Journal of Health Informatics

J Health Inform. 2026;18:e1893
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1893

ARTIGO COMPLETO
CBIS 2026

Uncovering determinants of anxiety among university students in Brazil: an explainable machine learning approach

Determinantes da ansiedade em estudantes universitários no Brasil: uma abordagem de aprendizado de máquina explicável

João Victor Chamarelli[1] ✉, Juliana Dias Lima[2], Maria Luiza Cantanhede Wuillaume[2], Felipe Schuch[3], Andrea Deslandes[2], Manoel Saísse[1], Andréa Nunes Carvalho[1]

ORCID: 0009-0002-2965-7594 · 0000-0002-8663-2767 · 0009-0001-2061-5243 · 0000-0002-5190-4515 · 0000-0001-5941-9111 · 0009-0000-7352-927X · 0000-0002-8294-4552

[1] Instituto Nacional de Tecnologia, Rio de Janeiro (RJ), Brasil.
[2] Universidade Federal do Rio de Janeiro, Rio de Janeiro (RJ), Brasil.
[3] Universidade Federal de Santa Maria, Santa Maria (RS), Brasil.

ARTICLE INFORMATION

TYPE
Artigo completo · CBIS 2026

IDENTIFIER
e1893

DOI
10.59681/2175-4411.v18.iEspecial.2026.1893

RECEIVED
17 May 2026

ACCEPTED
20 July 2026

REVIEW PROCESS
Double Blind Review

HANDLING EDITOR
Juliano de Souza Gaspar

ACCESS
Open, CC BY 4.0

CORRESPONDING AUTHOR
João Victor Chamarelli
jvchamarelli@gmail.com

LICENSE

This is an open access article published under the Creative Commons CC BY 4.0 license, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.

CC BY 4.0

ABSTRACT

Objective: Identify key modifiable and non-modifiable risk factors associated with anxiety symptoms among Brazilian university students using interpretable machine learning. **Method:** Data from 6,398 undergraduate students across 13 universities in all five Brazilian regions were drawn from the UNILIFE-M project. A binary classification task was defined using the GAD-7 clinical cut-off, and five models were trained and compared. **Results:** No statistically significant performance differences were found among classifiers, and LightGBM was selected on practical grounds. SHAP analysis revealed sleep quality, sex, social support, sexual orientation, and stress management as the most influential predictors. **Conclusion:** These findings highlight actionable targets for university mental health policies and demonstrate the value of combining predictive modeling with explainability techniques in population health research.

Keywords: Machine Learning; Data Science; Anxiety Disorders.

RESUMO

Objetivo: Identificar os principais fatores de risco modificáveis e não modificáveis associados a sintomas de ansiedade em estudantes universitários brasileiros utilizando aprendizado de máquina interpretável. **Método:** Dados de 6.398 estudantes de 13 universidades de todas as regiões brasileiras foram extraídos do projeto UNILIFE-M. Uma classificação binária foi definida pelo ponto de corte clínico do GAD-7, e cinco modelos foram treinados e comparados. **Resultados:** Sem diferenças significativas de desempenho entre classificadores, o LightGBM foi selecionado por razões práticas. A análise SHAP revelou que qualidade do sono, sexo, suporte social, orientação sexual e manejo do estresse são os fatores mais influentes. **Conclusão:** Os achados destacam alvos para políticas de saúde mental universitária e demonstram o valor de combinar modelagem preditiva com técnicas de explicabilidade em saúde populacional.

Descritores: Aprendizado de Máquina; Ciência de Dados; Transtornos de Ansiedade.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Paper presented at the 21st Brazilian Congress of Health Informatics (CBIS 2026), held from September 23 to 25, 2026, in Brasília (DF), Brazil, and published in the CBIS 2026 section of the Journal of Health Informatics.

HOW TO CITE THIS ARTICLE: Chamarelli JV, Lima JD, Wuillaume MLC, Schuch F, Deslandes A, Saísse M, et al. Uncovering determinants of anxiety among university students in Brazil: an explainable machine learning approach. J Health Inform 2026;18:e1893. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1893

1 INTRODUCTION

Mental health disorders represent a major and growing public health concern worldwide, with anxiety disorders among the most prevalent conditions [1], particularly affecting young populations [2]. Notably, many of these disorders emerge during adolescence and early adulthood [3], a critical developmental period that often coincides with entry into higher education, placing university students in a particularly vulnerable position. Anxiety symptoms have been associated with adverse academic outcomes, including lower performance, higher stress levels, and increased risk of dropout [4].

