DOI: 10.5748/16CONTECSI/KMG-6259

DISCOVERY OF KNOWLEDGE IN A DATABASE: APPLICATION OF THE PROCESS ON A CAGED BASIS

DESCOBERTA DE CONHECIMENTO EM UMA BASE DE DADOS: APLICAÇÃO DO PROCESSO EM UMA BASE DO CAGED

Yonara Costa Magalhães – Universidade Ceuma - yonara.magalhaes@ceuma.br
Will Ribamar Mendes Almeida - Universidade Ceuma - will75@gmail.com
Rodrigo Costa Matos - Universidade Ceuma - rodrigocostamatos2211@gmail.com
Emanoel Claudino Silva - Universidade Ceuma - emanoelclaudino@yahoo.com.br

Resumo. *As organizações possuem um fluxo intenso de atividades que geram uma grande quantidade de dados, os quais são registrados e armazenados para serem usados posteriormente. A utilização satisfatória desses registros pode contribuir em larga escala para essas empresas ou indivíduos que queiram extrair conhecimento e aplica-lo. Para isso foram desenvolvidos sistemas capazes de manipular esses dados adequadamente fazendo parte de um processo de extrair conhecimento chamado de KDD. O objetivo deste trabalho é descrever e analisar um processo de descoberta de conhecimento, apresentando a infraestrutura e os processos envolvidos, como a limpeza dos dados, a transformação desses dados em algo que possa ser compreendido, a mineração usando a técnica de clusterização, entre outros. Com estes resultados será possível demonstrar como a descoberta de conhecimento é um processo importante no tratamento de dados para gerar informações que subsidiam a tomada de decisão em diferentes áreas.*

Palavras-chave: Mineração de dados, KDD, Data Warehouse, Clusterização.

Abstract. *Organizations have an intense flow of activity that generates a large amount of data, which is recorded and stored for later use. The satisfactory use of these records can contribute to a large scale for those companies or individuals who want to extract knowledge and apply it. For this, systems were developed capable of manipulating this data adequately, being part of a process of extracting knowledge called KDD. The objective of this work is to describe and to analyze a process of knowledge discovery, presenting the infrastructure and the involved processes, like the cleaning data, the transformation of this data into something that can be understood, mining using the clustering technique, among others. With these results it will be possible to demonstrate how the discovery of knowledge is an import process in data processing to generate information that decision-making in different areas.*

Keywords: Data mining, KDD, Data Warehouse, Clustering.

1. INTRODUÇÃO

Na era da tecnologia o aumento exponencial do volume de dados é um fato evidente e as grandes empresas e instituições se constituem como grandes geradoras de dados. Essas organizações possuem um fluxo intenso de atividades que geram uma grande quantidade de registros, os quais são armazenados para serem usados posteriormente. A utilização satisfatória desses registros pode contribuir em larga escala para essas empresas ou instituições, assim como para qualquer indivíduo que queira, a partir de uma base de dados, extrair conhecimento e aplicá-lo.

Uma informação se constitui da relação que se faz entre dados e um contexto, onde os dados isolados necessitam ser relacionados e contextualizados. Para melhor compreensão, apresenta-se como exemplo um conjunto de dados sobre o mercado empregatício, utilizando dados de admissão e demissão. Esses dados quando apresentados de forma isolada mostram fatos que foram registrados, no entanto, se eles forem relacionados e contextualizados, podem elucidar questões sobre esse mercado de trabalho, evidenciar potencialidades e dificuldades e/ou confirmar um conjunto de hipóteses que se deseja validar. Desta forma, a utilização de informações direcionadas para o entendimento de uma situação é que gera conhecimento.

Para operacionalizar a extração de conhecimento a partir de um conjunto de dados foram desenvolvidos sistemas capazes de manipular esses dados adequadamente em um processo chamado *KDD (Knowledge-discovery in databases)*. Para Macedo (2010), o KDD se refere a um procedimento que busca informações e padrões em dados, com objetivo de extrair conhecimento a partir deles.

O processo de descoberta de conhecimento em base de dados tem sido muito utilizado tanto no meio acadêmico quanto no empresarial, como forma de obter conhecimento a partir de um banco de dados. A contextualização desses dados gera a compreensão de algumas situações, mas para isso esses dados passam antes por várias etapas, dentre elas a mineração de dados que consiste inicialmente na seleção dos dados relevantes para um determinado objetivo.

Neste trabalho, para demostrar a importância da descoberta de conhecimento e da mineração de dados escolheu-se, tendo em vista o atual cenário econômico do país, utilizar a base de dados do Cadastro Geral de Empregados e Desempregados (CAGED) do Ministério do Trabalho e Emprego (MTE) que armazena informações sobre a dinâmica do mercado de trabalho, admissão e demissão de empregados, dos salários, dentre outras informações, que podem contribuir para a compreensão de como andam os indicadores de trabalho neste panorama.

Para isso, nesta pesquisa, esse processo de descoberta foi direcionado aos registros de admissão e demissão aonde os resultados desta análise permitem comunicar para os gestores de empresas e demais públicos, como se encontra o perfil socioeconômico atualmente, quais os parâmetros têm pesado mais na faixa salarial e quais as diferenças sociais presentes. Os resultados desta pesquisa geram insumos que podem auxiliar na tomada de decisões pelos gestores e ao mesmo tempo podem servir para subsidiar o delineamento de políticas governamentais

relacionadas a esta área.

Assim, a intencionalidade neste trabalho é apresentar formas operacionais de gerar conhecimento e que por meio delas podem ser gerados resultados que servem de base para pontuar possíveis causas e apontar onde as mudanças podem ser feitas, bem como apresentar informações otimizadas. Quando aplicadas no contexto do CAGED permitirão compreender os aspectos relacionados a esta situação.

A estrutura deste trabalho encontra-se organizada da seguinte forma: na segunda seção deste trabalho serão expostos os embasamentos teóricos necessários para sua execução. Nesse capítulo tem uma explanação sobre o processo de KDD além de um estudo direcionado a mineração de dados e seu processo de clusterização. Está exposto também o conceito de *Data Warehouse* e ferramentas que auxiliam na mineração de dados. Na seção 3 será detalhadamente explicado as etapas enfrentadas para a obtenção dos resultados além de uma exposição dos resultados alcançados com a pesquisa. E na seção 4 serão apresentadas as conclusões sobre a utilização do KDD e da mineração de dados para gerar conhecimento.

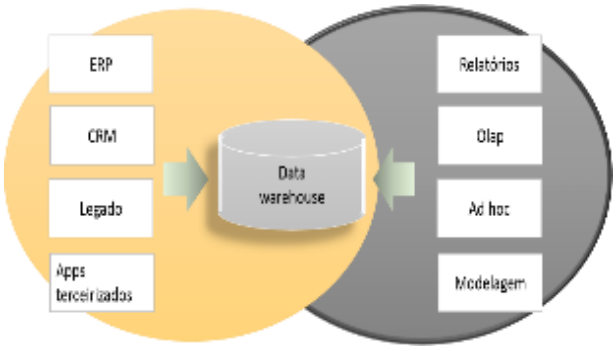
2. FUNDAMENTAÇÃO TEÓRICA

O avanço da tecnologia trouxe consigo o crescimento exponencial de dados e, em decorrência disto, a necessidade e a dificuldade de gerar conhecimento a partir deles. Por essa razão, se torna necessário a utilização de processos e técnicas cada vez mais otimizados com o objetivo de coletar e selecionar dados de forma a gerar informações relevantes em tempo hábil, ou seja, gerar informações contextualizadas a partir das relações estabelecidas os dados, tornando assim as informações úteis para quem as interpreta.

