Team members: Aksel Nodapera, Villem Scheler
Project title: College dropout analysis

Task 2. Business understanding

Identifying our business goals

Background

Student dropout is a growing problem for most universities. When a student quits their studies early, it affects not only the student, who loses time, tuition money and career opportunities, but also the institution, which loses funding and may see a decline in reputation.

The dataset we use comes from a Portuguese university called Polytechnic Institute of Portalegre and contains demographic, economic and academic information collected at enrollment, along with students' first-year performance and final outcome (dropout, enrolled, graduate). This allows us to explore which factors relate most strongly to dropout and success.

Business goals

Through this project, we aim to help the universities:

- Detect students who are at higher risk of dropping out as early as possible.
- Understand how different background, economic and performance factors influence dropout and graduation.
- Support academic advisors in making targeted interventions instead of using broad, generic strategies.
- Improve student retention and graduation rates.

Business success criteria

The project will be considered successful if:

- We can clearly identify student profiles or conditions associated with higher dropout risk.
- The results can be interpreted easily by professors, lecturers and support services.

- Our findings could realistically be used in enrollment or early-warning systems.
- The insights support helpful actions, such as offering tutoring, financial help or academic counseling to reduce dropout.

Assessing the situation

Inventory of resources

For this project, we will use:

- A structured dataset (~471 KB) containing a wide range of demographic, academic and economic variables.
- Documentation from Kaggle explaining variable codes.
- Our own experience in data analysis.

Requirements, assumptions and constraints

- We assume the dataset’s categorical codes and numeric values are valid and consistent.
- Missing information either appears as “Unknown” categories or can be dismissed.
- The target variable is imbalanced (one outcome is much more common), so we must use methods that handle imbalance properly.
- Since time is limited, we will focus on a manageable set of interpretable models and avoid unnecessary complexity.

Risks and contingencies

- There may be data quality issues, such as miscoded entries or outliers. We will address this to the best of our abilities using analysis and data cleaning.
- The dataset may reflect biases, meaning the model might work better for some universities than others. We will monitor this during evaluation.
- Because the data comes from one institution, the results may not generalise perfectly to other universities. We will acknowledge this in our conclusions.

Terminology

- **Dropout:** a student who leaves the studies before completing it.
- **Graduate:** a student who has successfully finished the degree.
- **Enrolled:** a student who is still studying at the time of data collection.
- **Success:** mainly refers to graduating, but also includes continuous enrollment without dropping out.

Costs and benefits

There are no financial costs for conducting this project; the data and tools are freely available. The main investment is our time.
The potential benefits are large: even a slight reduction in dropout can save the university money, increase student success, and improve academic planning. For students, better retention support could mean a better academic path and fewer disruptions.

Defining our data-mining goals

Data-mining goals

Our analytical goals are to:

- Build classification models that can predict whether a student will drop out, stay enrolled, or graduate.
- Explore which features, such as background, parents' education, financial situation or first-year results, are most strongly linked to dropout risk.
- Test different models (decision trees, logistic regression, random forest, k-nearest neighbours).
- Handle class imbalance appropriately.

Data-mining success criteria

We will consider the data-mining phase successful if:

- Our models clearly outperform simple baselines like predicting the majority class.

- Key metrics, especially recall and F1-score for the Dropout class show that the model identifies risk cases reliably.

- The model results are explainable, meaning we can show which factors matter most and why.

Task 3. Data understanding

Gathering Data

Outline data requirements

To understand student dropout patterns, we need data that covers three main areas:

1. **Student background** - demographic information (age, gender, nationality), economic indicators, and parental education/occupation.

2. **Academic history and performance** - previous qualification, application details, first-year academic results, and course information.

3. **Outcome variable** - final outcome of the student: dropout, enrolled or graduate.

These categories allow us to explore not only what happened (the final outcome) but also why it might have happened (the influencing factors).

Verify data availability

All required information is in the provided dataset. The dataset combines enrollment records, economic information, academic performance in the first and second semesters, and indicators such as unemployment rate, inflation rate, and GDP. No additional data sources are necessary. All fields come with clear descriptions, which help us interpret data point values correctly.

