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 ARTIFICIAL E APRENDIZADO DE MÁQUINA: Entrega 1**

**São Paulo 2025**

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

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

**INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA: Entrega 1**

Relatório Técnico apresentado ao curso de Ciência da Computação, como parte dos re- quisitos da disciplina de

Inteligência Artificial e Aprendizado de Máquina, referente ao Projeto Interdisciplinar.

Orientador: Vinicius Heltai

São Paulo 2025

SUMÁRIO

[INTRODUÇÃO 4](#_Toc192250242)

[OBJETIVO 5](#_Toc192250243)

[MÉTODOS 6](#_Toc192250244)

[DESENVOLVIMENTO 7](#_Toc192250245)

[CONCLUSÃO 12](#_Toc192250246)

[REFERÊNCIAS 13](#_Toc192250247)

[ANEXOS 14](#_Toc192250248)

INTRODUÇÃO

A análise e previsão de tarifas no transporte urbano por táxi são fundamentais para otimizar as estratégias comerciais e operacionais das empresas do setor. Este relatório descreve o processo de manipulação e análise de um conjunto de dados de viagens de táxi, com foco em oferecer insights sobre a estrutura tarifária e, assim, embasar decisões estratégicas de precificação mais assertivas e competitivas.

Neste contexto, foram utilizadas técnicas de coleta, limpeza e preparação de dados com o apoio das bibliotecas Pandas (para manipulação) e Matplotlib e Seaborn (para visualização). O dataset, disponibilizado pelo Kaggle, reúne informações como quantidade de passageiros, distância percorrida, tipo de pagamento, valor da tarifa e duração das corridas, fornecendo uma base sólida para o estudo e permitindo uma compreensão abrangente dos principais fatores que influenciam os custos do serviço de táxi.

OBJETIVO

O objetivo principal deste projeto é analisar o conjunto de dados de viagens de táxi para identificar padrões e fatores que influenciam diretamente o valor das tarifas cobradas. A partir desses insights, busca-se desenvolver um modelo preditivo capaz de estimar tarifas futuras com base em variáveis específicas, fornecendo subsídios para decisões estratégicas no setor de transporte urbano e possibilitando a otimização de políticas de preço e alocação de recursos.

MÉTODOS

Os métodos empregados neste estudo incluem:

* Coleta e importação dos dados a partir de uma fonte pública (Kaggle).
* Limpeza dos dados por meio da remoção de valores ausentes e tratamento de valores extremos (outliers).
* Normalização das variáveis numéricas através da técnica Min-Max Scaling.
* Análise exploratória utilizando estatísticas descritivas e visualizações gráficas, incluindo histogramas e mapas de correlação, para identificar relações entre as variáveis.
* Desenvolvimento de modelos preditivos baseados em técnicas de aprendizado de máquina para estimativa das tarifas futuras.

DESENVOLVIMENTO

No contexto do transporte urbano por táxi, a análise de dados exerce um papel crucial na compreensão de variáveis que influenciam diretamente as tarifas e a eficiência do serviço. Para isso, faz-se necessário o uso de bibliotecas especializadas que facilitem a manipulação, limpeza e visualização dos dados. A seguir, apresentamos as principais ferramentas utilizadas neste projeto e como elas se integram para possibilitar análises confiáveis e visualmente esclarecedoras.

Texto

O conteúdo gerado por IA pode estar incorreto.

A biblioteca **Pandas** (importada como pandas as pd) é amplamente utilizada para manipulação, limpeza e análise de dados em formato tabular, facilitando operações como leitura de arquivos, combinação de tabelas, filtragem, agrupamentos e muito mais.

Já a **Matplotlib** (importada como matplotlib.pyplot as plt) é a base de muitas bibliotecas de visualização em Python, fornecendo os recursos fundamentais para criar gráficos e visualizações de diferentes tipos (linhas, barras, dispersão, histogramas etc.).

