International Journal of Applied Science and Engineering

Prediction of metabolic ageing in higher education staff using machine learning: A pilot study

Pineda-Rico Zaira [1*], Rojas Mendoza Diana Luz de los Angeles [1], Pineda-Rico Ulises [2], Arguelles Ojeda Jose Luis [1], Martinez Lopez Francisco Javier [1]

[1] *Coordinacion Academica Region Altiplano, Universidad Autonoma de San Luis Potosi, Matehuala, S.L.P, 78700, Mexico*
[2] *Facultad de Ciencias, Universidad Autonoma de San Luis Potosi., Privadas del Pedregal, 78295, Mexico*

ABSTRACT

The detection of individuals with obesity or overweight allows to predict the prevalence of health risks, such as premature death, disabilities and other chronic diseases. This study describes a pilot conducted on the members of a higher education staff in the city of Matehuala, Mexico. It involved processing anthropometric measurements, health indicators and the results of bioelectrical impedance analysis using machine learning techniques. The goal was to identify the metabolic aging of individuals. The recorded data were used to create a database that was subsequently employed in four different classification models: decision tree, random forest, artificial neural networks and adaptive boosting. Additionally, four statistical techniques were utilized to determine variable importance scores: Pearson, Chi^2, Anova, recursive elimination method and the variance inflation factor. The variable importance score was employed to identify the features that were most consistently repeated across methods. This analysis concluded that both anthropometric measurements and the results of bioelectrical impedance analysis provide valuable references for identifying obesity and overweight in individuals. Among the anthropometric measurements that exhibited a greater impact on the models' predictions were waist-to-height ratio, hip and arm circumferences, body mass index, systolic and diastolic blood pressure and heart rate. Additionally, body fat and muscle mass also contributed significantly.

Keywords: Classification, Machine learning, Obesity prediction, Variable importance.

OPEN ACCESS

Received: July 18, 2023
Revised: August 25, 2023
Accepted: September 18, 2023

Corresponding Author:
Pineda-Rico Z.
zaira.pineda@uaslp.mx

Publisher:
Chaoyang University of Technology
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

1. INTRODUCTION

In 2022, the World Health Organization (WHO) identified that there are more than 1 billion people in the world with obesity, of which 650 million are adults. In Mexico, obesity is a significant concern. The 2020 National Health and Nutrition Survey on Covid-19 found that 76% of adult women are overweight or obese, compared to 72.1% of men (Shamah-Levy et al., 2021). Mexican adults aged between 29 and 69 engage in approximately 300 min per week of moderate to vigorous physical activity. However, about 29% do not meet the minimal recommendation of 150 min per week (Medina et al., 2013).

On the other hand, nearly 50% of Mexican adults have metabolic syndrome due to sedentary behaviors, physical inactivity, unhealthy dietary habits and poor sleep patterns (Macias et al., 2021). A study on the sociodemographic and anthropometric characteristics of adults aged 20 to 69 years in Mexico City revealed that the prevalence of participants classified in the highest sitting time category (≥ 420 min/day) increased by 8% over nine years. This increase had an impact, leading to a rise of 5.4% in overweight/obesity and a 1.3% increase in the diagnosis of diabetes (Medina et al., 2017).

https://doi.org/10.6703/IJASE.202312_20(4).009

In higher education institutions, staff work activities are predominantly centered in offices, inducing sedentary behaviors. Given this, monitoring the health condition of the staff is crucial for detecting potential health risks that could contribute to the development of chronic diseases.

Previous studies focused on higher education staff have shown that poor nutritional habits and lack of physical activity promotes the prevalence of overweight/obesity. Consequently, it is important to implement strategies aimed at reducing obesity and promoting well-being among the teaching population (Rodriguez-Guzman, 2006; Freedman et al., 2010; He et al., 2014; Rodrigues-Rodrigues et al., 2018).

According to the WHO, obesity and overweight can be identified from the anthropometric measurements of an individual. For adults, body mass index (BMI) and waist circumference (WA) serve as reliable indicators to discern obesity and overweight. Anthropometric measurements encompass a variety of body metrics, including weight, height, standing length, skin folds, circumferences (head, waist, hip, etc.), length of limbs and widths (shoulder, wrist, etc.). The Official Mexican Standard (NOM) defines the parameters and anthropometric criteria considered to determine abdominal obesity within the Mexican population (Shamah-Levy et al., 2017). Table 1 shows the values of BMI and WA values used for classifying obesity in Mexican adults.

In this study, we aimed to determine the prevalence of obesity and overweight among the staff of a higher education institution situated in the city of Matehuala, Mexico. The assessment was based on BMI and WA measurements. Considering WA, our observations indicate that 72% of males and 60.5% of females among the staff members are classified as abdominally obese. Evaluating the BMI of the observed group, we found that 60% of males and 31.5% of females are categorized as overweight, while 16% of males and 5.2% of females fall under the classification of obesity. Among those identified as obese, 12% of males and 5.2% of females belong to the obese class I category, and 4% of males are in the obese class II category, while no females fall within this category. None of the individuals belong to the obese class III category. When considering the entire monitored staff as a collective, 65% are abdominally obese, 42.8% are overweight, and 9.5% are obese. In total, 52.3% of the observed group are categorized as overweight or obese.

While the utilization of BMI, WA and waist-to-height ratio (WHtR) for predicting mortality remains effective, an alternative approach involves analyzing the outcomes of bioelectrical impedance analysis (BIA). This method has been suggested for monitoring and tracking the health status of individuals, including those with chronic conditions such as obesity (Ricciardi and Talbot, 2007; Heydari et al., 2011; de-Mateo-Silleras et al., 2019; Aldobali et al., 2022). BIA results have previously been utilized to estimate body fat percentage (BF) and correlate it with assessing the risk of diseases or mortality (Böhm and Heitmann, 2013). In 2021, the significance of the association between fat, visceral fat (VF) and muscle mass (MM) obtained through BIA in identifying metabolic syndrome as a health concern was recognized (Pouragha et al., 2021).

Machine learning (ML) is the science of programming computers so that they can learn from the information that has been provided to them (Geron, 2019). There are multiple ML techniques that can be used to build projects related to healthcare, with the aim of improving medical diagnosis or assisting health staff in the process of identifying a patient's condition (Sprogar et al., 2001; Javaid et al., 2022; Manickam et al., 2022; Payal et al., 2022). Common ML techniques used for these purposes are DTs, Logistic Regression (LR) and Support Vector Machine (SVM). Classification algorithms are effective to predict syndromes related to the prevalence of overweight and obesity (Chatterjee et al., 2020; Gutierrez-Esparza et al., 2020; Safaei et al., 2021; Crowson et al., 2022; Dhabarde et al., 2022; Strzelecki and Badura, 2022).

