JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 309 – 316 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4458
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

STUDENT DEPRESSION SCREENING BASED ON THE OPTIMUM DATA BALANCING AND RANDOM FOREST

M Sayyidul Adnan[1], Irwan Budi Santoso[1], Cahyo Crysdian[1*]
[1]Master of Information Technology, UIN Maulana Malik Ibrahim Malang
email: *cahyo@ti.uin-malang.ac.id

Abstract: Mental health issues, particularly depression among young adult university students, are often detected late due to stigma and reluctance to seek medical consultation. The objective of this study is to develop an early screening model employing machine learning techniques, specifically the random forest algorithm, on a dataset of 268 students (aged 17-29 years; consisting of 98 males and 170 females) within a multicultural educational setting. The principal challenges associated with this dataset are class imbalance and the potential for data leakage from clinical scores. This study implements a rigorous feature selection approach that involves the elimination of depression score features and the utilization of the Synthetic Minority Oversampling Technique (SMOTE) to balance the training data distribution. Furthermore, a Threshold Tuning strategy is employed to prioritize detection sensitivity (Recall). The findings indicate that reducing the decision threshold to an optimal value of 0.25 led to a substantial enhancement in the recall value, increasing it from 36% (baseline) to 77%. A feature importance analysis was conducted, the results of which indicated that Total Social Connectedness (ToSC) is the most dominant predictor. In summary, the present study corroborates the notion that optimizing sensitivity through threshold tuning is of paramount importance for medical screening. Furthermore, social isolation factors emerge as more significant indicators of depression risk than demographic attributes.

Keywords: data mining; depression; imbalanced data; random forest; smote; threshold tuning

Abstrak: Masalah kesehatan mental, khususnya depresi di kalangan mahasiswa dewasa muda, sering terdeteksi terlambat akibat stigma dan enggan mencari konsultasi medis. Tujuan studi ini adalah mengembangkan model skrining dini menggunakan teknik machine learning, khususnya algoritma random forest, pada dataset 268 mahasiswa (usia 17-29 tahun; terdiri dari 98 laki-laki dan 170 perempuan) dalam lingkungan pendidikan multikultural. Tantangan utama yang terkait dengan dataset ini adalah ketidakseimbangan kelas dan potensi kebocoran data dari skor klinis. Studi ini menerapkan pendekatan seleksi fitur yang ketat, yang melibatkan eliminasi fitur skor depresi dan penggunaan Teknik Over-sampling Minoritas Sintetis (SMOTE) untuk menyeimbangkan distribusi data pelatihan. Selain itu, strategi Penyesuaian Ambang Batas diterapkan untuk memprioritaskan sensitivitas deteksi (Recall). Hasil penelitian menunjukkan bahwa mengurangi ambang batas keputusan ke nilai optimal 0,25 menyebabkan peningkatan signifikan dalam nilai recall, dari 36% (dasar) menjadi 77%. Analisis pentingnya fitur dilakukan, hasilnya menunjukkan bahwa Total Social Connectedness (ToSC) adalah prediktor yang paling dominan. Secara ringkas, studi ini membenarkan bahwa mengoptimalkan sensitivitas melalui penyesuaian ambang batas sangat penting untuk skrining medis. Selain itu, faktor isolasi sosial muncul sebagai indikator risiko depresi yang lebih signifikan daripada atribut demografis.

Kata kunci: penambangan data; depresi; data tidak seimbang; hutan acak; smote; penyesuaian ambang batas

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 309 – 316 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4458
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Depression poses a significant global challenge [1], particularly within the context of higher education. International students encounter elevated levels of risk due to acculturative stress, language barriers, and feelings of isolation[2]. Early detection is of the essence, yet this is frequently impeded by social stigma and restricted access to services[3]. It is evident that delayed intervention has a detrimental effect on academic achievement and increases the risk of suicide[4]. Consequently, the utilisation of an objective, artificially intelligent screening system that employs demographic and social patterns is imperative.

Recent studies have indicated an escalation in depressive symptoms among students in the post-pandemic era, attributable to prolonged isolation and future uncertainty[5][6]. Furthermore, the transition to hybrid learning has given rise to novel stressors, including digital fatigue and diminished peer interaction[7].

As demonstrated in previous studies, machine learning has been shown to be an effective method for identifying mental disorders[8], [9]. While Convolutional Neural Networks (CNNs) and other deep learning models demonstrate high levels of accuracy, they necessitate substantial computational resources and extensive datasets, which are frequently not available in university clinical settings[10]. Consequently, conventional algorithms such as Random Forest continue to be the preferred approach for tabular data due to their interpretability and efficiency on smaller datasets[11]. Furthermore, Random Forest has been shown to be more effective in processing multi-dimensional medical data and is less susceptible to overfitting compared to single decision trees[12].

