

Avaliação de Técnicas de Aprendizado de Máquina Aplicadas à Análise de Risco de Crédito

Jane Thais Soares de Oliveira
Graduanda em Engenharia Física
Instituto de Engenharia,
Ciência e Tecnologia
Universidade Federal dos Vales
do Jequitinhonha e Mucuri
Janaúba MG, Brasil
email jane.oliveira@ufvjm.edu.br

Rogério Alves Santana
Instituto de Engenharia,
Ciência e Tecnologia
Universidade Federal dos Vales
do Jequitinhonha e Mucuri
Janaúba MG, Brasil
email rogerio.santana@ufvjm.edu.br

Honovan Paz Rocha
Instituto de Engenharia,
Ciência e Tecnologia
Universidade Federal dos Vales
do Jequitinhonha e Mucuri
Janaúba MG, Brasil
email honovan.rocha@ufvjm.edu.br

Resumo—O serviço de concessão de crédito ao consumidor final tem crescido de forma contundente nos últimos anos, sendo esta uma tendência se levarmos em consideração os juros mais baixos praticados com um cenário político/econômico cada vez mais conservador. Neste cenário, torna-se ainda mais relevante o processo de análise de risco de crédito, que ainda utiliza muitas técnicas arcaicas, com enfoque no tratamento individual e subjetivo dos dados do candidato à concessão. Visando otimizar a tarefa de análise de risco de crédito o presente trabalho tem como objetivo o estudo, implementação e avaliação de 10 dentre os principais algoritmos de classificação presentes na literatura aplicados à classificação de candidatos à concessão de crédito. Adicionalmente, a etapa de pré-processamento da base de dados utilizada incluiu a tarefa de seleção de características através de filtros univariados e multivariados, com o intuito de encontrar os atributos mais relevantes bem como a eliminação de redundâncias. Os experimentos realizados demonstraram a importância da seleção de características para melhoria dos algoritmos de classificação, além disso, foi possível verificar que os métodos do tipo *ensemble* obtiveram os melhores resultados de maneira geral considerando-se a base de dados utilizada.

Index Terms—Aprendizado de máquina, algoritmos de classificação, seleção de características, *ensemble*, análise de crédito

I. INTRODUÇÃO

Atualmente, as empresas têm encontrado um mercado cada vez mais competitivo, e a busca por sobrevivência se torna desleal com empresas que não conseguem se adaptar a novas tecnologias. Isso não seria diferente para as instituições financeiras, que lidam com grande quantidade de solicitações de concessão de crédito, onde deve-se levar em consideração que os potenciais clientes são oriundos de diferentes classes sociais, faixa etária, dentre outros diversos fatores socioeconômicos. Portanto, possuir modelos para análise de crédito que sejam rápidos e eficazes, poderia colocar este tipo de organização numa situação favorável em relação às concorrentes.

Com o crescimento da área de *Data Mining* nas últimas décadas, devido ao aumento da disponibilidade de dados, informações de alto valor agregado geradas a partir de dados brutos começaram a ser descobertas e utilizadas para gerar

mais valor nas organizações [1]. A área financeira não foge à regra, pois organizações deste segmento são detentoras de grande quantidade de dados ignorados ao longo dos anos, como por exemplo, dados fornecidos por um potencial cliente no ato de solicitação de crédito.

O processo de gerenciamento baseado em risco de crédito nas instituições financeiras vem passando por uma evolução ao longo dos últimos anos, pois os métodos usados na tomada de decisão tradicional, que é baseada exclusivamente em critérios julgamentais e subjetivos, têm perdido espaço dentro das instituições. Os novos cenários demandam a busca por instrumentos mais eficazes para o gerenciamento da exposição ao risco de crédito, visando minimizar perdas. [2]

As técnicas de classificação de padrões surgem nesse meio como uma alternativa para suprir a necessidade das instituições financeiras para tratar essas informações de maneira mais probabilística e generalista, evitando julgamentos individuais e subjetivos que desconsideram o histórico da empresa com relação a concessões de crédito. É importante salientar que, com base nas regras de decisão tradicionais, as chances de eventuais candidatos à inadimplência serem classificados de maneira distinta são relativamente altas. Entretanto, quando se efetua uma análise multivariada das variáveis junto à classificação automática baseada no histórico de indivíduos em situações semelhantes, o risco de classificar erroneamente pode diminuir consideravelmente. Neste contexto, o intuito deste trabalho é implementar, analisar e avaliar alguns dos principais algoritmos de aprendizado de máquina aplicados à análise de risco de crédito, considerando-se as tarefas de seleção de características e classificação.

A principal contribuição deste trabalho é fornecer a um gestor uma ampla análise com indicações dos métodos potencialmente mais adequados à automatização da classificação dos clientes na tarefa de análise de risco de crédito. O enfoque desta proposta é o auxílio à tomada de decisão, dada a diversidade de algoritmos avaliados quanto à eficiência para classificação de clientes. Vale ressaltar que, considerando-se a amplitude das nossas pesquisas, percebemos uma escassez de trabalhos na literatura que realizem a análise e avaliação

desta quantidade de combinações de métodos de seleção de características e classificação, aplicados à análise de risco de crédito.

Este trabalho está organizado como descrito a seguir. A Seção II apresenta os trabalhos relacionados. A Seção III apresenta uma breve descrição dos métodos de aprendizado de máquina utilizados. A Seção IV aborda o pré-processamento dos dados. A Seção V contém a metodologia utilizada nos experimentos. A Seção VI apresenta os resultados e discussões a cerca dos experimentos. Por fim, na Seção VII é apresentada a conclusão deste trabalho.

II. TRABALHOS RELACIONADOS

Trabalhos direcionados à análise de risco de crédito através da aplicação de técnicas de inteligência artificial não são novidade. Entretanto, devido ao conservadorismo inerente às instituições financeiras, continua havendo certa resistência por parte das mesmas para aplicação destas pesquisas em ambiente real. Estas dificuldades acabam por desacelerar algumas pesquisas, o que de certa forma abre campo para exploração. A seguir, serão abordados trabalhos aplicados nesta área.

A dissertação em [3] aborda o desenvolvimento de 3 algoritmos para classificação de clientes na área de crédito, ao qual são usadas as técnicas de Regressão Logística, Redes Neurais e Algoritmos Genéticos, com a justificativa de que qualquer avanço nas técnicas de classificação que resulte numa melhora de precisão para um modelo de previsão, acarretará em ganhos financeiros para a instituição.

