NON- COMMUNICABLE DISEASE PREDICATION BY USING MACHINE LEARNING APPROACH

L. Mahendiran

16APC2778

Bachelor of Science Honours Degree in Computing and Information Systems

Department of Computing and Information Systems

Sabaragamuwa University of Sri Lanka

January 2023

NON- COMMUNICABLE DISEASE PREDICATION BY USING MACHINE LEARNING APPROACH

L. Mahendiran

16APC2778

Thesis submitted in partial fulfillment of the requirements for the BSc. (Hons) in Computing and Information Systems

Department of Computing and Information Systems

Sabaragamuwa University of Sri Lanka

January 2023

DECLARATION

I hereby declare that this thesis does not, except where appropriate citation is made in the text, incorporate, without acknowledgment, any material previously submitted for a Degree or a Diploma in any university. I affirm that the work contained in this thesis is entirely original to me and was completed after I registered for the degree or diploma at Sabaragamuwa University of Sri Lanka. To the best of my knowledge and belief, this thesis also does not contain any material that I or another person have previously published or written. In addition, I hereby grant Sabaragamuwa University of Sri Lanka the non-exclusive right to publish, distribute, and otherwise use my thesis in print, electronic, and other media, in whole or in part. I reserve the right to use this material entirely or in part in other works (such as articles or books).

26-01-2023 M. Lenuga

……………………….. …………………………..

Date Signature

M. Lenuga (16APC2778)

CERTIFICATION OF APPROVAL

This is to certify that the thesis on "Non- Communicable Disease Predication By Using Machine Learning Approach" submitted by L. Mahendiran, in Partial fulfillment of the requirements for the award of the Degree of BSc. (Hons) in Computing and Information Systems is an original work carried out by her under our joint guidance. This thesis has been submitted with our approval.

Date

…………… ……………………………………

Mr. K. Banujan,

Lecturer,

Department of Computing and Information Systems,

Faculty of Computing,

Sabaragamuwa University of Sri Lanka.

Date

…………… ……………………………………

Dr. L.S. Lekamge,

Head of the Department,

Department of Computing and Information Systems,

Faculty of Computing,

Sabaragamuwa University of Sri Lanka.

DEDICATION

This study paper is dedicated to my wonderful family, who patiently supported me until all of my research was completed, and to my cherished mother, who has been actively encouraging me for months to complete my work with genuine self-confidence. Thank you to my academic adviser who guided me in this process and the committee who kept me on track. I dedicate this dissertation to my relations and friends who inspired my pursuit of economics.

ACKNOWLEDGEMENT

First, I would like to express my deepest gratitude to my supervisor **Mr. K. Banujan, Lecturer**, Department of Computing and Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, for his advice, support, and vast knowledge that he shared with me while I was studying and conducting research, as well as for the time he dedicated to helping the research team finish the project successfully.

Specially I must thanks to my research co-supervisor **Mrs. Y. Vasini (Nutritionist),** Jaffna Teaching Hospital. She is assisting me in locating a research topic, gathering information from the hospital, and offering assistance whenever I require it.

Furthermore, I would like to express my gratitude to **Dr. L.S. Lekamge**, Head of the Department, Department of Computing and Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, for her guidance and encouragement to successfully conduct this research project, **Dr. U.A.P. Ishanka**, the research project coordinator, and the senior lecturer of the Department of Computing and Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka for their active coordination and the tremendous support provided to us in finishing this work successfully.

The main objective of any computer science student is to get as much of doing research knowledge as possible. Being an able to have a research knowledge by developing a project is a lifetime experience. I'm thankful of having this project. Therefore, I also extend my thanks to my lecturers and to Department of Computing and Information Systems for giving me best education support which helped me in many ways for this project works.

I will be eternally grateful to my family members for their unwavering support, encouragement, and strength in helping me finish this research project and throughout my life. Finally, I would like to express my gratitude to all of the academic and non-academic staff at the Department of Computing and Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, as well as my friends, for their encouragement and best wishes for the successful completion of this research.

Lenuga Mahendiran.

ABSTRACT

Every person receives a life expectancy as a result of medical treatment. Good health services ensure people's well-being. Two diseases might harm a person's health: (a) communicable and (b) non-communicable. In recent years, heart disease, diabetes, and various types of cancer have been recognized as some of the leading causes of death (80%) in most nations worldwide. Obesity, frequently ascribed to excess body fat, is one of the most prevalent risk factors for these diseases. The researchers used knowledge discovery through predictive modeling to make the enormous amounts of data generated by health care information systems valuable to the potential. Anthropometric measurements were employed in this study as input data to several data mining techniques to forecast patients with non-communicable diseases. Data was collected at JTH and SUSL, and only adults over the age of 18 were considered, with the exception of children under the age of 18 and pregnant women. Ten variables: age, weight, height, gender (girls and boys), mass of body fat (MBF), total body water (TBW), percent of body fat (PBF), body mass index (BMI), visceral fat area (VFA), and waist-to-hip ratio (WHR) were utilized to represent the most valuable features in the prediction procedure, and researchers found the most affecting feature is VFA. Data mining techniques for binary classification include ensemble machine learning approaches such as Cat Boost, Ada Boost, and Extreme Gradient Boosting, as well as Support Vector Machine (SVM), Artificial Neural Networks (ANNs), Random Forest, Decision Tree, and Logistic Regression, which employ the chosen features as input and output. Yes/No is the final result. If the result is yes, that patient is an NCD patient. These eight algorithms provide more than 85% accuracy, and the highest algorithm is Random Forest, with an accuracy is 98.90%.

Keywords: Obesity, Data mining, Artificial Neural Network, Support Vector Machine, Artificial Neural Networks, Random Forest, Extreme Gradient Boosting, Decision Tree, AdaBoost Logistic Regression, and Cat Boost, Prediction, Non- Communicable Disease.

Table of Content

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML- Machine Learning

SVM- Support Vector Machine methods

AI- Artificial Intelligence

ANN- Artificial Neural Network

JTH- Jaffna Teaching Hospital

SUSL- Sabaragamuwa University of Sri Lanka

LBW- Lean Body Water

MBF- Mass Body Fat

TBW- Total Body Water

SLM- Soft Lean Mass

TBW- Total Body Water

PBF- Percentage Body Fat

BMI- Body Mass Index)

VFA- Visceral Fat Area

AC- Abdominal Circumference

WHR- Waist Hip Ratio

Lt.Arm- Left Arm

Rt.Arm- Right Arm

Lt.Leg- Left Leg

Rt.Leg- Right Leg

SLM- Soft Lean Mass

BMR- Basal Metabolic Rate

AMB- Age Matched of Body

NCD- Non- communicable Disease

LIST OF APPENDICES

1. CHAPTER ONE- INTRODUCTION

1.1. Background

A person’s quality of life is significantly impacted by maintaining excellent health and preventing disease. Health issues present significant economic and social obstacles in both low- and middle-income nations [1]. The terms “communicable” and “non-communicable” are widely used to describe diseases. In contrast to NCDs, which are primarily chronic diseases like malignancies, cardiovascular diseases, and diabetes, communicable diseases include infectious diseases like measles and tuberculosis [2]. Disease-carrying agents or vectors propagate infectious disease epidemics among susceptible population sectors in predictable ways. That brings up the concept of communicable. However, a lot of people lately disputed the term “non-communicable” and asserted that certain illnesses are indeed communicable [3]. One of the main causes of death and disability in low-income nations and marginalized groups is communicable diseases, which include HIV/AIDS, tuberculosis (TB), malaria, viral hepatitis, and neglected tropical diseases (NTDs) [4].

Non- communicable illness burden has been increasing globally for many years. Chronic diseases, also known as NCDs, are not contagious, have a propensity for a lengthy duration, and are caused by a confluence of genetic, physiological, environmental, and behavioral variables. [5]. In addition to other things, this trend has been substantially influenced by tobacco use, physical inactivity, bad diets, and excessive alcohol intake. The four primary categories of NCD include diabetes, cancer, chronic respiratory diseases like chronic obstructive pulmonary disease and asthma, and cardiovascular disorders like heart attacks and strokes [6]. Chronic diseases are directly linked to health risk behaviors including smoking, poor diet, physical inactivity, and excessive alcohol consumption; nevertheless, the factors that contribute to these risky behaviors are much too commonly ignored [7]. A significant risk factor for NCDs is malnutrition [8]. Approximately one in three individuals worldwide suffer from some kind of malnutrition4, and based on present trends, this number will increase to one in two by 2025. 5 Malnutrition comprises nutritional conditions such stunting, wasting, and micronutrient deficiencies that are brought on by insufficient intake of energy or nutrients. It also involves overeating and eating in an unbalanced manner, which can result in obesity, overweight, and diet-related NCDs [8]. High-level recognition of the connection between NCDs, poverty, and social and economic growth has resulted [9].

Nutritionists, healthcare experts, and sports scientists are all interested in human body composition assessments since they are impartial tools for evaluating person's nutritional status. There is an increased need for body composition measurements with improved sensitivity and precision due to the rising prevalence of obesity and lifestyle disorders [10]. In nutrition, the assessment of body composition is useful for describing growth and development from conception to adulthood and for understanding the developmental causes of health and disease. It is also useful for designing nutritional strategies and for keeping track of therapeutic interventions [11]. One of the most important aspects of a person's health and fitness profile is their body composition, which can be influenced by their surroundings (social and cultural), genetics, ethnicity, age, and sex [12]. Numerous studies have demonstrated how important body composition has become in recent years [13]. Various components of the human body are described and quantified by body composition [14]. Our body's composition determines what proportion of our overall weight is made up of bone, muscle, and fat [15]. Due to differences in their physical composition, two people with the same height and weight may experience distinct health problems [16].

The body composition analyzer takes a variety of measurements [17]. Those are Weight, LBW, MBF, SLM, Mineral, Protein, TBW, PBF, BMI, Fat Mass, VFA, AC, WHR, STDWT, Lt.Arm, Rt.Arm, Trunk, Lt.Leg, Rt.Leg, Body Fat, SLM, BMR, AMB, Control Kilocalories, Diet, Exercise [18-21]. This anthropometric assessment is employed to identify illnesses, nutritional status, and other factors [22].

Artificial intelligence is only recently being used in the healthcare industry [23]. Different statistical methods that enable computers to learn from experience without being explicitly programmed are referred to as ML[24]. ML is a branch of AI that focuses on how computers learn without explicit programming. AI technology that attempts to increase the efficiency and precision of doctors' work. AI offers great hope for countries now struggling with overcrowded healthcare systems and a physician shortage [25]. The two main subfields of ML are supervised learning and unsupervised learning [26]. One of the greatest sectors of the global economy that this technology may help is healthcare. Numerous studies have already suggested that AI is capable of performing important healthcare activities, such disease diagnosis, on par with or even better than humans [27-29].

Recently, data mining has gained acceptability on a global scale in practically all fields of life, and particularly in medical [30]. It is a method of knowledge discovery in which information is discovered by looking at data that may be concealed in very vast sources, these sources are then examined from various angles using various methodologies, and the information that was retrieved is then condensed into useable information [31]. This section describes the methods for identifying NCDs and the use of data mining in forecasting NCDs in light of the enormous potential of data mining in improving health care [32].

In this researcher is planned to predict the NCDs person with the help of the body composition analyzer. Researcher considers some parameters or attributes are Age, Weight, Height, Gender, MBF, TBW, PBF, BMI, VFA, WHR. The VFA and WHR both are relatively same. The above mention parameters normal range and risk range are mention on below table (See Tabel 1- 1, Table 1- 2). The researcher has those parameters to identify the NCDs persons. Researchers used some data mining algorithms like Random Forest, Extreme Gradient Boosting, ANN, Decision Tree, AdaBoost/ Adaptive Boosting, Logistic Regression, Cat Boost, SVM. These algorithms automatically identify the pattern of that dataset to classify theNCD patient.

Table 1-1: Parameters with normal condition and risk range

Parameter	Normal Range	Risk Range
Total Body Water (TBW)Male	(30.0~40.0)	>40%
Total Body Water (TBW) Female	(25.0~32.0)	>33%
Body Mass Index (BMI)	(18.5~25.0)	(23~27.0)&>27.50
Visceral Fat Area(VFA)-Male	(50~100)	>90 cm
Visceral Fat Area (VFA) -Female	(40~80)	>80cm

Waist Hip Ratio (WHR)-Male	(0.75~0.90)	>0.90m
Waist Hip Ratio (WHR) -Female	(0.70~0.85)	>0.85m

Table 1-2:Normal and Risk range for NCDs

Parameter	**Normal Range**	**Risk Range**
Percentage of Body Fat: Male		
20-39yrs	(8~20)%	>25%
40-59yrs	(11~22)%	>28%
60-79yrs	(13~25)%	>30%
Percentage of Body Fat: Female		
20-39yrs	(21~33)%	>39%
40-59yrs	(23~35)%	>40%
60-79yrs	(24~36)%	>42%

1.2. Purpose and motivation of the study

The most common causes of death in today's world are non- communicable diseases. According to a survey by the World Health Organization (WHO), non-communication disease as cardiovascular diseases (CVD) is the cause of almost a third of deaths worldwide. Early detection of the heart disease could help reduce the death rate. They are generally due to modern lifestyles. This researcher objective is to do a prediction about the non- communication disease using machine learning approach. For each approach we will present the objectives, the algorithms used, the data source, the risk factors or parameters used, and result obtained.

In this work researchers will show the diversity of these approaches through the different parameters or risk factors that are used in the prediction of NCD. Four main factors age, gender, percentage of body fat, BMI, WHR have been identified by health professionals.

Researchers will, thus note that in most approaches these factors are not all used and that nevertheless the result obtained.

So, we will identify all of the parameters that have made it possible to predict NCD and future work show the impact of each of them in predicting NCD diseases and improve dataset to predict NCD. Results of the test can help a doctor to: Who have or haven't NCD.

1.3. Research Objectives

1.3.1. Main objective

- To find the NCD person or factor for the particular anthropometric measurement along with the prediction model.

1.3.2. Sub objectives/ Specific Objectives

- To identify the significant methodology to incorporate the machine learning for analyzing the results
- To identify the highest accurate supervised learning technique in developing the prediction model
- To identify the highly influential anthropometric parameters towards for NCD

1.4. Problem statement and Research questions

1.4.1. Problem statement

NCD patient identification is a must wanted. Because, most of the deaths are due to the causes of the NCD like heart diseases, cancer, diabetes and etc. So, anthropometric measurement is help to identify the NCD person or patient along with some parameters. Experienced nutritionist and doctors can able to find patient's condition to identify the NCD patient or they can provide the medicine also. Inexperienced nutritionist or doctors can't be able to understand whether that the patient NCD or not. Everyone needs to check the body composition or analyze or measurement their body frequently. Because they don't know whether they are healthy people or not? provide suggestion or medicine. There will 3 types of measures are made by the nutritionist such as height, weight and body composition analyzer measurement. If an experience nutritionist isn't available at a particular condition, with the help of the prediction model, we can easily assist the experts as well as beginners in the medical field to find the appropriate solution for the problem. If a nutritionist/ doctors can't be able to find out the exact problem of a patient using manual process, they can easily utilize the outcomes of this approach.

The steady rise in NCDs has posed a serious challenge to Sri Lanka's healthcare system. So, in this research, the problem statement is how machine learning approach find the disease factor in a NCD prediction.

1.4.2. Major Research Questions

RQ1: What is the best method for predicting NCD prediction?

