Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v18n2p458

Numerical methods to predict health-related Metrics. A case of great economic impact: the Body Mass Index
By di Tollo et al.

15 October 2025

Electronic Journal of Applied Statistical Analysis
Vol. 18, Issue 2, October 2025, 458-479
DOI: 10.1285/i20705948v18n2p458

Numerical methods to predict health-related Metrics. A case of great economic impact: the Body Mass Index

Giacomo di Tollo*[a], Nora Kordic[b], Massimo Squillante[c], and Salvador Cruz Rambaud[d]

[a] *Department of Management, Polytechnic University of Marche, Piazzale Martelli, 8, Ancona I-60121, Italy*
[b] *University of Luxembourg, 2 Avenue de l'Universite, Esch-Belval Esch-sur-Alzette, 4365, Luxembourg*
[c] *Accademia Peloritana dei Pericolanti, Piazza Pugliatti, 1, Messina, 98122,Italy*
[d] *Department of Economics and Business, University of Almería, Carretera Sacramento s/n, La Canada de San Urbano, Almería, 04120, Spain*

15 October 2025

Public health represents nowadays one of the major global challenges, especially with regards to the predictive analysis of biometric data. In this framework, Body Mass Index (referred to as *BMI*) represents one of the most recognized indicators for monitoring, understanding and forecasting the general health of the population. An analysis of *BMI* could have important social and economic impacts, allowing policy-makers to foster more and more sustainable economic development and to develop preventive strategies (and effective health policies) aimed at promoting social well-being. This contribution proposes the use of computational approaches to model *BMI* using both traditional statistical methods and AI techniques, namely linear regression, random forest, decision trees and neural networks. In particular, a comparative analysis of the predictive performance of the models mentioned above is proposed by discussing the significance of different health-related indicators on *BMI*. The computational analysis is conducted on a dataset consisting of health parameters of a sample of women in Belgium collected between 2000 and 2001. The results demonstrate the effectiveness of linear and AI-based approaches in predicting *BMI* and provide valuable information for policy

*Corresponding author: g.ditollo@univpm.it.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

makers interested in quantitative assessment of health parameters and disease prediction to promote strategic choices that can contribute to improving collective well-being and reducing health and economic disparities, generating benefits at both the individual and community levels.

keywords: Body Mass Index, Artificial Intelligence, Computational Approach, Numerical methods, Sustainable economic development.

1 Introduction

Population health represents nowadays one of the most debated topics amongst the academic community, health professionals, and policy-makers (Turnock, 2012). In this context, the deployment of Big Data approaches in health-related topics plays a crucial role (Gamache et al., 2018), along with the development of methodological approaches for their analysis (Wells et al., 2016). In fact, data-driven approaches to map and model health-related dynamics would allow policy makers to define customized and sustainable interventions for the population (Kuo et al., 2014), leading to more efficient management of resources, improved outcomes in public health, and policy formulation directed towards certain demographic and epidemiological needs. Additionally, these methods can be employed to better predictive analytics: in this way it would be possible to early identify potential health threats, and to optimize healthcare provisions, leading to a sustainable and responsive healthcare system. In recent decades, a larger and larger public has paid more and more attention to the individuals' body weight: individuals' weight can be recognized as one of the most critical factors to define the individuals' general health and well-being, and many health problems could derive from weight, including hypertension, cancer, diabetes and heart disease (Khanna et al., 2022; Tanaka, 2020). In this contribution we focus on the individual's *Body Mass Index*, referred to as *BMI*, which is calculated as:

$$BMI = weight/height^2$$

where *weight* represents the individual's weight measured in kilograms, and *height* represents the individual's height measured in meters: this indicator is currently used to assess the risk of developing the afore mentioned diseases, and it is important to emphasize that it is intended as a population-level indicator, designed to provide a practical comparative measure across adults of varying stature, rather than a direct assessment of individual body fatness or health status. For instance, individuals with high muscle mass may exhibit elevated BMI values while maintaining low body fat levels. The normalization achieved by squaring height in the denominator enhances comparability between shorter and taller individuals. From a geometric perspective, human beings may be conceptualized as three-dimensional entities: in proportional growth, height scales linearly with one dimension, whereas weight (mass) scales approximately with volume, which is a cubic function of height. Nevertheless, body mass is also influenced by body density, whose inter-individual variability is generally moderate and not known a priori. Consequently, the use of height squared represents a pragmatic compromise for

normalizing weight relative to body size. This approach, as originally demonstrated in the seminal works of Adolphe Quetelet during the first half of the nineteenth century (Quetelet, 1835), provides the most stable relationship between weight and stature at the population level.

Based on the *BMI* values, it is possible to categorize adults into *underweight*, *healthy weight*, *overweight* and *obese* (Weir and Jan, 2019): this categorization can be summarised in Table 1 (Nuttall, 2015).

***BMI* value**	**Weight Status**
Below 18.5	Underweight
18.5 to 24.9	Healthy weight
25.0 to 29.9	Overweight
30 and higher	Obese

Table 1: Classification of the individual's weight status according to the individual's *BMI* values.

Since weight represents the projection of the individual's health, *BMI* is also used as a tool to identify the risk of the onset (or prevalence) of different health problems and as an indicator of the individual's energy balance. In this framework, we can assert that food is a source of energy that, when not used and burned, leads to the creation of fat tissue. According to Chambers and Swanson (Chambers and Swanson, 2010), the factors that influence the energy balance have an influence on the value of *BMI* and can be grouped as *Dietary* and *Activity* factors. *Dietary* factors include aspects of both eating habits and general behavior associated with eating such as healthy eating (i.e. reduced consumption of nutritious foods), emotional eating (i.e., the increased or excessive consumption of nutrients linked to particular emotional states), or social influences on eating. *Activity* factors refer instead to body movement and general physical activity. In addition to physical exercise and diet, there are other factors that can have an impact on *BMI* (Chambers and Swanson, 2010): in particular, reference can be made to *dietary behavior* (according to which a greater frequency and/or a lesser effectiveness in weight control can have both a positive and negative impact on *BMI*), *alcohol consumption* (an increase in alcohol intake can influence *BMI*), *amount of sleep* (the reduction in hours of sleep can have a negative impact on the *BMI*, making it either too low or too high), and *early maturation*, according to which children who grow faster than their peers can have a higher *BMI*.

