UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

BÁRBARA CAROLINE TURRA KUCHINISKI

APLICAÇÃO DE MÉTODOS DE MINERAÇÃO DE DADOS EM BASES DE DADOS DE CRÉDITO E SEGURO DE CLIENTES

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA

2018

BÁRBARA CAROLINE TURRA KUCHINISKI

APLICAÇÃO DE MÉTODOS DE MINERAÇÃO DE DADOS EM BASES DE DADOS DE CRÉDITO E SEGURO DE CLIENTES

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Engenharia de Produção, do Departamento de Engenharia de Produção, da Universidade Tecnológica Federal do Paraná – Campus Ponta Grossa.

Orientador: Prof. Dr. Antonio Carlos de Francisco

PONTA GROSSA

2018

	Ministério da Educação **UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ** **CÂMPUS PONTA GROSSA** Departamento Acadêmico de Engenharia de Produção	

TERMO DE APROVAÇÃO DE TCC

APLICAÇÃO DE MÉTODOS DE MINERAÇÃO DE DADOS EM BASES DE DADOS DE CRÉDITO E SEGURO DE CLIENTES

por

Bárbara Caroline Turra Kuchiniski

Este Trabalho de Conclusão de Curso (TCC) foi apresentado em 22 de Junho de 2018 como requisito parcial para a obtenção do título de Bacharel em Engenharia de Produção. A candidata foi arguida pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Prof. Dr. Antonio Carlos de Francisco
Prof. Orientador

_____
Prof. Dr. Fabio Neves Puglieri
Membro titular

_____
Prof. Me. Jovani Taveira de Souza
Membro titular

"A Folha de Aprovação assinada encontra-se na Coordenação do Curso".

AGRADECIMENTOS

Primeiramente a Deus que permitiu que tudo isso acontecesse.

Aos meus pais, Maurício e Cleusa, por me fornecerem suporte em todos os momentos de minha vida e por todos os sacrifícios que fizeram em prol de meu benefício. Espero um dia eu poder retribuir uma parte de tudo o que fizeram por mim. E aos meus irmãos, Gabriel e Camila, por sempre estarem ao meu lado me apoiando e me incentivando.

Ao meu orientador, Antonio Carlos de Francisco, pela confiança e paciência com que me acompanhou durante toda a minha graduação e aos desenvolvimentos dos trabalhos.

Ao Jovani Taveira de Souza que me auxiliou pacientemente no desenvolvimento deste trabalho.

E a todos aqueles que de alguma maneira contribuíram para que esse trabalho fosse realizado.

RESUMO

KUCHINISKI, Bárbara Caroline Turra. **Aplicação de Métodos de Mineração de Dados em Bases de Dados de Crédito e Seguro de Clientes**. 2018. 57p. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Produção) - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2018.

Para as empresas consideram importante definir o sistema de categorização dos clientes. Neste trabalho, foram abordados dois tipos de focos de clientes, um deles foi o mercado de seguro automobilístico, este serviço permite um alto grau de interação empresa-cliente, sendo considerado um mercado de alto potencial e em fase de crescimento intenso, porém os clientes podem mudar facilmente de seguradora dependendo de sua satisfação. Outro foco, são os de clientes de créditos, onde os clientes são permitidos a adquirem créditos de empréstimos pelos bancos dependendo do seu perfil, tendo como importância dos créditos como meio de impulsionar as atividades produtivas. Existem um grande leque de dados de todos os tipos de clientes, tendo que cada ramo a necessidade de traçar o perfil dos seus clientes. Para que as empresas possam saber quais as questões que são realmente necessárias para a tomada de decisão estratégicas aplicou-se o estudo da Mineração de Dados. Os métodos empregados foram: Projeção Aleatória e a Análise de Componentes Principais (PCA), ambos utilizando os algoritmos Naive Bayes, J48 e SVM, com o auxílio do *software* WEKA. Como resultado, foram mostradas melhoras significativas nas eficiências dos classificadores envolvendo os métodos empregados. A abordagem de Projeção Aleatória obteve os melhores resultados para as duas bases de dados analisadas. Os algoritmos J48 e SVM apresentaram melhor desempenho comparado com o Naive Bayes dentre as bases. Portanto, a partir dos subconjuntos escolhidos, podem ser submetidos a análises específicas, no intuito de direcionar uma identificação mais precisas.

Palavras-chave: Clientes. Bases de Dados. Mineração de Dados. Projeção Aleatória. Análise de Componentes Principais.

ABSTRACT

KUCHINISKI, Bárbara Caroline Turra Kuchiniski. **Application of Data Mining Methods in Credit Databases and Customer Insurance.** 2018. 57p. Work of Conclusion Course (Graduation in Production Engennier) - Federal Technology University Paraná. Ponta Grossa, 2018.

For companies, it is important to define the customer categorization system. In this work, two types of customer focus were address, one of which was the automobile insurance market, this service allows a high degree of interaction between company and customer, being consider a high potential market and in an intense growth phase, but the clients can easily switch from insurer depending on your satisfaction. Another focus is that of credit customers, where customers are allowed to borrow from banks depending on their profile, with credit as a means of boosting productive activities. There is a wide range of data from all types of customers, having each branch the need to profile their customers. In order for companies to know what issues are really need for strategic decision-making, the study of Data Mining was apply. The methods used were Random Projection and Principal Component Analysis (PCA), both using the Naive Bayes, J48 and SVM algorithms, with the help of WEKA software. As a result, significant improvements have been shown in the efficiencies of the classifiers involving the methods employed. The Random Projection approach obtained the best results for the two databases analyzed. The J48 and SVM algorithms presented better performance compared to Naive Bayes among the bases. Therefore, from the chosen subsets, they can be submitted to specific analyzes, in order to direct a more precise identification.

Keywords: Customer. Data Base. Data Mining. Random Projection. Principal Component Analysis.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

CRM	Customer Relationship Management (Gerenciamento de Relacionamento com o Cliente)
DM	Data Mining (Mineração de Dados)
KDD	Knowledge Discovery in Databases (Descoberta de Conhecimento em Bancos de Dados)
PCA	Principal Component Analysis (Análise de Componentes Principais)

SUMÁRIO

1 INTRODUÇÃO

Para as organizações obterem o sucesso que desejam é necessário desenvolver a capacidade de criar novos conhecimentos e estratégias, ainda mais com o avanço as transferências de dados e disponibilidade de informações que tornaram os custos de mudanças extremamente baixos (EICHORN, 2004). Segundo Menguc et al. (2007), a importância no desenvolvimento do conhecimento dentro das empresas leva à criação de novos produtos e serviços para competir frente aos concorrentes e atingir vantagens competitivas.

A qualidade do serviço em relação ao cliente tornou-se uma estratégia predominante de diferenciação e vantagem competitiva. Além disso, as empresas estão terceirizando cada vez mais as funções externas, associadas ao gerenciamento de relacionamento com o cliente, como *call centers* e *telemarketing*, e funções internas, como contabilidade e recursos humanos. Essa divisão apresenta desafios ainda maiores para garantir que os clientes tenham uma experiência positiva, consistente e que os sistemas internos estejam conectados e integrados (ARBACHE et al., 2011).

A qualidade do serviço está vinculada na percepção em que o cliente possui, quando relaciona uma marca, empresa, produto ou serviço, e as percepções associadas aos mesmos. Neste sentido, a implementação de uma pesquisa dirigida para uma análise do cliente permite a empresa, conhecer quem são os seus clientes e o que eles esperam da empresa (SIQUEIRA et al., 2014).

Estudos mostraram que 96% dos clientes quando estão insatisfeitos não fazem reclamações, simplesmente deixam de comprar, por isso a empresa deve atentar-se na diminuição da quantidade de fluxo dos seus clientes (CAMURÇA; MAGALHÃES, 2017).

Devido à grande preocupação com relação a esses fatores e a quantidade de informações existentes, as organizações estão optando pela utilização de ferramentas auxiliares, a fim de amparar na busca por estratégias adequadas e eficientes (SIQUEIRA et al., 2014).

Uma das abordagens utilizadas neste contexto, envolve a utilização de métodos de mineração de dados (*data mining*) que avaliam o valor dos clientes, a diferença entre as percepções dos clientes quanto aos benefícios e ao custo para adquirir o produto ou serviço. Os métodos mencionados analisam padrões com o

objetivo de extrair conhecimentos para otimizar os relacionamentos com os clientes (CAMURÇA; MAGALHÃES, 2017).

Observa-se, também, que as empresas veem os clientes como padrão de comparação de avaliação de qualidade de serviço. Entretanto, há muitas informações que as pessoas que são responsáveis pela tomada de decisão não utilizam, devido à falta de conhecimento sobre os dados (ZEITHAML et al., 2014).

A descoberta do conhecimento, escondida nas grandes bases de dados de empresas de diversos setores de maneira automática ou semiautomática, é o objetivo da mineração de dados, além de ser uma técnica que permite maior agilidade no processo de tomada de decisão (PASTA, 2011). A coleta e a armazenagem dos dados somente para si, não traz nenhuma contribuição para a melhoria estratégica da empresa.

O *Data Mining* tem um papel importante para que as grandes quantidades de dados sejam exploradas, encontrando padrões, regras ou dados ocultos nas bases de dados (KAMBER et al., 2012). Para Han e Kamber (2006), os métodos fornecem diversas metodologias para resolução de problemas, análise, planejamento, diagnóstico, aprendizagem e inovação. Além de ser um campo interdisciplinar, cujo processo envolve banco de dados, visualização de dados, aprendizado de máquinas, algoritmos matemáticos e técnicas estatísticas.

Há um grande interesse por partes das empresas e organizações em relação à base de dados de seguro devido principalmente, a margem de lucro fornecida. No ano de 2015, o mercado de seguros obteve uma margem bruta de lucro de R$11.800.000,00, o que corresponde 16,2% da receita do seguro de automóveis (TSS, 2018).

Também temos outro assunto de grande interesse por parte da comunidade acadêmico-científica para realização de estudos para que haja maior conhecimento sobre o assunto, a área de crédito para as pessoas físicas, que ocorreu um aumento nos financiamentos para veículos e cartão de crédito à vista. O saldo das operações de crédito do sistema financeiro atingiu R$3,1 trilhões em abril de 2018, a relação crédito/PIB atingiu 46,5% neste mesmo mês (BCB, 2018).

As escolhas e aplicações dos métodos devem ser criteriosamente estudados, pois é necessário proteger e assegurar a veracidade dos dados. Em alguns casos, entretanto, se essas técnicas forem utilizadas inadequadamente podem gerar

resultados inconsistentes. Assim, é preciso analisar se os métodos escolhidos são eficazes para o contexto em que são utilizados (LIERENA, 2013).

Portanto, diante desta temática, o estudo visa utilizar métodos de mineração de dados em bases de dados de clientes de crédito e seguro.

1.1 PROBLEMA

Esta pesquisa pretende responder a seguinte questão: Entre os métodos de Análise de Componentes Principais e de Projeção Aleatória qual possui maior taxa de acerto nos dados de créditos e seguros de clientes quando aplicados no software?

1.2 JUSTIFICATIVA

A essência mais importante de uma organização são os clientes. Não pode haver perspectivas de negócios sem clientes satisfeitos que permaneçam fiéis e desenvolvam seu relacionamento com a organização. Utilizar-se de boas estratégias é fundamental, pois além de melhorarem a fidelização de clientes, fornecem a lucratividade para essas organizações (ZIAFAT; SHAKERI, 2014).

Estudos conceituais e práticos têm destacado a confiança como elemento fundamental no desenvolvimento de fortes e longos relacionamentos entre clientes e organizações (SANTOS; FERNANDES, 2008).

Há cinco elementos, sugeridos por Ribeiro, Grisi e Saliby (1999), para o desenvolvimento de um relacionamento produtivo entre o cliente e a organização, primeiro desenvolver um serviço que será construído ao redor desse relacionamento, depois customizar o relacionamento para o cliente individual, posteriormente aumentar o serviço central com benefícios extras, quarto especializar o serviço de forma a encorajar a lealdade do cliente e último elemento, praticar o marketing com os empregados de forma a incentivá-los a fazer o melhor para os clientes.

Ou seja, fazendo com que as organizações que estão focadas nos clientes estabeleçam relacionamentos baseados no aprendizado de suas necessidades e

desejos, oferecendo produtos adequados e assim mantendo relações de longo prazo (CAMURÇA; MAGALHÃES, 2017).

