

[Entrar](#)[Cadastrar](#)

Esta é uma cópia de uma conversa entre ChatGPT e Anônimo.
[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 it according to your specifications. Let's begin.

The dataset contains 30,130 rows and 34 columns. Key columns include:

- `admission_period`: Represents the period when students were admitted.
- `enrollment_type`: Specifies the type of enrollment (e.g., "Normal", "Exempted").
- `situation`: Describes the student's academic status (e.g., "In Progress", "Approved").

Cleaning Steps:

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

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

The cleaned dataset (`df_filtered`) contains 16,563 rows and 34 columns after applying the specified filters. Let me know your specific analytical question, and I'll extract insights strictly using this dataset. [-]

Identify the key variables that differentiate between passed and failed students by evaluating feature importance. Propose three analytical approaches (e.g., decision tree importance, random forest, logistic regression coefficients) and select the optimal method.

To differentiate between passed (approved) and failed students, I'll analyze feature importance using three analytical approaches:

1. Decision Tree Feature Importance

- Decision trees rank features based on how effectively they split the data.
- Pros: Easy to interpret.
- Cons: Prone to overfitting.

2. Random Forest Feature Importance

- Random forests aggregate multiple decision trees, providing a more stable importance ranking.
- Pros: More robust and accurate.
- Cons: Harder to interpret than a single tree.

3. Logistic Regression Coefficients

- Measures how each variable affects the probability of passing/failing.
- Pros: Great for understanding relationships.
- Cons: Assumes linearity, less effective for complex patterns.

Approach Selection

I will run all three and compare their results. The optimal method will be chosen based on:

- Accuracy
- Stability of importance rankings
- Interpretability

I'll now preprocess the dataset and implement these methods.

Key Findings:

- **Final Average** is the most significant predictor across all models.
- **General Average** is the second most important factor.
- **Admission Period and Admission Age** have moderate importance.
- **Workload and Credits** have minimal influence.

Optimal Method:

- **Random Forest** provides the most balanced and reliable feature importance ranking.
- It reduces overfitting (compared to a single decision tree) while maintaining interpretability.

Recommendation:

To predict student success, focus on **Final Average** and **General Average**, as they are the strongest differentiators between passed and failed students.

Let me know if you need further analysis! 📊 [-]

Envie uma mensagem para o ChatGPT

📎 Anexar

🗣️ Voz

O ChatGPT pode cometer erros. Considere verificar informações importantes.

?