

Este é uma cópia do seu conteúdo entre ChatGPT e Android.

Desmarcar 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 will first load and inspect the dataset to understand its structure and contents. Then, I'll clean it according to your specified criteria. Let's begin by examining the dataset.

The dataset contains 30,130 rows and 34 columns. The key columns relevant to filtering are:

- `admission_period` (float64): Ranges from at least 1979.2 to 2017.2.
- `enrollment_type` (object): Includes "Exemption", which needs to be removed.
- `situation` (object): Includes "In Progress", which should be excluded.

Now, I'll clean the dataset based on the given criteria. [-]

After applying the filtering criteria, the cleaned dataset `df_filtered` contains 16,563 rows and 34 columns. Let me know the specific analytical question you want to address using this dataset. [-]

Determine whether a student's performance in PROGRAMMING II can be accurately estimated using their performance in PROGRAMMING I and PROGRAMMING LABORATORY I. Outline three analytical approaches and justify the optimal choice based on predictive accuracy.

To determine whether a student's performance in PROGRAMMING II can be accurately estimated using their performance in PROGRAMMING I and PROGRAMMING LABORATORY I, we will follow these steps:

Step 1: Extract Relevant Data

We need to filter `df_filtered` for students who have grades (`final_average`) in the three relevant subjects:

- PROGRAMMING I
- PROGRAMMING LABORATORY I
- PROGRAMMING II (our target variable)

Step 2: Analytical Approaches

We will consider three predictive modeling approaches:

- Linear Regression
 - Assumes a linear relationship between PROGRAMMING II grades and the other two subjects.
 - Useful for interpreting coefficients and understanding direct influence.
 - Assumption: Grades in PROGRAMMING I and PROGRAMMING LAB I contribute additively to PROGRAMMING II.
- Decision Tree Regression
 - Captures non-linear relationships.
 - Can identify thresholds (e.g., a minimum grade in PROGRAMMING I required for success in PROGRAMMING II).
 - Assumption: The relationship between grades may not be strictly linear.
- Random Forest Regression
 - An ensemble method that reduces overfitting common in decision trees.
 - Can handle complex interactions between subjects.
 - Assumption: Multiple decision trees working together can capture hidden patterns.

Step 3: Justification of Optimal Approach

We will evaluate these models based on predictive accuracy, measured using Mean Squared Error (MSE) and R^2 score. The model with the best performance will be chosen.

Now, let's extract the relevant data and proceed with the modeling.

The error indicates that some students have multiple grades recorded for the same subject, creating duplicate entries when attempting to pivot the table. To resolve this, I'll aggregate each student's grades by taking the average final score for each subject before pivoting. Let me fix this and proceed. [-]

The variable `subject` was lost due to the previous error. I'll redefine the models and re-run the training and evaluation. [-]

Model Evaluation and Optimal Choice

The performance of the models is summarized below:

Model	MSE (Lower is Better)	R^2 Score (Higher is Better)
Linear Regression	2.19	-41.26
Decision Tree	5.36	-102.37
Random Forest	1.92	-36.01

Analysis

- The Random Forest model achieved the lowest Mean Squared Error (1.92), meaning its predictions are closer to actual values.
- The Decision Tree performed the worst, with the highest MSE and a significantly negative R^2 score.
- Linear Regression performed better than Decision Tree but worse than Random Forest, indicating that a simple linear relationship might not fully capture the complexities.

Optimal Choice

The Random Forest model is the best choice because it has the lowest MSE, suggesting it generalizes better than the other models.

However, all models have negative R^2 scores, indicating that the predictors (grades in Programming I and Programming Lab I) do not explain much of the variance in Programming II grades. This suggests:

- Other factors might influence Programming II performance (e.g., study habits, attendance, prior coding experience).
- We might need more features to improve prediction accuracy.

Would you like to explore additional factors that could improve prediction, such as dropout rates or workload? [-]

Envie uma mensagem para o ChatGPT

Resposta

Enviar

O ChatGPT pode cometer erros. Considere verificar informações importantes.