Predictive model based on machine learning to identify sleep-related health problems

Laberiano Andrade-Arenas[1], Inoc Rubio Paucar[2], Margarita Giraldo Retuerto[1], Cesar Yactayo-Arias[3]
[1]Faculty of Science and Engineering, Universidad de Ciencias y Humanidades, Lima, Perú
[2]Faculty of Engineering and Business, Universidad Privada Norbert Wiener, Lima, Perú
[3]Department of General Studies, Universidad Continental, Lima, Perú

Article Info

Article history:

Received Jul 19, 2025
Revised Jul 13, 2026
Accepted Jul 21, 2026

Keywords:

Extreme gradient boosting
Knowledge discovery in databases methodology
Machine learning
Predictive model
Public health monitoring
Sleep quality

ABSTRACT

Sleep quality has become a growing public health issue worldwide, mainly due to a lack of awareness about its long-term consequences. Despite existing strategies to address this problem, there remains a need for more effective approaches. In this study, an early detection model for sleep disorders was implemented using the extreme gradient boosting (XGBoost) algorithm, following the knowledge discovery in databases (KDD) methodology, which includes the phases of selection, preprocessing, transformation, data mining, and interpretation. A dataset extracted from the Kaggle platform in CSV format was used, consisting of 374 records. With an overall accuracy of 91.5%, a recall of 100% for the insomnia class, and a precision of 100% for sleep apnea, the proposed model demonstrated exceptional performance. It also received an area under the curve (AUC) of 0.909 and an average F1-score of 0.913. With a mean accuracy of 91%, a 95% confidence interval (0.89–0.94), and a p-value of 0.0012, cross-validation confirmed its robustness and showed a statistically significant change from the baseline model. The error rates remained within clinically acceptable ranges, confirming its applicability as a diagnostic support tool. Overall, the results demonstrate the effectiveness of the model in identifying patterns related to sleep disorders.

Corresponding Author:

Cesar Yactayo-Arias
Department of General Studies, Universidad Continental
Lima, Perú
Email: cyactayo@continental.edu.pe

1. INTRODUCTION

The scientific community is currently becoming more interested in learning more about human sleep indicators. Numerous studies have demonstrated that sleep quality significantly affects one's physical and mental health. Lifestyle choices, eating habits, stress levels, and particularly the frequency and intensity of personal physical activity are among the primary determinants of this quality [1], [2]. In response to this issue, data science and machine learning (ML) have emerged as essential tools for addressing the large-scale analysis of biometric and behavioral data related to sleep. These technologies also enable the identification of complex patterns and the extraction of key variables that facilitate informed decision-making in the fields of public health and the prevention of sleep disorders. The availability of valuable data on sleep habits, the calculation of relevant physiological indicators, and the assessment of lifestyle choices provide an unprecedented opportunity to investigate risk factors that affect the overall health of the population.

Although there is extensive evidence linking sleep to various aspects of health, there are still challenges in accurately predicting the potential negative health consequences caused by poor sleep patterns. Moreover, many studies rely on traditional statistical analyses, whose main limitation lies in their inability to capture nonlinear relationships between the studied variables [3], [4]. For this reason, there is a growing interest in complementing these approaches with ML techniques, which enable the modeling of complex patterns and the delivery of more precise and in-depth knowledge.

Sleep is a vital physiological function for human beings, and its sustained deprivation or disruption can significantly compromise functional performance in daily activities and promote the onset of various non-communicable chronic diseases [5]. In this regard, scientific evidence has shown that sleep disorders are closely associated with an increased risk of developing hypertension, type 2 diabetes mellitus, obesity, as well as mental health disorders, including anxiety and depression. Despite this, many individuals prioritize their daily responsibilities without considering that chronic sleep restriction represents a significant risk factor for their overall well-being. This neglect has contributed to a high prevalence of chronic stress conditions and sleep-related pathologies [6]. In light of this, people must become aware of the key aspects of achieving proper sleep at appropriate times in order to maintain a healthy life.

In this context, the application of ML techniques, framed within the knowledge discovery in databases (KDD) methodology, enables the exploration and analysis of large volumes of data with the aim of identifying non-trivial patterns and developing highly accurate predictive models. This approach not only strengthens research in public health and behavioral sciences but also provides advanced analytical tools for prevention, early diagnosis, and the personalization of health interventions [7], [8]. It is important to consider that ML is a key factor in generating knowledge that adds value in addressing a public health issue [9], [10].

The present research is justified by its capacity to generate applied knowledge in the field of preventive health through the design and implementation of an algorithmic model capable of predicting health risks based on the analysis of variables related to sleep quality and lifestyle. This contribution not only expands the scientific evidence base but also provides relevant inputs for decision-making by healthcare professionals, public policy makers, and developers of technologies aimed at promoting population well-being. The objective of this research is to develop a predictive model using ML techniques specifically the extreme gradient boosting (XGBoost) algorithm that enables the identification and prediction of potential impacts on physical and mental health based on the analysis of sleep patterns.

2. LITERATURE REVIEW

This section focuses on related works by various authors on the topic, thus providing a holistic overview for the investigation. Likewise, the theoretical foundations support the identification and justification of the variables to be studied. This comprehensive approach ensures a solid framework for the current study.

2.1. Related works

The article evaluates various ML models for predicting sleep disorders, considering their significant impact on public health. The study utilizes a dataset of 400 individual records classified by sleep disorder status (none, insomnia, and sleep apnea) and incorporates demographic, lifestyle, and health metrics. Different ML models, including logistic regression, decision trees, ensemble methods (like random forest (RF) and gradient boosting), support vector machines, and neural networks are assessed for their diagnostic accuracy. Key findings indicate that ensemble methods, particularly RF and extreme gradient boosting classifier (XGBClassifier), outperform other models, achieving high accuracy and F1-scores (up to 0.93). The results highlight the importance of advanced ensemble techniques in accurately diagnosing sleep disorders, suggesting their integration into clinical practices for improved outcomes. The research methodology includes data collection and preprocessing, model development, training, and evaluation using metrics like accuracy, precision, recall, and F1-score. The study emphasizes the complexity of sleep disorders and the potential of ML to enhance diagnostic tools in sleep medicine, advocating for future research into optimizing these models for real-world applications [11], [12].

Sleep problems pose an important challenge for human health. In this study, a cross-validated XGBoost–bidirectional long short-term memory (BiLSTM) model was implemented using 75 variables and a dataset of 10,765 aging-related records. The proposed model reached 97% accuracy and demonstrated better performance than traditional long short-term memory (LSTM) models and other machine-learning approaches. The results demonstrate its effectiveness in the diagnosis and treatment of sleep apnea by identifying risk factors. In previous studies, the XGBoost algorithm was also employed, applying training strategies focused on the most relevant features of the dataset. As in the previous study, k-fold cross-validation was implemented to optimize model performance. The results were obtained by retraining on subsets of the training set using test data, achieving an accuracy of 90.6% in predicting the different stages of sleep [13], [14].

On the other hand, some authors investigated obstructive sleep apnea (OSA) in the pediatric population, linking it to developmental aspects such as physical growth, cardiovascular function, and cognition through ML techniques. Using a sample of 3,139 children, the XGBoost algorithm was applied after randomly splitting the data into training and testing sets. In the testing set, the model achieved area under the curve (AUC) of 0.95, 0.88, and 0.88 for the classification of mild, moderate, and severe OSA, respectively, with accuracies of 90.45%, 85.67%, and 89.81%, outperforming regression-based models. Similarly, complementary approaches have been explored using techniques such as entropy and linear feature analysis for the selection of relevant variables. In this context, a model based on the XGBoost algorithm was trained using the Sleep Heart Health Study 2 dataset. Through feature selection strategies, the most significant attributes for diagnosis were identified. When evaluating the model with a validation subset composed of 422 participants not included in the training phase, an accuracy of 68.83%, a precision of 71.61%, and a recall of 67.84% were achieved, demonstrating acceptable performance in scenarios involving previously unseen data, as reported by [15], [16].

