.2018.01.039.

[21] A. Jafarian and P. Rahimloo, “Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them,” *Bull. la Société R. des Sci. Liège*, pp. 1148–1164, 2016, doi: 10.25518/0037-9565.5938.

[22] N. S. Gill and P. Mittal, “A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease,” *J. Theor. Appl. Inf. Technol.*, vol. 87, no. 1, pp. 1–10, 2016.

[23] S. Joshi and M. Borse, “Detection and prediction of diabetes mellitus using back-propagation neural network,” *Proc. - 2016 Int. Conf. Micro-Electronics Telecommun. Eng. ICMETE 2016*, pp. 110–113, 2016, doi: 10.1109/ICMETE.2016.11.

[24] M. Maniruzzaman *et al.*, “Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm,” *Comput. Methods Programs Biomed.*, vol. 152, pp. 23–34, 2017, doi: 10.1016/j.cmpb.2017.09.004.

[25] R. Mirshahvalad and N. A. Zanjani, “Diabetes prediction using ensemble perceptron algorithm,” *Proc. - 9th Int. Conf. Comput. Intell. Commun. Networks, CICN 2017*, vol. 2018-January, pp. 190–194, 2018, doi: 10.1109/CICN.2017.8319383.

[26] X. Sun, X. Yu, J. Liu, and H. Wang, “Glucose prediction for type 1 diabetes

using KLMS algorithm," *Chinese Control Conf. CCC*, vol. 2, no. 2, pp. 1124–1128, 2017, doi: 10.23919/ChiCC.2017.8027498.

[27] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[28] A. Ashiquzzaman *et al.*, "Reduction of overfitting in diabetes prediction using deep learning neural network," *Lect. Notes Electr. Eng.*, vol. 449, pp. 35–43, 2017, doi: 10.1007/978-981-10-6451-7_5.

[29] G. Swapna, K. P. Soman, and R. Vinayakumar, "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1253–1262, 2018, doi: 10.1016/j.procs.2018.05.041.

[30] A. Mohebbi, T. B. Aradottir, A. R. Johansen, H. Bengtsson, M. Fraccaro, and M. Morup, "A deep learning approach to adherence detection for type 2 diabetics," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 2896–2899, 2017, doi: 10.1109/EMBC.2017.8037462.

[31] M. Nirmaladevi, S. A. Alias Balamurugan, and U. V. Swathi, "An amalgam KNN to predict diabetes mellitus," *2013 IEEE Int. Conf. Emerg. Trends Comput. Commun. Nanotechnology, ICE-CCN 2013*, no. Iceccn, pp. 691–695, 2013, doi: 10.1109/ICE-CCN.2013.6528591.

[32] H. Gylling *et al.*, "Insulin sensitivity regulates cholesterol metabolism to a greater extent than obesity: Lessons from the METSIM study," *J. Lipid Res.*, vol. 51, no. 8, pp. 2422–2427, 2010, doi: 10.1194/jlr.P006619.

[33] A. G. Jones and A. T. Hattersley, "The clinical utility of C-peptide measurement in the care of patients with diabetes," *Diabet. Med.*, vol. 30, no. 7, pp. 803–817, 2013, doi: 10.1111/dme.12159.

[34] X. Zheng *et al.*, "A new model to estimate insulin resistance via clinical parameters in adults with type 1 diabetes," *Diabetes. Metab. Res. Rev.*, vol. 33, no. 4, 2017, doi: 10.1002/dmrr.2880.

[35] Y. Liu, "Artificial intelligence-based neural network for the diagnosis of diabetes: Model development," *JMIR Med. Informatics*, vol. 8, no. 5, 2020, doi: 10.2196/18682.

[36] D. Rodbard, "Continuous glucose monitoring: A review of recent studies demonstrating improved glycemic outcomes," *Diabetes Technol. Ther.*, vol. 19, pp. S25–S37, 2017, doi: 10.1089/dia.2017.0035.

[37] L. W. Turner, D. Nartey, R. S. Stafford, and S. Singh, "Page 1 of 25 Diabetes Care," pp. 1997–2012, 2012.

[38] E. Krishnan, B. J. Pandya, L. Chung, A. Hariri, and O. Dabbous, "Hyperuricemia in young adults and risk of insulin resistance, prediabetes, and diabetes: a 15-year follow-up study.," *Am. J. Epidemiol.*, vol. 176, no. 2, pp. 108–116, 2012, doi: 10.1093/aje/kws002.

[39] M. A. De Vries *et al.*, "Glucose-dependent leukocyte activation in patients with type 2 diabetes mellitus, familial combined hyperlipidemia and healthy controls," *Metabolism.*, vol. 64, no. 2, pp. 213–217, 2015, doi: 10.1016/j.metabol.2014.10.011.

[40] D. J. Lee, J. S. Choi, K. M. Kim, N. S. Joo, S. H. Lee, and K. N. Kim, "Combined effect of serum Gamma-glutamyltransferase and uric acid on Framingham risk score," *Arch. Med. Res.*, vol. 45, no. 4, pp. 337–342, 2014, doi: 10.1016/j.arcmed.2014.04.004.

[41] S. Riaz, "Study of protein biomarkers of diabetes mellitus type 2 and therapy with vitamin B1," *J. Diabetes Res.*, vol. 2015, 2015, doi: 10.1155/2015/150176.

[42] K.D. Pagana, T.J. Pagana, T.N. Pagana, "Mosby's Diagnostic & Laboratory Test Reference," *Mo: Elsevier,* 14th ed. St. Louis, 2019.

[43] K. Stawiski, I. Pietrzak, W. Młynarski, W. Fendler, and A. Szadkowska, "NIRCa: An artificial neural network-based insulin resistance calculator," *Pediatr. Diabetes*, vol. 19, no. 2, pp. 231–235, 2018, doi: 10.1111/pedi.12551.

[44] B. G. Choi, S. W. Rha, S. W. Kim, J. H. Kang, J. Y. Park, and Y. K. Noh, "Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks," *Yonsei Med. J.*, vol. 60, no. 2, pp. 191–199, 2019, doi: 10.3349/ymj.2019.60.2.191.

[45] B. Farran, R. AlWotayan, H. Alkandari, D. Al-Abdulrazzaq, A. Channanath, and T. A. Thanaraj, "Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait," *Front. Endocrinol. (Lausanne).*, vol. 10, 2019, doi: 10.3389/fendo.2019.00624.

[46] W. E. Kraus *et al.*, "2 years of calorie restriction and cardiometabolic risk (CALERIE): exploratory outcomes of a multicentre, phase 2, randomised controlled trial," *Lancet Diabetes Endocrinol.*, vol. 7, no. 9, pp. 673–683, 2019, doi: 10.1016/S2213-8587(19)30151-2.

[47] M. S. Udler, M. I. McCarthy, J. C. Florez, and A. Mahajan, "Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine," *Endocr. Rev.*, vol. 40, no. 6, pp. 1500–1520, 2019, doi: 10.1210/er.2019-00088.

[48] D. Balboa, R. B. Prasad, L. Groop, and T. Otonkoski, "Genome editing of human pancreatic beta cell models: problems, possibilities and outlook," *Diabetologia*, vol. 62, no. 8, pp. 1329–1336, 2019, doi: 10.1007/s00125-019-4908-z.

[49] M. I. Friedman and S. Appel, "Energy expenditure and body composition changes after an isocaloric ketogenic diet in overweight and obese men: A secondary analysis of energy expenditure and physical activity," *bioRxiv*, pp. 324–333, 2018, doi: 10.1101/383752.

[50] R. Klén, M. Karhunen, and L. L. Elo, "Likelihood contrasts: a machine learning algorithm for binary classification of longitudinal data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-57924-9.

[51] P. Patil and S. Shinde, "Performance Analysis of Different Classification Algorithms: Naïve Bayes, Decision Tree and K-Star," *J. Crit. Rev.*, vol. 7, no. 19, pp. 1160–1164, 2020.

[52] I. Verkouter *et al.*, "The association between adultweight gain and insulin resistance at middle age: Mediation by visceral fat and liver fat," *J. Clin. Med.*, vol. 8, no. 10, 2019, doi: 10.3390/jcm8101559.

[53] Y. Singh, M. K. Garg, N. Tandon, and R. K. Marwaha, “A Study of insulin resistance by HOMA-IR and its cut-off value to identify metabolic syndrome in urban Indian adolescents,” *JCRPE J. Clin. Res. Pediatr. Endocrinol.*, vol. 5, no. 4, pp. 245–251, 2013, doi: 10.4274/Jcrpe.1127.

[54] M. Murguía-Romero *et al.*, “Plasma triglyceride/HDL-cholesterol ratio, insulin resistance, and cardiometabolic risk in young adults,” *J. Lipid Res.*, vol. 54, no. 10, pp. 2795–2799, 2013, doi: 10.1194/jlr.M040584.

[55] W. C. Yeh, Y. C. Tsao, W. C. Li, I. S. Tzeng, L. S. Chen, and J. Y. Chen, “Elevated triglyceride-to-HDL cholesterol ratio is an indicator for insulin resistance in middle-aged and elderly Taiwanese population: A cross-sectional study,” *Lipids Health Dis.*, vol. 18, no. 1, pp. 1–7, 2019, doi: 10.1186/s12944-019-1123-3.

[56] H. A. Khan, S. H. Sobki, A. Ekhzaimy, I. Khan, and M. A. Almusawi, “Biomarker potential of C-peptide for screening of insulin resistance in diabetic and non-diabetic individuals,” *Saudi J. Biol. Sci.*, vol. 25, no. 8, pp. 1729–1732, 2018, doi: 10.1016/j.sjbs.2018.05.027.

[57] F. L. Greenway, “Physiological adaptations to weight loss and factors favouring weight regain,” *Int. J. Obes.*, vol. 39, no. 8, pp. 1188–1196, 2015, doi: 10.1038/ijo.2015.59.

[58] T. McGrath, K. G. Murphy, and N. S. Jones, “Quantitative approaches to energy and glucose homeostasis: Machine learning and modelling for precision understanding and prediction,” *J. R. Soc. Interface*, vol. 15, no. 138, 2018, doi: 10.1098/rsif.2017.0736.

