DATA MINING AND STATISTICS

Assignment: Technical Report

Student: S244679

Contents

1 Abstract

This study explores the application of data mining and machine learning techniques to predict academic outcomes from a dataset containing pre-enrolment and first year information of Portuguese undergraduate students. It furthers the work of a study (Martins et al., 2021) which classified students, using only data available at the time of enrolment, into three categories of academic achievement after the normal duration of the course. The aim being to identify students at risk of academic failure. They provided the additional first year data and encouraged further development of the solution. This motivated the first research question to see if prediction accuracy and F1-score could be improved by incorporating first-year academic data. Other studies have success with binary classification, which motivated the second research question to see if the results from the first question can be improved by modelling and predicting each of the target classes separately.

During data quality testing, 4% of the data had to be removed. Data visualisation and hypothesis testing were used to aid feature selection. The target class of the data is imbalanced and various sampling techniques were tested to overcome this, SMOTE proving to be most effective. A Stratified 5-Fold cross-validation of each model was performed to collate performance metrics and determine models for hyperparameter tuning. This study first modelled the three-category classification and achieved results that improved on (Martins et al., 2021) but still struggled with the minority class. Logistic Regression and Gradient Boosting were most effective. Work continued with a set of binary classification models, which achieved higher accuracy and F1-scores. Future direction encourages the acquisition of more data, the testing of additional boosting models and exploration into the balance between performance and overfitting when hyperparameter tuning.

2 Introduction

A global challenge for universities is the high student dropout rates, with a significant proportion of students failing to complete their courses within the standard timeframe. Universities, aiming to mitigate this problem, wish to identify students at-risk of dropout using data collected at enrolment and during coursework. Limited research exists on which data features are most predictive, leaving institutions without clear guidance. This report aims to improve that knowledge by analysing and modelling a Portuguese undergraduate student dataset.

The introduction document for this Portuguese dataset (Martins et al., 2021) helps to define the problem facing universities and explains the features of the data. They employed machine learning models to try to classify student success into three groups. They used Random Forest (RF), Logistic Regression (LR), Decision Tree (DT) and Gradient Boosting (GB). Class imbalance was noted in the data and SMOTE sampling was used to address it. Their best results had F1-scores of 0.83 (Success), 0.68 (Failure), 0.44 (Relative success) with an average of 0.65 and an Accuracy of 0.73. They struggled when predicting the “Relative success” cases who were the students who remained enrolled but failed to graduate on time. Notably, the researchers excluded first-year performance data, suggesting its potential for future model improvement. That is a motivation for the first research question:

- **RQ1) Can the prediction accuracy and F1-score of the Portuguese study be improved by incorporating first-year academic data?**

(Delogu et al., 2023), relates to higher education in Italy. It uses a different dataset and focused on predicting a binary outcome of whether the student dropped out. They used first year academic performance to aid prediction and found it to be the most important feature of their models with the RF achieving a 90% accuracy. The dataset is not available for study so cannot be used to compare with the models generated during this research. They are the motivation of the second research question:

- **RQ2) Can the prediction accuracy and F1-score results of RQ1 be improved by modelling and predicting each of the target classes separately?**

(Yağcı, 2022) studied the grades of 1854 Turkish students and used midterm exam results to predict final exam grades using models including: RF, Support Vector Machines (SVM), LR, and Naïve Bayes (NB). They concluded that midterm grades were a good predictor of final results. RF had highest accuracy with 74.6% of samples classified correctly. They suggest their model could be used to identify students that might fail and drop-out, but didn't investigate further. Unlike other studies they limited their features to just grade, faculty and department. They did not use socio-demographics features, which this research will include.

(Al-Alawi et al., 2023) used supervised ML to study 11 years of student data from a university in Oman. They aimed to predict students that would underperform at university based on pre-enrolment data. InfoGain algorithm was used to discover the key features for prediction; university study duration and secondary school performance were found to be significant. Also, male students performed worse than female students. This was a complex investigation using an ensemble of ML algorithms, but its findings indicate that admission grade and gender from this data set could be important.

It is noted throughout these studies that the RF algorithm is effective. Although education systems vary throughout the world, the high rate of dropout at university is a common area of concern. If first-year academic results are a good predictor for the Portuguese dataset, as they were in Italy, then that can encourage other universities to collect and test data for their own students. Improving academic prediction could potentially improve the lives of thousands of students worldwide.

3 Data Description

The dataset consists of 4424 rows of student data from a Portuguese polytechnic. There are 36 independent features; 18 are categorical (8 of which are binary), 18 are continuous. The 18 categorical features are numerically encoded in the datafile, so care will be taken to treat them as categorical.

TARGET is the only dependant feature which has three possible values: Graduate, Enrolled, Dropout.

TARGET shows imbalance with 2209 Graduate students, 1421 Dropout, and 794 Enrolled.

GENDER is also imbalanced with 2868 females and 1556 males.

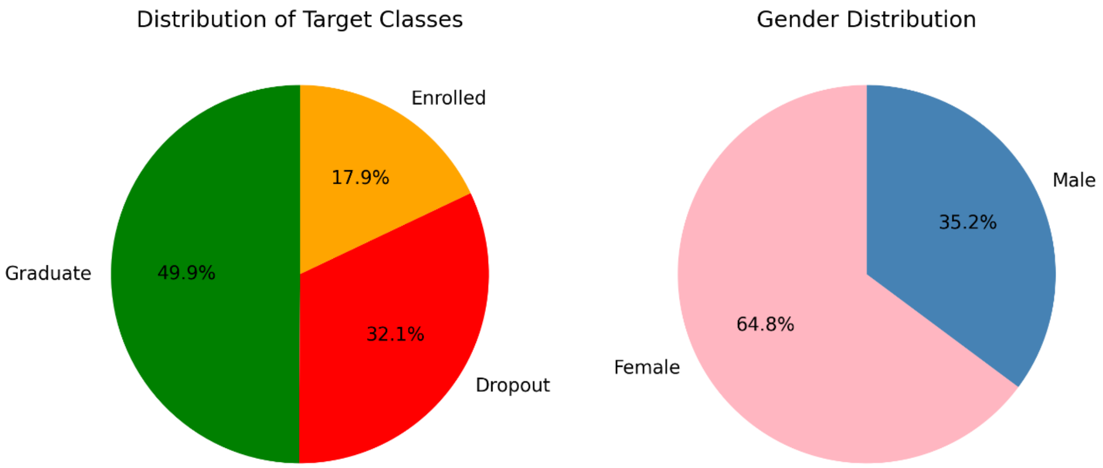

Figure 1 *– Pie charts showing Target and Gender Distribution*

Some categorical features have large numbers of underrepresented classes, they could be an issue when modelling. The dataset was declared as having no missing values, but this will need to be validated. AGE_AT_ENROLLMENT appears to be skewed based on the summary information so requires investigation. The _CREDITED and _WITHOUT_EVALUATIONS semester data also suggest either skew or the presence of outliers.

Column Name	Unique Classes	Data Type
MARITAL_STATUS	6	int64
APPLICATION_MODE	18	int64
APPLICATION_ORDER	8	int64
COURSE	17	int64
DAYTIME/EVENING_ATTENDANCE	2	int64
PREVIOUS_QUALIFICATION	17	int64
NATIONALITY	21	int64
MOTHER_QUALIFICATION	29	int64
FATHER_QUALIFICATION	34	int64
MOTHER_OCCUPATION	32	int64
FATHER_OCCUPATION	46	int64
DISPLACED	2	int64
EDUCATIONAL_SPECIAL_NEEDS	2	int64
DEBTOR	2	int64
TUITION_FEES_UP_TO_DATE	2	int64
GENDER	2	int64
SCHOLARSHIP_HOLDER	2	int64
INTERNATIONAL	2	int64
TARGET	3	object

