Análise de Crédito Utilizando uma Abordagem de Mineração de Dados

Credit Analysis Utilizing a Data Mining Approach

Joyce Maria do Carmo de Sá[1] orcid.org/0000-0001-8224-1323
Iago Richard Rodrigues Silva[1] orcid.org/0000-0002-8242-9059
Raniel Gomes da Silva[1] orcid.org/0000-0003-4874-3447
Luís Gustavo Arcoverde Souto[1] orcid.org/0000-0002-0410-0151
Paloma Gabriela Santos Silva[1] orcid.org/0000-0003-1477-9986

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

E-mail do autor principal: Joyce de Sá jmcs@ecomp.poli.br

Resumo

O crédito é um instrumento aplicado para incrementar e facilitar a realização de vendas de bens e serviços. Ele é o responsável por grande parte dos resultados auferidos nas empresas e pelo desenvolvimento e crescimento da economia do país. No entanto, faz-se necessário uma rígida avaliação para onde este crédito deve ir, uma vez que, sendo aplicado para empresas ou pessoas erradas, o credenciado pode acumular prejuízos. Desta forma, este trabalho propõe uma abordagem utilizando Mineração de Dados para análise de crédito através da aplicação de algoritmos de Inteligência Computacional, proporcionando uma tomada de decisão mais assertiva no momento de concessão do crédito.

Palavras-Chave: Análise de Crédito; Mineração de Dados; Inteligência Computacional;

Abstract

Creditis an instrument used to increase and facilitate sales of goods and services. He is responsible for a great part of the results obtained in the companies and for the development and growth of the economy of the country. However, a rigid assessment is necessary to where this credit should go, since, being applied to companies or wrong people, the credentialed can accumulate losses. In this way, this work proposes an approach using Data Mining for credit analysis through the application of Computational Intelligence algorithms, providing a more assertive decision making at the moment of credit granting.

Key-words: *Credit Analysis; Data Mining; Computational Intelligence;*

 http://dx.doi.org/10.25286/repa.v3i3.967

1 Introdução

Segundo Ross, Westerfield e Jordan (2002), a concessão de crédito é motivada pela necessidade de estimular vendas, mas isso acarreta para empresa concessora custos de imobilização do capital, bem como o risco do cliente não pagar, por isso é necessário definir como conceder e como cobrar, ou seja, uma política de crédito **[1]**. Entretanto, é necessário saber que política de crédito possui melhores resultados. Para isso, é essencial uma análise minuciosa das possíveis variáveis que venham a influenciar o bom do ruim credenciado.

Recentemente as necessidades dos clientes e a economia nacional têm sofrido diversas alterações. O processo de mudança de atitude, no que tange o crédito para pessoas físicas, dos tempos da inflação elevada para o momento de estabilidade, gerou uma desorientação para as pessoas e instituições financeiras, acarretando um aumento considerável na inadimplência **[2]**. Tal fato direcionou a necessidade de se ter, a cada dia, critérios mais precisos para a análise e concessão de crédito. Com os avanços tecnológicos foi possível verificar de diversas maneiras a análise do crédito. De acordo com Schrickel **[3]**, a análise de crédito envolve a habilidade de fazer uma decisão de crédito dentro de um cenário de incertezas e constantes mutações e transformações incompletas. Esta habilidade depende da capacidade de analisar logicamente situações complexas, e chegar a uma conclusão clara, prática e factível, de ser implementada.

Com a finalidade de obter uma análise mais rigorosa a pesquisa será guiada por metodologias de mineração de dados (*Data Mining*), como o CRISP-DM (*Cross Industry Standard Process for Data Mining*), o qual defende um modelo de processo que fornece uma estrutura para realização de projetos de mineração de dados que são independentes da indústria e da tecnologia utilizada, focando o descobrimento de padrões e regras significativos **[4]**. Pode-se descrever Mineração de Dados como parte do processo de descoberta de conhecimento em CRISP-DM, que tem por objetivo selecionar técnicas que serão utilizadas para localização de padrões nos dados, gerando por fim uma busca dos referidos padrões relacionados a um dado interesse **[5]**. Suas etapas podem apresentar-se de forma cognitiva, interativa e exploratória, compreendendo nos seguintes passos: entendimento do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e desenvolvimento **[6]**.

O presente trabalho tem por finalidade apresentar uma abordagem de análise de crédito através de Mineração de Dados, utilizando métodos de classificação para aprovação da concessão de crédito. O restante do trabalho está organizado da forma que segue. A seção 2 traz um embasamento teórico sobre análise de crédito e mineração de dados; a seção 3 apresenta e descreve a base, dicionário e alterações de dados juntamente com a parametrização das técnicas, além de descrever com detalhes os experimentos realizados durante a pesquisa; a seção 4 analisa os experimentos e apresenta conclusões.

2 Referencial Teórico

Nesta seção serão apresentados os principais temas que formam a base teórica para realização deste trabalho.

2.1 Análise de Crédito

Crédito é um conceito que está presente no cotidiano das pessoas e empresas, com o passar do tempo é perceptível uma maior necessidade da utilização desse conceito, que para Schrickel **[3]** crédito significa, "todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente, parte do seu patrimônio a um terceiro, com a expectativa de que esta parcela volte a sua posse integralmente, após decorrido o tempo estipulado". Silva **[7]** defende que a função do crédito consiste em avaliar a capacidade de pagamento do tomador, visando assegurar a reputação e a solidez do emprestador.

