Comparação de Algoritmos de Aprendizado de Máquina para Predição de Pontuação de Crédito

Renato De Sant'anna Lopes
renatosalopes@gmail.com
Instituto Federal do Espírito Santo
Árvore de DecisãotaddressRodovia ES-010 - Km 6,5, Manguinhos, Serra ES, Brasil

Leandro Resendo Colombi
leandro@ifes.edu.br
Programa de Pós-Graduação em Computação Aplicada (PPCOMP)
Instituto Federal do Espírito Santo
Árvore de DecisãotaddressRodovia ES-010 - Km 6,5, Manguinhos, Serra ES, Brasil

Filipe Mutz
filipe.mutz@ufes.br
Programa de Pós-Graduação em Computação Aplicada (PPCOMP)
Instituto Federal do Espírito Santo
Universidade Federal do Espírito Santo
Árvore de DecisãotaddressAv. Fernando Ferrari, 514 - Goiabeiras Vitória, ES, Brasil

ABSTRACT

According to the Central Bank of Brazil, the total value of credit operations in Brazil reached R$4.2 trillion in May 2021. Financial institutions must consider the risk of default associated with each operation. Credit analysis, which evaluates this risk, can be performed using machine learning algorithms. These algorithms compare new loan proposals to historical data to estimate the default probability based on the proposal and proponent characteristics. The accuracy of the model is critical to the profitability of institutions, so choosing the right algorithm is crucial. This study compares the performance of machine learning algorithms on three public datasets in the task of credit risk estimation. The results show that a stack of multiple classifiers achieved the highest accuracy at 81.41%, followed by XGBoost at 80.87% and Regressão Logística at 80.48%.

KEYWORDS

Artificial Intelligence. Machine Learning. Credit Scoring. Credit Risk. Classification.

1 INTRODUÇÃO

Há pelo menos 2000 anos antes de Cristo, o ato de emprestar bens esperando a devolução plena do valor emprestado com um adicional de juros já era registrado na Babilônia [1]. Com o passar do tempo, através consolidação dos bancos privados, por volta de 1800, e principalmente com grande demanda por financiamento para adquirir os primeiros automóveis, em 1920, houve uma maior necessidade de analisar o risco dos consumidores honrarem com o pagamento antes de aprovar um empréstimo [2]. O não pagamento de empréstimos pode causar prejuízos no orçamento de empresas, principalmente de micro, pequeno e médio porte, implicando no aumento do custo de seus produtos e serviços [3, 4].

Atualmente, o mercado de crédito tem alto impacto econômico nos países e ainda há espaço para crescimento. Segundo o Banco Central do Brasil [5], o saldo total em operações de crédito alcançou R$4,2 trilhões em maio de 2021 no Brasil. De acordo com o Instituto Locomotiva [6], 34 milhões de brasileiros (aproximadamente 21% da população), possuem pouco ou nenhum acesso a serviços bancários. Já nos Estados Unidos, no mesmo mês de maio foi registrado um total de aproximadamente US$4,2 trilhões, segundo The Fed [7], e o total de adultos americanos que são desbancarizados é 14,1 milhões [8].

No Brasil, em 2020, 43% dos empréstimos pessoais foram feitos por consequência da pandemia do Coronavírus e cerca de 70% desses empréstimos foram os primeiros feitos por estes clientes a fim de complementar a renda, comprar móveis para trabalhar em casa, ajudar um conhecido ou suprir gastos com saúde [9]. Em 2021, Serasa Experian [10] fez uma pesquisa sobre o uso de crédito na pandemia, e foi percebido que 79% dos entrevistados utilizaram alguma fonte de crédito, sendo que 32% usaram 6 vezes ou mais, para amenizar problemas financeiros.

Para analisar se um cliente pode receber crédito ou não, é comum o uso da técnica de pontuação de crédito [2]. Esta técnica foi desenvolvida na Segunda Guerra Mundial devido à escassez de trabalhadores especializados em analise de crédito, muitos dos quais foram requisitados para o serviço militar. A fim de permitir que funcionários inexperientes realizassem a atividade, analistas escreveram regras para definir a pontuação e, de forma relacionada, o risco de um pedido de empréstimo [11]. Para calcular essa pontuação, podem ser utilizados dados como a idade do cliente, renda mensal, se ele está empregado, se possui casa própria, além do valor e motivo do empréstimo [2]. Esses dados são associados a faixas de pontuações e quanto menor a pontuação, maior o risco de inadimplência atribuído ao pedido.

Embora o uso da pontuação de crédito ainda seja comum, com a disseminação dos computadores, foram propostos métodos automatizados para realização da tarefa [12]. Dois grandes grupos de métodos se destacam, a saber, aqueles baseados em sistemas especialistas e os sistemas baseados em aprendizado de máquina [12]. O segundo grupo atribui uma pontuação para pedidos de crédito utilizando dados históricos acerca de pedidos e seus respectivos resultados. A acurácia de tais sistemas está diretamente relacionada com a lucratividade das instituições credoras. De acordo com Steiner et al. [13], o custo de pontuações incorretas pode ultrapassar o ganho com diversas análises corretas.

Este trabalho compara algoritmos de aprendizado de máquina na tarefa de análise de risco de crédito com a finalidade de identificar métodos que levem à uma melhor performance. Para isto, são utilizados três bases de dados públicas relacionadas à análise de crédito *South German Credit (UPDATE) Data Set*, *Australian Credit Approval* e *Default of Credit Card Clients Data Set*, que estão disponíveis

em *UCI Machine Learning Repository* [14]. Foram comparados os algoritmos K-Vizinhos mais Próximos (*k-Nearest Neighbors* - KNN), Árvores de Decisão, Floresta Aleatória (*Random Forest*), *AdaBoost*, Extreme Gradient Boosting (*XGBoost*), Regressão Logística, *Support Vector Machines* (SVM) e Perceptron de Múltiplas Camadas (*Multilayer Perceptron*) e um *Ensemble* com todos os modelos anteriores empilhados (*Stack Ensemble*) [15–17]. Resultados experimentais mostraram que o *Stack* alcançou a melhor performance com uma acurácia média de 81,41%, seguido por *XGBoost* com 80,87% e Regressão Logística com 80,48%.

