Unversidade Federal do Maranhão

Mineração de Dados aplicada a dados educacionais de perfomance.

Aluno:
Rafael Freitas de Paula
rafael.fp@discente.ufma.br

Professor:
Thales Levi Azevedo Valente
thales.levi@ufma.br

SUMÁRIO

INTRODUÇÃO

Como melhorar o desempenho acadêmico entre os alunos?

Quais variáveis impactam mais ?

O nível de escolaridade dos pais impacta nas notas?

O gênero impacta nas notas?

Podemos usar modelo de machine learning para prever nota dos alunos?

*Essas são algumas perguntas que queremos responder com esse *dataset*.

OBJETIVO GERAL

O objetivo geral desse trabalho é **explorar** os fatores que influenciam o desempenho acadêmico dos alunos usando um conjunto de dados que inclui informações demográficas, socioeconômicas e educacionais.

OBJETIVOS ESPECÍFICOS

01 **Identificar**
padrões de desempenho acadêmico

02 **Descobrir**
relações entre fatores socioeconômicos e rendimento escolar

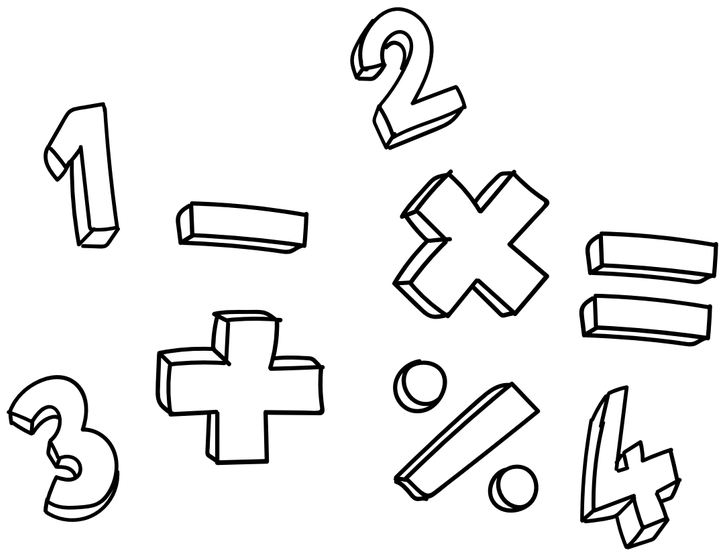
03 **Aplicar**
técnicas de mineração de dados para segmentar alunos com características semelhantes

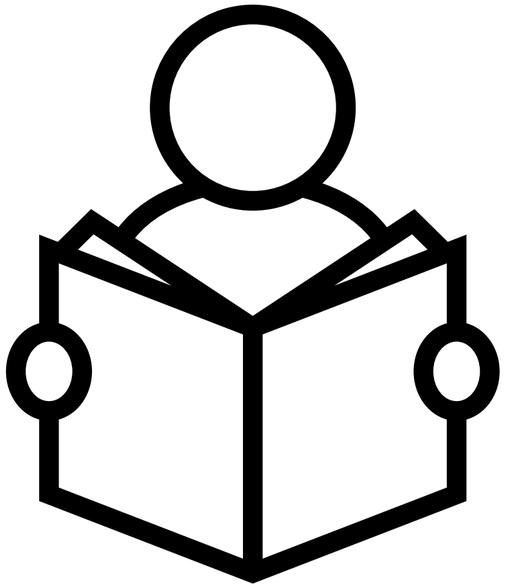
04 **Desenvolver**
modelos preditivos para estimar o desempenho dos alunos

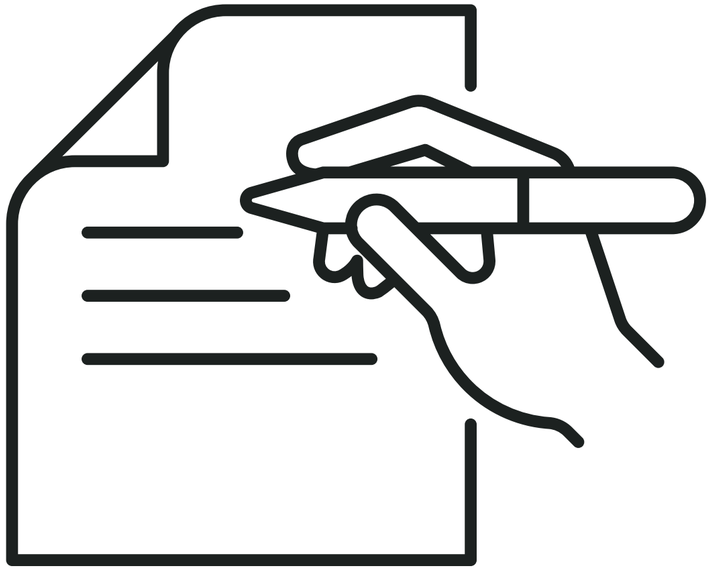
MATERIAIS E MÉTODOS

Base de Dados Utilizada

- O *dataset* "*StudentsPerformance.csv*" contém informações sobre o desempenho acadêmico de estudantes em três disciplinas: **matemática, leitura e escrita**.







