

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA DE CIÊNCIAS EXATAS E DA COMPUTAÇÃO
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO



Classificação de clientes utilizando mineração de dados

HISTEPENSON S. RIBEIRO

GOIÂNIA

2020

HISTEPENSON S. RIBEIRO

Classificação de clientes utilizando mineração de dados

Trabalho de Conclusão de Curso apresentado à Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Sibelius Lellis Vieira

Banca examinadora:

Prof. Me. Geraldo Valeriano Ribeiro

Prof. Me. Gustavo Siqueira Vinhal

GOIÂNIA

2020

HISTEPENSON DA SILVA RIBEIRO

Classificação de clientes utilizando mineração de dados

Trabalho de Conclusão de Curso aprovado em sua forma final pela Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás, para obtenção do título de Bacharel em Ciência da Computação, em ____/____/____.

Orientador: Prof. Dr. Sibelius Lellis Vieira

Prof. Me. Wilmar Oliveira de Queiroz

Prof. Me. Gustavo Siqueira Vinhal

Prof^a. Ma. Ludmilla Reis Pinheiro dos Santos
Coordenadora de Trabalho de Conclusão de
Curso

GOIÂNIA

2020

DEDICATÓRIA

Aos meus pais e amigos que foram os maiores apoiadores nessa trajetória.

AGRADECIMENTOS

Ao Professor Sibelius Lellis, orientador acadêmico, pelo apoio e paciência no decorrer do trabalho.

À Coordenação da Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás por ajudar, de forma inestimável, na execução deste projeto.

A todos que direta ou indiretamente colaboraram para materialização deste trabalho.

RESUMO

O desenvolvimento de descoberta de conhecimento em base de dados é um processo amplo que abrange diversos conceitos e processos. A mineração de dados é uma das várias etapas do processo, que contribui para tomadas de decisões nas empresas. No presente trabalho foram utilizados modelos de classificação para minerar dados de clientes relacionados a uma empresa atacadista. Mais especificamente, foram abordados os fatores referentes aos dados comerciais e financeiros dos clientes, considerando um período de 06 (meses). Após todo processamento de descoberta em base de dados, os clientes foram classificados considerando o método de árvore de decisão implementado no *software weka (Waikato Environment for Knowledge Analysis)*. Como resultado do trabalho, obteve-se a acurácia de 89 %, levando em consideração os clientes que compraram em ambos os períodos do ano estudado.

Palavras-Chave: classificação, árvore de decisão, descoberta de conhecimento, classificação de clientes, mineração de dados.

ABSTRACT

The development of knowledge discovery in databases is a broad process that encompasses several concepts and processes. Data mining is one of several stages in the process, which contributes to decision making in companies. In the present work, classification models were used to mine customer data related to a rice industry. More specifically, the factors related to customer purchase, regularity and default data were addressed, considering a period of 06 (six) months. After all the discovery processing in the database, the customers were classified considering the decision tree method implemented in the weka software (Waikato Environment for Knowledge Analysis), having been classified as: excellent, good, regular and bad. The result of the work, obtained an accuracy of 89%, taking into account the customers who bought in both quarter of the year studied.

Keywords: classification, decision tree, knowledge discovery, customer classification, data mining.

LISTA DE ILUSTRAÇÕES

Figura 1 – Etapas do processo de descoberta de conhecimento	16
Figura 2 – Árvore de decisão	19
Figura 3 – Proporção dos dados	20
Figura 4 – Tela inicial weka	24
Figura 5 – Tela principal do módulo <i>explorer</i>	25
Figura 6 – Informações após <i>upload</i> do arquivo	26
Figura 7 – Parâmetros de processamento weka	27
Figura 8 – Informações do resumo do processamento	27
Figura 9 – Informações do resumo do processamento segunda parte	28
Figura 10 – Distribuição do atributo R	30
Figura 11 – Distribuição do atributo V	31
Figura 12 – Distribuição do atributo A	32
Figura 13 – Informações básicas do arquivo para classificar	35
Figura 14 – Estrutura da árvore de decisão – modelo 1	36
Figura 15 – Matriz de confusão – modelo 1	36

LISTA DE TABELAS

Tabela 1 – Regra de negócio para atributo R	30
Tabela 2 – Regra de negócio para atributo V	31
Tabela 3 – Regra de negócio para atributo A	32
Tabela 4 – Tabela de pesos para o atributo R	33
Tabela 5 – Tabela de pesos para o atributo V	33
Tabela 6 – Tabela de pesos para o atributo A	33
Tabela 7 – Atributos do primeiro trimestre e classificação do segundo trimestre	34
Tabela 8 – Acurácia do modelo 1	35
Tabela 9 – Acurácia do modelo 2	35
Tabela 10 – Precisão e sensibilidade para o modelo 1	37

SUMÁRIO

1	INTRODUÇÃO	12
1.1.	Justificativa	12
1.2.	Objetivo	13
1.2.1	Objetivo geral	13
1.1.2	Objetivos específicos	13
1.3	Estrutura da monografia	14
2	REFERENCIAL TEÓRICO	15
2.1	Descoberta de conhecimento em base de dados - DCBD	15
2.2	Etapas do processo DCBD	15
2.3	Mineração de dados	16
2.4	Aprendizagem Supervisionada	17
2.5	Árvore de decisão	18
2.6	Ferramenta Weka	21
2.5	Aplicação de mineração de dados para classificação de clientes	21
2.6	Trabalhos relacionados	21
3	MATERIAS E MÉTODOS	23
3.1	Materiais	23
3.1.2	<i>Hardware</i>	23
3.1.3	<i>Software</i>	23
3.1.4	Weka	24
3.2	Métodos	28
3.2.1	Pesquisa bibliográfica	28
3.2.2	Pesquisa exploratória	28
4	RESULTADOS	29
4.1	Pré-processamento	29
4.2	Atributos	29
4.2.1	Atributo R	30
4.2.2	Atributos V e A	31
4.3	Divisão clientes em faixas	32
4.4	Experimentos	34
4.5	Análise dos resultados obtidos	37
5	CONCLUSÃO	38

5.1 Trabalhos futuros.....	38
REFERÊNCIAS BIBLIOGRÁFICAS.....	39

1 INTRODUÇÃO

O modelo de negócios no qual as empresas estão inseridas torna o mercado cada vez mais competitivo, exigindo tecnologias mais alinhadas ao negócio para se manter próspero neste ambiente. Os grandes bancos de dados das empresas contêm informações ocultas que podem ajudá-las a melhorar o relacionamento com os clientes, otimizar atribuição de crédito, prever vendas, traçar perfis de clientes, entre outros propósitos, e são exemplos destas tecnologias. O conceito de crédito segundo Schrickel (2000) é:

Crédito é 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.

Dessa forma, considera-se um interesse para as empresas a possibilidade de examinar grandes volumes de informação, para extrair informações assertivas para tomada de decisão no segmento de crédito.

Nesse processo, os dados estruturados podem assumir um papel importante para estabelecer normas de reestruturação e reposicionamento no mercado com maior rapidez e eficiência (LIRA et al., 2008).

A Mineração de Dados é uma das tecnologias mais promissoras da atualidade. Um dos fatores deste sucesso é o fato de dezenas, e muitas vezes centenas, de milhões de reais serem gastos pelas companhias na coleta dos dados e, no entanto, nenhuma informação útil é identificada (LAROSE, 2005). O processo de descoberta em banco de dados auxilia na geração de informações úteis, confiáveis e muitas vezes fáceis de serem interpretados. O volume de dados gerado pode ser processado em ferramentas de mineração de dados, concebendo informações úteis para o negócio.

Diante disso, o estudo tem como objetivo aplicar mineração de dados para auxiliar no processo decisório de fornecer crédito aos clientes e a realizar previsão do comportamento analisando dados históricos.

1.1. Justificativa

Na concepção das empresas que fornecem crédito para seus clientes para compra de mercadoria, a necessidade de ter informações assertivas dos clientes

aumenta. Informações como histórico de pagamento, inadimplência, volume de compras, entre outras informações podem definir o perfil do cliente.

