Students' dropout and academic success analysis

Hanna Mia Sooman, Johann Mattias Tamm

HannaSooman/Students-dropout-and-academic-success-analysis: Introduction to Data Science course project 2023 (github.com)

Business understanding

Identifying our Business Goals:

Background:

Probably every university struggles with students dropping out to some degree. Many students might just discover that their chosen degree wasn't a right fit for them personally but often the drop-out rates are influenced by many different social, economic, and demographic factors. By understanding these factors, universities and higher education institutions could get more accurate reasoning behind students' success or failure and then take proactive measures to improve overall retention rates.

Business Goals:

Reduce Dropout Rates: Identify factors contributing to student dropout and target minimising these factors' effect to students dropping out.
Enhance Academic Success: Understand the predictors of lower academic success and provide personalised support to students for better outcomes.

Business Success Criteria:

- Decrease overall student dropout rates over the next academic year by a percentage specified by the universities which might wish to use this project's results.
- Achieve a 80% or higher accuracy in predicting students who require targeted intervention.

Assessing the Situation:

Inventory of Resources:

- Comprehensive Kaggle dataset with demographic, socio-economic, and academic performance information.

Requirements, Assumptions, and Constraints:

This project is assuming that:

- the data is unbiased and represents the students accurately as a large group
- the economic and academic factors are relevant

This project has the constraints of:

- limited time
- limited expertise

Risks and Contingencies:

- Risk: Incomplete or inaccurate data.
  Contingency: Implement data cleaning processes.

- Risk: Model overfitting.
  Contingency: Implement overfitting reducing strategies such as regularisation or feature selection.

Terminology:

- Dropout - student who leaves a program, course, or educational institution before completing it.
- Academic success - meeting or exceeding the expected standards of achievement, completing their studies in nominal time
- Demographic data - information about the student including factors like: marital status, nationality, gender, age at enrollment etc.
- Socioeconomic data - information about the student including factors like: their mother's/father's occupation and education, tuition fees, scholarships etc.
- Macroeconomic data - information about the economy at the time of the students enrollment including factors like: unemployment rate, inflation rate, GDP.
- Academic data - information about the student et the time of enrollment, end of the 1st and 2nd semester including factors like: previous qualification, daytime vs evening attendance, grades, etc.

Costs and Benefits:

- Costs do not apply to this project since it is done as a part of a university course.
- Benefits: Decreased dropout rates, enhanced institution students' performance.

Defining our Data-Mining Goals:

Data-Mining Goals:

- Identify key predictors of student dropout.
- Develop predictive models for probable dropout.

Data-Mining Success Criteria:

- Achieve an accuracy rate of 80% or higher in predicting student dropout.
- Identify at least five significant predictors of low academic success.

Understanding our data

Gathering data

We sourced our dataset from Kaggle, the world's largest data science community with over 250 000 high-quality public datasets contributed by professionals and researchers. This dataset provides a comprehensive view of students enrolled in various undergraduate degrees offered at a higher education institution. It contains multiple disjoint databases consisting of relevant demographic data available about the students at the time of enrollment, such as gender, marital status, age of enrollment, nationality and so on. In addition, the dataset also includes social-economic factors such as unemployment rate, inflation rate and GDP from the region and of course students' performance at the end of their first and second semester. All this can be used to analyze what motivates students to stay in school or abandon their studies

The data refer to records of students enrolled in Portugal between the academic years 2008/2009 to 2018/2019. These include data from 17 undergraduate degrees from different fields of knowledge, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. In this dataset, dropouts are defined from a micro-perspective, where field and institution changes are considered dropouts, regardless of when these changes take place. This approach leads to much higher dropout rates than the macro-perspective, which considers only students who leave the higher education system without a degree.

