PROJETO FINAL DE MACHINE LEARNING

Previsão de Desempenho Acadêmico

Desenvolvimento de Modelo de Regressão para Pontuação_Prova_Final

O Desafio

Problema de Negócio

Quais fatores predizem a Pontuação Final de um estudante, e qual a acurácia de um modelo de Machine Learning para essa previsão?

Objetivo Técnico

Desenvolver um modelo de **Regressão** que minimize o MAE e maximize o R^2.

Etapas do Projeto

01

Análise Exploratória de Dados

Investigação inicial do dataset e identificação de padrões.

02

Pré-processamento

Limpeza, transformação e preparação dos dados.

03

Modelagem e Baseline

Estabelecimento do modelo base com $R^2 = 0.65$.

04

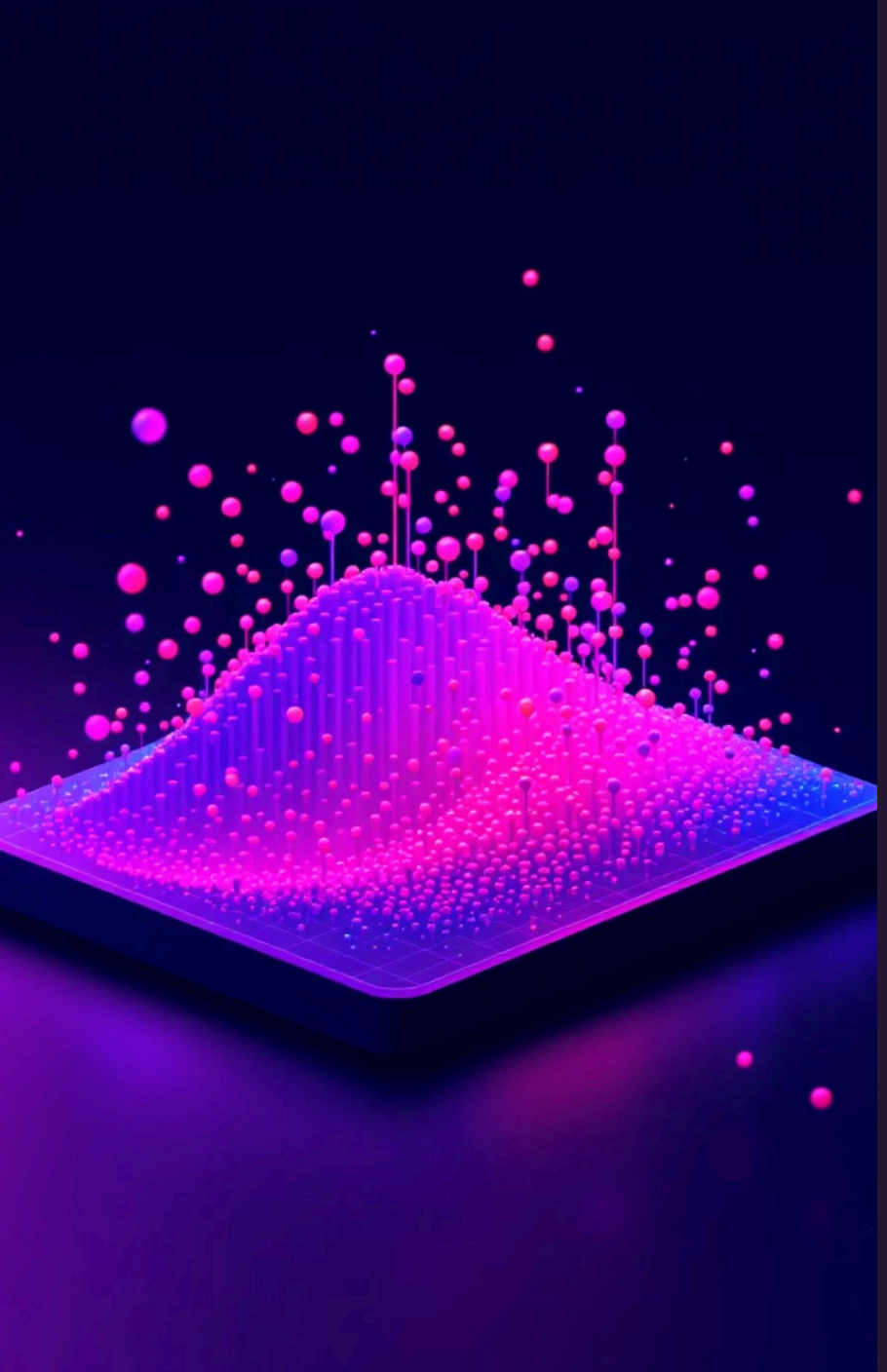
Otimização e Tuning

Ajuste de hiperparâmetros para melhor desempenho.

05

Relatório e Documentação

Consolidação dos resultados e análises finais.



Análise Exploratória: O Grande Preditivo

12K

Observações

Dataset limpo, sem valores ausentes ou outliers.

+0.97

Correlação

Entre Horas_Estudo_Semana e Pontuação Final.

A descoberta chave revela uma correlação **extremamente forte** ($r \approx +0.97$) entre horas de estudo semanais e o desempenho na prova final.

Preparação dos Dados

1

Feature Engineering

Criação da feature **Horas_por_Idade** para capturar relações não-lineares.

2

Encoding

Label Encoding aplicado à variável categórica Aprovado (True/False).

