UNIVERSIDADE SANTA CECÍLIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA MECÂNICA
MESTRADO EM ENGENHARIA MECÂNICA

DAVI SILVESTRE MOREIRA DOS REIS

SOFTWARE DE MINERAÇÃO DE DADOS PARA OBTENÇÃO DE MENORES DISTÂNCIAS ENTRE EMPRESAS FORNECEDORAS DE AUTOPEÇAS E EMPRESAS MONTADORAS E DE MANUTENÇÃO DE VEÍCULOS

SANTOS/SP
2016

DAVI SILVESTRE MOREIRA DOS REIS

SOFTWARE DE MINERAÇÃO DE DADOS PARA OBTENÇÃO DE MENORES DISTÂNCIAS ENTRE EMPRESAS FORNECEDORAS DE AUTOPEÇAS E EMPRESAS MONTADORAS E DE MANUTENÇÃO E DE VEÍCULOS

Dissertação apresentada à Universidade Santa Cecília como parte dos requisitos para obtenção de título de Mestre no Programa de Pós-Graduação em Engenharia Mecânica, sob orientação da Profª. Ma. Dorotéa Vilanova Garcia.

SANTOS/SP

2016

Reis, Davi Silvestre Moreira dos.

Software de Mineração de Dados para Obtenção de Menores Distâncias entre Empresas Fornecedoras de Autopeças e Empresas Montadoras e de Manutenção de Veículos/Davi Silvestre Moreira dos Reis, 2016.
55 p.

Orientadora: Profa. Ma. Dorotéa Vilanova Garcia.

Dissertação (Mestrado) - Universidade Santa Cecília, Programa de Pós-Graduação em Engenharia Mecânica, Santos, SP, 2016.

1. Mineração de Dados. 2. Data Mining. 3. Cálculo de distância em coordenadas geográficas. 4. Distância entre empresas da indústria automotiva. I. Garcia, Dorotéa Vilanova. II. *Software* de Mineração de Dados para Obtenção de Menores Distâncias entre Empresas Fornecedoras de Autopeças e Empresas Montadoras e de Manutenção de Veículos.

Elaborada pelo SiBi - Sistema Integrado de Bibliotecas - Unisanta

Dedico esse trabalho a Deus, à minha mãe, ao meu pai (in memoriam), aos meus irmãos e à minha esposa.

AGRADECIMENTOS

Primeiramente a Deus, por me abençoar e me proteger, me permitindo galgar degraus não imaginados.

Aos meus pais, Caetana e Horácio (*in memoriam*), por me ensinarem os valores básicos da vida e por sempre me direcionarem aos estudos, desde a mais tenra idade, estando sempre ao meu lado em todos os momentos.

Aos meus irmãos, Danilo e Roberta, pela companhia fraternal, mesmo à distância.

À minha querida esposa Daniela, por ser meu apoio e ponto de equilíbrio em todas as horas, dosando cuidadosamente os momentos de compreensão e cobrança.

À professora Ma. Dorotéa Vilanova Garcia, por suas inestimáveis orientação e compreensão, além de ideias ao longo da elaboração do presente.

Ao Professor Dr. João Inácio, por suas aulas e por sua disponibilidade em colaborar e compartilhar seu vasto conhecimento.

Ao Professor Dr. Maurício Conceição, por suas ricas aulas e por sua frequente disponibilidade em auxiliar a todo o corpo discente com atenção.

Aos queridos amigos e colegas de mestrado Alexandre Stucchi, Fernando Bacic, João Carlos, Joseffe Barroso, Mario Rocha, Ricardo Reiff e Tertuliano Paulo (*in memoriam*), pelo apoio durante o período do mestrado e pela luta diária na docência.

E, na figura da Sandra, da secretaria da pós-graduação *Scrictu Sensu*, agradeço a todos funcionários do programa de Mestrado da Universidade Santa Cecília, que me auxiliaram e me permitiram chegar até o presente trabalho.

"Se enxerguei mais longe, foi porque me apoiei sobre os ombros de gigantes."

(Isaac Newton)

RESUMO

Esta dissertação apresenta um sistema inteligente que foi desenvolvido especificamente para encontrar as menores distâncias geográficas na indústria automotiva, especificamente entre empresas montadoras de veículos e carrocerias e empresas fabricantes de autopeças e equipamentos. Os dados iniciais das empresas estavam em formato de arquivo-texto puro, sob o qual repousavam, inertes, informações importantes para auxiliar no processo de tomada de decisão sobre a escolha das fornecedoras. O único fator considerado na escolha das parceiras de produção, para este caso, foi a menor distância geográfica. Para obter as menores distâncias, foi desenvolvido um sistema que utiliza as bases do processo de KDD (*Knowledge Discovery In Databases*, ou Descoberta de Conhecimento em Bases de Dados), em especial o *Data Mining* (Mineração de Dados). Os dados iniciais passaram por um processo de limpeza, depuração, redução de quantidade e preparação, até ser possível aplicar técnicas de mineração de dados, a fim de conseguir demonstrar a capacidade da ferramenta desenvolvida em extrair informações ricas para auxiliar no processo decisório para escolha de potenciais fornecedores, baseado na menor distância entre as empresas.

Palavras-Chave: Mineração de Dados. Data Mining. Cálculo de distância com coordenadas geográficas. Distância entre empresas da indústria automotiva.

ABSTRACT

This dissertation presents an intelligent system that was developed specifically to find the smallest geographic distances in the automotive industry, specifically between vehicle assembling companies and manufacturers. The data of the companies were in pure text-file format, under which had inert and important information to assist in the decision-making process on the choice of the companies. In this case, just the smallest geographic distance was the factor considered in the choice of companies. To obtain the shortest distances, a system was developed using the bases of the KDD (Knowledge Discovery In Databases) process, especially Data Mining. The initial data went through a process of cleaning, debugging, quantity reduction and preparation, until it is possible to apply data mining techniques, in order to demonstrate the ability of the tool developed in extracting rich information to aid in the decision-making process for choosing potential suppliers, based on the shortest distance between companies.

Keywords: Data mining. Calculation of distance with geographic coordinates. Distance between companies in the automotive industry.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

API	*Application Programming Interface*
GB	Giga-Bytes
GPS	*Global Positioning System*
JUCESP	Junta Comercial do Estado de São Paulo
MB	Mega-Bytes
SQL	*Structured Query Language*

LISTA DE SÍMBOLOS

$\emptyset_i$	Latitude inicial
λ_i	Longitude inicial
$\emptyset_f$	Latitude final
λ_f	Longitude final
$\Delta\emptyset$	Diferença numérica entre latitudes final e inicial
$\Delta\lambda$	Diferença numérica entre longitudes final e inicial

SUMÁRIO

1. INTRODUÇÃO

A escolha de um fornecedor de equipamentos ou matérias-primas é uma decisão muito importante no dia a dia das empresas. Isto porque o valor gasto com aquisição de produtos ou serviços para a produção de um bem, segundo Martins e Alt (2011), “varia de 50 a 80% do total das receitas brutas”. Assim, torna-se importante que o processo de tomada de decisão para a escolha de um fornecedor esteja entre as principais estratégias de uma empresa.

Nesse processo decisório, um dos vários fatores que influenciam na seleção de um fornecedor é, dentre outros, sua localização geográfica, pois muitas vezes é desejável que o fornecedor esteja próximo do comprador - ou, ao menos, mantenha um estoque local -, já que uma localização próxima auxilia na redução dos tempos de entrega (ARNOLD, 1999). Afinal, quanto maior a distância, “mais representativo será o custo do frete em relação ao valor da mercadoria” (AYRES, 2009). Além de poder influenciar no custo, a distância pode auxiliar nas atividades técnicas e comerciais, facilitando a resolução de problemas onde o contato presencial seja mais interessante (AYRES, 2009).

Sabendo da importância da proximidade geográfica entre fornecedores de insumos e potenciais clientes, conforme supramencionado, vários podem ser os meios para se obter as menores distâncias entre tais envolvidos. Uma maneira simples – e pouco prática – é calcular a distância a partir do próprio deslocamento entre um cliente e um fornecedor, com a utilização de um meio de transporte qualquer – como um carro, por exemplo. Outra maneira – mais prática – é calcular a distância entre fornecedor e cliente baseado em seus respectivos endereços, utilizando-se, para tal fim, algum equipamento de GPS[1]. Em ambos os casos, tal pesquisa pode ser realizada para uma pequena quantidade de fornecedores e clientes.

Contudo, quando se tem uma quantidade grande de empresas e muitas outras fornecedoras o cálculo da distância entre elas torna-se bastante complexo.

Uma das maneiras de tratar uma grande quantidade de dados complexos para tomada de decisão é explorar um campo da ciência da computação

[1] *Global Positioning System* – ou Sistema de Posicionamento Global, sistema de posicionamento via satélite que informa a um equipamento receptor sua localização em qualquer ponto do globo terrestre.

conhecido como KDD[2], que, de maneira simplista e resumida, trata-se de um processo de extração não trivial de informações potencialmente úteis – e previamente desconhecidas – a partir de uma base de dados (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992). Uma das áreas mais importantes do processo de KDD é a tarefa de *Data Mining* – DM ou mineração de dados –, que envolve a aplicação de algoritmos de mineração de dados de acordo com o objetivo a ser alcançado (CORDEIRO; FALOUTSOS; JÚNIOR, 2013).

A fim de explorar parte do potencial do processo de KDD e de sua importante área de mineração de dados, o presente trabalho utiliza uma base de dados contendo informações meramente cadastrais de empresas da área de engenharia mecânica, especificamente da indústria automotiva. Essa base de dados contém informações de empresas montadoras e/ou fabricantes de veículos, doravante denominadas “montadoras”, e empresas fabricantes e/ou fornecedores de autopeças, doravante denominadas “fornecedoras”.

A partir de técnicas de mineração de dados, foi elaborado um programa de computador com interfaces visuais – que constitui um sistema inteligente – para facilitar o entendimento e potencializar a visualização dos resultados. A dissertação ora elaborada apresenta o resultado do processamento de um algoritmo de mineração de dados utilizado para localizar as empresas fornecedoras mais próximas das empresas montadoras, dentro do Estado de São Paulo – condição técnica limitante da base de dados ora obtida.

1.1. JUSTIFICATIVA

A localização rápida e assertiva de potenciais fornecedores, bem como a distância geográfica a que estão deles, é condição preponderante para auxiliar gestores de empresas compradoras em seu processo de tomada de decisão na escolha de tais fornecedores. Entretanto, quando a quantidade de fornecedores é volumosa e desconhecida, a tarefa de realizar tal localização de forma manual torna-se lenta e pouco prática.

Assim, é mister que a localização de potenciais fornecedores seja feita de forma ágil e confiável para os gestores de empresas que demandam de insumos

[2] *Knowledge Discovery in Databases* – ou Descoberta de Conhecimento em Bases de Dados

para a produção de seus produtos.

A agilidade e a confiabilidade supramencionadas podem ser obtidas com o desenvolvimento de um programa de computador específico para tal fim, no qual esteja presente um processo de descoberta de conhecimento em bases de dados (KDD), bem como seu inerente algoritmo de mineração de dados. Tal desenvolvimento e seus resultados são apresentados no presente trabalho – para o qual não foi encontrado projeto semelhante durante o levantamento bibliográfico realizado.

1.2. OBJETIVOS

O objetivo deste trabalho é localizar as empresas fornecedoras de peças mais próximas das empresas montadoras de veículos, a partir de uma massa de dados contendo empresas de todo o Estado de São Paulo. Para tal fim, serão utilizados os conceitos do processo de descoberta de conhecimento em bases de dados, conhecido como KDD, dentro do qual há um algoritmo de mineração de dados específico para realizar o cálculo das distâncias entre as empresas, apresentando os resultados de forma visual, com a utilização de mapas e respectivos marcadores de posição das empresas e as distâncias entre elas.

1.3. KDD – *KNOWLEDGE DISCOVERY IN DATABASES*

Uma das definições mais consolidadas sobre *Knowledge Discovery in Databases* – KDD ou Descoberta de Conhecimento em Bases de Dados – especifica que este trata-se do “processo não trivial, interativo e iterativo, para identificação válida, original e potencialmente útil, de padrões compreensíveis existentes nos dados” (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

De acordo com Mathias (2015), o termo KDD foi cunhado em 1989, com o objetivo de representar todo o processo de busca e extração de conhecimento que, em seu nível mais operacional, inclui a aplicação de técnicas e algoritmos de mineração de dados para manipular e encontrar indícios de correlação ou de implicação em grandes volumes de dados.