Such symptoms are influenced by a complex interplay of social, behavioral, and lifestyle determinants. Among university students, factors such as sleep patterns [5], physical activity and dietary habits [6], and social support are particularly relevant and modifiable [7]. Sociodemographic characteristics such as age, sex, ethnicity, family socioeconomic status, family structure, and urban upbringing have also been consistently associated with anxiety, including in Brazil [8]. In addition, mental health conditions are not equally distributed across social groups, with higher risks observed among more socially disadvantaged students, highlighting the role of intersecting inequalities [9].

Capturing these complex relationships relies on robust indicators, high-quality data, and the identification of the most relevant determinants, enabling effective monitoring and a deeper understanding of the conditions in which adolescents and young adults develop. However, fully leveraging such data remains challenging, as traditional statistical approaches are often limited in capturing non-linear interactions and complex dependencies inherent to these phenomena. Machine learning (ML) methods have emerged as powerful tools for predictive modeling in mental health research among university students [10]. Nevertheless, their limited interpretability remains a key constraint, particularly in applied and population health contexts. To address this limitation, explainable artificial intelligence (XAI) techniques, such as SHapley Additive exPlanations (SHAP) [11], have been proposed to provide consistent and theoretically grounded measures of feature contributions.

Building on this, despite recent advances, there remains a limited understanding of how sociodemographic and lifestyle factors jointly shape anxiety outcomes among university students, particularly within frameworks that effectively integrate predictive performance and interpretability. This gap reflects the persistent challenge of translating accurate predictions into meaningful insights about the mechanisms underlying mental health outcomes. Accordingly, this study develops and evaluates ML models to predict anxiety symptoms among Brazilian university students and applies SHAP values to interpret the contribution of key variables, aiming to generate actionable evidence on mental health determinants in this population. The main contribution of this study lies in advancing knowledge in this field by providing insights that may support the development of more effective public policies.

2 METHODS

2.1 Study design and dataset

The dataset used in this study was obtained from the UNIversity students' LIFEstyle behaviours and Mental health cohort (UNILIFE-M), a multicenter cohort study investigating the relationship between lifestyle behaviors and mental health among university students across multiple countries [12]. Data were collected through online surveys conducted in 2023 and 2024. For this study, only data from the Brazilian sample were used, comprising 6,398 undergraduate students from 13 public universities across all five regions of Brazil.

The UNILIFE-M survey includes over 200 items covering sociodemographic, socioeconomic, lifestyle, and mental health domains. A central component of the survey is the Short Multidimensional Inventory Lifestyle Evaluation (U-SMILE), which assesses lifestyle behaviors across seven domains, including diet and nutrition, substance use, physical activity, stress management, restorative sleep, social support, and environmental exposures [13].

The outcome variable of the predictor model is "gad7_class", a binary variable derived from the Generalized Anxiety Disorder 7-item scale (GAD-7) [14]. The GAD-7 is a widely validated self-report instrument used to screen and assess the severity of generalized anxiety disorder, where participants rate the frequency of seven anxiety-related symptoms over the past two weeks. To construct gad7_class, the total GAD-7 score was dichotomized according to its established clinical cut-off point (>9), yielding two categories: 0 ("non-anxious"), assigned to participants whose scores fall below the threshold, and 1 ("anxious"), assigned to those who meet or exceed it. This binary classification allows the model to frame anxiety detection as a supervised classification problem, with the goal of predicting whether a given student is likely to screen positive for clinically meaningful anxiety symptoms.

The predictor set comprises 36 variables spanning sociodemographic, socioeconomic, lifestyle, and mental health domains. Sociodemographic variables include categorical features such as sex, gender identity, sexual orientation, ethnic group, marital status, and student housing, as well as numerical variables such as age and body mass index (BMI). Socioeconomic status is represented by work status

and income. Lifestyle variables are primarily derived from the U-SMILE scale, with additional variables capturing sedentary behavior.

2.2 Data preprocessing

All preprocessing steps were embedded within a structured pipeline applied under a repeated random-split cross-validation scheme (ShuffleSplit; n=5 splits, test fraction = 0.20), incorporating both outer and inner cross-validation to reduce overfitting and mitigate data leakage. Features were partitioned into categorical (n=7, sociodemographic variables) and discrete (n=29: age, income, BMI, sedentary behavior measures, and 24 U-SMILE scale items) sets, each subjected to distinct imputation strategies.

