I.T.C

Content list available at ITC

Techno-Science Research Journal

Journal Homepage: http://techno-srj.itc.edu.kh/

Techno-Science Research Journal

Undergraduate Student Dropout Prediction with Class Balancing Techniques

Lyheng Sam [1*], Sokkhey Phauk [1], Dona Valy [2]

[1]*Department of Applied Mathematics and Statistics, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia*

[2]*Department of Information and Communication Engineering, Institute of Technology of Cambodia, Russian Federation Blvd., P.O Box 86, Phnom Penh, Cambodia*

Received:12 July 2024; Revised: 01 August 2024; Accepted: 11 September 2024; Available online: 30 August 2025

Abstract: *This study investigates how machine learning (ML) and deep learning (DL) techniques can be used to predict student dropouts, which is a major issue for higher education institutions. Using a dataset from Kaggle titled "Predict students' dropout and academic success," we analyzed data from 4424 students across 17 undergraduate programs. We used 35 different attributes for each student's profile, which gave us a strong basis for our predictive modeling. To handle the class imbalance in the dataset, we used three methods: oversampling, undersampling, and the Synthetic Minority Oversampling Technique (SMOTE). We tested several ML and DL models, such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Gaussian Naive Bayes, AdaBoost, XGBoost, 1D Convolutional Neural Network (CNN), Multiple Layer Perceptron (MLP), and Deep Belief Network (DBN). We evaluated these models based on accuracy, precision, recall, and F1 score. The Multiple Layer Perceptron (MLP) stood out, achieving the highest scores for accuracy 98.6%, precision 98%, recall 98%, and F1-score 98% with the oversampled dataset. This shows its great capability in managing complex data. The 1D Convolutional Neural Network (1D CNN) also performed well, particularly in recall and F1-score, with scores of 91.5% and 88.5%, respectively, with the original dataset. It maintained a strong recall of 91.4% and an F1-score of 87.7% with the undersampled dataset, and a recall of 89.2% and an F1-score of 88.1% with the SMOTE dataset, proving its effectiveness in identifying dropouts under various conditions. These results underscore the effectiveness of resampling techniques in enhancing model accuracy and the critical role of precise academic indicators in predicting student outcomes. Our study's contribution extends to informing educational strategies with practical evidence of the efficiency of ML and DL models supported by innovative resampling methods. By recording the exceptional performance of both the MLP and 1D CNN models, the research emphasizes the transformative potential of applying advanced analytical techniques to foster student retention and academic success. The insights derived from this work could lead to actionable, data-informed interventions tailored to support students at risk of dropout, thereby improving retention rates and shaping the future landscape of educational analytics.*

Keywords: Predictive Modeling; Student Dropout; Machine Learning; Deep Learning; Class Balancing

1. INTRODUCTION

High dropout rates are a serious problem, leading to millions of dollars in financial losses for schools due to lost tuition and increased recruitment costs. Students who leave their studies often find themselves with fewer career opportunities and lower earnings over their lifetimes. According to the National Center for Education Statistics, the dropout rate is 40% for undergraduates in the United States, 30% in European countries, and varies across Asia and Africa [1]. These numbers clearly show the need for predictive models that can identify students who are at risk of dropping out. With these models, schools can create targeted interventions to help keep students on track and support their success.

Our research applied machine learning (ML) and deep learning (DL) algorithms to predict student dropout risks. These modern techniques provide educational institutions with valuable insights to improve their student retention strategies. Compared to traditional methods, ML and DL are better at handling the complex and diverse data found in student records [2].

* Corresponding author: Lyheng Sam
E-mail: samlyheng@gmail.com; Tel: +855 10-567-274

While previous studies have identified factors that lead to student attrition, there is still a gap in effectively using ML and DL for datasets with class imbalances. Our study addresses this by employing resampling techniques to balance the data [3]. This approach enhances the accuracy of our predictive models and ensures a reliable analysis of dropout indicators [4]. This research seeks not only to validate the effectiveness of ML and DL models in predicting student dropouts but also to establish a precedent for the application of resampling in educational data analysis. Our findings have the potential to guide interventions by educational institutions, reduce dropout rates, and signal a shift toward a more data-informed educational framework.

