Hoang Thi Lam, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Machine Learning-Based Clustering and Prediction of Students' Mental Health Care Needs: Evidence from Hanoi National University of Education

Hoang Thi Lam[1], Pham Minh Phuong[2], Pham Quoc Thang[3]
[1,3]Tay Bac University, Vietnam
[3]Faculty of Psychology and Education, Hanoi National University of Education, Hanoi, Vietnam.

Abstract- This study extends an undergraduate thesis dataset on mental health care needs among 300 students at Hanoi National University of Education by adding machine learning analyses. Three need dimensions, namely cognitive/informational needs, emotional support needs, and access-to-service needs, were used for K-means clustering, while demographic variables, DASS-21 scores, and contextual factors were used to train Decision Tree and Random Forest models. Results showed a relatively high overall need level (M = 3.93/5). K-means with k = 3 identified low-need (25.7%), moderate-need (44.0%), and high-need (30.3%) groups. Random Forest outperformed Decision Tree, reaching 78.9% accuracy in three-class prediction and 85.6% accuracy in detecting high-need students. Family, learning-environment, socio-cultural, and personal factors were the strongest predictors. The findings suggest that machine learning can complement traditional statistics for early screening and targeted student support.

Keywords: Mental health care needs, students, DASS-21, K-means clustering, Decision Tree, Random Forest, educational data mining.

I. INTRODUCTION

The mental health of university students is widely recognized as a multidimensional issue that affects learning, social functioning, quality of life, and professional preparation. The World Health Organization conceptualizes mental health not merely as the absence of disorder, but as a state that enables individuals to cope with everyday stressors, work productively, and contribute to their communities [1]. Within higher education, mental health care should therefore include prevention, information, emotional support, and access to professional services rather than only treatment after serious symptoms occur.

Students are a particularly vulnerable population because they are in a transitional stage between adolescence and adulthood. They face academic workload, examination pressure, career orientation, financial constraints, family expectations, and social comparison through online networks. International studies have reported increasing mental health concerns and service utilization among university students, while also documenting gaps between students' needs and their actual help-seeking behaviors [3], [4], [5]. These gaps are often linked to stigma, lack of information, low service availability, and uncertainty about where to seek support.

In the context of teacher education, students' mental health care needs have special significance. Teacher trainees are future educators who require not only professional competence but also emotional regulation, resilience, and healthy interpersonal functioning. A previous undergraduate thesis at Hanoi National University of Education surveyed 300 students and examined their mental health care needs across three main dimensions: cognitive or informational needs, emotional support needs, and service access needs [14]. The thesis used questionnaire-based assessment, DASS-21, descriptive statistics, and inferential analysis.

The present paper develops that dataset further by applying machine learning methods. The aim is not to replace classical statistical analysis, which remains necessary for scale checking, description, and hypothesis testing. Instead, machine learning is used as a complementary approach to reveal hidden patterns, classify students into need-based groups, predict high-need students, and identify predictive factors that can inform more targeted support programs. Specifically, the study asks: (1) What groups of students can be identified according to their mental health care needs? (2) How accurately can Decision Tree and Random Forest models predict students' need-level groups? (3) Which factors are most important in predicting students' mental health care needs?

II. MATERIALS AND METHODS

Study design and dataset

This study used a secondary quantitative dataset derived from the undergraduate thesis entitled "The current situation of mental health care needs among students at Hanoi National University of Education" [14]. The sample consisted of 300 students. Data were collected from January to April 2026 through a questionnaire, the DASS-21 scale, and semi-structured interviews in the original thesis design. The present article focuses on the coded survey dataset and applies additional machine learning analyses.

The survey measured mental health care needs using a five-point Likert scale. Three composite dimensions were used for clustering: cognitive/informational needs, emotional support needs, and access-to-service needs. Mental health symptoms were represented by Depression, Anxiety and Stress scores from the Depression Anxiety Stress Scales-21 (DASS-21), a widely used screening instrument [6]. Contextual predictors included personal, family, learning-environment, and socio-cultural factors.

Variables and preprocessing

The clustering variables were the mean scores of the three need dimensions. For machine learning prediction, the input variables included demographic/general variables (A1-A5), DASS-21 component scores, personal factors, family factors, learning-environment factors, and socio-cultural factors. Variables directly used to form the clustering outcome were excluded from the prediction inputs to avoid circular prediction. Numeric variables were checked and standardized where required. The dataset was split into training and testing subsets using a 70:30 ratio with stratification by the need-level group.