[59] Ndahimana, D., & Kim, E.-K. (2017). Measurement Methods for Physical Activity and Energy Expenditure: a Review. *Clinical Nutrition Research*, *6*(2), 68. https://doi.org/10.7762/CNR.2017.6.2.68

[60] Ndahimana, D., Lee, S. H., Kim, Y. J., Son, H. R., Ishikawa-Takata, K., Park, J., & Kim, E. K. (2017). Accuracy of dietary reference intake predictive equation for estimated energy requirements in female tennis athletes and non-athlete college students: comparison with the doubly labeled water method. *Nutrition Research and Practice*, *11*(1), 51–56. https://doi.org/10.4162/NRP.2017.11.1.51

[61] Zhang, W.-S. (2010). Construction, calibration and testing of a decimeter-size heat-flow calorimeter. *Thermochimica Acta*, *499*, 128–132. https://doi.org/10.1016/j.tca.2009.11.013

[62] Schrack, J. A., Simonsick, E. M., & Ferrucci, L. (2010). Comparison of the Cosmed K4b2 Portable Metabolic System in Measuring Steady-State Walking Energy Expenditure. *PLOS ONE*, *5*(2), e9292. https://doi.org/10.1371/JOURNAL.PONE.0009292

[63] Lyden, K., Kozey, S. L., Staudenmeyer, J. W., & Freedson, P. S. (2011). A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. *European Journal of Applied Physiology*, *111*(2), 187–201. https://doi.org/10.1007/S00421-010-1639-8

[64] Hills, A. P., Mokhtar, N., & Byrne, N. M. (2014). Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. *Frontiers in Nutrition*, *1*. https://doi.org/10.3389/FNUT.2014.00005

[65] Welk, G. J., Differding, J. A., Thompson, R. W., Blair, S. N., Dziura, J., & Hart, P. (2000). The utility of the Digi-walker step counter to assess daily physical activity patterns. *Medicine and Science in Sports and Exercise*, *32*(9 Suppl). https://doi.org/10.1097/00005768-200009001-00007

[66] Greenway, F. L. (2015). Physiological adaptations to weight loss and factors favouring weight regain. *International Journal of Obesity (2005)*, *39*(8), 1188–1196. https://doi.org/10.1038/IJO.2015.59

[67] Tinsley, G. M., Moore, M. L., & Graybeal, A. J. (2018). Reliability of hunger-related assessments during 24-hour fasts and their relationship to body composition and subsequent energy compensation. *Physiology & Behavior*, *188*, 221–226. https://doi.org/10.1016/J.PHYSBEH.2018.02.017

[68] Martin, C. K., Nicklas, T., Gunturk, B., Correa, J. B., Allen, H. R., & Champagne, C. (2014). Measuring food intake with digital photography. *Journal of Human Nutrition and Dietetics : The Official Journal of the British Dietetic Association*, *27 Suppl 1*(0 1), 72–81. https://doi.org/10.1111/JHN.12014

[69] Williamson, D. A., Allen, H. R., Martin, P. D., Alfonso, A., Gerald, B., & Hunt, A. (2004). Digital photography: a new method for estimating food intake in cafeteria settings. *Eating and Weight Disorders : EWD*, *9*(1), 24–28. https://doi.org/10.1007/BF03325041

[70] Martin, C. K., Han, H., Coulon, S. M., Allen, H. R., Champagne, C. M., & Anton, S. D. (2009). A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method. *The British Journal of Nutrition*, *101*(3), 446–456. https://doi.org/10.1017/S0007114508027438

[71] Fontana, J. M., Higgins, J. A., Schuckers, S. C., Bellisle, F., Pan, Z., Melanson, E. L., Neuman, M. R., & Sazonov, E. (2015). Energy intake estimation from counts of chews and swallows. *Appetite*, *85*, 14. https://doi.org/10.1016/J.APPET.2014.11.003

[72] Bi, Y., Lv, M., Song, C., Xu, W., Guan, N., & Yi, W. (2016). AutoDietary: A Wearable Acoustic Sensor System for Food Intake Recognition in Daily Life. *IEEE Sensors Journal*, *16*(3), 806–816. https://doi.org/10.1109/JSEN.2015.2469095

[73] Kalantarian, H., Alshurafa, N., Le, T., & Sarrafzadeh, M. (2015). Monitoring eating habits using a piezoelectric sensor-based necklace. *Computers in Biology and Medicine*, *58*, 46–55. https://doi.org/10.1016/J.COMPBIOMED.2015.01.005

[74] Sun, M., Burke, L. E., Mao, Z. H., Chen, Y., Chen, H. C., Bai, Y., Li, Y., Li, C., & Jia, W. (2014). eButton: A Wearable Computer for Health Monitoring and Personal Assistance. *Proceedings. Design Automation Conference*, *2014*. https://doi.org/10.1145/2593069.2596678

[75] Jia, W., Chen, H. C., Yue, Y., Li, Z., Fernstrom, J., Bai, Y., Li, C., & Sun, M. (2014). Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera. *Public Health Nutrition*, *17*(8), 1671–1681. https://doi.org/10.1017/S1368980013003236

[76] Bell, B. M., Alam, R., Alshurafa, N., Thomaz, E., Mondol, A. S., de la Haye,

K., Stankovic, J. A., Lach, J., & Spruijt-Metz, D. (2020). Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review. *NPJ Digital Medicine*, *3*(1). https://doi.org/10.1038/S41746-020-0246-2

[77] Farooq, M., & Sazonov, E. (2016). A Novel Wearable Device for Food Intake and Physical Activity Recognition. *Sensors (Basel, Switzerland)*, *16*(7). https://doi.org/10.3390/S16071067

[78] Salley, J. N., Hoover, A. W., Wilson, M. L., & Muth, E. R. (2016). Comparison between Human and Bite-Based Methods of Estimating Caloric Intake. *Journal of the Academy of Nutrition and Dietetics*, *116*(10), 1568–1577. https://doi.org/10.1016/J.JAND.2016.03.007

[79] Dong, Y., Hoover, A., & Muth, E. (n.d.). *A Device for Detecting and Counting Bites of Food Taken by a Person During Eating*. Retrieved March 20, 2022, from www.isense.com

[80] Turner, L. W., Nartey, D., Stafford, R. S., & Singh, S. (2012). *Page 1 of 25 Diabetes Care*. 1997–2012.

[81] *Using Bite Counter for Weight Loss: A One-month Usability Trial to Test the Effectiveness of Using the Bite Counter - Full Text View - ClinicalTrials.gov*. (n.d.). Retrieved December 30, 2021, from https://clinicaltrials.gov/ct2/show/NCT02494674

[82] *A Clinical Study to Evaluate the Maximum Maxillary Bite Force (BF) When Using Two Novel Denture Adhesives Compared to Using No-Adhesive - Full Text View - ClinicalTrials.gov*. (n.d.). Retrieved March 20, 2022, from https://clinicaltrials.gov/ct2/show/NCT05173974

[83] Fontana, J. M., Farooq, M., & Sazonov, E. (2014). Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior. *IEEE Transactions on Bio-Medical Engineering*, *61*(6), 1772–1779. https://doi.org/10.1109/TBME.2014.2306773

[84] Vu, T., Lin, F., Alshurafa, N., & Xu, W. (2017). Wearable Food Intake Monitoring Technologies: A Comprehensive Review. *Computers 2017, Vol. 6, Page 4*, *6*(1), 4. https://doi.org/10.3390/COMPUTERS6010004

[85] Chung, J., Chung, J., Oh, W., Yoo, Y., Lee, W. G., & Bang, H. (2017). A glasses-type wearable device for monitoring the patterns of food intake and facial activity. *Scientific Reports 2017 7:1*, *7*(1), 1–8. https://doi.org/10.1038/srep41690

[86] Bedri, A., Li, R., Haynes, M., Kosaraju, R. P., Grover, I., Prioleau, T., Beh, M. Y., Goel, M., Starner, T., & Abowd, G. (2017). EarBit. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–20. https://doi.org/10.1145/3130902

[87] Martin, A. R. (n.d.). *Adaptive Thermogenesis and Metabolic Changes Following Diet- and Exercise- Induced Weight Loss*. Retrieved March 23, 2022, from http://digitalcommons.unl.edu/nutritiondiss/74

[88] Kaushik, B. K., & Majumder, M. K. (2015). Carbon Nanotube Based VLSI Interconnects: Analysis and Design. *SpringerBriefs in Applied Sciences and Technology*, 1–86. https://doi.org/10.1007/978-81-322-2047-3

[89] Kumar, B., Kaushik, B. K., & Negi, Y. S. (2014). Organic Thin Film Transistors: Structures, Models, Materials, Fabrication, and Applications: A Review. *Https://Doi.Org/10.1080/15583724.2013.848455*, *54*(1), 33–111. https://doi.org/10.1080/15583724.2013.848455

[90] Bimo Prakoso, A., Wang, J., -, al, Narayan Mishra, K., Kumar, S., Patel -, N. R., Salman, A. J., Al-Jawad, M., Al Tameemi -, W., Goyal, S. B., Bedi, P., Kumar Yadav, D., & Vakil, N. A. (2021). Internet of Things Information Analysis using Fusion based Learning with Deep Neural Network. *Journal of Physics: Conference Series*, *1714*(1), 012022. https://doi.org/10.1088/1742-6596/1714/1/012022.

[91] Bedi, P., Goyal, S. B., Sharma, R., Yadav, D. K., & Sharma, M. (2021). Smart Model for Big Data Classification Using Deep Learning in Wireless Body Area Networks. *Lecture Notes in Networks and Systems*, *179 LNNS*, 215–224. https://doi.org/10.1007/978-981-33-4687-1_21K.

[92] Alam, T., Qamar, S., Dixit, A., & Benaida, M. (n.d.). *Genetic Algorithm: Reviews, Implementations, and Applications*.

[93] Aggarwal, S., Tomar, S. K., & Aggarwal, A. (2012). On challenges and opportunities in second wave of ICT revolution for south Asian countries. *Proceedings of 2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, PDGC 2012*, 597–602. https://doi.org/10.1109/PDGC.2012.6449888

[94] Yadav, D., Akanksha, & Yadav, A. K. (2020). A novel convolutional neural network based model for recognition and classification of apple leaf diseases. *Traitement Du Signal*, *37*(6), 1093–1101. https://doi.org/10.18280/TS.370622

[95] Bhatia, M., Bansal, A., & Yadav, D. (2017). A proposed quantitative approach to classify brain MRI. *International Journal of Systems Assurance Engineering and Management*, *8*, 577–584. https://doi.org/10.1007/S13198-016-0465-8

[96] Lal, N., Shiwani, S., & Qamar, S. (2016). Information Retrieval System and challenges with Dataspace Information Retrieval System and challenges with Dataspace of Saudi Arabia. *Article in International Journal of Computer Applications*, *147*(8), 975–8887. https://doi.org/10.5120/ijca2016911128

[97] Verma, R., Sehgal, V. K., & Nitin. (2016). Implementation of information system in crisis management using modeling and simulation. *International Journal of Simulation: Systems, Science and Technology*, *17*(32). https://doi.org/10.5013/IJSSST.A.17.32.12

[98] Gill, H. K., Sehgal, V. K., & Verma, A. K. (2021). A deep neural network based context-aware smart epidemic surveillance in smart cities. *Library Hi Tech*. https://doi.org/10.1108/LHT-02-2021-0063/FULL/XML

[99] Rangra, A., Sehgal, V. K., & Shukla, S. (1 C.E.). A Novel Approach of Cloud Based Scheduling Using Deep-Learning Approach in E-Commerce Domain. *Https://Services.Igi-Global.Com/Resolvedoi/Resolve.Aspx?Doi=10.4018/IJISMD.2019070104*, *10*(3), 59–75. https://doi.org/10.4018/IJISMD.2019070104.

