U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Explore XAI techniques on the Student's Dropout Dataset

Group DS2-G3 – Report

Alexandre Sousa: 202206427
Magda Costa: 202207036
Rafael Pacheco: 202206258

Advanced Topics on Machine Learning (M.IA009) 2025/2026
Faculdade de Engenharia da Universidade do Porto
Faculdade de Ciências da Universidade do Porto

Table of Contents

CONTENTS

I Introduction 3

II Dataset 3

III Pre-Modelling Techniques 3
- III-A Exploratory Analysis and Visualization . . . 3
- III-B PCA and Kernel-PCA . . . 3
- III-C t-SNE . . . 4
- III-D Prototypes and Criticisms with MMD . . . 4

IV In-Modelling Techniques - Decision Tree 4
- IV-A No Limited Depth: With and Without Enrolled . . . 4
- IV-B Limited Depth: With and Without Enrolled . . . 5
- IV-C Comparing Full and Limited Trees Without Enrolled . . . 5

V Post-Modelling Techniques: XGBoost 5
- V-A Simplification-Based: Decision Tree . . . 5
- V-B Feature-Based Techniques . . . 5
  - V-B1 Shap Values (TreeShap) . . . 5
  - V-B2 Permutation Feature Importance (PFI) . . . 6
  - V-B3 Local Explanations - LIME . . . 6
  - V-B4 Local Explanations - SHAP . . . 6
- V-C Example-Based: Anchors . . . 7

VI Post-Modelling Techniques: Multilayer Perceptron 7
- VI-A Simplification Based: Decision Tree . . . 7
- VI-B Feature Based Techniques . . . 7
  - VI-B1 SHAP Values (KernelShap) . . . 7
  - VI-B2 Permutation Feature Importance (PFI) . . . 7
  - VI-B3 Local Explanations - LIME . . . 8
  - VI-B4 Local Explanations - SHAP . . . 8
- VI-C Example Based: Anchors . . . 8

VII Comparison of Post-Hoc Methods 8
- VII-A Faithfulness: Rank Correlation . . . 8
  - VII-A1 XGBoost . . . 8
  - VII-A2 Multilayer Perceptron . . . 9
  - VII-A3 Comparison . . . 9

VIII Conclusion 9

References 9

Appendix A: Distribution of the Different Features 10

Appendix B: Correlation between Features and Target 13

Appendix C: PCA and Kernel PCA 15

Appendix D: t-SNE and t-SNE using Kernel PCA 16

Appendix E: Prototypes and Criticisms with MMD 17

Appendix F: Full Depth Decision Tree: With VS Without Enrolled 18

Appendix G: Limited Depth Decision Tree: With VS Without Enrolled 18

Appendix H: Limited VS Unlimited DT: Without Enrolled 19

Appendix I: XGBoost Classification Report 20

Appendix J: XGBoost Surrogate Tree 20

Appendix K: XGBoost - Feature Importance - PFI 21

Appendix L: XGBoost - Local Explanations 22

Appendix M: MLP - Classification Report 24

Appendix N: MLP - Surrogate Tree 24

Appendix O: MLP - Global Assessment 25
O-A PFI . . . 25
O-B SHAP (KernelShap) . . . 25

Appendix P: MLP - Local Explanations 26
P-A LIME . . . 26
P-B SHAP . . . 27

Appendix Q: Anchors 28

I. Introduction

Student dropout in higher education is a persistent and costly challenge. When students interrupt their studies before completion, they incur personal losses in time, effort and financial investment. At a broader scale, high dropout rates undermine the role of higher education in promoting social mobility and economic development.

For there reasons, institutions are increasingly interested in data-driven tools that can help identify students at risk of dropout early enough to enable timely and targeted interventions.

The availability of detailed administrative and academic records has made it possible to frame dropout prediction as a supervised learning problem. In this work, we use the *Predict Students Dropout and Academic Success* dataset, which provides a rich set of student-level attributes.

In this work, we're going to address some explainability issues in black box models and apply different techniques of XAI to explain the dataset, the models trained and the problem it self.

II. Dataset

The dataset was obtained from *SATDAP - Capacitação da Administração Pública*, an institution that gathered a lot of data from various education institutions.

Each record corresponds to a single student and includes demographic information, socio-economic indicator, admission characteristics, prior schooling, academic performance and financial status (e.g. debtor status, tuition fees up to date, scholarship holder). The problem presented in this dataset is framed as a classification task divided into three categories: "Graduate", "Dropout" and "Enrolled".

From a modeling perspetive, this dataset is challenging for several reasons. First, the feature space is high-dimensional and heterogenous, combining numerical, ordinal and nominal attributes, many of which are correlated and non-linear, because they were labeled encoded previously. Second the **class distribution is imbalanced**, reflecting the fact that dropout, enrolled and graduation do not occur with equal frequency.

Table I: Class distribution of the target variable.

	1 (Graduate)	0 (Dropout)	2 (Enrolled)
Proportion (%)	49.93	32.12	17.95

III. Pre-Modelling Techniques

Pre-modelling techniques are essential to ensure transparency, fairness, and reliability from the very beginning of the AI development process. These techniques **help identify and mitigate bias in data**, **improve feature selection**, and **enhance model interpretability** before training. By integrating XAI early, projects can have more trustworthy, ethical, and effective AI systems.

A. Exploratory Analysis and Visualization

We began by **examining the marginal distribution and type of each feature**, categorizing them into different types. There are numerical variables, which are measured on an interval or ratio scale, such as: *"admission grade"*, *"age at enrollment"*, *"number of curricular units approved"*, etc. We also identified categorical variables that were label-encoded (e.g., *"course"*, *"application mode"*, *"previous qualification"*, or *"occupation"*) along with non-linear and binary variables (e.g., *"debtor status"*, *"tuition fees up to date"*, *"scholarship holder"*, *"international student"*, and *"educational special needs"*).

To identify skewed distributions, rare categories, and potential outliers, **we used univariate plots**: histograms for numerical features and bar charts for categorical ones, which can be viewed in Appendix A. This process helped us gain an initial understanding of the variables that might be informative regarding dropout risk.

We then **examined the correlation between the features and the target variable**. To accomplish this, we utilized different correlation methods since the dataset includes both numeric and categorical attributes. **For numeric-numeric pairs**, we calculated **the Pearson correlation coefficient**, focusing on linearly related variables. This resulted in a Pearson correlation matrix, which we visualized as a heatmap in Appendix B. **The heatmap highlighted a notable cluster of strongly correlated features**, such as the counts of curricular units enrolled, evaluated, and approved within the same semester, as well as semester grades and the corresponding number of approved units.

For categorical-categorical pairs, we computed **Cramér's V using contingency tables**. This analysis produced a second heatmap that illustrated the associations between categorical features, helping us identify variables that are strongly associated with the outcome, such as debtor status, tuition fee up to date, and scholarship holder status, as well as associations among the categorical variables themselves. Lastly, **for numeric-binary pairs, we again calculated the Pearson correlation**. All these graphs are available in Appendix B.

B. PCA and Kernel-PCA

The **exploratory analysis** revealed a **high degree of correlation and redundancy among many of the numeric features**, particularly those describing academic performance across semesters. These observations motivated the use of dimensionality reduction techniques in subsequent stages of the study. In particular, **we applied**

both linear PCA and Kernel PCA to obtain compact, low-dimensional representations of the feature space and later compared their behaviour, which can be seen in Appendix C. **The Kernel PCA projections revealed a more structured, non-linear separation among the classes compared with the linear PCA projections**. In **both cases**, however, **considerable overlap was observed between the Enrolled and Graduate classes**. We concluded that Kernel PCA better captured non-linear relationships, yielding a smaller number of components that explain more than 90% of the variance, and improves the global separability of the classes, producing a more coherent embedding.

Based on these findings, **we decided that the Enrolled class**, which is extremely overlapped with the Graduate class, **was not necessary for our analysis**. Keeping this class would mainly **introduce additional noise into the data**, **leading to poorer model performance** and making the problem more difficult to analyse and to explain.

C. t-SNE

To better analyse the data we also tried to visualize it with t-distributed Stochastic Neighboor Embedding (t-SNE). t-SNE is a nonlinear dimensionality-reduction method that projects high-dimensional data onto 2 or 3 dimensions while preserving local neighborhood structure. Unlike PCA, which preserves global variance linearly, t-SNE focuses on preserving local structure, meaning that points that are close together in the original space remain close in the visualization.

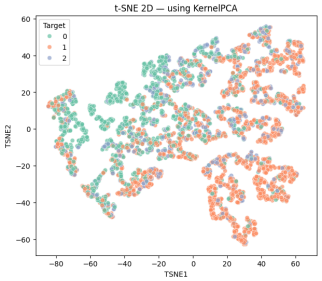


Figure 1: t-distributed Stochastic Neighborhood Estimation with Kernel PCA with Enrolled

t-SNE on the original data produces multiple compact local clusters. **Class 'Dropout' forms several dense groups**, while **class 'Graduated' is more widely dispersed** across space. When **t-SNE is applied to the two Kernel PCA components**, the resulting embedding becomes **more organised and more consistent** with the global non-linear structure. All graphs related to t-SNE and t-SNE using Kernel PCA can be seen in Appendix D.

D. Prototypes and Criticisms with MMD

Maximum Mean Discrepancy (MMD) gives us a kernel-based way to summarise the dataset through representative examples (prototypes) and to highlight underrepresented, atypical points (criticisms).

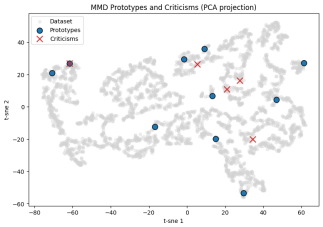


Figure 2: Visualization of Prototypes and Criticisms with Maximum Mean Discrepancy without Enrolled

From this graph we can **conclude that there's not a single "typical student"**, the data contains **several distinct profiles**, and each prototype acts as an archetype for one of these regions. The red crosses (criticisms) fall in regions that are not well summarised by the prototypes, or that lie near transitions between clusters. Three of them correspond to graduates with moderate academic performance in the second semester. Two criticisms are strong outliers in terms of academic trajectory: both have second-semester grade of 0, which represents a case where they've dropout of school around the first semester.

In Appendix E, it is possible to observe the prototypes and criticisms for the case where Enrolled was present in the dataset.

IV. In-Modelling Techniques - Decision Tree

Decision trees provide a transparent, rule-based decision process. Their interpretability depends heavily on model complexity: deep trees are hard to read, while shallow trees are easier to explain but may miss subtle patterns.

A. No Limited Depth: With and Without Enrolled

Our data analysis indicated that the 'Enrolled' attribute adds noise to the dataset and negatively impacts the classification capabilities of our models. To confirm this hypothesis, we trained two decision trees: one that included the 'Enrolled' attribute and another that excluded it. Both trees were trained without depth limits, and we obtained the following results:

Table II: Full-depth Decision Tree: With VS Without Enrolled

Metric	Accuracy	F1	Precision	Recall
With Enrolled	0.677	0.616	0.619	0.617
Without Enrolled	0.860	0.851	0.852	0.849

Performance shows significant improvement in the binary case, suggesting that 'Enrolled' is less separable and introduces overlap between outcomes. This reaffirms our initial expectations and strengthens the rationale for their exclusion. The confusion matrices for these trees are in Appendix F.

B. Limited Depth: With and Without Enrolled

To create a simpler, easier-to-understand tree, we limited it to a depth of 3. Out of curiosity, we decided to compare the results of using or excluding Enrolled again, obtaining the following results, with more information available in Appendix G.

Table III: Limited-depth Decision Tree: With VS Without Enrolled

Metric	Accuracy	F1	Precision	Recall
With Enrolled	0.715	0.616	0.663	0.609
Without Enrolled	0.891	0.881	0.900	0.870

Again, without enrolled, the results are far superior. However, limiting the tree's depth yielded slightly better results than using the complete tree, suggesting that not limiting the tree may have led to more particular cases rather than encompassing more.

C. Comparing Full and Limited Trees Without Enrolled

A very deep decision tree tends to be less suitable from an XAI perspective because its complexity makes human interpretation difficult.

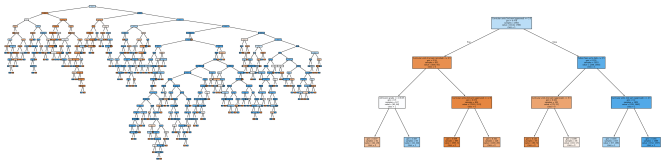


Figure 3: Full VS Limited Decision Tree Depth Without Enrolled

In a **full-depth decision tree**, the resulting model **tends to be large and challenging to interpret as a cohesive set of rules**. This presents a significant limitation of unconstrained decision trees as explanatory models, especially when many features are involved. In contrast, **restricting the depth of the tree simplifies its structure, resulting in fewer decision pathways**. This enhancement **improves interpretability** while still capturing the most relevant relationships within the data. In this simplified tree, the primary splits are associated with academic progression and financial or administrative status. This aligns with the expectation that consistent academic success and timely fee payments are strong indicators of graduation.

V. Post-Modelling Techniques: XGBoost

XGBoost was chosen as a black-box model because it typically excels with structured or tabular datasets and works well with various explanation techniques.

The classification report, which can be consulted in Appendix I, indicates balanced performance across both classes, demonstrating that the model does not simply favor the majority class. The high performance on each metric, suggests that this model is appropriate as a reference black box for future interpretability analysis.

A. Simplification-Based: Decision Tree

A simple Decision Tree was trained to imitate the XGBoost model by using its predictions as pseudo-labels. This approach offers a global, human-readable approximation of the complex decision-making process of the black-box model. The effectiveness of the surrogate model was assessed through two metrics: fidelity, which measures the agreement between the Decision Tree's predictions and those of the XGBoost model, and accuracy, which compares the predictions against the actual labels.

Table IV: Metrics of the Surrogate Tree for XGBoost

Metric	Value
Fidelity (tree VS XGBoost)	0.9339
Accuracy (surrogate VS true labels)	0.8912
Accuracy (XGBoost VS true labels)	0.8994

The surrogate tree closely aligns with XGBoost decisions while maintaining most of its predictive performance. This indicates that the surrogate can serve as a transparent summary of the black box. By analyzing the surrogate tree, which can be found in Appendix J, we conclude that **second-semester academic progress is the primary factor in predicting student dropout rates and that very low numbers of approved units strongly indicate a risk of dropout**. When students perform better academically, **the model takes into account** additional factors, such as **administrative and financial signals** (for example, whether tuition fees are up to date), and in some cases, considers grades or admission scores. As a simplification-based method, **the surrogate tree is easy to understand and communicate**. However, it **compresses the interactions of the ensemble** and **may provide misleading insights in areas where the decision boundaries are more complex**.

B. Feature-Based Techniques

1) Shap Values (TreeShap): TreeSHAP is used to obtain global feature attributions for the fitted XGBoost model. Unlike simple ranking methods, SHAP provides both a global importance ordering and a direction of effect.

