Predicting Students' Dropout and Academic Success

Arsalan Khan (210862640), Aliha Ali (210184090)
Wilfrid Laurier University

Abstract
In higher education, timely identification of at-risk students is crucial for improving retention and academic success rates. By predicting whether a student is likely to drop out, remain enrolled, or graduate, educational institutions can allocate support resources more effectively. This study employs a comprehensive dataset from a Portuguese higher education institution, integrating demographic, socioeconomic, and academic performance attributes. We preprocess the data, address class imbalance, conduct exploratory data analysis, and evaluate multiple machine learning models. Our results highlight the Random Forest classifier as the top performer, achieving an accuracy of approximately 81%, thus providing actionable insights into early intervention strategies and improving the overall student experience.

1 Introduction

Improving student retention and promoting academic success are key priorities for higher education institutions worldwide. Early identification of students who are at risk of dropping out enables universities to provide timely interventions—academic tutoring, financial support, and personalized counseling—that can significantly improve outcomes. The dataset used in this project, drawn from a Portuguese higher education institution,[1] includes a rich mixture of demographic, socioeconomic, and academic performance information, making it suitable for building predictive models.

The primary objective is to forecast student outcomes—classified as *Dropout*, *Enrolled*, or *Graduate*—using a multi-class classification approach. To achieve this, we implement a structured machine learning pipeline. Our key goals include:

- Data preparation and preprocessing to ensure analytical readiness.
- Exploratory Data Analysis (EDA) to uncover patterns and relationships.
- Class imbalance handling to ensure model fairness and robust prediction capabilities.
- Model development and comparison, testing Logistic Regression, Decision Tree, Random Forest, SVM, and Neural Network classifiers.
- Identifying the most influential features and refining models through hyperparameter tuning.
- Deriving insights that can guide early interventions, reducing dropout rates and improving graduation outcomes.

This paper proceeds by detailing our data exploration and preprocessing steps, followed by EDA, class imbalance handling, feature importance analysis, model building, and evaluation. We then discuss hyperparameter tuning, challenges encountered, key findings, and conclude with insights and suggestions for future work.

2 Data Exploration and Preprocessing

The dataset, sourced from the UCI Machine Learning Repository,[1] contains variables describing student demographics, parents' education and occupation, academic records, and economic indicators. The target variable indicates whether a student eventually dropped out (0), remained enrolled (1), or graduated (2).

2.1 Data Cleaning and Inspection

After loading the dataset, we standardized column names for consistency. No missing values were identified, ensuring a complete dataset. Summary statistics and initial inspections (e.g., `head()`, `info()`, `describe()`) provided an overview of distributions and data types.

2.2 Categorical Encoding

Categorical features such as *Marital_status*, *Course*, and *Gender* were label-encoded. The target variable was also transformed, mapping *Dropout* → 0, *Enrolled* → 1, and *Graduate* → 2. This encoding step ensured compatibility with machine learning algorithms.

2.3 Feature Scaling

Continuous variables, including *Age_at_enrollment*, *Admission_grade*, and *Previous_qualification*, were standardized to ensure all features contributed proportionally to the model and to prevent dominance by any one feature.

3 Exploratory Data Analysis (EDA)

EDA was conducted to understand data distributions, relationships between variables, and their association with the target classes.

3.1 Univariate Analysis

A class distribution plot revealed an imbalance among the three categories: graduates form the largest group, followed by dropouts, and then enrolled students. Such an imbalance can bias the predictive model if not addressed, prompting the use of resampling techniques later in the pipeline.

