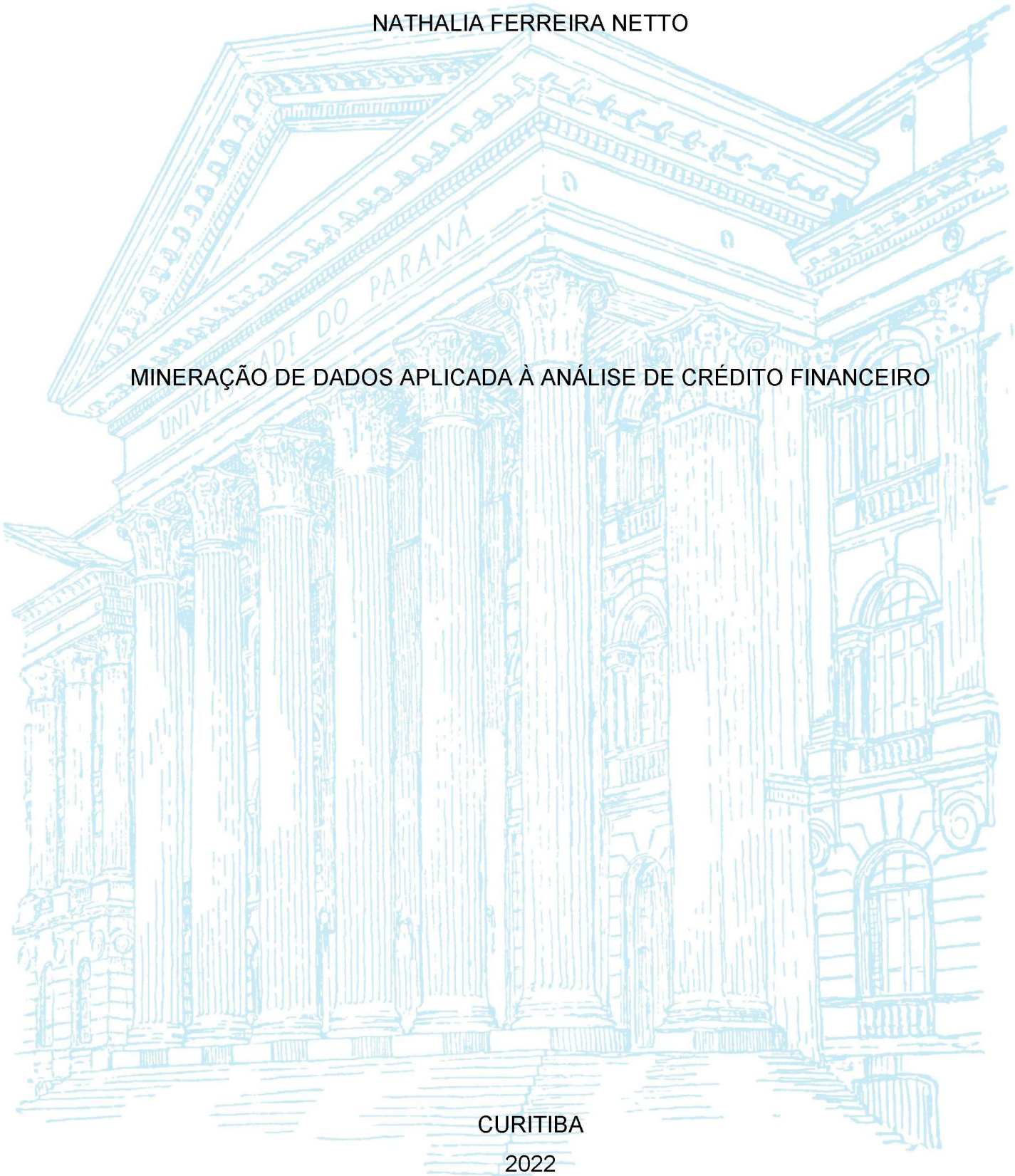
UNIVERSIDADE FEDERAL DO PARANÁ

NATHALIA FERREIRA NETTO



MINERAÇÃO DE DADOS APLICADA À ANÁLISE DE CRÉDITO FINANCEIRO

CURITIBA

2022

NATHALIA FERREIRA NETTO

MINERAÇÃO DE DADOS APLICADA À ANÁLISE DE CRÉDITO FINANCEIRO

Trabalho de Conclusão de Curso apresentado ao curso de Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Gestão da Informação.

Orientador: Prof. Dr. Luciano Heitor Gallegos Marin

CURITIBA

2022

TERMO DE APROVAÇÃO

NATHALIA FERREIRA NETTO

MINERAÇÃO DE DADOS APLICADA À ANÁLISE DE CRÉDITO FINANCEIRO

Trabalho de Conclusão de Curso apresentado ao curso de Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Gestão da Informação.

_____

Prof. Dr. Luciano Heitor Gallegos Marin
Orientador – Departamento de Ciência e Gestão da Informação, UFPR

_____

Profa. Dra. Denise Fukumi Tsunoda
Departamento de Ciência e Gestão da Informação, UFPR

_____

Prof. Dr. José Marcelo Almeida Prado Cestari
Departamento de Ciência e Gestão da Informação, UFPR

Curitiba, __ de _____ de 2022.

AGRADECIMENTOS

Aos meus amigos, de dentro e fora da universidade, que sempre me apoiaram, me animaram e motivaram.

Aos meus colegas de curso, que colaboraram em trabalhos, atividades e estudos ao longo de toda minha graduação.

Aos colegas e amigos de outros cursos, que colaboraram com diferentes visões e agregaram muito na amplitude dos meus conhecimentos complementares à graduação e conhecimentos pessoais.

Aos meus professores universitários, os quais eu admiro grandemente por seus conhecimentos e competências, e que foram essenciais para me guiar em minha formação e início de carreira, indo muito além da sala de aula.

À Universidade Federal do Paraná, que para mim desde muito nova representava um sonho, sendo hoje motivo de orgulho por eu poder fazer parte dessa que compõe a elite universitária da América do Sul.

Aos meus professores anteriores à universidade, que foram exemplares no esforço e competência para contribuir com minha educação pessoal e formal desde a infância, mesmo com as dificuldades enfrentadas na esfera da educação pública.

E especialmente aos meus pais e irmãos, que nunca hesitaram em me apoiar, ajudar, incentivar, motivar, ensinar, colaborar e guiar em toda minha vida, educação e carreira, e assim como a toda minha família, sou extremamente grata.

RESUMO

O presente relatório se propõe a demonstrar a aplicação de métodos de mineração de dados em uma base de dados de análise de crédito para verificar a efetividade na classificação de clientes por padrões de comportamento financeiro. Para isso, discorre-se sobre o crédito, a análise e risco de crédito, a inadimplência e introduz os conceitos de mineração de dados e sua participação no processo de Descoberta de Conhecimento em Bases de Dados. Descreve-se detalhadamente as aplicações deste processo para responder ao problema proposto de classificar clientes inadimplentes e não inadimplentes de forma preditiva. Verifica-se o desempenho de diferentes técnicas de mineração de dados e faz-se a comparação entre elas utilizando a acurácia, precisão e revocação de cada modelo. Por fim, conclui-se que é suficientemente eficaz a classificação proposta utilizando a técnica de árvore de decisão.

Palavras-chave: Descoberta de Conhecimento em Bases de Dados. Mineração de dados. Análise de crédito. Inadimplência. Árvore de decisão.

ABSTRACT

The present report proposes to demonstrate the application of data mining methods in a credit analysis database to verify the effectiveness in the classification of customers by financial behavior patterns. To this end, the report discusses credit, credit analysis, risk and default, and it introduces the concepts of data mining and its participation in the process of Knowledge Discovery in Databases. The applications of this process to answer the proposed research problem of classifying defaulting and non-defaulting customers in a predictive way are described in detail. The performance of different data mining techniques is verified and compared using the accuracy, precision and recall of each model. Finally, it is concluded that the proposed classification is sufficiently effective using the decision tree technique.

Keywords: Knowledge Discovery in Databases. Data mining. Credit analysis. Credit default. Decision tree.

