Um estudo sobre Publicações Relacionadas a Mineração de Dados

Carlos Alves de Souza Junior[1], Humberto F. Villela[1]

[1]Ciência da Computação – Universidade FUMEC
R. Cobre, 200 - Cruzeiro Caixa Postal 30310 – 190 – Belo Horizonte – MG – Brasil

carlos.alves.souza.junior@gmail.com, humberto.villela@fumec.br

Resumo.

A grande quantidade de dados acumulados nos bancos de dados informatizados das organizações pode esconder conhecimentos valiosos e úteis para a tomada de decisão. A mineração de dados é uma das técnicas adotadas para identificar estes padrões. O objetivo do trabalho é identificar o uso de algoritmos, softwares e linguagens aplicadas à mineração de dados. A metodologia utilizada foi através da revisão sistemática da literatura. Foram utilizados dissertações de mestrado, teses de doutorado e feitas pesquisas bibliográficas nas bases: biblioteca digital USP, Biblioteca digital UFMG e Sistema PUC MINAS, nas áreas: saúde, educação, vendas, agro economia, cooperativa de crédito, ciência da computação e bioinformática.

Abstract.

The large amount of data accumulated in the computerized databases of organizations can hide valuable and useful knowledge for decision making. Data mining is one of the techniques used to identify these patterns. The objective of this work is to identify the use of algorithms, software and languages applied to data mining. The methodology used was through the systematic review of the literature. Master's dissertations, doctoral theses and bibliographic research were used in the following databases: USP digital library, UFMG digital library and PUC MINAS System, in the areas of health, education, sales, agro economy, credit cooperative, computer science and bioinformatics.

1. Introdução

A mineração de dados nos últimos anos tem apresentado crescimento exponencial de dados gerados em praticamente quase todas as áreas de atividade humana, seja ela cientifica, comercial, lazer, industrial, entre outras (GONZALES, 2014).

Segundo Rezende (2005) a evolução da computação possibilitou um aumento na capacidade de processamento e armazenamento de dados. A facilidade atual que uma

aplicação cientifica e/ou comercial possui para gerar gigabytes ou terabytes de dados excede em muito a capacidade de pesquisadores e analistas de mercado em fazer análises sobre os mesmos. A mineração de dados pode ser vista como a sistematização de teorias, técnicas e algoritmos desenvolvidos em outras disciplinas já consagradas, como a Estatística, a Inteligência Artificial, o Aprendizado de Máquina, a base de dados etc. O propósito da mineração de dados é detectar automaticamente padrões de associação úteis e não óbvios em grandes quantidades de dados (REZENDE, 2005).

Atualmente, os analistas de negócios precisam usar ferramentas capazes de responder a perguntas complexas como: "qual produto de alta lucratividade venderia mais com a promoção de um item de baixa lucratividade, analisando os dados dos últimos cinco anos de vendas?" (REZENDE, 2005).

Han e Kamber (2006) mostra, na Figura 1, o posicionamento lógico de diferentes fases da tomada de decisão com seu valor potencial para as dimensões tática e estratégica. O valor estratégico da informação aumenta quando os dados estão altamente resumidos.

FIGURA 1 - Mineração de dados e Inteligência de negócios

Fonte: HAN; KAMBER, 2006, p. 12.

Diante desta deficiência para analisar e compreender grande volume dados, diversos estudos têm sido direcionados ao desenvolvimento de tecnologias de extração automática de conhecimento de base de dados. Uma das técnicas para extração do conhecimento geralmente referenciada na literatura é o Knowledge Discovery in Database (KDD) (BRANQUINHO, 2015).

Para Rezende (2005) as técnicas de análise de dados geralmente não extrapolam a realização de consultas SQL (Structured Query Language) simples, a utilização de

ferramentas OLAP ou mecanismos de visualização de dados. Todo o processo de mineração de dados é orientado em função de seu domínio de aplicação e dos repositórios de dados inerentes aos mesmos. Para usar os dados é necessário que estejam estruturados de forma a serem consultados e analisados adequadamente.

Os sistemas de aplicações, conhecidos por (On-line Transaction Processing) OLTP, processam dados armazenados em base de dados relacionais usadas para armazenar, consultar e alterar informações do negócio. Normalmente, não é possível aplicar as técnicas de mineração de dados diretamente a essas bases, pois isso poderia resultar numa sobrecarga de consultas podendo literalmente "travar" um sistema, impossibilitando qualquer outro tipo de operação transacional. Assim é recomendável que os dados a serem utilizados na descoberta de conhecimento estejam separados da base de dados operacional (REZENDE, 2005).