Table 1 *– Categorical features*

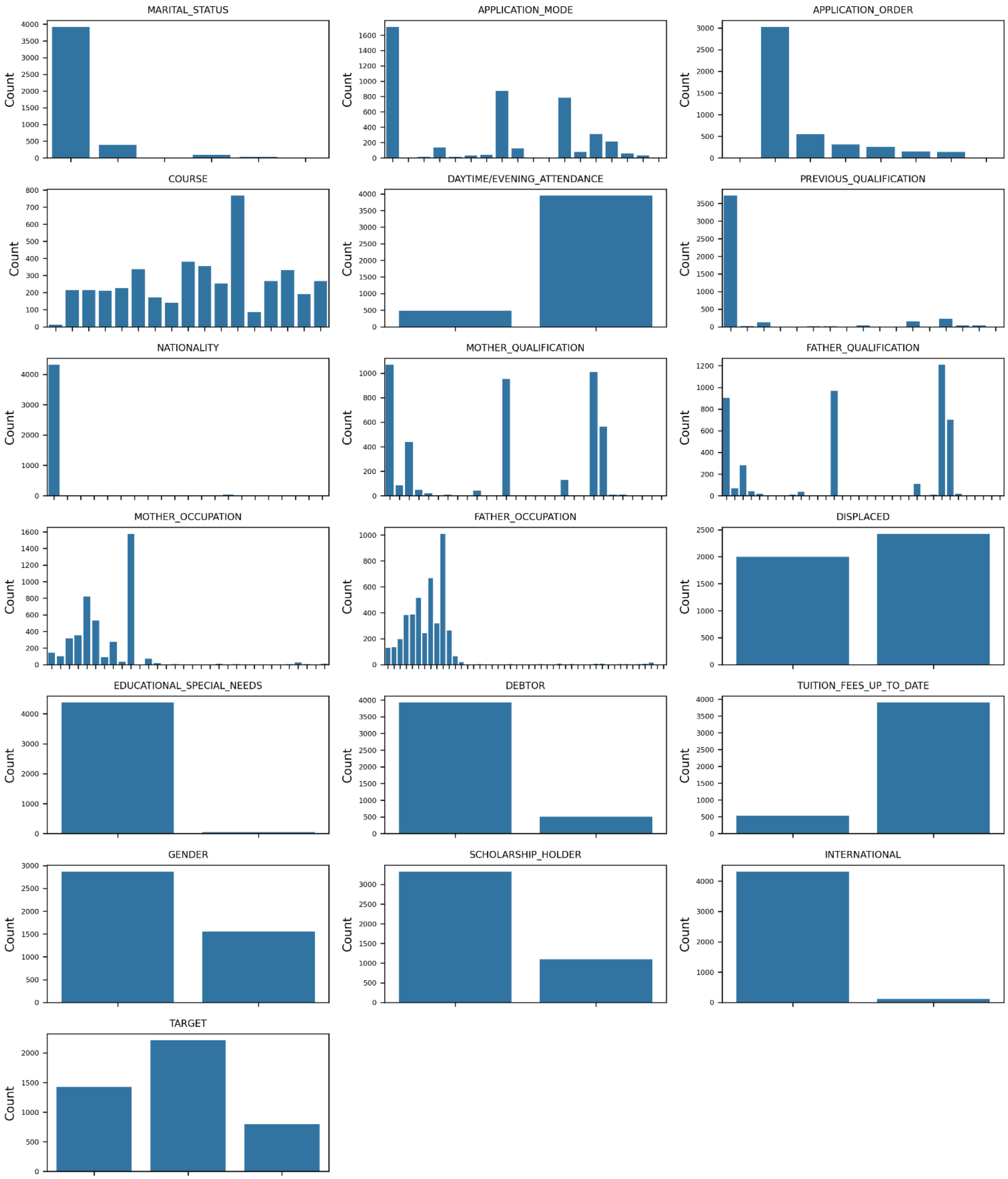

Figure 2 *– Class distribution of categorical features*

Column Name	Mean	Median	Mode	Min	Max	Range	Q1 (25%)	Q3 (75%)	IQR	Std Dev	Variance
PREVIOUS_ QUALIFICATION_GRADE	132.61	133.1	133.1	95	190	95	125	140	15	13.19	173.93
ADMISSION_GRADE	126.98	126.1	130	95	190	95	117.9	134.8	16.9	14.48	209.73
AGE_AT_ENROLLMENT	23.27	20	18	17	70	53	19	25	6	7.59	57.57
CURR_UNITS_1ST_SEM_ CREDITED	0.71	0	0	0	20	20	0	0	0	2.36	5.57
CURR_UNITS_1ST_SEM_ ENROLLED	6.27	6	6	0	26	26	5	7	2	2.48	6.15
CURR_UNITS_1ST_SEM_ EVALUATIONS	8.3	8	8	0	45	45	6	10	4	4.18	17.46
CURR_UNITS_1ST_SEM_ APPROVED	4.71	5	6	0	26	26	3	6	3	3.09	9.57
CURR_UNITS_1ST_SEM_ GRADE	10.64	12.29	0	0	18.88	18.88	11	13.4	2.4	4.84	23.46
CURR_UNITS_1ST_SEM_ WITHOUT_ EVALUATIONS	0.14	0	0	0	12	12	0	0	0	0.69	0.48
CURR_UNITS_2ND_SEM _CREDITED	0.54	0	0	0	19	19	0	0	0	1.92	3.68
CURR_UNITS_2ND_SEM _ENROLLED	6.23	6	6	0	23	23	5	7	2	2.2	4.82
CURR_UNITS_2ND_SEM _EVALUATIONS	8.06	8	8	0	33	33	6	10	4	3.95	15.59
CURR_UNITS_2ND_SEM _APPROVED	4.44	5	6	0	20	20	2	6	4	3.01	9.09
CURR_UNITS_2ND_SEM _GRADE	10.23	12.2	0	0	18.57	18.57	10.75	13.33	2.58	5.21	27.15
CURR_UNITS_2ND_SEM _WITHOUT_ EVALUATIONS	0.15	0	0	0	12	12	0	0	0	0.75	0.57
UNEMPLOYMENT_RATE	11.57	11.1	7.6	7.6	16.2	8.6	9.4	13.9	4.5	2.66	7.1
INFLATION_RATE	1.23	1.4	1.4	-0.8	3.7	4.5	0.3	2.6	2.3	1.38	1.91
GDP	0	0.32	0.32	-4.06	3.51	7.57	-1.7	1.79	3.49	2.27	5.15

Table 2 *– Continuous features*

4 Data Preprocessing

A series of validation checks were performed on the data to look for missing values. It was proven that there are no NaN or infinite values, however there are 180 students with a mix of Target data results but with no curricular unit data. As this data is required for RQ1 and RQ2 those rows were dropped to prevent model distortion.

A likely incorrect value with APPLICATION_ORDER data was identified and corrected.

The exact meaning of the DEBTOR feature is unclear, but it was proven to be independent from TUITION_FEES_UP_TO_DATE.

To address the underrepresented class issue, the categorical features were processed to group together any categories containing < 3% of the rows. E.g. FATHERS_OCCUPATION reduced from 46 classes to 11. It is expected that this will benefit the modelling process.

After grouping, it was found that NATIONALITY and INTERNATIONAL held the same information, so NATIONALITY was dropped.

Each of the continuous columns were processed to handle outliers - rather than removing rows, it was decided to cap outliers using the interquartile range. The capping boundaries were set at (Q1 – 1.5 * IQR) and (Q3 + 1.5 * IQR).

Histograms and Box Plots were used to visually check the processed data distributions; it was decided that the curricular units features _1ST_SEM_CREDITED, _2ND_SEM_CREDITED, _1ST_SEM_WITHOUT_EVALUATIONS, and _2ND_SEM_WITHOUT_EVALUATIONS where not useful and were dropped.

In preparation for machine learning, eight new features were created as described in Table 3, three of which being new dependant Target features specifically for RQ2.

Feature	**Description**
TARGET_IS_GRADUATE	1 if TARGET value is 'Graduate' otherwise 0
TARGET_IS_ENROLLED	1 if TARGET value is 'Enrolled' otherwise 0
TARGET_IS_DROPOUT	1 if TARGET value is 'Dropout' otherwise 0
AGE_GROUP_AT_ENROLLMENT	Grouping of Age_At_Enrollment into bins to address the bias in the data
CURR_UNITS_TOTAL_ENROLLED	Total of curricular units Enrolled for both semesters
CURR_UNITS_TOTAL_EVALUATIONS	Total of curricular unit Evaluations for both semesters
CURR_UNITS_TOTAL_APPROVED	Total of curricular units Approved for both semesters
CURR_UNITS_AVG_GRADE	Average curricular unit Grade over both semesters

Table 3 *– Features created*

5 Statistical and Machine Learning Methods

Both research questions are classification problems, RQ1 having three target classes, RQ2 having two target classes (1/0) for each of the three newly created Boolean Target features.

For Machine Learning (ML) models it is important to reduce dataset dimensionality to avoid overfitting and improve performance. Some features can be discarded visually or statistically, but others require significance testing to find statistical relevance.

For the categorical features, Chi-squared will be used to test if there is a relationship between the feature and the target(s) of RQ1 and RQ2. The Null Hypothesis will be that the two features are independent. If the p-value < = 0.05 then H_0 will be rejected.

For the continuous features, Kruskall-Wallis (KW) H-tests will be used to test if there is a relationship between the feature and the three RQ1 target categories. KW is a non-parametric test so will handle the imperfect distributions in the data and is suitable for three target groups. Mann-Whitney (MW) will be used to run the relationship test for the two group targets of RQ2. The Null Hypothesis will be that the two features are independent. If the p-value < = 0.05 then H_0 will be rejected.

The most important features of these tests will be selected for machine learning. Multicollinearity will be checked using a correlation heatmap and features with correlation above 70% will be rejected.

The dataset will be split 70:30 for training and testing. LR, RF, DT, GB models will be run against training data using Stratified K-Fold cross validation with five splits, due to the target class imbalance over/under sampling techniques will be tested including SMOTE and ADASYN. LR will require the use of a scaler, and all models will require categorical features to be One-Hot Encoded.

In addition to the modelling results provided by (Martins et al., 2021), an initial 'baseline' will be established by running the ML models against data that excludes all first-year academic features. Then the selected features for ML will be used for TARGET (for RQ1) and for TARGET_IS_GRADUATE, TARGET_IS_ENROLLED, TARGET_IS_DROPOUT (for RQ2).

The best models will proceed to a hyperparameter tuning stage to produce final results for RQ1. The best RQ1 classifier(s) will be used to evaluate RQ2.

Predictions will be made using the testing data and a classification report and confusion matrix will be displayed.

For RQ1 to be true, the accuracy and F1-score of the untuned model must be higher than the baseline results and those values of the tuned model must be higher than the results of (Martins et al., 2021). If RQ1 is true it provides evidence that first year academic data is important when predicting if a student will graduate in a timely manner.

For RQ2 to be true, the precision and F1-score must be higher than the results of RQ1 for a given target group.

6 Results

Following the data preprocessing, the data was explored using summary information and visualisations.

6.1 Categorical data

The number of unique classes has successfully been reduced; this has reduced the quantity of minority classes. However, the Boolean categories of EDUCATIONAL_SPECIAL_NEEDS and INTERNATIONAL still have minorities of 1.1% and 2.5% respectively, so are unlikely to be useful.