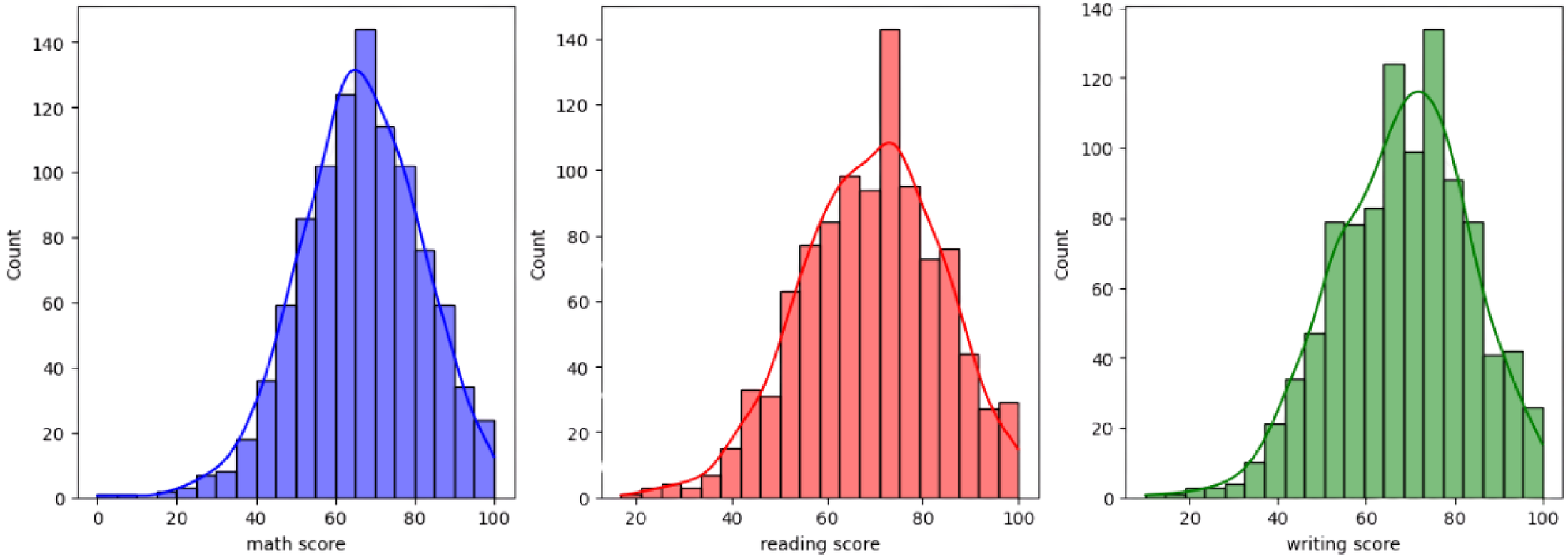
MATERIAIS E MÉTODOS

Estrutura da Base de Dados

- **Gender (gênero)**: Indica o gênero do estudante, podendo ser "male" (masculino) ou "female" (feminino).
- **Race/ethnicity (raça/etnia)**: Classificação do aluno em grupos raciais ou étnicos, representados por categorias como "group A", "group B", etc.
- **Parental level of education (nível de educação dos pais)**: Nível educacional mais alto alcançado pelos pais ou responsáveis pelo aluno, como "high school" (ensino médio), "associate's degree" (tecnólogo), "bachelor's degree" (graduação), entre outros.
- **Lunch (almoço)**: Tipo de almoço recebido pelo aluno, podendo ser "standard" (padrão) ou "free/reduced" (gratuito ou com preço reduzido).
- **Test preparation course (curso preparatório para o teste)**: Indica se o aluno completou um curso preparatório antes dos testes, com valores "completed" (completou) ou "none" (nenhum).
- **Math score (nota em matemática)**: Pontuação obtida pelo aluno na prova de matemática, variando de 0 a 100.
- **Reading score (nota em leitura)**: Pontuação obtida pelo aluno na prova de leitura, variando de 0 a 100.
- **Writing score (nota em escrita)**: Pontuação obtida pelo aluno na prova de escrita, variando de 0 a 100.