The classification process uses the features that have the biggest impact on the prediction of the objective. The selection of feature importance is very relevant, especially in classification problems with few samples (Mohd and Awang, 2021). In Archer and Kimes (2008), the authors evaluate the effectiveness of using the variable importance score in the Random Forest (RF) technique, concluding that this methodology is applicable in classification problems when the objective is to produce an accurate classifier. Also using RF, in Chen et al. (2020) different methods are

Table 1. Obesity classification by BMI and WA in Mexican adults, according to the Official Mexican Standard (NOM) and the WHO

Source	Underweight	Normal	Overweight		Obesity	
				Class I	Class II	Class III
WHO	< 18.5	18.5–24.9	25.0–29.9	30.0–34.9	35.0–39.9	> 40.0
NOM			25.0–29.9 ≥ 23 and < 25 in low height adults	≥ 30 or ≥ 25 in low height adults		
	Abdominal obesity according to the Official Mexican Standard					
Male	≥ 90 cm					
Female	≥ 80 cm					

BMI = Actual weight (kg)/ height (m) * Low height = Less than 1.50 meters in adult female and less than 1.60 meters in adult male. Source: INSP (2018).

presented in order to reduce the number of features based on the identification of the variable importance measures (VIMs); the authors evaluated and compared the accuracy of specific RF, SVM, K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) classification models. Additionally, in Gregorutti et al. (2017) and Senan et al. (2021), Recursive Feature Elimination (RFE) is used to identify VIMs. Misra and Singh Yadav (2020) suggest that a less complex algorithm can improve the accuracy of the classification, so they propose a method that analyzes each of the features and registers its importance with a predictor variable. Supporting the statement that using several methods to obtain VIMs in a classification problem, offers more reliability and consistency in the classification of the objective (Kiang, 2003; Nithya and Ilango, 2019). The selection of features has helped to obtain important results in ML biomedical applications. For example, Gutierrez-Esparza et al. (2020) used VIMs in the prediction of metabolic syndrome in a Mexican population. McLaren et al. (2019) used VIMs to predict malignant lesions in the breast with magnetic resonance imaging as features. Also, Ganggayah et al. (2019) used VIMs to identify the factors that predict the survival of patients with breast cancer.

Wilson et al. (2012) and Sparling et al. (2007) state that the factors contributing to overweight/obesity are diverse and require a comprehensive approach that takes into account environmental and cultural influences. They also emphasize the significance of early intervention in effectively reducing rates of overweight and obesity.

The problem addressed in this paper centers on the high prevalence of obesity and overweight. In Mexico, either the rate of obesity and the prevalence of metabolic syndrome are alarming, potentially leading to the suffering of long-term health conditions. The sedentary lifestyle, unhealthy dietary habits and poor sleep patterns among Mexican adults contribute to the high rates of obesity and metabolic syndrome. The problem is worsened by the increasing prevalence of overweight/obesity among higher education staff, who are primarily engaged in sedentary activities.

The motivation behind this study is based on the need to reduce the prevalence of obesity and metabolic syndrome among higher education staff in Mexico. By utilizing ML techniques, the paper aims to contribute to the development of effective strategies for identifying health risks and promoting wellbeing. The study's focus on higher education staff underlines the importance of creating interventions tailored to specific work environments to mitigate the adverse health effects associated with sedentary behaviors.

In the present study we use data obtained in a health condition monitoring initiative involving the staff of a higher education institution situated in Matehuala, Mexico. The aim is to identify health risks through the application of ML techniques. The features include individual records comprising anthropometric measurements, glucose levels, and results obtained from BIA. Python and Scikit Learn were used to implement four classification algorithms based on ML, and four statistical techniques, that helped to compute VIMs of the features in the prediction of the individual's risk of having obesity, by observing the body age or metabolic age.

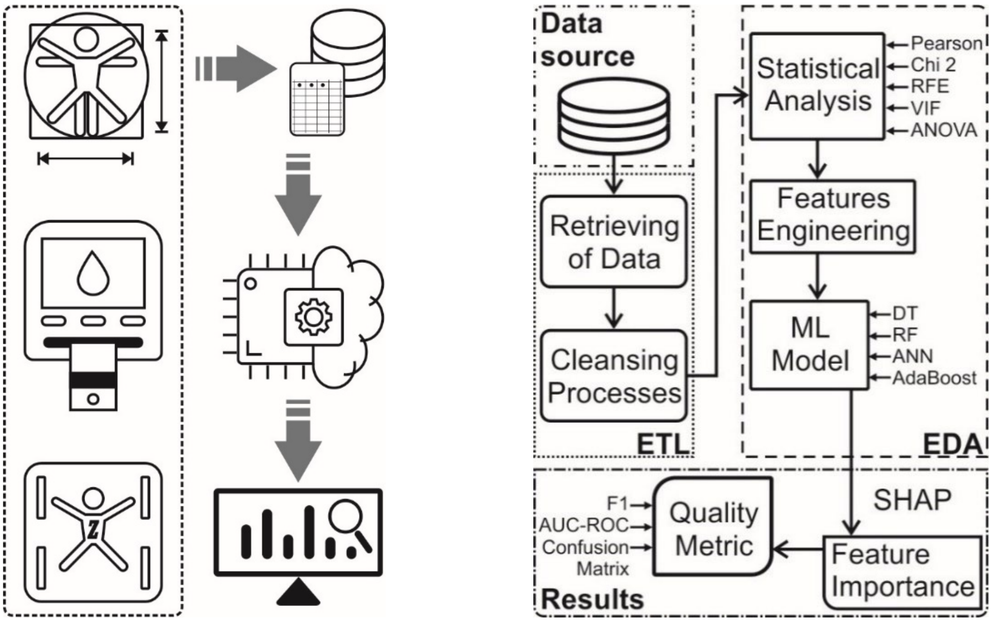
2. MATERIALS AND METHODS

In biomedical applications based on supervised learning, medical data are used to train the algorithm in accordance with its relation to the target. In the present pilot study, a database was created with 63 records, identifying anthropometric measurements, glucose levels and BIA results as features (Fig. 1(a)). Fig. 1(b) depicts the block diagram representing the process flow: during the Extract Transform Load (ETL) phase, data is retrieved from the database and cleaned by replacing missing values with the computed mean. Additionally, feature scaling is performed at this stage. In the Exploratory Data Analysis (EDA) phase, statistical analyses are conducted using the univariate methods (Pearson, Chi^2 and ANOVA), and both RFE and variable importance factor (VIF) methods. Within the ML model block, the following classifiers are implemented: DT, RF, artificial neural networks (ANN) and adaptive boosting (AdaBoost), aiming to obtain VIMs through Shapley additive explanations (SHAP) values. For quality assessment of the classifiers, the F1 score, the Area Under the Receiver Operating Characteristic curve (AUC-ROC), and the confusion matrix were utilized as metrics.

The recorded data include the following anthropometric measurements: age (AG), weight (WE), height (HE), BMI, WA, WHtR, arm circumference (AR), hip circumference (HP), systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate (HR). Additionally, the health indicator includes glucose (DX), and the following functional fitness parameters: MM, VF, body fat (BF) and body age. Ageing (AGG) is defined as the ratio age/body age.

