CENTRO UNIVERSITÁRIO DE ANÁPOLIS – UNIEVANGÉLICA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

ALLAN GONÇALVES DA CRUZ
ELSON BENTO DOS SANTOS

APLICAÇÃO DA MINERAÇÃO DE DADOS PARA MITIGAÇÃO DOS ACIDENTES DE TRÂNSITO NO MUNICÍPIO DE GOIÂNIA

ANÁPOLIS-GO
2018-1

ALLAN GONÇALVES DA CRUZ
ELSON BENTO DOS SANTOS

APLICAÇÃO DA MINERAÇÃO DE DADOS PARA MITIGAÇÃO DOS ACIDENTES DE TRÂNSITO NO MUNICÍPIO DE GOIÂNIA

Projeto de Pesquisa apresentado ao Curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – Uni EVANGÉLICA, sob orientação da Profa. Aline Dayany de Lemos.

Anápolis, GO, ____ de _____ de 2018.

BANCA EXAMINADORA

.....
Profa. Aline Dayany de Lemos
Orientadora

.....

Prof ( ). .....
Convidado

.....

Prof ( ). .....
Convidado

Nota: .....

ANÁPOLIS-GO
2018-1

RESUMO

A evolução da tecnologia tornou possível a produção de expressiva quantidade de informações que vem sendo armazenadas pelos banco de dados. Com o uso de ferramentas adequadas para análise, é possível encontrar nos dados armazenados informações úteis que não são perceptíveis em sua forma natural. Descobertas que podem auxiliar na tomada de decisões quando se busca por soluções para diversos problemas encontrados na sociedade, bem como apontar melhorias em diversos setores como o *marketing*, o comércio, a saúde, entre outros. O objetivo deste trabalho é identificar, através das técnicas de mineração de dados, novos conhecimentos sobre as causas dos acidentes de trânsito no município de Goiânia, no período de 2011 a 2017, a partir dos dados armazenados no banco da Secretaria de Segurança Pública do Estado de Goiás, que possam ser úteis na mitigação desses acidentes. Para auxiliar na pesquisa será utilizada a ferramenta *Weka*. Como resultado espera-se obter novas informações a respeito do número de acidentes, os períodos em que mais ocorrem, avaliar o número de vítimas e a faixa etária nos dados coletados para o trabalho. Com base nos resultados, foram encontradas associações que podem ser utilizadas nas ações de mitigação dos acidentes de trânsito e, se possível, apontar soluções para o problema.

Palavras-chave: Acidentes de trânsito, Mineração de dados, Banco de dados, Ferramenta *Weka*.

ABSTRACT

The evolution of technology has made it possible to produce an expressive amount of information that is being stored by the database. With the use of appropriate tools for analysis, it is possible to find in the stored data useful information that is not perceptible in its natural form. Discoveries that can aid in decision making when searching for solutions to various problems found in society, as well as pointing improvements in several sectors such as marketing, commerce, health, among others. The objective of this work is to identify, through the techniques of data mining, new knowledge about the causes of traffic accidents in the city of Goiânia, from 2011 to 2017, from the data stored in the bank of the State Public Security Secretariat of Goiás, that may be useful in mitigating these accidents. To assist in the research, the Weka tool will be used. As a result, it is expected to obtain new information regarding the number of accidents, the periods in which they occur most, and to assess the number of victims and the age group in the data collected for the work. Based on the results, associations were found that can be used in mitigation actions of traffic accidents and, if possible, to point out solutions to the problem.

Keywords: Traffic accidents, Data mining, Database, Weka tool.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CBM-GO	Corpo de Bombeiros Militar do Estado de Goiás
CTB	Código de Trânsito Brasileiro
*DDL	*Data Definition Language*
DENATRAN	Departamento Nacional de Trânsito
DETRAN-GO	Departamento Estadual de Trânsito de Goiás
*DML	*Data Manipulation Language*
DNIT	Departamento Nacional de Infraestrutura de Transportes
DPVAT	Seguro de Danos Pessoais Causados por Veículos Automotores de Vias Terrestres
HUGO	Hospital de Urgências de Goiânia
*KDD	*Knowledge Discovery in Databases*
OMS	Organização Mundial de Saúde
ONSV	Observatório Nacional de Segurança Viária
PC-GO	Polícia Civil do Estado de Goiás
PM-GO	Polícia Militar do Estado de Goiás
PRF	Polícia Rodoviária Federal
RAI	Registro de Atendimento Integrado
SES-GO	Secretaria Estadual de Saúde de Goiás
SGBD	Sistema Gerenciador de Banco de Dados
SSP-GO	Secretaria de Segurança Pública do Estado de Goiás
*WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Os acidentes de trânsito constituem um agravo à vida e saúde das pessoas. Este é um problema considerável dado o número de lesões e óbitos sofridos pelas vítimas do trânsito.

No ano de 2013, cerca de 41 mil pessoas foram mortas por causa de acidentes nas vias brasileiras. No período de 2009 a 2013 o número de acidentes de trânsito no Brasil teve um aumento de 19 por 100 mil habitantes para 23,4 por 100 mil habitantes. Este foi o maior registro da América do Sul, segundo informações de um relatório divulgado pela Organização das Nações Unidas no Brasil (ONUBR, 2015).

Em Goiânia, de acordo informações disponíveis no site do Hospital de Urgências (HUGO, 2017), Tabela 1, o número de vítimas atendidas que sofreram acidentes de carro foi de 766 no ano de 2012, 712 em 2013, 937 em 2014, 817 em 2015 e 716 em 2016. O que mostra cerca de 789 atendimento por ano, ou 65 pessoas atendidas a cada mês, em média, relacionadas somente aos acidentes de carro, sem contar os acidentes de motos, atropelamentos, ciclistas, etc.

Tabela 1 - Estatísticas HUGO, atendimentos às vítimas de trânsito.

HUGO – QUANTITATIVO DE ATEND. ÀS VÍT. DE ACIDENTES DE TRÂNSITO					
Acidentes de Trânsito	**2012**	**2013**	**2014**	**2015**	**2016**
Acid. carro – Estrada/Rodovia -BR	373	96	98	62	61
Acid. carro – Estrada/Rodovia -GO	110	294	256	104	75
Acid. carro – Goiânia	766	712	937	817	716
Acid. moto – em serviço	1.260	1.449	1.530	1.473	1.180
Acid. moto – no trajeto p/ o serviço	833	2.096	2.176	1.957	1.375
Acid. moto – passeio/lazer	3.399	2.708	2.854	2.006	1.803
Atropelamento	1.077	1.123	1.212	966	680
Bicicleta	585	589	620	528	385
Caminhão	94	63	86	68	31
Ônibus	58	70	54	72	36
Total	8.874	9.638	10.485	8.490	6.574
MÉDIA MENSAL	**740**	**803**	**874**	**708**	**548**
MÉDIA DIÁRIA	**25**	**27**	**29**	**23**	**18**

Fonte: HUGO, 2017.

Como visto na Tabela 1 o número das vítimas de acidentes de carro teve crescimento até o ano de 2014, com uma leve redução para os anos de 2015 e

2016. Queda que reflete os resultados de uma ação criada em 2014, coordenada entre o poder público e a sociedade civil, denominada "Maio Amarelo", que tem como objetivo mobilizar a sociedade e discutir a questão dos problemas no trânsito. Porém, apesar de ações visando diminuir esse quantitativo, os danos causados no trânsito ainda mostram números elevados, produzindo uma média mensal de 740 acidentes no ano de 2012, 803 em 2013, 874 em 2014, 708 em 2015 e 548 em 2016, o que faz dos acidentes de trânsito, ainda, um grande desafio para a sociedade goianiense.

O desenvolvimento tecnológico possibilitou avanços em muitos setores da sociedade: na medicina, na ciência, na educação, etc. Avanços como o armazenamento de informações, que nas últimas décadas, vem sendo organizadas sistematicamente nos grandes bancos de dados. Essas informações podem ser coletadas de várias fontes como computadores, celulares, *tablets*, na utilização dos *softwares*, nas páginas de redes sociais, em canais de vídeo, e muitos outros recursos. Esses dados por si só não apresentam significado relevante, porém quando interpretados corretamente geram informações de valor, que podem produzir a descoberta dos mais variados tipos de conhecimentos, como preferências dos clientes, tendências sociais, econômicas, políticas, etc. Dessa forma, com a utilização de ferramentas e processos especiais para "garimpar" essas informações, pode-se obter novos resultados que virão a ser úteis na solução de problemas ou mesmo previsão de futuras convergências. Através da mineração de dados, com seus algoritmos e técnicas, a lapidação e análise desses dados podem construir novos conhecimentos muitas vezes não percebidos pela simples observação.

Em Goiás, a Secretaria de Segurança Pública tem investido em equipamentos tecnológicos, *softwares*, bem como capacitado pessoas para implementação de programas que tornem possível o registro do maior número de informações sobre os serviços públicos prestados. No ano de 2016 foi criado o sistema de Registro de Atendimento Integrado (RAI), em substituição ao o Sistema Integrado de Atendimento a Emergências (SIAE). Ambos fazem parte de uma plataforma que agrega vários outros sistemas do setor de segurança pública. Dentre esses, destaca-se o M.Portal que foi criado em 2013 e nasceu com o intuito de integrar informações de diversos banco de dados confiáveis que permitem trazer dados sobre pessoas e veículos como por exemplo registros cíveis, criminais, veicular e autuações, e também o Sistema Geográfico de Informação (GisGestão) que é um

software de análise criminal e geoprocessamento, disponibilizando, em tempo real, todas as análises dos crimes considerados de alta prioridade e os convertidos em metas de redução pela Secretaria de Segurança Pública do Estado de Goiás (SSP-GO), para a elaboração das estratégias policiais, planejamento de emprego operacional e investigativo, entre outros.

Os dados coletados por esses sistemas podem ser manipulados, utilizando a mineração, para construir resultados que ajudem na compreensão das causas e dos efeitos dos acidentes de trânsito, tanto em Goiânia, como em outras cidades que existe serviço prestado por Unidades da Segurança Pública.

Diante disso, sabendo que o banco da Secretaria de Segurança Pública do Estado de Goiás (SSP-GO) dispõe de diversas informações sobre ocorrências de acidentes de trânsito, existem novos conhecimentos que podem ser obtidos a partir da análise por mineração de dados (MD) e vir a ser úteis aos gestores das políticas públicas na tomada de decisões para mitigar os acidentes de trânsito no município de Goiânia?

Dessa forma, torna-se o objetivo geral deste trabalho identificar informações específicas que possam apoiar as ações criadas para redução do número de vítimas do trânsito na capital de Goiás, construídas a partir dos dados armazenados no sistema para Registro de Atendimento Integrados (RAI), da SSP-GO. E para que esse objetivo se torne possível, definiu-se as seguintes metas:

- Obter o banco de dados, disponibilizado para fins de estudo, através de solicitação por ofício, junto à SSP-GO;
- Conhecer a base de dados disponibilizada;
- Conhecer os tipos de técnicas e tarefas de mineração de dados;
- Identificar a(s) melhor(es) técnica(s) e tarefa(s) para a base em questão;
- Aplicar a técnica de mineração de dados no banco, visando a identificação de regras de associação sobre os acidentes de trânsito;
- Examinar os resultados obtidos e destacar as melhores regras de associação identificadas que podem ser úteis nas tomadas de decisões, por parte dos gestores, relacionadas políticas públicas de mitigação dos acidentes de trânsito no município de Goiânia, como por exemplo no desenvolvimento de ações de propaganda direcionada aos motoristas segundo uma faixa etária na qual existem mais vítimas do trânsito.

Este trabalho se justifica pelo fato de que uma das maiores preocupações por parte das autoridades e gestores das cidades é minimizar o grande número de acidentes de trânsito que, em consequência do aumento da quantidade de veículos circulando pelas ruas, tendem a ocorrer em números assustadores. Esses acidentes vêm provocando danos a vida das pessoas como sequelas pós-traumáticas e prejuízos financeiros, tanto às vítimas, quando ao Estado. Segundo informações do Hospital de Urgências de Goiânia (HUGO, 2017) das 13.083 cirurgias que a instituição realizou em 2016, 5.857 foram relacionadas a acidentes de trânsito, o que equivale a 45% do total dos procedimentos cirúrgicos. Fato que evidencia a violência do trânsito nas ruas de Goiânia, bem como o gasto público dispensando com suas vítimas.

.

2 FUNDAMENTAÇÃO TEÓRICA

Como fundamentação teórica serão abordados assuntos acerca da tipificação dos acidentes de trânsito, sua a definição, seus índices de ocorrência, principalmente aqueles que trazem como consequência os óbitos no Brasil, em Goiás e na cidade de Goiânia. Nessa oportunidade, ainda serão apresentados conceitos de banco de dados e, por fim, a definição de mineração de dados, bem como suas tarefas e técnicas.

2.1 TIPIFICAÇÃO DOS ACIDENTES DE TRÂNSITO

Segundo o dicionário *MICHAELIS* (*2017*) um acidente é casual, fortuito, imprevisto, é um acontecimento infausto que envolve dado, estrago, sofrimento ou morte. Esses fatos se dão em razão das causas chamadas acidentais, como o trânsito, trabalho, quedas, envenenamentos, afogamentos, entre outros.

O acidente de trânsito é todo evento danoso que envolva o veículo, a via, o homem e/ou animais, contendo no mínimo dois desses fatores. (TRANSITOBR, 2017).

De acordo com o DNIT (2017) o acidente de trânsito é uma ocorrência que afeta diretamente o cidadão, porquanto a esse são impingidos aspectos relacionados com a morte, com a incapacitação física, perdas materiais, podendo provocar sérios comprometimentos de cunho psicológico, muitas vezes de difícil superação. Sendo classificado quanto a gravidade em acidente com morto, com ferido ou sem vítima.

Para fins de pesquisa neste trabalho deve-se entender por vítima qualquer pessoa que esteja envolvida em acidente de trânsito e tenha sofrido algum prejuízo, seja físico, psicológico ou financeiro.

Como descrição e classificação dos acidentes o Departamento Nacional de Trânsito (DENATRAN, 2010) adota os seguintes tipos: colisão, abalroamento, tombamento, capotagem, atropelamento e choque com objeto fixo.

Informações da Polícia Rodoviária Federal (PRF, 2017), apontam que entre as principais causas dos acidentes com mortes ocorridos em 2016 estão a falta de atenção (30,8% dos óbitos registrados); velocidade incompatível (21,9%); ingestão

de álcool (15,6%); desobediência à sinalização (10%); ultrapassagens indevidas (9,3%); e sono (6,7%).

A Organização Mundial de Saúde (OMS, 2004) registra que o Brasil está entre os países com os maiores índices de vítimas de acidentes de trânsito, com taxas que no ano de 2002 chegaram a 219,5 pessoas feridas por 100 mil habitantes e 18,7 mortes por 100 mil habitantes. O "Portal Estatísticas", lançado pelo Observatório Nacional de Segurança Viária (ONSV, 2015), informa que ocorreram 1.864 mortes tendo como causa os acidentes de trânsito. Fontes de dados do DETRAN-GO (2017) apontam que no período entre 2002 e 2012, na cidade de Goiânia, o número de pessoas mortas por acidentes de trânsito variou de 242 a 310, respectivamente. Números que mostram um crescimento na quantidade de vidas perdidas.

Somado ao número de mortes encontra-se o impacto econômico para o país em relação aos gastos com resgate, hospitalização, indenizações, tratamento de recuperação, perda de veículos e reparo de infraestrutura. Esses gastos, no Brasil, geraram um custo de mais de R$ 56 bilhões com acidentes de trânsito em todo país no ano de 2014 (ONSV, 2015). O Seguro de Danos Pessoais Causados por Veículos Automotores de Vias Terrestres DPVAT indenizou 42.500 pessoas por morte e 515.750 por invalidez em 2015. Em Goiás o valor chegou a R$ 2,7 bilhões no mesmo período (OPOPULAR, 2017), com gastos R$ 930 milhões na assistência à saúde e na concessão de benefícios previdenciários às vítimas do trânsito, segundo a Secretaria Estadual de Saúde de Goiás (SES-GO, 2017), tornando a maior parte dos gastos voltada ao fator humano.

Buscando implantar medidas eficazes no combate a violência de trânsito e para regulamentar o setor, visando oferecer maior segurança a pedestres, motoristas e passageiros, o Código de Trânsito Brasileiro (CTB) foi criado em 1997, através da Lei 9.503. Com ele o cinto de segurança se tornou obrigatório, as faixas de pedestres aumentaram a segurança na travessia das vias, e foi ampliado o rigor com relação a Lei Seca para punir o condutor embriagado.

Assim como a criação do CTB, várias ações vêm sendo tomadas por parte do poder público na tentativa de prevenir, conscientizar as pessoas e frear o crescimento dos acidentes. Soluções que vão desde a educação de trânsito nas escolas, melhorias na formação dos motoristas, campanhas impactantes, investimento no transporte coletivo, até melhoria nas estruturas das vias. O Maio

Amarelo (mês de luta contra acidentes de trânsito), que aborda os temas como o uso de celular ao volante, ingerir bebidas alcoólicas antes de dirigir, uso do cinto de segurança, respeito no trânsito, excesso de velocidade, é um exemplo de ação que o governo vem usando nesse sentido.

Dessa forma, diante dos enormes prejuízos que os acidentes de trânsito vêm causando às pessoas, às famílias e à sociedade, torna-se necessário um maior envolvimento de todas as partes, tanto do poder público, quanto do cidadão. Apontar novas soluções, através da descoberta de conhecimentos que possam auxiliar na redução do número de vítimas do trânsito torna-se uma responsabilidade individual de todos.

2.2 BANCO DE DADOS

Por volta de 1970, a empresa *IBM* apresentou o que seria os fundamentos de bancos de dados relacionais. Nessa época foram realizadas pesquisas sobre armazenamento de dados e foram descobertos os modelos hierárquicos, relacionais, baseado em objeto, semiestruturado.

Conforme define GUIMARÃES (2003) um banco de dados ou base de dados é uma coleção de dados ou informação que estão relacionados entre si, de forma a representar aspectos do mundo real, com significado próprio e que desejamos armazenar para utilização futura.

De acordo com ELMASRI (2011) esse aspecto do mundo real é denominado minimundo ou de universo de discurso. Um banco de dados pode ser construído e mantido manualmente ou através de um sistema computadorizado. Os sistemas são denominados Sistema Gerenciador de Banco de Dados (*SGBD – Database Management System*). Isto é, uma coleção de programas que permite aos usuários criar e manter um banco de dados.

Pode-se citar como exemplo de SGBDs o *Oralcle* e o *MySql* (lançado em 1996), ambos da empresa *Oracle;* o *SQL Serve*r da *Microsoft*, foi lançado em 1989; o *MongoDB*, baseado no conceito *NoSQL*, é da empresa *MongoDB* e foi lançado em 2009; o *PostgreSQL* da empresa *PostgreSQL Global Development,* foi lançado em 1989; *DB2* desenvolvido pela *IBM* durante os anos 70, por *Edgar Frank Codd*, este foi um dos primeiros bancos de dados; dentre outros.