A pesquisa proposta por [4] buscou Avaliar o Risco de Crédito utilizando alguns grupos de classificadores, os modelos utilizados foram: Regressão Logística, rede Bayesiana Ingênua, SVM, Árvore de decisão e aprendizagem baseada em instâncias. Caracterizou-se as principais vantagens e desvantagens dos modelos, concluiu que o grau de acerto das predições usando um grupo de classificadores ao invés de um único modelo de classificação propicia resultados superiores ao melhor classificador individual.

O artigo em [5] apresentou um novo modelo de classificação para avaliação de risco de crédito, baseado em pesquisas das quais afirmavam que conjuntos de classificadores são superiores a classificadores únicos, o artigo propõe um novo modelo de classificação usando da técnica de sobreamostragem (*SMOTE*) ao qual é usada para mitigar os efeitos negativos de conjuntos de dados desequilibrados ao rebalancear o conjunto de dados de treinamento, combinado com o algoritmo de otimização enxame de partículas, que tem como função pesquisar os pesos e desvios junto com a rede neural.

O trabalho proposto por [6] demonstra que a classificação voltada para problemas de análise de crédito apresentam muitos infortúnios, dado que os resultados sofrem interferência de fatores como desbalanceamento de uma das classes do conjunto de dados, desequilíbrio de dados nas amostras, dentre outros empecilhos. Portanto, para maximizar os resultados da classificação, o mesmo propôs a utilização de um algoritmo de reamostragem combinado baseado em um modelo de mistura gaussiana. Funcionando da seguinte maneira: A partir do fator

de amostragem, determina o número de amostras da classe majoritária e da classe minoritária, em seguida, o agrupamento da mistura gaussiana é usado para subamostragem da maioria das amostras, e a técnica de sobreamostragem minoritária sintética é usada para o resto das amostras, a fim de eliminar qualquer problema de desequilíbrio.

O artigo em [7] propõe uma análise do método vizinho mais próximo ponderado, em que o mesmo foi aplicado na avaliação de crédito. Para os experimentos utilizaram uma base de dados de um banco privado da Indonésia. Com o intuito de analisar o desempenho de algumas funções do *kernel*, tais como: retangular, triangular, gaussiano, *epanechnikov*, *triweight* e inversão. A pesquisa obteve como resultado que a utilização do *kernel* gaussiano é superior as demais.

A pesquisa proposta por [8] tem o objetivo de fazer o uso combinado de pontuação de crédito e da pontuação de lucro para aumentar a eficácia do processo de concessão de empréstimos em cooperativas de crédito. O artigo revela que o uso de métodos estatísticos melhoram significativamente a previsibilidade do padrão quando comparado ao uso de técnicas subjetivas. Além de demonstrar a superioridade do modelo *random forest* em estimar pontuação de crédito e de lucro em comparação com o método de regressão logística. Salienta-se que para fazer essa análise o trabalho utilizou uma base de dados de uma cooperativa de crédito brasileira.

Foram organizados na Tabela I alguns resultados encontrados em trabalhos relacionados a análise de risco de crédito. Contudo, só foram listados os resultados que utilizavam a base “*German Credit*”, que é a base utilizada neste trabalho.

Tabela I
COMPARAÇÃO DE RESULTADOS DA LITERATURA PARA PREVISÃO DE CRÉDITO PARA A BASE *German Credit*.

ordem	autor	método	Acurácia (%)
1	[5]	RNA+PSO	78,70
2	[9]	SVM (RBFK)	73,60
3	[10]	MLP	73,93
4	[11]	MLP	75,35
5	[12]	ifair	73,00
6	[6]	mistura gaussiana	75,21

III. CLASSIFICAÇÃO DE PADRÕES

A tarefa de classificar padrões é realizada baseada em um conjunto de dados contendo observações (atributos) e cujas categorias (classes) são conhecidas. Existem diversos algoritmos para classificação de padrões, os classificadores utilizados neste trabalho serão abordados nas próximas subseções.

A. *Naive Bayes*

O teorema de Bayes criado por Thomas Bayes no séc. XVIII foi a inspiração para elaboração do algoritmo *Naive Bayes*, que se tornou um classificador probabilístico popular na área de Aprendizado de máquina. O método *Naive Bayes* é um algoritmo de aprendizagem supervisionada baseados na aplicação do teorema de *Bayes* com a suposição “ingênua” de independência condicional entre cada par de características dado o valor da variável de classe [13].

B. Logistic Regression

A Regressão Logística (inglês: *Logistic regression*) é um dos algoritmos de *machine learning* mais conhecido e utilizado no mundo, este método trabalha com os conceitos de estatística e probabilidade. De forma técnica a Regressão Logística mede a relação entre a variável dependente categórica e uma ou mais variáveis independentes, estimando as probabilidades com base em uma função logística [14].

C. K-nearest-neighbors

O algoritmo K- vizinhos mais próximos (inglês: *K-nearest-neighbors* ou *KNN*) é um algoritmo de aprendizado de máquina supervisionado simples e fácil de se implementar que pode ser usado para resolver problemas de classificação e regressão. O aprendizado baseado em instâncias é um método não-paramétrico, que consiste em: Dado um banco de dados, com exemplos já rotulados e um novo elemento desconhecido, esse elemento assume as características dos K exemplos que estejam mais próximos a ele. Ou seja, quanto mais próximos os exemplos do elemento, maiores são as semelhanças, consequentemente a vizinhança do elemento passa a descrevê-lo [14].

D. Support vector machine

Uma máquina de vetores de suporte (inglês: *support vector machine* ou *SVM*) é um classificador linear binário não probabilístico que possui um aprendizado supervisionado que analisa os dados e reconhece padrões. O algoritmo de treinamento do SVM constrói um modelo que atribui novos exemplos a uma categoria ou outra. Portanto o modelo representa os exemplos como pontos no espaço, mapeados de maneira que os exemplos de cada categoria sejam divididos por um espaço claro que seja tão amplo quanto possível. O que uma SVM faz é encontrar uma linha de separação, mais comumente chamada de hiperplano entre dados de duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes [15].

