Universidade Federal de Ouro Preto
Escola de Minas
CECAU - Colegiado do Curso de
Engenharia de Controle e Automação

Sávio Reis Castellar

Exploring the Use of Machine Learning for Body Fat Prediction

Monografia de Graduação

Ouro Preto, 2023

Sávio Reis Castellar

Exploring the Use of Machine Learning for Body Fat Prediction

Trabalho apresentado ao Colegiado do Curso de Engenharia de Controle e Automação da Universidade Federal de Ouro Preto como parte dos requisitos para a obtenção do Grau de Engenheira(o) de Controle e Automação.

Universidade Federal de Ouro Preto

Supervisor: Prof. Ph.D. Rodrigo César Pedrosa Silva

Co-supervisor: Prof. Ph.D. Agnaldo José da Rocha Reis

Ouro Preto

2023

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Sávio Reis Castellar

Exploring the use of machine learning for body fat prediction

Monografia apresentada ao Curso de Engenharia de Controle e Automação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Engenheiro de Controle e Automação

Aprovada em 22 de Março de 2023

Membros da banca

Doutor - Rodrigo Cesar Pedrosa Silva - Orientador - Universidade Federal de Ouro Preto
Doutor - Agnaldo Jose da Rocha Reis - Coorientador - Universidade Federal de Ouro Preto
Doutora - Adrielle de Carvalho Santana - Universidade Federal de Ouro Preto
Mestre - Guilherme Augusto Lopes Silva - Universidade Federal de Ouro Preto

Rodrigo Cesar Pedrosa Silva, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 27/03/2023

Documento assinado eletronicamente por **Rodrigo Cesar Pedrosa Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 27/03/2023, às 09:10, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0497993** e o código CRC **931637A3**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Abstract

The present study investigates the application of machine learning algorithms to estimate body fat percentage. A publicly accessible sample consisting of adult males with their corresponding body fat percentage and anthropometric measurements was utilized to train the various models. The dataset was divided into six categories based on weight range. Five machine learning techniques, including linear regression, decision tree, and random forest, were employed to analyze the data and predict body fat percentage. The results indicated that the linear regression model demonstrated the highest accuracy and that the predictions were more accurate for the group with weight greater than 90 kg. Additionally, it was observed that the measurement of abdominal circumference alone was sufficient for an adequate prediction. In conclusion, this study suggests that machine learning can be a valuable tool for estimating body fat percentage. Further investigation is required to confirm these findings in a larger and more diverse sample population.

Key-words: machine learning; body fat; anthropometric measurements

Resumo

Este estudo investiga a aplicação de algoritmos de aprendizado de máquina para estimar o percentual de gordura corporal. Um conjunto de dados público de indivíduos com seus respectivos percentuais de gordura corporal e medidas antropométricas foi utilizado para treinar os vários modelos. O conjunto de dados foi dividido em seis categorias com base na faixa de peso. Cinco técnicas de aprendizado de máquina, incluindo regressão linear, árvore de decisão e floresta aleatória, foram utilizadas para analisar os dados e prever o percentual de gordura corporal. Os resultados indicaram que o modelo de regressão linear demonstrou a maior precisão e que as previsões foram mais precisas para o grupo com peso maior que 90 kg. Além disso, observou-se que apenas a medida da circunferência abdominal era suficiente para uma previsão adequada. Em conclusão, este estudo sugere que o aprendizado de máquina pode ser uma ferramenta valiosa para estimar o percentual de gordura corporal. Mais investigações são necessárias para confirmar estas conclusões em uma amostra mais ampla e diversificada.

Palavras-chaves: aprendizado de máquina; gordura corporal; medidas antropométricas.

List of Figures

List of Tables

Contents

1 Introduction

The human body exhibits several markers that highlight the importance of proper health care, one of which is the body fat percentage. Body fat plays a crucial role in regulating body temperature and providing energy. While it offers numerous benefits, excessive body fat can also pose a risk to health and increase the likelihood of cardiovascular disease. It is thus imperative to monitor and maintain a healthy body fat percentage (POWELL-WILEY et al., 2021).

Moreover, elevated body fat levels can also negatively impact physical appearance by obscuring muscle definition. For those seeking to improve their physical appearance, reducing body fat and enhancing muscle definition is the primary goal. This can be accomplished by reducing the amount of adipose tissue and making muscle definition more apparent.

Computing an individual's body fat percentage is not a straightforward task. Healthcare professionals often rely on techniques that utilize skinfold calipers to estimate body fat. The seven-site skinfold is the most widely used method of body composition assessment (BARREIRA et al.). As such, it is worthwhile to explain its workings.

The seven-site measurements are all taken on the same side of the body and include the triceps, chest, subscapular, midaxillary, suprailiac, abdominal, and thigh. After recording these measurements, the next step is to enter them into Pollock's equation (JACKSON; POLLOCK, 1978).

$$D = 1.112 - \left(0.00043499 \times \sum 7S\right) + \left[0.00000055 \times \left(\sum 7S\right)^2\right] - (0.00028826 \times A) \tag{1.1}$$

where:

D = body density
$\sum 7S$ = sum of the seven measured skinfolds
A = age

Equation 1.1 gives us the body density, then we need to plug the result into Siri's equation

$$BF = \left(\frac{4.95}{D} - 4.5\right) \times 100 \tag{1.2}$$

where:

BF = body fat

In addition to the challenges associated with the use of equipment, the level of expertise of the evaluator is also a crucial factor to consider. The procedure, which involves a combination of manual labor and precision equipment, typically takes at least 20 minutes to complete. However, there are often significant discrepancies between the results obtained from different methods and equipment options, making it difficult for laypeople to determine which professional is providing accurate results.

Accuracy is also a concern with the use of skinfold calipers, which have a reported accuracy range of around 1 mm (LEGER; LAMBERT; MARTIN, 1982). Despite the presence of an experienced evaluator, there is still inherent uncertainty in the results. To address these issues, bioelectrical impedance has emerged as a potential alternative for body composition assessment in fitness settings. However, the reliability of this method is often disputed, as there is a wide range of product quality in the market, and low-end options tend to provide results that cannot be trusted (BOSY-WESTPHAL et al., 2013).

The gold standard for body composition assessment and evaluation is the Multi-Compartment Models (MCM) (LUKASKI, 2017). MCM involves dividing the body into compartments such as body mass, volume, water, and bone and conducting individual tests for each compartment. For instance, bioelectrical impedance can be used to measure the amount of body water, while Dual Energy X-ray Absorptiometry measures bone density. The strength of this method lies in the combination of multiple tests to obtain comprehensive information about the different body compartments. However, MCM is also associated with disadvantages, such as time cost. With at least three tests required, the process is both expensive and time-consuming, making it less practical for widespread use.

In an attempt to address the challenge of inconsistent results in body fat assessments, a study was conducted to explore the use of machine learning techniques to predict body fat percentage using only anthropometric measurements as input data (UÇAR et al., 2021). This approach offers a dual benefit of obtaining two measurements, body circumferences and body fat, from a single test.

To understand this alternative, a basic understanding of Machine Learning and the process of anthropometric measurements is necessary. Machine Learning, as defined by Arthur Samuel (1959), is a field of study that enables computers to learn without being explicitly programmed. Anthropometric measurements involve taking several body measurements, including abdominal, chest and arm circumferences, weight, height, body mass index (BMI), and skinfold thickness.

The use of a Machine Learning model, as outlined in Uçar et al. (2021), represents an innovative approach in predicting the body fat percentage based solely on anthropometric measurements. This model is capable of learning from the input data and producing predictions by leveraging the information derived from the collected samples. As more

data is added, the model becomes more robust and capable of estimating the values of unseen samples.

However, to further enhance the consistency of the results, it is necessary to address certain limitations. Specifically, defining the relevance of each measured body feature and identifying irregularities in the dataset can help improve accuracy. Additionally, the selection of an appropriate machine learning algorithm is crucial in achieving better results. In this regard, five algorithms will be evaluated and analyzed: Linear Regression, Decision Tree Regression, Random Forest, XGBoost, and Support Vector Machine.