Segundo Goldschmidt e Passos (2005), o termo iterativo sugere a possibilidade de repetições integrais ou parciais do processo de KDD e a expressão “não trivial” atenta para a complexidade normalmente presente na execução de processos de KDD. Dizer que a identificação deve ser válida indica que o conhecimento tem de ser verdadeiro e adequado ao contexto da aplicação em questão, enquanto o termo “original” designa que a descoberta deve acrescentar novos conhecimentos aos existentes, para que todo esse processo gere conhecimento útil e possa ser aplicado de forma a proporcionar benefícios ao contexto de aplicação de KDD. Contudo, a extração de conhecimento a partir de uma grande base de dados, utilizando um processo de KDD, exige a melhor compreensão das diferenças entre dado, informação e conhecimento, conforme ilustra a Figura 1:

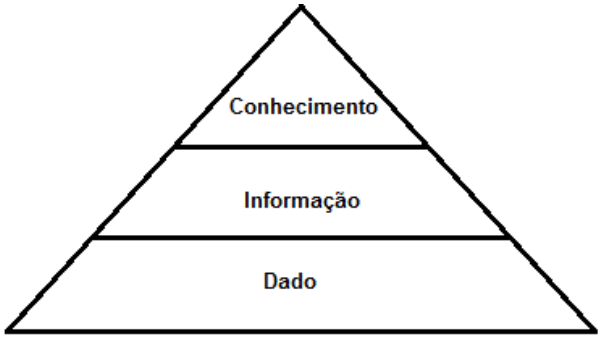

Figura 1 – Hierarquia entre dado, informação e conhecimento.
(Fonte: adaptada de REZENDE, 2003)

Os dados, na base da pirâmide, podem ser interpretados como itens elementares, captados e armazenados por recursos da tecnologia da informação (GOLDSCHMIDT; PASSOS, 2005), sendo elementos fora de contexto e sem significância. Já a "informação é o dado investido de relevância e propósito" (DRUCKER, 2001), que está contextualizado, organizado e com significado dentro desse contexto. Já o conhecimento é o padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações (GOLDSCHMIDT; PASSOS, 2005), sendo considerado a capacidade de interpretar as informações e, eventualmente, aplicá-las em um processo decisório.

De maneira geral, os passos que formam um processo de KDD costumam ter sempre uma mesma ordem e sequência. Inicialmente, Fayyad, Piatetsky-

Shapiro e Smyth (1996) propuseram serem nove os passos básicos que compõem um processo de KDD, a saber:

1. Desenvolver uma compreensão do domínio da aplicação;
2. Criar um conjunto de dados para analisar;
3. Limpeza de dados (*data cleaning*) e pré-processamento (*preprocessing*);
4. Redução de dados e projeção para procurar características relevantes;
5. Escolher a tarefa ou o objetivo do *data mining* (mineração de dados);
6. Escolher o(s) algoritmo(s) de *data mining* (mineração de dados);
7. *Data mining* (mineração de dados);
8. Interpretar padrões resultantes da procura e retornar para um dos passos anteriores para nova iteração;
9. Consolidar o conhecimento descoberto

Os nove passos apresentados acima são considerados como os básicos no processo de KDD, no entanto, não apresentam a dimensão de iterações e de ciclos que podem constituir todo seu processo. Os passos cinco, seis e sete são os que têm relação direta com a mineração de dados, sendo que a grande parte do trabalho num processo de KDD está centrada no sétimo passo.

Embora essa divisão em nove passos seja bastante conceituada e aceita, a evolução das tecnologias e as melhorias nos processos fazem com que autores mais recentes, como Rezende (2003), diminuam a quantidade de passos e os agrupem em três grandes etapas, conforme visto na Figura 2.

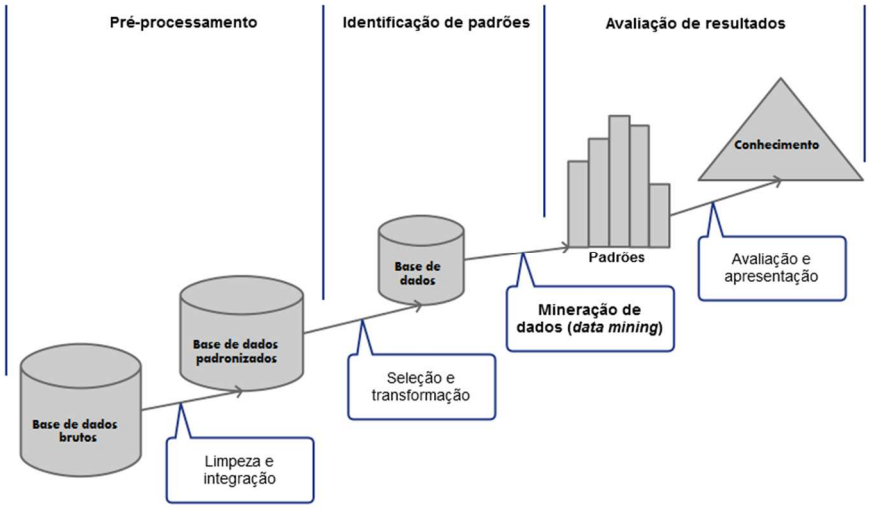

Figura 2 – Passos do processo de KDD.
(Fonte: adaptada de REZENDE, 2003)

De acordo com a Figura 2, a mineração de dados faz parte de uma etapa conhecida como "Identificação de Padrões", que é considerada a principal etapa do processo de KDD.

1.3.1. ETAPAS DO KDD – *KNOWLEDGE DISCOVERY IN DATABASES*

Basicamente, existe uma fase anterior ao processo de mineração de dados, que se refere ao conhecimento do domínio e identificação do problema, e uma fase posterior à mineração de dados, que se refere à utilização do conhecimento obtido. A seguir uma breve apresentação de cada etapa, de acordo com Rezende (2003):

a. Pré-processamento: caracteriza-se pela adequação dos dados para a extração de conhecimento. Nesta fase aplicam-se métodos que modificam parcialmente a massa de dados, com funções de limpeza, tratamento e redução do volume de dados, já que os dados disponíveis para análise nem sempre estão em um formato adequado. A realização das modificações que podem ser efetuadas na etapa de pré-processamento são: extração e integração, transformação, limpeza, seleção e redução de dados. Vale ressaltar que nem todas essas alterações são necessárias. Segue breve explicação sobre as alterações possíveis, ainda segundo Rezende (2003):
    1. Extração e integração: os dados disponíveis podem se apresentar em diversos formatos, como arquivo-texto, arquivos no formato de planilhas ou banco de dados. Dessa forma, há necessidade de unificação, a qual será aplicada como entrada para o algoritmo de extração de padrões;
    2. Transformação: após a extração e integração dos dados, algumas transformações podem ser realizadas aos dados, como resumos, agrupamentos ou padronizações;
    3. Limpeza: os dados disponíveis para aplicação dos algoritmos de extração de padrões podem apresentar problemas provenientes do processo de coleta de dados. Esses erros podem ser de digitação ou leitura de dados pelos sensores. Como o resultado do processo

de extração geralmente será aplicado em um processo de tomada de decisão, a qualidade dos dados é um fator extremamente relevante. A limpeza dos dados pode ser realizada utilizando o conhecimento do domínio que algum especialista tem. Por exemplo, pode-se encontrar registros com valor inválido em algum atributo, granularidade incorreta ou exemplos errôneos;

4. Seleção e redução de dados: em virtude das restrições de capacidade de memória ou desempenho (tempo de processamento), o número de exemplos e de atributos disponíveis para análise pode inviabilizar a utilização de algoritmos de extração de padrões. Como solução pode ser aplicado um método para redução dos dados antes de começar a busca de padrões. Esta redução pode ser realizada de três formas (WEISS; INDURKHYA, 1998): redução de atributos, de valores de um atributo ou número de exemplos.

b. Identificação de Padrões: objetiva o cumprimento das metas definidas na identificação do problema, inclusive com a escolha da tarefa de mineração de dados a ser empregada, a escolha do algoritmo a ser utilizado e a extração dos padrões propriamente dita. A tarefa de mineração pode ser classificada em preditiva ou descritiva. Uma atividade preditiva trata da generalização de exemplos (subdivididas em classificação e regressão), enquanto uma tarefa descritiva consiste na identificação de comportamentos inerentes ao conjunto de dados, sem que estes possuam classes específicas (atividades subdivididas em associação, sumarização e *clusterização* – atividade na qual o presente trabalho se baseia). Sendo um processo iterativo, pode ser necessário que esta etapa de identificação de padrões seja executada diversas vezes para ajustar o conjunto de parâmetros, tendo em vista a obtenção de resultados mais adequados aos objetivos pré-estabelecidos;
c. Avaliação de resultados (também conhecida como pós-processamento): Esta etapa é de suma importância para que o resultado final seja obtido. Ela visa propiciar que o conhecimento extraído seja utilizado na resolução de problemas reais, seja por meio de um sistema inteligente, seja por uso de um ser humano, auxiliando algum processo decisório (REZENDE,

2003). Caso tal conhecimento não atenda às necessidades do usuário final – em outras palavras, não esteja cumprindo com os objetivos propostos –, é necessário que o processo de extração de conhecimento seja repetido, ajustando-se assim os parâmetros ou melhorando o processo de escolha dos dados, a fim de obter resultados que realmente sejam úteis na iteração seguinte.

1.4. MINERAÇÃO DE DADOS OU *DATA MINING*

Conforme visto anteriormente, a mineração de dados faz parte do processo de descoberta de conhecimento em bases de dados (KDD), sendo considerada sua principal etapa.

São muitas as definições sobre mineração de dados encontradas na literatura. Dentre elas, seguem algumas interessantes:

"*Data mining* é uma ferramenta utilizada para descobrir novas correlações, padrões e tendências entre as informações de uma empresa, através da análise de grandes quantidades de dados armazenados em *Data Warehouse*[3] usando técnicas de reconhecimento de padrões, estatísticas e matemáticas" (NIMER; SPANDRI, 1998).

Segundo Rodrigues (2000), "d*ata mining* é um processo que encontra relações e modelos dentro de um grande volume de dados armazenados em um banco de dados".

Já Silva (2000) afirma ser o *data mining* uma técnica para determinar padrões de comportamento, em grandes bases de dados, auxiliando na tomada de decisão.

Outra definição interessante é feita por Possas et al. (1998): "*data mining* é um conjunto de técnicas que envolve métodos matemáticos, algoritmos e heurísticas para descobrir padrões e regularidades em grandes conjuntos de dados".

Contudo, a definição mais apropriada para o presente trabalho versa que a mineração de dados pode ser definida como uma técnica cujo objetivo é

[3] Segundo Date (2004), "*Data Warehouse* é um deposito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar as decisões gerenciais"

proporcionar ao usuário final suporte à sua tomada de decisão, para os casos em que haja uma grande base de dados a ser pesquisada e que, por seu tamanho, torna-se impossível a tomada de decisão ao analista humano por si só. Esse é um processo interativo entre homem e máquina, no qual um precisa do outro, ou seja, após a conclusão do processo de mineração, precisa-se de um especialista do domínio para fazer a análise dos dados válidos, estabelecendo-se uma relação de interdependência (MOUNT, 2004).

O intuito da utilização da mineração de dados é descobrir, de forma automática ou semiautomática, o conhecimento que existe e está oculto nas grandes quantidades de dados armazenados nos bancos de dados das empresas, auxiliando sobremaneira no processo de tomada de decisão. Uma organização que emprega ferramentas de *data mining* é capaz de criar parâmetros para entender o comportamento dos dados, identificar afinidades entre dados, prever hábitos ou comportamentos e analisar hábitos para detectar comportamentos fora do padrão, entre outros tipos de parâmetros.

Para Elmasri e Navathe (2005), os propósitos da mineração de dados podem ser classificados, de modo geral, como:

a. Predição: projeções feitas para identificar o comportamento de certos atributos no futuro;
b. Identificação: padrões de dados que podem identificar a presença de um item, um evento ou uma atividade;
c. Classificação: particionamento dos dados, onde as classes ou categorias podem ser identificadas através de combinações de parâmetros;
d. Otimização: realiza tarefas para otimizar recursos que sejam originalmente limitados, maximizando variáveis de saída.

A mineração de dados deve ser feita utilizando, dentre as técnicas disponíveis, a que melhor se aplica ao tipo de informação a ser encontrada, de acordo com a tarefa estabelecida para atender ao processo decisório.

1.4.1. TAREFAS DE MINERAÇÃO DE DADOS

Inicialmente, é importante separar os conceitos sobre o que é uma “tarefa” de mineração de dados e o que é uma “técnica” de mineração de dados. A tarefa trata, basicamente, da especificação sobre o que se deseja encontrar nos dados ou que tipo de comportamentos ou categoria de padrões pode conter informações relevantes. Já a técnica de mineração de dados consiste na especificação de métodos que garantam como encontrar os padrões que sejam interessantes à tomada de decisão.

De acordo com Reategui (2002), as tarefas de mineração de dados podem ser divididas, basicamente, em dois principais grupos:

a. Aprendizado supervisionado ou descoberta direta de conhecimento: mineração direcionada por objetivo, ou seja, explica o valor de determinado campo a partir de outros. Para isso, deve-se selecionar um campo alvo e solicitar ao sistema para estimá-lo, classificá-lo ou prevê-lo;
b. Aprendizado não-supervisionado ou descoberta indireta de conhecimento: não há campo alvo. Deve-se perguntar ao sistema como identificar padrões significativos nos dados. A partir disso, a técnica de mineração trabalha livremente na descoberta de padrões que podem ser úteis.