SUMÁRIO

1 INTRODUÇÃO

Uma das principais tendências que se definiram a partir do fim do século XX foi o deslocamento do paradigma de sociedade industrial para sociedade da informação (BORGES, 1995), o que significa que as principais economias mundiais se baseiam não mais na indústria, mas sim na gestão da informação.

Esta mudança de paradigma ocorreu em um contexto de aumento da competitividade entre as empresas no mercado, que viram na informação um recurso estratégico fundamental para a maximização da qualidade do processo decisório (VITAL; FLORIANI; VARVAKIS, 2010), consolidando-a como um dos bens mais valiosos para uma organização na atualidade (COSTA *et al.*, 2019).

Desde o surgimento dos sistemas computacionais, um dos principais objetivos das organizações tem sido coletar e armazenar dados e informações (CAMILO; SILVA, 2009). De acordo com Camilo e Silva (2009), nas últimas décadas essa tendência ficou ainda mais evidente devido à queda nos custos para a aquisição de hardware, sendo possível armazenar quantidades cada vez maiores de dados. E com o volume de dados crescendo diariamente, tornou-se crucial para as organizações entenderem o que fazer com estes dados armazenados para gerar informação e até o conhecimento agregador de valor para a tomada de decisão.

Com a finalidade de atender esta demanda de gerar informação e conhecimento a partir de grandes volumes de dados, foi proposta, no final da década de 80, a Mineração de Dados, do inglês *Data Mining* (CAMILO; SILVA, 2009). Em termos gerais, a mineração de dados é uma forma de descobrir informações ou conhecimentos de caráter prático a partir de dados que uma organização coleta, organiza e armazena (SHARDA; DELEN; TURBAN, 2019).

A Mineração de Dados é definida por Amo (2004) como a aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse em grandes volumes de dados. A autora destaca ainda que a mineração é uma etapa essencial do processo de descoberta de conhecimento em base de dados, do inglês *Knowledge Discovery in Databases* (KDD).

1.1 PROBLEMATIZAÇÃO

A organização escolhida para a aplicação do estudo, chamada no presente trabalho de Empresa X, é uma instituição financeira, de nível nacional, que concede crédito para pessoas físicas. A Empresa X não efetua a penhora de qualquer bem do cliente, tendo isso como fator facilitador de suas vendas, porém assumindo assim maiores riscos de inadimplência.

O nicho atendido pela Empresa X chega a consumidores menos qualificados por classificações de modelos estatísticos de *score* de crédito disponíveis no mercado financeiro, que indicam que estes clientes podem ter histórico de endividamentos, tendências a serem maus pagadores e/ou poucas habilidades para liquidar suas dívidas, o que, por sua vez, os impede muitas vezes de adquirirem crédito com qualquer outra instituição financeira, tornando o produto da Empresa X diferenciado no mercado. Silva (1988) já abordava sobre a concessão de crédito que se baseia puramente na confiança do credor em seu cliente e o que deve ser analisado pelo credor nestes casos:

> A concessão de crédito é baseada na confiança que o credor tem, na vontade e na capacidade do devedor de liquidar suas obrigações dentro das normas estabelecidas. A honesta intenção do devedor para pagar é de importância primária. Somente quando o credor tem certeza de que o prospectivo devedor tem vontade para liquidar suas obrigações, é que investigações mais aprofundadas são aconselhadas. Assumindo que o prospectivo devedor tem vontade para pagar a suas obrigações, a segunda pergunta, que deve ser respondida, é se ele tem habilidade para pagar (SILVA,1988, p. 74).

Dada a necessidade de (1) se adaptar à exigente realidade atual do mercado na sociedade da informação, (2) responder a estas questões que Silva (1988) destacou e, consequentemente, (3) tornar as análises de crédito mais eficientes, a Empresa X tem buscado cada vez mais coletar e utilizar o grande volume de dados que tem à sua disposição para tomar decisões e criar políticas de concessão de crédito que auxiliem no combate à inadimplência sem fugir de seu nicho de mercado, garantindo que o maior número de clientes desse nicho sejam atendidos mas ainda obtendo o retorno estipulado.

Cabe também citar que, por conta dos fatores supracitados, a Empresa X tem deixado de utilizar apenas o *score* fornecido pelo Serviço de Proteção ao Crédito (SPC) em sua análise de crédito e buscado internamente desenvolver

critérios próprios de classificação dos possíveis clientes para mitigar os riscos de se atender consumidores com um *score* mais baixo de mercado.

A Empresa X então adquiriu bases externas contendo alguns aspectos do comportamento financeiro de seus clientes, como as dívidas passada, renda, entre outros dados fornecidos pelo Banco Central do Brasil, e juntamente com informações internas de seus clientes (inadimplentes ou não), criou uma base para gerar uma análise preditiva em relação à habilidade e intenção de pagamento do cliente, se ele será ou não um bom pagador até o fim de seu contrato.

Neste contexto, a mineração de dados conta com diferentes técnicas realizadas por algoritmos para extração de padrões, que podem ser divididas entre preditivas e descritivas (GALVÃO; MARIN, 2009). Entre as tarefas de predição se destacam a regressão e classificação e entre as tarefas de descrição estão o agrupamento, a associação e a sumarização (GALVÃO; MARIN, 2009).

A partir disso, supõe-se que a aplicação de um processo de KDD (que contém como etapa essencial a mineração de dados) pode ser ideal para explorar a grande quantidade de dados e obter uma classificação que ajudará a empresa a identificar melhor seus bons clientes.

Logo, o problema de pesquisa do presente estudo consiste em se é **possível encontrar um padrão de comportamento financeiro que possibilite classificar consumidores de crédito da Empresa X por inadimplência, com a utilização de técnicas de mineração de dados?**

1.2 OBJETIVOS

1.2.1 Objetivo geral

O objetivo geral é então aplicar um processo de descoberta de conhecimento em uma base de dados de clientes da Empresa X para a classificação entre inadimplentes e não inadimplentes, avaliando a efetividade na classificação, por meio de técnicas de mineração de dados.

1.2.2 Objetivos específicos

I. Descrever todas as etapas da aplicação processo de KDD;
II. Definir técnicas e mineração de dados adequadas para o conjunto de dados;
III. Avaliar o desempenho do modelo de mineração de dados utilizado.

1.3 JUSTIFICATIVA

Como justificativa acadêmica a este trabalho destaca-se primeiramente a relevância e atualidade do tema de mineração de dados e da descoberta de conhecimento em bases de dados, principalmente para a gestão e ciência da informação.

Já no contexto de mercado, segundo o Instituto de Pesquisa Econômica Aplicada (IPEA), no início de 2022, houve um aumento na inadimplência para pessoas físicas e o comprometimento da renda e níveis de endividamento se encontram em patamares historicamente altos e em trajetória de crescimento.

Então, este estudo poderá demonstrar o uso da gestão da informação e da mineração de dados para a gestão de um negócio contra um problema chave que empresas de crédito financeiro vêm enfrentando, principalmente com o agravante da pandemia de Covid-19.

Este estudo, se bem sucedido, poderá ainda subsidiar a tomada de decisão sobre as políticas de aprovação de crédito que serão aplicadas no sistema de análise da Empresa X.

2 REVISÃO DE LITERATURA

2.1 CRÉDITO PARA PESSOAS FÍSICAS

Partindo da etimologia, a palavra "crédito" vem do latim *creditum*, que remete a algo emprestado, objeto passado em confiança a outrem, particípio passado de *credere*, que é "acreditar, confiar".

Neste contexto, sob o ponto de vista meramente empresarial, a concessão de crédito significa a transferência da posse de um bem ou de uma quantia em dinheiro, mediante a promessa de pagamento futuro (GUIMARÃES; NETO, 2002).

Alinhado a isto, de maneira geral, o crédito depende de duas partes: a credora é aquela que empresta o dinheiro a uma pessoa ou instituição, e crê que a contraparte devedora devolva o dinheiro com um prêmio de risco, chamado de juros (MOURA, 2018).

No caso do crédito pessoal, a parte devedora deve ser uma pessoa física, o consumidor final, que usará o crédito para pequenos projetos pessoais ou situações cotidianas, adiantando o consumo de bens e serviços.