In contrast to other studies [17], [18] conducted a longitudinal analysis on 239 patients with insomnia using the XGBoost algorithm, which achieved a multiclass AUC of 80% and a binary AUC of 83%, with adequate sensitivity and specificity, facilitating the identification of patients with low response to conventional depression treatments. In line with previous studies on the relationship between sleep-wake cycle disorders and major depressive disorders in adolescents, various ML models were evaluated, with XGBoost standing out by achieving an AUC of 85% in predicting the risk of depression. Multiple investigations have demonstrated that the XGBoost algorithm is highly effective in anticipating the onset of depressive disorders associated with sleep-wake rhythm disturbances, thus supporting the development of personalized treatments and more precise preventive strategies. Likewise, conditions such as OSA, the comorbidity of insomnia and sleep apnea (COMISA), and insomnia as an independent disorder represent significant manifestations of sleep disorders. In order to predict the risk of these conditions, XGBoost was implemented on clinical data collected from medical centers, selecting nine relevant variables through the Shapley additive explanations (SHAP) explainability method. These features were used to construct the schedule, light, eat and drink, environment, physiology, and stress (SLEEPS) model, which demonstrated outstanding predictive capacity, with AUC values exceeding 89% across all three analyzed disorders. As a result, SLEEPS is positioned as an innovative diagnostic tool that enables early and accessible detection of sleep disorders without relying on invasive methods such as polysomnography.

In the present study, three predictive models were developed using the Python programming language: logistic regression, XGBClassifier, and support vector classifier (SVC). After evaluating their performance, the XGBClassifier model was identified as the most effective. The results showed that occupation is the main factor associated with insomnia, while body mass index (BMI) category significantly influences the occurrence of sleep apnea. Additionally, BMI, stress level, blood pressure, and sleep duration were identified as key variables affecting sleep quality. These findings suggest that interventions aimed at controlling BMI and blood pressure could significantly contribute to improving nighttime rest [19], [20].

On the other hand, a study focused on estimating the risk of insomnia in breast cancer patients used data obtained through surveys. In the predictions performed, logistic regression models with L2 regularization and the XGBoost algorithm achieved predictive accuracies of 71.5% and 70.6%, respectively, with AUC values of 0.76 and 0.75. Additionally, population subgroups at high risk of insomnia were found by using the RuleFit algorithm. The findings demonstrated that among breast cancer survivors, cancer-related fatigue is a significant predictor of the onset of sleeplessness. In contrast to the study that examined insomnia in patients with breast cancer, the current study used the XGBoost algorithm to find risk factors linked to sleep disturbances. A total of 7,929 patients who satisfied the inclusion criteria were included in the sample; their mean age was 49.2 years (standard deviation (SD) = 18.4), with 4,055 (51%) for women and 3,874 (49%) for men. The ethnic composition was distributed as follows: 2,885 (36%) white individuals, 2,144 (27%) African Americans, 1,639 (21%) Hispanics, and 1,261 (16%) from other ethnic groups. Of the 684 variables analyzed, 64 were statistically significant $p < 0.0001$ and were incorporated into the predictive model. The performance of the XGBoost model was reflected in an area under the receiver operating characteristic curve (AUROC) of 0.87, with both sensitivity and specificity equal to 0.77, demonstrating a robust capacity to identify risk factors related to sleep disorders [21], [22].

In a different study, Haque *et al.* [23] created a predictive model to assess the connection between academic achievement and sleeplessness in kids and teenagers between the ages of 4 and 17. For this purpose, the "Young minds matter" database was used, applying the Boruta algorithm in combination with the RF classifier to select the most relevant features. Subsequently, the tree-based pipeline optimization tool-classifier (TPOTClassifier) was used to identify the best-performing supervised models, evaluating the RF, XGBoost, decision tree, and Gaussian naive Bayes algorithms. As a result, the RF model demonstrated the best performance, achieving a precision of 99% and an accuracy of 95%, outperforming the other evaluated models and showing greater effectiveness in predicting insomnia in relation to academic performance.

2.2. Theoretical bases

2.2.1. Extreme gradient boosting

XGBoost is defined as a supervised ML algorithm based on decision trees. Its main goal is to perform classification and regression tasks with a high level of accuracy. It is based on the idea of gradient boosting, a method that entails training models one after the other, with each new model fixing the mistakes of the preceding ones. What makes XGBoost particularly effective is that it is optimized to deliver high speed, performance, and computational efficiency. One of its most notable features is its ability to handle large volumes of data quickly, thanks to a highly optimized internal architecture that includes parallel processing, efficient memory usage, and advanced tree pruning techniques [24], [25].

In addition, XGBoost incorporates regularization mechanisms (L1 and L2 penalties), which help prevent overfitting and enhance the model's generalization ability when applied to new data. Another significant advantage is its ability to automatically handle missing values, making it a robust and practical tool for real-world applications where incomplete data is common. The algorithm also allows fine-tuning of multiple hyperparameters, such as tree depth, learning rate, and the number of estimators, providing a high degree of customization and facilitating model optimization according to the specific problem [26], [27].

2.2.2. Sleep quality

Sleep quality is a multidimensional construct that refers to the efficiency, continuity, depth, and subjective perception of nighttime rest. Medically, it is assessed by considering multiple objective and subjective parameters, including sleep latency, sleep efficiency, nocturnal awakenings, total sleep duration, and the presence or absence of sleep disorders such as insomnia, sleep apnea, or periodic limb movements. From a clinical standpoint, healthy sleep is typically defined as taking no longer than about 30 minutes to fall asleep and maintaining a sleep efficiency the percentage of time actually sleeping while in bed above 85%, a total duration of 7 to 9 hours per night in adults, and a reduced number of prolonged nocturnal awakenings. Additionally, the individual's subjective perception upon waking-such as feeling rested and alert is a relevant component in the clinical assessment of sleep quality [28], [29].

Sleep quality is closely related to multiple neurophysiological functions, including memory consolidation, emotional regulation, cellular repair, and endocrine balance. Disruptions in sleep have been linked to an increased likelihood of psychiatric conditions including major depressive disorder and anxiety as well as long-term medical problems such as hypertension, type 2 diabetes, and cardiovascular disease. For this reason, evaluating sleep quality is considered a key element both in clinical settings and in research on general health [30].

Blue light, with wavelengths between 450 and 495 nanometers, is emitted both by the sun and by electronic devices such as cell phones, computers, and televisions. While beneficial during the day, its exposure at night can negatively affect sleep. The circadian rhythm, which regulates the sleep-wake cycle, is controlled by the suprachiasmatic nucleus (SCN) of the hypothalamus and responds primarily to ambient light. Melatonin, a key hormone for inducing sleep, is secreted in darkness, and its production decreases when exposed to light. When the retina detects blue light at night, specialized cells (intrinsically photosensitive retinal ganglion cells (ipRGCs) containing melanopsin) are activated and send signals to the SCN indicating that light is still present. This inhibits melatonin secretion, delays sleep onset, and disrupts the circadian rhythm. As a result, using blue light-emitting devices before bedtime can lead to insomnia, fatigue, and a decrease in sleep quality [31], [32]. It is important to raise awareness of this issue, as the majority of people use electronic devices excessively, often ignoring the health risks associated with blue light exposure.