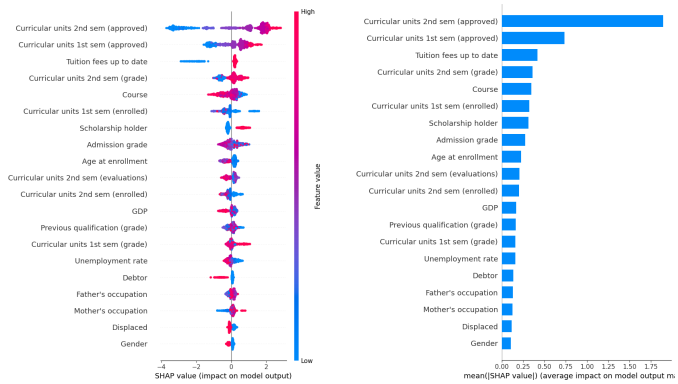


Figure 4: SHAP Values for Global Explanations

The analysis using SHAP reveals that **the model is primarily influenced by curriculum progression**. A higher number of approved units in both the first and second semesters is associated with a greater likelihood of graduation, whereas fewer approved units increases the risk of dropout. Additionally, tuition status has a clear impact, indicating a consistent financial influence on the predictions.

Lower-ranked variables have a smaller average contribution, suggesting that **the model does not heavily rely on demographic or socio-economic factors**.

Regarding interpretability, **SHAP is particularly effective**, as TreeSHAP is designed for tree ensemble models. However, its complexity can be a drawback as the plots require careful explanation, and attribution values can be less intuitive for non-technical audiences.

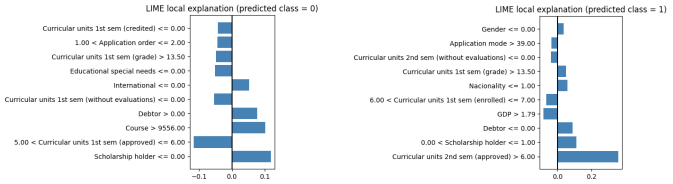
2) Permutation Feature Importance (PFI): PFI quantifies how much model performance drops when a single feature is randomly shuffled. Due to space limitations, the graph relating to this technique is available in Appendix K.

PFI strongly corroborates the SHAP results. Shuffling **Curricular units 2nd sem (approved) yields the largest drop in ROC–AUC by a wide margin**, confirming it as **the key feature**. Other top contributors include Curricular units 1st sem (approved) and Tuition fees up to date, with a long tail of smaller effects.

This pattern suggests **the model relies primarily on academic progression and administrative/financial stability**, with background variables playing a comparatively limited role.

PFI is easy to explain and model-agnostic, making it useful as a sanity check. However, it **does not provide directionality**, and it can **underestimate importance when predictors are correlated**.

3) Local Explanations - LIME: LIME provides local explanations by fitting a simple surrogate model around a specific instance. Two correctly predicted test students were selected (one from each class) to compare local drivers of dropout vs graduation:



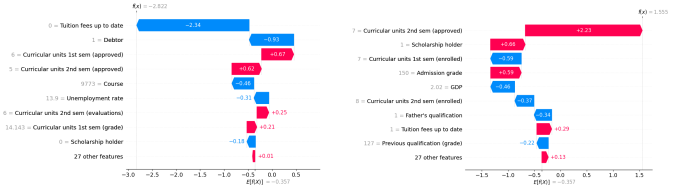
(a) Local LIME — classe 0 (b) Local LIME — classe 1

Figure 5: LIME for Local Explanations

Example predicted as class 0 (dropout): Even though this student shows positive academic signals (approvals and higher grades), the local explanation indicates that financial/administrative vulnerability (debtor, no scholarship) plus course-related context dominate, leading the model to a dropout prediction.

Example predicted as class 1 (graduate): strong positive contribution comes from high second-semester approvals, often supported by stability indicators, outweighing weaker negative factors.

4) Local Explanations - SHAP: Local SHAP explanations provide additive, instance-specific attributions consistent with the global SHAP framework. This section explains in detail the same two students selected in LIME. Two plots were created: the waterfall plots, which quantify the largest feature contributions, and the force plots, which provide an intuitive 'push/pull' view from the base value, shown in Appendix L.



(a) Local SHAP — classe 0 (b) Local SHAP — classe 1

Figure 6: SHAP for Local Explanations

Example predicted as class 0 (dropout): the prediction is pushed strongly towards class 0. Being *tuition fees up to date = 0* and *Dector = 1*, the most influential negative contributors. Positive academic signals push towards class 1 but are insufficient to overcome the strong financial risk indicators. For this student, the model treats financial instability as decisive.

Example predicted as class 1 (graduate): The prediction is pushed towards class 1, primarily due to: *Curricular units 2nd sem (approved)* and *Scholarship holder = 1*. Some factors push mildly towards class 0, but they are outweighed by strong academic progression and stability signals.

Across the two XGBoost local examples, **LIME and SHAP tell a consistent story**, with differences mainly in presentation. For the Dropout case, both methods show financial/administrative vulnerability (e.g., fees not up

to date/debtor) as the strongest push towards Dropout, outweighing weaker positive academic signals. For the Graduate case, both highlight academic progression, especially 2nd-semester approvals, as the dominant reason for a Graduate prediction. Overall, **these local explanations align with the other XAI techniques applied**, which also prioritise approvals and fee status as key drivers.

C. Example-Based: Anchors

The Anchors method can have an important role in this work because it provides **local, rule-based explanations** that are both human-readable. Unlike feature-attribution methods such as SHAP,LIME or PFI, which distribute importance scores over features but do not specify when a prediction is stable, Anchors are if-then rules together with two key quantities: precision and coverage. This makes the explanations more actionable. In Appendix X, we show the anchors of two students of the two different classes.

Then, we extended the analysis by generating 200 anchor rules for each class to obtain a more global picture of the model's decision logic. Across these rules there is a clear and consistent pattern: anchors for the Dropout class are dominated by conditions involving very low numbers of approved curricular units and low grades in the second semester, often combined with debtor status or unfavourable administrative conditions. Conversely, anchors for the Graduate class typically require a high number of approved units and high grades, frequently together with being up to date with tuition fees or holding a scholarship. Most of these rules achieve very high precision, confirming that, although each rule only covers a subset of students.

				anchor
class	rank	mean_precision	mean_coverage	
0	1	0.964193	0.283982	Curricular units 2nd sem (approved) <= 2.00
	2	0.991475	0.180400	Curricular units 2nd sem (approved) <= 2.00 AND Course > 9085.00
	3	0.952202	0.180667	Curricular units 2nd sem (approved) > 5.00 AND Curricular units 2nd sem (grade) > 13.50
	4	0.961143	0.096900	Curricular units 2nd sem (approved) > 6.00 AND Curricular units 2nd sem (grade) > 13.50
	5	0.992647	0.113300	Curricular units 2nd sem (approved) <= 2.00 AND Age at enrollment > 25.00
1	1	0.959110	0.178260	Curricular units 2nd sem (approved) > 5.00 AND Scholarship holder > 0.00
	2	0.963341	0.075319	Curricular units 2nd sem (approved) > 6.00 AND Scholarship holder > 0.00
	3	0.957454	0.182131	Curricular units 2nd sem (approved) > 5.00 AND Curricular units 2nd sem (grade) > 13.50
	4	0.952936	0.173525	Curricular units 2nd sem (approved) > 5.00 AND Curricular units 1st sem (grade) > 13.50
	5	0.964139	0.098925	Curricular units 2nd sem (approved) > 6.00 AND Curricular units 2nd sem (grade) > 13.50

Figure 7: Top 5 Anchors for each class

VI. Post-Modelling Techniques: Multilayer Perceptron

We decided to also apply all of the previous XAI techniques to a Multilayer Perceptron (MLP) (Appendix M), a well-known version of neural networks, and thus commonly used for types of classification tasks such as our own. Despite being more simplified than a regular NN, MLP's remain blackbox models. For this reason, it will undergo the same treatment as the XGBoost model.

A. Simplification Based: Decision Tree

A simple Decision Tree (Appendix N) was trained to imitate the MLP model by using its predictions as pseudo- labels.

The DT's predictions were compared against the MLP's predictions on the test set to determine the fidelity score, which returned at **approximately 0.904**, indicating that the DT successfully captured decision boundaries.

The Root Split uses *Curricular units approved in the 2nd semester*, implying it is the **most informative** feature globally. Other key decision paths indicate *Low 2nd-semester approvals* may lead to **High probability of non-graduation**, **High** *2nd-semester approvals* **+** *tuition fees up to date* are associated with **High probability of graduation**, and that **Intermediate approval levels** also affect overall decision process.

In essence, this surrogate model portrays that the MLP relies heavily on recent performance metrics, as well as financial consistency and intermediate academic background.

B. Feature Based Techniques

1) SHAP Values (KernelShap): Globally, the SHAP results (Appendix O-B) indicate that academic performance indicators yield a **strong direct proportion between high feature values and impact**, most notably in *approved* and *graded* curricular units (for both the 1st and 2nd semesters). Being a scholarship holder and the student's international status also share this property, although less intensely. On the other hand, there are also cases of features with **intense direct proportion of low feature values and high impact**, more specifically, credited curricular units (1st and 2nd semesters). The remaining features with discernible importance show **arbitrary and unidentifiable division of correlations**.

In comparison to XGBoost, the MLP prioritizes approval and grades above Tuition Fees.

These values indicate that **grades** and **approval** in curricular units are **highly important** for the model's decision making, but other factors like **nationality**, **scholarship** and **GDP** might still affect overall classification (with some degree of **identifiability of effect**).

2) Permutation Feature Importance (PFI): The PFI results (Appendix O-A) confirm all previous SHAP conclusions, by placing Curricular Unit approval and grade on the top of the importance hierarchy. Interestingly, PFI also associates importance with some sensitive features, such as *Debtor* and *Tuition fees up to date*, which could potentially be indicative of a certain level of bias in financial interest during the model's decision process.

When compared to XGBoost, the hierarchy is quite similar.

3) Local Explanations - LIME: All relevant results can be found in Appendix P-A.

Example predicted as class 0 (dropout): Absence of *debt* and *approval* (*curricular units in the 2nd semester*) pull the model towards Graduate. On the other hand, lack of *scholarship* for the specific *course* paired with *credited units for the 1st and 2nd semester* eventually overcome the previous information, ultimately classifying as Dropout. The *father's qualification* also contributes to this conclusion.

Example predicted as class 1 (graduate): multiple features contribute towards the Dropout conclusion, such as not being *international*, low *1st semester grades*, and *course*. However, the absence of *credited units* (*1st and 2nd semester*) and lack of *debt* with the *nationality* and *scholarship* eventually overcome the previous information, ultimately classifying this example as Graduate. Interestingly, the individual's *gender* still somewhat contributes to the Graduate conclusion.

4) Local Explanations - SHAP: All relevant results can be found in Appendix P-B

Example predicted as class 0 (dropout): the number of *approved curricular units in the 2nd semester*, paired with relatively high *admission grade*, pull the model towards the Graduate classification. However, the specific combination of low *credited units (both in 1st and 2nd semester)* and *evaluation* (*2nd semester*) with the *course* overcome the previous information, ultimately classifying this example as Dropout. Once again, *father's qualification* and *occupation* (as well as the individual's *gender*) still somewhat contribute to the Dropout conclusion.

Example predicted as class 1 (graduate): the *course* identifier strongly leads the model to classify as Dropout However, the specific combination of the student's *previous qualification* and high *admission grade*, with having a *scholarship* and a high number of *1st semester approved units*, ultimately guides the model to classify as Graduate. Similarly, *previous qualification*'s grade and no *debts* still contribute to the Graduate conclusion.

With this, we can draw the conclusion that the **global behavior adapts well to** specific, **local examples**, quite similar to that displayed with the LIME technique (despite placing **lower focus in the** *debtor* status of the student). Additionally, these examples show that **financial indicators** still play a role in classification, to some extent.

C. Example Based: Anchors

As with the XGBoost model, we also applied the Anchors method to the MLP. The same type of example-based comparisons was carried out in order to analyse the local decision rules learned by the neural network. The corresponding anchors for representative students of each class are reported in Appendix Q.

class	rank	anchor	n_instances	mean_precision	mean_coverage	mean_correct
0	1	Curricular units 2nd sem (approved) <= 2.00 AND Age at enrollment > 26.00	18	0.980677	0.106633	1.0
	2	Curricular units 2nd sem (approved) <= 2.00 AND Debtor > 0.00	16	0.984591	0.055419	1.0
	3	Curricular units 2nd sem (approved) <= 2.00 AND Course > 9556.00	11	0.971682	0.063936	1.0
	4	Curricular units 2nd sem (approved) <= 2.00 AND Application mode > 17.00	8	0.976292	0.139088	1.0
	5	Curricular units 2nd sem (approved) <= 2.00 AND Admission grade <= 118.00	7	0.986326	0.100386	1.0
1	1	Curricular units 2nd sem (approved) > 5.00 AND Curricular units 2nd sem (grade) > 13.50	35	0.958066	0.181031	1.0
	2	Curricular units 2nd sem (approved) > 6.00 AND Curricular units 2nd sem (grade) > 13.50	28	0.972382	0.097618	1.0
	3	Curricular units 2nd sem (approved) > 6.00 AND Curricular units 2nd sem (grade) > 12.33	15	0.955962	0.170080	1.0
	4	Curricular units 2nd sem (approved) > 5.00 AND Scholarship holder > 0.00 AND Curricular units 1st sem (approved) > 5.00	8	0.982347	0.163625	1.0
	5	Curricular units 2nd sem (approved) > 5.00 AND Curricular units 2nd sem (grade) > 12.33 AND Age at enrollment <= 19.00	5	0.988817	0.206460	1.0

Figure 8: Top 5 Anchors for each class - MLP

The MLP and XGBoost anchors tell essentially the same story: both models base their decisions mainly on second-semester academic performance, with few approved units anchoring dropout and many approved units with high grades anchoring graduation. The MLP anchors add slightly more nuanced conditions (age, admission grade, debtor status), but they reinforce rather than change the core patterns found with XGBoost.

VII. Comparison of Post-Hoc Methods

A. Faithfulness: Rank Correlation

A central requirement for post-hoc explanations is *faithfulness*: an explanation method should reflect the way the underlying model actually uses the features, rather than imposing an arbitrary or overly simplified view. To quantify this, we use a **rank-correlation analysis**: for each model we take a reference feature-importance ranking (given by the model itself or by permutation feature importance) and compare it with the global rankings produced by each XAI method. We use the *Spearman rank correlation* coefficient ρ, which measures the strength of a monotonic relationship between two rankings. Alongside ρ we report a p-value for the null hypothesis of $\rho = 0$ (no association): small p-values (e.g. $p < 0.01$) provide strong evidence that the observed agreement is not due to random variation.

1) XGBoost: To assess the global faithfulness of the explanation methods to the underlying XGBoost model, we computed the Spearman rank correlation between the feature-importance ranking given by the model itself (XGBoost gain) and the rankings produced by each XAI technique (LIME, TreeSHAP, permutation feature importance and surrogate tree). The results show that all four methods are positively and significantly correlated with the internal XGBoost ranking. LIME and TreeSHAP achieve the highest correlations ($\rho \approx 0.65$ and $\rho \approx 0.64$, respectively, both with $p \ll 0.01$), indicating a moderate–strong and statistically significant agreement with the model's notion of which features matter most. PFI also displays a positive but weaker correlation ($\rho \approx 0.50$), while the surrogate tree has the lowest agreement with the original ranking ($\rho \approx 0.44$), reflecting the fact that a single global tree provides only a coarse approximation of the boosted ensemble.

