

UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIAS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ANA LUISA SILVEIRA DA SILVA

**APLICAÇÃO DE ALGORITMOS DE ASSOCIAÇÃO PARA ANÁLISE DE
COMPORTAMENTO DE FALHAS EM ROLAMENTO DE MATERIAL RODANTE**

SÃO LUÍS

2019

ANA LUISA SILVEIRA DA SILVA

**APLICAÇÃO DE ALGORITMOS DE ASSOCIAÇÃO PARA ANÁLISE DE
COMPORTAMENTO DE FALHAS EM ROLAMENTO DE MATERIAL RODANTE**

Monografia apresentada ao Curso de Engenharia Elétrica da Universidade Federal do Maranhão como requisito para obtenção do grau de Bacharel em Engenharia Elétrica.

Orientador: Prof. Denivaldo Cícero Pavão Lopes

SÃO LUÍS

2019

Silva, Ana Luisa Silveira da.

Aplicação de algoritmos de associação para análise de comportamento de falhas em rolamento de material rodante / Ana Luisa Silveira da Silva. - 2019.

62 f.

Orientador(a): Denivaldo Cícero Pavão Lopes.

Monografia (Graduação) - Curso de Engenharia Elétrica, Universidade Federal do Maranhão, São Luís, 2019.

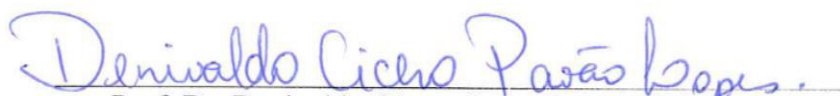
1. Big Data. 2. Manutenção preventiva. 3. Método de associação. 4. Mineração de dados. 5. Padrão de falha.
I. Lopes, Denivaldo Cícero Pavão. II. Título.

ANA LUISA SILVEIRA DA SILVA

**APLICAÇÃO DE ALGORITMOS DE ASSOCIAÇÃO PARA ANÁLISE DE
COMPORTAMENTO DE FALHAS EM ROLAMENTO DE MATERIAL RODANTE**

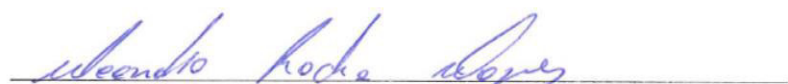
Aprovada em 22 / 02 / 2019

BANCA EXAMINADORA:


Prof. Dr. Denivaldo Cícero Pavão Lopes (Orientador)
Departamento de Engenharia Elétrica - UFMA


Prof. Dr. André Borges Cavalcante (Membro)
Departamento de Engenharia Elétrica - UFMA


Prof. Dr. José de Ribamar Braga Pinheiro Junior (Membro)
Departamento de Engenharia Elétrica – UFMA


Msc. Leandro Rocha Lopes (Membro)
Vale S. A.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus e à minha família por todo apoio, confiança depositada em mim e todo o suporte durante a minha jornada até este momento. E ao meu companheiro de vida, Gabriel Moreira, por todo o apoio imensurável.

Aos meus colegas de trabalho que me auxiliaram e criticaram este trabalho durante sua construção, em especial, Rhaisa Tavares, Leandro Rocha, Giovanni Dias, Suilan Maia e Ada Cristina, pois não mediram esforços para me ajudar mediante as dificuldades.

Enfim, agradeço a todos que participaram direta ou indiretamente durante minha formação acadêmica.

RESUMO

As interrupções de circulação de trem da Companhia Vale do Rio Doce acarretam em perdas de produtividade na cadeia logística de transporte de minério. No contexto do processo de controle de tráfego ferroviário, a tomada de decisão pela parada do trem em função dos problemas relacionados às falhas no rolamento de um vagão é algo problemático, gerando perdas financeiras para a Vale. Sendo assim, uma parada de vagão para manutenção preventiva é menos prejudicial, pois evita que toda a composição de vagões seja parada para uma manutenção corretiva. O fluxo de dados proveniente das medições realizadas pelo sistema de monitoramento da viagem do trem é intenso, gerando um conjunto de dados com considerável volume, variedade, velocidade e valor, caracterizando um cenário de *Big Data*. O método de associação é um procedimento conhecido para fazer mineração de dados, o qual torna possível a descoberta de padrões de falhas ao analisar um conjunto conhecido de dados. De acordo com isso, este trabalho propõe fazer um estudo a partir do conjunto de dados referente às falhas de temperatura e ruídos acústicos nos rolamentos de vagões da empresa Vale, para identificar padrões desconhecidos. Algoritmos de associação foram utilizados nos experimentos, com enfoque no aprendizado de máquina não supervisionado. A partir dos resultados obtidos, foi analisado quais regras são úteis para auxiliar no processo de manutenção preventiva de vagões.

Palavras-chave: Manutenção preventiva; Big Data; Mineração de dados; Método de Associação; Padrão de falha.

ABSTRACT

Interruptions in the circulation of the train of the Company Vale do Rio Doce result in losses of productivity in the logistic chain of transport of ore. In the context of the rail traffic control process, the decision to stop the train due to the problems related to the failure of a wagon bearing is problematic, generating financial losses for Vale. Therefore, a wagon stop for preventive maintenance is less damaging as it prevents the whole composition of wagons from being stopped for corrective maintenance. The data flow from the measurements made by the train travel monitoring system is intense, generating a set of data with considerable volume, variety, velocity and value, characterizing a Big Data scenario. The association method is a known procedure for doing data mining, which makes it possible to discover fault patterns when analyzing a known set of data. According to this, this work proposes to make a study based on the data set concerning the temperature and acoustic noise faults in the wagon bearings of the Vale, in order to identify unknown patterns.

Association algorithms were used in the experiments, with focus on the machine learning unsupervised. From the results obtained it was analyzed which rules are useful to assist in the process of preventive maintenance of wagons.

KEYWORDS: Predictive Maintenance; Big Data; Data Mining; Association Method; Failure Pattern.

LISTA DE FIGURAS

Figura 1 - Mapa da malha ferroviária brasileira	13
Figura 2 - Definições para Big Data	20
Figura 3 - Estrutura de Big Data.....	21
Figura 4 - Tela inicial do Power BI.....	22
Figura 5 - Tela inicial do software Weka	23
Figura 6 - Modo explorer do Weka	24
Figura 7 – Vagão Transportador de Minério	25
Figura 8 - Estrutura do rodeiro ferroviário composta por eixo, roda e rolamento	25
Figure 9 - Leitura da temperatura.....	26
Figura 10 - HotBox.....	27
Figura 11 - Railbam.....	27
Figura 12 - Rolamento degolado	28
Figura 13 - Hierarquia de aprendizado	32
Figura 14 - Tarefas do algoritmo Apriori	35
Figura 15 - Todos itemsets frequentes possíveis	36
Figura 16 - Conjuntos de itens frequentes para suporte mínimo de 30%.....	37
Figura 17 - Busca para regras de associação para um itemset	38
Figura 18 - Espaço de busca em profundidade.....	38
Figura 19 - Transações para construção da FP-tree.....	39
Figura 20 - Construção da FP-tree.....	40
Figura 21- Metodologia para descoberta de regras para manutenção preventiva de vagões de minério.....	43
Figura 22 - Tabela transacional dos dados correlacionados	47
Figura 23 - Tabela transacional somente com valores true	47
Figura 24 - Atributos e Instâncias do Weka	49

LISTA DE TABELAS

Tabela 1 - Categoria de gravidade associado aos tipos de falhas de enrolamento dos vagões	29
Tabela 2 - Regras geradas pelo Algoritmo Apriori para temperaturas menores que 20°C	50
Tabela 3 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas menores que 20°C	50
Tabela 4 - Regras geradas pelo Algoritmo Apriori para temperaturas entre 20°C e 30°C	51
Tabela 5 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas entre 20°C e 30°C	51
Tabela 6 - Regras geradas pelo Algoritmo FP-Growth para temperaturas entre 20°C e 30°C	52
Tabela 7 - Regras geradas pelo Algoritmo Apriori para temperaturas entre 30°C e 40°C	52
Tabela 8 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas entre 30°C e 40°C	53
Tabela 9 - Regras geradas pelo Algoritmo FP-Growth para temperaturas entre 30°C e 40°C	53
Tabela 10 - Regras geradas pelo Algoritmo Apriori para temperaturas entre 40°C e 50°C	53
Tabela 11 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas entre 40°C e 50°C	54
Tabela 12 - Regras geradas pelo Algoritmo FP-Growth para temperaturas entre 40°C e 50°C	54
Tabela 13 - Regras geradas pelo Algoritmo Apriori para temperaturas entre 50°C e 60°C	54
Tabela 14 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas entre 50°C e 60°C	54
Tabela 15 - Regras geradas pelo Algoritmo FP-Growth para temperaturas entre 50°C e 60°C	55
Tabela 16 - Regras geradas pelos Algoritmos Apriori e FP-Growth para temperaturas entre 60°C e 70°C	55

LISTA DE ABREVIATURAS E SIGLAS

CCO	Centro de Controle Operacional
DNIT	Departamento Nacional de Infraestruturas de Transportes
IOT	<i>Internet Of Things</i>
GPL	<i>General Public License</i>
EFC	Estrada de Ferro Carajás
EFVM	Estrada de Ferro Vitória a Minas
KDD	<i>Knowledge Discovery in Databases</i>
TMPM	Terminal Marítimo de Ponta da Madeira
UFMA	Universidade Federal do Maranhão

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1. INTRODUÇÃO	12
1.1. Objetivos	14
1.2. Motivação.....	15
1.3. Organização do trabalho	16
2. TRABALHOS RELACIONADOS	17
3. FUNDAMENTAÇÃO TEÓRICA	19
3.1. <i>Big Data</i>	19
3.2. <i>Softwares</i> para análise de dados.....	21
3.2.1. Power BI.....	21
3.2.2. Weka.....	22
3.3. Contexto Ferrovia	23
3.4. Aprendizado de Máquina	30
3.5. Algoritmos de associação	32
3.5.1. Apriori.....	35
3.5.2. FP Growth.....	37
3.5.3. Predictive Apriori.....	41
4. METODOLOGIA PROPOSTA PARA APLICAÇÃO DE ALGORITMOS DE ASSOCIAÇÃO	43
4.1. Base de dados.....	43
4.2. Pré-processamento dos dados	44
4.2.1. Limpeza dos dados	44
4.2.2. Redução dos dados	45
4.2.3. Representação dos dados	45
4.3. Regras de Associação	46
5. RESULTADOS.....	49
5.1. Experimento para temperaturas menores que 20°C.....	49
5.2. Experimento para temperaturas entre 20 e 30°C.....	51
5.3. Experimento para temperaturas entre 30 e 40°C.....	52
5.4. Experimento para temperaturas entre 40 e 50°C.....	53
5.5. Experimento para temperaturas entre 50 e 60°C.....	54
5.6. Experimento para temperaturas entre 60 e 70°C.....	55

5.7. Análise de resultados dos experimentos	55
6. CONCLUSÃO	59
REFERÊNCIAS BIBLIOGRÁFICAS	61

1. INTRODUÇÃO

Atualmente, a informação em tempo real é imprescindível para as organizações, sejam estas de pequeno, médio ou grande porte. A informação é importante para o acompanhamento dos indicadores internos de desempenho, reduzindo custos e aumentando a produtividade. Para obter informação útil, é necessário ter um fluxo de dados constante, resultando em uma necessidade de tratamento, de forma que se consiga analisar fatores e variáveis distintos, sendo correlacionados aos objetivos e metas que devem ser alcançados.

A necessidade de análise dos dados para obtenção de informação resulta no estudo sobre a melhor forma de correlacionar e apresentar esses dados. A importância deste trabalho está no fato de propor uma solução para problemas relacionados à dificuldade de análise de um banco de dados que manipula dados provenientes de sensores na malha ferroviária da Vale.

O sistema ferroviário no Brasil tem uma malha densa nas regiões do centro-oeste, sul e sudeste, com uma participação mais tímida na região norte do país, como pode ser visto na Figura 1. Dentre suas maiores vantagens, o sistema ferroviário se destaca pela sua capacidade de transportar grandes volumes de carga se comparado ao sistema rodoviário. Por isso, o transporte ferroviário tem alta produtividade, sendo a eficiência no transporte de cargas indispensável.

Esse tipo de transporte é utilizado como intermediário para escoar a produção agrícola de soja, milho e outros, transportar *commodities*, que representa a maior parcela na carga transportada, assim como o transporte de combustível, níquel, cobre, carvão, manganês e outros. Isso comprova a importância no desenvolvimento deste setor e sua representatividade para o Brasil.

O centro de controle operacional (CCO) coordena todo o tráfego de trens na malha ferroviária. O CCO é responsável por enviar todos os comandos por meio de um sistema de transmissão de dados, cuja função é executar os comandos de (BOZI, 2005):

- a) Movimentação dos aparelhos de mudança de via, de acordo com a operação das máquinas de chave;
- b) Conceder as rotas aos trens na malha;
- c) Licenciamento da circulação dos trens, através de sinais de cabine e/ou externos, como semáforos na via;

- d) Controle de acesso de entrada e saída da via férrea sinalizada e das passagens de nível¹;
- e) Controle das paradas dos trens quando ocorre alguma anomalia.

Figura 1 - Mapa da malha ferroviária brasileira



Fonte: <http://infologis.blogspot.com/2011/01/mapa-ferrovias-no-brasil.html>

¹ **passagem de nível** ou **passagem em nível** é um tipo de cruzamento ao mesmo nível entre uma **ferrovia** e um caminho ou estrada.

Neste estudo, o sistema de detecção de falhas ferroviárias é abordado. Ele é composto por inúmeros equipamentos de sensoriamento distribuídos ao longo da ferrovia, que têm como papel fundamental medir todas as condições dos ativos do sistema durante o traslado dos trens na malha ferroviária e garantir a segurança de todos os envolvidos no transporte.

A avaliação das propriedades dos ativos do trem é uma das principais condições de segurança durante o traslado. Como os vagões são responsáveis por carregar toda a carga transportada, seu bom funcionamento é essencial para o ciclo de transporte se manter constante. Portanto, o objeto desse estudo são os dados de falhas dos rolamentos dos vagões e suas características.

Os dados analisados são provenientes de equipamentos instalados na via que realizam medições de algumas características dos ativos presentes no trem ao decorrer da viagem. Os equipamentos utilizados neste trabalho são o equipamento Railbam, que é capaz de detectar os tipos de falhas acústicas² dos rolamentos dos vagões, e o equipamento Hot Box, que tem a tarefa de medir as temperaturas dos rolamentos dos vagões durante toda a viagem do trem.

A partir desses dados, foi construída uma base de dados correlacionando as informações dos dois equipamentos e utilizados algoritmos de associação para descoberta de padrões de falhas.

1.1. Objetivos

Os objetivos estão organizados em geral e específicos.

Esta monografia tem como objetivo geral realizar uma análise dos dados de medições de temperaturas e ruídos acústicos de equipamentos ferroviários, caracterizados como *Big Data* devido ao seu grande volume e variedade. A atualização desses dados acontece em tempo real, o que mostra a necessidade de tratamento e análise desses dados de forma dinâmica e rápida a fim de gerar informações úteis para tomadas de decisões assertivas.

Os dados serão utilizados em uma análise sobre as perdas operacionais geradas pelos eventos de falha dos ativos ao longo da ferrovia com o intuito de auxiliar na apuração de ocorrências ferroviárias, assim como, o tempo de reconhecimento da falha até a manutenção desses.

