UNIVERSITY OF PISA

Master's Degree in Artificial Intelligence and Data Engineering

Data Mining and Machine Learning

Predicting Student Dropout in Higher Education Using Supervised Learning

Candidate:

Martina Fabiani

Project GitHub Repository: https://github.com/martiFabia/DMML_project

ACADEMIC YEAR 2024/2025

Contents

1 Introduction

Predicting student dropout and **academic performance** represents a critical challenge for higher education institutions due to its significant impact on students, institutions, and broader society. Dropout and educational failure adversely affect economic growth, employment rates, competitiveness, and productivity, while profoundly impacting the personal and professional lives of students and their families.

This project addresses these challenges by leveraging data mining and machine learning techniques to analyze and predict **student behaviors**. The dataset utilized originates from diverse administrative sources and includes demographic, socioeconomic, macroeconomic, and academic performance data of students enrolled in various undergraduate programs at a **Portuguese higher education institution**.

The core objective of this initiative is to build robust supervised learning models capable of accurately **classifying students** into categories of graduates, currently enrolled, or at risk of dropout. Beyond predictive accuracy, the project aims to identify critical factors that significantly influence academic outcomes, thus enabling timely, targeted, and personalized interventions. Such predictive analytics form the foundation of an early warning system designed to provide educators and tutoring teams with actionable insights to mitigate dropout rates and enhance educational success.

By proactively identifying at-risk students, educational institutions can implement tailored support strategies, improve curriculum design, and ultimately foster better academic outcomes and societal integration.

2 Dataset

The dataset [1] analyzed in this study is essential for understanding student performance and predicting dropout risks, facilitating targeted and effective educational interventions. The comprehensive and multi-source nature of the dataset allows for a thorough investigation into the factors influencing student success.

2.1 Dataset Description

The dataset covers a range of factors that may influence a student's academic success, including demographic, socio-economic, and academic factors.

The data contains variables related to ***demographic factors*** (age at enrollment, gender, marital status, nationality, address code, special needs), ***socio-economic factors*** (the educational and employment background of their parents, whether they received a student grant or have student debt) ***student's academic path*** (admission grade, their admission grade, the order of choice for their enrolled course, and the type, of course, they took in high school) and data at the end of the first and second semesters.

The dataset covers student records from the academic years 2008/2009 to 2018/2019. It includes data from **17 undergraduate degree** programs across diverse fields such as agronomy, design, education, nursing, journalism, management, social services, and technology. The final dataset consists of **4424 records** with **37 attributes**, where each record represents an individual student, and contains no missing values.

2.2 Dataset Distribution

Each student record is classified into three categories based on the duration taken to complete their degree: Graduated, Enrolled, and Dropout. ***"Graduated"*** indicates students who obtained their degree within the expected timeframe, ***"Enrolled"*** refers to students who required up to three additional years, and ***"Dropout"*** represents students who took more than three extra years or did not graduate at all. These categories correlate with varying risk levels, where "Graduated" represents low-risk students, "Enrolled" indicates medium-risk students potentially benefiting from targeted institutional interventions, and "Dropout" highlights high-risk students with the greatest likelihood of failing academically.

The dataset is imbalanced, which presents significant challenges for predictive modeling. This imbalance can bias predictive models toward the majority class, potentially reducing their accuracy in effectively identifying and supporting high-risk students.

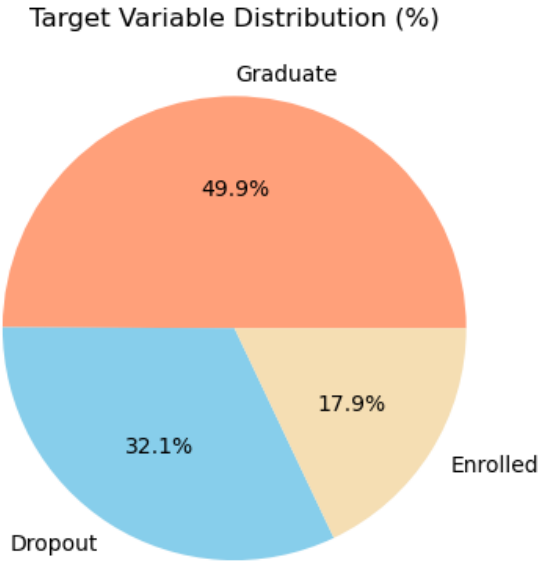

Figure 2.1: Distribution of student records among the three categories

3 Exploratory Data Analysis (EDA)

The exploratory data analysis was conducted to identify patterns and potential predictors of student outcomes (*"Dropout," "Enrolled," "Graduate"*) within the dataset. Specifically, the analysis focused on examining the distributions of the target variable across various features to better understand their predictive power for classification.

3.1 Data Visualization

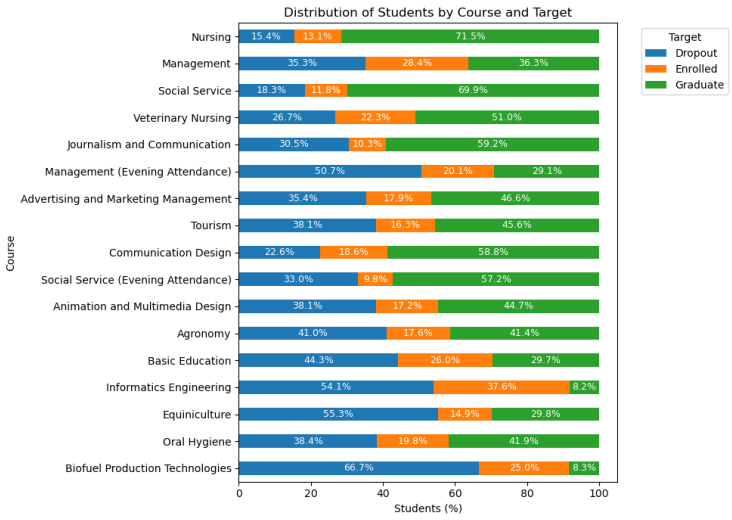

Figure 3.1: Student outcomes grouped by course

The provided chart illustrates the distribution of student outcomes ("Dropout," "Enrolled," "Graduate") across different academic programs. This visualization primarily serves to offer a **general overview** of the dataset, highlighting significant variations in student outcomes among different courses.

Notable variations can be observed: courses like Biofuel Production Technologies, Informatics Engineering, and Equiniculture have particularly **high dropout rates** (over 50%), suggesting these programs may have higher academic or practical demands. Conversely, programs such as Nursing and Social Service exhibit

high graduation rates (over 60%), possibly indicating stronger support systems or lower barriers to completion. Evening attendance programs like Management (Evening Attendance) show significant dropout percentages, likely due to challenges associated with balancing employment and study commitments.

