Student Dropout Classification

Course: Machine Learning and Deep Learning

Students - Copenhagen Business School

MSc. Business Administration and Data Science

Submission date: 16/05/2025

Number of characters: 34.108

Number of pages: 15

Group members:

Xiaosu Feng (176837) Wenxin Gou (176833)

Hans-Henrik hansen (153689) Frederik Skov Wacher (S145204)

Table of Contents

Abstract

This study explores the application of machine learning models to predict student dropout and academic success in higher education. The central problem is to identify students at risk of dropping out. This allows for timely interventions that can help reduce dropout rates and improve the overall quality of educational outcomes. Our research examines how a baseline (Random Forest) and two complex models (XGBoost, Multilayer Perceptron) predict student outcomes (Dropout, Graduation) in an imbalanced binary setup and pinpoints key predictive factors. The conceptual framework integrates supervised learning with comprehensive feature engineering, incorporating academic performance indicators, socioeconomic variables, and macroeconomic factors. Datasets we used contained 4,424 undergraduate student records from Portugal, employing classification algorithms with stratified sampling and class weighting to address data imbalance. Results indicate that XGBoost achieves best performance with a macro-F1 score of 0.90 and superior dropout class recall (0.83). Additionally, feature importance analysis indicates that first-year academic performance is the most influential predictor, followed by course type, age at enrollment, prior academic performance. We conclude that XGBoost offers the most suitable model for early warning systems in higher education, and recommend future work incorporating time-series behavioral features and advanced optimization techniques to address persistent classification ambiguity in boundary cases.

Keywords: Machine Learning in Education, XGBoost, Random Forest, Multilayer Perceptron, Class Imbalance

1. Introduction

Student dropout and academic underperformance remain critical challenges for higher education institutions, with far-reaching economic and social implications for students, universities, and policy-makers (Realinho et al., 2022; Martins et al., 2021). A variety of early-warning systems have been proposed that draw on a mix of demographic (e.g., age, gender), socio-economic (e.g., scholarship status, family income), pre-university (entrance exam scores, high-school GPA) and first-year academic performance indicators (e.g., credit completion, first-semester grades). To address the severe class imbalance inherent in multi-outcome prediction (dropout vs. still enrolled vs. graduate), many of these studies apply SMOTE oversampling or class-weighted loss functions (Martins et al., 2021; Villar & de Andrade, 2024). However, because most published work evaluates only one model family, often on proprietary institutional data, there is limited insight into how model complexity and feature engineering choices generalize across different algorithms and datasets.

To address this gap, we use the publicly available Portuguese undergraduate dataset (Shamim, 2023), comprising ~4424 student records, 36 attributes spanning demographics, socio-economic status, and course enrollments, to benchmark three classifiers of increasing complexity:

- Random Forest (RF): a widely used ensemble of decision trees offering interpretability and robust handling of mixed data types.
- XGBoost (XGBM): a gradient-boosted tree method known for its efficiency and superior baseline performance on tabular data.
- Multi Layer Perceptron-based Classification (MLP): A feedforward neural network trained to classify student profiles.

We evaluate each model under transparently engineered features, augmenting with first-year credit completion ratios, mean GPA, pass/fail course ratios, and macroeconomic indicators at enrollment (e.g., national GDP growth, inflation) that prior work has shown to be predictive and validated by us (Realinho et al., 2022).

Our primary evaluation metrics are macro-F1 score (slightly modified for MLP), ensuring equal weight to each outcome class, and per-class recall, to safeguard against under-detection of minority events such as dropouts vs graduates. Through a systematic comparison of one "non-complex" model (RF) and two "complex" approaches (XGBM, MLP), on both raw and enriched feature sets, we aim to identify the most effective combination for early identification of at-risk students, thus informing targeted retention strategies and resource allocation, focusing on a binary classification to maximize performance while balancing utility for universities:

1.1 Research question

How do one non-complex model (random forest) and two complex models (XGBoost, Multilayer Perceptron) perform in predicting student outcomes (dropout, graduate), in an imbalanced binary outcome setting and what are the most important factors for predicting the outcomes?

2. Related Work

Across recent studies from 2020-2025, A cohesive body of research casts student outcome prediction as a three-class classification problem (dropout, still enrolled, graduate), employing different input variables, including demographics gender, age at enrollment, nationality, parents' educational attainment (Martins et al., 2021; Realinho et al., 2022); socio-economic variables like scholarship or grant receipt, tuition fee payment

status, self-reported household income (Martins et al., 2021); academic involving pre-university GPA, admission exam scores, number of curricular units enrolled vs. approved in the first year (Realinho et al., 2022; Shamim, 2023); and the macroeconomic and program context such as country-level GDP per capita, inflation rate at time of enrollment, and degree program characteristics (discipline, full-time vs. part-time) (Villar & de Andrade, 2024). Many authors further engineer composite indicators, most notably first-year credit completion ratio (approved credits ÷ enrolled credits), mean semester GPA, and pass/fail ratios, which repeatedly emerge among the top predictors in feature-importance analyses (Realinho et al., 2022; Martins et al., 2021).

The algorithmic approaches include interpretable baselines such as logistic regression, single decision trees, support vector machines, valued for transparency but typically achieving lower macro-F1 scores in the 0.50-0.60 range on imbalanced data (Martins et al., 2021; Villar & de Andrade, 2024). Beyond this, they include Random Forest, XGBoost (most performative across studies), LightGBM, and CatBoost. These outperform simpler models, attaining macro-F1 scores in the mid-0.60s to low-0.70s. Their success owes to inherent capacity for modeling non-linear interactions, along with straightforward integration of class weights or SMOTE oversampling (Martins et al., 2021; Villar & de Andrade, 2024). Finally, Öztürk, 2023 employed deep learning feedforward neural networks and autoencoders, matching or slightly trailing tree-based methods. These require larger datasets and careful tuning (dropout layers, L2 regularization) to prevent overfitting, particularly when sample sizes for minority classes are small.