Weight, BMI, MM, MA, VF, BF and body age were derived from the BIA results obtained using an Omron HBF-514C body monitor. This device sends electrical currents through the hands via electrodes that the individual holds with both hands and through the feet via electrodes placed on the scale's surface. This combination allows for an analysis of both the upper and lower body (Pribyl et al., 2011). Participants were instructed not to exercise and to fast on the test day, including refraining from coffee. Blood pressure (BP) was measured using an inflatable cuff with a gauge around the arm, providing measurements in millimeters of mercury (mmHg) for DBP and SBP. Waist (WA), hip (HP) and arm (AR) circumferences were measured using a tape measure in centimeters. Heart rate (HR) or pulse was measured at the wrist on the radial artery in beats per minute.

The WHtR is calculated by dividing waist by height measurement in centimeters. GLU measurements were taken using a blood sugar meter, with blood samples collected from fingertip pricks, reported in millimoles per liter (mmol/L). The status of "aged" was utilized as the objective or label, determined based on the ratio between body age, and the subject's real age. Specifically, if AGG > 1.0, the subject is considered "aged." Table 2 displays the anthropometric measurements, blood glucose levels, and functional physical fitness indices obtained by BIA for the staff members. The data is presented with average values, standard deviations, as well as maximum and minimum values for each characteristic. On average, the staff members are 40 years old, with an average weight of 70 kg and height of 1.65 meters. According to Table 1, the staff is classified as overweight with an average BMI of 25.58 kg/m² (> 25), they exhibit normal glucose levels (< 99) and normal blood pressure (109/73). The BIA results gave an AGG of 1.13, indicating that the staff members are “aged” with a BF percentage of 32.9%.



(a) Anthropometries recording, BIA results and glucose compose the database to be processed with ML techniques and statistical analysis

(b) Data analysis block diagram that includes the ETL, the EDA and the results

Fig. 1. Block diagram that describes the work process

Table 2. General description of the anthropometries, blood glucose measurements, and functional physical fitness indices obtained by BIA for the staff members

Code	Mean	Standard deviation	Minimum	Maximum
AG	40.31	7.60	26	56
		Anthropometries		
WE	70.06	13.72	51.9	105.1
HE	165.04	9.98	149	189
BMI	25.58	3.66	20.5	35
WA	88.01	11.01	69	116
WHtR	0.53	0.06	0.44	0.7
AR	28.23	2.76	23	35
HP	103.37	5.98	95	120
SBP	109.40	13.03	90	145
DBP	73.13	10.33	60	90
HR	70.64	9.43	50	95
		Health indicator		
DX	95.81	11.87	69	130
		Functional fitness		
MM	29.11	5.56	20.5	41
AGG	1.13	0.25	0.6	1.7
VF	7.62	3.85	2	19
BF	32.96	7.28	20.5	49.7

2.1 Machine Learning

One of the goals of applying ML techniques to large datasets is to discover patterns among features. As listed above, some of the most important algorithms used in supervised learning are SVM, DT and RF. SVM and DT can be used for classification, and regression tasks on complex datasets, while the RF algorithm is built from many individual DT. DT learn the best way to divide the training dataset into smaller and even more smaller subsets until reaching the target prediction. In RF, the predictions from all the trees are used to make the final prediction of the target.

In ML algorithms, the relative importance of each feature is scored after training the algorithm. This method is helpful to get a better understanding of which characteristics are more important when a selection of features is required, in addition "to discovering complex relationships between predictors corresponding to interaction terms". In these algorithms, variable importance can be measured by observing the decrease in model accuracy if the values of a variable are randomly permuted (Peter et al., 2020). In this work, the following models were used: DT, RF, ANN, AdaBoost.

2.2 Statistical Analysis

In addition to the ML algorithms, other methods exist to identify VIMs. We implemented univariate analysis, the RFE method and the VIF calculation.

The univariate analysis method involves analyzing each variable in the dataset using Pearson, Chi^2 and ANOVA correlation tests. The value of 'p' is used as a criterion to determine the degree of importance of each characteristic. The Chi^2 correlation test determines whether the variables are related to the objective. The RFE method employs a ML model to iteratively remove variables with the least impact on the target prediction. Various models can serve as a basis for this technique, such as linear, SVM, DT, among others. The VIF factor provides a measure of collinearity that assesses whether two variables in the model are highly correlated and conveys similar information about the dataset's variance. In multiple regression, this helps to identify the most significant predictor variables.

Python and Scikit-Learn were utilized for the statistical analysis of the dataset, data processing, and modeling using ML techniques. Data normalization was performed before conducting the data analysis. For the classification modeling, the dataset was randomly split into two subsets: the training dataset (80% of the data) and the testing set (20% of the data). The prediction of whether an individual is aged or not aged was based on the AGG ratio.

3. RESULTS AND DISCUSSION

Table 3 shows the characteristics of the population by group: aged (AGG > 1.0) or not aged (AGG ≤ 1.0). According to the table, 60.3% of the population had a body age of 1.27 years older than the mean age of the staff. This group have an average weight of 74.37 kg, a BMI of 26.94 kg/m^2 and BF of 33%. Likewise, 39.6% of the population had a body age of 1.91 years younger than the mean age of the staff; with an average weight of 63.52 kg, a BMI of 23.51 kg/m^2 and BF of 32.88%. Both groups display normal glucose levels (< 99) and normal blood pressure (< 120/80).

The Pearson correlation coefficient was used to compare the characteristics of the two groups, the p value varies between 1 and -1 with 0 indicating that there is no correlation. The values of the anthropometries WE, BMI, WHtR, AR, HP, SBP and DBP are higher for the aged population, as well as the DX health indicator; and the measurements of MM, VF and BF obtained by BIA.

3.1 Classification Models

The F1-score from Scikit-Learn was used as a measure of accuracy for the classification tasks. The score is normalized, a value approaching 1.0 indicates the best performance. The accuracy of all the classification models was above 0.9, as follows: DT (1.0), RF (0.923), ANN (0.923), AdaBoost (1.0). Additionally, as a measure of performance, the AUC-ROC was computed for each classification model. A value close to 1.0 implies that the model is accurate. The ROC Curve is shown in Fig. 3. The AUC of all the classification models was above 0.9: DT (1.0), RF (0.9), Artificial ANN (0.95), AdaBoost (1.0).

Table 3. Anthropometries, blood glucose measurements and functional physical fitness indices of the staff members according to their classification as aged (AGG > 1.0) or not aged (AGG ≤ 1.0)

Code	Yes (N = 38 (60.3%))	No (N = 25 (39.6%))	p value
AG	41.42 (26–56)	38.40 (32–49)	0.0039
	Anthropometries		
WE	74.37 (56.3–105.1)	63.52 (51.9–102)	0.0000
HE	165.57 (149–189)	164.24 (150–178)	0.0000
BMI	26.94 (21.9–35)	23.51 (20.5–37.1)	0.0010
WA	92.02 (78–116)	82.24 (69–108)	0.0000
WHtR	0.55 (0.44–0.69)	0.5 (0.45–0.7)	0.0000
AR	29.31 (24–35)	26.6 (23–34)	0.0003
HP	105.1 (97–120)	100.89 (95–117)	0.0000
SBP	110.29 (90–145)	108.2 (90–140)	0.0000
DBP	74.26 (60–90)	71.6 (60–90)	0.0000
HR	70.78 (50–87)	70.44 (50–95)	0.0000
	Health indicator		
DX	98.11 (69–118)	92.37 (75–130)	0.0000
	Functional fitness		
MM	30 (20.5–41)	27.77 (20.5–37.1)	0.0011
AGG	1.28 (1.03–1.70)	0.88 (0.6–1)	0.0003
VF	8.78 (3–17)	5.92 (2–19)	0.0006
BF	33.01 (20.7–49.3)	32.88 (20.5–49.7)	0.0020

A confusion matrix was also used to visualize the specific accuracy for each class (aged or not aged). The confusion matrix helped to identify that all the models classified correctly 100% of the aged individuals. In addition, the RF and the ANN models classified correctly only 80% of the not aged labels, with the rest of the models scoring 100%.

