Relatório Técnico - Projeto de Machine Learning Aplicado ao Varejo

1. Introdução

Este projeto tem como objetivo aplicar técnicas de Machine Learning no setor de varejo, com foco na previsão do valor de venda de produtos com base em características históricas dos pedidos. Para isso, foi utilizado um conjunto de dados sintético realista e o algoritmo Random Forest.

2. Metodologia

O algoritmo Random Forest foi escolhido por sua robustez e capacidade de capturar não linearidades nos dados. Foram utilizadas bibliotecas como pandas, matplotlib, seaborn e scikit-learn para o tratamento e modelagem dos dados, além do Streamlit para criação de uma interface interativa.

O dataset foi gerado artificialmente com colunas como ID do pedido, data, segmento, localização, categoria e valor da venda.

3. Análise Exploratória

Foram realizadas análises descritivas para compreender a distribuição das vendas e o comportamento dos segmentos de clientes. Gráficos como histogramas e contagem por categoria foram gerados para melhor visualização dos padrões nos dados.

4. Pré-processamento e Modelagem

As variáveis categóricas foram codificadas com LabelEncoder e colunas como ano e mês foram extraídas da data do pedido. O modelo Random Forest foi treinado com 80% dos dados e avaliado com os 20% restantes, utilizando métricas como MSE e R^2. O modelo final foi salvo com joblib, assim como os encoders, permitindo reutilização futura sem reprocessamento.

5. Dashboard Interativo

Foi desenvolvido um dashboard interativo com Streamlit, onde o usuário pode:

- Visualizar distribuições e filtros por segmento e categoria

- Ver a importância das variáveis para o modelo
- Simular uma previsão de venda informando os dados do pedido

As entradas usam os encoders salvos para mostrar os valores originais legíveis.

6. Resultados e Conclusões

O modelo obteve bom desempenho, com R^2 superior a 0.80 na validação. A categoria do produto, o segmento do cliente e o estado foram algumas das variáveis mais relevantes. O sistema desenvolvido é facilmente adaptável para dados reais, e pode ser expandido com novas fontes de dados e algoritmos.

Como melhorias futuras, recomenda-se testar algoritmos como Gradient Boosting e implementar previsão em lote no dashboard.