[100] Verma, R., Sehgal, V. K., & Nitin. (2016). Implementation of control measure in the crisis using social networks. *International Journal of Simulation:*

Systems, Science and Technology, *17*(32). https://doi.org/10.5013/IJSSST.A.17.32.11

[101] Xie, M. ge, & Zheng, Z. (2022). Homeostasis phenomenon in conformal prediction and predictive distribution functions. *International Journal of Approximate Reasoning*, *141*, 131–145. https://doi.org/10.1016/J.IJAR.2021.09.001

[102] Alonso-Bastida, A., Adam-Medina, M., Posada-Gómez, R., Salazar-Piña, D. A., Osorio-Gordillo, G. L., & Vela-Valdés, L. G. (2022). Dynamic of Glucose Homeostasis in Virtual Patients: A Comparison between Different Behaviors. *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 716*, *19*(2), 716. https://doi.org/10.3390/IJERPH19020716

[103] Veen, L. van, Morra, J., Palanica, A., & Fossat, Y. (2020). Homeostasis as a proportional–integral control system. *Npj Digital Medicine 2020 3:1*, *3*(1), 1–7. https://doi.org/10.1038/s41746-020-0283-x

[104] Horie, T., Nakao, T., Miyasaka, Y., Nishino, T., Matsumura, S., Nakazeki, F., Ide, Y., Kimura, M., Tsuji, S., Rodriguez, R. R., Watanabe, T., Yamasaki, T., Xu, S., Otani, C., Miyagawa, S., Matsushita, K., Sowa, N., Omori, A., Tanaka, J., … Ono, K. (2021). microRNA-33 maintains adaptive thermogenesis via enhanced sympathetic nerve activity. *Nature Communications 2021 12:1*, *12*(1), 1–17. https://doi.org/10.1038/s41467-021-21107-5

[105] "Coronavirus disease (covid-19) situation reports." [Online]. Available: https://www.who.int/emergencies/diseases/ novel-coronavirus-2019/situation-reports

[106] S. Punitha, F. Al-Turjman, and T. Stephan, "Genetically optimized computer-aided diagnosis for detection and classification of covid-19," *AI-Powered IoT for COVID-19*, pp. 105–122, 2020.

[107] V. Hespanhol and C. Ba´rbara, "Pneumonia mortality, comorbidities matter?" *Pulmonology*, vol. 26, no. 3, pp. 123–129, 2020.

[108] Q. Zou, S. Zheng, X. Wang, S. Liu, J. Bao, F. Yu, W. Wu, X. Wang, B. Shen, T. Zhou, *et al.*, "Influenza a-associated severe pneumonia in hospitalized patients: Risk factors and nai treatments," *International Journal of Infectious Diseases*, vol. 92, pp. 208–213, 2020.

[109] A. Zumla, D. S. Hui, and S. Perlman, "Middle east respiratory syn- drome," *The Lancet*, vol. 386, no. 9997, pp. 995–1007, 2015.

[110] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, *et al.*, "Clinical characteristics of coronavirus disease 2019 in china," *New England journal of medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.

[111] J.-j. Zhang, X. Dong, Y.-y. Cao, Y.-d. Yuan, Y.-b. Yang, Y.-q. Yan, C. A. Akdis, and Y.-d. Gao, "Clinical characteristics of 140 patients infected with sars-cov-2 in wuhan, china," *Allergy*, 2020.

[112] J. Yang, Y. Zheng, X. Gou, K. Pu, Z. Chen, Q. Guo, R. Ji, H. Wang, Y. Wang, and Y. Zhou, "Prevalence of comorbidities in the novel wuhan coronavirus (covid-19) infection: a systematic review and meta- analysis," *International journal of infectious diseases*, 2020.

[113] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, *et al.*, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china," *Jama*, vol. 323, no. 11, pp. 1061–1069, 2020.

[114] Q. Ruan, K. Yang, W. Wang, L. Jiang, and J. Song, "Clinical predictors of mortality due to covid-19 based on an analysis of data of 150 patients from wuhan, china," *Intensive care medicine*, vol. 46, no. 5, pp. 846– 848, 2020.

[115] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, *et al.*, "Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study," *The lancet*, 2020.

[116] C. Wu, X. Chen, Y. Cai, X. Zhou, S. Xu, H. Huang, L. Zhang, X. Zhou, C. Du, Y. Zhang, *et al.*, "Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in wuhan, china," *JAMA internal medicine*, 2020.

[117] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: sum- mary of a report of 72 314 cases from the chinese center for disease control and prevention," *Jama*, vol. 323, no. 13, pp. 1239–1242, 2020.

[118] C. Leung, "Clinical features of deaths in the novel coronavirus epidemic in china," *Reviews in Medical Virology*, p. e2103, 2020.

[119] B. Li, J. Yang, F. Zhao, L. Zhi, X. Wang, L. Liu, Z. Bi, and Y. Zhao, "Prevalence and impact of cardiovascular metabolic diseases on covid-19 in china," *Clinical Research in Cardiology*, vol. 109, no. 5, pp. 531–538, 2020.

[120] A. Remuzzi and G. Remuzzi, "Covid-19 and italy: what next?" *The Lancet*, 2020.

[121] C. Iwendi, K. Mahboob, Z. Khalid, et al., "Classification of COVID-19 individuals using adaptive neuro-fuzzy inference system," *Multimedia Systems*, pp. 1-15, 2021, doi: https://doi.org/10.1007/s00530-021-00774-w.

[122] C. Iwendi, AK Bashir, A. Peshkar, R. Sujatha, JM Chatterjee, S. Pasupuleti, R. Mishra S. Pillai and O. Jo, "COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm," *Front. Public Health*, 8:357, 2020, doi: 10.3389/fpubh.2020.00357.

[123] Y. W. Kerk, C. Y. Teh, K. M. Tay and C. P. Lim, "Parametric Conditions for a Monotone TSK Fuzzy Inference System to be an n-Ary Aggregation Function," in *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 7, pp. 1864-1873, July 2021, doi: 10.1109/TFUZZ.2020.2986986.

[124] T. Stephan and K. Suresh Joseph, "Particle swarm optimization-based energy efficient channel assignment technique for clustered cognitive radio sensor networks," *The Computer Journal*, vol. 61, no. 6, pp. 926– 936, 2018.

[125] T. Stephan, K. Sharma, A. Shankar, S. Punitha, V. Varadarajan, and P. Liu, "Fuzzy-logic-inspired zone-based clustering algorithm for wire- less sensor networks," *International Journal of Fuzzy Systems*, pp. 1–12, 2020.

[126] T. Stephan, A. Rajappa, K. Sendhil Kumar, S. Gupta, A. Shankar, and V. Vijayakumar, "Modified fuzzy-based greedy routing protocol for vanets,"

Journal of Intelligent & Fuzzy Systems, no. Preprint, pp. 1–8.

[127] A. Gilchrist, "5 acute cough types and how to treat them appropriately." [Online]. Available: https://www.contemporaryclinic. com/view/5-acute-cough-types-and-how-to-treat-them-appropriately

[128] S. Dalal and D. S. Zhukovsky, "Pathophysiology and management of fever." *The journal of supportive oncology*, vol. 4, no. 1, pp. 9–16, 2006.

[129] B. Sissons, "Common cold: Stage by stage." [Online]. Available: https://www.medicalnewstoday.com/articles/327348

[130] "Heart failure: Understanding heart failure management and treatment." [Online]. Available: https://my.clevelandclinic. org/health/diseases/17069-heart-failure-understanding-heart-failure/ management-and-treatment.

[131] N. LeBrun, "Types of blood pressure problems," Nov 2019. [Online]. Available: https://www.healthgrades.com/right-care/ high-blood-pressure/types-of-blood-pressure-problem.

[132] "Covid-19: overview and resources – global health 50/50." [Online]. Available: https://globalhealth5050.org/COVID19/

APPENDIX

Application Interface Designed for Data Analysis and Modelling

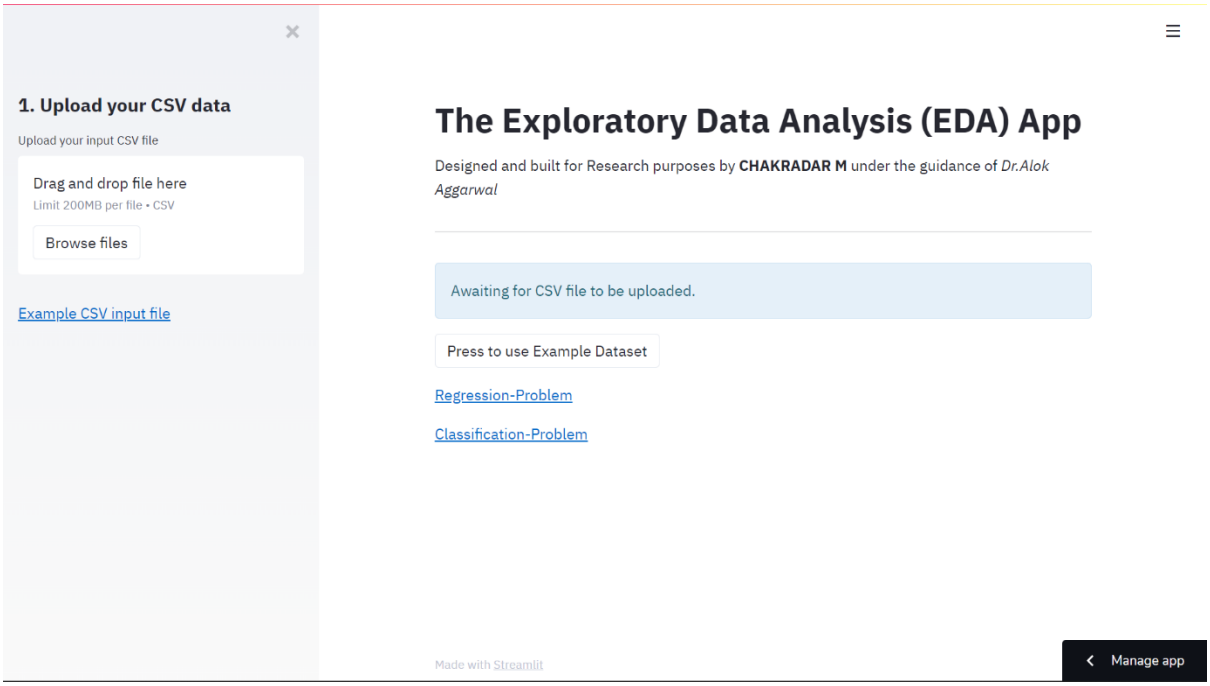


Figure 1 Exploratory data analysis (EDA) Home page

In the course of data filtering and slicing over such a large dataset as the CALERIE study. The initial pain of finding insights from unfiltered data at the start was genuinely challenging. So, an easy tool to scatter a dataset into numerous charts and the statistical properties of each element and most importantly to find whether that data is useful should save a lot of time. With these intentions, an effort was made to put forth a statistical profiling tool, where one could push the dataset in and extract fruitful insights based on the data distributions as shown in figures 1-4.