Given the importance mentioned above, *BMI* has been investigated considering different criteria, ranging from prediction of diseases to mortality based on *BMI*: Khanna et al. (Khanna et al., 2022) states that *BMI* is important in this sense, referring to the fact that it helps assess chronic conditions such as the risk of high blood pressure, diabetes, cancer, high cholesterol, and diseases in later life. (see Section 2 for more details).

Many contributions stress the importance of an analysis of *BMI* to provide general implications on public health (Wang et al., 2021), health economics (Müller-Riemenschneider

et al., 2008; Kouris-Blazos and Wahlqvist, 2007; Kelly et al., 2019) and social policies (Swinburn, 2008). From an economic point of view, studies on *BMI* have important implications on the containment of health costs (Kent et al., 2017), on the improvement of employee productivity (Finkelstein et al., 2010; Finkelstein and Trogdon, 2008; Dall et al., 2024), not to mention the use of *BMI* data to define policies aimed at optimizing public resources (Finkelstein and Trogdon, 2008). However, despite its importance for academics and practitioners is clearly addressed, the computational approaches to model, predict, or forecast *BMI* are still few and far between: some contributions focus on predicting obesity and *BMI* using childhood data through logarithmic transformed *BMI*, neural networks and penalized splines (Griggs, 2013), or contains basic definitions to lists the diseases that can come from an unhealthy *BMI* (Weir and Jan, 2019); in most cases, the experimental phase is not exhaustive, since there is a lack of comparisons amongst different techniques, and the assumptions behind the approaches are not well detailed.

Starting from these considerations, we aim to introduce a computational approach to predict *BMI* that handles a real case study coming from a publicly available set of data gathered between 2000 and 2001 about 489 women in Belgium, that has already been used to compare obesity and metabolic syndrome prevalence using *BMI* and other indicators (Wong et al., 2021). As for the computational approaches AI-related techniques (Random Forest, Decision Trees and Neural Networks) to test their capabilities to predict the variable of interest, and compared their performances with Linear Regression which, according to the Gauss–Markov theorem (Greene, 2018), guaranties that, under its assumptions, ordinary least squares regression provides the most precise (lowest variance) estimates among all possible linear and unbiased methods. Because of this property, linear regression serves as a theoretical gold standard or benchmark against which other, often more flexible or robust models are compared.

The paper is organized as follows: Section 2 describes the related literature, Section 3 outlines data used in our study, that will be given as input of the techniques outlined in Section 4; Section 5 reports the experimental analysis (whose results will be outlined in Section 5.3), before concluding the paper and discussing further research in Section 6.

2 Main Literature

The influence of an individual's weight on his/her overall health has been investigated from different points of view. In what follows, we are outlining the aspects dealt with by the literature that are relevant for our investigation. *BMI* has been introduced as an indicator to predict health related aspects, i.e., a screening metrics to assess risk for hypertension, diabetes, cancer, hypercholesterolemia, and other chronic diseases (Khanna et al., 2022), especially focusing on children and infants: it has been shown that early complementary food introduction before the age of 4 months is associated with increased *BMI* at the age of 5–7 years (Ha et al., 2023), and this finding led to predict the possibility of child obesity and different health risks. In particular, Khanna et al (Khanna et al., 2022) examine the relationships amongst *BMI*, waist circumference, and waist-hip

ratio measures, focusing on how they may be used to forecast chronic diseases (e.g., diabetes and high blood pressure). Many contributions explore the relationships amongst obesity, weight loss and health-related quality of life (referred to as *HRQoL*); Kolotkin and Andersen (Kolotkin and Andersen, 2017) defines a framework aimed to compare different stages of weight loss and gain related to bariatric surgery, providing evidence of increased *HRQoL* after this kind of therapy, and showing that physical (rather than mental) aspects of *HRQoL* are more correlated with obesity (that is associated to significantly lower *HRQoL* in all samples); Park et al. (Park et al., 2019) investigate the effect of low *BMI* on the development of chronic obstructive pulmonary diseases (referred to as *COPD*) and mortality, detecting that low *BMI* is a significant indicator. Other contributions investigate the correlation between *BMI* and tuberculosis (Choi et al., 2021) or asthma (Vortmann and Eisner, 2008), reporting that there is no apparent combined effect of *BMI* and *DM*[1] on the probability of incident tuberculosis. Additionally, compared to individuals with a healthy *BMI*, obese individuals show a greater likelihood of experiencing regular symptoms of asthma.

Another open research direction is given by using *BMI* for predicting eating disorders: the American Psychiatric Association (*APA*) uses *BMI* to define the severity of anorexia, categorizing it into mild, moderate, severe, or extreme, based on specific ranges(Himmerich and Treasure, 2024). This perspective stresses the need for more holistic approaches beyond *BMI* in understanding eating disorders, especially with respect to different professional categories: Ralph-Nearman et al. (Ralph-Nearman et al., 2020) investigate this aspect in professional female fashion models, and show that professional female fashion models exhibit lower BMI than non-models, but they also show higher fat percentage and muscle mass. This is correlated with a survey-based (alas, not based on clinical observations) eating disorder assessment, in which models reported higher eating and weight concerns than non-models. Please notice that higher BMI was associated with greater eating disorder symptoms in both models and nonmodels, and that models (contrarily to non models) showed a high correlation between muscle mass and eating disorders.

Some contributions recognize the limitations of *BMI* in predicting eating disorders: Ralph-Nearman et al. (Ralph-Nearman et al., 2024) highlights the *BMI*'s inability to accurately reflect health or eating disorders, and express criticism to *BMI* for not considering body fat, muscle mass, or bone density, thus questioning its efficacy in being used for diagnosing conditions like anorexia.

As pointed out by Nuttall (Nuttall, 2015), there are several social standards and expectations to target the ideal weight and its distribution across an individual's body, with respect to gender, age, and cultures: this is reflected in the USA population, that is classified with respect to *BMI*, and shows 63% of individuals that are overweight and 26% that are obese, on top of different changes in the statistics and the appearance of the obesity pandemic. In the same direction, Chambers and Swanson (Chambers and

[1]Diabetes mellitus: a condition characterized by sustained high blood sugar levels, which occurs when the pancreas does not produce enough insulin, or the cells of the body become unresponsive to insulin's effects.