RQ2: How to find a possible factors for NCD prediction?

RQ3: What are the main features important for your target?

RQ4: How the evaluation process of the implemented system can be carried out?

RQ5: How the proposed methodology has been different from existing methodologies?

1.5. Significance of the Research

NCDs such as heart disease, stroke, cancer, diabetes, and chronic lung disease cause over 80% of deaths in Sri Lanka. More than a quarter of Sri Lankan's are overweight, and one in four adults consume tobacco. Addressing NCDs requires a lifestyle change, and since individuals in the workforce spend a significant amount of time in the workplace, creating a healthy workplace is vital to our well-being.

The goal of this project is to compare and contrast the most successful methods of disease prediction, as well as test and evaluate various data mining algorithms that will aid in the prediction of NCDs.

The experiment is expected to provide doctors with a valuable tool for forecasting in complex medical cases involving anthropometric factors related to NCDs and providing parents with advice based on the predictions produced by precise algorithms, which will be extremely advantageous for the field of medical science. This study will show that physicians may successfully predict and forecast risky medical cases using the data mining technique, which is widely used in industrialized countries. This study will serve as a foundation for future data collection on patients in Sri Lanka.

1.6. Delimitations and Limitations and assumptions

This study has certain limitations such as it includes patients' records from machine and data collect from northern and Sabaragamuwa province in Sri Lanka. Once completed, the findings of the research would be communicated and discussed with the healthcare authorities in the Sri Lanka, where additional comprehensive research can be conducted for covering the whole Sri Lanka population.

The experiment is expected to provide doctors with a valuable tool for forecasting in complex medical cases involving anthropometric factors related to NCDs and providing parents with advice based on the predictions produced by precise algorithms, which will be extremely advantageous for the field of medical science. This study will show that physicians may successfully predict and forecast risky medical cases using the data mining technique, which is widely used in industrialized countries. This study will serve as a foundation for future data collection on patients in Sri Lanka.

1.7. Contribution to the Study

This study deal with the problem of NCD prediction based on the ML approach. First, Reacher formulate the problem of Raga classification task using traditional approaches.

Second, we explore highly affecting factor for the NCD based anthropometric measurement by using ML approaches. Finally, we show that Random Forest model can be used for NCD prediction. To conclude, the following are the major contributions of this work:

- We classification the NCD using ML approaches such as Random Forest, Extreme Gradient Boosting, ANN, Decision Tree, AdaBoost, Logistic Regression, Cat Boost, SVM.
- With our approach we obtain 98.90% high accuracy to classify 300 datasets for our NCD classification task.

1.8. Composition of the Thesis

Chapter 01 generally describes the research filed of ML and background of the NCD prediction task and the purpose of the study, motivation and the significance of conducting the research, and what research questions going to address through the study followed by subsections.

a. Background

b. Purpose of the study and research objectives

c. Research questions

d. Significance of the study

e. Delimitations and limitations

f. Contribution to the study

g. Composition of the Thesis

Chapter 02 provides an overview of the related works. It based on the literature review about past NCD classification methods and anthropometric studies carried out. NCD and finally a summary table of 15 past studies including title, purpose or objective, machine learning or techniques, limitations and weaknesses and future directions for NCD of the studies respectively.

Chapter 03 presents each aspect of the progression of this study. It focuses on the overall methodology to propose a ML approach to our classification task which includes following subsections,

a. Research design

b. Data collection

c. Data Cleaning

d. Feature extraction

e. Feature selection

f. Implementation

g. Evaluation

Chapter 04 provide all the descriptive data such as results of the most affecting factors for our research problem, evaluation results obtained by ML models, results of overfitting handling techniques, compiration results of proposed method with traditional approaches and findings from NCD disease prediction task.

Chapter 05 concludes the thesis and suggest potential directions for future development of this study.

2. CHAPTER TWO- LITERATURE REVIEW AND RELATED WORK

In this section are presented various studies and surveys that have been conducted in the prediction and diagnosis of NCDs in general. In the present times, an increased acceptance of the datamining on the international forums is being recognized, across different medicine and life spheres. Considering the dynamics scope of data mining related to its efficacy for enhancing the healthcare outcomes, this section intends to highlight the theoretical basis which assists in NCD determination as well as the application of the data mining in forecasting.

[33] presented the study which shows the NCD prediction system using Machine Learning. The proposed model diseases are abstract health parameter dataset collected from Kaggle and stored in database. That dataset considered attributed like user ID, Step count, mood, calories, burned, hours of sleep, bool of active, weight(kg), and BMI. The second step is user input and preprocessing. After that applied by Machine Learning algorithm such as K-Nearest Neighbor Clustering to find the nearest distance, sorting, data point selection in third step. The fourth step is entropy estimation. The fifth step is hidden Markov model. The last step is fuzzy classification. K- nearest neighbor algorithm are evaluated form the disease proneness factor and selected for the hidden factor estimation process by Shannon information gain theory and proposed model successfully predicts some diseases like Diabetes, Insomnia, Obesity, Anxiety, Hypertension and Cardiovascular [33].

Machine learning based heart disease prediction system for Indian population: An exploratory study has been prepared by [34]. During this research, a technique incorporated with specific predictive modeling was taken into consideration. To name them would be K-Nearest Neighbors, Naive Bayes, Logistic Regression, AdaBoost and Random Forest. This research carried out by collaboration of data scientists and specialist doctors. They randomly selected dataset and they exclude medical data sets corresponding to pregnant females, patients reporting chronic kidney disease, severe mental illness, atrial fibrillation, patients who reported the prolonged use of anti-depressants, antibiotics and medicines for asthma, tuberculosis and cancer, patients who are prescribed oral corticosteroids, antipsychotic drugs and immunosuppressants and patients younger than 20 years or older than 100 years. Totally collected 1670 medical records, age between 30-79, 881 males and 789 females. In that dataset 893 positive CVD and 777 negative CVD [34]. Considered lifestyle attributes like age, gender, weight, height, total cholesterol levels, smoking habits, alcohol, diabetes, hypertension, family history of CVDs, intake of healthy diet, physical activity/exercise habits and stress/anxiety in life, internal attribute is BMI. The prediction system was later deployed in the cloud for easy accessibility via Internet [34].

[35] This study evidenced that first study to applied ML approaches to predict the hypertension and its associated factors using population-representative data in three South Asia countries. Researchers publicly available datasets were analyzed got it from DHS website upon registration and request., identified seven risk factors associated with hypertension in the South Asian population: age, BMI, education, wealth status, ever measuring BP, being diagnosed by a doctor, and taking medication to lower BP. This founded study BMI to be a good predictor of hypertension. Previous studies have shown obesity is strongly related to the risk of developing hypertension, whereas waist circumference is correlated with cardiovascular diseases. This study suggested that using simple, non-invasive information, ML models can predict hypertension among the South Asian population with high accuracy. Age and BMI were the most significant risk factors associated with hypertension in our study population [35].

[36]. have researched “Association and Predictive Capability of body composition and diabetes mellitus using artificial intelligence” a cohort study. This study aimed the association between regional body fat distribution and the prevalence of Diabetes mellitus (DM) in adult populations using machine learning. The “Tanita Segmental Body Composition Analyzer BC-418 MA Tanita Corp, Japan” machine was used by researchers. They measure some Weight, Basal Metabolic Rate (BMR), Fat Percentage:(Fat Mass/Weight) × 100 (FATP), Fat Mass (FATM), Fat-Free Mass (FFM), TBW, Desirable body fat ranges, and Segmental body fat information. They collected total 4661 dataset from Fasa’s rural region residents, a city in the eastern portion of the Fars province in southwest Iran. Trained 80%dataset. They used some algorithms like SVM, SGD, KNN, MLP, Ada boost and EDINet. They classified two types of classification. Those are diabetic and non-diabetic. Finally, they used some ensemble learning algorithms: Gradient boosting, Ada boost, Stacking and Voting. Here, they got more accuracy than before. Fat

mass and fat percentage were related factors of being diabetic. fat mass and fat percentage were related factors of being diabetic are more associated with the diabetic [36].

[37]. have researched “Techniques for Cardiovascular Disease by Machine Learning”. They used two different datasets from previous research. The first dataset digs out of 11 attributes like cardiovascular disease are high blood pressure, smoking, high cholesterol, diabetes, inactivity, obese, family history of heart attack or stroke, age, gender, diet, alcohol. An algorithm is Naïve Bayes and got accuracy is 89.77%, SVM predicted accuracy is 100%. The second dataset considered age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, physical activity, presence or absence of cardiovascular disease. Researchers focused to analysis various parameters and factors responsible for human death due to cardiovascular disease. This paper described various machine learning techniques used for classification and analysis of parameters of cardiovascular disease that are helpful in predicting and forecasting any unfortunate happening due to Cardiovascular disease. Researchers defined future work is researchers can work on trying different Deep Learning models which would be having a better accuracy rate, and they can try their research on the larger datasets. They also considered air pollution as a factor in prediction CVD [37].

[38].have completed yet another survey concerning. Researchers considered 30 attributes and it was divided by two parts. The first phase developed Support Vector Machine, accuracy is 95% and second phase developed neural network ANN, accuracy value is 86.6%. These algorithms to classify the dataset as pre-diabetic, diabetic or nondiabetic. analysis implies that patients with obesity, nonalcoholic fatty liver disease can lead to diabetes mellitus disease and more violent illnesses such as cirrhosis and mortality are expected to occur[38].

Table 2-1: Related work review

No	Dataset	Dataset Collected way	Data size	Features	Classes	Algorithms
#1	Abstract health parameter dataset	Open-source dataset repository from Kaggle	Not mention	Step count, mood, calories burned, hours of sleep, bool of active, weight kg, and BMI	Diabetes, Insomnia, Obesity, Anxiety, Hypertension and Cardiovascular (Proneness for six diseases)	KNN(K-Nearest Neighbor) & HMM(Hidden Markov Model)
#2	Tertiary hospital in South India	Random selection record of the heart patients with exclusion criteria	1670 anonymized medical records, Age 30-79, 881 males and 789 females.	Age, Weight, Height, Total cholesterol levels, Gender, Hypertension, Diabetes, Alcohol, Smoking habits, Exercise/physical activity, Stress, Family history of cardiovascular diseases Healthy diet Risk of CVD, BMI(Internal)	Binary classification: CVD risk- high and CVD risk- low	ML- Python 5 algorithms: K-Nearest Neighbours(KNN), Naive Bayes(NB), Logistic Regression(LR), AdaBoost, Random Forest (RF).Libraries: NumPy, Pandas, Seaborn, Statsmodels.api, SciPy & Sklearn
#3	Three South Asia countries	DHS surveys received ethical approval from the ICF	818,603	Age of the respondent, Level of BMI, Level of education, Wealth status, ever measured blood pressure, told by a doctor to have high blood	Non- hypertension, Hypertensive	Random Forest, Gradient Boosting Machine (GBM), LR,LDA, DT,

		Institutional Review Board and country-specific review boards & DHS website upon registration and request.		pressure, Taking prescribed medicine to lower blood pressure, prescribed medicine to lower blood pressure		
#4	Random sample of 200 adults	In urban field practice areas of a Medical College in central district of Delhi from August 2011 to January 2012, 3 physicians, 1 research officer, and 2 field volunteers was trained in data collection	200	Alcohol, smoking, waist circumference, overweight or obese	No	WHO STEPS questionnaire for assessing NCDs.

#5	Fasa cohort	Fasa’s rural region residents, a city in the eastern portion of the Fars province in southwest Iran,	4661 participants (571 diabetic models, 4090 nondiabetics models), age 35 - 70	Age, fat mass, and percentages in arms, legs, and trunk area	Diabetic, Non-diabetic	SVM, SGD, KNN, MLP, Ada boost and EDINet
#6		Previous 2 dataset:	Type 1: 303, Type 2: 70,001	Cardiovascular Disease are High Blood Pressure, Smoking, High Cholesterol, Diabetes, Inactivity, Obese, Family History of Heart Attack or Stroke, Age, Gender, Diet, Alcohol	Not in the paper	ML algorihms such as Support Vector machine (SVM), Artificial Neural Network (ANN)
#7	Al-Kasr, Al-Aini, Faculty of Medicine, Cairo University.	Real dataset	Not in the paper	Age, Sex, Schistosomiasis (Shisto), Alanine Aminotransferase (ALT), An aspartate aminotransferase (AST), Alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT) and nonalcoholic fatty Liver disease. The second phase: Nonalcoholic Fatty Liver disease attribute (output of phase one), Weight, Height, Waist Circumference (WC), Fasting Blood	Non-diabetic, Pre-diabetic and Diabetic patients	Support Vector Machine (SVM) and Artificial Neural Network (ANN)

				Sugar (FBS), History of Hypertension, History of Diabetes and Hemoglobin A1C (HBA1C).		

For NCDs as: cardiovascular illnesses, diabetes, musculoskeletal disorders, and types of cancers [4], it is a significant risk factor. Weight gain and body mass are essential to type 1 and type two diabetes development and increased incidence [38], Overweight and obesity are powerful risk factors for type 2 diabetes.

3. CHAPTER THREE- METHODOLOGY

3.1. Research design

Figure 3- 1 provides the suggested research methodology. The various stages of our research are the identification of the relevant factors, the selection of the most efficient factors, the collection of anthropometric data, the data preprocessing (data cleaning, data transforming), and the prediction of NCD patients or individuals using machine learning techniques. The following sections provide explanations of the various steps.

