

# Trabalho Final de Introdução ao Aprendizado de Máquina

Aplicando Modelos Supervisionados para Classificação de  
Quitação de Empréstimos

João Pedro Sousa, Milton Salgado e Pedro Saito

Universidade Federal do Rio de Janeiro

2025-07-28



# 1. Introdução

---

Algoritmos de classificação supervisionada aprendem padrões em dados rotulados para classificar novas amostras. Possuem amplas aplicações em dados financeiros, especialmente para gerenciamento de risco em empréstimos bancários.

O Banco Mundial mantém dados históricos de empréstimos concedidos a países em desenvolvimento, incluindo quantias, datas, taxas de juros e tipos de empréstimos. Este trabalho utiliza registros de abril de 2011 a maio de 2025, agregando-os em duas categorias (“quitado” e “não-quitado”) para propor um problema de classificação binária, avaliando diferentes algoritmos de aprendizado supervisionado através de métricas como acurácia e F1-score.

## 2. Pré-processamento

---

**Base de Dados:** *IBRD Statement of Loans and Guarantees - Historical Data do Grupo do Banco Mundial*, (05/25).

O dataset reúne registros mensais desde abril de 2011, contendo empréstimos e garantias concedidos pelo *IBRD* para projetos internacionais.



Os dados incluem tipo de operação, valor contratado, *status* e características contratuais. Para cada dívida, identificou-se o primeiro registro cronológico e seu estado conclusivo, criando a variável alvo `last_loan_status`.

Os registros foram particionados em duas categorias: dívidas totalmente quitadas (Fully Repaid = 1) e dívidas em andamento ou concluídas sem quitação integral (= 0). Foi aplicada amostragem de *Bernoulli* para seleccionar aproximadamente 60% do grupo quitado e 40% do restante.

Status da Dívida	Descrição
<i>Approved</i>	Aprovado pelo Banco Mundial, ainda não assinado.
<i>Signed</i>	Contrato assinado, mas pendente de efetivação.
<i>Effective</i>	Contrado em vigor, habilitado para desembolso.
<i>Disbursing</i>	Valores liberados conforme o andamento do projeto.
<i>Disbursing &amp; Repaying</i>	Parte dos recursos e do desembolso foi liberada.
<i>Fully Disbursed</i>	Todo o valor contratado foi desembolsado.
<i>Repaying</i>	Somente pagamentos em andamento, mas sem desembolsos.
<i>Fully Repaid</i>	Toda a dívida quitada.
<i>Fully Cancelled</i>	Contrato totalmente cancelado.
<i>Fully Transferred</i>	Dívida realocada para outra unidade ou instrumento.
<i>Terminated</i>	Encerrado por rescisão ou fim do prazo contratual.

Figura 1: *Descrição do status de pagamento da dívida.*

## 2.2 Filtragem de Atributos

### **Critérios de Remoção:**

- Baixa Variância (60% de valores idênticos):
- Due 3rd Party, Undisbursed Amount, Exchange Adjustment
- Borrower's Obligation, Sold 3rd Party, Repaid 3rd Party
- Due to IBRD, Loans Held, Currency of Commitment

### **Variáveis de Identificação:**

- Loan Number, Project ID
- Features Relacionadas:
- Guarantor Country removida devido à redundância com Country Code



Foram utilizados dois métodos para calcular correlação com a variável alvo dicotômica:

**Variáveis Contínuas:** Coeficiente de Ponto Bisserial para calcular associação entre variável contínua e dicotômica.

**Variáveis Categóricas:** Coeficiente V de Cramér com teste  $\chi^2$ .

**Critério de Seleção:** Variáveis com magnitude de correlação superior a 0,2.

Variável	Correlação	p-valor	Tipo
loan_status	0.83	0.00	Chi Quadrado (Cramer V)
repaid_percentage	0.77	0.00	Point Bisserial
disbursed_percentage	0.76	0.00	Point Bisserial
loan_type_FSL	0.71	0.00	Chi Quadrado (Cramer V)
agreement_signing_date_timestamp	0.61	0.00	Point Bisserial
effective_date_timestamp	0.57	0.00	Point Bisserial
interest_rate	0.52	0.00	Point Bisserial
country_code	0.42	0.00	Chi Quadrado (Cramer V)

Tabela 1: *Análise de correlação.*

**Avaliação de Redundância:** Correlação de Spearman para pares numéricos, V de Cramér para categóricas e ponto bisserial para contínua-dicotômica.