A aplicação de mineração de dados pelas empresas de cartões de crédito na análise de Banco de dados de clientes para identificar seus diferentes grupos e predizer seu comportamento, de forma a direcionar as atividades de marketing (GONZALES, 2014).

Na área da saúde oferece inúmeras possibilidades de aplicações destas técnicas devido a complexidades dos processos e o grande volume de armazenamento de seus dados em uso pelos sistemas de informação.

Para Branquinho (2015) estes bancos de dados dos sistemas de saúde têm sido analisados para tornar mais efetiva a recuperação de informações sobre o comportamento de prescrição de testes laboratórios, no caso deste relacionados as hepatites virais.

Para o mercado de medicina diagnóstica é importante entender o comportamento de prescrição dos médicos no diagnóstico das doenças para antecipar tendências e assim realizar ações de marketing e vendas direcionadas ao mercado. Portanto, entender o domínio dos testes laboratoriais complementares ao diagnóstico possibilita aprimorar os padrões extraídos no processo de mineração de dados (BRANQUINHO, 2015).

Na educação a mineração de dados pode orientar os alunos no processo de aprendizagem, guiar os professores e tutores na melhoria do processo de ensino e, ainda, auxiliar coordenadores e gestores no processo administrativo do ambiente educacional (RODRIGUES, 2016).

Atualmente o processo-e-aprendizagem (EA) está fortemente suportado pelas tecnologias de aprendizagem e pelo aumento de conectividade, os quais são utilizados nos ambientes educacionais tanto presenciais como a distância (RODRIGUES, 2016).

Para Rodrigues (2016) dentro do cenário atual, a mineração de dados, do inglês Data-Mining (DM) é uma forte aliada para exportar esses dados com o objetivo de extrair conhecimento e informações uteis para tornar o processo de EA mais eficiente.

Com os resultados obtidos demonstrou-se, na prática, como as diversas tecnologias ligadas ao processo de mineração de dados e organização do conhecimento podem apoiar as tomadas de decisões, de forma a entender o comportamento das diversas áreas estudadas (BRANQUINHO, 2015).

Para Alves (2018) a mineração de dados tem um grande potencial para aumentar receitas e reduzir custos. Organizações inovadoras em todo mundo utilizam dados resultantes de processos de mineração para, por exemplo, aumentar as vendas ou minimizar as perdas.

O objetivo deste trabalho é identificar o uso de algoritmos, softwares e linguagens aplicadas à mineração de dados.

A justificativa deste trabalho é devido a uma quantidade incalculável de dados que é gerada na forma de registros de vendas, textos brutos, imagens, sons, gráficos nas diversas áreas, tanto por sistemas computacionais como por seres humanos, constituindo uma espécie de informação não estruturada. Embora esta forma de registro de dados seja adequada para o ser humano, quando se trata de analisar grandes quantidades de dados de forma automatizada, é comum e conveniente que se introduza mineração de dados para facilitar o acesso e o processamento sistemático para obter maior precisão na descoberta de conhecimento.

2. Referencial Teórico

Data Mining é uma técnica que se aplica a uma grande quantidade de dados e informações que muitas vezes estão escondidas nos bancos de dados das corporações. Essa técnica pode ser aplicada em qualquer segmento (medicina, vendas, marketing, entre outras) que trabalhe com um grande volume de dados armazenados, porém para que essas informações possam ser analisadas de forma coerente, um especialista no assunto não pode ser dispensado (CARVALHO, 2002).

Ainda segundo Carvalho (2002), apesar de ser uma técnica antiga, a utilização da mesma passou a ser aplicada por muitas empresas devido aos motivos como grande volume de dados, organização dos dados, recursos computacionais muito mais potentes e competição de mercado.

