Student's Dropout and Academic Success Predictor

Md Saeem Hossain Shanto
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
Saeem.Shanto@northsouth.edu

Md. Ariful Islam
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
ariful.islam8@northsouth.edu

***Abstract*—Many educational institutions put a high value on monitoring and supporting university freshmen. The high proportion of students who do not receive their free education exemplifies this, as do the full economic and social costs associated with them. Higher education institutions' challenge is developing and improving policies that will increase student retention, particularly in the beginning years. In this paper, we propose a methodology and a specific classification algorithm to find a better way to trace the success and dropout of the students, which would help universities to take the necessary steps for their respective students. We have used decision tree, logistic regression, naive bayesian, and KNN algorithms and compared them to get better prediction results from them. The dataset we used is information on undergrad degree students of Polytechnic Institute of Portalegre, Portugal and the result we found from the model was able to predict the students' future situation of their degree. This will help the education institution and other concern institutions to make early decisions about students for their betterment.**

***Index Terms*—monitoring, classification, early decisions**

I. Introduction

Predicting student dropout in high school is an essential task. Issues in education since it affects far too many kids individual schools and institutions across the world, and It frequently results in total financial loss and a reduced graduation rate. rates and a poor school reputation in the eyes of all included [1]. The sacred goal of every academic institution is to embed learning. But it can't be achieved if academic institutions don't monitor their student's dropout rates and by not reduce them. Dropout is defined differently by researchers, but in any case, if an institution loses a student by any means, the institution has a lower grade point average. Any school retention plan must be successful by early detection of students who are likely to quit their classes. And, in order to try to lessen the aforementioned issue, it is vital to identify students who are at risk as early as possible and provide them some attention in order to stop them from abandoning their studies and act early to support student retention [2]. Dropout prediction is concerned with predicting students who are unlikely to be retained in the academic institution depend on a variety of factors. These factors can be financial, educational, social, or psychological in nature. Students who are impacted by such factors are less likely to stay in the system, which is why academic institutions should focus more on this group of students to help them stay in the system. Seidman created a concept for student retention [3] that demonstrates that identifying at-risk students early and maintaining rigorous and ongoing intervention is essential to lowering dropout rates. To solve the issues, we applied a machine learning approach to identify the early dropout, smoothly graduate, and late graduate students that could be essential for any educational institution to take necessary steps to reduce their student dropout rate.

II. Literature Review

The purpose of this research is to forecast student academic status using the Decision Tree machine learning approach. But the recent research by Ofori et al. (2020) suggested that Depending on the applicability of the acquired data and the goals of the data analytics process, numerous Machine learning models, such as clustering, classification, and association rules mining, might be adopted to analyze the data. Students' dropout is one of the most significant and difficult difficulties that students and global institutions encounter across the world [4]. As a result, accurately anticipating student dropout might help reduce the financial and social costs. [5] Recent research have revealed that Machine Learning approaches are used to forecast students who are at danger of failing and dropout rates in order to enhance their academic performance. [6] Several machine learning models based on algorithms such as decision trees, neural networks, random forests, support vector machine, and logistic regression have been developed to predict student dropout. Machine learning uses classical statistical analysis to improve predicted accuracy over verified hypotheses. [4]. However, the efficiency of these models differs mostly according to the type and quantity of datasets analyzed, feature selection methodologies, performance evaluation criteria, and experimental procedures. Furthermore, various types of literature employed different approaches and chose predicted factors. The table below covers current research on students' academic success, dropout rate prediction, identifying at-risk students, and evaluating the effectiveness of the prediction model. Carlos et al. [7] use dataset consisting of 419 high school students from mexico. They applied data mining to conduct a series of experiments to predict dropout at various stages of the course. They demonstrate that their algorithm can predict student dropout during the first 4-6 weeks of the course and is reliable enough to be implemented in an early warning system. Limsathitwong et al. [8] did a case study

on the students at Thai-Nichi Institute of Technology between first-year and second-year that have a high rate dropout. They evaluate the data to determine the dropout rate and create a web application to anticipate the status of students based on their grades in each subject. To accomplish an improvement, the prediction models were created to use Decision Tree techniques as well as Random Forest Algorithms. Precision, recall, and F1-Measure for the Decision Tree classifier were 0.80, 0.92, and 0.85, respectively. Amare et al. [9] use a dataset consist of 472 students. By analyzing the dataset they aims to develop a prediction model for students' dropout prediction using machine learning techniques. They achieve 94% accuracy by using logistic regression and got 93% accuracy on both Decission tree and Random forest algorithm.