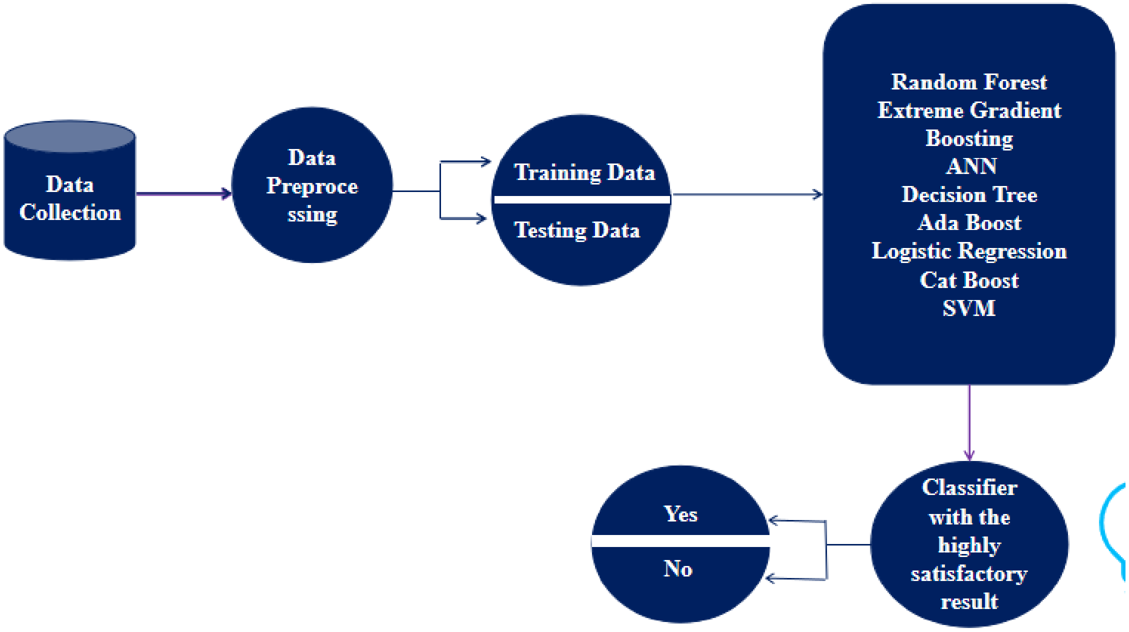


Figure 3-1: Propose methodology of our research

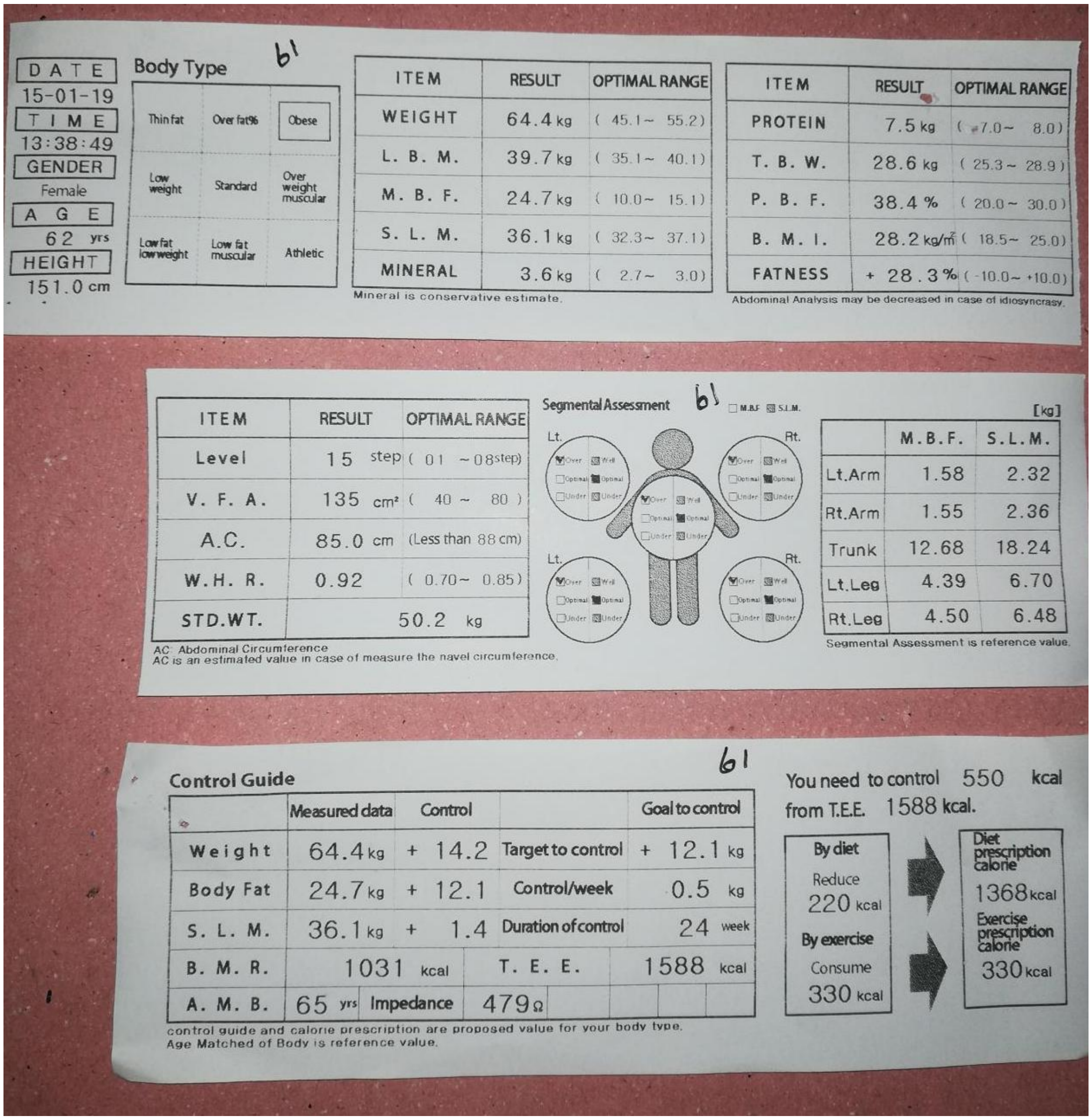
3.2. Data collection

The next stage in analyzing consumer expectations for a unique or customized product is to conduct a requirement analysis. Before designing any software project or model, it is critical to precisely identify the project's core needs/requirements. Capturing requirements may be done in a variety of ways. Many factors can be employed to attain the goal of obtaining facts. Interviews, brainstorming, questionnaires, web scraping, data consumption from a Web API, and so on are a few examples.

Researcher experiments are carried out with the anthropometric measurements for both female and male at Jaffna teaching hospital and Sabaragamuwa University of Sri Lanka.

Researcher collected 300 data set along with 164 male and 136 female. Researcher found the most effective factors for NCD with anthropometric assessment. Those factors are Age, Weight, Height, Gender, MBF, TBW, PBF, BMI, VFA, WHR. Researcher took these datasets from body composition analyzer machine. But the nutritionist took height weight from deferent scalar, then nutritionist input those two data sets in body composition analyzer.

Data collected with paper format, then researcher manually add those datasets in excel sheet to manually classify with the help of the nutritionist. After that, researcher converted this dataset in '.csv' file format.



DATE
15-01-19
TIME
13:38:49
GENDER
Female
AGE
62 yrs
HEIGHT
151.0 cm

Body Type 61

Thin fat	Over fat%	Obese
Low weight	Standard	Over weight muscular
Low fat low weight	Low fat muscular	Athletic

ITEM	RESULT	OPTIMAL RANGE
WEIGHT	64.4 kg	(45.1~ 55.2)
L. B. M.	39.7 kg	(35.1~ 40.1)
M. B. F.	24.7 kg	(10.0~ 15.1)
S. L. M.	36.1 kg	(32.3~ 37.1)
MINERAL	3.6 kg	(2.7~ 3.0)

Mineral is conservative estimate.

ITEM	RESULT	OPTIMAL RANGE
PROTEIN	7.5 kg	(7.0~ 8.0)
T. B. W.	28.6 kg	(25.3~ 28.9)
P. B. F.	38.4 %	(20.0~ 30.0)
B. M. I.	28.2 kg/m²	(18.5~ 25.0)
FATNESS	+ 28.3 %	(-10.0~ +10.0)

Abdominal Analysis may be decreased in case of idiosyncrasy.

ITEM	RESULT	OPTIMAL RANGE
Level	15 step	(01 ~08step)
V. F. A.	135 cm²	(40 ~ 80)
A.C.	85.0 cm	(Less than 88 cm)
W.H. R.	0.92	(0.70~ 0.85)
STD.WT.	50.2 kg	

AC: Abdominal Circumference
AC is an estimated value in case of measure the navel circumference.

Segmental Assessment 61 □ M.B.F ▨ S.L.M.

[kg]

	M.B.F.	S.L.M.
Lt.Arm	1.58	2.32
Rt.Arm	1.55	2.36
Trunk	12.68	18.24
Lt.Leg	4.39	6.70
Rt.Leg	4.50	6.48

Segmental Assessment is reference value.

Control Guide 61

	Measured data	Control		Goal to control
Weight	64.4 kg	+ 14.2	Target to control	+ 12.1 kg
Body Fat	24.7 kg	+ 12.1	Control/week	0.5 kg
S. L. M.	36.1 kg	+ 1.4	Duration of control	24 week
B. M. R.	1031 kcal		T. E. E.	1588 kcal
A. M. B.	65 yrs	Impedance	479Ω	

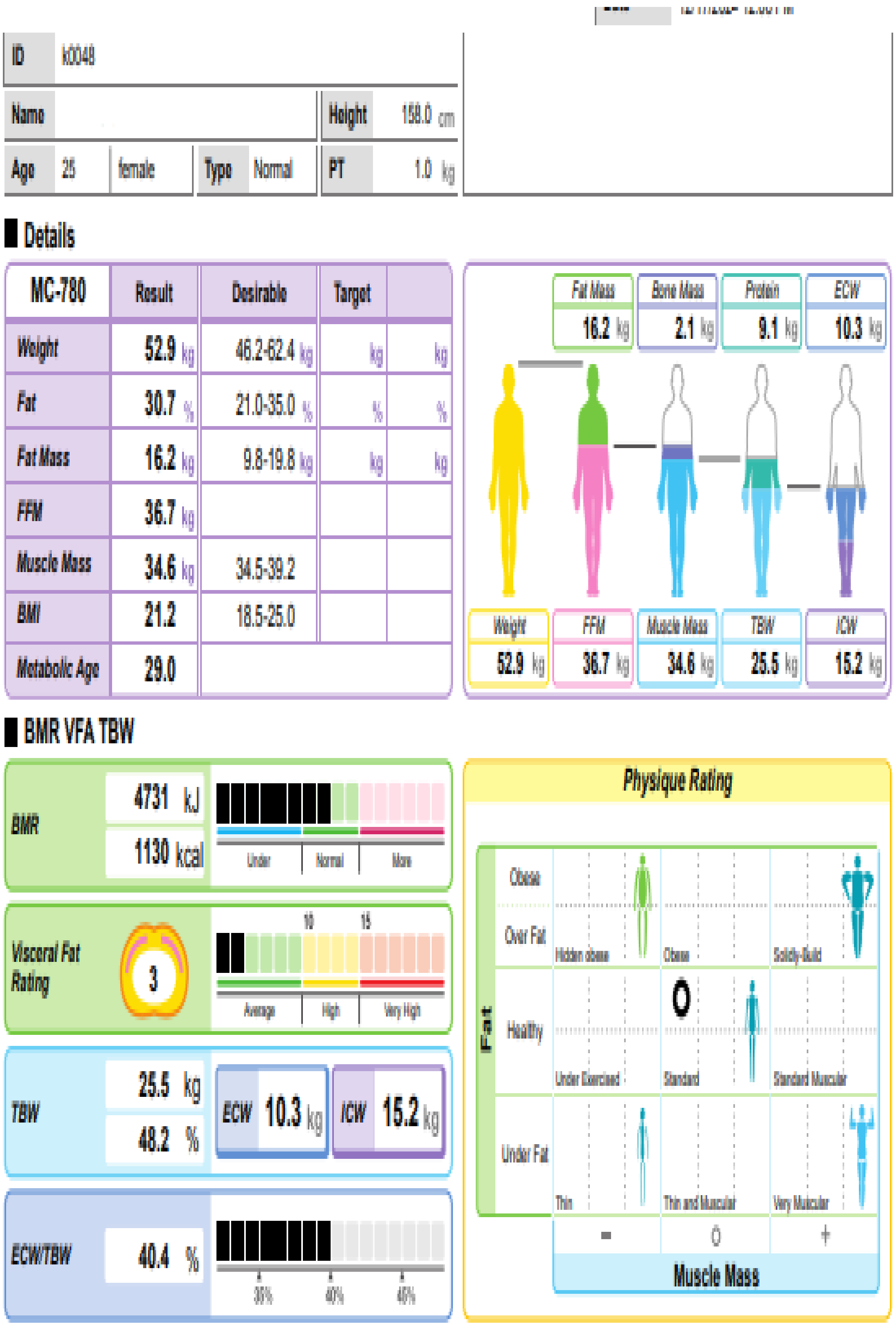
control guide and calorie prescription are proposed value for your body type.
Age Matched of Body is reference value.

You need to control 550 kcal
from T.E.E. 1588 kcal.

By diet
Reduce
220 kcal
By exercise
Consume
330 kcal

Diet prescription calorie
1368 kcal
Exercise prescription calorie
330 kcal

Figure 3-2: Body composition analyzer data format



ID	k0048				
Name				Height	158.0 cm
Age	25	female	Type Normal	PT	1.0 kg

Details

MC-780	Result	Desirable	Target	
Weight	52.9 kg	46.2-62.4 kg	kg	kg
Fat	30.7 %	21.0-35.0 %	%	%
Fat Mass	16.2 kg	9.8-19.8 kg	kg	kg
FFM	36.7 kg			
Muscle Mass	34.6 kg	34.5-39.2		
BMI	21.2	18.5-25.0		
Metabolic Age	29.0			

Fat Mass	Bone Mass	Protein	ECW
16.2 kg	2.1 kg	9.1 kg	10.3 kg

Weight	FFM	Muscle Mass	TBW	ICW
52.9 kg	36.7 kg	34.6 kg	25.5 kg	15.2 kg

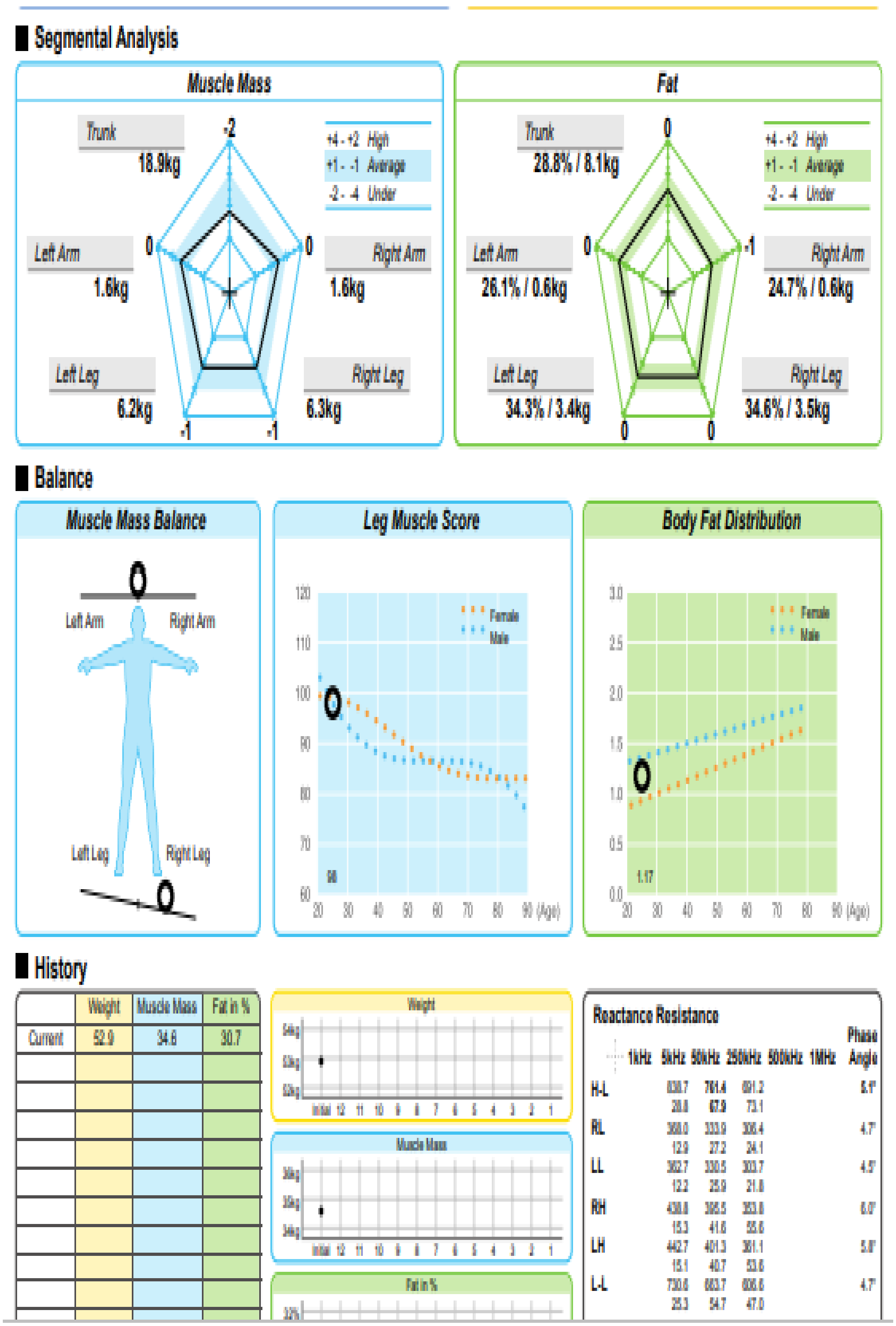
BMR VFA TBW

BMR	4731 kJ 1130 kcal	
Visceral Fat Rating	3	
TBW	25.5 kg 48.2 %	ECW 10.3 kg ICW 15.2 kg
ECW/TBW	40.4 %	