Swanson, 2010) develops a survey that explores the habits of a 18-77 years old aged sample (mostly women), showing that the amount of food consumed and the participation in physical exercise (rather than more advertised predictors such as healthy eating and use of mechanized transportation) are significant predictors of *BMI*. Several metrics have been investigated to classify obesity, including percentage body fat (*PBF*), *BMI*, and fat mass index (*FMI*). In a study by Wong et al. (Wong et al., 2021), *PBF* was found to classify more cases of obesity than both *BMI* and *FMI*, primarily due to its lower predetermined threshold. From the age of 50–59 onward, the differences between *BMI* and *PBF* classifications become more pronounced, due to age-related loss of lean mass. In contrast, *BMI* showed the lowest sensitivity for the diagnosis of obesity because it does not distinguish between fat and lean mass: this limitation led to *BMI* underclassifying some people with obesity into the category *overweight* compared to *FMI*. However, despite these classification differences, there was no significant variation in the prevalence of metabolic risk factors between individuals classified as *obese* by either *BMI* or *FMI*, suggesting that the distribution of body fat may play a more significant role in metabolic disregulation.

The influence of socioeconomic individual's and communities' characteristics are examined by Rundle et al. (Rundle et al., 2008), on a sample of 13,102 inhabitants from New York City, concluding that that local setting might affect how socioeconomic position and body size relate to individuals, especially women. Furthermore, as outlined by Chambers and Swanson (Chambers and Swanson, 2010), women are more likely to report eating in response to more harmful societal pressures and, generally, for reasons other than hunger, hence being more likely to engage in weight control behavior. In contrast, men tend to report eating less healthy, but the quality of their chosen food increases over years; men report higher alcohol use than females, and exhibit higher early maturation factor, which is a condition characterized by the early onset of puberty: it includes indicators such as heavier birth weight and being taller than peers at a young age (e.g., age 7 years). Early maturation can have various implications for physical and psychological development. For example, research has shown that early maturation is associated with an increased risk of obesity and that children who mature early are more likely to have higher *BMI* and *PBF* compared to their peers who mature later: This relationship is due to the early onset of puberty leading to changes in body composition and metabolism that favor fat accumulation. In addition, early maturation can have significant psychological and social effects. Adolescents who mature early can experience higher levels of stress and are more likely to engage in risky behaviors, including higher alcohol use, as noted in the current study. Social pressures and expectations placed on early-maturing children can contribute to these outcomes. Last, the underweight and healthy weight groups show better sleep than the other groups, although there is an increase in weight in children, and the sleep schedules are getting shorter. It is worthwhile to note, as a last finding, that Kamycheva et al. (Kamycheva et al., 2003) reports a strong negative correlation between vitamin D consumption and *BMI* in Northern Norway: According to this research, increasing vitamin D intake to the recommended levels may be able to lower population rates *BMI* and obesity.

As we have stated in the Introduction, contributions aimed to successfully predict *BMI* are still few and far between, and amongst the different proposed approaches, Heslehurst et al. (Heslehurst et al., 2019) show that mothers who are obese before having children have a 264% higher chance of having obese children compared to mothers with a normal weight before conception: this provides evidence for introducing pre-conception therapies to help future women maintain their weight in order to prevent intergenerational obesity. Baltasar et al. (Baltasar et al., 2011) aim to forecast the *BMI* of bariatric surgery patients three years following their operation, based on the assumption that *BMI* values are good indicators for evaluating and contrasting the outcomes of bariatric treatments. This align with the study's objective to create a mathematical formula to predict the final *BMI* (*FBMI*) based on the initial *BMI* (*IBMI*) for better comparison of bariatric surgery outcomes. In the same direction, Guo et al. (Guo et al., 2016) assess the relationship between *BMI* and risk of breast cancer frequency: they developed a weighted *BMI* genetic score made up of 84 *BMI*-related genetic variations, contradicting earlier studies (García-Estévez et al., 2021) which were reporting that genetically predicted *BMI* is inversely linked with breast cancer risk.

It is worthwhile to mention that, according to He et al. (He et al., 2020), males have higher levels of body admiration than females, and that both boys' and girls' body dissatisfaction are positively correlated with age. Interestingly, white women are more likely than women of other races to be unhappy with their bodies. However, the relationship between body appreciation and *BMI* does not show a significant difference between males and females. The meta-analysis indicates that body admiration and *BMI* are inversely related, but this relationship is not specifically stronger in males than in females.

The issue of mortality is a key topic of research, and several contributions investigate the relation between *BMI* and mortality in different scenarios. Bhaskaran et al. (Bhaskaran et al., 2018) examines the associations between *BMI* and all-cause mortality on a sample of 3,632,674 individuals, reporting that lower *BMI* was linked to an increased mortality risk for mental, behavioral, neurological, and external causes of death. *BMI* exhibited J-shaped correlations with overall mortality and the majority of particular causes of death. In the same direction, Flegal et al. (Flegal et al., 2013) explore the relationship between *BMI* and mortality and found that both underweight and severe obesity are associated with increased risk of mortality. However, they noted that overweight individuals had a lower mortality risk compared to normal-weight individuals, suggesting a potential protective effect of being slightly overweight. Another study by the Global *BMI* Mortality Collaboration involving 10.6 million participants from 239 prospective studies worldwide showed that both low and high *BMI* are associated with increased mortality. The lowest mortality risk was observed at a *BMI* of 20-25 kg/m^2. This study emphasized the importance of maintaining a healthy weight within this range to minimize the risk of premature death (Di Angelantonio et al., 2016).

These studies collectively underscore the nuanced relationship between *BMI* and mortality, illustrating that both low and high *BMI* values are linked to increased mortality risks, whereas maintaining a *BMI* within the 20-25 kg/m^2 range appears to be optimal for reducing the risk of various causes of death.

3 Data

The set of data used in this study consists of data gathered between 2000 and 2001 from 489 randomly chosen women from the electoral roll of Belgium. This publicly available dataset has been used by Wong et al. (Wong et al., 2021) to compare obesity and metabolic syndrome prevalence using *FMI*, *BMI*, and percentage body fat (*PBF*): the three main metrics analyzed are *BMI*, *FMI*, and *PBF*, and they were used to assess the prevalence of obesity and metabolic syndrome among individuals. Personal metrics such as height, weight, and body composition, were also gathered from the individuals to make computations, along with Dual-energy X-ray absorptiometry (*DXA*) body composition measurements and clinical data. The statistical analysis was performed using MedCalc[2], which is a statistical software package designed for biomedical research. The significance level for the analysis was set at 0.05 ($two - tailed$), indicating that results with a $p - value$ less than 0.05 were considered statistically significant. The statistical analysis included descriptive statistics, correlations, and comparisons between groups to determine the relationship between body composition metrics and the prevalence of metabolic conditions (Wong et al., 2021). For each woman, the dataset contains measurements of height, *BMI*, *FMI*, *PBF*, lean and fat mass of various body parts, as well as the occurrence of diseases such as diabetes mellitus (*DM*), Hypertension, dyslipidemia, and metabolic syndrome. Our set of data contains the following variables:

- *BMI* (Body Mass Index), whose definition has been provided in the Introduction.
- *FMI* (Fat Mass Index), which measures the amount of fat mass in relation to a person's height. It's a more specific measure of body composition compared to *BMI*, and is computed as:

$$FMI = \frac{FatMass(kg)}{Height^2(m^2)}$$

- *PBF* (Percent Body Fat), that represents the proportion of a person's weight that comes from fat tissue. It is defined as:

$$PBF = \frac{FatMass(kg)}{TotalBodyWeight(kg)}100$$

- *ALMI* (Appendicular Lean Mass Index), that measures the quantity of lean mass in the arms and legs in relation to a person's height. It can be computed as:

$$ALMI = \frac{AppendicularLeanMass(kg)}{Height^2(m^2)}$$

- *Height* (cm): The standing height of the individual measured in centimeters.
- *DM* (Diabetes Mellitus): A binary variable indicating the presence (1) or absence (0) of diabetes mellitus.

[2]Available at `https://www.medcalc.org`, accessed on Februry 26th, 2025.

- *Hypertension*: A binary variable indicating the presence (1) or absence (0) of hypertension.
- *Dyslipidemia*: A binary variable indicating the presence (1) or absence (0) of dyslipidemia.
- *Metabolic Syndrome*: A binary variable indicating the presence (1) or absence (0) of metabolic syndrome.
- *Age*: The age of the individual.
- *WC* (Waist Circumference): The circumference of the individual's waist.
- *TB BMC* (Total Body Bone Mineral Content): The total amount of bone mineral content in the body.
- *TBM* (Total Body Mass): The total mass of the body.
- *TB ST*(Total Body Soft Tissue): The total mass of soft tissue in the body.
- *TB FAT* (Total Body Fat Mass): The total mass of fat tissue in the body.
- *TB LEAN* (Total Body Lean Mass): The total mass of lean tissue in the body.
- *ARMS BMC* (Arms Bone Mineral Content): The bone mineral content in the arms.
- *ARMS ST* (Arms Soft Tissue): The mass of soft tissue in the arms.
- *ARMS FAT* (Arms Fat Mass): The mass of fat tissue in the arms.
- *ARMS LEAN* (Arms Lean Mass): The mass of lean tissue in the arms.
- *LEGS BMC* (Legs Bone Mineral Content): The bone mineral content in the legs.
- *LEGS ST* (Legs Soft Tissue): The mass of soft tissue in the legs.
- *LEGS FAT* (Legs Fat Mass): The mass of fat tissue in the legs.
- *LEGS LEAN*(Legs Lean Mass): The mass of lean tissue in the legs.
- *TRUNK BMC* (Trunk Bone Mineral Content): The bone mineral content in the trunk.
- *TRUNK ST* (Trunk Soft Tissue): The mass of soft tissue in the trunk.
- *TRUNK FAT* (Trunk Fat Mass): The mass of fat tissue in the trunk.
- *TRUNK LEAN* (Trunk Lean Mass): The mass of lean tissue in the trunk.

Table 2 reports the main statistics for all variables considered. Please notice that we have run preliminary experiments in order to assess collinearity and to reduce the input size by using Recursive Feature Elimination (RFE) (Guyon et al., 2002), which is a feature selection technique that provides interpretability while mitigating overfitting and reducing data noise: it first fits a model on the entire set of predictors (and assigns each feature an importance score); subsequently, it iteratively discards the least significant features according to these rankings, retraining the model at each step, until a predefined number of features remains (Brzezinski, 2020; Naseriparsa et al., 2014). The final model is thus trained exclusively on the subset of features deemed most relevant to the prediction task. In our case, we have compared our preliminary experiments (i.e., experiments of techniques introduced in what follows coupled with REF) with experiments without REF through a Wilcoxon signed rank test (tested against a 0.05 significance level), and this led us to reject the null hypothesis of difference between the two approaches. Hence, we will report results without REF in what follows, also because the introduced computational approaches (but the linear approach) do not reply on any underlying model.

4 Computational approach

In this section we are giving a brief overview of the different tools used to perform our prediction: linear regression will be outlined in Section 4.1, decision trees will be detailed in Section 4.2, random forest in Section 4.3, and eventually, neural networks in Section 4.4.

4.1 Linear Regression

Linear regression models aim to find the link between a set of predictors and one dependent variable by identifying the best-fitting line on the predictors dependent variables space, by minimizing the Residual Sum of Squares (*RSS*) (Gupta and Nagalakshmi, 2019): they rely on the assumptions of linearity, independence, homoscedasticity, normality, and non-multicollinearity. They still represent a widespread tool for the prediction exercise, though they suffer from assuming a model behind data. Amongst their clear advantages we found to be the fact that they are able to assess the significance of the different predictors w.r.t. the prediction exercise, and this will be useful in what follows. Alas, often the assumption of linearity is hard to be reflected on data at hand, and that is why the next introduced models do not need any assumptions on the underlying data.

4.2 Decision Trees

Decision trees (Rokach and Maimon, 2005) are flexible models for tasks involving regression and classification, and rely on the use of decision rules use to discriminate the output space: their hierarchical tree structure consists of root, internal, and leaf nodes, and their iterate action of splitting the data according to the decision rule taken into

account allow the end user to understand principles that control predictions. They demonstrate flexibility in handling several types of data, including missing values, without requiring the expansion of features. Anyhow, they suffer of some drawbacks, e.g., they can be computationally expensive during the training phase, particularly for large or complex structures. Furthermore, they are prone to overfitting, especially when dealing with complex datasets, which can lead to excessive variance and impair their capacity to generalize to new data. The next method attempts to mitigate these shortcomings by pooling a number of decision trees together.

4.3 Random Forest

Random Forest is a supervised learning method that relies on the bagging principle, meaning that a collection of models is trained on various dataset subsets, and the result is produced by combining the results of each individual model. The algorithm is made of decision trees (of which it represents a generalization), directly exploiting their capability of handling missing values and making optimal decisions (Rigatti, 2017).

The Random Forest algorithm can be summarized as follows:

Step 1: Each decision tree in the Random Forest model is built using a subset of features and a subset of data points. To put it simply, m features and n random records are selected from a data collection containing k records. This ensures diversity among the trees in the forest and reduces the risk of overfitting.

