

2.1 - presc

17 de julho de 2024

ⓘ Este chat contém arquivos ou imagens produzidos por Análise de Dados Avançada que ainda não estão visíveis nos Chats Compartilhados.



Analise os dados e limpe as colunas da seguinte forma:
periodo_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

- O arquivo de dados foi carregado corretamente e contém muitas colunas. Agora, vamos filtrar os dados conforme as especificações:

1. `periodo_ingresso` deve estar entre 2006.1 e 2019.2.
2. Remover linhas onde `tipo_matricula` é "Dispensa".
3. Remover linhas onde `situacao` é "Em Curso".

Vamos aplicar esses filtros agora. [•]

O dataframe `df_filtrado` foi criado e os dados foram filtrados conforme as especificações:

- Período de ingresso entre 2006.1 e 2019.2.
- Linhas com tipo de matrícula "Dispensa" foram removidas.
- Linhas com situação "Em Curso" também foram removidas.

Este dataframe contém 13,831 entradas, após a limpeza. Se precisar de mais alguma manipulação ou análise desses dados, estou à disposição para ajudar! [•]

De que forma o perfil do aluno que cursou disciplinas do tipo extra-curricular, nunca reprovou e nunca trancou o curso, influencia significativamente o seu sucesso acadêmico? Defina 3 opções de como solucionar essa questão e siga a melhor

- Para analisar como o perfil do aluno que cursou disciplinas do tipo extra-curricular, nunca reprovou e nunca trancou o curso influencia significativamente o sucesso acadêmico, precisamos primeiro definir claramente os critérios de "sucesso acadêmico". Por exemplo, isso poderia ser medido por:

1. **Média Geral de Notas:** A média das notas obtidas em todas as disciplinas cursadas.
2. **Taxa de Conclusão do Curso:** A porcentagem de alunos que concluem o curso no tempo regulamentar.
3. **Realizações Pós-Graduação:** Ingresso em programas de pós-graduação ou posições profissionais de destaque após a graduação.

Com essas definições em mente, podemos considerar três abordagens para responder à pergunta:

Opção 1: Análise Descritiva e Comparativa

Descrição: Realizar uma análise estatística básica, comparando as médias de notas e taxas de conclusão de curso entre alunos que se enquadram no perfil descrito (cursaram disciplinas extra-curriculares, nunca reprovaram, nunca trancaram) e os que não se enquadram.

Método: Usar testes estatísticos para verificar se as diferenças observadas são estatisticamente significativas (e.g., teste t de Student para médias e teste chi-quadrado para proporções).

Opção 2: Modelagem Preditiva

Descrição: Desenvolver um modelo preditivo para estimar o sucesso acadêmico com base em variáveis como tipos de disciplinas cursadas, histórico de reprovações e trancamentos, além de outras variáveis potencialmente relevantes.

Método: Utilizar técnicas de machine learning (regressão logística, árvores de decisão, ou florestas aleatórias) para prever a probabilidade de sucesso acadêmico, e avaliar a importância relativa de cada variável no modelo.

Opção 3: Análise Longitudinal

Descrição: Acompanhar uma coorte de estudantes ao longo do tempo para observar como os critérios definidos (extra-curricular, sem reprovações, sem trancamentos) influenciam o sucesso acadêmico a longo prazo.

Método: Análise de sobrevivência ou modelos de efeitos mistos para avaliar o impacto ao longo do tempo, ajustando para outros fatores confundidores.

Recomendação e Implementação

A **Opção 2**, a modelagem preditiva, é recomendada por ser mais abrangente e oferecer insights sobre como diferentes fatores contribuem para o sucesso acadêmico. Isso permite não apenas verificar a influência do perfil específico, mas também ajustar para outros fatores relevantes, proporcionando uma compreensão mais holística do sucesso acadêmico.

Podemos prosseguir com a implementação dessa opção, começando por preparar e explorar mais profundamente os dados para incluir as variáveis relevantes para a modelagem. Vamos iniciar essa preparação?