3. METHOD

3.1. Definition of the knowledge discovery in databases methodology

The KDD methodology is an iterative and interactive process that integrates expert knowledge about a specific problem with a diverse set of data analysis techniques, particularly those based on ML algorithms. This process consists of five fundamental stages: selection, preprocessing, transformation, data mining, and interpretation/evaluation of results. Figure 1 shows that each of these phases is articulated in a sequential and logical manner, allowing input data to be cleaned, transformed, and subsequently analyzed in order to extract useful patterns or knowledge. Finally, the results are interpreted and evaluated according to the initial objectives [33]. One of the main advantages of this approach is its flexibility, as it allows returning to previous stages of the process if the results are unsatisfactory or if it becomes necessary to adjust certain assumptions. This ensures greater accuracy and coherence in the extraction of knowledge from large volumes of data. Ultimately, this methodology represents a robust and adaptable framework for various projects and promotes continuous improvement thanks to its iterative nature.

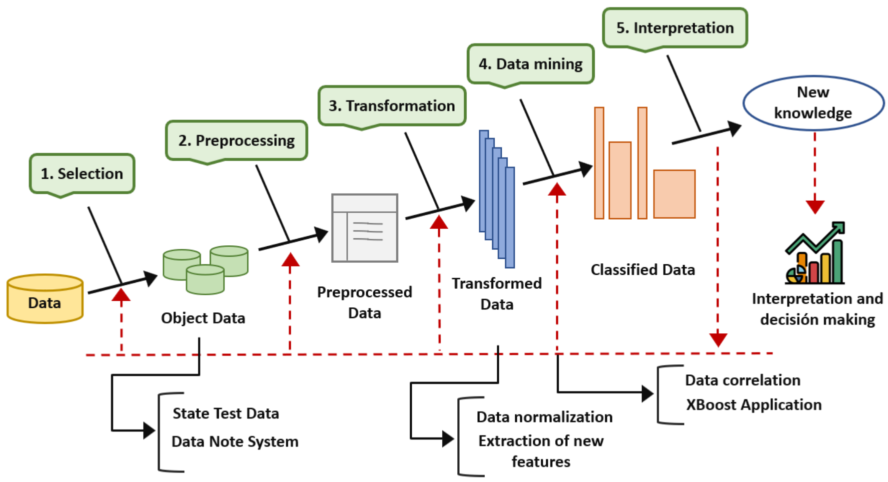


Figure 1. KDD methodology

3.1.1. Selection

In this section, an exhaustive search was conducted on the Kaggle platform, a widely used repository within the scientific and technological community due to its vast variety of public datasets from different fields of knowledge. During the search process, specific criteria were employed to ensure the relevance and quality of the selected dataset. Keywords such as sleep health, sleep disorders, sleep quality, mental health, among others, were used within Kaggle's search engine to find databases containing relevant variables related to sleep habits, associated disorders, and their possible connection to other dimensions of physical and mental health [34], [35]. Likewise, priority was given to selecting a dataset in CSV format due to its tabular structure, compatibility with statistical analysis tools (such as Python), and ease of manipulation. This format allows for efficient data processing, facilitating cleaning, segmentation, and descriptive or inferential analysis according to the objectives of the study.

3.1.2. Preprocessing

In the preprocessing stage, descriptive statistics reveal that the 374 participants have a mean age of 42.18 years (SD = 8.67; range = 27–59), with an average sleep duration of 7.13 hours (SD = 0.80; Q1 = 6.40; median = 7.20; Q3 = 7.80; minimum = 5.80; maximum = 8.50), and an average sleep quality score of 7.31 (SD = 1.20; range = 4–9). The average physical activity level is 59.17 (SD = 20.83; Q1 = 45; median = 60; Q3 = 75; range = 30–90), and the mean stress level is 5.39 (SD = 1.77; range = 3–8). The average heart rate is 70.17 bpm (SD = 4.14; Q1 = 68; median = 70; Q3 = 72; range = 65–86), and the average daily steps are 6,816.84 (SD = 1,617.92; Q1 = 5,600; median = 7,000; Q3 = 8,000; range = 3,000–10,000), as shown in Table 1. Finally, all numerical variables were standardized using the StandardScaler() function, and this information is presented in Table 2.

Table 1. Descriptive statistics of the dataset

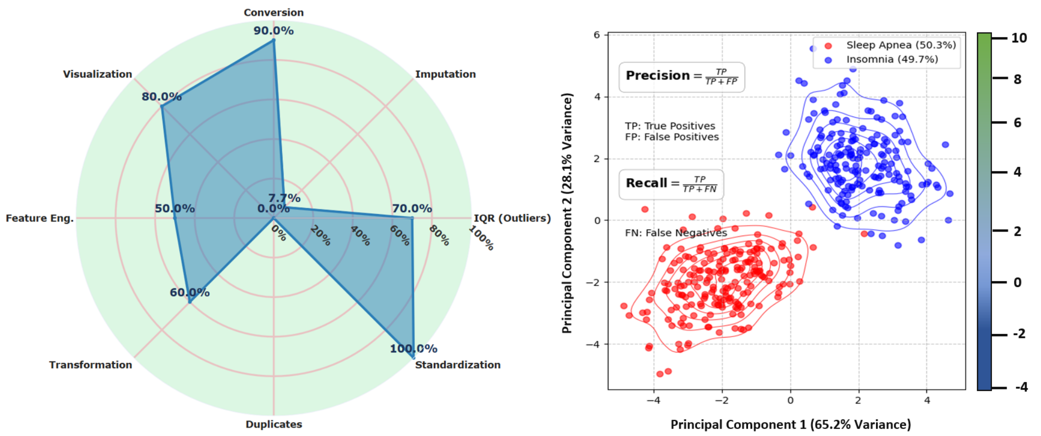
Statistics	Age	Sleep duration	Quality of sleep	Physical activity level	Stress level	Heart rate	Daily steps
Count	374.00	374.00	374.00	374.00	374.00	374.00	374.00
Mean	42.18	7.13	7.31	59.17	5.39	70.17	6816.84
SD	8.67	0.80	1.20	20.83	1.77	4.14	1617.92
Min	27.00	5.80	4.00	30.00	3.00	65.00	3000.00
Q1 (25%)	35.25	6.40	6.00	45.00	4.00	68.00	5600.00
Q2 (Median)	43.00	7.20	7.00	60.00	5.00	70.00	7000.00
Q3 (75%)	50.00	7.80	8.00	75.00	7.00	72.00	8000.00
Max	59.00	8.50	9.00	90.00	8.00	86.00	10000.00

Table 2. Technical steps for data preprocessing

Passed	Technical operation
1	Conversion to numeric: `pd.to_numeric(df['Age'], errors='coerce')`
2	Imputation of NA: `df['Sleep Duration'].fillna(median, inplace=True)`
3	Calculate IQR para `Heart Rate`: $\text{IQR} = Q_3 - Q_1$ Outliers = $\{x < Q_1 - 1.5 \times \text{IQR}\} \cup \{x > Q_3 + 1.5 \times \text{IQR}\}$
4	Standardization: `StandardScaler().fit_transform(df[column_numbers])`

3.1.3. Transformation

During the data transformation phase, several preprocessing techniques were applied to prepare the health and lifestyle dataset for analysis. Data conversion and standardization were performed extensively, ensuring that the variables were transformed into a consistent format suitable for model training. Additional procedures, including missing value imputation and duplicate record removal, were required only to a limited extent, while outlier detection, variable transformation, and feature engineering were applied selectively to improve data quality. Data visualization also played an important role in exploring the dataset, together with the analysis of the variance explained by the principal components, as illustrated in Figure 2. Specifically, Figure 2(a) summarizes the preprocessing workflow, showing that standardization (100%) and data conversion (90%) were the most extensively applied techniques, followed by data visualization (80%) and outlier detection using the interquartile range (IQR) method (70%). In contrast, missing value imputation (7.7%) was only occasionally necessary, and no duplicate records (0%) were detected, highlighting the high quality of the dataset prior to preprocessing. Figure 2(b) illustrates the two-dimensional representation generated using principal component analysis (PCA). The first two principal components explain 93.4% of the total variance (PC1 = 65.2% and PC2 = 28.2%), revealing a clear separation between the sleep apnea and insomnia classes with minimal overlap. This distribution suggests that the extracted features effectively preserve the underlying class structure, providing a solid foundation for the subsequent classification task.