Um dos focos dos clientes é o mercado do seguro, por causa de algumas considerações, o seguro é caracterizado como um serviço de relacionamento prolongado, que permite um alto grau de interação empresa-cliente; as barreiras de mudanças de fornecedores do serviço são baixas, uma vez que um cliente pode mudar facilmente de seguradora; o mercado segurador é de alto potencial e em fase de crescimento intenso e esse setor está enfrentando grandes desafios a fim de buscar alternativas para suas práticas tradicionais de atuação no mercado (RIBEIRO; GRISI; SALIBY, 1999).

Outro foco que não se pode ignorar, é o crédito liberado pelos bancos aos clientes, é de grande importância a liberação do crédito como meio impulsionador da atividade produtiva. Há países que disponibilizam para seus agentes econômicos créditos superiores ao volume das unidades de bens e serviços produzidos, num ciclo virtuoso entre a produção e o consumo que, sem maiores esforços do legislativo, estimula a geração de emprego e renda (SOARES; SOBRINHO, 2008).

Há um importante empecilho ao acesso das comunidades de baixa renda aos mecanismos de financiamento tradicionais, mesmo aqueles cobertos por linhas especiais de incentivos governamentais aos micro e pequenos negócios, é a falta de instrumentos eficientes de garantia. Nessa linha, Soares e Sobrinho (2008) apontam como fator determinante para essa escassez de crédito, a falta de estrutura legal e de justiça que permita a essas comunidades securitizarem seus ativos.

Portanto, em relação aos seguros, há reflexões acadêmicas nesse mercado a fim de apoiar a indústria na busca de alternativas mais produtivas de comercialização e de relacionamento com os clientes (RIBEIRO; GRISI; SALIBY, 1999). E também, no ponto de créditos concedidos aos clientes de bancos, visto ser de interesse público, ressalta-se que uma das preocupações fundamentais do governo é ampliar o acesso a serviços financeiros para grande parte da população (SOARES; SOBRINHO, 2008).

Com a modernização das tecnologias e as suas evoluções cada vez mais avançadas, visto que é de fundamental importância que as organizações do conhecimento disponham de técnicas e ferramentas para análise de dados e de informações, criadas para suportar as decisões estratégicas, táticas e operacionais. Nesse aspecto, a mineração de dados contribui com as descobertas de

conhecimentos, pois através de técnicas e ferramentas, ajudam a buscar correlações importantes entre os dados (FAYYAD et al., 1996).

As técnicas de mineração de dados não podem substituir o papel significativo dos especialistas em domínio e seu conhecimento comercial. Porém, pode-se obter resultados úteis combinando com essas técnicas. Por exemplo, combinar experiência pessoal no campo ou informações de negócios com um modelo de mineração de dados para gerar resultados mais bem-sucedidos. Além disso, esses resultados devem ser sempre avaliados por especialistas. Assim, os conhecimentos do negócio podem ajudar e enriquecer os resultados da mineração de dados (ZIAFAT; SHAKERI, 2014).

A aplicação desse estudo pretende contribuir para o entendimento dos gestores, pesquisadores e profissionais da área, sobre informações preponderantes, extraídas dos métodos empregados, que possam auxiliar no processo de tomada de decisão de empresas e organizações.

1.3 OBJETIVO GERAL

Avaliar a taxa de acerto na aplicação dos métodos de Projeção Aleatória e Análise de Componentes Principais em dados de Crédito e Seguro de clientes.

1.4 OBJETIVOS ESPECÍFICOS

Constituem como objetivos específicos para este Trabalho de Conclusão de Curso:

– Selecionar as bases de dados utilizadas para a aplicação do estudo;

– Aplicar o método de Projeção Aleatória e o método de Análise de Componentes Principais nas bases de dados de Crédito e Seguro;

– Realizar a classificação dos algoritmos nas bases de dados;

– Avaliar as informações que foram coletadas após a aplicação dos métodos adotados.

Após os objetivos definidos, serão apresentadas as delimitações do tema de pesquisa.

1.5 DELIMITAÇÃO DO TEMA

Este estudo delimita-se na aplicação dos métodos de Projeção Aleatória e de Análise de Componentes Principais em duas bases de dados, de Crédito e Seguro, avaliando as informações e conhecimentos gerados após a aplicação dos métodos.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta a fundamentação teórica dos principais conceitos sobre os clientes, pois é o tema principal que direciona o desenvolvimento da pesquisa, a Descoberta do Conhecimento em Banco de Dados, por meio da mineração de dados, com suas características, definições e utilidades, juntamente com o método de Projeção Aleatória e o método de Análise dos Componentes Principais que serão utilizados no presente estudo.

2.1 PERFIL DE CLIENTE

Para serem considerados clientes, é preciso fazer com que estes recebam ou envolver produtos, serviços ou processos, podendo caracteriza-los como clientes internos ou externos, onde os clientes internos são as pessoas que fazem parte da empresa e os clientes externos são alcançados pelos produtos porem não faz parte da empresa que o produz (CAMURÇA; MAGALHÃES, 2017).

Os clientes são como padrões de comparações das avaliações de qualidade dos serviços e dos produtos, pois são os clientes que decidem quais atendem as suas necessidades e suas expectativas (ZEITHAML et al., 2014).

O principal objetivo de cada indústria é entender cada cliente individualmente e usá-lo para tornar mais fácil para o cliente fazer negócios com eles, e não com os concorrentes (ZIAFAT; SHAKERI, 2014).

Os gestores devem aproveitar as oportunidades que possuem uma alta capacidade de retorno do capital dentro de um período de tempo. Tendo como intuito resgatar os dados durante as operações diárias e armazenados nos depósitos para fins do gerenciamento de relacionamento com o cliente (*Customer Relationship Management* – CRM) precisando ser transformados em conhecimentos úteis (DURSUN; CABER, 2016).

O CRM se concentra naturalmente em clientes estabelecidos. É a estratégia para construir, gerenciar e fortalecer relacionamentos leais e duradouros com os clientes (ZIAFAT; SHAKERI, 2014).

Para a implementação do sistema de CRM não se trata somente em desenvolver e utilizar um sistema, mas também de transformar a cultura da empresa,

tendo que estar totalmente pronta a atender as necessidades de seus clientes (SIQUEIRA et al., 2014).

Para que exista uma implementação do CRM bem sucedido é necessário considerar os fatores organizacionais, tecnologias, orientações aos clientes e experiências em CRM (DURSUN; CABER, 2016). Formando-se quatro dimensões, o gerenciamento de relacionamento com o cliente, primeiramente identifica, atraem, retém e por fim o ocorre o desenvolvimento da implementação (NGAI; XIU; CHAU, 2009).

A identificação dos clientes mais lucrativos e a segmentação desses clientes dependem de variáveis armazenadas nos conjuntos de dados que são essenciais, estudos anteriores no setor de serviços, em geral, mostram que apenas 15% dos clientes geram 45% de receita e 70% de lucro (DURSUN; CABER, 2016).

A lealdade e lucratividade dos clientes estão correlacionadas. Portanto, uma das principais premissas do CRM é satisfazer e criar relacionamentos de longo prazo com clientes lucrativos aumenta o sucesso dos negócios das empresas (RIBEIRO; GRISI; SALIBY, 1999).

O conhecimento obtido a partir dos dados pode minimizar os riscos gerenciais e aumentar a eficácia das estratégias de CRM. Assim, os métodos de mineração de dados mantêm a identificação das tendências significativas ocultas e os relacionamentos dentro dos dados (DURSUN; CABER, 2016).

Para a empresa é preciso definir o sistema para a categorização dos clientes, esses níveis podem ser identificados, incentivados e atendidos, possivelmente os diferentes níveis geram diferentes lucros para as empresas. As companhias aumentam as oportunidades para gerar lucros quando aumentam a proporção de compras de clientes (ZEITHAML et al., 2014).

Tendo que os acadêmicos geralmente adaptarem as abordagens quantitativas para a criação dos perfis e segmentar dos clientes, como mineração de dados, análise fatorial, análise conjunta, regressão linear ou análise de regressão logística, análise discriminante (DURSUN; CABER, 2016).

Detalha-se a seguir a mineração de dados e a Descoberta de Conhecimento em base de dados.

2.2 MINERAÇÃO DE DADOS E DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Em anos recentes, grandes quantidades de dados se tornaram cada vez mais disponíveis em volumes significativos (TREVISAN, 2017). Torna-se difícil a análise manual para realizar uma tomada de decisão estratégica, sendo necessário auxílio para realizar análises e estudos, como a mineração de dados, onde ocorre a obtenção de informações importantes a partir dos dados disponíveis (NAIK; SAMANT, 2016).

Muitas empresas costumam usar técnicas de mineração de dados para CRM, o que ajuda a fornecer um serviço personalizado, atendendo às necessidades individuais dos clientes, em vez do marketing em massa. Existem vários pacotes de *softwares de* CRM usados para rastrear as interações com os clientes, registrando o histórico de contato e armazenando informações valiosas do cliente. No entanto, esses pacotes são ferramentas nas quais devem ser utilizados para apoiar a estratégia de gerenciar efetivamente os clientes (ZIAFAT; SHAKERI, 2014).

As organizações precisam obter informações sobre os clientes, suas necessidades e desejos por meio da análise de dados para a obtenção do sucesso com o CRM. Em outras palavras, as organizações analisam as informações do cliente para melhor atender os objetivos do CRM e entregar a mensagem certa ao cliente certo. Envolvendo o uso de métodos de mineração de dados para avaliar o valor dos clientes, entender e prever seu comportamento. Eles analisam padrões para extrair conhecimento para otimizar os relacionamentos com os clientes. A mineração de dados ainda é uma questão estrangeira para muitos profissionais que confiam apenas em suas experiências (ZIAFAT; SHAKERI, 2014).

Segundo Costa (2012) existem dois tipos de metas para definir as funções dos objetivos na utilização do sistema para o processo de descoberta, a meta do tipo verificação, que o sistema se limita ao verificar as hipóteses definidas pelo usuário. E a meta da descoberta, onde os sistemas encontram novos padrões de forma autônoma das tarefas de mineração de dados.

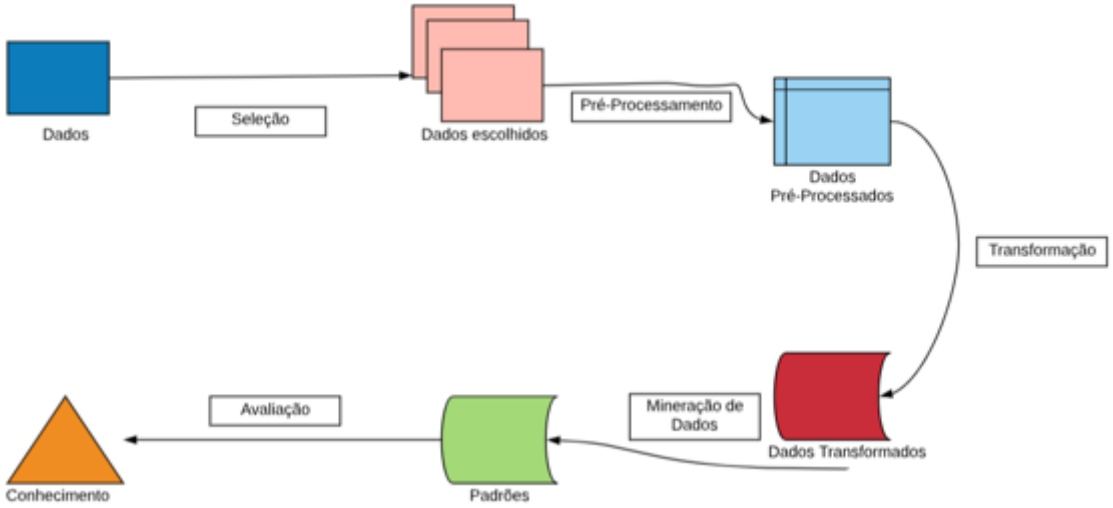
Em relação a extração de padrões para geração de conhecimento, se destaca o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD). Este continua evoluindo, a partir da pesquisa em áreas como banco de dados, aprendizado de máquinas, reconhecimento de padrões,

estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas (TREVISAN, 2017).

O termo Descoberta de Conhecimento em Bases de Dados busca otimizar o processo para que torne os dados de baixo nível, em conhecimento de alto nível e úteis. Tornou-se estrategicamente importante pois possibilita a produção de conhecimento a partir das grandes bases de dados, sendo útil para as grandes organizações de empresas compostas por múltiplas sub-organizações (TREVISAN, 2017).

