Um estudo do desempenho acadêmico de estudantes e fatores relevantes

Abid Lohan, Nicolly Zorzam, Vitória Nazareth

Instituto de Computação – Universidade Federal do Rio de Janeiro (UFRJ)

{abidlsfs, nicollyzm, vitoriamna}@dcc.ufrj.br

Abstract. *This report describes the experiments done by the group as final work of the course of Introduction to Machine Learning. In this report, it is shown and specified a database with factors that can affect the performance of students and it is also documented two experiments done with this database: an analysis between study routine and final scores and also an attempt of students group classification considering their academic performance.*

Resumo. *Este relatório descreve experimentos realizados pelo grupo como trabalho final do curso de Introdução à Aprendizado de Máquina. Nele, é apresentado e especificado uma base de dados que contém fatores que podem afetar o desempenho dos estudantes e também é documentado os dois experimentos realizados com essa base: uma análise da relação entre estudo e nota final e também a tentativa de classificação de grupos de estudantes levando em conta seu desempenho.*

1. Introdução

1.2. Sobre a Base Escolhida

A base escolhida contém diversas informações que podem influenciar a performance de um estudante durante as avaliações. Os dados incluem hábitos de estudos, percentual de presença, envolvimento dos pais, notas de exames e outros aspectos.

Embora diversos parâmetros tenham sido mapeados, identificamos a ausência de informações essenciais que não foram fornecidas pelo proprietário da base. Faltam detalhes como o local de coleta dos dados (país/região), a faixa etária e o nível de escolaridade dos estudantes (se estão no ensino fundamental ou médio). Além disso, alguns atributos carecem de descrições mais claras. Por exemplo, não está especificado se o campo "Previous_Scores" se refere à última avaliação feita pelo estudante antes do exame final ou à média de todos os exames realizados ao longo de um período não delimitado. Outros atributos também são excessivamente subjetivos por serem categóricos. Apesar das limitações mencionadas, o grupo decidiu utilizar essa base de dados por se tratar de um tema de grande interesse (performance acadêmica). Além disso, acreditamos que, mesmo com os problemas relatados, é possível realizar análises relevantes e aplicar modelos de previsão e clusterização de maneira satisfatória. Em outras palavras, as questões apontadas não comprometem significativamente os resultados que gostaríamos de encontrar.

1.3. Objetivo

O objetivo inicial do grupo é utilizar a base de dados para medir, classificar e prever aspectos relacionados à performance acadêmica. As análises que pretendemos realizar são:

- Examinar a relação entre hábitos de estudo (horas dedicadas, frequência às aulas e monitorias) e a performance acadêmica;
- Classificar os estudantes, buscando separá-los por alto, médio e baixo desempenho/rendimento;

2. Descrição da Base e Pré-Processamento

2.1. Descrição da Base

Como já informado anteriormente, a base escolhida contém diversas informações que podem influenciar a performance de um estudante durante as avaliações. Abaixo, temos uma lista dos atributos e suas respectivas descrições:

- Hours_Studied: horas de estudo por semana.
- Attendance: porcentagem de presença nas aulas.
- Parental_Involvement: nível de envolvimento dos pais (Low, Medium, High).
- Access_to_Resources: disponibilidade de recursos educacionais (Low, Medium, High).
- Extracurricular_Activities: participação em atividades extracurriculares (Yes, No).
- Sleep_Hours: número médio de horas de sono por noite.
- Previous_Scores: notas de exames anteriores.
- Motivation_Level: nível de motivação do estudante (Low, Medium, High).
- Internet_Access: acesso à internet (Yes, No).
- Tutoring_Sessions: número de monitorias frequentadas por mês.
- Family_Income: renda da família (Low, Medium, High).
- Teacher_Quality: qualidade dos professores (Low, Medium, High).
- School_Type: tipo de escola (Public, Private).
- Peer_Influence: influência dos colegas na performance acadêmica (Positive, Neutral, Negative).
- Physical_Activity: média de horas de atividades físicas por semana.
- Learning_Disabilities: presença de deficiência que prejudica o aprendizado (Yes, No).
- Parental_Education_Level: maior grau de escolaridade dos pais (High School, College, Postgraduate).
- Distance_from_Home: distância de casa para a escola (Near, Moderate, Far).
- Gender: gênero do estudante (Male, Female).
- Exam_Score: nota no exame final.