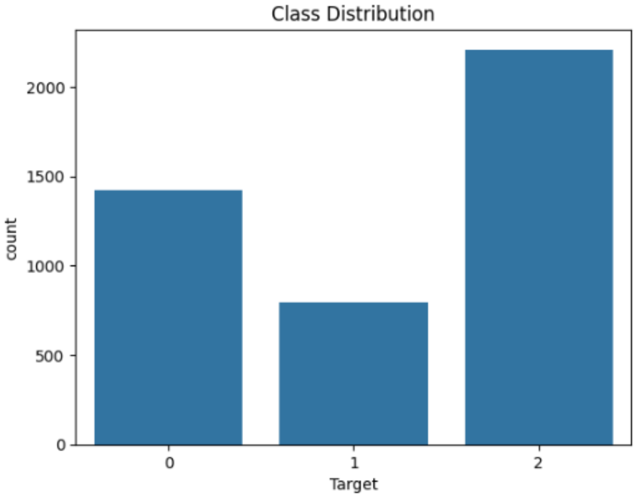

Figure 1: Class Distribution of the Target Variable. Graduates form the largest class, followed by Dropouts and Enrolled, indicating an imbalanced dataset.

3.2 Bivariate Analysis

The correlation matrix highlighted clusters of highly correlated features, especially among *Curricular_units* variables. This suggests that academic performance metrics (e.g., credits, evaluations, and approved grades) are interconnected and likely influential in determining student outcomes.

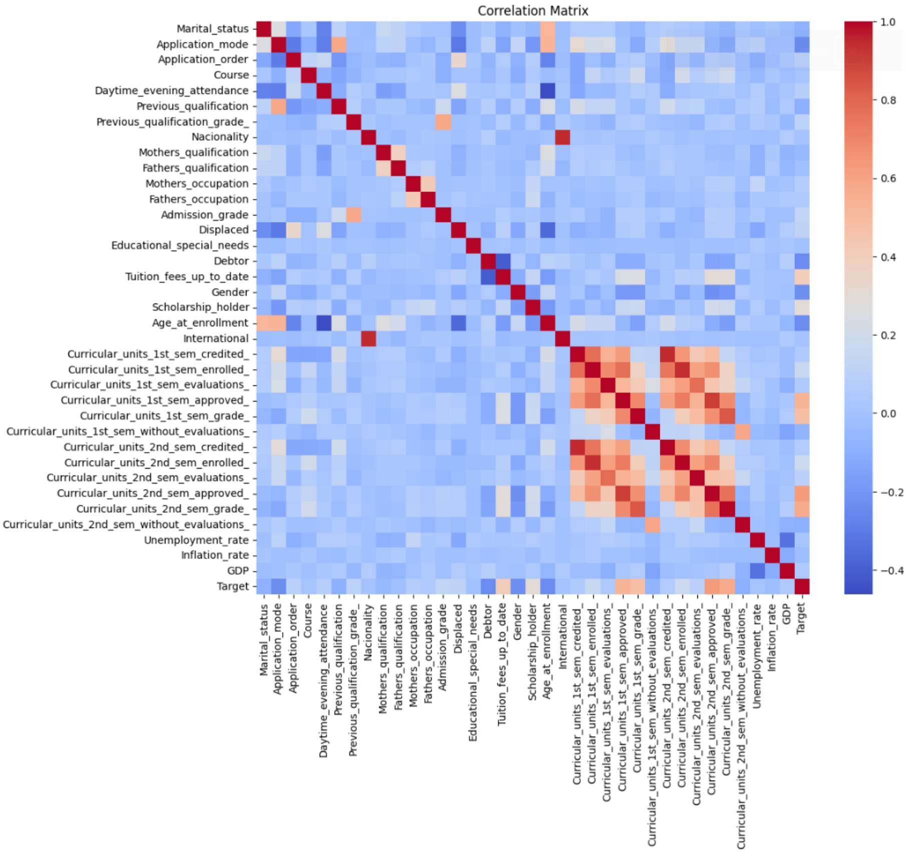

Figure 2: Correlation Matrix of Features. Blocks of high correlation among academic performance indicators suggest their strong influence on final outcomes.

3.3 Feature Relationships

A boxplot comparing *Admission_grade* across target classes showed that while the overlap is substantial, a slightly higher median admission grade is noted among graduates, indicating admission performance may offer a subtle predictive signal.