Dessa forma, para que o cliente possa fazer negócio com a empresa, o mesmo é avaliado através de uma análise de crédito e potencial de compra para saber se os dados informados são coerentes com a situação da empresa. Obviamente, no processo de liberação de crédito, embora entenda-se que o cliente tem a obrigação de pagar pelo serviço ou produto que recebeu, e no cotidiano das empresas, o recebimento às vezes requer muita negociação e cobrança, surgindo à necessidade de melhor avaliar os clientes antes da venda. Os maiores desafios das empresas perante o fornecimento de crédito são as informações comerciais e financeiras, importante indicador indireto de saúde financeira dos clientes, sendo interligado diretamente ao processo de análise de crédito. O índice de inadimplência é o indicador mais prejudicial para o negócio, tendo impacto em diversas áreas como planejamento e desenvolvimento.

Neste sentido, como o fornecimento de crédito para os clientes é de grande importância, a implementação de novas técnicas para melhorar a tomada de decisão no processo de crédito se torna crucial para o planejamento financeiro e estruturação comercial. Este processo de tomada de decisão deve levar em conta todos os dados pertinentes que a empresa possui sobre os clientes, tais como seu capital social, sua posição de mercado, tempo de relacionamento, entre outras. Como nem todas estas informações estão disponíveis para quem faz a análise de crédito, uma análise cuidadosa é bem importante neste sentido e a mineração de dados pode auxiliar neste processo, visto que oferece uma série de técnicas que podem prever ou classificar um cliente conforme o histórico de outros clientes, através da formação de um perfil.

1.2. Objetivo

1.2.1 Objetivo geral

Esse trabalho tem como objetivo a classificação de clientes utilizando atributos comerciais e financeiros padrão dos clientes, visando identificar perfis dos clientes utilizando mineração de dados, para possibilitar uma análise de comportamento e crédito. Esses clientes são classificados em 4(quatro) classes, conforme seu desempenho de atributos financeiros e comerciais.

1.1.2 Objetivos específicos

1. Identificar padrões de comportamento de clientes.
2. Classificar os clientes de acordo com atributos comerciais e financeiros.
3. Compreender melhor a carteira de clientes.
4. Ajudar no apoio ao fornecimento de crédito.

1.3 Estrutura da monografia

No capítulo 2 é apresentado o embasamento teórico para elaboração do trabalho e a descrição detalhada do objeto de estudo.

No capítulo 3 são apresentados os materiais (*software*, dados e *hardware*), e metodologias adotadas. No capítulo 4 são apresentados os resultados do trabalho. No capítulo 5 são apresentadas as conclusões pertinentes ao trabalho.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados os conteúdos utilizados para a fundamentação teórica do trabalho, abordando de forma objetiva o tema e seus conceitos gerais.

2.1 Descoberta de conhecimento em base de dados - DCBD

Em meio a um enorme volume de dados gerados pelas empresas diariamente, incide a necessidade de estruturação dos dados para que seja possível extrair conhecimento. As empresas estão cada vez mais dependentes dos dados para a melhoria na tomada de decisões (FEIJÓ, 2013).

O processo conhecido como *Knowledge Discovery from Databases* - KDD é o processo iterativo para identificar nos dados novos padrões que sejam válidos, potencialmente úteis e interpretáveis (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996). Em um nível abstrato, o campo KDD se preocupa com o desenvolvimento de métodos e técnicas para entender os dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Para implementar processo de DCBD faz-se necessário ter objetivos claros e definidos, saber quais resultados deseja-se atingir e ter conhecimento da regra de negócio e aplicação.

2.2 Etapas do processo DCBD

O processo de descoberta de conhecimento envolve várias etapas complexas, como a preparação dos dados, a busca por padrões, a avaliação e o refinamento do conhecimento (HALMENSCHLAGER, 2002). Conforme análise, o processo de descoberta de conhecimento foi dividido em cinco etapas, seleção dos dados, pré-processamento, transformação, mineração e interpretação, que são executadas de forma interativas e iterativas que tem como objetivo identificar padrões nos dados.

A etapa de seleção se desenvolve através do entendimento das regras de negócio e aplicação para mensurar o objetivo final do cliente (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Nesta etapa também é definido quais dados deverão ser utilizados para descoberta de conhecimento.

No pré-processamento, que conforme Fonseca (2014), é um processo de grande importância para remoção de dados fora do domínio, com ausência de valores

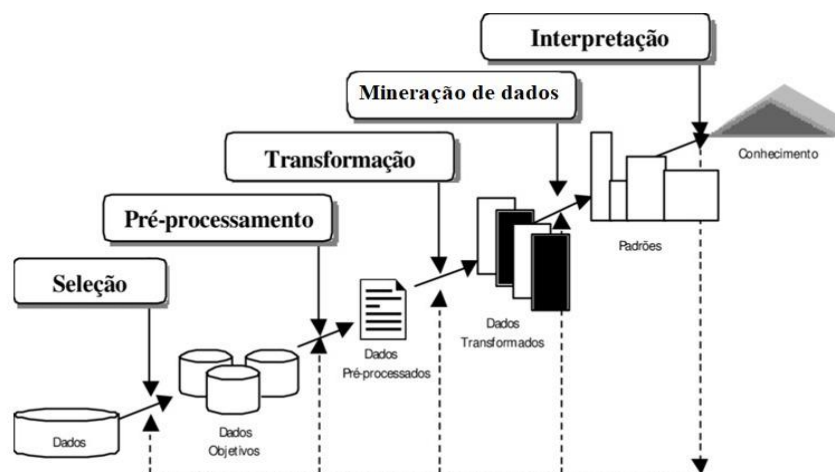
e valores que não condizente para o atributo. Deste modo, os dados tornam-se mais confiáveis no processo de mineração de dados.

Na transformação os dados são convertidos em um modelo para que possam ser submetidos à etapa posterior. A mineração de dados trata da aplicação de técnicas, implementadas por meio de algoritmos, capazes de receber, como entrada, um conjunto de fatos ocorridos no mundo real e devolver, como saída, um padrão de comportamento (SILVA; MARQUES; BOSCARIOLI, 2006).

Por fim na visualização e interpretação, que segue após da fase de mineração de dados deve-se interpretar o conhecimento apresentado, verificando a relevância (ou não) na obtenção dos padrões e, com isso, analisar a eficácia do método aplicado na etapa de mineração de dados (GÓES; STEINER, 2016)

Na Figura 1 são ilustradas as cinco etapas do processo para melhor entendimento.

Figura 1 - Etapas do processo de descoberta de conhecimento



Fonte: Fayyad; Piatetsky-Shapiro; Smyth, 1996

2.3 Mineração de dados

Os avanços da chamada “era da informação” têm colocado como desafio a implementação de técnicas que consigam mensurar e descobrir padrões relevantes na crescente massa de dados, resultante sobretudo, do aumento da complexidade

nas tarefas operacionais e decisórias (GONÇALVES; PACHECO; MORALES, 2001). Portanto, conforme Angeloni e Reis (2006) a informação pura e simples já não garante um diferencial competitivo. Baseado no requisito do cenário descrito que as empresas vivem, existe a necessidade de ser aderir a tecnologias e ferramentas que ajudam no processo de extração de conhecimento, e dessa forma, justifica-se o uso de ferramentas de mineração de dados.

A definição para mineração de dados é dada da perspectiva do aprendizado de máquina: "Mineração de Dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A mineração de dados também pode ser definida em termos do uso de técnicas estatísticas ou analíticas para processar e analisar dados brutos para encontrar padrões e tendências significativos (CHIU; TAVELLA, 2006). Para realização dos processos, as tarefas de mineração são divididas em 2 (dois) níveis (SILVA; MARQUES; BOSCARIOLI, 2006). Um deles é a tarefa de previsão na qual se prevê uma classe baseado na análise dos atributos. Já as tarefas descritivas têm o objetivo de encontrar padrões que descrevem os dados de maneira que o ser humano possa interpretar.

Na definição do escopo do problema é definido qual técnica será implementada, uma vez que cada método aborda uma técnica especializada, como classificação, regressão, associação ou agrupamento. Neste trabalho será usado o modelo de aprendizado supervisionado, para o qual os métodos são providos com um conjunto de dados que possuem uma variável alvo pré-definida e os registros são categorizados em relação a ela (CAMILO; SILVA, 2009). De acordo com o modelo escolhido, o método escolhido no processo de *data mining* neste trabalho é a classificação, que consiste em examinar uma certa característica nos dados e atribuir uma classe previamente definida (CORTÊS; PORCARO; LIFSHITZ, 2002).

