UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

CURSO DE GRADUAÇÃO DE ENGENHARIA DE PRODUÇÃO

JOÃO VITOR SOUBHIA

APLICAÇÃO DA METODOLOGIA KDD NA DESCOBERTA DE CONHECIMENTO EM DADOS RELACIONADOS AO SETOR DE MANUTENÇÃO DE UMA LOCADORA DE VEÍCULOS

LONDRINA
2018

JOÃO VITOR SOUBHIA

APLICAÇÃO DA METODOLOGIA KDD NA DESCOBERTA DE CONHECIMENTO EM DADOS RELACIONADOS AO SETOR DE MANUTENÇÃO DE UMA LOCADORA DE VEÍCULOS

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Produção como requisito parcial para obtenção do título de Engenheiro de Produção, do curso de Bacharelado em Engenharia de Produção da UTFPR - Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Me. Bruno Samways dos Santos

LONDRINA
2018

TERMO DE APROVAÇÃO

APLICAÇÃO DA METODOLOGIA KDD NA DESCOBERTA DE CONHECIMENTO EM DADOS RELACIONADOS AO SETOR DE MANUTENÇÃO DE UMA LOCADORA DE VEÍCULOS

POR

JOÃO VITOR SOUBHIA

Esta Monografia foi apresentada às 14 horas do dia 26 de Novembro de 2018 como requisito parcial para obtenção do título de bacharel em ENGENHARIA DE PRODUÇÃO, Universidade Tecnológica Federal do Paraná – Campus Londrina. O candidato foi arguido pela Banca Examinadora composta pelos professores relacionados abaixo. Após deliberação, a Banca Examinadora considerou o trabalho: **APROVADO**.

Prof. Dr. Rafael Henrique Palma Lima (UTFPR)
Banca Examidadora

Prof. Me. Eduardo Jose Pitelli (UTFPR)
Banca Examidadora

Prof. Me. Bruno Samways dos Santos (UTFPR)
Presidente da Banca Examinadora
Orientador

SOUBHIA, João V. **APLICAÇÃO DA METODOLOGIA KDD NA DESCOBERTA DE CONHECIMENTO EM DADOS RELACIONADOS AO SETOR DE MANUTENÇÃO DE UMA LOCADORA DE VEÍCULOS**. 2018. 43 f. Trabalho de conclusão de curso (Bacharelado em Engenharia de Produção) – UTFPR – Universidade Tecnológica Federal do Paraná. Londrina, 2018.

RESUMO

Visando que os processos de manutenção têm grande influência na rentabilidade das locadoras de veículos, o presente trabalho objetivou analisar características deste setor a partir da análise de agrupamento (ou clusterização). Para tal, foi aplicada a técnica *k-means* na qual os principais atributos da manutenção de veículos foram separados em diferentes grupos (clusters) então foram verificadas as similaridades e diferenças entre os atributos dos grupos. Neste contexto, esse trabalho contribuiu para um melhor entendimento das variáveis que influenciam nos processos de manutenção, permitindo agrupá-las de modo que possibilitou uma melhor visibilidade para a gestão da manutenção. Com isso, foram propostas novas medidas na empresa, com intuito de direcionar os investimentos de forma eficiente, melhorando a política de aplicação dos recursos do setor, minimizando o risco de custos desnecessários e melhor gerindo os custos variáveis.

Palavras-chave: Mineração de dados; *k-means;* KDD; aprendizado não supervisionado; manutenção de veículos.

SOUBHIA, João V. **APPLICATION OF THE KDD METHODOLOGY IN THE DISCOVERY OF KNOWLEDGE IN DATA RELATED TO THE MAINTENANCE SECTOR OF A VEHICLE LESSOR**. 2018. 43 f. Completion of course work (Bachelor of Engineering of Production) - UTFPR - Universidade Tecnológica Federal do Paraná. Londrina, 2018.

ABSTRACT

Aiming that the maintenance processes have a great influence on the profitability of the car rental companies, the present work aimed to analyze characteristics of this sector from the grouping (or clustering) analysis. To that end, the k-means technique was applied in which the main vehicle maintenance attributes were separated into different groups (clusters), so the similarities and differences between the attributes of the groups were verified. In this context, this work contributed to a better understanding of the variables that influence the maintenance processes, allowing them to be grouped in a way that allowed a better visibility for maintenance management. With this, new measures were proposed in the company, with the purpose of directing the investments in an efficient way, improving the policy of application of the resources of the sector, minimizing the risk of unnecessary costs and better managing the variable costs.

Key words: Data mining; *k-means;* KDD; non-supervised learning; vehicle maintenance.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O mercado de locadoras de veículos tem crescido significativamente no Brasil, segundo a ABLA (Associação Brasileira de Locadoras de Automóveis) as empresas locadoras de automóveis cresceram 12,3% em faturamento bruto, aumentaram a receita líquida em 11,6% e o número de clientes cresceu 17,2% de 2016 para 2017 (OLIVEIRA, 2017). Nesse contexto, ocorreu uma “guerra de preços”, onde as empresas líderes estão com preços cada vez mais competitivos, comparando-se aos preços de aluguel de carros em 1994. Com isso, torna-se necessário utilizar ferramentas inovadoras para se manter no mercado de alta concorrência e com grandes oportunidades de crescimento.

Para que uma companhia possa se destacar em meio a um cenário em que novos empreendimentos são criados todos os dias é necessário conhecer minuciosamente todos os aspectos referentes ao seu funcionamento, possibilitando uma visibilidade mais ampla para tomada de decisão. Utilizada como uma das principais matérias primas nas organizações, a informação pode ser comparada à energia que alimenta um sistema. Na busca do conhecimento a tecnologia deve ser empregada com efetividade, rapidez e qualidade, somando fatores decisivos para estar à frente. Ainda nesse contexto, tem-se o mercado de tecnologia da informação (TI) com crescimento de 7,7% comparado 2015 a 2016 (BASSANETO, 2017), ou seja, da maneira que o mercado evolui se manter numa boa posição perante os adversários se torna mais desafiador.

Quando se trata de TI e aprendizado para gestão empresarial, há um processo que se destaca dentre os outros. A descoberta de conhecimento em base de dados (*Knowledge Discovery in Databases* - KDD) é uma metodologia de grande relevância para exploração de bases de dados. Contudo, tal aplicação é complexa quanto à percepção e interpretação de padrões não triviais nos dados

Dado que o setor de manutenção tem grande influência na rentabilidade das locadoras de veículo, vale ressaltar que é necessária uma gestão apropriada para identificar pontos de melhoria e melhor gerenciamento de recursos. Neste contexto, tem-se o problema a ser investigado: Como aplicar a metodologia KDD para a descoberta de conhecimento em base de dados relacionados ao setor de manutenção de uma locadora de veículos?

1.1 OBJETIVO GERAL

Avaliar a aplicação do KDD no setor de manutenção automotiva para identificar oportunidades que propiciem o aumento da eficiência no setor.

1.2 OBJETIVOS ESPECÍFICOS

- Geração de indicadores para dar suporte à tomada de decisão;
- Aplicar metodologia KDD;
- Analisar estatisticamente as diferenças;
- Propor melhorias no setor de gestão da manutenção.

1.3 JUSTIFICATIVA

A manutenção em geral, seja em veículos ou em máquinas, tem como objetivo atingir um nível de continuidade no processo produtivo, assim como as grandes indústrias fazem a gestão da manutenção juntamente com suas atividades produtivas, antecipando os possíveis reparos necessários. Esse processo tem origem atrelada ao problema em ter um plano de contingência (prevenção), com procedimentos interligados à necessidade de manutenção. Com isso, percebe-se que esse conceito evolui principalmente devido a evolução da indústria (FEDELE, 2011). Ainda segundo Fedele (2011), os computadores e softwares utilizados na gestão da manutenção tem um papel fundamental no monitoramento dessa etapa, garantindo eficiência na coleta de dados e mensuração dos atributos necessários (km rodado, idade do veículo, nº histórico de manutenções). Com a crescente evolução da TI é possível a aquisição de dados em tempo real e, consequentemente, gerar análises da informação obtida.

Segundo McDowell (1991), a utilização de softwares, sistemas informatizados e outras variações de recursos da tecnologia da informação (inteligência instalada nos veículos para detecção de necessidade de manutenção, por exemplo), que crescem dia após dia, trazem uma carência de treinamento e estudo na mão-de-obra aplicada a esse processo. Portanto, aqui se cabe a aplicação de testes individuais (análise sobre tipo de veículo, qualidade aplicada na manutenção, gastos e tipos de manutenção) visando evidenciar possíveis deficiências presentes na mão-de-obra ou no processo definido pela empresa.