Table 1 summarizes the comparative analysis of model performance, highlighting different features, datasets, algorithms, and accuracy metrics from previous studies, underscoring the need for our innovative approach [5].

Table 1. Comparative Analysis of Model Performance

Reference	Feature	Dataset	Algorithms	Accuracy	
				Min	Max
Asif et al. (2017) [6]	Academic performance	210	DT, 1-NN, NB, NN, RF	NN (62.50%)	NB (83.65%)
Cruz-Jesus et al. (2020) [7]	Academic performance	1854	ANN, DT, ET, RF, SVM, kNN,LR	LR (81.1%)	SVM (51.2%)
Hoffait and Schyns (2017)[8]	Predicting students' performance	2244	RF, LR, ANN	ANN (70.4%)	RF (90%)
Ahmad (2018) [9]	Identify students at risk	300	MPNN		95%
Musso et al., (2020) [10]	Academic grade	655	ANN	60.5%	80.7%
Waheed et al., (2020) [11]	Academic grade,	32593	ANN, SVM, LR	84%	93%
Xu et al. (2019) [12]	Predicting students' performance	4000	DT, NN, SVM	71%	76%
Bernacki et al. (2020) [13]	Predict achievement	337	LR, NB, J-48 DT, J-Rip DT	J-48 (53.71%)	LR (67.36%)
Burgos et al. (2018) [14]	Drop out of a course	100	SVM, FFNN, PESFAM, LOGIT_Act	SVM (62.50)	LOGIT_ Act(90%)

2. METHODOLOGY

2.1 Dataset

The analysis is based on a large dataset gathered from Kaggle called "Predicting students' dropout and academic success." The dataset comprises information from 4424 students enrolled in 17 undergraduate programs. Each student profile contains 35 features, including demographic information (age, gender, nationality), socioeconomic status (family income, parental education level, parental career), academic achievement measures (grades, attendance rates, courses, and curricular units), and target variables (dropout and graduate). The dataset is imbalanced, with 1421 dropout cases and 2209 graduate cases. The class imbalance is a major challenge across research studies in educational data mining, as it can significantly affect predictive modeling. Models might become biased toward the majority class, leading to poor sensitivity in detecting dropout cases, which is critical for early intervention strategies.

A first examination found that the dataset is imbalanced, with much fewer dropout cases than successful continuations. This is a typical issue in educational datasets, and it might bias predictive models toward the majority class, resulting in an underestimation of dropout probability [15]. To ensure the dataset's reliability and efficiency in training predictive models, thorough cleaning methods were applied. The procedures involved resolving missing values, normalizing continuous attributes, and encoding categorical variables. Each record was thoroughly analyzed for completeness and consistency to ensure that the dataset was reliable and appropriate for testing various predictive algorithms. Additionally, we conducted exploratory data analysis to identify trends and correlations in the dataset, which is essential for developing precise predictive models to improve student retention and success. The study also addressed class imbalance by utilizing resampling techniques such as oversampling, undersampling, and SMOTE to balance the dataset and enhance model accuracy [16]. Table 2 summarizes the dataset features, which categorizes the attributes into demographic, socioeconomic, macroeconomic, and academic data.