Missing values in categorical features were imputed using the most frequent category, aiming to preserve marginal distributions, while numerical features were imputed using k-nearest neighbor imputation with k=1, as the feature with the most missing values was numerical. Following imputation, categorical features were one-hot encoded and discrete features scaled to the [0,1] range using MinMaxScaler. All preprocessing steps were fitted exclusively on the training data and applied to the test data without re-fitting, mitigating the risk of data leakage.

To support SHAP-based interpretability, an additional version of the numerical features, imputed but not scaled, was retained to preserve the original measurement scale. Finally, chi-squared and Kolmogorov–Smirnov tests were used to assess whether imputed distributions differed significantly from the original data.

2.3 Machine learning models

Five classification models were evaluated: Random Forest, Support Vector Machine (SVM), Logistic Regression, XGBoost, and LightGBM. This selection spans models with different modeling assumptions, increasing the likelihood of identifying the best-performing approach. Model performance was assessed using precision, recall, and F1-score. Given the binary classification task and potential class imbalance, macro-averaged F1-score was used as the primary optimization criterion.

Hyperparameter optimization was performed for all five classifiers using the Tree-structured Parzen Estimator (TPE) algorithm, implemented via the Hyperopt library. Each model was assigned a classifier-specific search space over its key hyperparameters. Model configurations were evaluated using cross-validation, with Synthetic Minority Over-sampling Technique (SMOTE) applied to the training folds within each iteration to address class imbalance while mitigating data leakage. The mean macro-F1 score across folds was used as the optimization objective, with a budget of 50 evaluations per model. The best-performing configuration for each classifier was then selected for further comparison.

The optimized models were compared using the non-parametric framework proposed by Demšar [15]. Performance was evaluated based on per-fold macro-F1 scores across cross-validation splits. A Friedman test was conducted to assess overall differences among classifiers at a significance level of $\alpha = 0.05$. When significant differences were detected, the Nemenyi post-hoc test was applied to identify pairwise differences between models, using the critical difference threshold derived from the studentized range statistic for the corresponding number of classifiers and folds.

2.4 Model interpretability (SHAP)

Understanding feature importance was central to this study; thus, interpretability tools were required to translate model outputs into actionable insights. High predictive performance alone is insufficient if model behavior remains opaque. To address this, SHAP values were employed [11]. As a model-agnostic framework grounded in cooperative game theory, SHAP attributes predictions to individual features through Shapley values, enabling the validation of learned relationships, the detection of potential biases, and the communication of results in a transparent and consistent manner.

2.5 Ethical considerations

The UNILIFE-M study complies with the Declaration of Helsinki. Participants provided electronic informed consent, and data were pseudonymized in accordance with applicable regulations, including the Brazilian Lei Geral de Proteção de Dados. Participants could withdraw at any time and were referred to specialized services when needed. The study was approved by the Research Ethics Committee of the Federal University of Santa Maria (CEP-UFSM; CAAE: 63025822.8.1001.5346).

3 RESULTS

3.1 Performance and selection of machine learning models

The results of the training process for each of the five models are presented in Table 1, which reports all three performance metrics before and after hyperparameter optimization (shown with “opt.”).

Table 1. Performance metrics for each model, before vs after optimization

Metrics	Logistic Regression	SVM	XGBoost	LightGBM	Random Forest
Precision	0.671	0.666	0.641	0.656	0.665
F1-score	0.670	0.666	0.642	0.651	0.661
Recall	0.670	0.665	0.641	0.648	0.662
Precision (opt.)	0.697	0.689	0.695	0.695	0.687
F1-score (opt.)	0.693	0.694	0.693	0.693	0.684
Recall (opt.)	0.693	0.694	0.693	0.693	0.684

Source: Prepared by the authors (2026).

The Friedman test revealed no statistically significant differences among the five classifiers ($\chi^2(4) = 1.09$, $p = 0.896$), indicating that the null hypothesis of equal performance across models cannot be rejected at the $\alpha = 0.05$ level. These results suggest that, for the present dataset, classifier choice has no statistically meaningful impact on macro-F1 score performance, and all five models may be considered equivalent under this metric.

Given this equivalence, the selection of the final model was guided by practical considerations. LightGBM was chosen based on three main factors. First, its gradient-boosted tree architecture imposes no linearity assumptions on the decision boundary, providing a higher expressive capacity ceiling than Logistic Regression under potential data complexity or dataset growth. Second, it offers computational advantages over Random Forest and SVM in terms of training speed and memory efficiency, which are relevant under periodic retraining requirements. Lastly, it provides native support for feature importance estimation and SHAP-based explainability, facilitating model auditing and interpretability requirements common in applied and regulated contexts.