Figure 1 is the home page which presents exploratory data analysis and why it was made and on the left-hand side, one can find upload your CSV data. One could drag and drop as well and in the center, there is an option that says “Press to use Example Dataset”. In the initial phase in order to learn the whereabouts of the application one could make use of this feature. After exploring the dataset using this tool as shown in figures 2-4 one could form a problem statement and

understand the pathways to solve the problem. This opens up a parameter of interest that had to predict based on the other data points in the dataset. This parameter of interest opens up an opportunity for an approach to the problem. In figure 2 we can see an example dataset of the flower classification problem where the last column decides the flower. This data is in a fairly continuous type format so one could approach a regression approach. But if examined carefully the problem we are solving is the classification of flowers amongst 5 different flowers.

Therefore, a feature engineering approach proceeds to scale down these flowers into 5 distinct classifications. Based on the findings from figures 3 and 4 one prefers any approach and showcase their prediction rate based on any of the two approaches at the bottom of the application. If it is a continuous term then click on Regression-Problem, else click on the Classification-Problem as shown in figure 1.

The choice of development for this tool is python 3.8 and in figure 3 a data profiling library was used which is called Pandas. This library reads the data in the form of multidimensional data frames. Which enables us to establish a relationship individually and in combinations. This technique provided the research to scavenge insights from complex datasets and enhance user understandability. Every parameter is statistically graphed with their distribution along with their characteristic type. For suppose if the data point talks about the participant's gender then this data point would be binary either 0 or 1. In the case of explaining the day of the week, one could just pick numbers from 1-7 if the interest is purely focused on the day of the week irrespective of the date and time.

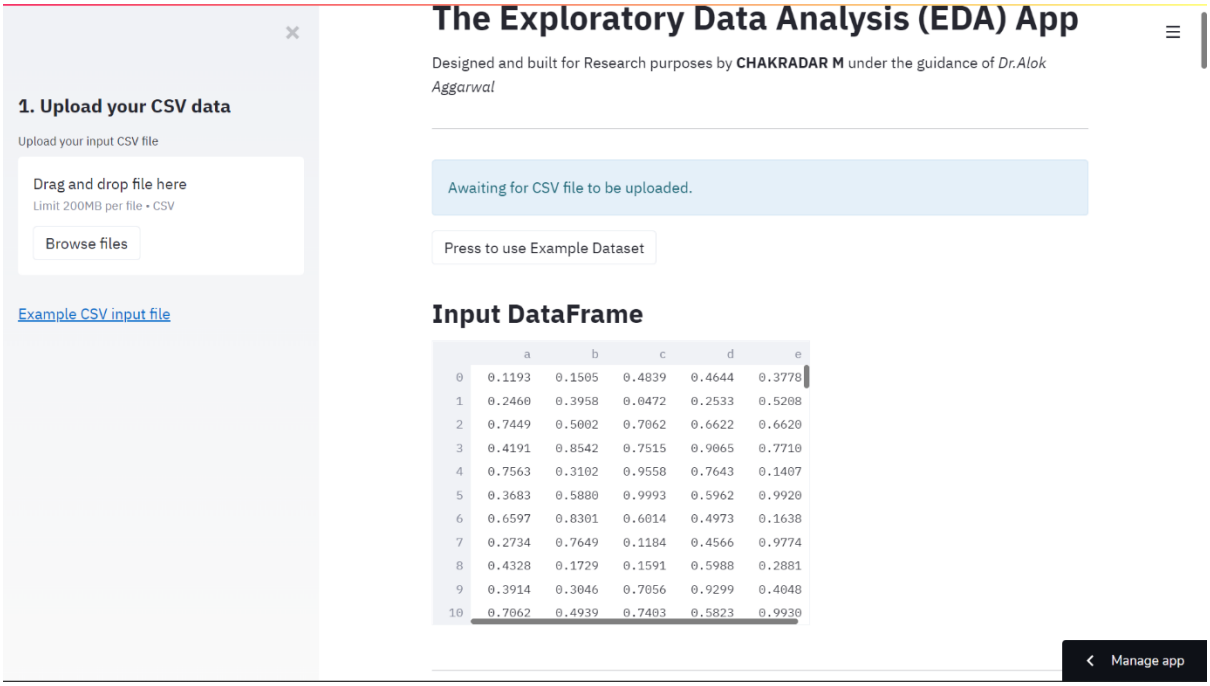


Figure 2 EDA with example dataset

Figure 4 shows the relationships between the data points in the form of explaining correlations and all 5 types of correlation techniques with a heatmap graphical representation.