Dada a alta no mercado de locação de veículos, crescente evolução e busca por novas tecnologias nas empresas do ramo e, consequentemente, na disputa por preços e posições, é necessário um diferencial para ter êxito. A ótima gestão empresarial mostra-se extremamente necessária, visto que há uma carência pela estratégia no mercado para sobrevivência e disputa nesse cenário. Sem o devido conhecimento a respeito dos processos internos e uma qualidade na transformação de dados em informações úteis, não se entende claramente qual a dimensão e as limitações do negócio. Portanto, percebe-se que não há uma visão clara nos pontos de melhoria e oportunidades de corte de gastos na empresa, ou seja, há uma evidente carência de gestão devido à falta de informação de qualidade.

1.4 ESTRUTURA DO TRABALHO

Após o presente capítulo 1 apresentado, o trabalho segue a seguinte estrutura: No capítulo 2, foi levantado fundamentação teórica a respeito dos processos de KDD, manutenção e análises estatísticas. No capítulo 3 é contextualizado o sistema de negócio da locadora e levantado o problema do estudo. No capítulo 4 foram descritos, tanto a pesquisa, quanto os materiais e métodos utilizados, contando com os procedimentos de coleta e análise de dados, além dos softwares utilizados. No capítulo 5 são apresentados os resultados das análises e as conclusões sobre as mesmas. O capítulo 6 é um breve compilado e conclusão final do trabalho.

2 REFERENCIAL TEÓRICO

O presente capítulo apresenta os conceitos sobre a descoberta de conhecimento em bases de dados, a descrição das técnicas que serão utilizadas no desenvolvimento da pesquisa e algumas definições importantes quanto às características do setor de manutenção.

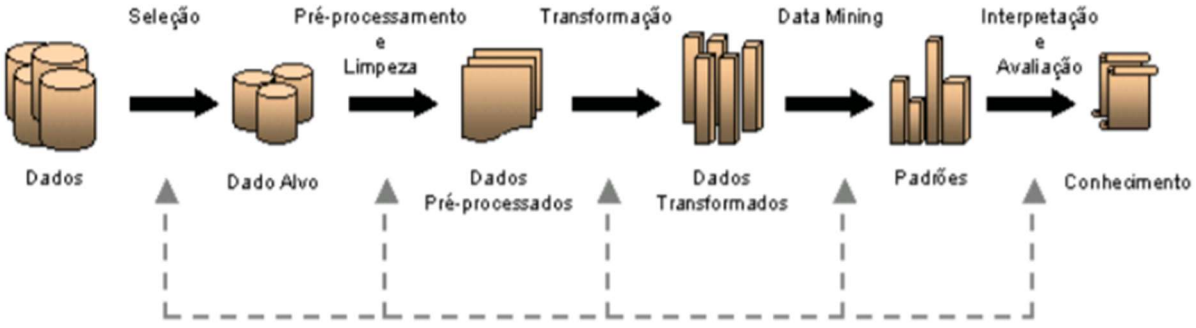
2.1 *KNOWLEDGE DISCOVERY IN DATABASES* – KDD

Segundo Fayyad et al. (1996, p. 40-41), “KDD é um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados”.

Para Goldschimidt e Passos (2005), o termo “iterativo” sugere que possam haver repetições integrais ou parciais, enquanto que a expressão “não trivial”, implica que há certa complexidade no processo, envolvendo métodos específicos na identificação de padrões. O termo “dados” trata-se de itens captados e armazenados, sem significância e fora de contexto e “padrões” refere-se a um subconjunto de dados que descrevem uma situação (por exemplo, “lojas com mais produtos tendem a vender mais” ou “regiões com solos mais férteis tem melhor produtividade agrícola”). Por último, quando se refere a “padrão válido” indica-se que a informação deve ser real e adequada ao contexto de aplicação da KDD e os conhecimentos devem ser “novos e potencialmente úteis” trazendo algum resultado que agregue ao propósito do estudo.

De acordo com Fayyad et al. (1996), o processo KDD pode ser divido em cinco etapas: seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; Mineração de Dados (*Data Mining* – DM, principal etapa do processo KDD); interpretação e avaliação dos resultados. A Figura 1 ilustra o processo.

Figura 1 - Fases do Processo de KDD



Fonte: Adaptação de FAYYAD et al. (1996).

Nos seguintes itens serão descritas mais detalhadamente as fases mostradas na figura 1.

A. SELEÇÃO DE DADOS

Na fase de seleção, primeira fase do processo de descoberta de informação, serão escolhidos os conjuntos de dados, contendo todas as variáveis (atributos que influenciam nos padrões) que farão parte da análise. A seleção dos dados pode ser complexa, quando o banco de dados é extenso ou possui fontes diferenciadas. É importante garantir uma coleta de qualidade, pois são a partir desses dados que as conclusões posteriores tomarão forma.

B. PRÉ PROCESSAMENTO E LIMPEZA DE DADOS

Nesta fase utilizam-se técnicas para eliminar dados redundantes e/ou inconsistentes, restaurar dados incompletos avaliar possíveis dados discrepantes (*outliers,* que interferem de maneira inadequada ao processo). Para esta fase é necessário um conhecimento ainda mais aprofundado sobre a base de dados tanto quanto sobre técnicas de limpeza de dados, pois assim como a coleta, esse processo influenciará totalmente nas informações que serão descobertas.

Uma das técnicas de pré-processamento é a de substituição de valores faltantes (*missing values),* esse processo tem o intuito de tratar os dados que estão faltando na amostra (e devem ser preenchidos), seja excluindo-os, estimando-os ou apenas ignorando as informações faltantes. Por exemplo numa pesquisa com um grupo de pessoas, em alguns casos o entrevistado pode recusar-se a falar sua idade,

resultando em uma informação necessária ficando em branco (nesse caso pode ser utilizada uma exclusão ou estimação do atributo falho). Outra técnica usada nessa etapa do KDD é a detecção de *outliers*, que trata valores que se diferenciam muito do restante da amostra. Quando se identifica um valor anômalo, deve-se tratar a instância em que ele aparece e, dependendo do tamanho da amostra, normalmente é interessante excluir um *outlier* do restante da base para ter uma informação mais precisa (TAN;STEINBACH; KUMAR, 2009).

C. TRANSFORMAÇÃO DE DADOS

Após a coleta e pré-processamento dos dados, estes precisam ser formatados adequadamente de acordo com o que é mais relevante para o aprendizado que se deseja obter. A transformação de variáveis ocorre quando é preciso alterar um atributo visando facilitar o usa das técnicas de mineração de dados, por exemplo, se na base de dados em uma pesquisa de uma vegetação um dos atributos for o “número de folhas”, mas para o estudo fosse mais relevante o “número de folhas por galho”, então deve-se transformar o atributo inicial, dividindo pelo número de galhos. Tan, Steinbach e Kumar (2009) retratam alguns exemplos mais comuns de transformação de dados:

Um dos processos que podem ser aplicados nessa fase é a “função simples”, que é a simples aplicação de uma função à variável que deve ser transformada. Se a variável for x , pode-se transformá-la através da função simples em $\sqrt{x}, e^x, \frac{x}{2}$, etc. Um exemplo de transformação por função simples é uma base de dados onde um dos atributos é “número de bytes” e esse número varia de um a um milhão. Nesse caso, seria vantajoso utilizar um log_{10} na transformação, por ser um atributo com amplitude muito grande.

Outro exemplo de transformação é a “normalização”, onde o objetivo desse processo é fazer com que uma série de dados tenha uma propriedade particular, muito comum quando se utiliza a normalização de variáveis na estatística:

“Se $\bar{x}$ é a média dos valores do atributo e s_x é o seu desvio padrão, então a transformação $x' = (x - \bar{x})/s_x$ cria uma nova variável que tem mediana 0 e desvio padrão 1” (TAN; STEINBACH; KUMAR, 2009, p. 64).

D. MINERAÇÃO DE DADOS

Como citado anteriormente, todas as etapas do processo KDD são necessárias e importantes para o êxito no mesmo, porém é na fase de mineração de dados (do inglês, *Data Mining* - DM) que se define qual será o algoritmo utilizado na descoberta do conhecimento (regressão, classificação, clusterização ou associação). A mineração é o processo de análise mais detalhado e abrange amplos métodos para buscar a informação necessária. Segundo Fayyad et al. (1996), DM é uma das etapas do KDD, onde tem-se uma descoberta de padrões únicos a partir da aplicação de algoritmos específicos e análise de dados.

Witten e Frank (2005) ressaltam que, DM é sobre resolver problemas a partir da análise de uma base de dados já existente, ou ainda, descobrir padrões nos dados de maneira automática ou semiautomática. Para Tan, Steinbach e Kumar (2009), mineração de dados é o processo automático de descoberta de informação valiosa em grandes bases de dados para identificar padrões úteis e que eram desconhecidos anteriormente, possibilitando até a previsão de um resultado a partir do aprendizado com os dados. Segundo Berry e Linoff (1997), data mining é a exploração e análise, de forma automática ou semiautomática, de grandes bases de dados com objetivo de descobrir padrões e regras.

