Machine Learning

Sistema de Intervenção Estudantil

EIACD

Afonso Santos | Carolina Jesus | Rafael Santos

2024/25

Objetivo do Trabalho

O trabalho tem como objetivo desenvolver um sistema capaz de prever se um estudante do ensino secundário irá passar no exame final, com base nas suas características demográficas, sociais e escolares.

Está apresentado em formato Jupyter Notebook, acompanhado por um ficheiro Python (‘utils.py’). A construção de um sistema de intervenção estudantil utiliza um pipeline de ciência de dados que inclui:

1. Exploração dos Dados
2. Pré-processamento dos Dados
3. Balanceamento de Classes
4. Modelação e Avaliação

Exploração dos Dados

Carregamento dos Dados

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...
1	GP	F	17	U	GT3	T	1	1	at_home	other	...
2	GP	F	15	U	LE3	T	1	1	at_home	other	...
3	GP	F	15	U	GT3	T	4	2	health	services	...
4	GP	F	16	U	GT3	T	3	3	other	other	...
...	...	...	...	...	...	...	...	...	...	...	...

Análise de Valores Ausentes e Desconhecidos

Ausência de Valores “Unknown” e valores nulos (NaN)

Análise da Estrutura dos Dados

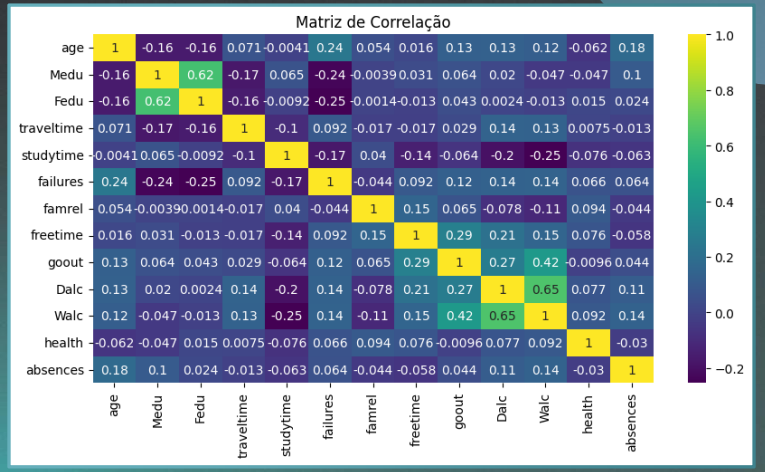
```
Number of rows:  395
Number of columns:  31
school        object
sex           object
age            int64
address       object
famsize       object
```

Exploração dos Dados

Análise de Variáveis Categóricas

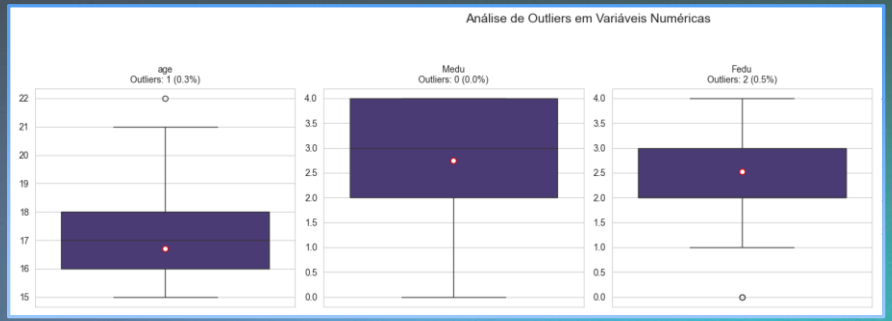
Análise de Correlação

Não existe nenhuma variável redundante.



Análise de Outliers

Os valores dos outliers estão dentro do expectável.



Pré-processamento de Dados

Codificação de Variáveis Categóricas

codificação one-hot às variáveis categóricas, transformando-as em formato numérico que os algoritmos de machine learning possam interpretar.