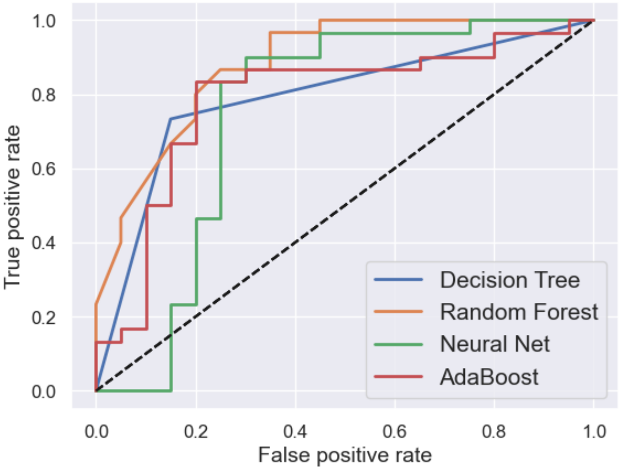


Fig. 3. ROC curve of the different classification models

3.2 Variable Importance

Since DT, RF and AdaBoost performed better among the prediction models based on ML, SHAP values were used to identify the importance of each feature and its impact on the prediction. A SHAP value of zero indicates little contribution to the prediction, so the further from zero declares higher contribution. For the Pearson and Chi^2 correlation tests, Anova, RFE method and the VIF factor, the features are ordered by importance according to the scores given by the statistical analysis. Table 4 shows the top nine features characterizing metabolic age as a result of applying each method.

The anthropometric measurements that exhibit a greater impact on the prediction of obesity and overweight include WHtR, SBP and DBP, HP and AR, HR and BMI. Additionally, the BIA results, specifically BF and MM, also show a significant impact on the models' predictions.

Within the complete dataset, the classification models identified the most significant features as BMI, BF, WHtR, DBP and HE. Meanwhile, the statistical methods highlighted BF, SBP, WHtR and HP as the most important features. The features that demonstrated greater significance in predicting metabolic aging, considering the eight proposed methods, include BMI, BF and WHtR.

These findings align with previous studies conducted on populations of various ethnicities, where BMI, WA and WHtR are suggested indicators for assessing abdominal obesity and cardiometabolic risk. However, the authors of these studies have acknowledged certain limitations when using each parameter separately.

Devajit and Haradhan (2023) studied BMI as one of the most popular anthropometric tools to measure body fitness in order with the intention of uncovering its constraints in accurately assessing obesity in individuals of different ethnicity. The authors found that does not capture effectively and proficiently status of overweight/obesity across all populations, regardless of sex, age, socio-economic standing, and ethnic background. Ashwell et al. (2011) completed a study on individuals with different ethnicity, about the utilization of WHtR in detecting abdominal obesity, along with the possible health risks associated with it. The study's results indicate that WHtR surpasses WA as a more accurate predictor for diabetes, dyslipidemia, HR, and the risk of cardiovascular disease; and that abdominal obesity offers more effective instruments for discerning cardiometabolic risks linked to obesity compared to BMI. On the other hand, previous research has studied the relation of BF with obesity and metabolic AGG. Sandeep et al. (2010), produced comprehensive gene expression profiles across both visceral and subcutaneous fat stores in Asian Indian individuals with and without diabetes. Additionally, the researchers assessed multiple intermediary phenotypic traits related to diabetes, including distinct anthropometric attributes, indicators of insulin resistance and secretion, glycemic control metrics, distribution of BF, among others. The authors conclude that adipose tissue pathology is linked to diabetes in both subcutaneous and VF deposits holding a crucial role in the development of metabolic syndrome.

In regards of BF as an indicator for obesity, Jensen (2008) study the roles of distinct fat deposits concerning the storage and release of fatty acids in both healthy individuals and

Table 4. Importance of the features characterizing metabolic ageing by method.

	BMI	BF	HE	WHtR	DBP	SBP	VF	AR	HE	HP	MM	HR	AG	WA	DX	SEX	WE
1	*	*	*														
2	*			*	*	*	*	*	*	*	*						
3	*	*		*	*				*			*	*	*			*
4	*			*			*	*		*			*	*	*		*
5		*	*	*		*		*		*	*	*				*	
6		*	*		*	*	*			*	*	*			*		
7	*	*		*	*	*		*			*		*			*	
8	*	*		*		*			*	*		*		*			*
9	6	6	3	6	4	5	3	4	3	5	4	4	3	3	2	2	3

1. DT, 2. RF, 3. AdaBoost, 4. Pearson, 5. Chi^2, 6. Anova, 7. RFE, 8. VIF, 9. Repetition count

https://doi.org/10.6703/IJASE.202312_20(4).009

those with obesity; with the aim to discuss the disagreement regarding to the fact that upper body or visceral obesity increases the risk for conditions such as type 2 diabetes and that elevated quantities of lower BF are independently linked to a decreased risk of metabolic issues. Also, that VF mass has a more pronounced correlation with an abnormal metabolic profile compared to subcutaneous fat in the upper body. The results concludes that abdominal fat accumulation in individuals with overweight is highly associated with the metabolic complications of obesity.

Previous studies performed on Mexican population also discuss the use of BMI, WA, BF and WHtR as indicators for obesity. Sanchez Soto et al. (2012) found that 80% of people with obesity had high percentage of BF.

In Gutierrez-Esparza et al. (2020), the authors used ML algorithms to prioritize health parameters, aiming to identify the most suitable variables for classifying Metabolic Syndrome (MetS) within the Mexican population of the city of Tlalpan. They used Correlation-based Feature Selection (CFS) and Chi^2 filter methods to identify pertinent features for diagnosing MetS. In their results, WHtR, coupled with the Adult Treatment Panel III (ATP III) variables (excluding waist measurement), outperforms WAIST and BMI in terms of classification accuracy, in the prediction of metabolic syndrome in Mexican population.

In Barquera et al. (2020), the authors analyzed the data of 16,256 individuals to study the prevalence of obesity among Mexican adults while considering various physical and sociodemographic factors, and subsequently, to assess trends in these prevalence rates over time. The classification considered obesity (according to WHO standards), abdominal adiposity (as per IFD criteria), and short stature (following NOM-008-SSA3-2017). The researchers used LR models to identify the correlation between obesity and various risk factors. The results showed that heigh plays an important role in identification of obesity in Mexican women and men, although it was more notorious in women, along with WA as a complementary index that allows the evaluation of VF accumulation. The authors recognize BMI as an indicator of the risk of comorbidities associated with excessive adipose tissue, although, they state that this indicator is not very accurate for assessing adiposity at an individual level.

