# Predict Students' Dropout and Academic Success

HarvardX PH125.9x - Data Science Capstone 2

Jasmine Zhang

*11 June, 2025*

# Contents

# 1 Introduction

Student retention and academic success remain critical challenges in higher education worldwide. As student populations diversify and institutions strive to maintain financial sustainability, understanding and mitigating the drivers of student dropout have become central to effective educational management. Educational Data Mining (EDM) has emerged as a powerful interdisciplinary field, enabling universities to harness data-driven methods to identify at-risk students and optimize retention strategies. The ability to accurately predict student dropout not only helps institutions improve academic outcomes but also reduces financial losses and supports broader educational equity goals.

This project uses an open dataset from the UCI Machine Learning Repository. This dataset is supported by program SATDAP - Capacitação da Administração Pública under grant POCI-05-5762-FSE-000191, Portugal.The dataset comprises 4,424 anonymized records from a Portuguese higher education institution, integrating information from multiple administrative sources. It includes macroeconomic context, student demographics, academic preparation and performance, and program characteristics, capturing a wide array of factors that are hypothesized to influence dropout and academic success. It is important to note that certain variables, such as credit hour, major classification, and application mode, differ from those commonly used in the US system. These distinctions will be explored in the subsequent exploratory data analysis steps.

The core objective of this study is to develop and evaluate predictive models that classify students into risk categories at the end of the normal course duration. Specifically, the target is formulated as a binary outcome—distinguishing between students who drop out and those who are considered on-track (including both enrolled and graduated students). Recognizing the complexity of dropout as a phenomenon influenced by both internal (endogenous) and external (exogenous) factors, the project incorporates features such as prior academic achievement, parental education, economic context, and course characteristics to improve predictive performance.

The workflow for this project consists of several stages: data preparation and cleaning, exploratory analysis, and model building, model selection, and performance validation. We compare several machine learning approaches including Random Forests, k-Nearest Neighbors, Naive Bayes, and Gradient Boosted Trees, using cross-validation and hold-out testing for robust performance assessment. To improve predictive performance, two ensemble approaches are employed: first, a simple majority voting ensemble incorporating all models; second, a refined ensemble that includes only the three best-performing models. Assuming that dropout is coded as the positive class (dropout = 1), in the end, the chosen ensemble model achieves a sensitivity of 0.72, a specificity of 0.94, and an overall accuracy of 0.87.

## 1.1 Dataset Summary

The initial distribution of the target variable is presented in Table 1. This breakdown shows that roughly one-third of students dropped out, while the remainder either remained enrolled or successfully graduated.

However, for the purpose of predictive modeling and interpretability, we recode the target variable into two categories:

-**dropout**: students who left their studies before completion

-**on_track**: students who are either still enrolled or have graduated

Table 1: Distribution of Target Results

| Target | Count | Percentage |
|---|---|---|
| Dropout | 1421 | 32.1 |
| Enrolled | 794 | 17.9 |
| Graduate | 2209 | 49.9 |

This binary classification allows us to focus on the most actionable question for institutional intervention: it streamlines the analysis by focusing on the distinction between students who disengage and those who persist in the academic pipeline.

## 1.2 Variable Processing

To ensure clarity, all 37 variables in the original dataset were carefully reviewed, cleaned, and organized into four major categories. Below, we detail the treatment of each variable, specifying transformations, recoding, or removal as appropriate.

**Category 1. Macroeconomic Context**

- **`Unemployment rate`**
  Retained as a numeric variable and rescaled for interpretability.
- **`Inflation rate`**
  Retained as a numeric variable, with type conversions as needed.
- **`GDP`**
  *Deleted.* Removed due to redundancy, as unemployment and inflation rates more directly reflect students' economic context.

**Category 2. Student Demographics**

- **`Gender`**
  Recoded to a binary indicator (`Is_Male`) for interpretability.
- **`Age at enrollment`**
  Retained as a numeric variable.
- **`Marital status`**
  Recoded from numeric codes to descriptive categories (e.g., *single*, *married*).
- **`Displaced`**
  Retained as a binary indicator.
- **`International`**
  Retained as a binary indicator (`Is_International`). The more granular `Nacionality` column was removed to avoid fragmentation and simplify modeling.
- **`Educational special needs`**
  Retained as a binary variable (`SpecialNeeds`).

- **`Parent education and occupation`**
  All detailed columns were consolidated into a single binary indicator (`Parent_Higher_Edu`), reflecting whether either parent attained postsecondary education. The original detailed columns were removed to reduce dimensionality and prevent overfitting.

- **`Scholarship holder`**
  Retained as a binary variable.

- **`Debtor`**
  Retained as a binary variable.

### Category 3. Academic Preparation and Performance

- **`Previous qualification & Previous qualification (grade)`**
  Only "Previous qualification (grade)" was retained (renamed `PriorGrade`), as it more directly reflects academic preparation. The categorical "Previous qualification" was deleted.

- **`Admission grade`**
  Retained and renamed (`AdmissionGrade`) for clarity.

- **`Curricular units (1st and 2nd semester)`**
  For both semesters, only the average grade (`Sem1_AvgGrade`, `Sem2_AvgGrade`) and units enrolled (`Sem1_UnitsEnrolled`, `Sem2_UnitsEnrolled`) were retained, as they best capture academic engagement and performance. All other related variables (credits, approvals, evaluations, etc.) were removed due to redundancy or limited predictive value.

- **`Application mode`**
  Recoded from numeric codes to descriptive labels and renamed (`Application_mode`) for interpretability.

- **`Application order`**
  Retained as an integer variable, reflecting the priority ranking of each student's application.

### Category 4. Program Characteristics

- **`Course`**
  Renamed as `Major` and recoded to descriptive program names for interpretability.

- **`Daytime/evening attendance`**
  Renamed as `Is_Daytime_Attendance` and recoded as a binary variable indicating attendance mode.

### Deleted Variables

Several variables were removed from the dataset during preprocessing. The rationale for each removal is detailed below:

- **`Previous qualification`**
  Removed because this categorical variable was redundant given the retention of "Previous qualification (grade)", which provides more direct and actionable information on academic preparedness.

- **`Nacionality`**
  Removed because analysis focused on domestic versus international status using the binary "International" indicator, rather than tracking specific nationalities. This reduces fragmentation and simplifies modeling.

- **`Mother's qualification, Father's qualification, Mother's occupation, Father's occupation`**
  Removed because these detailed parental background variables were consolidated into a single binary indicator (`Parent_Higher_Edu`) capturing whether either parent attained postsecondary education. This reduces dimensionality and minimizes the risk of overfitting.

- **`Tuition fees up to date`**
  Removed because this variable may reflect student outcomes or institutional processes after admission, introducing potential data leakage and offering little value for early prediction of dropout or academic success.

- **`Curricular units 1st sem (credited), Curricular units 1st sem (evaluations), Curricular units 1st sem (approved), Curricular units 1st sem (without evaluations), Curricular units 2nd sem (credited), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (approved), Curricular units 2nd sem (without evaluations)`**
  Removed because these variables are highly correlated with retained measures of academic engagement ("average grade" and "units enrolled"), add redundancy, and offer limited incremental predictive value.

- **`GDP`**
  Removed because in the presence of unemployment and inflation rates, GDP does not provide additional explanatory power and is less directly related to immediate student outcomes.

