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 examining its structure. Then, I'll clean the data based on your specified criteria:

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

Let me process this now.

The dataset contains **30,130 entries** and **34 columns**. Now, I'll clean the data based on the specified conditions:

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

Let me proceed with the data cleaning. [›]

After cleaning, the dataset now contains **16,563 entries** while retaining all **34 columns**. The data is now ready for analysis. Let me know the specific analytical question you'd like me to address! [›]

Using advanced analysis of grade sequences and academic records, identify key periods in a student's academic trajectory where the dropout risk is significantly elevated. Propose three analytical approaches and select the optimal method.

To identify key periods where the **dropout risk** is significantly elevated, we need to analyze students' **grade sequences and academic records**. Below are three advanced analytical approaches:

---

1. Survival Analysis (Kaplan-Meier & Cox Regression)

Concept:

- This method is used to model **time-to-event data**, where the event is a **dropout**.
- The **Kaplan-Meier estimator** helps visualize the probability of a student remaining enrolled over time.
- The **Cox Proportional Hazards Model** can quantify risk factors that influence dropout (e.g., GPA trends, course difficulty).

Pros:

- ✔ Captures **when** dropout is most likely.
- ✔ Handles **censored data** (students still enrolled).
- ✔ Identifies **high-risk academic periods** (e.g., 1st year, last semesters).

Cons:

- ✖ Requires a well-defined **event occurrence (dropout vs. graduation)**.
- ✖ Assumes proportional hazard rates, which may not always hold.

---

2. Hidden Markov Models (HMM) for Grade Sequences

Concept:

- This probabilistic model detects **hidden states** in a student's academic journey (e.g., **stable, struggling, at-risk, dropout**).
- It learns transitions between states based on **grade sequences and course performance trends**.

Pros:

- ✔ Captures the **dynamic nature** of academic performance.
- ✔ Identifies students who **improve or decline** before dropping out.
- ✔ Can be applied to **early-warning systems**.

Cons:

- ✖ Requires **significant computational resources**.
- ✖ Needs well-defined **state labels** (e.g., What constitutes "at-risk"?).

---

3. Machine Learning Classification (Random Forests / XGBoost)

Concept:

- Train a **predictive model** to classify students as "**Likely to Dropout**" vs. "**Likely to Persist**".
- Use features like **GPA trends, failed courses, credit load, enrollment gaps**.

Pros:

- ✔ High **predictive accuracy** with models like **XGBoost**.
- ✔ Identifies **nonlinear interactions** between academic factors.
- ✔ Can provide **feature importance** (e.g., failing certain courses increases risk).

Cons:

- ✖ Requires **large labeled datasets** (dropout vs. non-dropout).
- ✖ May act as a "black box" with limited interpretability.

Optimal Method: Survival Analysis

Why?

- ☑ Best suited for analyzing **time-based dropout risks**.
- ☑ Provides **actionable insights** on **when dropout risk peaks**.
- ☑ Works well with **academic progression data**.

Next Steps

I'll proceed with a **Kaplan-Meier survival analysis** to pinpoint the semesters where dropout risk is highest. Let me generate the survival curve now.

It looks like the `lifelines` package, which is required for survival analysis, isn't available in this environment. As an alternative, I can analyze dropout trends using descriptive statistics and visualizations, such as **dropout rates per academic period**.

Would you like me to proceed with this alternative approach, or would you prefer a different method? [›]