Centro Universitário de Excelência
Sistemas de Informação

Implementação Decision Tree Modelo de Classificação

Autores: João Vitor da Conceição de Almeida
Deivid Souza dos Santos Oliveira Flávio
André Almeida Gomes Neto

Feira de Santana, 2025

Agenda

- O objetivo dessa apresentação é o resultado da implementação e desenvolvimento do modelo de classificação basedo no algoritimo Decision tree.

1. Contextualização:
Cenário e Dor do Cliente.

2. Solução:
Classificação do Pedidos (Entregue / Cancelado).

3. Base de Dados:
Caracteristicas e Pré-Processamento.

4. Algoritimo:
O que é, e como funciona uma Decision Tree.

5. Treinamento:
Estratégias e Métricas utilizadas.

6. Resultados:
O que foi Obtido e qual o seu valor.

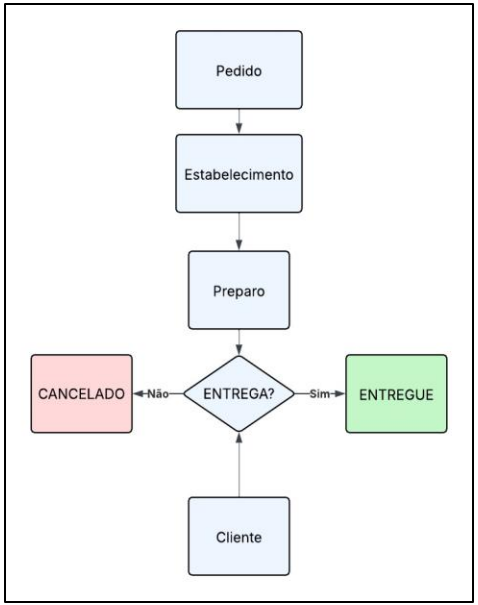
7. Considerações Finais:
Conclusões e Referências.

Contextualização

- Estabelecimentos parceiros sofrem com desistência de pedidos, absorvem os custos dos pedidos cancelados.
- Um restaurante recebe 100 pedidos/dia. Alguns pedidos não são entregues por problemas logísticos, atraso de entregadores, falha no sistema de pagamento ou devolução pelo cliente.
- “Como prever se um pedido será entregue ou cancelado?”

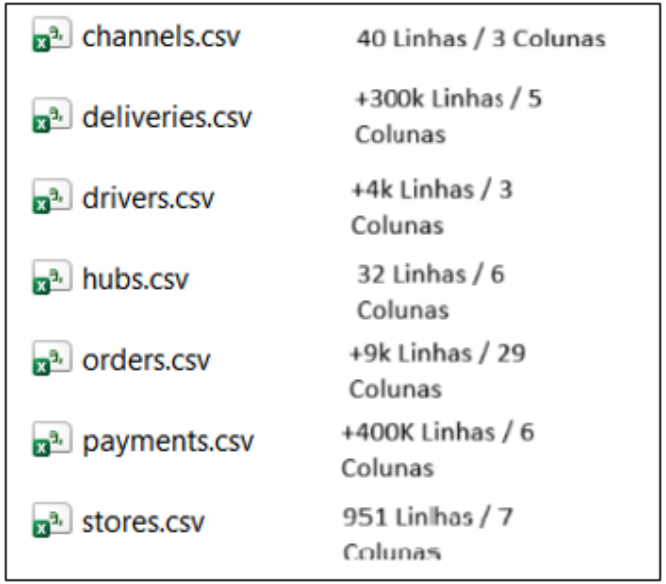
Solução

- Classificar pedidos em **ENTREGUE** ou **CANCELADO**.
- Importância: antecipar cancelamentos → ações preventivas.
- Vários algoritmos possíveis (Naive Bayes, SVM, Decision Tree).
- O algoritmo utilizado para essa solução é o Decision Tree.