De acordo com a Serasa Experian (2019), nestes casos o consumidor de crédito conta com as seguintes opções, que são as principais modalidades de se adquirir crédito:

I. empréstimo pessoal: é uma das modalidades de crédito mais comuns, em que o devedor recebe a quantia do credor e posteriormente paga com taxas de juros que variam de acordo com o banco, garantias e até nível de burocracia;
II. financiamento: está dentro da linha do empréstimo pessoal, no entanto, ele é usado para uma situação específica, como aquisição de um imóvel, por exemplo, do qual o devedor não tenha o montante do valor do bem no momento da necessidade para o consumo;
III. consignado: costuma ter juros menores, pois suas parcelas são descontadas diretamente da folha salarial. Já que o risco de crédito é menor, seus juros se tornam mais baixos;

IV. consórcio: o valor é pago em parcelas previamente ao recebimento do crédito; e
V. cartão de crédito: a forma mais utilizada de se adquirir crédito em todo o mundo. Trata-se de um crédito rotativo idealizado para o uso cotidiano em um período usualmente mensal.

Para cada modalidade de crédito existem procedimentos, garantias e burocracias diferentes que afetam a percepção de risco de perda por parte do credor, influenciando diretamente nos juros que o devedor irá pagar.

Como exemplo, uma vez que no financiamento existe uma intenção de aquisição de crédito especificamente para o consumo de um bem, normalmente prioritário para o devedor, o credor pressupõe que há planejamento financeiro da parte devedora para quitar a dívida, ou seja, é considerada a possível boa intenção e habilidade para o cumprimento do contrato. Porém, em todo caso, no mercado de crédito é necessário sempre que se realize uma análise prévia do consumidor de crédito, chamada análise de crédito. Ela ocorre para avaliar os principais fatores de risco de crédito, elencados por Silva (1988) como a falta de intenção e a falta de habilidade do devedor para pagar sua dívida, que desencadearia na perda para o credor, chamada de inadimplência.

É a partir da análise de crédito que se define o grau de confiança do credor no comprometimento futuro do devedor e, consequentemente, tomar a decisão de finalmente conceder ou não o crédito.

2.1.1 O risco e a análise de crédito

O risco é constituído pela ocorrência de qualquer fato adverso para uma dada situação esperada. No caso específico, o risco de crédito é a probabilidade de ocorrência de perdas por inadimplência (BORGES, 2001).

Para mitigar este risco, que é um dos principais fatores de atenção das empresas de crédito, é estabelecido um processo de análise que consiste na utilização sistêmica das informações disponíveis sobre o solicitante do crédito para avaliar se existem indícios de que lhe falta habilidade ou intenção para quitar sua futura dívida, com o objetivo principal de decidir se haverá a concessão de crédito.

Weber *et al.* (1993) destacam que desde os pequenos empreendedores às grandes instituições financeiras, todos analisam crédito de pessoas físicas. E assim como varia o sujeito, variam também os métodos utilizados, que podem ir da simples confiança, verificação de comportamento em outras instituições ao uso do SPC (Serviço de Proteção ao Crédito), como bem citados pelos autores.

No caso das instituições financeiras, que são organizações regulamentadas para concessão de crédito, essas análises são de fato metódicas, sistêmicas e regidas por políticas de crédito.

Maia (2007) explica que uma boa política de crédito deve estabelecer:

I. Quem decide quem pode receber crédito;
II. Quem decide os limites de crédito;
III. Que fatores controlam a decisão inicial de crédito;
IV. Que fatores determinam os limites de crédito;
V. Quais as condições de cada tipo de crédito;
VI. Como deve ser precificado o crédito;
VII. Qual a política aplicável para descontos; e
VIII. Como deve ser tratada a inadimplência.

Dentre os principais fatores que controlam a decisão e os limites de crédito está o *score*. Guimarães e Neto (2002) afirmam ainda que:

> O principal meio de controle do risco, ou pelo menos o mais utilizado, é o sistema de *score*. Este sistema consiste basicamente em avaliar características do novo cliente, atribuindo um determinado valor a cada característica. Em seguida os dados obtidos são usados na elaboração de um *score*. Com base no escore obtido pelo cliente toma-se a decisão de conceder, ou não, o crédito. Para tomar tal decisão, o *score* é comparado com um valor previamente estabelecido, chamado valor de corte. É na obtenção deste último que reside a maior parte dos problemas enfrentados pelos profissionais envolvidos. A questão a ser resolvida neste ponto pode ser colocada da seguinte forma: como obter um valor de corte confiável a ponto de evitar perdas para a empresa, tanto pela aceitação, errada, de clientes que venham a se tornar inadimplentes quanto pela rejeição, igualmente errada, de clientes adimplentes (GUIMARÃES; NETO, 2002).

No Brasil, o próprio SPC, entre outras instituições, como a Serasa Experian ou Quod, já fornecem um *credit score* para o mercado financeiro para que qualquer empreendimento possa utilizá-lo, sob um custo por consulta, e minimizar seus riscos de crédito para um cliente novo. Este *score* trata de uma pontuação atribuída, geralmente de 0 a 1000, para cada consumidor de crédito, em que 1000 tende a ser

bom pagador e 0 tende a ser um mau pagador. Cada uma dessas instituições tem uma metodologia própria de coleta e análise de dados para a definição do *credit score*. Estas instituições que fornecem estas consultas são então intituladas birôs.

No entanto, além da possibilidade de consultar um birô de crédito, os credores podem ainda fazer a própria análise com critérios mais bem definidos para seu modelo de negócio, usando informações disponibilizadas pelo SCR (Sistema de Informações de Crédito do Banco Central do Brasil) mescladas com informações internas baseadas no comportamento financeiro do cliente e analisadas por um modelo estatístico, sendo este o *behaviour score* (MANFIO, 2007).

Ao fazer-se o uso correto de métodos e dos dados na análise de crédito, várias são as vantagens obtidas, dentre as quais Lemos, Steiner e Nievola (2005) elencaram:

I. necessidade de menos pessoas envolvidas com a análise do crédito, as quais podem ser aproveitadas em outras atividades;
II. maior rapidez no processamento dos pedidos de crédito;
III. menor subjetividade no processo;
IV. direcionamento mais eficaz do crédito.

Quando o analista de crédito tem à sua disposição uma regra de reconhecimento de padrões e classificação que indique previamente a possibilidade de inadimplência de um futuro cliente, a decisão de concessão de crédito é facilitada; esse profissional pode então utilizar argumentos quantitativos em substituição a argumentos subjetivos e decidir com maior eficiência (GUIMARÃES; NETO, 2002). E congruente a isto, a aplicação da regra pode até ser automatizada dentro do processo de análise de crédito para poupar tempo e custo de operação.

2.1.2 Inadimplência

A inadimplência é um dos grandes fatores de risco para as instituições financeiras e, portanto, deve ser mapeada e "prevista". Mas para isso é preciso saber definir o que é e quando um cliente de fato está inadimplente.

Um dos indicadores usados no mercado é o Sinistro. Ele representa o percentual da quantidade de contratos que ultrapassaram 60 dias de atraso histórico, ou seja, que possuem ou já possuíram atraso acima de 60 dias (MANFIO,

2007). Assim sendo, uma vez que o cliente entra para o conjunto de sinistrados ele não deixa de ser considerado inadimplente, pois já demonstrou ser um risco para a retomada do valor de crédito pelo credor.

Esta, porém, é apenas uma das diversas formas de definir e medir a inadimplência. Entre as outras, muda-se o prazo para ser considerado inadimplente ou número de parcelas em atraso da dívida, por exemplo, pois apesar do atraso, o pagamento total da dívida ainda pode ocorrer. O perfil do inadimplente, assim como o risco para a cooperativa de crédito, pode mudar de acordo com o contexto.

Herling *et al.* classificou estes perfis de inadimplentes, ou maus pagadores, da seguinte forma:

> O mau pagador pode ter vários perfis, sendo eles o "verdadeiro mau pagador" que é definido como uma pessoa com a intenção de lesar o credor, e recusa pagar o débito ou tentar prolongar ao máximo o pagamento. Já o "mau pagador ocasional" não tem a intenção de enganar o credor, porém por motivos pessoais não tem condições de honrar seus compromissos. Existe também o "devedor Crônico" que é aquele que sempre atrasa o pagamento, mas acaba pagando. Se esse devedor for bem administrado e controlado pelo credor pode ser uma excelente fonte de lucro (HERLING *et al.*, 2013).