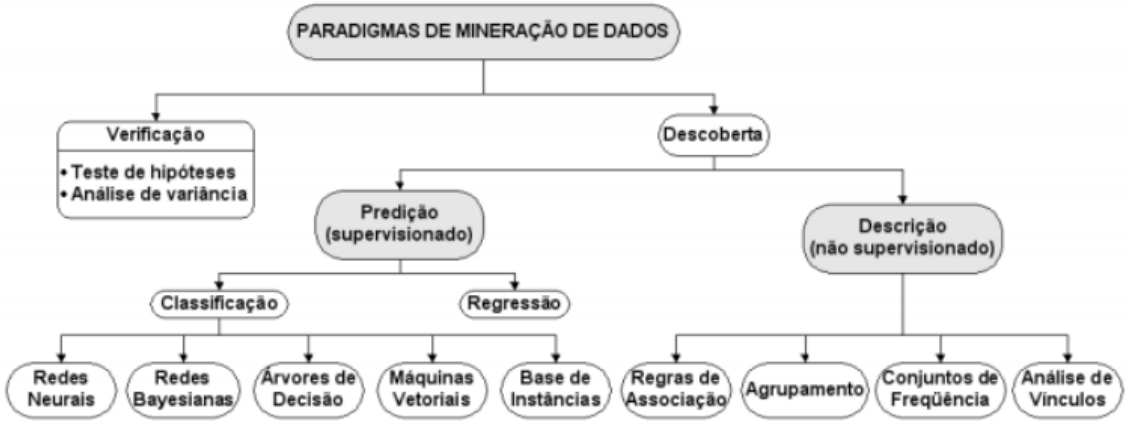
O Data Mining apenas se aplica nas grandes massas de dados, isso para que os algoritmos de mineração de dados tenham a sua real contribuição e, por fim, extrair as informações necessárias de forma confiável.

As organizações atuais então padronizando seus dados utilizando (Data Warehouse) DW auxiliando na tomada de decisões, mas muitas vezes ainda não são extraídas as informações da forma mais inteligente, por isso a partir dessa organização e padronização dos dados, aplicam-se as técnicas de Data Mining, buscando de forma inteligente os melhores resultados.

O Data Mining exige um poder computacional muito elevado para executar os seus algoritmos, porém com a evolução da microeletrônica e o baixo custo dos equipamentos, isso possibilita a aplicação do Data Mining. Outro fator que favoreceu muito foram os bancos de dados bem distribuídos, onde as informações estão mais coerentes e robustas.

As empresas geram um grande volume de dados nos seus sistemas, porém não sabem como utilizar a informação. A mineração de dados auxilia essas empresas nas tomadas de decisões, tais como promoção de vendas, CRM (sistema de relacionamento com o cliente), pró-atividade no tratamento de saúde, entre vários outros segmentos, tornando assim a empresa mais forte perante o mercado extremamente competitivo.

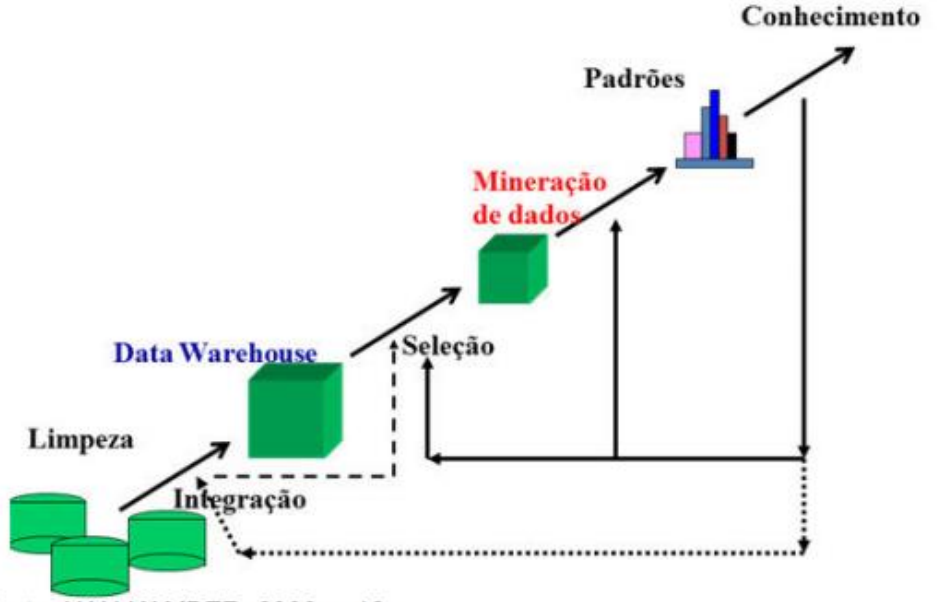
De acordo com Branquinho (2015) a mineração de dados possui várias técnicas que variam de acordo com o objetivo a ser alcançado. A descoberta de padrões pode ser pré-definida, ou seja, supervisionada com características prescritivas e preditivas ou não supervisionada, onde não há conhecimento prévio, com características descritivas.