This dataset was created by Valentim Realinho, Jorge Machado, Luís Baptista, & Mónica V. Martins and released in 2021. The work received funding from a program called SATDAP - Capacitação da Administração Pública under grant POCI-05-5762-FSE-000191, Portugal. The authors dedicated the work to the public domain by waiving all of their rights to the work

worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. Copying, modifying, distributing, and performing the work are all allowed, even for commercial purposes, without the need for permission.

Describing data

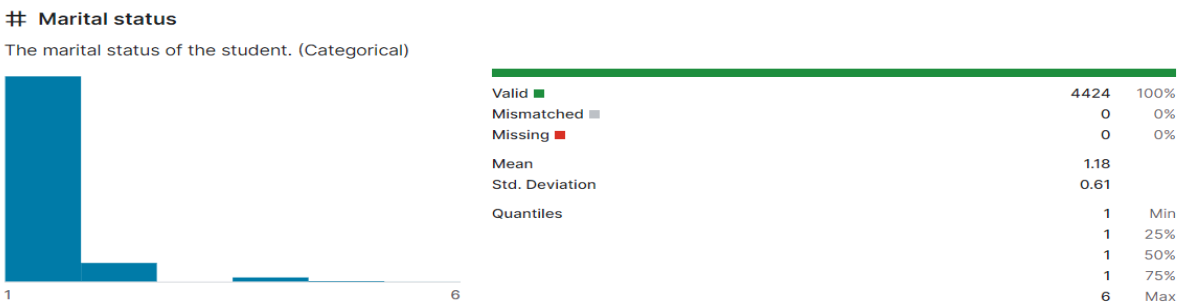
Our data comes in a single .csv file that, as mentioned before, we sourced from Kaggle. The dataset consists of 35 columns and 4424 instances which should be more than sufficient for our analysis. All the expected variables are present in the file and suit our analytical purpose well. The data is of high quality and there are no missing values. In addition to numerical variables, the dataset consists of many categorical variables that make it extremely informative. Fortunately, all of these categories are well defined in a separate file, contributing to a clear and easily understandable classification system across various columns.

While most columns in the dataset are well-defined and comprehensible, a few present challenges in terms of interpretation. Specifically, the columns Curricular units (credited) and Curricular units (without evaluations) lack clear documentation or definitions. The patterns in their distributions, making it challenging to confidently ascertain their intended meanings.

Our primary objectives center around understanding the factors influencing students' academic success and developing a predictive model for anticipating whether a student will complete their degree or potentially drop out. To achieve these goals, our focal point is the target variable. This categorical variable encapsulates the diverse academic trajectories of students and takes on values representing whether a student has successfully completed their degree, dropped out, or is still actively engaged in their studies. As mentioned before, dropouts are defined from a micro-perspective, where field and institution changes are considered dropouts.

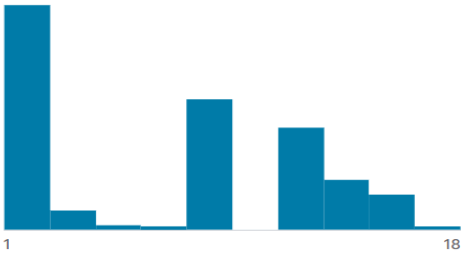
Exploring data

Here are the descriptions, distributions, means, std. deviations and quantiles for all the columns in our dataset:



Application mode

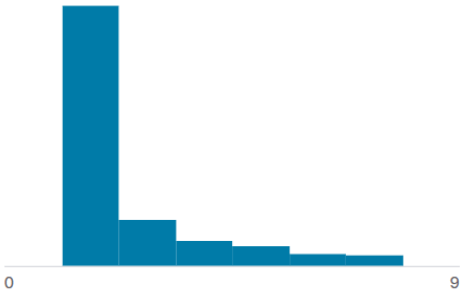
The method of application used by the student. (Categorical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	6.89	
Std. Deviation	5.3	
Quantiles	1	Min
	1	25%
	8	50%
	12	75%
	18	Max

Application order

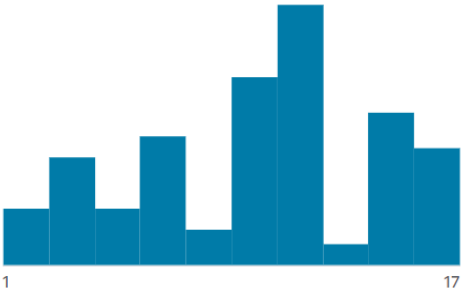
The order in which the student applied. (Numerical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.73	
Std. Deviation	1.31	
Quantiles	0	Min
	1	25%
	1	50%
	2	75%
	9	Max

Course

The course taken by the student. (Categorical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	9.9	
Std. Deviation	4.33	
Quantiles	1	Min
	6	25%
	10	50%
	13	75%
	17	Max

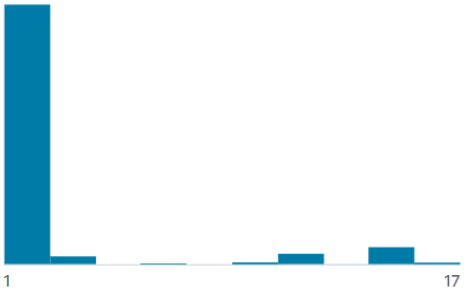
Daytime/evening attendance

Whether the student attends classes during the day or in the evening. (Categorical)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.89	
Std. Deviation	0.31	
Quantiles	0	Min
	1	25%
	1	50%
	1	75%
	1	Max

Previous qualification

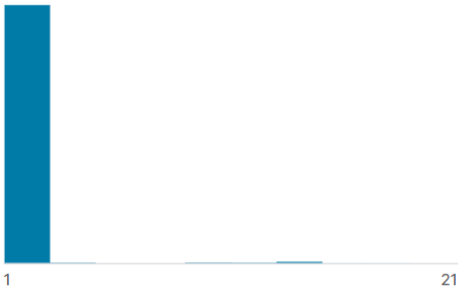
The qualification obtained by the student before enrolling in higher education. (Categorical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	2.53	
Std. Deviation	3.96	
Quantiles	1	Min
	1	25%
	1	50%
	1	75%
	17	Max

Nacionality

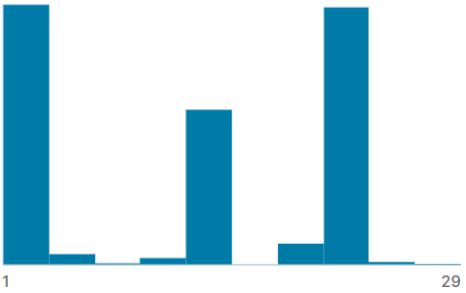
The nationality of the student. (Categorical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.25	
Std. Deviation	1.75	
Quantiles	1	Min
	1	25%
	1	50%
	1	75%
	21	Max

Mother's qualification

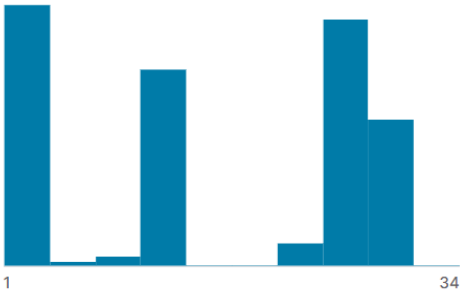
The qualification of the student's mother. (Categorical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	12.3	
Std. Deviation	9.03	
Quantiles	1	Min
	2	25%
	13	50%
	22	75%
	29	Max

Father's qualification

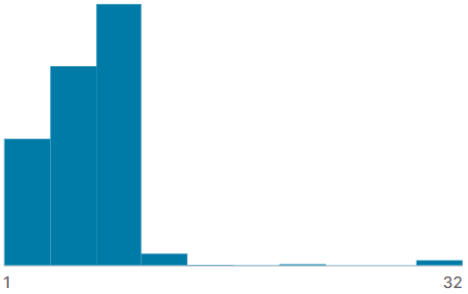
The qualification of the student's father. (Categorical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	16.5	
Std. Deviation	11	
Quantiles	1	Min
	3	25%
	14	50%
	27	75%
	34	Max

Mother's occupation

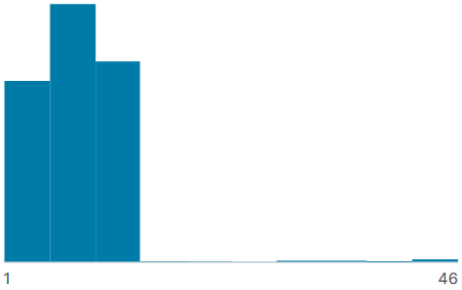
The occupation of the student's mother. (Categorical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	7.32	
Std. Deviation	4	
Quantiles	1	Min
	5	25%
	6	50%
	10	75%
	32	Max

Father's occupation

The occupation of the student's father. (Categorical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	7.82	
Std. Deviation	4.86	
Quantiles	1	Min
	5	25%
	8	50%
	10	75%
	46	Max

Displaced

Whether the student is a displaced person. (Categorical)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.55	
Std. Deviation	0.5	
Quantiles	0	Min
	0	25%
	1	50%
	1	75%
	1	Max

Educational special needs

Whether the student has any special educational needs. (Categorical)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.01	
Std. Deviation	0.11	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

Debtor

Whether the student is a debtor. (Categorical)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.11	
Std. Deviation	0.32	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

Tuition fees up to date

Whether the student's tuition fees are up to date. (Categorical)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.88	
Std. Deviation	0.32	
Quantiles	0	Min
	1	25%
	1	50%
	1	75%
	1	Max

Gender

The gender of the student. (Categorical)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.35	
Std. Deviation	0.48	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	1	Max

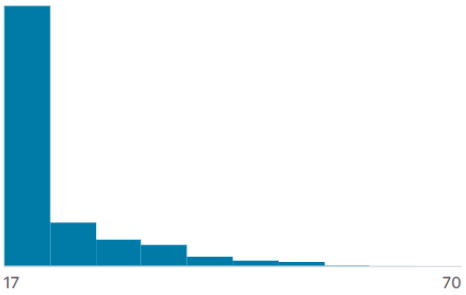
Scholarship holder

Whether the student is a scholarship holder. (Categorical)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.25	
Std. Deviation	0.43	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

Age at enrollment

The age of the student at the time of enrollment. (Numerical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	23.3	
Std. Deviation	7.59	
Quantiles	17	Min
	19	25%
	20	50%
	25	75%
	70	Max

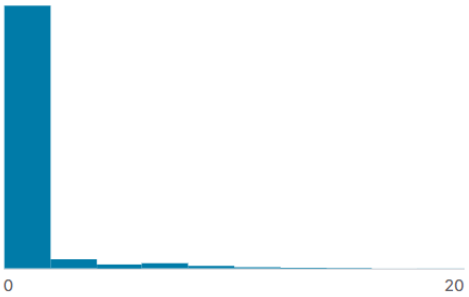
International

Whether the student is an international student. (Categorical)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.02	
Std. Deviation	0.16	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

Curricular units 1st sem (credited)

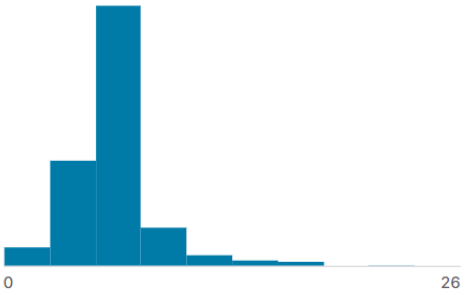
The number of curricular units credited by the student in the first semester. (Numerical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.71	
Std. Deviation	2.36	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	20	Max

Curricular units 1st sem (enrolled)

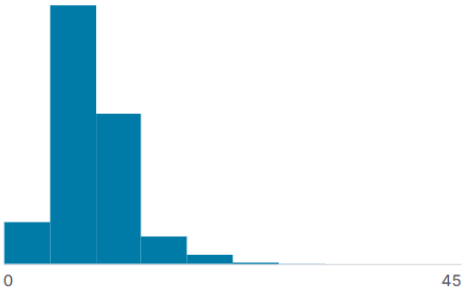
The number of curricular units enrolled by the student in the first semester. (Numerical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	6.27	
Std. Deviation	2.48	
Quantiles	0	Min
	5	25%
	6	50%
	7	75%
	26	Max

Curricular units 1st sem (evaluations)

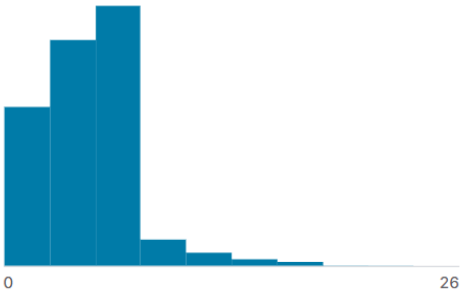
The number of curricular units evaluated by the student in the first semester. (Numerical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	8.3	
Std. Deviation	4.18	
Quantiles	0	Min
	6	25%
	8	50%
	10	75%
	45	Max

Curricular units 1st sem (approved)

The number of curricular units approved by the student in the first semester. (Numerical)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	4.71	
Std. Deviation	3.09	
Quantiles	0	Min
	3	25%
	5	50%
	6	75%
	26	Max

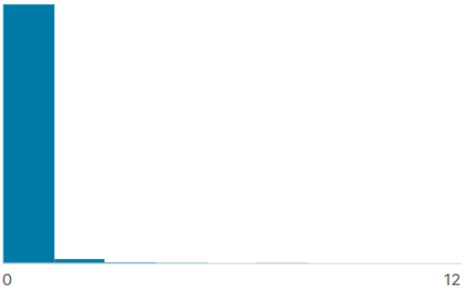
Curricular units 1st sem (grade)

Curricular Units 1st Sem (grade)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	10.6	
Std. Deviation	4.84	
Quantiles	0	Min
	11	25%
	12.3	50%
	13.4	75%
	18.9	Max

Curricular units 1st sem (without evaluations)

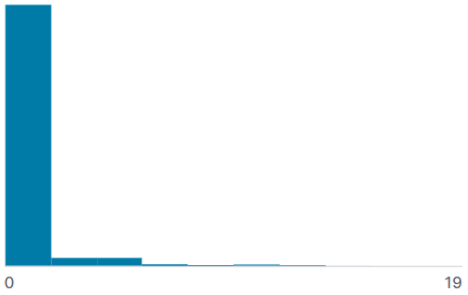
Curricular Units 1st Sem (without Evaluations)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.14	
Std. Deviation	0.69	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	12	Max

Curricular units 2nd sem (credited)

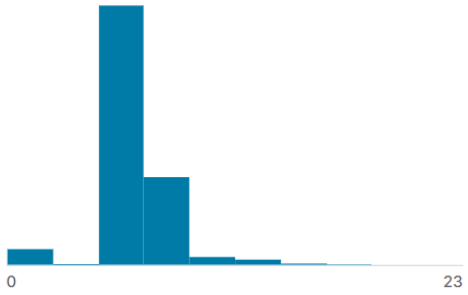
Curricular Units 2nd Sem (credited)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.54	
Std. Deviation	1.92	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	19	Max

Curricular units 2nd sem (enrolled)