Column Name	Unique Classes	Data Type
MARITAL_STATUS	3	int64
APPLICATION_MODE	7	int64
APPLICATION_ORDER	7	int64
COURSE	16	int64
DAYTIME/EVENING_ATTENDANCE	2	int64
PREVIOUS_QUALIFICATION	4	int64
NATIONALITY	2	int64
MOTHER_QUALIFICATION	6	int64
FATHER_QUALIFICATION	6	int64
MOTHER_OCCUPATION	8	int64
FATHER_OCCUPATION	11	int64
DISPLACED	2	int64
EDUCATIONAL_SPECIAL_NEEDS	2	int64
DEBTOR	2	int64
TUITION_FEES_UP_TO_DATE	2	int64
GENDER	2	int64
SCHOLARSHIP_HOLDER	2	int64
INTERNATIONAL	2	int64
TARGET	3	object

Table 4 *–categorical features after processing*

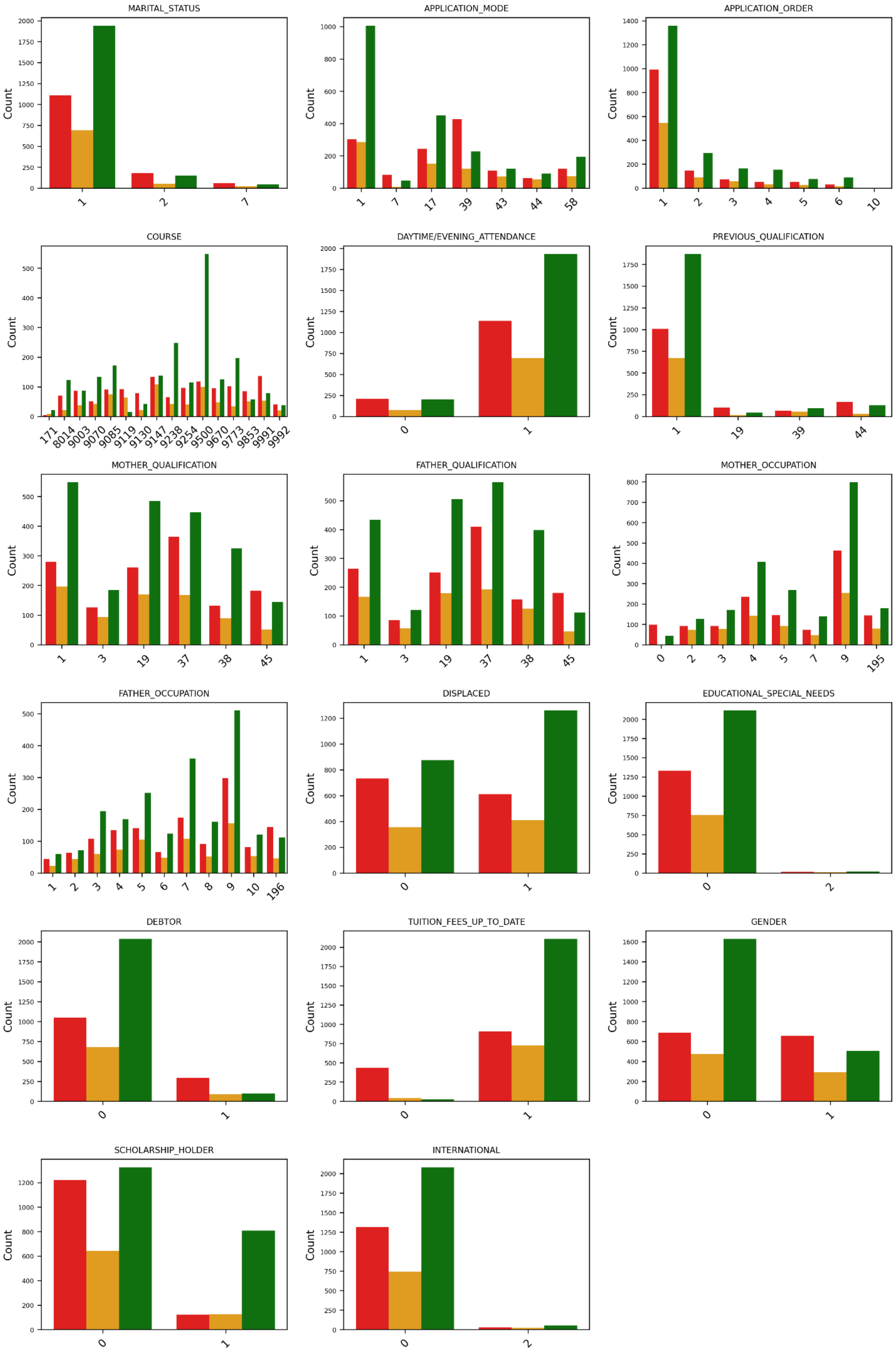

Figure 4 *– Categorical feature class distribution by target*

It was noted that most students are:

- single
- studying at their first choice university in the daytime
- not classed as educational special needs
- not classed as DEBTOR
- not late with tuition fee payments.

When looking in terms relative counts of each Target:

- COURSE 9500 (Nursing) stands out as having a high proportion of Graduate students.
- DEBTOR students and ones late with tuition fees have a relatively high proportion of Dropout.
- Being a SCHOLARSHIP_HOLDER shows a high probability of graduating.

From the Gender percentage chart

- There are more females (65.7%) than males (34.3%).
- COURSE 9119 (Informatics Engineering) is primarily male and had a very low proportion of graduates.
- 76% of graduates are female.

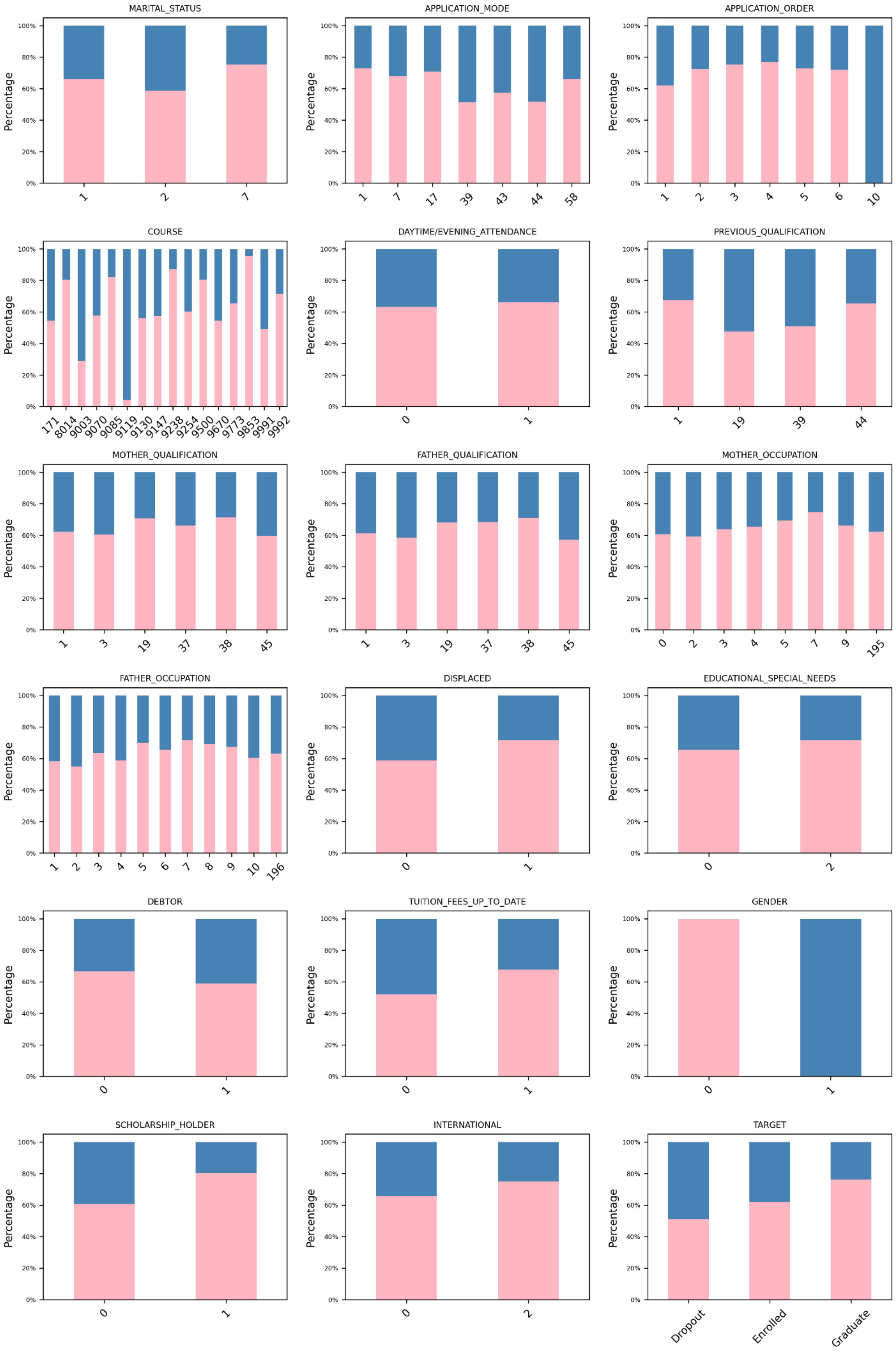

***Figure 5** – Categorical feature relative percentage distribution by gender*

6.2 Continuous data

Column Name	Mean	Median	Mode	Min	Max	Range	Q1 (25%)	Q3 (75%)	IQR	Std Dev	Variance
PREVIOUS_QUALIFICATION_GRADE	132.26	133	133.1	100	164	64	124	140	16	12.69	161.01
ADMISSION_GRADE	126.41	125.8	130	95	158.55	63.55	117.8	134.1	16.3	13.73	188.65
AGE_AT_ENROLLMENT	22.56	20	18	17	34	17	19	25	6	5.46	29.8
CURR_UNITS_1ST_SEM_CREDITED	0	0	0	0	0	0	0	0	0	0	0
CURR_UNITS_1ST_SEM_ENROLLED	6.23	6	6	4.5	8.5	4	6	7	1	1.06	1.12
CURR_UNITS_1ST_SEM_EVALUATIONS	8.5	8	8	2.5	14.5	12	7	10	3	2.93	8.57
CURR_UNITS_1ST_SEM_APPROVED	4.78	5	6	0	10.5	10.5	3	6	3	2.61	6.82
CURR_UNITS_1ST_SEM_GRADE	12.1	12.33	7.98	7.98	16.7	8.71	11.25	13.43	2.18	2.01	4.03
CURR_UNITS_1ST_SEM_WITHOUT_EVALUATIONS	0	0	0	0	0	0	0	0	0	0	0
CURR_UNITS_2ND_SEM_CREDITED	0	0	0	0	0	0	0	0	0	0	0
CURR_UNITS_2ND_SEM_ENROLLED	6.37	6	6	2	10	8	5	7	2	1.36	1.85
CURR_UNITS_2ND_SEM_EVALUATIONS	8.31	8	8	0	16	16	6	10	4	3.36	11.31
CURR_UNITS_2ND_SEM_APPROVED	4.56	5	6	0	10.5	10.5	3	6	3	2.75	7.57
CURR_UNITS_2ND_SEM_GRADE	11.86	12.33	7.38	7.38	17.04	9.67	11	13.42	2.42	2.34	5.49
CURR_UNITS_2ND_SEM_WITHOUT_EVALUATIONS	0	0	0	0	0	0	0	0	0	0	0
UNEMPLOYMENT_RATE	11.56	11.1	7.6	7.6	16.2	8.6	9.4	13.9	4.5	2.66	7.1
INFLATION_RATE	1.23	1.4	1.4	-0.8	3.7	4.5	0.3	2.6	2.3	1.38	1.9
GDP	0.02	0.32	0.32	-4.06	3.51	7.57	-1.7	1.79	3.49	2.27	5.17

***Table 5** –continuous features after processing*

The capping of outliers has reduced the long tail of AGE_AT_ENROLLMENT.