## 2.4 Features Finais

Após uma análise de redundâncias entre features, selecionamos os seguintes atributos para compor as variáveis de entrada.

- loan\_status
- repaid\_percentage
- agreement\_signing\_date\_timestamp
- interest\_rate — Taxa de juros aplicada ao empréstimo.
- loan\_type (SNGL CRNCY, SCP USD, POOL LOAN, FSL, NON POOL)
- first\_repayment\_date\_timestamp
- original\_principal\_amount

### Codificação de Variáveis:

- One-hot encoding para variáveis categóricas
- Codificação ordinal para loan\_status respeitando progressão temporal
- StandardScaler para padronização de variáveis numéricas

### Validação Cruzada:

- Repeated Stratified K-Fold: 5 folds  $\times$  6 repetições = 30 splits
- Manutenção da proporção de classes em cada fold
- Random state fixo para reprodutibilidade

### Dataset Final:

- 6.945 registros, 11 features
- Distribuição: 39,58% classe 0, 60,42% classe 1

### Principais métricas utilizadas:

- Acurácia.
- F1-Score.

# 3. Experimentos

---

## 3.1 Ambiente Computacional

### Especificações do Sistema:

CPU: Intel(R) Core(TM) i7-10700 *t*@ 2.90GHz

- Sistema: Debian GNU/Linux 12 (bookworm)
- Memória: 94GB
- Cores: 16
- Disco: 907GB

### Garantia de Homogeneidade:

- Todos os experimentos executados no mesmo ambiente
- Controle de variáveis externas
- Reprodutibilidade dos resultados

## 3.2 Hiperparâmetros Otimizados

### SVM:

- Kernel: Radial Basis Function (rbf)
- Parâmetro de regularização:  $C = 100$
- Máximo de iterações: 10.000
- Alcance de influência:  $\gamma = 1, 0$

### Redes Neurais:

- Função de ativação: tangente hiperbólica
- Camadas ocultas: 2 (100, 50 neurônios)
- Máximo de iterações: 500



## 3.2 Hiperparâmetros Otimizados

### Árvore de Decisão:

- Critério: Entropia
- Profundidade máxima: 10
- Amostras mínimas para divisão: 10

*Os demais parâmetros seguem o padrão das implementações do scikit-learn*

Performance dos Modelos Otimizados:

Modelo	Acurácia	Precisão	Recall	F1-score	Tempo
SVM	0.9534	0.9540	0.9534	0.9535	0min 7.73s
Rede Neural	0.9565	0.9567	0.9563	0.9565	3min 51s
Árvore de Decisão	0.9626	0.9626	0.9626	0.9626	0min 0.37s