A Figura 1 ilustra o processo da Descoberta de Conhecimento em Base de Dados:

Figura 1 - Processo da Descoberta de Conhecimento em Bases de Dados (KDD)



Fonte: Adaptado de Fayyad et al. (1996)

Para iniciar, segundo a Figura 1, o processo KDD é feita a escolha das fontes de dados que serão utilizados e os objetivos para a redução de dimensionamento. A partir do problema é possível aprender e desenvolver o conhecimento por meio das ferramentas capazes de extrair informações úteis em uma determinada base de dados (YAMAGUCHI et al., 2010).

Posteriormente, o pré-processamento é onde ocorre a preparação de dados, identificando o conjunto de dados, que fará com que ocorra a possível realização e a aplicação das técnicas para a extração do conhecimento. É nesta fase que ocorre a busca por padronização ou modelos para o tratamento do conhecimento obtido (SOUZA, 2017).

A próxima fase é a mineração de dados (*Data Mining* – DM) na qual extrai padrões a partir dos dados observados. Nesta fase, são identificados os métodos e os algoritmos que irão realizar a busca pelo conhecimento que é considerado útil (SOUZA, 2017).

A mineração de dados pode ser considerada como a principal etapa de um processo de KDD, cujo o papel é incluir as tarefas de seleção, preparação e exploração das informações, e a análise e interpretação dos resultados, assimilando o conhecimento extraído do processo. Os padrões citados devem ser novos, compreensíveis e úteis, devendo trazer algum novo benefício que possa ser compreendido (COSTA, 2012).

A mineração de dados se justifica com estudos que, por meio da aplicação de técnicas, influenciam no apoio ao planejamento. Buscando o uso estratégico da informação, possibilitando a extração de informações implícitas existentes nos bancos de dados, contribuindo com esse processo identificar e classificar novos padrões (PASTA, 2011).

As etapas do processo de mineração de dados são: seleção, pré-processamento, mineração de dados e pós-processamento, que quando aplicados concomitantemente, permitem a descoberta do conhecimento (BORGES; NIEVOLA, 2012).

Os principais objetivos da mineração de dados são descobrir relacionamentos entre dados e fornecer uma contribuição para realizar uma previsão de tendências futuras baseadas no passado (DIAS, 2002).

Um dos problemas na mineração de dados é a classificação que envolve encontrar os parâmetros para predefinição das classes dos dados (NAIK; SAMANT, 2016).

Os resultados obtidos com a mineração de dados podem ser empregados no gerenciamento de informação, processamento de pedidos de informação, tomada de decisão, controle de processo e outras aplicações. Os dados contidos nas bases de dados são usados para aprender um determinado conceito alvo ou padrão (PASTA, 2011).

Algumas das principais tarefas estão descritas no Quadro 1:

Quadro 1 - Tarefas realizadas por técnicas de mineração de dados

Tarefa	Descrição	Exemplo
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes, tendo como objetivo relacionar o atributo meta (cujo valor será previsto) e um conjunto de atributos de previsão.	Classificar pedidos de crédito Esclarecer pedidos de seguros fraudulentos. Identificar a melhor forma de tratamento de um paciente.
Estimativa (ou Regressão)	Usada para definir um valor para alguma variável contínua desconhecida.	Estimar o número de filhos ou a renda total de uma família Estimar o valor em tempo de vida de um cliente Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos Médicos Prever a demanda de um consumidor para um novo produto
Associação	Usada para determinar quais itens tendem a ser adquiridos juntos em uma mesma transação.	Determinar que produtos costumam ser colocados juntos em um carrinho de supermercado
Segmentação (ou *Clustering*)	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos.	Agrupar clientes por região do país Agrupar clientes com comportamento de compra similar Agrupar seções de usuários Web para prever comportamento futuro de usuário
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados.	Tabular o significado e desvios padrão para todos os itens de dados Derivar regras de síntese.

Fonte: Dias (2002).

No Quadro 1, encontra-se as cinco principais tarefas realizadas por técnicas de mineração com suas descrições e exemplos. Assim podendo ajudar a saber quais as tarefas que podem ser utilizadas conforme cada tipo de base de dados.

E o Quadro 2 abaixo, mostra as técnicas de mineração de dados com as descrições as tarefas correlacionadas a cada técnica.

Quadro 2 - Técnicas de mineração de dados

Técnica	Descrição	Tarefas
Descoberta de Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjunto de dados.	Associação
Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos.	Classificação Regressão
Raciocínio Baseado em Casos ou MBR	Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança.	Classificação Segmentação
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores tem mais chance de ter "descendentes".	Classificação Segmentação
Redes Neurais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neurais e dos pesos dessas conexões.	Classificação Segmentação

Fonte: Adaptado de Dias (2002).

O Quadro 2, auxiliam com sugestões das técnicas para resolver os problemas de mineração de dados, dependendo da área de interesse para a utilização da mineração de dados. Cada técnica oferece vantagens e desvantagens, com o estudo das técnicas escolhidas, facilitando a escolha de uma delas de acordo com o problema apresentado (DIAS, 2002).

Por fim, a última etapa da mineração de dados, o pós-processamento realiza a filtragem, a estruturação e a ordenação os resultados, avaliando a qualidade e utilidade adquirido dos resultados obtidos na mineração de dados, determinando a viabilidade de sua utilização no apoio a algum processo de decisão para então serem apresentados ao usuário (COSTA, 2012).

A escolha das técnicas de mineração de dados a ser aplicada é uma tarefa complexa, pois a escolha das técnicas dependerá da tarefa específica a ser executada e dos dados disponíveis para análise. Pode-se dividir a seleção das técnicas de mineração de dados em dois passos, o primeiro, é traduzir o problema de negócio a ser resolvido em séries de tarefas de mineração de dados, e o segundo, é

compreender a natureza dos dados disponíveis em termos de conteúdo e tipos de campos de dados e estrutura das relações entre os registros (DIAS, 2002).

A primeira tarefa é a classificação, com um conjunto de dados de treinamento, posteriormente, selecionar a técnica de mineração de dados que minimiza o número e dificuldades de transformação de dados para obter bons resultados e atingir a meta. Existindo uma lista de características de dados baseada no que ajudará na escolha de uma abordagem de mineração de dados, como citado no Quadro 3 (DIAS, 2002).

Quadro 3 - Características de dados

Característica	Descrição	Técnicas de Mineração de Dados
Variáveis de categorias	São campos que apresentam valores de um conjunto de possibilidades limitado e predeterminado	Descoberta de regras de associação Árvores de Decisão
Variáveis numéricas	São aquelas que podem ser somadas e ordenadas	Raciocínio baseado em casos (MBR) Árvores de Decisão
Muitos campos por registro	Este pode ser um fator de decisão da técnica correta para uma aplicação específica, uma vez que os métodos de mineração de dados variam na capacidade de processar grandes números de campos de entrada	Árvores de Decisão
Variáveis dependentes múltiplas	Caso em que é desejado prever várias variáveis diferentes baseadas nos mesmos dados de entrada	Redes neurais
Registro de comprimento variável	Apresentam dificuldades na maioria das técnicas de mineração de dados, mas existem situações em que a transformação para registros de comprimento fixo não é desejada	Descoberta de regras de associação
Dados ordenados cronologicamente	Apresentam dificuldades para todas as técnicas e, geralmente, requerem aumento dos dados de teste com marcas ou avisos, variáveis de diferença etc.	Rede neural intervalar (time-delay) Descoberta de regras de associação
Texto sem formatação	A maioria das técnicas de mineração de dados é incapaz de manipular texto sem formação	Raciocínio baseado em casos (MBR)

Fonte: Dias (2002).

Com as características de dados, mostrado no Quadro 3, pode-se encontrar a técnica mais apropriada, cada base de dados possui características em suas variáveis.

Os padrões descobertos são avaliados para verificar a satisfação do critério necessário para constituir um elemento importante para o apoio à tomada de decisão. Sendo avaliadas e facilmente interpretadas pelos usuários. Os padrões descobertos são quando os dados redundantes e irrelevantes são removidos e também após a correção dos erros que não foram vistos anteriormente (SOUZA, 2017).

A extração de dados pode ser categorizada de acordo com as técnicas de mineração de dados subordinadas. E também, de acordo com a abordagem de mineração de dados subordinada, como a extração de dados baseada em generalização, baseada em padrões, baseada em teorias estatísticas ou matemáticas, abordagens integradas, etc. A descoberta de regras de associação parece ser uma das técnicas de mineração de dados mais utilizadas (DIAS, 2002).

2.2.1 Algoritmos de Classificação

A mineração de dados abrange alguns algoritmos utilizando tarefas para a classificação, porém cada algoritmo possui um objetivo específico (WU et al., 2008). Foram selecionados três algoritmos considerados como melhores classificadores.

Sendo apresentados no Quadro 4 abaixo, os algoritmos empregados no estudo e seus principais conceitos.

Quadro 4 - Algoritmos utilizados na respectiva pesquisa

Algoritmo	Conceito
Naive Bayes (JOHN, 1995)	Os algoritmos Naive Bayes são classificadores estatísticos baseados no Teorema de Bayes, que predizem a probabilidade de um determinado dado pertencer a uma classe em particular. De acordo com Mitchell (2010), o algoritmo Naive Bayes é muito utilizado, tanto para variáveis discretas ou contínuas, pois é de fácil aplicação em um conjunto de amostras. As probabilidades são estimadas de acordo com a frequência de cada valor para os registros de treino. Assim, dada uma nova instância, o classificador faz a estimativa de probabilidade de o registro feito pertencer a uma nova classe específica, considerando que os atributos são condicionalmente independentes (BERTON, 2011).
J48 (QUINLAN, 1993)	O algoritmo J48 permite a criação de modelos de decisão em árvore. O modelo de árvore de decisão é feito a partir da análise dos dados de treino e pelo modelo utilizado para classificar dados ainda não classificados. O algoritmo gera árvores de decisão, a qual cada nó da árvore avalia individualmente a existência ou significância de cada atributo de maneira individual (FRUTUOSO, 2014). As árvores são geradas através da escolha do atributo mais adequado para cada situação e são construídas do topo para a base. Para Tavares, Bozza e Kono (2007), o algoritmo constrói uma árvore de decisão a partir do atributo mais significativo, por meio da abordagem *top-down*. Neste caso, o atributo mais global é escolhido para ser a raiz da árvore, comparando-o com todos os atributos do conjunto. Com isso, para prosseguimento da construção, é considerado o segundo atributo como sendo o próximo nó da árvore, e assim até que se gere o nó folha, que representa o atributo alvo da instância.
SVM (HASTIE, 1998)	É baseado em modelos lineares, abordando aspectos referentes ao aprendizado para problemas de reconhecimento de padrão. Tem como objetivo a determinação de limites de decisão que produzem uma separação ótima entre classes, por meio da minimização de erros, além de realizar a separação entre duas classes distintas, por meio de um hiperplano de separação (VAPNIK, 1995). O algoritmo SVM mapeia cada dado analisado, utilizando um mapeamento fixo, usando os dados de treino, construindo dessa forma um hiperplano com margem de separação máxima, utilizado para classificar exemplos desconhecidos. O algoritmo trabalha com dados linearmente separáveis, no entanto, existe a possibilidade de adaptação para conjuntos não lineares através das funções kernel não

	lineares (ALVES; FRAGAL, 2011). Mediante essa função, é possível trabalhar com problemas não separáveis linearmente.

Fonte: Adaptado de Souza (2017).

As técnicas de mineração de dados não podem substituir o papel significativo dos especialistas em domínio e seu conhecimento comercial. Podemos obter resultados úteis combinando técnicas de mineração de dados e especialização em negócios. Esses resultados devem ser sempre avaliados por especialistas em negócios. Assim, o conhecimento do negócio pode ajudar e enriquecer os resultados da mineração de dados (ZIAFAT; SHAKERI, 2014).

Por outro lado, as técnicas de mineração de dados podem descobrir padrões que até mesmo as pessoas de negócios mais experientes podem não ter percebido. Como resultado, a combinação do conhecimento do domínio de negócios com o poder das técnicas de mineração de dados pode ajudar as organizações a obter uma vantagem competitiva em seus esforços para otimizar o gerenciamento de clientes (ZIAFAT; SHAKERI, 2014). A seguir, serão apresentadas as principais ferramentas para mineração de dados.

2.2.2 Ferramentas para a Mineração de Dados

Como os dados que poderão ser analisados são de grande quantidade, o tempo de execução pode tornar-se demorado. Havendo uma necessidade de ferramentas, como os *softwares*, que auxiliam na transformação desses dados em informações úteis e com maior garantia de qualidade nelas (NAIK; SAMANT, 2016).