# 2 Exploratory Data Analysis

## 2.1 Macroeconomic Context

### 2.1.1 Inflation Rate

Figure 1 shows how students in the dataset are distributed across different inflation rate cohorts. It is important to note that the inflation rate variable is not smoothly continuous in this dataset; rather, it appears in a set of discrete steps, likely corresponding to one value for each cohort or year. Most inflation rate categories have between 300 and 500 students, but there is a clear peak at 1.4%, where nearly 800 students are observed. This indicates that, during the years covered by the dataset, a significant portion of students experienced an inflation rate around this value, while fewer students were present in periods with either lower or higher inflation.

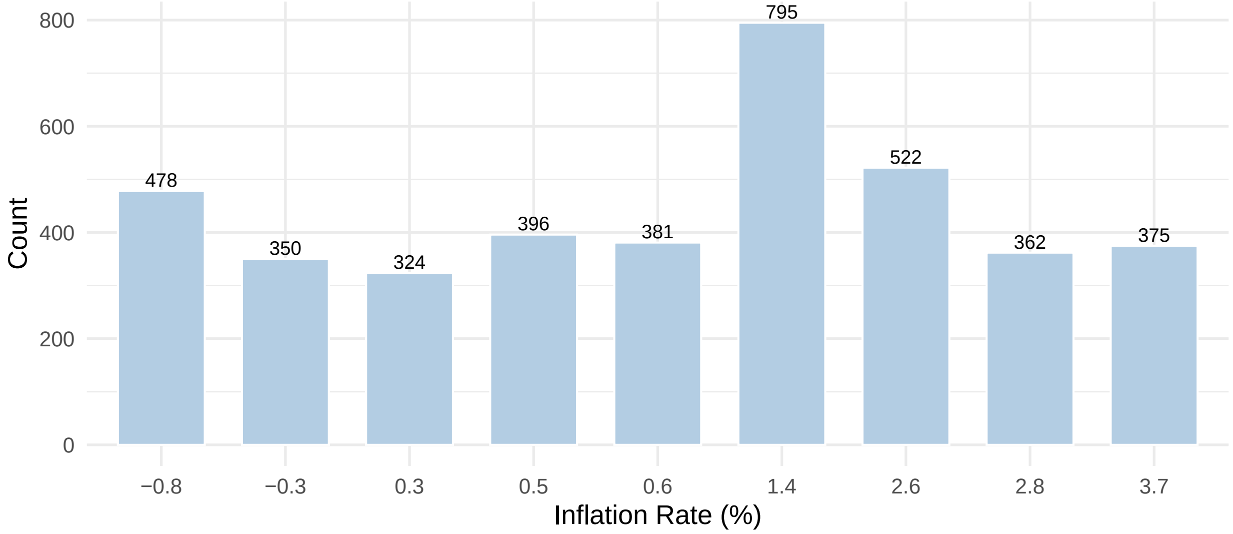

Figure 1: Counts by Unique Inflation Rate

Figure 2 provides a closer look at how the student outcome—specifically, whether a student dropped out or stayed on track—varies by inflation rate. For each inflation rate, the bar is divided into percentages of "dropout" and "on_track" students. In most cohorts, the majority of students remain on track, but the proportion of dropouts fluctuates noticeably. For example, at an inflation rate of -0.3%, the dropout rate rises to over 40%, while at 0.5%, it drops to just 23%. Overall, there is no clear linear relationship between inflation rate and student dropout, but certain cohorts demonstrate notably higher or lower dropout proportions, suggesting that macroeconomic context may influence student persistence in specific circumstances.

### 2.1.2 Unemployment rate

Figure 3 shows the relationship between unemployment rate and student outcomes. It displays the percentage of students classified as "dropout" or "on_track" for each unique unemployment rate value in the dataset. Similar to the inflation rate, the unemployment rate variable is not continuous;

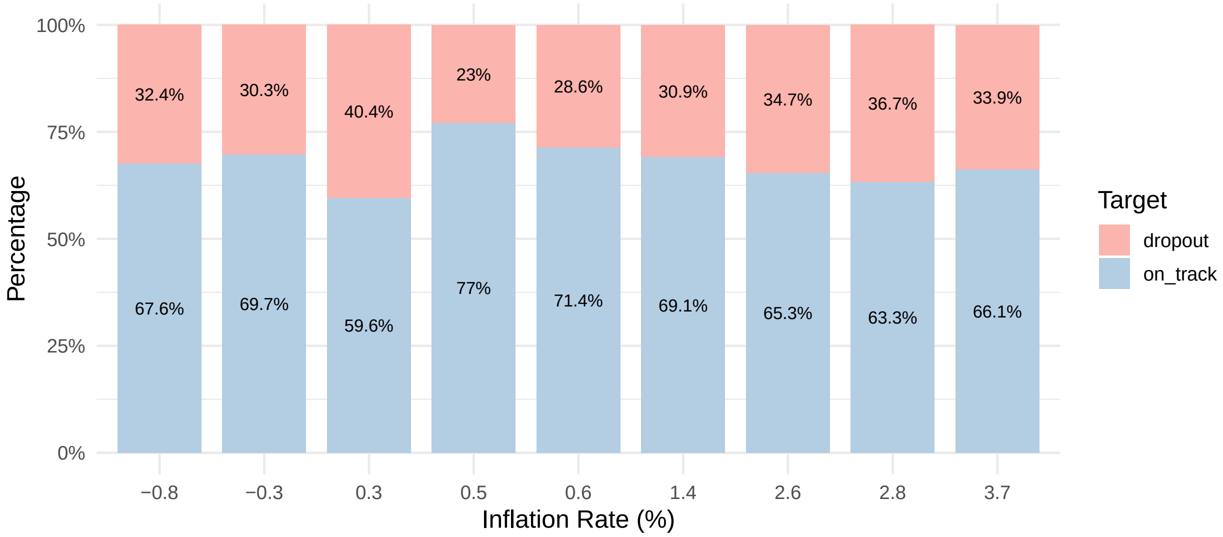

Figure 2: Target Status Breakdown by Inflation Rate

instead, it appears in discrete increments. This likely reflects economic conditions that correspond to particular years or cohorts.

In most unemployment rate categories, a larger proportion of students remain on track compared to those who drop out. However, the dropout rate varies noticeably across unemployment levels. For example, the dropout percentage is lowest at an unemployment rate of 12.4 percent (23 percent) and increases as unemployment rises, reaching 40.4 percent at 16.2 percent. This pattern suggests a potential association between higher unemployment rates and an increased risk of student dropout.

Including unemployment rate as a predictor in dropout modeling is justified for several reasons. The unemployment rate captures broader macroeconomic conditions that can influence students' financial stability, perceived job prospects, and overall stress. These factors are known to affect educational persistence. By accounting for changes in the economic environment, the model can better distinguish between individual-level effects and those driven by external economic pressures.

## 2.2 Student Demographics

### 2.2.1 Gender, Displaced Status, International Status, Special Needs, Parent Education, Financial Aid, Loan Borrowing

Figure 4 shows the percentage of students classified as "dropout" and "on_track" within each subgroup of several key demographic variables. Each subplot represents a different binary demographic variable. Across most variables, students in the "on_track" category make up the majority. However, the proportion of dropouts differs noticeably depending on demographic group.

Among students with outstanding debt, 38.7% dropped out compared to 28.4% of those without debt. This suggests that financial stress may contribute significantly to dropout risk. For displaced students, the dropout rate is 27.5%, lower than the 37.7% observed among non-displaced students, possibly reflecting the presence of institutional support mechanisms that benefit displaced students.