A razão pela qual o devedor crônico é tão valorizado é devido ao fato que ele irá quitar suas dívidas ainda com um bônus de juros pelo atraso. Ao definir o que é um mau pagador ou a inadimplência é preciso considerar este tipo de fator, e por isso a inadimplência pode ser considerada algo subjetivo para as instituições financeiras.

Em contrapartida, este estudo vai de encontro com a definição do Sinistro para inadimplentes.

Ademais, a inadimplência, por ser um risco de perda, está diretamente ligada às avaliações de políticas de crédito, ajustes de taxas de juros para compensar o risco e definição de estratégias de vendas das instituições financeiras.

2.2 GESTÃO DA INFORMAÇÃO

Como já introduzido, as organizações hoje entendem cada vez mais que os dados e informações são importantes para o bom desempenho de um negócio. Mas, diferente dos desafios enfrentados no passado pela falta de informação, hoje

enfrentam dificuldades para administrar todo volume que está disponível e que ainda é gerado a cada dia.

Devido à intensidade de produção e dispersão, pesquisadores de épocas, países e formações diferentes estudaram modos de "racionalizar" e democratizar o acesso a todo o emaranhado de informações dispersas pelo mundo (MONTEIRO; DUARTE, 2018).

A Gestão da Informação tem sua origem em áreas clássicas da Biblioteconomia, Documentação e Ciência da Informação (WILSON, 2002). E não escapando disso, hoje ela visa incrementar a competitividade empresarial e os processos de modernização organizacional (MARCHIORI, 2002), através do estudo dos processos informacionais, do modo como a informação pode ser organizada, armazenada, recuperada e utilizada para a tomada de decisões e para a construção do conhecimento (DUARTE, 2011).

Para que haja de fato valor na informação, acima dela, as organizações precisam da gestão dessa informação.

2.2.1 Análise de dados

Respeitando a área de Gestão da Informação, quando se fala em análise de dados o foco está no processamento de um grande volume de dados brutos para se obter informação a partir de uma análise.

Sharda, Delen e Turban (2019) destacam que os dados podem ser encarados como a matéria-prima daquilo que as tecnologias de decisão produzem: informações, idéias e conhecimento. Sob essa ótica, as técnicas de análise de dados se baseiam geralmente nos estudos de estatística, conjugando análises descritivas e preditivas.

As análises de dados descritivas permitem obter um diagnóstico situacional para a base de dados e seu contexto através de relatórios que compilam as informações presentes nessas bases.

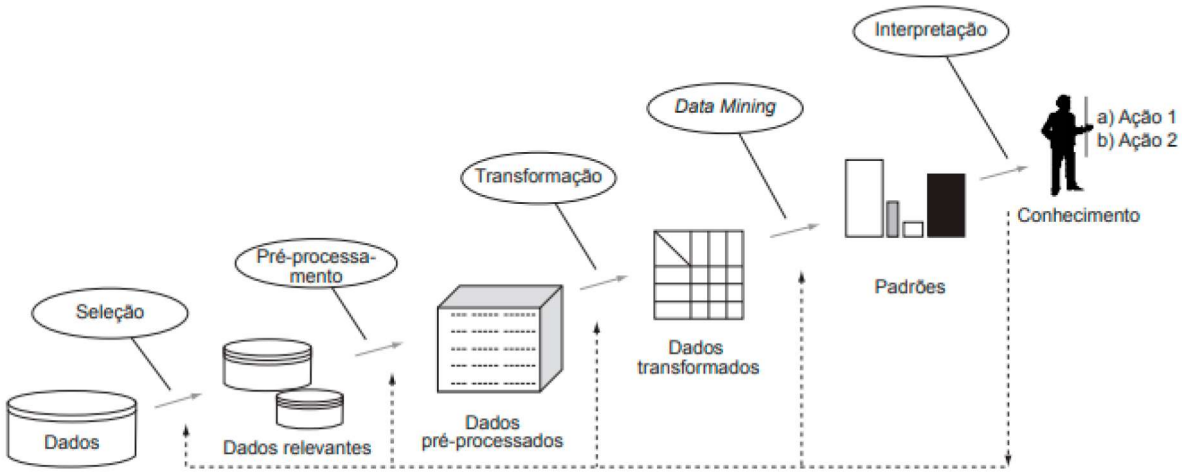
Já no caso de análises preditivas, que contam com a utilização de algoritmos e técnicas mais avançadas, procura-se entender o que irá acontecer observando padrões nos conjuntos de dados históricos. Neste caso, enquadra-se a mineração de dados.

2.2.1.1 KDD e a Mineração de Dados

A Descoberta de Conhecimento em Bases de Dados, do inglês *Knowledge Discovery in Databases* (KDD) é um processo interativo para identificar nos dados novos padrões que sejam válidos, novos, potencialmente úteis e interpretáveis (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Ele conta com 5 etapas, elucidadas por Steiner *et al.* (2006) na figura 1.

FIGURA 1 - PROCESSO DE KDD



FONTE: Steiner *et al.* (2006).

Durante a fase de seleção é que se procura entender o contexto e os dados que serão utilizados. É importante que se tenha dados pertinentes ao problema e por isso esta etapa não deve ser negligenciada. No pré-processamento, ocorre a limpeza dos dados para reduzir o ruído e isso irá interferir muito no desempenho de todo o processo. Com a transformação, a base é preparada para a aplicação do algoritmo de mineração de dados de maneira adequada e nesta etapa são realizadas conversões, discretização, formatação e alterações dos dados para que o modelo possa atingir um bom desempenho.

Como Sharda, Delen e Turban (2019) comentam, se não houver adequação dos dados e a execução correta de todas as etapas, pode-se chegar a uma resposta errada para um problema correto.

A fase seguinte é a mineração de dados, e entre as etapas, esta é a essencial em que ocorre de fato a aplicação de um algoritmo para evidenciar os padrões. Um algoritmo de mineração de dados é um conjunto de heurística e

cálculos usado para encontrar padrões em determinadas bases de dados. Cada algoritmo respeita uma tarefa assumindo diferentes formas e se adaptando a diferentes conjuntos de dados e problemas. Assim, tem-se uma classificação dos algoritmos de acordo com as principais tarefas de mineração de dados, que podem ser elencadas da seguinte forma:

I. Classificação: identificar a classe de um registro de forma supervisionada, ou seja, esta tarefa analisa a base fornecida para treinamento, já contendo a indicação à qual classe cada registro pertence, a fim de identificar padrões para classificar um novo registro;
II. Regressão: utilizada para estimar os valores numéricos consequentes de outras variáveis de um conjunto, também de forma supervisionada;
III. Agrupamento: tarefa de aprendizado não supervisionado que busca agrupar os registros de acordo com seus atributos;
IV. Associação: tarefa que cria regras para indicar a relação entre dois atributos.

Para o sucesso do KDD é necessária a aplicação da tarefa mais adequada de mineração de dados para o conjunto e contexto dos dados.

E, finalmente, no KDD tem-se a etapa de interpretação que consiste em entender o que foi extraído de informação para a tomada de decisão. E essa não é uma tarefa simples. Dantas *et al.* (2008) argumentam que o valor de uma decisão estratégica para o negócio depende da capacidade do gestor de interpretá-las e da experiência para associá-las de maneira conveniente, não bastando apenas ter a informação e metodologia adequadas.

3 METODOLOGIA

Dado o objetivo de aplicação do processo de KDD no presente trabalho, e uma vez que a metodologia tem alta relação com a sequência de etapas deste processo apoiado por uma revisão teórica sobre o tema, esta pesquisa pode ser caracterizada como descritiva.

A pesquisa também tem caráter quantitativo por lidar com métodos estatísticos de análises de dados para solucionar o problema de pesquisa.

Portanto, as etapas da metodologia deste trabalho foram divididas da seguinte forma:

I. descrição da seleção da base pertinente ao tema de análise de crédito, contendo informações sobre o comportamento financeiro, que possa ser útil ao objetivo do trabalho;
II. preparação da base, que consistiu apenas na extração e conversão do arquivo para *Comma Separated Values* (CSV), por ser mais otimizado e retirada dos espaços em branco nos nomes de colunas;
III. entendimento, pré-processamento e transformação da base, detalhados no item 3.1.3 deste trabalho;
IV. aplicação de algoritmos de mineração de dados de forma a testar o desempenho de cada um;
V. escolha de um algoritmo e revisão dos parâmetros para melhorias nos resultados; e
VI. análise dos resultados.