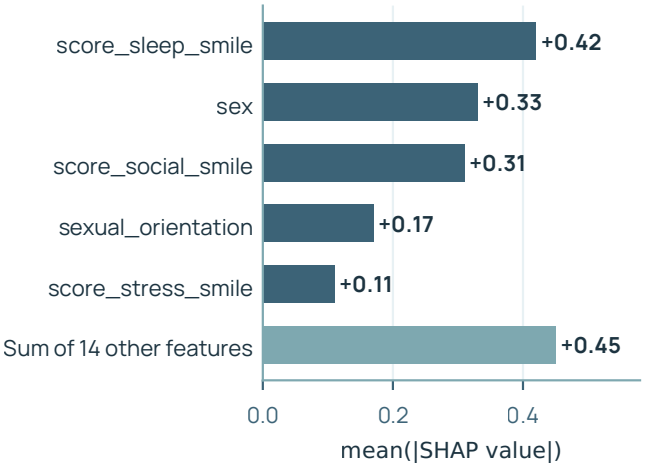
3.2 Feature importance

To assess the contribution of individual predictors to the LightGBM model's output, SHAP values were computed for all observations in the test set. Figure 1 presents the mean absolute SHAP values for the five most influential features, alongside the aggregated contribution of the remaining 14 predictors. Restorative sleep emerged as the single most impactful individual predictor, followed by sex and social support. Sexual orientation and stress management contributed more modestly.

3.3 Effects of key predictors

To complement the feature importance analysis and characterize the directional and magnitude effects of each predictor on model output, SHAP beeswarm and decision plots were examined. Together, these visualizations move beyond aggregate importance rankings to reveal *how* and *for whom* each feature drives predictions. The beeswarm plot (Figure 2) displays the distribution of SHAP values across all observations for the five most impactful features, where each point represents one observation, its horizontal position encodes the direction and magnitude of its contribution to the model output, and its color encodes the raw feature value (red = high; blue = low).

Figure 1. Bar plot for SHAP values



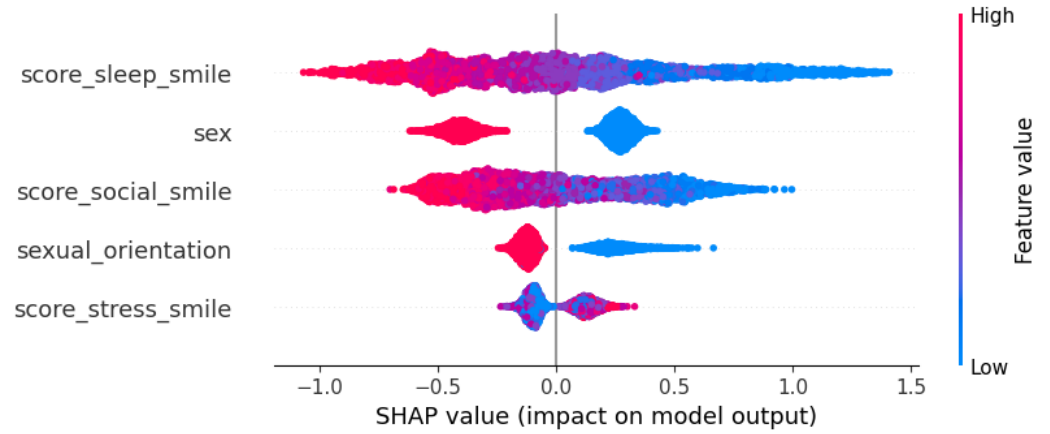
Source: Prepared by the authors (2026).

The most prominent feature, *restorative sleep*, exhibits a strong and consistent directional effect: high scores (red) are associated with large negative SHAP values, pushing predictions toward the negative class (non-anxious), while low scores (blue) generate large positive SHAP values, increasing the likelihood of a positive GAD-7 classification (anxious). This pattern, combined with the widest horizontal spread among all features, indicates that sleep is both the strongest and most variable individual predictor in the model. A qualitatively similar directional pattern is observed for *social support*, where high social functioning scores lead to more negative-class predictions, and low scores correspond with positive-class predictions. The *sex* feature presents a notably asymmetric distribution: low values (0; female sex) concentrate strongly on the positive SHAP side, indicating it is systematically associated with a higher predicted probability of GAD-7 positivity, while high values (1; male sex) cluster near zero with modest negative contributions. *Sexual orientation* displays a pattern with a slight negative skew for

high values (1; heterosexual) versus the positive orientation of low values (0; LGBTQ+ individuals), indicating a present but modest and categorical effect.