Step 2: For every sample, a separate decision tree is built. This is done by recursively splitting the data based on feature values, aiming to group similar data points together.

Step 3: An output will be produced by each decision tree. The purpose here is to leverage the diversity of the forest, since each tree has been trained on slightly different data with unique features, their combined outputs can cover a broader spectrum of the input space more accurately than any single tree could.

Step 4: The final product involves combining the outputs of all individual trees to get the final prediction, and it is evaluated using either regression or classification-based majority voting or averaging. For classification tasks, this is typically done through majority voting where the predicted class is the one that a majority of the trees have agreed upon. For regression tasks, the final prediction is usually the average of all the tree outputs (Rigatti, 2017).

Please note that the singular decision trees may be trained independently by one another, hence the training process can be completed quickly.

4.4 Neural Network

Neural Networks (Wu and Feng, 2018) are black-box based algorithm that mimic the behavior of the human brain to solve complex problem, and are composed of neurons

and synapses, that can be organized according to different topologies, amongst which the most widespread architecture is (still) the feed-forward, in which neurons are organized in layers: this network is often trained by using Back- Propagation, which is an algorithm based on the gradient descent to change the synapses weights to perform learning.

All the afore mentioned algorithms have several parameters to be set: this represents a key topic in experimental algorithmics, since the algorithms behaviors may be affected by these parameters. Many approaches have been proposed by the literature to set the parameter values (referred to as parameter setting), and these can be classified in parameter tuning, in which the parameter values are determined before the run, and parameter control, in which the parameter values are dynamically set during the algorithm's run. In this contribution, we have resorted to the Iterated Race method (López-Ibáñez et al., 2016), which is a generalization of F-Race method for parameter tuning of optimization algorithms (Birattari et al., 2010), that is, the tuning of their parameters by finding the most appropriate setting given a set of instances of an optimization problem. To this extent we have used the on- line available implementation (in Python) by M. Lopez-Ibanez and L. Perez Caceres that can be found at the link https://mlopez-ibanez.github.io/irace/. All tools outlined in this section have been implemented in Python.

5 Experimental Analysis

In this section we are outlining our experimental analysis: first we are defining the main building blocks of our approach, namely data pre-processing operations in Section 5.1, the guidelines to partition our data amongst training and test set in Section 5.2, and the results of the experiments performed (by using the tools outlined in Section 4) in Section 5.3.

5.1 Data Pre-Processing

Data pre-processing is important in every experimental framework, to preserve the most information as possible and to feed the computational tool with the most appropriate input data. We have followed the guidelines reported in Angelini et al. (Angelini et al., 2008) and Corazza et al. (Corazza et al., 2021), and performed the following tasks:

- *Removal and replacement.* Detection of missing and wrong values and their handling should always be performed as the first task of a sound experimental analysis. In our case, the set of data at hand is already clean, since it is already being used by Wong et al. (Wong et al., 2021). Hence, differently from Angelini et al. (Angelini et al., 2008) and Corazza et al. (Corazza et al., 2021), we have neither removed all entries with missing data, nor replaced the missing values by meaningful values, since data is robust w.r.t. this aspect.

- *Normalization.* Data normalization is desirable for neural network in order to back-propagate errors belonging to the same interval over the different variables, whilst

not strictly necessary for all other tools identified in Section 4. Hence, we have normalized data only to feed the neural network, by using the logarithmic normalization used by di Tollo et al. (Di Tollo et al., 2015)(Di Tollo et al., 2012)(di Tollo et al., 2014), in which the formula used is the following:

$$\bar{x}_i = \log_u(x_i + 1) \tag{1}$$

where $\bar{x}_i$ is the post-normalised value, x_i the pre-normalised one and, to ensure $x_i \in [0, 1]$], u has been set to $x_{max} + 1$.

- *Correlation analysis.* Correlation analysis is performed to better understand data at hand, to avoid considering predictors that are too correlated, and to avoid collinearity issues. We have performed Pearson correlation (Cohen et al., 2009) analysis since our biggest concern has been devoted to linear regression. Spearman (Hauke and Kossowski, 2011)(i.e., rank-based) correlation analysis led to comparable results. Figure 1 reports the Pearson Correlations found over all possible pairs of variables.

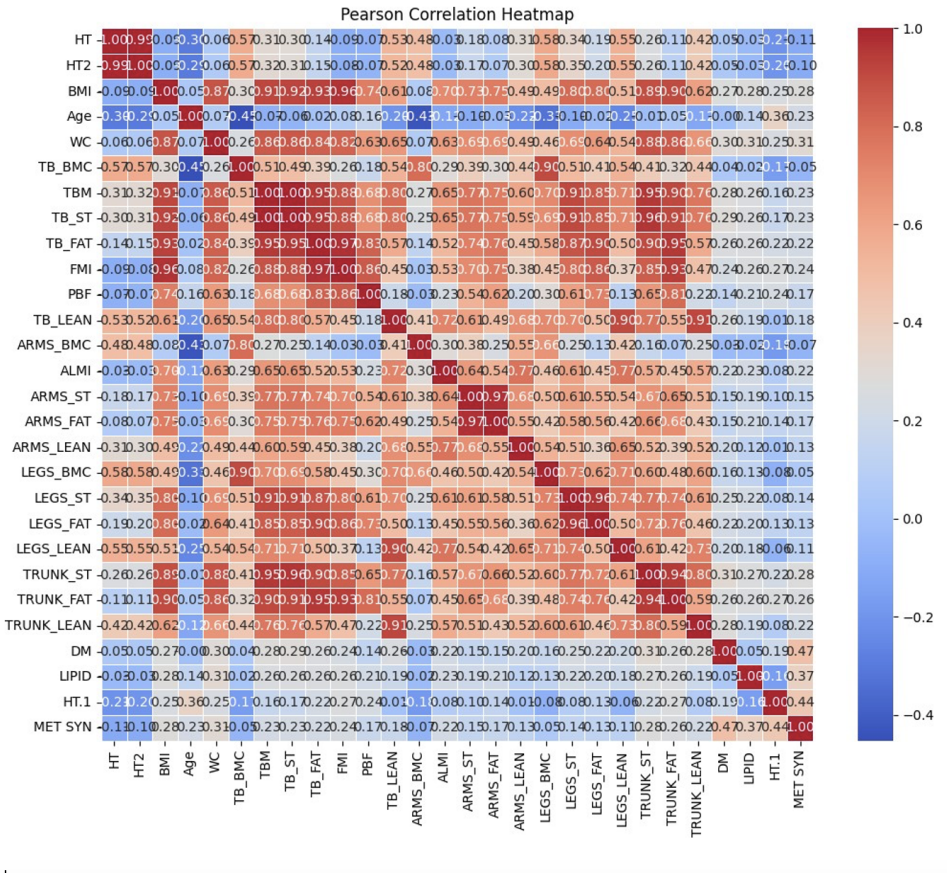

Figure 1: Pearson correlation matrix amongst all pairs of variables in data at hand