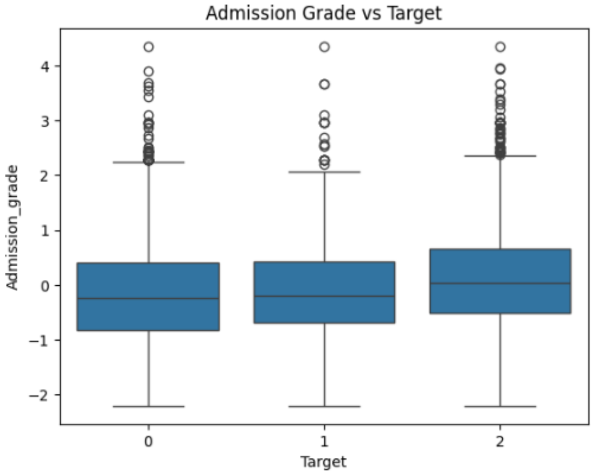

Figure 3: Admission Grade vs. Target. Graduates tend to have marginally higher admission grades, though distributions overlap significantly.

4 Handling Class Imbalance

Initial class distribution analysis indicated a skewed target variable, with the majority of students being graduates. Without addressing this imbalance, models might underpredict minority classes, reducing their usefulness.

We applied SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic examples of the minority classes (Enrolled), achieving a balanced dataset. Post-resampling, the model had a fair chance to learn from all classes, improving recall and overall performance.

5 Feature Engineering and Selection

To identify the most influential features, a Random Forest model was initially trained, and feature importance scores were extracted. Top predictors included:

- *Curricular_units_2nd_sem_approved_*
- *Curricular_units_1st_sem_grade_*
- *Curricular_units_1st_sem_approved_*
- *Curricular_units_2nd_sem_grade_*
- *Tuition_fees_up_to_date*
- *Age_at_enrollment*