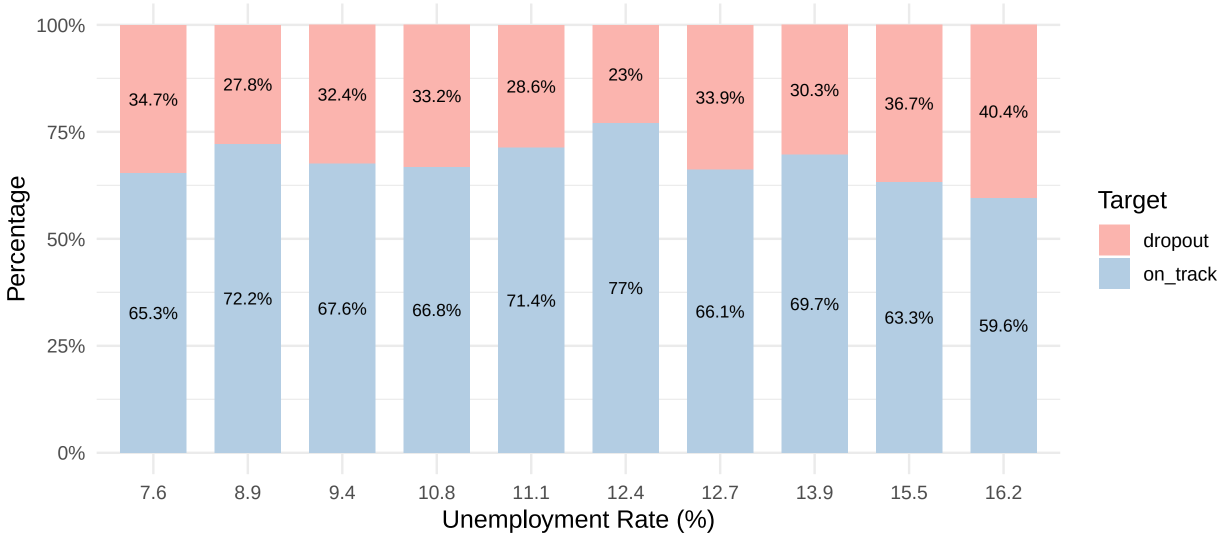

Figure 3: Target Status Breakdown by Unemployment Rate

Scholarship status has a particularly strong effect. Only 12.3% of scholarship recipients dropped out, in contrast to a much higher 38.7% among those without scholarships, highlighting the important role of financial aid in promoting retention. When considering international status, 28.7% of international students dropped out compared to 32.2% of domestic students. This may indicate that international students are, on average, more motivated or benefit from additional institutional support.

The analysis also reveals marked gender differences. Male students had a dropout rate of 44.9%, nearly twice the 25.3% rate seen among female students, underscoring gender as a significant predictor of persistence. Family educational background also appears to matter: students whose parents have postsecondary education (`Parent_Higher_Edu: Yes`) have a dropout rate of 30.5%, lower than the 32.5% for those whose parents do not have such a background.

Students with special needs exhibit a dropout rate of 37%, higher than the 32.1% for students without special needs. This suggests that additional challenges faced by these students may not be fully addressed by current institutional support structures.

### 2.2.2 Age at Enrollment

Figure 5 shows the distribution of age at enrollment for students by their eventual outcome. The density plot reveals that most students enroll at a traditional college-going age, with a pronounced peak around 18 to 20 years old. Within this age group, the proportion of students who stay on track is noticeably higher than those who drop out, as indicated by the larger blue area under the curve.

As age at enrollment increases, the density for both groups declines, but a key difference emerges: the dropout group has a relatively higher density among older entrants compared to the on-track group. In other words, students who begin their studies at a later age are more likely to drop out than those who enroll immediately after high school. The "dropout" curve remains consistently above the "on_track" curve for most ages above 25, suggesting that non-traditional or adult learners

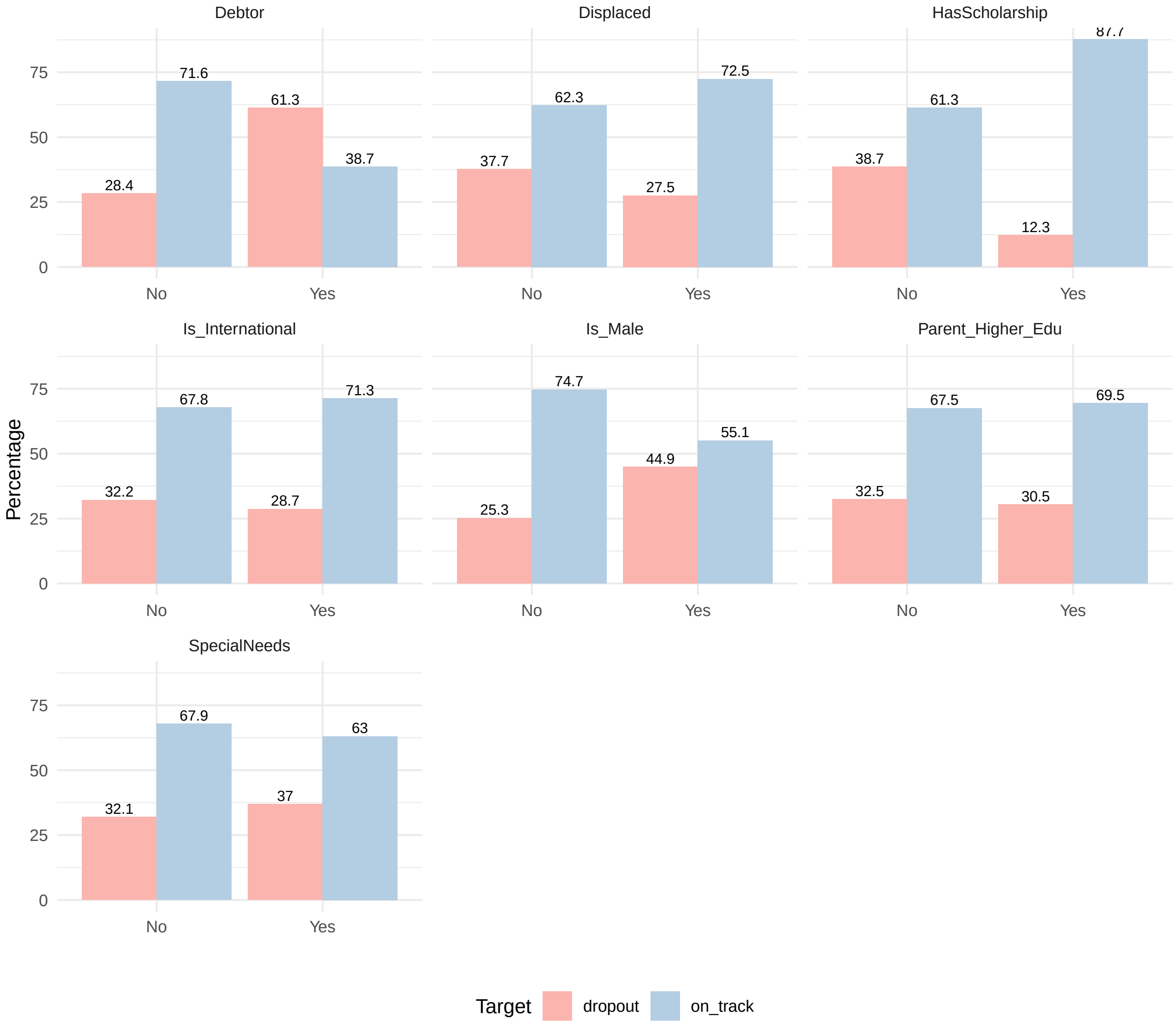

Figure 4: Percentage of Target Status by Student Demographic Variables

face additional challenges in persisting to completion.

These results highlight age at enrollment as an important predictor of student success. The increased risk of dropout among older entrants may reflect factors such as work and family obligations, competing priorities, or gaps in academic preparation. Including age at enrollment in predictive models can improve the accuracy of risk identification and support targeted interventions for students who may benefit from additional resources.