GRADE features are normally distributed, but the 1st and 2nd semester grades have a high quantity of low scores that spoils the distribution but cannot be considered an outlier as they appear indicative of Dropout (see figure 6).

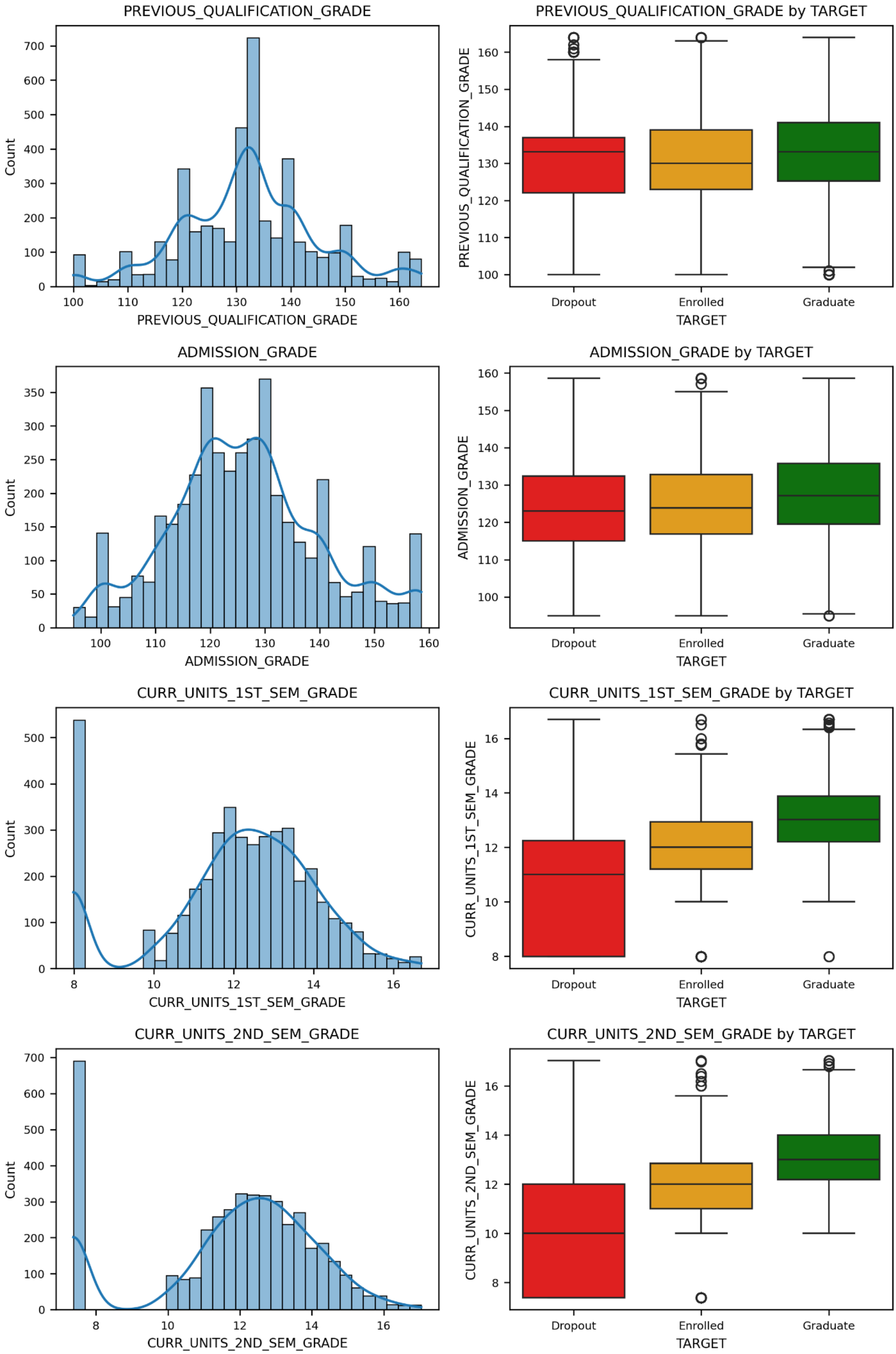

***Figure 6** – Grade distribution and box plots by target*

Units ENROLLED and APPROVED for both semesters show good separation of targets on the box plots in ascending order: Dropout, Enrolled, Graduate.

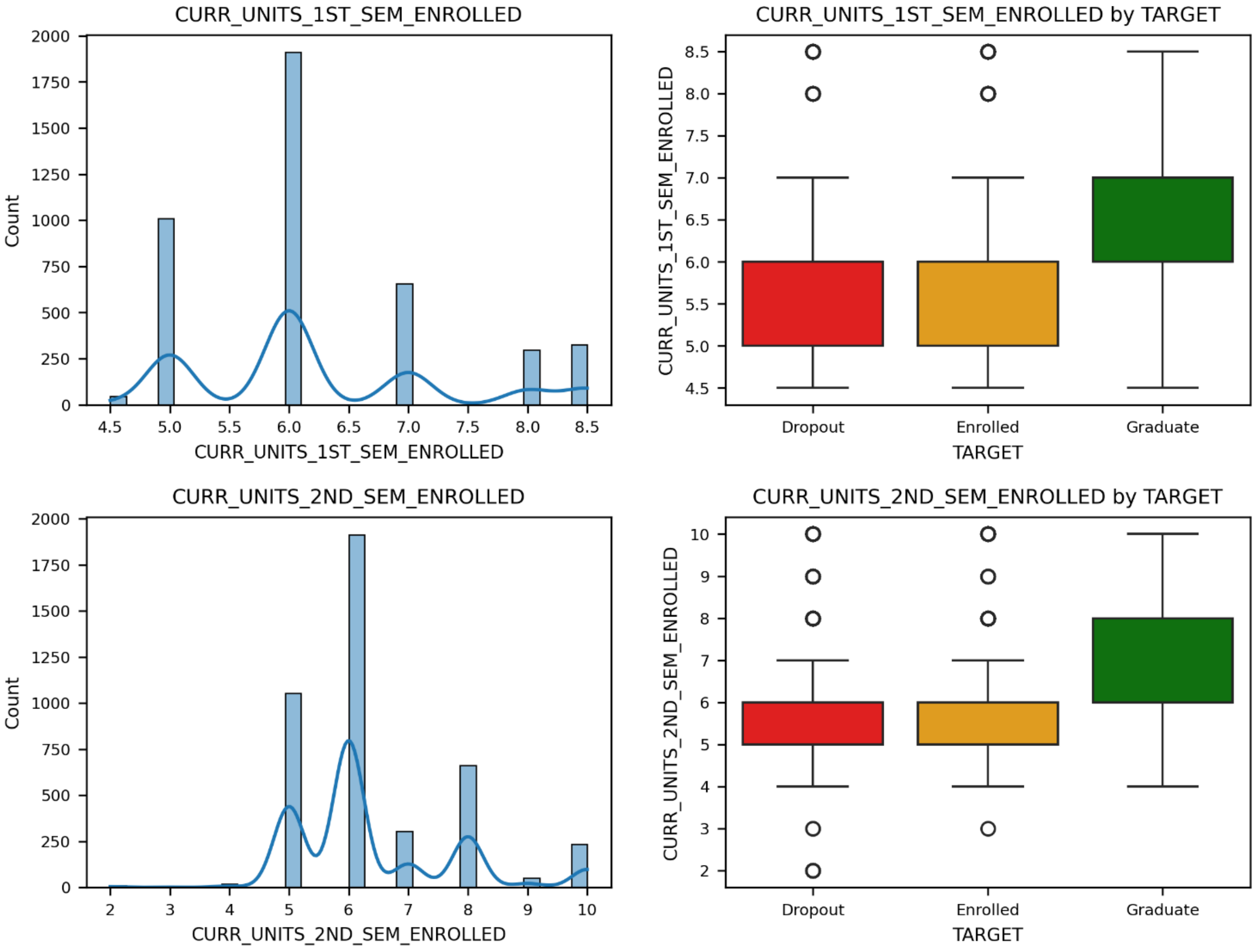

***Figure 7** – Enrolled units, distribution and box plots by target*

AGE_AT_ENROLMENT is heavily skewed towards younger students and indicates that younger students are more likely to graduate.