E. Decision Tree

Árvore de decisão (inglês: *Decision tree*) é um método de aprendizado supervisionado não paramétrico usado em classificações. Tem como objetivo criar um modelo que preveja o valor de uma variável de destino, aprendendo regras de decisão simples inferidas dos recursos de dados. Uma árvore pode ser vista como uma aproximação constante por partes. Ou seja, ela aprende com os dados ao se aproximar de uma curva senoidal com um conjunto de regras de decisão *if-then-else*. Quanto mais profunda a árvore, mais complexas são as regras de decisão e mais adequado é o modelo [16].

F. Bagging Decision Tree

Bagging Decision Tree é um algoritmo de aprendizagem supervisionada, que cria um conjunto de árvores de decisão usando *Bagging*. Sendo que *Bagging* (é um acrônimo para agregação de *bootstrap*), é entendido como um meta-algoritmo

da área de aprendizado de máquina, proposto por Breiman [17] em 1996 com o intuito de melhorar a precisão da predição.

O algoritmo *Bagging* gera várias amostras de *bootstraps* diferentes e de forma aleatória, baseado em amostragem que substitui o conjunto de dados original. Ou seja, cada instância selecionada pode ser repetida várias vezes na mesma amostra, consequentemente aumenta-se os dados de treinamento. Sendo a previsão final uma média de todos os modelos preditivos. É muito utilizado quando o objetivo é reduzir a variância dos classificadores individuais.

G. Boosting Decision Tree

Boosted Decision Tree é um algoritmo de aprendizagem supervisionada, que cria um conjunto de árvores de decisão usando *boosting*. Portanto, é um procedimento que agrega muitos classificadores “fracos” para alcançar um alto desempenho de classificação. Além disso, o *boosting* ajuda a estabilizar a resposta dos classificadores em relação às mudanças na amostra de treinamento. Já que o algoritmo aprende ajustando o resíduo das árvores que o precederam [18].

H. Random Forest

Criado por Tim Kan Ho [19] em 1995 as florestas aleatórias (inglês: *Random forest*) é um algoritmo que cria muitas árvores de decisão, de maneira aleatória, formando uma floresta, onde cada árvore será utilizada na escolha do resultado final. Florestas aleatórias surgem como meio de calcular a média de várias árvores de decisão profundas, treinadas em diferentes partes do mesmo conjunto de treinamento com o objetivo de reduzir a variância, essa redução ocorre com um custo que é um pequeno aumento do viés e perda de interpretabilidade, mas geralmente aumenta significativamente o desempenho no modelo final. Cabe ressaltar que esse modelo surgiu pois árvores de decisão que crescem muito fundo tendem a aprender padrões altamente irregulares, consequentemente elas se ajustam demais ao conjunto de treinamento, ou seja, tem baixa polarização e alta variância.

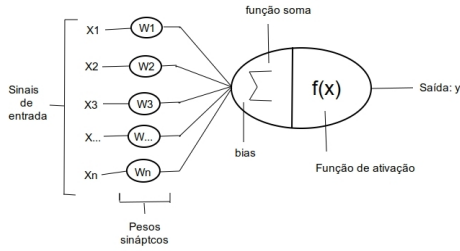
I. Voting Classification

O classificador de votação (inglês: *Voting classification*) é um modelo de aprendizado de máquina que treina um conjunto de vários modelos e prevê uma saída (classe) com base em sua probabilidade mais alta da classe escolhida como a saída. Ou seja, ele combina classificadores de aprendizado de máquina conceitualmente diferentes e usa um voto majoritário ou as probabilidades médias previstas (voto suave) para prever os rótulos das classes. Tal classificador pode ser útil para um conjunto de modelos de desempenho igualmente bom, a fim de equilibrar suas fraquezas individuais [20]. Portanto, optou-se neste trabalho por utilizar a combinação de 3 classificadores, sendo estes: KNN, SVM e *Random Forest*.

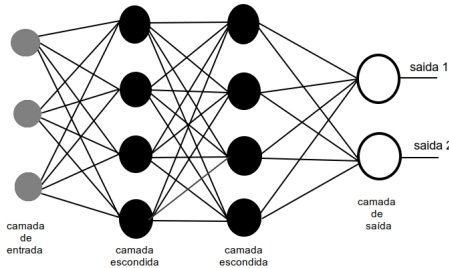
J. Neural Network

Redes neurais Artificiais (inglês: *Neural Network*) são modelos computacionais inspirados no sistema nervoso humano e que o intuito é aprender, podendo ser utilizadas em tarefas de

classificação. Em outras palavras, dado um problema, a RNA aprende quando acha uma possível solução geral (chamada de generalização), que é capaz de prever novos dados [21]. Uma rede neural é composta por vários neurônios como da Fig. 1 (a), conectados em camadas, que calculam determinadas funções matemáticas (conhecidas como funções de ativação) são organizadas de modo a ter camadas de entrada, ocultas e de saída, lembrando que uma RNA pode possuir uma ou múltiplas camadas ocultas, modelo de RNA na Fig. 1 (b) .



(a) Esquema de um neurônio.



(b) Esquema de uma rede neural.

Figura 1. Rede neural

1) *Multilayer perceptron*: Existem diversos modelos de Redes Neurais, para este trabalho utilizou-se as redes de múltiplas camadas (inglês: *Multilayer Perceptron — MLP*). Sendo um modelo que apresenta uma ou mais camadas de neurônios entre as camadas de entrada e de saída da rede, também conhecidas como intermediárias ou ocultas, e o seu treinamento é feito utilizando o algoritmo de Retropropagação (*Backpropagation*). Por conseguinte, o treinamento de uma rede MLP consiste em ajustar os pesos e os vies ou bias de suas unidades para que a classificação desejada seja obtida.

IV. PRÉ-PROCESSAMENTO DOS DADOS

O trabalho proposto em [22] mostra que o “Pré-processamento de dados” consiste na área que trabalha os dados para maximizar as possibilidades em se encontrar padrões úteis, sendo de grande importância no processo de mineração, pois um algoritmo aplicado a uma base de dados que contenha valores discrepantes e que não tenha passado por um pré-processamento pode identificar padrões de dados incongruentes.