Define selection criteria

We decided to use the entire dataset, since the sample size is relatively small and all variables could potentially contain useful indicators. However, during later stages (data preparation and modeling), we may exclude variables that show extremely low variance or no meaningful relationship to the target outcome.

Describing Data

The dataset contains a wide variety of variables:

- **Demographics:** gender, age, nationality, marital status.

- **Academic background:** previous qualification type, application, course, evening/day attendance.

- **Economic indicators:** parents' education and occupation, scholarship, tuition fee payment status.

- **Academic performance:** number of enrolled, approved and average grades.

- **Macroeconomic variables:** unemployment rate, GDP.

- **Target variable:** output (dropout, enrolled, graduate), which forms a three-class classification problem.

Many features are categorical with numerical codes. Some categories are detailed and contain many levels (e.g., parental occupation), meaning we will likely need encoding during preparation.

Exploring Data

Early exploration focused on understanding distributions, identifying dominant categories, and checking relationships with the target variable.

Initial observations

- The dataset is highly imbalanced. Dropout is significantly more frequent than Graduate or Enrolled. This will affect the choice of evaluation metrics and model strategies.

- **Age distribution** is positively skewed, with most students between 18 and 25. A smaller group consists of older or returning students.

- **Gender** is fairly balanced, though slight differences might appear when examining dropout rates.

- **Parental education and occupation** variables have many categories, and some appear infrequently. These may need grouping or simplification.

- **Academic performance variables** show the clearest separations between outcomes: students with high dropout risk tend to have lower numbers of approved units, more failed evaluations, and lower grades in the first semester.

Verifying Data Quality

Missing values and anomalies

A few variables appear to have very low or zero variance in certain categories, which may limit their usefulness.

Consistency checks

Relationships between variables appear consistent (for example approved units never exceed enrolled units).
Macroeconomic indicators are the same across all students, meaning they do not provide variation at the individual level. We may exclude them later.

Data suitability for modeling

The dataset is sufficiently complete for classification tasks. The main quality consideration is class imbalance, which will require techniques like resampling, class weighting, or careful metric selection. Apart from that, the dataset is well-structured and rich enough to support meaningful analysis.

Task 4. Planning Your Project

To complete our student dropout analysis project, we created a structured plan with clearly defined tasks, responsibilities, and tools. Our goal is to divide the workload fairly while making sure both team members contribute to each stage.

Project Tasks and Time Allocation

Data Understanding (10 hours total)

- Aksel: 5 hours - exploring variables, checking distributions.
- Villem: 5 hours - reviewing data descriptions and verifying data quality.

Data Preparation (12 hours total)

- ○ Aksel: 6 hours - cleaning, handling missing values, encoding categories.
- ○ Villem: 6 hours - feature selection, scaling, train/test split.

Modeling (20 hours total)

- ○ Aksel: 10 hours - training baseline models (Decision Tree, Logistic Regression).
- ○ Villem: 10 hours – tuning models, handling class imbalance (e.g., oversampling).

Evaluation and Interpretation (10 hours total)

- ○ Aksel: 5 hours - analyzing performance metrics (accuracy, F1, recall).
- ○ Villem: 5 hours - interpreting feature importance and explaining results.

Report Writing and Presentation (6 hours total)

- ○ Aksel: 3 hours - writing CRISP-DM sections.
- ○ Villem: 3 hours - preparing visualizations and presentation slides.

Extras (4 hours total)

- ○ Aksel: 2 hours - organizing catering for the team (role not final)
- ○ Villem: 2 hours - making jokes for morale boost (role not final)

Methods and Tools

We plan to use Python, along with pandas, NumPy, scikit-learn, and matplotlib/seaborn for visualizations. For handling class imbalance we may use class weighting. GitHub will be used for version control and collaboration.

This plan ensures we follow the CRISP-DM process clearly while sharing responsibilities evenly throughout the project.

Link to our GitHub repository: https://github.com/scheler1/dropoutds