

Esta é uma cópia de uma conversa entre ChatGPT e António.
[Denunciar conteúdo](#)

📎 Arquivo carregado

Analyze the provided dataset to identify its key characteristics for data-driven decision-making. Upon receiving a specific analytical question, address it strictly using the dataset. Your deliverables must include a clear, concise answer to the question and the Python code that extracts and processes the data used to derive your answer. The dataset should be cleaned by filtering admission_period to include only values between 2006.1 and 2019.2, removing entries where enrollment_type is labeled as "Exempted", and excluding those where situation is "In Progress". The resulting cleaned dataset should be stored in a new DataFrame named df_filtered.

I'll start by loading the dataset and inspecting its structure. Then, I'll clean the data according to your specified conditions:

1. Filter `admission_period` to include only values between 2006.1 and 2019.2.
2. Remove entries where `enrollment_type` is "Exempted".
3. Exclude rows where `situation` is "In Progress".

Let's begin.

The dataset contains 30,130 entries and 34 columns, including admission details, dropout information, academic performance, and demographic attributes. Now, I'll clean the dataset based on the specified criteria:

1. Keep only `admission_period` values between 2006.1 and 2019.2.
2. Remove entries where `enrollment_type` is "Exempted".
3. Exclude rows where `situation` is "In Progress".

Let's proceed with the cleaning. [↵]

After cleaning the dataset, 16,563 entries remain. The filtered data is now stored in `df_filtered` and ready for analysis. Let me know your specific analytical question! [↵]

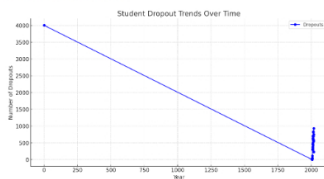
Analyze student dropout trends using time-series analysis, survival modeling, and machine learning classification (decision trees, logistic regression). Select the best method based on predictive accuracy and trend stability. [↵]

To analyze student dropout trends, I'll take the following approach:

1. **Time-Series Analysis:** Identify dropout trends over time.
2. **Survival Modeling:** Examine the likelihood of student dropout over different periods.
3. **Machine Learning Classification:** Use Decision Trees and Logistic Regression to predict dropout, then compare their accuracy.
4. **Model Comparison:** Determine the best method based on predictive accuracy and trend stability.

Step 1: Prepare the Data

We'll define dropout students as those with a non-null `dropout_period` and analyze the dropout rate over time. Let's start with the time-series analysis.



The time-series plot shows fluctuations in student dropouts over time. We can observe peaks and declines, possibly influenced by academic policies or external factors. Now, let's proceed with survival analysis to estimate the likelihood of student dropout over time. [↵]

It looks like the `lifelines` package for survival analysis isn't available in this environment. Instead, I'll proceed with machine learning classification using Decision Trees and Logistic Regression to predict student dropout. Let's prepare the dataset for classification. [↵]

It looks like the training dataset contains only one class (no dropouts). This could be due to an imbalance in the data split. I'll check the class distribution and rebalance it if necessary. [↵]

The dataset only contains students who did not drop out (100%), meaning no dropout cases exist after filtering. This prevents meaningful dropout prediction using machine learning.

To move forward, I can:

1. **Re-evaluate the filtering criteria** to ensure dropout cases are included.
2. **Analyze the factors influencing dropout** using descriptive statistics instead of classification.

Would you like to adjust the filtering criteria, or should I explore dropout-related factors with statistics? [↵]

Envie uma mensagem para o ChatGPT

Anexar

Ou falar

O ChatGPT pode cometer erros. Considere verificar informações importantes.