Table 2. Comparative Analysis of Model Performance

Attribute Class	Attribute	Type
Demographic Data	Martial Status	Numeric/Discrete
	Nationality	Numeric/Discrete
	Displaced	Numeric/Binary
	Gender	Numeric/Binary
	Age	Numeric/Discrete
	International	Numeric/Binary
Socioeconomic Data	Mother’s Qualification	Numeric/Discrete
	Father’s Qualification	Numeric/Discrete
	Mother’s Occupation	Numeric/Discrete
	Father’s Occupation	Numeric/Discrete
	Education special needs	Numeric/Binary
	Debtor	Numeric/Binary
	Tuition fee up to date	Numeric/Binary
	Scholarship Holder	Numeric/Binary
	Umemployment rate	Numeric/Continuous
Macroeconomic Data	Inflation Rate	Numeric/Continuous
	GDP	Numeric/Continuous
	Application Mode	Numeric/Discrete
Academic data at enrollment	Application Order	Numeric/Ordinal
	Course	Numeric/Discrete
	Attendance	Numeric/Binary
	Previous Qulification	Numeric/Discrete
Academic Data at the end 1st and 2nd of semester	Curricular unit (Credited)	Numeric/Discrete
	Curricular unit (enrolled)	Numeric/Discrete
	Curricular unit (evaluation)	Numeric/Discrete
	Curricular unit (approved)	Numeric/Discrete
	Curricular unit (grade)	Numeric/Continuous
	Curricular unit (no evaluation)	Numeric/Discrete
Target	Target	Categorical

2.2. Resampled Methods

To address the dataset's detected class imbalance, we used three resampling strategies: oversampling, undersampling, and the Synthetic Minority Over-sampling Technique (SMOTE). Each strategy was chosen based on its capacity to improve class balance, hence improving predictive model performance and generalizability.

Oversampling: Using this method, we expanded the size of the minority class ("Dropped Out") by randomly reproducing instances until the number of dropout cases equaled the number of continuing students. This strategy helps to prevent data loss, which is a danger linked with undersampling [4]. Oversampling ensures that the model is not biased towards the majority class and can learn from the intricacies found in the minority class.

Undersampling: In contrast, undersampling includes removing instances at random from the majority class ("graduate") in order to equal the number of dropout instances. When the dataset is large enough to maintain important information after reduction, this method is known for its computational efficacy and efficiency [16]. Undersampling can speed up model training and reduce the possibility of overfitting. However, it can also result in the loss of important data.

Synthetic Minority Over-sampling Technique (SMOTE): SMOTE generates new synthetic instances of the minority class by interpolating between existing ones [17]. This more nuanced representation of the minority class aids models in learning a broader range of features associated with dropouts, which can lead to improved detection of at-risk students. SMOTE has the advantage of creating more diverse synthetic samples, which can enhance the model's ability to generalize to new, unseen data.

By resolving the class imbalance using these resampling strategies, we saw considerable increases in the performance of our prediction models. Models trained on resampled datasets have higher accuracy, precision, recall, and F1 scores than those trained on the original imbalanced dataset. This suggests that the models performed better in recognizing both dropout and continuation cases, resulting in more trustworthy and generalizable predictions.

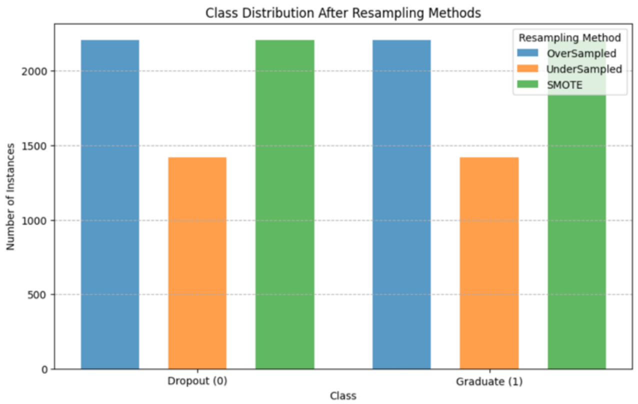

Fig 2. Comparative Analysis of Model Performance

The class distributions after applying the resampling methods are illustrated in Fig. 1, which shows the number of instances for each class under different resampling techniques.

2.3 Model Selection

The selection of predictive models is an important phase in the research process since it affects the performance and interpretability of the results. In our research, we purposefully selected a varied range of algorithms, including both traditional machine-learning techniques and new deep-learning methods. This decision is supported by the research, which shows that depending on the characteristics of the dataset and the specifics of the situation at hand, different approaches may perform differently.

We included Logistic Regression (LR) to improve interpretability and simplicity. For our experiments, LR was configured with a maximum of 1000 iterations .

