UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

FINAL DEGREE THESIS

Bachelor's Degree in Biomedical Engineering

MACHINE-LEARNING MODELS FOR OBESE PATIENTS STRATIFICATION

Report and Annex

Author: Nidà Farooq Akhtar

Director: Flavio Palmieri

Co-Director: Pau Gama Pérez

Call: June 2023

Resum

L'obesitat és un problema mundial en constant creixement que augmenta el risc de malalties cròniques i té un impacte significatiu en el sistema sanitari. El tractament i la prevenció de l'obesitat són essencials per reduir aquest impacte en la salut individual i pública. Un factor determinant en els riscos associats amb l'obesitat és la distribució del teixit adipós, també conegut com a greix. Més específicament, el teixit adipós visceral (VAT), que es troba en la cavitat abdominal del cos envoltant òrgans, s'ha assenyalat en diversos estudis com el greix que més riscos associats té.

Aquest treball té com a objectiu avaluar la quantitat de VAT en dones obeses utilitzant algoritmes d'aprenentatge automàtic supervisat. Es treballa amb una base de dades de pacients obesos que conté dades clíniques, resultats d'analítiques de sang i dades antropomètriques (relacionades amb les mesures corporals). El treball es desenvolupa en tres parts: preprocessament de les dades, classificació dels pacients segons la quantitat de VAT i la predicció de la quantitat de VAT fent servir xarxes neuronals.

Mitjançant aquestes tasques i l'ús dels diferents tipus de dades, s'avalua la qualitat de les classificacions i prediccions, obtenint informació rellevant sobre les variables i el seu impacte en l'obesitat. Un resultat positiu en la predicció i classificació seria crucial, ja que permetria la possibilitat de crear una eina econòmica per a l'aproximació inicial dels riscos relacionats amb l'obesitat, especialment en situacions on les eines convencionals per estudiar la distribució del greix no són fàcilment accessibles.

Resumen

La obesidad es un problema mundial en constante crecimiento que aumenta el riesgo de enfermedades crónicas y tiene un impacto significativo en el sistema sanitario. El tratamiento y la prevención de la obesidad son esenciales para reducir este impacto tanto en la salud individual como en los sistemas de salud pública. Un factor determinante en los riesgos asociados con la obesidad es la distribución del tejido adiposo, también conocido como grasa. Específicamente, el tejido adiposo visceral (VAT), que se encuentra en la cavidad abdominal envolviendo órganos, ha sido señalado en varios estudios como la grasa con más riesgos asociados.

Este trabajo tiene como objetivo evaluar la cantidad de VAT en mujeres obesas utilizando algoritmos de aprendizaje automático supervisado. Se trabaja con una base de datos de pacientes obesos que contiene datos clínicos, resultados de análisis de sangre y datos antropométricos (relacionados con las medidas corporales). El trabajo se desarrolla en tres partes: preprocesamiento de los datos, clasificación de los pacientes según la cantidad de VAT y la predicción de la cantidad de VAT usando redes neuronales.

A través de estas tareas y el uso de los diferentes tipos de datos, se evalúa la calidad de las clasificaciones y predicciones, obteniendo información relevante sobre las variables y su impacto en la obesidad. Un resultado positivo en la predicción y clasificación sería crucial, ya que permitiría la posibilidad de crear una herramienta económica para una aproximación inicial de los riesgos relacionados con la obesidad, especialmente en situaciones donde las herramientas convencionales para estudiar la distribución de grasa no son fácilmente accesibles.

Abstract

Obesity is a growing worldwide problem that increases the risk of chronic diseases, significantly impacting healthcare systems. Treatment and prevention of obesity are essential to reduce this impact on individual and public health. One determining factor in the risks associated with obesity is the distribution of adipose tissue, also known as fat. Specifically, visceral adipose tissue (VAT), which is found in the abdominal cavity surrounding organs, has been identified in several studies as the fat with the highest associated risks.

This study aims to evaluate the amount of VAT in obese women using supervised machine learning algorithms. By using a database of obese patients that includes clinical data, blood test results, and anthropometric data (related to body measurements). The study is divided into three distinct parts: data preprocessing, classification of patients based on the amount of VAT, and prediction of VAT quantity using neural networks.

Through these tasks and the use of the different types of data, the quality of classifications and predictions is assessed, obtaining relevant information about the variables and their impact on obesity. A positive result in the classifications and predictions would be crucial as it would allow the possibility of developing a cost-effective tool for the initial assessment of obesity-related risks, particularly in situations where conventional tools for studying fat distribution are not easily accessible.

Acknowledgements

The creation of this project has been a remarkable learning experience, particularly in the field of Machine Learning (ML) and its applications in Biomedical Engineering.

First and foremost, I would like to extend my appreciation to my supervisor, Flavio Palmieri. For offering me the opportunity to participate in this project and for providing me with all the tools, knowledge, and guidance through these months.

I would also like to acknowledge my cotutor, Pau Gama Pérez, for providing me with comprehensive knowledge about the biological basis of this study, and Pablo Miguel Garcia for his supervision and insights about this project.

Furthermore, I extend thanks to the Department of Biophysics of Universitat de Barcelona in Bellvitge and Hospital Clinic for providing me with access to a database which has been the foundation of this thesis.

Lastly, I would like to express my gratitude to family and friends for their unconditional support throughout the completion of this thesis.

Glossary

AT: Adipose Tissue

AUC: Area Under the Roc Curve

BMI: Body Mass Index

DEXA: Dual-Energy X-Ray Absorptiometry

DF: Data Frame

DT: Decision Tree

EDA: Exploratory Data Analysis

KNN: K-Nearest Neighbours

LR: Logistic Regression

ML: Machine Learning

MRI: Magnetic Resonance Imaging

MSE: Mean Squared Error

NB: Naïve Bayes

NN: Neural Network

RF: Random Forest

ROC: Receiver Operating Characteristic Curve

SAT: Subcutaneous Adipose Tissue

SVM: Support Vector Machine

VAT: Visceral Adipose Tissue

Anthropometric data: Data that origins from non-invasive measurements related to the dimensions, size, and proportions of the human body.

Hyperparameter Tuning: Process of finding the optimal set of hyperparameters for a machine learning model.

Index

Index of figures

Index of Tables

1. Introduction

Obesity is an emerging global problem that is often accompanied by various other health complications such as cardiovascular issues, type II diabetes and hypertension, amongst others. As a result of this escalating obesity epidemic, it is vital to identify individuals with higher stakes of suffering the associated pathologies [1]. The profile of obese patients exhibits significant heterogeneity, and consequently, the risk associated with it varies correspondingly. Previous studies have suggested the distribution of fat depots within the body as a key point to correctly assess the risk of obesity-related problems. Giving special importance to Visceral Adipose Tissue, or VAT, as an indicator of such risks [2]. In Figure 1.1. the location of VAT can be seen.

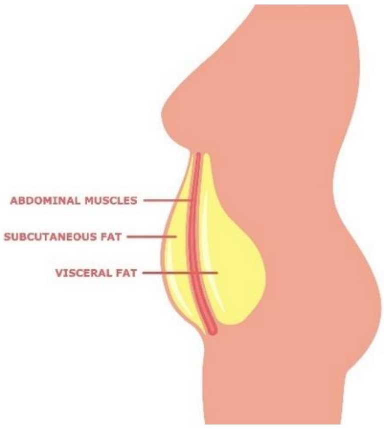


Figure 1.1. Visceral Adipose Tissue (Visceral Fat) location. **(**Source: [3])

Measuring VAT in individuals plays a significant role in assessing their potential risk of developing obesity-related conditions. This step is crucial for identifying and treating these conditions promptly. VAT quantity assessment in the form of classification and prediction can be proven useful in situations where the standard tools to quantify and measure VAT are not available. By developing these models, the project seeks to enhance our comprehension of the intricate relationship between VAT and the variables chosen for the predictive and classification models.

1.1. Origin of the study

The Universitat de Barcelona's Biophysics Research Department, in Bellvitge, and Hospital Clinic de Barcelona have provided a raw database of obese patients containing data of their visceral adipose tissue, mass and volume, and other anthropometric, blood-test-related, and clinical data.

This study originates from the need to evaluate obesity-related risks in an ever-growing obese population globally. To accomplish this, a stratification of patients based on their visceral adipose tissue (VAT) levels can be done to categorize patients based on health risks. Additionally, the impact of the usage of anthropometric and blood-test-related data in the stratification process can be assessed.

1.2. Motivation

The motivation for this project comes from the possibility of having a cost-effective alternative (i.e. machine learning techniques) to assess VAT and therefore obesity-related risks without the need of using imaging techniques such as DEXA scan and MRI. While these imaging techniques are considered the gold standard for assessing VAT, they come with drawbacks such as high costs and the need for well-trained clinicians to operate and provide diagnosis.

1.3. Objective and scope

The objective of this study is to use supervised classification and prediction techniques to estimate VAT in female obese patients. This will be achieved by analysing anthropometric data (i.e. related to body measurements), blood samples and clinical data.

For this purpose, various machine learning (ML) algorithms will be implemented using anthropometric, blood-test and clinical data, and a combination of all of them. By examining these distinct sets of variables, the goal is to analyse the impact of anthropometric and non-anthropometric factors on VAT quantities. The structure of the ML analysis can be seen in the figure 1.2.

The predictions generated by the ML models will offer valuable insights into understanding the relationship between different types of data and VAT in obese patients.

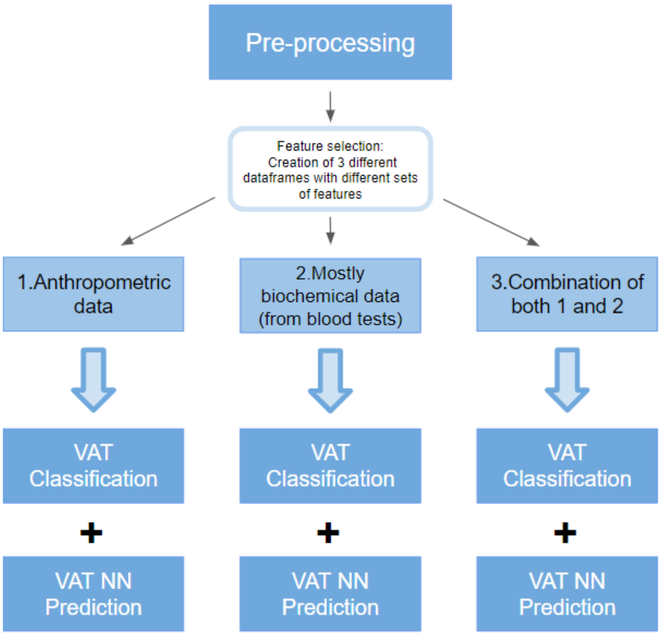

Pre-processing
Feature selection: Creation of 3 different dataframes with different sets of features
1.Anthropometric data
2.Mostly biochemical data (from blood tests)
3.Combination of both 1 and 2
VAT Classification
+
VAT NN Prediction
VAT Classification
+
VAT NN Prediction
VAT Classification
+
VAT NN Prediction

Figure 1.2. Diagram of the ML steps and objectives of the thesis. **(**Source: Own elaboration)

2. Theoretical framework

2.1. Obesity

According to the World Health Organization, obesity is defined as a complex condition that is characterized by an excessive body fat (adipose tissue) accumulation that poses a threat to one's health. Obesity is commonly diagnosed when having a Body Mass Index (BMI), which takes into account body weight and height, over 30kg/m2 [4].

The prevalence of obesity arises from a combination of different factors, such as environmental, genetic and lifestyle. When energy intake surpasses energy expenditure the imbalance results in excess storage of the remaining unused calories in the form of adipose tissue.

This condition was once considered only a high-income country problem [5]. Nevertheless, it has increased globally in the last decades, reaching epidemic-like magnitudes. In some areas like North America, one-third of the population has been reported obese [4]. Moreover, in recent years over 4 million people are reported each year to die as a result of being overweight or obese [6].

The situation in Spain is far from different. As of 2020, 16.5% of adult men and 15.5% of adult women suffered from obesity [7]. Several investigations suggest these numbers won't improve, as it is expected that there will be an increase in obesity if the current trends continue [8][9].

While BMI is a commonly used measure to assess obesity, it may not provide a comprehensive understanding of the health risks associated with it. In this context, a body composition analysis emerges as a more suitable indicator [10].

2.1.1. Complications

Obesity alone poses a significant threat to public health as it not only increases the risk of numerous chronic conditions such as type 2 diabetes, cardiovascular diseases, fatty liver disease, hypertension, certain types of cancers, respiratory problems, and mental health issues but also contributes to a wide array of other conditions. Consequently, individuals affected by obesity experience a decline in both quality of life and life expectancy.

Furthermore, the implications of obesity extend beyond personal health. On one hand, the growing number of obese individuals directly impacts the costs incurred by healthcare systems, placing a substantial burden on resources. On the other hand, obesity has been associated with unemployment and reduced socio-economic productivity, creating additional economic challenges [11]. Both situations contribute to an overall economic burden.

Considering the escalating prevalence of obesity and the complex complications it entails; it is evident that this condition poses a significant threat to public health [12].

2.1.2. Obesity treatment

Due to the complications of obesity, addressing it becomes a necessity. There is not one universal way to treat obesity, as the effectiveness of obesity treatment varies from person to person due to individual factors such as genetics, metabolism, lifestyle, and underlying health conditions.

There are several approaches to treating obesity, with lifestyle changes being the primary method. This involves regulating one's diet and eating habits and increasing physical activity. However, if these lifestyle changes are not successful or cannot be effectively implemented, other options such as medication may be considered. In extreme cases, medical interventions can be performed to address it.

Medical interventions to treat obesity involve making changes to the digestive system to aid the weight loss process and improve the overall excess of adipose tissue. These surgeries are called bariatric surgeries. There are different varieties of bariatric surgery, such as gastric bypass, sleeve gastrectomy, gastric banding, or gastric balloon to name some of the more popular procedures. All of the above use different mechanisms to reduce, or modify in some way, the stomach or the intestines [13].

2.2. Adipose Tissue

Adipose tissue or AT is one of the body’s largest endocrine organs and an active tissue for cellular reactions and metabolic homeostasis. The dysfunctionality of adipose tissue is often associated with pathologies such as diabetes, obesity, cardiovascular disease, and dyslipidemia to name a few [14].

2.2.1. Distribution and functionality

Adipose tissue (AT) is a specialized connective tissue that mainly consists of cells called adipocytes, which are rich in lipids, however, it is important to consider the diversity in cellular components of adipose tissue, particularly in the context of obesity. AT constitutes around 20-25% of the total body weight in healthy individuals. The primary function of adipose tissue is to store energy in the form of lipids.

