Disponível em: http://periodicos.ufes.br/BJPE/index

ISSN: 2447-5580

Brazilian Journal of Production Engineering
BJPE - Revista Brasileira de Engenharia de Produção

Campus São Mateus
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

ARTIGO ORIGINAL OPEN ACCESS

DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS ABERTA DA POLÍCIA RODOVIÁRIA FEDERAL: IDENTIFICAÇÃO DE PONTOS CRÍTICOS NA RODOVIA BR 101 NO MUNICÍPIO DE SÃO MATEUS/ES

KNOWLEDGE DISCOVERY IN THE FEDERAL HIGHWAY POLICE'S OPEN DATABASE: IDENTIFICATION OF CRITICAL POINTS ON HIGHWAY BR 101 IN THE MUNICIPALITY OF SÃO MATEUS/ES

Flavia da Silva Nogueira[1]; Luciana Lee[2]; Silvia das Dores Rissino[3]

[1 2 3] Departamento de Computação e Eletrônica do Centro Universitário Norte do Espírito Santo da Universidade Federal do Espírito Santo, Rodovia BR 101 Norte, Km. 60, Bairro Litorâneo, CEP 29932-540, São Mateus.
flavia-sn@hotmail.com, lucianalee@gmail.com, srissino@gmail.com

ARTIGO INFO.

Recebido em: 16/10/2018

Aprovado em: 07/11/2018

Disponibilizado em: 15/12/2018

PALAVRAS-CHAVE:

Descoberta de Conhecimento; Dados Abertos; Pontos Críticos; Acidentes Rodoviários; Rodovia BR 101.

KEYWORDS:

Discovery of Knowledge; Open Data; Critical Points; Road Accidents; Highway BR 101.

***Autor Correspondente:** Flavia da Silva Nogueira.

RESUMO

Este trabalho propõe o uso do Processo de Descoberta de Conhecimento em Base de Dados, com o objetivo de analisar a base de dados aberta da Polícia Rodoviária Federal e apresentar os pontos críticos da Rodovia BR 101, no trecho entre o km 55 até km 90 Norte, do Município de São Mateus/ES. O conjunto de dados em uso é o de acidentes ocorridos no ano de 2016, na BR 101. Na fase de pré-processamento os dados são limpos e adequados para a próxima fase através do uso das ferramentas Calc e WEKA. Na fase de mineração de dados utiliza-se o Weka para análise dos dados com aplicação do Algoritmo Apriori. O resultado da análise exibe o conhecimento útil através de infográficos e mapas que abrange os pontos críticos no trecho avaliado, quais dias da semana com maior incidência de acidentes e quais as principais causas.

ABSTRACT

This work proposes the use of the Knowledge Discovery Process in a database, with the objective of analyzing the Federal Highway Police's open database and presenting the critical points of the BR 101 Highway, in the stretch from km 55 to km 90 North, from Municipality of São Mateus/ES. The data set in use is that of accidents occurring in the year 2016, in BR 101. In the pre-processing phase the data is clean and suitable for the next phase using of the Calc and WEKA tools. In the data mining phase, the Weka is used to analyze the data using the Apriori algorithm. The result of the analysis shows the useful knowledge through infographics and maps that covers the critical points in the section evaluated, which days of the week with the highest incidence of accidents and which are the main causes.

Citação (APA): NOGUEIRA, F. da S., LEE, L. & RISSINO, S. das D. (2018). Descoberta de Conhecimento na base de dados aberta da Polícia Rodoviária Federal: Identificação de Pontos Críticos na Rodovia BR 101 no Município de São Mateus/ES. Brazilian Journal of Production Engineering, 4(4): 70-90.

INTRODUÇÃO

A BR 101 é uma rodovia federal, longitudinal brasileira, onde o ponto inicial está localizado na cidade de Touros no Rio Grande do Norte e o final na cidade de São José do Norte no Rio Grande do Sul, como mostra a Fig. 1. Esta BR é a mais extensa rodovia brasileira, com comprimento de pista de 4.615 km, em que diariamente acidentes rodoviários são registrados (BIT, 2017).

Figura 1. Mapa com a extensão da BR 101 no Brasil

Fonte – BIT, 2017

No ano de 2016 ocorreram 96.356 acidentes em rodovias federais, na BR 101 ocorreram 14.567 acidentes, mas no estado do Espírito Santo – ES ocorreram 15%, do total destes (PRF, 2017). Em extensão territorial, o ES não corresponde a 15% dos 4.615 km da BR 101. Portanto, identificar os pontos críticos em determinadas regiões pode ser o diferencial para a diminuição dos acidentes.

Para os fins deste trabalho, serão explorados dados secundários provenientes dos dados abertos no sítio eletrônico da Polícia Rodoviária Federal (PRF) do ano de 2016. Nestes dados, são aplicadas o Processo de Descoberta de Conhecimento (KDD), com ênfase na fase de mineração de dados, a fim de encontrar padrões e associações entre as variáveis relacionadas entre acidentes de trânsito. O conjunto de dados selecionados é o relacionado aos acidentes ocorridos no ano de 2016, na BR 101, no trecho correspondente ao km 55 até km 90 Norte, do município de São Mateus-ES. Para este estudo, é considerado como trecho crítico a região

com maior índice de acidentes, sendo analisados os fatores que podem contribuir para a frequência dos mesmos.

A definição dos pontos críticos em uma rodovia é de suma importância, uma vez que, com tal conhecimento, pode-se identificar necessidades de investimentos ou mesmo de campanhas de educação para o trânsito (Branco, 1999).

Há estudos que analisam a influência da qualidade das rodovias com o índice de acidentes e pode-se encontrar conclusões opostas sobre tal assunto, como no estudo apresentado por Shikida, no qual o autor tenta provar que o investimento em infraestrutura viária pode induzir o condutor a uma direção sem prudência. Eles concluíram que trechos bem sinalizados e com boas condições são considerados críticos, uma vez que quando as condições de tráfego e tempo são favoráveis, os motoristas tendem a ser mais agressivos (Shikida et al., 2008).

No trabalho de (Branco, 1999), são apresentadas conclusões que são contrárias àquela obtida por (Shikida et al., 2008). No primeiro trabalho, o autor conclui que uma estrada segura pode reduzir a gravidade dos acidentes, mas o condutor pode ser mais imprudente. Para o segundo autor, as rodovias que possuem dispositivos adequados de proteção podem interferir positivamente na diminuição do número de acidentes bem como minimizar os fatores contribuintes.

Embora não haja uma definição universal para trechos críticos, para (Geuters e Wets, 2003), esses locais normalmente estão ligados a um alto grau de acidentes. Os autores ainda citam que as regras para encontrar os trechos críticos podem ser baseadas no total de acidentes em um local.