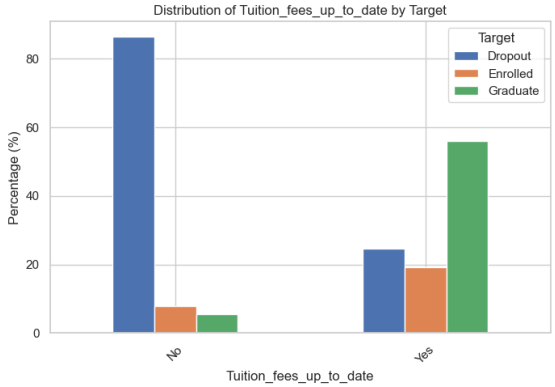

Figure 3.2: Distribution of *Tuition fees up to date* By Target

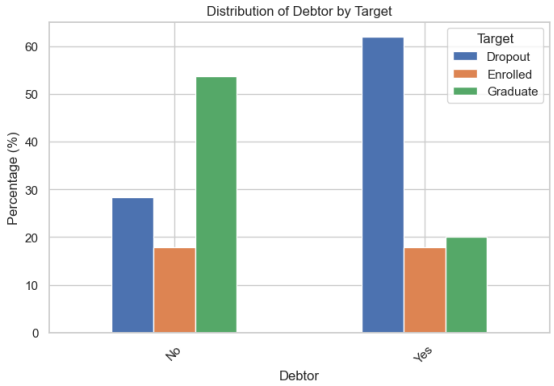

Figure 3.3: Distribution of *Debtor* By Target

The above graphs presented highlight the significant influence **financial factors** may have on predicting student outcomes for classification purposes. Students who did not keep tuition fees current predominantly experienced dropout outcomes, suggesting that **tuition fee status could be a strong predictor** in identifying at-risk students. Similarly, **debtor status provides a valuable indicator**, as students with outstanding debts demonstrated markedly higher dropout rates. These insights imply that incorporating financial indicators into predictive models could greatly enhance the accuracy of student outcome classification, particularly in identifying students at risk of dropping out.

The two sets of distribution plots for *Curricular units 1st sem approved* and *Curricular units 2nd sem approved* (Figure 3.4) reveal a highly informative relationship between **academic performance and student outcomes.** These visualizations show distinct patterns across the three classes, suggesting that early academic achievement is a critical feature for classification models.

Dropout students tend to cluster around very low approval counts (mostly 0–2 units), while **graduates** show consistent peaks at higher approval levels (around 6–8 units). **Enrolled students** generally fall in between, indicating moderate academic progress. This separation between distributions is particularly pronounced and stable across both semesters, making these features reliable and strong indicators for distinguishing among the outcome classes.

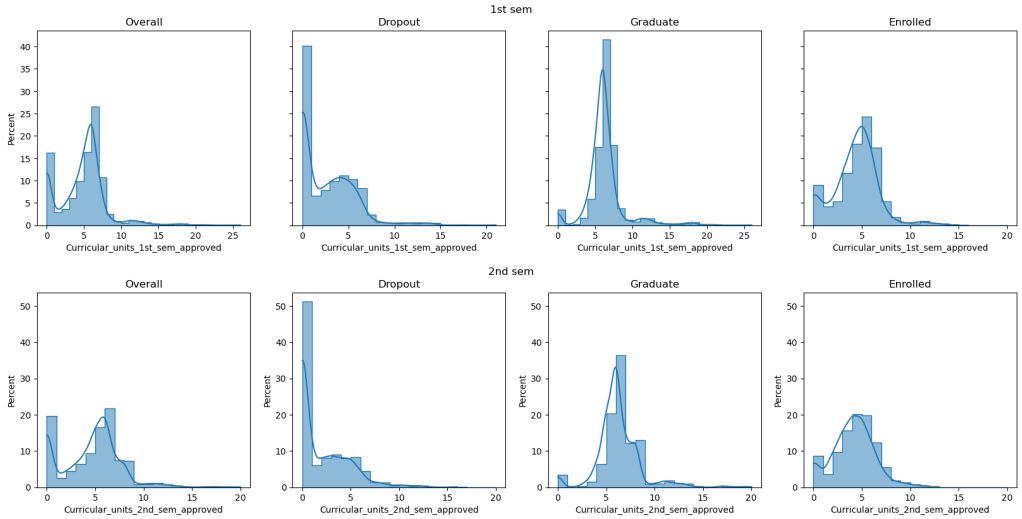

Figure 3.4: Distribution of approved 1 and 2 semester curricular units

Furthermore, an additional boxplot showing the grade distribution per semester by target further supports the **predictive power of academic performance**. It reveals that **graduates** tend to have consistently higher median grades in both semesters compared to enrolled or dropout students. **Dropout students** exhibit a wider spread and lower medians, including a high frequency of failing grades. These grade differences reinforce the idea that not only course completion but also **performance quality** contributes meaningfully to outcome prediction.

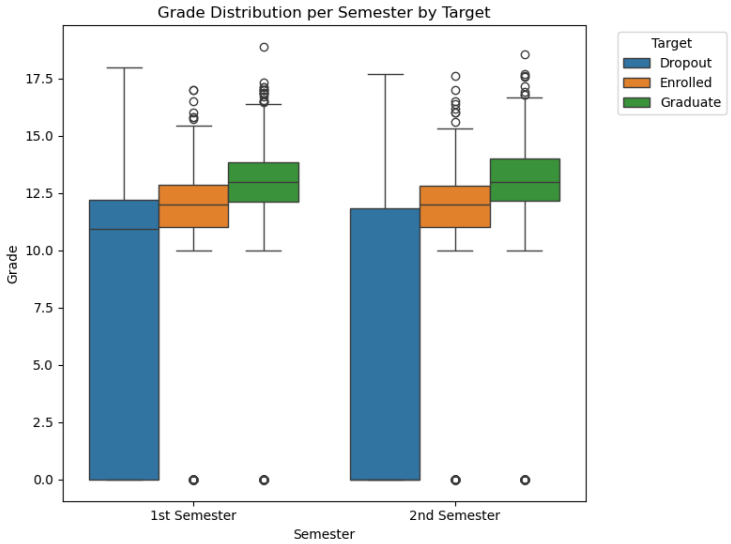

Figure 3.5: Grade distribution per semester by Target

Therefore, both the **number of approved curricular units** and **the grades earned** stand out as key predictive variables. Their ability to clearly separate the classes highlights their importance in improving model precision and in identifying students at risk based on early performance trajectories.

3.2 Correlation analysis

The correlation analysis was performed using both Cramér's V for categorical variables and Pearson correlation for numerical ones, complemented by ANOVA F-test scores to assess feature relevance with respect to the target variable.

The **Cramér's V heatmap** indicates that most categorical variables have weak correlations with each other. In particular, there is a very low correlation between Marital_status and the target variable, as well as between Nationality and the target variable.

