Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP
IX SAEPRO – 25 e 26 de novembro de 2025

Machine Learning Aplicado à Concessão de Crédito: Comparação de Desempenho entre KNN e Random Forest

Rafael Nonato Lima de Oliveira
Universidade Federal do Vale do São Francisco
(rafael.nonato@discente.univasf.edu.br)

Thiago Magalhães Amaral
Universidade Federal do Vale do São Francisco
(thiago.magalhaes@univasf.edu.br)

CONTEXTUALIZAÇÃO: A concessão de crédito é um dos pilares do sistema financeiro moderno, desempenhando papel essencial no estímulo à economia ao permitir que indivíduos e empresas tenham acesso a recursos para consumo e investimento (MESTER, 1997). Entretanto, o processo de análise de crédito envolve riscos significativos, uma vez que a decisão de aprovar ou negar um empréstimo depende da capacidade de prever, de forma precisa, a probabilidade de inadimplência do solicitante (THOMAS, 2000). A necessidade de modelos preditivos eficazes é urgente, considerando que o volume de crédito no Brasil atingiu aproximadamente R$ 6,4 trilhões em 2024, com crescimento de 11% em relação ao ano anterior, e que a inadimplência média em operações de crédito com atraso superior a 90 dias alcançou 3,0% no total do sistema, sendo 3,5% para pessoas físicas (BANCO CENTRAL DO BRASIL, 2025; AGÊNCIA BRASIL, 2025). Esses números refletem não apenas a importância econômica da concessão de crédito, mas também os riscos financeiros bilionários associados ao aumento da inadimplência. Nesse cenário, a ascensão da Indústria 4.0 e da transformação

digital tem impulsionado o uso de técnicas de Aprendizado de Máquina (ML), que se destacam como ferramentas poderosas para apoiar a tomada de decisão automatizada e baseada em dados (LESSMANN et al., 2015). O ML é um subcampo da inteligência artificial que se baseia na ideia de que sistemas computacionais podem aprender a partir de dados sem programação explícita (ALPAYDIN, 2020), sendo a classificação supervisionada a abordagem adotada neste trabalho. Entre os algoritmos supervisionados empregados, estão o K-Nearest Neighbors (KNN), que classifica por similaridade e distância (HAN; KAMBER; PEI, 2011), e o Random Forest, um algoritmo de aprendizado de conjunto (*ensemble learning*) que combina múltiplas árvores de decisão para melhorar a acurácia e reduzir o sobreajuste (BREIMAN, 2001). O foco do presente estudo foi o desenvolvimento de um modelo preditivo para análise de aprovação de crédito no Orange Data Mining, buscando, primariamente, a avaliação comparativa do desempenho dos algoritmos K-Nearest Neighbors (KNN) e Random Forest em um problema de classificação supervisionada.

OBJETIVO: Analisar a aprovação de crédito no setor financeiro, utilizando técnicas de aprendizado de máquina.

MÉTODO: Para a realização deste estudo, utilizou-se a base de dados *Loan Approval Dataset* disponível na plataforma Kaggle, composta por 2.000 registros, cada um representando uma solicitação de empréstimo. As variáveis preditoras incluíram atributos demográficos e financeiros dos solicitantes, tais como renda anual (income), pontuação de crédito (credit_score), valor do empréstimo solicitado (loan_amount), tempo de emprego (years_employed) e pontuação de relacionamento com a instituição (points), sendo que as variáveis de identificação e localização (name e city) foram removidas por não contribuírem para o processo de classificação. O alvo (target) do problema foi a variável loan_approved, configurada como um problema de classificação binária

(aprovado = True, negado = False). No pré-processamento, constatou-se que não havia valores ausentes no conjunto de dados, o que dispensou procedimentos de imputação ou remoção de registros, e as variáveis numéricas foram normalizadas para o intervalo [0,1], a fim de evitar que atributos com escalas diferentes influenciassem de forma desproporcional os algoritmos, especialmente o KNN. Após o pré-processamento, os dados foram divididos em conjuntos de treino (70%) e teste (30%) para garantir uma avaliação imparcial. Em seguida, aplicaram-se os seguintes algoritmos de classificação: o KNN testado com diferentes valores de k (k=3, k=5, k=7) utilizando a distância Euclidiana como métrica, e o Random Forest, implementado como um classificador ensemble com ajustes no número de árvores e profundidade máxima para otimização do desempenho e redução do sobreajuste. A avaliação dos modelos foi realizada por meio de validação cruzada (10-fold cross-validation), e as métricas consideradas incluíram Acurácia, Precisão, Recall (Sensibilidade), F1-Score e AUC (Área sob a Curva ROC), permitindo analisar tanto o desempenho geral quanto a capacidade dos modelos de distinguir entre classes.