As tarefas de mineração de dados se encaixam em diferentes situações, dependendo do objetivo de sua aplicação: classificação; regressão; associação, ou; análise de agrupamentos (ou clusterização) (TAN; STEINBACH; KUMAR, 2009):

- Classificação: é uma função de aprendizado supervisionado que mapeia (classifica) uma série de dados em várias classes pré-definidas (WEISS, KULIKOWSKI 1991; HAND 1981). Após descoberta a função, pode-se aplicar a novos dados, de maneira a prever a classe que esses se encaixam. A classe de saída sempre será um dado categórico (não numérico), como por exemplo: sim/não, alto/médio/baixo, nível A/B, etc.;
- Associação: Segundo Tan, Steinbach e Kumar (2009) associação é o método utilizado na descoberta de padrões onde hajam atributos fortemente associados na base de dados, representados normalmente por subconjuntos de características, por exemplo um grupo de pessoas

fumantes dentro de uma população em análise. Aqui, caracteriza-se o uso de método de aprendizado não supervisionado;

- Análise de agrupamento (clusterização): A análise de cluster, diz respeito a agrupar ou segmentar os dados oferecidos em subconjuntos. Diniz e Louzada Neto (2000), definem um cluster como um subconjunto, dentre todos os outros formados pela base em estudo, no qual os objetos desse conjunto (*cluster*) estão mais próximos (semelhantes entre si) do que qualquer outro objeto alocado em outro cluster, por exemplo uma separação de um grupo de pessoas onde a média da idade das pessoas do grupo 1 é muito maior que as agrupadas para o grupo 2. Trabalha com método de aprendizado não supervisionado, assim como a tarefa de associação;
- Regressão (estimativa): segundo Fayyad et al. (1996), a regressão é o aprendizado mapeando um item de saída para um atributo de predição real estimada. Semelhante à classificação, a regressão também propõe uma estimativa de alguma variável, porém esta lida com resultados discretos (numéricos, por exemplo o volume de chuva de uma semana no inverno após uma seca). O método utilizado para esta tarefa é de aprendizado supervisionado.

E. INTERPRETAÇÃO E AVALIAÇÃO DOS DADOS

Nesta fase é feita a interpretação e avaliação dos dados, para que o objetivo do trabalho seja cumprido. Se caso os dados, ao fim dessa fase, não condizerem com a realidade o processo pode voltar a qualquer etapa anterior, assim como ser refeito desde o início, até que a informação relevante fique evidente e seja condizente. Esta fase também exige alto nível de entendimento do processo em análise, para avaliação dos resultados obtidos através do DM.

2.2 *MACHINE LEARNING*

Faz parte de um processo de descoberta em base dados o *Machine Learning,* que se traduz “aprendizado de máquina” (AM). Witten e Frank (2005) citam

aprendizado como adquirir conhecimento por estudo, averiguar, gravar na memória ou receber instrução. Contudo, busca-se a definição do aprendizado de máquina propriamente dito, Tom Mitchell (1997) definiu aprendizado de máquina como: "Diz-se que um programa de computador aprende pela experiência E, com respeito a algum tipo de tarefa T e performance P, se sua performance P nas tarefas em T, na forma medida por P, melhoram com a experiência E".

Antes de citar técnicas do AM tem-se alguns conceitos básicos para melhor aproximação do assunto (MITCHELL, 1997):

- Exemplo (padrão, instância): é um objeto único, do qual se aprenderá um modelo, ou sobre o qual um modelo será usado (em casos de predição, por exemplo). As instâncias normalmente serão um conjunto de características;
- Característica (atributo, variável): uma quantidade descrevendo um exemplo. Uma variável tem seu tipo definido pelo seu domínio, que designa valores que ele pode assumir;
- Acurácia (erro): é a taxa de predições corretas (ou incorretas) mostrada pelo modelo estudado para um conjunto de dados;
- Classe: é um atributo especial, que descreve o fenômeno de interesse do estudo.

Quando se trata de AM, pode-se dividir em dois tipos de aprendizado, que serão aplicados conforme a expectativa e área de aplicação, são elas: aprendizado supervisionado e aprendizado não supervisionado.

No aprendizado supervisionado, é necessário um fator externo à mineração aplicada nos dados, de modo que este se combine com a estrutura em estudo. Por exemplo, pode-se utilizar o aprendizado supervisionado em uma classificação, onde, o software vai receber as informações para a classificação, mas para realizar o aprendizado é necessário definir quais instâncias pertencem a qual classe. Ou seja, é como se em um estudo diversos tipos de flores aleatórias foram agrupados em classes ("rosa", "girassol", "lírio", por exemplo) via AM a partir de suas propriedades químicas. O aprendizado supervisionado é aplicado nesse tipo de estudo fazendo uma comparação do resultado do software contra o verdadeiro tipo da flor (TAN; STEINBACH; KUMAR, 2009).

O aprendizado não supervisionado é uma técnica de aprendizado que classifica dados previamente sem classificação automática, através de técnicas convenientes a ocasião. Ao contrário do aprendizado supervisionado, nesse método, não é utilizado nenhuma fonte externa para o aprendizado, utilizando-se apenas a base de dados fornecida para o AM (REVISTABW, 2015). Uma das utilizações desse processo é um agrupamento de instâncias, de acordo com suas variáveis, resultando então em um filtro na base de dados inicial, formando grupos que tem características em comum. Por exemplo, como será mais detalhado nos próximos itens, utilizar técnicas de aprendizado não supervisionado em uma base de dados composta por manutenções de veículos e quilometragem rodada, separando-os em grupos de acordo com suas semelhanças (dentro de uma base seria possível separar carros de tratores ou caminhões, apenas utilizando as informações dispostas).

Neste trabalho será utilizado apenas a aprendizagem não supervisionada, com a tarefa de agrupamento.

2.2.1 ANÁLISE DE AGRUPAMENTO

Dentre os tipos de mineração de dados citados anteriormente, o foco deste trabalho será o método de clusterização, nos próximos itens será desenvolvido mais detalhadamente sobre esse processo.

A análise de grupos ou clusterização pode ser realizada por diversas metodologias, dependendo da aplicação em questão. No presente trabalho serão utilizadas as técnicas *K-means* e maximização de expectativas (*Expectation Maximization* – EM) onde o algoritmo EM será utilizado para definir a quantidade de clusters que serão agrupados pelo *K-means.*

2.2.2 EXPECTATION MAXIMIZATION (EM)

O *expectation maximization* (maximização de expectativas) é um método que busca maximizar a probabilidade, descrito nos softwares como *likelihood,* que pode se aproximar ao português por verossimilhança (parâmetro de confiabilidade de um modelo estatístico). A parte “*expectation”* ou “expectativa”, refere-se ao primeiro passo do algoritmo, onde calcula-se a probabilidade do cluster (probabilidade de uma

instância estar em um cluster), quando se fala sobre "*maximization"* ou "maximização" refere-se ao segundo passo, onde calcula-se a distribuição dos parâmetros (média e desvio padrão) esperando uma maximização da verossimilhança (WITTEN; FRANK, 2005; COUVREUR, 1996).

O algoritmo funciona trocando as instâncias de clusters e recalculando os parâmetros até que a maior verossimilhança seja atingida. Para verificar o nível de confiança no algoritmo, deve-se utilizar o *log likelihood* que deve ser o maior possível no processo aplicado. Esse tipo de agrupamento também permite gerar um número de clusters para uma amostra, a partir de uma validação cruzada, seguindo o passo a passo abaixo (software WEKA, 2017):

I. Número inicial de clusters é definido como 1;
II. Os dados são separados dez vezes, em 90% para aprendizado e 10% para teste;
III. O algoritmo EM é aplicado nos 10 modelos;
IV. É feito uma média da verossimilhança dos 10 modelos criados;
V. Se a verossimilhança aumentou, é somado um cluster ao número inicial e então o algoritmo se repete, até que a verossimilhança não aumente.

2.2.3 K-MEANS

O algoritmo *K-means* (MacQUEEN, 1967; ANDERBERG, 1973) pode ser resumido em quatro passos básicos:

I. Escolha do número inicial de clusters *K*;
II. Cálculo da dissimilaridade entre o objeto e a média de um cluster;
III. Alocação do objeto no cluster no qual a média é a mais próxima;
IV. Recalcular a média de um cluster a partir dos objetos que foram alocados para que a dissimilaridade no cluster seja minimizada;

Com exceção do item (I), todos os três outros métodos são repetidos até que o algoritmo convirja para que não seja mais necessário a realocação de objetos. É importante ressaltar alguns pontos sobre o método:

I. O algoritmo é eficiente em bases de dados onde o número de variáveis, clusters e interações deve ser muito menor do que o número de instâncias estudados (MURTAGH 1992);

II. Será utilizado apenas com valores numéricos, pois trabalha em função de minimizar a dissimilaridade (numérica) nos clusters;
III. A decisão do número (*K*) de clusters pode ser uma escolha difícil e alguns algoritmos paralelos podem ser utilizados para aplicação do método.

