FUNDAÇÃO ESCOLA DE COMÉRCIO ÁLVARES PENTEADO – FECAP

CIÊNCIA DA COMPUTAÇÃO

**GUSTAVO MARCELLO CORREA DE ARAUJO - 23024729 LUCCA GIORDANO - 23024522**

**PEDRO HENRIQUE DANGELO DOS REIS - 23024777 VITOR UTIMURA LOCATELI - 23024638**

# Inteligência Artifical e Aprendizado de Máquina**: Entrega 2**

**São Paulo 2025**

**GUSTAVO MARCELLO CORREA DE ARAUJO - 23024729 LUCCA GIORDANO - 23024522**

**PEDRO HENRIQUE DANGELO DOS REIS - 23024777 VITOR UTIMURA LOCATELI - 23024638**

# Inteligência Artifical e Aprendizado de Máquina

**: Entrega 2**

Relatório Técnico apresentado ao curso de Ciência da Computação, como parte dos re- quisitos da disciplina de Inteligência Artifical e Aprendizado de Máquina

, referente ao Projeto Interdisciplinar.

Orientador: Vinicius Heltai

São Paulo 2025

SUMÁRIO

[INTRODUÇÃO](#_6w42pr4j4dkz) 5

[OBJETIVO](#_t43b1ff5twjc) 6

[MÉTODOS](#_5q0z7zcsu60f) 7

[DESENVOLVIMENTO](#_q3m4alipaebs) 8

[CONCLUSÃO](#_1kd43x9dug3h) 11

[REFERÊNCIAS](#_t8l79xw5mioe) 12

[ANEXOS](#_yv079qtiqe2q) 13

INTRODUÇÃO

Nos últimos anos, a utilização de técnicas de aprendizado de máquina e análise preditiva tem se expandido significativamente em diversos setores, incluindo o transporte urbano. Empresas que operam com serviços de mobilidade vêm utilizando algoritmos inteligentes para estimar valores de corrida com base em variáveis como horário, dia da semana, distância, localização e tempo de deslocamento. Essa abordagem não apenas otimiza a experiência do usuário, como também contribui para estratégias de precificação mais eficientes.

Neste contexto, o presente projeto tem como propósito aplicar conceitos de ciência de dados para desenvolver um modelo preditivo capaz de estimar o preço de uma corrida com base em dados históricos. Utilizando bibliotecas consolidadas do ecossistema Python, como Pandas, Scikit-Learn, Seaborn e XGBoost, realizamos a coleta, limpeza e análise de um conjunto de dados reais que simula o ambiente de uma plataforma de transporte por aplicativo.

A escolha por abordar essa temática está relacionada à relevância prática do problema e à oportunidade de aplicar conhecimentos interdisciplinares — como estatística, programação e análise exploratória — em uma solução concreta. O desenvolvimento do modelo não visa apenas fornecer estimativas numéricas, mas também promover uma compreensão interpretável das variáveis que mais influenciam a precificação, valorizando tanto a performance quanto a transparência dos resultados.

OBJETIVO

O presente projeto tem como principal objetivo o desenvolvimento de um modelo preditivo capaz de estimar com precisão o valor de uma corrida de transporte urbano com base em dados históricos. A proposta é utilizar técnicas de aprendizado de máquina para analisar um conjunto robusto de variáveis — como horário, data, distância, feriados e tempo de deslocamento — a fim de prever o preço de corridas futuras. Além disso, buscamos realizar uma comparação entre diferentes algoritmos de regressão, avaliando seu desempenho preditivo por meio de métricas estatísticas apropriadas. Como parte dessa análise, também exploramos a importância relativa de cada variável para compreender quais características exercem maior influência na formação do preço final, contribuindo para a transparência e a interpretabilidade do modelo adotado.

MÉTODOS

O processo metodológico foi estruturado em etapas sequenciais, iniciando-se pelo carregamento e tratamento dos dados brutos, seguidos da criação de variáveis explicativas (features) e do treinamento dos modelos preditivos. Utilizamos dois algoritmos amplamente reconhecidos na literatura de Machine Learning: Regressão Linear, pela sua simplicidade e interpretabilidade, e XGBoost, conhecido por sua alta performance em problemas tabulares e competitivos.