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	...
0	18	4	4	2	2	0	4	3	4	1	...
1	17	1	1	1	2	0	5	3	3	1	...
2	15	1	1	1	2	3	4	3	2	2	...
3	15	4	2	1	3	0	3	2	2	1	...
4	16	3	3	1	2	0	4	3	2	1	...

Divisão dos Dados

Divisão dos dados pré-processados em conjuntos de treino (70%) e teste (30%) - crucial para treinar os modelos com a maioria dos dados e avaliá-los de forma imparcial em dados não vistos.

Balanceamento de Classes

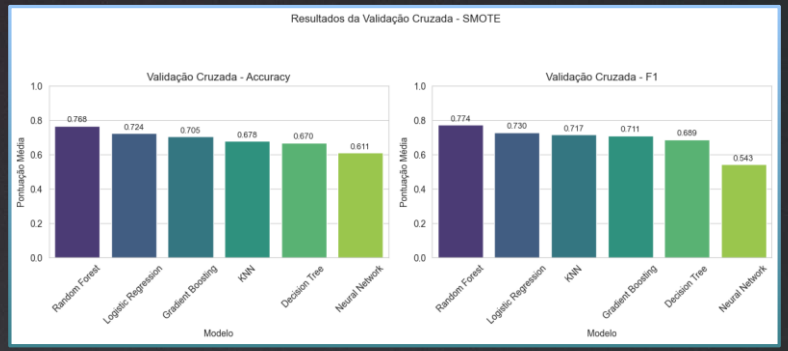