A pesquisa em [23] complementa ao afirmar que para se obter todo o conhecimento disponível dentro de uma base de

dados, o primeiro passo é preparar essa base, a fim de eliminar redundâncias. A normalização dos dados, por exemplo, mantém os dados padronizados dentro de uma mesma faixa de valores, facilitando tarefas como aquelas pertencentes à área de aprendizado de máquina. A seguir serão abordados dois assuntos importantes na área de pré-processamento, que são: Normalização e Seleção de características.

A. Normalização

A normalização trata-se de um dos passos mais básicos quando se quer trabalhar com bases de dados em que seus atributos numéricos estejam em faixas numéricas diferentes, e conseqüentemente, um atributo possa sobrepor outro, gerando valores discrepantes ou até mesmo impossibilitando que o classificador obtenha resultados precisos. Existem diversas métricas de normalização para uma base de dados, a escolhida neste projeto é dada por (1) que mantém a faixa de todos os atributos entre [-1, 1].

$$x_{ij} = 2 * \left(\frac{x_{ij} - x_{j_{min}}}{x_{j_{max}} - x_{j_{min}}} \right) - 1, \quad (1)$$

onde x_{ij} é o valor do atributo j para o cliente i , $x_{j_{min}}$ e $x_{j_{max}}$ respectivamente são: o valor mínimo e máximo encontrado no atributo j .

B. Seleção de Características

Selecionar características é uma tarefa extremamente necessária quando se lida com bases de dados, já que possui inúmeras informações e muitas delas podem ser redundantes. O trabalho proposto por [24] aborda o fato de que as bases de dados por muitas vezes podem conter atributos irrelevantes bem como um número reduzido de amostras, provocando um aumento de complexidade computacional e a perda de exatidão na tarefa de classificação, a partir disto, a seleção de características, surge para minimizar o problema, e tornar o algoritmo mais eficiente, dado que remove atributos irrelevantes para a classificação.

Neste trabalho foram utilizado 3 métodos de seleção de características: *F-score*, Coeficiente de correlação de Pearson, e um Algoritmo Genético.

1) *F-score*: O método de seleção *F-Score* (*Fisher score*) é simples de implementar e também eficiente, pois consegue medir a relevância de cada atributo para as classes, sendo uma análise univariada dos atributos. Considerando um problema de classificação binário, onde as classes são C_1 e C_2 , a fórmula é definida por:

$$f(i) = \frac{(\mu_i^{C_1} - \mu_i) + (\mu_i^{C_2} - \mu_i)}{\sigma_i^{C_1} + \sigma_i^{C_2}}, \quad (2)$$

onde μ_i^C e σ_i^C correspondem, respectivamente, à média e o desvio padrão para a classe C com relação à i -ésima característica.

2) *Coefficiente de correlação de Pearson* : O coeficiente de Correlação de Pearson pode ser definido como um filtro univariado para seleção de características, dado que ele gera um ranking que pontua cada atributo de acordo com sua capacidade individual de discriminar duas classes. Com base no ranking gerado, pode-se eliminar os atributos com as piores pontuações do ranking, através da definição de um limiar. O coeficiente é dado pela fórmula:

$$C(j) = \frac{\sum_{i=1}^p (x_{ij} - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_{x_j} \cdot \sigma_y}, \quad (3)$$

onde x_{ij} é o valor do atributo na j -ésima característica e no i -ésimo padrão de entrada, \bar{x} , \bar{y} , σ_x e σ_y são respectivamente as médias e desvios-padrão de x e y .

C. Algoritmo Genético

O algoritmo Genético (inglês: *genetic algorithm* - GA) foi proposto na década de 1960 por John H. Holland um pesquisador da Universidade de Michigan, com o intuito de otimizar sistemas complexos. O trabalho [25] descreve o GA como sendo um algoritmo matemático inspirado no princípio Darwiniano, onde os mais aptos tendem a sobreviver e reproduzir. Com base nisso, o algoritmo é um mecanismo de busca adaptativa, que leva em consideração os princípios de seleção natural e recombinação genética.

Por ser um algoritmo muito versátil tem inúmeras utilidades para as mais diversas áreas, uma delas é na seleção de características, ao fazer uma análise multivariada dos atributos, com o intuito de selecionar apenas características relevantes à classificação.

Cada iteração do algoritmo genético corresponde à aplicação de um conjunto de quatro operações básicas: cálculo de aptidão, seleção, cruzamento e mutação. A implementação das duas últimas operações citadas anteriormente, podem ser vistas na Fig. 2. Para a operação de seleção foi utilizado o método da roleta. Ao final de cada iteração, também chamada de geração, cria-se uma nova população que tende a representar uma melhor aproximação da solução do problema de otimização do que a população anterior.

V. METODOLOGIA

Os algoritmos avaliados neste trabalho foram elaborados em linguagem Python, utilizando o IDE (ambiente de desenvolvimento integrado) Spyder 3.7, sistema operacional Ubuntu 18.04 Lts e processados em um computador com as seguintes configurações: Intel(R) Core(TM) i5-7200U CPU 2.50GHz, 8Gb de RAM, 64-bit.

A. Base de dados

A base de dados utilizada nesse estudo foi a *German Credit*. Esta base foi obtida no repositório UCI [26] elaborada pelo professor Hofmann [27] na cidade de Hamburgo e consiste em uma base de dados real com informações de 1000 clientes de uma instituição bancária. Esta base de dados, foi originalmente disponibilizada contendo atributos categóricos e numéricos, contudo, a Universidade Strathclyde produziu uma versão

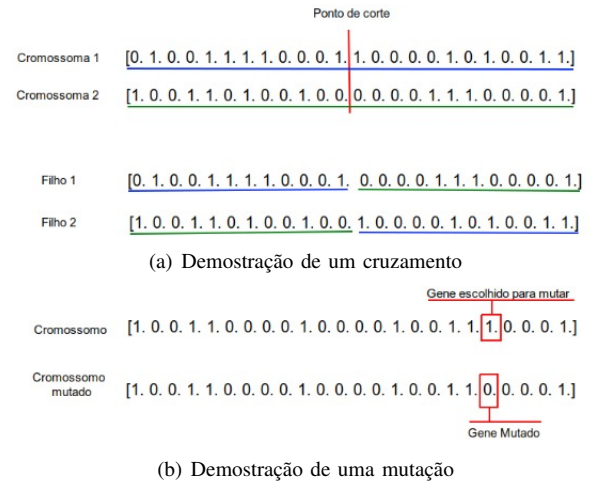


Figura 2. Algoritmo Genético implementado

totalmente numérica desse conjunto de dados que também foi disponibilizada no mesmo local. Para facilitar a classificação foi utilizada a base de dados modificada que contém apenas atributos numéricos e não há dados faltantes.