As tarefas que compõem os grupos acima descritos podem ser visualizadas na Figura 3:

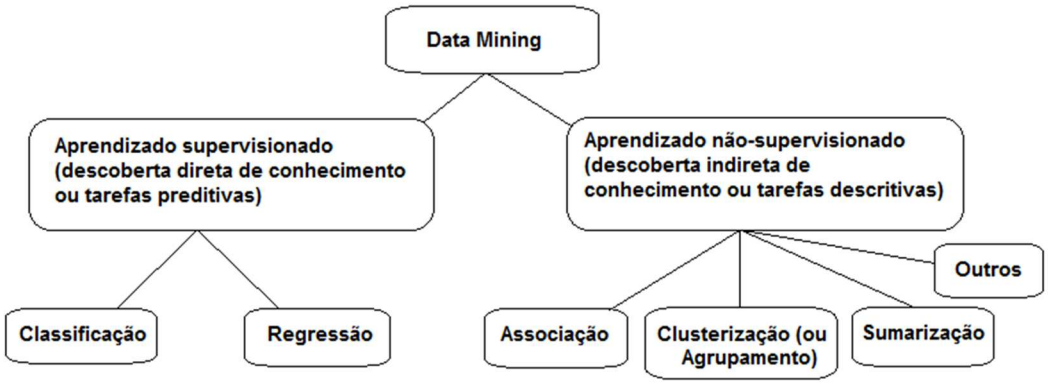

Figura 3 – Tarefas de *Data Mining*.
(Fonte: adaptada de REZENDE, 2003)

A mineração de dados é frequentemente classificada pela sua capacidade em realizar determinadas tarefas. Segundo Larose (2005), as tarefas mais comuns utilizadas no processo de mineração de dados são:

a. **Descrição**: É a tarefa utilizada para descrever os padrões e tendências revelados pelos dados. A descrição geralmente oferece uma possível interpretação para os resultados obtidos. A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido.

b. **Classificação**: Uma das tarefas mais comuns, a classificação visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos e o compara com cada registro que previamente já esteja categorizado em uma classe, com o intuito de "aprender" como classificar um novo registro. A tarefa de classificação pode ser usada, por exemplo, para:
    1. Determinar quando uma transação de cartão de crédito pode ser uma fraude;
    2. Identificar em uma escola, qual a turma mais indicada para um determinado aluno;
    3. Diagnosticar onde uma determinada doença pode estar presente.

c. **Regressão**: A regressão, também chamada estimação, é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. A tarefa de estimação pode ser usada, por exemplo, para:
    1. Estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas;
    2. Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal.

d. **Predição**: A tarefa de predição é similar às tarefas de classificação e

estimação, porém, ela visa descobrir o valor futuro de um determinado atributo. Exemplos:

a. Predizer o valor de uma ação três meses adiante (Figura 4);
b. Predizer o percentual de aumento de tráfego na rede se a velocidade aumentar.

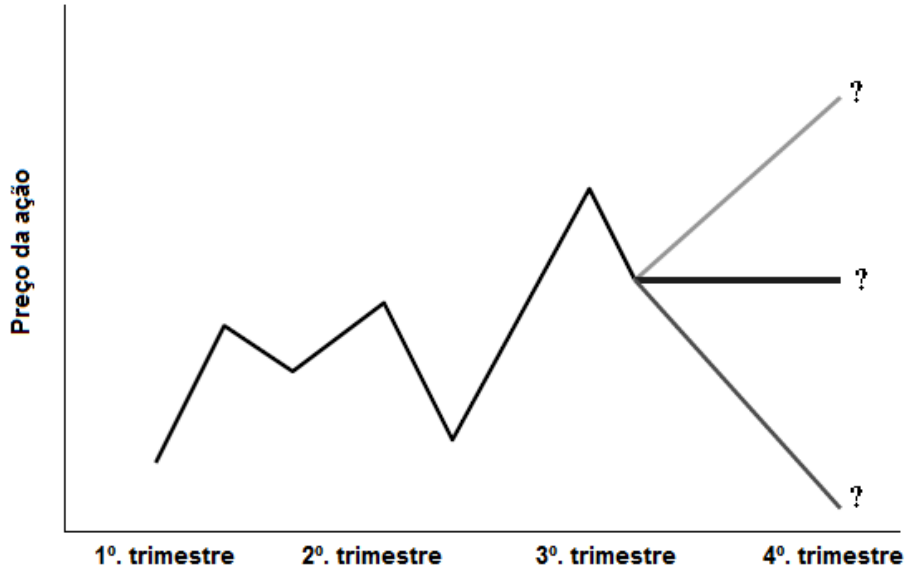

Figura 4 – Exemplo de Predição: preço de uma ação em três meses.
(Fonte: adaptada de LAROSE, 2005)

Alguns métodos de classificação e regressão podem ser usados para predição, com as devidas considerações.

e. **Agrupamento (ou *Clustering*)**: A tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou *cluster*) é uma coleção de registros similares entre si, porém, diferentes dos outros registros nos demais agrupamentos, como apresentado de forma meramente ilustrativa na Figura 5. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados. Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares. Exemplos:

1. Segmentação de mercado para um nicho de produtos;
2. Reduzir para um conjunto de atributos similares registros com centenas de atributos.

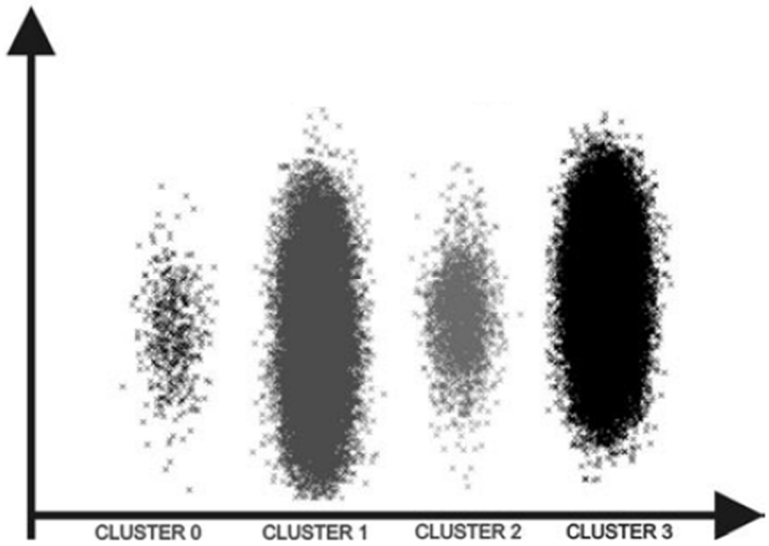

Figura 5 – Exemplo de agrupamento (*clustering*).
(Fonte: SALVADOR; MARQUES; CUNHA, 2009)

As aplicações das tarefas de agrupamento são as mais variadas possíveis: pesquisa de mercado, reconhecimento de padrões, processamento de imagens, análise de dados, segmentação de mercado, taxonomia de plantas e animais, pesquisas geográficas – como o proposto no presente trabalho -, detecção de comportamentos atípicos (fraudes), entre outras (OLIVEIRA; CARVALHO, 2008). Geralmente a tarefa de agrupamento é combinada com outras tarefas, além de ser usada na fase de preparação dos dados.

f. **Associação**: A tarefa de associação tem a função básica de associar elementos, identificando quais atributos estão relacionados entre si. Uma regra de associação possui a forma A ⇒ B (SE atributo A, ENTÃO atributo B), denota-se por (A) o conjunto de itens no antecedente da regra e por (B) o conjunto de itens no consequente da regra. Dessa forma, de acordo com Goldschmidt e Passos (2005), uma regra de associação indica que o conjunto de itens do antecedente das regras tem propensão a ocorrer juntamente com o conjunto de itens do consequente. Alguns exemplos:
    1. Determinar os casos onde um novo medicamento pode apresentar efeitos colaterais;

2. Identificar quais produtos são levados juntos pelos consumidores.

1.4.2. TÉCNICAS (OU MÉTODOS) DE MINERAÇÃO DE DADOS

As tarefas de mineração de dados são realizadas por técnicas de mineração de dados, sendo que diferentes técnicas servem para diferentes propósitos (HARRISON, 1998). A seguir são descritas brevemente as técnicas de mineração de dados mais utilizadas:

a. **Regras (ou Análise) de Associação**: identifica conjuntos de itens que ocorrem simultaneamente e de forma frequente em banco de dados, estabelecendo uma correlação estatística entre esses atributos (HARRISON, 1998). A aplicação de técnicas de análise de associação num conjunto de dados pode revelar afinidades entre uma coleção de itens. Essas afinidades entre itens são representadas por regras de associação. Uma regra expõe, textualmente, quais itens implicam a presença de outros itens.

b. **Regras de Indução (ou Classificação)**: trata-se de uma das técnicas mais comuns em mineração de dados. É altamente automatizada e, possivelmente, é a melhor técnica para de mineração para expor todas as possibilidades de padrões existentes em um banco de dados (BERSON; SMITH; THEARLING, 1999). Consiste, basicamente, na localização de propriedades comuns entre um conjunto dados e, posteriormente, os classifica em diferentes classes pré-definidas. As regras geradas possuem a estrutura de "se <condição> então <conclusão>", ou, em outras palavras, “se <isto> então <aquilo>” (por exemplo, “**se** comprou queijo e presunto **então** comprou pão também”). Frequentemente as regras de classificação são usadas para representar conhecimentos em sistemas especialistas, podendo ser interpretadas por especialistas humanos sem maiores dificuldades.

c. **Árvores de Decisão**: são usadas para mineração de dados dirigida, particularmente a classificação. Dividem os registros do conjunto de

dados de treinamento em subconjuntos separados, cada um descrito por uma regra simples em um ou mais campos, o que acaba por criar uma árvore de decisão. Cada nó não terminal desta árvore representa um teste ou decisão sobre o item dado. Uma das principais vantagens das árvores de decisão é que o modelo é bem explicável, uma vez que tem a forma de regras explícitas (KIMBALL, 1998)(HARRISON, 1998). Isso permite às pessoas avaliarem os resultados, identificando atributos-chave no processo. As próprias regras podem ser expressas facilmente como declarações lógicas em linguagens de programação simples.

d. **Raciocínio Baseado em Casos (ou MBR, *Memory Based Reasoning*)**: procura solucionar problemas fazendo uso direto de experiências e soluções passadas. Procura solucionar problemas fazendo uso direto de experiências e soluções passadas. Trata-se de uma técnica de mineração de dados dirigido, que usa exemplos conhecidos como modelo para fazer previsões sobre exemplos desconhecidos (HARRISON, 1998). Uma das maiores vantagens do MBR é a habilidade de ser executado em qualquer fonte de dados, mesmo sem modificações (HARRISON, 1998). Os dois elementos-chave no MBR são a função de distância usada apara encontrar os vizinhos mais próximos e a função de combinação, que combina valores dos vizinhos para fazer uma previsão. Outra vantagem do MBR é sua habilidade de aprender sobre novas classificações, simplesmente introduzindo novos exemplos no banco de dados.

e. **Algoritmos Genéticos**: muito útil para problemas que envolvem otimização (GOLDSCHMIDT; PASSOS, 2005), trata-se de um procedimento iterativo para construção de hipóteses sobre a dependência entre as variáveis. Basicamente, os algoritmos genéticos aplicam mecanismos de seleção genéticos e naturais para uma busca usada para encontrar os melhores conjuntos de parâmetros que descrevem uma função de previsão.

f. **Redes Neurais Artificiais (RNA)**: são técnicas que procuram reproduzir, de maneira simplificada, as conexões do sistema biológico neural,

formando neurônios artificiais interconectados, organizados em camadas que aprendem pela modificação de suas conexões. As redes neurais não-supervisionadas são as mais adequados para realização das tarefas de agrupamento (GOLDSCHMIDT; PASSOS, 2005). Segundo Harrison (1998), as redes neurais artificias são provavelmente a técnica de mineração de dados mais comum, talvez sinônimo de data mining para algumas pessoas. Em sua forma mais comum, aprendem com um conjunto de dados de treinamento, generalizando modelos para classificação e previsão.

g. **Agrupamentos ou Clusterização (*Clustering*)**: Segundo Harrison (1998), esta técnica é definida como a construção de modelos que encontram registros de dados semelhantes. Essas reuniões por semelhança são chamadas grupos (*clusters*). Diferentemente da classificação, em que os dados estão previamente rotulados, a análise de *clusters* trabalha sobre dados em que as classes não estão definidas. A técnica consiste em identificar novos agrupamentos, que contenham características similares e agrupar os registros, ou seja, particionar (segmentar) uma dada população de objetos ou itens em conjuntos. Existem vários métodos clusterização descritos na literatura, sendo que a escolha do método ideal depende do tipo de dado a ser analisado, assim como do propósito e da aplicação da análise. Dentre esses métodos, destacam-se o particionamento, o hierárquico e os fundamentados em densidade. Entretanto, as técnicas de agrupamento têm sido cada vez mais utilizadas em bases de dados geográficos - tal qual no presente trabalho -, devido à grande quantidade de dados coletados. Por agrupamento é possível identificar regiões mais densas e mais esparsas, logo, é possível descobrir padrões de distribuição global e as correlações interessantes entre os atributos dos dados (SEIXAS, 2011). De acordo com Goldschmidt e Passos (2005), o algoritmo mais popular para realização da tarefa de agrupamento é o K-médias (ou *K-Means*), que utiliza o método de partição.

1.4.3. MINERAÇÃO DE DADOS ESPACIAIS (*SPATIAL DATA MINING*)

Mineração de dados espaciais, ou *spatial data mining*, é uma extensão da mineração de dados voltada para domínios de aplicação onde a consideração da dimensão espacial é essencial na extração de conhecimento (NEVES; FREITAS; CÂMARA, 2001).