Fonte: BRANQUINHO, 2015, p.19.

Para Branquinho (2015) a figura 2 representa KDD sobre a perspectiva de Han e Kamber (2006). Para um melhor entendimento o processo de KDD pode ser agrupado em três fases: pré-processamento, mineração de dados e pós-processamento. O pré-processamento compreende a captação, organização e tratamento dos dados; já a mineração de dados, os algoritmos e as técnicas para busca dos padrões; o pós-processamento abrange o resultado obtido na mineração de dados e a sua interpretação.

FIGURA 2 – Processo de descoberta de conhecimento (KDD)



Fonte: HAN;KAMBER, 2006, p.12.

O processo de KDD não se resume apenas a mineração de dados, mas o uso destas técnicas é elementar para geração do conhecimento a partir de bases de dados.

Para Correa (2015) para a fase da pré-seleção foram se utilizados o uso do software R e o algoritmo de cluster no processo de mineração de dados, métodos estatísticos para a análise integrada de preços de comercialização do agronegócio.

Após o processamento de tratamento estatístico, as aplicações dos métodos de redução de dimensionalidade e discretização de dados foram implementadas com rotinas no software R (CORREA, 2015).

Conforme Fonseca (2014) uma análise exploratória para nortear melhorias no Sistema educacional brasileiro com mineração de dados foi utilizada uma implementação do algoritmo Naïve Bayes disponibilizada no software Weka (ver HALL et al., 2009), que é uma ferramenta gratuita de referência reconhecida pela comunidade cientifica, composta de uma coleção de algoritmos de mineração de dados e uma série de funcionalidades que auxiliam na etapa de pré-processamento.

De acordo com Sousa (2014) o desenvolvimento de modelos para analisar a capacidade dos associados de uma cooperativa de crédito de saldar os seus compromissos. Para tal, foram utilizadas técnicas de Mineração de Dados (Data Mining).

Sousa (2014) Optou-se, em sua pesquisa, por utilizar a implementação do software WEKA do algoritmo da árvore de decisão. Segundo Goldschmidt e Passos (2005), esta árvore é amplamente utilizada e aceita. Tomou-se um modelo gerado pela técnica Árvore de Decisão para exemplificar as regras e matriz de confusão.

A mineração de dados por regras de associação é uma das técnicas mais utilizada sendo a tarefa de associação realizada por meio de algoritmos que geram regras que caracterizam o quanto a presença de um conjunto de itens, nos registros de uma base de dados, implica na presença de outro conjunto distinto de itens, nos mesmos registros. (BRANQUINHO, 2015).

Há algumas ferramentas de mineração de dados que auxiliam o processo de KDD. O uso do software Linguagem R devido licença livre (open source) e facilidade de uso em pesquisas acadêmicas a área da saúde. (BRANQUINHO, 2015).

Segundo Alves (2019) o uso de mineração de dados vem sendo usadas nas mais diversas áreas incluindo a área medica a fim de classificar pacientes de acordo com quatro tipos de doenças mentais. Para alcançar o objetivo proposto os seguintes algoritmos foram utilizados.

O algoritmo do vizinho mais próximo k-NN, do inglês Nearest Neighbor (NN), tem como objetivo caracterizar (rotular ou achar uma classe) uma amostra, a partir de uma ou mais medidas (normalmente utiliza-se medidas de dissimilaridade e de similaridade) de uma amostra, com base nas informações de um ou mais indivíduos previamente rotulados no espaço de busca. (ALVES, 2019).

Para Anghinoni (2018) o algoritmo k-Nearest Neighbor (k-nn) é de fácil implementação e requer pouco tempo computacional para a fase de treinamento, porém mais tempo para a fase de classificação. O nível de generalização do algoritmo pode ser ajustado através do valor de k (número de vizinhos) e, de acordo com o problema estudado, diversas medidas de distâncias podem ser testadas, como a Manhattan, Euclideana, Camberra, Minkowsky, entre outras.