For evaluation, researchers emphasize per-class precision and recall, confusion matrices and macro-F1 rather than overall accuracy, to ensure that rare but critical outcomes (actual dropouts) are not overshadowed by majority classes (Öztürk, 2023; Villar & de Andrade, 2024). Even so, misclassification rates for the smallest classes, such as delayed graduates, remain high (>30% in many cases), highlighting the ongoing need for advanced resampling techniques, cost-sensitive learning, and novel performance metrics in next-generation early-warning systems (Martins et al., 2021; Öztürk, 2023).

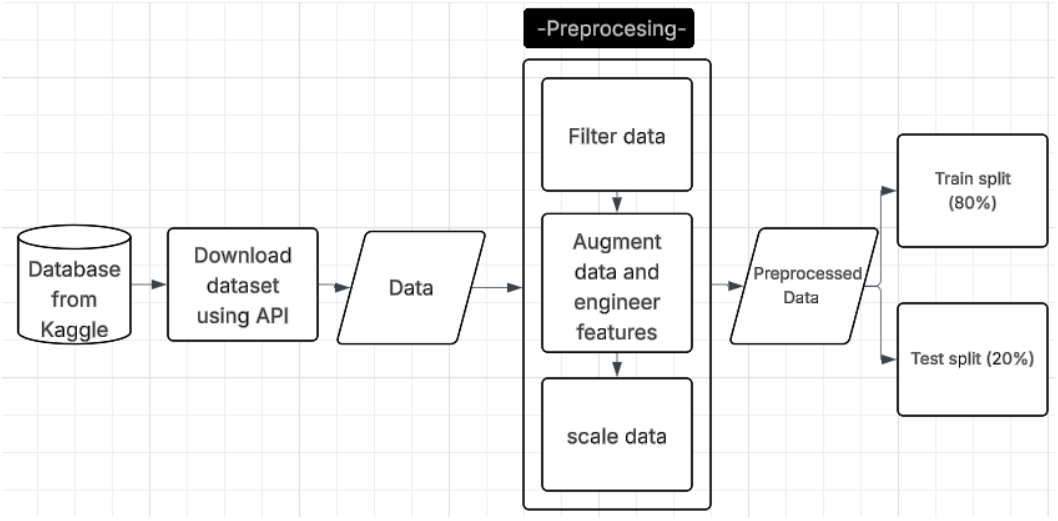
3. Conceptual Framework

3.1 Data Preparation

The data preparation process comprises a mix of data filtering, data augmentation,, and normalization, for one of the models being used.

The data that is being used has already been thoroughly preprocessed by the researchers that have collected it, so it does not contain any NA values and they have to a large extent handled all outliers in the data. Therefore, our preprocessing steps have primarily been to improve the data's suitability for the specific models that have been implemented in this analysis.

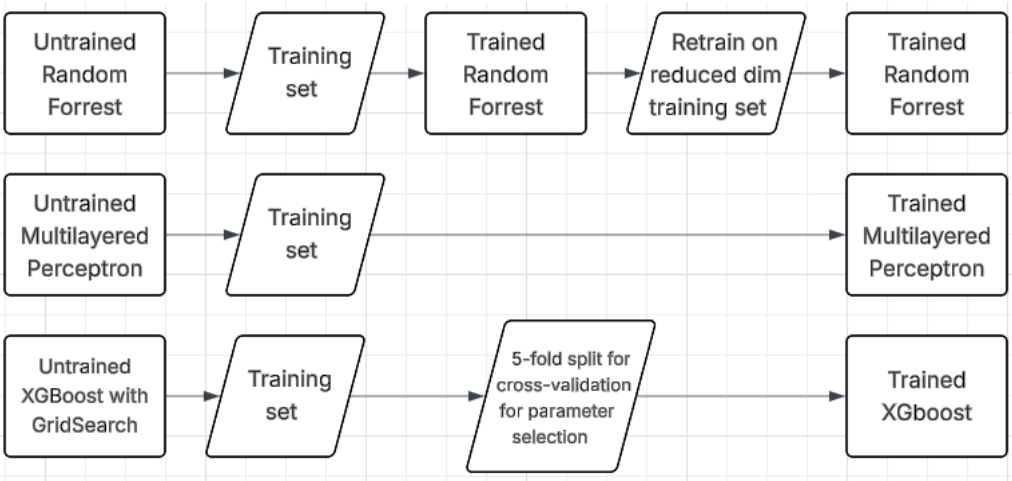
Below is a conceptual overview of the data preparation steps that has been applied:



3.2 Training Strategy

We utilize supervised learning models, and trained all the models on the same labelled data with the same training and test splits to ensure comparability. The baseline model utilizes dimensionality reduction to only maintain the 15 most relevant features, which was found to improve the models performance. The XGBoost model and the Multilayer Perceptron Neural Network does not use a reduced dimensionality but are trained on all the variables since these are well suited for determining the most important variables by themselves and, thus, does not need dimensionality reduction.

Below is an overview of the training strategies for the 3 models that has been applied:



4. Methodology

4.1 Dataset Description

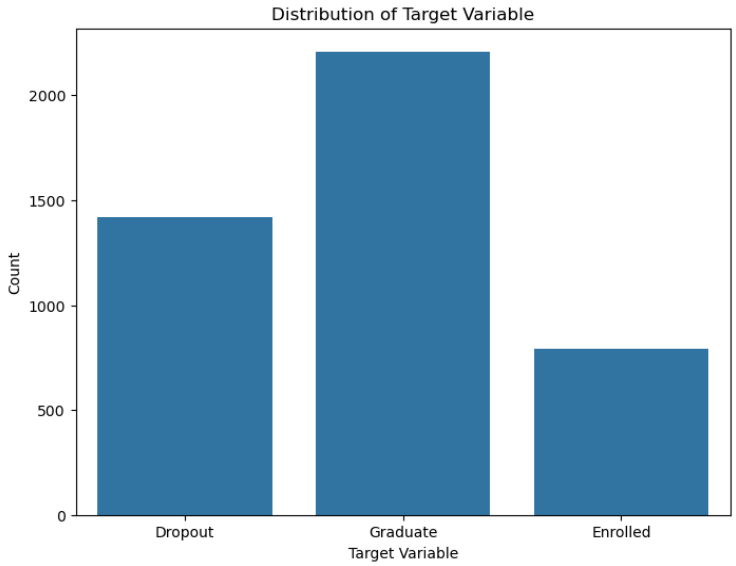
The study draws on the “Predict Students’ Dropout and Academic Success” data set first published on Kaggle by Shamim (2023). It contains 4,424 undergraduate student records collected at a Portuguese polytechnic institution and spans a wide range of degrees. There are in total 36 raw attribute columns plus the outcome column (total 37) with the target variable being a three-class label with values Dropout, Enrolled, and

Graduate. In the raw file the classes are distributed at roughly 32% Dropout, 18% Enrolled, 50% Graduate, making class imbalance an inherent characteristic of the data. There are no missing values as the the creators already removed NAs and obvious anomalies before release (ibid.). The variable groups before any preprocessing are described in Appendix 1 - Data Overview, including demographics, socio-economic and financial indicators, how and where the student was admitted, first-year performance, and the macroeconomic context. The breadth of attributes lets us test hypotheses about who is at risk of dropping out (demographics & socio-economics), where risk materialises first (first-year metrics), and when external shocks may amplify attrition (macroeconomics).

4.2 Data Preprocessing

4.2.1 Data Filtering

The class balance of the original dataset has been examined, and it originally contains 3 classes, “graduate”, “dropout” and “enrolled”. The balance of these classes can be illustrated as below:



We see that the original data contains a serious imbalance with the enrolled class constituting only 17% of the observations in the data and the graduate being overrepresented. Our initial research question and the intention of the analysis was to predict risk factors for students in terms of high likely they are to graduate or dropout as well as establishing a good model for forecasting the dropout rate based on the students that are currently studying in a given university. The “enrolled” class does not provide any real value to this analysis since students that are enrolled has not yet dropped out or graduated, which means they are delayed and can potentially end up either dropping out or graduating. Therefore, the enrolled class must be seen as a noisy element to the data in light of the purpose of the analysis. Since it only represents 17% of the data we shall filter out the observations where the student is classified as “enrolled” to maintain a focused and useful dataset for the purpose of the analysis.

After filtering out enrolled students we end up with a distribution of 40% of students being labelled “dropout” and 60% of students being labelled “graduate”.

4.2.2 Data augmentation

We ran a check for multicollinearity between our variables since we have several columns that potentially contain a lot of the same information. For this we visualised the correlations in a matrix extracting the correlations above 0.8 (see appendix 2 for reference).

We found that there is an extremely (>95) high correlation between curricular units (1st semester) and curricular units (2nd semester), and the 3 types of curricular units (enrolled, evaluations and approved in 1st and 2nd semester respectively). The variables containing curricular units that have evaluations also seem unnecessary since it highly resembles those variables containing curricular units that have been approved. Because of this we have dropped the two variables containing curricular units that have evaluations and instead keep the curricular units that have been approved.

Furthermore, since the variables for curricular units for 1st semester and for 2nd semester are highly correlated we will combine them into averages since we want to avoid including features containing multicollinearity. The same logic applies to grades achieved in 1st and 2nd semester. After having constructed the averages for curricular units and grades, we drop the original columns pertaining to these.

4.2.3 Data standardization

Data standardization has not been necessary for the random forest and the XG boost model since they are not sensible to variance in the variables. However, for the neural network all the columns containing true continuous numerical values have been scaled using the standardscaler method which transforms the variables so that they have a mean of 0 and a standard deviation of 1. However, most of the variables in the dataset have been encoded during the preprocessing performed by the data scientists collecting the original data, and these have thus not been appropriate to scale.

4.3 Modelling Framework

4.3.1 Splitting data and taking into account class imbalance

The models are using a train and test split with random_state=42 to ensure that the same splits are used for the models. Furthermore, the neural network uses a train, validation and test split since the validation split is an integral part of the way we run the neural network. Initially, we used a 5-fold cross validation method for the Random forest but the difference between the result of the model trained with cross-validation and the model trained on a train split was minimal and did not improve performance in any mentionable way. Therefore, we ended up using a simple train/test split instead since this is less computationally demanding.

Importantly, the “stratified” method has been used to perform the splits since we are dealing with unbalanced data, and this method ensures that we get the same proportions of our categorical target variable in each of the splits as the proportions that also apply in the original data. This hinders us from accidently introducing bias in terms of ensuring that we do not run the risk of overrepresenting a given class in the test, train or validation split.

However, even though the stratified method ensures that we do not introduce bias in the data when splitting it, it does not actually handle the class imbalance problem. For this several methods could be applied such as creating synthetic data that increases the proportion of observations belonging to the minority class using the

SMOTE method (oversampling). Alternatively, implementing an undersampling method could also have been chosen, but this is a controversial method since it runs the risk of dropping relevant information from the data.

We did not want to introduce synthetic data in the dataset since it does not handle potential cross class overlapping (observations of the two labels that highly resemble each other), has a risk of overfitting data and might decrease the representability of the data.

Instead, we have utilised class weights in all the models which alters how the loss function is evaluated based on the class weights instead of altering the data. The data is also not severely imbalanced (40/60) so the class weights method is suitable to use for it. The class-weight automatically adjusts the weight of dropout and graduate based on how frequent these classes are in the training data, and by doing so, it gives more importance to the underrepresented class (dropout) and if dropout is misclassified during training it will have a larger contribution to the loss-function.

4.3.2 Random Forest - baseline model

We have applied the ensemble learning method Random forest as a baseline model for our analysis. This model is made up of multiple smaller models (decision trees using each of their own subset of the data and a different number of features) where the predictions of each tree are combined taking the popular predictions amongst the trees (Schafi, 2020).