2.2. Pré-Processamento

O pré-processamento da base foi organizado em quatro passos: limpeza de dados

e tratamento de outliers, integração de dados (não realizado, pois nossos dados são provenientes de uma única base), redução de dados e transformação de dados.

A limpeza de dados e tratamento de outliers consistiu, inicialmente, em identificar dados ausentes e tratá-los devidamente. Nesta etapa, identificamos que haviam 78, 90 e 67 dados ausentes nas colunas “Teacher_Quality”, “Parental_Education_Level” e “Distance_from_Home”, respectivamente. Uma opção seria remover essas linhas da nossa base, contudo, para evitar remover linhas que poderiam conter características significantes para os modelos posteriormente, o grupo optou por calcular e substituir pela moda. Como os dados eram categóricos, substituímos os dados ausentes dessas linhas pela categoria mais comum de cada coluna. A segunda etapa foi procurar por outliers nos dados numéricos, ou seja, vamos analisar as colunas das pontuações dos exames, das horas estudadas, da presença, das horas dormidas e também das visitas às monitorias. Considerando o contexto do nosso banco de dados, o grupo decidiu primeiro o que seria "normal" para cada variável analisada, por exemplo: não seria anormal um aluno tirar 0 em determinado exame, porém valores negativos ou acima do máximo (100) seriam estranhos. Isso pode significar um erro na hora de popular o banco e esses valores devem ser tratados/removidos, pois podem atrapalhar nossos modelos na hora de aprender o padrão dos dados. Durante essa análise, identificamos um aluno que supostamente teria pontuado 101 no exame final, o que não é “normal”, mas apenas configura um possível erro de digitação. Com isso, identificamos a linha em que se encontrava esse outlier e corrigimos a nota para 100. Uma última etapa na limpeza foi usar o “drop_duplicates()” para remover qualquer linha duplicada.

Como o passo de integração não foi necessário, passamos para o terceiro passo: a redução de dados. O grupo começou reduzindo o tamanho da matriz removendo as colunas "Teacher_Quality" e "Motivation_Level". O grupo escolheu remover essas colunas, pois acreditamos que são muito subjetivas e difíceis de quantificar. Além disso, não acreditamos que seriam necessárias para as análises que serão feitas posteriormente (apresentadas no Objetivo). Alteramos também os tipos dos dados representados na tabela. Existem dados que podem ser do tipo 'bool' (Yes/No) e dados categóricos (Medium, Low, High). Dados categóricos são textos e, para os algoritmos que queremos utilizar, não serão bem tratados. Com isso, mudamos os tipos dos dados para 'int' com um mapeamento manual. As colunas “Extracurricular_Activities”, “Internet_Access” e “Learning_Disabilities” que tem valores “Yes’' ou “No” foram convertidas para 0 ou 1. As colunas que possuem valores com uma relação de ordem entre si (“Medium-Low-High”, “Near-Moderate-Far” e “High School-College-Postgraduate”) receberam valores de 1 a 3. A coluna "Peer_Influence" com valores “Positive-Neutral-Negative” recebeu os valores 1, 0 e -1, respectivamente. Já para as colunas de “Gender” e “School_Type” que não possuem essa relação, usamos o método “get_dummies()” para criar novas colunas com as respectivas classificações.

No último passo, transformação de dados, normalizamos e padronizamos a nossa base com os métodos “MinMaxScaler” e “StandardScaler”, respectivamente. Dessa forma separamos 3 bases (com os dados brutos, com os dados normalizados e com os dados padronizados), para podermos usar dependendo dos modelos escolhidos mais adiante.