2 TRABALHOS CORRELATOS

Lessmann et al. [18] realizaram uma comparação de desempenho entre 41 algoritmos de classificação baseados em aprendizado de máquina. Dentre os algoritmos utilizados, eles podem ser classificados em três tipos diferentes: classificadores individuais, *ensembles* homogêneos e *ensembles* heterogêneos. Para a comparação foram utilizadas 8 bases de dados, sendo elas: *Australian Credit Approval* e *German Credit* [14], Th02 [2], Bene-1, Bene-2, UK-1 a UK-4 [19], PAK [20] e GMC [21]. Em dados reais, existe um número substancialmente maior de propostas aprovadas que de rejeitadas. Contudo, na avaliação não foram utilizadas técnicas de balanceamento para que a capacidade dos algoritmos de lidar com o desbalanceamento natural dos dados pudesse ser avaliada. Foi utilizada a técnica de validação cruzada *N×2-fold* [22] onde *N* vezes os dados são aleatoriamente organizados em dois conjuntos de mesmo tamanho, um deles para treinamento e outro para avaliação dos modelos e, em seguida, é feita a avaliação permutando os papéis dos conjuntos. Para obter dados de validação foi aplicada validação cruzada de 5 *folds*.

Os resultados mostraram que classificadores avançados e recentes alcançam bom desempenho, mas não superam classificadores tradicionais. Também foi possível atestar que diversos classificadores obtiveram um desempenho maior que a Regressão Logística, que é considerado um classificador padrão na área, principalmente os *ensembles* heterogêneos. As MLPs têm melhor desempenho do que *Extreme Learning Machines*, *Random Forests* se saíram melhor que *Rotation Forests* e os *Dynamic Selective Ensembles* obtiveram resultados piores do que quase todos os outros classificadores. Vale notar, contudo, que os resultados desta comparação de desempenho não representam, necessariamente, o desempenho definitivo de cada classificador. Análises laboratoriais, como esta, podem superestimar as vantagens de classificadores avançados mas, ainda assim, é possível que eles ganhem momento na indústria.

Louzada et al. [12] realizou uma revisão sistemática da literatura de técnicas para análise de crédito utilizando aprendizado de máquina. Foram analisados 187 artigos publicados entre 1992 e 2015 e identificados nas bases *ScienceDirect*, *Engineering Information*, *Reaxys* e *Scopus*. Foram identificados trabalhos utilizando, dentre outros, Redes Neurais Artificiais, SVMs, Regressão Linear, Árvores de Decisão, Programação Genética e Redes Bayesianas. Contudo, o autor observou que muitas comparações feitas na literatura utilizavam métricas e bases de dados diferentes. Assim, o trabalho comparou usando uma metodologia padronizada a performance dos métodos e utilizando as bases de dados *Australian Credit Approval*, *German Credit* e *Japanese Credit Screening* [14]. SVMs e sistemas baseados em lógica *Fuzzy* alcançaram os melhores desempenhos [12].

Bequé and Lessmann [23] comparam *Extreme Learning Machines* (ELMs) com modelos já consolidados em pontuação automática de crédito como redes neurais e KNN. A motivação foram resultados positivos alcançados com ELMs em diversas tarefas na época. A análise considerou as bases de dados *Australian Credit Approval* e *German Credit* [14] e três métricas: facilidade de uso, consumo de recursos computacionais e performance preditiva. A facilidade de uso foi dada pela quantidade de hiperparâmetros e o quão sensíveis os modelos são a mudanças nos seus valores. No consumo de recursos, foi medido o uso de memória RAM e os tempos necessários para treino e predição. A terceira métrica mede a taxa de acerto dos classificadores. Constatou-se que a ELM alcança acurácia satisfatória, além de ser mais fácil de usar que a rede neural e mais eficiente que rede neural e SVM.

Nascimento et al. [24] desenvolveram uma arquitetura de aprendizado de máquina para avaliar propostas de empréstimo pessoal utilizando algoritmos de aprendizado de máquina para uma triagem de casos de aprovação e rejeição óbvias. Os casos resultantes são enviados para avaliação por especialistas. A análise automática é feita usando dois modelos em sequência referidos como modelos M1 e M2. Apenas as propostas não classificados por M1 são enviadas para M2. Similarmente, apenas propostas não classificadas por M2 são transmitidas para os especialistas humanos. Experimentos avaliaram além da taxa de acerto dos classificadores, o percentual de amostras classificadas automaticamente. Resultados mostraram que os modelos alcançaram alta acurácia e classificaram um volume considerável de propostas. Tomando a acurácia de 97%, o limiar de 0,75 do Modelo M1 classifica 86,56% das propostas e o Modelo M2 mais 4,04% do restante das propostas.

3 DESENVOLVIMENTO

Esta seção descreve as etapas para o desenvolvimento deste trabalho. Inicialmente, as bases de dados e o seu pré-processamento são descritas. Em sequência, as métricas de avaliação são apresentadas. Por fim, é detalhada a metodologia de experimentação, incluindo ferramentas e técnicas utilizadas.

3.1 BASES DE DADOS

Neste trabalho, foram analisadas três bases de dados, *German Credit Data*, *Australian Credit Approval* e *Default of Credit Card Clients Data Set*, sendo esta última referente a inadimplência de clientes taiwaneses, que estão disponíveis em *UCI Machine Learning Repository* [14]. A versão utilizada da base *German Credit Data* é a *South German Credit (UPDATE) Data Set*, onde todos os dados, que eram categorias em palavras, foram codificadas para números. Deste ponto em diante, as bases serão referenciados por seu país de origem, i.e., base alemã, australiana e taiwanesa.