The random forest model has 100 trees in the forest and there is no max_dept. It uses sampling with replacement (bootstrapping) and the splits are performed based on the gini impurity criterion.

It is first used to extract the most important features for predicting whether a student will graduate or dropout, and here the 15 most important features are extracted. Based on this dimensionality reduction the model is then retrained on the 15 most important features to create predictions for whether a student will be a graduate or a dropout.

4.3.3 XGBoost

An XGBoost model has been implemented to check if the baseline model could be outperformed. XGBoost is a more computationally heavy model and introduces significantly more complexity compared to the Random Forest. XGBoost has more hyperparameters that allows for fine tuning of the model and it therefore offers more customisation, that in term allows for better adaptation to the data that the model is being trained on. However, it is nearly impossible to find the optimal tuning hyperparameters manually due to the large number of parameters and their complex effects on the model. Therefore, a GridSearch method (inbuilt in SciKit) has been used to find the optimal parameter selection. The GridSearch works by providing a grid of values for each parameter than is then evaluated based on a stratified K-fold cross validation

The parameter grid that has been provided to the gridsearch includes a large amount of parameters that each has a broad array of grid values that the search can run through. Basically, the gridsearch runs through all the possible combinations of parameter values from the grid and evaluates the optimal parameter combination based on a chosen evaluation metric. There are a total of 5184 potential combinations of parameter values and we have implemented a 5-fold cross validation method for testing the optimal parameters which leads to a total of 25920 since the model is being trained and evaluated on each of the 5 different datasplits.

We have used the logloss evaluation metric for the search, which is a good metric when dealing with imbalanced data since logloss does not get fooled by always predicting the majority class (as accuracy would have) since logloss exposes very confident predictions that are wrong by penalizing these harshly.

Below is an overview of the grid that was chosen as well as the values that the gridsearch found to be optimal parameters for the model:

Parameter	Array of grid values	Optimal parameter value based on gridsearch	Parameter description
learning_rate	[0.01, 0.05, 0.1]	0.1	The contribution of each tree (the lower the more cautious the model)
n_estimators	[100, 300, 500]	300	Number of trees (boosting rounds)
max_depth	[3, 5, 6, 8]	8	The maximum dept of each tree (higher dept allows for more complex splits)
subsample	[0.6, 0.7, 0.8, 1]	1	The fraction of data that is used per tree (a 100% of the data is used per tree)
colsample_bytree	[0.8, 1]	1	How high a percentage of features that are to be used per tree. (all features are used per tree)
gamma	[0, 0.05, 0.1]	0	The minimum loss reduction that is required for a split. It is zero so the trees are very easily split
min_child_weight	[1, 2, 3, 4, 5, 7]	2	Minimum sum of instance weights in a child

4.3.3 Multilayer Perceptron - Deep Neural Network

To explore whether a high-capacity model can outperform the tree-based baselines, we trained a fully connected Multilayer Perceptron (MLP). Neural networks are widely used in educational data mining because they can capture the non-linear interactions that often underlie student success and failure (Shahiri et al., 2015; Srivastava et al., 2014; Xing & Du, 2019). Our method followed the recommendation of these recent dropout studies, combined with our own testing of what improved model performance.

We kept all 27 engineered features and allowed the hidden layers to discover which ones matter most, rather than pruning the input beforehand. Thus, the final MLP consists of an input layer with 27 neurons, two hidden layers of 128 and 64 neurons with ReLU activation, and a single sigmoid output that returns the dropout probability. Batch-normalisation layers follow each hidden layer to stabilise internal feature distributions, and dropout layers (rates 0.4 then 0.3) randomly disable neurons during training to reduce overfitting. We optimised the weights with the Adam algorithm because its adaptive learning rates speed convergence on tabular data (Kingma & Ba, 2015). The learning rate was set to 0.0005 and the model learned from mini-batches of 32 observations. A stratified 80/20 split preserved the original class imbalance, and 10% of the training fold served as a validation set. Early-stopping halted training when validation loss failed to improve for ten epochs, so runs typically finished after roughly 30 epochs which was well before the 100-epoch cap.

On the unseen test fold the network achieved 0.87 accuracy, 0.90 macro-F1, and an AUC of 0.82, outperforming the simpler baseline while maintaining recall above 0.90 for the dropout class. Training on a single GPU took under two minutes, and inference for a new student is effectively instantaneous. Taken together, the regularised MLP provides a practical yet flexible early-warning engine that complements the more interpretable tree models in our study.

4.4 Evaluation Metrics

The evaluation metric that has been used for the analysis in this paper is the f1-score which is combining the metrics, recall and precision. Since our dataset is imbalanced the accuracy metric would have been unsuitable to use for the evaluation of our models. With imbalance data our prediction models have a risk of being biased by favoring the "graduate" target variable and achieving a high accuracy by simply going with this class since it is the most frequent one. Because of this, F1-Score is being favored.

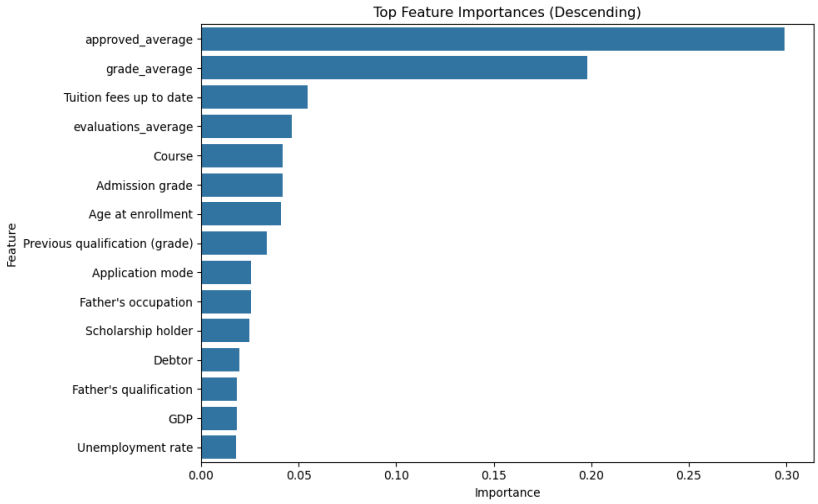
Since the F1-Score relies on precision and recall it is important to also understand how these metrics work. The precision metric measures how many of all the predicted positives (label = 1) were actually positive, whereas the recall metric measures how many of all the true positives that were correctly predicted as positive. Thus, precision is good for highlighting false positives and recall is good for highlighting false negatives, and these two measures then have a trade off between what should be the focus for evaluation of the model.

