Predicting students' dropout and academic success

BIG DATA AND DATABASES
GROUP PROJECT

Amedeo Federica 3160248
Benedetti Rebecca 3152342
Bonanno Susanna 3159805
Caputi Giulio 3153584
Hasbani Isaac Ethan 3153080

Report outline

1	2	3	4	5	6	7
Introduction	Data description	Data preparation	Data description	Data modeling	Outcomes	Managerial implications
	• UNIVARIATE ANALYSIS	• MISSING VALUES • OUTLIERS • VARIABLE ENCODING • NEW FEATURE CREATION • FEATURE SELECTION	• BIVARIATE ANALYSIS • SECOND FEATURE SELECTION	• LOGISTIC REGRESSION • DECISION TREE • RANDOM FOREST • GRADIENT BOOSTING	• RESULTS OF OUR MODELS	• IMPLICATIONS FOR UNIVERSITIES • IMPLICATIONS FOR TUTORING COMPANIESI

Introduction

Overview

Our analysis is based on a dataset containing information on 4425 students enrolled in various undergraduate degree programs in Portuguese universities. Each entry includes 37 features, detailing both the socio-economic status of the students and their academic performance.

We aim to use this data for predicting students' academic outcomes. Specifically, our focus is on predicting the main dependent variable (referred to as "Target"), which represents the status of each student at the end of the standard duration of the undergraduate program. This variable can take three different values depending on whether the student has successfully graduated, dropped out of the course, or is still enrolled.

Our investigation has several benefits:

- First and foremost, by identifying crucial predictors of academic success, our model can suggest to students which aspects to prioritize, enhancing their likelihood of success. Additionally, these models can detect individuals at risk of dropping out before they do so.

- Secondly, from a managerial point of view, being aware of at-risk students is beneficial for companies providing tutoring or other kinds of academic help, as targeting those people with ads is likely to result in an increase in customers for these businesses.

- Thirdly, and quite clearly, identifying at-risk students is relevant for universities themselves, as this empowers them to take preventive measures to help struggling students.

Data description

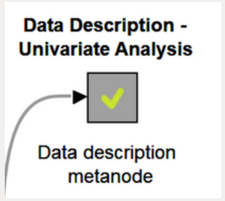
UNIVARIATE ANALYSIS

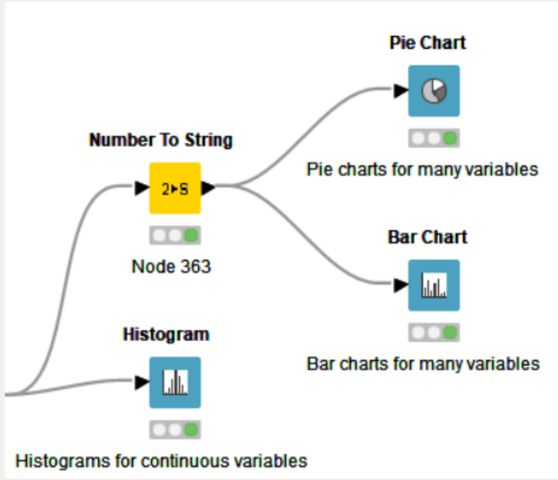
Overview of univariate analysis

In this section, we offer a brief description of each variable along with the key insights that can be extracted from each one of them. Additionally, we provide a distribution plot for each variable in our sample.

We initially convert numbers to strings using a *Number to String* node. For categorical variables with a limited number of categories, we display a pie chart as it effectively illustrates the distribution. However, for categorical variables with numerous classes, we choose to use a bar chart.

As for numerical variables, both continuous and discrete, we represented their distributions using histograms, employing 16 bins for each variable.





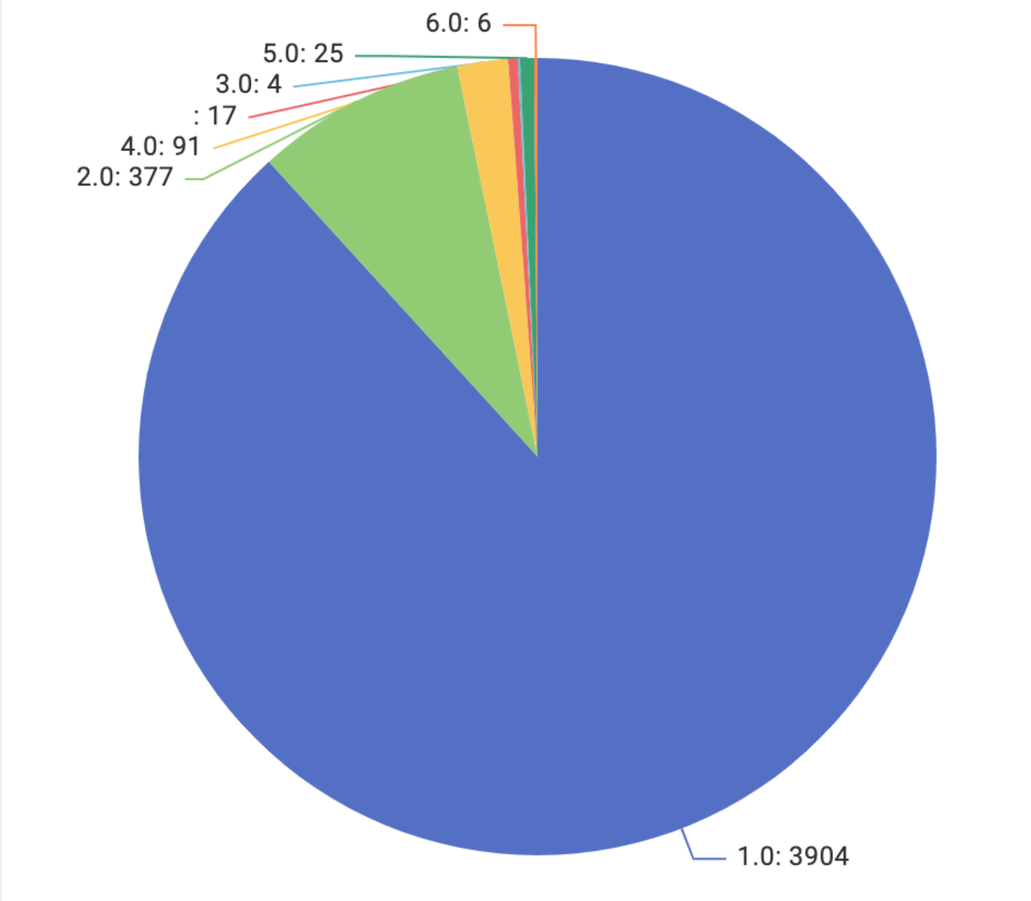
Marital status

Nature: Categorical, nominal

Description: the marital status of the students

Categories: 1-single; 2-married; 3-widower; 4-divorced; 5-facto union; 6-legally separated

Insights: Most of the students (88,59%) are single



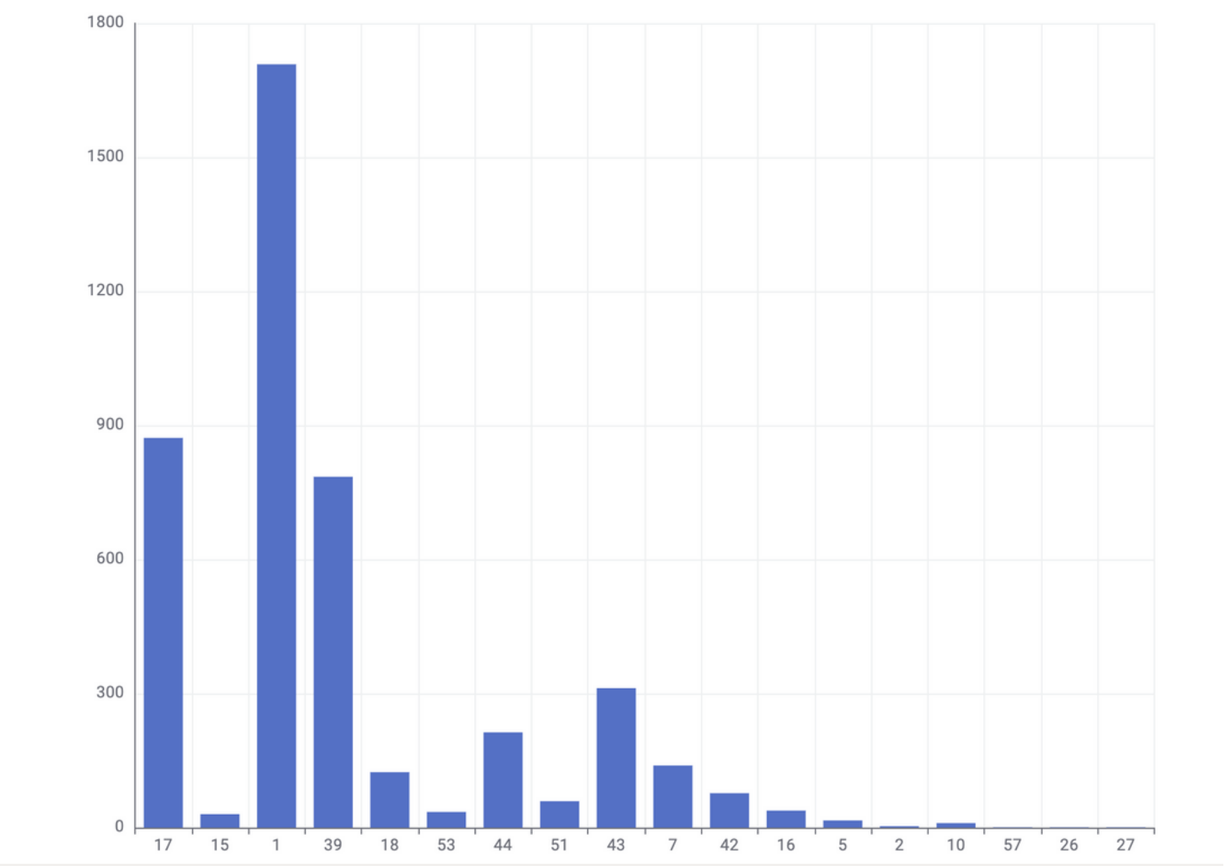
Application mode

Nature: Categorical, nominal

Description: the application mode of the students

Categories: 1 - 1st phase - general contingent; 2 - Ordinance No. 612/93; 5 - 1st phase - special contingent (Azores Island); 7 - Holders of other higher courses; 10 - Ordinance No. 854-B/99; 15 - International student (bachelor); 16 - 1st phase - special contingent (Madeira Island); 17 - 2nd phase - general contingent; 18 - 3rd phase - general contingent; 26 - Ordinance No. 533-A/99, item b2) (Different Plan); 27 - Ordinance No. 533-A/99, item b3 (Other Institution); 39 - Over 23 years old; 42 – Transfer; 43 - Change of course; 44 - Technological specialization diploma holders; 51 - Change of institution/course; 53 - Short cycle diploma holders; 57 - Change of institution/course (International)

Insights: Many students (1708, or 38.6%) applied in the 1st phase – general contingent, whereas the number of the students who applied in the second phase – general contingent is 872 (roughly 20%)



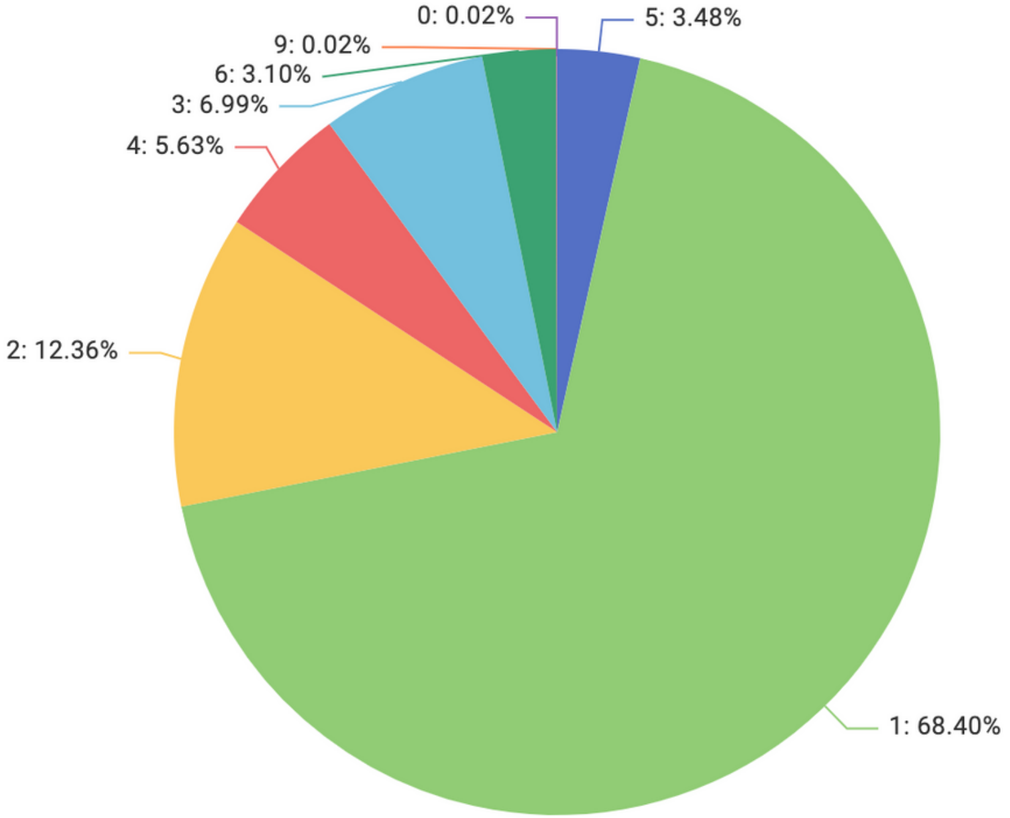
Application order

Nature: Categorical, nominal

Description: the ordinal ranking of students' preferences for courses

Categories: between 0 - first choice; and 9 - last choice

Insights: Most of the students (3026, or 68.4%) have been accepted to their second choice, while only one was admitted to its first choice, and still only one was admitted to its last choice



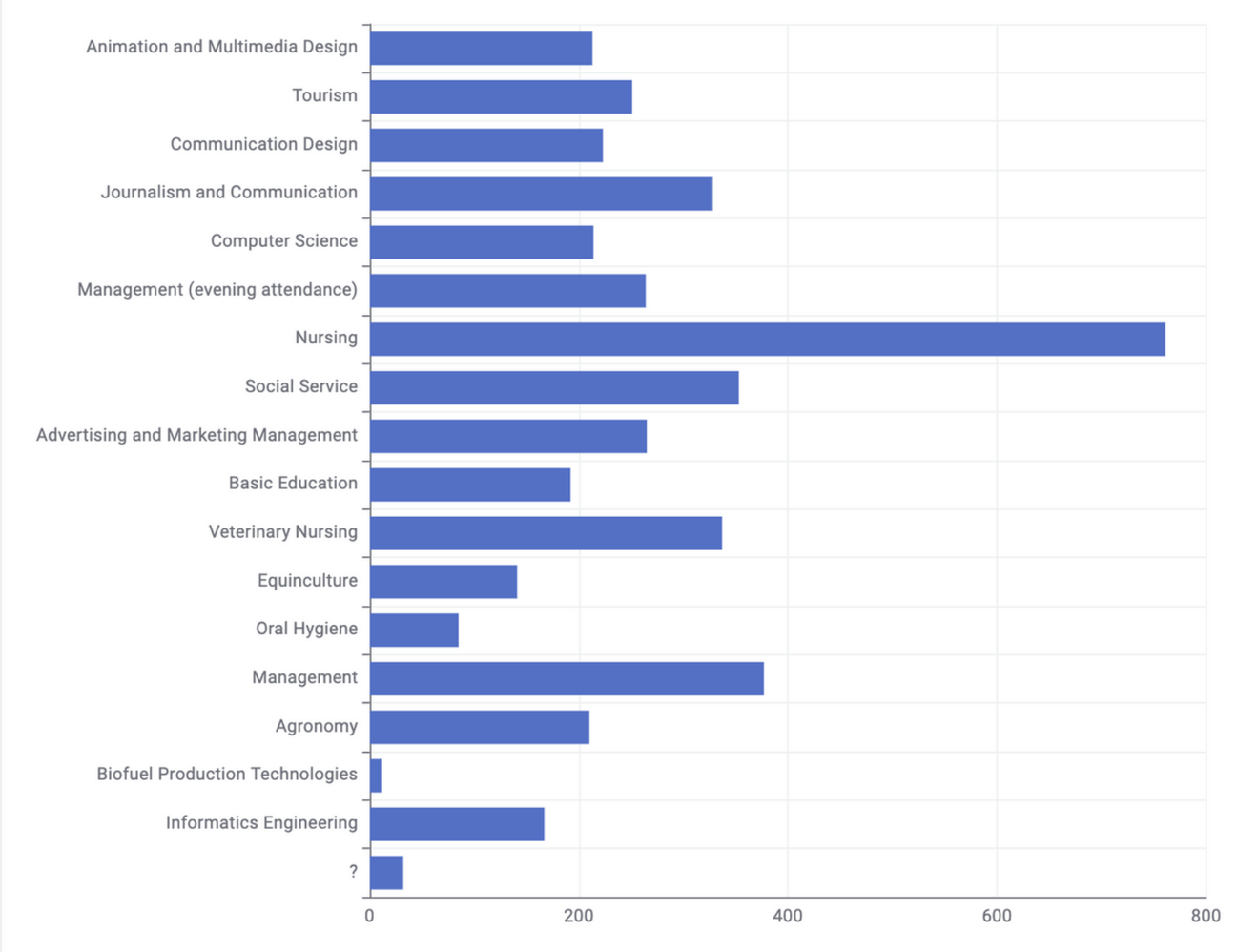
Course

Nature: Categorical, nominal

Description: the course attended by the students

Categories: Biofuel Production Technologies, Animation and Multimedia Design, Social Service, Agronomy, Communication Design, Veterinary Nursing, Informatics Engineering, Equinculture, Management, Social Service, Tourism, Nursing, Oral Hygiene, Advertising and Marketing Management, Journalism and Communication, Basic Education, Management (evening attendance)

Insights: The most popular course is Nursing, attended by 17.2% of students



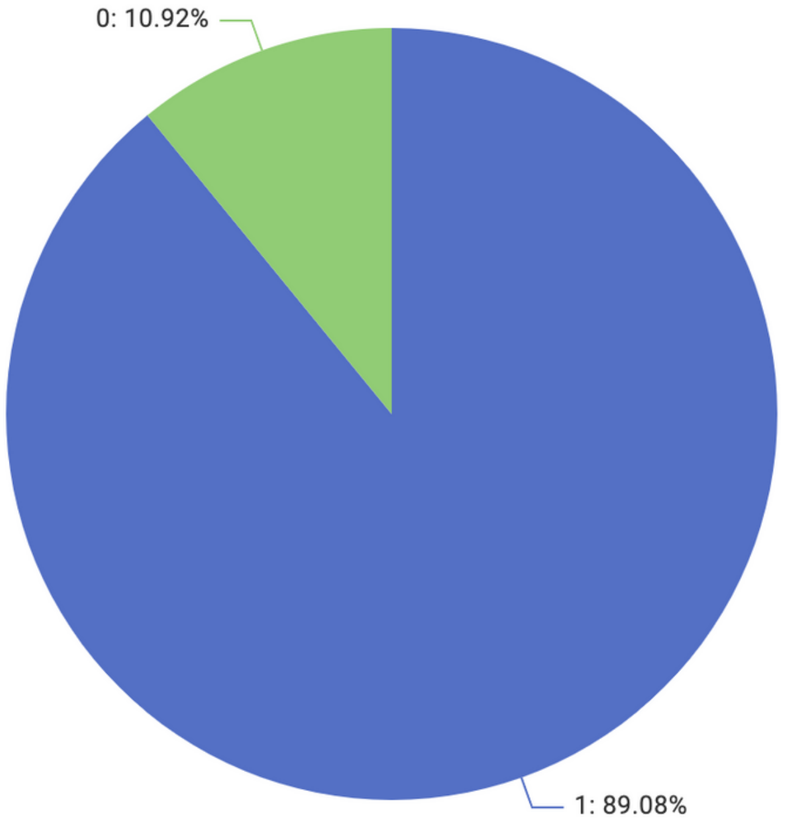
Daytime/evening attendance

Nature: Categorical, nominal

Description: whether the students attended a course during the daytime or in the evening.

Categories: 1 – daytime; 0 - evening

Insights: Most of the students (89.08%) attend classes during the day



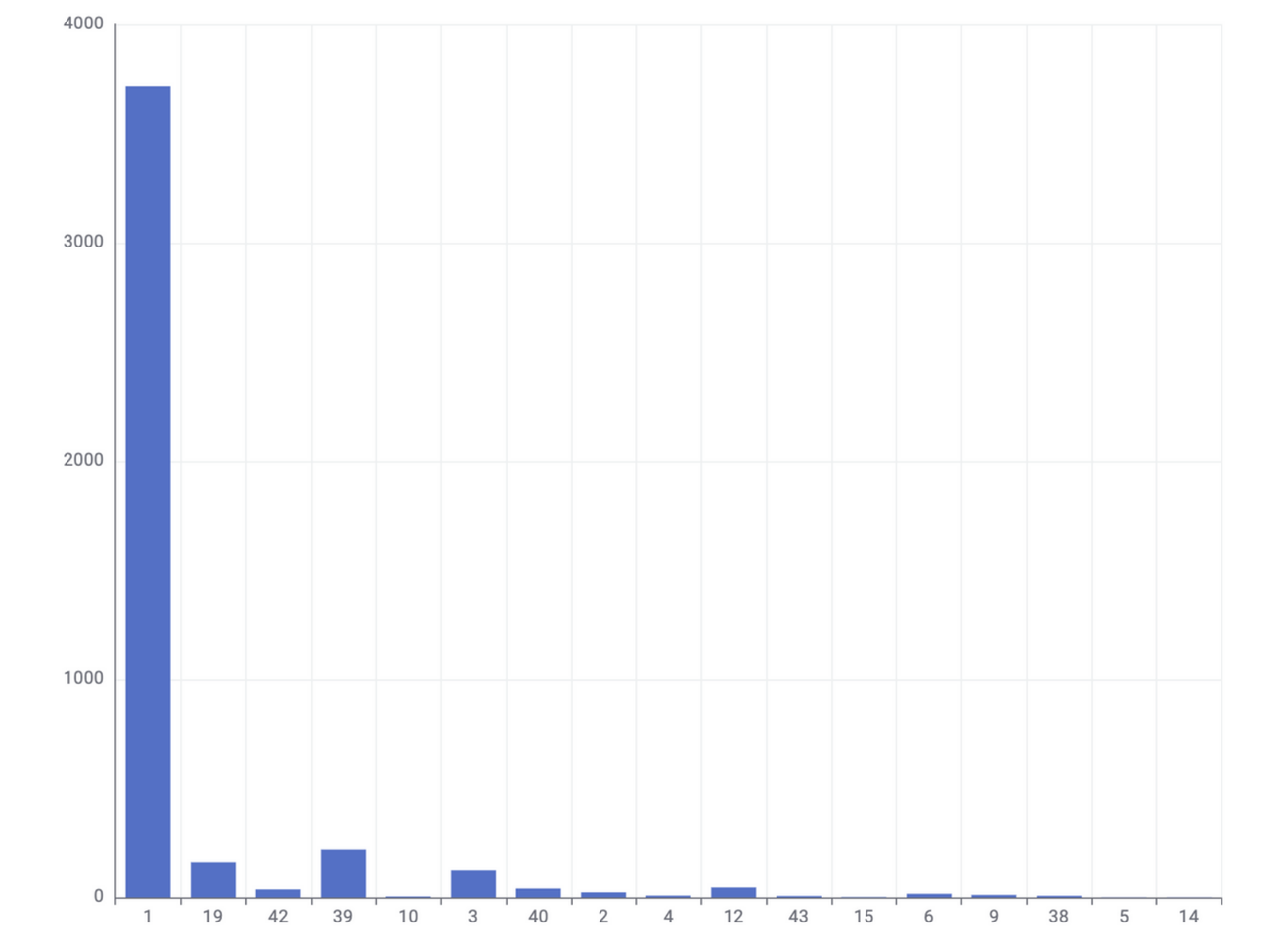
Previous qualification

Nature: Categorical, nominal

Description: the previous qualification (education level) of the students

Categories: 1 - Secondary education; 2 - Higher education - bachelor's degree; 3 - Higher education – degree; 4 - Higher education - master's; 5 - Higher education – doctorate; 6 - Frequency of higher education; 9 - 12th year of schooling - not completed; 10 - 11th year of schooling - not completed; 12 - Other - 11th year of schooling; 14 - 10th year of schooling; 15 - 10th year of schooling - not completed; 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv.; 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv.; 39 - Technological specialization course; 40 - Higher education - degree (1st cycle); 42 - Professional higher technical course; 43 - Higher education - master (2nd cycle)

I**nsights**: Most of the students (3717, or 84%) have a secondary education.

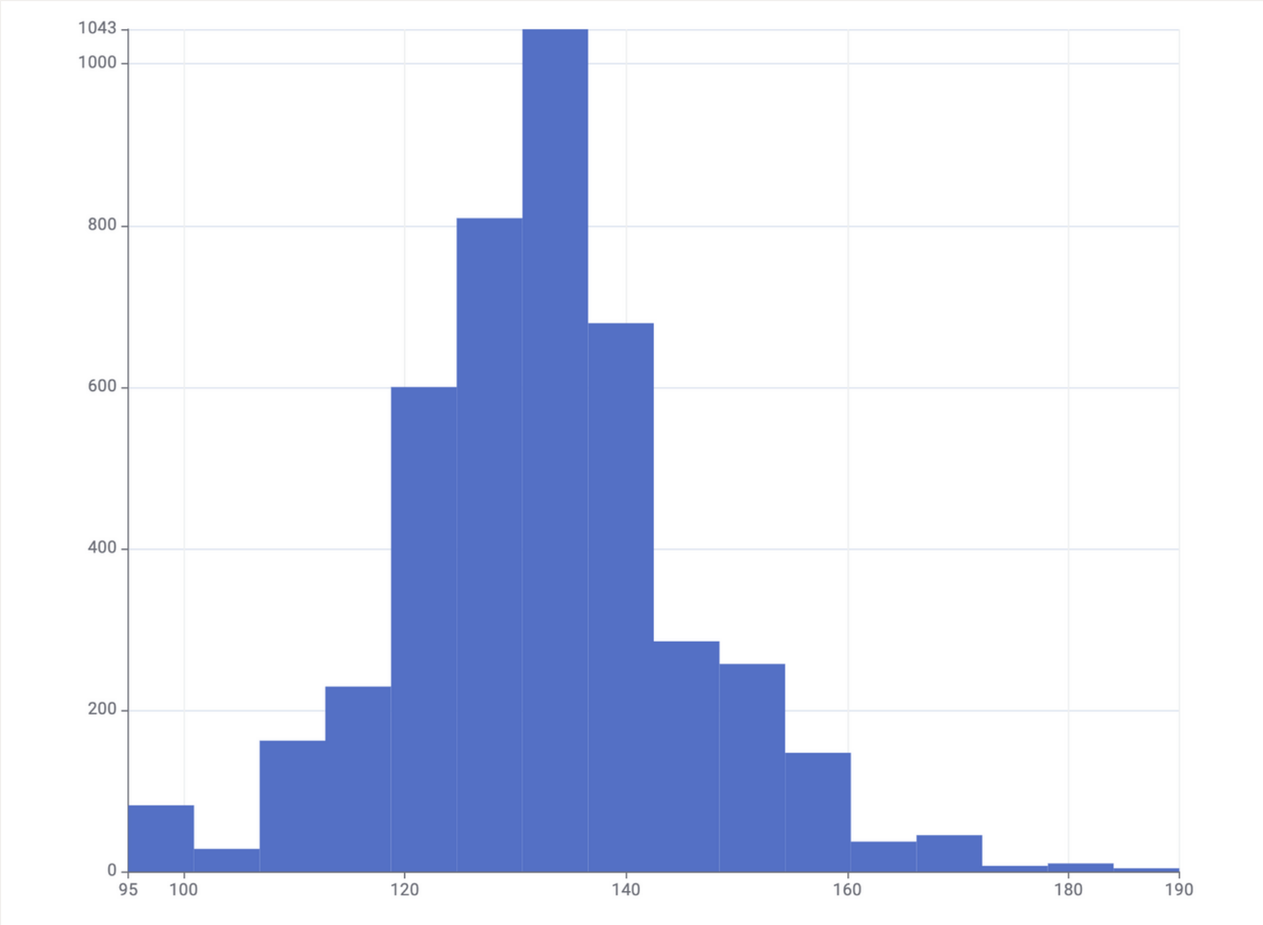


Previous qualification (grade)

Nature: Numerical, continuous

Description: the grade of the previous qualification of the students
Range: between 0 and 200

Insights: Many students (1444, or 32.6%) obtained a grade between 133 and 143 in their previous qualification.

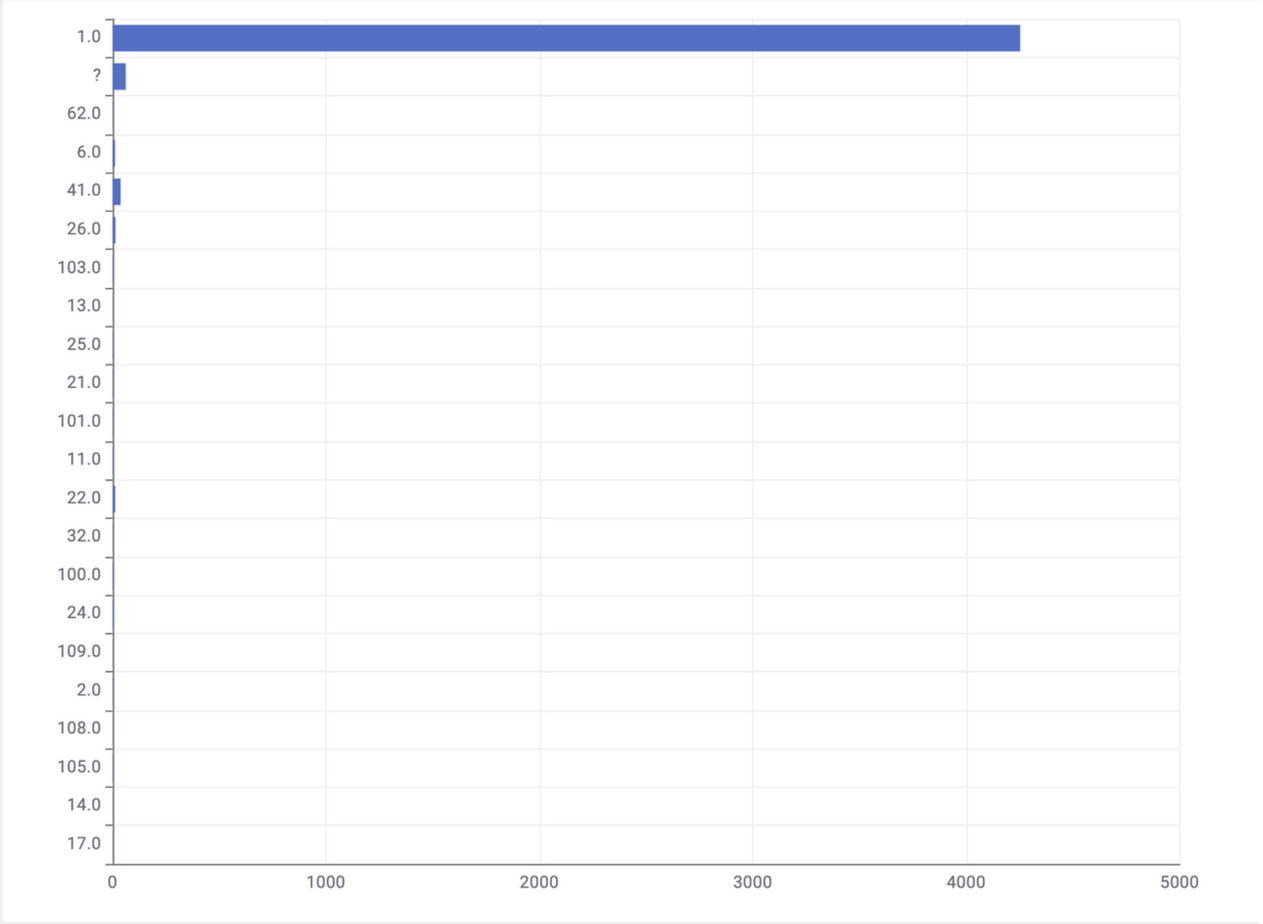


Nationality

Nature: Categorical, nominal

Description: nationality of the students
Categories: 1 - Portuguese; 2 - German; 6 - Spanish; 11 - Italian; 13 - Dutch; 14 - English; 17 - Lithuanian; 21 - Angolan; 22 - Cape Verdean; 24 - Guinean; 25 - Mozambican; 26 - Santomean; 32 - Turkish; 41 - Brazilian; 62 - Romanian; 100 - Moldovan; 101 - Mexican; 103 - Ukrainian; 105 - Russian; 108 - Cuban; 109 - Colombian

Insights: The vast majority ofstudents (96,13%) are Portuguese.

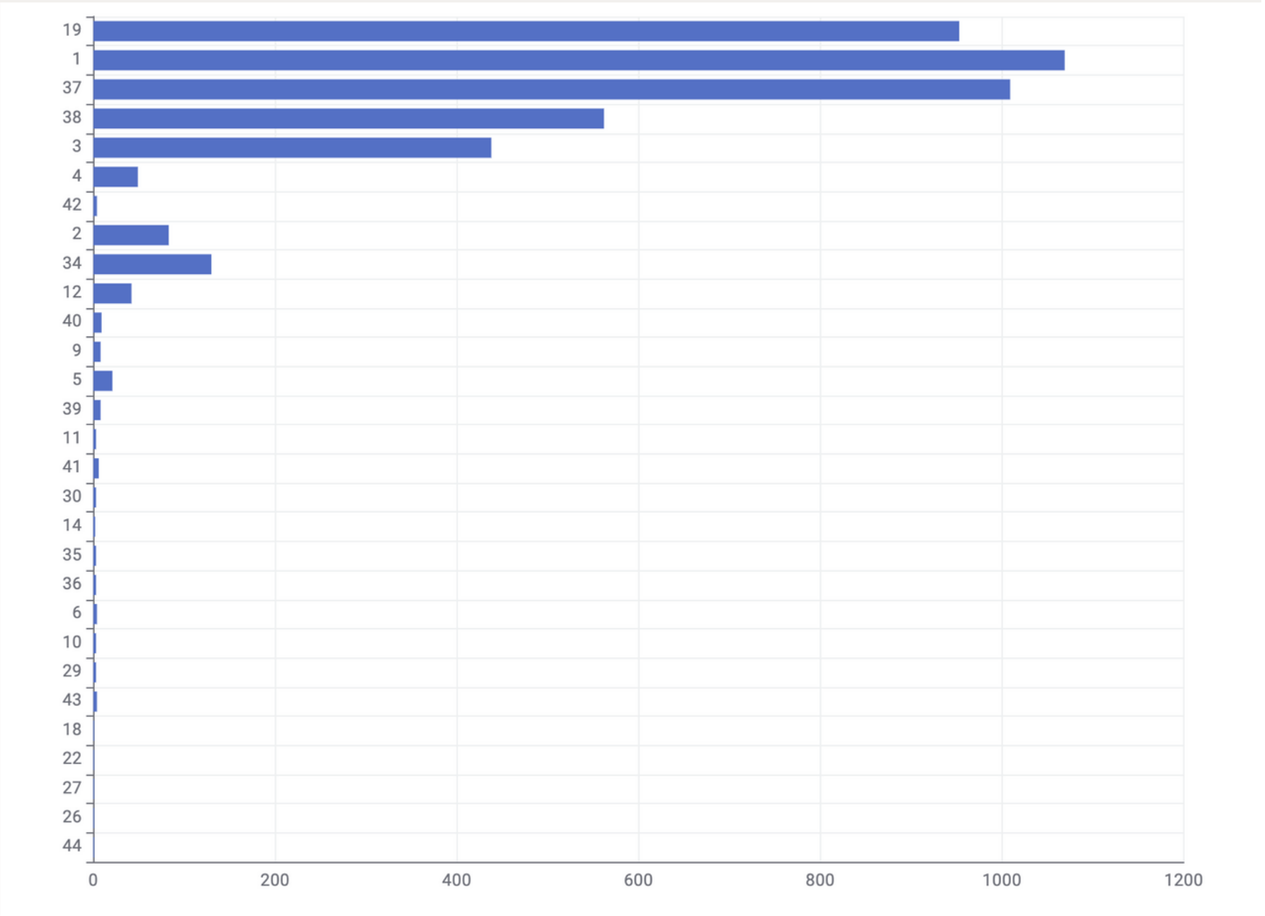


Mother’s qualification

Nature: Categorical, nominal

Description: qualification of the students' mothers

Categories: 1 - Secondary Education - 12th Year of Schooling or Eq.; 2 - Higher Education - Bachelor's Degree; 3 - Higher Education – Degree; 4 - Higher Education - Master's; 5 - Higher Education – Doctorate; 6 - Frequency of Higher Education; 9 - 12th Year of Schooling - Not Completed; 10 - 11th Year of Schooling - Not Completed; 11 - 7th Year (Old); 12 - Other - 11th Year of Schooling; 14 - 10th Year of Schooling; 18 - General commerce course; 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.; 22 - Technical-professional course; 26 - 7th year of schooling; 27 - 2nd cycle of the general high school course; 29 - 9th Year of Schooling - Not Completed; 30 - 8th year of schooling; 34 – Unknown; 35 - Can't read or write; 36 - Can read without having a 4th year of schooling; 37 - Basic education 1st cycle (4th/5th year) or equiv.; 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.; 39 - Technological specialization course; 40 - Higher education - degree (1st cycle); 41 - Specialized higher studies course; 42 - Professional higher technical course; 43 - Higher Education - Master (2nd cycle); 44 - Higher Education - Doctorate (3rd cycle)



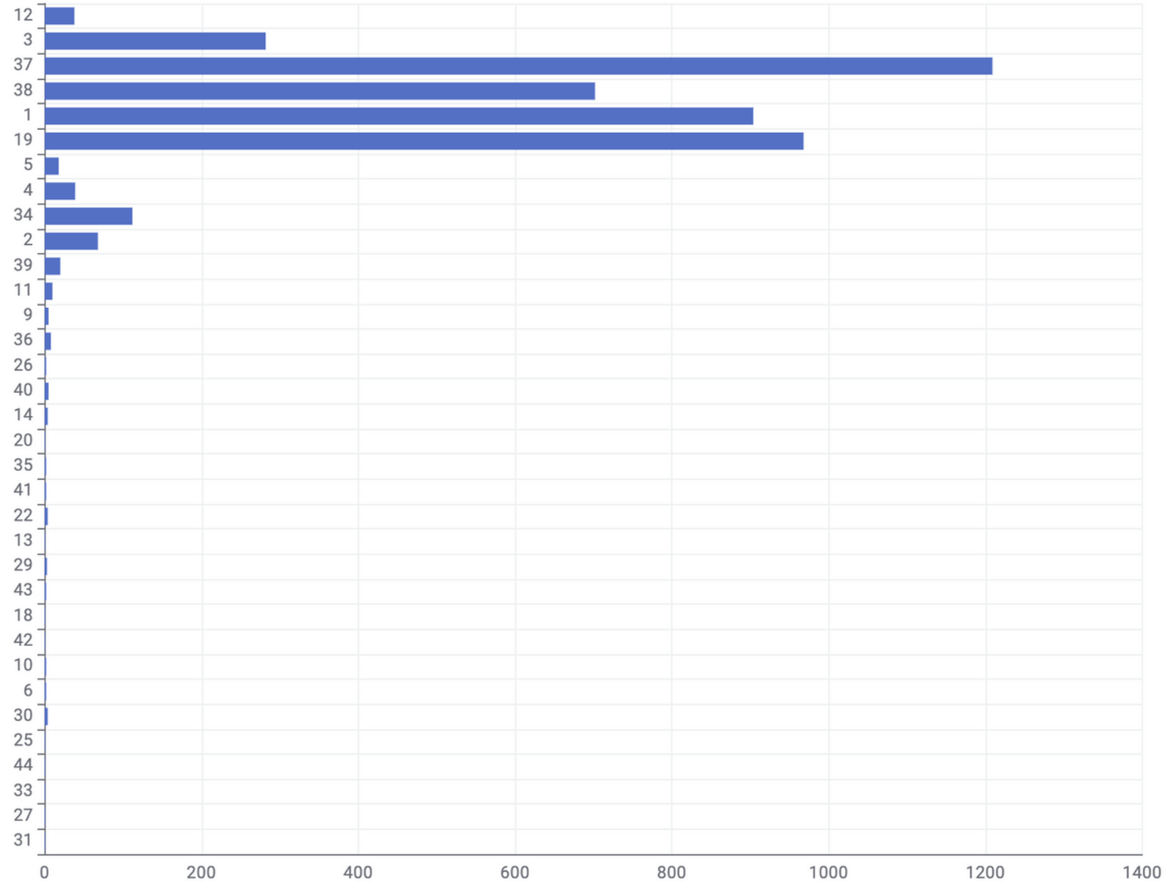
Insights: In many cases (1069, or 24.1%) the students’ mothers have a secondary education. Many others (953) have a basic education 3rd cycle or a basic education 1st cycle (1009).

Father’s qualification

Nature: Categorical, nominal

Description: the qualification of the students' fathers

Categories: 1 - Secondary Education - 12th Year of Schooling or Eq.; 2 - Higher Education - Bachelor's Degree; 3 - Higher Education – Degree; 4 - Higher Education - Master's; 5 - Higher Education – Doctorate; 6 - Frequency of Higher Education; 9 - 12th Year of Schooling - Not Completed; 10 - 11th Year of Schooling - Not Completed; 11 - 7th Year (Old); 12 - Other - 11th Year of Schooling; 13 - 2nd year complementary high school course; 14 - 10th Year of Schooling; 18 - General commerce course; 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.; 20 - Complementary High School Course; 22 - Technical-professional course; 25 - Complementary High School Course - not concluded; 26 - 7th year of schooling; 27 - 2nd cycle of the general high school course; 29 - 9th Year of Schooling - Not Completed; 30 - 8th year of schooling; 31 - General Course of Administration and Commerce; 33 - Supplementary Accounting and Administration; 34 – Unknown; 35 - Can't read or write; 36 - Can read without having a 4th year of schooling; 37 - Basic education 1st cycle (4th/5th year) or equiv.; 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.; 39 - Technological specialization course; 40 - Higher education - degree (1st cycle); 41 - Specialized higher studies course; 42 - Professional higher technical course; 43 - Higher Education - Master (2nd cycle); 44 - Higher Education - Doctorate (3rd cycle)



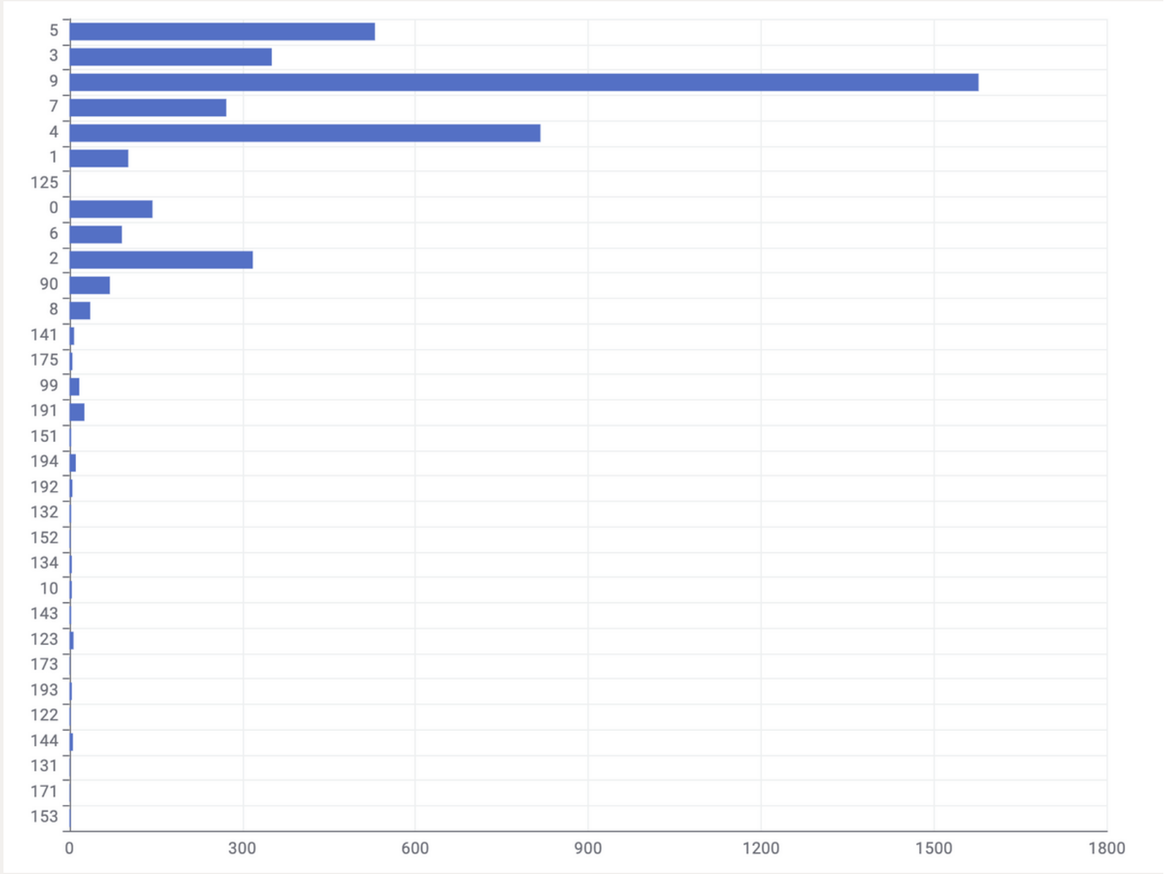
Insights: In many cases (1209, or 27.3%) the students’ fathers have a basic education 1st cycle. Many others have a secondary education (904) or a basic education 3rd cycle (968).