A Tabela 1 contém o número de amostras, características categóricas e numéricas das bases, além da taxa de inadimplência. A base de dados taiwanesa possui uma quantidade maior de amostras com dados de 30000 clientes. Ela contém 3 atributos categóricos e 20 numéricos, totalizando 23 atributos por cliente. Nesta base, cerca de 22% dos clientes são inadimplentes, apresentando um desbalanceamento moderado [25]. A base alemã possui um total de 1000

Tabela 1: Metadados das bases utilizadas

Características	Australiana	Alemã	Taiwanesa
Amostras	690	1000	30000
Categóricas	8	16	3
Numéricas	6	4	20
Total	14	20	23
% Inadimplência	44%	30%	22%

amostras, sendo 30% de inadimplentes. Esta base possui 20 características sendo 16 categóricas e 4 numéricas. A base australiana é composta por 690 amostras, onde cerca de 44% é inadimplente. Esta base conta com 14 atributos em que 8 são categóricos e 6 são numéricos. Para simplificar a experimentação foi padronizado nas três que a classe inadimplente terá valor 1 e a adimplente terá valor 0. Em cerca de 5% das instâncias da base *Australian Credit Approval* haviam valores faltantes. Atributos categóricos foram substituídos pela moda e numéricos pela média.

Na etapa de pré-processamento, foi realizada a padronização dos dados. Esta técnica tem como objetivo colocar diferentes características em uma mesma escala [15]. Neste trabalho foi utilizada a padronização *Z-Score*, representada na Equação 1, em que os dados são ajustados para que tenham média (μ) 0 e desvio padrão (σ) de 1 [26]. Seja x o valor original das variáveis e x' o valor após a normalização, então:

$$x' = \frac{(x - \mu)}{\sigma} \tag{1}$$

3.2 MÉTRICAS

Nas avaliações foram utilizadas a matriz de confusão e as métricas derivadas acurácia, precisão, revocação e *f1 score* [15]. Na matriz confusão, as linhas representam valores preditos pelo modelo e as colunas representam os valores conhecidos e previamente anotados. A diagonal principal mostra os valores corretamente classificados pelo algoritmo, sendo eles verdadeiros positivos (VP) e verdadeiros negativos (VN). As demais células da matriz contém os valores classificados incorretamente, os falsos positivos (FP) e os falsos negativos (FN).

A acurácia (Equação 2) é responsável por informar a taxa de classificação correta do modelo. Ela indica qual o número de predições corretas sobre o total de predições, sejam positivas ou negativas.

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2}$$

A precisão (Equação 3) indica a habilidade do classificador de classificar corretamente uma amostra positiva [15]. De todos os casos classificados como positivos pelo modelo, a precisão indica a porcentagem de verdadeiros positivos.

$$Precisao = \frac{VP}{VP + FP} \tag{3}$$

Revocação (Equação 4), também chamada de sensibilidade, representa a capacidade do modelo em classificar elementos positivamente [15]. Do total de casos positivos, a revocação indica a porcentagem de casos detectados pelo modelo.

$$Revocacao = \frac{VP}{VP + FN} \tag{4}$$

A métrica *F1 Score* (Equação 5) também é considerada um modelo de acurácia na classificação [15]. Esta métrica consiste na média harmônica entre precisão e revocação, que são grandezas inversamente proporcionais.

Por fim, é importante salientar que os erros de classificação, falsos positivos e falsos negativos, não são simétricos na tarefa de predição de *score* de crédito. Para uma empresa é muito pior dar crédito para um cliente que vai se tornar inadimplente. Em conjunto com a acurácia também é interessante observar a revocação, que identifica a taxa de acertos em todos os casos anotados de maneira positiva.

$$F1Score = \frac{2 * (Precisao * Revocacao)}{Precisao + Revocacao} \tag{5}$$

3.3 METODOLOGIA DE AVALIAÇÃO

Foram feitas duas avaliações para cada modelo e cada base de dados. Na primeira avaliação, os modelos foram treinados e avaliados com o mesmo conjunto de dados para avaliar a possibilidade de *underfitting*. Na segunda, para avaliar a generalização dos algoritmos e a ocorrência de *overfitting*, eles foram submetidos à técnica de validação cruzada *k-fold* com 10 *folds* estratificados [15].

Na avaliação usando dados de treino ocorre o confronto entre o modelo já treinado com a mesma base de dados em que aprendeu a identificar os padrões, testando seu conhecimento em apenas dados vistos previamente. O desempenho desta comparação traz indícios da capacidade de aprendizado do modelo. Experimentos usando validação cruzada com *k-folds* tem o propósito de treinar o modelo com um subconjunto de dados da base e testar seu aprendizado em um subconjunto de dados previamente desconhecido. Ao usar novos dados para serem classificados, esta técnica avalia a generalização do modelo, ela estima como seria o desempenho do modelo em uma situação real.

Se a avaliação usando dados de treino levar à uma baixa acurácia, então é possível que o modelo esteja sofrendo de subajuste (*underfitting*). Este fenômeno acontece quando não possui capacidade suficiente para modelar os dados de treinamento ou quando os dados seguem uma distribuição que o modelo não é capaz de representar. Já uma acurácia muito alta na avaliação usando dados de treino pode apontar para uma especialização para aquele conjunto de dados em conjunto com uma acurácia reduzida na validação cruzada pode indicar superajuste (*overfitting*) [15]. Isso ocorre pois ao invés do modelo generalizar para abranger e classificar novos dados de forma adequada, ele se especializa nos dados fornecidos e encontra dificuldades em classificar dados inéditos.