2.1 *DATA WAREHOUSE* (DW)

O termo DATA WAREHOUSE (DW) e é definido por Inmon (2005) como um sendo conjunto de dados direcionado por conteúdo, integrado, que não pode ser alterado, que varia com o tempo e tem o objetivo de apoiar as decisões da gerência, como esta ilustrada na figura abaixo:

Figura 1 – *DATA WAREHOUSE*



Fonte: Baseado em (BONEL, 2015)

Dalfovo e Tamborlin (2010) complementam o conceito de DW ao defini-lo como um banco de dados especialista que age integrando e gerenciando os dados a partir de uma base de dados relacional, sendo elaborado para que os dados sejam consultados de forma rápida e sem as limitações impostas pelas tradicionais tabelas relacionais. Para Bonel (2015), DW funciona como uma pilha de dados organizadas de forma multidimensional para otimizar as consultas como se fosse um container de dados originado de várias fontes, formando assim uma visão mais detalhada do negócio.

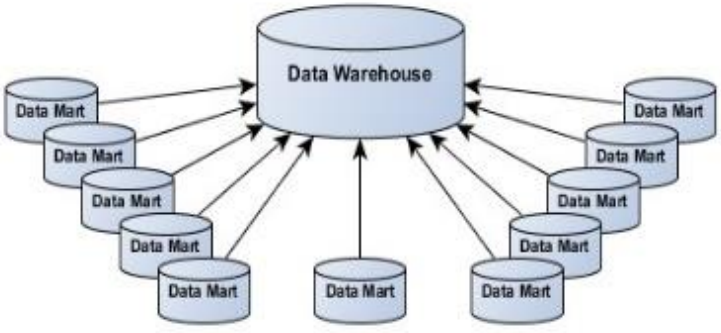
Uma consulta em modelos relacionais muitas vezes se torna inviável tanto pela demora na execução de queries quanto pela falta de conhecimento do usuário para com as tabelas. Já nos modelos multidimensionais esse problema não existe, pois, os mesmos dados estão expostos em uma só tabela, fornecendo apenas uma origem para os dados desejados, o que facilita na tomada de decisão em relação a esses dados.

Para Han (et al., 2011), existem três principais processos utilizados juntamente com DW: processamento de informações (consultas, análises e relatórios), processamento analítico (OLAP e suas ferramentas) e processamento de mineração de dados (busca padrões entre os dados através de técnicas específicas para isso).

Apesar de controvérsias entre autores o DATA MART pode ser considerado como subconjunto dentro de um DATA WAREHOUSE e também consiste em depósitos de dados. Segundo Inmon (2005) esses depósitos se constituem como partes menores, coleções de assuntos voltados para áreas específicas que se baseiam nas necessidades ali existentes para assim ajudar no apoio a tomada de decisões.

A estrutura do *DATA MART* está ilustrada na Figura 2, a seguir.

Figura 2 – *Data Mart*



Fonte: O autor

Dessa forma, o DATA MART se restringe a tratar de problemas locais, diferenciando- se do DATA WAREHOUSE no que tange ao volume e complexidade e se aproximando quanto ao uso da tecnologia (ver Quadro 1, a seguir).

Quadro 1 – Propriedades obtidas após o processamento

DATA MART	*DATA WAREHOUSE*
Departamental	Corporativo
Nível tático	Nível estratégico
Otimizado para acesso e análise	Gerenciamento de grande volume de dados
Poucas fontes de dados	Muitas fontes de dados
Pequenos estágios de implementação	Múltiplos estágios de implementação

Fonte: adaptado de Inmon (2005)

2.2 Relação entre *DATA WAREHOUSE* e *DATA MINING*

De acordo com Keen (1991) os Sistemas de Apoio a Decisão (SADs) ou Decision Support System (DSS) são elaborados com o objetivo de oferecer suporte aos profissionais dentro das empresas, elevando assim o progresso e a produtividade. O autor ainda aponta que esse suporte não pretende substituir as análises do profissional responsável pelas decisões, pelo contrário, como os SADs não torna automático esse processo ele na verdade auxilia nas decisões sem impor análises.

Para dar suporte aos SADs e as ferramentas que o colocam em prática surge o DW, que armazena uma grande quantidade de dados em um mesmo repositório, facilitando a busca por esses dados, tarefa essa que pode ser executada por uma das ferramentas desses sistemas que é o data mining.

O data mining tem a função de encontrar padrões em dados para extrair conhecimento deles e seu processo pode ser mais produtivo e rápido se ele contar com uma base de dados DW que fornece dados limpos e consistentes. Sanches (2003) confirma isso quando diz que o DW fornece dados integrados, detalhados, resumidos, dados históricos e metadados, os quais colaboram com a performance do processo de data mining.

O processo de mineração requer que anteriormente haja um condicionamento e refinamento dos dados e isso requer tempo. Dessa forma os dados integrados facilitam a visualização e poupam mais tempo, permitindo que o agente minerador esteja direcionado para os algoritmos de mineração. Já os dados detalhados corroboram para a possível exigência de uma análise mais minuciosa onde são necessárias buscas mais aprofundadas e que também levam tempo. No entanto, para que seja evitado um processamento de dados repetitivo e extenso pode-se contar com os dados resumidos que trazem a garantia de que uma prévia foi realizada, tornando desnecessário diversos processamentos.

O objetivo de identificar tendências e padrões de comportamento nos dados pode ser comprometida se forem utilizadas apenas informações atuais. Assim, os dados históricos fornecidos pelo DW permitem que possam ser analisadas uma extensa quantidade de informações implicitamente armazenadas, colaborando para

a compreensão dos negócios e das circunstâncias em que eles se apresentam. O processo de mineração de dados não está subordinado ao DW, no entanto quando executados em conjunto alcança maior desempenho em sua tarefa.

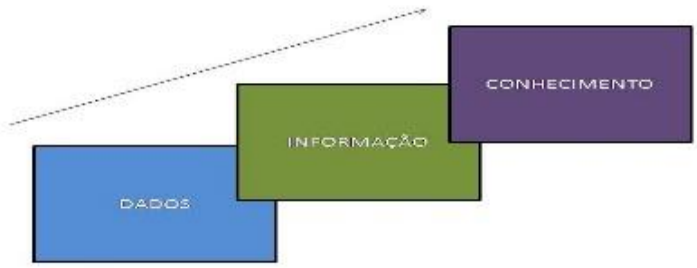
2.3 KDD

Segundo Quilici e Zampirolli (2014) devido ao crescente aumento do volume de dados foi necessária a criação de sistemas inteligentes capazes de manipular satisfatoriamente esses dados por meio de ferramentas que conseguissem administrar o conhecimento obtido e aperfeiçoar sua execução baseado em sua própria experiência, otimizando assim esse processo.

O termo KDD (Knowledge-discovery in databases) se refere a descoberta de conhecimento em banco de dados e é utilizado para nomear o processo geral de busca de informações e padrões úteis em dados. No entanto, anteriormente a esta nomenclatura foram utilizadas outras, como: mineração de dados, extração de conhecimento, descoberta de informação e processamento de padrões em dados etc. Apenas em 1989 o termo foi utilizado para se referir especificamente a esse processo. Na época também foi pontuado que a mineração de dados é parte integrante do KDD, e por mais que ainda seja confundida com este, desempenham funções distintas que colaboram para que ambos alcancem o mesmo objetivo final (FAYYAD et al., 1996).

De acordo com Fayyad (et al., 1996), KDD consiste em um processo significativo de identificação de padrões em dados que abrange várias etapas buscando padrões anteriormente desconhecidos, que sejam válidos, que possam ser úteis, que podem ser compreendidos e tudo isso a partir de um amplo conjunto de dados. Baseado nisso, a descoberta de conhecimento em banco de dados visa analisar uma grande quantidade de dados, a fim de converter todas essas informações em conhecimento como ilustrado na Figura 3, a seguir.