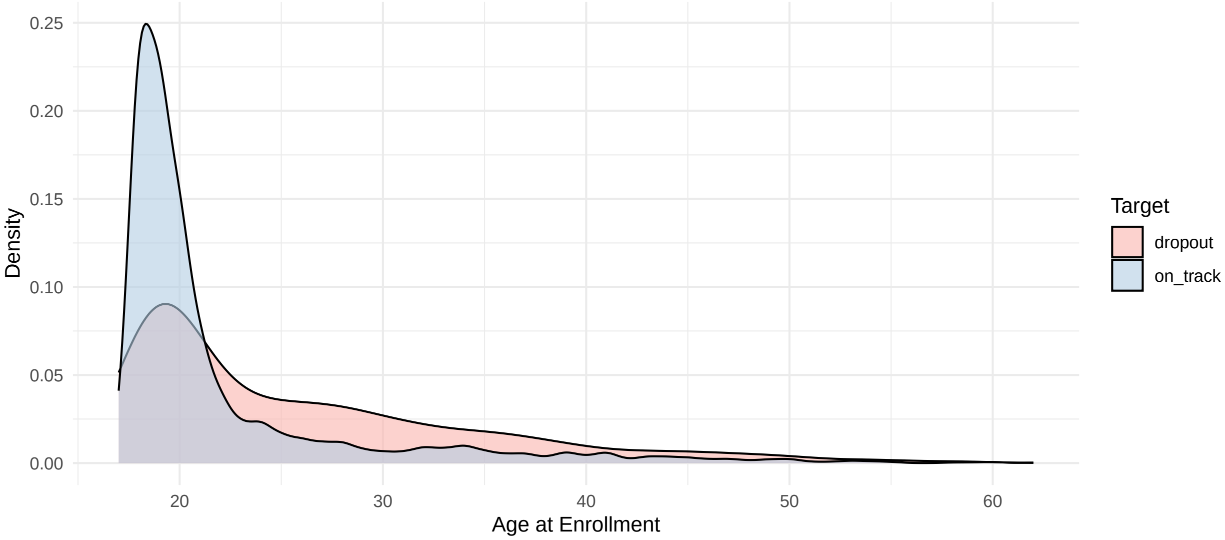

Figure 5: Distribution of Age at Enrollment by Target Status

### 2.2.3 Marital Status

Figure 6 shows the proportion of students across different categories of marital status. The plot reveals substantial variation in dropout rates among these groups. For example, students who are single have a dropout rate of 30.3 percent, while those who are married experience a much higher dropout rate of 46.2 percent. Similarly, the dropout rates for widowed, divorced, and those in a facto union are all above 33 percent, and the rate for legally separated students reaches as high as 75 percent.

This considerable variability suggests that marital status may affect student persistence, potentially reflecting differences in life responsibilities, support systems, and competing obligations outside of school. Given the wide range of dropout rates observed between categories, the marital status is included as a predictor in the model.

## 2.3 Academic Preparation and Performance

### 2.3.1 PriorGrade

Figure 7 shows the density distribution of prior grades for students. The plot reveals that the distribution for both groups peaks in the same general range, but there are noticeable differences in the shape and height of the curves.

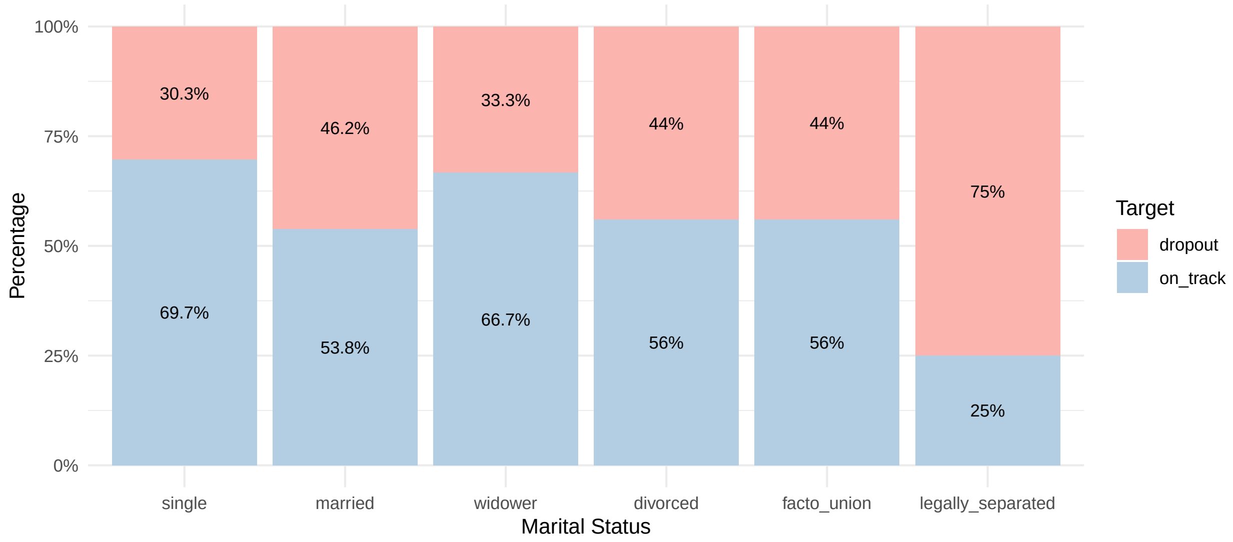

Figure 6: Proportion of Target Status by Marital Status

Students who dropped out tend to have a higher density at lower prior grade values, with the dropout curve peaking more sharply at lower grades compared to the on-track group. In contrast, students who remained on track show a relatively higher density at moderate and higher prior grade values. This indicates that students who enter higher education with stronger academic preparation, as reflected by higher prior grades, are more likely to persist and succeed.

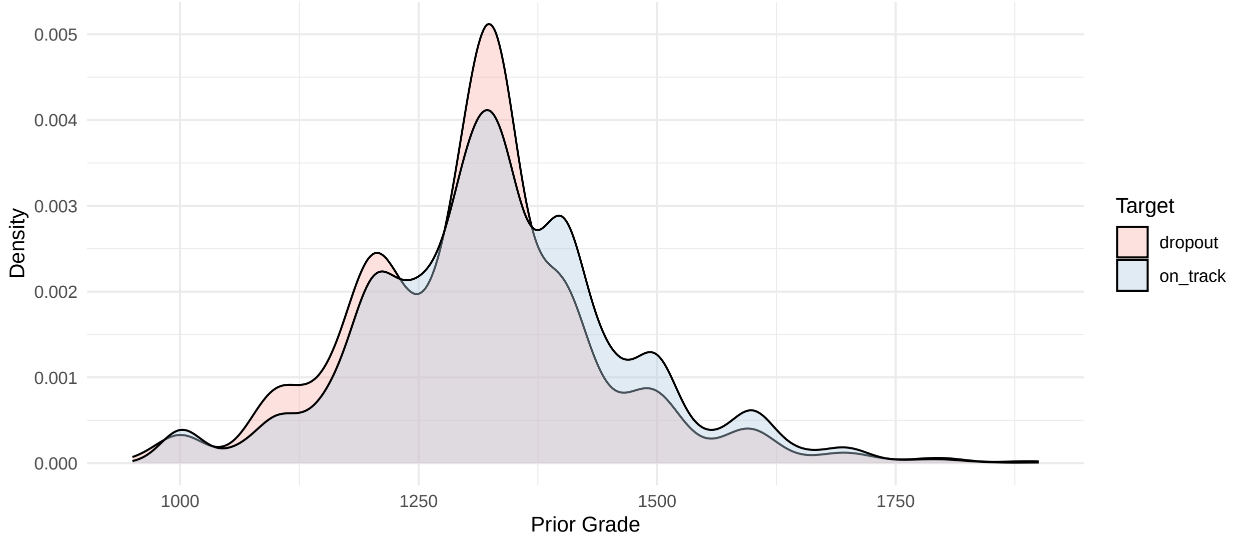

Figure 7: Prior Grade Density by Target Status

### 2.3.2 AdmissionGrade

Figure 8 shows the density distribution of admission grades for students. The two curves reveal important differences in academic profiles at entry. Students who dropped out generally had lower admission grades, as indicated by the higher density of the dropout curve at the lower end of the scale. In contrast, the on-track group shows a peak at higher admission grades and maintains higher density values throughout the upper grade ranges.