Physique Rating

Fat	-	0	+
Obese Over Fat	Hidden obese	Obese	Solidly-Build
Healthy	Under Exercised	Standard	Standard Muscular
Under Fat	Thin	Thin and Muscular	Very Muscular
		Muscle Mass	

Figure 3-3: TANITA body composition analyzer dataset format I



Segmental Analysis

Balance

History

	Weight	Muscle Mass	Fat in %
Current	52.9	34.6	30.7

Weight
54kg
53kg
52kg
Initial 12 11 10 9 8 7 6 5 4 3 2 1
Muscle Mass
36kg
35kg
34kg
Initial 12 11 10 9 8 7 6 5 4 3 2 1
Fat in %
32%

Reactance Resistance

	1kHz	5kHz	50kHz	250kHz	500kHz	1MHz	Phase Angle
H-L		838.7	761.4	691.2			5.1°
		28.8	67.9	73.1			
RL		368.0	333.9	306.4			4.7°
		12.9	27.2	24.1			
LL		362.7	330.5	303.7			4.5°
		12.2	25.9	21.8			
RH		438.8	395.5	353.8			6.0°
		15.3	41.6	55.6			
LH		442.7	401.3	361.1			5.8°
		15.1	40.7	53.6			
L-L		730.6	663.7	606.6			4.7°
		25.3	54.7	47.0			

Figure 3-4: TANITA body composition analyzer dataset format II

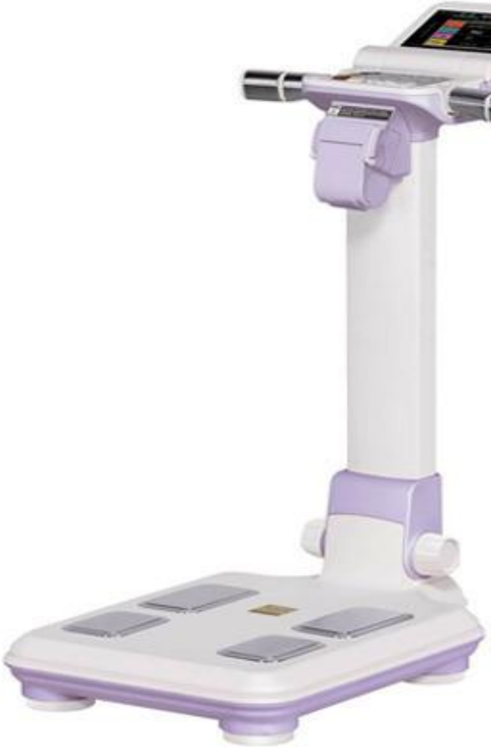


Figure 3-5: Body composition analyzer

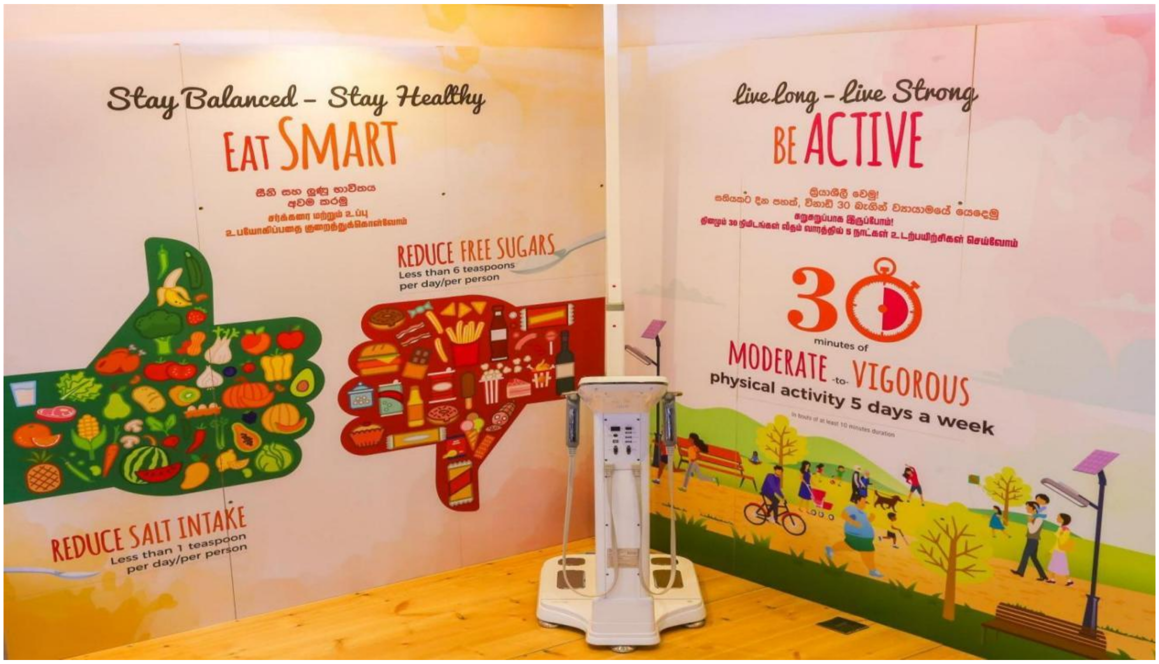

Stay Balanced – Stay Healthy
EAT SMART
සීනි සහ ලුණු භාවිතය අවම කරමු
சர்க்கரை மற்றும் உப்பு உபயோகிப்பதை குறைத்துக்கொள்வோம்
REDUCE FREE SUGARS
Less than 6 teaspoons per day/per person
REDUCE SALT INTAKE
Less than 1 teaspoon per day/per person
Live Long – Live Strong
BE ACTIVE
ක්‍රියාශීලී වෙමු!
සතියකට දින පහක්, විනාඩි 30 බැගින් ව්‍යායාමයේ යෙදෙමු
சுறுசுறுப்பாக இருப்போம்!
தினமும் 30 நிமிடங்கள் வீதம் வாரத்தில் 5 நாட்கள் உடற்பயிற்சிகள் செய்வோம்
30
minutes of
MODERATE -to- VIGOROUS
physical activity 5 days a week
In bouts of at least 10 minutes duration

Figure 3-6: Body composition analyzer front view

3.3. Factors identification

The NCDs are most affecting the factors are Age, Weight, Height, Gender, MBF, TBW, PBF, BMI, VFA, WHR. All are not healthy people, NCD most of the time will go through the human death. Researcher mention ten factors are most affecting the NCD with one person. So, that we no need to take the different biological tests like blood testing, urine testing and etc.

3.4. Identifying suitable measurement mechanism

There are different methods are available for NCD prediction. Anthropometric measurement is the unique way of the prediction of the NCDs.

3.5. Data preprocessing

The chosen collection of data needs to be filtered and cleaned before processing. The goal is to strengthen and increase the validity of the selected set of data. The element is crucial for giving the chosen data set the highest level of trustworthiness. Eliminating the specified barriers is essential in this situation because the cleansing procedure involves getting rid of any noise or obstacles that can hinder processing.

This may be a time-consuming process, but it effectively conveys the highest level of certainty on the quality and dependability of the data. Insufficient data sets can be obtained under some circumstances. However, in circumstances when precise forecasts are used to offset poor data sets, deliberate efforts are needed to achieve desired results. Consider a scenario where patient information that is extraneous to the procedure could be a hindrance. As a result, it is crucial to delete any such data at this time.

Data Pre-processing is the process of removing inaccurate, noisy, or inconsistent data from a data collection. Any prediction model development process must include this stage. There isn’t a pre-processing stage that applies to all data kinds. because different data sets are different from one another. Data produced by sounds and sights, for instance, are different. However, normalization is one of the pre-processing techniques that can be applied in a range of situations. In order to put the quality inside a predetermined range and reduce the variance between the available data, normalization has been planned. During the inquiry, data pre-processing was carried out utilizing the “python pandas library”. The aforementioned library was used to execute the following tasks.

1. Creating a CSV file from the raw data.
2. Using pandas to load the CSV.
3. Counting how many columns and rows there are
4. Removing columns that aren’t needed
5. Checking for null values and duplicates
6. Changing all variables to the same type
7. Detecting and replacing outliers
    - Outlier values for each and every column were initially found. (In this case, all columns’ First Quartile(Q1), Third Quartile(Q3), and Inter Quartile range (IQR) values were employed.)
    - For outlier discovered cells, random values are replaced with values within the quartile range.
    - Each column’s values have been validated to be within the quartile range limit.
    - The outlier values are removed from a fresh CSV.
8. The data set is normalized between a set of preset ranges.
    - In this case, the Min Max Scaling approach was used

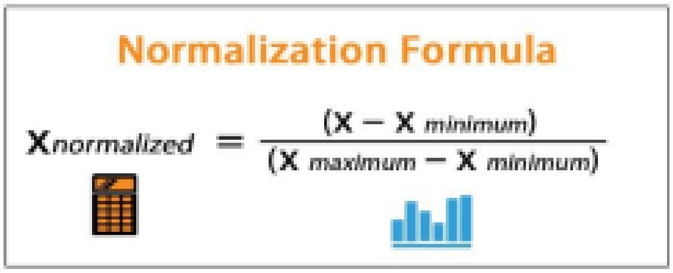

Normalization Formula
$X_{normalized} = \frac{(X - X_{minimum})}{(X_{maximum} - X_{minimum})}$

Figure 3-7: Normalization formula

9. Using a common scale to scale the data
    - In this case, the Python “StandardScaler” library has been modified

3.6. ML approach

3.6.1. Choose the adequate data mining task

The next step after following the above approach is to choose the appropriate data mining type for the project. The decision was made based on what could happen. Additionally, the choice between classifications, grouping, and regression is made based on the intended result. Typically, the results of data mining can either be projected, where the process is controlled by supervision, or described, where the process is visualized.

3.6.2. Choosing the right algorithm for data mining

The next step after following the above approach is to choose the appropriate data mining type for the project. The decision was made based on what could happen. Additionally, the choice between classifications, grouping, and regression is made based on the intended result. Typically, the results of data mining can either be projected, where the process is controlled by supervision, or described, where the process is visualized.

3.6.3. Choosing the right algorithm for data mining

The next step after following the above approach is to choose the appropriate data mining type for the project. The decision was made based on what could happen. Additionally, the choice between classifications, grouping, and regression is made based on the intended result. Typically, the results of data mining can either be projected, where the process is controlled by supervision, or described, where the process is visualized.

It appears that this stage is where the techniques to be employed in achieving the strategic goals are discussed. Is the employment of neural networks sufficient? Or is using judgments preferable? In actuality, the choice is made based on the type of searching pattern that is discovered to be consistent with the current project and the anticipated outcome. Does it have anything to do with the project objective, which is based on the outcomes' accuracy, comprehensibility, or decipherability? While decision trees are more suitable when getting a knowledge of patterns and trends is needed, neural networks are preferred more when the accuracy of the findings is necessary

3.7. Implementation

This section provides the details of the model implementation steps. We have proposed approaches for our NCD prediction patient. After preprocessing data, we applied ML algorithms. Researcher has applied most famous algorithm like Random Forest, Extreme Gradient Boosting, Artificial Neutral Network, Decision Tree, AdaBoost, Logistic Regression, Cat Boost and Support Vector Machine.

3.8. Testing and evaluation of the models

The anthropometric related 300 JTH (Jaffna Teaching Hospital) and Sabaragamuwa University of Sri Lanka dataset was divided into 8: 2. The 80 percentage of the data was trained and 20 percentage of the data tested. Testing data was used to evaluate the ML algorithm. The sample split was used for all experiments to ensure a fair comparison of the proposed models.

3.9. Prediction

ML algorithms learned the pattern of the dataset and then if we add the different dataset, it will automatically predict the NCD patient. It'll give the Yes/No values or zero/ one at the end of the prediction.

3.10. Research Instruments

A. Python

In recent years, Python has risen to become one of the most widely used programming languages worldwide. Everything from website development to software testing to machine learning uses it. Both developers and non-developers can use it.

The software that runs self-driving cars and Netflix's recommendation engine were both developed using Python, one of the most well-liked programming languages in the world. Since Python is a general-purpose language, it may be used for a variety of tasks, including data research, software and web development, automation, and everyday task completion.

Python is a popular computer programming language used to create software and websites, automate processes, and analyze data. Python is a general-purpose language, which means it may be used to make many various types of applications and isn't tailored for any particular issues. Its adaptability and beginner-friendliness have elevated it to the top of the list of programming languages in use today. It was ranked as the second-most popular programming language among developers in a poll by market research company Red Monk.

Some things consist of:

- Data analysis and artificial intelligence

- Scripting or automation in web development
- Software prototyping and testing
- Routine tasks

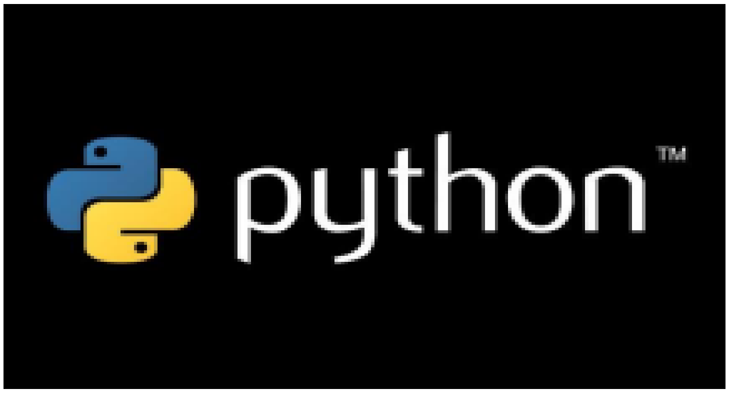


Figure 3-8: Python

B. Scikit-learn library

The most effective and reliable Python machine learning library is called Sklearn (Skit-Learn). Through a Python consistency interface, it offers a variety of effective tools for statistical modeling and machine learning, including classification, regression, clustering, and dimensionality reduction. This library is based on NumPy, SciPy, and Matplotlib and was written primarily in Python.

Figure 3-9: Scikit-learn library

C. PyCharm community edition:

An open-source IDE tool is PyCharm Community Edition, which is free. All popular operating systems are compatible with the software, which was created by JetBrain.

A slightly condensed version of the pricey Professional Edition is the Community Edition. Anyone learning or practicing Python coding will find it to be great. It is equipped with every tool required to write, debug, and run programs. You may easily test and inspect the code on the cutting-edge, user-friendly platform so that you can move your code into production.

