A Research on Students' Academic Success Using Machine Learning

Faishal Monir
Dept of
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
faishal.monir@g.bracu.ac.bd

Umma Salma Mim
Dept of
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
umma.salma.mim@g.bracu.ac.bd

***Abstract*—In the present world, the number of dropout rate is currently alarming and poses significant challenges to students, institutions, and societies. At the tertiary or higher education level, a student can drop out of higher education because of various reasons. Some common reasons can be financial issues, challenges in life, surroundings, migration, etc. If the cause of dropout cannot be identified at an early stage, it can cause serious problems not only to the institution, society, or the individual's life but also to the economy or the future of that nation. So, in order to find the root cause, we examined various datasets and conducted this research on a dataset of students at higher education levels in Portugal with a vision to identify the students in danger of dropping out and minimize the number of dropouts. In this research we have used six different machine learning models: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Ada Boost, Kth-Nearest Neighbor, Neural Network (MLP Based) and used a technique to mix some models (Ensemble Method) and predict the outcome in a separate implementation. In our research, we used several types of evaluation metrics, but the highest accuracy and F1 score is achieved in the Random Forest Model. The remaining models also tried to outperform the best model, but they lacked behind by 1-3% from the RF model.**

***Index Terms*— Student dropout prediction, Tertiary education, Machine learning, Models, Random Forest, Higher Education analytics, Ensemble Learning, Classification modelling, Predictive modeling.**

I. Introduction

In recent times, many higher education institutions are concerned about the academic status of their students. Not long ago, humankind faced the deadly clutches of COVID-19, and during that period, educational methods came to a standstill. Various researchers and institutions inferred and published various studies about online education and multimodal education systems. But only a few Higher Educational Institutions (HEIs) conducted studies about the dropout rates at this academic level. Experts now believe that student dropout is the most significant and complex issue that needs attention in our educational system [1]. In the context of our country, each year, 1.2 million students enroll in higher education, but only 0.25 successfully graduate, resulting in 0.95 million students who are either pursuing their education or have dropped out [2]. This creates various risks for the development of a nation. Researchers found out that dropout rates can directly influence unemployment and poverty rates [3]. Till now, there has been much research that has been conducted on multimodal learning, especially during or after the pandemic. But a few research studies are done identifying the causes and factors of tertiary education dropout. In this factor, Educational Data Mining (EDM) plays a prime factor where it deals with various educational methods and techniques that explore data gained from various educational contexts[4].

Unfortunately, in Bangladesh, very little research like this has been conducted.

Although the reasons for dropping out can be various and hard to identify, we aim that this research report will be helpful in identifying the in-risk students of dropping out with the help of machine learning (ML). We have created this research to very precisely leverage computational data using statistical Machine learning models to analyze the educational, social, and other factors in order to identify their academic status. The identification can be more precise and reliable thanks to the help of these diverse algorithms to save a student's academic future.

II. Literature Review

Numerous studies around the world have been conducted to identify the factors contributing to student dropout. A common approach in these studies is to analyze historical academic data. Research by Rabelo and Zárate [4] suggests that the first step in addressing this problem is domain understanding and feature selection. In their study, the researchers implemented three models—logistic regression, decision tree, and neural network. Using random sampling and bootstrapping with 5-fold cross-validation, they achieved a maximum average accuracy rate of 89%. However, the study did not classify anything about academic success; rather, it is specialized in identifying the dropouts with a success rate of 98.1%.

To address this issue more deeply in another research, they took a different approach. They preprocessed the data and used an optimization algorithm named Learning Vector Quantization (LVQ) [5]. This is basically a feature selection process where the algorithm chooses the most related and influential predictors initially. A 10-fold cross-validation technique was used with popular statistical models such as Knn, DT (using cart), and a naive Bayesian classifier. Later, the models were stacked to further enhance the accuracy by using the ensemble method. The results showed a drastic turn this time; here, the average accuracy was 98.29%, and the ensemble method worked in correctly identifying the target values.

