Innovative Integration of Machine Learning Techniques for Early Prediction of Metabolic Syndrome Risk Factors

Shendry Balmore Vásquez Rosero[1,2(✉)]

[1] Universidad Estatal Península de Santa Elena, Av Principal Santa Elena, La Libertad, Ecuador
srosero@upse.edu.ec

[2] Universidad Politécnica Salesiana, Campus El Vecino, Sede Cuenca, Ecuador

Abstract. Over the past two decades, chronic degenerative diseases have risen to prominence in global and national morbidity and mortality statistics. Notably, type 2 diabetes mellitus, arterial hypertension, and metabolic syndrome have been highlighted for their prevalence and have been identified by the World Health Organization (WHO) as potential causes of 50% of worldwide fatalities. Despite increased awareness driven by internet dissemination about risks associated with sedentary lifestyles and poor diets, and the subsequent shift in public perception towards healthier living, it remains a reality that individual concern typically arises following the initial symptomatology of these conditions. In response to this situation, the current study proposes the development of an early warning system, underpinned by advanced machine learning algorithms such as LightGBM, XGBoost, and ensemble methods based on Random Forests that employ gradient boosting techniques to enhance predictive accuracy. This model processes data efficiently, requiring minimal computational resources, to provide personalized risk predictions based on categorical characteristics, as well as biometric and clinical variables. In light of the findings, the deployment of a web platform is anticipated, which will enable individuals to assess their health risk using readily available indicators, thereby promoting greater proactivity in personal health management

Keywords: Síndrome metabólico · Machine Learning · LightGBM · XGBoost y Random Forest

1 Introduction

Sedentary behavior, understood as the lack of regular physical activity, has become an increasing concern in modern society. Specifically, in Ecuador, as in many parts of the world, sedentariness is a public health issue affecting diverse population groups, including students, workers, and university faculty. This sedentary trend carries with it a series of negative health consequences,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14818, pp. 20–36, 2024.
https://doi.org/10.1007/978-3-031-65273-8_2

one of the most prominent being the emergence and rise of metabolic syndrome. It is necessary to understand that metabolic syndrome is a complex medical condition characterized by the presence of various cardiovascular risk factors, including abdominal obesity, high blood pressure, elevated blood glucose levels, low levels of high-density lipoprotein (HDL), and elevated blood triglyceride levels. This combination of factors significantly increases the risk of cardiovascular diseases and type 2 diabetes.

Within the Ecuadorian university context, this issue worsens due to the long hours of sedentary study and work faced by students, workers, and faculty. The lack of regular physical activity, combined with unhealthy eating habits, creates a conducive environment for the development of metabolic syndrome. That is why many universities in Ecuador, especially the State University of Santa Elena Peninsula since 2010, have been carrying out a program and various occupational health proposals aimed at understanding the relationship between cardiovascular risk indicators and the sedentariness of the main actors in the university community. Therefore, the present study aims to gather part of that research conducted between 2018 and 2022 to implement a risk prediction model based on robust machine learning models that later serve as a diagnostic and prediction guide with the aim of achieving an early diagnosis of metabolic syndrome based on risk indicators. For this purpose, the following sections have been proposed, which have allowed the implementation of a novel risk prediction method for metabolic syndrome based on risk indicators:

1. Theoretical Framework, which includes the foundations of Machine Learning and the justification for applying these methods in predicting the risk of metabolic syndrome
2. Explanation of the Proposed Method
3. Experimental Design
4. Experimentation
5. Results and Discussion
6. Conclusions and, finally,
7. Bibliographic References

2 Theoretical Framework

During the last twenty years, it has been observed that chronic and degenerative diseases lead the statistics of morbidity and mortality both globally and nationally. Among these diseases are type 2 diabetes, hypertension, metabolic syndrome, and lipid disorders, which are widely distributed and have a high global prevalence. The World Health Organization (WHO) has indicated that these diseases could be responsible for 50% of all deaths globally, anticipating a 17% increase in the next decade [5,11].

In America, it is estimated that non-communicable chronic diseases are responsible for about 5.5 million deaths annually, which represents 81% of all deaths. Of these, 2.2 million occur in people under 70 years of age. Breaking

down by regions, North America records approximately 2.7 million deaths; in Mexico, Latin America, and the Caribbean, deaths total 868,600; and in South America, about 1.94 million. Cardiovascular diseases and diabetes, along with other ailments, account for 28.1% and 20.8% of these deaths, respectively [12]. These conditions have also contributed to an increase in premature mortality, with the risk of dying from one of these diseases between the ages of 30 and 70 estimated at 15% to 18% in men and 13% in women throughout the American continent as evidenced in [4].

On the other hand, within the field of computer science, the art of programming has played an important role over the last 20 years [2,10], a role that has been characterized by the introduction of variables in a structured way into the computer, which then delivers results to its programmer. It can be said that the computer program has enabled data processing to obtain information, but this process would not be executed without the programmer's intervention. Therefore, in recent years at a practical level, new paradigms have emerged in which the programmer plays a passive role since it is intended that the computer "learn" based on the analysis of the set of variables that the programmer has entered. It is then said that this is the basis of "*Machine Learning*" (ML) [3] or automated learning, where the computer learns by itself. For decades, traditional programming both in computers and online has been preeminent, with the development of algorithms and techniques to optimize program efficiency. But recently, machine learning has revolutionized this field. This approach has modified the previously explained paradigm to a new one. Now, instead of simply inputting data into the program, these data are used along with the obtained results to create the program, or 'model'. Hence, through techniques such as prediction, regression, clustering, and classification, names like supervised learning and unsupervised learning appear as the main algorithms of this ML.