Os modelos de aprendizado de máquina avaliados foram KNN, Árvores de Decisão, *Random Forest*, *XGBoost*, *AdaBoost*, Regressão Logística, *Support Vector Machines*, *Multilayer Perceptron* e *Ensemble Stack*, com o modelo de Regressão Logística como agregador.

Todas as etapas do processo, do pré-processamento até a avaliação dos métodos, foram feitas usando o *software Orange3* (disponível em https://orangedatamining.com/), um *software* de código aberto voltado para aprendizado de máquina e visualização de dados. Nele,

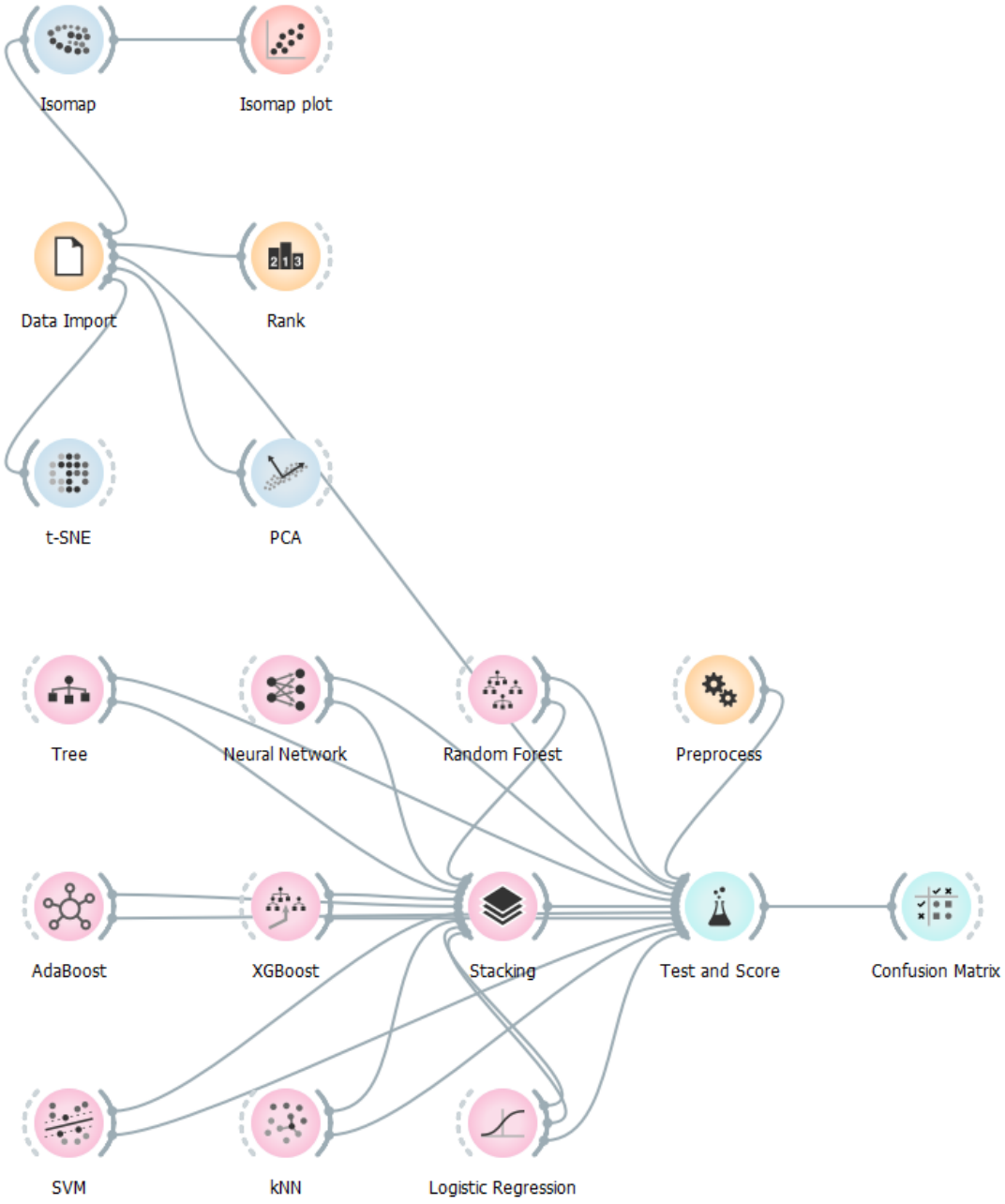

Figura 1: Fluxo de trabalho no *software Orange3*

é possível, de forma simples e intuitiva, gerar gráficos, além de treinar, testar e comparar o desempenho de modelos de aprendizado de máquina. O *Orange3* possui uma programação visual baseada em componentes, chamados de *widgets*. Cada *widget* possui uma funcionalidade, entradas e saídas, e eles podem ser conectados entre si a fim de alcançar um resultado específico, como visualizar ou pré-processar os dados [27].

A Figura 1 ilustra o fluxo de trabalho do *Orange3* utilizado neste trabalho. Este fluxo foi aplicado para as 3 bases de dados. Os nós, *widgets*, representam transformações aplicadas aos dados e arestas representam fluxos de dados entre *widgets*. São eles:

- *Data Import*: este componente é responsável por carregar os dados para serem utilizados. Aqui são definidos os tipos de dados de cada coluna, podendo ser categóricos, numéricos ou a classe em que os dados são classificados. Neste trabalho a importação foi feita através de arquivos TAB e CSV, porém também é possível importar diretamente de URLs ou tabelas SQL.