Os objetivos específicos são os seguintes:

- a) Construir uma base de dados contendo histórico de temperatura, tipos de falhas acústicas, velocidade de cada rolamento, presença de ruído e direção do trem;

² Falhas acústicas são diagnosticadas a partir da medição do equipamento Railbam, de acordo com a leitura da frequência e amplitude dos sons emitidos pelos rolamentos de um trem durante a viagem.

- b) Descobrir os padrões de falhas associados a características específicas intrínsecas ao processo de transporte de carga realizado pela viagem de trem;
- c) Verificar se as variáveis escolhidas são satisfatórias para utilizar os algoritmos de associação;
- d) Avaliar os resultados obtidos de acordo com a utilização dos algoritmos de associação e fazer análise comparativa entre essas abordagens.

1.2. Motivação

Observou-se uma necessidade de aperfeiçoar as técnicas de análise de dados para aumentar a eficiência e a qualidade no processo de manutenção preventiva dos vagões de trem de minério da Vale. Visto que a Vale é uma empresa inserida em um contexto global, sendo a principal vantagem da Vale a qualidade do seu produto por apresentar 67% de teor de minério de ferro, proveniente da principal mina de extração de minério localizada em Carajás no Pará.³

No entanto, a Vale está em desvantagem em relação as suas concorrentes quanto a sua localização geográfica. Neste caso, a redução do custo de transporte é um fator importante para a empresa aumentar a competitividade internacional. Por isso, é importante analisar todos os dados, gerando informação, a fim de realizar um processo confiável e com baixas perdas.

De acordo com o Departamento Nacional de Infraestruturas de Transportes (DNIT), o cenário de crescimento é favorável para o transporte ferroviário com novos investimentos no setor. O crescente volume transportado na malha ferroviária da Vale é, infelizmente, acompanhado por um volume de perdas do sistema, seja por falhas na composição de vagões, seja pela intervenção da comunidade por meio de bloqueios, seja por vândalos que destroem sensores e atuadores, impactando no desempenho operacional do transporte ferroviário de minério.

Esses impactos que geram perdas no sistema ocasionam interrupções de circulação dos trens na malha ferroviária, visto que para cada evento desse é necessário realizar as devidas tratativas em função da origem das falhas. Independente se aconteceu uma falha em um ativo do trem ou atuação de vândalos é necessário realizar os devidos acionamentos necessários e resolver o problema momentaneamente.

As interrupções de circulação dos trens são decorrentes de diversos tipos de problemas na ferrovia, as quais são avaliadas mediante cada cenário específico. Cada interrupção de

³ Site: <http://www.vale.com/brasil/PT/business/mining/iron-ore-pellets/Paginas/default.aspx>

circulação gera parada de um ou mais trens, logo, se não há circulação de trens em um determinado local da via férrea o transporte de minério não é realizado no tempo previsto.

Além disso, existem inúmeros fatores que ocasionam indisponibilidade da linha férrea ou redução da capacidade de transporte momentânea, por exemplo, restrição de velocidade em determinado local. Portanto, tudo o que gera perdas para o sistema é mensurado em quantidade e duração da interrupção de circulação dos trens.

Os impactos operacionais em decorrência de falhas em vagões serão objeto de estudo deste trabalho, devido ao grande volume de dados registrados e a necessidade de interpretação destes. Portanto, observou-se a oportunidade de correlacionar as perdas com os tipos de eventos ocorridos e verificar a recorrência dos eventos de acordo com o local, tipo de falha, tempo de interrupção da circulação e/ou de energia, a data, responsabilidade do evento e causa da ocorrência.

O intuito desse estudo é descobrir comportamentos recorrentes, os quais não são possíveis de determinar no cotidiano, devido ao alto fluxo de dados. Então, esse estudo visa viabilizar a geração de informação útil, afim de relacionar os impactos operacionais com as causas dos eventos. Estes impactos têm ligação direta com os custos de produção, transporte e manutenção da empresa.

1.3. Organização do trabalho

Este manuscrito está organizado conforme a seguir:

- Capítulo 1: Introdução e apresentação dos objetivos;
- Capítulo 2: Fundamentação teórica, abordando os diversos temas que estiveram relacionados ao realizar o trabalho, tais como o contexto sobre a ferrovia, séries temporais, Weka, Power BI e algoritmos de associação;
- Capítulo 3: Descrição da metodologia proposta para descoberta de regras para detectar falhas em vagões de trem, apresentando as etapas realizadas durante o trabalho;
- Capítulo 4: Apresentação dos resultados obtidos com a utilização dos algoritmos de associação;
- Capítulo 5: Conclusões a respeito dos resultados obtidos pelos algoritmos de associação.

2. TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos relacionados ao tema aprendizado de máquina, em específico, algoritmos de associação e ferramentas para suporte a extração de conhecimento a partir de banco de dados, focando na descoberta de padrões, por exemplo padrões de falhas ou até mesmo de doenças.

Dentre essas diversas aplicações possíveis, o objetivo principal é a descoberta de informações implícitas do conjunto de dados e, conseqüentemente, a geração de regras de associações. As regras úteis encontradas são utilizadas para suporte à tomada de decisão, previsão financeira, políticas de marketing, diagnósticos médicos, dentre outras.

Diversos trabalhos foram feitos com intuito de prever e classificar as características intrínsecas à ocorrência dos eventos de Hot Box em ferrovias pelo mundo todo. É o caso do estudo realizado em uma ferrovia dos Estados Unidos (LI; PARIKH; HE,2014) que utilizou dados de temperaturas provenientes do equipamento de Hot Box e de acústica provenientes do equipamento Railbam. Esse trabalho desenvolveu um modelo para classificar os eventos em duas classes, alarmado ou não alarmado, com antecedência de 3 e 7 dias tendo o *Support Vector Machine* como técnica de classificação escolhida. Foram 55 características escolhidas para análise e foram obtidos resultados satisfatórios para previsão de 7 dias antes, a maior taxa de acertos positivos foi de 91,54%, e para previsão de 3 dias antes a maior taxa de acerto foi de 92,56%.

Esse tipo de análise requer processamentos de grandes bancos de dados. Para avaliar o desempenho dos algoritmos de associação, um estudo de comparação entre alguns desses algoritmos para mineração de dados (GYORODI; GYORODI; HOLBAN, 2004) foi realizado, comparando os algoritmos Apriori e FP-Growth. A principal diferença entre os dois algoritmos está no método utilizado para a geração das regras, o Apriori se baseia na geração de combinações dos conjuntos de itens frequentes, enquanto o FP-Growth utiliza-se do conceito de dividir para conquistar e gerar as regras de forma mais eficiente que o Apriori.

Em (GONÇALVES, 2005), um estudo é apresentado sobre a avaliação das medidas de interesse utilizadas para determinar a qualidade dessas regras de associação, que descrevem os padrões encontrados entre os itens de um conjunto de dados. As medidas de interesse podem ser subdivididas em objetivas e subjetivas, de tal forma que medidas objetivas utilizam-se de

técnicas e modelos matemáticos da estatística para identificar a força de uma regra⁴, enquanto as medidas subjetivas são baseadas na opinião de um analista para determinar a força da regra.

Em outro estudo (PATIL et al., 2011), a geração de regras de associação é utilizada na medicina para auxiliar médicos durante a tomada de decisão rotineira. Nesse trabalho, os dados são referentes aos casos de pessoas diagnosticadas com diabetes, sendo os dados discretizados através da aproximação e divisão dos dados em intervalos categóricos determinados pela equipe médica. As regras foram geradas utilizando os algoritmos Apriori e *predictive* Apriori. A partir dos resultados encontrados, o *predictive* Apriori demonstrou sua capacidade em extrair regras de maior qualidade, porém com uma performance de processamento demorada em relação ao Apriori.

Através dos trabalhos apresentados, percebe-se que os algoritmos de associação são aplicados para diversos tipos de análise. Em relação ao objetivo proposto nesse trabalho é interessante ressaltar que esse tipo de estudo no contexto ferroviário é uma aplicação nova.

O estudo citado anteriormente (LI; PARIKH; HE,2014) aplicou algoritmos para predição da classificação dos eventos, enquanto este trabalho aplicou algoritmos de associação para encontrar padrões, correlacionando o dados equipamentos Hot Box e Railbam.

⁴ Força de uma regra é um termo utilizado para indicar qual o potencial de uma regra gerada por um algoritmo, em função das medidas calculadas pelo algoritmo indicando qual a regra mais forte, ou seja, mais adequada.

3. FUNDAMENTAÇÃO TEÓRICA

A tecnologia progrediu consideravelmente, como exemplos, temos IOT (*Internet of Things*), *smart grid*, dentre outras inúmeras tecnologias que exercem grande influência na vida cotidiana, no trabalho ou em quaisquer outras situações presentes na vida do ser humano. O que não é possível enxergar a olho nu é a grande quantidade de dados que estão sendo gerados, transmitidos e armazenados em todos os instantes. Por isso, observou-se a necessidade de estudar como analisar esses dados, chamados de *Big Data*.

Sendo essa nova concepção de dados amplamente utilizada como objeto de estudo para desenvolvimento e aprimoramento de novas tecnologias. Bem como aplicações de sistemas de energia, como a *smart grid*, pois esta faz uso da análise de *Big Data* proporcionando acessibilidade do consumidor aos dados de consumo em tempo real, assim como, previne o mesmo quanto à interrupções no fornecimento de energia e informa ao usuário a relação entre o horário e o preço cobrado pela energia, assegurando a qualidade e eficiência do serviço.

3.1. *Big Data*

Big Data é definido por um conjunto de dados que não consegue ser processado por sistemas de base de dados convencionais, devido ao seu grande volume de dados que são gerados a todo instante. Informações de diferentes fontes, tipos e formatos que não conseguem ser analisadas por hardwares e softwares tradicionais, sendo um desafio desde o processamento, armazenamento, até a garantia da segurança de tais informações.

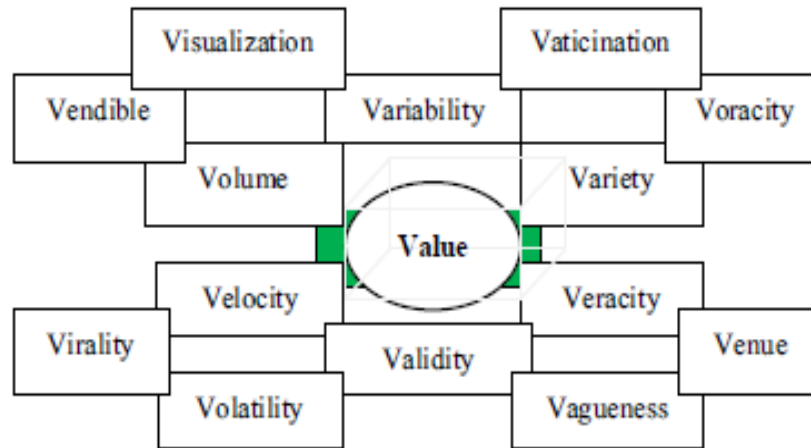
Esse termo é instantaneamente associado à definição dos três V, que significam volume, variedade e velocidade. Porém, nem sempre se tem somente essas referências como definição para um determinado conjunto de dados, logo, pode-se ter outras definições características associadas, como valor, variabilidade, visualização, veracidade, dentre outros, conforme ilustra a Figura 2 (DAVE, 2017).

A informação gerada em tempo real é característica das organizações, sejam estas de pequeno, médio ou grande porte. Esse fluxo de dados constante revela a necessidade de um tratamento de informações, de forma que se consiga analisar fatores e variáveis distintas sendo correlacionadas aos objetivos e metas que devem ser alcançados mutuamente.

Observa-se a importância de identificar as oportunidades para realizar análise de *Big Data*, determinando como e quando este tipo de análise poderá ser utilizado, como sua implementação dentro das empresas para auxiliar na tomada de decisão e na otimização de processos, proporcionando diferentes tipos de vantagens competitivas para determinada organização. Essa análise pode ser realizada em 5 fases como, aquisição/gravação,

extração/limpeza/anotação, integração/agregação/representação, análise/modelagem, interpretação (DAVE, 2017).

Figura 2 - Definições para Big Data



Fonte: (DAVE, 2017)

Devido à complexidade, diversidade e necessidade de grande espaço disponível para armazenamento de dados não deve ser considerada uma só definição específica para *Big Data*. Portanto, o termo *Big Data* pode ser definido de inúmeras formas, de acordo com a complexidade e particularidades de cada situação problema que se deseja solucionar.

Atualmente, com o desenvolvimento de novas tecnologias os dados têm certos níveis de complexidade e podem ser armazenados em diversos formatos de arquivos. Observa-se que cada situação tem sua estrutura de dados particular, a qual indica os formatos de arquivos em que os dados são gerados e armazenados.

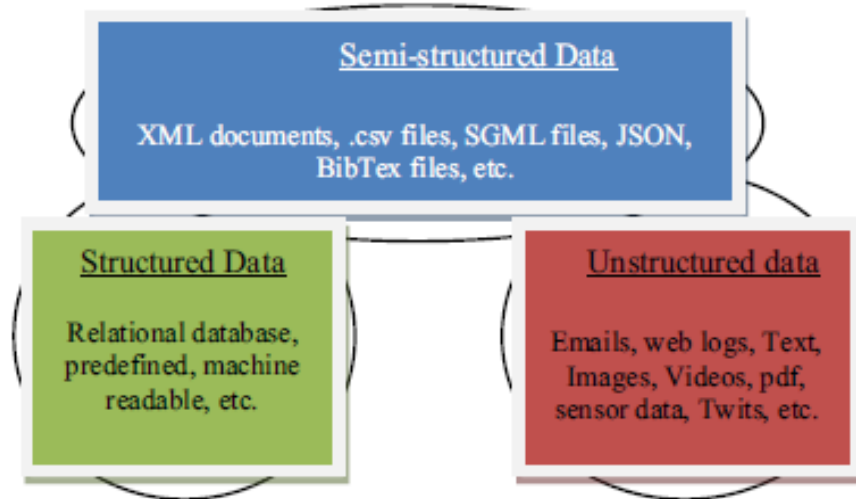
A alta frequência de geração de dados e seu grande volume tem desafiado os profissionais do ramo quanto a gestão, aumento da capacidade de armazenamento, processamento e análise do problema proposto. E de acordo com o crescimento de volume de dados armazenados é necessário aumentar a capacidade de armazenamento, porém isso demonstra um menor grau de conhecimento do usuário sobre este grande volume de dados armazenados.

Por isso, é necessário definir a estrutura de dados para a situação estudada. Geralmente, o termo 'dados estruturados' é aplicado a bases de dados, enquanto 'dados não estruturados' remete ao restante que é produzido através de tecnologias diferentes, como *e-mails*, páginas *web*, vídeos, pdf, dentre outros, conforme ilustra a Figura 3 (DAVE, 2017).

Para realizar a análise de *Big Data* é necessário a utilização de alguns métodos e programas para desvendar o valor e o significado dos dados, afim de descobrir as correlações entre as informações relevantes. Tais métodos como, evitar movimentação intensa de dados

realizando o armazenamento adequado e processamento das informações no mesmo local. É imprescindível garantir a segurança das informações, o que é mais difícil em base de dados não estruturadas e ferramentas de análise de código aberto.

Figura 3 - Estrutura de Big Data



Fonte: (DAVE, 2017)

3.2. Softwares para análise de dados

Os softwares Power BI e Weka são destinados à análise de conjuntos de dados. O primeiro foi utilizado para analisar, manipular e adaptar o conjunto de dados para o formato requerido pelo segundo software, o Weka.

