BODY FAT PREDICTION USING MACHINE LEARNING MODELS

A PROJECT REPORT SUBMITTED BY

S.N.M. NAJWAN
S/16/821

to the

DEPARTMENT OF MATHEMATICS

in partial fulfillment of the requirement
for the award of the degree of

BSc. (Honors) in Statistics and Operations Research

of the

UNIVERSITY OF PERADENIYA
SRI LANKA
2022

DECLARATION

I do hereby declare that the work reported in this project report was exclusively carried out by me under the supervision of Dr. Jagath Senarathne. It describes the results of my independent work except where due reference has been made in the text. No part of this project report has been submitted earlier or concurrently for the same or any other degree.

Date: ……. …….

Signature of the Candidate

Certified by:

1. Supervisor: Dr. Jagath Senarathne

Date: ……. Signature: …….

2. Head of the Department: Dr. Rekha De Silva

Date: ……. Signature: …….

S. N. M. NAJWAN (S/16/821)

Department of Mathematics, University of Peradeniya,

Peradeniya, Sri Lanka

Body Fat is called adipose tissue. It is all over the body and it is contributing so well to the functions of the body. Body fat will be burned to provide energy to perform the body's basic operations. There are three types of body fats: Essential, Subcutaneous, and Visceral in the body by the way they store Body Fat. essential fat is important for maintaining normal physiological functions, Visceral fat can be found around the organs, and the subcutaneous fat lies underneath the skin. And it is all over the body. To keep track of our Body Fat Percentage, we calculate subcutaneous fat. Bioelectrical Impedance Analysis, Dual-energy X-ray absorptiometry, Hydrostatic Underwater weighing, and the Skinfold Caliper method are widely used to calculate the body fat percentage. Other than the Skinfold Caliper method these methods are too expensive and time-consuming, only hydrostatic underwater weighing gives accurate results. So, prediction is used to solve the problem. The dataset of 252 subjects with 15 observations is used to create the model. The variables are density, body fat, age, height, weight, and circumferences of the neck, hip, abdomen, chest, knee, ankle, forearm, wrist, thigh, and biceps. To make the prediction. The main objective of this research is to identify a better model that predicts the Body Fat data with higher accuracy. As the first step of the analysis dataset was sent through the preliminary analysis. This analysis resulted in giving us the data without outliers, misinterpreted values, or null values. And the density variable is removed from the dataset because it has a high negative correlation of almost one with body fat and it gives redundancy information. So, the rest of the data with 242 subjects and 14 observations were split into training and testing sets for further analysis for model creation. Multiple linear regression, partial least squares regression, and Artificial neural networks are used in this research. In the multiple linear regression analysis variance inflation factor and backward stepwise elimination, methods are used to identify the significant variables that contribute to the prediction of body fat. Only height, abdomen circumference, and wrist circumference are identified. As a next step partial least squares, and neural networks are used to fit the models using these three variables. Finally, accuracy scores were calculated for every model for the testing set and the training set. We can conclude that multiple linear regression performed better than the other two methods for predicting body fat percentage.

ACKNOWLEDGMENTS

The success of this project required a lot of guidance and assistance from many people, and I won't be able to finish this research work without their guidance and the correct direction towards the research. Therefore, I would like to express my sincere gratitude to them for the continuous support of my research, and their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

First and foremost, I offer my sincere gratitude and deep appreciation to my supervisor, Dr. Jagath Senarthne, a senior lecturer in the Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, for his full-time guidance and constant supervision, as well as unfailing encouragement and enthusiasm for the success of the project and providing necessary information regarding the project.

I would like to offer my thanks to Dr. Rekha De Silva, the Co-coordinator of the program (B.Sc. in Statistics and Operations Research), the Department of Mathematics, Faculty of Science, University of Peradeniya, and all the Lecturers of the program of the bachelor's degree in Statistics and Operations Research, Faculty of Science, University of Peradeniya.

I would like to thank my fellow batch mates who helped me from a single word during my research period for the stimulating discussions, the sleepless nights we were working together before deadlines, and for all the memories, we have had in the last four years. who helped me in various ways to succeed in the project and throughout my studies. Finally, I wish to express my love for my parents for their inspiration and the sacrifices they have made throughout my university career.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBRIVIARIONS

ABBREVIATION	DEFINITION
ANN	ARTIFICIAL NEURAL NETWORK
BD	BODY DENSITY
BFP	BODY FAT PERCENTAGE
BMI	BODY MASS INDEX
BIA	BIOELECTRICAL IMPEDANCE ANALYSIS
CT	COMPUTERIZED TOMOGRAPHY
DXA	DUAL ENERGY X-RAY ABSORPTIOMETRY
MRI	MAGNETIC RESONANCE IMAGING
ML	MACHINE LEARNING
MLR	MULTIPLE LINEAR REGRESSION
OLS	ORDINARY LEAST SQUARES
PCA	PRINCIPAL COMPONENT ANALYSIS
PLSR	PARTIAL LEAST SQUARES REGRESSION
RMSE	ROOT MEAN SQUARE ERROR
RES	BIOELECTRICAL RESISTANCE
SAT	SUBCUTANEOUS ADIPOSE TISSUE
VIF	VARIANCE INFLATION FACTOR
VAT	VISCERAL ADIPOSE TISSUE

CHAPTER 01

INTRODUCTION

The presence of body fat (BF) is an essential asset of life. The BF in the body is usually called adipose tissue. There are commonly three adipose tissues in the human body. One is white adipose tissue (white fat). White fat is located in the upper legs, chest, and abdomen. functions of white fat are commonly known as BF functions like storing the fats from derived food providing insulation against cold and mainly converting the very small amount of fat into energy. This process is functioning continuously if not we would have to eat continuously to stay alive. The second adipose tissue is brown adipose tissue (brown fat). Brown fat is located in the neck, chest and abdomen, shoulders. Brown fat can be found in the body very less than white fat. Brown fat's functions are keeping the body warm by burning the fat it stores. Third adipose tissue is a beige fat cell, it is derived from white fat, and this process is called browning. This process is triggered by exposure to low temperatures. The functions of beige fat cells are burning energy to produce heat when the temperature inside the body dips.

BF can be divided into three types by the way they store fat; those are essential, subcutaneous, and visceral. Essential fat is necessary to maintain healthy and reproductive functions and can be found in bone marrow, muscles, nerve cells, heart, and liver. subcutaneous fat makes the most of the body fat and can be found under the skin. The visceral fat could be found as wrappers around organs, inside the belly, intestines, and liver its major function is protecting internal organs. Other widely known functions of BF are it helps in giving our body energy, helps the body absorb vital nutrients, supports cell growth, protects the inner body organs, and keeps the cholesterol and blood pressure under control. Due to the demands of childbearing and other hormonal functions, the percentage of essential body fat for women is greater than that for men. Nowadays, Body fat percentage (BFP) has been used in the sports sector too. Athletic performance might be affected by body fat. The ideal body fat percentage for athletic performance is 12–18% for women and 6–15% for men; this was indicated by a study by the University of Arizona. Bodybuilders may compete in the essential body fat range certified

personal trainers will suggest they keep that extremely low level of BF. 3–6% is generally considered a minimum BFP for human males.

We got to know how well BF contributes to the body. Even though BF is a necessary substance for the body when it exceeds the limit, it would bring serious health problems. When excess BF is stored in the body it will bring infectious diseases like heart disease, metabolic syndrome, stroke, sleep apnea, fatty liver diseases, high blood pressure, etc. if we see these diseases then excessive BF would be the primary suspect of course. Metabolic syndrome alone can cause heart disease, stroke, and diabetes under these conditions when your body contains high blood pressure, too much fat around the waist, and high blood glucose level. Fatty liver disease is caused when too much fat builds in the liver. Cardiovascular disease is the reason for almost three fourth of all worldwide deaths and this cardiovascular disease is majorly contributed by obesity which is an effect of excess body fat. It is also important to state what it will bring when it is lower than the limit it should contain. Essential fat level is when physical and physiological health is negatively affected, like the weaker immune system, colon cancer, hormonal imbalance, and vitamin deficiency diseases. and below the limit in which death is certain.

Then you should know the importance of predicting your body's fat and keeping it at a proper level. Subcutaneous BF consists of fat accumulation several methods are available for determining BFP, such as measurement using bioelectrical impedance analysis (BIA), dual-energy x-ray absorptiometry (DXA), and underwater weighing There are many methods even though explored to measure the BF but using anthropometric measurements is still the best technique so far. In anthropometric methods, measurements of body composition such as body length, width, circumference, and skinfold thickness. So, it is quite easy to take the measurements and less costly than other methods. For example, the circumference can easily be measured with a caliper, known as the skinfold caliper method. Many researchers have worked on anthropometric datasets using different machine learning (ML) models to predict the BF. To find out the best models which perform at a higher accuracy level.

1.1 Objectives

The main objective of this study is to develop a model that will predict the BFP using only simple anthropometric measurements. And the second objective is to identify the better performing model from the three machine learning models such as multiple linear regression (MLR), partial least squares regression (PLSR), and artificial neural network (ANN). To find the model that predicts the BFP for the variables with better accuracy. To develop a clinical tool that comes with the most accurate results and improves the existing prediction methods. R statistical software has been used for the whole analysis.

1.2 Problem Statement

Ages before all people were healthy because they had to work hard to live and get their needs done. They used their physical strength mostly. But nowadays wherever you go you can see the power of technology. How easily our works are being done. We differ from ancient people for mainly one reason: technology. We mostly use our mental strengths for work. That is where the problem arises because we are becoming unhealthy because of this laziness. Our body is just storing fat more and more because we are using our energy rarely. And if we don't use the energy then BF will not be used. Then it will lead to an unhealthy life. So very few people understood the situation and used their time wisely by doing daily exercise, sports, and building the body to stay fit. It is important to keep the BFP at a certain level to prevent diseases that will take place with low BFP. To do that we should measure the BF occasionally. To measure the BF there are numerous methods have been used including dual-energy X-ray absorptiometry (DXA), magnetic resonance imaging (MRI), and computerized tomography (CT) are available. But these are very advanced technologies and very expensive to use at once. Then we need a way or a method to use as a solution. Then the technology strikes again, this time with an advantage. That is, we can use prediction as a way out. Once we give body measurements to a model then it will give the exact BFP. So how to predict the BF effectively? this is the question we are going to go through in this research.

Effectively predicting means this will be effective in every way. It is not expensive and easier to do so. To perform prediction first we should have the dataset. And the dataset should contain variables that should be easier to measure. Then only we can talk about anthropometric measurements. Many BFP measuring techniques even though explored More than several decades ago anthropometry was the only technique available for quantifying body size and proportions, that is these measurements can be measured even with a caliper. Not all the measurements but many body circumferences can be measured and not any extra supports can measure all alone and other anthropometric variables like height, and weight are also easily measurable. We have a big challenge in front of us that is how the predictions are going to be. Then we can propose the use of ML here by training the dataset with the proper ML models Then to make predictions there are numerous ML methods available like linear regression, multiple linear regression, ridge regression, partial least squares regression, neural network, etc. Using those we can easily predict the BFP. Since our data is based on the anthropometry method of body measurements. ML is a method of data analysis that automates analytical model building. It is a subset of artificial intelligence based on the type of model that systems can learn from data, identify patterns, and make decisions with minimal human intervention. It can automatically produce models that can analyze more complex data and deliver faster, more accurate results even on a very large scale. Our dependent variables are body measurements like body length, width, circumference, and skinfold thickness. So, the prediction of BF for the random dataset is based on the anthropometry method of data collection and efficient use of ML methods. Once we create the model whenever you need to measure the BFP you just must give measurements to the model, and you will get the predicted BFP in return.

1.3 Data Analysis

We accumulated this data from the Kaggle web page. This data set was collected from men population. R statistical software has been used for the whole analysis. In this study first Exploratory data analysis was used for all the 15 variables. and then Multiple Linear Regression will be used to find the model. Then furthermore for the variables that were selected in the MLR model Partial Least Squares Regression and Artificial Neural Network analysis will be used.

1.4 Research Outline

In our first Chapter, the introduction to the study, along with the background information about the BF is given. In addition, the importance of the study, the objectives of the study, and the outline of the upcoming chapters are discussed. In chapter 2, a discussion of the past studies of Multiple Linear Regression, Partial Least Squares Regression, and Artificial Neural networks of BFP prediction is considered. Chapter 3, under the Methodology, covers the topics of the methods used in this analysis such as Exploratory Data Analysis (EDA), how to do the MLR, PLSR, and ANN analysis. After that under chapter 4, results will be discussed. Finally, it will be discussed about the conclusion of the study in Chapter 5.