Previous research mostly used structured methods to evaluate students' academic success. In this study, however, a more straightforward and hands-on approach was taken, which surprisingly led to even better accuracy. In this research, a deep learning model was used with popular statistical machine learning models[6]. Other than this, a number of special models, such as XG-Boost, AdaBoost, and the combination of these, are used as ensemble models. This time the accuracy is shocking as the model outperformed with a score of 99%, and random forest was used in this scoring classification. Being this a classification problem, the object-based classifier outperformed others.

Now, if we look at a mining approach where the researcher used a mining approach to collect data. Researchers Dake and Andoh followed a different approach known as modified crisp-dm methodology, where data mining is done to understand the data preparation and selection of features [7]. In feature selection, they followed a method named chi-square, which is very effective to ranking and selecting the dominant factors. Finally, a 10 and a second pass 5-fold cross validation technique was used that resulted in an identification score of 70.98% where MLP and SVM outperformed others. Using the 5-fold cross validation, furthermore validated the score by 69.74%, which is significantly near the original predicted score.

Lastly, in this research, a notable model was used, named Feed Forward Neural Network (FNN), where the outputs are fed into several metaclassifiers[8]. This approach helps widely to stop overfitting of data in the learning process. To implement this similarly, a 10-fold cross-validation technique was used on the data of some students in Slovakia. These types of neural networks highly rely on forward and backward propagation of data, where the iterations can improve the overall learning if the dataset is evenly distributed. If the model is used with other object-based classifiers in an ensemble method, the results can be closer to perfection. As a characteristic, the research resulted in the model's overall accuracy being 96.67%.

Analyzing these papers, we concluded with the statement that not much academic research is done on the Asian or sub-Asian continents due to a lack of infrastructure or datasets used in EDM. These studies failed to identify the dropout students of the Eastern region, as these studies are mostly done in the western countries of the world. Our research aims to build a foundational structure for these researches in this region in the future to successfully identify these at-risk students.

III. METHODOLOGY

Dataset Description:

The dataset we used here is from an open-source platform named Kaggle. The dataset comes from a student research project that aims to reduce academic dropout among higher education students in Portugal under the SAT-DAP program [9]. The dataset contains data about 76 thousand students, with an additional testing dataset for effective classification testing. These data were collected in the time span of the 2021-2022 academic year in Portugal.

In the dataset, there are various factors such as a student's results, gender, if he/she is an international student or not, evening/daytime attendance, GDP, if he/she is displaced or not. In total, there are a number of 37 feature columns and a target column that indicates if a student will graduate, drop out, or enroll in another institute based on these correlational data. In summary, the dataset is a type of student mining dataset to analyze the core reasons for student dropouts by using machine learning.

IV . DATA PREPROCESSING AND MODEL DESCRIPTION

The dataset was based on real-life data, so it is a matter of fact that in our dataset, there were absolutely no null values. The amount of raw data was 76 thousand, and every instance represents a student's academic qualifications. The main problem here was that our database was biased and imbalanced. So, in order to fix this issue, we had to scale it using a standard scaler. Some of the lower-correlated columns were dropped as they almost contributed nothing to the learning process.

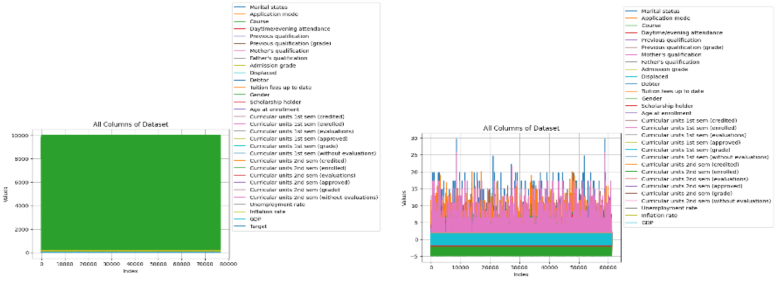

Fig.1 Imbalanced datasets side by side

Our target column was also feature engineered as it had numerical values. The feature column was analyzed, and the values were mapped to address the classification tasks. Later, a correlation was analyzed between them, ensuring the ML model had balanced and highly influential data in the learning phase.