- *Rank*: este componente classifica cada atributo da base acerca da capacidade delas influenciarem na classificação do cliente. Dentre as métricas disponíveis estão Ganho de Informação e Coeficiente de Gini.
- *Isomap*: este *widget* gera os dados necessários para traçar o gráfico homônimo. O *Isomap plot* é um Gráfico de Dispersão para exibir as informações de forma visual.
- *t-SNE*: neste componente são ajustados os parâmetros da configuração do *t-SNE*.
- *PCA*: este *widget* calcula os componentes principais da base fornecida. Nele é possível observar a variância acumulada, de forma gráfica, ao se utilizar uma determinada quantidade de componentes principais.
- *Preprocess*: neste *widget* é possível selecionar e aplicar técnicas de pré-processamento variadas como padronizar os dados, selecionar características relevantes, preencher ou remover dados faltantes. Neste trabalho apenas a padronização dos dados foi aplicada.
- *Test and Score*: este *widget* recebe como entrada diversos modelos de aprendizado de máquina e os dados utilizados para treino e teste. Aqui também são exibidas as métricas de desempenho para cada modelo.
- *Confusion Matrix*: este componente fornece a matriz de confusão para cada modelo, indicando a quantidade de verdadeiros e falsos positivos, bem como verdadeiros e falsos negativos.
- *Stack*: este *widget* representa um *ensemble* de diferentes modelos de aprendizado de máquina. Diversos modelos são utilizados para predizer o resultado final e um modelo adicional é responsável por agregar os valores fornecidos por cada modelo.

O fluxo do trabalho inicial com a importação das bases de dados. Uma análise preliminar é feita através dos *widgets* de *Isomap*, *t-SNE*, *PCA* e *Rank*. Estas análises foram utilizadas apenas para avaliação qualitativa da distribuição dos dados. Em seguida, a avaliação dos modelos ocorre no *widget Test and Score*. Este é *widget* é responsável por receber como entrada todos os modelos de aprendizado de máquina e o *Preprocess* para padronizar os dados a cada divisão da validação cruzada. Dentro do *widget Test and Score* é possível visualizar os resultados das métricas obtidas por cada avaliação modelo e exportar relatórios. Por fim, o *widget Confusion Matrix* recebe os resultados de classificação de cada modelo e compara com os dados anotados, gerando uma matriz confusão para cada modelo e a respectiva avaliação executada.

Os valores dos hiperparâmetros foram definidos como os valores padrão do *Orange3*. Os modelos que possuem a opção de permitir a replicabilidade dos resultados ou utilizar uma semente para um gerador de números aleatórios, tiveram essa opção habilitada.

4 EXPERIMENTOS

Esta seção apresenta e discute os resultados obtidos nas avalições experimentais. As Tabelas de 2 até 8 foram geradas a partir do *widget Test and Score*. As métricas de precisão, revocação e *F1 Score* foram calculadas relativas a classe de clientes inadimplentes (1).

A Tabela 2 contém os resultados obtidos utilizando dados de treino da base australiana. Nesta avaliação, *AdaBoost* e *XGBoost* atingiram 100% em todas as métricas. Os três modelos com menor acurácia foram Regressão Logística com 87,68%, KNN com 88,41% e SVM com 91,74%. A Tabela 5 lista os resultados obtidos através da validação cruzada nesta base. Na validação cruzada, o método de *Ensemble Stack* teve a maior acurácia na validação cruzada, 86,81%, seguido pelo *XGBoost* com 86,52% e *Random Forest* com 86,38%. O modelo com menor acurácia foi *AdaBoost* com 81,30%, seguido por Árvore de Decisão 83,48% e KNN com 84,06%. A diferença entre os modelos com maior e menor acurácia foi de 5,51%. *Adaboost* demonstrou um aprendizado perfeito na avaliação com a base de

Tabela 2: Avaliação usando dados de treino na base australiana

Modelo	Acurácia	Precisão	Revocação	F1 Score
AdaBoost	**100,00%**	**100,00%**	**100,00%**	**100,00%**
KNN	88,41%	87,46%	86,32%	86,89%
Regressão Logística	87,68%	83,84%	89,58%	86,61%
MLP	96,96%	96,43%	96,74%	96,59%
Random Forest	96,52%	96,70%	95,44%	96,07%
Stack	98,12%	99,00%	96,74%	97,86%
SVM	91,74%	88,82%	93,16%	90,94%
Árvore de Decisão	96,09%	96,36%	94,79%	95,57%
XGBoost	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Tabela 3: Avaliação usando dados de treino na base alemã

Modelo	Acurácia	Precisão	Revocação	*F1 Score*
AdaBoost	**100,00%**	**100,00%**	**100,00%**	**100,00%**
KNN	80,80%	77,84%	50,33%	61,13%
Regressão Logística	78,80%	68,80%	53,67%	60,30%
MLP	**100,00%**	**100,00%**	**100,00%**	**100,00%**
Random Forest	95,30%	96,34%	87,67%	91,80%
Stack	95,80%	99,24%	86,67%	92,53%
SVM	82,80%	87,21%	50,00%	63,56%
Árvore de Decisão	94,80%	98,06%	84,33%	90,68%
XGBoost	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Tabela 4: Avaliação usando dados de treino na base taiwanesa

Modelo	Acurácia	Precisão	Revocação	*F1 Score*
AdaBoost	**99,93%**	**99,83%**	**99,85%**	**99,84%**
KNN	84,46%	72,57%	47,83%	57,66%
Regressão Logística	81,14%	71,80%	24,25%	36,25%
MLP	83,30%	70,37%	42,30%	52,84%
Random Forest	95,81%	97,53%	83,17%	89,78%
Stack	86,57%	84,26%	48,33%	61,43%
SVM	74,06%	23,48%	7,64%	11,53%
Árvore de Decisão	96,06%	96,39%	85,40%	90,56%
XGBoost	86,80%	83,31%	50,41%	62,81%

Tabela 5: Resultado dos modelos na base australiana com validação cruzada

Modelo	Acurácia	Precisão	Revocação	F1 Score
AdaBoost	81,30%	78,71%	79,48%	79,09%
KNN	84,20%	83,22%	80,78%	81,98%
Regressão Logística	86,09%	82,46%	**87,30%**	84,81%
MLP	85,51%	84,16%	83,06%	83,61%
Random Forest	86,38%	83,39%	86,64%	84,98%
Stack	**86,81%**	84,18%	86,64%	**85,39%**
SVM	85,22%	81,54%	86,32%	83,86%
Árvore de Decisão	83,48%	82,06%	80,46%	81,25%
XGBoost	86,52%	**84,52%**	85,34%	84,93%

treino porém apresentou a menor acurácia na validação cruzada. *XGBoost* também atingiu uma acurácia de 100% na avaliação de treino e conseguiu obter a segunda melhor acurácia na validação cruzada, com 86,52%.

