Integrating Feature Selection and Extraction with Tuned Multilayer Perceptrons for Predicting Student Dropout Risk in Higher Education

Zannatul Ferdousee
Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh
methela1603060@gmail.com

Md Rakibul Haque
Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh
rakibulhaq56@gmail.com

***Abstract*—Every year, thousands of students drop out of educational institutions, significantly undermining the effectiveness of our education systems. Identifying factors contributing to dropout and implementing early interventions based on these factors is crucial. However, the acquisition of institutional data poses significant challenges due to privacy and security concerns. Moreover, while the availability of educational datasets is increasing, these often suffer from poor structure or organisation or lack relevance to specific problems. Additionally, the tools commonly used in pedagogy may lack optimised processes. In this paper, we have proposed a novel approach to predict dropout-related factors based on socio-economic and academic features. Our method combines Random Forest for feature selection and Principal Component Analysis for feature extraction, followed by training a finely tuned neural network to classify potential dropouts effectively. We have evaluated our model on two benchmark datasets SATDAP and ICFES, achieving accuracies of 90.08% and 80.03%, respectively, thereby outperforming state-of-the-art machine learning classifiers. This demonstrates the potential of our approach in addressing the dropout challenge in educational settings.**

***Index Terms*—Random Forest, PCA, Educational Data Mining, Multi-Layer Perceptron**

I. Introduction

The dropout crisis in higher education is a significant global concern, with dropout rates varying widely but affecting millions of students each year. For instance, in the United States [1], approximately 40% of undergraduate students do not complete their degree programs within six years, while in Europe, dropout rates can exceed 20% in some countries [2]. Several factors contribute to this issue, including financial difficulties, academic challenges, lack of engagement, and personal circumstances that hinder a student's ability to sustain their education. These dynamics underscore the need for comprehensive strategies and interventions to improve student retention and success in higher education institutions worldwide.

Educational data mining has provided vast academic data, yielding new insights [3]. However, the reliance on statistical methods and automated tools [4] often limits practical feature engineering, reducing model precision. Advanced machine learning and sophisticated feature engineering are crucial for maximizing predictive capabilities. Fine-tuning features related to socio-economic, demographic, and family backgrounds can more accurately predict dropout rates, facilitating targeted interventions to improve student retention.

Researchers have used feature selection and extraction techniques to identify key factors in student dropout. Feature selection [5] isolates relevant features from the dataset, reducing redundancy and noise, while feature extraction [6], such as PCA, transforms features into fewer dimensions, capturing significant variance. Our study combines Random Forest-based feature selection with PCA, leveraging both interpretability and efficient data summarization for a comprehensive dropout predictor analysis.

Traditional dropout prediction has relied on classifiers like Decision Trees, Random Forests, and Support Vector Machines [7]. However, these methods struggle with large, complex datasets. With growing educational data, deep neural networks have shown superior accuracy and predictive power. Hence, we employed a finely tuned neural network for classification, enhancing our model's ability to learn complex patterns and improve student outcome predictions.

The main contributions of this paper are threefold: (1) developing an extensive feature space by integrating Random Forest-based selection with PCA, (2) optimizing a multi-layer perceptron for robust classification, and (3) surpassing state-of-the-art methods using two benchmark datasets. Our approach enhances predictive accuracy and sets a new standard in educational data analysis.

The paper is structured as follows: Section II details the dataset and methodology, Section III discusses the results and Section IV concludes with key findings and implications.

II. Proposed Approach

To tackle the challenge of identifying students at heightened dropout risk, we implemented a comprehensive feature selection approach utilizing Random Forest alongside a prominent feature extraction method, Principal Component Analysis (PCA). This combination was strategically employed to refine and enhance the quality of the feature space. By integrating these two methods, we crafted a superior feature space enriched with high-quality features. Subsequently, classification was performed using a finely tuned Multi-Layer Perceptron (MLP), optimized to effectively leverage this refined feature