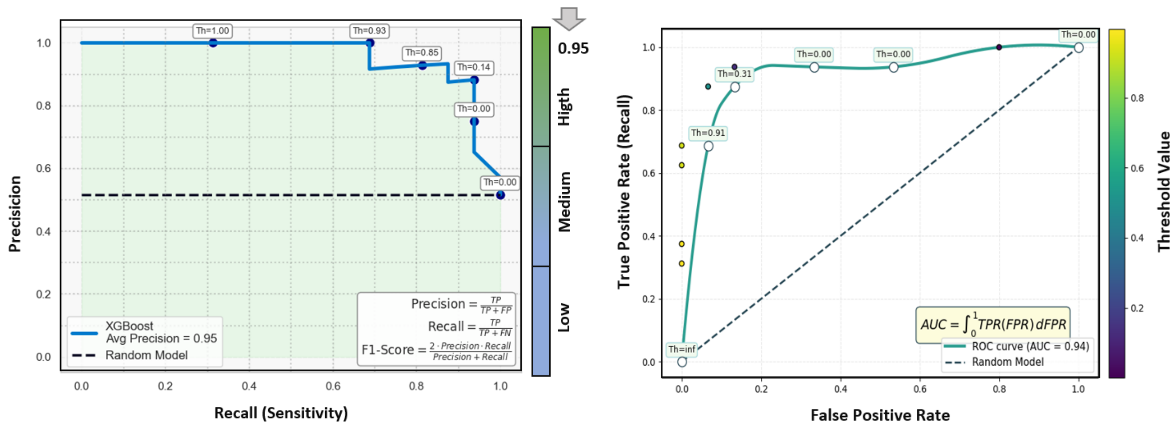
(a) (b)

Figure 2. Data manipulation and transformation of (a) percentages of features in the data and (b) percentage of variance in components

3.1.4. Data mining

During the data mining procedure, the required XGBoost algorithm was applied. The results in Figure 3 indicate that the XGBoost model demonstrates exceptional performance according to the precision-recall and receiver operating characteristic (ROC) curves. The precision-recall curve, with an average precision of 0.95, indicates that the model maintains high precision even with a high recall, making it highly effective at identifying positive classes, as in Figure 3(a). Simultaneously, the ROC curve, with an AUC of 0.94, confirms excellent discriminative ability, meaning that the model is very capable of distinguishing between positive and negative classes across a wide range of thresholds, as in Figure 3(b). Together, these results suggest a robust and reliable classifier for the task at hand.

Both plots demonstrate that the XGBoost model performs exceptionally well for binary classification tasks. The precision-recall curve, with an average precision of 0.95, shows that the model maintains high precision (few incorrect positive predictions) even while achieving high recall (detecting most actual positive cases). Complementarily, the ROC curve, with an AUC of 0.94, confirms the model's excellent capacity to distinguish between positive and negative classes over a broad range of thresholds, significantly outperforming a random classifier [36], [37].



(a) (b)

Figure 3. Classifier performance metrics for XGBoost model: (a) precision-recall curve and (b) ROC curve

4. MATHEMATICAL BASIS OF THE EXTREME GRADIENT BOOSTING MODEL

The following section presents the mathematical foundation of the XGBoost algorithm, adapted to the binary classification problem aimed at identifying health-related conditions associated with sleep. It describes the key concepts and equations that govern the model's learning process. The section also explains how the algorithm optimizes its objective function to improve predictive performance.

4.1. General objective function

XGBoost builds an ensemble of additive decision trees by optimizing an objective function that combines a loss function with a regularization term to prevent overfitting as (1) [38]. This approach allows the model to sequentially correct the errors of previous trees. It also balances model complexity and predictive accuracy to achieve robust performance on unseen data.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{1}$$

Where $l(y_i, \hat{y}_i^{(t)})$ is loss function between the prediction and the actual label; and $\Omega(f_k)$ is penalty on tree complexity f_k.

4.2. Model regularization

Regularization controls the complexity of the model by penalizing the depth and the weight of the leaves in each tree as in (2) [39]. This helps prevent overfitting by discouraging overly complex trees. It ensures that the model generalizes well to new, unseen data.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2}$$

Where T: number of leaves; w_j is predicted value on the sheet j; γ is penalty for each leaf; and λ is L2 regularization parameter.

4.3. Logistic loss function

For binary classification, the logistic loss (log-loss) is used [40], which measures the error between the predicted probability and the actual class, as in (3). This loss function penalizes incorrect predictions more heavily when the model is confident but wrong. It provides a smooth gradient that facilitates optimization during the training of XGBoost.

$$l(y_i, \hat{y}_i) = -\left[y_i \log(\sigma(\hat{y}_i)) + (1 - y_i)\log(1 - \sigma(\hat{y}_i))\right] \tag{3}$$

Where the sigmoid function is:

$$\sigma(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}}$$

4.4. Optimization by second order expansion

To improve training efficiency, XGBoost uses a second-order (Taylor) expansion [41] around the current prediction, as in (4). This approximation allows the model to consider both the gradient and the Hessian when updating tree leaves. It accelerates convergence and enhances the accuracy of each iteration during training.

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2\right] + \Omega(f_t) \tag{4}$$

Where $g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$ is gradient and $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)2}}$ is Hessian.

4.5. Optimal weight of a sheet

The optimal value that a leaf should predict is calculated as (5). This value is derived by minimizing the regularized objective function for that specific leaf. It ensures that each leaf contributes optimally to reducing the overall prediction error of the model.

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{5}$$

Where $G_j = \sum_{i \in I_j} g_i$ is sum of gradients; $H_j = \sum_{i \in I_j} h_i$ is sum of Hessians; and I_j is set of samples on the sheet j.

4.6. Profit from a division

To decide whether to perform a split at a node, the resulting information gain is calculated (6). This gain measures how much the split reduces the overall loss of the model. It helps the algorithm choose the most informative splits, improving predictive accuracy while controlling model complexity.