2.4 Aprendizagem Supervisionada

Conforme Lorena e Carvalho (2007), as técnicas de aprendizado de máquinas empregam um princípio de inferência denominado indução, no qual é possível obter

conclusões genéricas a partir de um conjunto particular de exemplos. Esta classe de algoritmos possui esta designação devido a sua aprendizagem ser orientada para encaixar em um rótulo de classe, por exemplo.

Estes algoritmos são preditivos, devido suas tarefas de mineração desempenhar conclusões nos dados com o objetivo de fornecer previsões ou tendências, adquirindo informações não disponíveis a partir dos dados disponíveis. Com a classificação, através de algoritmos supervisionados é possível determinar o valor de um atributo através dos valores de um subconjunto dos demais atributos da base de dados.

2.5 Árvore de decisão

Existem diversos algoritmos de aprendizado que provocam segmentação nos dados. Algoritmos que induzem árvores de decisão pertencem ao grupo de algoritmos *Top Down Induction of Decision Trees* — TDIDT. Uma definição objetiva para árvore de decisão é uma estrutura de dados definida recursivamente como um nó folha que corresponde a uma classe ou um nó de decisão que contém uma validação sobre o atributo. Para cada resultado da validação existe uma aresta para uma sub árvore até que seja atingido o final da árvore, ou simplesmente o nó folha (BREIMAN et al,1984).

Uma das técnicas mais utilizadas em mineração de dados é a árvore de decisão (SILVA; MARQUES; BOSCARIOLI, 2006). Segundo (CREPALDI et al., 2010), concluiu que:

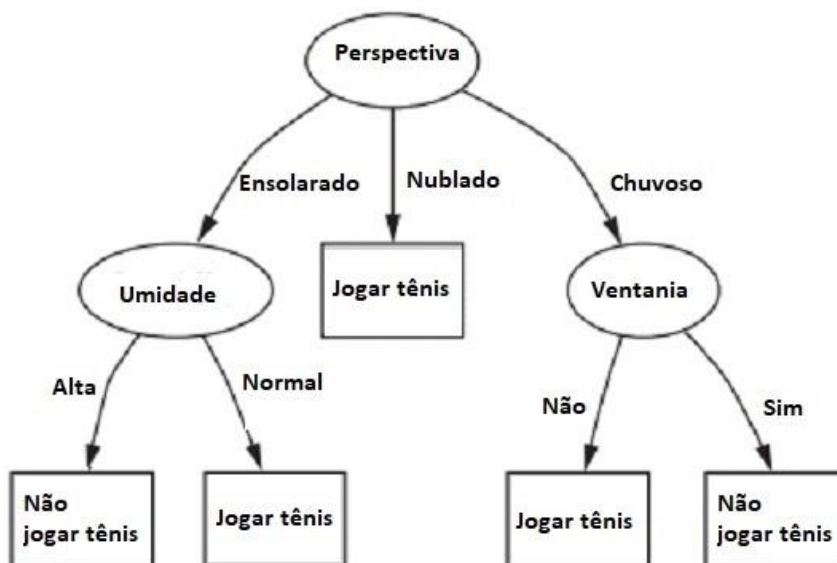
Árvores de Decisão são métodos de classificação de dados no contexto da chamada Mineração de Dados (*Data Mining*). Podem ser usadas em conjunto com a tecnologia de indução de regras, mas são as únicas a apresentar os resultados hierarquicamente (com priorização). Nelas, o atributo mais importante é apresentado na árvore como o primeiro nó, e os atributos menos relevantes são mostradas nos nós subsequentes. A vantagem principal das Árvores de Decisão é a tomada de decisões levando em consideração os atributos mais relevantes, além de compreensíveis para a maioria das pessoas.

Desta forma, este algoritmo é bastante utilizado, podendo trabalhar com atributos quantitativos e qualitativos, aumentando o nível de compreensão da análise.

Essa estratégia permite estimar como a árvore generaliza os dados e adapta-se a novas situações, além de determinar a proporção de erros e acertos ocorridos na construção da árvore (BRADZIL, 1999).

A Figura 2, mostra uma árvore de decisão para classificar uma indicação para jogar ou não jogar, com atributos quantitativos referente a clima, umidade e temperatura e vento. Através dessas informações, o jogador decidirá se ira ou não jogar tênis.

Figura 2 – Árvore de decisão



Fonte: Witten; Frank and Hall, 2011)

A chave para o sucesso de um algoritmo de aprendizado por árvore de decisão depende do critério utilizado para escolher o atributo que particiona o conjunto de exemplos em cada iteração (MONARD; BARANAUSKAS, 2003). Neste sentido, as possibilidades para escolha dos atributos são:

- 1 Aleatória: seleciona qualquer atributo, independentemente de qualquer fator;
- 2 Menos valores: seleciona o atributo com a menor quantidade de valores possíveis;
- 3 Mais valores: seleciona o atributo com a maior quantidade de valores possíveis;

- 4 Ganho máximo: seleciona o atributo que possui o maior ganho de informação utilizando o cálculo de entropia.

No contexto de análise dos atributos, considera o cálculo da entropia do conjunto de dados. Para realizar o cálculo da entropia é considerado a probabilidade de determinado atributo aparecer. A Figura 3 mostra probabilidade dos valores dos atributos, para árvore de decisão mostrada na Figura 2.

Figura 3 – Proporção dos dados

Perspectiva	Temperatura		Umidade			Ventania		Jogar Tênis					
	Sim	Não	Classificação	Sim	Não	Sim	Não	Sim	Não				
ensolarado	2	3	Quente	2	2	Alta	3	4	Sim	6	2	9	5
nublado	4	0	Medio	4	2	Baixa	6	1	Não	3	3		
chuvoso	3	2	Agradável	3	1								
Proporção dos atributos													
ensolarado	02/09	3/5	Quente	2/9	2/5	Alta	3/9	4/5	Sim	6/9	2/5	9/14	5/14
nublado	4/09	0/5	Medio	4/9	2/5	Baixa	6/9	1/5	Sim	3/9	3/5		
chuvoso	3/09	2/5	Agradável	3/9	1/5								

Fonte: Witten, Frank and Hall, 2011

Essa etapa considera as proporções dos atributos, que possibilita o cálculo de entropia e ganho de informação para o conjunto de dados.

Para a verificação da proximidade do resultado do modelo analisado com seu valor referência real, é utilizado o conceito de acurácia. A acurácia é definida como proximidade de concordância entre um valor de quantidade obtido por medição e o verdadeiro valor do mensurando (MENDITTO; PATRIARCA; MAGNUSSON, 2007).

Analisando apenas a taxa de acerto não é possível identificar se um modelo é bom ou não. A matriz de confusão vem para auxiliar na avaliação de um modelo. Existem outros itens que podem ser utilizados, porém a matriz é ponto de partida para os demais.

A matriz de confusão é a forma de representação da qualidade obtida de uma classificação, sendo expressa por meio da correlação de informações dos dados de referência (compreendido como verdadeiro) com os dados classificados (PRINA; TRENTIN, 2020). Esse recurso permite tirar conclusões mais assertivas para classificação, auxiliando também na melhor análise dos resultados obtidos.

2.6 Ferramenta Weka

Existem diversas ferramentas que dão apoio na execução dos processos de mineração de dados, que implementam os algoritmos, e portanto, essa tarefa de escolha do *software* nem sempre é trivial. Neste trabalho foi utilizado o *software* Weka (*Waikato Environment for Knowledge Analysis*), que é uma solução composta de diversos algoritmos implementados, responsáveis por realizar diversas tarefas como:

1. Classificação
2. Regressão
3. Clusterização
4. Regras de associação

2.7 Aplicação de mineração de dados para classificação de clientes

Em qualquer ramo de atividade econômica é necessário ter um bom conhecimento de como seus clientes se comportam e quais estão mais propensos a trazerem lucro ou prejuízo. A classificação de clientes em grupos proporciona a melhor compreensão da relevância para o negócio. Conforme colocado na Seção 2.3, a técnica de classificação é de grande relevância para ajudar no apoio de análise de crédito. O crédito cedido pelas empresas é o meio que facilita o volume de vendas, sendo ele o maior responsável pelos resultados das organizações (CARMO et al., 2018).