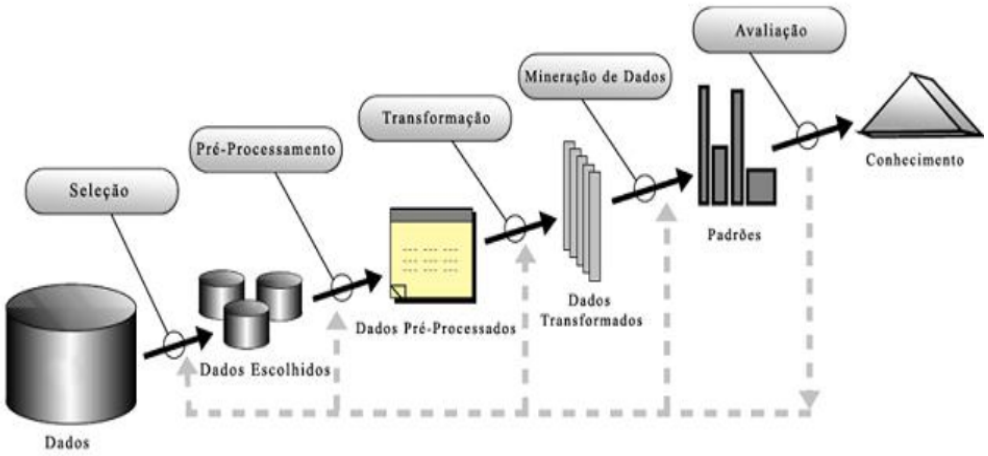
2. DESCOBERTA DE CONHECIMENTO EM BASE DADOS

O processo de Descoberta de Conhecimento em Base de Dados (*Discovery of Knowledge in Database* - KDD) é um processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados (Fayyad et al., 1996).

Toda aplicação de KDD possui três tipos de componentes: o problema em que será aplicado o processo de KDD, os recursos disponíveis para a solução do problema e os resultados obtidos a partir da aplicação dos recursos disponíveis em busca da solução do problema (GOLDSCHMIDT; PASSOS, 2005).

OKDD é composto pelas fases de Pré-processamento, Mineração de Dados e Pós-processamento, além das fases intermediárias, conforme como mostra a Fig. 2, e a descrição das mesmas apresentadas a seguir:

Figura2. Processo de Descoberta de Conhecimento em Base de Dados - KDD



Fonte - Adaptado de Fayyad, 1996

- ***Seleção***: Nessa fase o conjunto de dados são escolhidos para as fases posteriores;
- ***Pré-processamento***: Ocorre a limpeza do conjunto de dados para a garantia da qualidade da análise. Pode-se eliminar dados faltantes ou substituí-los por algo pré-definido. Atributos contendo zero podem ser eliminados de acordo com a decisão do especialista em KDD;
- ***Transformação***: Acontece logo após o pré-processamento, deixando os dados em um formato aceitável para aplicar o algoritmo de mineração;
- ***Mineração de Dados***: Fase em que os algoritmos são executados a fim de encontrar padrões relevantes na base de dados;
- ***Pós-Processamento e Avaliação***: Os padrões extraídos são interpretados e avaliados a fim de obter novos padrões de conhecimento e posteriormente utilizá-los. Nesta fase é avaliado a utilidade do conhecimento extraído na etapa anterior, para que possa ser utilizado como fator para a tomada de decisão de um especialista ou de um sistema especialista

3. CONTEXTUALIZAÇÃO DA BASE DE DADOS

3.1 DADOS ABERTOS

Dados Abertos são dados que estão livremente disponíveis para todos utilizarem e redistribuírem como desejarem, sem restrição de licenças, patentes ou mecanismos de controle. Segundo o Portal Brasileiro de Dados Aberto, para serem considerados dados abertos, os mesmos devem ser publicados em um formato legível por máquina (Brasil, 2017).

O Governo é um principal contribuinte neste contexto. “No Brasil, o direito de cada cidadão ter acesso aos dados está previsto na Lei Federal 12.527/2011, conhecida como Lei de Acesso à Informação” (Brasil, 2011).

O Manual dos dados abertos apresenta um conjunto de áreas e atividades em que os dados abertos estão gerando valor, entre as quais tem-se: Transparência e controle democrático; Participação popular; Empoderamento dos cidadãos; Melhores ou novos produtos e serviços privados; Inovação; Melhora na eficiência dos serviços governamentais; Conhecimento novo a partir da combinação de fontes de dados e padrões (NIC, 2017).

3.2 DADOS ABERTOS DA POLÍCIA RODOVIÁRIA FEDERAL

As estatísticas publicadas pelos órgãos governamentais brasileiros apresentam um aumento significativo quanto ao número de acidentes de trânsito nas rodovias, que podem ser consequência de diversos fatores como trechos críticos, condições do condutor do veículo entre outros, por isso a importância da análise dos dados destas fontes (PRF, 2017).

A base de dados ou o conjunto de dados a ser analisado foi obtido do sitio eletrônico da PRF, o qual está disponibilizado no formato de planilhas cvs (*common value separated*) e apresenta os dados do ano de 2016 com todos os acidentes ocorridos em todas nas rodovias brasileiras.

4. FERRAMENTAS E ALGORITMOS

4.1 FERRAMENTA DO PRÉ-PROCESSAMENTO

Os dados disponibilizados no sitio eletrônico da PRF estão no formato de planilhas, utilizam-se as ferramentas Calc (Planilha da suíte do *LibreOffice*) e o Software Weka para a fase inicial dos trabalhos de pré-processamento.

O Calc é um componente de planilha da suíte do *LibreOffice*, está disponível para diversos Sistemas Operacionais, dentre eles Windows, Solaris, Mac e Linux. Essa suíte utiliza o

formato *OpenDocument*, que é um formato de documento aberto para aplicações. A licença do *LibreOffice* é do tipo MPLv2.0 (*secondary license GPL, LGPLv3+ or Apache License* 2.0).

Os aplicativos presentes no *LibreOffice* são: *Calc, Write, Base, Math, Draw e Impress*, neste trabalho utilizou-se a versão 5.4.1.2 do *LibreOffice* (*LibreOffice*, 2017).

4.2 FERRAMENTA PARA MINERAÇÃO DE DADOS

O Software Weka (*Waikato Environment for Knowledge Analysis*)é uma ferramenta de código aberto desenvolvida pela Universidade de Waikato na Nova Zelândia. É muito utilizada no meio acadêmico, além de ser bem didática. Encontra-se licenciada como GPL (*General Public License*), sendo possível estudar e alterar o respectivo código fonte (WEKA, 2016).

O Weka foi escolhido por ser uma referência na área de mineração de dados, possuir diversos algoritmos implementados em sua biblioteca, ser um bom ambiente didático e oferecer portabilidade em função ser desenvolvida em Java(Galvão, 2010, p. 38).Neste trabalho utilizou-se a versão 3.8, a qual apresenta vários algoritmos implementados o que torna possível realizar testes entre algoritmos no mesmo ambiente, possibilitando atingir os objetivos do trabalho.

Para a realização de testes no ambiente exploratório do WEKA, é necessário realizar o pré-processamento, que é a adequação dos dados através de algoritmos de limpeza de dados que removem ruídos, valores faltantes ou inconsistentes no banco de dados.

4.3 FERRAMENTA DE PÓS-PROCESSAMENTO

Com a fase de mineração de dados finalizada, é necessário organizar e apresentar o resultado encontrado em forma de conhecimento. Nesta fase, foi utilizado a plataforma Canva que permite a visualização da informação em formato de conhecimento e o *Google Maps*® para a construção dos mapas com os locais e trecho críticos.

O Canva é uma plataforma online gratuita de criação de conteúdos gráficos, o qual disponibiliza um ambiente no formato arrastar e soltar, além de fornecer acesso a uma grande quantidade de fotografias, gráficos e fontes (Canva, 2017).