In an effort to simplify the assessment process and reduce the stress on the individual being assessed, it is proposed to minimize the number of necessary features for a good prediction. The partitioning of the dataset into six parts, to understand how the model performs for each weight range, is also a key step in the analysis.

1.1 Objectives

The general objective is to predict the body fat percentage of new samples with the highest degree of accuracy and efficiency, using the minimal number of anthropometric measurements. To attain this goal, various algorithms and techniques will be utilized to optimize the machine learning model and improve its performance.

The specific objectives are:

1. **Apply data cleaning concepts.**
2. **Test different machine learning algorithms.**
3. **Test different feature selection methods.**
4. **Analyze accuracy of models for different weight ranges.**

2 Machine Learning for Body Fat Prediction

2.1 Machine Learning Models

The term "learning" in the context of machine learning refers to the process of improvement through training and experience. A machine learning system can be viewed as a computer system that acquires knowledge and develops new skills from data. This can raise questions such as what constitutes the machine's past experiences and how it is able to gain knowledge without explicit programming. The answer is that the machine's experiences are represented by the data it has been exposed to.

In the same way that humans use memories and associated emotions to form judgments, machines do so by using data. This is the central principle of the learning process. For each piece of information, a new bias is formed. An analogy can be made to the growth of students' grades as they progress in their studies. In machine learning, the performance of the model is measured by its experiences, as described in (MITCHELL, 1997). In essence, machine learning models can be seen as a solution that aims to improve the performance measure P from experience E for a given task T.

With a clear understanding of the underlying principles of machine learning, we can delve into the advantages and disadvantages of utilizing this approach to address real-world problems. One of the main benefits of this method is its ability to efficiently process vast amounts of data, allowing the computer to perform the arduous task of analysis and provide data-driven recommendations. However, it is crucial to note that the quality of the model's predictions is highly dependent on the quality of the input data. In cases where the data is of poor quality, the model's predictions can be significantly impacted. This underscores the importance of investing effort into data corpus development to ensure that high-quality data is fed into the machine learning algorithm.

When it comes to learning tasks, the first step is to comprehend the available data. In the case where both the target and input features are provided, the method used is referred to as supervised learning. For the problem we are trying to solve, which is predicting body fat, a dataset with such structure can be found at [1] and it will be used for our analysis.

It is also essential to note that there are two main types of problems encountered in machine learning: Regression and Classification. The objective of Regression tasks is to predict continuous values, such as price, salary, and age. Conversely, in Classification tasks, the aim is to predict discrete values, such as true or false, or to classify instances into

[1] kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset

predefined categories, such as dog or cat, or spam or not spam. Based on these definitions, it can be inferred that the problem of predicting body fat percentage is a Regression task.

2.1.1 Linear Regression

The regression analysis is a statistical method and, because of its wide applicability in almost every field, it is seen as the basic tool of data science and the the most widely used statistical technique (MONTGOMERY; PECK; VINING, 2021).

In machine learning, linear regression is used to model the linear relationship between a dependent variable and an independent one. In other words, it is a way of finding the best straight line for a dataset. Once the coefficients are determined, the equation can be used to make predictions for new data points. A scatter graph is commonly used to determine this relation as we can see at Figure 1.

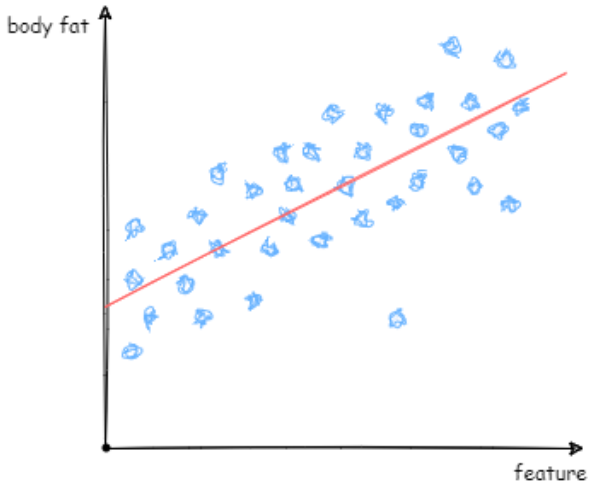

Figure 1 – Linear Regression model.

A model using this method makes a prediction by computing the weighted sum of the input features.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{2.1}$$

where:

y = predicted value
n = number of features
x_i = i^{th} feature value
θ_0 = bias
θ_j = j^{th} feature weight

To better illustrate the usage of this algorithm, we can take the sample from Figure 2 as example. To determine its body fat, we can replace the x values from Equation 2.1. It gives us:

$$y = \theta_0 + \theta_1 \times 0.82 + \theta_2 \times 1.00 + \theta_3 \times 1.93 + \theta_4 \times 87 \tag{2.2}$$

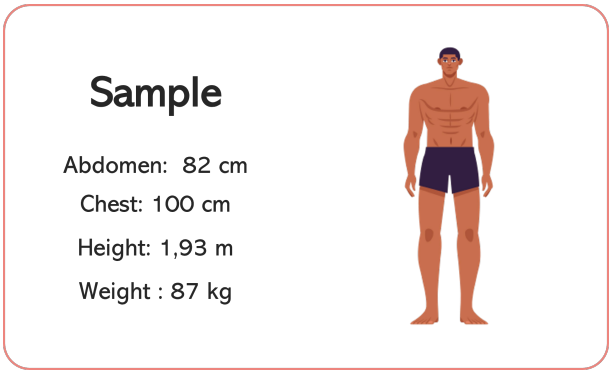

Figure 2 – Example of sample.

2.1.2 Decision Tree

The decision tree model is structured based on a hierarchy of nodes, which divide the dataset according to a comparison criterion defined at each node. Based on the values of the input features, the samples are separated into groups through branches, as seen in the accompanying illustration (Figure 3). This process continues, from branch to branch, until a leaf node is reached, at which point the model generates a prediction based on the input features present at that node.

For regression tasks, such as the prediction of body fat percentage, the predicted value at each leaf node is usually the mean of the target values for all training examples that end up at that node. The final prediction for a new input is made by traversing the tree and computing the mean of the target values at the reached leaf node.

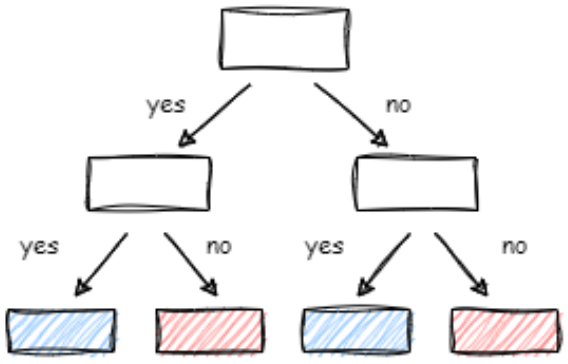

Figure 3 – Example of a Decision Tree structure.

The use of Decision Trees as a method for solving problems provides a clear advantage in terms of interpretability and comprehension of the process. This is achieved through the ease of plotting the trees, which facilitates visual representation of the decision making process. However, it is important to note that deep trees are prone to overfitting, which occurs when a model has a high accuracy on training data but performs poorly on new and unseen data. This highlights the need to balance interpretability with the risk of overfitting.

As depicted in Algorithm 1, the Decision Tree algorithm operates with the set of possible conditions (Cs), the target features (Y), and the set of training examples (Es) as inputs. The output of this algorithm is a function that predicts the value of Y.

To make a prediction in a decision tree for a regression problem, the algorithm follows a series of splitting rules based on the feature values of the input data. At each node of the tree, the algorithm selects the feature that provides the best split, based on the reduction of the mean squared error (MSE) or mean absolute error (MAE). Once the tree is built, a new data point is passed down the tree from the root node, and the splitting rules are applied based on the values of its features, until the prediction is made at a leaf node. The prediction at the leaf node is typically the mean or median value of the target variable for the training examples that reached that node. The resulting prediction is the output of the decision tree algorithm for the given input data point.

