AVALIAÇÃO DO RISCO DE CRÉDITO: COMPARAÇÃO DA PERFORMANCE DE UMA REDE NEURAL ARTIFICIAL COM O SUPPORT VECTOR MACHINE COM OTIMIZAÇÃO BAYESIANA

Victor Hugo de Oliveira Mota [2], Analucia Vieira Fantin[3].

[1] Vinculado ao projeto ": Machine Learning: estudo de aplicações no setor financeiro"
[2]Acadêmico do Curso de Ciências Econômicas – ESAG – Bolsista PIBITI/CNPQ
[3] Orientadora, Departamento de Ciências Econômicas – ESAG – analucia.fantin@edu.udesc.br.

Estabilidade financeira é um dos mecanismos mais importantes para o rápido e sustentado progresso econômico dos bancos (Bhatore et.al, 2020). Entre os diversos indicadores de estabilidade financeira, destaca-se o risco de crédito, que é definido como o risco de perda financeira quando o tomador de empréstimos falha ao cumprir um acordo financeiro (Bhatore et. al, 2020). Anteriormente, o risco de crédito era mensurado através de avaliação subjetiva dos profissionais dos setores bancários, o que conferia baixa confiabilidade dos resultados (Hu e Tang, 2006). Com o aprimoramento das capacidades dos computadores e o notável desenvolvimento da área de inteligência artificial nos últimos vinte anos, mais especificamente a área de Machine Learning, diversas técnicas de mensuração de risco financeiro têm sido utilizadas (Bhatore et al., 2020).

O objetivo do trabalho consiste em comparar os dois métodos de avaliação de risco de crédito que aparentam ser mais promissores na literatura, o *Support Vector Machine* (SVM) e a Rede Neural Artificial (*Artificial Neural Network* ou "ANN"). Para o SVM, será utilizado a metodologia de otimização bayesiana para otimizar os hiper parâmetros do modelo.

Para abordagem, foi utilizado o conjunto de dados *German Credit Risk*. O conjunto possui 1000 observações, com 10 variáveis. As variáveis independentes utilizadas foram idade, gênero, tipo de trabalho, tipo de moradia, uma variável categórica que representa o valor da conta poupança e outra da conta corrente, a duração do empréstimo e uma variável *dummy* que representa o propósito do dinheiro ter sido tomado emprestado. Com base nessas nove variáveis, espera-se predizer se o risco de crédito de um determinado indivíduo é baixo ou alto. No primeiro caso, a variável dependente "risco" recebe valor um e valor zero no segundo caso.

As variáveis "conta poupança" e "conta corrente" apresentavam 183 e 394 valores faltantes, respectivamente. Para imputação, os valores categóricos foram substituídos por 0, no caso de *little*; 1 no caso de *moderate*; 2 em *rich*; e 3 em *quite rich*. Em seguida, os valores faltantes foram substituídos pela média desses valores. As variáveis categóricas gênero, tipo de moradia e propósito foram transformadas em valores binários utilizando a função *OneHotEnconder*, da biblioteca *Sckit-Learn*. As outras variáveis categóricas foram mantidas, dado que seus valores eram entradas numéricas e representavam uma escala de intensidade de cada variável. Os dados foram divididos em conjuntos de treino, teste e validação, nas proporções 70%, 15% e 15%, respectivamente. Para as divisões foram utilizadas, a biblioteca *Sckit - Learn* da linguagem Python, utilizando a função *train_test_split.*

Para ANN foi utilizado uma rede com três camadas. Na primeira (entrada) foram utilizados dezenove neurônios, correspondente ao número de colunas do conjunto de dados. Na camada

oculta foram dezesseis neurônios, usando função de ativação "relu" enquanto na camada de saída foi apenas um neurônio, utilizando função "*sigmoid*". Em relação aos hiper parâmetros, foi utilizando o otimizador "*Adam"* com taxa de aprendizado de 0,001 e regularizador L1 com *lambda* de 0,01. O erro ("*loss*") para o treinamento da rede foi o "binary cross-entropy" e a métrica de avaliação foi a acurácia. A arquitetura e os hiper parâmetros foram escolhidos pelo método de erro e tentativa.

Para o SVM foi utilizado a técnica de Otimização Bayesiana para encontrar os hiper parâmetros ótimos. O parâmetro ótimo encontrado foi de kernel polinomial com os hiper parâmetros "coeficiente" (coef0) igual a setenta e quatro, "C" igual a dez e grau igual a três.

Os resultados obtidos, referentes as diferentes métricas de avaliação, estão organizados na Tabela 1. Os resultados mostram que na comparação entre os dois modelos, com base em todas as métricas, com exceção da especificidade, o SVM apresenta uma performance melhor. Considerando que especificidade é uma medida da taxa de verdadeiros negativos, aliado ao fato de o conjunto de dados ser composto majoritariamente por indivíduos de elevado risco de crédito, esse resultado pode ser explicado por uma alta tendência da ANN treinada indicar que o indivíduo é um mal pagador. Apesar da melhor performance do SVM na maioria dos aspectos, deve-se considerar o elevado custo computacional da metodologia de otimização de parâmetros.

Tabela 1. Resultados da ANN e SVM

Métrica	**ANN**	**SVM com otimização bayesiana**
Acurácia	0,6381	0,6857
Especificidade	0,9384	0,7231
Sensibilidade	0,1500	0,6250
F1 score	0,2400	0,6024

Palavras-chave: Risco de crédito. *Machine Learning*. Redes Neurais. *Support Vector Machine*. Otimização Bayesiana.