Within these supervised and unsupervised techniques or algorithms [10], some models like *LightGBM*, *XGBoost*, and *Random Forest* [1] stand out for their exceptional performance in a variety of classification and regression tasks [8]. They have proven their effectiveness in numerous studies and competitions, making them solid options for tackling complex problems such as the prediction of risk factors [9], together with their ability to handle complex data and extract relevant patterns from multiple variables. It's also worth mentioning that while *LightGBM* and *XGBoost* are complex ensemble models, model interpretation techniques can be used to understand their decision-making. This is crucial for healthcare professionals and researchers to understand the predictions and take appropriate actions.

3 Proposed Method

To develop a novel approach that assesses the risk of individuals based on their occupation (workers, teachers, students) and gender, while using machine learning, a predictive model was considered that takes into account the interaction between these categorical characteristics and biometric and health variables. Below, the set of proposed steps for achieving the objectives is listed:

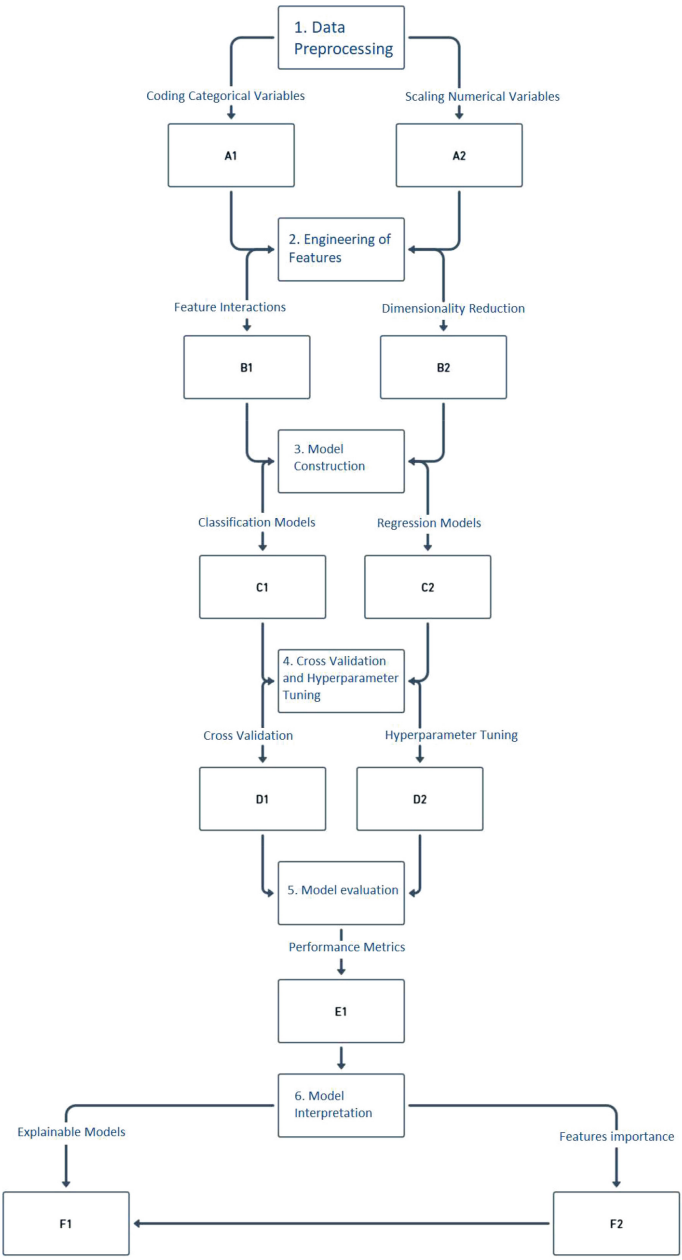


Fig. 1. Summary of the proposed model.

- Data Preprocessing: This step involved the encoding of categorical variables, i.e., converting categorical variables (gender, type of person) into numeric formats using one-hot encoding or label encoding. Additionally, scaling numeric variables, which entailed ensuring that all numeric variables are appropriately scaled so that the model is not biased towards variables with greater magnitude.
- Feature Engineering: This involves the interactions of features and the possibility of generating new features that represent the interaction between gender and occupation, since the risk might vary between a female student compared to a male worker. Likewise, dimensionality reduction with techniques such as Principal Component Analysis (PCA).
- Model Construction: Work was done on classification models, given the risk classification, classification models such as Random Forest and Gradient Boosting were used to predict the risk category. Regression models were also used because the risk is measured on a continuous scale, making regression models potentially the most appropriate.
- Cross-Validation and Hyperparameter Tuning: Cross-Validation to robustly evaluate the model's performance and Hyperparameter Tuning such as grid search or random search to find the best hyperparameters for the selected models.
- Model Evaluation: Performance Metrics such as recall, F1-score, ROC-AUC for classification, or R-squared, RMSE for regression, depending on the type of problem.
- Model Interpretation: Finally, the importance of features was analyzed to determine which are the most influential factors in predicting risk.