3.1 MATERIAIS E MÉTODOS

3.1.1 A base de dados

A Empresa X extraiu dados de seus clientes com contrato em aberto no momento da extração, em novembro de 2021, e buscou adquirir externamente dados sobre o histórico destes clientes antes de tomarem crédito para verificar se havia alguma maneira de prever o potencial de inadimplência destes clientes ao

solicitarem crédito com a Empresa X. A partir daí, ocorreu a seleção da base que seria utilizada no presente trabalho.

3.1.1.1 Seleção da base

A base de dados escolhida para realizar a análise foi gerada a partir do cruzamento dos dados do Sistema de Informações de Crédito (SCR) com informações internas de vendas para uma amostra de clientes da Empresa X que estavam com contrato em andamento no momento da extração da base.

O SCR é um sistema do Banco Central, e corresponde a um instrumento de registro e consulta de informações sobre as operações de crédito, avais e fianças prestados e limites de crédito concedidos por instituições financeiras a pessoas físicas e jurídicas no país. É regulado pela Resolução do Banco Central do Brasil (BCB) nº4.571/2017. O SCR é alimentado mensalmente pelas instituições financeiras, mediante coleta de informações sobre as operações concedidas.

É pertinente ressaltar que todos os dados foram extraídos, manipulados e analisados mantendo o anonimato dos clientes, sem qualquer informação que possibilitasse a identificação destes.

3.1.1.2 Descrição da base

A variável de classificação divide a base em duas classes: os inadimplentes e os não inadimplentes. O indicador utilizado para classificar a inadimplência dos clientes é o Sinistro 60, que acusa os clientes que possuem ou já possuíram atraso acima de 60 dias nos pagamentos de suas dívidas de crédito com a Empresa X. O Sinistro representa o percentual da quantidade de contratos que ultrapassaram os 60 dias de atraso histórico. O Sinistro desta base é então de 63,8%.

Como principais características da base, tem-se:

I. 5003 observações (ou linhas);
   a. 1811 clientes com o pagamento em dia e 3192 clientes sinistrados, ou seja, inadimplentes por mais de 60 dias até o momento da extração da base;
II. 390 variáveis (ou colunas), sendo:
   a. 1 variável de classificação;

b. 16 variáveis com informações internas sobre o cliente e sobre a venda, tal qual idade, renda, região, distância de seu endereço até o agente que realizou a proposta de crédito, idade do agente, canal da venda, se o cliente é novo, entre outros dados;
c. 373 variáveis com informações do SCR, como por exemplo, quantidade de operações a vencer daquele cliente, montante em dívidas vencidas, crédito liberado, entre outras.

3.1.2 Ferramentas

As principais ferramentas utilizadas foram: (1) o Excel, um editor de planilhas da empresa desenvolvedora de sistemas Microsoft; e (2) o Python, que é uma linguagem de programação de alto nível orientada a objetos, interpretada de s*cript*, imperativa, funcional, de tipagem dinâmica e forte.

O Excel, enquanto um recurso para visualizar e manipular a base de dados, é o editor de planilhas mais performático e, não à toa, mais utilizado. O Python, por ser uma linguagem de alto nível e voltada para a análise de dados, conta com diversas bibliotecas facilitadoras para a manipulação dos dados e aplicação de algumas etapas do processo de KDD.

Para cada etapa de aplicação do KDD foram utilizadas estas diferentes ferramentas mais adequadas para cada tarefa:

I. Para o pré-processamento da base, foi utilizado então o Microsoft Excel;
II. Já para o tratamento dos dados, foi utilizada a linguagem Python, e suas bibliotecas Pandas e *Scikit Learn* (SKLearn). O ambiente de execução foi o Google Colaboratory, um serviço de armazenamento em nuvem de *notebook*s voltados à criação e execução de códigos em Python para análises de dados;
III. Por fim, para a aplicação do algoritmo de mineração de dados e análise dos resultados foi utilizado novamente o Python.

3.1.3 Métodos

Dado o objetivo definido para o trabalho, que consiste em classificar os clientes de acordo com categorias de inadimplência, bem como as características da

base, que já etiqueta os clientes (registros) de forma binária como inadimplente ou não, a tarefa de mineração de dados escolhida como mais adequada foi a classificação. A seguir são descritas as etapas do KDD para este fim.

3.1.3.1 Entendimento e pré-processamento da base

Nesta primeira etapa do processo, buscou-se o entendimento de cada variável da base.

De forma preliminar, uma característica da base que foi percebida como um possível problema para o desempenho do modelo de mineração que seria usado foi a redundância dos dados do SCR pela presença de variáveis dependentes entre si, podendo gerar uma multicolinearidade.

É possível citar como exemplo uma variável x com o montante total de dívidas a vencer de um cliente e outras variáveis com o montante de operações a vencer divididas por prazo de vencimento, que somadas resultam na variável x.

A figura 2 apresenta uma amostra destas colunas para exemplificar.

FIGURA 2 - AMOSTRA DA BASE ANTES DO PRÉ-PROCESSAMENTO

Valor dívidas a vencer da pessoa em R$	Valor operações a vencer até 30d	Valor operações a vencer até 31 a 60d	Valor operações a vencer até 61 a 90d	Valor operações a vencer até 91 a 180d	Valor operações a vencer até 181 a 360d
2254,47	1493,74	435,05	54,28	162,84	108,56

FONTE: tela do Google Colaboratory, a autora (2022).

Além do caso destas variáveis, era também redundante manter a variável de região do cliente e a da região do vendedor, ambas na mesma base para análise, uma vez que analisando a regra do negócio de que o vendedor só vende para clientes do mesmo estado, conclui-se que são exceções os casos que a região do cliente e do vendedor são diferentes, então foi desconsiderada a região do vendedor.

Por fim, foram retiradas as colunas que apresentavam menos de 1% dos registros preenchidos, dadas como irrelevantes para o modelo.

Dessa maneira a base foi reduzida para 61 colunas.

3.1.3.2 Transformação dos dados

A primeira etapa de transformação dos dados foi a discretização das variáveis numéricas, categorizando-as por faixas de valores. A discretização foi realizada para preparar o modelo para a aplicação de diferentes algoritmos de classificação de forma padronizada na mesma base para efeito de comparação, visto que algumas técnicas se adaptam melhor a bases estritamente categóricas.

A discretização conta com diferentes métodos, e para este trabalho o método escolhido foi por quantil, que divide a variável em intervalos de tamanhos iguais com base nos quantis de amostra. Dessa forma, as variáveis numéricas foram divididas em 4 faixas, codificadas de 0 a 3.

Já as variáveis categóricas foram todas codificadas em variáveis binárias para que os algoritmos utilizados não as interpretassem como valores ordinais. Esta codificação ocorreu da seguinte forma: se uma categoria fosse binária, seriam mantidos os códigos 0 e 1 para seus valores. No entanto, para variáveis com mais de dois valores possíveis, foram criadas variáveis fictícias com o codificador O*ne-Hot,* formando uma matriz binária.

Para elucidar: a variável “REGIAO”, que podia conter os valores “Sul”, “Sudeste, “Norte”, “Centro-Oeste” e “Nordeste”, foi transformada em 5 variáveis: “REGIAO_SUL”, “REGIAO_SUDESTE”, “REGIAO_NORTE”, “REGIAO_CENTRO_OESTE” e “REGIAO_NORDESTE”, sendo os valores para estas colunas binários, apenas 1 ou 0, representando respectivamente se é verdade ou não que o registro pertence àquela região. A figura 3 mostra o resultado desta codificação.

FIGURA 3 - MATRIZ DAS REGIÕES DOS CLIENTES

REGIAO_SUL	REGIAO_SUDESTE	REGIAO_CENTRO_OESTE	REGIAO_NORDESTE	REGIAO_NORTE
0.0	0.0	0.0	1.0	0.0
0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0

FONTE: tela do Google Colaboratory, a autora (2022).