BMI, WHtR, WA and BF are useful to assess cardiovascular disease risk, metabolic syndrome and obesity. Also, BMI is a relevant predictor associated with mortality due to chronic kidney disease and cardiovascular peril in diabetic patients (Sanabria-Arenas, 2015; Mendoza-Niño et al., 2023; Russo et. al., 2023).

The findings emerging from this investigation could offer valuable insights for shaping healthcare initiatives for Mexican population, especially those working in higher education institutions. Including the staff's behaviors in future studies, such as sedentary lifestyles, reduced sleeping hours, lack of health awareness and long working hours, may enhance the efficiency of healthcare supervision and the design of strategies for supervision, preemptive measures and active involvement.

4. CONCLUSION

Sedentary behaviors in people can lead to obesity or overweight. Therefore, monitoring an individual's health condition is essential for detecting potential health risks that could progress into chronic diseases. This work described the results of data modelling focused on anthropometric measurements collected from members of a higher education staff. The anthropometric measurements included age, waist, hip and ARs; heigh, BP, HR, BMI, among others. Additionally, the results of BIA such as BF, VF, MM and body age were incorporated. The health indicator glucose was also considered. These parameters were used as features in four classification models. Also, the data was analyzed using the univariate method, RFE and VIF. The objective was to determine the variable importance to identify which features played a more crucial role in predicting metabolic aging within the group.

The contributions of this work that collectively enrich the understanding of obesity, its assessment, and its links to metabolic aging, particularly within the Mexican population and higher education staff are:

- An in-depth analysis of a specific population's health characteristics is provided. A detailed statistics about the population's body age in relation to their mean age, along with their average weight, BMI, BF percentage, glucose levels and BP is presented. This comprehensive exploration emphasize the variations and potential health implications within the studied population.
- A correlation analysis using Pearson correlation coefficients to identify relationships between various characteristics of the population is conducted. This analysis reveals which attributes are positively or negatively correlated and offers insights into potential connections between different health indicators.
- The performance of different ML classification models for predicting metabolic aging is evaluated. The F1-score and AUC-ROC as measures of accuracy and performance are applied. All classification models performed an excellent discrimination, achieving high accuracy scores (above 0.9): DT (1.0), RF (1.0), ANN (0.923), AdaBoost (1.0). A ROC curve is also provided to visualize the accuracy of each model, supporting the effectiveness of ML techniques in predicting metabolic aging.
- The SHAP values to interpret the importance of features in the prediction models was introduced. They are used to measure the impact in the prediction for each feature and the results are compared to find coincidence to the variable importance obtained from the statistical methods. Both anthropometric measurements and the results of BIA provide valuable references for identifying obesity and overweight in individuals.

- The study employs various statistical methods for feature selection, such as Pearson and Chi^2 correlation tests, Anova, RFE and VIF. By comparing the top features selected by each method, the study displays the robustness and consistency of the identified important features, contributing to a more comprehensive understanding of the factors influencing metabolic aging.
- Specific features that hold the most significance in predicting metabolic aging across different methods are identified. The nine features that consistently appeared among the eight methods were BMI, BF, WHtR, diastolic blood pressure, systolic blood pressure, hip and ARs, MM and heart ratio. Out of these, the top five features were BMI, BF, WHtR, systolic blood pressure and HP. Anthropometric measurements like BMI, BF and WHtR as consistently influential in predicting metabolic aging.

REFERENCES

Aldobali, M., Pal, K., Chhabra, H. 2022. Noninvasive health monitoring using bioelectrical impedance analysis. Computational Intelligence in Healthcare Applications, 209–236.

Archer, K.J., Kimes, R.V. 2008. Empirical characterization of random forest variable importance measures. Computational Statistics and Data Analysis, 52, 2249–2260.

Ashwell, M., Gunn, P., Gibson, S. 2011. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: Systematic review and meta-analysis. Obesity Reviews, 13, 275–286.

Barquera, S., Hernandez-Barrera, L., Trejo-Valdivia, B., Shamah, T., Campos-Nonato, I., Rivera-Dommarco, J. 2020. Obesidad en México, prevalencia y tendencias en adultos. Ensanut 2018-19. Salud Pública de México, 62, 682–692.

Bohm, A., Heitmann, B.L. 2013. The use of bioelectrical impedance analysis for body composition in epidemiological studies. European Journal of Clinical Nutrition, 67, 79–85.

Chen, R.C., Dewi, C., Huang, S.W., Caraka, R.E. 2020. Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7, 1–26.

Crowson, M.G., Moukheiber, D., Arevalo, A.R., Lam, B.D., Mantena, S., Rana, A., Goss, D., Bates, D.W., Celi, L. A. 2022. A systematic review of federated learning applications for biomedical data. PLOS Digital Health, 1, 1–14.

Chatterjee, A., Gerdes, M.W., Martinez, S.G. 2020. Identification of risk factors associated with obesity and overweight-a machine learning overview. Sensors (Basel), 20, 2734.

de-Mateo-Silleras, B., de-la-Cruz-Marcos, S., Alonso-Izquierdo, L., Camina-Martín, M.A., Marugán-de-Miguelsanz, J.M., Redondo-Del-Río, M.P. 2019. Bioelectrical impedance vector analysis in obese and overweight children. PLoS One. 14, e0211148.

Devajit, M., Haradhan, K.M. 2023. Body mass index (BMI) is a popular anthropometric tool to measure obesity among adults. Journal of Innovations in Medical Research, 2, 25–33.

Dhabarde, S., Mahajan, R., Mishra, S., Chaudhari, S., Manelu, S., Shelke, N.S. 2022. Disease prediction using machine learning algorithms. 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22), 1–4.

Freedman, M.R., Rubinstein, R.J. 2010. Obesity and food choices among faculty and staff at a large urban university. Journal of American College Health. 59, 205–210.

Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P., Dhillon, S.K. 2019. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Medical Informatics and Decision Making, 19, 1–17.

Geron, A. 2019. Hands-on machine learning with scikit-learn and TensorFlow. O'Reilly Media. U.S.A.

Gregorutti, B., Michel, B., Saint-Pierre, P. 2017. Correlation and variable importance in random forests. Statistics and Computing, 27, 659–678.

Gutierrez-Esparza, G.O., Vazquez, O.I., Vallejo, M., Hernandez-Torruco, J. 2020. Prediction of metabolic syndrome in a Mexican population applying machine learning algorithms. Symmetry, 12. 581–596.

He, L., Ren, X., Qian, Y., Jin, Y., Chen, Y., Guo, D., Yao, Y. 2014. Prevalence of overweight and obesity among a university faculty and staffs from 2004 to 2010, China. Nutrición Hospitalaria, 29, 1033–1037.