Com o aumento da procura por crédito, ocasionou também aumentou do índice de inadimplência, tornando-se necessário que as empresas buscassem ferramentas para auxiliar nas decisões de riscos, como a análise de crédito. Para que essas instituições possam mensurar o risco da concessão de crédito, faz-se necessário o uso de inteligência computacional (IC), que prover redução de custos, aumento de

produtividade, precisão e flexibilidade na operacionalização de mudanças na estratégia de concessão de crédito **[8]**, conseguindo análises mais precisas através das suas abordagens e técnicas de mineração de dados que podem extrair informações importantes de um conjunto de dados.

2.2 Mineração de Dados

Com a exacerbada quantidade de dados crescendo diariamente, responder uma questão tornou-se necessário **[9]**: O que fazer com os dados armazenados? As técnicas tradicionais de exploração de dados não são mais adequadas para tratar a grande maioria dos repositórios. Com a finalidade de responder a esta questão, foi proposta, no final da década de 80, a Mineração de Dados, do inglês *Data Mining*.

Para Fayyad et al. **[10]**. A extração de conhecimento de base de dados (mineração de dados) é o processo de identificação de padrões válidos, novos potencialmente úteis e compreensíveis embutidos nos dados. Portanto, mineração de dados nada mais é do que a procura de respostas para perguntas que ainda não existem em um grande volume de dados, extração de conhecimento, sabedoria.

2.2.1 CRISP-DM

Atualmente diversos processos definem e padronizam as fases e atividades da Mineração de Dados. Apesar das particularidades, todos em geral contém a mesma estrutura. Neste trabalho, escolhemos o CRISP-DM (*Cross-Industry Standard Processof Data Mining*) como modelo, devido à vasta literatura disponível e por atualmente ser considerado o padrão de maior aceitação.

O processo CRISP-DM consiste de seis fases organizadas de maneira cíclica, conforme mostra a figura abaixo. Além disto, apesar de ser composto por fases, o fluxo não é unidirecional, podendo ir e voltar entre as fases.

As fases do processo CRISP-DM são:

I. **Entendimento do Negócio:** Nessa etapa, o foco é entender qual o objetivo que se deseja atingir com a mineração de dados. O entendimento do negócio irá ajudar nas próximas etapas.

II. **Entendimento dos Dados:** as fontes fornecedoras dos dados podem vir de diversos locais e possuírem diversos formatos. Após definir os objetivos, é necessário conhecer os dados visando:

   a. Descrever de forma clara o problema;

   b. Identificar os dados relevantes para o problema em questão;

   c. Certificar-se de que as variáveis relevantes para o projeto não são interdependentes

   Normalmente as técnicas de agrupamento e de exploração visual também são utilizadas nesta etapa.

III. **Preparação dos Dados:** devido às diversas origens possíveis, é comum que os dados não estejam preparados para que os métodos de Mineração de Dados sejam aplicados diretamente. Dependendo da qualidade desses dados, algumas ações podem ser necessárias. Este processo de limpeza dos dados geralmente envolve filtrar, combinar e preencher valores vazios.

IV. **Modelagem:** é nesta fase que as técnicas (algoritmos) de mineração serão aplicadas. A escolha da(s) técnica(s) depende dos objetivos desejados.

V. **Avaliação:** considerada uma fase crítica do processo de mineração, nesta etapa é necessária a participação de especialistas nos dados, conhecedores do negócio e tomadores de decisão. Diversas ferramentas gráficas são utilizadas para a visualização e análise dos resultados (modelos). Testes e validações, visando obter a confiabilidade nos modelos, devem ser executados (*crossvalidation*, *suppliedtest set*, use *training set*, *percentage Split*).

VI. **Desenvolvimento:** Após executado o modelo com os dados reais e completos é necessário que os envolvidos conheçam os resultados.

http://dx.doi.org/10.25286/repa.v3i3.967

2.2.2 Classificação

A etapa de classificação pode ajudar no planejamento e na tomada de decisão, consiste em preparar os dados utilizados para treinamento, onde também será analisado o critério de parada que pode influenciar na qualidade final da previsão e testes. A classificação é aplicada na etapa da modelagem do CRISP-DM.

Uma abordagem geral para o aprendizado deste modelo consiste, primeiramente, em fornecer dados de treinamento, cujos resultados são conhecidos. Os dados de treinamento são então usados para gerar o modelo de classificação, que é posteriormente aplicado aos dados de teste, cujos resultados são desconhecidos. O objetivo é criar um modelo capaz de categorizar corretamente tanto os dados utilizados em seu treinamento, como dados nunca vistos antes, ou seja, um modelo com boa capacidade de generalização **[6]**.

As próximas subseções apresentam os algoritmos para classificação dos dados utilizados neste trabalho.

2.2.2.1 *NaiveBayes*

O *NaiveBayes* mostra ser uma ótima alternativa devido à sua utilidade para grandes volumes de dados e rapidez na execução quando comparados com outros algoritmos de classificação.

O *NaiveBayes* é uma técnica de aprendizado probabilístico supervisionado baseado no teorema de *Bayes* com uma suposição de independência entre os preditores. Basicamente, um classificador *NaiveBayes* assume que a presença de uma característica particular em uma classe não está relacionada com a presença de qualquer outro recurso.