III. Methodology

The following are the main methods we have used to develop the students' dropout and academic success predictor model:

1) Data Collection - includes the collection and integration process.
2) Data Pre-processing - includes the methods of pre-processing the input data before training the model.
3) Data-set splitting - includes the method of splitting the data-set for training, testing.
4) Model Building - includes the technique or algorithms used for training and testing the data-set.
5) Prediction and performance evaluation - includes prediction, performance comparison, and evaluation using accuracy, precision,F1-score, and recall performance.

A. Data collection

In our dataset we have performed preprocessing to handle data from anomalies, unexplainable outliers,missing values, and dropped records that couldn't be classified as explained. The final dataset consisted of 4424, Among them, 2209 data are from graduates, 1421 data are from dropout, and 794 are from enrolled. The data contains major features like age at enrollment, gender, marital status, nationality, address code, and special needs) and socio-economic factors like studentworker, parent's habilitations, parent's professions, parent's employment situation, student grant, and student's debt) and student's academic path like admission grade, retention years at high school, order of choice for enrolled course, type of course at high school [10]. We used institutional data from students enrolled in undergraduate courses at the Polytechnic Institute of Portalegre, Portugal, for this study. The data applies to student records from the academic years 2008/09 to 2018/2019 and from various undergraduate degrees such as agronomy, design, education, nursing, journalism, management, social service, and technology. This dataset is publicly accessible. [11] Depending on how long it took the student to complete her degree, each record was classified as Graduate, Enrolled, or Dropout. These three are the target class of our dataset. Graduate means the student completed the degree on time; Enrolled means the student took up to three extra years to complete the degree; Dropout means the student took more than three extra years to complete the degree or did not complete the degree at all.

B. Data Pre-processing

Data-set pre-processing is one of the important tasks for better prediction in machine learning. The gathered data is cleaned through pre-processing in order to make it ready for machine input. To improve the performance efficiency of the predictive models, entries with missing values and nulls should be removed from the data-set. To do the initial data processing, we check the dataset using python's pandas,seaborn, and matplotlib library. But could not find any null or missing values in the data-set. The data-set was already preprocessed by Realinho et al.

C. Data-set Splitting

Using the standard process, data was separated into two sets: training (80%) and test (20%). To avoid over-fitting, a 10-fold cross-validation approach was performed for each model. This indicates that the training data set was divided into 10 blocks, and each model was trained using 9 of them, with the remaining one utilized for validation. The technique was performed 10 times, once for each block, to maximize the total, a number of validation observations while avoiding over-fitting. The estimator with the highest average cross-validation score was chosen. This technique also included a process for ensuring that each class was adequately represented in each fold. The overall performance of each chosen model was then evaluated using the test sets. Python Version- 3.10.4 was used in this process

IV. RESULTS

Configuration: Intel(R) Core(TM) i5-8300H CPU 2.30 GHz RAM : 8.00 GB (7.85 GB usable) GPU:1050Ti.

Technical Specification: The model training and testing was done using Anaconda Navigator and Jupyter Notebook and Python was the programming language we used. Designed to make package management and deployment easier, In order to predict whether a student will graduate, enrolled or dropout, we have decided to use Decision Tree. Result evaluation for the decision tree is given below-

A. Performance Table

	Precision	Recall	F1-Score
Dropout	88%	62%	73%
Enrolled	43%	20%	28%
Graduate	70%	97%	82%

B. Confusion Matrix

		Predicted		
		Dropout	*Enrolled*	*Graduate*
Actual	*Dropout*	169	34	432
	Enrolled	588	673	259
	Graduate	105	133	12

C. Sensitivity Rate

	TPR	TNR	FPR	FNR	ACC
Dropout	61.68 %	96.24 %	3.76%	38.32%	85.53%
Enrolled	20.36%	93.73%	6.27%	41.27%	79.89%
Graduate	97.30%	58.73%	41.27%	2.70%	78.08%