O perfil de cada um destes clientes é definido através de 24 variáveis numéricas e uma variável de rótulo, indicando se este cliente se tornou inadimplente ou não. Vale ressaltar que esta é uma base desbalanceada, contendo 30% de inadimplentes e 70% adimplentes.

B. pré-processamento e seleção de características aplicado a base

Inicialmente, com intuito de padronizar todas as variáveis da base de dados para uma escala comum, utilizou-se (1) para efetuar a normalização da base, deixando todas as variáveis na faixa [-1, 1]. Com relação à variável rótulo a base de dados foi alterada da seguinte forma: uma das classes (classe de inadimplentes) que era representada pelo valor 2, foi alterada para o valor -1.

1) *Seleção de Características*: Nesta etapa, utilizou-se dois filtros univariados (*F-score* e *Pearson*) e um método multivariado (Algoritmo Genético) para efetuar a seleção de características.

Ao aplicar o filtro *F-score* na base de dados, foi obtida uma pontuação, que pode ser observada na Fig. 3. No topo de cada coluna encontra-se o número de identificação do atributo, que está listado na segunda coluna da Tabela III. Pode-se observar na Fig. 3 que os 6 melhores atributos se destacam, dado esse fato, os mesmos foram escolhidos de forma empírica para serem utilizados na classificação.

O método Coeficiente de Pearson foi aplicado na base de dados e gerou um ranking, que pode ser visto na Fig. 4. Com o auxílio da segunda coluna da Tabela III é possível identificar todos os atributos, pois no topo de cada coluna encontra-se a numeração que identifica o mesmo. Nota-se que, para o Coeficiente de Pearson os 4 primeiros atributos se destacam dos demais, portanto, os mesmos foram escolhidos de forma empírica para serem utilizados pelos classificadores.

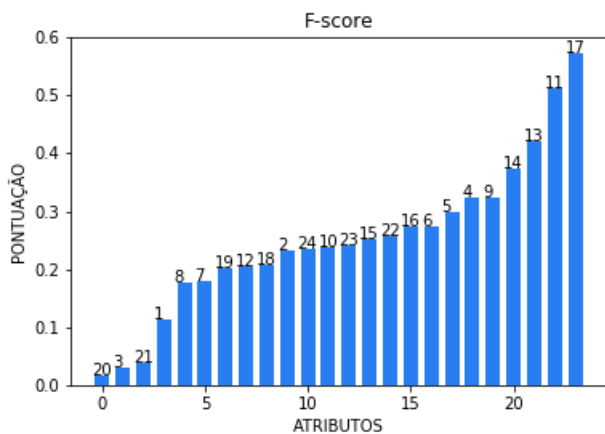


Figura 3. Pontuação gerada pelo F-score.

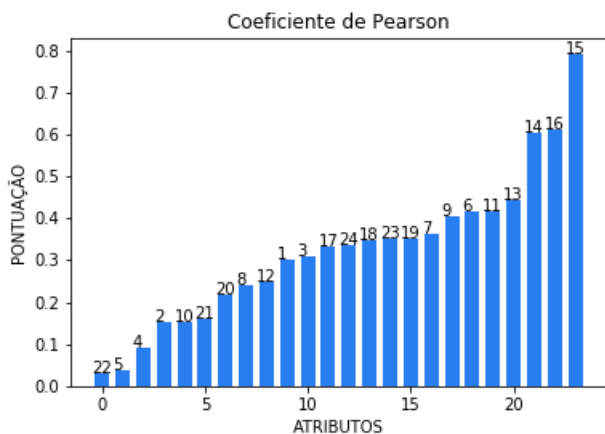


Figura 4. Pontuação gerada pelo Coeficiente de Pearson.

Para o presente trabalho, o GA foi projetado para encontrar a melhor combinação de atributos para cada método aplicado, fazendo uma análise multivariada, a fim de otimizar a função objetivo. Neste caso, temos: O vetor de atributos como sendo o indivíduo, a função objetivo assumindo a acurácia de cada modelo e foram utilizados 3 critérios de convergência para o GA (Atingir um dado numero de gerações, diferença entre melhor indivíduo, e a média da população, percentual 95% da população convergir para um mesmo valor), um deles sendo satisfeito, a melhor solução foi encontrada. O algoritmo genético foi aplicado utilizando os parâmetros definidos na Tabela II que foram baseados no trabalho [28].

A maioria dos métodos de aprendizado de máquina necessitam do ajuste de parâmetros ao serem implementados. A escolha dos parâmetros de vários dos métodos implementados neste trabalho foi feita com base na literatura. A segunda coluna das Tabelas VI, VII e VIII contem as referências aos trabalhos de onde os parâmetros foram obtidos para cada algoritmo, excluindo-se os classificadores *Naive Bayes* e *Voting* que não possuem parâmetros a serem ajustados. Este último, é

composto pelos classificadores KNN, SVM e *Random Forest*, já parametrizados com base em outros trabalhos da literatura.

Tabela II
PARÂMETROS ESCOLHIDOS PARA O GA.

População	Chance de Mutação	Chance de Cruzamento
30	0,01	0,80

VI. EXPERIMENTOS

Os experimentos apresentados nesta seção englobam a combinação dos métodos de seleção de características e os 10 algoritmos de classificação implementados.

A. Resultados

A divisão dos dados de treinamento e teste foi executada mediante a utilização do método de validação cruzada *k-fold*. O valor escolhido foi $k = 10$, seguindo diversos trabalhos da literatura.

Com o intuito de se ter uma visão prévia de quais variáveis possuem maior relevância para classificação na base "*German Credit*", os filtros univariados *F-score* e Coeficiente de Pearson foram aplicados. Gerando uma pontuação capaz de auxiliar na seleção de características. Contudo, isso nos traz apenas uma visão inicial e individualizada, uma vez que estes filtros não levam em consideração a informação gerada pela combinação de atributos.