2.3 MINERAÇÃO DE DADOS

A mineração de dados é um processo "sistemático, interativo e iterativo de preparação e extração de conhecimentos a partir de grandes bancos de dados". Essa ação de procurar por informações em meio a uma diversidade de dados se assemelha ao serviço dos mineiros que buscam minerais preciosos em suas escavações. Um trabalho minucioso e que se for realizado com sucesso recompensará o minerador com algo muito valioso e que não era visto ou encontrado facilmente (CASTRO, 2016).

A respeito de dado e informação SILVA (2016) diz: "o dado é um fato, um valor documentado ou um resultado de medição. Quando um sentido semântico ou um significado é atribuído aos dados, gera-se a informação." Empresas vêm utilizando a mineração de dados para interpretar dados e retirar informações que os auxiliem a melhorar suas vendas, entender as necessidades e decisões de seus clientes, descobrir e prever tendências.

A mineração de dados é parte integrante de um processo mais amplo, conhecido como descoberta de conhecimento em bases de dados (*Knouledge Discovery in Databases,* ou *KDD*) (CASTRO, 2016).

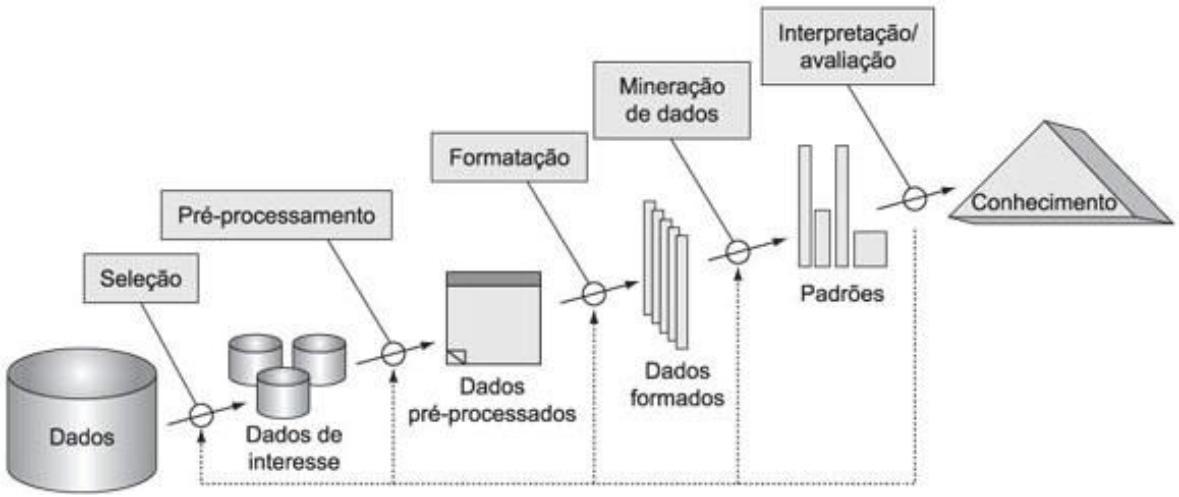
O processo KDD pode ser dividido em quatro partes:

> **"Bases de dados**: coleção organizada de dados [...]; **Preparação ou pré-processamento de dados:** são etapas anteriores à mineração que visam preparar os dados para uma análise eficiente e eficaz, essa etapa inclui a limpeza(remoção de ruídos e dados inconsistentes), a integração (combinação de dados obtidos a partir de múltiplas fontes), a seleção ou redução (escolha dos dados relevantes à análise) e a transformação (transformação ou consolidação dos dados em formatos apropriados para a mineração); **Mineração de dados**: [...] corresponde à aplicação de algoritmos capazes de extrair conhecimentos a partir dos dados pré-processados.[...] **Avaliação ou validação do conhecimento**: avaliação dos resultados da mineração objetivando identificar conhecimentos verdadeiramente úteis e não triviais." (CASTRO, 2016, p. 5 e 6).

Uma base de dados torna-se mais interessante quando deixa de servir apenas como um espaço para guardar um amontoado de dados para consultas futuras e passa produzir conhecimentos sobre os dados armazenados. As

informações encontradas através da mineração têm revolucionado a maneira como as empresas vendem, como a medicina trata seus pacientes, como os investidores fazem suas escolhas, entre outros exemplos. Através do processo *KDD* pode-se chegar a descobertas importantes e inesperadas de novos conhecimentos adquiridos na interpretação dos dados minerados. Na Figura 1 podem ser vistas todas as fases deste processo, desde a seleção dos dados até a produção do conhecimento.

Figura 1 - Fases do processo *KDD*



Fonte: FAYYAD et al., 1996

O processo *KDD*, possui cinco fases: seleção dos dados, pré-processamento, formatação ou transformação, mineração dos dados e interpretação/avaliação. A seleção dos dados é a fase onde são escolhidos os registros a serem trabalhados. Para ajudar na escolha dos atributos que irão compor a base a ser minerada é importante que se tenha uma pessoa conhecedora do domínio da aplicação. Um funcionário que possa orientar a maneira como esses dados são registrados, tendo conhecimento do contexto das ações que geram as informações no sistema. Outro objetivo desta fase é reunir os registros a serem estudados em um só banco de dados. Depois de cumprida essa etapa, os dados poderão ser submetidos à próxima fase, o pré-processamento dos dados. Nessa nova etapa, os dados passarão por um processo de limpeza, verificação de consistência, correção de erros, preenchimento ou eliminação de valores não conhecidos, redundantes ou aqueles que não fazem parte do domínio. Todo esse processo é realizado para que o

desempenho do algoritmo de análise dos dados se torne mais eficiente. Findada essa fase, inicia-se a transformação dos dados, onde os dados serão formatados, combinados de diferentes bases de dados, tendo seus formatos alterados. Nessa etapa, define-se o algoritmo que será utilizado, de acordo com o volume de dados e a finalidade da pesquisa. Existem várias técnicas disponíveis para o pré-processamento dos dados, como: suavização, que remove os valores errados dos dados; agrupamento, agrupa os valores em faixas suavizadas; generalização, converte valores específicos para valores mais genéricos (FABRICIO, 2016).

Com os dados apresentando uniformidade e organizados para o processo de mineração, aplica-se o algoritmo definido, com a utilização de uma ferramenta específica. Nesta fase, o objetivo é encontrar padrões em meio aos dados, através de técnicas de inteligência computacional. Como última etapa do processo, ocorre a análise dos resultados através da interpretação ou avaliação do que foi minerado, com o objetivo de encontrar conhecimento novo que possa ser útil. Nesse momento o que era apenas um conjunto de dados, se transforma em informação, através dos padrões evidenciados nas etapas anteriores.

2.3.1 As principais tarefas da mineração de dados

De acordo com SILVA (2016, p. 12):

> "A depender do tipo de dado à disposição e do tipo de conhecimento demandado, a mineração de dados oferece diferentes soluções e possibilidades. Assim, a área de mineração de dados é comumente dividida em tarefas, as quais ajudam a entender como situar um problema real junto aos diferentes algoritmos de análise de dados disponíveis e também a que tipo de padrão, e consequentemente a que tipo de conhecimento é possível descobrir."

“Em geral essas tarefas podem ser classificadas em dois grupos: (1) descritivas: caracterizam as propriedades gerais dos dados; e (2) preditivas: fazem inferência a partir dos dados objetivando predições.” (CASTRO, 2016).

Segundo SILVA (2016), existe ainda um segundo nível de divisão, onde no grupo das tarefas preditivas aparecem a classificação e a regressão, enquanto que

nas descritivas são inseridas as tarefas de agrupamento, sumarização, modelagem de dependências e detecção de desvio padrão. Dessa forma, com a divisão das tarefas nesses dois grupos, a escolha das ferramentas e algoritmos vem a ser mais assertiva.

2.3.1.1 Análise exploratória ou descritiva dos dados

Na tarefa de análise descritiva ocorre a organização dos dados com "o uso de ferramentas capazes de medir, explorar e descrever características intrínsecas dos dados" CASTRO (2016).

> "A estatística descritiva pode ser entendida como uma ferramenta capaz de descrever ou resumir dados, mostrando aspectos importantes do conjunto de dados, como o tipo de distribuição associada e os valores mais representativos do conjunto, permitindo criar visualizações referentes a tais aspectos". (SILVA, 2016, p. 12).

A estatística descritiva é uma das disciplinas da matemática, segundo CHAVANTE (2016) "[...] ela é composta por técnicas, cujo objetivo é descrever, analisar e interpretar dados numéricos de uma população ou amostra". Ou seja, quando se tem um conjunto de dados pode-se utilizar a análise descritiva para obter uma melhor compreensão das informações existentes no banco de dados.

2.3.1.2 Análise de grupos

A análise de grupos, também conhecida como agrupamento (*clustering*), é definida por SILVA (2016, p. 147) como "um processo pelo qual se estuda relações similares entre os exemplares, determinando como estão organizados em grupos."

"É o nome dado ao processo de separar (particionar ou segmentar) um conjunto de objetos em grupos (do inglês *clusters*) de objetos similares. [...] cada grupo formado pode ser visto como uma classe de objetos". (CASTRO, 2016, p. 9).

Durante o processo de agrupamento o algoritmo procura organizar os dados de acordo com seus atributos (características dos objetos). Dessa forma, torna-se

interessante pelo fato de reunir os objetos semelhantes em grupos de características similares, e consequentemente, afastar os objetos incompatíveis, mesmo sem ter que identificá-los nominalmente.

Existem algumas estratégias para a realização do agrupamento, conforme SILVA (2016):

- As "estratégias hierárquicas: os exemplares são divididos hierarquicamente em grupos", usando uma "abordagem *top-down*" ou "*botton-up*", conforme o modo de início do processo;

- As "estratégias por partição: partições (que representam os grupos) do espaço, no qual o conjunto de dados está inserido, são criadas de acordo com um critério de particionamento", onde pode-se aplicar o algoritmo k-médias MACQUEEN (1967);

- As estratégias de agrupamento por densidade são úteis para aplicação em conjuntos de dados com um grande número de exemplares. Um dos principais algoritmos utilizados para analisar aspectos de densidade é o *DBSCAN (Density Based Spatial Clustering of Applications with Noise).* Uma das principais características é a capacidade de trabalhar com ruídos presentes nos dados. (SILVA, 2016).

2.3.1.3 Análise Preditiva ou Classificação de Dados

A classificação pode ser definida como:

"Um processo pelo qual se determina um mapeamento capaz de indicar a qual classe pertence qualquer exemplar de um domínio sob análise, com base em um conjunto já classificado" (SILVA, 2016, p. 79). A partir do momento que se tem um conjunto de dados já classificado, o algoritmo pode "aprender", através desses dados, e classificar outros, indicando a qual das classes a informação encontrada pertence.

Já a análise preditiva dos dados está relacionada a previsão de resultados futuros, considerando os dados históricos. De acordo com SILVA (2016) a análise preditiva é "um processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos) e os rótulos a eles associados (atributo de classe)". Como

exemplo de utilização da análise preditiva são: análise de comportamento e expressão de emoções em redes sociais, na biometria, no mercado financeiro, na detecção de fraudes, no *marketing*, em operações como previsão de estoque e análise de risco pelas empresas de crédito.

2.3.1.4 Regras de associação

Os bancos de dados são caracterizados por apresentar objetos com atributos semelhantes. Uma base de dados que contém informações dos acidentes de trânsito, possui os objetos (natureza do acidente), cada um deles com diferentes atributos (local, horário, tipo de veículos envolvidos, número de vítimas, etc.).

Outro exemplo são as bases de dados transacionais Tabela 2, aquelas comuns em ambientes empresariais, que armazenam informações sobre itens movimentados em transações. Um exemplo desse tipo de banco de dados é o carrinho de supermercado, onde é registrado o valor, a quantidade, os produtos que cada cliente adquiriu. Essas informações são importantes, pois permitem uma análise futura do comportamento de seus clientes, tornando possível ações para melhorias do negócio de acordo com CASTRO (2016). A associação ocorre quando o objetivo é encontrar relações entre os atributos do banco de dados.

"Descoberta de regras de associação é o processo de analisar os relacionamentos existentes entre atributos de uma base de dados transacional, com o objetivo de encontrar associações ou correlações." (SILVA, 2016, p. 199).

Dentre as utilizações mais comuns está a área de compras, quando se procura saber quais produtos são comprados associados. Por exemplo, pais que compram fraldas para seus bebês e que acabam por levarem também cervejas nos finais de semana. Com essa informação, a rede de supermercado pode colocar cervejas próximo às fraldas (GUROVITZ, 2011).

A análise das bases de dados transacionais permite a busca de relações entre seus itens. Na Tabela 2 é possível dizer que se uma pessoa compra leite, também compra pão, sendo representado pela seguinte regra:

{Leite} -> {Pão} (“Leia-se leite implica pão”).

Tabela 2 - Exemplo de base de dados transacional

TID	Itens
1	**{Leite, pão, açúcar, café, manteiga}**
2	**{Mamão, banana, maçã}**
3	**{Leite, pão}**
4	**{Leite, pão, manteiga, banana}**

Fonte: CASTRO, 2016

2.3.2 Técnicas de mineração de dados

As técnicas de mineração de dados são os algoritmos utilizados no processo de mineração. Cada técnica ou algoritmo produz resultados de formatos diferentes e são escolhidos conforme o conhecimento que se busca na massa de dados. Nos subtópicos a seguir pode se ver algumas dessas técnicas.

2.3.2.1 Árvore de decisão

A técnica de árvore de decisão é mais utilizada na tarefa de classificação de dados e consiste na organização hierárquica, através de nós internos e nós folhas (semelhante a estrutura de dados do tipo árvore). (SILVA, 2016).

Há um interesse especial pelas árvores de decisão por utilizarem representações simbólicas e intuitivas, o que torna sua interpretação mais fácil (GONÇALVES, 2017):

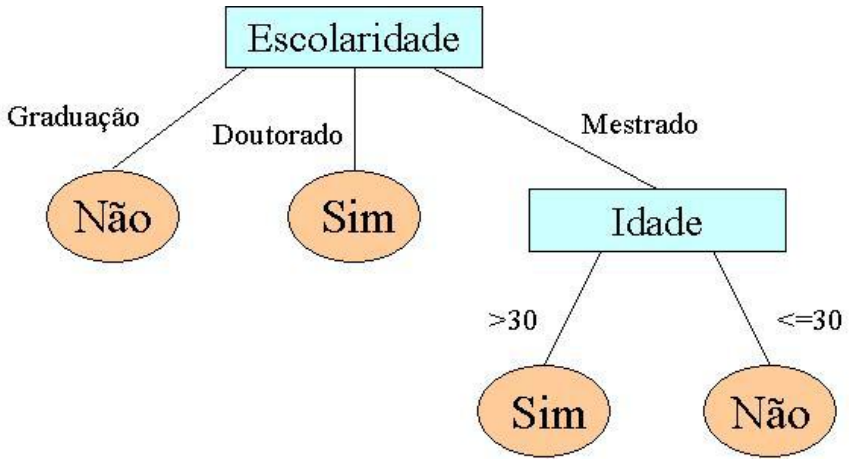
Tabela 3 - Tabela de exemplo para base de dados censitários

NOME	ESCOLARIDADE	IDADE	RICO (*atributo classe*)
Alva	Mestrado	>30	Sim
Amanda	Doutorado	<=30	Sim
Ana	Mestrado	<=30	Não
Eduardo	Doutorado	>30	Sim
Inês	Graduação	<=30	Não
Joaquim	Graduação	>30	Não
Maria	Mestrado	>30	Sim
Raphael	Mestrado	<=30	Não

Fonte: GONÇALVES, 2017.

De posse do classificador, GONÇALVES (2017) diz que “inicia-se a etapa de teste” com o objetivo de medir a acurácia, para isso é utilizada uma coleção “de dados censitários de teste” (Tabela 3), contendo observações escolhidas aleatoriamente na base de dados, diferentes das que foram escolhidas para a seleção de treinamento. “A acurácia do classificador representa a porcentagem de observações do conjunto de teste que são corretamente classificadas por ele.” O modelo de classificação será considerando eficiente quando essa porcentagem for alta, por exemplo, acima de 70%.

Figura 2 - Árvore de decisão a partir do conjunto de dados da Tabela 3



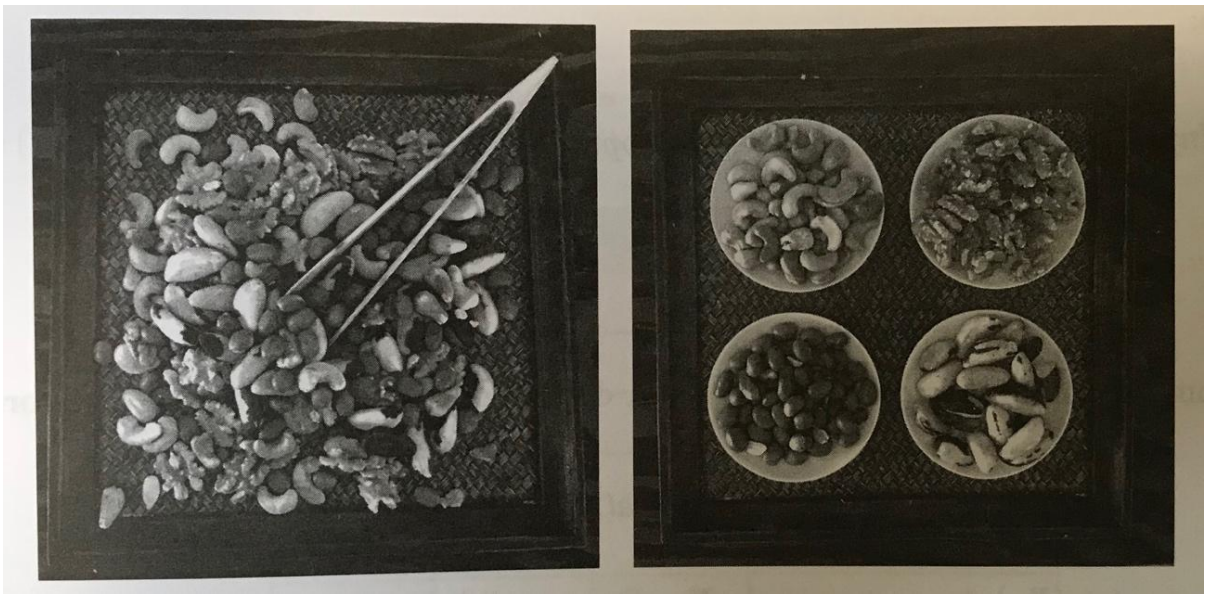
Fonte: GONÇALVES, 2016.

A Figura 2 mostra o exemplo de uma árvore de decisão gerada, a partir da Tabela 3, onde Escolaridade é o nó raiz e Idade nó filho ou galho, os "Sim” e os “Não” representam as folhas. Seguindo o caminho da árvore da raiz até a folha, a estrutura mostra que para garantir ser rico, segundo a Escolaridade, deve-se possuir doutorado. Pelo lado da Idade, para se garantir como resultado ser rio, deve-se possuir no mínimo mestrado e idade maior que 30 anos.