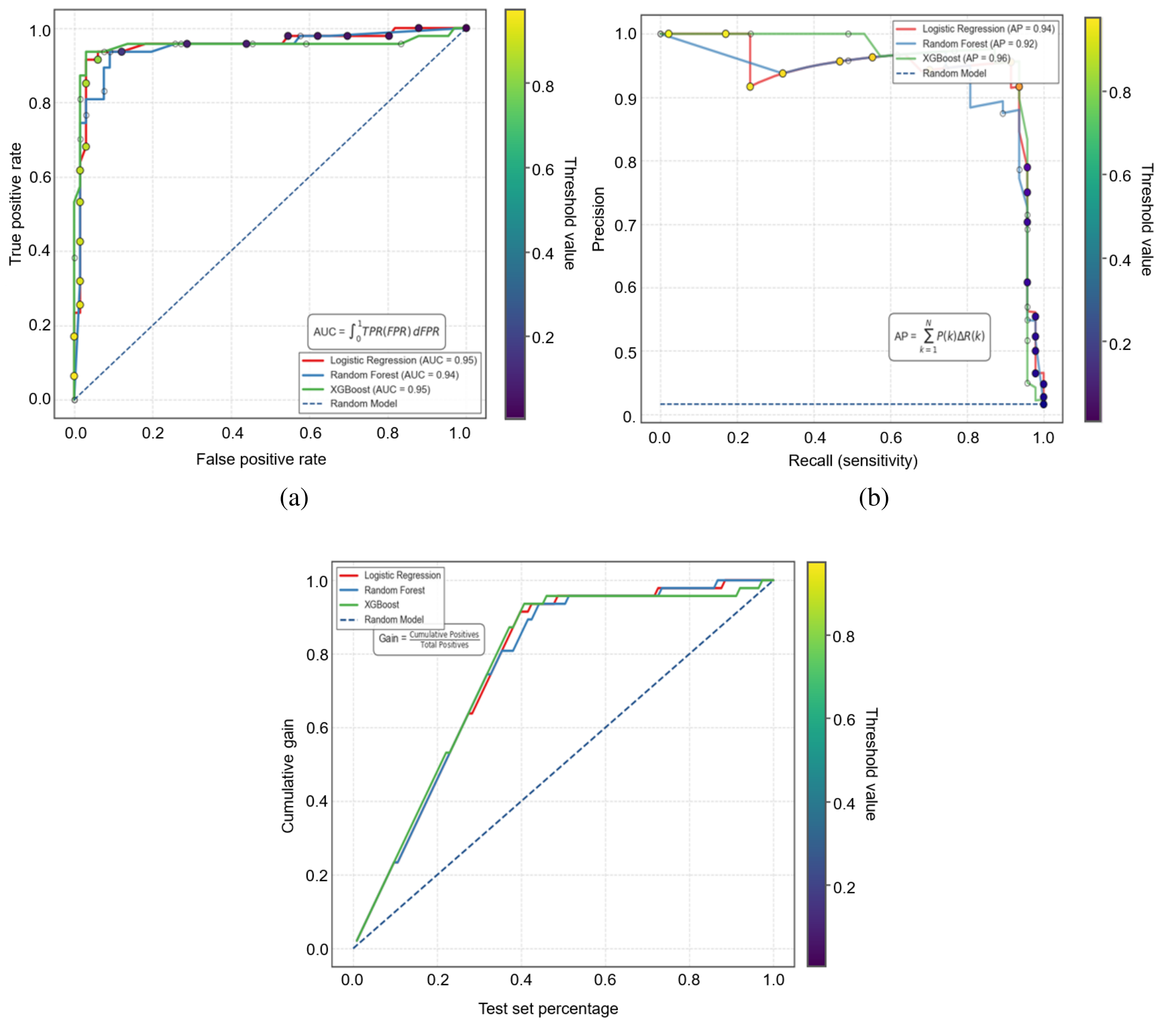
$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{6}$$

Where G_L, H_L is sum of gradients and Hessians to the left of the split; G_R, H_R is sum to the right of the split; and γ is penalty for creating an additional sheet.

5. RESULTS

5.1. Evaluation of result

Figure 4 compares three classification models (logistic regression, RF, and XGBoost), showing that all three exhibit excellent and comparable performance [42]. The ROC curve demonstrates that all models are equally capable of distinguishing between classes (AUC = 0.95), as shown in Figure 4(a). The precision-recall curve indicates that XGBoost is slightly superior in maintaining a balance between precision and recall (AP = 0.92), as shown in Figure 4(b), while the cumulative gain chart confirms that all models are highly efficient at quickly identifying the majority of positive cases, as shown in Figure 4(c).



(a) (b)

(c)

Figure 4. Model performance comparison: (a) logistic regression, (b) RF, and (c) XGBoost

Furthermore, other results shown in Figure 5 include normalized confusion matrices for the three ML models evaluated for the binary classification task, allowing for a detailed comparison between the predicted and actual class labels. Figure 5(a) shows the performance of the logistic regression model, which correctly classified 91% of the class 0 samples and 94% of the class 1 samples, with misclassification rates of 9% and 6%, respectively. On the other hand, Figure 5(b) illustrates the results obtained with the RF classifier, which achieved correct classification rates of 88% for class 0 and 94% for class 1, although it had a slightly higher false positive rate (12%) than the other models. Finally, Figure 5(c) presents the confusion matrix for the XGBoost model, which offered the best overall performance by correctly identifying 97% of class 0 instances and 94% of class 1 instances, while reducing the false positive rate to 3% and maintaining a low false negative rate (6%). Overall, the results indicate that all three classifiers achieved high predictive performance, with XGBoost demonstrating the most accurate and balanced classification.

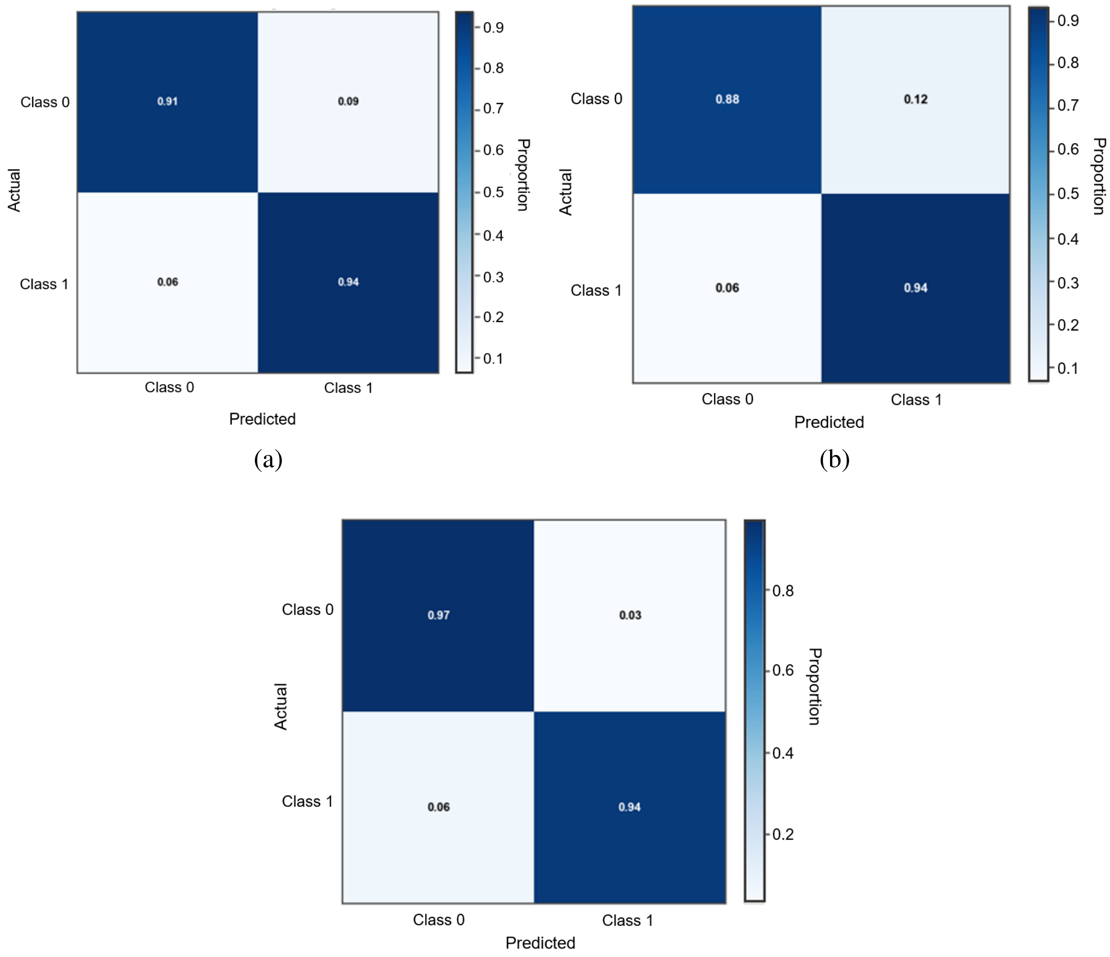


Figure 5. Confusion matrices of classification models: (a) logistic regression, (b) RF, and (c) XGBoost