Base de Dados

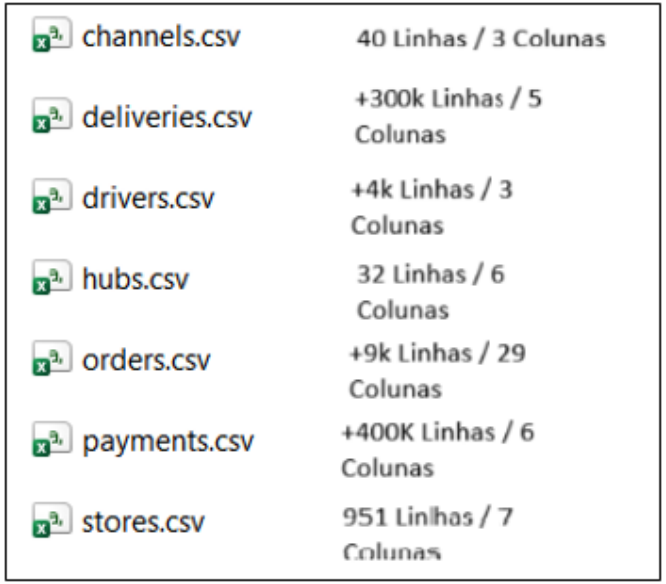
- Origem: 7 arquivos → channels, deliveries, drivers, hubs, orders, payments e stores.
- Total: ~59 características após integração.
- Tipos de atributos: numéricos discretos(order_created_year), numéricos contínuos(order_amount), categóricos nominais(hub_city), categoricos ordinais (hub_latitude).



Base de Dados

Problemas encontrados:

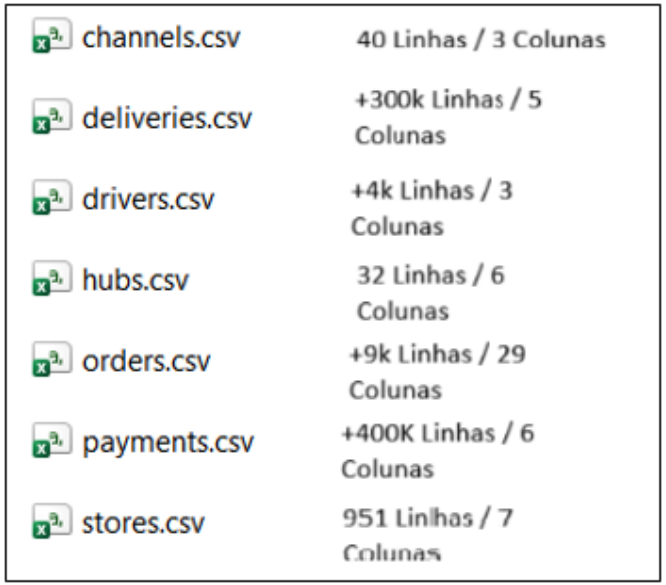
- Dados ausentes (ex.: tempo estimado não informado).
- Atributos únicos ou IDs sem valor preditivo.
- Necessidade de transformação (One-Hot para nominais, Ordinal Encoding para ordinais).



Base de Dados

Pré-processamento aplicado:

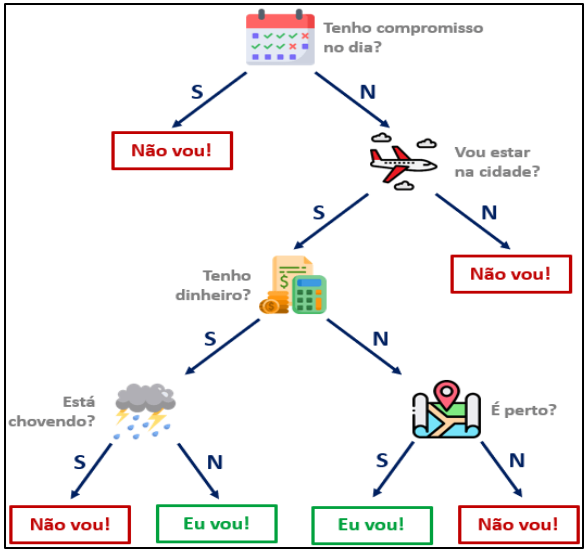
- Imputação de valores ausentes.
- Codificação adequada para cada tipo de variável.
- Seleção de atributos relevantes.



Algoritmo

O que é uma Decision Tree:

- Modelo baseado em **árvore de decisões.**
- Ideal para problemas de classificação binária.
- Apresenta raiz, nós internos, ramos e folhas.