Adipose tissue distribution is important because it can have profound implications for both health and the development of diseases. The two main types of adipose tissue distribution are subcutaneous adipose tissue (SAT) and visceral adipose tissue (VAT). SAT refers to fat located directly beneath the skin, while VAT refers to fat located around the organs in the abdominal cavity. Both SAT and VAT have different roles. On the one hand, SAT serves to maintain body temperature, store energy and cushion organs, while on the other, VAT is located deep within the body, and it releases various bioactive substances.

The AT distribution, regarding sex, differs between men and women, on one hand, women have a higher proportion of SAT whereas men have a higher percentage of VAT [15].

Adipose tissue, including VAT, is not only an inert fat storage depot but also has an important role in endocrine functions as it secretes cytokines and hormones that influence metabolism, and inflammation and play an important role in physiological homeostasis.

2.2.2. VAT and obesity

Surely to describe obesity as an increase in adipose tissue mass is an oversimplification. The obesity-associated morbidity and mortality in humans is associated with fat accumulation and other factors such as sex, genetics, environment, and the different types of adipose tissue. The variation in fat distribution is a potential explanation for the cardiometabolic risk differences between Individuals with the same BMI. That being the case, VAT is an interesting fat depot concerning the heterogeneity in obesity and has a significant role in metabolic risk.

Many studies have differentiated the risks associated with obesity, and therefore adipose tissue, with the amount of VAT, as people with the same BMI have been shown to have a higher risk of complications if they possess more VAT mass. This concept of variability between obese subjects is further explained in the 2005 study done by Haslam & James to assess treatment for obesity and the need to focus on the remarkable heterogeneity found in metabolic risk between individuals of different body mass index, as some patients with higher BMI but lower VAT mass show less risk than patients with lower BMI but higher VAT mass [16].

Moreover, a study done in the UK with the data of over 40.000 participants demonstrated that a deep learning approach based on two-dimensional MRI projections was adequate to predict and quantify VAT, SAT, and GFAT (Gluteofemoral Adipose Tissue) volumes at scale and suggested that VAT was linked to an increased risk of type 2 diabetes and coronary artery disease, in contrast with the SAT and GFAT which were mostly neutral in the matter [12].

Therefore, understanding the distribution of AT holds significant importance as it offers valuable information about an individual's overall health and disease risk. Specifically, evaluating the distribution and quantity of VAT enables healthcare professionals to identify individuals who may face increased risks of obesity-related complications. Thus, quantifying VAT serves as a great indicator to assess the potential risks associated with obesity.

2.3. Quantification of Visceral Adipose Tissue

2.3.1. Magnetic Imaging Resonance

MRI is a non-invasive cross-sectional tomographic imaging method that accurately assesses body fat distribution and composition as it can generate detailed and segmented images of internal body structures.

Nowadays, MRI is the gold standard for VAT quantification, as it has been proven to provide accurate quantification consistently and efficiently [17]. The main reasons for MRI to be the gold Standard are:

- The high accuracy with which it provides measurements of VAT as it offers soft tissue contrast that allows for differentiation between all types of tissue, VAT included.
- MRI provides imaging in coronal, axial and sagittal planes, which allows for a broad evaluation of VAT distribution in the body.
- It has been validated numerous times through clinical analysis and research studies [18].

2.3.2. Dual-energy X-ray Absorptiometry

Nowadays, DEXA or Dual-energy X-ray absorptiometry is increasingly being used in the healthcare sector to quantify fat due to its minor preparation, relative simplicity, and inexpensiveness compared to other imaging methods [19].

Dual-energy X-ray absorptiometry is an imaging technique that uses 2-dimensional projection data created by low-dose X-rays to create a model that can be used to differentiate between fat, bone, and lean tissue; it is used to measure body composition.

More specifically, it can be used to estimate the quantity of AT compartments in different regions of the human body. Also, by analysing the differential absorption of X-rays by different tissues, DEXA scans can provide estimates of anthropometric measures.

Despite the convenience of DEXA, the gold standard to quantify fat depots is magnetic resonance imaging or MRI, as mentioned in the previous section. However, the specific use of DEXA and MRI for VAT measurement has been shown to be highly correlated as stated in recent studies [19].

The choice between DEXA and MRI depends on the specific use the data obtained from these techniques is going to take. For large-scale studies that may require cost-effective assessments of VAT, DEXA is a preferable option.

3. Machine Learning

Machine learning is a branch of Artificial Intelligence that centres around the use of data and algorithms to imitate the human way of learning, progressively improving accuracy. There are two main types of machine learning depending on whether the value (or values) to be predicted are available or not; it is supervised ML if the expected output of the ML models is known, and if the expected output is unknown then it is unsupervised ML.

Supervised ML algorithms are built using sample data, known as the training set and then the accuracy is tested using a testing data set [20]. These models can be used to make predictions, classifications, or decisions with their corresponding accuracies and other metrics thanks to the train-test data division. A common way to divide the sets is to apply an 80-20 proportion to the train-test split as shown in Figure 3.1. [21]

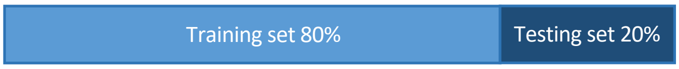


Figure 3.1. Training and testing set proportions visualisation. **(**Source: Own elaboration)

Usually, the values to be predicted are called target data, and the variables that are used to predict and classify can be named features.

In the provided database the expected output data is available so supervised ML algorithms will be applied and, in this section, the steps followed to apply ML for data classifications and predictions will be explained.

3.1. Python

The programming language chosen to compute and create Machine Learning algorithms in this project is Python. Python is an interpreted and high-level programming language that is easily available and has easy-to-learn syntax. Moreover, it has already multiple built-in libraries that come in handy with ML programming. Some of these libraries will be further explained later in the thesis.

3.2. Pre-processing

The first step to any ML project is the pre-processing of the data, i.e., proper cleaning of data, which means having a coherent database without any missing data, with only meaningful variables, standardized values, and other various steps depending on the format and origin of the database.

In this subsection, some of the common pre-processing practices will be mentioned.

3.2.1. Exploratory data analysis

One of the initial steps for the pre-processing is the exploratory data analysis or EDA. This analysis involves the examination and proper understanding of the dataset, including the characteristics and intricate relations between data variables. Also helps identifying missing values or outlier values [22].

Some common tasks done during EDA are getting to know which kind of variables are there in the database, as well as the shape of such database. Data can be classified in two main groups, qualitative and quantitative data, as can be seen in Figure 3.2:

- Quantitative data: This data refers to numerical data, and can be further classified into two groups:
  - Continuous: includes decimals, typically comes from measurements, e.g., height.
  - Discrete: Integers and countable numbers, e.g., nº of siblings.

- Qualitative data: Non-numerical data, typically comes from observations.
  - Nominal: Values without a specific order, e.g., eye colour.
  - Ordinal: Values with a natural ordering, e.g., clothing size.

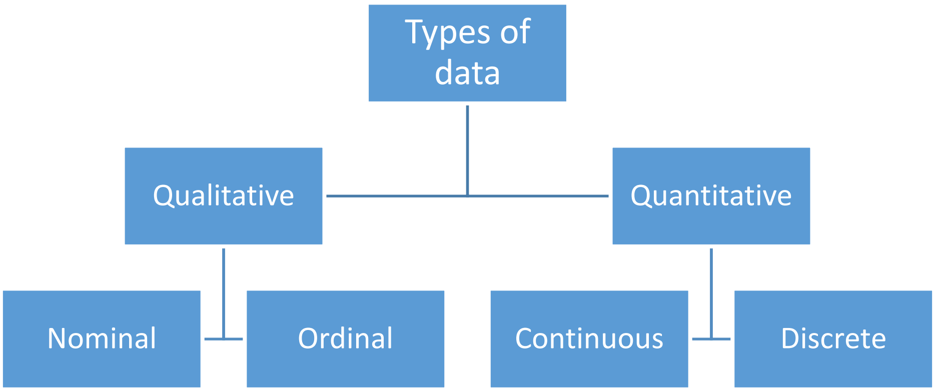


Figure 3.2. Types of data **(**Source: Own elaboration)

An exploratory analysis usually offers data visualization to properly comprehend the distribution and properties of the variables in the dataset, this includes all types of visual representations such as scatter plots, histograms, boxplots, etc.

Furthermore, to assess the posterior feature selection, a heat plot can be provided to better understand the association between variables.

Overall, EDA offers an initial understanding of the dataset to be worked with, it helps detect potential issues and visualise patterns, and therefore, is a key tool for decision making.

3.2.2. Data imputation

Another step is the treatment of missing data. If a raw dataset has missing values or data, data imputation is performed. Missing values can happen for many reasons, either from unavailability of a specific value, human error, etc. Handling missing values is an important part of the pre-processing because many of the ML algorithms require a complete dataset to accurately function.

Data imputation works by filling the missing values based on the existing ones, which can be done in multiple ways as there is not one proper method to do it. Some of the most common imputation methods are the following:

- Mean, mode or median value imputation, in which the missing values are replaced with one of these measures of the variable with missing data.
- Regression imputation, in which a regression model is used to calculate and predict the missing values based on the existing ones. [23]
- K-nearest neighbours, in which the missing value is replaced with the average value of its k-nearest neighbours, taking into consideration the existing data's distance between samples. K is a hyperparameter to be manually chosen. [23]
- Hot deck imputation, in which a randomly selected value is chosen from another similar record without missing data. This technique maintains the distributional characteristics of the dataset. [23].

3.2.3. Outlier data removal

Sometimes DF can have outlier values, which are data points or observations that significantly deviate from most of the data points in a dataset.

The detection and removal of these abnormal values is an important step in ML as they can negatively impact the ML model's performance and accuracy by introducing noise. They can be a source of distortion of the data; however, outliers can also provide valuable insights into unique values of biological importance. By handling outliers appropriately, data quality and reliability can be improved, leading to more reliable models. All in all, the detection and removal of some outlier values is good to enhance the model's performance, but the origin of the data must be taken into consideration, especially in the case of medical data. [24]

3.2.4. Data standardisation

Standardizing the data is an important step because databases can have variables of different origins and measurements, which means that each variable might have a different scale from the rest. For that reason, it is sometimes necessary to apply data normalization o standardization to size all the elements in a dataset to the same scale to properly compare them. Note that not all datasets need standardization, as all the variables from a specific dataset could be from the same measurement, that is why it is necessary to study the variables beforehand to properly determine the need for normalization. Moreover, prediction and classification models require all data to be in a common scale to function correctly. [25]

The standardisation of data should be applied to non-categorical data, as binary and categorical values might lose meaning by standardization.

All in all, the scaling of data helps with the interpretability and the application of some ML algorithms.

3.2.5. Feature selection

Feature selection refers to choosing the subset of the most relevant features from a dataset, this step improves the performance of ML models as non-important data delays or obstacles the functionality of such algorithms.

Feature selection should be done considering both the mathematical association of a variable with the target and the knowledge on the variables and their meaning regarding the output in the specific data's domain or field, for this project, a proper understanding of VAT and the relation between each feature is necessary.

The main techniques used for this project are knowledge driven feature selection, Spearman correlation coefficient between the target values and the features, and Cramer's V for the few categorical variables:

- Correlation coefficient: Correlation coefficients measure the covariance between variables to determine the strength of their relationship, this measure as it is standardized, varies from -1 to 1, -1 being strong negative correlation, 0 being no correlation at all and 1 being strong positive correlation.
- Cramer's V: This algorithm measures the strength of association between categorical variables, it is an extension of the chi-square test, and the measure goes from 0 to 1, 0 being no association at all, and 1 being high association. [26]

To sum it up, feature selection plays a crucial role in ML as it reduces the computational complexity of models and therefore improves the efficiency. It helps identifying the most informative features in a dataset, allowing for a more focused analysis. By understanding which features have the greatest influence on the output, it becomes easier to comprehend the relationship between input and output, and in the specific case of this project, it is of interest to understand which input variables are more relevant in the prediction and classification of VAT. [27]

3.2.6. Label data encoding

For the successful application of a classification model, the label data must be in a categorical format rather than continuous. This means that the data should be organized into distinct categories or classes, allowing the model to classify new instances accurately. Continuous data, on the other hand, represents a range of values and may not be suitable for classification tasks.

Transforming continuous label variables into categorical ones simply implies dividing the data into distinct groups or intervals of said continuous values. For example, VAT mass can be divided into three groups: one containing low VAT mass, another medium VAT mass and the third one containing high VAT mass.

However, for categorical variables to be used effectively in ML algorithms, they need to be encoded in a numerical format. This transformation is necessary as most ML algorithms operate on numerical data. By encoding categorical variables into numerical values, the algorithms can process and analyse the data accurately. When the order of the label data must be preserved, it is important to assign appropriate numerical values to avoid losing valuable ordinal information. Continuing with the previous example, each group should be encoded maintaining the natural order, one way to do it would be assigning the values of 0, 1 and 2 to each group respectively.

All in all, the categorization of data is an important pre-processing step to convert categorical variables into a suitable format for ML classification algorithms. It involves encoding and handling ordinal variables, as this ensures that the categorical data can be effectively utilized in the following modelling stages.

3.3. Classification

3.3.1. Classification models

In this section, an overview of the classification models applied in the ML process will be done.

3.3.1.1. Support Vector Machine

The support Vector Machine algorithm finds an optimal hyperplane that properly separates classes in a high-dimensional feature space. The main objective of SVM is to find the hyperplane that maximizes the margin, which represents the region of maximum separation between classes.[29]

A visual representation of the SVM mechanism is shown in Figure 3.3.

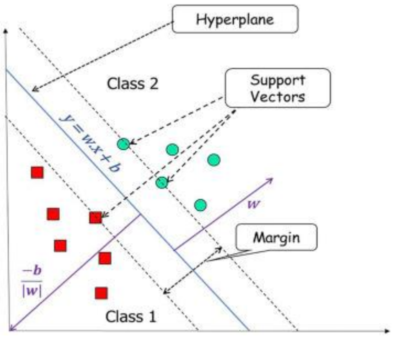


Figure 3.3. Support Vector Machine representation **(**Source: [28])

3.3.1.2. Logistic Regression

Logistic regression algorithms determine the probability of the label data belonging to a certain class by assessing the relationship of the label data with the features.

The probability of a data point belonging to a class is estimated by using a logistic function, also known as a sigmoid function. The logistic function transforms the input into a value between 0 and 1, which represents the probability of the data belonging to the positive class being tested (0 being less probable and 1 being highly probable). This method can be also applied to multiclass classification rather than only binary classification by using various methods, such as the one-vs-rest approach, where an individual logistic regression model is trained for each class and the model with the highest probability is selected as the predicted class.