979-8-3315-0822-7/24/$31.00 ©2024 IEEE

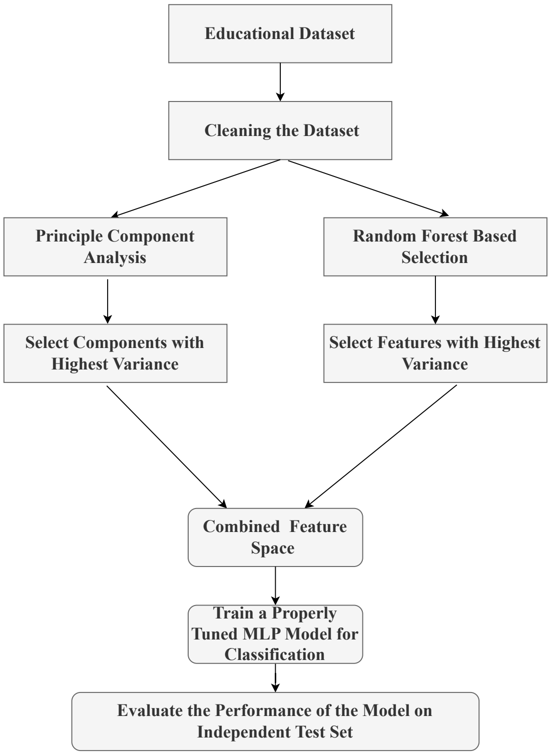

Educational Dataset
Cleaning the Dataset
Principle Component Analysis
Random Forest Based Selection
Select Components with Highest Variance
Select Features with Highest Variance
Combined Feature Space
Train a Properly Tuned MLP Model for Classification
Evaluate the Performance of the Model on Independent Test Set

Fig. 1: An Overview of the Proposed Approach

set. An illustrative overview of our methodological framework is provided in Figure 1, which depicts the sequential integration of feature selection, feature extraction, and classification processes.

A. Dataset Description

For proper evaluation of a method, a stringent and benchmark dataset is need. For this, we have used two benchmark dataset to evaluate the robustness and performance of our model.

1) SATDAP Dataset: The SATDAP [8] dataset was initially developed as part of a project aimed at reducing academic dropout and failure in higher education. The project's primary goal is to employ machine learning techniques to identify students at risk of dropping out early in their academic careers, thereby enabling the implementation of supportive strategies to improve retention rates. This dataset comprises 4424 samples and includes 35 features encompassing a comprehensive range of data available during student enrollment. These features cover aspects of the students' academic paths, demographics, and socio-economic factors, providing a rich basis for analyzing the determinants of student outcomes. The data captures diverse elements that might influence academic success or failure, including prior academic performance, family background, financial status, and more, making it ideal for developing predictive models that can forecast student graduation and dropout. Initially representing three classes of student outcomes, our study focuses on two primary classes: 'Graduated' and 'Dropout', to directly address our research objectives concerning student retention and performance enhancement.

2) ICFES Dataset: We have also used the ICFES [9] dataset to evaluate our proposed approach, which was primarily designed to survey engineering students. This dataset contains 12,411 observations, each representing an individual student, with a total of 44 variables. The dataset features two critical test scores: one from a national standardized test and another from an engineering subject-specific examination. We aggregated these scores to classify the dropout risk and segmented the entire sample space into four quartiles. Students in the lowest two quartiles (Q1 and Q2) were identified as being at risk of dropout or performing weakly, while those in the upper two quartiles (Q3 and Q4) were classified as not at risk. For our analysis, the dataset was refined to retain 26 features that encompass socio-economic and academic factors pertinent to our study. The target class was defined to indicate whether students are at risk of dropping out.

B. Dataset Cleaning

The datasets required extensive preprocessing to ensure they were primed for detailed analysis. Initially, we addressed any missing values by calculating and imputing them to maintain the data's completeness and integrity. Additionally, categorical variables have been systematically encoded to make them suitable for use in machine learning algorithms. We used a target encoder to convert the categorical variable to a numerical one. To facilitate the development and validation of our models, we partitioned the dataset into training and testing sets using a 70:30 ratio. Furthermore, 20% of the training set was designated for validation to refine the model's performance during the training phase. This thorough approach to data cleaning and preparation was meticulously designed to aid in the robust identification of critical predictors of student success, allowing for the development of precise interventions to boost retention rates in higher education.