Decision Trees (DT) and their ensemble counterpart, Random Forest (RF), have been incorporated to capture the non-linear correlations and interactions among features, [18]. Decision Trees were used with a maximum depth of 10, minimum samples split of 5, and minimum samples leaf of 2, while Random Forests utilized 100 estimators, a maximum depth of 20, and minimum samples split of 2.

The K-Nearest Neighbors (KNN) technique, a non-parametric method, is used to utilize the localization of data points in the feature space [19]. KNN was configured with 9 neighbors for the original dataset and 3 neighbors for oversampled and SMOTE datasets.

The effectiveness of Gaussian Naive Bayes (GNB), a model based on the independence of features in our potentially correlated dataset, is also investigated [20]. GNB was used with a variance smoothing parameter of 1e-08 for original, undersampled, and SMOTE datasets.

Boosting algorithms like AdaBoost and XGBoost are chosen for their excellence in the sequential improvement of weak learners' errors and their outstanding prediction performance track record [20], [21]. AdaBoost was configured with 50 estimators and a learning rate of 0.5 for original and undersampled datasets, while for SMOTE, it used 100 estimators. XGBoost was used with a learning rate of 0.05 for original and undersampled datasets and 0.1 for SMOTE, with a maximum depth of 6 for original and 8 for oversampled, undersampled, and SMOTE datasets.

For deep learning, a 1D Convolutional Neural Network (1D CNN) is expected to perform well with sequential or time-series data because of its feature extraction capabilities [22], [23]. The 1D CNN model included 2 convolutional layers with 64 and 32 filters respectively, kernel size of 3, pooling size of 2, and was trained with 100 epochs, batch size of 128, ReLU activation functions, and Adam optimizer.

The Multiple Layer Perceptron (MLP), a fundamental yet adaptable neural network model, was chosen for its architecture's capacity to describe complex functions [9], [13]. The MLP model included 2 hidden layers with 64 and 32 units respectively, ReLU activation function, 100 epochs, and batch size of 32.

Finally, Deep Belief Networks (DBN), with their multi-layered latent variables, have the potential to discover complicated data patterns [24]. The DBN model included 3 hidden layers with 128, 256, and 512 units respectively, learning rates of 0.01, 0.05, and 0.1, and a maximum of 10000 iterations. Collectively, the models range from the simplicity of logistic regression to the complexities of deep neural networks, allowing us to examine the balance of interpretability, computing load, and predictive power. This diverse method ensures a full comparative examination, which serves the main purpose of analyzing the predictive task from several analytical perspectives.

2.4. Evaluation Metrics

The efficacy of the predictive models was assessed using a set of indicators that provide a comprehensive perspective of performance. The metrics used to account for a variety of features of prediction quality, including accuracy, the balance of precision and recall, and the harmonic mean of the two. These measures, Accuracy, Precision, Recall, and F1-Score, were chosen because they are widely accepted in classification tasks and are relevant to the context of forecasting student dropouts, where class imbalance is a major concern [25].

To evaluate the models, we employed a train-test split method. The dataset was split into training and testing sets with a ratio of 80% for training and 20% for testing, resulting in 3540 training samples and 885 testing samples. The data was split randomly, ensuring that the representativeness of the class distribution was maintained in both the training and testing sets. This method ensures that both sets have a similar proportion of dropout and graduate cases, which is crucial for evaluating model performance in a realistic manner .

To address the class imbalance in the training set, we utilized various resampling techniques, including Random OverSampling, Random UnderSampling, and SMOTE [8], [16]. These methods help in balancing the dataset and enhancing model accuracy by preventing the predictive models from being biased towards the majority class. By applying these techniques, we aimed to improve the model's ability to accurately predict dropout cases, which are critical for early intervention strategies.

The test set was derived from the original dataset before applying any class balancing techniques. This approach ensures that the evaluation results reflect realistic performance, as using a balanced test set might yield high accuracy but may not represent real-world conditions accurately.