Para análise dos clusters obtidos no processo, é necessário utilizar de técnicas estatísticas posteriores, visto que não é trivial tomar conclusões nos clusters obtidos. Para avaliação da qualidade e melhor entendimento dos grupos formados, normalmente utilizam-se análises estatísticas, como verificação de normalidade dos dados, comparação de médias, índices de correlação, etc. Com isso é possível tomar melhores conclusões (além de serem embasadas estatisticamente) quanto a qualidade do algoritmo realizado e dos clusters formados, ou até mesmo tomar decisões de novas formatações da base de dados (se o resultado não for relevante para o estudo).

2.3 ANÁLISE ESTATÍSTICA

Como citado no capítulo 2.2, para que se possa analisar os clusters formados pelo processo de mineração de dados, pode-se utilizar testes estatísticos de modo a verificar as similaridades e diferenças nos grupos obtidos. Neste trabalho, serão utilizados testes não paramétricos para comparação dos clusters.

Segundo Triola (2008) e Mitra (2008), uma amostra com menos de 30 instâncias não é confiável para ser tratada como uma distribuição normal, assim, nesses casos, utiliza-se a distribuição não paramétrica. Quando se trata de uma distribuição paramétrica, tem-se uma base com parâmetros bem definidos (média, desvio padrão, proporção, etc.), contudo, a distribuição não paramétrica não possui esses parâmetros bem definidos. Esta pode ser definida como um “arranjo livre”, pois trata-se de uma base de dados onde não se sabe qual a real organização dos elementos na amostra. Para esse tipo de distribuição existem testes estatísticos específicos para comparações estatísticas (por exemplo verificar se um grupo tem uma mediana estatisticamente diferente a outro).

Um dos testes não paramétricos utilizados para comparação de duas amostras diferentes é o teste *Mann Whitney U* (MWU). Esse método pode ser aplicado somente para comparação de amostras de dois grupos independentes (dois tipos de carros diferentes, duas pessoas diferentes, dois objetos de dados diferentes). O teste reorganiza os dois grupos em estudo por um ranking crescente dos dados, utilizando um ranking para cada exemplo e a soma dos rankings de cada grupo. Então, o método compara a soma dos rankings e, para análise do resultado, deve-se analisar se o nível de significância é menor que 0,05. Se o resultado do nível de significância do teste for maior que 0,05, rejeita-se a hipótese nula (H_0) que os dois grupos possuem medidas de posição diferentes, concluindo que os dois grupos não são estatisticamente diferentes e, consequentemente, não tem suas medianas diferentes, quando se trata da variável em questão (MANN; WHITNEY, 1947).

Sabe-se que é possível utilizar o teste MWU para tomar conclusões sobre dois grupos independentes, porém esse teste não pode ser usado para comparação entre mais grupos. Para comparação de mais de dois grupos pode-se aplicar o teste de *Kruskal-Wallis,* este é um teste não paramétrico que utiliza rankings das amostras de três ou mais grupos não dependentes. O resultado do teste é aceitar ou rejeitar a hipótese de os grupos terem as medianas estatisticamente diferentes. Assim como o MWU, o algoritmo nesse teste faz um ranking entre as variáveis (de todos os grupos no mesmo ranking) e utiliza a soma dos rankings para trazer um nível de significância, assim como no teste anterior, busca-se um valor maior que 0,05 (TRIOLA, 2008).

2.4 MANUTENÇÃO DE VEÍCULOS

Pode-se definir manutenção de diversas maneiras, além de conceituar tipos, componentes e áreas que a manutenção aborda. Pode-se referenciar algumas definições por alguns órgãos de normatização e grupos coordenadores de diversificados ramos industriais.

Uma das definições é evidenciada na Associação Brasileira de Normas e regras (ABNT, 1971), onde manutenção são todos os processos envolvidos nas ações para que um item seja restaurado ou conservado sendo que esse permaneça (ou volta a permanecer) dentro das normas definidas. Ainda segundo a ABNT (1971), defeito é uma ocorrência que não impossibilita o uso do equipamento, mas pode acarretar em

indisponibilidade a curto ou longo prazo. Falhas, diferente dos defeitos é quando o equipamento fica inutilizável.

Vieira (1991) destaca duas características relevantes quanto à manutenção, (i) é um item caro na composição dos custos, somando gastos significativos ao longo da vida útil do equipamento (ou carro, no caso do trabalho em questão), (ii) sempre será um serviço de mão de obra intensiva, mesmo que a tecnologia avance na área.

Diversas políticas de manutenção aplicam-se a uma empresa, sejam isoladas ou combinadas, proporcionando uma gestão/plano de manutenção, são algumas delas (KELLY & HARRIS 1980):

- Em intervalos pré-fixados, onde ocorre tanto a substituição individual quanto a de grupos de componentes;
- Baseada nas condições pré-estabelecidas, sendo continua ou periodicamente;
- Manutenção corretiva, que é a ação a partir do defeito/falha;
- Manutenção de oportunidade, quando o processo pode ser muito longo (como por exemplo em grandes maquinas industriais).

Para veículos é necessário evidenciar alguns fatores que podem influenciar no processo de manutenção da frota. São eles: número de passageiros, quilometragem, tipo de combustível, ano do veículo, tipo de freio, tipo de motor, dentre outros fatores menos relevantes.

3 DESCRIÇÃO DO PROBLEMA

Atualmente a locadora de veículos possui em torno de 70 veículos, que em sua grande maioria (aproximadamente 66 automóveis) estão alugados em contratos. Os contratos da empresa se diferenciam em contratos de frota de veículos para outras empresas e contratos de carros para uso particular, contudo a grande maioria dos veículos é alugado para empresas (cerca de 85% da demanda é para empresas e o restante para particular).

Devido à grande quantidade de veículos e para uma melhor gestão dos custos e contratos, a empresa adotou, em 2016 um sistema para registro de dados de todos os automóveis que foram comprados ou vendidos e se estes foram alugados. Contudo, o sistema é usado atualmente apenas para uma melhor organização dos dados e contratos (que ficavam anteriormente em papéis/arquivos e pastas “isoladas” nos computadores da empresa).

Devido à crescente demanda da empresa, que começou em 2014 com menos de 10 veículos e hoje cresceu a ponto de não conseguir atender toda a demanda da região devido à falta de capital para investimento na frota, nota-se uma necessidade para um melhor gerenciamento de custos. Parte da renda da empresa e a maioria dos seus custos estão focados no setor de manutenção, onde a empresa tem de arcar com custos de rotina (por exemplo, troca de filtro de óleo, revisões, etc.).

Portanto, percebe-se que a falta de uma gestão adequada na área de manutenção em uma locadora de veículos gera custos excessivos, além da não visibilidade de possíveis investimentos e melhorias no negócio.

Para se obter uma gestão apropriada ao uso e melhor direcionamento de gastos/investimentos, deve-se estudar o processo de manutenção e definir parâmetros para base do estudo. Atualmente, a locadora de veículos não tem um sistema de gestão bem estruturado, além de não possuir nenhum processo de análise para melhoria.

4 MÉTODOS DE PESQUISA

Neste capítulo serão discutidos as técnicas e tipos de pesquisas do estudo, além de discorrer sobre o fluxo em que a pesquisa ocorrerá.

4.1 MATERIAIS E MÉTODOS

4.1.1 VISITAS E ENTREVISTA

Quanto ao fluxo da pesquisa, primeiramente foi realizada uma visita *in loco,* na empresa em questão para verificar a viabilidade do estudo e da disponibilidade de informações que seriam proporcionadas pela empresa. Foi autorizado total acesso à base de dados da empresa (que possui dados a partir de 2016), onde esta utiliza um sistema próprio para registro de informações quanto a movimentação de veículos. Vale ressaltar que a empresa possui um processo de qualidade quando se trata de *inputs* no sistema, ou seja, faz uma boa coleta de dados, sempre alimentando o software instalado na maneira que for necessário (manutenções pendentes, novos carros, novos contratos, venda de veículos, etc.). Destacam-se alguns registros feitos pelos funcionários da empresa:

- Dados de manutenção: todos os processos de manutenção e seus custos foram registrados em sistema;
- Compra e venda de veículos com seus respectivos custos;
- Registro de impostos sobre veículos;
- Todos os dados relevantes referentes aos veículos que a empresa possui ou já possuiu: ano do carro, modelo, km rodados, situação atual (alugado/não alugado), cliente atual, dentre outros.