Table -1: Research variables and analytical purposes

Component	Variables	Analytical purpose
Need dimensions	Cognitive, emotional, service-access need scores	K-means clustering
Mental health indicators	Depression, anxiety, stress scores	Predictor variables
Contextual factors	Personal, family, learning environment, socio-cultural	Predictor variables
General information	A1-A5 demographic/general variables	Control predictors

Machine learning procedure

K-means clustering was applied to standardized need scores to identify student groups. The number of clusters was examined from k = 2 to k = 5 using the Silhouette coefficient [8]. Although k = 2 had the highest Silhouette value, k = 3 was selected because it produced three educationally meaningful groups: low, moderate, and high need. This three-level segmentation is more useful for differentiated student support than a simple low/high split.

For supervised prediction, Decision Tree and Random Forest classifiers were trained to predict the three K-means need groups. Decision Tree was included because it provides simple interpretable rules [10]. Random Forest was included because ensemble learning can improve generalization by aggregating multiple decision trees [9]. Performance was evaluated using accuracy, precision, recall, and F1-score. A second binary Random Forest model was built to distinguish the high-need group from the remaining students, reflecting the practical screening objective of identifying students who may require prioritized support.

Ethical considerations

Because the study concerns mental health-related data, ethical caution is essential. The analysis used anonymized survey data and the results are reported at the group level. Machine learning outputs were interpreted only as screening and decision-support evidence, not as clinical diagnosis. The feature importance values represent predictive contributions within the model and should not be interpreted as proof of causality. Any institutional use of similar models should ensure informed consent, confidentiality, fairness, human oversight, and access to professional psychological support.

III. RESULTS

Descriptive statistics of need dimensions

The overall mean score of students' mental health care needs was 3.93 out of 5, indicating a relatively high level of need. The three dimensions were highly similar: cognitive/informational needs had a mean of 3.921, emotional support needs had a mean of 3.944, and access-to-service needs had a mean of 3.933. This pattern indicates that students' needs were not limited to emotional help after distress appeared; they also involved information, self-recognition, preventive skills, and accessible support services.

Table -2: Descriptive statistics of mental health care need dimensions

Need dimension	Mean	SD
Cognitive/informational needs	3.921	0.586
Emotional support needs	3.944	0.595
Access-to-service needs	3.933	0.608
Overall need score	3.933	0.574

K-means clustering of students

The Silhouette coefficient was 0.496 for k = 2, 0.475 for k = 3, 0.397 for k = 4, and 0.359 for k = 5. Although k = 2 produced the highest coefficient, the k = 3 solution was retained because it separated students into low, moderate, and high need groups, which correspond better to tiered support planning in educational settings.

The final K-means solution identified 77 students in the low-need group (25.7%), 132 students in the moderate-need group (44.0%), and 91 students in the high-need group (30.3%). Therefore, nearly one-third of the surveyed students belonged to a group that may require more active attention and targeted psychological support.

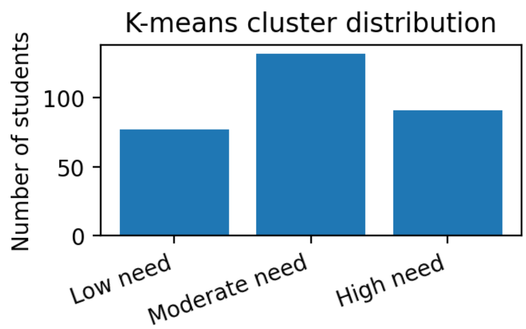

Chart -1: Distribution of K-means mental health care need clusters

Table -3: Cluster profile by psychological and contextual indicators

Indicator	Low	Moderate	High
DASS-21 total score	12.883	22.470	29.319
Depression score	4.896	8.492	11.066
Anxiety score	1.545	6.273	11.385
Stress score	1.792	8.326	12.495
Personal factors	3.227	3.930	4.602
Family factors	3.200	3.877	4.545
Learning environment	3.218	3.881	4.574
Socio-cultural factors	3.143	3.873	4.571

The cluster profile showed a clear gradient. Students in the high-need group had higher DASS-21 total scores and higher scores on personal, family, learning-environment, and socio-cultural factors than the moderate- and low-need groups. This suggests that high need is associated not only with symptom scores but also with broader contextual pressures.