“A árvore de decisão é um modelo que pode ser interpretado como regras de SE ENTÃO”, onde “[…] cada caminho percorrido em profundidade na árvore gera uma regra SE ENTÃO” (SILVA, 2016). Na árvore representada na Figura 2, podemos observar que SE a Escolaridade for apenas Graduação ENTÃO Não é rico. Existe também a variedade de classes presente e em um grupo de dados, e quando ela aparece em grande quantidade define-se impureza. Quanto maior for o número de classes diferentes, maior a impureza.

Além da impureza, os dados podem aparecer desordenados. A respeito disso SILVA (2016, p. 105) define entropia como sendo “uma forma de medir a desordem em um sistema fechado”. Dessa forma, pode-se concluir que o banco de dados com uma variedade de classes muito grande, caso esteja desordenado, apresentará uma entropia alta.

Figura 3 - Entropia versus desordem



Fonte: SILVA, 2016.

De acordo com SILVA (2016) quanto maior for a entropia, maior será a desordem. Assim, é possível concluir que, quanto maior for a entropia, de uma partição de dados, mais esforço será necessário para organizá-la. Na Figura 3 está representado através de imagem um conceito de desordem e ordem, o que mostra que se os dados, mesmo quando em grande variedade, estiverem classificados ou ordenados, apresentam baixa entropia e, consequentemente, menor custo no que se refere ao processo de mineração de dados.

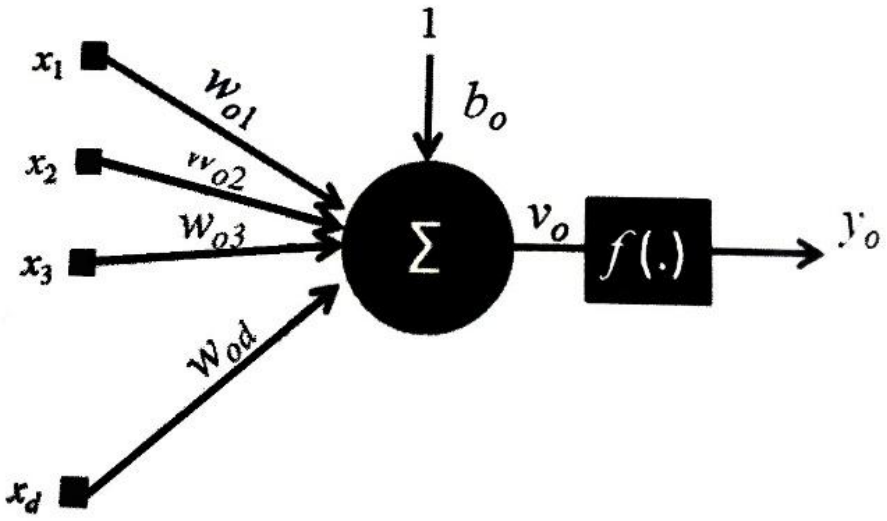
2.3.2.2 Redes Neurais

É uma técnica que se inspira no cérebro humano e em suas ligações através dos neurônios para criar um modelo matemático, com a capacidade de aprendizado, generalização, associação e abstração.

> "A simulação do funcionamento dessas unidades de processamentos - chamadas de neurônios - por meio de um computador permite, inclusive, contribuir para o entendimento dos mistérios do cérebro, embora, para os propósitos da mineração de dados, a vantagem de usar tal simulação é poder construir modelos capazes de resolver tarefas de classificação ou de agrupamento." (SILVA, 2016, p. 86).

Um **neurônio artificial** tem a finalidade de mapeia entradas e saída, o modelo *Perceptron* é capaz de realizar aprendizado de máquina e reconhecimento de padrões. Esse algoritmo de aprendizagem supervisionada considera um período de treinamento para definir uma nova entrada pertence a alguma classe selecionada ou não segundo SILVA (2016). A figura a seguir demostra a representação da estrutura de neurônio artificial do tipo *Perceptron.*

Figura 4 - Estrutura neurônio artificial do tipo Perceptron



Fonte: SILVA, 2016.

SILVA (2018) descreve a Figura 4 referenciando os sinais de entrada {x1, x2, x3, xd} são ponderados/multiplicados respectivamente por {w1, w2, w3, wd}. A função agregadora {Σ} recebe todos os sinais e realiza a soma dos produtos dos sinais. Ao resultado, é somado o limiar de ativação {b} (também chamado de *bias* ou parâmetro polarizador), soma essa conhecida como potencial de ativação {v}; o *bias* é uma constante que serve para aumentar ou diminuir a entrada líquida {v}, de forma a transladar a função de ativação no eixo de {v}. A função de ativação {f} é aplicada sobre o potencial de ativação {v} para deixar o sinal passar ou não. Todas as saídas

da rede são trocadas no início de intervalos discretos chamados de interações. No início de cada interação, a soma das entradas de cada neurônio é somada e aplicada a função de ativação. Essa função de ativação pode ser uma função bipolar (somente dois valores de saída), uma reta (função linear) ou até uma função gaussiana, hiperbólica.

O processo de treinamento tem como objetivo calibrar os pesos de modo iterativo, partindo de valores aleatórios (geralmente entre 0 e 1). A taxa de aprendizagem {η} (também um valor entre 0 e 1) diz o quão rápido a rede chega ao seu processo de classificação: um valor muito inferior causa demora a convergir, enquanto que um valor muito alto pode levar para valores fora do ajuste e nunca convergir.

Está técnica não foi utilizada no trabalho pela necessidade de conhecimento em modelos matemáticos e deveria ser montada de acordo com o problema proposto, o que demandaria mais tempo de estudo. Além disso, o processo de aprendizagem consumiria tempo para que a rede fosse treinada.

2.3.2.3 Naive Bayes

Este algoritmo é um classificador estatístico baseado no *Theorema de Bays* que trabalha com classificação de dados e apresenta bom desempenho em dados categóricos e numéricos. (SILVA, 2016). A implementação pode ser feita por ferramentas como *MALLET*, *Apache Mahout* e JLTK (LUCCA, 2013).

Uma das características do *Naive Bayes* se dá pela independência dos valores dos atributos dentro de uma classe. É uma premissa chamada de independência condicional da classe e seu principal objetivo se dá pela possibilidade de tornar os cálculos mais simples (CASTRO, 2016).

O algoritmo *Naive Bayes* não foi aplicado pelo motivo de possuir uma característica que se a variável categórica tem uma categoria (referente a amostra de dados) que não foi observada no conjunto de dados de treinamento, então o modelo, ira atribuir uma probabilidade de 0 (zero) e não será capaz de fazer uma previsão. Conforme SUNIL (2016) descreve esse contra do algoritmo *Naive Bayes* e conhecido como “*Zero Frequency*”, outro ponto relevante é a característica de suposição de preditores independentes, em cenários reais a probabilidade da existência de conjunto de indicadores que sejam completamente independentes.

2.3.2.4 Apriori

Segundo SILVA (2016) a existência de algoritmos tais como *Apriori*, que aplicam a resolução da tarefa de descoberta de regras de associação é composta de duas etapas, sendo a primeira o cálculo do suporte o qual tem o objetivo de medir os itens que frequentemente aparecem juntos, e o segundo é a confiança para a geração de regras.

> Outra característica interessante de tais algoritmos é que critérios de qualidade do resultado estão incorporados em seu processo de execução, diferentemente de outros algoritmos usados em mineração de dados, em que é necessário aplicar critérios de qualidade de avaliação de resultados após sua execução (SILVA, 2016, p. 200).

A estrutura dos dados para que o algoritmo *Apriori* seja aplicado, necessita de um conjunto de dados transacionais onde cada exemplar representa uma transação realizada, assim as transações são compostas por uma série de itens. Cada item pertence a um tipo ou conjunto de domínio previamente existente na aplicação também denominado *itemset*.(SILVA, 2016).

> A medida **suporte** quando aplicada a um *itemset* diz respeito à frequência com que os itens que o compõem aparecem juntos em transações individuais da base de dados transacionais. Essa frequência é geralmente expressa em termos percentuais [...]. Quando a medida suporte é aplicada a uma regra, ela diz respeito à frequência com que todos os itens dos dois *itemsets* envolvidos na regra aparecem juntos em transações individuais da base de dados transacional. (SILVA, 2016, p. 202).

A Tabela 4 demonstra um exemplo de transações que corresponde a produtos de um supermercado e se houve a compra do mesmo. Nesta tabela os produtos como leite, café, cerveja, entre outros, são os itemsets, e cada linha representa uma compra realizada por um cliente.

O valor do suporte é definido pelo usuário que está aplicando o algoritmo, sendo assim um valor de referência para que sejam filtrados os *itemsets* gerados após o cálculo:

Suporte = Número de registro com X e Y / Número total de registros.

Tabela 4 - Relação de produtos do mercado

N°	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	Não	Sim	Não	Sim	Sim	Não	Não
2	Sim	Não	Sim	Sim	Sim	Não	Não
3	Não	Sim	Não	Sim	Sim	Não	Não
4	Sim	Sim	Não	Sim	Sim	Não	Não
5	Não	Não	Sim	Não	Não	Não	Não
6	Não	Não	Não	Não	Sim	Não	Não
7	Não	Não	Não	Sim	Não	Não	Não
8	Não	Não	Não	Não	Não	Não	Sim
9	Não	Não	Não	Não	Não	Sim	Sim
10	Não	Não	Não	Não	Não	Sim	Não

Fonte: GRANATYR, 2018.

Para calcular o suporte serão utilizados os *itemsets* onde o valor de suporte *maior ou igual 0.3,* o que equivale a 30%. De início será aplicado para todos os produtos, e em seguida aplicado novamente para a combinações (transação) dos *itemsets* que satisfaçam a condição. Essas combinações irão ocorrer até que não haja mais nenhuma opção.

Passo 1: Calcular o suporte de conjuntos com apenas 1 item (Tabela 5).

Tabela 5 - Cálculo do suporte com um item

Itemsets	Suporte
Leite	2/10 = 0,2
Café	3/10 = 0,3
Cerveja	2/10 = 0,2
Pão	5/10 = 0,5
Manteiga	5/10 = 0,5
Arroz	2/10 = 0,2
Feijão	2/10 = 0,2

Fonte: GRANATYR, 2018.

Passo 2: Calcular o suporte de conjuntos com 2 itens (Tabela 6).

Tabela 6 - Cálculo do suporte com dois itens

Itemset	Suporte
Café, Pão	3/10 = 0,3
Café, Manteiga	3/10 = 0,3
Pão, Manteiga	4/10 = 0,3

Fonte: GRANATYR, 2018.

Passo 3: Calcular o suporte de conjuntos com 3 itens (Tabela 7).

Tabela 7 - Cálculo do suporte com três itens

Itemset	Suporte
Café, Pão, Manteiga	3/10 = 0,3

Fonte: GRANATYR, 2018.

A partir dos conjuntos de itens frequentes encontrados, torna-se necessário descobrir regras de associação com o fator de confiança maior ou igual ao definido pelo usuário que está aplicando o algoritmo.

> A medida **confiança** se aplica apenas às regras e tem o objetivo de expressar uma noção da importância e da confiabilidade de uma regra, dada a possibilidade de sua ocorrência. A medida também é geralmente expressa por meio de percentual e dada pela razão entre o suporte da regra e o suporte da premissa da regra. [...], a confiança é 65% (28/43). O significado por trás dessa porcentagem é que “em 65% das vezes em que a premissa da regra ocorre, a conclusão também ocorre”. (SILVA, 2016, p. 202).

O cálculo da confiança é baseado nos *itemsets* encontrados. As regras definidas abaixo serão demonstradas com as combinações de produtos acima de um *itemset*, além de definir um valor de confiança maior ou igual a 0.8, que equivale a 80%. As regras que não satisfazem a confiança mínima serão desconsideradas para o resultado do algoritmo *Apriori*. O cálculo para encontrar à medida que define a regra está demonstrado abaixo:

Confiança = Número de registro com X e Y / Número total de registro com X.
Por exemplo:

{café, pão}

SE café ENTÃO pão – confiança = 3 / 3 = 1.0 => 100% de confiança.

SE pão ENTÃO café – confiança = 3 / 5 = 0,6 => 60% de confiança. Então este é descartado.

{café, manteiga}

SE café ENTÃO manteiga – confiança = 3 / 3 = 1,0 => 100% de confiança.

SE manteiga ENTÃO café – confiança = 3 / 5 = 0,6 => 60% de confiança. Este também será descartado.

{pão, manteiga}
SE pão ENTÃO manteiga – confiança = 4 / 5 = 0,8 => 80% de confiança.
SE manteiga ENTÃO pão – confiança = 4 / 5 = 0,8 => 80% de confiança.

Nos exemplos acima, tem-se a confiança de 100% de certeza que quem compra café, também compra pão, porém quando se olha para o resultado invertido, o resultado é diferente, pois quem compra pão, nem sempre compra café (apenas 60% dos casos). De forma semelhante, quem compra café, sempre compra manteiga (com 100% de certeza). Já quem compra manteiga, nem sempre irá comprar café (60% dos casos). Por fim, na relação entre pão e manteiga, a confiança é a mesma (80%), tanto para quem compra pão e leva a manteiga junto, quanto para o contrário.

Os valores do suporte e da confiança são medidas que irão garantir a eficiência dos resultados obtidos pela mineração dos dados. São necessárias várias alterações no valor, tanto do suporte como da confiança, dependendo do resultado que se busca. Em alguns casos, onde a confiança está muito alta, pode ocorrer de não se encontrar nenhum resultado. A correta manipulação desses valores é que irá proporcionar a descoberta de novos conhecimentos.

2.3.3 Sistemas comerciais de mineração de dados

Na Tabela 8 são apresentados alguns exemplos de sistemas para mineração de dados, apontando seus fabricantes, suas principais funções/algoritmos, com destaque para algumas de suas vantagens. Como opção mais adequada para a realização deste trabalho foi escolhida a ferramenta *Weka* por trabalhar com o algoritmo *Apriori* e estar disponibilizada de forma *open source*. Esta é uma das ferramentas mais utilizadas em pesquisas com mineração de dados, especialmente pelos acadêmicos, nos trabalhos de conclusão de cursos de graduação, pós-graduação, mestrado e doutorado, fato comprovado pela existência de muitas referências em relação ao *Weka* nos trabalhos disponíveis na *web*.

Tabela 8 - Sistemas para mineração de dados

Nome	Fabricante	Funções	Destaque
Intelligent Miner	*IBM*	algoritmos para regras de associação, regressão, padrões sequenciais, *clustering*.	Integrado com o SGBD *DB2* da *IBM*. Grande escalabilidade dos algoritmos.
MineSet	*Silicon Graphics Inc.*	algoritmos para regras de associação, classificação, análise estatística.	Um robusto conjunto de ferramentas avançadas de visualização.
DBMiner	*DBMiner Technology Inc.*	algoritmos de regras de associação, classificação, *clustering*.	*Data Mining* utilizando *OLAP*
Genamics Expression	*Genamics Developer*	alogritmos de análise de sequências.	Análise de proteínas e de sequencias de DNA
Weka	Universidade de *Waikato*,	*Apriori* em tarefas de associação.	Desenvolvida em linguagem *java*, é *Open Source*.

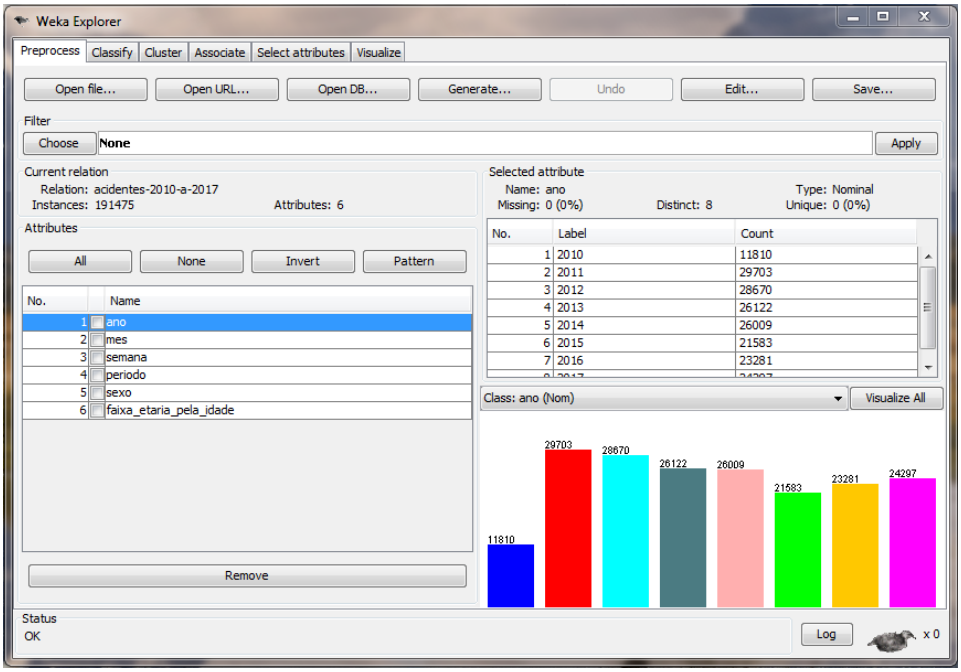
Fonte: CRUZ; SANTOS, 2017.

Os demais sistemas, diferentes do *Weka,* não foram utilizados, principalmente pelo fato de serem ferramentas comerciais pagas. Foram realizadas algumas tentativas com versões de testes, porém sem sucesso, pois quando se buscava por resultados, havia a necessidade de pagamento.

2.3.4 A Ferramenta Weka

A ferramenta *Weka (Waikato Environment for Knowledge Analysis*) versão 3.6, é uma coleção de algoritmos de aprendizado de máquina desenvolvida para executar tarefas de mineração de dados (FRANK, 2016). O *software* é *open source* e foi desenvolvido na linguagem *JAVA* por um grupo de pesquisadores da universidade de *Wikato* na Nova Zelândia desde 1993. Tem como principal característica a portabilidade, sendo executável em vários Sistemas Operacionais como *Windows, Linux, MacOS*. O *Weka.* Possui cinco módulos diferentes, *Experimenter, KnowledgeFlow, Workbench, Simple CLI* e *Explorer* (Figura 5).