Social factors have also been identified as playing a critical role in this regard. Theories propose that social connectedness constitutes a fundamental human need[13], while discrimination or social isolation have been shown to have a significant impact on students' well-being[14]. Furthermore, it has been evidenced that problematic internet use is a comorbidity of depression in young adults [15].

However, extant datasets frequently exhibit vulnerabilities with respect to data leakage risks and severe class imbalance[16]. In addressing these gaps, the present study employs a secondary dataset from a multicultural setting. The primary objective is to develop a robust screening model that is optimised for sensitivity. The objective of this research is to address the issue of class imbalance through the implementation of the Synthetic Minority Oversampling Technique (SMOTE), with the ultimate goal of enhancing the identification of at-risk students (Recall) by leveraging Threshold Tuning. Additionally, the research seeks to analyse significant social predictors, thereby supporting the development of preventive interventions.

METHOD

The method delineated in this section adheres to the Knowledge Discovery in Databases (KDD) standard, with modifications to accommodate imbalanced classification. As illustrated in Figure 1, the research process encompasses a series of stages, including data collection and performance evaluation.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 309 – 316 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4458
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

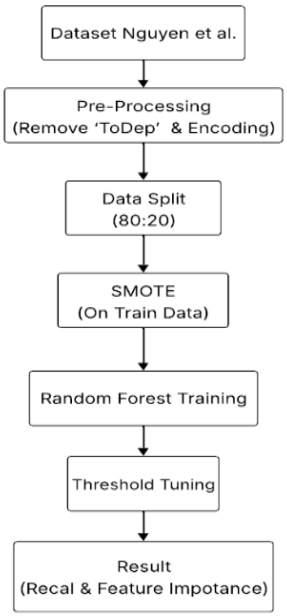

Figure 1. Proposed Research Methodology

The dataset used in this study is sourced from Nguyen et al.(2019), collected through a survey at Ritsumeikan Asia Pacific University (APU), Japan. It consists of 268 respondents, comprising 75% international students and 25% domestic students, predominantly falling within the young adult age category (17-29 years). The dependent variable is depression status (Dep), derived from the PHQ-9 instrument. The independent variables include demographics and social-psychological factors.

The preprocessing of data entailed a rigorous procedure of feature selection and encoding. The "Total Depression Score" (ToDep) variable was eliminated to avert data leakage, thereby ensuring that the model predicts based on social profiles rather than clinical scores. The conversion of categorical variables was executed through the utilization of Label Encoding. Imbalanced datasets present a considerable challenge in the context of medical diagnosis, as conventional classifiers often exhibit bias toward the majority class, thereby overlooking the minority positive cases[17]. In order to remediate the class imbalance, characterized by a preponderance of "Healthy" respondents relative to "Depressed" respondents, the Synthetic Minority Over-sampling Technique (SMOTE)[18] was implemented. SMOTE is an algorithm that generates synthetic data for the minority class based on k-nearest neighbors in the feature space.

Table 1. Description of Dataset Attributes

Attribute Group	Variable Description
Demographics	Includes Gender (Sex), Age (Years), Academic (Level of study), and Stay (Duration of stay in years).
Social Metrics	ToSC (Total Social Connectedness): A score measuring the sense of belonging and social bond.
Help-Seeking	Ten variables measuring preference to seek help from: Partner, Friends, Parents, Relative, Profess (Professional), Doctor, Reli (Religion), Alone (Self-help), Others, and Internet.
Clinical Score	ToDep (Total Depression Score): The raw PHQ-9 score used to label the target. (Note: This feature is removed during training to prevent data leakage).
Target Class	Dep (Depression Status): The binary classification target (0 = Healthy, 1 = Depressed).

This process was applied exclusively to the training set, while the origi-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 309 – 316 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4458
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

nal testing set was maintained for the purpose of valid evaluation.

The classification model employed is Random Forest [18], an ensemble learning technique that combines predictions from multiple decision trees. The mechanism in question involves two distinct processes: Bootstrap Aggregating (Bagging) and Random Feature Selection. The purpose of these processes is to reduce the correlation between trees. In the context of tree construction, the Gini Impurity index serves as a metric for evaluating the quality of node splits. The calculation of impurity for a node t is delineated in Equation (1).

$$Gini(t) = 1 - \sum_{i=0}^{c-1}[p(i|t)]^2 \quad (1)$$

In this context, $p(i|t)$ signifies the proportion of records that fall into class *i* at node *t*, with *c* representing the total number of classes. The final prediction is subsequently determined through a Majority Voting mechanism, as expressed in Equation (2).