2.2.2.2 *K-NearestNeighbors*

Os K vizinhos mais próximos ou (K-NN) é um algoritmo simples e é um classificador onde o aprendizado é baseado na analogia. O conjunto de treinamento é formado por vetores n-dimensionais e cada elemento deste conjunto representa um ponto no espaço n-dimensional. A ideia principal deste algoritmo é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento. Para determinar a classe de um elemento que não pertença ao conjunto de treinamento, o classificador K-NN procura K elementos do conjunto de treinamento que estejam mais próximos deste elemento desconhecido, ou seja, que tenham a menor distância. Estes K elementos são chamados de K-vizinhos mais próximos. Verifica-se quais são as classes desses K vizinhos e a classe mais frequente será atribuída à classe do elemento desconhecido. Dois pontos chaves que devem ser determinados para aplicação do K-NN são: a forma como se calcula a distância e o valor do K.

2.2.2.3 Regressão Logística

A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em detrimento de um conjunto de variáveis categóricas. Busca estimar a probabilidade da variável dependente dela assumir um determinado valor em função dos conhecidos de outras variáveis. Os resultados da análise ficam contidos no intervalo de zero a um.

O modelo de regressão logística (RL) é um modelo linear generalizado, sendo um tipo de análise de regressão muito utilizado para realizar predições ou explicar a ocorrência de um evento específico quando a variável dependente (variável resposta) é de natureza binária. Quanto às variáveis independentes, estas podem ser tanto quantitativas quanto qualitativas.

Por se tratar de um modelo linear generalizado, a RL apresenta três componentes: uma componente aleatória, que consiste em uma combinação das variáveis independentes (preditoras); um componente sistemático, que relaciona as variáveis independentes com os parâmetros do modelo correspondente à variável resposta que se quer modelar; e uma função de ligação. Como a regressão logística funciona com modelos preditivos binários, ela pode ser utilizada em classificação de dados cuja saída sejam binárias.

2.2.2.4 *Deep Learning*

Uma Rede Neural Artificial (RNA) consiste de vários neurônios simples conectados, cada um

produzindo uma sequência de valores de ativação. O aprendizado, ou a tarefa que a RNA se propõe a ser utilizada depende dos pesos encontrados que fazem a rede ter o comportamento desejado. Dependendo do problema e como os neurônios estão conectados tais comportamentos podem precisar de grandes blocos computacionais com vários neurônios interligados, onde cada bloco realiza transformações, normalmente de forma não linear. O *Deep Learning* (DL) se propõe a, de forma precisa, encontrar os pesos para cada bloco ou camada de neurônios **[16]**.

Podem se caracterizar por DL redes neurais que: usam uma cascata de diversas camadas, com unidades de processamento, normalmente, não-linear para a extração e transformação de características; cada camada sucessiva usa a saída da camada anterior como entrada; são baseados na aprendizagem (supervisionada) de vários níveis de características ou representações dos dados; realizam parte de uma área da aprendizagem de máquina mais ampla que é a aprendizagem de representações de dados; aprendem vários níveis de representações que correspondem a diferentes níveis de abstração; os níveis formam uma hierarquia de conceitos **[17]**.

Durante o treinamento da rede neural pode ocorrer dois problemas distintos ligados a fase de treinamento, são eles o *overfitting* e *underfitting*. O *overfitting* é o treinamento excessivo, fazendo com que a rede memorize padrões da base de dados atual e perca sua capacidade de generalização com entradas de dados novas a rede. O *underfitting* é o treinamento insuficiente, fazendo com que a rede não aprenda os padrões e comportamentos e assim não possa generalizar para novos dados de entrada.

Sendo assim é necessário determinar um critério de parada, e um dos mais usados é a validação cruzada. Esta é a verificação da diferença entre a saída encontrada e a saída desejada, onde os pesos são inéditos a cada ciclo de validação. Enquanto o erro de validação estiver diminuindo, a rede continua treinando, isto é, no momento em que o erro da validação começar a aumentar e o de treinamento continuar a diminuir, a rede está começando a memorizar padrões, sendo este o ponto de parada para o treinamento.

2.3 Teste de *Kolmogorov-Smirnov*

O teste estatístico de *Kolmogorov-Smirnov* (KS) foi proposto pelos soviéticos A.N. Kolmogorov e N.V. Smirnov **[11]**, é uma técnica não paramétrica, usada para testar se duas amostras podem ser provenientes de uma mesma função de distribuição **[13]**. A estatística KS é definida como a máxima diferença entre as distribuições acumuladas dos scores dos "bons" e "maus" pagadores **[11]**. Pode ser definida pela equação 1:

$$KS = \max_{s} \left\{ \left| F_M(s) - F_B(s) \right| \right\} \quad (1)$$

O KS mede a máxima separação entre a frequência relativa acumulada de maus pagadores, Fm(s) e a frequência relativa acumulada de bons pagadores, Fb(s). Sob a hipótese que as distribuições sejam iguais, o p-valor indica se esta hipótese é rejeitada o não a um nível de significância.

3 Materiais e Métodos

A metodologia utilizada neste trabalho foi a CRISP-DM. Suas etapas serão descritas nas próximas subseções.

3.1 Entendimento do Negócio

Essa é a primeira etapa para buscar compreensão adequada do problema, onde se definem os objetivos do projeto de mineração de dados. Esta etapa também estabelece os critérios para definição e interpretação dos resultados obtidos do processo de mineração de dados.

A base de dados a ser utilizada contém informações sobre clientes que podem determinar na análise de crédito, identificando se o cliente é um bom ou mau pagador. Para esta abordagem foi escolhido o método de classificação como técnica de mineração de dados.

