symmetry

MDPI

Article

Prediction of Metabolic Syndrome in a Mexican Population Applying Machine Learning Algorithms

Guadalupe Obdulia Gutiérrez-Esparza [1,2], Oscar Infante Vázquez [2], Maite Vallejo [2,*] and José Hernández-Torruco [3,*]

[1] Cátedras CONACYT Consejo Nacional de Ciencia y Tecnología, Ciudad de México 08400, Mexico; ggutierreze@conacyt.mx

[2] Instituto Nacional de Cardiología Ignacio Chávez, Ciudad de México 14080, Mexico; osinfa@yahoo.com

[3] División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco 86690, Mexico

* Correspondence: maite_vallejo@yahoo.com.mx (M.V.); jose.hernandezt@ujat.mx (J.H.-T.); Tel.: +52-5555732911 (M.V.)

Received: 15 March 2020; Accepted: 26 March 2020; Published: 7 April 2020

Abstract: Metabolic syndrome is a health condition that increases the risk of heart diseases, diabetes, and stroke. The prognostic variables that identify this syndrome have already been defined by the World Health Organization (WHO), the National Cholesterol Education Program Third Adult Treatment Panel (ATP III) as well as by the International Diabetes Federation. According to these guides, there is some symmetry among anthropometric prognostic variables to classify abdominal obesity in people with metabolic syndrome. However, some appear to be more sensitive than others, nevertheless, these proposed definitions have failed to appropriately classify a specific population or ethnic group. In this work, we used the ATP III criteria as the framework with the purpose to rank the health parameters (clinical and anthropometric measurements, lifestyle data, and blood tests) from a data set of 2942 participants of Mexico City Tlalpan 2020 cohort, applying machine learning algorithms. We aimed to find the most appropriate prognostic variables to classify Mexicans with metabolic syndrome. The criteria of sensitivity, specificity, and balanced accuracy were used for validation. The ATP III using Waist-to-Height-Ratio (WHtR) as an anthropometric index for the diagnosis of abdominal obesity achieved better performance in classification than waist or body mass index. Further work is needed to assess its precision as a classification tool for Metabolic Syndrome in a Mexican population.

Keywords: metabolic syndrome; Random Forest; Youden Index; Mexico City; cohort study; waist to height ratio

1. Introduction

Metabolic Syndrome (MetS) encompasses a group of cardiovascular risk factors that increase the likelihood of suffering heart and other metabolic illnesses, such as cerebrovascular stroke and diabetes.

MetS was first described by Kylin in 1920 as the coexistence of hypertension, hyperglycemia, and gout [1]. In 1940, the central obesity component was added [2]; since then, several definitions have been used to describe it, even different names have been given, such as the X syndrome, the insulin resistance syndrome [3] or the deadly quartet [4].

Due to the controversy regarding a worldwide definition, in 1998, an international initiative gather-up in an attempt to achieve an agreement on this matter. The World Health Organization (WHO 1999) proposed a set of criteria [5], then the National Cholesterol Education Program's Adult Treatment Panel III (NCEP: ATP III-2004) [6] and the European Group on the Study of Insulin Resistance (IDF) [7]

(Table 1). Even though these definitions agree on the essential components (glucose intolerance, obesity, hypertension, and dyslipidemia), there is a disagreement in the cutoff points of some components as well as in the cluster of components that should be included, an example of this is the anthropometric indexes that have been used to define obesity and central obesity. Also, because the distribution of adipose tissue may vary concerning age, gender and ethnicity, these proposed definitions have failed to appropriately classify a specific population or ethnic group, such as Africans, Latin-Americans or Japanese, among others.

On the other hand, machine learning, a sub-discipline of artificial intelligence has had a high tendency in health research, providing methods and techniques that have been successfully applied in a variety of medical domains and the early diagnosis of several diseases such as hypertension [8,9] diabetes [10,11], and MetS [12,13].

One of the first medical research applying machine learning algorithms [14], used the blood pressure data of 300 clinically healthy participants and 85 subjects with hypertension. An expert system applying neural networks was developed to diagnose and treat high blood pressure; the final system acieved 94% accuracy, this means that 94 out of every 100 participants were correctly diagnosed, either positives or negatives.

The use of machine learning to predict the MetS, applying algorithms such as decision tree [15,16], logistic regression [17], Naïve Bayes [18] and support vector machine (SVM) [19] among others, have achieved high performance. Karimi-Alavijeh et al. [20] used a decision tree and their results showed that SVM had a better performance, with an accuracy of 75%. Barakat et al. [21] also used SVM and other algorithms, such as the rule-based RIPPER (JRip), Classification and Regression Trees (CART), and C5.0 tree (C5) for diagnosis of MetS; however, the SVM achieved the best performance with an accuracy of 97%.

Artificial Neural Networks (ANN) have as well obtained a high performance in the prediction of MetS. Hirose et al. [13] predicted successfully the 6-year incidence of MetS using an ANN, with a sensitivity of 0.93 and a specificity of 0.91. Lin et al. [17] used ANN and logistic regression models to identify MetS in patients with second-generation antipsychotics treatment; as a result, the ANN (88.3%) achieved the best performance in accuracy, though logistic regression (83.6%) does not differ much from the ANN. Sedehi et al. [22] concluded that machine learning algorithms compared to logistic regression and discriminant analysis has better performance to predict MetS with higher accuracy.

Random Forest, another machine learning algorithm, has been applied in the prediction of MetS [23,24]. This algorithm performs a classification and regression process and the ranking of prognostic variables to support the early diagnosis or prediction of a specific disease. Apilak Worachartcheewan et al. [25] determined the prevalence of MetS and achieved an accuracy above 98%. According to related research, the value obtained with Random Forest to predict MetS was higher than SVM and ANN.

In this study, machine learning algorithms were used to rank the health parameters to determine the most appropriate variables for the classification of MetS in the Mexican population, using the Mexico City Tlalpan 2020 cohort [26] data set. We look for some new information, although there are symmetric relationships between findings in the established literature. Correlation-based Feature Selection (CFS) was applied as well as, chi.squared filter methods to select relevant features in MetS diagnosis. Several experiments were performed to create the predictive models, applying JRip, C4.5 and Linear SVM classifiers. The performance among the different models was compared.

Filter methods help to identify those variables that were the most important for classification and discard those that were not important. Their advantages are simplicity, speed, and low computational cost. Filter methods used in this study were obtained from the FSelector R package [27].

JRip, C4.5, and Linear SVM are known for getting good results in classification tasks [28–30]. JRip and C4.5 also provide predictive models, understandable by humans. JRip and C4.5 were taken from the RWeka R Package [31,32].

This paper is structured as follows: in Section 2, the materials and methods are introduced. In Section 3, the experiment's performance is presented. A discussion (Section 4) and conclusion (Section 5) complete this paper, as well as some ideas for future works.

Table 1. Criteria for MetS defined by the IDF, ATP III (2004) and WHO (1999) used in this study.