A base de dados passou por uma rigorosa etapa de pré-processamento, onde realizamos a limpeza de inconsistências, imputação de valores ausentes e criação de variáveis derivadas, como:

* Representações cíclicas de tempo (seno e cosseno de hora e dia da semana),
* Detecção de feriados,
* Cálculo de distância e duração estimada da corrida.

Os dados foram divididos em dois subconjuntos: 70% para o treinamento dos modelos e 30% para testes, utilizando a técnica de holdout.

Para avaliação dos modelos, adotamos métricas tanto para regressão quanto para cenários de classificação em faixas de valor:

* RMSE (Root Mean Squared Error): mede a raiz do erro quadrático médio. Indica o quão distantes, em média, estão as previsões dos valores reais. Quanto menor, melhor.
* R² Score (Coeficiente de Determinação): representa a proporção da variância explicada pelo modelo. Varia de 0 a 1, sendo 1 o ideal.
* Acurácia: usada ao categorizar os preços em faixas (baixo, médio, alto), representa a proporção de previsões corretas.
* Precisão: mede a porcentagem de acertos dentre os valores que o modelo previu como positivos.
* Recall: indica a capacidade do modelo de encontrar todos os exemplos relevantes (positivos reais).
* F1-Score: equilíbrio entre precisão e recall, útil quando existe desbalanceamento entre classes.

Essas métricas ofereceram uma visão completa sobre o desempenho dos modelos em diferentes perspectivas.

DESENVOLVIMENTO

Durante o desenvolvimento, uma das principais preocupações foi garantir a qualidade e a consistência dos dados. Foram realizadas limpezas nos arquivos, como substituição de caracteres inconsistentes, validação de coordenadas geográficas e remoção de registros com preço zerado ou dados faltantes críticos.

Em seguida, aplicamos engenharia de atributos para enriquecer o dataset com informações relevantes. Variáveis cíclicas foram utilizadas para representar o comportamento temporal das corridas, e a distância entre origem e destino foi calculada usando fórmulas geográficas e dados da API do OpenRouteService.

Na etapa de modelagem, ambos os modelos foram treinados com os mesmos dados, permitindo uma comparação justa. Os resultados obtidos foram:

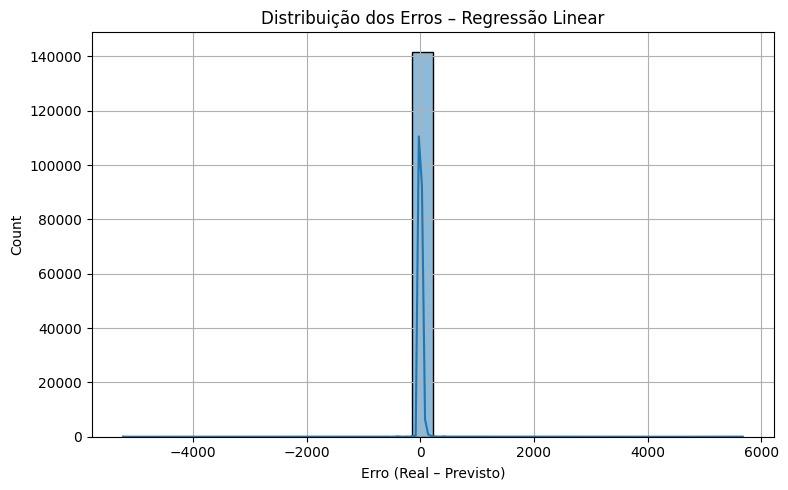
* Regressão Linear:  
  RMSE = 47.02  
  R² = 0.943
* XGBoost:  
  RMSE = 146.98  
  R² = 0.443