Mother’s occupation

Nature: Categorical, nominal

Description: the occupation of the students' mothers

Categories: 0 – Student; 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers; 2 - Specialists in Intellectual and Scientific Activities; 3 - Intermediate Level Technicians and Professions; 4 - Administrative staff; 5 - Personal Services, Security and Safety Workers and Sellers; 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry; 7 - Skilled Workers in Industry, Construction and Craftsmen; 8 - Installation and Machine Operators and Assembly Workers; 9 - Unskilled Workers; 10 - Armed Forces Professions; 90 - Other Situation; 99 - (blank); 122 - Health professionals; 123 – teachers; 125 - Specialists in information and communication technologies (ICT); 131 - Intermediate level science and engineering technicians and professions; 132 - Technicians and professionals, of intermediate level of health; 134 - Intermediate level technicians from legal, social, sports, cultural and similar services; 141 - Office workers, secretaries in general and data processing operators; 143 - Data, accounting, statistical, financial services and registry-related operators; 144 - Other administrative support staff; 151 - personal service workers; 152 – sellers; 153 - Personal care workers and the like; 171 - Skilled construction workers and the like, except electricians; 173 - Skilled workers in printing, precision instrument manufacturing, jewelers, artisans and the like; 175 - Workers in food processing, woodworking, clothing and other industries and crafts; 191 - cleaning workers; 192 - Unskilled workers in agriculture, animal production, fisheries and forestry; 193 - Unskilled workers in extractive industry, construction, manufacturing and transport; 194 - Meal preparation assistants



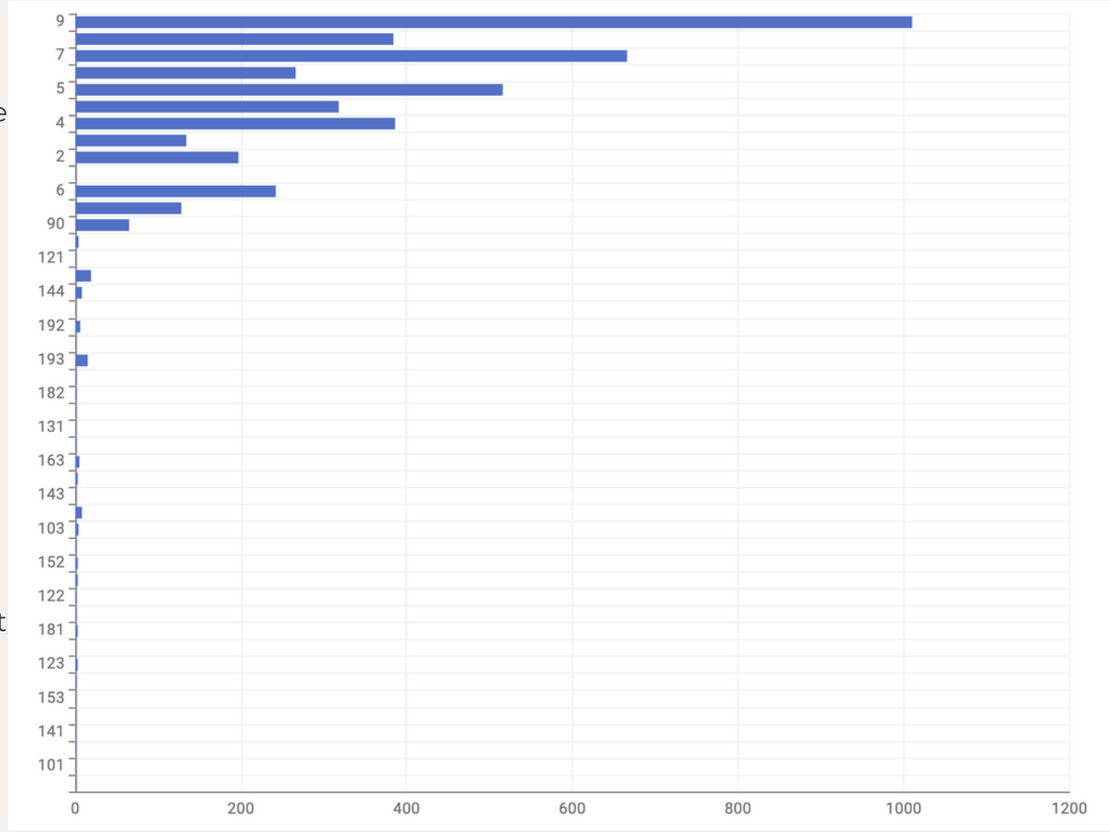
Insights: Many of the students' mothers (1577, or 35.6%) are unskilled workers, while many others (817, or 18.5%) are administrative staff members

Father’s occupation

Nature: Categorical, nominal

Description: the occupation of the students' fathers

Categories: 0 – Student; 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers; 2 - Specialists in Intellectual and Scientific Activities; 3 - Intermediate Level Technicians and Professions; 4 - Administrative staff; 5 - Personal Services, Security and Safety Workers and Sellers; 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry; 7 - Skilled Workers in Industry, Construction and Craftsmen; 8 - Installation and Machine Operators and Assembly Workers; 9 - Unskilled Workers; 10 - Armed Forces Professions; 90 - Other Situation; 99 - (blank); 101 - Armed Forces Officers; 102 - Armed Forces Sergeants; 103 - Other Armed Forces personnel; 112 - Directors of administrative and commercial services; 114 - Hotel, catering, trade and other services directors; 121 - Specialists in the physical sciences, mathematics, engineering and related techniques; 122 - Health professionals; 123 – teachers; 124 - Specialists in finance, accounting, administrative organization, public and commercial relations; 131 - Intermediate level science and engineering technicians and professions; 132 - Technicians and professionals, of intermediate level of health; 134 - Intermediate level technicians from legal, social, sports, cultural and similar services; 135 - Information and communication technology technicians; 141 - Office workers, secretaries in general and data processing operators; 143 - Data, accounting, statistical, financial services and registry-related operators; 144 - Other administrative support staff; 151 - personal service workers; 152 – sellers; 153 - Personal care workers and the like; 154 - Protection and security services personnel; 161 - Market-oriented farmers and skilled agricultural and animal production workers; 163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence; 171 - Skilled construction workers and the like, except electricians; 172 - Skilled workers in metallurgy, metalworking and similar; 174 - Skilled workers in electricity and electronics; 175 - Workers in food processing, woodworking, clothing and other industries and crafts; 181 - Fixed plant and machine operators; 182 - assembly workers; 183 - Vehicle drivers and mobile equipment operators; 192 - Unskilled workers in agriculture, animal production, fisheries and forestry; 193 - Unskilled workers in extractive industry, construction, manufacturing and transport; 194 - Meal preparation assistants; 195 - Street vendors (except food) and street service providers



Insights: Many of the students’ fathers (1010, or 22.8%) are unskilled workers. Many others are skilled workers in Industry, Construction and Crafts (15%) or Security and Safety Workers and Sellers (11.7%). A few are Directors and Executive Managers (3%).

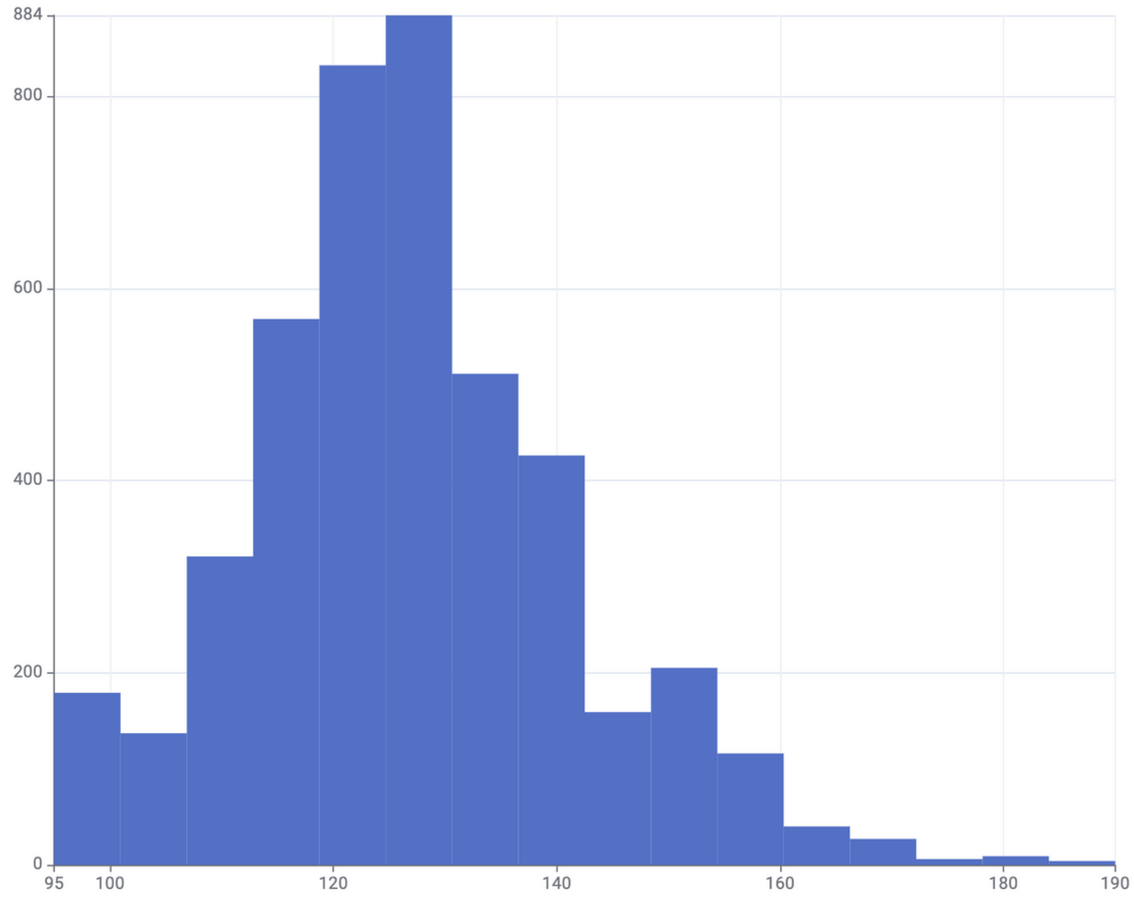
Admission grade

Nature: Numerical, continuous

Description: the admission grade of the students

Range: between 0 and 200

Insights: Many students (882, or 20%) got admitted with a grade between 127 and 133. Many other students got admitted with a grade between 114 and 126.



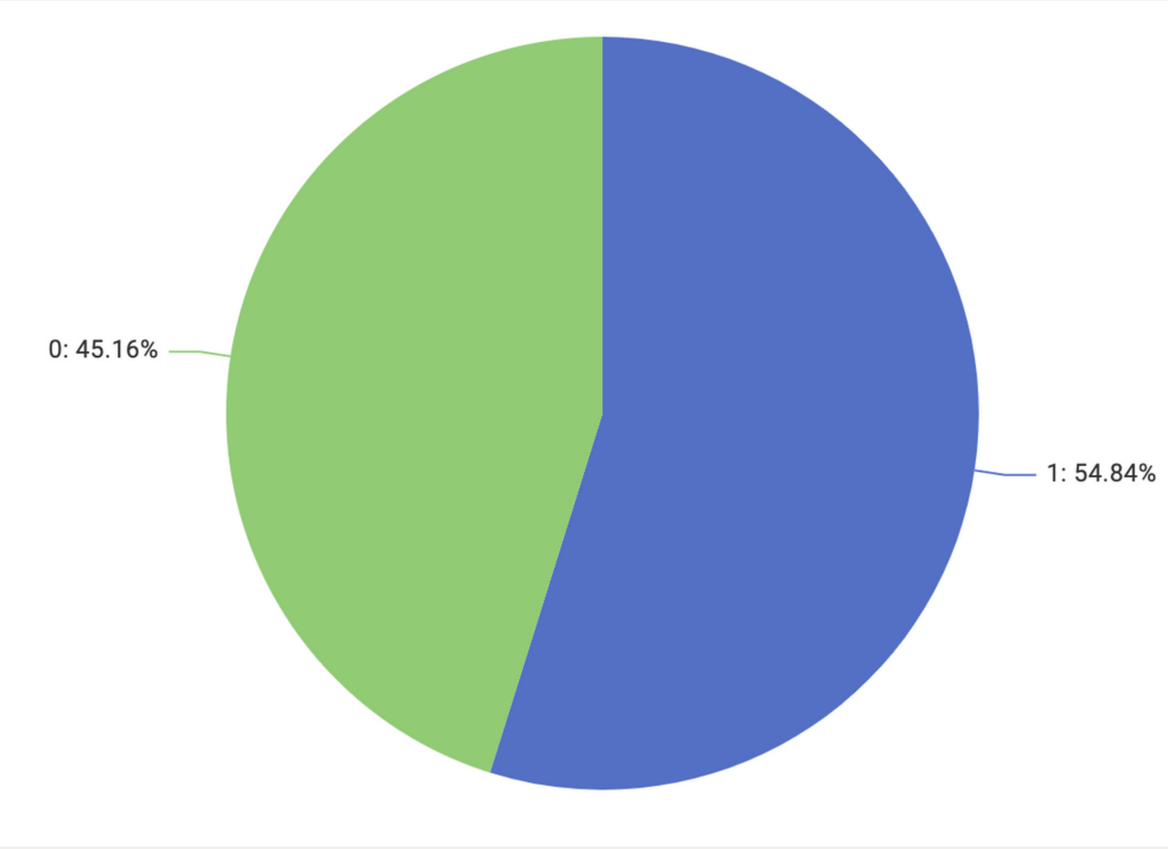
Displaced

Nature: Categorical, nominal

Description: whether the students are displaced or not

Categories: 1 – yes; 0 – no

Insights: Most of the students (54.84%) are displaced.



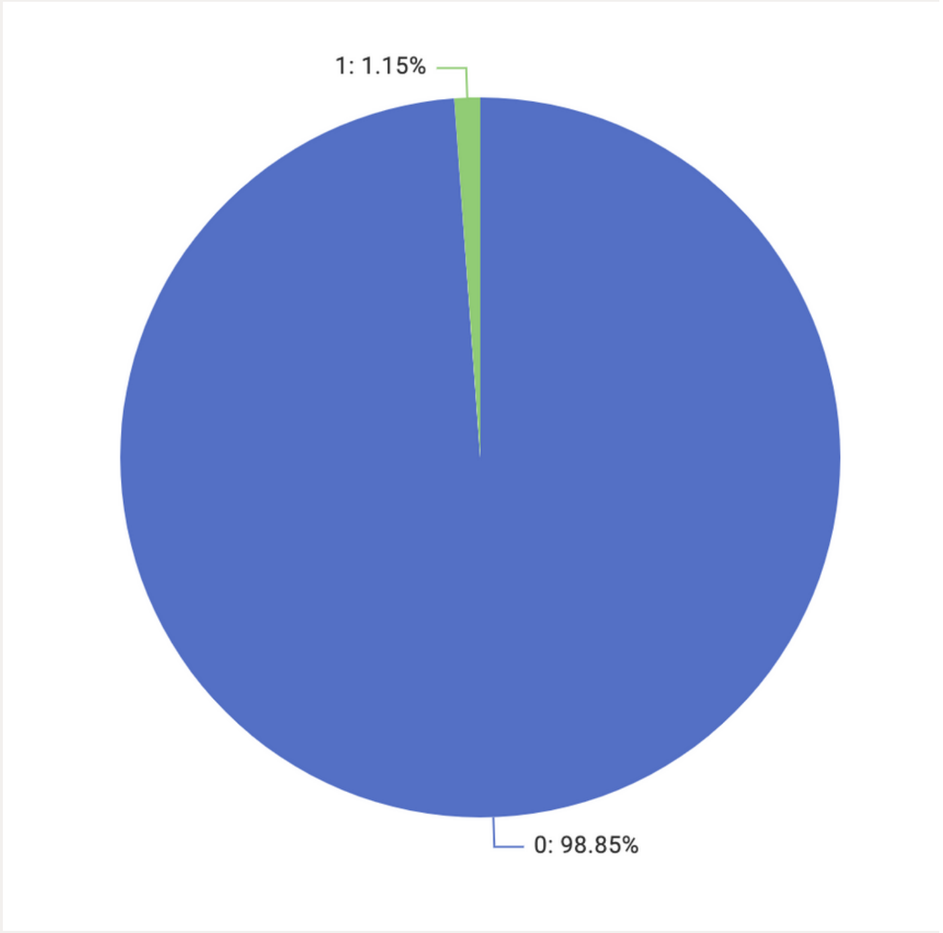
Educational special needs

Nature: Categorical, nominal

Description: whether the students have educational special needs or not

Categories: 1 – yes; 0 – no

Insights: The vast majority of students (98.85%) do not have educational special needs.



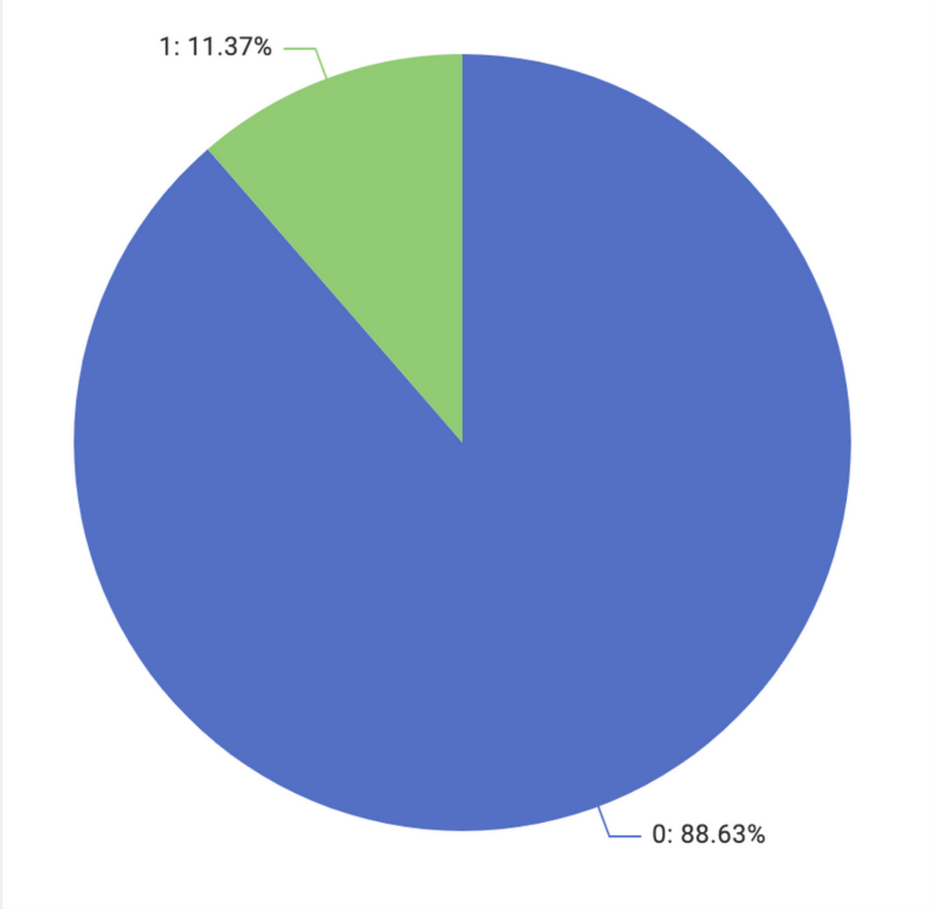
Debtor

Nature: Categorical, integer

Description: whether the students are debtors or not

Categories: 1 – yes; 0 – no

Insights: Most of the students (88.63%) are not in debt.



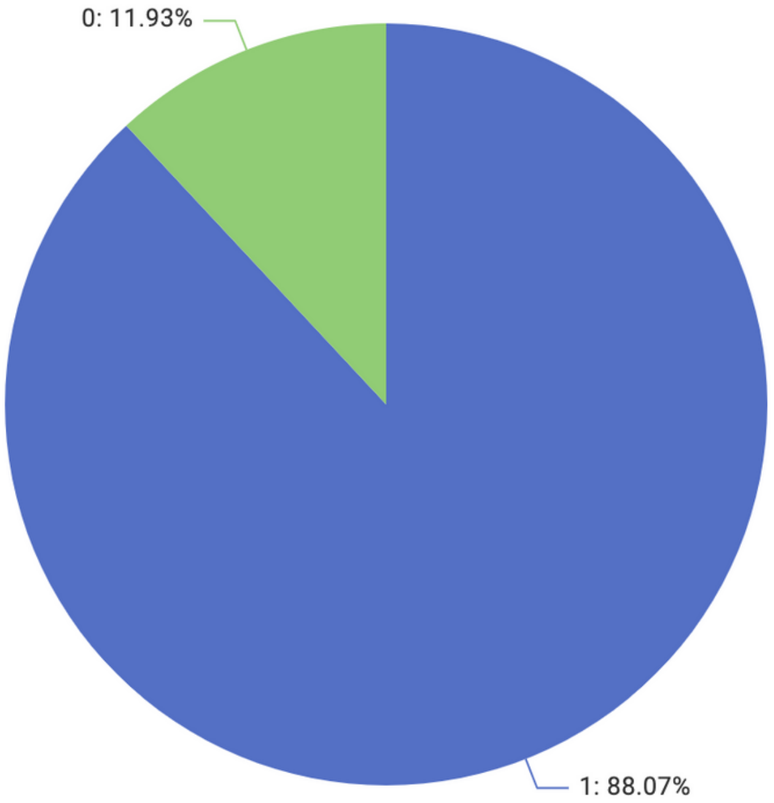
Tuition fees up to date

Nature: Categorical, nominal

Description: whether the students have their tuition fees paid up to date or not

Categories: 1 – yes; 0 – no

Insights: Most of the students (88.07%) are up to date with their tuition fees.



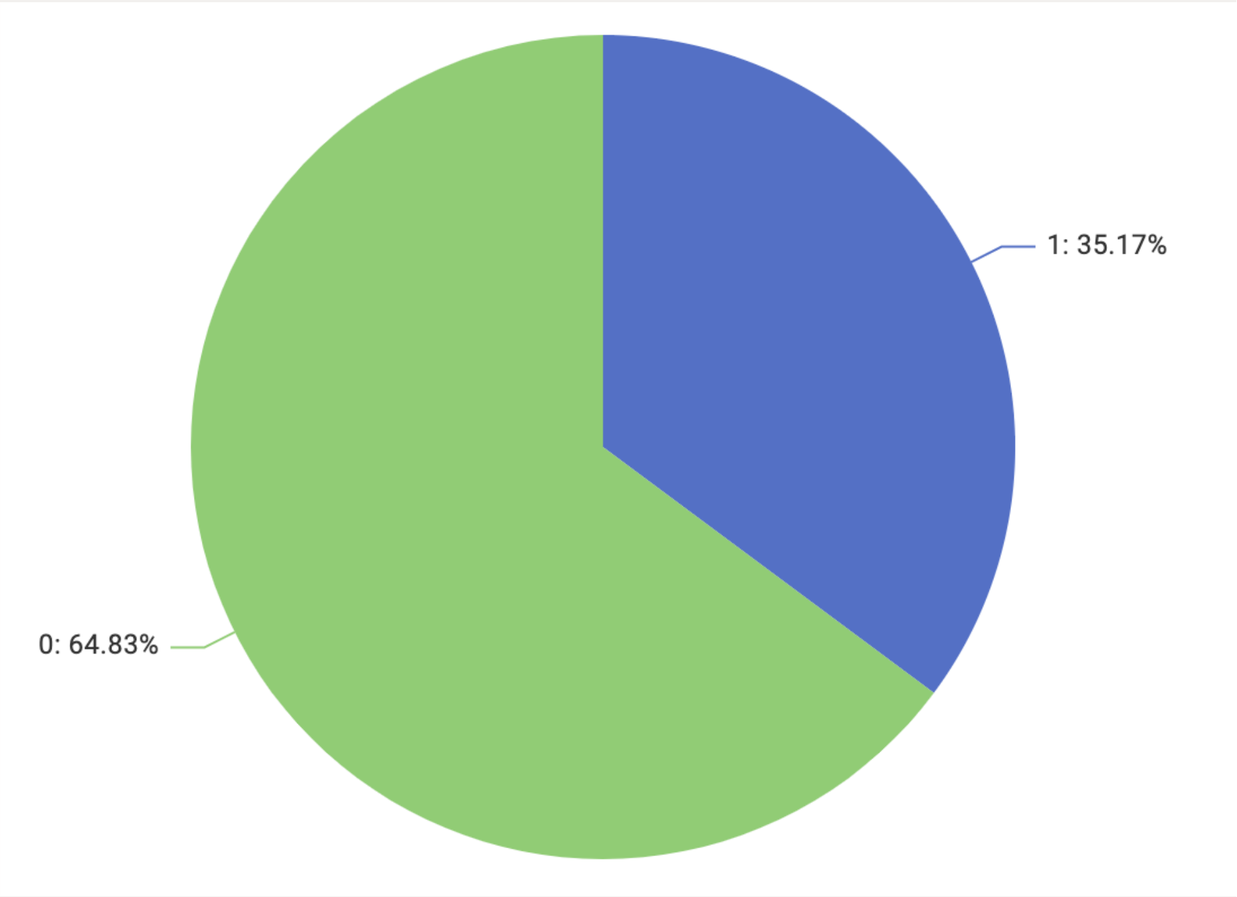
Gender

Nature: Categorical, nominal

Description: gender of the students

Categories: 1 – male; 0 – female

Insights: Most of the students (64.83%) are female.



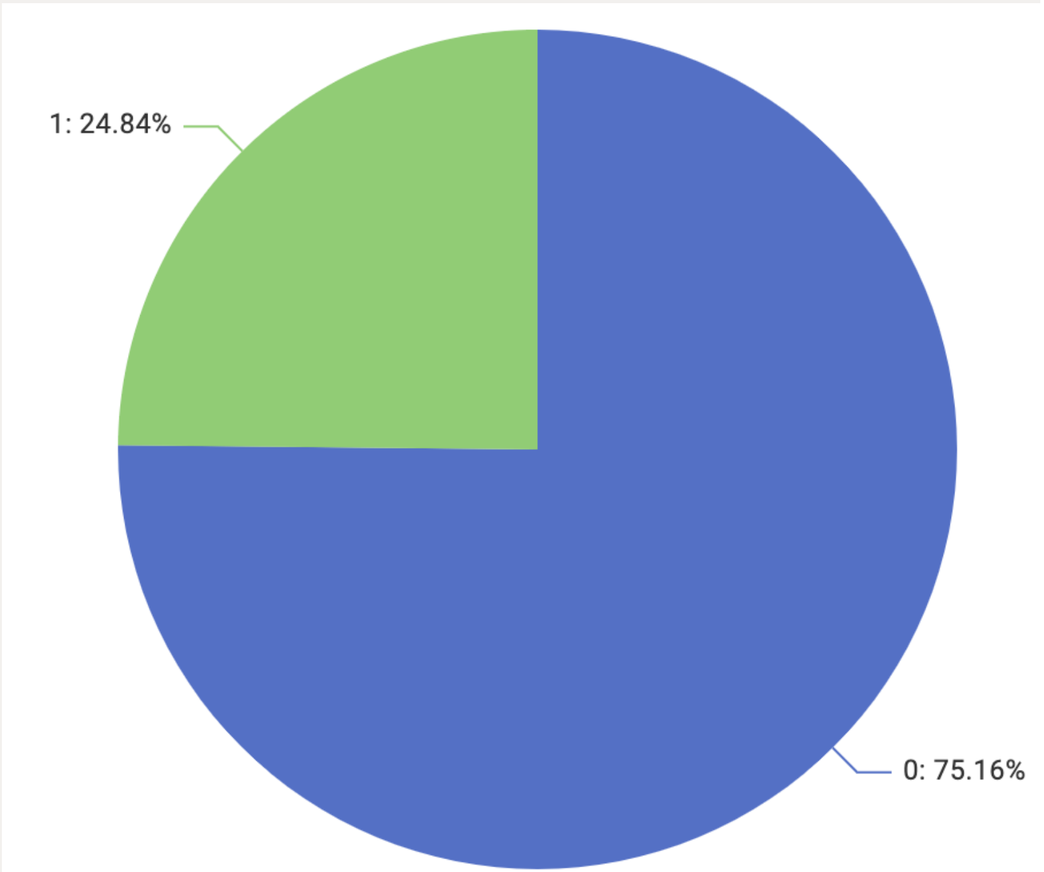
Scholarship holder

Nature: Categorical, nominal

Description: whether the students are scholarship holders or not

Categories: 1 – yes; 0 – no

Insights: Only 24.84% of the students hold a scholarship.



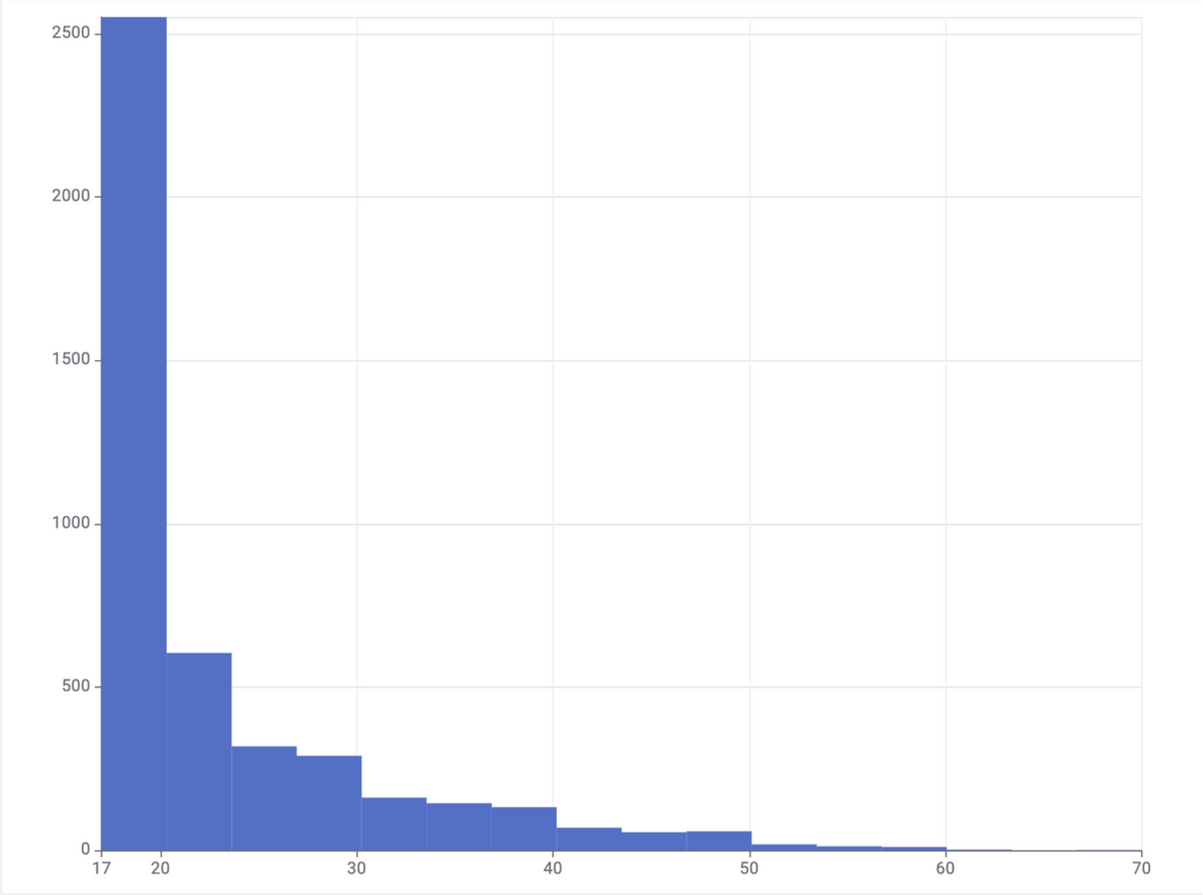
Age at enrollment

Nature: Numerical, discrete

Description: the age of the students at enrollment

Range: between 17 and 70

Insights: Most of the students (2551, or 57.7%) were between 17 and 20 years old at enrollment.



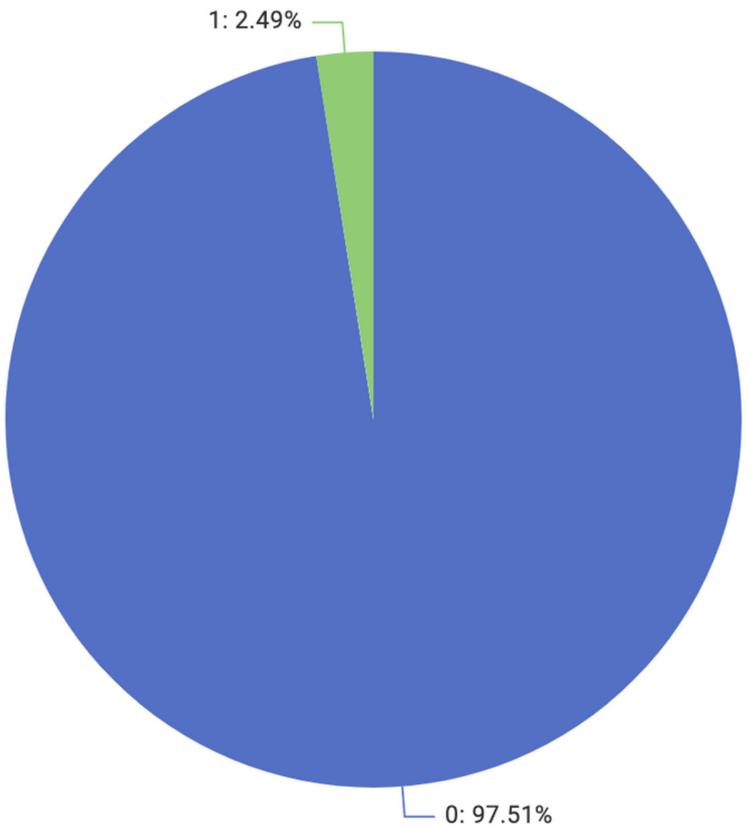
International

Nature: Categorical, nominal

Description: whether the students are international or not

Categories: 1 – yes; 0 – no

Insights: Only 2.49% of the students are international.



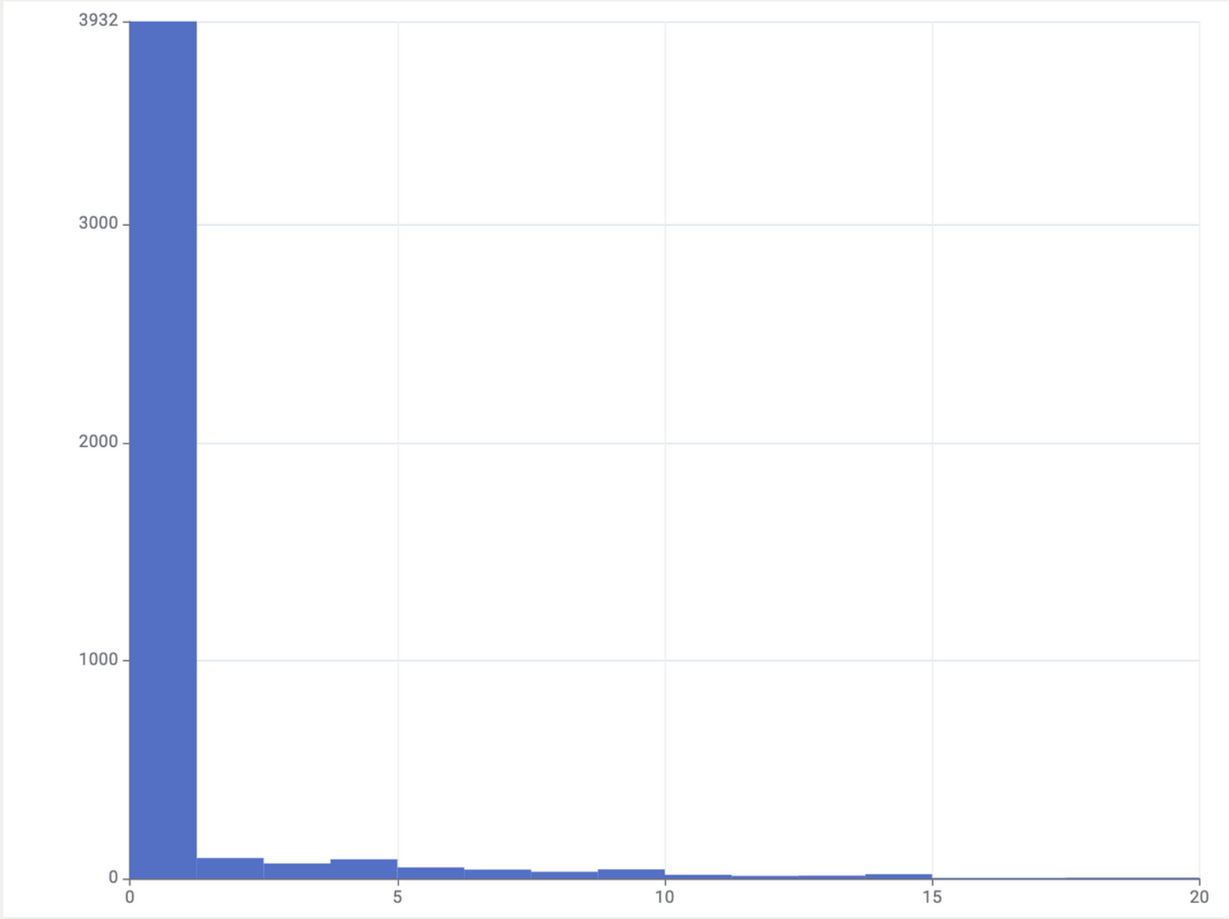
Curricular unites 1st sem (credited)

Nature: Numerical, discrete

Description: number of curricular units that students earned credits for during the first semester

Range: between 0 and 20

Insights: Most students (3847, or 87%) were credited 0 curricular units in the first semester.



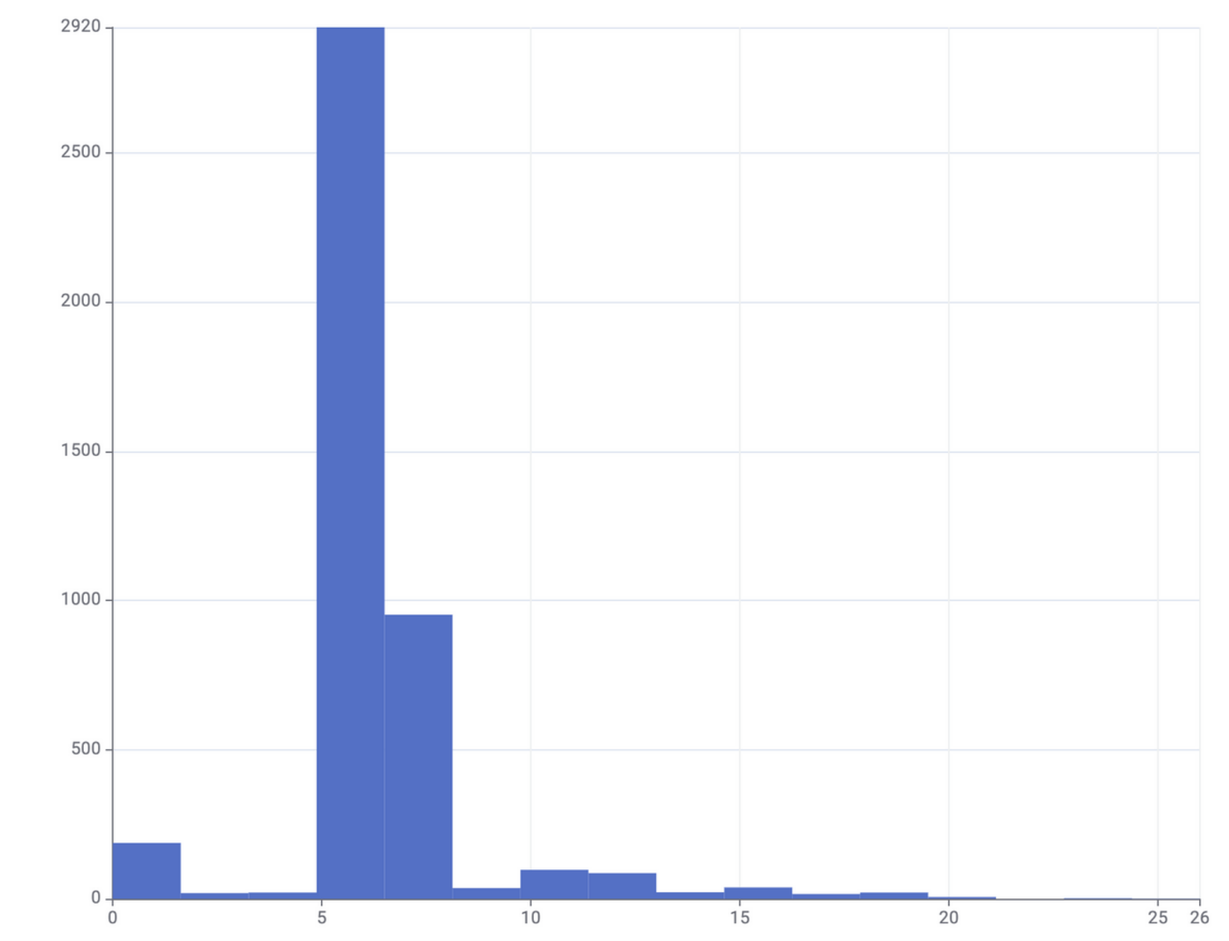
Curricular unites 1st sem (enrolled)

Nature: Numerical, discrete

Description: number of curricular units that students are enrolled in during the first semester

Range: between 0 and 26

Insights: Many students (1910, or 43.2%) are enrolled in 6 curricular units during the first semester. Many others (1010, or 22.8%) are enrolled in 5.



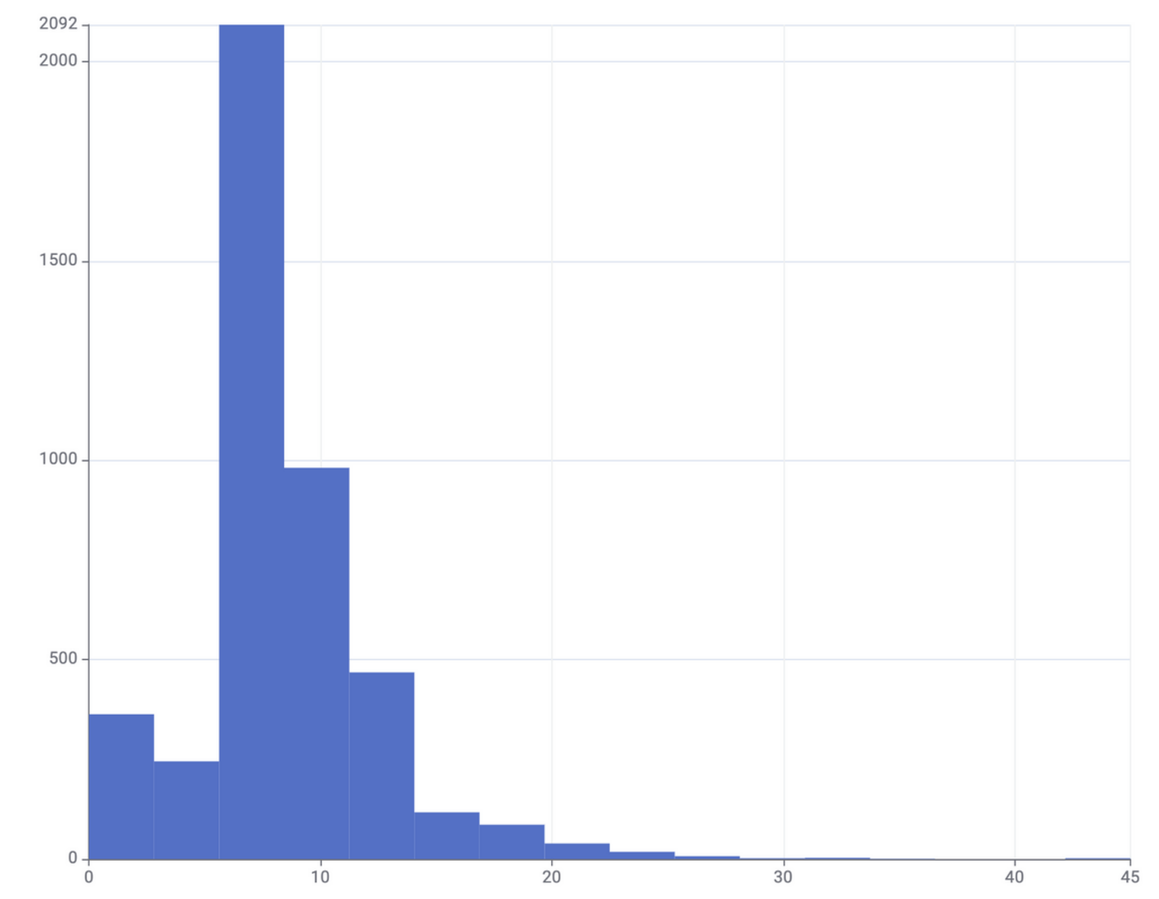
Curricular unites 1st sem (evaluations)

Nature: Numerical, discrete

Description: number of curricular units evaluated in the first semester

Range: between 0 and 45

Insights: Many students (791, or 17.9%) had 8 curricular units evaluated.



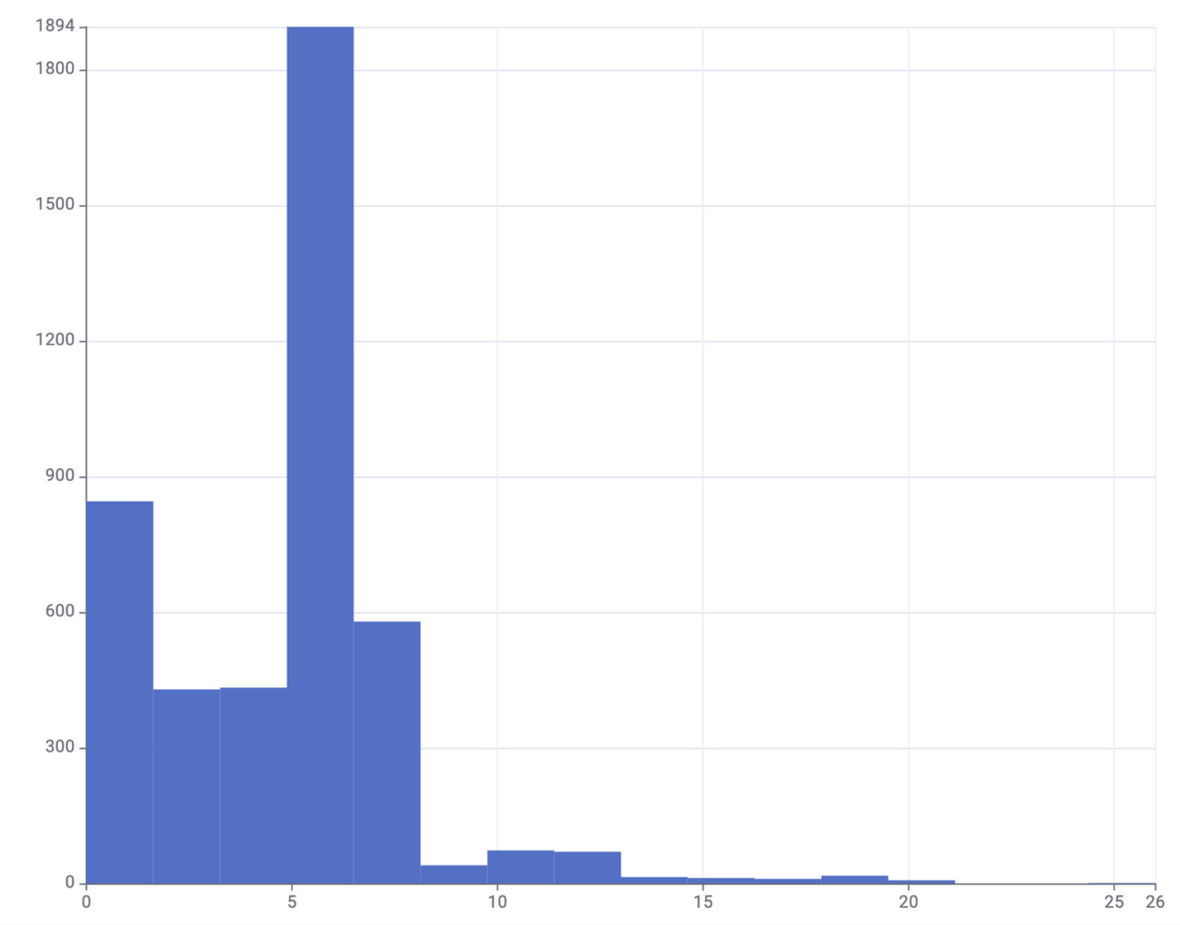
Curricular unites 1st sem (approved)

Nature: Numerical, discrete

Description: number of curricular units for which students gained approval during the first semester

Range: between 0 and 26

Insights: Many students (1171, or 26.5%) had 6 curricular units approves. Around 120 students had either 0 or 5 curricular units approved.