Siga a melhor opção para solucionar esse problema



Para prosseguir com a **Opção 2: Modelagem Preditiva**, precisaremos seguir algumas etapas fundamentais. Aqui está o plano:

1. Preparação dos Dados

- **Selecionar as variáveis relevantes:** Incluir variáveis como tipo de matrícula, situação de aprovação nas disciplinas, eventuais trancamentos, e notas finais.
- **Definir o critério de sucesso acadêmico:** Podemos considerar a média final como uma variável contínua ou criar uma variável categórica que indica se o aluno é considerado bem-sucedido (por exemplo, média final acima de um certo limiar).

2. Filtrar Dados

- Selecionar alunos que cursaram disciplinas do tipo extra-curricular.
- Excluir alunos com histórico de reprovações ou trancamentos.

3. Análise Exploratória

- Verificar distribuições das variáveis.
- Calcular estatísticas descritivas para entender melhor o perfil dos alunos.

4. Modelagem

- Dividir os dados em conjuntos de treinamento e teste.
- Treinar modelos de machine learning, como regressão logística, árvores de decisão, ou florestas aleatórias.
- Avaliar os modelos com base em métricas como precisão, recall, e AUC-ROC.

5. Interpretação

- Interpretar os coeficientes/modelos para entender quais variáveis têm mais impacto no sucesso acadêmico.
- Utilizar a importância das variáveis para propor melhorias ou intervenções focadas.

6. Validação

- Validar os resultados com um conjunto de dados de teste para verificar a generalização do modelo.

Antes de iniciarmos, precisamos de mais informações sobre as variáveis disponíveis que podem ser utilizadas para filtrar e definir os alunos de acordo com os critérios mencionados (tipo extra-curricular, nunca reprovaram, nunca trancaram). Vamos examinar novamente os dados para identificar essas variáveis e preparar o dataset adequadamente.

Para proceder com a análise, precisamos identificar e filtrar os dados com base nas seguintes

características:

1. **Disciplinas do Tipo Extra-Curricular:** Não temos uma coluna claramente identificada para isso. Precisamos verificar se "tipo_matricula" ou outra coluna pode indicar disciplinas extra-curriculares.
2. **Nunca Reprovaram:** Podemos usar a coluna `situacao` para verificar se o aluno nunca foi reprovado.
3. **Nunca Trancaram o Curso:** Precisaremos verificar se existe uma coluna que indica trancamento de curso. Talvez `forma_saida` ou `forma_evasao` possam oferecer insights.

Vou verificar primeiramente as categorias únicas em `tipo_matricula`, `situacao`, `forma_saida`, e `forma_evasao` para entender melhor como podemos filtrar os alunos conforme os critérios mencionados. [→]

Agora que temos as categorias das colunas relevantes, podemos definir como filtrar os dados:

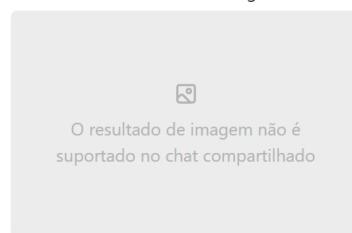
1. **Disciplinas do Tipo Extra-Curricular:** Filtraremos por `tipo_matricula` == "Extra Curricular".
2. **Nunca Reprovaram:** Excluiremos alunos com `situacao` == "Reprovado" ou "Reprovado por Falta".
3. **Nunca Trancaram o Curso:** Excluiremos alunos com `situacao` == "Trancado".

Com esses filtros, podemos preparar o conjunto de dados e depois realizar a modelagem preditiva para analisar o impacto desses fatores no sucesso acadêmico.

Vamos prosseguir com a filtragem dos dados e preparação para a modelagem? [→]

Após aplicar os filtros necessários, temos um conjunto de dados com 673 entradas para análise. As variáveis selecionadas incluem idade de ingresso, sexo, cor/raça, tipo de ensino médio, forma de ingresso, período de ingresso, média final das disciplinas e média geral de aprovação.

Antes de modelar, vamos realizar uma rápida análise exploratória para entender as distribuições das variáveis e a relação potencial com a média final, que consideraremos como indicador de sucesso acadêmico. Vamos visualizar algumas dessas relações? [→]



As visualizações proporcionam uma visão inicial sobre como algumas variáveis se relacionam com a média final, que estamos usando como um indicador de sucesso acadêmico:

- **Distribuição da Média Final:** A maioria dos alunos tem médias agrupadas em torno de valores mais altos, indicando um desempenho geralmente bom.
- **Idade de Ingresso vs. Média Final:** Não parece haver uma correlação clara entre a idade de ingresso e a média final.
- **Média Geral de Aprovação vs. Média Final:** Existe uma relação positiva aparente; alunos com maiores médias gerais tendem a ter médias finais mais altas.
- **Média Final por Sexo:** Não há diferenças significativas na distribuição de médias finais entre os sexos.