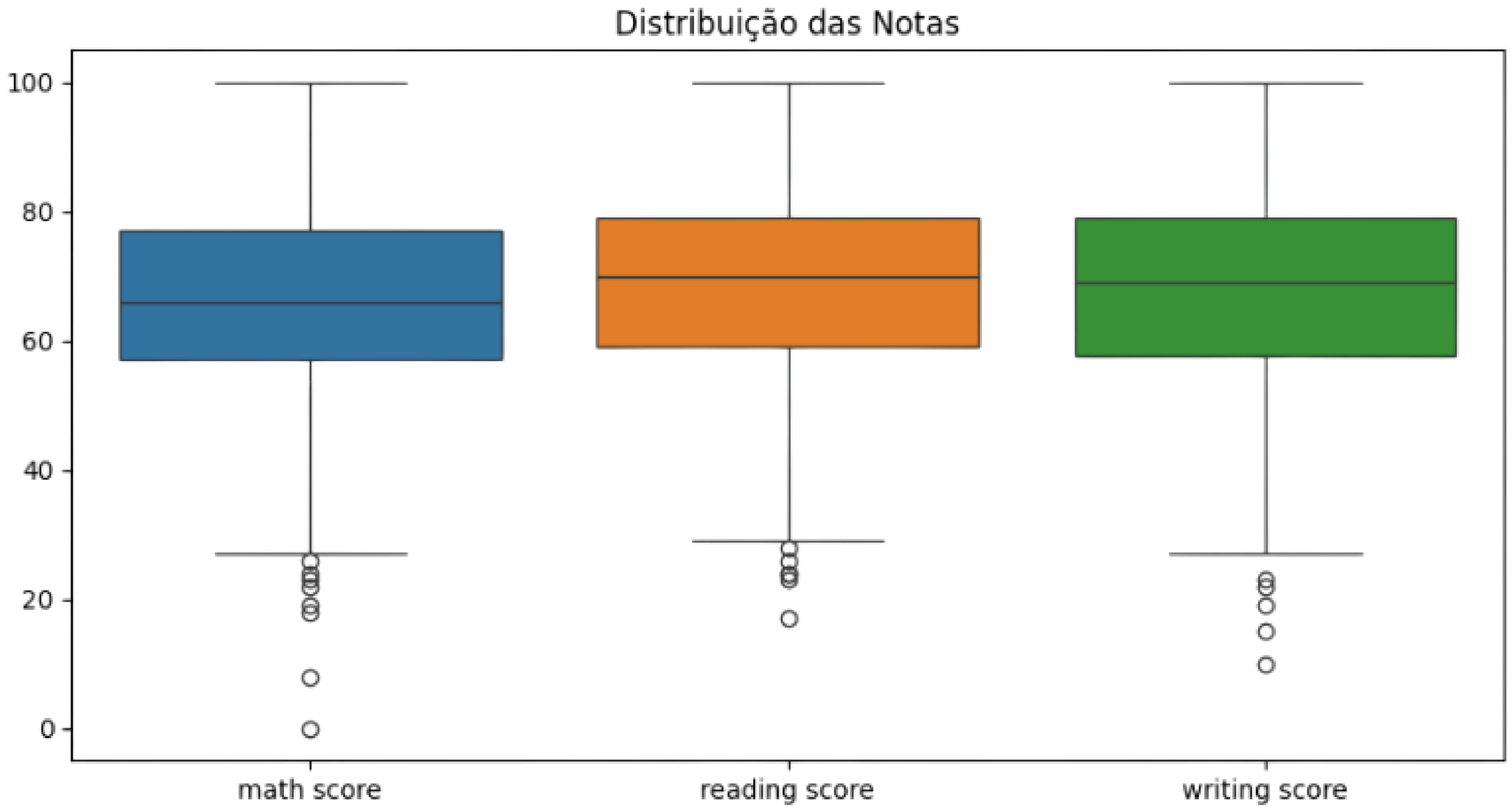
MATERIAIS E MÉTODOS

Análise Exploratória



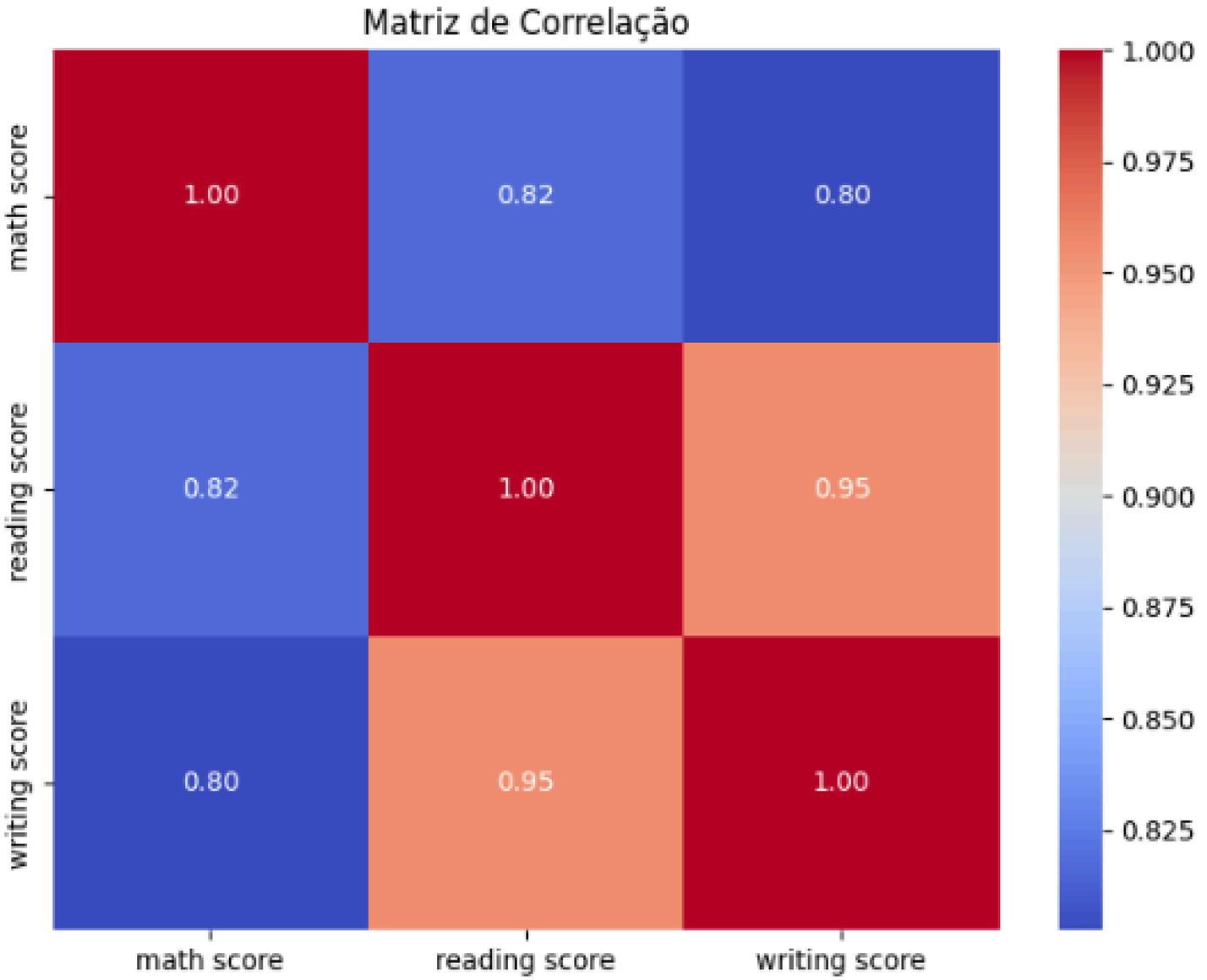
OBSERVAÇÕES

Essa plotagem permite visualizar a distribuição das notas dos alunos. Identifica se os dados seguem uma distribuição normal ou se há viés. Ajuda a detectar outliers (valores extremos)