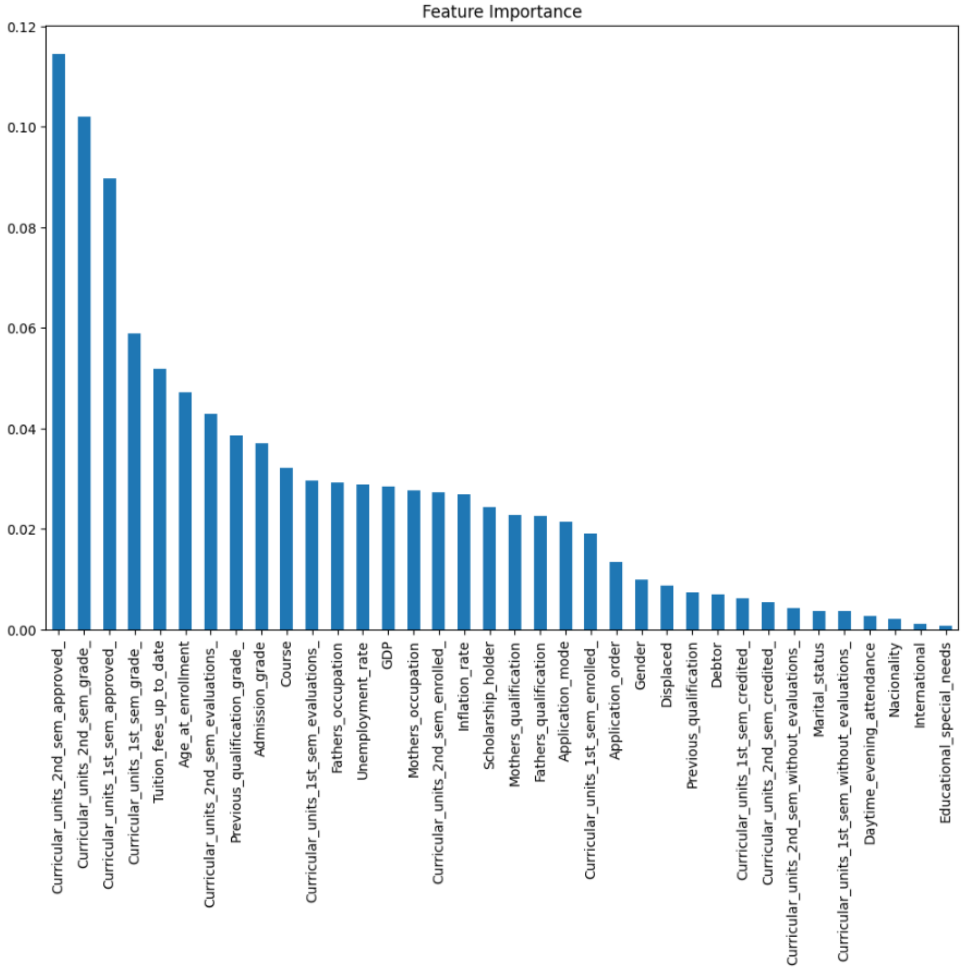

Figure 4: Feature Importance from Random Forest. Academic performance metrics and financial stability are top predictors.

These findings underscore the importance of academic performance in the first and second semesters, as well as financial stability, in shaping final outcomes.

6 Model Building

After splitting the balanced dataset into training and testing sets (80/20), several classification models were trained and evaluated.

6.1 Basic Models

Logistic Regression provided a strong baseline (approximately 74% accuracy), offering straightforward interpretability. **Decision Tree** delivered a comparable performance (73%), but tended to overfit, reducing its generalization capability.

6.2 Advanced Models

Random Forest emerged as the top-performing model, achieving roughly 81% accuracy. Its ensemble nature reduced overfitting and improved recall and precision across all classes.

To better understand each model's performance, confusion matrices were examined right after they were discussed:

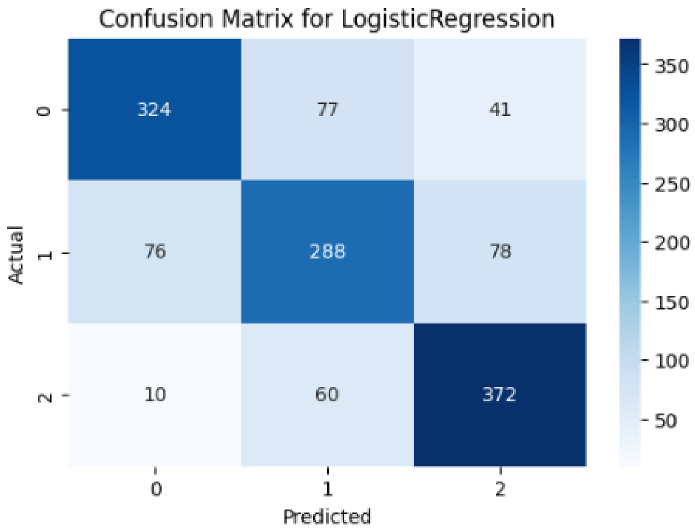

Figure 5: Confusion Matrix for Logistic Regression. Balanced but moderate performance.

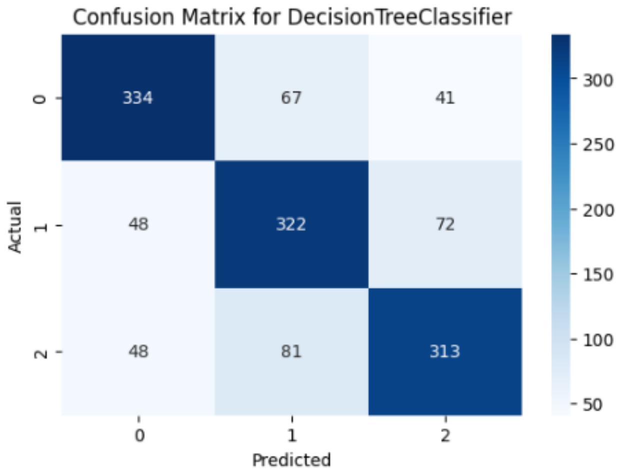

Figure 6: Confusion Matrix for Decision Tree Classifier. Slight improvements in some areas compared to LR, but still limited.

Support Vector Machine (SVM) struggled, achieving 66% accuracy, likely due to complex feature interactions and initial class imbalance challenges. **Neural Network (MLPClassifier)** offered about 75% accuracy, capturing non-linear relationships but not surpassing Random Forest without further tuning.