Figure 3.6: Cramer's V Heatmap

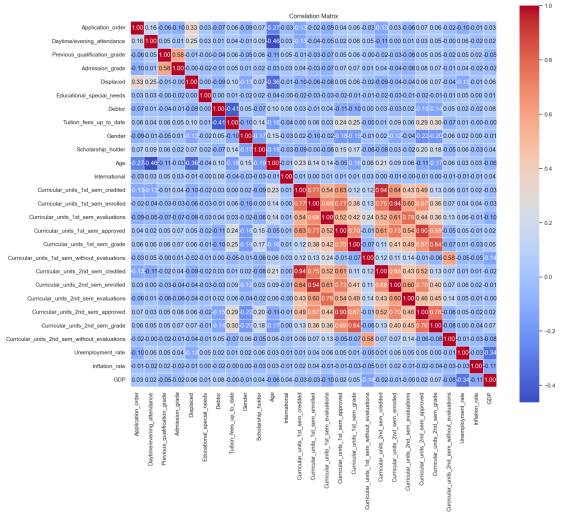

Figure 3.7: Correlation matrix

Similarly, the **Pearson correlation matrix** reveals that numerical features related to academic performance—such as the number of approved curricular units and grades—show the highest correlations with each other.

The **ANOVA F-test** further validates feature importance. The top-ranked features by F-value include **academic performance indicators** (e.g., Curricular_units_2nd sem_approved, Curricular_units_1st_sem_approved, and grades), followed by **financial and demographic attributes** such as Tuition_fees_up_to_date, Scholarship_holder, and Age.

```
====================== ANOVA Results (All Features) ====================

                                    Feature      F-value        p-value
21            Curricular_units_2nd_sem_approved  1410.732938   0.000000e+00
22               Curricular_units_2nd_sem_grade  1134.109544   0.000000e+00
15            Curricular_units_1st_sem_approved   859.866768  3.649472e-316
16               Curricular_units_1st_sem_grade   713.517328  2.803052e-269
7                       Tuition_fees_up_to_date   505.621429  1.784950e-198
9                            Scholarship_holder   225.751437   4.436825e-94
10                                          Age   154.712071   1.138849e-65
6                                        Debtor   137.647527   1.018223e-58
8                                        Gender   123.041811   9.950346e-53
20         Curricular_units_2nd_sem_evaluations    87.801092   4.039137e-38
19            Curricular_units_2nd_sem_enrolled    75.591910   5.244430e-33
13            Curricular_units_1st_sem_enrolled    59.467391   3.272852e-26
14         Curricular_units_1st_sem_evaluations    37.527840   6.897115e-17
3                            Admission_grade    35.648604   4.380466e-16
4                                  Displaced    29.239226   2.425582e-13
2                 Previous_qualification_grade    27.728589   1.077783e-12
23 Curricular_units_2nd_sem_without_evaluations    20.185531   1.876375e-09
0                          Application_order    19.727174   2.955293e-09
1                  Daytime/evening_attendance    14.454123   5.534625e-07
17 Curricular_units_1st_sem_without_evaluations    11.437319   1.110815e-05
18            Curricular_units_2nd_sem_credited     9.974542   4.762728e-05
12            Curricular_units_1st_sem_credited     7.979355   3.474158e-04
24                         Unemployment_rate     5.922513   2.699757e-03
26                                       GDP     4.799009   8.280870e-03
25                          Inflation_rate     1.741990   1.752917e-01
11                           International     0.639709   5.274945e-01
5                    Educational_special_needs     0.320854   7.255460e-01
```

Figure 3.8: ANOVA results

Based on this comprehensive evaluation, the following features were removed due to their limited relevance:

- **Nationality**: Shows very low correlation with the target (Cramér's V = 0.07)
- **Marital_status**: Correlates weakly with the target.
- **International**: Exhibits one of the lowest F-values and negligible correlation.
- **Educational_special_needs**: Has the lowest p-value.
- **Inflation_rate**: Statistically irrelevant, with an insignificant F-value.

Removing these variables allows for dimensionality reduction, decreases noise, and improves model interpretability without compromising predictive accuracy.

4 Student Dropout Classification

4.1 Data Splitting

To evaluate the predictive performance of the model and prevent data leakage, the dataset was split into two subsets: **80% for training** and **20% for testing**. This **stratified split** ensures that the distribution of the target variable is preserved across both subsets, as illustrated in the chart below. Maintaining a consistent class balance in both sets is crucial for reliable evaluation and helps the model generalize better to unseen data. This approach supports the development of robust and unbiased predictive models for student dropout classification.

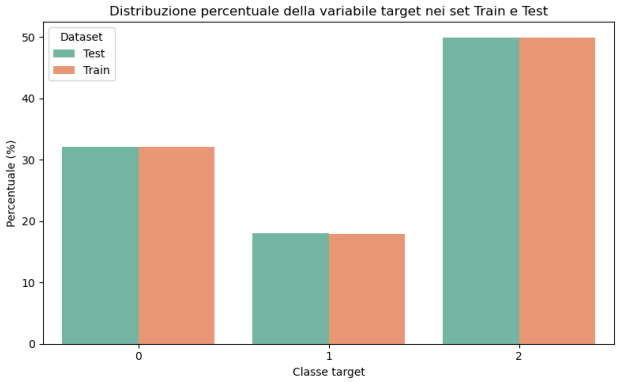

Figure 4.1: Percentage distribution of target classes in the training and test sets.

4.2 Preprocessing phase

To ensure the dataset was suitable for machine learning algorithms, several preprocessing steps were applied. This involved transforming categorical features into a numerical format, scaling numerical variables, and engineering new features to better capture the relationships in the data.

4.2.1 Feature engineering

To enhance the model's ability to capture meaningful patterns, a comprehensive feature engineering process was conducted. Several **new variables were created** to

summarize academic progression, performance dynamics, and socioeconomic background in a more compact and informative way.

- Two new variables were created to represent the ***pass rate*** **for each semester**. These were computed as the ratio between the number of curricular units a student passed and the number of units in which they were enrolled in that semester. To ensure robustness and prevent division-by-zero errors, the pass rate was set to zero for students not enrolled in any unit during the corresponding semester. These features capture **students' academic efficiency** and may serve as early indicators of potential dropout or successful progression.

- A ***weighted average grade*** across the first and second semesters was computed, where each semester's grade was weighted by the number of approved curricular units. This provides a more representative indicator of overall academic performance.

- The difference in approved curricular units between semesters (***approved_ delta***) was introduced to reflect performance over time—potentially revealing early signs of academic decline or improvement.

- The ***total number of enrolled curricular*** units across semesters was also included as a proxy for workload and persistence.

- To incorporate demographic insights, the ***age_ bin*** variable was introduced to segment students into categorical age groups, simplifying the model's ability to learn age-related patterns. The continuous "Age" variable was discretized into three bins: **young** ($\leq$ 20 years) , **medium** (21–25 years), and **adult** ($>$ 25 years) reflecting typical academic stages and life contexts. Importantly, the original "Age" feature was retained in the dataset to preserve its full numerical precision and allow the model to benefit from both representations.