Diante do cenário e na sua grande importância para o negócio, a técnica de classificação pode ser utilizada para definir classes para os clientes tais como bom, diamante, ouro, prata e bronze com base nos seguintes atributos:

1. Quantidade da compra realizada.
2. Prazos cumpridos
3. Tempo de relacionamento com cliente

Como resultado do processo a classificação, que também é uma forma de previsão, possibilita entender como os clientes se comportarão no futuro.

2.8 Trabalhos relacionados

Esta seção visa apresentar alguns trabalhos relacionados à Mineração utilizando algoritmos de classificação.

No trabalho de Cardoso (2017) classifica-se os clientes de uma distribuidora de medicamentos em cliente Diamante, Ouro e Bronze, utilizando mineração de dados baseado em árvore de decisão, implementado na ferramenta RapidMiner. Para realizar da tarefa de classificação, foram definidos os atributos de valor de venda bruta, quantidade de dias que são efetuadas compras e quantidade de produto que são comprados. Os dados analisados são do trimestre do ano de 2017. Aplicando os conceitos de descoberta de conhecimento, os resultados foram ilustrados na forma gráfica, utilizando ferramenta de inteligência de negócios.

Para o trabalho em Fonseca & Namen (2013) são utilizados modelos de classificação para minerar dados relacionados a aprendizagem de Matemática e ao perfil de professores do ensino fundamental. Especificamente, foram abordados os fatores referentes aos educadores do Estado do Rio de Janeiro que influenciam absolutamente e desfavoravelmente no desempenho dos alunos do ensino básico nas provas de Matemática.

Os dados utilizados para extrair estas informações são disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira que avalia o sistema educacional brasileiro em diversos níveis e modalidades de ensino, incluindo a Educação Básica, cuja avaliação, que foi objetivo do estudo, é realizada pela Prova Brasil. A partir desta base, foi aplicado o processo de Descoberta de Conhecimento em Bancos de Dados (KDD - *Knowledge Discovery in Databases*). Os padrões foram extraídos dos modelos de classificação gerados pelas técnicas árvore de decisão, indução de regras e classificadores Bayesianos, no qual os algoritmos estão implementados no *software*.

3 MATERIAS E MÉTODOS

Neste capítulo são apresentados materiais e métodos que são utilizados para o desenvolvimento do trabalho.

3.1 Materiais

Os dados foram utilizados são baseados em um padrão de atributos de uma empresa atacadista de porte grande. São dados comerciais e financeiros, que devem ser materializados como atributos indicando posição financeira dos clientes, da forma como é entendida pela empresa. São relacionados os dados do primeiro semestre do ano de 20XX, totalizando um total de N clientes que compraram nesse período, com N normalmente na ordem de milhar.

A mineração de dados possui alguns métodos e atividades, sendo nesse trabalho utilizado uma técnica de classificação, com a criação de classes de clientes baseadas nos atributos tais como financeiros e comerciais, que serão denominados atributo V, atributo A e atributo R.

Essa técnica de mineração de dados é implementada através de um algoritmo baseado em árvore de decisão que conforme Garcia (2003), refere-se, a uma técnica estatística que se emprega da divisão de um conjunto de exemplos em diversos subconjuntos, utilizados no treinamento do modelo, até que cada subconjunto pertença a uma única classe ou que a mesma seja majoritária, não sendo necessárias novas divisões.

3.1.2 Hardware

Foi utilizado notebook dell XPS com as seguintes configurações: Processador core i7 e memória de 16 Gb.

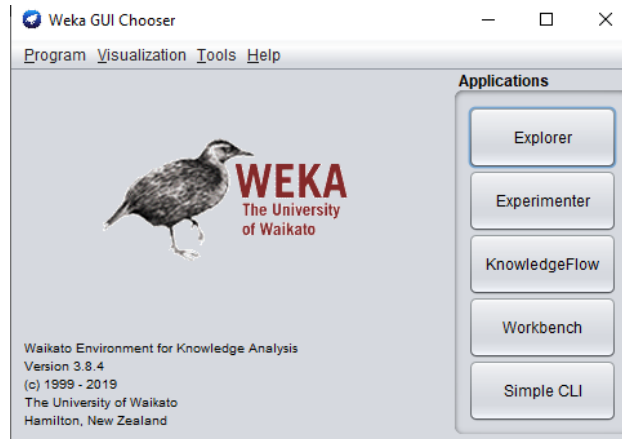
3.1.3 Software

O banco de dados utilizado pela empresa é *SQL SERVER*. Para estruturação das consultas em SQL - *Structured Query Language* foi utilizado o Sistema de gerenciamento de banco – SGBD dbvisualizer. Para o processo de mineração será usado o *software* Weka.

3.1.4 Weka

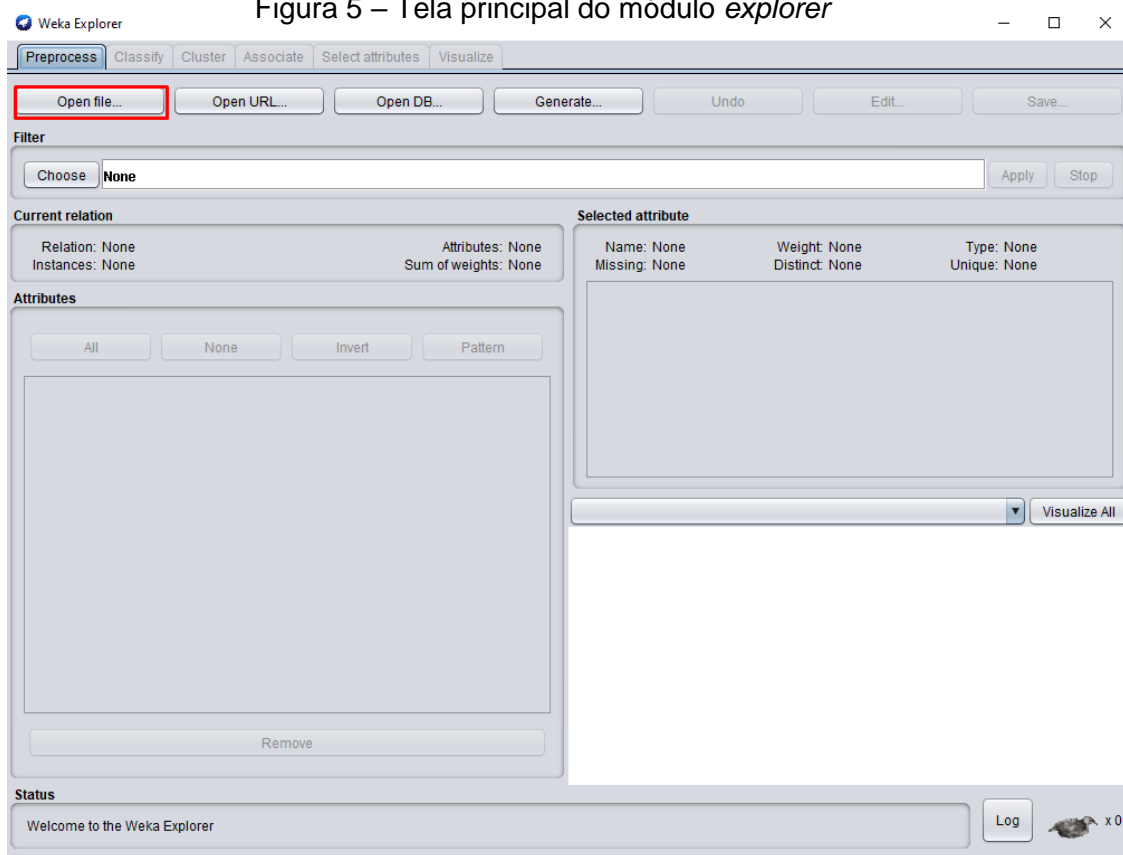
Foi utilizado a ferramenta Weka para realizar os processamentos, cuja tela inicial é mostrada na Figura 4.

Figura 4: Tela inicial weka.