3.3.1.3. K-nearest neighbours

K-nearest neighbours classifier or KNN, is a simple algorithm that stores all the possible variables and classifies them by their measure of similarity. The nearest neighbours are determined by calculating the distance between the observation to be predicted and each observation in the training dataset. The number of neighbours to consider is k, as it can variate depending on the classification and the distance can be computed using various methods such as Euclidean, Manhattan, Minkowski, etc. [30]

3.3.1.4. Naïve Bayes

Naïve Bayes (NB) classifier is a probability-based algorithm, the Bayes theorem is used with the assumption of feature independence. NB calculates the conditional probability for each class, and it assigns a class based on the highest probability of the data belonging to a specific class. [30]

3.3.1.5. Decision Tree and Random Forest

Decision Tree (DT) classifiers, as their name implies, are tree-like structures designed to make classification decisions. They consist of nodes, branches and leaves, where nodes represent features, branches represent decision rules, and the leaf nodes represent class labels. In a decision tree, the most informative features are selected at each node, and this process continues until a stopping criterion is reached. The stopping criterion, often defined by a maximum depth level of the DT, depends on the dataset, and helps control the complexity of the tree. The final leaf nodes represent the predicted class labels. A visual representation of a DT can be seen in Figure 3.4.

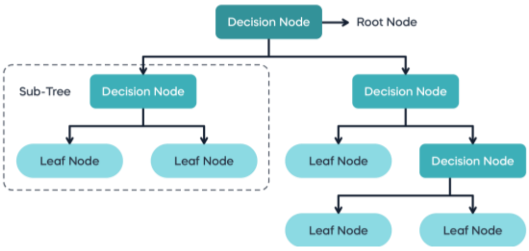


Figure 3.4. Representation of a Decision Tree classifier **(**Source: [31])

If DT becomes too complex, it can lead to overfitting, which means that the model is only tailored for the training data set, and therefore become less effective with other sets of data. One solution to this problem is the Random Forest classifier, which can be considered an extension of DT, as it combines multiple DTs. Each DT of a Random Forest is trained with a random subset of features which helps to reduce overfitting.[32]

3.3.2. Validation

Once the models are trained with the training dataset, a validation process must be applied to determine the correctness of the classifications. The classification validation process for this project includes the following steps:

3.3.2.1. K-cross validation

To ensure that the classification model will properly work with data external to the training set, k-cross validation is applied. This validation divides the input data into k folds or subsets in which one of them is used as the testing data set and the rest is used as the training data set, this process is repeated multiple times with each subset has been served as both the testing set and training set, then average metrics are extracted from all the iterations [33]. K-fold cross-validation provides a more robust estimation of model performance by utilizing the entire dataset and avoiding potential biases caused by a specific train-test split. It addresses the limitations of evaluating models based on a single partitioning of the data. This approach allows for a more reliable assessment of the model's performance, as it considers the variability that may arise from different data partitions.

3.3.2.2. Confusion matrix

The confusion matrix helps with the comprehension of the performance of the classification models as it offers a visual representation of the results of the execution of the classification task. It comes in the form of a, usually, square-shaped matrix in which the number of correct and incorrect predictions is shown in comparison to the actual values. If we take as an example a binary classification, the shape of the confusion matrix could be as the following one in Table 3.1.

	Predicted positive (0)	Predicted negative (1)
Actually positive (0)	True Positive (TP)	False Negative (FN)
Actually negative (1)	False Positive (FP)	True Negative (TN)

Table 3.1. Confusion matrix example **(**Source: Own elaboration)

The size of a confusion matrix is determined by the number of classes we want to predict, so if we want to predict three VAT classes, the matrix would be 3x3, the diagonal will represent the number of true values and the rest of the values represent false values. [34]

3.3.2.3. Receiver Operating Characteristic Curve (ROC) and Area Under the Roc Curve (AUC).

ROC curves are useful tools for comparing and evaluating the performance of classification models. They plot the true positive rate (TPR) against the false positive rate (FPR) by adjusting the decision threshold of the model. A diagonal ROC curve represents the performance of a model that cannot properly discriminate between classes and is essentially random. In contrast, an ideal classifier would be situated in the top-left corner, with a TPR of 1 and an FPR of 0.

ROC curves allow the calculation of the area under the curve or AUC, which is a valuable metric to summarize the overall accuracy of a model. AUC ranges from 0 to 1, where 0 indicates an inaccurate model and 1 represents perfect accuracy. Therefore, it provides an effective measure of the model's ability to discriminate between different classes. Considering that the diagonal ROC indicates the randomness of the classification model, an AUC of 0.5 also suggests no discrimination. [35] Representation of different ROC curves can be seen in Figure 3.5.

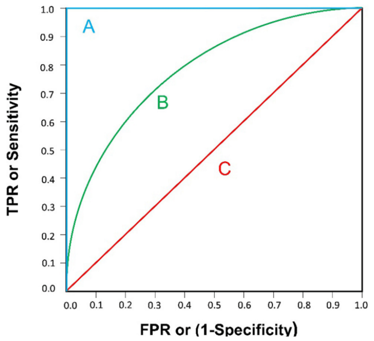


Figure 3.5. Three examples of ROC curve, A (AUC=1) is an ideal classifier, B is a realistic classifier that performs better than a random classifier and C (AUC=0.5) is a random classification (Source: [36])

3.3.2.4. Evaluation metrics

The statistical metrics used to evaluate the performance of the classification models are the accuracy, precision, recall and f-1 score [37], in the following equations, the variables shown are the ones from Table 3.1 :

- Accuracy: It is the metric that calculates the overall correctness of a model by dividing the total amount of correctly computed values by the total amount of computed values.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag*{\textbf{(Eq. 3.1)}}$$

- Precision: Evaluates the performance of a model in correctly assigning positive labels to truly positive instances. It quantifies the model's ability to make accurate positive predictions and avoid false positives.

$$Precision = \frac{TP}{TP + FP} \tag*{\textbf{(Eq. 3.2)}}$$

- Recall or sensitivity: Quantifies the proportion of positive instances correctly identified by a classification model compared to the actual number of positive instances in the data.

  It measures the model's ability to accurately capture and identify positive cases.

$$Recall = \frac{TP}{TP + FN} \tag*{\textbf{(Eq.3.3)}}$$

- F1 score combines the precision and recall metrics in one.

3.4. Neural Networks

Neural Networks (NN) models are machine learning algorithms inspired by the biological neural networks of the brain. These models are designed to learn and identify patterns and relationships within data, enabling them to make accurate predictions and perform various tasks.

The fundamental unit in NN is the neuron or node, which serves as the processor of data: they receive inputs and produce outputs similarly to biological neurons. Each of these inputs can come from another node, or an external source in the case of the first layer of the NN, that receives external features.

The node represents a mathematical operation that produces outputs from inputs. They take in inputs associated with a weight, which represents the significance or importance of that input, and therefore, weights determine the contribution of an input in the calculation of the output. The weights at first are randomly assigned and during the training process are adjusted according to the desired output.

The step size at which the NN algorithm updates the weights during the training is called the learning rate, it controls at which rate the weights are changed.

Then an activation function is applied to the weighted sum of the inputs, to finally generate an output. The output of a node can become an input for other nodes in the network, forming a network of interconnected nodes or layers [38]. The usual architecture of a NN is formed by an input layer, hidden layers, and the output layer, as shown in Figure 3.6.

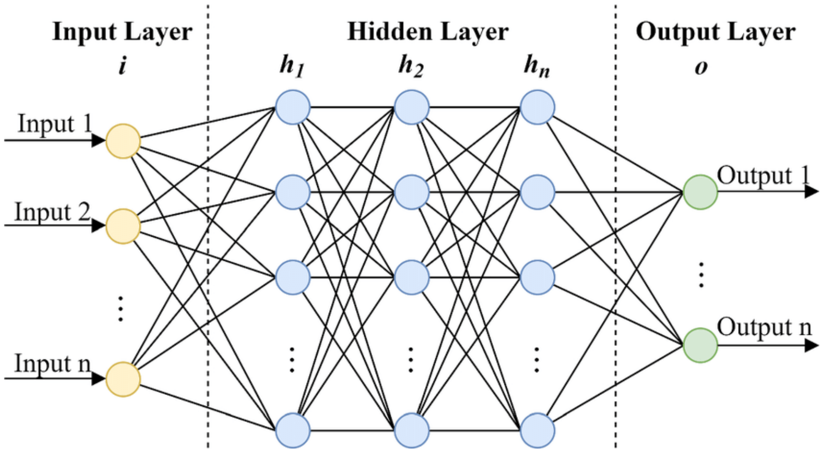


Figure 3.6 Architecture of a NN. **(**Source: [39])

In the context of this project, NN models will be used to perform a regression analysis. Regression analysis establishes functional relationships between independent variables, the features, and a dependent variable, the target variable [40]. Therefore, the NN will take input features to learn to predict a continuous numerical output. To accomplish this prediction, the NN output layer is designed to be a single node or neuron, this enables the production of an unrestricted continuous value as the final prediction.

All in all, NN networks use the combined behaviour of interconnected nodes to analyse and gain insights from data. This ability allows them to tackle intricate problems and provide predictions.

3.4.1. Validation

Once the models are trained with the training dataset, a validation process must be applied to determine the correctness of the predictions. The NN prediction validation process for this project, apart from having a graphic comparison between the real values and the predicted values, includes the following metrics:

3.4.1.1. R-squared and Mean squared error

The coefficient of determination or R-squared is a statistical metric used to assess how well a predictive regressive model approximates to actual data. It can be defined as the fraction of the variance in the dependent variable that can be explained, or is predictable, from the independent variables [41].

$$R^2 = \frac{\sum(y_{pred} - y_{mean})}{\sum(y_{actual} - y_{mean})} \qquad \textbf{(Eq. 3.4)}$$

In the equation 3.4, a simplified definition of the R-squared can be seen, being y-pred the predicted values, y-mean the average of the actual values to be predicted, and y-actual the actual values to be predicted.

R-squared usually ranges from 0 to 1, with 0 indicating the regression not fitting the data at all, and 1 representing a perfect fit. The R-squared metric of most interest will be the testing dataset one because it will provide an estimation of how well the trained model will perform on unseen data.

The Mean Squared Error of the testing data set will also be computed, MSE quantifies the average squared difference between predicted and actual values. It is used to evaluate regression models because it assigns greater importance to larger errors through the squaring process. A lower MSE signifies that the model's predictions closely align with the actual values, indicating superior performance.

For the validation process, k-cross validation (See *4.3.2.1.K-Cross Validation* section) will be used to compute the R-squared, of the training and testing sets, and the MSE.

It's worth noting that there are other evaluation metrics available for regression models depending on the specific problem and requirements, nonetheless R-squared has been suggested to be the most informative in regression analysis. [41]

4. Methodology

In this section, the steps and processes used to classify and predict VAT quantity, as shown in Figure 4.1, will be explained. All the code used for this project can be found in the following link:

https://github.com/Nidafarooqq/TFG/tree/main

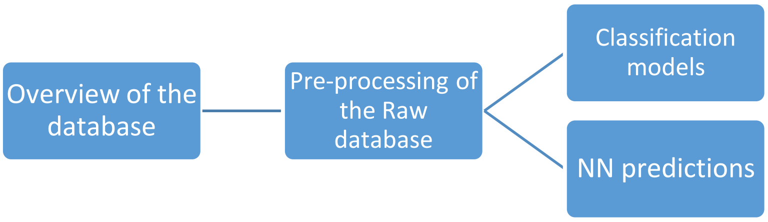


Figure 4.1. General outline of the methodology (Source: Own elaboration)

4.1. Database description

The database to work with is a raw data frame from obese patients with scheduled bariatric surgery, this means these patients are likely to be morbidly obese (i.e. BMI higher than 40kg/m2).

It consists of the data of 169 patients and contains 167 variables or columns (169x167), the variables consist of anthropometric data from a DEXA scan, blood-test-related data and data from medical conditions. The target data are the variables “Masa_VAT_g” and “Vol_VAT_cm^3”, which contain information about the mass and volume of VAT of each patient respectively. Both variables can be used interchangeably as the target data as they have an almost perfect correlation.

The following tables will help to get a better insight into the database:

Nº of female patients	Nº of male patients
156	13

Nº of anthropometric data	Nº of biochemical and non-anthropometric data	Nº of ID-related data
145	20	2

Nº of continuous variables	Nº of categorical variables	Nº of non-numerical variables
162	4	2 (1 is ID)

The data was collected within the framework of an ongoing study and is not published yet. All the data will be made available once the study finishes.

4.2. Pre-processing

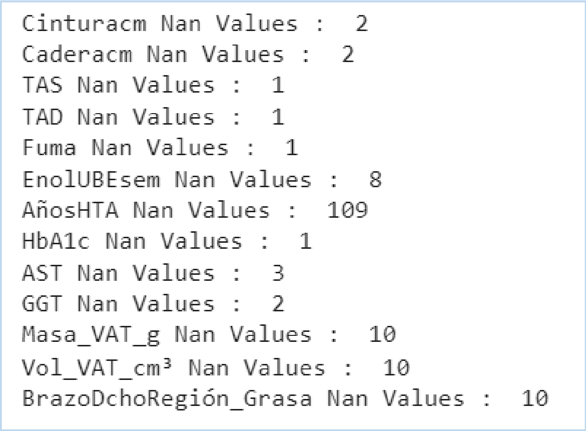
The first step for the pre-processing is importing the necessary Python libraries, here's a summary of the libraries used:

- Pandas
- NumPy
- Seaborn
- Matplotlib
- Scikit-learn
- Dython

4.2.1. Imputation and cleaning of data.

The next step is identifying and correcting any missing data in the data frame as well as the removal of redundant or unnecessary data.

The detection of null data in the data frame consists of a simple Python loop, in which variables that have missing data (Nan) and how many there are, can be obtained, as shown in Figure 4.2.



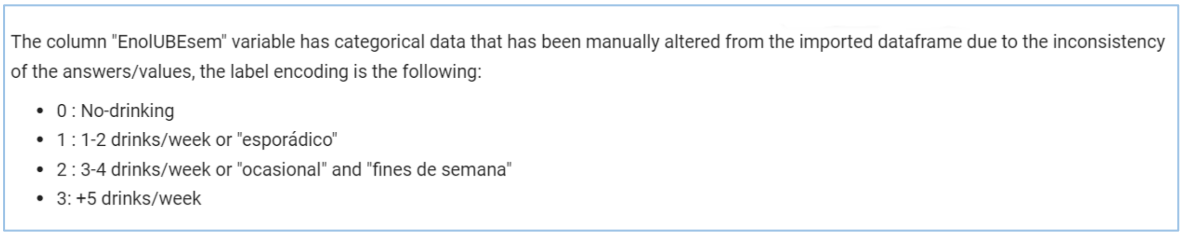
```
Cinturacm Nan Values :  2
Caderacm Nan Values :  2
TAS Nan Values :  1
TAD Nan Values :  1
Fuma Nan Values :  1
EnolUBEsem Nan Values :  8
AñosHTA Nan Values :  109
HbA1c Nan Values :  1
AST Nan Values :  3
GGT Nan Values :  2
Masa_VAT_g Nan Values :  10
Vol_VAT_cm³ Nan Values :  10
BrazoDchoRegión_Grasa Nan Values :  10
```

Figure 4.2. Missing data summary (Source: Own elaboration)

It can be seen see how one of the variables, "AñosHTA", has 109 missing values, which means more than half of the patient's data is null, that being the case the column is to be deleted. The rest of the variables have a relatively low amount of Nan values so imputation techniques can be applied.