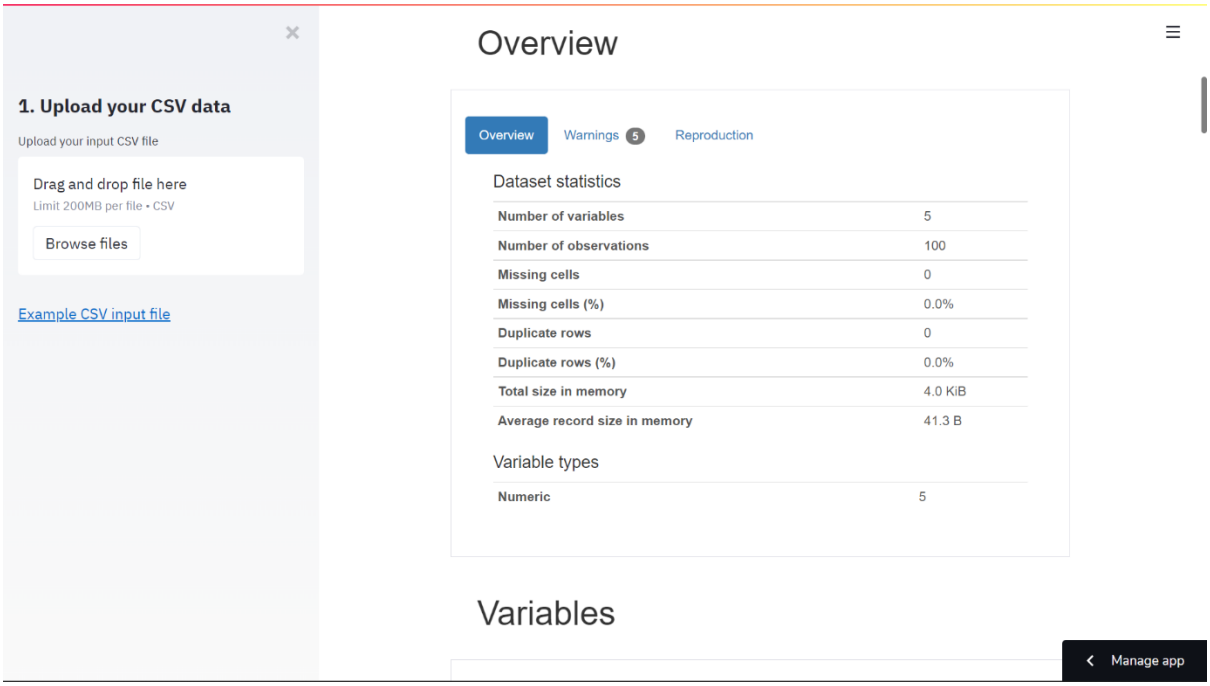


Figure 3 EDA statistical data profiling

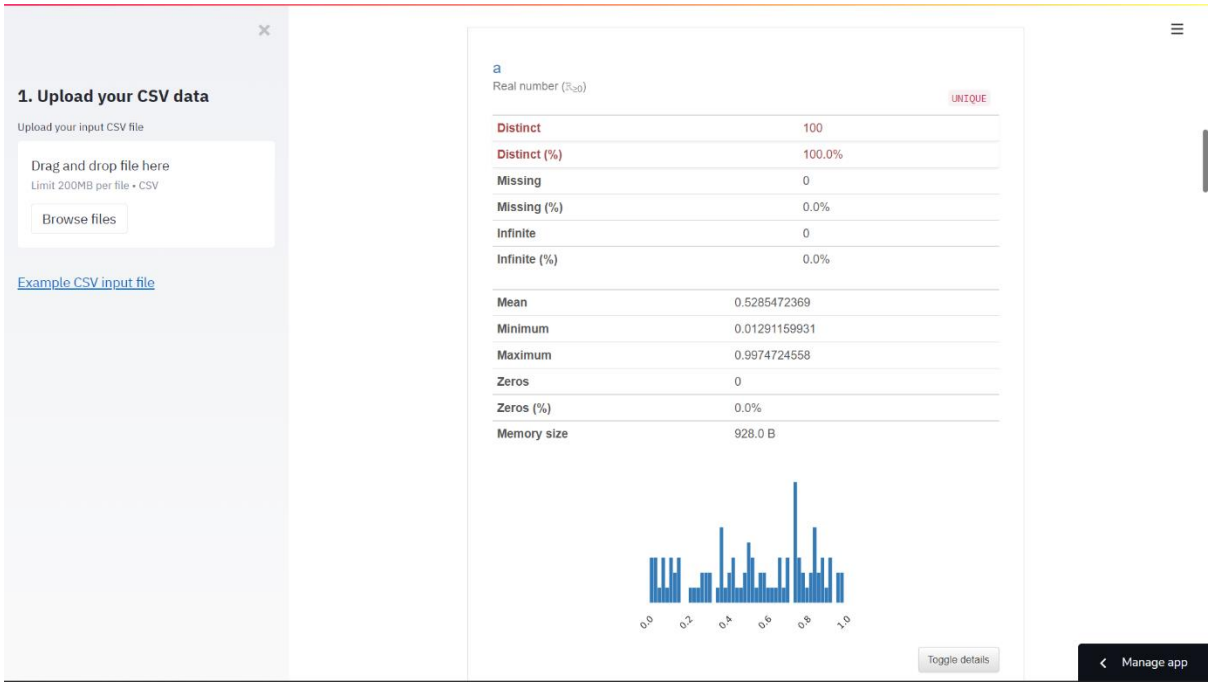


Figure 4 EDA parameter wise statistical data profiling

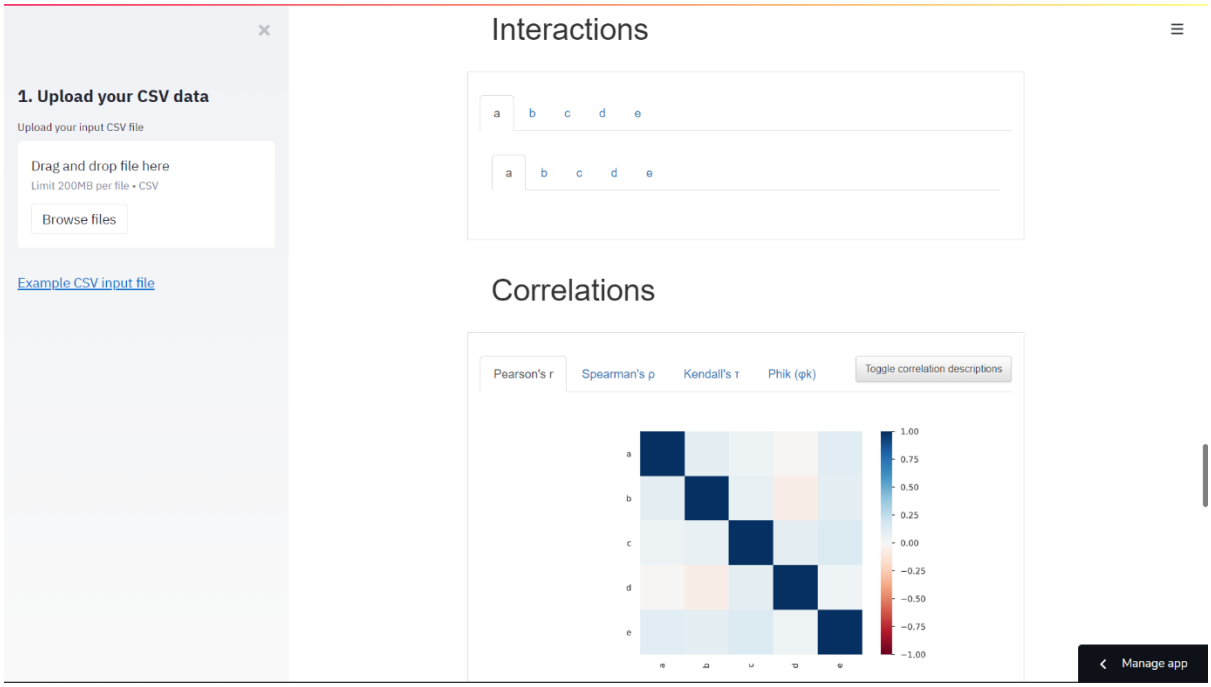


Figure 5 EDA correlation analysis and data interactions

Regression Analysis Interface:

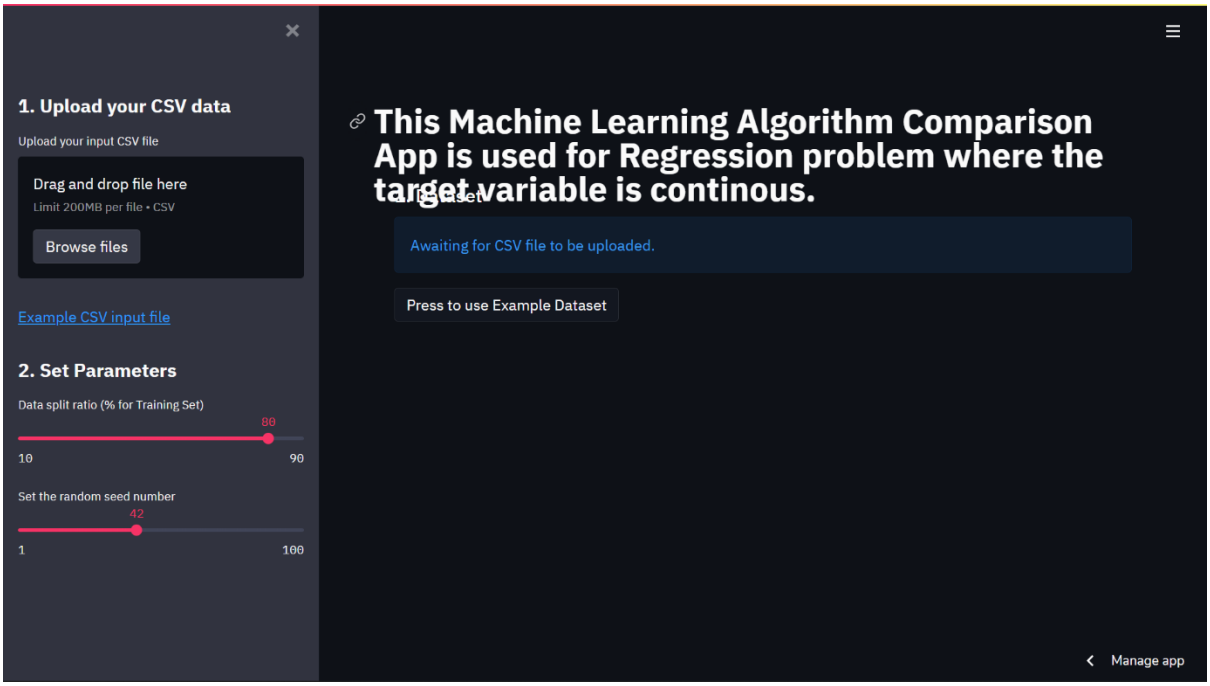


Figure 6 Regression analysis Page

This interface is primarily designed for regression-based problems which has a continuous target parameter. On the top left side of the interface there is an option to upload your dataset in the csv file format through "Upload your CSV data". One can refer to the example file in the interface as shown in the figure 6. One can drag and drop their csv file from their current device of option as this interface is web based. Maximum file size for the dataset is limited to 200MB. In the future there is a plan to add support to more database formats such as SQL.

After dropping the file in the 1st option, there are couple of slider buttons given just under them termed as "Set Parameters". As the name suggests this option is to set parameters of the models which are supposed to be generated. The first option in the set parameters is to split the entire dataset into 2 sets, one for training and one for testing. By default, training data is set to 90% and test data is set to 10%. Users can make use of this slider and access more scenarios for their datasets. The second option is to decide a random seed number, a random seed number helps users to reproduce their results. In this case picking exact datasets for training and testing.

In the center page there is an option called “Press to use Example Dataset”, in the case of regression analysis the boston housing prices dataset is added into the interface. One can click this button and see the possible analysis of the dataset and identify the best performing models of the dataset. By using the set parameters users can generate numerous scenarios based on training and testing datasets and reproduce the analysis with the help of random seed number. From an amateur data analyst to full fledged data scientist this interface should produce some value inorder to be able to read datasets and get insights out of this.

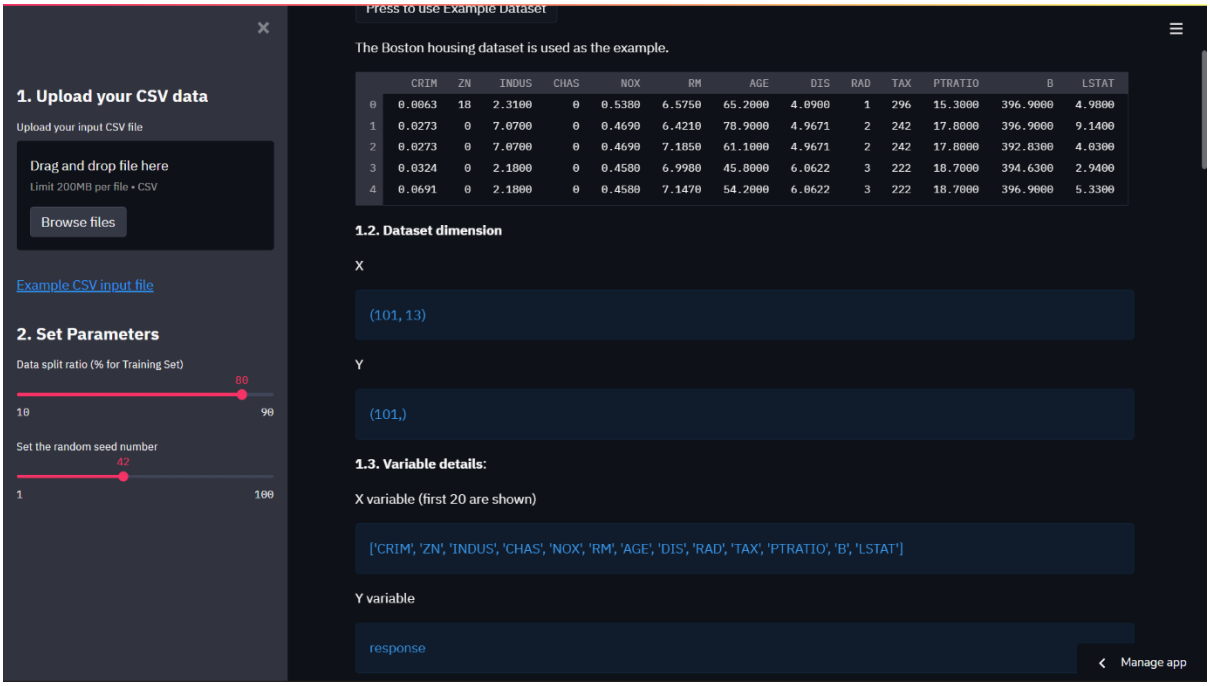


Figure 7 Regression analysis with example data

In the figure 7 on the first 4 rows of Boston housing data are displayed like spreadsheet and 1.2 explains the dimensions of the dataset. Generally, the last parameter in the spreadsheet is considered as target variable Y and the parameters before that are the dependent variables X. In the 1.3 the parameter heads and names are displayed. And so on can be explained in the interface and is self-explanatory through the interface.

Based on the dataset and why this is done two model screens are generated. One model screen is for training data and another for test data. Each table shows the few of the basic regression techniques with their cross-validation approach from

the basic regression technique to the advanced XGBoost regressor. But unfortunately, there are latest techniques which are yet to added into the system. Each of this model is explained with their performance metrics R-squared, RSME and time taken to build the model. These techniques are scrollable and the interface is equipped with over 40 regression techniques packaged into one execution.

