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.

The cleaned dataset `df_filtered` contains **16,563 records** and **34 columns** after applying the specified filters:

- `admission_period` is between **2006.1** and **2019.2**
- `enrollment_type` is not `"Exempted"`
- `situation` is not `"In Progress"`

Let me know the specific analytical question you'd like me to answer using this filtered dataset. (>_)

Using advanced analysis of characteristics from previously graduated students, develop hypothetical student profiles that are associated with a higher probability of graduation. Propose three approaches (e.g., cluster analysis, generative models, predictive modeling) and select the best option.

To identify student profiles associated with a **higher probability of graduation**, we can explore three advanced analytical approaches using the filtered dataset (`df_filtered`). Here's a summary of each approach and a recommendation on the best one for this task:

---

1. Predictive Modeling (Supervised Learning - Classification)

✅ **Recommended**

Objective:
Train a model to predict whether a student will graduate based on their characteristics (e.g., gender, admission type, admission age, subject performance, etc.).

How it works:

- Label records as "Graduated" or "Not Graduated" using the `exit_form` column.
- Use classification algorithms (e.g., Random Forest, XGBoost, Logistic Regression).
- Analyze **feature importance** to understand which attributes most influence graduation.
- Generate **hypothetical profiles** with high graduation probability using top feature combinations.

Strengths:

- Directly quantifies impact of variables on graduation.
- Allows simulation of profiles and probability scoring.
- Produces actionable insights and is easy to validate.

Weaknesses:

- Requires well-labeled data.
- May overfit if data imbalance isn't handled carefully.

---

2. Cluster Analysis (Unsupervised Learning)

Objective:
Group students into clusters based on shared characteristics to see if specific clusters have higher graduation rates.

How it works:

- Apply K-Means or Hierarchical Clustering on standardized features.
- Label clusters post-hoc with graduation outcomes.
- Identify clusters with higher graduation percentages and analyze their traits.

Strengths:

- Doesn't require predefined labels.
- Can reveal unexpected student subgroups.

Weaknesses:

- Graduation not directly optimized during clustering.
- Harder to translate clusters into practical decision rules.

---

3. Generative Models (Synthetic Profile Generation)

Objective:
Use generative algorithms (e.g., Variational Autoencoders or GANs) to create synthetic student profiles conditioned on "Graduated" outcome.

How it works:

- Train model to generate synthetic data matching the distribution of real graduates.

- Sample and interpret generated profiles.

Strengths:

- Useful for simulations and exploring the "space" of graduates.
- Can generate diverse and non-obvious combinations.

Weaknesses:

- Complex and computationally expensive.
- Less interpretable than traditional models.

🏆 Recommended Approach: Predictive Modeling

Predictive modeling is best suited for this objective because:

- It directly predicts graduation.
- It quantifies the influence of different features.
- It can simulate hypothetical high-likelihood profiles with explainability.

Would you like me to proceed with building a predictive model (e.g., Random Forest or Logistic Regression), and then extract the top characteristics of students who are most likely to graduate?

Follow the best approach to solve this problem based on the dataset.

✅ Model Summary and Evaluation

We trained a **Random Forest Classifier** to predict stude graduation based on several characteristics. Here's ho... it performed:

🎯 Model Performance

Metric	Graduated (1)	Not Graduated (0)
Precision	89.5%	93.3%
Recall	92.1%	90.9%
F1-Score	90.8%	92.1%
Accuracy	**91.5%**	

🔍 Top Features Contributing to Graduation

Below are the most influential features based on feature importance:

Feature	Importance
`general_average`	**34.9%**
`final_average`	20.5%
`admission_form`	14.5%
`admission_age`	13.0%
`color_name`	5.1%
`high_school`	4.7%

📘 Hypothetical High-Graduation Student Profiles

Based on these insights, we can construct hypothetical profiles likely associated with graduation:

🎓 **Profile 1: Academic Excellence**

- General Average: ≥ 8.5
- Final Average: ≥ 7.0
- Admission Form: "Entrance Exam / ENEM"
- Age at Admission: 17–20
- High School: Traditional or Private
- Gender: Any
- Color/Race: Not a determining factor but included for completeness

🎓 **Profile 2: Early Entry with Strong Performance**

- General Average: ≥ 7.5
- Final Average: ≥ 6.5
- Admission Age: ≤ 18
- Admission Form: "Unified Selection System (SISU)"
- Strong coursework completion (high credits and workload)

🎓 **Profile 3: Well-Prepared Affirmative Action Entrant**

- General Average: ≥ 7.0
- Final Average: ≥ 6.5
- Beneficiary of Affirmative Action
- Admission Form: ENEM-based
- Age: 18–20

- High School: With technical or preparatory curriculum

Would you like me to visualize these profiles or simulate probability scores for custom inputs? (>-)

Envie uma mensagem para o ChatGPT

Anexar | Buscar | Refletir | Voz

O ChatGPT pode cometer erros. Considere verificar informações importantes.