Figura 3 – Dados, Informação e Conhecimento



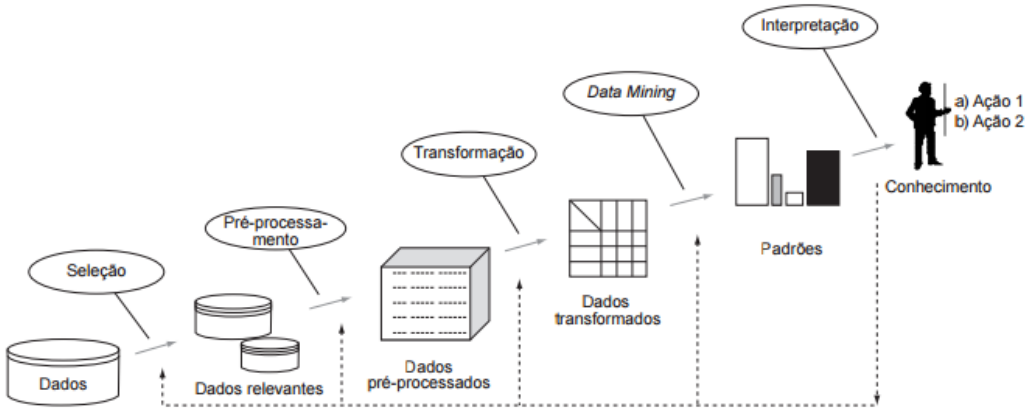
Fonte: O autor

Para Quilici e Zampiroli (2014) o termo dado é definido como um fato, uma quantidade ou qualidade registrada. Já a informação se refere a dados que apresentam um contexto, que possuem um significado. E o termo conhecimento se forma a partir de informações úteis para a compreensão de uma situação. Assim, os dados são fatos que quando contextualizados e associados a outros dados assumem significados, transformando-se em informações que geram conhecimento. Partindo disso, compreende-se que o KDD é um processo geral de descoberta de conhecimentos proveniente de dados, onde são analisados, relacionados e contextualizados produzindo conhecimentos.

O processo de KDD compreende o percurso que o dado faz desde a triagem até o final, quando ele se transforma em informação. Caracteriza-se como um processo iterativo, pois se trata de uma relação de dependência entre os resultados e interativo uma vez que possibilita a intervenção e o controle do andamento das atividades, permitindo uma comunicação, uma ação mútua.

O processo de KDD, ilustrado acima, requer uma definição dos objetivos desejados, onde é necessário um conhecimento aprofundado do domínio abordado, por isso geralmente no início do processo há um estudo bem detalhado feito por especialistas de um determinado domínio, definindo assim a viabilidade da aplicação, ou seja, verifica-se se os dados provenientes do KDD serão reconhecidos após a geração do conhecimento. Esse processo abrange as seguintes etapas: seleção, pré-processamento e limpeza, transformação, mineração de dados *Data Mining* e interpretação/avaliação (FAYYAD *et al.*, 1996), ver a Figura 4, a seguir.

Figura 4 – Etapas do processo KDD



Fonte: Fayyad et al. (1996)

De acordo com Fayyad et al. (1996), as etapas do KDD estão descritas a seguir:

a) Seleção dos Dados - há uma seleção de dados que irão compor a análise, esta triagem é feita a partir de critérios elencados com o objetivo de formar um conjunto de dados que contenham características e casos que digam respeito aos critérios escolhidos para a análise. Essa etapa tem influência direta sobre a qualidade do resultado final, pois dados mal selecionados geralmente levam a resultados indesejados, em alguns casos se faz necessário criar a sua própria base de dados, originada de várias outras distintas, onde se deve assegurar a homogeneidade e consistência dos dados originais. Os dados podem ser obtidos de fontes internas: originarias da aplicação do próprio domínio como, por exemplo, base de dados de empresas, históricos do sistema e etc. Seu recolhimento e análise são mais ágeis, pois, os dados já estão estruturados. As fontes externas são aquelas que geralmente não estão integradas a nenhum tipo de sistema, exigindo um maior tempo de análise e processamento, pois precisam de um filtro ainda mais restrito para a sua obtenção. Estas fontes quase sempre estão disponíveis em livros, pesquisas e entrevistas.

b) Pré-processamento dos dados - após a obtenção dos dados é necessário um processo chamado de “limpeza dos dados”, onde há uma eliminação de possíveis dados que possam comprometer o resultado da geração de conhecimento, tais como: valores nulos, dados viciados, duplicados, dados inconsistentes entre outros. Segundo Quilici e Zampirolli (2014), em grandes tarefas de KDD que envolvem diferentes fontes de dados a etapa de pré-processamento é determinante para o resultado, pois ela compreende a limpeza e a fusão dos dados e a eliminação de ruídos e dados redundantes. Assim, nesta etapa acontece a transformação dos dados a fim de deixá-los propícios para o uso das técnicas de mineração de dados.

c) Mineração de Dados – nesta etapa ocorre a aplicação de algoritmos de inteligência artificial afim de encontrar padrões entre os dados. É considerada uma das principais no processo de KDD, por esta razão essa etapa está descrita de forma mais detalhada nas próximas seções.

d) Interpretação e Avaliação dos Resultados – nesta etapa os resultados da mineração são avaliados e julgados quanto a sua utilidade e com objetivo de tornar os dados uteis compreensíveis. O processo de *KDD*, muitas das vezes não é um processo único, sendo necessário em muitos casos a execução de forma repetida, até que seja encontrado o resultado desejado (FAYYAD *et al.*, 1996).

e) Utilização do Conhecimento - nesse passo o conhecimento gerado pelos passos anteriores é implantado em aplicações ou são documentados. É necessária ainda uma análise desses resultados afim de evitar ou solucionar conflitos entre conhecimentos já adquiridos e o gerado pelo processo (FAYYAD *et al.*, 1996).

O campo de pesquisa em KDD tem se expandido bastante, e tem sido desenvolvido com forte orientação para responder a necessidades práticas, sociais e econômicas. O KDD utiliza ainda ferramentas muito eficientes para executar a coleta, armazenamento e gerenciamento de um extenso volume de dados. Dessa forma, os dados que evidenciam padrões e tendências podem ser usados para garantir assertividade nas escolhas dentro dos negócios entre outras disposições, trazendo vantagem para quem utiliza esse processo. Tudo isso tem impulsionado o crescimento e desenvolvimento deste campo de pesquisa, que tem sido muito explorado pelas organizações e instituições empresariais, entre outras, pois traz processos automáticos e otimizados para a manipulação satisfatória desses dados (REZENDE, 2003).

2.4 Mineração de Dados

2.4.1 Conceitos e tarefas

Atualmente é evidente que a produção de informações tem acontecido de uma forma muito rápida e em grande quantidade, e isso junto com o advento de tanta tecnologia tem culminado em uma grande quantidade de armazenamento de dados. No entanto, é fato que tantas informações não tem tanta utilidade se não for possível extrair conhecimento delas e isso em um tempo ágil para acompanhar o fluxo com que se propagam.

Diante deste cenário surge um novo desafio e novas buscas para formulações de ferramentas capazes de responder a essa nova demanda existente, daí surgiu o processo de mineração de dados que propõe a descobrir padrões de conhecimento implícito nessa grande e mutável base de dados.

A mineração de dados possui variações em suas definições de acordo com o campo de atuação. Segundo Fayyad (et al., 1996) a mineração de dados tem a finalidade de extrair padrões em base de dados aplicando técnicas e algoritmos específicos, sendo considerada como um processo fundamental da descoberta de conhecimento. Hand (et al., 2001) afirma que mineração de dados tem o objetivo de encontrar relacionamentos e resumir de uma forma útil e compreensível um grande conjunto de dados.

A mineração de dados não deve ser confundida com KDD, pois este último é um processo maior da qual a mineração faz parte. Assim, ela se caracteriza como uma das etapas principais do KDD e que busca descobrir relações novas que ainda não foram identificadas.