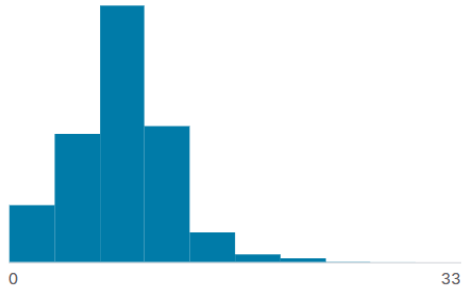
Curricular Units 2nd Sem (enrolled)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	6.23	
Std. Deviation	2.2	
Quantiles	0	Min
	5	25%
	6	50%
	7	75%
	23	Max

Curricular units 2nd sem (evaluations)

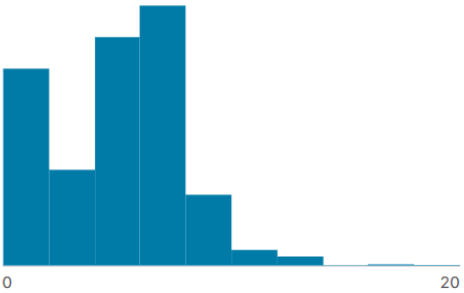
Curricular Units 2nd Sem (evaluations)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	8.06	
Std. Deviation	3.95	
Quantiles	0	Min
	6	25%
	8	50%
	10	75%
	33	Max

Curricular units 2nd sem (approved)

Curricular Units 2nd Sem (approved)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	4.44	
Std. Deviation	3.01	
Quantiles	0	Min
	2	25%
	5	50%
	6	75%
	20	Max

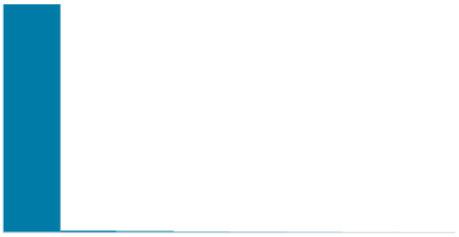
Curricular units 2nd sem (grade)

Curricular Units 2nd Sem (grade)

Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	10.2	
Std. Deviation	5.21	
Quantiles	0	Min
	10.8	25%
	12.2	50%
	13.3	75%
	18.6	Max

Curricular units 2nd sem (without evaluations)

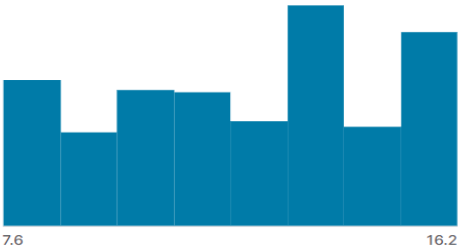
Curricular Units 2nd Sem (without Evaluations)



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.15	
Std. Deviation	0.75	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	12	Max

Unemployment rate

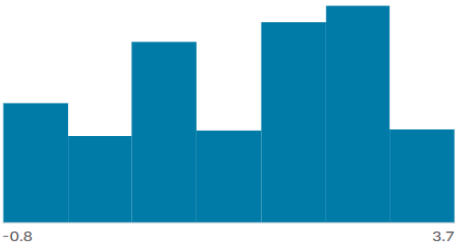
Unemployment Rate



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	11.6	
Std. Deviation	2.66	
Quantiles	7.6	Min
	9.4	25%
	11.1	50%
	13.9	75%
	16.2	Max