Apart from deleting "AñosHTA", the ID-related columns, "Base_origen" and "NHC", do not provide meaningful information for the ML techniques and will also be deleted. Once this data is handled a 169x164 data frame is left.

Before finally imputing data, the only non-numerical variable of the DF (apart from the ID ones) must be encoded. It must be considered that this DF comes from a hospital and some variables have been manually written, that is the case for the "EnolUBEsem" variable, which indicates the alcohol consumption of the patient. The data encoding has followed the next rules, as indicated in the main code shown in Figure 4.3.



The column "EnolUBEsem" variable has categorical data that has been manually altered from the imported dataframe due to the inconsistency of the answers/values, the label encoding is the following:

- 0 : No-drinking
- 1 : 1-2 drinks/week or "esporádico"
- 2 : 3-4 drinks/week or "ocasional" and "fines de semana"
- 3: +5 drinks/week

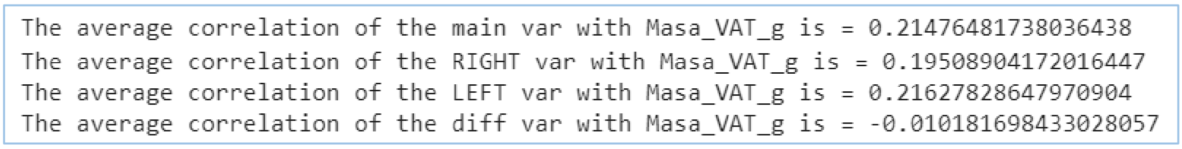
Figure 4.3. Ordinal data encoding of the "EnolUBEsem" variable (Source: Own elaboration)

Once all the DF has numerical values, missing data is imputed using a KNN imputer from the Scikit-learn library and a simple loop to iterate over the missing or Nan values. The result is a DF without any missing data.

The subsequent stage is handling excess redundant data. The anthropometric data in the DF comes from a DEXA scan, which estimates anthropometric values through calculations of differential absorption of X-rays by different tissues in the body. It's important to note that the scan captures data from both the right and left sides of the body. As a result, there is redundant data in the DF, as for each variable, there is data from the right side, the left side, the total value (which is the sum of left and right-side data), and the differential value (which is the difference between left and right side data). An example of this is the next variables from the DF:

- *"BrazosMO_g"*: Total bone mass of the arms in grams
- *"BrazosDcho_MO_g"*: Bone mass of the right arm in grams
- "*BrazosIzq_MO_g"*: Bone mass of the left arm in grams
- *"ArmsDiff_MO_g"*: Difference in bone mass between both arms in grams

To solve this issue of surplus of information, the correlation coefficient between these measurements and the target data has been computed, to decide whether to eliminate some of these columns or not. A similar correlation between the target data and each of these variables would mean that no significant information would be lost when removing excess variables. An example of the correlation results is shown in Figure 4.4.



```
The average correlation of the main var with Masa_VAT_g is = 0.21476481738036438
The average correlation of the RIGHT var with Masa_VAT_g is = 0.19508904172016447
The average correlation of the LEFT var with Masa_VAT_g is = 0.21627828647970904
The average correlation of the diff var with Masa_VAT_g is = -0.010181698433028057
```

Figure 4.4. Correlation coefficients between estimations of the same variable in different regions of the body, "left" and "right", the sum of "right" and "left" sides (main) and the difference between them "diff" (Source: Own elaboration)

It is clear that the correlation coefficients are highly analogous between measurements, which is why it has been decided to only maintain the total or "main" value, which is the sum of the right and left side estimates of a variable. With the removal of variables containing estimates of a variable from the right and left side and the difference between sides, a 169x75 DF is left, which means 89 columns have been removed.

Finally, it is necessary to remove the male patients as the VAT distribution between men and women varies significantly, and 13 men are not enough data to apply ML models or to create synthetic data. After the removal of male patients, the variable "sexo" that indicates the sex of the patient becomes unnecessary, as all the data is from the same sex. The result is a DF of 156 female patients and 75 variables (156x75).

4.2.2. Handling outlier values

The detection of outlier values has been a combination of non-automated and automated tasks:

Non-automated outlier handling

Each extreme value has been carefully considered taking into account the human origin of the data. Only values considered "impossible" should be treated. An example of an "impossible" value would be a negative weight.

The process to detect outlier values has been the following:

- Reviewing the maximum and minimum values for each variable
- If a maximum or minimum is abnormal, further investigate the rest of the variable and asses the possibility of the abnormal values.
- If a value is considered "impossible" review the rest of the information of the patient with an anomaly. If the rest of the values are correct, transform the anomaly into a null value.
- Impute the null value.

The method to treat these outliers would be to impute instead of eliminating the patients with such values, to avoid the loss of information. No obvious outliers have been detected in this manner.

Automated outlier detection

To avoid extreme values that negatively affect the subsequent ML models, the Isolation Forest [42] algorithm from Scikit-Learn has been used. Isolation Forest isolates anomalies using an ensemble of decision trees.

This method has been chosen through a systematic approach, with an iterative process of experimentation and evaluation, the exploration of different combinations of outlier removal algorithms (as well as not removing any at all) have been tested and the outlier handling technique that has proven to be most effective, in terms of the performance in the subsequent ML algorithms, has been chosen.

4.2.3. Standardisation of data

As previously seen, the DF has data from diverse origins, and therefore different scales, which indicates the necessity of standardising the data to the same scale for interpretability reasons.

A min-max scaler is applied to all the values apart from the categorical or binary variables: "Fuma", "EnolUBEsem", "DM2" and "HTA". The result is all the continuous variables having values in the range of 0 to 1.

4.2.4. Exploratory Data Analysis

Part of the exploratory analysis is mentioned in *5.1. Database description* section. The first part of the EDA consists of the description of the raw database and the types of variables it contains. Moreover, a detailed examination of the variables is recommended. Once a basic comprehension of the DF is gained, a visual exploratory analysis is done.

Three kinds of visual representations have been made, with each variable and the target value:

Scatter plots

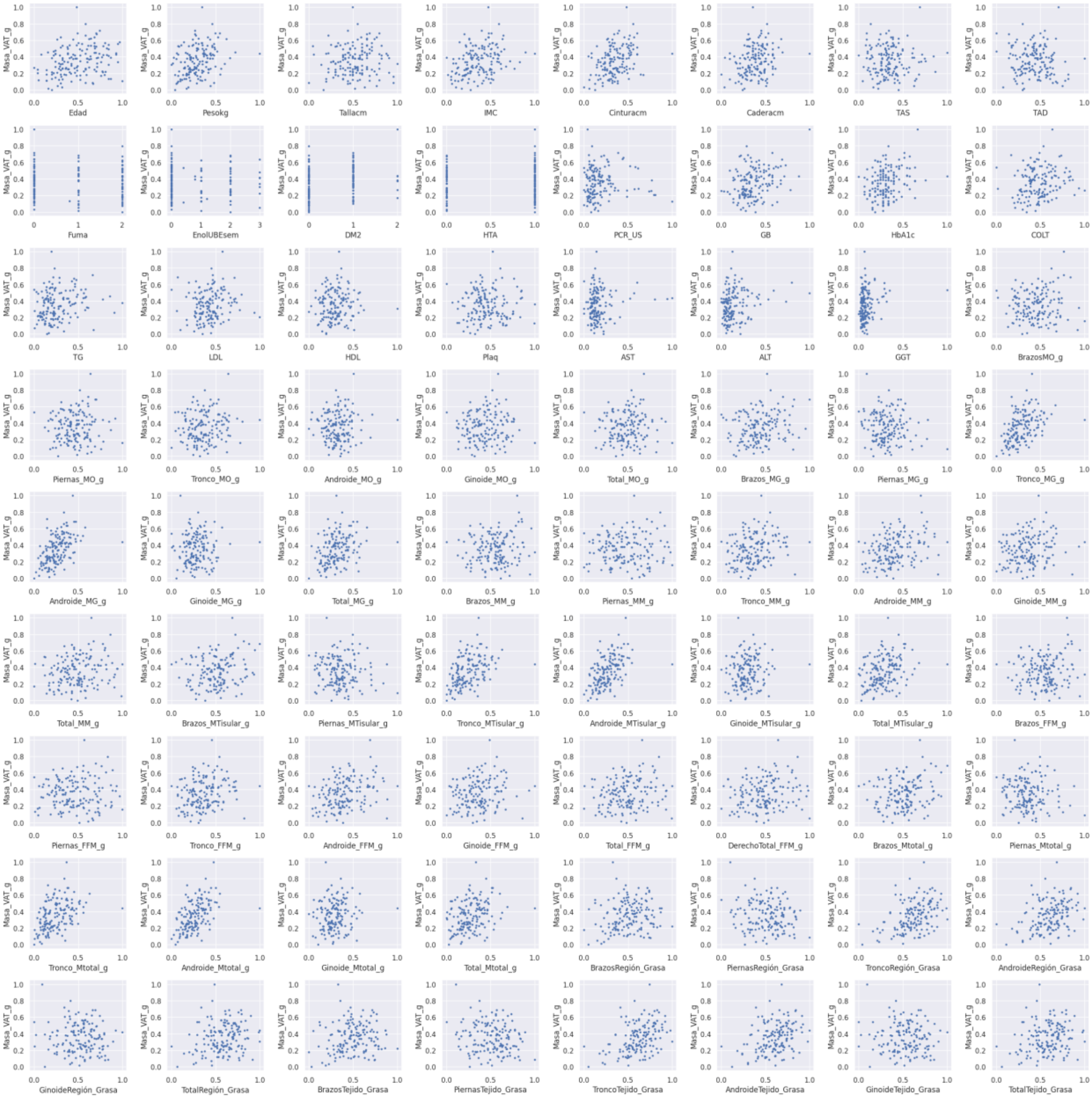


Figure 4.5. Compilation of all the scatter plots of the variables vs. target data ("Masa_VAT_g") (Source: Own elaboration)

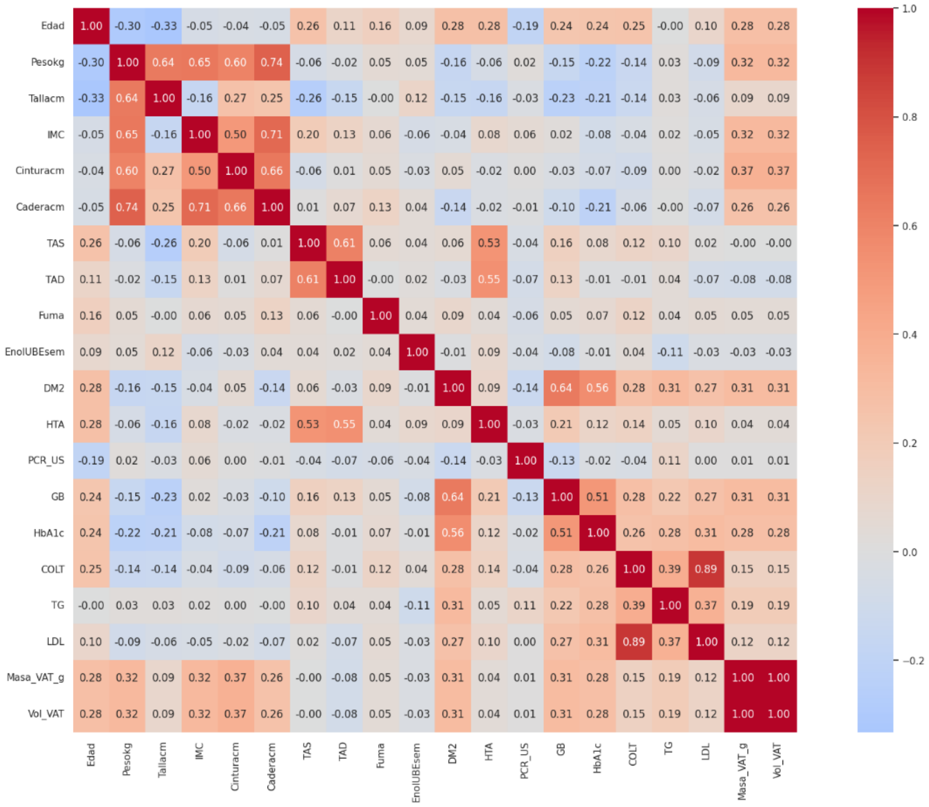
As seen in figure 4.5, most of the variables show high dispersion of values, nonetheless, some of them have an observable linear behaviour.

For that reason, it is of interest to look into the correlation matrix and coefficients of the variables with the label data. To visualize the correlation coefficients heatmaps can be used.

Heatmaps:

Heatmaps are graphical representations to visualise data in a matrix format using colour-coded cells, that in this case, represent the value of the correlation coefficient.

The heatmaps have been divided into subgroups for better interpretability. An example of a heatmap computed with the seaborn library [43] can be seen in Figure 4.6, the rows of interest are the last two ones, as they represent the correlations with the target variables.