Amongst the observations to be reported, a strong positive correlation between *TBM* and Total Body Soft Tissue Mass (*TB ST*) (0.97) has been remarked, indicating a solid linear link that may affect both total bone mineral content and lean mass. Other high positive correlations have been found over the following pairs of variables: Legs Bone Mineral Content (*LEGS BMC*) and *TB BMC* (0.9), Legs Soft Tissue Mass (*LEGS ST*) and Total Body Bone Mineral Content (*TB ST*) (0.91), Total Body Lean Mass (*TB LEAN*) and Legs Body Lean Mass (*LEGS LEAN*) (0.91): this suggests that these variables may be related as a result of physical activity, the development of muscles, or overall body composition. Furthermore, a moderate negative connection between *Age* and *TB BMC* (0.45) was also found, which may have been caused by aging-related changes in bone density or overall health. *TB ST*, *TB FAT*, and *FMI* showed strong positive relations with *BMI* (0.92, 0.93, and 0.96 respectively), which is likely due to the effect of both lean mass and fat mass on *BMI*. Additionally, a moderate positive correlation (0.47) was observed between Metabolic Syndrome (*MET SYN*) and Diabetes Mellitus, suggesting a possible link that might be attributed to shared metabolic or lifestyle variables affecting both conditions.

Based on the correlation analysis, we have removed variables that appear in at least a couple of variables showing correlation greater that 0.8. The variables which were removed are the following: *TB FAT*, *TBM*, *TB ST*, *FMI*, *TRUNK ST*, *TRUNK FAT*.

5.2 Training and Test Set

We have sampled two disjoint sets of individuals (i.e., the training set for learning, and the test set to assess the results) out of the total number of observations, by randomly allocating individuals to sets, with the only constraint that the training set has to contain 70% of the overall set, and the test set 30%. This has been done 30 times to produce 30 different partitions to validate our approaches over different sets of data and test their robustness, as we will see in Section 5.3.

5.3 Results

As a first analysis, we want to report the results of the linear regression model, and the scope of its application is twofold: first, it has been used to predict *BMI*; then, it is useful to assess the significance of variables in the predictor set by analyzing their $p-values$, taking in mind that $p-values$ lower than 0.05 indicates that it is possible to reject the null hypothesis that the predictor is not significant). The variables *WC*, *TB BMC*, *PBF*, *TB LEAN*, *ARMS BMC*, *ALMI*, *ARMS ST*, *ARMS FAT*, *ARMS LEAN*, *LEGS BMC*, *LEGS ST*, *LEGS FAT* , *LEGS LEAN*, *TRUNK LEAN*, *DM*, *LIPID*, *HT*, *METSY N* have $p-values$ smaller than 0.01, thus suggesting a high significance of many predictors and suggesting a great level of complexity, hence justifying the application of more complex heuristics. We can remark that *Age* shows a $p-value$ of about 0.31, appearing to have little significance in explaining the variance in *BMI* . Waist Circumference (*WC*) and

indicators such as *TB BMC*, *PBF*, *TB LEAN*, *ARMS ST*, *ARMS FAT*, *LEGS ST*, *LEGS FAT*, *LEGS LEAN*, *TRUNK LEAN*, and *ALMI* show extremely low $p-values$ (< 0.01), indicating their strong significance in relation to *BMI*. After assessing the significance of the predictors, we report the results from the application of the different tools introduced in Section 4: Table 3 reports the statistics of R^2 found, for each tool, over the thirty different partitions defined in Section 5.2. Please note that only the statistics over the test sets are reported.

It is interesting to note that the linear model alone is capable of producing good results, also at the cost of a really low standard deviation. The standard deviation is also kept low by random forests and decision trees, even though the overall performances of these two methods show an average behavior that is worse than the linear model. Neural networks, instead, are able to reach the best performance obtained over the database at hand, but alas, their performances are affected by a high standard deviation. However, given the low computational time required by all these methods, we may conclude that they can be used jointly.

6 Conclusion

Population health studies are one of the key elements for healthcare, economics, and demography as they are capable of influencing public policies, prevention strategies, and resource allocation. For these reasons, there is a growing need to analyze Big Data in the healthcare field with methodological approaches that allow their analysis. This work addressed one of the topics that has shown increasing attention in the literature: the body mass index. Although its importance is noted for academics and practitioners, computational approaches to model, predict, or forecast *BMI* are still few and far between; moreover, in most cases, the experimental phase is not exhaustive, as there is a lack of comparisons between different techniques and the assumptions underlying the approaches are not well detailed. Our contribution proposed a computational approach to model and predict *BMI*, proposing a comparative analysis amongst different forecasting techniques. The analysis was conducted on a real case by analyzing in particular a publicly available dataset collected between 2000 and 2001 in Belgium to compare the prevalence of obesity and metabolic syndrome using body fat index (FMI), body mass index (BMI) and body fat percentage (BFP). Both traditional methods (linear regression) and AI-related techniques (random forest, decision trees and neural networks) are used, both associated with significant preprocessing. Our results show that a linear trend is visible on the data and that linear models perform well. We note that AI-based approaches also perform well, but require computational time for learning, thus affecting their performance when a maximum time constraint is imposed.

We note that this work highlights the importance of integrating traditional approaches with more advanced AI-based techniques to improve predictive accuracy and decision support in healthcare settings. As technological advances continue to evolve, integrating sophisticated computational models into healthcare decision-making will be essential to address complex epidemiological challenges, formulate public policies to address certain

economic and demographic needs.

Future research will be devoted to the development of an algorithmic portfolio selection method to select the algorithm to be used based on the value assumed by some hyperparameter during execution. In particular, we want to design a method that can promote diversification or intensification depending on the needs of the search and switch between these two paradigms by adaptively modifying its behavior based on the information found during the search. Finally, we want to assess the robustness of the approach over other sets of data, with respect to the same population over time and with respect to a different population.

