Navigating Success: Harnessing Academic Data for Student Outcome Projections

Isha Mahadalkar
Computer Graphics Technology
Purdue University
West Lafayette, IN, USA
imahadal@purdue.edu

Shashank Thandri
Computer Information Technology
Purdue University
West Lafayette, IN, USA
sthandr@purdue.edu

Fan Yang
Construction Management Technology
Purdue University
West Lafayette, IN, USA
yang2352@purdue.edu

***Abstract*—An often underrated yet crucial concept in education is an educator's ability to identify and help students before they reach a point in their academic career where they are forced with failing or dropping out of school. While it can often be difficult to make predictions, some works have demonstrated the effectiveness of machine learning in predicting student performance. The issue with making predictions for education data is that there is not a one size fits all solution. Every dataset is different due to the complexity of the parameters and even the education systems that exist within each country. The importance of using academic data to make predictions is well known, however very few have actually implemented this train of thought. Our paper investigates the importance of the different features in predicting academic performance. We specifically focus on the impact of college students' first-year grade on their academic predictions. We assess traditional classification models, including but not limited to Logistic Regression, Decision Trees, and SVMs, while also creating Ensemble models using these classification methods. Additionally, we conduct hyperparameter tuning and incorporate evaluation metrics to rigorously test the models on our dataset.**

***Index Terms*—Education intervention, Machine learning, Performance prediction**

I. Introduction

Education is a dynamic journey, one that is filled with peaks and troughs. It is heavily influenced by the timely support of educators and teachers alike. There have been a number of instances where students have shown considerable improvement in their academic performance when early intervention has been employed. The proactiveness of educators to help students is crucial for academic success. This proactiveness can help students finish their education, without fear of them dropping out or failing entirely. Early intervention plays a pivotal role in this and will allow educators to help their students.

Predictive analysis can play a huge role in determining whether a student needs early intervention or not. The issue with predictive analysis, however, is that it hinges on identifying the right features that indicates a student's risk level. The performance of a predictive model depends on the ability to select features that are indicative of a students' potential academic trajectory. These features include demographic information, socio-economic, personal, etc.

As a result of this information, this study will jump into the realm of predictive analysis, with a specific focus on identifying the features that most accurately predict academic risks. Our aim with this study is to add to the current literature, regarding predictive models in an academic sense. We also hope to create a simple but robust tool that will highlight the most important features in a dataset for predicting students' outcomes. As mentioned before, a lot of the studies have focused on non-academic information in order to make predictions. We aim to extend beyond this scope and introduce an additional feature. This additional feature would be academic information for making predictions. The first year of college is critical for students' success and we aim to use this information in order to make predictions. By analyzing how this set of academic data is paired with other information in the dataset, we look to enable educators to make more targeted and effective interventions for students. So, with this study we aim to answer the following research questions: 1. What are the most effective features and factors in predicting a student's academic performance? 2. Does performance during the first year of enrollment affects a student's final academic outcome?

II. Literature Review

As previously mentioned there several studies that have worked with education datasets and creating predictive models based on the information provided. [1] created a two model study that was compared in order to see the importance of specific features. One model incorporated only academic features and the other incorporated both academic and non-academic features. Their research was conducted on a dataset that contained 6,807 datapoints of students from a technical college in India. They used multiple classification algorithms in order to determine the most effective parameters to make predictions on students' outcomes. They found that the model that used both set of parameters provided the best results. However, tests also showed the effectiveness of the academic parameters. In their study they showed that the model that only used academic parameters produced results of 79.3% accuracy. Studies have also focused on using non-academic parameters only in order to make predictions. [2] conducted a study focused on making predictions based on socio-economic and demographic data. Their study evaluated the models using a variety of metrics which include: confusion matrix, classification accuracy, precision, recall, F1-score, and area under the curve (AUC). They built their model using a number of classifiers including: Random Forest [3], Neural Networks [4], Support Vector Machines [5]. The researchers achieved a classification accuracy of 73% on average. Another study conducted by [6] focused on making predictions on non-academic parameters. They employed XGBoost [7] and other various boosting methods to make their predictions. They achieved an average classification accuracy of 71%. [8] was another study conducted that focused on each students lifestyle rather than their academic information. Their study used social participation as a method of predicting how well a student will do in school. They conducted 18 experiments using 2 datasets and 5 different classifiers. In their results they discovered that the ability to correctly process the data can lead to improved outcomes. Deep learning has also been implemented in predictive analysis regarding education. [9] addresses this idea in their own work. In their work they aim to predict student performance using online coursework, a task previously unexplored. In this study the researchers propose a method called GritNet. GritNet is an algorithm that is based on bidirectional long short-term memory. This method was shown to surpass traditional Logistic Regression models [10]. This study was one of the first to show the effectiveness of deep learning in predictive educational models.