Prediction of need-level groups

The three-class prediction results are summarized in Table -4. The Decision Tree achieved 73.3% accuracy and a macro-F1 score of 0.735. The Random Forest model performed better, with 78.9% accuracy and a macro-F1 score of 0.793. The improvement supports the usefulness of ensemble learning for classifying students' mental health care need levels.

Table -4: Three-class prediction performance

Model	Accuracy	Macro-F1
Decision Tree	0.733	0.735

Random Forest	0.789	0.793

For the high-need group, Random Forest achieved precision of 0.818 and recall of 0.667. In practical terms, when the model predicted a student as high need, it was usually correct; however, some high-need students were still classified as moderate need. This means that machine learning screening should not be used as the only basis for student support decisions.

In the binary task that distinguished the high-need group from the low/moderate groups, Random Forest achieved 85.6% accuracy and an F1-score of 0.735 for the high-need class. This binary formulation is practically useful because student support centers often need to identify individuals or groups who may require prioritized contact, further screening, or access to counseling services.

Table -5: Binary prediction of high-need students

Task	Accuracy	F1 high need
High need vs. non-high need	0.856	0.735

Important predictive factors

Random Forest feature importance showed that contextual factors had the largest predictive contribution. Family factors were the most important variable (0.3403), followed by the learning environment (0.1825), socio-cultural factors (0.1561), and personal factors (0.1468). DASS-21 component scores had lower importance values: depression (0.0560), stress (0.0497), and anxiety (0.0415). General information variables had only small contributions.

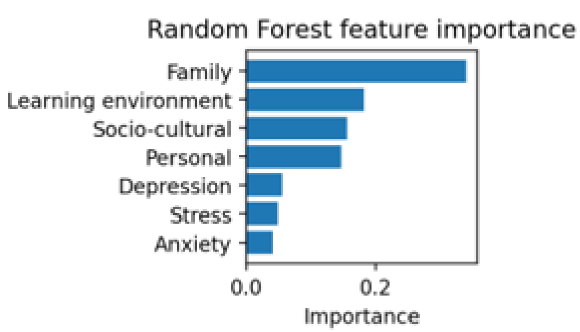

Figure 1: Important predictive factors in the Random Forest model

Table -6: Top predictive factors from Random Forest

Rank	Feature	Importance
1	Family factors	0.3403
2	Learning environment	0.1825
3	Socio-cultural factors	0.1561
4	Personal factors	0.1468
5	Depression score	0.0560
6	Stress score	0.0497
7	Anxiety score	0.0415

Correlation analysis was consistent with the feature-importance pattern. Overall need score was strongly correlated with family factors ($r = 0.906$), learning environment ($r = 0.885$), socio-cultural factors ($r = 0.884$), and personal factors ($r = 0.874$). DASS-21 scores had moderate to fairly strong correlations: depression ($r = 0.663$), stress ($r = 0.587$), and anxiety ($r = 0.576$).

The Decision Tree also provided interpretable rules. For example, when family factor scores were low, students were more likely to belong to the low-need group. When family factors and personal factors were both high, students were more likely to be classified into the high-need group. These rules are useful for interpretation but should be regarded as preliminary patterns that require validation on independent samples.

IV. DISCUSSION

The results demonstrate that machine learning can expand the value of questionnaire-based mental health care research. Traditional statistics describe average need levels and relationships between variables, while clustering and prediction models support segmentation and early screening. The three clusters identified in this study correspond to meaningful levels of support: universal mental health literacy for all students, skills workshops and group support for moderate-need students, and more active counseling or referral options for high-need students.

A notable finding is that contextual factors were more predictive than DASS-21 scores alone. This does not mean that depression, anxiety, or stress are unimportant. Rather, it suggests that mental health care needs are broader than symptom severity. Students may need information, emotional support, or access to services because of family expectations, academic pressure, social stigma, or limited coping skills, even when symptom scores are not extremely high.

The importance of family factors is particularly relevant in the Vietnamese cultural context. Family can be a major source of emotional and financial support, but it can also create pressure through academic expectations, career choices, and concerns about achievement. Mental health care programs in universities should therefore not only focus on individual students but also promote family communication and reduce unnecessary pressure from family expectations.

The learning environment and socio-cultural factors were also important. Academic workload, assessment pressure, relationships with lecturers and peers, and the availability of school-based support may shape students' perceived needs. At the same time, stigma and fear of being judged can prevent students from seeking help. Universities should therefore provide confidential, accessible, and flexible support channels, including online and anonymous options.

