

1.2 - 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



Vamos começar analisando o conteúdo do arquivo que você forneceu para entender sua estrutura e realizar as limpezas solicitadas nas colunas ``período_ingresso``, ``tipo_matricula`` e ``situacao``. Vou ler o arquivo e verificar as primeiras linhas.

O arquivo possui muitas colunas, mas vamos focar nas que você mencionou:

1. ``período_ingresso``: O período de ingresso dos alunos, com tipo float.
2. ``tipo_matricula``: Descreve o tipo de matrícula dos alunos, como "Dispensa" ou "Normal".
3. ``situacao``: O estado atual do aluno no curso, como "Em Curso" ou "Aprovado".

Vamos realizar as seguintes limpezas:

- Filtrar ``período_ingresso`` para incluir apenas de 2006.1 a 2019.2.
- Remover linhas onde ``tipo_matricula`` é "Dispensa".
- Remover linhas onde ``situacao`` é "Em Curso".

Procederei com essas limpezas e criei o novo DataFrame chamado ``df_filtrado``. [?]

O DataFrame ``df_filtrado`` foi criado conforme suas instruções. Ele contém 13.831 registros e 34 colunas. A limpeza envolveu:

- Filtragem de ``período_ingresso`` para o intervalo de 2006.1 a 2019.2.
- Remoção de registros onde ``tipo_matricula`` é "Dispensa".
- Remoção de registros onde ``situacao`` é "Em Curso".

Se precisar de mais alguma alteração ou análise nesses dados, fique à vontade para pedir! [?]

Qual é a probabilidade da forma de saída do aluno ser graduado versus evadido, baseando-se na forma de ingresso, período de ingresso e situação acadêmica? Defina 3 opções de como solucionar essa questão e siga a melhor



Para responder a sua pergunta sobre a probabilidade de um aluno ser graduado versus evadido com base na forma de ingresso, período de ingresso e situação acadêmica, podemos adotar várias abordagens. Aqui estão três possíveis métodos:

Opção 1: Análise Exploratória Simples

Realizar uma análise exploratória simples, usando agrupamentos (``groupby``) para calcular as proporções de alunos graduados versus evadidos para cada combinação de forma de ingresso, período de ingresso e situação acadêmica. Esta abordagem é direta e fornece uma visão rápida das relações nos dados, mas não modela interações complexas entre as variáveis.

Opção 2: Modelo de Regressão Logística

Construir um modelo de regressão logística onde a variável dependente é se o aluno é graduado ou evadido, e as variáveis independentes incluem a forma de ingresso, período de ingresso e situação acadêmica. Este método permite uma análise quantitativa da influência de cada variável e suas interações sobre a probabilidade de ser graduado versus evadido, oferecendo insights mais precisos sobre as relações causais.

Opção 3: Análise de Sobrevivência

Aplicar uma análise de sobrevivência (ou análise de risco) para estimar a probabilidade de um aluno ser graduado versus evadido ao longo do tempo. Este método é especialmente útil se estivermos interessados em como o tempo até a graduação ou evasão varia com base nas características do aluno e pode incorporar alunos que ainda estão em curso como dados censurados.

Escolha da Melhor Opção

Dadas as opções, a **Opção 2: Modelo de Regressão Logística** parece ser a mais adequada. Ela não apenas fornece uma visão geral das probabilidades condicionais de um aluno ser graduado versus evadido com base nas variáveis especificadas, mas também permite controlar outros fatores potenciais e avaliar a importância relativa de cada preditor. Isso nos dá uma compreensão mais aprofundada e quantitativa das dinâmicas envolvidas.

Vamos prosseguir com a criação de um modelo de regressão logística para analisar a probabilidade de saída ser "Graduado" versus "Evadido". Antes de construir o modelo, precisamos preparar os dados e verificar se é necessário transformar ou codificar variáveis categóricas. Vamos começar essa preparação.

Com base na análise inicial, aqui estão algumas observações e os próximos passos para construir o modelo de regressão logística:

Valores Únicos nas Colunas:

- **Forma de Ingresso:** Inclui várias categorias como 'VESTIBULAR / ENEM', 'SISU', 'TRANSFERENCIA', etc.
- **Situação Acadêmica:** Categorias incluem 'Aprovado', 'Reprovado por Falta', 'Trancado', etc.
- **Forma de Saída:** Inclui 'Graduado', 'Evadido', 'Reingresso', entre outros.

Preparação dos Dados:

- Criamos uma variável binária chamada `saida_binaria` para representar se o aluno foi graduado (1) ou evadido (0).
- As contagens mostram que temos 5.757 alunos graduados e 8.074 evadidos em nosso conjunto filtrado.

Próximos Passos:

1. **Codificação de Variáveis Categóricas:** Precisaremos transformar as variáveis categóricas (forma de ingresso e situação acadêmica) em variáveis numéricas usando one-hot encoding, pois modelos de regressão logística não podem interpretar diretamente dados categóricos.
2. **Construção do Modelo:** Após a transformação, podemos construir o modelo de regressão logística utilizando as variáveis transformadas como preditores.

Vou proceder com a codificação das variáveis categóricas e em seguida construir o modelo. [\[?\]](#)

O modelo de regressão logística foi treinado e avaliado. Aqui estão os resultados:

Precisão do Modelo

- **Acurácia Geral:** 76.0%
- **Precisão para 'Evadido' (0):** 82% (alta precisão em prever evasão)
- **Recall para 'Evadido' (0):** 75% (proporção de positivos reais identificados corretamente)
- **Precisão para 'Graduado' (1):** 69% (precisão moderada para prever graduação)
- **Recall para 'Graduado' (1):** 77% (boa capacidade de identificar graduados corretamente)

Interpretação

O modelo mostra que pode prever com razoável precisão se um aluno será graduado ou evadido com base na forma de ingresso, período de ingresso e situação acadêmica. O fato de a precisão para prever evasão ser maior sugere que os padrões associados a evasão podem ser mais consistentes ou distintos em nosso conjunto de dados.

Estes resultados podem ajudar a instituição a identificar alunos em risco de evasão e desenvolver estratégias mais eficazes de retenção. Se precisar de mais análises ou ajustes no modelo, estou à disposição para ajudar! [\[?\]](#)