III. Methodology

In this section we start by detailing the dataset we are working with, outlining the data preprocessing techniques we employed in the study, our exploratory data analysis techniques, and the baseline modeling approach. Additionally, we provide a concise overview of our advanced modeling techniques and evaluation metrics.

A. Data

Building upon the introductory study conducted by Martins et al. [6], we utilized the dataset they collected [11], which originates from the Polytechnic Institute of Portalegre, Portugal referring to students enrolled in various different degree programs. This dataset comprises 4424 records with 37 independent variables. It captures a diverse set of dimensions, including demographic data (such as age, gender, marital status), socio-economic data (including parental qualifications, loans, scholarships, etc.), and academic path factors (such as first-year curriculum and grades).

The full list of variables employed in our study includes ’Marital Status’, ’Application Mode’, ’Application Order’, ’Course’, ’Attendance’, ’Prev Qual’, ’Prev Qual Grade’, ’Nationality’, ’Mom Qual’, ’Dad Qual’, ’Mom Occ’, ’Dad Occ’, ’Admission Grade’, ’Displaced’, ’Edu Needs’, ’Debtor’, ’Tuition Date’, ’Gender’, ’Scholarship’, ’Age’, ’International’, ’CU 1 Credited’, ’CU 1 Enrolled’, ’CU 1 Eval’, ’CU 1 Appr’, ’CU 1 Grade’, ’CU 1 Without Eval’, ’CU 2 Credited’, ’CU 2 Enrolled’, ’CU 2 Eval’, ’CU 2 Appr’, ’CU 2 Grade’, ’CU 2 Without Eval’, ’Unemployment Rate’, ’Inflation Rate’, and ’GDP’ [11].

Our investigation is centered around examining the crucial role of academic grades in the first year for predicting student success, along with the consideration of demographic and socioeconomic factors. This comprehensive dataset provides us with the opportunity to thoroughly explore and analyze these various facets.

The target variable categorizes students into “Graduate”, “Enrolled” and “Dropout” classifications, representing different risk profiles – ranging from ‘low risk’ to ‘medium risk’ to ‘high risk’ [6]. Low-risk students are those who graduate within the standard four years, medium-risk students take more than three years to complete the course of study, and high-risk students are more likely to drop out without completing the course [6].

B. Exploratory Data Analysis

We initiated our analysis, using Python’s scikit-learn library [12], by cleaning the data, scrutinizing for any null values or duplicate rows, and renaming the columns for enhanced readability. Categorical variables underwent encoding using One-Hot Encoding [13]. We decided to encode the Target variable to assign specific integers to each individual label: Graduate - 0, Enrolled - 1, and Dropout - 2. We use Standard Scaling techniques [14] to ensure uniform scaling of numerical variables.

We continued to explore the dataset to understand its structure and gain insights into its distribution of features. We did a thorough analysis of its summary statistics, and we visualized correlations among numerical features and the distribution between numerical and categorical features.

C. Data Sampling and Feature Selection

While performing the data analysis, we saw that we had a class imbalance, as seen in Figure 1. Different sampling strategies are commonly used to address this issue of class imbalance based on the data and the problem you are trying to assess. This can be done by either removing some data from the majority class (under-sampling), duplicating data from the minority class (over-sampling), or introducing artificially generated data to the minority class [6].

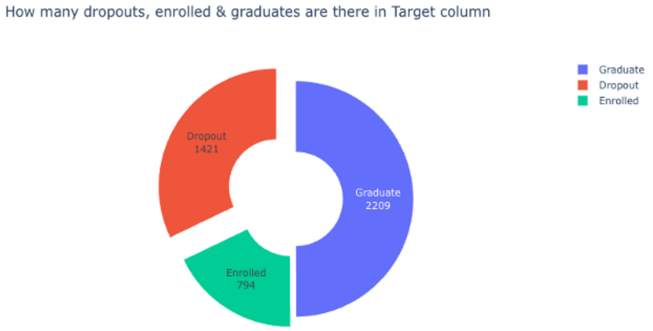

Fig. 1. Distribution of classes in the Target variable.