Para Quilici e Zampirolli (2014) fica evidente que nesta etapa é onde acontece a descoberta de conhecimento, pois é nesse processo que os dados são combinados para gerar resultados antes não conhecidos. Nessa etapa são usados algoritmos de aprendizagem de máquina para criar modelos e padrões capazes de gerar uma heurística que permita a análise dos resultados.

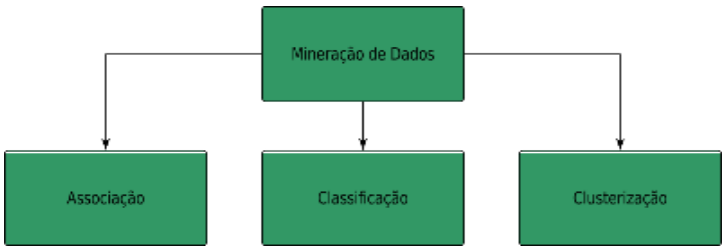
De acordo com Amo (2015) é importante que seja feita a devida diferenciação entre tarefa e técnicas de mineração, pois a tarefa compreende a especificação do que se procura encontrar nos dados, como regularidades ou classe de padrões, e a técnica consiste na identificação de métodos responsáveis pela descoberta desses padrões.

As tarefas de mineração de dados estão se diversificando ainda mais devido ao desenvolvimento de inúmeros sistemas de Mineração de Dados funcionando para domínios diferentes, proporcionando assim uma variabilidade de domínios a serem resolvidos (REZENDE, 2003).

Fayyad (et al., 1996) ainda colabora explicitando que de maneira bem ampla, as tarefas básicas de mineração de dados podem ser distribuídas em duas categorias: a descritiva e a preditiva. As tarefas básicas descritivas abrangem a descoberta de padrões que podem ser interpretados por pessoas que relatem fatos cadastrados na base de dados, apresentando características gerais importantes dos dados. Já as preditivas utilizam variáveis específicas como forma de previsão de valores desconhecidos de outras variáveis relevantes. As tarefas de predição estão relacionadas à tomada de decisão e as descritivas ao suporte nessas decisões.

A mineração de dados possui muitas tarefas entre as quais se destacam a classificação, clusterização e associação, como ilustrado na Figura 5, a seguir.

Figura 5 – Principais tarefas de Mineração de dados



Fonte: O autor

A execução dessas tarefas está intimamente ligada ao tipo de conhecimento que será gerado, onde serão aplicadas a situações especificas visando alcançar os objetivos traçados (QUILICI; ZAMPIROLLI, 2014), ou seja, ela tem a capacidade de abstrair vários tipos de conhecimento, tornando necessário que seja especificado através de critérios, qual o tipo de conhecimento a ser gerado. As principais tarefas de mineração de dados são:

a) Classificação - tem a função de classificar um dado novo a uma classe não conhecida, procurando encontrar relações entre atributos e classes. Para isso, é feito um mapeamento dos dados (ou dos conjuntos deles) de entrada em um número fixo de categorias, onde serão distribuídos de acordo com seus atributos. O objetivo final desse processo é que o algoritmo de classificação seja capaz de predizer a classe de um novo dado (REZENDE, 2003). Logo, trata-se de uma atividade de aprendizagem, pois busca-se por informações indisponíveis através dos dados apresentados, onde nesse contexto são empregados algoritmos preditivos para prever tendências e a partir daí auxiliar na tomada de decisões;

b) Associação - consiste na busca por um padrão de relações de dependência entre dois atributos, procurando descobrir atributos que estão relacionados, ou seja, regras de associação entre eles que acontecem concomitantemente. Assim, por exemplo, em uma análise no histórico de compras de uma loja de adereços buscam-se quais os itens que são comprados concomitantemente, e o resultado disso pode ser aplicado na criação de estratégias de marketing ou mesmo de disposição desses itens, com o objetivo de atender melhor as demandas dos clientes;

c) Clusterização (Segmentação) - tem o objetivo de descrever como alguns dados se agrupam, sendo dessa forma caracterizada como uma tarefa descritiva. Difere da classificação no que tange a predição, pois se trata de um aprendizado não-supervisionado, onde o objetivo é a formação de grupos baseados na aproximação das características similares. Nesse contexto, não é necessária uma definição de registros em classes predeterminadas para a tarefa, pois o critério aqui é a similaridade, seja entre os atributos ou entre as características, e onde o algoritmo se baseia nas alternativas encontradas na base de dados para descobrir classes, tendo a função de agrupar os dados de acordo com as características equivalentes.

d) Estimativa (Regressão) - se assemelha a tarefa de classificação, onde são empregados algoritmos de predição com o objetivo de prever propensões. Essa ação tem como base os dados históricos em uma base de dados, no entanto, diferentemente da tarefa de classificação a tarefa de estimativa trata de registros

distinguidos por valores numéricos ao invés de categóricos, lidando com valores reais.

2.4.2 Técnicas de mineração de dados

Segundo Quilici e Zampirolli (2014) a estatística já possuía algumas técnicas capazes de analisar dados antes mesmo do desenvolvimento da mineração de dados. No entanto, elas eram aplicadas manualmente, o que restringia consideravelmente o seu repertório a uma base pequena de dados. Dessa forma muitos dados ficavam imperceptíveis, passavam despercebidos, e eram apenas dados que não produziam conhecimento.

Com o desenvolvimento de sistemas computacionais foram sendo criadas cada vez mais formas de automatizar a análise de dados. Algumas técnicas respondem melhor que outras a determinados objetivos, dessa forma é necessária a aplicação de vários testes no processo de mineração de dados para então escolher a técnica (ou combinações destas) que mais se adeque ao que se procura. Algumas técnicas de mineração de dados que são geralmente usadas são:

a) Árvores de Decisão - consiste em uma estrutura de árvore, que age com base em um treinamento antecipado que utiliza regras básicas de decisão para segmentar em grupos menores um extenso conjunto de registros, promovendo assim a predição da classificação de um objeto. Tem aplicação muito propicia para as tarefas de classificação e regressão. Na sua estrutura de árvore de decisão, cada nó interno se refere a um atributo pertencente ao banco de dados de amostras, que se diferenciam do atributo-classe, as folhas dizem respeito a valores do atributo-classe, e os ramos individualmente fazem a ligação entre um nó-filho e um nó-pai, onde recebem uma etiquetação com o valor do atributo que está contido no nó-pai, onde um atributo que se apresenta em um nó não pode constar nos nós que descendem deste (AMO, 2015).

b) Raciocínio Baseado em Casos ou MBR (*Memory-Based Reasoning* – raciocínio baseado em memória) – é bastante favorável sua aplicação em tarefas de classificação e segmentação onde conta com os algoritmos BIRCH (Zhang et al, 1996), CLARANS (Chen et al, 1996) e CLIQUE (Agrawal et al, 1998) para efetivá-la. Baseada no método do vizinho mais próximo, ou seja, se baseia em casos já registrados para fazer previsões de novos exemplos, onde esses registros antigos se tornam modelos para predições nesse processo. Para Harrison (1998), se destaca duas habilidades desta técnica: habilidade de execução em qualquer fonte de dados, e capacidade de aprendizagem de novas classes a partir da inserção de novos exemplos no banco de dados. De acordo com Berry e Linoff (*et al.*, 1997), é relevante a pontuação de quatro passos importantes na utilização do Raciocínio Baseado em Casos: seleção do conjunto de dados de treinamento; determinar a função de distância, escolha do número de vizinhos mais próximos e determinação da função de combinação.

c) Algoritmos Genéticos - são utilizados para elaborar suposições sobre relações de dependência entre variáveis, consistindo em uma busca generalizada pela otimização agindo com base nos padrões naturais de evolução (GOEBEL; GRUENWALD, 1999). Ao simular os processos de evolução natural há a execução de operadores de seleção, cruzamento e mutação que objetivam o

desenvolvimento de frequentes gerações de soluções, promovendo a interação entre os atributos de um objeto. Com o tempo o algoritmo evolui de modo que apenas persistirão as soluções com maior poder de previsão (HARRISON, 1998). É também bem adequada para o desenvolvimento de novos sistemas de aprendizagem máquina.

d) Redes Neurais Artificiais – é uma classe especial de sistemas modelados com base no desempenho do cérebro humano e composta por neurônios que são semelhantes aos do cérebro, pode ter suas interconexões alteradas com o uso do algoritmo de aprendizagem, a partir de estímulos ou uma saída que permitam a aprendizagem (GOEBEL; GRUENWALD, 1999). Pode ser aplicada de variadas formas, no entanto tem dados de entrada e modelos complexos (HARRISON, 1998). É bem aplicada as tarefas de classificação, estimativa e segmentação.