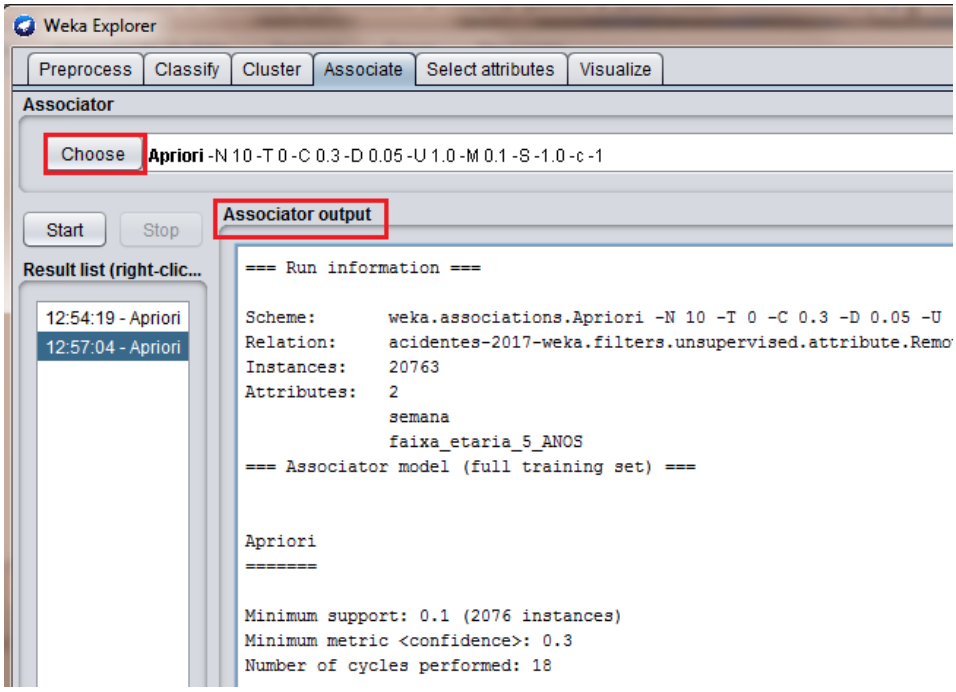
Figura 5 - Exemplo de tela Explorer da ferramenta *Weka*



Fonte: CRUZ; SANTOS, 2018.

O *software* permite realizar tarefas de Classificação (*Classify*), Agrupamento (*Cluster*), Associação (*Associate*) e Seleção de atributos (*Select atributes*). O *Weka* possibilita escolher o algoritmo, na parte de *Associator output,* (Figura 6).

Figura 6 - Opções na tela do Weka

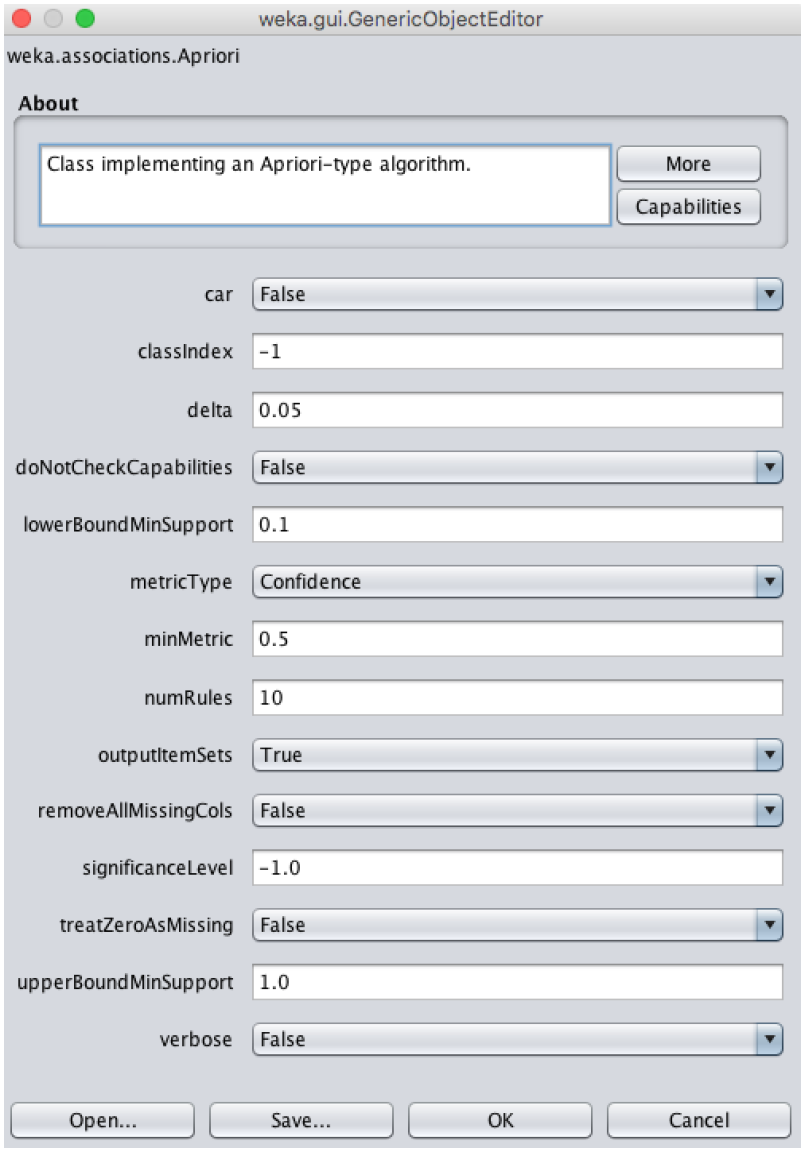


Fonte: CRUZ; SANTOS, 2018.

A configuração dos parâmetros do algoritmo é realizada na tela apresentada na Figura 7. Os parâmetros, conforme documentação do *Weka* dessa interface são descritos abaixo:

- *car*: Se habilitado, as regras de associação de classe são extraídas em vez das regras de associação (gerais).
- *classIndex*: Índice do atributo de classe. Se definido como -1, o último atributo é considerado atributo de classe.
- *delta*: Diminuir iterativamente o suporte por este fator. Reduz o suporte até que o suporte mínimo seja atingido ou o número necessário de regras tenha sido gerado.
- *lowerBoundMinSupport*: Limite inferior para suporte mínimo.
- *metricType*: Define o tipo de métrica pelo qual se pode classificar as regras.
- *minMetric*: Pontuação métrica mínima.
- *numRules*: Número de regras para encontrar.
- *outputItemSets*: Se ativada, os conjuntos de itens também são enviados.
- *removeAllMissingCols*: Remove colunas com todos os valores ausentes.
- *significanceLeve*: Nível de significância. Teste de significância (apenas métrica de confiança).
- *upperBoundMinSupport*: Limite superior para suporte mínimo. Comece a diminuir iterativamente o suporte mínimo deste valor.
- *verbose*: Se habilitado, o algoritmo será executado no modo detalhado.

Figura 7 - Tela de configuração de parâmetros do Weka



Fonte: CRUZ; SANTOS, 2018.

A tela de resultados (*Associator output*) é apresentada na Figura 8 são apresentados na Tabela 8.

Figura 8 - Exemplo de resultado na tela Associator output

```
1  === Run information ===
2
3  Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
4  Relation:     acidentes-2011-2017
5  Instances:    89780
6  Attributes:   6
7                ano
8                mes
9                semana
10               periodo
11               sexo
12               faixa_etaria_pela_idade
13 === Associator model (full training set) ===
14
15
16 Apriori
17 =======
18
19 Minimum support: 0.1 (8978 instances)
20 Minimum metric <confidence>: 0.5
21 Number of cycles performed: 18
22
23 Generated sets of large itemsets:
24
25 Size of set of large itemsets L(1): 22
26
27 Size of set of large itemsets L(2): 17
28
29 Best rules found:
30
31  1. ano=2017 16578 ==> sexo=MASCULINO 11590    <conf:(0.7)> lift:(1.06) lev:(0.01) [627] conv:(1.13)
32  2. semana=SABADO 14922 ==> sexo=MASCULINO 10183    <conf:(0.68)> lift:(1.03) lev:(0) [315] conv:(1.07)
33  3. ano=2016 13251 ==> sexo=MASCULINO 9035    <conf:(0.68)> lift:(1.03) lev:(0) [272] conv:(1.06)
34  4. periodo=NOITE 26079 ==> sexo=MASCULINO 17612    <conf:(0.68)> lift:(1.02) lev:(0) [367] conv:(1.04)
35  5. faixa_etaria_pela_idade=31 A 40 ANOS 25371 ==> sexo=MASCULINO 16961    <conf:(0.67)> lift:(1.01) lev:(0) [184] conv:(1.02)
36  6. faixa_etaria_pela_idade=41 A 50 ANOS 14103 ==> sexo=MASCULINO 9422    <conf:(0.67)> lift:(1.01) lev:(0) [96] conv:(1.02)
37  7. faixa_etaria_pela_idade=21 A 30 ANOS 29979 ==> sexo=MASCULINO 19836    <conf:(0.66)> lift:(1) lev:(0) [12] conv:(1)
38  8. periodo=TARDE 36535 ==> sexo=MASCULINO 24162    <conf:(0.66)> lift:(1) lev:(0) [3] conv:(1)
39  9. semana=SEXTA 14152 ==> sexo=MASCULINO 9295    <conf:(0.66)> lift:(0.99) lev:(-0) [-63] conv:(0.99)
40 10. ano=2012 13698 ==> sexo=MASCULINO 8978    <conf:(0.66)> lift:(0.99) lev:(-0) [-79] conv:(0.98)
```

Fonte: CRUZ; SANTOS, 2018.

Na Tabela 9 tem-se a descrição das principais linhas que compõe um resultado de mineração com a ferramenta *Weka* utilizando Regras de Associação com o algoritmo *Apriori*.

Tabela 9 - Descrição das linhas de resultados do *Weka*

Linha	Descrição
3	mostrado o algoritmo em uso, bem como um resumo dos parâmetros definidos
19	Suporte mínimo
20	Confiança
21	Número de combinações
25 e 27	Número de *itemsets* gerados
31 a 40	Melhores combinações encontradas

Fonte: CRUZ; SANTOS, 2018.

3 DESENVOLVIMENTO E ANÁLISE

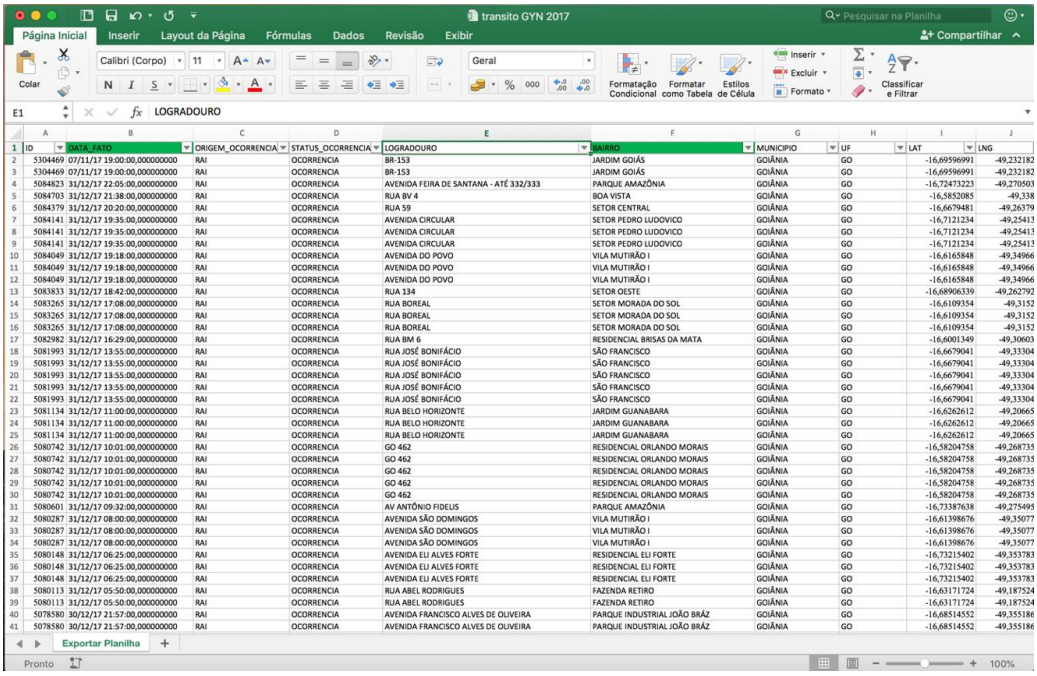
A proposta deste trabalho de pesquisa teve como cenário os dados armazenados no banco da Secretaria de Segurança Pública do Estado de Goiás. O objetivo foi produzir novos conhecimentos sobre os acidentes de trânsito no município de Goiânia, através da utilização das técnicas de mineração de dados.

Após pesquisas e testes com as tarefas e técnicas de mineração de dados citadas nos itens 2.3.1 e 2.3.2 deste trabalho, optou-se pela tarefa Regras de Associação com uso da técnica *Apriori*, por apresentarem resultados com interpretação mais coerentes com os objetivos procurados. Durante a mineração dos dados procurou-se por relacionamentos entre os diversos tipos de acidentes, os horários ou períodos em que eles vem ocorrendo, os lugares, as características dos motoristas, como sexo, idade, etc. Nas regras de associação dois conceitos tornam-se importantes, o suporte e a confiança, pois foi através da manipulação desses valores que obtivemos os melhores resultados.

3.1 SELEÇÃO DOS DADOS

De posse de uma amostra dos dados recebidos em planilha eletrônica (Figura 9), foi possível conhecer melhor a ferramenta *Weka*.

Figura 9 - Exemplo de arquivo recebido em planilha de *Microsoft excel*



ID	DATA_FATO	ORIGEM_OCORRENCIA	STATUS_OCORRENCIA	LOGRADOURO	BAIRRO	MUNICIPIO	UF	LAT	LNG
5304469	07/11/17 19:00:00,000000000	RAI	OCORRENCIA	BR-153	JARDIM GOIÁS	GOIÂNIA	GO	-16,69596991	-49,232182
5304469	07/11/17 19:00:00,000000000	RAI	OCORRENCIA	BR-153	JARDIM GOIÁS	GOIÂNIA	GO	-16,69596991	-49,232182
5084823	31/12/17 22:05:00,000000000	RAI	OCORRENCIA	AVENIDA FEIRA DE SANTANA - ATÉ 332/333	PARQUE AMAZÔNIA	GOIÂNIA	GO	-16,72473223	-49,270503
5084703	31/12/17 21:38:00,000000000	RAI	OCORRENCIA	RUA BV 4	BOA VISTA	GOIÂNIA	GO	-16,5852085	-49,338
5084379	31/12/17 20:20:00,000000000	RAI	OCORRENCIA	RUA 59	SETOR CENTRAL	GOIÂNIA	GO	-16,6679481	-49,26379
5084141	31/12/17 19:35:00,000000000	RAI	OCORRENCIA	AVENIDA CIRCULAR	SETOR PEDRO LUDOVICO	GOIÂNIA	GO	-16,7121234	-49,25413
5084141	31/12/17 19:35:00,000000000	RAI	OCORRENCIA	AVENIDA CIRCULAR	SETOR PEDRO LUDOVICO	GOIÂNIA	GO	-16,7121234	-49,25413
5084141	31/12/17 19:35:00,000000000	RAI	OCORRENCIA	AVENIDA CIRCULAR	SETOR PEDRO LUDOVICO	GOIÂNIA	GO	-16,7121234	-49,25413
5084049	31/12/17 19:18:00,000000000	RAI	OCORRENCIA	AVENIDA DO POVO	VILA MUTIRÃO I	GOIÂNIA	GO	-16,6165848	-49,34966
5084049	31/12/17 19:18:00,000000000	RAI	OCORRENCIA	AVENIDA DO POVO	VILA MUTIRÃO I	GOIÂNIA	GO	-16,6165848	-49,34966
5084049	31/12/17 19:18:00,000000000	RAI	OCORRENCIA	AVENIDA DO POVO	VILA MUTIRÃO I	GOIÂNIA	GO	-16,6165848	-49,34966
5083833	31/12/17 18:42:00,000000000	RAI	OCORRENCIA	RUA 134	SETOR OESTE	GOIÂNIA	GO	-16,68906339	-49,262792
5083265	31/12/17 17:08:00,000000000	RAI	OCORRENCIA	RUA BOREAL	SETOR MORADA DO SOL	GOIÂNIA	GO	-16,6109354	-49,3152
5083265	31/12/17 17:08:00,000000000	RAI	OCORRENCIA	RUA BOREAL	SETOR MORADA DO SOL	GOIÂNIA	GO	-16,6109354	-49,3152
5083265	31/12/17 17:08:00,000000000	RAI	OCORRENCIA	RUA BOREAL	SETOR MORADA DO SOL	GOIÂNIA	GO	-16,6109354	-49,3152
5082982	31/12/17 16:29:00,000000000	RAI	OCORRENCIA	RUA BM 6	RESIDENCIAL BRISAS DA MATA	GOIÂNIA	GO	-16,6001349	-49,30603
5081993	31/12/17 13:55:00,000000000	RAI	OCORRENCIA	RUA JOSÉ BONIFÁCIO	SÃO FRANCISCO	GOIÂNIA	GO	-16,6679041	-49,33304
5081993	31/12/17 13:55:00,000000000	RAI	OCORRENCIA	RUA JOSÉ BONIFÁCIO	SÃO FRANCISCO	GOIÂNIA	GO	-16,6679041	-49,33304
5081993	31/12/17 13:55:00,000000000	RAI	OCORRENCIA	RUA JOSÉ BONIFÁCIO	SÃO FRANCISCO	GOIÂNIA	GO	-16,6679041	-49,33304
5081993	31/12/17 13:55:00,000000000	RAI	OCORRENCIA	RUA JOSÉ BONIFÁCIO	SÃO FRANCISCO	GOIÂNIA	GO	-16,6679041	-49,33304
5081993	31/12/17 13:55:00,000000000	RAI	OCORRENCIA	RUA JOSÉ BONIFÁCIO	SÃO FRANCISCO	GOIÂNIA	GO	-16,6679041	-49,33304
5081134	31/12/17 11:00:00,000000000	RAI	OCORRENCIA	RUA BELO HORIZONTE	JARDIM GUANABARA	GOIÂNIA	GO	-16,6262612	-49,20665
5081134	31/12/17 11:00:00,000000000	RAI	OCORRENCIA	RUA BELO HORIZONTE	JARDIM GUANABARA	GOIÂNIA	GO	-16,6262612	-49,20665
5081134	31/12/17 11:00:00,000000000	RAI	OCORRENCIA	RUA BELO HORIZONTE	JARDIM GUANABARA	GOIÂNIA	GO	-16,6262612	-49,20665
5080742	31/12/17 10:01:00,000000000	RAI	OCORRENCIA	GO 462	RESIDENCIAL ORLANDO MORAIS	GOIÂNIA	GO	-16,58204758	-49,268735
5080742	31/12/17 10:01:00,000000000	RAI	OCORRENCIA	GO 462	RESIDENCIAL ORLANDO MORAIS	GOIÂNIA	GO	-16,58204758	-49,268735
5080742	31/12/17 10:01:00,000000000	RAI	OCORRENCIA	GO 462	RESIDENCIAL ORLANDO MORAIS	GOIÂNIA	GO	-16,58204758	-49,268735
5080742	31/12/17 10:01:00,000000000	RAI	OCORRENCIA	GO 462	RESIDENCIAL ORLANDO MORAIS	GOIÂNIA	GO	-16,58204758	-49,268735
5080742	31/12/17 10:01:00,000000000	RAI	OCORRENCIA	GO 462	RESIDENCIAL ORLANDO MORAIS	GOIÂNIA	GO	-16,58204758	-49,268735
5080601	31/12/17 09:32:00,000000000	RAI	OCORRENCIA	AV ANTÔNIO FIDELIS	PARQUE AMAZÔNIA	GOIÂNIA	GO	-16,73387638	-49,275495
5080287	31/12/17 08:00:00,000000000	RAI	OCORRENCIA	AVENIDA SÃO DOMINGOS	VILA MUTIRÃO I	GOIÂNIA	GO	-16,61398676	-49,35077
5080287	31/12/17 08:00:00,000000000	RAI	OCORRENCIA	AVENIDA SÃO DOMINGOS	VILA MUTIRÃO I	GOIÂNIA	GO	-16,61398676	-49,35077
5080287	31/12/17 08:00:00,000000000	RAI	OCORRENCIA	AVENIDA SÃO DOMINGOS	VILA MUTIRÃO I	GOIÂNIA	GO	-16,61398676	-49,35077
5080148	31/12/17 06:25:00,000000000	RAI	OCORRENCIA	AVENIDA ELI ALVES FORTE	RESIDENCIAL ELI FORTE	GOIÂNIA	GO	-16,73215402	-49,353783
5080148	31/12/17 06:25:00,000000000	RAI	OCORRENCIA	AVENIDA ELI ALVES FORTE	RESIDENCIAL ELI FORTE	GOIÂNIA	GO	-16,73215402	-49,353783
5080148	31/12/17 06:25:00,000000000	RAI	OCORRENCIA	AVENIDA ELI ALVES FORTE	RESIDENCIAL ELI FORTE	GOIÂNIA	GO	-16,73215402	-49,353783
5080113	31/12/17 05:50:00,000000000	RAI	OCORRENCIA	RUA ABEL RODRIGUES	FAZENDA RETIRO	GOIÂNIA	GO	-16,63171724	-49,187524
5080113	31/12/17 05:50:00,000000000	RAI	OCORRENCIA	RUA ABEL RODRIGUES	FAZENDA RETIRO	GOIÂNIA	GO	-16,63171724	-49,187524
5078580	30/12/17 21:57:00,000000000	RAI	OCORRENCIA	AVENIDA FRANCISCO ALVES DE OLIVEIRA	PARQUE INDUSTRIAL JOÃO BRÁZ	GOIÂNIA	GO	-16,68514552	-49,355186
5078580	30/12/17 21:57:00,000000000	RAI	OCORRENCIA	AVENIDA FRANCISCO ALVES DE OLIVEIRA	PARQUE INDUSTRIAL JOÃO BRÁZ	GOIÂNIA	GO	-16,68514552	-49,355186

Fonte: CRUZ; SANTOS, 2018.