Segundo Alves (2019) O algoritmo Naive Bayes é muito utilizado na prática por resultar em classificação com elevada performance, sendo robusto a atributos irrelevantes.

Para Anghinoni (2018) os algoritmos de classificação como o Naive Bayes, o Multilayer Perceptron, as Árvores de decisão e assim por diante também são conhecidos como

algoritmos de indução, pois a previsão de um novo dado é induzida com base na classificação dos dados conhecidos e podem ser selecionados para resolver previsões de séries temporais.

Considerando este cenário, novas técnicas de mineração de dados têm sido constantemente desenvolvidas para lidar com esta situação. O estudo de séries temporais baseado em suas características topológicas, observadas em uma rede complexa gerada com os dados da série temporal. O modelo proposto apresenta algumas vantagens em relação a métodos tradicionais, como o número adaptativo de classes, com força mensurável, e uma melhor absorção de ruídos (ANGHINONI, 2018).

As árvores de decisão são representações simples do conhecimento e têm sido amplamente aplicadas como, por exemplo, em diagnósticos médicos, análise de risco em créditos e outros exemplos. (BARROS; CARVALHO; FREITAS, 2015).

Os algoritmos de árvores de decisões consistem em um método não paramétrico que pode ser utilizado tanto para problemas de classificação como de regressão. São estruturas hierárquicas do aprendizado supervisionado por onde o espaço de entrada é divido em regiões locais de modo a prever a variável dependente (ALVES, 2019).

A descoberta de conhecimento, por meio de análise exploratória de dados e técnicas de mineração de dados (análise descritiva) em repositórios de processos de software para a identificação de benefícios de abordagens distintas de construção (BASTOS JUNIOR, 2016).

Bastos Junior (2016) utilizou a técnica de agrupamento clustering utilizando o algoritmo K-Means (HALL et al., 2009) e o software Weka, buscando identificar grupos de atividades em ambas as etapas de desenvolvimento do software (PRAXIS e Scrum) que reunissem observações de uma forma geral, agrupadas por semelhança.

Na área de saúde, de forma específica, há certa precariedade nas informações que estão sendo geradas para pacientes em tratamento, ou até mesmo para consultas rotineiras. O que se sabe muitas vezes é o diagnóstico do paciente naquele período e não dados históricos desse paciente (GREGORY, 2016).

Os dados contemplados no sistema de saúde atualmente estão distribuídos de forma incoerente, pois muitos deles não estão preenchidos ou estão informados de forma errada. Dificultando assim a manipulação dessas informações. Pensando então em uma estratégia para solucionar esse problema, observou-se a possibilidade de trabalhar com a mineração desses dados (Data Mining) para a geração de resultados concretos e de muita valia para ações futuras, de forma proativa ao tratamento (GREGORY, 2016).

Para Gregory (2016) o uso de mineração de dados na área da saúde foi utilizado à metodologia de Estudo de Caso. Este projeto se trata de caso um estudo específico com dados extraídos dos usuários da área de promoção à saúde de uma empresa de saúde.

Conforme Riberio (2017) os algoritmos de classificação é a tarefa de mapear um conjunto de atributos de entrada em um rótulo, através de uma função-alvo aprendida pelo algoritmo. Em modelos de classificação, o rótulo precisa ser discreto, o que os diferencia de modelos de regressão.

Encontrar a árvore de decisão ótima é um problema computacionalmente exponencial, devido ao número de caminhos possíveis a serem criados do nó raiz aos nós folha. Porém,

as heurísticas existentes para criação de árvores têm baixo custo computacional tanto para construção quanto para teste, e são fáceis de serem interpretadas, especialmente quando são menores. Além disso, árvores de decisão são não paramétricas (não requerem informação sobre a distribuição de probabilidade dos dados), resistentes a ruído e atributos redundantes não prejudicam sua acurácia (WITTEN et al., 2016).

Para Anjos (2018) a análise de agrupamento de dados é uma tarefa fundamental em mineração de dados. Ela tem por objetivo encontrar um conjunto finito de categorias que evidencie as relações entre os objetos (registros, instâncias, observações, exemplos) de um conjunto de dados de interesse. Os algoritmos de agrupamento podem ser divididos em particionais e hierárquicos.

Uma das vantagens dos algoritmos hierárquicos é conseguir representar agrupamentos em diferentes níveis de granularidade e ainda serem capazes de produzir partições planas como aquelas produzidas pelos algoritmos particionais, mas para isso é necessário que seja realizado um corte (por exemplo horizontal) sobre o dendrograma ou hierarquia dos grupos (ANJOS, 2018).