Random Forest outperformed the single Decision Tree, which is consistent with the general advantage of ensemble methods. However, Random Forest is less transparent than a single decision tree. In practice, Decision Tree rules can be used to communicate simple patterns to educators, while Random Forest can be used for more robust screening and feature ranking. Both approaches should remain under human supervision, especially because the dataset is modest in size and based on self-reported responses.

V. CONCLUSION

This study developed an undergraduate thesis dataset by adding machine learning analyses to cluster and predict students' mental health care needs. The findings show that students at Hanoi National University of Education reported relatively high mental health care needs, with an overall mean of 3.93/5. K-means clustering identified three interpretable groups: low need (25.7%), moderate need (44.0%), and high need (30.3%).

Random Forest performed better than Decision Tree in predicting the three groups, achieving 78.9% accuracy and a macro-F1 score of 0.793. In the binary screening task for the high-need group, Random Forest achieved 85.6% accuracy and an F1-score of 0.735 for the high-need class. The most important predictive factors were family factors, learning environment, socio-cultural factors, and personal factors.

The practical implication is that universities can use machine learning as a complementary tool to support needs-based mental health care planning. However, such models should not replace psychological assessment, counseling expertise, or ethical human decision-making. Future studies should validate the models on larger and multi-institutional datasets, include longitudinal data, and examine actual service-use behaviors.

VI. LIMITATIONS

This study has several limitations. First, the sample included 300 students from one university, so generalization should be made cautiously. Second, the dataset was based on self-reported questionnaire responses, which may be influenced by temporary mood, social desirability, and willingness to disclose psychological difficulties. Third, the machine learning models were evaluated using an internal train-test split and were not externally validated. Fourth, feature importance reflects prediction within the model and should not be interpreted as causal impact.

VII. RECOMMENDATIONS

Based on the findings, universities should consider a tiered support model. Universal programs should provide mental health literacy, stress-management skills, and information about available services. Moderate-need students may benefit from workshops, peer support, and group counseling. High-need students should have access to confidential individual counseling, online or anonymous consultation channels, and timely referral to professional services where necessary. Family communication and academic-environment improvement should be integrated into student mental health strategies.

Acknowledgements
The authors thank Hanoi National University of Education and the students who participated in the original survey. The present manuscript was developed from an undergraduate thesis dataset and extends the analysis by applying machine learning methods.

Conflict of Interest
The authors declare no conflict of interest. The results are intended for academic and educational support purposes only.

Data Availability
The anonymized dataset may be made available by the corresponding author upon reasonable request and subject to institutional and ethical restrictions.

REFERENCES

1. World Health Organization. (2022). World mental health report: Transforming mental health for all. Geneva: WHO.
2. Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. Journal of Health and Social Behavior, 43(2), 207-222.
3. Eisenberg, D., Hunt, J., Speer, N., & Zivin, K. (2011). Mental health service utilization among college students in the United States. Journal of Nervous and Mental Disease, 199(5), 301-308.
4. Lipson, S. K., Lattie, E. G., & Eisenberg, D. (2019). Increased rates of mental health service utilization by U.S. college students: 2007-2017. Psychiatric Services, 70(1), 60-63.
5. Yorgason, J. B., Linville, D., & Zitzman, B. (2008). Mental health among college students: Do those who need services know about and use them? Journal of American College Health, 57(2), 173-182.
6. Lovibond, S. H., & Lovibond, P. F. (1995). Manual for the Depression Anxiety Stress Scales. Psychology Foundation of Australia.
7. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297.
8. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.
9. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.
10. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
12. Oswalt, S. B., Lederer, A. M., Chestnut-Steich, K., Day, C., Halbritter, A., & Ortiz, D. (2019). Trends in college students' mental health diagnoses and utilization of services, 2009-2015. Journal of American College Health, 68(1), 41-51.
13. Cornish, P. A., Berry, G., Benton, S., Barros-Gomes, P., Johnson, D., Ginsburg, R., et al. (2017). Meeting the mental health needs of today's college student: Reinventing services through Stepped Care 2.0. Psychological Services, 14(4), 428-442.
14. Pham, M. P. (2026). The current situation of mental health care needs among students at Hanoi National University of Education. Unpublished undergraduate thesis, Faculty of Psychology and Education, Hanoi National University of Education, Vietnam.