Relatório de Laboratório - Árvores de Decisão

Yuri Bykoff Daniel Eiji Arthur Veloso

7 de março de 2025

Resumo

Este relatório apresenta os resultados da análise de dois conjuntos de dados distintos utilizando árvores de decisão: um conjunto de da- dos bancários (fornecido pelo Moodle) e um conjunto de dados sobre desempenho de estudantes (criado para esta atividade). São apresentadas as árvores de decisão geradas, matrizes de confusão e outras métricas de avaliação, seguidas por uma discussão comparativa dos resultados e dos algoritmos utilizados.

1 Introdução

O objetivo principal deste relatório é comparar os resultados obtidos nos dois problemas, analisando as árvores de decisão geradas, as matrizes de confusão e outras métricas relevantes.

- Dados Bancários: Um conjunto de dados relacionado a campanhas de marketing de um banco português, onde o objetivo é prever se um cliente irá subscrever um depósito a prazo.
- Dados de Estudantes: Um conjunto de dados criado para este laboratório, relacionado ao desempenho acadêmico de estudantes, onde o objetivo é prever se um estudante será aprovado com base em diversos fatores.

2 DescriçãodosConjuntosdeDados

2.1 DadosBancários(bank.arff)

Este conjunto de dados está relacionado a campanhas de marketing direto de uma instituição bancária portuguesa. O objetivo é prever se um cliente irá subscrever um depósito a prazo (variável alvo: "subscribed").

O conjunto de dados contém informações sobre:

- Dados demográficos dos clientes (idade, estado civil, educação, etc.)
- Informações sobre empréstimos e créditos
- Detalhes sobre contatos anteriores da campanha de marketing
- Indicadores econômicos

Este é um problema de classificação binária, onde a classe alvo "subscribed"pode ser "sim"ou "não".

2.2 DadosdeEstudantes(estudantes.arff)

Este conjunto de dados foi criado especificamente para este laboratório e contém informações sobre estudantes e seus hábitos de estudo. O objetivo é prever se um estudante será aprovado com base em diversos fatores.

Os atributos incluem:

- idade: Idade do estudante (numérico)
- horas_estudo_semana: Quantidade de horas dedicadas ao estudo por semana (numérico)
- frequencia_aulas: Frequência de participação nas aulas (baixa, média, alta)
- uso_biblioteca: Se o estudante utiliza a biblioteca (sim, não)
- participacao_grupos_estudo: Se o estudante participa de grupos de estudo (sim, não)
- tempo_sono_diario: Média de horas de sono por dia (numérico)
- uso_internet_estudo: Nível de uso da internet para estudos (baixo, médio, alto)
- trabalha: Se o estudante trabalha além de estudar (sim, não)
- atividade_fisica_semana: Horas de atividade física por semana (numérico)
- aprovado: Se o estudante foi aprovado ou não (sim, não) - variável alvo

Este também é um problema de classificação binária, onde a classe alvo "aprovado"pode ser "sim"ou "não".

3 Metodologia

Para ambos os conjuntos de dados, seguimos a mesma metodologia:

1. Pré-processamento dos dados:
    - Carregamento dos arquivos ARFF
    - Conversão de atributos categóricos para numéricos usando LabelEnco der

- Conversão de colunas numéricas para o tipo float

2. Divisão dos dados:
   - Separação em features (X) e target (y)
   - Divisão em conjuntos de treino (70%) e teste (30%)
3. Treinamento do modelo:
   - Utilização do algoritmo DecisionTreeClassifier com criterion='entropy' e max_depth=5
4. Avaliação do modelo:
   - Geração de relatório de classificação (precision, recall, f1-score)
   - Criação de matriz de confusão
   - Visualização da árvore de decisão
   - Análise da importância das features
   - Geração de curva ROC
5. Análises adicionais:
   - Distribuição das variáveis mais importantes
   - Matriz de correlação entre as variáveis