Declarations

- Funding: This research received no external funding.
- Conflict of interest/Competing interests: None of the authors have a conflict of interest to disclose.
- Ethics approval: Not applicable
- Consent to participate: Not applicable
- Consent for publication: yes
- Availability of data and materials: Data from literature and 3rd party.
- Code availability: Not applicable
- Authors' contributions: the authors contributed equally to this work

References

Angelini, E., Di Tollo, G., and Roli, A. (2008). A neural network approach for credit risk evaluation. *The quarterly review of economics and finance*, 48(4):733–755.

Baltasar, A., Perez, N., Serra, C., Bou, R., Bengochea, M., and Borrás, F. (2011). Weight loss reporting: predicted body mass index after bariatric surgery. *Obesity surgery*, 21(3):367–372.

Bhaskaran, K., dos Santos-Silva, I., Leon, D. A., Douglas, I. J., and Smeeth, L. (2018). Association of bmi with overall and cause-specific mortality: a population-based cohort study of 3· 6 million adults in the uk. *The lancet Diabetes & endocrinology*, 6(12):944–953.

Birattari, M., Yuan, Z., Balaprakash, P., and Stützle, T. (2010). F-race and iterated f-race: An overview. *Experimental methods for the analysis of optimization algorithms*, pages 311–336.

Brzezinski, D. (2020). Fibonacci and k-subsecting recursive feature elimination. *arXiv preprint arXiv:2007.14920*.

Chambers, J. A. and Swanson, V. (2010). A health assessment tool for multiple risk factors for obesity: age and sex differences in the prediction of body mass index. *British journal of nutrition*, 104(2):298–307.

Choi, H., Yoo, J. E., Han, K., Choi, W., Rhee, S. Y., Lee, H., and Shin, D. W. (2021). Body mass index, diabetes, and risk of tuberculosis: a retrospective cohort study. *Frontiers in Nutrition*, 8:739766.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Corazza, M., De March, D., and Di Tollo, G. (2021). Design of adaptive elman networks for credit risk assessment. *Quantitative Finance*, 21(2):323–340.

Dall, T. M., Sapra, T., Natale, Z., Livingston, T., and Chen, F. (2024). Assessing the economic impact of obesity and overweight on employers: identifying opportunities to improve work force health and well-being. *Nutrition & Diabetes*, 14(1):1–8.

Di Angelantonio, E., Bhupathiraju, S. N., Wormser, D., Gao, P., Kaptoge, S., De Gonzalez, A. B., Cairns, B. J., Huxley, R., Jackson, C. L., Joshy, G., et al. (2016). Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *The lancet*, 388(10046):776–786.

Di Tollo, G., Tanev, S., Liotta, G., and De March, D. (2015). Using online textual data, principal component analysis and artificial neural networks to study business and innovation practices in technology-driven firms. *Computers in Industry*, 74:16–28.

Di Tollo, G., Tanev, S., Ma, Z., et al. (2012). Neural networks to model the innovativeness perception of co-creative firms. *Expert Systems with Applications*, 39(16):12719–12726.

di Tollo, G., Tanev, S., Slim, K. M., and De March, D. (2014). Determining the relationship between co-creation and innovation by neural networks. *Complexity in Economics: Cutting Edge Research*, pages 49–62.

Finkelstein, E. A., daCosta DiBonaventura, M., Burgess, S. M., Hale, B. C., et al. (2010). The costs of obesity in the workplace. *Journal of Occupational and Environmental Medicine*, 52(10):971–976.

Finkelstein, E. A. and Trogdon, J. G. (2008). Public health interventions for addressing childhood overweight: analysis of the business case. *American Journal of Public Health*, 98(3):411–415.

Flegal, K. M., Kit, B. K., Orpana, H., and Graubard, B. I. (2013). Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *Jama*, 309(1):71–82.

Gamache, R., Kharrazi, H., and Weiner, J. P. (2018). Public and population health informatics: the bridging of big data to benefit communities. *Yearbook of medical informatics*, 27(01):199–206.

García-Estévez, L., Cortés, J., Pérez, S., Calvo, I., Gallegos, I., and Moreno-Bueno,

G. (2021). Obesity and breast cancer: a paradoxical and controversial relationship influenced by menopausal status. *Frontiers in oncology*, 11:705911.

Greene, W. H. (2018). *Econometric Analysis*. Pearson Education, New York, 8th edition. See Section 2.4 for the Gauss–Markov Theorem and properties of the OLS estimator.

Griggs, W. (2013). Penalized spline regression and its applications. *Whitman College Report. Available online at: https://www. whitman. edu/Documents/Academics/Mathematics/Griggs. pdf*.

Guo, Y., Warren Andersen, S., Shu, X.-O., Michailidou, K., Bolla, M. K., Wang, Q., Garcia-Closas, M., Milne, R. L., Schmidt, M. K., Chang-Claude, J., et al. (2016). Genetically predicted body mass index and breast cancer risk: Mendelian randomization analyses of data from 145,000 women of european descent. *PLoS medicine*, 13(8):e1002105.

Gupta, A. and Nagalakshmi, T. (2019). Stock price prediction using linear regression in machine learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(12):1382–1385.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

Ha, E. K., Kim, J. H., Baek, H.-S., Lee, E., Baek, J. H., Shim, S., Kim, Y. H., Moreno, L. A., Lee, S. W., and Han, M. Y. (2023). Association between complementary food introduction before age 4 months and body mass index at age 5–7 years: A retrospective population-based longitudinal cohort study. *Journal of Human Nutrition and Dietetics*, 36(3):787–797.

Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93.

He, J., Sun, S., Lin, Z., and Fan, X. (2020). The association between body appreciation and body mass index among males and females: A meta-analysis. *Body image*, 34:10–26.

Heslehurst, N., Vieira, R., Akhter, Z., Bailey, H., Slack, E., Ngongalah, L., Pemu, A., and Rankin, J. (2019). The association between maternal body mass index and child obesity: A systematic review and meta-analysis. *PLoS medicine*, 16(6):e1002817.

Himmerich, H. and Treasure, J. (2024). Anorexia nervosa: diagnostic, therapeutic, and risk biomarkers in clinical practice. *Trends in Molecular Medicine*, 30(4):350–360. Special issue: Eating disorders.

Kamycheva, E., Joakimsen, R. M., and Jorde, R. (2003). Intakes of calcium and vitamin d predict body mass index in the population of northern norway. *The Journal of nutrition*, 133(1):102–106.

Kelly, I. R., Doytch, N., and Dave, D. (2019). How does body mass index affect economic growth? a comparative analysis of countries by levels of economic development. *Economics & Human Biology*, 34:58–73.