Posteriormente o restante dos dados foram disponibilizados em vários arquivos, todos no formato de planilhas eletrônicas, com os seguintes nomes: siae2010.xlsx, siae2011.xlsx, siae2012.xlsx, siae2013.xlsx, siae2014.xlsx, siae2015.xlsx, siae2016.xlsx, até março de 2016 vieram do Sistema Integrado de Atendimento de Emergência de Goiás (SIAE); já os arquivos transito_GYN_2016.xlsx e transito_GYN_2017.xlsx, a partir de 2016 (abril) até 2017, foram adquiridos do sistema do RAI.

Após análise do arquivo referente ao ano de 2010, os dados foram desprezados por não contemplar os doze meses do ano, uma vez que o objetivo foi trabalhar com os anos com informações completas, restando as informações do período compreendido entre os anos de 2011 a 2017. Os registros vindos do banco de dados do SIAE apresentaram algumas inconsistências, principalmente em relação à data de nascimento ou faixa etária das vítimas com a ocorrência de vários campos nulos. Por esse motivo alguns registros precisaram ser descartados para uma melhor apresentação dos resultados. Na Tabela 10 pode ser observado o dicionário de dados das informações adquiridas no sistema SIAE.

Tabela 10 -Dicionário de dados do sistema SIAE

VARIÁVEL	DESCRIÇÃO	CATEGORIA
OCORRENCIA_ID	Código da ocorrência	Numérico
DATA_HORA_INICIO	Data e hora da ocorrência	Data
TIPO_ATENDIMENTO	Forma de registro da ocorrência	OFF-LINE, ONLINE
NATUREZA_ID	Código da natureza	Numérico
CHAMADA	Tipo de ação na ocorrência	CANCELADA, OCORRÊNCIA, RECURSOS DE ATENDIMENTO, SEM ATUAÇÃO, TROTE
ENDERECO	Endereço	TEXTO
CIDADE	Nome da cidade	TEXTO
BAIRRO	Nome do bairro	TEXTO
LATITUDE	Número de referência da latitude	Numérico
LONGITUDE	Número de referência da longitude	Numérico
SEXO	Gênero da vítima	Masculino,

		Feminino
DATA_NASCIMENTO	Data de nascimento da vítima	Data
QUALIFICACAO	Qualificação das pessoas envolvidas	Assistido, Autor, Condutor, Vítima, Pessoa Recusou Atendimento, Testemunha, Proprietário
DESCRICAO	Descrição do nível de consciência da vítima	Consciente, Inconsciente, Óbito
POSICAO_VITIMA	Posição em que se encontra a vítima na chegada à ocorrência	Deambulando, Decúbito dorsal, Decúbito L.D, Decúbito L.E, Decúbito ventral, Sentada
HOSPITAL_CLASSIFICACAO_ID	Código da capacidade de atendimento da unidade hospitalar (primário, secundário, terciário)	1, 2, 3
HOSP_NOME	Nome da unidade hospitalar	TEXTO

Fonte: CRUZ; SANTOS, 2018.

Em seguida, apresentamos na Tabela 11 a relação completa dos dados vindos do sistema RAI.

Tabela 11 - Dicionário de dados do sistema RAI

VARIÁVEL	**DESCRIÇÃO**	**CATEGORIA**
ID	Código da ocorrência	Numérico
DATA_FATO	Data e hora da ocorrência	Data
ORIGEM_OCORRENCIA	Sistema que registra a entrada dos dados da ocorrência	I9X, K9, RAI
STATUS_OCORRENCIA	Ação sobre o tipo da ocorrência	DUPLICATA, INFORMAÇÃO, OCORRÊNCIA
LOGRADOURO	Endereço da ocorrência	TEXTO
BAIRRO	Bairro da ocorrência	TEXTO
MUNICIPIO	Cidade da ocorrência	TEXTO
UF	Unidade Federativa	TEXTO

LAT	Número de referência da latitude	Numérico
LNG	Número de referência da longitude	Numérico
NATUREZA	Natureza da ocorrência	TEXTO
SEXO	Gênero da vítima	Masculino, Feminino
DATA_NASCIMENTO	Data de nascimento da vítima	Data
FAIXA ETÁRIA	Intervalo de idade das vítimas	0 A 11, 12 A 17, 18 A 24, 25 A 29, 30 A 34, 35 A 64, ACIMA DE 65, NÃO INFORMADO
QUALIFICACAO	Qualificação das pessoas envolvidas	Assistido, Autor, Condutor, Vítima, Pessoa Recusou Atendimento, Testemunha, Comunicante, Proprietário, Abordado, Envolvido
TOTALOGCS	Nível de consciência da vítima segundo a escala de Glasgow	3, 4, 5, 6, 7, 8, 9, 10, 11
SITUACAO_VITIMA_GERAL	Nível de consciência da vítima	Consciente, Inconsciente, Óbito
POSICAO_VITMA	Posição em que se encontra a vítima	Deambulando, Decúbito dorsal, Decúbito L.D, Decúbito L.E, Decúbito ventral, Sentada
TIPO_ACIDENTE	Característica do tipo de acidente	Doméstico, Trabalho
UNIDADE_SAUDE_CLASSIFICACAO	Capacitação de atendimento da unidade hospitalar	Primário, Secundário, Terciário
UNIDADE_SAUDE_REDE_PUBLICA	Informa onde o atendimento hospitalar foi realizado, rede pública(1) ou privado(0)	0, 1
UNIDADE_SAUDE	Nome da unidade hospitalar	TEXTO

Fonte: CRUZ; SANTOS, 2018.

3.2 O PRÉ-PROCESSAMENTO DOS DADOS

Após gerado o subconjunto das informações recebidas foi iniciado o pré-processamentos dos dados, onde a limpeza deve ocorrer. Foram verificadas e removidas as informações inconsistentes, como por exemplo os campos nulos, incompletos, bem como os irrelevantes (o nome da cidade, por exemplo, que não sofre variação, o nome da unidade federativa, que também será sempre o mesmo) com o objetivo de aprimorar a eficiência do algoritmo de mineração e posteriormente obter um resultado final mais eficiente.

Dos dados que foram disponibilizados do sistema SIAE, citados no tópico anterior, foram desconsideradas algumas colunas, conforme Tabela 12.

Tabela 12 - Relação dos dados desconsiderados do sistema SIAE

VARIÁVEL	**MOTIVO**
OCORRENCIA_ID	Os números das chaves das ocorrências não são necessários à pesquisa.
TIPO_ATENDIMENTO	Saber se a ocorrência foi digitada em tempo real, ou posteriormente, não é informação viável a pesquisa.
CHAMADA	O tipo da chamada não é objeto do trabalho, pois o foco está nos acidentes atendidos.
ENDERECO	Não foi priorizado o endereço, pois esta etapa foi deixada para trabalhos futuros.
CIDADE	Como a pesquisa está sendo realizada apenas no município de Goiânia, não há variedade de cidade.
LATITUDE	A localização do acidente ficou para trabalhos futuros.
LONGITUDE	Idem item anterior.
DESCRICAO	Texto onde a ocorrência é narrada, não continha informação adequada ao foco da pesquisa.
POSICAO_VITIMA	Saber como a vítima foi encontrada pelo socorrista não é interessante para o momento desse trabalho.
HOSPITAL_CLASSIFICACAO_ID	Identificação dos hospitais de Goiânia. Não há necessidade de se saber para onde a vítima foi transportada no foco deste trabalho.
HOSP_NOME	Idem item anterior.

Fonte: CRUZ; SANTOS, 2018.

Do sistema RAI, foram desconsideradas as colunas conforme Tabela 13.

Tabela 13 - Dados desconsiderados do sistema RAI

VARIÁVEL	**MOTIVO**
ID	Não é interessante para o trabalho de pesquisa, saber o número da chave primária da ocorrência.
ORIGEM_OCORRENCIA	Mostra informações de onde as informações surgiram, sistemas diferentes do SIAE e RAI. Não estão presentes na pesquisa.
STATUS_OCORRENCIA	Informações utilizadas pela Seção de Informática para controle.
LOGRADOURO	Não foi priorizado o endereço, pois esta etapa foi deixada para trabalhos futuros.
MUNICIPIO	Como a pesquisa está sendo realizada apenas no município de Goiânia, não há variedade de cidade.
UF	Não existe outro Estado além de Goiás.
LAT	A localização do acidente ficou para trabalhos futuros.
LNG	Idem item anterior
FAIXA ETÁRIA	Foi utilizado o campo com a Data de Nascimento, onde fracionamos nossa própria faixa etária.
TOTALOGCS	Números classificatórios do nível de consciência não são interessantes para a pesquisa pois trazem informações não importantes para o trabalho.
SITUACAO_VITIMA_GERAL	A gravidade da vitima não foi classificado como importante para a pesquisa, pois o nível de consciência levaria para outro foco.
POSICAO_VITMA	Saber como a vítima foi encontrada pelo socorrista não é interessante para o momento desse trabalho.
TIPO_ACIDENTE	O foco são os acidentes de trânsito, não o doméstico ou de trabalho.
UNIDADE_SAUDE_CLASSIFICACAO	Não pertinente a pesquisa saber o tipo de unidade de saúde para onde a vítima foi transportada.
UNIDADE_SAUDE_REDE_PUBLICA	Identificação dos hospitais de Goiânia. Não há necessidade de se saber para onde a vítima foi transportada no foco deste trabalho.

Fonte: CRUZ; SANTOS, 2018.

3.3 TRANSFORMAÇÃO DOS DADOS

A situação dos dados, por arquivos recebidos, antes (registros iniciais) e após (registros finais) as duas primeiras fases do processo *KDD*, seleção, pré-processamento dos dados ocorreram conforme a Tabela 14.

Tabela 14 - Situação dos dados antes e após a exclusão dos campos nulos

ARQUIVO	TOTAL DE REGISTROS INICIAIS	TOTAL DE REGISTROS FINAIS	SISTEMA
siae2010.xlsx	11811	3135	SIAE
siae2011.xlsx	29704	13020	SIAE
siae2012.xlsx	28670	13668	SIAE
siae2013.xlsx	26122	12096	SIAE
siae2014.xlsx	26009	10939	SIAE
siae2015.xlsx	21583	10053	SIAE
siae2016.xlsx	5144	2228	SIAE
Transito_GYN_2016.xlsx	18137	10996	RAI
Transito_GYN_2017.xlsx	24298	15672	RAI
TOTAL	**191478 (100%)**	**91807 (47%)**	

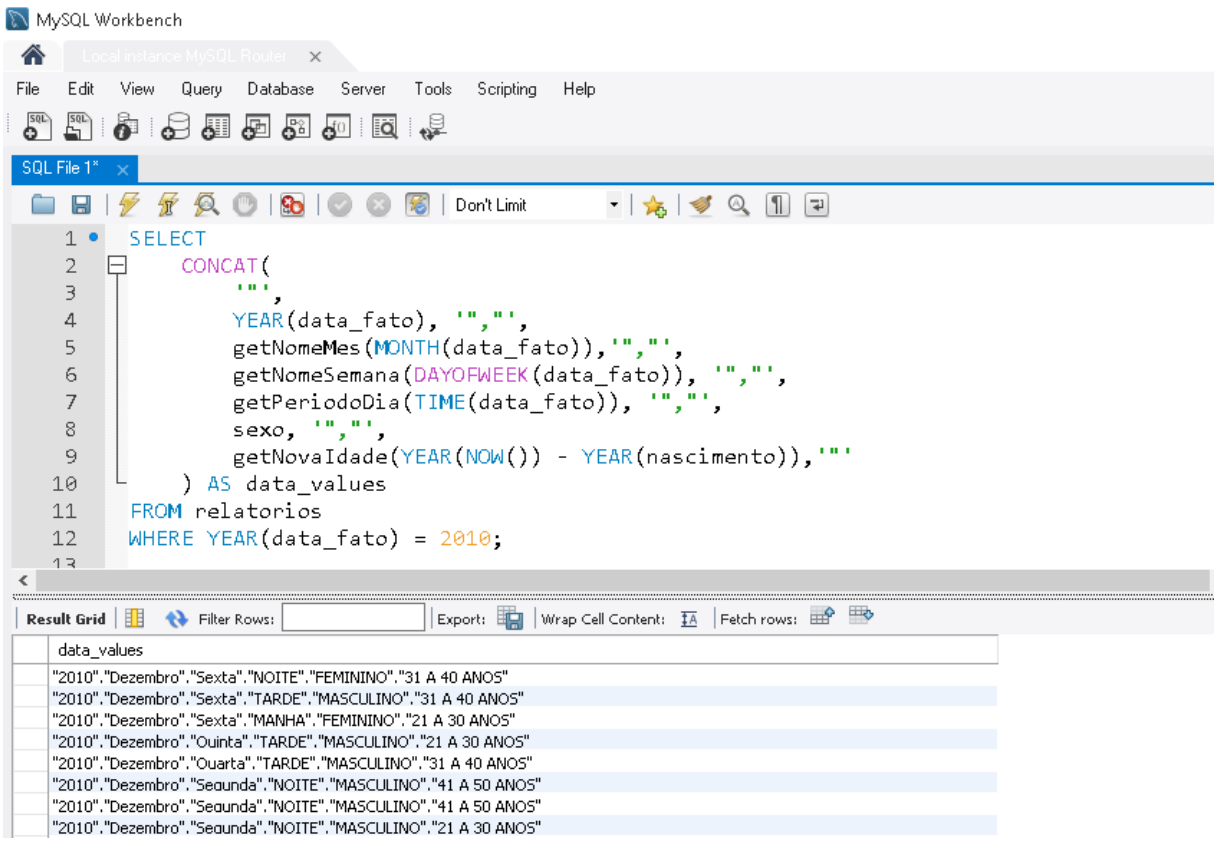
Fonte: CRUZ; SANTOS, 2018.

Observa-se, dessa forma, que de um total de 191478 registros recebidos de ambos os sistemas, restaram apenas 91807 (47%). Com isso, a proporção de registros inválidos tornou-se maior que a dos registros válidos. Logo, os resultados tendiam para o número dos inválidos, tornando a pesquisa inviável. Para solucionar essa deficiência, foi implementado, durante o pré-processamento dos dados, um método para que o algoritmo *Apriori* ignorasse os campos nulos dos dados, aproveitando apenas os dados válidos. Dessa forma, foi possível trabalhar com 100% os dados sem a necessidade de excluí-los do banco por causa dos campos nulos. Essa implementação foi realizada inserindo sinais de interrogação (?) nos campos nulos. Com isso, o algoritmo entende que naqueles campos com esse sinal existe ausência de valores, e esses campos devem ser ignorados durante a mineração dos dados.

Após concluídas as duas fases anteriores de seleção e pré-processamento dos dados restaram apenas as informações interessantes à pesquisa, ainda no formato de planilha eletrônica. Com os registros organizados, foi desenvolvido um

script em linguagem *DML (Data Manipulation Language)* para a inserção dos dados no SGBD *MySQL* 8.0 (Figura 10), da *Oracle,* os *scripts* se encontram se no APÊNDICE A. Esta ferramenta foi uma opção para a manipulação dos registros por ser mais completa e robusta que o *Microsoft Excel,* uma vez que o volume de dados é 190 mil registros aproximadamente, e quando manipulados no *excel* torna o processo mais lento. Foi feita a transformação dos dados, por exemplo, a alteração do tipo de data do formato brasileiro (DD/MM/YYYY HH:MM:SS) para o formato americano (YYYY-MM-DD HH:MM:SS). Através do campo DATA_FATO foram ordenados os dados em ano, mês, dia da semana e período do dia para que. As datas de nascimento foram divididas em intervalos de 10 anos (0 a 10 ..., ...51 a 60 e acima de 60).

Figura 10 - Dados no SGBD *MySQL 8.0* para serem transformados



Fonte: CRUZ; SANTOS, 2018.

Após os dados serem inseridos no *MySQL* foi gerada uma *QUERY* com a finalidade de concatenar todos os valores para compor as informações selecionadas no processo de mineração. Através da execução dessa *QUERY*, será gerado um conjunto de dados a serem salvos de acordo com o padrão exigido pela ferramenta

Weka, ou seja no formato **.arff**. Este tipo de arquivo possui três seções (Figura 11), sendo elas *@relation* (título da massa de dados), *@attribute* (local onde são definidos os nomes dos atributos e os tipos de dados) e *@data* (a partir dessa seção são inseridos os dados).