Table 3 reports the overall error metrics. The model shows a low false positive rate (FPR) of 0.065 and a true positive rate (TPR) of 0.909, indicating high overall sensitivity. The true negative rate (TNR) is also high (0.935), reflecting good specificity. The overall error rate is relatively low (0.085), which supports the model's general reliability.

Table 3. Global error metrics

Metric	Value
FPR	0.065
FNR	0.182
TPR	0.909
TNR	0.935
Error rate	0.085

The class-wise metrics presented in Table 4 reflect a balanced performance of the model in classifying insomnia and sleep apnea. The insomnia class shows a precision of 0.862, a perfect recall of 1.000, and an F1-score of 0.926, indicating that all insomnia cases were correctly identified with no false negatives.

In contrast, the sleep apnea class exhibits perfect precision (1.000) and a recall of 0.818, with an F1-score of 0.900, meaning there were no false positives but some actual cases were missed. Both classes achieve an AUC of 0.909, suggesting a good discriminative capability of the model.

Table 4. Per-class evaluation metrics

Class	Precision	Recall	Specificity	F1-score	AUC
Insomnia	0.862	1.000	0.818	0.926	0.909
Sleep apnea	1.000	0.818	1.000	0.900	0.909

Table 5 presents the statistical analysis of the proposed model. The cross-validation results show an average accuracy of 0.91 with a standard deviation of ±0.03, which reflects consistent performance. Moreover, the p-value obtained when comparing the model with a baseline classifier is 0.0012, suggesting that the improvement is statistically significant. The 95% confidence interval, ranging from 0.89 to 0.94, supports the robustness and capacity of the model to generalize across multiple data partitions.

Table 5. Statistical evaluation summary

Statistic	Value
Cross-validation accuracy (mean ± std)	0.91±0.03
p-value (vs baseline)	0.0012
95% confidence interval (accuracy)	[0.89, 0.94]

Table 6 provides a summary of the classification model's performance in detecting sleep disorders. The model demonstrates outstanding performance in distinguishing between insomnia and sleep apnea, with a precision of 0.862 and a perfect recall (1.000) for insomnia, as well as a perfect precision (1.000) and a recall of 0.818 for sleep apnea. The overall accuracy reaches 91.5%, indicating high model reliability. The macro-averaged metrics precision, recall, and F1-score achieve values of 0.931, 0.909, and 0.913 respectively, suggesting balanced performance across classes. The results presented in Table 7 further reinforces these findings, showing strong and consistent metrics that reflect the model's effectiveness and generalization capability in classifying sleep related conditions.

Table 6. Model classification report

Class	Precision	Recall	F1-score	Support
Insomnia	0.862	1.000	0.926	25
Sleep apnea	1.000	0.818	0.900	22
Accuracy		0.915		
Macro average	0.931	0.909	0.913	47
Weighted average	0.927	0.915	0.914	47

Table 7. Summary of overall metrics

Metric	Value
Accuracy	0.915
Macro precision	0.931
Macro recall	0.909
Macro F1-score	0.913

5.2. Discussion

In this section, the results obtained from the literature review of other researchers who developed the same line of research are analyzed. The authors in [11], [12] emphasize that advanced ML algorithms play a crucial role in integrating clinical practice with reliable outcomes. Their work emphasized the relevance of properly training models and assessing them with metrics such as accuracy, precision, recall, and F1-score. In contrast, the proposal adopts a more systematic framework grounded in the KDD methodology, which helps produce results that are both more reliable and methodologically sound.

Previous studies have highlighted that the use of advanced ML algorithms has been key to effectively integrating predictive tools into clinical practice, producing reliable results. These studies emphasize the importance of a rigorous training and evaluation process using standard metrics such as accuracy, precision, recall, and F1-score. In comparison to those works, the proposal is based on a more robust structure supported by the KDD process, which encompasses everything from data selection and transformation to model evaluation and statistical validation. This approach resulted in precision metrics of 0.915 and an AUC of 0.909, as well as consistent cross-validation (0.91±0.03).

The authors in [13], [14] apply the same strategy as the aforementioned authors, including training and other relevant parameters. Their model achieved a precision of 90.6%, whereas our model surpasses this performance with a 91% accuracy in sleep quality prediction particularly in parameters such as insomnia or sleep apnea, as noted by [28], [29]. It is also important to note that one of the most significant sleep-related issues is OSA. The authors in [15], [16] developed a model evaluation in which they reported relatively low performance metrics, such as a precision of 68.83% and a recall of 67.84%. However, their study objectives differ from those of our model, which achieved a higher level of performance. In another instance, the authors in [17], [18] used a different methodology to select the most important variables, employing the SHAP method, which resulted in an AUC value above 89% for the described disorders. This aspect is reflected in the transformation phase of our study model, where key features such as insomnia and sleep apnea are emphasized. Although these investigations follow different approaches, they use the same model XGBoost for each specific objective.

In a different context, the authors in [19], [20] developed predictive models using Python and the algorithms logistic regression, XGBClassifier, and SVC. After evaluating their performance, the best-performing model was XGBClassifier, which highlighted insomnia as the primary factor associated with BMI. This suggests that, similar to the previous study, the model performs algorithmic comparisons to determine the most suitable approach. However, the key improvement lies in fine-tuning values during the results process to obtain the expected outcome. It is important to consider that sleep disorders are often associated with other conditions such as breast cancer. The authors in [21], [22] demonstrate the performance of their model with metrics such as an ROC of 0.87, and both sensitivity and specificity equal to 0.77, highlighting a robust ability to identify risk factors related to sleep disorders. While the model in this study exhibits superior performance in these metrics, the comparison suggests the advanced capacity of the XGBoost algorithm in detecting sleep-related health conditions.

6. CONCLUSION

One of the major issues affecting various populations is sleep quality. When poor, it can lead to serious long-term consequences such as stress, fatigue, concentration difficulties, and memory loss, among other factors. These disturbances can develop into severe complications if awareness is not raised about the importance of maintaining healthy sleep habits and adopting appropriate preventive measures. In this regard, a CSV-format dataset extracted from the Kaggle platform was consolidated, comprising relevant clinical variables used as analytical input for the development of the predictive model. The five-stage KDD approach was used. During the selection process, a number of sources were investigated and assessed until an appropriate dataset that matched the goals of the study was discovered. The data were then statistically examined in the preparation stage using mean, mode, and standard deviation computations, and data cleaning was carried out to deal with discrepancies or missing values. In the transformation phase, significant variables such as sleep apnea and insomnia were selected and prepared for analysis. During the data mining phase, the XGBoost ML algorithm was implemented and its performance was evaluated using metrics such as ROC curve, precision, and recall. Finally, in the interpretation and evaluation phase, the results were analyzed and compared with other models, which validated the effectiveness of the proposed approach. Moreover, the predictive model developed using the XGBoost algorithm demonstrated high performance in identifying sleep disorders, achieving strong precision scores. Based on the results obtained, there is a need to explore new lines of research to complement and expand the current findings. For future work, predictive models can be integrated into mobile or web applications through exposure via a representational state transfer (REST) application programming interface (API), enabling the model's logic to be implemented using a backend language (such as Python or Node.js) and visualized using frontend technologies (such as Angular, React, or Vue.js). It is also possible to connect the model to data storage systems, whether relational (structured query language (SQL)) or non-relational (not only structured query language (NoSQL)) databases, to record and manage the results generated based on user input. In scenarios requiring scalability, the model can be extended within big data architectures. Additionally, it can be deployed on cloud platforms such as Amazon web services (AWS), Google Cloud, or Azure, which support training, versioning, and scaling of ML models, allowing for a robust and sustainable implementation in clinical or commercial environments.