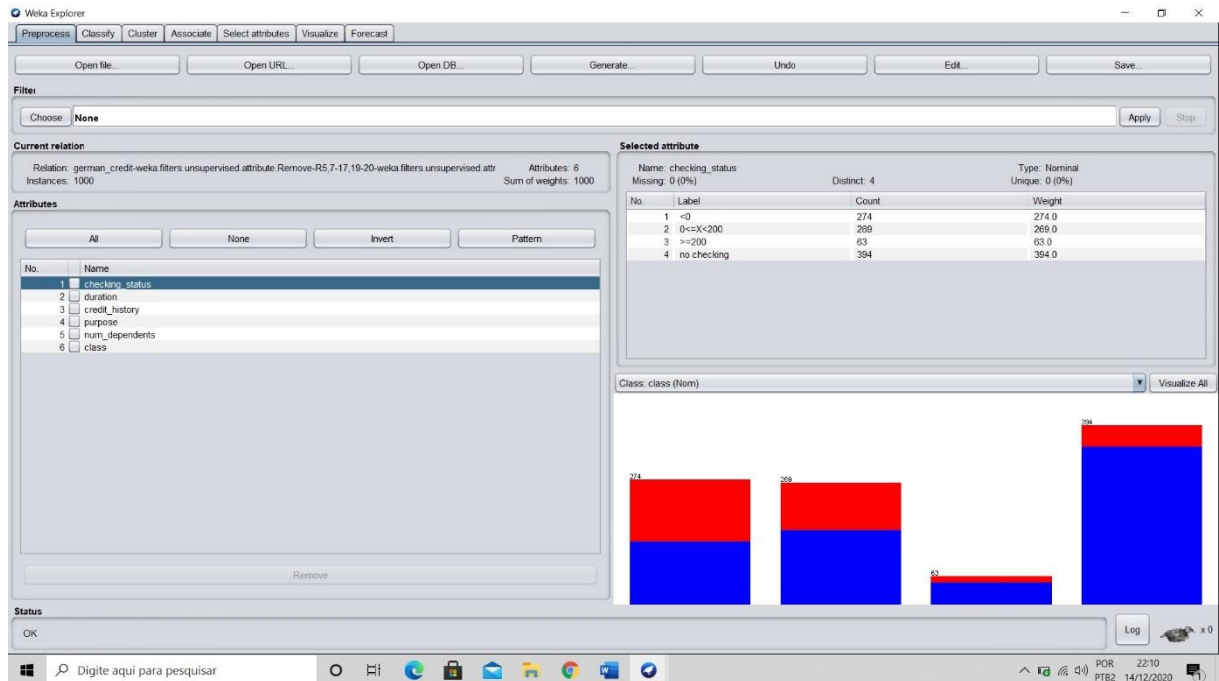
Fonte: Elaborado pelo autor.

Neste trabalho foi utilizado o primeiro módulo da ferramenta. Dentro do módulo *Explorer* foi possível realizar as operações de configuração para processamento. Dentro do módulo, existem todos os recursos para classificação, clusterização, associação, seleção de atributos e visualização dos dados. Neste trabalho foi utilizado apenas a parte de classificação. O primeiro passo após escolher o módulo é selecionar o arquivo no formato arff. Este arquivo é disponibilizado pela própria ferramenta responsável pela estruturação das informações que devem ser processadas. Na Figura 5 é mostrada a opção *open file*, recurso utilizado para fazer iniciar o trabalho com o arquivo.

Figura 5 – Tela principal do módulo *explorer*

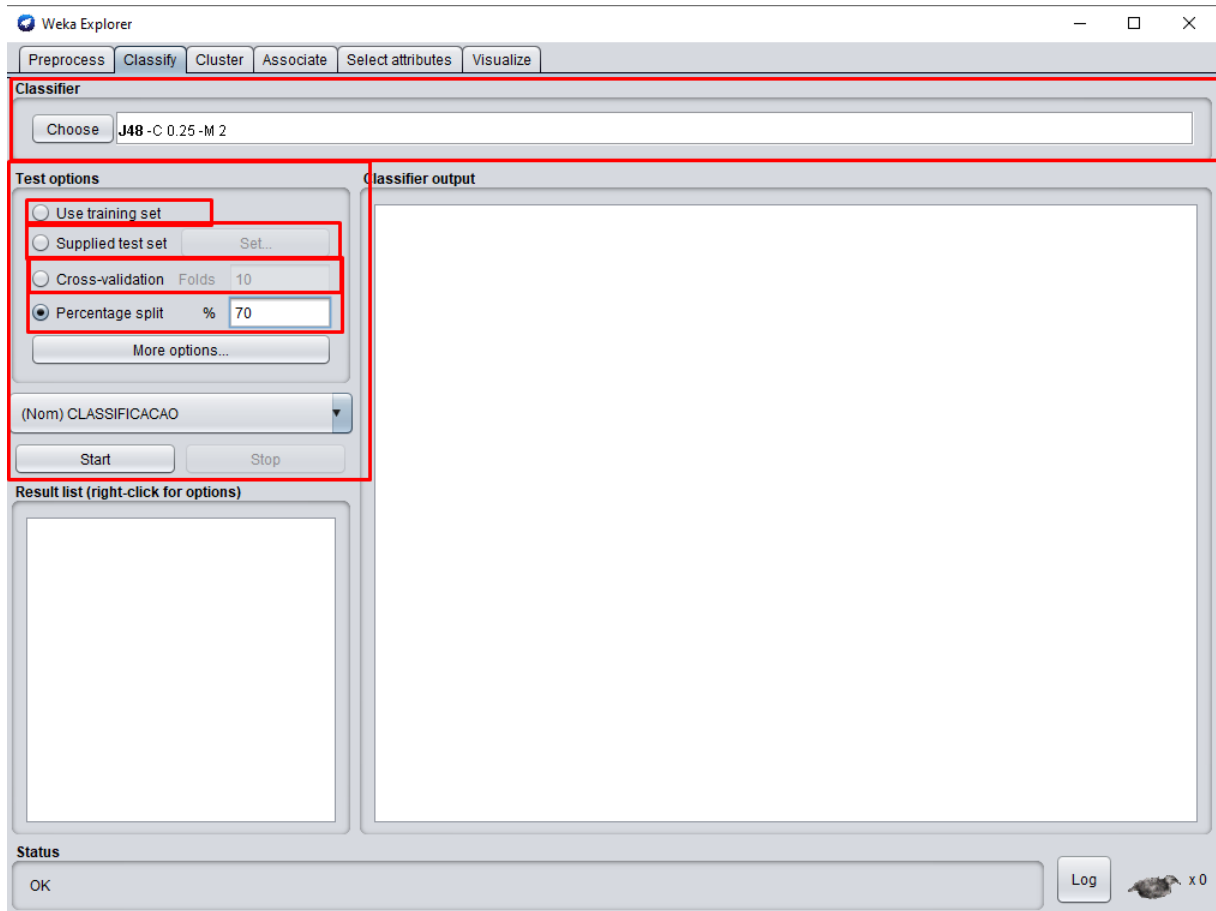
Fonte: Elaborado pelo autor

Após os dados serem incluído na ferramenta, é possível realizar a conferência dos dados, verificação dos atributos, visualização e distribuição dos atributos comparado com suas classes, edição de alguma informação dos registros, entre outros recursos conforme Figura 6.

Figura 6 – Informações após *upload* do arquivo.

A Figura 7 mostra os parâmetros que devem ser configurados para treinamento, sendo possível usar o arquivo inteiro para treino - *use training set*, fazer um novo *upload* de um arquivo para ser utilizado como teste - *supplied test set*, realização do processo de *cross-validation* e dividir o arquivo em porcentagem, uma parte para treino e uma parte para teste - *percentage split*. Neste trabalho foi utilizado a opção *Percentage Split*. Nesta etapa, também é feita a seleção da técnica de mineração de dados. A técnica escolhida foi árvore de decisão, implementado em java. O algoritmo usado foi o J48.

Figura 7 – Parâmetros de processamento weka.



Fonte: Elaborado pelo autor

Após parametrização, os resultados são mostrados na ferramenta, bem como informações como a árvore de decisão no formato texto e gráfica, porcentagem de acerto, quantidade de registro utilizado para treino e teste, resultado da classificação em percentual, entre outras. Nas Figuras 8 e 9 são apresentados os resultados do processamento.

Figura 8 – Informações do resumo do processamento.