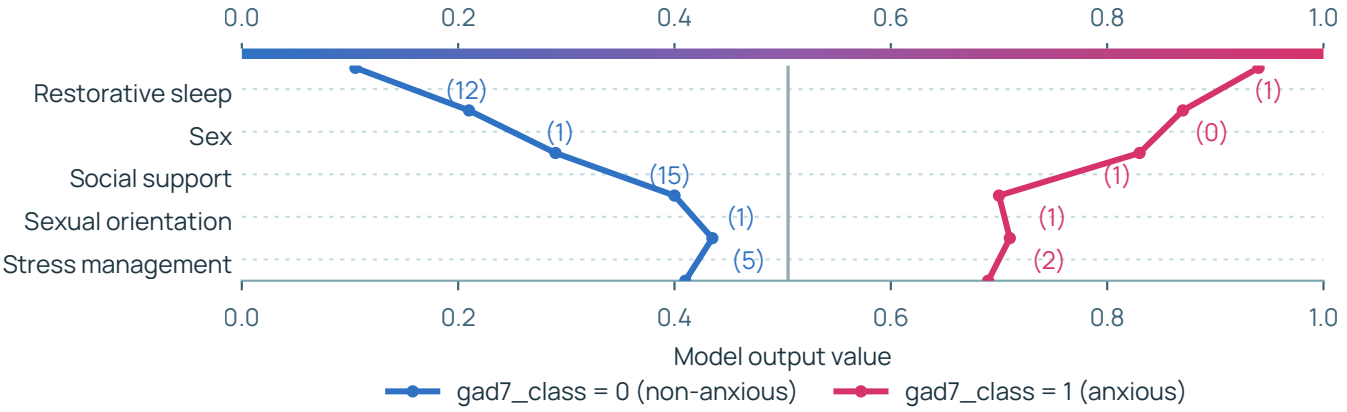
Finally, *stress management* contributes the smallest and most tightly clustered effects among the shown features, centered near zero with limited spread,

Figure 2. Beeswarm plot for SHAP values



Source: Prepared by the authors (2026).

Figure 3. Decision plot for gad7_class = 0 and gad7_class = 1



Source: Prepared by the authors (2026).

consistent with its lower global importance ranking.

The decision plot (Figure 3) for one non-anxious and one anxious observation (gad7_class=0 and gad7_class=1) provides an instance-level perspective by tracing how each feature cumulatively shifts the model output from its base value to the final prediction for each representative observation. Reading the plot from bottom to top, each feature either pushes the output leftward (toward class 0) or rightward (toward class 1), with the final position of the line relative to the 0.5 threshold determining the predicted class.

For class 0, indicated by the blue line on the left (Figure 3), the model output begins near 0.4 and is driven progressively leftward, terminating at approximately 0.1 and correctly falling well below the decision threshold. The dominant contributors to this displacement are *restorative sleep* (value = 12) and *social support* (value = 15), high scores that the model interprets as evidence against GAD-7 positivity. For class 1, indicated by the pink line on the right (Figure 3), a different observation starts near the 0.7 value and is pushed sharply rightward to approximately 1.0, driven by the opposite profile: low values on *restorative sleep* (1) and *social support* (1), both of which the model treats as strong positive evidence for GAD-7 classification. Across both, no single feature alone determines the outcome; rather, it is the coherent accumulation of domain scores, reinforced by the sex variable, that produces a confident classification. This is consistent with the population-level patterns observed in the beeswarm plot.

4 DISCUSSION

4.1 Interpretation of findings

This study examined the contribution of sociodemographic and lifestyle factors to predict anxiety symptoms among university students, using machine learning models combined with SHAP-based interpretation. The findings reinforce the multifactorial nature of anxiety and highlight the relevance of both behavioral and social determinants.

Restorative sleep emerged as the most influential predictor of GAD-7 classification, consistent with the well-established bidirectional relationship between sleep disturbance and anxiety. Poor sleep quality has been repeatedly associated with heightened anxiety symptoms in university students, with circadian dysregulation and insufficient sleep duration identified as key mechanisms [5]. SHAP value analysis further supports this finding, showing that lower sleep scores are associated with higher predicted probability of GAD-7 positivity, whereas higher scores are associated with lower risk. These results reinforce sleep health as a key target for intervention in this population.