CHAPTER 02

LITERATURE REVIEW

In this chapter, the results of a literature review on Body Fat (BF) prediction are presented. The various body fat measuring research methods and their findings are all discussed in the literature about BF prediction. The literature review was primarily focused on the anthropometry variables, skinfold caliper method, and other similar body fat measuring techniques, similar research subjects, results as well as test methodologies. These acceptable findings aided in understanding the research concept and developing a proper study approach**.**

It is important to select the proper dataset when we conduct research from existing data because if we select our dataset from less standard measurements then it will lead to bad results. So, when we talk about measuring body fat (BF) then there should be a valid method and variables. Then we can talk about anthropometric variables. Anthropometry was the first technique available for measuring body proportions. Anthropometric variables have been used in equations for predicting BF. Those measurements are weight, height, body circumferences, Body Mass Index (BMI), and skinfold thickness. We can see so many researchers using anthropometric variables like In Merrill (2019) developed and validated body fat prediction models in American adults by creating the model for men and women separately. Specifically, they have used statistical models to predict body fat percentage (BFP) which rely on skinfold measures, anthropometric measures, or the wide ranges of age and BMI present in the American adult population, and finally constructed a statistical regression model that included age, BMI, anthropometric measures, and skinfold measures with significant effects. This is not the only research that uses anthropometric variables Similarly Gomez (2012). They have modeled an equation to estimate the BF that assesses the predictive capacity of a model that describes based on BMI, sex, and age for estimating body fat and to study its clinical usefulness. Methods and researchers that have compared the developed equation with many other anthropometric indices regarding its correlation with actual BFP in a large population of 6,510 data from both sexes

representing a wide range and adiposity. And some of the research results show that anthropometric variables can be used to predict the BF all alone in Hodgdon (1987) One of the major findings of this study is that the inclusion of anthropometric variables in equations involving whole body electrical resistance (RES) can improve the prediction of BFP over that offered by RES, stature (Height), and body weight alone.

As well to using body composition variables some other variables are also used to predict body fat (BF). They also correlated with our predictor variable. Those are widely used BMI and density. We saw the importance of BMI above **(**Javier et al., 2012; Merrill, 2019). We all know that Body Mass Index (BMI) is a measure of weight and height. Even though it is a widely used measure some researchers have shown BMI is not a valid measure Weisbran (2010) identified a nonlinear relationship between BMI and BFP, suggesting that BMI does not adequately define obesity based on body fat. In another analysis by Meevsan (2010), they established the effects of age, and gender interactions on BMI, and BF relationships over a wide range of BMI and age. They also wanted to examine controversies regarding linear or curvilinear BMI and BF interrelationships. Body composition was measured using validated bio-impedance equipment (Bodystat) in a large sample of 23,627 UK adults of middle-aged, of which 11,582 were males finally, they concluded that the relationship between BMI and body fat was not linear, and they were more curvilinear. The association between BMI and body fat is not strong, particularly in their desired BMI range, and is also affected by age. (2010 Elsevier Ltd and European Society for Clinical Nutrition and Metabolism), Meevsan (2019). Next, we are going to see about body density (BD). If we are measuring BD, then we must know about the Siri equation which was established in 1961 to convert body density (BD) into body fat (BF). It is usually measured by underwater weighing Withers (1987) also measured BD by underwater weighing and in this analysis, they took 207 male densities and circumferences of members of South Australian representative squads in 18 sports and were tested to provide relative BF. And from the results, their models outperformed some models which were established earlier. In the analysis Katch and McArdle (1973), (Pollock et al., 1976; Sloan, 1967) seem appropriate for a sample as their standard deviations for predicted BD are approximately equivalent to their measured BD, and their total errors are roughly equal to the standard errors of estimate for their original equations and the prediction error analyses are non-significant when to compare with results of Withers

(1987). Other than underwater weighing, Garcia **(2015)** used the air displacement plethysmography (ADP) method to calculate the body density and they predicted BF.

Above we have seen some methods to measure body density (BD). Now we will look at measuring body circumferences. When we talk about body circumference, the skinfold caliper method is the best and easiest to obtain and the most common indirect method to assess body fat (BF). This idea is based on a measure of subcutaneous adipose tissue which is the greatest storage of body fat. Then it will provide the total body fat accurately. Using the caliper we will measure Circumferences of body compositions like waist, hip, abdomen, wrist circumferences, and so on. To predict body fat, the variables must have a relationship with it. The lower body skinfolds are highly related to the percent of Body fat in fit and healthy young men and women Eston (2005). One of the most used and widely accepted skinfold equations for assessing body composition (percentage of body fat and fat-free mass) accurately by the formula of Durnin and Womersley (1974). This formula uses the logarithmic sum of four upper body sites (biceps, triceps, subscapular, and the iliac crest). We can see similar models which use the logarithmic sum of circumference ideology in the Navy method developed by Hodgdon uses logarithmic terms including the abdomen circumference for men, the waist and hip circumference for women, and height and neck circumference for both genders. The relative fat mask (RFM) method created by Woolcott avoids using skinfold measurements and only employs height and waist circumferences. These results show us the interrelationships of BF and the body composition circumferences. And, other than body circumferences such as age, gender, sex, height, and weight contribute to the prediction of BF. Body fat (BF) is dependent on age, gender, and race (Heyward & Stolarczyk et al., 1996; Hawes & Martin, 2001). If we look at Garcia (2015). In that analysis, they tend to develop the improved prediction of body fat by measuring common anthropometric measurements. They took 117 healthy German subjects 46 men and 71 women from the middle-aged population, and they used two different procedures: validation and cross-validation obtaining the common anthropometric measurements and body composition by Dual-energy X-ray absorptiometry (DXA) and concluded that combining skinfold thickness with circumference provides a more precise Prediction of BF. You may wonder what this DXA is. We already discussed skinfold calipers as indirect body fat (BF) measuring method. Dual energy-X-ray absorptiometry (DXA) is also one of the indirect BF measuring methods. DXA is accepted as one of the valid methods of body composition analysis

(Prior et al, 1997; Kohrt, 1998). We can see some research that uses DXA. In that matter we can talk about the topic of prediction of DXA determined whole body fat from skinfold. Combining skinfold thickness with circumference provides a more precise Prediction of body fat Garcia (2015) in this analysis they have used two different procedures: validation and cross-validation obtained from the common anthropometric measurements and body composition by DXA, predicting BFM for anthropometric measurements and developed regression models. Even though the validity of this method has remained subject to question, particularly about concerns over tissue thickness and hydration levels (Laskey et al, 1992; Jebb et al, 1995; Pietrobelli et al, 1996, 1998; Wang et al, 1998; van der Ploeg et al, 2003), which will vary between individuals and groups of subjects. We understood the variables and methods. Now we should move on to the most important part of the research which is the results of the above-mentioned researchers and their findings.

Once we got our dataset then we should try to solve the problem that we took, which means we should know in which way we are going to perform the analysis or predict the problem. If we look at Merrill (2019) they used multiple linear regression analysis. in this study, they tend to develop multiple regression models to predict BFP in working men and women using all these ages, BMI, several anthropometric and skinfold measurements and performed backward stepwise regression analysis, so they can develop a clinical tool that will provide the most accurate results in body fat (BF) prediction. Similarly, (Garcia, 2005) also used multiple linear regression analyses with backward elimination of independent variables to avoid collinearity they performed with the tolerance level of 0.3. in another study to predict body fat with the new equation for clinical usefulness, they performed a multiple regression analysis (Gomez, 2012). As we saw before (Eston, 2005) they have conducted forward stepwise multiple regression analysis to assess the individual and multiple relationships with body fat and DXA in which combination of factors best-predicted body fat in this group of young men and women, all skinfolds, gender, age, height, and mass were made available in this analysis and they made sure the assumptions of normality and homoscedasticity of the criterion and predictor variables were satisfied to perform multiple linear regression. The Siri equation converts body density into body fat, but this is time-consuming and needs expensive equipment so using multiple linear regression equations would be great (Wethers, 1987). However, these multiple regression equations have been demonstrated to be population-specific (Durnin and Womersley 1974;

Pollock et al. 1976; Smith and Mansfield 1984). In (Swartz, 2012) predicted the BF in older adults by time spent in sedentary behavior, and the analysis was performed using linear-regression models. Due to the high correlation between the measures, principal component analysis (PCA) was applied to convert their variables of physical activity and sedentary behavior into uncorrelated principal components. And they concluded that limiting sedentary time may benefit BF levels and body size of adults aged 50 and older.

When we discuss the results of the analyses then it is much better to include the results obtained from some other methods. Still, this is a prediction of BF then other than MLR, PLS regression and ANN are also being used to analyze the BF data from the past. Szymanska used MLR and PLS methods in two different analyses. The waist circumference and the total BF are sufficient to predict the Visceral Adipose Tissue (VAT), Subcutaneous Adipose Tissue (SAT) in women (Szymanska, 2012). Moreover, he concluded that in healthy overweight men when plasma metabolites are included then only SAT can be used to predict. Only a few different metabolites are associated with BF distribution parameters such as android gynoid ratio, VAT, and SAT (Szymanska, 2012). To find these associations he used PLSR and correlation analysis. He evaluated the associations for profiles of metabolic to develop diagnostics for obesity and some metabolic disorders using the data from 32, and 83 overweight men and women respectively. In an analysis of Caucasian children, aged 6 to 17 anthropometric measures got the coefficients of determination (R2) in the range of 0.869 to 0.936 for men and women in the range of 0.900 to 0.979 with higher accuracy (Flavel, 2012). This is a study of modeling the equations to predict total and regional BF in children. the variables of this study are some anthropometric measurements such as height, weight, BMI, waist, and hip girth, and 8 skinfold thickness. The dependent variables that they used are total BFP, and total body fat mass, trunk, and abdominal region of interest, measured using dual-energy X-ray absorptiometry (DXA) used to measure arms and legs. In this study PLS regression has been used from the variables to determine the best predictive equation for BFP.

ANN is also widely used to make predictions on data. In (Duran, 2018) Duran analyzed to predict BF of children who are aged from 8 to 19 years. In that analysis whole body measurements were measured using DXA scans and then performed ANN to predict excess BF of 1999 children and 856 of them are female. And the sensitivity of the variables BMI, and waist

circumference in the analysis is 0.751, and 0.523 for females. And for the male's sensitivity of the BMI, waist circumferences are 0.721, and 0.572 respectively. BMI is less sensitive to predicting BFP (Duran, 2018). (Aleksander, 2014) analyzed predicting BFP using ANN. They used the subjects of 1332 women and a total of 2755 subjects. The data was collected using the measurements of Bioelectrical Impedance Analysis (BIA). The dataset contains data aged 18 to 88 and a BMI range of 16.6 to 64.6 kg/m^2. and finally got an accuracy score of 80.43%.

Many methods have been used for predicting Body fat (BF) or conducting analysis on BF datasets. Multiple Linear Regression is one of the widely used methods in this case. It is helpful to identify the significant variable which seriously affects Body fat and how closely they are correlated with each other variables. But the results when it can be improved more and reduce the random error that will be a much better model. So Partial Least Squares (PLS) regression and Artificial Neural Network (ANN) can be used to check the model performance for the randomly chosen Body fat dataset in this study.

CHAPTER 03

METHODOLOGY

In this methodology chapter, we are going to discuss the subjects that have been used throughout this research, statistical methods, data analysis, the theory behind the method approach, and how we have used the methods to find the results of the analysis.

3.1 Data Collection

We obtained our Subject data from the Kaggle web page. This dataset already has been used by (K.W. Penrose, A.G. Nelson, and A.G. Fisher,1985) to predict the body composition equation for men by using simple measurement techniques. In that analysis, only the first 143 data were used to predict the equation, but it contains 252 data of men ranging in age from 22 to 81 years. And the measurements were the same standards as apparently those listed in (Behnke and Wilmore,1974). The measurements consisted of Density, Percent body fat, Age (years), Weight (lbs.), Height (inches), and 10 body circumferences those are Neck circumference (cm), Chest circumference (cm), Abdomen 2 circumference (cm), Hip circumference (cm), Thigh circumference (cm), Knee circumference (cm), Ankle circumference (cm), Biceps (extended) circumference (cm), Forearm circumference (cm), Wrist circumference (cm). These body circumferences are measured by the skinfold caliper method using a caliper, and body density is measured by the method that is most widely used and more accurate hydrostatic underwater weighing. This technique computes body volume as body weight difference in air and water displacement. And our independent variable BF was computed from the BD variable using Siri's equation (1956) and then we will get the BFP.

Before starting the statistical analysis, we first tried to understand our data by doing data preparation and simple preliminary analysis. To get the knowledge that the data

represents. Because it is important to fit the right data to get a better model. On behalf of the data preparation, we tried to find out if there are any null values. A null value means if we cannot find the value for a particular variable and find only the empty cell. After that, we checked the data set for if it has some misinterpreted values. Misinterpreted values mean if our dataset contains any practically impossible value, there are never any values like that recorded before in the world. In our preliminary analysis, we have calculated some summary statistics such as Mean, Median, Quantiles, Minimum, and Maximum. The mean is the average of the data points that represents the central value of the dataset. And Median and Quantiles are referring to the data that is in the positions of 50th, 25th, and 75th when the data is in ascending order. Finally Minimum and Maximum are representing the range of the data. These summary statistics give us an understanding of the data. Moreover, we tried to find the relationship between the dependent and every independent variable using the correlation. To use them in further analysis.

3.2 Preliminary Analysis

It is a primary analysis to get a better understanding of data. In this analysis, the following summary statistics will be checked and visualized to find the outliers. The summary statistics are Minimum, Maximum, Mean, Median, 1st Quantile, and 3rd Quantile. These statistics will be checked before and after data cleaning. After finding the outliers, null values, and misinterpreted values these values will be excluded from the primary dataset. Then to identify new summary statistics it will be conducted again.