Para que seja possível uma análise da relevância dos atributos considerando-os de maneira combinada, um Algoritmo Genético foi implementado e ajustado para realizar a seleção de características de forma multivariada. A Tabela III apresenta todos os atributos selecionados pelo GA para cada modelo de classificação. Diferente dos filtros *F-score* e Coeficiente de Pearson o Algoritmo Genético não gera um ranking, consequentemente, não é possível determinar quais os melhores atributos. Para definir os atributos selecionados pelo GA, ele foi executado 10 vezes para cada classificador, de forma que, a solução final foi definida pelos atributos encontrados em pelo menos 50% das execuções. A Tabela III está organizada da seguinte maneira: a coluna 1 contem o ID do atributo, a coluna 2 descreve o atributo, e as colunas de 3 a 12 contêm um valor binário, onde o valor 1 informa que o atributo foi selecionado pelo GA considerando-se o classificador indicado no cabeçalho da coluna e 0 caso o atributo não esteja entre os selecionados.

A Tabela IV mostra a quantidade de atributos selecionados por cada método de seleção de características para cada classificador utilizado no trabalho.

Com o intuito de facilitar a visualização e identificação dos resultados contidos nas Tabelas V, VI, VII e VIII destacou-se em negrito os melhores resultados.

A Tabela V foi montada com o intuito de propiciar uma análise comparativa entre os filtros univariados, bem como a combinação dos dois filtros utilizados. O melhor resultado entre eles será confrontado posteriormente com o Algoritmo Genético, que é um método multivariado.

Tabela III
ATRIBUTOS SELECIONADOS PELO GA PARA CADA MÉTODO UTILIZADO.

n°	Atributos	KNN	Bayes	R. Logística	A. decisão	SVM	R. forest	bagging	boosting	Voting	MLP
1	Verificação saldo	1	1	1	0	1	1	1	1	1	1
2	Nº meses do empréstimo	1	1	1	0	1	0	1	0	1	0
3	Historico de credito	1	0	0	1	0	1	0	1	0	0
4	Crédito de imposto mín. alternativo - AMT	0	0	0	1	0	1	1	1	1	1
5	Saldo poupança	1	1	0	1	0	0	1	0	1	0
6	Trabalho Atual	1	0	0	0	0	0	1	0	1	0
7	Sexo	1	0	1	1	0	0	1	0	1	0
8	Tempo de residencia atual	1	1	1	0	1	0	1	0	1	0
9	Propriedade	0	0	0	0	1	1	1	1	1	1
10	Idade	0	1	1	1	0	0	0	0	0	0
11	Outros parcelamentos	0	0	1	0	0	0	1	0	1	0
12	Crédito em outro banco	0	0	0	1	0	1	1	1	1	1
13	Conta individual/conjunta	0	0	1	0	0	1	0	0	0	0
14	Telefone	0	0	1	1	1	1	0	1	0	1
15	Trabalho estrangeiro	1	0	1	1	1	1	1	1	1	1
16	Compra carro novo	0	1	1	1	0	0	1	0	1	0
17	Compra carro usado	0	0	1	1	1	1	1	1	1	1
18	devedor s/ avalista	0	0	1	0	1	1	1	1	1	1
19	Devedor c/ avalista	0	0	0	0	0	1	0	0	0	0
20	Aluguel casa	0	1	1	0	1	0	0	0	0	0
21	Possui casa	1	1	1	0	0	1	0	1	0	1
22	Desempregado	0	1	1	1	0	1	1	1	0	1
23	Trabalho informal	0	0	0	0	1	0	1	1	1	0
24	Trabalho formal	1	0	1	1	1	0	1	1	1	0

Tabela IV
NÚMERO DE ATRIBUTOS SELECIONADOS POR CADA MÉTODO DE SELEÇÃO DE CARACTERÍSTICAS.

Modelo	Nº atributos
Sem seleção (todos os modelos)	24
<i>F-score</i> (todos os modelos)	6
Pearson (todos os modelos)	4
<i>F-score</i> + Pearson (todos os modelos)	8
KNN (GA)	10
Bayes (GA)	9
R. Logística (GA)	16
SVM (GA)	11
Árvore de decisão (GA)	12
Random forest (GA)	13
Bagging (GA)	17
Boosting (GA)	13
Voting (GA)	16
MLP (GA)	10

Com o objetivo de se analisar os possíveis benefícios da seleção de características, todos os classificadores foram aplicados à base de dados sem nenhuma seleção. A Tabela VI mostra os resultados obtidos para os 10 classificadores.

A Tabela VII mostra os resultados obtidos para os 10 classificadores, usando na seleção de características a combinação dos dois filtros univariados, pois esta combinação se mostrou mais eficiente que cada filtro aplicado individualmente, como pode ser observado através da Tabela V.

Por fim, os 10 classificadores foram aplicados à base de dados após a seleção de características utilizando o Algoritmo Genético. A Tabela VIII mostra os resultados obtidos.

Tabela V
RESULTADOS OBTIDOS ENTRE OS FILTROS UNIVARIADOS.

Modelo	Acurácia		
	<i>F-score</i>	C. Pearson	<i>F-score</i> + C. Pearson
KNN	0,631	0,677	0,691
Bayes	0,687	0,645	0,689
R. logística	0,692	0,691	0,730
SVM	0,699	0,692	0,728
A. decisão	0,683	0,680	0,686
R. Forest	0,714	0,714	0,720
Bagging	0,715	0,716	0,740
Boosting	0,711	0,715	0,758
Voting	0,721	0,713	0,729
MLP	0,706	0,701	0,735

Tabela VI
RESULTADOS OBTIDOS SEM SELEÇÃO DE CARACTERÍSTICAS

Modelo	parâ- metros	Acurácia (%)	Desvio padrão	Máximo (%)	Mínimo (%)	Tempo médio(s)
KNN	[7]	0,711	0,032	0,719	0,688	0,012
Bayes	-	0,688	0,041	0,740	0,650	0,003
R. logística	[8]	0,752	0,033	0,780	0,699	0,010
SVM	[9]	0,738	0,029	0,760	0,700	0,129
A. decisão	[29]	0,699	0,027	0,719	0,650	0,004
R. Forest	[8]	0,767	0,037	0,810	0,740	3,312
Bagging	[30]	0,785	0,036	0,839	0,740	3,529
Boosting	[31]	0,739	0,034	0,798	0,710	0,827
Voting	-	0,746	0,042	0,800	0,709	0,191
MLP	[11]	0,740	0,031	0,794	0,700	0,879

B. Discussões

Analisando-se a Tabela V pode-se verificar que ao combinar os dois filtros obteve-se uma melhoria significativa na classificação com relação à utilização individualizada deles.