Here, our first chosen model was Decision Tree. We can observe the model performance from the result section using Decision Tree, and by doing 10-fold cross-validation, we achieved 72.9% accuracy. After that, we analyse the performance using precision, recall and F1-score. We addressed the problem as a three-category classification task. That's accuracy alone is not a good performance metric. We need to rely on various performance metrics like precision, recall and F1-score. Our model accuracy is 72.94% and checking the cross-validation using 10-fold the maximum validation accuracy score was 74.4% and the lowest was 70.8% and these are near to the model accuracy. So we can consider that our model is not overfitted. But As we applied the decision tree algorithm, we tried to optimise it using Gini impurity and ID3 functions. They helped us to determine which splitter is best to build a pure tree.

D. Prediction Models

Performance Measures	Decision Tree	Naive Bayes	Logistic Regression	KNN
Accuracy	72.9%	69.0%	78.0%	64.0%
Precision	67.0%	61.7%	73.0%	57.0%
Recall	59.7%	60.7%	68.0%	55.05%
F1 Score	61.0%	61.0%	69.0%	55.0%

After that, We compare the result of our initial machine learning model Decision tree with other popular model such as Naive Bayes, Logistic Regression and K-Nearest Neighbour (KNN). The accuracy of these models are 69%, 78%, and 64% respectively. With an accuracy of 78%, 73% of precision, 68% Recall, and 69% F1 score in predicting students' success or dropout, Logistic Regression is clearly the winning model here. The decision Tree model is the second best model here with an accuracy of 72.9% and 61% F1 Score. Naive Bayes is close to the result of the Decision tree but KNN performs worst here.

V. Conclusion

This research includes a systematic review that compiles information on several types of system solutions that use ML approaches to predict and minimize student dropout. our research implemented several machine learning algorithms to predict students' academic success or dropout. The model was trained and tested using Decision Tree, Naive Bayes, Logistic Regression, and K-Nearest Neighbor. The suggested prediction approach assists course instructors, institutes, and the University in making decisions about students' performance and implementing suitable interventions to improve students' academic performance in advance. This study discovered that the Logistic Regression model predicted students' premature dropouts better than the other models used in this study. Future research needs to include unstructured datasets from students' online activity and using deep learning methods would be essential to assess the effectiveness of the prediction. We also want to build an UI based software which will be user friendly and easy to use for prediction. Machine learning accuracy is subjective. However, anything better than 70%, in our perspective, is excellent model performance. In reality, an exact measurement of 70%-90% is not only desirable but also attainable.

References

[1] R. C. Neild, R. Balfanz, and L. Herzog, "An early warning system," *Educational leadership*, vol. 65, no. 2, pp. 28–33, 2007.

[2] J. B. Heppen and S. B. Therriault, "Developing early warning systems to identify potential high school dropouts. issue brief." *National High School Center*, 2008.

[3] A. Seidman, "Retention revisited: R= e, id+ e & in, iv." *College and University*, vol. 71, no. 4, pp. 18–20, 1996.

[4] F. Ofori, E. Maina, and R. Gitonga, "Using machine learning algorithms to predict studentsâ€™ performance and improve learning outcome: a literature based review," *Journal of Information and Technology*, vol. 4, no. 1, 2020.

[5] B. R. Cuji Chacha, W. L. Gavilanes López, V. X. Vicente Guerrero, and W. G. Villacis Villacis, "Student dropout model based on logistic regression," in *International Conference on Applied Technologies*. Springer, 2019, pp. 321–333.

[6] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student'performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, p. 552, 2021.

[7] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: a case study with high school students," *Expert Systems*, vol. 33, no. 1, pp. 107–124, 2016.

[8] K. Limsathitwong, K. Tiwatthanont, and T. Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," in *2018 5th International Conference on Business and Industrial Research (ICBIR)*. IEEE, 2018, pp. 110–114.

[9] M. Yihun, "Global challenges of students dropout: A prediction model development using machine learning algorithms on higher education datasets," in *SHS Web of Conferences*, vol. 129, 2021, p. 09001.

[10] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predict students' dropout and academic success," Dec. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5777340

[11] 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 *World Conference on Information Systems and Technologies*. Springer, 2021, pp. 166–175.