Risk Factors	IDF	ATP III	WHO
Metabolic Syndrome	Waist circumference plus two or more of the following factors:	Three or more of the following factors:	Glucose intolerance and/or insulin resistance plus, at least two of the following features:
Central obesity	Ethnic specific values or BMI $\geq$ 30 kg/m ^{2}	Waist circumference: Male: >102 cm, Female: >88 cm	Waist-to-hip ratio: male (0.9), female (0.85), or BMI > 30 kg/m ^{2}
Raised blood pressure	≥130/85 mmHg or treatment for hypertension	≥130/85 mmHg	≥140/90 mmHg or treatment for hypertension
Raised triglycerides	≥150 mg/dL or treatment for dyslipidemia	≥150 mg/dL	≥150 mg/dL
Raised fasting plasma glucose	>100 mg/dL or previously diagnosed diabetes type 2	≥100 mg/dL or previously diagnosed diabetes type 2	≥110 mg/dL or previously diagnosed diabetes type 2
High density lipoprotein cholesterol (HDL-C)	Male: <40 mg/dL Female: <50 mg/dL, or treatment for dyslipidemia	Male: <40 mg/dL Female: <50 mg/dL	
Microalbuminuria albumin/creatinine ratio			≥30 mg/g

2. Materials and Methods

2.1. Data

The data set incorporated in this research was obtained from the Tlalpan 2020 cohort, and this study is conducted by the Instituto Nacional de Cardiología Ignacio Chávez in Mexico City [26]. The data were collected from 2942 subjects, 1869 women (64%) and 1073 men (36%), aged 20–50 years. The data set included different cardiovascular risk factors (clinical and anthropometric measurements, lifestyle habits and biomedical evaluation).

Clinical and anthropometric measurements. Systolic and diastolic blood pressure, measurements were made according to standard procedure JNC 7 [33], as well as Waist Circumference (WAIST), height and weight (The International Society for the Advancement of Kinanthropometry (ISAK)) [34], Body Mass Index (BMI) was estimated as (weight/height^2), and Waist-to-Height-Ratio (WHtR) was calculated dividing the waist by the height (waist/height) (cm/cm).

Lifestyles habits. Variable related to lifestyle habits such as alcohol consumption, smoking, and physical activity (measured with the long version of the International Physical Activity Questionnaire (IPAQ)) were obtained with validated questionnaires [35].

Biochemical evaluation. Blood samples were taken after 12 h of overnight fasting and the following laboratory tests were obtained: fasting plasma glucose (FPG), triglycerides (TGs), HDL cholesterol (HDL-C), LDL cholesterol (LDL-C), total cholesterol (T-Cho), uric acid (UrAc), creatinine (Cre) and sodium (Na).

2.2. Methods

Random Forest. It was introduced by Breiman and Adele Cutler [36], as a predictive algorithm that creates a set of CART classification trees and the class assigned to the instance is made by the majority vote, this is known as a classifier assembly. This algorithm can be applied to a wide range of prediction problems and can achieve a better prediction accuracy as compared to individual classification trees [37]. Random Forest has two important parameters: *mtry* and *ntree*. The *mtry* is the size of the random subsets of variables considered for splitting, being the default value $\sqrt[2]{p}$ for classification and $\frac{2}{3}$ for regression, where p is the number of variables in the data set [38]. The *ntree* parameter refers to tree size. In this work, we used randomForest library [38] available too in R [39].

Random Forest provides a method called Variable Importance Measures (VIMs) to rank the importance of variables in cases of regression or classification.

Variable Importance Measures. [40] Variable importance measures for Random Forest can be used to rank variables by their relevance in regression or classification cases. This method has been successfully applied in many applications [23,37,41]. There are two ways to identify relevant features or perform variable: (1) mean decrease of impurity (MDI), which is based on the Gini index [42], and (2) mean decrease of accuracy (MDA) based on permutation importance. MDI is typically used in classification [42] and is more robust than MDA [43,44]. MDA is more suitable for regression problems.

In this study, we use MDI, which typically uses Gini index (measure commonly chosen for classification-type cases [42]), this process is given by

$$Jimpurity(X_k) = \frac{\sum_{i \in N^k}(impbni - impani)}{\mid N^{(k)} \mid} \tag{1}$$

where *impbni* is impurity before node *i*, *impani* is impurity after node *i* and N^k represents the set of nodes in which a split based on X_k is made. This method is implemented by R with the function importance (type 2) as a part of the randomForest package [45].

JRip. It is a version of the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm [46]. JRip generates a rule set to identify the categories of the instances, and at the same time to optimize the classification error. The syntax of a rule is as follows:

if attribute1 <relational operator> value1 <logical operator>
attribute2 <relational operator> value2 < … > **then** decision-value

C4.5. [47] It creates a classification tree using training data through repeated splits. In each repetition, the most relevant predictor variable is identified using the gain ratio as a measure and using this variable the tree is bifurcated. This process is repeated until all training instances are classified. In the end, only the most important predictors are used to create the classification tree. This results in a more simplified tree.

Correlation-Based Feature Selection (CFS). [48] It assesses the capacity to predict the class and the correlation between the features in feature subsets, aiming at maximizing the former and minimizing the latter. As a result, the best predictive feature subset and with the least correlation between members is found. Having a feature subset S with k features, CFS computes the goodness of S, denoted M_s with the equation:

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2}$$

where $\overline{r_{ff}}$ is the average correlation of all feature-feature pairs. $k\overline{r_{cf}}$ is the average correlation of all feature-class pairs.

Chi-Squared. This filter calculates the chi-squared statistic of each variable taken individually concerning the class [49]. It gives a feature ranking as a result. Taken a feature f and the class c, the chi-squared test is computed with the equation:

$$X^2(f,c) = \frac{N[P(f,c)P(\overline{f},\overline{c}) - P(f,\overline{c})P(\overline{f},c)]^2}{P(f)P(\overline{f})P(c)P(\overline{c})} \tag{3}$$

where N is the number of records in the data-set. $P(x,y)$ is the joint probability of x and y. $P(x)$ is the marginal probability of x. For example: $P(\overline{f},\overline{c})$ is the joint probability of $\overline{f}$ and $\overline{c}$, $P(f)$ is the marginal probability of f. $\overline{f}$ is the complement of f. $\overline{c}$ is the complement of c.

2.3. Metrics

To evaluate the classifier performance, sensitivity (SENS), specificity (SPC), and balanced accuracy (BACC) were used and computed based on the confusion matrix, as well as on the Kappa index [50].

$$SENS = \frac{TP}{TP + FN} \tag{4}$$

$$SPC = \frac{TN}{FP + TN} \tag{5}$$

$$BACC = \left(\frac{1}{2}\right)\left(\frac{TP}{P} + \frac{TN}{N}\right) \tag{6}$$

where P = *Positive*, N = *Negative*, TP = *True Positive*, FN = *False Negative*, TN = *True Negative* and FP = *False Positive*, respectively.