Distribuição das Notas
100
80
60
40
20
0
math score
reading score
writing score

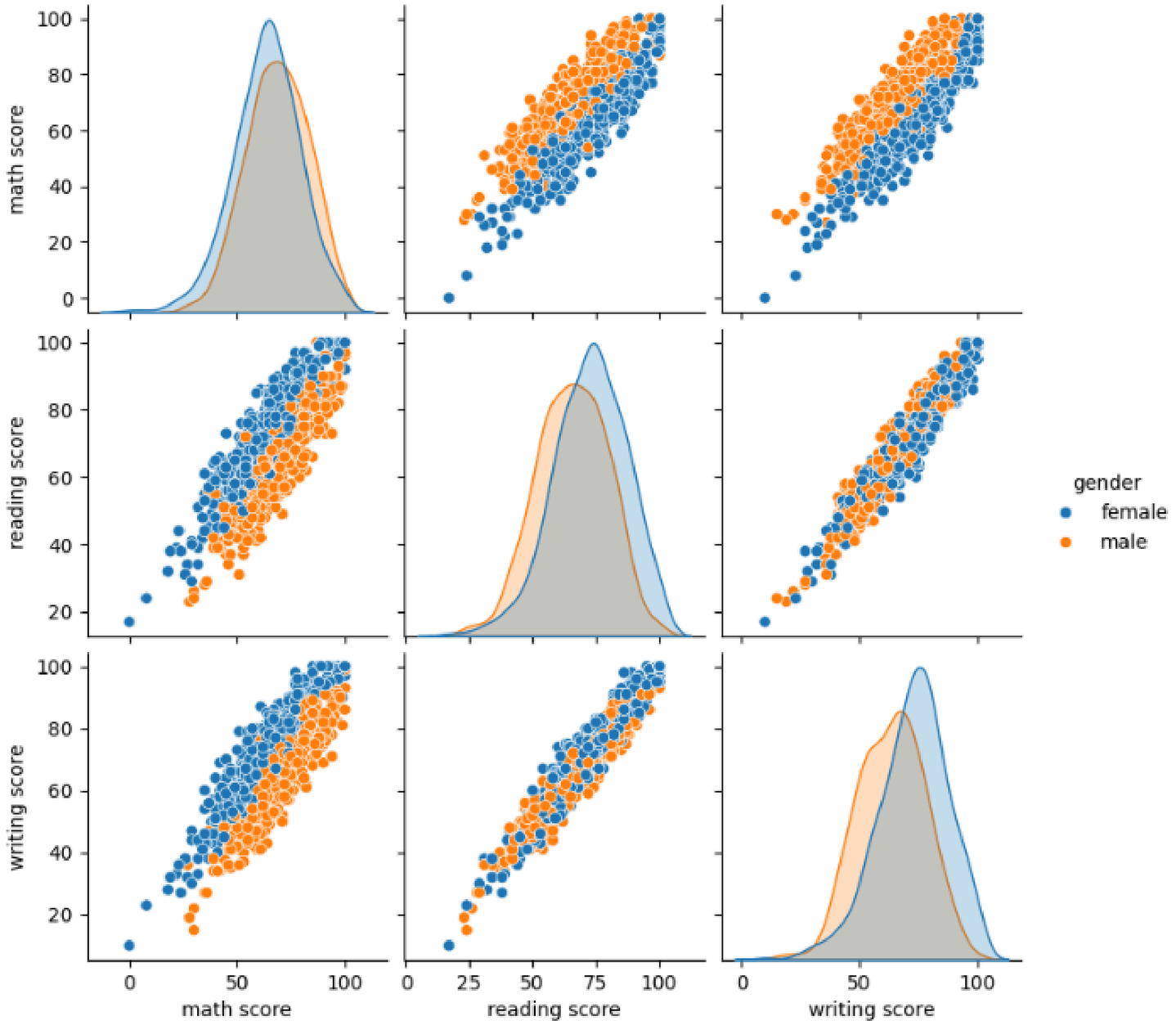
OBSERVAÇÕES

O boxplot mostra a mediana, quartis e possíveis outliers. Ajuda a comparar a distribuição das notas entre diferentes matérias.


Matriz de Correlação
math score
reading score
writing score
1.00
0.82
0.80
0.82
1.00
0.95
0.80
0.95
1.00
math score
reading score
writing score
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825

OBSERVAÇÕES

Foi analisado se há relações entre variáveis. Para ajudar a entender quais notas estão mais relacionadas entre si. Pode indicar colinearidade (se duas variáveis são muito correlacionadas, pode ser necessário remover uma para evitar redundância no modelo).


math score
reading score
writing score
gender
female
male
0
20
40
60
80
100
0
25
50
75
100