Algorithm 1 Decision Tree

procedure DECISION_TREE_LEARNER(Cs, Y, Es)
if stopping criterion is true **then**
  let $v = point\_estimate(Y, Es)$
  define $T(e) = v$
  return T
else
  pick condition $c \in Cs$
  $true\_examples := \{e \in Es : c(e)\}$
  $t1 := Decision\_tree\_learner\ (Cs \setminus \{c\}, \text{Y}, true\_examples)$;
  $false\_examples := \{e \in Es : \neg c(e)\}$
  $t1 := Decision\_tree\_learner\ (Cs \setminus \{c\}, \text{Y}, false\_examples)$;
  define $T(e) =$ if $c(e)$ then $t_1(e)$ else $t_0(e)$
  return T
end if
end procedure=0

To demonstrate the flow of this algorithm, we may utilize the aforementioned example, as shown in Figure 2. By utilizing the decision tree depicted in Figure 4, our sample would traverse the nodes highlighted in red. The quartet of values highlighted in green denote the body fat percentage of the training samples that underwent this identical flow. Thus, we can deduce that our sample shares a physical resemblance with the aforementioned training samples. Consequently, we can posit that calculating the mean of these four values would furnish us with a reliable estimate for the body fat of our sample.

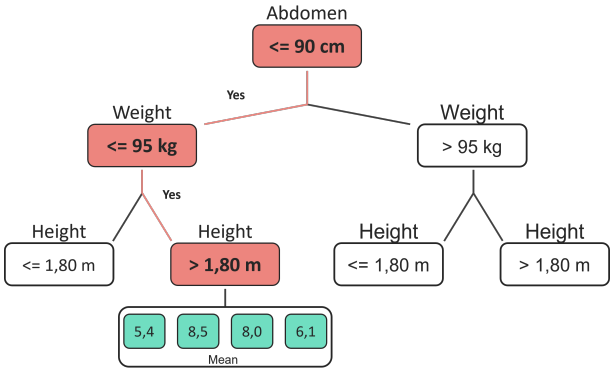

Figure 4 – Example of flow of the Decision Tree.

2.1.3 Random Forest

The Random Forest algorithm is an ensemble model that comprises multiple Decision Trees (as illustrated in Figure 5). This model functions by training multiple Decision Trees on randomized subsets of the data and aggregating the predictions produced by each tree to produce the final prediction. The utilization of an ensemble approach results in improved performance, although this improvement comes with an increase in computational cost.

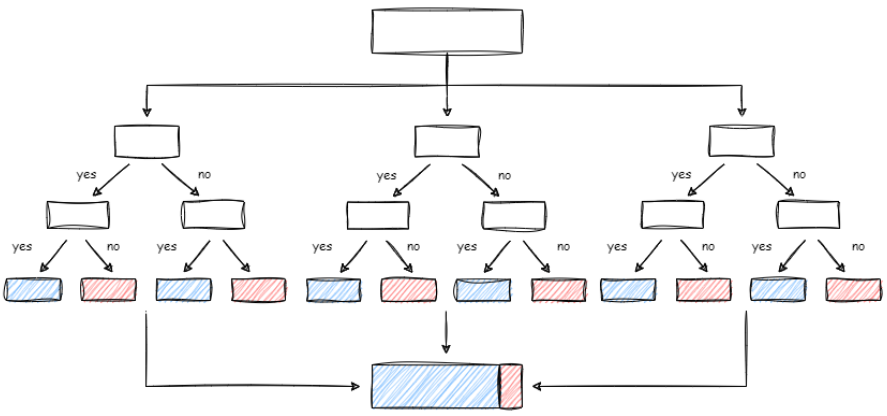

Figure 5 – Example of a Random Forest structure.

The primary benefit of utilizing the Random Forest algorithm is the reduction of overfitting, achieved through the training of multiple Decision Trees on different subsets of the data, which introduces randomness into the model.

For regression problems, the prediction is made by aggregating the predictions of multiple decision trees. Each decision tree is constructed using a random subset of the training data and a random subset of the features. When making a prediction for a new data point, the random forest takes the average of the predictions from all decision trees to arrive at the final prediction.

The decision tree structure depicted in Figure 2 shall now serve as a basis for elucidating the workings of the Random Forest. As an ensemble of trees, the algorithm creates distinct trees, each of which specializes in appraising a particular feature. As a consequence of this, the sample undergoes diverse nodes and criteria, leading to the derivation of distinct means for each tree. To compute the body fat percentage of the given sample, we calculate the mean value across all trees. This example is depicted in Figure 6.

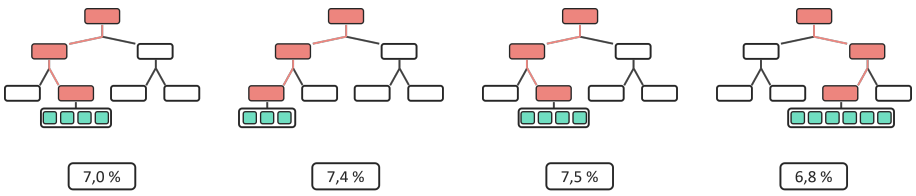

Figure 6 – Usage of a Random Forest.

2.1.4 XGBoost

The XGBoost (eXtreme Gradient Boosting) is a sophisticated decision tree algorithm that leverages the technique of gradient boosting to achieve improved performance. Unlike traditional decision tree algorithms, XGBoost builds trees in a sequential manner, with each tree correcting the errors made by the previous one, leading to a weighted sum of individual weak learners as the final model (CHEN; GUESTRIN, 2016).

To better understand the process, we can follow some steps:

1. The training data is fed into the XGBoost model. The data includes a set of input variables (features) and a target variable (label).

2. The model uses the input variables to make predictions about the target variable. Initially, these predictions will be poor, because the model has not been trained yet.

3. The model then compares the predicted values to the actual values of the target variable in the training data.

4. The model then builds a decision tree based on the input variables, with the goal of minimizing the difference between the predicted values and the actual values.
5. The model then repeats the process, building additional decision trees and using them to correct the mistakes made by the previous trees.
6. The process is repeated until the model has built a sufficient number of trees, at which point the training process is complete.
7. The trained model can then be used to make predictions on new data, using the input variables to predict the value of the target variable.

The model can be represented by the following algorithm, where L represents the loss function $L(y_i, F(x))$, M is the number of weak learners, α is the learning rate and the input is the training set $(x_i, y_i)_{i=1}^N$

Algorithm:

1. Initialize the model

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta) \tag{2.3}$$

2. For m = 1 to M

   a) Compute the 'gradients' and 'hessians':

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{2.4}$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{2.5}$$

   b) Fit a base learner (or weak learner, e.g. tree) using the following training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^N$ by solving the optimization problem below:

$$\hat{\phi}_m = \arg\min_{\phi \in \Phi} \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2 \tag{2.6}$$

$$\hat{f}_{(m)}(x) = \alpha \hat{\phi}_m(x) \tag{2.7}$$

   c) Update the model:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_{(m)}(x) \tag{2.8}$$

3. Output:

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x) \tag{2.9}$$

2.1.5 Support Vector Machine

The Support Vector Machine (SVM) is a type of supervised machine learning algorithm that is utilized for both classification and regression problems. It operates by identifying a hyperplane in a high-dimensional space that maximizes the separation between different classes. This line is found through the analysis of a set of training data. The selected hyperplane serves as the boundary between the different classes and can be used to classify new data points in classification problems. In regression tasks, SVM seeks to find the line that best fits the data and predicts continuous values, as shown at the Figure 7.