O *Google Maps* é um serviço de pesquisa, de visualização e localização geográfica através de mapas e satélites da Terra de forma gratuita na web, desenvolvido e fornecido pela empresa americana Google a partir de 2005. Este serviço permite traçar rotas de vários destinos, entre o destino atual e o desejado ou entre dois locais quaisquer. Além disso, informa o tempo entre

os locais, rotas alternativas e oferece também o serviço de zoom, principalmente em grandes metrópoles, além do serviço de fotos de satélites que permitem visualizar claramente um lugar, com auxílio do cursor do mouse é possível navegar através do local escolhido, e obter informações a respeito do local desejado (Google, 2017).

4.4 ALGORITMO APRIORI

Algoritmo Apriori é um dos mais famosos algoritmos para mineração de dados, utiliza um hash sobre uma árvore para coletar informações em um banco de dados. Por ser um clássico da tarefa de associação, realiza "buscas" sucessivas na base de dados mantendo um ótimo desempenho em termos de tempo de processamento" (HAYKIN, 1999).

Com aplicação do Aprioi é possível observar a frequência com que o item aparece dentro do conjunto de dados, utilizando o método candidato, possibilitando determinar quais relações são importantes ou não para o usuário, além de exibir as associações relevantes. Neste caso, o algoritmo usa um modelo matemático, onde as regras de associação geradas devem atender a um suporte e confiança mínimos especificados pelo analista (Agrawal et al, 1993). O Apriori parte do princípio de analisar os itens que mais aparecem na base para gerar as regras, diminuindo assim o número de candidatos que serão comparados em cada transação e levando o algoritmo a um bom desempenho computacional (SOUZA, 2015).

5. METODOLOGIA

5.1. SELEÇÃO DOS DADOS

Visando descobrir associações entre fatores envolvidos nos acidentes do trecho definido, a Descoberta de Conhecimento foi realizada seguindo as fases deste processo utilizando o conjunto de dados do ano de 2016 do sítio eletrônico da PRF, contendo todos os acidentes nas rodovias brasileiras. Os dados estão contidos em um arquivo no formato de uma planilha cvs, o que facilita a seleção dos atributos relevantes para a análise.

Com o objetivo de selecionar apenas os dados necessários, foram utilizados recursos do Calc que possibilitam escolher os dados dos acidentes ocorridos apenas na BR 101, depois somente os ocorridos no estado do Espírito Santo e por fim os do município de São Mateus (trecho em estudo).

O número inicial de acidentes era de 96.363 (total de acidentes no Brasil), após a seleção diminui para 154 (total de acidentes no trecho em destaque), sendo que este conjunto de dados encontrado foi utilizado na próxima fase do processo de KDD.

5.2 - PRÉ-PROCESSAMENTO

5.2.1.PRIMEIRA ETAPA DO PRÉ-PROCESSAMENTO

O Calc foi utilizado para realizar o carregamento da planilha contendo os dados de todos os acidentes ocorridos no ano de 2016 no trecho em estudo (seleção da fase anterior). Com os dados carregado, inicia-se a adequação do conjunto de dados para análise utilizando os recursos de:

- Substituição: o qual substitui MAIÚSCULA/ MINÚSCULA, o que resultou em todos os caracteres foram colocados em maiúsculo.

- Localizar/substituir: o qual todas as letras do conjunto de dados, que estavam acentuadas, foram substituídas por vogais sem acento.

Essas ações tiveram o objetivo de padronizar a escrita de todas as palavras no arquivo, para posterior carregamento no Weka, já que esta ferramenta é *case sensitive*, isto é, caso uma mesma palavra ocorra com escrita semelhante, é considerada mais de uma palavra sendo analisada como um dado diferente, o que prejudica a credibilidade da mineração.

5.2.2. SEGUNDA ETAPA DO PRÉ-PROCESSAMENTO

A planilha resultante, com apenas os acidentes ocorridos no trecho em analise, é carregada no Weka, para finalização da etapa de pré-processamento como:

-Remover atributos que não são relevantes para o trabalho;

- Converter atributos numéricos a nominais para poder aplicar o algoritmo de associação, que trabalha apenas com dados nominais, e o atributo “KM” que indica em qual quilômetro do município ocorreu o acidente estava como atributo numérico.

5.3 TRANSFORMAÇÃO DO CONJUNTO DE DADOS

A planilha original, utilizada na fase de seleção apresentava (20) atributos, dos quais identificava-se BR e CIDADE (esses dois por se tratarem de critérios de seleção), são excluídos. CONDICAO_METEOROLOGICA, USO_SOLO, MORTOS, FERIDOS, entre outros foram retirados, pois não influenciavam nas associações interessadas. A Fig. 3, apresenta um pequeno recorte do conjunto de dados antes dos procedimentos de limpeza e, a

Fig. 4 apresenta a planilha com o conjunto de dados resultantes após a limpeza da etapa anterior.

Observa-se que ocorreu a redução vertical (atributos) e horizontal (dados), os quais foram objetos da análise pelo Weka. O conjunto de dados resultante é denominado de dadosSAOMATEUS.cvs e possui um conjunto de seis atributos (Dia_Semana, KM, Causa_Acidente, Classificacao_Acidente, Tipo_Pista e Tracado_Via).

Figura3: Planilha de dados original antes da limpeza dos dados

	A	B	C	D	E	F	G	H	I	J
1	id	data_inversa	dia_semana	horario	uf	br	km	municipio	causa_acidente	tipo_acidente
2	36727	10/06/16	Sexta	18:30:00	RJ	101	66	CAMPOS DOS GOYTACAZES	Outras	Colisão com objeto fixo
3	83425846	01/01/16	Sexta	01:30:00	SC	101	135,5	BALNEARIO CAMBORIU	Falta de atenção	Colisão traseira
4	83425850	01/01/16	Sexta	01:00:00	PR	277	2	PARANAGUA	Ingestão de álcool	Colisão traseira
5	83425852	01/01/16	Sexta	01:45:00	PR	476	357	UNIAO DA VITORIA	Outras	Colisão com bicicleta
6	83425853	01/01/16	Sexta	02:00:00	SE	101	94,2	NOSSA SENHORA DO SOCORRO	Falta de atenção	Queda de motocicleta / bicicleta / veículo
7	83425855	01/01/16	Sexta	02:20:00	SC	282	3	FLORIANOPOLIS	Falta de atenção	Colisão lateral
8	83425856	01/01/16	Sexta	01:10:00	MG	50	61	UBERLANDIA	Velocidade incompatível	Colisão lateral
9	83425858	01/01/16	Sexta	02:10:00	PR	476	152	ARAUCARIA	Velocidade incompatível	Colisão traseira
10	83425860	01/01/16	Sexta	01:45:00	MS	163	257,3	DOURADOS	Velocidade incompatível	Colisão traseira
11	83425861	01/01/16	Sexta	02:30:00	SC	101	207,8	SAO JOSE	Falta de atenção	Colisão lateral
12	83425862	01/01/16	Sexta	02:15:00	MA	230	2	BARAO DE GRAJAU	Ultrapassagem indevida	Colisão lateral
13	83425864	01/01/16	Sexta	01:20:00	RR	174	507	BOA VISTA	Falta de atenção	Colisão traseira
14	83425866	01/01/16	Sexta	04:40:00	MG	50	201,3	DELTA	Outras	Capotamento
15	83425867	01/01/16	Sexta	02:00:00	GO	40	2,8	VALPARAISO DE GOIAS	Outras	Colisão Transversal
16	83425868	01/01/16	Sexta	03:50:00	GO	60	137,8	GOIANIA	Ingestão de álcool	Colisão traseira
17	83425871	01/01/16	Sexta	04:00:00	MG	116	513,3	UBAPORANGA	Dormindo	Queda de motocicleta / bicicleta / veículo
18	83425873	01/01/16	Sexta	01:00:00	SE	101	94	NOSSA SENHORA DO SOCORRO	Ingestão de álcool	Queda de motocicleta / bicicleta / veículo
19	83425875	01/01/16	Sexta	01:00:00	MA	222	677	ACAILANDIA	Falta de atenção	Colisão Transversal
20	83425876	01/01/16	Sexta	05:00:00	SC	101	213	PALHOCA	Dormindo	Capotamento