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_k(x)\} \quad (2)$$

In this context, $h_1(x)$ signifies the prediction output of the *i*-th decision tree, while *k* represents the total number of trees in the forest. In this study, the model was configured with 200 trees (n_estimators=200) and a maximum depth of 5 (max_depth=5) to prevent overfitting. The class that manifests most frequently among these *k* predictions is selected as the final output. This ensemble approach effectively reduces variance and improves model stability in comparison with the use of a single decision tree.

In order to validate the model, the dataset was divided into a training set, which constituted 80% of the data, and a testing set, which constituted 20% of the data. This division was accomplished through the use of stratified sampling, a technique that was employed to ensure the preservation of the class distribution. The experiment was executed in its entirety through the utilization of the Python 3.9 programming language in conjunction with the Scikit-learn library. Moreover, given that the conventional classification threshold (0.5) frequently proves to be inadequate for imbalanced data, this study employs a Threshold Tuning strategy. This process entails the adjustment of the decision boundary, predicated on the estimation of probabilities, with the objective of optimizing the recall while maintaining a reasonable level of precision.

RESULT AND DISCUSSION

This section undertakes an evaluation of the classification results on the 20% testing set, with a focus on prioritising recall (sensitivity) over accuracy, a strategy necessitated by the imbalanced nature of the dataset. The experiment compares the standard Random Forest model with the proposed model that has been optimised with SMOTE and Threshold Tuning.

Performance Analysis and Threshold Tuning

Initially, the Random Forest model with the standard decision threshold (0.50) demonstrated unsatisfactory performance in detecting the minority class. As demonstrated in Figure 2(a), the standard model failed to identify a substantial number of depressed students, leading to 14 false negatives and a low recall of 36%. This is particularly problematic in the context of medical screening, where the failure to adequately rec-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 309 – 316 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4458
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

ord a patient's vital signs can have serious consequences.

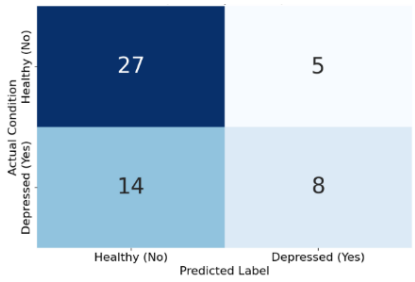

Figure 2 Confusion Matrix at Threshold 0.5

In order to rectify this issue, a threshold tuning strategy was initiated, which entailed the reduction of the decision boundary to 0.25. As illustrated in Figure 2(b), the confusion matrix of the optimized model is presented.

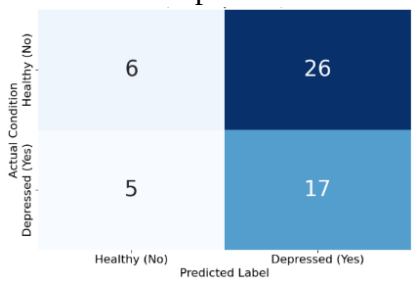

Figure 3 Confusion Matrix at Threshold 0.25

The optimized model demonstrates high efficacy, identifying 17 out of 22 depressed cases (True Positives), and significantly reducing false negatives to 5. Quantitative analysis of these results is presented in Table 2. While the standard threshold (0.50) offered superior precision (0.70), its recall (0.36) was inadequate for medical screening. Conversely, an adjustment of the threshold to 0.25 resulted in a surge to 0.77. Despite the rise in false positives, this strategic trade-off is deemed acceptable in order to minimise the number of missed diagnoses[11].

Table 2. Model Performance at Different Decision Thresholds

Threshold	Recall	Precision	False Negatif
0.50	0.36	0.62	14
0.45	0.50	0.58	11
0.40	0.50	0.48	11
0.35	0.64	0.48	8
0.30	0.68	0.45	7
0.25	0.77	0.40	5

This study also expanded the experiment to lower thresholds, specifically below 0.25, to assess the limits of the model. Nevertheless, the findings demonstrated a substantial decline in precision, a phenomenon often referred to as a "precision crash." At thresholds below 0.25, the model exhibited a tendency to categorize the majority of the test population as the positive class, designated as "Depressed." This condition resulted in a substantial increase in the number of false positives, thereby diminishing the model's capacity to differentiate between subjects with healthy and depressed states. Therefore, it was determined that the threshold of 0.25 was the optimal equilibrium point termed the "sweet spot" in this context where detection sensitivity (i.e., recall) is maximized just before the rate of misdiagnosis (i.e., false positives) becomes unmanageable.