2.4.3 Ferramentas de Mineração de Dados

A utilização de ferramentas de mineração de dados tem a capacidade de facilitar consideravelmente o processo geral de KDD, dando suporte a diversas técnicas e tarefas de data mining. Dessa forma, com o avanço e a grande utilização desse processo de descoberta de conhecimento tornou-se necessário a utilização de ferramentas que possam agilizar ainda mais esse processo. Isso tem gerado um aumento na criação deste tipo de ferramentas com o ituito de fornecer aos tomadores de decisões das organizações uma maneira compreensível e amigável de suporte que não são geralmente especialistas em data mining (REZENDE, 2003).

Algumas ferramentas de mineração de dados são apresentadas a seguir:

a) *PolyAnalyst* - de acordo com Megaputer (2015) o *PolyAnalyst* é um software constituído por um conjunto de ferramentas, que tem finalidade de minerar os dados, permitindo que seus usuários tenham disponíveis várias técnicas, como: categorização, clusterização, predição, análise *link, d*escoberta *pattern e* Detecção de anomalias.

b) See5 e C5.0 - São ferramentas desenvolvidas para analisar imensas bases de dados, que usufruem de computadores de até oito núcleos acelerando o processamento e está disponível para sistemas operacionais *WINDOWS/LINUX*. Desta forma, estas ferramentas têm como técnica principal a classificação (RULEQUEST, 2015);

c) KEEL (*Knowledge Extraction based on Evolutionary Learning*) - segundo Keel (2015), *KEEl* é uma ferramenta de software *open source*, que pode ser usado por um diversificado número de tarefas de descoberta de conhecimento. *KEEl* fornece uma interface gráfica simples, baseada em fluxo de dados para projetar experimentos com diferentes conjuntos de dados e algoritmos de inteligência computacional. Contém uma grande variedade de algoritmos como: algoritmo de extração de conhecimentos clássicos, Técnicas (seleção de conjunto de treinamento, seleção de recurso, discretização e Métodos de imputação de valores em falta), inteligência com base algoritmos de aprendizagem computacional, modelos híbridos e metodologias estatísticas pré-processamento para contrastar experiências. Permite executar uma análise completa de inteligência computacional em relação com os já existentes (KEEL, 2015).

d) WEKA (*Waikato Enviroment for Knowledge Analysis*) - consiste em

uma ferramenta livre de mineração de dados que foi desenvolvida na Nova Zelândia, na Universidade de Waikato em 1999 e atualmente tem sido uma das ferramentas mais utilizadas no meio acadêmico. Essa grande utilização se deve a própria estrutura da ferramenta, que possui uma gama de algoritmos que dão suporte a diversas técnicas, dando apoio maior na resolução de problemas (WEKA, 2015). Weka é um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Os algoritmos podem ser aplicados diretamente para um conjunto de dados ou chamado a partir de seu próprio código Java. Weka contém ferramentas para pré-processamento de dados, classificação, regressão, *clustering*, regras de associação e visualização.

2.5 Clusterização *E K-MEANS*

Nesta seção será exposto um breve estudo sobre um algoritmo de mineração de dados K-Means. Esse algoritmo tem como pilar principal, o agrupamento ou segmentação dos dados. O K-Means ou K-Médias almeja gerar uma classificação de informações de acordo com os próprios dados contidos no problema, essa análise é baseada em comparações entre os valores numéricos dos dados. Desta forma o algoritmo fornecerá resultados sem precisar de nenhuma supervisão, ou seja, não se faz necessário o pré-processamento dos dados para que se chegue a um resultado (PICHILIANI, 2008).

Segundo (GOLDSCHMIDT; BEZERRA, 2015), o *K-Means* seleciona pontos de conjunto de dados chamados de k. Esses pontos são denominados sementes. As sementes são os primeiros representantes dos grupos, ou centróides, dos clusters formados. Logo após, calcula-se para cada registro de dado, a distância deste ponto para cada um dos centróides já formados. Este registro será atribuído ao grupo no qual possui a menor distância calculada. O resultado desse processo é que cada registro ficará ligado a apenas um K grupo. Logo em seguida o algoritmo continua em execução, atualizando os centróides e a realocação de cada registro aos centróides mais próximos. Os novos centróides de cada grupo são calculados pelas médias dos pontos já alocados. Esse processo só termina, quando os números de interações preestabelecidas são alcançadas ou quando os centróides param de se modificar (GOLDSCHMIDT, 2013).

O *K-Means* divide todos os objetos em seus respectivos grupos, com objetivo de que a similaridade entre os indivíduos do mesmo grupo seja grande e a similaridade entre objetos de grupos distintos sejam as mínimas. Segundo Pichiliani (2008), o algoritmo pode ser descrito em cinco passos:

1) Iniciar os centróides com valores, onde geralmente são escolhidos os primeiros dados para fornecer seus atributos aos centroides;

2) Gerar uma matriz de distância entre cada ponto dos centróides. Aqui a distância entre cada ponto e os seus centróides são calculados, nessa parte é que ocorrem as maiores quantidades de cálculos, por exemplo, se temos N pontos e K centróides será calculado N x K distâncias neste passo;

3) Alocar cada ponto em seus respectivos centróides, classificando-os de acordo com suas distâncias. Essa classificação funciona assim: o centróide que possui a menor distância desse ponto vai se apropriar dele, o ponto de agora em diante pertencerá a classe que o centróide representa

até que o ponto encontre outro centróide com menor distância;

4) Calcular novos centróides para cada grupo, refinando assim os valores das coordenadas, para cada classe que possui mais de um ponto, um novo valor será calculado usando a média entre todos os pertencentes dessa classe;

5) Repetir até que os centróides não mudem mais ou até que o número de interações seja alcançado. Caso nenhuma dessas condições seja sanada, o algoritmo volta para o passo 2 repetindo iterativamente o refinamento do cálculo das coordenadas dos centróides.