3.2.1. Power BI

O Power BI é um software da Microsoft que permite a aquisição e relacionamento de dados de diversas fontes e tipos variados, realizando análises das informações de forma dinâmica e interativa. Assim como, é possível implementar uma rotina para atualização dos dados automaticamente. Na Figura 4 é mostrado a tela inicial do Power BI⁵.

Por exemplo, utilizando essa ferramenta é possível conectar bancos de dados, arquivos em excel, servidores, serviços em nuvem, dentre outras fontes, sendo possível criar relacionamento entre essas diversas fontes e a partir do gerenciamento desses dados criar relatórios interativos, publicar tais relatórios online e compartilhar com diversas pessoas.

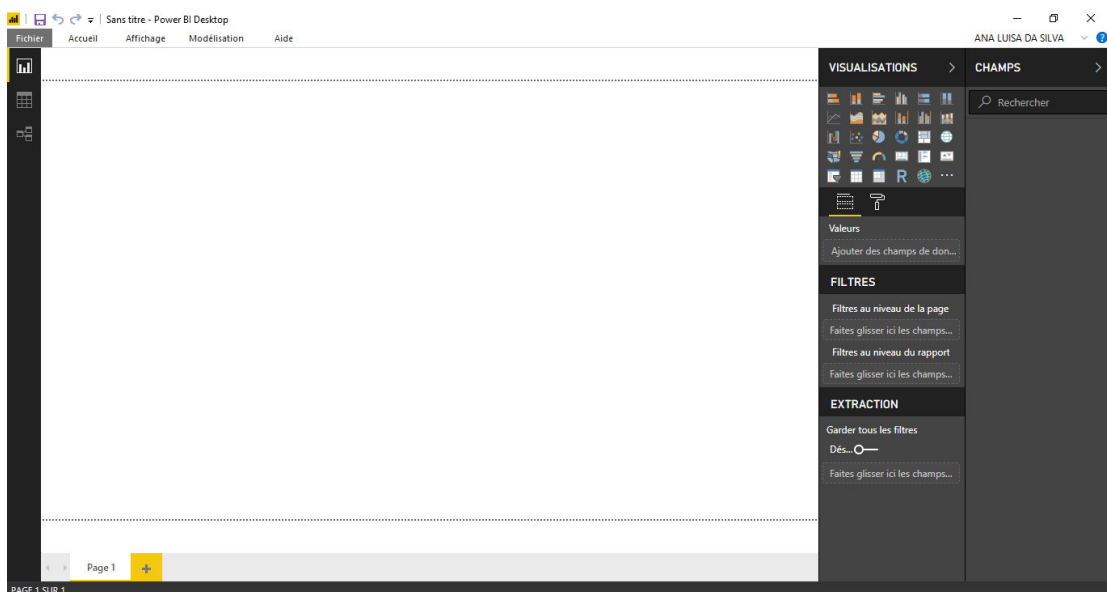
⁵ Site: <https://powerbi.microsoft.com/pt-br/>

Esses relatórios são úteis para criação de uma interface visual com o usuário, permitindo uma análise de dados de acordo com a perspectiva individual de cada usuário através de filtros disponíveis no ambiente. Para os relatórios é possível criar dados dinâmicos baseados nas entradas escolhidas, em função da necessidade do usuário.

Essa ferramenta permite que o usuário gerencie seus dados, possibilitando uma manipulação mais fácil dos dados e, da mesma forma, a análise destes, sendo possível construir modelos e correlacionar diversas bases de dados, de acordo com a necessidade.

O objetivo de utilizar essa ferramenta neste trabalho é a possibilidade de correlacionar os dados das medições em um só local. Além de ser possível manipular os dados, criar funções personalizadas com os dados pré-existent e alterar os dados para o formato desejado, adaptando a base de dados para o algoritmo que será utilizado na análise dos dados.

Figura 4 - Tela inicial do Power BI



Fonte: Autor.

3.2.2. Weka

Weka é um *software* de código aberto desenvolvido na linguagem Java pela Universidade de Waikato, situada na Nova Zelândia. O software é capaz de realizar diversas tarefas de mineração de dados e utiliza a GNU *General Public License* (GPL), o que torna possível o seu uso para fins estudantis⁶. Na Figura 5 é mostrado a tela inicial do Weka.

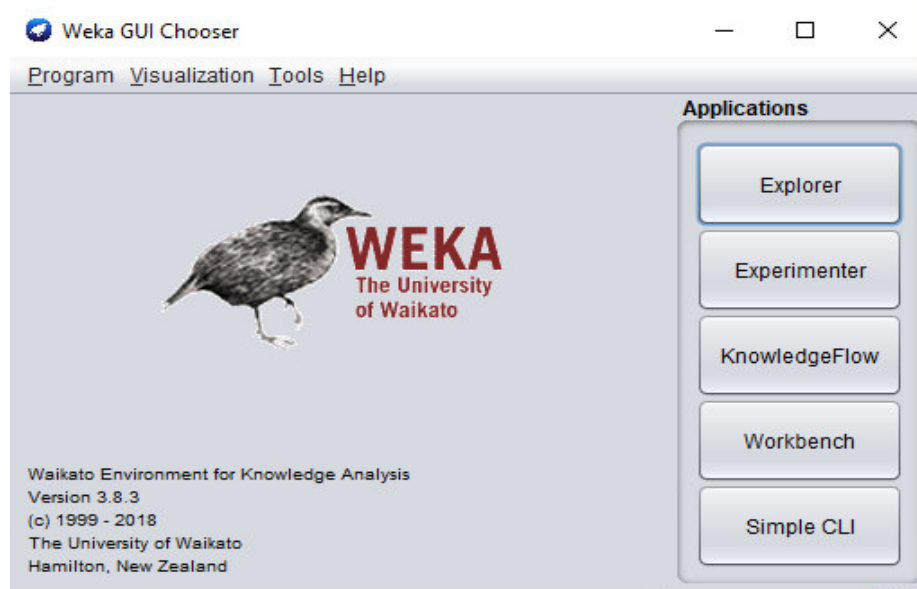
⁶ Site: <https://www.cs.waikato.ac.nz/ml/weka/>

A partir da tela de inicialização do software é possível escolher qual modo de trabalho o usuário deseja trabalhar com o seu conjunto de dados no programa, conforme pode ser visto na Figura 5.

Neste trabalho, o modo de trabalho fornecido pela opção *explorer* foi utilizado, conforme é visto na Figura 6. O Weka permite que o usuário insira seus dados de entrada na aba de pré-processamento, observe as instâncias e atributos gerais do conjunto de dados inicial, selecione os atributos e a classe desejada para análise. Após essas definições o usuário seleciona qual será o modelo de análise realizado, tais como classificação, agrupamento, associação, dentre outros.

Nesse estudo o Weka foi utilizado para realizar tarefas de associação do conjunto de dados, utilizando os algoritmos de associação fornecidos pelo *software* Weka.

Figura 5 - Tela inicial do *software* Weka



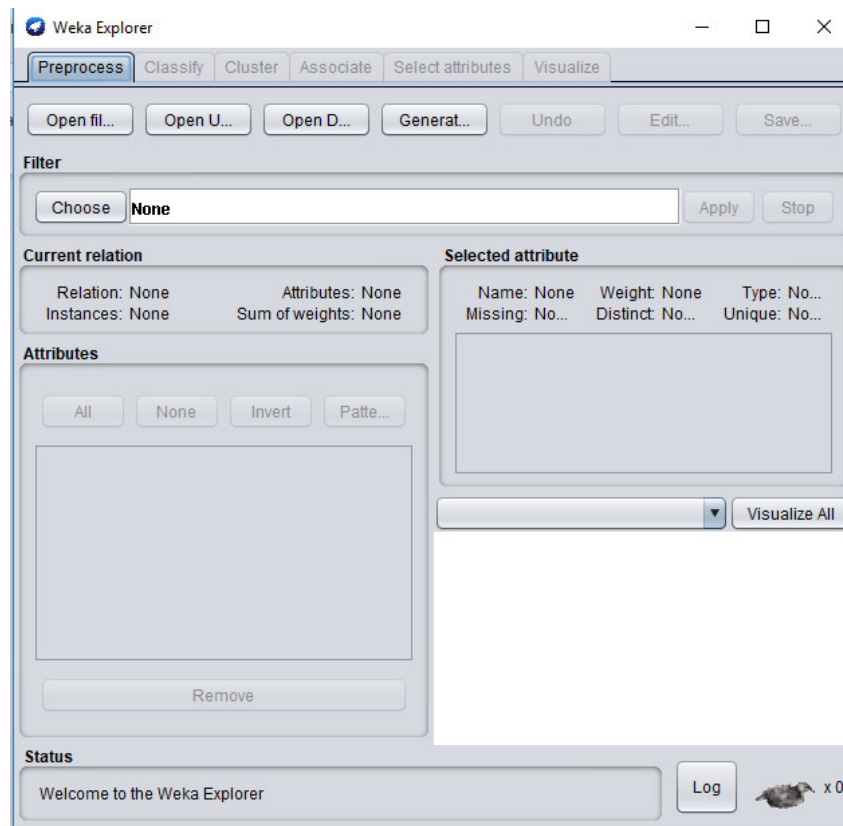
Fonte: Autor.

3.3. Contexto Ferrovia

A Vale opera duas ferrovias no Brasil, a Estrada de Ferro Carajás (EFC) e a Estrada de Ferro Vitória a Minas (EFVM), com cerca de 2 mil quilômetros de malha ferroviária. A EFC tem duas linhas com 972 quilômetros de extensão e 85 quilômetros de extensão do ramal ferroviário no sudeste do Pará, conectando o terminal marítimo de Ponta da Madeira até as três minas localizadas no Pará⁷.

⁷ Site: <http://www.vale.com/brasil/PT/business/logistics/railways/Paginas/default.aspx>

Figura 6 - Modo explorer do Weka



Fonte: Autor.

A ferrovia é utilizada para o transporte em larga escala de minério de ferro, no entanto, existem outros materiais que são transportados em menor quantidade, tais como combustível, soja, carvão, celulose, dentre diversos outros.

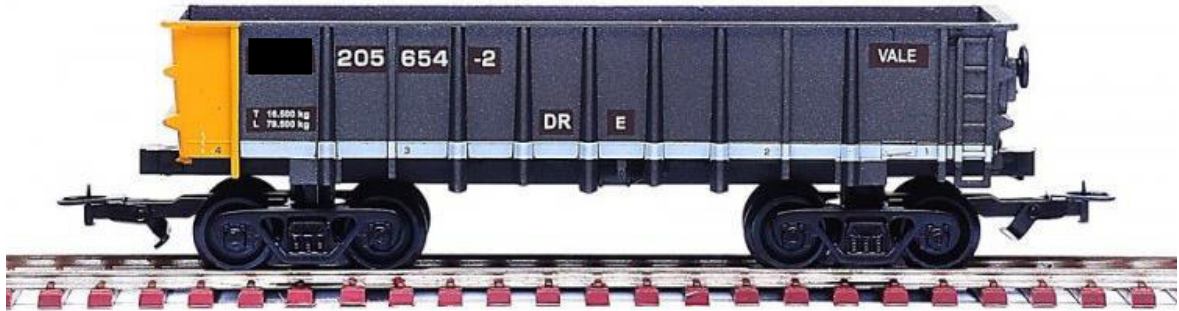
Os trens responsáveis pelo transporte de minério de ferro têm em média 3.3 quilômetros de comprimento, sendo compostos por 3 ou 4 locomotivas e 330 vagões. A empresa possui cerca de 19.000 vagões. Para este estudo, os vagões de tipo GDT e GDU são considerados, que são utilizados para o transporte de minério de ferro capazes de transportar entre 130 a 150 toneladas de peso bruto. Na Figura 7 é visto um modelo de vagão.

Cada vagão possui 4 eixos, cada eixo possui 2 rodas e 2 rolamentos, totalizando 8 rodas e 8 rolamentos por vagão que são monitorados por equipamentos distribuídos ao longo da ferrovia, conforme a Figura 8. Rodeiro ferroviário é a estrutura composta por duas rodas, um eixo e um par de rolamentos.

Neste estudo o foco é o comportamento das medições de temperatura e ruídos acústicos dos rolamentos desses vagões. Em decorrência do esforço do vagão durante o deslocamento do

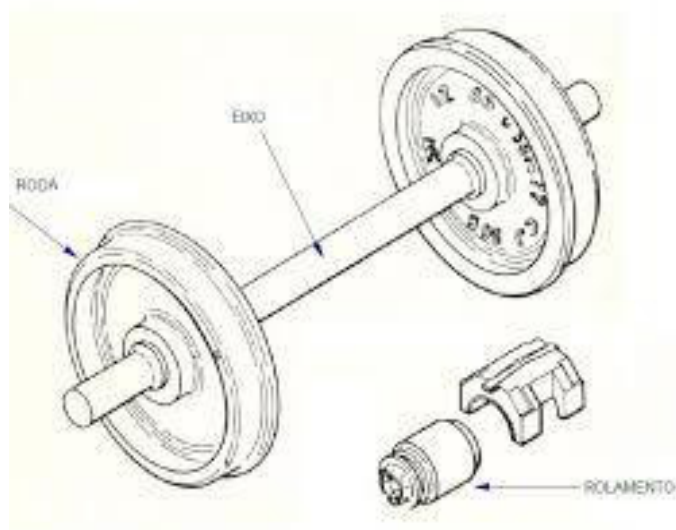
transporte da carga, além de outras influências devido ao comportamento dinâmico da condução do trem e estrutura da via.

Figura 7 – Vagão Transportador de Minério



Fonte: <http://shopferreo.com.br/produto/vagao-gondola-de-minerio-cvrd-frateschi-2091/6652>

Figura 8 - Estrutura do rodeiro ferroviário composta por eixo, roda e rolamento



Fonte: <http://www.brasilferroviario.com.br/partes-dos-vagoes/>

É possível observar diversos tipos de avarias nas rodas e rolamentos que podem gerar um aumento de temperatura, como: fadiga, corrosão por atrito ou sob tensão, contaminação por óleo lubrificante, dentre outras causas, decorrentes do processo de transporte de cargas.

Para realizar a medição de temperatura do rolamento o equipamento utilizado é chamado de Hot Box e está presente em 14 pontos específicos ao longo da ferrovia.

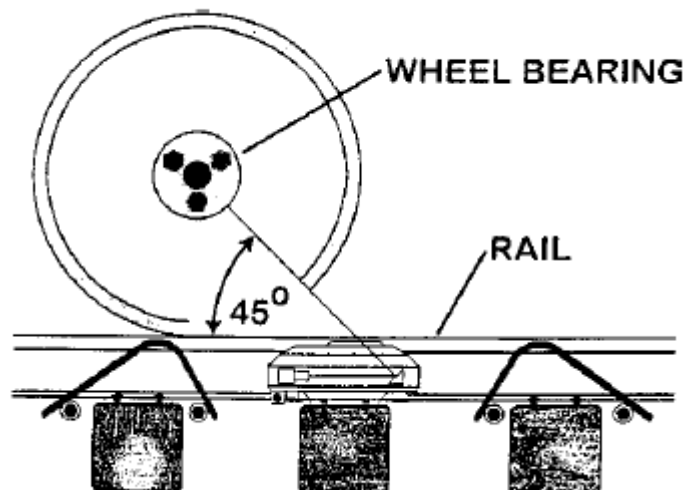
Este dispositivo tem como seu princípio de funcionamento a atuação de transdutor que produz um fluxo magnético constante, a medida que uma roda atravessa o fluxo, então o fluxo

é cessado momentaneamente, sendo possível detectar a passagem do trem, contagem dos eixos associando a temperatura medida de cada eixo, velocidade e sentido de deslocamento do trem.