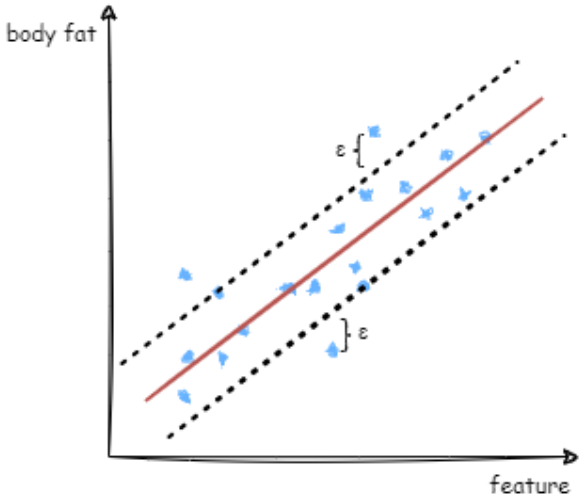

Figure 7 – Example of Support Vector Machine.

Here are the steps of the SVM algorithm:

1. Select a kernel function (e.g. linear, polynomial, radial basis function) to map the input data into a higher dimensional space where it can be linearly separated.

2. Find the hyperplane that best fits the data while minimizing the margin violations, which is the difference between the predicted value and the actual value of the target variable.

3. Use the support vectors to define the hyperplane equation, which can then be used for making predictions on unseen data.

The optimization problem in the Support Vector Machine (SVM) algorithm aims to minimize the discrepancy between the predicted and actual values by utilizing the epsilon-insensitive loss function as the cost metric. The epsilon-insensitive loss function enables the algorithm to tolerate errors within a specified threshold, which is represented by the epsilon (ϵ) parameter (SHARP, 2020).

SVM is a robust machine learning algorithm that can effectively handle non-linear problems, handle high-dimensional spaces and deal with a large number of features. Despite its strengths, SVM may entail a longer processing time when training on large datasets and does not provide probability estimates. It is imperative to consider the complexity of the problem and data when selecting SVM as the solution.

For the linear kernel, the objective function is the following weighted sum:

$$y = x'\theta' + \theta_0 \tag{2.10}$$

where:

y = predicted value
x' = feature value vector
θ' = feature weight vector
θ_0 = bias

During the training, the algorithm aims to minimize the vector of weights and the sum of acceptable errors. The objective in doing it is to find the line that best fits the data.

$$\min \frac{1}{2}\theta'\theta + \theta_0 \sum(\varepsilon + \varepsilon^*) \tag{2.11}$$

where:

ε = distance from data and soft margin

Subject to the following restrictions:

$$\forall : y_n - (x'_n\theta' + \theta_0) \leq \epsilon + \varepsilon_n \tag{2.12}$$

$$\forall : (x'_n\theta' + \theta_0) - y_n \leq \epsilon + \varepsilon^*_n \tag{2.13}$$

$$\forall : \varepsilon^*_n \geq 0 \tag{2.14}$$

$$\forall : \varepsilon_n \geq 0 \tag{2.15}$$

2.2 Feature Selection

In the field of machine learning, feature selection refers to the process of determining a subset of input features that are most relevant to the model being trained. The objective of feature selection is to enhance the model's generalization capacity, reduce

computational resources required for model building and application, and to improve the model's learning performance and decrease memory storage (LI et al., 2017).

Besides that, there are several reasons why feature selection is important:

1. Reducing the number of input features can reduce the risk of overfitting, which occurs when a model is too complex and fits the training data too closely, resulting in poor generalization to unseen data.

2. A smaller number of features can make the model easier to interpret and understand, which can be useful for explaining the model's predictions to stakeholders.

3. Using a smaller number of features can also make the model faster to train and apply, which can be important when working with large datasets or when the model needs to be deployed in real-time.

There are various techniques available for feature selection, including manual selection, wrapper methods, and filter methods. In the present case, for the selection of the most suitable features from the anthropometric measurements, a range of methods were tested, including Pearson Correlation, Recursive Feature Elimination, and Lasso.

2.2.1 Pearson Correlation

Pearson correlation is a measure of the linear relationship between two continuous variables. It ranges from -1 to 1, where -1 represents a perfectly negative correlation, 0 represents no correlation, and 1 represents a perfectly positive correlation.

A positive correlation means that as the value of one variable increases, the value of the other variable also increases. For example, there may be a positive correlation between the number of hours a student studies and their test scores. As the number of hours studied increases, the test scores may also increase. For the negative correlation, we have the value of one variable increasing while the value of the other decreases.

With respect to the prediction of body fat percentage, the correlation matrix (Figure 8) reveals a positive correlation between body fat percentage and abdominal circumference, meaning that as one increases, so does the other.

The Pearson Correlation can be calculated using the following formula:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} \tag{2.16}$$

where:

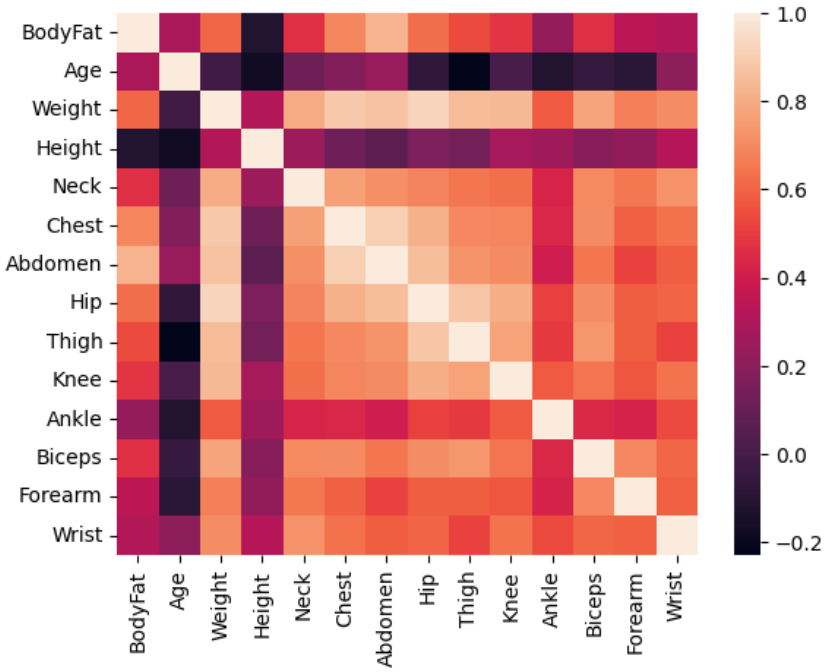

Figure 8 – Correlation matrix.

r = predicted value
x_i = values of the x-variable in a sample
$\overline{x}$ = mean of the x values
y_i = values of the y-variable in a sample
$\overline{y}$ = mean of the y values

2.2.2 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a method of feature selection that involves recursively removing features from the model until a desired level of performance is achieved.

Here is a simple example of how RFE might work:

1. Begin with a set of input features and train a model using all of the features.
2. Calculate the performance of the model using some evaluation metric (e.g. mean squared error or mean absolute error).
3. Remove the feature that has the least impact on model performance, as measured by the evaluation metric.
4. Train a new model using the remaining features.
5. Repeat steps 2-4 until a desired level of performance is achieved or until there are no more features to remove.

The Recursive Feature Elimination (RFE) method is an effective tool for selecting a relevant subset of features for a model, taking into account the interplay between features. Despite its benefits, it can be computationally intensive, as it necessitates the training of multiple models during the process of adding and removing features.

2.2.3 LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) equation is a mathematical formula utilized in the context of regression analysis for the purpose of feature selection. This equation involves the addition of a regularization term to the linear regression objective function, where the regularization term is calculated as the sum of the absolute values of the coefficients multiplied by a penalty parameter, denoted as lambda (λ). The optimization of the objective function, with the added LASSO term, results in a sparse solution where some coefficients are set to zero, thereby selecting only a subset of the features as important predictors in the model. The equation can be represented mathematically as:

$$\min_{\beta} \frac{1}{2n} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \tag{2.17}$$

Where y represents the response variable, X denotes the feature matrix, β represents the vector of coefficients, n is the number of samples, and $||.||_1$ and $||.||_2$ represent the L1 and L2 norms, respectively. The L1 norm encourages the coefficients to have as small a magnitude as possible, while the penalty parameter λ controls the balance between the fit of the model to the data and the magnitude of the coefficients.