A visita e as entrevistas foram focadas em entender os processos internos da empresa, de modo a poder identificar falhas de gestão e oportunidades de melhoria. Após a visita e verificação da base de dados a tratar, foram estudados os processos da empresa, via entrevista com o gerente da locadora e também o próprio dono do negócio. Assim que realizadas as primeiras entrevistas e análise dos dados a serem tratados, foi levantada a ideia de realizar uma mineração de dados no setor de

manutenção, onde não havia uma gestão apropriada. Para melhor embasar a decisão, foram utilizadas diversas bibliografias para decisão do tipo de estudo e quais técnicas seriam aplicadas para desenvolver a pesquisa. Com a leitura e análise das opções de mineração de dados, optou-se por utilizar uma análise de clusters com o método *k-means* e para definição do número de clusters utilizou-se o algoritmo EM.

4.1.2 BASE DE DADOS

Como citado anteriormente, a base de dados utilizada será focada principalmente no relatório de manutenção dos veículos (chamado de "manutenção frota – relatório analítico"), que possuí leituras a partir de janeiro de 2016. Assim, foi fixada a base de dados de manutenção (base do estudo) de janeiro de 2016 a setembro de 2018, visando utilizar a maior quantidade de dados possível (quanto mais dados, melhor será a qualidade do estudo).

O relatório retirado do sistema segue o modelo retratado na imagem abaixo:

Figura 2 - Relatório de Manutenção Completo

Placa	Modelo	Grupo	Reemb.	Sub.Grupo Serv.	O.S.
BWD-4427	FORD F 4000	001 - MECANICA		001 - SUBSTITUIÇÃO DE PEÇA	1199
BWD-4427	FORD F 4000	001 - MECANICA		001 - SUBSTITUIÇÃO DE PEÇA	1712
BWD-4427	FORD F 4000	001 - MECANICA		001 - SUBSTITUIÇÃO DE PEÇA	1915
BWD-4427	FORD F 4000	001 - MECANICA		001 - SUBSTITUIÇÃO DE PEÇA	2808

Data Fechamento	Km Exec.	Serviço Exec.	Fornecedor	Valor
10/05/17	10.166	CRUZETA DO CARDAN	000269-MAURICIO VISMARA - ME	50,00
12/07/17	17.326	ROLAMENTO DA COLU	000312-RENE FERRARI COMERCIC	55,00
10/08/17	17.326	TAMPA DO RADIADOR	000019-MAMED AUTO ELETRICO	12,00
15/02/18	19.391	BATERIA 150 A CAMINI	000065-M C MASSI ME	520,00

Fonte: Relatório de Manutenção JNSAL.

O relatório de manutenção, como mostra a figura, conta com as informações:

- Placa (placa do veículo);
- Modelo (modelo do veículo, constam 29 modelos nesse relatório, por exemplo: gol 1.0, saveiro, etc.);
- Grupo (grupo de manutenção, segmentada em 8 grupos: mecânica, funilaria, eletricista, tapeçaria, borracharia, acessórios, vidraçaria e outros);

- Reemb. (é separa em vazio ou “S”, onde as linhas marcadas com “S” são manutenções que serão pagas a empresa pelo locatário);
- Sub. Grupo Serv. (é o tipo de serviço que foi realizado, tipo de manutenção, somam 40 tipos no total);
- O. S. (número da ordem de serviço);
- Data Fechamento (data em que a manutenção foi realizada);
- Km Exec. (quilometragem do veículo no momento da manutenção);
- Serviço Executado (descrição da manutenção realizada);
- Fornecedor (lugar em que foi realizada a manutenção);
- Valor (custo, em reais, do serviço).

Com os dados de manutenção já coletados, depois foi necessário coletar também a idade do veículo e a contagem de quilômetros rodados. Para isso utilizou-se o relatório “idade frota”, que conta com as informações como: quilometragem, ano do veículo, idade do veículo (em meses). Como um relatório complementar apenas, neste foram utilizados apenas os dados de quilometragem rodada total e idade do veículo.

Com isso, foi possível obter todas as informações provenientes da empresa, que seriam necessárias para realizar o estudo de KDD no setor de manutenção. Contudo para que se obtenha uma informação de qualidade com a DM, é necessário montar uma base de dados que possua informações relevantes, onde as aplicações das técnicas de mineração resultem em conhecimentos que façam sentido para o negócio. Para montar a base final do estudo foi necessária uma análise qualitativa sobre os itens mostrados no relatório de manutenção, verificando a relevância da base montada de acordo com a visão da empresa. Com isso, foi definido que a base seria filtrada de modo que cada linha no estudo seria uma placa (um veículo) e as colunas seriam as informações de idade, quilometragem e manutenção.

Apesar de a decisão do formato da base de dados estar definida, ainda é necessário decidir quais colunas seriam utilizadas na clusterização. Como o número de informações é muito extenso, foi montada uma base completa, com todas as possibilidades relevantes de atributos para cada veículo. A base completa seguiu o modelo:

Tabela 1 - Parte 1 Base de Dados

Informações do Veículo				
Placa	**Modelo**	**Tipo**	**Km rodado**	**Idade**
FNK-0213	SAVEIRO ROBUST	Carro	44478	5
GBH-9513	GOL 1.0	Carro	90419	11
FOY-1707	SAVEIRO ROBUST	Carro	40871	5
FOP-3351	GOL 1.0	Carro	261443	34
FSO-8737	GOL 1.0	Carro	185222	25
FYS-5707	GOL 1.0	Carro	241721	34

Fonte: Autoria própria.

Tabela 2 - Parte 2 Base de Dados

Nº de Manutenções					
Mecânica	**Funilaria**	**Elétrica**	**Borracharia**	**Vidraçaria**	**Manut TT**
30	0	1	2	0	33
75	0	0	1	0	76
30	0	2	0	0	32
216	12	11	12	0	251
93	29	15	5	3	145
166	4	17	11	0	198

Fonte: Autoria própria.

Tabela 3 - Parte 3 Base de Dados

Valor Gasto em Manutenção (Reais)					
R$ Mecânica	**R$ Funilaria**	**R$ Elétrica**	**R$ Borracharia**	**R$ Vidraçaria**	**R$ TT**
1712	0	6	738	0	2457
3591	0	0	410	0	4001
1785	0	14	0	0	1799
22512	2449	1431	3154	0	29546
15221	5740	1845	355	740	23901
13403	802	2039	2121	0	18365

Fonte: Autoria própria.

Como mostrados nas tabelas acima, o relatório final foi dividido em nº de manutenções (contagem de manutenções por carro, divididos por tipo de manutenção), onde a coluna “Manut TT” é a somatória de todas as manutenções realizadas, valor gasto em manutenção (soma de reais gastos em cada tipo de manutenção, além do total, na coluna “R$ TT”) e informações do veículo: tipo de carro

(classificação definida pelo time da locadora), Km rodado (quilometragem total) e idade (meses).

Além das colunas mostradas nas tabelas, foram replicadas todas as colunas com os valores mostrados divididos pela idade do veículo, resultando no valor por mês. Por exemplo, se o carro tem 30 na coluna “Manut TT” e idade 5, ele terá 6 (30 ÷ 5 = 6) “Manut TT/mês”, isso foi feito para todos os indicadores mostrados. Feito isso, a base final para análise contou com uma tabela de 155 linhas (número de carros que fazem ou fizeram parte da frota a partir de 2016) e 27 colunas, contando com: idade, número de manutenções, número de manutenções por mês, valor gasto em manutenção, valor gasto em manutenção por mês, quilometragem total e quilometragem rodada por mês.

Com as informações compiladas e a base de dados completa, o próximo passo é definir qual será o modelo final para estudo da mineração de dados. Para isso, foi necessário realizar uma limpeza na base selecionada, pois aplicar a mineração sem nenhum critério de filtro (mesmo com os dados organizados por carro) não seria eficiente e nem relevante para análise. Assim, foram retirados da base as informações que não agregariam valor ao estudo:

- Manutenções do tipo “outros”, “acessórios” e “tapeçaria” (esse tipo de manutenção não é o foco do estudo, visto que não são influenciados pela idade ou quilometragem do carro);
- Manter apenas veículos do tipo “Carro” (visto que esse tipo de veículo forma a maior parte da base de dados, além de ser a principal fonte de renda da locadora);
- Carros com menos de 100 quilômetros rodados (foram retirados os carros novos, pois não possuem manutenções relevantes para a pesquisa, que tem foco em relacionar a manutenção com a quilometragem rodada).