3. Definição dos Métodos Usados e Justificativa

Os métodos utilizados nos experimentos apresentados neste relatório são:

- Previsão:
    - Regressão Linear
        - MSE
        - R^2

Como dispúnhamos de uma base de dados que relaciona diversos atributos com a variável final ("Exam_Score"), decidimos utilizar a regressão linear para verificar a eficácia do modelo treinado com esses atributos na previsão das notas finais.

- Classificação:
    - Random Forest
        - Acurácia
        - Recall
        - K-Fold Estratificado

Para a classificação, optamos por usar o Random Forest, uma vez que ele é um ensemble de árvores de decisão que trabalha de forma a reduzir a variabilidade do modelo, aumentando a precisão e a generalização. Além disso, ele é capaz de lidar bem com dados desbalanceados, o que é relevante em nosso contexto em que a distribuição das classes não é uniforme.

4. Trabalhos Relacionados

Diversos estudos se desdobraram sobre os fatores que influenciam o desempenho acadêmico de estudantes em geral, utilizando diferentes metodologias e focando em variáveis diversas.

Gómez-Sánchez et al. (2011) realizaram uma pesquisa com o objetivo de determinar o desempenho acadêmico percebido dos estudantes universitários e sua relação com variáveis como sexo, curso, semestre, média de notas e satisfação com a carreira escolhida. O conjunto utilizado é composto por 26 itens, incluindo dados sociodemográficos e escalas para medir satisfação e desempenho percebido. As análises estatísticas aplicadas foram correlação de Pearson, Rho de Spearman, Teste T para amostras independentes e análise de variância unidirecional. Os resultados mostraram que o desempenho acadêmico percebido está relacionado ao semestre cursado, ao sexo do aluno, à média de notas e à satisfação com a carreira escolhida.

Gong, Beck e Heffernan (2011) compararam dois modelos para prever o desempenho dos alunos: rastreamento de conhecimento (KT) e análise de fatores de desempenho (PFA). O estudo avaliou a precisão preditiva de cada modelo em relação ao desempenho dos alunos durante oportunidades de prática individual. Os autores exploraram diferentes decisões para cada abordagem e identificaram um conjunto de “melhores práticas” para cada modelo. Os resultados mostraram que o PFA apresentou maior qualidade preditiva que o KT.

No contexto da educação médica, o estudo de Andrade et al. (2020) analisou a correlação entre notas de testes de progresso, testes de aptidão e atitude, desempenho

em estágio e notas em exames de residência médica. A pesquisa utilizou correlação de Pearson e análise de regressão linear para identificar fatores associados ao desempenho em exames de residência. Os resultados mostraram que as pontuações nos testes de progresso e nos testes de atitude e atitude tiveram correlações significativas entre si. Além disso, o desempenho nos rodízios e nos testes de atitude e postura associou-se positivamente às notas dos exames de residência médica. Esses estudos destacam a importância de diversas variáveis, como satisfação no trabalho, métodos de previsão de desempenho e avaliações contínuas, para compreender os fatores que influenciam o desempenho acadêmico dos alunos.

Nosso trabalho se insere nesse contexto, na direção de analisar com ainda mais objetividade a correlação entre diferentes variáveis que possam afetar o desempenho do estudante, e tentar aferir uma nota final baseada apenas em tais parâmetros.

5. Experimentos Realizados

5.1. Treinar modelo para previsão das notas;

Este experimento tem como objetivo treinar um modelo para prever as notas finais dos alunos ("Exam_Score"). Para isso, utilizamos um modelo de Regressão Linear, que foi alimentado com todos os atributos disponíveis. O treinamento foi realizado tanto com os dados normalizados quanto com os dados padronizados, a fim de avaliar o impacto de cada abordagem na precisão das previsões.

O modelo treinado com os dados normalizados demonstrou uma alta precisão, com o erro médio quadrado em torno de 0,0023. Em contraste, o modelo treinado com os dados padronizados apresentou uma precisão inferior, com erro médio quadrado de 0,2713.