This pattern suggests that students entering with stronger admission grades are more likely to stay on track and complete their studies, while those admitted with lower grades are at greater risk of leaving early. The overlap between the two distributions also highlights that admission grade alone does not fully explain student outcomes, but it is nonetheless an indicator of future academic success.

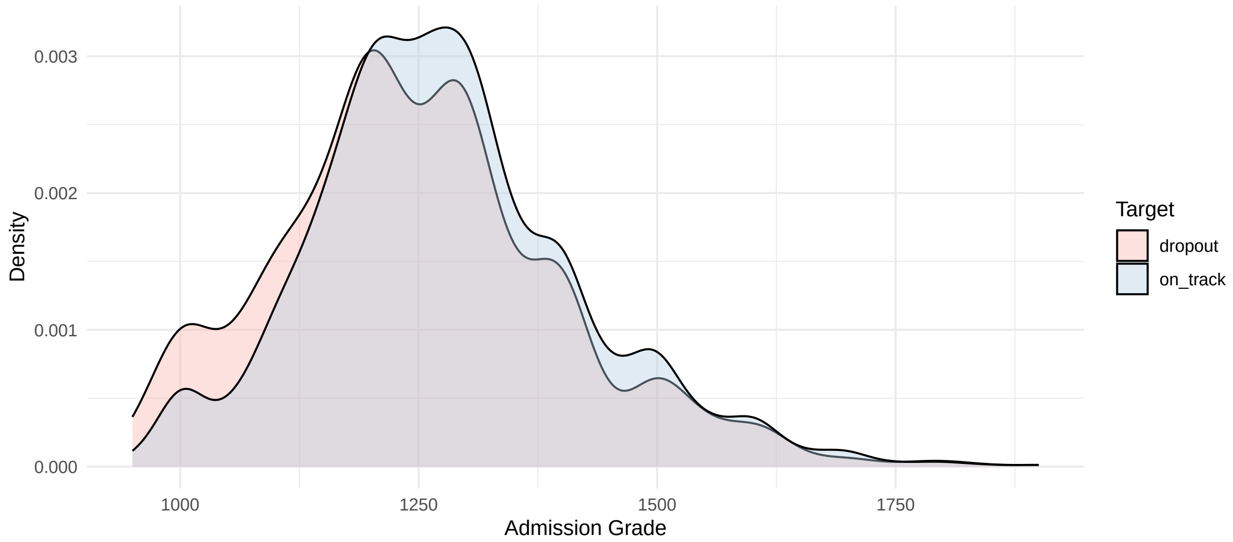

Figure 8: Admission Grade Density by Target Status

#### 2.3.3 1st Semester Academic Performance

Before interpreting the gradig results, please note that the grading system used in this dataset differs from the U.S. system. Table 2 summarizes the distribution of Semester 1 average grades. The median grade is 12.3, with most grades falling between 11 (Q1) and 13.4 (Q3), and the maximum observed grade is 18.9. The mean is approximately 10.6, and the standard deviation is 4.84. These summary statistics highlight that the grading scale and grade distribution follow a different structure than commonly seen in U.S. institutions, and should be interpreted accordingly.

Figure 9 visualizes the relationship between the number of units students enrolled in during their first semester and their average grade, separated by final outcome. Each point represents an individual student, and the smoothed lines show the general trend for each group.

The plot reveals that students who remained on track generally achieved higher average grades than those who dropped out, across almost all levels of course load. For students who enrolled in more units, the average grade for the "on_track" group remains consistently higher and shows a slight upward trend as units increase, while the "dropout" group's average grade is lower and tends to level off or even decline at higher unit counts. This suggests that students who are able to successfully manage a larger course load tend to perform better academically and are more likely to persist.

Table 2: Summary Statistics for Semester 1 Average Grade

| Statistic | Value |
|---|---|
| Min | 0.000000 |
| Q1 | 11.000000 |
| Median | 12.300000 |
| Mean | 10.649059 |
| Q3 | 13.400000 |
| Max | 18.900000 |
| SD | 4.844823 |
| N | 3983.000000 |

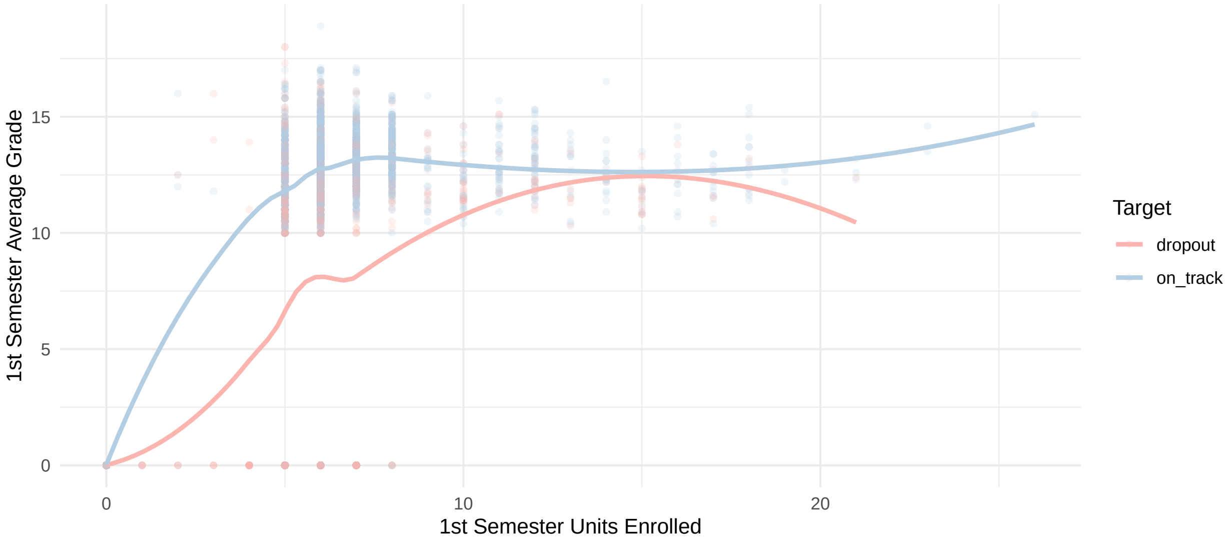

Figure 9: Trend: 1st Sem Avg Grade vs. Units Enrolled by Target

### 2.3.4 2nd Semester Academic Performance

Figure 10 plots 2nd semester average grade against units enrolled by student outcome, shows a trend similar to that seen in Figure 9 for the first semester. In both figures, students who remain on track consistently earn higher average grades than those who drop out, regardless of course load. The "on_track" group's performance remains strong even as the number of units increases, while the "dropout" group's average grades are lower and tend to plateau or decline with heavier course loads.

### 2.3.5 Application Mode

Figure 11 shows the breakdown of student outcomes by application mode, with each bar illustrating the proportion of students who were either "on track" or had dropped out. Notably, the dropout rates vary widely across different application modes; for example, certain categories such as "Ordinance_854_B_99" and "Ordinance_533_A_b2/b3" exhibit dropout rates as high as 100%, whereas others like "1st_phase_special_Azores" and "international_student" show much lower

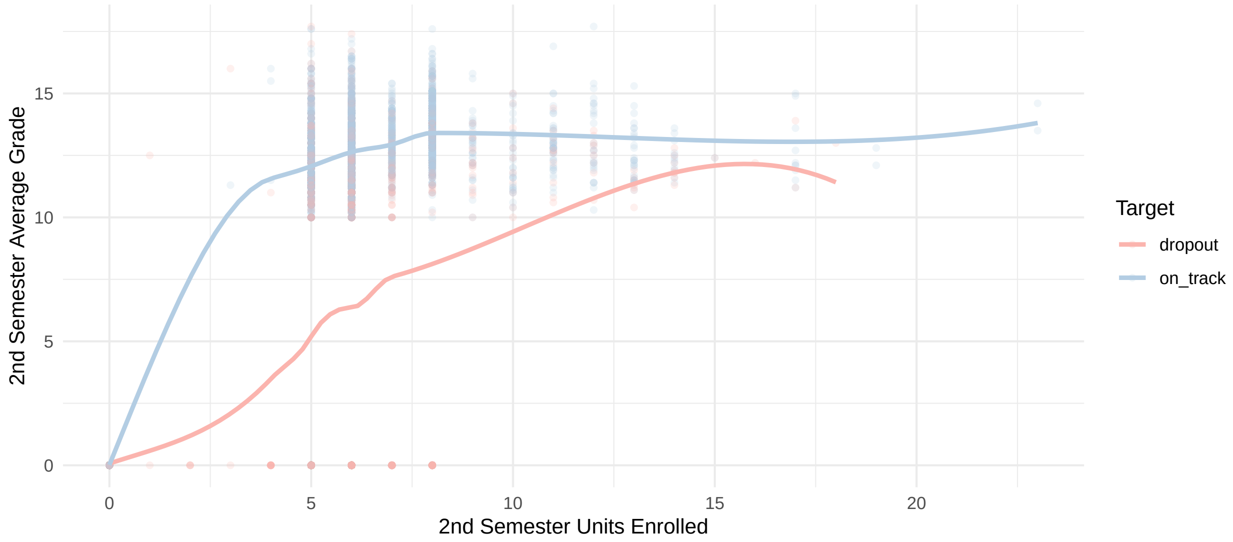

Figure 10: Trend: 2nd Sem Avg Grade vs. Units Enrolled by Target

dropout rates. This variation highlights that the way in which students enter the institution—through different application modes—is associated with their subsequent academic outcomes. The differences across modes may reflect a range of underlying factors, such as student motivation, or specific institutional pathways, which can significantly impact persistence.

### 2.3.6 Application Order

Figure 12 shows the relationship between application order and subsequent student outcomes. The on-track rate, shown in blue, increases as the application order rises, while the dropout rate, depicted in red, correspondingly decreases. Interestingly, this trend goes somewhat against common expectations, as one might assume students admitted through their first-choice application would be more likely to succeed. However, this pattern suggests that in the Portuguese context, students who enter through later application choices are actually more likely to persist. This may be due to differences in the national application process or other unique factors within the Portuguese higher education system. In any case, we respect what the data shows us here and recognize that application order is a meaningful factor in predicting student persistence.

## 2.4 Program Characteristics

### 2.4.1 Major

Figure 13 shows the distribution of student outcomes across different majors. The results show a clear variation in persistence and dropout rates among different fields of study. For example, majors such as Nursing and Social Service have the highest on-track rates (84.3% and 81.3%, respectively), while programs like Biofuel Production Technology, Equinculture, and Informatics Engineering have much lower on-track rates and correspondingly higher dropout rates. This variation is likely influenced by the innate nature and demands of different academic programs. Some fields may be