Figure 3-10: PyCharm Community Edition

3.10.1. Machine learning techniques

We have realized that the Supervised Machine Learning approach might be a best suitable for our problem during our analysis. The JTHs & SUSL dataset of collected and then classified using ML classifiers in “Scikit-learn”. It is the most useful and robust library for ML in python. Researcher used ensemble algorithm like Random Forest, Extreme Gradient Boosting, Artificial Neutral Network, Decision Tree, AdaBoost, Logistic Regression, Cat Boost and Support Vector Machine. algorithms for this classification/ prediction. Grid Search Cross Validation method was used to find out the best parameters for each classifier.

3.10.2. Data mining algorithms

There are numerous varieties of prediction algorithms, according on past studies. The three most well-known data mining algorithms with the best chance of success are:

1. Random Forest
2. Extreme Gradient Boosting
3. Artificial Neutral Network
4. Decision Tree

5. Ada Boost
6. Logistic Regression
7. Cat Boost
8. Support Vector Machine methods

A. Support Vector Machine

As it is based on the Vapnik-Chervonenkis theory in calculating the linear regression function, where regression modeling derives the coefficients through minimizing the square error, Support Vector Machine (SVM) is one of the most potent supervised theoretical machine learning techniques. Medical diagnosis is one area in which SVM has been used. To locate the ideal hyperplane that divides the dataset classes in the middle of the maximum margin, the robustness of the SVM approach depends on finding the maximum geometrical margins between hyperplanes. An n-dimensional feature space is required in order to create a classification model from an n-dimensional collection of features. The ideal hyperplane divides the sample vectors into classes. The nearest points of both positive and negative classes to the ideal hyperplane, the support vectors, must be separated by a maximum margin. The greatest margins represent the least chance of misclassifying fresh data.

The goal of the SVM is to increase the space between the hyper-plane and the data point. the

function that helps maximize the margin is called loss function c(x,y,f(x)) given by training data (xi, yi) for i = 1 ...N, with xi ∈ Rd and yi ∈ {−1, 1}, learn a classifier f(x).

$$c(x, y, f(x)) = \begin{cases} 0, \text{if}\, y * f(x) \geq 1 \\ 1 - y * f(x), \text{if}\, y * f(x) < 1 \\ \end{cases}$$

$$c(x, y, f(x)) = (1 - y * f(x)) +$$

The price is $0. It determines the loss value using the cost function if the actual value and the anticipated value have the same sign and if the two values are not equal to zero. When regularization parameters are added to the cost function, the result is as shown below. The regularization parameter uses the formula (3.22), where Xi is the input, w is the weight, and λ is the regularization parameters, to balance margin maximization and loss:

$$\mathrm{minw}\lambda\|\mathrm{w}\|\ 2\ +\sum_{i=1}^{n}(1-\ \mathrm{yi}\ \langle\mathrm{xi}\ ,\mathrm{w}\rangle)+$$

B. Artificial Neural Networks (ANNs)

The usage of ANNs for modeling non-static processes has grown as a result of their high noise tolerance and capacity to classify non-visual patterns.

Three levels can be distinguished amongst the nodes in ANNs: input, output, and one or more hidden layers. The simulation of the model to access data from these units for classification, prediction, analysis, or any other treatment of input data is the fundamental idea behind artificial neural networks. Neural networks have demonstrated their ability to solve prediction difficulties. The hidden neurons decide what should be done with the input data that comes from the input layer. Additionally, these weights which are created through learning play a significant influence in the choice. By using the activation functions built into the neurons, the data may later be transmitted to the output layer. Each input value has a corresponding weight W estimation. This respect is a significant aspect because it updates during the neural training process so that it can provide indicators of improved behaviour. Figure (3-11) below provides an illustration of the general structure of ANNs:

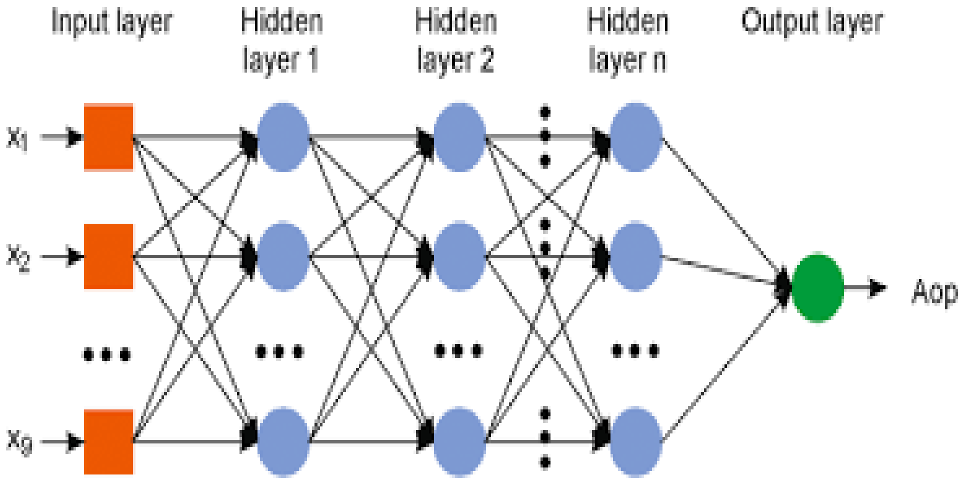


Figure 3-11: General model of multilayer artificial neural networks.

Three layers are represented by feed forward, multi-input multi-output in ANNs using the feed-forward back-propagation method, as seen in the following:

- Layer Xi, where I = 1, 2,..., n. the number of input nodes is indicated by n.
- Node I is the input node, and node j is the hidden layer node. Each node in the hidden layer j is a neuron, and each neuron is coupled to the input layer by the processing unit known as weights $w_{ij.}$
- Output layer k: Consists of the nodes that generate the output, which is represented by a number of neurons, and is dependent on a number of outputs, Y_k.

The total of weights from the input to the jth node in the hidden layer when the training phase was transmitted to the input layer is given by:

$$y = \sum W_{ij}X_i + \theta_j$$

θj: When employing ANNs, a node known as the bias node, which always has a value of 1, effectively calculates gradients using the back propagation process. A sigmoid function is commonly employed as the activation function in the back-propagation technique, which always starts from the output layer and propagates backward to update the weights. The j^{th} real output is:

$$Yj = Xk \ = \frac{1}{1 - e - y}$$

The output node actual value is Y_k, while the desired value is t_k, and X_k is the node that serves as the input to the following layer. The difference between these two values is expressed as k.

δ_k: Testing the models by using Δk and the sigmoid derivative, the error signal of the output layer is determined.

The change in the weight between node j and node k is done by multiplying the error at node k by the output of node j by using the delta rule

$$\Delta w_{jk} = l \ \delta_k X_k$$

w_{jk}: The following formula should be used to update the weight between nodes j and k, where l is the learning rate:

The error signal δ_j for node j in the hidden layer is calculated as follows: The following formula is used to determine the error signal for node j in the hidden layer:

$$\delta_j = (t_k - Y_k)\, Y_k \sum w_{jk}\, \delta_k$$

By employing 14 and 15, the weights between the input node I and the node j, $w_{ij,}$ can be modified.

$$\Delta w_{ij} = l\, \delta_j X_j$$

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

The back-propagation technique iterates until the output error is reduced to the absolute minimum.

C. Random Forest

Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. On various samples, it constructs decision trees and uses their average for classification and majority vote for regression.

The Random Forest is ability to handle data sets with both continuous variables, as in regression, and categorical variables, as in classification, is one of its most crucial qualities. In terms of classification issues, it delivers superior outcomes.

A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting numerous decision tree classifiers to distinct dataset subsamples. If bootstrap = True (the default), the size of the sub-sample is determined by the max samples argument; otherwise, each tree is constructed using the entire dataset.

To uncover patterns in massive data, data scientists employ a wide range of machine learning methods. They transform such information into useful insights for the businesses they work with. a data scientist will learn how to select the best algorithm for each problem as they gain experience. Random Forest, which is utilized for both classification and regression applications, is a very helpful method.

The random forest steps are as follows:

- Step 1: From a data set with k records, n random records are selected at random and used in the Random Forest algorithm.
- Step 2: For each sample, a unique decision tree is built.
- Step 3: An output will be produced by each decision tree.

- Step 4: For classification and regression, the final result is evaluated using a majority vote or an average.

Important Features of Random Forest

1. Diversity: Because each tree is unique, not all characteristics, variables, or features are taken into account when creating a particular tree.
2. Immune to the dimensionality curse: Since each tree only takes into account a subset of the features, the feature space is condensed.
3. Parallelization: Using various data and attributes, each tree is individually built. This implies that we can create random forests by using the CPU to its fullest extent.
4. In a random forest, we need to partition the data into train and test groups because there will always be 30% of the data that the decision tree cannot see.
5. Because the outcome is based on majority vote or average, there is stability.

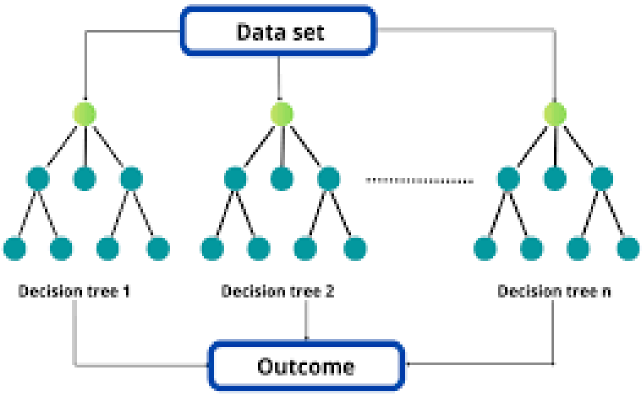


Figure 3-12: Random Forest

D. Extreme Gradient Boosting

The supervised branch of machine learning includes the tree-based method known as Extreme Gradient Boosting. All of the formulas and examples in this narrative refer to the use for classification problems, even though it may be applied for both classification and regression issues.

Basics of XGBoost recap the fundamentals of these algorithms before delving into the specifics.

- Decision trees are the basis estimators for both XGBoost tree-based methods.

- Prediction target: residuals rather than actual class labels are used to build the trees. The base estimators in these techniques are therefore regression trees rather than classification trees, despite the fact that we are focusing on classification problems. Because residuals are continuous rather than discrete, this is the case. However, several of the formulas you will read here are specific to classification; therefore, assume that they are also applicable to regression issues.
- To reduce the chance of overfitting the data, you can set the maximum size of the trees using either tree depth parameter.
- Similar to Random Forest or AdaBoost, ensemble approaches build numerous trees simultaneously. In the end, all of the trees contribute to the final prediction.
- Learning rate: Each value is scaled according to its learning rate. This makes it possible for the algorithm to advance more gradually and steadily with each step.
- Process map: Finally, here is a little explanation of how Gradient Boosting and XGBoost operate.

E. Decision Tree

The non-parametric supervised learning approach used for classification and regression applications is the decision tree. It is organized hierarchically and has a root node, branches, internal nodes, and leaf nodes.

A decision tree is a tool for modeling probable outcomes, resource costs, utility costs, and potential consequences. It has a tree-like structure. The presentation of algorithms with conditional control statements can be done using decision trees. They have branches that stand in for choices that might end in a good outcome.

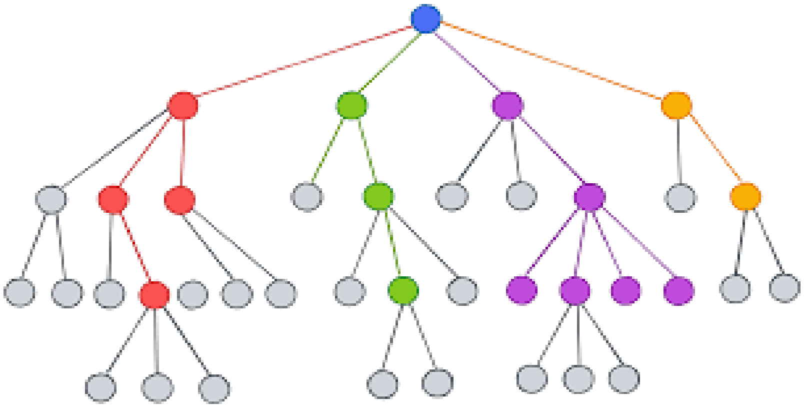


Figure 3-13: Decision Tree

F. Ada Boost

Since it was first introduced by Freund and Schapire in 1997, boosting is an ensemble modeling method that is frequently used to solve binary classification issues. By transforming a number of weak learners into strong learners, these methods increase prediction ability.

The basic idea behind boosting methods is that after creating a model using the training dataset, we create a second model to fix any mistakes in the original one. This process is repeated until the mistakes are reduced and the dataset can be accurately forecasted.

AdaBoost, also known as Adaptive Boosting, is a machine learning method used in an ensemble setting. Decision trees with one level, or Decision trees with only one split, are the most popular algorithm used with AdaBoost. Another name for these trees is Decision Stumps.

G. Logistic Regression

Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logit model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent range is 0 to 1. In logistic regression, the odds that is, the probability of success divided by the probability of failures are transformed using the logit formula. The following formulas are used to represent this logistic function, which is sometimes referred to as the log odds or the natural logarithm of odds:

$$\text{Logit(pi)} = \frac{1}{1 + \exp(-\text{pi})}$$

$$\ln(\text{pi}/(1\text{-pi})) = \text{Beta\_0} + \text{Beta\_1}*\text{X\_1} + \ldots + \text{B\_k}*\text{K\_k}$$

The dependent or response variable in this logistic regression equation is logit(pi), and the independent variable is x. Most frequently, maximum likelihood estimation is used to calculate the beta parameter, or coefficient, in this model (MLE). This approach evaluates various beta values over a number of iterations to get the best match for the log odds. Logistic regression aims to maximize this function after each of these iterations in order to determine the optimal parameter estimate. Once the best coefficient (or coefficients, if there are multiple independent variables) has been identified, the conditional probabilities for each observation can be computed, logged, and added to produce a predicted probability. For binary classification, a probability of.5 or less will forecast 0 whereas a chance of more will forecast. If the categorization is binary, a probability of less than.5 predicts 0 and a probability of more than 0 predicts 1. It is recommended to assess the goodness of fit, or how well it predicts the dependent variable, once the model has been computed. One common technique for evaluating model fit is the Hosmer-Leme show test.

H. Cat Boost

Yandex's Cat Boost is a recently released machine learning algorithm. It is simple to interface with deep learning frameworks such asApple's Core ML andGoogle's TensorFlow. It can operate with various data formats to assist in resolving a variety of issues that businesses are currently facing. Additionally, it offers industry-leading precision. According to Mikhail Bilenko, director of machine learning and research at Yandex, this is the first open source Russian machine learning technology.

It is particularly effective in two ways: It produces cutting-edge results without the substantial data training that other machine learning techniques normally demand, and it offers strong out-of-the-box support for the more descriptive data formats that go along with many commercial problems.

The name "Cat Boost" is a combination of the phrases "Category" and "Boosting. As was mentioned, the library functions effectively with a variety of data categories, including audio, text, and image files as well as historical data.

As this library is based on the gradient boosting library, the word "boost" is derived from the machine learning algorithm for gradient boosting. Gradient boosting is a potent machine learning method that has been successfully used for a variety of commercial difficulties, including fraud detection, product recommendations, and forecasting. In contrast to DL models, which must learn from enormous amounts of data, it can also produce excellent results with relatively little data.