A. The learning process:

For this research, a total of 7 machine learning models were used. The models worked independently in the majority of the cases, but we also used an ensemble model to analyze our data further. The models that we have used are :

1. *Decision Tree (DT):*

   Decision trees are built upon the partitioning approach and are followed top to bottom while doing so. It is highly effective in classifying object-based results[4]. The main tree is divided into subtrees and branches in the learning phase, and the data is driven into them. As a result, the tree can identify the class depending on the stopping criteria and make effective classification decisions following this approach. We used this approach to identify the non-linear factors in students dropping out.

2. *Random forest (RF):*

   Random forest is a tree-based algorithm, but it is more complex in nature. The algorithm works by clustering several smaller trees and boosts the overall performance. This approach to data selection lowers down the overfitting of data and ensures a higher performance with adequate robustness[11]. Complex and higher-dimensional data can easily be analyzed by this model. To further increase the metrics, an entropy-based loss function was chosen.

3. *Logistic Regression (LR):*

As our dataset is fairly complex in nature, we could not run or establish a linear relation between the classes. But the data were numerical in nature, so in order to classify them, we have used a logistic regression model. Logistic regression is a statistical approach widely used in statistical research to examine how one or more predictor variables are associated with a binary outcome variable, in our case, the dropping students [12]. We have used this in our research to identify the regression aspect of the target class predictors.

4. *Ada Boost:*

Ada Boots is vastly known as a boosting algorithm that can increase the predictive score by sequential analysis. The model is a stump-based model that expands several depths of stumps and works on various subsets of data to predict the target class. It works by clustering several weak classifiers that make a strong classifier by the ensemble method. The major characteristic is that it assigns the higher weights to the misclassified instances, allowing the subsequent classifiers to focus on harder cases[13]. We have used this in our dataset to comparatively analyze if a clustering-based algorithm can predict more accurately or not upon our target class.

5. *Kth nearest Neighbor (KNN):*

KNN is a classification object-based predictor that works in a slightly different way. The learning process heavily depends on majority voting. The majority voting is done by calculating the nearest nodes and evaluating the metrics using distance calculations such as Euclidean or Manhattan distance. The majority vote helps the model to predict effectively upon the final classification result. It is highly effective for classification tasks with well-defined data clusters[14]. As our tasks deal with this type of classification, Knn was used to identify the peer influence that our data might have from its neighbor instances.

6. *Neural Network:*

An input layer, one or more hidden layers, and an output layer are among the several layers of neurons that make up the Multilayer Perceptron (MLP), a fundamental architecture for artificial neural networks. A fully connected network is created when every neuron in one layer is connected to every other neuron in the layer above it. MLPs are trained using supervised learning methods, mainly the backpropagation algorithm, and employ nonlinear activation functions. MLPs are useful for tasks like pattern recognition, regression, and classification because of their ability to model complex, nonlinear relationships within data [15]. For having these robust characteristics, a neural model was used to identify the students at risk in our research.

7. *The Ensembled Model:*

The ensemble methods are a hybrid of the classical models, where one or more models are clustered and chosen upon majority voting. It not only improves overall predictions but also reduces variance and increases the robustness of the result. It is comparatively efficient by using a multimodal approach of bagging, boosting, and stacking the diverse models or the learners. Bagging works by training multiple models on random subsets of data and averaging their results. Boosting builds models sequentially, focusing on correcting errors made by previous models. Stacking blends predictions from several models using a meta-learner to achieve better performance[16]. To improve our model's accuracy and to cross-validate the results from the other models, ensemble methods were used.

IV. Result and Discussion

In this section, we will elaborately discuss the results of our model and the problems that we faced during the learning process. To start with, our initial learning dataset was very well organized with no null values. As it was based on real-life data and every instance represents a student, there were no null values. The major issue we faced was that our dataset was highly biased towards course numbers, as it has a large number of values and data. To mitigate this, we scaled the overall data with a standard scaling. The results were clear, with data now being limited to a range of -5 to 30 maximum range.

For each model, we rigorously tested the model with bootstrapping and bagging data with ideal iterations. The model parameters were set using hyperparameter testing to ensure a result above 60%. Moreover, to ensure the models are capable of classifying data, they were saved as (pkl) files to run unsupervised learning later on the trained models.