FUNDING INFORMATION

This research was financially supported by Universidad de Ciencias y Humanidades. The institution provided funding for the development and completion of this scientific study. No specific grant number was assigned to this financial support.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Laberiano Andrade-Arenas	✓		✓		✓		✓			✓	✓	✓	✓	✓
Inoc Rubio Paucar		✓		✓	✓		✓		✓		✓	✓		
Margarita Giraldo Retuerto	✓		✓	✓						✓		✓	✓	✓
Cesar Yactayo-Arias	✓	✓	✓	✓			✓		✓		✓	✓		

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject Administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding Acquisition
Fo	: **Fo**rmal Analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data supporting the findings of this study are publicly available on the Kaggle platform at https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset. The dataset is provided in CSV format and contains variables related to sleep habits, sleep quality, sleep disorders, and mental health. The data can be accessed and reused in accordance with the terms established by the Kaggle repository.

REFERENCES

[1] P. L. S. Laborde, C. T. Pizzuti, and M. S. C. Pesce, "Sleep quality and chronotype and their relationship with obesity in the adult population. narrative literature review," *Nursing: Humanized Care*, vol. 12, no. 2, 2023, doi:10.22235/ech.v12i2.3213.

[2] J. M. M. Aguilar and M. O. I. D. Castro, "Self-medication and sleep hygiene in patients with systemic lupus erythematosus: a salutogenic perspective," *Revista De Investigación*, vol. 2, no. 14, pp. 53–65, 2023, doi:10.5377/revunivo.v2I14.17024.

[3] L. A. L. Muciño, I. A. G. Torres, M. E. A. Hernández, and F. G. García, "¿Cómo influye la falta de sueño sobre la conducta adictiva?: notes for the analysis of bias in research in health science" *Revista Médica de la Universidad Veracruzana* , vol. 21, no. 1, pp. 58–66, 2022, doi:10.25009/RMUV.2021.1.17.

[4] J. F. D. Osorio, "Descripción de la estructura interna del mesencéfalo de aotus (primates: simiiformes: aotidae) con referencia a homo (primates: simiiformes: hominidae): posibles implicaciones en la evolución de los ciclos circadianos de mamíferos," *Revista De La Academia Colombiana De Ciencias Exactas, Físicas Y Naturales,*, vol. 33, no. 128, pp. 425–444, 2023, doi:10.18257/raccefyn.33(128).2009.2374.

[5] K. Spiegel, E. Tasali, P. Penev, and E. V. Cauter, "Brief communication: sleep curtailment in healthy young men is associated with decreased leptin levels, elevated ghrelin levels, and increased hunger and appetite," *Annals of Internal Medicine*, vol. 141, no. 11, pp. 846–850, 2004, doi:10.7326/0003-4819-141-11-200412070-00008.

[6] J. R.-Munell, J. P.-Cruz, R. M. Al.-Rosa, and F. Á. Barbosa, "Effects of a physical activity intervention in the school setting on sleep quality and quantity in preschoolers: a systematic review," *Retos*, vol. 51, no. 51, pp. 235–242, 2023, doi:10.47197/RETOS.V51.99792.

[7] D. C. Vanegas, G. A. Amórtegui, and V. B. Ortiz, "Calidad del sueño en el personal administrativo de la Fundación Universitaria del Área Andina," *Documentos de trabajo Areandina*, no. 2, pp. 42–61, 2021, doi:10.33132/26654644.1974.

[8] G. Airlangga, "Evaluating machine learning models for predicting sleep disorders in a lifestyle and health data context," *JIKO (Jurnal Informatika dan Komputer)*, vol. 7, no. 1, pp. 51–57, 2024, doi:10.33387/jiko.v7i1.7870.

[9] D. E.-Calvo, "Filosofía de la Salud Pública," *Memoria Del Posgrado*, vol. 4, no. 1, pp. 62–68, 2023, doi:10.53287/SDSW5980SK82P.

[10] M. E. F. Carrillo and C. X. D. Garifias, "Public health: essential functions of public health," *TEPEXI Boletín Científico de la Escuela Superior Tepeji del Río*, vol. 7, no. 4, pp. 91–93, 2020, doi:10.29057/ESTR.V7I14.5661.

[11] W. Su *et al.*, "An XGBoost-based knowledge tracing model," *International Journal of Computational Intelligence Systems*, vol. 16, no. 13, pp. 1 – 9, 2023, doi:10.1007/S44196-023-00192-Y.

[12] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *International Journal of Distributed Sensor Networks*, vol. 18, no. 19, pp. 9 – 14, 2022, doi:10.1177/15501329221106935.

[13] S. B. Jabeur, S. M.-Wali, and J. L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," *Annals of Operations Research*, vol. 334, pp. 679–699, 2024, doi: 10.1007/S10479-021-04187-W.

[14] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," *Pattern Recognition Letters*, vol. 136, no. 42, pp. 190–197, 2020, doi:10.1016/j.patrec.2020.05.035.

[15] J. Wang and S. Zhou, "Particle swarm optimization-XGBoost-based modeling of radio-frequency power amplifier under different temperatures," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 37, no. 37, pp. 22–35, 2024, doi:10.1002/jnm.3168.

[16] J. Zhang, R. Wang, Y. Lu, and J. Huang, "Prediction of compressive strength of geopolymer concrete landscape design: application of the novel hybrid RF–GWO–XGBoost algorithm," *Buildings*, vol. 14, no. 3, 2024, doi:10.3390/buildings14030591.

[17] J. S. Vetter, K. Schultebraucks, I. G.-Levy, H. Boeker, A. Brühl, E. Seifritz, and B. Kleim, "Predicting non-response to multimodal day clinic treatment in severely impaired depressed patients: a machine learning approach," *Scientific Reports*, vol. 12, 2022, doi: 10.1038/s41598-022-09226-5.

[18] J. Luo *et al.*, "Major depressive disorder prediction based on sleep-wake disorders symptoms in us adolescents: a machine learning approach from national sleep research resource," *Psychology Research and Behavior Management*, vol. 17, pp. 691–703, 2024, doi: 10.2147/PRBM.S453046.

[19] H. Gong, "Sleep quality analysis based on logistical, SVC and XGBoost," *Highlights in Science, Engineering and Technology*, vol. 76, pp. 399–404, 2023, doi: 10.54097/7bfdy131.

[20] E. Holler *et al.*, "Development and temporal validation of an electronic medical record-based insomnia prediction model using data from a statewide health information exchange," *Journal of Clinical Medicine*, vol. 12, 2023, doi: 10.3390/jcm12093286.