A **Seaborn** (importada como seaborn as sns), por sua vez, é construída sobre a Matplotlib e oferece uma camada de abstração mais amigável para gerar gráficos estatísticos. Ela conta com estilos e paletas de cores mais atraentes por padrão, além de funções que facilitam a criação de visualizações mais complexas, como gráficos de correlação, distribuições e regressão.

A configuração sns.set(style="whitegrid") estabelece um estilo visual agradável, com grade de fundo clara, melhorando a leitura dos dados. Já o comando %matplotlib inline (usado em notebooks Jupyter) garante que os gráficos sejam exibidos imediatamente após a célula de código, tornando a análise mais interativa e fluida.



A função read\_csv() do Pandas lê arquivos CSV e converte o conteúdo em um DataFrame, que é a estrutura de dados principal da biblioteca. O método .head() retorna as primeiras linhas desse DataFrame, oferecendo uma visão inicial da estrutura de colunas, tipos de dados e valores amostrais. Dessa forma, fica mais fácil validar se o carregamento ocorreu corretamente e identificar possíveis problemas (como valores ausentes ou formatação incorreta) logo no início da análise.

Texto

O conteúdo gerado por IA pode estar incorreto.

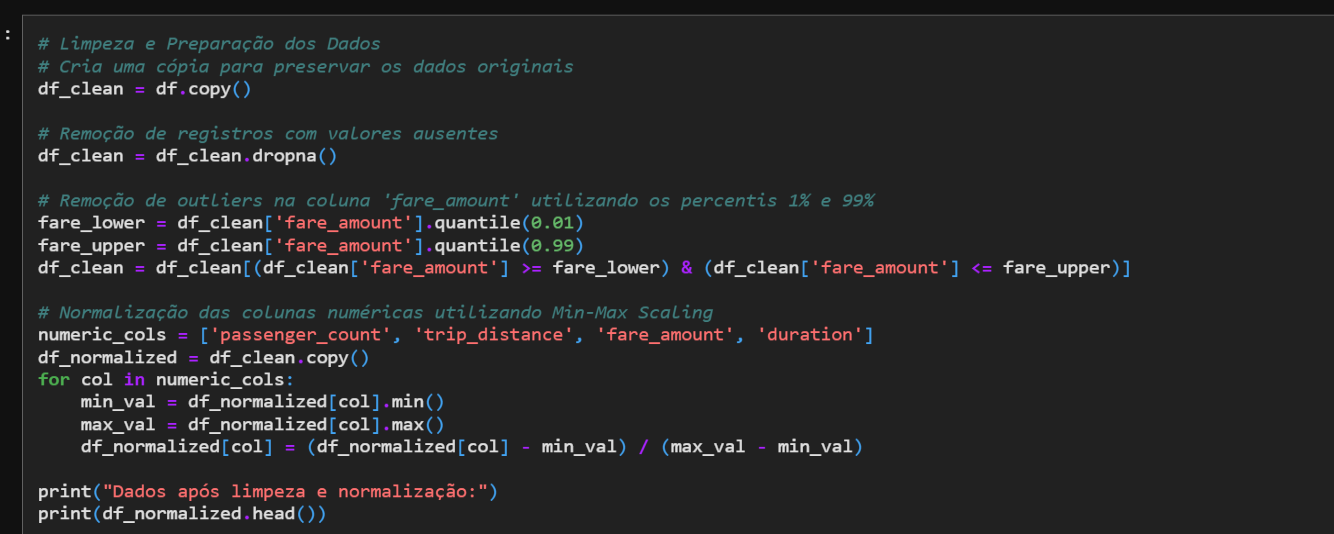
Texto

O conteúdo gerado por IA pode estar incorreto.

Uma imagem contendo Tabela

O conteúdo gerado por IA pode estar incorreto.