Para tomar a decisão final de qual seria a base final do estudo, foram montados cinco modelos para análise, obedecendo os critérios acima:

- MODELO 1: oito variáveis (nº de manutenções mecânica, nº de manutenções funilaria, nº de manutenções elétrica, nº de manutenções borracharia, nº de manutenções vidraçaria, valor gasto total, quilometragem total e quilometragem rodada por mês);

- MODELO 2: oito variáveis (valor gasto em manutenções mecânica, valor total gasto em manutenções funilaria, valor total gasto em manutenções elétrica, valor total gasto em manutenções borracharia, valor total gasto em manutenções vidraçaria, nº total de manutenções, quilometragem total e quilometragem rodada por mês), iguais ao MODELO 1, trocando nº de manutenções por valor total gasto em manutenções;
- MODELO 3: cinco variáveis (nº de manutenções total, valor gasto com manutenções total, quilometragem total, quilometragem rodada por mês e idade do veículo);
- MODELO 4: sete variáveis (nº de manutenções por mês mecânica, nº de manutenções por mês funilaria, nº de manutenções por mês borracharia, nº de manutenções por mês vidraçaria, quilometragem rodada por mês e valor gasto por mês em manutenções), iguais ao MODELO 1, com todos os atributos divididos pela idade do carro;
- MODELO 5: sete variáveis (reais gastos em manutenção por mês mecânica, reais gastos em manutenção por mês funilaria, reais gastos em manutenção por mês elétrica, reais gastos em manutenção por mês borracharia, reais gastos em manutenção por mês vidraçaria, nº de manutenções por mês e quilometragem rodada por mês), iguais ao MODELO 2, com todos os atributos divididos pela idade do carro. DATA MINING.

Os modelos apresentados na lista anterior são todos plausíveis para serem agrupados e analisados conforme sua distribuição, contudo, apenas um deles será o objeto de estudo deste trabalho. Enfim, de modo a escolher dentre os modelos montados previamente, todas as cinco bases foram submetidas ao processo de mineração de dados (clusterização). Para esta fase da metodologia foi estipulado que o método *k-means* seria o algoritmo selecionado para distribuir os carros em clusters para análise final dos grupos formados. Entretanto, o número de clusters deve ser pré-definido para esse tipo de clusterização, assim, foi aplicado o processo de EM na escolha do número K de clusters, visto que esse algoritmo possibilita a escolha automática de clusters (no software utilizado).

O software escolhido para aplicação da clusterização foi o WEKA (*Waikato Environment for Knowledge Analysis*), desenvolvido na Nova Zelândia pela *University of Waikato* (Universidade de Waikato) é um software programado na linguagem JAVA que possui diversas possibilidades de mineração e visualização de dados (informação obtida no site da Univeridade de Waikato). No presente trabalho os únicos artifícios utilizados foram os dois algoritmos de DM, *k-means* e EM.

O resultado da mineração de dados é mostrado na tabela 4:

Tabela 4 - Comparativo Modelos

	nº variáveis	**nº clusters**	**Erro K-means**	**log likelihood (EM)**
MODELO 1	8	4	14,38	-36,6
MODELO 2	8	5	13,67	-60,62
MODELO 3	5	6	5,48	-37,68
MODELO 4	7	5	11,48	-10,22
MODELO 5	7	5	12,51	-29,67

Fonte: Autoria própria.

Como se pode observar na tabela, o único modelo que resultou em um *log likelihood* muito baixo foi o MODELO 2 e ao mesmo tempo teve o segundo maior erro quando aplicado o método *k-means*, tornando-o não elegível para o modelo final de estudo. Contudo, o critério para seleção da melhor clusterização será o erro do *k-means,* visto que esse foi o algoritmo selecionado para classificar os grupos finais para análise completa. Além disso, vale destacar que o MODELO 3 traz indicadores mais globais do processo, tratando as variáveis de manutenção como um todo, assim sendo um destaque para conclusões sobre o processo geral da manutenção. Com isso, o MODELO 3 mostrou-se com um erro muito menor que os outros modelos (menos da metade) e é a melhor para uma análise da manutenção relacionada ao custo total, visando a descoberta de conhecimento.

4.1.3 DESCOBERTA DE CONHECIMENTO

Após a definição do modelo final e da formação dos clusters, é necessária uma análise mais detalhada para entender o agrupamento realizado pelo *k-means*. Como

visto no capítulo 2.2, de modo a ter uma visão mais estratificada dos grupos obtidos, pode-se utilizar testes estatísticos de comparação para tomar conclusões.

Nessa etapa foi utilizado o software SPSS (*Statistical Package for the Social Sciences*), é um programa que opera na plataforma JAVA lançado em 1968 por Norman H. Nie, C. Hadlai e Dale H. Bent. Hoje é um programa conhecido mundialmente, líder em análises preditivas (combinação de análises avançadas e otimização de decisão). Dentre suas inúmeras funções, para o presente trabalho serão utilizados dois testes estatísticos para comparação de amostras não paramétricas: *Mann Whitney* e *Kruskal-Wallis*.

Para comparar os clusters formados, primeiramente é necessário verificar o tamanho dos clusters formados e analisar qual a sua distribuição de dados, onde, caso os grupos formarem uma distribuição normal, serão utilizados testes paramétricos e, caso contrário, são utilizados testes não paramétricos. No resultado dos grupos obtidos nesse estudo, constam apenas clusters com menos de 30 elementos, representando amostras muito pequenas para serem tratadas como uma distribuição normal. Com isso, deve-se aplicar primeiro o teste de *Kruskal-Wallis*, que é robusto suficiente para comparar todos os seis clusters formados, verificando se estes possuem uma mediana estatisticamente diferentes.

O processo foi feito para todas as 5 variáveis e então em cada variável onde a hipótese de que as medianas dos grupos são diferentes foi aceita, o teste *Mann Whitney* foi aplicado comparando dois grupos por vez. Assim foi possível tomar conclusões finais sobre a distribuição dos clusters.

5 RESULTADOS E DISCUSSÃO

Neste capítulo serão mostrados os resultados dos métodos descritos no capítulo 4 e estes serão discutidos quantitativamente e qualitativamente.

A base final do estudo foi definida no MODELO 3, na tabela 5 foram recortadas as primeiras linhas do relatório:

Tabela 5 - Modelo Final para Clusterização

Placa	Manut TT	R$ TT	Km TT	Idade	Km/Mês
FNK-0213	33	R$ 2.457,00	44.478	5	8.896
GBH-9513	76	R$ 4.001,00	90.419	11	8.220
FOY-1707	32	R$ 1.799,00	4.871	5	8.174
FOP-3351	251	R$29.546,00	261.443	34	7.690
FSO-8737	15	R$23.901,00	185.222	25	7.409
FWS-5707	198	R$18.365,00	241.721	34	7.109

Fonte: Autoria própria.

5.1 RESULTADOS

Após ser aplicada a mineração de dados (*k-means*) obtiveram-se seis clusters (com erro quadrado de 5,48), distribuídos como mostra a tabela, vale ressaltar que os valores dos atributos na tabela são as medianas dos respectivos clusters:

Tabela 6 - Clusters Modelo Final

	Instâncias	Manut TT	R$ TT	Km TT	idade	Km/Mês
cluster1	13	132	R$ 17.433,84	141.250	45	3.060
cluster2	21	13	R$ 2.321,37	71.932	45	1.621
cluster3	17	194	R$ 24.868,55	200.565	34	6.110
cluster4	13	73	R$ 4.000,72	79.178	17	5.084
cluster5	12	59	R$ 5.538,23	115.155	45	2.651
cluster6	29	17	R$ 1.134,43	26.989	7	4.168

Fonte: Autoria própria.

Os clusters formados foram inseridos no software SPSS para as análises estatísticas, o primeiro teste aplicado é o *Kruskal-Wallis,* para verificar se é aceita a hipótese de que os clusters tem medianas diferentes (em cada variável). Quando se aceita a hipótese, não é preciso realizar mais nenhum teste estatístico para provar a

similaridade dentre os grupos, contudo, se a hipótese for rejeitada é necessário realizar mais uma série de testes para identificar as diferenças nos grupos. Os resultados do primeiro teste *Kruskal-Wallis* são mostrados nas imagens abaixo:

Tabela 7 - Teste Kruskal-Wallis

	Manut TT	**R$ TT**	**Km TT**	**Idade**	**Km/Mês**
Chi-square	88,838	83,809	90,67	85,29	63,676
df	5	5	5	5	5
Asymp. Sig.	0	0	0	0	0

Fonte: Software SPSS.

