Previsão de Demanda por Componente

Disciplina: Ciência de Dados

Universidade Estadual Paulista - UNESP

Aluno: **Marco Antonio Cerqueira De Queiroz**
RA: **211026141**

Data de Entrega: 25/06/2025

1 Introdução

A previsão de demanda é uma atividade essencial para o planejamento de produção, controle de estoque, logística e estratégias de marketing em empresas. Saber antecipadamente quais produtos terão maior ou menor saída permite otimizar recursos, evitar faltas e excessos, e melhorar o atendimento ao cliente.

Neste trabalho, desenvolvemos um modelo preditivo baseado em técnicas de ciência de dados para prever a demanda futura de componentes. Utilizamos um conjunto de dados com informações diárias sobre vendas, promoções, preços, clima, regiões, entre outros fatores. Modelos como Naive Bayes e Random Forest foram testados e comparados para verificar qual apresenta melhor desempenho na classificação de alta ou baixa demanda futura.

2 Objetivo

O principal objetivo é aplicar técnicas estatísticas e computacionais para prever, com base em dados históricos, se um determinado componente terá alta demanda em um período futuro. Isso é feito com a criação de um modelo de classificação binária, que rotula os registros como "alta demanda" ou "baixa demanda" conforme a variável `future_demand`. Pretende-se avaliar a performance dos modelos, compreender os atributos mais relevantes e verificar a viabilidade prática da abordagem em cenários reais.

3 Descrição do Conjunto de Dados

O dataset utilizado foi obtido do Kaggle e contém 4.999 registros. Cada linha representa a venda de um produto em um dia. A base contém atributos como:

- **sales_units**: unidades vendidas no dia;
- **promotion_applied**: se houve promoção (1) ou não (0);
- **economic_index**: índice que reflete a situação econômica;
- **weather_impact**: variável que indica impacto do clima;
- **price** e **discount_percentage**;
- **region_Europe**, **region_North America**;

- **store_type_Retail**, **store_type_Wholesale**;
- **Categorias**: Sofás, Cadeiras, Mesas e Armários;
- **future_demand**: valor da demanda futura (alvo).

A variável alvo foi transformada em binária (0 ou 1) com base na mediana de `sales_units`, marcando com 1 os registros com `future_demand` acima da mediana.

4 Pré-processamento dos Dados

Os principais passos de tratamento foram:

- Conversão de datas;
- Remoção de colunas irrelevantes para predição;
- Criação da variável `High_Demand`;
- Normalização das variáveis numéricas com `StandardScaler`;
- Divisão em treino (80%) e teste (20%) com `train_test_split`.

Todos os passos foram realizados com cuidado para evitar vazamento de dados (data leakage).

5 Análise Exploratória

Foram feitas visualizações para melhor compreensão do comportamento das variáveis. A distribuição de vendas é assimétrica à direita, indicando maior frequência de vendas baixas. A análise de correlação mostrou associações relevantes entre preço, desconto e vendas.