Análise do Desequilíbrio de Classes

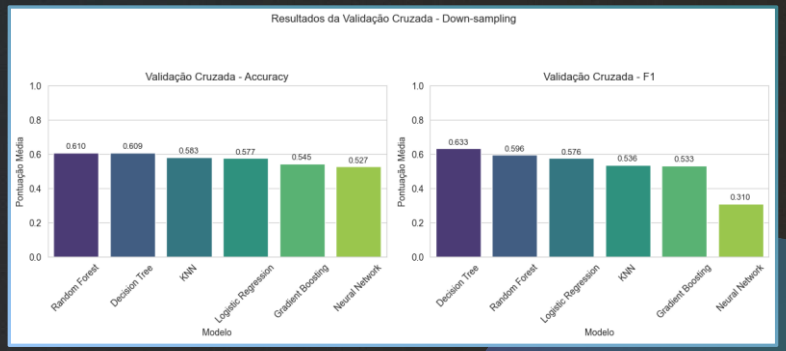
passed
yes | 67.088608
no | 32.911392

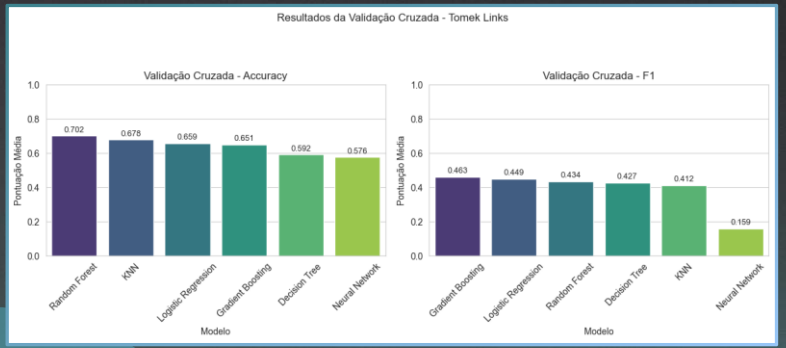
Aplicação de Técnicas de Balanceamento

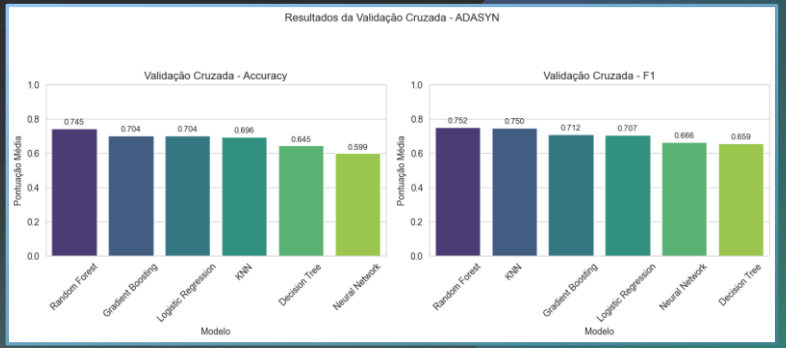
Down-sampling
SMOTE
ADASYN

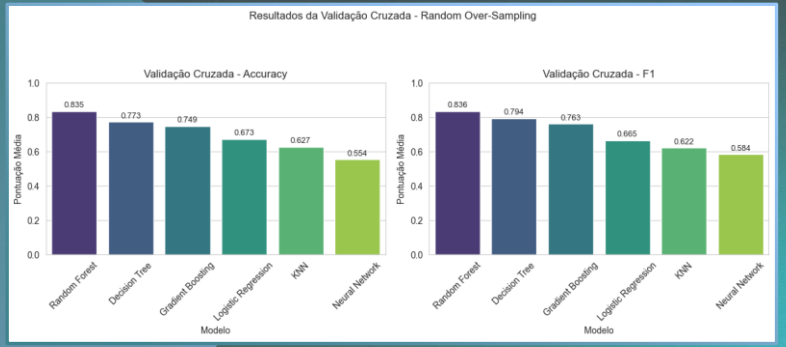
Modelação e Avaliação

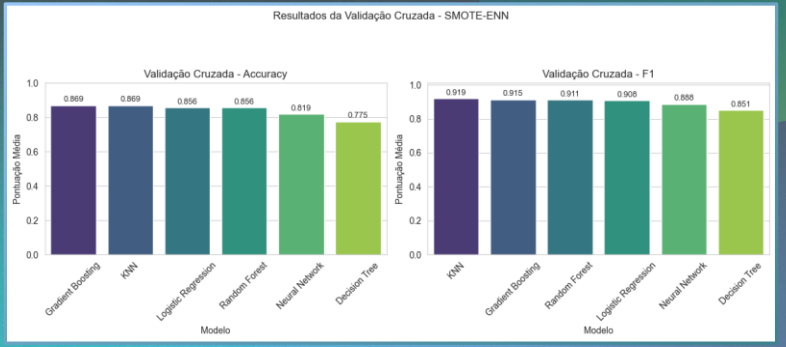




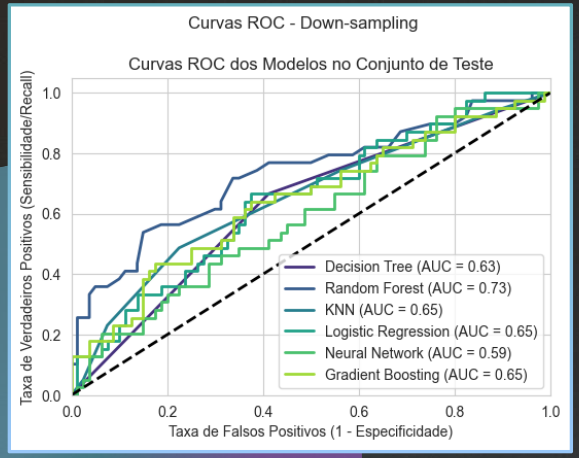


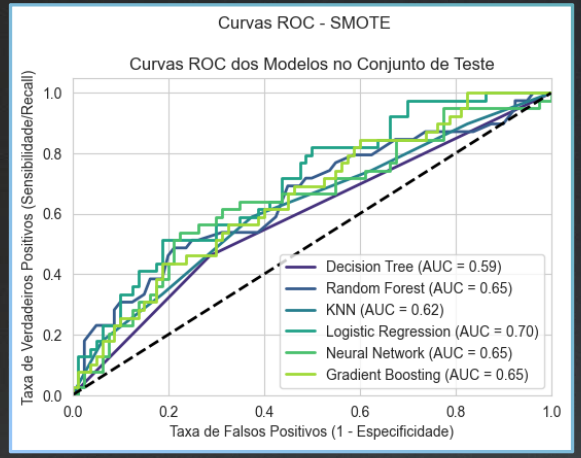


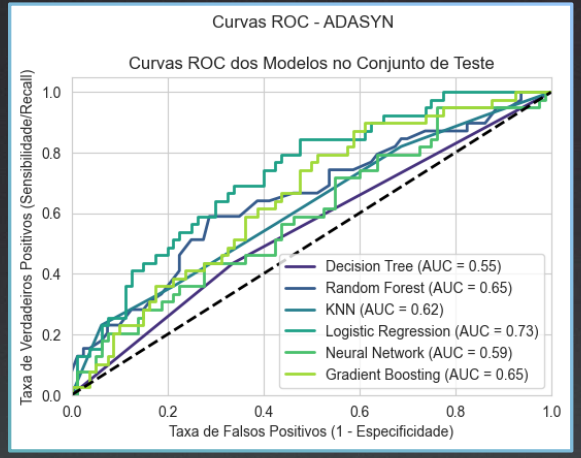