Sex was the second most influential predictor of GAD-7 classification. SHAP analysis indicated that the female category was associated with a higher predicted probability of GAD-7 positivity compared to males. This pattern aligns with the epidemiological literature documenting that women are approximately twice as likely to meet criteria for generalized anxiety disorder compared to men [16]. Several mechanisms have been proposed to explain this difference, including neurobiological factors such as hormonal modulation of stress reactivity [17], as well as psychosocial influences related to gender-specific socialization processes and academic performance pressures [18]. The prominence of sex in the model underscores the importance of gender-sensitive approaches in both screening and intervention design.

Social support ranked third, with the SHAP analysis showing a clear directional effect, with higher social functioning scores associated with a lower predicted probability of anxiety, whereas lower scores were associated with increased risk. This is consistent with extensive evidence positioning perceived social support as a protective factor against psychological distress, and social isolation as an independent risk factor for anxiety and depression [7]. In the Brazilian context, where familial and communal bonds play a central cultural role, geographic relocation for undergraduate students may intensify feelings of disconnection.

Sexual orientation was the fourth most influential predictor, with sexual minorities' identities associated with a higher predicted probability of anxiety. This pattern is consistent with Meyer's minority stress model [19], which conceptualizes stigma, discrimination, and identity concealment as chronic stressors. In university settings, these stressors may be intensified by identity development processes and the lack of supportive institutional environments [20]. In Brazil, the persistent structural discrimination against LGBTQ+ individuals [21] reinforces the need for inclusive and affirming policies within higher education institutions.

Stress management showed the smallest contribution among the five most influential predictors, and its direction was counterintuitive: higher stress management scores, which indicate better self-reported stress management, were associated with a slightly higher predicted probability of GAD-7 positivity (Figure 2). Given the cross-sectional design, this most likely reflects reverse causation rather than a harmful effect. Coping efforts are mobilized in response to appraised threat [22], so students experiencing anxiety symptoms may be more likely to adopt or report stress-management strategies such as relaxation techniques or help-seeking. Residual confounding is also plausible: women tend to report better stress management while also presenting higher anxiety, and SHAP values for this feature are conditional on sleep and social support, which absorb much of the shared variance. Given its small magnitude, this association should be interpreted with caution and examined in longitudinal analyses.

Beyond ranking predictors by importance, the SHAP framework enabled a detailed characterization of *how* each feature operates within the model. At the global level, the beeswarm plot revealed not only the relative contribution of each predictor but also its directionality: lower sleep and social scores consistently pushed predictions toward GAD-7 positivity, while higher scores were systematically protective, confirming monotonic, dose-dependent relationships aligned with theoretical expectations. At the local level, the decision plots illustrated how these effects accumulate across features for individual observations, demonstrating that confident positive classifications arise not from any single dominant predictor but from the coherent co-occurrence of adverse profiles across multiple domains: low sleep, low social support, and female sex acting in concert. This instance-level resolution is particularly valuable in clinical and public health contexts, where understanding *who* is at risk and *why* is as important as aggregate prediction accuracy.

The convergence of modifiable lifestyle factors, namely sleep quality, social functioning, and stress management, among the most influential predictors, has important implications for university mental health policy. Sleep, identified as the most influential predictor, highlights the relevance of sleep hygiene interventions, including psychoeducation, structured scheduling support, and the reduction of nocturnal academic demands, as priority targets for intervention.

The identification of sex and sexual orientation as influential sociodemographic factors suggests the need for differentiated, identity-informed screening approaches, rather than universal, one-size-fits-all strategies. Social connectedness also emerged as a key factor, indicating the potential value of peer support networks, mentorship programs, and inclusive campus environments as preventive measures. Collectively, these findings suggest that effective university mental health strategies should be multilevel, combining

individual behavioral interventions with broader structural and sociocultural approaches, and prioritizing the populations and domains identified as most influential by the model.

4.2 Methodological considerations and contribution of the pipeline

The use of ML, rather than conventional statistical approaches, was motivated by the high-dimensional and potentially non-linear nature of anxiety determinants, for which gradient-boosted models such as LightGBM are well suited. The final model selection was guided by statistical comparison: applying the non-parametric framework proposed by Demšar [15] revealed no significant performance differences among the five evaluated classifiers, supporting a choice based on practical considerations, including computational efficiency, feature importance capabilities, and compatibility with interpretability tools.