Since the previous research [6] indicates that utilizing SMOTE [15] to address the class imbalance in this specific problem produces optimal results, we decided to follow the same approach.

We also do feature selection where we initially, looked at feature correlations, examining the relationships between different attributes to see which features were highly correlated to the target variable. Figure 2 shows the top 10 features that have the highest correaltions to the Target variable. Simultaneously, we also performed a manual selection process, where we handpicked 3-4 features, driven by their relevance to the problem, correlation with the target variable, and our collective domain knowledge. This method allowed us to get insights from the different attribute and then also decide to drop 8-10 features which were the least important allowing us to reduce the dataset's dimensionality. By focusing on the most relevant attributes, we aimed to enhance the reproducibility of our findings in the diverse landscape of educational data, laying a foundation for advancing future research in the realm of education.

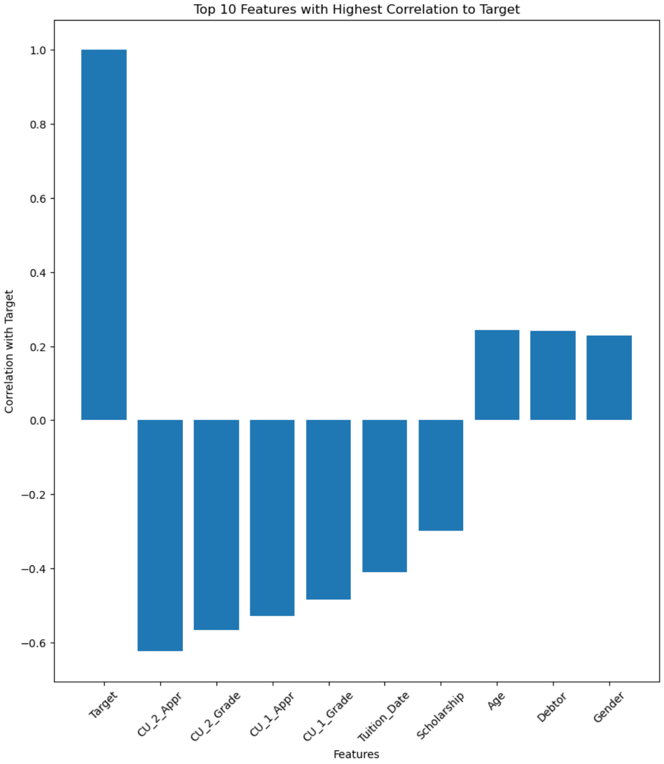

Fig. 2. The top 10 features with the highest correlation to the Target variable.

After this process, we decided to drop the following columns: ['Course', 'Nationality', 'Mom Qual', 'Dad Qual', 'Mom Occ', 'Dad Occ', 'Edu Needs', 'International', 'CU 1 Credited', 'CU 1 Eval', 'Unemployment Rate', 'Inflation Rate', 'GDP']. Despite this pruning, we chose to retain the "CU*" columns, preserving the first-year grade and class information, since it was crucial to our analysis.

D. Classification Models

To facilitate comparison and evaluation, we constructed a Logistic Regression model [10] using the original dataset, implementing all preprocessing steps. Additionally, we incorporated curriculum data for the first and second semesters in our Logistic Regression model to assess early intervention and ascertain the significance of first-year grades in predicting student success. We considered results from the introductory paper [6] as well as our own findings in this process, forming the baseline for our analysis.

For our analysis, we used the typical models used for classification tasks as an initial approach to test their usefulness before trying various other techniques including boosting, stacking and ensembles. The models that we utilized which are well suited for multi-class classification are: Decision Trees [16], Random Forests [3] which is an ensemble method, LogisticRegression [10] K-Nearest Neighbors [17] and Linear Support Vector Machines [5]. We also apply boosting classification models. Boosting techniques belong to the ensemble methods category, where a robust model is constructed through the sequential training of weaker models [18]. We use the AdaBoost [19] and the Extreme Boosting Classifier [7] for our analysis since they are reported to provide good results for multi-class classification.