Quando o transdutor tem seu fluxo cessado, a leitura das temperaturas é acionada através da atuação de scanners tanto da caixa do vagão, quanto da roda do mesmo. Os scanners são constituídos de sensores infravermelhos chamados de pirômetros, que detectam os raios infravermelhos emitidos pelas caixas de rolamento e rodas dos vagões.

Os scanners utilizam um ângulo de 45 graus para medir a temperatura dos rolamentos de acordo com o manual do equipamento Hot Box, como pode ser visto na Figura 9 e o equipamento na Figura 10.

Figure 9 - Leitura da temperatura



Fonte: Manual Técnico do Hot Box, 2011

Para realizar as medições de ruídos sonoros, o equipamento utilizado é o Railbam que se encontra em um local específico, devido a sua robustez. Esse equipamento é composto por diversos sensores, sendo estes capazes de detectar se o som proveniente dos rolamentos dos vagões é característico de alguma falha já parametrizada. O equipamento é visto na Figura 11.

As falhas são classificadas de acordo com um prefixo que indica se o rolamento está emitindo algum barulho estranho, interferindo na detecção da falha. Além disso, o sistema caracteriza o tipo, nível e descrição da falha através do som captado.

Enquanto o HotBox mede a temperatura dos rolamentos dos vagões por 14 vezes diferentes ao decorrer da viagem do trem, o Railbam capta o som dos rolamentos apenas 1 vez ao decorrer da viagem, pois este equipamento está presente em um único local da ferrovia.

Figura 10 - HotBox



Fonte: Autor.

Figura 11 - Railbam



Fonte: Autor.

A velocidade ideal deve ser acima de 30km/h para passagem nos locais com esses equipamentos presentes para que a leitura seja no momento em que o ativo esteja realizando um esforço significativo durante a viagem do trem.

Há casos em que a temperatura identificada é tão elevada que o rolamento é diagnosticado como perda total, dificultando até sua retirada do ativo. Nesses casos, ocorreu a degola do rolamento, conforme a Figura 12.

Figura 12 - Rolamento degolado



Fonte: Autor.

O manual do equipamento Railbam descreve as notações das falhas no item 3, conforme a seguir:

PREFIXO – Ruído que pode interferir na detecção da falha

- Clpd – Sinal cortado;
- Shrk – Ruído tonal;
- FBS – Flanqueamento, branqueamento e batendo;
- NOISY – Ruído desconhecido;

TIPO – Tipo da falha

- RS – Falha na superfície do rolamento;
- LF – Folga ou desgaste do rolamento;
- WHLFLT – Roda plana;

NIVEL – Severidade da falha

- 1 – Falha severa;
- 2 – Falha moderada;
- 3 – Falha pequena;
- 4 – Nenhuma falha identificada;

DESCRITOR – Detalhe sobre a falha

- _r – Falha no rolo;
- _p – Falha na capa;

- *_n* – Falha no cone;
- *_m* – Múltiplas falhas;
- *_e* – Falha estendida;
- *_s* – Falha estendida;

Além desses tipos de descrição das falhas que já foram pré-configuradas, é possível criar classificações adicionais ao sistema se desejado pelo administrador. De acordo com esse conceito, a severidade da falha pode ser vista como uma forma de nomear múltiplos tipos de falhas de mesma categoria, auxiliando no reconhecimento rápido. Conforme pode ser visto na Tabela 1.

Tabela 1 - Categoria de gravidade associado aos tipos de falhas de enrolamento dos vagões

Categoria da gravidade	Tipos	Níveis	Descritores	Prefixos
Clear Level 1	RS	1	<i>_p, _n, _r, _m, _e, _s</i>	Shrk ou sem prefixo
Clear Level 2	RS	2	<i>_p, _n, _r, _m, _e, _s</i>	Shrk ou sem prefixo
Wheel Flats	WHLFLT	1 ou 2		Sem prefixo
Potential 1&2	RS	1 ou 2		Sem prefixo

Fonte: Manual do equipamento Railbam de 2011

O alarme *Clear Level 1* refere-se aos tipos de falhas RS – falhas na superfície do rolamento com nível 1 e associação dos descritores, enquanto o alarme *Clear Level 2* refere-se aos tipos de falha RS – falha na superfície do rolamento com nível 2 e associação dos descritores, ambos podem ou não ter prefixos correspondentes aos ruídos detectados.

O alarme *Wheel Flats* refere-se ao tipo de falha WHLFLT – roda plana com nível 1 ou 2 e sem associação de descritores ou prefixos. O alarme *Potential 1&2* refere-se aos tipos de falhas RS – falhas na superfície do rolamento com nível 1 ou 2 sem associação de descritores e prefixos.

Neste estudo o alarme *Potential 1&2* foi desconsiderado. Contudo, outros dois alarmes foram parametrizados a partir da base de dados. Esses dois alarmes são *Potential 1* e *Potential 2* referentes as categorias de falhas RS – falhas na superfície do rolamento com nível 1 e nível 2, respectivamente, sem associação de descritores e prefixos.

Além disso, para este estudo foi considerado os alarmes *Potential 3* e *Clear Level 3*, sendo que o primeiro representa a categoria de falha RS – falhas na superfície do rolamento com nível

3 sem associação de descritores da falha e o segundo representa a mesma categoria de falha com associação dos descritores da falha detectada.

No entanto, esses alarmes, *Potential 3* e *Clear Level 3*, não estão parametrizados no equipamento Railbam. Esse equipamento permite a configuração de novos parâmetros, por isso foram criadas essas duas categorias de gravidade ou alarmes a mais para verificar a frequência e relevância de ocorrência desses. Além desses alarmes novos, as categorias de gravidade mostradas na Tabela 1 foram aplicadas durante a realização desse trabalho.

O sistema de monitoramento dos vagões do trem é composto pelos equipamentos citados anteriormente, HotBox e Railbam, além de outros que verificam erros de alinhamento, medição de perfil e detecção de impacto dos vagões.

Cada dispositivo desse já tem valores pré-definidos de alarmes para os sensores e de acordo com a gravidade do alarme é necessário tomar a decisão de parar a composição. Caso a decisão de parar o trem seja necessária, isso significa uma perda para o sistema de transporte.

Embora a decisão de parar o trem acarreta em uma perda para o sistema, esse tipo de deliberação é fundamental para casos de alarmes, principalmente, com gravidades elevadas, pois a parada do trem pode prevenir contra problemas futuros.

Em certas situações, o ativo é retirado de circulação quando há quebra e/ou não há solução momentânea. Problemas como esses inutilizam a linha férrea e, conseqüentemente, podem gerar vários outros problemas de perda para o sistema, como:

- Parada indevida dos trens;
- Perda de produtividade;
- Hora extra da equipe de manutenção que será acionada para corrigir o problema;
- Riscos de acidentes devido à parada indevida.

3.4. Aprendizado de Máquina

Essa área de estudo da inteligência artificial vem aprimorando a capacidade e o desenvolvimento das máquinas para o aprendizado. Inúmeros métodos e algoritmos foram criados para solucionar problemas complexos de acordo com o comportamento dos dados históricos.

A máquina aprende por meio de funções e técnicas o comportamento do problema e reaplica o seu aprendizado. O intuito é capacitar os sistemas modernos a aprender de forma automática, sendo esses capazes até de aprender com dados imperfeitos.

Segundo citou CARVALHO (2011), o estudo de aprendizado de máquina pode ser desenvolvido de acordo com dois modelos: a) preditivo e b) descritivo.

- a) O modelo preditivo é baseado no aprendizado de uma função aproximada da função desconhecida, que representa o comportamento dos dados, e essa função aproximada permite estimar os futuros valores para valores observados posteriormente. Em função da natureza dos dados é possível determinar se o problema é de classificação ou regressão. E de acordo com o custo associado às previsões do modelo é possível determinar a qualidade de determinado modelo preditivo.
- b) O objetivo do modelo descritivo é encontrar padrões ou tendências para auxiliar a tomada de decisão, visto que não é possível determinar uma meta a ser perseguida. Diante disso é possível utilizar três tipos de tarefas: sumarização, associação e agrupamento. Esse tipo de modelo não requer conhecimento prévio sobre os dados, por isso está intensamente associado a mineração de dados para a descoberta de conhecimento em base de dados.

A Figura 13 mostra os tipos de aprendizado de máquina, sendo esses o supervisionado e o não supervisionado. Esses tipos de aprendizados são aplicados conforme o objetivo do trabalho a ser realizado, basicamente o modelo preditivo se aplica as tarefas com intuito de prever alguma característica específica, enquanto o modelo descritivo deve ser aplicado para tarefas com intuito de descobrir algum tipo de padrão, a partir do comportamento dos dados estudados.

Figura 13 - Hierarquia de aprendizado



Fonte: (CARVALHO, 2011)

A partir do modelo escolhido é possível determinar a capacidade de compreensão e generalização desse modelo que é imprescindível, sendo possível determinar essa capacidade baseado na influência dos dados de erro, ou seja, dados de comportamentos bem distintos.

A diferença entre os dois modelos de aprendizado é devido ao fato de o aprendizado supervisionado precisar conhecer a saída rotulada de cada exemplo do conjunto de dados de entrada, para assim, prever o valor de saída para os novos exemplos. Utilizado para tarefas de classificação, no qual os rótulos são classes de valores discretos; regressão, no qual os rótulos são classes de valores contínuos.

Enquanto o modelo de aprendizado não supervisionado não necessita de rótulo de saída, visto que este tem o objetivo de explorar e descrever o conjunto de dados. Com a finalidade de extrair padrões comportamentais dos dados sem um supervisor. Utilizado para tarefas de agrupamento, sendo os dados separados em grupos; associação, capaz de definir padrões de acordo com a associação dos dados em função da frequência; sumarização, tem a meta de encontrar uma descrição de acordo com o comportamento dos dados.

3.5. Algoritmos de associação

Uma vez que esses algoritmos não possuem o atributo classificar, estes são derivados do método de descrição. A caracterização de aprendizagem não supervisionado dá-se ao fato de que as regras são geradas a partir da observação de comportamento dos atributos. Esses

atributos podem conter somente dois valores, sendo esses valores *true* ou *false*, o que caracteriza a tabela transacional.

A tabela transacional é composta por uma quantidade de atributos, todos de valores booleanos. Cada atributo é verificado de acordo com a mudança do seu valor, o que é chamado de transação, ou seja, uma mudança do valor daquela variável. Essas transações determinam as condições de ocorrência de padrões de acordo com os valores dos atributos que podem ser associados. Esse tipo de algoritmo analisa as mudanças de transações entre as linhas da tabela, onde cada linha dessa tabela significa uma ocorrência dos eventos estudados de acordo com determinados atributos, ou seja, variáveis do processo podem ocorrer ou não, baseados no seu valor.

A partir dessa condição dos atributos, o algoritmo analisa a frequência em que aquele conjunto de atributos aparecem com valores *true* para inferir algum comportamento característico baseado nessa condição.

O número de possíveis associações de regras aumenta em função da quantidade de atributos presentes na base de dados estudada. A Equação 1 a seguir mostra como pode ser feito o cálculo do número de possíveis associações de regras, onde k é a quantidade de atributos.

$$n^{\circ} \text{ de associações de regras} = k * 2^{k-1} \quad (1)$$

Segundo CASTRO (2014), esses algoritmos de associação, em sua maioria, separam a tarefa principal de gerar regras de associação em duas subtarefas, que são divididas pela geração de conjunto de dados frequentes e a geração das regras de associação ao final.

Os parâmetros utilizados para análise dos resultados do algoritmo são os valores de suporte e confiança previamente estabelecidos pelo usuário e esses mesmos valores são calculados para cada regra gerada pelo algoritmo, de acordo com o objetivo almejado. Segue descrição dos parâmetros abaixo:

Suporte – Esse parâmetro é estabelecido de acordo com a quantidade de vezes que o atributo ou conjunto de atributos tinham o valor verdadeiro, *true*, significando uma transação, dividido pelo total de transações do conjunto de dados. Dado pela Equação 2.

$$\text{suporte} = \frac{\text{qtd de transações verdadeiras do atributo}}{\text{qtd de transações total}} \quad (2)$$

Confiança – Esse parâmetro significa a probabilidade condicional de um atributo ocorrer em função da ocorrência de outro atributo associado. Dado pela Equação 3.

$$\text{confiança}(A \rightarrow B) = \frac{P(A \cup B)}{P(A)} = \frac{\text{suporte}(A \cup B)}{\text{suporte}(A)} \quad (3)$$

A tarefa de geração de conjunto de dados ou *itemsets* frequentes realizando uma busca no banco de dados e elege os *itemsets* a partir do valor mínimo de suporte preestabelecido. Enquanto a tarefa de geração de regras de associação extrai as regras de acordo com o valor de confiança preestabelecido a partir dos *itemsets* frequentes encontrados na etapa anterior.

As regras provenientes desse tipo de algoritmo são dadas a partir de uma condição precedente gerando uma consequência, baseado em cálculos probabilísticos a partir do conjunto de dados estudados.

De acordo com CARVALHO (2011), a principal propriedade para algoritmos de mineração de conjuntos de itens frequentes é que o suporte é monotonicamente decrescente com relação ao número de itens de um *itemset*.

Exemplificando, a medida que um novo item é acrescentado, o suporte de um conjunto de itens diminui. Para dois conjuntos de itens A e B em um banco de dados com X transações sobre H, então $X, Y \subset H$, sendo $X \subset Y$, logo, o suporte $(Y) \leq \text{suporte}(X)$.

A partir dessa propriedade é possível concluir que se um *itemset* é pouco frequente, todos os conjuntos derivados deste serão pouco frequentes. Assim como, se um *itemset* é frequente todos os seus conjuntos derivados serão frequentes. Essa é a propriedade da monotonicidade do suporte.

Como os itens do banco de dados não têm a mesma frequência na tabela transacional, os parâmetros de suporte mínimo e confiança precisam ser variados com o objetivo de encontrar regras com atributos tanto de suporte alto quanto de suporte baixo dependendo do tipo de análise requerida pelo usuário.

O grau de interesse nas combinações geradas é representado pela confiança, dessa forma é interessante observar as combinações possíveis com um nível de interesse estabelecido pelo usuário para as regras de associações encontradas. Após isso, o algoritmo mostra ao usuário

quais são as regras selecionadas de acordo com a confiança superior a confiança mínima estabelecida pelo usuário.