Segundo Cruz (2007) identificou que há uma grande disponibilidade de ferramentas de mineração de dados, no Quadro 5 mostra algumas das ferramentas:

Quadro 5 - Principais ferramentas de mineração de dados

Ferramenta	Licença	Uso
Alyuda Neuro Inteligence	Comercial	Comercial
BrainMaker	Comercial	Acadêmica/Comercial
BSVM	*Freeware e shareware*	Acadêmica
Clementine	Comercial	Comercial
DTREG	Comercial	Acadêmica/Comercial
EQUBITS Foresight ™	Comercial	Acadêmica/Comercial
EWA Systems	Comercial	Acadêmica/Comercial
GhostMiner	Comercial	Acadêmica/Comercial
Gist	*Freeware e shareware*	Acadêmica
Gornik	Comercial	Comercial
Insightful Miner	Comercial	Acadêmica/Comercial
Kernel Machines	*Freeware e shareware*	Acadêmica
Knowledge Miner	Comercial	Acadêmica/Comercial
KXEN	Comercial	Comercial
LIBSVM	*Freeware e shareware*	Acadêmica
MATLAB NN Toolbox	Comercial	Acadêmica
MCubiX from Diagnos	Comercial	Comercial
MemBrain	*Freeware e shareware*	Acadêmica
NeuralWorks Predict	Comercial	Comercial
NeuroSolutions	Comercial	Acadêmica/Comercial
NeuroXL	Comercial	Comercial
IPNNL Software	*Freeware e shareware*	Acadêmica
Oracle Data Mining	Comercial	Comercial
Orange	*Freeware e shareware*	Acadêmica
PcSVM	Pública	Acadêmica
R	Pública	Acadêmica
SAS Enterprise Miner	Comercial	Acadêmica/Comercial
StarProbe	Comercial	Acadêmica/Comercial
STATISTICA NN	Comercial	Acadêmica
SvmFu 3	Pública	Acadêmica
SVM-light	*Freeware e shareware*	Acadêmica
TANAGRA	*Freeware e shareware*	Acadêmica
HhinkAnalitics	Comercial	Comercial

Tiberius	Comercial	Acadêmica/Comercial
Weka	Pública	Acadêmica
XLMiner	Comercial	Acadêmica/Comercial

Fonte: Adaptado de Cruz (2007, p.45).

Estas ferramentas, mostradas no Quadro 5, fornecem um conjunto de métodos e algoritmos que auxiliam na análise dos dados. Auxiliando na análise de grupo, visualização de dados, análise de regressão, árvores de decisão, análise preditiva, mineração de texto, etc. (NAIK; SAMANT, 2016).

A ferramenta de mineração de dados que foi escolhida para ser utilizado neste estudo foi o *Waikato Environment for Knowledge Analysis* (WEKA) versão 3.8.1 (WEKA, 2017), sendo de licença pública e de uso acadêmico. No próximo item será explanado sobre o software WEKA.

2.2.2.1 Software Waikato Environment for Knowledge Analysis (WEKA)

É um conjunto de ferramentas amplamente utilizado para conhecimento de máquinas e a mineração de dados, originalmente desenvolvido na Universidade de Waikato, na Nova Zelândia (NAIK; SAMANT, 2016).

O *software Waikato Environment for Knowledge Analysis* (WEKA) é gratuito, contendo uma vasta coleção de informações de mineração de dados e algoritmos escritos em Java. O WEKA contém ferramentas para regressão, classificação, agrupamento, regras de associação, visualização e pré-processamento de dados. Tornando-se muito popular com os pesquisadores acadêmicos e industriais (NAIK; SAMANT, 2016).

Para evitar vieses e ajustes excessivos, utiliza-se a configuração padrão de parâmetros do WEKA. O s*oftware* WEKA pode ser acessado pelo site http://www.cs.waikato.ac.nz/ml/weka. Considerado umas das ferramentas mais completas na mineração de dados. Contendo uma plataforma pública de bases de dados, que reúnem muitos algoritmos de aprendizado para extração de dados, abrangendo o pré-tratamento de dados, como classificação, regressão, classe de cluster, associação e ainda também inclui regra de mineração e visualização em nova interface (ZHONG, 2011).

O WEKA contém ferramentas para ambas as tarefas da mineração de dados, podendo declarar mais antiga e bem-sucedida biblioteca de dados de código aberto neste âmbito (SOUZA, 2017).

O WEKA usa uma série de interface gráfica unificada com tecnologia de aprendizagem de *software* padrão, pode unificar muitos métodos de pré-tratamentos e pós-processamento, que muitos algoritmos de diferentes estudos são aplicados em conjunto de dados e avaliam os resultados correspondentes (ZHONG, 2011).

A interação do utilizador com o WEKA resultará na combinação dos módulos de modo a produzir a saída desejada (GARNER, 1995).

Os dois próximos itens a seguir encontram-se os dois Métodos que serão utilizados para o estudo, na literatura.

2.3 MÉTODO DE PROJEÇÃO ALEATÓRIA

Em 1984, surgiu o lema de Johnson e Lindenstrauss, o método de projeção aleatória como um forte método para redução de dimensionalidade.

Em muitas aplicações de mineração de dados, alguns métodos de redução de dimensionalidade são restringidos devido à alta dimensão dos dados. A projeção aleatória é um método que pode ser aplicado em vários tipos de dados como texto, imagem, áudio, entre outros (LIN; GUNOPULOS, 2003).

A ideia do método é simples, segundo os mesmos autores, dada uma matriz *X*, a dimensionalidade dos dados pode ser reduzida pela projeção de uma matriz formada por valores aleatórios: $A[n*k] = X[n*m]*R[m*k]$, onde *k* representa a quantidade de colunas da matriz reduzida.

O método de projeção aleatória é motivado pelo Teorema de Johnson-Lindenstrauss (OLIVEIRA, 2018).

Teorema de Johnson-Lindenstrauss:

Para qualquer $0<\varepsilon<1$ e qualquer inteiro *n*, sendo *k* um inteiro positivo, tal que se forma a equação (1):

$$k \geq 4\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)^{-k} \ln n \qquad (1)$$

Então, para qualquer conjunto V de n pontos em R^m, há uma função de mapeamento f: $R^m \rightarrow R^k$ *tal que para todo* u, v $\in$ V, essas restrições refere-se a equação (2):

$$(1 - \varepsilon)\ ||u - v||^2 \leq ||f(u) - f(v)||^2 \leq (1 + \varepsilon)\ ||u - v||^2 \qquad (2)$$

Considerando o teorema, o desenvolvimento da equação proporciona concluir que um conjunto de *n* pontos em um espaço Euclidiano de alta dimensionalidade pode ser definido como $O(\log n / \varepsilon^2)$ no subespaço dimensional tal que as distâncias entre os pontos são aproximadamente mantidos (CUMPA, 2013).

O teorema de Johnson-Lindenstrauss, segundo o mesmo, mostra que a geometria de um conjunto V com n pontos não é prejudicial para certas projeções ortogonais sobre subespaços de dimensão logarítmica de n, ou seja, é possível projetar V em subespaços de dimensões baixa preservando bem a distância entre eles.

Tipicamente, os elementos em *R* são distribuições Gaussianas, onde uma distribuição Gaussiana é definida por: $G(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, onde μ é a média e σ é o desvio padrão da distribuição.

Achlioptas (2001) propôs duas distribuições a equação (3) ou a equação (4):

$$r_{i,j} = \begin{cases} +1\ com\ probabilidade\ 1/2 \\ -1\ com\ probabilidade\ 1/2 \end{cases} \qquad (3)$$

Ou

$$r_{i,j} = \sqrt{3} * \begin{cases} +1\ com\ probabilidade\ 1/6 \\ 0\ com\ probabilidade\ 2/3 \\ -1\ com\ probabilidade\ 1/6 \end{cases} \qquad (4)$$

Essas distribuições reduzem o tempo computacional para o cálculo de X∗R. Com esse método, segundo Bingham e Mannila (2001) os dados originais de dimensão m são projetados em um subconjunto k (k<<d). Dessa maneira a matriz original X_{n*m} é projetada pela matriz aleatória R_{m*k} obtendo a matriz reduzida A_{n*k}.

2.4 ANÁLISE DE COMPONENTES PRINCIPAIS

Analisando o método da Análise de Componentes Principais (PCA), observa-se que existem vários estudos de aplicações desta técnica estatística cuja dimensionalidade é alta, podendo propor múltiplas opções de aplicações. Esse método elimina informações redundantes, destacando os recursos escondidos, provenientes das informações contidas nas bases e visualiza as relações existentes entre as observações vistas (SOUZA, 2017).

Tendo o mesmo objetivo que o método acima, que é a redução de dimensionalidade, tem um importante papel para o processamento de dados de alta dimensão da Mineração de Dados, reduzindo o volume de informações, mais precisamente o número de atributos, retirando os dados redundantes e irrelevantes de uma determinada base de dados (ZHANG et al., 2010).

O PCA começou a ser chamado como componente principal, criada por Karl Pearson em 1901, e posteriormente consolidada por Harold Hottelling em 1933, sendo utilizada em diversas áreas do conhecimento com o objetivo de reduzir a dimensionalidade e interpretação dos dados do conjunto, transformando subsequentemente em um novo conjunto de variáveis denominado de componentes principais, preservando ao máximo as informações originais (SCHIMITT, 2005).

O método Análise de Componentes Principais (*Principal Component Analysis*), é considerado por Santana (2013), o melhor extrator de características linear conhecido, além de proporcionar a redução da dimensionalidade do conjunto original de dados sem perda significativa das informações.

Este método de aprendizagem não supervisionada encontra a combinação de condições que explicam a maior variação de dados utilizando vários tipos de análises (SOUZA, 2017).

Aplicando o método de PCA converte à obtenção de um novo conjunto de coordenadas, menor que a original, a fim de utilizar para descrição dos dados, mesmo para ser utilizado em outras técnicas de análise ou de mineração de dados. É uma decomposição de valores próprios da matriz de covariância dos dados, utilizado para aproximação de baixa classificação, que compara os dados através de uma função linear de variáveis (SCHIMITT, 2005).

Definindo como uma transformação linear ortogonal que transforma os dados para um novo sistema de coordenada, de forma que a maior variância por qualquer

projeção dos dados fica ao longo da primeira coordenada, chamada primeiro componente, a segunda maior variância fica ao longo da segunda coordenada, chamada segundo componente, e assim por diante (SANTANA, 2013).

Segundo Xu e Wang (2005), matematicamente, os componentes principais são calculados resolvendo o problema do autovalor da matriz de covariância C, como apresentada na equação (5):

$$C_{vi} = \lambda_i v_i \quad (5)$$

A matriz de covariância dos vetores dos dados originais X, ou seja, a matriz quadrada que contém as variâncias e covariâncias associadas a diversas variáveis, é representada por C, λi onde refere-se aos autovalores da matriz C e vi corresponde aos autovetores correspondentes. Consecutivamente até o fim da redução de dimensionalidade dos dados, os autovetores *k*, que correspondem aos maiores autovalores *k*, sendo necessário ser computadorizados. Considerando E_k=[v_1,v_2,v_3,...,v_k] e Λ=[λ_1,λ_2,λ_3,..,λ_k], logo tem-se CEk= $E_K\Lambda$ (XU e WANG, 2005). Com isso, pode-se obter a seguinte equação:

$$X^{PCA} = E_K^T X \quad (6)$$

O número das características da matriz de dados original *X* é reduzido pela multiplicação com a matriz *d* x *k* E_k que tem autovetores *k* correspondentes aos maiores autovalores *k*, em relação a Equação (6), tendo como resultado da matriz é X^{PCA} (SOUZA, 2017).

Quando aplicado o método PCA com o objetivo de redução do número de características, espera-se que os primeiros componentes expliquem uma proporção significativa da variância total dos dados (SANTANA, 2013).

O próximo capítulo refere-se à metodologia empregada nesta pesquisa, onde trata-se das etapas do desenvolvimento do trabalho.

3 METODOLOGIA

Este capítulo descreve a classificação da pesquisa e são apresentadas as etapas propostas para a execução da aplicação dos procedimentos metodológicos utilizados para atingir os objetivos propostos neste trabalho.

3.1 CLASSIFICAÇÃO DA PESQUISA

De acordo com as classificações da pesquisa existentes, o presente estudo mostra abaixo como pode ser classificado os métodos.

A natureza da pesquisa é considerada aplicada, pois, segundo Gil (2008), a pesquisa apresenta como característica principal a aplicação dos conhecimentos, a utilização e consequências práticas destes, pois foram realizadas minerações de dados.