- Finally, a ***parental background score*** was constructed to capture the influence of family education levels. This score was obtained by summing the numerical codes associated with the mother's and father's educational qualifications, provided they were valid numeric values. The resulting composite indicator serves as **a proxy for the student's socioeconomic and educational environment**, which is known to influence academic performance and persistence.

To reduce redundancy and multicollinearity, the original columns used to compute these new features—such as parental qualifications—were removed from the dataset. This process ensured that only the most informative and synthesized representations of the data were retained for modeling.

4.2.2 Numerical and Categorical Feature Transformation

As part of the preprocessing pipeline, tailored transformations were applied to different feature types to ensure compatibility with machine learning algorithms and improve model performance. A **ColumnTransformer** was used to handle numerical, binary, and categorical variables in a structured and modular way.

- Numerical features were standardized using *StandardScaler*, which transforms each variable to have zero mean and unit variance.
- Binary variables were passed through without modification, as they were already in a suitable format for modeling.
- Categorical features were encoded using *One-Hot encoding*, with the first category dropped to avoid multicollinearity. The encoder was configured to ignore unknown categories during inference, thus ensuring robustness when encountering unseen values in the test set or in production.

This preprocessing strategy ensures that all input variables are converted into a consistent numeric format, enabling effective learning while preserving the interpretability and integrity of the original data.

4.3 Pipeline building

To streamline the training process and ensure reproducibility, a complete machine learning pipeline was constructed by chaining together all essential preprocessing and modeling steps. The pipeline consists of four main components, executed sequentially.

1. **Feature Engineering:** the first step is a custom feature engineering transformer, which applies the previously defined transformations and removes the original features from which they were derived, as specified by the drop_originals=True parameter. This encapsulation ensures that feature engineering is consistently applied during both training and inference.
2. **Preprocessing:** the second step applies the previously defined preprocessing pipeline, where numerical features are standardized, binary features are passed through, and categorical features are encoded using one-hot encoding.
3. **Handling imbalanced data:** to address class imbalance in the target variable, the third step incorporates **SMOTE (Synthetic Minority Oversampling Technique)**, which generates synthetic examples of minority classes within the training data, thereby helping the model learn more effectively from underrepresented outcomes.

4. **Model Integration:** finally, the fourth step integrates the machine learning model to be trained on the transformed data.

This modular pipeline ensures that all preprocessing steps are applied identically during cross-validation and deployment, reducing the risk of data leakage and improving overall model robustness.

4.4 Model Comparison

To identify the most effective algorithm for predicting student outcomes, a systematic model comparison strategy was implemented. Multiple supervised learning models were evaluated, including:

- Decision Tree,
- Random Forest,
- Support Vector Machine (SVM),
- Gradient Boosting,
- XGBoost,
- LightGBM,
- CatBoost

These algorithms were selected based on their use in the literature for educational data mining [3] and their proven effectiveness in similar predictive tasks. Moreover, they are better suited for **handling multicollinearity**, either by being inherently robust to correlated features or through regularization mechanisms. Each model was paired with a refined hyperparameter grid tailored to its architecture and training dynamics.

To optimize hyperparameters efficiently, the pipeline was wrapped within a **HalvingGridSearchCV** procedure. This method is a resource-efficient alternative to traditional grid search. It begins by evaluating all parameter combinations on a small subset of data. At each iteration, only the most promising candidates are retained and evaluated on increasingly larger portions of the data. This successive halving approach allows the search to converge faster on optimal configurations, while significantly reducing computational cost. The search process was guided by **5-fold stratified cross-validation**, with macro-averaged F1-score ***(f1_macro)*** used as the primary scoring metric to account for class imbalance across the three target classes. The **F1-score** is a suitable evaluation metric for imbalanced datasets, as it balances precision and recall. In this study, the F1-score was chosen to assess

the model's accuracy, considering both false positives and false negatives.

The **hyperparameter grids** tested for each model are summarized in the following table.

Table 4.1: Hyperparameter grids tested for each model

Model	**Hyperparameters Tested**
Decision Tree	`max_depth` = [None, **5**, 10, 15]; `min_samples_split` = [2, **5**, 10]; `min_samples_leaf` = [**1**, 2, 4]
Random Forest	`n_estimators` = [100, 200, **300**]; `max_depth` = [None, **10**, 20]; `max_features` = [**'sqrt'**, 'log2']
SVM	`C` = [0.1, 1, **10**]; `kernel` = ['linear', **'rbf'**]; `gamma` = ['scale', **'auto'**]
Gradient Boosting	`n_estimators` = [100, **200**]; `learning_rate` = [**0.01**, 0.1]; `max_depth` = [**3**, 5]
XGBoost	`n_estimators` = [**100**, 200]; `learning_rate` = [0.01, **0.1**]; `max_depth` = [**3**, 5]; `subsample` = [0.8, **1**]
LightGBM	`n_estimators` = [100, **200**]; `learning_rate` = [**0.01**, 0.1]; `num_leaves` = [31, **50**]; `max_depth` = [-1, **5**]
CatBoost	`iterations` = [**100**, 200]; `learning_rate` = [0.01, **0.1**]; `depth` = [6, **10**]

The best model from each search was then evaluated on the **held-out test set** using multiple performance metrics, including balanced accuracy, macro F1-score, and macro-averaged ROC AUC. Additionally, **class-specific F1-scores** were computed to assess per-class performance. The best-performing pipeline for each algorithm was saved in the folder *Models* for further analysis and deployment.

4.4.1 Results

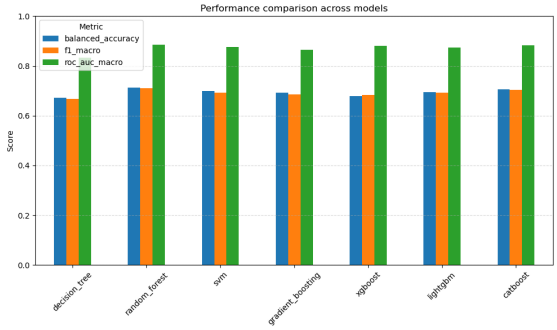

Figure 4.2: Performance comparison

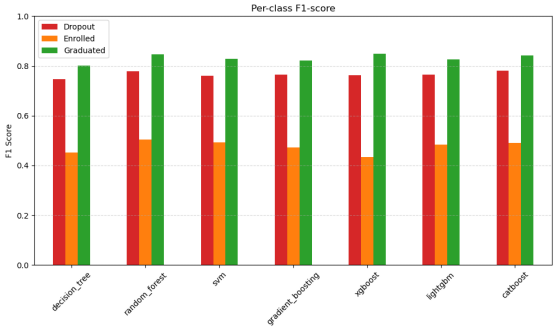

Figure 4.3: Per-class F1-score

The *Figure 4.2* illustrates the overall performance of each model across the three selected metrics: **balanced accuracy, macro F1-score, and macro-averaged ROC AUC**. All models performed reasonably well, with Random Forest, XGBoost, LightGBM, and CatBoost consistently achieving higher scores across all metrics. In particular, Random Forest achieved the best trade-off between precision and recall, reflected in its superior F1-score **(0.7104)**.