Correctly Classified Instances	758	89.917 %
Incorrectly Classified Instances	85	10.083 %
Kappa statistic	0.8458	
Mean absolute error	0.0827	
Root mean squared error	0.2069	
Relative absolute error	24.9396 %	
Root relative squared error	50.8172 %	
Total Number of Instances	843	

Fonte: Elaborado pelo autor

Figura 9 – Informações do resumo do processamento segunda parte.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,965	0,041	0,825	0,965	0,890	0,869	0,978	0,839	DIAMANTE
0,926	0,066	0,923	0,926	0,925	0,859	0,955	0,915	OURO
0,984	0,042	0,908	0,984	0,945	0,921	0,963	0,858	PRATA
0,203	0,001	0,923	0,203	0,333	0,418	0,702	0,295	BRONZE
0,899	0,050	0,902	0,899	0,883	0,849	0,944	0,842	

Fonte: Elaborado pelo autor

Nesta tela é possível analisar as informações pertinentes aos resultados do modelo, tais como precisão, sensibilidade, falsos positivos, falsos negativos entre outros indicadores.

3.2 Métodos

Nesta seção são apresentados os métodos para uma proposta de estudo de caso prático referente à mineração de dados.

3.2.1 Pesquisa bibliográfica

O primeiro passo foi o estudo e análise de trabalhos científicos relacionados a descoberta de conhecimento em banco de dados – DCBD, focado em apoio na decisão no segmento empresarial. Após análise inicial, foi estudado impacto da aplicação de mineração de dados aplicado no processo DCBD com embasamento em diversos autores referência na área. O material bibliográfico que contempla o trabalho relaciona os principais conceitos do assunto, com abordagens práticas e teóricas.

3.2.2 Pesquisa exploratória

Foi realizado o levantamento de quais são os dados importantes para uma empresa de grande porte de uma empresa comercial, para melhor entendimento do problema. Com o escopo do problema definido, foi executado o levantamento das informações utilizando atributos padrão para o problema.

4 RESULTADOS

Neste capítulo são mostrados os processos que resultaram nos dados finais do trabalho, assim como a aplicação do algoritmo de árvore de decisão estabelecido, e as informações referentes ao processo de classificação feito pela ferramenta Weka. A análise foi realizada utilizando as informações de caráter comercial e financeiro dos clientes.

4.1 Pré-processamento

Os atributos utilizados para o processamento foram gerados através de uma análise do primeiro trimestre de 20XX. Para o objetivo de classificação, foram utilizados os atributos do primeiro trimestre e classificação prévia do segundo trimestre.

O primeiro trimestre de 20XX é usado para verificar ou testar a acurácia da previsão, e portanto, são dados de treino e teste mas não são usados para classificar o cliente. O segundo trimestre de 20XX é usado para classificar o cliente, portanto, apenas para classificar, não sendo usados como atributos para treino ou teste, e sim para, através dos valores dos atributos, calcular a classe do cliente.

Para o processo de classificação, foram definidos 2 modelos: modelo 1 para os clientes com compram nos dois períodos, e modelo 2 clientes que não necessariamente compram nos dois períodos. Os clientes que não compraram no segundo trimestre foram indicados como clientes da classe bronze, visto que todos os clientes que deixam de comprar são motivados por dificuldades financeiras no período anterior ou por escolha de outro fornecedor.

4.2 Atributos

De acordo com as regras da empresa foram analisados os melhores atributos que descrevem o comportamento do cliente, a partir do primeiro trimestre. Foram definidos 3 atributos:

- 1 R.
- 2 V.
- 3 A.

Para representação gráfica da distribuição dos atributos será considerado o modelo 2, clientes que não necessariamente compraram nos dois trimestres. Porém, nos dois modelos mantém-se a mesma proporção e heterogeneidade.

4.2.1 Atributo R

Para o atributo R, foi considerado a possibilidade definida na quantidade de elementos de regularidade. Foi transformado o atributo em valor nominal através de condições definidas por uma situação de compra regular. A Tabela 1 mostra regra utilizada para transformação do atributo quantitativo em qualitativo de acordo regra da empresa.

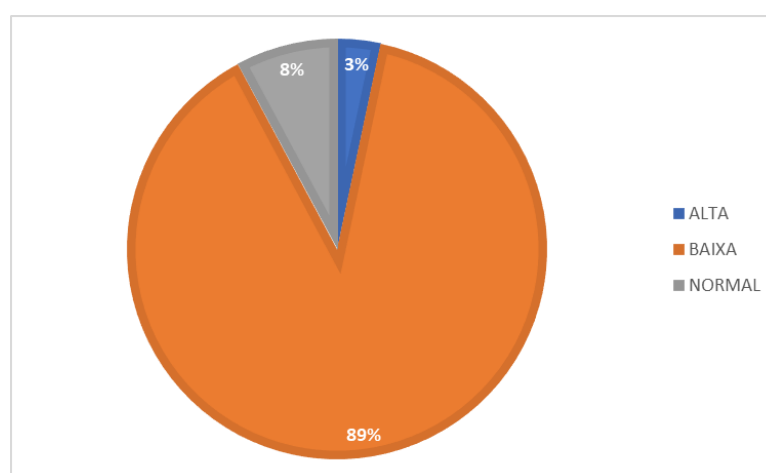
Tabela 1 – Regra de negócio para atributo R

Atributo R	Quantidade
Baixa	1
Normal	2
Alta	>3

Fonte: Elaborado pelo autor.

O atributo R foi definido para os clientes de forma que um atributo R alta é muito bom, indicando uma boa relação com a empresa, um atributo R Normal é a situação desejável e o atributo R baixa tem uma relação fraca com a empresa. A Figura 10 mostra a distribuição do atributo em forma gráfica.

Figura 10 – Distribuição do atributo R.



Fonte: Elaborado pelo autor.

Através do gráfico de distribuição do atributo R, percebe-se que o atributo possui baixa heterogeneidade, mostrando que os clientes do primeiro trimestre têm um relacionamento de baixa frequência.

4.2.2 Atributos V e A

O atributo V é calculado de acordo com a média do valor deste atributo do primeiro trimestre. A Tabela 2 mostra regra utilizada para transformação do atributo quantitativo.

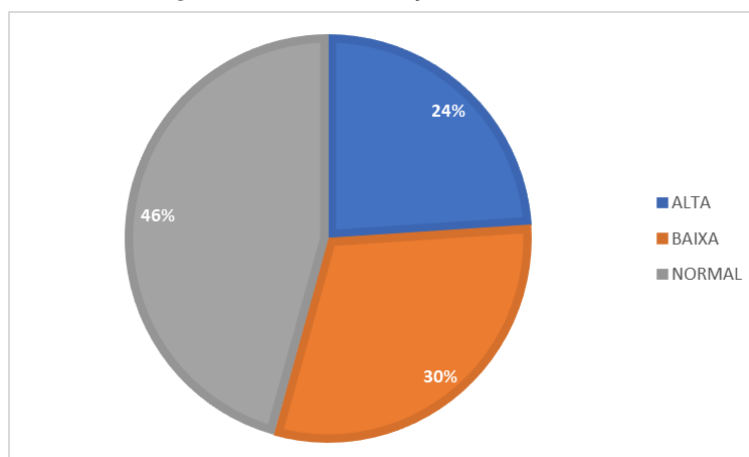
Tabela 2 – Regra de negócio para o atributo V.

Atributo V	Média de compra
Baixo	Menor ou igual a R\$ 600,00
Normal	Entre R\$ 601,00 e R\$ 3600,00
Alta	Maior R\$ 3601,00

Fonte: Elaborado pelo autor.

Os valores monetários para o atributo V foram definidos através de uma análise da melhor forma de distribuição dos valores que separa os clientes com maior heterogeneidade. A Figura 11 mostra a distribuição do atributo V na forma gráfica.

Figura 11 – Distribuição do atributo V



Fonte: Elaborado pelo autor.

Nota-se que 70 % dos clientes tem atributo V acima do mínimo definido pela empresa e para os clientes com atributo V baixo representam 30 %.

Para cálculo do atributo A foi considerado questões relativas ao prazo, tempo de relacionamento, tempos diversos, levando em consideração a média. A Tabela 3 mostra regra utilizada para transformação do atributo quantitativo em qualitativo de acordo com a regra temporal.