Feature Importance Analysis

A comprehensive understanding of the determinants of depression is imperative for the development of effective intervention strategies. As illustrated in Figure 3, the Feature Importance scores are based on Gini impurity reduction. The analysis indicates that 'Total Social Connectedness" (ToSC) is the dominant predictor.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 309 – 316 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4458
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

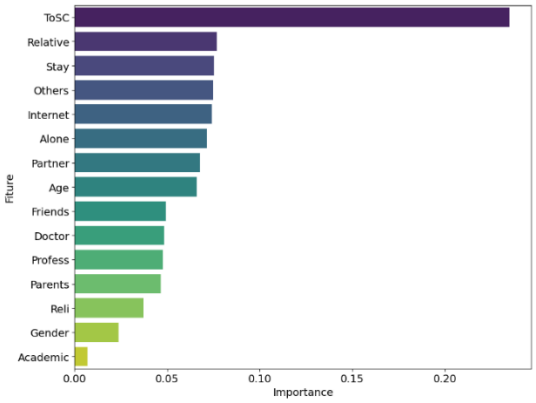

Figure 4. Feature Importance Score identifying key predictors of depression

The analysis indicates that "Total Social Connectedness" (ToSC) is the most dominant predictor, as indicated by its highest importance score. This finding aligns with psychological theories that posit a lack of belongingness as a fundamental precursor to depression[13]. Students who exhibit low social connectedness scores demonstrate a heightened degree of vulnerability. In accordance with the findings of ToSC, the variables associated with help-seeking behaviors namely, "Internet" and "Alone" emerge as preeminent predictors. This finding suggests a concerning pattern among students at risk, who often exhibit a tendency to withdraw from social interactions and rely on non-interpersonal coping mechanisms, such as the internet. The present findings align with the conclusions of Shannon *et al.*[15], which establish a correlation between problematic internet use and depressive symptoms.

A salient finding of the study was the relative insignificance of demographic variables such as gender and academic level in comparison to social-psychological factors. This finding indicates that, within this particular multicultural student population, the likelihood of depression is predominantly influenced by the quality of social interaction and coping mechanisms, rather than biological gender or academic strata. This insight is of significant value to university policymakers, as it underscores the importance of fostering inclusive social environments rather than merely targeting specific demographic groups.

CONCLUSION

The objective of this study was to develop an early depression detection model that is optimized for imbalanced data in a multicultural student context. The study was successful in achieving this objective. The integration of SMOTE and Random Forest was found to be an effective method for recognizing patterns characteristic of minority classes. A salient finding of this research is that the conventional decision threshold proves inadequate for medical screening. The application of a Threshold Tuning strategy at 0.25 was identified as the optimal equilibrium point, thereby significantly enhancing recall to 77% while maintaining a manageable false positive rate. Lowering the threshold to this lower limit resulted in a precision crash, thereby confirming 0.25 as the effective limit for this dataset. In addition, the feature analysis indicates that social isolation, as indicated by low Total Social Connectedness (ToSC) and non-interpersonal help-seeking behaviors (e.g., reliance on the internet), is the primary determinant of depression risk. These factors have been found to be more significant than biological or academic factors in this regard.

In the interest of advancing the field, it is advised that future endeavors involve the validation of this model on data sets with a more extensive cultural background. This will facilitate an assessment of its generalizability. Further-

more, subsequent studies may investigate the potential of hybrid deep learning methods to enhance precision without compromising recall. From a pragmatic standpoint, higher education institutions are strongly encouraged to prioritize intervention programs that emphasize fostering a sense of social belonging. Additionally, it is recommended that these institutions consider leveraging the optimized screening model as a proactive instrument to facilitate early mental health support.

BIBLIOGRAPHY

[1] D. Phiri, F. Makowa, V. L. Amelia, Y. V. A. Phiri, L. P. Dlamini, and M. H. Chung, "Text-Based Depression Prediction on Social Media Using Machine Learning: Systematic Review and Meta-Analysis," *J. Med. Internet Res.*, vol. 27, 2025, doi: 10.2196/59002.

[2] I. F. Kristiana, N. A. Karyanta, E. Simanjuntak, U. Prihatsanti, T. M. Ingarianti, and M. Shohib, "Social Support and Acculturative Stress of International Students," Jun. 01, 2022, *MDPI*. doi: 10.3390/ijerph19116568.