I. Testing the models

The accuracy, precision, recall, and F1-score metrics are computed and compared in order to further assess the overall classification performance of the suggested models after they have been successfully developed and trained. These performance measures are simply explained with the equations below, where TP and FN stand for the proportion of properly classified and incorrectly classified raga tasks, respectively; FP stands for the number of times a particular raga classification task is incorrectly classified as this type; TN stands for the number of raga classification tasks not belonging to a specific class that are not classified as this class.

a) Accuracy

One of the often-used assessment measures for classification problems is accuracy, which computes the total number of accurate predictions divided by the total number of predictions for a dataset.

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

b) Precision

A measure of precision counts how many correctly positive forecasts were made. Therefore, precision determines the accuracy for the minority class. It is determined by dividing the total number of correctly anticipated positive examples by the ratio of correctly predicted positive examples.

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

c) Recall

Recall is a metric that measures the proportion of accurate positive predictions among all possible positive predictions. Recall gives an indicator of missed positive predictions, unlike precision, which only comments on the accurate positive predictions out of all positive predictions. Recall gives an idea of the positive coverage in this way.

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

d) F1-score

Precision and recall can be combined into one metric using F-Measure, which covers both characteristics. A way to communicate both concerns with a single score is offered by the F-measure. Calculating precision and recall is possible using the confusion matrix. After that, F-score7 is calculated as the harmonic mean of recall and precision.

$$F1 - score = 2 \times \frac{\text{precision} \times \text{recal}l}{\text{precision} + \text{recal}l}$$

4. CHAPTER FOUR- RESULT AND FINDING

Results obtained during the performance comparison of several algorithms in predicting the NCD patients.

Numerous algorithms have been tweaked and assessed against certain assessment measures in order to predict the final results. Later, a more thorough explanation of the comparative results obtained using this methodology will be provided.

This approach involved looking at 300 patient anthropometric test data with 10 features and one target variable. 70% of the training datasets and 30% of the testing datasets are randomly selected.

The language of choice for handling machine learning algorithms is Python. In this work, Scikit-learn was used as the machine learning library. Open-source Scikit-learn is a Python library.

The table below displays the models'evaluation metrics findings.

Table 4-1: Algorithm's accuracy, mean squared error and absolute squared error

Algorithms	Accuracy(%)	Mean Squared Error [MSE] (%)	Absolute Squared Error [ASE] (%)
Random Forest	98.90	1.09	1.09
Extreme Gradient Boosting	97.80	2.19	2.19
ANN	97.80	2.19	2.19
Decision Tree	96.05	3.94	3.94
Ada Boost	93.40	6.59	6.59
Logistic Regression	88.52	11.47	11.47
Cat Boost	87.91	12.08	12.08
SVM	85.24	14.75	14.75

4.1. Evaluation of the machine learning models

Researcher below mention evaluation table (precision recall f1-score and support) zero means the patient/ person don’t have NCD, one means person have NCD.

Table 4-2: Evaluation of Random Forest

	Precision	**Recall**	**F1-Score**	**Support**
0	0.97	1.00	0.99	33
1	1.00	0.98	0.99	58
Accuracy			0.99	91
Macro Avg	0.99	0.99	0.99	91
Weighted Avg	0.99	0.99	0.99	91

Table 4-3: Evaluation of Extreme Gradient Boosting

	Precision	**Recall**	**F1-Score**	**Support**
0	0.97	0.97	0.97	33
1	0.98	0.98	0.98	58
Accuracy			0.98	91
Macro Avg	0.98	0.98	0.98	91
Weighted Avg	0.98	0.98	0.98	91

Table 4-4: Evaluation of Artificial Neural Network

	Precision	Recall	F1-Score	Support
0	1.00	0.94	0.97	35
1	0.97	1.00	0.98	56
Accuracy			0.98	91
Macro Avg	0.98	0.97	0.98	91
Weighted Avg	0.98	0.98	0.98	91

Table 4-5: Evaluation of Decision Tree

	Precision	Recall	F1-Score	Support
0	0.87	0.96	0.92	28
1	0.98	0.92	0.95	48
Accuracy			0.93	76
Macro Avg	0.92	0.94	0.93	76
Weighted Avg	0.94	0.93	0.93	76

Table 4-6: Evaluation of AdaBoost

	Precision	Recall	F1-Score	Support
0	0.89	0.97	0.93	40
1	0.98	0.90	0.94	51
Accuracy			0.93	91

Macro Avg	0.93	0.94	0.93	91
Weighted Avg	0.94	0.93	0.93	91

Table 4-7: Evaluation of Logistic Regression

	Precision	**Recall**	**F1-Score**	**Support**
0	0.83	0.93	0.88	27
1	0.94	0.85	0.89	34
Accuracy			0.89	61
Macro Avg	0.88	0.89	0.88	61
Weighted Avg	0.89	0.89	0.89	61

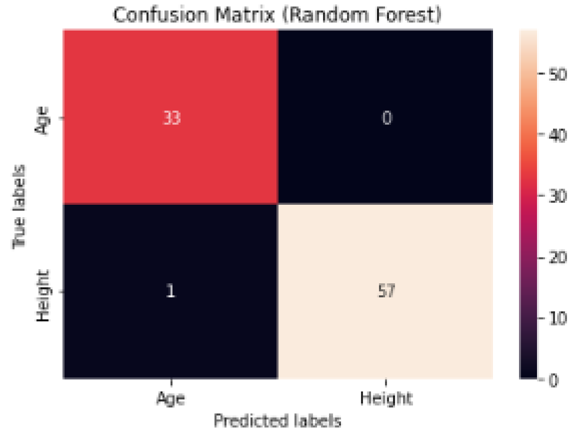
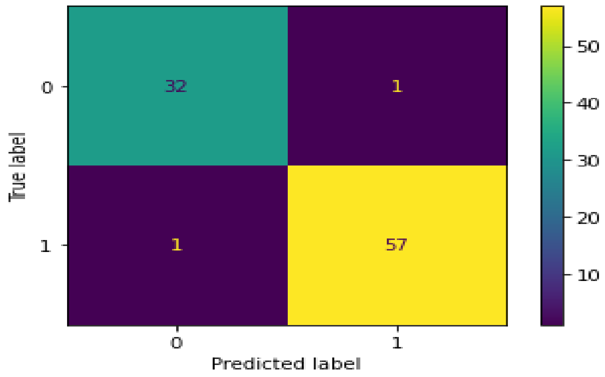
Table 4-8: Evaluation of Cat Boost

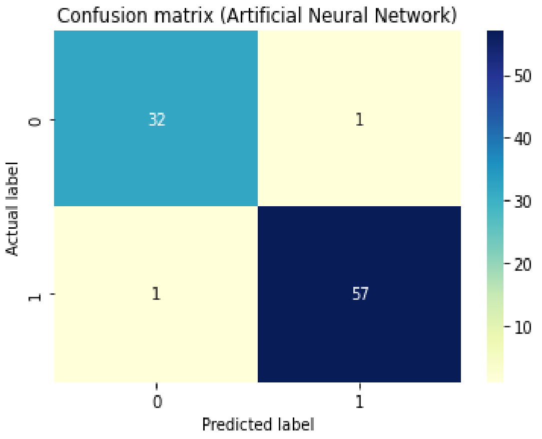
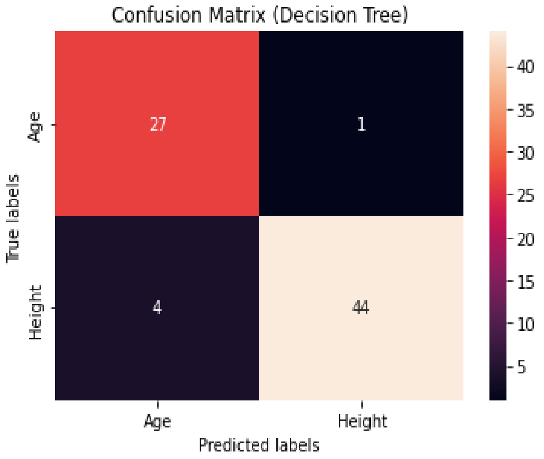
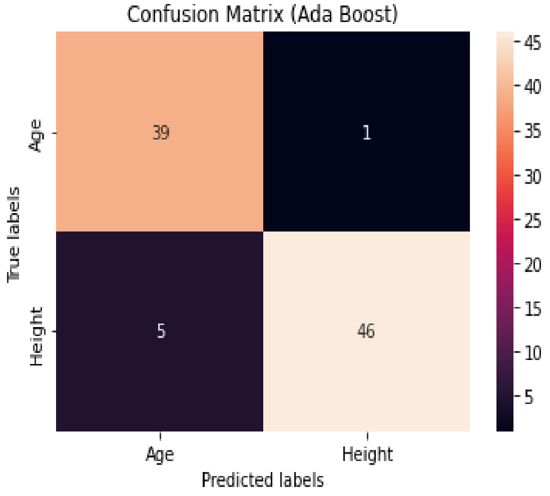
	Precision	**Recall**	**F1-Score**	**Support**
0	0.84	0.90	0.87	40
1	0.92	0.86	0.89	51
Accuracy			0.88	91
Macro Avg	0.88	0.88	0.88	91
Weighted Avg	0.88	0.88	0.88	91

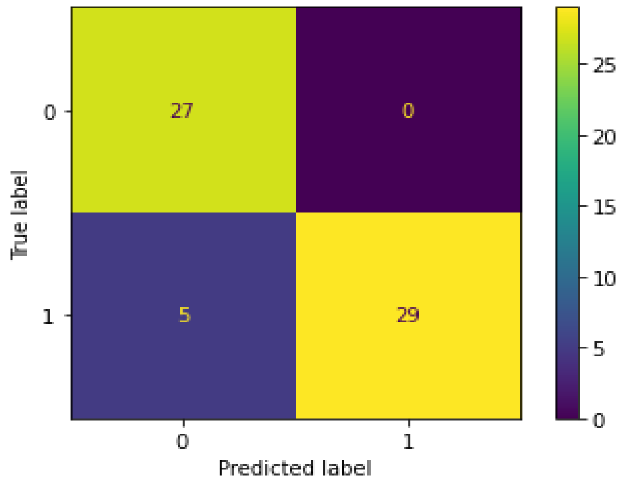
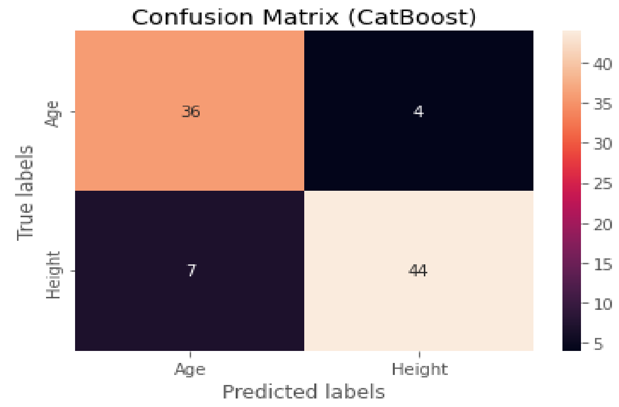
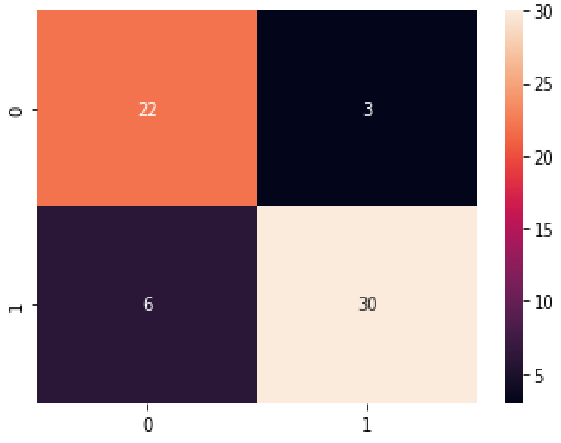
Table 4-9: Evaluation of Support Vector Machine

	Precision	Recall	F1-Score	Support
0	0.76	0.79	0.78	24
1	0.86	0.84	0.85	37
Accuracy			0.82	61
Macro avg	0.81	0.81	0.81	61
Weighted avg	0.82	0.82	0.82	61

Table 4-10: Confusion matrix obtained for individual classifier

Classifier	Confusion Matrix
Random Forest	
Extreme Gradient Boosting	Confusion Matrix (Extreme Gradient Boost) 

ANN	
Decision Tree	
AdaBoost	

<table>
<tr>
<td>Logistic Regression</td>
<td>
</td>
</tr>
<tr>
<td>Cat Boost</td>
<td>
</td>
</tr>
<tr>
<td>SVM</td>
<td>
</td>
</tr>
</table>