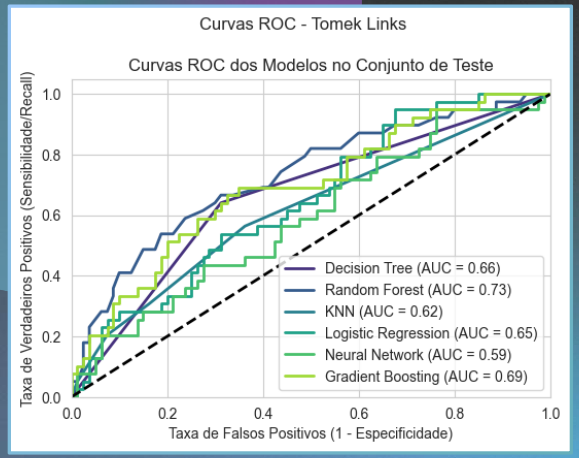


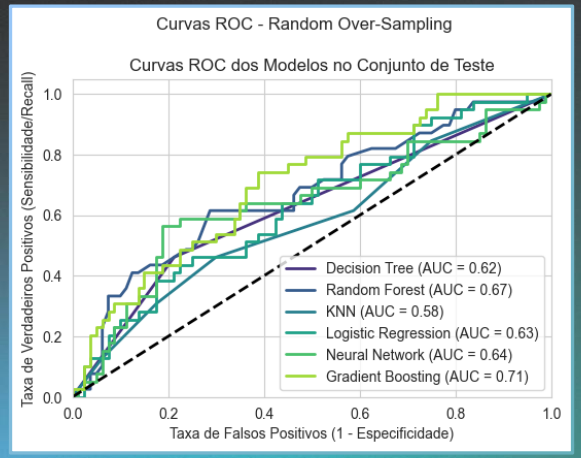
Curvas Roc

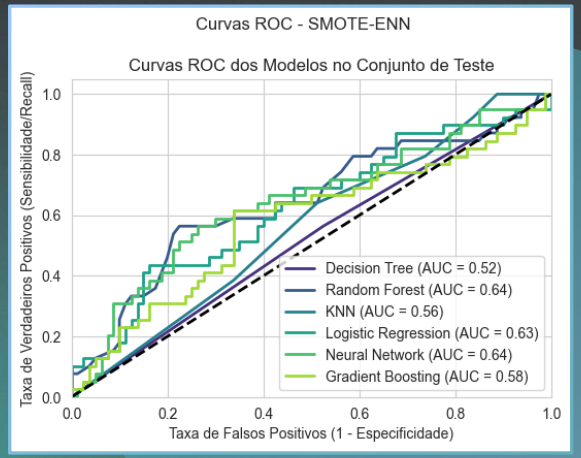

Curvas ROC - Down-sampling
Curvas ROC dos Modelos no Conjunto de Teste
Taxa de Verdadeiros Positivos (Sensibilidade/Recall)
Taxa de Falsos Positivos (1 - Especificidade)
Decision Tree (AUC = 0.63)
Random Forest (AUC = 0.73)
KNN (AUC = 0.65)
Logistic Regression (AUC = 0.65)
Neural Network (AUC = 0.59)
Gradient Boosting (AUC = 0.65)