A summary of the proposed model can be seen in Fig. 1 on the following page.

3.1 Simplified Model Algorithm

- PHASE 1.- Data Set Loading, starts by loading the data set to inspect it and understand its structure; an Exploratory Data Analysis (EDA) is generated to gain a deep understanding of the data, including variable distributions, descriptive statistics, and possible correlations.
- PHASE 2.- Data Preprocessing, this involves cleaning the data before determining missing values or anomalies; proceeding to convert categorical variables to numeric formats using appropriate encoding. Normalize or standardize numerical variables to pave the way for Feature Engineering, which implies creating feature interactions if relevant. Finally, consider dimensionality reduction for cases with many variables.
- PHASE 3.- Construction and Selection of Models; Divide the data into training and test sets. Select candidate models and train them on the training set. Adjust hyperparameters using cross-validation. Finally, Model Evaluation: Evaluate each model on the test set. Compare performance metrics and select the most promising model.

– PHASE 4.- Model Interpretation: Analyze the importance of features. Use explainability techniques to understand the model's predictions.

Algorithm 1. General Algorithm

Input: Dataset loading

1: Phase 1: EDA Analysis
2: Phase 2: Preprocessing
3: Phase 3: Construction, selection, and optimization of models:
Repeat this phase until conclusive results are obtained

Output: Presentation (Accuracy, Precision, Recall, F1-Score)

In [6], similar works can be found where a similarity measure within the conventional process of the Forest and others algorithm is proposed. In fact they say:*"When the risk for new-onset diabetes was examined for the Framingham cohort, in both men and women, the presence of metabolic syndrome was highly predictive of new-onset diabetes. Almost half of the population-attributable risk for diabetes could be explained by the presence of ATP III metabolic syndrome."*. In the other hand the main idea is that a rating is unique for a group or for a user when that rating is different from the majority of ratings issued by other users, similar to [13] where the study collects data from a chain of health examination institution in Urumqi from 2017–2019, and performs 39,134 remaining data after preprocessing such as deletion and filling, concluding that *"The interpretability methods can be used as a novel means of identifying variables that are more likely to be good predictors. These predictors can be evaluated as features in other models developed with more appropriate datasets. In addition, it can also provide a methodological reference for the feasibility of applying the model interpretability method in health examination data mining"* .

4 Experimental Design

4.1 Dataset Characteristics

The dataset to be used consists of data collected between August 23, 2021, and May 31, 2022, as part of the Occupational Health program of the State University of Santa Elena (UPSE) led by Dr. Bertha Maggi, in charge of the Department of Well-being and Occupational Health at the University; to maintain the confidentiality of the data, all types of identifiers of the individuals to whom the control was directed have been removed in advance, so unless otherwise indicated, the data can be considered for general purposes without there being variables or indicators that could generate a bias that allows precisely determining a characteristic of the institution, characteristics that the author and collaborators in this preliminary study reserve for future formal and detailed analyses of occupational health at the University of Santa Elena. Under this context and based on a preliminary analysis, the data found in the dataset include:

- The"gender" column indicates the gender of the person (M/F).
- The column "type_person(W/T/S)" indicates whether the person is a Worker, Teacher, or Student.
- QS appears to be a qualitative marker, possibly an indicator of health status or a specific test; until the conclusion of this study, it was not possible to interview the person in charge of the data to provide a detailed definition of the QS field.
- Glycosylated is a value that does not seem to be correctly defined since all records show 0.0, which could be a data error; similarly to the previous field, until the conclusion of this study, it was not possible to interview the person in charge of the data to provide a detailed definition of the glycosylated field, from the literature reviewed and for subsequent analysis this field will be ignored, focusing on the "glycosylated hemoglobin" field which potentially represents greater importance when generating correlations with other variables.
- Glucose mg/dL shows the blood glucose level.
- Total Cholesterol, HDL (High-Density Lipoproteins), LDL (Low-Density Lipoproteins), and Triglycerides are measures related to the lipid profile.
- Glycosylated Hemoglobin is a measure of long-term blood glucose control.
- Weight Kg, Height meters, BMI (Body Mass Index), Abdominal Perimeter are anthropometric measures.
- SysBP and DiaBP are the systolic and diastolic blood pressure, respectively.
- Pulse and SO2 (probably oxygen saturation) are vital parameters.
- Diagnosis indicates whether the person is at risk or not at risk, probably in relation to a specific health condition.

The detail of the descriptive statistics is shown in Table 1.