Inflation rate

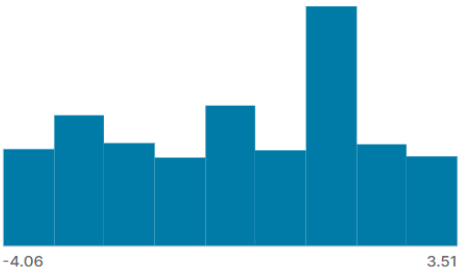
Inflation Rate



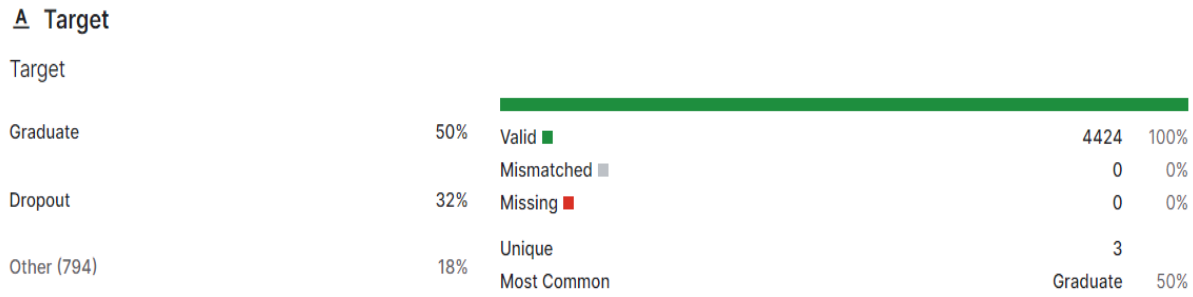
Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.23	
Std. Deviation	1.38	
Quantiles	-0.8	Min
	0.3	25%
	1.4	50%
	2.6	75%
	3.7	Max

GDP

Gdp



Valid	4424	100%
Mismatched	0	0%
Missing	0	0%
Mean	0	
Std. Deviation	2.27	
Quantiles	-4.06	Min
	-1.7	25%
	0.32	50%
	1.79	75%
	3.51	Max



The distribution of the majority of columns demonstrates a coherent pattern, with mean and standard deviation values aligning closely with expected ranges, affirming the dataset's consistency and reliability for our analysis. However, there are a few columns that have an unusual distribution. Columns about the curricular units in the first and second semester called ‘credited’ and ‘without evaluation’ have a distribution where 90% or more values are 0. As mentioned before, the given definitions for these columns also present serious ambiguity, and because of these factors it is unclear at this point whether we are going to use them in our project.

Verifying data quality

After a thorough examination of the dataset, we are pleased to report that the majority of the data is of high quality, with no missing or null values and no outstanding anomalies. Most of the columns and values are clearly defined and easily understandable. Accessing the data is seamless, facilitated by its consolidated form in a single .csv file. This dataset aligns very well with our analytical goals and provides a solid foundation for our project.

However, we have encountered a challenge related to a few columns, specifically 'credited' and 'without evaluation,' concerning curricular units in the first and second semesters. Unfortunately, these columns lack clear definitions and have an unusual distribution, making it difficult for us to interpret their meaning accurately. To address this, we have reached out to the moderators of the dataset to unveil the true significance of these columns. Even if we are unable to ascertain the intended purpose of these columns, the dataset's overall size and the excellent quality of other columns empower us to confidently move forward with our research.

Project plan

1. Data Exploration and Variable Analysis:

- Conduct an in-depth exploration of the dataset, examining distributions, patterns, and relationships among variables.

- Identify key variables that may contribute to students' academic success by analyzing descriptive statistics and visualizations.
- Assess correlations between variables to understand potential dependencies and interactions.

2. Variable Selection and Feature Engineering:

- Select a subset of variables that exhibit significant influence on students' academic success based on insights gained from data exploration.
- Explore the possibility of feature engineering to create new variables that might enhance the predictive power of the model.

3. Data Preprocessing and Cleaning:

- Address any missing or inconsistent values in the dataset through appropriate imputation or removal strategies (if there are any).
- Standardize or normalize numerical features to ensure consistent scaling.
- Encode categorical variables for compatibility with machine learning models.

4. Model Development and Evaluation:

- Split the dataset into training and testing sets for model development and evaluation.
- Choose suitable machine learning algorithms for predictive modeling (like logistic regression, decision trees, ensemble methods).
- Train the model on the training set and evaluate its performance on the testing set using appropriate metrics (accuracy, precision, recall, etc.).

5. Fine-Tuning and Interpretability:

- Fine-tune the model parameters to optimize predictive performance.
- Assess the interpretability of the model by examining feature importance scores.
- Validate the model's performance through cross-validation or other methods.