RESULTADOS E DISCUSSÕES: Após a aplicação dos algoritmos de classificação KNN e Random Forest na base de dados de empréstimos e a avaliação de seus desempenhos por meio de validação cruzada estratificada com 10 folds, os resultados demonstraram clara superioridade do Random Forest, o qual apresentou desempenho perfeito, atingindo 100% em todas as métricas avaliadas (AUC, Acurácia, Precisão, Recall e F1-Score). Em contrapartida, o modelo KNN (k=5) obteve resultados ligeiramente inferiores, com 0.997 em Acurácia, 0.980 em Precisão e 0.998 em AUC. O desempenho do Random Forest sugeriu extrema eficácia na aprendizagem e generalização dos padrões complexos do conjunto de dados, atributo decorrente de sua natureza ensemble, que contribui para a redução da variância e evita o sobreajuste característico de árvores singulares. O AUC de 100% evidenciou a capacidade de distinção perfeita entre solicitantes aprovados e reprovados, enquanto o Recall de 100% se destacou como resultado mais crítico no contexto financeiro, pois significou que o modelo identificou corretamente todos os casos positivos reais, minimizando a ocorrência

de falsos negativos (classificação de indivíduos de alto risco como seguros), aspecto essencial para mitigação de perdas financeiras. Embora o KNN tenha apresentado desempenho elevado, a pequena margem de erro evidenciou que a classificação baseada em similaridade de instâncias mostrou-se menos robusta do que a combinação de múltiplas árvores de decisão. Por fim, discutiu-se como limitação que resultados podem refletir baixa complexidade do conjunto de dados ou possível *leakage* de variáveis, ressaltando a importância de validação em bases maiores e diversificadas para confirmar a capacidade de generalização dos modelos.

CONSIDERAÇÕES FINAIS: Este estudo explorou a aplicação de algoritmos de Machine Learning, especificamente KNN e Random Forest, para a tarefa de classificação em uma base de dados de análise de empréstimos. Os resultados demonstraram que o Random Forest obteve um desempenho superior, atingindo 100% em todas as métricas de avaliação (Acurácia, Precisão, Recall, F1-Score e AUC), o que indicou sua eficácia total na previsão do risco de empréstimos para o conjunto de dados utilizado, enquanto o KNN, embora com alto desempenho, mostrou-se ligeiramente inferior. A metodologia empregada, que incluiu o pré-processamento de dados e a avaliação robusta por meio de validação cruzada, foi fundamental para garantir a confiabilidade dos modelos preditivos. A relevância prática deste estudo reside na capacidade de as instituições financeiras utilizarem o modelo Random Forest para automatizar as decisões de concessão de crédito, reduzir custos operacionais e mitigar proativamente o risco de inadimplência, aproveitando a alta precisão preditiva demonstrada. Para trabalhos futuros, sugere-se a expansão deste estudo com a utilização de bases de dados de empréstimos maiores e mais representativas, que permitam uma análise mais aprofundada do comportamento dos modelos em cenários reais e complexos. Além disso, a investigação de técnicas avançadas de balanceamento de classes para lidar com a possível desproporção entre empréstimos aprovados e inadimplentes, bem como a otimização de hiperparâmetros, são caminhos promissores para aprimorar ainda mais a performance e a capacidade de generalização dos modelos.

PALAVRAS-CHAVE: Análise de Crédito, Aprendizado de Máquina, Classificação Supervisionada, KNN, Random Forest, Risco de Inadimplência

REFERÊNCIAS

ALPAYDIN, Ethem. *Introduction to Machine Learning*. 4. ed. Cambridge: MIT Press, 2020.

AGÊNCIA BRASIL. Brasil registra alta de 11% na concessão de crédito em 2024. *Agência Brasil – EBC*, 27 jan. 2025. Disponível em: https://agenciabrasil.ebc.com.br/radioagencia-nacional/economia/audio/2025-01/brasil-registra-alta-de-11-na-concessao-de-credito. Acesso em: 10 out. 2025.

BANCO CENTRAL DO BRASIL. *Estatísticas monetárias e de crédito*. Dados abertos. Disponível em: https://dadosabertos.bcb.gov.br/. Acesso em: 10 out. 2025.

BREIMAN, Leo. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data Mining: Concepts and Techniques*. 3. ed. San Francisco: Morgan Kaufmann, 2011.

KAGGLE. *Loan Approval Dataset*. Disponível em: https://www.kaggle.com/. Acesso em: 8 out. 2025.

LESSMANN, Stefan; BAUER, A.; SEOW, H. V.; THOMAS, Lyn C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, v. 247, n. 1, p. 124–136, 2015.

MESTER, Loretta J. What's the point of credit scoring? *Business Review*, Federal Reserve Bank of Philadelphia, p. 3–16, 1997.

ORANGE DATA MINING. *Orange – Data Mining Fruitful & Fun*. 2024. Disponível em: https://orange.biolab.si/. Acesso em: 6 de outubro 2025.

THOMAS, Lyn C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, v. 16, n. 2, p. 149–172, 2000.