Os algoritmos de associação utilizados neste estudo serão descritos a seguir:

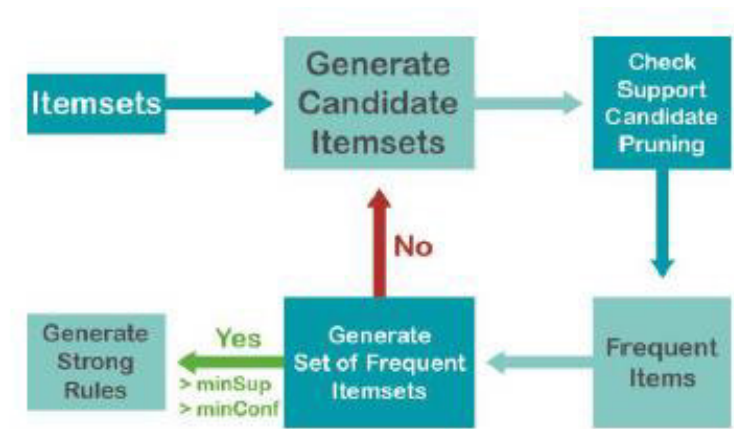
- *Apriori*
- *Fp-Growth*
- *Predictive Apriori*

3.5.1. Apriori

Esse foi um dos primeiros algoritmos desenvolvidos para mineração de dados e, atualmente, é um dos mais conhecidos e utilizado dentre dessa área de estudo. As tarefas principais desse algoritmo são descritas a seguir e apresentados na Figura 14, segundo (POONSIRIVONG et al., 2018). O Apriori apresenta:

- Conceito de itens frequentes – Atributos com número de ocorrências maior que o suporte mínimo;
- Conceito de *itemsets* frequentes – Conjuntos de atributos com número de ocorrências maior que o suporte mínimo;
- Geração de regras de associação – As regras são produzidas de acordo com os limites de suporte e confiança estabelecidos;
- Antecedente (se) e consequente (então) – A regra gerada tem um antecedente que é um item do conjunto de dados e um consequente que é uma co-ocorrência com o antecedente respeitando os limites mínimos de suporte e confiança.

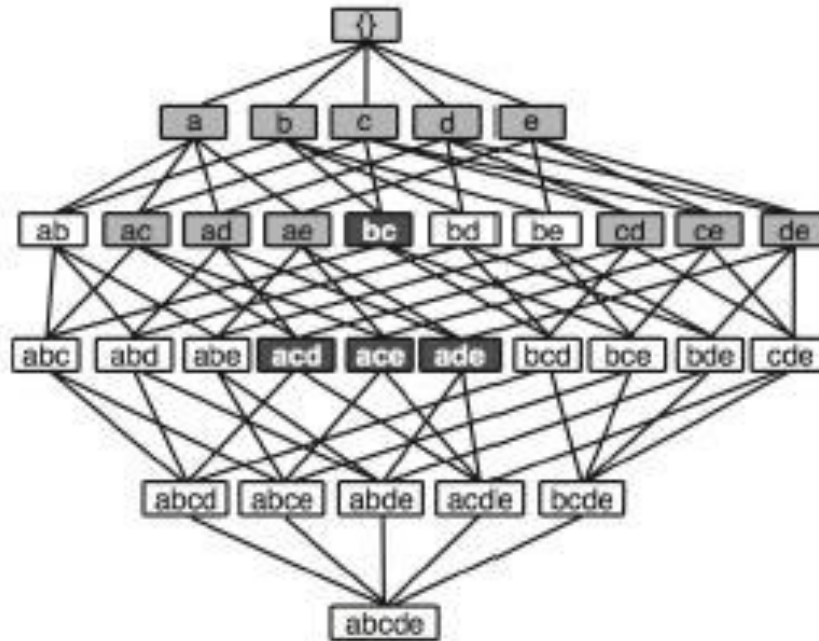
Figura 14 - Tarefas do algoritmo Apriori



Fonte: (POONSIRIVONG et al, 2018)

Os conjuntos de dados frequentes, os *itemsets*, são gerados com todas as possibilidades de associação entre as variáveis presentes na base de dados até se esgotar as associações possíveis. Conforme mostrado na Figura 15.

Figura 15 - Todos *itemsets* frequentes possíveis



Fonte: (CARVALHO, 2011)

O algoritmo Apriori utiliza uma estratégia de busca em largura, *breath-first*, com um algoritmo de geração e teste. Em cada nível são gerados *itemsets* possíveis, tendo em conta os *itemsets* frequentes gerados no nível anterior. Após serem gerados, a frequência desses *itemsets* é testada, percorrendo novamente a base de dados de transações (CARVALHO, 2011, p.183).

O objetivo dessa abordagem é encontrar os conjuntos de itens frequentes, de modo que o suporte relativo de cada conjunto de itens seja maior ou igual ao suporte mínimo parametrizado pelo usuário e encontrar o conjunto de regras de associação com confiança maior que a confiança mínima parametrizada pelo usuário.

O suporte mínimo determinado previamente é o parâmetro utilizado pelo algoritmo para escolher os conjuntos de itens frequentes, de 1 até n elementos por conjunto, que poderão ser associados para descoberta de padrões.

Por exemplo, para um suporte mínimo de 30%, conforme mostrado na Figura 16, é apresentado um banco de dados de 5 itens, com a ocorrência de 10 transações. Todos os conjuntos de itens frequentes são enumerados para 0 itens (é considerado somente para iniciar a construção da árvore de associação dos atributos), 1 item, 2 itens e 3 itens. Notou-se que para

conjunto de 4 itens não há mais combinações possíveis de itens frequentes segundo o suporte mínimo determinado de 30%.

Os conjuntos de itens possíveis para esse exemplo com 5 itens é de $2^5 = 32$ conjuntos de itens. No entanto, como pode ser visto na Figura 16, para o suporte mínimo de 30% há 15 conjuntos de itens frequentes.

Figura 16 - Conjuntos de itens frequentes para suporte mínimo de 30%

TID	Itens
1	{a,d,e}
2	{b,c,d}
3	{a,c,e}
4	{a,c,d,e}
5	{a,e}
6	{a,c,d}
7	{b,c}
8	{a,c,d,e}
9	{b,c,e}
10	{a,d,e}

0 itens	1 item	2 itens	3 itens
∅: 10	{a}: 7	{a,c}: 4	{a,c,d}: 3
	{b}: 3	{a,d}: 5	{a,c,e}: 3
	{c}: 7	{a,e}: 6	{a,d,e}: 4
	{d}: 6	{b,c}: 3	
	{e}: 7	{c,d}: 4	
		{c,e}: 4	
		{d,e}: 4	

Fonte: (CARVALHO, 2011)

Esse processo requer o teste de todas as combinações que possam formar regras, gerando diversas varreduras para recalculando os suportes dos *itemsets* frequentes. Após essa fase, o algoritmo seleciona as regras baseado na confiança calculada de cada regra em relação a confiança mínima estipulada pelo usuário.

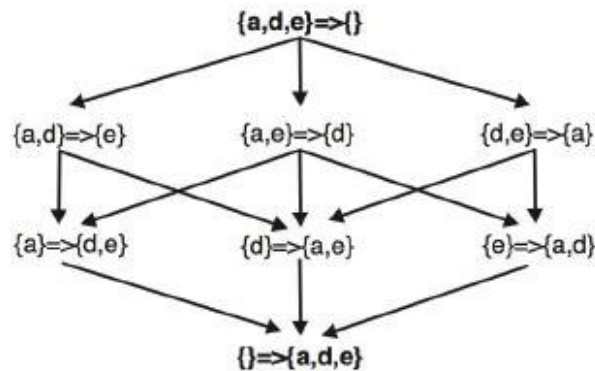
Como pode ser visto na Figura 17, por exemplo, a procura de regras de associação demonstrando o processo a partir do *itemset* frequente {a,d,e} e de acordo com a confiança predeterminada, as regras serão escolhidas a partir de cada combinação possível que for testada.

3.5.2. FP Growth

O algoritmo FP Growth tem como objetivo principal a redução das varreduras do banco de dados, utilizando uma estratégia de busca em profundidade ao invés da busca em largura usada pelo algoritmo apriori descrito na seção 2.6.1. Assim como as métricas para seleção de regras de associação, o suporte e a confiança, também são utilizadas para seleção das regras de associação.

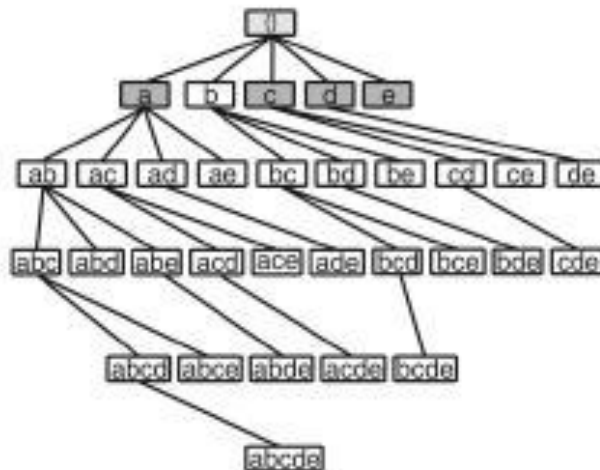
A construção dessa árvore inicial é fundamental para evitar as auto combinações feitas pelo algoritmo apriori, pois é uma árvore de prefixos que armazena os conjuntos de dados. Conforme mostrado na Figura 18.

Figura 17 - Busca para regras de associação para um *itemset*



Fonte: (CARVALHO, 2011)

Figura 18 - Espaço de busca em profundidade



Fonte: (CARVALHO, 2011)

Para construção de regras de associação, o algoritmo FP-growth procede em duas fases. Na primeira fase a estrutura de dados a FP-tree é construída percorrendo a base de dados duas vezes. A FP-tree é então usada para encontrar as regras de associação (CARVALHO, 2011, p.185).

O conjunto de dados de itens frequentes e a contagem do suporte destes é feito durante a primeira varredura do algoritmo no conjunto de dados. Esses conjuntos são ordenados de acordo com cada suporte calculado e armazenado. Em seguida, a árvore de padrões frequentes, *FP-tree*, é construída em função da base de transações.

Para exemplificar a construção da *FP-tree* tem-se como exemplo as transações mostradas na Figura 19. Baseado nessas transações foi verificado os itens mais frequentes, logo, para os dados da Figura 19 tem-se que o item mais frequente é o a, seguido de b, c, d e e.

A construção da *FP-tree* pode ser vista na Figura 19, inicialmente, foi criado o nó *null*, como a raiz da árvore. Para cada transação presente no banco de dados, os itens são dispostos em ordem decrescente de suporte. Dada a primeira transação, {a, b} apresentada na Figura 19, observou-se a criação do primeiro percurso representado na Figura 20(a).

Em seguida, a segunda transação, {b, c, d} apresentada na Figura 19, foi processada gerando um novo conjunto de nós a parte, gerando novo percurso mostrado na Figura 20(b), pois a primeira e segunda transações não compartilharam prefixos comuns. Como b é o único elemento em comum nas duas transações, foi estabelecido uma ligação viabilizando o cálculo da frequência de b.

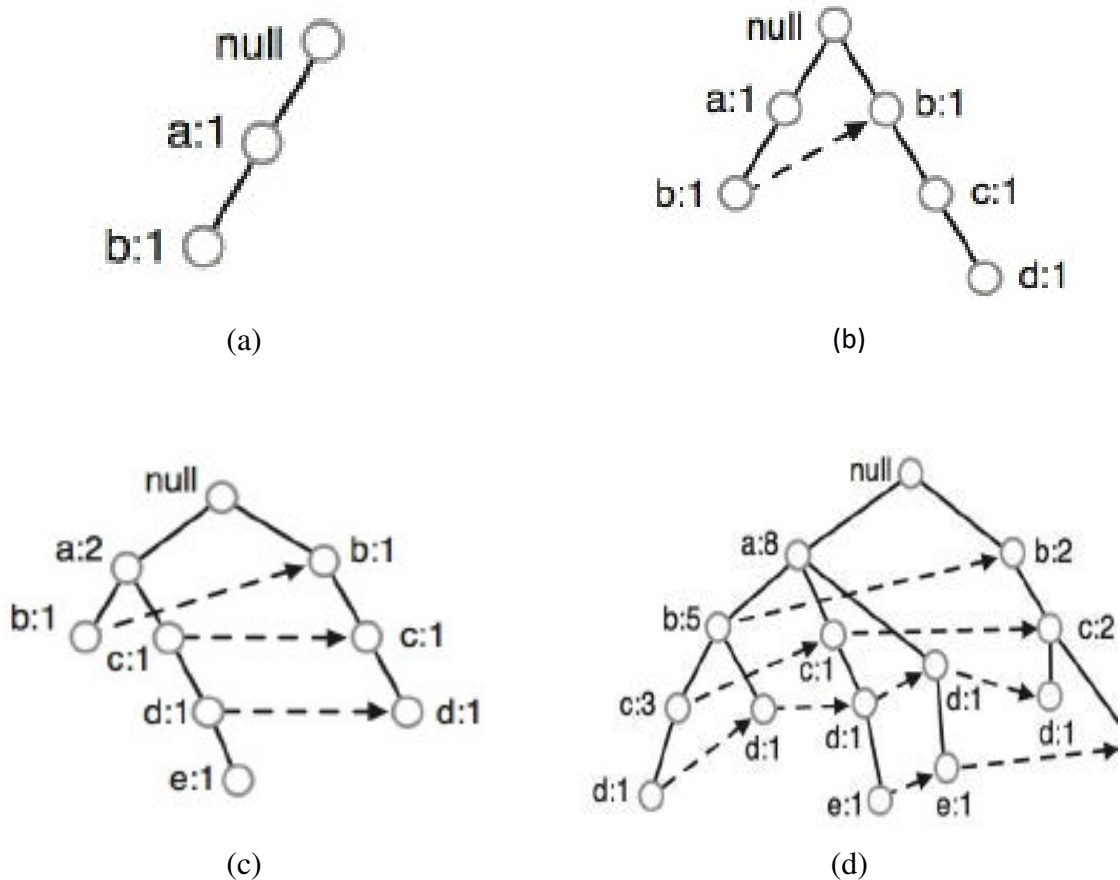
A terceira transação {a, c, d, e} apresentada na Figura 19, tem o item a como prefixo comum entre as transações 1 e 3, então o percurso gerado se sobrepõe ao primeiro percurso para que a frequência de a seja atualizada, o que pode ser visto na Figura 20(c).

Dessa forma o processamento de todas as transações continua até finalizar a construção da *FP-tree* que para esse exemplo é vista na Figura 20(d).

A otimização da *FP-tree* é feita a partir da disposição dos itens mais frequentes próximos à raiz da árvore de acordo com a primeira varredura realizada pelo algoritmo em função da ordem decrescente de suporte da base de transação. Essa característica permite que a árvore seja mantida a menor possível.

Figura 19 - Transações para construção da *FP-tree*

Transações	
TID	Itens
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

Figura 20 - Construção da *FP-tree*

Fonte: (CARVALHO, 2011)

Além do suporte e da confiança utilizam-se como parâmetros para selecionar as regras de associação o coeficiente de interesse ou *lift*, a convicção e o *leverage* para os algoritmos FP Gowth e Apriori.

O *Lift* ou coeficiente de interesse é considerado uma medida para avaliar o nível de independência estatística entre os dois elementos da regra, o antecessor e o consequente. Dado pela Equação 3.

$$lift(A \rightarrow B) = \frac{confiança(A \rightarrow B)}{suporte(B)} = \frac{suporte(A \cup B)}{suporte(A) \times suporte(B)} \quad (3)$$

O valor de *lift* igual a 1 indica que A e B são independentes. Valores de *lift* inferiores a 1 indicam que A e B são negativamente correlacionados, enquanto valores superiores a 1 indicam uma correlação positiva (CARVALHO, 2016, p.190).

A convicção tem o objetivo de medir o nível de convencimento da regra, evidenciando a frequência de erro da regra, supondo que os dois elementos da regra fossem independentes. Dada pela Equação 4.

$$conv(A \rightarrow B) = \frac{1 - suporte(B)}{1 - confiança(A \rightarrow B)} \quad (4)$$

A convicção pode ser interpretada como o quociente da frequência esperada de A ocorrer sem B (ou seja, a frequência de erro da regra) como se A e B fossem independentes dividido pela frequência de previsões incorretas (CARVALHO, 2016, p.190).

Para valores dessa medida inferiores a 1 a associação entre as variáveis A e B é aleatória e para valores superiores a 1 a associação entre as variáveis A e B é positiva.