Quanto aos objetivos desta pesquisa será descritiva que, segundo o mesmo autor, considera esta pesquisa uma descoberta existente de associações entre as variáveis com os dados coletados e este trabalho foi realizado juntamente com um *software* para melhor analisar e avaliar os resultados encontrados.

Os procedimentos são de pesquisas documental, baseando-se em Gil (2008), a pesquisa documental coleta dados referente a pessoas, produtos, empresas de maneira indireta, que tomam a forma de documentos, como a base de dados já existente que foi utilizado para o desenvolvimento deste Trabalho de Conclusão de Curso. Neste trabalho, utilizou-se duas bases de dados diferentes para o estudo.

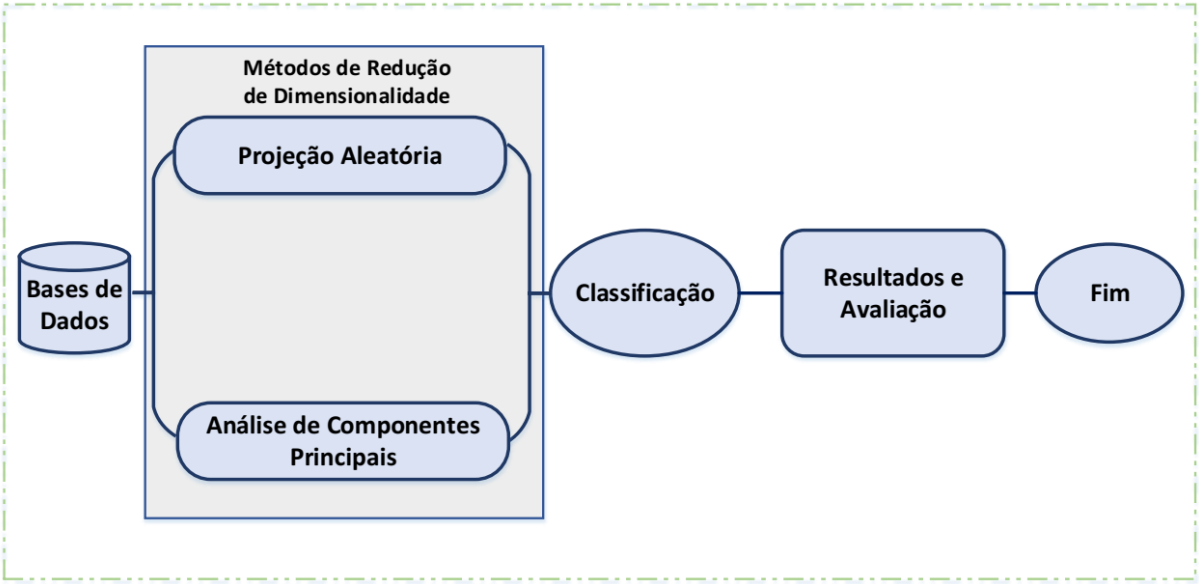
A próxima etapa do processo refere-se à escolha das bases de dados que serão estudadas.

3.2 ETAPAS PARA REALIZAÇÃO DA PESQUISA

Para a realização desta pesquisa, iniciou-se com a escolha das bases de dados para serem estudadas. A segunda etapa, são as aplicações dos Métodos de Projeção Aleatória e Análise de Componentes Principais. A terceira etapa, refere-se a

tarefa de Execução dos Algoritmos de Classificação, tanto nos dados originais quanto nos dados resultantes gerados pelos dois métodos utilizados. E, finalmente, a análise e avaliação dos resultados, junto com a comparação entre os métodos aplicados, como mostrado na Figura 2.

Figura 2 - Etapas para realização do estudo



Fonte: Autoria Própria (2018)

A ferramenta de mineração de dados escolhida para ser utilizado neste estudo foi o *Waikato Environment for Knowledge Analysis* (WEKA) versão 3.8.2 (WEKA, 2017). Para a utilização deste *software*, os dados utilizados devem ser preparados no formato adequado. Por padrão, o formato de arquivo para o WEKA é o ARFF, porém o *software* aceita arquivos CSV. As bases de dados analisadas, encontravam-se neste formato.

3.2.1 Descrição das Bases de Dados

Os dados coletados encontram-se no Repositório *Machine Learning* (UCI, 2017). Considerando que não existe um mínimo de base de dados estabelecido para mineração de dados, selecionou-se duas bases de dados para a aplicação deste trabalho.

A primeira base de dados escolhida são os Dados da Companhia de Seguros, o nome da base é *The Insurance Company (TIC) Benchmark.* Neste trabalho será chamada como “Seguro”*.* Com a obtenção da base foi disponibilizado um conjunto de dados, com aproximadamente 5822 instâncias, contento informações dos clientes referentes as informações de compra ou não das apólices de seguros e o arquivo contém 86 atributos (UCI, 2017).

A segunda base de dados selecionada são os dados de Créditos de um Banco Alemão. A base será chamada de “Crédito” no decorrer deste trabalho, porém o nome original, segundo UCI (2017), é *Stalog*. A base inclui dados sobre os clientes que os classifica conforme o risco de créditos que essas podem apresentar. A base possui ao total 1000 instâncias, com 20 atributos.

A Tabela 1, descreve a quantidade total de atributos e instâncias de cada base para melhor compreensão dos dados.

Tabela 1 – Quantidade total de atributos e instâncias de cada base

Bases de Dados	Atributos	Instâncias
Seguro	86	5822
Crédito	20	1000

Fonte: Autoria Própria

A próxima seção irá tratar da segunda etapa do experimento, que aborda a aplicação dos métodos escolhidos.

3.2.2 Aplicação do Método de Projeção Aleatória

Para a execução do método de projeção aleatória a dimensão do novo conjunto foi definida de acordo com os seguintes critérios: número fixo de atributos e porcentagem de atributos. Para o primeiro critério, utilizou-se os seguintes valores: 10, 40 e 80 atributos. Já para o segundo, foram escolhidos como porcentagem: 10%, 40% e 80%.

3.2.3 Aplicação do Método de Análise de Componentes Principais

Os critérios para utilização do método de Análise de Componentes Principais, deu-se a partir da porcentagem de variância dos dados originais, ou seja, de acordo com o percentual de utilização de dados da base original. Para este trabalho, foram definidas porcentagens com 90%, 95% e 99%.

Para a transformação dos dados em componentes principais, o método não computa o atributo classe. Portanto, os valores correspondentes as classes são juntamente recolocadas para os dados transformados.

Posteriormente, é descrito a etapa de classificação.

3.2.4 Classificação

As bases de dados e os subconjuntos gerados pela aplicação dos métodos foram submetidos à classificação, utilizando os algoritmos Naive Bayes, J48 e SVM.

Na aplicação dos algoritmos é necessário empregar métodos para validar os modelos, para evitar resultados parciais ou tendenciosos. Este método é a chamado de Validação Cruzada Estratificada, divide-se aleatoriamente as bases de dados originais em 10 partições iguais. Após todas as execuções, gera-se a média das avaliações (taxa de acerto).

A taxa de acerto refere-se ao número de instâncias classificados corretamente dividido pelo número total de instâncias, que é necessário quando se trata de modelos preditivos, quanto maior a taxa de acerto, maior a eficiência da base de dados no algoritmo. Um valor considerado ideal para taxa de acerto deve estar entre 70% a 100% (BORGES, 2006).

Como os dados ocorrem alterações dos dados pode haver um conflito entre perda de privacidade e perda de informações. Por isso, é preciso analisar a eficiência de cada técnica no cenário em que ela será aplicada, para verificar se há precisão e confiabilidade nos resultados com confiança (NASCIMENTO, 2017).

3.2.5 Resultados e Avaliação

A última etapa, de resultados e avaliação, visa medir o desempenho dos modelos previstos, verificando a taxa de acerto. Porém, para avaliação dos métodos foram executados primeiramente sobre as bases com todos os atributos para as devidas comparações, ou seja, sem a utilização de nenhum método. Posteriormente, é descrito a etapa de classificação que foi aplicado no *software* para realização da aplicação do estudo.

4 ANÁLISE DOS RESULTADOS

Este capítulo apresenta os principais resultados encontrados. A seção 4.1 refere-se às bases de dados com todos os atributos, a seção 4.2 os resultados referentes a utilização do Método de Projeção Aleatória e a seção 4.3 os resultados referentes a utilização do Método de Análise dos Componentes Principais.

4.1 RESULTADOS COM TODOS OS ATRIBUTOS

Aplicou-se, primeiramente, os algoritmos classificadores Naive Bayes, J48 e SVM nas bases de dados escolhidas, sem a utilização dos métodos. Os resultados obtidos são apresentados na Tabela 2.

Tabela 2 - Resultados da classificação com todos os atributos das bases

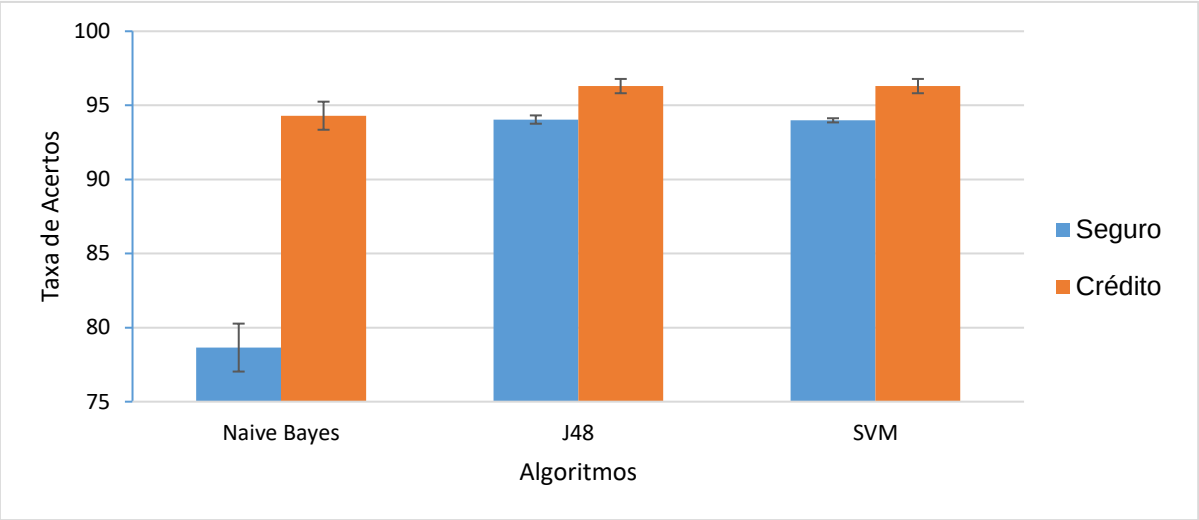
	Algoritmos		
Bases de Dados	**Naive Bayes**	**J48**	**SVM**
Seguro	78,65 ± 1,62	94,04 ± 0,28	93,99 ± 0,14
Crédito	94,30 ± 0,95	96,30 ± 0,48	96,30 ± 0,48

Fonte: Autoria Própria (2018)

Observa-se que na Tabela 2, há uma grande taxa de acerto (acima de 90%) na base de Crédito, na qual nota-se alta eficiência na aplicação dos algoritmos mesmo sem utilizar os métodos de mineração. Em relação à base Seguro, percebe-se que o algoritmo Naive Bayes apresentou desempenho médio inferior, comparando-se aos demais algoritmos.

O Gráfico 1 apresenta um comparativo entre as médias de taxa de acerto utilizando todos os atributos.

Gráfico 1 – Média das taxas de acerto utilizando todos os atributos, nas duas bases de dados analisadas



Fonte: Autoria Própria (2018)

Com relação do Gráfico 1, pode-se perceber que não houve uma diferença significativa nas bases de dados entre os algoritmos J48 e SVM e observa-se que a base de dados Crédito tem maior eficiência do que a base de dados Seguro quando analisados pelo algoritmo Naive Bayes. Nas próximas seções encontram-se os resultados com os métodos sendo aplicados nas bases de dados.

4.2 RESULTADOS DO MÉTODO DE PROJEÇÃO ALEATÓRIA SOBRE AS BASES DE DADOS

Para o método de Projeção Aleatória os resultados foram avaliados de acordo com dois critérios: primeiro, utilizando um número fixo de atributos para a formação do novo subconjunto de atributos e ao outro da porcentagem de atributos.

A seguir encontra-se a Tabela 3, com os resultados obtidos para a base de dados Seguro.