4.1.1. Correlation of Aba Boost

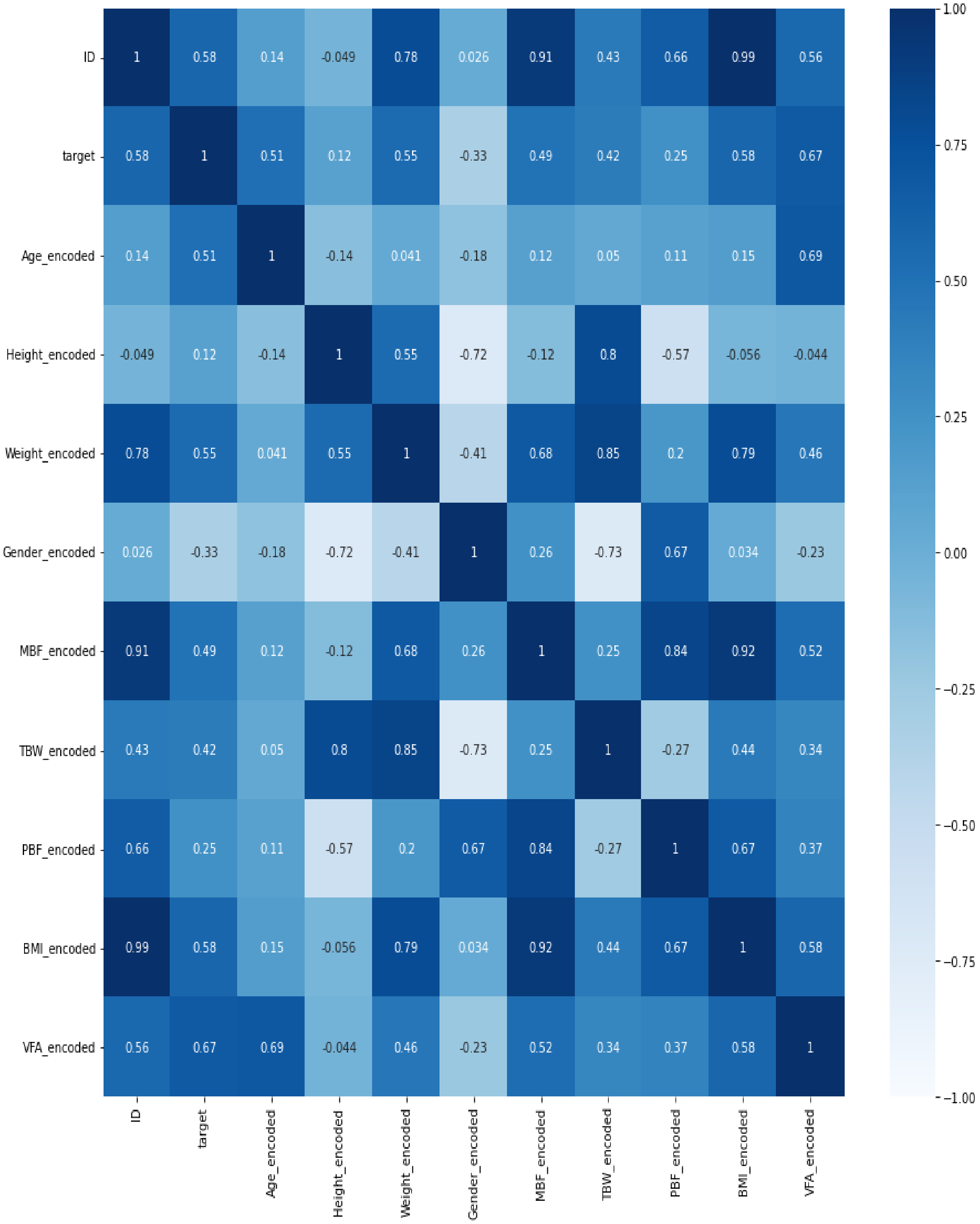


Figure 4-1: Correlation of the SVM

4.1.2. Features

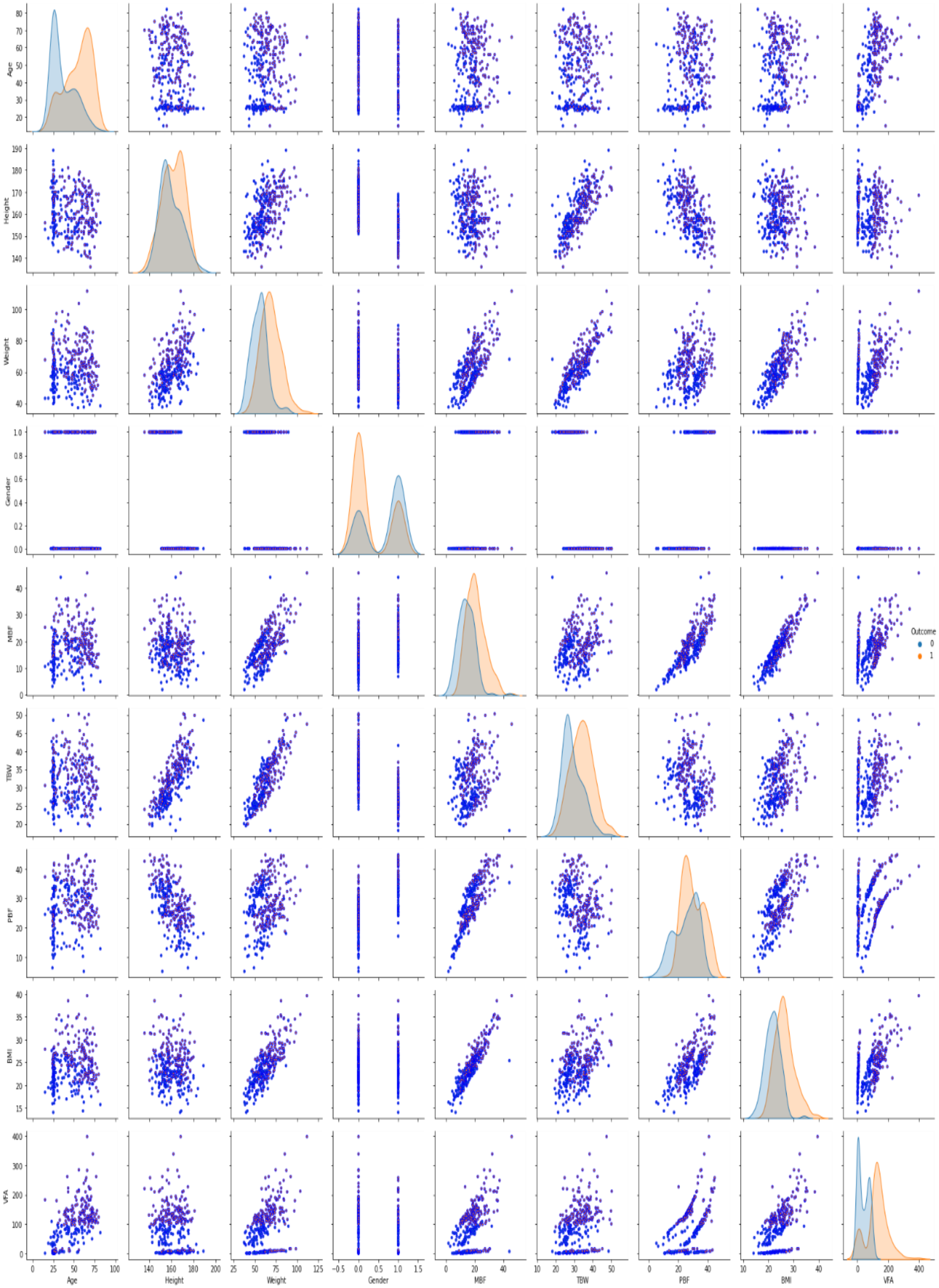


Figure 4-2: Features extraction from Random Forest

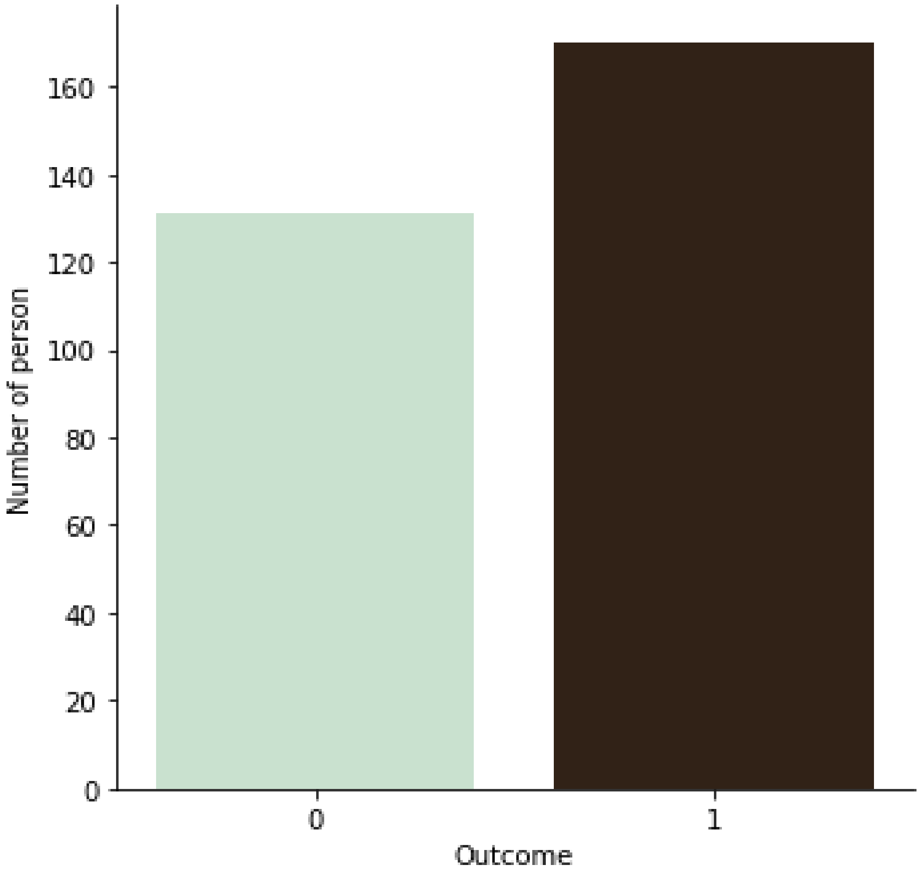


Figure 4-3: Final outcome visualization

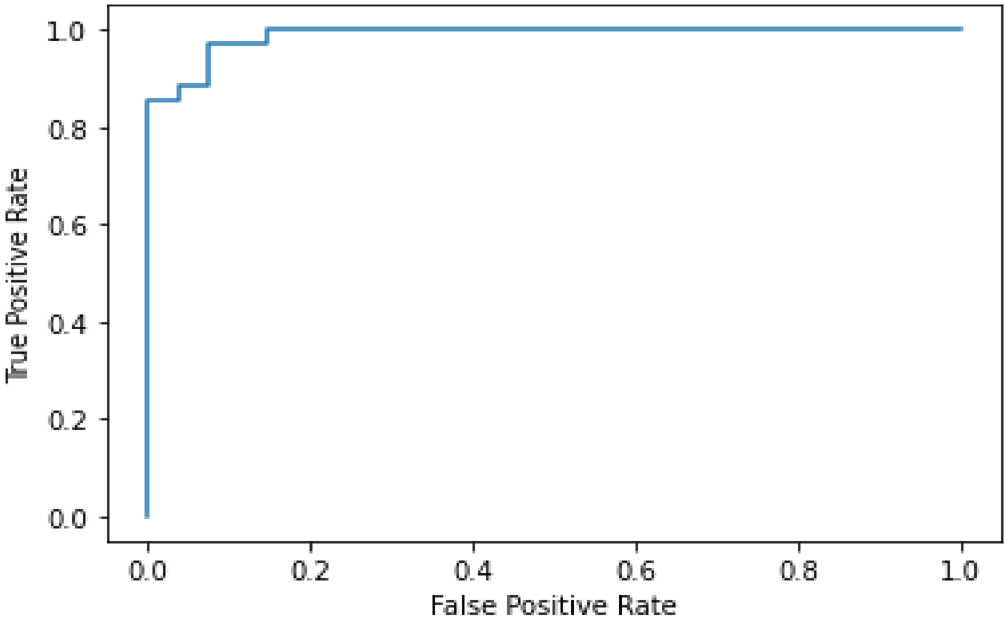


Figure 4-4: Logistic Regression true positive rate vs false positive rate

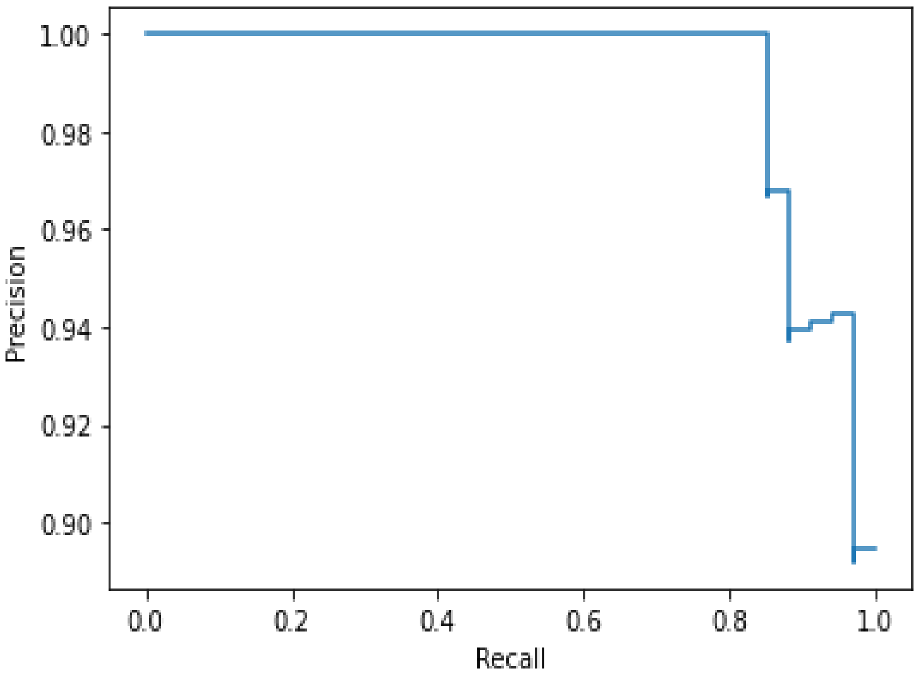


Figure 4-5: Logistic regression precision vs recall

5. CHAPTER FIVE- CONCLUSION, DISCUSSION AND RECOMMENDATION

The researchers proposed a method for predicting NCD through ML algorithms and the data was collected two different province and totally collected data were 300 at Jaffna Teaching Hospital, Sabaragamuwa University of Sri Lanka to verify the performance of the proposed method. The researcher developed using the eight algorithm of ML such as Random Forest, Extreme Gradient Boosting, ANN, Decision Tree, AdaBoost, Logistic Regression, Cat Boosting, in addition to the SVM and the researcher got accuracies are 98.90%, 98.90%, 97.80%, 97.80%, 96.05%, 93.40%, 88.52%, 87.91%, 85.24%. The highest accuracy is 98.90%. The researcher also evaluates the evaluation metrics precision, recall, f1-score and support. The random forest model achieved mean squared error [MSE] is 1.09% and absolute squared error [ASE] also 1.09%.

According to the modeling results, the developed Random Forest approach outperformed other methods used to predict the NCD person with ten selected anthropometric parameters. Due to the high efficiency of prediction, the resulting knowledge was understandable and implementable so that it could be used and applied to a dataset of the Sri Lankan population.

5.1. Discussion

The study aimed to predict the NCD person or patient using minimal data and a high accuracy rate using data mining algorithms, such as Random Forest, Extreme Gradient Boosting, ANN, Decision Tree, AdaBoost, Logistic Regression, Cat Boosting, in addition to the SVM.

In recent years, NCD analysis is one of most essential tests because some leading causes of death (80%) in most nations throughout the world such as heart disease, diabetes, and various types of cancer have been recognized and used to evaluate MBF (Mass of Body Fat), TBW (Total Body Water), PBF (Percentage Body Fat), BMI (Body Mass Index), VFA (Visceral Fat Area), WHR (Waist Hip Ratio) finally compare the all values with Gender, Age categorical along with the WHO (World Health Organization) values for

NCD. Anthropometric measurement is not frequently conducted in a medical lab or clinic, but it is better to measure frequently once a month.

By employing the suggested method, additional cost savings in healthcare can be achieved while maximizing the use of vital medical resources and removing the need to manually calculate NCD patients or manually record information on paper. All patients will have access to this program as a one-stop shop for accurate and calculated health information. The use of this application is seen as a benefit for doctors so that they can ensure the accuracy and professionalism of their detection or inferences. Additionally, this data can be used globally to provide medical histories to people who need to travel abroad but have health issues.

5.2. Future work

Still this research has drawbacks that must be addressed in the future research will focus on two areas. First, the proposed model only has 300 datasets. Collecting enough labeled data, on the other hand, takes time and effort. Based on this limitation, the proposed model can be further increased the dataset, which will improve its classification ability on the data sets. Second one predicts the personal nutritional level by using machine learning approach for Asian people. This would be a promising start for a future research scholar to adapt the hybrid techniques and the deep learning techniques to identify the NCD patient/ person along with the prediction of the type of the personal nutritional level.

References

[1] P. E. Petersen,"The World Oral Health Report 2003: continuous improvement of oral health in the 21st century–the approach of the WHO Global Oral Health Programme" *Community Dentistry and oral epidemiology,* vol. 31, pp. 3-24, 2003.

[2] M. Ackland, B. Choi, and P. Puska,"Rethinking the terms non- communicable disease and chronic disease" *Journal of Epidemiology & Community Health,* vol. 57, pp. 838-839, 2003.

[3] S. Li and I. Laher,"Rethinking "Exercise is Medicine"" *EXCLI journal,* vol. 19, p. 1169, 2020.

[4] W. H. Organization *communicable and non- communicable diseases, and mental health*. Available: https://www.who.int/our-work/ communicable-and-non communicable-diseases-and-mental-health

[5] D. Abegunde and A. Stanciole,"An estimation of the economic impact of chronic non- communicable diseases in selected countries, " *World Health Organization, Department of Chronic Diseases and Health Promotion,* vol. 2006, 2006.