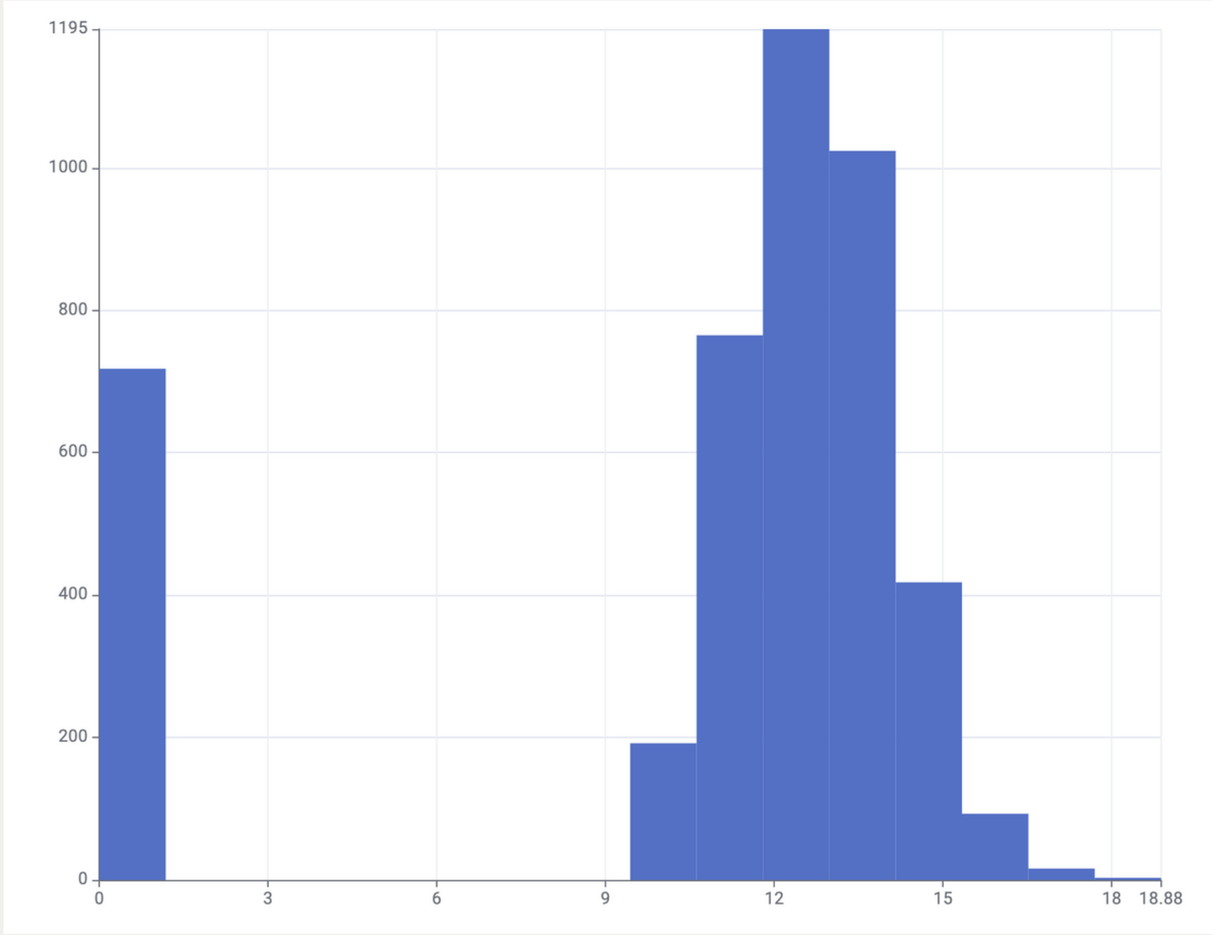
Curricular unites 1st sem (grade)

Nature: Numerical, discrete

Description: the grade associated with the curricular units that a student completed during the first semester

Range: between 0 and 20

Insights: Many students (963, or 21.8%) got a grade between 12 and 13.



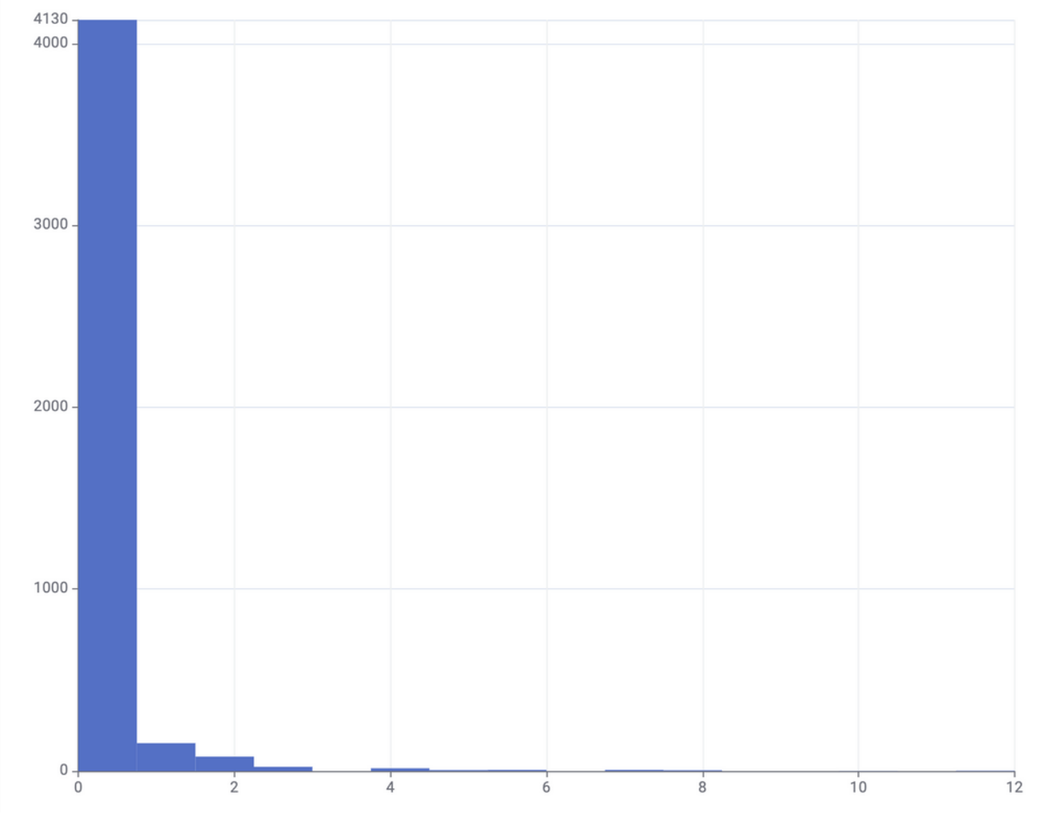
Curricular unites 1st sem (without evaluations)

Nature: Numerical, discrete

Description: number of curricular units that were not evaluated during the first semester

Range: between 0 and 12

Insights: Most of the students (4130, or 93.4%) have not undergone evaluation for any curricular units during the first semester.



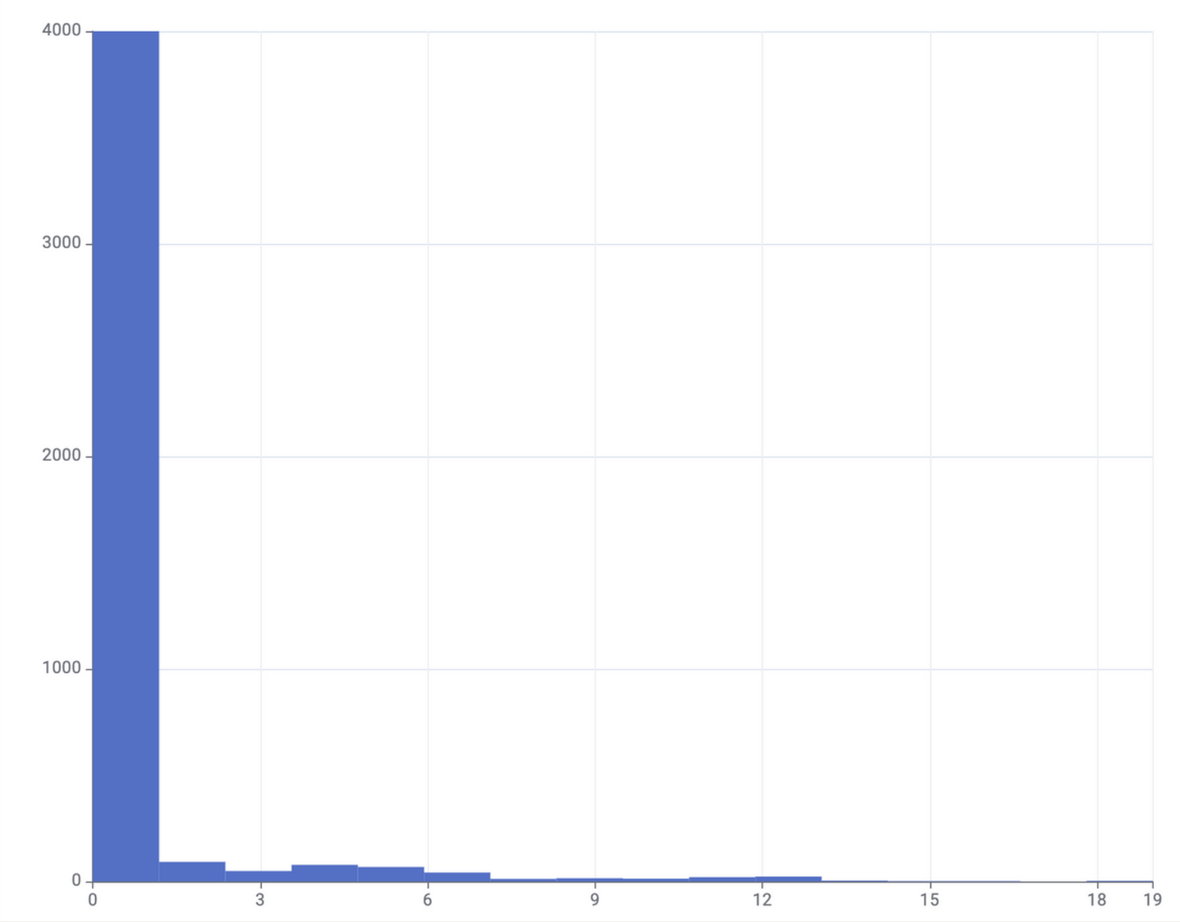
Curricular unites 2nd sem (credited)

Nature: Numerical, discrete

Description: number of curricular units that students earned credits for during the second semester

Range: between 0 and 19

Insights: Most of the students (3894, or 88%) were credited 0 curricular units during the second semester.



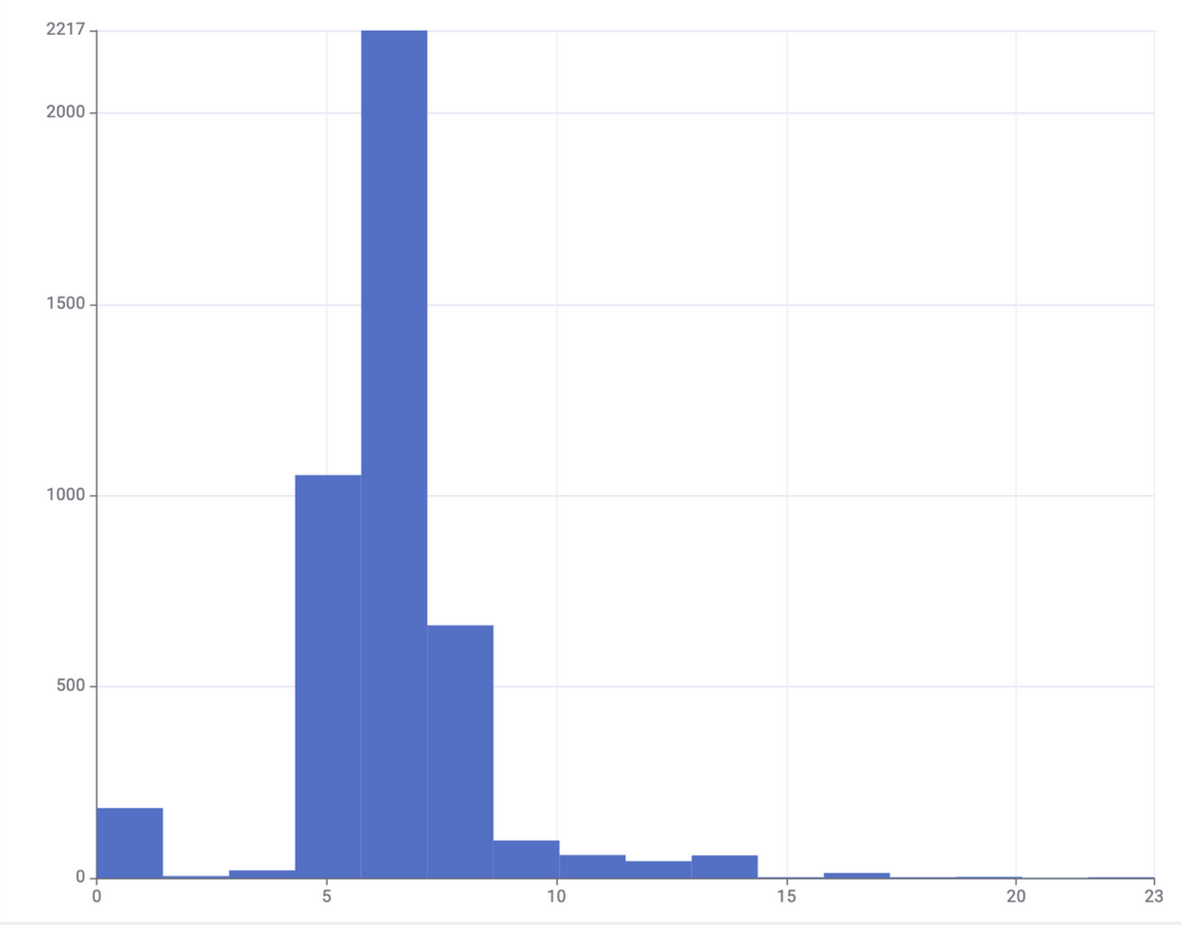
Curricular unites 2nd sem (enrolled)

Nature: Numerical, discrete

Description: number of curricular units that students are enrolled in during the second semester.

Range: between 0 and 23

Insights: Many students (1913, or 43.2%) are enrolled in 6 curricular units.



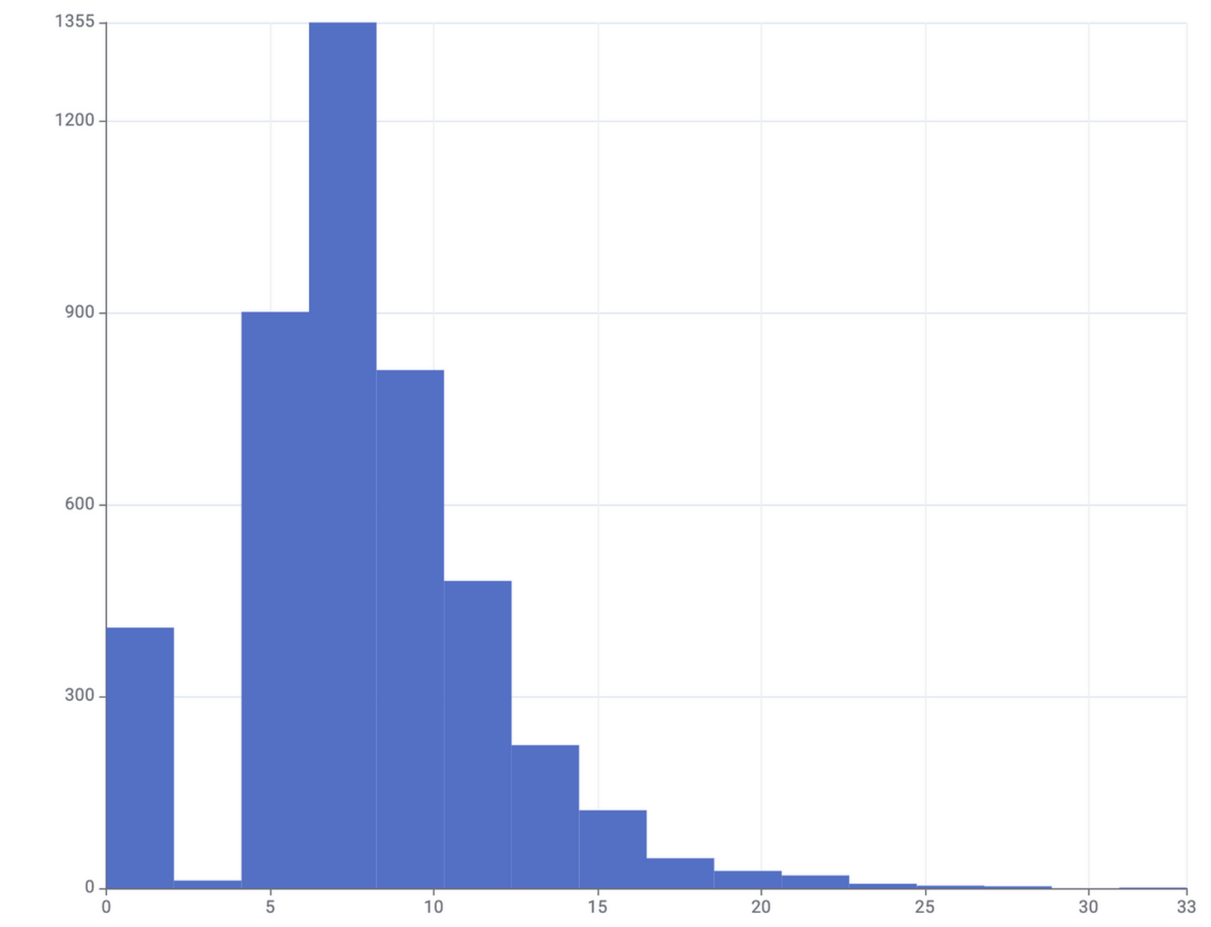
Curricular unites 2nd sem (evaluations)

Nature: Numerical, discrete

Description: number of curricular units evaluated in the second semester

Range: between 0 and 33

Insights: Many students (792, or 17.9%) had 8 curricular units evaluated.



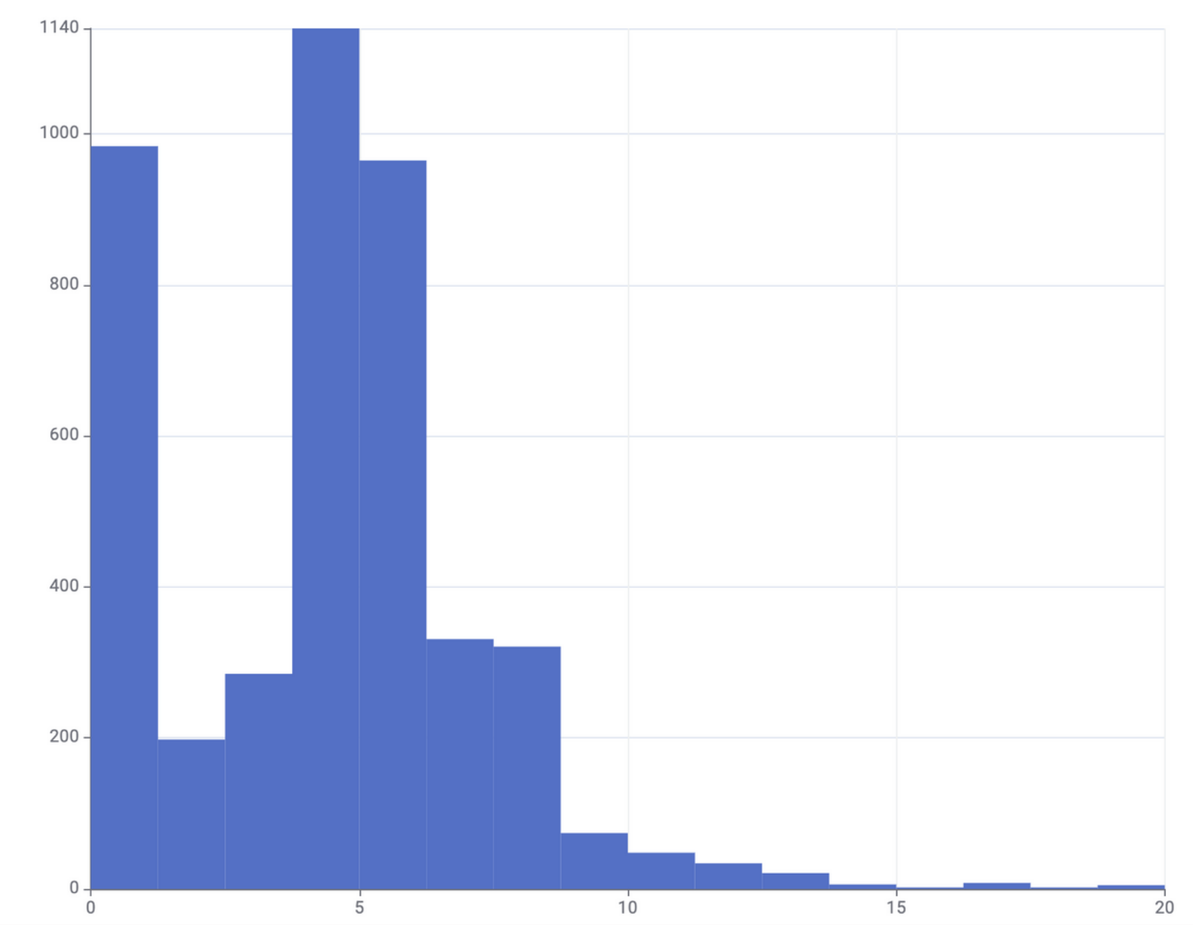
Curricular unites 2nd sem (approved)

Nature: Numerical, discrete

Description: number of curricular units for which students gained approval during the second semester

Range: between 0 and 20

Insights: Many students (965, or 21.8%) had 6 curricular units approved, many others (870, or 19.7%) had 0 approved.



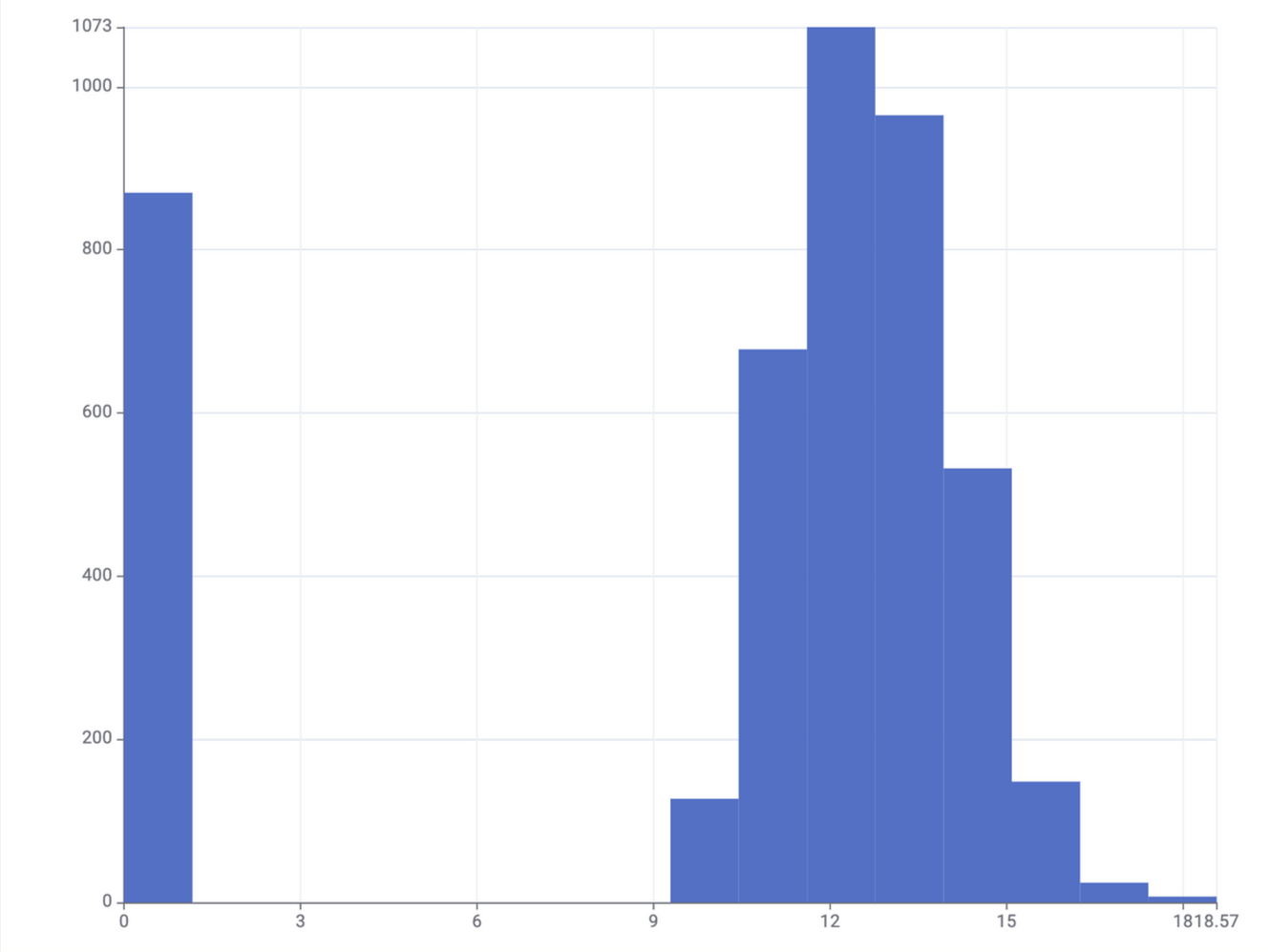
Curricular unites 2nd sem (grade)

Nature: Numerical, discrete

Description: the grade associated with the curricular units that a student completed during the second semester

Range: between 0 and 20

Insights: Many students (1009, or 22.8%) got a grade between 12.396 and 13.429.



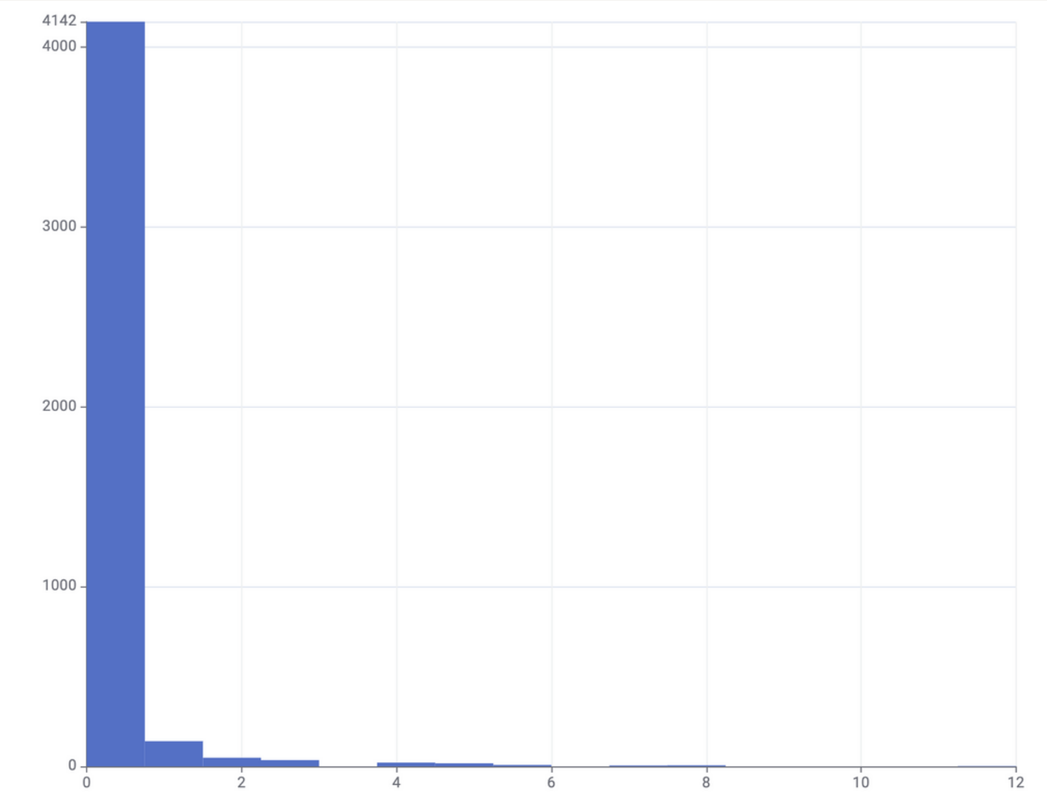
Curricular unites 2nd sem (without evaluations)

Nature: Numerical, discrete

Description: number of curricular units that were not evaluated during the second semester

Range: between 0 and 12

Insights: The vast majority of students (4142, or 93.6%) have not undergone evaluation for any curricular units during the second semester.



Unemployment rate

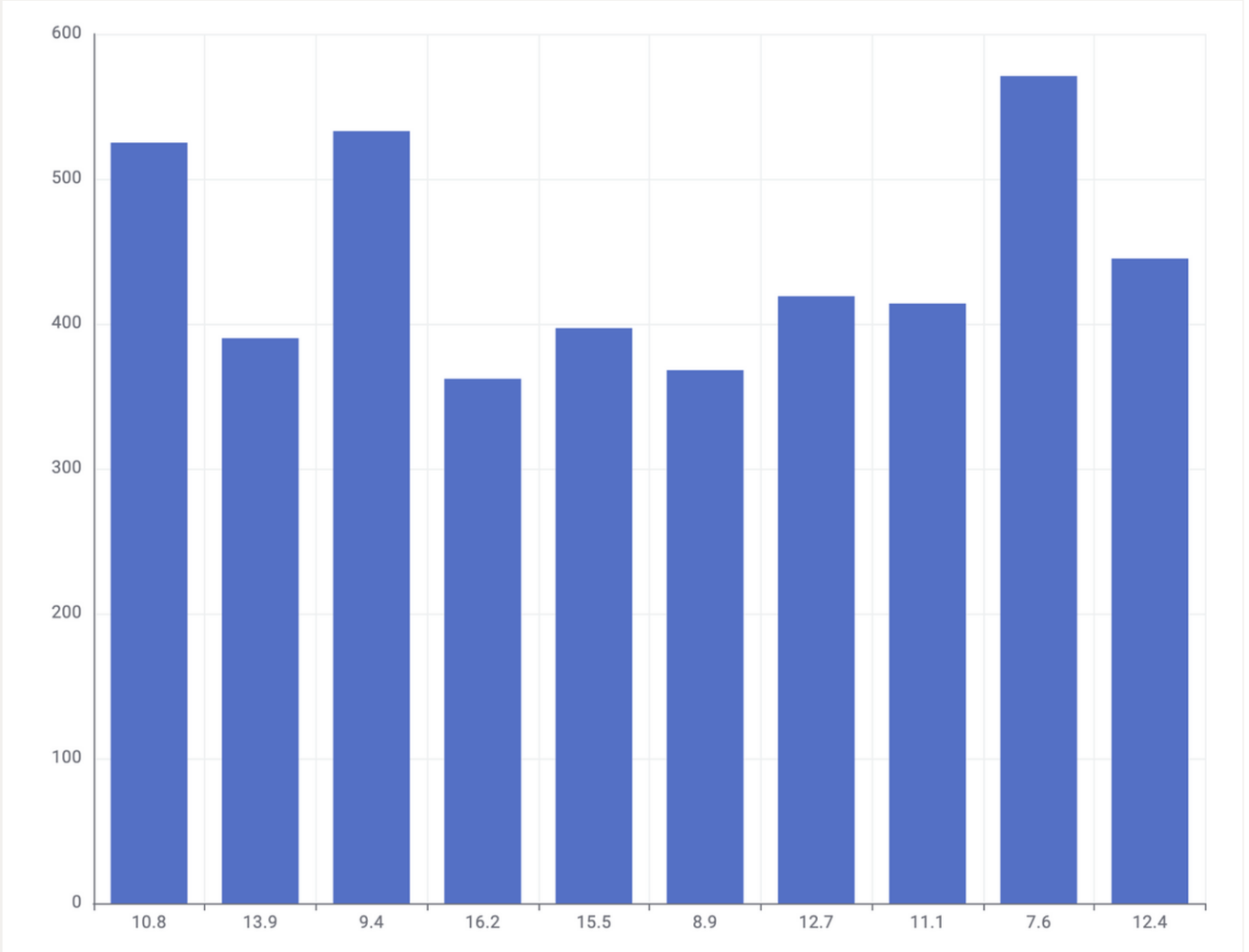
Nature: Numerical, continuous

Description: the unemployment rate in Portugal when each record was created

Range: between 7.6% and 16.2%

Insights: The unemployment rate in most (864) is between 12.19% and 12.77%.

Note: for this variable, we chose a bar chart instead of a histogram. While it's theoretically a continuous variable, the limited information available only covers discrete time periods. Consequently, this variable exhibits numerous repeated values, as depicted in the bar chart.



Inflation rate

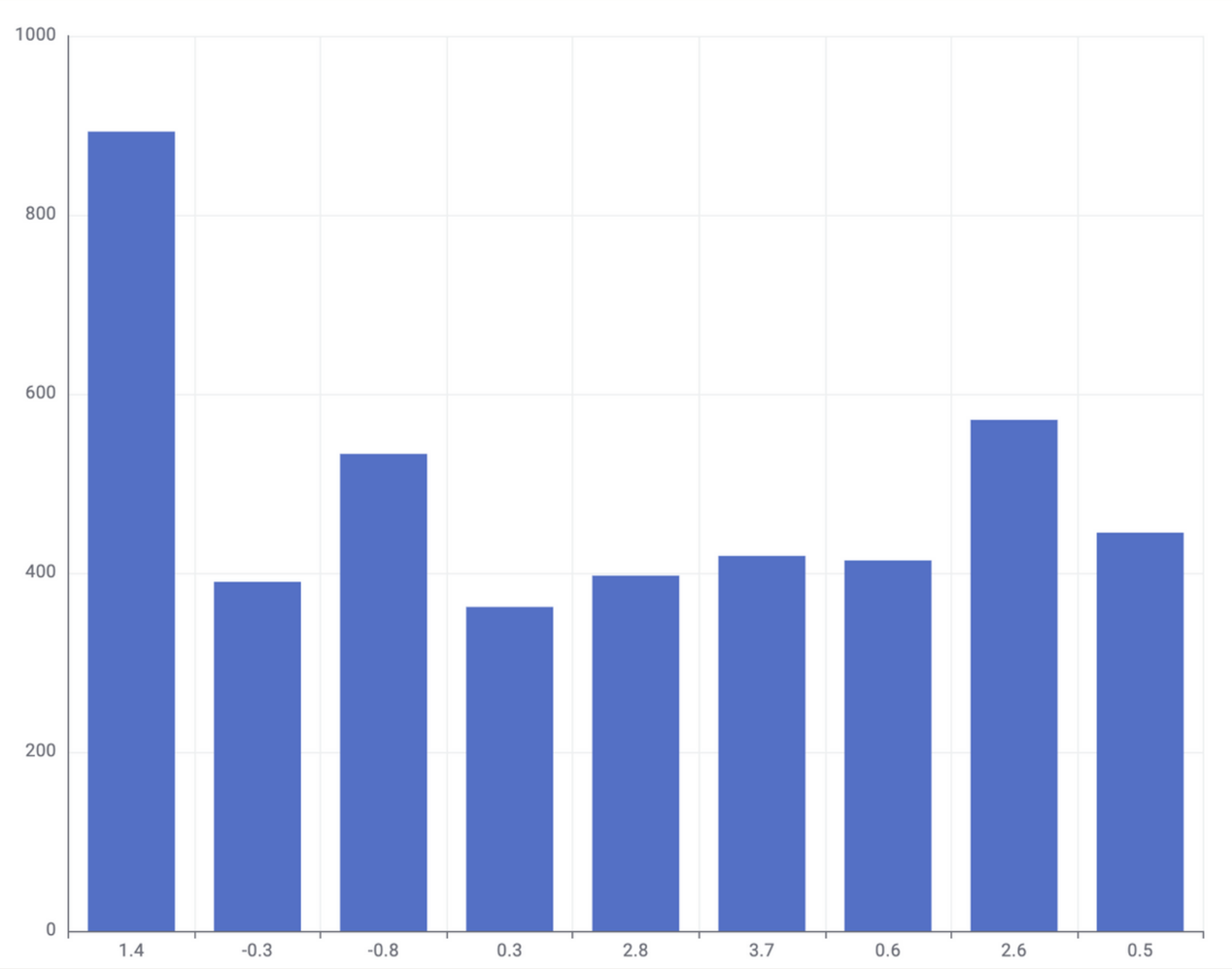
Nature: Numerical, continuous

Description: the inflation rate of the Eurozone when each record was created

Range: between –0.8% and 3.7%

Insights: The inflation rate in most cases (893) is between 1.376% and 1.558%. For many other students (859), the inflation rate is between 0.466% and 0.648%.

Note: same considerations as before



GDP

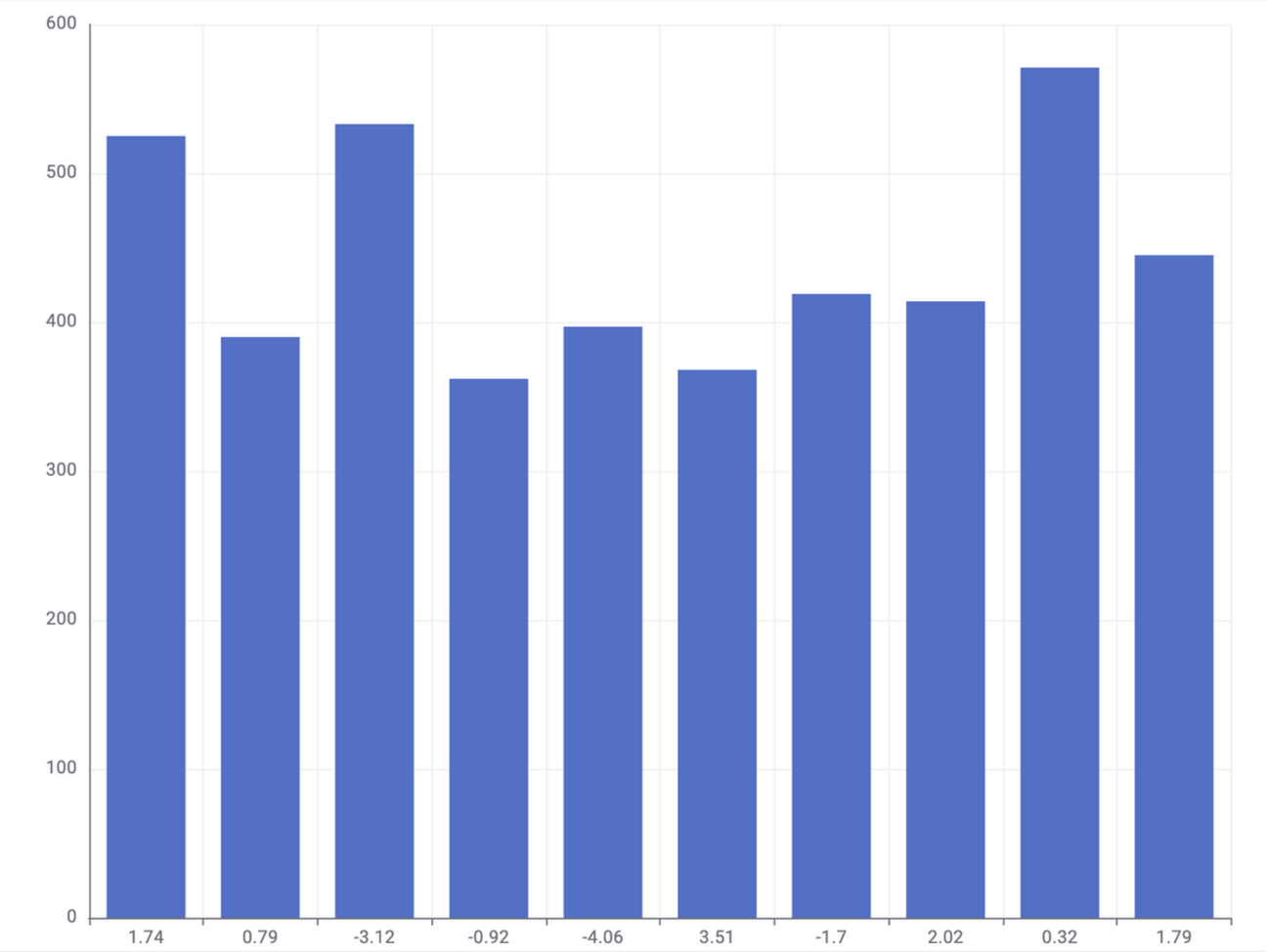
Nature: Numerical, continuous

Description: Percentage change in Portugal GDP relative to the previous year when each record was created

Range: between –4,06% and 3,51%

Insights: The percentage GDP change in most cases (976) is between 1,532% and 1,817%.

Note: same considerations as before



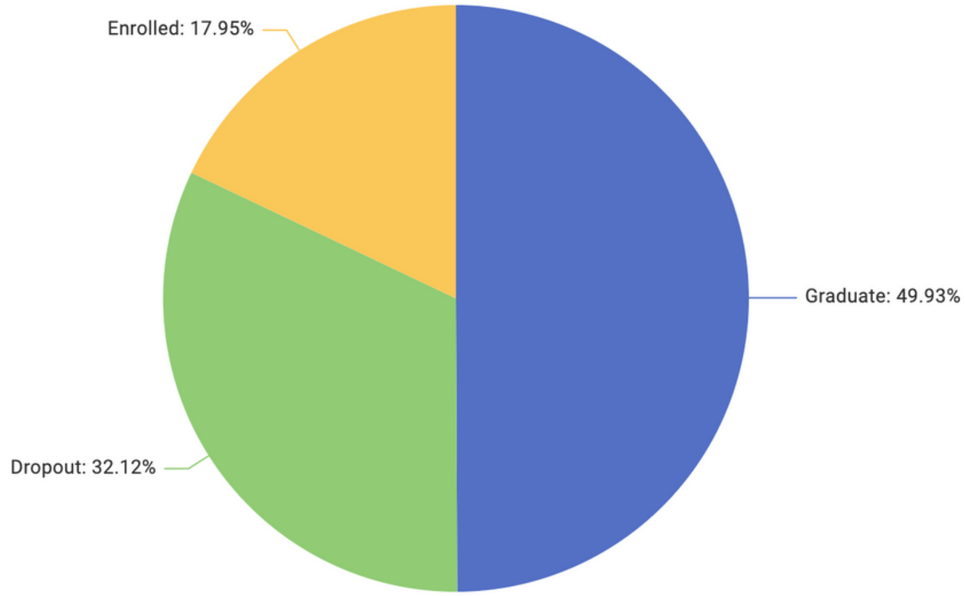
Target

Nature: Categorical

Description: the classification of the students at the end of the normal duration of the course

Categories: 1 – dropout; 2 – enrolled; 3 - graduate

Insights: Nearly half of the students are graduates, a small portion (17.95%) are currently enrolled, and 32.12% have dropped out.



Data preparation

MISSING VALUES
OUTLIERS
STRINGS TO INTEGERS
VARIABLE ENCODING
NEW FEATURE CREATION
FEATURE SELECTION

Overview of data preparation

In this section we accomplish several important feature-engineering tasks.

- We first identify and deal with missing values.
- We later present our approach towards handling outliers, explaining why we decide to keep them.
- We go on by transforming two variables from strings to integers, and to explaining the rationale behind this decision.
- Subsequently, we group together the values of some variables (both categorical and numerically-meaningful), to later have simpler and easier-to-interpret models.
- We later use available variables to create new and potentially-relevant columns in the dataset.
- We conclude by examining the correlations among the different predictors in our dataset. Additionally, we remove a few variables to prevent potential multicollinearity issues in our later analysis and to streamline our study.

Missing values

- The presence of missing data in the training data set may reduce the fit and predictive ability of the model, possibly leading to misclassifications.

- By using a *Statistics* node, we notice that the only missing values are 62 observations in Nationality, 32 in Course (0.7% of data) and 17 in Marital Status (0.3% of data).