3.2.1 Correlation

Correlation is an interrelationship between two variables. It can indicate any relationship between two variables whether it is linear or nonlinear. We usually use correlation to identify how one variable is dependent on another. Because when the correlation between two independent variables is higher than it is technically impossible to change one variable without changing the other. The correlation coefficient is being used to measure correlation. By dividing

the products of covariance by the respective standard deviations gives us the measure of the correlation coefficient.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$

Figure 3.1. Equation of correlation

n = number of elements

$\sum x$ = sum of one of the independent variables in the dataset

$\sum y$ = sum of the dependent variable in the dataset

$\sum x\,y$ = sum of the product of the dependent and one of the independent variables

$\sum x^2$ = sum of the squares of the independent variable

$\sum y^2$ = sum of the squares of the dependent variable

3.3 Data Analysis

In this section, we are going to discuss some of the Machine Learning methods which are used to predict the BF. These methods are Multiple Linear Regression (MLR), Partial Least Squares Regression (PLSR), and Artificial Neural Network (ANN). Here, we will explain MLR analysis, assumption, and how it works step by step.

3.3.1 Multiple Linear Regression

MLR is a statistical method that is used to predict the outcome of a variable based on the value of two or more variables. Sometimes, we must know simple linear regression (LR), and this is an extension of LR. The variable that we need to predict is known as the dependent variable, while the variables which we use to predict the variable are known as independent variables. The main difference between LR and MLR is in simple linear regression (LR) the variable will be predicted by only one variable and in MLR more than one variable will be used to predict the

respective variable. MLR will establish a connection between the dependent and independent variable, where the dependent variable follows a straight line then this is a linear relationship with two or more independent variables. If not, then it is nonlinear. Even though the relationship divides into two linear and non-linear regressions it can track a particular dependent variable using two or more variables graphically. But nonlinear regression is difficult to use. By using this method regression problems are being solved. The general equation of the MLR is presented below.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i$$
$$\text{for } i = 1,2, \ldots n.$$

y = dependent variable

x = independent variable

β_0 = The Intercept

β_1 = The coefficient of the first independent variable

β_2 = The coefficient of the second independent variable

ε = estimate of error

3.3.1.1 Assumptions

1. Linearity condition: There should be a linear relationship between a dependent variable and independent variables this is called the linearity condition.
2. Independence condition: The residuals should not be correlated with each other.
3. Homoscedasticity condition: The variance of the residual must be constant at every level of X.
4. Normality: The corresponding data must follow a normal distribution.

These assumptions can be checked by using some conditions before conducting the main analysis. The normality condition can be checked by doing a preliminary analysis of the dataset. If we are clearer use Bar plots and skewness tests easily identify the normality of the dataset. The linearity condition can be checked by creating scatter plots if they don't seem linear then the linearity assumption has been broken. The homoscedasticity condition also can be checked by the scatter plot of predicted values vs residuals. Finally, our independence assumptions can be checked visually by creating histograms.

3.3.1.2 Variance Inflation Factor

The first step of the MLR is identifying the multicollinearity because as we mentioned before if collinearity is present among the independent variable, then it is violating the model then it should be solved. In this matter, we use the variance inflation factor (VIF). The VIF value can be computed using the ratio of the overall model variance by the variance of a model of a single independent variable. This ratio is calculated for each independent variable. Then the variable with a VIF value greater than 10 will be removed from the model. Then only we can get the better model accuracy

3.3.1.3 F-test

one of the most useful statistical tests in which under the use of F-distribution is the F test. We use this test to compare different statistical models which were fitted from the dataset. If we want to test the variability of the two independent samples drawn from the same population then the F test will arise. For example, suppose that a medical trial compares four treatments. Alternatively, we could carry out pairwise tests. The advantage of the F test is that we do not need to pre-specify which treatments are to be compared, and we do not need to adjust for making multiple comparisons. The disadvantage of the F-test is that if we reject the null hypothesis, we do not know which treatments can be said to be significantly different from the others, nor, if the F-test is performed at level α, we can state that the treatment pair with the greatest mean difference is significantly different at level α.

3.3.1.4 RMSE

To find the accuracy of the model we use Root Mean Square Error (RMSE). It is the standard deviation between the predicted values and the actual values. These residuals measure the data points are how far from the line of the regression. By using the scatter plots, we can identify how these residuals are spread out. RMSE measure will tell us how much the model is fitted better.

$$\text{RMSE} = \sqrt{(F - O)^2}$$

F = predicted value

O = actual value

3.3.2 Partial Least Squares Regression (PLSR)

PLSR is a similar statistical technique to principal component regression (PCR). This method is widely used when we have more variables than the observations and where we face high-level multicollinearity. This is used to find the relations between the matrix of independent and dependent variables. Simply, we can introduce PLSR as a method that performs part of the principal component analysis and reduces the number of variables, and then perform regression to predict. The thing that makes PLSR a better performer than PCR is that the new model cannot explain the dependent variable but in PLSR it is modified to explain both the independent and dependent variable.

3.3.2.1 Steps of PLSR for every component

1. As a first step we must find the Y (U_h) scores.

2. Next using Uh we should find the loadings of X (P_h).

3. Then we must use Ph to find the scores of X (t_h).

4. After that we use the (t h) to find the loading of Y (q_h).

5. Finally, use (q_h) to calculate (U_h).

6. Repeat until the convergence.

The equation of scores vectors is given below

$$U_h = (b_h) (t_h) (U = T B)$$

And the final X, Y related equation is seen as a below-mentioned equation

$$Y = T B (Q_t) + F$$

3.3.2.2 Determine the number of components

When we obtain the optimal model there will be enough components to fit our dataset. And can be used to predict simply. To assess the utility when a new component added to the model there are 3 equations are available,

Sum of squares of variable X: $R^2 X = 1 - \frac{\sum(X\ model - Xobs)^2}{(\sum Xobs^2)}$

Sum of squares of variable Y: $R^2 X = 1 - \frac{\sum(Ymodel - Yobs)^2}{\sum(Yobs^2)}$

The ratio of the total variation in Y: $Q^2 Y = [1 - \prod(PRESS/SS)\ a]$

PRESS = Prediction Error Sum of Squares

1. At first remove an individual element (I, k)
2. Second, fit the model

3. The third step is predicting the element i, k that was withheld

   $(\text{observed}_{i,k} - \text{predicted}_{i,k})^2$

4. Finally, repeat until each element is withheld once

3.3.2.3 Decomposition of independent and dependent variables

Pls regression decomposes each X and Y as a product of a standard set of orthogonal factors and a collection of specific loadings. So, the X variables are decomposed as X = TPT with TTT = I with I the unit matrix. We compute the score matrix as like in Principal Component Analysis (PCA) T represents the score matrix, and P is the loading matrix. Like that, Y can be calculated as Yb = TBCT wherever B could be a square matrix with the regression weights as diagonal parts (see below for a lot of details on these weights). The columns of T square measure the latent vectors. once their range is up to the rank of X, they perform a precise decomposition of X. Note, however, that they estimate Y alone.

3.3.2.4 R-square analysis

R-square (R^2) also indicates how well the model fits like RMSE. R^2 is called the coefficient of determination in statistics. It usually measures in a regression model the ratio of dependent variables that is explained by the independent variables.

3.3.3 Neural Network (NN)

NN is a series of algorithms that mimics the functions of the human brain called a Neural Network (NN). It goes through the given data and recognizes its underlying relationship of it. If we go more precisely it is performing the same as the biological neuron and the connections that are usually called synapses in the brain. NN is a subset of Machine learning (ML). This neuron

in NN collects and classifies the information under some mathematical procedure. NN has a strong resemblance to regression analysis and curve fitting. So, it is being used widely in statistical analysis. According to the process and application, few types of NN are there. These are Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Deep Neural Network (DNN), and Artificial Neural Network (ANN).

3.3.3.1 Artificial Neural Network (ANN)

ANN is a type of NN. A NN contains layers of interconnected nodes In NN there are 3 types of layers: an input layer, an output layer, and a hidden layer. every node of NN is called a perceptron and it is more like MLR. The perceptron gives the signal made by an MLR towards an activation function that will be nonlinear. Hidden layers fine-tune the input weightings till the neural network's margin of error becomes minimum. It is hypothesized that hidden layers extrapolate salient options within the input file that have prophetic power concerning the outputs. This describes feature extraction that accomplishes a utility like applied math techniques like PCA. ANN first needs to learn the dataset before making predictions. There is a particular learning algorithm in ANN that is called the perceptron learning algorithm. A perceptron is a unit of NN that is used in computation and detecting features of input data. The input variable will be multiplied by the learned weight coefficient. The Perceptron algorithm is given below.

1. As an Initialization step. Set w (0) = 0. Then perform the following computations for the time step (one training sample) t = 1, 2, ... n

2. Next activation. For each input example t, activate the perceptron by applying

continuous-valued input vector x(t) and desired response d(t).

3. Computation of Actual Response/output. Compute the actual response of the

perceptron as $y(t) = f(w^T(t)x(t))$ where $f(\cdot)$ is the activation function.

4. Adaptation of Weight Vector. Update the weight vector of the perceptron to obtain $w(t + 1) = w(t) + \eta e(t)x(t)$ where $e(t) = d(t) - y(t)$

5. Continuation. Increment time step m by one and go back to step 2 until convergence.

There is a specific rule for Weight update rule:

$$w(t + 1) = w(t) + \eta[d(t) - y(t)]x(t).$$

Learning rate $0 < \eta \leq 1$.

The initial weights are set to small random values.

ANN is a weighted directed graph of nodes and neurons. In ANN information is received as a pattern and image which is in vector form. Those input notations are designated as of X(n). and then every input will be multiplied by its weights. The weights are the strength of the neuron's connection and represent the information that the NN uses to solve the problem. After that, each weighted input will be summed up into a computational unit. The sum can be any value from 0 to infinity, but when we have 0, a bias value will be added. The threshold value is used to limit through the activation function to get the output of desired value.

3.3.3.2 Backpropagation

Backpropagation is a delta learning rule based on gradient. Here, after we found the difference between actual and predicted this error is propagated backward from the output layer to the input layer via the hidden layer. This is mostly used when we have multiple layers for NN.

1. Initialize the network by setting the weights into small random values.

2. Then calculate the output **y p** and the error this is called Forward Propagate.

3. After finding Back Propagate Error by finding the amount of weight update for a given input.

4. Train Network by performing the weight after (example, epoch, batch).

5. Predict and calculate accuracy.

CHAPTER 04

RESULTS AND DISCUSSION

In this chapter of results and discussion, we have included detailed information about the results obtained by the analysis method that was described in chapter 3 and the interpretation of those results. We have organized this chapter in the following way: first preliminary data analysis and data preparation, then the data analysis.

4.1 Preliminary Analysis

4.1.1 Density

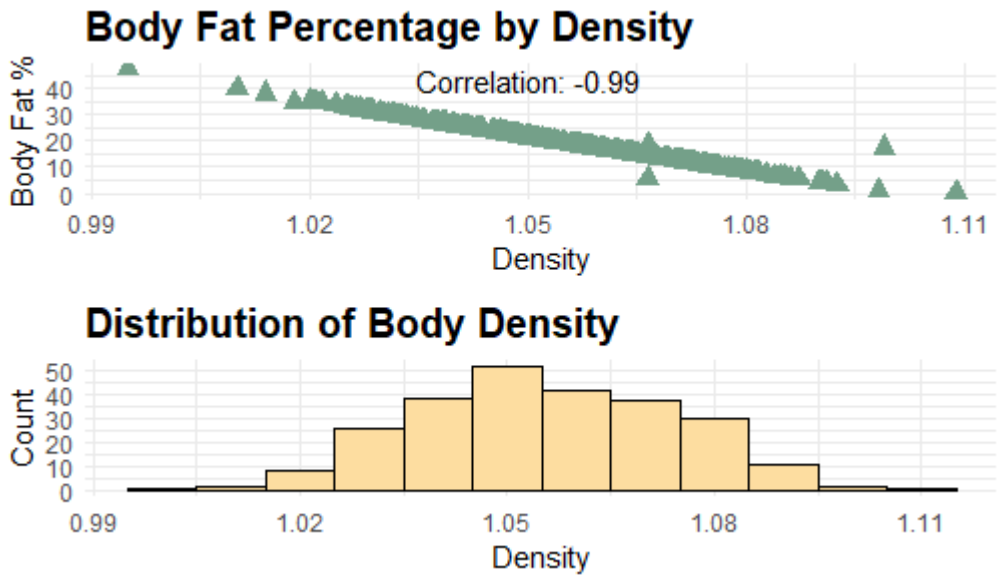

Figure 4.1. Distribution of Body Density

Table 4.1 Summary Statistics of Density variable from the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum

0.995	1.041	1.055	1.056	1.070	1.109

In the above **fig** of density variable from the raw data. There are 6 outliers in the density variable. The minimum data that was recorded is 0.995 Kgm^{-3} and the maximum data is recorded as 1.109. Since the mean density is 1.056 kgm^{-3} can conclude that most of the subjects' Body Density is around 1.056Kgm^{-3} another important thing to consider is the density variable correlation between BFP and BD recorded as 0.99Kgm^{-3}. This is a high negative correlation. Most of the data centered around the value 1.056Kgm^{-3}.