Tabela 3 - Resultados do método de Projeção Aleatória na base de dados Seguro quando utilizado um número fixo de atributos para a formação dos subconjuntos de atributos

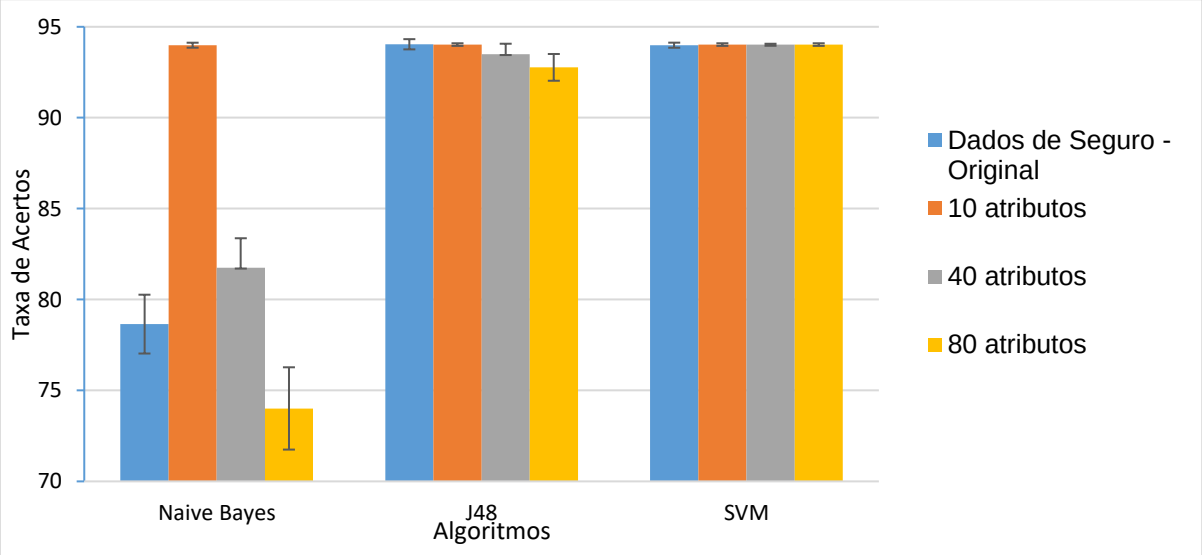
	Algoritmos		
Números de Atributos	**Naive Bayes**	**J48**	**SVM**
10 atributos	93,99 ± 0,14	94,02 ± 0,07	94,02 ± 0,07
40 atributos	81,76 ± 1,61	93,50 ± 0,57	94,01 ± 0,05
80 atributos	74,01 ± 2,26	92,77 ± 0,74	94,02 ± 0,07

Fonte: Autoria Própria (2018)

Observa-se, pelos dados da Tabela 3, que para os algoritmos Naive Bayes e J48, os melhores resultados foram com um subconjunto de atributos formando 10 atributos. Já, a taxa de acerto para os subconjuntos formados por 40 atributos e 80 atributos, apresentaram redução à medida que a formação de atributos era maior, para esses mesmos algoritmos. Em relação ao algoritmo SVM, nota-se que não houve uma diferença significativa na taxa de acerto.

O Gráfico 2 apresenta um comparativo entre as médias das taxas de acerto obtidas por meio de um número fixo de atributos para a base de dados Seguro.

Gráfico 2 - Comparação da base de dados Seguro com o método de Projeção Aleatória utilizado um número fixo de atributos



Fonte: Autoria Própria (2018)

A partir dos dados do Gráfico 2, nota-se que houve diferenças estatisticamente significativas para o algoritmo Naive Bayes, comparando-se o resultado da base de dados original com o resultado do método de projeção aleatória,

mais especificamente na execução do conjunto com 10 atributos. Observa-se, que não houve diferenças significativas nos demais algoritmos (J48 e SVM).

A seguir, a Tabela 4 apresenta os resultados referentes a base de dados Seguro, aplicando uma porcentagem de atributos.

Tabela 4 - Resultados do método de Projeção Aleatória na base de dados Seguro quando utilizado a porcentagem de atributos para a formação dos subconjuntos de atributos

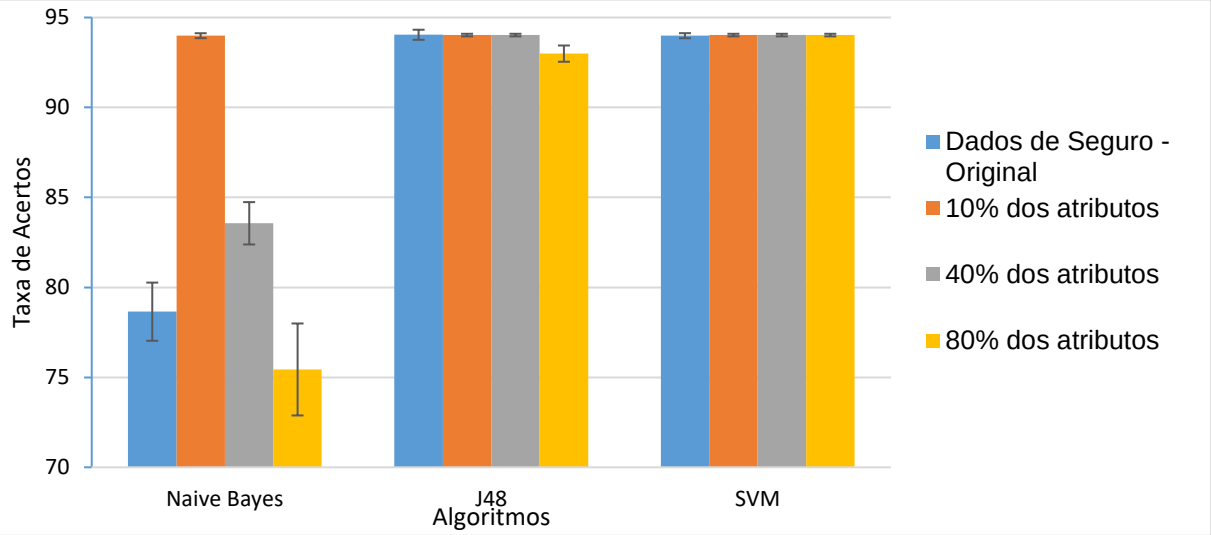
	Algoritmos		
Porcentagem de Atributos	**Naive Bayes**	**J48**	**SVM**
10% dos atributos	93,99 ± 0,14	94,02 ± 0,07	94,02 ± 0,07
40% dos atributos	83,56 ± 1,18	94,02 ± 0,07	94,02 ± 0,07
80% dos atributos	75,44 ± 2,56	92,99 ± 0,45	94,02 ± 0,07

Fonte: Autoria Própria (2018)

Para a Tabela 4, é visto que apenas o algoritmo Naive Bayes apresentou diminuições significativas nas taxas de acerto, quando se aumenta a porcentagem de atributos, o mesmo não ocorrendo para os algoritmos J48 e SVM.

O Gráfico 3 mostra a comparação entre as médias das taxas de acerto, utilizando uma porcentagem de atributos para a base Seguro.

Gráfico 3 - Comparação da base de dados Seguro com o método de Projeção Aleatória utilizado porcentagem dos atributos



Fonte: Autoria Própria (2018)

O Gráfico 3, encontra-se uma semelhança com o Gráfico 2, onde encontrou-se diferenças estatisticamente significativas para o algoritmo Naive Bayes,

comparando-se o resultado da base de dados original com o resultado do método de projeção aleatória, mais especificamente na execução do conjunto com 10 atributos. Considera-se, que não houve diferenças significativas nos demais algoritmos (J48 e SVM).

Na Tabela 5 encontram-se os resultados da base de dados Crédito, quando utilizado um número fixo de atributos para a formação dos subconjuntos de atributos.

Tabela 5 - Resultados do método de Projeção Aleatória na base de dados Crédito quando utilizado um número fixo de atributos para a formação dos subconjuntos de atributos

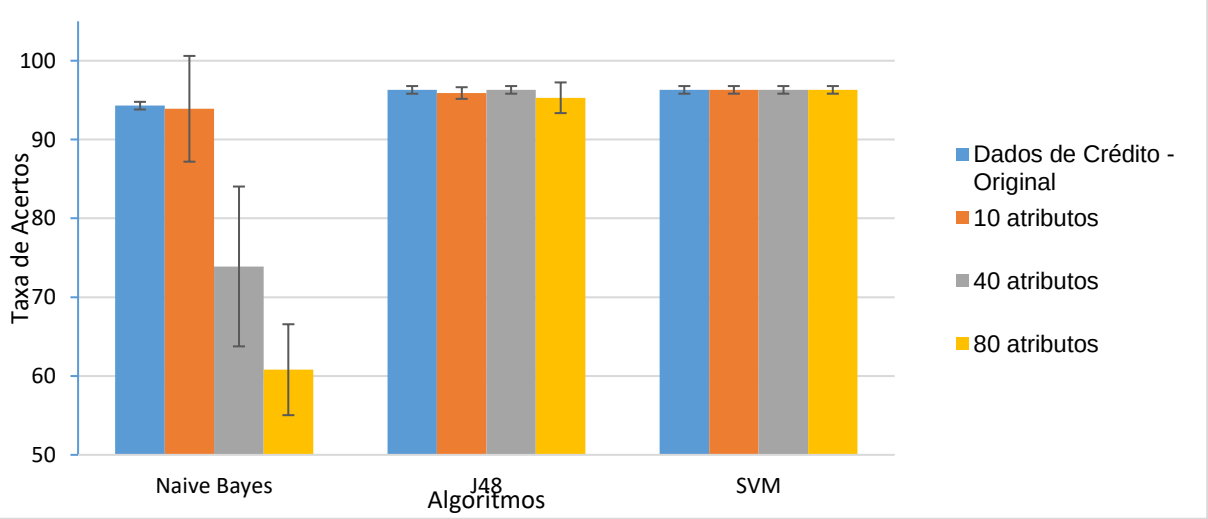
	Algoritmos		
Números de Atributos	**Naive Bayes**	**J48**	**SVM**
10 atributos	93,90 ± 6,71	95,9000 ± 0,74	96,30 ± 0,48
40 atributos	73,90 ± 10,14	96,3000 ± 0,48	96,30 ± 0,48
80 atributos	60,80 ± 5,77	95,3000 ± 1,95	96,30 ± 0,48

Fonte: Autoria Própria (2018)

Analisando a Tabela 5, identifica-se uma grande variação da taxa de acerto quando ocorre o aumento do número de atributos, mais precisamente para o algoritmo Naive Bayes. Para os resultados obtidos com a utilização dos algoritmos J48 e SVM, constata-se que não há diferença significativa da taxa de acerto, independentemente da quantidade de números de atributos.

O Gráfico 4, encontrado abaixo, obtém os resultados comprando as taxas de acerto com a taxa de acerto da base original.

Gráfico 4 - Comparação da base de dados Crédito com o método de Projeção Aleatória utilizado número fixo de atributos



Fonte: Autoria Própria (2018)

Nota-se que no Gráfico 4, a taxa de acerto manteve-se constante para os algoritmos J48 e SVM. Entretanto, para o algoritmo Naive Bayes, percebe-se uma variação estatisticamente significativa entre o resultado da base original com o resultado do método de projeção aleatória na execução do conjunto com 40 atributos e 80 atributos.

A Tabela 6, abaixo, possui os resultados da base de dados Crédito utilizando uma porcentagem de atributos para a formação dos subconjuntos de atributos.