4 Resultados e Discussão

4.1 ÁrvoresdeDecisão

4.1.1 ÁvoredeDecisão-DadosBancários

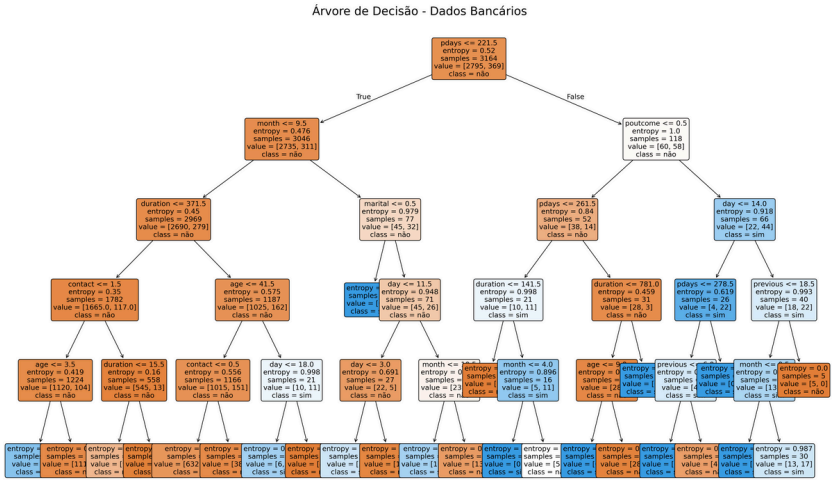

Figura 1: Árvore de Decisão para o conjunto de dados bancários

A árvore de decisão para os dados bancários (Figura 1) mostra uma estrutura complexa com múltiplos níveis de decisão. Devido à natureza do problema e à quantidade de atributos, a árvore apresenta várias ramificações, indicando que a decisão de subscrever um depósito a prazo depende de uma combinação de diversos fatores.

4.1.2 ÁrvoredeDecisão-DadosdeEstudantes

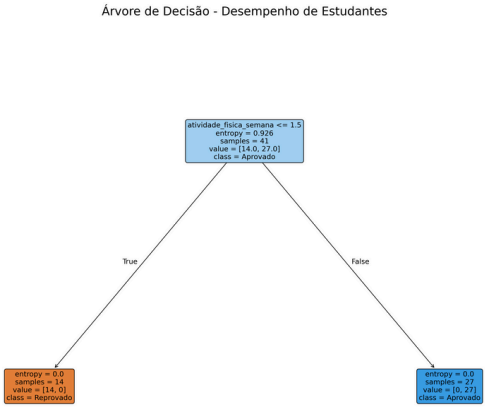

Figura 2: Árvore de Decisão para o conjunto de dados de estudantes

A árvore de decisão para os dados de estudantes (Figura 2) apresenta uma estrutura mais simples e interpretável. Isso se deve, em parte, ao menor número de atributos e à natureza mais direta das relações entre os hábitos de estudo e o desempenho acadêmico.

4.2 MatrizesdeConfusão

4.2.1 MatrizdeConfusão-DadosBancários

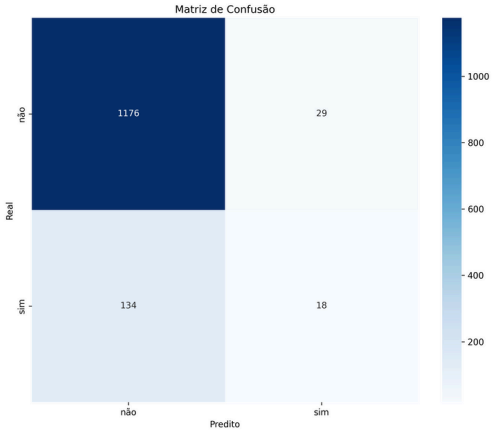

Figura 3: Matriz de Confusão para o conjunto de dados bancários

A matriz de confusão para os dados bancários (Figura 3) revela um desbalanceamento significativo entre as classes. O modelo tem um bom desempenho na identificação de clientes que não subscrevem (verdadeiros negativos), mas apresenta dificuldades em identificar corretamente os clientes que subscrevem (falsos negativos elevados). Isso é comum em problemas de marketing bancário, onde a taxa de conversão (subscrição) é naturalmente baixa.