4.1.2 Body Fat

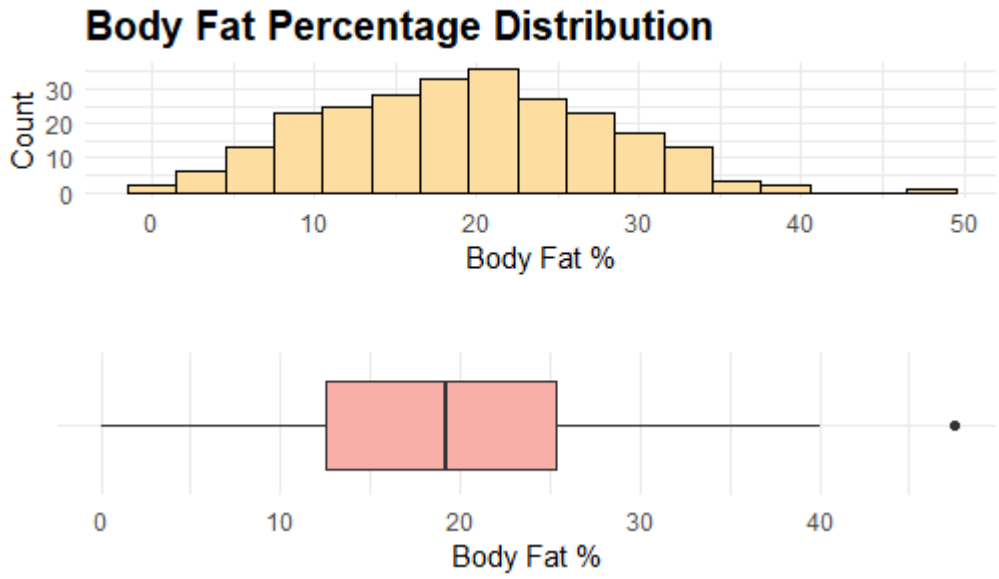

Figure 4.2. Body Fat Percentage Distribution

Table 4.2 Summary Statistics of Body Fat variable for original data

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
47.50	12.47	19.20	19.15	25.30	47.50

Using the above fig, we can understand the BFP. BFP follows normal distribution because the distribution nearly follows a bell shape. And there is an outlier data near 50. The minimum value for BFP was recorded as 0. And the data was distributed from 0 to 40. The maximum value for BFP is 47.5. Most of the data is around 20.

4.1.3 Age

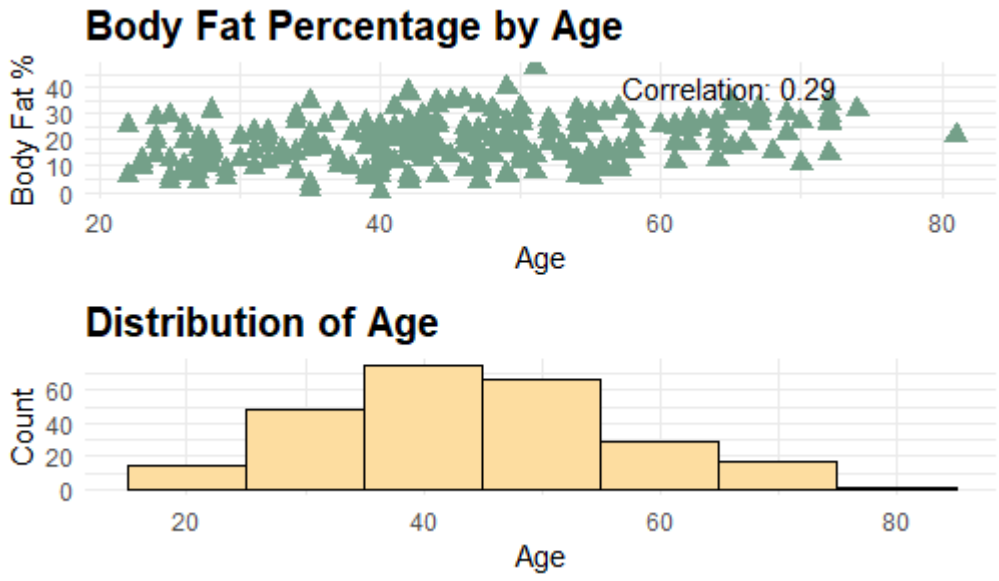

Figure 4.3. Distribution of Age

Table 4.3 Summary Statistics of Age Variable for the original dataset

Min	1st Quantile	Median	Mean	3rd Quantile	Maximum
22.00	35.75	43.00	44.88	54.00	81.00

Variable age shows a positive correlation of 0.29. since there is a positive correlation then there is a linear relationship between BF and Age. The age variable also reasonably follows the normal distribution. Age is measured in years. From the whole dataset, the subjects were taken in the

age range of 22 and 81 so the data is distributed in that range. There is only one outlier for variable age. Most of our subjects' age are around 44 and 45. Very few subjects were selected above age 60.

4.1.4 Weight

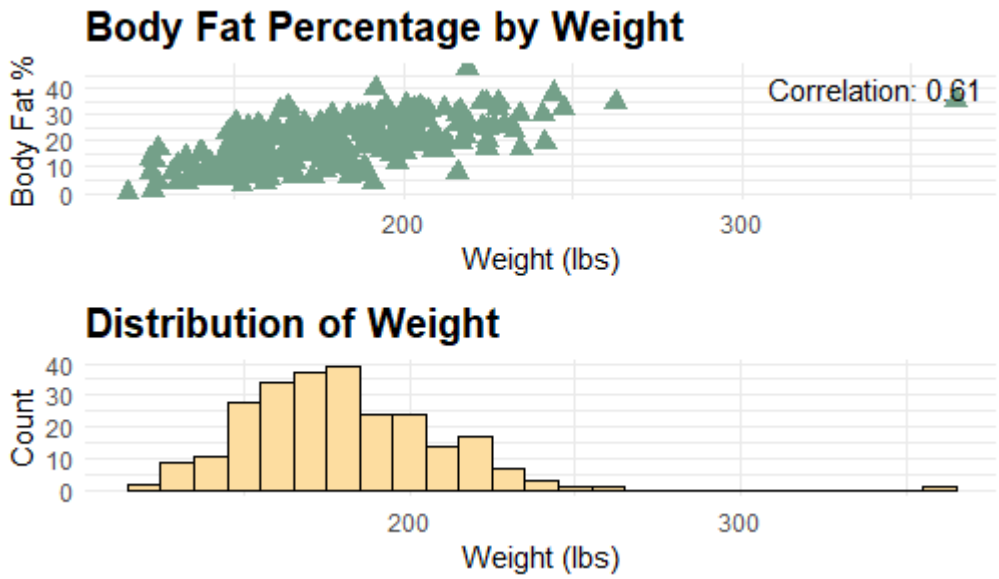

Figure 4.4. Distribution of Weight

Table 4.4 Summary Statistics of Weight for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
118.5	159.0	176.5	178.9	197.0	363.1

Variable weight also shows a high positive correlation with the target variable. The correlation between weight and BFP is 0.61. Since there is a high correlation then there can be a linear relationship with the BF. Weight is measured in **(lbs.).** The maximum weight that was recorded was 363.15 lbs. which is almost 164 kilograms, and the minimum weight is 118.5 lbs. which is nearly 54 kilograms. And most of the subjects are around 80 kgs (178.92 lbs.).

4.1.5 Height

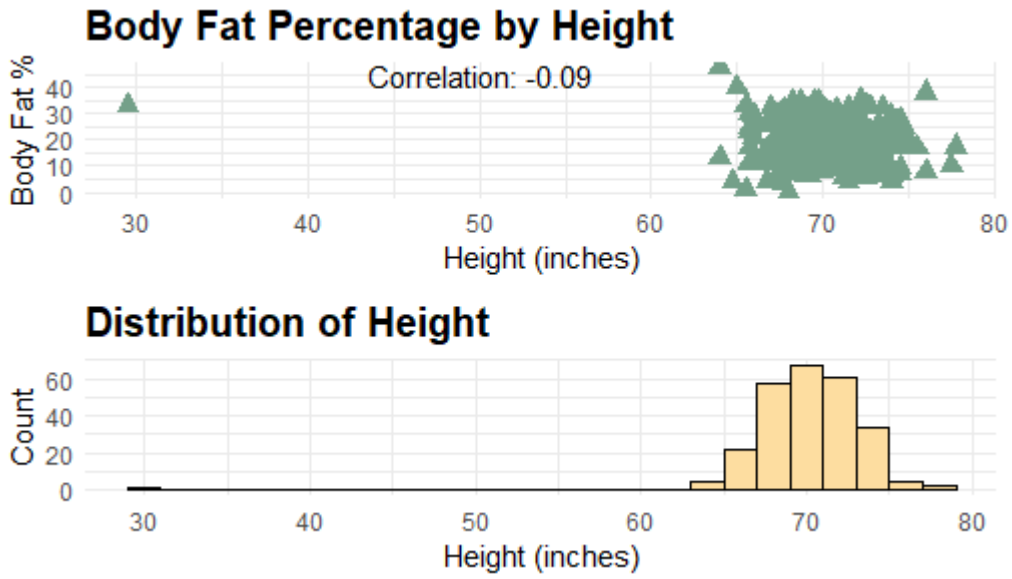

Figure 4.5. Distribution of Height

Table 4.5 Summary Statistics of Height for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
29.50	68.25	70.00	70.15	72.25	77.75

In these graphs of the height variable, we can see that there is an extreme outlier at 29.5. Because Height is measured in inches. That is also the minimum of the height variable. Since other data are around 70 inches, this data must be recorded incorrectly. And other than extreme outliers this data follows a normal distribution. The correlation between height and the BFP is very weak; it is negative 0.09.

4.1.6 Neck Circumference

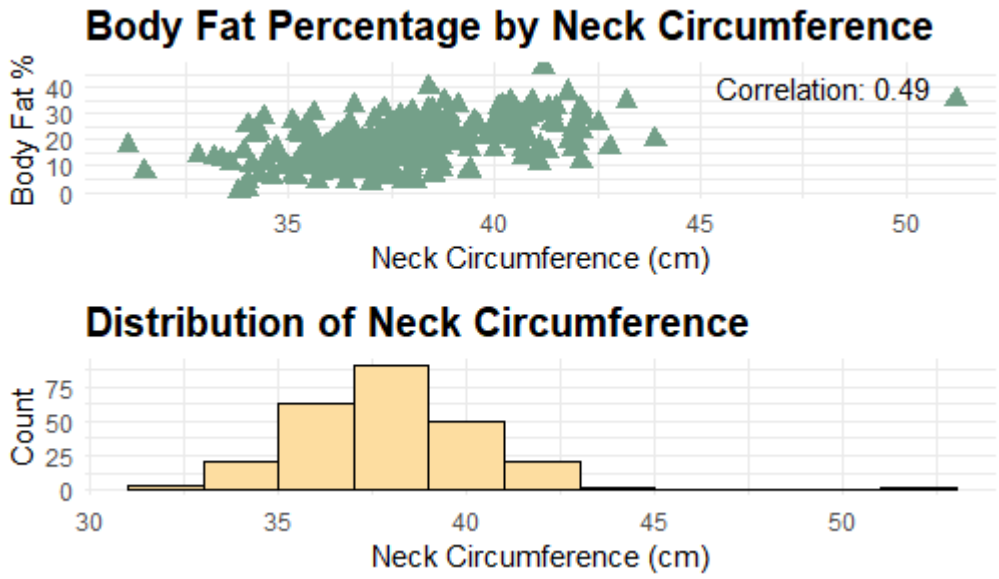

Figure 4.6. Distribution of Neck Circumference

Table 4.6 Summary Statistics of Neck Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
31.10	36.40	38.00	37.99	39.42	51.00

In the distribution of neck circumference, three outliers can be seen. Minimum of the neck circumference is 31.1cm and the maximum is 51.2cm. And the data is distributed from around 31 cm to 51cmThis maximum data belongs to the same subject which person had the largest weight. Neck circumference shows a positive 0.49 correlation. The neck variable also follows the normal distribution.