Importantly, the top-ranked features are very similar across all methods: *Curricular units 2nd sem (approved)* is consistently identified as the most influential variable, followed by first- and second-semester academic performance (units enrolled/approved and grades), tuition-fee status, debtor status, scholarship holder and course. This convergence suggests that the explanations are not only statistically faithful to the XGBoost ranking, but also robust across different families of XAI methods. Overall, LIME and TreeSHAP provide the most faithful global summaries of the XGBoost model, with PFI and the surrogate tree still broadly consistent but less tightly aligned with the reference ranking.

2) Multilayer Perceptron: We repeated the rank-correlation analysis for the Multilayer Perceptron (MLP) model. Since the MLP does not expose a built-in feature-importance measure analogous to XGBoost gain, we used permutation feature importance (PFI) computed on the trained network as the reference ranking. We then calculated the Spearman correlation between this PFI ranking and the global rankings induced by KernelSHAP, LIME and the surrogate tree fitted to the MLP predictions.

The correlations are again positive and statistically significant for all methods. LIME achieves the highest agreement with the PFI ranking, with $\rho = 0.72$ and $p \approx 7.7 \times 10^{-7}$, followed by KernelSHAP with $\rho = 0.63$ and $p \approx 3.5 \times 10^{-5}$. The surrogate tree shows a weaker but still positive correlation of $\rho = 0.49$ ($p \approx 2.4 \times 10^{-3}$), reflecting the limited capacity of a single global tree to approximate the behaviour of the MLP. As in the XGBoost case, there is a clear overlap in the most important variables: second-semester academic performance (approved units and grades) and tuition-fee status dominate the top positions, with scholarship holder, debtor status and first-semester performance also appearing prominently across methods. This confirms that, although the MLP is a black-box model, the different XAI methods provide globally coherent and reasonably faithful explanations of its decisions, highlighting the same key academic and financial factors that emerged in the XGBoost analysis.

3) Comparison: Taken together, the rank-correlation results show that the post-hoc methods are *consistently faithful* across both architectures. For XGBoost, TreeSHAP and LIME achieve moderate–strong agreement with the model's own importance ranking, while for the MLP, LIME and KernelSHAP reach slightly higher correlations with the PFI reference. In both cases the surrogate tree lags behind, as expected for a simple global surrogate, but still preserves a non-negligible amount of the original ranking structure. Most importantly, the same set of features emerges as dominant for both XGBoost and MLP, indicating that the explanations are capturing stable patterns in the data rather than artefacts of a particular model or XAI technique.

VIII. Conclusion

Exploratory analysis and correlation studies provided global transparency into the dataset, revealing strong redundancy among academic performance features, as well as some strong associations between certain socioeconomic variables and the target. Dimensionality reduction methods (PCA) helped identify the strong class overlap and how it would represent a challenge, even after producing globally structured embeddings. t-SNE was a great visual tool, and its effectiveness only improved after removing the Enrolled class. Prototypes and Criticisms were ideal for identifying atypical/abnormal instances. By combining all of these insights, we were able to reduce ambiguity, redundancy and noise in out data considerably.

In chapter IV, we learnt that more isn't always better: removing Enrolled proved to be the best possible choice. Not only was it amazing for result quality and efficiency, it also allowed us to get more meaningful and consistent visualizations and explanations. The restrained depth decision tree displayed good performance and great explainability through rule path visualizations.

The Decision Tree for surrogate worked better when applied to XGBoost, which is only natural as it is a tree based model. However, it still provided some valuable insights, which were further confirmed through other XAI techniques. At a global level, both models converge on academic progression as the dominant factor. The main difference here is the driving socieoconomic factor in XGBoost is usually tuition fees up to date, while MLP often identifies scholarship status as important. This distinction is more prominent in local explanations, where XGBoost shows that socioeconomic features are capable of overpowering academic success, while MLP prioritizes academic consistency and success indicators.

Overall, no single XAI method is sufficient on its own. Simplification-based methods display simplified versions of decision processes, feature-based methods expose global rule sets, and Local and example-based methods help explain individual decisions. The intra-model (and sometimes inter-model) Consistency is a strong indicator that both models rely on stable, interpretable triggers rather than completely abstract patterns, and also exposes some potentially sensitive bias in socioeconomic factors.