Here are the steps of the method:

1. Begin with a set of input features and train a model using all of the features.
2. Calculate the importance of each feature using the coefficients of the model.
3. For each feature, add a penalty term to the cost function that shrinks the coefficient of the feature towards zero if it is not important.
4. Train a new model using the modified cost function.
5. Repeat steps 2-4 until the desired level of feature selection is achieved.

Comparing the steps to the aforementioned method (2.2.2), we can see that the difference is at the third step. The LASSO method adds a penalty term to the equation.

2.3 Model Evaluation

In the field of machine learning, model evaluation refers to the systematic assessment of the performance of a model on a specified dataset. The objective of model evaluation is to quantify the discrepancy between the predicted output of the model and the actual output. This is achieved by comparing the two outputs and calculating an appropriate metric that reflects the difference. This chapter will provide an overview of some commonly used metrics for model evaluation.

2.3.1 Cross-validation

The method of Cross-validation is employed to evaluate the effectiveness of a model by dividing the available data into multiple subsets. The model is trained on one subset and tested on another, with the final result being the average of all iterations. This repetition of the process with different partitions leads to an improvement in the model's ability to estimate unseen data, resulting in a more robust and accurate evaluation.

There are two commonly used approaches to Cross-validation, namely k-fold and leave-one-out. Both approaches have their own advantages and disadvantages, which will be discussed in the subsequent section.

2.3.2 K-fold

The K-fold cross-validation consists in splitting the dataset into k random folders with approximately equal size. By each iteration, the model is trained and tested with different sets. There are four factors impact the accuracy of the model and must be considered when using this method of cross-validation (WONG, 2015):

- The number of folds.
- The number of samples in a fold.
- The level of averaging.
- The repetition of k-fold cross validation.

2.3.3 Leave-One-Out

Leave-one-out cross validation (LOO-CV) is a special type of k-fold cross validation. In LOO-CV, the data is partitioned into subsets by leaving out one sample at a time. The model is then trained on all samples except for the one that was left out, and, then, it is tested on that single one. This process is repeated for every data point in the dataset, so that each point is used once as a test point and the remaining points are used for training.

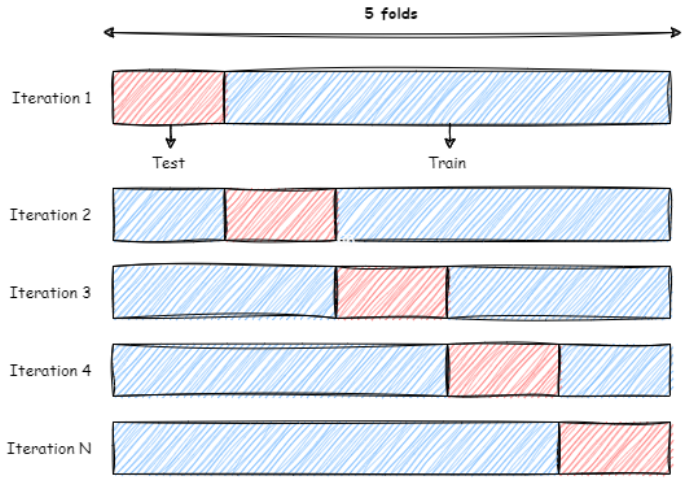

Figure 9 – K-fold cross-validation.

This method is useful when the dataset is small and the goal is to maximize the number of test points while still having a sufficient number of training points. By using each sample as the test set, the process is more computational intensive in comparison to k-folds. The advantage is that there is no randomness in the evaluation.

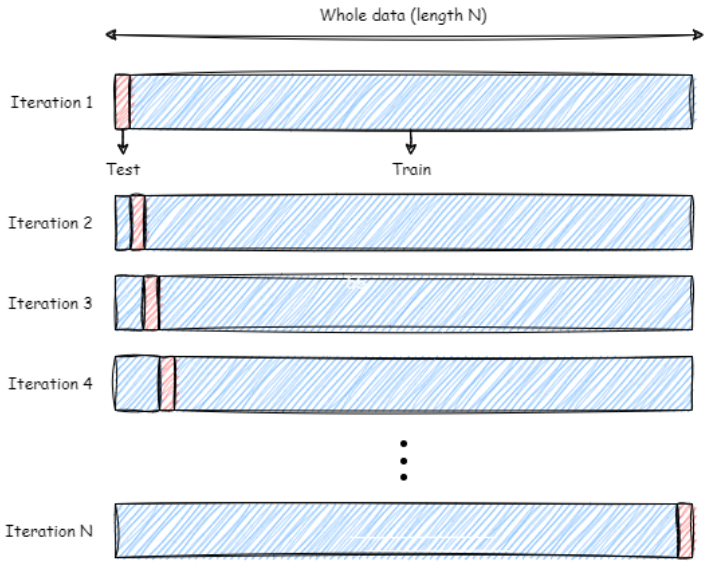

Figure 10 – Leave-one-out cross-validation.

2.4 Evaluation Criteria

An evaluation criterion is used to evaluate the performance of a model in predicting an outcome variable. That means it is a measure of the difference between predicted and actual values. For regression problems, two common criteria are recommended: The Root Mean Squared Error and the Mean Absolute Error (JAMES et al., 2013).

2.4.1 Root Mean Squared Error

The Root Mean Squared Error (RMSE) is the square root of the average of the squared differences between the predicted and observed values. In simple terms, it is a way to measure how far the predicted values are from the actual values, on average, and it is expressed in the same units as the outcome variable. A lower RMSE indicates a better fit of the model to the data (WILLMOTT; MATSUURA, 2005).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \tag{2.18}$$

where:

y_i = predicted value
x_i = true value
n = sample size

2.4.2 Mean Absolute Error

Mean Absolute Error (MAE) is the average of the absolute differences between the predicted and actual values. Unlike the Root Mean Squared Error(RMSE), which gives more weight to larger errors, MAE gives equal weight to all errors. A lower MAE indicates a better fit of the model to the data (WILLMOTT; MATSUURA, 2005).

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \tag{2.19}$$

where:

y_i = predicted value
x_i = true value
n = sample size

2.4.2.1 Mean Absolute Percentage Error

As MAE measures the average absolute difference between predicted and actual values, without taking into account the scale of the data. For example, if the MAE of a model is 5, it means that the average absolute error in the predictions is 5 units.

MAPE, on the other hand, measures the percentage difference between the predicted and actual values. This means that it takes into account the scale of the data and allows for a more meaningful comparison across different data sets. For example, if the MAPE of a model is 10%, it means that the average percentage error in the predictions is 10%.

To relate MAPE and MAE, we can say that MAPE is a normalized version of MAE that expresses the errors as a percentage of the actual values. In other words, MAPE is MAE divided by the actual value, multiplied by 100

Mathematically, we can express this relationship as:

$$MAPE = \frac{MAE}{ActualValue} \times 100\% \tag{2.20}$$

2.5 Machine Learning for Body Fat Estimation: Proposed Approach

The objective of this study was to model a machine learning problem by utilizing five learning algorithms and three feature selection techniques, in accordance with the flowchart depicted in Figure 11. The primary purpose was to assess the performance of the various combinations of learning algorithms and feature selection methods.

2.5.1 The dataset

The dataset utilized in this study encompasses 252 observations and 15 variables, comprising estimates of the percentage of body fat obtained through underwater and anthropometric measurements. It is crucial to note that the dataset is comprised of adult males and, therefore, the results should not be generalized beyond this population.

In the table below we can see variables available:

Id	Variable
1	Density determined from underwater weighing
2	Percent body fat from Siri's equation (1.2)
3	Age (years)
4	Weight (kg)
5	Height (inches)
6	Neck circumference (cm)
7	Chest circumference (cm)
8	Abdomen 2 circumference (cm)
9	Hip circumference (cm)
10	Thigh circumference (cm)
11	Knee circumference (cm)
12	Ankle circumference (cm)
13	Biceps (extended) circumference (cm)
14	Forearm circumference (cm)
15	Wrist circumference (cm)

Table 1 – Variables in dataset.