2.4. Statistical Analysis

The statistical analysis was performed with the Stata package, version 13.0. The distribution of numerical variables was tested with Shapiro France Test ($P > 0.05$). Mann-Whitney U test and Chi-squared or exact Fisher tests were used to compare the studied groups (with and without MetS). An alpha index of ≤ 0.05 was considered statistically significant.

3. Results

In this study, we used a data set from a cohort study called Tlalpan 2020 (the study protocol for this cohort was published elsewhere [26]). The ATP III criteria were applied (see Table 1) to identify influential cardiovascular risk factors and classify participants with or without MetS. Figure 1 shows a general diagram of our proposed model, where the first step was to identify the variable importance of all data set applying Random Forest, chi-squared and CFS. The results obtained indicate the most important variables (features), which were used to train different models using Random Forest, C45, and JRip. Models were created using 30 independent iterations, as it is the typical number used in the literature for fair comparisons among experiments [51,52]. Then, their performance was compared considering balanced accuracy, sensitivity, and specificity.

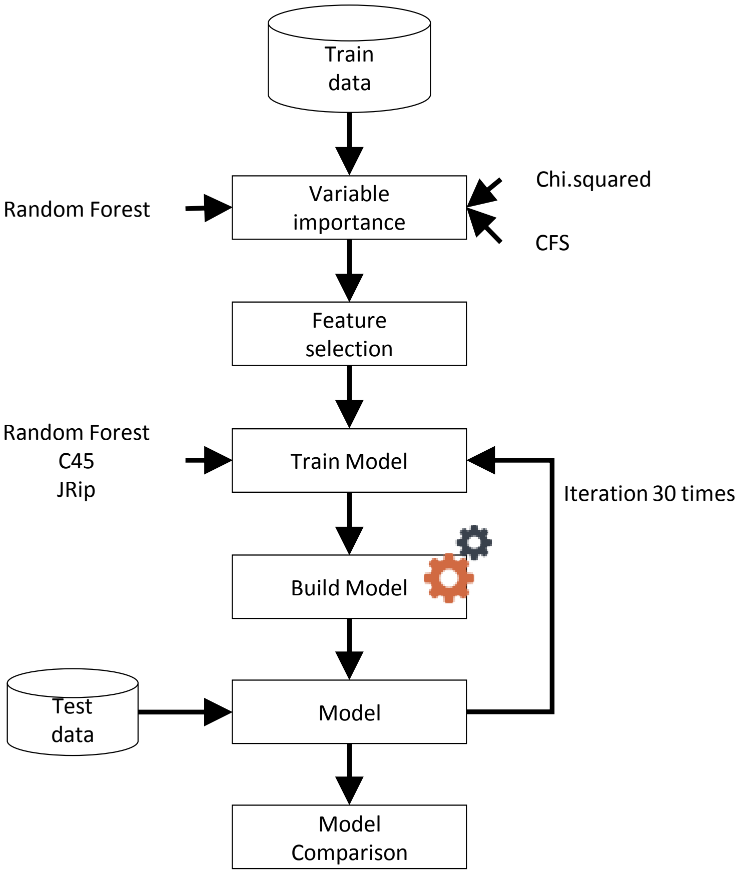

Figure 1. Diagram to define best cutoff points to predict MetS in Tlalpan 2020.

The prevalence of MetS according to ATP III criteria was 20.5% (603 participants), and no significant differences were identified between sexes (with MetS: women 20.6% vs. men 20.4%, without MetS: women 79.4% vs. men 79.6%).

The median and interquartile range (IQR) of anthropometric, clinical, and biochemical parameters are shown in Table 2. Participants with MetS were significantly older than those without MetS and showed higher values of all anthropometric and clinical parameters. Concerning biochemical parameters, MetS participants had substantially higher values than those without MetS.

Table 2. Description of anthropometric, clinical and biochemical parameters between MetS and non-MetS groups.

	Metabolic Syndrome		***p* Value**
	Yes	**No**	
	(N = 603 (20.5%))	**(N = 2339 (79.5%))**	
Anthropometric and Clinical parameters *			
Age (years)	42 (36–47)	38 (29–45)	0.0000
Weight (Kg)	81 (71.0–92.8)	66 (58.5–76)	0.0000
Height (m)	1.61 (1.55–1.68)	1.61 (1.56–1.68)	0.6407
BMI (Kg/m^2)	31.1 (28.27–33.4)	25.5 (23.1–28)	0.0000
WC (cm)	100 (94–108)	86 (80–94)	0.0000
WHtR (cm/cm)	0.62 (0.59–0.66)	0.53 (0.49–0.58)	0.0000
SBP (mmHg)	111.3 (104.7–120)	105.3 (98–112.7)	0.0000
DBP (mmHg)	77.3 (70.7–82.7)	70.7 (64.7–76.7)	0.0000
Biochemical parameters *			
FPG (mg/dL)	102 (94–108)	91 (86–96)	0.0000
CHOL (mg/dL)	188.1 (169.4–211.8)	178.3 (158.7–201.7)	0.0000
LDL-C (mg/dL)	123.4 (105.3–144.1)	115 (97.3–135.9)	0.0000
HDL-C (mg/dL)	38.2 (33.9–44)	49.6 (42.4–58)	0.0000
TGs (mg/dL)	194.1 (157.6–255.7)	108 (79.7–145.5)	0.0000
Uric Acid (mg/dL)	5.8 (4.8–6.8)	5.1 (4.3–6)	0.0000
Creatinine (mg/dL)	0.76 (0.66–0.89)	0.77 (0.67–0.90)	0.0410
Sodium (mmol/L)	140.9 (140–141.6)	140.9 (140–141.6)	0.0780
Lifestyle **			
Smoking habit			
- Never	214 (35.5)	899 (38.4)	
- Former	227 (37.6)	921 (39.4)	0.0500
- Present	162 (26.9)	519 (22.2)	
Alcohol consumption			
- Yes	376 (62.4)	1644 (70.3)	
- No	227 (37.6)	695 (29.7)	0.0000
Physical Activity			
- Low	80 (13.2)	257 (11.0)	
- Medium	276 (45.8)	1054 (45.0)	0.2000
- High	247 (41.0)	1028 (44.0)	

* Numerical data were expressed as the median (interquartile range (IQR)) and ** categorical as the number of cases and its corresponding percentage (n (%)). BMI: body mass index, WC: waist circumference, WHtR: Waist-to-Height-Ratio, SBP: systolic blood pressure, DBP: diastolic blood pressure, HR: heart rate, FPG: fasting plasma glucose, CHOL: total cholesterol, LDL-C: low-density lipoprotein cholesterol, HDL-C: high-density lipoprotein cholesterol, TGs: triglycerides.

