

RELATÓRIO DE CLASSIFICAÇÃO

Monte Carmelo, 13 de fevereiro de 2026

Por:

Kensley Alves de Oliveira – 32411BSI037

Pedro Paulo Cunha

1. Introdução

Esta análise técnica detalha o desempenho dos modelos de Árvore de Decisão e K-NN para a classificação de risco de crédito, fundamentada nos relatórios de desempenho, matrizes de confusão e Curva ROC fornecidos. A classificação busca encontrar uma função que mapeie os atributos preditores ao atributo classe (*TARGET*).

2. Análise Comparativa do Relatório de Desempenho

O relatório de desempenho permite avaliar a eficácia dos modelos em identificar proponentes adimplentes (Classe 0) e inadimplentes (Classe 1). Os resultados apresentados pelos algoritmos supervisionados K-NN e Árvore de Decisão foram sumarizados nas tabelas 01 e 02.

Tabela 01: Resultado de classificação pelo algoritmo K-NN

Classe de cliente	Precisão	Recall	F1-Score	Suporte
0	0,93	0,67	0,78	45904
1	0,09	0,37	0,14	3906
Acurácia	-	-	0,65	49810
Média (macro)	0,51	0,52	0,46	49810
Média (peso)	0,86	0,65	0,73	49810

Tabela 02: Resultado de classificação pelo algoritmo Árvore de Decisão

Classe de cliente	Precisão	Recall	F1-Score	Suporte
0	0,94	0,70	0,80	45904
1	0,12	0,49	0,20	3906
Acurácia	-	-	0,69	49810
Média (macro)	0,53	0,60	0,50	49810
Média (peso)	0,88	0,69	0,76	49810

RELATÓRIO DE CLASSIFICAÇÃO

A Árvore de Decisão superou o K-NN em todas as métricas críticas. O *Recall* superior (0,49 vs 0,37) indica que a Árvore é mais eficaz em capturar o risco, embora ambos os modelos enfrentem desafios com a baixa *Precision*, comum em problemas de crédito altamente desbalanceados.

Tabela 03: Análise Comparativa do Relatório de Desempenho dos algoritmos

Métrica	Importância e Influência	Árvore de Decisão (Classe 1)	K-NN (Classe 1)	Melhor Desempenho
Precision	Proporção de acertos entre as previsões positivas. Indica o custo de "falsos alarmes" (negar crédito a bons pagadores).	0,13	0,09	Árvore
Recall	Capacidade de identificar todos os inadimplentes reais. Essencial para minimizar o prejuízo financeiro.	0,53	0,37	Árvore
F1-Score	Média harmônica entre <i>precision</i> e <i>recall</i> . Ideal para bases desbalanceadas como esta.	0,20	0,14	Árvore
Support	Número real de ocorrências da classe no conjunto de teste. Indica a base estatística da métrica.	3.906	3.906	Empate

Conforme resultados sumarizados na tabela 03, a Árvore de Decisão mostrou-se superior. Apresentou consistência tanto para prever bons pagadores (Classe 0) quanto inadimplentes (Classe 1).

3. Análise da Matriz de Confusão

A matriz de confusão traduz os erros e acertos em termos de volume e percentuais, refletindo a capacidade de generalização do modelo (Figura 01).