To further analyze model behavior, the *Figure 4.3* presents the **F1-scores for each class**: Dropout, Enrolled, and Graduated. While most models performed well for the "Graduated" class, the "Enrolled" class was generally more challenging to predict accurately, with lower F1-scores across the board. **Random Forest and CatBoost** demonstrated more balanced performance across all classes, suggesting better generalization.

These results highlight the value of using multiple evaluation metrics and per-class analysis to identify the most reliable and interpretable model in an imbalanced multi-class classification setting.

The complete set of numerical evaluation results for all models and metrics can be found in the supplementary file *model_ comparison_ results_ SMOTE.csv*.

4.4.2 Statistical comparison

To assess which model performs best, a statistical comparison was conducted between the **Random Forest and CatBoost classifiers**. In order to obtain a meaningful comparison, multiple runs of 10-fold cross-validation were performed.

Before conducting the test, the distribution of F1-score differences was assessed using Q-Q plots, histograms and Shapiro-Wilk Test to verify the normality assumption. As the differences did not appear to follow a normal distribution, a **Wilcoxon signed-rank test** was used as a non-parametric alternative.

The resulting p-value was **0.6265**, indicating that the performance difference between the two models is not statistically significant.

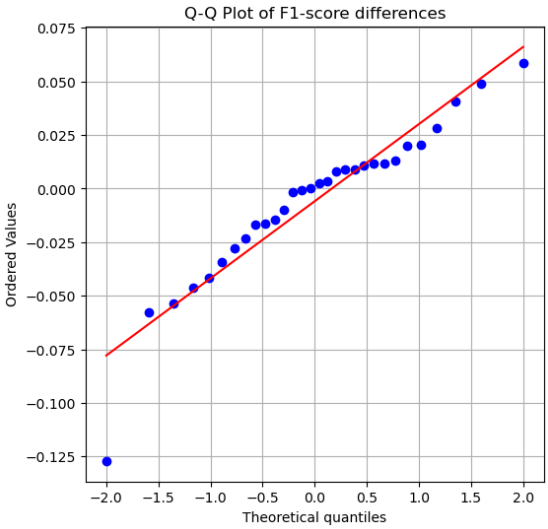

Figure 4.4: Q-Q plot

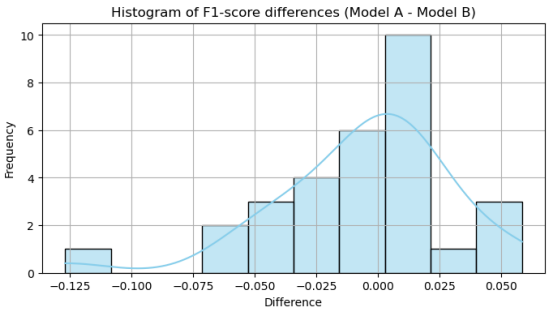

Figure 4.5: Histogram

4.4.3 Conclusion

Among the models evaluated, **Random Forest** and **CatBoost** emerged as the top-performing classifiers, achieving the highest scores across multiple performance metrics. However, a statistical comparison revealed that the difference between them is not significant. Therefore, both models can be considered equally effective for predicting student outcomes in this context.

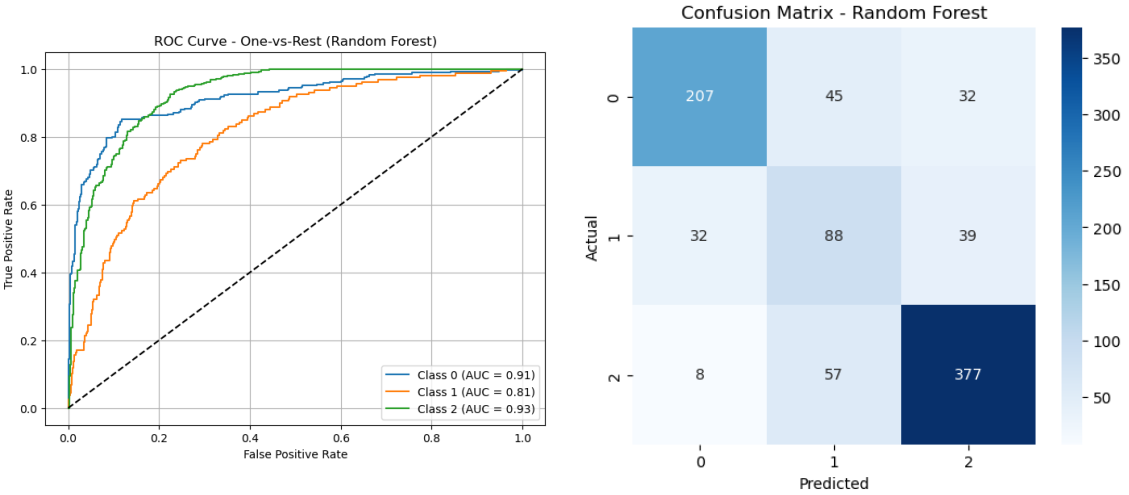

Figure 4.6: ROC curve and Confusion Matrix for the fine-tuned **Random Forest model**