E o *leverage* tem o objetivo de medir o valor da diferença entre o suporte real e o esperado de uma regra de associação. Dada pela Equação 5.

$$Leverage(A \rightarrow B) = suporte(A \rightarrow B) - (suporte(A) \times suporte(B)) \quad (5)$$

3.5.3. Predictive Apriori

De acordo com PATIL et al. (2011), a principal diferença entre o algoritmo predictive Apriori e o Apriori são as medidas de interesse das regras de associação geradas. O Apriori gera mais regras em função da confiança e as ordena conforme a confiança calculada para cada regra, sendo que as regras geradas possuem alto suporte e baixa confiança e regras específicas com baixo suporte e alta confiança.

Já o predictive Apriori avalia a confiança das regras baseadas em seus respectivos suportes. Esse algoritmo prefere as regras mais gerais, em função do objetivo que é a aquisição de conhecimento sob nova perspectiva através dos dados não vistos, e assim, melhorando a precisão.

E, ainda assim, ambos algoritmos possuem passos em comum, por exemplo, a geração de *itemsets* frequentes funciona da mesma forma. Para o algoritmo Apriori a seleção dos atributos para geração das regras é baseada no valor de suporte mínimo definido e se a confiança é suficiente. Para o predictive Apriori o objetivo é maximizar a precisão esperada de uma regra ao invés da precisão da confiança dos dados estudados.

Essa probabilidade de gerar uma precisão correta da regra de associação é chamada de predictive accuracy.

Os parâmetros de suporte e confiança são utilizados como critérios de avaliação dos dados, da mesma forma que o algoritmo Apriori. A diferença está em uma nova medida chamada de predictive accuracy, que faz uma combinação entre o suporte e a confiança definindo as n melhores regras de associação em ordem de importância de acordo com a predictive accuracy.

A medida predictive accuracy ajusta os valores de suporte e confiança, relacionando esses parâmetros através de uma relação binomial, com aumento gradativo do suporte mínimo ordenando as regras de associação. Essa medida representa o grau de previsibilidade mais precisamente para a regra gerada.

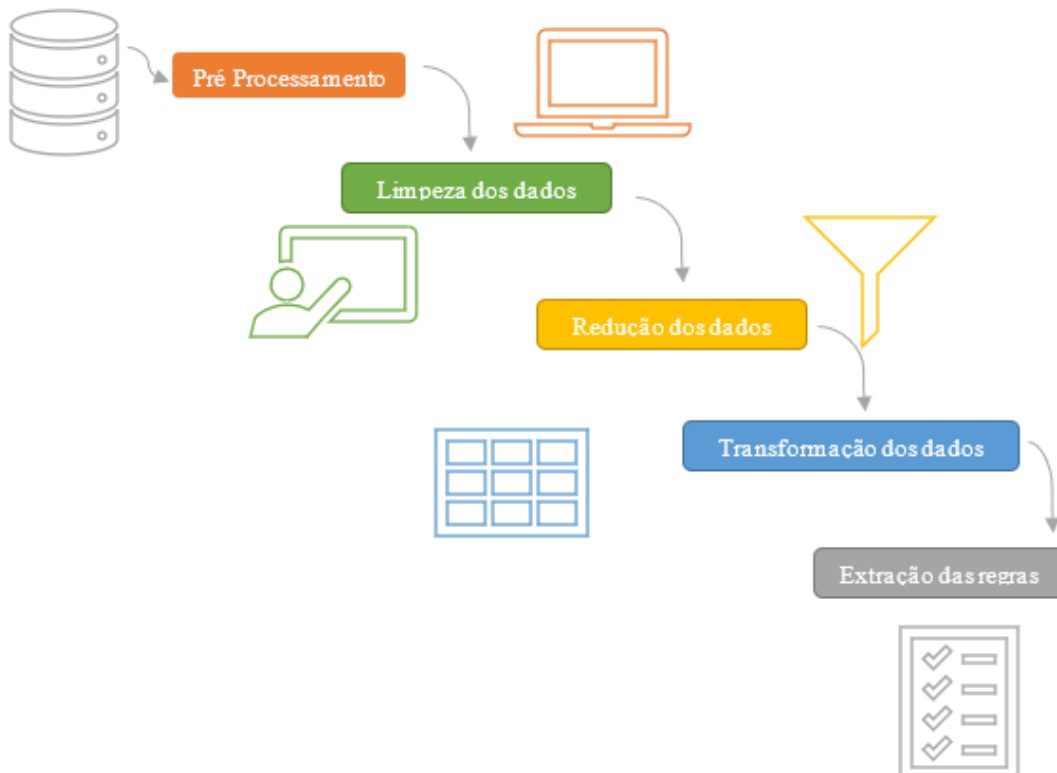
4. METODOLOGIA PROPOSTA PARA APLICAÇÃO DE ALGORITMOS DE ASSOCIAÇÃO

O processo KDD - *Knowledge Discovery in Databases*, que segundo FAYYAD et al., é realizado para extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir de um banco de dados. As etapas do processo devem ser seguidas para construir o fluxo até a descoberta da informação.

Essas etapas foram realizadas para o desenvolvimento deste trabalho com intuito de encontrar padrões de falhas escondidos em grande volume de dados. Fundamental a verificação das etapas de pré-processamento referentes a limpeza, redução e transformação dos dados. Conforme é mostrado na Figura 21.

Os dados estudados são característicos da Estrada de Ferro Carajás – EFC, pois tem relação direta com o perfil da via e normas de operação de trens praticadas nessa ferrovia.

Figura 21- Metodologia para descoberta de regras para manutenção preventiva de vagões de minério



Fonte: Autor.

4.1. Base de dados

A base de dados utilizada neste trabalho é composta pelas medições obtidas dos equipamentos HotBox e Railbam instalados na ferrovia Estrada de Ferro Carajás no período de

2017 até meados de 2018. Durante este período ocorreram 12 alarmes de temperatura de rolamento no ano de 2017 e 5 alarmes até meados de 2018.

O equipamento Hotbox realiza as medições de temperatura de cada rolamento do trem, tanto das locomotivas quanto dos vagões da composição, através de radiação infravermelha e após a leitura a esta radiação é codificada em temperatura.

O equipamento Railbam realiza as medições dos ruídos provenientes dos rolamentos gerados durante a passagem dos trens. Esse equipamento classifica as falhas de acordo com a amplitude e frequência dos ruídos captados.

A base de dados é composta pelas variáveis referente aos números de identificação do vagão, data e hora da medição, tipos de falhas, tipos ruídos, valores de temperaturas, velocidade do eixo do vagão durante a passagem pelo equipamento de leitura, eixo do vagão e direção do trem.

Além disso, quando ocorrem evento de Hot Box, geralmente, os vagões que apresentaram a temperatura acima do limite são retidos, ou seja, retirados do trem e não continuam a viagem em função da criticidade do evento e possibilidade de ocorrência de acidentes.

4.2. Pré-processamento dos dados

Os resultados do trabalho dependem da qualidade dos dados estudados. A análise de dados requer uma boa avaliação dos dados para a retirada de problemas que possam interferir nos resultados do trabalho proposto. Tais problemas podem ser ruídos, valores não significativos para o estudo, valores desconhecidos, dentre outros.

Por isso, essa etapa de realizar um pré-processamento dos dados é tão importante, pois tem o objetivo de aperfeiçoar a qualidade dos dados.

4.2.1. Limpeza dos dados

A limpeza dos dados é o primeiro passo para a retirada de dados inconsistentes, incompletos e ruidosos. Tais problemas são encontrados devido a defeitos na leitura, armazenamento ou transmissão dos dados. É necessário compreender o processo envolvido para identificar os erros e retirá-los.

Para realizar essa identificação de erros nos dados é preciso verificar as divergências, duplicação de valores, falta de valores, dentre outros tipos de erros. Como nesse estudo foi necessário correlacionar as leituras provenientes de dois equipamentos, Hotbox e Railbam, foram observados casos de ausência de valores, muitas vezes devido ao fato de um dos equipamentos ter realizado a leitura e o outro não naquele mesmo período cronológico.

O fato da ausência de medição por um dos equipamentos impossibilita a correlação dos dados, logo, esses casos foram excluídos da base de dados. Além disso, aconteceram casos de eliminação quando encontrados valores duplicados provenientes de leituras errôneas.

4.2.2. Redução dos dados

A análise de grandes volumes de dados muitas vezes não é tão eficiente se o analista não escolher amostras de boa qualidade para o estudo. Além da dificuldade existente em processar grandes volumes de dados, nem todos os métodos de análise conseguem expurgar os dados indesejados sem a intervenção manual do analista.

Em relação a este trabalho tem-se cerca de 20 trens circulando diariamente. A cada leitura do equipamento dos 4 eixos dos 330 vagões presentes em um único trem geram cerca de 1.320 linhas referentes a leitura de temperatura dos vagões por trem, enquanto a leitura dos ruídos acústicos dos vagões gera mais 1.320 linhas por trem no banco de dados da empresa.

Resultando em um valor diário de 2.640 linhas geradas pelos dois equipamentos por trem, logo, com 20 trens na ferrovia tem-se um resultado de 52.800 linhas de dados. Por esse motivo, para esse estudo foram escolhidas amostras de dados específicas de vagões alarmados ou pré-alarmados e tipos de acústicas para reduzir a base de dados.

4.2.3. Representação dos dados

Os algoritmos de regras de associação tratam dados com modelo de tabela transacional, ou seja, que possuem dois valores possíveis, neste caso esses valores são *true* ou *false*. Os valores de medições dos equipamentos foram relacionados entre si, associando os valores de leitura dos ruídos acústicos com as temperaturas observadas durante a viagem dos trens.

As medições dos ruídos acústicos são realizadas duas vezes por viagem, uma vez quando o trem vazio sai de São Luís em direção à mina e uma outra vez quando o trem carregado retorna para São Luís. Enquanto a medição da temperatura é realizada durante o deslocamento do trem por toda a extensão da malha em 14 pontos distintos.

Dado este fato, para cada leitura de temperatura foi necessário repetir os dados acústicos lidos de acordo com data, horário e sentido da viagem do trem para todas as leituras de temperaturas medidas ao longo da viagem. Dessa forma, as informações foram agrupadas em ordem cronológica. Esses dados foram inseridos no software Power BI para adaptar a base de dados relacional em uma base de dados transacional, formato este que os algoritmos de associação aceitam por meio do *software* Weka.

Os atributos são referentes a ocorrência dos dados medidos pelos equipamentos Railbam e Hot Box em função da presença de ruído, eixo correspondente a falha, tipos de falhas acústicas, velocidade e temperatura de acordo com a base relacional originária.

Para adaptar os dados ao modelo de tabela transacional, as medições de temperaturas, que são grandezas numéricas, foram divididas em intervalos com valores de temperaturas menores que 20°C, maior que 20°C e menor ou igual a 30°C, maior que 30°C e menor ou igual a 40°C, maior que 40°C e menor ou igual a 50°C, maior que 60°C e menor ou igual a 60°C e maior que 60°C e menor ou igual a 70°C. Utilizando o mesmo princípio, os valores numéricos referentes a velocidade de passagem do eixo durante a leitura do equipamento foram divididos em intervalos com valores menor que 30km/h, entre 30 a 50km/h e 50 a 80km/h.

Os tipos de falhas foram utilizados conforme apresentado na Tabela 1 na seção 2.3 do Capítulo 2, sendo esses *Clear Level 1*, *Clear Level 2*, *Clear Level 3 Potential 1*, *Potential 2*, *Potential 3* e *Wheel Flats*. No entanto, para o tipo de falhas *Wheel Flats* não foi encontrado correlação com nenhum dos casos estudados de Hot Box, ou seja, para os casos estudados de Hot Box não ocorreu o alarme *Wheel Flats*.

Os demais valores categóricos distintos referentes ao eixo do vagão relacionado a falha, presença de ruído e lado do equipamento de leitura foram transpostos para aderir ao modelo da tabela transacional. Com isso a tabela de modelo transacional foi concluída, sendo composta por 16 atributos e 9032 instâncias com valores possíveis *true* ou *false*.

Os valores dessa tabela transacional são nominais, admitindo valores *true* ou *false*, sendo o valor de um atributo correspondente ao par (atributo, valor).

O valor correspondente na tabela transacional será *true* se o atributo tiver valor na base de dados relacional e será *false* se o atributo não tiver valor na base de dados relacional. Conforme é mostrado na Figura 22.

Como o objetivo desse estudo é encontrar regras com valores '*true*', significando a ocorrência das falhas de acordo com os atributos, os valores '*false*' foram excluídos porque as regras com esse tipo de valor não contribuem para o trabalho, visto que o *false* indica a não ocorrência do atributo. Conforme é mostrado na Figura 23.

4.3. Regras de Associação

Após finalizar o processo de pré-processamento, a etapa seguinte é a utilização dos algoritmos de associação. Esse tipo de algoritmo requer a participação do usuário para definir quais dados serão analisados e verificar se as regras resultantes da aplicação dos algoritmos de associação escolhidos são úteis ou inúteis avaliando qualitativamente as medidas de interesse.

Figura 22 - Tabela transacional dos dados correlacionados

No.	1: Temp	20: 2	30: 3040	4: 4050	5: 5060	6: 6070	7: RuÁ-do	8: 1R	9: 2R	10: 3R	11: 4R	12: 1L	13: 2L	14: 3L	15: 4L	16: Clear Level 1	17: Clear Level 2	18: Potential 1	19: Potential 2	20: V=30	21: 30 V=50	22: 50 V=80	23: LEIT
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Non
1	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true
2	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
3	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
4	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
5	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true
6	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true
7	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	true
8	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	true
9	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true
10	true	false	false	false	false	false	false	true	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
11	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
12	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
13	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true
14	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
15	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
16	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
17	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true
18	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
19	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
20	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
21	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true
22	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
23	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
24	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
25	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
26	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
27	false	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
28	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
29	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true
30	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
31	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
32	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
33	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
34	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	true	false	true
35	true	true	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
36	true	false	false	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	true	false	true
37	true	false	false	false	false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	true	false	true

Fonte: Autor.

Figura 23 - Tabela transacional somente com valores true

No.	1: Temp	20: 2	30: 3040	4: 4050	5: 5060	6: 6070	7: Ruído	8: 1R	9: 2R	10: 3R	11: 4R	12: 1L	13: 2L	14: 3L	15: 4L	16: Clear Level 1	17: Clear Level 2	18: Potential 1	19: Potential 2	20: V=30	21: 30 V=50	22: 50 V=80	23: LEIT
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	true							true													true		true
2	true								true												true		true
3	true									true											true		true
4	true										true										true		true
5	true											true									true		true
6	true												true								true		true
7	true													true							true		true
8	true														true						true		true
9	true															true					true		true
10	true																true				true		true
11	true																	true			true		true
12	true																		true		true		true
13	true																			true	true		true
14	true																				true		true
15	true																				true		true
16	true																				true		true
17	true																				true		true
18	true																				true		true
19	true																				true		true
20	true																				true		true
21	true																				true		true
22	true																				true		true
23	true																				true		true
24	true																				true		true
25	true																				true		true
26	true																				true		true
27		true																			true		true
28	true																				true		true
29	true																				true		true
30	true																				true		true
31	true																				true		true
32	true																				true		true
33	true																				true		true
34	true																				true		true
35		true																			true		true
36	true																				true		true
37	true																				true		true

Fonte: Autor.