RELATÓRIO DE CLASSIFICAÇÃO

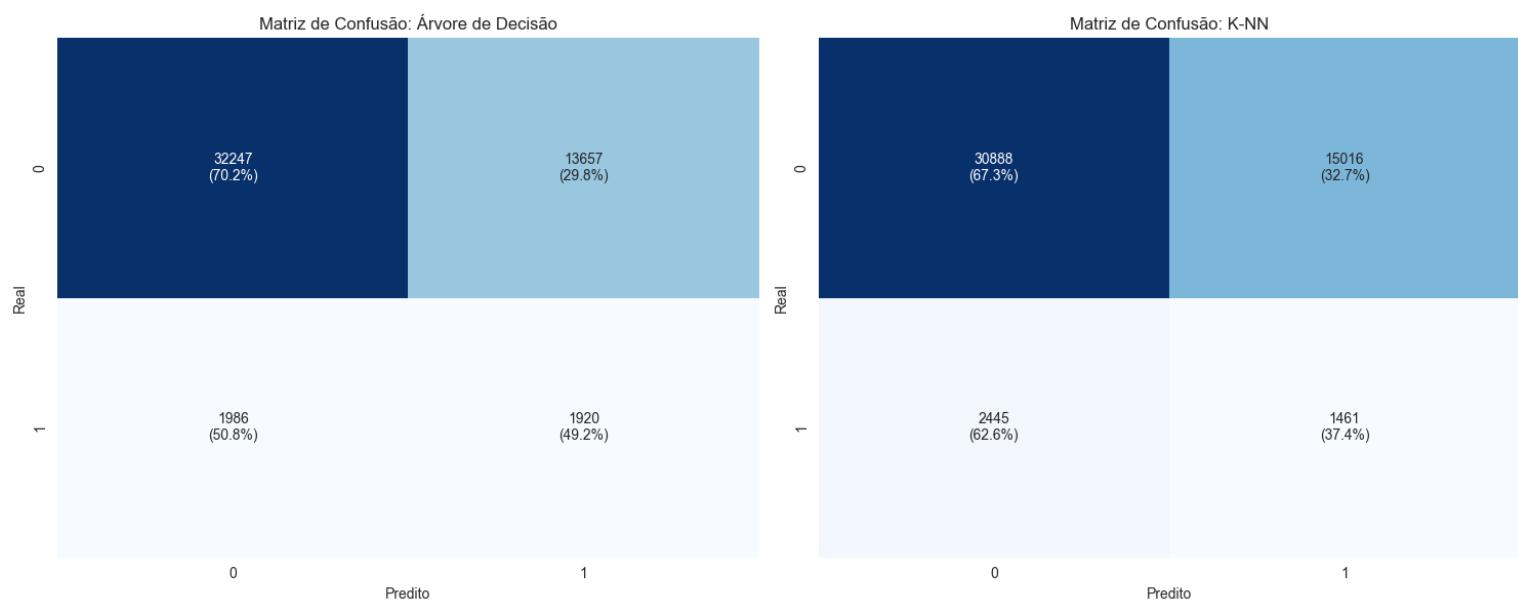


FIGURA 01: Matriz de Confusão para os algoritmos de Árvore de Decisão e K-NN

A análise da Figura 01 mostrou que:

- ✓ **Verdadeiros Positivos (Classe 1 - Inadimplentes):** A Árvore de Decisão identificou corretamente 1.920 (49,2 %) dos inadimplentes, enquanto o K-NN identificou apenas 1.461 (37,4 %);
- ✓ **Falsos Negativos (Risco Crítico):** O K-NN deixou passar 62,6 % dos inadimplentes como se fossem bons pagadores, gerando um risco financeiro significativamente maior que a Árvore (50,8 %).
- ✓ **Falsos Positivos (Custo de Oportunidade):** Ambos os modelos apresentam taxas de erro elevadas para a bons pagadores (29,8 % para Árvore e 32,7% para K-NN), o que resultaria em muitas recusas indevidas de crédito.

4. Análise da Curva ROC e AUC

A Curva ROC e a área sob a curva (AUC) medem o poder discriminatório do modelo, ou seja, sua habilidade em separar as classes independentemente do ponto de corte.