2.5.2 Methodology

1. The initial step in preparing the data for modeling involved the removal of inconsistent samples from the dataset, leaving 249 observations for further processing. In addition, a column containing "density" values was dropped from the dataset as it was dependent on the target value (body fat) for calculation, yielding a total of 13 features.

2. As apparent from Table 1, there are at least four distinct measurement units. Hence, the normalization of the data is essential to facilitate their assimilation into a model. Consequently, the subsequent step involved scaling all the data within the range of 0 to 1.

3. To analyze the efficiency of the model over different weight ranges, the dataset was splitted as follows:

   - All samples.
   - Less than 70 kg - 45 samples.
   - 70 kg to 75 kg - 39 samples.
   - 75 kg to 80 kg - 40 samples.
   - 80 kg to 90 kg - 66 samples.
   - Greater than 90 kg - 59 samples.

4. For each partition, Leave-One-Out cross-validation was executed to compute the mean of the evaluation metrics.

5. Subsequently, an examination of three feature selection methods, namely Pearson Correlation, Recursive Feature Elimination (RFE), and Least Absolute Shrinkage and Selection Operator (LASSO), was conducted to produce a rank. This rank was then applied to every training set of the partitions to establish a hierarchy of features.

6. In the following step, the aforementioned five learning algorithms, namely Linear Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Machine, were utilized to train the datasets using the selected features obtained through the rank.

The full factorial design of variations leaded to ninety experiments. All the analyzes were conducted in Python using the scikit-learn library and all the implementation of models used the default parameters.

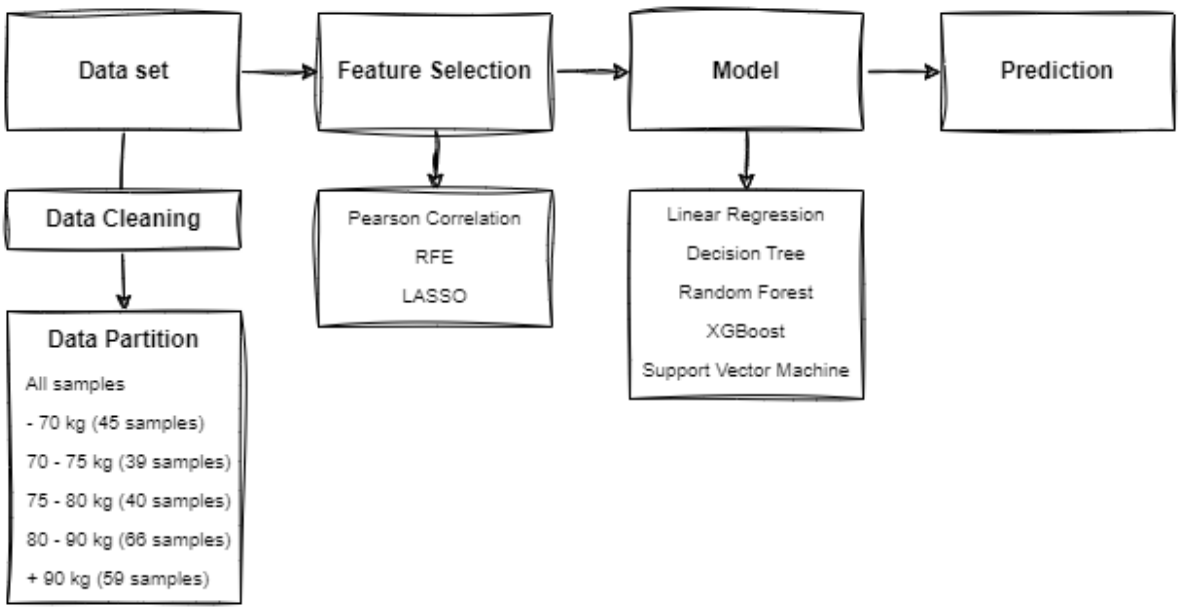

Figure 11 – Methodology flowchart.

3 Results

3.1 Algorithms

In accordance with the aforementioned methodology, we conducted an evaluation of the performance of five distinct machine learning algorithms for predicting body fat percentage. Figure 12 illustrates the performance of the each algorithm, using MAPE as evaluation criteria. The results obtained from our study demonstrate significant disparities in the performance of the algorithms.

The Linear Regression (LR) algorithm demonstrated the best performance, with a mean absolute percentage error of 26.9%. The Random Forest (RF) and XGBoost (XGB) algorithms also showed good results, with a mean MAPE of 28.6% and 29.9%, respectively, which were slightly larger than the performance of the Linear Regression algorithm. However, the standard deviation for both of them is smaller, which compensates for the higher mean. For the SVR algorithm, with a mean absolute error of 35.5%, it was found to be the worst performing algorithm.

In summary, the results of this study suggest that the Linear Regression algorithm is the best performing algorithm for predicting body fat percentage, followed closely by the random forest and XGBoost algorithms. The decision tree algorithm was found to perform worse than the other algorithms, and the SVR algorithm had the highest error rate.

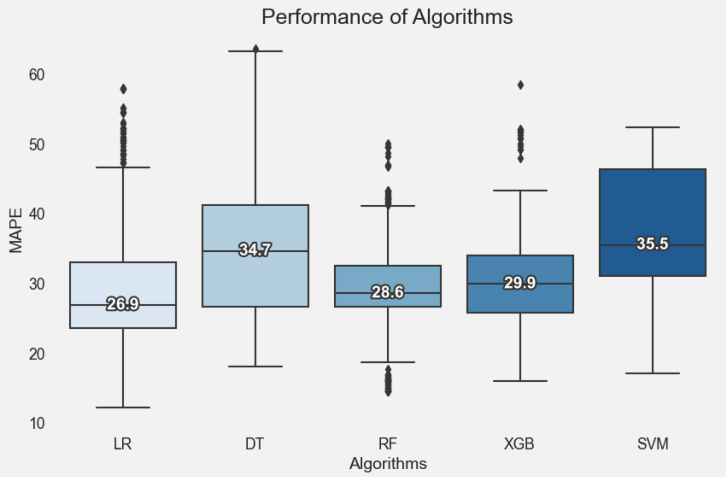

Figure 12 – MAPE for each ML algorithm.

3.2 Partitions

Continuing with the analysis, the performance of machine learning models was evaluated by partitioning the data according to weight ranges. The dataset was divided into five weight ranges, namely, less than 70 kg, 70 to 75 kg, 75 to 80 kg, 80 to 90 kg, and greater than 90 kg.

The results illustrated by Figure 13 showed that the best-performing models (LR, RF and XGB) were observed in the weight range of more than 90 kg, with a mean MAPE of 17.8%. When using all the available 249 samples, the mean MAPE achieved was 29.1%. However, the partition with samples weighing less than 70 kg showed the worst results (46.5%).

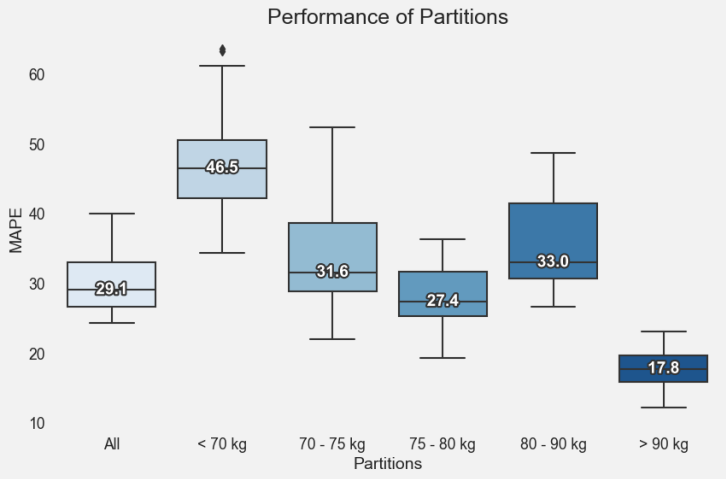

Figure 13 – MAPE for each partition.

