01

Predicting Student Dropout in Higher Education Using Supervised Learning

Data Mining and Machine Learning

Martina Fabiani

Introduction

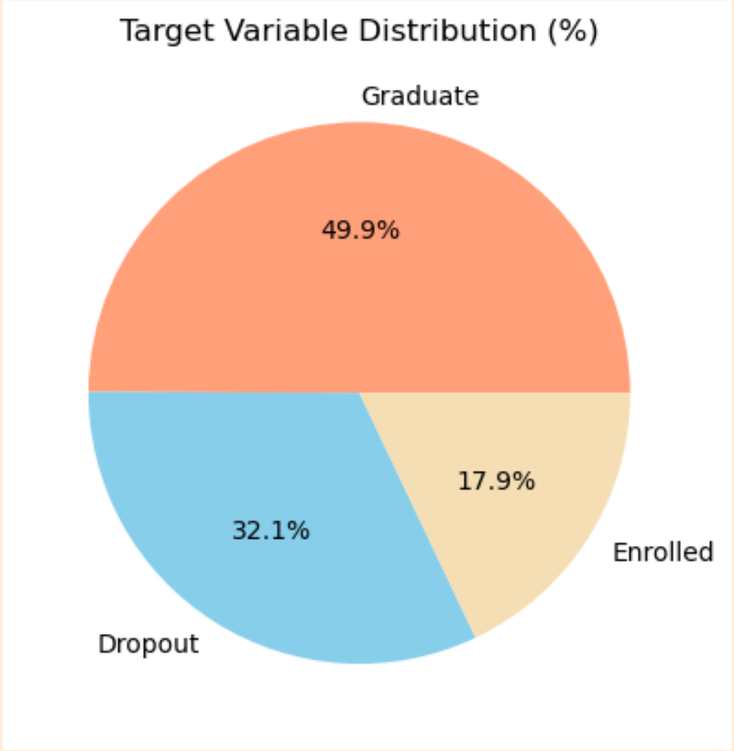
- **Student dropout** is a major concern in higher education, with broad social and economic impacts.
- This project applies **supervised machine learning** to predict student outcomes.
- The goal is to **classify students** as Graduate, Enrolled, or Dropout.
- **Early prediction** enables timely, personalized interventions to support at-risk students.
- Insights from the model can inform **data-driven** educational strategies.

Dataset

DATASET OVERVIEW

- **Labeled dataset** containing data from a **Portuguese higher education institution**.
- It covers a **range of factors**, including demographic, socio-economic, and academic factors.
- It covers student records from the academic years 2008/2009 to 2018/2019.
- Data from **17 undergraduate degree** programs across diverse fields
- **4424 records** with **37 attributes**, where each record represents an individual student, and contains **no missing values**

Unbalanced dataset



Dataset

CORRELATION ANALYSIS

Categorial features

```
======================== 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
```

Numerical features

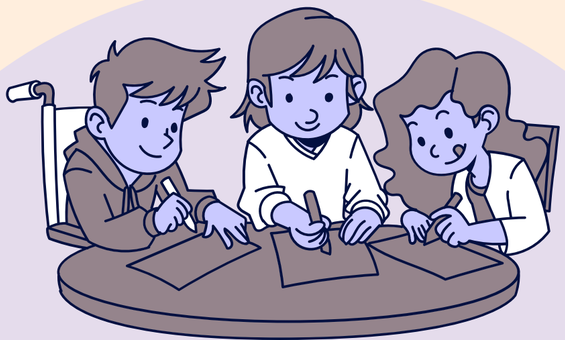
Preprocessing & Pipeline Building

Feature Deletion

- Nationality
- International
- Educational_special_needs
- Marital_status
- Inflation_rate

Feature Engineering

- Pass rate (1st & 2nd semester)
- Weighted average grade *(calculated by weighting each semester's grade by the number of approved curricular units)*
- Delta approved units.
- Total enrolled unit
- Age binning
- Parental background score



FEATURE ENGINEERING

↓

STANDARD SCALER
ONE-HOT ENCODER

↓

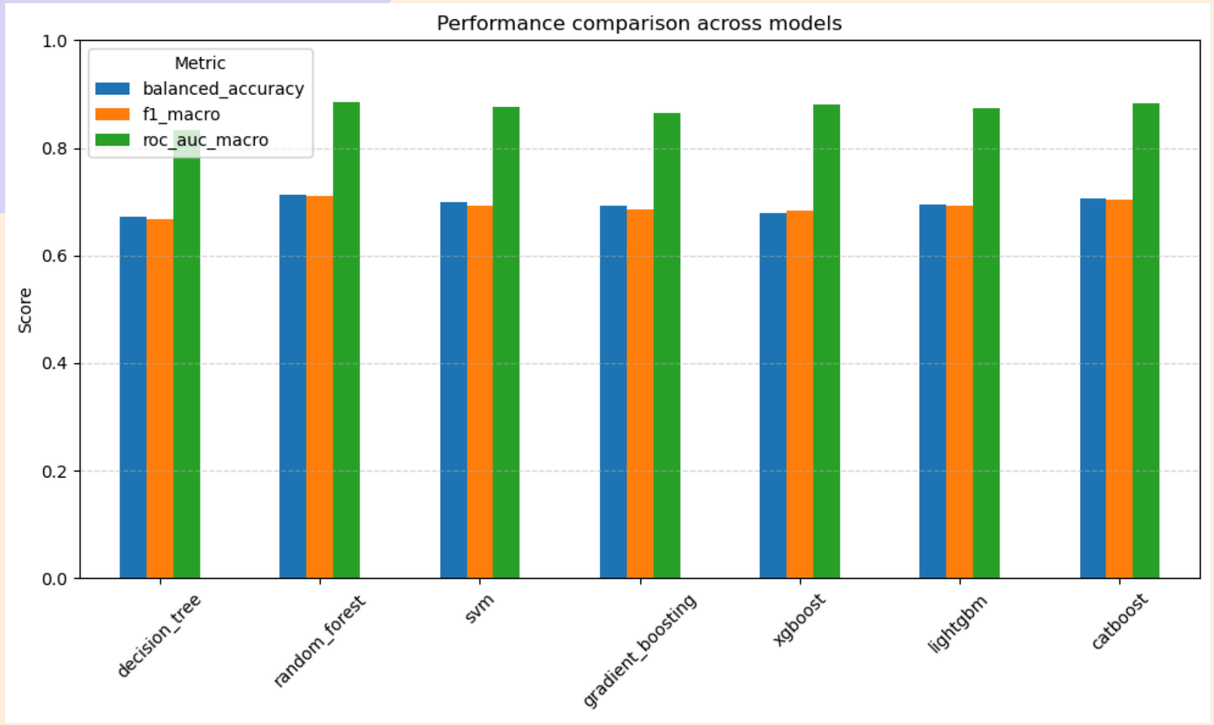
SMOTE

↓

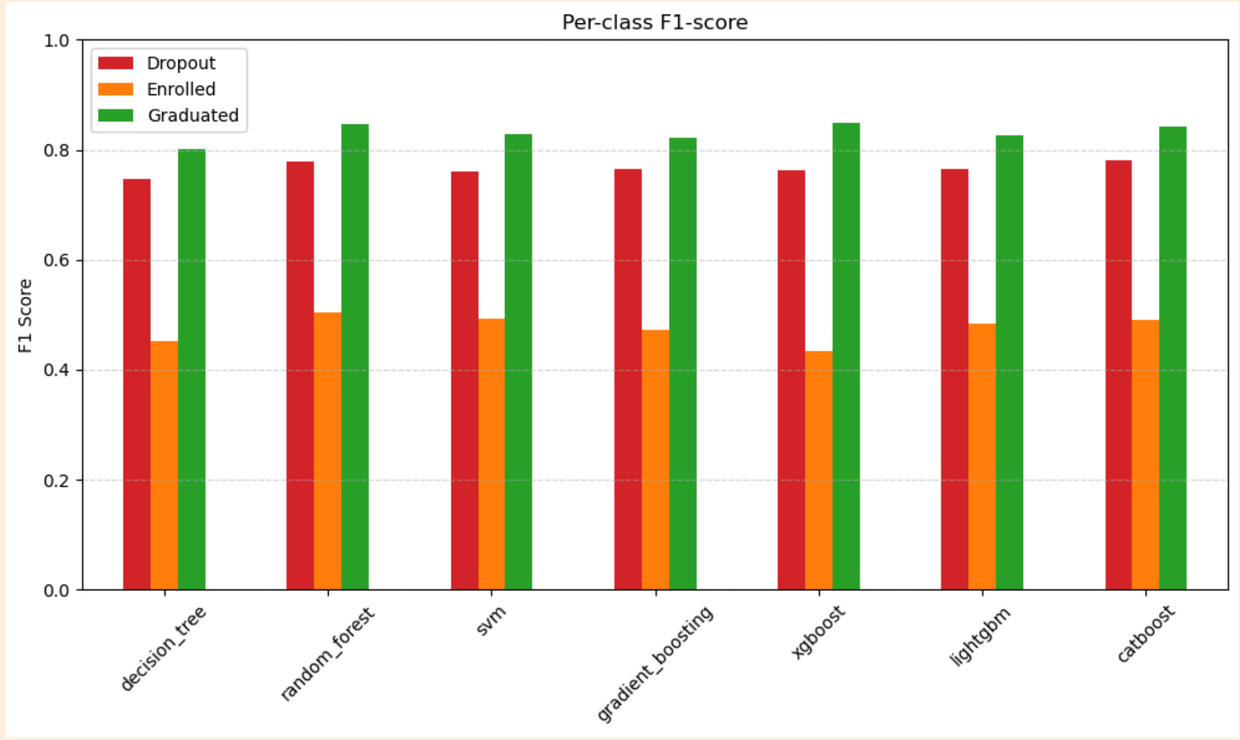
MODEL

Model comparison

RESULTS



- **Random Forest** showed the best precision-recall trade-off, achieving the highest F1-score **(0.71)**

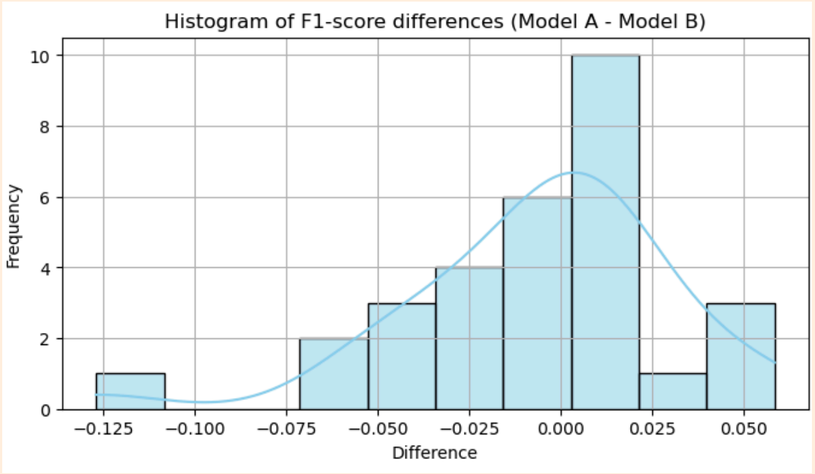


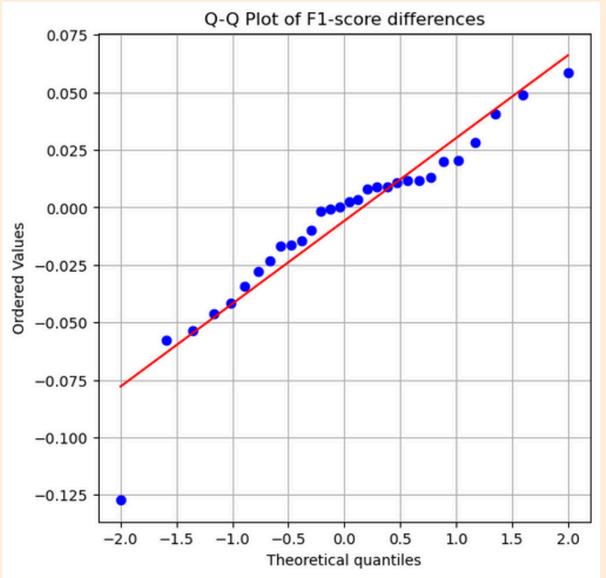
- **Random Forest** and **CatBoost** had the most balanced performance across classes.
- The "**Enrolled**" class was the most difficult to predict for all models.

Statistical comparison

Comparison between **Random Forest** and **CatBoost**

- To have more data, I performed three 10-fold cross-validations
- Assumption of normality not supported → **Wilcoxon signed-rank test** applied

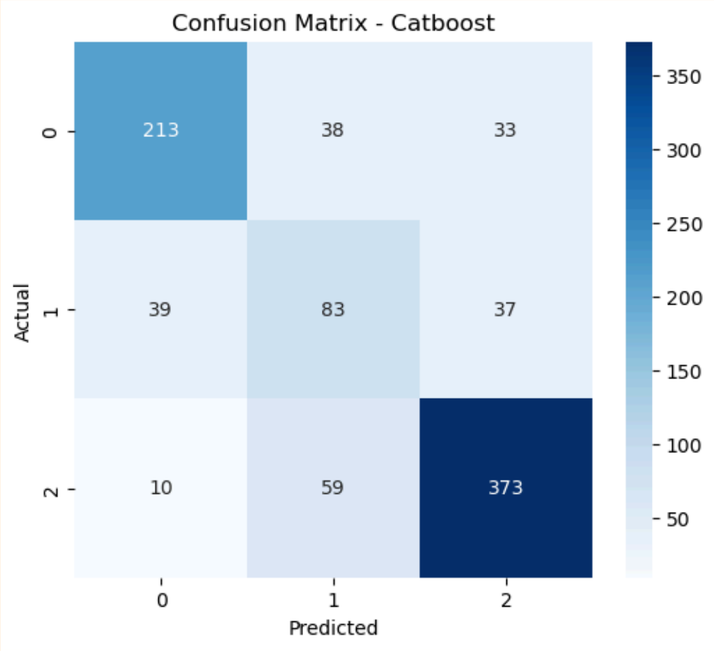


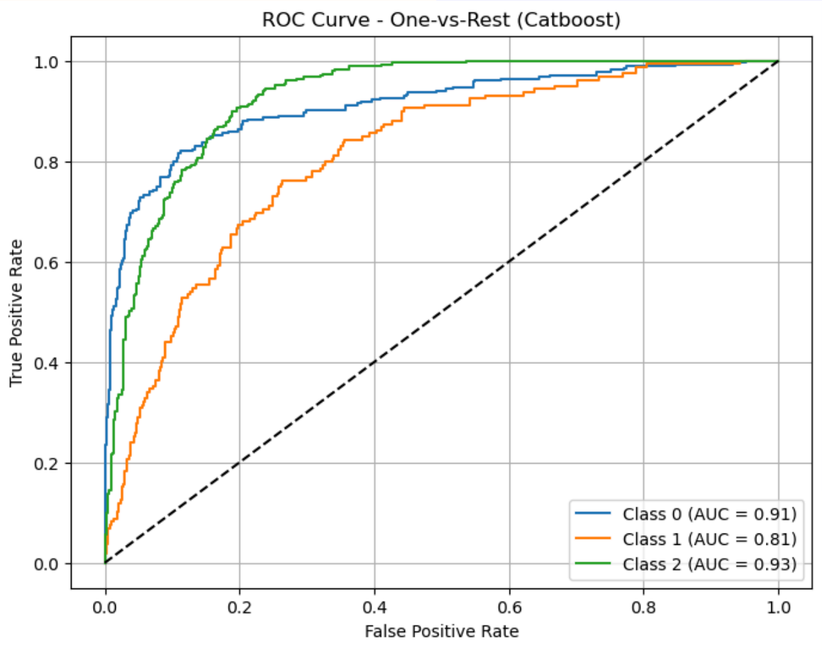


p-value = 0.6265 → No statistically significant difference between the two models.

Model evaluation

CATBOOST

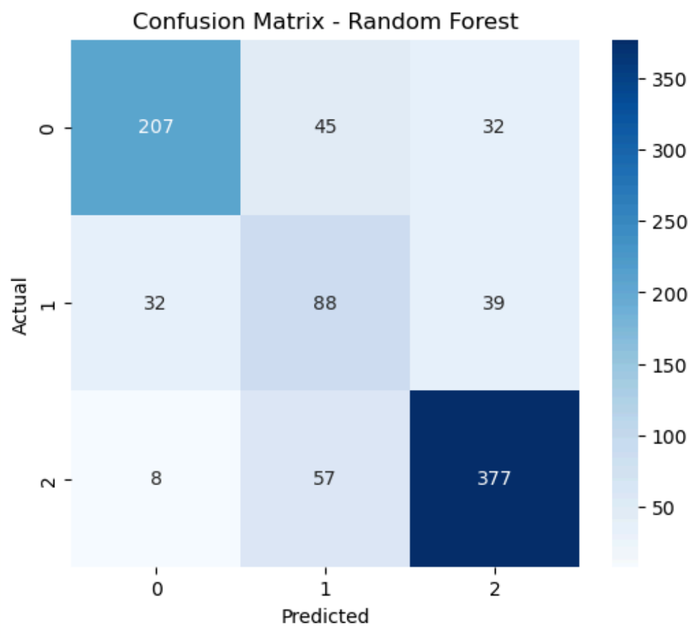


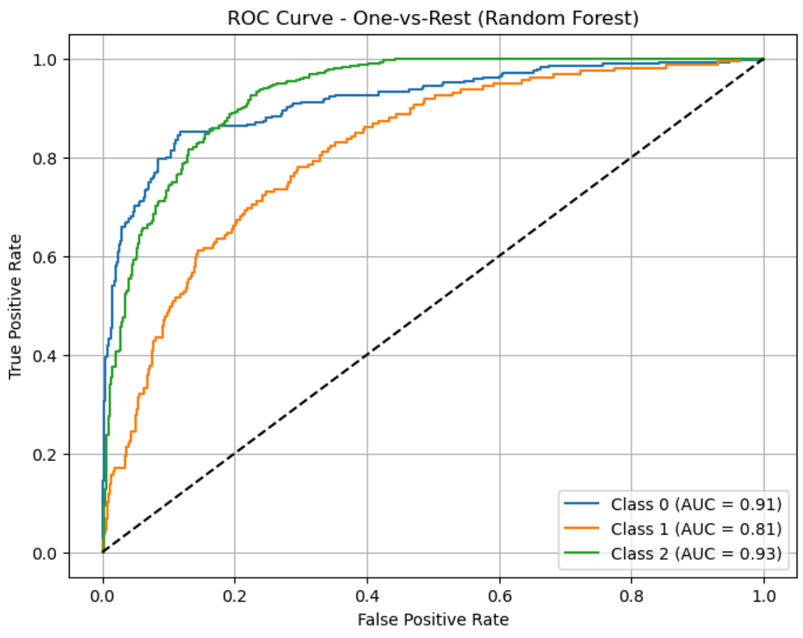



```
=== Classification Report: Catboost ===

              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
```

Model evaluation

RANDOM FOREST



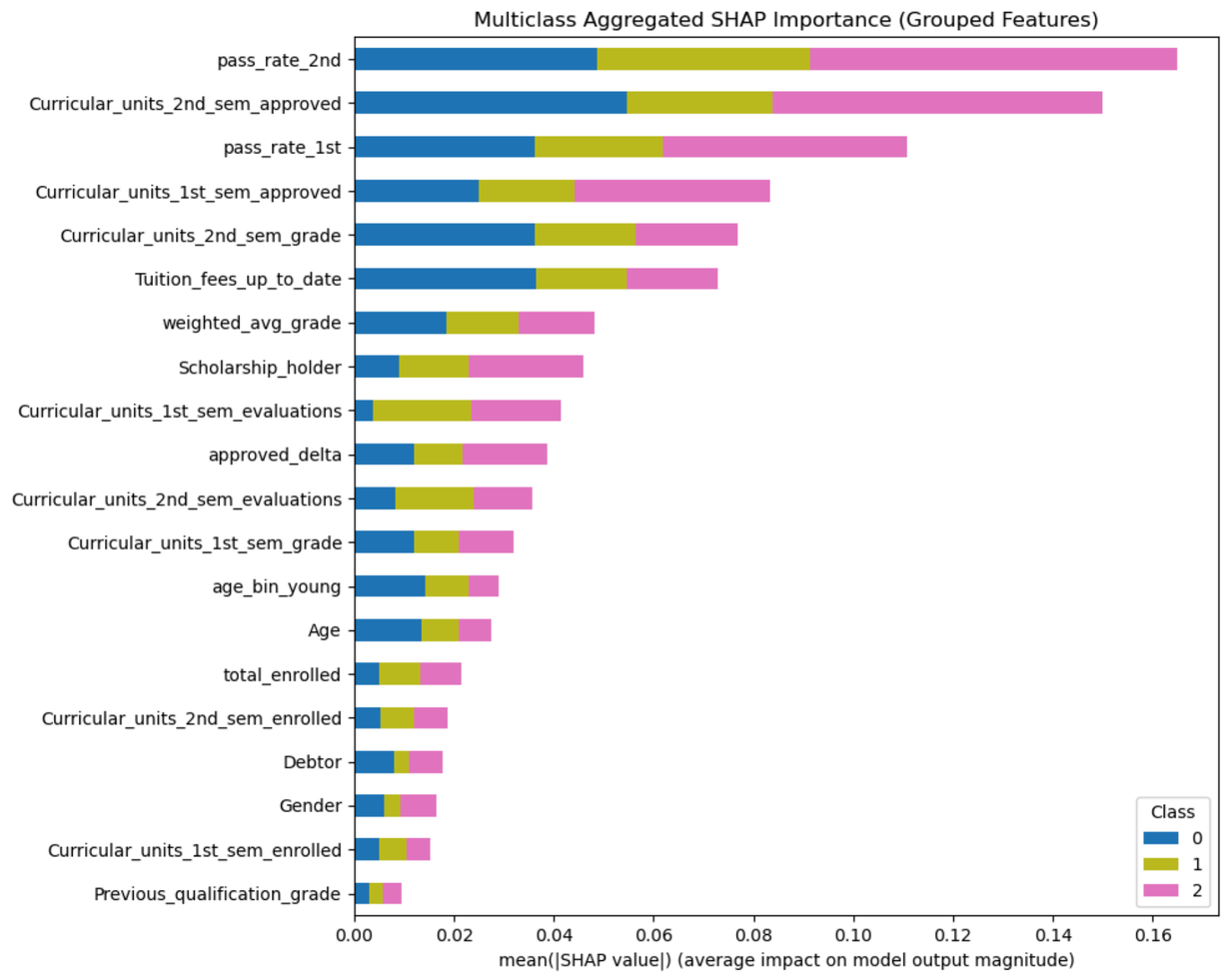



```
=== Classification Report: Random Forest ===

              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
```

Model Explainability

RANDOM FOREST

SHAP summary

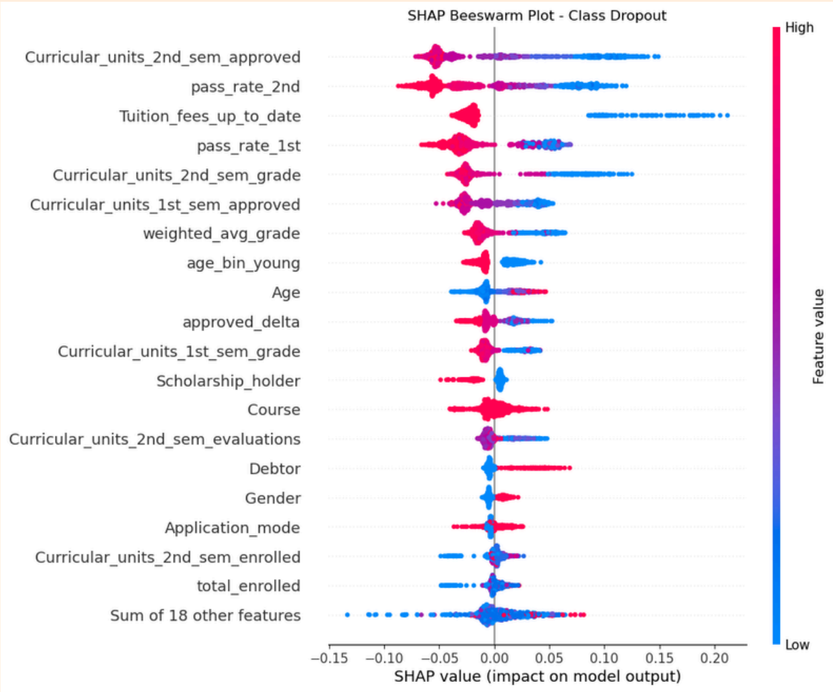


Model Explainability

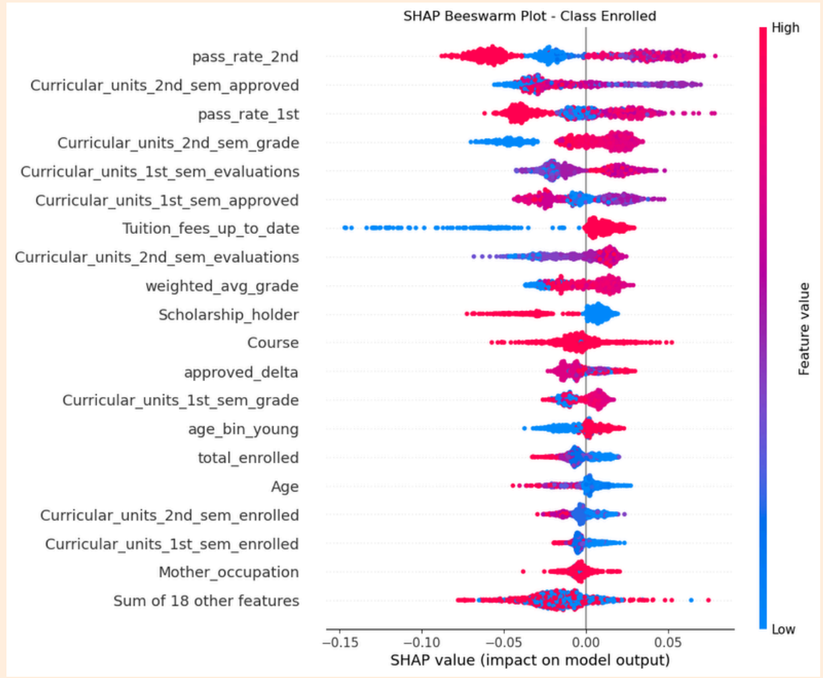
RANDOM FOREST

Global explanations

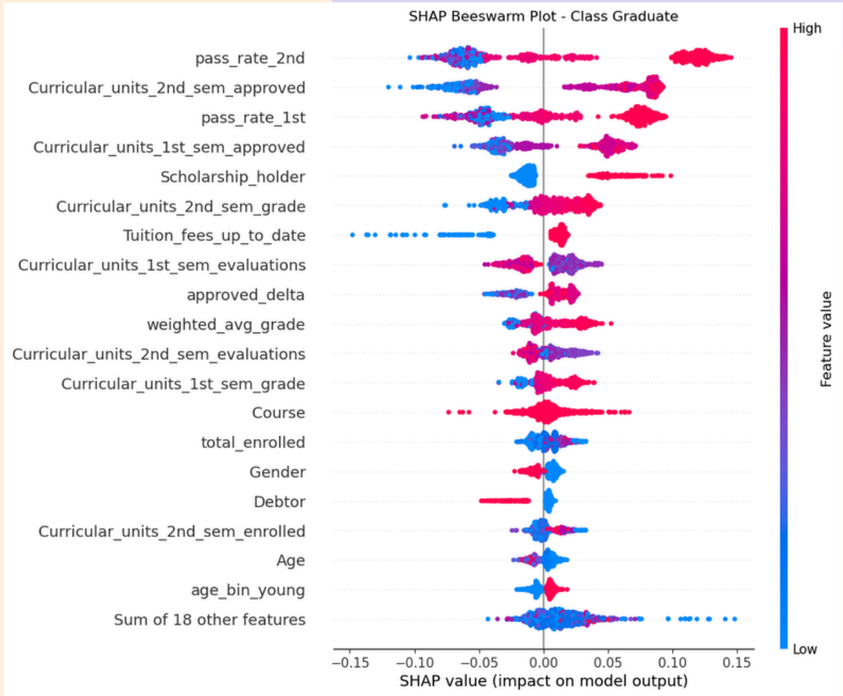
Dropout



Enrolled



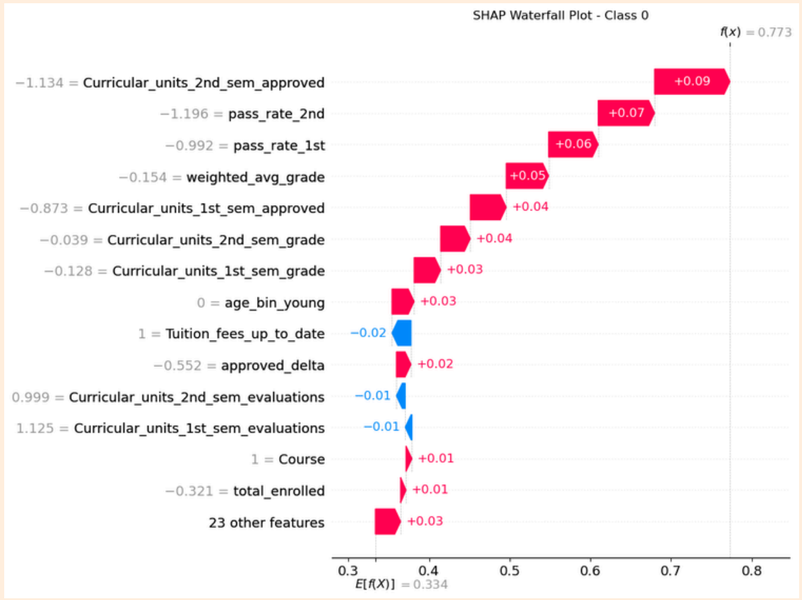
Graduated



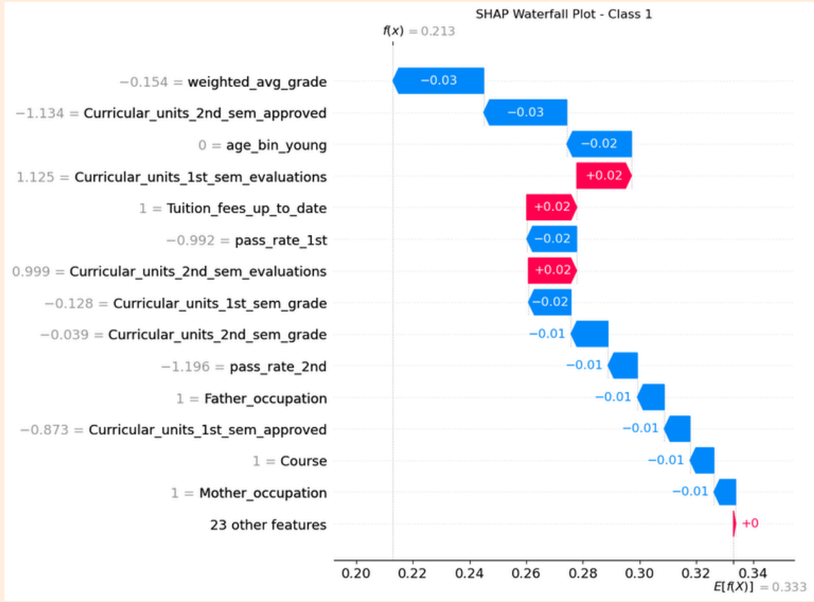
Model Explainability

RANDOM FOREST

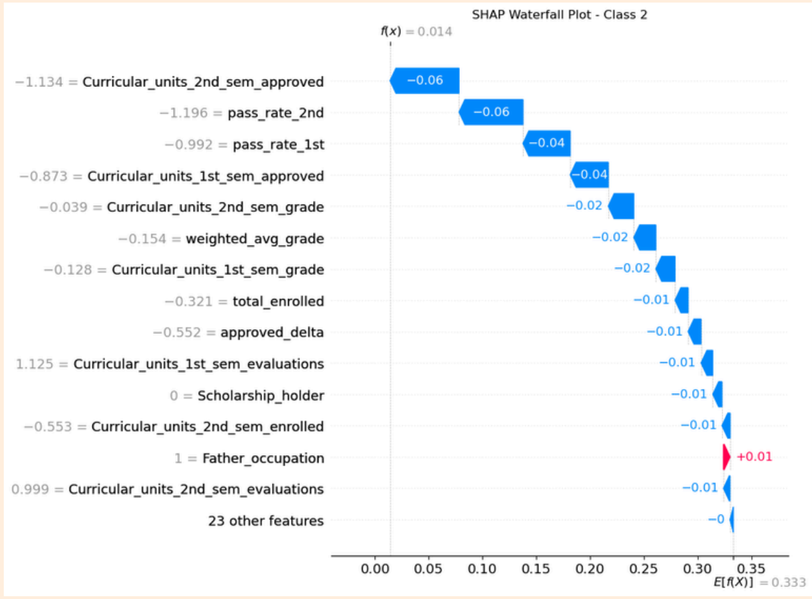
Dropout

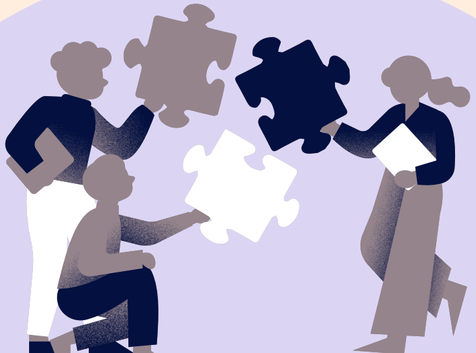


Enrolled



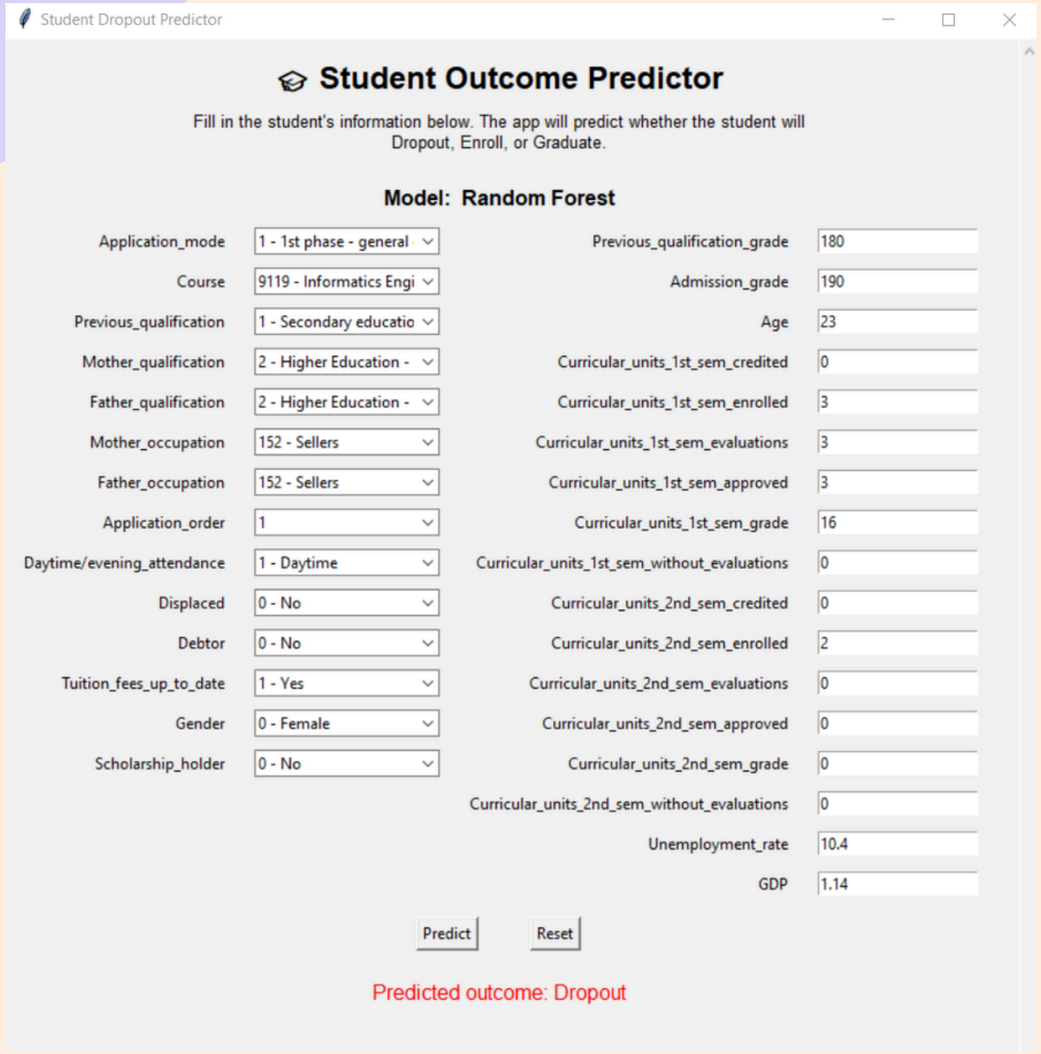
Graduated

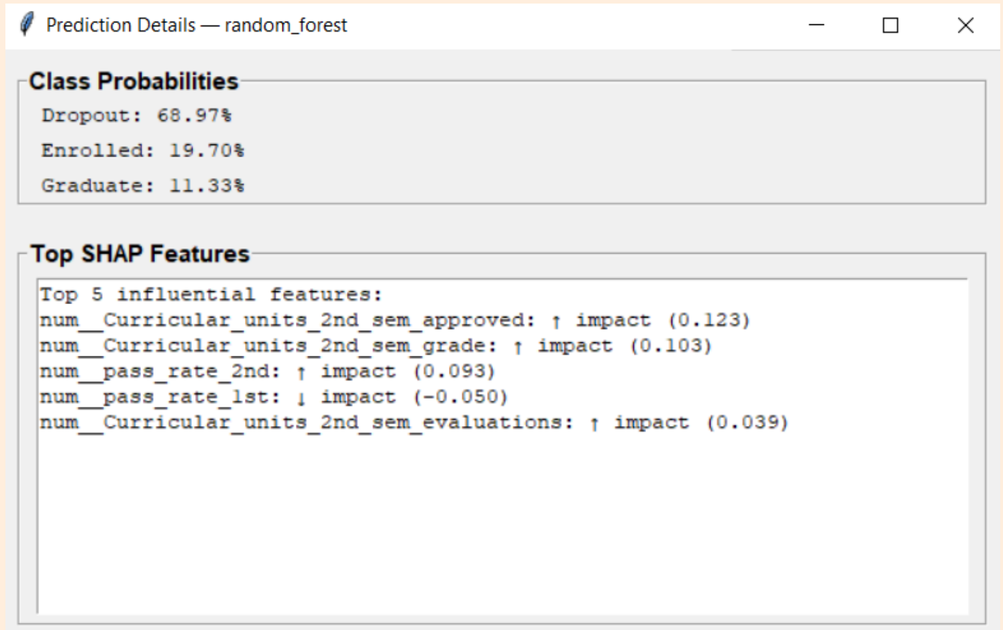




- These plots illustrate how the model arrives at **individual predictions** for each class using SHAP values.
- **Local explanations** help us to understand how individual student profiles influence the model's decision-making.

Interface


Student Dropout Predictor
Student Outcome Predictor
Fill in the student's information below. The app will predict whether the student will Dropout, Enroll, or Graduate.
Model: Random Forest
Application_mode
1 - 1st phase - general
Course
9119 - Informatics Engi
Previous_qualification
1 - Secondary educatio
Mother_qualification
2 - Higher Education -
Father_qualification
2 - Higher Education -
Mother_occupation
152 - Sellers
Father_occupation
152 - Sellers
Application_order
1
Daytime/evening_attendance
1 - Daytime
Displaced
0 - No
Debtor
0 - No
Tuition_fees_up_to_date
1 - Yes
Gender
0 - Female
Scholarship_holder
0 - No
Previous_qualification_grade
180
Admission_grade
190
Age
23
Curricular_units_1st_sem_credited
0
Curricular_units_1st_sem_enrolled
3
Curricular_units_1st_sem_evaluations
3
Curricular_units_1st_sem_approved
3
Curricular_units_1st_sem_grade
16
Curricular_units_1st_sem_without_evaluations
0
Curricular_units_2nd_sem_credited
0
Curricular_units_2nd_sem_enrolled
2
Curricular_units_2nd_sem_evaluations
0
Curricular_units_2nd_sem_approved
0
Curricular_units_2nd_sem_grade
0
Curricular_units_2nd_sem_without_evaluations
0
Unemployment_rate
10.4
GDP
1.14
Predict
Reset
Predicted outcome: Dropout


Prediction Details — random_forest
Class Probabilities
Dropout: 68.97%
Enrolled: 19.70%
Graduate: 11.33%
Top SHAP Features
Top 5 influential features:
num__Curricular_units_2nd_sem_approved: ↑ impact (0.123)
num__Curricular_units_2nd_sem_grade: ↑ impact (0.103)
num__pass_rate_2nd: ↑ impact (0.093)
num__pass_rate_1st: ↓ impact (-0.050)
num__Curricular_units_2nd_sem_evaluations: ↑ impact (0.039)

Bibliography

- 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

- 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

- 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

THANKS FOR YOUR ATTENTION