Variable Importance and Prediction Model

As a first step, we identify the most important variables of the data set using Random Forest algorithm to construct the corresponding model, where the number of trees (ntree) varied between 100 to 1000 (ntree = 100, 200, 300, 500, 800, and 1000) and the mtry value varied between 1 to 10, applying the grid search method proposed by Hsu et al. [53]. Also, 10-fold cross-validation with ten repeats to train the model was used to ensure all data. Once the training process was finished and the best parameters were found and applied, the variable importance was obtained. Figure 2 shows the features attained by the model, where the best value in mtry was 10 and in ntree was 1000, to achieve a balanced accuracy of 0.9675 and a standard deviation (SD) of 0.0006.

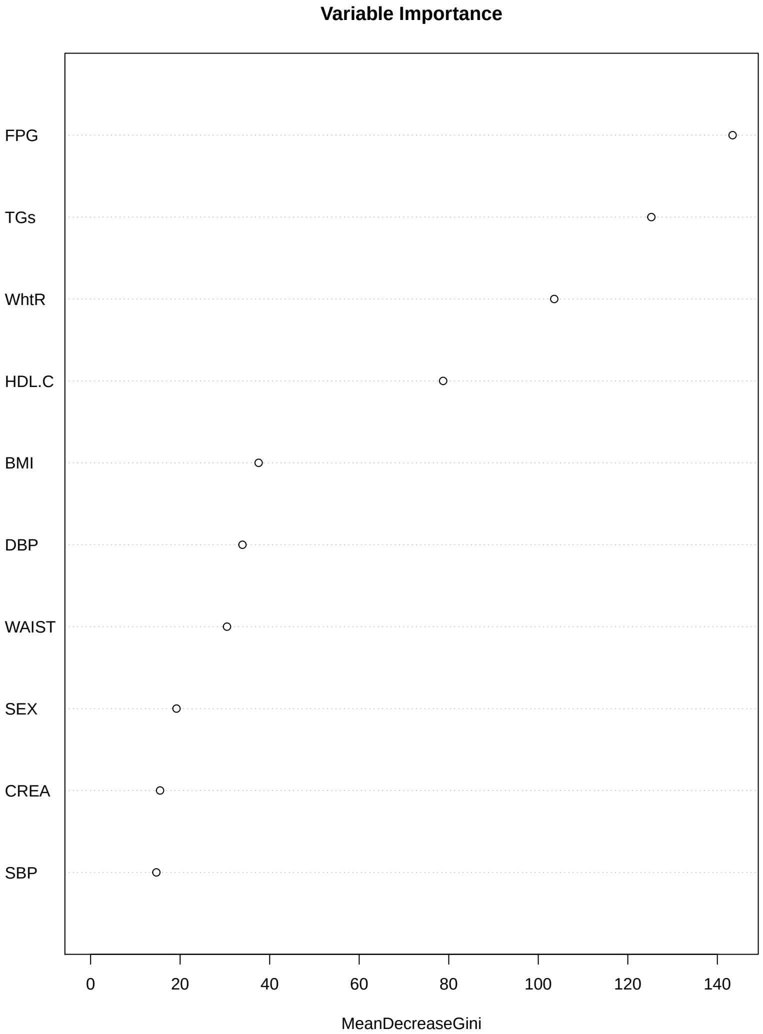

Figure 2. Variable Importance in the complete data set. FPG: glucose, TGs: triglycerides, WHtR: Waist-to-Height-Ratio, HDL.C: high-density lipoprotein cholesterol, BMI: body mass index, DBP: diastolic blood pressure, WAIST: waist circumference, CREA: Creatinine SBP: systolic blood pressure, CHOL: Total cholesterol, LDL.C: low-density lipoprotein.

In Figure 2, the variable importance is shown. FPG displayed the highest value of importance, followed by TGs, WHtR, and HDL.C; then there was a second group (BMI, DBP, and waist) and SEX, CREA and SBP showed the lowest values. Three of the four variables within the first group are considered to be indicators used by ATP III to identify MetS. However, the anthropometric index that ATP III uses to identify abdominal obesity is the waist. In our results, this index showed a lower value of importance than WHtR and even than BMI. Therefore, considering the role that abdominal obesity has as a cardiometabolic risk factor, the importance of each of the three anthropometric indexes (WHtR, BMI and waist) was tested. We performed experiments where WEIGHT, HEIGHT, and BMI, waist or WHtR were eliminated depending on the case, and a separate algorithm was built applying Random Forest and chi.squared.

Table 3 shows the variable importance of BMI, WHtR and WAIST using Random Forest. WHtR was placed in the second position with a value higher (146.4299) than BMI (118.7353) and WAIST (118.6575), which were placed in the third position.

Table 3. Variable Importance by Random Forest.

BMI_RF		WhtR_RF		WAIST_RF	
Variable	**attr_importance**	**Variable**	**attr_importance**	**Variable**	**attr_importance**
FPG	161.24255	**FPG**	154.933891	**FPG**	164.075832
TGs	133.419239	**WhtR**	146.429954	**TGs**	128.552016
BMI	118.735349	**TGs**	130.930428	**WAIST**	118.657588
HDL.C	85.128492	**HDL.C**	86.519673	**HDL.C**	96.989354
DBP	41.962949	**DBP**	40.292102	**SEX**	45.629925
CREA	25.827319	**SEX**	19.934242	**DBP**	41.41832
SEX	19.913751	**SBP**	17.330333	**CREA**	18.910847
SBP	19.120117	**CREA**	17.207579	**SBP**	14.264248
CHOL	13.730005	**CHOL**	13.408404	**URIC**	10.246256
URIC	13.240346	**URIC**	11.439449	**CHOL**	9.573875
LDL.C	12.541307	**LDL.C**	9.80734	**LDL.C**	7.160062
AGE	9.899638	**SODIUM**	8.417866	**SODIUM**	6.650695
SODIUM	9.205347	**AGE**	7.628182	**AGE**	5.868129
TOBACCO	3.266133	**TOBACCO**	2.863531	**TOBACCO**	1.640314
PHYSICAL_ACT	3.056807	**PHYSICAL_ACT**	2.193845	**PHYSICAL_ACT**	1.387089
ALCOHL	1.570261	**ALCOHL**	1.266757	**ALCOHL**	1.125397

Table 4, shows the variable importance of BMI, WHtR and WAIST, using chi.squared. Even though the three anthropometric indexes were placed in the third position, WHtR achieved a higher value (0.5118) than WAIST (0.5068) and BMI (0.4975). As for the last six variables for which the importance was 0, it means that they are not important for diagnosing MetS according to chi.squared filter, therefore they can be discarded from the models.

Table 4. Variable Importance by Chi.squared.