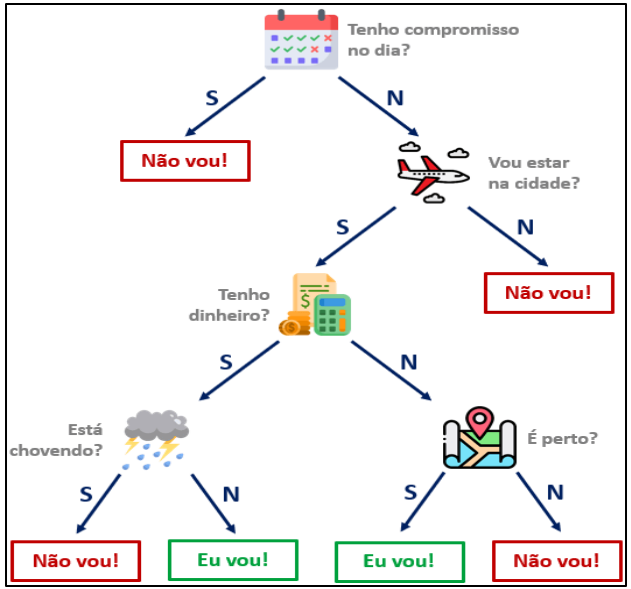
Algoritmo

Funcionamento: divide dados em regras “se-então” até chegar a uma classificação.

Critério de divisão: Gini (Pureza) ou Entropia (Quantidade) .

Parâmetros principais:

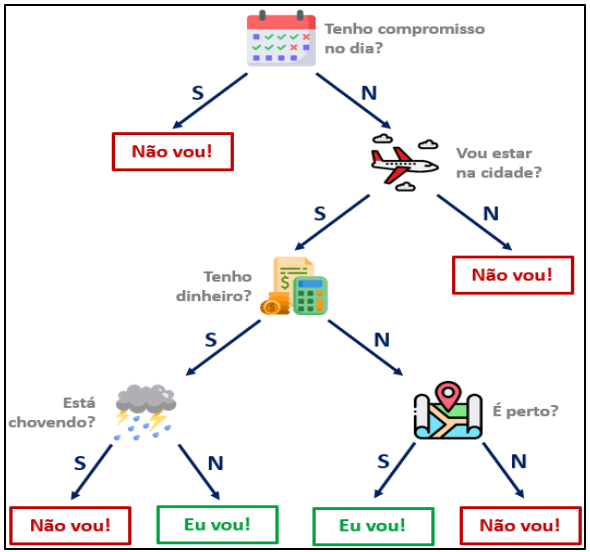
- max_depth → controla profundidade da árvore.
- min_samples_leaf → tamanho mínimo de folhas.
- class_weight → balancear classes desiguais.



Algoritmo

Restrições:

- Pode sofrer overfitting sem limitação de profundidade.
- Sensível a variáveis irrelevantes → necessidade de seleção de features.
- Funciona bem com dados categóricos codificados.



Algoritmo

Gini:

- Mede o quão “puro” ou “misturado” está um conjunto de dados em relação às classes.
- Se todas as amostras de um nó pertencem à mesma classe → **Gini = 0** (nó puro).
- Quanto mais misturado entre classes → maior o valor do Gini

$$Gini = 1 - \sum_{i=1}^{k} p_i^2$$

Onde: pip_i = proporção de elementos da classe ii no nó.

Algoritmo

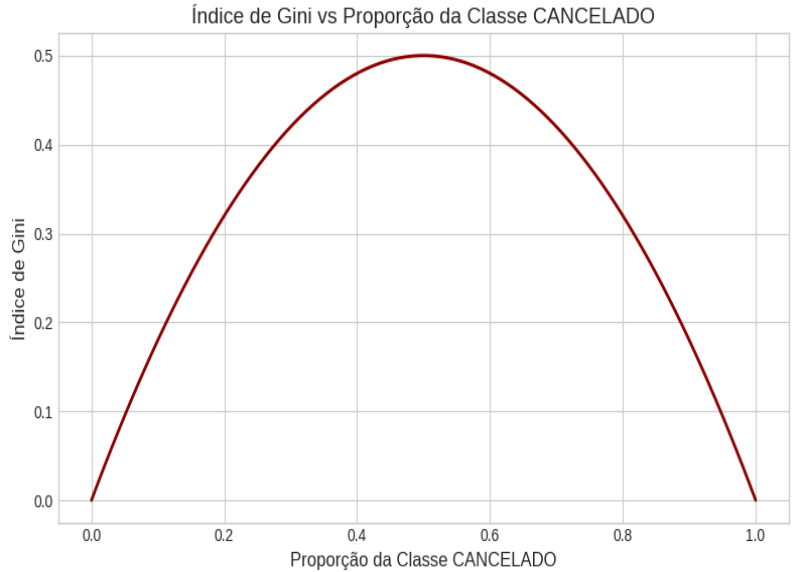
Exemplo:

•Temos um nó com **10 pedidos**:
6 **ENTREGUES**
4 **CANCELADOS**

•Proporções:
pENTREGUE=0.6p_{ENTREGUE} = 0.6
pCANCELADO=0.4p_{CANCELADO} = 0.4

•Fórmula:
Gini=1−(pENTREGUE2+pCANCELADO2)Gini = 1 - (p_{ENTREGUE}^2 + p_{CANCELADO}^2)

•Cálculo:
Gini=1−(0.36+0.16)=0.48



Treinamento

Técnica: **Stratified K-Fold Cross-Validation**.

- Manter proporção de ENTREGUE/CANCELADO em cada fold.

Métricas utilizadas:

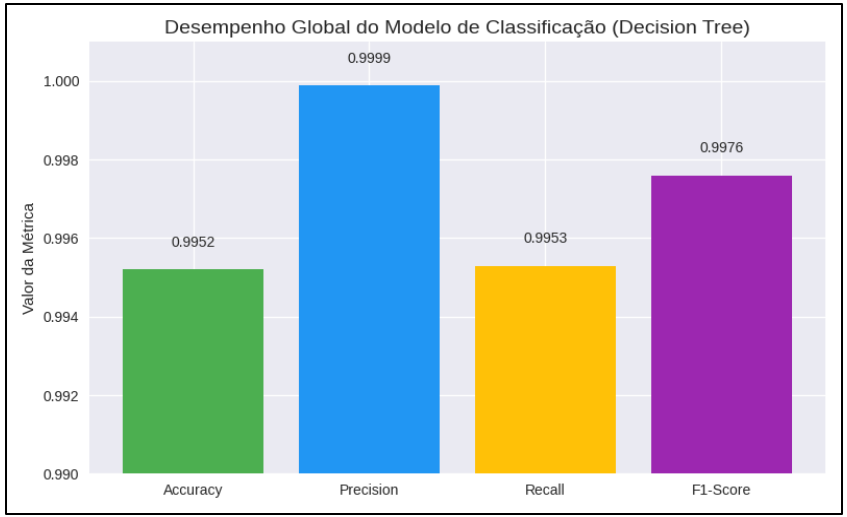
- Precisão (Precision): entre os pedidos previstos como CANCELADOS, quantos realmente foram cancelados.
- Revocação (Recall): entre os pedidos realmente CANCELADOS, quantos foram corretamente previstos.
- F1-Score: equilíbrio entre precisão e recall.

Resultados

- Valores médios de precisão, recall e F1 nos folds.
- Matriz de confusão → erros mais comuns.
- Importância das variáveis → principais fatores que influenciam cancelamentos (ex.: tempo estimado, distância, canal de venda).

Ganhos:

- Antecipação de cancelamentos.
- Possibilidade de ações preventivas.
- Redução de custos e aumento da satisfação do cliente.



Implementação Decision Tree – Modelo de Classificação

Resultados

- Metricas Globais

Métrica	Valor
Accuracy	0.9952
Precision	0.9999
Recall	0.9953
F1-Score	0.9976

Resultados

- Classification Report por Classe

Classe	Precision	Recall	F1-Score	Support
ENTREGUE ⓪	0.78	0.99	0.88	1,453
CANCELADO ①	1.00	1.00	1.00	84,736
Accuracy			1.00	86,189
Macro Avg	0.89	0.99	0.94	86,189
Weighted Avg	1.00	1.00	1.00	86,189

Implementação Decision Tree – Modelo de Classificação

Resultados

- Matriz de Confusão

	Previsto ENTREGUE	Previsto CANCELADO
Real ENTREGUE ⓪	1,443	10
Real CANCELADO ①	402	84,334

Resultados

- AUC (Curva ROC)

Métrica	Valor
AUC	0.9942

Resultados

- Cross-Validation (5 folds)

Métrica	Média	Desvio Padrão
F1-Score	0.9936	0.0023

Considerações finais

- Modelo ajuda restaurantes a reduzir cancelamentos e custos, melhorando a experiência do cliente.
- Previsibilidade, otimização logística, maior satisfação.
- Dependência da qualidade dos dados e risco de overfitting.
- Decision Tree é eficaz, interpretável e abre caminho para soluções mais robustas no futuro.