Curvas ROC - SMOTE
Curvas ROC dos Modelos no Conjunto de Teste
Taxa de Verdadeiros Positivos (Sensibilidade/Recall)
Taxa de Falsos Positivos (1 - Especificidade)
Decision Tree (AUC = 0.59)
Random Forest (AUC = 0.65)
KNN (AUC = 0.62)
Logistic Regression (AUC = 0.70)
Neural Network (AUC = 0.65)
Gradient Boosting (AUC = 0.65)

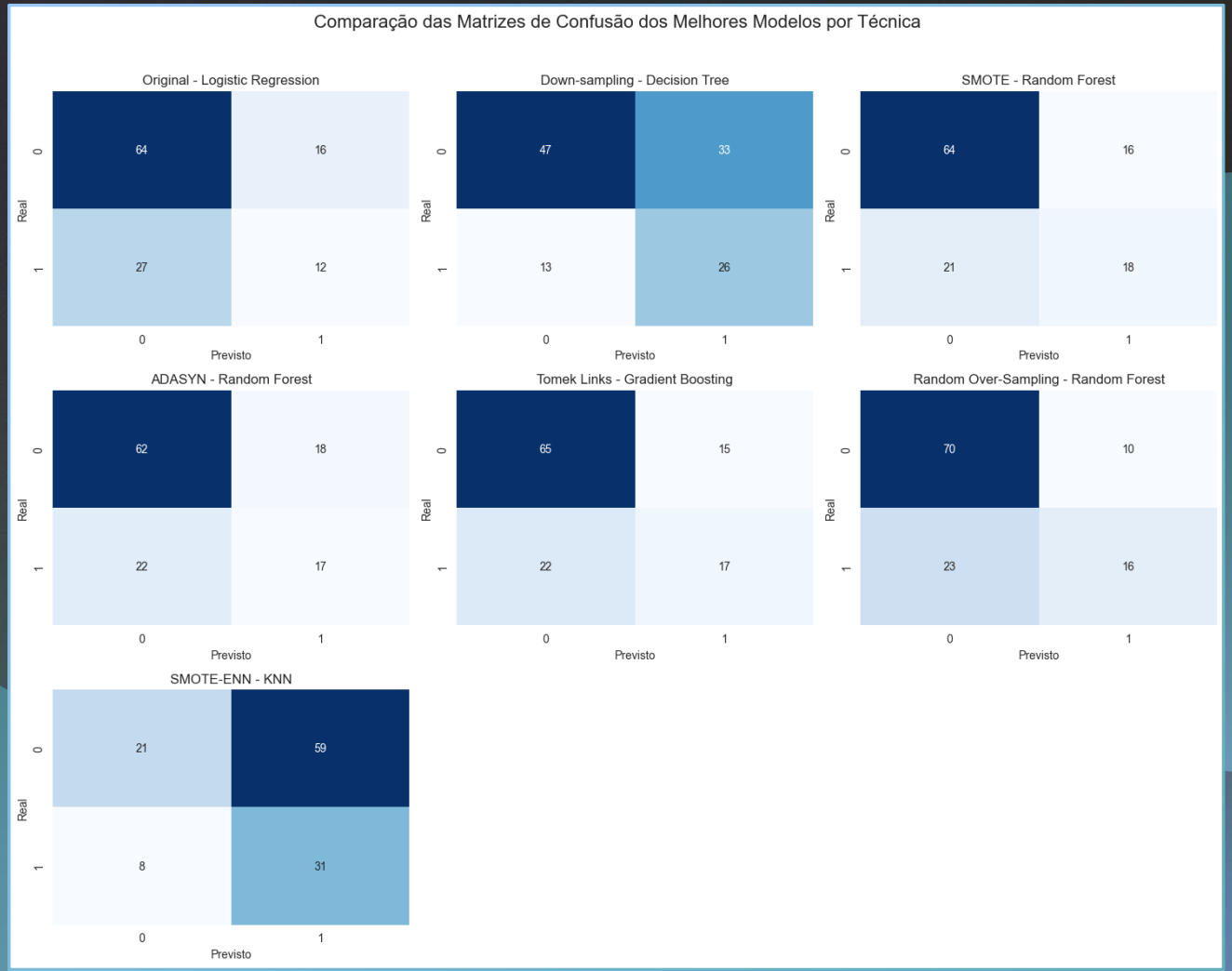

Curvas ROC - ADASYN
Curvas ROC dos Modelos no Conjunto de Teste
Taxa de Verdadeiros Positivos (Sensibilidade/Recall)
Taxa de Falsos Positivos (1 - Especificidade)
Decision Tree (AUC = 0.55)
Random Forest (AUC = 0.65)
KNN (AUC = 0.62)
Logistic Regression (AUC = 0.73)
Neural Network (AUC = 0.59)
Gradient Boosting (AUC = 0.65)


Curvas ROC - Tomek Links
Curvas ROC dos Modelos no Conjunto de Teste
Taxa de Verdadeiros Positivos (Sensibilidade/Recall)
Taxa de Falsos Positivos (1 - Especificidade)