RELATÓRIO DE CLASSIFICAÇÃO

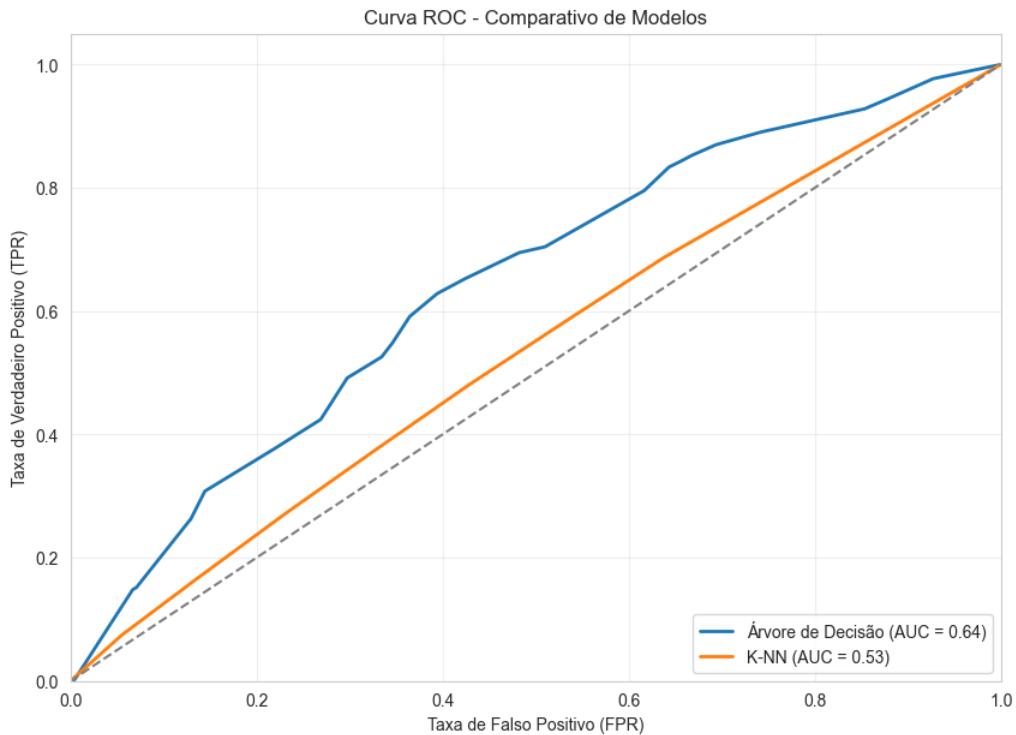


FIGURA 02: Gráfico de *Receiver Operating Characteristics* (ROC)

Da Figura 02 pode-se extrair o seguinte:

- ✓ **Árvore de Decisão (AUC = 0.64):** Demonstra uma capacidade moderada de separação. O gráfico mostra que a curva da Árvore permanece consistentemente acima da linha do K-NN.
- ✓ **K-NN (AUC = 0.53):** Apresenta um desempenho próximo ao acaso (diagonal pontilhada, AUC = 0.50), indicando que o modelo tem dificuldade extrema em distinguir entre bons e maus pagadores nesta base de dados.

5. Validação Cruzada

A análise da validação cruzada aponta que a Árvore de Decisão com Profundidade Limitada foi a melhor estratégia para evitar o *overfitting* e garantir estabilidade. A Validação Cruzada Estratificada confirmou que os resultados (Tabela 04) não são aleatórios e que a integração das bases (Serasa e Histórico Interno) forneceu atributos com real poder preditivo. O modelo de Árvore de Decisão deve ser o escolhido para o sistema de produção, preferencialmente, acompanhado de uma análise humana para os casos de "Risco" indicados, dado o alto volume de falsos positivos.

RELATÓRIO DE CLASSIFICAÇÃO

Tabela 04: Resultados da Análise da validação cruzada

Critério	Árvore de decisão	K-NN
AUC Médio	0,7104	0,5363
Desvio Padrão	0,0047	0,0030
	0,7102	0,5392
	0,7109	0,5372
Socre por Fold	0,7181	0,5321
	0,7034	0,5334
	0,7092	0,5393
Veredito da Validação Cruzada		
Árvore de Decisão: AUC de 0,71 com estabilidade Alta.		
K-NN: AUC de 0,54 com estabilidade Alta		

6. Simulação de Crédito

A simulação de crédito (registro Apêndice 01) apresentada, resultando em "NEGADO" com uma probabilidade de inadimplência de 55,47%, reflete a aplicação prática dos modelos preditivos desenvolvidos.

O veredito NEGADO é a escolha mais segura para a instituição financeira sob a ótica da Árvore de Decisão, dado o seu AUC de 0,71. No entanto, do ponto de vista de negócio, o valor de 55,47% indica um cliente que poderia ser reavaliado manualmente, considerando que o modelo tende a ser conservador, devido ao alto volume de falsos positivos gerados para proteger o capital contra a inadimplência real.

7. Conclusão Sugerida

Os resultados confirmam que a Árvore de Decisão é o algoritmo mais adequado para este problema. Sua natureza não-paramétrica permite lidar melhor com os outliers mantidos na base, que representam sinais vitais de risco. Conforme McKinney (2018), o K-NN é altamente sensível à escala e densidade, o que explica sua dificuldade em uma base com forte desbalanceamento e presença de valores extremos.