Figura 11 - Seção de um arquivo no formato arff

```
@relation 'acidentes-2011-2017'

@attribute ano {"2011", "2012", "2013", "2014", "2015", "2016", "2017"}
@attribute mes {"JANEIRO","FEVEREIRO","MARCO","ABRIL","MAIO","JUNHO","JULHO","AGOSTO","
@attribute semana {"DOMINGO","SEGUNDA","TERCA","QUARTA","QUINTA","SEXTA","SABADO"}
@attribute periodo {"MANHA", "TARDE", "NOITE"}
@attribute sexo {FEMININO, MASCULINO}
@attribute faixa_etaria_10_ANOS {"0 A 10 ANOS", "11 A 20 ANOS", "21 A 30 ANOS", "31 A 4

@data
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO","41 A 50 ANOS"
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO",?
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO","31 A 40 ANOS"
"2011","DEZEMBRO","SABADO","NOITE","FEMININO",?
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO",?
"2011","DEZEMBRO","SABADO","NOITE","FEMININO",?
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO","31 A 40 ANOS"
"2011","DEZEMBRO","SABADO","NOITE",?,?
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO",?
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO",?
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO","11 A 20 ANOS"
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO",?
"2011","DEZEMBRO","SABADO","NOITE","FEMININO",?
"2011","DEZEMBRO","SABADO","NOITE","FEMININO","11 A 20 ANOS"
"2011","DEZEMBRO","SABADO","NOITE",?,?
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO","0 A 10 ANOS"
"2011","DEZEMBRO","SABADO","NOITE","MASCULINO","51 A 60 ANOS"
"2011","DEZEMBRO","SABADO","NOITE","FEMININO",?
```

Fonte: CRUZ; SANTOS, 2018.

3.4 MINERAÇÃO DE DADOS

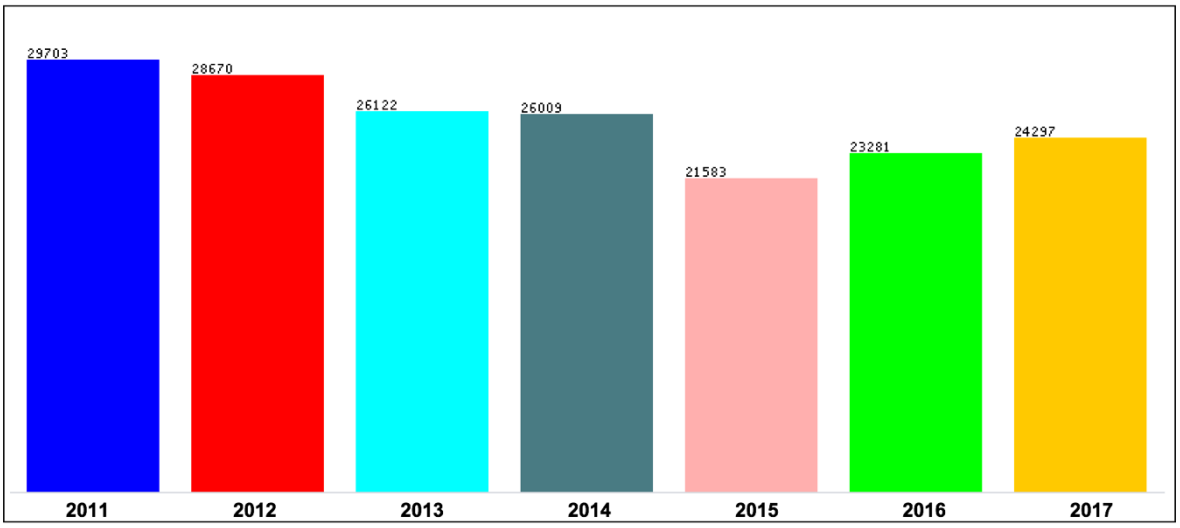
3.4.1. Análise superficial dos resultados

Pela observação dos gráficos gerados automaticamente, após a inserção dos dados na ferramenta *Weka*, é possível a percepção de alguns resultados relevantes:

- No período 2011 a 2017 (Figura 12), o número das ocorrências de acidentes de trânsito sofreu redução nos anos de 2011, 2012 e 2013. No ano de 2014 o número de acidentes foi semelhante a 2013. Em 2015 houve mais uma queda, no número dos acidentes, porém constatou-se aumento nos períodos de 2016 e 2017. Assim tem-se

uma média de aproximadamente 25666 vítimas de acidentes de trânsito por ano.

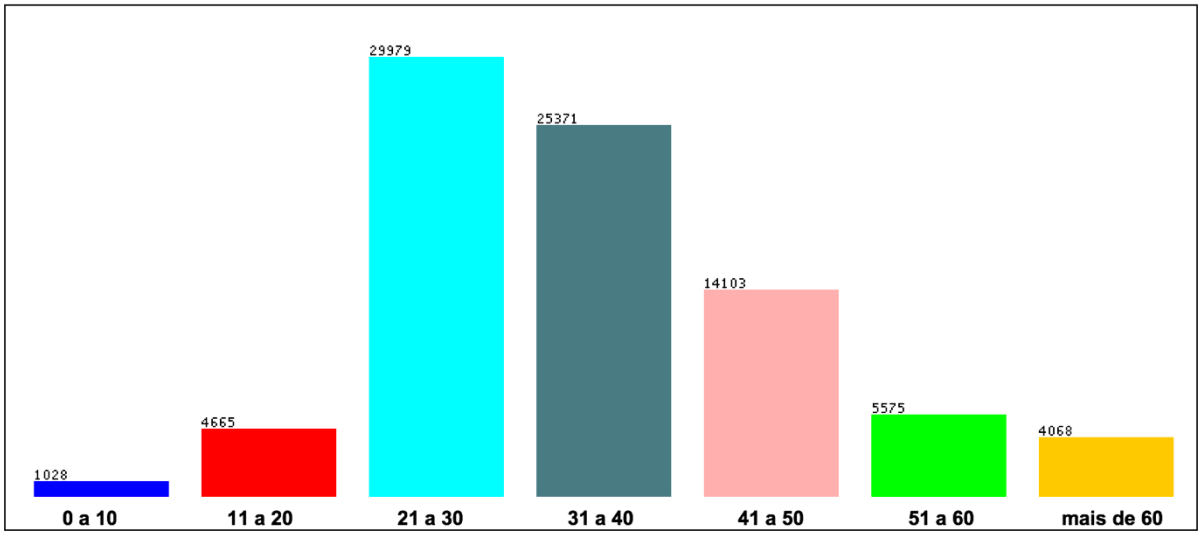
Figura 12 - Gráfico de acidentes por ano.



Fonte: CRUZ; SANTOS, 2018.

- No gráfico com resultados por faixa etária (Figura 13) verificou-se o maior número de ocorrências está para as vítimas com idade entre 21 e 30 anos, o que representa 17% de um total de 179665 casos. A segunda faixa etária mais afetada foi a de 31 a 40 anos com 25371 pessoas, ou seja 14% do total.

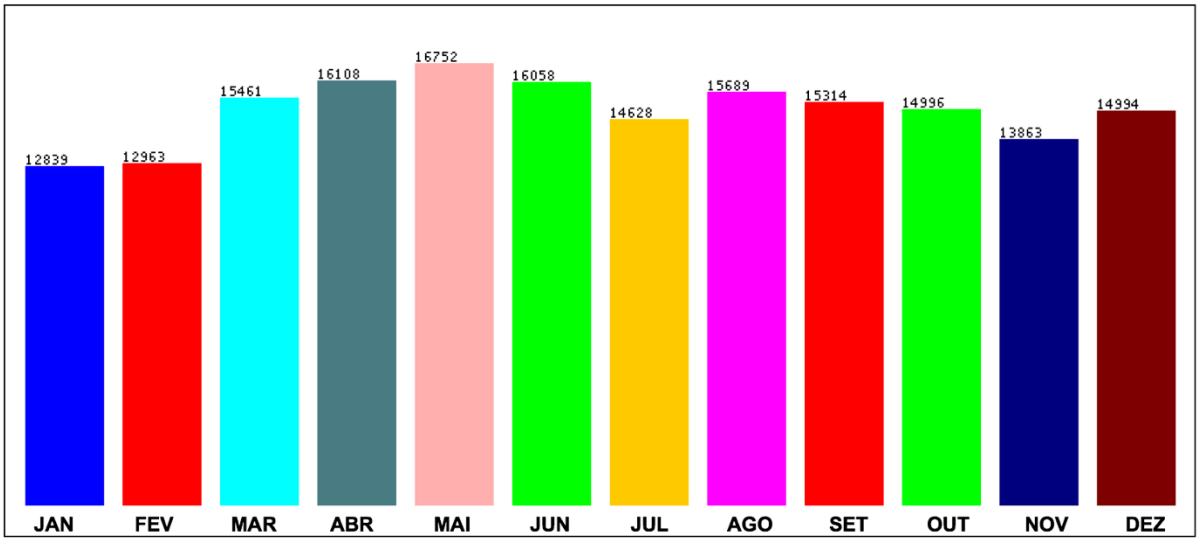
Figura 13 - Gráfico de vítimas por faixa etária.



Fonte: CRUZ; SANTOS, 2018.

- Nos resultados do total de acidentes divididos em períodos mensais (Figura 14) identificou-se que o período maior de ocorrências está no mês de maio, com 16752 casos, de um total de 179665 ocorrências, o que equivale a 9% dos acidentes. Logo em seguida vieram os meses de abril e junho. Esses números representam um média de 14972 vítimas por mês no período compreendido entre os anos de 2011 a 2017.

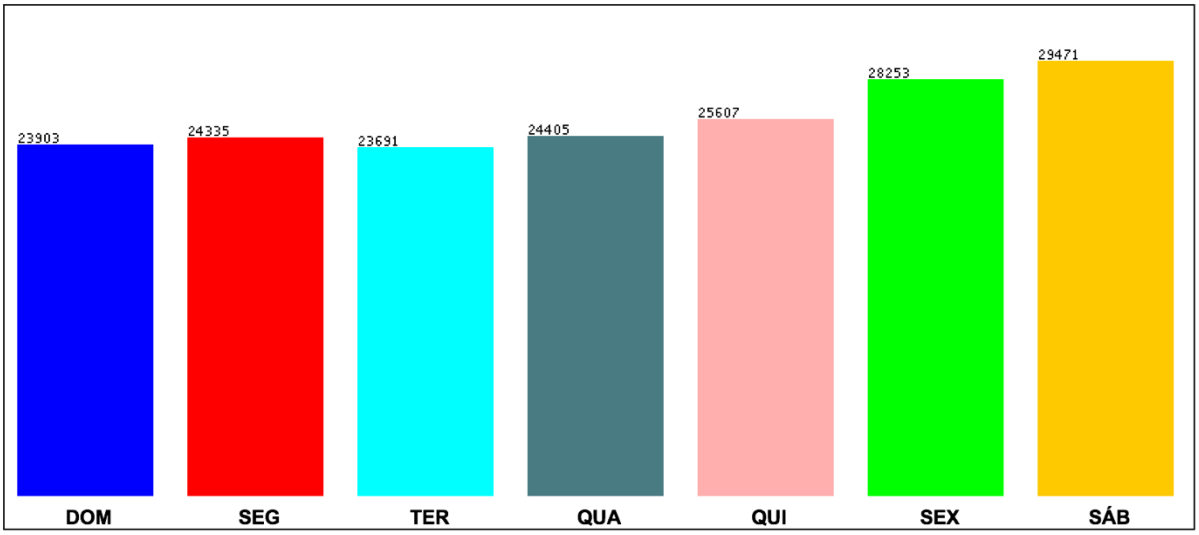
Figura 14 - Gráfico do total de acidentes por mês



Fonte: CRUZ; SANTOS, 2018.

- Na distribuição do total de acidentes por dias da semana (Figura 15), o maior número de ocorrências ficou no sábado, com 29471, o que equivale a 16% do total, seguido pelos dias de quinta e sexta-feira. O domingo apresentou o menor números de registros, 23903, ou seja, 13% do total.

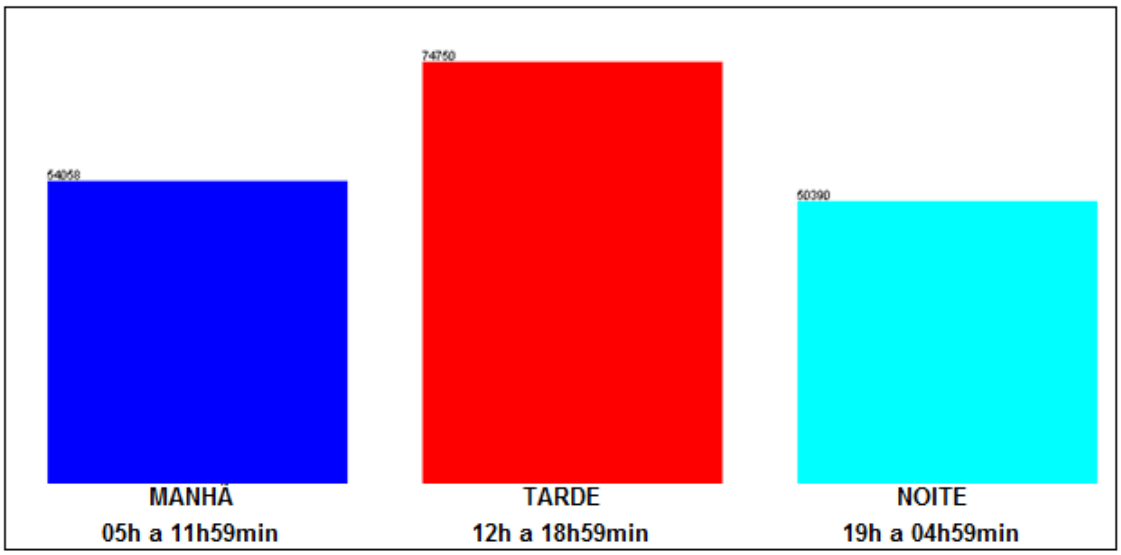
Figura 15 - Gráfico do total de acidentes por dia da semana



Fonte: CRUZ; SANTOS, 2018.

- Quando o número de vítimas foi analisado por períodos do dia, sendo eles definidos como período da manhã (05h00min a 11h59min), tarde (12h a 18h59min) e noite (19h a 04h59min), obteve-se os seguintes resultados: o período da tarde foi o que apresentou maior número de acidentes, 74750, o que equivale a 42% do total nos anos pesquisados (Figura 16).

Figura 16 - Gráfico com o total de vítimas dividido por período



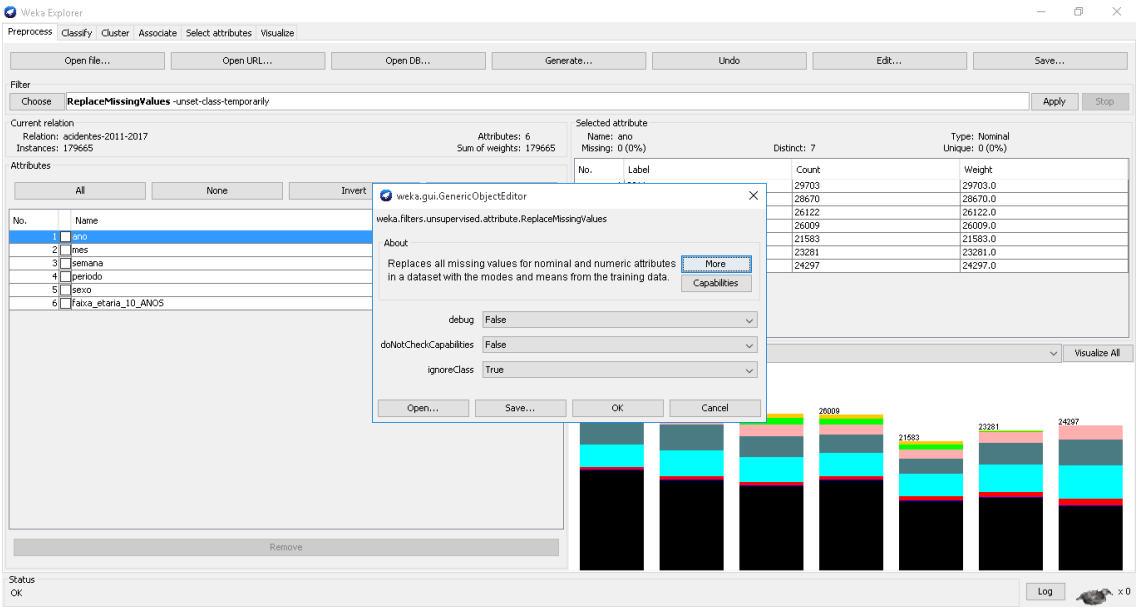
Fonte: CRUZ; SANTOS, 2018.

3.4.2. Análise dos resultados com a aplicação do algoritmo *Apriori*

Após o arquivo **acidentes_2011_2017.arff** ser aberto na ferramenta *Weka*, foi aplicado o procedimento *ReplaceMissingValues* nos dados para remover valores ausentes (Figura 17). Essa medida foi necessária para remover vícios de tendências causados por itens com suporte acima da média, e com isso obter resultado mais adequados. O *ReplaceMissingValues* possui as seguintes propriedades:

- *debug* - Se definido como verdadeiro, o filtro pode gerar informações adicionais para o console.
- *doNotCheckCapabilities* - Se definido, os recursos do filtro não serão verificados antes de serem criados. (Use com cuidado para reduzir o tempo de execução.)
- *ignoreClass* - O índice da classe será temporariamente cancelado antes que o filtro seja aplicado.

Figura 17 - Configuração do ReplaceMissingValues

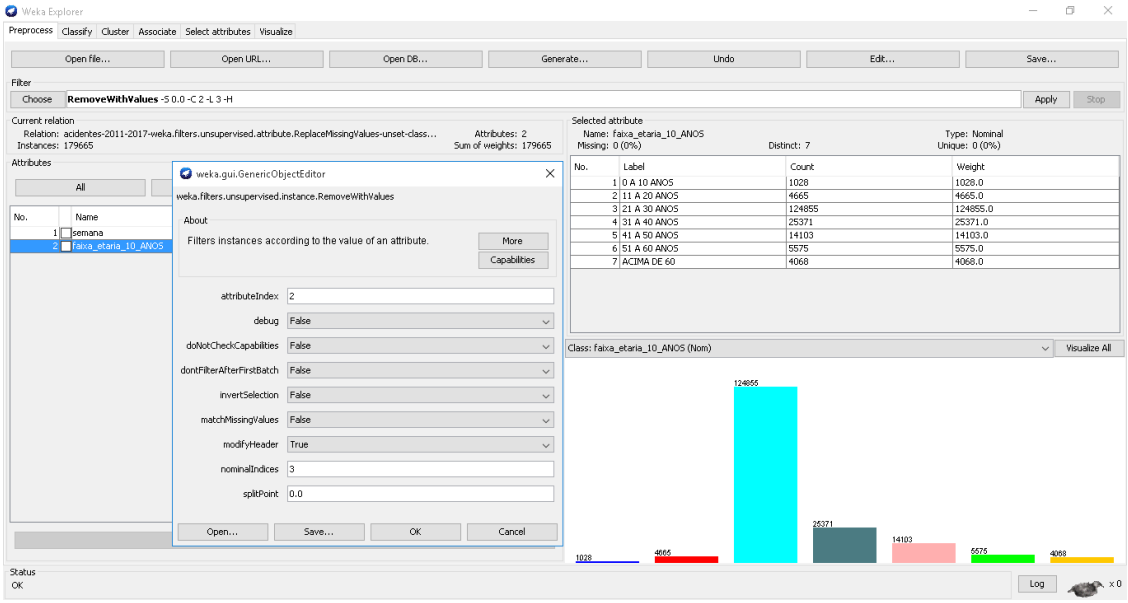


Fonte: CRUZ; SANTOS, 2018.