This harmonic mean is a way of mitigating the trade-off between the precision and recall and it balances false positives and false negatives in the model. Because of this, we deem the F1-Score best fit for evaluating our models due to its harmonic approach that allows for maximising both precision and recall through a maximisation of the F1-score.

In terms of evaluating the F1-Score, both a macro average and a micro-average is provided in the classification reports in our models. Here we choose to focus on the macro-average which calculates the F1 for each class (dropout and graduate) individually and then scores the average of those two. Thus, it treats the classes equally even though their representations in the dataset are imbalanced. The micro average is basically an average of the F1 score that is proportional to the support. However, we want both graduate and dropout to matter equally and will therefore use the macro average.

5. Results

The random forest was utilized for extracting the 15 most important features, and the feature importance can be visualized as below:



From this we see that the most important features are the average of curricular units approved (1st and 2nd semester) as well as the average of the grades, which makes a lot of sense since this is a direct indication that the student has passed an important milestone towards graduating when having finished 2nd semester and the average of the grades that the student achieved during those two semesters will also have a high impact on the student's change of graduating. The grade that the student had from previous education is also a relatively important factor. A less important but still relevant factor is also whether the student is a scholarship holder. Thus, the variables pertaining to the students academic performance plays the highest role for determining whether the student will successfully graduate or dropout.

Furthermore, we see that the student's previous grade (before enrollment) plays a relatively important role, and that the father's occupation also seems to have an important impact. This suggests that there is an element of social heritage in the student's chances of graduating successfully.

Thus, academic performance (curricular units and grades), social heritage and macroeconomic- factors and debt are the most important features in the model.Based on the table below, we notice that both XGBoost and Random Forest exhibit highly comparable performance, each achieving a macro F1-score of 0.90. While Random Forest slightly outperformed XGBoost in the graduate class (recall: 0.97 vs. 0.95), XGBoost achieved a higher recall in the minority dropout class (recall: 0.83 vs. 0.81). Given that the primary goal of this study is the early identification of at-risk students, this makes XGBoost the preferred model. In conclusion, we consider XGBoost to be the best-performing model due to its better handling of class imbalance and stronger performance in identifying students at risk of dropping out.

The MLP model demonstrated the weakest performance (Macro F1 = 0.85). Although it performed well in the “graduate” class（recall: 0.96), its ability to identify “dropout” students was limited, achieving a recall of only 0.72.

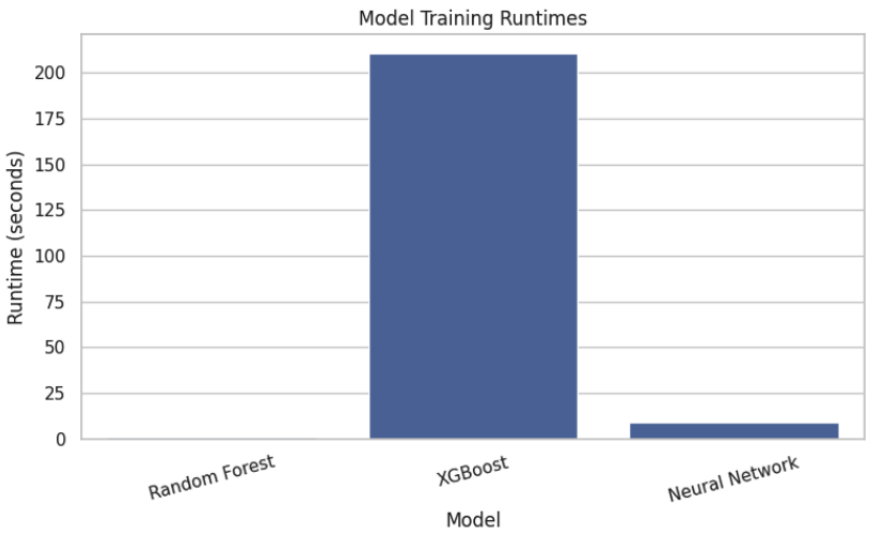
Model	Class	Precision	Recall	F1-Score
Random Forest (Macro F1 = 0.90)	Dropout	0.94	0.81	0.87
	Graduate	0.89	0.97	0.93
XGBoost (Macro F1= 0.90)	Dropout	0.92	0.83	0.87
	Graduate	0.90	0.95	0.93
MLP (Macro F1 = 0.85)	Dropout	0.92	0.72	0.81
	Graduate	0.84	0.96	0.90

The confusion matrix in the Appendix further supports the above conclusion. XGBoost correctly identified 237 dropout students and misclassified only 47, making it the best performing model in this aspect. Random Forest correctly identified 231 dropout students and misclassified 53, thus slightly underperforming compared to XGBoost. In contrast, the MLP model demonstrated the weakest ability to detect dropout cases, correctly identifying only 204 students while misclassifying as many as 80. Overall, these results suggest that XGBoost is the most suitable model for the early identification of high-risk dropout students. Detailed classification results for each model are presented in Table.