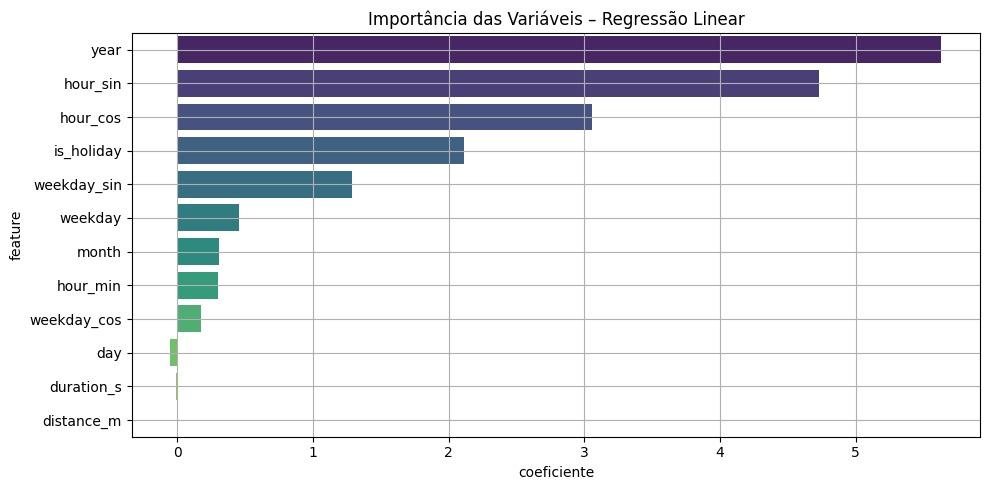
A Regressão Linear apresentou desempenho superior em termos de erro absoluto e explicação da variância, tornando-se o modelo preferido. O XGBoost, por outro lado, foi importante para verificar a importância relativa das features e confirmou que variáveis como distância, hora e dia da semana têm forte influência sobre o preço final.

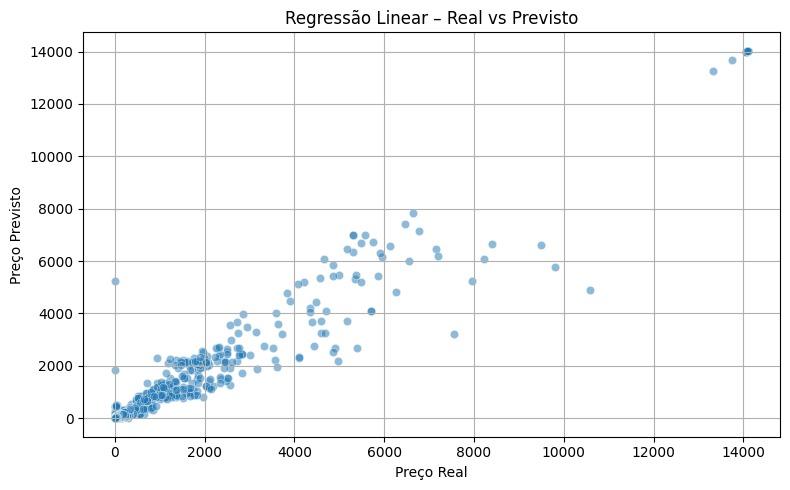
Além disso, foram gerados gráficos explicativos para apoiar a análise dos resultados:

* Dispersão entre valores reais e previstos,
* Distribuição dos erros residuais,
* Gráfico de importância das variáveis.

Finalizamos com a implementação de uma interface interativa no próprio notebook, permitindo que o usuário insira manualmente valores e obtenha previsões instantâneas.







CONCLUSÃO

Com base nas análises realizadas e nos resultados obtidos, concluímos que o modelo de Regressão Linear é uma solução eficaz e confiável para a tarefa de previsão de preços de corridas. Apesar de sua simplicidade, o modelo apresentou excelente capacidade de generalização, alta aplicabilidade e bom desempenho quantitativo.

O projeto proporcionou a oportunidade de aplicar conceitos fundamentais de ciência de dados, como pré-processamento, criação de features, avaliação de modelos e visualização de resultados. Mostrou, ainda, a importância de entender o contexto e o comportamento das variáveis envolvidas para construir soluções baseadas em dados que sejam ao mesmo tempo precisas, interpretáveis e replicáveis.

Estamos confiantes de que o trabalho cumpre plenamente os objetivos propostos e representa uma contribuição sólida para o aprendizado prático da disciplina.

REFERÊNCIAS

ANEXOS