



Entrar

Cadastrar

Esta é uma cópia de uma conversa entre ChatGPT e Andréino.  
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 to understand its contents. Then, I'll apply the specified cleaning steps:

1. **Filter admission\_period** to include only values between 2006.1 and 2019.2.
2. **Remove rows where enrollment\_type** is "Exempted".
3. **Exclude entries where situation** is "In Progress".

Let's proceed with these steps.

The dataset contains 30,130 rows and 34 columns with various attributes related to student admissions, enrollments, and academic progress. Key observations:

- `admission_period` is a float column, so filtering between 2006.1 and 2019.2 will be straightforward.
- `enrollment_type` contains values like "Exempted" that need to be removed.
- `situation` includes "In Progress," which must be excluded.

I'll

The dataset has been successfully filtered, reducing it to 16,563 rows. The cleaned dataset (`df_filtered`) excludes:

- Entries before 2006.1 and after 2019.2 (based on `admission_period`).
- Rows where `enrollment_type` is "Exempted".
- Entries where `situation` is "In Progress".

Now, please specify an analytical question you'd like to address using this cleaned dataset. 🤖 (↩)

Analyze whether there are identifiable patterns that distinguish students who dropped out from those who graduated, based on mode of admission and course workload. Propose three analytical techniques (e.g., clustering, principal component analysis, discriminant analysis) and select the most effective method.

To analyze the differences between students who **dropped out** and those who **graduated**, I'll explore how **mode of admission** and **course workload** impact their outcomes. Here's how I'll approach the problem:

#### Step 1: Define the Key Variables

- **Target Variable:** `exit_form` (either "Graduate" or "Dropout")
- **Predictor Variables:**
  - `admission_form` (categorical: mode of admission)
  - `workload` (numerical: total course workload)

#### Step 2: Select Three Analytical Technique. ↓

I'll consider the following methods:

1. **Clustering (K-Means or Hierarchical Clustering)**
  - Helps group students based on similarities in workload and admission mode.
  - Can reveal natural groupings between graduates and dropouts.
2. **Principal Component Analysis (PCA) with Logistic Regression**
  - PCA reduces dimensionality and identifies the most influential components.
  - Logistic regression will then predict dropout vs. graduation.
3. **Linear Discriminant Analysis (LDA)**
  - Optimizes separation between graduates and dropouts.
  - Works well for classification problems where variables are continuous and categorical.

#### Step 3: Choose the Most Effective Method

- **Clustering** may not be ideal since we already have labels (dropout vs. graduate).
- **PCA + Logistic Regression** can help, but PCA may not capture categorical effects well.
- **LDA** is the best choice because:
  - It directly models the distinction between **graduates** and **dropouts**.
  - It works well when we have both **numerical (workload)** and **categorical (admission mode)** data.

#### Step 4: Apply Linear Discriminant Analysis (LDA)

I'll now implement LDA to determine how well `admission_form` and `workload` distinguish between dropouts and graduates. Let's analyze! 🤖

#### LDA Analysis Results

- **Accuracy:** 56.9%
- **Classification Report Insights:**
  - The model predicts **graduates** (79% recall) and **regular students** (95% recall) reasonably well.
  - However, it fails to identify **dropouts**, **reentries**, and **others**, leading to poor precision and F1 scores.

#### Interpretation

- **Course workload** and **mode of admission** alone are not strong predictors of dropout vs. graduation.
- The model **struggles to differentiate dropouts from other categories**, indicating that additional features (e.g., academic performance, socioeconomic background) may be needed.

#### Next Steps

- Introduce additional variables (e.g., **general average**, **final average**, **dropout period**).
- Consider a **logistic regression model** with interaction terms.
- Explore **decision trees** or **random forests** for better classification.

Would you like to refine the analysis with more features or try a different method? 🤖 (↩)

Envie uma mensagem para o ChatGPT

Anexar

Voz

O ChatGPT pode cometer erros. Considere verificar informações importantes.