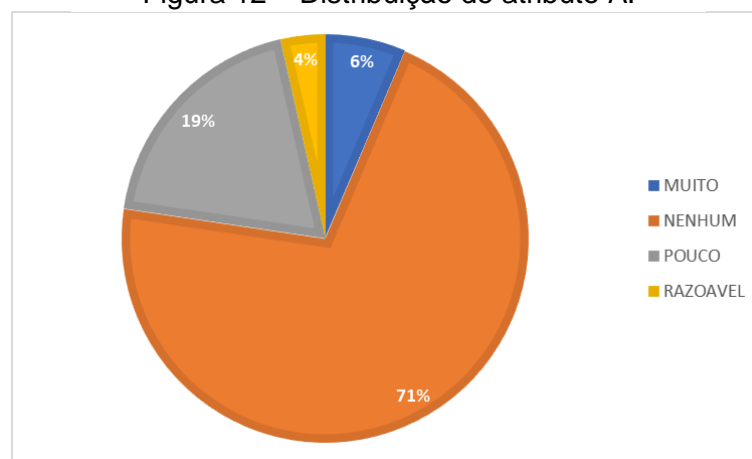
Tabela 3 – Regra de negócio atributo A

A	Unidade tempo
Nenhum	0
Pouco	Entre 1 a 5
Razoável	Entre 6 a 16
Muito	Maior que 16

Fonte: Elaborado pelo

A Figura 12 mostra a distribuição dos valores em forma gráfica.

Figura 12 – Distribuição do atributo A.



Fonte: Elaborado pelo autor.

Percebe-se que a porcentagem de clientes com atributo A ruim é baixa, representando 6 % da base de clientes que compraram no primeiro trimestre.

4.3 Divisão dos clientes em faixas

Para cada conteúdo dos atributos, neste caso do segundo trimestre, que é realmente quando se conhece o cliente, foi definido um valor de acordo com o impacto

para o negócio. As Tabelas 4,5 e 6 mostram os pesos. Os valores dos atributos foram escolhidos através do processo de tentativa e erro, levando em consideração a importância do atributo.

Tabela 4 – Tabela de pesos para o atributo R.

Atributo A	Pesos
Baixa	100
Normal	200
Alta	300

Fonte: Elaborado pelo autor.

Tabela 5 – Tabela de pesos para o atributo V.

Atributo V	Pesos
Baixa	100
Normal	200
Alto	300

Fonte: elaborado pelo autor.

Tabela 6 – Tabela de pesos para o atributo A.

Atributo A	Pesos
Nenhum	350
Pouco	300
Razoável	150
Muito	50

Fonte: Elaborado pelo autor.

Para separar os clientes em classes, foi realizada divisão de acordo com a pontuação definida através da soma dos atributos de cada cliente, e aplicada a regra para associação dos clientes de acordo com sua pontuação.

Após aplicadas as regras sobre o conjunto de dados do primeiro e segundo trimestre do ano de 20XX e, posteriormente, associando a classificação do segundo semestre com os atributos de cada cliente do primeiro trimestre, obteve-se como resultado uma tabela com a identificação dos clientes e seus atributos e classificação conforme Tabela 7.

Tabela 7 – Atributos do primeiro trimestre e classificação do segundo trimestre.

Cliente	R	V	A	Classificação
401	BAIXA	BAIXA	NENHUM	Prata
1601	BAIXA	ALTA	NENHUM	Diamante
2301	BAIXA	NORMAL	NENHUM	Ouro
5701	ALTA	ALTA	NENHUM	Diamante
5801	BAIXA	ALTA	NENHUM	Diamante
6901	BAIXA	BAIXA	POUCO	Prata
7501	BAIXA	BAIXA	NENHUM	Prata
8801	BAIXA	NORMAL	NENHUM	Ouro
9301	BAIXA	NORMAL	NENHUM	Ouro
9501	BAIXA	NORMAL	NENHUM	Ouro
11501	BAIXA	NORMAL	NENHUM	Ouro
12901	BAIXA	BAIXA	NENHUM	Prata
13101	BAIXA	NORMAL	POUCO	Ouro

Fonte: Elaborado pelo autor.

Em seguida, após estrutura de arquivo pronto, foi anexado na ferramenta de mineração de dados para realizar os experimentos.

4.4 Experimentos

Foi escolhido J48, este algoritmo está implementando a técnica de classificação por árvore de decisão. Os resultados obtidos foram aproximados, sendo o escolhido o algoritmo J48. A Figura 13 mostra o arquivo carregado no weka.

Para o processo de classificação, foi realizado a divisão em 80 % para treino e 20 % validação. A Tabela 8, exibe os resultados para esse processo considerando 2 modelos: modelo 1 para os clientes com compra nos dois períodos, e modelo 2 clientes que não necessariamente compraram nos dois períodos.

Figura 13 – Informações básicas do arquivo para classificar.

The screenshot shows the Weka Explorer interface. The 'Selected attribute' section is active, displaying the following table:

No.	Label	Count	Weight
1	DIAMANTE	142	142.0
2	OURO	391	391.0
3	PRATA	251	251.0
4	BRONZE	59	59.0

Below the table is a bar chart with the following data series:

- DIAMANTE: 142 (blue bar)
- OURO: 391 (red bar)
- PRATA: 251 (cyan bar)
- BRONZE: 59 (dark green bar)

Fonte: Elaborado pelo autor.

. A base de dados para o modelo 1 possui N registros, já do modelo 2 possui M, ambos próximos ao milhar, para melhor tratamento dos dados. As Tabelas 8 e 9 mostram a acurácia para ambos os modelos.

Tabela 8 – Acurácia do modelo 1.

Base	Acurácia	Total da amostra- validação
Validação	89,34%	169

Fonte: Elaborado pelo autor.

Tabela 9 – Acurácia do modelo 2.

Base	Acurácia	Total da amostra- validação
Validação	72,22%	216

Fonte: Elaborado pelo autor.

Como o melhor resultado de classificação foi para o modelo 1, deve-se considerar apenas este modelo para os demais exemplos de resultados. A Figura 14 mostra a estrutura da árvore gerada. Os números para os nós-folha representam as quantidades de acerto e quantidades de erro para aquele caminho da árvore. A árvore gerada possui 13 nós folhas.

Figura 14 – Estrutura da árvore de decisão - modelo 1.

```
V = BAIXA: PRATA (256.0/23.0)
V = NORMAL
| A = NENHUM
| | R = BAIXA: OURO (253.0/12.0)
| | R = NORMAL: OURO (22.0)
| | R = ALTA: DIAMANTE (5.0)
| A = POUCO: OURO (76.0/7.0)
| A = RAZOAVEL: OURO (15.0/5.0)
| A = MUITO: BRONZE (14.0/1.0)
V = ALTA
| A = NENHUM: DIAMANTE (134.0/16.0)
| A = POUCO: DIAMANTE (31.0/13.0)
| A = RAZOAVEL: OURO (5.0)
| A = MUITO
| | R = BAIXA: PRATA (16.0/2.0)
| | R = NORMAL: OURO (7.0)
| | R = ALTA: OURO (9.0/1.0)

Number of Leaves :      13

Size of the tree :      18
```

Fonte: Elaborado pelo autor.

A Figura 15 mostram o resultado da matriz de confusão, indicando com mais detalhes da classificação por classes.

Figura 15 - Matriz de confusão – modelo 1.

```
  a   b   c   d  <-- classified
137   5   0   0 |  a = DIAMANTE
 26 362   2   1 |  b = OURO
  3   1 247   0 |  c = PRATA
  0  24  23  12 |  d = BRONZE
```

Fonte: Elaborado pelo autor

Observe que o maior número de erro de classificação é na classe dos clientes bronze, e isso ocorre devido a regra de classificação ser baseada no próximo trimestre, portanto o cliente pode ter bons atributos no primeiro trimestre, mas uma classificação ruim no próximo trimestre. Na Tabela 10, mostra a precisão e sensibilidade do modelo 1.

Tabela 10 – Precisão e sensibilidade para o modelo 1.

Precisão	Sensibilidade	Classe
0,825	0,965	Diamante
0,923	0,926	Ouro
0,908	0,984	Prata
0,923	0,203	Bronze

Fonte: Elaborado pelo autor.

4.5 Análise dos resultados obtidos

Após o processo de classificação realizado pela ferramenta os resultados obtidos foram analisados graficamente e foi possível notar diversas relações entre R, V e A, sendo importante para prever quais clientes podem ser classificados de acordo com seu comportamento no primeiro semestre.

Através da classificação, podem ser definidos faixas de limite de crédito para cada perfil de clientes, facilitando o processo de análise de crédito, uma vez que o processo de avaliação é feito em geral, de forma integralmente manual nas empresas.