A escolha de como realizar esse corte é um problema clássico que vem sendo investigado há décadas. Mais recentemente, este problema tem ganhado especial importância no contexto de algoritmos hierárquicos baseados em densidade, pois somente estratégias mais sofisticadas de corte, em particular cortes não horizontais denominados cortes locais (ao invés de globais) conseguem selecionar grupos de densidades diferentes para compor a solução final (ANJOS, 2018).

Entre as principais vantagens dos algoritmos baseados em densidade está sua robustez à interferência de dados anômalos, que são detectados e deixados de fora da partição final, rotulados como ruído, além da capacidade de detectar clusters de formas arbitrárias (ANJOS, 2018).

Para Anjos (2018) o algoritmo HDBSCAN representa conceitualmente o estado-da-arte em agrupamento de dados e tem se tornado a escolha de facto em termos de algoritmos baseados em densidade na literatura, como implementações muito eficientes já disponíveis nas principais bibliotecas de linguagem como R (disponível, por exemplo, no pacote R “dbscan”) e Python (disponível, por exemplo, na biblioteca “scikit-learn”). Na pratica, o algoritmo já foi mostrado experimentalmente ser superior a vários algoritmos bem estabelecidos.

Segundo Ferreti (2015) A massificação dos estudos da medicina translacional permite aos pesquisadores que usufruam de fontes de dados das mais diversas áreas. Uma área de suma importância é a bioinformática, que agrega alta capacidade de processamento computacional disponível atualmente, com a infindável quantidade de dados gerada por métodos de sequenciamento de ultima geração, para entregar aos pesquisadores uma quantidade rica de dados para serem analisados.

Apesar da disponibilidade desses dados, a expertise necessária para analisa-los dificulta que profissionais com pouco conhecimento em bioinformática, estatística e ciência da computação possam realizar pesquisas e analises com estes dados (FERRETI, 2015).

Rodrigues (2016) utiliza diversos algoritmos de mineração de dados com o objetivo de obter o melhor ajuste/configuração no processo de ensino, a fim de tornar mais precisa a avaliação de desempenho do aluno.

Os algoritmos utilizados são o algoritmo Gray Relational Analysis (GRA), que busca identificar os principais fatores de aprendizagem que afetam a nota final do aluno; o algoritmo de agrupamento K-Means usado para determinar logicamente a função de pertinência (grau de participação), utilizada pelo algoritmo de regras de associação Fuzzy, para avaliar o desempenho do aluno; e, por fim, o algoritmo de inferência Fuzzy usado para classificar o desempenho da aprendizagem do aluno (RODRIGES, 2016).

3. Metodologia

A metodologia utilizada para a realização da revisão foi através de levantamento bibliográfico. Foram utilizados dissertações de mestrado, teses de doutorado e feitas pesquisas bibliográficas nas bases: biblioteca digital USP, Biblioteca digital UFMG e Sistema PUC MINAS, utilizando mineração de dados nas áreas: saúde, educação, vendas, agro economia, cooperativa de crédito, ciência da computação e bioinformática.

4. Resultados e análise dos resultados

Identificar o mercado e se posicionar corretamente na oferta de produtos e serviços constituem um grande desafio para os tomadores de decisão. A partir do entendimento das necessidades dos clientes é possível promover ações para direcionar as estratégias de mineração de dados.

Há algumas ferramentas de mineração de dados que auxiliam o processo de KDD.

A partir da tabela 1 foi realizada uma classificação dos algoritmos mais utilizados em relação à área pesquisada.

Tabela 1 – Área e Algoritmos.

Área	**Algoritmo**
Saúde	Apriori, Naive Bayes, k-NN, Arvores de decisão,DBSCAN
Educação	Naive Bayes, ID3(Neuro-FDT), (HMM), (SVM), Fuzzy, Árvores de decisão
Vendas	OneR, ID3, PRISM
Agroeconomia	Algoritmo de cluster
Cooperativa de crédito	C4.5, CART, QUEST, CHAID, Multilayer Perceptrom
Ciência da computação	K-Means, DBSCAN
Computação Aplicada	SVM, Naive Bayes, k-NN
Bioinformática	Apriori

Fonte: Dados da pesquisa.