BMI_chi.squared		WhtR_chi.squared		Waist_chi.squared	
Variable	**attr_importance**	**Variable**	**attr_importance**	**Variable**	**attr_importance**
FPG	0.536249993	**FPG**	0.536249993	**FPG**	0.536249993
TGs	0.518863961	**TGs**	0.518863961	**TGs**	0.518863961
BMI	0.49751679	**WhtR**	0.511873169	**WAIST**	0.506820302
HDL.C	0.439123538	**HDL.C**	0.439123538	**HDL.C**	0.439123538
DBP	0.320697596	**DBP**	0.320697596	**DBP**	0.320697596
SBP	0.269691951	**SBP**	0.269691951	**SBP**	0.269691951
URIC	0.198714284	**URIC**	0.198714284	**URIC**	0.198714284
AGE	0.159260928	**AGE**	0.159260928	**AGE**	0.159260928
CHOL	0.123288978	**CHOL**	0.123288978	**CHOL**	0.123288978
LDL.C	0.117543224	**LDL.C**	0.117543224	**LDL.C**	0.117543224
SEX	0	**SEX**	0	**SEX**	0
TOBACCO	0	**TOBACCO**	0	**TOBACCO**	0
ALCOHL	0	**ALCOHL**	0	**ALCOHL**	0
PHYSICAL_ACT	0	**PHYSICAL_ACT**	0	**PHYSICAL_ACT**	0
CREA	0	**CREA**	0	**CREA**	0
SODIUM	0	**SODIUM**	0	**SODIUM**	0

Once the importance of the variables was obtained with Random Forest (see Table 3) and chi.squared (Table 4), the models for each anthropometric index (WHtR, WAIST, and BMI) using Random Forest, C45 and JRIP as classifiers were developed.

In Table 5, the performance of the 30 models developed for each anthropometric index (WHtR, WAIST, and BMI) using Random Forest, C45 and JRIP as classifiers are shown. The classifier that performed best was the Random Forest for the three anthropometric indexes; however, WAIST showed the highest importance. On the other hand, C45 and JRIP obtained lower importance for the three anthropometric indexes, and the highest importance was observed for the WHtR.

Table 5. Performance of the models constructed for each case (WhtR, WAIST and BMI).

Variable	Classifiers	Attribute Selection Method	Balanced Accuracy	SD Balanced Accuracy	Sensitivity	SD Sensitivity	Specificity	SD Specificity
WhtR	RF (mtry = 9 ntree = 800)	RF	0.9403	0.0018	0.9669	0.0010	0.9137	0.0037
	C45	CFS	0.8655	0.0182	0.9524	0.0110	0.7786	0.0420
	JRIP	CFS	0.8606	0.0182	0.9546	0.0139	0.7667	0.0469
WAIST	RF (mtry = 10 ntree = 300)	RF	0.9772	0.0025	0.9813	0.0015	0.9731	0.0050
	JRIP	CFS	0.8388	0.0323	0.9331	0.7444	0.0863	
	C45	CFS	0.8043	0.0309	0.9593	0.0205	0.6493	0.0797
BMI	RF (mtry = 9 ntree = 300)	RF	0.9101	0.0032	0.9645	0.0019	0.8557	0.0061
	JRIP	CFS	0.8475	0.0154	0.9443	0.0141	0.7507	0.0391
	C45	CFS	0.8354	0.0233	0.9519	0.0154	0.7189	0.0573

Since ATP III is one of the most used guidelines in Latin America to define the MetS, we constructed three models, one to measure the performance of variables used by ATP III (see Table 6), another to measure the same ATP III variables, replacing WAIST for WHtR (see Table 7) and the last one to measure the same ATP III variables using the BMI instead of WAIST.

Tables 6–8 show sensitivity, specificity, and the balanced accuracy, as well as their respective standard deviations of the average performance for the 30 models generated for each case. In the case of the model using ATP III variables (Table 6), the best performance was attained by Random Forest with a Balanced Accuracy of 0.8754 and an SD of 0.0036, followed by JRip (0.8723, 0.0203). The worst average performance was attained by SVM linear with a cost = 100, where the Balanced Accuracy was 0.7561 and the SD was 0.0136. The model in which WAIST was replaced with WHtR achieved the best performance with JRip with a balanced accuracy of 0.8926 and an SD of 0.0142, followed by Random Forest (0.8905, 0.0022), the worst average performance was attained by SVM linear with a cost = 50 (0.7812, 0.0154).

Table 6. Results of model performance using ATP III variables.

Classifiers	Avg Balanced Accuracy	SD Balanced Accuracy	Avg Sensitivity	SD Sensitivity	Avg Specificity	SD Specificity
Random Forest	0.8754	0.0036	0.9512	0.0007	0.7996	0.0072
Jrip	0.8723	0.0203	0.9428	0.0164	0.8018	0.0513
C4.5	0.8592	0.0207	0.9525	0.0130	0.7658	0.0488
knn (k = 44 d = 2)	0.7696	0.0136	0.9633	0.0090	0.5758	0.0305
SVM Linear c = 100	0.7561	0.0137	0.9562	0.0093	0.5561	0.0305

Table 7. Results of model performance using ATP III variables, replacing WAIST with WhtR.

Classifiers	Avg Balanced Accuracy	SD Balanced Accuracy	Avg Sensitivity	SD Sensitivity	Avg Specificity	SD Specificity
Jrip	0.8926	0.0142	0.9483	0.0132	0.8370	0.0333
Random Forest	0.8905	0.0022	0.9534	0.0012	0.8275	0.0044
C4.5	0.8775	0.0187	0.9573	0.0105	0.7977	0.0434
knn (k = 44 d = 1)	0.8004	0.0114	0.9733	0.0072	0.6275	0.0247
SVM Linear c = 50	0.7812	0.0154	0.9575	0.0099	0.6050	0.0323

Table 8. Results of model performance using ATP III variables, replacing WAIST with BMI.

Classifiers	Avg Balanced Accuracy	SD Balanced Accuracy	Avg Sensitivity	SD Sensitivity	Avg Specificity	SD Specificity
Jrip	0.8691	0.0168	0.9421	0.0144	0.7960	0.0420
Random Forest	0.8650	0.0033	0.9407	0.0020	0.7894	0.0060
C4.5	0.8534	0.0185	0.9518	0.0133	0.7551	0.0450
knn (k = 44 d = 1)	0.7809	0.0139	0.9729	0.0062	0.5889	0.0315
SVM Linear c = 50	0.7694	0.0153	0.9568	0.0083	0.5821	0.0330

Finally, the model using ATP III variables, replacing WAIST with BMI, achieved the best performance model with JRip with a balanced accuracy of 0.8691 and an SD of 0.0168, followed by Random Forest (0.8650, 0.0033), the worst average performance was attained by SVM linear with a cost = 50 (0.7694, 0.0153).

The executed experiments showed that the model using the ATP III variables with WHtR instead of WAIST achieved the best performance, whereby could be a useful index for the identification of MetS in a Mexican population, along with the variables already proposed by ATP III.

4. Discussion

In this study, a set of health parameters was ranked applying Random Forest and compared with chi.squared and CFS filter methods to obtain the variable importance. These results showed that the main prognostic variables of MetS in our cohort of the Mexico City population according to its importance were: FPG, TGs, WHtR, HDL-C, and BMI, four out of these five variables are among those proposed by the WHO, IDF and ATP III criteria for the classification of people with MetS; however, not taking into consideration its predictability importance. Other studies have also found these prognostic variables; however, using different classification methods [15,20,54].