A. *Correlation and Feature Importance*

In order to correctly identify the dominant classes, a correlation matrix was generated that gave us insight into the highly correlated classes. The correlation heat map visually portrays the information as: deep blue indicates highly positive correlation, light blue indicates low to moderate correlation, and the white blocks indicate a negative correlation among the classes.

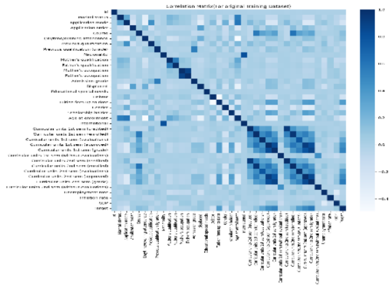

Fig. 2. Heatmap of the dataset

But the heat map fails to quantify the values of the correlated factors. So we had to run a person's correlation-based method where the values of the correlations were organized in a declining form. The results indicated that Curricular units 2nd Sem (grade) had the highest relation score of 0.5790, and Age at enrollment had the lowest score of -0.2568.

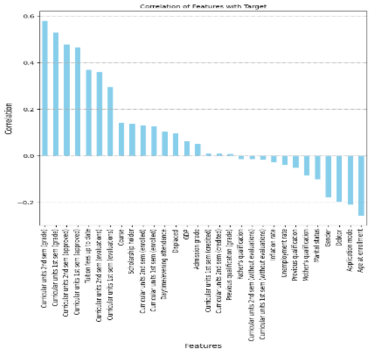

Fig. 3. Pearsons's correlation plot

To further analyze the most dominant features among the model, a majority vote was done from where the majority classes were distinguished successfully based on the top 3 results. This helps us to understand what highly influences a student's academic success the most.

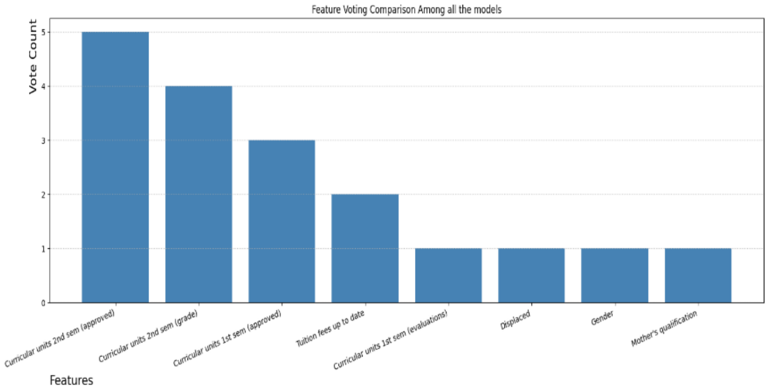

Fig. 4. Important Features based on voting

And lastly for result verification cross validation was also done to reassure the models predictive metric scores were highly accurate.

B. Performance Metrics

As discussed, various machine learning models were run over the dataset to identify the at-risk dropout students. The systematic analysis started with the models as Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Ada Boost, etc., except for the ensemble model and Knn All the models were assessed on their individual accuracy scores and F1 scores.

The ratio of correctly predicted instances to all instances is known as accuracy, which measures how accurate the model's predictions are overall, whereas the F1 score is the harmonic mean of recall and precision, which offers a balance between the two measurements [17].

Among our models, Random Forest achieved the highest accuracy of 83.02%, and the rest of the models were nearly as close to it, but not less than 75%.The ensembled method performed poorly, as it might be due to data biasness or class distributions. As earlier, the model became biased towards random forest, resulting in an accuracy of 79.2% in ensemble method analysis.

Further analyzing the precision, recall and the F1 scores made the observation clear, and it was clear that the dominant model among them was Random Forest (RF).

Table I

MODEL PERFORMANCE SUMMARY

Model	Performance Metrics		
	Accuracy	F1 Score	Weighted Average
Decision Tree	0.811879	0.8	0.81
Random Forest	0.8302	0.83	0.832
Logistic Regression	0.818544	0.82	0.82
Ada Boost	0.815	-	-
KNN	0.8	-	-
Neural Network	0.8153	0.82	0.83
Ensembled Model	0.7921	-	-

And to further test if our models can distinguish the difference or not, the previous learning models were called, and it was run on our test dataset, and it classified them, but we could not verify the results as there were no target columns, as this was a part of unsupervised learning.