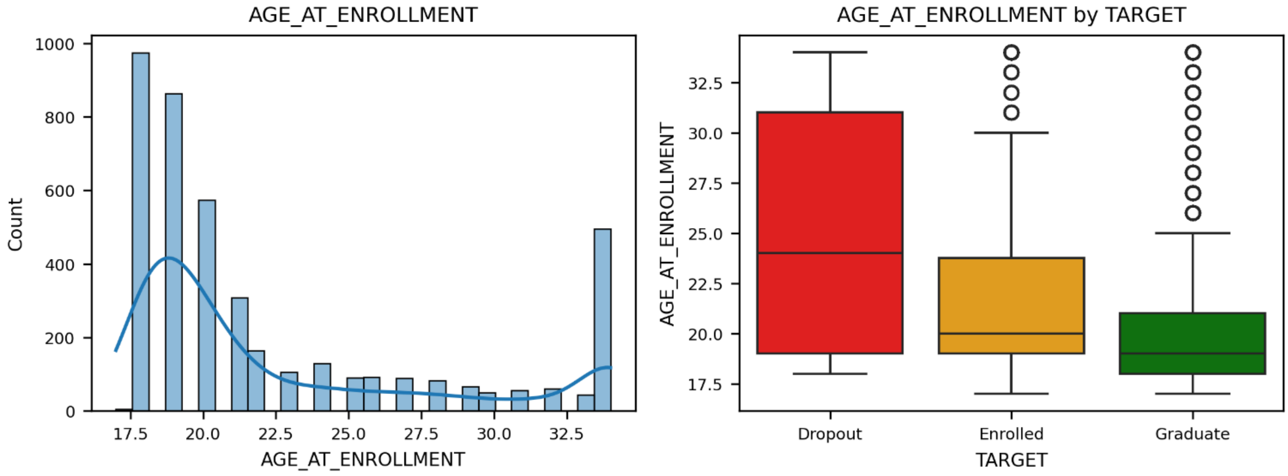

***Figure 8** – AGE_AT_ENROLMENT distribution and box plots by target*

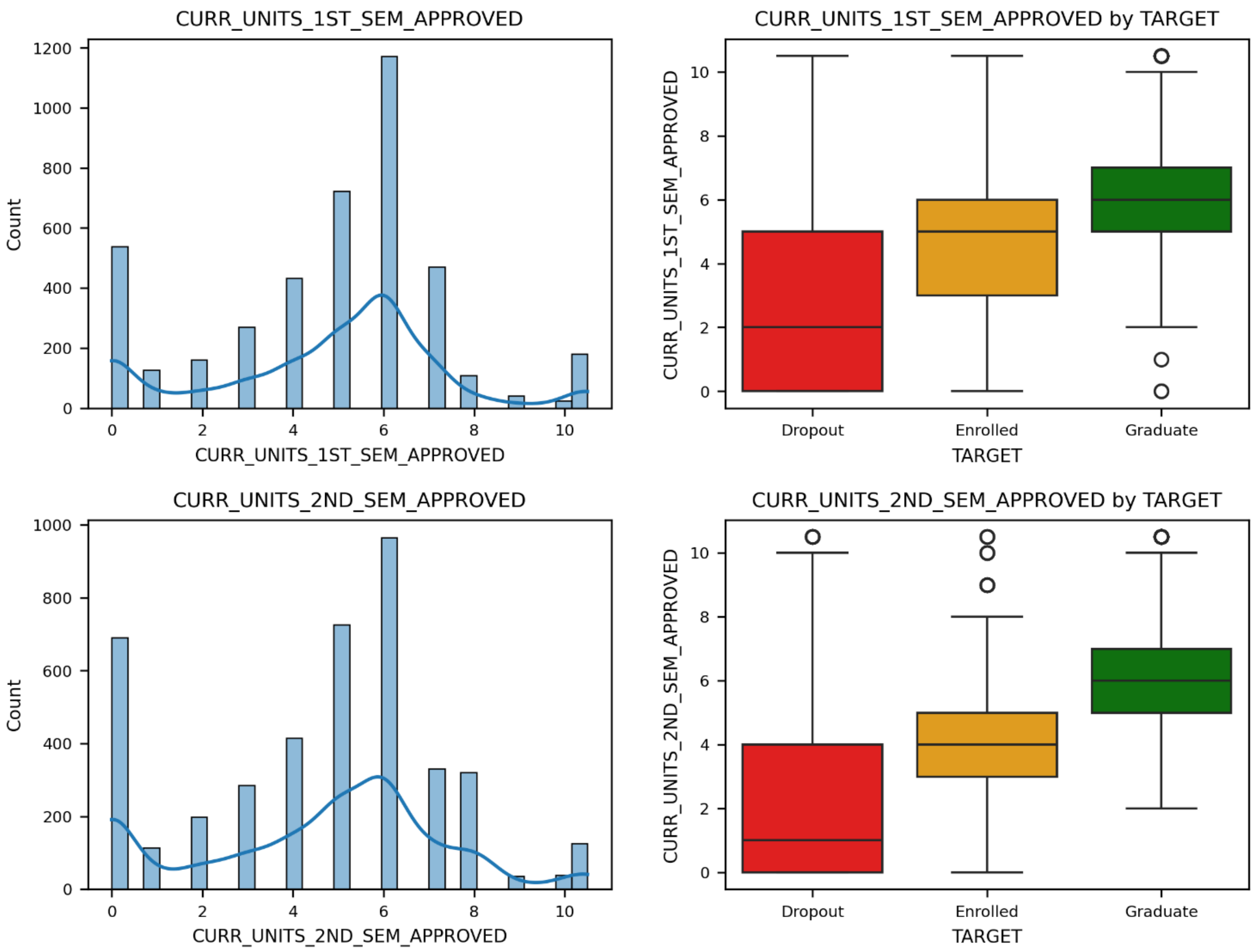

Figure 9 *– Approved units, distribution and box plots by target*

6.3 Feature Selection for Machine Learning

The Chi-Squared tests against the multi-class TARGET for RQ1 or individual targets for RQ2 helped to determine the important categorical features.

INTERNATIONAL and EDUCATIONAL_SPECIAL_NEEDS are not related to any target category. These features help classify Graduate and Dropout, but not Enrolled: FATHER_QUALIFICATION, MARITAL_STATUS, FATHER_OCCUPATION, DAYTIME/EVENING_ATTENDANCE, DISPLACED, DEBTOR

Plotting a Pearson correlation heatmap showed high correlation between first and second semester features. The economic features do not appear to be correlated to any of the Targets. TARGET_IS_ENROLLED shows very low correlation to all independent features except for EVALUATIONS.

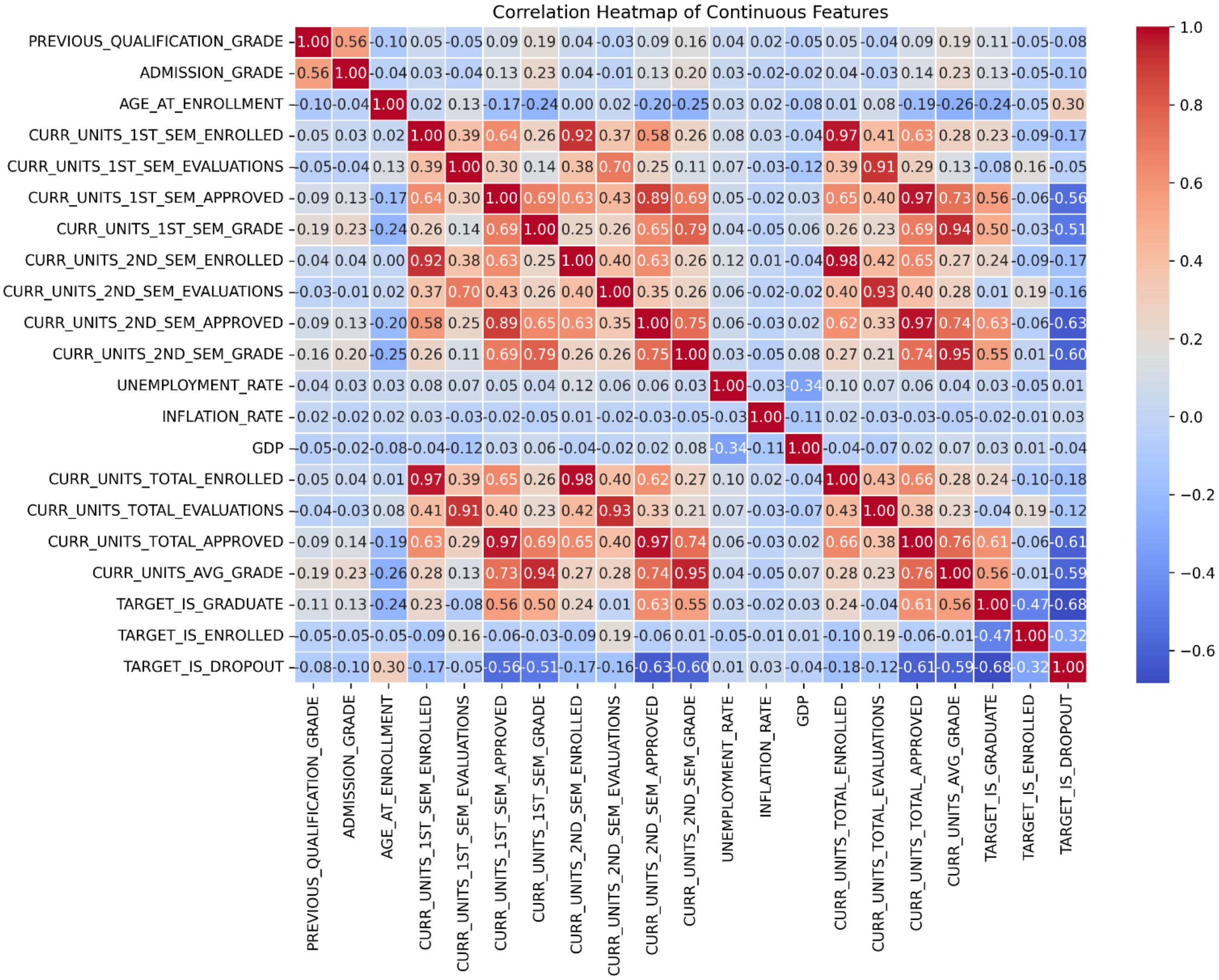

Figure 10 – Pearson Correlation Heatmap of continuous features

First and Second semester grades also showed correlation in the pair-plots. The spike of low grades leading to dropout is apparent.