4.1.7 Chest Circumference

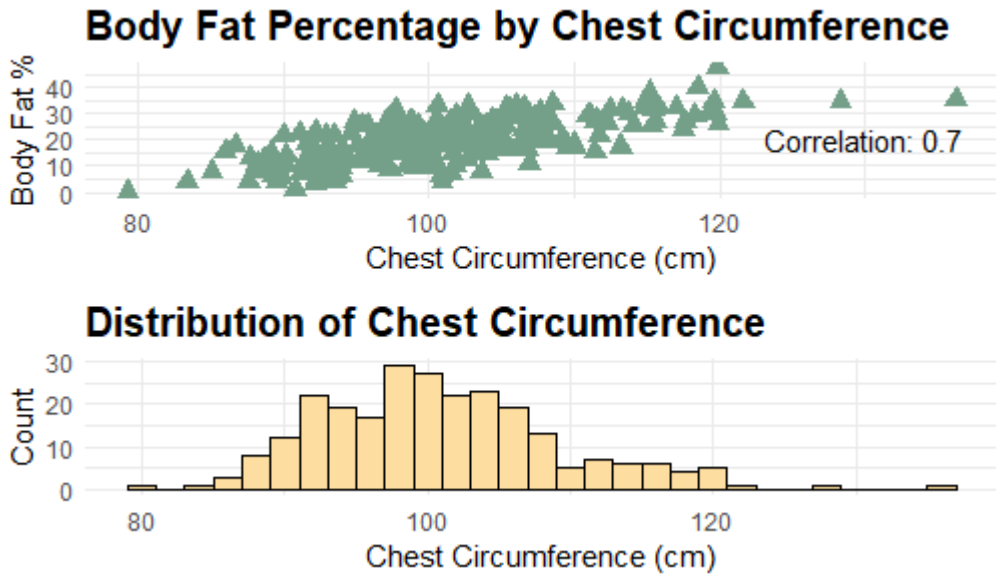

Figure 4.7. Distribution of Chest Circumference

Table 4.7 Summary Statistics of Chest Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
79.30	94.35	99.65	100.82	105.38	136.20

This variable shows a high positive correlation (0.7). chest circumference has three outliers. And it is nearly following a normal distribution. Many subjects' chest circumference is around 100. And maximum chest circumference in cm recorded is 136.2cm. The lowest size for chest circumference recorded is 79.3cm.

4.1.8 Abdomen Circumference

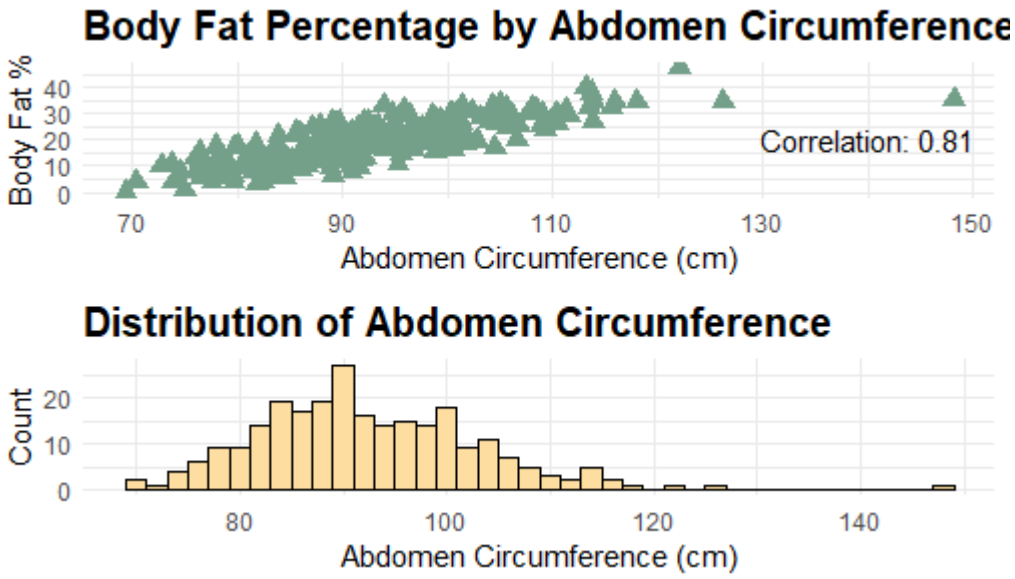

Figure 4.8. Distribution of Abdomen Circumference

Table 4.8 Summary Statistics of Abdomen Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
69.40	84.58	90.95	92.56	99.33	148.10

There is one extreme outlier in this variable. Abdomen circumference is also measured in cm. The minimum abdomen circumference is 69.4cm. The maximum of the abdomen variable is 148.1cm. Since the mean value is 92.6cm most of the subject's size of abdomen circumference is around 90cm. Moreover, the abdomen variable shows a high positive correlation with the BFP. It is 0.81. So, there can be a linear relationship between the Abdomen and the BF.

4.1.8 Hip Circumference

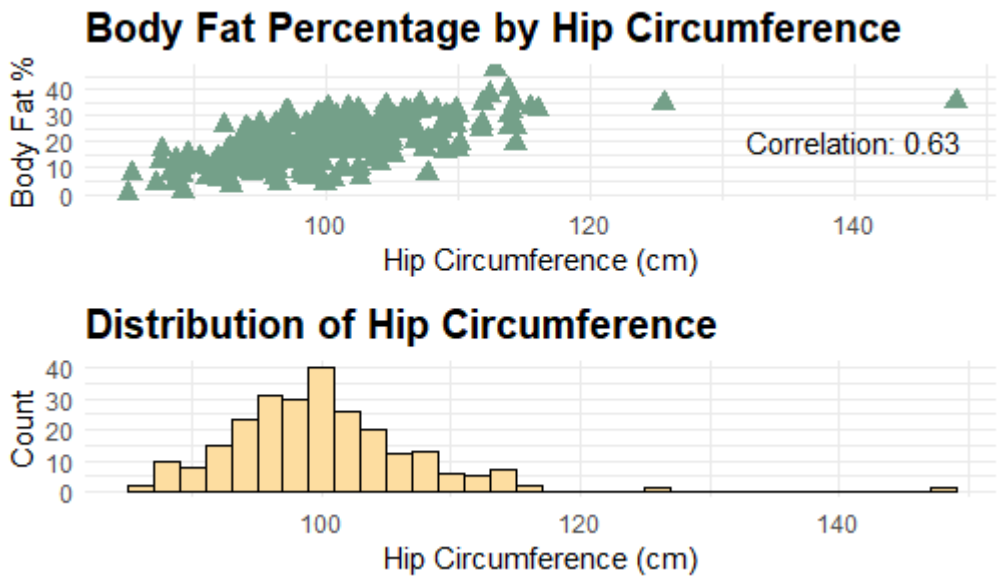

Figure 4.9. Distribution of Hip Circumference

Table 4.9 Summary Statistics of Hip Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
85.0	95.5	99.3	99.9	103.5	147.7

shows a positive correlation of 0.63. It has two outliers. Even though hip circumference is somewhat following a normal distribution. 147.7cms and 85 cms are the maximum and minimum hip circumferences respectively. And the mean value of hip circumference is 99.9 cm. most of the data centered around 100 cm for hip variables.

4.1.10 Thigh Circumference

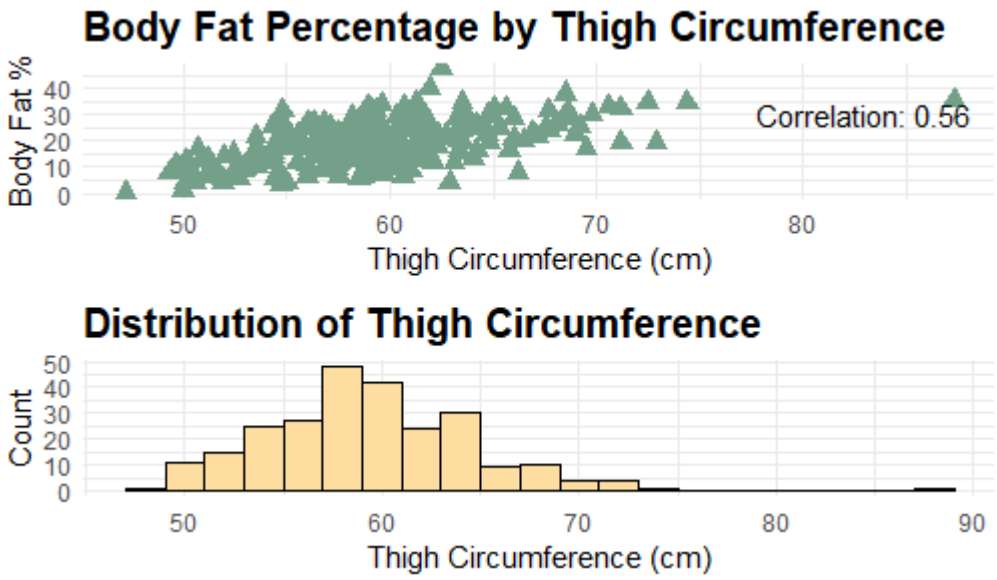

Figure 4.10. Distribution of Thigh Circumference

Table 4.10 Summary Statistics of Thigh Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
47.20	56.00	59.00	59.41	62.35	87.30

The correlation between thigh circumference and Body fat is 0.56. This variable also has a few outliers. The minimum thigh circumference is 47.2cms and the maximum thigh circumference is 87.3cms. and this variable also reasonably follows a normal distribution. And the mean of the Thigh variable is 59.41cms. because most of the data is centered around 59cms.

4.1.11 Knee Circumference

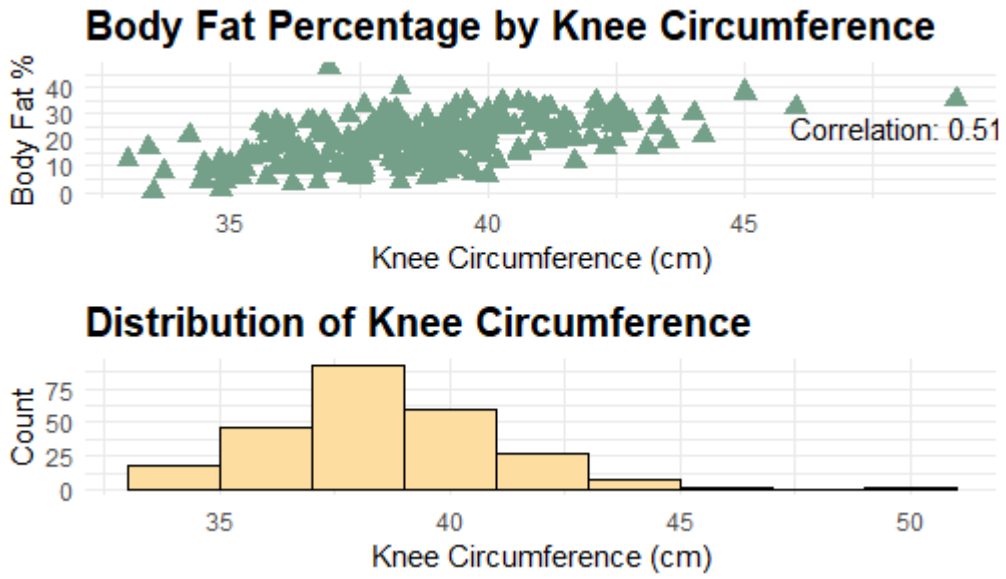

Figure 4.11. Distribution of Knee Circumference

Table 4.11 Summary Statistics of Knee Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
33.00	36.98	38.50	38.59	39.92	49.10

Knee circumference shows a 0.51 correlation. There can be a linear relationship between Knee and Body Fat. Since the mean value is 38.59cms most of the subject's knee circumference value is around 38cms. The minimum recorded value for knee circumference is 33cms and the maximum value that was recorded is 49.1cms. Here also the same subject is the extreme outlier.

4.1.12 Ankle Circumference

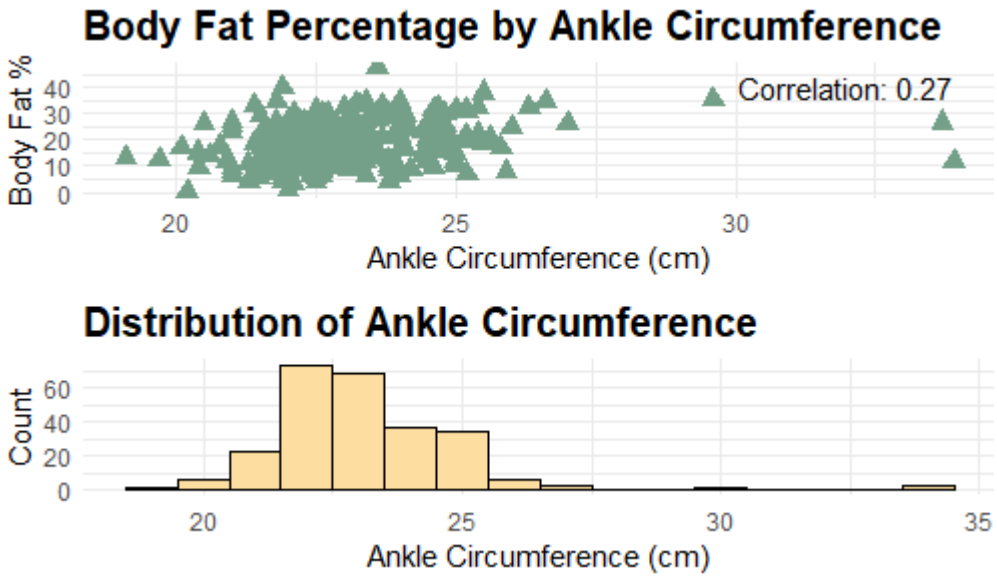

Figure 4.12. Distribution of Ankle Circumference

Table 4.12 Summary Statistics of Ankle Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
19.1	22.0	22.8	23.1	24.0	33.9

shows a positive correlation of 0.27. The ankle variable has a few outliers above 30cms. Even though hip circumference is somewhat following a normal distribution. 19.1cms and 33.9 cms are the maximum and minimum ankle circumferences respectively. And the central value of ankle circumference is 23.1 cms. most of the data collected were around 23cms.