Tabela VII
RESULTADOS OBTIDOS COM SELEÇÃO DE CARACTERÍSTICAS (*F-score* +
COEFICIENTE DE PEARSON)

Modelo	parâ- metros	Acurácia (%)	Desvio padrão	Máximo (%)	Mínimo (%)	Tempo médio(s)
KNN	[7]	0,691	0,042	0,720	0,649	0,001
Bayes	-	0,689	0,031	0,711	0,644	0,002
R. logística	[8]	0,730	0,041	0,778	0,698	0,004
SVM	[9]	0,728	0,036	0,799	0,701	0,067
A. decisão	[29]	0,686	0,054	0,721	0,630	0,002
R. Forest	[8]	0,731	0,034	0,806	0,701	0,770
Bagging	[30]	0,740	0,043	0,811	0,699	1,760
Boosting	[31]	0,758	0,046	0,825	0,711	0,740
Voting	-	0,729	0,048	0,801	0,699	0,070
MLP	[11]	0,735	0,029	0,798	0,703	0,739

Tabela VIII
RESULTADOS OBTIDOS COM SELEÇÃO DE CARACTERÍSTICAS
(ALGORÍTIMO GENÉTICO)

Modelo	parâ- metros	Acurácia (%)	Desvio padrão	Máximo (%)	Mínimo (%)	Tempo médio(s)
KNN	[7]	0,731	0,038	0,800	0,670	0,003
Bayes	-	0,728	0,039	0,780	0,650	0,002
R. logística	[8]	0,785	0,032	0,839	0,719	0,032
SVM	[9]	0,759	0,028	0,810	0,709	0,069
A. decisão	[29]	0,706	0,050	0,770	0,640	0,003
R. Forest	[8]	0,766	0,039	0,869	0,729	1,467
Bagging	[30]	0,792	0,034	0,819	0,731	2,34
Boosting	[31]	0,765	0,038	0,829	0,728	0,807
Voting	-	0,771	0,033	0,801	0,719	0,095
MLP	[11]	0,762	0,044	0,831	0,711	0,369

Pegando como exemplo o *Boosting Decision Tree*, é possível verificar que o mesmo aumentou sua acurácia em aproximadamente 4,3% ao combinar os filtros.

Entretanto, ao compararmos as Tabelas VI e VII podemos observar que a seleção de características com os filtros univariados gerou maiores valores de acurácia em poucos casos, quando efetuamos a comparação com os resultados obtidos sem qualquer seleção de atributos. Contudo, observa-se que mesmo selecionando somente os atributos de maior destaque para combinação dos filtros, sendo relativamente um número pequeno de atributos, foi possível a obtenção de resultados satisfatórios. Portanto, podemos considerar que existem benefícios ao utilizar estes filtros, pois são resultados satisfatórios obtidos com baixo custo computacional. Além disso os filtros são de fácil implementação e reduzem o custo computacional na etapa de classificação, consistindo em uma alternativa viável para seleção de características em bases de dados da área análise de risco de crédito. O algoritmo de *Boosting Decision Tree*, por exemplo, aumentou sua acurácia em 1,9% e reduziu seu tempo médio de execução em 0,087 segundos quando comparamos os resultado sem seleção de características e com a seleção por combinação dos filtros *F-score* e Coeficiente de Pearson.

Fazendo uma análise dos resultados da Tabela VIII nota-se que o algoritmo genético foi um excelente método de seleção de características. Mesmo tendo um custo computacional

relativamente alto, vale a pena sua implementação para selecionar os melhores atributos da base de dados de acordo com cada classificador. Por culminar na melhoria significativa dos classificadores, seja na avaliação da acurácia quanto na redução do custo computacional na etapa de classificação.

Analisando por exemplo: o algoritmo *Bagging Decision Tree*, nota-se, que o mesmo aumentou sua acurácia em 0,7% e reduziu seu tempo médio de execução em 1,18 segundos quando se compara o resultado sem seleção de características e com seleção utilizando o GA. Observa-se o mesmo comportamento para os demais métodos, como o algoritmo *Boosting Decision Tree* que aumentou sua acurácia em 2,6% e reduziu seu tempo médio de execução em 0,020 segundos.

Ao comparar as Tabelas VI, VII e VIII verifica-se que os métodos tendem a melhorar seu desempenho a medida que se retira atributos redundantes da base de dados e que o comportamento dos classificadores tendem a se manter constantes no ranking, sofrendo variações mínimas de posição. Tanto que os 3 piores métodos (KNN, bayes, árvore de decisão), mesmo tendo melhorias nos resultados continuam ruins se comparados aos demais, permanecendo nas últimas 3 posições no decorrer dos experimentos. Observa-se que os métodos do tipo *ensemble* foram bastante regulares mantendo-se sempre entre os primeiros colocados. Nota-se que o melhor resultado geral (ver Tabela VIII) foi obtido pelo método *Bagging Decision Tree* com 79,2% de acurácia, seguido pela Regressão Logística com 78,5%, os demais métodos de *ensemble*, a rede MLP com 76,2%, e a SVM 75,9%.

Contudo, cabe ressaltar que mesmo os métodos do tipo *ensemble* obtendo os melhores resultados, os mesmos possuem os maiores custos computacionais. Enquanto que outros métodos com custo computacionais relativamente baixos também classificam bem. Consequentemente, para analisar a viabilidade do método, fatores como esse precisam ser considerados, a fim de determinar qual método melhor se encaixa as necessidades do gestor.

VII. CONCLUSÃO

As técnicas de aprendizado de máquina utilizadas neste trabalho demonstraram ser de grande valia para a tarefa de análise de risco de crédito. Sendo que todos os 10 classificadores implementados mostraram precisão semelhante a trabalhos da literatura, considerando-se a base de dados utilizada.

