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 first inspect the dataset to understand its structure, then apply the necessary cleaning steps to create `df_filtered`. Let's start by loading and examining the dataset.

The dataset contains **30,130 rows and 34 columns.** Now, I'll apply the cleaning steps:

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

Let's proceed with these steps. [>_]

The cleaned dataset, `df_filtered`, now contains **16,563 entries** after applying the filtering criteria.

Let me know your specific analytical question, and I'll extract insights accordingly! [>_]

Using advanced analytical techniques, predict the mode of admission based on variables such as period of admission, gender, affirmative action status, and age of entry. Propose three predictive models and justify the best choice based on performance criteria.

To predict the **mode of admission** (`admission_form`) based on:

- **Admission period** (`admission_period`)
- **Gender** (`gender`)
- **Affirmative action status** (`affirmative_action`)
- **Age of entry** (`admission_age`)

Approach:

1. **Data Preprocessing:**
    - Encode categorical variables.
    - Handle missing values if any.
    - Split data into training and test sets.
2. **Model Selection:**
   I will train and compare three models:
    - **Logistic Regression** (baseline for interpretability)
    - **Random Forest** (handles non-linearity, robust to overfitting)
    - **Gradient Boosting (XGBoost)** (typically best for structured data)
3. **Evaluation Metrics:**
    - **Accuracy** (primary metric)
    - **F1-score** (for class balance assessment)
    - **Confusion Matrix** (to analyze misclassifications)

Now, let's implement these steps.

Model Performance Summary

Model	Accuracy	F1-score
Logistic Regression	46.42%	39.11%
Random Forest	96.62%	95.89%
Gradient Boosting	96.29%	95.04%

Best Model: Random Forest

- **Accuracy:** 96.62%
- **F1-score:** 95.89%
- **Justification:**
    - Significantly outperforms Logistic Regression, which struggles with complex relationships.
    - Slightly better than Gradient Boosting in accuracy and F1-score.
    - Handles categorical data and non-linearity well.

Would you like me to further analyze feature importance or optimize the model? [>_]

Envie uma mensagem para o ChatGPT

Anexar Voz

O ChatGPT pode cometer erros. Considere verificar informações importantes.