Figura 2: *Tabela de performance dos modelos com otimização.*

## 4. Discussão dos Resultados

---

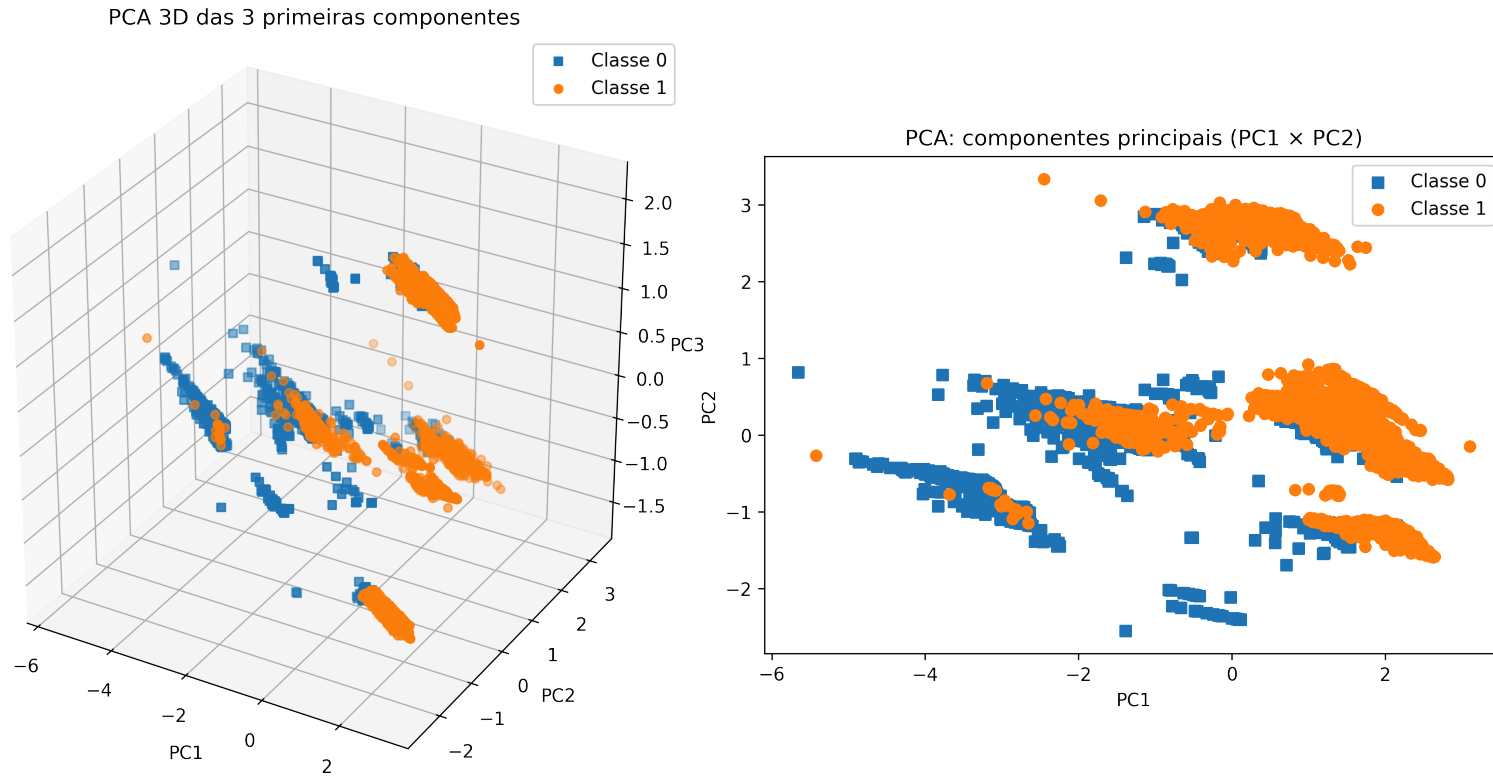


Figura 3: *Análise de duas e três primeiras componentes principais (PCA).*

### **Análise de Separabilidade:**

- Baixo grau de separabilidade entre quitadas (1) e não quitadas (0).
- Classe 1 concentra-se em valores positivos de PC1 e PC2

### **Largura da Margem:**

- $\gamma = \frac{2}{\|w\|} = 0,0002$
- Confirma quantitativamente a estreita região de separação
- Justifica uso da função RBF (não-linear)

### **Limitações:**

- Máximo de 10.000 iterações para evitar treinamento prolongado
- Necessidade de kernel não-linear devido à baixa separabilidade linear