OBSERVAÇÕES

Assim como no plot anterior esse gráfico mostra a relação entre as diferentes notas.
Permite identificar padrões entre gêneros (por exemplo, se um gênero tende a ter notas mais altas em uma matéria específica).
Ajuda a visualizar se há correlações lineares entre as variáveis.

MATERIAIS E MÉTODOS

Integração de bases e limpeza

Essa etapa é essencial para garantir análises e modelos precisos.

Permite identificar colunas com dados ausentes.

Se houver muitas informações ausentes, pode ser necessário tratar os dados (preenchendo ou removendo valores).

Dados duplicados podem distorcer análises e modelos preditivos.

Mantém a qualidade e confiabilidade do dataset. Por isso foi decidido remover dados duplicados

MATERIAIS E MÉTODOS

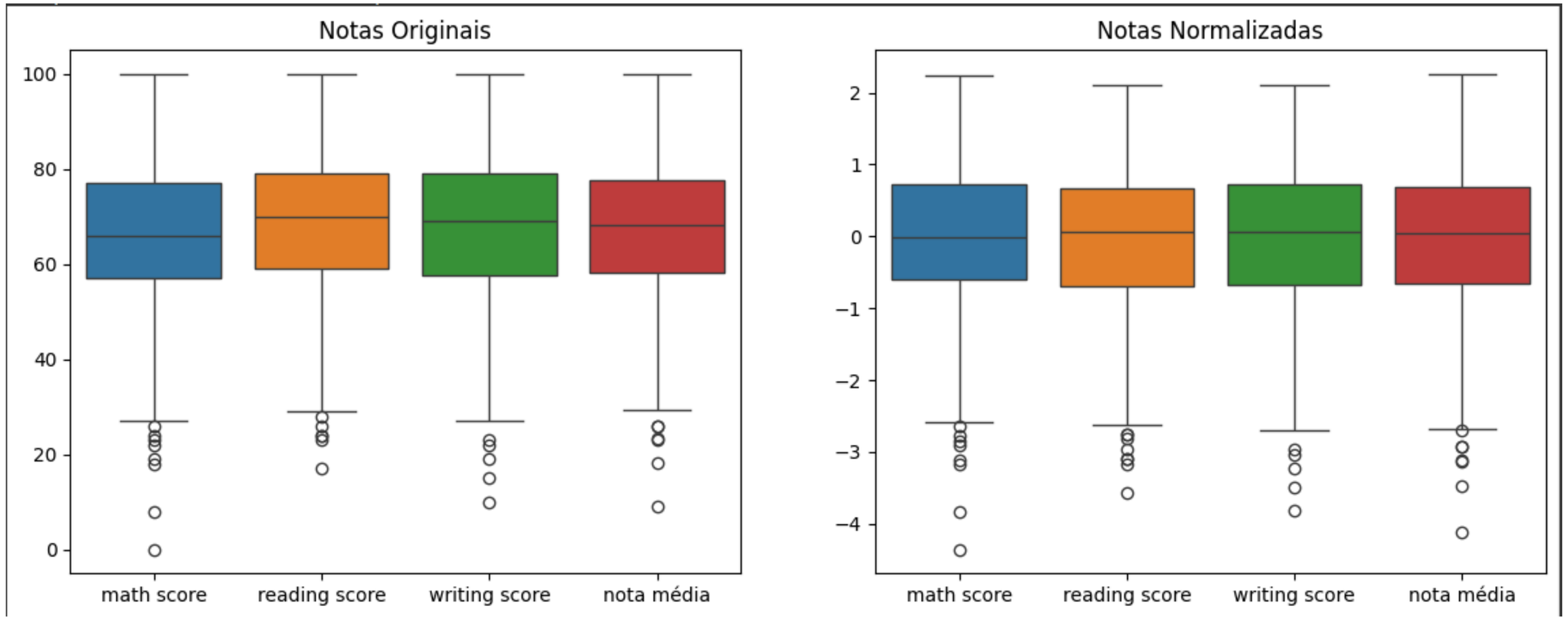
Normalização e Engenharia de características