[3] S. Shahwan *et al.*, "The potential impact of an anti-stigma intervention on mental health help-seeking attitudes among university students," *BMC Psychiatry*, vol. 20, no. 1, Dec. 2020, doi: 10.1186/s12888-020-02960-y.

[4] Sindi Putri Ayu, Finkah Sabillah, Nurhidayah Nurhidayah, Bilqis Salsabila, Risma Anita Puriani, and Rizki Novirson, "Deteksi Dini Perilaku Depresi pada Siswa Sekolah Menengah," *WISSEN : Jurnal Ilmu Sosial dan Humaniora*, vol. 3, no. 2, pp. 191–201, May 2025, doi: 10.62383/wissen.v3i2.754.

[5] W. Luo, B.-L. Zhong, and H. F.-K. Chiu, "Prevalence of depressive symptoms among Chinese university students amid the COVID-19 pandemic: a systematic review and meta-analysis," *Epidemiol. Psychiatr. Sci.*, vol. 30, p. e31, Mar. 2021, doi: 10.1017/S2045796021000202.

[6] J. Deng *et al.*, "The prevalence of depressive symptoms, anxiety symptoms and sleep disturbance in higher education students during the COVID-19 pandemic: A systematic review and meta-analysis," *Psychiatry Res.*, vol. 301, p. 113863, Jul. 2021, doi: 10.1016/j.psychres.2021.113863.

[7] A. Abbas, G. B. Rincón, L. Wang, and M. K. Siddiqui, "Investigating the Impact of Technostress on Perceived Hybrid Learning Environment and Academic Performance," *Electronic Journal of e-Learning*, vol. 21, no. 4, pp. 382–393, Nov. 2023, doi: 10.34190/ejel.21.4.3084.

[8] M. Abdullah and N. Negied, "Detection and prediction of Future Mental disorder from Social Media Data using Machine Learning, Ensemble Learning, and Large Language Models.," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3406469.

[9] M. Tabares Tabares, C. Vélez Álvarez, J. Bernal Salcedo, and S. Murillo Rendón, "Anxiety in young people: Analysis from a machine learning model," *Acta Psychol. (Amst).*, vol. 248, Aug. 2024, doi: 10.1016/j.actpsy.2024.104410.

[10] N. Wang, R. Kamil, S. A. R. Al-Haddad, N. Ibrahim, and Z. Zhao, "Enhancing AI Depression Detec tion Using Transfer Learning," *Contemporary Mathematics*, pp. 3054–3080, May 2025, doi: 10.37256/cm.6320256184.

[11] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, Jun. 2024, doi: 10.1109/TNNLS.2022.3229161.

[12] A. N. S. Kinasih, A. N. Handayani, J. T. Ardiansah, and N. S. Damanhuri, "Comparative analysis of decision tree and rand om forest classifiers for structured data classification in machine learning," *Science in Information Technology Letters*, vol. 5, no. 2, pp. 13–24, Nov. 2024, doi: 10.31763/sitech.v5i2.1746.

[13] K.-A. Allen, M. L. Kern, C. S. Rozek, D. M. McInerney, and G. M. Slavich, "Belonging: a review of conceptual issues, an integrative framework, and directions for future research," *Aust. J. Psychol.*, vol. 73, no. 1, pp. 87–102, Jan. 2021, doi: 10.1080/00049530 .2021.1883409.

[14] A. Maleku *et al.*, "The hidden minority: Discrimination and ment al health among international students in the US during the COVID-19 pandemic," *Health Soc. Care Community*, vol. 30, no. 5, Sep. 2022, doi: 10.1111/hsc .13683.

[15] H. Shannon, K. Bush, P. J. Villeneuve, K. G. Hellemans, and S. Guimond, "Problematic Social Media Use in Adolescents and Young Adults: Systematic Review and Meta-analysis," *JMIR Ment. Health*, vol. 9, no. 4, p. e33450, Apr. 2022, doi: 10.2196/33450.

[16] M. Kavitha, "Enhanced Cost-sensitive Ensemble Learning for Imbalanced Class in Medical Data," *Journal of Electrical Systems*, vol. 20, no. 7s, pp. 1043–1053, May 2024, doi: 10.5278 3/jes.3520.

[17] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," in *2021 6th International Conference on Informatics and Computing, ICIC 2021*, Institute of Electrical and Electronics Engine ers Inc., 2021. doi: 10.1109/ ICIC54025.2021.9632912

[18] I. D. Mienye and Y. Sun, "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data," *Inform. Med. Unlocked*, vol. 25, p. 100690, 2021, doi: 10.1016/j.imu.2021.100690.