Table 1. Descriptive statistical analysis of the dataset in general

	Blood Glucose	Total Cholesterol	HDL	LDL	Triglycerides	Glycated Hemoglobin	Weight Kg	Height mts	BMI	Abdominal Perimeter	SBP (Systolic Blood Pressure)	DBP (Diastolic Blood Pressure)	Pulse	Oxygen Saturation
count	1079	1079	1079	1079	1079	1079	1079	1079	1079	1079	1079	1079	1079	1079
mean	96,48	180,31	40,67	124,63	127,97	4,90	58,28	1,70	22,21	79,40	112,40	68,70	74,05	97,78
std	15,28	35,78	8,61	37,24	72,59	0,63	11,30	4,88	7,09	8,05	8,22	5,04	7,85	1,20
min	0,00	0,00	0,00	0,00	0,00	0,00	37,10	0,00	0,00	61,00	82,00	50,00	47,00	76,00
0,25	90,00	155,40	35,00	101,60	80,90	4,80	54,00	1,54	22,21	77,00	112,00	68,70	72,00	97,78
0,50	95,20	176,50	39,20	117,10	113,00	4,80	54,00	1,54	22,21	77,00	112,00	68,70	72,00	97,78
0,75	100,80	201,00	45,05	145,20	152,40	4,80	58,00	1,54	24,05	77,95	112,00	68,70	73,00	98,00
max	352,80	478,00	100,00	423,40	791,00	13,80	132,70	161,70	62,10	132,00	193,00	99,00	116,00	99,00

Summary of Numerical Variables: Most of the numerical variables appear to have no missing values. Some variables have a minimum value of 0, which could be a missing or atypical value for certain biometric measures such as Glucose mg/dL, Total Cholesterol, HDL, LDL, and Triglycerides. This may require further investigation. The HEIGHT meters variable has a maximum value of 161.7, which is evidently incorrect and may be a data entry error. The variables WEIGHT Kg, BMI, and ABDOMINAL PERIMETER seem to have values that could be realistic, but we would need to confirm if the range of values is expected

for the study population. The blood pressure variables (systolic, diastolic) and PULSE have very consistent values for the 25%, 50%, and 75% percentiles, suggesting that they could have been imputed or that the data is not as varied as expected. The SO2 (oxygen saturation) variable has a very low standard deviation and an average very close to the maximum of 99%, which could indicate a lack of variability or possible outliers.

Summary of Categorical Variables:

There are two genders (M, F) and three types of person (Worker, Teacher, Student). The diagnosis category has two values (At Risk, Not At Risk), and most individuals are labeled as Not At Risk.

Missing Values:

There do not appear to be any explicit missing values in the dataset.

Next Steps:

- Verify outliers and errors: It's important to check for outliers or possible entry errors, especially for values that are 0 or exceptionally high where they shouldn't be.
- Distribution analysis: Proceed to visualize the distribution of both numerical and categorical variables to better understand the characteristics of the study population.
- Investigate Consistent Variables: Decide to thoroughly analyze variables with little variability to determine if these are imputed values or if they reflect the reality of the dataset.
- Verify the Scale of Variables: Given that some variables might need to be rescaled or transformed to be effectively used in a machine learning model, proceed with the analysis of outliers and potential entry errors, and then visualize the distribution of some of the key variables to gain a better understanding of the data Figs. 3, 4, 5.

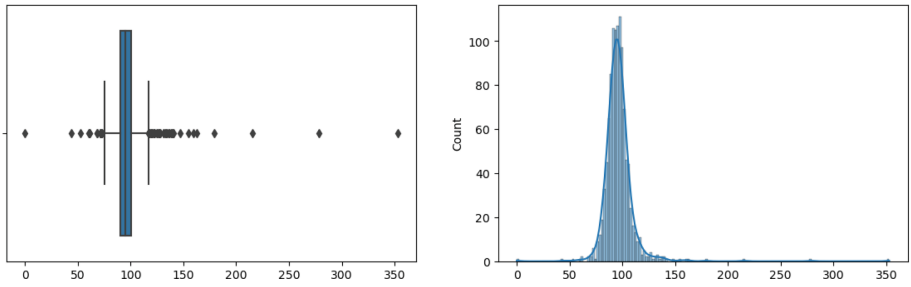


Fig. 2. Glucose analysis

The visualization of variables can be seen in the Figs. 10, 7, 8, 9 among the most important:

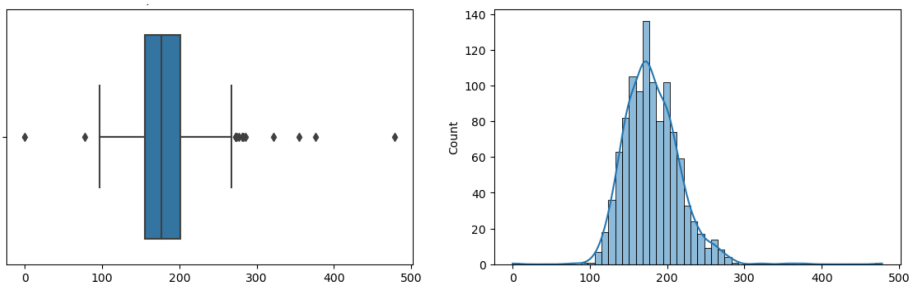


Fig. 3. Cholesterol analysis

The visualizations of the distributions of the selected variables show:

- Glucose mg/dL: Most values are within a normal range, but there are some very high outliers that could represent real cases of elevated glucose levels or possible entry errors.
- Total Cholesterol: Similar to glucose, there is a relatively normal distribution with some high outliers.
- HEIGHT meters: Most heights are within the expected range, though there are extremely high values surpassing realistic values for human height, which clearly indicates a data entry error.
- WEIGHT Kg: Weight appears to have a normal distribution with some outliers at the upper end.