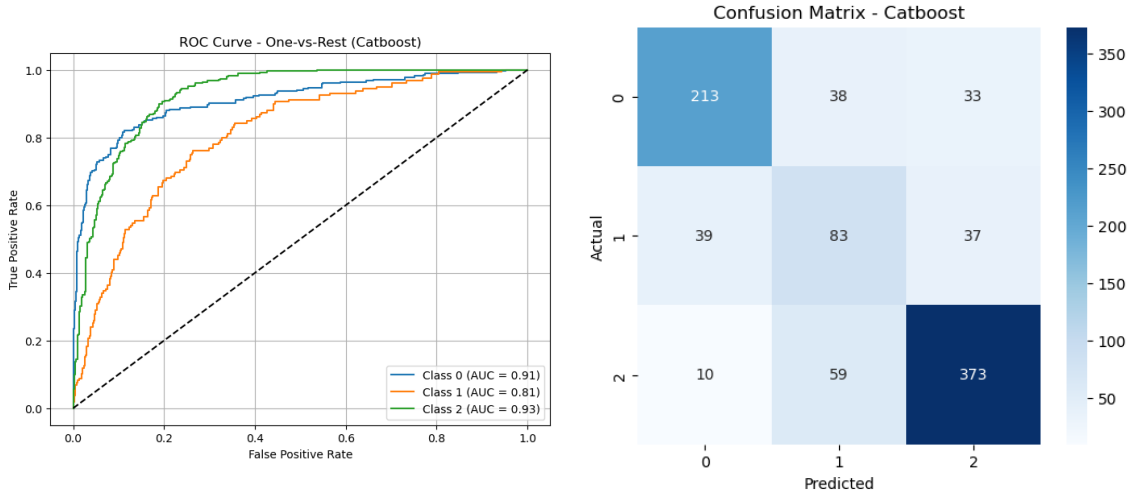

Figure 4.7: ROC curve and Confusion Matrix for the fine-tuned **CatBoost model**

Figures below provide additional insights into the performance of the fine-tuned Random Forest and CatBoost models through **ROC curves** and **confusion matrices**. Both models achieved high AUC scores across all classes, indicating strong discriminatory power. The confusion matrices also reveal similar patterns of misclassification, particularly for the **"Enrolled" class**. These visual results reinforce the conclusion that both classifiers perform comparably well.

```
=== Classification Report: Random Forest ===

              precision    recall  f1-score   support

           0     0.8381    0.7289    0.7797       284
           1     0.4632    0.5535    0.5043       159
           2     0.8415    0.8529    0.8472       442

    accuracy                         0.7593       885
   macro avg     0.7142    0.7118    0.7104       885
weighted avg     0.7724    0.7593    0.7639       885

Balanced Accuracy: 0.7118
```

```
=== Classification Report: Catboost ===

              precision    recall  f1-score   support

           0     0.8130    0.7500    0.7802       284
           1     0.4611    0.5220    0.4897       159
           2     0.8420    0.8439    0.8429       442

    accuracy                         0.7559       885
   macro avg     0.7054    0.7053    0.7043       885
weighted avg     0.7642    0.7559    0.7593       885

Balanced Accuracy: 0.7053
```

Figure 4.8: Model Performance Comparison

5 Model explainability

5.1 SHAP analysis

To gain a deeper understanding of the model's behavior and the underlying factors influencing its predictions, **SHAP (SHapley Additive exPlanations)** values were employed. SHAP offers a game-theoretic approach to explain the output of machine learning models and is particularly effective in interpreting complex non-linear models like CatBoost.

Global Explanations are provided through the aggregated bar chart and beeswarm plots for each class. These representations summarize the average absolute SHAP values per feature across the entire dataset, indicating which features have the greatest impact on the model's output overall.

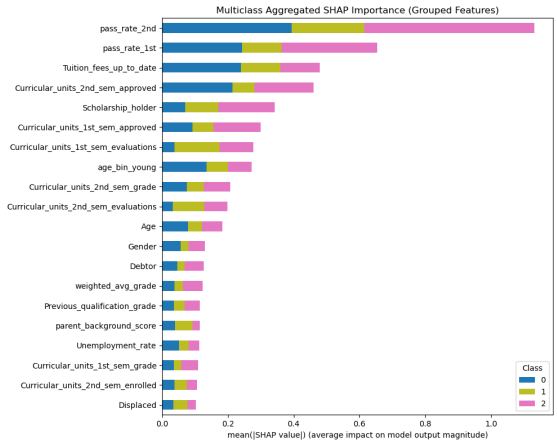

Figure 5.1: CatBoost Model

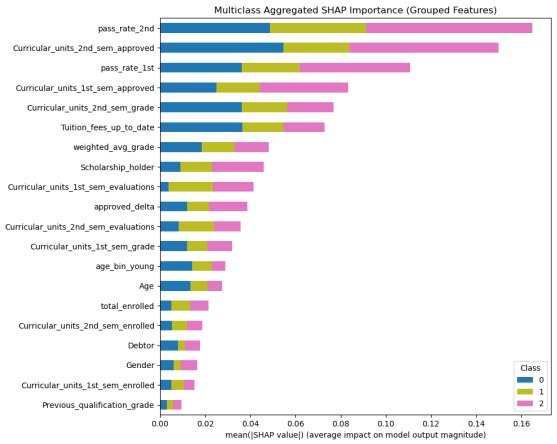

Figure 5.2: Random Forest Model

The first plot (CatBoost) and the second plot (Random Forest) both visualize the mean absolute SHAP values across classes, providing a global importance ranking of features for a multiclass classification task.

The comparison between CatBoost and Random Forest reveals consistent identification of key features—such as ***pass_rate_2nd***, ***Curricular_units_2nd_sem_approved***, and ***pass_rate_1st*** —as the most influential across classes. This consistency indicates the robustness and reliability of these variables in influencing the model's output across different algorithms. Finally, the color distribution highlights that

certain features specifically influence only particular classes (e.g., pass_rate_2nd for "Graduate"), while others have a more transversal impact across all classes. Moreover, it appears that the models struggle more to identify strong patterns for the **"Enrolled" class.**