Fonte - Próprio autor, 2018

Figura4: Planilha como os dados resultantes após limpeza – dadosSAOMATEUS.CSV

	A	B	C	D	E	F
1	DIA_SEMANA	KM	CAUSA_ACIDENTE	CLASSIFICACAO_ACIDENTE	TIPO_PISTA	TRACADO_VIA
2	QUARTA	69	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
3	SABADO	68	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	CRUZAMENTO
4	SEGUNDA	81.8	ULTRAPASSAGEM INDEVIDA	COM VITIMAS FERIDAS	SIMPLES	RETA
5	DOMINGO	65	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	RETA
6	QUINTA	66	FALTA DE ATENCAO	SEM VITIMAS	SIMPLES	RETA
7	SEXTA	70	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	RETA
8	SEXTA	68.5	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
9	QUARTA	61	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
10	SABADO	66	DESOBEDIENCIA DA SINALIZACAO	COM VITIMAS FERIDAS	SIMPLES	RETA
11	SEGUNDA	58	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	RETA
12	SEGUNDA	67.2	ANIMAIS NA PISTA	SEM VITIMAS	SIMPLES	RETA
13	QUINTA	74	DESOBEDIENCIA DA SINALIZACAO	COM VITIMAS FERIDAS	SIMPLES	RETA
14	SEXTA	66.5	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	RETA
15	QUARTA	65	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
16	QUARTA	70.2	DESOBEDIENCIA DA SINALIZACAO	COM VITIMAS FERIDAS	SIMPLES	CRUZAMENTO
17	DOMINGO	64.8	FALTA DE ATENCAO	COM VITIMAS FATAIS	SIMPLES	RETA
18	QUINTA	72.4	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
19	QUARTA	66	NAO GUARDAR DISTANCIA DE SEGURANCA	COM VITIMAS FERIDAS	SIMPLES	RETA
20	SEXTA	87.2	DEFEITO MECANICO EM VEICULO	SEM VITIMAS	SIMPLES	CRUZAMENTO
21	SABADO	68	FALTA DE ATENCAO	COM VITIMAS FERIDAS	DUPLA	RETA
22	DOMINGO	66	OUTRAS	COM VITIMAS FERIDAS	MULTIPLA	RETA
23	SABADO	79.6	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
24	SABADO	67	FALTA DE ATENCAO	COM VITIMAS FERIDAS	MULTIPLA	RETA
25	SEGUNDA	76	DORMINDO	COM VITIMAS FERIDAS	SIMPLES	RETA
26	SABADO	67	DESOBEDIENCIA DA SINALIZACAO	COM VITIMAS FERIDAS	DUPLA	CRUZAMENTO
27	QUARTA	66.9	FALTA DE ATENCAO	COM VITIMAS FERIDAS	MULTIPLA	RETA
28	QUARTA	85	OUTRAS	IGNORADO	SIMPLES	RETA

Fonte - Próprio autor, 2018.

5.4 MINERAÇÃO DE DADOS

Na fase de Mineração de Dados utiliza-se a ferramenta Weka, por possuir uma biblioteca com diversos algoritmos de aprendizagem de máquina e de mineração de dados já implementados.

Em mineração de dados e aprendizado de tratamento, regras de associação são usadas para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados. Logo, para gerar as regras de associações necessárias para o conjunto de dados, utiliza-se o Algoritmo Apriori, por ser muito eficiente na geração de regras de associação em relação aos parâmetros de suporte e confiança, além de ter bom desempenho em tempo de execução para o conjunto de dados em análise.

Esta fase inicia-se com a carga do conjunto de dados do arquivo dadosSAOMATEUS.cvs, obtido na fase anterior, na ferramenta Weka, tendo como objetivo a escolha dos atributos relevantes que indicam associações entre os fatores envolvidos na causa de um acidente.

O arquivo dados SAOMATEUS.cvs, com tais atributos, é salvo no formato *arff* para fins de segurança e praticidade, pois este é o padrão de arquivo de dados para uso na ferramenta Weka.

O Algoritmo Apriori foi aplicado nos dados contidos no arquivo dados SAOMATEUS.arff, após os devidos ajustes nos parâmetros da janela de configuração do Weka. Os parâmetros que, normalmente, são configurados são: min Metric (valor mínimo de métrica); numRules (número de regras a serem encontradas); lowerBoundMinSupport (limite inferior para o suporte mínimo); treat Zero As Missing (se habilitado, o zero é tratado como valor faltante); metricType (Configura um tipo de regras através de um ranking de regras, confiança, peso, alavancagem e convicção; upperBoundMinSupport (Limite superior para o mínimo suporte. Começa decrementando o suporte a partir deste valor).

A configuração deve ser realizada considerando o suporte e confiança mínimos para a geração das regras de associação, assim como, indicar no campo *numRules* o número de regras que devem ser geradas no máximo, caso haja mais regras que satisfaça as condições de suporte e confiança mínimos.

Os parâmetros num Rule, min Metric e lower Bound Min Suppor, neste caso, foram configurados, considerando sempre a Confiança como tipo de métrica.

Uma boa medida de confiança, está em 0.9 ou 90%, o suporte de 0.1 ou 10% e com numRule configurado em 100 regras.

Com os ajustes dos parâmetros, foram obtidas 14 regras de associação, como mostrado na Tabela 1.

Tabela 1: Regras de associação geradas com base em confiança igual a 0.9.

Número	MELHORES REGRAS
1	DIA_SEMANA=SABADO TIPO_PISTA=SIMPLES TRACADO_VIA=RETA 19 ==> CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS 19 <conf:(1)> lift:(1.21) lev:(0.02) [3] conv:(3.33).
2	TIPO_PISTA=DUPLA 18 ==> CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS 18 <conf:(1)> lift:(1.21) lev:(0.02) [3] conv:(3.16).
3	DIA_SEMANA=QUINTA 22 ==> TRACADO_VIA=RETA 21 <conf:(0.95)> lift:(1.16) lev:(0.02) [2] conv:(1.93).
4	DIA_SEMANA=SABADO TIPO_PISTA=SIMPLES 22 ==> CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS 21 <conf:(0.95)> lift:(1.16) lev:(0.02) [2] conv:(1.93).
5	DIA_SEMANA=SEXTA CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS 19 ==> TRACADO_VIA=RETA 18 <conf:(0.95)> lift:(1.15) lev:(0.02) [2] conv:(1.67).
6	TIPO_PISTA=MULTIPLA 18 ==> TRACADO_VIA=RETA 17 <conf:(0.94)> lift:(1.15) lev:(0.01) [2] conv:(1.58).
7	DIA_SEMANA=QUINTA CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS 17 ==> TRACADO_VIA=RETA 16 <conf:(0.94)> lift:(1.14) lev:(0.01) [1] conv:(1.49).
8	DIA_SEMANA=QUINTA TIPO_PISTA=SIMPLES 17 ==> TRACADO_VIA=RETA 16 <conf:(0.94)> lift:(1.14) lev:(0.01) [1] conv:(1.49).
9	CLASSIFICACAO_ACIDENTE=SEM VITIMAS 16 ==> TIPO_PISTA=SIMPLES 15 <conf:(0.94)> lift:(1.22) lev:(0.02) [2] conv:(1.87).
10	CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS TIPO_PISTA=MULTIPLA 16 ==> TRACADO_VIA=RETA 15 <conf:(0.94)> lift:(1.14) lev:(0.01) [1] conv:(1.4).
11	DIA_SEMANA=SEXTA 24 ==> TRACADO_VIA=RETA 22 <conf:(0.92)> lift:(1.11) lev:(0.01) [2] conv:(1.4).
12	DIA_SEMANA=SABADO TRACADO_VIA=RETA 24 ==> CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS 22 <conf:(0.92)> lift:(1.11) lev:(0.01) [2] conv:(1.4).
13	DIA_SEMANA=SABADO CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS TIPO_PISTA=SIMPLES 21 ==> TRACADO_VIA=RETA 19 <conf:(0.9)> lift:(1.1) lev:(0.01) [1] conv:(1.23).
14	DIA_SEMANA=SABADO 31 ==> CLASSIFICACAO_ACIDENTE=COM VITIMAS FERIDAS 28 <conf:(0.9)> lift:(1.1) lev:(0.02) [2] conv:(1.36)