Quanto ao coeficiente de determinação, que mede a variabilidade explicada pelo modelo, o valor obtido para os dados normalizados foi de 0,6946, indicando que cerca de 69% da variação nos dados de saída é explicada pelo modelo. Para os dados padronizados, o coeficiente foi ligeiramente superior, em torno de 0,73, sugerindo uma explicação um pouco maior da variabilidade dos dados.

Além disso, o modelo treinado com os dados padronizados gerou previsões fora do intervalo esperado, incluindo notas negativas e valores superiores a 100.

5.2. Examinar a relação entre hábitos de estudo (horas dedicadas, frequência às aulas e monitorias) e a performance acadêmica;

Este experimento tem como objetivo entender a relação entre entre os hábitos de estudo e a performance acadêmica final (levando em consideração apenas o “Exam_Score”) dos alunos. Para isso, utilizamos um modelo de Regressão Linear para prever as notas finais com base nos seguintes atributos:

- Hours_Studied: horas de estudo por semana.
- Attendance: porcentagem de presença nas aulas.

- Tutoring_Sessions: número de monitorias frequentadas por mês.

Esses atributos foram escolhidos para este experimento, pois acreditamos que são os atributos mais relacionados com a ideia de "hábitos de estudo". Utilizamos os dados no formato bruto para facilitar a compreensão das notas na visualização do gráfico. Com o modelo treinado, o passo seguinte foi exibir um gráfico de dispersão entre os valores reais e os valores previstos pelo modelo a fim de avaliar a relação entre esses valores. Além do gráfico, também realizamos as medidas de MSE (5.66) e R^2 (0.61).

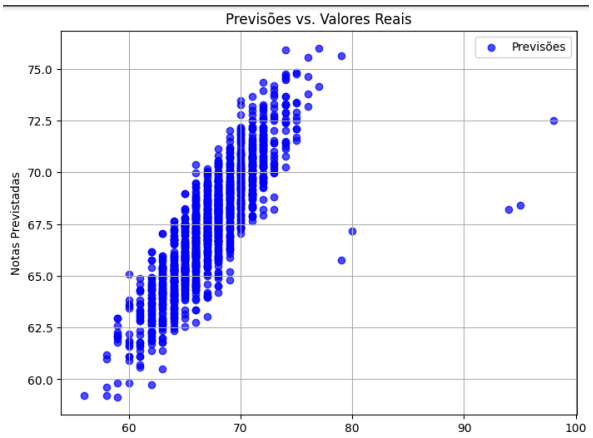

Figura 1. Gráfico de dispersão

O gráfico nos mostra o esboço de uma reta, onde os pontos plotados estão relativamente agrupados nesse padrão com poucos pontos realmente distantes. Isso nos permite concluir uma correlação minimamente significativa entre os valores previstos pelo modelo e os valores reais. O erro quadrático médio (MSE) de 5.66 nos diz que, em média, nosso modelo está errando a pontuação real em 5.66 pontos e em relação ao valor de R^2, podemos concluir que nossos atributos conseguem explicar 61% da variância dos dados. Apesar de poder parecer um modelo razoável à primeira vista, temos que levar em consideração que, com apenas 3 atributos, nosso modelo foi capaz de explicar a maior parte dos dados que desejávamos prever.

5.3. Classificação dos estudantes: Classificar estudantes de alto, médio e baixo rendimento.

O objetivo deste experimento era classificar os estudantes em grupos já definidos (aprendizado supervisionado). Inicialmente definimos os grupos da seguinte forma:

- Classe A: alunos de alto desempenho, instâncias com M >= 80
- Classe B: alunos de médio desempenho, instâncias com M >= 60 e < 80
- Classe C: alunos de baixo desempenho, instâncias com M < 60

Onde M é a média aritmética entre os valores de *'Previous_Scores'* e *'Exam_Score'*. Nosso intuito era treinar um modelo para, a partir de features de entrada, conseguir classificar as instâncias, identificando se o estudante era de alto, médio ou baixo desempenho/rendimento. Incluímos as colunas *'Mean_Scores'* com as médias calculadas (retirando as duas colunas anteriores utilizadas para o cálculo da média) e *'Class'* que atribuía o valor correto para cada instância. Separamos então a nova versão

do banco construído (df_Means) em duas partes: df_X que continha todas as features independentes e df_y contendo a feature dependente (coluna 'Class'). Escolhemos usar o método *RandomForestClassifier*.

Ao dividirmos os grupos para teste e treino notamos que havia um grande desbalanceamento entre as classes, onde a classe B era aproximadamente 75% de todo o banco de dados. Considerando isso, utilizamos o parâmetro de treino 'stratify' na divisão dos conjuntos para garantir a proporção e impedir qualquer tipo de viés no nosso modelo. Após a previsão do modelo, realizamos alguns cálculos para avaliá-lo. A primeira métrica que utilizamos foi a acurácia que chegou a 76%, o que pode ser considerado um bom desempenho no nosso cenário. Contudo, levando em consideração o grande desbalanceamento das classes, chegamos a conclusão de que o cálculo da acurácia não era a melhor opção, visto que se o modelo chutasse apenas a classe B, ele já teria algo próximo desse desempenho. Assim, plotamos a matriz de confusão do nosso modelo.

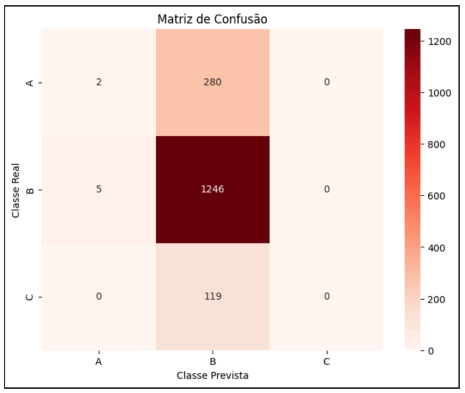

Figura 2. Matriz de confusão - classificação 1

Pela matriz acima, nosso modelo chuta a maioria das instâncias no grupo B, o que não gostaríamos que estivesse acontecendo. Como tentativa de contornar o problema, realizamos o mesmo experimento utilizando o K-fold Stratified com 5 splits para tentar observar alguma diferença em relação ao parâmetro utilizado na tentativa anterior.

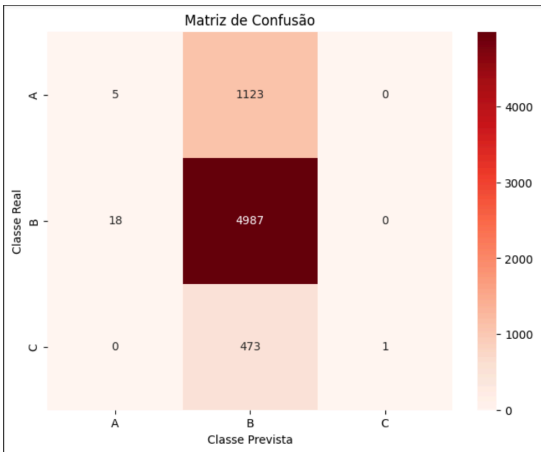

Figura 4. Matriz de confusão - classificação 2

Contudo, como exibido na figura acima, o problema persistiu. O modelo continuou chutando a maioria das instâncias na maior classe.

6. Discussão dos Resultados

O primeiro experimento gerou um modelo altamente preciso utilizando os dados normalizados, enquanto o modelo treinado com os dados padronizados apresentou maior coeficiente de determinação, mas foi menos preciso e gerou notas fora do intervalo esperado.

Em relação ao segundo experimento (subseção 5.2), o gráfico de dispersão revelou uma forte correlação entre os features de entrada e o de saída. Para confirmar isso, plotamos o heatmap de correlação exibido na figura abaixo.

