

2.3 - pred

17 de julho de 2024



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



Vou começar carregando e analisando os dados do arquivo que você forneceu para entender melhor a estrutura do dataset. Em seguida, procederei com a limpeza conforme as suas instruções.

O dataset contém 34 colunas e 30,130 registros. Aqui estão os passos que vou seguir para limpar os dados conforme as suas instruções:

1. Selecionar registros no intervalo de período de ingresso de 2006.1 até 2019.2.
2. Remover registros onde `tipo_matricula` é igual a "Dispensa".
3. Remover registros onde `situacao` é igual a "Em Curso".

Vou aplicar essas limpezas agora. [x]

Os dados foram filtrados com sucesso conforme as suas instruções:

- O intervalo do período de ingresso foi limitado a entre 2006.1 e 2019.2.
- Registros com o tipo de matrícula "Dispensa" foram removidos.
- Registros com a situação "Em Curso" também foram removidos.

O novo dataframe `df_filtrado` agora contém 13,831 registros e 34 colunas.

Se precisar de mais ajustes ou análises, estou à disposição! [x]

É possível determinar a forma de saída de um aluno com base em características como o número de créditos cursados, tipo de matrícula e situação das disciplinas? Defina 3 opções de como solucionar essa questão e siga a melhor



Para determinar a forma de saída de um aluno com base em características como o número de créditos cursados, tipo de matrícula e situação das disciplinas, podemos considerar as seguintes abordagens:

1. **Análise Estatística Descritiva e Visualização:**
 - Examinar a relação entre as características mencionadas e a forma de saída através de estatísticas descritivas e visualizações, como gráficos de barras ou boxplots. Isso pode fornecer insights iniciais sobre padrões e tendências, como a frequência de determinadas formas de saída em relação a diferentes tipos de matrícula ou situações de disciplina.
2. **Modelagem Preditiva (Machine Learning):**
 - Utilizar técnicas de aprendizado de máquina para criar um modelo preditivo que possa estimar a forma de saída com base nas variáveis fornecidas. Modelos como árvores de decisão, florestas aleatórias ou máquinas de vetores de suporte podem ser aplicados para entender melhor as relações entre as variáveis e prever a forma de saída.
3. **Análise de Agrupamento:**
 - Realizar uma análise de cluster para agrupar alunos com características similares e observar se existem padrões comuns na forma de saída dentro de cada cluster. Isso pode ajudar a identificar perfis de alunos que tendem a seguir trajetórias de saída específicas.

Sugestão e Ação:

Fazer uma breve descrição das opções sugeridas e como elas podem ser implementadas no projeto.

Entre essas opções, a abordagem de **modelagem preditiva** (**Machine Learning**) pode ser a mais robusta e informativa, pois permite não apenas entender as relações entre as características e a forma de saída, mas também criar um modelo que pode ser utilizado para prever a forma de saída de alunos com base em suas características observadas. Isso oferece tanto insights quanto uma ferramenta prática de previsão.

Vou proceder com a preparação dos dados para criar um modelo de machine learning, utilizando uma técnica adequada para classificação. Vou verificar a distribuição da variável 'forma_saida' e a disponibilidade das outras características no dataframe para garantir que podemos construir um modelo eficaz.

A variável '**forma_saida**' tem as seguintes categorias principais: Graduado, Evasão, Regular, Reingresso, e algumas outras categorias sob 'Outros'. Aqui está um resumo das características relevantes:

- 'numero_creditos' varia principalmente entre 0 a 10, com a maioria dos cursos valendo 4 créditos.
- 'tipo_matricula' inclui 4 categorias únicas, com 'Normal' sendo a mais comum.
- 'situacao' apresenta 7 estados diferentes, com 'Aprovado' predominando.

Para construir um modelo preditivo, precisaremos:

1. Codificar variáveis categóricas (tipo_matricula, situacao) em números para que possam ser utilizadas em algoritmos de machine learning.
2. Tratar os dados para garantir que não haja valores faltantes nas variáveis de entrada.
3. Escolher um algoritmo apropriado para classificação multiclasse.

Vou começar a preparação dos dados e, em seguida, selecionar e treinar um modelo adequado.

Vamos prosseguir? [\[+\]](#)

sim



O modelo de **Random Forest** alcançou uma precisão geral de aproximadamente 54.5% no conjunto de teste. Aqui estão alguns detalhes da avaliação do modelo:

- **Evasão:** Precision de 62%, Recall de 60%, F1-Score de 61%.
- **Graduado:** Precision de 51%, Recall de 89%, F1-Score de 65%.
- **Regular:** Precision de 63%, Recall de 16%, F1-Score de 26%.
- **Reingresso:** Não houve acertos para esta categoria, com todos os scores sendo 0.
- **Outros:** Também não houve acertos para esta categoria, com todos os scores sendo 0.