Decision Tree (AUC = 0.66)
Random Forest (AUC = 0.73)
KNN (AUC = 0.62)
Logistic Regression (AUC = 0.65)
Neural Network (AUC = 0.59)
Gradient Boosting (AUC = 0.69)

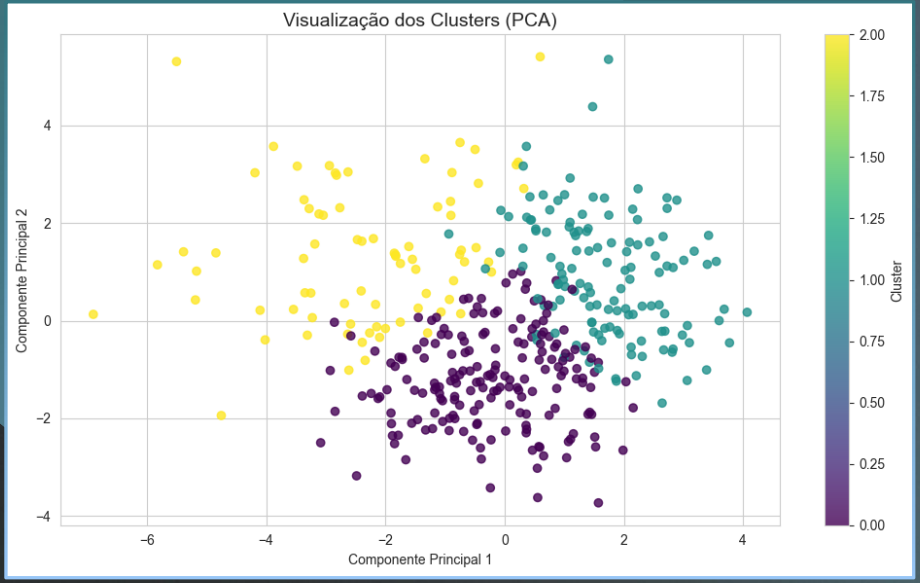

Curvas ROC - Random Over-Sampling
Curvas ROC dos Modelos no Conjunto de Teste
Taxa de Verdadeiros Positivos (Sensibilidade/Recall)
Taxa de Falsos Positivos (1 - Especificidade)
Decision Tree (AUC = 0.62)
Random Forest (AUC = 0.67)
KNN (AUC = 0.58)
Logistic Regression (AUC = 0.63)
Neural Network (AUC = 0.64)
Gradient Boosting (AUC = 0.71)


Curvas ROC - SMOTE-ENN
Curvas ROC dos Modelos no Conjunto de Teste
Taxa de Verdadeiros Positivos (Sensibilidade/Recall)
Taxa de Falsos Positivos (1 - Especificidade)
Decision Tree (AUC = 0.52)
Random Forest (AUC = 0.64)
KNN (AUC = 0.56)
Logistic Regression (AUC = 0.63)
Neural Network (AUC = 0.64)
Gradient Boosting (AUC = 0.58)