References

APPENDIX A
DISTRIBUTION OF THE DIFFERENT FEATURES

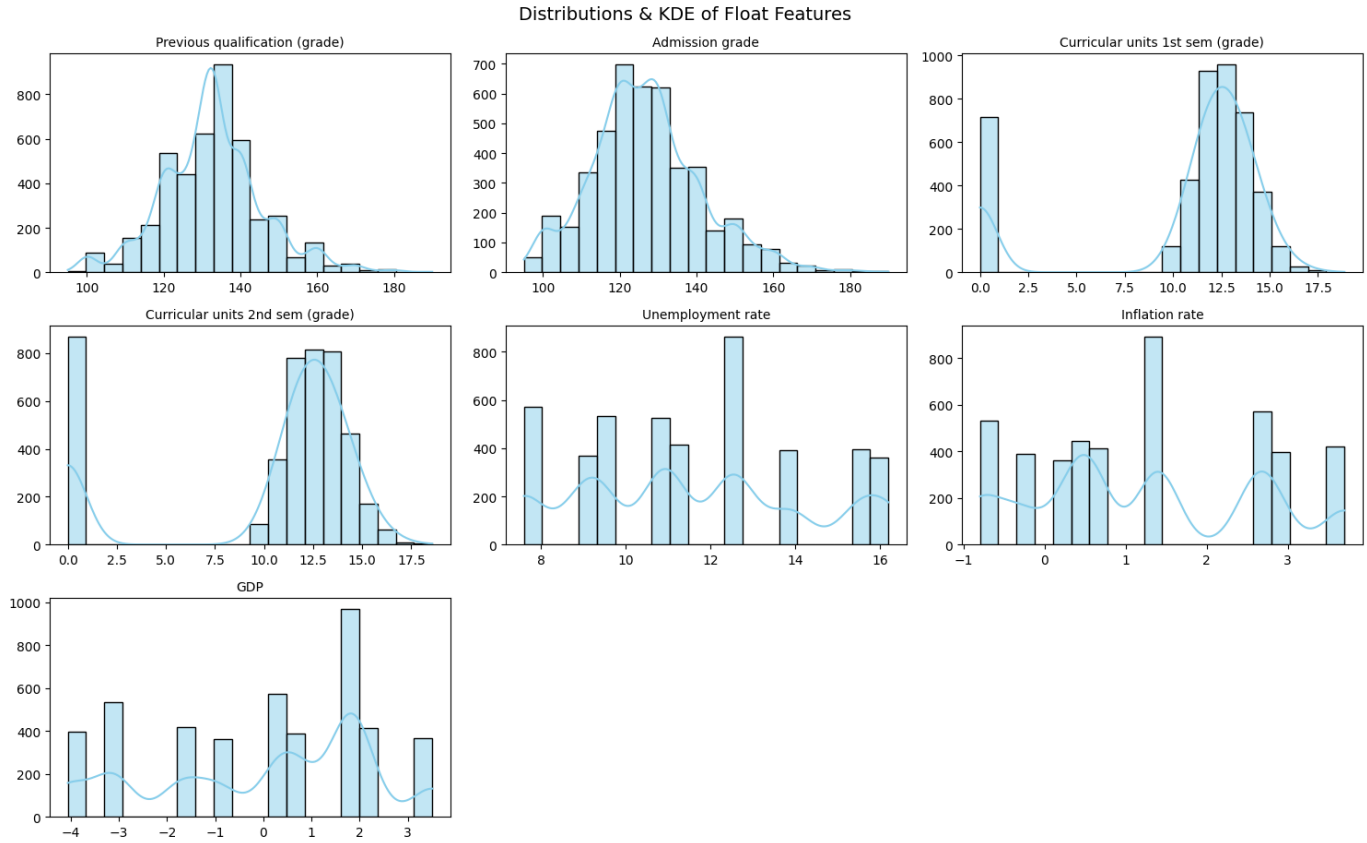


Figure 9: Distribution KDE of Float Features

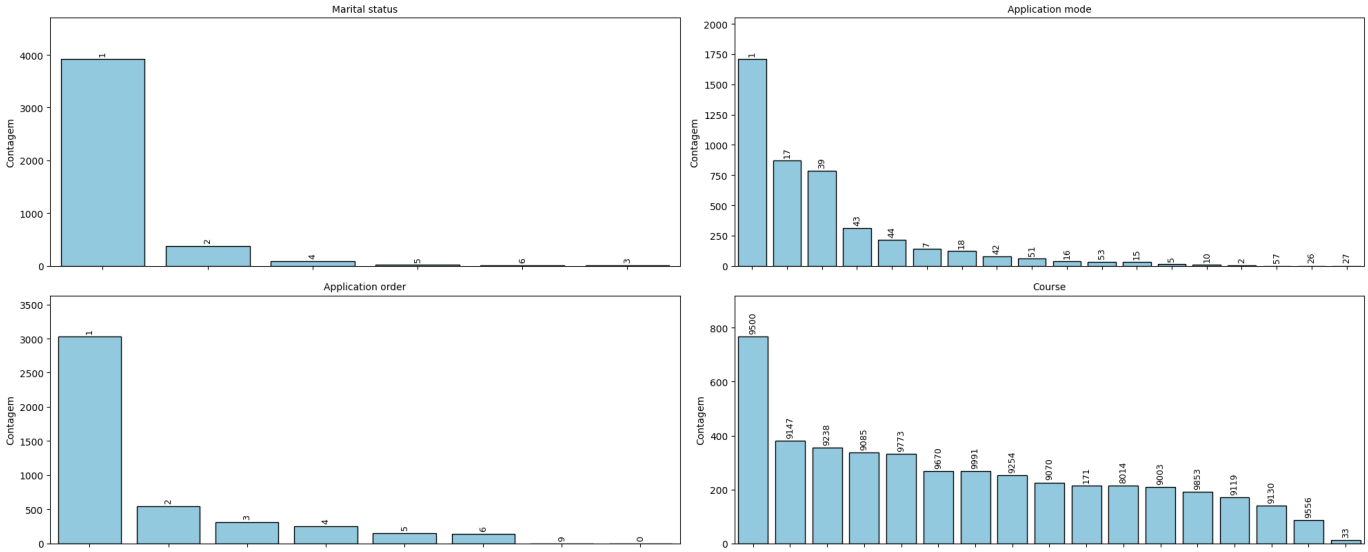


Figure 10: Bar Plots of Categorical Features: Part 1

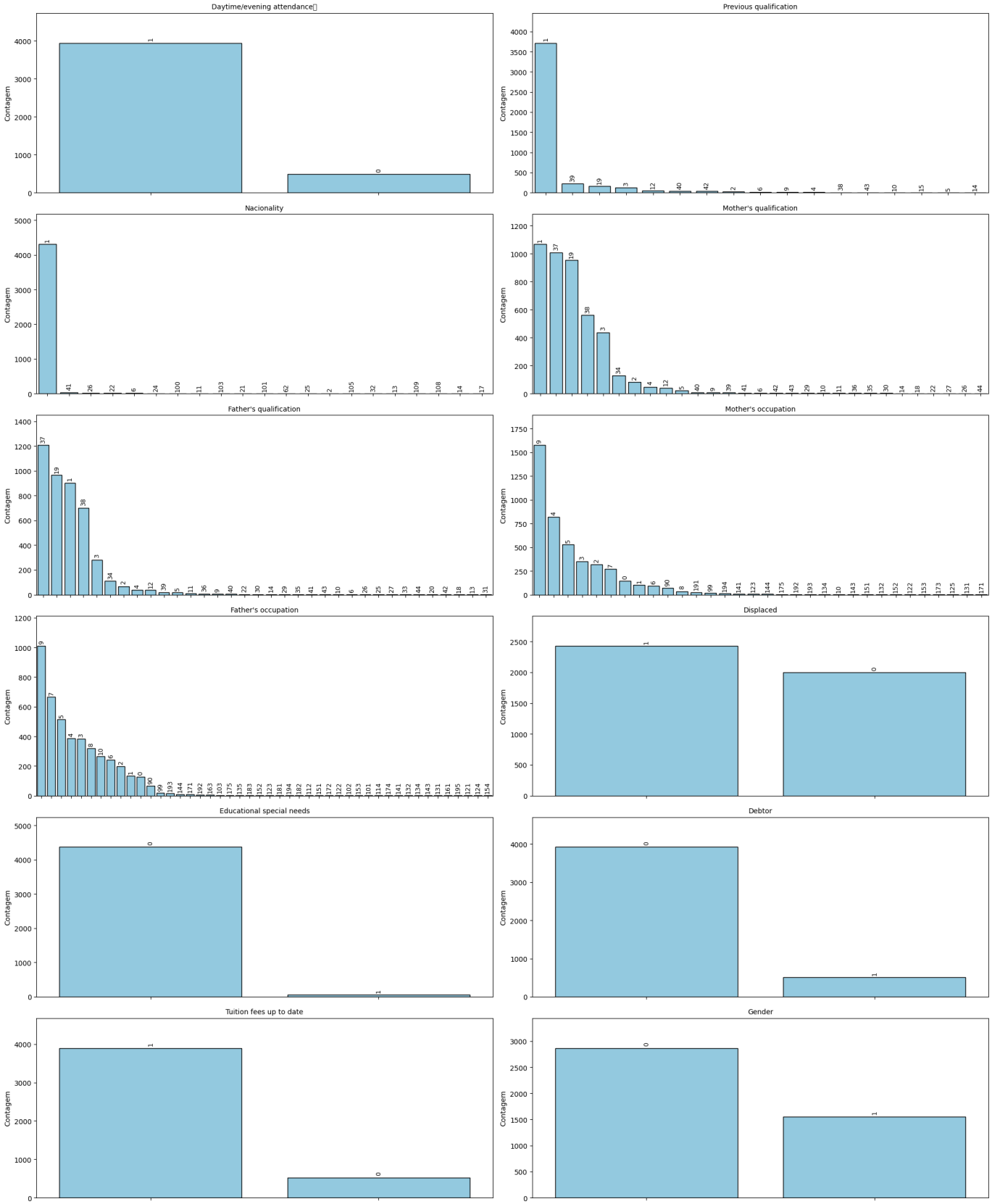


Figure 11: Bar Plots of Categorical Features: Part 2

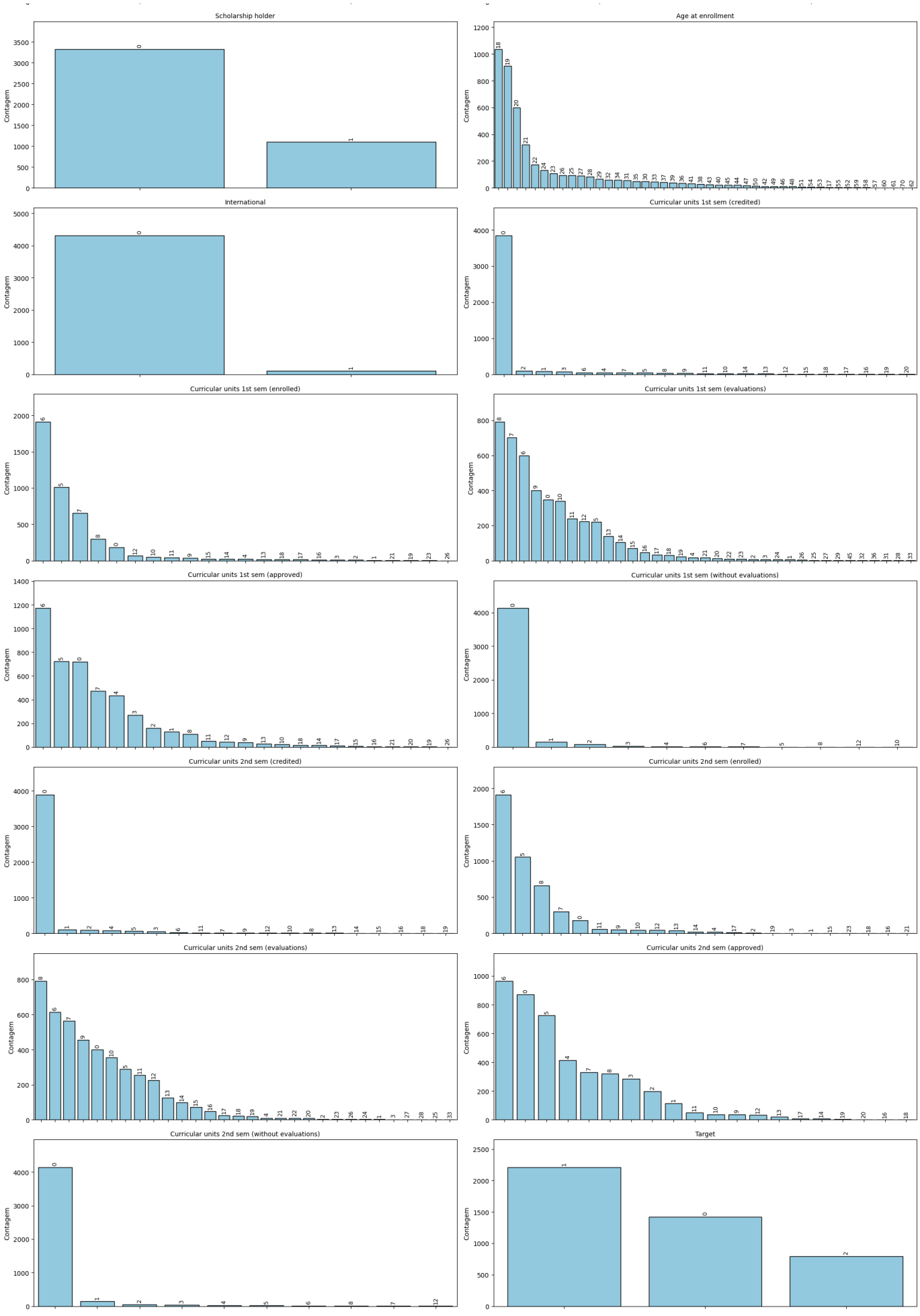


Figure 12: Bar Plots of Categorical Features: Part 3

APPENDIX B
CORRELATION BETWEEN FEATURES AND TARGET

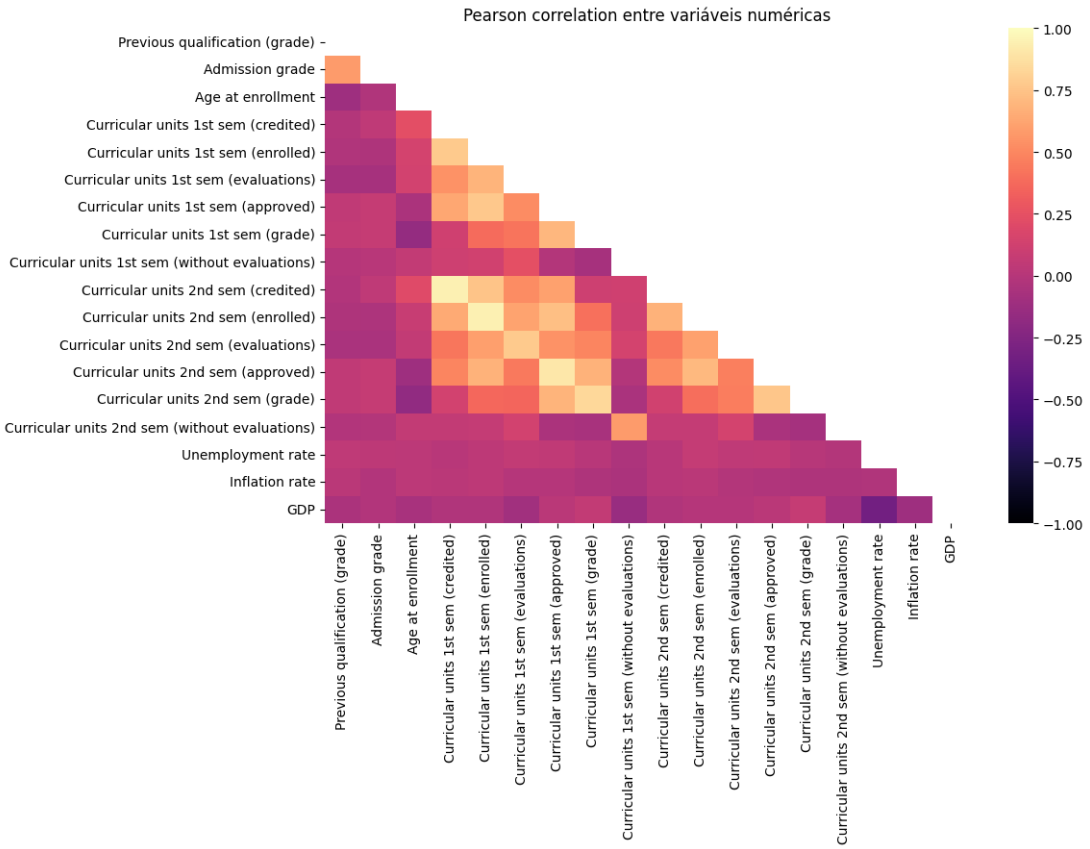


Figure 13: Pearson Correlation between Numerical Features

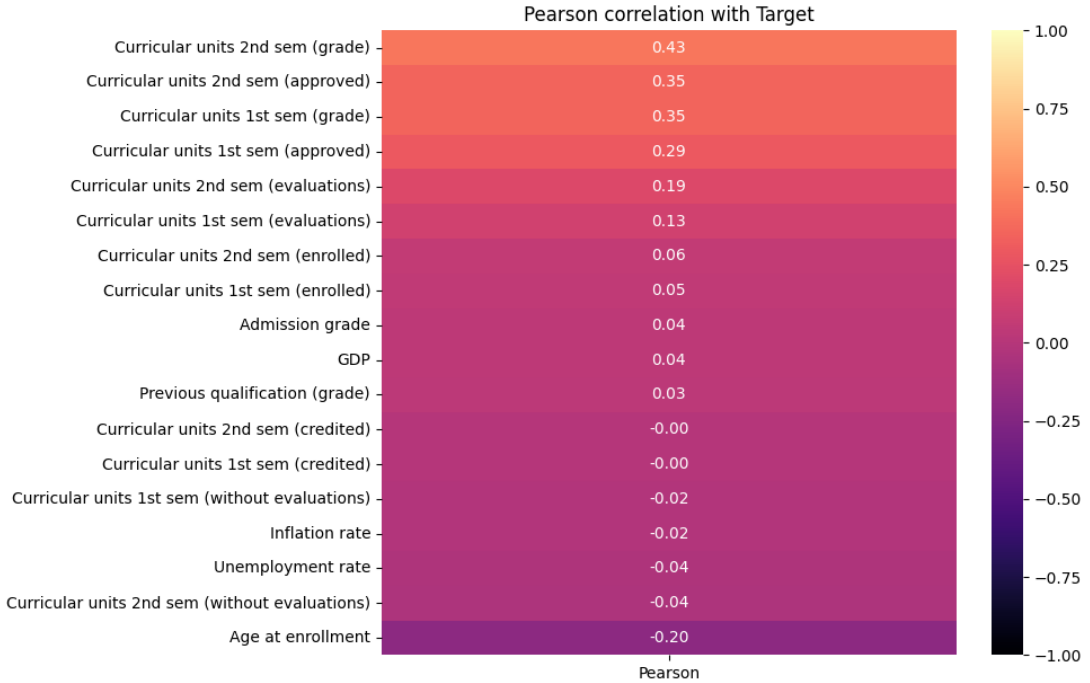


Figure 14: Pearson Correlation between Numerical Features and Target

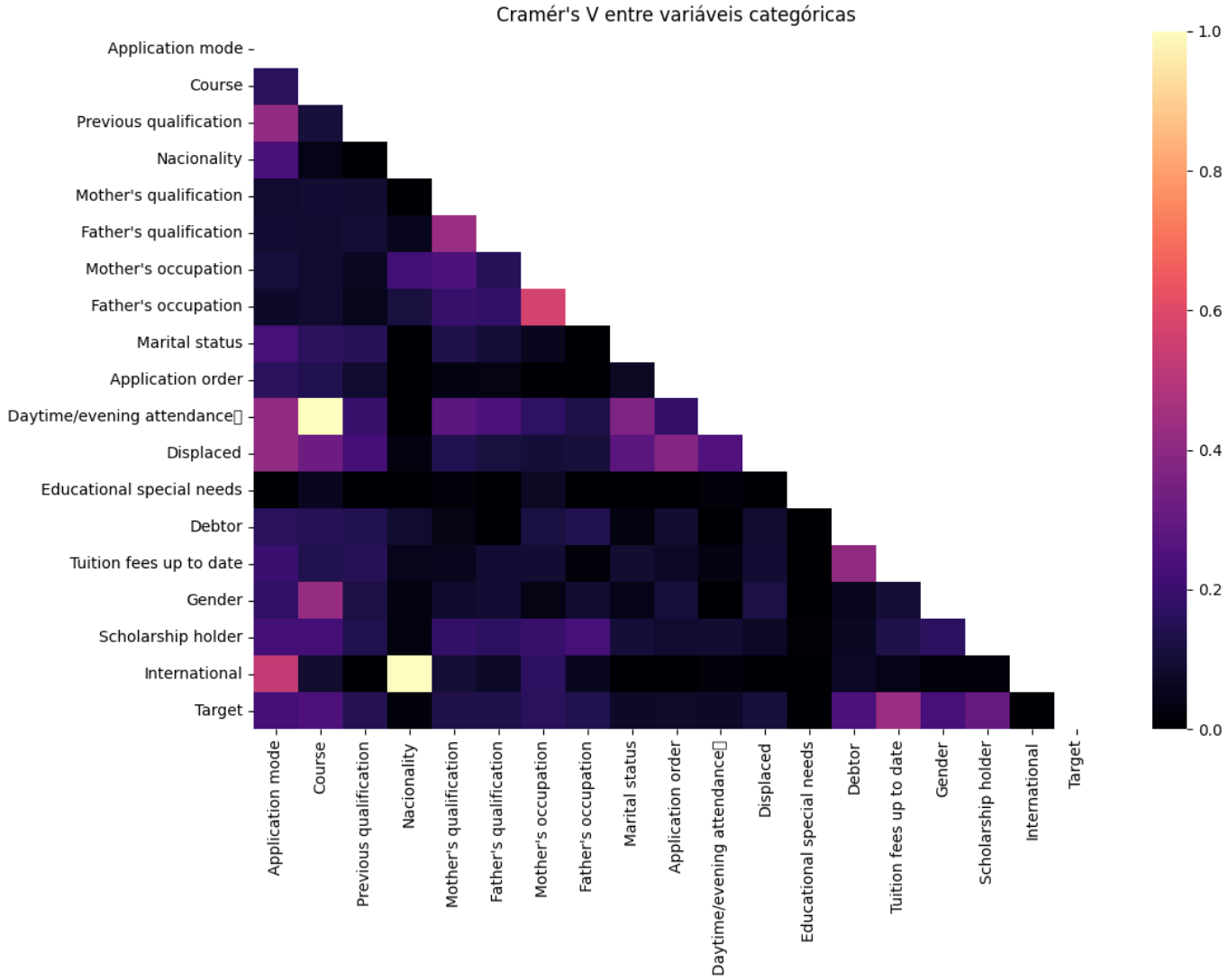


Figure 15: Cramér's V correlation between numerical and non-linear features

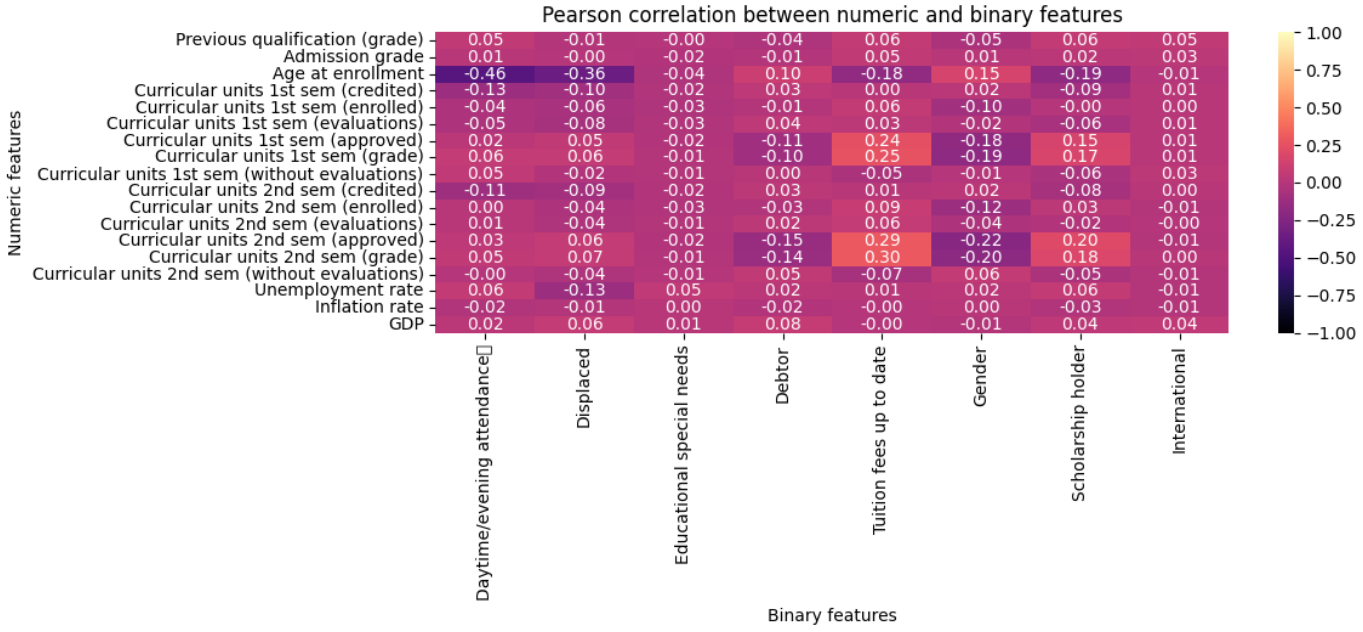
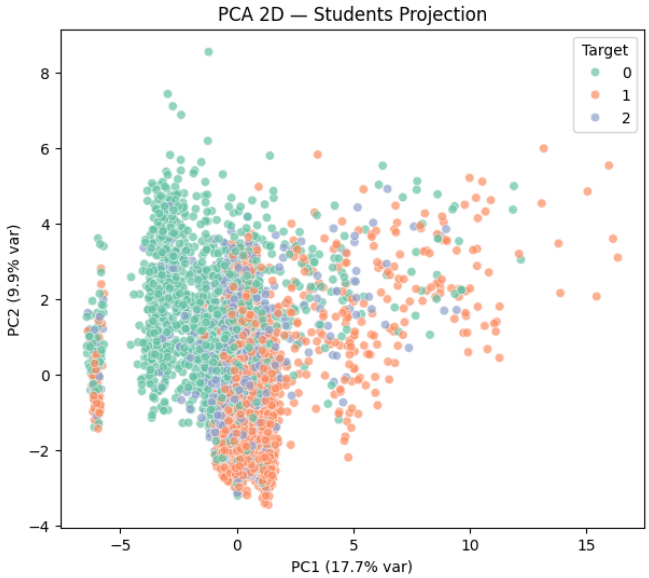
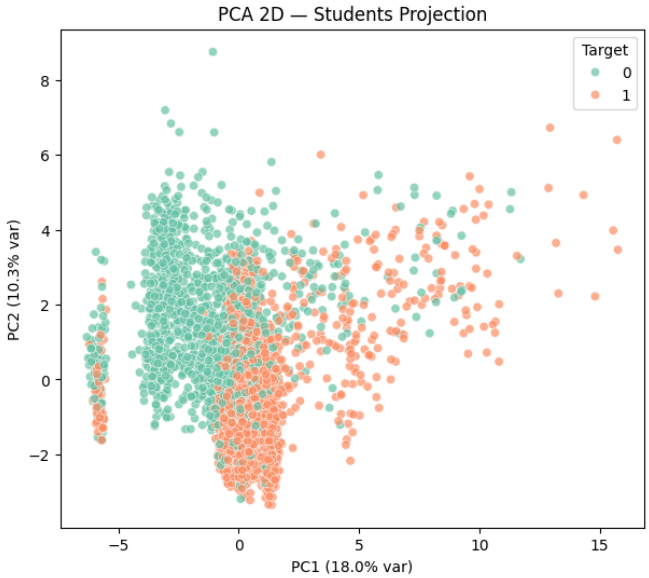