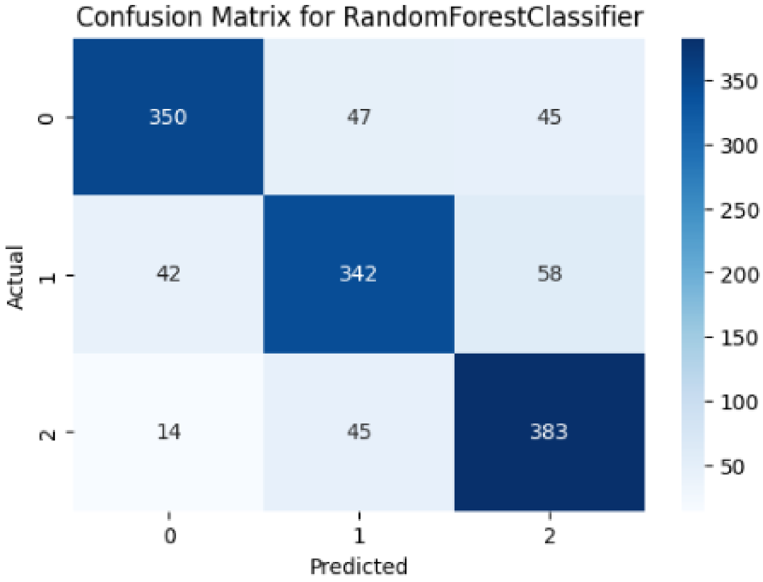

Figure 7: Confusion Matrix for Random Forest Classifier. Shows balanced predictions across classes.

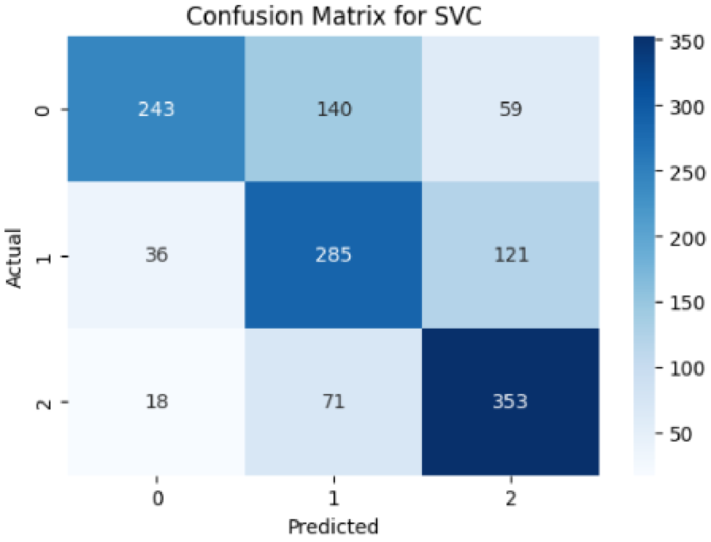

Figure 8: Confusion Matrix for SVM. Lower accuracy and imbalance in predictions.

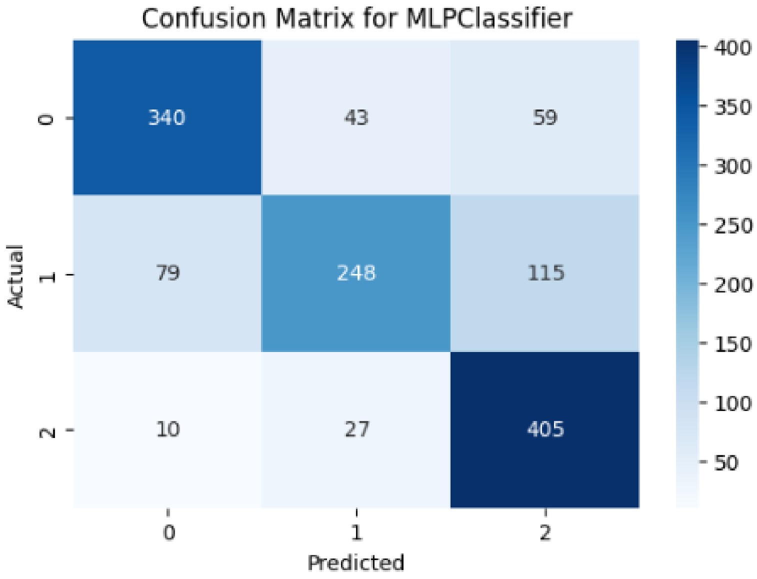

Figure 9: Confusion Matrix for MLP Classifier. Improved over basic models but not better than RF.