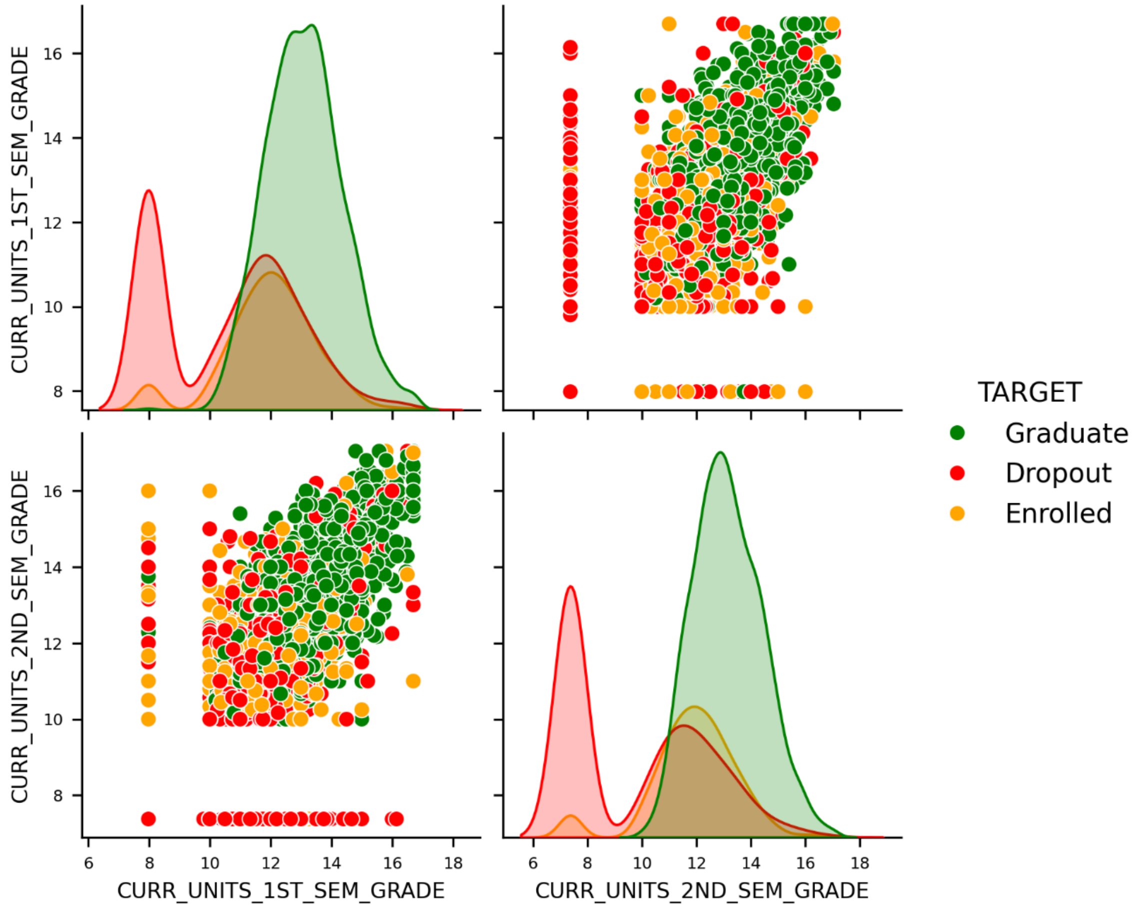

***Figure 11** – 1st and 2nd Semester Grades pair-plots*

To avoid multicollinearity only one feature of each type of curricular unit could be selected for machine learning; the visualisations in conjunction with Kruskall-Wallis and Mann-Whitney tests were used to achieve this.

6.4 Machine Learning

While developing the LR model a scaler was required, MinMaxScaler and RobustScaler where both tested but StandardScaler gave the best results. This is likely due to outliers being handled during data preprocessing.

Due to the TARGET imbalance, it was expected that resampling would be required to achieve a quality model, this proved to be the case with all the best performing models using a sampler.

For RQ1, the Machine Learning before Hyperparameter Tuning (HPT) showed GB with SMOTE to have a slightly better F1-score than LR and RF. They exceeded both the baseline scores (which models with the first-year academic features excluded) and the best scores of the Portuguese study. DT underperformed the other models.

RQ1 Machine Learning results before Hyperparameter tuning

- All three Target classes (Graduate, Enrolled, Dropout)

Feature	Accuracy	Precision	Recall	F1-score
Dummy classifier	0.503	0.168	0.333	0.223
Logistic regression	0.771	0.710	0.677	0.683
Logistic regression (Random Under-sampling)	0.740	0.706	0.712	0.701
Logistic regression (Random Over-sampling)	0.746	0.710	0.712	0.704
Logistic regression (SMOTE)	0.753	0.716	0.720	0.712
Logistic regression (ADASYN)	0.745	0.703	0.710	0.702
Random forest	0.770	0.717	0.676	0.685
Random forest (Random Under-sampling)	0.744	0.720	**0.722**	0.710
Random Forest (Random Over-sampling)	0.774	0.725	0.701	0.710
Random Forest (SMOTE)	0.776	0.729	0.697	0.708
Random Forest (ADASYN)	0.775	0.724	0.699	0.708
Decision Tree	0.753	0.716	0.643	0.644
Decision Tree (Random Under-sampling)	0.707	0.691	0.682	0.673
Decision Tree (Random Over-sampling)	0.721	0.711	0.696	0.686
Decision Tree (SMOTE)	0.723	0.701	0.685	0.682
Decision Tree (ADASYN)	0.721	0.704	0.677	0.673
Gradient Boosting	0.776	0.725	0.696	0.706
Gradient Boosting (Random Under-sampling)	0.737	0.711	0.715	0.702
Gradient Boosting (Random Over-sampling)	0.766	0.725	0.721	**0.719**
Gradient Boosting (SMOTE)	**0.781**	**0.734**	0.711	**0.719**
Gradient Boosting (ADASYN)	0.780	0.730	0.709	0.717

Table 6 *–* ***RQ1*** *Machine Learning results before Hyperparameter tuning*

Baseline for RQ1 (no first-year academic features)

- All three Target classes (Graduate, Enrolled, Dropout)

Feature	Accuracy	Precision	Recall	F1-score
Dummy classifier	0.503	0.168	0.333	0.223
Logistic regression	**0.651**	0.578	0.536	0.531
Logistic regression (Random Under-sampling)	0.595	0.579	0.579	0.565
Logistic regression (Random Over-sampling)	0.605	0.580	0.580	0.570
Logistic regression (SMOTE)	0.600	0.575	0.574	0.565
Logistic regression (ADASYN)	0.603	0.569	0.571	0.564
Random forest	0.646	0.584	0.547	0.550
Random forest (Random Under-sampling)	0.585	0.577	**0.581**	0.563
Random Forest (Random Over-sampling)	0.636	0.569	0.558	0.560
Random Forest (SMOTE)	0.642	0.575	0.549	0.553
Random Forest (ADASYN)	0.640	0.574	0.551	0.554
Decision Tree	0.625	0.504	0.490	0.454
Decision Tree (Random Under-sampling)	0.500	0.543	0.513	0.486
Decision Tree (Random Over-sampling)	0.475	0.575	0.507	0.467
Decision Tree (SMOTE)	0.555	0.589	0.477	0.463

Decision Tree (ADASYN)	0.522	**0.596**	0.481	0.464
Gradient Boosting	0.649	0.575	0.542	0.542
Gradient Boosting (Random Under-sampling)	0.584	0.576	0.577	0.559
Gradient Boosting (Random Over-sampling)	0.620	0.579	0.579	**0.575**
Gradient Boosting (SMOTE)	0.647	0.579	0.558	0.562
Gradient Boosting (ADASYN)	0.647	0.574	0.556	0.558

Table 7** – Baseline for **RQ1

After HPT, GB with SMOTE had the best accuracy and average scores, but LR with SMOTE was a close second and had better results for the Enrolled minority class.

RQ1 Model Hyperparameter Tuning Classification Results

Model	Target	Precision	Recall	F1-score	Support
Logistic Regression with SMOTE	Dropout	0.83	0.69	0.75	403
	Enrolled	0.43	**0.60**	**0.50**	230
	Graduate	**0.85**	0.82	0.84	641
	accuracy			0.74	1274
	macro avg	0.70	**0.70**	0.70	1274
	weighted avg	**0.77**	0.74	0.75	1274
Decision Tree with Random Over Sampler	Dropout	0.85	0.61	0.71	403
	Enrolled	0.38	0.50	0.43	230
	Graduate	0.80	0.85	0.82	641
	accuracy			0.71	1274
	macro avg	0.68	0.65	0.65	1274
	weighted avg	0.74	0.71	0.71	1274
Random Forest with Random Under Sampler	Dropout	**0.86**	0.58	0.69	403
	Enrolled	0.38	0.56	0.45	230
	Graduate	0.82	0.84	0.83	641
	accuracy			0.71	1274
	macro avg	0.69	**0.70**	0.66	1274
	weighted avg	0.75	0.71	0.72	1274
Gradient Boosting with SMOTE	Dropout	0.83	**0.75**	**0.79**	403
	Enrolled	**0.49**	0.42	0.45	230
	Graduate	0.80	**0.90**	**0.85**	641
	accuracy			**0.76**	1274
	macro avg	**0.71**	0.69	**0.70**	1274
	weighted avg	0.76	**0.76**	**0.76**	1274