Figure 16: Pearson Correlation between Numerical and linear Features and Binary Features

APPENDIX C
PCA AND KERNEL PCA

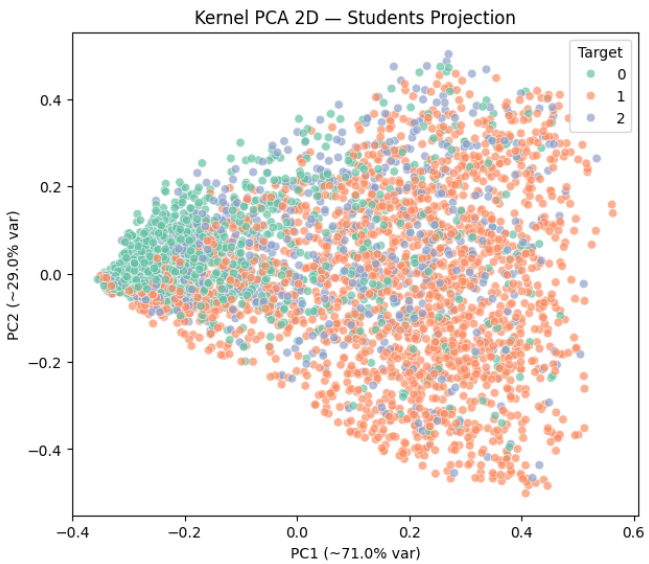


(a) PCA with Enrolled

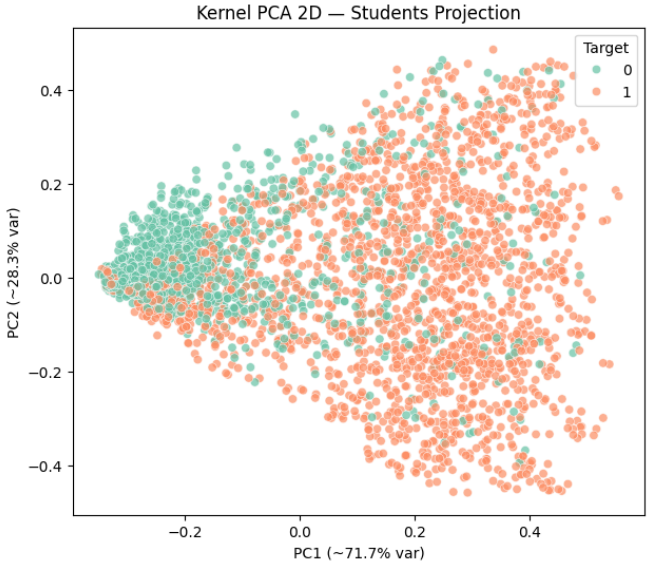


(b) PCA without Enrolled

Figure 17: PCA Analysis



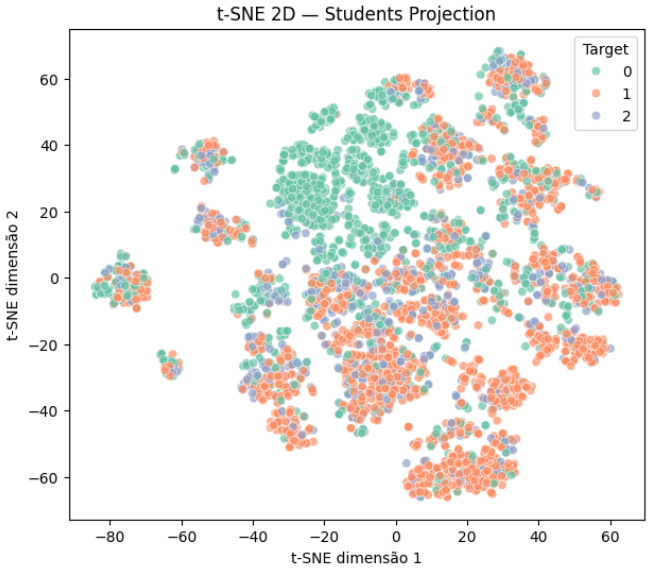
(a) Kernel PCA with Enrolled



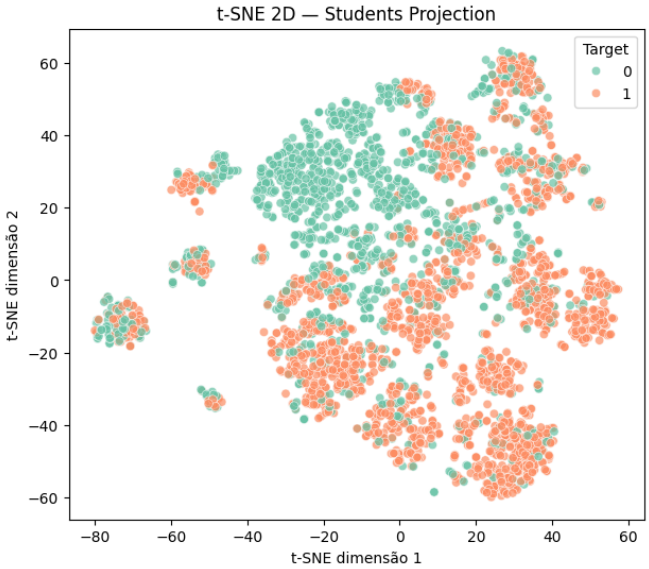
(b) Kernel PCA without Enrolled

Figure 18: Kernel PCA Analysis

APPENDIX D
T-SNE AND T-SNE USING KERNEL PCA

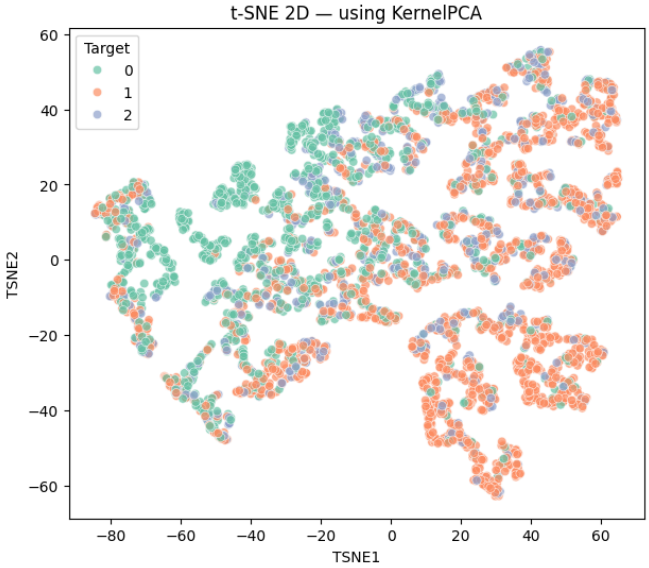


(a) t-SNE with Enrolled

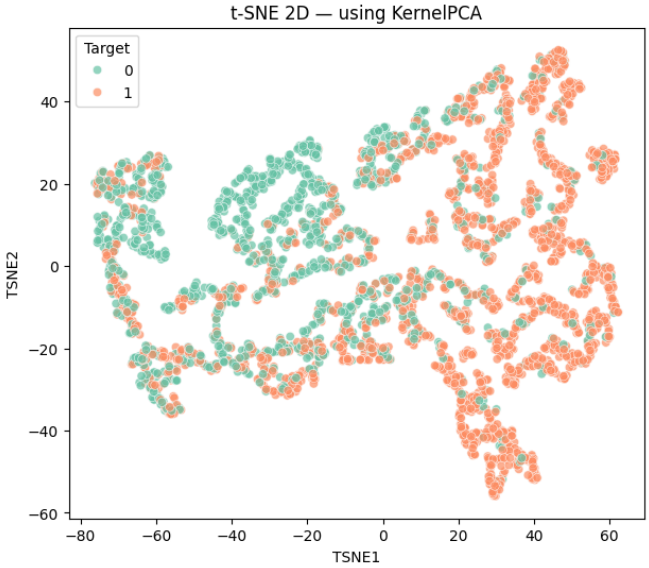


(b) t-SNE without Enrolled

Figure 19: t-SNE Analysis



(a) t-SNE using Kernel PCA with Enrolled



(b) t-SNE using Kernel PCA without Enrolled

Figure 20: t-SNE using Kernel PCA Analysis

APPENDIX E
PROTOTYPES AND CRITICISMS WITH MMD

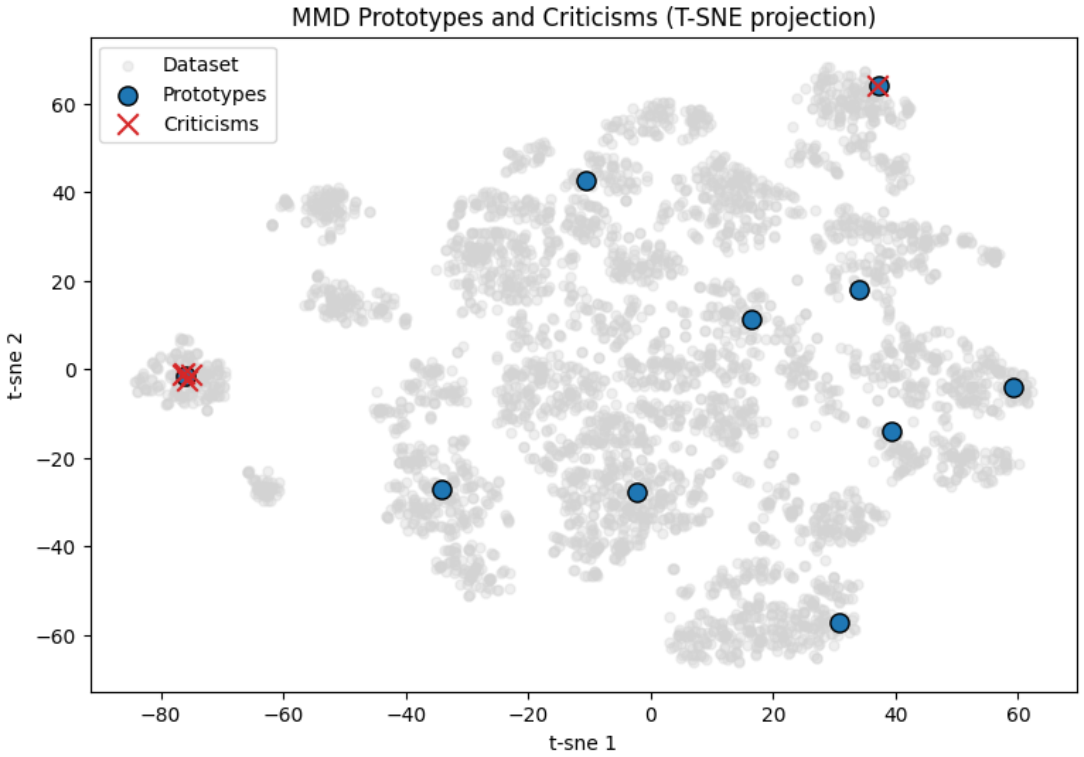


Figure 21: Prototypes and Criticisms with MMD with Enrolled

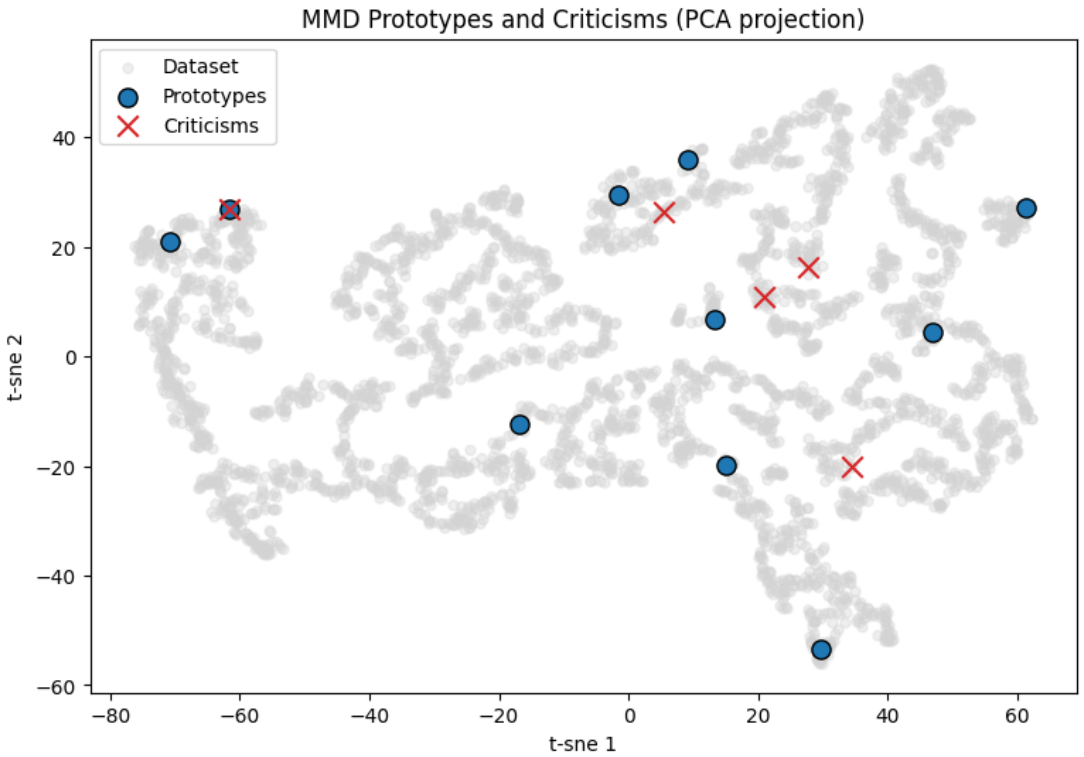


Figure 22: Prototypes and Criticisms with MMD without Enrolled

APPENDIX F
FULL DEPTH DECISION TREE: WITH VS WITHOUT ENROLLED

Table I: Full-depth Decision Tree: With VS Without Enrolled

Metric	Accuracy	F1	Precision	Recall
With Enrolled (3 classes)	0.677	0.616	0.619	0.617
Without Enrolled (2 classes)	0.860	0.851	0.852	0.849

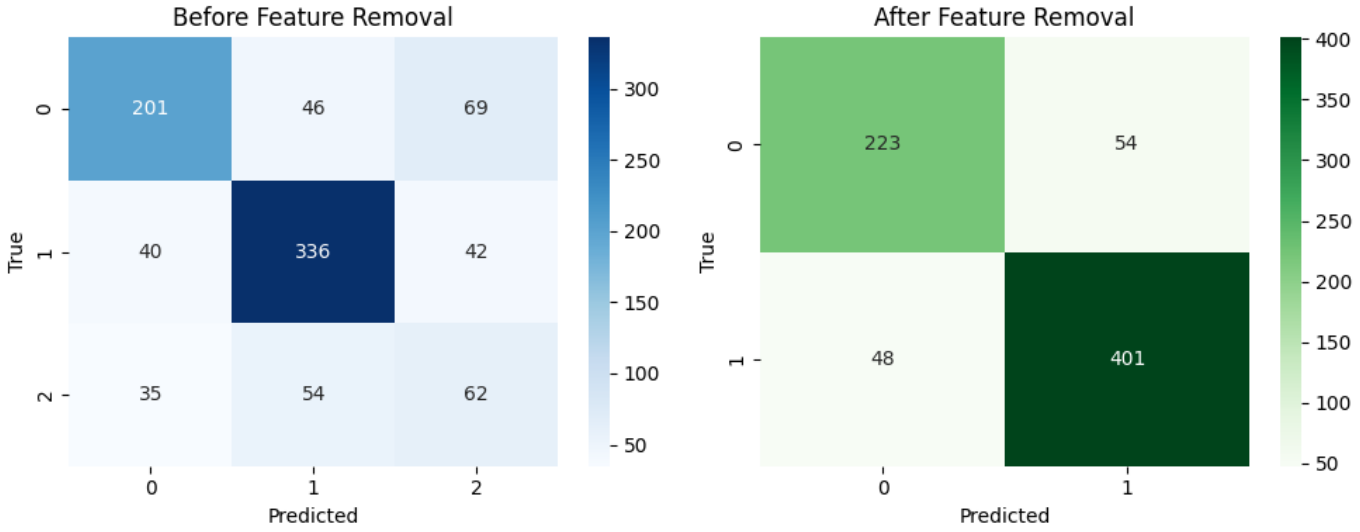


Figure 23: Correlation Matrices of Full-depth Decision Trees

APPENDIX G
LIMITED DEPTH DECISION TREE: WITH VS WITHOUT ENROLLED

Table II: Limited-depth Decision Tree: With VS Without Enrolled

Metric	Accuracy	F1	Precision	Recall
With Enrolled (3 classes)	0.715	0.616	0.663	0.609
Without Enrolled (2 classes)	0.891	0.881	0.900	0.870

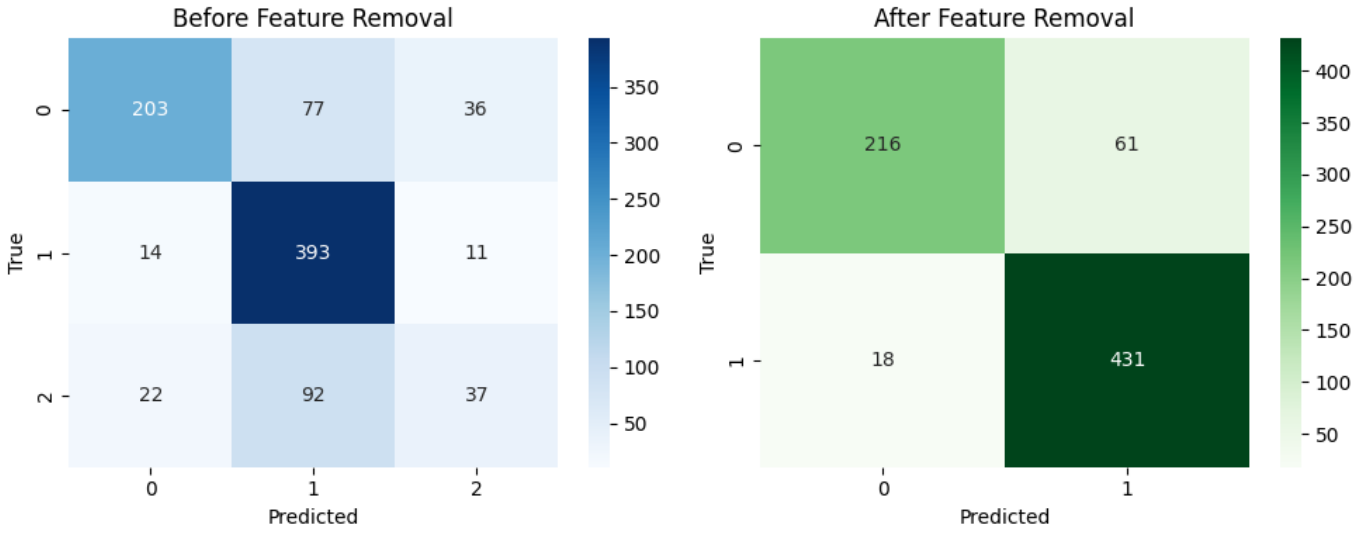


Figure 24: Correlation Matrices of Limited-depth Decision Trees

APPENDIX H
LIMITED VS UNLIMITED DT: WITHOUT ENROLLED

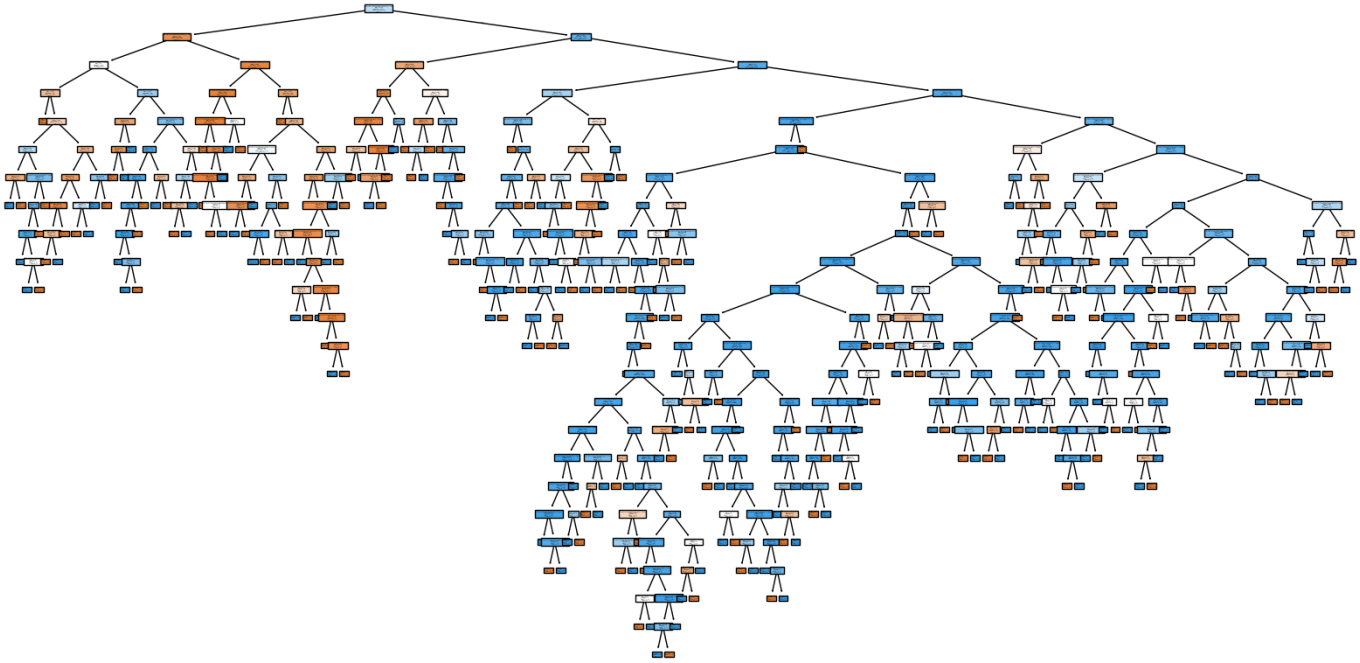


Figure 25: Full-depth Decision Tree Without Enrolled

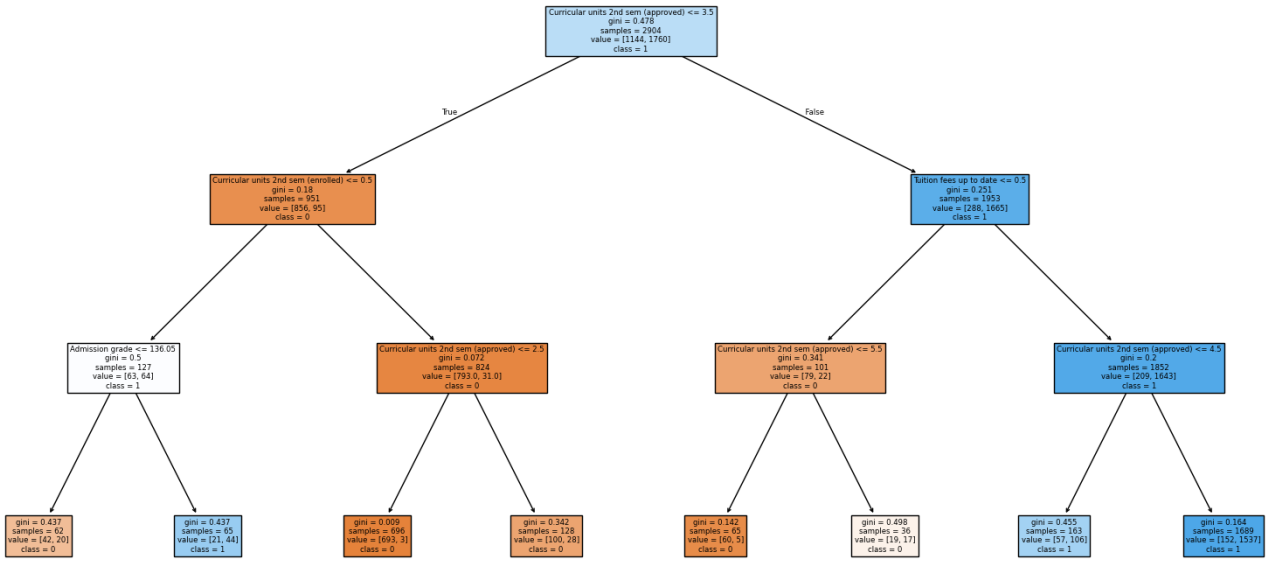


Figure 26: Limited-depth Decision Tree Without Enrolled

APPENDIX I
XGBOOST CLASSIFICATION REPORT

Table III: XGBoost Overall Classification Report

Metric	Accuracy	FI-Score	ROC-AUC
Value	0.8994	0.9195	0.9562

Table IV: XGBoost Detailed Classification Report

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	0.88	0.85	0.87	277
1	0.91	0.93	0.92	449
Accuracy		0.90		726
Macro avg	0.90	0.89	0.89	726
Weighted avg	0.90	0.90	0.90	726

APPENDIX J
XGBOOST SURROGATE TREE

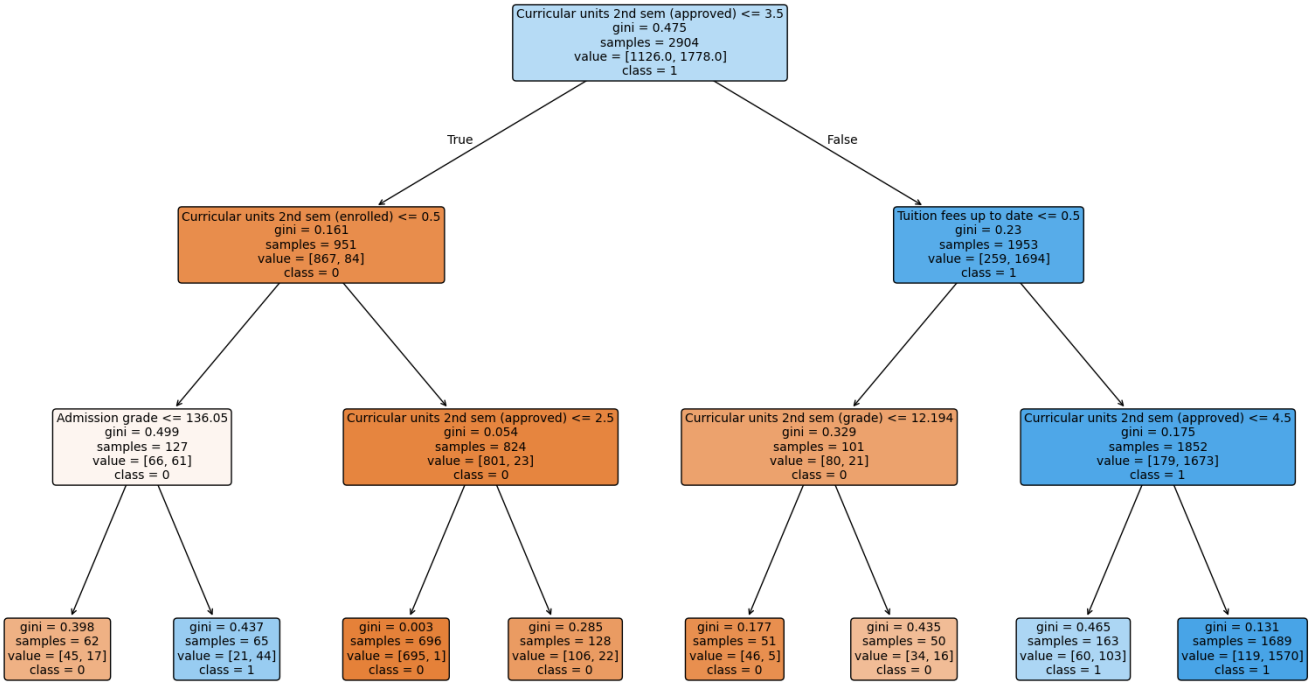


Figure 27: XGBoost Surrogate Tree

APPENDIX K
XGBOOST - FEATURE IMPORTANCE - PFI

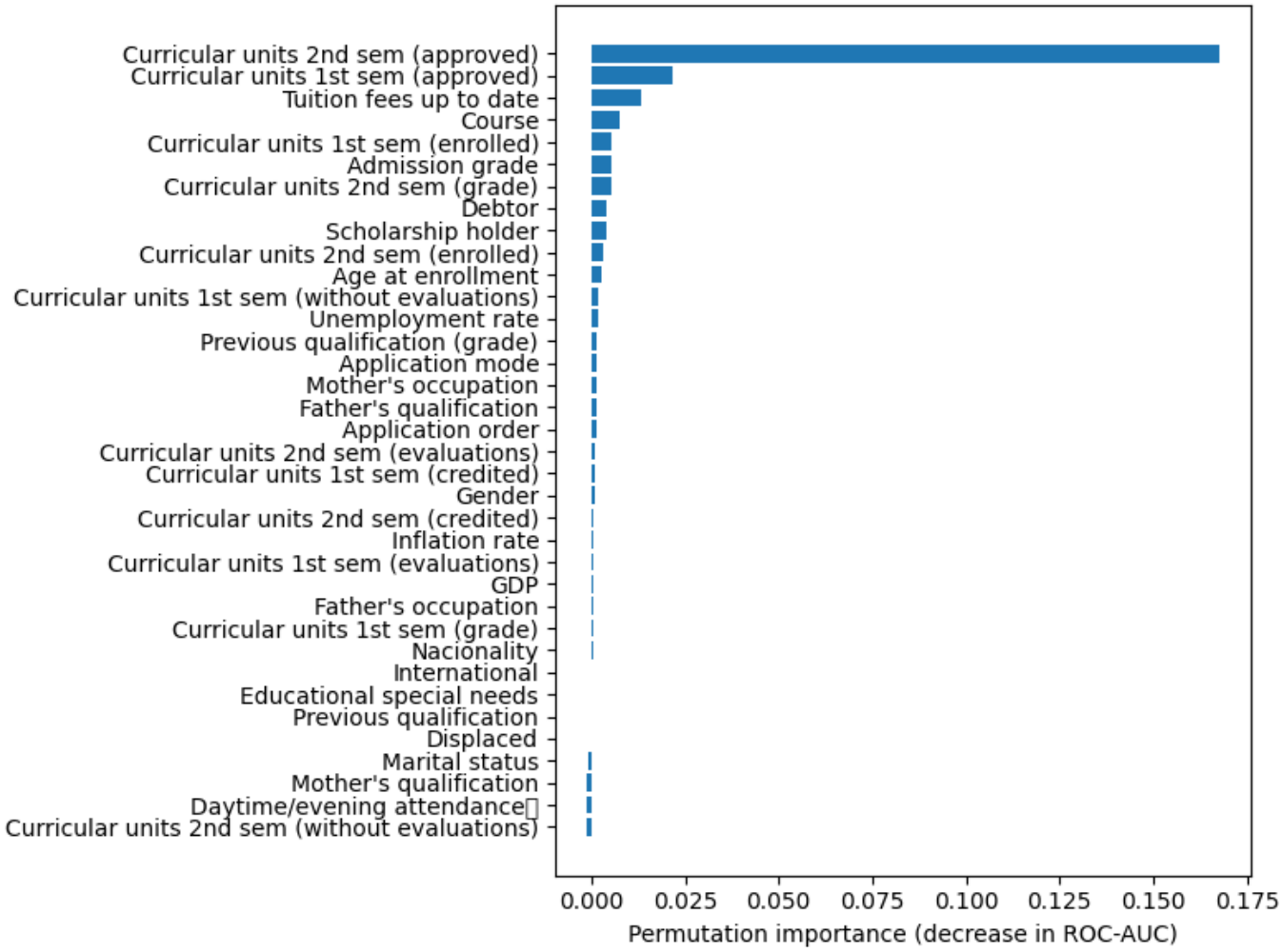


Figure 28: PFI (decrease in ROC-AUC)

APPENDIX L
XGBOOST - LOCAL EXPLANATIONS

To ensure a consistent analysis of the local explanations provided by LIME and SHAP, we applied these techniques to the same examples. Specifically, we chose the first case where the model's prediction was correct for both the Graduate and Dropout examples. After selecting the instances for analysis, we applied the XAI techniques.

LIME - Dropout Example:

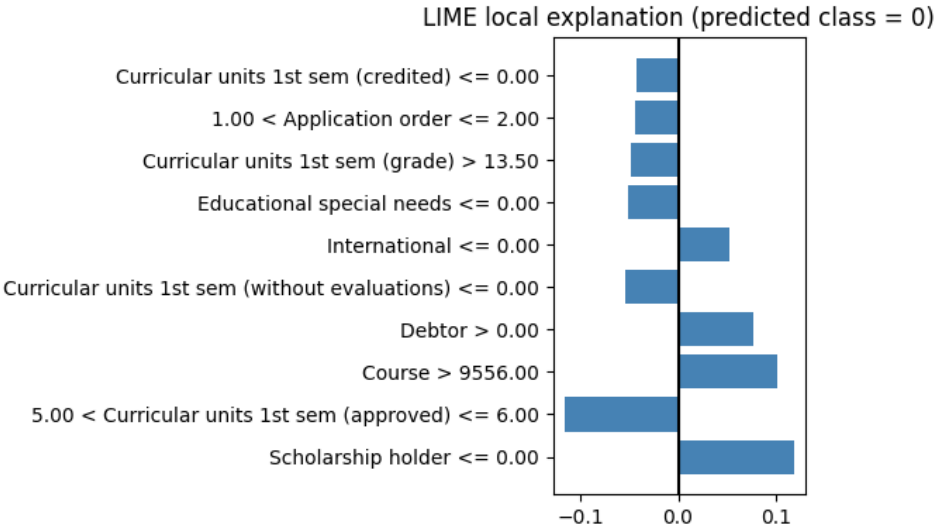


Figure 29: LIME Dropout Example

LIME - Graduate Example:

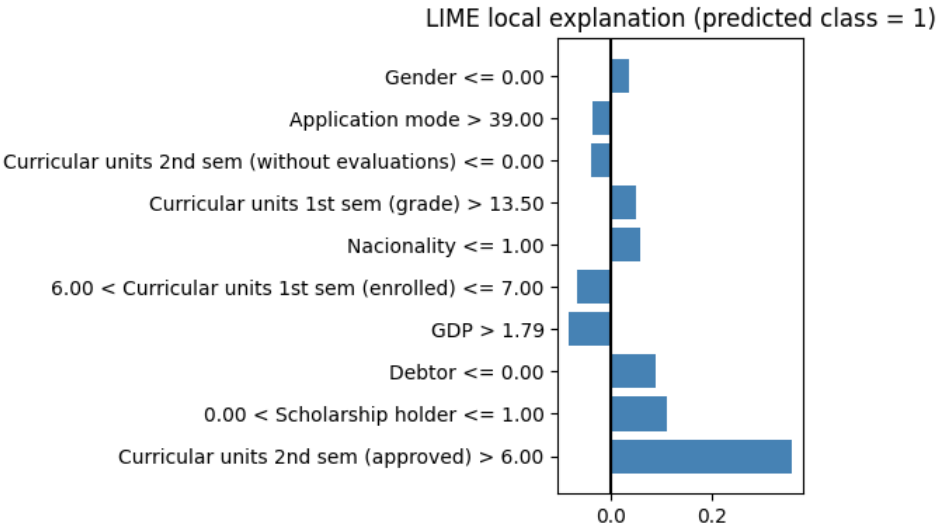


Figure 30: LIME Graduate Example

SHAP - Dropout Example:

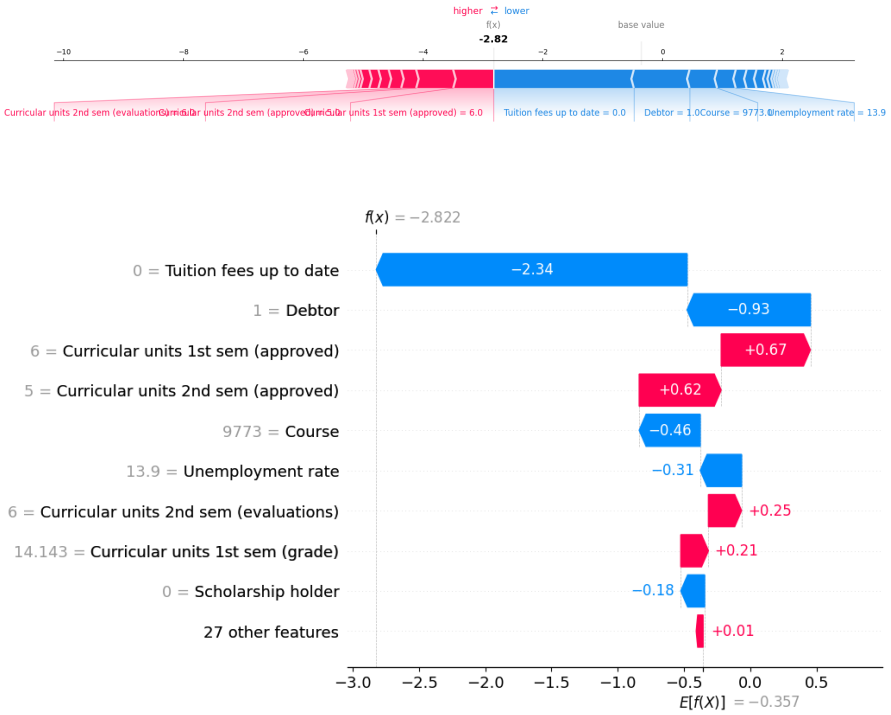


Figure 31: SHAP for Local Explanations

SHAP - Graduate Example:

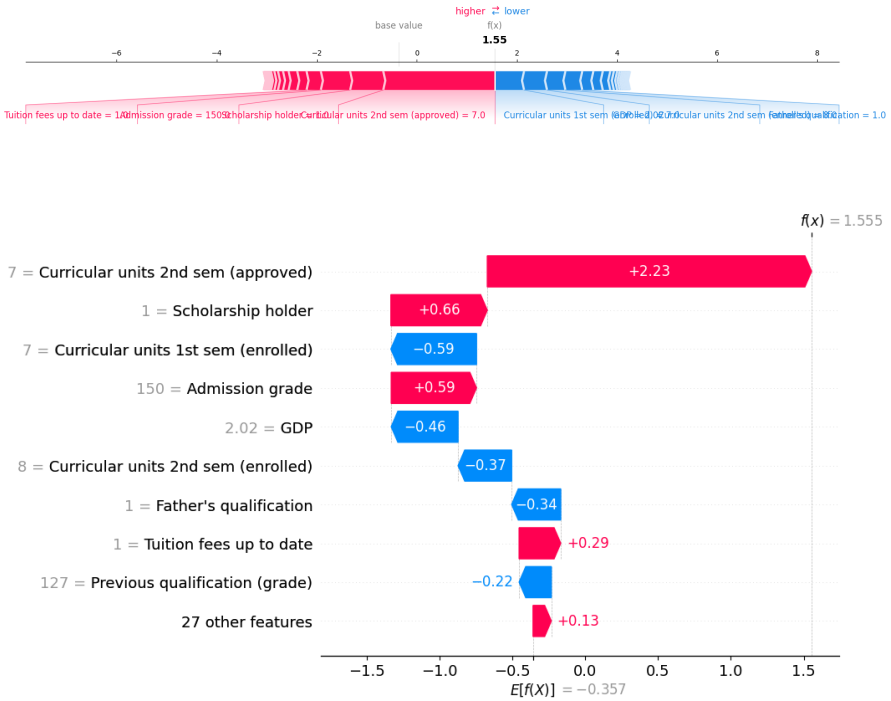


Figure 32: SHAP for Local Explanations

APPENDIX M
MLP - CLASSIFICATION REPORT

Table V: MLP Overall Classification Report

Metric	Accuracy	F1-Score	ROC-AUC
Value	0.9008	0.9205	0.9390

Table VI: MLP Detailed Classification Report

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	0.91	0.84	0.87	284
1	0.90	0.94	0.92	442
Accuracy		0.90		726
Macro avg	0.90	0.89	0.89	726
Weighted avg	0.90	0.90	0.90	726

APPENDIX N
MLP - SURROGATE TREE

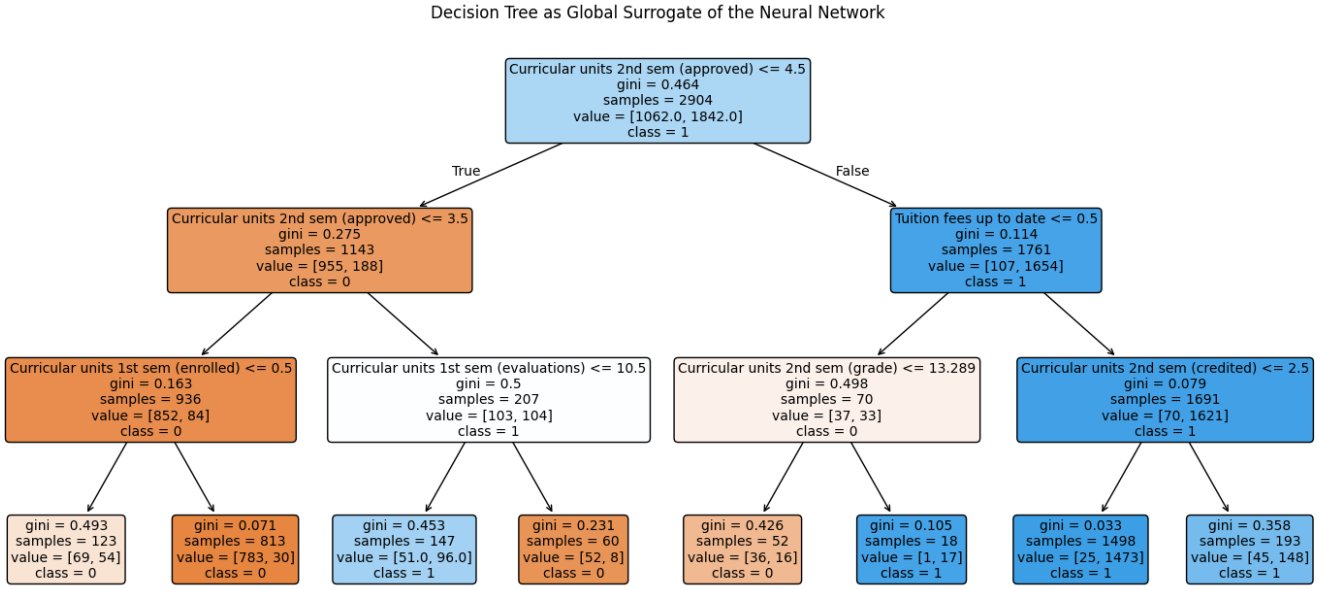


Figure 33: XGBoost Surrogate Tree

APPENDIX O
MLP - GLOBAL ASSESSMENT

A. *PFI*

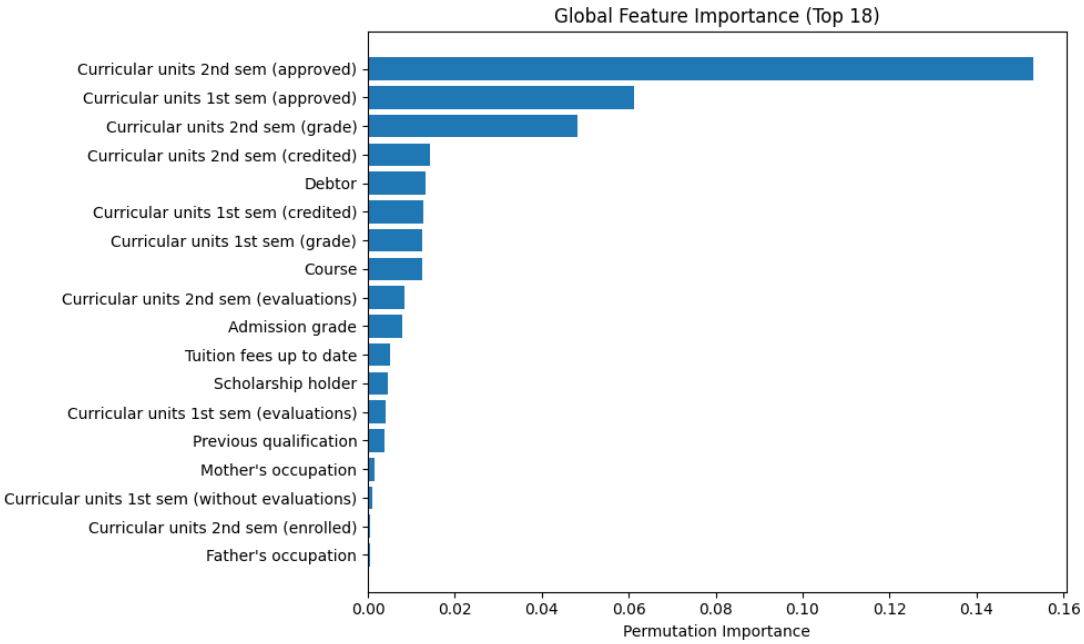


Figure 34: PFI (decrease in ROC-AUC)