Accuracy represents the proportion of correctly predicted instances out of the total instances [25]. The formula is:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + FP} \qquad \text{(Eq.1)}$$

Recall, also known as the true positive rate, indicates how well the model identifies positive instances [25]. The formula is:

$$Recall = \frac{TP}{TP + FN} \qquad \text{(Eq.2)}$$

Precision measures the correctness of positive predictions by the model [25]. The formula is:

$$Precision = \frac{TP}{TP + FP} \qquad \text{(Eq.3)}$$

The F1-Score is the harmonic mean of precision an recall, providing a balance between the two [25]. The formula is:

$$F1\ Score = \frac{2 \times TP}{2 \times TP + FP + FN} \qquad \text{(Eq.4)}$$

Table 3 illustrates the confusion matrix shows the relationship between actual and predicted classifications, including true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) [25].

Table 3. Comparative Analysis of Model Performance

		Actual	
		Positive	**Negative**
Predicted	**Positive**	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

3. RESULTS AND DISCUSSION

The graphs in Figs. 2 through 5 illustrate the accuracy, precision, recall, and F1-score of various machine learning and deep learning models using a range of resampling strategies. They underscore the crucial role of selecting the right resampling technique and model based on the unique performance metric of interest.

The analysis of various machine learning models using different resampling methods reveals key insights into model performance across metrics like accuracy, precision, recall, and F1-score. The Multiple Layer Perceptron (MLP) consistently performs the best, particularly with the oversampled dataset, achieving the highest scores across all metrics [25]. The 1D Convolutional Neural Network (CNN) also demonstrates strong and stable performance across different resampling methods, maintaining high recall and F1-scores [26].

Table 5 shows that the Multiple Layer Perceptron (MLP) performed consistently better than the other models on all datasets, performing especially well on the oversampled dataset. With an accuracy of 0.986, precision of 0.980, recall of 0.980, and F1-Score of 0.980, it obtained the highest metrics. This suggests that MLP gains a great deal from the oversampling method, as evidenced by the noteworthy improvements in all assessed measures. On the other hand, different datasets showed variations in the effectiveness of alternative models. As can be seen in Tables 4, 6, and 7, some models improved when specific sampling strategies were used, whereas other models did not consistently demonstrate benefits.