	Edad	Pesokg	Tallacm	IMC	Cinturacm	Caderacm	TAS	TAD	Fuma	EnolUBEsem	DM2	HTA	PCR_US	GB	HbA1c	COLT	TG	LDL	Masa_VAT_g	Vol_VAT
Edad	1.00	-0.30	-0.33	-0.05	-0.04	-0.05	0.26	0.11	0.16	0.09	0.28	0.28	-0.19	0.24	0.24	0.25	-0.00	0.10	0.28	0.28
Pesokg	-0.30	1.00	0.64	0.65	0.60	0.74	-0.06	-0.02	0.05	0.05	-0.16	-0.06	0.02	-0.15	-0.22	-0.14	0.03	-0.09	0.32	0.32
Tallacm	-0.33	0.64	1.00	-0.16	0.27	0.25	-0.26	-0.15	-0.00	0.12	-0.15	-0.16	-0.03	-0.23	-0.21	-0.14	0.03	-0.06	0.09	0.09
IMC	-0.05	0.65	-0.16	1.00	0.50	0.71	0.20	0.13	0.06	-0.06	-0.04	0.08	0.06	0.02	-0.08	-0.04	0.02	-0.05	0.32	0.32
Cinturacm	-0.04	0.60	0.27	0.50	1.00	0.66	-0.06	0.01	0.05	-0.03	0.05	-0.02	0.00	-0.03	-0.07	-0.09	0.00	-0.02	0.37	0.37
Caderacm	-0.05	0.74	0.25	0.71	0.66	1.00	0.01	0.07	0.13	0.04	-0.14	-0.02	-0.01	-0.10	-0.21	-0.06	-0.00	-0.07	0.26	0.26
TAS	0.26	-0.06	-0.26	0.20	-0.06	0.01	1.00	0.61	0.06	0.04	0.06	0.53	-0.04	0.16	0.08	0.12	0.10	0.02	-0.00	-0.00
TAD	0.11	-0.02	-0.15	0.13	0.01	0.07	0.61	1.00	-0.00	0.02	-0.03	0.55	-0.07	0.13	-0.01	-0.01	0.04	-0.07	-0.08	-0.08
Fuma	0.16	0.05	-0.00	0.06	0.05	0.13	0.06	-0.00	1.00	0.04	0.09	0.04	-0.06	0.05	0.07	0.12	0.04	0.05	0.05	0.05
EnolUBEsem	0.09	0.05	0.12	-0.06	-0.03	0.04	0.04	0.02	0.04	1.00	-0.01	0.09	-0.04	-0.08	-0.01	0.04	-0.11	-0.03	-0.03	-0.03
DM2	0.28	-0.16	-0.15	-0.04	0.05	-0.14	0.06	-0.03	0.09	-0.01	1.00	0.09	-0.14	0.64	0.56	0.28	0.31	0.27	0.31	0.31
HTA	0.28	-0.06	-0.16	0.08	-0.02	-0.02	0.53	0.55	0.04	0.09	0.09	1.00	-0.03	0.21	0.12	0.14	0.05	0.10	0.04	0.04
PCR_US	-0.19	0.02	-0.03	0.06	0.00	-0.01	-0.04	-0.07	-0.06	-0.04	-0.14	-0.03	1.00	-0.13	-0.02	-0.04	0.11	0.00	0.01	0.01
GB	0.24	-0.15	-0.23	0.02	-0.03	-0.10	0.16	0.13	0.05	-0.08	0.64	0.21	-0.13	1.00	0.51	0.28	0.22	0.27	0.31	0.31
HbA1c	0.24	-0.22	-0.21	-0.08	-0.07	-0.21	0.08	-0.01	0.07	-0.01	0.56	0.12	-0.02	0.51	1.00	0.26	0.28	0.31	0.28	0.28
COLT	0.25	-0.14	-0.14	-0.04	-0.09	-0.06	0.12	-0.01	0.12	0.04	0.28	0.14	-0.04	0.28	0.26	1.00	0.39	0.89	0.15	0.15
TG	-0.00	0.03	0.03	0.02	0.00	-0.00	0.10	0.04	0.04	-0.11	0.31	0.05	0.11	0.22	0.28	0.39	1.00	0.37	0.19	0.19
LDL	0.10	-0.09	-0.06	-0.05	-0.02	-0.07	0.02	-0.07	0.05	-0.03	0.27	0.10	0.00	0.27	0.31	0.89	0.37	1.00	0.12	0.12
Masa_VAT_g	0.28	0.32	0.09	0.32	0.37	0.26	-0.00	-0.08	0.05	-0.03	0.31	0.04	0.01	0.31	0.28	0.15	0.19	0.12	1.00	1.00
Vol_VAT	0.28	0.32	0.09	0.32	0.37	0.26	-0.00	-0.08	0.05	-0.03	0.31	0.04	0.01	0.31	0.28	0.15	0.19	0.12	1.00	1.00

Figure 4.6. Heatmap of some of the features and label data (Source: Own elaboration)

Overall, the correlations can be qualified with either no correlation or low to medium positive correlation considering the following criteria:

- Correlation coefficient = 0: no-correlation
- 0 < Correlation coefficient < 0.3: low correlation
- 0.3 < Correlation coefficient < 0.5: medium correlation
- Correlation coefficient > 0.5: High correlation

Note: This criterion works the same for negative correlations by simply using the absolute value of such coefficients.

Moreover, the correlation coefficients have been independently computed and the variables that best correlate (Correlation coefficient > 0.3), which are 19, have been stored in a separate DF for the subsequent feature selection.

Scatterplots of the variables that correlate the best can be seen in figure 4.7.

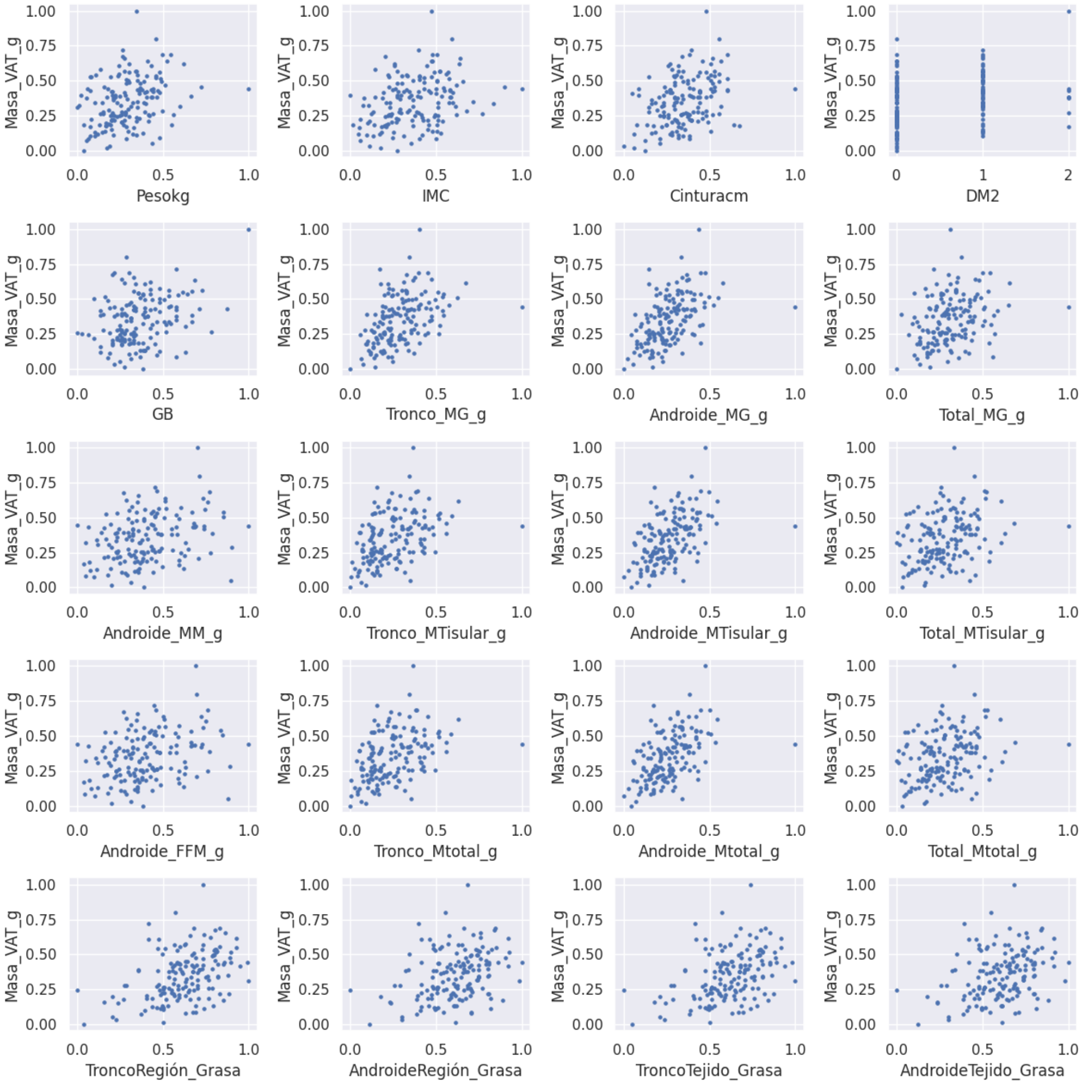


Figure 4.7. Compilation of all the scatter plots of the variables that correlate the best with the target data "Masa_VAT_g" (Source: Own elaboration)

Furthermore, boxplots have been computed to assure that all the outlier values have been considered, as boxplots provide information about the distribution of the data as well as potential outliers.

4.2.5. Transforming the target data into categorical data

The first ML algorithms that will be applied are aimed at classifying VAT quantities. To perform this classification, the target data cannot be in a quantitative data format; instead, it needs to be in a categorical or qualitative format. To achieve this, the target variable 'Masa_VAT_g' has been divided into three groups using tertiles. This division has been done using the Panda's library 'quantile' and 'cut' functions, and the result is the following:

- 0: Values smaller than the first tertile, lower than 33.33% of the data ⟶ Low VAT
- 1: Values between the first and second tertiles, between 33.33% and 66.67% of the data ⟶ Medium VAT
- 2: Values higher than the second tertile, above 66.67% of the data ⟶ High VAT

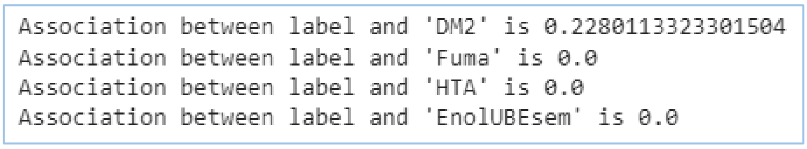
This transformation has been stored in a new "Masa_VAT_cat" variable, that has been added to the data frame. Please note that while these three categories, 0, 1, and 2, have been labelled as Low, Medium, and High VAT, respectively, it is important to understand that these names are relative. Technically, all VAT quantities in the data frame are above average since they are obtained from obese patients.

4.2.6. Feature selection

Part of the feature selection has been a gradual process throughout the pre-processing, by progressively cleaning the data, the DF has already been reduced from 169x167 to a 156x75 dimension.

Nonetheless, features can be further narrowed down by using association metrics between the variables and the target data. One association metric that has been already reviewed in *5.3.4. The exploratory Data Analysis* section is the correlation coefficient, which quantifies the degree to which quantitative variables are linearly related or associated with the target data.

Another association metric implemented is the Cramer's V algorithm from the Dython library, this algorithm computes the association between categorical data, the measures go from no association, 0, to high association, 1. The only categorical variables "Fuma", "DM2", "EnolUBEsem" and "HTA" have had the following results shown in Figure 4.8.



```
Association between label and 'DM2' is 0.2280113323301504
Association between label and 'Fuma' is 0.0
Association between label and 'HTA' is 0.0
Association between label and 'EnolUBEsem' is 0.0
```

Figure 4.8. Cramer V algorithm used between categorical variables and target data (Source: Own elaboration)

It can be seen from the figure that the only variable that provides evidence of association with the target is "DM2", consequently the rest won't be considered as features for now.

To recapitulate, the main points for feature selection discussed up to this point are the following:

- Feature selection with the correlation coefficient between quantitative variables and the target variable: 19 possible features.
- Feature selection with the association between categorical variables and the target variable: 1 possible feature.

Nonetheless, a final feature selection strategy is a knowledge-driven selection, mostly non-anthropometric variables (blood-related data and clinical data) have been selected based on their known biological importance in obesity. They can be seen in Table 4.1.

Selected feature	Reasoning	Selected feature	Reasoning
"Pesokg" Weight	Directly correlated to obesity	"COLT" Total cholesterol	Obesity is known to increase total cholesterol levels
"TAS" Peak systolic blood pressure	Blood pressure-related data, obesity is known to increase high blood pressure risk.	"TG" Triglycerides	Obesity is linked to elevated triglyceride levels
"TAD" Minimum diastolic pressure	Blood pressure-related data, obesity is known to increase high blood pressure risk.	"LDL" Low-density lipoprotein	"bad-cholesterol", LDL levels tend to increase with obesity
"DM2" Diabetes	Obesity is known to increase diabetes risk.	"HDL" High-density lipoprotein	"good-cholesterol", HDL levels tend to decrease with obesity.
"HTA" Arterial hypertension	Blood pressure-related data, obesity is known to increase high blood pressure risk.	"Plaq" Platelets	Obesity is linked to elevated platelet levels.
"PCR_US" Ultrasensitive c-reactive protein	Biomarker used to evaluate cardiovascular risk. Obesity is known to increase cardiovascular disease risk.	"AST" Aspartate aminotransferase	Liver disease-related data, Obesity is known to increase non-alcoholic fatty liver disease risk.
"GB" Basal glucose	Diabetes-related data, obesity is known to increase diabetes risk.	"ALT" Alanine aminotransferase	Liver disease-related data, Obesity is known to increase non-alcoholic fatty liver disease risk.
"HbA1c" Glycosylated hemoglobin	Diabetes-related data, obesity is known to increase diabetes risk.	"GGT" Gamma-glutamyl transpeptidase	Liver disease-related data, Obesity is known to increase non-alcoholic fatty liver disease risk.

Table 4.1. Knowledge-driven feature selection variables (Source: Own elaboration)

4.2.7. Pre-processing outcome

As it is of interest to capture a more comprehensive view of the underlying mechanisms for VAT distribution and a deeper understanding of obesity, a comparison between the results of the usage of anthropometric data and non-anthropometric data will be done with the following final data frames, described in table 4.2. and figure 4.9.

(To see the specific features of these data frames see ANNEX A1 and A2)

Data Frame name	Information	Nº of features
DataAP	DF with only anthropometric data from all the feature selection methods explained.	17
DataPHY	DF with non-anthropometric data from knowledge-driven selection.	16
DataALL	DF that combines both DataAP and DataPHY.	33

Table 4.2: Dataframes obtained after the pre-processing (Source: Own elaboration)

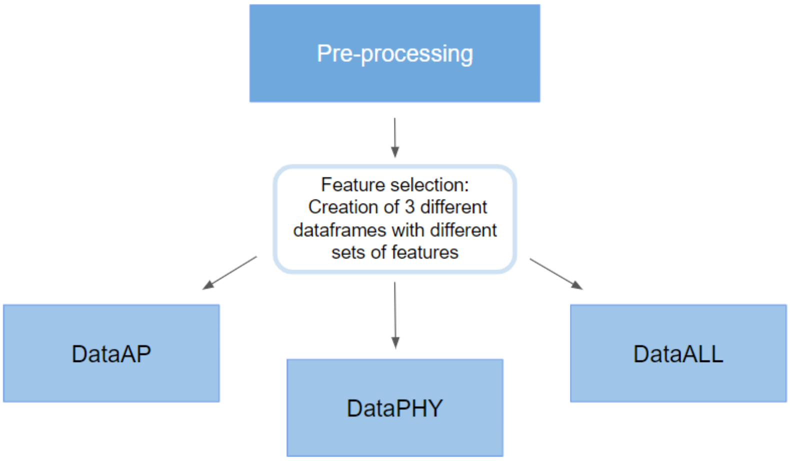

Pre-processing
Feature selection: Creation of 3 different dataframes with different sets of features
DataAP
DataPHY
DataALL

Figure 4.9. Pre-processing results diagram (Source: Own elaboration)

4.3. Classification

The classification process consists of using the “Masa_VAT_cat” variable as the target data, as it is the target data in a categorical form, and using the features from each dataset, DataAP, DataPHY and DataALL, as input data for the classifications.