5 Experimentation

As part of the experimentation, possible errors in the data were corrected. This involved correcting the extremely high values in HEIGHT (meters) which could obviously be considered as errors in data entry during digitization. On the other hand, for the outliers, it was necessary to decide whether to trim them, transform them, or keep them depending on whether they are representative of the population or not. Additionally, the consistency of variables was verified, that is, reviewing variables with little variability, such as SO2, to make the decision whether to keep them in their current form. These corrections can be seen in Figures 10, 7, 8, 9:

After this process, the corrections and adjustments were applied successfully to the data, which involved:

- Height (meters): After correcting values that seemed to be in centimeters, the distribution now appears more realistic and aligns with what is expected for human height.

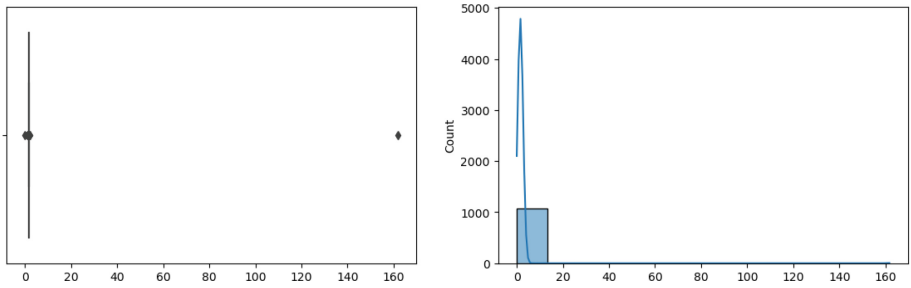


Fig. 4. Height analysis

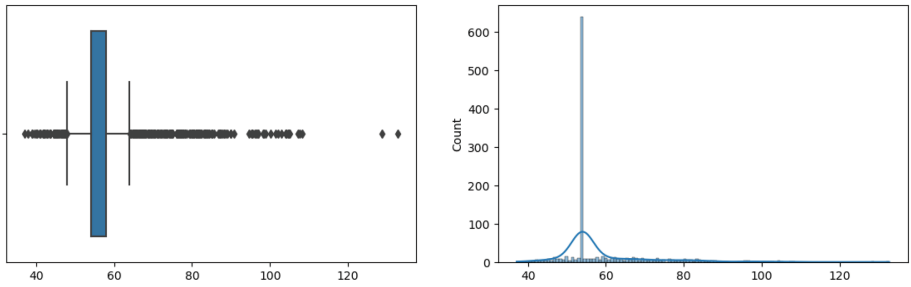


Fig. 5. Weight analysis

– Glucose mg/dL, Total Cholesterol, Weight Kg: By trimming values at the 1st and 99th percentiles, we managed outliers without eliminating corresponding data. The distributions of these variables now show less extreme variation and are more consistent with what we would expect in a typical population.

The next step was to investigate the SO2 variable due to its low variability and any other variables that might require further inspection. Additionally, preparing categorical variables for modeling through appropriate encoding and normalizing numerical variables as necessary.

Below are the data for the SO2 variable in the Fig. 10:

The visualization of the SO2 variable reveals very low variability, and it has been confirmed that it is not a variable with variance close to zero, which means that there could be some useful variability in it. However, the low variability suggests that it may not be very informative for machine learning models.

Regarding the normalization of numerical variables, errors were encountered when attempting to exclude certain variables that do not exist under the'float64' or'int64' data types. Steps were taken to correct this and ensure that only relevant numerical variables were normalized. Additionally, the presence of the'QS' and'glycosylated' variables was verified before proceeding with the normalization of the remaining numerical variables.

Once the numerical variables were successfully normalized, and the categorical variables were encoded, the SO2 variable was reanalyzed, where it was found