O problema proposto refere-se a avaliar se o cliente é um bom ou mau pagador para a concessão de crédito através da base de dados proposta. O objetivo deste trabalho é de aplicar técnicas de mineração de dados com diversos

http://dx.doi.org/10.25286/repa.v3i3.967

algoritmos de classificação a fim de encontrar o melhor desempenho.

3.2 Entendimento dos Dados

A base de dados foi obtida com a empresa Neurotech, a qual se dispôs a oferecer uma parte de sua real base de dados. A base de dados possui 176 atributos e 500000 instâncias e nela encontram-se informações de clientes que podem ou não serem bons para receberem o crédito. Os tipos de dados encontram-se da seguinte forma:

A. **Cadastrais:** dados relacionados ao local onde o indivíduo vive como classe social, vizinhança, entre outros.
B. **Demográficos:** comparações com a vizinhança como renda, por exemplo.
C. **Financeiros:** atividade do indivíduo como consumidor.
D. **Geográficos:** exposição do indivíduo em locais considerados importantes.
E. **Partidos:** possíveis filiações a partidos políticos.
F. **Programas Sociais:** o indivíduo faz parte de ONGs, bolsa família, prouni, etc.
G. **Riscos:** verifica a exposição do indivíduo a alguns fatores considerados de risco.
H. **Servidor:** verifica se o indivíduo faz parte de algum serviço militar ou público.
I. **Web:** analisa a exposição do indivíduo em sites na internet com temas pré-selecionados.

Informações como classe social do consumidor, renda da vizinhança, atividade do consumidor no mercado financeiro, exposição a endereço de hotéis, filiação política, *flag* bolsa família, exposição risco web, *flag* servidor militar, *flag* servidor civil, etc. podem ser levadas em consideração no momento de escolha para aplicação de crédito, inclusive o seu montante.

3.2.1 Análise Estatística Bivariada

O mercado está cada vez mais competitivo e não existe mais espaço para interpretações errôneas e/ou incompletas. Mais do que nunca é preciso a obtenção de informações sólidas para a tomada de decisões. Por este motivo a análise de dados torna-se essencial para qualquer negócio que almeja ter sucesso.

Após a realização da PCA verificamos que a variável **RENDA** tem uma grande correlação (aproximadamente 0.6) com o alvo e a partir disso foi realizada a sua análise. Ela é do tipo categórica**,** suas categorias são: "ATE 2 SM" , "2 A 4 SM","4 A 10 SM", "10 A 20 SM", "ACIMA DE 20 SM" ou "NULL" quando não há informações de renda para aquele indivíduo. Foram realizados os seguintes procedimentos:

A. **Agrupamento:** O agrupamento foi realizado para determinar a quantidade de 'indivíduos' por categoria de renda e como eles estão relacionados com o alvo (**INDICE_BOM_CLI**).
B. **Porcentagem Total:** É a porcentagem da quantidade de indivíduos por categoria.
C. **Porcentagem da Taxa de Maus:** Foi realizada para calcular a 'taxa de maus', variável que indica quantos indivíduos por categoria pertencem ao alvo = '1'.

A partir dessas informações foi possível gerar o gráfico apresentado na Figura 1. No eixo Y estão as variáveis que representam a PORCENTAGEM TOTAL, no eixo X estão as categorias, e a curva indica a taxa de maus em porcentagem, em relação à porcentagem total.

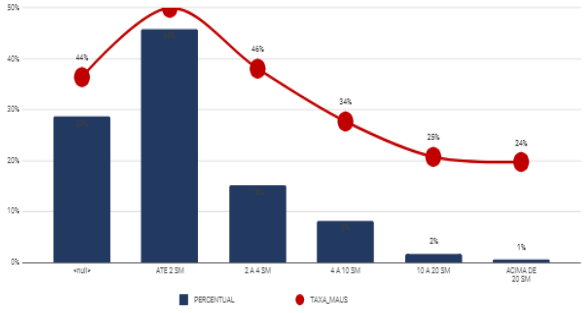


Figura 1 - Gráfico referente à análise bivariada da renda em relação ao alvo**.**

Após as transformações, formação de gráficos e análise, foi possível obter as seguintes conclusões:

- Quanto maior a renda do indivíduo, menor o 'risco' ou a probabilidade de ele ser um mal pagador (PERTENCER AO ALVO '1');
- A quantidade expressiva de nulos (143421) pode atrapalhar no desempenho do modelo final, uma forma de contornar essa situação é considerar que os

indivíduos dessa categoria se enquadram na categoria "ATÉ DOIS SM", realizar o cálculo da média entre essas variáveis.

Como os resultados obtidos após os processamentos não foram suficientemente relevantes, foi necessário o reajuste e utilização de outras técnicas de preenchimento de dados faltosos e pré-processamento. Para certificar que os dados estavam coerentes e obter um melhor entendimento do seu comportamento, após gerar o KS de cada uma foi realizada a análise bivariada. Com ela, foi possível visualizar outras formas de preencher os dados faltosos e como cada variável se comporta em relação ao alvo (**INDICE_BOM_CLI)**.

3.2.2 Análise de Correlação Entre as Variáveis

Para critério de análise dos atributos mais significativos, foi aplicada a técnica de correlação de Spearman. A técnica de spearman realiza a correlação entre duas variáveis para determinar as *features*mais relevantes da base de dados **[18]**. A execução de um algoritmo *featureselection*é de suma importância no processo de pré-processamento, pois é possível remover atributos que não trarão bons resultados na acurácia, além de diminuir o custo computacional para os algoritmos de classificação de dados.