A tabela 2 mostra o algoritmo mais utilizado independente da área.

Tabela 2 – Algoritmos mais utilizados independente da área.

Algoritmos mais utilizados
Apriori
Naive Bayes
Árvores de decisão
K-nn
K-Means
ID3
Algoritmo de cluster
DBSCAN
OneR
Fuzzy
PRISM
C4.5
SVM
CART
QUEST
CHAID
Multilayer Perceptrom
HMM
SVM

Fonte: Dados da pesquisa.

A tabela 3 mostra os softwares e linguagens de programação de acordo com os autores da pesquisa.

Tabela 3 – Softwares, autores e linguagens de programação.

Software	Autor	Linguagem
RStudio	Branquinho (2015)	Linguagem R
Weka	Fonseca (2014)	Linguagem JAVA/Python
Software R	Alves (2019)	Linguagem R
Weka	Gregory e Pretto (2016)	Linguagem JAVA
Real Statistics Resource Pack	Ribeiro (2017)	Excel, C
Conexp	Rodrigues (2016)	Linguagem JAVA
Weka e MATLAB	Gonzales (2014)	Linguagem JAVA
Software R	Correa (2015)	Linguagem R
Weka	Souza (2014)	Linguagem JAVA
Weka e Software R	Júnior (2016)	Linguagem JAVA e R
Python	Anghinoni (2018)	Python
Software R	Anjos (2018)	Linguagem R
Python	Ferreti (2015)	Python

Fonte: Dados da pesquisa.

A tabela 4 mostra os softwares mais utilizados de acordo com os autores da pesquisa.

Tabela 4 – Softwares mais utilizados.

Software mais utilizado
Software R / RStudio
Weka
Python
Conexp
Real Statistics Resource Pack
MATLAB

Fonte: Dados da pesquisa.

A tabela 5 mostra as linguagens mais utilizadas de acordo com os autores da pesquisa.

Tabela 5 – linguagem mais utilizada.

Linguagem mais utilizada
Linguagem JAVA
Linguagem R
Python
Excel / C

Fonte: Dados da pesquisa.

Com os resultados obtidos nas tabelas demonstraram-se, como as diversas tecnologias ligadas ao processo de KDD e organização do conhecimento podem apoiar as tomadas de decisões, de forma a entender o comportamento nas áreas da saúde, vendas, educação, agro economia, cooperativa de crédito, ciência da computação e bioinformática.

De acordo com os experimentos, foi possível obter resultados para responder ao objetivo do trabalho que é identificar o uso de algoritmos, softwares e linguagens aplicadas a mineração de dados.

5. Conclusão

Diante do estudo sobre publicações relacionadas à mineração de dados, foram identificadas ações estratégicas para transformar dados puros em informações úteis a potenciais áreas de pesquisa por meio de revisões de literaturas, por meio de livros, teses e dissertações utilizando-se as bases: biblioteca digital USP, Biblioteca digital UFMG, Sistema PUC MINAS a fazer uso de mineração de dados, analisando-os por meio de metodologias, ferramentas e algoritmos que propiciassem a extração de informações que oferecessem rotas para a definição de ações e contribuições voltadas nas áreas da saúde, vendas, educação, agro economia, cooperativa de crédito, ciência da computação e bioinformática. Propõe-se estudar e comparar os algoritmos mais usados, identificar por meio de um questionário o motivo pela escolha do algoritmo.

6. Referências

ALVES, Caroline Lourenço. **Diagnóstico de doenças mentais baseado em mineração de dados e redes complexas.**2019. 137 f. Dissertação (Mestrado) - Curso de Ciência da Computação e Matemática Computacional, Instituto de Ciências Matemáticas e de Computação, São Carlos, 2019.

ALVES, Ricardo Brito. **REDUÇÃO DA INFLUÊNCIA DE CONFUNDIDORES EM MODELOS DE APRENDIZADO DE MÁQUINA.** 2018. 78 f. Dissertação (Mestrado) - Curso de Pós-graduação em Engenharia Elétrica, PontifÍcia Universidade CatÓlica de Minas Gerais, Belo Horizonte, 2018.

ANGHINONI, Leandro. **Classificação e previsão de séries temporais através de redes complexas.** 2018. 87 f. Dissertação (Mestrado) - Curso de Computação Aplicada, Universidade de São Paulo, Ribeirão Preto, 2018.