V. Conclusion

Our model demonstrated positive results in predicting student dropout in higher education, despite the difficulties we encountered. This provides us a solid starting point and guides our future model refinement. We think the model could develop into a reliable tool for academic professionals and colleges with better data quality and wider sampling. A student's educational journey doesn't have to end when they drop out. By spotting early warning signs and offering opportunities for intervention, we hope to assist both institutions and students. We can strive towards a more precise and significant dropout prediction system with more research and a closer relationship with our own educational systems.

References

[1] Kim, D., & Kim, S. (2018b). Sustainable Education: Analyzing the determinants of university student dropout by nonlinear panel data models.*Sustainability*,*10*(4),954.https://doi.org/10.3390/su10040954

[2] B. M. S. Hossain and A. Naeema, “Employers perception of university graduates’,” *ResearchGate*, Aug. 2012, [Online]. Available: https://www.researchgate.net/publication/353884468_Employers_Perception_of_University_Graduates'

[3] A. F. Vignoles and N. Powdthavee, “The socioeconomic gap in university dropouts,” *The B E Journal of Economic Analysis & Policy*, vol. 9, no. 1, Apr. 2009, doi: 10.2202/1935-1682.2051.

[4] Rabelo, A. M., & Zárate, L. E. (2024). A model for predicting dropout of higher education students. *Data Science and Management*. https://doi.org/10.1016/j.dsm.2024.07.001

[5] N. Hutagaol and S. Suharjito, “Predictive modelling of student dropout using ensemble classifier method in higher education,” *Advances in Science Technology and Engineering Systems Journal*, vol. 4, no. 4, pp. 206–211, Jan. 2019, doi: 10.25046/aj040425.

[6] D. Andrade-Girón *et al.*, “Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review,” *ICST Transactions on Scalable Information Systems*, Jul. 2023, doi: 10.4108/eetsis.3586.

[7] D. K. Dake and C. Buabeng-Andoh, “Using machine learning techniques to predict learner drop-out rate in higher educational institutions,” *Mobile Information Systems*, vol. 2022, pp. 1–9, Nov. 2022, doi: 10.1155/2022/2670562.

[8] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, “Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization,” *Computers and Education Artificial Intelligence*, vol. 3, p. 100066, Jan. 2022, doi: 10.1016/j.caeai.2022.100066.

[9] V. Realinho, M. Vieira Martins, J. Machado, and L. Baptista. "Predict Students' Dropout and Academic Success," UCI Machine Learning Repository, 2021. [Online]. Available: https://doi.org/10.24432/C5MC89.

[10] “Classification with an Academic Success Dataset,” *Kaggle*. https://www.kaggle.com/competitions/playground-series-s4e6/data

[11] Breiman, L. ”Random Forests,” Machine Learning, 2001, 45, pp. 5–32, doi: https://doi.org/10.1023/A:1010933404324

[12] P. Schober and T. R. Vetter, "Logistic Regression in Medical Research," *Anesthesia & Analgesia*, vol. 132, no. 2, pp. 365–366, Jan. 2021. [Online]. Available: https://doi.org/10.1213/ANE.0000000000005247

[13] Wikipedia contributors, “AdaBoost,” *Wikipedia*, Nov. 23, 2024. https://en.wikipedia.org/wiki/AdaBoost

[14] IBIMA Publishing, “From Data to Decision: Machine learning and Explainable AI in student dropout prediction,” *IBIMA Publishing*, Jan. 15, 2025. https://ibimapublishing.com/articles/JELHE/2024/246301/

[15] M. W. Gardner and S. R. Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences,” *Atmospheric Environment*, vol. 32, no. 14–15, pp. 2627–2636, Aug. 1998, doi: 10.1016/s1352-2310(97)00447-0.

[16] Z.-H. Zhou, Ensemble methods: Foundations and Algorithms. CRC Press, 2012.

[17] “3.4. Metrics and scoring: quantifying the quality of predictions,” *Scikit-learn*. https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics