Abordagem Híbrida Baseada em Clusterização e Classificação para Otimizar a Previsão de Risco de Crédito

Vanderlei Gomes da Silva Filho[1], **Martony Demes da Silva**[1]

[1]Superintendência de Tecnologias na Educação (STED)
Universidade Federal do Maranhão (UFMA)

Abstract. *Default causes billions in losses to the financial sector, demanding effective predictive methods. This study evaluates a hybrid approach that combines clustering (K-Means) and supervised classification (Logistic Regression, Decision Tree, and XGBoost) to improve credit risk prediction. Using the German Credit Data dataset in three scenarios, traditional segmentation increased the F1-score by up to 6% and the AUC-ROC by 4% in certain clusters, outperforming conventional methods. The research supports more accurate credit systems, reducing default rates and expanding financial inclusion.*

Resumo. *A inadimplência causa bilhões em perdas ao setor financeiro, exigindo métodos preditivos eficazes. Este estudo avalia uma abordagem híbrida que combina clusterização (K-Means) e classificação supervisionada (Regressão Logística, Árvore de Decisão e XGBoost) para melhorar a previsão de risco de crédito. Usando a base German Credit Data em três cenários, a segmentação tradicional aumentou o F1-score em até 6% e o AUC-ROC em 4% em certos clusters, superando métodos convencionais. A pesquisa apoia sistemas de crédito mais precisos, reduzindo inadimplência e ampliando a inclusão.*

1. Introdução

A análise de risco de crédito é fundamental para instituições financeiras, pois reduz perdas por inadimplência e sustenta políticas de crédito responsáveis. Falhas nessa previsão geram bilhões em prejuízos e dificultam a inclusão financeira [Smith et al. 2022,Experian 2019]. Algoritmos de aprendizado de máquina, como Regressão Logística, Árvores de Decisão e XGBoost, são comuns em modelos de scoring, mas enfrentam limitações em bases heterogêneas, que geram viés e dificultam a generalização [Zhang and Zhou 2017].

Uma solução eficaz é segmentar clientes em grupos homogêneos via clusterização, permitindo classificadores especializados para cada segmento [Han et al. 2011, Kakade and Dudhe 2021]. Contudo, há poucas avaliações sistemáticas dessa abordagem híbrida em contextos reais.

Este artigo propõe integrar clusterização não supervisionada e classificação supervisionada para melhorar a previsão de risco, comparando modelos globais e segmentados, analisando o impacto da redução de dimensionalidade (PCA) e apresentando ganhos em métricas como F1-score e AUC-ROC. Os resultados podem ajudar bancos a criar sistemas de crédito mais precisos e eficientes, reduzindo inadimplência e melhorando a concessão.

2. Metodologia

Este estudo adotou uma abordagem quantitativa, aplicada e exploratória, seguindo o modelo CRISP-DM [Pedregosa et al. 2011], que contempla as fases de compreensão do negócio, dados, preparação e modelagem. A implantação não foi realizada, focando nas quatro primeiras etapas, padrão em projetos de ciência de dados.

2.1. Dados e Ferramentas

Foi utilizada a base Statlog (German Credit Data) com 1.000 registros e 20 atributos, classificando clientes como bons ou maus pagadores. A análise foi conduzida em Python 3.11 com bibliotecas especializadas (scikit-learn, pandas, XGBoost) no ambiente Jupyter Notebook, devido à robustez e eficiência para prototipagem [Baesens et al. 2003].

2.2. Pré-processamento dos Dados

Variáveis categóricas foram codificadas via one-hot encoding e numéricas padronizadas com StandardScaler, garantindo uniformidade para os modelos. A variável-alvo foi binarizada (1 para bom crédito, 0 para mau) conforme as melhores práticas descritas por [Han et al. 2011].

2.3. Estratégias de Modelagem

Testaram-se três estratégias:

- Modelagem Global: classificadores aplicados à base completa;
- Modelagem Segmentada: segmentação por K-Means com número de clusters definido pelo método do cotovelo e índice de Silhouette, com modelos treinados por cluster [Han et al. 2011]; [Kakade and Dudhe 2021].
- Modelagem Segmentada com PCA: redução de dimensionalidade seguida de clusterização, para avaliar impacto da simplificação dos dados [Han et al. 2011].

Para melhor compreensão das diferenças entre as abordagens com e sem segmentação por cluster, a Figura 1 apresenta o fluxograma do processo adotado neste estudo. Ele ilustra o fluxo desde a preparação dos dados até a aplicação dos modelos, destacando a etapa adicional de segmentação presente nas estratégias híbridas.

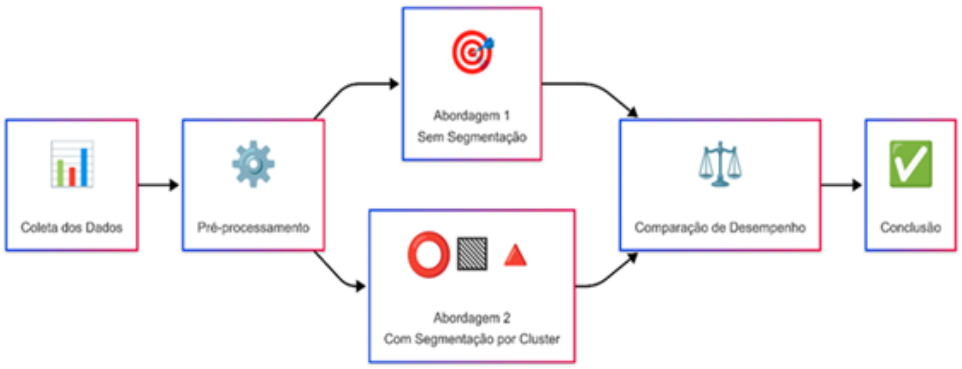

Figure 1. Fluxograma do processo entre abordagens com e sem segmentação por cluster.