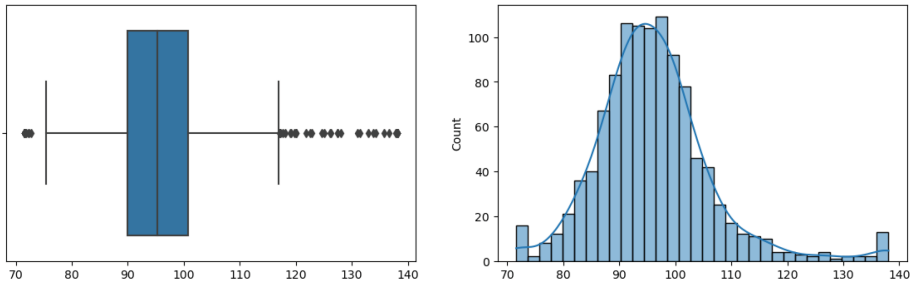


Fig. 6. Glucose analysis

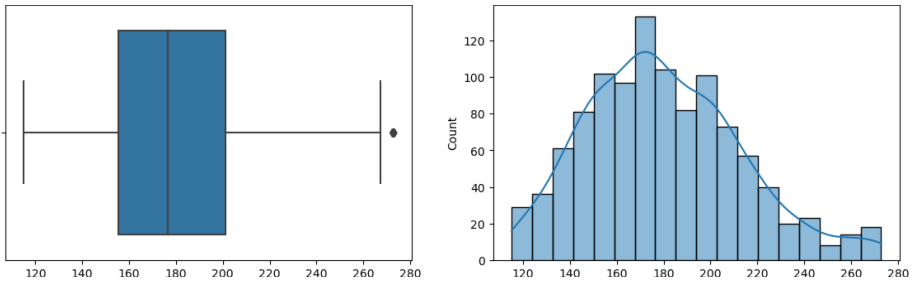


Fig. 7. Cholesterol analysis

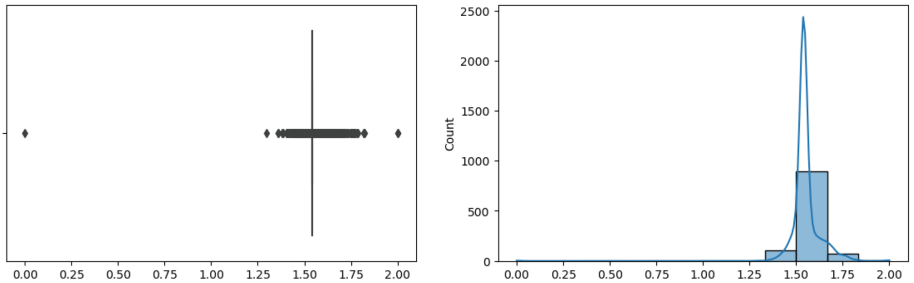


Fig. 8. Height analysis

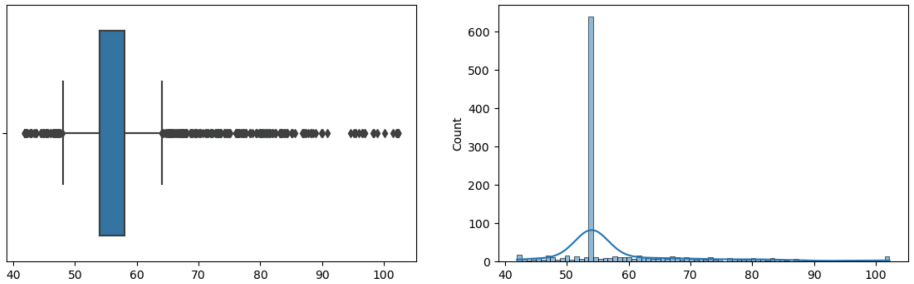


Fig. 9. Weight analysis

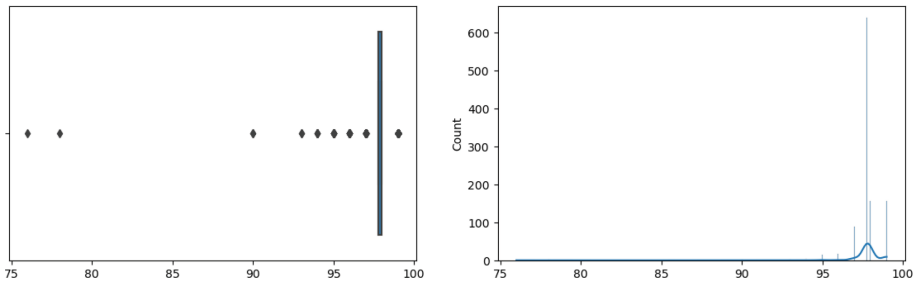


Fig. 10. Oxygen Saturation Analysis

to have low variability and not have variance close to zero, so it was estimated that it could still provide information to the models. With this background, the following steps were taken:

- Handling the target variable; that is, encoding the target diagnostic variable appropriately for modeling. It should be noted that if it is a binary classification problem, it must be ensured that it is encoded as 0 and 1.
- Splitting the data into training and test sets to validate the performance of the machine learning models.
- Selecting and training models after choosing the appropriate machine learning models for the binary classification task and training them using the training data.

Once the data has been successfully divided into training and test sets, we obtained 863 samples in the training set and 216 samples in the test set. The features (X) have 19 columns, which are the independent variables after encoding and normalization. The target variable (y) is a binary vector representing the presence or absence of risk.

The next step was to select and train suitable machine learning models for the binary classification task. Ensemble models of decision trees were chosen, specifically the classifier *Random Forest*, *LightGBM* and *XGBoost*.