Aliás, a principal diferença entre *data mining* e *spatial data mining* é consideração dos relacionamentos espaciais existentes entre as entidades do mundo real. Ester et al. (1999, apud NEVES; FREITAS; CÂMARA, 2001) apresenta três tipos básicos de relações espaciais: relações topológicas, de distâncias e de direção.

A presente dissertação envolve conceitos relacionados às distâncias entre objetos presentes no globo terrestre, levando em conta, para isso, os dados de coordenadas geográficas de cada ponto que a ser analisado.

1.5. LOCALIZAÇÃO FÍSICA DE OBJETOS NO GLOBO TERRESTRE

A localização física de um objeto baseia-se na sua localização espacial, que, por sua vez, é expressa pelo sistema de coordenadas geográficas. Através das coordenadas geográficas é possível expressar qualquer posição horizontal no globo terrestre através de duas das três coordenadas existentes num sistema de coordenadas esférico, alinhadas com o eixo de rotação da Terra.

1.5.1. SISTEMA DE COORDENADAS GEOGRÁFICAS

Segundo Silva (2012), o sistema de coordenadas geográficas faz uso das coordenadas latitude e longitude. Tais coordenadas podem ser expressas de várias formas, como segue:

a. Graus - Minutos - Segundos, onde cada grau é dividido em 60 minutos, que por sua vez se subdividem, cada um, em 60 segundos. Exemplo: 22o 54’ 21.64"S 47o 03’ 38.06"W;

b. Graus - Minutos decimais, onde cada grau é dividido em 60 minutos, que por sua vez são divididos decimalmente. Exemplo: 22o 54.361’ S 47o 3.634’ W;

c. Graus decimais, onde a latitude recebe a abreviatura *lat* e a longitude é abreviada como *lon*; os valores positivos são para Norte (latitude) e Leste (longitude) e os valores negativos são para Sul (latitude) e Oeste (longitude). Exemplo: lat -22.906014º lon -47.060571º.

A latitude geográfica de um ponto na superfície da Terra equivale ao ângulo entre o plano equatorial e uma linha que passa por esse ponto e é normal à superfície de referência que aproxima a forma da Terra. A latitude mede-se para Norte e para Sul do Equador, entre -90º no Polo Sul e +90º no Polo Norte. A longitude descreve a localização de um lugar medido em graus, de 0º a -180º para Oeste ou a +180º para Leste, a partir do Meridiano de Greenwich. Portanto, ao se combinar estes dois ângulos, latitude e longitude, é possível indicar qualquer localização na superfície terrestre.

1.5.2. CÁLCULO DA DISTÂNCIA ENTRE DOIS PONTOS GEOGRÁFICOS

Devido ao fato de a latitude e a longitude não terem o mesmo tamanho ao longo de suas respectivas distribuições no globo terrestre, não é possível realizar a medição da distância entre quaisquer 2 pontos geográficos com precisão recorrendo-se apenas a unidades de medida angulares.

Assim, para determinar a distância entre dois pontos de acesso, foi utilizada a fórmula de Haversine (NORDIN et al., 2012). Essa função é bastante usada em sistemas de navegação e é capaz de fornecer a distância entre dois pontos de uma esfera utilizando coordenadas geográficas (latitude e longitude). Realizando uma aproximação da Terra como uma esfera perfeita, é possível apresentar um erro médio de 0.3% nos cálculos.

A fórmula de Haversine é apresentada na Equação 1:

Equação 1 – Fórmula de Haversine

$$d = 2\arcsin\left(\sqrt{sin^2\left(\frac{\Delta\emptyset}{2}\right) + \cos\emptyset_i \,.\cos\emptyset_f \,.sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

(Fonte: SILVA, 2012)

Onde $\emptyset_i$, λ_i, $\emptyset_f$ e λ_f representam as coordenadas, latitude e longitude, dos pontos inicial e final, respectivamente, e $\Delta\emptyset$ e $\Delta\lambda$ representam as diferenças em latitude e longitude, respetivamente, entre os 2 pontos.

O valor calculado na fórmula da Equação 1, “d”, representa a distância angular entre os 2 pontos na esfera terrestre. Portanto, a distância entre eles, em quilômetros (km), é dada por $r\,.d$, onde r representa o raio da esfera terrestre, em quilômetros – no caso da presente dissertação, foi considerado o valor de 6.372,8km (LAGARINHOS; TENÓRIO, 2012). Para utilização da fórmula de Haversine, as coordenadas geográficas estão representadas em graus decimais ao longo deste trabalho.

1.5.3. DADOS ESPACIAIS

Também chamados dados geoespaciais ou informação geográfica, os dados espaciais são dados que se relacionam com objetos que ocupam o espaço. Podem ser descritos através de propriedades geométricas como, por exemplo, a localização e a área, ou através de propriedades topológicas.

Os dados espaciais são normalmente acessados, manipulados e analisados através de um Sistema de Informação Geográfica (SIG), que permite e facilita a análise, gestão e representação do espaço e dos fenômenos que nele ocorrem.

Devido às suas caraterísticas, os dados espaciais devem ser guardados em bases de dados propriamente preparadas, a fim de permitir o armazenamento e a manipulação dos dados de forma eficiente. Além de informações de posicionamento e localização como latitude, longitude e altitude, pode ser necessário gravar informações sobre itens utilizados para plotagens de informações gráficas em mapas, como pontos de localização, retas, polígonos etc.

2. MATERIAIS E MÉTODOS

Para o desenvolvimento da pesquisa foram utilizados dados meramente cadastrais de empresas da área de engenharia mecânica, especificamente da indústria automotiva. Essa base de dados contém informações de empresas montadoras e/ou fabricantes de veículos, doravante denominadas “montadoras”, e empresas fabricantes e/ou fornecedores de autopeças, doravante denominadas “fornecedoras”.

Esses dados cadastrais foram obtidos perante a Junta Comercial do Estado de São Paulo – JUCESP –, em formato de arquivo de textos simples, contendo os seguintes dados (de maior relevância para a pesquisa): NIRE, Data da constituição, Razão Social, Logradouro, Número, Bairro, Município, UF, Objeto e CNAE. Uma amostra do arquivo pode ser vista na Figura 6:

NIRE	DATA INICIO ATIVIDADES	RAZAO SOCIAL	CNPJ	TIPO JURÍDICO	LOGRADOURO	NUMERO	BAIRRO	MUNICIPIO	UF	CEP	CAPITAL OBJETO	CNAE	NOME	IDENTIFICACAO	COTAS	CARGO
35212466797	NULL	C.C.F. COMERCIO IMPORTACAO EXPORTACAO LTDA	30339000130	Sociedade Limitada	RUA LUS NOLBERO FREIRE	637	JARDIM BRASILIA	Sao Paulo	SP	3585150	2.000.000.000.00					
35212466797	NULL	C.C.F. COMERCIO IMPORTACAO EXPORTACAO LTDA	30339000130	Sociedade Limitada	RUA LUS NOLBERO FREIRE	637	JARDIM BRASILIA	Sao Paulo	SP	3585150	2.000.000.000.00					
35221142613	NULL	ECO LAVAGEM DE VEICULOS LTDA	8597669000114	Sociedade Limitada	RUA OTAVIANO ALVES DE LIMA	1824	BAIRRO DO LIMAO	Sao Paulo	SP	2701000	100.000.000.000.000					
35221142613	NULL	ECO LAVAGEM DE VEICULOS LTDA	8597669000114	Sociedade Limitada	RUA OTAVIANO ALVES DE LIMA	1824	BAIRRO DO LIMAO	Sao Paulo	SP	2701000	100.000.000.000.000					
35220115795	NULL	B & B COMERCIO DE VEICULOS LTDA.	7619942000100	Sociedade Limitada	RUA BARAO DE MOTA PAES	730	CENTRO	Espirito Santo do Pinhal	SP	13990000	200.000.					
35220115795	NULL	B & B COMERCIO DE VEICULOS LTDA.	7619942000100	Sociedade Limitada	RUA BARAO DE MOTA PAES	730	CENTRO	Espirito Santo do Pinhal	SP	13990000	200.000.					
35220115795	NULL	B & B COMERCIO DE VEICULOS LTDA.	7619942000100	Sociedade Limitada	RUA BARAO DE MOTA PAES	730	CENTRO	Espirito Santo do Pinhal	SP	13990000	200.000.					
35225995319	20/09/2011	V.R.D. MULTIMARCAS LTDA	NULL	Sociedade Limitada	AVENIDA MARIA AUGUSTA FAGUNDES GOMES	106	PARQUE SANTA CRUZ D	Jacarei	SP	12322300	500.000.000.000.					
35217579859	26/04/2002	AUTO CENTER RIBEIRAO LTDA	50864050000108	Sociedade Limitada	AV BRSIL	500	CENTRO	Ribeirão Pires	SP	9400005	45.000.000.000.000		NULL	4520005		
35217579859	26/04/2002	AUTO CENTER RIBEIRAO LTDA	50864050000108	Sociedade Limitada	AV BRSIL	500	CENTRO	Ribeirão Pires	SP	9400005	45.000.000.000.000		NULL	4520005		
35224771816	31/03/2010	VALE MOREIRA - SERVICOS DE LAVA JATO LTDA.	NULL	Sociedade Limitada	AVENIDA ARMANDO SALES DE OLIVEIRA	262	TAQUARAL	Campinas	SP	13076015						
35224771816	31/03/2010	VALE MOREIRA - SERVICOS DE LAVA JATO LTDA.	NULL	Sociedade Limitada	AVENIDA ARMANDO SALES DE OLIVEIRA	262	TAQUARAL	Campinas	SP	13076015						
35216141965	08/02/2000	ZINHO'S FUNILARIA E PINTURA LTDA.	3671141000124	Sociedade Limitada	R. 28	3465	PQ. DO ESTADO II	Mogi Mirim	SP	13800000	50.000.000.000.0					
35220115795	NULL	B & B COMERCIO DE VEICULOS LTDA.	7619942000100	Sociedade Limitada	RUA BARAO DE MOTA PAES	730	CENTRO	Espirito Santo do Pinhal	SP	13990000	200.000.					
35229382419	05/08/2015	MARQUES MODESTO & HILARIO LTDA	NULL	Sociedade Limitada	RUA SETE DE JANEIRO	87	JARDIM SANTANA	Tremembé	SP	12120000	200.000.000.000.000					
35229382419	05/08/2015	MARQUES MODESTO & HILARIO LTDA	NULL	Sociedade Limitada	RUA SETE DE JANEIRO	87	JARDIM SANTANA	Tremembé	SP	12120000	200.000.000.000.000					
35229382419	05/08/2015	MARQUES MODESTO & HILARIO LTDA	NULL	Sociedade Limitada	RUA SETE DE JANEIRO	87	JARDIM SANTANA	Tremembé	SP	12120000	200.000.000.000.000					
35229382419	05/08/2015	MARQUES MODESTO & HILARIO LTDA	NULL	Sociedade Limitada	RUA SETE DE JANEIRO	87	JARDIM SANTANA	Tremembé	SP	12120000	200.000.000.000.000					
35203092529	12/04/1985	EMPORIUM DITRIBUIDORA NACIONAL DE PECAS LTDA.	54391446000113	Sociedade Limitada	RUA MANOEL BACELAR	266	JARDIM HELEAN	Sao Paulo	SP	8265120	600.000.					
35203092529	12/04/1985	EMPORIUM DITRIBUIDORA NACIONAL DE PECAS LTDA.	54391446000113	Sociedade Limitada	RUA MANOEL BACELAR	266	JARDIM HELEAN	Sao Paulo	SP	8265120	600.000.					
35215100238	24/03/1998	N.A.B.R.A NUCLEO AUTOMOTIVO BRASILEIRO LTDA	2480608000196	Sociedade Limitada	RUA JEAN ATLAN	28	VL. SANTA TEREZA	Sao Paulo	SP	4187080	700.000.					
35215100238	24/03/1998	N.A.B.R.A NUCLEO AUTOMOTIVO BRASILEIRO LTDA	2480608000196	Sociedade Limitada	RUA JEAN ATLAN	28	VL. SANTA TEREZA	Sao Paulo	SP	4187080	700.000.					
35223873046	19/01/2010	T.P. COMERCIO DE AUTO PECAS LTDA	11589627000146	Sociedade Limitada	RUA CRISTIANOPOLIS	460	MOOCA	Sao Paulo	SP	3128030	600.000.000.000.000					
35220128340	01/07/2005	REDRASFER INDUSTRIA DE AUTO PECAS LTDA.	75892470000134	Sociedade Limitada	ESTRADA MUNICIPAL RFD 258		111	D.I. AUGUSTO SCARAS	Rio das Pedras	SP	13390970					
35226264903	13/12/2011	LS COMERCIO E MANUTENCAO DE EMPILHADEIRAS LTDA	NULL	Sociedade Limitada	TRAVESSA GESSY ALVES MEIRA	108	JARDIM ROSA	Francisco Morato	SP	7991045						
35226264903	13/12/2011	LS COMERCIO E MANUTENCAO DE EMPILHADEIRAS LTDA	NULL	Sociedade Limitada	TRAVESSA GESSY ALVES MEIRA	108	JARDIM ROSA	Francisco Morato	SP	7991045						
35226264903	13/12/2011	LS COMERCIO E MANUTENCAO DE EMPILHADEIRAS LTDA	NULL	Sociedade Limitada	TRAVESSA GESSY ALVES MEIRA	108	JARDIM ROSA	Francisco Morato	SP	7991045						
35217276295	NULL	FABIO BIZARRIA COMERCIO DE PECAS E ACESSORIOS PARA VEICULOS AUTOMOTORES LTDA	4903275000196	Sociedade Limitada	RUA SOLDADO JOSE ALVES DE ABREU	271	VILA PANTALEAO	Caçapava								

Figura 6 – Fração do arquivo original em formato texto.