An interesting result was that WHtR was considered the third variable in order of importance, which is an important finding especially concerning the obesity epidemic in our country [55], and its relationship with cardiovascular disease, which is the first cause of morbidity and mortality worldwide and in Mexico.

Abdominal obesity has become an indicator of cardiometabolic risk. Therefore, significant efforts have been made to find the proper anthropometric measurement that reflects the accumulation of fat tissue in the abdominal area and can be easily obtained without high technology equipment.

It is also true that anthropometric indexes are importantly influenced by age, gender, and ethnicity, among other factors, and therefore, finding the appropriate one could be an overwhelming task. BMI has been used as an indicator of body fatness; however, it does not reflect abdominal obesity. Furthermore, BMI might scale to height with other power than 2, and therefore erroneous conclusions might be made regarding the adipose composition in people with different heights [56].

In recent years, abdominal obesity indexes such as BMI, WAIST, and recently the WHtR have been proposed as indicators of a high cardiometabolic risk [57,58].

A systematic review that included seventy-eight cross-sectional and prospective studies analyzed the predicting capability of WHtR, WAIST, and BMI to identify the risk of diabetes and CVD, and found that WHtR, WAIST, and BMI are useful predictors for this matter, furthermore, balance and adjusted data suggested that WHtR and WAIST are stronger predictors than BMI [59]. Browning et al. [59] suggest that "Keep your waist circumference to less than half your height", could be a suitable cutoff point for all ethnic groups.

In a more recent systematic review and meta-analysis, Ashwell et al. [57], aimed to differentiate the screening potential of WHtR and WAIST for adult cardiometabolic risk (hypertension, diabetes, dyslipidemia, MetS, and overall cardiovascular outcomes) and found that WHtR had significantly higher discriminatory power compared with BMI. However, most importantly, statistical analysis of the with-in study showed that WHtR was a better predictor than WAIST for hypertension, diabetes, cardiovascular disease, and all outcomes in both genders.

The predictive capability of WHtR has been tested in several populations [58,60–63].

Comparing WHtR, WAIST, and BMI

To compare the importance value of WHtR, WAIST, and BMI separately in the complete data set we applied Random Forest and chi.squared. The results showed that WHtR is the most important variable since it obtained the highest values with Random Forest (see Table 3) and chi.squared (Table 4). Likewise, in the results shown in Table 5, WHtR achieved the best performance in balanced accuracy, sensitivity, and specificity with C45 and JRip, using CFS as a feature selection method, even if WAIST achieves better performance with Random Forest.

The ATP III guidelines are the most used to diagnose MetS; however, ethnic and regional characteristics need to be recognized to adjust the parameters for the diagnosis of abdominal obesity. Thus, the performance of WHtR and BMI using the variables of ATP III except for the WAIST was proved. The results in Table 6 show the performance of the model using only ATP III variables, where Random Forest achieves the highest value (0.8754). In Table 7, BMI reached the best performance with JRip (0.8691); however, it fails to reach the value obtained by WAIST. The values attained by WHtR showed the best performance using all classifiers, highlighting Random Forest with the highest values. This shows that for our study, using data from the Mexico City Tlalpan 2020 cohort participants, the WHtR in combination with the variables of ATP III (except for the waist) achieves a better performance in classification than the WAIST and BMI.

5. Conclusions

Machine learning algorithms have become a useful prognostic tool in medicine [64] to predict different medical outcomes such as treatment response to (chemo)radiotherapy [65], study metabolomic [66], and to identify the association between microbes, metabolites and abdominal pain in children with irritable bowel syndrome [67]. In our case, we used Random Forest to rank health parameters evaluating the prediction performance of the algorithm by accuracy (97%), sensitivity (97%) and specificity (93%). Even though the results of Apilak Worachartcheewan et al. [25] are similar to ours, they obtained an accuracy of 98% using Random Forest to determine MetS prevalence. However, when using other algorithms, such as SVM, results have shown an important variability, for instance, Karimi Alavijeh et al. [20] achieved an accuracy of 75%, while Barakat et al. [21] achieved an accuracy of 97%. Similar results were published using ANN; Hirose et al. [13] reported a sensitivity of 93% and a specificity of 91%. Lin et al. [17] achieved a lower accuracy (88.3%) using the same technique and 83.6% applying logistic regression. However, it is necessary to emphasize that an adequate feature selection and feature ranking significantly impacts the performance and computational burden of machine learning algorithms [11].

In this study, we only included a population living in Mexico City. Nevertheless, MetS encompasses chronic degenerative diseases with a significant genetic burden. Also, Mexico is a country with a wide variety of ethnic groups. Therefore, it will be essential to include populations from other regions of Mexico to have these ethnicities, cultures, customs, lifestyles, diet, and anthropometric characteristics represented and to develop an algorithm that can be applied throughout Mexico to detect and predict the MetS.

Finally, machine learning algorithms have potential applicability in medicine for diagnosis, being Random Forest the most useful algorithm for prediction and ranking variables; in our Tlalpan 2020 cohort, FPG, TGs, WHtR, HDL-C, BMI, DBP, and WAIST were the most important variables to diagnose (or predict) MetS, these results were similar to those found in other cohorts [15,25,54]. Likewise, results using JRip, C4.5, Knn, SVM and Random Forest, showed that WHtR could be a useful index for the identification of MetS, along with other variables proposed by ATP III.

Author Contributions: Conceptualization—G.O.G.-E., M.V.; Methodology—G.O.G.-E., J.H.-T.; Software—G.O.G.-E., J.H.-T.; Supervision—M.V.; Validation—G.O.G.-E.; formal analysis—J.H.-T.; writing—original draft preparation—G.O.G.-E., J.H.-T., M.V.; writing—review and editing—G.O.G.-E., J.H.-T., M.V., O.I.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Council of Science and Technology (CONACYT, México), Cátedras CONACYT 1591.