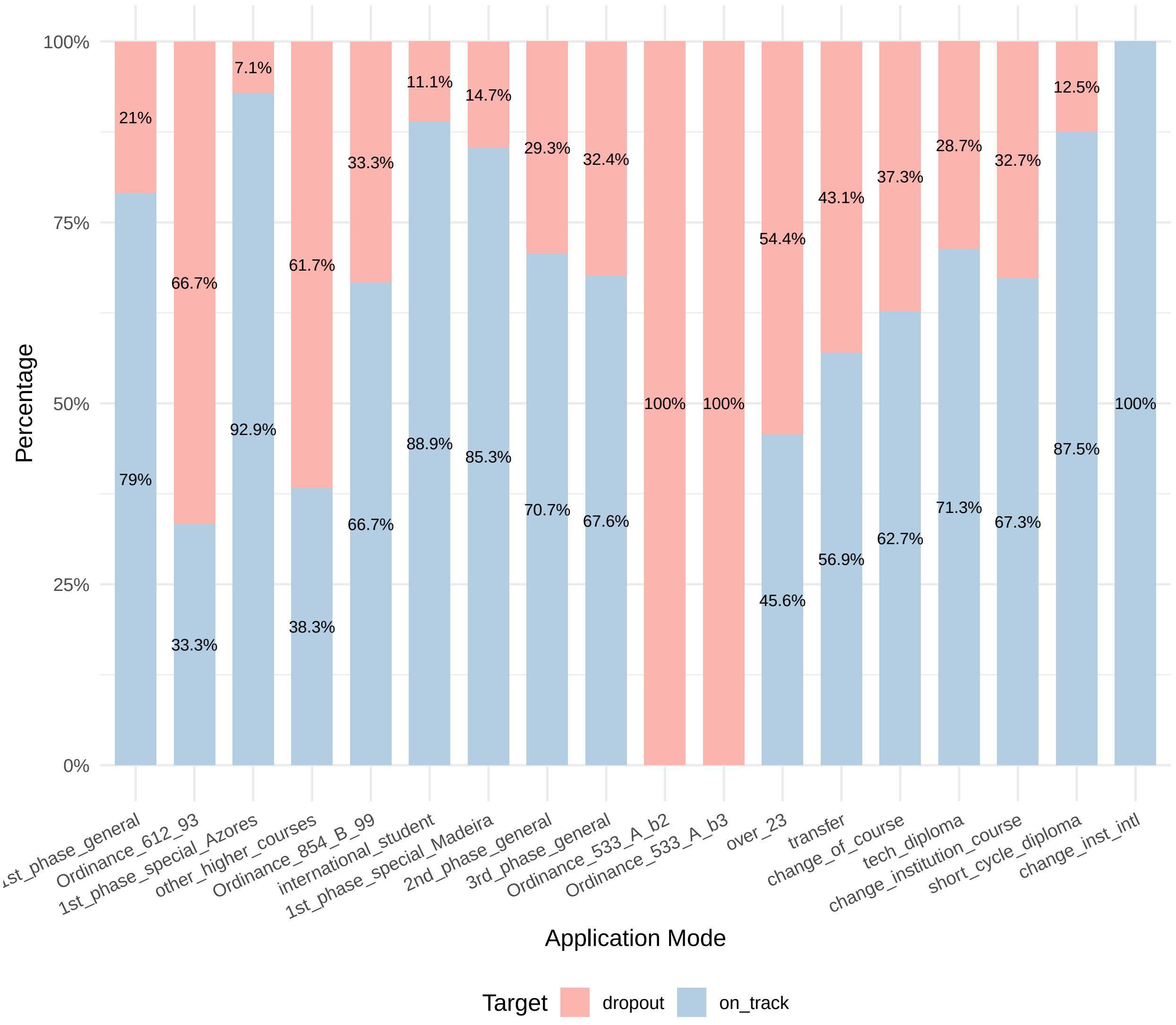

Figure 11: Target Result Breakdown by Application Mode

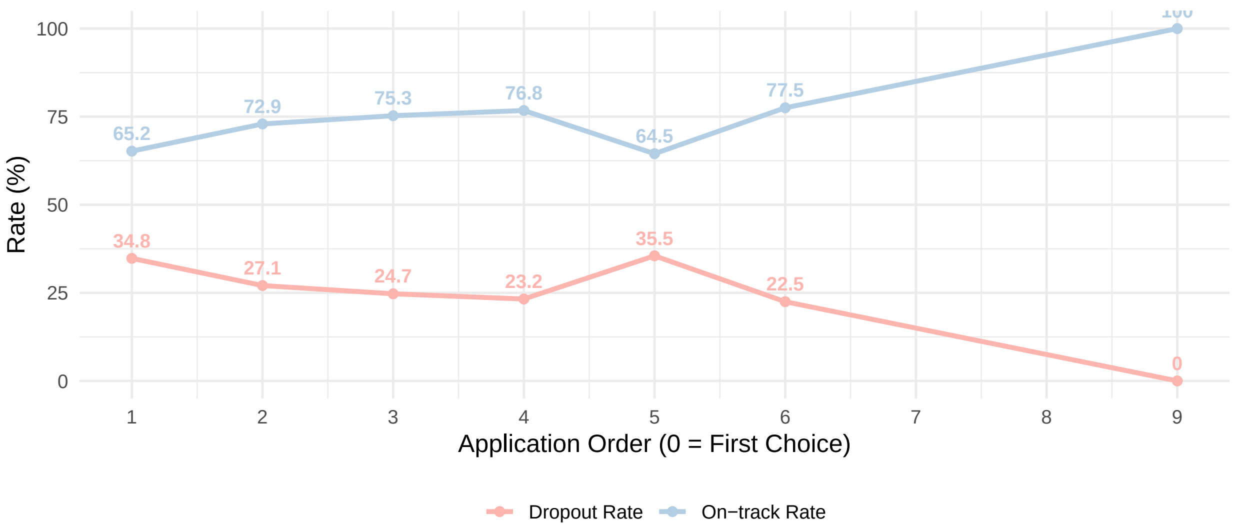

Figure 12: On-track and Dropout Rates by Application Order

inherently more challenging, either academically or in terms of required skills, leading to higher dropout rates.

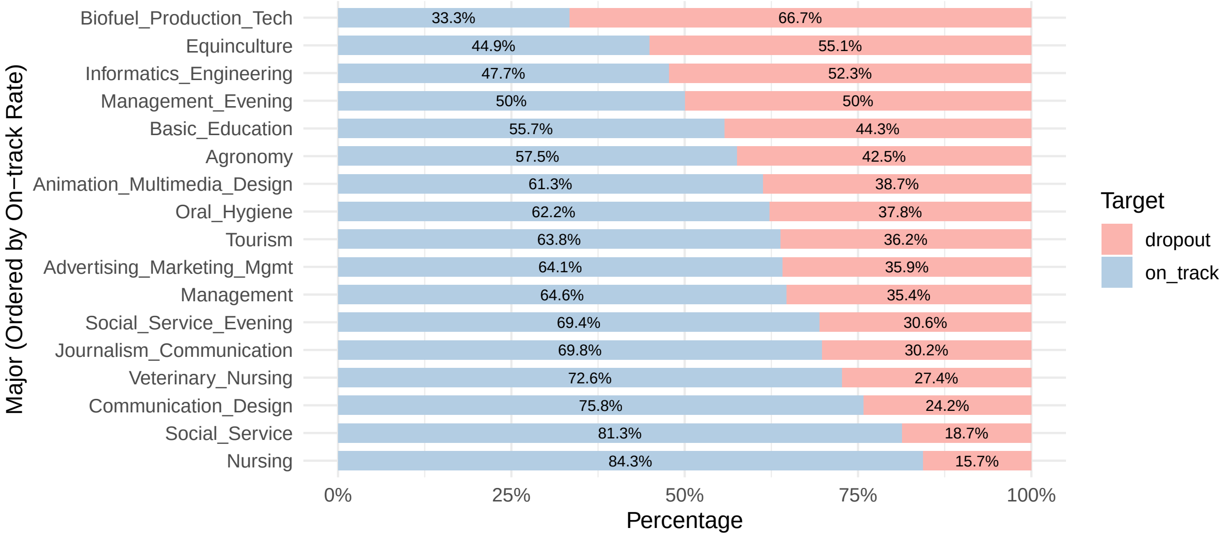

Figure 13: Target Status Breakdown by Major (Ordered by Graduation Rate), fig.height = 8

### 2.4.2 Daytime/Evening Attendance

Figure 14 shows the breakdown of student outcomes by attendance type. For daytime students, 69% remain on track while 31% drop out. In contrast, evening students have a lower on-track rate at 58.6% and a higher dropout rate at 41.4%. This visual shows that students attending in the evening are more likely to drop out compared to those attending during the day.

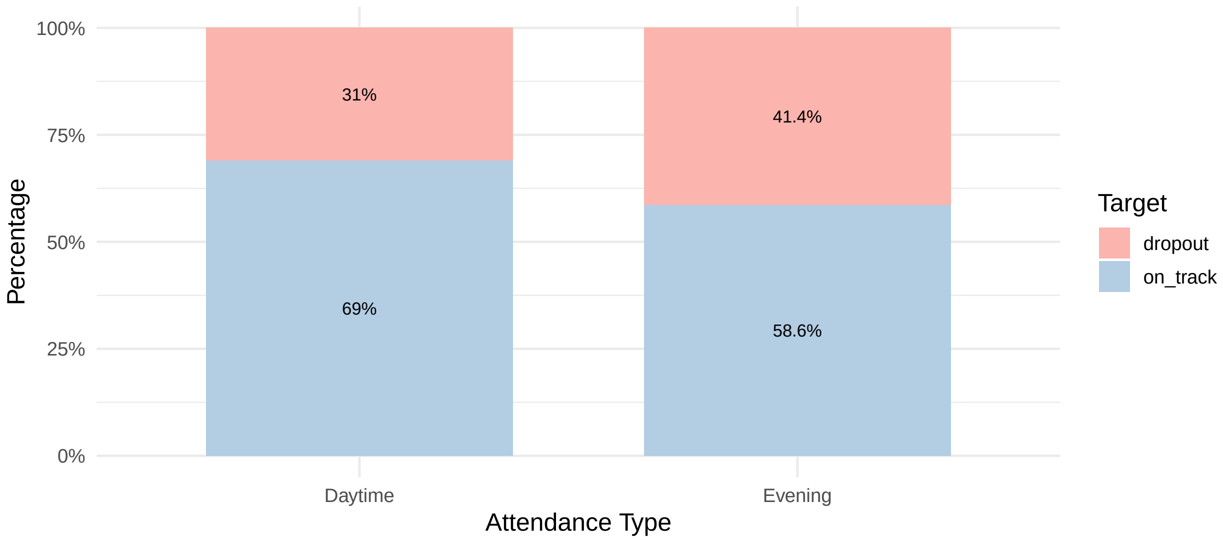

Figure 14: Target Status Breakdown by Attendance Type

# 3 Methods

## 3.1 Data Partition

To ensure robust model development and unbiased performance evaluation, the dataset was partitioned into distinct training, validation, and final hold-out sets. The `createDataPartition` function from the `caret` package was used to randomly assign 90% of the student records to the training/validation set, with the remaining 10% reserved as a final hold-out set for assessing generalizability. To avoid issues arising from unseen categorical values, the final hold-out set was filtered so that all levels of key variables, such as major, application mode, marital status, attendance type, parental education, special needs, and international status, matched those present in the training/validation set. This was accomplished using a series of `semi_join` operations, which restricted the hold-out set to only those students whose attributes were already represented in the training data.