Nessa etapa o objetivo é realizar algumas etapas essenciais da ciência de dados:

1. Criação de uma nota média para cada aluno com base em suas três notas.
2. Codificação de variáveis categóricas (como gênero e nível educacional dos pais) usando Label Encoding para facilitar o uso desses dados em modelos de aprendizado de máquina.
3. Normalização das notas utilizando ***StandardScaler***, garantindo que todas tenham uma escala semelhante, o que melhora a performance de certos modelos estatísticos e de aprendizado de máquina.
4. Visualização dos dados antes e depois da normalização através de boxplots e histogramas, permitindo a comparação entre a distribuição original e normalizada.

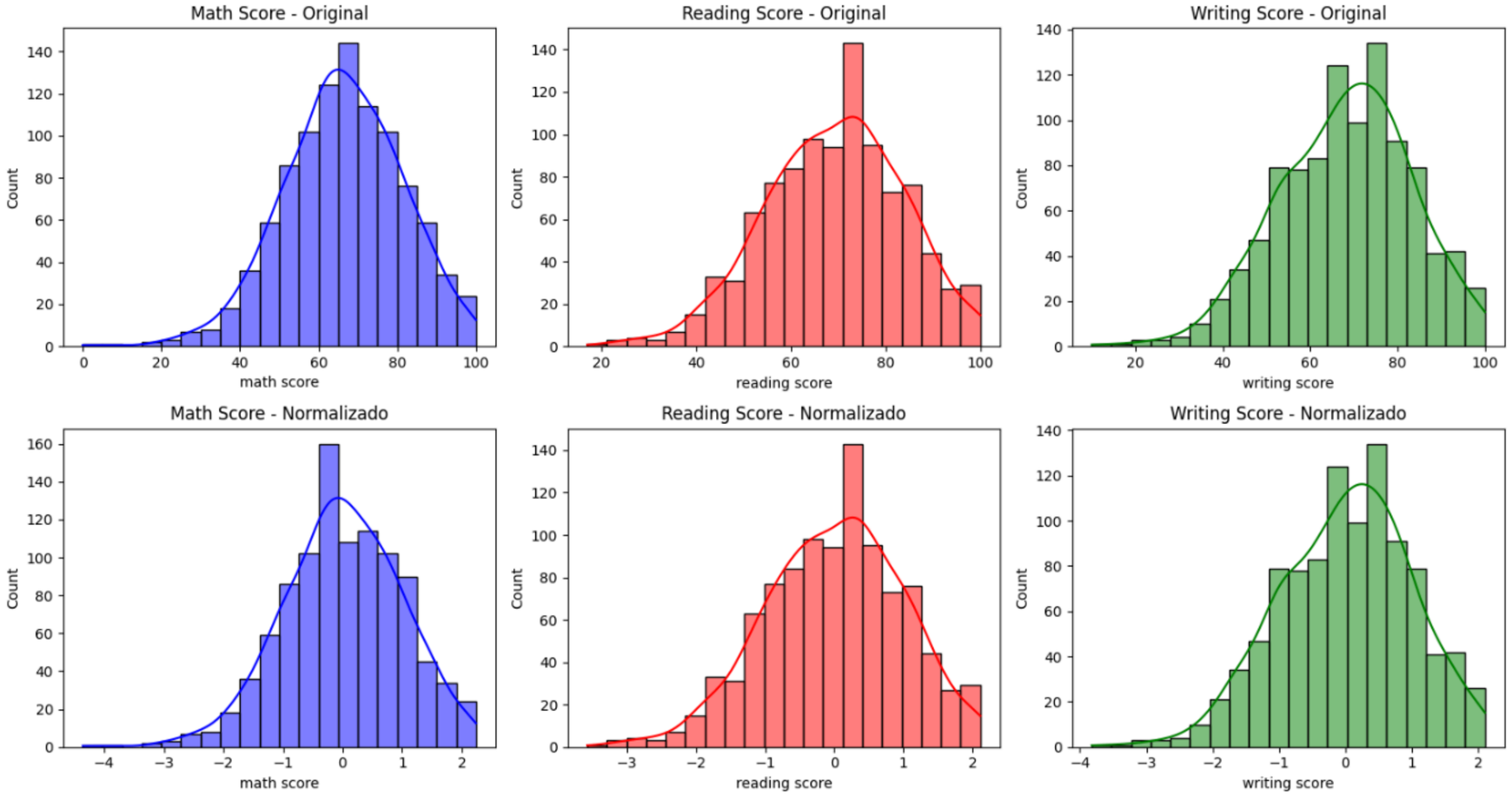
MATERIAIS E MÉTODOS

Normalização e Engenharia de características



MATERIAIS E MÉTODOS

Normalização e Engenharia de características



MATERIAIS E MÉTODOS

Normalização e Engenharia de características

Por que normalizar um DataFrame?