Each dataset has been split into a training set and a testing set, following an 80-20 proportion respectively, using the “train_test_split” [44] function from the Scikit-Learn library.

4.3.1. GridSearch

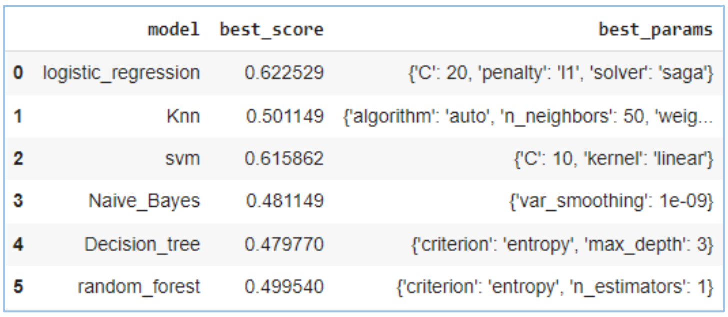
Once the pre-processed data is imported, the GridSearch algorithm [45] from the Scikit-Learn library is employed to find the best classification models and their respective optimal hyperparameters.

GridSearch essentially works as a cross-Validation technique, used to determine the optimal hyperparameter values for a given model by exhaustively searching through a predefined set of hyperparameter values and evaluating the performance of the model iterating over each combination of values.

The classification models evaluated with GridSearch are the following:

- Logistic Regression
- K-nearest neighbours
- Support Vector Machine
- Naïve Bayes
- Decision Tree
- Random Forest

Finding the best model and hyperparameter tuning with GridSearch results, shown in Figure 4.10.



	model	best_score	best_params
0	logistic_regression	0.622529	{'C': 20, 'penalty': 'l1', 'solver': 'saga'}
1	Knn	0.501149	{'algorithm': 'auto', 'n_neighbors': 50, 'weig...
2	svm	0.615862	{'C': 10, 'kernel': 'linear'}
3	Naive_Bayes	0.481149	{'var_smoothing': 1e-09}
4	Decision_tree	0.479770	{'criterion': 'entropy', 'max_depth': 3}
5	random_forest	0.499540	{'criterion': 'entropy', 'n_estimators': 1}

Figure 4.10. Results of the GridSearch search for the best model with the dataAP data frame (Source: Own elaboration)

From the GridSearch result it can be seen that some models perform better than others, the models with the highest scores will be further evaluated.

The GridSearch algorithm has been used a total of three times, one for each subset of features.

4.3.2. Evaluation

In the evaluation process of the classification models, the confusion matrix, the ROC curve and evaluation metrics using k-cross validation will be computed to assess the performance of each model. In the following lines, an example of the evaluation of the logistic regression model for the dataAP dataset will be shown.

Confusion matrix:

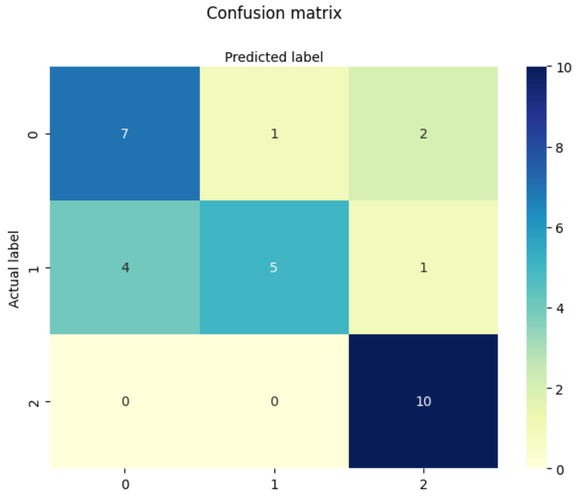


Figure 4.11. Confusion matrix of the LR model with *dataAP* features. (Source: Own elaboration)

As seen in figure 4.11, the specific model shown has an acceptable performance when classifying “Low” and “High” VAT quantities.

Mean ROC curve:

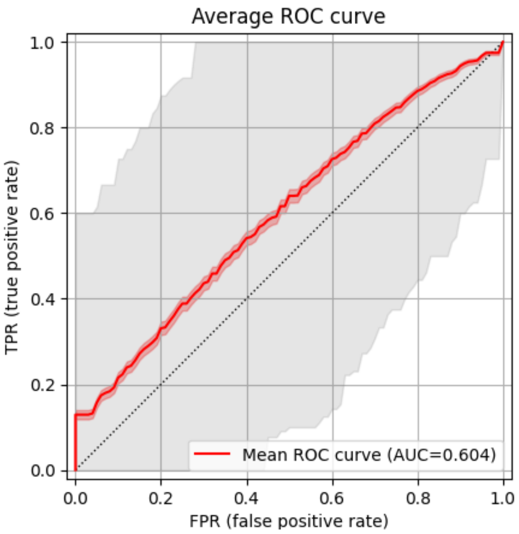
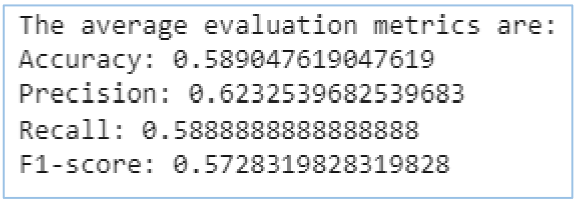


Figure 4.12. Mean ROC curve of the LR model with *dataAP* features. (Source: Own elaboration)

As seen in Figure 4.12, the model has an AUC > 0.5, which indicates that it performs better than a random classifier.

Evaluation metrics using k-cross-validation:

To better calculate evaluation metrics, a 10-fold cross-validation approach has been employed. In this process, the dataset has been divided into ten equally sized subsets or "folds." The model has been trained and evaluated ten times, each time using a different fold as the testing set and the remaining nine folds as the training set. This allows for a comprehensive assessment of the model's performance across different data samples. For each fold, the accuracy, precision, recall, and F1 score have been computed, resulting in ten individual scores for each metric. By averaging these scores, a more robust and reliable estimate has been obtained. This 10-fold cross-validation provides a more comprehensive evaluation of the classification model's performance, an example of an outcome of a k-cross evaluation is shown in Figure 4.13.



```
The average evaluation metrics are:
Accuracy: 0.589047619047619
Precision: 0.6232539682539683
Recall: 0.5888888888888888
F1-score: 0.5728319828319828
```

Figure 4.13. Evaluation metrics with k-cross validation of the LR model with *dataAP* features. (Source: Own elaboration)

4.4. Neural Networks

The prediction process consists of using the “Masa_VAT_g” variable as the target data, as it is the target data in a continuous form, and using the features from each dataset, DataAP, DataPHY and DataALL, as input data for the predictions.

Each dataset has been split into a training set and a testing set, following an 80-20 proportion respectively, using the “train_test_split” [44] function from the Scikit-Learn library.

4.4.1. Keras Tuner

Once the pre-processed data is imported, the Keras tuner algorithm [46] from the TensorFlow library is utilized to optimize the NN hyperparameter search. Keras tuner is an optimization framework that solves the hyperparameter search for NN models by iterating over each hyperparameter combination until finding the one that performs the best.

However, it is important to note that the main architecture of the neural network needs to be manually defined. In this case, a simple four-layer network has been chosen. The network consists of an input layer, which contains as many input neurons as the number of features there are, followed by two hidden layers. Finally, there is an output layer with a single neuron responsible for predicting the quantity of VAT. An example of a NN architecture can be seen in Figure 4.14.

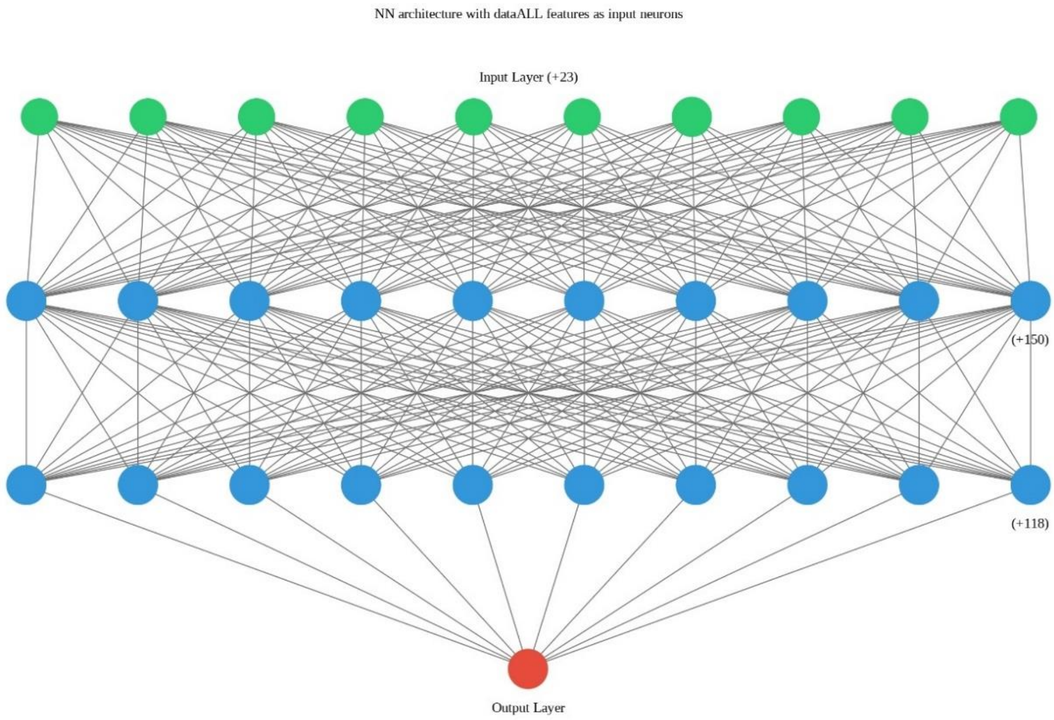
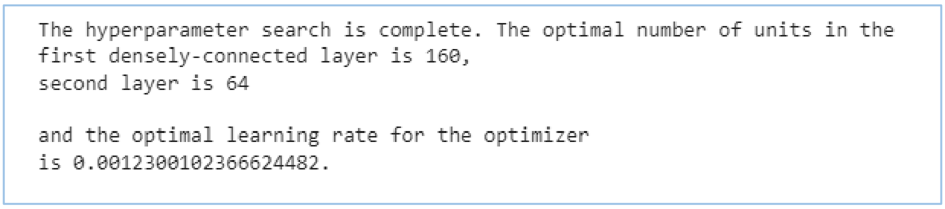


Figure 4.14. Example of a NN architecture, the result of an already optimized network for dataALL features. (Source: Own elaboration)

The hyperparameter search will define the number of hidden layer neurons and the Learning Rate of the NN algorithm. Example of hyperparameter tuning output using Keras Tuner, using features from DataPHY as input neurons is shown in Figure 4.15.



```
The hyperparameter search is complete. The optimal number of units in the
first densely-connected layer is 160,
second layer is 64

and the optimal learning rate for the optimizer
is 0.0012300102366624482.
```

Figure 4.15. Keras Tuner hyperparameter tuning outcome for a 4-layer NN using dataPHY features. (Source: Own elaboration)

A total of three NN algorithms have been modelled, one for each subset of feature values (dataAP, dataPHY and dataALL). Once the NN models are defined, an evaluation is done to further assess the quality of the predictions.

4.4.2. Evaluation

The evaluation process of the prediction consists of a visual analysis of the results and a summary of R-squared and the MSE using k-cross validation for each of the three prediction models. In the following lines, an example of the evaluation of the NN model for the DataALL features will be shown.

Visual analysis

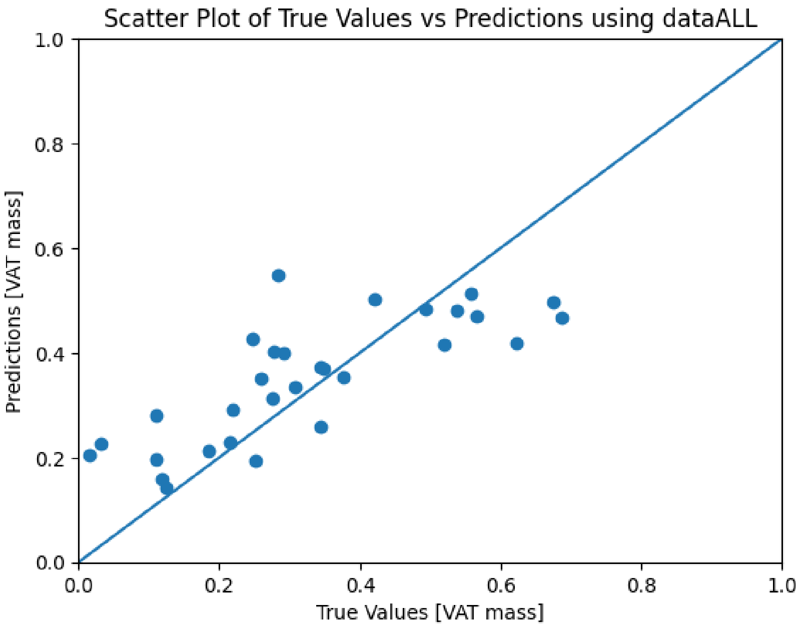


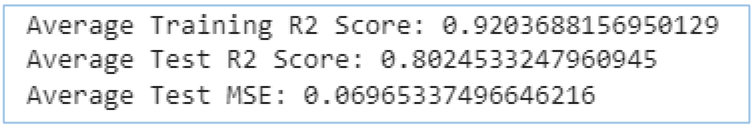
Figure 4.16. Visual representation of the prediction of the testing set vs. the actual values of the testing set. (Source: Own elaboration)

The scatter plot is shown in Figure 4.16. helps to visualize the prediction error of the model, as the diagonal line represents a perfect prediction, and therefore, the distance between the scatter plot values and the line represents the prediction error.