The choice of *Random Forest, LightGBM,* and *XGBoost* as training models for classification was made because *Random Forest* maintains robustness and accuracy as this method is known for its high accuracy and robustness. By using multiple decision trees and averaging their predictions, this model reduces the risk of overfitting, which is especially useful in datasets with a lot of variability. Similarly, concerning the handling of nonlinear data, it is effective for classifying nonlinear data, something common in many real-world problems. Additionally, due to the feature importance, it provides a good indication of feature importance, which was useful for understanding the model and the problem, not to mention that it is less sensitive to outliers, as by its ensemble nature, it is less sensitive to outliers compared to other algorithms. On the other hand, *LightGBM* maintains efficiency in large datasets as it uses gradient-based algorithms. Likewise, due to its low memory usage and training speed, much faster than

Table 2. Comparison of Classifier Results

Method	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9352	0.94	0.89	0.94
XGBoost	0.9537	0.96	0.95	0.96
LightGBM	0.9630	0.97	0.95	0.96

other models, which is crucial when working with large volumes of data. It also efficiently handles categorical features without extensive preprocessing, simplifying the data preparation process, and finally, due to its high performance with shallow trees, as it performs well even with shallow trees, which is beneficial for avoiding overfitting. Lastly, *XGBoost* was chosen for its performance and speed as *XGBoost* is another gradient boosting algorithm, known for its performance and speed in machine learning model competitions. Alongside this, it was chosen because it includes regularization (L1 and L2), which helps prevent overfitting and improves the overall model accuracy, as well as its flexibility because it allows users to define custom optimization objectives and evaluation criteria, increasing its flexibility to address different classification problems.

6 Results and Discussion

In this section, the most relevant results following the phases of the proposed method are presented. The calculation of singular values and measures was carried out using the Train set. The best parameter values for the methods have been highlighted in the Table.2.

Random Forest results

Overall Accuracy (accuracy): 94.44%, indicating that the model has correctly predicted the risk classification in approximately 94.44% of the cases in the test set.

For the"No Risk" (sin riesgo) class (encoded as 0):

- Accuracy: 97%, meaning that when the model predicts "No Risk," it is correct 97% of the time.
- Recall: 96%, indicating that the model captures 96% of the actual "No Risk" cases.
- F1-Score: 97%, a weighted average of precision and recall.

For the "Risk" class (encoded as 1):

- Accuracy: 81%, meaning that when the model predicts "Risk," it is correct 81% of the time.
- Recall: 85%, indicating that the model captures 85% of the actual "Risk" cases.
- F1-Score: 83%, a weighted average of precision and recall.

Conclusion. These results suggest that the model performs well overall, with excellent ability to identify "No Risk" cases. However, there is a slight room for improvement in identifying "Risk" cases, as indicated by the lower precision and recall for that class

LightGBM Results

The LightGBM model has shown outstanding performance in binary classification, as detailed below: **Overall Accuracy**

The model's accuracy is 96.30%, indicating a high efficacy in correctly predicting classification in the test set.

For the "No Risk" class (encoded as 0):

- Precision: 99% (The model correctly predicts 99% of instances).
- Recall: 97% (The model identifies 97% of all actual instances of the class).
- F1-Score: 98% (High weighted average of precision and recall).
- Support: 182 (Number of actual instances in the test set).

For the "Risk" class (encoded as 1):

- Precision: 84% (The model correctly predicts 84% of instances).
- Recall: 94% (The model identifies 94% of all actual instances of the class).
- F1-Score: 89% (Good performance but not as high as for class 0).
- Support: 34 (Number of actual instances in the test set).

Averages

- Macro Avg:
  - Precision: 0.92
  - Recall: 0.95
  - F1-Score: 0.93
- Weighted Avg:
  - Precision: 0.97
  - Recall: 0.96
  - F1-Score: 0.96

Conclusion These results suggest that the **LightGBM** model possesses a remarkable ability to identify classes, with particularly strong performance in the "No Risk" class. The effectiveness of gradient boosting tree-based ensemble models is demonstrated for this type of classification tasks.

XGBoost Results

The results of the XGBoost model in binary classification are detailed below:

Overall Accuracy

The accuracy of the model is 95.83%, indicating high efficacy in correctly predicting classification in the test set.

For the "No Risk" Class (coded as 0):

- Precision: 0.99 (The model correctly predicts 99% of instances).
- Recall: 0.96 (The model identifies 96% of all real instances of the class).
- F1-Score: 0.97 (High weighted average of precision and recall).
- Support: 182 (Number of real instances of class 0 in the test set).

For the "Risk" Class (coded as 1):

- Precision: 0.82 (The model correctly predicts 82% of instances).
- Recall: 0.94 (The model identifies 94% of all real instances of the class).
- F1-Score: 0.88 (Good performance but not as high as for class 0).
- Support: 34 (Number of real instances of class 1 in the test set).

Averages

- Macro Avg:
  - Precision: 0.90
  - Recall: 0.95
  - F1-Score: 0.93
- Weighted Avg:
  - Precision: 0.96
  - Recall: 0.96
  - F1-Score: 0.96

Conclusion. The XGBoost model has proven to be highly effective in classifying the "No Risk" class and quite effective in classifying the "Risk" class. Although the precision for the "Risk" class is lower than for the "No Risk" class, the high recall indicates that the model is capable of identifying most real risk cases, which is positive especially if it is critical not to overlook risk cases. The difference between precision and recall for the "Risk" class could be an area of focus for improving the model.

Discussion

These results indicate that the ***LightGBM*** model has an even greater ability than the ***XGBoost*** model to correctly identify both risk and "No Risk" cases, with slightly better precision and F1-Score for the "Risk" class.

This demonstrates the effectiveness of ensemble models based on gradient boosting trees for this type of classification task. Both ***XGBoost*** and ***LightGBM*** are known for their high precision and performance in classification tasks and are robust choices for binary classification.