[21] T. Ueno *et al.*, "Comorbid insomnia among breast cancer survivors and its prediction using machine learning: a nationwide study in Japan," *Japanese Journal of Clinical Oncology*, vol. 52, pp. 39–46, 2022, doi: 10.1093/jjco/hyab169.

[22] A. A. Huang and S. Y. Huang, "Use of machine learning to identify risk factors for insomnia," *PLoS ONE*, vol. 18, no. 4, 2023, doi: 10.1371/JOURNAL.PONE.0282622.

[23] U. M. Haque, E. Kabir, and R. Khanam, "Detection of child depression using machine learning methods," *PLoS ONE*, vol. 16, no. 12, 2021, doi: 10.1371/JOURNAL.PONE.0261131.

[24] N. Higashiyama *et al.*, "Heart rate variability and the quality of life of patients receiving chemotherapy against gynecologic malignancies." *JCO Global Oncology*, vol. 9, no. 1, pp. 106–106, 2023, doi: 10.1200/GO.2023.9.supplement 1.106.

[25] C. Rao, Y. Liu, and M. Goh, "Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost model," *Complex and Intelligent Systems*, vol. 9, pp. 1391–1414, 2023, doi:10.1007/S40747-022-00854-Y.

[26] J. Zhou, Y. Qiu, S. Zhu, D. J. Armaghani, M. Khandelwal, and E. T. Mohamad, "Estimation of the TBM advance rate under hard rock conditions using XGBoost and Bayesian optimization," *Underground Space*, vol. 6, no. 5, pp. 506–515, 2021, doi: 10.1016/J.UNDSP.2020.05.008.

[27] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, "An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach," *MethodsX*, vol. 10, 2023, doi:10.1016/J.MEX.2023.102119.

[28] G. Qu *et al.*, "Association between adverse childhood experiences and sleep quality, emotional and behavioral problems and academic achievement of children and adolescents," *European Child and Adolescent Psychiatry*, vol. 33, no. 2, pp. 527–538, 2024, doi: 10.1007/S00787-023-02185-W.

[29] Z. Khalili, E. Habibi, N. Kamyari, S. Tohidi, and V. Yousofvand, "Correlation between spiritual health, anxiety, and sleep quality among cancer patients," *International Journal of Africa Nursing Sciences*, vol. 20, 2024, doi: 10.1016/J.IJANS.2024.100668.

[30] S. Seo and M. K. Mattos, "The relationship between social support and sleep quality in older adults: a review of the evidence," *Archives of Gerontology and Geriatrics*, vol. 117, 2024, doi: 10.1016/J.ARCHGER.2023.105179.

[31] G. B. Matos *et al.*, "Corante natural azul da flor comestível clitoria (Clitoria ternatea): extração, caracterização e estabilidade," *Observatory of the Latin American Economy*, vol. 22, no. 2, 2024, doi: 10.55905/OELV22N2-202.

[32] S. S. Jesus and C. G. A. Asensio, "EEfecto de la distancia de la planta al foco emisor de luz durante la fase de multiplicación del cultivo in vitro de Limonium sinuatum," III Workshop en Investigación Agroalimentaria: WiA3.14, vol. 3, pp. 95–98, 2024, doi: 10.31428/10317/10830.

[33] G. A. Romero, C. A. G. Prieto, M. A. D. Barriosnuevos, and N. A. R. Menjura, "Review and perspectives for the construction of robust databases with missing data: case applied to financial information," *Tecnura*, vol. 27, pp. 12–37, 2023, doi: 10.14483/22487638.18268.

[34] M. Rochmawati, G. W. C. Bagaskara, I. A. Adha, Y. Umaidah, and A. Voutama, "Implementation of the k-means algorithm in sales clustering at a company using the KDD methodology," *SISTEMASI:Jurnal Sistem Informasi*, vol. 13, no. 1, 2024, doi: 10.32520/STMSI.V13I1.3074.

[35] M. H. Behiry and M. Aly, "Cyberattack detection in wireless sensor networks using a hybrid feature reduction technique with AI and machine learning methods," *Journal of Big Data*, vol. 11, no. 16, 2024, doi: 10.1186/S40537-023-00870-W.

[36] X. Li *et al.*, "Probabilistic solar irradiance forecasting based on XGBoost," *Energy Reports*, vol. 8, pp. 1087–1095, 2022, doi: 10.1016/J.EGYR.2022.02.251.

[37] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network intrusion detection based on PSO-XGBoost model," *IEEE Access*, vol. 8, pp. 58 392–58 401, 2020, doi: 10.1109/ACCESS.2020.2982418.

[38] Y. Qiu, J. Zhou, M. Khandelwal, H. Yang, P. Yang, and C. Li, "Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration," *Engineering with Computers*, vol. 38, no. 5, pp. 4145–4162, 2022, doi: 10.1007/S00366-021-01393-9.

[39] D. Zhang and Y. Gong, "The comparison of lightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure," *IEEE Access*, vol. 8, pp. 220 990–221 003, 2020, doi: 10.1109/ACCESS.2020.3042848.

[40] C. Dong *et al.*, "Non-contact screening system based for COVID-19 on XGBoost and logistic regression," *Computers in Biology and Medicine*, vol. 141, 2022, doi: 10.1016/J.COMPBIOMED.2021.105003.

[41] S. Sankar, A. Potti, G. N. Chandrika, and S. Ramasubbareddy, "Thyroid disease prediction using XGBoost algorithms," *Journal of Mobile Multimedia*, vol. 18, pp. 917–934, 2022, doi: 10.13052/JMM1550-4646.18322.

[42] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, pp. 4514–4523, 2022, doi: 10.1016/J.JKSUCI.2020.10.013.

BIOGRAPHIES OF AUTHORS

Laberiano Andrade-Arenas is doctor in Systems and Computer Engineering and a master in Systems Engineering. He graduated with a master's degree in University Teaching and graduated with a master's degree in Accreditation and Evaluation of Educational Quality. He is currently the dean of the School of Science and Engineering (Lima) at the University of Sciences and Humanities in Lima, Peru. He is a systems engineer certified in scrum fundamentals, a research professor with publications in Scopus-indexed journals. He can be contacted at email: landrade@uch.edu.pe.

Inoc Rubio Paucar is bachelor in Systems and Computer Engineering at Universidad Privada Norbert Wiener, Lima, Peru. He has a background in database management and computer system design, with a focus on artificial intelligence applications, ML, and data science. His research interests are in the area of computer science. He can be contacted at email: Enoc.Rubio06@hotmail.com.

Margarita Giraldo Retuerto is a systems engineer at the Universidad de Ciencias y Humanidades, Lima, Peru with a solid background in software engineering, specialized in the analysis, design, and development of technological solutions. She has experience applying software development methodologies and quality best practices. She shows a strong interest in scientific research, particularly in areas related to emerging technologies and process improvement. She has contributed to academic and technical projects with an analytical and methodical approach. Her professional profile combines technical expertise, critical thinking, and a strong research orientation. She can be contacted at email: mgiraldo@uch.edu.pe.

Cesar Yactayo-Arias bachelor's degree in administration from Universidad Inca Garcilazo de la Vega and a master's degree in education from Universidad Nacional de Educación Enrique Guzmán y Valle, he is a doctoral candidate in administration at Universidad Nacional Federico Villarreal. Since 2016 he has been teaching administration and mathematics subjects at the Universidad de Ciencias y Humanidades and since 2021 at the Universidad Continental. Currently, he also works as an administrator of educational services at the higher level, he is the author and co-author of several refereed articles in journals, and his research focuses on TIC applications to education, as well as management using computer science and the internet. He can be contacted at email: yactayocesar@gmail.com.