Evaluation metrics using k-cross-validation

To better calculate evaluation metrics, a 10-fold cross-validation approach has been employed. In this process, the dataset has been divided into ten equally sized subsets or "folds." The model has been trained and evaluated ten times, each time using a different fold as the testing set and the remaining nine folds as the training set. This allows for a comprehensive assessment of the NN model's performance across different data samples. For each fold, the R-squared and MSE have been computed, resulting in ten individual scores for each metric. By averaging these scores, a more robust and reliable estimate has been obtained. This 10-fold cross-validation provides a more comprehensive evaluation of the NN model's performance.

Next in figure 4.17. there is an example of the evaluation metrics, the result of applying the metrics using k-cross validation.



```
Average Training R2 Score: 0.9203688156950129
Average Test R2 Score: 0.8024533247960945
Average Test MSE: 0.06965337496646216
```

Figure 4.17. K-cross validation average training R-squared, average testing R-squared and MSE for the NN, modelled with dataALL features. (Source: Own elaboration)

Overall, R-squared can be used to assess the fit of a neural network regressor by considering the following criteria:

- 0-0.25: Little to no variance explained
- 0.25-0.50: Small amount of variance explained
- 0.5-0.75: Good amount of variance explained
- 0.75-1: Significant amount of variance explained.

When R-squared explains a good amount of variance, it means that a larger proportion of the variability in the VAT quantities can be explained by the feature variables in the NN. In other words, a higher R-squared implies that the model is a better fit for the data and has a stronger ability to explain or predict the outcomes of VAT quantity. On the other hand, a lower MSE indicates that the model has better predictive capacity and is closer to the true values.

Therefore, the example provided in Figure 4.17, can be interpreted as having a high predictive value, R2>0.75, and a relatively low MSE.

5. Results and discussion

Results 1: CLASSIFICATIONS

For the discussion of the results, the best classification models for each database, dataAP, dataPHY and dataALL will be provided with their respective evaluation metrics.

Best model trained with anthropometric data, from **dataAP** data frame

Logistic regression model, as defined with Python: `LogisticRegression(solver='saga', C=20, penalty='l1')`

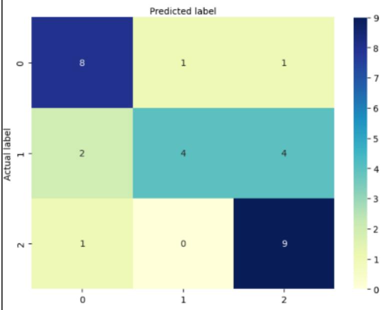
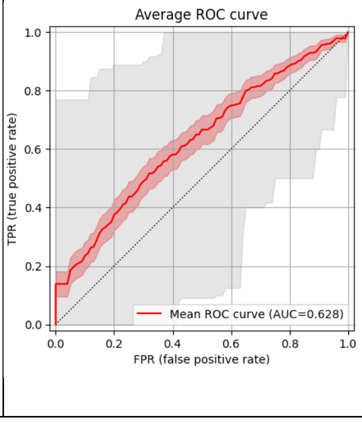
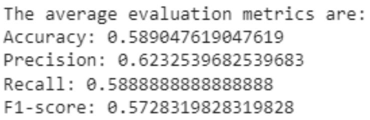
Confusion matrix	ROC curve and AUC	Evaluation metrics
 	 	 The average evaluation metrics are: Accuracy: 0.589047619047619 Precision: 0.6232539682539683 Recall: 0.5888888888888888 F1-score: 0.5728319828319828

Table 5.1. Summary of the LR classification model with dataAP. (Source: Own elaboration)

Best model trained with biochemical and clinical data, from **dataPHY** data frame

Support Vector Machine model, as defined with Python: `svm.SVC(gamma='auto', C = 10, kernel = 'linear')`

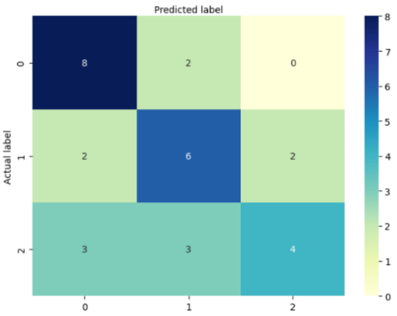
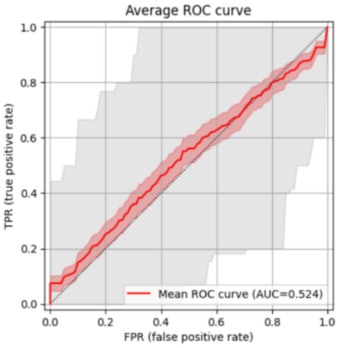
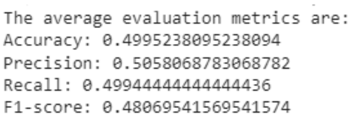
Confusion matrix	ROC curve and AUC	Evaluation metrics
 	 	 The average evaluation metrics are: Accuracy: 0.4995238095238094 Precision: 0.5058068783068782 Recall: 0.49944444444444436 F1-score: 0.48069541569541574

Table 5.2. Summary of the SVM classification model with dataPHY. (Source: Own elaboration)

Best model trained with all types of data, from **dataALL** data frame

Support Vector Machine model, as defined with Python: `svm.SVC(gamma='auto', C = 20, kernel = 'linear')`

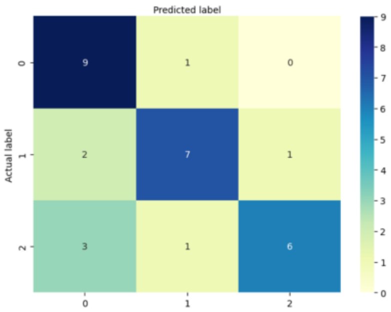
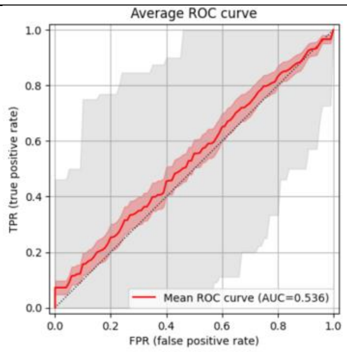
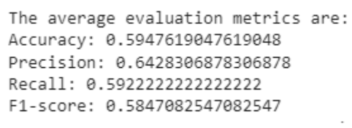
Confusion matrix	ROC curve and AUC	Evaluation metrics
 	 	 The average evaluation metrics are: Accuracy: 0.5947619047619048 Precision: 0.6428306878306878 Recall: 0.5922222222222222 F1-score: 0.5847082547082547

Table 5.3. Summary of the SVM classification model with dataALL. (Source: Own elaboration)

Discussion 1

As seen in tables 5.1, 5.2 and 5.3, the classification model's performance varies to a certain extent between the different databases used. However, it is clear that the worst performance is for the dataPHY database, as the AUC is very close to 0.5, performing similarly to a random classifier, and its evaluation metrics are also unfavourable. Nevertheless, the evaluation metrics between models that used dataAP and dataALL to train are very similar, both models correctly predicted the outcome class for approximately 60% of the testing data. Nonetheless, the AUC of 0.536 from the ROC curve of the model that used dataALL also suggests the behaviour of a random classifier. On the contrary, the AUC from the ROC curve of the model that used dataAP of 0.628 suggests a higher discriminatory power and better ability to distinguish between the given instances.

To sum it up, the best classification model overall is the logistic regression model trained with anthropometric data from dataAP.

Results 2: PREDICTIONS

For the discussion of the results, each Neural Network model created, three in total, trained with data from dataAP, dataPHY and dataALL, will be provided with their respective evaluation metrics.

<u>NN model trained with anthropometric data, from **dataAP** data frame</u>

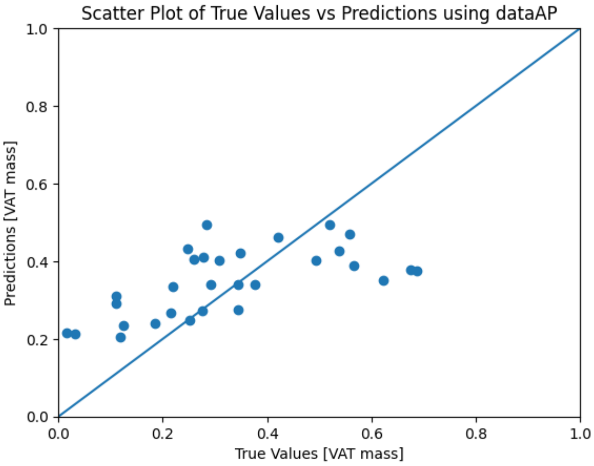
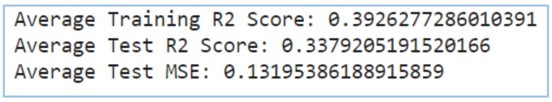


Figure 5.1. Scatter plot comparing the true values with the predicted values from the model trained with dataAP. (Source: own elaboration)



```
Average Training R2 Score: 0.3926277286010391
Average Test R2 Score: 0.3379205191520166
Average Test MSE: 0.13195386188915859
```

Figure 5.2. K-cross validation average metrics for the NN modelled with dataAP features. (Source: own elaboration)

<u>NN model trained with biochemical and clinical data, from **dataPHY** data frame</u>

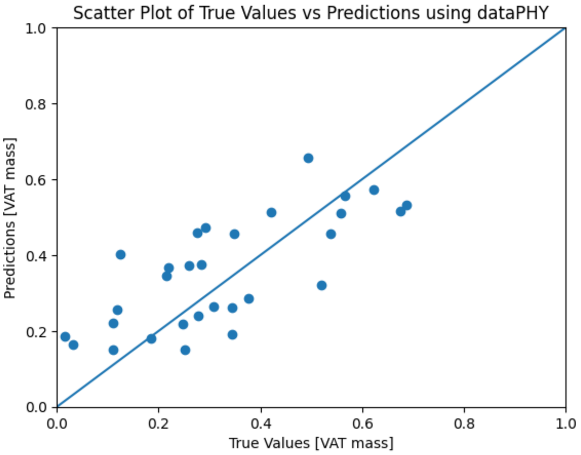
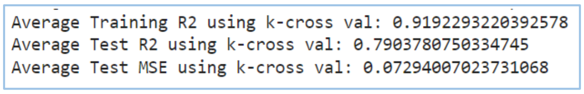


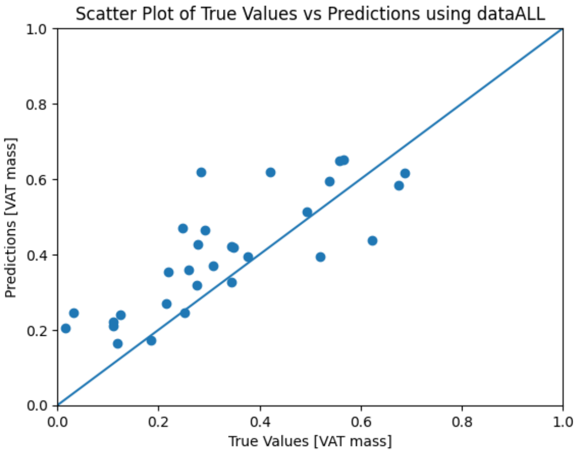
Figure 5.3. Scatter plot comparing the true values with the predicted values from the model trained with dataPHY. (Source: own elaboration)

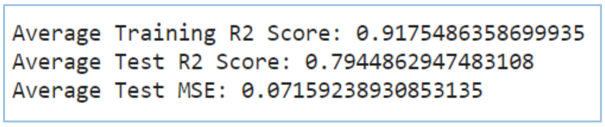


```
Average Training R2 using k-cross val: 0.9192293220392578
Average Test R2 using k-cross val: 0.7903780750334745
Average Test MSE using k-cross val: 0.07294007023731068
```

Figure 5.4. K-cross validation average metrics for the NN modelled with dataPHY features. (Source: own elaboration)

NN model trained with all types of data, from **dataALL** data frame





```
Average Training R2 Score: 0.9175486358699935
Average Test R2 Score: 0.7944862947483108
Average Test MSE: 0.07159238930853135
```

Figure 5.6. K-cross validation average metrics for the NN modelled with dataALL features. (Source: own elaboration)

Figure 5.5. Scatter plot comparing the true values with the predicted values from the model trained with dataALL. (Source: own elaboration)

Discussion 2

As seen in the figures 5.1 to 5.6, from the three NN, the worst results are undoubtedly from the NN model trained with anthropometric features from the dataAP database, as it is observable in figure 5.2 the R-squared value of the testing dataset is 0.34, which suggests a moderate degree of explanatory power, indicating that the anthropometric variables have some degree of influence on the variation in VAT. Nonetheless, there is still a significant amount of unexplained variability.

On the other hand, the NN models trained with data from dataPHY and dataALL have roughly identical performance. Their testing R-squared (Fig 5.4 and 5.6) is almost 0.8, which indicates a high degree of explanatory power, suggesting that the features contained in both data frames have a strong influence on the variation of VAT.

It is important to consider then, that the performance of the NN defined with mostly biochemical data from blood tests (DataPHY) is the same as the one containing the same information plus anthropometric data (DataALL). This suggests that the influence of anthropometric data in the NN trained with dataALL is very low or null. Therefore, the model with the best result and cost-effectiveness, computationally speaking, is the NN trained with the dataPHY dataset.

6. Environmental impact analysis

This project has had a minimal environmental impact, as for the realization of this project only a laptop has been used (hp pavilion x360 convertible 14-dw1098nr). These kinds of laptops do not have a considerable electricity consumption, tough they still leave a carbon footprint. Considering the usage time of the laptop for the completion of the project to be 600h, the average consumption of energy of the laptop is approximately 40W/h, and the emission factor for the electricity grid in Catalonia is 259gCO2eq/kWh [47], the total carbon footprint is approximately of 6,216 kg of CO2.

However, other factors can as well be considered, such as the carbon footprint that running code in Google Colab and storing data in the cloud can have. When computation tasks are performed on Google Colab and data is stored in servers, the underlying infrastructure and servers consume electricity, which also generates carbon emissions.

Conclusions and future projects

After the pre-processing, classification and prediction steps of this project, valuable insights can be obtained from the best performing models.

Firstly, considering that the classification model that portrayed better results is the one trained with the dataAP dataset, we can extract the somewhat linear or direct relation between VAT and anthropometric data, due to the interpretability of logistic regression models in comparison with NN models. Logistic regression results directly represent the relationship between input data (features) and the output (VAT), allowing for a clear understanding of the impact of the variables chosen.

Conversely, NN encompass complex computations across multiple layers, complicating the interpretation of the contribution of specific input variables. However, NN can capture nonlinear and complex patterns and interactions in the data, which results in more flexible approaches to prediction modelling. Therefore, from the results of the NN prediction that performed the best, we can extract the significant influence of biochemical data and clinical data in VAT assessment. Moreover, the data used reassures the outcomes of several studies [12] [16] where VAT has been shown to be linked with the prevalence of some of the conditions that dataPHY features are associated to. (See Table 4.1.)