4.1.13 Biceps Circumference

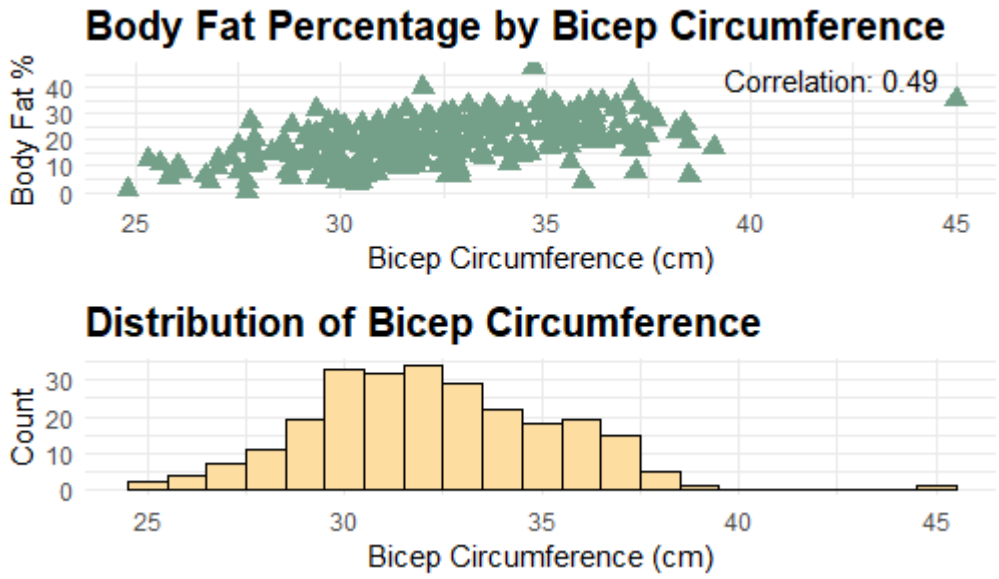

Figure 4.13. Distribution of Bicep Circumference

Table 4.13 Summary Statistics of Bicep Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
24.80	30.20	32.05	32.27	34.33	45.00

In the distribution of biceps circumference, there is only one outlier can be seen. The minimum of the bicep's circumference is 24.8cms and the maximum is 45cms. This maximum data belongs to the same subject which had the largest weight. Neck circumference shows a positive 0.49 correlation. biceps variable also follows a normal distribution.

4.1.14 Forearm Circumference

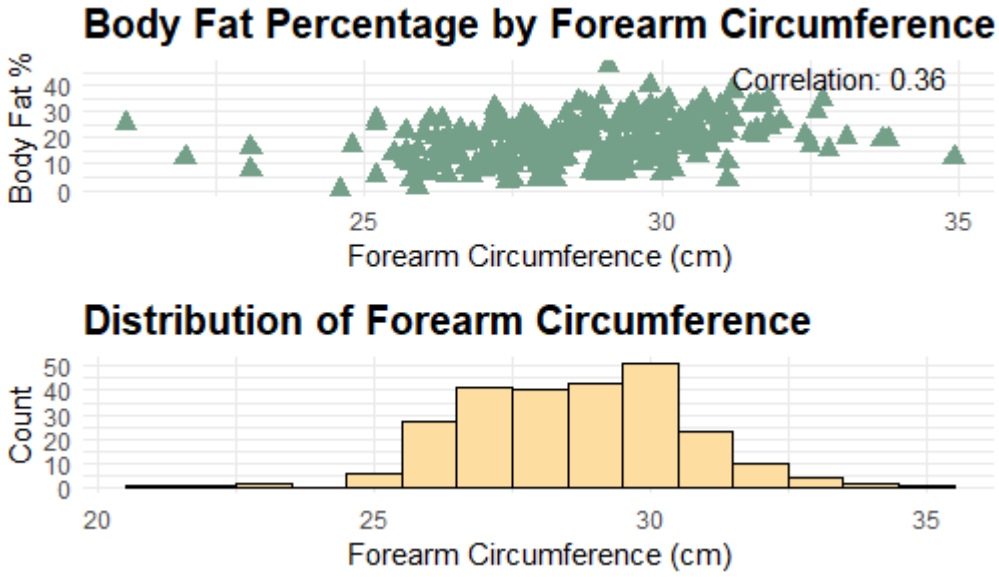

Figure 4.14. Distribution of Forearm Circumference

Table 4.14 Summary Statistics of Forearm Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
21.00	27.30	28.70	28.66	30.00	34.90

Forearm circumference shows a positive correlation of 0.36. forearm circumference has few outliers. And it is nearly following a normal distribution. Many subjects' sizes of forearm circumference are around 28cms. And maximum chest circumference in cm recorded is 34.9cms. The lowest size of the chest circumference recorded is 21cms.

4.1.14 Wrist Circumference

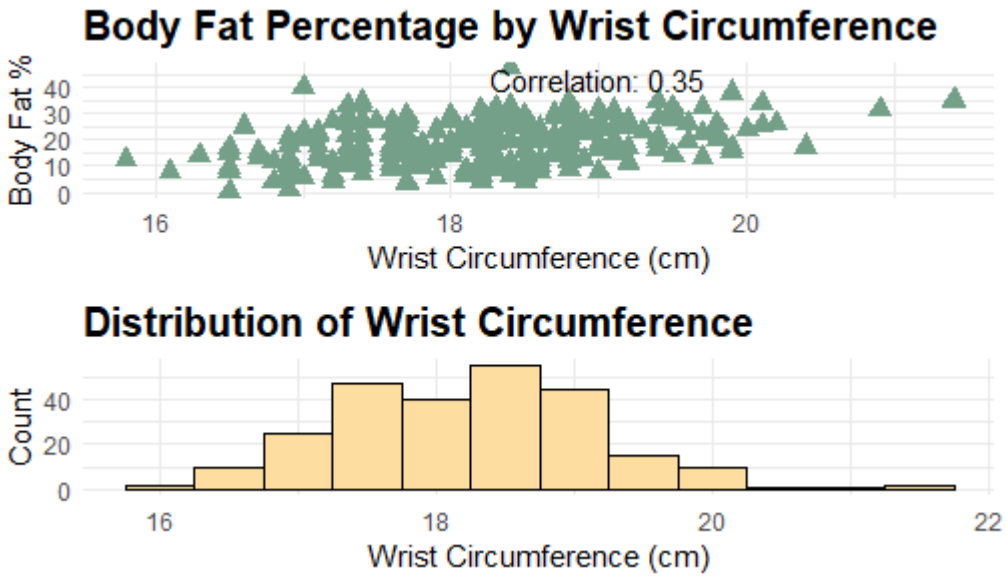

Figure 4.15. Distribution of Wrist Circumference

Table 4.15 Summary Statistics of Wrist Circumference for the original dataset

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
15.80	17.60	18.30	18.23	18.80	21.40

Our wrist variable shows a positive correlation with the target variable. The correlation between weight and BFP is 0.35. maximum wrist circumference was recorded at 21.4cms and the minimum wrist circumference is 15.8cms. And most of the subject's wrist circumference is around 18.23cms.

4.2 Data Preparation

This table refers to the summary statistics of the dataset after performing the Preliminary Data Analysis and removing extreme outliers, and misinterpreted values.

Table 4.16. summary statistics of the dataset after data preparation

Variables	**Minimum**	**1st Quantile**	**Median**	**3rd Quantile**	**Mean**	**Maximum**
Body Fat%	3.0	12.6	19.2	24.9	19.0	40.1
Age	22	35.25	43.00	54.00	44.88	81
Weight	125.0	159.2	176.1	196.6	178.1	247.2
Height	64.0	68.50	70.25	72.25	70.37	77.75
Hip	85.3	95.53	99.30	103.10	99.60	116.1
Thigh	49.3	56.10	59.00	62.10	59.29	74.4

Abdomen	70.4	84.75	90.95	99.05	92.20	118
Forearm	23.1	27.30	28.80	30.00	28.74	34.9
Wrist	16.1	17.60	18.30	18.80	18.22	20.9
Neck	31.1	36.40	38.00	39.40	37.95	43.9
Chest	83.4	94.45	99.60	105.28	100.59	121.6
Knee	33.4	37.10	38.50	39.90	38.58	46
Ankle	19.1	22.00	22.80	23.98	23.00	27
Biceps	25.6	30.23	32.00	34.25	32.25	39.1

On behalf of the data preparation, we used the results that we obtained from Exploratory data analysis. The analysis found out luckily there is not any null value. And checked out for if it has any outliers or misinterpreted values. To do that we used the summary statistics that we mentioned earlier in the previous chapter. Outlier and misinterpreted values of our dataset are from height and BFP. The height variable is measured in inches and the minimum height is 29.5 inches. It is not possible to have 165.61kg/m^2 as a Body Mass Index (BMI). And the minimum value for BFP we have found is 0. As we mentioned before, even if a subject is a bodybuilder, then the subject should at least contain BFP in a range of 3-4%. Since these misinterpreted

values are practically impossible to have the corresponding data excluded from the dataset. And filtered the BFP variable to at least 3%. We had some extreme outliers, and they are also removed from the dataset.

Table 4.17. Correlation values of variables

Density	Age	Weight	Height	Neck	Chest	Abdomen
-0.99	0.29	0.61	-0.09	0.49	0.7	0.81

Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
0.63	0.56	0.51	0.27	0.49	0.36	0.35

In the above table Density, Abdomen, Chest, Hip, and Weight variables are highly correlated more than 0.6. In those variables, only Density is negatively correlated. Another main point to state here is since our dependent variable BFP is measured from the Density variable then there can be a very strong relationship between BD and BFP. Siri's equation was used to compute BFP from the Density variable which was measured from the hydrostatic underwater weighing method. The correlation of density variables is almost negative 1. Then to improve performance measures we must remove one of the redundant variables. So, we excluded the density variable from the dataset. Even though the other three variables are highly correlated we don't have enough evidence to conclude that these variables will affect the results. So only the Density variable is dropped from the dataset. After performing data cleaning finally, the dataset contains data from 242 subjects and 14 variables.

4.2.1 Data Splitting

This is an important aspect of machine learning. Data is divided into two sets: Training, Testing. A training test is used to train the model with the data set and initialize the model creation. And the testing set is used to test or evaluate the model which was developed using the train set. Splitting is used for making sure the model is not overfitted because when we train the data and test it using the same dataset then the model can overfit the data. So, when we give a new dataset then it will perform purely. Then splitting can be used to test the model more accurately.

4.3 Data Analysis

4.3.1 Multiple Linear Regression

Dataset was randomly split into two subgroups: the training set, which contains 190 subjects, and the testing set, which contains the remaining 52, with each set with 14 variables. The purpose of splitting the whole data set into the testing and training sets is that the predictive models can be independently developed and validated on separate data sets so that the model performance on the testing set would be representative of the model performance on real-world data. Developing and validating the models on separate sets also ensure that any overfitting of the models does not occur.

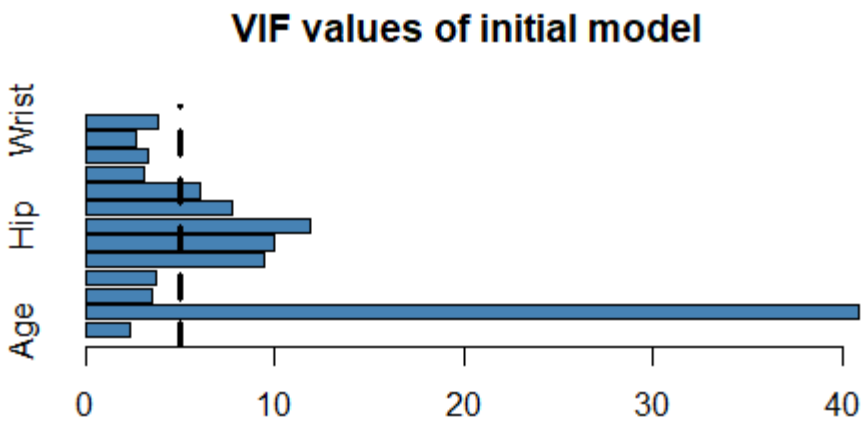

Figure 4.16. VIF values of the initial model

In this analysis, we are going to predict BF using anthropometric measurements. Since this is a predictive problem, we can use two types of regressions. If we have a single independent variable, then we would use a linear regression model, but we must find the model with multiple independent variables. We did MLR. Our dependent variable is BFP, and independent variables are age, height, weight, and 10 body circumference measurements as we mentioned above. Specifically, at first, we tried to fit the model backward stepwise regression analysis was performed on the training set. And using variance inflation factor (VIF) looked for multicollinear variables then it was removed if they exceeded the VIF limit of 10. Since from above **table 1** Weight, Abdomen, Hip, and Chest variables are highly correlated with other independent variables. We used the VIF step by step because one variable with a higher VIF value could be the reason for other variables with a higher VIF value. So, the Weight variable was excluded from the model due to the highest VIF values. After the removing Weight variable, the VIF values limit is satisfied under value 10 as presented in the below graph.