A normalização é importante porque:

- Evita que variáveis com escalas diferentes dominem a análise: Por exemplo, se um modelo de Machine Learning usa distância (como KNN ou regressão logística), variáveis com valores muito maiores podem influenciar mais do que outras.
- Melhora a convergência de modelos baseados em gradiente, como redes neurais e regressão logística.
- Ajuda a interpretar os dados mais facilmente, especialmente quando comparando diferentes características com escalas distintas.

Neste caso, a normalização das notas faz com que todas fiquem dentro de uma mesma escala (distribuição com média 0 e desvio padrão 1), facilitando a análise e garantindo que nenhuma disciplina tenha um peso desproporcional no modelo.

MATERIAIS E MÉTODOS

Seleção de Características

A seleção de características (feature selection) melhora a eficiência e a precisão do modelo porque:

- Reduz a dimensionalidade, tornando o modelo mais rápido e menos propenso a overfitting.
- Remove informações irrelevantes ou redundantes, melhorando a interpretabilidade dos resultados.
- Evita viés e colinearidade, garantindo que o modelo aprenda padrões significativos e não relações artificiais entre variáveis.

MATERIAIS E MÉTODOS

Experimentos

- Divisão dos Dados:

Para garantir uma avaliação justa do modelo, os dados foram divididos em:

- 80% para treino
- 20% para teste

Essa separação foi feita com a função train_test_split() da biblioteca sklearn, garantindo que o modelo aprenda com um conjunto e seja testado em dados nunca vistos antes.

- Modelo Escolhido:

O **Random Forest Regressor** foi utilizado por sua capacidade de lidar com dados complexos e capturar relações não-lineares entre variáveis

É baseado em múltiplas árvores de decisão, reduzindo o risco de overfitting e melhorando a precisão das previsões

MATERIAIS E MÉTODOS

Experimentos

- Treinamento do Modelo:

O modelo foi treinado com 100 árvores de decisão (n_estimators=100), garantindo um bom equilíbrio entre desempenho e tempo de execução.
O treinamento foi feito no conjunto de dados normalizado, com as variáveis categóricas transformadas via LabelEncoder

MATERIAIS E MÉTODOS

Avaliação

Métricas Utilizadas:

- Erro Médio Absoluto (MAE): Mede a diferença média entre as previsões e os valores reais. Quanto menor, melhor.
- Coeficiente de Determinação (R^2): Indica o quanto o modelo explica a variabilidade dos dados. Valores próximos de 1 indicam um bom ajuste.

MAE: 0.81

indica a diferença média entre as previsões e os valores reais

R^2: 0.03

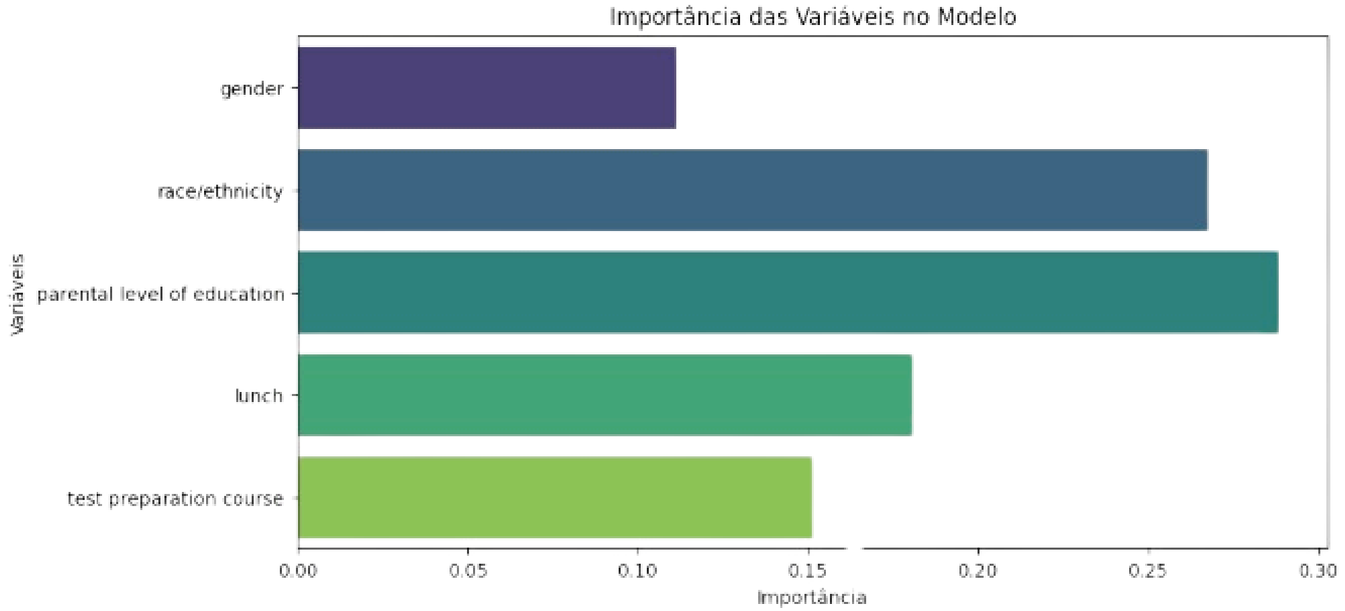
valor negativo indica que o modelo não conseguiu explicar a variabilidade dos dados

MATERIAIS E MÉTODOS

Avaliação

Análise do resultados

- O MAE é relativamente baixo, sugerindo que os erros individuais não são grandes.
- O R^2 negativo indica que o modelo não conseguiu capturar bem a relação entre as variáveis preditoras e o desempenho acadêmico.