3

Escalonamento

Uso do **StandardScaler** para padronizar todas as features numéricas.

Modelo Base e Seleção

1

Baseline

Regressão Linear estabeleceu $R^2 = 0.65$ como ponto de partida.

2

Comparação

Testamos Linear Regression vs. Random Forest Regressor.

3

Seleção Final

Random Forest Regressor atingiu **R^2 = 0.9567** no conjunto de validação.

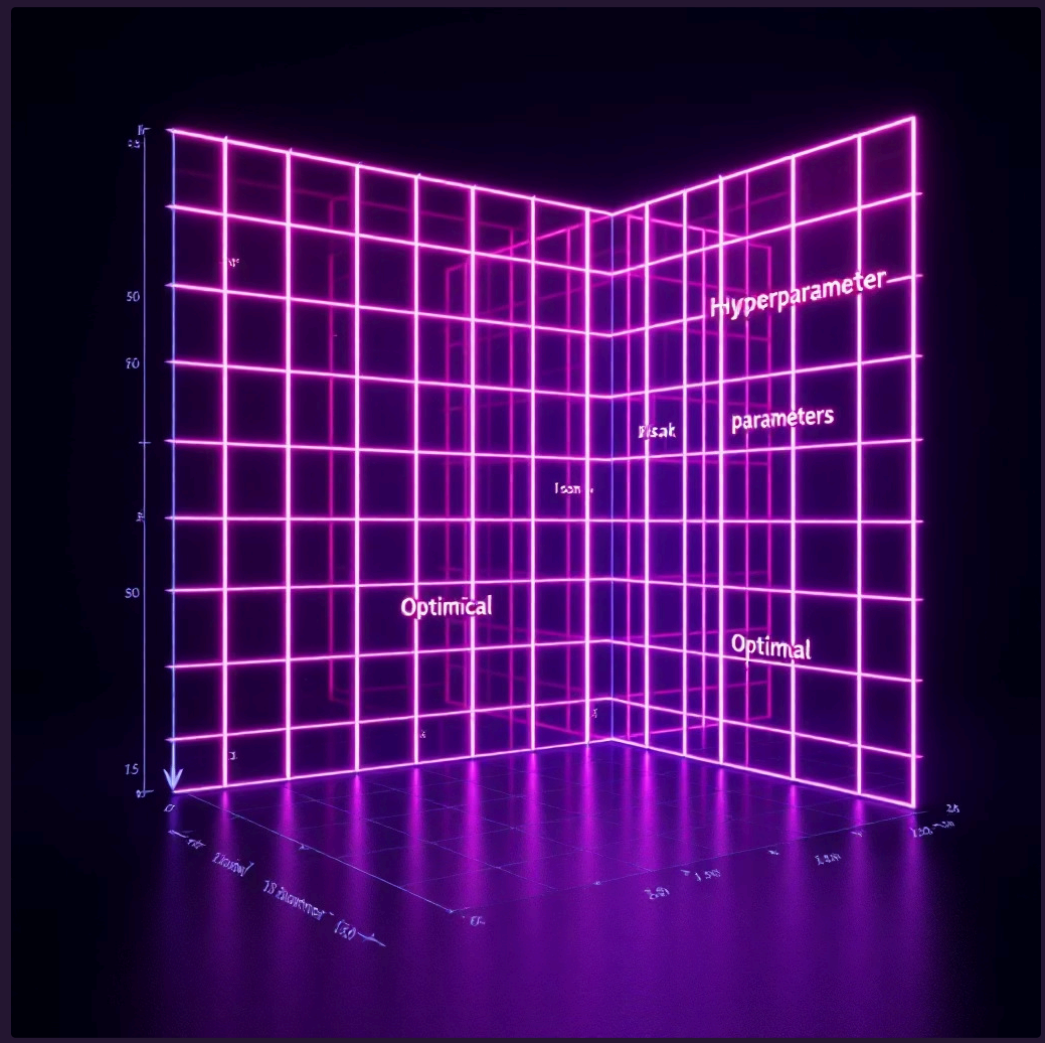
Otimização via Grid Search

Técnica Utilizada

Aplicamos **Grid Search com Cross-Validation** para ajustar hiperparâmetros e maximizar o desempenho do modelo.

Melhores Parâmetros

- max_depth: 5
- min_samples_leaf: 1
- min_samples_split: 2
- n_estimators: 100



Desempenho Final no Teste

Comparação entre o modelo baseline e o Random Forest otimizado:

Métrica	Baseline (R^2 = 0.65)	RF Otimizado (TESTE)
R^2	0.65	**0.9562**
MAE	N/A	4.65

Interpretação dos Resultados

Alta Precisão

O modelo Random Forest Otimizado é **altamente preditivo** com $R^2 = 0.9562$.

Erro Mínimo

O **MAE de 4.65** significa que a previsão erra, em média, por apenas 4.65 pontos na nota final.

Otimização

A otimização não gerou ganho significativo, pois o modelo padrão já estava próximo do ideal.

Robustez e Próximos Passos

Análise de Resíduos

A distribuição é **aleatória e centrada em zero**, indicando robustez do modelo.

Limitações

O modelo tende a **subestimar valores reais muito altos** da pontuação - limitação típica do Random Forest.

Trabalhos Futuros

1. Testar algoritmos de boosting avançados (**XGBoost** e **LightGBM**)
2. Explorar features de interação mais complexas para ganho marginal

Repositório GitHub