Fonte –Autores, 2018.

A Tabela 2 apresenta a descrição das regras de associação geradas em formato de texto com o respectivo número da regra, conforme as regras exibidas na Tabela 1.

Tabela 2: Descrição das regras de associação geradas em formato texto.

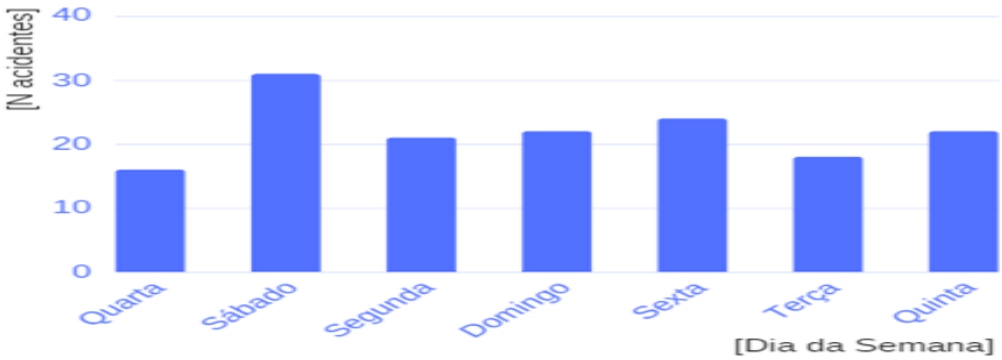
Número Regra	Descrição
1	Se o dia da semana for SÁBADO, o tipo de pista for SIMPLES e o traçado da pista for uma RETA, então em todos os casos que esses três fatores aparecem juntos, o acidente é classificado com VÍTIMAS FERIDAS, apresentando uma confiança de 1 ou 100%.
2	Dos 18 acidentes ocorridos em pista DUPLA, todos deixaram VÍTIMAS FERIDAS. Confiança 1 ou 100%.
3	Dos 22 acidentes ocorridos na QUINTA, 21 deles foram em pista com o traçado de RETA. Configurando uma confiança de 0.95 ou 95%.
4	Dos 22 acidentes ocorridos no SÁBADO em pista SIMPLES, 21 deles tiveram VÍTIMAS FERIDAS. Configurando uma confiança de 0.95 ou 95%.
5	Dos 19 acidentes ocorridos na SEXTA com VÍTIMAS FERIDAS, 18 deles ocorreram em pistas cujo traçado era uma RETA. Configurando uma confiança de 0.95 ou 95%.
6	De todos os acidentes ocorridos em pista MÚLTIPLA (18), 17 ocorreram em pista cujo traçado era RETA. Configurando uma confiança de 0.94 ou 94%.
7	Dos 17 acidentes ocorridos na QUINTA que possuíram VÍTIMAS FERIDAS, 16 deles ocorreram em pista cujo traçado era RETA. Confiança de 0.94 ou 94%.
8	Dos 17 acidentes ocorridos na QUINTA em pista SIMPLES, 16 deles ocorreram em pista cujo traçado era RETA. Confiança de 0.94 ou 94%.
9	Dos 17 acidentes ocorridos SEM VÍTIMAS, 16 deles aconteceu em pista SIMPLES. Confiança de 0.94 ou 94%.
10	De 16 acidentes ocorridos em pista MÚLTIPLA e COM VÍTIMAS FERIDAS, 15 deles ocorreu em RETA. Confiança de 0.94 ou 94%.
11	Dos 24 acidentes ocorridos na SEXTA, 22 deles ocorreu em uma RETA. Confiança de 0.92 ou 92%.
12	Dos 24 acidentes ocorridos no SÁBADO em traçados tipo RETA, 22 possuem VÍTIMAS FERIDAS. Confiança de 0.92 ou 92%.
13	Dos 21 acidentes ocorridos no SÁBADO, COM VÍTIMAS FERIDAS e num tipo de pista SIMPLES, 19 ocorreram em RETA. Confiança de 0.90 ou 90%.
14	Dos 31 acidentes ocorridos no SÁBADO, 29 deles tiveram VÍTIMAS FERIDAS. Confiança de 0.90 ou 90%.

Fonte – Autores, 2018

5.3 PÓS-PROCESSAMENTO E AVALIAÇÃO DO RESULTADO DA MINERAÇÃO

Com o objetivo de apresentar os resultados da mineração de dados, isto é, transformar as regras de associações em conhecimento, são gerados os infográficos e mapas, exibindo cada classe separadamente com os respectivos números de acidentes e os locais de maior incidência de problemas. Nas Fig.5 até a Fig. 10 são exibidos os infográficos com as informações obtidas na mineração de dados. Observando que os infográficos foram produzidos pelos autores com auxílio da Plataforma Canva.

Figura 5. Infográfico com o número de acidentes por dia da semana.