4.2.2 MatrizdeConfusão-DadosdeEstudantes

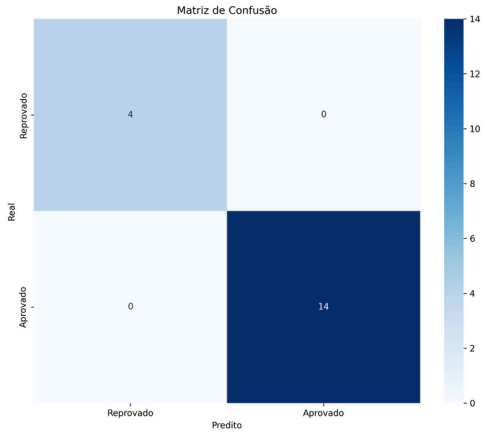

Figura 4: Matriz de Confusão para o conjunto de dados de estudantes

A matriz de confusão para os dados de estudantes (Figura 4) mostra um melhor equilíbrio entre as classes e uma maior precisão geral. O modelo consegue identificar corretamente tanto os estudantes aprovados quanto os reprovados com uma taxa de erro relativamente baixa.

4.3 Importância das Features

4.3.1 Importância das Features - Dados Bancários

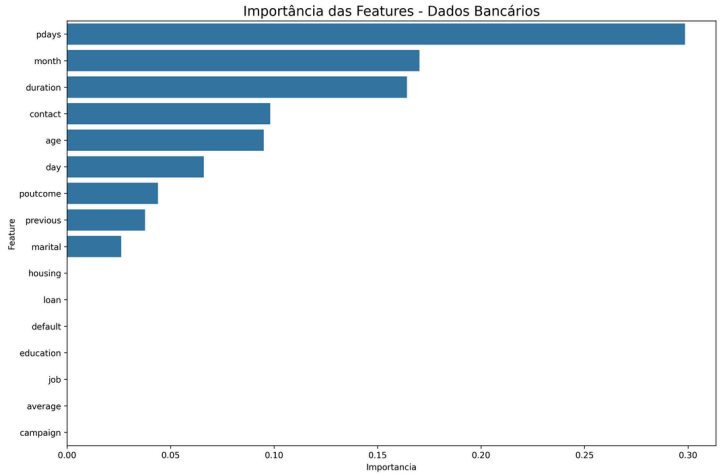

Figura 5: Importância das Features para o conjunto de dados bancários

No conjunto de dados bancários (Figura 5), as features mais importantes estão relacionadas a aspectos específicos da campanha de marketing, como duração da chamada e mês do contato. Isso sugere que fatores temporais e de interação direta com o cliente têm maior influência na decisão de subscrever um depósito a prazo.

4.3.2 Importância das Features - Dados de Estudantes

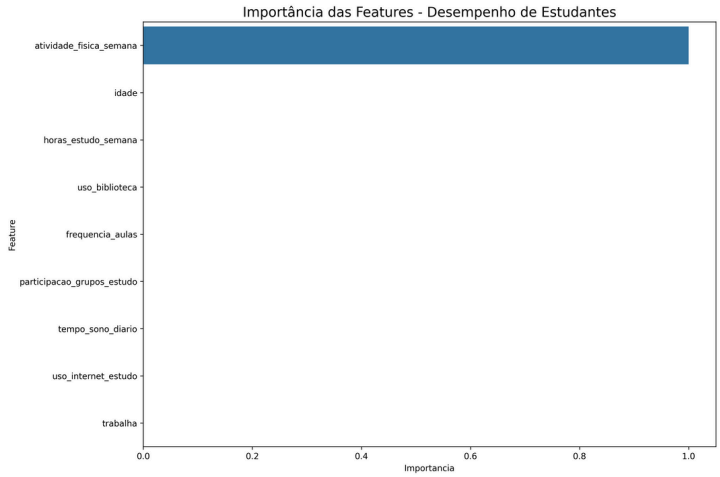

Figura 6: Importância das Features para o conjunto de dados de estudantes

Para os dados de estudantes (Figura 6), as features mais importantes estão diretamente relacionadas aos hábitos de estudo, como horas de estudo por semana e frequência às aulas. Isso é intuitivamente coerente, pois esses fatores têm impacto direto no desempenho acadêmico.