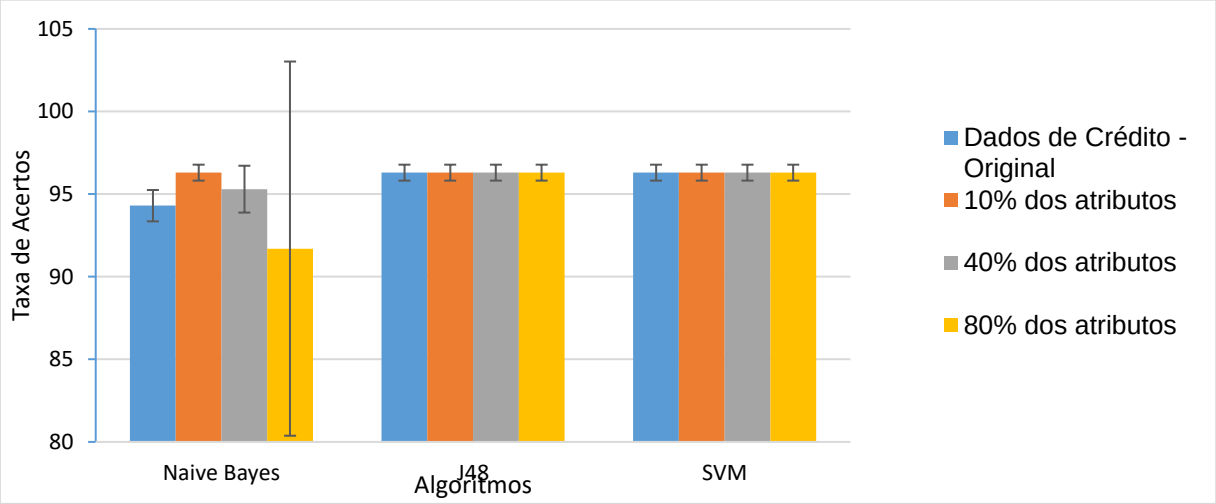
Tabela 6 - Resultados do método de Projeção Aleatória na base de dados Crédito quando utilizado a porcentagem de atributos para a formação dos subconjuntos de atributos

	Algoritmos		
Porcentagem de Atributos	**Naive Bayes**	**J48**	**SVM**
10% dos atributos	96,30 ± 0,48	96,30 ± 0,48	96,30 ± 0,48
40% dos atributos	95,30 ± 1,42	96,30 ± 0,48	96,30 ± 0,48
80% dos atributos	91,70 ± 11,32	96,30 ± 0,48	96,30 ± 0,48

Fonte: Autoria Própria (2018)

A Tabela 6 mostra que todos as taxas de acerto encontrados foram acima de 90%, sendo que nos algoritmos J48 e SVM apresentaram a mesma taxa (96,30%). Portanto, não houve diferenças significativas independentemente da quantidade da porcentagem de atributos e do algoritmo utilizado. A seguir, encontram-se os resultados expostos no Gráfico 5, comparando com a taxa de acerto da base original.

Gráfico 5 - Comparação da base de dados Crédito com o método de Projeção Aleatória utilizado porcentagem de atributos



Fonte: Autoria Própria (2018)

O Gráfico 5 mostra que não há diferenças mínimas significativas entre o resultado da base original com os resultados obtidos pela aplicação do método de Projeção Aleatória utilizando porcentagem de atributos.

4.3 RESULTADOS DO MÉTODO DE ANÁLISE DE COMPONENTES PRINCIPAIS SOBRE AS BASES DE DADOS

Neste tópico, serão apresentados os resultados pertencentes ao método de Análise de Componentes Principais. Os critérios para esse método foram de acordo com a porcentagem de variância, que nesse estudo foi de 90%, 95% e 99% dos atributos.

A Tabela 7, possui os resultados do método de Análise dos Componentes Principais na base de dados Seguro quando utilizado a porcentagem dos atributos para a formação dos subconjuntos de atributos.

Tabela 7 - Resultados do método de Análise dos Componentes Principais na base de dados Seguro quando utilizado a porcentagem dos atributos para a formação dos subconjuntos de atributos

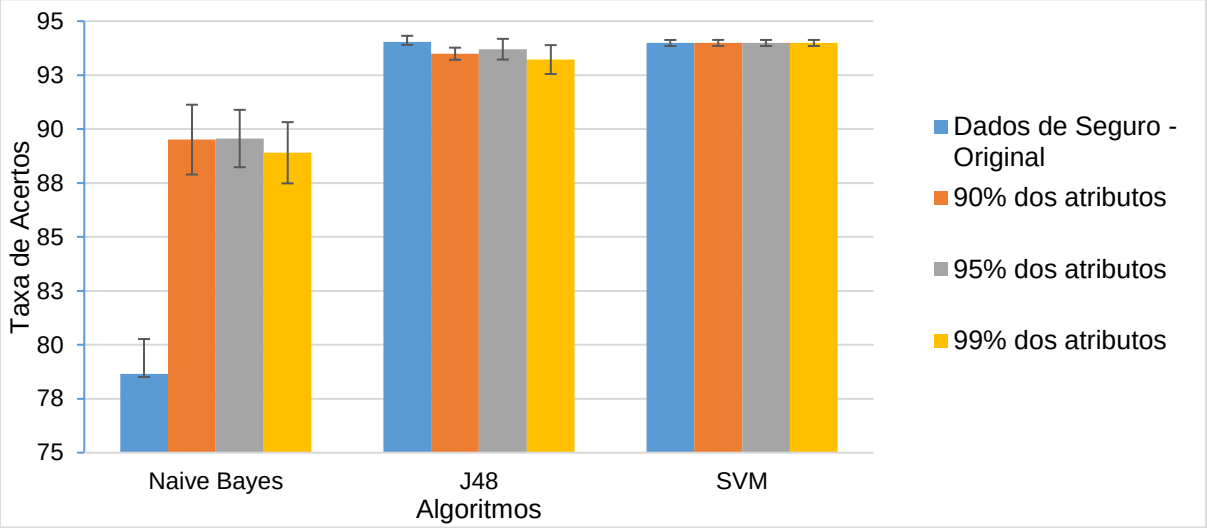
	Algoritmos		
Porcentagem de Atributos	**Naive Bayes**	**J48**	**SVM**
90% dos atributos	89,51 ± 1,31	93,49 ± 0,69	93,99 ± 0,14
95% dos atributos	89,56 ± 1,33	93,70 ± 0,48	93,99 ± 0,14
99% dos atributos	88,90 ± 1,42	93,22 ± 0,67	93,99 ± 0,14

Fonte: Autoria Própria (2018)

Analisando os resultados da Tabela 7, nota-se que para todos os algoritmos as taxas de acerto apresentam resultados favoráveis. O algoritmo com melhores resultados foi o SVM, tendo como taxa de acerto 93,99% para ambas as porcentagens de atributos.

O Gráfico 6 ilustra os resultados obtidos através da aplicação do método, além dos resultados da base original.

Gráfico 6 - Comparação da base de dados Seguro com o método de Análise dos Componentes Principais utilizado porcentagem dos atributos



Fonte: Autoria Própria (2018)

A partir dos dados do Gráfico 6, evidencia-se que não houve melhora nas taxas de acertos, tanto para o algoritmo J48 quanto para o algoritmo SVM, porém houve um aumento estatisticamente significativo quando comparado o resultado da base original com os resultados obtidos pelo método PCA para o algoritmo Naive Bayes.

Posteriormente, a Tabela 8 mostra os resultados para a base de dados Crédito.

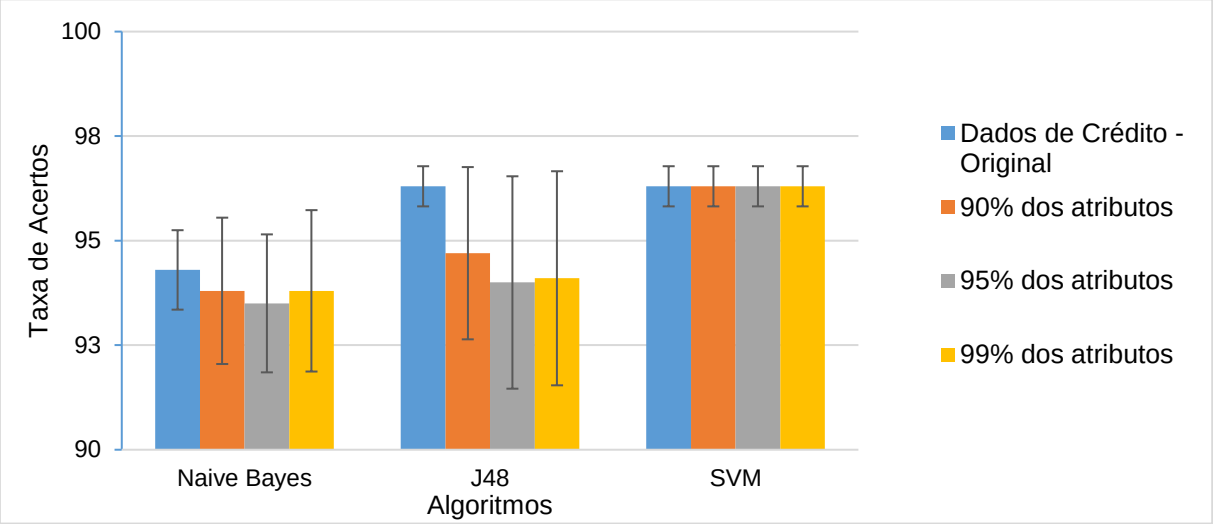
Tabela 8 - Resultados do método de Análise de Componentes Principais na base de dados Crédito quando utilizado a porcentagem de atributos para a formação dos subconjuntos de atributos

	Algoritmos		
Porcentagem de Atributos	**Naive Bayes**	**J48**	**SVM**
90% dos atributos	93,80 ± 1,75	94,70 ± 2,06	96,30 ± 0,48
95% dos atributos	93,50 ± 1,65	94,00 ± 2,54	96,30 ± 0,48
99% dos atributos	93,80 ± 1,93	94,10 ± 2,56	96,30 ± 0,48

Fonte: Autoria Própria (2018)

Para a Tabela 8, percebe-se que não houve uma diferença significativa entre as porcentagens para ambos os algoritmos. O algoritmo com desempenho melhor foi o SVM, com uma média de 96,30%. Para melhor visualização tem-se o Gráfico 7 abaixo.

Gráfico 7 - Comparação da base de dados Crédito com o método de Análise dos Componentes Principais utilizado porcentagem de atributos



Fonte: Autoria Própria (2018)

Pelos resultados do Gráfico 7, constata-se que não houve melhoras significativas nas taxas de acerto ao se aplicar o método PCA para ambos os algoritmos, no entanto as taxas de acerto tiveram média superior a 90%, o que se apresenta como sendo um bom método de análise.

Comparando os dois métodos de redução de dimensionalidade, Projeção Aleatória e Análise de Componentes Principais, através dos dados encontrados observa-se que os resultados foram melhores comparados com os obtidos quando aplicados nas bases de dados com todos os atributos.

Os resultados obtidos em comparação com o Métodos de Projeção Aleatória e do PCA, observa-se que os resultados da Projeção Aleatória há um aumento muito significativo na taxa de acerto com os atributos em 10 e 10%, tendo uma melhora na seleção dos resultados. Já o PCA não houve uma variação significativa comparando com a taxa de acerto com os resultados sem a aplicação do método. Sendo assim, conclui-se que o Método de Projeção Aleatória com resultados mais aparentes que o PCA.

5 CONCLUSÃO

Neste trabalho, o objetivo geral foi avaliar com a aplicação dos métodos de Projeção Aleatória e Análise de Componentes Principais em dados de Crédito e Seguro de clientes. Tendo a necessidade de realizar com os objetivos específicos propostos.

Os objetivos específicos foram realizados a partir do processo de conhecimento e descoberta dos dados. O primeiro objetivo específico foi selecionar as bases de dados utilizadas para a aplicação do estudo, nesse estudo, a partir de dados de clientes de Crédito e Seguro.

O segundo objetivo específico foi aplicar o método de Projeção Aleatória e o método de Análise de Componentes Principais nas respectivas bases de dados. Para o método de Projeção aleatória as experiências foram de 10, 40 e 80 atributos e por porcentagem de 10%, 40% e 80%, e para o método de Análise de Componentes Principais foram por porcentagens de 90%, 95% e 99%.

Outro objetivo específico era realizar a classificação dos algoritmos nas bases de dados, com a ferramenta WEKA, onde utilizou-se três algoritmos, sendo eles, Naive Bayes, J48 e SVM, desejando analisar as taxas de acerto destes.

O último objetivo específico foi avaliar as informações que foram coletadas após a aplicação dos métodos adotados. E realizando a redução de dimensionalidade busca padrões e conhecimentos úteis, eliminando os atributos que não são significativos.

Diante dos resultados encontrados, praticamente todos encontram-se na variação considerada ideal (70% a 100%) não sendo necessário descartar as aplicações dos mesmos, somente um dos resultados do método de projeção aleatória na base de dados de crédito quando utilizado uma quantidade de 80 dos atributos para a formação dos subconjuntos de atributos foi encontrado um valor inferior a esse ideal.

Analisando primeiramente pelo Método de Projeção Aleatória conclui-se que os algoritmos SVM e J48 foram os que produziram os melhores resultados, com pequenas variações na taxa de acerto quando mudava o número de atributos ou a porcentagem dos atributos.