Acknowledgments: We want to extend our appreciation to Consejo Nacional de Ciencia y Tecnología (CONACYT) (National Council for Science and Technology) under the 'Cátedras CONACYT' programme, No. 1591.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kylin, E. Studien ueber das Hypertonie-Hyperglyka "mie-Hyperurika" miesyndrom. *Zentralblatt Für Inn. Med.* **1923**, *44*, 105–127.
2. Vague, J. La différentiation sexuelle facteur déterminant des formes de l'obésité. *Presse Med.* **1947**, *30*, 339–340.
3. Ferrannini, E.; Haffner, S.; Mitchell, B.; Stern, M. Hyperinsulinaemia: The key feature of a cardiovascular and metabolic syndrome. *Diabetologia* **1991**, *34*, 416–422. [CrossRef] [PubMed]
4. Kaplan, N.M. The deadly quartet: Upper-body obesity, glucose intolerance, hypertriglyceridemia, and hypertension. *Arch. Intern. Med.* **1989**, *149*, 1514–1520. [CrossRef]
5. World Health Organization. *Waist Circumference and Waist-Hip Ratio: Report of a WHO Expert Consultation, Geneva, 8–11 December 2008*; World Health Organization: Geneva, Switzerland, 2011.
6. Grundy, S.M.; Brewer, H.B., Jr.; Cleeman, J.I.; Smith, S.C., Jr.; Lenfant, C. Definition of metabolic syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation* **2004**, *109*, 433–438. [CrossRef]

7. Alberti, K.G.M.; Zimmet, P.; Shaw, J. The metabolic syndrome—A new worldwide definition. *Lancet* **2005**, *366*, 1059–1062. [CrossRef]
8. Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access* **2017**, *5*, 8869–8879. [CrossRef]
9. Magoulas, G.D.; Prentza, A. Machine learning in medical applications. In *Advanced Course on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 300–307.
10. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116. [CrossRef]
11. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; González-Vidal, A.; Zamora, M.Á. Feature Selection for Blood Glucose Level Prediction in Type 1 Diabetes Mellitus by Using the Sequential Input Selection Algorithm (SISAL). *Symmetry* **2019**, *11*, 1164. [CrossRef]
12. de Edelenyi, F.S.; Goumidi, L.; Bertrais, S.; Phillips, C.; MacManus, R.; Roche, H.; Planells, R.; Lairon, D. Prediction of the metabolic syndrome status based on dietary and genetic parameters, using Random Forest. *Genes Nutr.* **2008**, *3*, 173. [CrossRef]
13. Hirose, H.; Takayama, T.; Hozawa, S.; Hibi, T.; Saito, I. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. *Comput. Biol. Med.* **2011**, *41*, 1051–1056. [CrossRef] [PubMed]
14. Poli, R.; Cagnoni, S.; Livi, R.; Coppini, G.; Valli, G. A neural network expert system for diagnosing and treating hypertension. *Computer* **1991**, *24*, 64–71. [CrossRef]
15. Worachartcheewan, A.; Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Pidetcha, P.; Prachayasittikul, V. Identification of metabolic syndrome using decision tree analysis. *Diabetes Res. Clin. Pract.* **2010**, *90*, e15–e18. [CrossRef] [PubMed]
16. Babič, F.; Majnarić, L.; Lukáčová, A.; Paralič, J.; Holzinger, A. On patient's characteristics extraction for metabolic syndrome diagnosis: Predictive modelling based on machine learning. In *International Conference on Information Technology in Bio-and Medical Informatics*; Springer: Cham, Switzerland, 2014; pp. 118–132.
17. Lin, C.C.; Bai, Y.M.; Chen, J.Y.; Hwang, T.J.; Chen, T.T.; Chiu, H.W.; Li, Y.C. Easy and low-cost identification of metabolic syndrome in patients treated with second-generation antipsychotics: Artificial neural network and logistic regression models. *J. Clin. Psychiatry* **2010**, *71*, 225. [CrossRef] [PubMed]
18. Perveen, S.; Shahbaz, M.; Keshavjee, K.; Guergachi, A. Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques. *IEEE Access* **2019**, *7*, 1365–1375. [CrossRef]
19. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef] [PubMed]
20. Karimi-Alavijeh, F.; Jalili, S.; Sadeghi, M. Predicting metabolic syndrome using decision tree and support vector machine methods. *ARYA Atheroscler.* **2016**, *12*, 146.
21. Barakat, N. Diagnosis of Metabolic Syndrome: A Diversity Based Hybrid Model. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*; Springer: Cham, Switzerland, 2016; pp. 185–198.
22. Sedehi, M.; Mehrabi, Y.; Kazemnejad, A.; Hadaegh, F. Comparison of artificial neural network, logistic regression and discriminant analysis methods in prediction of metabolic syndrome. *Iran. J. Endocrinol. Metab.* **2010**, *11*, 638–646.
23. Janitza, S.; Tutz, G.; Boulesteix, A.L. Random forest for ordinal responses: Prediction and variable selection. *Comput. Stat. Data Anal.* **2016**, *96*, 57–73. [CrossRef]
24. Boulesteix, A.L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 493–507. [CrossRef]
25. Worachartcheewan, A.; Shoombuatong, W.; Pidetcha, P.; Nopnithipat, W.; Prachayasittikul, V.; Nantasenamat, C. Predicting metabolic syndrome using the random forest method. *Sci. World J.* **2015**, *2015*, 581501. [CrossRef] [PubMed]
26. Colín-Ramírez, E.; Rivera-Mancía, S.; Infante-Vázquez, O.; Cartas-Rosado, R.; Vargas-Barrón, J.; Madero, M.; Vallejo, M. Protocol for a prospective longitudinal study of risk factors for hypertension incidence in a Mexico City population: The Tlalpan 2020 cohort. *BMJ Open* **2017**, *7*, e016773. [CrossRef] [PubMed]