Furthermore, the performance of algorithms varied with different weight ranges. Figure 14 illustrates the algorithmic performance for each weight range. For the partitions with samples weighing less than 70 kg and from 75 to 80 kg, Random Forest outperformed the other algorithms. For all the other weight ranges, Linear Regression performed better.

3.3 Feature Sets

The table denoting feature identifiers, presented in Table 2, provides a numerical identification to each feature of the training dataset to aid in comprehending the subsequent tables. An examination of Table 3 reveals that all three feature selection methods agree that abdominal circumference is the most significant feature in the training set. However, the importance of the other features varies among the methods. Therefore, it is necessary to rank these features based on their relevance in each method, as presented in Table 4.

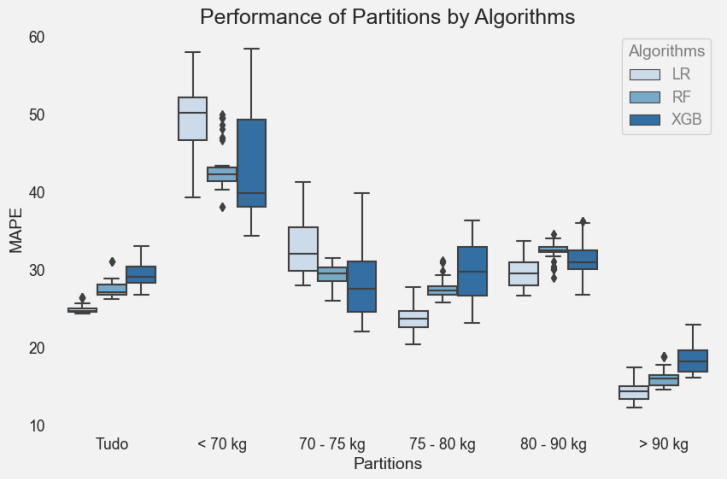

Figure 14 – MAPE for each partition by algorithm.

Id	Feature (Name)
1	Age
2	Weight
3	Height
4	Neck
5	Chest
6	Abdomen
7	Hip
8	Thigh
9	Knee
10	Ankle
11	Biceps
12	Forearm
13	Wrist

Table 2 – Feature Id.

Id	LASSO	RFE	Pearson
1	6	6	6
2	13	2	5
3	2	13	7
4	4	7	2
5	1	8	8
6	12	1	9
7	3	4	11
8	7	5	4
9	8	12	12
10	11	11	13
11	5	3	1
12	9	10	10
13	10	9	3

Table 3 – Order of feature importance for each method.

The relationship between the performance of each feature set for the rank from Table 4 is illustrated in Figure 15. The conclusion drawn from the image is that adding more features generally worsens the performance of the models after the third feature. This observation can be attributed to the varying importance of each feature in relation to body fat. Certain features are more closely related to body fat than others, which may account for this trend.

3.4 Feature Selection Methods

The performance of each feature selection method is depicted in Figure 16, where all available features in the dataset are used. The inference drawn from the figure is that the choice of method will not have a substantial influence on our models since we have a difference of less than one percent between their means and their deviation is similar.

Rank	Feature	Score
1	Abdomen	3
2	Weight	9
3	Wrist	15
4	Hip	15
5	Thigh	19
6	Age	19
7	Neck	21
8	Chest	22
9	Forearm	24
10	Biceps	27
11	Height	31
12	Knee	31
13	Ankle	37

Table 4 – Rank of features by score.

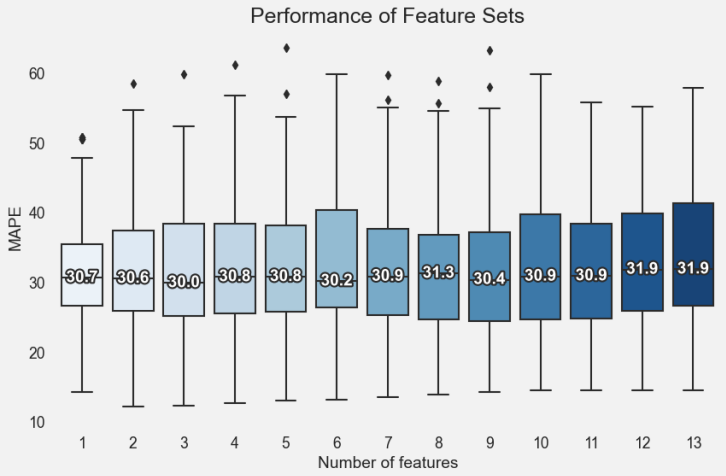

Figure 15 – MAPE for each feature set.

The performance of each method according by each algorithm is illustrated by Figure 17. From that, we can notice that changing the algorithm is more impactful than chnaging the feature selection method.

Based on the results presented in Figure 15, it can be inferred that superior outcomes can be achieved by considering solely the initial three features (abdomen, weight and wrist). Additionally, Figure 18 showcases the performance of feature sets separated by the feature selection techniques employed. It can be discerned that the Pearson method yields the most optimal models, registering a mean accuracy of 27.7% while only utilizing the first two features. On the other hand, the other techniques require the inclusion of the third feature to enhance performance, leading to a mean accuracy of 28.5%.

The association between feature selection methods and algorithms is demonstrated in Figure 19. Based on the findings, it can be inferred that selecting an appropriate algorithm contributes more significantly to the outcome than the feature selection method.

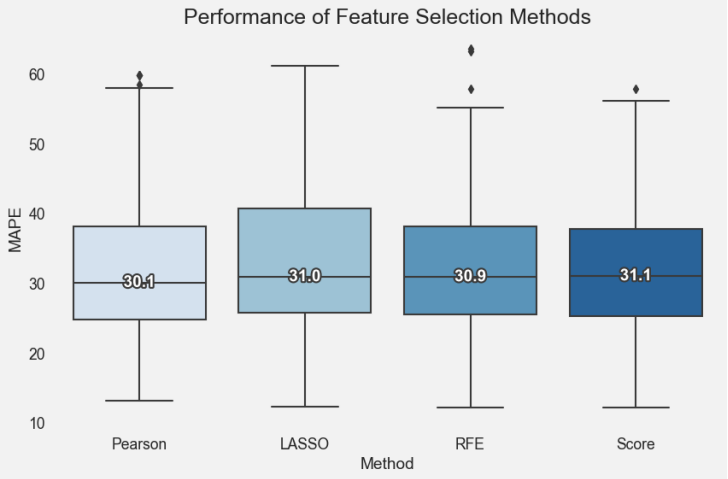

Figure 16 – MAPE for each feature selection method.

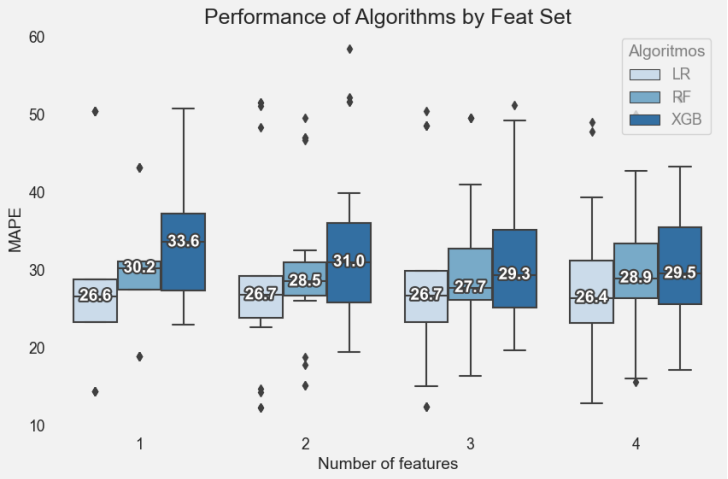

Figure 17 – MAPE for each feature selection method.