Heydari, S.T., Ayatollahi, S.M., Zare, N. 2011. Diagnostic value of bioelectrical impedance analysis versus body mass index for detection of obesity among students. Asian Journal of Sports Medicine. 2, 68–74.

Javaid, M., Haleem, A., Singh, R.P., Suman, R., Rab, S. 2022. Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks, 3, 58–73.

Jensen, M.D. 2008. Role of body fat distribution and the metabolic complications of obesity. Journal of Clinical Endocrinology and Metabolism, 93, 57–63.

Jiang, M., Yin, S. 2023. Facial expression recognition based on convolutional block attention module and multi-feature fusion. International Journal of Computational Vision and Robotics. 13, 21–37.

Jiang, Y., Yin, S. 2023. Heterogenous-view occluded expression data recognition based on cycle-consistent adversarial network and K-SVD dictionary learning under intelligent cooperative robot environment. Computer Science and Information Systems, 34.

Kiang, M.Y. 2003. A comparative assessment of classification methods. Decision Support Systems, 35, 441–454.

Laghari, A.A., Yin, S. 2022. How to collect and interpret medical pictures captured in highly challenging environments that range from nanoscale to hyperspectral imaging. Current Medical Imaging, 54, 36582065.

Laghari, A.A., He, H., Shafiq, M., Khan, A. 2018. Assessment of quality of experience (QoE) of image compression in social cloud computing. Multiagent Grid Systems, 14, 125–143.

Laghari, A. A., Shahid, S., Yadav, R., Karim, S., Khan, A., Li, H., Yin, S. 2023. The state of art and review on video streaming. Journal of High Speed Networks, (Preprint), 1–26.

Macias, N., Espinosa-Montero, J., Monterrubio-Flores, E., Hernandez-Barrera, L., Medina-Garcia, C., Gallegos-Carrillo, K. 2021. Screen-based sedentary behaviors and their association with Metabolic Syndrome components among adults in Mexico. Preventing Chronic Disease, 18, 1–12.

Manickam, P., Mariappan, S.A., Murugesan, S.M., Hansda, S., Kaushik, A., Shinde, R., Thipperudraswamy, S.P. 2022. Artificial intelligence (AI) and internet of medical things (IoMT) assisted biomedical systems for intelligent healthcare. Biosensors, 12, 562–591.

McLaren, C.E., Chen, W.P., Nie, K., Su, M.Y. 2009. Prediction of malignant breast lesions from MRI features: A comparison of artificial neural network and logistic regression Techniques. Academic Radiology, 16, 842–851.

Medina, C., Janssen, I., Campos, I., Barquera, S. 2013. Physical inactivity prevalence and trends among Mexican adults: Results from the national health and nutrition survey (ENSANUT) 2006 and 2012. BMC Public Health, 13, 1–10.

Medina, C., Tolentino-Mayo, L., Lopez-Ridaura, R., Barquera, S. 2017. Evidence of increasing sedentarism in Mexico City during the last decade: Sitting time prevalence, trends, and associations with obesity and diabetes. Plos One, 12, 1–15.

Mendoza-Niño, C., Martinez-Robles, J.D., Gallardo-Garcia, I. 2023. Relationship between overweight and obesity with the progression of chronic kidney disease in patients at the Naval Medical Center in Mexico. Enfermería Nefrologica, 26, 60–66.

Meng, X., Wang, X., Yin, S. Li, H. 2023. Few-shot image classification algorithm based on attention mechanism and weight fusion. Journal of Engineering and Applied Science, 70, 14.

Misra, P., Yadav, A.S. 2020. Improving the classification accuracy using recursive feature elimination with cross-validation. International Journal on Emerging Technologies, 11, 659–665.

Nafis, N.S.M., Awang, S. 2021. An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification. IEEE Access, 9, 52177–52192.

Nithya, B., Ilango, V. 2019. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. Applied Sciences, 1, 1–16.

Payal, M., Kumar, K.S., Kumar, T.A. 2022. Recent advances of Machine Learning Techniques in Biomedicine. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 5, 772–779.

Peter, B., Bruce, A., Gedeck, P. 2020. Practical Statistics for Data Scientists. 2nd Edition. O'Reilly Media, Inc. U.S.A.

Pouragha, H., Amiri, M., Saraei, M., Pouryaghoub, G., Mehrdad, R. 2021. Body impedance analyzer and anthropometric indicators; Predictors of metabolic syndrome. Journal of Diabetes and Metabolic Disorders, 20, 1169–1178.

Pribyl, M.I., Smith, J.D., Grimes, G.R. 2011. Accuracy of the Omron HBF-500 body composition monitor in male and female college students. International Journal of Exercise Science, 4, 93–101.

Ricciardi, R., Talbot, L.A. 2007. Use of bioelectrical impedance analysis in the evaluation, treatment, and prevention of overweight and obesity. Journal of the American Academy of Nurse Practitioners, 19, 235–241.

Rodriguez-Guzman, L., Diaz-Cisneros, F., Rodriguez-Guzman, E. 2006. Overweight and obesity in teachers. Anales de la Facultad de Medicina, 67, 224–229.

Rodrigues-Rodrigues, T., Viera Gomes, A.C, Rodrigues Neto, G. 2018. Nutritional status and eating habits of professors of health area. International Journal of Sport Studies for Health, 1, e64335.

Russo, M.P., Grande-Ratti, M.F., Burgos, M.A., Molaro, A.A., Bonella, M.B. 2023. Prevalence of diabetes, epidemiological characteristics and vascular complications. Archivos de Cardiología de México, 93, 30–36.

Safaei, M., Sundararajan, E.A., Driss, M., Boulila, W., Shapi'i, A. 2021. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity, Computers in Biology and Medicine, 136, 104754.

Sanabria-Arenas, M., Paz-Wilches, J., Laganis-Valcarcel, S., Muñoz-Porras, F., Lopez-Jaramillo, P., Vesga-Guald, J., Perea-Buenaventura, D., Sanchez-Pedraza, R. 2015. Dialysis initiation and mortality in a population with chronic kidney disease in Colombia. Revista de la Facultad de Medicina, 63, 209–216.

Sanchez Soto, J.M., Martinez Reyes, M., Quintero Soto, M.L., Padilla Loredo, S. 2012. Determinación de obesidad a personal de salud de primer nivel de la Jurisdicción de Nezahualcótotl (México) por medio del índice de masa corporal. Medwave, 12, e5464.

Sandeep, S., Gokulakrishnan, K., Velmurugan, K., Deepa, M., Mohan, V. 2010. Visceral & subcutaneous abdominal fat in relation to insulin resistance & metabolic syndrome in non-diabetic south Indians. Indian Journal of Medical Research, 131, 629–635.

Senan, E.M., Al-Adhaileh, M.H., Alsaade, F.W., Aldhyani, T.H.H., Alqarni, A.A., Alsharif, N., Uddin, M.I., Alahmadi, A.H., Jadhav, M.E., Alzahrani, M.Y. 2021. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering, 2021, 1–10.

Karim, S., Qadir, A., Farooq, U., Shakir, M., Laghari, A.A. 2023. Hyperspectral imaging: A review and trends towards medical imaging. Current Medical Imaging, 19, 417–427.