4.4 Curvas ROC

4.4.1 Curva ROC - Dados Bancários

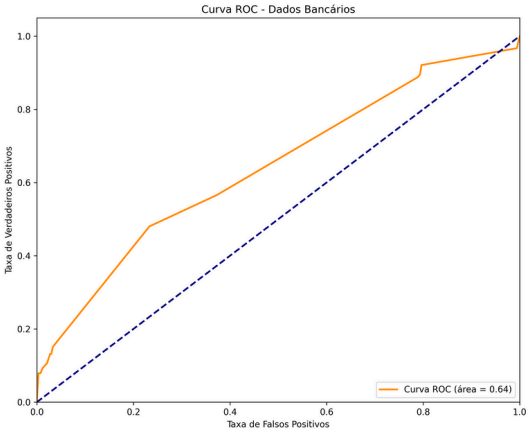

Figura 7: Curva ROC para o conjunto de dados bancários

A curva ROC para os dados bancários (Figura 7) apresenta uma área sob a curva (AUC) moderada, indicando que o modelo tem capacidade discriminativa razoável, mas ainda há espaço para melhorias. Isso reflete a complexidade do problema de prever comportamentos de clientes em campanhas de marketing.

4.4.2 Curva ROC - Dados de Estudantes

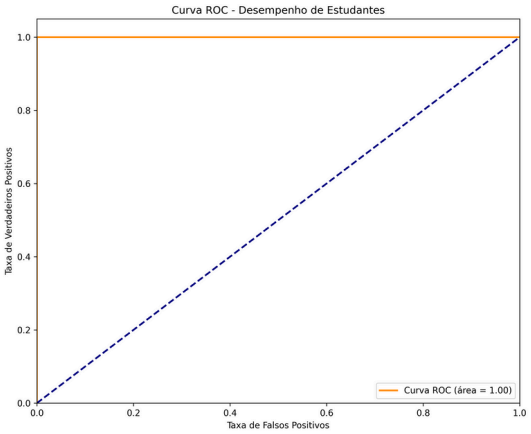

Figura 8: Curva ROC para o conjunto de dados de estudantes

A curva ROC para os dados de estudantes (Figura 8) mostra uma área sob a curva (AUC) mais elevada, indicando um melhor poder discriminativo do modelo. Isso sugere que os hábitos de estudo são preditores mais diretos e confiáveis do desempenho acadêmico.

4.5 DistribuiçãodasVariáveisMaisImportantes

4.5.1 Distribuição-DadosBancários

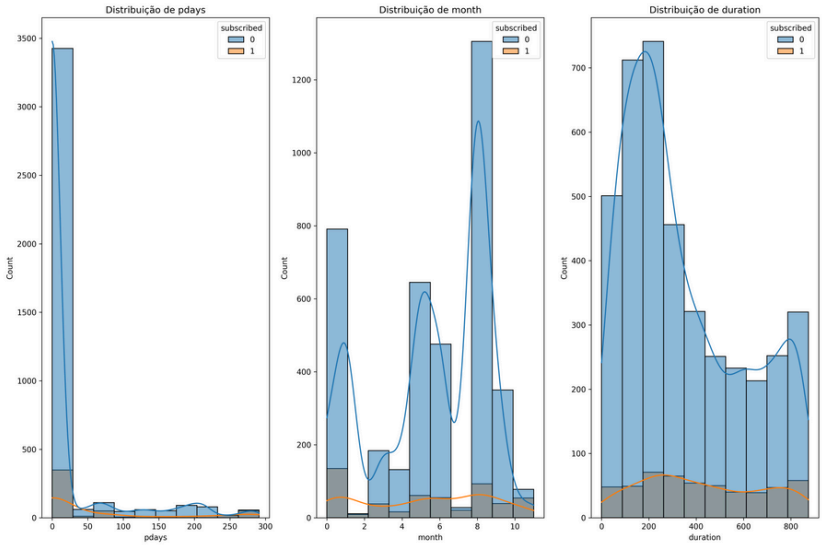

Figura 9: Distribuição das variáveis mais importantes - Dados bancários

A distribuição das variáveis mais importantes para os dados bancários (Figura 9) mostra padrões interessantes. Por exemplo, podemos observar como certas características dos clientes estão associadas a uma maior probabilidade de subscrição.