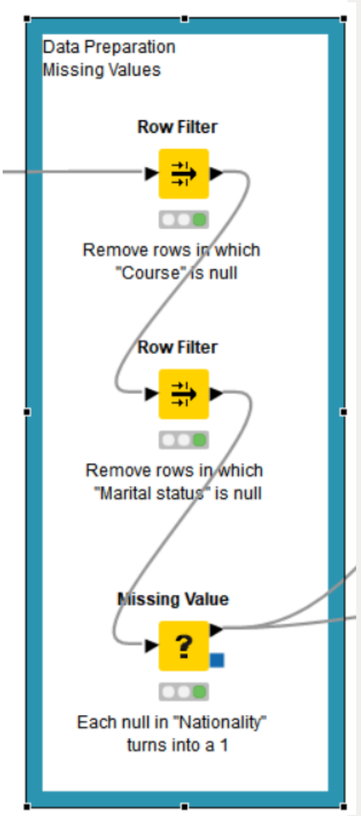
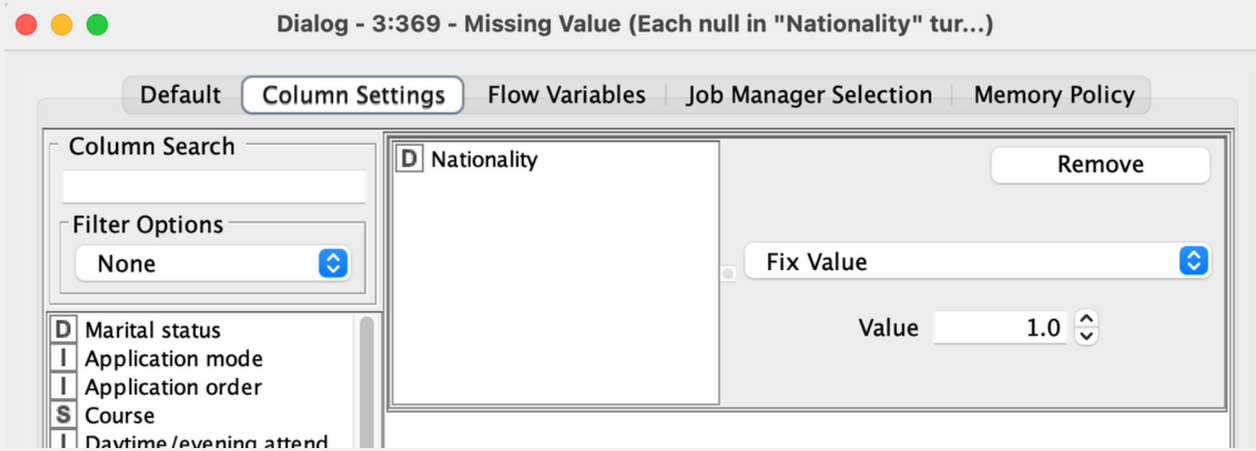


Table View

Rows: 37 | Columns: 23

Row...	Name *String*	Type *String*	# Mi... *Number (long)*	# Unique... *Number (long)*
Nati...	Nationality	Number (dou...	62	21
Cour...	Course	String	32	17
Marit...	Marital status	Number (dou...	17	6
Appli...	Application ...	Number (inte...	0	18

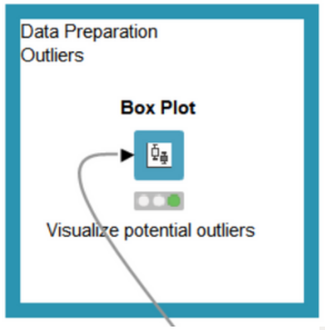


- We decide to apply a list-wise deletion for both “Marital Status” and “Course”, by using a *Row Filter* node. Instead, for “Nationality”, since the vast majority (96%) of the observations is 1, we change the missing values in this column with this value through a *Missing Value* node, in order to avoid reducing the sample size any further.

Outliers

When looking for outliers, we only focus on numerical predictors. Moreover, we do not analyze all the columns related to academic units because (spoiler alert) many of them will be eliminated later on, to avoid multicollinearity issues.

As evidenced by the box plots in the following slide, the only variables with outliers are the ones in the second plot, namely “Previous qualification (grade)”, “Admission grade”, and “Age at enrollment”. As it is obvious from the considerations we made in the previous section about the variables “Unemployment rate”, “Inflation rate”, and “GDP”, these predictors do not have outliers.



Regarding how we deal with outliers, we follow a different approach depending on whether the outliers are upward or downward. Upward outliers are values greater than 1.5 * third quartile, and downward outliers are values less than 0.5 * first quartile. Specifically:

- for the “Previous qualification (grade)” and “Admission grade” columns, we only remove upward outliers
- for the “Age at enrollment” column, we only remove downward outliers

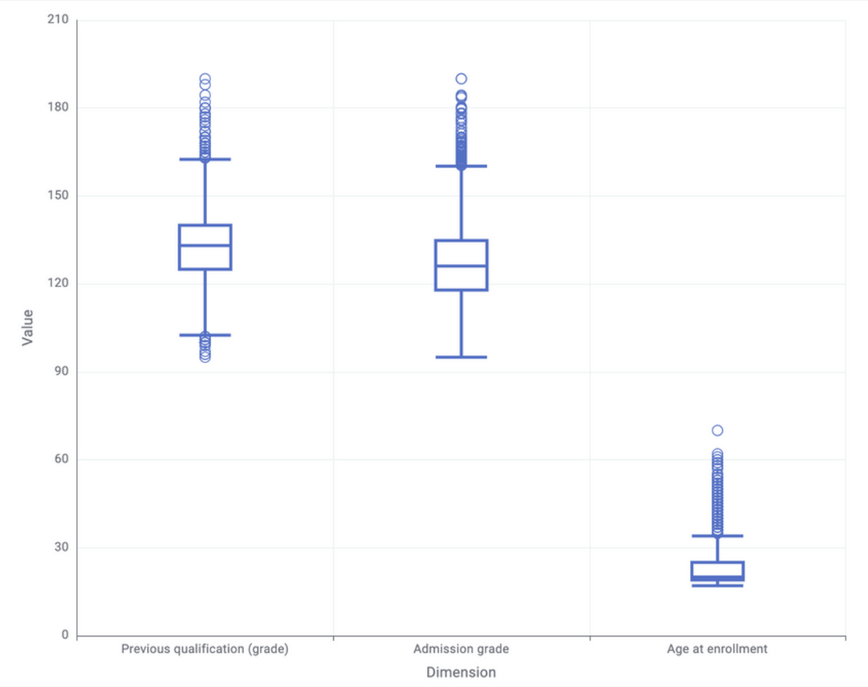
The reason behind this peculiar approach is straightforward. Given the nature of our analysis, we want to build models to identify at-risk students, and we expect many of them to have either low grades or high ages (or both). Thus, not training our models with these data might seriously undermine the effectiveness of our work.

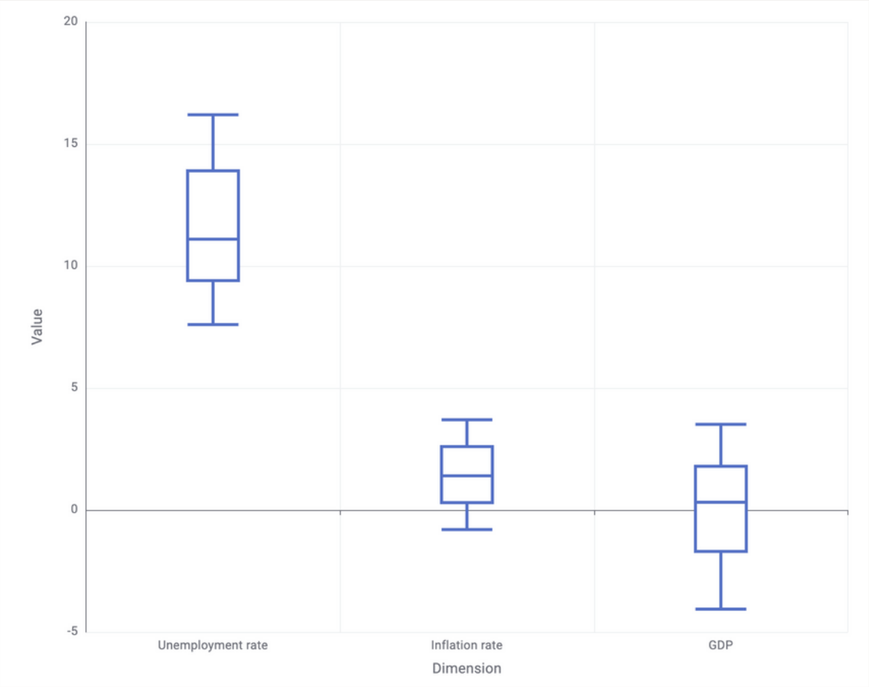
Outliers

- For “Previous qualification (grade)”, 1.5 * third quartile corresponds to a value of 210. Being 200 the maximum value of this column, we do not have any outliers to eliminate here.

- For “Admission grade”, 1.5 * third quartile corresponds to a value of 202.2. Also for this variable the maximum value is 200, thus there are no outliers to drop here.

- For “Age at enrollment”, being 17 the minimum value (a reasonable age to enrol at university, given that many countries have a 4-year high school), we do not eliminate any value.



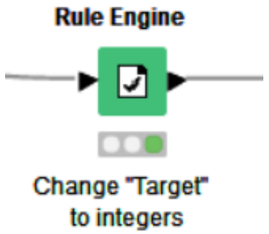


Thus, eventually, we do not have any outlier to eliminate

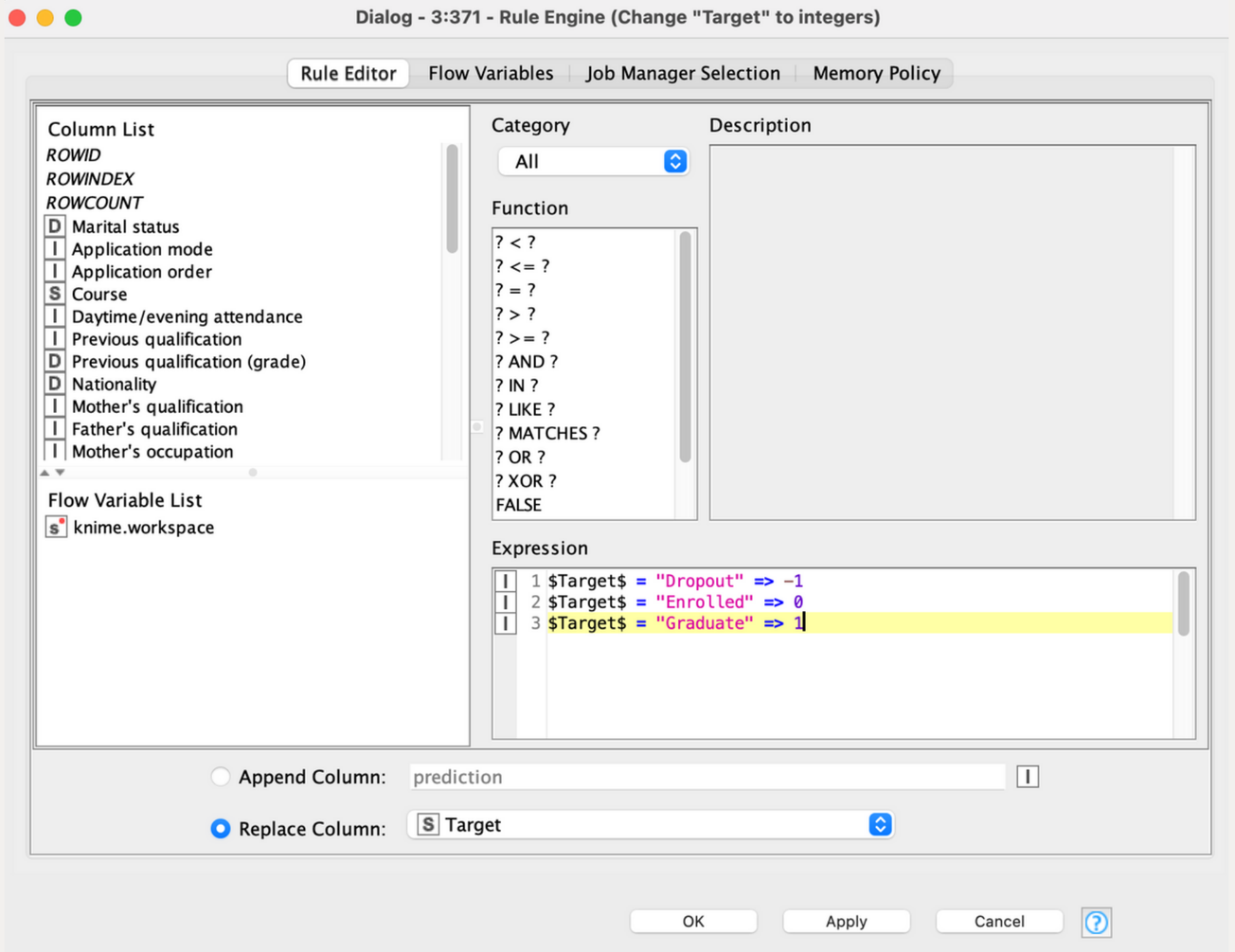
Turning strings to integers -**Target**

We use a *Rule Engine* node to turn the values in the “Target” column from strings to integers, according to the following scheme:

- Graduate –> 1
- Enrolled –> 0
- Dropout –> -1

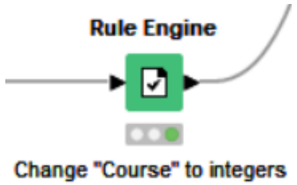


Target Number (inte...
-1
1
-1
1
1
1
1
-1

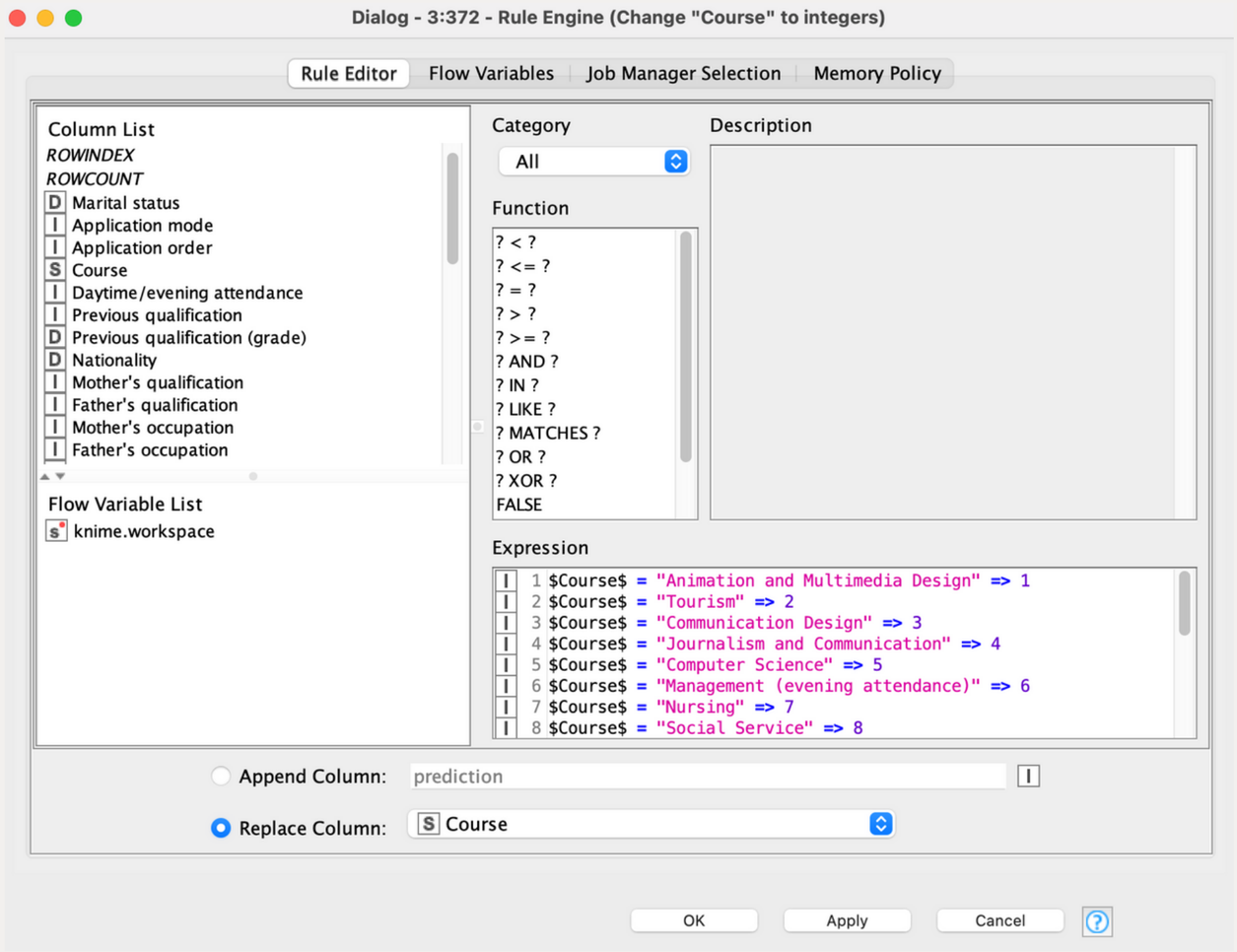


The reason is that this will later allow us to compute the correlation between any of our predictors and the target variable, offering a way to establish the strength of the effects of our variables on “Target”

Turning strings to integers -**Course**



We do the same thing for the “Course” variable, following this scheme:

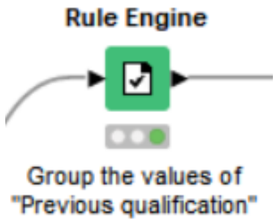


- Animation and Multimedia Design –> 1
- Tourism –> 2
- Communication Design –> 3
- Journalism and Communication –> 4
- Computer Science –> 5
- Management (evening attendance) –> 6
- Nursing –> 7
- Social Service –> 8
- Advertising and Marketing Management –>
- Basic Education –> 10
- Veterinary Nursing –> 11
- Equinculture –> 12
- Oral Hygiene –> 13
- Management –> 14
- Agronomy –> 15
- Biofuel Production Technologies –> 16
- Informatics Engineering –> 17

Variable encoding -**Previous qualification**

We now proceed by turning some continuous variables into discrete ones. We do this for several reasons:

1) Students in the same category are expected to behave similarly
2) This makes the dataset simpler to analyze, and thus drawing conclusions should be more straightforward
3) The increase in simplicity should also benefit the efficacy of the machine learning models we develop, as the number of instances we are working with is not enormous (slightly more than 4000 students)

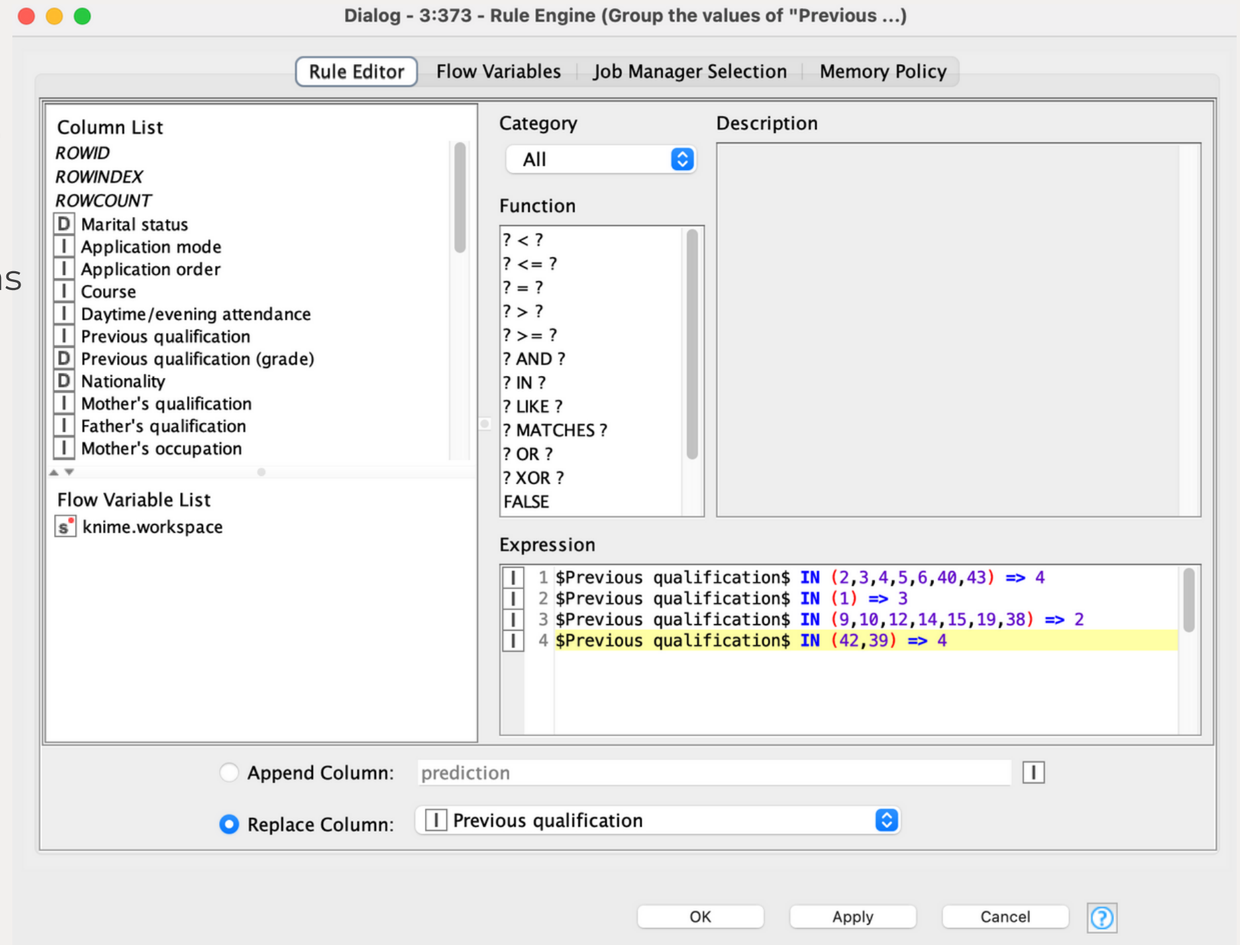


We later follow a similar rationale for encoding continuous (or, anyway, numerically-meaningful) variables

For this task we use other *Rule Engine* nodes.

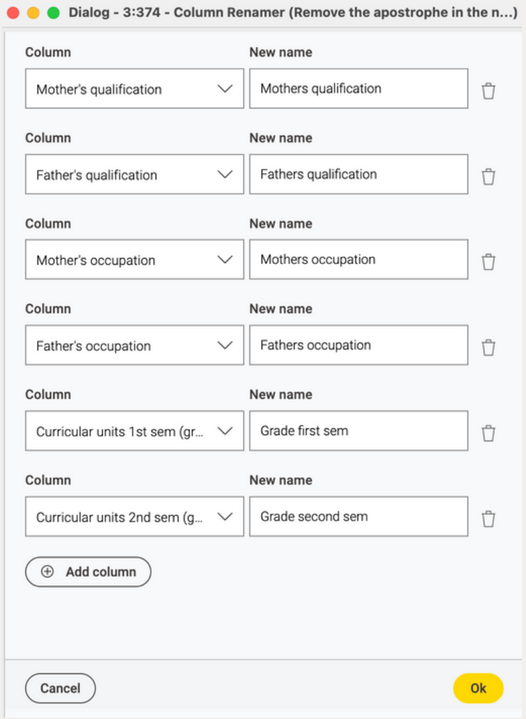
We start by grouping the values of “Previous qualification” into 4 sets based on the Portuguese education system:

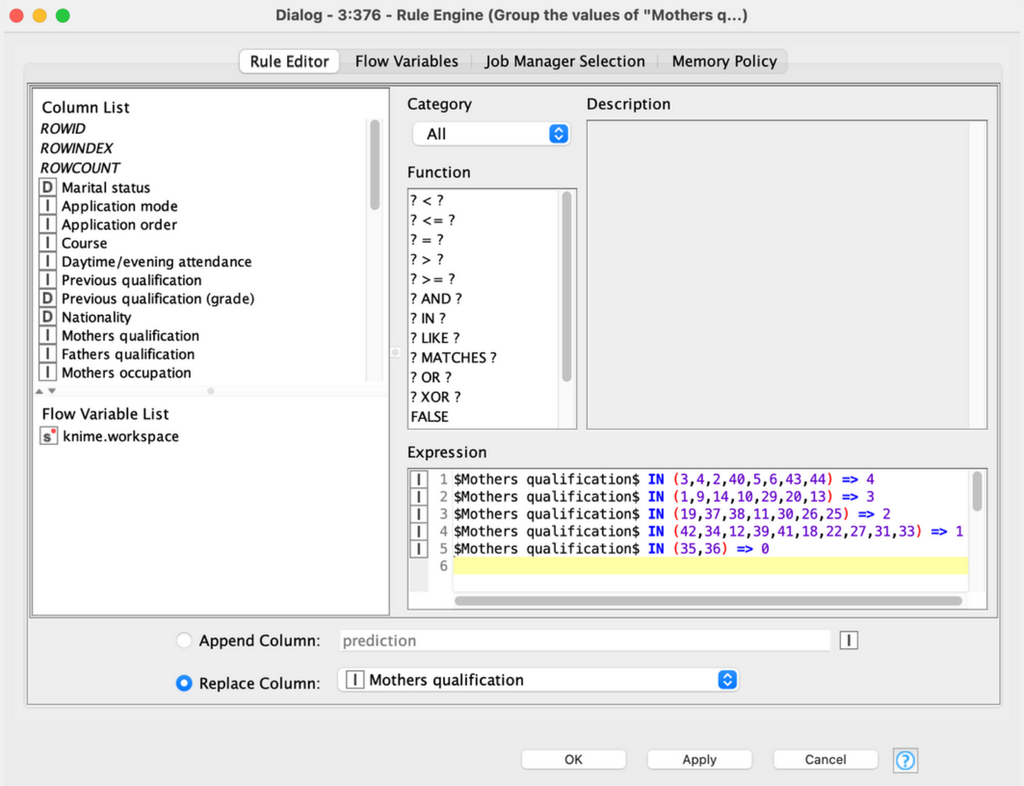
- "Higher Education", 4
- "Secondary Education", 3
- "Basic Education", 2
- "Other Education", 1



Variable encoding -**Mother's and father's qualification**

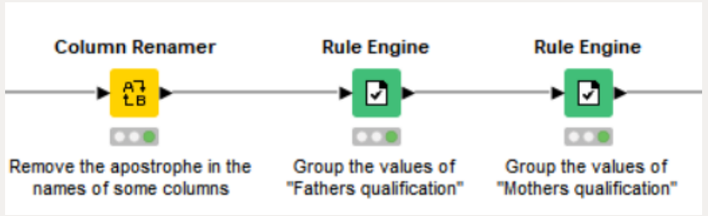
We proceed by doing the same with mother's and father's qualification (given the presence of the apostrophe in the names of these columns, we first change their names to avoid problems with the *Rule Engine* node). We also rename others columns whose original names would cause problems later on.





We group them in the following categories:

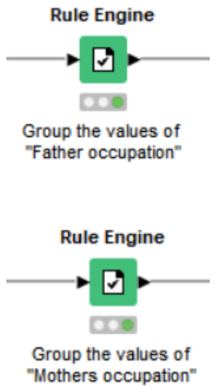
- "Higher Education", 4
- "Secondary Education", 3
- "Basic Education", 2
- "Other Education", 1
- "No Education", 0

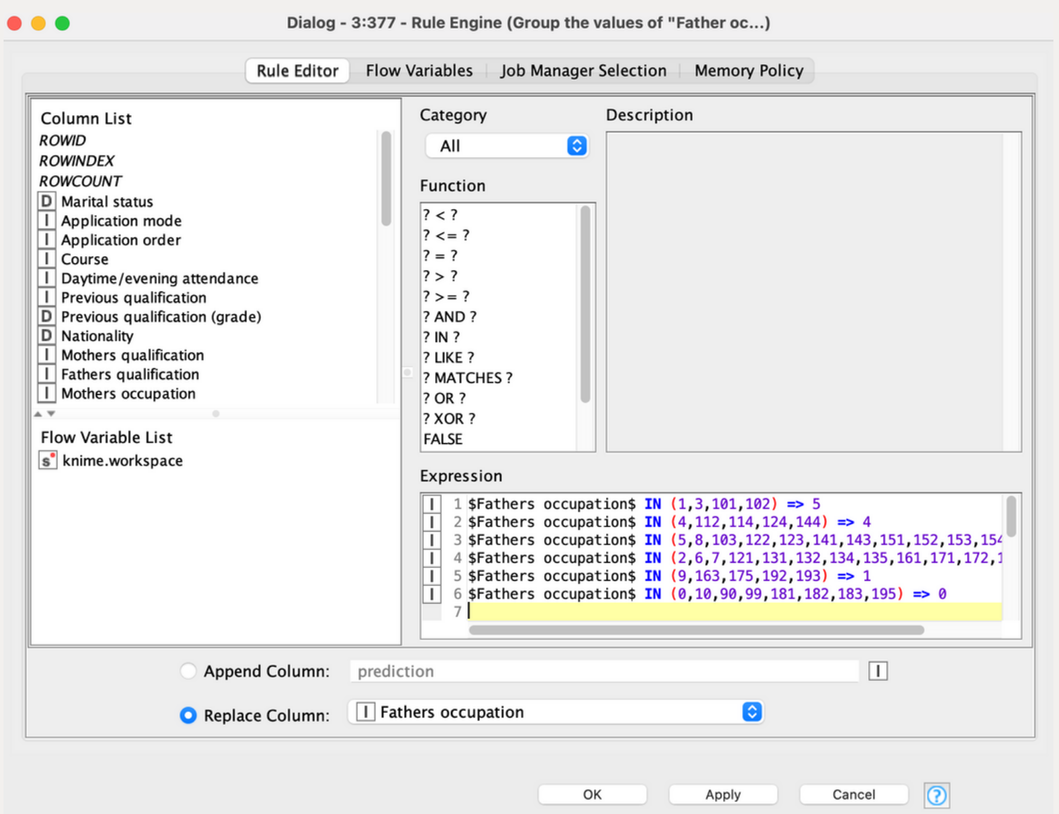


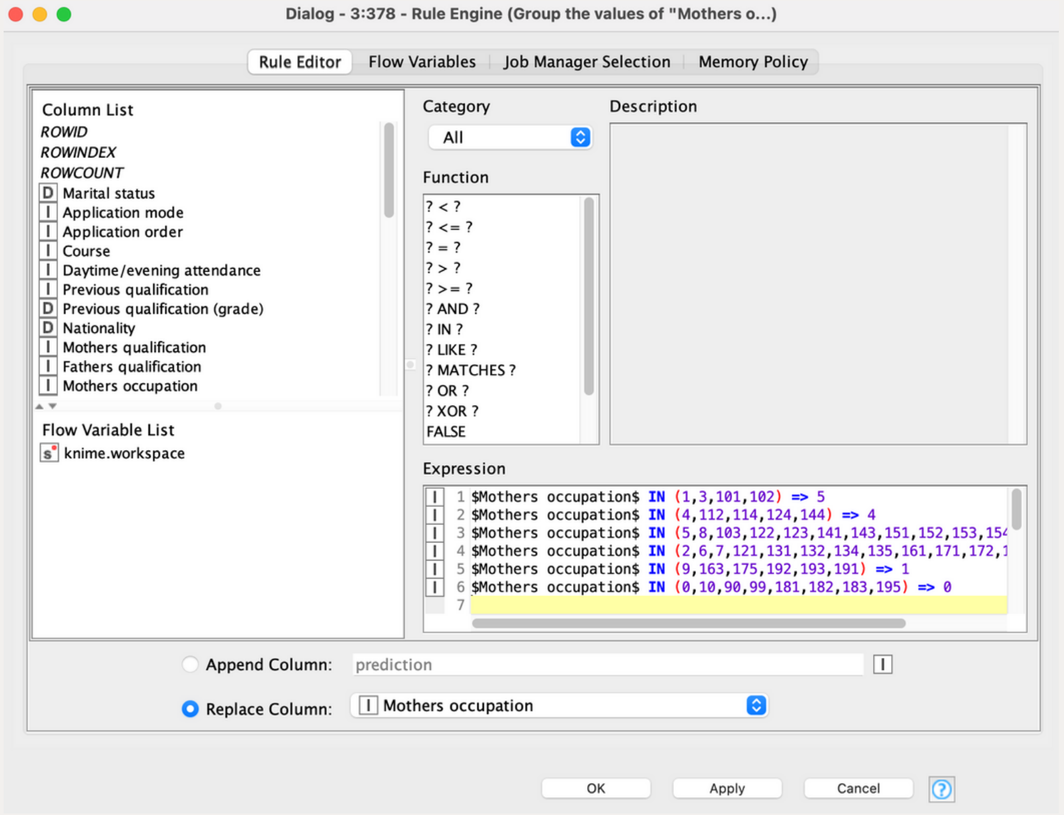
Variable encoding -**Mother's and father's occupation**

Regarding mother and father's occupation, we group them according to the type of job or sector:

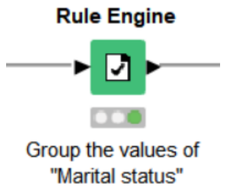
- "Professional and Managerial occupations", 5
- "Administrative occupations", 4
- "Service and Sales occupations", 3
- "Skilled workers", 2
- "Unskilled workers" 1
- "Other", 0





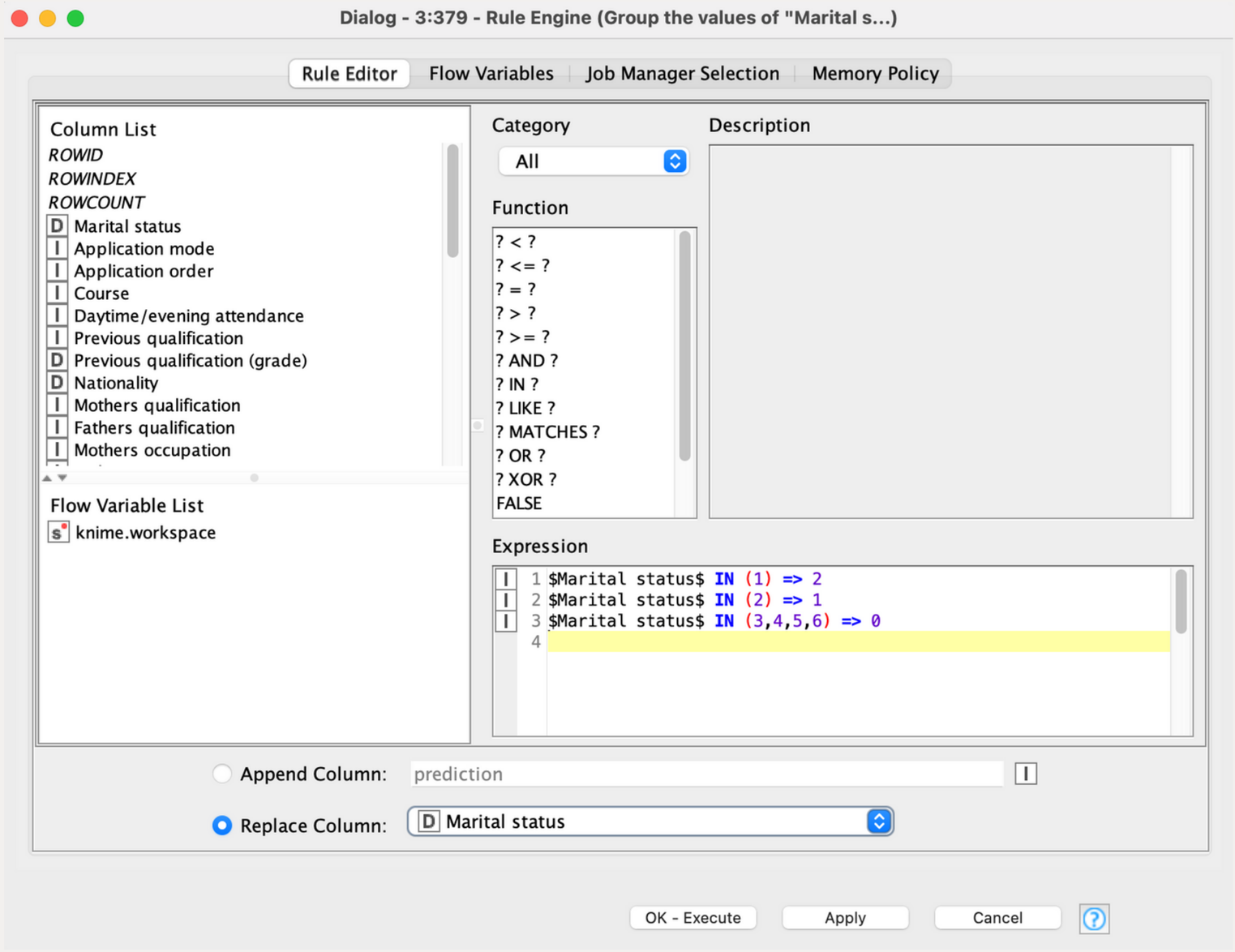


Variable encoding -**Marital status**



For the variable “Marital Status”, we create three categories:

- Single, 2
- Married, 1
- Other, 0

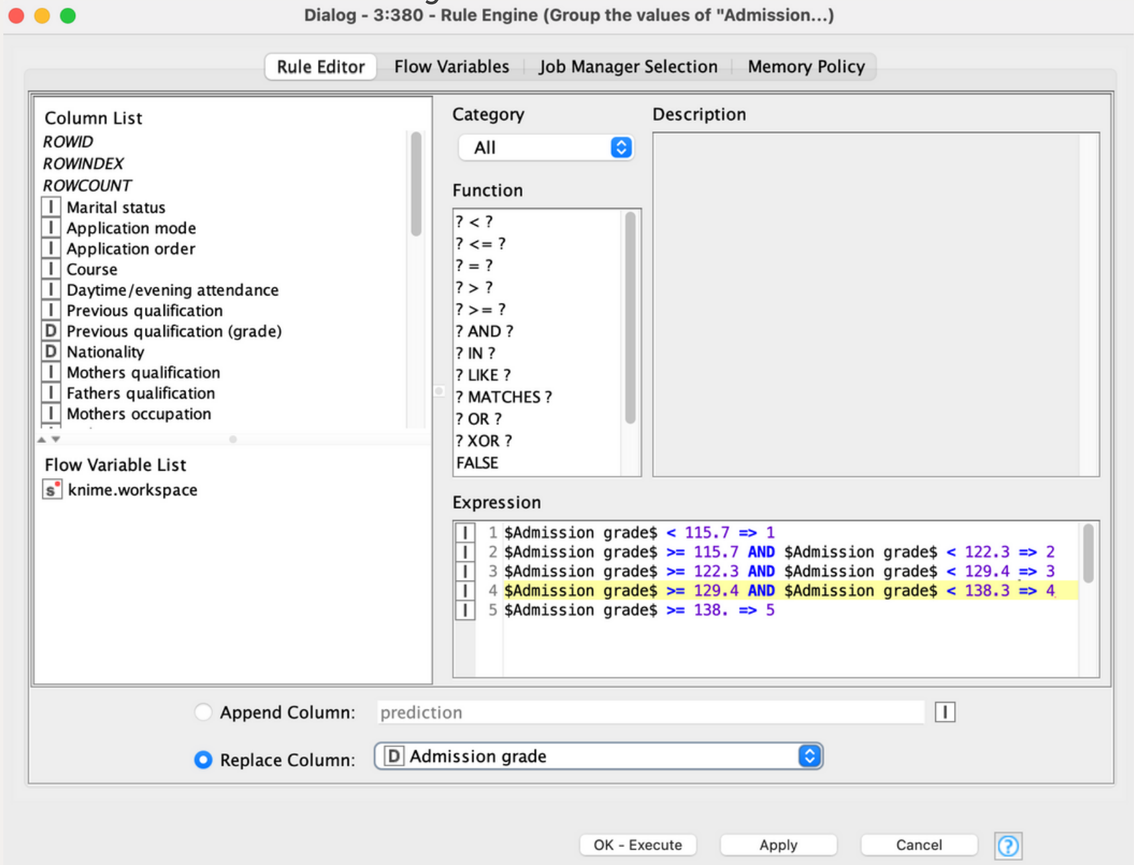


Variable encoding -**Admission grade**

We now do something similar to what we have just done, but with numerically-meaningful variables. Specifically, we group their values in different categories, under the assumption that students within the same category should behave similarly.

We start by turning the continuous variable “Admission grade” into a categorical one.
The scheme we follow is:

- Admission grade < 115.7 —> 1
- 115.7 <= Admission grade < 122.3 —> 2
- 122.3 <= Admission grade < 129.4 —> 3
- 129.4 <= Admission grade < 138.3 —> 4
- Admission grade >= 138.3 —> 5



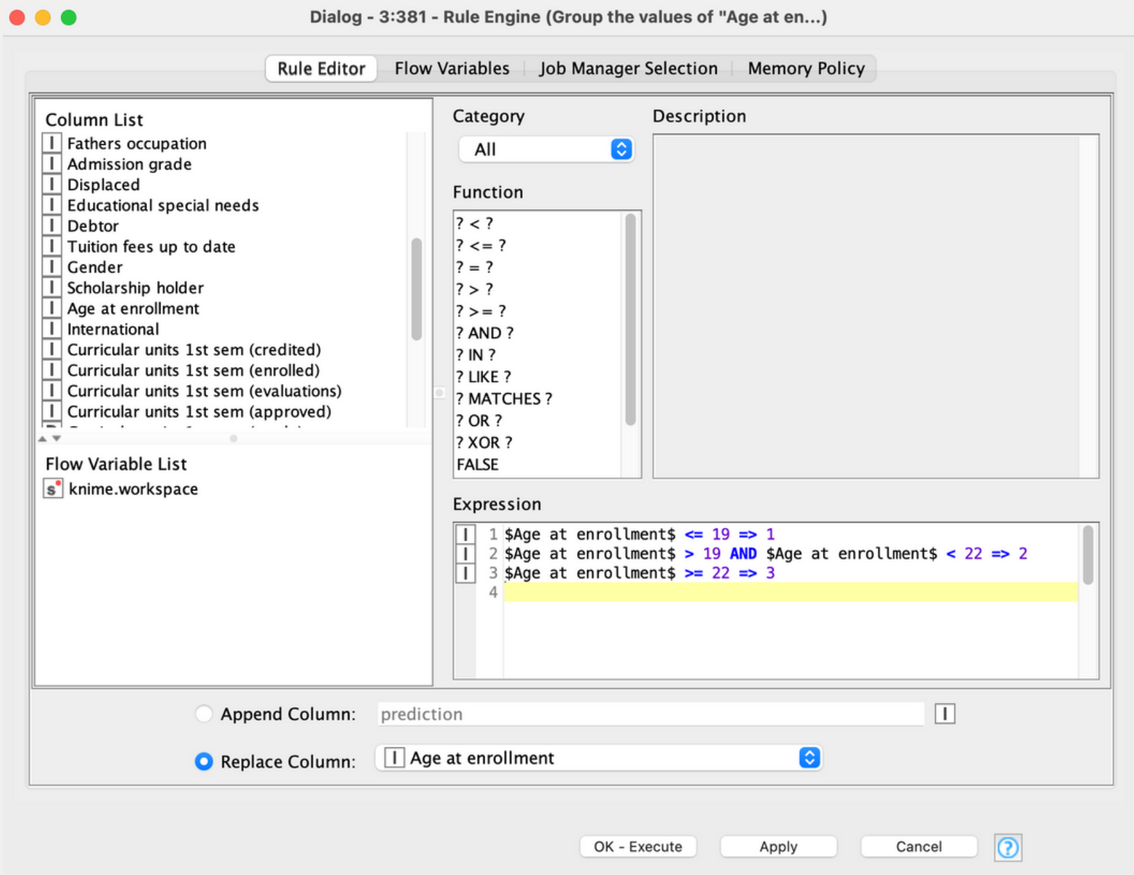
The cut-off values we use are, respectively, the 20th, 40th, 60th, and 80th percentile. This is done in order to ensure an equal number of students within each category, which in turn makes the creation of categories meaningful.

Variable encoding -**Age at enrollment**

We proceed similarly for “Age at enrollment”, but this time we employ the 33th and the 66th as the cut-off values for our newly-created categories. These values correspond to, respectively, 19 and 22 years.

So, the scheme is:

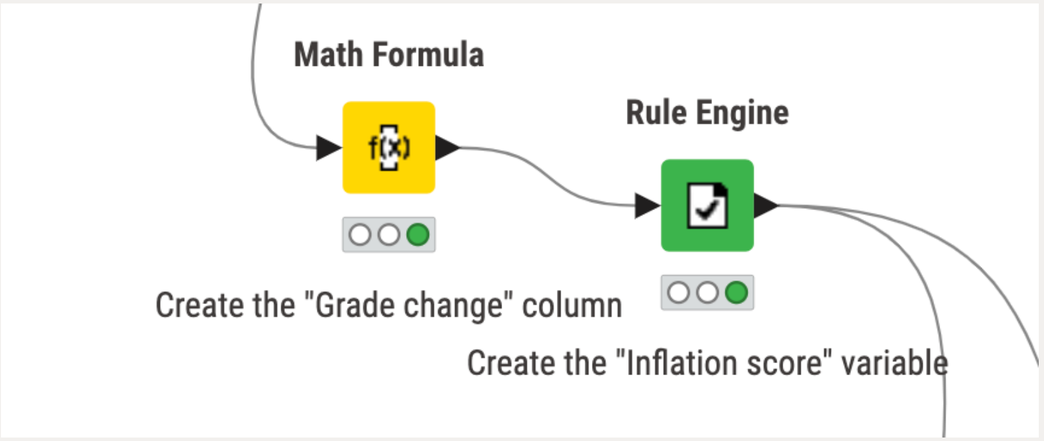
- Age <= 19 —> 1
- 19 < Age < 22 —> 2
- Age >= 22 —> 3



The rationale is that there is probably a big difference between, for instance, 18-year old and 22-year old enrolling students, but if a student enrols at 40 years old, that does not necessarily make them differ substantially from another student enrolling at 47 years old.

Creating new variables

A fundamental task in feature engineering is being able to use the variables at hand to extract valuable information and to therefore create new variables from existing ones. In the following slides we propose our approach in this regards and, specifically, why and how we create a “Grade change” and an “Inflation score” variable.



Creating new variables -**Grade change**

Sometimes for students the trends of grades has a greater influence on future academic performance than actual GPA. This is why we create a new variable called “Grade change”, which is the difference between the grade each student obtained in the second semester and in the first one. This newly-created variable has a 20% correlation with “Target”.