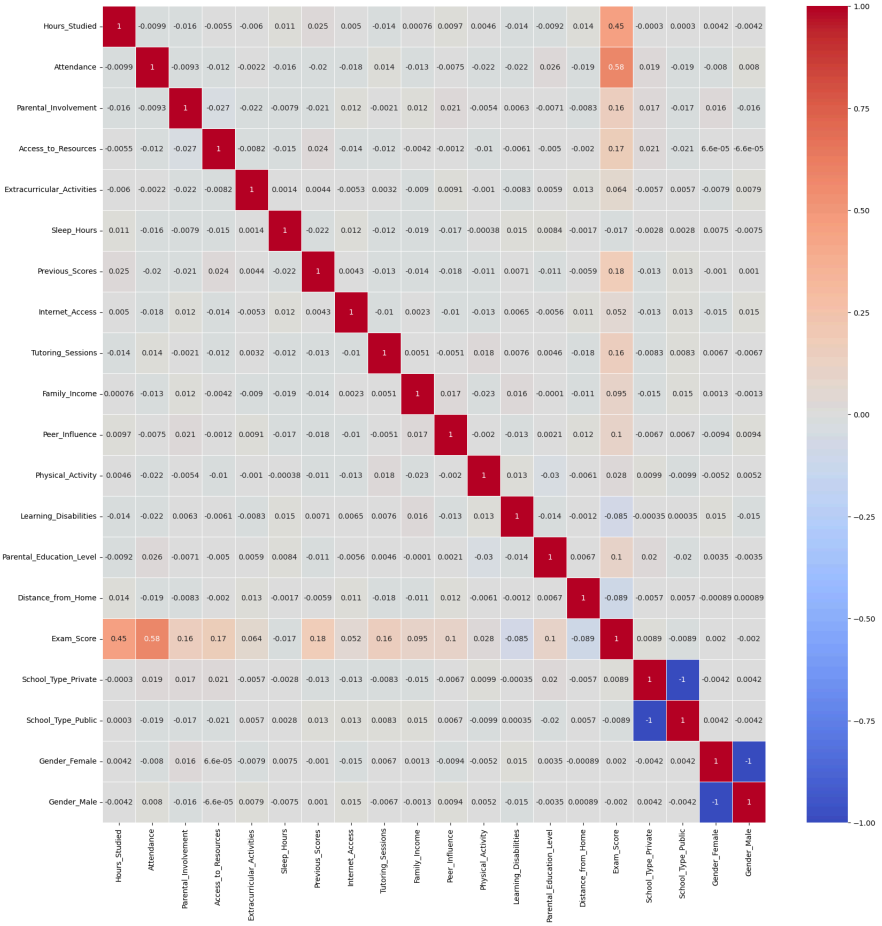

Figura 4. Heatmap

Os features utilizados na entrada são features que, de acordo com o heatmap, tem mais correlação com nossa saída (0.45 do "Hours_Studied", 0.58 do "Attendance" e 0.16 do "Tutoring_Sessions"), o que explica o motivo do nosso modelo conseguir capturar o padrão para a maioria dos dados. Considerando nossas métricas, como as notas variam de 0 a 100, um erro médio de 5.66 é consideravelmente pequeno. Se apenas com essas features, o modelo obteve um desempenho razoável, acreditamos que isso poderia ser melhorado adicionando mais variáveis, como por exemplo

“Previous_Scores” que tem correlação 0.18.

Com o nosso terceiro experimento (seção 5.3), não conseguimos gerar um modelo minimamente efetivo mesmo utilizando técnicas para tratar o desbalanceamento de dados. O experimento evidenciou desafios no treinamento do modelo para classificar os estudantes em classes de desempenho (A, B e C). Apesar de uma acurácia de 76%, esse resultado é enviesado devido ao desbalanceamento das classes, com 75% dos dados pertencendo à Classe B. A matriz de confusão mostrou que o modelo prioriza a classificação na maior classe, prejudicando a identificação correta das classes minoritárias. Mesmo com o uso de estratégias como o parâmetro stratify e o K-fold Stratified, o problema de enviesamento permaneceu, sugerindo a necessidade de abordagens adicionais para lidar com o desbalanceamento.