Junto com a técnica de *Spearman*, é possível aliar recursos de visualização de dados, como o *heatmap*para que o cientista de dados consiga visualizar de maneira geral, a relevância dos atributos na base de dados. A Figura 2 apresenta o resultado da correlação de *Spearman* aplicado ao *heatmap*.

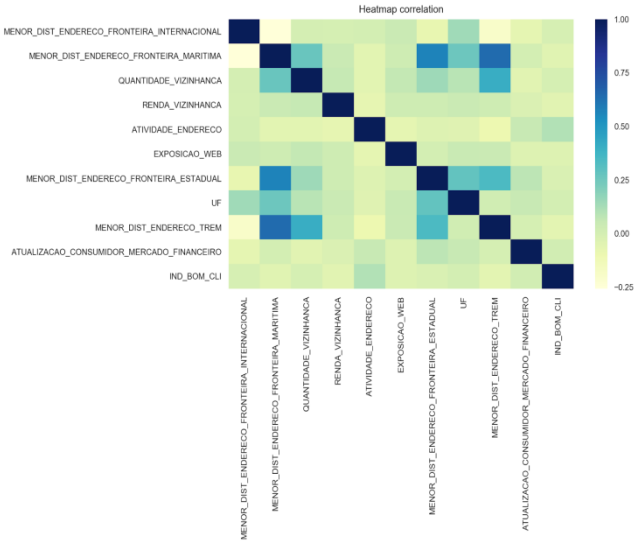


Figura 2 - *Heatmap* para as 10 primeiras variáveis mais relevantes.

3.2 Preparação da Base de Dados

Ao realizar a análise da base de dados utilizada no projeto, foram encontrados alguns problemas como dados faltosos, atributos contendo strings, alta amplitude dos valores e alta quantidade de atributos. Para a preparação dos dados, foram utilizados diversos softwares e plataformas, tais quais: Python, Excel e R, cada um utilizado para uma finalidade específica.

A resolução dos problemas citados serão discutidos nos pontos a seguir:

A. Tratamento de Dados Faltosos: Em 168 atributos faltaram dados a serem preenchidos, em todas as instâncias da base de dados esse problema foi detectado. Para resolução deste problema, foram analisados previamente os tipos de dados que compõem os atributos para posterior aplicação de um método de estimação de tendência central da base de dados, como mediana e moda. Nos espaços vazios correspondentes a dados categóricos que não poderiam conter dados faltosos foram aplicados a moda (16 atributos), nos dados categóricos que poderiam assumir dados faltosos foram preenchidos com zero (2 atributos), e nos dados contínuos foram aplicados à mediana (150). Apenas em

http://dx.doi.org/10.25286/repa.v3i3.967

oito atributos não foi feito o tratamento de dados faltosos.

B. Transformação de Dados: Alguns atributos contendo valores do tipo String estão presentes na base de dados, como por exemplo, Estados Brasileiros, Bancos, Classe Social, etc. Estes valores foram convertidos em dados numérico-categóricos para posterior execução dos algoritmos de classificação de dados, visto que estes utilizam dados numéricos (contínuos ou não) em suas execuções. Por exemplo, no atributo **UF**, que corresponde ao Estado do indivíduo analisado, os valores possíveis são as siglas correspondentes a eles e para transformação cada sigla recebeu um identificador único, de 1 a 27 (Unidades Federativas existentes no Brasil) e assim foram substituídos na base de dados. Processo igual a este foi realizados nos outros atributos, alterando apenas a quantidade de identificadores únicos.

Além disso, os dados da base não se encontram normalizados, existindo um desbalanceamento na amplitude dos valores dos atributos. Por exemplo, um dos atributos do tipo demográfico, que geralmente são nomeados com o prefixo **EXPOSICAO_ENDERECO** possuem valor mínimo igual a 0 e valor máximo de 2850 em uma de suas variáveis, enquanto o atributo **MENOR_DIST_BANCO** possui o mesmo valor mínimo, entretanto valor máximo igual a 998886. Isso ocorre na maioria dos atributos da base de dados, sendo necessário a aplicação de uma função de normalização em todos os atributos para que seus *ranges* tornem-se entre 0 e 1, esta função é descrita na equação 2.

$$F(i) = \frac{Xi - min(x)}{max(x) - min(x)} \quad \textbf{(2)}$$

Onde:F(i) = Valor a ser setado na cédula atual; Xi = Valor cédula atual; min(x) = menor valor do atributo; max(x) = maior valor do atributo.

Desta forma, com os dados normalizados, é possível executar os algoritmos de classificação sem a interferência de variáveis menos relevantes (pelo fato de ter uma amplitude maior) sobre as variáveis mais relevantes, proporcionando uma classificação estatisticamente mais confiável.

C. Seleção de atributos: A quantidade de atributos é de 176. A análise estatística de correlação com a aplicação do algoritmo PCA (*Principal Components Analysis*) torna-se essencial para redução de redundância e informações desnecessárias contidas na base de dados. Uma das vantagens que a redução de características pode proporcionar é uma mais rápida classificação e taxa de acurácia quase equivalente a que se obtém rodando o classificador na base original. Antes da aplicação do PCA, foi observado que o primeiro atributo da base de dados é um identificador (ID) assumindo-se chave primária da base de dados, o que é irrelevante para o processo de classificação, sendo esta excluída da execução dos experimentos subsequentes. Outra variável removida foi a variável demográfica**CAD_DEMOGRAFICO_VAR_35_1** porque a mesma estava duplicada.