Unemplo... Number (dou...	Inflation ... Number (dou...	GDP Number (dou...	Target Number (inte...	Grade ch... Number (dou...
10.8	1.4	1.74	-1	0
13.9	-0.3	0.79	1	-0.333
10.8	1.4	1.74	-1	0
9.4	-0.8	-3.12	1	-1.029
13.9	-0.3	0.79	1	0.667
16.2	0.3	-0.92	1	-0.357
15.5	2.8	-4.06	1	1.045
15.5	2.8	-4.06	-1	0

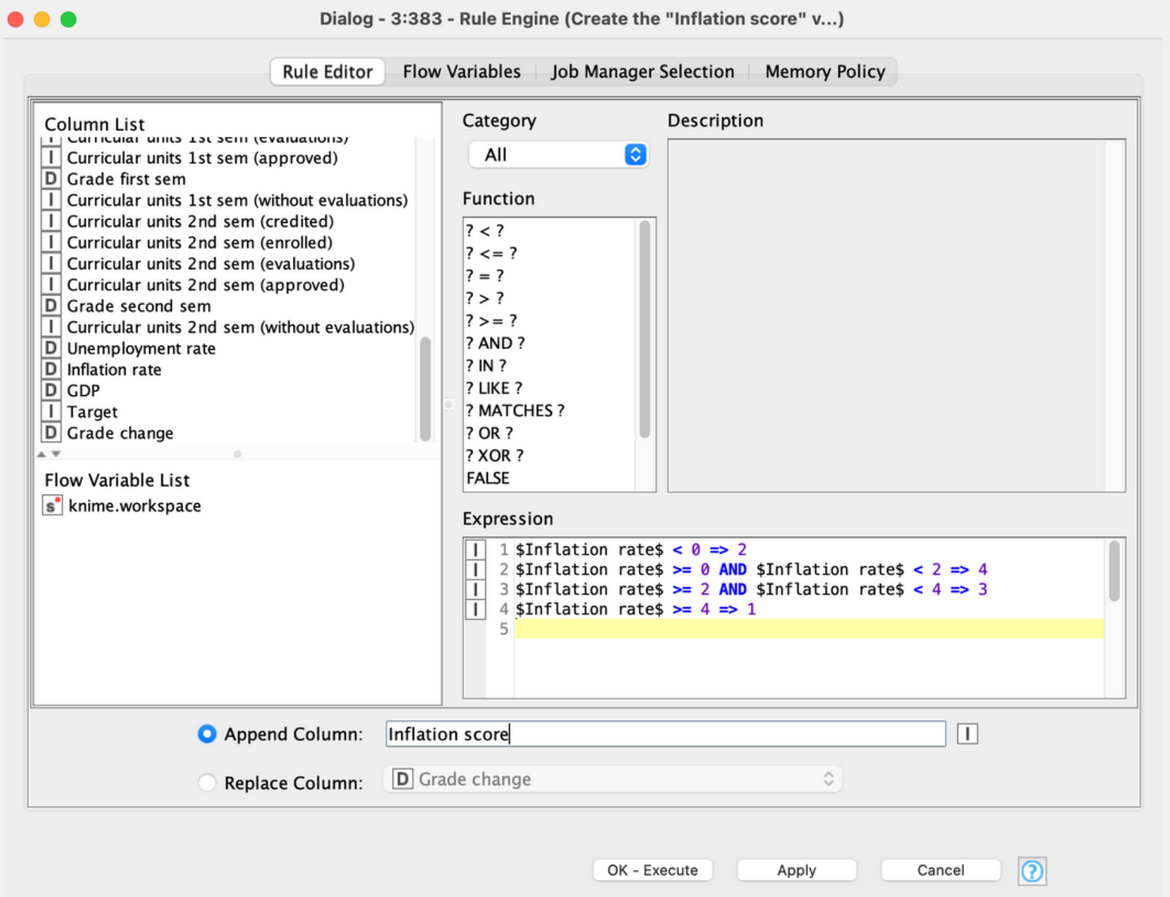
Creating new variables -**Inflation score**

We also experiment with creating another variable, called “Inflation score”, by grouping the values of “Inflation rate” in five categories, namely:

- high inflation (inflation rate greater than 4%), 1
- deflation (negative inflation), 2
- moderate inflation (inflation rate between 2% and 4%), 3
- low inflation (inflation rate between 0% and 2%), 4

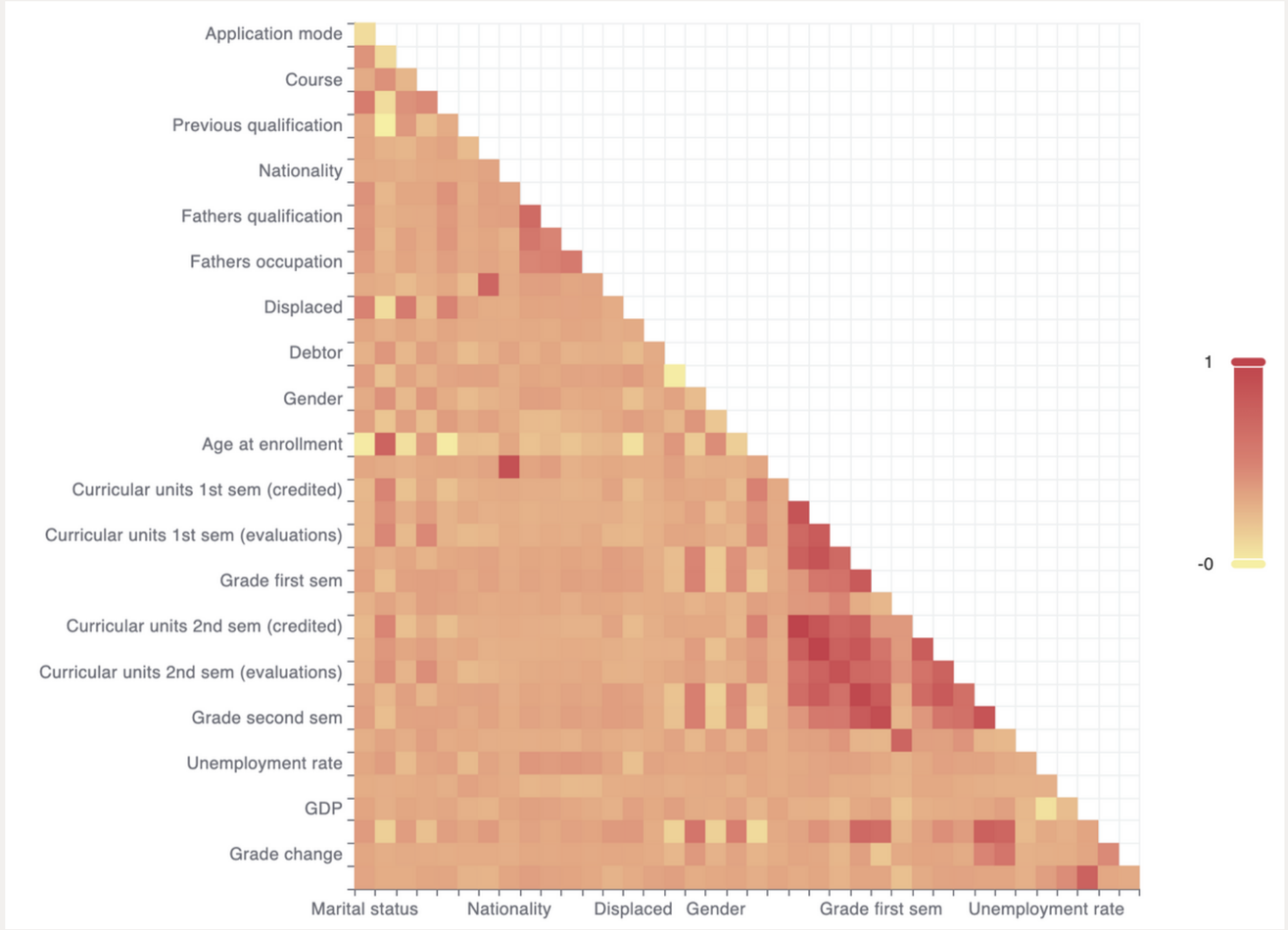
The most preferable category (meaning, the one that should benefit the job market more) is low inflation, followed by moderate inflation then deflation, and finally high inflation (the worst category). The numbers assigned depend on the expected benefits on the job market.

Note that including the last category high inflation rate) has no effect on our data, since the maximum inflation rate in our sample is 3.7%. However, we include it for demonstration purposes (it would surely turn out to be useful in a larger dataset).



Feature selection

For feature selection, we start by plotting a correlation matrix of our dataset (we only plot the lower-diagonal part of it, since this is a symmetric matrix and plotting it all would just yield a second part specular to the one shown here).

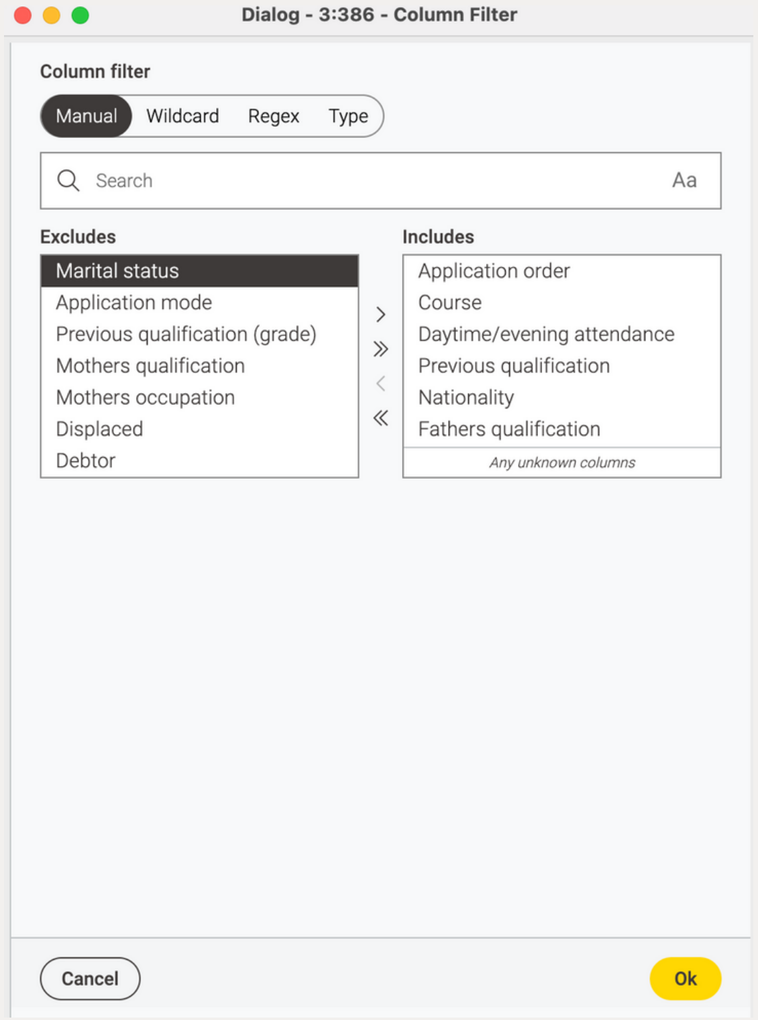


We notice several highly-correlated variables. To avoid multicollinearity, we remove the following predictors from our dataset:

"Marital status", “Displaced”, "Application mode", "Previous qualification (grade)", "Mother's occupation", "Mother's qualification", "International", "Debtor", "Unemployment rate", “Inflation score”, "Curricular units 1st sem (credited)", "Curricular units 1st sem (enrolled)", "Curricular units 1st sem (evaluations)", "Curricular units 1st sem (approved)", "Curricular units 2nd sem (credited)", "Curricular units 2nd sem (enrolled)", "Curricular units 2nd sem (evaluations)", "Curricular units 2nd sem (approved).

Feature selection

With a *Column Filter* node, we eliminate the previously-mentioned columns.



There were many variables dealing with academic units, so we decided to keep only a few of them. Moreover, there were some variables with clearly-repeating information (for instance, there was no reason to keep both “International” and “Nationality”, since 96% of the students in our sample are Portuguese).

Data description

BIVARIATE ANALYSIS

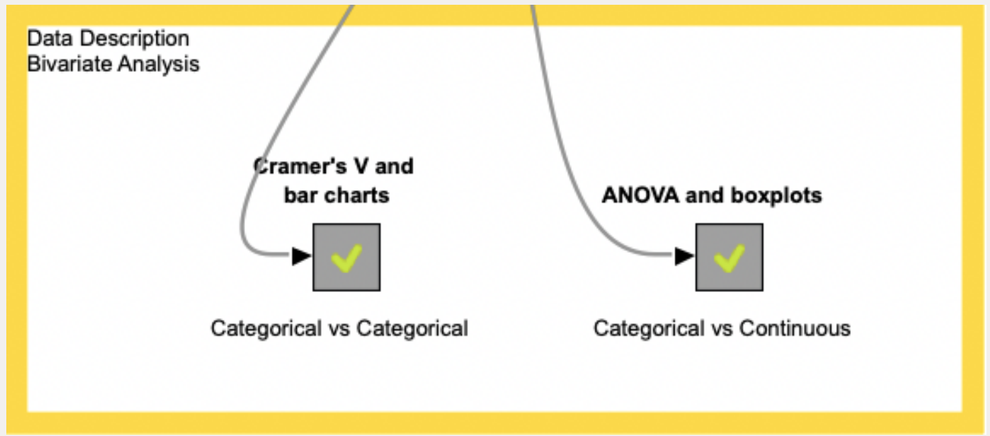
Overview of bivariate analysis

In this section we analyze each predictor together with “Target”, to identify cases of low relevancy of some variables.

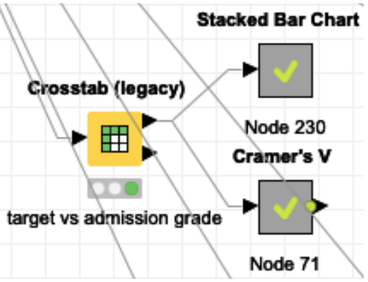
Specifically, for each categorical predictor, we run a Chi-Square test, and we compute its Cramer’s V with “Target” and correlation with “Target”.

Based on the results we obtained, we decided to eliminate four additional variables from our dataset.

Subsequently, we computed the One-Way ANOVA between each continuous predictor and “Target”, and we showed both the p-value and the f-value.



Reading the output of the Chi-Square test



The following few slides display the output of some “Crosstab” nodes, used to run the Chi-Square test. The aim of this section is to compute some statistical measures to assess whether a categorical variable is significantly associated with our “Target” variable.

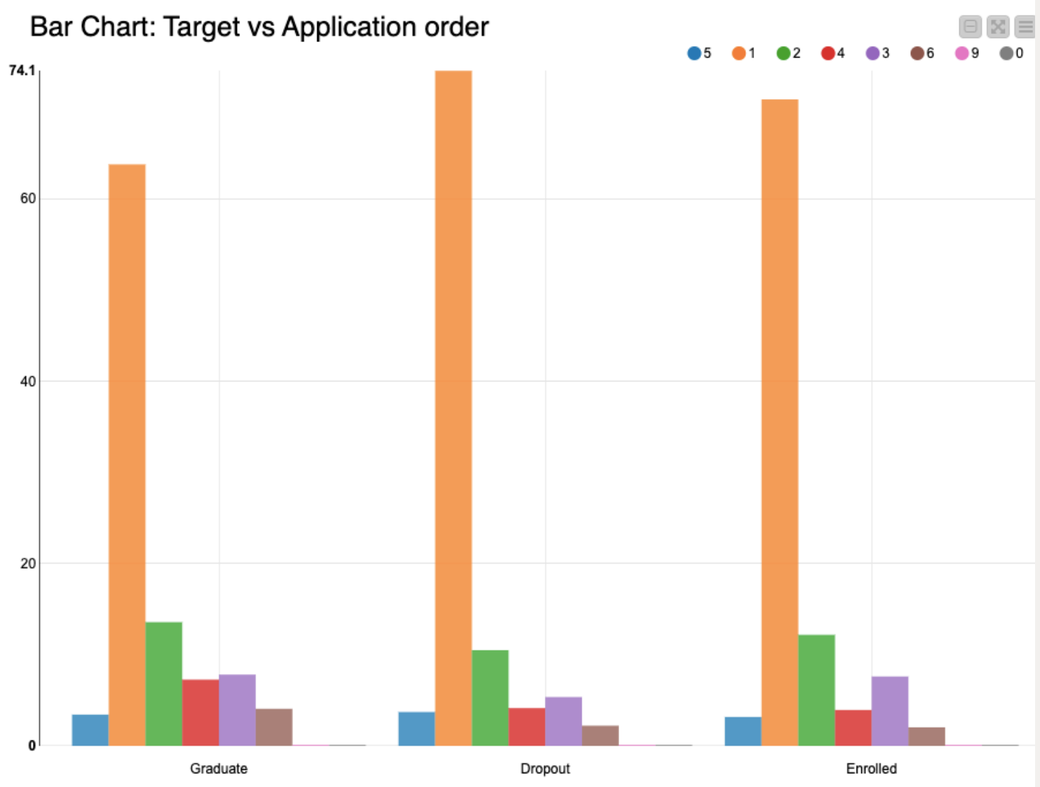
The first table is a cross-tabulation table, and it displays the observed frequency of occurrences for each combination of the categories of the two variables. For each possible couple of categories, we have information regarding the actual frequency, the expected frequency (in case of no relation between the two variables), and some percentages that place the observed frequencies in context:

- Percent: The percentage that this cell's observed frequency represents out of the total number of observations.
- Row Percent: The percentage of the row total that this cell's observed frequency represents.
- Column Percent: The percentage of the column total that this cell's observed frequency represents.
- Cell Chi-Square: A value that contributes to the overall Chi-Square statistic, indicating how much the observed frequency deviates from the expected frequency.

The second table instead shows Chi-Square: (the sum of all the individual cell Chi-Square values, which indicates the overall discrepancy between the observed and expected frequencies), DF (the degrees of freedom for the test), Value (the actual value of the Chi-Square statistic), and Prob (the p-value, which is the probability of observing a Chi-square statistic at least as extreme as the observed value).

To this table we also add the Cramer’s V between the two variables analyzed (computed with the Chi-Square statistic and degrees of freedom coming from the table), and the correlation between the two variables (computed in the previous section).

Bivariate analysis of categorical variables -**Application order**



Cross Tabulation of Target by Application order

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	2	3	4	5	6	9	Total
-1		1,037	147	75	58	52	31		1,400
		957.2212	172.7605	97.8976	79.3419	48.6289	43.5101		
		23.6974%	3.3592%	1.7139%	1.3254%	1.1883%	0.7084%		31.9927%
		74.0714%	10.5%	5.3571%	4.1429%	3.7143%	2.2143%		
		6.6491	3.8412	5.3556	5.7407	0.2337	3.5969		
0		559	96	60	31	25	16	1	788
		538.7788	97.2395	55.1024	44.6581	27.3711	24.4899	0.1801	
		12.7742%	2.1938%	1.3711%	0.7084%	0.5713%	0.3656%	0.0229%	18.0073%
		70.9391%	12.1827%	7.6142%	3.934%	3.1726%	2.0305%	0.1269%	
		0.7589	0.0158	0.4353	4.1772	0.2054	2.9432	3.7334	
1	1	1,396	297	171	159	75	89		2,188
	0.5	1,496	270	153	124	76	68		
	0.0229%	31.9013%	6.787%	3.9077%	3.6335%	1.7139%	2.0338%		50%
	0.0457%	63.8026%	13.574%	7.8154%	7.2669%	3.4278%	4.0676%		
	0.5	6.6845	2.7	2.1176	9.879	0.0132	6.4853		
Total	1	2,992	540	306	248	152	136	1	4,376
	0.0229%	68.3729%	12.34%	6.9927%	5.6673%	3.4735%	3.1079%	0.0229%	100%

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Application order

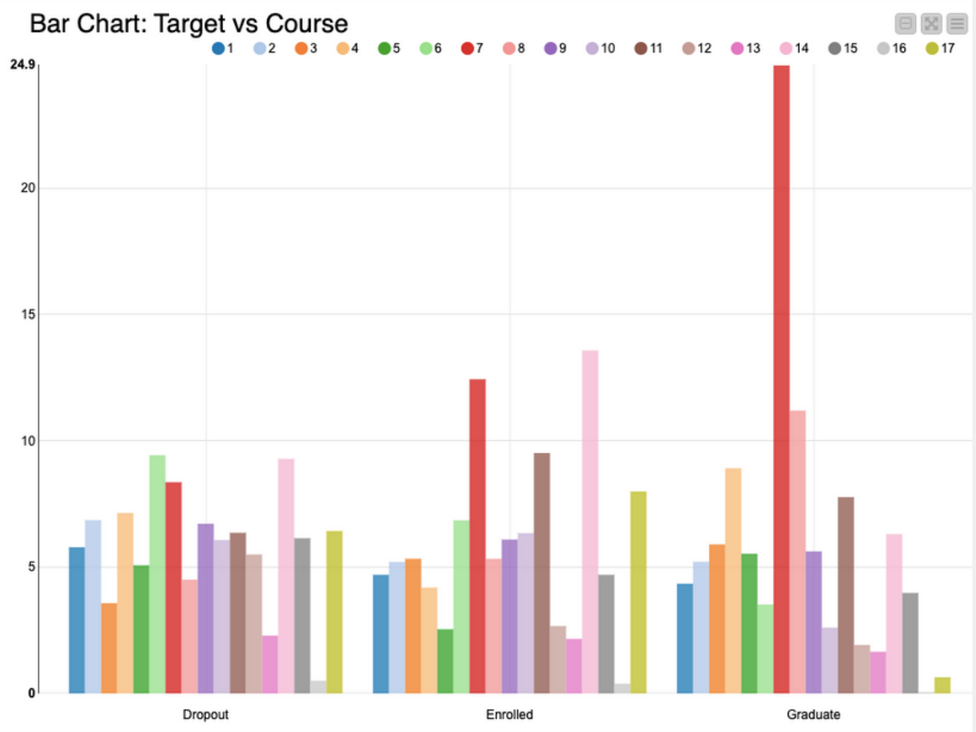
Statistic	DF	Value	Prob
Chi-Square	14	66.066	9.86E-9

Total sample size: 4376.0

Cramer’s V = 0.07093976 (weak)

Correlation = 0.088

Bivariate analysis of categorical variables -**Course**



Frequency Expected Percent Row Percent Cell Chi-Square	1	2	3	4	5	6	7	8	9	10	... (7)	Total
-1	81	96	50	100	71	132	117	63	94	85		1,400
	68.1444	80.3016	70.7038	104.936	67.8245	84.1408	242.8245	111.9744	84.7806	61.426		
	1.851%	2.1938%	1.1426%	2.2852%	1.6225%	3.0165%	2.6737%	1.4397%	2.1481%	1.9424%		31.9927%
	5.7857%	6.8571%	3.5714%	7.1429%	5.0714%	9.4286%	8.3571%	4.5%	6.7143%	6.0714%		
	2.4252	3.0689	6.0626	0.2322	0.1487	27.2223	65.1985	21.42	1.0026	9.0472		
0												788
												18.0073%
1												2,188
												50%
Total	213	251	221	328	212	263	759	350	265	192		4,376
	4.8675%	5.7358%	5.0503%	7.4954%	4.8446%	6.0101%	17.3446%	7.9982%	6.0558%	4.3876%		100%

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Course

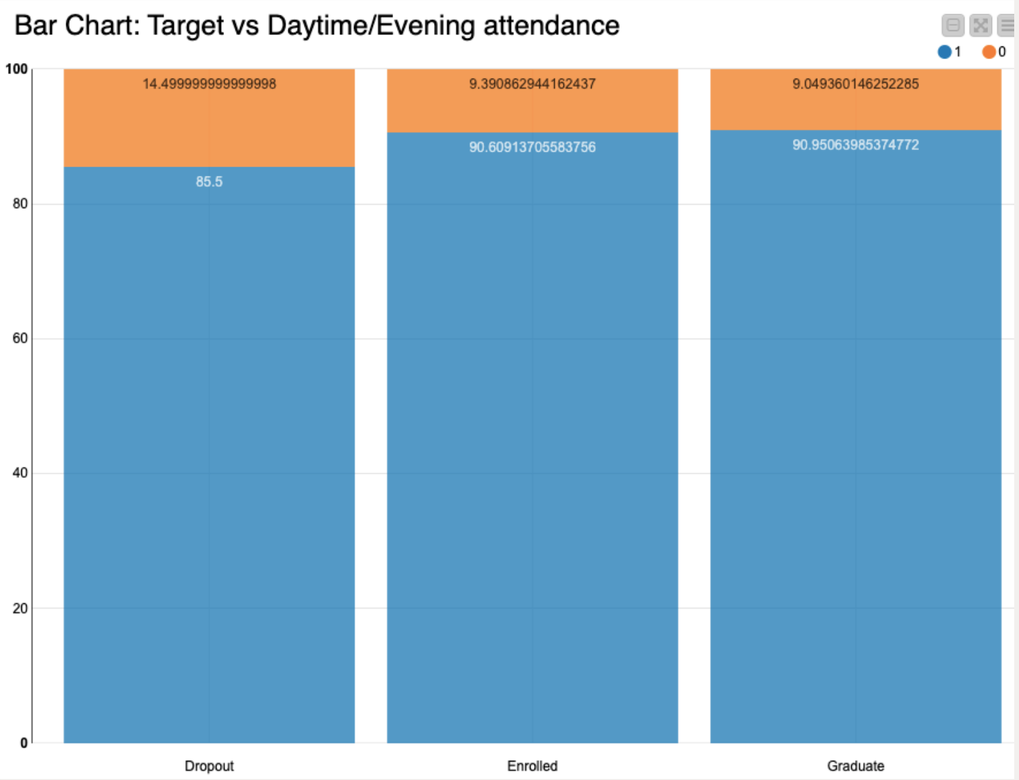
Statistic	DF	Value	Prob
Chi-Square	32	549.8516	1.25E-95

Total sample size: 4376.0

Cramer’s V = 0.20465552 (strong)

Correlation = -0.129

Bivariate analysis of categorical variables -**Daytime/ evening attendance**



Cross Tabulation of Target by Daytime/evening attendance

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	Total
-1	203	1,197	1,400
	151.9653	1,248.0347	
	4.6389%	27.3537%	31.9927%
	14.5%	85.5%	
	17.1391	2.0869	
0	74	714	788
	85.5347	702.4653	
	1.691%	16.3163%	18.0073%
	9.3909%	90.6091%	
	1.5555	0.1894	
1	198	1,990	2,188
	237.5	1,950.5	
	4.5247%	45.4753%	50%
	9.0494%	90.9506%	
	6.5695	0.7999	
Total	475	3,901	4,376
	10.8547%	89.1453%	100%

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Daytime/evening attendance

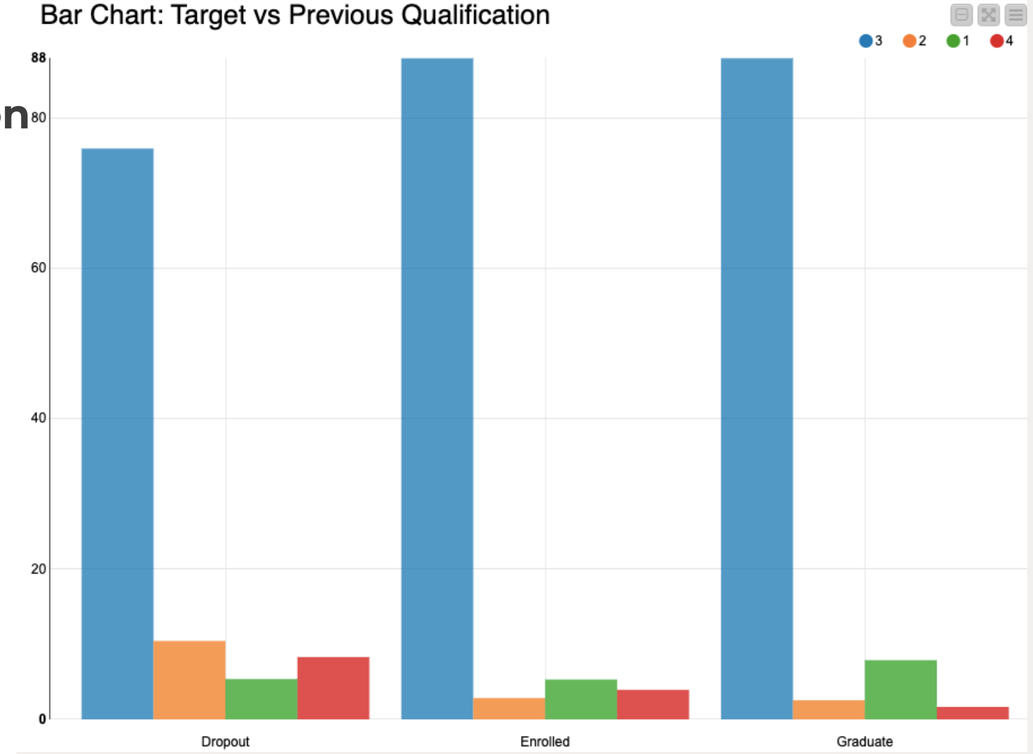
Statistic	DF	Value	Prob
Chi-Square	2	28.3403	7.01E-7

Total sample size: 4376.0

Cramer’s V = 0.05690476 (weak)

Correlation = 0.075

Bivariate analysis of categorical variables -**Previous qualification**



Cross Tabulation of Target by Previous qualification

Frequency Expected Percent Row Percent Cell Chi-Square	1	2	3	4	Total
-1	75	146	1,063	116	1,400
	80.9415	72.9433	1,177.3309	68.7843	
	1.7139%	3.3364%	24.2916%	2.6508%	31.9927%
	5.3571%	10.4286%	75.9286%	8.2857%	
	0.4361	73.1702	11.1027	32.4104	
0	62	20	693	13	788
	45.5585	41.0567	662.6691	38.7157	
	1.4168%	0.457%	15.8364%	0.2971%	18.0073%
	7.868%	2.5381%	87.9442%	1.6497%	
	5.9335	10.7993	1.3883	17.0809	
1	116	62	1,924	86	2,188
	126.5	114	1,840	107.5	
	2.6508%	1.4168%	43.9671%	1.9653%	50%
	5.3016%	2.8336%	87.9342%	3.9305%	
	0.8715	23.7193	3.8348	4.3	
Total	253	228	3,680	215	4,376
	5.7815%	5.2102%	84.0951%	4.9132%	100%

- [x] Frequency
- [x] Expected
- [ ] Deviation
- [x] Percent
- [x] Row Percent
- [ ] Column Percent
- [x] Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Previous qualification

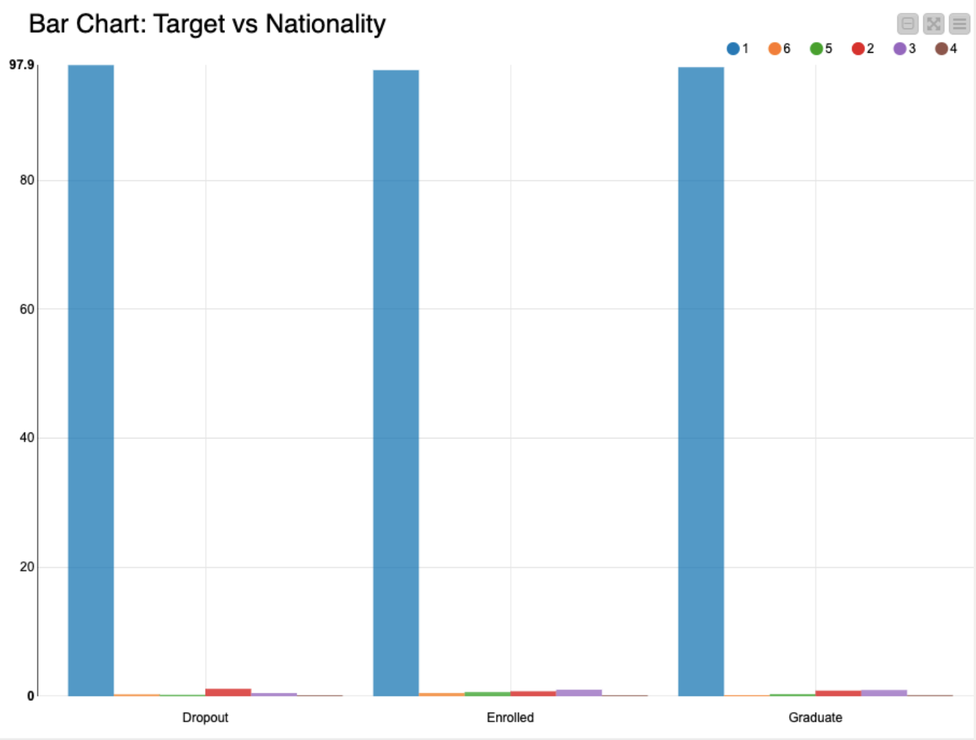
Statistic	DF	Value	Prob
Chi-Square	6	185.047	2.87E-37

Total sample size: 4376.0

Cramer’s V = 0.11872488 (medium)

Correlation = 0.03

Bivariate analysis of categorical variables -**Nationality**



Cross Tabulation of Target by Nationality

Frequency Expected Percent Row Percent Cell Chi-Square	1.0	2.0	6.0	11.0	13.0	14.0	17.0	21.0	22.0	24.0	... (10)	Total
-1	1,370		3				1	1	4	1		1,400
	1,365.7678		3.8391				0.3199	0.6399	4.159	1.5996		
	31.3071%		0.0686%				0.0229%	0.0229%	0.0914%	0.0229%		31.9927%
	97.8571%		0.2143%				0.0714%	0.0714%	0.2857%	0.0714%		
	0.0131		0.1834				1.4456	0.2027	0.0061	0.2248		
0												788
												18.0073%
1												2,188
												50%
Total	4,269	2	12	3	1	1	1	2	13	5		4,376
	97.5548%	0.0457%	0.2742%	0.0686%	0.0229%	0.0229%	0.0229%	0.0457%	0.2971%	0.1143%		100%

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Nationality

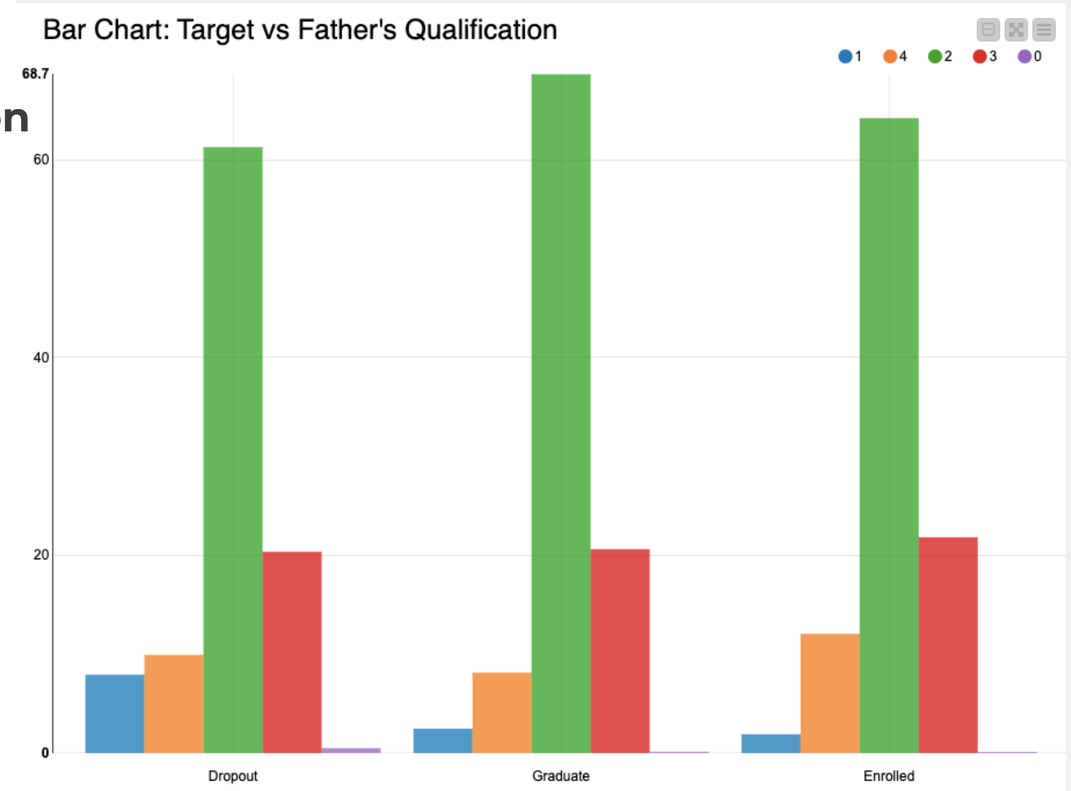
Statistic	DF	Value	Prob
Chi-Square	38	31.9362	0.745

Total sample size: 4376.0

Cramer’s V = 0.02932219 (weak)

Correlation = -0.01

Bivariate analysis of categorical variables -**Father's qualification**



Cross Tabulation of Target by Fathers qualification

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	2	3	4	Total
-1	7	111	858	285	139	1,400
	3.1993	57.5868	916.9104	290.4936	131.8099	
	0.16%	2.5366%	19.6069%	6.5128%	3.1764%	31.9927%
	0.5%	7.9286%	61.2857%	20.3571%	9.9286%	
	4.5153	49.542	3.7849	0.1039	0.3922	
0		15	506	172	95	788
		32.4132	516.0896	163.5064	74.1901	
		0.3428%	11.5631%	3.9305%	2.1709%	18.0073%
		1.9036%	64.2132%	21.8274%	12.0558%	
		9.3548	0.1973	0.4412	5.837	
1	3	54	1,502	451	178	2,188
	5	90	1,433	454	206	
	0.0686%	1.234%	34.3236%	10.3062%	4.0676%	50%
	0.1371%	2.468%	68.6472%	20.6124%	8.1353%	
	0.8	14.4	3.3224	0.0198	3.8058	
Total	10	180	2,866	908	412	4,376
	0.2285%	4.1133%	65.4936%	20.7495%	9.415%	100%

- [x] Frequency
- [x] Expected
- [ ] Deviation
- [x] Percent
- [x] Row Percent
- [ ] Column Percent
- [x] Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Fathers qualification

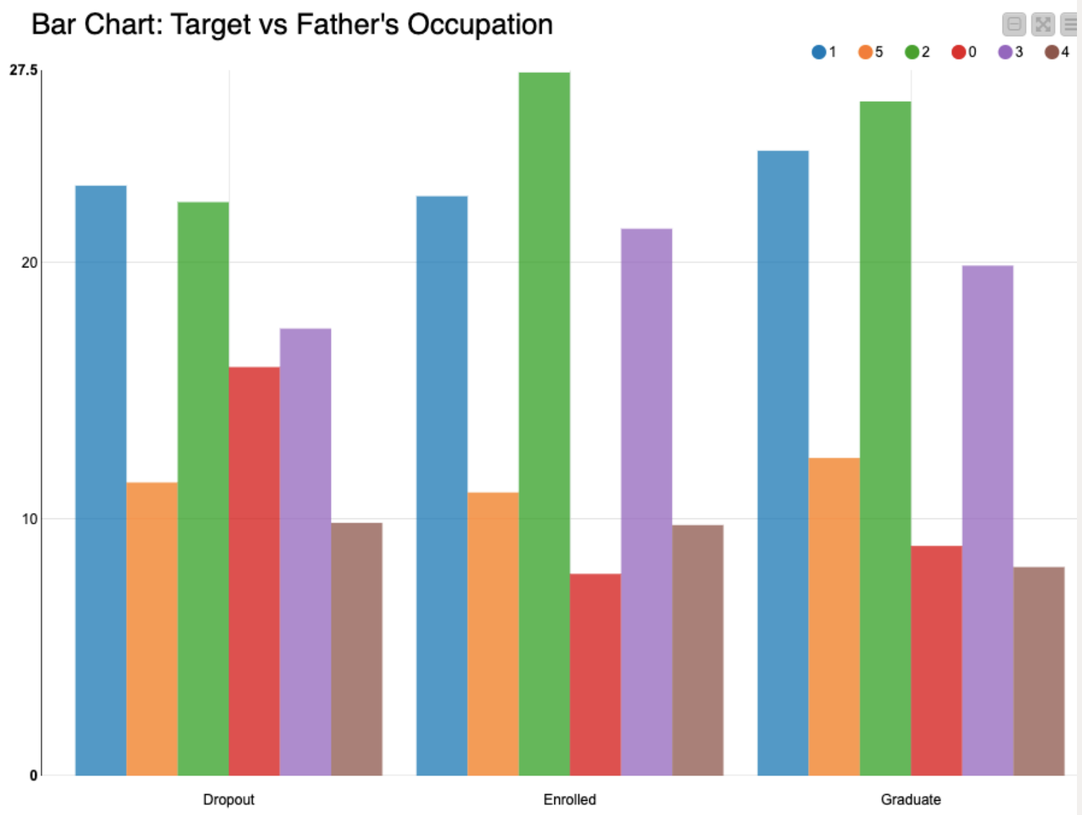
Statistic	DF	Value	Prob
Chi-Square	8	96.5166	2.20E-17

Total sample size: 4376.0

Cramer's V = 0.0857436 (weak)

Correlation = 0.012

Bivariate analysis of categorical variables -**Father's occupation**



Cross Tabulation of Target by Fathers occupation

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	2	3	4	5	Total
-1	223	322	313	244	138	160	1,400
	153.8848	330.4845	353.1993	270.9781	125.7313	165.7221	
	5.096%	7.3583%	7.1527%	5.5759%	3.1536%	3.6563%	31.9927%
	15.9286%	23%	22.3571%	17.4286%	9.8571%	11.4286%	
	31.0421	0.2178	4.5753	2.6859	1.1972	0.1976	
0	62	178	216	168	77	87	788
	86.6152	186.0155	198.8007	152.5219	70.7687	93.2779	
	1.4168%	4.0676%	4.936%	3.8391%	1.7596%	1.9881%	18.0073%
	7.868%	22.5888%	27.4112%	21.3198%	9.7716%	11.0406%	
	6.9954	0.3454	1.488	1.5707	0.5487	0.4225	
1	196	533	575	435	178	271	2,188
	240.5	516.5	552	423.5	196.5	259	
	4.479%	12.1801%	13.1399%	9.9406%	4.0676%	6.1929%	50%
	8.958%	24.3601%	26.2797%	19.8812%	8.1353%	12.3857%	
	8.2339	0.5271	0.9583	0.3123	1.7417	0.556	
Total	481	1,033	1,104	847	393	518	4,376
	10.9918%	23.606%	25.2285%	19.3556%	8.9808%	11.8373%	100%

☑ Frequency
☑ Expected
☐ Deviation
☑ Percent
☑ Row Percent
☐ Column Percent
☑ Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Fathers occupation

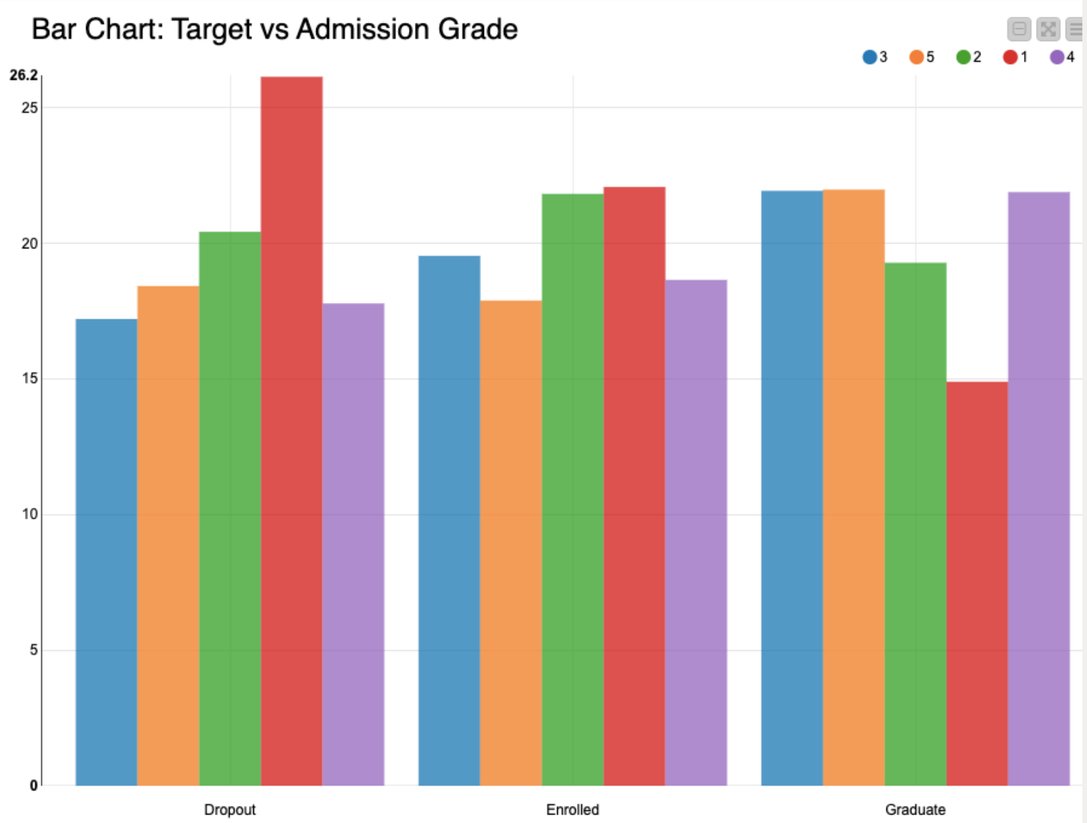
Statistic	DF	Value	Prob
Chi-Square	10	63.6158	7.45E-10

Total sample size: 4376.0

Cramer's V = 0.06961186 (weak)

Correlation = 0.04

Bivariate analysis of categorical variables -**Admission grade**



Cross Tabulation of Target by Admission grade

Frequency Expected Percent Row Percent Cell Chi-Square	1	2	3	4	5	Total
-1	366	286	241	249	258	1,400
	277.0567	281.5356	279.936	279.936	281.5356	
	8.3638%	6.5356%	5.5073%	5.6901%	5.8958%	31.9927%
	26.1429%	20.4286%	17.2143%	17.7857%	18.4286%	
	28.5534	0.0708	5.4156	3.4188	1.9675	
0	174	172	154	147	141	788
	155.9433	158.4644	157.564	157.564	158.4644	
	3.9762%	3.9305%	3.5192%	3.3592%	3.2221%	18.0073%
	22.0812%	21.8274%	19.5431%	18.6548%	17.8934%	
	2.0908	1.1562	0.0806	0.7083	1.9247	
1	326	422	480	479	481	2,188
	433	440	437.5	437.5	440	
	7.4497%	9.6435%	10.9689%	10.9461%	10.9918%	50%
	14.8995%	19.287%	21.9378%	21.8921%	21.9835%	
	26.4411	0.7364	4.1286	3.9366	3.8205	
Total	866	880	875	875	880	4,376
	19.7898%	20.1097%	19.9954%	19.9954%	20.1097%	100%

- [x] Frequency
- [x] Expected
- [ ] Deviation
- [x] Percent
- [x] Row Percent
- [ ] Column Percent
- [x] Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Admission grade

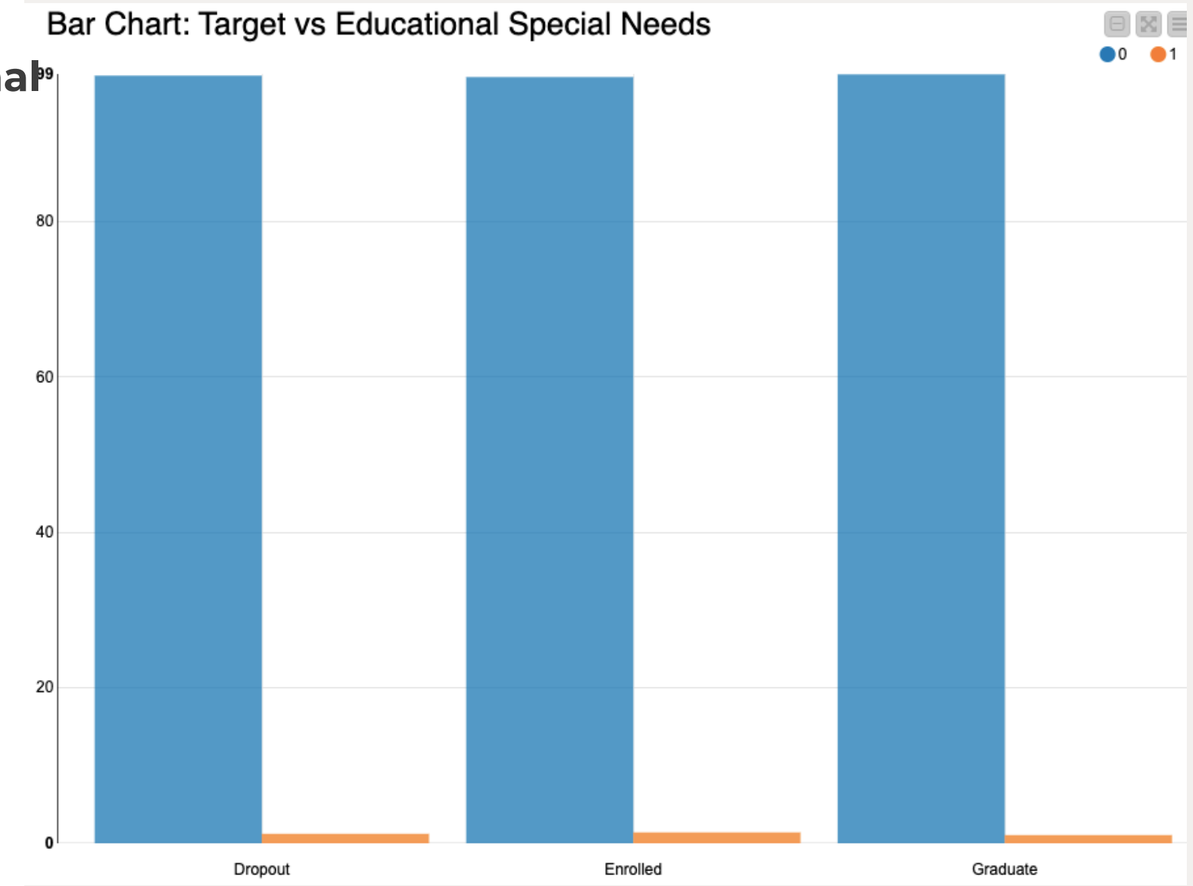
Statistic	DF	Value	Prob
Chi-Square	8	84.4497	6.19E-15