3. RESULTADOS E DISCUSSÕES

Trata-se de um estudo de caso fundamentado na metodologia para a realização de um processo de KDD. A coleta de dados ocorreu durante o ano de 2015 e foram realizados e 5 experimentos na base de dados CAGED - MA. Inicialmente foi realizada uma pesquisa para compreender e analisar de forma mais concisa todos os campos referentes ao CAGED. De acordo com (MTE, 2015) os dados disponíveis contêm as seguintes informações, conforme Quadro 1, abaixo:

Quadro 1 – Campos e descrição do CAGED

CAMPO	DESCRIÇÃO
Admitidos_Desligados	Contratação ou desligamento dos funcionários;
Competência_Declarada	Mês ano em que a movimentação do funcionário foi efetuada;
Município	Município da localização do estabelecimento;
Ano_Declarado	Ano em que a movimentação foi declarada;
CBO_2002_Ocupação	Classificação brasileira de ocupações criada em 2002;
CNAE_1_0_Classe	Classe de Atividade Econômica segundo a classificação CNAE/95, convertida a partir da Classe CNAE 20, disponível a partir de 01/2008;
CNAE_2_0_Classe	Classe de Atividade Econômica, segundo CNAE -versão 2.0
CNAE_2_0_Subclas	Subclasse de Atividade Econômica, segundo classificação CNAE -versão 2.0 a partir de 2006;
Faixa_Empr_Inicio_Jan	Tamanho do estabelecimento em janeiro do ano de referência;
Grau_Instrução	Grau de instrução ou escolaridade do funcionário;
Qtd_Hora_Contrat	Quantidade de horas contratuais por semana;
IBGE_Subsetor	Subsetor Econômico segundo IBGE;
Idade	Idade do trabalhador (quando acumulada representa a soma das idades);
Ind_Aprendiz	Indicador de movimentação referente a contrato de aprendizagem;
Ind_Portador_Defic	Indicador se o empregado/servidor de portador de deficiência habilitado ou beneficiário reabilitado;
Raça_Cor	Raça e Cor do Trabalhador - disponível a partir de 01/2007;
Salario_Mensal	Salário mensal em moeda corrente;
Saldo_Mov	Saldo de movimentação (1 para admissão e -1 para desligamento);
Sexo	Sexo do funcionário;
Tempo_Emprego	Tempo de emprego do trabalhador (quando acumulada representa a soma dos meses);
Tipo_Estab	Tipo de estabelecimento;

CAMPO	DESCRIÇÃO
Tipo_Defic	Tipo de deficiência/Beneficiário habilitado;
Tipo_Mov_Desagregado	Tipo de movimento;
UF:	Estado de localização do estabelecimento.

Fonte: O autor

Após a compreensão e definição foi iniciado um processo de seleção dos dados a fim de eliminar possíveis dados duplicados ou com registros que não ajudariam nos experimentos, como por exemplo dados do mesmo funcionário enviado mais de uma vez ou possíveis promoções ocorridas durante o mês escolhido.

Figura 6 – Processos do KDD

```
@relation weather

@attribute Admitidos_Desligados real
@attribute Competencia_Declarada real
@attribute Município real
@attribute Ano_Declarado real
@attribute CBO_2002_Ocupação real
@attribute CNAE_1_0_Classe real
@attribute CNAE_2_0_Classe real
@attribute CNAE_2_0_Subclas real
@attribute Faixa_Empr_Inicio_Jan real
@attribute Grau_Instrução real
@attribute Qtd_Hora_Contrat real
@attribute IBGE_Subsetor real
@attribute Idade real
@attribute Ind_Aprendiz real
@attribute Ind_Portador_Defic real
@attribute Raça_Cor real
@attribute Salario_Mensal real
@attribute Saldo_Mov real
@attribute Sexo real
@attribute Tempo_Emprego real
@attribute Tipo_Estab real
@attribute Tipo_Defic real
@attribute Tipo_Mov_Desagregado real
@attribute UF real
@attribute Mesorregiao_Microrregiao real


@data

2,201508,110028,2015,514320,74705,81214,8121400,7,7,44,21,34,0,0,9,819,-1,2,1,1,0,11,11,1102,11006
2,201508,110028,2015,514320,74705,81214,8121400,7,7,44,21,26,0,0,9,980,-1,2,1,1,0,6,11,1102,11006
2,201508,110028,2015,514320,74705,81214,8121400,7,7,44,21,43,0,0,8,878,-1,2,23,1,0,4,11,1102,11006
2,201508,110028,2015,514320,74705,81214,8121400,7,7,44,21,37,0,0,8,878,-1,2,1,1,0,11,11,1102,11006
2,201508,110028,2015,514225,74705,81214,8121400,7,7,44,21,22,0,0,9,1075,-1,2,3,1,0,11,11,1102,11006
```

Fonte: O autor

Com os dados em mãos foram aplicados alguns filtros: Estado= Maranhão, Ano= 2015 e Mês= Agosto. Com os filtros aplicados foram separados todos os registros em demitidos e admitidos. Os dados foram divididos com o objetivo de aplicar de forma separada o processo de clusterização em cada um dos casos. Foi preparado os arquivos ARFF para a leitura através da ferramenta WEKA. Como pode ser visto na figura abaixo o arquivo gerado a partir das bases selecionadas.

A clusterização foi executada com algumas configurações que serão relatadas no decorrer deste capítulo, mas todas utilizaram o algoritmo K-MEANS.

O primeiro experimento demostrado na Figura 7, foi realizado com os registros referentes aos funcionários demitidos, onde o número de interações para se chegar a um resultado aceitável foi 2.

Figura 7 – Primeiro experimento

```
40
41   Number of iterations: 2
42   Within cluster sum of squared errors: 488607.85778241453
43   Missing values globally replaced with mean/mode
44
45   Cluster centroids:
46                                      Cluster#
47   Attribute          Full Data          0          1
48                       (527702)   (141867)   (385835)
49   ================================================
50   CNAE_1_0_Classe   52378.1616 52690.1506  52263.447
51   Grau_Instrucao             7          9          7
52   Idade                 32.188    34.3491    31.3934
53   Raca_Cor               5.1063     2.8452     5.9376
54   Salario_Mensal     1521.1248  2231.2586  1260.0169
55   Sexo                       1          1          1
56   Tempo_Emprego        22.0233    29.6549    19.2173
57
58
59
60
61   Time taken to build model (full training data) : 6.3 seconds
62
63   === Model and evaluation on training set ===
64
65   Clustered Instances
66
67   0      141867 ( 27%)
68   1      385835 ( 73%)
69
70
```

Fonte: O autor

Neste experimento inicial, foram formados dois clusters (cluster 0 e cluster 1) compostos por funcionários do sexo masculino.

O cluster 0 (141.867), corresponde a 27% dos funcionários demitidos no setor do comércio varejista, com grau de instrução em Ensino Superior completo, com idade de 34 anos, da raça branca, possuindo salários de 2231. 2586 e com tempo de emprego correspondente a 29.6549 (aproximadamente 2 anos e 6 meses).

O cluster 1 (385.835), corresponde a 73% dos funcionários demitidos no setor do comércio varejista de bebidas. Essa mostra possui o grau de instrução no Ensino Médio completo, com idade de 31 anos, da raça amarela, com salários de 1260.0169 e com tempo de emprego de 19.2173 (aproximados 1 ano e 7 meses).

Com base nos resultados dessa primeira clusterização, pode-se inferir que a maioria dos funcionários demitidos é do sexo masculino. Do total dessa mostra, a maioria que foi demitida, correspondente a 73% dos funcionários, possuem grau de instrução apenas de Ensino Médio, e tem menos tempo de serviços prestados no comércio varejista, ou seja, passaram menos tempo nesse emprego. Outra evidência se mostra nos valores salariais que se apresentam abaixo daqueles que possuem ensino superior completo, tornando evidente o peso do grau de instrução na diferenciação salarial.

O segundo experimento ilustrado na Figura 8, também contém os registros de funcionários demitidos, onde o número de interações para se chegar a um resultado aceitável foi 24. Para este experimento foram formados quatro clusters, compostos

por funcionários do sexo masculino.

Figura 8 – Segundo experimento

```
kMeans
======

Number of iterations: 24
Within cluster sum of squared errors: 430782.23144026997
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute          Full Data          0          1          2          3
                    (527702)    (82628)    (49235)   (195525)   (200314)
=============================================================================
CNAE_1_0_Classe   52378.1616 46915.5522 18533.6568 59661.5652 55840.7813
Grau_Instrucao             7          9          7          7          7
Idade                 32.188    36.0155    31.6467    31.3191    31.5904
Raca_Cor               5.1063     3.4219     8.0663     8.1171     2.1346
Salario_Mensal     1521.1248  2894.2936  1236.1064  1170.5982   1366.904
Sexo                       1          1          1          1          1
Tempo_Emprego        22.0233    35.9645    24.4408    16.0177    21.5405



Time taken to build model (full training data) : 63.95 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       82628 ( 16%)
1       49235 (  9%)
2      195525 ( 37%)
3      200314 ( 38%)
```

Fonte: O autor

O cluster 0 (82628), corresponde a 16% dos funcionários demitidos no setor de construção, que possuem o grau de instrução em Ensino Superior completo, com idade de 36 anos, da raça branca, com salários de 2894.2936, e com tempo de emprego de 35.9645 (aproximadamente 2 anos e 9 meses).