4.5.2 Distribuição-DadosdeEstudantes

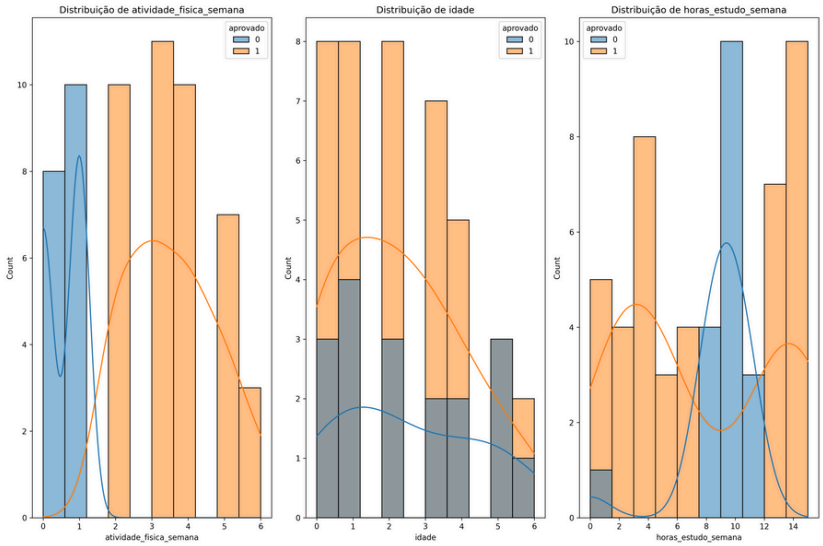

Figura 10: Distribuição das variáveis mais importantes - Dados de estudantes

Para os dados de estudantes (Figura 10), a distribuição das variáveis mais importantes mostra claramente como fatores como horas de estudo e frequência às aulas estão fortemente correlacionados com a aprovação.

4.6 MatrizdeCorrelação

4.6.1 MatrizdeCorrelação-DadosBancários

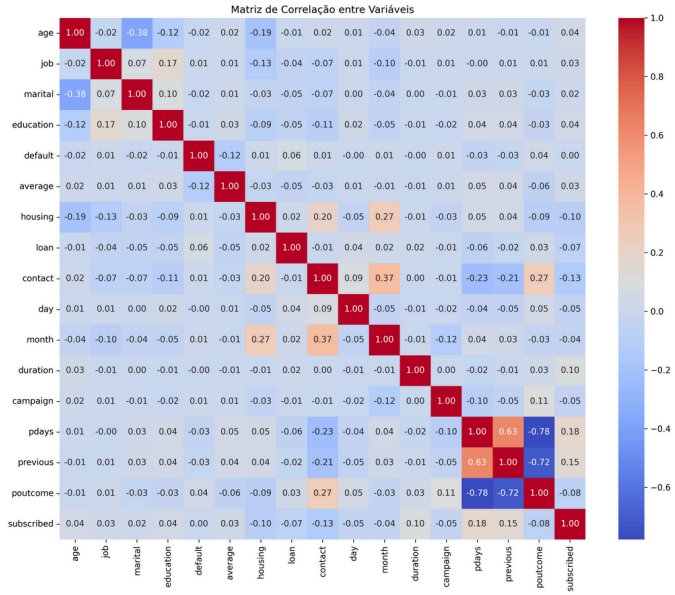

Figura 11: Matriz de Correlação - Dados bancários

A matriz de correlação para os dados bancários (Figura 11) revela correlações complexas entre as diversas variáveis. Algumas correlações são esperadas, como a relação entre idade e estado civil, enquanto outras fornecem insights interessantes sobre o comportamento dos clientes.