Observando a linha “Asymp. Sig.” (esse é o nível de significância do teste, onde com um valor maior que 0,05 é rejeitada a hipótese de que as medianas dos grupos são estatisticamente diferentes). Portanto, percebe-se que para todas as variáveis as medianas são diferentes entre os seis grupos, assim é necessário analisar mais afundo (comparando grupo a grupo) para tomar uma conclusão definitiva quanto a distribuição dos dados. Para isso, aplicam-se diversos testes de *Mann Whitney U*, verificando as diferenças e similaridades dentre os grupos. As tabelas abaixo foram organizadas de modo que cada linha ou coluna seja um cluster e na sua intersecção mostra-se o valor de significância no teste aplicado:

Tabela 8 - Mann Whitney U "Manut TT"

Cluster	**1**	**2**	**3**	**4**	**5**	**6**
1	-	0	0	0	0	0
2	0	-	0	0	0	0,05
3	0	0	-	0	0	0
4	0	0	0	-	0,007	0
5	0	0	0	0,007	-	0
6	0	0,05	0	0	0	-

Fonte: Autoria própria.

Tabela 9 - Mann Whitney U "R$ TT"

Cluster	1	2	3	4	5	6
1	-	0	0,001	0	0	0
2	0	-	0	0,001	0	0,018
3	0,001	0	-	0	0	0
4	0	0,001	0	-	0,142	0
5	0	0	0	0,142	-	0
6	0	0,018	0	0	0	-

Fonte: Autoria própria.

Tabela 10 - Mann Whitney U "Km TT"

Cluster	1	2	3	4	5	6
1	-	0	0	0	0,034	0
2	0	-	0	0,045	0	0
3	0	0	-	0	0	0
4	0	0,045	0	-	0,001	0
5	0,034	0	0	0,001	-	0
6	0	0	0	0	0	-

Fonte: Autoria própria.

Tabela 11 - Mann Whitney U "idade"

Cluster	1	2	3	4	5	6
1	-	0,472	0,31	0	0,017	0
2	0,472	-	0,021	0	0,015	0
3	0,31	0,021	-	0	0	0
4	0	0	0	-	0	0
5	0,017	0,015	0	0	-	0
6	0	0	0	0	0	-

Fonte: Autoria própria.

Tabela 12 - Mann Whitney U "Km/mês"

Cluster	1	2	3	4	5	6
1	-	0	0,003	0,118	0,014	0,124
2	0	-	0	0	0,001	0
3	0,003	0	-	0,098	0	0,014
4	0,118	0	0,098	-	0	0,438
5	0,014	0,001	0	0	-	0
6	0,124	0	0,014	0,438	0	-

Fonte: Autoria própria.

Para melhor compreensão do resultado da análise estatística foi elaborado um resumo, onde são comparados dois grupos por vez envolvendo todas as variáveis. O resumo foi elaborado de forma que grupo a grupo são analisadas as semelhanças e diferenças, onde um grupo por vez é fixado (destaque em amarelo) e, se caso houver algum grupo/variável que não são diferentes, este está em destaque verde.

Tabela 13 - Análise Cluster 1

Cluster	Manut TT	R$ TT	KM TT	Idade	KM/mês
1	132	R$ 17.433,84	141.250	45	3.060
2	13	R$ 2.321,37	71.932	45	1.621
3	194	R$ 24.868,55	200.565	34	6.110
4	73	R$ 4.000,72	79.178	17	5.084
5	59	R$ 5.538,23	115.155	45	2.651
6	17	R$ 1.134,43	26.989	7	4.168

Fonte: Autoria própria.

Tabela 14 - Análise Cluster 2

Cluster	Manut TT	R$ TT	KM TT	Idade	Cluster
1	132	R$ 17.433,84	141.250	45	3.060
2	13	R$ 2.321,37	71.932	45	1.621
3	194	R$ 24.868,55	200.565	34	6.110

(continua)

(continuação)

Cluster	Manut TT	R$ TT	KM TT	Idade	Cluster
4	73	R$ 4.000,72	79.178	17	5.084
5	59	R$ 5.538,23	115.155	45	2.651
6	17	R$ 1.134,43	26.989	7	4.168

Fonte: Autoria própria.

Tabela 15 - Análise Cluster 3

Cluster	Manut TT	R$ TT	KM TT	Idade	KM/mês
1	132	R$ 17.433,84	141.250	45	3.060
2	13	R$ 2.321,37	71.932	45	1.621
3	194	R$ 24.868,55	200.565	34	6.110
4	73	R$ 4.000,72	79.178	17	5.084
5	59	R$ 5.538,23	115.155	45	2.651
6	17	R$ 1.134,43	26.989	7	4.168

Fonte: Autoria própria.

Tabela 16 - Análise Cluster 4

Cluster	Manut TT	R$ TT	KM TT	Idade	KM/mês
1	132	R$ 17.433,84	141.250	45	3.060
2	13	R$ 2.321,37	71.932	45	1.621
3	194	R$ 24.868,55	200.565	34	6.110
4	73	R$ 4.000,72	79.178	17	5.084
5	59	R$ 5.538,23	115.155	45	2.651
6	17	R$ 1.134,43	26.989	7	4.168

Fonte: Autoria própria.

Tabela 17 - Análise Cluster 5

Cluster	Manut TT	R$ TT	KM TT	Idade	KM/mês
1	132	R$17.433,84	141.250	45	3.060
2	13	R$ 2.321,37	71.932	45	1.621
3	194	R$24.868,55	200.565	34	6.110
4	73	R$ 4.000,72	79.178	17	5.084

(continua)

(continuação)

Cluster	Manut TT	R$ TT	KM TT	Idade	KM/mês
5	59	R$ 5.538,23	115.155	45	2.651
6	17	R$ 1.134,43	26.989	7	4.168

Fonte: Autoria própria.

Tabela 18 - Análise Cluster 6

Cluster	Manut TT	R$ TT	KM TT	Idade	KM/mês
1	132	R$17.433,84	141.250	45	3.060
2	13	R$ 2.321,37	71.932	45	1.621
3	194	R$24.868,55	200.565	34	6.110
4	73	R$ 4.000,72	79.178	17	5.084
5	59	R$ 5.538,23	115.155	45	2.651
6	17	R$ 1.134,43	26.989	7	4.168

Fonte: Autoria própria.

Com todos os testes estatísticos realizados, fica evidente quais grupos têm de fato diferenças ou semelhanças quanto às suas medianas, possibilitando tomar conclusões sobre o estudo realizado.

5.2 ANÁLISE E DISCUSSÃO

A análise principal dos resultados obtidos deve ser relacionada ao total de valor gasto, visto que a manutenção interfere diretamente na lucratividade, que é o foco da empresa. Assim, é válido comparar os clusters onde o gasto total foi maior, cluster 1 (mediana reais gastos igual a R$17.433,84) e cluster 3 (mediana reais gastos igual a R$24.868,55), visto que o valor gasto é muito maior que o terceiro cluster nesse ranking (mais de três vezes).

De imediato, percebe-se que os grupos 1 e 3 não têm idade do carro estatisticamente diferentes, ou seja, suas medianas podem ser comparadas (apesar de os números serem diferentes), porém isso não significa que sua quilometragem total nem o total gasto também seguem essa semelhança. Nota-se que o grupo 1 tem uma quilometragem rodada por mês muito menor que o grupo 3 (3060km/mês contra

6.110km/mês, praticamente metade) e o grupo 3 possui gasto e quilometragem total muito maior que 1 (-29,6% de quilometragem total e -29,9% de gasto total). Isso permite concluir que o valor de quilômetro rodado por mês influenciou diretamente no gasto em manutenção dos carros desse grupo.

Essa evidência fica mais clara quando se observa os grupos 4 e 5, onde ambos não têm as medianas de reais total gasto estatisticamente diferentes, porém quando se observa a idade e a quilometragem rodada por mês, os números se diferem. No grupo 4, os carros têm 17 meses de idade enquanto no grupo 5, 45; olhando para a quilometragem rodada, o grupo 4 roda aproximadamente 50% mais que o grupo 5. Ou seja, novamente fica claro que a quilometragem rodada por mês influencia diretamente o total gasto, pois os carros com média de idade de 17 meses já têm um total gasto igual aos que tem em média 45 meses de idade.

Ainda utilizando a quilometragem rodada por mês como critério de avaliação, quando se observa os grupos 2 e 1, que possuem a mesma distribuição de idade, nota-se uma grande diferença no valor total gasto (cluster 1 gasta, em média aproximadamente 17 mil reais enquanto o grupo 2 gasta menos de três mil). O número também é muito discrepante quando se trata da quilometragem total, onde o grupo 2 andou aproximadamente 50% menos que o grupo 1, mesmo tendo a mesma média de idade. A grande diferença nesse caso se reflete novamente na quilometragem rodada por mês, onde o grupo 1 andou 53% a mais que o grupo 2 e gasta 87% a mais.

Quando se observa um exemplo onde a mediana de quilometragem rodada por mês não é estatisticamente diferente em dois grupos, novamente os valores de gasto e quilometragem total se conversam. Por exemplo, olhando para os grupos 4 e 6, que não possuem os valores de quilometragem rodada por mês diferentes estatisticamente. Observa-se que o grupo 4 gastou 71% a mais que a média dos carros do cluster 6 e andou 66% a mais também. Novamente é evidenciado que os carros que andam mais por mês, com o passar do tempo aumentam sua quilometragem total mais rapidamente, logo, gastam mais em manutenção.