O cluster 1 (49.235), corresponde a 9% dos funcionários demitidos no setor de fabricação de acessórios para segurança. Esses funcionários possuem o grau de instrução de Ensino Médio completo, com idade de 31 anos, da raça parda, com salários de 1236.1064 e com tempo de emprego de 24.4408 (aproximados 2 anos e 1 mês).

O cluster 2 (195525), corresponde a 37% dos funcionários demitidos no setor de restaurantes. Esses funcionários possuem o grau de instrução de Ensino Médio completo, com idade de 31 anos, da raça parda, com salários de 1170.5982 e com tempo de emprego de 16.0177 (aproximados 1 ano e 5 meses).

O cluster 3 (200314), corresponde a 38% dos funcionários demitidos no setor de restaurantes. Esses funcionários possuem o grau de instrução de Ensino Médio completo, com idade de 31 anos, da raça branca, com salários de 1366.904 e com tempo de emprego correspondente a 21.5405 (aproximados 1 ano e 10 meses).

Os clusters 2 e 3 são compostos por funcionários que estão no mesmo setor, na área de restaurantes. Pode-se inferir assim que, esse setor teve mais demissões em comparação com os setores de construção e de fabricação de acessórios para segurança e que o tempo de permanência no emprego é menor.

O terceiro experimento, Figura 9, foi realizado com os registros de funcionários admitidos, com 9 interações para se chegar a um resultado aceitável. Para este experimento foram formados dois clusters (clusters 0 e 1), com funcionários do sexo masculino.

Figura 9 – Terceiro experimento

```
kMeans
======

Number of iterations: 9
Within cluster sum of squared errors: 515104.29852336564
Missing values globally replaced with mean/mode

Cluster centroids:
                                  Cluster#
Attribute          Full Data          0          1
                    (520871)    (70520)   (450351)
=================================================
CNAE_1_0_Classe   52732.3847 18521.3536 58089.4549
Grau_Instrucao             7          2          7
Idade                30.7441    33.2132    30.3575
Raça_Cor              5.2542     5.3714     5.2358
Salario_Mensal     1331.6135  1297.6949  1336.9248
Sexo                       1          1          1
Tempo_Emprego              0          0          0



Time taken to build model (full training data) : 18.08 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       70520 ( 14%)
1      450351 ( 86%)
```

Fonte: O autor

O cluster 0 (70520), corresponde a 14% dos funcionários admitidos no setor de fabricação de acessórios do vestuário e de segurança, que possuem o grau de instrução de até o 5º ano incompleto do Ensino Fundamental (antiga 4º série) que se tenha alfabetizado sem ter frequentado escola regular, possuindo a idade de 33 anos, da raça amarela, com salários de 1297.6949, e com tempo de emprego de 0.

O cluster 1 (450351), corresponde a 86% dos funcionários admitidos no setor de restaurantes, que possuem o grau de instrução em Ensino Médio completo, com idade de 30 anos, da raça amarela, com salários de 1336.9248, e com tempo de emprego de 0.

Assim nota-se que a grande maioria da amostra (86%- cluster 1), composta por funcionários admitidos se encontra no setor de restaurantes, possuindo o Ensino Médio completo e salário superior aos que possuem apenas o grau de instrução de até o 5º ano incompleto do Ensino Fundamental (antiga 4ª série) e que se tenha alfabetizado sem ter frequentado escola regular do setor de fabricação de acessórios do vestuário e de segurança.

O quarto experimento, ilustrado na Figura 10, foi realizado com os registros de funcionários admitidos, onde foram feitas 11 interações para se chegar a um resultado aceitável. Para este experimento foram formados quatro clusters, onde três deles foram formados por funcionários do sexo masculino e um cluster pelo sexo feminino.

Figura 10 – Quarto experimento

```
kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 284964.1930004743
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute          Full Data          0          1          2          3
                    (520871)    (40356)   (138836)   (194109)   (147570)
=============================================================================
CNAE_1_0_Classe   52732.3847 18624.3568 50943.1002 60116.0941 54030.9973
Grau_Instrucao             7          2          7          7          7
Idade                30.7441    33.9821    31.2711    29.8304    30.5649
Raça_Cor              5.2542      6.847     2.2021     4.9169     8.1337
Salario_Mensal     1331.6135   1064.733  1598.8679  1246.1083  1265.6316
Sexo                       1          1          1          2          1
Tempo_Emprego              0          0          0          0          0



Time taken to build model (full training data) : 25.07 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       40356 (  8%)
1      138836 ( 27%)
2      194109 ( 37%)
3      147570 ( 28%)
```

Fonte: O autor

O cluster 0 (40356) como mostra a figura 13, corresponde a 8% dos funcionários admitidos no setor de fabricação de acessórios do vestuário e de segurança, do sexo masculino os quais possuem o grau de instrução de até o 5º ano incompleto do Ensino Fundamental (antiga 4ª série) que se tenha alfabetizado sem ter frequentado escola regular, possuindo a idade de 33 anos, da raça amarela, com salários de 1064.733, e com tempo de emprego de 0.

O cluster 1 (138836), corresponde a 27% dos funcionários admitidos no setor de comércio a varejo de combustíveis, do sexo masculino, que possuem o grau de instrução de Ensino Médio completo, possuindo a idade de 31 anos, da raça branca, com salários de 1598.8679, e com tempo de emprego de 0.

O cluster 2 (194109), corresponde a 37% dos funcionários admitidos no setor de transporte terrestre, do sexo feminino, que possuem o grau de instrução de Ensino Médio completo, idade de 29 anos, da raça preta, com salários de

1246.1083, e com tempo de emprego de 0.

O cluster 3 (147570), corresponde a 28% dos funcionários admitidos no setor de alojamento e alimentação, do sexo masculino, que possuem o grau de instrução de Ensino Médio completo, idade de 30 anos, da raça pardo, com salários de 1365.6316 e com tempo de emprego de 0.

Esse experimento contém a clusterização mais diversificada, onde abrange vários setores de ocupação, raça e sexo. Infere-se que o setor que menos admitiu funcionários foi fabricação de acessórios do vestuário e de segurança (8%), e o que mais admitiu foi o setor de transporte terrestre (37%), sendo esse o setor que tem admitidos mais mulheres do que os outros setores.

No quinto experimento ilustrado na Figura 11, o número de interações para que se chegasse a um resultado aceitável foi 13. Foram formados 6 clusters, onde três são compostos por funcionários do sexo feminino e três do sexo masculino, onde um desses grupos conta com registros de funcionários demitidos e cinco de admitidos.

