

## 1.2 - presc

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 carregando o arquivo que você enviou e dar uma olhada nos dados para entender sua estrutura. Em seguida, prosseguirei com a limpeza de acordo com suas instruções.

O arquivo carregado contém 34 colunas e 30.130 linhas. Vamos agora filtrar os dados de acordo com os critérios especificados:

1. **período\_ingresso:** Deve estar entre 2006.1 e 2019.2.
2. **tipo\_matricula:** Remover entradas com o valor "Dispensa".
3. **situacao:** Remover entradas com o valor "Em Curso".

Vou proceder com essa filtragem. [↗]

Os dados foram filtrados de acordo com os critérios especificados, e o dataframe resultante, chamado `df_filtrado`, agora contém 13.831 entradas. As colunas "tipo\_matricula" e "situacao" foram ajustadas para remover os valores "Dispensa" e "Em Curso", respectivamente, e os períodos de ingresso foram limitados entre 2006.1 e 2019.2.

Se precisar de mais alguma análise ou ação com esses dados, estou à disposição! [↗]

Quais são as variáveis que mais impactam na diferenciação entre alunos aprovados e reprovados? Defina 3 opções de como solucionar essa questão e siga a melhor



Para determinar quais variáveis mais impactam na diferenciação entre alunos aprovados e reprovados, podemos seguir algumas abordagens analíticas distintas:

### Opções para Análise

1. **Análise Estatística Descritiva:** Usar estatísticas descritivas para comparar as médias, medianas, e outras medidas estatísticas entre grupos de alunos aprovados e reprovados. Isso pode incluir análise de variância (ANOVA) para determinar se as diferenças entre grupos são estatisticamente significativas.
2. **Análise de Correlação:** Identificar correlações entre a variável 'situacao' (aprovado/reprovado) e outras variáveis numéricas. Isso poderia incluir análise de correlação de Pearson ou Spearman, dependendo da distribuição dos dados.
3. **Modelagem Preditiva:** Utilizar técnicas de aprendizado de máquina para construir um modelo preditivo que classifique os alunos como aprovados ou reprovados. Este modelo poderia então ser analisado para identificar quais características (variáveis) são mais importantes para a previsão, por meio da importância das características no modelo.

### Melhor Abordagem

A **Modelagem Preditiva** é a opção mais robusta para entender o impacto relativo das variáveis na diferenciação entre alunos aprovados e reprovados, pois permite não só prever os resultados com base nas características, mas também avaliar a importância de cada característica no modelo. Vamos proceder com esta abordagem, construindo um modelo de classificação simples e analisando a importância das variáveis.

Vou começar preparando os dados para o modelo de classificação, selecionando algumas variáveis que podem ser relevantes e criando um conjunto de dados adequado para treinamento.

O modelo de classificação Random Forest alcançou uma alta precisão de aproximadamente 99.6% no conjunto de teste. Isso indica que o modelo é bastante eficaz em distinguir entre alunos aprovados e reprovados com base nas variáveis selecionadas.

A seguir, vamos analisar quais variáveis foram as mais importantes para o modelo na tomada dessa decisão. Isso nos ajudará a entender quais características têm maior impacto na diferenciação entre alunos aprovados e reprovados. [~]

As variáveis que mais impactam na diferenciação entre alunos aprovados e reprovados, segundo o modelo Random Forest, são:

1. **media\_final**: Esta variável é, de longe, a mais importante, com cerca de 93.6% de importância. Isso faz sentido, já que a média final diretamente influencia se um aluno é considerado aprovado ou reprovado em um curso.
2. **periodo\_matricula**: A segunda variável mais importante, com cerca de 3.1% de importância. O período de matrícula pode influenciar a aprovação ou reprovação devido a diversos fatores, como mudanças curriculares ou a dificuldade de cursos em períodos específicos.
3. **idade\_ingresso**: Com aproximadamente 1.2% de importância, a idade com que o aluno ingressa no curso pode ter um impacto menor, mas ainda relevante, na sua performance acadêmica.

Essas informações podem ser utilizadas para desenvolver intervenções focadas ou programas de suporte para aumentar as taxas de aprovação, especialmente considerando a enorme influência da média final dos alunos. Se precisar de mais alguma análise ou ação baseada nestes resultados, estou à disposição! [~]

Comece com Data Analyst

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

?