7 Model Evaluation

In addition to confusion matrices, we examined precision, recall, and F1-scores for each model and class. Table 1 summarizes the results. The ”Conclusion” column provides a brief interpretation of what the scores suggest for each class and overall model performance.

Table 1: Comparison of Precision, Recall, and F1-scores for Each Model and Class

Model	**Class**	**Precision**	**Recall**	**F1-score**	**Conclusion**
Logistic Regression	0 (Dropout)	0.79	0.73	0.76	Decent at identifying dropouts, but may miss some at-risk students.
	1 (Enrolled)	0.68	0.65	0.66	Struggles slightly with enrolled students, moderate balance.
	2 (Graduate)	0.76	0.84	0.80	Strong in identifying graduates accurately.
	Macro Avg	0.74	0.74	0.74	Overall balanced performance, no class heavily favored.
	Accuracy	0.74			Solid baseline, but room for improvement.
Decision Tree	0 (Dropout)	0.78	0.76	0.77	Similar to LR, reliable on dropouts, slight improvement in recall.
	1 (Enrolled)	0.69	0.73	0.71	Slightly better at enrolled than LR, improving recall.
	2 (Graduate)	0.73	0.71	0.72	Balanced for graduates, still not top-tier.
	Macro Avg	0.73	0.73	0.73	Overall similar to LR, slightly more balanced.
	Accuracy	0.73			Comparable baseline, marginal improvement in some classes.
Random Forest	0 (Dropout)	0.86	0.79	0.83	Strong precision, good at identifying dropouts accurately.
	1 (Enrolled)	0.79	0.77	0.78	Much improved on enrolled, more balanced detection.
	2 (Graduate)	0.79	0.87	0.83	Excellent at recognizing graduates, high recall.
	Macro Avg	0.81	0.81	0.81	Best overall balance and high performance across classes.
	Accuracy	0.81			Top performer, most reliable model tested.
SVM	0 (Dropout)	0.82	0.55	0.66	Good precision but low recall means it misses many dropouts.
	1 (Enrolled)	0.57	0.64	0.61	Struggles with enrolled; low precision affects reliability.
	2 (Graduate)	0.66	0.80	0.72	Better at graduates, but still not as good as RF.
	Macro Avg	0.69	0.66	0.66	Weaker overall, imbalanced detection of classes.
	Accuracy	0.66			Lowest accuracy; less suitable for balanced predictions.
Neural Network (MLP)	0 (Dropout)	0.79	0.77	0.78	Similar to LR/DT, reasonably strong on dropouts.
	1 (Enrolled)	0.78	0.56	0.65	Improved precision for enrolled but recall still low.
	2 (Graduate)	0.70	0.92	0.79	Very good at catching graduates, high recall stands out.
	Macro Avg	0.76	0.75	0.74	Good balance, though not surpassing RF.
	Accuracy	0.75			Improvement over baseline models, second best after RF.

Precision, recall, and F1-scores from classification reports, along with confusion matrices (Figures 5 through 9), provided a nuanced understanding of each model's strengths and weaknesses. The Random Forest model's robust performance and balanced predictions support its selection as the primary predictive tool.

8 Hyperparameter Tuning

A Grid Search was performed on the Random Forest model to refine hyperparameters (*n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*). This tuning process slightly improved performance.