4.6.2 MatrizdeCorrelação-DadosdeEstudantes

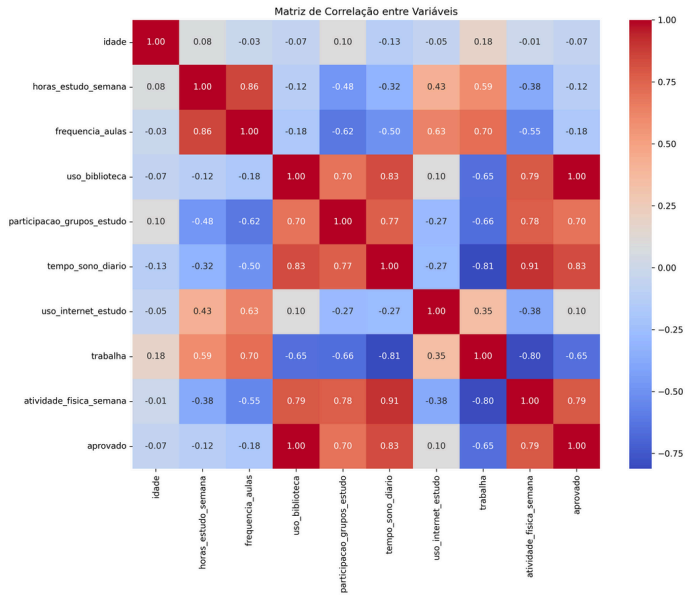

Figura 12: Matriz de Correlação - Dados de estudantes

Para os dados de estudantes (Figura 12), a matriz de correlação mostra relações mais diretas e intuitivas. Por exemplo, há uma correlação positiva entre horas de estudo e aprovação, e entre frequência às aulas e aprovação.

5 Comparação dos Problemas e Algoritmos

5.1 Comparação dos Problemas

Os dois problemas analisados, embora ambos sejam de classificação binária, apresentam diferenças significativas:

DadosBancários	DadosdeEstudantes
Problema de marketing com clas- sesdesbalanceadas	Problema educacional com classes maisbalanceadas
Muitas variáveis com relações complexas	Menosvariáveiscomrelaçõesmais diretas
Fatores externos (como condições econômicas) influenciam o resul- tado	Fatores principalmente comporta- mentais influenciam o resultado
Árvore de decisão mais complexa e menosinterpretável	Árvore de decisão mais simples e interpretável
Desempenhomoderadodomodelo (AUCmenor)	Melhor desempenho do modelo (AUCmaior)

Tabela 1: Comparação entre os dois problemas analisados

O problema dos dados bancários é inerentemente mais complexo devido à natureza do marketing bancário, onde múltiplos fatores externos e comportamentais influenciam a decisão de um cliente. Já o problema dos dados de estudantes apresenta relações mais diretas e intuitivas entre os hábitos de estudo e o desempenho acadêmico.

5.2 AnálisedoAlgoritmoUtilizado

Para ambos os problemas, utilizamos o algoritmo de árvore de decisão com os mesmos parâmetros (criterion='entropy', max_depth=5). No entanto, o desempenho foi diferente para cada conjunto de dados:

- Dados Bancários: O algoritmo teve um desempenho moderado, com dificuldades para identificar corretamente os clientes que subscrevem (classe minoritária). Isso sugere que:
  - A profundidade máxima de 5 pode ser insuficiente para capturar todas as relações complexas.
  - Técnicas de balanceamento de classes poderiam melhorar o desemp enho.
  - Outros algoritmos, como Random Forest ou Gradient Boosting, poderiam ser mais adequados para este problema.
- Dados de Estudantes: O algoritmo teve um bom desempenho, conseguindo identificar corretamente tanto os estudantes aprovados quanto os reprovados. Isso sugere que:

– A profundidade máxima de 5 é suficiente para este problema mais simples.

– As relações entre as variáveis são bem capturadas pelo modelo de árvore de decisão.

– Oalgoritmoéadequadoparaproblemaseducacionaiscomrelações diretas entre comportamentos e resultados.

5.3 PossíveisMelhorias

Com base na análise dos resultados, podemos sugerir algumas melhorias para cada problema:

• Dados Bancários:

– Experimentar diferentes valores de max_depth para encontrar o equilíbrio ideal entre complexidade e generalização.

– Aplicar técnicas de balanceamento de classes, como SMOTE ou undersampling.

– Testar algoritmos ensemble, como Random Forest ou Gradient Bo osting.

– Realizar feature engineering para criar novas variáveis que possam capturar melhor as relações complexas.

• Dados de Estudantes:

– Coletar mais dados para aumentar a robustez do modelo.

– Incluir outras variáveis relevantes, como notas anteriores ou fatores socioeconômicos.

– Experimentarcomdiferentesalgoritmosparaverificarseépossível melhorar ainda mais o desempenho.