A Tabela 3 representa os resultados obtidos após avaliação usando dados de treino na base alemã. Nesta avaliação, *AdaBoost*, *XGBoost* e MLP alcançaram 100% de acurácia, classificando todas as amostras corretamente. Já Regressão Logística obteve a menor acurácia nesta avaliação do aprendizado, com 78,80%. A Tabela 6 lista resultados gerados pela validação cruzada. Neste experimento, os três modelos com maior acurácia foram, respectivamente, o método *Ensemble Stack* com 75,60%, *XGBoost* com 74,50% e Regressão Logística com 74,30%. Os três modelos com desempenho inferior foram *AdaBoost* com 68,20%, SVM com 68,60% e Árvore de Decisão com 69,30%. A diferença entre a melhor e pior acurácia foi de 7,7%. Novamente, *AdaBoost* teve a pior performance na validação cruzada e *XGBoost* obteve a segunda melhor performance, sendo que ambos atingiram 100% de acurácia na avaliação com apenas base de treino. Já a rede MLP, que também atingiu 100% de acurácia com a base de treino, obteve uma performance próxima à média, comparada aos outros modelos.

A Tabela 4 possui os resultados da avaliação usando dados de treino na base taiwanesa. Nesta base, *AdaBoost* alcançou a maior acurácia com 99,93%, seguido por Árvore de Decisão com 96,06%, e *Random Forest* com 95,81%. O SVM obteve o pior desempenho na acurácia, durante a avaliação com o conjunto de treino com 74,06%. A Tabela 7 traz os resultados da validação cruzada na base de dados taiwanesa. Durante a validação cruzada, o método *Ensemble Stack* obteve a maior acurácia, 81,81%, seguido por *XGBoost* 81,58% e MLP com 81,25%. Ainda na validação cruzada, SVM obteve 63,59% de acurácia, seguido por Árvore de Decisão, 74,82% e KNN, 79,30%, sendo os modelos com pior desempenho nesta modalidade. A diferença entre a pior e melhor acurácia foi de 18,23%. *AdaBoost* atingiu a melhor performance na base de treino e teve um desempenho intermediário na validação cruzada.

A Tabela 8 contém o desempenho médio dos modelos nas três bases de dados. Na média dos resultados pode ser observado que, nas três bases, o método *Ensemble Stack* obteve a melhor performance na validação cruzada com uma acurácia média de 81,41%.

Tabela 6: Resultado dos modelos na base alemã com validação cruzada

Modelo	Acurácia	Precisão	Revocação	*F1 Score*
AdaBoost	68,10%	46,93%	48,33%	47,62%
KNN	72,80%	57,61%	35,33%	43,80%
Regressão Logística	74,30%	59,07%	46,67%	52,14%
MLP	73,70%	56,88%	**51,00%**	**53,78%**
Random Forest	74,20%	60,00%	42,00%	49,41%
Stack	**75,60%**	**63,21%**	44,67%	52,34%
SVM	68,60%	47,71%	48,67%	48,18%
Árvore de Decisão	69,30%	48,79%	47,00%	47,88%
XGBoost	74,50%	59,18%	48,33%	53,21%

Tabela 7: Resultado dos modelos na base taiwanesa com validação cruzada

Modelo	Acurácia	Precisão	Revocação	F1 Score
AdaBoost	79,02%	54,22%	33,09%	41,10%
KNN	79,30%	54,95%	35,74%	43,31%
Regressão Logística	81,06%	**71,24%**	24,11%	36,03%
MLP	81,25%	63,38%	36,12%	46,02%
Random Forest	80,62%	60,34%	36,11%	45,18%
Stack	**81,81%**	66,80%	35,32%	46,21%
SVM	63,59%	26,17%	35,47%	30,12%
Árvore de Decisão	74,82%	42,35%	**38,28%**	40,21%
XGBoost	81,58%	64,82%	36,62%	**46,80%**

Este resultado é esperado dado que ele utiliza o poder preditivo de todos os modelos para classificar os dados. *XGBoost* demonstrou um boa capacidade de aprendizado e ficou em segundo lugar nas três bases, marcando uma acurácia média de 80,87%. Tendo uma das melhores performances na base de treinto e estando entre as piores da validação cruzada, há indícios que *AdaBoost* sofre com sobreajuste, principalmente na base alemã.

A base australiana teve a maior acurácia média entre os modelos na validação cruzada, sendo 85,04%, em comparação à base 78,11% na base taiwanesa e 72,37% da base alemã. Para tentar explicar este fenômeno, a Figura 2 compara o gráfico t-SNE [15] das bases de dados. O t-SNE é um método de aprendizado de *manifolds* que pode ser usado para redução de dimensionalidade enquanto retém as informações originais no processo [28]. Esta técnica busca projetar os dados em altas dimensões em um espaço com baixa dimensão de forma que os agrupamentos sejam preservados, aproximando dados semelhantes por afinidade através de uma distribuição t [29]. A distribuição t é semelhante a uma distribuição normal, porém com o meio com uma altura menor e as bordas com uma altura maior, ela tem o propósito de facilitar a separação dos agrupamentos, de forma que eles não se sobreponham.