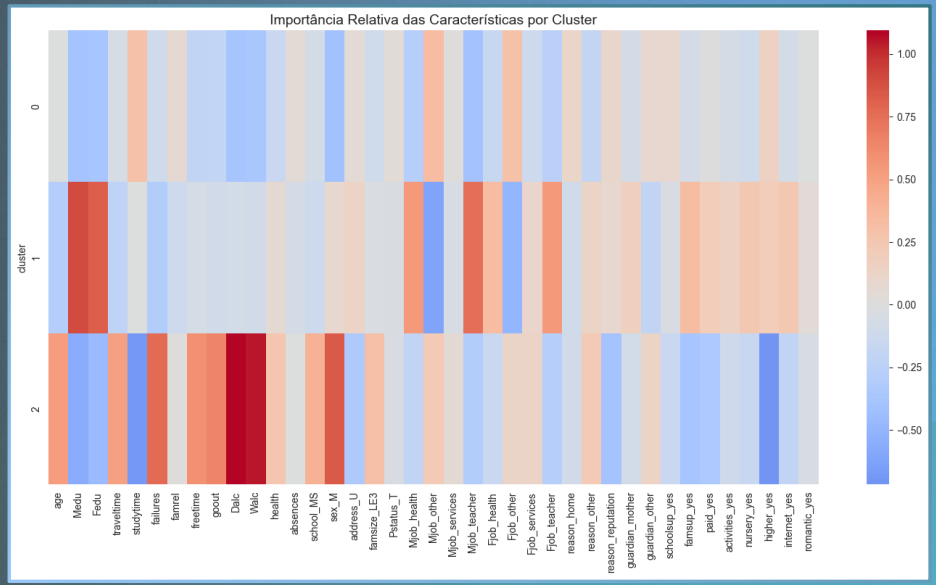
Matrizes de Confusão – Original

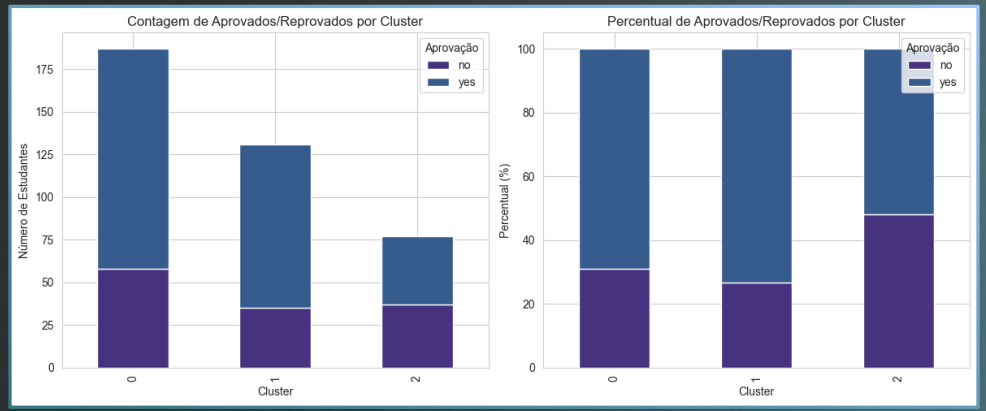


Clustering



Importância das características





Conclusão

- Recomendações:
  - Cluster 0 (Estudantes em Risco): Combinação de ações psicossociais (álcool + família); Suporte acadêmico direto (tutoria + workshops); Mecanismo proativo (alertas).
  - Cluster 1 (Estudantes de Alto Desempenho): Desenvolvimento acadêmico (enriquecimento curricular); Responsabilidade social (mentoria) + foco futuro (ensino superior).
  - Cluster 2 (Estudantes de Perfil Intermediário): Abordagem dupla: prevenção (intervenções + monitoramento) + personalização (aconselhamento + planos).

- Comparação final dos algoritmos: classe positiva (1) - reprovação ('no'); classe negativa (0) - aprovação ('yes').
- Melhor combinação: SMOTE-ENN e KNN (accuracy e F1-score).

Bibliografia: Notebooks e python cheat sheets da cadeira; deepseek ; ChatGPT .