C. Random Forest Based Feature Selection

Our methodology employs the Random Forest Classifier to assess and prioritise the significance of various predictors in our dataset [10]. The Random Forest, an ensemble learning method comprising multiple decision trees, evaluates the importance of each feature by calculating the decrease in model accuracy when that feature is omitted from the model. The importance metric ϕ for a feature i is quantified via Equation1 :

$$\phi(i) = \frac{\sum(\gamma(i))}{N_t} \tag{1}$$

Here, $\gamma(i)$ refers to a decrease in model accuracy when feature i is excluded, and N_t refers to the number of trees in the forest.

Our Random Forest model is configured with 100 estimators for a comprehensive assessment and a fixed seed for reproducibility, subsequently informing the feature selection process. The parameters of the models have been hypertuned. After that, we have selected a threshold of feature importance, which is defined mathematically in Equation and then used that threshold as a selection criteria 2:

$$\theta = \frac{\sum \phi(i)}{n} \tag{2}$$

$$\text{Selected Features} = \{x \mid \phi(x) > \theta\} \tag{3}$$

We have selected only those features whose importance surpasses the mean threshold. In the case of the SATDAP dataset, seven features were identified, encapsulating critical student-related information such as **Tuition fees up to date**, **Age at enrollment**, and various metrics concerning **curricular units** performance across semesters. Whereas in the case of the ICFES, the following features have been selected: '**Education of the Father**', '**Education of the Mother**, '**Occupation of Father**', '**Name of the School**', and '**Name of the University**'. These features are instrumental in refining the predictive model's accuracy and reducing complexity, as evidenced by the focused subset of variables that significantly influence the target outcome. This streamlined feature set underscores the efficacy of integrating robust machine-learning techniques in educational data analysis.

D. Principal Component Analysis

In our study, we employed Principal Component Analysis (PCA) [11] on two benchmark datasets to reduce dimensionality, aiming to capture significant patterns while reducing complexity. PCA operates by first standardizing the feature matrix, X_{scaled}, where each feature is centred around zero with unit variance. This preprocessing step is crucial for PCA because it ensures that the PCA algorithm treats all features equally, especially when they are measured on different scales. The key mathematical steps involved in PCA after scaling are as follows:

1) Compute the Covariance Matrix: : The covariance matrix, Σ, is computed to understand how the variables of the X_{scaled} relate to each other. The covariance matrix is defined as:

$$\Sigma = \frac{1}{n-1} X_{\text{scaled}}^T X_{\text{scaled}} \tag{4}$$

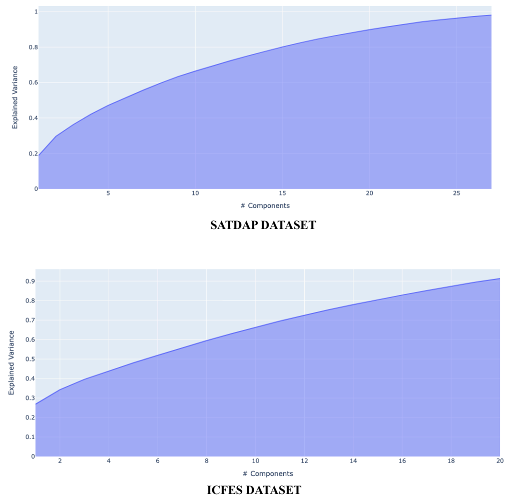
where X_{scaled}^T is the transpose of the scaled feature matrix. This matrix helps in understanding the correlation structure of the data.

2) Eigen Decomposition: : PCA finds the eigenvalues and eigenvectors of the covariance matrix. The eigenvalues and eigenvectors are computed as solutions to the equation:

$$\Sigma \mathbf{v} = \lambda \mathbf{v} \tag{5}$$

where λ represents the eigenvalues, and $\mathbf{v}$ are the corresponding eigenvectors. The eigenvalues represent the magnitude of the variance captured by each principal component, and the eigenvectors define the direction of these components. The eigenvectors are sorted by decreasing eigenvalues to prioritize components with maximum variance.

3) Select Principal Components: : Based on the sorted eigenvalues, several components are selected to capture a desired amount of total variance. For the SATDAP dataset, selecting 27 principal components accounted for 98% of the total variance. For the ICFES dataset, 20 components captured 90% of the variance, as shown in Figure 2.