Shamah-Levy, T., Romero-Martínez, M., Barrientos-Gutiérrez, T., Cuevas-Nasu, L., Bautista-Arredondo, S., Colchero, M.A., GaonaPineda, E.B., Lazcano-Ponce, E., Martinez-Barnetche, J., Alpuche-Arana, C., Rivera-Dommarco, J. 2021. Encuesta nacional de salud y nutrición 2020 sobre Covid-19. Resultados nacionales. Cuernavaca, México: Instituto Nacional de Salud Pública, 135–152.

Shamah-Levy, T., Vielma-Orozco, E., Heredia-Hernández, O., Romero-Martínez, M., Mojica-Cuevas, J., Cuevas-Nasu, L., Santaella-Castell, J.A., Rivera-Dommarco, J. 2020. Encuesta nacional de salud y nutrición 2018-19: Resultados nacionales. Cuernavaca, México: Instituto Nacional de Salud Pública, 171–172.

Das, S., Adhikary, A., Laghari, A. A., Mitra, S. 2023. Eldo-care: EEG with kinect sensor based telehealthcare for the disabled and the elderly. Neuroscience Informatics, 100130.

Sparling, P.B. 2007. Obesity on campus. Preventing Chronic Disease, 4, A72.

Šprogar, M., Kokol, P., Zorman, M., Podgorelec, V., Yamamoto, R., Masuda, G., Sakamoto, N. 2001. Supporting medical decisions with vector decision trees. In MEDINFO 2001, 552–556. IOS Press.

Strzelecki, M., Badura, P. 2022. Machine Learning for Biomedical Application. Applied Sciences, 12, 1–5.

Teng, L., Qiao, Y., Shafiq, M., Srivastava, G., Javed, A.R., Gadekallu, T.R, Yin, S. 2023. FLPK-BiSeNet: Federated learning based on priori knowledge and bilateral segmentation network for image edge extraction. IEEE Transactions on Network and Service Management, 20, 1529–1542.

Wilson, S.L., Gallivan, A., Kratzke, C., Amatya, A. 2012. Nutritional status and socio-ecological factors associated with overweight/obesity at a rural-serving US-Mexico border university. Rural and Remote Health, 12, 1–15.