E. Model development and optimization

We divided our data into training (80%), testing (5%), validation (5%), and holdout (10%) sets. The Pipeline class from the scikit-learn library [12] was employed to construct the model pipeline for both the individual classifier models and the eventual ensemble model. To build the model pipeline, we utilized Python's Column Transformer as a preprocessor, wherein the StandardScaler and OneHotEncoder were fitted on the training data before applying it to the model. These techniques can help to improve model performance, reduce the impact of outliers, and ensure that the data is on the same scale. This approach was adopted to transform the testing data, thereby preventing any data leakage during the model testing process [20]. The test set was utilized to produce a classification report for evaluating the model's performance. Since we had an imbalanced dataset we focused on both the accuracy and individual and average F1 scores. As F1 scores account for both precision and recall, they serve as a robust metric for evaluating our model's effectiveness [21].

We used the four best-performing classifiers as base estimators to create the ensemble using the VotingClassifier [12], and we followed the same pipeline model for its evaluation. Random Forests [3], Logistic Regression [10], Extreme Boosting [7], and SVC [5] were utilized to form the final ensemble. Figure 3 visualizes our model pipeline.

Additionally, we perform hyperparameter tuning on the classifiers of the final model where we tune the hyperparameters to maximize the F1 Score. We use GridSearch to test multiple configurations and select the sets that performs better with cross-validation. The holdout set is utilized to optimize training on a small subset of the data, and the validation set is used to conduct an unbiased evaluation of the model's performance.

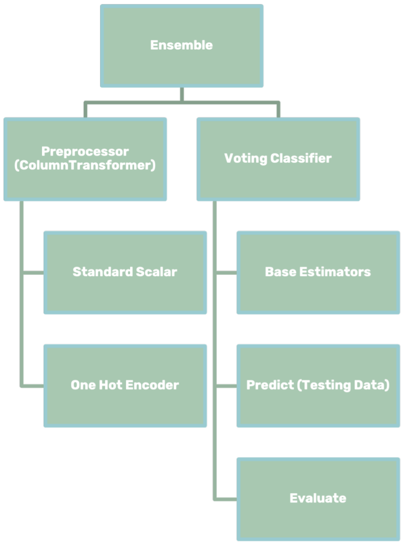

Fig. 3. Model Pipeline

IV. Results and Analysis

Table I shows the results of our analysis of the baseline classifiers obtained from the test set. This shows that utilizing features that include first-year academic performance is vital to accurately predicting student success. The results show a 74% accuracy and 0.71 average F1 score, average and individual, for our baseline model which included curriculum features perform better than using the best model without academic features [6].

TABLE I
Baseline Classification Performance

	Accuracy	Avg F1 Score	F1 Score (Dropout)
Logistic Regression [6]	0.68	0.49	0.61
XGBoost [6]	0.73	0.65	0.68
Logistic Regression w/ curriculum data (Baseline)	**0.74**	**0.71**	**0.75**

The evaluation metrics used for the individual classification models and the ensemble were accuracy, average F1 score and the F1 Score for the Dropout class. We decided to put additionally focus on the Dropout class since we were focusing on early intervention for at risk students. Table II shows the results for the individual classifiers. XGBoost was one of the top performing classifiers which was also showed by previous researchers [6]. Random Forests, Logistic Regression and SVMs were the other best performing classifiers.

The top four performing classifiers were utilized to create the ensemble. The final ensemble model pipeline achieved our highest accuracy of 77%, with an individual Dropout F1 score of 0.76. Table III presents the results of the final pipeline, including both hard and soft voting scores for our final findings. Even though, multi-class classification is a complex task we found that using ensemble models perform better than using only individual models.