Foi utilizada a técnica de *Kolmogorov-Smirnov*, a qual forneceu uma visão geral da base de dados e propôs uma melhor e menor seleção de variáveis contendo um número de 43 colunas. Para implementar esse teste estatístico, foi utilizado a linguagem python. O algoritmo foi aplicado na base de dados, gerando um csv com todas as variáveis e resultados do KS, sendo possível montar um ranking com as variáveis que possuem melhor valor.

D. Balanceamento das Classes: Em relação à quantidade de instâncias da classe 0 (aptos para receber o crédito) e da classe 1 (não aptos), pode-se afirmar que a base de dados encontra-se balanceada, não sendo necessária a aplicação de algum algoritmo para o balanceamento das classes. A classe 0 (aprovados na análise de crédito) possui 253804 instâncias, enquanto a classe 1 (não aprovados) possui 246196 instâncias.

3.4 Modelagem

Nesta seção serão apresentados os algoritmos utilizados para o processo de modelagem. A base de dados foi gerada através do teste KS, resultando em um ranking com as variáveis de maior correlação, sendo 4 bases de dados com 10, 20, 30 e 40 variáveis de maior correlação respectivamente. O conjunto de dados foi separado em 80% para treinamento e 20% para teste. Cada algoritmo foi executado 30 vezes para obtenção de um resultado estatisticamente mais confiável.

3.4.1 *Naive Bayes*

O algoritmo *Naive Bayes* foi utilizado com o auxílio da biblioteca *scikit-learn* para Python, utilizando suas configurações padrão. Para criação do modelo de classificação, foi utilizada a função Gaussian NB.

3.4.2 K-NN

Para execução do treinamento com KNN, foi utilizado com o auxílio da biblioteca *scikit-learn* para Python. O K-NN também possui a necessidade de inicialização de alguns parâmetros como a distância a ser utilizada e o valor K. A configuração do K para os experimentos foi de K=5, devido o número de classes ser par foi definido um número ímpar e maior que 3 para uma maior capacidade de consideração de vizinhos com características similares. Como o vetor de dimensões foi diminuído, foi utilizada a distância euclidiana para cálculo da distância de todos os pontos entre si.

3.4.3 Regressão Logística

A Regressão Logística foi aplicada utilizando a biblioteca *scikit-learn*para Python, utilizando suas configurações padrão, sendo que esta não oferece opções para personalização de parâmetros.

3.4.4 *Deep Learning*

Para o treinamento da *Deep Learning* (DL) é necessário definir os dados de entrada da rede neural, a quantidade de neurônios na camada de entrada, a quantidade de neurônios nas camadas intermediárias e na camada de saída, definir a taxa de conectividade, o número de ciclos do warmup, a função de ativação na camada intermediária e a equação para o cálculo do erro. A Tabela 1 apresenta os valores utilizados para os parâmetros citados.

Tabela 1 - Valores dos parâmetros utilizados na Deep Learning.

Parâmetro	Valor
Quantidade de Neurônios na Camada de Entrada	18
Quantidade deNeurônios nas Camadas Escondidas	Qtdneurônios entrada * 1.5 (progressivamente)
Quantidade de Neurônios na Camada de Saída	136,6875
Função de Ativaçãona Camada Intermediária	sigmoide

4 Resultados

Esta seção descreve os resultados para cada algoritmo utilizado para as bases de dados.

4.1 *Naive Bayes*

O Quadro 1 apresenta os resultados obtidos com a aplicação do *Naive Bayes.*

Quadro 1 - Resumo dos resultados obtidos com a aplicação do Naive Bayes.

BASE	ACURÁCIA (MÉDIA)	DESVIO PADRÃO
10	0,554776	0,001
20	0,56265	0,001
30	0,554799333	0,001
40	0,559786	0,01

Com a execução do *Naive Bayes*, foi possível observar que o número de atributos em relação às bases em estudo não possui nenhuma relevância para o modelo, onde há uma oscilação para mais ou menos conforme o número de atributos é aumentado. Este é outro modelo

http://dx.doi.org/10.25286/repa.v3i3.967

preciso, pois o desvio padrão do mesmo foi relativamente baixa, em contrapartida o modelo proporcionou também uma baixa acurácia.

4.2 K-NN

O Quadro 2 apresenta os resultados obtidos com a aplicação do NaiveBayes.

Quadro 2 - Resumo dos resultados obtidos com a aplicação do Naive Bayes.

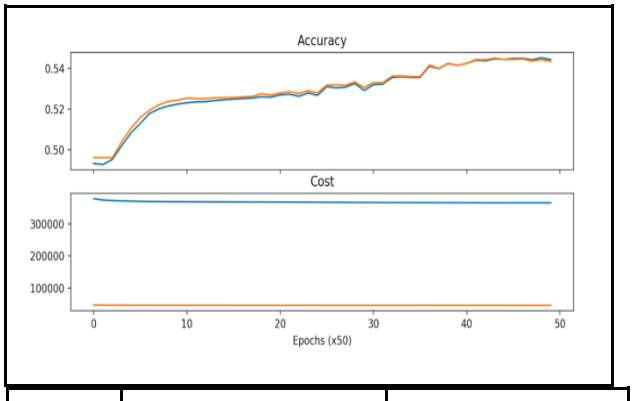
BASE	ACURÁCIA (MÉDIA)	DESVIO PADRÃO
10	0,5488	0,00112015393
20	0,550019	0,00114078119
30	0,55390	0,00116039911
40	0,55621	0,00116178119

O algoritmo K-NN seguiu o mesmo padrão em relação a acurácia dos outros modelos, permanecendo na faixa de 54% a 55%. O mesmo cenário de avaliação da acurácia dos outros modelos é também válida para este modelo. Semelhantemente a regressão logística, este algoritmo pode apresentar resultados melhores de acordo quando o número de atributos é aumentado.