### **Configuração e Convergência:**

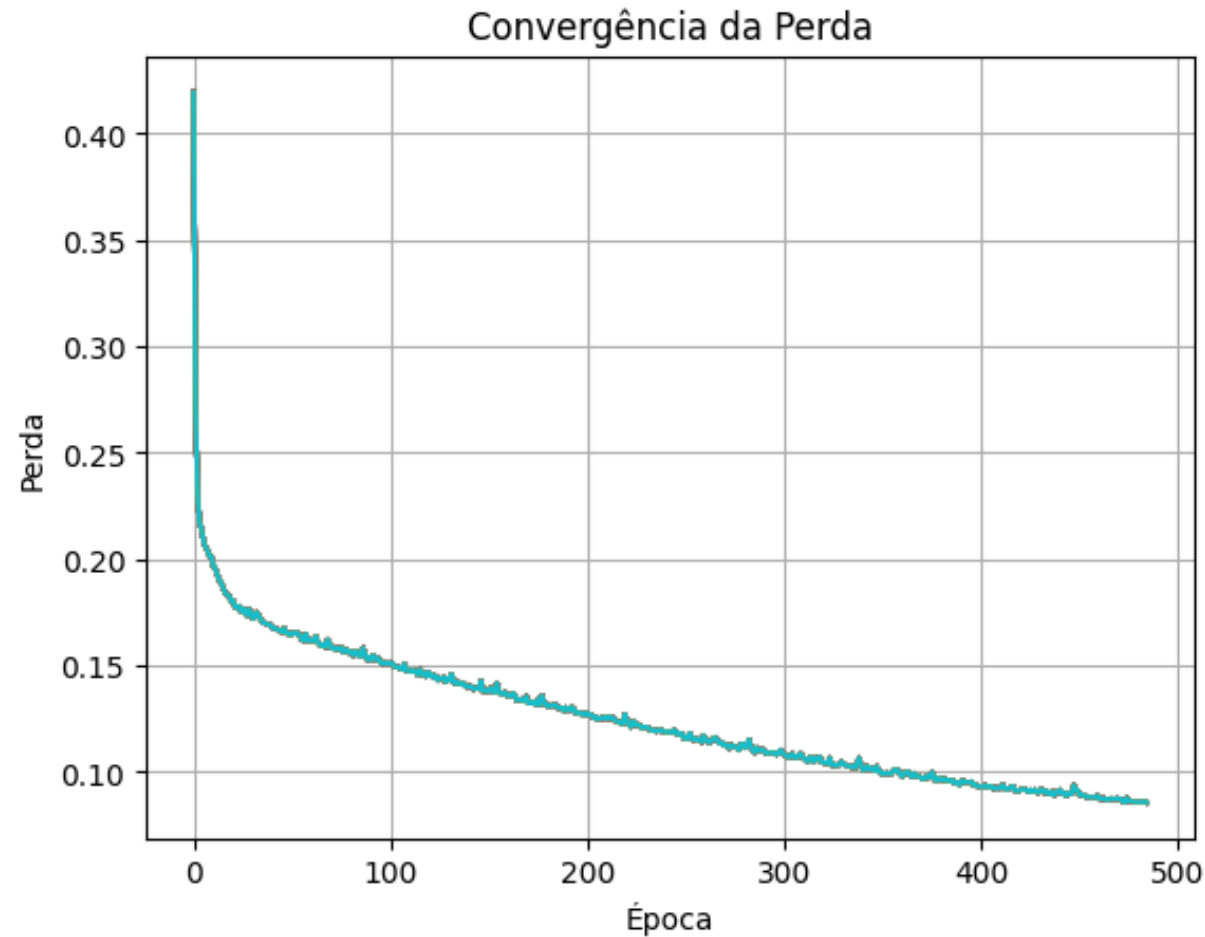
- Média de 485 iterações até convergência
- Função de perda final: aproximadamente 0,0856
- Tangente hiperbólica capturou não-linearidades adequadamente

### **Análise de Performance:**

- Arquitetura (100, 50 neurônios) foi adequada para modelar complexidade
- Proximidade das iterações com o máximo (500)
- Sugere necessidade de mais iterações ou ajuste na taxa de aprendizado

### **Desafios:**

- Múltiplos hiperparâmetros podem causar flutuações no comportamento
- Alto custo computacional comparado aos demais modelos





### Performance Robusta:

- Configuração ótima apareceu em 7 dos 30 folds
- Critério de entropia permitiu divisões mais balanceadas
- Profundidade 10: captura complexidade sem **overfitting** excessivo

### Estabilidade:

- Desvio padrão: apenas 0,0060
- Mínimo 5 amostras: evita regras muito específicas
- Tempo médio: 0,01s por fold (total: 0,33s)

### Eficiência:

- Robustez em Métricas
- Tempo de execução de 0.37s em 30 folds

Teste de Friedman ( $\alpha = 5\%$ ):

Métrica	$\chi^2$	P-valor
Acurácia	16.9402	0,0002
F1-score	16.0678	0,0002

Figura 4: Resultados Teste de Friedman

Matriz de p-valores do teste pareado de Neminyi ( $\alpha = 5\%$ ):

	Acurácia			F1-Score		
	Árvore	SVM	MPL	Árvore	SVM	MPL
Árvore	1.0000	0.0002	0.0125	1.0000	0.0003	0.0184
SVM	0.0002	1.0000	0.5157	0.0003	1.0000	0.4761
MPL	0.0125	0.5157	1.0000	0.0184	0.4761	1.0000

Figura 5: Resultados dos testes de hipótese usando a acurácia e F1-Score.