Model	Class	Correctly Classified	Misclassified
Random Forest	Dropout	231	53 (→ Graduate)
	Graduate	428	14 (→ Dropout)
XGBoost	Dropout	237	47 (→ Graduate)
	Graduate	421	21 (→ Dropout)
MLP	Dropout	204	80 (→ Graduate)
	Graduate	425	17 (→ Dropout)

5.1 Complexity & Running Time Analysis

All the models have been run on 376,53 RAM and 64v GPU's and their run times can be illustrated in the visual below:



Based on this we see that XGboost is by far the most computationally heavy model and takes over 200 seconds to run. The neural network has a runtime of 9 seconds whereas random forest is taking below 5 seconds. The random forest is by far the most simple model out of the 3 and this is clearly illustrated when we consider run times of our 3 models. The neural network's increased runtime compared to the random forest is not significant

and this is also expected due to the fact that it only has 14.337 trainable parameters based on the amount of neurons added to each hidden layer.

6. Discussion

6.1 Comparison of the Models

XGBoost and Random Forest demonstrate comparable performance in macro-F1 score, both showing strong predictive capabilities, particularly for the graduate category. This result is consistent with previous research findings (Villar & de Andrade, 2024), further validating the effectiveness of ensemble tree models in handling imbalanced educational data.

While XGBoost holds a slight advantage in precision for this category, both models are practically feasible for implementation. When controlling false positives is a key metric (such as in graduate misclassification scenarios), XGBoost shows superior performance. Meanwhile, Random Forest's faster training speed makes it more suitable for big data iterative scenarios. In such cases, its computational efficiency may outweigh minor performance differences.

Contrastly, MLP's macro-F1 score is lower, especially in dropout category recall, exposing limitations in its feature abstraction capability, likely due to insufficient network depth or feature encoding deficiencies. Although performance might be improved through deepening the architecture or adding more data, results still support XGBoost as the preferred solution. So our study focuses on XGBoost optimization, employing grid search (covering parameters like learning rate, tree depth, regularization, and subsampling) and sample weighting strategies to address class imbalance with macro-F1 as the optimization target.

6.2 Error Analysis

Model error pattern reveals that all three models exhibit false negative problems in the "Dropout" category, systematically misclassifying numbers of actual dropout cases as graduates. Quantitatively, the MLP demonstrates the most pronounced misclassification (80 instances), followed by Random Forest (53 instances), with XGBoost showing the best performance (47 instances). From an ethical perspective, false negatives in the "Dropout" category may lead to missed opportunities for timely intervention, thereby exacerbating educational inequality. This issue highlights the real-world impact of model errors beyond accuracy metrics.

It is widely accepted that such errors are largely caused by ambiguity in class boundaries, meaning that the 'Dropout' and 'Graduate' samples share considerable overlap in key feature dimensions. This overlap makes it difficult for classifiers to make definitive judgments based on limited rules. For tree-based models, this feature overlap directly affects their ability to achieve "partition purity." Both XGBoost and Random Forest construct classification boundaries by recursively splitting features, relying on clear threshold divisions that generate information gain. However, when multiple feature dimensions fail to provide stable distinguishing signals, leaf nodes end up containing mixed samples from different classes, resulting in blurred decision boundaries and ultimately leading to misclassification. This aligns theoretically with the research by Domingos (1999) and Hooker, G. (2021), which shows that decision tree-based models are highly sensitive to boundary ambiguity and class overlap, particularly tending to favor the majority class when inter-class differences are indistinct.

Furthermore, our current data has limitations in feature representation, lacking time-series features that could reflect dynamic changes in student behavior. Research has shown that the continuity of student behavior over

time is a critical factor in identifying dropout risks (Baker & Yacef, 2009; Brooks et al., 2015). Incorporating such time-series data could enhance the model's ability to detect patterns that static features may overlook, improving predictive accuracy for dropout prediction.

To reduce errors caused by ambiguous class boundaries, future research could optimize the model through multiple approaches. First, introducing time-series features represents a critical breakthrough. For example, temporal trends in behavioral indicators like attendance rates, course completion rates, and classroom engagement levels may better reflect students' true learning states than static numerical values, thereby improving the model's sensitivity in identifying hard-to-classify samples. In terms of modeling methods, more complex loss functions such as focal loss could be considered. By dynamically adjusting sample weights, it focuses the learning process on hard-to-classify samples, particularly enhancing the model's ability to recognize marginal classes. Finally, to further mitigate potential biases from single models when handling boundary samples, future work could explore multimodal fusion strategies like soft voting to integrate judgments from multiple models, maintaining overall performance while improving robustness in ambiguous decision regions.

7. Conclusion & Future Work

This study compared the performance of three classification models: Random Forest, XGBoost, and MLP to predict student dropout based on a publicly available Portuguese higher education dataset.

We focused on a binary outcome to improve interpretability and practical relevance. The results showed that XGBoost performed best, particularly in identifying students at risk of dropping out. It achieved a recall of 0.83 for the dropout class and a macro-average F1 score of 0.90. Although Random Forest delivered similar overall results and trained more quickly, it was slightly less precise in detecting dropout cases, which may limit its value in early warning systems. MLP, while theoretically capable of capturing non-linear relationships, performed the worst in this task, with the highest misclassification rate and the lowest recall for the dropout class, suggesting that it is not well suited for this dataset.

Feature importance analysis reveals that students' likelihood of graduation is shaped by a combination of academic, socioeconomic, and macroeconomic factors. Academic performance during the first year emerged as the most significant predictor of graduation. This is intuitive, as strong performance in the first two semesters reflects a solid academic foundation. Other important factors include the course of study and age at enrollment, while prior academic achievement and father's occupation also contribute, suggesting the influence of social background. Additionally, macroeconomic indicators such as GDP also play meaningful roles. These insights can help universities develop targeted interventions. For example, institutions could offer academic or financial support to struggling first-year students or those from disadvantaged backgrounds.