Como concluímos na seção 5.3, a acurácia não era a melhor métrica a ser utilizada. Para definir isso, levamos em consideração um cenário em que alunos classificados como baixo desempenho receberiam mais atenção/auxílio. Aqueles da classe C classificados erroneamente (falsos negativos) pelo modelo poderiam ser prejudicados. Nesse contexto, optamos por utilizar a métrica de recall (onde os falsos negativos são mais prejudiciais) para avaliar nosso modelo. Para a classe C o modelo não conseguiu identificar nenhuma instância real (recall = 0) o que significa que, apesar da acurácia de 76%, nosso modelo está muito longe do efetivo.

Os resultados do nosso trabalho reforçam em certa parte os achados de Gómez-Sánchez et al. (2011), que identificaram associações significativas entre variáveis acadêmicas, como média de notas e satisfação com a carreira, com o desempenho percebido dos estudantes. Entretanto, nosso estudo possui o diferencial de ter realizado uma predição quantitativa (para a nota dos alunos) baseando-se em múltiplos fatores como horas estudadas e frequência. Adicionalmente, diferente do modelo de análise dos fatores de performance (PFA) avaliado por Gong, Beck e Heffernan (2011), que apresentou alta acurácia em predições individuais, nosso modelo enfrentou desafios em relação à classificação desbalanceada de acordo com as classes escolhidas.

7. Conclusão

Podemos concluir que os hábitos de estudos analisados pelo grupo realmente possuem um impacto significativo nas notas finais dos estudantes, visto o desempenho do nosso modelo linear, destacando a relevância das horas de estudo e, principalmente, a frequência às aulas para o desempenho acadêmico. Mesmo sendo um modelo relativamente simples, ele apresenta resultados relevantes, sugerindo que os hábitos de estudo analisados têm influência significativa no sucesso acadêmico, embora existam outros fatores que podem ser explorados para aprimorar a previsão.

Em relação ao segundo experimento, acreditamos que seria possível construir um modelo minimamente razoável para classificação dos alunos (considerando o contexto da seção anterior) utilizando de outras técnicas que não foram experimentadas pelo grupo, como por exemplo ajuste de pesos das classes, o uso de métodos de balanceamento no pré-processamento, outro modelo (usando densidade dos dados ao invés de distância) e até mesmo classificar 2 ao invés de 3 grupos usando Regressão Logística para uma classificação binária. Embora o modelo tenha alcançado uma

acurácia inicial considerável, a predominância de previsões na Classe B compromete sua eficácia, como evidenciado no cálculo do recall. O experimento ressalta a importância de técnicas específicas para tratar dados desbalanceados. Vale pontuar que o grupo não considera as sugestões acima necessariamente compatíveis com o cenário do problema, mas sim ações que poderiam ser tomadas para experimentar mais com os dados.

8. Referências:

1. Códigos de terceiros: ChatGPT, código do treino do modelo RandomForest

```
seed = 10

X_training, X_test, y_training, y_test = train_test_split(df_X,
df_y, random_state=seed, test_size=0.25, stratify=df_y)

modelo_forest = RandomForestClassifier(random_state=42)

modelo_forest.fit(X_training, y_training)

y_pred = modelo_forest.predict(X_test)

acuracia = accuracy_score(y_test, y_pred)

print(f"Acurácia do modelo: {acuracia:.2f}")
```

2. Notebook desenvolvido pelo grupo:

   https://colab.research.google.com/drive/1CdrNss73uauxsOXEjZ8NIyvcaSWmglD8?usp=sharing

3. Gómez-Sánchez et al. (2011) "Factores que influyen en el rendimiento académico del estudiante universitario", TECNOCIENCIA Chihuahua, 5(2), 90–97.

4. Gong, Beck e Heffernan (2011) "How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factor Analysis", International Journal of Artificial Intelligence in Education, vol. 21, no. 1-2, pp. 27-46.

5. Andrade et al. (2020) "Factors associated with student performance on the medical residency test", Revista Da Associação Médica Brasileira, 66(10), 1376–1382.