RESULTADOS E DISCUSSÕES

Coleta e pré-processamento: Remoção de valores ausentes, normalização de variáveis e seleção de atributos.

Análise exploratória: Distribuição de variáveis, detecção de outliers e análise de correlação entre features.

Modelagem: Teste de diferentes algoritmos, comparação de métricas como acurácia e F1-score.

Avaliação: Desempenho do modelo em dados de teste, limitações e melhorias possíveis.

RESULTADOS E DISCUSSÕES

O nível educacional dos pais é o maior fator para as maiores notas dos alunos, seguido pela variável de raça/etnia.

A variável de raça/etnia não define que alguma raça é superior a outra, porém mostra que algumas raças/etnias estão ligadas a níveis educacionais e isso influencia as maiores médias dos alunos.

O gênero é a variável menos influente na relação das maiores notas dos alunos.

CONCLUSÃO

Objetivos Atingidos

Identificação de padrões de desempenho:
A análise exploratória revelou tendências e comportamentos no rendimento acadêmico.
.

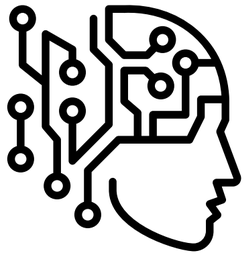
Relação entre fatores socioeconômicos e desempenho:
Análises correlacionais confirmaram essa influência.

Segmentação de alunos:
Técnicas de mineração de dados, como clustering, agruparam alunos com características semelhantes.

Desenvolver modelos preditivos para estimar o desempenho dos alunos:
Algoritmos de regressão e classificação foram aplicados para estimar o desempenho acadêmico.

CONCLUSÃO

Limtações do trabalho



01

O dataset utilizado possui um número limitado de variáveis, o que restringe a análise de outros fatores externos (como métodos de ensino e ambiente escolar)

02

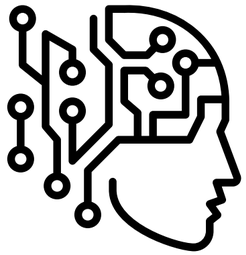
A pesquisa se baseou em um conjunto de dados específico, o que pode limitar a generalização dos resultados para outras populações estudantis

03

O modelo de machine learning utilizado, embora eficiente, poderia ser comparado com outros modelos mais avançados para verificar possíveis melhorias

CONCLUSÃO

Trabalhos Futuros



01

Expansão do Dataset

Coletar novos dados que incluam mais informações sobre os alunos, como hábitos de estudo, presença em sala de aula e nível de envolvimento dos professores

02

Testar Outros Modelos de Machine Learning

Explorar redes neurais, XGBoost ou modelos de regressão mais complexos para melhorar a acurácia das previsões.

03

Aplicação em Ambientes Educacionais

Desenvolver um sistema baseado nessas análises para ajudar escolas a identificar alunos em risco e oferecer suporte personalizado

PRINCIPAIS REFERÊNCIAS

01 COSTA, E.; BAKER, R. S.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. Jornada de Atualização em Informática na Educação, v. 1, n. 1, p. 1-29, 2012.

02 SIMON, A.; CAZELLA, S. Mineração de Dados Educacionais nos Resultados do ENEM de 2015. In: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 6., 2017, Recife. Anais dos Workshops do Congresso Brasileiro de Informática na Educação. Recife: SBC, 2017. p. 754

03 SOUZA, V. F. de. Mineração de dados educacionais com aprendizagem de máquina. Revista Educar Mais, v. 5, n. 4, p. 766-787, 2021.

04 BATISTA, M. R.; ARAÚJO FAGUNDES, R. A. de. Mineração de dados educacionais aplicada à performance de estudantes: uma revisão sistemática da literatura. Revista Novas Tecnologias na Educação, v. 21, n. 1, p. 271-280, 2023.

05 SILVA, J. C. S.; RODRIGUES, R. L.; RAMOS, J. L. C.; SOUZA, F. D. F.; GOMES, A. S. Mineração de dados educacionais orientada por atividades de aprendizagem. Revista Novas Tecnologias na Educação, v. 14, n. 1, 2016.

06 SILVA, L. A.; MORINO, A. H.; SATO, T. M. C. Prática de mineração de dados no Exame Nacional do Ensino Médio. In: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 3., 2014. Anais dos Workshops do Congresso Brasileiro de Informática na Educação. p. 651.

OBRIGADO!

www.linkedin.com/in/rafaelfreitasdados

OBRIGADO!

www.linkedin.com/in/rafaelfreitasdados