Para remover instâncias de atributos que possuem quantitativo acima da média, foi aplicado o filtro de pré-processamento *RemoveWithValues* (Figura 18), o qual possui os seguintes parâmetros de configuração:

- *dontFilterAfterFirstBatch* - Se deve aplicar o processo de filtragem a instâncias que são inseridas após o primeiro lote (de treinamento). O padrão é false, portanto, instâncias em lotes subsequentes podem ser "consumidas" pelo filtro.
- *debug* - Se definido como verdadeiro, o filtro pode gerar informações adicionais para o console.
- *splitPoint* - Valor numérico a ser usado para seleção no atributo numérico. Instâncias com valores menores que o valor dado serão selecionadas.
- *matchMissingValues* - Os valores ausentes são contados como uma correspondência. Essa configuração é independente da opção *invertSelection*.
- *nominalIndices* - Faixa de índices de etiqueta a serem usados para seleção no atributo nominal. Primeiro e último são índices válidos.
- *modifyHeader* - Ao selecionar em atributos nominais, remove referências de cabeçalho para valores excluídos.
- *attributeIndex* - Escolha o atributo a ser usado para seleção (padrão por último).
- *doNotCheckCapabilities* - Se definido, os recursos do filtro não serão verificados antes de serem criados. (Use com cuidado para reduzir o tempo de execução.)
- *invertSelection* - Inverte o sentido de correspondência.

Figura 18 - Configuração do filtro RemoveWithValues.



Fonte: CRUZ; SANTOS, 2018.

Com os dados configurados, foram aplicados vários testes em situações distintas, onde foram obtidos os seguintes resultados:

Cenário 1

Foram desprezados os atributos ano, mês, sexo, período do dia, para os dias da semana foram selecionados apenas sexta e sábado, faixa etária foram eleitas as faixas 0 a 10, 11 a 20, 51 a 60 e acima e 60 anos (correspondente a crianças, jovens e idosos), apresentados na Tabela 15.

Instances: 4986

Attributes: 2

semana (sexta e sábado)

faixa_etaria_10_ANOS (0 a 10, 11 a 20, 51 a 60 e acima de 60 anos)

Minimum support: 0.05 (249 *instances*)

Minimum metric <confidence>: 0.1

Number of cycles performed: 19

NumRules: 15

Tabela 15 - Resultado aplicação do Apriori - Cenário 1

Nº	RESULTADO	CONF
1	**faixa_etaria_10_ANOS=ACIMA DE 60 1309 ==> semana=SEXTA 703**	**54%**
2	**faixa_etaria_10_ANOS=11 A 20 ANOS 1627 ==> semana=SABADO 853**	**52%**
3	**faixa_etaria_10_ANOS=51 A 60 ANOS 1668 ==> semana=SEXTA 847**	**51%**
4	faixa_etaria_10_ANOS=51 A 60 ANOS 1668 ==> semana=SABADO 821	49%
5	faixa_etaria_10_ANOS=11 A 20 ANOS 1627 ==> semana=SEXTA 774	48%
6	faixa_etaria_10_ANOS=ACIMA DE 60 1309 ==> semana=SABADO 606	46%
7	semana=SABADO 2471 ==> faixa_etaria_10_ANOS=11 A 20 ANOS 853	35%
8	semana=SEXTA 2515 ==> faixa_etaria_10_ANOS=51 A 60 ANOS 847	34%
9	semana=SABADO 2471 ==> faixa_etaria_10_ANOS=51 A 60 ANOS 821	33%
10	**semana=SEXTA 2515 ==> faixa_etaria_10_ANOS=11 A 20 ANOS 774**	**31%**
11	**semana=SEXTA 2515 ==> faixa_etaria_10_ANOS=ACIMA DE 60 703**	**28%**
12	**semana=SABADO 2471 ==> faixa_etaria_10_ANOS=ACIMA DE 60 606**	**25%**

Fonte: CRUZ; SANTOS, 2018.

Este Cenário 1 foi escolhido por apresentar resultados diferentes dos encontrados naturalmente. As faixas etárias definidas neste Cenário apresentaram conhecimento novo em relação às vítimas de acidentes. As vítimas que aparecem nos resultados não estão presentes nos indicadores quantitativos.

Para as regras com a confiança inferior a 50% que se encontram em negrito, as mesmas não serão consideradas, por motivos da relação entre quantidade de ocorrência do valor encontrado na condição do **SE** ser 70% superior a quantidade encontrar no **ENTÃO**, sendo assim um valor que possui uma confiança indesejável.

Cenário 2

Na etapa de pré-processamento, foram desconsiderados os atributos de ano, faixa etária e sexo. Para os dias de semana foram selecionas os apenas de domingo a quinta feira, período manhã e tarde e abrangendo todos os meses (Tabela 16).

Instances: 69829

Attributes: 3

mês (janeiro a dezembro)

semana (domingo a quinta-feira)

período (manhã e noite)

Minimum support: 0.01 (698 instances)
Minimum metric <confidence>: 0.2
Number of cycles performed: 20
NumRules: 100

Tabela 16 - Resultado aplicação do Apriori - Cenário 2

Nº	RESULTADO	CONF
1	mes=JANEIRO semana=DOMINGO 1168 ==> periodo=NOITE 862	74%
2	mes=MARCO semana=DOMINGO 1184 ==> periodo=NOITE 864	73%
3	mes=OUTUBRO semana=TERCA 1593 ==> periodo=MANHA 1138	71%
4	mes=DEZEMBRO semana=DOMINGO 1371 ==> periodo=NOITE 970	71%
5	mes=OUTUBRO semana=DOMINGO 1037 ==> periodo=NOITE 722	70%
6	mes=NOVEMBRO semana=DOMINGO 1176 ==> periodo=NOITE 814	69%
7	mes=JULHO semana=DOMINGO 1202 ==> periodo=NOITE 831	69%
8	mes=ABRIL semana=DOMINGO 1343 ==> periodo=NOITE 928	69%
9	semana=DOMINGO 14453 ==> periodo=NOITE 9978	69%
10	mes=SETEMBRO semana=DOMINGO 1264 ==> periodo=NOITE 855	68%
	...	
85	**mes=JANEIRO 4846 ==> semana=DOMINGO 1168**	**24%**
87	**mes=DEZEMBRO 5805 ==> semana=DOMINGO 1371**	**23%**
95	**mes=NOVEMBRO 5228 ==> semana=DOMINGO 1176**	**22%**

Fonte: CRUZ; SANTOS, 2018.

O principal motivo da escolha do Cenário 2 foi o nível de confiança apresentado, chegando a 74% na melhor solução e apresentando um bom número de resultados acima de 50%. Fato que mostra a qualidade das informações encontradas. Pois em comparação as confianças em negrito correspondentes a 24%, 23% e 22% na parte final da Tabela 16, que são os menores valores de confiança gerados, percebe-se que 74% é um bom resultado**.**

Cenário 3

Para presente cenário na etapa de pré-processamento, foram aplicados filtros nos atributos, removendo sexo masculino com o objetivo de realizar uma mineração voltada para o público feminino, foram considerados apenas as faixas etárias de 0 a

10 anos (faixa etária referente a crianças), referente ao período de 7 anos (2011 a 2017), para os período do dia foram removidos o período da tarde, onde os resultados apresentam-se na Tabela 17.

Instances: 173
Attributes: 3
mes (janeiro a dezembro)
semana (domingo a quinta-feira)
periodo (manha e noite)

Minimum support: 0.1 (17 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 18
NumRules: 10

Tabela 17 - Resultado aplicação do Apriori - Cenário 3

Nº	RESULTADO	CONF
1	semana=DOMINGO 34 ==> periodo=NOITE 28	82%
2	semana=SABADO 40 ==> periodo=NOITE 32	80%
3	semana=QUARTA 25 ==> periodo=MANHA 18	72%
4	semana=SEXTA 32 ==> periodo=NOITE 18	56%

Fonte: CRUZ; SANTOS, 2018.

Este Cenário 3 foi escolhido por apresentar ótimos resultados em relação ao dia da semana e o período do dia em que mais ocorrem acidentes.

3.5 RESULTADOS OBTIDOS

De acordo com a proposta deste trabalho foi solicitado junto à Seção de Informática do Corpo de Bombeiros Militar do Estado de Goiás a possibilidade de que fossem disponibilizadas informações do banco de dados do sistema de registro das ocorrências atendidas por este órgão. A instituição atendeu à solicitação de disponibilizou as informações.

Os dados estavam divididos em dois bancos, um do sistema SIAE e outro do sistema RAI, o mais atual. Os dados foram repassados no formato de arquivos de

planilhas *Microsoft Excel 2016*. Onde cada ano veio em um arquivo diferente, e posteriormente transformados em um único arquivo. Para que pudesse ser aberto pela ferramenta *Weka*, o arquivo foi transformado no tipo **arff**.

Com os dados lidos pelo *Weka*, tornou-se possível a análise de vários gráficos gerados, de onde foram elencadas várias informações sobre o tema ora em análise. Também foi possibilitado a verificação de resultados de alguns algoritmos diferentes do *Apriori*, como o EM de clusterização. Esses resultados não se mostraram tão eficazes para este trabalho quanto resultados do algoritmo *Apriori.*

Para os resultados deste trabalho foram desprezados os dados do período de 2010, pois apresentavam informações apenas do segundo semestre. Através do processo *KDD* para a mineração dos dados foram obtidos os seguintes resultados:

- A análise das ocorrências de acidentes de trânsito em grupos anuais mostrou um decréscimo no período de 2011 a 2015, porém um crescimento nos anos de 2016 e 2017;
- No intervalo das vítimas por faixas etárias (10 em 10 anos), o grupo que mais se destacou foi o 21 a 30 anos, o que correspondeu a 17% do total;
- Para o conjunto total de acidentes distribuídos nos 12 meses do ano, o mês que apresentou maior número de acidentes de trânsito maio;
- Os resultados da distribuição por dia da semana apontaram que os dias de maior incidência dos acidentes são o sábado e a sexta-feira, onde o sábado ficou com maior número, 16% do total;
- Na divisão dos resultados por períodos do dia, em manhã, tarde e noite, observou-se que a maioria dos acidentes ocorrem no período da tarde, com 42% do total.

Após a realização mineração dos dados, foram selecionados 3 cenários para análise. Os resultados encontrados no Cenário 1 identificaram que:

- Idosos com idade acima de 60 anos, possuem a confiança de 54% de chances, sofrerem acidentes na sexta-feira.
- Jovens com idade de 11 a 20 anos possuem a confiança de 52% de chances, sofrerem acidentes no sábado.
- Homens na faixa etária de 51 a 60 anos possuem a confiança de 51% de chances, sofrerem acidentes na sexta-feira.

Para o Cenário 2 foram identificadas 10 regras, sendo as mesmas geradas com 3 atributos, todas acima de 60 % de confiança, as presentes descobertas demonstram conhecimento das ocorrências referente aos meses, semanas e turno e suas respectivas relações, abaixo estão descritas as regras referente a Tabela 14 que se encontram em negrito:

- Para a combinação do mês de Janeiro e dia da semana Domingo, possuem a confiança de 74% de chances de sofrerem acidentes no período da Noite.
- Mês de Março e Domingo, possuem confiança de 73% de chances de sofrerem acidentes no período da Noite.
- Mês de Outubro no dia da semana Terça, possuem confiança de 71% de sofrerem acidentes no período da Noite.

O cenário 3 foram utilizados filtros mais aprofundados a um contexto de faixa etária voltada a crianças (0 a 10 anos) do sexo feminino, apenas do período da manhã e noite, foram obtidas as seguintes regras abaixo:

- Para o dia da semana Domingo, possuem confiança de 84% de chances do acidente de trânsito no período da Noite.
- Aos Sábados, possuem confiança de 80% de chances do acidente de trânsito no período da Noite.
- As Quarta feiras, possuem 72% de chances de ocorrem acidente de trânsito no período da Manhã.
- Para o dia da semana Sexta feira, possuem 58% de chances de ocorrem acidente de trânsito no período da Noite.

3.6 TRABALHOS FUTUROS

Nesta seção apresenta-se uma lista com a propostas para trabalhos futuros:

- Desenvolvimento de um sistema para a integração entre a ferramenta *Weka* e o banco de dados do RAI, possibilitando resultados instantâneos dos dados inseridos;
- Implementação para as demais cidades do Estado de Goiás;

- Incorporação de dados oriundos de outras fontes para ampliação e produção de resultados mais amplos.

4 CONSIDERAÇÕES FINAIS

Diante dos resultados observados torna-se notável que a tecnologia de mineração de dados se apresenta como uma valiosa ferramenta a ser utilizada na compreensão de dados. Os conhecimentos revelados pelos processos de mineração de dados são úteis para auxiliar, tanto no setor privado, como no público, nas tomadas de decisões.

Esta análise permitirá aos gestores públicos do município de Goiânia estabelecer estratégias e tomar decisões mais assertivas nas medidas de mitigação dos acidentes de trânsito ocorridos naquele contexto. Pois, como exemplo, se é conhecida a faixa etária dos condutores que mais sofrem acidentes, pode-se criar medidas voltadas a conscientização de tais motoristas, observando temas adequados aquela faixa etária. Medidas como propagandas de divulgação nas mídias sociais, no rádio, na televisão podem ser positivas para reduzir o número de acidentes. Outro exemplo é o conhecimento dos dias da semana e período do dia em que mais ocorrem os acidentes. Diante disso pode-se buscar medidas como aumentar o policiamento nesses dias e períodos do dia e forma preventiva e para orientação.

Por fim, com os resultados encontrados através desta pesquisa, espera-se que os gestores públicos sejam despertados para o assunto e tenham na mineração de dados mais uma ferramenta útil na produção de informações para ampliar a eficiência das ações de mitigação dos acidentes de trânsito no município de Goiânia, estendendo-se aos demais municípios do Estado.

REFERENCIAL BIBLIOGRÁFICO

ALMEIDA, M. S. **Elaboração de projeto, TCC, dissertação e tese**: Uma abordagem simples, prática e objetiva. São Paulo: Atlas, 2011.

BRASIL. **Código de trânsito brasileiro**. 4. ed. Brasília: Câmara dos Deputados, Edições Câmara, 2010.

CÂMARA DOS DEPUTADOS. **Impacto anual dos acidentes de trânsito é R$ 2,3 bilhões no Brasil, diz ministro da Saúde**. Disponível em: <http://www2.camara.leg.br/camaranoticias/noticias/TRANSPORTE-E-TRANSITO/514460-IMPACTO-ANUAL-DOS-ACIDENTES-DE-TRANSITO-E-R$-2,3-BILHOES-NO-BRASIL,-DIZ-MINISTRO-DA-SAUDE.html >. Acesso em: 30 out. 2017.

CASTRO, L. N.; FERRARI, D. G. **Introdução à mineração de dados**. São Paulo: Saraiva, 2016.

CHAVANTE, E.; PRESTES, D. **Quadrante matemática, 3º ano: ensino médio.** 1. ed. São Paulo: Edições SM, 2016.

DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TRANSPORTES. **Anuário estatístico das rodovias federais 2010.** Disponível em: <http://www.dnit.gov.br/download/rodovias/operacoes-rodoviarias/estatisticas-de-acidentes/anuario-2010.pdf >. Acesso em: 30 out. 2017.

DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TTRANSPORTES. **Estatísticas de acidentes.** Disponível em: <http://www.dnit.gov.br/rodovias/operacoes-rodoviarias/estatisticas-de-acidentes>. Acesso em 30 out. 2017.

DEVMEDIA. **Mineração de Dados: Tarefas e Técnicas.** Disponível em: <https://www.devmedia.com.br/mineracao-de-dados-tarefas-e-tecnicas/30919>. Acesso em: 15 nov. 2017.

E. Frank.; M. A. Hall; I. H. Witten. **The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"**, Morgan Kaufmann, Fourth Edition, 2016.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados.** 6. ed. São Paulo: Pearson Addison Wesley, 2011.

FABRICIO, J. A.; GRANATYR, J.; GOMES, H. M. **Descoberta de conhecimento utilizando o processo KDD.** Disponível em: < https://www.devmedia.com.br/descoberta-de-conhecimento-utilizando-o-processo-kdd/38709>. Acesso em: 10 abr. 2018.

GONÇALVES, E. C. Extração de Árvores de Decisão com a Ferramenta de Data Mining Weka. **DEVMEDIA**. Disponível em: <https://www.devmedia.com.br/extracao-de-arvores-de-decisao-com-a-ferramenta-de-data-mining-weka/3388>. Acesso em: 17 nov. 2017.

GRANATYR, J.; **Mineração de Regras de Associação com Weka, Apriori e Java**. Disponível em: < https://www.udemy.com/mineracao-de-regras-de-associacao-com-weka-apriori-e-java>. Acesso em: 18 maio 2018.

GUIMARÃES, C. C. **Fundamentos de bancos de dados: modelagem, projeto e linguagem SQL.** Campinas, SP: Editora UNICAMP, 2003.

GUROVITZ, H. **O que cerveja tem a ver com fraldas?** Disponível em< https://exame.abril.com.br/revista-exame/o-que-cerveja-tem-a-ver-com-fraldas-m0053931/ >. Acesso em: 12 abr. 2018

HOSPITAL DE URGÊNCIAS DE GOIÂNIA (HUGO). **Estatísticas**. Disponível em: < http://hugo.org.br/estatisticas/ > Acesso em: 20 set. 2017.

LUCCA G.; PEREIRA I. A.; PRISCO A.; BORGES E. N. Uma implementação do algoritmo Naïve Bayes para classificação de texto. Centro de Ciências Computacionais – Universidade Federal do Rio Grande (FURG), Rio Grande – RS – Brasil. 2013. Disponível em: < http://www.lbd.dcc.ufmg.br/colecoes/erbd/2013/0019.pdf>. Acesso em: 18 nov. 2017.

MELO, I. V. A. **Normalização de Bancos de Dados Relacionais**. Disponível em <http://www.dsc.ufcg.edu.br/~pet/jornal/maio2011/materias/recapitulando.html>. Acesso em: 26 out. 2017.

METRO. **Brasil é o quinto país do mundo em mortes no trânsito, segundo OMS**. Disponível em: <https://www.metrojornal.com.br/foco/2017/05/01/brasil-e-o-quinto-pais-mundo-em-mortes-no-transito-segundo-oms.html>. Acesso em: 30 out. 2017.

MICHAELIS. **Dicionário brasileiro da língua portuguesa.** Disponível em: <http://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/acidente/>. Acesso em 29 out. 2017.

NETO, M. A. S.; VILLWOCK, R; SCHEER, S.; STEINER, M. T. A.; DYMINSKI, A. S. **Técnicas de mineração visual de dados aplicadas aos dados de instrumentação da barragem de Itaipu**. Disponível em: <http://dx.doi.org/10.1590/S0104-530X2010000400007>. Acesso em: 19 nov. 2017.