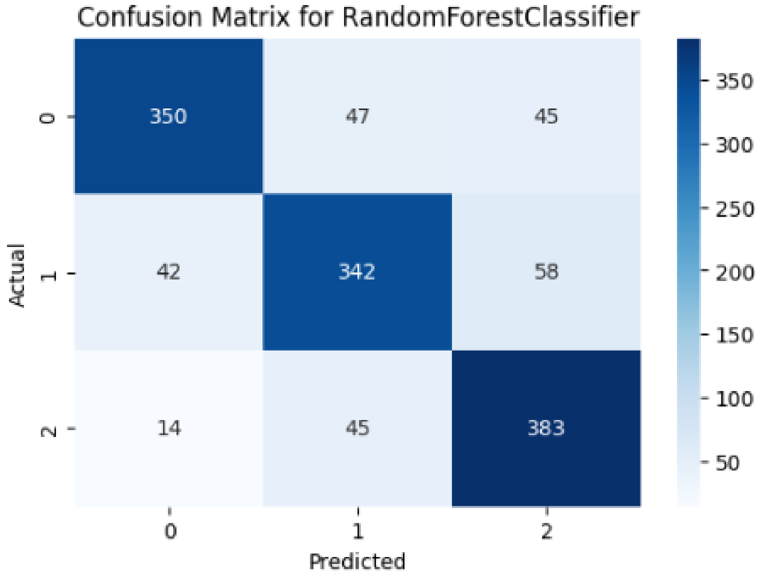

Figure 10: Confusion Matrix for Tuned Random Forest Classifier. Further refined performance after hyperparameter optimization.

9 Challenges and Considerations

- **Class Imbalance:** Without SMOTE, the minority class (Enrolled) would remain underrepresented.
- **Feature Selection Complexity:** Determining which performance metrics matter most required careful analysis. Advanced feature engineering could further improve results.
- **Generalization:** Models trained on a single institution's data might not generalize to different educational contexts or regions.
- **Data Quality:** Although no missing values were present, interpreting certain categorical encodings demanded domain expertise.

10 Key Findings

The results of this study highlight several critical insights:

- **Early Academic Performance is Crucial:** Metrics related to curricular units, particularly those approved in the first and second semesters, emerged as strong predictors. Students lagging in these areas may require immediate support to prevent dropout.
- **Financial Stability Matters:** The *Tuition_fees_up_to_date* feature significantly influenced outcomes, suggesting that timely fee payments correlate with consistent academic engagement and eventual success.
- **Baseline Models vs. Advanced Techniques:** Logistic Regression and Decision Tree models offered baseline performance, but advanced ensemble methods like Random Forest provided more balanced and accurate predictions.
- **Impact of Class Imbalance Handling:** Employing SMOTE to balance classes improved the ability of models to identify at-risk students (e.g., those who might drop out), ensuring more equitable predictions.
- **Incremental Gains from Hyperparameter Tuning:** While Random Forest performed well initially, careful tuning further enhanced its predictive ability, demonstrating the value of model refinement.

11 Conclusion

This study demonstrates a comprehensive, data-driven approach to predicting student outcomes in higher education. By addressing class imbalance, evaluating multiple models, and focusing on academic performance indicators, we identified the Random Forest classifier as the most effective predictive model, achieving approximately 81% accuracy. Students struggling in the first and second semesters, or those with delayed tuition payments, represent clear intervention points.

These insights enable more proactive strategies, such as early academic counseling and financial support, to reduce dropout rates and enhance overall student success. The methodologies and findings presented here offer a starting point for continued refinement, as educational institutions seek data-driven approaches to improving academic outcomes.

12 Future Work

Future directions include:

- **Comprehensive Tuning:** Applying grid search and other optimization methods (e.g., Bayesian optimization) to SVM and Neural Networks.
- **Advanced Models:** Experimenting with Gradient Boosting methods (XGBoost, LightGBM) or ensemble stacking.
- **Feature Engineering:** Incorporating attendance records, engagement metrics, or external socioeconomic indicators.
- **Explainability Tools:** Utilizing SHAP or LIME to understand model predictions at an individual level.

- **Early Prediction:** Focusing on earlier semesters to provide timely interventions and prevent dropout escalation.

References

[1] UCI Machine Learning Repository: Predict Students' Dropout and Academic Success Dataset. https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

[2] Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). *Early prediction of student's performance in higher education: a case study*. Trends and Applications in Information Systems and Technologies.