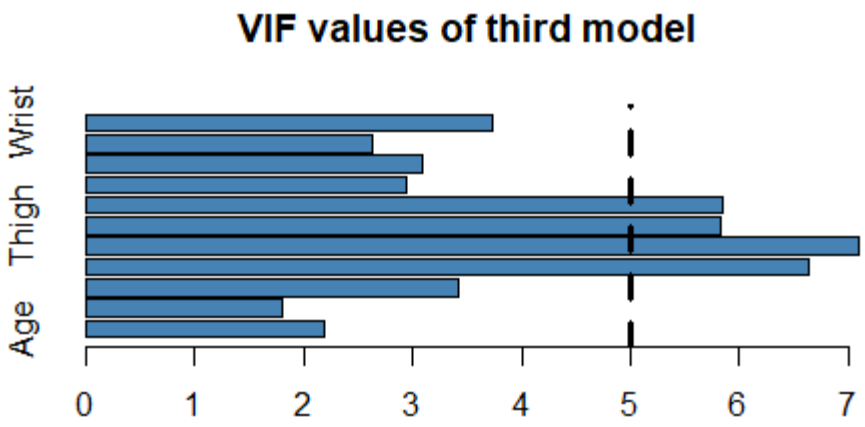

Figure 4.17. VIF values of 3rd model

In each step of the analysis, the predictor with the largest P value was removed, and the analysis was repeated. This process of removing the least significant predictor and repeating the analysis continued until the P values for all predictors were below 0.001. once met the model coefficients which satisfy the requirements.

Table 4.18. Intercept values of the final model

(Intercept)	Height	Abdomen	Wrist
0.306142	0.003539	< 2e-16	0.000136

Then only four variables satisfied the 0.001 significance level. Those variables are Height, Abdomen, and Wrist. After finalizing the training model, we should check the model validity by applying new data, then test data has been given to the existing model, so the data can be

predicted and already have the actual value of the test set. Then the final model looked like this with four variables. And obtained F statistic as 141.4 this is higher than previous models.

BF% = 9.23157 – 0.39179 × (Height) – 1.91943 × (Wrist) + 0.78344 ×(Abdomen)

The model of Multiple Linear Regression must satisfy the assumptions to fit the data more accurately. The assumptions of MLR: normality, linearity, independence, and homoscedasticity can be checked using the plots given below.

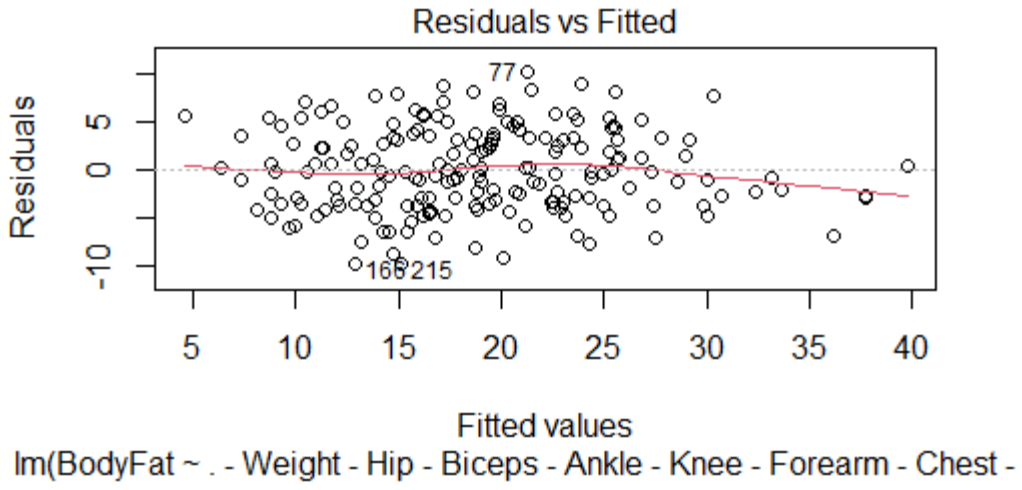

Figure 4.18. Residuals vs Plotted Values

This scatters plot shows if the residuals have nonlinear patterns or nonlinear relationships between predictors. This plot visualizes residuals on the y-axis and fitted values on the x-axis. And can detect outliers, non-linearity, and unequal error variances. In this plot, there is a kink at the end of the red line. And mostly this red line fits the central line. It is relatively evenly distributed on both sides.

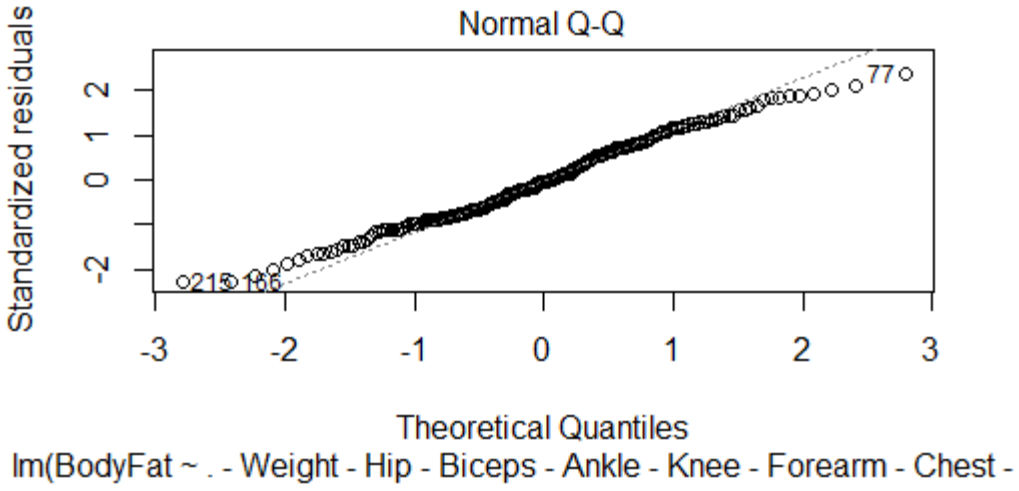

Figure 4.19. Standardized Residuals Plot

This plot shows whether residuals are normally distributed or not. The distribution of the residual will follow a straight line. These deviations from the start and the end of the line show there is some severe deviation. Since the residuals have followed near normal distribution.

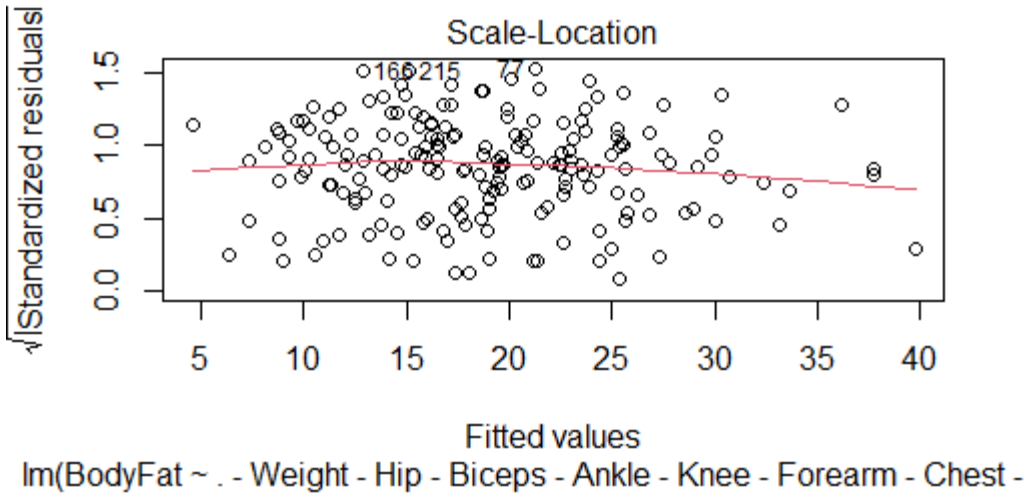

Figure 4.20. Scale Location Plot

Using this plot, we can check homoscedasticity or equal variance. This plot is also called a spread location. Standardized residuals are evenly spread across this graph. The red line is roughly horizontal across the plot, so the assumption is likely to be satisfied for the final regression model. The residuals spread are roughly equal at all fitted values. Can verify there is no clear pattern among residuals. The residuals should be scattered randomly around the line with roughly equal variability at all fitted values.

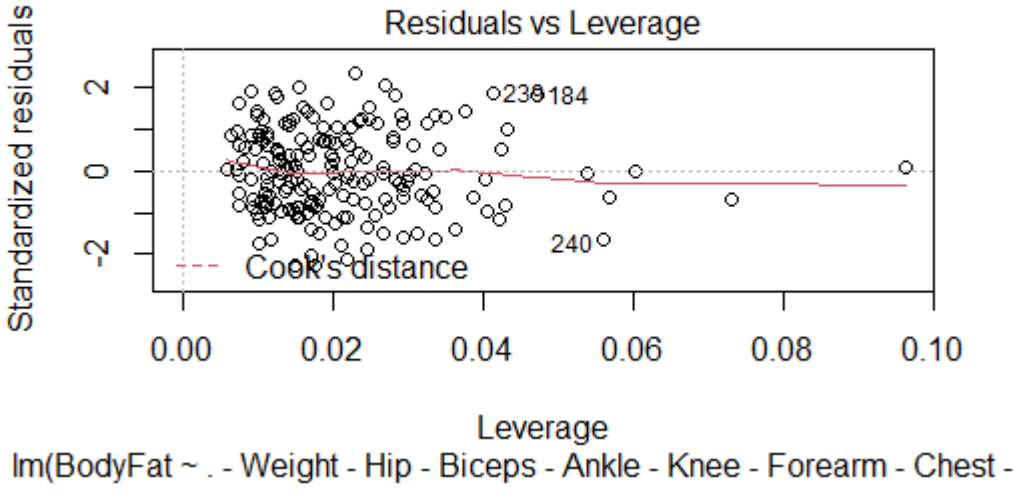

Figure 4.21. Residuals vs Leverage

This graph will allow us to identify any influential cases. There is no influential point in the plot because every point falls behind the cook's distance. Having Influential observations could indicate that the model does not provide a good fit for the data.

After checking the model could satisfy the MLR assumptions the accuracy score for the testing set is calculated to validate the model. Then root means squared error (RMSE) has been used to determine the difference between test sets' actual value and the predicted value. Then got the RMSE as 3.91111. That is a larger value for the better fit RMSE should be in the range of 0 to 1. We will get an RMSE value of more than 1 when the model overfits. And the accuracy score

that I obtained from MLR analysis for the training set is 89.28% and for the test set, it is increased to 95.67%.

4.3.2 Partial Least Squares Regression

In this analysis, we used PLSR to find a model. For this study, PLSR was preferred over other Ordinary Least Squares Regression (OLS). Because OLS will be easily biased even for one outlier if it is found in the dataset. Even though we removed several outliers from the dataset still some outliers left in the dataset. We didn't remove them because excluding more data will lead to a loss of information. So, we preferred to do the analysis using a method that is less sensitive to outliers. To fit the data for the PLSR model we used the variables that we obtained from the MLR final model. So, we don't need to consider every variable again in this analysis too.

For this analysis our dependent variable is BFP, and the independent variables are Abdomen, Height, and Wrist. Since PLSR is an extension of Principal Component Regression as a first step, PLSR performs Principal Component Analysis (PCA) to reduce the dimension.

Table 4.19. The variance explained by Partial Least Squares components

1^{st} component	2^{nd} component	3^{rd} component
52.45	85.11	100.00

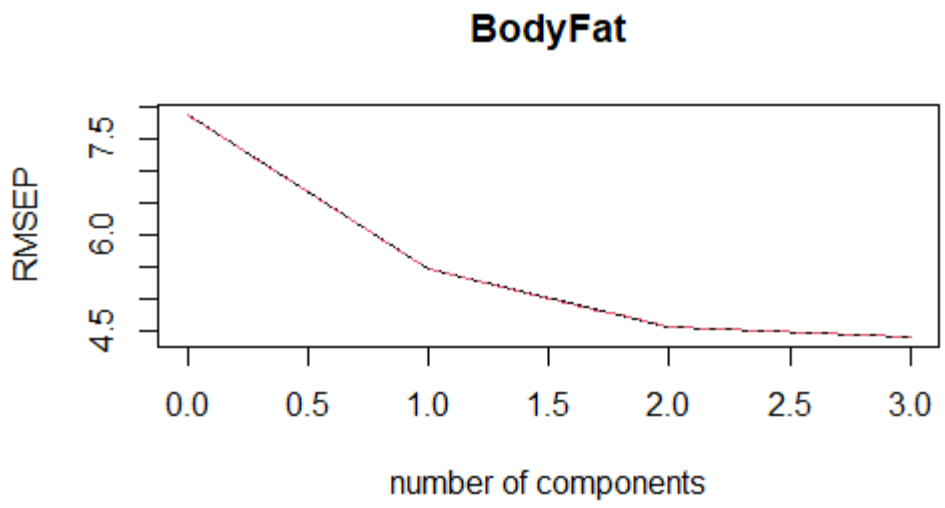

Figure 4.22. Root Mean Square Error of Prediction Plot

As you can see above, the table and the plot are evidence enough to say only 2 components are enough to explain more than 85% of BFP. After finalizing the number of components, we predicted the BFP for the test set using the PLSR model. After the comparison of actual and predicted values. We found the RMSE value of the model. The RMSE value of the PLSR model is equal to 3.8513. and the accuracy for the train set is 88.47%. And in the test set accuracy score is expected to improve to 94.96%. Even though the RMSE value of the PLSR model is less than the MLR model it is still greater than 1.