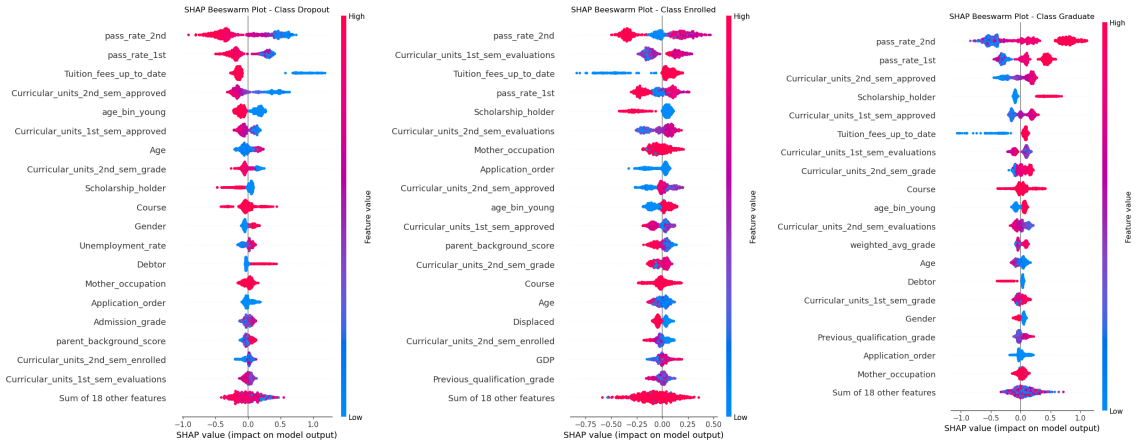

Figure 5.3: SHAP plots for each class using the CatBoost model

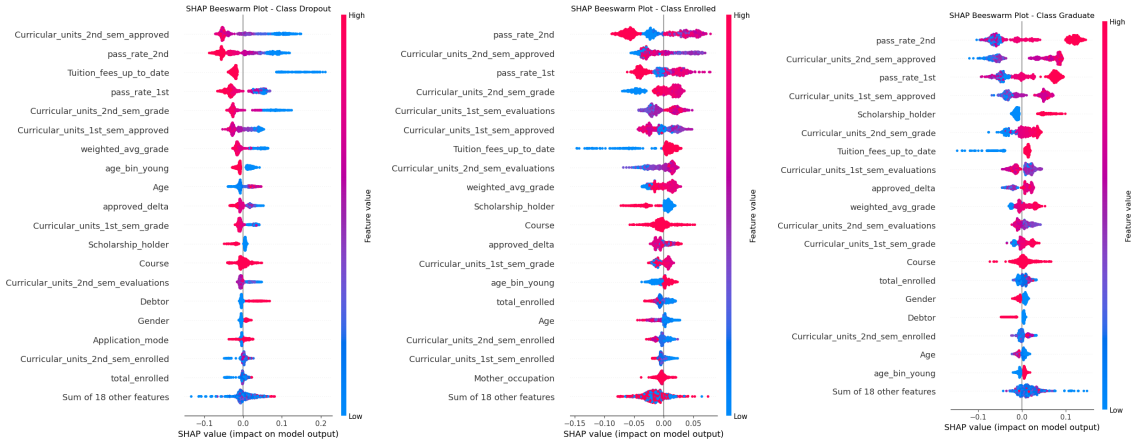

Figure 5.4: SHAP plots for each class using the Random Forest model

Figure 5.3 presents the SHAP beeswarm plots for each class using the **CatBoost model**, while Figure 5.4 shows the corresponding plots for the **Random Forest model**. Each plot displays the **top 20 features** ranked by importance, along with the distribution of SHAP values across the dataset.

The results obtained from both models demonstrate **strong consistency:** variables such as *pass_rate_2nd, pass_rate_1st, Curricular_units_2nd_sem_approved, Curricular_units_1st_sem_approved, Tuition_fees_up_to_date,* repeatedly appear among the most influential across all three classes.These variables reflect **students' academic progress and engagement** during the semesters, making them

intuitive predictors of dropout and graduation. This alignment across algorithms confirms the robustness of these predictors and their relevance in determining student outcomes.

Notably, *Tuition_fees_up_to_date* and *Scholarship_holder* also emerge as important features, suggesting that **financial support plays a key role** in shaping student outcomes. In contrast, demographic features like *Gender* and *Age* tend to have a limited impact across classes.

The colors in the beeswarm plots reflect actual feature values. In both models, we see that low values of academic performance features (e.g., low pass_rate_2nd or Tuition_fees_up_to_date) tend to increase the likelihood of dropout, while high values of the same features push predictions toward graduation.

Overall, the SHAP analysis confirms the dominance of academic performance indicators in predicting student outcomes and provides transparency on how specific features influence the classification for each class.

Local Explanations are represented using SHAP waterfall plots. These plots break down the contribution of each individual feature for a single prediction instance. They provide transparency into how the model arrived at a specific decision for a given student.

In the **CatBoost plots**, many features contribute to the prediction, each with a small effect. For example, a low pass rate in the second semester and few approved subjects push the prediction toward "Dropout".

In the **Random Forest plots**, the model focuses more on just a few strong features. For "Dropout", the most important ones are again low pass rate and few approved subjects, which have a big influence.

In summary, both models rely heavily on **academic performance,** especially in the second semester.

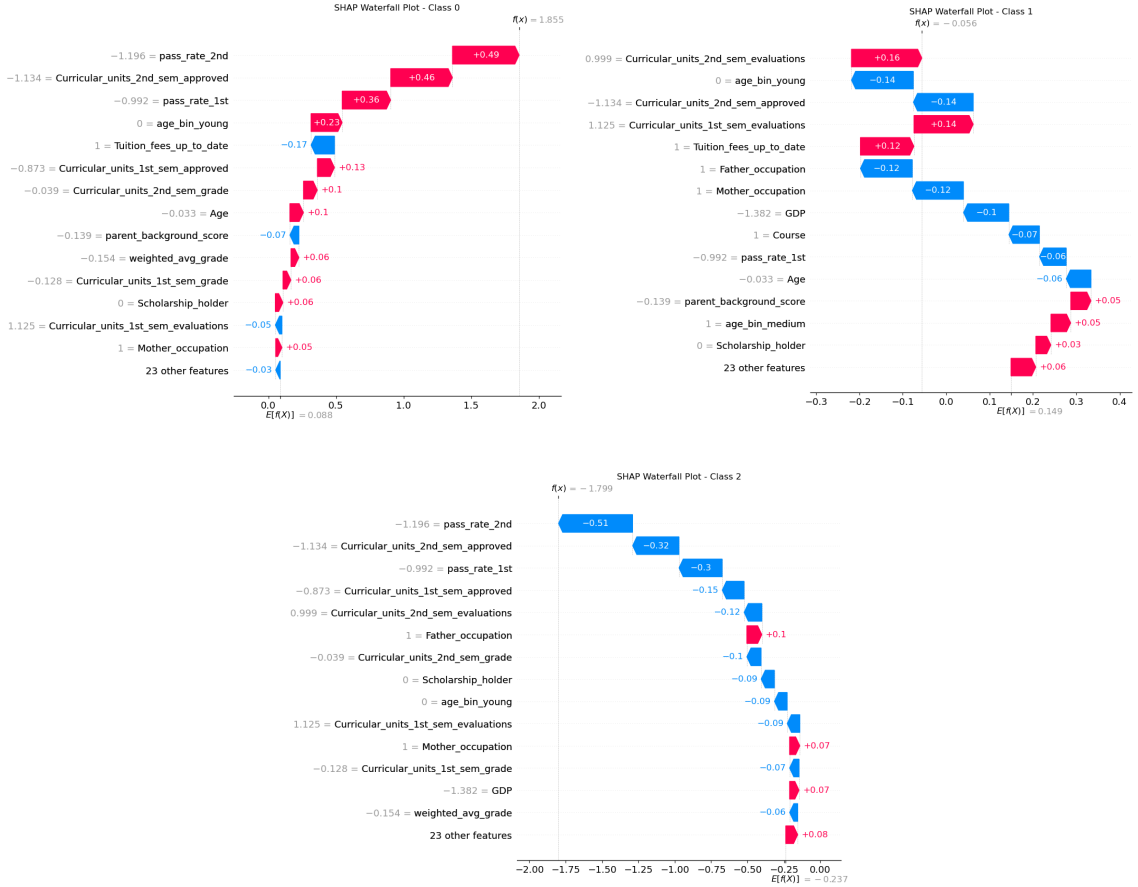

Figure 5.5: SHAP waterfall plots showing the logit-level feature contributions of the **CatBoost model** for a single student across the three classes