SATDAP DATASET

ICFES DATASET

Fig. 2: Cumulative Varince Graph of both the datasets

4) Transform Data: : The original data X_{scaled} is then projected onto the space defined by the selected eigenvectors, transforming it into a new set of uncorrelated variables, which are the principal components. This transformation is performed using the equation:

$$X_{\text{new}} = X_{\text{scaled}} \mathbf{V}_{\text{selected}} \tag{6}$$

Where $\mathbf{V}_{\text{selected}}$ is the matrix containing the eigenvectors that correspond to the selected eigenvalues.

E. Multilayer Perceptron

In our study, we developed a deep learning model to classify student dropout risks, employing a neural network architecture designed to handle the complexities of educational data. The model consists of a sequential architecture with multiple layers of neurons, each followed by a ReLU activation function to introduce non-linearity, enhancing the model's ability to learn complex patterns in the data [12]. Specifically, the architecture includes an input layer that matches the number of features in X_{train}, followed by three hidden layers with 128, 64, and 32 neurons, respectively. The output layer uses a sigmoid activation function with one neuron, corresponding to the binary classification task of predicting whether a student will drop out. The forward pass can be defined using Equation 7

$$\begin{aligned} \mathbf{z}_i &= \mathbf{W}_i \mathbf{x} + \mathbf{b}_i \\ \mathbf{a}_i &= \text{ReLU}(\mathbf{z}_i) \end{aligned} \tag{7}$$

where $\mathbf{W}_i$ is the weight matrix of the $i'th$ layer, $\mathbf{b}_i$ is the bias vector , $\mathbf{z}_i$ is the linear combination before activation, and $\mathbf{a}_i$ is the output after activation.

In our study, we implemented the binary cross-entropy loss function to measure discrepancies between actual labels and predicted probabilities, providing a clear quantification of model accuracy for binary classification tasks. Gradient calculations guided the optimization process, efficiently executed using the gradient descent algorithm enhanced by the Adam

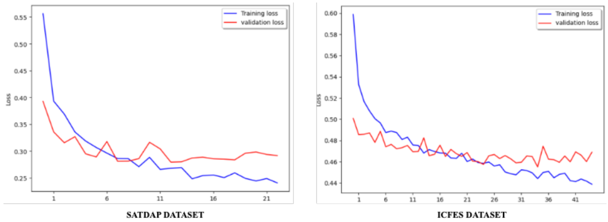


Fig. 3: Graphical Depiction of Training and Validation Loss for both the datasets.

TABLE I: Evaluation of our proposed approach on SATDAP dataset based on different Evaluation Metric

	Accuracy	Precision	Recall	F1-Score
Dropout	83.39	90.00	83.00	87.00
Graduate	94.20	90.00	94.00	92.00
Overall Accuracy	**90.08**	89.20	88.90	**89.04**

Optimizer, known for its adaptive learning rate capabilities. The optimizer helps swiftly converge to the optimal solution, with updates from the loss function backpropagated throughout the network to allow systematic adjustments to model weights.

Our training regimen spanned 100 epochs with batches of 16 samples to manage computational load effectively. We incorporated a 20% validation split to monitor the model's performance against unseen data and mitigate overfitting. An early stopping mechanism was also integrated, tracking validation loss and halting training if no improvement was noted over 3 epochs. This approach ensures the optimal use of computational resources and maintains the model's robustness and generalizability in predicting student dropout risks. The models have been trained for 29 epochs on SATDAP Dataset and 14 epochs for the ICFES dataset. From Figure 4 we can see after that; the validation loss gets stable.

III. Result Analysis

We have furnished a comprehensive analysis of the outcomes realized via our methodology and the underlying rationales. Additionally, we have presented ablation studies and comparative evaluations against other benchmark techniques.

A. Evaluation Based on Different Metric

While many researchers primarily rely on accuracy as the evaluation metric, we have recognized the importance of employing a broader set of metrics to more comprehensively assess our model's performance in predicting dropout. To this end, we have incorporated Precision, Recall, and F1-Score alongside accuracy. These metrics provide deeper insights into the model's ability to correctly identify true positives and balance the trade-offs between sensitivity and specificity, ensuring a more robust and nuanced evaluation.