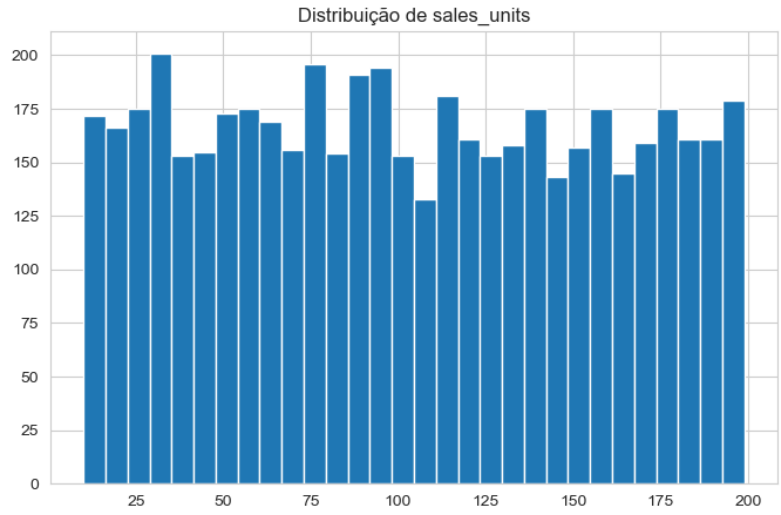

Figura 1: Distribuição de unidades vendidas

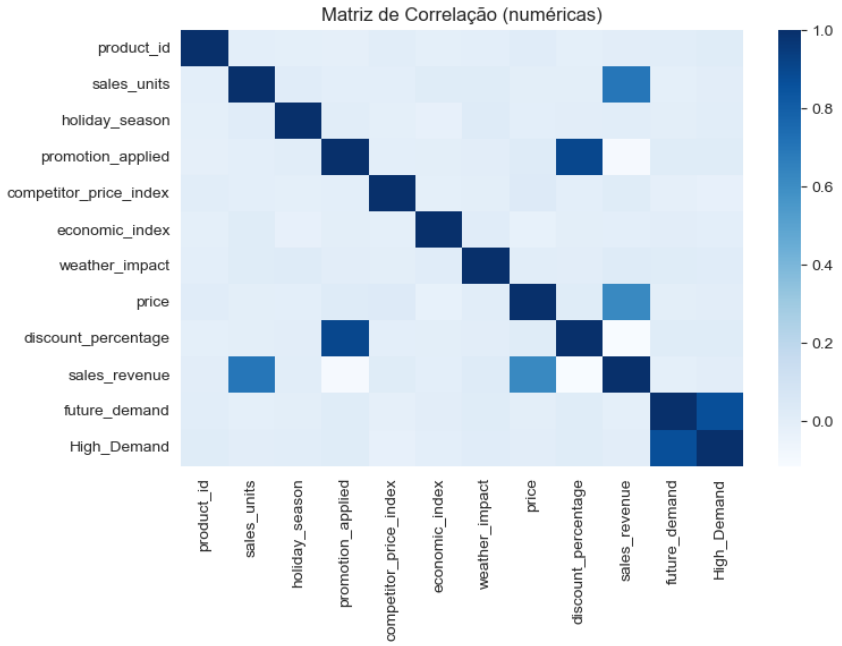

Figura 2: Mapa de correlação entre variáveis numéricas

6 Modelos Utilizados

6.1 Naive Bayes

O modelo Naive Bayes assume independência entre os atributos e distribuição normal. Após treinamento, apresentou acurácia de cerca de 50%, classificando todas as instâncias como classe 1 (alta demanda). Isso mostra uma limitação prática desse modelo em contextos mais complexos.

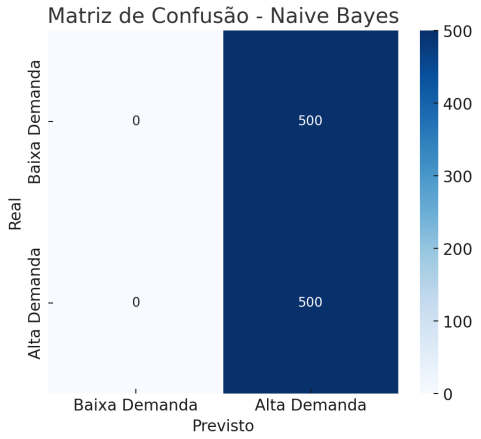

Figura 3: Matriz de confusão - Naive Bayes

6.2 Random Forest

A floresta aleatória é composta por várias árvores de decisão e foi treinada com 300 estimadores. Apresentou resultados perfeitos no conjunto de teste.

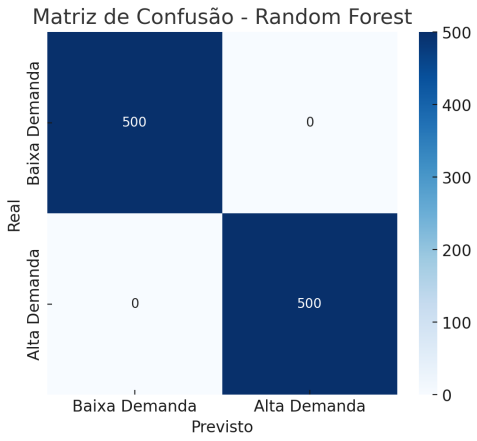

Figura 4: Matriz de confusão - Random Forest

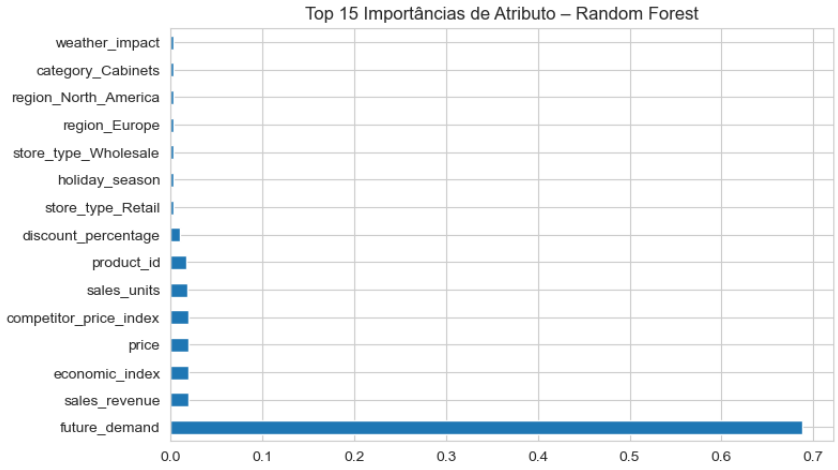

Figura 5: Importância das variáveis - Random Forest

7 Discussão dos Resultados

O Naive Bayes apresentou baixa eficácia por não conseguir separar corretamente as classes. Já o Random Forest demonstrou excelente desempenho e capacidade de generalização, indicando que consegue capturar interações e padrões relevantes nos dados.

A análise da importância das variáveis revelou que `sales_units`, `discount_percentage` e `economic_index` são os atributos mais impactantes, o que faz sentido no contexto comercial.

8 Conclusão

A tarefa de previsão de demanda foi conduzida com sucesso usando técnicas de ciência de dados. A comparação entre os modelos mostrou que o Random Forest é mais apropriado para esse tipo de dado. A metodologia aplicada respeitou o fluxo típico de um projeto de machine learning: entendimento dos dados, tratamento, análise, modelagem e interpretação.

Este estudo evidencia como ferramentas estatísticas e computacionais podem ser aliadas no processo decisório das empresas, aumentando eficiência, reduzindo custos e melhorando o atendimento ao mercado.