Total sample size: 4376.0

Cramer’s V = 0.08020469 (weak)

Correlation = 0.112

Bivariate analysis of categorical variables -**Educational special needs**



Cross Tabulation of Target by Educational special needs

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	Total
-1	1,383	17	1,400
	1,383.6837	16.3163	
	31.6042%	0.3885%	31.9927%
	98.7857%	1.2143%	
	0.0003	0.0287	
0	777	11	788
	778.8163	9.1837	
	17.7559%	0.2514%	18.0073%
	98.6041%	1.3959%	
	0.0042	0.3592	
1	2,165	23	2,188
	2,162.5	25.5	
	49.4744%	0.5256%	50%
	98.9488%	1.0512%	
	0.0029	0.2451	
Total	4,325	51	4,376
	98.8346%	1.1654%	100%

- [x] Frequency
- [x] Expected
- [ ] Deviation
- [x] Percent
- [x] Row Percent
- [ ] Column Percent
- [x] Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Educational special needs

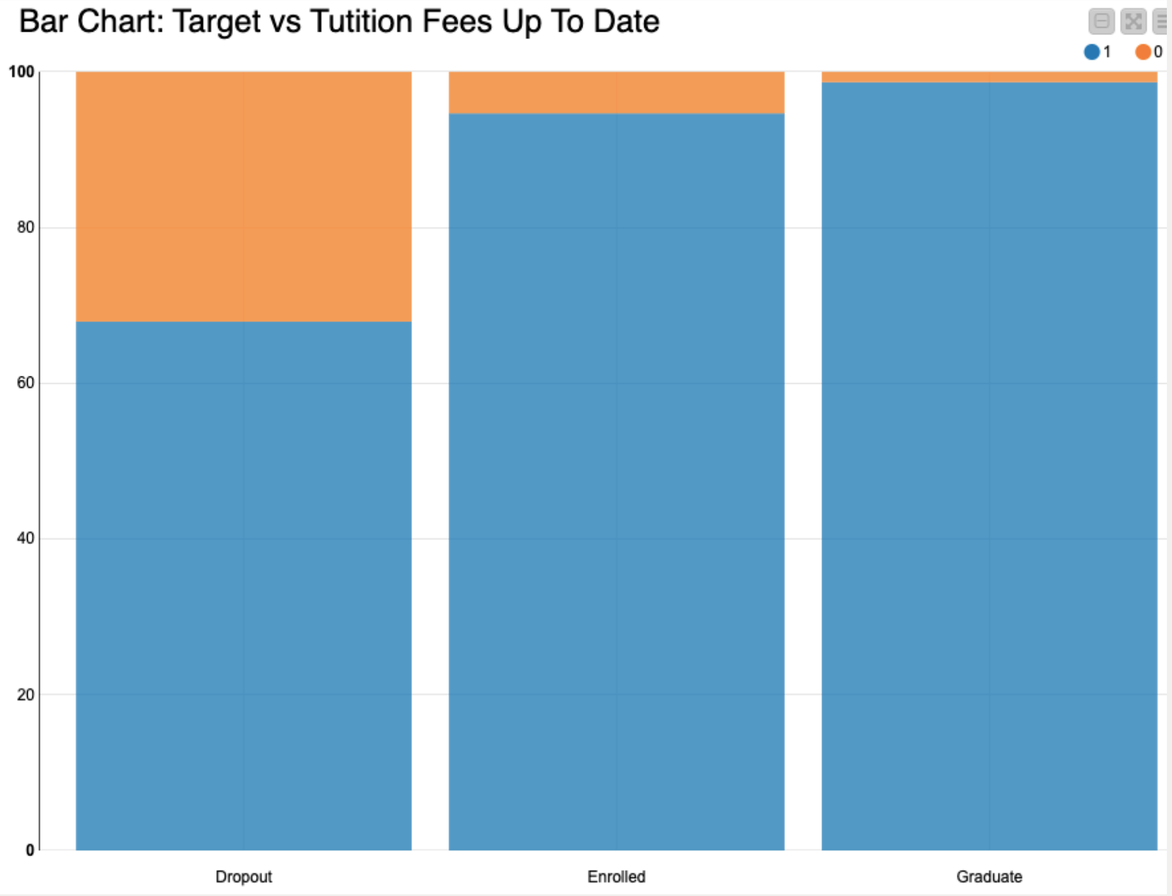
Statistic	DF	Value	Prob
Chi-Square	2	0.6404	0.726

Total sample size: 4376.0

Cramer’s V = 0.00855405 (weak)

Correlation = -0.008

Bivariate analysis of categorical variables -**Tuition fees up to date**



Cross Tabulation of Target by Tuition fees up to date

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	Total
-1	449	951	1,400
	166.362	1,233.638	
	10.2605%	21.7322%	31.9927%
	32.0714%	67.9286%	
	480.1834	64.755	
0	42	746	788
	93.638	694.362	
	0.9598%	17.0475%	18.0073%
	5.3299%	94.6701%	
	28.4765	3.8402	
1	29	2,159	2,188
	260	1,928	
	0.6627%	49.3373%	50%
	1.3254%	98.6746%	
	205.2346	27.6769	
Total	520	3,856	4,376
	11.883%	88.117%	100%

- [x] Frequency
- [x] Expected
- [ ] Deviation
- [x] Percent
- [x] Row Percent
- [ ] Column Percent
- [x] Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Tuition fees up to date

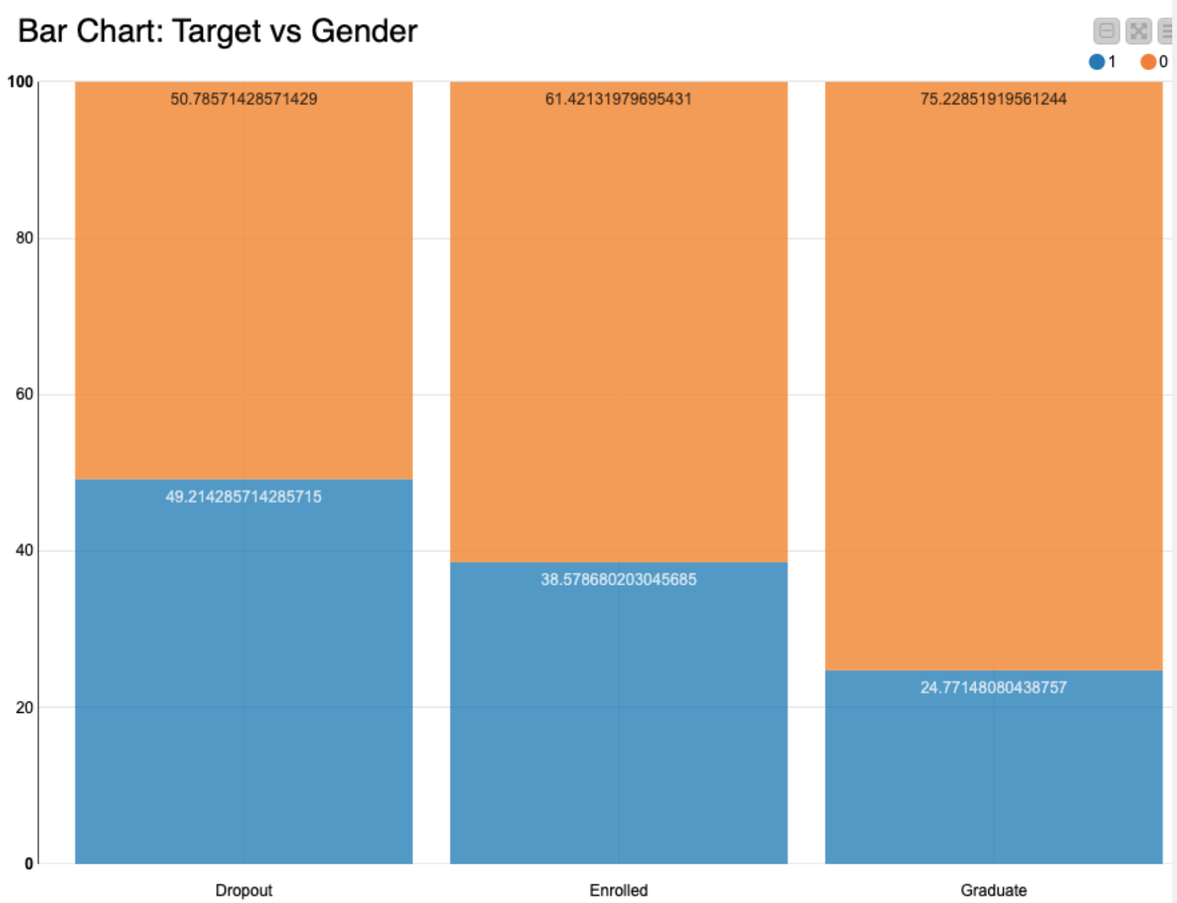
Statistic	DF	Value	Prob
Chi-Square	2	810.1666	1.19E-176

Total sample size: 4376.0

Cramer’s V = 0.30425205 (strong)

Correlation = 0.409

Bivariate analysis of categorical variables -**Gender**



Cross Tabulation of Target by Gender

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	Total
-1	711	689	1,400
	908.9122	491.0878	
	16.2477%	15.745%	31.9927%
	50.7857%	49.2143%	
	43.0947	79.7602	
0	484	304	788
	511.5878	276.4122	
	11.0603%	6.947%	18.0073%
	61.4213%	38.5787%	
	1.4877	2.7534	
1	1,646	542	2,188
	1,420.5	767.5	
	37.6143%	12.3857%	50%
	75.2285%	24.7715%	
	35.7974	66.2544	
Total	2,841	1,535	4,376
	64.9223%	35.0777%	100%

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Gender

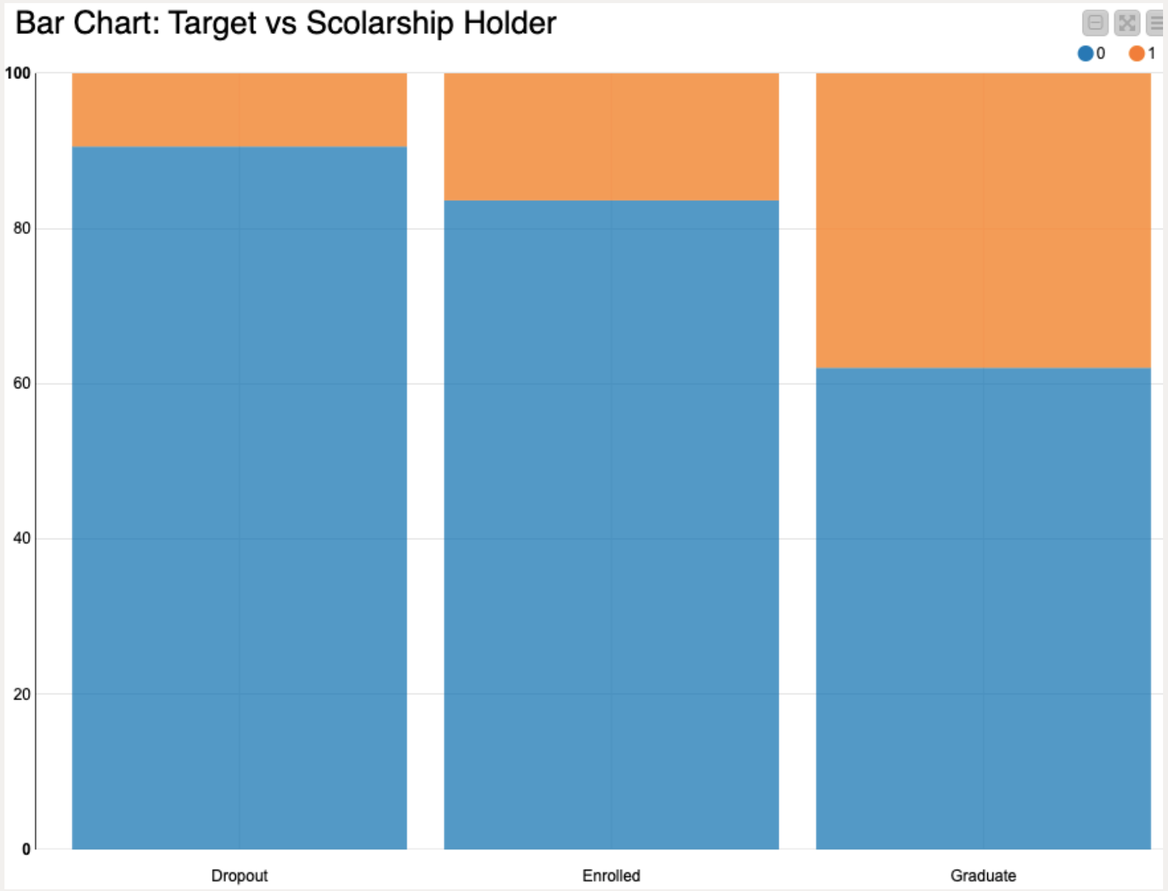
Statistic	DF	Value	Prob
Chi-Square	2	229.1478	1.74E-50

Total sample size: 4376.0

Cramer’s V = 0.16180957 (medium)

Correlation = -0.228

Bivariate analysis of categorical variables -**Scholarship holder**



Cross Tabulation of Target by Scholarship holder

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	Total
-1	1,268	132	1,400
	1,050.9598	349.0402	
	28.9762%	3.0165%	31.9927%
	90.5714%	9.4286%	
	44.8223	134.96	
0	659	129	788
	591.5402	196.4598	
	15.0594%	2.9479%	18.0073%
	83.6294%	16.3706%	
	7.6932	23.1641	
1	1,358	830	2,188
	1,642.5	545.5	
	31.0329%	18.9671%	50%
	62.0658%	37.9342%	
	49.2787	148.3781	
Total	3,285	1,091	4,376
	75.0686%	24.9314%	100%

- [x] Frequency
- [x] Expected
- [ ] Deviation
- [x] Percent
- [x] Row Percent
- [ ] Column Percent
- [x] Cell Chi-Square

Max rows: 10

Max columns: 10

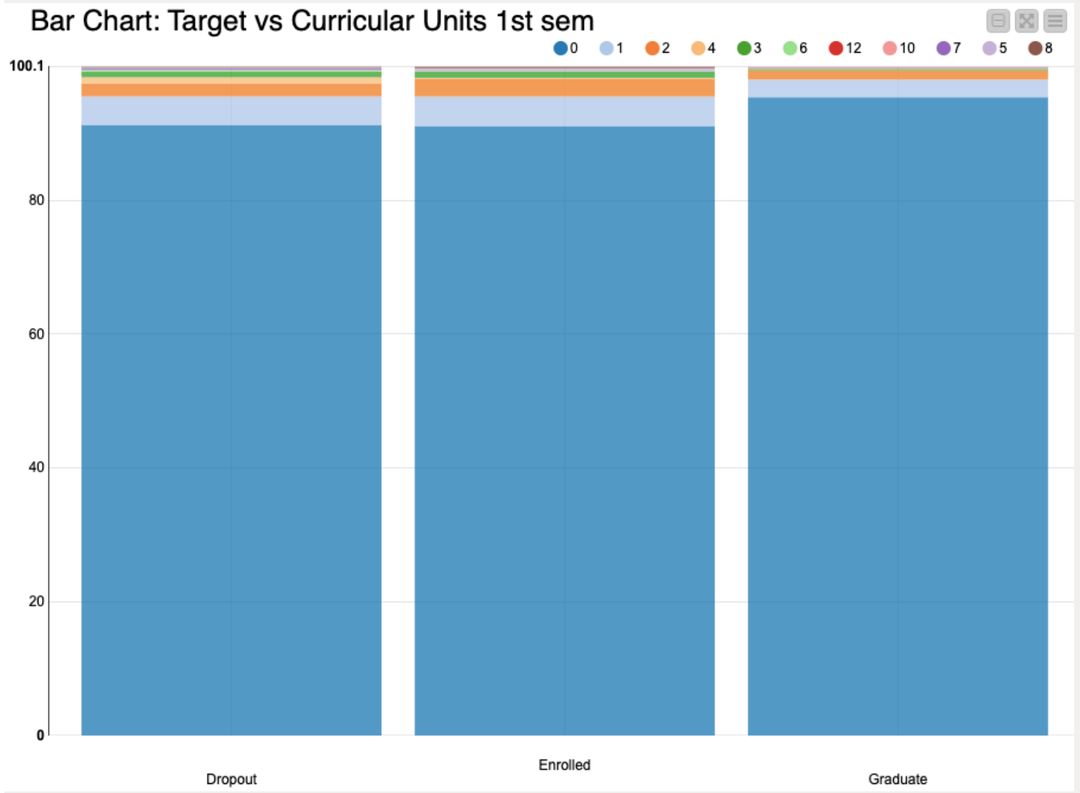
Statistics for Table of Target by Scholarship holder

Statistic	DF	Value	Prob
Chi-Square	2	408.2964	2.19E-89

Cramer's V = 0.21599024 (strong)

Correlation = 0.299

Bivariate analysis of categorical variables -**Curricular units 1st sem (without evaluation)**



Cross tabulation - 3:387 - Crosstab (Chi-Square test)

File

Cross Tabulation of Target by Curricular units 1st sem (without evaluations)

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	2	3	4	5	6	7	8	10	... (1)	Total
-1	1,278	60	27	12	13	1	3	4	2			1,400
	1,306.5814	48.9488	24.6344	7.3583	4.7989	1.5996	1.9196	1.9196	1.2797			
	29.2048%	1.3711%	0.617%	0.2742%	0.2971%	0.0229%	0.0686%	0.0914%	0.0457%			31.9927%
	91.2857%	4.2857%	1.9286%	0.8571%	0.9286%	0.0714%	0.2143%	0.2857%	0.1429%			
	0.6252	2.495	0.2272	2.928	14.0153	0.2248	0.6081	2.2548	0.4054			
0	718	35	21	7	1	2	1	1	2			788
	735.4186	27.5512	13.8656	4.1417	2.7011	0.9004	1.0804	1.0804	0.7203			
	16.4077%	0.7998%	0.4799%	0.16%	0.0229%	0.0457%	0.0229%	0.0229%	0.0457%			18.0073%
	91.1168%	4.4416%	2.665%	0.8883%	0.1269%	0.2538%	0.1269%	0.1269%	0.2538%			
	0.4126	2.0139	3.6709	1.9726	1.0713	1.343	0.006	0.006	2.2736			
1	2,088	58	29	4	1	2	2	1		1		2,188
	2,042	76.5	38.5	11.5	7.5	2.5	3	3		0.5		
	47.7148%	1.3254%	0.6627%	0.0914%	0.0229%	0.0457%	0.0457%	0.0229%		0.0229%		50%
	95.4296%	2.6508%	1.3254%	0.1828%	0.0457%	0.0914%	0.0914%	0.0457%		0.0457%		
	1.0362	4.4739	2.3442	4.8913	5.6333	0.1	0.3333	1.3333		0.5		
Total	4,084	153	77	23	15	5	6	6	4	1		4,376
	93.3272%	3.4963%	1.7596%	0.5256%	0.3428%	0.1143%	0.1371%	0.1371%	0.0914%	0.0229%		100%

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Curricular units 1st sem (without evaluations)

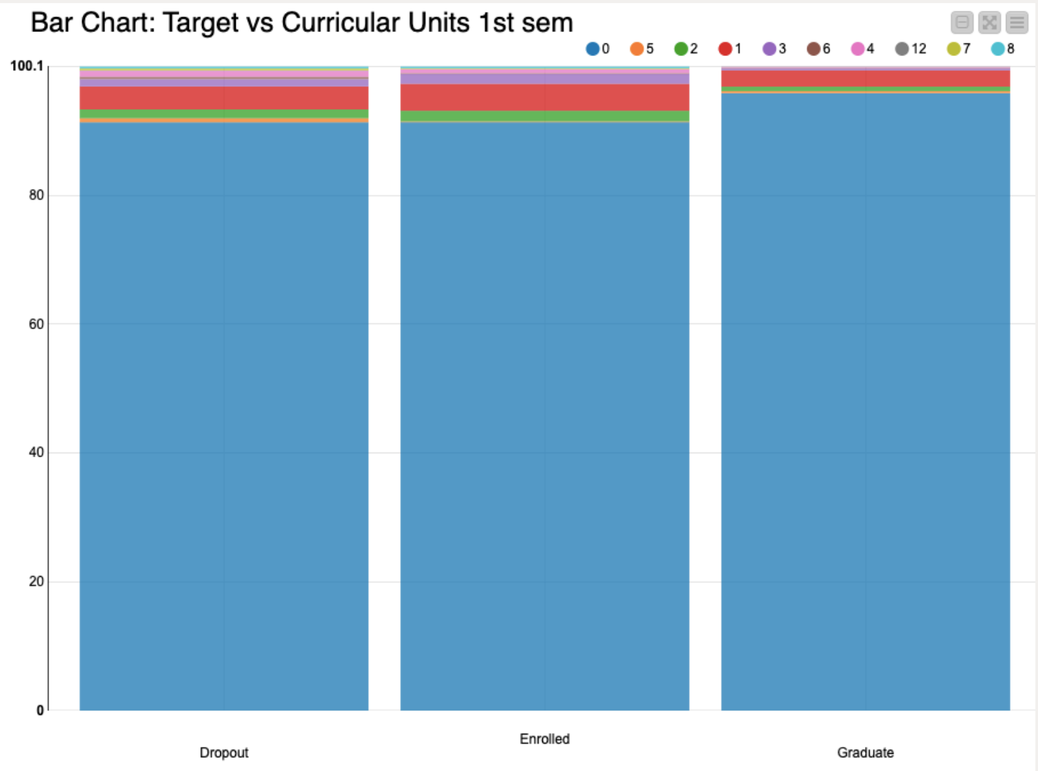
Statistic	DF	Value	Prob
Chi-Square	20	58.1992	1.35E-5

Total sample size: 4376.0

Cramer’s V = 0.06658237, weak

Correlation = 0.009

Bivariate analysis of categorical variables -**Curricular units 2nd sem (without evaluation)**



Cross tabulation - 3:387 - Crosstab (Chi-Square test)

File

Frequency Expected Percent Row Percent Cell Chi-Square	0	1	2	3	4	5	6	7	8	12	Total
-1	1,279	50	19	16	13	9	5	4	4	1	1,400
	1,310.7404	44.1499	15.0366	11.1974	6.7185	5.4388	2.5594	1.5996	1.9196	0.6399	
	29.2276%	1.1426%	0.4342%	0.3656%	0.2971%	0.2057%	0.1143%	0.0914%	0.0914%	0.0229%	31.9927%
	91.3571%	3.5714%	1.3571%	1.1429%	0.9286%	0.6429%	0.3571%	0.2857%	0.2857%	0.0714%	
	0.7686	0.7752	1.0447	2.0598	5.873	2.3319	2.3273	3.6019	2.2548	0.2027	
0	720	33	13	12	5	1	1	1	2		788
	737.7596	24.8501	8.4634	6.3026	3.7815	3.0612	1.4406	0.9004	1.0804		
	16.4534%	0.7541%	0.2971%	0.2742%	0.1143%	0.0229%	0.0229%	0.0229%	0.0457%		18.0073%
	91.3706%	4.1878%	1.6497%	1.5228%	0.6345%	0.1269%	0.1269%	0.1269%	0.2538%		
	0.4275	2.6729	2.4317	5.1504	0.3926	1.3879	0.1347	0.011	0.7826		
1	2,098	55	15	7	3	7	2			1	2,188
	2,048.5	69	23.5	17.5	10.5	8.5	4			1	
	47.9433%	1.2569%	0.3428%	0.16%	0.0686%	0.16%	0.0457%			0.0229%	50%
	95.8867%	2.5137%	0.6856%	0.3199%	0.1371%	0.3199%	0.0914%			0.0457%	
	1.1961	2.8406	3.0745	6.3	5.3571	0.2647	1			0.0	
Total	4,097	138	47	35	21	17	8	5	6	2	4,376
	93.6243%	3.1536%	1.074%	0.7998%	0.4799%	0.3885%	0.1828%	0.1143%	0.1371%	0.0457%	100%

- [x] Frequency
- [x] Expected
- [ ] Deviation
- [x] Percent
- [x] Row Percent
- [ ] Column Percent
- [x] Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Target by Curricular units 2nd sem (without evaluations)

Statistic	DF	Value	Prob
Chi-Square	18	54.6643	1.45E-5

Total sample size: 4376.0

Cramer’s V = 0.06452865, weak

Correlation = -0.032

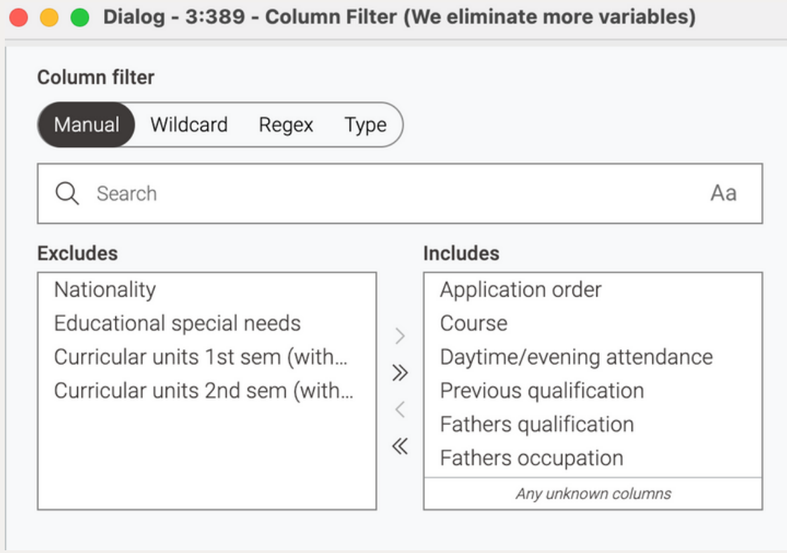
Removing four more variables

Given the results in the previous slides, we opt for removing “Nationality”, “Educational special needs", "Curricular Units 1st sem (without evaluation)", and "Curricular Units 2nd sem (without evaluation)”, as all four of them show a very low Cramer’s V and correlation (in absolute terms) with “Target”.
We interpret the irrelevancies of them as follows:

- Nationality –> almost all students in our sample (96%) are from Portugal. This most likely explains why the nationality of students is, in our case, virtually irrelevant for predicting our target variable.

- Educational special needs –> in many Universities (for instance, at Bocconi), students with special educational needs face different examination conditions than students without special educational needs. The rationale is to have a system in which each student is required (approximately) them same level of effort to reach any given grade. This probably explains the irrelevance of this variable in predicting the outcome of students’ academic careers.

- Curricular Units (without evaluation), for first and second semester –> for the vast majority of students these two values (that of the first and second semester) are 0, and for almost all of them they are anyway very low (the means are indeed 0.14 and 0.15, respectively).



Reading the output of the One-Way ANOVA

The following slides show the output of the One-Way Analysis of Variance (ANOVA), which is a statistical technique used to compare the means of three or more groups to see if there is a statistically significant difference between them. It implemented between a categorical variable (which, in our case, is “Target”) and a continuous variable, and its final purpose is to establish the strength of the relation between the two variables.

The following tables have a section called “Descriptive Statistics”, in which we find a summary of the data for each group in our categorical variable.

We subsequently have the results of a Levene test, with the test statistic, the degrees of freedom of each of the two variables, and the associated p-value.

The third and last section displays the results of the One-Way ANOVA test itself, focusing on variations between groups, within groups, and throughout the whole sample.

Bivariate analysis of continuous variables -**Age at enrollment**

One-way analysis of variance (ANOVA)

Descriptive Statistics
Confidence Interval (CI) Probability: 95.0%

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
Age at enrollment	-1	1400	0	0	2.2664	0.8768	0.0234	2.2205	2.3124	1	3
Age at enrollment	1	2188	0	0	1.6819	0.8239	0.0176	1.6474	1.7164	1	3
Age at enrollment	0	788	0	0	1.9023	0.8521	0.0304	1.8427	1.9619	1	3
Age at enrollment	Total	4376	0	0	1.9086	0.8846	0.0134	1.8824	1.9348	1	3

Levene Test
The Levene Test is used to test for the equality of variances.

	F	df 1	df 2	p-Value
Age at enrollment	14.7667	2	4373	4.06E-7

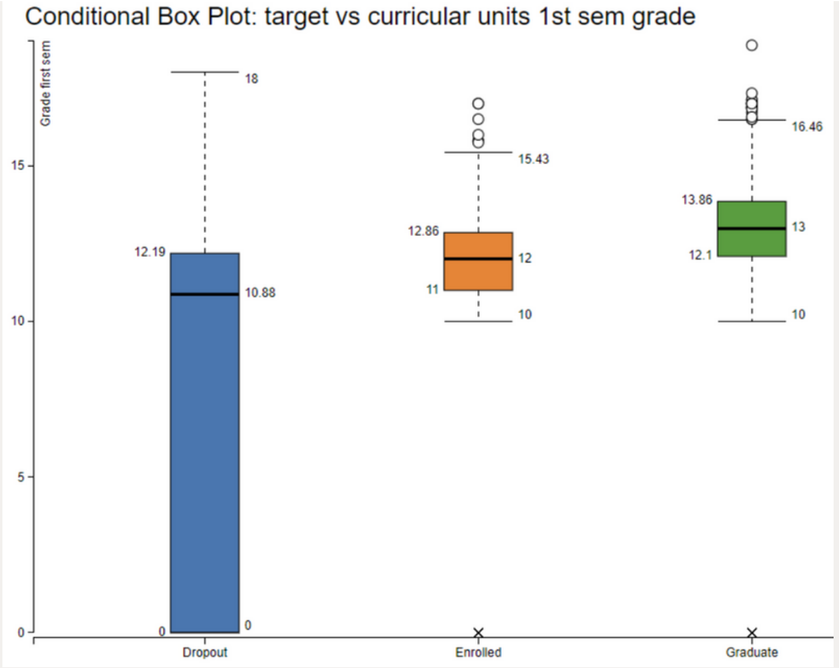
ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
Age at enrollment	Between Groups	291.7356	2	145.8678	203.6848	0.0
Age at enrollment	Within Groups	3,131.7013	4373	0.7161		
Age at enrollment	Total	3,423.4369	4375			

Correlation = -0.29

Note: “Age at enrollment” is actually composed only of integers, but we have included it among the continuous variables as it is numerically meaningful (we cannot see each age as a category).

Bivariate analysis of continuous variables -**Grade first sem**



Correlation = 0.486

One-way analysis of variance (ANOVA)

Descriptive Statistics
Confidence Interval (CI) Probability: 95.0%

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
Grade first sem	-1	1400	0	0	7.2284	6.0379	0.1614	6.9118	7.5449	0.0	18
Grade first sem	1	2188	0	0	12.6393	2.7099	0.0579	12.5257	12.7529	0.0	18.875
Grade first sem	0	788	0	0	11.1329	3.6661	0.1306	10.8765	11.3892	0.0	17
Grade first sem	Total	4376	0	0	10.6369	4.8491	0.0733	10.4932	10.7806	0.0	18.875

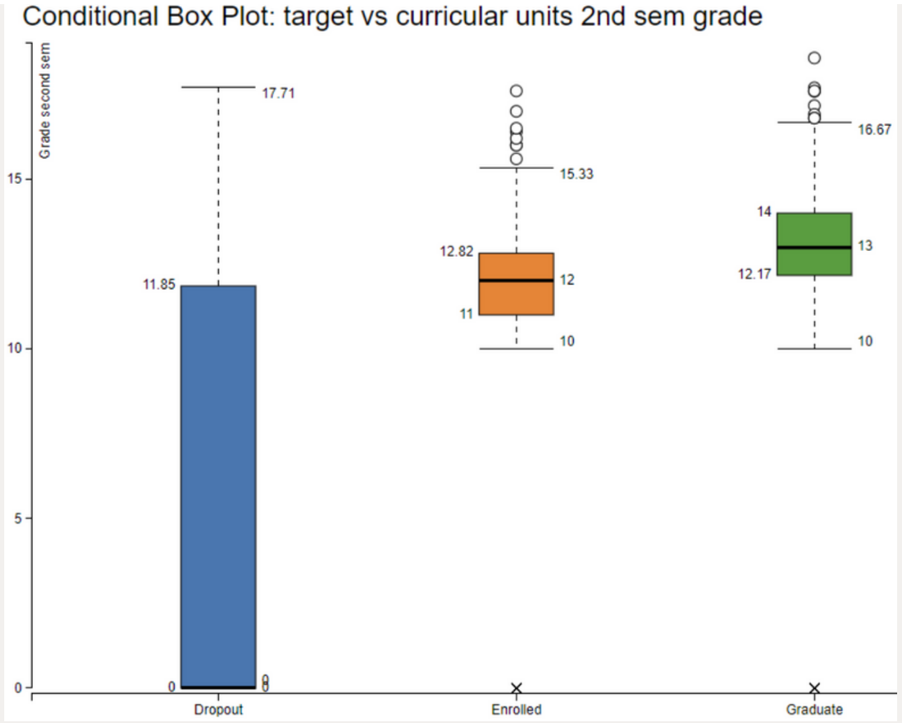
Levene Test
The Levene Test is used to test for the equality of variances.

	F	df 1	df 2	p-Value
Grade first sem	1,715.3209	2	4373	0.0

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
Grade first sem	Between Groups	25,232.3459	2	12,616.173	710.5886	0.0
Grade first sem	Within Groups	77,640.5949	4373	17.7545		
Grade first sem	Total	102,872.9408	4375			

Bivariate analysis of continuous variables -**Grade second sem**



Correlation = 0.567

One-way analysis of variance (ANOVA)

Descriptive Statistics
Confidence Interval (CI) Probability: 95.0%

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
Grade second sem	-1	1400	0	0	5.872	6.1226	0.1636	5.551	6.193	0.0	17.7143
Grade second sem	1	2188	0	0	12.6939	2.6958	0.0576	12.5809	12.807	0.0	18.5714
Grade second sem	0	788	0	0	11.109	3.6124	0.1287	10.8564	11.3616	0.0	17.6
Grade second sem	Total	4376	0	0	10.226	5.2173	0.0789	10.0714	10.3806	0.0	18.5714

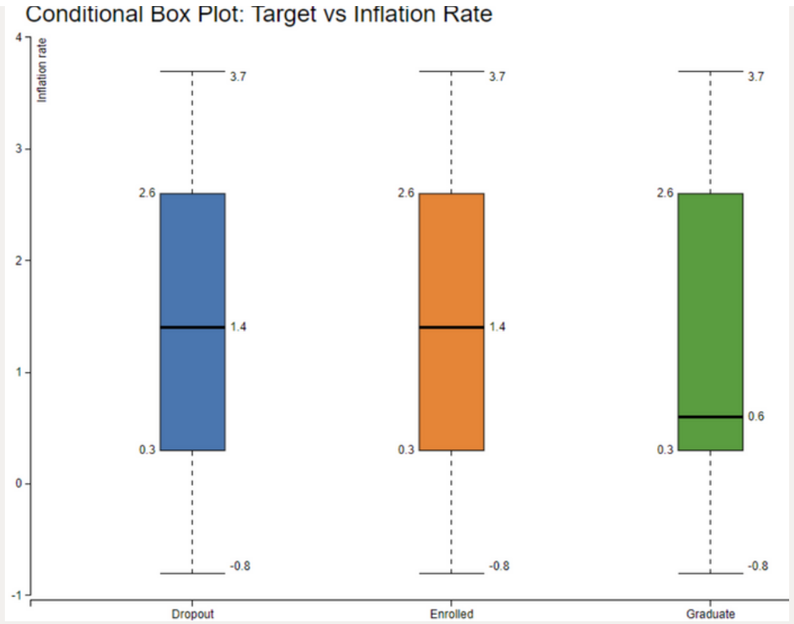
Levene Test
The Levene Test is used to test for the equality of variances.

	F	df 1	df 2	p-Value
Grade second sem	2,116.6085	2	4373	0.0

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
Grade second sem	Between Groups	40,481.2109	2	20,240.6055	1,126.0047	0.0
Grade second sem	Within Groups	78,607.2811	4373	17.9756		
Grade second sem	Total	119,088.4921	4375			

Bivariate analysis of continuous variables -**Inflation rate**



Correlation = -0.029

One-way analysis of variance (ANOVA)

Descriptive Statistics
Confidence Interval (CI) Probability: 95.0%

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
Inflation rate	-1	1400	0	0	1.2907	1.4029	0.0375	1.2172	1.3643	-0.8	3.7
Inflation rate	1	2188	0	0	1.1987	1.3732	0.0294	1.1412	1.2563	-0.8	3.7
Inflation rate	0	788	0	0	1.2122	1.372	0.0489	1.1162	1.3081	-0.8	3.7
Inflation rate	Total	4376	0	0	1.2306	1.3829	0.0209	1.1896	1.2716	-0.8	3.7

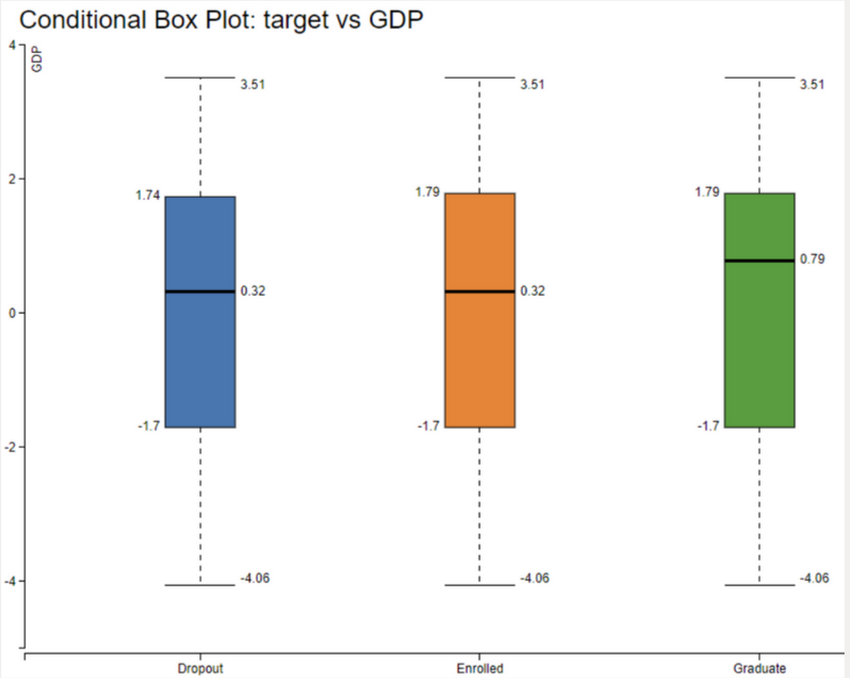
Levene Test
The Levene Test is used to test for the equality of variances.

	F	df 1	df 2	p-Value
Inflation rate	0.8286	2	4373	0.4367

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
Inflation rate	Between Groups	7.5502	2	3.7751	1.975	0.1389
Inflation rate	Within Groups	8,358.6587	4373	1.9114		
Inflation rate	Total	8,366.2089	4375			

Bivariate analysis of continuous variables -**GDP**



Correlation = 0.041

One-way analysis of variance (ANOVA)

Descriptive Statistics
Confidence Interval (CI) Probability: 95.0%

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
GDP	-1	1400	0	0	-0.1366	2.2451	0.06	-0.2543	-0.0189	-4.06	3.51
GDP	1	2188	0	0	0.0783	2.2625	0.0484	-0.0166	0.1731	-4.06	3.51
GDP	0	788	0	0	0.0545	2.317	0.0825	-0.1076	0.2165	-4.06	3.51
GDP	Total	4376	0	0	0.0052	2.2685	0.0343	-0.062	0.0725	-4.06	3.51

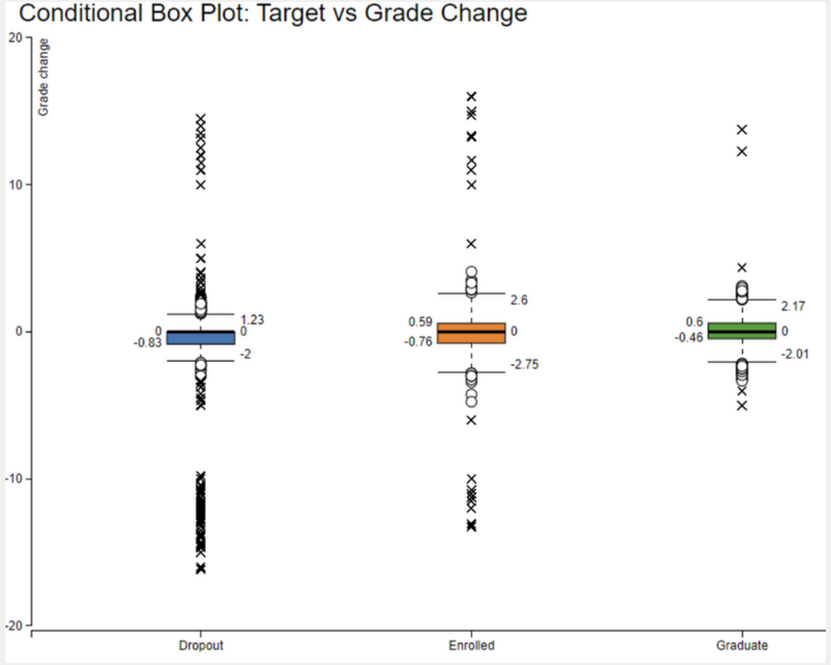
Levene Test
The Levene Test is used to test for the equality of variances.

	F	df 1	df 2	p-Value
GDP	0.3251	2	4373	0.7225

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
GDP	Between Groups	41.7633	2	20.8816	4.0635	0.0173
GDP	Within Groups	22,471.8949	4373	5.1388		
GDP	Total	22,513.6582	4375			

Bivariate analysis of continuous variables -**Grade change**



Correlation = 0.209

One-way analysis of variance (ANOVA)

Descriptive Statistics
Confidence Interval (CI) Probability: 95.0%

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
Grade change	-1	1400	0	0	-1.3564	4.4176	0.1181	-1.588	-1.1247	-16.1429	14.5
Grade change	1	2188	0	0	0.0546	1.0212	0.0218	0.0118	0.0975	-5	13.75
Grade change	0	788	0	0	-0.0238	2.5234	0.0899	-0.2003	0.1526	-13.25	16
Grade change	Total	4376	0	0	-0.4109	2.8859	0.0436	-0.4964	-0.3254	-16.1429	16

Levene Test
The Levene Test is used to test for the equality of variances.

	F	df 1	df 2	p-Value
Grade change	379.891	2	4373	0.0

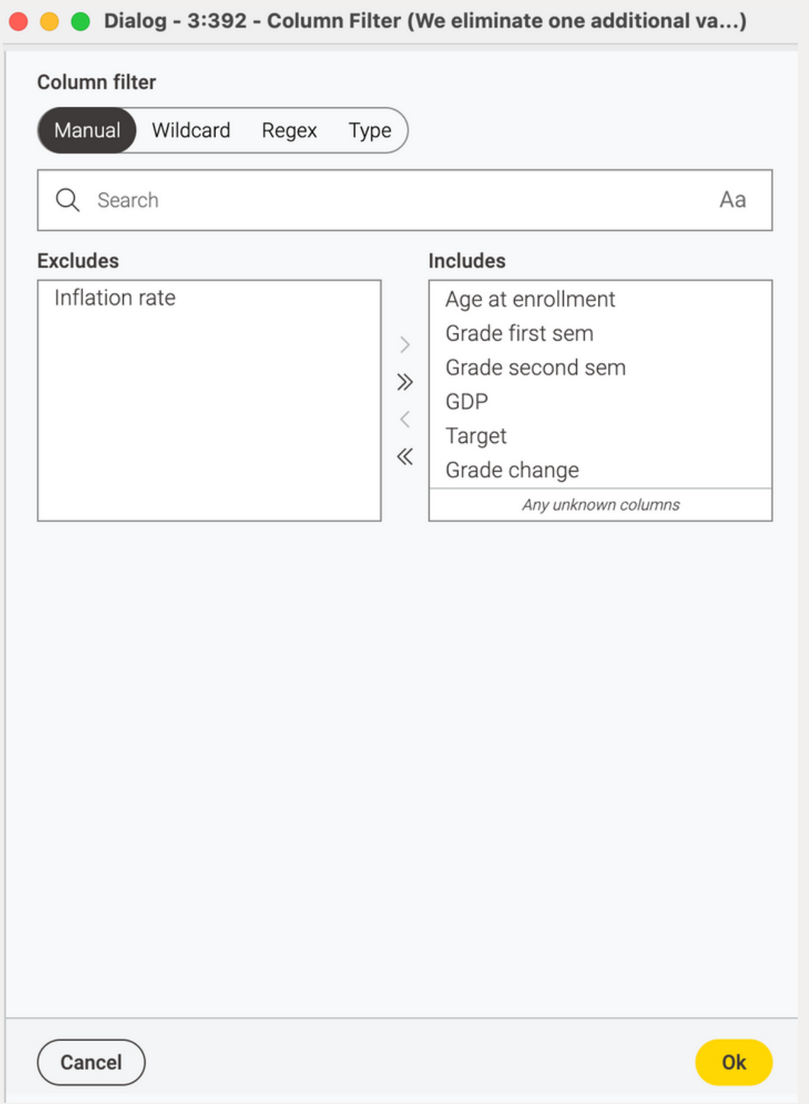
ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
Grade change	Between Groups	1,843.6858	2	921.8429	116.5303	0.0
Grade change	Within Groups	34,593.7548	4373	7.9108		
Grade change	Total	36,437.4406	4375			

Eliminating one last predictor

The null hypothesis for ANOVA states that all group means are equal. Therefore, a high p-value indicates that the differences in group means are likely due to random chance rather than a systematic effect or intervention.