Figura 11 – Quinto experimento

```
======

Number of iterations: 13
Within cluster sum of squared errors: 294879.7557037058
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute               Full Data          0          1          2          3          4          5
                         (459027)    (37300)    (30030)    (31836)    (80105)   (148355)   (131401)
=============================================================================================================
Admitidos_Desligados       1.3712     1.7121     1.0824     1.6844          2          1     1.3002
CNAE_1_0_Classe        54613.0663 58547.3268 56605.9655 47284.7423  50054.574 52259.4426 60252.5949
Grau_Instrucao                7          6          5          4          7          7          7
Idade                   31.2897    30.5375    33.8022    34.0382    32.6898    30.7998    29.9625
Raca_Cor                 4.2672     3.7386     6.7263      6.311     3.5238     4.2293      3.856
Salario_Mensal         1510.1102  1506.5249  1343.1862  1460.192  1744.093   1600.469  1316.7121
Sexo                          1          2          2          1          1          1          2
Tempo_Emprego            7.8816    17.7408     1.3845    14.1871    22.3026          0     5.1472



Time taken to build model (full training data) : 34.59 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       37300 (  8%)
1       30030 (  7%)
2       31836 (  7%)
3       80105 ( 17%)
4      148355 ( 32%)
5      131401 ( 29%)
```

Fonte: O autor

O cluster 0 (37300), corresponde a 8% dos funcionários admitidos no setor de restaurantes, do sexo feminino, que possuem o grau de instrução de Ensino Médio incompleto, idade de 30 anos, da raça preta, com salários de 1506.5249 e com tempo de emprego de 17.7408 (aproximados 1 ano e 6 meses).

O cluster 1 (30030), corresponde a 7% dos funcionários admitidos no setor de

restaurantes, do sexo feminino, que possuem o grau de instrução de Ensino Fundamental completo, idade de 33 anos, da raça amarela, com salários de 1343.1862 e com tempo de emprego de 1.3845 (aproximadamente 1 ano).

O cluster 2 (31836), corresponde a 7% dos funcionários admitidos no setor de aluguel de equipamentos de construção, do sexo masculino, que possuem o grau de instrução o 6 ao 9 ano do e Ensino Fundamental incompleto (antiga 5 a 8 série), idade de 34 anos, da raça amarela, com salários de 1460.192 e com tempo de emprego de 14.1871 (aproximadamente 1 ano e 2 meses).

O cluster 3 (80105), corresponde a 17% dos funcionários demitidos no setor de comércio; preparação de veículos automotores, objetos pessoais e domésticos, eles são do sexo masculino, que possuem o grau de instrução de Ensino Médio completo, idade de 32 anos, da raça preta, com salários de 1744.093 e com tempo de emprego de 22.3026 (aproximadamente 1 ano e 10 meses).

O cluster 4 (148355), corresponde a 32% dos funcionários admitidos no setor de comércio varejista, eles são do sexo masculino, que possuem o grau de instrução de Ensino Médio completo, idade de 30 anos, da raça preta, com salários de 1600.469 e com tempo de emprego de 0.

O cluster 5 (131401), corresponde a 29% dos funcionários admitidos no setor de transporte terrestre, eles são do sexo feminino, que possuem o grau de instrução de Ensino Médio completo, idade de 29anos, da raça preta, com salários de 1316.7121 e com tempo desemprego de 5.1472 (aproximados 5 meses).

Pode-se perceber que as mulheres têm se concentrado nos setores de restaurantes e transportes terrestres, onde aquelas que tem um salário mais elevado em relação as outras estão no setor de restaurantes. Fica evidente que mesmo possuindo o ensino médio incompleto as mulheres que trabalham em restaurantes ganham mais do que as que estão no setor de transporte. Mostra-se ainda que os homens, em sua maioria, ainda compõem o número maior de admissões em relação as mulheres, possuindo salários mais elevados mesmo quando estão em igualdade de raça e grau de instrução.

5. CONCLUSÃO

Conclui-se que, a mineração de dados consiste em apenas uma parte de um complexo processo de descoberta de conhecimento, pois trata da etapa de refinamento e filtragem dos dados, para que possam ser extraídos padrões de dados capazes de gerar conhecimento. Este processo combinado com as ferramentas de *Data Warehouse* possuem um poder de resultado ainda maior tornando o processo como um todo mais eficiente.

Foram descritas algumas técnicas e ferramentas capazes de facilitar e tornar possível a mineração de dados até mesmo para pessoas que não são especialistas em descoberta de conhecimento em base de dados.

O trabalho torna evidente a importância do processo de mineração de dados, pois ele é capaz de executar de uma forma operacional e satisfatória esse processo de encontrar relações entre dados que ainda não foram achadas, e isso partindo de uma base extensa de registros, que caso contrário levaria muito tempo para extrair conhecimento dela.

6. REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. **Fast algorithms for mining association rules**. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, 20., 1994, Santiago do Chile. Proceedings… Santiago do Chile, 1994. p. 487-499.

AMO, S. de. Técnicas de mineração de dados. São Paulo, Brazil, p. 44, 2015. Disponível em:<http://www.researchgate.net/profile/Sandra_Amo/publication/260300816_Tcnicas_de_ Minerao_de_Dados/links/54230bd80cf290c9e3ae25e3.pdf>.

BERRY, M. J.; LINOFF *et al.* **Data Mining Techniques For Marketing**. [S.l.: s.n.], 1997. BONEL, C. **Afinal, o que é Business Intelligence?** Rio de Janeiro, Brazil: [s.n.], 2015.

BRAGA, L. P. V. **Introdução à Mineração de Dados - 2a edição: Edição ampliada e revisad**. São Paulo, Brazil: [s.n.], 2005. 212 p.

CHEN, M.S. et al. **Data mining: an overview from database perspective**. IEEE Transactions on Knowledge and Data Engineering,v. 8, n.6, p. 866-883, 1996.

DALFOVO, O.; TAMBORLIN, N. **Business Intelligence**. Blumenau, Brazil: [s.n.], 2010.

FAYYAD, U. *et al.* From data mining to knowledge discovery in databases. 1996. Disponível em: <http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>.

GOEBEL, M.; GRUENWALD, L. **A survey of data mining and knowledge discovery software tools**. University of Auckland, p. 33, 1999. Disponível em: <http://kdd.org/exploration_files/ survey.pdf>.

GOLDSCHMIDT, R.; BEZERRA, E. **Data mining: conceitos, técnicas, algoritmos, orienta- ções e aplicações**. [S.l.: s.n.], 2015.

HAN, J. *et al.* **Data Mining Concepts and Techniques**. [S.l.: s.n.], 2011. HAND, D. *et al.* **Principles of Data Mining**. [S.l.: s.n.], 2001.

HARRISON, T. H. **Intranet data warehouse**. São Paulo, Brazil: [s.n.], 1998.

INMON, W. H. **Building the Data WareHouse: Gettiing Started. 4a Edição.** [S.l.: s.n.], 2005.

KEEL. **KEEL (Knowledge Extraction based on Evolutionary Learning). @ONLINE**. 2015. Disponível em: <http://www.keel.es/>.

MACEDO, D. C. e MATOS, S. N., **Extração de Conhecimento Através da Mineração de Dados**, Revista de Engenharia e Tecnologia. v. 2, n. 2, p.22-30, 2010.

MEGAPUTER. **Data Mining. Text Mining. All in a single, intuitive package. @ONLINE**. 2015. Disponível em: <http://www.megaputer.com/site/polyanalyst.php>.

MTE. **Portal CAGED @ONLINE**. 2015. Disponível em: <https://https://www.caged.gov.br>. PICHILIANI, M. C. **Conversando sobre Banco De Dados**. [S.l.: s.n.], 2008. 687 p.

QUILICI, J. A.; ZAMPIROLLI, F. A. **Sistemas inteligentes e mineração de dados**. São Paulo, Brazil: [s.n.], 2014.

REZENDE, S. O. **Sistemas Inteligentes: fundamentos e aplicações**. São Paulo, Brazil: [s.n.], 2003.

RULEQUEST. **Data Mining Tools See5 e C5.0.** @**ONLINE**. 2015. Disponível em: <https://www.rulequest.com/see5-info.html>.

WEKA. **Weka 3: Data Mining Software in Java.** @**ONLINE**. 2015. Disponível em: <http: //www.cs.waikato.ac.nz/ml/weka/>.

ZHANG, T. et al. **BIRCH: an efficient data clustering method for very large database**. In: ACM SIGMOD CONFERENCE, 1996, Montreal. Proceedings… Montreal: Le Centre Sheraton, 1996. p. 103-114