***Table 8** – **RQ1** Model Hyperparameter Tuning Classification Results*

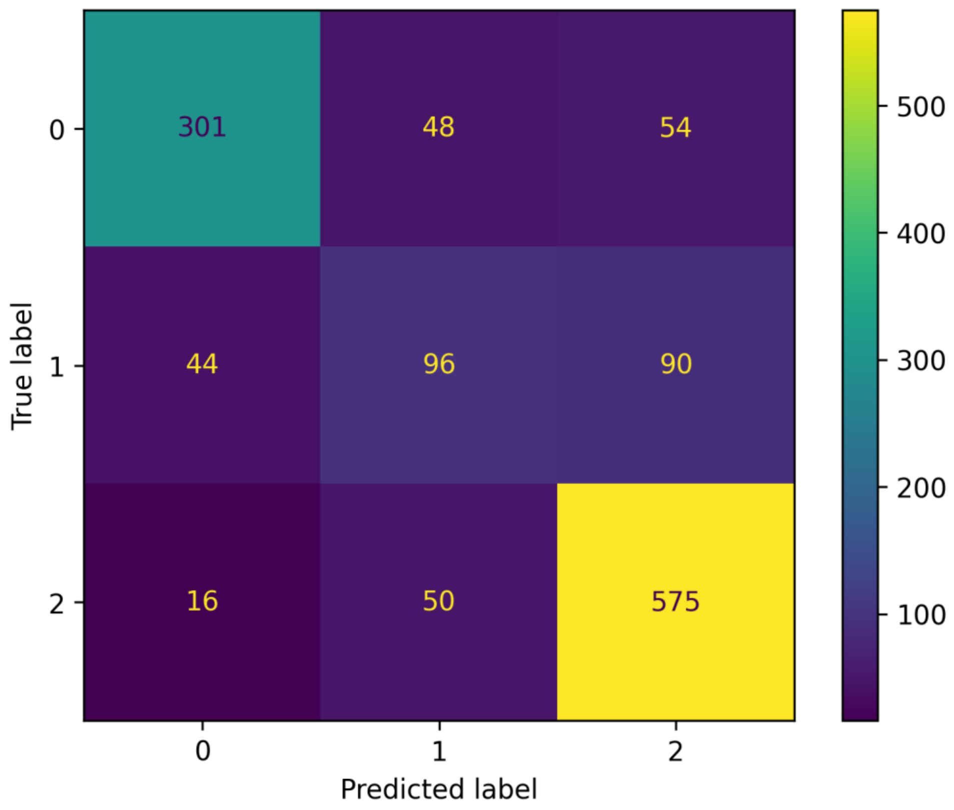

***Figure 12** – RQ1 - Gradient Boosting with SMOTE confusion matrix (0:Dropout, 1:Enrolled, 2:Graduate)*

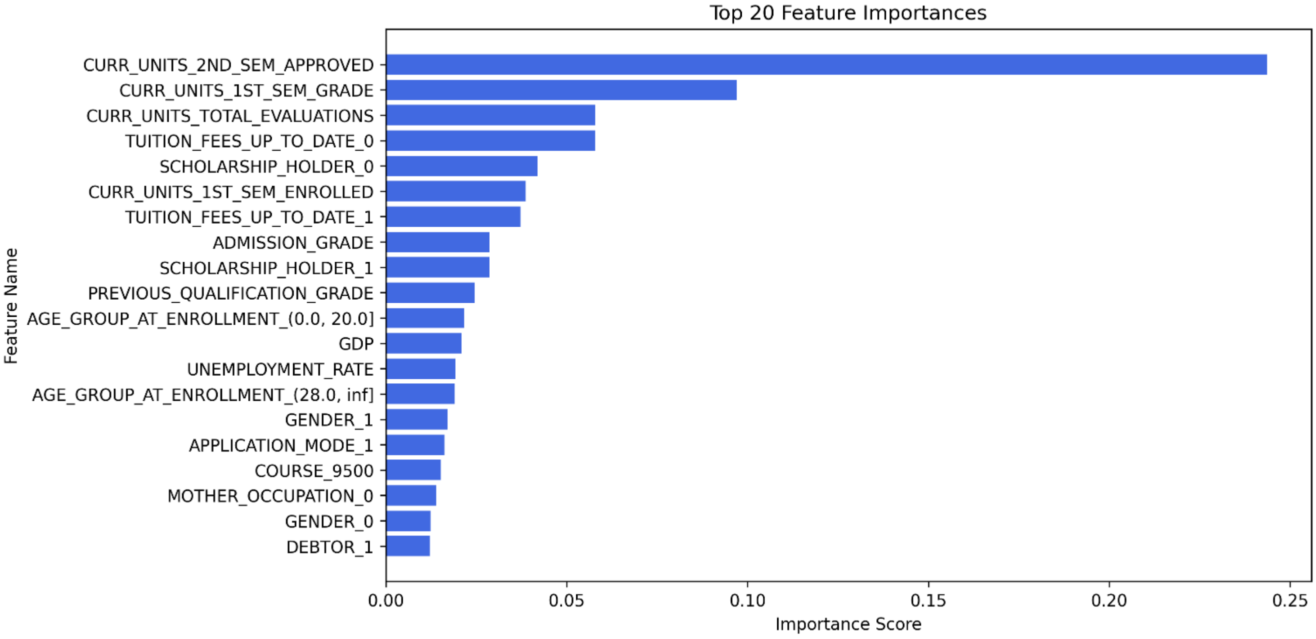

***Figure 13** – RQ1 - Gradient Boosting with SMOTE feature importance*

The GB with SMOTE and LR with SMOTE models were taken forward to RQ2. For the Graduate & Dropout class tests, the precision and overall accuracy and F1-score beat the RQ1 results. From the scores point of view, LR was the best model for Graduate and GB was best for Dropout. Enrolled is less clear as GB achieved higher precision than RQ1 but has a lower recall meaning that it missed many positive cases. When looking at the confusion matrices (figures 14,15,16) if the objective is to find the highest number of true positives, then LR would be chosen for all three models.

Model Hyperparameter Tuning Classification Results Applied to RQ2

Model	Target	Precision	Recall	F1-score	Support
Gradient Boosting with SMOTE targeting IS_GRADUATE	NOT Graduate	**0.89**	0.80	0.84	633
	Graduate	0.82	**0.90**	0.86	641
	accuracy			0.85	1274
	macro avg	0.85	0.85	0.85	1274
	weighted avg	0.85	0.85	0.85	1274
Logistic Regression with SMOTE targeting IS_GRADUATE	NOT Graduate	**0.89**	**0.81**	**0.85**	633
	Graduate	**0.83**	**0.90**	**0.86**	641
	accuracy			**0.86**	1274
	macro avg	**0.86**	**0.86**	**0.86**	1274
	weighted avg	**0.86**	**0.86**	**0.86**	1274
Model	**Target**	**Precision**	**Recall**	**F1-score**	**Support**
Gradient Boosting with SMOTE targeting IS_ENROLLED	NOT Enrolled	0.86	**0.93**	**0.90**	1044
	Enrolled	**0.52**	0.33	0.40	230
	accuracy			**0.82**	1274
	macro avg	**0.69**	0.63	**0.65**	1274
	weighted avg	**0.80**	**0.82**	**0.81**	1274
Logistic Regression with SMOTE targeting IS_ENROLLED	NOT Enrolled	**0.90**	0.73	0.81	1044
	Enrolled	0.34	**0.65**	**0.45**	230
	accuracy			0.71	1274
	macro avg	0.62	**0.69**	0.63	1274
	weighted avg	**0.80**	0.71	0.74	1274
Model	**Target**	**Precision**	**Recall**	**F1-score**	**Support**
Gradient Boosting with SMOTE targeting IS_DROPOUT	NOT Dropout	0.88	**0.95**	**0.91**	871
	Dropout	**0.86**	0.72	0.78	403
	accuracy			**0.87**	1274
	macro avg	**0.87**	0.83	**0.85**	1274
	weighted avg	**0.87**	**0.87**	**0.87**	1274
Logistic Regression with SMOTE targeting IS_DROPOUT	NOT Dropout	**0.91**	0.89	0.90	871
	Dropout	0.77	**0.80**	**0.79**	403
	accuracy			0.86	1274
	macro avg	0.84	**0.85**	0.84	1274
	weighted avg	0.86	0.86	0.86	1274

Table 9** – Model Hyperparameter Tuning Classification Results Applied to **RQ2

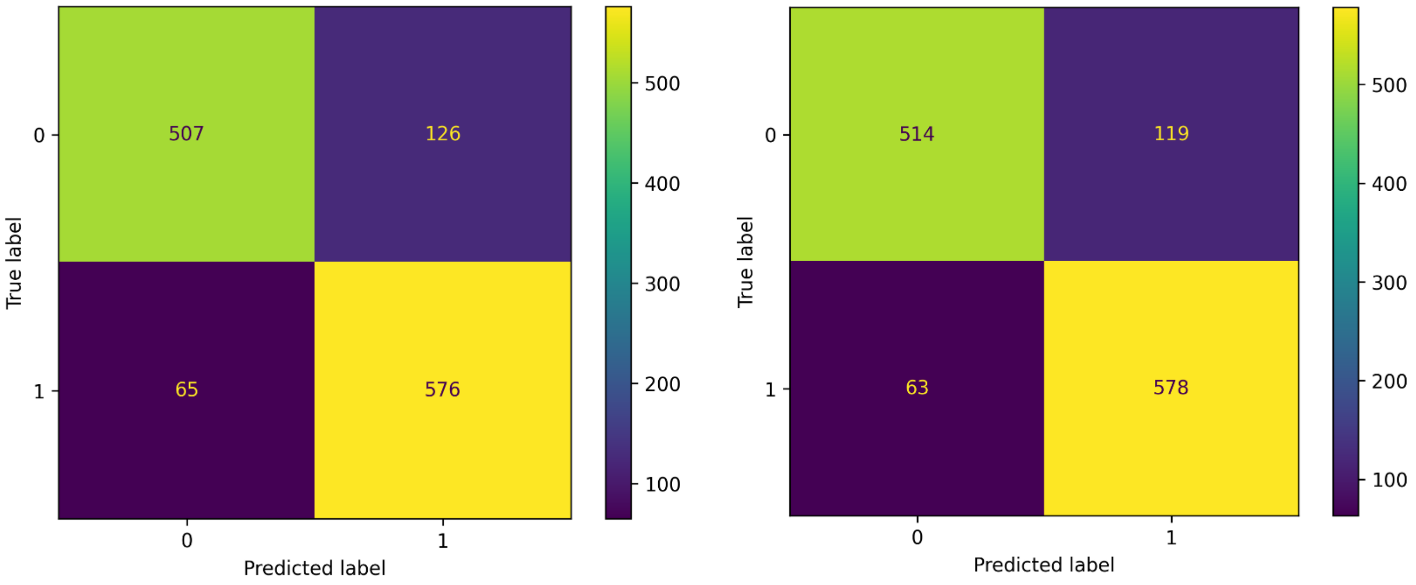

Figure 14 *– RQ2 (Graduate) confusion matrices – Gradient Boosting (left), Logistic Regression (right)*