Any observations in the hold-out split with novel or unmatched categories were returned to the training/validation set using `anti_join` and `rbind`, thereby maximizing the data available for model fitting while maintaining evaluation integrity. The training/validation set was further split into internal training and validation subsets to support parameter tuning and model selection. Throughout this process, care was taken to ensure that the model was always evaluated on student cases for which all relevant contextual information was available in the training data, thereby preventing data leakage and ensuring a reliable estimate of real-world model performance.

## 3.2 Model Building

### 3.2.1 Random

To begin our modeling process, we trained a Random Forest classifier using the training subset of our data. The Random Forest algorithm is a powerful ensemble method that constructs multiple decision trees and aggregates their predictions for improved accuracy and robustness. We used 500 trees and applied cross-validation to optimize model performance and prevent overfitting. After training, the model's predictions were evaluated on a hold-out validation set, with an accuracy of 0.8211587.

```
##  Accuracy
## 0.8211587
```

```
##           Reference
## Prediction dropout on_track
##   dropout       76       20
##   on_track      51      250
```

### 3.2.2 KNN

Next, we implemented the k-Nearest Neighbors (KNN) algorithm, which classifies each student based on the majority outcome of its closest neighbors in the feature space. The optimal value of k was determined via cross-validation within the caret package. KNN's performance was also assessed using the validation set, resulting in an accuracy of 0.7178841.

```
##  Accuracy
## 0.7178841
```

```
##           Reference
## Prediction dropout on_track
##   dropout       35       20
##   on_track      92      250
```

### 3.2.3 Naive Bayes

We then built a Naive Bayes classifier, a probabilistic model based on Bayes' Theorem and the assumption of feature independence. This simple yet effective model is often competitive for classification tasks involving categorical data. The model's accuracy on the validation set was 0.7329975.

```
##  Accuracy
## 0.7329975
```

```
##           Reference
## Prediction dropout on_track
##   dropout       85       64
##   on_track      42      206
```

### 3.2.4 Gradient Boosted Trees (xgboost)

Additionally, we trained a Gradient Boosted Trees model using the xgboost implementation, which builds an ensemble of trees sequentially, each one learning from the errors of the previous. This model is known for its high predictive power, especially with tabular data. Cross-validation was used to select the best hyperparameters. The validation accuracy for this model was 0.8161209.

```
##  Accuracy
## 0.8161209
```

```
##           Reference
## Prediction dropout on_track
##   dropout       78       24
##   on_track      49      246
```

### 3.2.5 Ensemble (Majority Vote)

To leverage the strengths of individual models, we constructed an ensemble model using a majority voting scheme. The predictions from Random Forest, KNN, Naive Bayes, and Gradient Boosted Trees were combined, and the final prediction was assigned based on the majority vote. This approach yielded an ensemble accuracy of 0.8136020.

```
## [1] 0.813602
```

### 3.2.6 Ensemble for the 2nd try: omit KNN model

Upon reviewing individual model performance, we noticed that the KNN model had relatively lower accuracy, which negatively impacted the overall ensemble. Therefore, we constructed a second ensemble model using only Random Forest, Naive Bayes, and Gradient Boosted Trees. This refined ensemble demonstrated improved performance, achieving an accuracy of 0.8211587.

```
##                    Model  Accuracy
## 1          Random Forest 0.8211587
## 2                    KNN 0.7178841
## 3            Naive Bayes 0.7329975
## 4 Gradient Boosted Trees 0.8161209
## 5    Ensemble (4 models) 0.8136020
## 6    Ensemble (3 models) 0.8211587
```

# 4 Result

## 4.1 Model Result

To evaluate the final models, we applied each to the independent validation set and compared their predictive accuracy. The following table summarizes the performance of all models and ensemble combinations.

Though both the four-model and three-model ensemble approaches achieved the highest overall accuracy (0.8684807), the three-model ensemble (exluding KNN) is selected as the final model. This decision was based on the consistently lower performance of the KNN model (accuracy of 0.7823129) relative to the other models. Including KNN in the ensemble did not yield any improvement in predictive performance and could introduce unnecessary complexity or potential instability.

```
##                     Model  Accuracy
## 1           Random Forest 0.8639456
## 2                     KNN 0.7823129
## 3             Naive Bayes 0.8004535
## 4 Gradient Boosted Trees 0.8662132
## 5    Ensemble (4 models) 0.8684807
## 6    Ensemble (3 models) 0.8684807
```

### 4.2 Model Performance

To further evaluate the model's effectiveness, we examined the confusion matrix for the three-model ensemble on the validation set below.The ensemble model correctly identifies about 71.8% of students who will drop out and 94.0% of those who will stay on track. Such performance indicates the ensemble is effective both at detecting at-risk students and minimizing false alarms, supporting its use for targeted intervention strategies.

```
##            Reference
## Prediction dropout on_track
##   dropout      102       18
##   on_track      40      281
```

# 5 Conclusion

## 5.1 Summary

This project explored predictive modeling for student dropout and academic success using a comprehensive, institution-level dataset from the UCI Machine Learning Repository. The analysis included careful data cleaning and feature engineering, followed by systematic model building using Random Forest, k-Nearest Neighbors, Naive Bayes, and Gradient Boosting. Model performance was assessed using a final hold-out validation set, with Random Forest and Gradient Boosting achieving the highest individual accuracies (0.8639 and 0.8662 respectively). In addition, ensemble approaches combining Random Forest, Gradient Boosting, and Naive Bayes were evaluated and demonstrated superior performance compared to the best single models. Attempts to include all four models in the ensemble did not improve overall accuracy, primarily due to the relatively low performance of the k-Nearest Neighbors model. As a result, excluding k-Nearest Neighbors from the ensemble is chosen as the optimal model

## 5.2 Potential Impact

The potential impact of this work lies in its ability to help educational institutions proactively identify and support students who may be at risk of dropping out, thereby improving both student outcomes and institutional retention rates. By leveraging multiple data sources such as macroeconomic context, demographics, academic history, and program characteristics, the models offer a holistic approach to early warning and intervention.

## 5.3 Limitations

Several limitations of this analysis should be acknowledged. First, while the validation set provides useful insights into model performance, its relatively modest size (441 records) may not fully reflect the diversity of student experiences within the broader population. Additionally, the target variable is imbalanced, with a considerably larger proportion of “on_track” cases (3,003) compared to “dropout” cases (1,421). This imbalance may lead the models to favor the majority class, potentially resulting in an underestimation of dropout risk. Furthermore, the dataset is derived from a single institution, which may limit the generalizability of the findings to other educational contexts. Additional research using data from multiple institutions would be necessary to confirm the robustness and broader applicability of these results.

## 5.4 Future Work

Future work could address these limitations by collecting more extensive and diverse datasets, applying advanced resampling or class-weighting techniques to address imbalance, and exploring other machine learning approaches such as matrix factorization or neural networks. Collaborating with institutional stakeholders to incorporate additional features such as student engagement data, financial information, or qualitative feedback could further enhance predictive performance and actionable insights for student success initiatives.