We see that the variable “Inflation rate” has a high (13.89%) p-value for the ANOVA test, together with quite a low (-2.9%) correlation with the variable “Target”. Therefore, we opt for deleting this variable from our dataset. All other variables show either a sufficiently-low ANOVA p-value, or a sufficiently-high (in absolute terms) correlation with “Target”, or both.

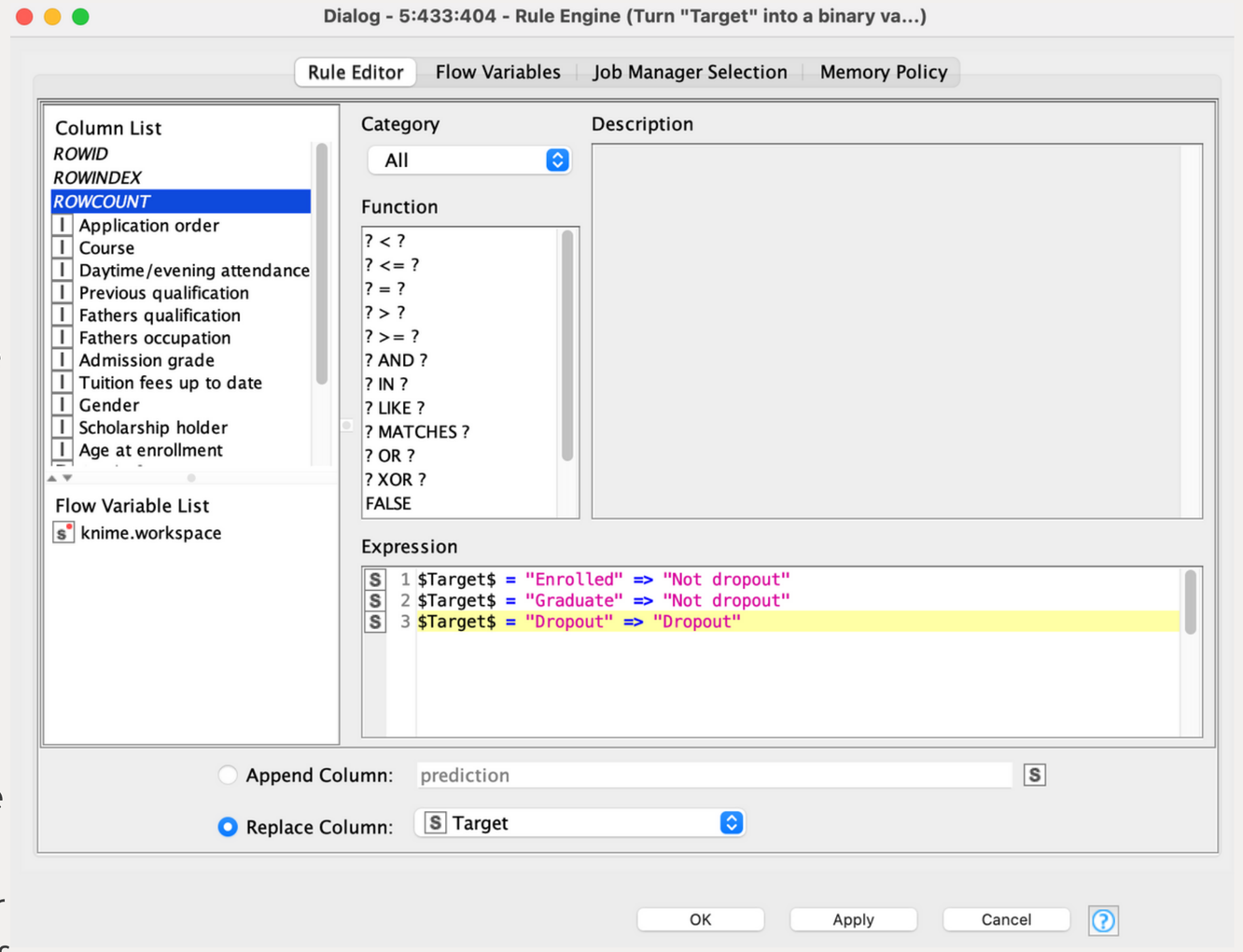


Data modeling

Logistic Regression -**Introduction**

The first machine learning algorithm we implement is a Logistic Regression. This method is used for binary classification and, indeed, in this case we are interested in predicting whether a given student is likely to drop out of college or not, as opposed to classify him/her in one of three categories. We first turn our “Target” variable into a binary one, where the two classes are now “Dropout” and “Not Dropout”.

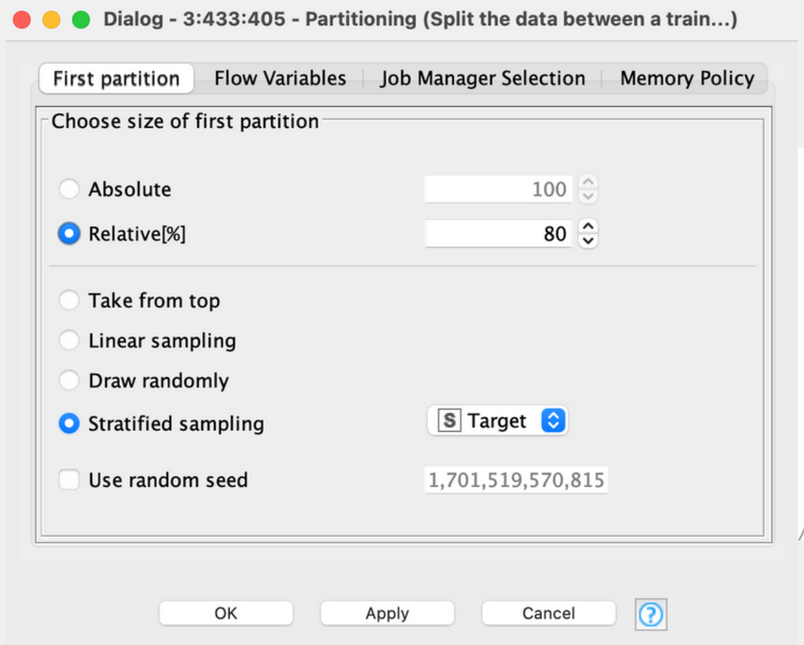
Logistic Regression models the probability that a given input vector (in our case, a student) belongs to one of the two classes. This algorithm uses a logistic function to squeeze the output of a linear equation between 0 and 1. The weights of this model are estimated using maximum likelihood estimation. The probability threshold is the standard 0.5.

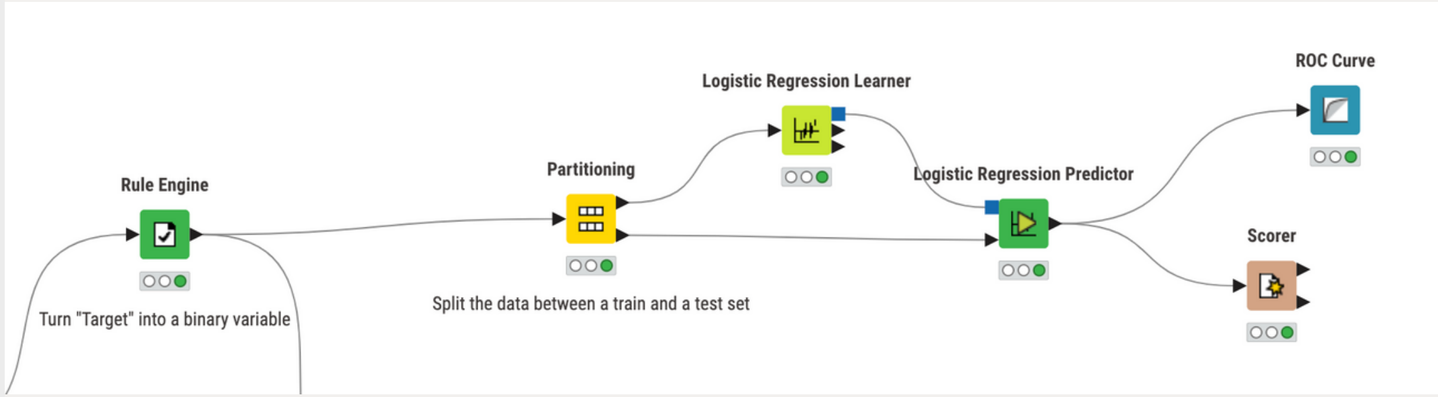


Logistic Regression. -**Partitioning the data**

We initially implement the Logistic Regression model without one-hot encoding and normalization.

The first step is partitioning our dataset between a train set and a test set. The proportion we used are 80-20, given the relatively-limited number of instances in our dataset (~4000). For this task, we perform a stratified sampling with respect to the “Target” column (which essentially means that the frequency of each class of “Target” is the same in the two subsets created).

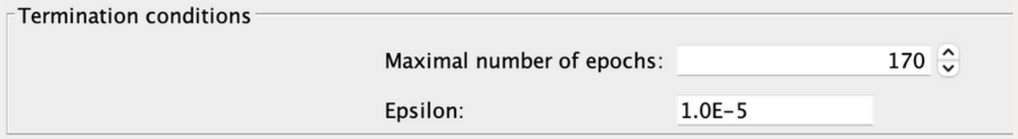


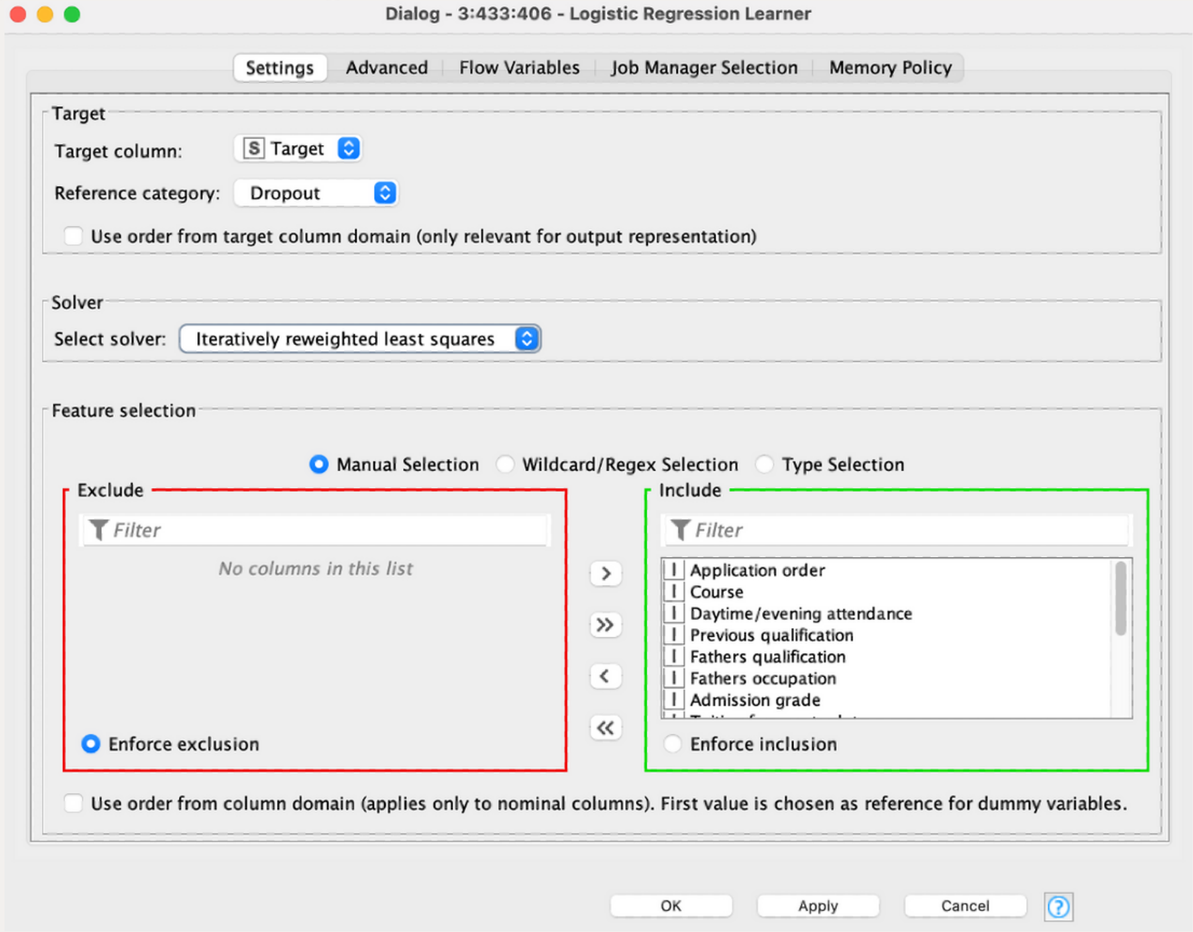


Logistic Regression -**Training the model**

We input our training dataset into a *Logistic Regression Learner* node. In this case we choose the Iteratively reweighted least squares algorithm as our solver (aka the optimization technique used to find the best-fitting parameters). This technique iteratively updates parameter estimates by solving a series of weighted least squares problems, efficiently handling the logistic function's non-linearity, until convergence.

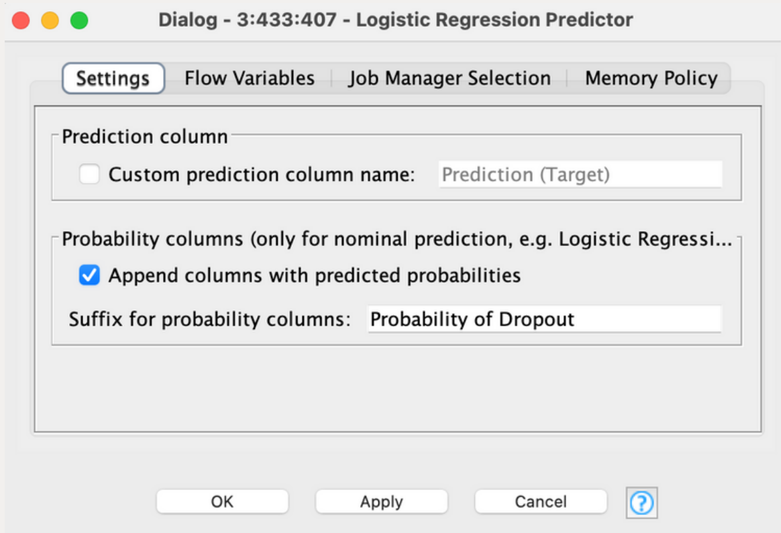
We clearly train the model with all regressors, since we already eliminated variables we did not want to include in our models.
We select a (hopefully) quasi-optimal value of 170 for the maximal number of epochs, and an epsilon value of 10^-5, which should guarantee convergence.
To arrive at this values, a basic trial and error procedure has been put in place.





Logistic Regression -**Evaluating the model**

We proceed by inputing the trained model and the test dataset into a *Logistic Regression Predictor* node. Apart from the actual class predictions, we also append to the test dataset a column containing the predicted probability of each student belonging to a certain class (meaning, the confidence with which each prediction has been made).



We use a *Scorer* node to find the most relevant metrics for the evaluation of the model. We see that the algorithm has an accuracy of 85.731% and a Cohen's kappa of 0.655%. This metric quantifies the agreement between the predicted and observed classifications while adjusting for agreement that occurs by chance.

Prediction ...	Dropout	Not dropout
Dropout	192	37
Not dropout	88	559

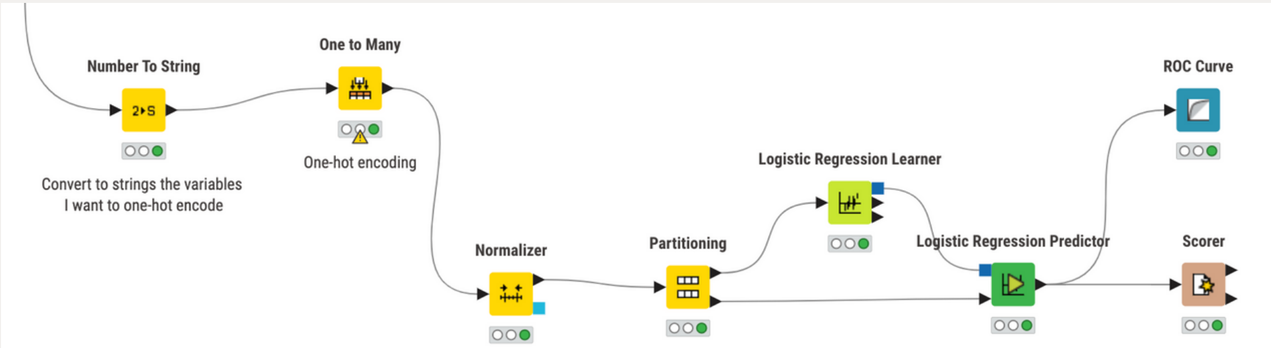
Correct classified: 751 Wrong classified: 125

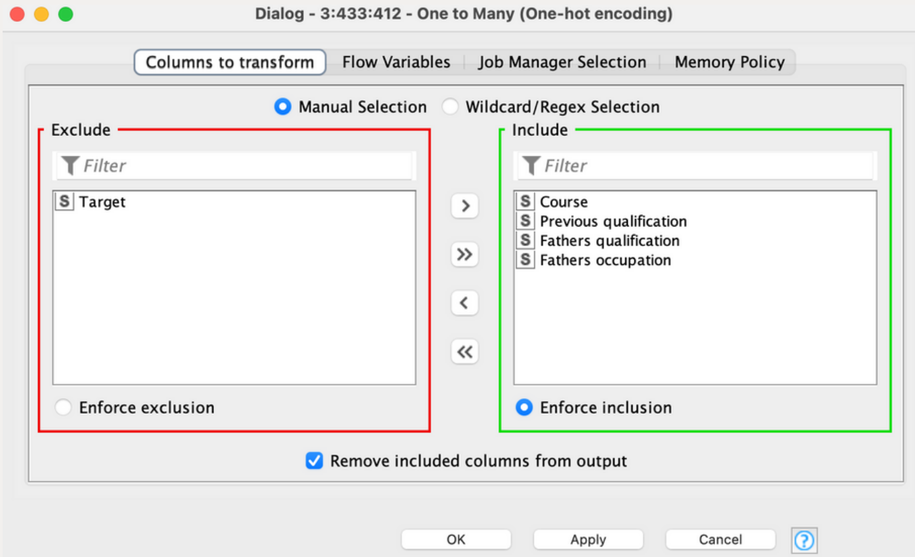
Accuracy: 85.731% Error: 14.269%

Cohen's kappa (κ): 0.655%

Logistic Regression -**Enhancing the model**

We now implement the same steps as before, but we first perform some adjustments on the train and test data. Specifically, we perform the one-hot encoding of the variables “Course”, “Previous qualification”, “Father’s qualification”, and “Father’s occupation”. This techniques implies substituting each column containing categorical variables with k-1 columns, where k is the number of categories of the dropped column. Each new variable is binary, and specifies whether the individual belongs to that category or not. This should enhance the model's ability to analyze data with categorical features.

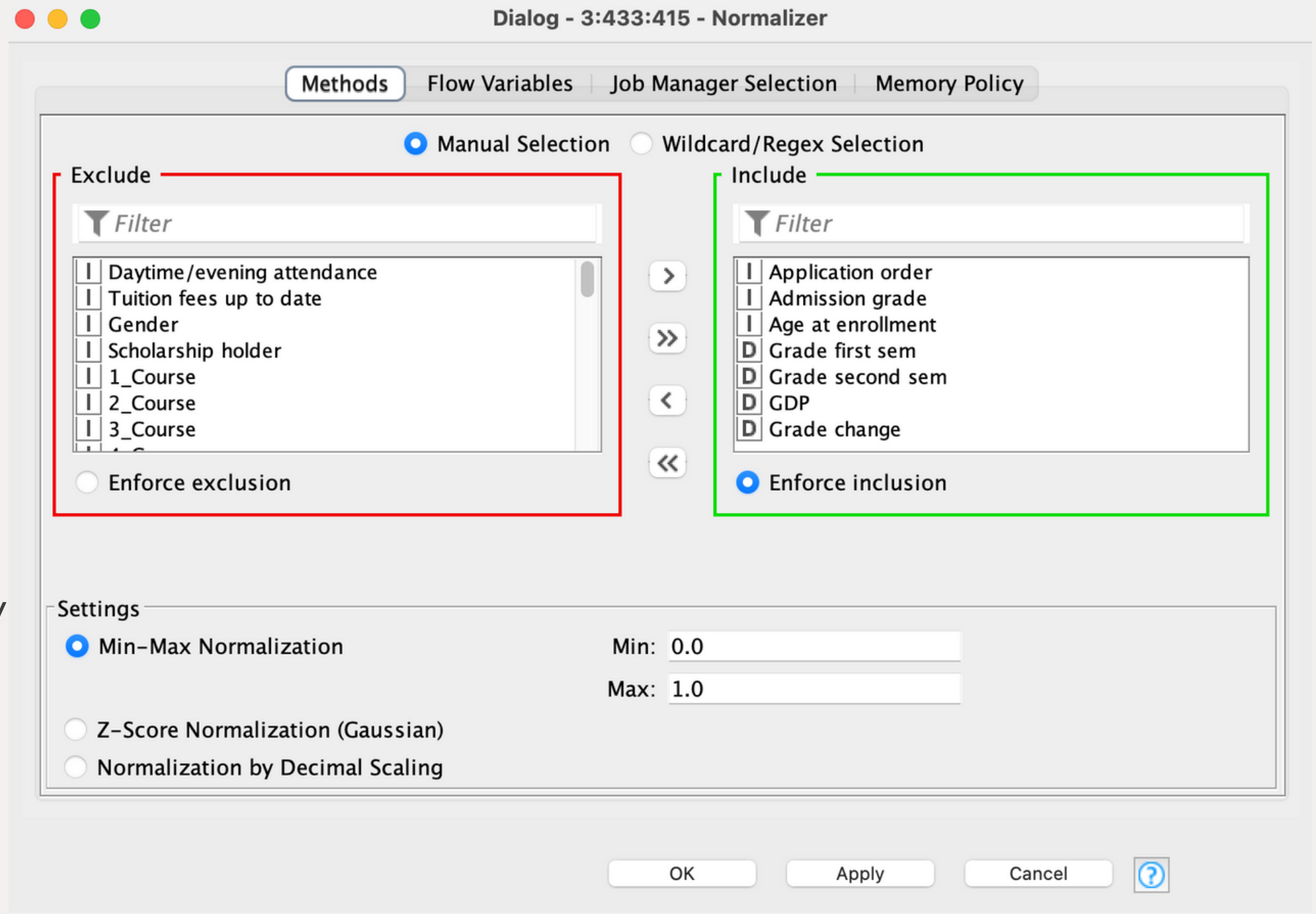




Logistic Regression -**Enhancing the model**

Thanks to one-hot encoding, each categorical variable is now either 0 or 1. To make each continuous variable fall within the [0,1] interval as well, we normalize them through a Normalizer node, choosing a min-max normalization.

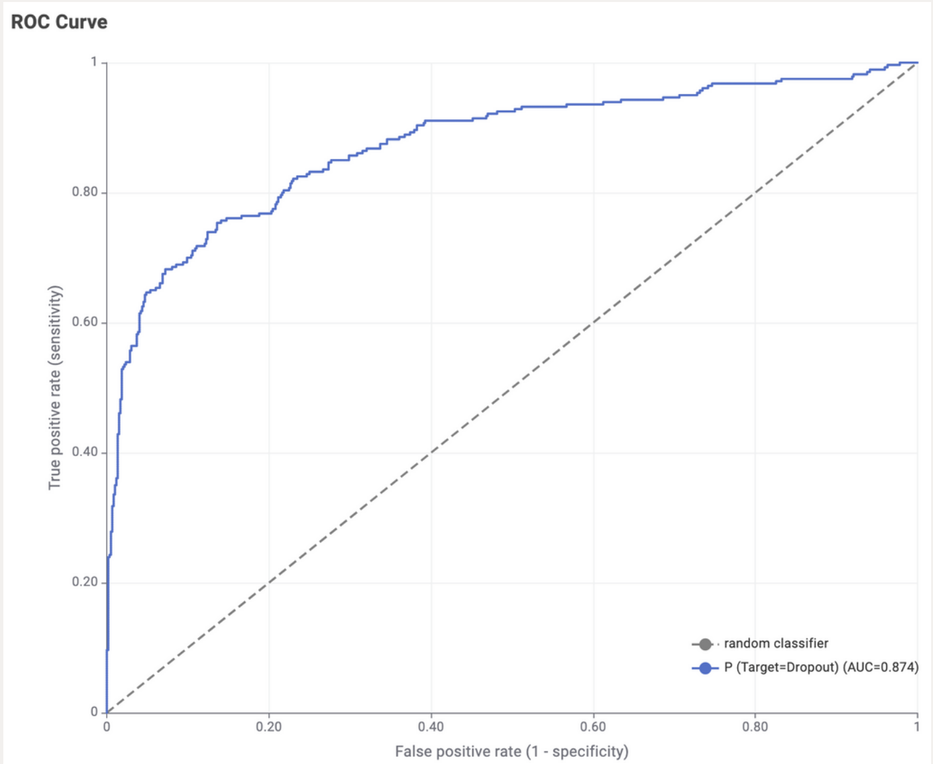
The steps we follow from now on are the same as in the previous section, apart from the solver we choose when training our Logistic Regression model, which in this case is a stochastic average gradient (we also experimented with the IRLS solver, and we did not notice any significant difference in the explanatory power of the model).



Logistic Regression -**Enhancing the model**

When evaluating the Logistic Regression model with one-hot encoding and regularization, we notice that its performance is very similar to the one of the simpler model, even slightly worse. We think this might be due to several reason:

- First of all, the creation of many additionaly binary variables (aka the one-hot encoding we performed) might increase the complexity of the model to a point in which we have overfitting, especially if we take into account the relatively-small size of our dataset
- Secondly, since the effectiveness of one-hot encoding and normalization largely depends on how these techniques align with the underlying patterns in the data, it might be that in this case these preprocessing steps do not align well with the data's structure, making the simpler model slightly more powerful
- Finally, these difference might also be due, at least in part, to random variation. If we had a larger dataset, we could probably assess better which Logistic Regression model is the best.



Target \ Ta...	Dropout	Not dropout
Dropout	183	97
Not dropout	37	559

Correct classified: 742 Wrong classified: 134

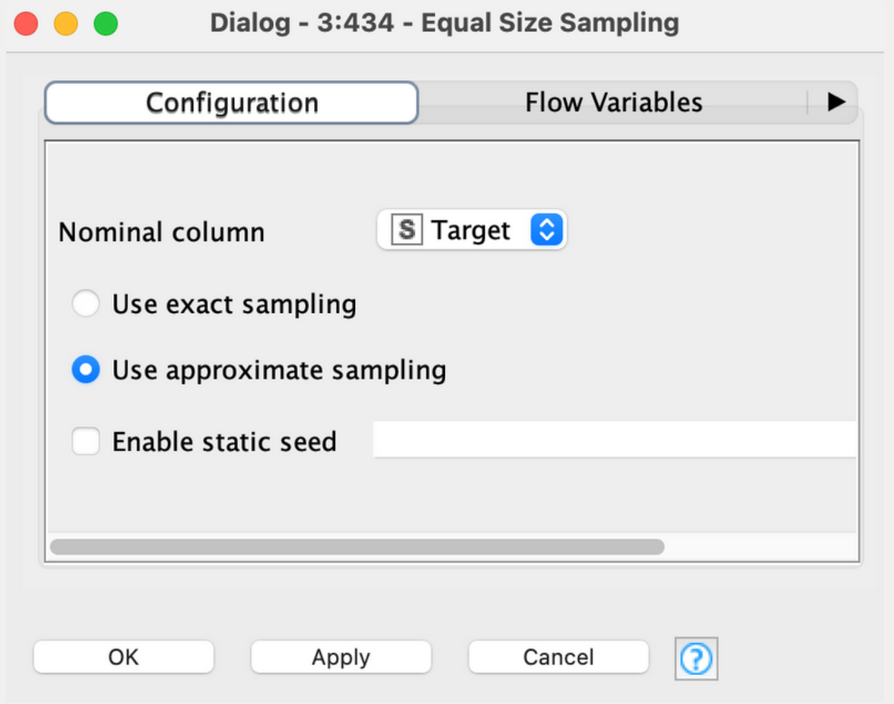
Accuracy: 84.703% Error: 15.297%

Cohen's kappa (κ): 0.627%

Decision Tree -**Introduction**

The second model we implement is a Decision Tree. This algorithm starts by splitting the input data into subsets according to certain decision criteria, which are based on the attributes of the data. Each split forms a branch of the tree, leading to more splits moving down. This process continues until either a certain pre-determined maximum depth is reached, or when further splitting adds no significant value to the predictions.

We experiment using an Equal Size Sampling node on the train set to make sure that the relative frequencies of the three classes of the “Target” variable are similar. However, we end up not adopting this strategy, as the fact that we have a non-negligible class imbalance (roughly 18% of students belong to the “Enrolled” class, whereas almost 50% to the “Graduate” class) forces this node to dramatically reduce the size of our training set, therefore decreasing the learning power of the Decision Tree. It is peculiar that the exact problem that this strategy aims at solving (that is, class imbalance) is actually the reason why employing this strategy is pointless, but this is precisely the way it is. As further proof of this, the accuracy of the model significantly improves when the *Equal Size Sampling* node is removed.

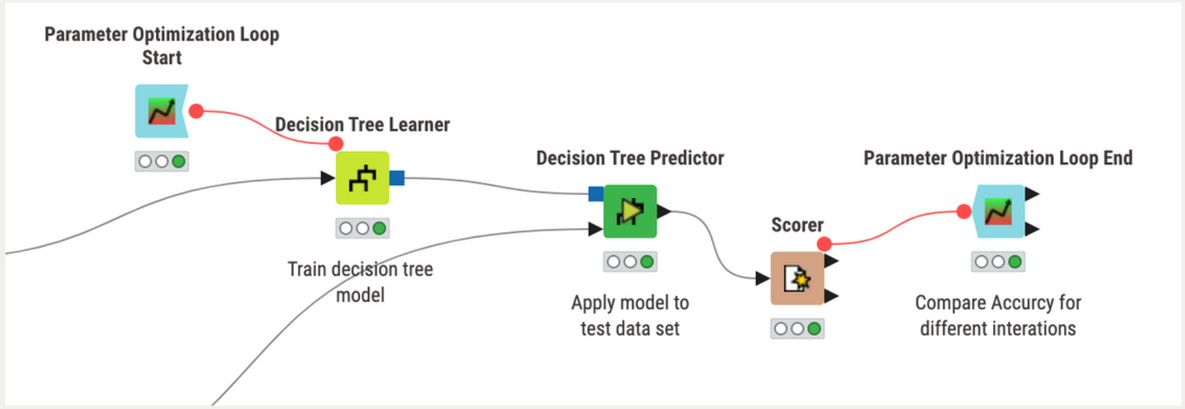


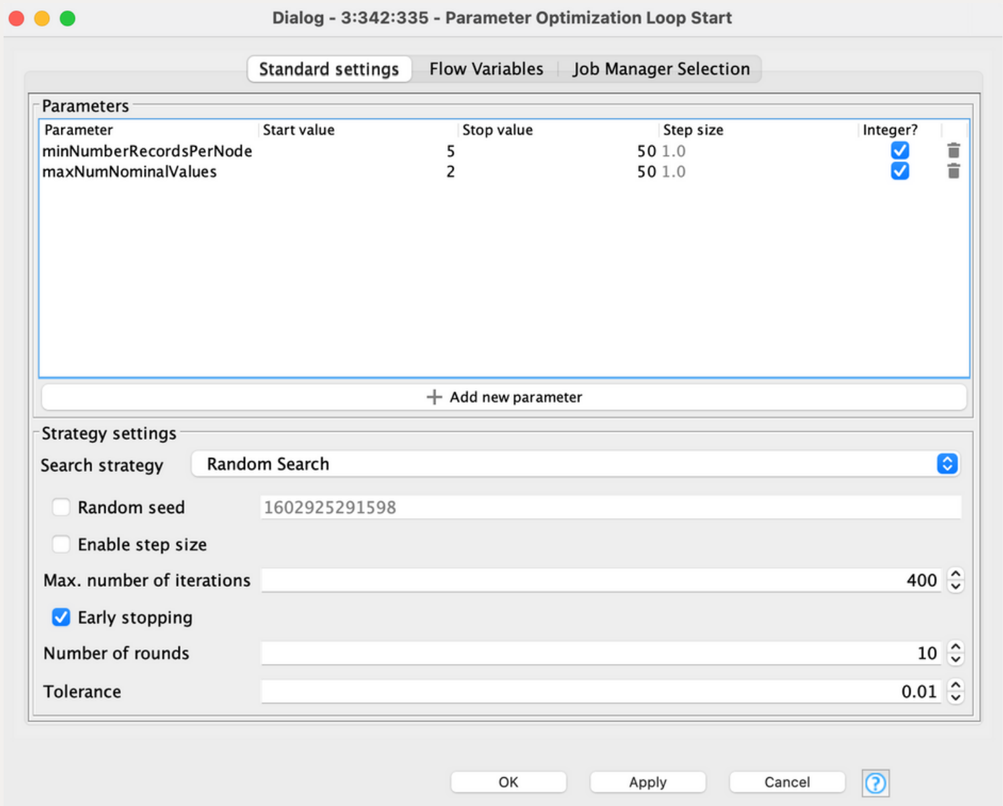
Decision Tree -**Training the model**

To choose the optimal parameters of the *Decision Tree Learner* node, we implement an optimization loop. Specifically, we want to optimize both the minimum number of records per node, and the maximum number of nominal values.

After experimenting with different search strategies, we opt for Random Search. We set the maximum number of iterations to 400, the number of rounds (aka optimization sessions) to 10, and the tolerance to 0.01.

This results in a minimum number of records per node of 49, and in a maximum number of nominal values of 34.





Decision Tree -**Evaluating the model**

The parameters found by the optimization loop, the output of which is shown below, result in an accuracy of roughly 71.9%, and a Cohen's kappa of 0.512%.

Target \ Pr...	Dropout	Graduate	Enrolled
Dropout	190	64	26
Graduate	13	407	18
Enrolled	28	97	33

Correct classified: 630 Wrong classified: 246

Accuracy: 71.918% Error: 28.082%

Cohen's kappa (κ): 0.512%

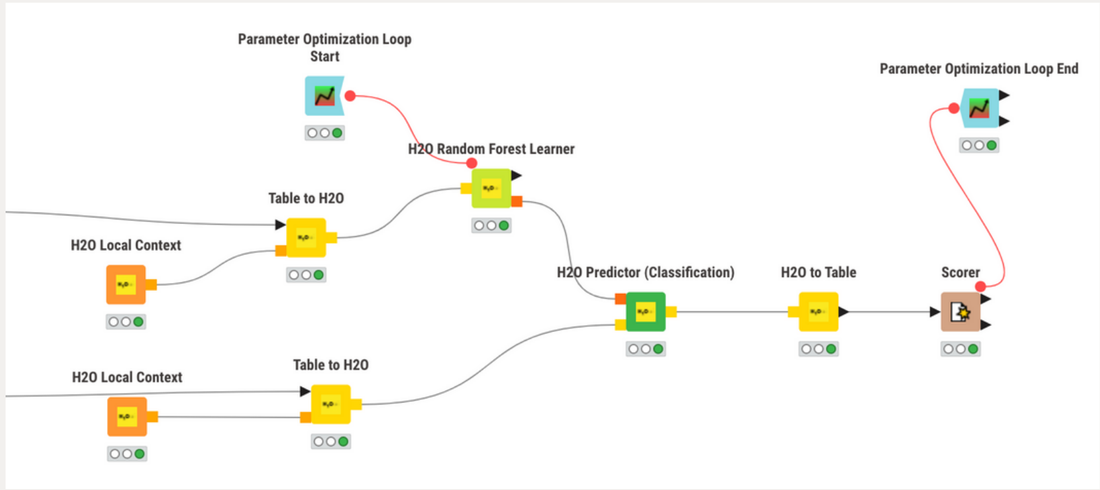
In the following section, we try to improve on these results by using many decision trees together.

Row ID	minNumberRecordsPerNode	maxNumNominalValues	Objective value
Best parameters	11	19	0.719

Random Forest -**Introduction**

Indeed, the model we implement here is a type of ensemble learning method that joins together multiple Decision Trees, with the aim of improving the overall performance of each single tree. This algorithm starts by creating multiple Decision Trees from randomly selected subsets of the training dataset. Each tree is built from a sample drawn with replacement (aka Bootstrap sample) from the training set. During the construction of these trees, a random subset of features is chosen at each node to determine the split. When the model makes predictions (for classification tasks), each Decision Tree “votes” for one class, and the most-voted class is the output of the Random Forest.

Here we implement an H2O version of the Random Forest, as this allows for better optimization of the parameters. This explains the presence of *Table to H2O* and *H2O to Table* nodes in the image below.

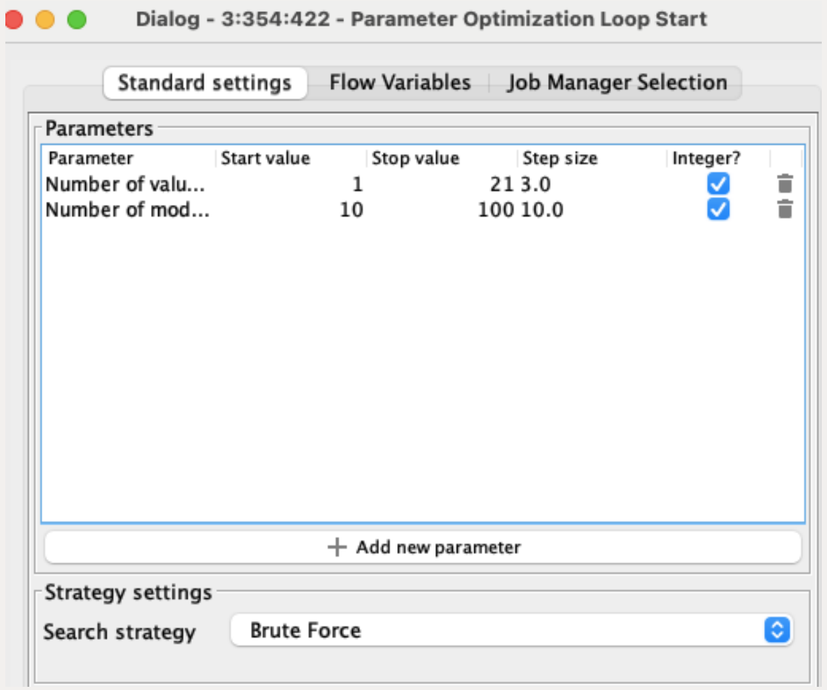


We again experiment with applying an *Equal Size Sampling* node on the train data but, for the same reasons explained before, we opt for not using it in the end

Random Forest -**Training the model**

As we did when training the Decision Tree model, also in this case we implement an optimization loop, with the hope of selecting (quasi) optimal parameters for our Random Forest model. Specifically, after experimenting with different search strategies, we choose a Brute Force optimization method (which essentially means trying out all the combinations of parameters within the range we specify, and keep the configuration that maximizes the accuracy of the model).

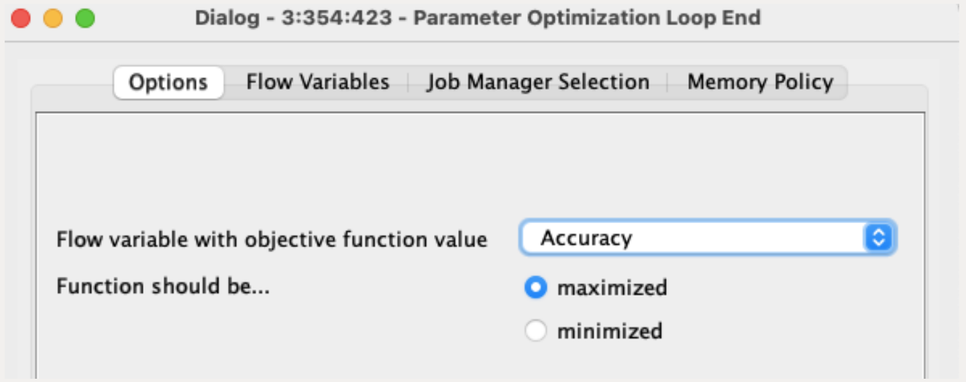
We optimize both the number of models and the max tree depth. For the first parameter, we try values between 10 and 100 with a step of 10. For the second parameter, we try values between 1 and 21 with a step size of 3. This creates a total of 10 * 7 = 70 configurations to try.



Random Forest -**Evaluating the model**

The parameter optimization loop outlined in the previous slide returns an optimal value of 10 for the number of models, and 1 for the tree depth. This is peculiar, as improving the complexity of the model does not increase accuracy (which is the measure that the optimization loop maximizes). The main reason why this happens is probably that the model is particularly inclined to overfit on the training data (again, the relatively small number of instances poses some limitations here). Another possible reason is that the nature of our data does not require complex models. Indeed, the accuracy of the Decision Tree model is very similar to the one of the Random Forest one (something that does not happen frequently, and that suggests that simpler models are better suited in this case).

Clearly, the choice of which metric to optimize plays an important role. We choose accuracy, since for us it is the most important statistic, but other valid choices might be metrics like F1-score, precision, Cohen's kappa, or AUC.

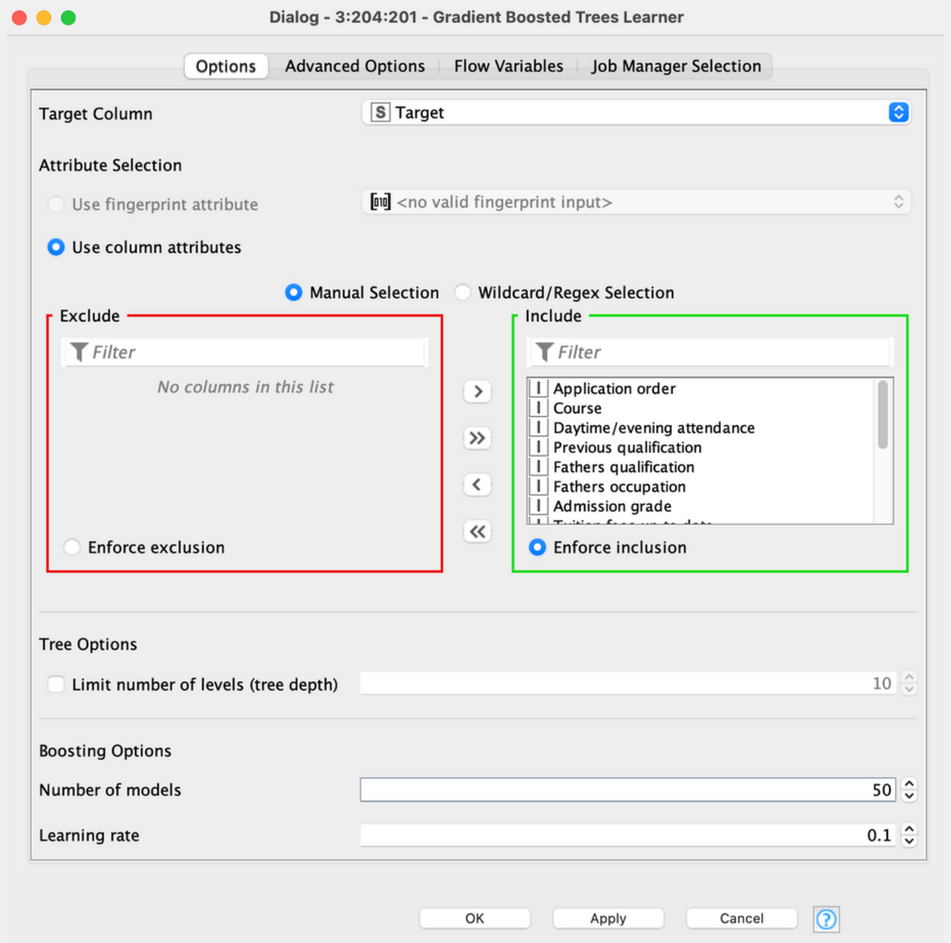


#	RowID	Number of values (tree depth) *Number (integer)*	Number of models *Number (integer)*	Objective value *Number (double)*
1	Best parameters	1	10	0.716

Gradient Boosting -**Introduction and model training**

The last algorithm we use is a Gradient Boosted Trees model. This algorithm builds on the principles of boosting, an ensemble technique that combines the predictions from multiple models to improve overall performance. The model starts with a simple model (in our case, a simple Decision Tree), and it later iteratively builds more-and-more complex models to predict both the actual values of the “Target” variable and the errors that the classification models will make. By proceeding sequentially until the improvement in performance becomes negligible, the algorithm arrives at a final model.

The first approach we try is a simple one: we do not limit the tree depths, so that the algorithm stops expanding the trees only when the improvement in performance is very small, and we manually try different values for the number of models. We arrive at a value of 50, which corresponds to an accuracy of roughly 71%.

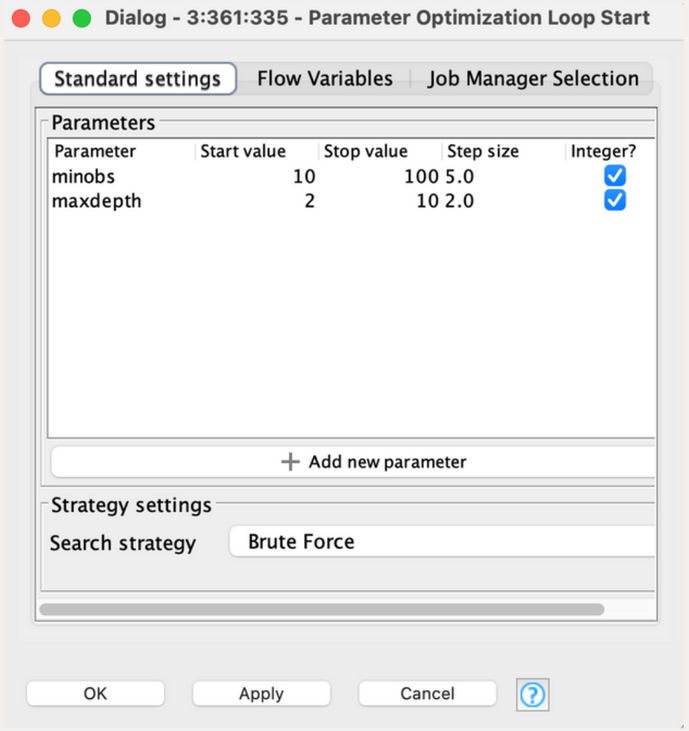


Gradient Boosting -**Enhancing the model**

At this point, in order to improve the performance of our Gradient Boosting model, we implement some changes.
First of all, we employ an H2O version of the model (similarly to what we did for the Random Forest algorithm), hoping to better optimize the parameters.

Moreover, we run a parameter optimization loop, in order to find optimal values for both minobs and maxdepth. We again experiment with different searching methods, ending up to opt again for a brute-force approach.

