Análise Preditiva da Nota Final em Ciência de Dados

Membros: Victor Matheus (01716714),
José Humberto Silva de Araújo – 01589405,
Naeliton Chavez - 01594737

Contexto e Objetivos

O Problema de Negócio

Como podemos prever o desempenho acadêmico de estudantes em cursos superiores de ciência de dados? Identificar fatores-chave e construir um modelo robusto para antecipar a nota final é crucial para intervenções pedagógicas e otimização do ensino.

Objetivo do Projeto

Desenvolver um modelo preditivo capaz de estimar a nota final de estudantes, utilizando técnicas de aprendizado de máquina. A precisão na previsão permitirá a identificação precoce de alunos em risco e a personalização de estratégias de apoio.

Metodologia: Uma Jornada em Quatro Etapas

Nossa abordagem seguiu um fluxo de trabalho estruturado para garantir a robustez e a confiabilidade do modelo preditivo.

1

1. Exploração de Dados (EDA)

Compreensão profunda das características do dataset, identificação de padrões e anomalias.

2

2. Pré-processamento

Limpeza, transformação e preparação dos dados para otimizar a performance dos modelos.

3

3. Modelagem

Avaliação e seleção dos algoritmos de machine learning mais adequados ao problema.

4

4. Otimização

Ajuste fino dos hiperparâmetros e validação do modelo para resultados precisos.

Exploração de Dados (EDA): Desvendando o Dataset

Visão Geral do Dataset

- Total de **2.510 registros** de estudantes de cursos superiores.
- Variáveis diversas incluindo notas anteriores, demografia e engajamento.
- Identificação de tipos de dados (numéricos, categóricos) e distribuição inicial.

Análise da Variável Alvo: final_grade

A distribuição da nota final mostra uma tendência de concentração em notas intermediárias, com poucos outliers em extremos, indicando um cenário de avaliação comum.

Principais Correlações e Insights

- `previous_scores`: O preditor mais forte para a nota final, sublinhando a importância do desempenho prévio.
- Engajamento em atividades: Correlação moderada, sugerindo que a participação ativa influencia o resultado.
- Variáveis demográficas: Impacto sutil, mas presente, na distribuição das notas.

EDA: Desafios e Próximos Passos

A fase de Exploração de Dados revelou pontos críticos que demandaram atenção no pré-processamento.

Missing Values

Identificação de dados faltantes em colunas-chave, exigindo uma estratégia de imputação cuidadosa para não distorcer as análises.

Outliers

Presença de valores atípicos que poderiam influenciar negativamente o treinamento do modelo, necessitando de tratamento específico.

Variáveis Categóricas

Variáveis não numéricas que precisaram ser convertidas para um formato compreensível pelos algoritmos de Machine Learning.

Pré-processamento: Refinando os Dados

A qualidade dos dados é fundamental para a performance do modelo. Esta etapa focou em tornar os dados utilizáveis e otimizados.

Tratamento de Dados

- **Missing Values:** Imputação pela mediana para variáveis numéricas e pela moda para categóricas, preservando a distribuição original.
- **Outliers:** Análise e capping ou remoção de valores extremos, evitando que o modelo aprendesse com ruídos.

Transformações Essenciais

- **Encoding:** Utilização de One-Hot Encoding para variáveis categóricas nominais e Label Encoding para ordinais.
- **Padronização/Normalização:** Aplicação de StandardScaler para garantir que todas as features numéricas contribuíssem igualmente para o modelo, evitando o domínio de variáveis com escalas maiores.

Modelagem: Escolha e Avaliação dos Algoritmos

Para a tarefa de regressão, exploramos uma variedade de modelos preditivos, cada um com suas características.

1

Regressão Linear

Modelo base, simples e interpretável, serviu como ponto de comparação inicial.

2

Random Forest

Algoritmo ensemble robusto, conhecido por sua capacidade de lidar com a não-linearidade e reduzir overfitting.

3

XGBoost

Modelo de boosting de gradiente, de alta performance e amplamente utilizado em competições de Machine Learning.