O campo NIRE (Número de Identificação de Registro de Empresas) serve para identificar, de maneira única, uma empresa em nível nacional. Seu valor é atribuído a cada empresa no momento em que ela é registrada na junta comercial de cada Estado da Federação.

Já o campo CNAE (Classificação Nacional de Atividades Econômicas) (IBGE, 2016) contém a informação sobre o tipo de atividade econômica exercida pela empresa, sendo uma forma de padronizar, em todo o território nacional, as atividades econômicas executadas. Esse atributo é de fundamental importância para o presente trabalho, visto que a partir dele foi possível realizar uma primeira parte do processo de KDD, filtrando as empresas desejadas para tal pesquisa.

2.1. *SOFTWARES* UTILIZADOS NA CONSTRUÇÃO DA APLICAÇÃO

A seguir são descritos brevemente os *softwares* empregados para a construção da aplicação utilizada para demonstrar os conceitos estudados no presente trabalho.

2.1.1. SISTEMA GERENCIADOR DE BANCO DE DADOS MYSQL

Para o gerenciamento do banco de dados foi utilizado o *software* MySQL, em sua versão 5.6. A ferramenta foi escolhida por se tratar de um programa robusto para manipulação de grandes massas de dados, além de ter sua licença gratuita. O MySQL utiliza a linguagem SQL como interface, permitindo fácil acesso para consulta, gravação, alteração e exclusão de dados armazenados em suas tabelas.

2.1.2. AMBIENTE E LINGUAGEM DE PROGRAMAÇÃO

Para o desenvolvimento do sistema foi utilizada a ferramenta Microsoft® Visual Studio 2013 e a linguagem de programação C# (C Sharp). Optou-se por este ambiente por ser um ambiente consolidado, já com versões gratuitas, que permite o desenvolvimento de programação orientada a objeto e com total integração ao banco de dados.

O sistema foi desenvolvido em camadas, cada qual com sua responsabilidade e função específicas, visando facilitar o desenvolvimento e a manutenção do projeto. Seguem-nas:

a. Camada de acesso a dados: onde é feita a comunicação com o banco de dados e a construção dos objetos que interagem com o mesmo;
b. Camada de regras de negócios: onde são feitas e armazenadas as regras de negócio e as validações atinentes ao domínio do problema; essa camada realiza a comunicação entre a camada de acesso a dados e a camada de interface gráfica com o usuário;

c. Camada de interface gráfica: responsável por permitir a interação com o usuário, recebendo dados e ações deste e apresentando resultados. Comunica-se diretamente com a camada de negócios e, em alguns casos.

2.2. PREPARAÇÃO E DESENVOLVIMENTO DO BANCO DE DADOS

O arquivo-texto contendo os dados das empresas foi originalmente gerado com alguns dados duplicados, faltantes ou despadronizados, além de espaços que ocupavam tamanho desnecessário. O arquivo original ocupava 0,99 GB (1.071.899.995 bytes) de espaço, contendo um total de 830.203 registros.

Dentro do conceito de pré-processamento do KDD, após uma primeira limpeza do arquivo, na qual foram eliminados os espaços em branco, seu tamanho foi reduzido para pouco mais da metade do tamanho original, passando a ocupar 506 MB.

Em seguida, no banco de dados foram criadas as tabelas necessárias para armazenar os registros e permitir o funcionamento do sistema. Dentre essas tabelas de banco de dados está a tabela “tb_empresa_geral” (Figura 7), para a qual todos os 830.203 registros do arquivo-texto foram importados.

Figura 7 – Tabela de banco de dados com dados não filtrados.

Após a importação de todos os registros, foram aplicados mais filtros e processos de melhoria nos dados originais, eliminando registros duplicados e obtendo somente os registros de empresas com atividades econômicas realmente pertinentes à pesquisa. Para obter esses registros, foram consideradas as empresas com CNAEs de números referentes a algumas atividades de engenharia mecânica, especificamente da indústria automotiva. Na Tabela 1 é possível verificar as empresas fabricantes e/ou montadoras de veículos.

Tabela 1 – CNAEs relacionados a fabricantes e montadores de veículos.

CNAE	Descrição da atividade econômica
29.10-7	Fabricação de automóveis, camionetas e utilitários
29.20-4	Fabricação de caminhões e ônibus
29.30-1	Fabricação de cabines, carrocerias e reboques para veículos automotores

Para fornecimento de peças e equipamentos para as montadoras, foram listadas as empresas com CNAEs relacionados a fabricação de peças e acessórios para a indústria automotiva, conforme apresentado na Tabela 2:

Tabela 2 – CNAEs relacionados à fabricantes e fornecedores de autopeças.

CNAE	Descrição da atividade econômica
29.41-7	Fabricação de peças e acessórios para o sistema motor de veículos automotores
29.42-5	Fabricação de peças e acessórios para os sistemas de marcha e transmissão de veículos automotores
29.43-3	Fabricação de peças e acessórios para o sistema de freios de veículos automotores
29.44-1	Fabricação de peças e acessórios para o sistema de direção e suspensão de veículos automotores
29.45-0	Fabricação de material elétrico e eletrônico para veículos automotores, exceto baterias
29.49-2	Fabricação de peças e acessórios para veículos automotores não especificados anteriormente

A partir dessa segmentação e do processo de limpeza e filtragem feitos anteriormente, restaram separadas 1.095 empresas consideradas fabricantes e/ou montadoras de veículos automotores e 3.515 empresas fabricantes de autopeças, sendo consideradas potenciais fornecedoras de insumos para as montadoras do primeiro grupo.

Após esse processo de limpeza e categorização, os registros dessas empresas foram armazenados em outra tabela de banco de dados (“tb_empresa”), enquanto uma nova tabela (“tb_empresa_cnae”) recebeu os códigos das atividades econômicas (CNAEs), vinculados à cada empresa. Por fim, outra tabela foi criada para armazenar as informações dos municípios existentes no Estado de São Paulo, bem como as regiões administrativas a cada qual município pertence (SEADE, 2016a).

Essa última tabela (“tb_regiao_administrativa”) serve de apoio ao usuário, a fim de permitir a aplicação de filtros de pesquisa quanto à localização das empresas em cada região administrativa do Estado de São Paulo.

As tabelas de banco de dados acima descritas são vistas na Figura 8:

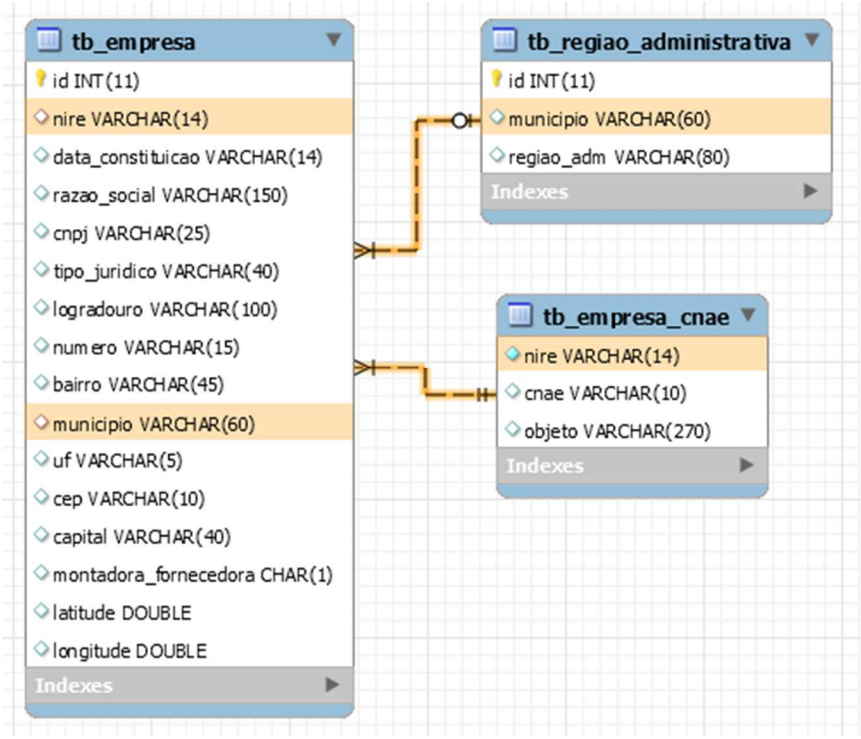

Figura 8 – Relacionamento entre tabelas do banco de dados.

Já as regiões administrativas do Estado de São Paulo encontram-se na Tabela 3:

Tabela 3 – Regiões Administrativas do Estado de São Paulo.

Regiões do Estado de SP
Administrativa Central
Administrativa de Araçatuba
Administrativa de Barretos
Administrativa de Bauru
Administrativa de Campinas
Administrativa de Franca
Administrativa de Itapeva
Administrativa de Marília
Administrativa de Presidente Prudente
Administrativa de Registro
Administrativa de Ribeirão Preto
Administrativa de Santos
Administrativa de São José do Rio Preto
Administrativa de São José dos Campos
Administrativa de Sorocaba
Metropolitana de São Paulo

(Fonte: adaptada de SEADE, 2016)

2.3. DESENVOLVIMENTO DO SISTEMA

Os dados obtidos, armazenados e que passaram por um processo inicial de preparação – etapa de pré-processamento – precisam ser analisados, pois o real valor desses dados reside na informação que se pode extrair deles, formando o conhecimento, essencial para o processo de tomada de decisão em uma organização. Além disso, esses dados cadastrais possuem somente parte da informação para a execução do processo de mineração de dados espacial, não possuindo uma das partes principais: as coordenadas geográficas. Para obter as coordenadas para cada empresa, e a fim de conseguir obter o conhecimento potencial que esses dados possuem, mas que até então permanecia inerte, foi desenvolvida uma aplicação que permite utilizar, filtrar e descobrir informações da base de dados pré-processada.

A tela inicial de abertura da aplicação pode ser observada na Figura 9.

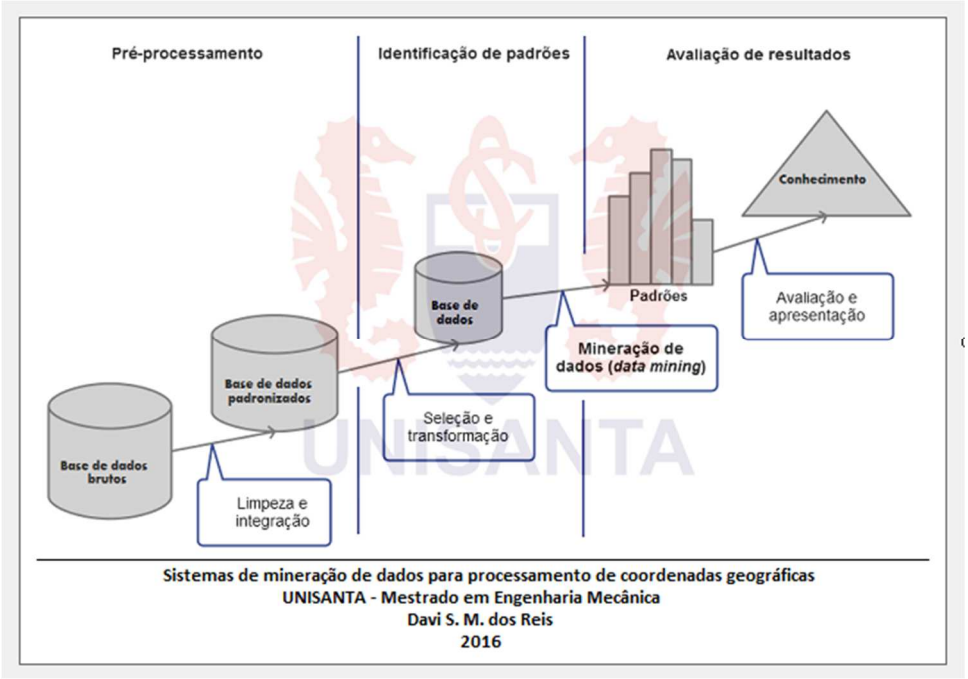

Figura 9 – Tela inicial (*Splash Screen*) do sistema.