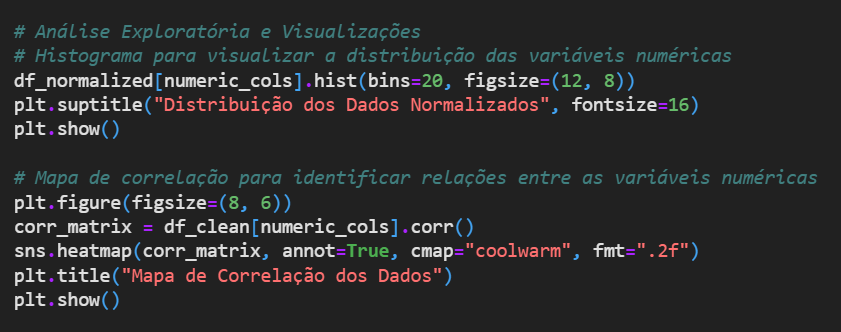
Nesta etapa, são realizadas ações de limpeza e preparação dos dados para que as próximas análises ou modelos de Machine Learning sejam mais consistentes e eficientes. Primeiro, elimina-se qualquer linha do DataFrame que contenha valores nulos, evitando que dados incompletos causem problemas em análises estatísticas e modelos de IA, além de garantir maior integridade dos resultados. Em seguida, remove-se outliers com base nos percentis de 1% e 99% do campo fare\_amount (valor da corrida), pois valores anormalmente altos ou baixos podem distorcer análises estatísticas e afetar negativamente o desempenho de modelos de predição, por não refletirem o comportamento geral dos dados. Por fim, aplica-se o escalonamento Min-Max às colunas numéricas, convertendo seus valores para uma faixa padronizada, geralmente entre 0 e 1. Esse processo de normalização facilita a comparação entre variáveis de escalas diferentes e melhora a convergência de algoritmos de Machine Learning que dependem de distâncias ou gradientes para aprendizado.