RELATÓRIO DE CLASSIFICAÇÃO

REFERÊNCIA BIBLIOGRÁFICA

- FAWCETT, Tom; PROVOST, Foster. **Data Science para Negócios**. Rio de Janeiro: Alta Books, 2016.
- GÉRON, Aurélien. Mão à Obra: **Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow**. 2. ed. Rio de Janeiro: Alta Books, 2021.
- MCKINNEY, Wes. **Python para Análise de Dados**. Rio de Janeiro: Novatec, 2018.
- TAN, Pang-Ning et al. **Introdução ao Data Mining Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna, 2009.
- DE CASTRO, Leandro N. e Ferrari, Daniel G. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. São Paulo: Saraiva, 2016.

Apêndice 01

--- EXEMPLO DE USO DA SIMULAÇÃO ---

RELATÓRIO DE CLASSIFICAÇÃO

```
# Cliente fictício (baseado nos atributos integrados)
cliente_teste = {

    # --- DADOS DEMOGRÁFICOS E DA PROPOSTA (emprestimos.csv) ---
    'AMT_INCOME_TOTAL': 3500.00,
    'AMT_CREDIT': 20000.00,
    'NAME_EDUCATION_TYPE': 'Higher education',
    'DAYS_BIRTH': -12500, # Idade em dias (ex: ~34 anos)
    'CODE_GENDER': 'F',
    'FLAG_OWN_CAR': 'N',
    'FLAG_OWN_REALTY': 'Y',
    'CNT_CHILDREN': 1,
    'NAME_INCOME_TYPE': 'Working',
    'NAME_FAMILY_STATUS': 'Married',
    'NAME_HOUSING_TYPE': 'House / apartment',
    'FLAG_MOBIL': 1,
    'FLAG_EMP_PHONE': 1,
    'FLAG_EMAIL': 0,
    'OCCUPATION_TYPE': 'Core staff',
    'REGION_RATING_CLIENT': 2,
    'ORGANIZATION_TYPE': 'Business Entity Type 3',
    'EXT_SOURCE_1': 0.5,
    'EXT_SOURCE_2': 0.62,
    'EXT_SOURCE_3': 0.48,


    # --- HISTÓRICO NO MERCADO (Agregados do serasa.csv) ---
    'Qtde_emprestimos': 5,      # Contagem de SK_ID_BUREAU
    'Divida_atrasada': 250.00,   # Soma de AMT_CREDIT_SUM_OVERDUE
    'Creditos_Aктивы': 2,       # Contagem de registros 'Active'
    'Media_Dias_Credito': 450.0, # Média de DAYS_CREDIT
    'Max_Dias_Atraso': 15,      # Máximo de CREDIT_DAY_OVERDUE
    'Soma_Max_Atrasado': 300.0,  # Soma de AMT_CREDIT_MAX_OVERDUE
    'AMT_CREDIT_SUM': 50000.0,   # Valor total de crédito no mercado
    'AMT_CREDIT_SUM_DEBT': 12000.0, # Valor atual da dívida no mercado}
```

RELATÓRIO DE CLASSIFICAÇÃO

```
'CNT_CREDIT_PROLONG': 0,      # Vezes que prorrogou crédito no bureau

# --- HISTÓRICO INTERNO (Agregados de emprestimos_anteriores.csv) ---
'Total_Contratos_Passados': 3,  # Contagem de SK_ID_PREV
'Taxa_Recusa': 0.15,          # % de status 'Refused' em NAME_CONTRACT_STATUS
'Soma_Solicitado': 18000.0,    # Soma de AMT_APPLICATION
'Soma_Concedido': 15000.0,     # Soma de AMT_CREDIT (anterior)
'Dif_Aprovado': 3000.0,       # Diferença entre solicitado vs aprovado
'Motivo_Rejeicao_Principal': 'LIMIT', # Moda de CODE_REJECT_REASON
'Categoria_Bem_Frequente': 'XNA',   # Moda de NAME_GOODS_CATEGORY
'NAME_CLIENT_TYPE': 'Repeater',    # Tipo de cliente na última solicitação
'NFLAG_INSURED_ON_APPROVAL': 1     # Se solicitou seguro anteriormente
}
```