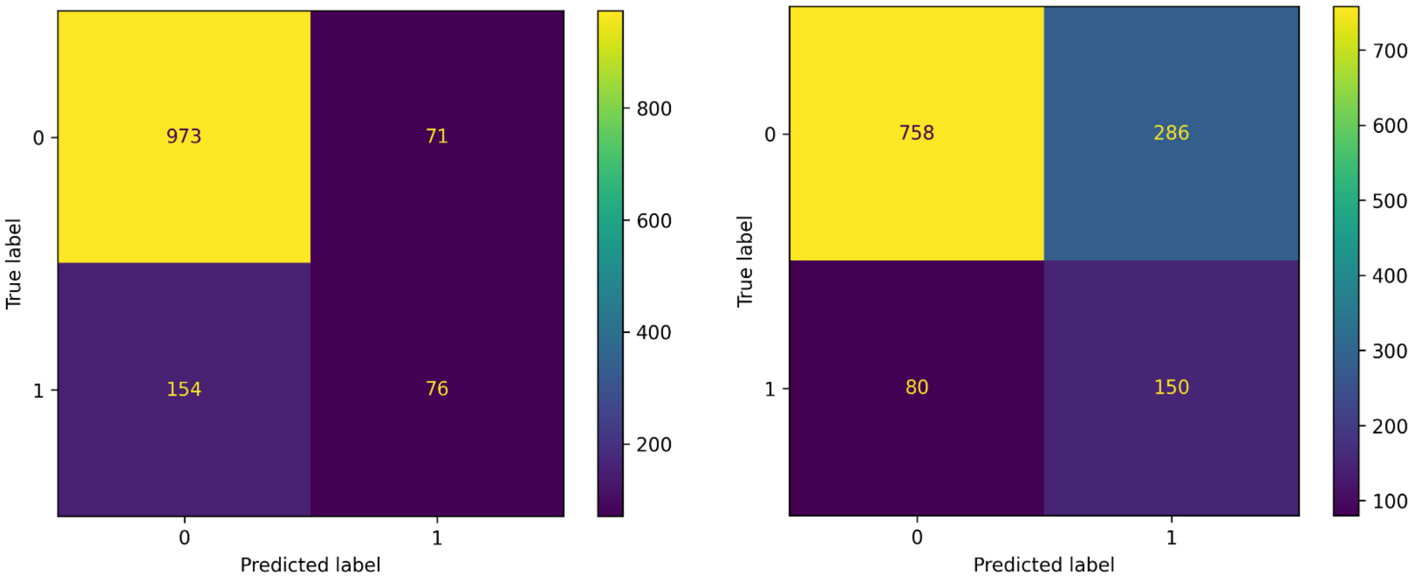

Figure 15 *– RQ2 (Enrolled) confusion matrices – Gradient Boosting (left), Logistic Regression (right)*

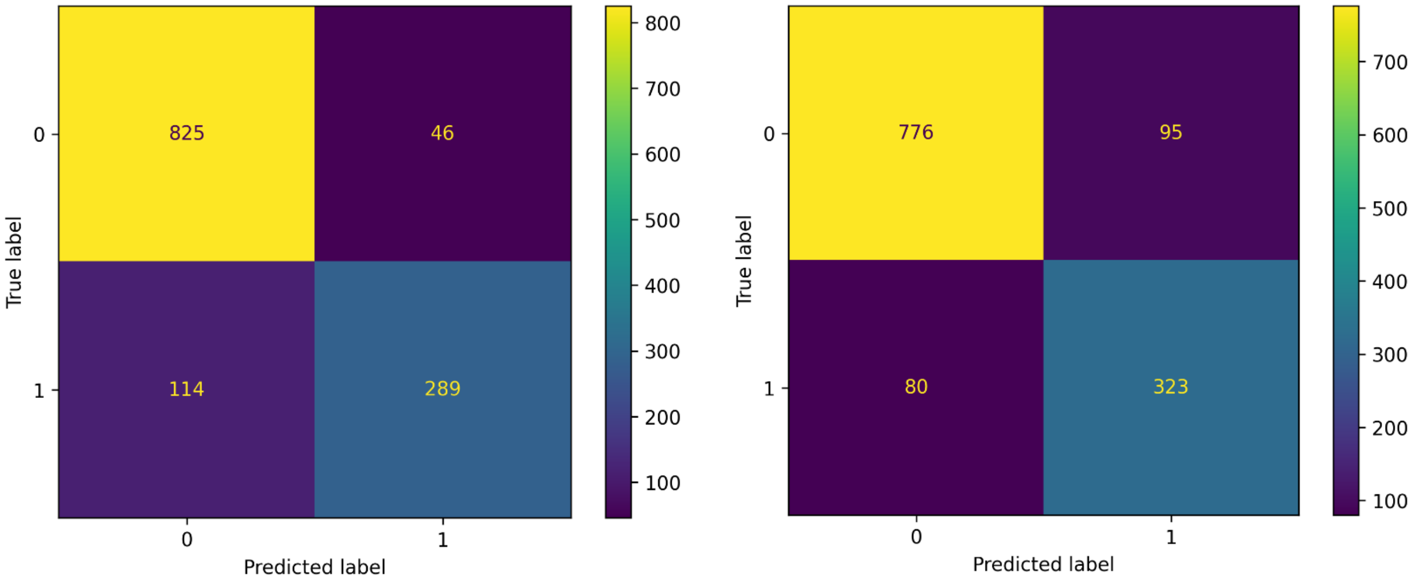

Figure 16 *– RQ2 (Dropout) confusion matrices – Gradient Boosting (left), Logistic Regression (right)*

During hyperparameter tuning the models showed signs of overfitting. Allowing a model to train with unlimited depth produced high training scores but much lower testing scores. More work is required in this area to understand overfitting and find a balance.

Semester Grades, Units Approved, and Evaluated consistently rank high as important features for prediction as does Tuition fees being up to date.

The minority Enrolled class remains difficult to predict. With more data from other universities perhaps the model could be improved, but these results suggest the dataset is lacking key features that would be required to identify students of this class.

Throughout this study I have learnt how to check data quality, engineer features and develop classification models splitting the data for training, validation and evaluation purposes. I would recommend the solution as the RQ1 classification model improved on the Portuguese study and the concept of modelling each class separately for RQ2 has worked well. With GB and LR each performing better for different target classes, a multi-model approach to prediction is recommended to give an overall probability of a student failing into one class or another.

7 Discussion and Conclusion

This study provides evidence that the use of first-year academic performance significantly enhances the predictive accuracy of machine learning models when forecasting student dropout and academic success. Although the minority Enrolled class is difficult to predict, using a multi-model approach improves the identification of students at-risk of failing to graduate on time. As expected, with the imbalanced nature of the target classes, a resampling method was required and this study confirms the (Martins et al., 2021) findings that SMOTE gives the best results.

The multicollinearity of the 1st and 2nd semester data caused issues during this study, and I did not have the experience to know whether including two features with >70% correlation into the Machine Learning stage would improve performance or corrupt the model. Summary features were created to aggregate the 1st and 2nd semesters, with the aim of not losing vital information, however even that caused issues; the average grade for both semesters became too correlated with the total units approved. The final selection needing to be two 1st semester, one 2nd semester and one aggregate value.

The only difference between the Baseline models and full RQ1 models was that the Baseline excluded the four first-year academic features, and yet the Baseline performed poorly. Those features consistently ranked high in feature importance, as seen in the accompanying notebook, with Units Approved always ranking first.

It was interesting to discover during the investigation of RQ2 that no single classification model performed best and that using Logistic Regression alongside Gradient Boosting could provide the best predictive outcome. With LR being better at detecting true positives at the cost of higher false positives it will depend on the priorities of the user as to which model they prefer. Perhaps falsely giving extra help to a student that would have graduated anyway is preferred compared to missing a student that is about to dropout.

Given more time and computing power, this study could have investigated a wider range of classification models and included more hyperparameter tuning. The boosting models seemed particularly computationally expensive, but gave the best F1 scores, so would be worth further investigation. An area for future research would be to develop a stacking or voting system where multiple models contribute to the overall prediction (So, 2020).

A future direction for this study would be to acquire data from a wider range of universities with more focus on the academic performance of the students. Ideally it should include data from all three years of study. In the case of drop-out or delayed graduation, a reason should be captured; a coarse-grain categorisation such as 'Health'/'Financial'/'Family'/'Location' would allow better modelling of the problem. Where universities have intervened to help a student risking dropout, some categorical information about that intervention could be provided.

Tuition Fees being up-to date ranked highly in feature importance, so a better understanding of student's financial situation could improve prediction.

In conclusion, this study contributes to ongoing research on predicting student dropout and academic success by emphasising the importance of monitoring academic performance. However, without broader and more detailed data, these models may be approaching their predictive limits. Their potential highlights the need for universities to collect and share more data with the data science community to further enhance predictive accuracy.

8 References

M.V.Martins, D. Tolledo, J. Machado, L. M.T. Baptista, V.Realinho (2021). Early prediction of student's performance in higher education: a case study: Trends and Applications in Information Systems and Technologies, vol.1, in Advances in Intelligent Systems and Computing series. Springer. DOI: 10.1007/978-3-030-72657-7_16

Delogu, M., Raffaele Lagravinese, Paolini, D. and Giuliano Resce (2023). Predicting dropout from higher education: Evidence from Italy. *Economic Modelling*, 130, pp.106583–106583. doi:https://doi.org/10.1016/j.econmod.2023.106583.

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1). doi:https://doi.org/10.1186/s40561-022-00192-z.

Al-Alawi, L., Al Shaqsi, J., Tarhini, A. and Al-Busaidi, A.S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*. doi:https://doi.org/10.1007/s10639-023-11700-0.

So, A, V., JT, Thas, JR, Worsley, A, & Asare, S (2020), The Data Science Workshop : A New, Interactive Approach to Learning Data Science, Packt Publishing, Limited, Birmingham.

9 Use of AI

I attest that this project made use of AI in the following ways:

Usage	**Tool Used**	**How the output was edited**	**Conversation Link**
Search of relevant literature	Elicit	Output was not copied, only read to find appropriate literature for review	Not available
Learning tool – to clarify the meaning of Data Science concepts – used to supplement Google searching	Chat GPT	Not copied just read	Not available
Code assistance – Copilot helped with the syntax of the python library functions required to visualise the data	GitHub Copilot in VS Code	Small snippets of code were placed into Visual Studio which I then heavily edited to achieve the necessary display. Saved the time searching for relevant examples on the internet.	Not available