TABLE II: Evaluation of our proposed approach on ICFES dataset based on different Evaluation Metric

	Accuracy	Precision	Recall	F1-Score
Dropout	83.93	80.00	84.00	82.00
Not Dropout	76.76	81.00	77.00	79.00
Overall Accuracy	**80.30**	81.00	81.00	**80.00**

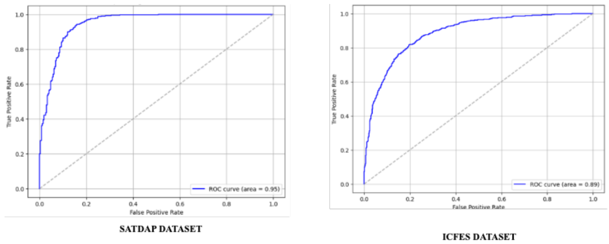


Fig. 4: ROC curve of our Proposed Approach on both the datasets

From Table I and II, it is patent that by achieving 90.08% accuracy on the SATDAP dataset and 80.05% accuracy on the ICFES dataset, our model has demonstrated its efficacy in predicting dropout students with high precision and recall across both datasets. Utilizing a Random Forest-based feature selector and Principal Component Analysis for feature extraction, we have effectively combined these techniques to create a more generalized feature space. This integration has facilitated the capture of crucial patterns and relationships within the data, enhancing the predictive capability of our model. Following this, employing a finely tuned Multi-Layer Perceptron (MLP) for classification has enabled superior performance, leveraging the refined feature space for robust and accurate classifications. This approach underscores the importance of sophisticated feature handling and advanced machine learning techniques in improving prediction outcomes in educational settings.

The robustness of our model has been rigorously evaluated using the Receiver Operating Characteristic (ROC) curve, further substantiating its efficacy in classifying student dropout risk [13]. We achieved impressive AUC (Area Under the Curve) scores of 95.00 for the SATDAP dataset and 85.00 for the ICFES dataset. These high AUC scores illustrate the model's strong discriminatory ability to distinguish between students who are at risk of dropping out and those who are not. The scores reflect the model's effective utilization of the refined feature space, achieved through our comprehensive approach of combining advanced feature selection and extraction techniques.

B. Ablation Study

An ablation study is a critical component in the evaluation of machine learning models as it systematically assesses the impact of various components and their combinations on the overall performance. In our research, the ablation study distinctly illustrated the added value of integrating Principal Component Analysis (PCA) and Random Forest (RF) in feature preparation before applying a Multi-Layer Perceptron (MLP) for classification. Our results, as shown in the table III, demonstrate that the combined approach of PCA and RF prior to MLP classification achieves the highest accuracy scores, with 90.08% on the SATDAP dataset and 80.03% on the ICFES dataset. This contrasts with the individual application of PCA+MLP and RF+MLP, which yielded lower accuracies. Specifically, PCA+MLP scored 88.97% and 78.45% , and

TABLE III: Results of the Ablation study of our proposed approach on both the dataset.

Method	SATDAP (Accuracy)	ICFES Dataset(Accuracy)
PCA+MLP	88.97	78.45
RF +MLP	87.45	76.45
PCA+RF+MLP	**90.08**	**80.03**

TABLE IV: Competitive Comparison with other benchmark methods methods

Method	SATDP	ICFES
	(Accuracy)	(Accuracy)
Decision Tree	84.02	74.08
Balanced Random Forest	84.60	76.28
Stacked Classifier	87.60	77.45
Support Vector Machine (RBF)	88.90	78.80
Proposed Approach	90.08	80.03

RF+MLP scored 87.45% and 76.45% on the SATDAP and ICFES datasets, respectively.

C. Competitive Comparison

In our comparative analysis demonstrated in Table IV, our proposed model outperformed established methods like Balanced Random Forest, Decision Tree, Stacked Classifiers, and SVM [14] [15]. This superior performance can be traced back to the effective combination of Principal Component Analysis (PCA) and Random Forest (RF) for feature engineering, followed by classification with a Multi-Layer Perceptron (MLP).

PCA streamlines the data by capturing the most significant variances, reducing dimensionality and noise, while RF prioritizes the most informative features. This process results in a highly optimized feature set for the MLP, which is finely tuned to exploit these features for enhanced prediction accuracy. This strategic integration allows our model to more effectively capture complex patterns and relationships in the data, resulting in more incredible predictive performance and generalization than traditional methods.