TABLE II
Individual Classification Performance

	Accuracy	Avg F1 Score	F1 Score (Dropout)
Decision Trees	0.71	0.67	0.72
Random Forest	**0.75**	**0.71**	**0.77**
Logistic Regression	**0.74**	**0.71**	**0.77**
K-Nearest Neighbours	0.69	0.65	0.63
Adaboost	0.71	0.67	0.71
XGBoost	**0.75**	**0.71**	**0.74**
SVM	**0.73**	**0.70**	**0.73**

TABLE III
Classification Performance for Ensemble Models

	Accuracy	Avg F1 Score	F1 Score (Dropout)
Hard Voting	0.76	0.72	0.76
Soft Voting	**0.77**	**0.73**	**0.76**

V. Limitations and Future Work

Over the course of this study, we came to the realization about certain limitations that exist within the context of this study. These limitations hampered our study in ways that we did not foresee before starting. One of the limitations that we concluded was that the data size and the information in the dataset can limit the generalizability of the conclusions. The data that we worked with was focused on a specific group of people. Therefore, it can limit the findings that we have to broader populations. If we used a larger more diverse dataset, then we would be able to apply our findings to a broader more general population.

The second limitation that we came across was that the model would assume independence across features. In most situations this would not have affected the findings, however, in the context of our study, it is inappropriate to assume feature independence. When it comes to education data it is important to understand that features are often not independent of each other. There are certain attributes that can affect other ones, and a machine learning model is often times not nuanced enough to understand how it works.

The third limitation that we came across was that our dataset does not account for temporal changes in a student performance. Our dataset is a snapshot of how well students are doing at a very specific time. However, the performance can change over time and our dataset does not account for this change that may or may not occur. Being able to account for temporal changes is an important part of predictive analytics that our study does not account for.

Along with limitations we have also concluded that there is a lot of future work that can occur regarding this study. One of the directions that this study could go would be the implementation of deep learning architectures. The implementation of these deep learning architectures could uncover more complex relationships within educational data. Another future direction this study could head is improving the interpretability and explainability. The development of methods for enhancing the interpretability of the models can help lead to more transparent insights. Finally, the ability to transfer this study to other areas of education is another future work of this study.

VI. Conclusion

In this work we used a dataset from a higher education institute in Portugal. The premise of this study was to build a classification model that would be able to accurately predict the outcome of students based on their first year performance. We address the idea that using academic information we can make accurate and effective predictions that can be used for early intervention. We used 7 different classifiers in order to determine the effectiveness of this data across multiple models. Our results show that academic parameters are crucial to predict the outcome of students in higher education. We also show that creating an ensemble with the various base classifiers is the best method for accurately making these predictions.

References

[1] D. Aggarwal, S. Mittal, and V. Bali, "Significance of non-academic parameters for predicting student performance using ensemble learning techniques," *International Journal of System Dynamics Applications (IJSDA)*, vol. 10, no. 3, pp. 38–49, 2021.

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

[3] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[4] L. Hardesty, "Explained: Neural networks," Apr 2017. [Online]. Available: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[6] 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.

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[8] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student performance prediction and classification using machine learning algorithms," in *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, 2019, pp. 7–11.

[9] B.-H. Kim, E. Vizitei, and V. Ganapathi, "Gritnet: Student performance prediction with deep learning," 2018.

[10] T. J. Hastie and D. Pregibon, "Generalized linear models," in *Statistical models in S*. Routledge, 2017, pp. 195–247.

[11] V. Realinho, M. V. Martins, D. Tolledo, J. Machado, and L. M. Baptista, "Predict students' dropout and academic success," UCI Machine Learning Repository, 2021, DOI: https://doi.org/10.24432/C5MC89.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Courty, E. Duchesne, J. Eskine, M. Kevin, O. Louppe, R. Nguyen, and J. Vanderplas, "scikit-learn: Machine learning in python," pp. 2825–2830, 2011.

[13] J. Brownlee, "Why one-hot encode data in machine learning?" Jun 2020. [Online]. Available: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[14] S. Mulani, "Using standardscaler() function to standardize python data," Aug 2022. [Online]. Available: https://www.digitalocean.com/community/tutorials/standardscaler-function-in-python

[15] G. LemaÃŽtre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of machine learning research*, vol. 18, no. 17, pp. 1–5, 2017.

[16] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[17] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. Springer, 2003, pp. 986–996.

[18] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Systems with Applications*, vol. 41, no. 2, pp. 321–330, 2014.

[19] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[20] A. Bhandari, "Feature engineering: Scaling, normalization, and standardization (updated 2023)," Oct 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

[21] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *Journal of Big Data*, vol. 7, pp. 1–47, 2020.