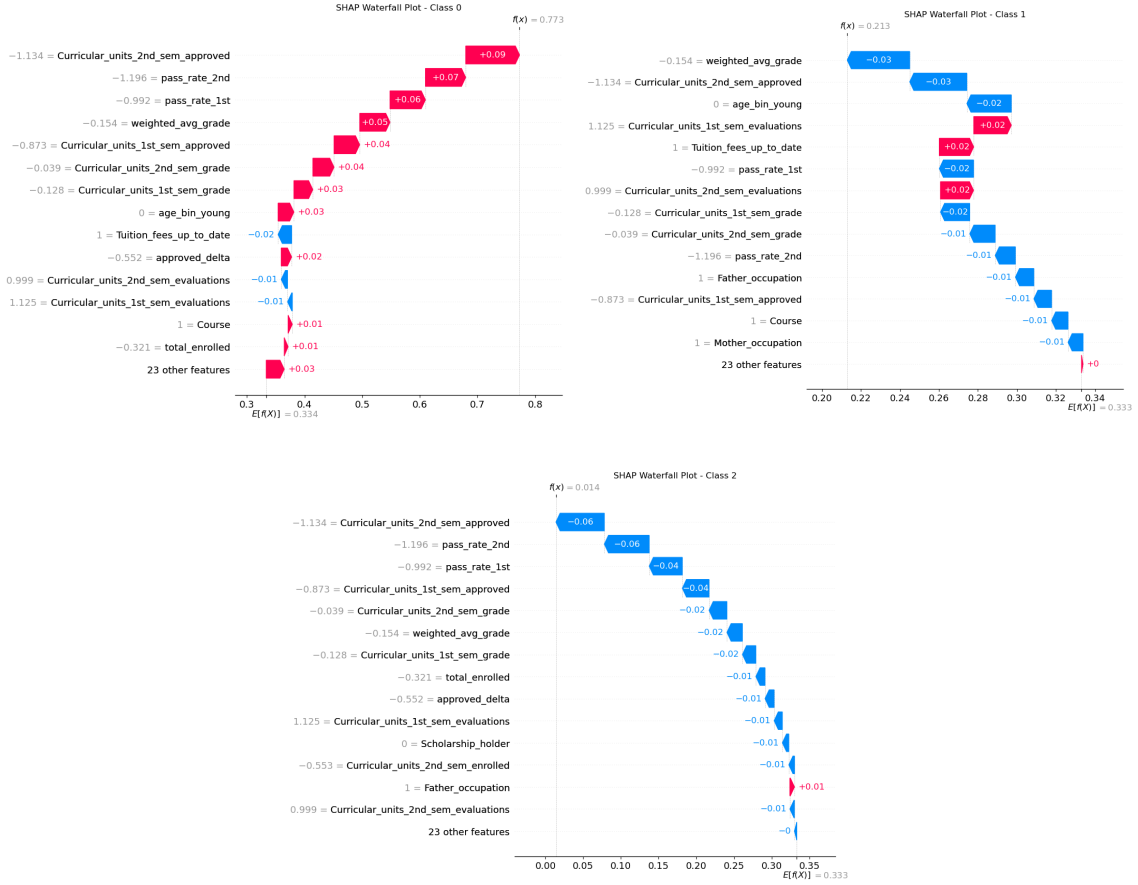

Figure 5.6: SHAP waterfall plots showing the probability-level feature contributions of the **Random Forest model** for a single student across the three classes.

In summary, SHAP global explanations help identify the most influential features across the dataset, guiding policy-level insights and interventions. Local explanations, on the other hand, support individual-level diagnosis, enabling personalized support for students at risk.

The SHAP feature importance results from the paper [3] **confirm the relevance of core academic and administrative variables**, such as *Curricular_units_2nd_sem_approved, Curricular_units_1st_sem_approved,* and *Scolarship_holder*, which also emerge as top predictors in our SHAP analysis. However, some key differences arise: in our model, aggregated features such as *pass_rate_2nd, pass_rate_1st, and weighted_avg_grade* play a dominant role. These variables were not included in the paper, as their analysis was based only on raw features. This highlights how feature engineering expanded the explanatory power of the model, allowing SHAP to capture interactions and derived patterns.

6 Graphical User Interface

To support model interpretability and real-world usability, a graphical user interface (GUI) was developed using Python's Tkinter library. The interface allows users to manually input a student's demographic, academic, and socioeconomic characteristics.

Once the data is entered, the interface processes the input through the corresponding fine-tuned prediction pipeline and **outputs a predicted class label** (Dropout, Enrolled, or Graduate).

In addition to the predicted outcome, the interface opens a separate window that displays detailed **prediction insights**. This includes the class probabilities across all three classes, allowing users to evaluate the model's confidence, as well as a textual explanation based on SHAP values that highlights the **top five most influential feature**s contributing to the prediction.

This enriched feedback mechanism enhances transparency and fosters trust in model outputs, while also allowing users to explore "what-if" scenarios by modifying input values.

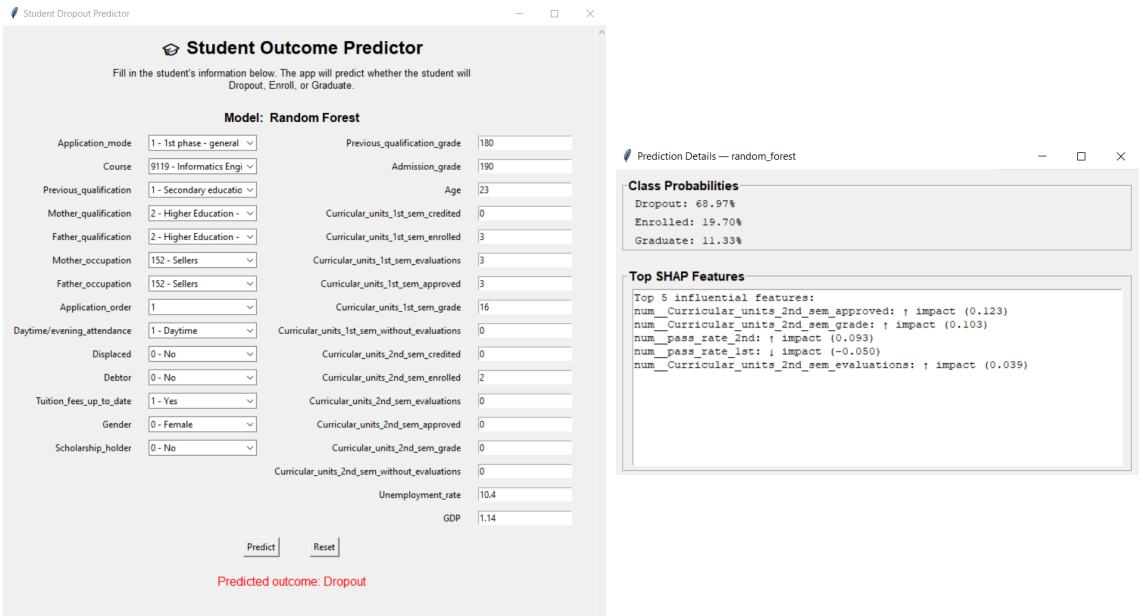

Figure 6.1: Graphical User Interface

Bibliography

[1] UCI Machine Learning Repository, "Predict students dropout and academic success data set," 2023, dataset used for student dropout and academic success prediction studies. [Online]. Available: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

[2] V. Realinho, J. Machado, L. Baptista, and M. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, p. 146, 2022. [Online]. Available: https://doi.org/10.3390/data7110146

[3] A. Villar and C. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study," *Discover Artificial Intelligence*, vol. 4, no. 2, 2024. [Online]. Available: https://doi.org/10.1007/s44163-023-00079-z