OBSERVATÓRIO NACIONAL DE SEGURANÇA VIÁRIA. **Mortes por modal**. Disponível em: <http://iris.onsv.org.br/iris-beta/#/stats/profiles/52/death>. Acesso em 30 out. 2017.

ORGANIZACIÓN MUNDIAL DE LA SALUD. **Informe mundial sobre prevención de los traumatismos causados por el tránsito: resumen**. Washington DC: Organización Mundial de la Salud; 2004.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS NO BRASIL. **OMS: Brasil é o país com maior número de mortes de trânsito por habitante da América do Sul.** Disponível em: < https://nacoesunidas.org/oms-brasil-e-o-pais-com-maior-numero-de-mortes-de-transito-por-habitante-da-america-do-sul/>. Acesso em: 19 nov. 2017.

OPOPULAR. **Goiás gastou R$ 2,7 bilhões com acidentes no trânsito em 2014, aponta pesquisa**. Disponível em: <https://www.opopular.com.br/editorias/cidade/goi%C3%A1s-gastou-r-2-7-bilh%C3%B5es-com-acidentes-no-tr%C3%A2nsito-em-2014-aponta-pesquisa-1.1271004>. Acesso em: 02 nov. 2017.

POLÍCIA RODOVIÁRIA FEDEREAL. **Após queda de 25% em 2015, mortes em rodovias federais no Paraná sobem 12% em 2016**. Disponível em < https://www.prf.gov.br/portal/estados/parana/noticias/apos-queda-de-25-em-2015-mortes-em-rodovias-federais-no-parana-sobem-12-em-2016>. Acesso em 02 nov. 2017.

SES-GO. **Encontro conscientiza população para redução dos acidentes de trânsito**. Disponível em: <http://www.saude.go.gov.br/encontro-conscientiza-populacao-para-reducao-dos-acidentes-de-transito/>. Acesso em: 02 nov. 2017.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R.** 1. ed. Rio de Janeiro: Elsevier, 2016.

SUNIL, RAY.6 passos fáceis para aprender o algoritmo Naive Bayes (com o código em Python). Disponível em: <https://www.vooo.pro/insights/6-passos-faceis-para-aprender-o-algoritmo-naive-bayes-com-o-codigo-em-python/>. Acesso em: 30 jun 2018.

SSPAP. Plataforma de sistemas integrados inova segurança pública em goiás. Disponível em: <http://www.ssp.go.gov.br/destaques/plataforma-de-sistemas-integrados-inova-seguranca-publica-em-goias.html>. Acesso em: 17 out. 2017.

TRANSITOBR.**Acidentes - causas.** Disponível em: <http://www.transitobr.com.br/index2.php?id_conteudo=8>. Acesso em 28 out. 2017.

VIAS SEGURAS. **Acidentes no Estado de Goiás: estatísticas do DETRAN.** Disponível em: < http://www.vias-seguras.com/os_acidentes/estatisticas/estatisticas_estaduais/estatisticas_de_acidentes_no_estado_de_goias/acidentes_no_estado_de_goias_estatisticas_do_detran>. Acesso em: 30 out. 2017.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação**. 6. ed. Rio de Janeiro: Elsevier, 2009.

WILLIAM, S. **Trânsito - mortes em acidentes de trânsito em Goiânia chega a 42 neste ano**. Disponível em: <http://www.blogjadirgomes.com.br/transito-mortes-goiania-estatistica-2017> Acesso em: 19 set 2017.

APÊNDICE

APÊNDICE A – SCRIPTS DDL E SQL – TRANSFORMAÇÃO DE DADOS

SCRITPS DDL E SQL– TRANSFORMAÇÃO DE DADOS

Os scripts abaixo foram utilizados para auxiliar nas etapas de pré-processamento e transformação dos dados.

1. Criação da base de dados e tabela.

Figura 19 - Criação da Estrutura do Banco de Dados

```
1  -- SCRITPS DE DML REFERENTE A CRIAÇÃO DAS ESTRUTURAS DE TABELA E FUNCTIONS QUE
2  -- AUXILIAM NA TRANSFORMAÇÃO DOS DADOS E GERAÇÃO DO ARQUIVO ARFF.
3
4  set GLOBAL log_bin_trust_function_creators = 1;
5
6  -- EXCLUI A BASE CASO A MESMA EXISTA.
7  drop database if exists weka_mineracao;
8
9  -- CRIA UMA NOVA BASE CASO A MESMA NÃO EXISTA
10 create database if not exists weka_mineracao;
11
12 -- SELECIONA A BASE DE DADOS
13 use weka_mineracao;
14
15 -- CRIA TABELA COM OS DADOS SELECIONADO, CASO A TABELA NAO EXISTA
16 create table if not exists relatorios (
17     id int not null primary key auto_increment,
18     data_fato datetime,
19     natureza text,
20     qualificacao text,
21     bairro text,
22     sexo text,
23     nascimento datetime
24 )ENGINE=InnoDB;
25
```

Fonte: CRUZ; SANTOS, 2018.

2. Criação da função que transforma a idade da vítima em intervalos de 10 anos.

Figura 20 - Criação da função de intervalo de idade

```
-- EXCLUI A FUNCAO CASO ELA EXISTA
drop function if exists getNovaIdade;

-- CRIA FUNCAO REPONSAVEL POR RETORNA A FAIXA ETARIA DA IDADE, A CADA 10 ANOS
delimiter $$
create function getNovaIdade(idade int) returns varchar(200)
begin
    declare novo_nome varchar(200);

    if (idade >= 0 and idade <= 10) then
        set novo_nome = '0 A 10 ANOS';
    elseif (idade >= 11 and idade <= 20) then
        set novo_nome = '11 A 20 ANOS';
    elseif (idade >= 21 and idade <= 30) then
        set novo_nome = '21 A 30 ANOS';
    elseif (idade >= 31 and idade <= 40) then
        set novo_nome = '31 A 40 ANOS';
    elseif (idade >= 41 and idade <= 50) then
        set novo_nome = '41 A 50 ANOS';
    elseif (idade >= 51 and idade <= 60) then
        set novo_nome = '51 A 60 ANOS';
    elseif (idade > 60) then
        set novo_nome = 'ACIMA DE 60';
    else
        set novo_nome = 'NULL';
    end if;
    return novo_nome;
end $$
delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

3. Criação da função que recupera o nome dos meses conforme o número mês informado.

Figura 21 - Criação da função que recupera o nome do mês

```
120  -- EXCLUI A FUNCAO CASO ELA EXISTA
121  drop function if exists getNomeMes;
122
123  -- CRIA A FUNCAO QUE RETORNA O NOME DO MES CONFORME O NUMERO DO MES INFORMADO
124  delimiter $$
125  create function getNomeMes(mes int) returns varchar(200)
126  begin
127      declare novo_nome varchar(200);
128
129      if (mes = 1) then
130          set novo_nome = 'Janeiro';
131      elseif (mes = 2) then
132          set novo_nome = 'Fevereiro';
133      elseif (mes = 3) then
134          set novo_nome = 'Marco';
135      elseif (mes = 4) then
136          set novo_nome = 'Abril';
137      elseif (mes = 5) then
138          set novo_nome = 'Maio';
139      elseif (mes = 6) then
140          set novo_nome = 'Junho';
141      elseif (mes = 7) then
142          set novo_nome = 'Julho';
143      elseif (mes = 8) then
144          set novo_nome = 'Agosto';
145      elseif (mes = 9) then
146          set novo_nome = 'Setembro';
147      elseif (mes = 10) then
148          set novo_nome = 'Outubro';
149      elseif (mes = 11) then
150          set novo_nome = 'Novembro';
151      elseif (mes = 12) then
152          set novo_nome = 'Dezembro';
153      else
154          set novo_nome = 'Mes Invalido';
155      end if;
156      return novo_nome;
157  end $$
158  delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

4. Criação da função que recupera o nome da semana conforme o número da semana extraído da data da ocorrência.

Figura 22 - Cria a função que recupera o nome da semana

```
160  -- EXCLUI A FUNCAO CASO ELA EXISTA
161  drop function if exists getNomeSemana;
162
163  -- CRIA A FUNCAO QUE RETORNA O NOME DA SEMANA, CONFORME O NUMERO DA SEMANA INFORMADO
164  delimiter $$
165  create function getNomeSemana(semana int) returns varchar(200)
166  begin
167      declare novo_nome varchar(200);
168
169      if (semana = 1) then
170          set novo_nome = 'Domingo';
171      elseif (semana = 2) then
172          set novo_nome = 'Segunda';
173      elseif (semana = 3) then
174          set novo_nome = 'Terca';
175      elseif (semana = 4) then
176          set novo_nome = 'Quarta';
177      elseif (semana = 5) then
178          set novo_nome = 'Quinta';
179      elseif (semana = 6) then
180          set novo_nome = 'Sexta';
181      elseif (semana = 7) then
182          set novo_nome = 'Sabado';
183      else
184          set novo_nome = 'Semana Invalido';
185      end if;
186      return novo_nome;
187  end $$
188  delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

5. Criação da função que corrigi a data da ocorrência removendo a descrição do mês e substituindo pelo padrão que o SGBD aceita.

Figura 23 - Cria a função que corrigida data para o valor que o SGBD solicita

```
-- EXCLUI A FUNCAO CASO ELA EXISTA
drop function if exists correcaoMesData;

-- CRIA A FUNCAO REPONSAVEL POR CORRIGIR A DATA DA OCORRENCIA, REMOVENDO A DESCRIÇÃO DO MES E SUBSTITUINDO PELO PADRAO QUE O SGBD ACEITA.
delimiter $$
create function correcaoMesData(data varchar(200)) returns varchar(200)
begin
    declare mes varchar(100);
    declare mesValor varchar(100);
    declare nova_data varchar(200);

    set nova_data = '';

    if (data != '') then
        set mes = SUBSTRING(TRIM(data), 4, 3);

        if (mes = 'JAN') then
            set mesValor = '01';
        elseif (mes = 'FEV') then
            set mesValor = '02';
        elseif (mes = 'MAR') then
            set mesValor = '03';
        elseif (mes = 'ABR') then
            set mesValor = '04';
        elseif (mes = 'MAI') then
            set mesValor = '05';
        elseif (mes = 'JUN') then
            set mesValor = '06';
        elseif (mes = 'JUL') then
            set mesValor = '07';
        elseif (mes = 'AGO') then
            set mesValor = '08';
        elseif (mes = 'SET') then
            set mesValor = '09';
        elseif (mes = 'OUT') then
            set mesValor = '10';
        elseif (mes = 'NOV') then
            set mesValor = '11';
        elseif (mes = 'DEZ') then
            set mesValor = '12';
        else
            set mesValor = '00';
        end if;

        set nova_data = REPLACE(data, mes, mesValor);
    end if;

    return nova_data;
end $$
delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

6. Criação da função que converte os valores da data de ocorrência para o padrão do SGBD, para os dados do sistema SIAE.

Figura 24 - Cria a função que converte a data para o padrão do SGBD

```
308  -- EXCLUI A FUNCAO CASO ELA EXISTA
309  drop function if exists getDataOcorrenciaToBase;
310
311  -- CRIA A FUNCAO REPONSAVEL POR CONVERTER A DATA DO PADRAO PT_BR PARA O ACEITAVEL PELA BASE DE DADOS.
312  delimiter $$
313  create function getDataOcorrenciaToBase(data_ocorrencia varchar(200)) returns datetime
314  begin
315      declare nova_data datetime;
316      set nova_data = null;
317      set data_ocorrencia = correcaoMesData(data_ocorrencia);
318
319      if (data_ocorrencia != '') then
320          set nova_data = STR_TO_DATE(SUBSTRING(TRIM(data_ocorrencia), 1, 19), '%d-%m-%Y %H:%i:%s');
321      end if;
322
323      return nova_data;
324  end $$
325  delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

7. Criação da função que converte os valores da data de nascimento para o padrão do SGBD.

Figura 25 - Cria a função que converte a data para o padrão do SGBD

```
-- EXCLUI A FUNCAO CASO ELA EXISTA
drop function if exists getDataNascimentoBrToUs;

-- CRIA A FUNCAO REPONSAVEL POR CONVERTER A DATA DE NASCIMENTO DO PADRAO PT_BR PARA O ACEITAVEL PELA BASE DE DADOS.
delimiter $$
create function getDataNascimentoBrToUs(data_nascimento varchar(200)) returns datetime
begin
    declare nova_data varchar(200);
    set nova_data = null;

    if (data_nascimento != '') then
        set nova_data = STR_TO_DATE(SUBSTRING(TRIM(data_nascimento), 1, 19), '%d/%m/%Y %H:%i:%s');
    end if;

    return nova_data;
end $$
delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

8. Criação da função que remove os caracteres especiais.

Figura 26 - Remove os caracteres especias de uma palavra.

```
-- EXCLUI A FUNCAO CASO ELA EXISTA
drop function if exists removerAcento;

-- CRIA A FUNCAO REPONSAVEL POR REMOVER TODOS OS CARACTERES ESPECIAIS.
delimiter $$
CREATE FUNCTION removerAcento(Texto VARCHAR(500)) RETURNS varchar(500)
   DETERMINISTIC
   BEGIN
       declare semAcento varchar(500);
       SELECT lower(Texto) INTO semAcento;
       SELECT REPLACE(semAcento,'ã','a') INTO semAcento;
       SELECT REPLACE(semAcento,'á','a') INTO semAcento;
       SELECT REPLACE(semAcento,'â','a') INTO semAcento;
       SELECT REPLACE(semAcento,'à','a') INTO semAcento;
       SELECT REPLACE(semAcento,'ê','e') INTO semAcento;
       SELECT REPLACE(semAcento,'é','e') INTO semAcento;
       SELECT REPLACE(semAcento,'ë','e') INTO semAcento;
       SELECT REPLACE(semAcento,'ï','i') INTO semAcento;
       SELECT REPLACE(semAcento,'í','i') INTO semAcento;
       SELECT REPLACE(semAcento,'ó','o') INTO semAcento;
       SELECT REPLACE(semAcento,'ô','o') INTO semAcento;
       SELECT REPLACE(semAcento,'ö','o') INTO semAcento;
       SELECT REPLACE(semAcento,'ú','u') INTO semAcento;
       SELECT REPLACE(semAcento,'ü','u') INTO semAcento;
       SELECT REPLACE(semAcento,'ç','c') INTO semAcento;
       SELECT REPLACE(semAcento,'ñ','n') INTO semAcento;
       SELECT upper(semAcento) INTO semAcento;
       RETURN semAcento;
 end $$
delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

9. Criação da função que retorna o intervalo do período do dia.

Figura 27 - Cria a função que recupera o período do dia

```
-- EXCLUI A FUNCAO CASO ELA EXISTA
drop function if exists getPeriodoDia;

-- CRIA A FUNCAO REPONSAVEL POR RECUPERAR O PERIODO DO DIA, CONFORME DA DATA DA OCORRENCIA INFORMADA.
delimiter $$
create function getPeriodoDia(periodo time) returns varchar(200)
begin
    declare novo_nome varchar(200);

    if (periodo BETWEEN time('05:00:00') AND time('12:59:00')) then
        set novo_nome = 'MANHA';
    elseif (periodo BETWEEN time('13:00:00') AND time('20:59:00')) then
        set novo_nome = 'TARDE';
    elseif ((periodo BETWEEN time('21:00:00') AND time('23:59:00')) OR(periodo BETWEEN time('00:00:00') AND time('04:59:00'))) then
        set novo_nome = 'NOITE';
    else
        set novo_nome = 'PERIODO INVALIDO';
    end if;
    return novo_nome;
end $$
delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

10. Criação da função que converte os valores da data de ocorrência para o padrão do SGDB, para os dados do sistema RAI.

Figura 28 - Cria a função que formata a data para o padrão do SGBD referente ao sistema RAI

```
398  -- EXCLUI A FUNCAO CASO ELA EXISTA
399  drop function if exists getDataOcorrenciaToBaseRAI;
400
401  -- CRIA A FUNCAO REPONSAVEL POR CONVERTER A DATA DO PADRAO PT_BR PARA O ACEITAVEL PELA BASE DE DADOS.
402  delimiter $$
403  create function getDataOcorrenciaToBaseRAI(data_ocorrencia varchar(200)) returns datetime
404  begin
405      declare nova_data datetime;
406      set nova_data = null;
407
408      if (data_ocorrencia != '') then
409          set nova_data = STR_TO_DATE(SUBSTRING(TRIM(data_ocorrencia), 1, 17), '%d/%m/%Y %H:%i:%s');
410      end if;
411
412      return nova_data;
413  end $$
414  delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

11. Criação da função que converte os valores da data de nascimento para o padrão do SGBD, para os dados do sistema RAI.

Figura 29 - Cria a função que formata a data para o padrão do SGBD referente ao sistema RAI

```
416  -- EXCLUI A FUNCAO CASO ELA EXISTA
417  drop function if exists getDataNascimentoBrToUsRAI;
418
419  -- CRIA A FUNCAO REPONSAVEL POR CONVERTER A DATA DE NASCIMENTO DO PADRAO PT_BR PARA O ACEITAVEL PELA BASE DE DADOS.
420  delimiter $$
421  create function getDataNascimentoBrToUsRAI(data_nascimento varchar(200)) returns datetime
422  begin
423      declare nova_data varchar(200);
424      set nova_data = null;
425
426      if (data_nascimento != '') then
427          set nova_data = STR_TO_DATE(SUBSTRING(TRIM(data_nascimento), 1, 17), '%d/%m/%Y %H:%i:%s');
428      end if;
429
430      return nova_data;
431  end $$
432  delimiter ;
```

Fonte: CRUZ; SANTOS, 2018.

12. Script SQL responsável por recuperar os dados correspondente a criação do arquivo .arff

Figura 30 - Script SQL responsável por gerar os valores formatados do arquivo .arff

```
SELECT
    CONCAT(
        '"',
        YEAR(data_fato), '","',
        getNomeMes(MONTH(data_fato)),'","',
    getNomeSemana(DAYOFWEEK(data_fato)), '","',
    getPeriodoDia(TIME(data_fato)), '","',
        sexo, '","',
        getNovaIdade(YEAR(NOW()) - YEAR(nascimento)),'"'
    ) AS data_values
FROM relatorios
WHERE YEAR(data_fato) = 2010;
```

Fonte: CRUZ; SANTOS, 2018.

ANEXOS

ANEXO A - Declaração de Disponibilização dos Dados.

SECRETARIA DE SEGURANÇA PÚBLICA E ADM. PENITENCIÁRIA
CORPO DE BOMBEIROS MILITAR
6ª SEÇÃO DO ESTADO-MAIOR GERAL

GOIÁS

DECLARAÇÃO

Declaro para os devidos fins que o 2º Sgt Elson Bento dos Santos, lotado no 3º Batalhão Bombeiro Militar, do Corpo de Bombeiros Militar do Estado de Goiás, solicitou a esta 6ª Seção do Estado Maior Geral informações de acidentes de trânsito do Sistema de Registro de Atendimento Integrado (RAI) para fins de pesquisa acadêmica, as quais serão disponibilizadas à medida que possível.

Goiânia, 20 de novembro de 2017.

Ricardo de Souza Oliveira – Ten QOC
Subchefe da BM/6