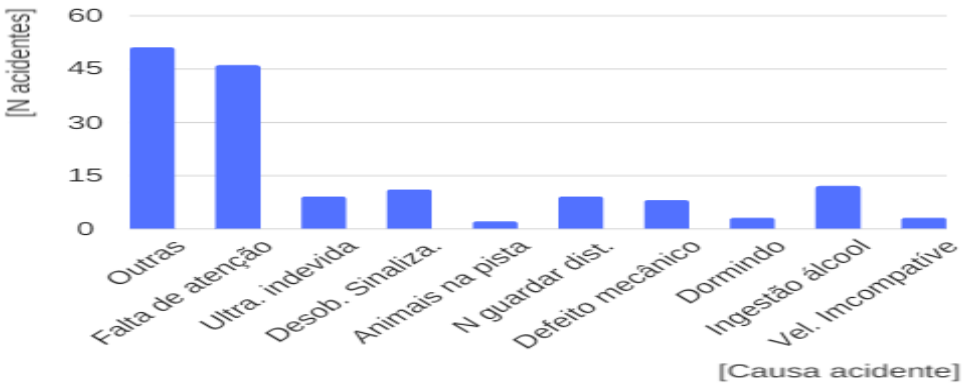


Fonte: elaborada pelos autores

Figura 6. Infográfico como número de acidentes por Km de ocorrência

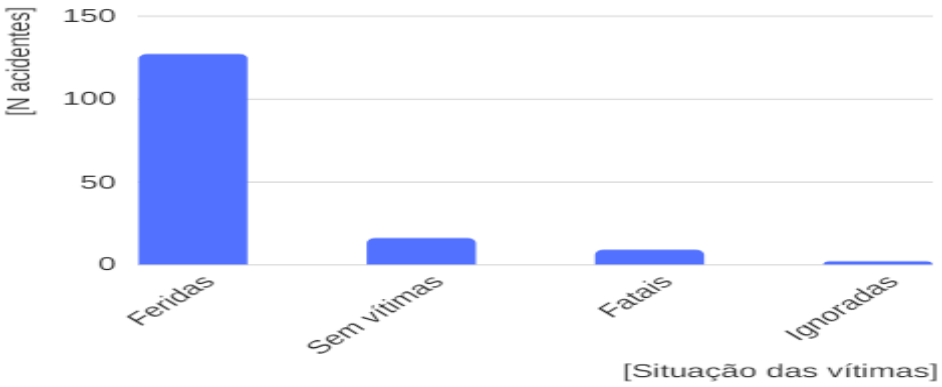
Fonte: elaborada pelos autores

Figura7. Infográfico com o número de acidentes por Causa do acidente



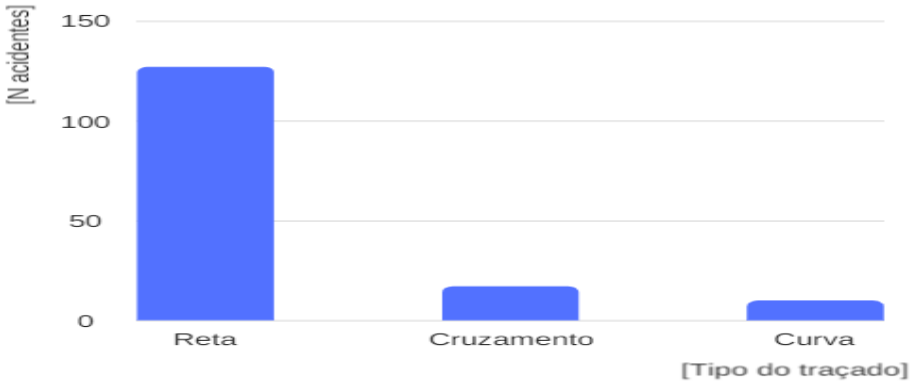
Fonte: elaborada pelos autores

Figura 8: Número de acidentes por classificação da situação da vítima



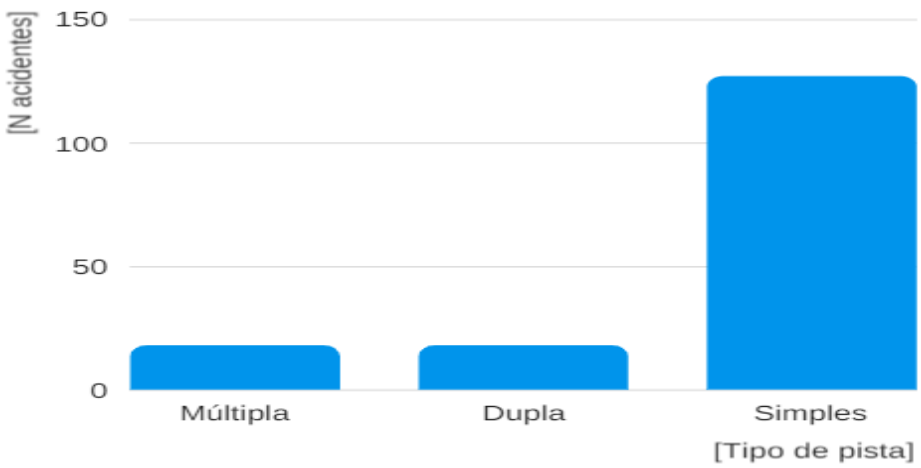
Fonte: elaborada pelos autores

Figura9: Número de acidentes por tipo de traçado da pista



Fonte: elaborada pelos autores

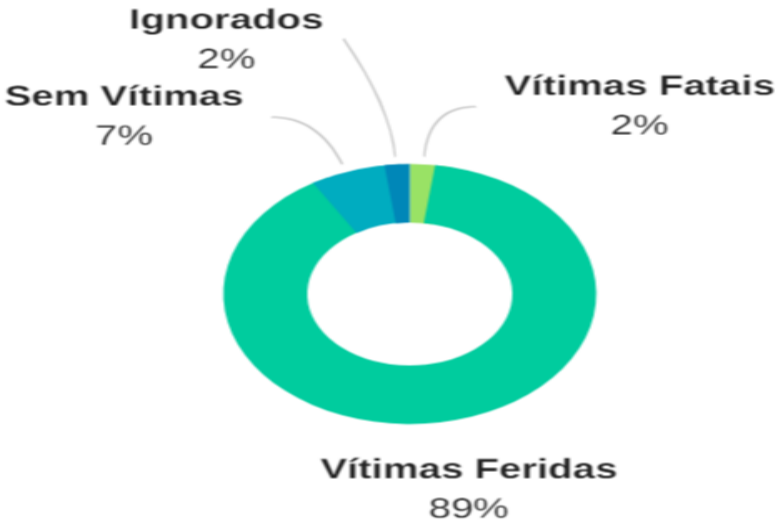
Figura10: Número de acidentes por tipo de pista.



Fonte: elaborada pelos autores

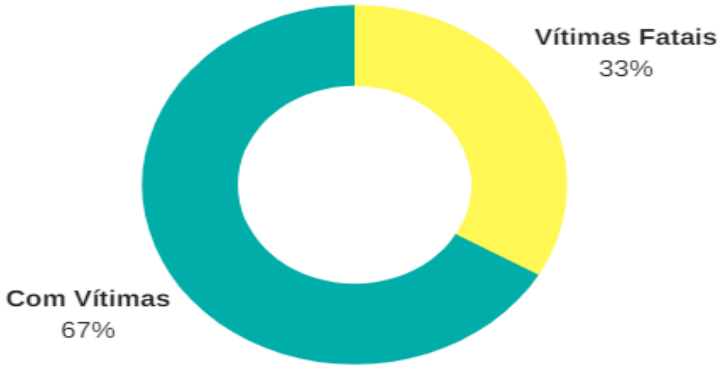
Os próximos infográficos apresentam a relação entre o percentual de vítimas com as quatro principais causas de acidentes: falta de atenção do condutor na Fig.11, ultrapassagem indevida na Fig. 12, desobediência a sinalização na Fig. 13 e ingestão de bebida alcóolica na Fig. 14.