No t-SNE da base de dados australiana (Figura 2(a)), é possível ver duas regiões com maior concentração de cada tipo de cliente. Na parte direita há uma maior densidade de clientes inadimplentes

Tabela 8: Desempenho médio dos modelos em validação cruzada nas três bases

Modelo	Acurácia	Precisão	Revocação	*F1 Score*
AdaBoost	76,14%	59,95%	53,63%	55,94%
KNN	78,77%	65,26%	50,62%	56,36%
Regressão Logística	80,48%	70,92%	52,69%	57,66%
MLP	80,15%	68,14%	56,73%	61,14%
Random Forest	80,40%	67,91%	54,92%	59,86%
Stack	**81,41%**	**71,40%**	55,54%	61,31%
SVM	72,47%	51,81%	**56,82%**	54,05%
Árvore de Decisão	75,87%	57,73%	55,25%	56,45%
XGBoost	80,87%	69,51%	56,76%	**61,65%**

(1), enquanto à esquerda possui uma maior incidência de clientes com bom crédito, mesmo havendo clientes inadimplentes. Tanto no t-SNE da base alemã quanto taiwanesa (Figuras 2(b) e 2(c)), os dois tipos de clientes coexistem nas mesmas regiões, dificultando a separação visual dos mesmos. A separação das classes em dois núcleos principais na base australiana pode explicar a maior acurácia média nesta base.

Buscando identificar as características mais significativas da base de dados, a Tabela 9 compara os atributos com maior índice de impureza de Gini nas bases e quanto maior a porcentagem, mais provável é que esta característica tenha mais influência na classificação dos dados. O índice de impureza Gini é uma métrica que indica a capacidade de uma determinada característica em separar os clientes inadimplentes de adimplentes de forma mais homogênea.

Na base de dados australiana as características A8, A10 e A9 tem uma melhor distinção entre as classes, acima de 10%. Na base taiwanesa, apenas PAY_6 se posiciona acima de 1%, e todos os PAY_AMT aparecem na tabela, que indicam a quantidade do último pagamento e possuem uma correlação inversamente proporcional com a inadimplência. Já na base alemã, status, histórico de crédito e a duração do empréstimo consegue discernir melhor os clientes que ficarão inadimplentes, embora suas porcentagens sejam baixas, sendo de 5,20% o primeiro e 1,92% o último. É possível que a base taiwanesa alcançasse uma melhor acurácia média entre os modelos caso houvesse uma engenharia de características. A engenharia de características tem como objetivo remover atributos com pouco ou nenhum impacto no aprendizado e classificação [15].

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho comparou o desempenho entre nove modelos de aprendizado de máquina a fim de encontrar o modelo com melhor performance na identificação de possíveis clientes inadimplentes, no âmbito de análise de crédito. Os modelos utilizados foram *k-Nearest Neighbors*, Árvores de Decisão, *Random Forest*, *XGBoost*, *AdaBoost*, Regressão Logística, *Support Vector Machines*, *Multilayer Perceptron* e um *Ensemble Stack* que combina as predições dos modelos citados anteriormente.

Com os resultados dos experimentos foi possível observar que o método *Ensemble Stack* obteve o maior desempenho nas três bases, com uma acurácia média de 81,41%. O *XGBoost*, ocupou a segunda posição de acurácia média nas três bases de dados utilizadas, 80,87%, seguido por Regressão Logística com 80,48%. O SVM obteve o pior desempenho médio nas três bases, marcando 72,47% de acurácia média. *AdaBoost* alcançou um alto desempenho de aprendizado, porém obteve a terceira pior acurácia média na validação cruzada com 76,16% e foi o modelo com menor acurácia em duas das três bases.

Em trabalhos futuros serão aplicadas técnicas de engenharia de características e otimização de hiperparâmetros. A engenharia de características pode reduzir os atributos utilizados, simplificando modelos, melhorando o aprendizado e, por consequência, a acurácia em alguns casos. O uso de técnicas para otimização de hiperparâmetros pode melhorar a performance preditiva dos modelos ao definir parâmetros com melhor desempenho para cada classificador.

AGRADECIMENTOS

Este estudo foi financiado em parte pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil); e pela Fundação de Amparo à Pesquisa do Espírito Santo (FAPES, Brazil) - processo 2021-07KJ2; Os autores agradecem ainda ao apoio da FAPES e da CAPES (processo 2021-2S6CD, nºFAPES 132/2021) por meio do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