Agora, estamos prontos para seguir para a etapa de modelagem preditiva. Podemos codificar variáveis categóricas, dividir os dados em conjuntos de treinamento e teste, e treinar um modelo para prever a média final com base nas variáveis disponíveis. Vamos prosseguir com essa etapa? [→]

Siga a melhor opção para solucionar esse problema

[→] Para seguir com a modelagem preditiva, vamos primeiro codificar as variáveis categóricas e dividir os dados em conjuntos de treinamento e teste. Em seguida, utilizaremos um modelo de machine learning para prever a média final, usando as variáveis selecionadas.

Passos para Modelagem:

1. **Codificação de Variáveis Categóricas:** Transformar variáveis categóricas em numéricas usando

codificação one-hot ou label encoding.

2. **Divisão de Dados:** Separar os dados em conjuntos de treinamento e teste.
3. **Seleção de Modelo:** Treinar um modelo inicial, como a regressão linear, para estabelecer uma linha de base.
4. **Treinamento e Avaliação:** Treinar o modelo e avaliar seu desempenho usando métricas adequadas.

Vamos começar com a codificação de variáveis e a divisão dos dados.

Os resultados da modelagem preditiva com uma regressão linear mostram o seguinte desempenho:

- RMSE (Root Mean Squared Error) no Treinamento: 1.275
- RMSE no Teste: 1.184
- R² (coeficiente de determinação) no Treinamento: 0.052
- R² no Teste: 0.047

Interpretação:

- RMSE indica o erro médio em termos de média final; valores mais baixos são melhores. Os valores obtidos sugerem uma variação média de cerca de 1.2 pontos na previsão da média final, o que é moderadamente alto considerando que as notas podem variar de 0 a 10.
- R² mede a quantidade de variância nos dados de resposta que é explicada pelo modelo. Os valores próximos de zero indicam que o modelo explica muito pouco da variância, sugerindo que pode não ser o modelo mais adequado ou que faltam variáveis explicativas importantes.

Próximos Passos:

Dada a baixa performance do modelo linear, poderíamos considerar as seguintes etapas para melhorar a análise:

1. **Explorar Modelos Mais Complexos:** Modelos não-lineares como árvores de decisão, florestas aleatórias ou gradient boosting podem captar relações mais complexas entre as variáveis.
2. **Engenharia de Recursos Adicional:** Incluir mais variáveis ou transformações dos dados que possam ter influência significativa na média final.
3. **Avaliação de Overfitting e Underfitting:** Refinar o modelo para melhorar o equilíbrio entre bias e variância, ajustando parâmetros ou usando técnicas de regularização.

Caso deseje explorar alguma dessas sugestões ou outra direção, estou à disposição para ajudar! [\[+\]](#)

Siga a melhor opção para solucionar esse problema



Para melhorar nosso entendimento e previsão do sucesso acadêmico com base nas variáveis disponíveis, exploraremos um modelo de machine learning mais complexo que possa captar melhor as relações não lineares entre as variáveis. Neste caso, utilizaremos uma **Floresta Aleatória (Random Forest)**, que é eficaz para lidar com overfitting e capturar complexidades nos dados sem necessidade de extensa pré-configuração.

Etapas:

1. **Modelagem com Floresta Aleatória:** Configurar e treinar um modelo de floresta aleatória.
2. **Avaliação de Desempenho:** Avaliar o novo modelo usando métricas adequadas, como RMSE e R².
3. **Interpretação:** Analisar a importância das variáveis para entender quais contribuem mais para o sucesso acadêmico.

Vamos proceder com a configuração e treinamento do modelo de Floresta Aleatória.