ANJOS, Francisco de Assis Rodrigues dos. **Seleção de grupos a partir de hierarquias:** uma modelagem baseada em grafos. 2018. 86 f. Dissertação (Mestrado) -

Curso de Ciência de Computação e Matemática Computacional, Universidade de São Paulo, São Carlos, 2018.

BASTOS JÚNIOR, José Luiz Gonçalves. **MINERAÇÃO EM REPOSITÓRIOS DE PROCESSOS DE SOFTWARE PARA IDENTIFICAÇÃO DE BENEFÍCIOS DE ABORDAGENS DISTINTAS DE CONSTRUÇÃO.** 2016. 133 f. Dissertação (Mestrado) - Curso de Pós Graduação, Universidade Católica de Minas Gerais, Belo Horizonte, 2016.

BARROS, R. C.; CARVALHO, A. C. de; FREITAS, A. A. **Automatic design of decision-tree induction algorithms.** [S.1]: Springer, 2015. Citado na página 53.

BRANQUINHO, LucÉlia Pinto. **MODELO PARA SUPORTE À DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS (KDD):** APLICAÇÃO EM ESTRATÉGIAS NO MERCADO DE MEDICINA DIAGNÓSTICA. 2015. 122 f. Dissertação (Mestrado) - Curso de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

CARVALHO, Luís A. V. de. DATAMINING: **A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.** 2a Edição. São Paulo-SP: Érica, 2002.

CORREA, Fernando Elias. **Modelo integrado de mineração de dados para análise de séries temporais de preços de indicadores agroeconômicos.** 2015. 99 f. Tese (Doutorado) - Curso de Sistemas Digitais, Universidade de São Paulo, São Paulo, 2015.

FERRETI, Yuri. **Ferramenta Computacional para Análise Integrada de Dados Clínicos e Biomoleculares.** 2015. 60 f. Dissertação (Mestrado) - Curso de Bioinformática, Universidade de São Paulo, Ribeirão Preto, 2015.

FONSECA, S. O. **Utilização de modelos de classificação para mineração de dados relacionados à aprendizagem de matemática e ao perfil de professores do ensino fundamental.** 2014. 121 f. Dissertação (Mestrado em Modelagem Computacional) – Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2014.

GONZALES, José Artur Quilici; ZAMPIROLLI, Francisco de Assis. **Sistemas inteligentes e mineração de dados.** São Paulo: Triunfal Gráfica e Editora, 2014.

GOLDSCHMIDT, R., & Passos, E. (2005). Data mining: Um guia prático. Rio de Janeiro: Elsevier.

GREGORY, Guilherme; PRETTO, Fabrício. MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO EM DADOS DE PROMOÇÃO À SAÚDE. **Revista Destaques Acadêmicos**, [s.l.], v. 8, n. 4, p.51-65, 29 dez. 2016. Editora Univates. http://dx.doi.org/10.22410/issn.2176-3070.v8i4a2016.1234.

HAN, J.; KAMBER, M. **Datamining:** concepts and techniques. 2006.

QUONIAM, L. **INTELIGÊNCIA OBTIDA PELA APLICAÇÃO DE DATA MINING EM BASE DE TESES FRANCESAS SOBRE O BRASIL**. Ciência da Informação, Brasília, v.30, n.2, maio/ago. 2001.Disponível em: . Acesso em: 24 mar. 2019.

REZENDE, Solange Oliveira. **Sistemas inteligentes:** fundamentos e aplicações. Barueri: Manole Ltda, 2005.

RODRIGUES, Marcos Wander. **MINERAÇÃO DE DADOS EDUCACIONAIS:** CENÁRIO DE DUAS DÉCADAS. 2016. 177 f. Dissertação (Mestrado) - Curso de Pós-graduação em Informática, Pontifícia Universidade CatÓlica de Minas Gerais, Belo Horizonte, 2016.

SOUSA, Marcos de Moraes; FIGUEIREDO, Reginaldo Santana. CREDIT ANALYSIS USING DATA MINING: APPLICATION IN THE CASE OF A CREDIT UNION. **Journal Of Information Systems And Technology Management**, [s.l.], v. 11, n. 2, p.379-396, 21 ago. 2014. TECSI. http://dx.doi.org/10.4301/s1807-17752014000200009.

WITTEN, I. H. et al. Data Mining: **Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016.