Predictive performance alone may be insufficient to meet the objectives of this study, as opaque models provide limited insight for researchers and policymakers. In this context, SHAP offers a principled approach to feature attribution grounded in cooperative game theory. While traditional statistical approaches evaluate the influence of a single feature, the SHAP method provides a way to measure the importance of a feature in conjunction with other features. It enables both global interpretation, through aggregated feature importance measures, and local interpretation, by explaining individual predictions (Figure 3). This dual perspective supports a more transparent understanding of model behavior in population health applications. An additional advantage of SHAP values lies in their model-agnostic nature. By decoupling the explanatory layer from the predictive model, interpretability can be maintained across different algorithms, providing flexibility in model selection without compromising transparency.

4.3 Strengths and limitations

This study benefits from several notable strengths. The sample, comprising 6,398 undergraduate students drawn from 13 universities across all five Brazilian regions, provides broad national coverage that is rarely achieved in this literature. The use of validated instruments (GAD-7; U-SMILE) ensures cross-study comparability, while the pipeline's strict leakage-prevention protocol, embedding all preprocessing within cross-validation folds, strengthens result reliability. The combined use of global and local SHAP value analyses represents a methodological contribution, enabling a detailed level of interpretability that is less commonly applied in studies of anxiety among Brazilian university students.

Nonetheless, several limitations warrant consideration. The cross-sectional design precludes causal inference, and the directional effects identified by SHAP values, although theoretically consistent, should not be interpreted as causal or temporally stable. The dichotomization of GAD-7 scores reduces outcome granularity and may obscure meaningful variation in symptom severity. Finally, reliance on self-reported data may introduce reporting bias, including potential underreporting of anxiety symptoms, particularly among groups for whom stigma may act as a barrier to disclosure. This consideration is especially relevant in light of the findings related to sexual orientation.

5 CONCLUSION

This study identified key determinants of anxiety among Brazilian university students using ML combined with SHAP-based interpretability. Sleep quality, sex, social functioning, sexual orientation, and stress management emerged as the most influential predictors, with both global and local SHAP analyses revealing their direction and relative contributions. By applying interpretable ML to a large, nationally distributed sample of 6,398 students, this work advances understanding of how modifiable and non-modifiable risk factors jointly shape anxiety risk in this population. The findings support targeted interventions such as sleep hygiene programs, peer support networks, and identity-sensitive screening within universities. Future research should adopt longitudinal designs to assess causal pathways and extend this framework to other mental health outcomes.

Author contributions. The authors contributed equally to all stages of the article: conception and design of the study; data collection, analysis, and interpretation; drafting and critical revision of the manuscript; and approval of the final version.

Acknowledgments. The authors gratefully acknowledge the UNILIFE-M Team and all participating universities and members for their contribution to this research.

Funding. This work was supported by the Instituto Nacional de Tecnologia and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (Call No. 21/2023: Transdisciplinary Studies in Public Health, Grant 444435/2023-9), which also funded a PIBIC undergraduate research scholarship.

Data and materials availability. The authors did not provide a data and materials availability statement.

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

Use of artificial intelligence. ChatGPT (OpenAI, GPT-5.3) was used for language editing and to improve the clarity and organization of the manuscript. It was not used to develop scientific content; all conceptual, methodological, and analytical decisions were made by the authors.