The results of the predictions are of significant importance, because the variables used for the training (dataPHY) are of relatively easy acquisition through blood test analysis and medical records. This opens the door, with the help of further investigation, for the possibility of creating a cost-effective tool for an initial approximation of VAT, and consequently of obesity-related risks, in situations where the common tools to study fat distribution, such as DEXA and MRI, are not easily accessible. Therefore enabling healthcare professionals to identify high-risk obese individuals.

Limitations

The development of this project has been limited by the fact that only data from female patients has been used due to the scarcity of male patient data. Considering the difference in VAT distribution between men and women (*see section 2.2.1. Distribution and functionality*), the created models might not be able to generalize well to this segment of the population. It will most likely properly work to predict female VAT quantities, and consequently allowing solely to stratify female obese patients.

Another limitation of the project has been the overall low number of patients in the dataset (156 patients whose data has been used). Generally, in ML, it is preferable to have a larger number of patients for the models to learn from a broader range of patterns and variations in the data. This helps to make the models more robust, reducing their sensitivity to small variations in data.

Future improvements

Following the previous statement, a future improvement would be the addition of more patients in the dataset, as well as having the same proportion of males and females to properly assess each segment of the obese population. With an increased dimension of the data, ML training can be improved, and the models obtained can be further validated with external datasets that have not been used for the training. In addition, if the data set were large enough, dimensionality reduction techniques, such as Principal Component Analysis (PCA), could be applied to enhance the performance and efficiency of the ML models.

Another crucial improvement would be the validation of the ML models with clinicians and domain experts to evaluate the clinical applications of the models through their expertise and domain knowledge.

Lastly, another future addition to the project would be the extension of the project itself by adding image data obtained through DEXA imaging and applying ML techniques, such as convolutional neural networks (CNN), to segment the images to directly measure VAT from the segmentations.

Economic analysis

The main categories to be considered for the economic analysis of this project are human resources, hardware, software, and services.

For the realization of this project, Google Colaboratory has been used, which is a cloud-hosted version of Jupiter-notebook that allows easy access to Python libraries, this programming environment is free of charge and therefore the cost of Software in this project is null.

Human resources

Description	**Hours**	**Cost (€/h)**	**TOTAL**
Engineering student	600h	8€/h	4800€
Tutor	50h	30€/h	1500€
Co-tutor	15h	30€/h	450€
TOTAL:			**6750€**

This project has been mainly created with the participation of three individuals; an engineering student, a tutor and a co-tutor. For the simplification of costs, the author will be considered to have the minimum required salary of an intern from the Universitat Politècnica de Catalunya.

Hardware

Description	**Quantity**	**Cost**
Laptop (hp pavilion x360 convertible 14-dw1098nr)	1	999€
TOTAL:		**999€**

Services

By estimation, the cost of the electricity used by a laptop in 600h of usage is roughly 6€, therefore it will be neglected from the total costs.

TOTAL COST = 7.749€

Bibliography

[1] Shuster, A., Patlas, M., Pinthus, J. H., & Mourtzakis, M. (2012). The clinical importance of visceral adiposity: a critical review of methods for visceral adipose tissue analysis. *The British Journal of Radiology*, *85*(1009), 1. https://doi.org/10.1259/BJR/38447238

[2] Iacobellis, G. (2005). Imaging of visceral adipose tissue: An emerging diagnostic tool and therapeutic target. *Current Drug Targets. Cardiovascular & Haematological Disorders*, *5*(4), 345–353. https://doi.org/10.2174/1568006054553408

[3] *Visceral Fat Vs. Subcutaneous Fat* (n.d.). Retrieved May 12, 2023, from https://elitebodysculpture.com/airsculpt-daily/visceral-fat-vs-subcutaneous-fat-will-lipo-help-either/

[4] Lavie, C. J., de Schutter, A., Parto, P., Jahangir, E., Kokkinos, P., Ortega, F. B., Arena, R., & Milani, R. v. (2016). Obesity and Prevalence of Cardiovascular Diseases and Prognosis—The Obesity Paradox Updated. *Progress in Cardiovascular Diseases*, *58*(5), 537–547. https://doi.org/10.1016/J.PCAD.2016.01.008

[5] World Heart Federation. *"Obesity"*. Retrieved May 10, 2023, from https://world-heart-federation.org/what-we-do/obesity/

[6] World Health Organization. "Obesity." Who.int. Accessed May 11, 2023. https://www.who.int/health-topics/obesity.

[7] Instituto Nacional de Estadística. "4.6 Determinantes de salud (sobrepeso, consumo de fruta y verdura, tipo de lactancia, actividad física)" Accessed May 11, 2023. https://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259926457058&p=%5C&pagename=ProductosYServicios%2FPYSLayout&param1=PYSDetalle&param3=1259924822888.

[8] Hernáez, Á., Zomeño, M. D., Dégano, I. R., Pérez-Fernández, S., Goday, A., Vila, J., Civeira, F., Moure, R., & Marrugat, J. (2019). Excess Weight in Spain: Current Situation, Projections for 2030, and Estimated Direct Extra Cost for the Spanish Health System. *Revista Española de Cardiología (English Edition)*, *72*(11), 916–924. https://doi.org/10.1016/J.REC.2018.10.010

[9] *Obesity in Spain: 80% of men and 55% of women in Spain will be overweight by 2030: study | Spain | EL PAÍS English*. (n.d.). Retrieved May 30, 2023, from https://english.elpais.com/elpais/2019/01/10/inenglish/1547131751_501777.html

[10] Frank, A. P., de Souza Santos, R., Palmer, B. F., & Clegg, D. J. (2019). Determinants of body fat distribution in humans may provide insight about obesity-related health risks. *Journal of Lipid Research*, *60*(10), 1710–1719. https://doi.org/10.1194/JLR.R086975

[11] Blüher, M. (2019). Obesity: global epidemiology and pathogenesis. In *Nature Reviews Endocrinology* (Vol. 15, Issue 5, pp. 288–298). Nature Publishing Group. https://doi.org/10.1038/s41574-019-0176-8

[12] Agrawal, S., Klarqvist, M. D. R., Diamant, N., Stanley, T. L., Ellinor, P. T., Mehta, N. N., Philippakis, A., Ng, K., Claussnitzer, M., Grinspoon, S. K., Batra, P., & Khera, A. v. (2023). BMI-

adjusted adipose tissue volumes exhibit depot-specific and divergent associations with cardiometabolic diseases. *Nature Communications 2023 14:1*, *14*(1), 1–10. https://doi.org/10.1038/s41467-022-35704-5

[13] National Health Service “Weight loss surgery”. Retrieved May 22, 2023, from https://www.nhs.uk/conditions/weight-loss-surgery/

[14] Pi-Sunyer, X. (2009). “The Medical Risks of Obesity.” *Postgraduate Medicine* 121 (6): 21–33. https://doi.org/10.3810/pgm.2009.11.2074.

[15] Lumish, H. S., O’Reilly, M., & Reilly, M. P. (2020). Sex Differences in Genomic Drivers of Adipose Distribution and Related Cardiometabolic Disorders. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *40*(1), 45–60. https://doi.org/10.1161/ATVBAHA.119.313154

[16] Després, J. P., Lemieux, I., & Prud’homme, D. (2001). Treatment of obesity: need to focus on high risk abdominally obese patients. *BMJ : British Medical Journal*, *322*(7288), 716. https://doi.org/10.1136/BMJ.322.7288.716

[17] Armao, D., Guyon, J. P., Firat, Z., Brown, M. A., & Semelka, R. C. (2006). Accurate quantification of visceral adipose tissue (VAT) using water-saturation MRI and computer segmentation: Preliminary results. *Journal of Magnetic Resonance Imaging*, *23*(5), 736–741. https://doi.org/10.1002/JMRI.20551

[18] Hu, H. H., Chen, J., & Shen, W. (2016). Segmentation and quantification of adipose tissue by magnetic resonance imaging. *Magma (New York, N.Y.)*, *29*(2), 259. https://doi.org/10.1007/S10334-015-0498-Z

[19] Crabtree, C. D., Lafountain, R. A., Hyde, P. N., Chen, C., Pan, Y., Lamba, N., Sapper, T. N., Short, J. A., Kackley, M. L., Buga, A., Miller, V. J., Scandling, D., Andersson, I., Barker, S., Hu, H. H., Volek, J. S., & Simonetti, O. P. (2019). Quantification of Human Central Adipose Tissue Depots: An Anatomically Matched Comparison Between DXA and MRI. *Tomography (Ann Arbor, Mich.)*, *5*(4), 358–366. https://doi.org/10.18383/J.TOM.2019.00018

[20] IBM “What is Machine Learning?” Retrieved May 20, 2023, from https://www.ibm.com /topics/machine-learning

[21] *Conjuntos de entrenamiento y prueba: División de datos | Machine Learning | Google for Developers*. (n.d.). Retrieved May 29, 2023, from https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data?hl=es-419

[22] Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 33–44. https://doi.org/10.1002/WICS.2

[23] Song, Q., & Shepperd, M. (2007). Missing data imputation techniques. *International Journal of Business Intelligence and Data Mining*, *2*(3), 261–291. https://doi.org/10.1504/IJBIDM. 2007.015485

[24] Cousineau, D., & Chartier, S. (2010). ISSN impresa (printed) 2011-2084 ISSN electrónica (electronic). In *International Journal of Psychological Research* (Vol. 3, Issue 1).

[25] Kuhn, M., Raton, K. J. B., Butcher, B., & Smith, B. J. (2020). Feature Engineering and Selection: A Practical Approach for Predictive Models 124 *74*(3), 308–309. https://doi.org/10.1080/00031305.2020.1790217

[26] *Cramér's V - IBM Documentation*. (n.d.). Retrieved June 1, 2023, from https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-cramrs-v

[27] Li, Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. (2018). "Feature Selection: A Data Perspective." *ACM Computing Surveys* 50 (6): 1–45. https://doi.org/10.1145/3136625.

[28] Rani, A., Kumar, N., Kumar, J., & Sinha, N. K. (2022). Machine learning for soil moisture assessment. *Deep Learning for Sustainable Agriculture*, 143–168. https://doi.org/10.1016/B978-0-323-85214-2.00001-X

[29] Shmilovici, A. (2005). Support Vector Machines. *Data Mining and Knowledge Discovery Handbook*, 257–276. https://doi.org/10.1007/0-387-25465-X_12

[30] Soofi, A. A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, *13*, 459–465.

[31] *Introduction to Decision Trees: Why Should You Use Them? | 365 Data Science*. (n.d.). Retrieved June 1, 2023, from https://365datascience.com/tutorials/machine-learning-tutorials/decision-trees/

[32] *What is a Decision Tree | IBM*. (n.d.). Retrieved June 12, 2023, from https://www.ibm.com/topics/decision-trees

[33] Schaffer, C., & Edu, S. A. H. C. (1993). Selecting a classification method by cross-validation. *Machine Learning 1993 13:1*, *13*(1), 135–143. https://doi.org/10.1007/BF00993106

[34] Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, 83–106. https://doi.org/10.1016/B978-0-12-818366-3.00005-8

[35] Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0B013E3181EC173D

[36] Zhang, C., Zhao, J., Zhu, Z., Li, Y., Li, K., Wang, Y., & Zheng, Y. (2022). Applications of Artificial Intelligence in Myopia: Current and Future Directions. In *Frontiers in Medicine* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fmed.2022.840498

[37] Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, 83–106. https://doi.org/10.1016/B978-0-12-818366-3.00005-8

[38] Alsaadi, A. (2019). *Overview of Neural Networks*. https://www.researchgate.net/publication/332655457

[39] Saffari, A., & Khishe, M. (n.d.). *Classification of marine mammals using trained multilayer perceptron neural network with whale algorithm developed with fuzzy system*.

[40] John', S., Labrador, C., Younas, A., & Ali, P. (2021). Understanding and interpreting regression analysis. *Evidence-Based Nursing*, *24*(4), 116–118. https://doi.org/10.1136/EBNURS-2021-103425

[41] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623/SUPP-1

[42] *sklearn.ensemble.IsolationForest — scikit-learn 1.2.2 documentation*. (n.d.). Retrieved May 24, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

[43] Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

[44] *sklearn.model_selection.train_test_split — scikit-learn 1.2.2 documentation*. (n.d.). Retrieved June 1, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[45] *sklearn.model_selection.GridSearchCV — scikit-learn 1.2.2 documentation*. (n.d.). Retrieved June 1, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[46] *KerasTuner*. (n.d.). Retrieved June 2, 2023, from https://keras.io/keras_tuner/

[47] *Factor de emisión de la energía eléctrica: el mix eléctrico. Cambio climático*. (n.d.). Retrieved June 2, 2023, from https://canviclimatic.gencat.cat/es/actua/factors_demissio_associats_a_lenergia/

Annex

All the code used for this project can be found at the following link: https://github.com/Nidafaroogq/TFG/tree/main

A1. DataAP variables

Variable	Meaning	Variable	Meaning
“Pesokg”	Weight	"Total_MTisular_g"	Total tissue mass in grams
"IMC”	Body Mass Index	"Androide_FFM_g"	Fat free mass around the waist in grams
"Cinturacm"	Diameter of the waist in cm	"Tronco_Mtotal_g"	Total mass in the torso in grams
"Tronco_MG_g"	Torso fat mass in grams	"Total_Mtotal_g"	Total mass of the body in grams
"Androide_MG_g"	Fat mass around the waist in grams	"TroncoRegión_Grasa"	Amount of fat in % identified in the trunk region
"Total_MG_g"	Total fat mass in grams	“AndroideRegión_Grasa”	Amount of fat in % identified around the waist region
"Androide_MM_g"	Muscle mass around the waist in grams	“TroncoTejido_Grasa”	Amount of fat tissue in % identified in the trunk, result of dividing fat mass by tissue mass.
"Tronco_MTisular_g"	Torso tissue mass in grams	“AndroideTejido_Grasa”	Amount of fat tissue in % identified in the waist, result of dividing fat mass by tissue mass.
"Androide_MTisular_g"	Tissue mass around the waist in grams		

A2. DataPHY variables

Variable	Meaning	Variable	Meaning
“Pesokg”	Weight	"COLT"	Total cholesterol
"TAS”	Peak systolic blood pressure	"TG"	Triglycerides
"TAD"	Minimum diastolic pressure.	"LDL"	Low-density lipoprotein
"DM2"	Type 2 Diabetes	"HDL"	High-density lipoprotein
"HTA"	Arterial hypertension	"Plaq"	Platelets

"PCR_US"	Ultrasensitive c-reactive protein	"AST"	Aspartate aminotransferase
"GB"	Basal glucose	“ALT”	Alanine aminotransferase
"HbA1c"	Glycosylated hemoglobin	“GGT”	Gamma-glutamyl transpeptidase