REFERÊNCIAS

[1] Edward M Lewis. *An introduction to credit scoring*. Fair, Isaac and Company, 1992.
[2] Lyn Thomas, Jonathan Crook, and David Edelman. *Credit scoring and its applications*. SIAM, 2017.
[3] Serasa Experian. Qual é a melhor estratégia para lidar com clientes inadimplentes?, 2020. URL https://empresas.serasaexperian.com.br/blog/clientes-inadimplentes/.
[4] Serasa Experian. 5 em cada 10 pmes sofreram com inadimplência de clientes durante a pandemia, 2021. URL https://www.serasaexperian.com.br/conteudos/estudos-e-pesquisas/5-em-cada-10-pmes-sofreram-com-inadimplencia-de-clientes-durante-a-pandemia/.
[5] Banco Central do Brasil. Estatísticas monetárias e de crédito, 2021. URL https://www.bcb.gov.br/estatisticas/estatisticasmonetariascredito.
[6] Instituto Locomotiva. Labs news: 34 milhões de brasileiros não têm acesso a serviços bancários, 2021. URL https://www.ilocomotiva.com.br/single-post/labs-news-34-milh%C3%B5es-de-brasileiros-n%C3%A3o-t%C3%AAm-acesso-a-servi%C3%A7os-banc%C3%A1rios.
[7] The Fed. Consumer credit - g.19, 2021. URL https://www.federalreserve.gov/releases/g19/current/.
[8] NextAdvisor TIME. What you should know if you are unbanked right now, 2021. URL https://time.com/nextadvisor/banking/what-to-know-if-you-are-unbanked/.
[9] Valor Investe. Pandemia motivou 43% dos empréstimos pessoais em 2020, mostra pesquisa, 2021. URL https://valorinveste.globo.com/produtos/credito/noticia/2021/02/19/pandemia-motivou-43percent-dos-emprestimos-pessoais-em-2020-mostra-pesquisa.ghtml.
[10] Serasa Experian. O papel do crÉdito em um momento de retomada, 2021. URL https://www.serasa.com.br/assets/cms/2021/Pesquisa-de-cre%CC%81dito-para-retomada-2021.pdf.
[11] Robert W Johnson. Legal, social and economic issues in implementing scoring in the us. *Credit scoring and credit control*, 19:32, 1992.
[12] Francisco Louzada, Anderson Ara, and Guilherme B Fernandes. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2):117–134, 2016.
[13] Maria Teresinha Arns Steiner, Celso Carnieri, Bruno H Kopittke, and Pedro J Steiner Neto. Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário. *Revista de Administra&ccdeil; ão da Universidade de São Paulo*, 34(3), 1999.
[14] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.
[15] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
[16] CM Bishop. Bishop-pattern recognition and machine learning-springer 2006. *Antimicrob. Agents Chemother*, pages 03728–14, 2014.

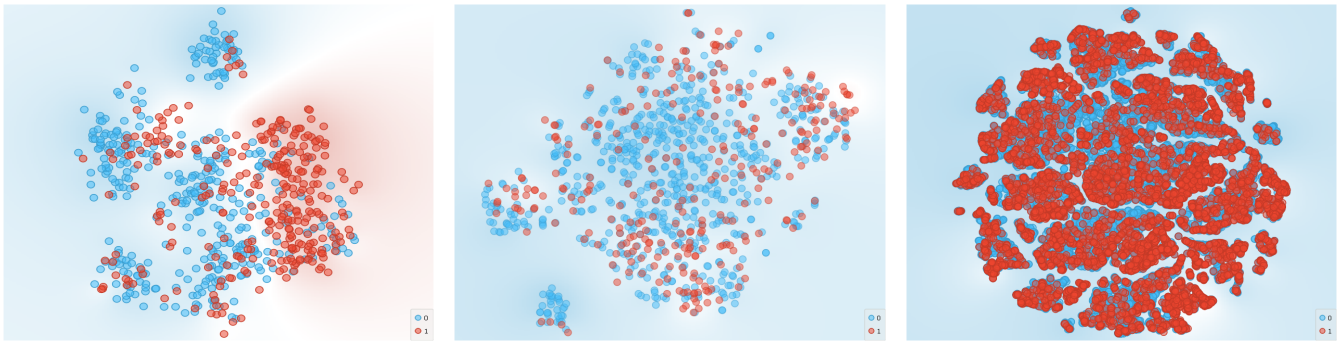

Figura 2: Gráficos t-SNE das bases de dados Australiana (esquerda), Alemã (meio) e Taiwanesa (direita). Cada ponto representa uma proposta de crédito, sendo que pontos azuis são aquelas que foram pagas no prazo e as em vermelho resultaram em inadimplência.

Tabela 9: Índice de impureza Gini das características da base de dados Australiana (esquerda), Alemã (meio) e Taiwanesa (direita).

Característica	Gini
A8	25,63%
A10	13,79%
A9	10,37%
A14	7,80%
A7	7,65%
A5	7,04%
A6	3,22%
A3	2,69%
A13	2,63%
A4	1,91%

Característica	Gini
status	5,20%
credit_history	2,59%
duration	1,92%
savings	1,52%
purpose	1,40%
amount	1,15%
property	1,00%
housing	0,78%
employment_duration	0,77%
age	0,65%

Característica	Gini
PAY_6	2,10%
LIMIT_BAL	0,94%
PAY_AMT1	0,83%
PAY_AMT2	0,74%
PAY_AMT3	0,66%
PAY_AMT4	0,54%
PAY_AMT6	0,50%
PAY_AMT5	0,44%
EDUCATION	0,19%
BILL_AMT6	0,07%

[17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[18] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[19] Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

[20] Mohammed Javeed Zaki, Jeffrey Xu Yu, Balaraman Ravindran, and Vikram Pudi, editors. *Advances in Knowledge Discovery and Data Mining, 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I*, volume 6118 of *Lecture Notes in Computer Science*, 2010. Springer. ISBN 978-3-642-13656-6. doi: 10.1007/978-3-642-13657-3. URL https://doi.org/10.1007/978-3-642-13657-3.

[21] Kaggle. Give me some credit, 2011. URL https://www.kaggle.com/c/GiveMeSomeCredit/.

[22] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[23] Artem Bequé and Stefan Lessmann. Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, 86:42–53, 2017.

[24] Pablo Simões Nascimento, Karin Satie Komati, and Jefferson Oliveira Andrade. Avaliação de crédito de empréstimos pessoais usando técnicas de aprendizado de máquina. *Revista de Sistemas de Informação da FSMA*, 2020.

[25] Google. Imbalanced data | data preparation and feature engineering for machine learning | google developers, 2021. URL https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data.

[26] Google. Normalization | data preparation and feature engineering for machine learning | google developers, 2021. URL https://developers.google.com/machine-learning/data-prep/transform/normalization.

[27] Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353, 2013. URL http://jmlr.org/papers/v14/demsar13a.html.

[28] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[29] Morris H DeGroot. *Probability and statistics*. Pearson, 2012.