

Relatório de Análise de Machine Learning - Dataset Diabetes

Resumo Executivo

Este relatório apresenta uma análise completa de machine learning aplicada ao dataset de diabetes, comparando dois algoritmos de classificação: Decision Tree e Random Forest. O objetivo é desenvolver um modelo capaz de prever a presença de diabetes com base em variáveis clínicas.

Principais Resultados:

- Melhor Modelo:** Decision Tree (max_depth=3) com 75% de acurácia no teste
- Feature Mais Importante:** Dobra cutânea do tríceps (56.85% de importância)
- Desafio Principal:** Dataset desbalanceado (67% não-diabéticos vs 33% diabéticos)

1. Análise Exploratória dos Dados

Características do Dataset:

- Tamanho:** 394 amostras, 6 variáveis
- Variáveis Explicativas:** glicemia, pressão sanguínea, dobra cutânea do tríceps, insulina, IMC
- Variável Alvo:** diabetes (binária: 0 = não, 1 = sim)
- Qualidade:** Dataset completo, sem valores nulos

Distribuição da Variável Alvo:

- Classe 0 (Não diabético):** 264 casos (67.0%)
- Classe 1 (Diabético):** 130 casos (33.0%)
- Status:** Dataset moderadamente desbalanceado

Estatísticas Descritivas:

Variável	Média	Desvio Padrão	Min	Max
Glicemia	70.65	12.47	24	110
Pressão San- guínea	29.11	10.50	7	63
Dobra Cutânea Tríceps	155.89	168.79	7	846
Insulina	31.99	5.18	14.7	43.3
IMC	0.53	0.35	0.085	2.42

2. Metodologia

Divisão dos Dados:

Seguindo as especificações solicitadas:

- **Treino:** 280 amostras (71.1%)
- **Validação:** 94 amostras (23.9%)
- **Teste:** 20 amostras (5.1%)

A estratificação foi aplicada para manter as proporções das classes em todos os conjuntos.

Modelos Implementados:

1. Decision Tree Classifier

- **Parâmetros:** max_depth=3, random_state=42
- **Justificativa:** Profundidade limitada para evitar overfitting e manter interpretabilidade

2. Random Forest Classifier

- **Parâmetros:** max_depth=2, n_estimators=100, random_state=42
- **Justificativa:** Ensemble method com árvores ainda mais rasas para maior generalização

3. Resultados e Performance

3.1 Acurácias por Conjunto

Modelo	Treino	Validação	Teste
Decision Tree	81.07%	67.02%	75.00%
Random Forest	78.57%	62.77%	65.00%

3.2 Análise de Overfitting

Decision Tree:

- Diferença treino-validação: 14.05%
- ⚠ Indica possível overfitting moderado

Random Forest:

- Diferença treino-validação: 15.81%
- ⚠ Indica possível overfitting mais acentuado

3.3 Métricas Completas (Conjunto de Teste)

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC
Decision Tree	0.750	0.667	0.571	0.615	0.731
Random Forest	0.650	0.500	0.143	0.222	0.725

3.4 Análise das Matrizes de Confusão (Teste)

Decision Tree:




- Verdadeiros Negativos: 11
- Falsos Positivos: 2
- Falsos Negativos: 3
- Verdadeiros Positivos: 4

Random Forest:




- Verdadeiros Negativos: 12
- Falsos Positivos: 1
- Falsos Negativos: 6
- Verdadeiros Positivos: 1

3.5 Interpretação das Métricas

Decision Tree:

-  Melhor balance entre precisão e recall
-  F1-Score superior (0.615 vs 0.222)
-  Melhor capacidade de detectar casos positivos (recall = 57.1%)

Random Forest:

-  Muito conservador na predição de casos positivos
-  Recall extremamente baixo (14.3%)
-  Subestima significativamente a presença de diabetes

4. Importância das Features

Ranking por Modelo:

Posição	Feature	Decision Tree	Random Forest
1º	Dobra Cutânea Trí-ceps	56.85%	42.39%
2º	Insulina	23.04%	20.43%
3º	IMC	12.60%	13.21%
4º	Glicemia	7.51%	14.09%
5º	Pressão Sanguínea	0.00%	9.88%

Insights sobre Features:

1. **Dobra Cutânea do Tríceps:** Feature dominante em ambos os modelos
 - Indicador importante de gordura corporal
 - Correlação forte com resistência à insulina

2. **Insulina:** Segunda feature mais importante
 - Diretamente relacionada ao diabetes
 - Consistente entre ambos os modelos
 3. **Pressão Sanguínea:** Não utilizada pela Decision Tree
 - Pode indicar redundância com outras variáveis
 - Random Forest consegue extrair algum valor (9.88%)
-

5. Análise ROC e AUC

Curvas ROC:

- **Decision Tree AUC:** 0.731
- **Random Forest AUC:** 0.725

Ambos os modelos apresentam performance similar em termos de AUC, indicando capacidade discriminativa moderada. A diferença mínima (0.006) sugere que ambos têm potencial similar para distinguir entre classes.