27. Romanski, P.; Kotthoff, L. *FSelector: Selecting Attributes*; R Package Version 0.31. 2018. Available online: http://freebsd.yz.yamagata-u.ac.jp/pub/cran/web/packages/FSelector/FSelector.pdf (accessed on 15 March 2020).
28. Gutiérrez Esparza, G.; Vallejo, M.; Hernandez, J. Classification of Cyber-Aggression Cases Applying Machine Learning. *Appl. Sci.* **2019**, *9*, 1828. [CrossRef]
29. Asri, H.; Mousannif, H.; Al Moatassime, H.; Noël, T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Comput. Sci.* **2016**, *83*, 1064–1069. [CrossRef]
30. Kourou, K.; Exarchos, T.; Exarchos, K.; Karamouzis, M.; Fotiadis, D. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2014**, *13*. [CrossRef] [PubMed]
31. Hornik, K.; Buchta, C.; Zeileis, A. Open-Source Machine Learning: R Meets Weka. *Comput. Stat.* **2009**, *24*, 225–232. [CrossRef]
32. Witten, I.H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, USA, 2005.
33. Chobanian, A.V.; Bakris, G.L.; Black, H.R.; Cushman, W.C.; Green, L.A.; Izzo, J.L., Jr.; Jones, D.W.; Materson, B.J.; Oparil, S.; Wright, J.T., Jr.; et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension* **2003**, *42*, 1206–1252. [CrossRef]
34. Marfell-Jones, M.J.; Stewart, A.; De Ridder, J. *International Standards for Anthropometric Assessment*; International Society for the Advancement of Kinanthropometry: Wellington, New Zealand, 2012.
35. Craig, C.L.; Marshall, A.L.; Sjöström, M.; Bauman, A.E.; Booth, M.L.; Ainsworth, B.E.; Pratt, M.; Ekelund, U.; Yngve, A.; Sallis, J.F.; et al. International physical activity questionnaire: 12-country reliability and validity. *Med. Sci. Sport. Exerc.* **2003**, *35*, 1381–1395. [CrossRef]
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef]
38. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2001; Volume 1.
39. Team, R.C. R: A Language and Environment for Statistical Computing. 2013. Available online: https://repo.bppt.go.id/cran/web/packages/dplR/vignettes/intro-dplR.pdf (accessed on 4 April 2020).
40. Hjerpe, A. Computing Random Forests Variable Importance Measures (vim) on Mixed Numerical and Categorical Data. 2016. Available online: http://www.diva-portal.org/smash/record.jsf?pid=diva2 (accessed on 4 April 2020).
41. Shi, T.; Horvath, S. Unsupervised learning with random forest predictors. *J. Comput. Graph. Stat.* **2006**, *15*, 118–138. [CrossRef]
42. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Classification and regression trees. *Wadsworth Int. Group* **1984**, *37*, 237–251.
43. Nembrini, S.; König, I.R.; Wright, M.N. The revival of the Gini importance? *Bioinformatics* **2018**, *34*, 3711–3718. [CrossRef]
44. Calle, M.L.; Urrea, V. Letter to the editor: Stability of random forest importance measures. *Briefings Bioinform.* **2010**, *12*, 86–89. [CrossRef]
45. Liaw, A.; Wiener, M. RandomForest: Breiman and Cutler's random forests for classification and regression. *R Package Version* **2015**, *4*, 6–10.
46. Cohen, W.W. Fast Effective Rule Induction. In Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; pp. 115–123.
47. Quinlan, J. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Francisco, CA, USA, 1993.
48. Hall, M.A. *Correlation-based Feature Selection for Machine Learning*; Technical Report; The University of Waikato: Hamilton, New Zealand, 1999.
49. Zheng, Z.; Wu, X.; Srihari, R.K. Feature selection for text categorization on imbalanced data. *SIGKDD Explor.* **2004**, *6*, 80–89. [CrossRef]
50. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
51. Wayne, D. *Bioestadística: Base para el Análisis de las Ciencias de la Salud*; Technical Report; Limusa: Mexico City, Mexico, 1983.
52. GECCO. *Genetic and Evolutionary Computation-GECCO 2003: Genetic and Evolutionary Computation Conference; Chicago, IL, USA, July 12–16; Proceedings. 1 (2003)*; Springer: Berlin/Heidelberg, Germany, 2003.

53. Hsu, C.W.; Chang, C.C.; Lin, C.J. *A Practical Guide to Support Vector Classification*; National Taiwan University: Taipei, Taiwan, 2003.
54. Kim, T.; Kim, J.; Won, J.; Park, M.; Lee, S.; Yoon, S.; Kim, H.R.; Ko, K.; Rhee, B. A decision tree-based approach for identifying urban-rural differences in metabolic syndrome risk factors in the adult Korean population. *J. Endocrinol. Investig.* **2012**, *35*, 847–852.
55. Shamah-Levy, T.; Cuevas-Nasu, L.; Rivera-Dommarco, J.; Hernández-Ávila, M. Encuesta Nacional de Nutrición y Salud de Medio Camino 2016 (ENSANUT MC 2016). In *Informe Final de Resultados*. 2016. Available online: https://www.insp.mx/ensanut/medio-camino-16 (accessed on 15 March 2020).
56. Heymsfield, S.B.; Peterson, C.M.; Thomas, D.M.; Heo, M.; Schuna, J., Jr. Why are there race/ethnic differences in adult body mass index–adiposity relationships? A quantitative critical review. *Obes. Rev.* **2016**, *17*, 262–275. [CrossRef]
57. Ashwell, M.; Gunn, P.; Gibson, S. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: Systematic review and meta-analysis. *Obes. Rev.* **2012**, *13*, 275–286. [CrossRef]
58. Khader, Y.S.; Batieha, A.; Jaddou, H.; Batieha, Z.; El-Khateeb, M.; Ajlouni, K. Anthropometric cutoff values for detecting metabolic abnormalities in Jordanian adults. *Diabetes Metab. Syndr. Obes. Targets Ther.* **2010**, *3*, 395. [CrossRef]
59. Browning, L.M.; Hsieh, S.D.; Ashwell, M. A systematic review of waist-to-height ratio as a screening tool for the prediction of cardiovascular disease and diabetes: 0·5 could be a suitable global boundary value. *Nutr. Res. Rev.* **2010**, *23*, 247–269. [CrossRef] [PubMed]
60. Latifi, S.M.; Rashidi, H.; Shahbazian, H. The most appropriate cut-off point of anthropometric indices in predicting the incidence of metabolic syndrome and its components. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2019**, *13*, 2739–2745.
61. Rezende, A.C.; Souza, L.G.; Jardim, T.V.; Perillo, N.B.; Araújo, Y.C.L.; de Souza, S.G.; Sousa, A.L.L.; Moreira, H.G.; de Souza, W.K.S.B.; Peixoto, M.d.R.G.; et al. Is waist-to-height ratio the best predictive indicator of hypertension incidence? A cohort study. *BMC Public Health* **2018**, *18*, 281. [CrossRef]
62. Pavanello, C.; Zanaboni, A.M.; Gaito, S.; Botta, M.; Mombelli, G.; Sirtori, C.R.; Ruscica, M. Influence of body variables in the development of metabolic syndrome—A long term follow-up study. *PLoS ONE* **2018**, *13*, e0192751. [CrossRef]
63. Romero-Saldaña, M.; Fuentes-Jiménez, F.J.; Vaquero-Abellán, M.; Álvarez-Fernández, C.; Aguilera-López, M.D.; Molina-Recio, G. Predictive Capacity and Cutoff Value of Waist-to-Height Ratio in the Incidence of Metabolic Syndrome. *Clin. Nurs. Res.* **2019**, *28*, 676–691. [CrossRef] [PubMed]
64. Deo, R.C. Machine learning in medicine. *Circulation* **2015**, *132*, 1920–1930. [CrossRef] [PubMed]
65. Deist, T.M.; Dankers, F.J.; Valdes, G.; Wijsman, R.; Hsu, I.C.; Oberije, C.; Lustberg, T.; van Soest, J.; Hoebers, F.; Jochems, A.; et al. Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers. *Med. Phys.* **2018**, *45*, 3449–3459. [CrossRef] [PubMed]
66. Acharjee, A.; Ament, Z.; West, J.A.; Stanley, E.; Griffin, J.L. Integration of metabolomics, lipidomics and clinical data using a machine learning method. *BMC Bioinform.* **2016**, *17*, 440. [CrossRef]
67. Hollister, E.B.; Oezguen, N.; Chumpitazi, B.P.; Luna, R.A.; Weidler, E.M.; Rubio-Gonzales, M.; Dahdouli, M.; Cope, J.L.; Mistretta, T.A.; Raza, S.; et al. Leveraging Human Microbiome Features to Diagnose and Stratify Children with Irritable Bowel Syndrome. *J. Mol. Diagn.* **2019**, *21*, 449–461. [CrossRef]