

3.2 - pred

17 de julho de 2024



historico_estratificado.csv
Planilha

Analise os dados e limpe as colunas da seguinte forma:
período_ingresso: de 2006.1 até 2019.2
tipo_matricula: remover "Dispensa"
situacao: remover "Em Curso"
o novo dataframe após a limpeza deve se chamar df_filtrado



Vou começar carregando e analisando os dados do arquivo `historico_estratificado.csv` que você enviou. Isso me ajudará a entender como aplicar os filtros que você especificou. Vamos começar lendo o arquivo e dando uma olhada nas primeiras linhas para entender a estrutura dos dados.

Agora, vamos aplicar os filtros especificados:

1. **período_ingresso**: Deve estar entre 2006.1 e 2019.2.
2. **tipo_matricula**: Remover as linhas onde o valor é "Dispensa".
3. **situacao**: Remover as linhas onde o valor é "Em Curso".

Após aplicar esses filtros, vamos criar um novo DataFrame chamado `df_filtrado`. Vamos prosseguir com isso. [\[↗\]](#)

O novo DataFrame `df_filtrado` foi criado aplicando os filtros especificados:

- Mantido apenas os registros com **período_ingresso** entre 2006.1 e 2019.2.
- Removidas as linhas onde **tipo_matricula** é "Dispensa".
- Removidas as linhas onde **situacao** é "Em Curso".

Se precisar de mais alguma análise ou alteração nesses dados, é só me avisar! [\[↗\]](#)

Por meio de uma análise avançada, é possível prever a forma de ingresso com base no período de ingresso, sexo, ação afirmativa e idade de ingresso? Defina 3 opções de como solucionar essa questão e siga a melhor



Para prever a forma de ingresso com base no período de ingresso, sexo, ação afirmativa e idade de ingresso, podemos adotar uma abordagem de aprendizado de máquina para classificação. Aqui estão três opções para abordar essa tarefa:

Opção 1: Regressão Logística

A regressão logística é um modelo estatístico que, apesar de seu nome, é usado para problemas de classificação. Ela pode ser eficaz em casos onde as relações entre as variáveis são aproximadamente lineares e o problema não é muito complexo.

Vantagens:

- Simplicidade e eficiência computacional.
- Facilidade de interpretação dos resultados.

Desvantagens:

- Pode não lidar bem com relações complexas e não-lineares entre variáveis.
- Pode ser limitado se a distribuição de classes for muito desequilibrada.

Opção 2: Árvores de Decisão e Florestas Aleatórias

As árvores de decisão são métodos não lineares que podem capturar relações mais complexas entre as variáveis. A floresta aleatória, que é um ensemble de árvores de decisão, geralmente oferece melhor desempenho por reduzir a variância (overfitting).

Vantagens:

- Capaz de modelar relações não-lineares e interações entre variáveis.
- Robusto a outliers e capaz de lidar com variáveis categóricas diretamente.

Desvantagens:

- Mais complexo e computacionalmente intensivo que a regressão logística.
- Pode ser difícil de interpretar, especialmente as florestas aleatórias.

Opção 3: Redes Neurais

Redes neurais são modelos altamente flexíveis e capazes de capturar padrões muito complexos nos dados. Elas são adequadas para grandes volumes de dados e podem ser ajustadas para melhorar o desempenho através de várias técnicas.

Vantagens:

- Capacidade de modelar relações extremamente complexas e não-lineares.
- Altamente flexível e escalável.

Desvantagens:

- Requer uma grande quantidade de dados para treinamento eficaz.
- Mais difícil de interpretar e requer ajuste fino dos parâmetros (hyperparameters).

Recomendação

Considerando a natureza dos dados e a complexidade típica de problemas educacionais, a **Opção 2 (Árvores de Decisão e Florestas Aleatórias)** pode ser a melhor escolha inicial. Esta abordagem oferece um bom equilíbrio entre complexidade, capacidade de modelagem e interpretabilidade. Além disso, as árvores de decisão e as florestas aleatórias são menos suscetíveis a overfitting em comparação com redes neurais e são capazes de lidar diretamente com variáveis categóricas e numéricas.

Se isso parece bom para você, podemos prosseguir preparando os dados para treinar um modelo de floresta aleatória e verificar seu desempenho.

Vamos seguir com a preparação dos dados e treinamento de um modelo de Floresta Aleatória. Antes de começarmos, vamos verificar a existência e formatar corretamente as variáveis 'sexo', 'ação afirmativa' e 'idade de ingresso' no seu dataset, além de preparar a variável alvo 'forma de ingresso'. Primeiro, vou checar se essas colunas estão presentes e visualizar alguns de seus dados.

Ótimo! As colunas necessárias para nossa análise estão presentes no dataset. São elas:

```
`forma_ingresso`, `periodo_ingresso`, `sexo`, `acao_afirmativa` e `idade_ingresso`.
```

Agora, vamos seguir com a preparação dos dados. Isso inclui:

1. **Codificação das variáveis categóricas:** Transformar variáveis categóricas como ``sexo`` e ``acao_afirmativa`` em formatos numéricos que os modelos de machine learning podem processar.
2. **Divisão dos dados:** Separar o dataset em um conjunto de treinamento e outro de teste para avaliar o desempenho do modelo.
3. **Treinamento do modelo:** Usar o modelo de Floresta Aleatória para treinar com o conjunto de dados de treinamento.

Vamos começar com a codificação das variáveis e preparação dos dados. [\[>\]](#)

O modelo de Floresta Aleatória foi treinado com sucesso e alcançou uma acurácia de aproximadamente 96.12% no conjunto de teste. Isso sugere que o modelo é bastante eficaz para prever a forma de ingresso com base nas variáveis selecionadas (período de ingresso, sexo, ação afirmativa e idade de ingresso).

Se precisar de mais análises, como avaliar a importância das variáveis ou ajustar o modelo para melhorar a performance, fique à vontade para pedir! [\[>\]](#)