Kent, S., Fusco, F., Gray, A., Jebb, S. A., Cairns, B. J., and Mihaylova, B. (2017).

Body mass index and healthcare costs: a systematic literature review of individual participant data studies. *Obesity Reviews*, 18(8):869–879.

Khanna, D., Peltzer, C., Kahar, P., and Parmar, M. S. (2022). Body mass index (bmi): a screening tool analysis. *Cureus*, 14(2).

Kolotkin, R. L. and Andersen, J. R. (2017). A systematic review of reviews: exploring the relationship between obesity, weight loss and health-related quality of life. *Clinical obesity*, 7(5):273–289.

Kouris-Blazos, A. and Wahlqvist, M. L. (2007). Health economics of weight management: evidence and cost. *Asia Pacific journal of clinical nutrition*, 16(S1):329–338.

Kuo, M.-H., Sahama, T., Kushniruk, A. W., Borycki, E. M., and Grunwell, D. K. (2014). Health big data analytics: current perspectives, challenges and potential solutions. *International Journal of Big Data Intelligence*, 1(1-2):114–126.

López-Ibáñez, M., Dubois-Lacoste, J., Cáceres, L. P., Birattari, M., and Stützle, T. (2016). The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43–58.

Müller-Riemenschneider, F., Reinhold, T., Berghöfer, A., and Willich, S. N. (2008). Health-economic burden of obesity in europe. *European journal of epidemiology*, 23:499–509.

Naseriparsa, M., Bidgoli, A.-M., and Varaee, T. (2014). A hybrid feature selection method to improve performance of a group of classification algorithms. *arXiv preprint arXiv:1403.2372*.

Nuttall, F. Q. (2015). Body mass index: obesity, bmi, and health: a critical review. *Nutrition today*, 50(3):117–128.

Park, H. J., Cho, J. H., Kim, H. J., Park, J.-Y., Lee, H. S., and Byun, M. K. (2019). The effect of low body mass index on the development of chronic obstructive pulmonary disease and mortality. *Journal of internal medicine*, 286(5):573–582.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*. Bachelier, Paris. One of the earliest works introducing the concept later known as the Body Mass Index (BMI).

Ralph-Nearman, C., Hooper, M. A., Achee, M., Tomarken, A., and Filik, R. (2024). Is there a relationship between body mass index and eating disorder symptoms in professional male fashion models? *American Journal of Men's Health*, 18(5):15579883241279507.

Ralph-Nearman, C., Yeh, H.-w., Khalsa, S. S., Feusner, J. D., and Filik, R. (2020). What is the relationship between body mass index and eating disorder symptomatology in professional female fashion models? *Psychiatry research*, 293:113358.

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, 47(1):31–39.

Rokach, L. and Maimon, O. (2005). Decision trees. *Data mining and knowledge discovery handbook*, pages 165–192.

Rundle, A., Field, S., Park, Y., Freeman, L., Weiss, C. C., and Neckerman, K. (2008). Personal and neighborhood socioeconomic status and indices of neighborhood

walk-ability predict body mass index in new york city. *Social science & medicine*, 67(12):1951–1958.

Swinburn, B. A. (2008). Obesity prevention: the role of policies, laws and regulations. *Australia and New Zealand health policy*, 5(1).

Tanaka, M. (2020). Improving obesity and blood pressure. *Hypertension Research*, 43(2):79–89.

Turnock, B. (2012). *Public health*. Jones & Bartlett Publishers, Burlington, Massachusetts, Stati Uniti.

Vortmann, M. and Eisner, M. D. (2008). Bmi and health status among adults with asthma. *Obesity*, 16(1):146–152.

Wang, Y., Zhao, L., Gao, L., Pan, A., and Xue, H. (2021). Health policy and public health implications of obesity in china. *The lancet Diabetes & endocrinology*, 9(7):446–461.

Weir, C. B. and Jan, A. (2019). Bmi classification percentile and cut off points.

Wells, T. S., Ozminkowski, R. J., Hawkins, K., Bhattarai, G. R., and Armstrong, D. G. (2016). Leveraging big data in population health management. *Big Data Analytics*, 1:1–14.

Wong, J. C., O'Neill, S., Beck, B. R., Forwood, M. R., and Khoo, S. K. (2021). Comparison of obesity and metabolic syndrome prevalence using fat mass index, body mass index and percentage body fat. *Plos one*, 16(1):e0245436.

Wu, Y.-c. and Feng, J.-w. (2018). Development and application of artificial neural network. *Wireless Personal Communications*, 102:1645–1656.

Variable	**Min**	**Max**	**Avg**	**Std.dev**
BMI	15.9	46.1	26.6	4.89
FMI	2.6	27.4	10.9	3.72
PBF	0.1	0.6	0.4	0.08
ALMI	4.5	12.8	6.3	0.78
Height	143	179	162,1	6.35
DM	0	1	0.05	0.22
Hypertension	0	1	0.38	0.36
Dyslipidemia	0	1	0.43	0.50
Metabolic Syndrome	0	1	0.16	0.37
Age	41.4	79.9	60.2	10.88
WC	51.5	131	83.7	11.97
TB BMC	1353.5	3328.1	2329.8	351.98
TBM	43.1	130.4	69.9	13.61
TB ST	41238	127066	67604.9	13427.84
TB FAT	7109.3	70914.7	28646.5	9897.86
TB LEAN	28050.3	61943.4	38954.6	5028.51
ARMS BMC	136.3	458.8	288.8	53.51
ARMS ST	3524	23739	7356.7	2290.97
ARMS FAT	360.6	16126.1	3200.3	1750.82
ARMS LEAN	2622.5	16719.3	4175.5	923.21
LEGS BMC	493.7	1758.2	902.4	162.94
LEGS ST	12601	52579	22789.4	5021.63
LEGS FAT	3383.4	32477.9	10453.9	3925.97
LEGS LEAN	8332.9	20101.1	12335.5	1724.73
TRUNK ST	18183	65372	33728.3	7272.80
TRUNK FAT	2350.1	33241.1	14000.6	5035.31
TRUNK LEAN	1845.8	31859.8	19689.6	3046.40

Table 2: Main Statistics of variables considered.

Model	**Min**	**Max**	**Avg**	**Std.dev**
Linear Regression	0.93	0.98	0.951	0.021
Random Forest	0.82	0.91	0.861	0.024
Decision Trees	0.63	0.80	0.722	0.035
Neural Networks	0.57	1	0.674	0.113

Table 3: Main Statistics of the R^2 reported over 30 iterations of the different methods outlined in the first column.