O objetivo da utilização desses algoritmos é encontrar informações válidas para auxiliar na tomada de decisão, ou seja, informações não óbvias e inerentes ao processo. Os parâmetros utilizados para análise dos resultados dos algoritmos são os valores de suporte e confiança previamente estabelecidos pelo usuário, de acordo com o objetivo especificado para o algoritmo.

Baseado nesses conceitos, o *software* permite ao usuário inserir um valor de suporte mínimo e confiança de acordo com o objetivo de análise do mesmo. Os atributos foram

utilizados para gerar padrões, somente se tiverem um valor acima do suporte mínimo, o que significa que esses atributos estão presentes em uma quantidade de transações maior que o valor de suporte mínimo determinado.

Baseado nisso, o suporte mínimo foi variado à medida que a temperatura aumentava, pois quanto maior a temperatura menor a quantidade de transações associadas aos tipos de falhas acústicas. Esse comportamento é considerado normal, em função da baixa quantidade de dados de vagões alarmados, em relação a uma frota com cerca de 19.000 vagões.

Foi visto na Seção 3.5 o número de possíveis associações de regras. Para este estudo com 16 atributos presentes no conjunto de dados teremos o número de associações de regras de acordo com a Equação 6.

$$n^{\circ} \text{ de associações de regras} = 16 * 2^{16-1} = 16 * 32.768 = 524.288 \quad (6)$$

Portanto, existem aproximadamente 524.288 possíveis associações de regras.

O algoritmo tem como produto uma relação de regras geradas em função de todas essas associações possíveis, ao final do processamento da base de dados por cada algoritmo. No entanto, não podemos afirmar que todas as regras descobertas serão úteis para o usuário final.

Por exemplo, é possível que o objetivo seja encontrar regras com itens raros, logo os valores de suporte e confiança devem ser adequados a esta finalidade.

Nesse estudo, como o objetivo era encontrar possíveis associações entre os dados de temperatura e tipos de falhas acústicas, as medidas de interesse das regras resultantes foram analisadas, dado que essas medidas são calculadas pelos algoritmos para cada regra gerada.

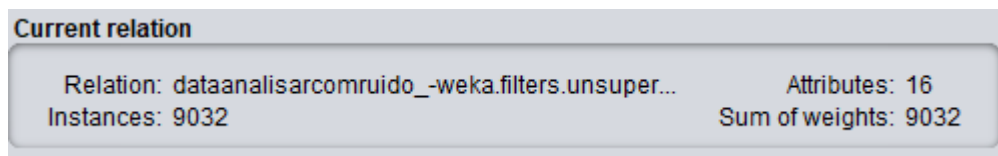
As regras foram obtidas a partir desses três algoritmos mencionados na Seção 3.5, sendo esses Apriori, predictive Apriori e FP-Growth com parametrização de um suporte mínimo de valor 0.0 para descobrir as regras devido ao fato de alguns atributos referentes aos tipos de falhas acústicas e medições de temperatura representarem uma porcentagem pequena de dados em relação a base de dados completa.

Além disso, a avaliação do usuário final tem grande importância. Visto que esse tem o conhecimento específico sobre o assunto estudado, porém esse tipo de análise não está no escopo desse trabalho.

5. RESULTADOS

Este capítulo apresenta os testes realizados com a aplicação dos algoritmos de associação escolhidos que foram Apriori, Predictive Apriori e FP Growth. Inicialmente, a base de dados construída possui 16 atributos e 9032 instâncias relacionadas, conforme mostrado na Figura 24.

Figura 24 - Atributos e Instâncias do Weka



Fonte: Autor.

Os atributos como o número do vagão e a data e hora da medição foram removidos porque não apresentaram os valores requeridos pelo modelo de tabela transicional, sendo esses utilizados somente para critérios de organização dos dados na construção da tabela transacional. E os dois atributos referentes ao lado do equipamento na via férrea também foram retirados.

Esse fato evidencia a ausência de medições feitas pelo equipamento Railbam, presente somente em local específico da ferrovia, logo após esse tipo de ocorrência. O que implicou em baixa quantidade de dados para correlacionar as medições de temperaturas elevadas com os tipos de falhas acústicas para os casos críticos.

Os algoritmos Apriori, predictive Apriori e FP-Growth com os 16 atributos e as 9032 instâncias foram utilizados para geração de regras de associação. Os experimentos foram divididos de acordo com as regras de temperaturas geradas, devido a diferença de criticidade dada para cada temperatura.

O suporte mínimo foi definido em 0.0 decorrente da presença de atributos de tipos de falhas e temperatura com suporte muito baixo, ou seja, quantidade de registros frequentes sobre o total de registros do conjunto de *itemsets*.

5.1. Experimento para temperaturas menores que 20°C

Para temperaturas menores que 20°C foram apresentadas nas Tabelas 2 e 3 a seguir as regras produzidas pelos algoritmos.

Essa temperatura é considerada a temperatura normal para os rolamentos, sendo essa temperatura a mais frequente no conjunto de dados analisado, o que já seria esperado devido essa condição ser normal para a performance do ativo. Os parâmetros utilizados para os algoritmos foram de 0.9 para a confiança mínima, 0.0 para o suporte mínimo, 1.0 para o limite superior de suporte mínimo e quantidade de regras distinta para cada algoritmo.

A apresentação das regras segue abaixo conforme o algoritmo utilizado, o n° da regra gerada, a descrição da regra encontrada e as medidas de interesse calculadas pelos algoritmos.

Tabela 2 - Regras geradas pelo Algoritmo Apriori para temperaturas menores que 20°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Apriori	1	Ruido=true Clear Level 1=true 30 < V <= 50=true 10 ==> TEMP<20=true 10	conf:(1) lift:(1.18)
Apriori	2	Clear Level 3=true 30 < V <= 50=true 8 ==> TEMP<20=true 8	conf:(1) lift:(33.83)
Apriori	12	Clear Level 3=true V<=30=true 1 ==> TEMP<20=true 1	conf:(1) lift:(1.18)
Apriori	19	Ruido=true Clear Level 1=true 21 ==> TEMP<20=true 20	conf:(1) lift:(1.18)
Apriori	20	Potential 3=true V<=30=true 154 ==> TEMP<20=true 141	conf:(0.92) lift:(1.08)
Apriori	22	V<=30=true 725 ==> TEMP<20=true 656	conf:(0.9) lift:(1.07)

Tabela 3 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas menores que 20°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Predictive Apriori	1	Ruido=true Clear Level 1=true 30 < V <= 50=true 10 ==> TEMP<20=true 10	acc:(0.91635)
Predictive Apriori	2	Ruido=true Clear Level 1=true 21 ==> TEMP<20=true 20	acc:(0.913)
Predictive Apriori	3	Potential 3=true V<=30=true 154 ==> TEMP<20=true 141	acc:(0.91022)
Predictive Apriori	4	Clear Level 3=true 30 < V <= 50=true 8 ==> TEMP<20=true 8	acc:(0.89965)
Predictive Apriori	5	Ruido=true Potential 2=true 30 < V <= 50=true 56 ==> TEMP<20=true 51	acc:(0.8964)
Predictive Apriori	6	Potential 2=true V<=30=true 50 ==> TEMP<20=true 45	acc:(0.88441)
Predictive Apriori	7	Ruido=true Potential 2=true 88 ==> TEMP<20=true 78	acc:(0.87763)

Para o algoritmo FP-Growth os mesmos parâmetros foram utilizados inicialmente para geração de regras, porém não foram encontradas regras correlacionando a condição da temperatura com os demais fatores. Então, a confiança foi variada para obtenção de novos resultados, contudo, o algoritmo não encontrou regras correlacionando a temperatura menor que 20°C e os demais atributos presentes na base de dados utilizando esse algoritmo.

5.2. Experimento para temperaturas entre 20 e 30°C

Para temperaturas entre 20°C e 30°C foram apresentadas nas Tabelas 4, 5 e 6 a seguir as regras produzidas pelos algoritmos.

Tabela 4 - Regras geradas pelo Algoritmo Apriori para temperaturas entre 20°C e 30°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Apriori	3	20<TEMP<30=true Clear Level 3=true 3 ==> 50 < V <= 80=true 3	conf:(1) lift:(2.2)
Apriori	7	Ruido=true Potential 3=true V<=30=true 2 ==> 20<TEMP<30=true 2	conf:(1) lift:(7.78)
Apriori	17	20<TEMP<30=true Ruido=true Clear Level 1=true 1 ==> 50 < V <= 80=true 1	conf:(1) lift:(2.2)
Apriori	18	20<TEMP<30=true Potential 1=true V<=30=true 1 ==> Ruido=true 1	conf:(1) lift:(33.83)

Tabela 5 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas entre 20°C e 30°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Predictive Apriori	14	20<TEMP<30=true Clear Level 3=true 3 ==> 50 < V <= 80=true 3	acc:(0.8)
Predictive Apriori	18	20<TEMP<30=true Potential 1=true 50 < V <= 80=true 6 ==> Ruido=true 5	acc:(0.75005)
Predictive Apriori	21	Ruido=true Potential 3=true V<=30=true 2 ==> 20<TEMP<30=true 2	acc:(0.75)
Predictive Apriori	41	20<TEMP<30=true Potential 3=true 248 ==> 50 < V <= 80=true 145	acc:(0.584)
Predictive Apriori	42	20<TEMP<30=true Potential 1=true 17 ==> 30 < V <= 50=true 10	acc:(0.57895)
Predictive Apriori	59	20<TEMP<30=true Potential 2=true 60 ==> 50 < V <= 80=true 29	acc:(0.48387)

Para o algoritmo FP-Growth o parâmetro da confiança foi variado, como é possível ver abaixo na Tabela 6 regras com confiança ente 1 até 0.5.

Tabela 6 - Regras geradas pelo Algoritmo FP-Growth para temperaturas entre 20°C e 30°C

Algoritmo	N°	Regra (se ==> então)	Medidas
FP-Growth	66	[Potential 3=true, V<=30=true, Ruido=true]: 2 ==> [20<TEMP<30=true]: 2	conf:(1) lift:(7.78)
FP-Growth	72	[20<TEMP<30=true, Potential 1=true]: 17 ==> [Ruido=true]: 11	conf:(0.65) lift:(21.89)
FP-Growth	74	[20<TEMP<30=true]: 1161 ==> [50 < V <= 80=true]: 686	conf:(0.59) lift:(1.3)
FP-Growth	75	[Potential 3=true, 20<TEMP<30=true]: 248 ==> [50 < V <= 80=true]: 145	conf:(0.58) lift:(1.29)
FP-Growth	80	[20<TEMP<30=true, V<=30=true, Potential 2=true]: 4 ==> [Ruido=true]: 2	conf:(0.5) lift:(16.91)

5.3. Experimento para temperaturas entre 30 e 40°C

Para temperaturas entre 30°C e 40°C foram apresentadas nas Tabelas 7, 8 e 9 a seguir as regras produzidas pelos algoritmos.

A partir desse experimento também foi necessário variar o parâmetro confiança em todos os algoritmos implementados para descoberta de novas regras.

Tabela 7 - Regras geradas pelo Algoritmo Apriori para temperaturas entre 30°C e 40°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Apriori	43	30<TEMP<40=true Ruido=true Potential 3=true 5 ==> 30 < V <= 50=true 4	conf:(0.8) lift:(1.72)
Apriori	66	30<TEMP<40=true Clear Level 1=true 3 ==> 50 < V <= 80=true 2	conf:(0.67) lift:(1.47)
Apriori	70	30<TEMP<40=true Ruido=true 50 < V <= 80=true 3 ==> Potential 1=true 2	conf:(0.67) lift:(44.94)
Apriori	97	30<TEMP<40=true Potential 3=true 37 ==> 30 < V <= 50=true 19	conf:(0.51) lift:(1.1)

Tabela 8 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas entre 30°C e 40°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Predictive Apriori	23	30<TEMP<40=true Ruido=true Potential 3=true 5 ==> 30 < V <= 50=true 4	acc:(0.71432)
Predictive Apriori	35	30<TEMP<40=true Ruido=true 30 < V <= 50=true 6 ==> Potential 3=true 4	acc:(0.625)
Predictive Apriori	39	30<TEMP<40=true Ruido=true 50 < V <= 80=true 3 ==> Potential 1=true 2	acc:(0.60002)
Predictive Apriori	49	30<TEMP<40=true Ruido=true 9 ==> Potential 3=true 5	acc:(0.54545)
Predictive Apriori	54	30<TEMP<40=true Potential 3=true 37 ==> 30 < V <= 50=true 19	acc:(0.51282)
Predictive Apriori	57	30<TEMP<40=true Potential 2=true 6 ==> 30 < V <= 50=true 3	acc:(0.5)

Tabela 9 - Regras geradas pelo Algoritmo FP-Growth para temperaturas entre 30°C e 40°C

Algoritmo	N°	Regra (se ==> então)	Medidas
FP-Growth	79	[30<TEMP<40=true, Potential 1=true]: 12 ==> [50 < V <= 80=true]: 6	conf:(0.5) lift:(1.1)
FP-Growth	71	[50 < V <= 80=true, Ruido=true, 30<TEMP<40=true]: 3 ==> [Potential 1=true]: 2	conf:(0.67) lift:(44.94)

5.4. Experimento para temperaturas entre 40 e 50°C

Para temperaturas entre 40°C e 50°C foram apresentadas nas Tabelas 10, 11 e 12 a seguir as regras produzidas pelos algoritmos.

Tabela 10 - Regras geradas pelo Algoritmo Apriori para temperaturas entre 40°C e 50°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Apriori	5	40<TEMP<50=true Ruido=true 2 ==> Potential 1=true 2	conf:(1) lift:(67.4)
Apriori	8	40<TEMP<50=true Potential 1=true V<=30=true 2 ==> Ruido=true 2	conf:(1) lift:(33.83)
Apriori	9	40<TEMP<50=true Ruido=true V<=30=true 2 ==> Potential 1=true 2	conf:(1) lift:(67.4)
Apriori	78	40<TEMP<50=true Potential 2=true 5 ==> 30 < V <= 50=true 3	conf:(0.6) lift:(1.29)

Tabela 11 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas entre 40°C e 50°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Predictive Apriori	19	40<TEMP<50=true Ruido=true 2 ==> Potential 1=true V<=30=true 2	acc:(0.75)
Predictive Apriori	44	40<TEMP<50=true Potential 3=true 12 ==> 50 < V <= 80=true 7	acc:(0.57143)
Predictive Apriori	45	40<TEMP<50=true Potential 2=true 5 ==> 30 < V <= 50=true 3	acc:(0.57142)

Tabela 12 - Regras geradas pelo Algoritmo FP-Growth para temperaturas entre 40°C e 50°C

Algoritmo	N°	Regra (se ==> então)	Medidas
FP-Growth	76	[Potential 3=true, 40<TEMP<50=true]: 12 ==> [50 < V <= <= 80=true]: 7	conf:(0.58) lift:(1.28)

5.5. Experimento para temperaturas entre 50 e 60°C

Para temperaturas entre 50°C e 60°C foram apresentadas nas Tabelas 13, 14 e 15 a seguir as regras produzidas pelos algoritmos.