Após este processo é necessário fazer uma revalidação da multicolinearidade na base, uma vez que estas variáveis fictícias apresentam uma correlação natural, por um valor sempre depender dos outros – se de três variáveis fictícias herdadas de uma mesma variável, duas apresentarem o valor 0 então a última com certeza apresentará o valor 1.

Desta vez foi aplicado um algoritmo para que fosse utilizada uma métrica, chamada fator de inflação da variância (VIF), para verificar todas as variáveis redundantes no conjunto. Ao aplicar esta métrica na base sem a variável de classificação, uma variável preditora em que se observa grande correlação com outras variáveis preditoras resultará em um alto valor de VIF e será redundante. Estas variáveis foram retiradas do modelo por se tratar de ruído.

Como resultado, não só variáveis fictícias foram excluídas - “REGIAO_SUDESTE” e “CANAL_CORBAN” - mas também a variável “CONSTA_BACEN” e a variável que apresentava o montante total de dívidas a vencer do cliente.

3.1.3.3 Aplicação dos algoritmos de mineração de dados

Por ser um modelo de aprendizagem supervisionada, a base foi dividida em base de treino, para os algoritmos analisarem e aprenderem, e base de testes para verificar o desempenho do aprendizado. Para a base de testes foram selecionados e separados 20% dos registros totais de forma aleatória.

A princípio, foram elencadas algumas das principais técnicas de classificação para realizar os testes de desempenho, observando qual seria a mais adequada para o modelo. São elas, descritas por Camilo e Silva (2009):

Classificação Bayesiana, Árvore de Decisão e *Support Vector Machines* (SVM). Os principais critérios para escolha destas técnicas foram as características da base, desempenho e facilidade na interpretação e comparação dos resultados. Além disso, estas então entre as técnicas mais tradicionais para a tarefa de classificação na área de análise de crédito.

Para a Classificação Bayesiana, o algoritmo utilizado foi o *MultinomialNB* (Naive Bayes Multinominal), da biblioteca SKLearn. Este algoritmo foi escolhido por ser o mais indicado para bases com dados discretizados, com mais de dois valores possíveis em cada atributo.

A Árvore de Decisão, por sua vez, foi construída com o algoritmo *Extreme Gradient Boosting* (XGBoost). Este algoritmo, de biblioteca própria, foi escolhido por sua rapidez e desempenho. Este algoritmo ganhou notoriedade e relevância por isso.

E baseado em SVM, também do *SKLearn*, foi aplicado o algoritmo *Support Vector Classification* (SVC) tendo como parâmetro a função *Radial Basis Function* (RBF), baseada na Distância Euclidiana Quadrática, por ser mais adaptativa a conjuntos de dados não lineares.

3.1.3.4 Avaliação de resultados

A base de teste contou com um suporte de 362 negativos para inadimplência (0) e 639 registros de sinistrados, ou seja, clientes inadimplentes (1).

A matriz de confusão, utilizada para visualização dos resultados, funciona como um mapa de erros e acertos da predição do modelo na base de teste. Sua leitura se dá da seguinte maneira: no eixo y do quadro que representa a matriz estão os valores reais da base de testes e no eixo x estão os valores previstos pelo modelo para a mesma base.

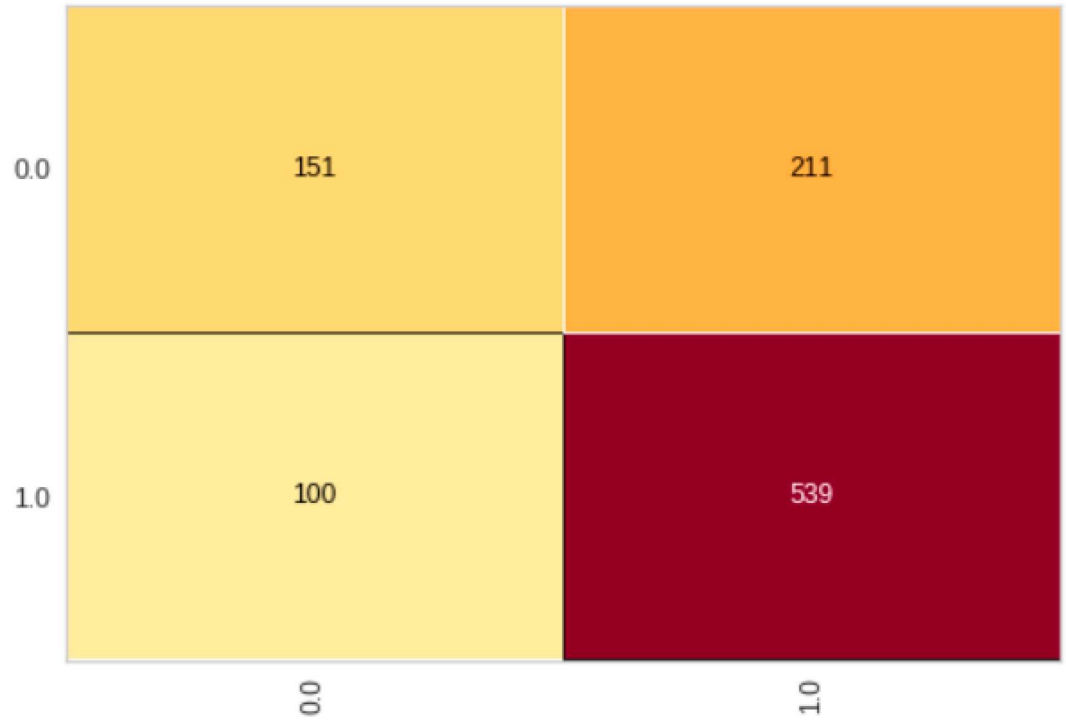
Estes dados são cruzados de forma que possibilita a verificação dos verdadeiros negativos (quantidade de valores reais 0 que o modelo previu como 0), falsos positivos (valores 0 que o modelo classificou como 1 erroneamente), falsos negativos (0 classificados como 1 erroneamente pelo modelo) e verdadeiros positivos (valores reais 1 que o modelo acertadamente previu como 1). Com base nisso são calculadas a acurácia, a precisão e a revocação.

A acurácia é calculada dividindo os acertos sobre o suporte total. A precisão é calculada a partir da divisão dos verdadeiros positivos sobre a soma dos verdadeiros e falsos positivos, ou da divisão dos verdadeiros negativos sobre a soma dos verdadeiros e falsos negativos. Já a revocação revela-se na divisão dos verdadeiros positivos pelos verdadeiros positivos somados aos falsos negativos, ou dos verdadeiros negativos sobre a soma dos verdadeiros negativos e falsos positivos.

Para a avaliação dos resultados, foram então consideradas a acurácia (*accuracy*), precisão (*precision*) e revocação (*recall*) do modelo demonstradas com a matriz de confusão.

A figura 4 apresenta a matriz de confusão do modelo de classificação bayesiana aplicado.