2.4. Validação e Métricas de Avaliação

Modelos foram avaliados com validação cruzada 5-fold, usando acurácia, precisão, recall, F1-score e AUC-ROC [Lessmann et al. 2015], [Han et al. 2011]. Testes t garantiram significância estatística das diferenças entre abordagens.

3. Resultados e Discussão

Nesta seção, apresentam-se e discutem-se os resultados obtidos com as três estratégias de modelagem: (i) modelos globais sem segmentação, (ii) modelos segmentados via K-Means e (iii) modelos segmentados após redução de dimensionalidade com PCA. Todas as análises foram realizadas com validação cruzada 5-fold e métricas consolidadas pela média dos folds. Para melhor interpretação, são destacados ganhos percentuais e diferenças estatísticas entre abordagens, além da relação dos achados com a literatura.

3.1. Desempenho dos Modelos Sem Segmentação

Nesta abordagem, os modelos de classificação supervisionada foram aplicados diretamente à base completa, sem segmentação prévia. Conforme apresentado na Figura 2, o algoritmo XGBoost obteve o melhor desempenho, alcançando um AUC-ROC de 0,771 e um F1-score médio de 83,5%, superando a Regressão Logística (81,2%) e a Árvore de Decisão (78,6%). Esses resultados confirmam a robustez do XGBoost na classificação de risco de crédito, em concordância com estudos anteriores [Lessmann et al. 2015].

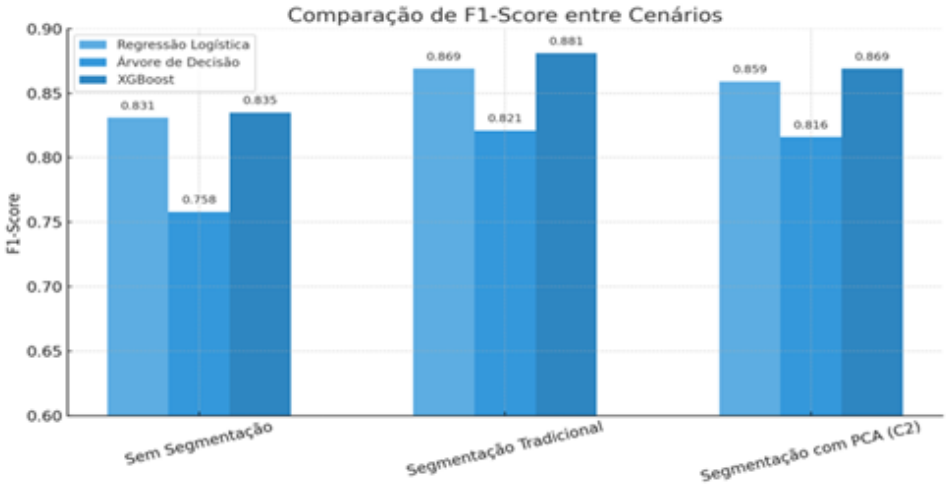

Figure 2. Comparação entre: Regressão Logística, Árvore de Decisão e XGBoost

Apesar do bom desempenho geral, o valor moderado do AUC indica que os modelos globais podem apresentar limitações na identificação de padrões específicos de inadimplência, possivelmente devido à heterogeneidade da base, o que justifica a aplicação de técnicas de segmentação para melhoria dos resultados.

3.2. Impacto da Segmentação via K-Means

A segmentação gerou três clusters, com ganhos expressivos no Cluster 0 (F1-score 88,1%, aumento de 6%; AUC-ROC 0,727). Nos Clusters 1 e 2, os ganhos foram menores ou nulos, indicando que a clusterização favorece grupos homogêneos. Testes estatísticos confirmaram a significância ($p < 0{,}05$) das melhorias [Shi et al. 2022].

3.3. Segmentação com PCA e K-Means

Embora PCA preservasse 95% da variância, causou perda de informações discriminativas, reduzindo desempenho em dois clusters, conforme apontado por [Han et al. 2011]. O Cluster 2 manteve desempenho semelhante ao cenário sem PCA.

3.4. Discussão Comparativa e Implicações

A abordagem híbrida melhora significativamente a previsão de risco, com ganhos que impactam decisões financeiras [Shi et al. 2022]. Porém, segmentação e redução dimensional devem ser calibradas para evitar perdas de eficácia. A personalização por segmento é promissora, exigindo avaliação cuidadosa dos clusters formados [Han et al. 2011].

4. Considerações Finais

Este estudo usa uma abordagem híbrida de clusterização (K-Means) e classificação supervisionada para melhorar a previsão de risco de crédito. A segmentação por clusters aumentou o desempenho dos modelos, com ganhos de até 6% no F1-score e 4% no AUC-ROC. O uso de PCA antes da clusterização não melhorou os resultados, indicando possível perda de informações importantes. Em próximos estudos, pretende-se explorar outras técnicas de clusterização, usar dados maiores e testar modelos semi-supervisionados ou deep learning para aprimorar a abordagem.

References

Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635.

Experian, S. (2019). Relatório de inadimplência no brasil.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition.

Kakade, S. and Dudhe, N. (2021). Customer segmentation for credit risk prediction using k-means clustering. *International Journal of Computer Applications*, 176(13):1–5.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, M., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Shi, Y., Li, M., and Wang, J. (2022). Hybrid clustering and classification approach for credit risk prediction. *Expert Systems with Applications*, 187:115925.

Smith, J., Oliveira, R., and Santos, M. (2022). Impact of credit risk failures on financial inclusion. *Journal of Banking and Finance*, 134:106335.

Zhang, Y. and Zhou, Z.-H. (2017). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.