Future work should address the false negative issue, which is likely caused by ambiguity in class boundaries and significant overlap in feature distributions. One promising direction is to add time-series behavioral features, such as attendance trends which could provide a more comprehensive representation of students' academic trajectories than static variables. Additionally, implementing optimization techniques that address

class boundary ambiguity, such as focal loss or cost-sensitive learning, may enhance the model's ability to handle hard-to-classify cases and improve detection of students from minority outcome groups. Ensemble methods such as soft voting or model stacking could further enhance robustness. Lastly, expanding the dataset to include multiple institutions or academic years would support testing the generalizability of the approach and ensure fair and effective deployment in diverse educational context.

References

Martins, T. C., Correa, C., Oliveira, F. L., & Iglesias, Í. L. (2021). Comunicação pública para a compreensão de políticas culturais: O software Elum no Rio Grande do Sul, Brasil. Revista Internacional de Comunicação e Cultura, 14(1), 1–20.

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. Data, 7(11), 146. https://doi.org/10.3390/data7110146

Shamim, A. (2023). Predict students' dropout and academic success [Data set]. Kaggle. https://www.kaggle.com/datasets/adilshamim8/ predict-students-dropout-and-academic-success/data

Öztürk, Ç. (2023). Predicting student dropout using multiclass classification. Tilburg University Repository. https://arno.uvt.nl/ show.cgi?fid=181524

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. Discover Artificial Intelligence, 4(2), Article 2. https://doi.org/10.1007/s44163-023-00079-z

Kundu, R. (2022). F1 Score in Machine Learning: Intro & Calculation. V7. https://www.v7labs.com/blog/f1-score-guide

Schafi, A. (2020). Random Forest Classification with SciKit Learn. Datacamp. https://www.datacamp.com/tutorial/random-forests-classifier-python

Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99), 155-164. https://doi.org/10.1145/312129.312220

Hooker, G. (2021). Unbiased measurement of feature importance in tree-based methods. Annals of Statistics, 49(6), 3368-3392. https://doi.org/10.1214/21-AOS2073

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1). https://jedm.educationaldatamining.org/index.php/JEDM/issue/view/1

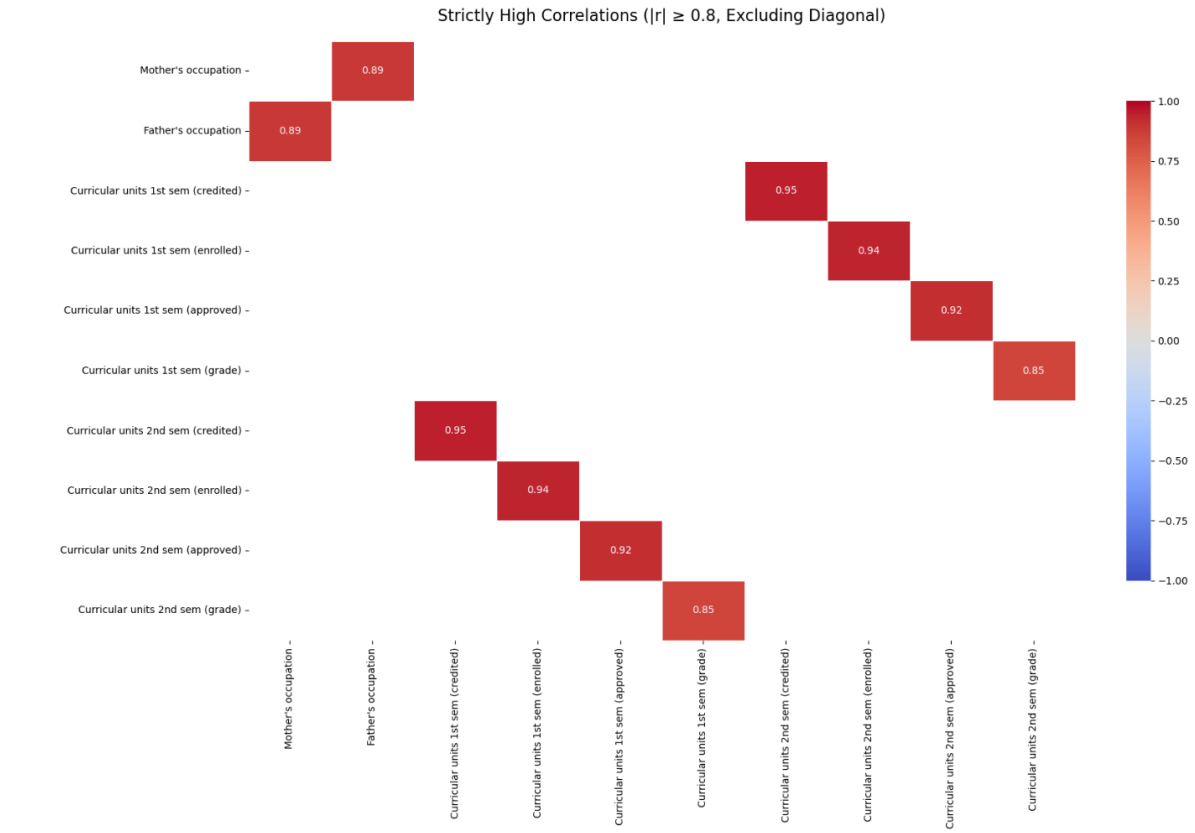
Brooks, C., Thompson, C., & Teasley, S. (2015). A time series interaction analysis method for building predictive models of learners. LAK '15: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge. https://doi.org/10.1145/2723576.2723580