FIGURA 4 - MATRIZ DE CONFUSÃO DA CLASSIFICAÇÃO BAYESIANA

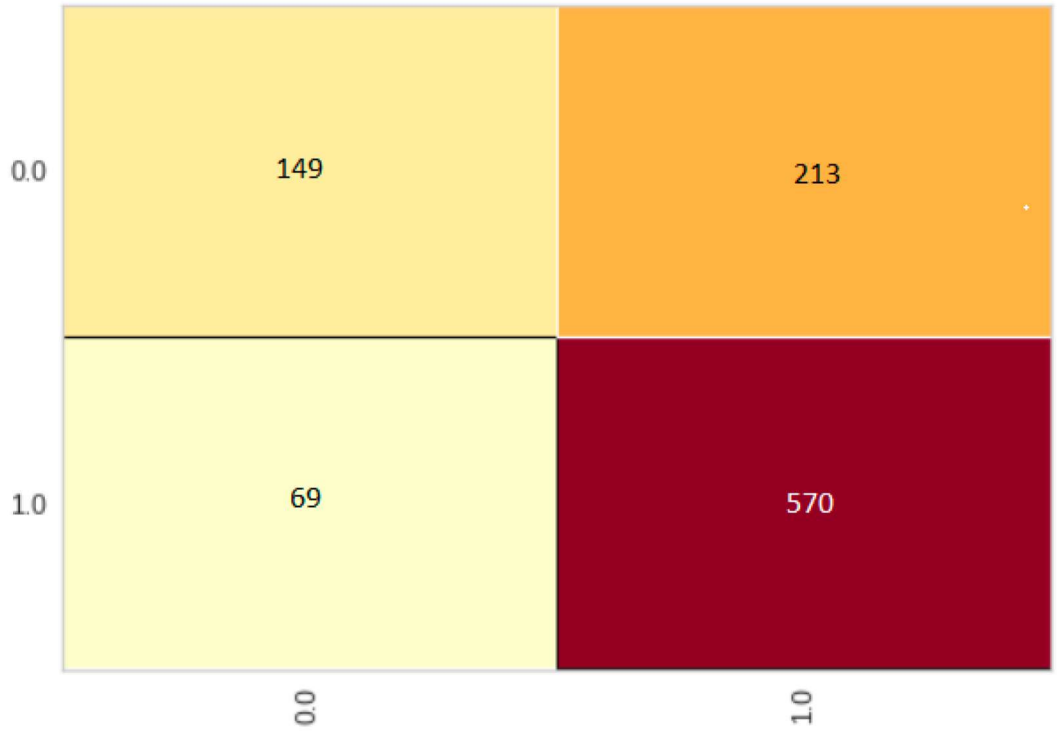


FONTE: a autora (2022).

Conforme é possível calcular com a matriz de confusão, para a classificação com Naive Bayes Multinominal, a acurácia foi de 68,93%. Obteve precisão de 71,87% nos positivos para inadimplência e 60,15% nos negativos. A revocação dos positivos foi de 84,35%, enquanto a dos negativos foi de 41,71%.

A figura 5, abaixo, mostra a matriz de confusão obtida com os resultados do teste de eficácia do modelo de Árvore de Decisão.

FIGURA 5 - MATRIZ DE CONFUSÃO DA ÁRVORE DE DECISÃO

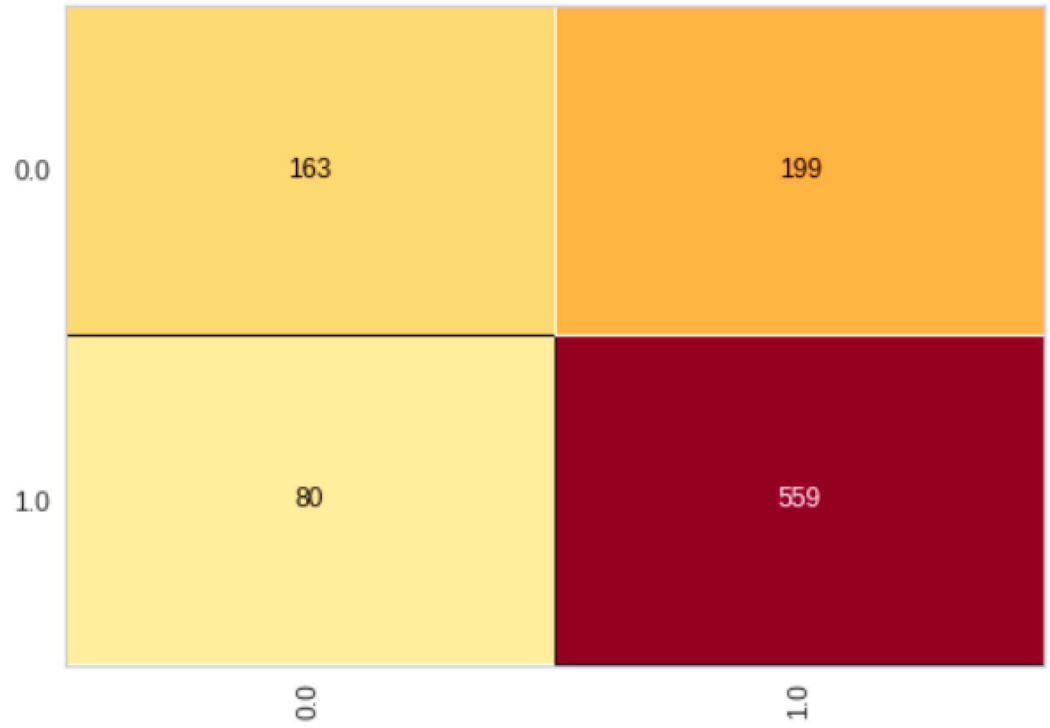


FONTE: a autora (2022).

Classificando pelo XGBoost, a acurácia foi de 71,82%, a precisão nos positivos foi de 72,79% e nos negativos foi de 68,34%. A revocação dos positivos foi de 89,20% e a dos negativos foi de 41,16%.

E finalmente apresenta-se a figura 6, com a matriz de confusão para o modelo SVM, a seguir.

FIGURA 6 - MATRIZ DE CONFUSÃO DO SVM



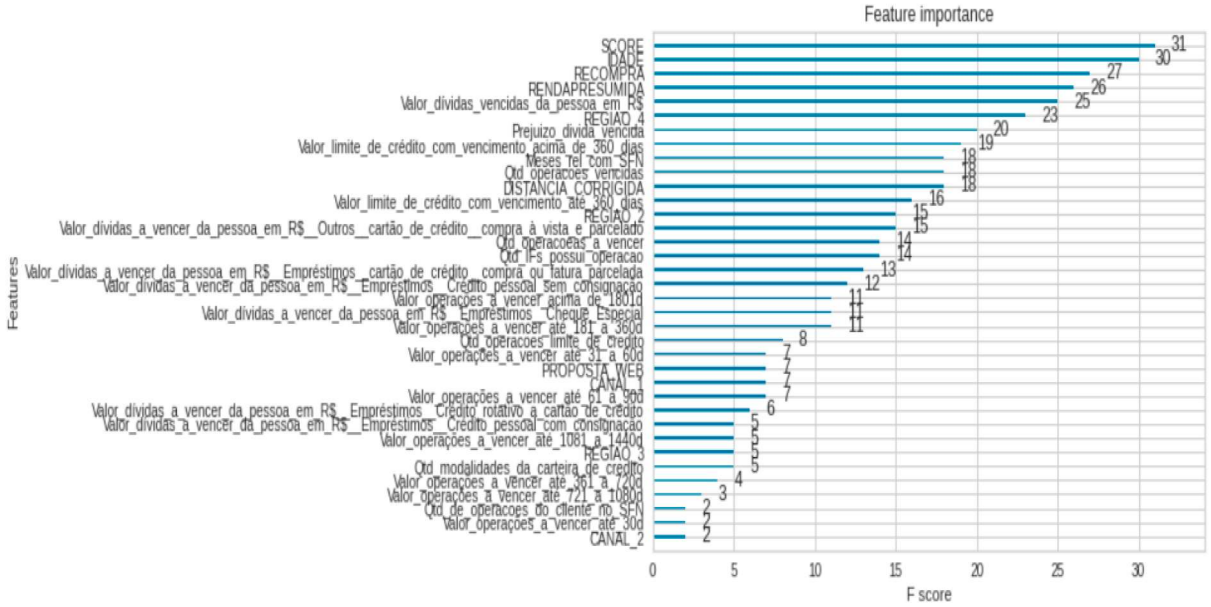
FONTE: a autora (2022).

O SVC alcançou 72,12% de acurácia, obtendo uma precisão de 73,74% nos inadimplentes e 67,07% nos não inadimplentes. A revocação foi de 87,48% para os positivos e de 45,02% para os negativos.

É possível analisar que ambos, SVC e XGBoost, tiveram acurácia muito próxima, mais alta que o Naive Bayes Multinomial. No entanto, o XGBoost atingiu uma revocação maior na identificação de clientes inadimplentes, que, analisando o contexto dos dados e o problema a ser resolvido, pode ser mais interessante.

Ao desdobrar-se nos resultados do XGBoost, então, é possível avaliar as variáveis de maior importância para a classificação dos clientes. A figura 7 apresenta estas variáveis por grau de importância.