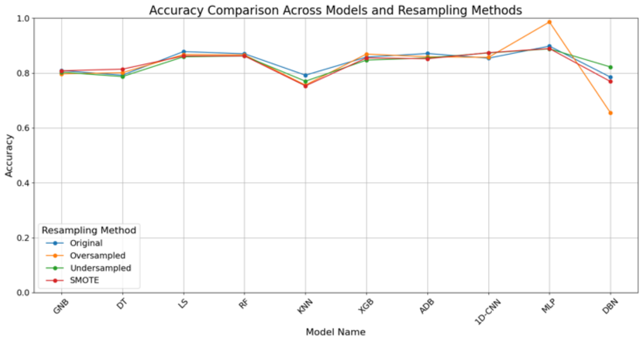

Fig 2. Accuracy by Resampling

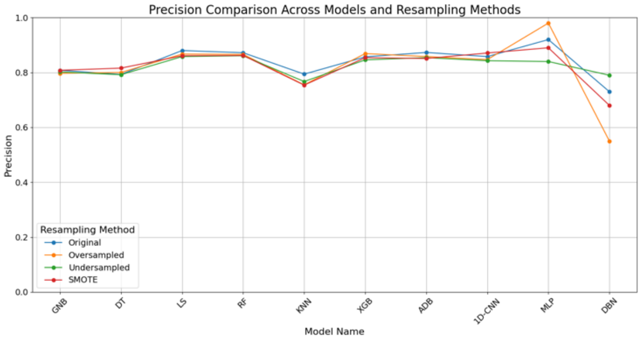

Fig 3. Precision by Resampling

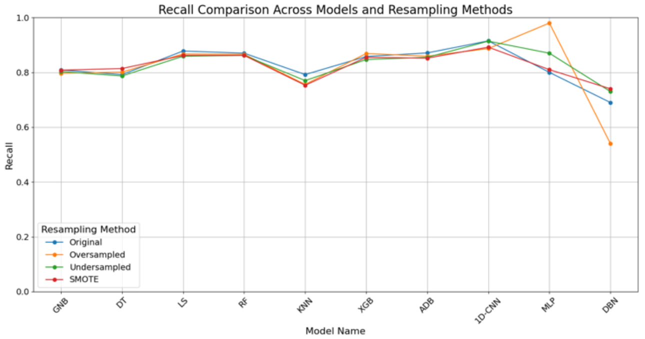

Fig 4. Recall by Resampling

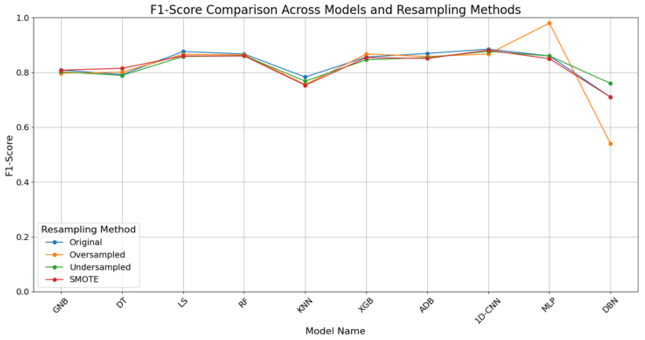

Fig 5. Score by Resampling

Table 4. Dropout Results in 'Original' Dataset:

Model Name	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.792	0.793	0.792	0.792
Random Forest	0.870	0.872	0.870	0.867
Logistic Regression	0.878	0.880	0.878	0.876
K-Nearest Neighbor	0.792	0.794	0.792	0.783
Gussian Naïve Bayes	0.809	0.808	0.809	0.809
AdaBoost	0.871	0.873	0.871	0.869
XGBoost	0.858	0.857	0.858	0.856
1D CNN	0.854	0.858	**0.915**	**0.885**
Multiple Layer Perceptron	**0.898**	**0.920**	0.800	0.860
Deep Belief Network	0.785	0.730	0.690	0.710

An important finding highlighted in the feature importance chart (Fig. 6) is the significance of the *'Tuition fees up to date'* feature. As depicted in the accompanying bar chart, this attribute holds the highest importance among all features when predicting student dropouts. The importance of the feature highlights how crucial financial stability is in keeping students in school [27].

Additionally, the *'Curricular units 2nd semester (grade)'* feature is also notably significant, suggesting that academic performance in the second semester is a key indicator of a student's likelihood to drop out. This can provide educational institutions with actionable insights into identifying students at risk based on their financial and academic status.

Furthermore, being a *'Scholarship holder'* and whether a student is *'International'* are also important factors, indicating that both financial support and the challenges faced by international students play crucial roles in predicting dropout rates [28]. These insights can help institutions focus their support efforts more effectively to improve student retention.

Table 5. Dropout Results in 'Oversampled' dataset

Model Name	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.801	0.800	0.801	0.801
Random Forest	0.866	0.866	0.866	0.864
Logistic Regression	0.867	0.867	0.867	0.866
K-Nearest Neighbor	0.756	0.755	0.756	0.755
Gussian Naïve Bayes	0.797	0.797	0.797	0.797
AdaBoost	0.859	0.858	0.859	0.858
XGBoost	0.869	0.869	0.869	0.867
1D CNN	0.857	0.847	0.887	0.867
Multiple Layer Perceptron	**0.986**	**0.980**	**0.980**	**0.980**
Deep Belief Network	0.655	0.550	0.540	0.540

Table 6. Dropout Results in 'Undersampled' dataset

Model Name	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.787	0.792	0.787	0.789
Random Forest	0.860	0.840	0.840	0.840
Logistic Regression	0.859	0.858	0.859	0.858
K-Nearest Neighbor	0.770	0.767	0.770	0.768
Gussian Naïve Bayes	0.803	0.802	0.803	0.802
AdaBoost	0.855	0.854	0.855	0.854
XGBoost	0.847	0.846	0.847	0.846
1D CNN	0.873	**0.843**	**0.914**	**0.877**
Multiple Layer Perceptron	**0.888**	0.840	0.870	0.860
Deep Belief Network	0.822	0.790	0.730	0.760