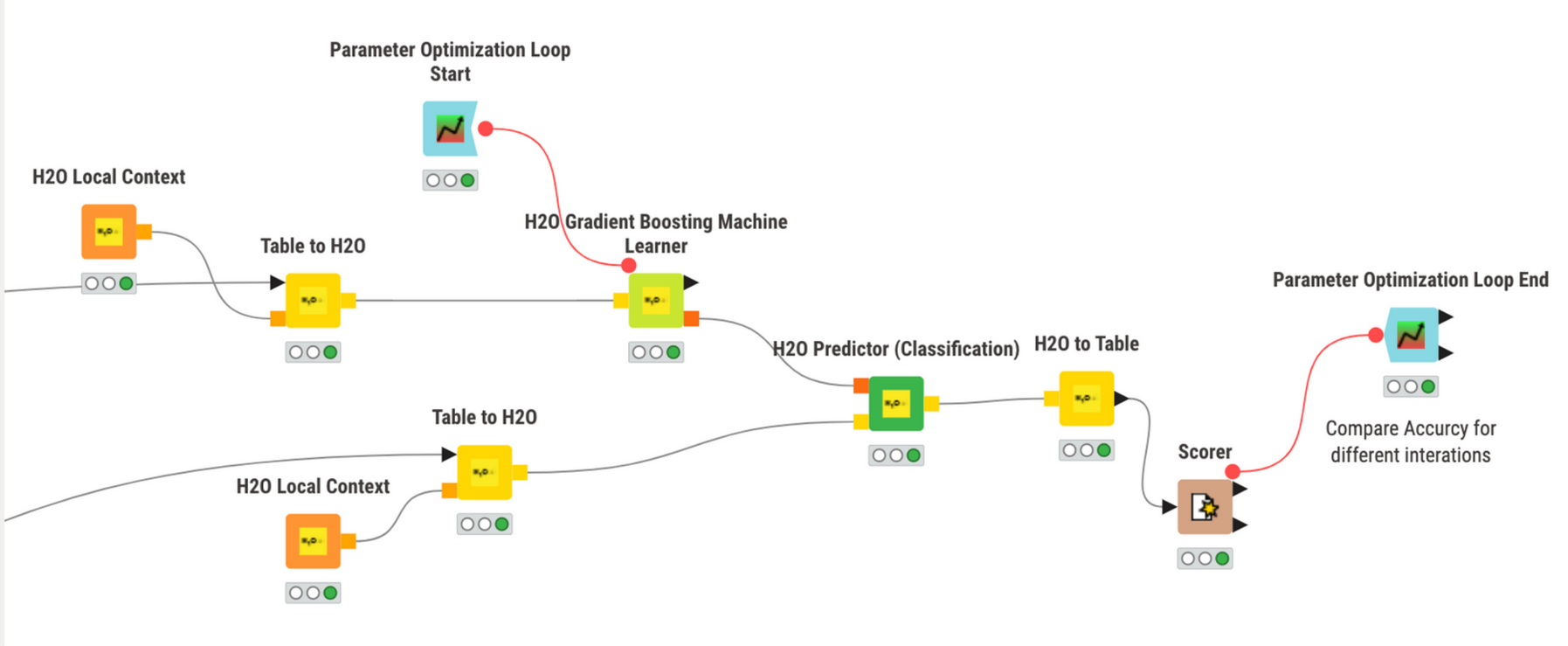
After several rangesand step size tried, we test minobs values between 10 and 100 with a step size of 5, and maxdepth values between 2 and 10 with a step size of 2. This implies that the optimization loops tests 19*5=95 different combination of parameters.



Gradient Boosting -**Evaluating the model**

The optimal parameters that we obtain with this approach are 20 for minobs and 6 for maxdepth, which are associated with an accuracy of roughly 72%.

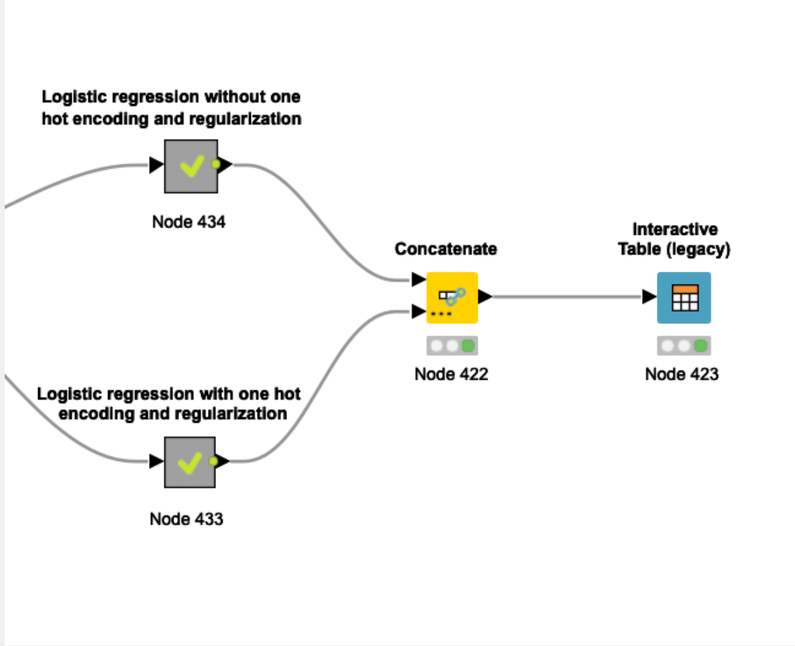
Row ID	minobs	maxdepth	Objective value
Best parameters	20	6	0.724



Outcomes

Logistic Regression Outcomes

In order to determine which is the best Logistic Regression model, we concatenate the performance measures relative to our two models and display them in an "Interactive Table" for comparison.

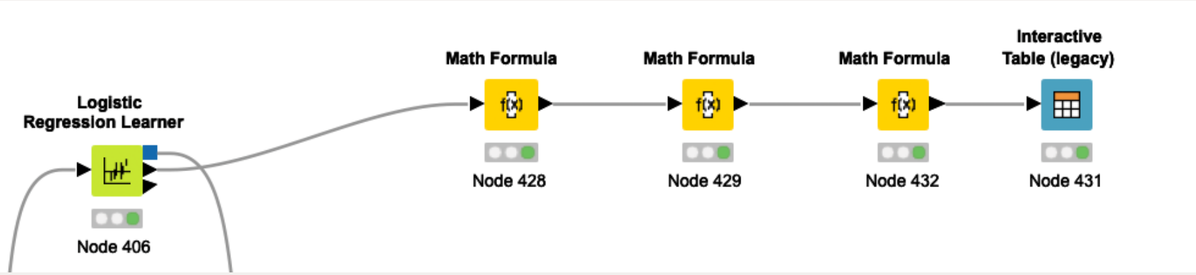


Row ID	Accuracy	Precision	Sensitivity	Specificity	AUC
P (Target=Not dropout)	0.857	0.864	0.938	0.686	0.892
P (Target=Not dropout)_one-hot-enc&norm	0.847	0.852	0.938	0.654	0.874

Analyzing the table, we notice that the performance of the second model is very similar to the performance of the first one. The model with one-hot encoding and regularization has a slightly higher accuracy (0.857) compared to the one without them, higher precision (0.864) and specificity (0.686) . It also has a higher AUC (0.892). Hence we can say that our first model is slightly better, even though clearly these differences are at least partially explained by randomness. In the following slides we proceed by analyzing the outcome of the Logistic Regression Model we choose.

Logistic Regression Outcomes

We analyze the “Coefficients and Statistics” output of the *Logistic Regression Learner* node. We include three math formulae for computing the odds ratio and the extreme values of a 95% confidence interval.



The lower and upper bounds of the 95% confidence interval provide a range within which we can be reasonably confident that the true odds ratio lies.
A confidence interval that does not include 1 is generally considered statistically significant.

We obtain the following table, which we will discuss in the next slide.

Row ID	Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|	odds ratio	low_95%	upp_95%
Row1	Not dropout	Application order	-0.064	0.041	-1.584	0.113	0.938	-0.144	0.015
Row2	Not dropout	Course	-0.054	0.012	-4.576	0	0.947	-0.078	-0.031
Row3	Not dropout	Daytime/evening attendance	0.213	0.167	1.27	0.204	1.237	-0.116	0.541
Row4	Not dropout	Previous qualification	-0.082	0.086	-0.959	0.338	0.921	-0.251	0.086
Row5	Not dropout	Fathers qualification	0.076	0.07	1.077	0.281	1.079	-0.062	0.214
Row6	Not dropout	Fathers occupation	0.017	0.033	0.5	0.617	1.017	-0.049	0.082
Row7	Not dropout	Admission grade	0.078	0.035	2.264	0.024	1.081	0.011	0.146
Row8	Not dropout	Tuition fees up to date	2.49	0.168	14.863	0	12.058	2.161	2.818
Row9	Not dropout	Gender	-0.376	0.103	-3.665	0	0.687	-0.577	-0.175
Row10	Not dropout	Scholarship holder	0.95	0.137	6.922	0	2.586	0.681	1.219
Row11	Not dropout	Age at enrollment	-0.396	0.065	-6.068	0	0.673	-0.524	-0.268
Row12	Not dropout	Grade first sem	0.081	162,325....	0	1	1.084	-318,157.476	318,157.637
Row13	Not dropout	Grade second sem	0.163	162,325....	0	1	1.177	-318,157.393	318,157.719
Row14	Not dropout	GDP	0.011	0.022	0.514	0.607	1.011	-0.032	0.054
Row15	Not dropout	Grade change	0.082	162,325....	0	1	1.086	-318,157.473	318,157.638
Row16	Not dropout	Constant	-2.943	0.456	-6.452	0	0.053	-3.837	-2.049

Logistic Regression Outcomes

The explanatory variables "Course", “Admission grade”, "Tuition fees up to date", "Gender", "Scholarship holder" and "Age at enrollment" are significant at the 5% level. The odds ratio column in logistic regression output provides valuable insights into the impact of each predictor variable on the odds of the event (not dropout) occurring. An odds ratio greater than 1 suggests that as the predictor variable increases, the odds of the event occurring also increase.

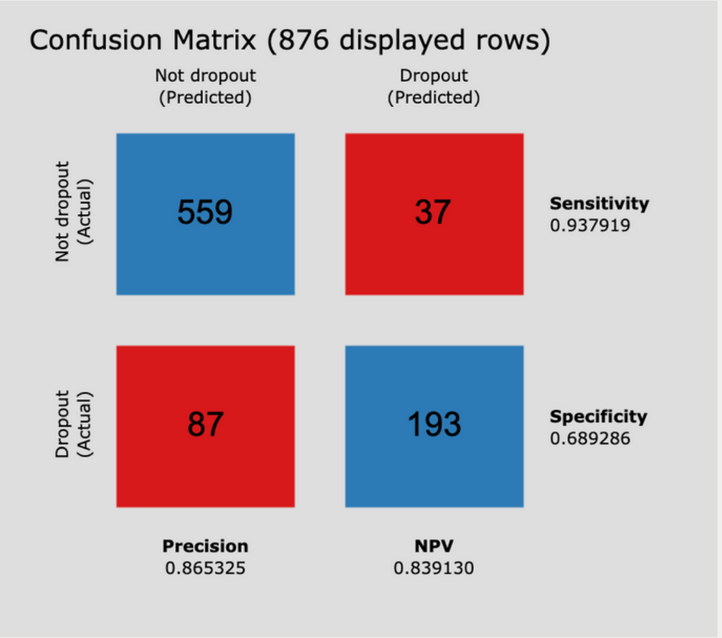
The odds ratio underlines the importance of "Tuition fees up to date" for not dropping out. Keeping all other variables constant, the odds of a student not dropping out increases by about 15 times if the “Tuition fees up to date” variable is 1, as opposed to the base case of when it is 0.

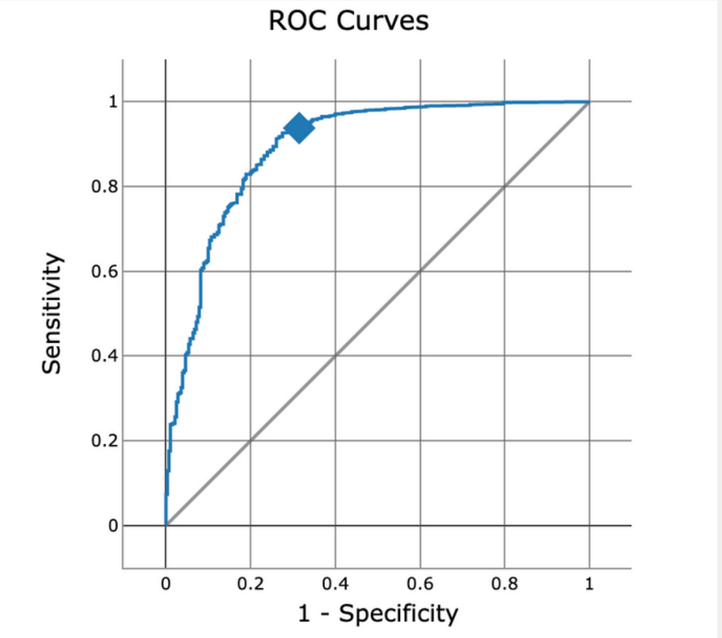
We can also observe:

- For binary variables like "Gender", the odds ratio represents the change in odds when the variable changes from 0 to 1 (Female to Male). For example, in the "Gender" variable, being male (0) is associated with a 37.6% decrease in the odds of not dropping out compared to being female (1).
- The odds of a student not dropping out when holding a scholarship is 2.59 times the odds for a student not holding one.
- For a one-unit increase in the “Age at enrollment” variable, the odds of not dropping out decrease by approximately 39.6%.
- The odds of a student not dropping out when attending during the daytime is 1.23 times the odds for a student attending in the evening.
- For a one-unit increase in the admission grade variable, the odds of not dropping out increase by approximately 7.8%.

Logistic Regression Outcomes

The threshold of 0.5 chosen for the "Binary Classification Inspector" node, results in a sensitivity of 0.9379 and a specificity of 0.6893. A False Positive occurs when the model predicts a positive class (Not Dropout) when the actual class is negative (Dropout). In other words, it is a case where the model fails to identify students who are actually at risk of dropping out. A False Negative occurs when the model predicts a negative class (Dropout) when the actual class is positive (Not Dropout). In other words, it is a case where the model incorrectly identifies a student as a potential dropout when, in reality, they do not drop out. A false negative means that resources might be allocated to students who are predicted to drop out but do not, potentially leading to unnecessary interventions or support. The desire for higher specificity is justified because classifying a student as "Not Dropout" when they actually drop out (False Positive) is considered more critical than classifying a student as "Dropout" when they do not drop out (False Negative). In other words, the cost of missing a student who might drop out is considered higher than the cost of intervening with a student who ultimately does not drop out. To increase the specificity of our model we could increase our threshold.





Random Forest Outcomes

Now we move on to analyzing the Random Forest Output by looking at the “Variable importance measure” table, from the *H20 Random Forest Learner* output.

This table provides a ranking of features based on their contribution to the model's predictive accuracy. Features at the top of the list are considered more important in making predictions.

If a feature has a high “Percentage” value, it means that this feature contributes more to the decision-making process of the Random Forest model. On the other hand, features with lower percentages are considered less important.

Row ID	Relative Importance	Scaled Importance	Percentage
Grade second sem	28,224.006	1	0.361
Grade first sem	17,557.314	0.622	0.224
Tuition fees up to date	10,978.074	0.389	0.14
Scholarship holder	5,485.625	0.194	0.07
Age at enrollment	4,094.624	0.145	0.052
Grade change	4,085.571	0.145	0.052
Course	4,009.046	0.142	0.051
Gender	1,400.119	0.05	0.018
Admission grade	706.174	0.025	0.009
Fathers qualification	486.198	0.017	0.006
GDP	372.656	0.013	0.005
Fathers occupation	313.065	0.011	0.004
Application order	228.085	0.008	0.003
Previous qualification	210.232	0.007	0.003
Daytime/evening attendance	81.801	0.003	0.001

We can notice that in this example, the feature “Grade score second sem” is considered the most important, contributing 36.1% to the model's predictive accuracy. This underlines the importance of grades for a successful academic career. Feature “Daytime/evening attendance”, with only 0.1%, is considered the least important.

Random Forest Outcomes

We now evaluate the ability of our Random Forest model to correctly predict the target variable through the Confusion matrix output. By simply looking at the Overall Accuracy, we notice our model results in 70.7% of correct predictions with 619 correct classifications and 29.3% of Error with 257 wrong classifications. This figure might seem high, however, accuracy and error are not the only relevant measures in a predictive model. In the following slides, we therefore provide a in-depth analysis of other metrics that we consider important for evaluating the effectiveness of our predictive model.

Target \ Pr...	Enrolled	Graduate	Dropout
Enrolled	10	104	44
Graduate	2	409	27
Dropout	14	66	200

Correct classified: 619

Wrong classified: 257

Accuracy: 70.662%

Error: 29.338%

Cohen's kappa (κ): 0.481%

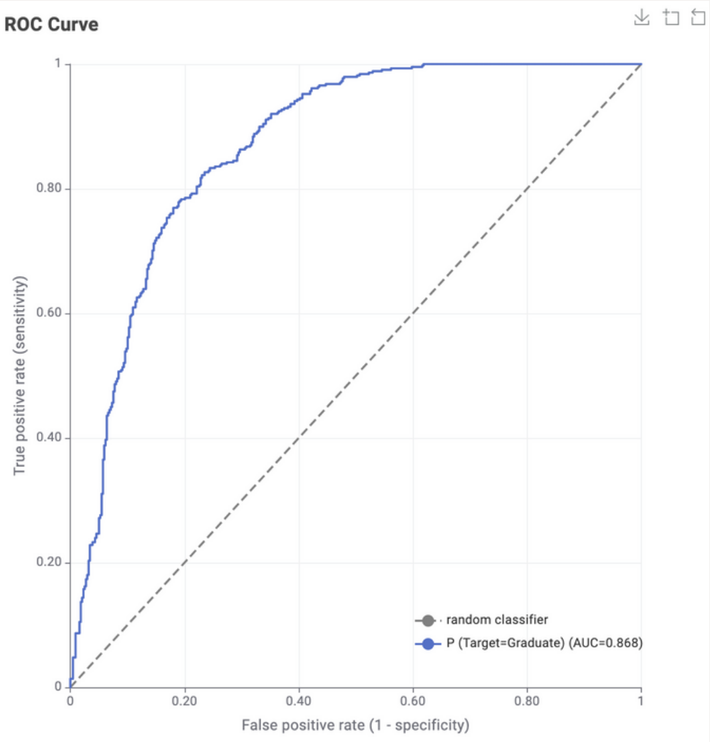
Random Forest Outcomes

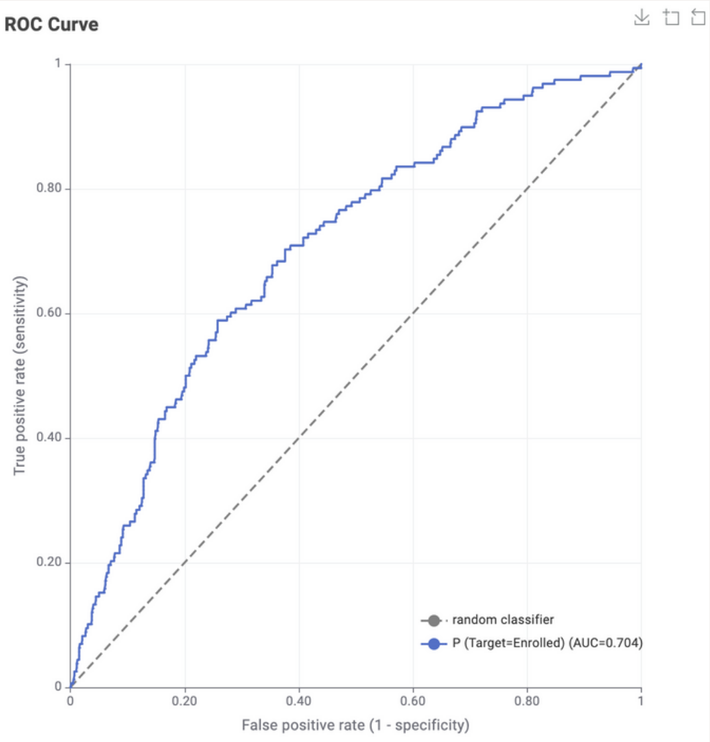
The model's precision is quite low for Enrolled (38%) , while it is 71% for Graduate and 74% for Dropout. Looking at specificity, we have 98% for Enrolled, 62% for Graduate and 88% for Dropout, which are more-than-satisfactory values.
Additionally, another relevant piece of information is the detection of False Negatives, or sensitivity, in particular of the Enrolled students for whom we want to correctly predict the odds of Dropout or Graduate.
Unfortunately, our model has a very low sensitivity of 6% for the Enrolled class, but for the other two classes, the sensitivity values are 93% (Graduate) and 71% (Dropout).

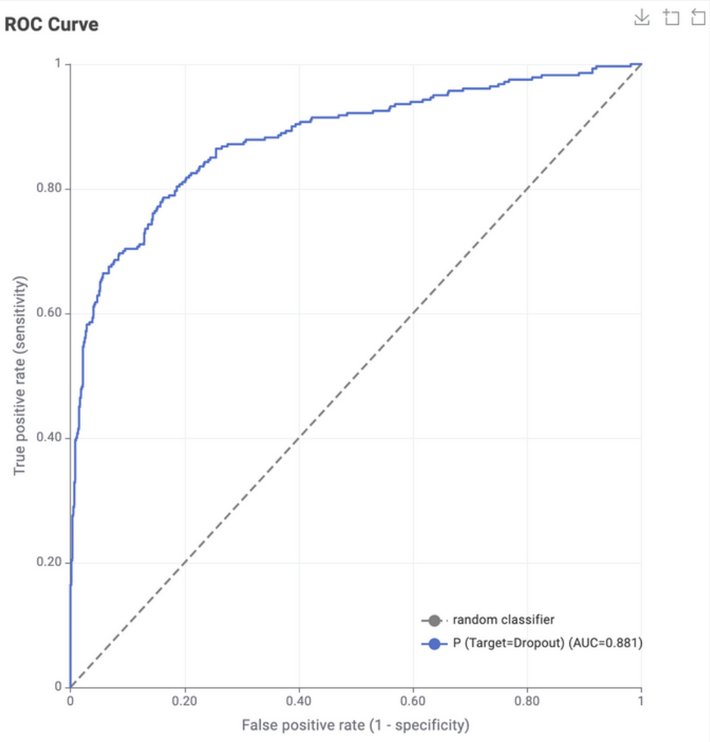
Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Precision	D Sensitivity	D Specificity
Enrolled	10	16	702	148	0.385	0.063	0.978
Graduate	409	170	268	29	0.706	0.934	0.612
Dropout	200	71	525	80	0.738	0.714	0.881

To complete the outcome description, we need a Receiver Operating Characteristic (ROC) curve. However, this curve is specifically designed for binary-classification tasks, as it essentially plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. We therefore plot multiple ROC curves, one for each class, treating each class as the positive class in turn and the rest as the negative class. This way, we hope to get a clearer picture of how the model performs for each individual class.

Random Forest Outcomes

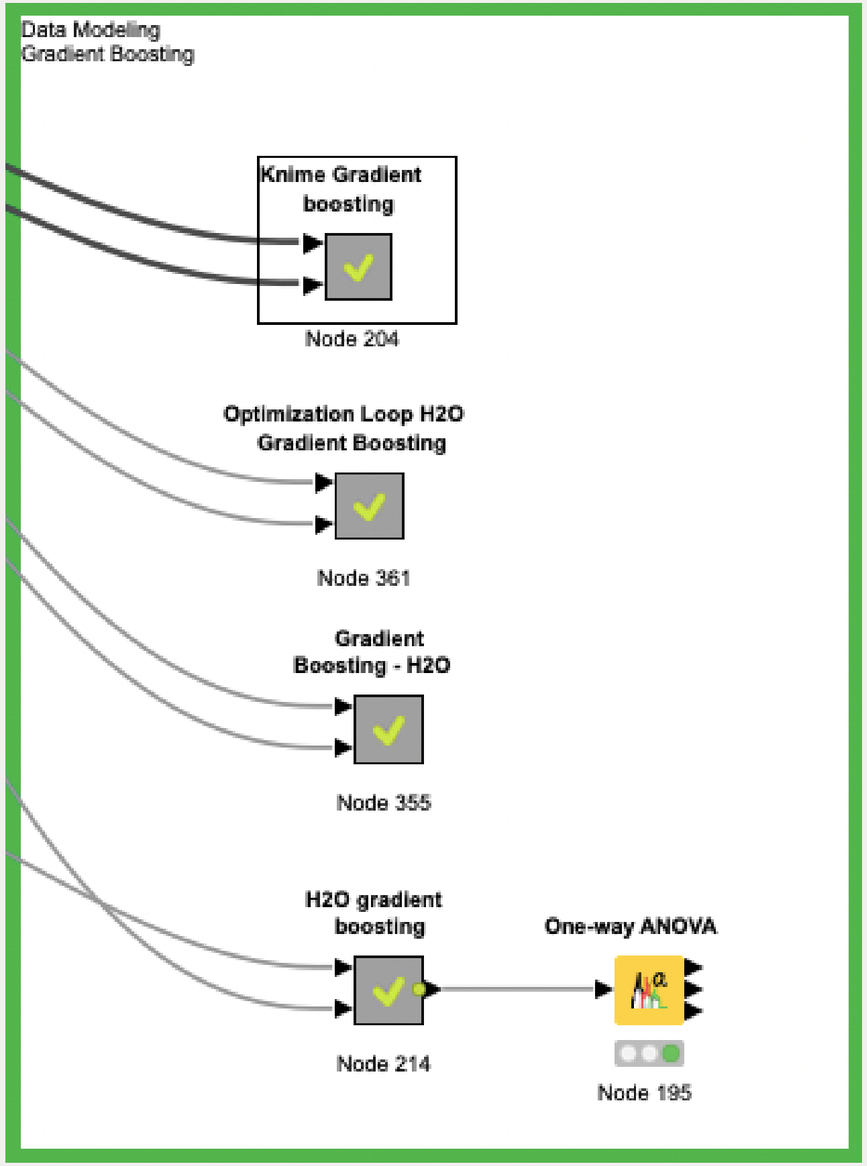






Each curve above is created by plotting the true positive rate against the false positive rate at different threshold values for the classifier. Each point on the ROC curve represents a different trade-off between sensitivity and specificity. The diagonal line (45-degree line) represents the performance of a random classifier. According to the plotted curves, our model shows a great performance for all the classifiers. In particular, for the Graduate students, the Area Under the Curve (AUC) is 0.868, for Enrolled and Dropout the AUC's are slightly lower (0.704 and 0.881, respectively), but anyway satisfactory.

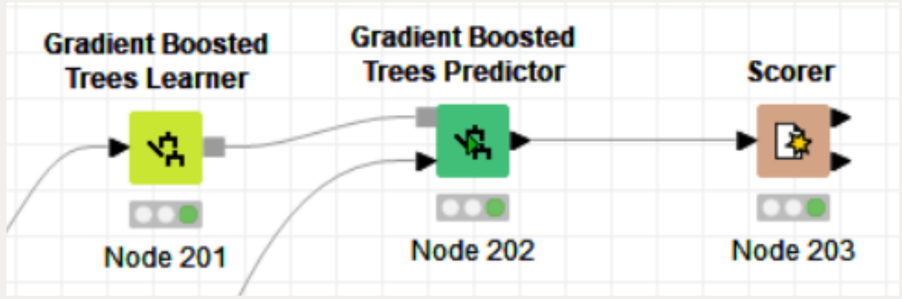
Gradient Boosting Outcomes



We propose 3 types of Gradient Boosting models:

- The one inside the *Knime Gradient boosting* metanode, which is the most basic version of this model, with a simple *Gradient Boosting Trees Learner* node, with an associated predictor and scorer
- The one inside the *Gradient Boosting - H2O* metanode, supported by the optimization loop inside the *Optimization Loop H20 Gradient Boosting* metanode, which uses algorithms based on the H2O.ai platform (an open-source, distributed in-memory machine learning platform) and that are particularly beneficial for training complex models
- The one inside the *H2O gradient boosting* metanode, which uses the exact same model as the previous one, but is connected to a sampling node called *SMOTE,* which oversamples from the initial data to enrich the training data.

Gradient Boosting Outcomes



Target \ Pr...	Dropout	Graduate	Enrolled
Dropout	197	58	25
Graduate	23	399	16
Enrolled	46	83	29

Correct classified: 625 Wrong classified: 251

Accuracy: 71.347% Error: 28.653%

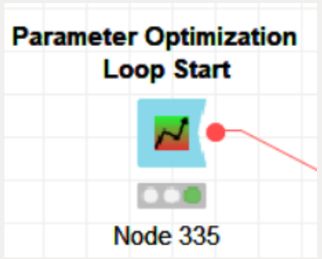
Cohen's kappa (κ): 0.506%

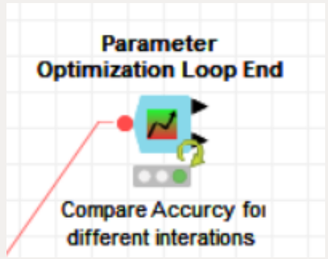
When analyzing the results obtained from the first model, the one trained and evaluated inside the *Knime Gradient boosting* metanode, we see that the overall accuracy is 71,347% with a percentage of error of 28,653%.

In addition to these results, we have other important information such as values for precision, sensitivity, and specificity, obtained from the “Accuracy Statistics” table. In particular, precision is quite high (around 74%) for both Dropout and Graduate classes while is way lower for the Enrolled category (only 41%). Moreover, also the sensitivity measure reports higher values for the Dropout and Graduate classes and a much lower one for the Enrolled.
However, the same is not true for specificity, with 94.3% for enrolled, 88.4% for the dropouts and 67,8% for the graduates.

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Precision	Sensitivity	Specificity
Dropout	197	69	527	83	0.741	0.704	0.884
Graduate	399	141	297	39	0.739	0.911	0.678
Enrolled	29	41	677	129	0.414	0.184	0.943

Gradient Boosting Outcomes





Parameter	Start value	Stop value	Step size	Integer?
minobs	10	100	5.0	☑
maxdepth	2	10	2.0	☑

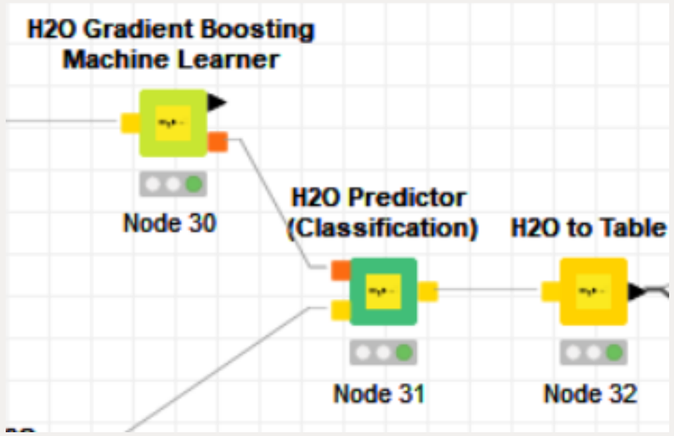
Row ID	I minobs	I maxdepth	D Objective value
Best parameters	20	6	0.724

Before implementing the H2O Gradient Boosting model, we decide to run an optimization loop which will help us with setting the right values for our parameters. This step, widely discussed in the previous section, involves specifying a range of potential values and a step size and then making the algorithm train several models with different parameter combinations, to later keep the best one in terms of accuracy.

Specifically, our optimization loop explores variations in the minimum number of observations and the maximum number of levels (tree depth).

The most effective combination of parameters suggested setting the minimum number of observations to 20 and the maximum levels in the tree to 6.

Gradient Boosting Outcomes



Target \ Pr...	Enrolled	Graduate	Dropout
Enrolled	43	77	38
Graduate	34	380	24
Dropout	33	54	193

Correct classified: 616 Wrong classified: 260

Accuracy: 70.32% Error: 29.68%

Cohen's kappa (κ): 0.499%

In evaluating the outcomes of the H2O Gradient Boosting model, it's evident that the overall accuracy and the error measure align closely with the results obtained from the previous model. However, notable changes have appeared when examining the precision, sensitivity, and specificity measures across different classes.

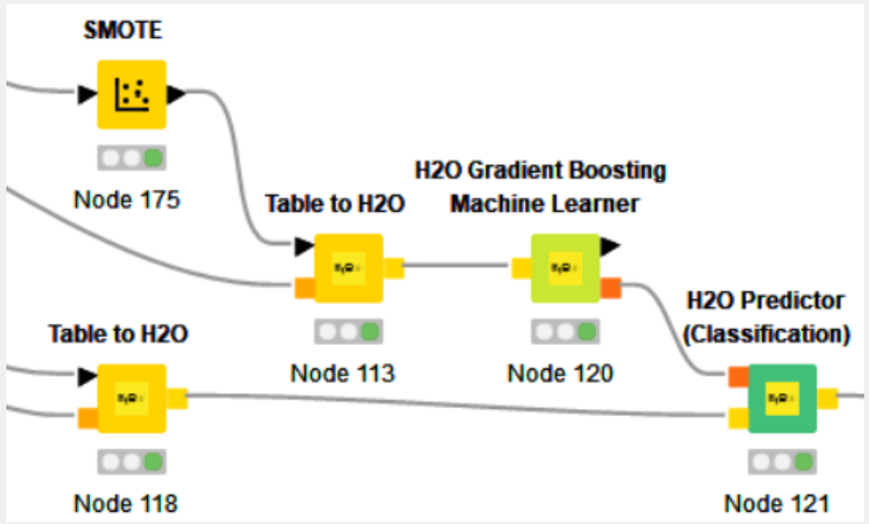
The Dropout class, in particular, reached an impressive 75.7% in precision, making it the category with the highest precision value. On the other side, the Enrolled category experiences a decrease in precision from the previous model, with a value of 39.1%.

In terms of sensitivity, the Graduate class maintains its position with the highest value at 86.8%. Meanwhile, the Dropout class experiences a decrease in sensitivity with respect to the previous model, lowering to 68.9%.

Specificity remains consistently high for both the Enrolled and Dropout categories, floating around 90%. Moreover, for the Graduate category, specificity shows a slightly higher value at 70.1%.

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Precision	D Sensitivity	D Specificity
Enrolled	43	67	651	115	0.391	0.272	0.907
Graduate	380	131	307	58	0.744	0.868	0.701
Dropout	193	62	534	87	0.757	0.689	0.896

Gradient Boosting Outcomes



Target \ Pr...	Enrolled	Graduate	Dropout
Enrolled	46	67	45
Graduate	36	377	25
Dropout	33	48	199

Correct classified: 622 Wrong classified: 254

Accuracy: 71.005% Error: 28.995%

Cohen's kappa (κ): 0.515%

When analyzing the gradient boosting model combined with the SMOTE oversampling method, it is evident that this hybrid model performs a little bit better than the gradient boosting employing the H2O classification model, but falls below the basic model, particularly in terms of accuracy.

Precision measures demonstrate consistency at approximately 75% for both the graduate and dropout categories, indicating a reliable performance in correctly identifying instances of these classes.

In the sensitivity metric, the graduate class holds the best position with an 86.1% accuracy rate, followed by the dropout class at 71.1%. However, the enrolledclass exhibits a lower sensitivity at 29.1%, highlighting an area for potential improvement.

For specificity, high values are observed for both enrolled and dropout categories, standing at approximately 89%, while graduates register a slightly lower specificity at 73.7%.

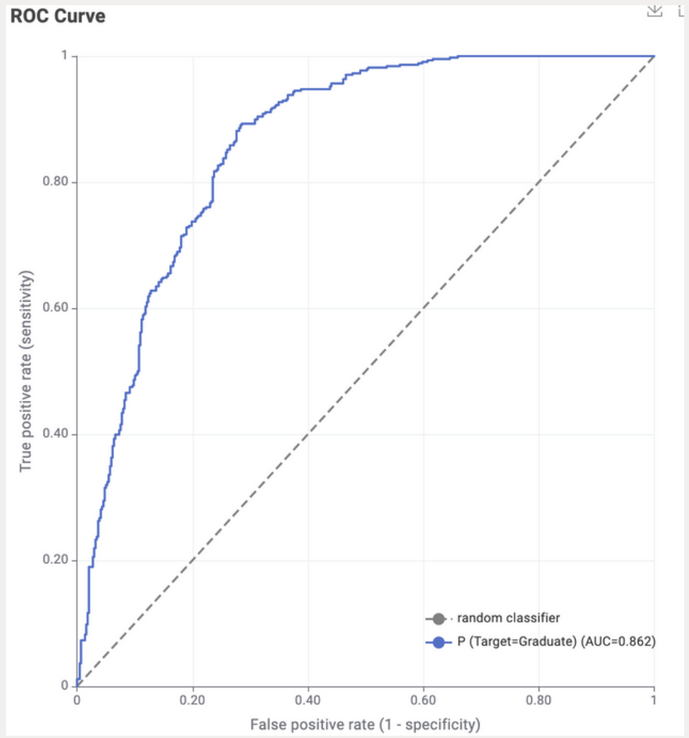
Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Precision	D Sensitivity	D Specificity
Enrolled	46	69	649	112	0.4	0.291	0.904
Graduate	377	115	323	61	0.766	0.861	0.737
Dropout	199	70	526	81	0.74	0.711	0.883

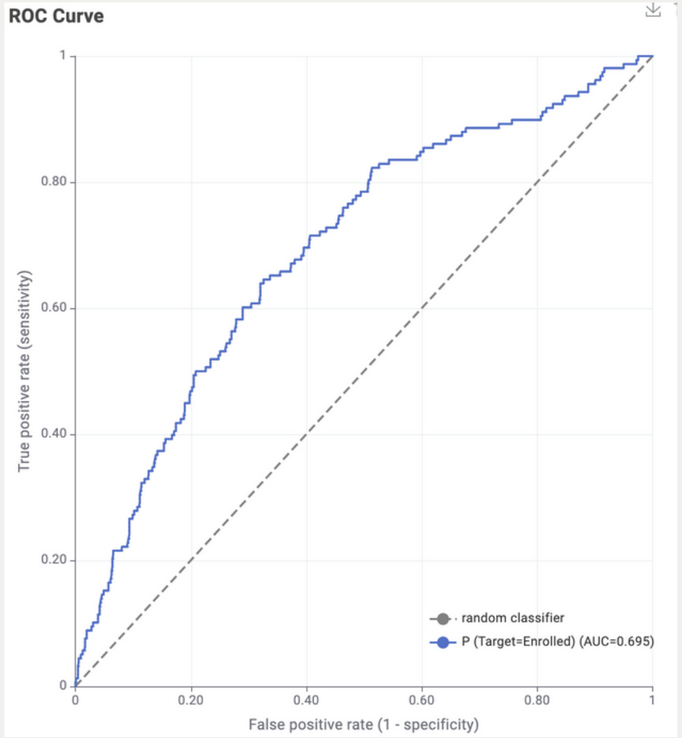
Gradient Boosting Outcomes

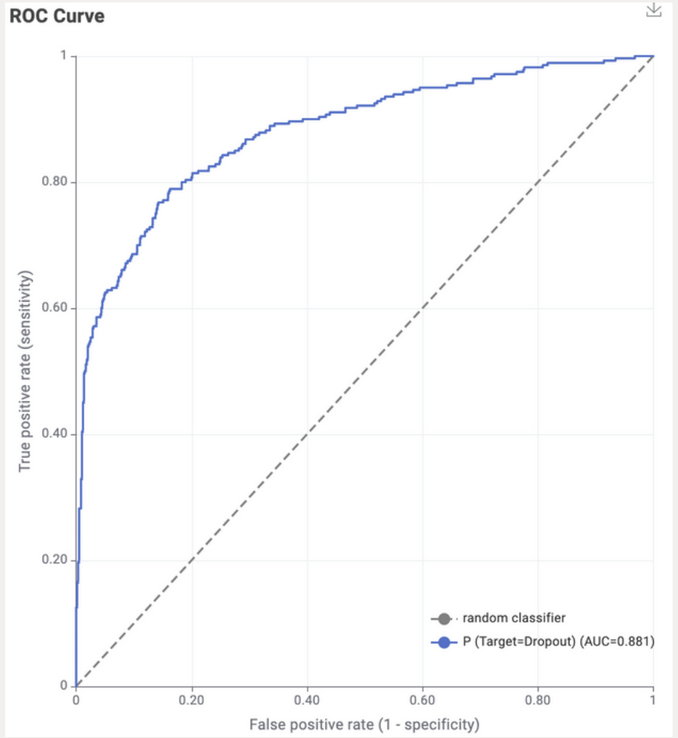
As we did in the Random Forest model, we use the following table to evaluate the most relevant predictors for the accuracy of our model. (this table actually refers to the H2O version of the model). We notice again that “Grade score second sem” is considered the most important variable, contributing 20.5% to the model's predictive accuracy. Feature “Daytime/evening attendance”, with only 0.5% of contribution, is again the lowest one in terms of importance.

Row ID	D Relative Importance	D Scaled Importance	D Percentage
Grade second sem	2,650.022	1	0.205
Grade first sem	1,976.928	0.746	0.153
Course	1,296.746	0.489	0.1
Tuition fees up to date	1,134.527	0.428	0.088
Scholarship holder	956.047	0.361	0.074
Grade change	755.005	0.285	0.058
GDP	726.66	0.274	0.056
Age at enrollment	689.641	0.26	0.053
Fathers occupation	668.477	0.252	0.052
Admission grade	587.673	0.222	0.045
Gender	544.271	0.205	0.042
Fathers qualification	450.841	0.17	0.035
Application order	299.116	0.113	0.023
Previous qualification	128.575	0.049	0.01
Daytime/evening attendance	59.751	0.023	0.005

Gradient Boosting Outcomes







As we previously did, we again plot three distinct ROC curves to evaluate our model's performance in a multiclass classification scenario. Each curve represents one class, treating it as the positive class, while the other two are considered negative. The obtained AUC values for each class are as follows:

- AUC = 0.862 for the Graduate class
- AUC = 0.695 for the Enrolled class
- AUC = 0.881 for the Dropout class

Our expectations were met, as the AUC values align with the precision values observed for each class. The Graduate and Dropout classes, which show higher precision values, also have higher AUC values. This consistency reinforces our confidence in the model's ability to distinguish between the classes, particularly with superior accuracy in predicting graduates and potential dropouts compared to enrolled students.

Model comparison

Logistic regression*

- accuracy: 0.857
- precision: 0.864
- sensitivity: 0.938
- specificity: 0.686
- AUC: 0.892

Prediction ...	Dropout	Not dropout
Dropout	192	37
Not dropout	88	559

Correct classified: 751 Wrong classified: 125

Accuracy: 85.731% Error: 14.269%

Cohen's kappa (κ): 0.655%

Random Forest

- accuracy: 0.707
- precision (wtd avg): 0.658
- sensitivity (wtd avg): 0.707
- specificity (wtd avg): 0.764
- AUC (wtd avg): 0.845

Target \ Pr...	Enrolled	Graduate	Dropout
Enrolled	10	104	44
Graduate	2	409	27
Dropout	14	66	200

Correct classified: 619 Wrong classified: 257

Accuracy: 70.662% Error: 29.338%

Cohen's kappa (κ): 0.481%

Gradient boosting

- accuracy: 0.713
- precision (wtd avg): 0.681
- sensitivity (wtd avg): 0.714
- specificity (wtd avg): 0.792
- AUC (wtd avg): 0.838

Target \ Pr...	Dropout	Graduate	Enrolled
Dropout	197	58	25
Graduate	23	399	16
Enrolled	46	83	29

Correct classified: 625 Wrong classified: 251

Accuracy: 71.347% Error: 28.653%

Cohen's kappa (κ): 0.506%

*Clearly the relatively-high values associated to the Logistic Regression model are partially explained by the fact that this algorithm predicts one of two classes, not one of three. In this setting, a random classifier would have an accuracy of 50%, while in the case of a three-class “Target” variablem, the accuracy of a random model would be 33%

Model comparison

At the end of our analysis, we recall that a direct comparison among the models based on the provided metrics is flawed. In fact, for the Logistic Regression, we have built a binary classifier by combining the Enrolled and the Graduate classes. As a consequence, the performance of this model is better than the other two with an accuracy of 0.857, a sensitivity of 0.938 and and AUC of 0.892.

On the other hand, for the Random Forest and the Gradient Boosting models, with three classes, we decide to explore them separately within each model and compare their ROC curves. Finally, we calculate the weighted average for specificity, sensitivity, and precision to get an overall understanding of the goodness of the model.

As reported on the previous slide, the two models perform similarly, yet the Gradient Boosting one has a higher accuracy (0.713) and sensitivity (0.714) and a relatively satisfactory AUC (0.838), which makes it the best model for the multi-class case.

Managerial implications

Managerial Implications -**Main findings**

As evidenced by the results obtained by our machine learning models, the most influential factors in predicting student dropout rates are the grades attained by students, particularly those achieved in the second semester, whether the student is up-to-date with tuition fee payments, and whether he/she holds a scholarship.

It is important to note that, in the case of “Tuition fees up to date”, we might have a case of reverse causality. Indeed, most likely, if a student is considering dropping out, they will stop paying tuition fees, more than the other way around. This means that it is less likely that not being able to afford tuition fees causes students to dropout, since our data regards public universities in Portugal, and the maximum tuition fee for public universities in Portugal is currently set at €697 per year.

In any case, the findings of our analysis represent valuable information that can be utilized not only by universities, which have a direct interest in preventing student dropouts (for both economic and reputation reasons), but also by tutoring companies. These companies can make use of student data to strategically targeting their advertising efforts towards students at a higher risk of dropping out.

Managerial Implications -**Implications for Universities**

From the perspective of universities, preventing dropouts is crucial. When students leave, universities lose a source of funding (both direct and indirect). Additionally, a high dropout rate can signal issues within the educational system and adversely affect the university's graduation rate, along with its prestige, and consequently the reputation of the professors and students of the university. This rate is a pivotal measure of success and quality, also influencing the institution's ability to attract donations and secure state funding.

A countermeasure for universities to reduce their dropout rates could involve implementing a free internal incentive system. Struggling students would receive free tutoring from high-achieving peers, while those with the highest grades would earn extra credits, further enhancing their academic standing. This solution could particularly aid students facing financial difficulties who might otherwise choose to permanently abandon their studies rather than invest more money in tutoring. Our findings may be beneficial in this regard as they could suggest which variables to consider when choosing students to target with ads about these initiatives.

Another measure to tackle the dropout rates among students facing financial challenges could involve establishing one-on-one counseling sessions focused on educating students about managing student loans and comprehending available financial aid packages. Given that data indicates many students lack a proper understanding of their financing options, this initiative would empower them with the skills to make more informed decisions regarding university expenses.

Managerial Implications -**Implications for tutoring companies**

Regarding how our work can benefit tutoring companies, they can use our findings to better target at-risk students, therefore increasing the ratio of actual customers to total number of people they target with advertisement.

Moreover, by knowing which variables end up affecting students' academic careers the most, they can implement specific programs meant at improving those variables specifically. For instance, since the grades of the second semester appear more relevant than those of the first semester in predicting the academic success of students, tutoring companies might employ more resources in helping students with their second-semester courses.

Other valuable approach could be to provide personalized tutoring sessions tailored to individual student needs.

THANK YOU