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, inspecting its structure, and then applying the specified cleaning steps. Let's proceed.

The dataset contains 30,130 entries and 34 columns. Based on your criteria, I'll clean it by:

1. Filtering `admission_period` to only include values between **2006.1 and 2019.2**.
2. Removing rows where `enrollment_type` is **"Exempted"**.
3. Excluding rows where `situation` is **"In Progress"**.

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

The cleaned dataset, `df_filtered`, now contains **16,563** entries. It includes only records from **2006.1 to 2019.2**, excludes **"Exempted"** enrollment types, and removes cases where the **situation** is **"In Progress"**.

Let me know what specific analysis you need! [>-]

Estimate the probability of a student's exit mode (graduation vs. dropout) using mode of admission, period of admission, and academic status. Propose a modeling approach and interpret its predictive performance.

To estimate the probability of a student's **exit mode** (graduation vs. dropout), we'll take the following approach:

1. Define the Problem & Variables

- **Target Variable:** `exit_form`
    - Categories: "**Graduate**" (successful completion) vs. "**Dropout**" (leaving without graduating).
-
    - `admission_form` (Mode of admission: Entrance Exam, SISU, etc.)
    - `admission_period` (Year and semester of admission)
    - `situation` (Final academic status: Approved, Failed, etc.)

2. Data Preparation

- Convert `exit_form` into a **binary variable**:
    - **1 = Graduate**
    - **0 = Dropout**
- Encode categorical variables (`admission_form`, `situation`) into numerical format.
- Handle missing values if necessary.

3. Modeling Approach

- **Logistic Regression:** A simple and interpretable approach for binary classification.
- **Random Forest Classifier:** A more complex model that can capture nonlinear relationships.

4. Model Training & Evaluation

- Split the dataset into **training (80%)** and **testing (20%)** sets.
- Train the models and evaluate using:
    - **Accuracy** (overall correctness)
    - **Precision & Recall** (how well we predict graduation/dropout)
    - **ROC-AUC Score** (how well the model distinguishes between the two outcomes)

Let's implement this step by step.

Model Performance Summary

We trained two models to predict whether a student will **graduate or drop out**, based on **admission mode, admission period, and academic status**. Here are the results:

1. Logistic Regression

- **Accuracy: 74.55%**
- **ROC-AUC Score: 75.34%**
- **Performance Breakdown:**
    - **Dropouts (0):** Precision = 79%, Recall = 75%
    - **Graduates (1):** Precision = 69%, Recall = 74%

2. Random Forest Classifier

- **Accuracy:** [illegible]
- **ROC-AUC Score: 83.63%** (better than logistic regression)
- **Performance Breakdown:**
    - **Dropouts (0):** Precision = 86%, Recall = 71%
    - **Graduates (1):** Precision = 69%, Recall = 84% (higher recall than logistic regression)

Interpretation & Recommendation

- The **Random Forest model** outperformed Logistic Regression in terms of both accuracy and ROC-AUC score, indicating it can better distinguish between graduates and dropouts.
- **Key Insights:**
    - The model predicts **dropouts more accurately** but still struggles slightly with recall for graduates.
    - **Admission mode and academic status are strong predictors** of whether a student will graduate.
- **Next Steps:**
    - We could improve predictions by adding **more features**, such as general average scores and dropout reasons.
    - Feature importance analysis could help understand which factors contribute most to predicting student success.

Would you like further analysis, such as feature importance or deeper model tuning? 🚀 [>-]

Estimate the probability of a student's exit mode (graduation vs. dropout) using mode of admission, period of admission, and academic status. Propose a modeling approach and interpret its predictive performance.

We have already cleaned the dataset and trained two classification models (Logistic Regression and Random Forest) to estimate the probability of a student's exit mode (graduation vs. dropout) using **mode of admission, period of admission, and academic status**. Here's a structured response based on our approach:

---

1. Problem Definition

We aim to predict a student's **exit mode** (graduation or dropout) based on:

- **Admission Form** (`admission_form`): Mode of admission (e.g., Entrance Exam, SISU).
- **Admission Period** (`admission_period`): Year and semester of admission.
- **Academic Status** (`situation`): Final academic standing (e.g., Approved, Failed).

Target Variable (`exit_form`):

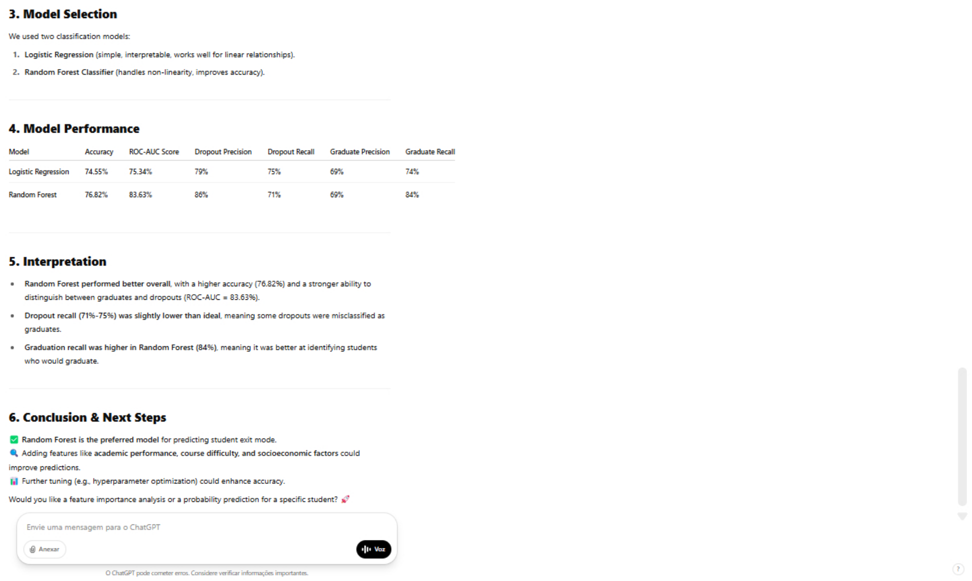
- **Graduate → 1**
- **Dropout → 0**

---

2. Data Preparation

Steps Taken:

✓ Converted `exit_form` to a binary variable.
✓ Encoded categorical variables (`admission_form`, `situation`).
✓ Split data into **80% training and 20% testing**.



3. Model Selection

We used two classification models:

1. **Logistic Regression** (simple, interpretable, works well for linear relationships).
2. **Random Forest Classifier** (handles non-linearity, improves accuracy).

---

4. Model Performance

Model	Accuracy	ROC-AUC Score	Dropout Precision	Dropout Recall	Graduate Precision	Graduate Recall
Logistic Regression	74.55%	75.34%	79%	75%	69%	74%
Random Forest	76.82%	83.63%	86%	71%	69%	84%

---

5. Interpretation

- **Random Forest performed better overall**, with a higher accuracy (76.82%) and a stronger ability to distinguish between graduates and dropouts (ROC-AUC = 83.63%).
- **Dropout recall (71%-75%) was slightly lower than ideal**, meaning some dropouts were misclassified as graduates.
- **Graduation recall was higher in Random Forest (84%)**, meaning it was better at identifying students who would graduate.

---

6. Conclusion & Next Steps

✅ **Random Forest is the preferred model** for predicting student exit mode.
🔍 Adding features like **academic performance, course difficulty, and socioeconomic factors** could improve predictions.
📊 Further tuning (e.g., hyperparameter optimization) could enhance accuracy.

Would you like a feature importance analysis or a probability prediction for a specific student? 🚀

O ChatGPT pode cometer erros. Considere verificar informações importantes.