Carros Usados

Luis Henrique Turra Ramos

17 de setembro de 2025

Objetivo: Projeto teve foco em limpeza de dados, analise, dashboard e predição de preço de carros usados.

Metodologia:

Fontes de Dados https://www.kaggle.com/datasets/pratyushpuri/used-car-sales-listings-dataset-2025

Técnicas de Análise:

- Inserção de Dados: python, Jupyter Notebook;
- Limpeza de Dados: mysql;
- Analise de Dados: python,Jupyter Notebook;
- Dashboard: python, Jupyter Notebook e biblioteca Streamlit;
- Predição: python, Jupyter Notebook e biblioteca Streamlit.

Limitações: Base de dados sintética e limitada a 2,068 linhas.

Processo: A base de dados em formato CSV contem informação de venda de carros usados da América do norte, Europa, oriente médio, Ásia-pacifico e américa latina. Contem carros de luxo e premium, tipo de venda, modelo do carro, tipo do carro, tipo de combustão, transmissão, quilometragem, preço, localização e recursos extras.

A inserção dos dados em mysql foi feito usando python em jupyter notebook usando a biblioteca pymysql, em mysql foi feita uma análise para localização de erros e duplicados. Após disso usando jupyter notebook foi realizado o pré-processamento conectando com mysql, o pré-processamento consistiu alteração do tipo de dados, separação dos recursos extras em múltiplas colunas passando valor booleano e remoção de colunas não necessárias.

Na analise foi feito cruzamento de dados com foco na variação de preço, com base de montadora, o ano do carro, tipo de combustível, o tipo de venda, qual recurso extra do carro mais afeta o preço e média de preço por pais.

Usando a biblioteca Streamlit foi possível host dashboard que foi feita através de gráficos na análise e site da predição usando algoritmo machine learning Random Forest que através de entrada de dados do usuário fazer uma predição de um valor total do carro.

Resultados: Os dados da base já estavam tratados, a localização e recurso extras estão em uma coluna singular, durante pré-processamento foi feita a separação dessas colunas gerando mais colunas com recursos mais abranges com cidade, estado e pais, e recursos extra em valores booleano que ajuda na focalização do impacto desses recursos no preço dos carros.

A predição usando o algoritmo Random Forest pode não ser o mais complexo algoritmo, porem se enquadra melhor que algoritmos mais poderosos como redes neurais que tem melhor resultados em base de dados mais complexas. Os resultados da predição tiveram uma performance de 81%.