Quando a aplicação está em execução, a primeira tela disponível é apresentada na Figura 10:

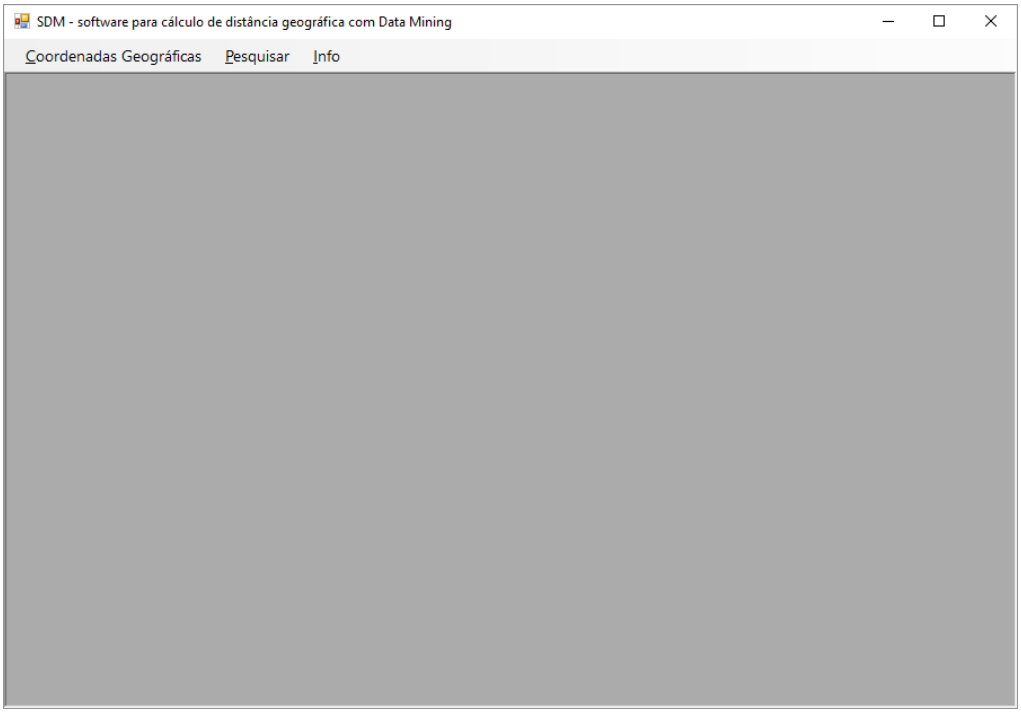

Figura 10 – Tela base da aplicação e seu menu.

Na opção de menu “Coordenadas Geográficas”, é possível acessar uma tela demonstrativa de exemplo, conforme apresentado na Figura 11:

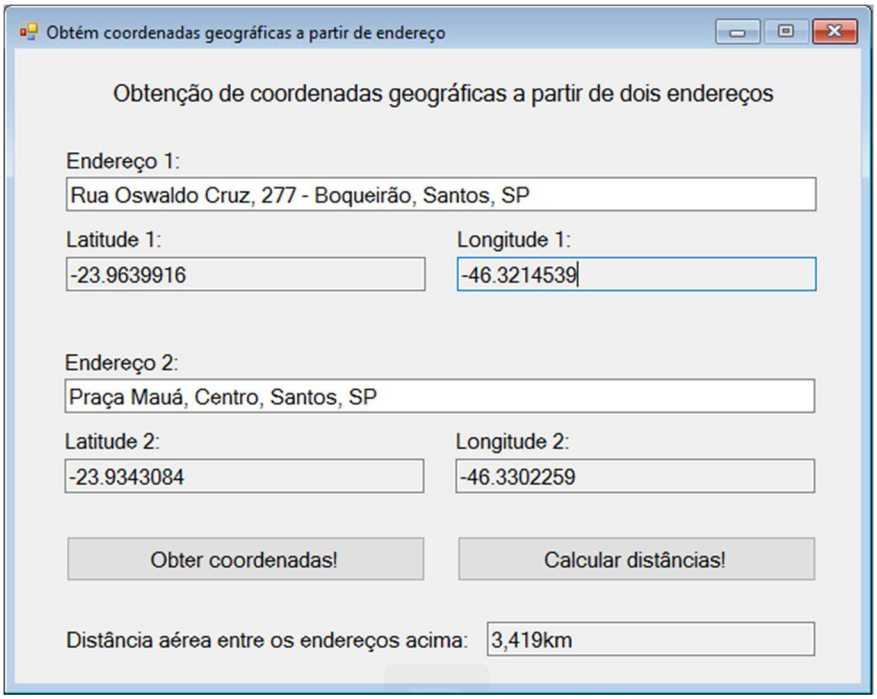

Figura 11 – Tela de exemplo para manipulação de coordenadas geográficas.

Na tela exibida Figura 11 é apresentado, em caráter de exemplo, os endereços da Universidade Santa Cecília (campo “Endereço 1”) e da Prefeitura Municipal de Santos (campo “Endereço 2”). Para fins de exemplo, ao clicar no botão “Obter coordenadas”, o aplicativo obtém as coordenadas geográficas de ambos os endereços, preenchendo-as nos campos de latitude e longitude dos respectivos endereços. De posse desses dados, é possível calcular a distância geográfica entre ambos endereços, clicando no botão “Calcular distâncias”.

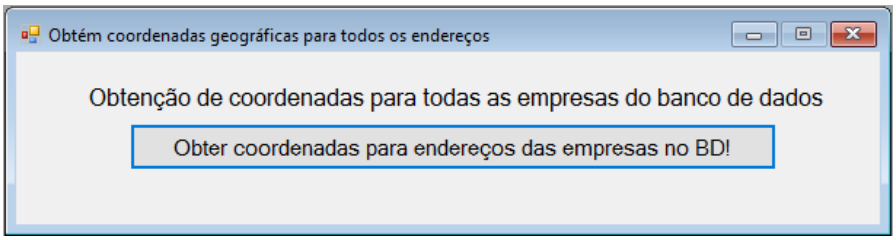

Figura 12 – Tela para obtenção de coordenadas geográficas.

Na tela exibida Figura 12 é apresentada a tela que permite a obtenção de coordenadas geográficas para todos os registros armazenados no banco de

dados do sistema. Uma vez que essa obtenção seja realizada, não é necessária executá-la novamente, pois as coordenadas geográficas obtidas são gravadas na tabela de empresas do banco de dados.

A obtenção das coordenadas geográficas a partir do endereço das empresas é realizada por intermédio de um serviço do Google Maps, explicado na seção 2.3.1 adiante.

2.3.1. API DO GOOGLE MAPS PARA COORDENADAS GEOGRÁFICAS

O Google Maps é um serviço gratuito fornecido pelo Google que, dentre outras funções, permite a realização de busca de endereços, coordenadas geográficas e visualização de mapas.

O Google Maps possui uma API[4] que permite o uso de seus recursos e sua incorporação em sites e aplicativos. A utilização da API também é gratuita, podendo ser utilizada mesmo em sites e aplicativos comerciais. A única limitação no uso gratuito da API é que esta permite apenas a execução de 2.500 consultas por dia. O Google fornece uma seção para desenvolvedores em seu site, contendo ampla documentação e exemplos de utilização de seus serviços e API.

O serviço da API do Google Maps que foi utilizado no desenvolvimento do projeto foi o que permite a obtenção das coordenadas geográficas a partir de um endereço fornecido.

O trecho de programação referente à utilização da API do Google Maps para obtenção das coordenadas geográficas é apresentado no Quadro 1:

[4] API – *Application Programming Interface*, ou Interface de Programação de Aplicações, é um conjunto de rotinas e funções prontas para serem utilizadas por aplicativos que não pretendem envolver-se em detalhes da implementação do software em questão, mas apenas utilizar suas funcionalidades.

```
// URL com o serviço de API do Google Maps para obtenção de coordenadas
string url = "https://maps.google.com/maps/api/geocode/xml?address=" + endereco +
"&sensor=false&key=AIzaSyC5aATF7YYuI3oOAKhMVOcPgSI0adHkk6I";

// cria objeto de requisição para solicitação de retorno dos dados da URL
WebRequest request = WebRequest.Create(url);
// cria objeto de DataTable para armazenar os retornos
DataTable dtbCoord = new DataTable();

// recupera o retorno do processamento da API do Google, para o endereço fornecido
using (WebResponse response = (HttpWebResponse)request.GetResponse())
{
  // configura objeto para leitura do retorno do processsamento
  using (StreamReader reader = new StreamReader(response.GetResponseStream(),
  Encoding.UTF8))
  {
    DataSet dsResult = new DataSet();
    // converte o resultado da leitura para um objeto DataSet
    dsResult.ReadXml(reader);
    // monta objeto DataTable para ler cada parâmetro retornado pelo Google Maps
    dtbCoord.Columns.AddRange(new DataColumn[4] { new DataColumn("Id", typeof(int)),
                    new DataColumn("Address", typeof(string)),
                    new DataColumn("Latitude",typeof(string)),
                    new DataColumn("Longitude",typeof(string)) });

    // algoritmo iterativo para preencher DataTable com cada parâmetro retornado
    foreach (DataRow row in dsResult.Tables["result"].Rows)
    {
      string geometry_id = dsResult.Tables["geometry"].Select("result_id = " +
      row["result_id"].ToString())[0]["geometry_id"].ToString();
      DataRow local = dsResult.Tables["location"].Select("geometry_id = " +
      geometry_id)[0];
      dtbCoord.Rows.Add(row["result_id"], row["formatted_address"], local["lat"],
      local ["lng"]);
    }
  }
}
```

Quadro 1 – Código-fonte de programação para obtenção das coordenadas.

De posse das coordenadas geométricas para cada empresa, a distância entre duas ou mais delas é calculada com a utilização da fórmula de Haversine, explicada previamente na seção 1.5.2.

O código-fonte com a fórmula de Haversine pode ser visto no Quadro 2. No mesmo quadro é possível visualizar, ainda, o código-fonte necessário para converter os valores de latitude e longitude – originalmente formatados em graus – para radianos, a fim de que seja possível realizar os cálculos e expressar seus resultados em quilômetros.

.

```
public double calcularDistancia(double latA, double longA, double latB,
double longB)
{
    // define valor para o raio da Terra
    double raioTerra = 6372.8;
    // diferença entre latitude final e inicial; o mesmo para longit.
    double deltaLat = GrausParaRadianos(latB - latA);
    double deltaLong = GrausParaRadianos(longB - longA);

    // realiza o cálculo principal da fórmula de Haversine
    double a = Math.Pow(Math.Sin(deltaLat / 2), 2) +
               Math.Cos(GrausParaRadianos(latA)) *
               Math.Cos(GrausParaRadianos(latB)) *
               Math.Pow(Math.Sin(deltaLong / 2), 2);
    // conclui o cálculo da fórmula
    double haversine = 2 * Math.Atan2(Math.Sqrt(a), Math.Sqrt(1 - a));
    // retorna o valor calculado, multiplicando pelo raio da Terra
    return raioTerra * haversine;
}

public double GrausParaRadianos(double graus)
{
    // cálculo para converter graus (medida original da latitude e
    // longitude) em radianos
    return (Math.PI / 180) * graus;
}
```

Quadro 2 – Código-fonte de programação com a função de Haversine.

2.3.2. PESQUISA E LISTAGEM DE EMPRESAS

A tela de pesquisa e listagem de empresas é apresentada na Figura 13. Nessa tela – que é a principal interface com o usuário que precisa de informações sobre a distância entre empresas para tomada de decisão –, é possível realizar diversos filtros, a fim de conseguir obter o melhor resultado.

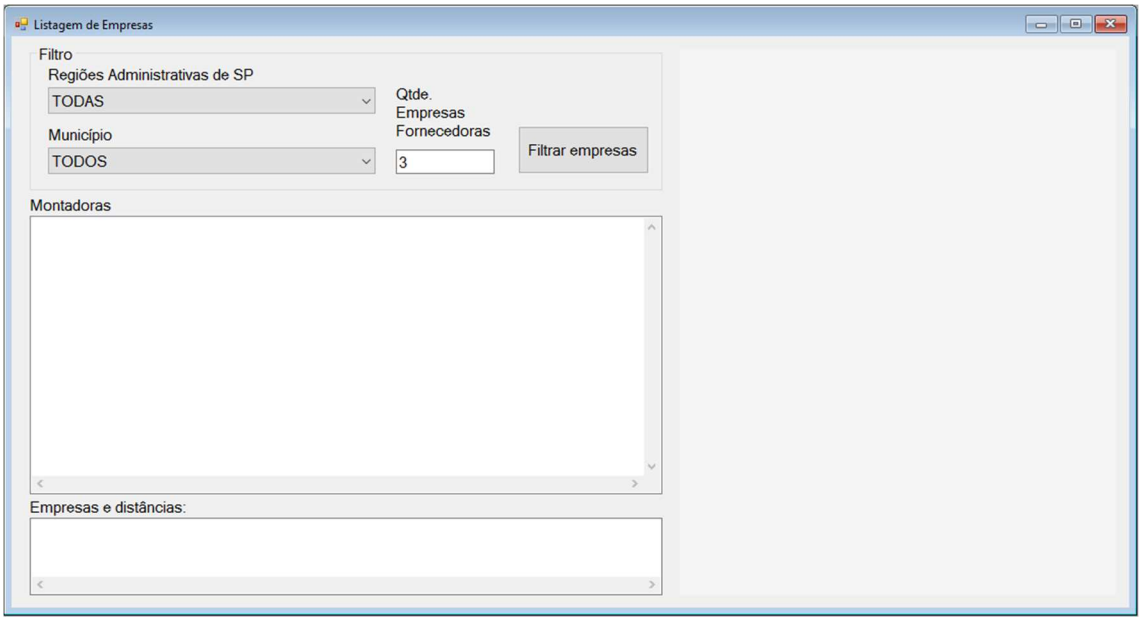

Figura 13 – Tela de pesquisa e listagem de empresas, com seus filtros.