REFERENCES

1. World Health Organization. World mental health today: latest data [Internet]. Geneva: WHO; 2025 [cited 2026 Oct 15]. Available from: https://www.who.int/publications/i/item/9789240113817
2. UNICEF. The State of the World's Children 2024: The Future of Childhood in a Changing World [Internet]. New York: UNICEF; 2024 [cited 2026 Oct 15]. Available from: https://www.unicef.org/reports/state-of-worlds-children/2024
3. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, et al. WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders. J Abnorm Psychol. 2018;127(7):623-38. doi: https://doi.org/10.1037/abn0000362
4. Bruffaerts R, Mortier P, Kiekens G, Auerbach RP, Cuijpers P, Demyttenaere K, et al. Mental health problems in college freshmen: Prevalence and academic functioning. J Affect Disord. 2018;225:97-103. doi: https://doi.org/10.1016/j.jad.2017.07.044
5. Becker SP, Jarrett MA, Luebbe AM, Garner AA, Burns GL, Kofler MJ. Sleep in a large, multi-university sample of college students: sleep problem prevalence, sex differences, and mental health correlates. Sleep Health. 2018;4(2):174-81. doi: https://doi.org/10.1016/j.sleh.2018.01.001
6. Sheldon E, Simmonds-Buckley M, Bone C, Mascarenhas T, Chan N, Wincott M, et al. Prevalence and risk factors for mental health problems in university undergraduate students: A systematic review with meta-analysis. J Affect Disord. 2021;287:282-92. doi: https://doi.org/10.1016/j.jad.2021.03.054
7. Noble N, Paul C, Turon H, Oldmeadow C. Which modifiable health risk behaviours are related? A systematic review of the clustering of Smoking, Nutrition, Alcohol and Physical activity ('SNAP') health risk factors. Prev Med. 2015;81:16-41. doi: https://doi.org/10.1016/j.ypmed.2015.07.003
8. Mangolini VI, Andrade LH, Wang YP. Epidemiologia dos transtornos de ansiedade em regiões do Brasil. Rev Med (São Paulo). 2019;98(6):415-22. doi: https://doi.org/10.11606/issn.1679-9836.v98i6p415-422
9. Vieira VMSA, Torrenté MON. Saúde mental e interseccionalidade entre estudantes em uma universidade pública brasileira. Interface (Botucatu). 2022;26:e210674. doi: https://doi.org/10.1590/interface.210674
10. Daza A, Saboya N, Necochea-Chamorro JI, Zavaleta Ramos K, Vásquez Valencia YR. Systematic review of machine learning techniques to predict anxiety and stress in college students. Inform Med Unlocked. 2023;43:101391. doi: https://doi.org/10.1016/j.imu.2023.101391
11. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017); 2017 Dec 4-9; Long Beach, CA, EUA. Red Hook (NY): Curran Associates; 2017. p. 4765-74. [cited 2026 Oct 15]. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
12. Schuch FB, Waclawoscky A, Tornquist D, Oyeyemi AL, Sadarangani KP, Takano K, et al. UNIversity students' LIFEstyle behaviours and Mental health cohort (UNILIFE-M): study protocol of a multicentre, prospective cohort study. BMJ Open. 2026;16(1):e085006. doi: https://doi.org/10.1136/bmjopen-2024-085006
13. De Boni RB, Mota JC, Schuch FB, Pires DA, Matias TS, Monteiro-Junior RS, et al. U-SMILE: a brief version of the Short Multidimensional Inventory on Lifestyle Evaluation. Trends Psychiatry Psychother. 2025;47:e20230722. doi: https://doi.org/10.47626/2237-6089-2023-0722
14. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006;166(10):1092-7. doi: https://doi.org/10.1001/archinte.166.10.1092
15. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7(1):1-30.
16. McLean CP, Asnaani A, Litz BT, Hofmann SG. Gender differences in anxiety disorders: prevalence, course of illness, comorbidity and burden of illness. J Psychiatr Res. 2011;45(8):1027-35. doi: https://doi.org/10.1016/j.jpsychires.2011.03.006
17. Seeman MV. Psychopathology in women and men: focus on female hormones. Am J Psychiatry. 1997;154(12):1641-7. doi: https://doi.org/10.1176/ajp.154.12.1641
18. Chaplin TM. Gender and Emotion Expression: A Developmental Contextual Perspective. Emot Rev. 2015;7(1):14-21. doi: https://doi.org/10.1177/1754073914544408
19. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull. 2003;129(5):674-97. doi: https://doi.org/10.1037/0033-2909.129.5.674
20. Woodford MR, Kulick A, Sinco BR, Hong JS. Contemporary heterosexism on campus and psychological distress among LGBQ students: the mediating role of self-acceptance. Am J Orthopsychiatry. 2014;84(5):519-29. doi: https://doi.org/10.1037/ort0000015
21. Alves RF. Atitudes dos/as estudantes universitários/as face à homossexualidade: tradução e validação de uma escala de medida. Educ Rev. 2018;34(71):191-204. doi: https://doi.org/10.1590/0104-4060.58396
22. Lazarus RS, Folkman S. Stress, appraisal, and coping. New York: Springer Publishing Company; 1984.

EDITORIAL NOTES

The Journal of Health Informatics is a scientific journal maintained by the Brazilian Society of Health Informatics (Sociedade Brasileira de Informática em Saúde), a non-profit scientific association dedicated to the development of Health Informatics and Digital Health in Brazil. The editorial process is conducted with scientific independence, free from institutional, commercial, or sponsor interference, and follows international best practices in publication ethics.



This article was double-blind peer reviewed by specialists in the field. The editorial team used artificial intelligence tools, under human supervision, in the initial screening, in checking the text, references and numbers and, when needed, in redrawing figures, charts and tables, which the authors approved in the proof. The tools and safeguards are described in the journal's Policy on the use of artificial intelligence. Editorial decisions were made by people.