Portanto, conclui-se que a idade (nos carros em exemplo) não afeta o gasto total em manutenção, o verdadeiro fator que tem alta influência no total gasto por carro é a quilometragem rodada por mês. Dados os exemplos acima, é possível observar que mesmo que um carro tenha uma idade avançada (até 45 meses, por exemplo) este pode não ter sido um alvo de grandes gastos com manutenção. Logicamente, os

carros que tem uma maior quilometragem rodada, gastam mais com manutenção e isso é mais um exemplo para embasar que a quilometragem por mês é um fator chave quando se trata de gastos em manutenção.

Contudo, é necessário levar alguns pontos em consideração também. Nesse estudo não foi considerado o tempo parado do veículo, o que, de acordo com os dados estudados, reduz o gasto em manutenção no curto prazo. Além disso, vale ressaltar que a amostra de estudo foi pequena (105 carros, clusterizados em 6 grupos diferentes, onde o maior grupo possui 29 carros, apenas), com bases de dados muito pequenas as conclusões podem estar um pouco destorcidas. Outro ponto importante é relacionar também ao estudo, em uma futura oportunidade, o preço do aluguel do veículo. Pois, dada as fortes evidências da influência da quilometragem rodada por mês no valor final gasto, podem surgir novas possibilidades no formato de contratos de locação, embasados em estudos estatísticos. Esses são fatores importantes para um próximo estudo na área que podem proporcionar ainda mais informação para empresa.

6 CONCLUSÕES

No presente trabalho, foram estudadas as variáveis de manutenção em uma locadora de veículos com o objetivo de relacioná-las por meio de um agrupamento que fosse relevante para os processos da empresa. Primeiro foi discorrida uma breve introdução no contexto do mercado das locadoras no Brasil e a importância de uma boa gestão do conhecimento em uma empresa. Com isso, foram definidos os objetivos (geral e específicos), buscando responder à pergunta: “Como aplicar a metodologia KDD para a descoberta de conhecimento em base de dados relacionados ao setor de manutenção de uma locadora de veículos? ”.

Para que o estudo fosse realizado de maneira coerente com a literatura, no capítulo 2 foram utilizados principais autores, referenciando cada assunto tratado neste estudo. Foi relatado o processo de KDD e a aplicação da mineração de dados nesse quesito. Contudo, é necessário um embasamento estatístico para tomar conclusões reais do estudo, visto isso, foram referenciados também métodos de comparação estatísticos, utilizados no capítulo 4 e 5. Para tratar de manutenção de veículos, foi dada uma breve introdução ao assunto e retomados os principais conceitos, além de estratificar os itens que mais influenciam no processo de manutenção de veículos (objeto do estudo).

Dado o embasamento teórico, foram realizadas visitas na empresa, além de diversas entrevistas com os funcionários e o dono do negócio para embasar a decisão de qual área e método seria mais relevante na descoberta de conhecimento. Foi aplicada uma limpeza e transformação na base de dados do setor de manutenção para que então, posteriormente, fosse aplicado uma clusterização. Depois de aplicada a mineração de dados, que resultou no agrupamento dos veículos em seis grupos, foram utilizados testes estatísticos para uma análise mais detalhada da nova informação obtida.

Com a DM e aplicação de testes estatísticos, foi possível concluir que na base de dados do estudo a idade do carro não é um fator tão relevante quanto a quilometragem rodada por mês, quando se trata de gasto total em manutenção. Fica claro que a quilometragem rodada por mês deve ser um fator decisório no momento de montar contratos de aluguel, visto que os carros que andam mais em menos tempo, gastam mais com manutenção.

Por fim, conclui-se que para futuros estudos, vale acrescentar alguns fatores que não foram tratados aqui. É interessante utilizar e analisar o preço de aluguel do contrato de cada carro, relacionando com sua quilometragem total e mensal, também é crucial que seja mensurado o tempo parado do veículo, para que se tenha uma precisão melhor na informação. Outro ponto notável para um futuro estudo é o tamanho da base, onde é interessante comparar o mesmo estudo em uma locadora (ou qualquer empresa que tenha veículos as informações necessárias) em que haja uma base de dados mais extensa.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR: 5456: Termos fundamentais de eletricidade**. Rio de janeiro, 1971.

ANDERBERG, Michael R. **Cluster analysis for applications**. Office of the Assistant for Study Support Kirtland AFB N MEX, 1973.

ANGERS, Maurice. Initiation pratique à la méthodologie des sciences humaines. Montreal: Centre Educatif et Culturel (CEC), 1992.

BASSANETO, Renata. **Mercado de TI tem perspectivas de crescimento em 2018**, 2017. Disponível em: <http://www.administradores.com.br/noticias/negocios/mercado-de-ti-tem-perspectivas-de-crescimento-em-2018/122725/>. Acesso em: 03 de jul. 2018.

BERRY, Michael J.; LINOFF, Gordon. **Data mining techniques: for marketing, sales, and customer support**. John Wiley & Sons, Inc., 1997.

COUVREUR, Christophe. The EM algorithm: A guided tour. In: **Computer Intensive Methods in Control and Signal Processing**. Birkhäuser, Boston, MA, 1997. p. 209-222.

DESLAURIERS. **Recherche qualítative: Guide pratique**. Montreal: McGraw-Hill, 1997.

DINIZ, C. A. R.; LOUZADA, Neto F. **Análise Multivariada de Dados**. 5ª ed. Porto Alegre: Bookman, 2000.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowlegde discovery**. American Association for Artificial Intelligence, 1996.

FEDELE, Lorenzo. **Methodologies and techniques for advanced maintenance**. Springer Science & Business Media, 2011.

FONSECA, João José Saraiva. Metodologia da Pesquisa Científica. 2002.

GIL, Antonio Carlos. Como elaborar projetos de pesquisa. **São Paulo**, v. 5, n. 61, p. 16-17, 2002.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. **Rio de Janeiro: Campus**, v. 1, 2005.

HAND, David J. Discrimination and classification. **Wiley Series in Probability and Mathematical Statistics, Chichester: Wiley, 1981**, 1981.

WEKA, W. E. K. A. 3: data mining software in Java. **University of Waikato, Hamilton, New Zealand (www. cs. waikato. ac. nz/ml/weka)**, v. 19, p. 52, 2011. Acesso em: 03 de jul. 2018.

KELLY, Arthur; HARRIS, M. J. Administração da manutenção industrial. **Rio de Janeiro: IBP**, 1980.

MACQUEEN, James et al. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. 1967. p. 281-297.

MANN, Henry B.; WHITNEY, Donald R. On a test of whether one of two random variables is stochastically larger than the other. **The annals of mathematical statistics**, p. 50-60, 1947.

MCDOWELL, Jim. **Electronics Maintenance in Truck and Bus Fleets**. SAE Technical Paper, 1991.

MITRA, Amitava. **Fundamentals of Quality Control and Improvement.** 3th Auburn University College of Business: Auburn, Alabama, 2008.

MURTAUGH, Fionn. Comments on. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, n. 10, p. 1056-1057, 1992.

MITCHELL, Tom M. et al. Machine learning. 1997. **Burr Ridge, IL: McGraw Hill**, v. 45, n. 37, p. 870-877, 1997.

OLIVEIRA, João. **Valor econômico**. 2017. Disponível em: <http://www.valor.com.br/empresas/5384437/setor-de-locacao-de-veiculos-cresce-123-em-2017-aponta-associacao>. Acesso em: 03 de jul. 2018.

SITEWARE. **O Domínio da informação como aliado para o crescimento de negócios. Disponível em: <**https://www.siteware.com.br/comunicacao/crescimento-de-negocios/>. Acesso em: 22 jun. 2018.

REVISTABW. **Aprendizado de Máquina: Aprendizado Não Supervisionado**. Revista Brasileira de Web: Tecnologia. 2018. Disponível em: <http://www.revistabw.com.br/revistabw/aprendizado-de-maquina-aprendizado-nao-supervisionado/>. Acesso em: 22 de jun. 2018.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao datamining: mineração de dados**. Ciência Moderna, 2009.

TRIOLA, Mario F. **Introdução à Estatística.** Rio de Janeiro: LTC, 2008.

VIEIRA, M.G. **Introdução à manutenção**. EESC-SEM: São Carlos, 1991.

WEISS, Sholom M.; KULIKOWSKI, Casimir A. **Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems**. Morgan Kaufmann Publishers Inc., 1991.

WITTEN, I. H.; FRANK, E. **Data Mining**: **Practical Machine Learning Tools and Techniques**. 2 ed. San Francisco, CA: Elsevier, 2005.