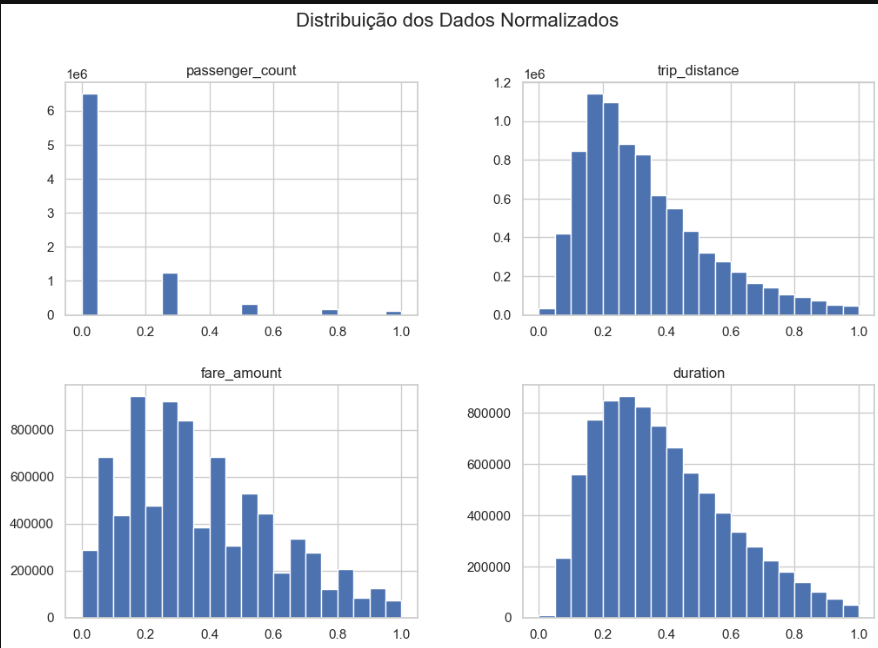


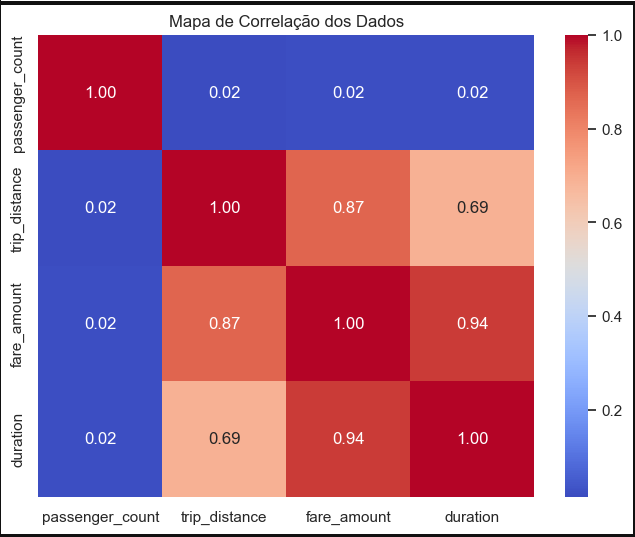
Texto

O conteúdo gerado por IA pode estar incorreto.

Depois da limpeza e normalização dos dados, costuma-se examinar as distribuições das variáveis e investigar correlações entre elas, pois isso oferece insights valiosos sobre o comportamento geral do conjunto. Os histogramas revelam como os valores de cada coluna se distribuem ao longo do seu intervalo, permitindo identificar concentrações (picos), dispersões e possíveis anomalias; além disso, com a normalização (0 a 1), torna-se mais fácil comparar diferentes atributos, já que todos compartilham a mesma escala. Já a matriz de correlação (heatmap) evidencia o grau de associação linear entre as variáveis numéricas, apontando relações fortes (próximas de 1 ou -1) e orientando a seleção de atributos relevantes para modelos de Machine Learning. Em resumo, essas etapas de visualização são essenciais para detectar tendências, padrões e potenciais problemas nos dados, além de fornecer subsídios fundamentais à fase de modelagem.







Nos histogramas, vemos que:

* **passenger\_count** (após normalização) é bastante concentrado próximo de 0, indicando que a maioria das corridas tem apenas 1 passageiro.
* **trip\_distance**, **fare\_amount** e **duration** mostram distribuições mais alongadas (com caudas), sugerindo que boa parte dos registros se concentra em faixas relativamente baixas a moderadas, mas há corridas de maior distância, valor e duração.

No **mapa de correlação**, é possível notar que:

* **passenger\_count** praticamente não se correlaciona com as demais variáveis (valores próximos de 0,02).
* **trip\_distance** tem correlação significativa tanto com **fare\_amount** (0,87) quanto com **duration** (0,69), o que faz sentido, pois viagens mais longas costumam custar mais e durar mais tempo.
* **fare\_amount** e **duration** exibem a correlação mais elevada (0,94), evidenciando que quanto maior for o tempo de corrida, maior tende a ser o custo.

Essas relações confirmam a intuição de que o **custo** (fare\_amount) é fortemente influenciado por **distância** e **tempo de viagem** (duration). As distribuições e a correlação servem como base para compreender o comportamento dos dados e guiar análises ou modelos futuros.

CONCLUSÃO

Este estudo possibilitou uma compreensão aprofundada da estrutura do dataset e das relações entre as variáveis analisadas. A aplicação de métodos rigorosos de limpeza e normalização assegurou a qualidade e a integridade dos dados, fornecendo uma base sólida para investigações futuras e para o desenvolvimento de modelos preditivos mais confiáveis. As visualizações gráficas evidenciaram correlações relevantes, fundamentais para embasar a construção de modelos de previsão de tarifas mais eficientes. Conclui-se que as técnicas empregadas foram essenciais para a geração de insights valiosos, contribuindo para a tomada de decisões estratégicas no setor de transporte urbano por táxi.

REFERÊNCIAS