FIGURA 7 - ANÁLISE DA IMPORTÂNCIA DAS VARIÁVEIS

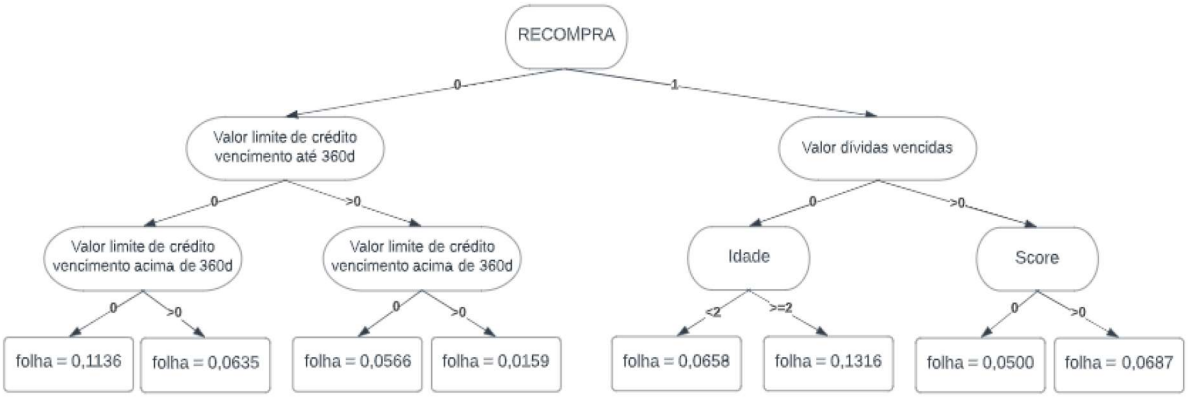


FONTE: a autora (2022).

A variável de maior importância segundo este modelo é o Score, que foi consultado no Sistema de Proteção ao Crédito (SPC), mas que não deve ser tomado unicamente como critério para a classificação. Este modelo destaca também a importância de variáveis como idade, recorrência na compra do crédito, renda presumida, dívidas vencidas do cliente, entre outras.

E por fim, com isso tem-se a construção da árvore de decisão a partir dos padrões descobertos pelo algoritmo XGBoost, que se encontra esquematizada na figura 8, a seguir. Ela foi visualmente reconstruída pela autora de maneira mais interpretável para humanos.

FIGURA 8 - ÁRVORE DE DECISÃO GERADA PELO XGBOOST



FONTE: a autora (2022).

Nota-se, visualizando a árvore, que os atributos de maior importância aparecem determinando os valores das folhas. E as folhas, por sua vez, representam a probabilidade do registro que apresenta as características dos ramos ser um bom cliente. Por exemplo, se um cliente é de recompra, tem a menor faixa de valor de dívidas vencidas e é um cliente com a faixa etária 2 ou 3 ele é considerado um cliente com melhor potencial de pagamento que qualquer cliente que não é de recompra.

4 CONSIDERAÇÕES FINAIS

A acurácia de aproximadamente 72% alcançada pelos modelos de Árvore de Decisão e SVM pode ser considerada baixa para determinadas aplicações, mas sob a ótica da análise de crédito é comum obter-se resultados um pouco menos eficazes devido à complexidade do comportamento financeiro dos indivíduos tomadores de crédito. Portanto, em decorrência do fato da base contar com valores reais, foi um resultado suficientemente significativo para responder ao problema da pesquisa.

A partir da análise dos resultados obtidos conclui-se, então, que é possível classificar os clientes inadimplentes da Empresa X de forma preditiva com base em seu comportamento financeiro e características pessoais, tal qual a idade, o *score*, o limite de crédito e o fato de ser um cliente de recompra (que já tomou crédito com a Empresa X anteriormente) ou não.

Neste trabalho também foi descrito todo o processo de Descoberta de Conhecimento na Base de Dados (KDD) de forma detalhada, e comparou-se mais de um algoritmo em busca do melhor desempenho atendendo também ao que foi proposto nos objetivos específicos.

Como trabalhos futuros, pretende-se aplicar ainda outros algoritmos amplamente usados no mercado de crédito, principalmente de Redes Neurais Artificiais (RNA), buscando cada vez mais efetividade na classificação. Da mesma maneira, é necessário também explorar outras variáveis com dados internos e externos dos clientes da Empresa X para verificar um cenário mais eficiente para a classificação dos clientes. De forma geral, ainda assim, com os métodos e resultados aqui demonstrados já seria possível revisar a política de crédito em busca de oportunidades de melhoria com uso de mineração de dados.

REFERÊNCIAS

BORGES, Mônica Erichsen Nassif. A informação como recurso gerencial das organizações na sociedade do conhecimento. **Ciência da informação**, v. 24, n. 2, 1995.

BORGES, Luiz Ferreira Xavier; BERGAMINI JUNIOR, Sebastião. O risco legal na análise de crédito. **Revista do BNDES**, Rio de Janeiro, v. 8, n. 16, dez. 2001.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1-29, 2009. Disponível em: https://rozero.webcindario.com/disciplinas/fbmg/dm/RT-INF_001-09.pdf. Acesso em 10 de mar. de 2022.

COSTA, Cláudio N. *et al*. Descoberta de conhecimento em bases de dados. **Revista Eletrônica: Faculdade Santos Dumont**, v. 2, p. 20, 2019.

DUARTE, E. N. Conexões temáticas em gestão da informação e do conhecimento no campo da Ciência da Informação. **Informação & Sociedade: Estudos**., João Pessoa, v. 21, n. 1, p. 159-173, jan./abr. 2011.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery. **AI Magazine**, p. 37-54,1996.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, v. 22, p. 686-690, 2009.

GUIMARÃES, Inácio Andruski; CHAVES NETO, Anselmo. Reconhecimento de padrões: metodologias estatísticas em crédito ao consumidor. **RAE eletrônica**, v. 1, p. 1-14, 2002.

HERLING, Luiz Henrique et al. A inadimplência nas instituições de ensino superior: um estudo de caso na instituição XZX. **Revista Gestão Universitária na América Latina-GUAL**, v. 6, n. 2, p. 126-142, 2013.

LEMOS, Eliane Prezepiorski; STEINER, Maria Teresinha Arns; NIEVOLA, Julio César. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. **Revista de Administração-RAUSP**, v. 40, n. 3, p. 225-234, 2005.

MAIA, Andréa do Socorro Rosa Silva. Inadimplência e recuperação de créditos. 2007.

MANFIO, Fernando. O risco nosso de cada dia: uma orientação objetiva para os profissionais da área. Barueri: **Estação das Letras Editora**, 2007.

MARCHIORI, P. Z. A ciência e a gestão da informação: compatibilidades no espaço profissional. **Ciência da Informação**, Brasília, v.31, n.2, maio/ago. 2002

MONTEIRO, Samuel Alves; DUARTE, Emeide Nóbrega. Bases teóricas da gestão da informação: Da gênese às relações interdisciplinares. **InCID: Revista de Ciência da Informação e Documentação**, v. 9, n. 2, p. 89-106, 2018.

MOURA, Gabriela Machado. Regressão Logística aplicada à análise de risco de crédito. 2018.

SERASA EXPERIAN. Tipos de Crédito: Qual é o ideal para você? Youtube, 10 de julho de 2019. Disponível em: https://www.youtube.com/watch?v=bOQuebmdoFk&t=93s. Acesso em: 02 de fev. de 2022.

SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business Intelligence e Análise de Dados para Gestão do Negócio**. 4 ed. Porto Alegre: Bookman, 2019.

SILVA, José Pereira. **Análise e decisão de crédito.** São Paulo: Atlas, 1988.

STEINER, Maria Teresinha Arns et al. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. **Gest Prod**, v. 13, n. 2, p. 325-37, 2006. Disponível em: http://www.scielo.br/pdf/gp/v13n2/31177.pdf. Acesso em: 11 de dez. de 2021.

VITAL, Luciane Paula; FLORIANI, Vivian Mengarda; VARVAKIS, Gregório. Gerenciamento do fluxo de informação como suporte ao processo de tomada de decisão: revisão. **Informação & Informação**, v. 15, n. 1, p. 85-103, 2010.

WEBER, Rosina de Oliveira et al. Sistema especialista difuso para análise de crédito. 1993.

WILSON, T. D. Information management. In:FEATHER, J.; STURGES, P. (Ed.). **International Encyclopedia of Information and Library Science**. Londres: Rout leg, 2002.