4.3 Regressão Logística

O Quadro 3 apresenta os resultados obtidos com a aplicação da regressão Regressão Logística.

Quadro 1 - Resumo dos resultados obtidos com a aplicação da Regressão Logística

BASE	ACURÁCIA (MÉDIA)	DESVIO PADRÃO
		
10	0,551208666	0,001553385032
20	0,556811666	0,001319963645
30	0,562282	0,001521241326
40	0,576364	0,001419870319

Foi observado que o número de amostras proporcionou pouco impacto na acurácia do modelo, obtendo apenas um ganho de 2% na base de dados de 40 atributos, em relação à execução na base de 10 atributos. Foi observado que o desvio padrão foi considerado baixo, e o modelo apesar da baixa acurácia é preciso. De todos os modelos testados neste trabalho, este modelo proporcionou a maior acurácia, tendo aproximadamente 57,64% nesta métrica utilizando na base de dados com 40 atributos.

4.4 *Deep Learning*

As Figuras 3, 4, 5, 6 e 7 apresentam os resultados (em forma de gráficos) obtidos com a aplicação da *Deep Learning*.

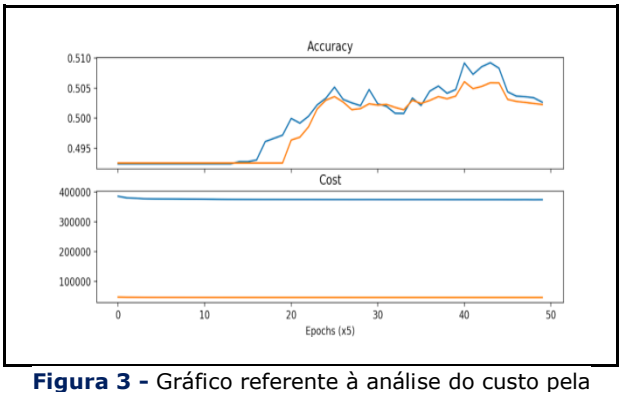


Figura 3 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando apenas 10 variáveis.

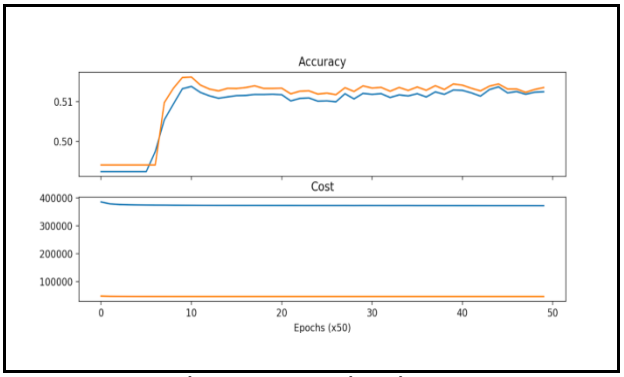


Figura 4 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando apenas 20 variáveis.

Figura 5 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando apenas 30 variáveis.

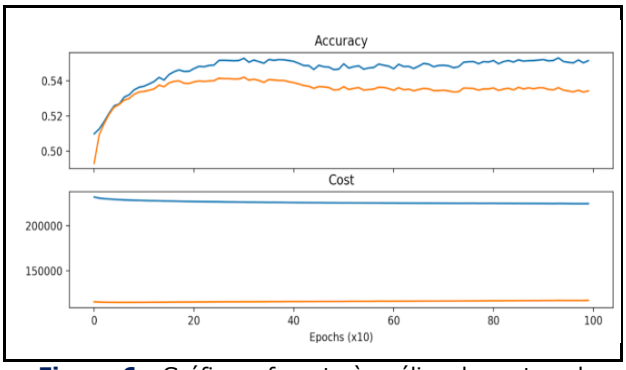


Figura 6 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando apenas 40 variáveis.

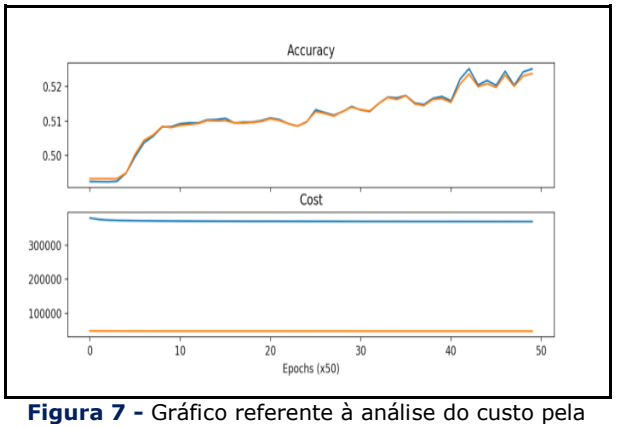


Figura 7 - Gráfico referente à análise do custo pela acurácia dos resultados utilizando todo o resto da base.