B. *SHAP (KernelShap)*

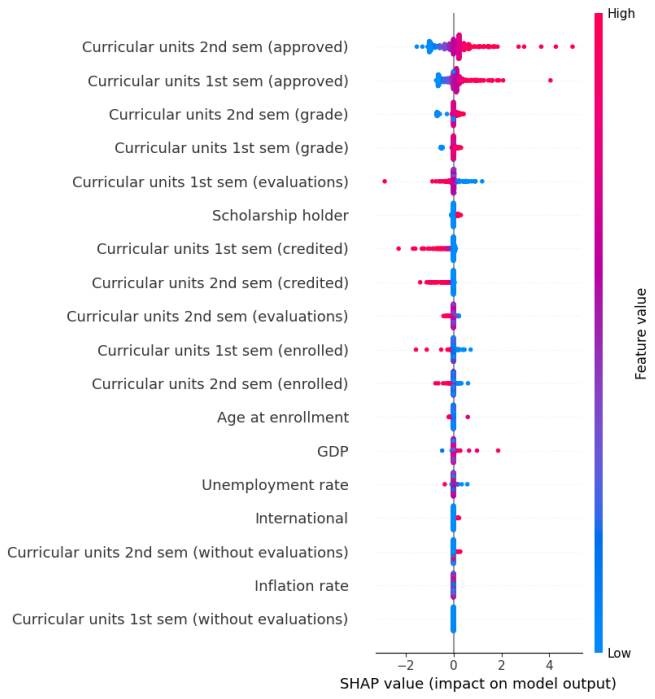


Figure 35: SHAP Values (Directional)

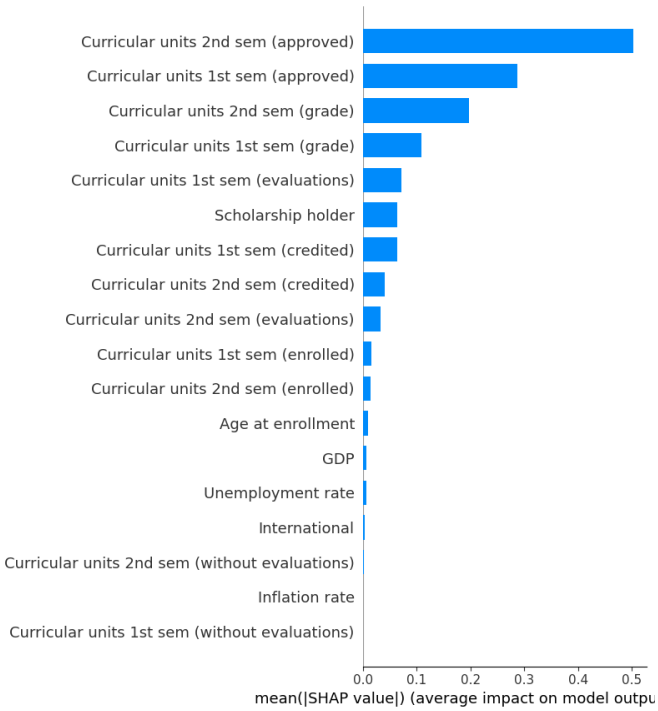


Figure 36: SHAP Values

APPENDIX P
MLP - LOCAL EXPLANATIONS

A. *LIME*

To preserve result compatibility, the examples analyzed in XGBoost and MLP will be the same. In this case, both classes were correctly classified by the MLP.

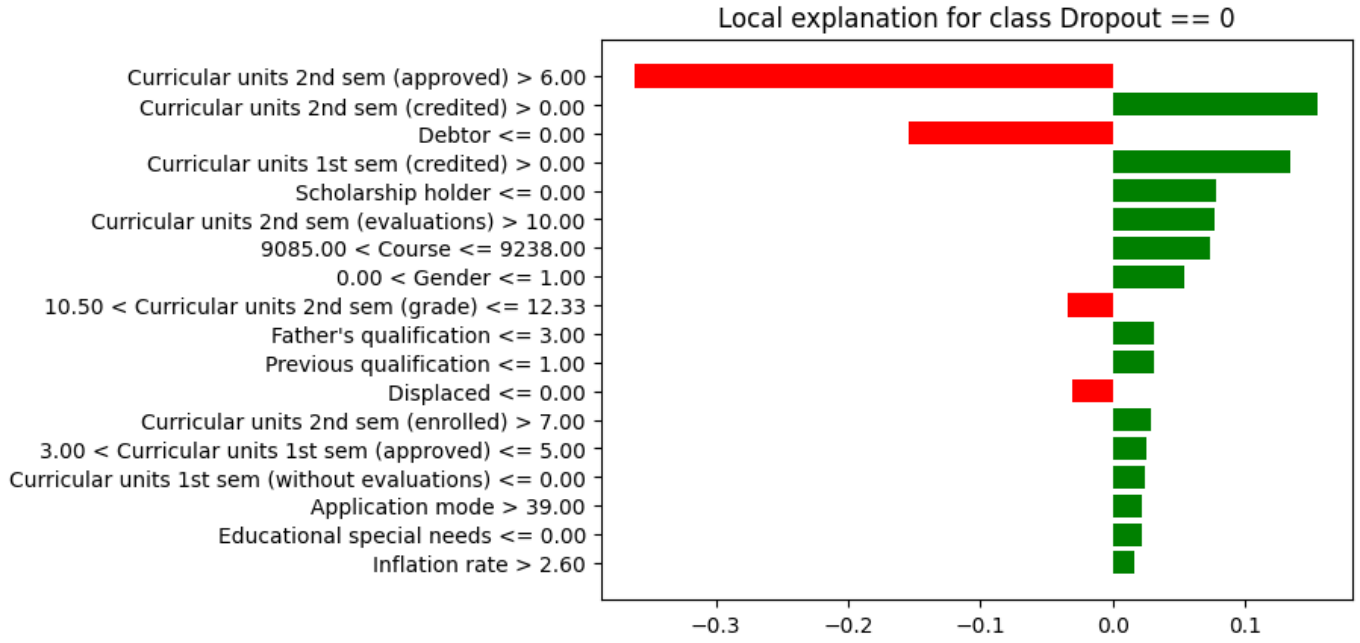


Figure 37: LIME - Dropout (0)

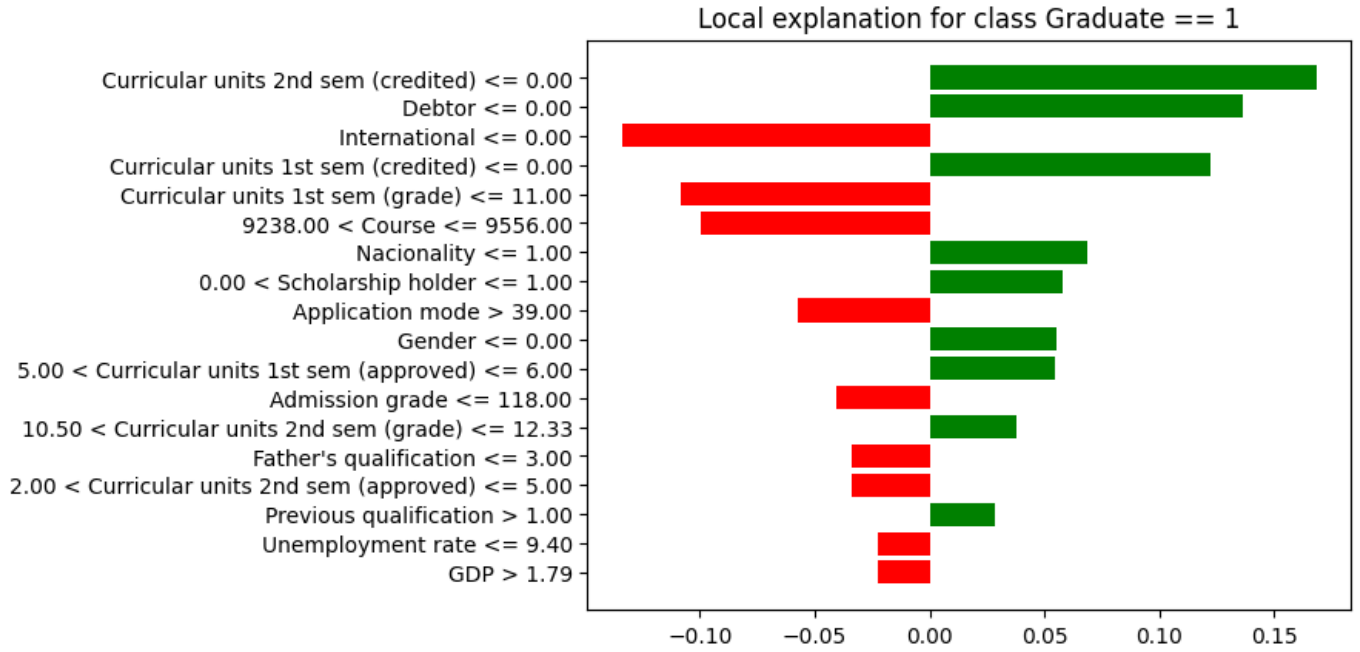


Figure 38: LIME - Graduate (1)

B. SHAP

SHAP - Dropout Example:

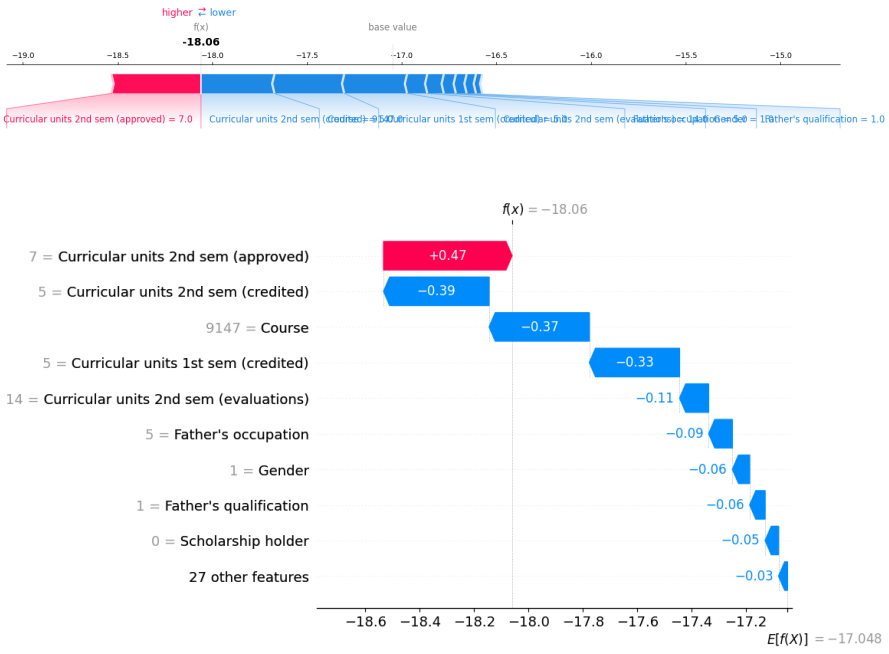


Figure 39: SHAP for Local Explanations

SHAP - Graduate Example:

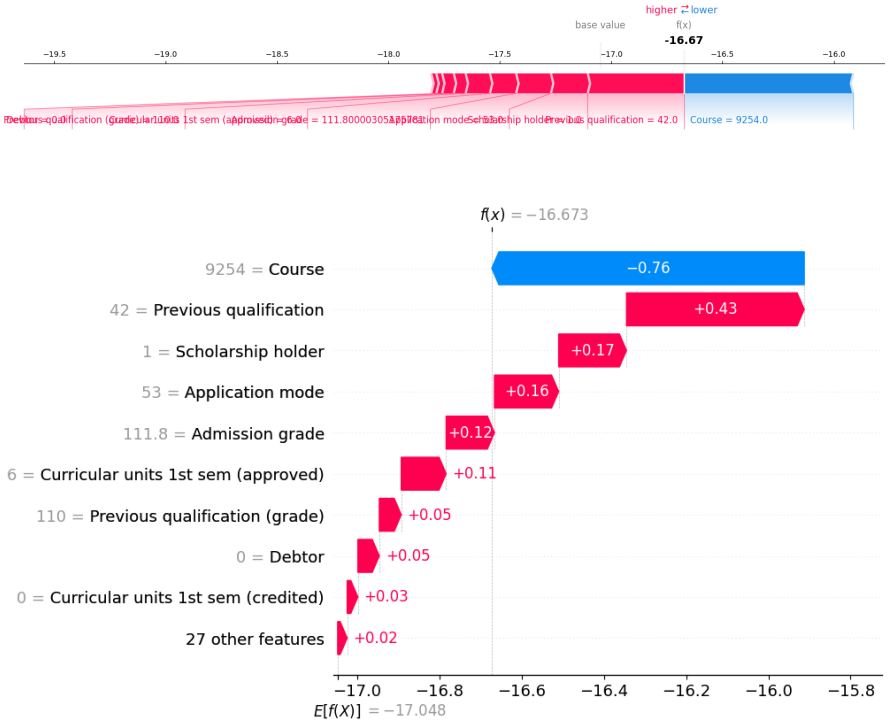


Figure 40: SHAP for Local Explanations

APPENDIX Q
ANCHORS

```
=== Student A (idx = 1) ===
Anchor: Curricular units 2nd sem (approved) <= 5.00 AND Debtor > 0.00 AND Curricular units 2nd sem (grade) <= 12.33 AND Course > 9238.00 AND Scholarship holder <= 0.00
Precision: 0.976
Coverage: 0.028
Prediction: 0
True label: 0
=== Student B (idx = 0) ===
Anchor: Curricular units 2nd sem (approved) > 6.00 AND Scholarship holder > 0.00
Precision: 0.959
Coverage: 0.070
Prediction: 1
True label: 1
```

Figure 41: Two Anchors for different class students - XGBoost

```
=== Student A (idx = 1) ===
Anchor: Curricular units 2nd sem (approved) <= 5.00 AND Debtor > 0.00 AND Curricular units 2nd sem (grade) <= 12.33 AND Course > 9238.00 AND Scholarship holder <= 0.00
Precision: 0.976
Coverage: 0.028
Prediction: 0
True label: 0
=== Student B (idx = 0) ===
Anchor: Curricular units 2nd sem (approved) > 6.00 AND Scholarship holder > 0.00
Precision: 0.959
Coverage: 0.070
Prediction: 1
True label: 1
```

Figure 42: Two Anchors for different class students - MLP