E o algoritmo Naive Bayes teve uma grande variação de resultados durante a classificação, onde encontrou-se melhores resultados desse algoritmo em 10 atributos

e 10% dos atributos, não havendo vantagem no aumento da porcentagem dos atributos para a aplicação dos algoritmos.

Os atributos mais relevantes são de 10 atributo e 10% dos atributos através dos métodos de redução de dimensionalidade em bases de dados de clientes, utilizando os métodos Naive Bayes, J48 e SVM visto o Método de Projeção Aleatória.

Observando o Método de Análise dos Componentes Principais, não se obteve uma variação significativa entre os algoritmos, e nem uma melhora significativa com a variação de porcentagem de atributos encolhidos para a quantidade que houve a aplicação. Somente na Base de Dados de Seguro teve uma melhora em um dos algoritmos, Naive Bayes*,* não havendo melhorias de taxa de acerto nos demais.

A aplicação dos métodos de projeção aleatória e análise dos componentes principais teve o objetivo de aumentar a eficiência das bases de dados Seguro e Crédito produzindo bons resultados. De modo geral, os resultados dos experimentos comprovam que as aplicações desses métodos de redução de dimensionalidade produzem uma taxa de acerto do classificador maior do que quando aplicado somente o algoritmo de mineração sobre as bases de dados com todos os atributos.

Do ponto de vista dos colaboradores das organizações (gerente bancário, analistas de créditos, seguradores, corretores), sempre existe a vantagem em se utilizar esta ferramenta, porque a mesma mostra seus resultados, regras de classificação, em uma forma fácil de compreender, detalhando quais os atributos, ou seja, as informações das empresas analisadas foram mais relevantes para as suas classificações com a taxa de acerto satisfatória.

Desta forma, a empresa pode conferir se os resultados obtidos por esta técnica combinam, ou não, com a sua experiência e utilizá-la na análise de novas propostas de crédito e seguros com uma margem de segurança satisfatória como um apoio adicional as suas tomadas de decisões.

REFERÊNCIAS

ACHLIOPTAS, D. Database-friendly random projections. In Proc. ACM Symp. **On the Principles of Database Systems**, p. 274-281, 2001.

ALVES, F. C.; FRAGAL, E. H. Avaliação dos Algoritmos MAXVER e SVM na Classificação da Cobertura Vegetacional da Planície de Inundação do Alto Rio Paraná**.** In: **Encontro Estadual De Geografia E Ensino, II, Semana De Geografia**, XX, 2011, Maringá. Anais... Maringá, p. 621-632. 2011.

ARBACHE, F. S. et al. **Gestão de logística, distribuição e trade marketing.** Editora FGV, 4° ed. Rio de Janeiro – RJ, 2011.

BCB, BANCO CENTRAL DO BRASIL. **Estatísticas Monetárias e de crédito**. Disponível em: <https://www.bcb.gov.br/htms/notecon2-p.asp>. Acesso em: 12 mai. 2018.

BERTON, L. **Caracterização de classes e detecção de outliers em redes complexas*.*** 2011. 89f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Universidade de São Paulo. São Carlos, 2011.

BINGHAM, E.; MANNILA, H. Random projection in dimensionality reduction: applications to image and text data. **In Knowledge Discovery and Data Mining**, p. 245-250, 2001.

BORGES, H. B. **Redução de dimensionalidade em bases de dados de expressão gênica*.*** *2006.* 123 f. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Paraná. Curitiba, 2006.

BORGES, H. B.; NIEVOLA, J. C. Comparing the dimensionality reduction methods in gene expression databases. **Expert Systems with Applications**, vol. 39, p. 10780–10795, 2012.

CAMURÇA, J. O.; MAGALHÃES, L. L. A satisfação dos clientes atacadistas em uma empresa de confecção atacadista cearense: um estudo de caso da New Impact. **Revista de Administração da UNI7**, v. 1, n. 1, p. 089-119, 2017.

COSTA, E. et al. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, p. 1-29, 2012.

CRUZ, A. J. R. **Data Mining via redes neuronais artificiais e máquinas de vectores de suporte.** Lisboa: UM, 2007. 123f. Dissertação (Mestrado em Sistemas de Informação), Universidade do Moinho, Lisboa, 2007.

CUMPA, M. A. O. **Análise em Grassmannianas e o Teorema de Johnson-Lindenstrauss.** Dissertação de Mestrado. Pontifícia Universidade Católica do Rio de Janeiro. PUC-RJ. 2013.

DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. Acta Scientiarum. **Technology**, v. 24, p. 1715-1725, 2002.

DURSUN, A.; CABER, M. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. **Tourism management perspectives,** v. 18, p. 153–160, 2016.

EICHORN, F. L. Internal Customer Relationship Management (IntCRM): A Framework for Achieving Customer Relationship Management from the Inside Out. **Problems and Perspectives in Management**, v.2, n.1, 2004.

FAYYAD, U. et al. From Data Mining to Knowledge Discovery in Databases. **American Association for Artificial Intelligence**, 1996.

FRUTUOSO, D. G. **Recuperação de informação e classificação de entidades organizacionais em textos não estruturados.** 2014. 86 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Pernambuco. Recife, 2014.

GARNER, S. R. et al. Weka: The waikato environment for knowledge analysis. In: **Proceedings of the New Zealand computer science research students conference,.** p. 57-64. 1995.

GIL, A. C. **Como elaborar projetos de pesquisa.** São Paulo: Atlas, 2008.

HAN, J; KAMBER, M. Data Mining: Concepts & Techniques. University of Ilinois at Urbana-Champaign: **Elsevier**, 2006.

HASTIE, T. et al. Classification by pairwise coupling. **Annals of statistics**, v. 26, n. 2, p. 451-471, 1998.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. **Morgan Kaufmann Publishers Inc.**, 1995. p. 338-345.

KAMBER, M. et al. Data mining: Concepts and techniques. **Elsevier**, 2012.

LIERENA, S. E. **Redução dimensional de dados de alta dimensão e poucas amostras usando Projection Pursuit**. Tese (Doutorado em Engenharia Elétrica) – Universidade de São Paulo. São Paulo – SP. 2013.

LIN, J.; GUNOPULOS, D. Dimensionality Reduction by Random Projection and Latent Semantic Indexing. In proceedings of the Data Mining Workshop, at the **3th SIAM International Conference on Data Mining**. San Francisco, CA. 03 Mai. 2003.

MENGUC, B.; AUH, S.; SHIH. Transformational Leasership and Market Orientation: Implications for the Implementation of Competitive Strategies and Business Unit Performance. **Journal of Business Research**, v. 60, p. 314-321, 2007.

MITCHELL, T. M. **Machine learning.** Boston: WCB/McGraw-Hill, 2010.

NAIK, A.; SAMANT, L. Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. **Procedia Computer Science**, v. 85, p. 662-668, 2016.

NASCIMENTO, F. R. S. **Um Estudo Comparativo entre Algoritmos de Proteção da Privacidade e Segurança de Dados Aplicado à Bases de Dados na Área de Saúde.** 2017. 52 f. Trabalho de Conclusão de Curso (Graduação em Sistema de Informação) - Universidade Federal do Rio Grande do Norte. Caicó – RN. 2017.

NGAI, E, W, T.; XIU, L.; CHAU, D. C. K. Application of data mining techniques in customer relationship management: a literature review and classification. **Expert Systems with Application**, v. 36, n. 2, part 1, p. 2592-2602, mar. 2009.

OLIVEIRA, S. R. **Métodos Usados para Redução e Sintetização de Dados**. Disponível em:

<https://www.ime.unicamp.br/~wanderson/Aulas/MT803_Aula3_Reducao_Sintetizacao_Dados.pdf>. Acesso em: 12 Mai. 2018.

PASTA, A. **Aplicação da técnica de Data Mining na base de dados do ambiente de gestão educacional:** um estudo de caso de uma instituição de ensino superior de Blumenau-SC. Blumenau: UVI, 2011. Dissertação (Mestrado em Computação Aplicada), Universidade do Vale do Itajaí, São José-SC, 2011.

PLATTS, K. et al. Testing Manufacturing Strategy Formulation Processes. **Internacional Journal of Production Economics**, v.56-57, p. 517-523, 1998.

PRAHALAD, C. K.; KRISHNAN, M. S. A nova era da inovação: a inovação focada no relacionamento com o cliente. Rio de Janeiro: **Elsevier**, 2008.

RIBEIRO, A. H. P.; GRISI, C. C. H.; SALIBY, P. E. Marketing de relacionamento como fator-chave de sucesso no mercado de seguros. **Revista de Administração de Empresas**, v. 39, n. 1, p. 31-41, 1999.

ROMDHANE, L. B. et al. An efficient approach for building customer profiles from business data. **Expert Systems with Applications,** v.37, p.1573-1585, 2010.

SANTANA, G. A. **Uma abordagem para a identificação automática de problemas de usabilidade em interfaces de sistemas web através de reconhecimento de padrões**. 2013. 141 f. Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Tecnológica Federal do Paraná. Cornélio Procópio-PR, 2013.

SANTOS, C. P.; FERNANDES, D. V. D. H**.** A recuperação de serviços como ferramenta de relacionamento e seu impacto na confiança e lealdade dos clientes. **Revista de administração de empresas**, v. 48, n. 1, p. 10–24, 2008.

SCHMITT, J. et al. **Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo.** Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Santa Catarina. Florianópolis–SC, 2005.

SCHUCH, R. et al. **Mineração de dados em uma subestação de energia elétrica.** In: Proceedings Of The 9th Brazilian Conference On Dynamics, Controland Their Applications – Dincon'10. Anais... Serra Negra, p. 804–810, jun. 2010.

SIQUEIRA, D. M. R. et al. A Pesquisa e Análise de Satisfação como Ferramenta de Gestão do Relacionamento com o Consumidor. **Revista FAIPE**, v. 4, n. 1, p. 12-18, jun. 2014.

SOARES, M. M.; SOBRINHO, M. **Microfinanças**: O papel do Banco Central do Brasil e a importância do cooperativismo de crédito**.** Brasília: BCB, 2008.

SOUZA, J. T. de. **Métodos de Seleção de Atributos e Análise de Componentes Principais:** um estudo comparativo. 2017. 78 f. Dissertação (Mestrado em Engenharia de Produção) - Universidade Tecnológica Federal do Paraná. Ponta Grossa-PR. 2017.

TREVISAN, G. M. M. **O uso da Mineração de Dados na descoberta de conhecimento em empresa do setor agrícola.** 2017. Dissertação (Mestrado em Engenharia de Produção) – Universidade de Araraquara. Araraquara-SP. 2017.

TSS, TUDO SOBRE SEGUROS. **Seguro de Automóveis.** Disponível em <http://www.tudosobreseguros.org.br/portal/pagina.php?c=1215>. Acesso em: 12 mai. 2018.

UCI Machine Learning Repository**. Browse Through:** 5 data sets. Disponível em: <http://archive.ics.uci.edu/ml/datasets.html?format=&task=&att=&area=bus&numAtt=&numIns=&type=&sort=nameUp&view=table>http://archive.ics.uci.edu/ml/>. Acesso em: 20 nov. 2017.

WEKA. **The University of Waikato.** Software. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 22 ago. 2017.

WU, X.; et al. Top 10 algorithms in data mining. **Knowledge and information systems***,* v. 14, n. 1, p. 1–37, 2008.

XU, X.; WANG, X**. An adaptive network intrusion detection method based on PCA and support vector machines**. In X. LI, S.; WANG, Z. Y. Dong (Eds.), Advanced data mining and applications, first international conference (ADMA 2005), July 22–24, 2005. Wuhan, China: Proceedings. Lecture notes in computer science (v. 3584, p. 696–703). Springer, 2005.

YAMAGUCHI, J. K. **Diretrizes para a escolha de técnicas de visualização aplicadas no processo de extração do conhecimento**. 2010. 182f. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Maringá, Maringá, 2010.

ZEITHAML, V. et al. **Marketing de Serviços.** A empresa com foco no cliente. 6. Ed. Parte II, p. 49, Nova York: AMGH, 2014.

ZHANG, Z. et al. Effective multiplicative updates for non-negative discriminative learning in multimodal dimensionality reduction. **Artificial Intelligence Review**, v.34, n.3, p. 235–260, 2010.

ZHONG, X. The research and application of web log mining based on the platform weka. **Procedia engineering**, v. 15, p. 4073–4078, 2011.

ZIAFAT, H.; SHAKERI, M. Using Data Mining Techniques in Customer Segmentation. **International Journal of Engineering Research and Applications**, v. 4, n.9, p. 70–79, 2014.