Portanto, o melhor resultado foi obtido utilizando a base de dados completa, como pode ser visto na Figura 5, pois além de obter o melhor resultado de acurácia foi o que obteve a melhor relação entre o custo com a acurácia.

5 Conclusão

Este trabalho apresentou uma abordagem utilizando aprendizado de máquina para extração de conhecimento em uma base de dados, com a finalidade de classificar indivíduos aptos ou não para receber crédito, sendo este um sistema para análise de crédito. Esta abordagem seguiu a metodologia CRISP-DM, onde foram testadas diversas abordagens para preenchimento de dados faltosos, seleção de atributos e modelagem dos experimentos.

Foram utilizados diversos algoritmos clássicos para criação de modelos de predição para análise de crédito, tais quais: NaiveBayes, Regressão Logística, Deep Learning e K-NN, sendo estes analisados observando a métrica de acurácia. Cada algoritmo foi executado 30 vezes, tendo sido coletados em cada execução o dado correspondente a acurácia, sendo este coletado e para análise foi calculado sua média e desvio padrão. De acordo com os resultados obtidos, não foi possível obter uma alta acurácia de classificação dos dados separados para teste (20% da base).

A baixa acurácia avaliada neste projeto, ocorreu pela baixa correlação das variáveis. A *feature*mais conceituada foi a de ATIVIDADE_EMAIL, cujo *ranking*, de acordo com Spearman, foi de 0.11. Para um algoritmo de classificação, esse valor é considerado extremamente baixo e com isso, torna-se inviável o reconhecimento de padrões **[19]**.

Apesar das inconsistências identificadas nessa base de dados, é possível obter resultados mais significativos através do Deep Learning, mas será necessário aplicar alguns experimentos, como monitorar a acurácia a partir do acréscimo e remoção das *hiddenlayers* e neurônios, adicionar mais dados para que os neurônios consigam aprender mais rapidamente e avaliar outros algoritmos do *Framework Tensorflow*.

Referências

[1] ROSS, Stephen A. et al. **Administração financeira**. São Paulo: Editora Atlas, 1995.

[2] ALMEIDA, Hamilton. Políticas econômicas serão iguais até 95. **Zero Hora**, Porto Alegre, 24 mai 1992.

[3] SCHRICKEL, W. K. **Análise de crédito**: Concessão e Gerência de Empréstimos, São Paulo: Atlas, 1994.

[4] BERRY, Michael J. A.; LINOFF, **Gordon. Data mining techniques**. USA : Wiley Computer Publishing, 1997.

http://dx.doi.org/10.25286/repa.v3i3.967

[5] NARENDRAN, C. R. Data Mining-Classification Algorithm–Evaluation. May 8th, 2009.

[6] WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: International conference on the practical applications of knowledge discovery and data mining, 4., 2000, Citeseer. **Proceedings**... Citesser, 2000. p. 29-39.

[7] SILVA, José Pereira da. **Gestão e análise de risco de crédito**. 6 ed. São Paulo: Atlas, 2008.

[8] YU, L.; WANG, S.; LAI, K. K.; ZHOU, L. **Bio-Inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machines**. Berlim, Heidelberg: Springer-Verlag, 2008.

[9] LAROSE, Daniel T. **Discovering Knowledge in Data**: An Introduction to Data Mining. New Jersey: John Wiley & Sons, Inc, 2005.

[10] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge Discovery and Data Mining: towards a unifying framework. In: FELIX, Priscila et al. **Proceedings of second international conference on electrical system , tecnology and informacion.** Lecture Notes in Electrical Engineering, v.365. Berlim: Springer, 2015. p. 82-88.

[11] ANDERSON, R. **The Credit Scoring Toolkit**: Theory and Practice for Retail Credit Risk Management and Decision Automation. New York: Oxford University Press, 2007.

[12] ARSENAULT, MARC-OLIVER. Kolmogorov-Smirnov tes: a needed tool in your data science toolbox. **Towards data science**. 21 Nov. 2017. Disponível em: <https://towardsdatascience.com/kolmogorov-smirnov-test-84c92fb4158d>. Acesso em: 5 jun. 2018.

[13] CONOVER, W. J. **Practical Nonparametric Statistics**. 3. ed. New York: John Wiley and Sons, 1999.

[14] FRIEDMAN, Nir; GEIGER, Dan; GOLDSZMIDT, Moises. Bayesian network classifiers. **Machine learning**, v. 29,, 1997.

[15] FRIEDMAN, N.; GOLDSZMIDT,M. Building Classifiers Using Bayesian Networks. In: National Conference on Artificial Intelligence (AAAI96), 30., 1996, Portland. **Proceedings**... Prontland, AAAI Press,1996. v.2, p.1277-1284.

[16] SCHMIDHUBER, Jürgen. Deep learning in neural networks: An overview. **Neural networks**, v. 61, p. 85-117, 2015

[17] DENG, Li et al. Deep learning: methods and applications. **Foundations and Trends in Signal Processing**, v. 7, n. 3–4, p. 197-387, 2014.

[18] Correlation (Pearson, Kendall, Spearman). Disponivelem: <http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/>. Acesso em :8 djul. 2018.

[19] BROWNLEE, Jason. How to Use Correlation to Understand the Relationship Between Variables. Statistical Methods, **Machine Learning Mastery**, 27 April 2018. Disponivel em: <https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/>. Acesso em: 8 jul. 2018.