Como exemplo de filtros que podem ser aplicados, a Figura 14 apresenta selecionada a Região Administrativa de Santos (também conhecida como Região Metropolitana da Baixada Santista) e, dentro desta, o município de Santos. Assim, após clicar no botão “Filtrar empresas”, são listadas as montadoras e empresas de manutenção de veículos de Santos.

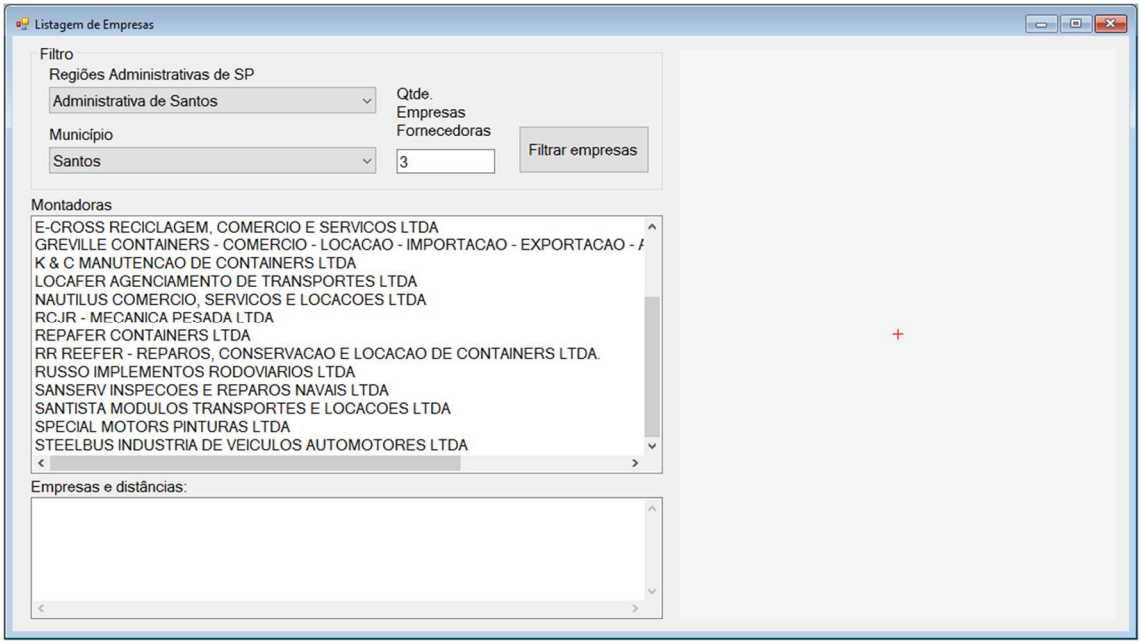

Figura 14 – Tela de pesquisa com filtro das empresas de Santos.

Após clicar no botão “Filtrar Empresas”, conforme descrito na Figura 14, a listagem de empresas montadoras foi obtida após a execução do comando SQL apresentado Quadro 3:

```
select distinct razao_social, e.municipio, latitude, longitude
  from tb_empresa e
 inner join tb_regiao_administrativa r
    on upper(e.municipio) = (r.municipio)
 where razao_social is not null
   and cnpj is not null
   and montadora_fornecedora = 'm'
   and r.regiao_adm like '%Administrativa de Santos%'
   and e.municipio like '%Santos%'
 order by razao_social, município;
```

Quadro 3 – Comando SQL para listagem de empresas em Santos.

Depois de obter todas as empresas montadoras, conforme descrito anteriormente, o usuário tem a opção de escolher quantas empresas fornecedoras ele pretende visualizar na listagem de empresas fornecedoras e também no mapa que será exibido com a indicação visual de cada empresa. Para tal opção, basta o usuário informar o número no campo “Qtde. Empresas Fornecedoras”, conforme apresentado na Figura 15. O valor inicial do campo está configurado para 3 empresas, podendo ser alterado antes de clicar no botão “Filtrar empresas”.

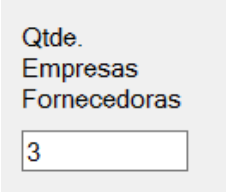

Figura 15 – Filtro de Quantidade de Empresas.

2.4. TESTES REALIZADOS

Após o desenvolvimento do programa, foram realizados diversos testes para verificar a validade da proposição do presente trabalho. Dentro desses testes, houve variação da região administrativa selecionada, bem como da cidade escolhida, da quantidade de empresas a serem visualizadas na resposta e, por fim, das empresas montadoras selecionadas.

Algumas das variações dos testes realizados são apresentadas nas Tabelas 4, 5, 6 e 7 a seguir, para as quais os resultados são mostrados no capítulo 3:

Tabela 4 – Parâmetros de Pesquisa para Mercedes-Benz do Brasil.

Parâmetro	Valor
Região Administrativa	Metropolitana de São Paulo
Município	São Bernardo do Campo
Qtde. de Empresas Fornecedoras	3
Montadora Escolhida	Mercedes-Benz do Brasil

Tabela 5 – Parâmetros de Pesquisa para Gil Equipamentos Industriais Ltda.

Parâmetro	Valor
Região Administrativa	Administrativa de Ribeirão Preto
Município	Ribeirão Preto
Qtde. de Empresas Fornecedoras	3
Montadora Escolhida	Gil Equipamentos Industriais Ltda.

Tabela 6 – Parâmetros de Pesquisa para AllTech Veículos Especiais Ltda.

Parâmetro	Valor
Região Administrativa	Metropolitana de São Paulo
Município	TODOS
Qtde. de Empresas Fornecedoras	3
Montadora Escolhida	AllTech Veículos Especiais Ltda.

Tabela 7 - Parâmetros de Pesquisa para Carrocerias Torrezan Ltda.

Parâmetro	Valor
Região Administrativa	Metropolitana de Sorocaba
Município	TODOS
Qtde. de Empresas Fornecedoras	3
Montadora Escolhida	Carrocerias Torrezan Ltda.

3. RESULTADOS E DISCUSSÕES

Após realizar o filtro das empresas montadoras, conforme explicado na seção 2.3.2 – e utilizando os valores descritos na seção 2.4 –, o usuário que necessita saber quais empresas fornecedoras estão mais próximas da montadora escolhida precisa realizar apenas uma tarefa simples: clicar na empresa montadora para a qual deseja obter os resultados, como descrito na seção 3.1.

3.1. RESULTADOS

Após realizar os filtros conforme descritos na seção 2.3.2, o resultado das empresas mais próximas é apresentado no campo “Empresas e distâncias”, em ordem crescente da mais próxima para a mais distante, bem como a distância em quilômetros de cada fornecedora até a montadora. Ao mesmo tempo, cada uma dessas empresas fornecedoras é apresentada, de maneira agrupada (“*clusterizada*”), no mapa interativo ao lado direito da tela, conforme apresentado na Figura 16 (em resposta aos dados de filtro de pesquisa apresentados na Tabela 4):

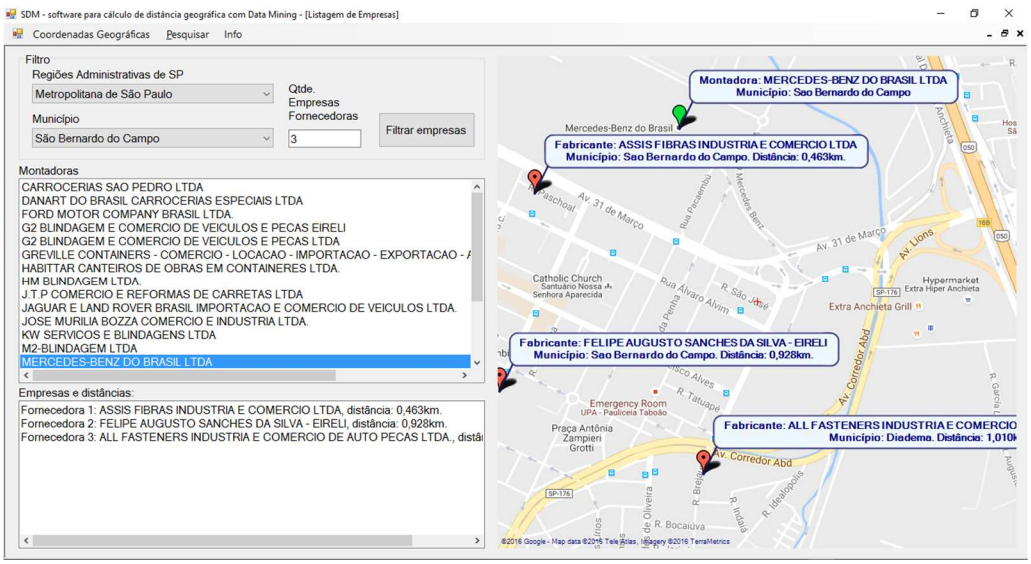

Figura 16 – Fornecedores mais próximos à empresa Mercedes-Benz.

Considerando que a base de dados disponível não possui uma listagem sobre os itens que de fato são produzidos e comercializados por cada empresa –

somente a atividade econômica à qual cada empresa está vinculada –, a informação da menor distância entre montadoras e fornecedoras é a informação disponível mais relevante e passível de ser “minerada” a partir dos dados originalmente fornecidos.

Segue, na Figura 17, os resultados apresentados para os filtros de pesquisa dos testes da Tabela 5:

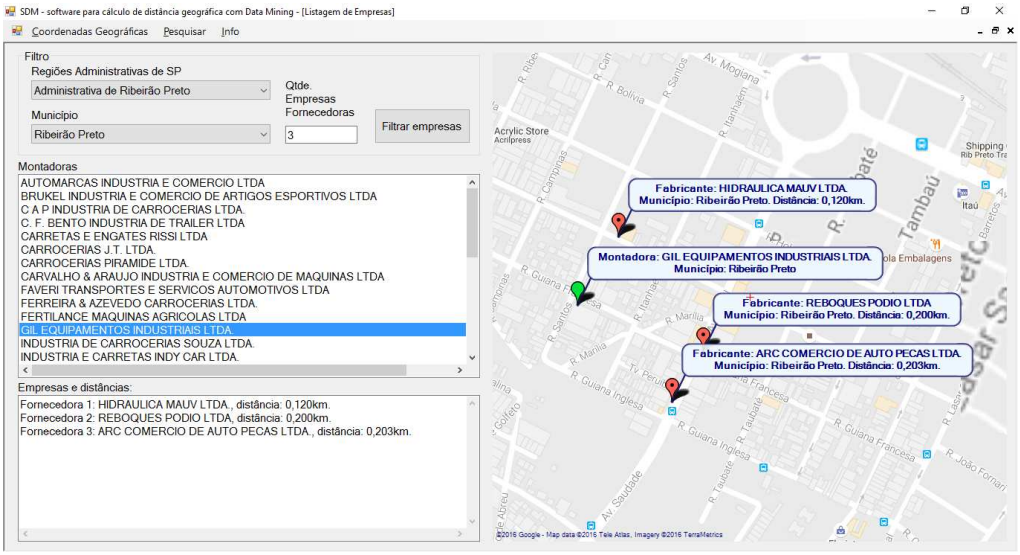

Figura 17 – Fornecedores mais próximos à empresa Gil Equipamentos Industriais Ltda.

Na Figura 18 são apresentados os resultados para os filtros de pesquisa dos testes da Tabela 6:

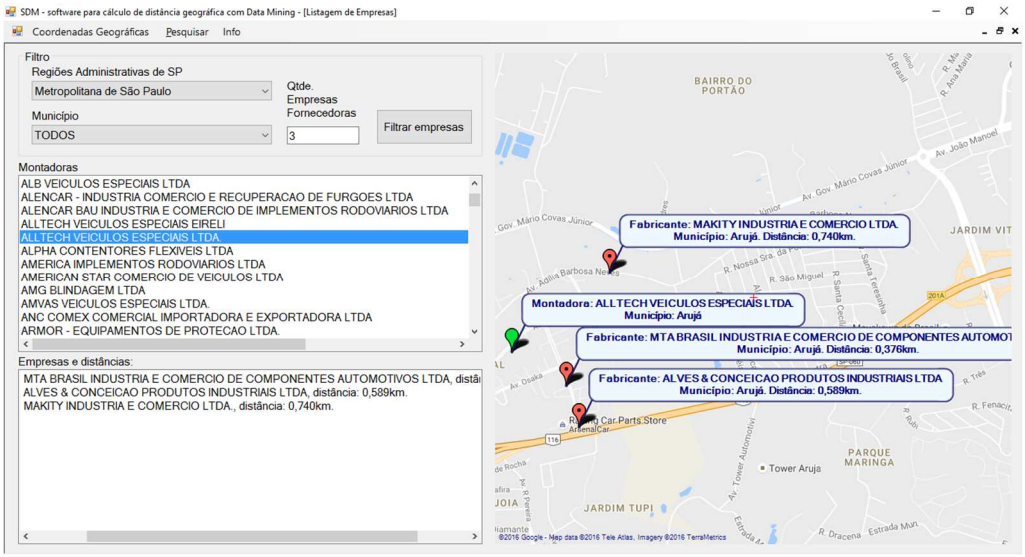

Figura 18 – Fornecedores mais próximos à empresa AllTech Veículos Especiais Ltda.