Foi possível verificar que, mesmo sendo um classificador simples, a Regressão Logística obteve um bom desempenho nos experimentos, o que pode indicar benefícios da seleção de características. Os métodos de *ensemble* obtiveram os melhores resultados de forma geral para o problemas em questão, seguidos pela rede MLP e SVM. Cabe ressaltar que, todos os métodos de modo geral obtiveram resultados semelhantes a vários trabalhos da literatura (ver Tabela I).

Os experimentos também demonstraram a efetividade dos métodos de seleção de características, quanto a redução do custo computacional e a melhoria significativa da classificação, onde foi possível verificar que o Algoritmo Genético foi o melhor método para seleção dos atributos.

O trabalho demonstrou que, as técnicas avaliadas são capazes de contribuir para uma maior compreensão do processo de concessão de crédito, auxiliando assim o processo de tomada de decisão do gestor. Desta forma, o gestor pode reduzir tempo e custos para traçar políticas ou estratégias que reduzam o nível de inadimplência dentro da instituição credora, através da utilização de técnicas de aprendizado de máquina.

Para trabalhos futuros, exploraremos o refinamento de hiper-parâmetros e da arquitetura da rede MLP e dos métodos *ensemble*, a fim de encontrar a melhor estrutura e configuração dos métodos para o problema descrito, bem como, implementação de métodos que tratem os problemas de desbalanceamento da base de dados. Adicionalmente, estenderemos os experimentos, adicionando outros classificadores baseados em aprendizado profundo (em inglês, *Deep Learning*) e bases de dados do mesmo domínio, além de implementar testes estatísticos para fundamentar melhor as comparações.

AGRADECIMENTOS

Os autores agradecem à UFVJM por todo o suporte prestado no desenvolvimento deste trabalho.

REFERÊNCIAS

- [1] E. P. Lemos, M. T. A. Steiner, and J. C. Nievola, "Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining," *Revista de Administração-RAUSP*, vol. 40, no. 3, pp. 225–234, 2005.
- [2] G. A. S. Brito, A. Assaf Neto, and L. J. Corrar, "Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil," *Revista contabilidade & finanças*, vol. 20, pp. 28–43, 2009.
- [3] E. B. Gonçalves, "Análise de risco de crédito com o uso de modelos de regressão logística, redes neurais e algoritmos genéticos," Ph.D. dissertation, Universidade de São Paulo, 2005.
- [4] R. H. de Andrade, "Avaliação de risco de crédito utilizando grupo de classificadores," Master's thesis, Universidade Federal de Minas Gerais, 2008.
- [5] F. Shen, X. Zhao, Z. Li, K. Li, and Z. Meng, "A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation," *Physica A: Statistical Mechanics and its Applications*, vol. 526, p. 121073, 2019.
- [6] X. Han, R. Cui, Y. Lan, Y. Kang, J. Deng, and N. Jia, "A gaussian mixture model based combined resampling algorithm for classification of imbalanced credit data sets," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 12, pp. 3687–3699, 2019.
- [7] M. Mukid, T. Widiharhi, A. Rusgiyono, and A. Prahutama, "Credit scoring analysis using weighted k nearest neighbor," in *Journal of Physics: Conference Series*, vol. 1025, no. 1. IOP Publishing, 2018, p. 012114.
- [8] D. A. V. de Paula, R. Artes, F. Ayres, and A. M. A. F. Minardi, "Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques," *RAUSP Management Journal*, 2019.
- [9] Z. Zhang, J. He, G. Gao, and Y. Tian, "Sparse multi-criteria optimization classifier for credit risk evaluation," *Soft Computing*, vol. 23, no. 9, pp. 3053–3066, 2019.
- [10] L. Nanni and A. Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring," *Expert systems with applications*, vol. 36, no. 2, pp. 3028–3033, 2009.
- [11] C.-F. Tsai and J.-W. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring," *Expert systems with applications*, vol. 34, no. 4, pp. 2639–2649, 2008.
- [12] P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *2019 IEEE 35th international conference on data engineering (icde)*. IEEE, 2019, pp. 1334–1345.
- [13] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1. Citeseer, 1998, pp. 41–48.
- [14] M. Svensén and C. M. Bishop, "Pattern recognition and machine learning," 2007.
- [15] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *the Journal of machine Learning research*, vol. 9, pp. 1871–1874, 2008.
- [16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.
- [17] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [18] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, pp. 3146–3154, 2017.
- [19] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.
- [20] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.
- [21] S. Haykin, *Redes neurais: princípios e prática*. Bookman Editora, 2007.
- [22] F. N. Bonifácio, "Comparação entre as redes neurais artificiais mlp, rbf e lvq na classificação de dados," *Paraná: Universidade Estadual do Oeste do Paraná*, 2010.
- [23] E. Sobrinho, J. Araújo, L. A. Guedes, and R. Francês, "Descoberta de conhecimento em uma base de dados de bilhetes de tarifação: Estudo de caso em telefonia celular," 2005.
- [24] H. P. Rocha and A. P. Braga, "Seleção clonal de características rankeadas por filtros univariados para classificação de tipos de leucemia aguda," *Proceedings Seminário Interno da disciplina de Técnicas Clássicas de Reconhecimento de Padrões*, p. 64, 2010.
- [25] M. A. C. Pacheco *et al.*, "Algoritmos genéticos: princípios e aplicações," *ICA: Laboratório de Inteligência Computacional Aplicada. Departamento de Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro. Fonte desconhecida*, p. 28, 1999.
- [26] D. Dua and C. Graff, "Uci - repositório de aprendizado de máquina," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [27] H. Hofmann, "German credit data," *UCI - Repositório de aprendizado de máquina*, 2000.
- [28] G. L. PAPPÁ, "Seleção de atributos utilizando algoritmos genéticos multiobjetivos," Ph.D. dissertation, Pontifícia Universidade Católica do Paraná, 2002.
- [29] S. Y. Sohn and J. W. Kim, "Decision tree-based technology credit scoring for start-up firms: Korean case," *Expert Systems with Applications*, vol. 39, no. 4, pp. 4007–4012, 2012.
- [30] D. Zhang, X. Zhou, S. C. Leung, and J. Zheng, "Vertical bagging decision trees model for credit scoring," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7838–7843, 2010.
- [31] G. Wang and J. Ma, "Study of corporate credit risk prediction based on integrating boosting and random subspace," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13 871–13 878, 2011.