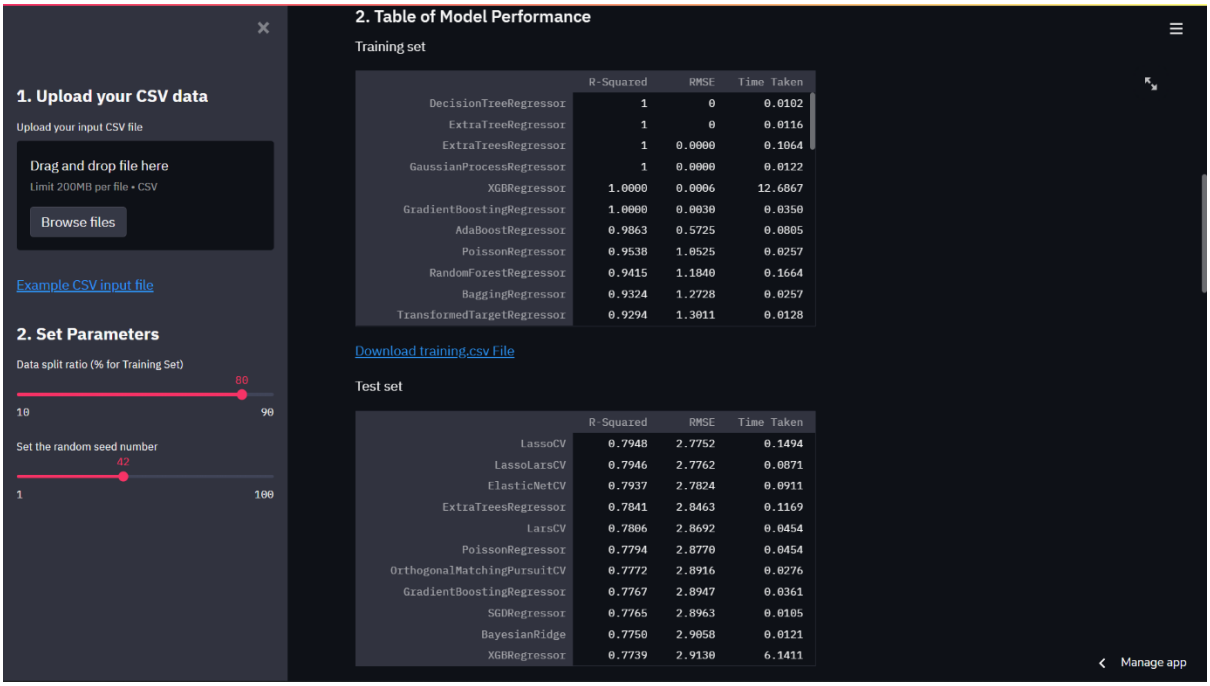


Figure 8 Table of model's performance

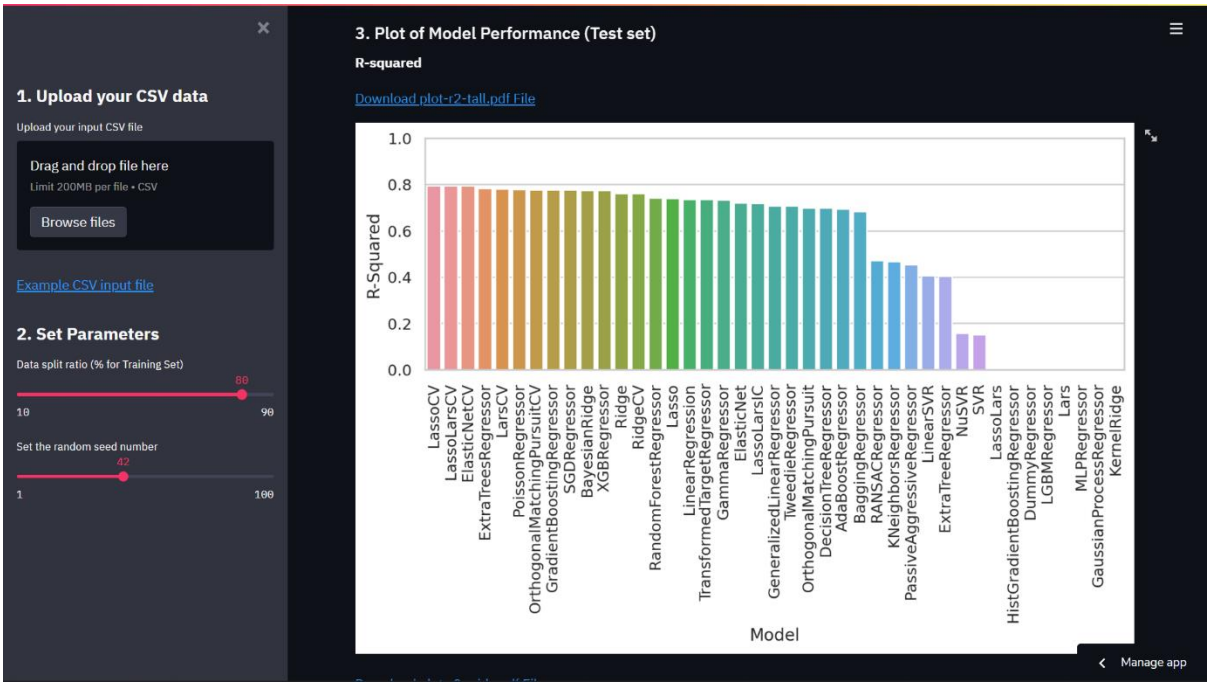


Figure 9 plot of models based on R-squared metric

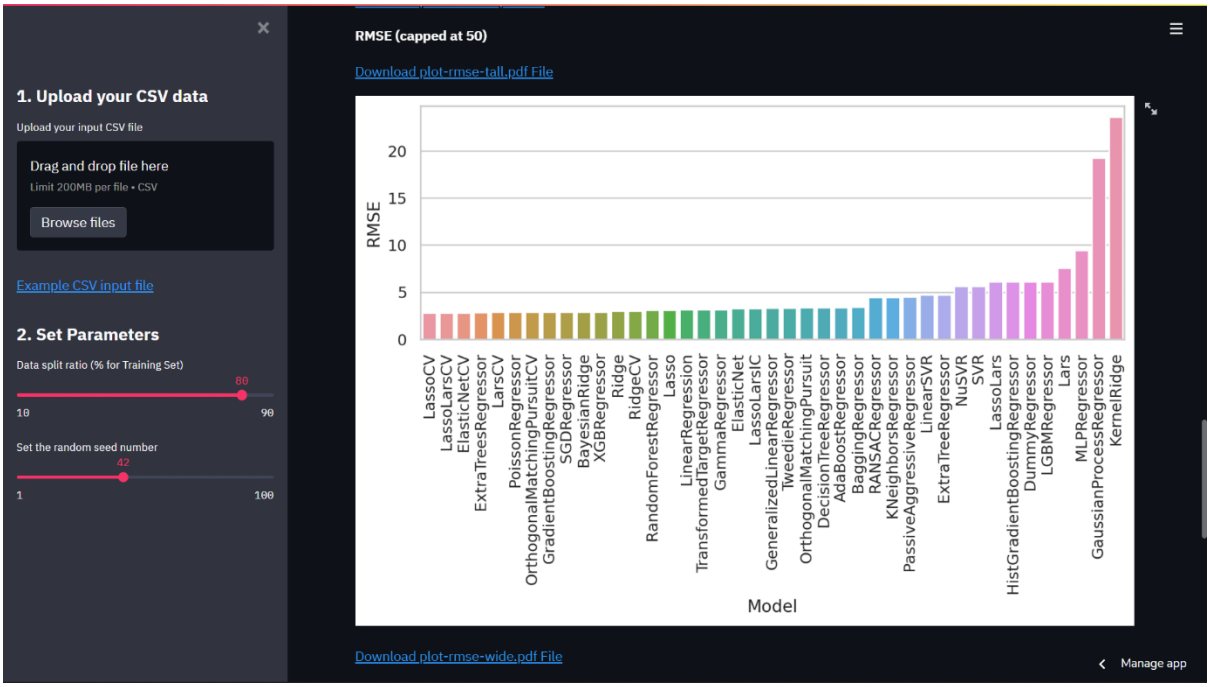


Figure 10 plot of models based on RSME metric

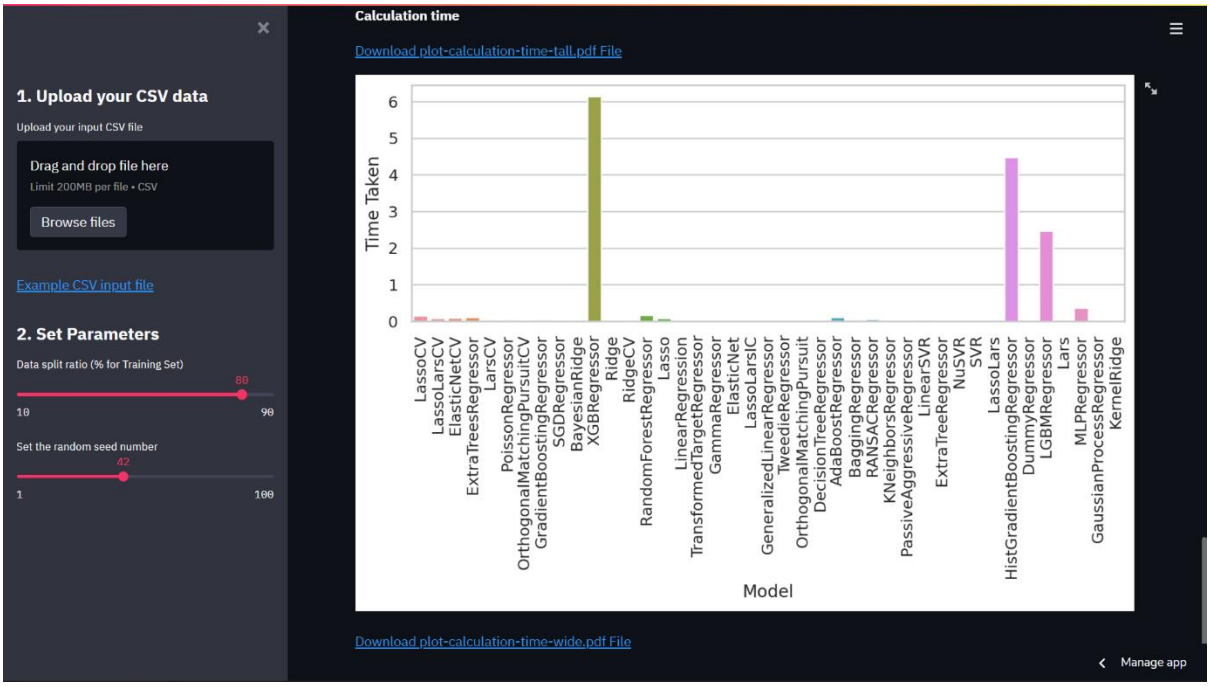


Figure 11 plot of models based on most time taken

Classification Analysis Interface:

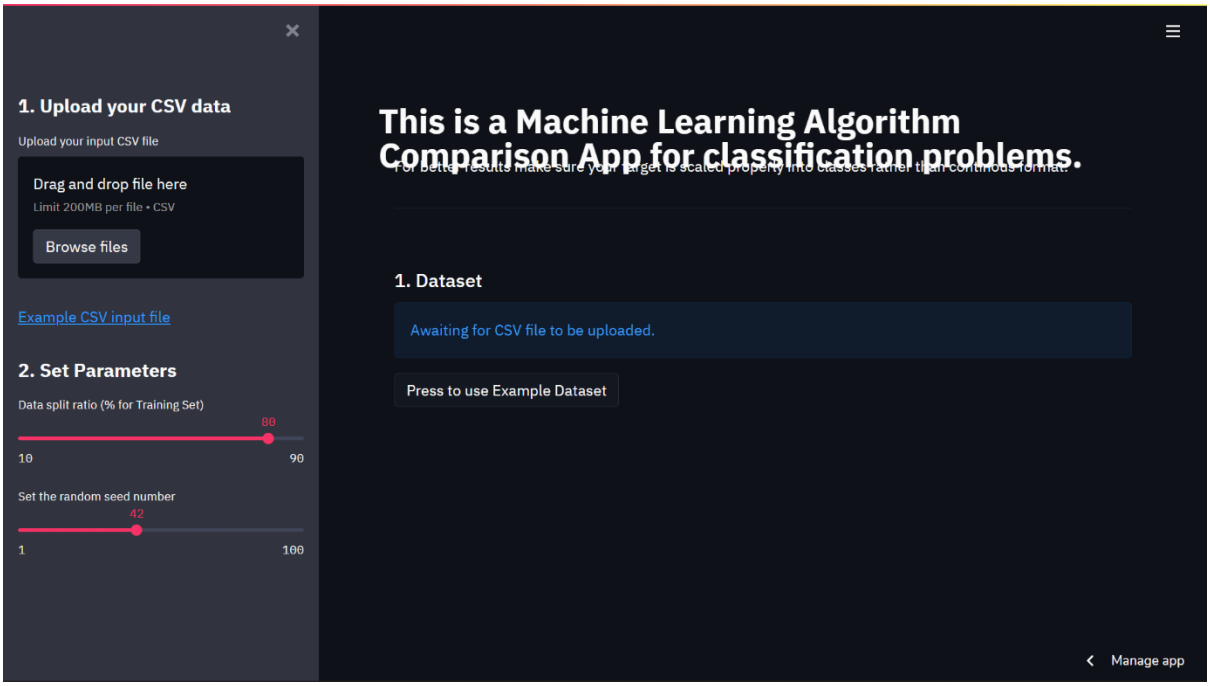


Figure 12 classification analysis Page

This interface is primarily designed for classification-based problems which has a discrete values for target parameter or else it can be within a range of classes. On the top left side of the interface there is an option to upload your dataset in the csv file format through "Upload your CSV data". One can refer to the example file in the interface as shown in the figure 13. One can drag and drop their csv file from their current device of option as this interface is web based. Maximum file size for the dataset is limited to 200MB. In the future there is a plan to add support to more database formats such as SQL.

After dropping the file in the 1st option, there are couple of slider buttons given just under them termed as "Set Parameters". As the name suggests this option is to set parameters of the models which are supposed to be generated. The first option in the set parameters is to split the entire dataset into 2 sets, one for training and one for testing. By default, training data is set to 90% and test data is set to 10%. Users can make use of this slider and access more scenarios for their datasets. The second option is to decide a random seed number, a random seed number helps users to reproduce their results. In this case picking exact datasets for training and testing.

In the center page there is an option called “Press to use Example Dataset”, in the case of classification analysis the pima Indians diabetes dataset is added into the interface. One can click this button and see the possible analysis of the dataset and identify the best performing models of the dataset. By using the set parameters users can generate numerous scenarios based on training and testing datasets and reproduce the analysis with the help of random seed number. From an amateur data analyst to full-fledged data scientist this interface should produce some value inorder to be able to read datasets and get insights out of this.

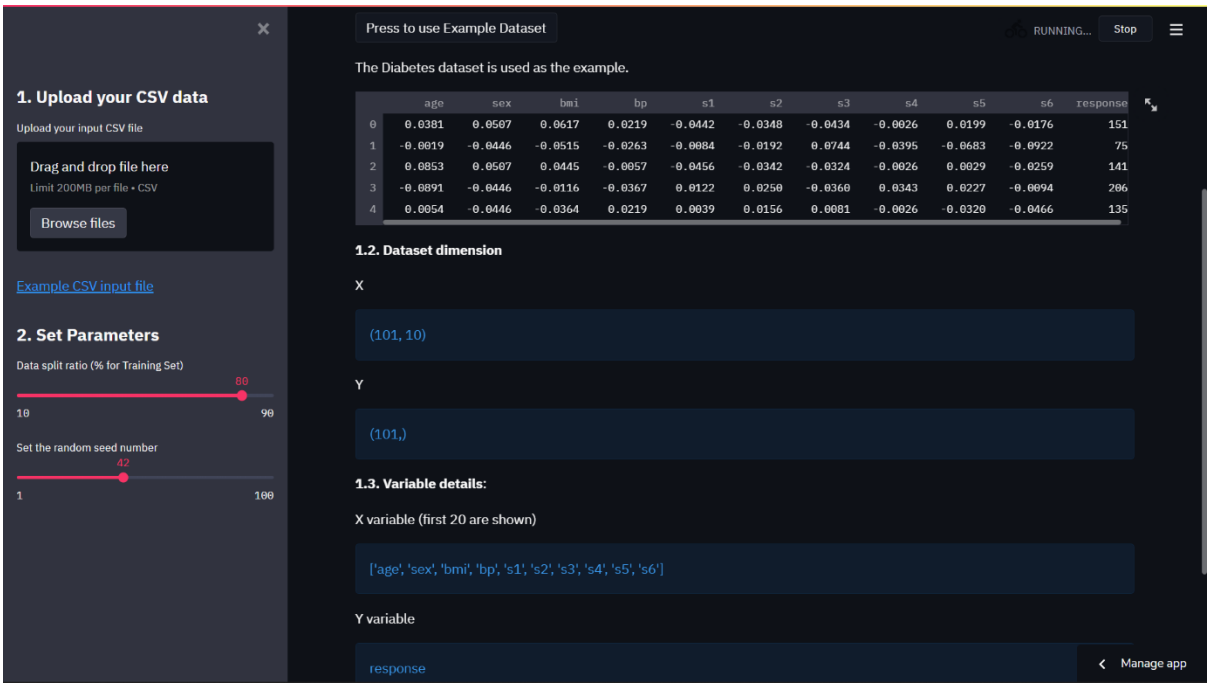


Figure 13 Regression analysis with example data

If observed in figure 13, the example is providing the basic explanation of the dataset followed by running the state-of-the-art classification techniques. In the future newer and more advanced techniques shall be added to the interface. But these models provide their results in 2 segments as shown in the regression analysis interface in the earlier segment which are based on train and test split ratio of the dataset the user can pick from the slider in the left-hand side of the application interface.

There are about 35-40 machine learning techniques added into the interface whereas their performance characteristics as follows:

1. Accuracy
2. Balanced Accuracy
3. ROC & AUC scores
4. F1-scores
5. Time taken

If the dataset has a comparable balance which means if the dataset is normally distributed, accuracy can be a valuable metric. Otherwise, Balanced Accuracy could be required. The accuracy paradox might be related to a model having high accuracy with poor performance or low accuracy with excellent performance. That's why both accuracy and balanced accuracy is being calculated in this case and also for both training data and testing data just the check dataset spread is similar in both the scenarios. This interface has undergone initial brainstorming sessions followed by rigorous testing such as unit testing for each machine learning module and all of them together. Libraries for this application are auto-updated as it checks for the latest updates on the web every once in a month. Any crash reports are forwarded to the admin saying if they are incompatible requirements with latest versions in the library requirements.