Figura 11: Acidentes causados por falta de atenção do condutor



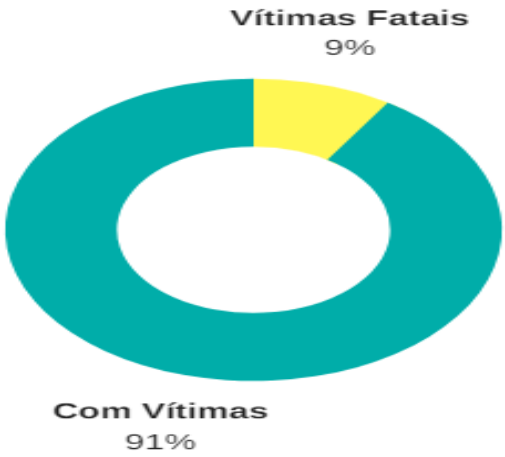
Fonte: elaborada pelos autores

Figura 12: Acidentes causados por ultrapassagem indevida



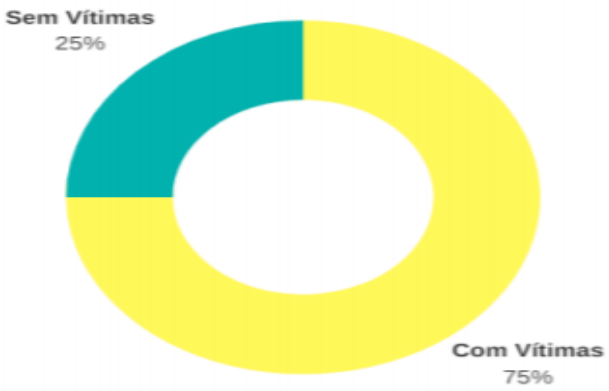
Fonte: elaborada pelos autores

Figura 13: Acidentes causados por desobediência a sinalização



Fonte: elaborada pelos autores

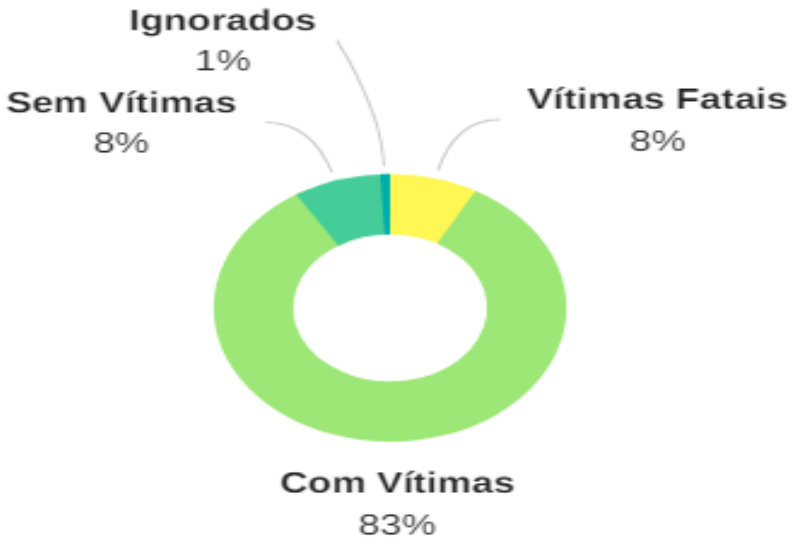
Figura 14: Acidentes causados por ingestão de bebida alcóolica.



Fonte: elaborada pelos autores

A Fig. 15 apresenta um infográfico relacionando os acidentes em pistas simples com a situação das vítimas (fatais – óbitos, com vítimas – feridos, sem vítimas - sem lesões, ignorados – não registrado com ocorrência de qualquer tipo de vítima).

Figura 15: Acidentes em pista simples com situação das vítimas.



Fonte: elaborada pelos autores

Com a visualização das informações em formato de infográficos, pode-se perceber um padrão nos acidentes ocorridos dentre as melhores regras geradas na mineração:

- Dos 31 acidentes ocorridos aos sábados, 29 tiveram vítimas feridas. Dentre estes, 24 foram em pistas do tipo reta, sendo 19 destes em pista simples;
- Dos 24 acidentes ocorridos na sexta, 22 foram em retas e 18 destes resultaram em vítimas feridas;
- Dos 22 acidentes ocorridos na quinta, 21 foram em reta e, dentre estes, 16 acidentes deixaram vítimas feridas;
- Todos os acidentes ocorridos em pista múltipla, num total de 18, deixaram vítimas feridas. Destes, 15 deles aconteceram em retas;
- Apenas 16 acidentes ocorridos em pista simples, sem vítimas.

6. ANÁLISE DOS RESULTADOS

A análise das informações através dos infográficos e das regras geradas, possibilita identificar que a maioria dos acidentes ocorreram próximo e durante o fim de semana (quinta, sexta, sábado e domingo). A quantidade significativa dos acidentes ocorre em "Retas" com "Vítimas Feridas", sendo a "Ultrapassagem Indevida" a causa de acidentes que mais resulta em vítimas fatais. É possível identificar na Fig. 16, que apresenta a planilha todos os acidentes que resultaram em vítimas fatais ocorreram em pistas simples.

Fig. 16: Planilha de todos os acidentes que resultaram em "Vítimas Fatais".

	A	B	C	D	E	F
16	DOMINGO	64.8	FALTA DE ATENCAO	COM VITIMAS FATAIS	SIMPLES	RETA
94	SABADO	62.4	ULTRAPASSAGEM INDEVIDA	COM VITIMAS FATAIS	SIMPLES	CURVA
121	SEGUNDA	51.8	DESOBEDIENCIA DA SINALIZACAO	COM VITIMAS FATAIS	SIMPLES	RETA
123	TERCA	42.9	ULTRAPASSAGEM INDEVIDA	COM VITIMAS FATAIS	SIMPLES	RETA
127	QUARTA	64	OUTRAS	COM VITIMAS FATAIS	SIMPLES	CURVA
132	DOMINGO	89.7	OUTRAS	COM VITIMAS FATAIS	SIMPLES	RETA
140	TERCA	68	OUTRAS	COM VITIMAS FATAIS	SIMPLES	RETA
143	SEXTA	76	ULTRAPASSAGEM INDEVIDA	COM VITIMAS FATAIS	SIMPLES	RETA
146	QUINTA	67	OUTRAS	COM VITIMAS FATAIS	SIMPLES	RETA

Fonte -Próprio autor, 2018

Como indicado na Fig.6, o trecho crítico da cidade de São Mateus, encontra-se entre os quilômetros 65 a 68 sentido Norte da BR 101. O trecho em epígrafe está localizado entre os bairros Cacique e Sernamby do município, como mostra a Fig. 17, a qual mostra que o trecho é composto principalmente por pista reta, possuindo algumas curvas e cruzamentos.

Figura 17:Trecho crítico da cidade de São Mateus, compreendido entre os Km 65 e Km 68

Fonte - Google Maps, 2018.

A Fig.18 apresenta o ponto mais crítico (Km 66 Norte), obtido através de imagens de satélites. Neste local ocorreram 17 acidentes no ano de 2016, o trecho é uma reta dividida em dois fluxos opostos, com sinalização vertical e horizontal, possui faixa dupla contínua, o que indica a proibição de ultrapassagem (DETRAN, 2017).

Figura 18: Foto de Satélite do Km 66 Norte, São Mateus-ES

Fonte – *GoogleMaps*, 2018

Utiliza-se o Calc, novamente, com o objetivo de gerar uma planilha com os acidentes ocorridos nos trechos críticos indicados. A Fig. 19 apresenta a planilha com os dados que comprovam que as regras encontradas estão corretas, pois indicam que em pistas retas os condutores são mais desatenciosos.