IV. Conclusion

Concluding our research, the integrated approach of combining Principal Component Analysis (PCA) and Random Forest (RF) for feature engineering, followed by the application of a finely tuned Multi-Layer Perceptron (MLP) for classification, is proving to be a superior method in the realm of predicting student dropouts in higher education settings. Demonstrating remarkable predictive accuracy with scores of 90.08% on the SATDAP dataset and 80.03% on the ICFES dataset, our model distinctly outperforms traditional methods such as Balanced Random Forest, Decision Tree, Stacked Classifiers, and SVM. This enhanced performance is attributable to the strategic refinement of the feature space through PCA and RF, effectively reducing noise and highlighting crucial predictive features. This optimized feature set enables the MLP to effectively capture complex nonlinear patterns, significantly boosting the model's ability to generalize across diverse educational data. The robustness and adaptability of our approach establish it as a potent tool in the academic sector, offering a reliable means to identify at-risk students early and implement timely interventions.

References

[1] J. McFarland, J. Cui, J. Holmes, and X. Wang, "Trends in high school dropout and completion rates in the united states: 2019. compendium report. nces 2020-117." *National Center for Education Statistics*, 2020.

[2] A. Tayebi, J. Gómez, and C. Delgado, "Analysis on the lack of motivation and dropout in engineering students in spain," *IEEE Access*, vol. 9, pp. 66 253–66 265, 2021.

[3] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 10, no. 3, p. e1355, 2020.

[4] K. P. S. Attwal and A. S. Dhiman, "Exploring data mining tool-weka and using weka to build and evaluate predictive models," *Advances and Applications in Mathematical Sciences*, vol. 19, no. 6, pp. 451–469, 2020.

[5] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022.

[6] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471–1483, 2020.

[7] A. Alam and A. Mohanty, "Predicting students' performance employing educational data mining techniques, machine learning, and learning analytics," in *International Conference on Communication, Networks and Computing*. Springer, 2022, pp. 166–177.

[8] M. V. Martins, D. Tolledo, J. Machado, L. M. Baptista, and V. Realinho, "Early prediction of student's performance in higher education: a case study," in *Trends and Applications in Information Systems and Technologies: Volume 1 9*. Springer, 2021, pp. 166–175.

[9] E. Delahoz-Dominguez, R. Zuluaga, and T. Fontalvo-Herrera, "Dataset of academic performance evolution for engineering students," *Data in brief*, vol. 30, p. 105537, 2020.

[10] E. M. Senan, I. Abunadi, M. E. Jadhav, and S. M. Fati, "Score and correlation coefficient-based feature selection for predicting heart failure diagnosis by using machine learning algorithms," *Computational and Mathematical Methods in Medicine*, vol. 2021, no. 1, p. 8500314, 2021.

[11] M. Li, H. Wang, L. Yang, Y. Liang, Z. Shang, and H. Wan, "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction," *Expert Systems with Applications*, vol. 150, p. 113277, 2020.

[12] S. Ranjeeth, T. Latchoumi, and P. V. Paul, "Optimal stochastic gradient descent with multilayer perceptron based student's academic performance prediction model," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 14, no. 6, pp. 1728–1741, 2021.

[13] K. Nahar, B. I. Shova, T. Ria, H. B. Rashid, and A. S. Islam, "Mining educational data to predict students performance: A comparative study of data mining techniques," *Education and Information Technologies*, vol. 26, no. 5, pp. 6051–6067, 2021.

[14] M. V. Martins, L. Baptista, J. Machado, and V. Realinho, "Multi-class phased prediction of academic performance and dropout in higher education," *Applied Sciences*, vol. 13, no. 8, p. 4702, 2023.

[15] T. R. Noviandy, Z. Zahriah, E. Yandri, Z. Jalil, M. Yusuf, N. I. S. M. Yusof, A. Lala, and R. Idroes, "Machine learning for early detection of dropout risks and academic excellence: A stacked classifier approach," *Journal of Educational Management and Learning*, vol. 2, no. 1, pp. 28–34, 2024.