## 5. Trabalhos Relacionado

---

1. **Dados:** 326,000 empréstimos *IBRD* (1980–2018), sendo 18,000 cancelados.
2. **Pré-processamento:** remoção de colunas com muitos faltantes, imputação de juros por país e criação de métricas temporais
3. **Variáveis:** 7 no total (4 numéricas, 3 categóricas), todas com diferenças estatisticamente significativas entre “repaid” e “cancelled”
4. **Modelos:** *Decision Tree* teve robustez maior que SVC e Gradient Boosting, alcançando acurácia e F1-Score de aproximadamente 0,99.

<i>Decision Tree</i>	<i>Predicted Repaid</i>	<i>Predicted Cancelled</i>
<i>True Repaid</i>	1757	3
<i>True Cancelled</i>	0	1760

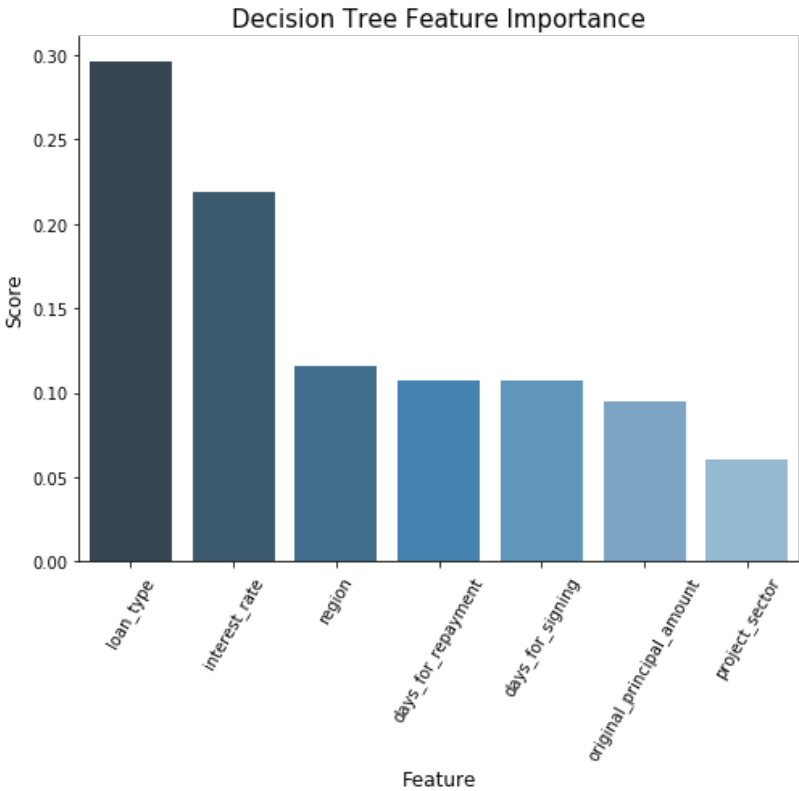


Figura 6: Features ordenadas por relevância.

## 6. Conclusões

---

### Ranking Final:

1. **Árvore de Decisão:** 96,26% (performance robusta + eficiência)
2. **Redes Neurais:** 95,65% (boa performance, alto custo computacional)
3. **SVM:** 95,34% (performance adequada, baixo custo computacional)

**Recomendação:** Árvore de Decisão combina alta acurácia, estabilidade entre folds e eficiência computacional.

### Considerações Finais:

- Todas as diferenças são estatisticamente significativas
- Interpretabilidade da Árvore de Decisão é vantagem adicional
- Tempo de processamento varia drasticamente entre os métodos



# 7. Obrigado!

---

## Referências

- (1) Kornbrot, D. Point Biserial Correlation; 2005; pp 12–13. <https://doi.org/10.1002/0470013192.bsa485>.
- (2) Singhal, R.; Rana, R. Chi-square test and its application in hypothesis testing. *Journal of the Practice of Cardiovascular Sciences* **2015**. <https://doi.org/10.4103/2395-5414.157577>.
- (3) Soleymani, F.; Masnavi, H.; Shateyi, S. Classifying a Lending Portfolio of Loans with Dynamic Updates via a Machine Learning Technique. *Mathematics* **2021**, 9 (1).
- (4) Hamad, R. A.; Kimura, M.; Lundström, J. Efficacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments. *SN Computer Science* **2020**, 1 , 204. <https://doi.org/10.1007/s42979-020-00211-1>.
- (5) James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. *Introduction to Statistical Learning*; Springer, 2023.
- (6) Géron, A. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras e Tensorflow*; O'Rilley, 2021.