Tabela 13 - Regras geradas pelo Algoritmo Apriori para temperaturas entre 50°C e 60°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Apriori	4	50<TEMP<60=true Potential 3=true 3 ==> 50 < V <= 80=true 3	conf:(1) lift:(67.4)
Apriori	13	50<TEMP<60=true V<=30=true 1 ==> Clear Level 1=true 1	conf:(1) lift:(88.55)
Apriori	69	50<TEMP<60=true Potential 1=true 3 ==> 30 < V <= 50=true 2	conf:(0.67) lift:(1.43)

Tabela 14 - Regras geradas pelo Algoritmo predictive Apriori para temperaturas entre 50°C e 60°C

Algoritmo	N°	Regra (se ==> então)	Medidas
Predictive Apriori	15	50<TEMP<60=true Potential 3=true 3 ==> 50 < V <= 80=true 3	acc:(0.8)
Predictive Apriori	38	50<TEMP<60=true Potential 1=true 3 ==> 30 < V <= 50=true 2	acc:(0.60002)
Predictive Apriori	73	50<TEMP<60=true 50 < V <= 80=true 7 ==> Potential 3=true 3	acc:(0.44444)

Tabela 15 - Regras geradas pelo Algoritmo FP-Growth para temperaturas entre 50°C e 60°C

Algoritmo	Nº	Regra (se ==> então)	Medidas
FP-Growth	64	[Potential 3=true, 50<TEMP<60=true]: 3 ==> [50 < V <= 80=true]: 3	conf:(1) lift:(2.2)
FP-Growth	65	[Clear Level 1=true, 50<TEMP<60=true]: 1 ==> [V<=30=true]: 1	conf:(1) lift:(12.46)

5.6. Experimento para temperaturas entre 60 e 70°C

Para temperaturas entre 60°C e 70°C foram apresentadas na Tabela 16 a seguir as regras produzidas pelos algoritmos.

Tabela 16 - Regras geradas pelos Algoritmos Apriori e FP-Growth para temperaturas entre 60°C e 70°C

Algoritmo	Nº	Regra (se ==> então)	Medidas
Apriori	15	60<TEMP<70=true Potential 1=true 1 ==> 50 < V <= 80=true 1	conf:(1) lift:(2.2)
FP-Growth	102	[50 < V <= 80=true, 60<TEMP<70=true]: 3 ==> [Potential 1=true]: 1	conf:(0.33) lift:(22.47)

5.7. Análise de resultados dos experimentos

Nesta Seção, uma comparação é feita com os resultados apresentados nas Seções 4.1 até 4.6.

Para os eventos de Hot Box a temperatura maior ou igual a 70°C é considerada crítica, quando esse valor de temperatura ocorreu em um dos rolamentos do vagão, o trem é parado para inspeção, medição da temperatura do rolamento e avaliação da condição física do vagão alarmado. Os vagões com temperatura acima de 50°C e menor que 70°C são considerados pré-alarmados, ou seja, condição anterior a falha.

Dado essas considerações, como não foi possível correlacionar os dados referentes as temperaturas acima de 70°C devido à ausência de informação relacionada a decisão de retirada de circulação dos vagões nessas condições, os dados foram analisados para temperaturas até 70°C que foram possíveis correlacionar com os demais dados do Railbam.

Por isso, os experimentos foram realizados de acordo com intervalos de temperaturas, como condição para essas avaliações e suas possíveis associações. E dessa forma, observar o comportamento das falhas acústicas para todos os intervalos de temperatura.

É importante salientar que a frequência das temperaturas presentes no banco de dados é maior para temperaturas menores, ou seja, quanto maior a temperatura, menor será sua

frequência nesse conjunto de dados. Por isso, quanto maior a temperatura menor será o seu suporte.

Nesse caso, as regras têm maior precisão devido ao suporte baixo, enquanto um suporte elevado é a causa de uma maior quantidade de regras geradas, porém, de baixa precisão. Portanto, quanto maior a temperatura, menor será o suporte daquele conjunto de dados referente ao intervalo de temperaturas e, conseqüentemente, a regra gerada tem maior precisão.

Outro fator importante ajustado durante os experimentos foi a variação do parâmetro de confiança para que o algoritmo encontrasse novas regras associando atributos de diferentes suportes.

Para temperaturas menores que 20°C, experimentos apresentados na Seção 5.1, as regras encontradas pelo algoritmo Apriori tiveram confiança 100%, 92% e 90%, enquanto para o algoritmo Predictive Apriori a acurácia oscilou entre 91% até 87%.

Ainda para esse mesmo intervalo, os algoritmos geraram algumas regras similares, como é o caso das regras de número 1 de ambos os algoritmos, em seguida a regra de número 2 do Apriori é a mesma que a de número 4 do Predictive Apriori, a regra de número 19 do Apriori é a mesma que a de número 2 do Predictive Apriori, dentre outras, essas regras foram apresentadas na Tabela 2 e na Tabela 3 da Seção 5.1.

Para o intervalo de temperatura entre 20°C e 30°C, experimentos apresentados na Seção 5.2, as regras encontradas pelo algoritmo Apriori todas têm uma confiança de 100%. As regras descobertas pelo algoritmo Predictive Apriori têm uma acurácia variando de 80% até 48% e as regras geradas pelo algoritmo FP-Growth têm uma confiança variando de 100% até 50%.

Algumas regras geradas para esse intervalo de temperatura pelos algoritmos foram idênticas, como é o caso da regra de número 3 do Apriori e a regra de número 14 do Predictive Apriori, a regra de número 7 do Apriori e a regra de número 21 e a regra de número 41 do Apriori e a regra de número 75 do FP-Growth, observadas nas Tabela 4, Tabela 5 e Tabela 6 da Seção 5.2.

Para o intervalo de temperatura entre 30°C e 40°C, experimentos apresentados na Seção 5.3, as regras geradas pelo algoritmo Apriori tiveram a confiança variando entre 80% e 51%, enquanto as regras geradas pelo algoritmo Predictive Apriori teve uma variação da acurácia entre 71% até 50% e as regras geradas encontradas pelo algoritmo FP-Growth tiveram confiança 50% e 67%.

Algumas regras geradas para o intervalo de temperatura entre 30°C e 40°C foram idênticas, como é o caso da regra de número 43 do Apriori e a regra de número 23 do Predictive Apriori, a regra de número 70 do Apriori, a regra de número 39 do Predictive Apriori e a regra

de número 71 do FP-Growth, dentre outras que são apresentadas na Tabela 7, Tabela 8 e Tabela 9 da Seção 5.3.

Para o intervalo de temperatura entre 40°C e 50°C, experimentos apresentados na Seção 5.4, as regras geradas pelo algoritmo Apriori tiveram uma variação de confiança entre 100% até 60%, enquanto as regras geradas pelo Predictive Apriori teve a acuraria variando entre 75% e 57% e o algoritmo FP-Growth encontrou somente uma regra para esse intervalo de confiança 58%.

De acordo com as regras geradas para esse intervalo, algumas regras produzidas foram semelhantes, como é o caso da regra de número 78 do Apriori e a de número 45 do algoritmo Predictive Apriori e a regra de número 44 do Predictive Apriori e a de número 76 do FP-Growth, essas regras foram apresentadas na Tabela 10, Tabela 11 e Tabela 12 da Seção 5.4.

Para o intervalo de temperatura entre 50°C e 60°C, experimentos apresentados na Seção 5.5, as regras encontradas pelo algoritmo Apriori tiveram confiança variando de 100% até 67%, enquanto as regras geradas pelo algoritmo Predictive Apriori teve uma variação da acurácia de 80% até 44% e as regras produzidas pelo algoritmo FP-Growth tiveram confiança 100%.

Ainda para o intervalo de temperatura entre 50°C e 60°C foram encontradas algumas regras semelhantes geradas por algoritmos diferentes, como é o caso da regra de número 4 do Apriori, a de número 15 do Predictive Apriori e a de número 64 do algoritmo FP-Growth, a regra de número 13 do algoritmo Apriori e a de número 65 do algoritmo FP-Growth, dentre outras, apresentadas nas Tabela 13, Tabela 14 e Tabela 15 da Seção 5.5.

Para o intervalo de temperatura entre 60°C e 70°C, experimento apresentado na Seção 5.6, as regras encontradas são idênticas e foram geradas pelos algoritmos Apriori e FP-Growth, respectivamente, com confiança de 100% e 33%.

Notou-se que para alguns intervalos de temperatura em específico os algoritmos geraram uma maior quantidade de regras que para outros intervalos, isso se deve ao maior suporte ou frequência de certos intervalos de temperaturas que os outros intervalos presentes na base de dados estudada.

Para os casos de regras semelhantes obtidas por algoritmos diferentes em que as medidas de interesse calculadas são diferentes, o número da confiança reflete o quanto a consequência da regra é precisa em relação a condição da regra, enquanto a medida de acurácia calculada pelo Predictive Apriori é uma relação calculada entre o suporte e a confiança da regra.

Durante a realização dos experimentos notou-se a geração de regras redundantes, principalmente, geradas pelo algoritmo Apriori devido a sua característica de geração de itens candidatos antes de produzir as regras de associação.

Baseado nas regras geradas pelos experimentos realizados foi possível notar uma diminuição na quantidade de regras geradas conforme a temperatura aumentara. Essa conclusão é decorrente do decremento do suporte em função do aumento da temperatura. Além disso, esse fator também é decorrente da baixa quantidade associações possíveis para altas temperaturas.

Assim como, também foi possível perceber um tempo maior no processamento para geração de regras do algoritmo Predictive Apriori, o tempo não foi cronometrado. Essa afirmação foi concluída durante a realização dos experimentos e em função da ínfima quantidade de segundos requerida pelos demais algoritmos enquanto o Predictive Apriori demorava um pouco mais que os demais.

6. CONCLUSÃO

O processo de transporte de carga na Estrada de Ferro Carajás – EFC sofre influência de inúmeros fatores ligados aos próprios ativos, além de condições externas. Então, é indispensável o acompanhamento de todas as variáveis do processo para garantir a confiabilidade dos ativos evitando paradas de trens na malha, e assim, realizar o objetivo principal que é o transporte de cargas até o Terminal Marítimo de ponta da Madeira – TFPM sem perder produtividade.

Este documento apresentou a aplicação de técnica de aprendizado de máquina utilizando regras de associação para descoberta de padrões de falhas. O intuito do estudo foi definir padrões de falhas despercebidos devido ao grande volume de dados, e dessa forma auxiliar a equipe de manutenção na identificação dos gatilhos antes do evento ocorrer, evitando a parada de trens na ferrovia.

No entanto, existem diversos fatores que podem influenciar também na ocorrência de falhas na composição de um trem. Por exemplo, os trens não são eximidos de influências relacionadas as condições climáticas sazonais.

Visto que no período de chuva a parada do trem em local inapropriado, ou seja, com uma rampa ascendente e com chuva é necessário utilizar artifícios, tal como jogar areia nos trilhos para que o trem consiga movimentar, gerando atrito entre o trilho e a roda, e sair do local, pois o trem realiza um esforço muito maior para movimentar.

E como a ferrovia tem quase 1000 km de extensão, somente 14 locais para leitura das temperaturas durante a viagem do trem é considerado pouco, sendo o tempo e a distância entre uma leitura e a próxima grande.

Os objetivos iniciais desse estudo eram a construção da base de dados histórica correlacionando os diversos dados de leitura dos equipamentos Railbam e Hot Box e a verificação das variáveis escolhidas. Esses objetivos foram alcançados, porém com algumas dificuldades em correlacionar os dados dos equipamentos entre si.

Vale ressaltar que esse tipo de correlação entre esses dois equipamentos foi feito pela primeira vez. Até para questões de priorização de manutenção nos vagões a avaliação dos dados é feita separada, ou seja, não é realizado correlação entre os dados desses dois equipamentos para priorização de manutenção.

Tais dificuldade de correlacionar os dados das medições de temperatura e tipos de falhas acústicas dos rolamentos, em função das leituras dos equipamentos, são afetadas justamente pelas decisões cotidianas. E como os eventos de Hot Box são críticos, a opção de retirar os

vagões alarmados de circulação quando ocorre esse tipo de evento impede que as demais leituras sejam feitas.

Esses tipos de decisões são recorrentes durante o processo de controle de tráfego. Dessa forma, o trabalho foi realizado para entender quais fatores influenciam de forma crítica na tomada de decisão.

Os atributos de interesse desse estudo apresentaram baixo suporte devido à baixa quantidade de exemplos disponíveis para efetuar essa correlação em função da ausência de leitura dos equipamentos no momento ou após a ocorrência do evento, reforçando o que já foi citado acima.

Outros objetivos desse trabalho eram a descoberta e avaliação dos padrões de falhas descobertos intrínsecos ao processo de transporte, que foi apresentado no Capítulo 5, demonstrando os resultados a partir dos experimentos realizados com a implementação dos algoritmos de associação, resultando em uma série de regras apresentadas.

Como conforme a temperatura aumentava menor o suporte dos atributos, a confiança, como parâmetro de entrada, teve que ser variada para que os algoritmos encontrassem a maioria das associações possíveis de acordo com os intervalos de temperaturas determinadas. Verificou-se que o algoritmo Apriori induz a geração de algumas regras redundantes devido ao seu tipo de busca das associações na base de dados.

E a partir de todos os experimentos realizados foi possível perceber que muitas falhas de rolamentos, que são detectadas e categorizadas pelo Railbam, ocorrem ainda com uma temperatura abaixo dos limites de 50°C, caracterizado como pré-alarme, e do limite crítico de 70°C, considerado como alarme.

Apesar das dificuldades, a base de dados permitiu treinar os algoritmos de aprendizado de máquina selecionados para descoberta de regras de associação entre os atributos fornecidos.

Assim, pode se concluir que esse método pode ser utilizado para descobrir novas regras ainda não exploradas para auxiliar no processo de manutenção.

Por fim, sugere-se como trabalhos futuros:

- a) Realização de estudo utilizando outros equipamentos do *Wayside*, tais como o detector de impacto para detecção de falhas de rolamento;
- b) Inclusão dos dados de frequência e amplitude captados pelo Railbam para geração de padrões de falhas e descoberta de novos tipos de alarmes possíveis;
- c) Construção de uma base de dados com maior quantidade de dados correlacionando os dados de todos os equipamentos do *Wayside* com as leituras de temperatura do Hot Box.

REFERÊNCIAS BIBLIOGRÁFICAS

BOZI, M. **Detector de valores fora da faixa de equipamentos de supervisão de velocidade de trens.** Revista Ferroviária, São Paulo, out. 2005.

CARVALHO, A. C. de. **Inteligência Artificial. Uma abordagem de aprendizado de máquina.** LTC, 2011.

CASTRO, R. F. V. **Análise de desempenho dos algoritmos Apriori e Fuzzy Apriori na extração de regras de associação aplicados a um sistema de Detecção de intrusos.** Universidade do Estado do Rio de Janeiro, 2014.

DAVE, M.; KAMAL, J. **Identifying Big Data Dimensions and Structure.** Jaipur : Jagan Nath University, 2017.

FAYYAD, U.; PIATESKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery: An Overview.** In: Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

GYORODI, C. ;GYORODI, R. ;HOLBAN, S. **A Comparative Study of Association Rules Mining Algorithms.** Department of Computer Science, University of Oradea, 2004.

GONÇALVES, E. C. **Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas.** Universidade Federal Fluminense, 2005.

LI, H.; PARIKH, D.; HE, Q. **Improving rail network velocity: A machine learning approach to predictive maintenance.** Elsevier, 2014. United States.

Manual técnico do equipamento Railbam, 2011.

Manual técnico do equipamento Hot Box, 2011.

PATIL, B.M.; RAMESH, C. J.; DURGA, T. **Classification of type-2 diabetic patients by using Apriori and Predictive Apriori.** *International J. Co,putational Vision and Robotics*, Vol 2, No 3, 2011.

POONSIRIVONG, K.; JITTAWIRIAYNUKON, C. **Big Data Analytics Using Association Rules in eLearning.** *3rd International Conference on Big Data Analysis*, 2018.