APPENDIX

Pseudocode of the algorithm employed for statistical analysis and classification models.

```
# Ensure Python version compatibility
IF sys.version_info < (3, 5):
    raise Exception("Python version 3.5 or higher is required")
```

```
# Ensure Scikit-Learn version compatibility
IF sklearn.__version__ < "0.20":
    raise Exception("Scikit-Learn version 0.20 or higher is required")
# Import required libraries
IMPORT sys
IMPORT sklearn
IMPORT numpy as np
IMPORT os
IMPORT matplotlib as mpl
IMPORT matplotlib.pyplot as plt
IMPORT pandas as pd
IMPORT datetime
from scipy.stats IMPORT trim_mean
from statsmodels IMPORT robust
IMPORT seaborn as sns
# Import necessary libraries for machine learning and statistical analysis
from sklearn.impute IMPORT SimpleImputer
from sklearn.pipeline IMPORT Pipeline
from sklearn.preprocessing IMPORT StandardScaler
from sklearn.compose IMPORT ColumnTransformer
from sklearn.preprocessing IMPORT scale
from scipy.stats IMPORT shapiro
from scipy.stats IMPORT Chi$^2$
from sklearn.svm IMPORT LinearSVC
from itertools IMPORT compress
from sklearn.feature_selection IMPORT RFE, r_regression, SelectKBest, f_classif
from sklearn.svm IMPORT SVR
from statsmodels.stats.outliers_influence IMPORT variance_inflation_factor
from sklearn.ensemble IMPORT RandomForestClassifier
from sklearn.inspection IMPORT permutation_importance
from sklearn.neural_network IMPORT MLPClassifier
from sklearn.datasets IMPORT make_classification
from sklearn.model_selection IMPORT train_test_split
from sklearn.pipeline IMPORT make_pipeline
# Import libraries to obtain accuracy metrics
IMPORT shap
from sklearn.model_selection IMPORT cross_val_predict
from sklearn.metrics IMPORT roc_curve
from sklearn.metrics IMPORT f1_score
from sklearn.metrics IMPORT accuracy_score
from sklearn.metrics IMPORT confusion_matrix
# Load data from a suitable source (replace with actual function)
SET data TO load_data()
# Create a pandas DataFrame from the loaded data
SET df TO pd.DataFrame(data)
# PRE-PROCESSSING DATA
# Calculate the trimmed mean and robust scale using scipy.stats functions
SET trimmed_mean TO trim_mean(df['column_name'], proportiontocut=0.1)
SET robust_scale TO robust.scale.mad(df['column_name'])
# Compute p-values between AGED and NOT AGED groups
# Define the list of columns to analyze
SET columns_to_analyze TO ['AGE', 'WEIGHT', 'HEIGHT', 'BMI', 'WAIST', 'PULSE', 'WHtR', 'ARM', 'HIP', 'BPSY', 'BPDI', 'DX', 'MUSCLE', 'METABOLIC_AGE', 'VISCERAL_FAT', 'BODY_FAT','AGED']
# Create an empty DataFrame to store p-values
SET p_vals TO pd.DataFrame()
# Loop through the list of columns and calculate Pearson correlation p-values of data considering 'AGED' status
FOR column IN columns_to_analyze:
    SET stats, p_val TO stats.pearsonr(df[df['AGED'] EQUALS 0][column].describe(),
                        df[df['AGED'] EQUALS 1][column].describe())
    SET p_vals[column] TO p_val
# Print the DataFrame with p-values
OUTPUT(p_vals)
# PRE-PROCESSING DATA FOR STATISTIC ANALYSIS
# Identify numeric columns
SET numeric_columns TO df.select_dtypes(include=['float64', 'int']).columns
# Create a DataFrame for numeric data
SET df_num TO pd.DataFrame()
# Extract numeric columns into the new DataFrame
FOR column IN numeric_columns:
    SET df_num[column] TO df[column]
# Calculate the correlation matrix for numeric data
SET corr_matrix TO df_num.corr()
# Identify categorical columns
SET categorical_columns TO df.select_dtypes(include=['object']).columns
# Create a DataFrame for categorical data
```

```
SET df_cat TO pd.DataFrame()

# Extract categorical columns into the new DataFrame

FOR column IN categorical_columns:

    SET df_cat[column] TO df[column]

# Calculate descriptive statistics for the dataset

df_stats.describe()

# Calculate the correlation matrix for the dataset

df_stats.corr()

# Fill missing NaN values with the median

SET df_stats TO df_stats.fillna(df_stats.median())

# MACHINE LEARNING

# Create a copy of the dataset to analyse using machine learning techniques

SET df TO df_stats.copy()

# Perform stratified train-test split

from sklearn.model_selection IMPORT train_test_split, StratifiedShuffleSplit

SET split TO StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)

FOR train_index, test_index IN split.split(df, df["AGED"]):

    SET strat_train_set TO df.loc[train_index]

    SET strat_test_set TO df.loc[test_index]

# Function to calculate category proportions

DEFINE FUNCTION variable_cat_proportions(data):

    RETURN data["AGED"].value_counts() / len(data)

# Perform random train-test split FOR comparison

SET train_set, test_set TO train_test_split(df, test_size=0.2, random_state=42)

# Calculate and compare category proportions

SET compare_props TO pd.DataFrame({

    "Overall": variable_cat_proportions(df),

    "Stratified": variable_cat_proportions(strat_test_set),

    "Random": variable_cat_proportions(test_set),

}).sort_index()

SET compare_props["Rand. %error"] TO 100 * compare_props["Random"] / compare_props["Overall"] - 100

SET compare_props["Strat. %error"] TO 100 * compare_props["Stratified"] / compare_props["Overall"] - 100


# Drop labels from the training set

SET df TO strat_train_set.drop("AGED", axis=1)

SET df_labels TO strat_train_set["AGED"].copy()

# Identify rows with missing data

SET sample_incomplete_rows TO df[df.isnull().any(axis=1)].head()

# Create a pipeline for numerical data preprocessing

SET num_pipeline TO Pipeline([

        ('imputer', SimpleImputer(strategy="median")),

    ])

# Create a full preprocessing pipeline

SET num_attribs TO list(df)

SET full_pipeline TO ColumnTransformer([

        ("num", num_pipeline, num_attribs),

        ("cat", OneHotEncoder(), cat_attribs),

    ])

SET df_prepared TO full_pipeline.fit_transform(df)

# Convert prepared data to DataFrame

SET X_train TO pd.DataFrame(df_prepared, columns=df.columns)

# Scale the training data
```

```
SET X_train_scaled TO scale(X=X_train, axis=0, with_mean=True, with_std=True)

SET X_train_scaled TO pd.DataFrame(X_train_scaled, columns=X_train.columns, index=X_train.index)

# STATISTICAL SIGNIFICANCE AND P-VALUES

# Perform Shapiro-Wilk test

shapiro(X_train['WHtR'])

# Perform Chi$^2$ test for feature selection

SET X TO X_train.copy()

SET y TO y_train.copy()

SET chi_scores TO Chi$^2$ (X, y)

SET p_values TO pd.Series(chi_scores[1], index=X.columns)

p_values.sort_values(ascending=False, inplace=True)

# Perform ANOVA test

SET X TO X_train.copy()

SET y TO y_train.copy()

SET F, pvals TO f_classif(X, y)

SET p_values TO pd.Series(pvals, index=X.columns)

p_values.sort_values(ascending=False, inplace=True)

p_values.plot.bar()

# CORRELATION BASED FUTURE SELECTION

# Pearson Correlation

SET selector TO SelectKBest(r_regression, k=8).fit(X, y)

# Random Forest Estimator RFE

SET estimator TO SVR(kernel="linear")

SET selector TO RFE(estimator, n_features_to_select=8, step=1)

SET selector TO selector.fit(X, y)

# Variance Inflation Factor VIF

SET vif TO pd.DataFrame()

SET vif["VIF Factor"] TO [variance_inflation_factor(X.values, i) FOR i IN range(X.shape[1])]

SET vif["features"] TO X.columns

vif.sort_values(by=['VIF Factor'])

# CLASSIFICATION MODELS

# Training a Random Forest Classifier

SET RF_clf TO RandomForestClassifier(n_estimators=200, max_leaf_nodes=16, random_state=42)

SET X TO X_train.copy()

SET y TO y_train.copy()

SET X_test TO strat_test_set.drop("AGED", axis=1)

RF_clf.fit(X, y)

SET y_pred TO RF_clf.predict(X_test)

SET accuracy TO accuracy_score(y_test, y_pred)

SET conf_matrix TO confusion_matrix(y_test, y_pred)

OUTPUT("Random Forest Classifier Accuracy:", accuracy)

OUTPUT("Confusion Matrix:\n", conf_matrix)

# Permutation Importance with Random Forest

SET rnd_clf TO RandomForestClassifier(n_estimators=200, max_leaf_nodes=16, random_state=42)

SET X TO X_train.copy()

SET y TO y_train.copy()

rnd_clf.fit(X, y)

SET result TO permutation_importance(rnd_clf, X, y, n_repeats=10, random_state=42, n_jobs=2)

SET feat_labels TO X.select_dtypes(include=['float64', 'int64']).columns

SET forest_importances TO pd.Series(result.importances_mean, index=feat_labels)
SET sorted_indices TO np.argsort(forest_importances)[::-1]

# Sorting indices

FOR f IN range(X.shape[1]):

    OUTPUT("%2d) %-*s %f" % (f + 1, 30, feat_labels[sorted_indices[f]], forest_importances[sorted_indices[f]]))

# Fitting the selected classification models: Decision Tree, Random Forest, Neural Networks, AdaBoosting using FOR loop

# List of classifier names and corresponding classifier instances

SET names TO [

    "Decision Tree",

    "Random Forest",

    "Neural Net",

    "AdaBoost",
]

SET classifiers TO [

    DecisionTreeClassifier(max_depth=5),

    RandomForestClassifier(max_depth=5, n_estimators=10, max_features=1),

    MLPClassifier(alpha=1, max_iter=1000),

    AdaBoostClassifier(),
]

# Function to plot ROC curve

DEFINE FUNCTION plot_roc_curve(fpr, tpr, label=None):

    plt.plot(fpr, tpr, linewidth=2, label=label)

    plt.plot([0, 1], [0, 1], 'k--')

    plt.axis([-0.05, 1.05, -0.05, 1.05])

    plt.xlabel('False positive rate')

    plt.ylabel('True positive rate')
```

```
# Loop over classifiers

FOR name, clf IN zip(names, classifiers):

    SET X TO X_train.copy()

    SET y TO y_train.copy()

    SET X_test TO strat_test_set.drop("AGED", axis=1)

    # Create a pipeline with standard scaling and the classifier

    SET clf TO make_pipeline(StandardScaler(), clf)

    clf.fit(X, y)

    SET y_pred TO clf.predict(X_test)

    SET accuracy TO accuracy_score(y_test, y_pred)

    SET conf_matrix TO confusion_matrix(y_test, y_pred)

    SET f1_macro TO f1_score(y_test, y_pred, average='macro')

    OUTPUT(name)

    OUTPUT("Accuracy:", accuracy)

    OUTPUT("Confusion Matrix:\n", conf_matrix)

    OUTPUT("F1 Score (Macro):", f1_macro)

    # SHAP explanation and visualization

    SET explainer TO shap.Explainer(clf.predict, X_test)

    SET shap_values TO explainer(X_test)

    # Compute ROC curve FOR the classifier

    SET y_probas TO cross_val_predict(clf, X, y, cv=3, method='predict_proba')

    SET y_scores TO y_probas[:, 1]

    SET fpr, tpr, thresholds TO roc_curve(y, y_scores)

    # Plot ROC curve and add legend

    plot_roc_curve(fpr, tpr, name)

    plt.legend(loc="lower right")

plt.show()
```

```
# Print SHAP values

OUTPUT(shap_values)
```