6. Limitações e Considerações

6.1 Limitações do Estudo:

1. **Tamanho da Amostra de Teste:** Apenas 20 casos
 - Pode gerar variabilidade alta nas métricas
 - Recomenda-se validação cruzada para maior robustez
2. **Desbalanceamento das Classes:**
 - 67% vs 33% pode enviesar os modelos
 - Técnicas de balanceamento poderiam melhorar o recall
3. **Overfitting Observado:**
 - Ambos os modelos mostram sinais de overfitting
 - Regularização adicional pode ser necessária

6.2 Qualidade dos Dados:

✅ Pontos Positivos:

- Dataset completo (sem valores nulos)
- Variáveis clinicamente relevantes
- Estratificação adequada

⚠️ Pontos de Atenção:

- Possíveis outliers na dobra cutânea (máximo de 846)
 - Escala muito diferente entre variáveis
 - Normalização poderia beneficiar alguns algoritmos
-

7. Conclusões e Recomendações

7.1 Modelo Recomendado: Decision Tree (max_depth=3)

Justificativas:

1. **Performance Superior:** 75% de acurácia vs 65% do Random Forest
2. **Melhor Balance:** F1-Score de 0.615 vs 0.222
3. **Maior Sensibilidade:** Detecta 57.1% dos casos positivos vs 14.3%
4. **Interpretabilidade:** Estrutura de árvore facilita explicação clínica

7.2 Recomendações Estratégicas:

Curto Prazo:

1. **Implementar o Decision Tree** como modelo de triagem inicial
2. **Validar com dados externos** antes da aplicação clínica
3. **Estabelecer protocolo** de revisão médica para casos positivos

Médio Prazo:

1. **Coletar mais dados** para aumentar o conjunto de teste
2. **Implementar validação cruzada** para métricas mais robustas
3. **Explorar técnicas de balanceamento** (SMOTE, undersampling)

Longo Prazo:

1. **Testar algoritmos avançados** (Gradient Boosting, SVM, Neural Networks)
2. **Incluir novas features** (histórico familiar, idade, etc.)
3. **Desenvolver ensemble personalizado** combinando múltiplos modelos

7.3 Aplicação Prática:

Cenário de Uso Recomendado:

- Ferramenta de **triagem inicial** em unidades básicas de saúde
- **Complemento** à avaliação médica, não substituto
- **Identificação de pacientes** para exames mais detalhados

Protocolo Sugerido:

1. Aplicar o modelo nos dados do paciente
2. Se positivo: encaminhar para avaliação médica detalhada
3. Se negativo: manter acompanhamento de rotina
4. Sempre considerar contexto clínico completo

7.4 Métricas de Sucesso:

Para implementação em produção, monitorar:

- **Acurácia** $\geq 70\%$
 - **Recall** $\geq 50\%$ (para não perder casos positivos)
 - **Precisão** $\geq 60\%$ (para evitar alarmes falsos excessivos)
-

8. Próximos Passos

8.1 Melhorias Técnicas:

1. **Feature Engineering:** Criar variáveis derivadas (ratios, interações)

2. **Normalização:** Aplicar StandardScaler ou MinMaxScaler
3. **Seleção de Features:** Usar técnicas como RFE ou LASSO
4. **Hyperparameter Tuning:** Grid Search ou Random Search

8.2 Validação Adicional:

1. **Validação Cruzada Estratificada** (k=5 ou k=10)
2. **Teste em dados externos** de outras instituições
3. **Análise de estabilidade temporal** dos modelos

8.3 Expansão do Estudo:

1. **Incluir mais variáveis** (demográficas, laboratoriais)
2. **Análise de subgrupos** (idade, sexo, etc.)
3. **Estudo longitudinal** para predição de risco futuro

Anexos

Arquivos Gerados:

- `analise_diabetes_ml.ipynb` - Notebook completo com código
- `metricas_modelos.csv` - Todas as métricas por modelo e conjunto
- `importancia_features.csv` - Importância das features
- `resumo_final_modelos.csv` - Resumo das métricas de teste
- `distribuicao_variaveis.html` - Gráficos de distribuição
- `matrizes_confusao.html` - Matrizes de confusão interativas
- `comparacao_metricas.html` - Comparação visual das métricas
- `curvas_roc.html` - Curvas ROC interativas
- `importancia_features.html` - Gráfico de importância das features

Contato e Suporte:

Para dúvidas sobre a implementação ou interpretação dos resultados, consulte a documentação técnica no notebook Jupyter ou entre em contato com a equipe de Data Science.

Relatório gerado em: 20 de julho de 2025

Versão: 1.0

Autor: Sistema de Análise ML - Abacus.AI