Table 7. Dropout Results in ‘SMOTE’ dataset:

Model Name	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.814	0.816	0.814	0.815
Random Forest	0.862	0.862	0.862	0.860
Logistic Regression	0.862	0.861	0.862	0.860
K-Nearest Neighbor	0.753	0.754	0.753	0.753
Gussian Naïve Bayes	0.808	0.808	0.808	0.808
AdaBoost	0.852	0.851	0.852	0.851
XGBoost	0.855	0.854	0.855	0.854
1D CNN	0.874	0.871	**0.892**	**0.881**
Multiple Layer Perceptron	**0.889**	**0.890**	0.810	0.850
Deep Belief Network	0.769	0.680	0.740	0.710

Additionally, the role of socio-economic factors, including access to educational resources, cannot be overlooked. Students with limited access to textbooks and online materials often face additional challenges, impacting their ability to succeed academically [27]. This highlights the need for institutions to ensure all students have equitable access to necessary resources. Moreover, the impact of extracurricular activities on student retention is significant. Participation in sports, clubs, and other non-academic pursuits fosters a sense of belonging and community, which is essential for student motivation and persistence. Encouraging involvement in such activities can positively influence students' academic journeys and retention rates.

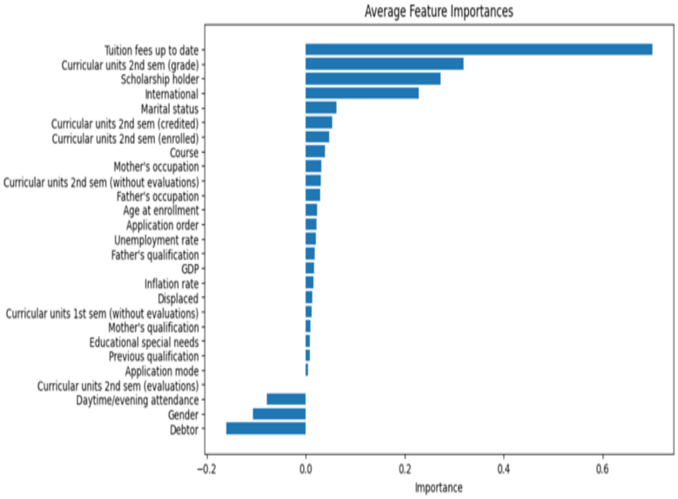

Fig 6. Important Feature

4. CONCLUSIONS

The study makes significant strides in predictive analytics in higher education, employing machine learning (ML) and deep learning (DL) approaches to assess student dropout rates. The research confirms that resampling strategies such as oversampling, undersampling, and SMOTE efficiently correct class imbalances, thereby increasing predictive model accuracy. The Multiple Layer Perceptron (MLP), a deep learning model, consistently performs well, particularly on the oversampled dataset. However, the 1D Convolutional Neural Network (CNN) consistently outperforms the MLP in terms of recall and F1-Score, except for the oversampled dataset. This highlights the efficacy of 1D CNN in managing imbalanced datasets, notably in terms of recall and F1-Score.

The novelty of this research lies in its comprehensive integration of both machine learning (ML) and deep learning (DL) techniques to predict student dropouts, coupled with the innovative application of class balancing methods, including oversampling, undersampling, and the Synthetic Minority Oversampling Technique (SMOTE). This dual approach not only enables a detailed comparison of various predictive models but also significantly enhances predictive accuracy and reliability. By addressing the common issue of class imbalance, the study sets a new benchmark in educational data analysis. The methodological framework established in this research—encompassing extensive data cleaning, feature engineering, and advanced resampling techniques—provides a robust template for future studies.