Given the high performance of both models, the choice between them could be based on other factors such as training time, interpretability, or performance on an additional validation dataset if available.

7 Conclusions

In this study, the performance of three machine learning models (*LightGBM, XGBoost, and Random Forest*) for predicting the risk of metabolic syndrome in workers, students, and teachers from an institution with a rigorous occupational health program has been evaluated. The results obtained reveal a high degree of effectiveness in classifying workers into risk and no-risk categories for developing metabolic syndrome.

The *LightGBM* model stands out as the most effective in the classification task, achieving an overall accuracy of 96.30%. Furthermore, it demonstrates exceptional performance in identifying risk cases, with a precision of 99% and a recall of 97% for the "No Risk" class. This indicates that the model is highly precise and capable of identifying the majority of real "No Risk" cases. For the "Risk" class, the *LightGBM* model achieves a precision of 84% and a recall of 94%, indicating significant capability in identifying risk cases.

The *XGBoost* model also presents notable results, with an overall accuracy of 95.83%. Although its precision for the "Risk" class is slightly lower than that of the *LightGBM* model, its high recall of 94% indicates a solid ability to capture real "Risk" cases. This is essential for early detection of potential cases of metabolic syndrome.

Lastly, the *Random Forest* model shows respectable performance, with an overall accuracy of 94.44%. Its results indicate a reasonable ability to classify workers into risk and no-risk categories. However, its precision and recall for the "Risk" class are lower than those of the other evaluated models.

Collectively, these results support the effectiveness of ensemble models based on gradient boosting trees (*LightGBM and XGBoost*) in predicting the risk of metabolic syndrome in workers from institutions with occupational health programs. These models have high potential to enhance early detection and risk management of metabolic syndrome, which can be crucial for the health and well-being of workers.

Furthermore, it is worth noting that the choice between the *LightGBM* and *XGBoost* models may depend on other factors such as training time, interpretability, and performance on an additional validation dataset if available. Both models represent solid and reliable options for binary classification in this context.

This study provides a solid foundation for future research and practical applications in the field of occupational health and metabolic syndrome risk management.

References

1. Ahn, J.M., Kim, J., Kim, K.: Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting. Toxins **15**(10), 608 (2023). https://doi.org/10.3390/toxins15100608
2. Canelo Sotelo, C.A., Espinoza Haro, P.C. (Dir.).: Redes neuronales artificiales y máquina con soporte vectorial para clasificar a los solicitantes de microcrédito (ed.). Universidad Nacional de Ingeniería (2021)
3. Dell'Aversana, P.: A Global Approach to Data Value Maximization: Integration, Machine Learning and Multimodal Analysis (ed.). Cambridge Scholars Publishing (2019)
4. Gómez Mendoza, R.A.(Il.)., Carrillo Esper, R. (Il.)., González Chávez, A.(Il.).: (2022). Síndrome metabólico y su impacto, CMMI Vol. 08: (1 ed.). Editorial Alfil, S. A. de C. V
5. González Chávez, A., López Espinosa, R., Mesa Pérez, J.A.: Manual práctico para la detección y el tratamiento integral del síndrome metabólico: (2 ed.). Editorial Alfil, S.A. de C.V (2019)
6. Horvat, T., Havaš, L., Srpak, D., Havaš, J., Srpak, J., Crčić, D.: predicting metabolic syndrome risks using machine learning methods. In: 2023 IEEE 21st Jubilee International Symposium on Intelligent Systems and Informatics (SISY), Pula, Croatia, 2023, pp. 000017–000022. https://doi.org/10.1109/SISY60376.2023.10417918. keywords: Sociology;Medical services;Predictive models;Prediction algorithms;Statistics;Biomedical monitoring;Medical diagnostic imaging;artificial intelligence;machine learning;metabolic syndrome;WEKA
7. Lee, W.: Python Machine Learning: (ed.). Wiley (2019). https://bibliotecas.ups.edu.ec:3488/es/lc/bibliotecaups/titulos/186090
8. Liang, W., et al.: Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. Mathematics (2227-7390), [s. l.], vol. 8, n. 5, pp. 765 (2020). https://doi.org/10.3390/math8050765
9. Ortega Candel, J.M.: Big data, machine learning y data science en Python: (1 ed.). RA-MA Editorial (2022)
10. Restrepo Leal, D.A., Viloria Porto, J.P., Robles Algarín, C.A.: El camino a las redes neuronales artificiales: (1 ed.). Editorial Unimagdalena (2021)
11. Samaniego Bañuelos, F.: ¿De qué diabetes estamos hablando?: (ed.). Editorial ITACA (2019)
12. Villegas Abrill, C.B., Paz Aliaga, A. (Dir.).: Efecto de una intervención educativa-nutricional en el síndrome metabólico premórbido y estado nutricional de docentes de la Universidad Nacional del Altiplano de Puno: (ed.). D - Universidad Nacional San Agustín de Arequipa (2022)
13. Zhang, Y., Zhang, X., Razbek, J., et al.: Opening the black box: interpretable machine learning for predictor finding of metabolic syndrome. BMC Endocr. Disord. **22**(1), 214 (2022). https://doi.org/10.1186/s12902-022-01121-4