Os resultados do modelo de Floresta Aleatória são os seguintes:

- RMSE no Treinamento: 0.910
- RMSE no Teste: 1.337
- R² no Treinamento: 0.517
- R² no Teste: -0.215

Interpretação:

- A redução do **RMSE no treinamento** indica uma melhor capacidade do modelo em ajustar-se aos dados de treinamento em comparação com o modelo linear anterior.
- No entanto, o **RMSE aumentou no conjunto de teste** e o **R² negativo** sugere que o modelo não está generalizando bem para novos dados, indicando um possível overfitting.

Próximos Passos:

1. **Ajuste de Hiperparâmetros:** Ajustar os parâmetros do modelo de floresta aleatória, como o número de árvores (estimadores) e a profundidade máxima das árvores, para tentar melhorar a generalização.
2. **Validação Cruzada:** Utilizar técnicas de validação cruzada para garantir que o modelo é robusto e estável em diferentes subconjuntos dos dados.
3. **Inclusão de Mais Variáveis:** Explorar a adição de mais variáveis que podem influenciar o sucesso acadêmico, ou revisar a forma como as variáveis categóricas estão sendo codificadas e utilizadas no modelo.

Se você quiser que eu continue com uma dessas etapas, como o ajuste de hiperparâmetros ou a validação cruzada, posso ajudar com isso! [\[+\]](#)

Siga a melhor opção para solucionar esse problema



Para resolver o problema de overfitting e melhorar a generalização do modelo de floresta aleatória, vamos proceder com o ajuste de hiperparâmetros. Isso envolverá explorar diferentes configurações dos parâmetros do modelo para encontrar a melhor combinação que minimize o erro no conjunto de teste e mantenha um desempenho consistente.

Etapas para o Ajuste de Hiperparâmetros:

1. **Definição dos Parâmetros:** Escolheremos alguns parâmetros chave para ajustar, como o número de árvores (`n_estimators`), a profundidade máxima das árvores (`max_depth`), e o número mínimo de amostras necessárias para dividir um nó (`min_samples_split`).
2. **Grid Search com Validação Cruzada:** Usaremos o método de Grid Search combinado com validação cruzada para explorar sistematicamente várias combinações de parâmetros e encontrar a melhor configuração.
3. **Avaliação do Modelo:** Avaliar o modelo ajustado usando métricas de desempenho para garantir que a generalização foi melhorada.

Vamos começar configurando e executando o Grid Search para ajustar os hiperparâmetros.

Parece que houve um problema técnico durante a execução do Grid Search. Vamos tentar executar novamente a configuração e o treinamento do Grid Search para ajustar os hiperparâmetros do modelo de floresta aleatória. Vou iniciar o processo mais uma vez. [\[+\]](#)

Parece que continuamos enfrentando um problema técnico com o Grid Search. Enquanto trabalhamos nessa questão, podemos tentar uma abordagem alternativa para ajustar manualmente alguns parâmetros baseados em heurísticas ou usar uma validação mais simples sem o Grid Search completo para evitar sobreregar o sistema.

Vou ajustar manualmente alguns parâmetros do modelo de floresta aleatória para ver se conseguimos melhorar o desempenho do modelo de forma mais direta. Isso incluirá modificar o número de árvores e a profundidade máxima. Vamos tentar isso agora. [\[+\]](#)

Parece que continuamos enfrentando problemas técnicos. Enquanto isso, posso sugerir algumas abordagens que você pode experimentar localmente ou discutir outras maneiras de explorar e analisar os dados sem depender de métodos computacionalmente intensos.

1. **Validação Cruzada Simples:** Você pode realizar uma validação cruzada manual, treinando o modelo em diferentes subconjuntos dos dados para verificar a consistência dos resultados.
2. **Ajuste Heurístico de Parâmetros:** Ajustar os parâmetros baseados em intuição e testar mudanças incrementais pode ser uma alternativa para encontrar uma boa configuração sem usar métodos automatizados como Grid Search.
3. **Exploração de Modelos Alternativos:** Além da floresta aleatória, modelos como Gradient Boosting ou SVM podem ser explorados para ver se oferecem melhor generalização.

Se precisar de ajuda com uma dessas abordagens ou se tiver outra ideia em mente, estou aqui para ajudar! [\[+\]](#)

Comece com Data Analyst

[Denunciar conteúdo](#) | [Termos de uso](#) | [Política de privacidade](#) |

