



Este é uma cópia de uma conversa entre ChatGPT e o usuário.
Desmarcar 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, examining its structure, and then applying the necessary data cleaning steps. Let's begin.

The dataset consists of 30,130 entries and 34 columns. Key columns for filtering include:

- admission_period (float): Must be between 2006.1 and 2019.2.
- enrollment_type (object): Entries labeled as "Exempted" should be removed.
- situation (object): Entries labeled as "In Progress" should be removed.

I'll now apply these filters and create df_filtered. [-]

The dataset has been successfully filtered, reducing the number of records from 30,130 to 16,563. It now excludes:

- Entries where admission_period is outside the range 2006.1 to 2019.2.
- Entries with enrollment_type labeled as "Exempted".
- Entries where situation is "In Progress".

The cleaned dataset is stored in df_filtered. Now, please provide a specific analytical question you'd like me to address using this dataset. [-]

Given the overall average (mean of non-null grades), determine the probability that a student with an average below 7.0 and more than 3 failures will pass the next course. Propose three analytical approaches and select the optimal one.

To address this, I'll proceed with the following steps:

1. Define Key Variables

- The overall average is the mean of all non-null final_average values.
- A student is classified as "below 7.0" if their overall average is < 7.0.
- "More than 3 failures" means having more than 3 occurrences of failing grades.
- A student "passes the next course" if their next final_average is ≥ 5.0.

2. Propose Three Analytical Approaches

- Approach 1: Frequentist Probability Estimation
 - Compute the proportion of students meeting the criteria who passed their next course.
- Approach 2: Logistic Regression Model
 - Use logistic regression to predict passing probability based on final_average and failure count.
- Approach 3: Bayesian Inference
 - Use a Bayesian model to estimate probability, incorporating prior knowledge.

3. Select the Optimal Approach

- The Frequentist Probability Estimation is the most direct and interpretable.
- Logistic regression would be useful if more predictors were involved.
- Bayesian inference is useful with domain expertise but requires prior assumptions.

Thus, I will implement Approach 1 (Frequentist Probability Estimation) to determine the probability. Let's compute it.

Results

- The overall average (mean of non-null final grades) is 6.77.
- The probability that a student with an average below 7.0 and more than 3 failures will pass the next course is 45%.

This was determined using the Frequentist Probability Estimation approach, which directly computes the proportion of students meeting the criteria who passed their next course. Let me know if you need further insights! [-]

Envie uma mensagem para o ChatGPT

📎 Anexar

🗣️ Voz

O ChatGPT pode cometer erros. Considere verificar informações importantes.