[6] W. H. Organization,"WHO package of essential non- communicable (PEN) disease interventions for primary health care" 2020.

[7] Y. Yang, S. Wang, L. Chen, M. Luo, L. Xue, D. Cui*, et al.,*"Socioeconomic status, social capital, health risk behaviors, and health-related quality of life among Chinese older adults" *Health and Quality of Life Outcomes,* vol. 18, pp. 1-8, 2020.

[8] F. Branca, A. Lartey, S. Oenema, V. Aguayo, G. A. Stordalen, R. Richardson*, et al.,*"Transforming the food system to fight non- communicable diseases" *Bmj,* vol. 364, 2019.

[9] K. Juma, P. A. Juma, S. F. Mohamed, J. Owuor, A. Wanyoike, D. Mulabi*, et al.,*"First Africa non- communicable disease research conference 2017: sharing evidence and identifying research priorities" *Journal of global health,* vol. 8, 2019.

[10] R. Thibault, L. Genton, and C. Pichard,"Body composition: why, when and for who?" *Clinical nutrition,* vol. 31, pp. 435-447, 2012.

[11] L. C. Ward, "Human body composition: yesterday, today, and tomorrow" *European journal of clinical nutrition,* vol. 72, pp. 1201-1207, 2018.

[12] X. He, Z. Li, X. Tang, L. Zhang, L. Wang, Y. He*, et al.*,"Age-and sex-related differences in body composition in healthy subjects aged 18 to 82 years" *Medicine,* vol. 97, 2018.

[13] K. Lee, Y. Shin, J. Huh, Y. S. Sung, I.-S. Lee, K.-H. Yoon*, et al.*,"Recent issues on body composition imaging for sarcopenia evaluation" Korean journal of radiology, vol. 20, pp. 205-217, 2019.

[14] F. Campa, S. Toselli, M. Mazzilli, L. A. Gobbo, and G. Coratella,"Assessment of body composition in athletes: A narrative review of available methods with special reference to quantitative and qualitative bioimpedance analysis"*Nutrients,* vol. 13, p. 1620, 2021.

[15] F. M. Stich, S. Huwiler, G. D'Hulst, and C. Lustenberger,"The Potential Role of Sleep in Promoting a Healthy Body Composition: Underlying Mechanisms Determining Muscle, Fat, and Bone Mass and Their Association with Sleep"*Neuroendocrinology,* vol. 112, pp. 673-701, 2022.

[16] R. Kuriyan,"Body composition techniques" *The Indian journal of medical research,* vol. 148, p. 648, 2018.

[17] B. Masanovic,"Comparative study of anthropometric measurement and body composition between junior basketball and volleyball players from Serbian national league" *Sport Mont,* vol. 16, pp. 19-24, 2018.

[18] L.-L. B. Mass,"A Comparative Study on Health Status of Male and Female Agricultural Workers" *Productivity with Health, Safety, and Environment: Select Proceedings of HWWE 2019,* p. 79, 2022.

[19] N. Nešić, V. Šeper, and E. D. Cvetko, "Body composition changes during eight weeks of aerobic, strength or combined aerobic-strength training" *Editors-in-Chief: Dario Škegro*.

[20] S. Ghannadi, F. Halabchi, F. Maleklou, Z. Tavakol, M. Rajabian Tabesh, D. Bala*, et al.*,"The effect of 6 weeks electrical muscle stimulation training and aerobic exercise on body composition of overweight women: a randomized controlled study" *Sport Sciences for Health,* pp. 1-9, 2022.

[21] A. Bagheri, S. M. Nachvak, H. Abdollahzad, and M. Rezaei, “Inflammatory Potential of Diet and the Risk of Prostate Cancer: A Case-control Study in the West of Iran” *Current Nutrition & Food Science,* vol. 15, pp. 718-724, 2019.

[22] J. Mason, “Nutritional principles and assessment of the gastroenterology patient” 2015.

[23] A. Bohr and K. Memarzadeh, “The rise of artificial intelligence in healthcare applications” in *Artificial Intelligence in healthcare*, ed: Elsevier, 2020, pp. 25-60.

[24] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, “Significance of machine learning in healthcare: Features, pillars and applications” *International Journal of Intelligent Networks,* 2022.

[25] K. Santosh and L. Gaur, *Artificial Intelligence and Machine Learning in Public Healthcare: Opportunities and Societal Impact*: Springer Nature, 2022.

[26] S. Sun, Z. Cao, H. Zhu, and J. Zhao, “A survey of optimization methods from a machine learning perspective” *IEEE transactions on cybernetics,* vol. 50, pp. 3668-3681, 2019.

[27] A. Abdelaziz, M. Elhoseny, A. S. Salama, and A. Riad, “A machine learning model for improving healthcare services on cloud computing environment” *Measurement,* vol. 119, pp. 117-128, 2018.

[28] M. D. Abràmoff, B. Cunningham, B. Patel, M. B. Eydelman, T. Leng, T. Sakamoto*, et al.*,“Foundational considerations for artificial intelligence using ophthalmic images” *Ophthalmology,* vol. 129, pp. e14-e32, 2022.

[29] M. A. Ahmad, C. Eckert, and A. Teredesai, “Interpretable machine learning in healthcare” in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559-560.

[30] A. Aldallal and A. A. A. Al-Moosa, “Using data mining techniques to predict diabetes and heart diseases” in *2018 4th International Conference on Frontiers of Signal Processing (ICFSP)*, 2018, pp. 150-154.

[31] M. Nabeel, S. Majeed, M. J. Awan, H. Muslih-ud-Din, M. Wasique, and R. Nasir, “Review on Effective Disease Prediction through Data Mining Techniques” *International Journal on Electrical Engineering & Informatics,* vol. 13, 2021.

[32] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda,“Machine learning and data mining methods in diabetes research” *Computational and structural biotechnology journal,* vol. 15, pp. 104-116, 2017.

[33] P. Kadu and A. Buchade, “Non communicable Disease Prediction System Using Machine Learning”

[34] E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, “Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India” *Medical Journal Armed Forces India,* vol. 77, pp. 302-311, 2021.

[35] S. M. S. Islam, A. Talukder, M. A. Awal, M. M. U. Siddiqui, M. M. Ahamad, B. Ahammed*, et al.*, “Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries” *Frontiers in Cardiovascular Medicine,* vol. 9, 2022.

[36] M. A. Nematollahi, A. Askarinejad, A. Asadollahi, M. Salimi, M. Moghadami, S. Sasannia*, et al.*, “Association and predictive capability of body composition and diabetes mellitus using artificial intelligence: a cohort study” 2022.

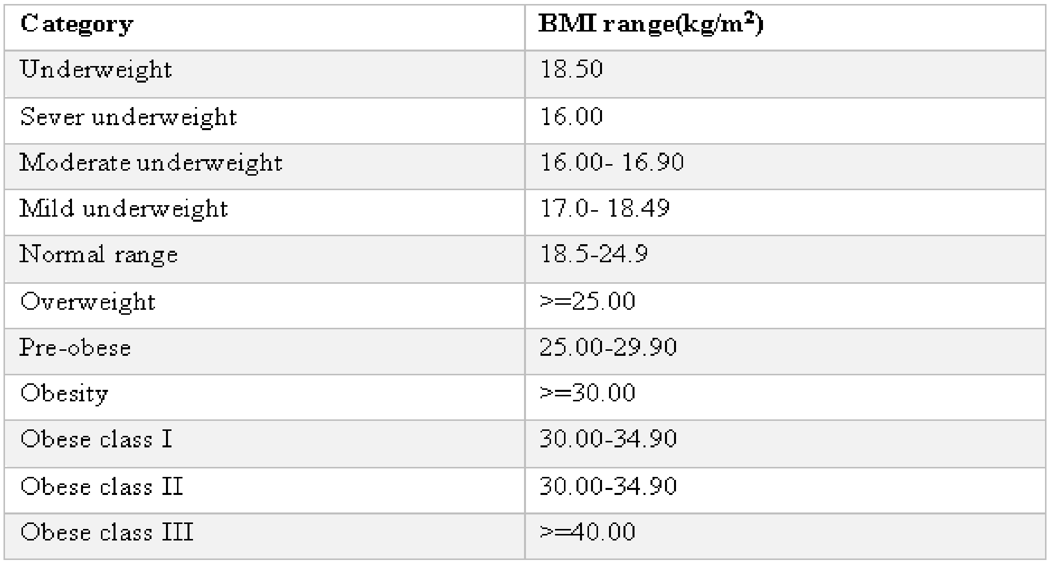
[37] K. Tripathi and H. Garg, “Machine Learning techniques for Cardiovascular Disease” in *IOP Conference Series: Materials Science and Engineering*, 2021, p. 012140.

[38] R. E. Ali, H. El-Kadi, S. S. Labib, and Y. I. Saad, “Prediction of potential-diabetic obese-patients using machine learning techniques” 2019.

Appendix – A: Normal Range for BMI of Asian People

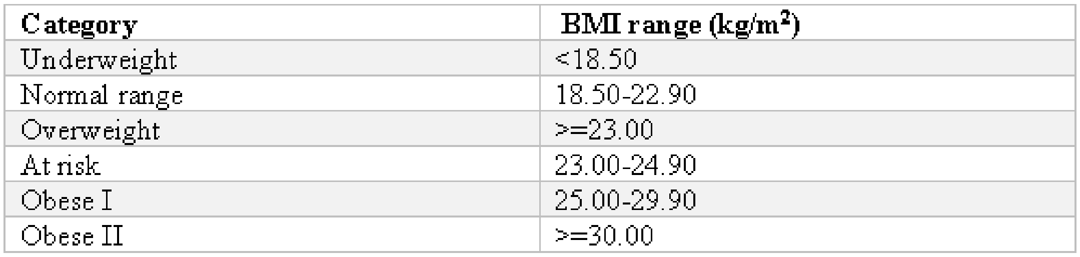
The BMI is different for every country people, because their food habits and culture.

Table 1: BMI normal range of adults for Asian



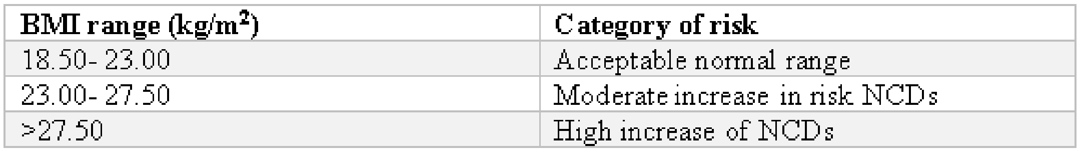
Category	BMI range(kg/m^2)
Underweight	18.50
Sever underweight	16.00
Moderate underweight	16.00- 16.90
Mild underweight	17.0- 18.49
Normal range	18.5-24.9
Overweight	>=25.00
Pre-obese	25.00-29.90
Obesity	>=30.00
Obese class I	30.00-34.90
Obese class II	30.00-34.90
Obese class III	>=40.00

Table 2: BMI classification for Adult Asians of obesity, WHO



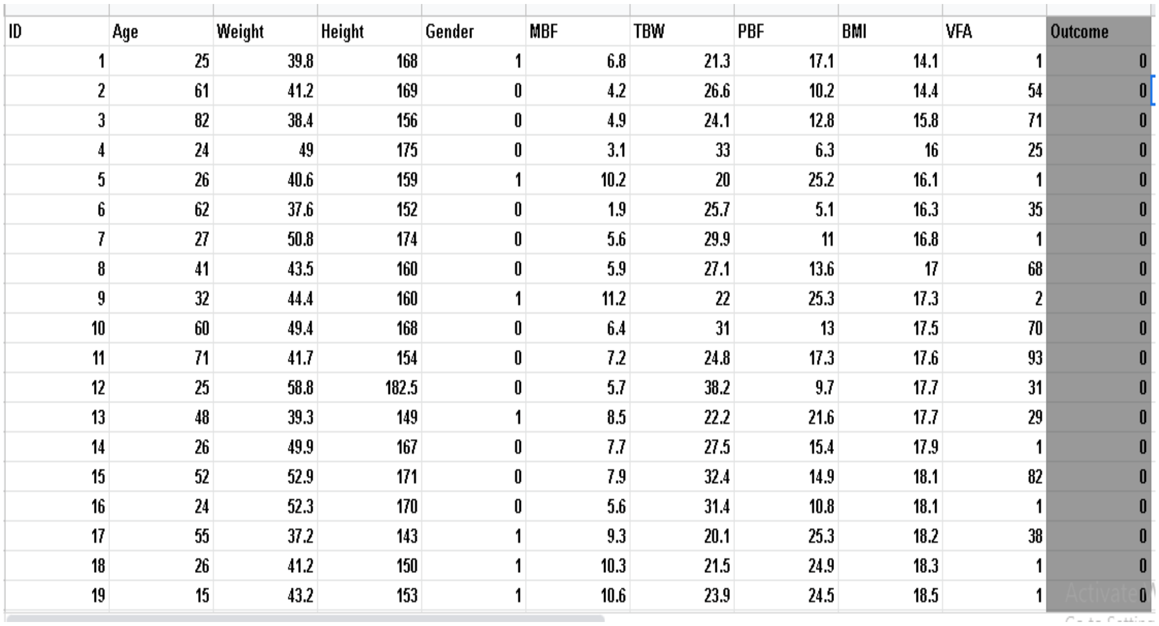
Category	BMI range (kg/m^2)
Underweight	<18.50
Normal range	18.50-22.90
Overweight	>=23.00
At risk	23.00-24.90
Obese I	25.00-29.90
Obese II	>=30.00

Table 3: BMI cutoff point indicating high riskoof NCDs in Asians, WHO



BMI range (kg/m^2)	Category of risk
18.50- 23.00	Acceptable normal range
23.00- 27.50	Moderate increase in risk NCDs
>27.50	High increase of NCDs

APPENDIX B: Dataset Sample File Format



ID	Age	Weight	Height	Gender	MBF	TBW	PBF	BMI	VFA	Outcome
1	25	39.8	168	1	6.8	21.3	17.1	14.1	1	0
2	61	41.2	169	0	4.2	26.6	10.2	14.4	54	0
3	82	38.4	156	0	4.9	24.1	12.8	15.8	71	0
4	24	49	175	0	3.1	33	6.3	16	25	0
5	26	40.6	159	1	10.2	20	25.2	16.1	1	0
6	62	37.6	152	0	1.9	25.7	5.1	16.3	35	0
7	27	50.8	174	0	5.6	29.9	11	16.8	1	0
8	41	43.5	160	0	5.9	27.1	13.6	17	68	0
9	32	44.4	160	1	11.2	22	25.3	17.3	2	0
10	60	49.4	168	0	6.4	31	13	17.5	70	0
11	71	41.7	154	0	7.2	24.8	17.3	17.6	93	0
12	25	58.8	182.5	0	5.7	38.2	9.7	17.7	31	0
13	48	39.3	149	1	8.5	22.2	21.6	17.7	29	0
14	26	49.9	167	0	7.7	27.5	15.4	17.9	1	0
15	52	52.9	171	0	7.9	32.4	14.9	18.1	82	0
16	24	52.3	170	0	5.6	31.4	10.8	18.1	1	0
17	55	37.2	143	1	9.3	20.1	25.3	18.2	38	0
18	26	41.2	150	1	10.3	21.5	24.9	18.3	1	0
19	15	43.2	153	1	10.6	23.9	24.5	18.5	1	0