Outra análise é verificar se houve queda na venda para algum cliente, ou se mudou de categoria. Isso é importante caso algum cliente tenha reduzido suas compras, sua classificação irá rebaixar. Portanto a equipe comercial deve fazer o levantamento daqueles clientes regrediram na sua classificação, para melhorar o atendimento, produtos ou preço. Diante de todo o estudo realizado ao longo deste trabalho, obteve-se um resultado positivo, na medida que uma tomada de decisão pode ser baseada nos dados que conseguem obter boa acurácia.

5 CONCLUSÃO

O processo de mineração de dados traz um grande poder decisório no processo empresarial, apoiando em diversas áreas. No cenário deste trabalho, teve o objetivo de aplicar um conjunto de regras de negócio internas, aplicadas em dados históricos que permitiram a análise dos dados de compra dos clientes e inadimplência e classificá-los utilizando algoritmo de classificação.

Os processos de classificação foram realizados através do *software* Weka, o qual supriu os requisitos da mineração de dados. O desenvolvimento do trabalho atendeu às expectativas gerando resultados satisfatórios e permitindo o uso em diversas áreas da empresa, como Marketing e Financeiro.

Através dos resultados é possível extrair diversas relações entre os atributos escolhidos. A análise mostrou os clientes que possui atributo V alto tem mais probabilidade de ser atributo A alto, e essas informações podem ser usada na atribuição de crédito e direcionamento de marketing para segmentos de clientes específicos.

Ações de marketing podem ser direcionadas para os clientes que estão com atributos indesejáveis, por exemplo, atributo V baixo, podendo ser oferecido conjunto de promoções, um mix de produtos diferentes para fortalecer a variedade de produtos vendidos no ponto de venda, tornando os atributos mais importantes com valores desejáveis.

Para atribuição de crédito será útil poder montar uma faixa de crédito de acordo com a classificação durante o período, no qual os clientes de determinada categoria possuem um intervalo de limite a para ser atribuído, levando sempre em consideração os fatores externos e internos ao negócio.

5.1 TRABALHOS FUTUROS

Recomenda-se para trabalhos futuros a incorporação de atributos do cliente e o cruzamento de informações locais com informações externas, como Serasa, Sintegra, Receita Federal entre outras instituições que possuem dados aberto da pessoa jurídica. Sugere-se também a aplicação de outros algoritmos de aprendizagem para realizara a previsibilidade de faturamento de clientes no percorrer do trimestre ou ano.

REFERÊNCIAS BIBLIOGRÁFICAS

ANGELONI, Maria; REIS, Eduardo S. **Business Intelligence como Tecnologia de Suporte a Definição de estratégias para melhoria da qualidade do ensino**. In: Encontro da ANPAD, 2006, Salvador. XXX Encontro Nacional de Pós-Graduação em Administração, 2006, 2006. v. 1. <Disponível em: http://www.anpad.org.br/diversos/down_zips/10/enanpad2006-adid-0815.pdf>.

Acesso em 23 mai. 2020.

BRAZDIL, P. **Construção de Modelos de Decisão a partir de Dados**. Disponível por WWW em: <http://www.ncc.up.pt/~pbrazdil/Ensino/ML/DecTrees.html>, 1999.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C. J. (1984). **Classification and Regression Trees**. Wadsworth.

CARDOSO, Letícia. **Análise de clientes de uma distribuidora de produtos farmacêuticos com Mineração de dados baseada em Árvore de Decisão**. 2017. Trabalho de conclusão de curso (Bacharel em Ciência da Computação) - Faculdade de Computação da Universidade Federal de Uberlândia, [S. l.], 2017.

CAMILO, Cássio; SILVA, João. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**, [s. l.], 2009. Disponível em: http://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf. Acesso em: 24 maio 2020.

CARMO, Joyce; SILVA, Iago; SILVA, Raniel; SOUTO, Luís; SILVA, Paloma. Análise de Crédito Utilizando uma Abordagem de Mineração de Dados. **Análise de Crédito Utilizando uma Abordagem de Mineração de Dados**, [s. l.], v. 3, 2018.

CHIU, Susan; TAVELLA, Domingo. **Data Mining and Market Intelligence for Optimal Marketing Returns**. Burlington: Elsevier, 2006. 295 p. v. 1. ISBN 978-0-7506-8234-3.

CORTÊS, Sergio; PORCARO, Rosa; LIFSHITZ, Sergio. **Mineração de Dados - Funcionalidades, Técnicas e Abordagens**, [s. l.], 2002. Disponível em:

ftp://obaluae.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf. Acesso em: 24 maio 2020.

CREPALDI, Paola; AVILA, Renato; OLIVEIRA, Jaqueline; RODRIGUES, Paulo; MARTINS, Renan. **UM ESTUDO SOBRE A ÁRVORE DE DECISÃO E SUA IMPORTÂNCIA NA HABILIDADE DE APRENDIZADO**, [s. l.], 2010. Disponível em: https://www.inesul.edu.br/revista/arquivos/arq-idvol_15_1320100263.pdf. Acesso em: 24 maio 2020.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery**. *AI Magazine*, 1996, p. 37-54. <Disponível em:<https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>>.

FEIJÓ, Bruno Vieira. **A Revolução dos Dados**. *Revista Exame PME – Pequenas e Médias Empresas*, São Paulo, p. 30-43, set. 2013. <Disponível em: <https://exame.com/pme/a-revolucao-dos-dados/9/>> Acesso em 23 mai. 2020>.

FONSECA, S. O; NAMEN, A. A. **Um modelo de classificação para a descoberta de conhecimento relacionado a professores de matemática do ensino básico**. In: **ENCONTRO DE MODELAGEM COMPUTACIONAL, 16. 2013**, Ilhéus, Bahia, Ilhéus. Rio de Janeiro: Rede Sirius, UERJ, 2013, v.1. Disponível em <http://www.btdtd.uerj.br/tde_busca/arquivo.php?codArquivo=6723>. Acesso em 23 mai. 2020.>

GÓES, Anderson; STEINER, Roges. **Proposta de metodologia para a criação de etiqueta de classificação – estudo de caso: desempenho escolar**. Artigo, [s. l.], 2016. DOI <https://doi.org/10.1590/0104-530X810-13>. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-530X2016000100177&lng=pt&tlng=pt. Acesso em: 15 dez. 2020.

HALMENSCHLAGER, Carine. **Um algoritmo para indução de árvores e regras de decisão**. 112 f. Tse (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Sul, 2002.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. John Wiley and Sons, Inc, 2005.

LORENA, A. C.; CARVALHO, A. C. P. L. F. **Uma Introdução às Support Vector Machines. Revista de Informática Teórica e Aplicada**, vol.14, no2, pp 43-67, 2007. Disponível em: . Acesso em: 16 out. 2010.

LIRA, Waleska; CÂNDIDO, Gesinaldo; ARAÚJO, Geraldo; BARROS, Marcelo. **A busca e sua informação nas organizações**, [s. l.], 2008. Disponível em: <https://www.scielo.br/pdf/pci/v13n1/v13n1a11.pdf>. Acesso em: 29 maio 2020.

PRINA, Bruno; TRENTIN, Romario. Geração de Matriz de Confusão a partir de uma classificação digital de imagem do ArcGIS. **Geração de Matriz de Confusão a partir de uma classificação**, João Pessoa-PB, 18 nov. 2020.

MENDITTO, A., PATRIARCA, M. & MAGNUSSON, B. **Understanding the meaning of accuracy, trueness and precision. Accred Qual Assur** **12**, 45–47 (2007). <https://doi.org/10.1007/s00769-006-0191-z>

MONARD, Maria Carolina, BARANAUSKAS, José Augusto. Conceitos Sobre Aprendizado de Máquina. **Sistemas Inteligentes Fundamentos e Aplicações**. 1 ed. Barueri-SP: Manole Ltda, 2003. p. 89--114. ISBN 85-204-168.

SCHRICKEL, W. K. **Análise de Crédito: Concessão e Gerência de Empréstimos**. 4 ed. São Paulo: Editora Atlas S.A. 2000.

SILVA, Leandro; MARQUES, Sarajane; BOSCARIOLI, Clodis. **Introdução a Mineração de Dados com Aplicações em R**. 2016. ed. Rio de Janeiro: Elsevier, 2006. 468 p. v. 1. ISBN 978-85-352-8446-1.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. Data Mining: **Practical Machine Learning Tools and Techniques**. 3. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.