The findings have practical implications for educational institutions, administrators, educators, and support staff, all of whom play important roles in reducing dropout rates. By implementing predictive models, institutions can identify at-risk students and execute targeted early intervention programs such as academic guidance, tutoring, financial aid, and flexible payment options. The integration of predictive models into student information systems enables real-time monitoring and interventions, a task that requires the collective effort of all involved.

Future studies might explore the use of predictive models in a range of educational settings and include additional data pieces, such as student engagement indicators from learning management systems, to increase forecast accuracy. Continuous research can track the long-term effects of predictive interventions on student retention and achievement levels. Furthermore, the research delivers actionable insights for educational institutions to develop targeted interventions, ultimately improving student retention. This scalable model offers a practical tool for real-time monitoring and support, with broad applicability across diverse educational settings.

REFERENCES

[1] National Center for Education Statistics, "Fast facts," U.S. Department of Education, Washington, DC, USA. [Online]. Available: https://nces.ed.gov/fastfacts/. [Accessed: Jun. 10, 2024].

[2] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," Appl. Sci., vol. 10, no. 3, p. 1042, Feb. 2020.

[3] F. Dalipi, A. S. Imran, and Z. Kastrati, "MOOC dropout prediction using machine learning techniques: Review and research challenges," in 2018 IEEE Global Engineering Education Conf. (EDUCON), Apr. 2018, pp. 1007–1014.

[4] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[5] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," Data, vol. 7, no. 146, pp. 1–17, Jul. 2022.

[6] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, Oct. 2017.

[7] S. Cruz-Jesus, et al., "Predicting academic performance and student success in higher education: A data mining approach," Appl. Sci., vol. 10, no. 3, p. 943, Feb. 2020.

[8] A. Hoffait and B. Schyns, "Early detection of university students with potential difficulties," Decis. Support Syst., vol. 101, pp. 1–11, Nov. 2017.

[9] Z. Ahmad and E. Shahzadi, "Prediction of students' academic performance using artificial neural network," Bull. Educ. Res., vol. 40, no. 3, pp. 157–164, Sep. 2018.

[10] M. F. Musso, C. F. R. Hernández, and E. C. Cascallar, "Predicting key educational outcomes in academic trajectories: A machine-learning approach," High. Educ., vol. 80, no. 5, pp. 875–894, Nov. 2020.

[11] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," Comput. Hum. Behav., vol. 104, p. 106189, Mar. 2020.

[12] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," Comput. Hum. Behav., vol. 98, pp. 166–173, Oct. 2019.

[13] M. L. Bernacki, M. M. Chavez, and P. M. Uesbeck, "Predicting achievement and providing support before STEM majors begin to fail," Comput. Educ., vol. 158, p. 103999, Jan. 2020.

[14] C. Burgos, M. L. Campanario, J. A. De Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," Comput. Electr. Eng., vol. 66, pp. 541–556, Feb. 2018.

[15] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," Data, vol. 7, no. 146, pp. 1–17, Jul. 2022. (Kaggle dataset)

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[17] D. Delen, "Predicting student attrition with data mining methods," J. Coll. Stud. Retent. Res. Theory Pract., vol. 13, no. 1, pp. 17–35, May 2011.

[18] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[19] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[20] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, Aug. 1997.

[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[23] Z. Song, S.-H. Sung, D.-M. Park, and B.-K. Park, "All-year dropout prediction modeling and analysis for university students," Appl. Sci., vol. 13, no. 1143, pp. 1–14, Jan. 2023.

[24] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.

[25] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," Comput. Hum. Behav., vol. 47, pp. 168–181, Jun. 2015.

[26] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," Comput. Hum. Behav., vol. 104, p. 106189, Mar. 2020.

[27] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," J. Bus. Res., vol. 94, pp. 335–343, Mar. 2019.

[28] K. A. Oqaidi, S. Aouhassi, and K. Mansouri, "Towards a students' dropout prediction model in higher education institutions using machine learning algorithms," Int. J. Emerg. Technol. Learn., vol. 17, no. 18, pp. 103–117, Sep. 2022.