4.3.3 Artificial Neural Network

Artificial Neural Network is a well-known method. This structure was inspired by the biological nervous system. ANN contains 3 layers. The first one is the input layer from the left side and on the right side there is the output layer, and the middle layer is called the hidden layer. The output layer always contains one neuron. And the input layer contains the same number of neurons as

the number of input variables. In this analysis to we are going to use the same input variables. That we obtained from the MLR model. The variables are Height, Abdomen, Wrist, and Neck. So, these 4 variables will be on the left side of the input variables. And there is not any specified technique to choose the hidden layer and we chose just 2 neurons of one layer. And the output layer will give us the predicted BFP. The value or activity level of each node of the hidden and output layer depends on the sum of its inputs. Each input is included in the sum with its specific weight. The output of the node is the value of its activation function for its activity level.

The activation functions structure of an ANN will be set in the beginning, and the starting weights are usually initialized with random values. In our training set for the random weights, input and output variables will be adjusted to fit. We created only one hidden layer with 2 neurons to predict BFP using the Abdomen, Wrist, and Height nodes. ANN can find and derive the statistical relationship of the input and output parameters from the training data set without the specifications of a mathematical formula. The ANN used a training dataset to find the best mathematical relationship between input and output variables. Therefore, we used an ANN with our four input variables. All the input variables were normalized before training the model to make the model faster. A logistic function is used as the activation function and the threshold function is 0.01 as in the default settings. In the weight backtracking, the error backpropagation algorithm was used in the model.

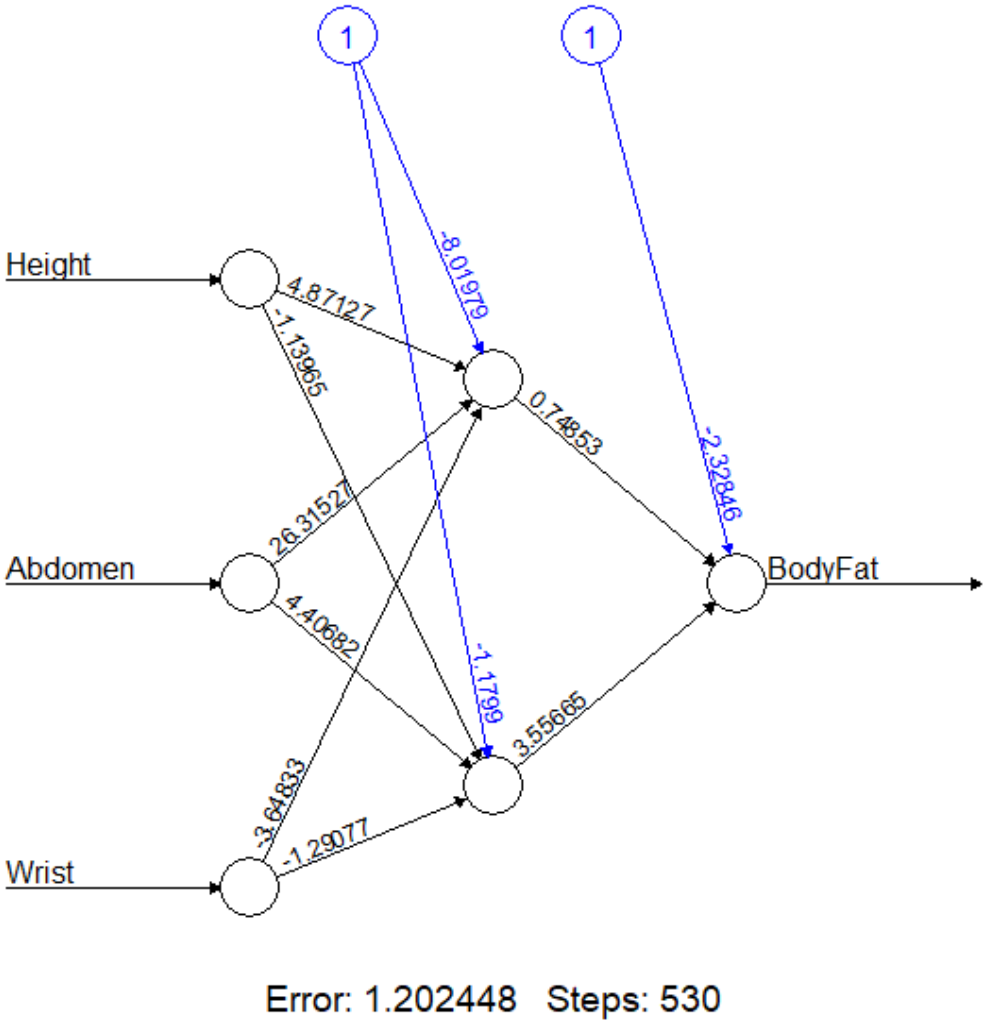

Figure 4.23. Artificial Neural Network Plot

As shown in the graph above 530 steps took to fit the model. The RMSE value of ANN is 4.205. if we use more hidden layers and neurons RMSE can be reduced. As we find the accuracy for the training set is 90.73%. And it is expected to turn out to be 94.98% accuracy in the test model. So, the model is not overfitted because it is predicted better in the test set than in the train set.

CHAPTER 05

CONCLUSIONS

Everyone should understand the importance of predicting BF. Because it was mostly used before for medical needs but now it is being used in the sports sector because the bodybuilding market is very high. And in the gym, we can see all the men and women trying to spend more time getting fit. For even medical purposes measuring BFP is too expensive. So, the people who are trying to get fit and stay healthy are struggling to keep a record of their BF level. To reduce these types of problems it is better to build a clinical tool that will predict the BFP more accurately with just easy measurements.

In this study, the main properties of the most used techniques to measure BFP, and methods to predict the BFP were discussed and have been presented. Initially, the dataset was sent through some preliminary analysis to identify the main aspects of the dataset. Using the summary statistics almost 10 outliers were cleaned from the total of 252 observations. Furthermore, MLR was used to identify the most significant variables that we used in the further analysis. Multiple Linear Regression is mostly used to identify the relationship between a dependent variable and two or more independent variables. In the MLR analysis, we didn't use the Density variable because it was already used in **Siri's equation** to compute BFP so it has a very serious relationship with our dependent variable so using it will reduce the chance of selecting other variables and measuring body density is not easier compared to take measurements of skinfold thickness of other body parts. And only a few assumptions are there to use MLR and all assumptions were satisfied during the analysis.

The next method used was PLSR, and this is used when there are more predictors than the observation. But PLS regression is better than the other Ordinary Least Squares Regression (OLS) when there are outliers in the data. For this PLSR analysis, only four variables are used because analyzing all the variables again is waste of time and non-necessary. In this section of

the analysis of four components, only three principal components are used to fit the model because those three components can explain 85% of the total variance.

And then analysis of Artificial Neural networks (ANN) can be used when we have to model against complex and nonlinear relationships. In this analysis, only 2 neurons are used as hidden layers because we only have 4 input variables so we should use fewer neurons than the input variables. First, the weights are randomly selected and fitted with the dataset. Then using backpropagate error weights were updated. Then the final model and results were computed.

Our training dataset contained 190 observations. For this set model accuracy was high almost 90%. The accuracy scores of MLR, PLSR and ANN are 89.28%, 88.47%, and 90.73% respectively. From these results, we can conclude that for the training set ANN performed better than MLR and PLSR. These are the same dataset where the models initialized so it is better to check the models in the new dataset. If the model overfitting the data, then we will get the accuracy score for the testing set below these accuracy scores that we obtained for the training set. Then we checked the model and validated it using the test set. Our test set contains 52 observations. In this set, the accuracy scores of MLR, PLSR, and ANN are 95.67%, 94.96%, and 94.98% respectively. These scores made sure that our models didn't overfit. And for the testing set Multiple Linear Regression model performed better than the other two models. ANN model still performed better than the PLSR model in the test set also. We can still use PLSR and ANN in prediction because their accuracy scores are in the testing set more than 94%. RMSE values for MLR, PLSR, and ANN are 3.911, 3.851, and 4.174 in the testing set respectively. When we talk about the RMSE values lower the value better the model. And RMSE measures positional accuracy. The PLSR model performed far better than the other two models. MLR is the only model that gives us better accuracy and RMSE value. So, we can conclude that MLR is the better model to predict body fat using simple measurements.

REFERENCES

https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset

Ann M. Swartz, Sergey Tarima, Nora E. Miller, Teresa L. Hart, Elizabeth K. Grimm, Aubrianne E. Rote, and Scott J. Strath. (2012). Prediction of Body Fat in Older Adults by Time Spent in Sedentary Behavior. Journal of Aging and Physical Activity, 2012, 20, 332-344

Merrill Z, Chambers A, Cham R. Development, and validation of body fat prediction models in American adults. Obes Sci Pract. 2020; 6:189–195. HTTPS:// doi.org/10.1002/osp4.392

GARCIA, ADA L., KAREN WAGNER, TORSTEN HOTHORN, CORINNA KOEBNICK, HANS-JOACHIM F. ZUNFT, AND ULRIKE TRIPPO. 2005. Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths. Obes Res.2005; 13:626 – 634

Fletcher and Alan N. Peiris Lynell C. Collins, Phillip D. Hoberty, Jerome F. Walker, Eugene C. T. (1995). The Effect of Body Fat Distribution on Pulmonary Function Tests. Chest 1995;107; 1298-1302. DOI 10.1378/chest.107.5.1298

JAVIER GÓMEZ-AMBROSI, PHD CAMILO SILVA, MD VICTORIA CATALÁN, PHD AMAIA RODRÍGUEZ, PHD JUAN CARLOS GALOFRÉ, MD, PHD JAVIER ESCALADA, MD, PHD. (2012). Clinical Usefulness of a New Equation for Estimating Body Fat. Diabetes Care 35:383–388

RG Eston, AV Rowlands, S Charlesworth, A Davies, and T Hoppitt. (2005). Prediction of DXA-determined whole body fat from skinfolds: the importance of including skinfolds from the thigh and calf in young, healthy men and women. European Journal of Clinical Nutrition 59, 695–702

J. WANG, a J. C. THORNTON, S. KOLESNIK, AND R. N. PIERSON JR. Anthropometry in Body Composition an Overview. Body Composition Unit, St. Luke's/Roosevelt Hospital, Columbia University, New York, New York 10025, USA

R. T. Withers, N. P. Craig, P. C. Bourdon, and K. I. Norton. (1987). Relative body fat and anthropometric prediction of body density of male athletes. Eur J Appl Physiol (1987) 56:191 200

Hiroshi Shimokata, Jordan D. Tobin,1 Denis C. Muller, Dariush Elahi, Patricia J. Coon, and Reubin Andres. (1989). Studies in the Distribution of Body Fat: I. Effects of Age, Sex, and Obesity. Journal of Gerontology: MEDICAL SCIENCES 1989. Vol. 44. No. 2. M66-73

S. Meeuwsen a, G.W. Horgan b, M. Elia. (2010). The relationship between BMI and percent body fat, measured by bioelectrical impedance, in a large adult sample, is curvilinear and influenced by age and sex. Clinical Nutrition 29 (2010) 560-566

Nicole A Flavel, Timothy S Olds, Jonathan D Buckley (jon.buckley@unisa.edu.au), Matthew T Haren, John Petkov. (2012). Anthropometric estimates of total and regional body fat in children aged 6–17 years, Foundation Acta Pædiatrica 2012 101, pp. 1253–1259

Szyman´ska E, Bouwman J, Strassburg K, Vervoort J, Antti J. Kangas, Soininen P, Mika Ala-Korpela, Westerhuis J, John P.M. van Duynhoven, David J. Mela, Ian A. Macdonald, Rob J. Vreeken, Age K. Smilde, and Doris M. Jacobs. (2012). Gender-Dependent Associations of Metabolite Profiles and Body Fat Distribution in a Healthy Population with Central Obesity: Towards Metabolomics Diagnostics. DOI: 10.1089/omi.2012.0062

Duran I, Martakis K, Rehberg M, Semler O, Schoenau E. Diagnostic performance of an artificial neural network to predict excess body fat in children. Pediatric Obesity. 2018; e12494. https://doi.org/10.1111/ ijpo.12494

Tamas Ferenci, Kovacs L. (2017). Predicting body fat ´ percentage from anthropometric and laboratory measurements using artificial neural networks, (2017), http://dx.doi.org/10.1016/j.asoc.2017.05.063

Kupusinac A, Stokic´ E, Doroslovacki R. (2014). Predicting body fat percentage based on gender, age, and BMI by using artificial neural networks. computer methods and programs in biomedicine 1 1 3 (2 0 1 4) 610–619

Timothy Z. Keith. (2015). Multiple Regression and Beyond: An Introduction to Multiple Regression and Structural Equation Modeling. The second edition was published in 2015 by Routledge 711 Third Avenue, New York, NY 10017.