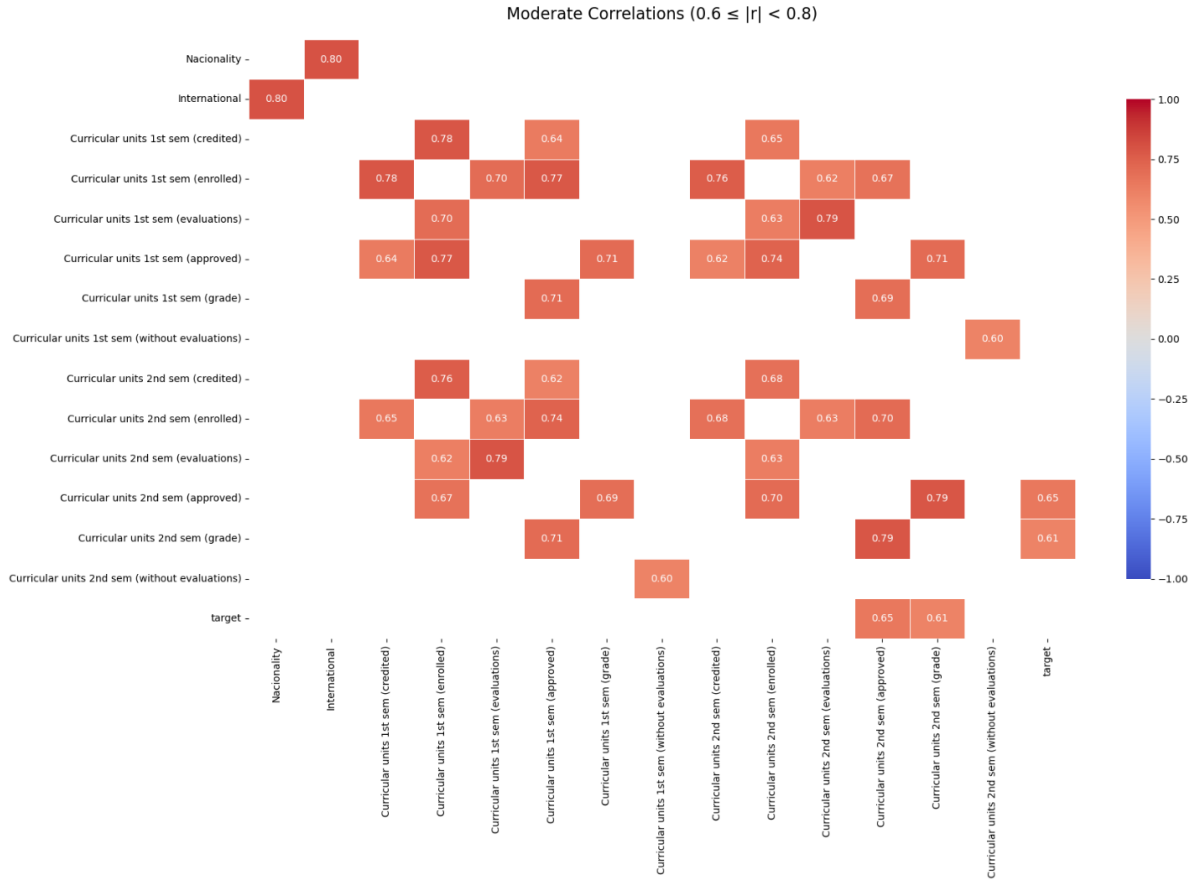
Appendix

Appendix 1 - Data Overview

Domain	Typical columns	Notes
Demographics	Gender, Age at enrollment, Marital Status, Nationality	Static information captured at matriculation.
Socio-economic & financial	Scholarship holder, Tuition fees up to date, Debtor, parents' qualification and occupation	Proxy for household resources and financial stressors.
Academic path at entry	Course, Application mode, Application order, Daytime/evening attendance, Previous qualification (+ grade)	Describe how and where the student was admitted.
First-year performance	Course, Application mode, Application order, Daytime/evening attendance, Previous qualification (+ grade)	Provide an early snapshot of academic engagement.
Macroeconomic context	Course, Application mode, Application order, Daytime/evening attendance, Previous qualification (+ grade)	Year-specific national indicators matched to each enrolment cohort.

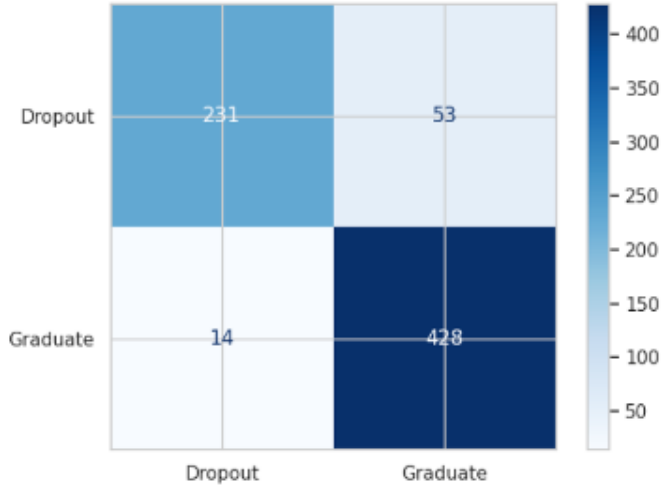
Appendix 2 - Correlation matrix



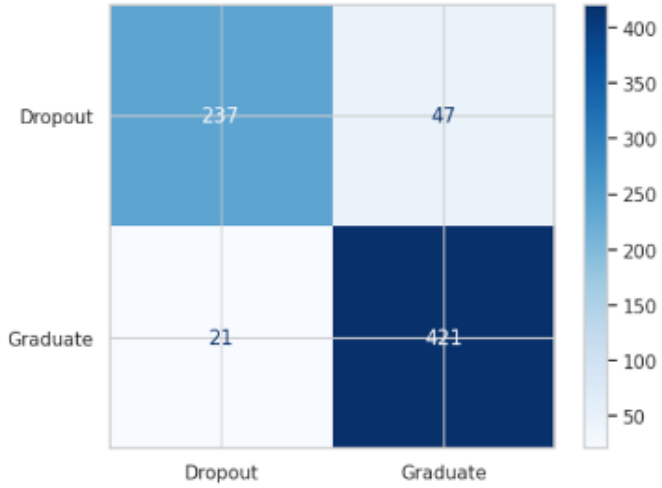


Appendix 3 - Confusion Matrices

Random Forest Confusion Matrix



XGBoost Confusion Matrix



MLP Confusion Matrix