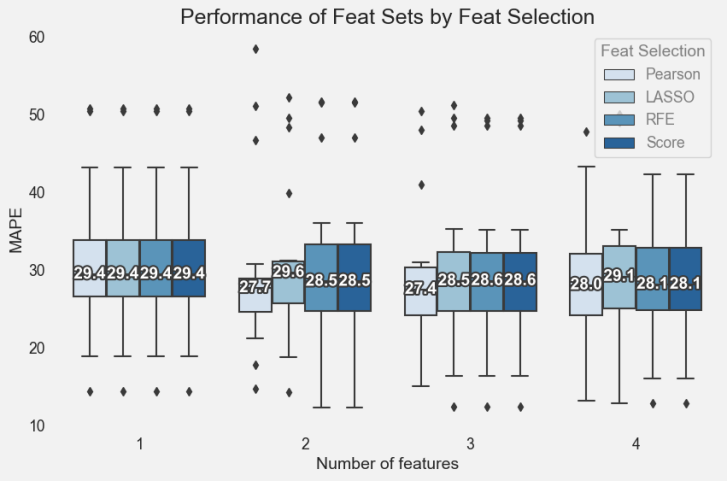

Figure 18 – MAPE for each feature selection method.

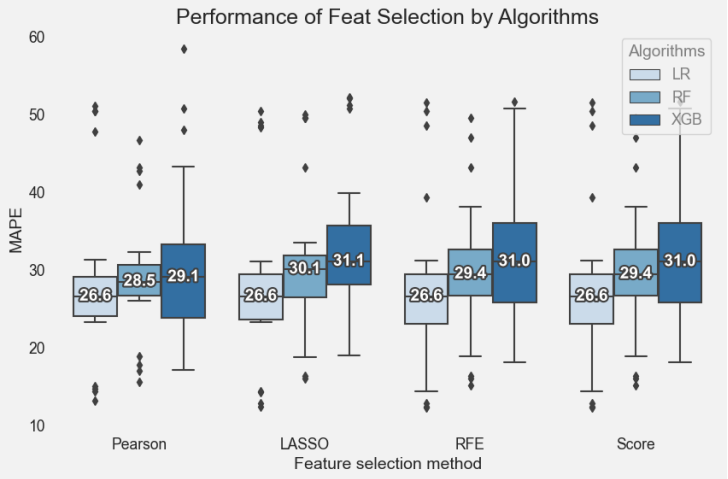

Figure 19 – MAPE for each feature selection method.

4 Conclusion and Future Work

In conclusion, the present study aimed to assess the accuracy of machine learning models in predicting body fat percentage through the examination of various combinations of algorithms, feature selection methods, data partitions, and feature sets. The results indicated that the solution performed well for the partition with more than 90 kg, with a mean absolute percentage error of 12.22 % for the best model (Linear Regression + Pearson + abdomen circumference). The best model for each partition can be seen at Table 5. While the values generated by the model can serve as a reference, further research is necessary to validate its efficacy on larger and more diverse populations. This investigation underscores the viability of using machine learning for body fat prediction and paves the way for further exploration in the field.

Partition	Algorithm	Feat Selection	No Feats	MAPE (%)
All	LR	Score	2	24.35
$<$ 70 kg	XGB	Pearson	8	34.38
70 - 75 kg	XGB	Score	7	22.05
75 - 80 kg	DT	Score	11	19.38
80 - 90 kg	LR	Pearson	1	26.68
$>$ 90 kg	LR	Score	2	12.22

Table 5 – Best model by partition.

Linear Regression showing the better results can be explained by two factors:

- Linearity between the most important feature (abdomen circumference) and the target (body fat).

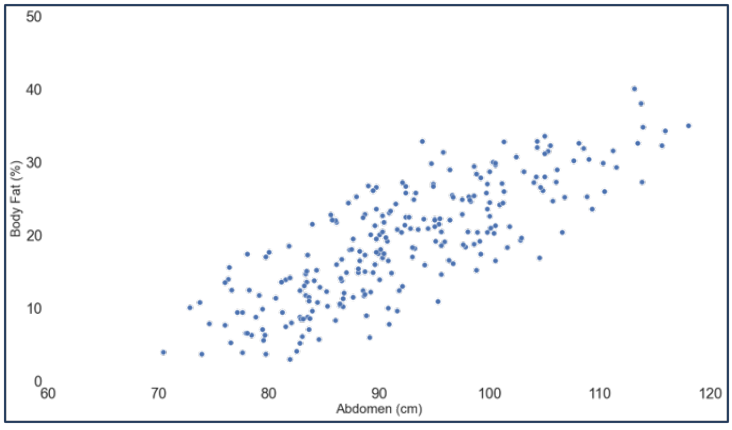

Figure 20 – Linearity between Abdomen and Body Fat.

- Lack of fine-tuning of the other algorithms.

As future work, it is recommended to validate the model on a larger and more diverse population and to conduct the following:

- A comparison of the machine learning model with alternative methods of body fat prediction, such as skinfold thickness measurements or bioelectrical impedance analysis.
- The tune of the models by setting specific parameters to achieve better accuracy in predicting body fat percentage.
- The hypothesis about the motives for the best predictions for greater 90 kg cases.

References

BARREIRA, Tiago V et al. The validity of 7-site skinfold measurements taken by exercise science students. *International Journal of Exercise Science*, v. 6, n. 1, p. 4, 2013. **citecountpage** 8.

BOSY-WESTPHAL, Anja et al. What makes a BIA equation unique? Validity of eight-electrode multifrequency BIA to estimate body composition in a healthy adult population. *European journal of clinical nutrition*, Nature Publishing Group, v. 67, n. 1, s14–s21, 2013. **citecountpage** 9.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: PROCEEDINGS of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. P. 785–794. **citecountpage** 16.

JACKSON, Andrew S; POLLOCK, Michael L. Generalized equations for predicting body density of men. *British journal of nutrition*, Cambridge University Press, v. 40, n. 3, p. 497–504, 1978. **citecountpage** 8.

JAMES, Gareth et al. *An introduction to statistical learning*. Springer, 2013. v. 112. **citecountpage** 25.

LEGER, Luc; LAMBERT, J; MARTIN, P. Validity of plastic skinfold caliper measurements. Human Biology, 54(3): 667-675, 1982. *Human biology*, v. 54, p. 667–75, Dec. 1982. **citecountpage** 9.

LI, Jundong et al. Feature Selection: A Data Perspective. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 50, n. 6, Dec. 2017. ISSN 0360-0300. DOI: 10.1145/3136625. Available from: https://doi.org/10.1145/3136625. **citecountpage** 20.

LUKASKI, Henry C. *Body composition: health and performance in exercise and sport*. CRC Press, 2017. **citecountpage** 9.

MITCHELL, Tom M. *Machine Learning*. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2. **citecountpage** 11.

MONTGOMERY, Douglas C; PECK, Elizabeth A; VINING, G Geoffrey. *Introduction to linear regression analysis*. John Wiley & Sons, 2021. **citecountpage** 12.

POWELL-WILEY, Tiffany M et al. Obesity and cardiovascular disease: a scientific statement from the American Heart Association. *Circulation*, Am Heart Assoc, v. 143, n. 21, e984–e1010, 2021. **citecountpage** 8.

SHARP, Tom. An introduction to support vector regression (svr). *Towards Data Science*, 2020. **citecountpage** 18.

UÇAR, Muhammed Kürşad et al. Estimation of body fat percentage using hybrid machine learning algorithms. *Measurement*, v. 167, p. 108173, 2021. ISSN 0263-2241. DOI: https://doi.org/10.1016/j.measurement.2020.108173. Available from: https://www.sciencedirect.com/science/article/pii/S0263224120307119. **citecountpage** 9.

WILLMOTT, Cort J; MATSUURA, Kenji. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, v. 30, n. 1, p. 79–82, 2005. **citecountpage** 25.

WONG, Tzu-Tsung. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, Elsevier, v. 48, n. 9, p. 2839–2846, 2015. **citecountpage** 23.