O alto recall para "Graduado" sugere que o modelo é eficaz em identificar essa categoria, mas muitas vezes à custa de classificar incorretamente outras categorias como "Graduado". A falta de previsões corretas para "Reingresso" e "Outros" indica que o modelo pode precisar de mais dados ou características diferenciadas para essas categorias.

Aqui estão algumas sugestões para melhorar o desempenho do modelo:

1. **Aumentar o conjunto de características:** Incluir mais variáveis que possam estar correlacionadas com a forma de saída, como idade de ingresso, desempenho acadêmico, ou histórico de matrículas anteriores.
2. **Balanceamento de Classes:** As classes "Outros" e "Reingresso" estão sub-representadas, o que pode ser mitigado através de técnicas de oversampling ou undersampling.
3. **Experimentar outros modelos:** Modelos como XGBoost ou redes neurais podem capturar relações mais complexas entre as características.

Gostaria de explorar alguma dessas melhorias ou há algo mais que você gostaria de investigar? [\[+\]](#)

Siga a melhor opção para solucionar esse problema



Dada a distribuição desigual das classes e a necessidade de incorporar mais características para melhorar a precisão do modelo, a sugestão mais promissora é **aumentar o conjunto de características** e também considerar o **balanceamento de classes**.

Vou começar com estas etapas:

1. Incluir mais variáveis que possam estar correlacionadas com a forma de saída.
2. Aplicar técnicas de balanceamento de classes para melhorar a representação das classes minoritárias no modelo.

Vamos começar pela inclusão de novas variáveis. Analisarei as colunas disponíveis para identificar quais podem ser relevantes para adicionar ao modelo. Em seguida, tratarei o balanceamento das classes.

Baseado nas colunas disponíveis, aqui estão algumas variáveis adicionais que podemos considerar incluir no modelo:

1. **idade_ingresso**: A idade na qual o aluno ingressou pode ser relevante, pois pode influenciar a trajetória educacional.
2. **apv_media_geral**: A média geral do aluno até o momento pode indicar seu desempenho acadêmico e influenciar a forma de saída.
3. **periodo_relativo**: O período relativo no curso pode dar uma ideia de quanto avançado o aluno está em seu currículo acadêmico.
4. **sexo**: Características demográficas como o sexo podem oferecer insights sobre padrões de evasão ou graduação.
5. **carga_horaria**: A carga horária total pode refletir o comprometimento do aluno com o curso.

Agora, vou preparar os dados incluindo essas variáveis adicionais e também aplicar técnicas de balanceamento de classes. Vamos iniciar pelo tratamento e inclusão dessas novas características no conjunto de dados. [\[+\]](#)

Parece que o ambiente atual não tem a biblioteca `imblearn` disponível, que é essencial para aplicar a técnica de balanceamento de classes através de SMOTE. No entanto, posso seguir com a adição das variáveis e normalização dos dados para melhorar o modelo. Vamos continuar sem o balanceamento de classes neste momento e avaliar a melhoria potencial com as novas características. Vou proceder com essas etapas e reavaliar o modelo. [\[+\]](#)

Parece que ocorreu um erro porque a variável `scaler` não foi definida no contexto atual devido à interrupção anterior. Vou definir e aplicar o `StandardScaler` novamente junto com o modelo para garantir que tudo esteja configurado corretamente. Vamos proceder com isso. [\[+\]](#)

O modelo atualizado, agora incluindo novas características, manteve uma precisão de aproximadamente 54.5%, similar ao modelo anterior sem as novas variáveis. Aqui estão os detalhes da avaliação:

- **Evasido**: Precision de 62%, Recall de 60%, F1-Score de 61%.
- **Graduado**: Precision de 51%, Recall de 89%, F1-Score de 65%.
- **Regular**: Precision de 63%, Recall de 16%, F1-Score de 26%.
- **Reingresso e Outros**: Ambos ainda apresentam scores de 0 para todas as métricas.

Esses resultados indicam que, mesmo com a adição de mais informações, o desempenho do modelo não melhorou significativamente, o que pode ser atribuído em parte à falta de balanceamento adequado das classes. Sem o uso de técnicas de balanceamento de classes como SMOTE (que não pudemos aplicar devido às limitações do ambiente), há um limite para a melhoria que podemos alcançar dada a predominância de certas classes sobre outras.

Como próximos passos, poderíamos:

1. Experimentar outros algoritmos de machine learning, que possam lidar melhor com dados desbalanceados.
2. Aplicar técnicas de engenharia de recursos mais avançadas, possivelmente criando interações entre características ou utilizando análise de componentes principais (PCA) para redução de dimensionalidade.
3. Adquirir mais dados ou informações adicionais que possam ajudar a distinguir entre as categorias de forma de saída.

[\[+\]](#)

Comece com Data Analyst

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