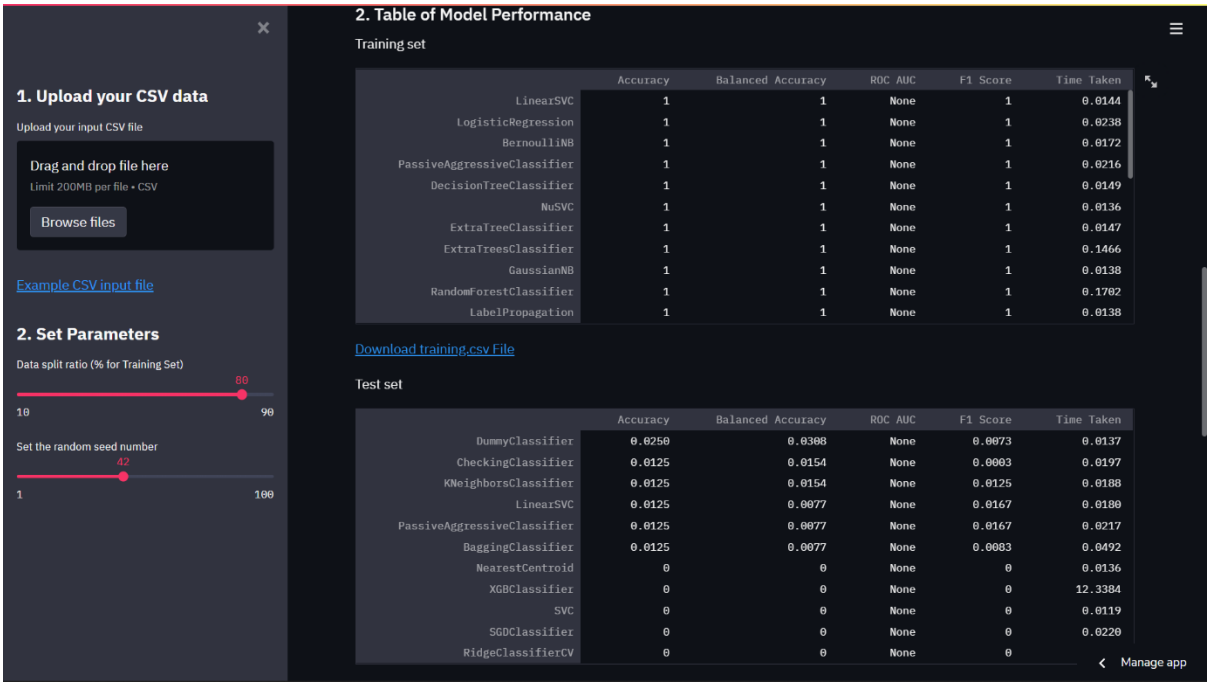


Figure 14 Model metrics based on Regression analysis

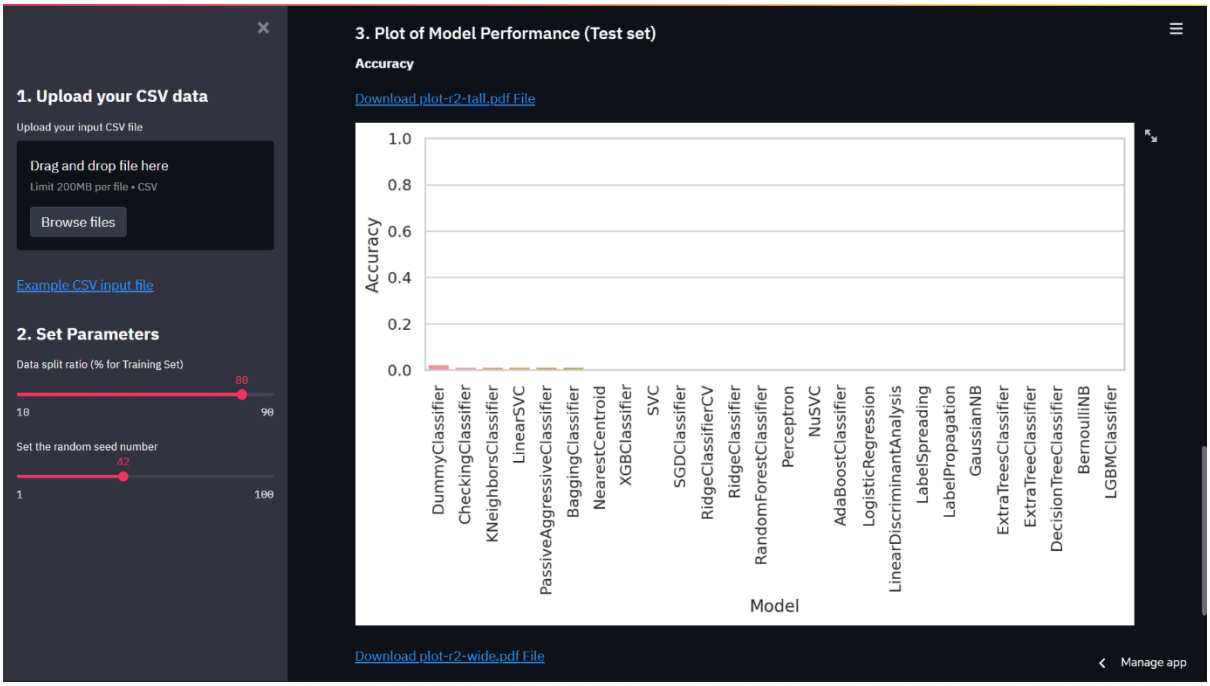


Figure 15 Plot based on accuracy scores of the models created

As in when after identifying a better performing machine learning technique, these algorithms are accessed runs separately on the dataset inorder to tweak hyper-parameters of the techniques to generate a better functioning model. The real challenges in making this interface useful has been observed when using EBM (Explainable Boosting Machines) techniques. As this technique is about explaining the parametric importance of each individual parameter to the dataset that would considerably effect on the target variable. Inshort this algorithm is nested on decision boosting technique which would also explain the parameter dependency of the dataset towards the desired target parameter. Hence just the plain performance parameters wouldn't sustained justice towards the parameter's graphs with the current application.

Madam Chakradar

EDUCATION

2014-2016	M.Tech in Automation and Robotics Engineering, University of Petroleum & Energy Studies (UPES), Dehradun
2009-2013	B.Tech in Electronics and Instrumentation Engineering, RGMCET, JNTUA

List of Publications

Patent:

1. "System and Method for non-invasive Prediction of Insulin resistance" by Alok Aggarwal, Madam Chakradar (Patent Application No. 202011016050), The Patent Office Journal No. 25/2020, Published dated 26th June 2020 under University of Petroleum & Energy Studies, Dehradun, India.

Journal Publications:

SCI Indexed:

2. Madam Chakradar, Alok Aggarwal et al., "A Non-invasive Approach to Identify Insulin Resistance with Triglycerides and HDL-c Ratio using Machine learning," *Neural Processing Letters (NEPL)*, vol. 52, no. 3, Dec. 2020. (Impact Factor: 2.891, H-Index: 50) [Index: SCI]
   (https://doi.org/10.1007/s11063-021-10461-6)
3. Chakradar M, Aggarwal A et al., "COVID-19 Risk Prediction for Diabetic Patients Using Fuzzy Inference System and Machine Learning Approaches," J Healthc Eng. 2022 Apr 1; 2022:4096950. doi: 10.1155/2022/4096950. PMID: 35368915; PMCID: PMC8974235.
   (Impact Factor: 3.822, H-Index: 37) [Index: SCI]
   (https://pubmed.ncbi.nlm.nih.gov/35368915/)
4. Madam Chakradar, Alok Aggarwal, "Identification of Adaptive Thermogenesis in Humans Using Machine Learning from CALERIE Dataset", International Journal of Biomedical Imaging, MS 8398736. (Impact Factor: 3.246, H-Index: 43) [Index: SCI] {Under Final review/decision phase on EiC, two revisions completed}
5. Chakradar Madam, Alok Aggarwal, "Designing a Machine learning model to detect insulin resistance using xgboost technique," *Turkish Journal of Electrical Engineering and computer Sciences*, vol. 29, no. 05, 2021. [Indexed SCI, IF: 0.89] {Under Review Phase}
6. Chakradar Madam, Alok Aggarwal, "Interface design for machine learning model's lifecycle based on dataset," *Turkish Journal of Electrical Engineering and computer Sciences*, vol. 29, no. 05, 2021. [Indexed SCI, IF: 0.89] {Under Review Phase}

Scopus Indexed:

7. Madam Chakradar, Alok Aggarwal, "A Machine Learning based approach for the identification of insulin resistance with non-invasive parameters using Homa-IR," *International Journal of Emerging Trends in Engineering Research (IJETER)*, vol. 8, no. 5, May 2020 [Index: Scopus]
   (http://www.warse.org/IJETER/static/pdf/file/ijeter95852020.pdf)

8. Madam Chakradar, Alok Aggarwal, “A Regression based machine learning model to estimate the missing fat-mass parameter in CALERIE study dataset”, *J Arch. Egyptol*, vol. 17, no. 12, pp. 1533-1546, Mar. 2021. [Index: Scopus]
   (https://archives.palarch.nl/index.php/jae/article/view/7054)
9. Madam Chakradar, Alok Aggarwal, “Feature selection for insulin resistance using random forest based approach”, *J Arch. Egyptol*, vol. 18, no. 4, pp. 4861-4879, Jan. 2021. [Index: Scopus]
   (https://archives.palarch.nl/index.php/jae/article/view/7053)
10. Madam Chakradar, Alok Aggarwal, A case study to validate the machine learning model built to detect insulin resistance in humans using synthetic dataset, *Recent Advances in Computer Science and Communications*, [Index: Scopus] {Under Review Phase}

International Conferences:

11. Chakradar, M., & Aggarwal, A., “A Machine Learning Model to Identify Insulin Resistance in Humans,” *IEEE International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications*-2021, pp. 351–354, 2021, IEEE Conference ID: 52345. (https://ieeexplore.ieee.org/document/9688098)
12. Madam Chakradhar, Alok Aggarwal, “Exploring parametric influence in weight loose dynamics and thermogenesis in obese adult human being through data science: a review*,” Doctoral Colloquium - 2019 in Management, Engineering, Computer Science, Design, Life Sciences & Law organized by University of Petroleum and Energy Studies*, School of Law, 24th May 2019.
13. Madam Chakradar, Alok Aggarwal, “ML and fuzzy based risk prediction model for diabetic patients for COVID-19,” *Inter. Conf. on Advancement in Interdisciplinary Research-2020*, July 24-26, 2020, Agra, India.
14. Madam Chakradar, Alok Aggarwal, “ML and fuzzy based risk prediction model for cardiovascular patients for COVID-19,” *Inter. Conf. on Advancement in Interdisciplinary Research-2020*, July 24-26, 2020, Agra, India.
15. Madam Chakradar, Alok Aggarwal, “Understanding feature dependent variables of CALERIE study dataset using EBM models,” *2nd Inter. Conf. on Advancements in Interdisciplinary Research (ICAIR-2021),* organized by Dr. Bhimrao Ambedkar University, Agra, India & FIRMS India, October 29 - 31, 2021.
16. Madam Chakradar, Alok Aggarwal, “Designing a Machine learning model to detect insulin resistance using xgboost technique,” 2nd Inter. Conf. on Advancements in Interdisciplinary Research (ICAIR-2021), organized by Dr. Bhimrao Ambedkar University, Agra, India & FIRMS India, October 29 - 31, 2021.
17. Madam Chakradar, Alok Aggarwal, “Interface design for machine learning model's lifecycle based on dataset,” *2nd Inter. Conf. on Advancements in Interdisciplinary Research (ICAIR-2021),* organized by Dr. Bhimrao Ambedkar University, Agra, India & FIRMS India, October 29 - 31, 2021.
18. Madam Chakradar, Alok Aggarwal, “Identification of major bio-markers from CALERIE study dataset using Explainable Boosting Machines(EBM),” *2nd Inter. Conf. on Advancements in Interdisciplinary Research (ICAIR-2021),* organized by Department of Mathematics, Dr. Bhimrao Ambedkar University, Agra, India & FIRMS India, October 29 - 31, 2021.

19. Madam Chakradar, Alok Aggarwal, "A Literature review of Adaptive Thermogenesis and why insulin resistance is a bio-marker from machine learning analysis of CALERIE study dataset," *2022 8th International Conference on Signal Processing and Communication (ICSC)*. {Under Review Phase}
20. Madam Chakradar, Alok Aggarwal, "An observation on advancements in wearable sensors to identify adaptive thermogenesis in humans," *IEEE Int. Conf. CISCT*. {Under Review Phase}

UPES
UNIVERSITY WITH A PURPOSE

PLAGIARISM CERTIFICATE

1. We Dr. Alok Aggarwal (Internal Guide), -- (Co Guide/ External Guide) certify that the Thesis titled

A Machine Learning Model To Identify and Prevent the Occurrence of Adaptive Thermogenesis

submitted by Scholar Mr/ Ms Madam Chakradar having SAP ID 500057009 has been run through a Plagiarism Check Software and the Plagiarism Percentage is reported to be 6 %.

2. Plagiarism Report generated by the Plagiarism Software is attached .

Signature of the Internal Guide

Signature of External Guide/ Co Guide

Signature of the Scholar

19.10.2022 Eve thesis-plag

ORIGINALITY REPORT

6%	5%	3%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

2%

★ doaj.org

Internet Source

Exclude quotes	Off	Exclude matches	< 14 words
Exclude bibliography	Off		