Fig.19. Planilha com os acidentes entre os Km 65 e Km 67

	A	B	C	D	E	F
1	DIA_SEMANA	KM	CAUSA_ACIDENTE	CLASSIFICACAO_ACIDENTE	TIPO_PISTA	TRACADO_VIA
5	DOMINGO	65	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	RETA
6	QUINTA	66	FALTA DE ATENCAO	SEM VITIMAS	SIMPLES	RETA
10	SABADO	66	DESOBEDIENCIA DA SINALIZACAO	COM VITIMAS FERIDAS	SIMPLES	RETA
15	QUARTA	65	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
19	QUARTA	66	NAO GUARDAR DISTANCIA DE SEGURANCA	COM VITIMAS FERIDAS	SIMPLES	RETA
22	DOMINGO	66	OUTRAS	COM VITIMAS FERIDAS	MULTIPLA	RETA
24	SABADO	67	FALTA DE ATENCAO	COM VITIMAS FERIDAS	MULTIPLA	RETA
26	SABADO	67	DESOBEDIENCIA DA SINALIZACAO	COM VITIMAS FERIDAS	DUPLA	CRUZAMENTO
29	QUINTA	66	FALTA DE ATENCAO	COM VITIMAS FERIDAS	DUPLA	RETA
31	SABADO	66	OUTRAS	COM VITIMAS FERIDAS	DUPLA	CRUZAMENTO
35	SABADO	65	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
41	DOMINGO	67	OUTRAS	COM VITIMAS FERIDAS	MULTIPLA	RETA
42	QUINTA	66	VELOCIDADE INCOMPATIVEL	SEM VITIMAS	SIMPLES	RETA
43	TERCA	66	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	RETA
52	QUINTA	65	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	RETA
53	SEXTA	65	ANIMAIS NA PISTA	COM VITIMAS FERIDAS	SIMPLES	RETA
54	SEXTA	66	DESOBEDIENCIA DA SINALIZACAO	COM VITIMAS FERIDAS	SIMPLES	RETA
58	DOMINGO	66	FALTA DE ATENCAO	COM VITIMAS FERIDAS	DUPLA	RETA
61	SEGUNDA	66	NAO GUARDAR DISTANCIA DE SEGURANCA	COM VITIMAS FERIDAS	MULTIPLA	RETA
65	SABADO	66	INGESTAO DE ALCOOL	SEM VITIMAS	MULTIPLA	RETA
72	SEGUNDA	67	FALTA DE ATENCAO	COM VITIMAS FERIDAS	SIMPLES	RETA
77	SEXTA	66	OUTRAS	COM VITIMAS FERIDAS	SIMPLES	RETA
86	DOMINGO	65	FALTA DE ATENCAO	COM VITIMAS FERIDAS	DUPLA	RETA
89	QUARTA	65	FALTA DE ATENCAO	COM VITIMAS FERIDAS	DUPLA	CRUZAMENTO
90	TERCA	65	FALTA DE ATENCAO	COM VITIMAS FERIDAS	DUPLA	RETA
91	DOMINGO	66	INGESTAO DE ALCOOL	COM VITIMAS FERIDAS	DUPLA	RETA

Fonte - Autores, 2018.

CONCLUSÕES

A aplicação do processo de descoberta de conhecimento em base de dados no conjunto de dados dos acidentes ocorridos no ano de 2016, na BR 101, no trecho correspondente ao km 55 até km 90 Norte, do município de São Mateus-ES, proporcionou o entendimento de cada fase do KDD.

Na fase de seleção e pré-processamento, utiliza-se o Calc para identificar os atributos relevantes, realizar a limpeza e adequação inicial do conjunto de dados selecionado da base de dados da PRF para a fase de Mineração de dados, sendo a que a finalização desta etapa e obtida através do uso do Weka.

Na fase de mineração de dados implementou-se o Algoritmo Apriori com auxílio da ferramenta Weka. Este algoritmo apresentou-se muito eficiente na geração das regras de associação, pois a sua utilização resultou em regras coerentes e significativas como a maior frequência de ocorrência de acidentes, quais as principais causas dos acidentes entre outras.

Na fase de pós-processamento e avaliação das informações, apresenta-se o resultado da mineração de dados, isto é, transformar as regras de associações sem conhecimento útil no formato de infográficos e de mapas com auxílio da Plataforma Canva e Google Maps, o que possibilitou identificar na rodovia BR 101, no trecho em análise, quais são pontos críticos.
A descoberta de conhecimento na base de dados aberta da PRF no trecho em epígrafe indicou que o trecho crítico é composto por segmento contínuo, pista reta, com traçado simples, possui sinalização vertical e horizontal, faixa dupla contínua o que indica proibição da ultrapassagem, além de que os dias próximos e durante o final de semana (quinta, sexta, sábado e domingo), são os que têm maior ocorrência de acidentes nos locais identificados.

REFERÊNCIAS

AGRAWAL, R., SHAFER, J. C. Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering, vol. 8, NO. 6, December 1996.

BIT. BANCO DE INFORMAÇÕES E MAPAS DE TRANSPORTES. Disponível em: <http://www2.transportes.gov.br/bit/01-inicial/index.html>. Acesso em 27 de dez de 2017.

BRANCO, A. M., Segurança rodoviária. São Paulo: CL-A Cultural, 1999. 108p.

BRASIL. Lei de Acesso a Informação – LAI (Lei 12527/2011). Disponível em: <http://www2.camara.leg.br/transparencia/acesso-a-informacao>. Acesso em 02 de julde 2017.

BRASIL. Portal Brasileiro de Dados Aberto. Disponível em <http://dados.gov.br/>. Acesso em 20 de jun de 2017.

CANVA. Disponível em https://www.canva.com/pt_br/. Acesso em 01 de dez de 2017.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. Knowledge Discovery and Data Mining, Menlo Park: AAAI Press, 1996 ª.

DETRAN. Disponível em: <http://www.detran.se.gov.br/educ_sinalizacao_horizontal.asp>, acessado em 23 de dez de 2017.

GALVÃO, N. D.; MARIN, H. F. Características das Vítimas de acidentes de trânsito por meio da técnica de Mineração de Dados. Journal of Health Informatics. v. 2, n.4, p 102-107. 2010. ISSN 2175-4411.

GOLDSCHMIDT,L R.; PASSOS, E. Data Mining: Um guia prático. Rio de Janeiro: Elsevier, 2005. ISBN 85-3521877-7.

GEUTERS, K.; WETS, G. Black Spot Analysis Methods: Literature Review. Diepenbeek(Belgium): Centre for Traffic Safety Upward Mobility, 2003.

GOOGLE. Google Maps. Disponível em: <https://www.google.com/maps/about/>. Acesso em 14 de dez de 2017.

HAYKIN, S. Neural Networks: a Comprehensive Foundation. Prentice Hall, 1999.

LIBREOFFICE. Calc. Disponível em: <https://www.libreoffice.org>. Acesso em 01 de dez de 2017.

NIC. Núcleo de Informação e Coordenação do Ponto BR. Manual de Dados Aberto.2011. Disponível em<http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf.> Acesso em 20/jun/2017.

PRF. POLÍCIA RODOVIÁRIA FEDERAL. Dados abertos. Disponível em: <https://www.prf.gov.br/portal/dados-abertos/>. Acesso em: 20 de Jun de 2017.

SHIKIDA, C. Dj.; CASTRO, G.; ARAUJO JR., A. F. Economic Determinants of Driver's Behavior in Minas Gerais. Economics Bulletin, [S. l.], v. 8, n. 10, p. 1-7, 2008.

SOUZA, M. N. V. Comparação de Algoritmos de Aprendizado de Máquina Aplicados na Mineração de Dados Educacionais. Universidade Federal Rural de Pernambuco, Recife- Brasil, 2015

WEKA. Waikato Environment for Knowledge Analysis - Information. Disponível em <weka.associations.Apriori> Universidade de Waikato, 2016.