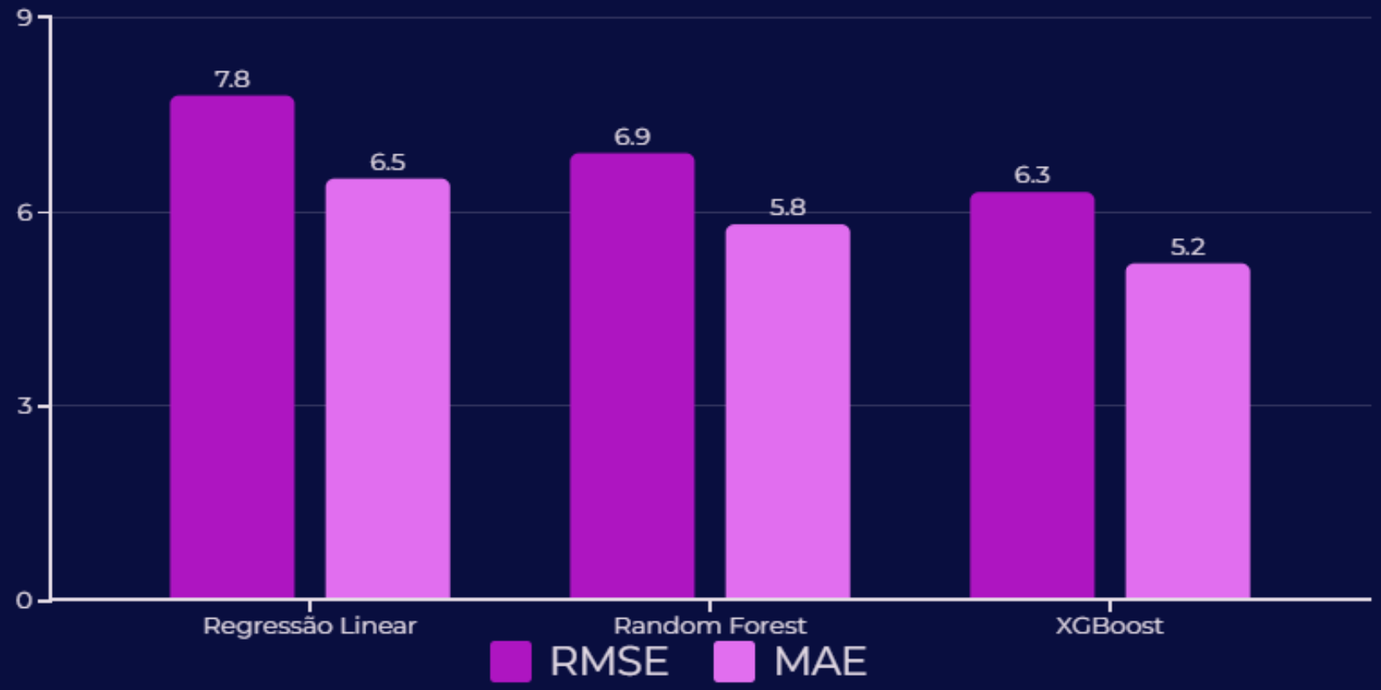
Métricas e Comparação de Modelos

A escolha do melhor modelo foi baseada em métricas de avaliação que refletem a performance real para predição da nota final.

Métricas de Avaliação

- **MAE (Mean Absolute Error):** Mede a média dos erros absolutos, oferecendo uma visão direta da magnitude dos erros.
- **RMSE (Root Mean Squared Error):** Penaliza erros maiores, sendo útil quando grandes desvios são indesejáveis.
- **R^2 (Coeficiente de Determinação):** Indica a proporção da variância na variável dependente que é previsível a partir das variáveis independentes.

Desempenho no Conjunto de Validação



O **XGBoost** demonstrou consistentemente o melhor desempenho nas métricas MAE e RMSE, justificando sua seleção como modelo final.

Resultados Finais e Análise Aprofundada

Após a otimização, o modelo XGBoost apresentou um desempenho robusto no conjunto de teste.

6.3

MAE Final

Em média, o erro absoluto na predição da nota final é de apenas 6.3 pontos.

0.84

R^2 Ajustado

O modelo explica 84% da variância na nota final, indicando um alto poder preditivo.

Otimização: GridSearchCV

Utilizamos **GridSearchCV** para explorar sistematicamente diferentes combinações de hiperparâmetros do XGBoost, garantindo que o modelo alcançasse seu potencial máximo.

Análise de Resíduos

A distribuição dos resíduos demonstrou aleatoriedade e ausência de padrões, confirmando a robustez e a adequação do modelo aos dados.

Conclusões e Próximos Passos

Principais Resultados

- Modelo XGBoost com alta precisão na previsão da nota final (MAE 6.3, R^2 0.84).
- `previous_scores` é o preditor mais influente, seguido por engajamento e participação.
- Potencial para intervenções pedagógicas direcionadas.

Limitações

- Dataset específico, pode não generalizar para outras instituições sem re-treinamento.
- Variáveis subjetivas (motivação, inteligência) não foram incluídas devido à falta de dados.
- Modelo não captura fatores exógenos imprevistos.

Trabalhos Futuros

- Incorporar dados de engajamento em tempo real (plataforma de e-learning).
- Explorar técnicas de Deep Learning para dados textuais (fóruns, etc.).
- Construir uma interface de usuário para fácil utilização do modelo.