O usuário do sistema, ao fazer uso do aplicativo para encontrar as empresas fornecedoras mais próximas de uma empresa montadora, não precisa despender de tempo para encontrar os endereços e calcular as distâncias para localizar as empresas mais próximas: basta realizar o filtro da região administrativa e, eventualmente, do município, para que consiga encontrar as montadoras e, a partir dessa listagem, as “n” fornecedoras mais próximas para a empresa selecionada – sendo “n” um número facilmente configurável na área de filtros da tela de listagem de empresas, conforme apresentado na seção 2.3.2 (Figura 15). Tal operação, se executada de forma manual, demandaria uma grande quantidade de tempo, além de ser um trabalho custoso.

Outro ponto de destaque na utilização do aplicativo diz respeito ao fato de não haver limitação de município para a localização das fornecedoras mais próximas: uma vez que o cálculo utiliza coordenadas geográficas e a fórmula de Haversine, as empresas mais próximas serão localizadas de maneira absoluta, como é possível visualizar na Figura 19, onde parte das empresas resultantes são do mesmo município da montadora e outra parte localiza-se em outro município – dados de pesquisa dos testes da Tabela 7.

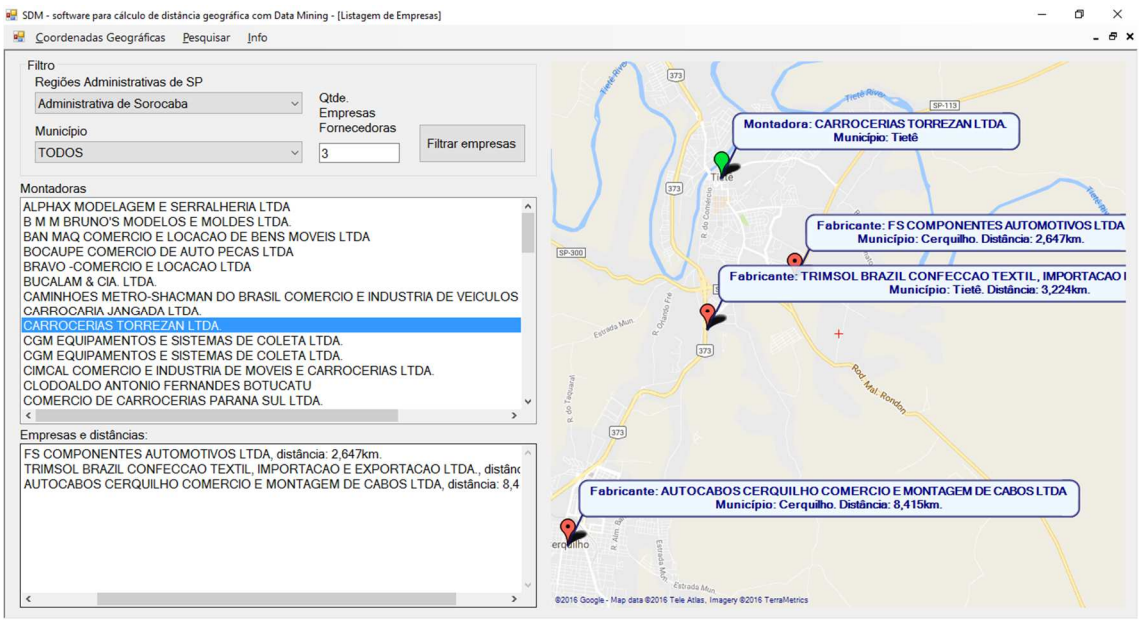

Figura 19 – Exemplo: fornecedores em municípios diferentes.

Tais resultados são possíveis graças à mineração de dados, que permite perscrutar massas de dados volumosas no intuito de extrair conhecimento desses dados que, à primeira vista, repousam inertes.

Afinal, o conhecimento é o padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações (GOLDSCHMIDT; PASSOS, 2005). E esse relacionamento é possível graças ao entendimento correto da necessidade dos envolvidos num processo decisório e da aplicação da ferramenta apropriada para extração de informação e conhecimento dos dados disponíveis nas bases de dados disponíveis para análise.

3.2. DISCUSSÕES

Ainda que a experiência humana, no papel do analista conhecedor das particularidades do negócio em questão, seja necessária para poder avaliar quais empresas são as mais relevantes para o cenário proposto – a fim de que possa haver, de fato, a escolha de qual empresa poderá atuar como fornecedora de insumos para uma montadora –, o emprego da tecnologia e da descoberta de conhecimento em bases de dados (ou KDD, como explicado na seção 1.3) é de suma importância para auxiliar no processo decisório, visto que ambos fornecem subsídios para permitir a obtenção de conhecimento que permita a escolha das opções disponíveis maneira assertiva.

Tais ferramentas – a tecnologia da informação e a execução de um processo de KDD –, sendo utilizadas em conjunto para compor um sistema inteligente, como apresentado a partir da seção 2.3, permitem explorar uma base de dados com milhares de informações e, com apenas poucos cliques e interações com o usuário, apresentar resultados de forma rápida e confiável, conforme foi possível observar nas figuras da seção 3.1.

4. CONCLUSÃO

À primeira vista, aos olhos menos atentos, a tarefa de analisar dados pode parecer simples. Entretanto, trata-se de uma atividade difícil, principalmente quando envolve uma grande quantidade de dados. Afinal, a ação humana demonstra-se limitada quando exigida de forma massiva, uma vez que sofre influência do cansaço e da exaustão – dentre outros fatores externos –, o que pode comprometer seu desempenho e a assertividade em suas respostas. Esse cenário é agravado quando o trabalho de análise envolve dados diversos e em grandes quantidades, ou quando há etapas que se relacionam inúmeras vezes – o que ocorre com frequência quando se trata de grandes massas de dados. Portanto, é de grande importância ter o apoio de sistemas inteligentes e apropriados para a necessidade a ser “minerada”, tornando o tratamento e a exploração dos dados mais simplificados e assertivos para obtenção de conhecimento.

O presente trabalho apresentou, de maneira teórica e prática, o grande potencial do processo de exploração de dados denominado Descoberta de Conhecimento em Base de Dados (KDD), conseguindo extrair, de uma vasta massa de dados bruta e inerte, com dados aparentemente desconexos, informações importantes para o processo decisório de escolher empresas fornecedoras de materiais e insumos a empresas compradoras – neste caso, baseando-se apenas na proximidade e no tipo de atividade econômica dos potenciais fornecedores, visto que a base de dados obtida não continha informações de produtos comercializados. Cumpre informar que se houvesse dados a respeito dos produtos fabricados por cada empresa, o sistema também seria capaz de cruzá-los, analisá-los e processá-los.

Assim, o resultado da pesquisa e do desenvolvimento do trabalho, na forma de um sistema de mineração de dados para obtenção de menores distâncias entre empresas fornecedoras e montadoras da indústria automotiva, demonstrou ter capacidade para ser uma importante ferramenta de apoio no processo decisório de escolha de fornecedores, apresentado parte do potencial da área de descoberta de conhecimento em bases de dados, em geral, e da área de mineração de dados, em particular, atingindo os objetivos propostos.

Cumpre informar, por fim, que o programa de computador ora desenvolvido e utilizado para demonstrar parte do potencial do *data mining* pode ser utilizado com diversos outros tipos de dados – por exemplo, dados relacionados às peças produzidas, ao custo de cada produto, à demanda por cada item comercializado, dentre outros –, além de ser capaz de realizar o mesmo tipo de análise para outras áreas de atividades econômicas (por exemplo, distribuição da população de um município, dados relacionados à saúde pública ou incidência de uma determinada doença, informações sobre potenciais candidatos a egressos de uma instituição de ensino, enfim), desde que, neste caso, os dados a serem trabalhados sejam devidamente preparados para serem armazenados nas tabelas de bancos de dados descritas na seção 2.2, o que demonstra a versatilidade, a utilidade e a alta empregabilidade da mineração de dados e da ferramenta desenvolvida.

Como trabalhos futuros, pretende-se adaptar o sistema inteligente ora desenvolvido para permitir que sejam utilizados dados de outras empresas, os quais serão passíveis de serem importados para as tabelas de bancos de dados descritas na seção 2.2. Assim, poderão ser exploradas outras áreas de atividades econômica, de acordo com os dados disponíveis.

REFERÊNCIAS BIBLIOGRÁFICAS

ARNOLD, J. R. T. **Administração de Materiais: Uma Introdução**. 1. ed. São Paulo: Atlas, 1999.

AYRES, A. DE P. S. **Gestão de Logística E Operações**. 1. ed. Curitiba: IESDE, 2009.

BERSON, A.; SMITH, S.; THEARLING, K. **Building Data Mining Applications for CRM**. New York: MacGrawHill, 1999.

CORDEIRO, R. L. F.; FALOUTSOS, C.; JÚNIOR, C. T. Data Mining in Large Sets of Complex Data. 2013.

DATE, C. J. **Introdução a sistemas de bancos de dados**. 8a. ed. Rio de Janeiro: Elsevier Brasil, 2004.

DRUCKER, P. F. O Advento da Nova Organização. In: ELSEVIER BRASIL (Ed.). . **Gestão Do Conhecimento**. Rio de Janeiro: Harvard Business Review, 2001.

ELMASRI, R.; NAVATHE, S. **Conceitos de Data Mining**. São Paulo: Pearson Addison Wesley, 2005.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, p. 37–54, 1996.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge Discovery in Databases : An Overview. **AI Magazine**, v. 13, n. 3, p. 57–70, 1992.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: Um Guia Prático – Conceitos, técnicas, ferramentas, orientações e aplicações**. Rio de Janeiro: Campus, 2005.

HARRISON, T. H. **Intranet Data Warehouse - Ferramentas e técnicas para a utilização do data warehouse na Intranet**. São Paulo: Berkeley, 1998.

IBGE, I. B. DE G. E E. **CONCLA - Comissão Nacional de Classificação**. Disponível em: <http://www.cnae.ibge.gov.br/>. Acesso em: 10 jul. 2016.

KIMBALL, R. **Data Warehouse Toolkit - técnicas para construção de data warehouses dimensionais**. [s.l.] Makron Books, 1998.

LAGARINHOS, C. A. F.; TENÓRIO, J. A. S. **Modelo de Logística Reversa com a Utilização do Algoritmo Genético**Congresso Brasileiro de Engenharia e Ciência dos Materiais. **Anais**...2012

LAROSE, D. T. **Discovering knowledge in data**. New Jersey: John Wiley and Sons, 2005.

MARTINS, P. G.; ALT, P. R. C. **Administração de Materiais e Recursos Patrimoniais**. 3. ed. São Paulo: Saraiva, 2011.

MATHIAS, L. A. F. **Mineração De Dados Em Sistemas De Energia Elétrica Utilizando Algoritmos Fundamentados Em Lógica Paraconsistente Anotada - LPA**. [s.l.] Universidade Santa Cecília, 2015.

MOUNT, D. W. **Bioinformatics: Sequence and Genome Analysis**. [s.l.] University of Arizona, 2004.

NEVES, M. C.; FREITAS, C. C.; CÂMARA, G. **Mineração de Dados em Grandes Coleções de Imagens**. [s.l: s.n.].

NIMER, F.; SPANDRI, L. C. Data Mining. **Revista Developers**, p. 32, fev. 1998.

NORDIN, N. A. M. et al. Finding Shortest Path of the Ambulance Routing : Interface of A * Algorithm using C # Programming. p. 1569–1573, 2012.

OLIVEIRA, R. R.; CARVALHO, C. L. **Algoritmos de agrupamento e suas aplicações**. [s.l: s.n.].

POSSAS, B. A. V. et al. **Data Mining: Técnicas para Exploração de Dados**. [s.l.] Universidade Federal de Minas Gerais, 1998.

REATEGUI, E. **Data Mining e Personalização Dinâmica** (L. Nedel, Ed.)Anais da X Escola de informática da SBC-Sul. **Anais**...Criciúma: Sociedade Brasileira de Computação, 2002

REZENDE, S. O. **Sistemas Inteligentes: Fundamentos e Aplicações**. 1. ed. Barueri: Manole, 2003.

RODRIGUES, A. M. **Escavando Dados no Varejo**. Rio de Janeiro: COPPEAD - Universidade Federal do Rio de Janeiro, 2000.

SALVADOR, H. G.; MARQUES, A.; CUNHA, D. A. Vedalogic um método de Verificação de Dados Climatológicos Apoiado em Modelos Minerados. p. 448–460, 2009.

SEADE, F. S. E. DE A. DE D. **Regiões Administrativas do Estado de SP**. Disponível em: <http://produtos.seade.gov.br/produtos/divpolitica/index.php?page=tabela&action=load&nivel=30>. Acesso em: 12 nov. 2016a.

SEADE, F. S. E. DE A. DE D. **Regiões Administrativas do Estado de SP**.

SEIXAS, J. A. DE. **Integração De Mineração De Dados E Visualização De Informações Geográficas Aplicados À Saúde Pública**. [s.l: s.n.].

SILVA, E. M. **Avaliação do Estado da Arte e Produtos Data Mining**. Brasília: Universidade Católica de Brasília, 2000.

SILVA, T. Data mining de dados geo-temporais para suporte à mobilidade. 2012a.

SILVA, T. **Data mining de dados geo-temporais para suporte à mobilidade**. [s.l.] Faculdade de Engenharia da Universidade do Porto, 2012b.

WEISS, S. M.; INDURKHYA, N. **Predictive Data Mining: A Practical Guide**. 1. ed. San Francisco: Academic Press, 1998.