CENTRO UNIVERSITÁRIO DE ANÁPOLIS –UniEVANGÉLICA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

GUILHERME MACHADO DE OLIVEIRA
LUCAS DA SILVA PEREIRA

APLICAÇÃO DE MINERAÇÃO DE DADOS NA DESCOBERTA DE PADRÕES EM CRIMES DE FURTO E ROUBO DE VEÍCULOS NA CIDADE DE SÃO PAULO - SP

ANÁPOLIS
2020-01

GUILHERME MACHADO DE OLIVEIRA
LUCAS DA SILVA PEREIRA

APLICAÇÃO DE MINERAÇÃO DE DADOS NA DESCOBERTA DE PADRÕES EM CRIMES DE FURTO E ROUBO DE VEÍCULOS NA CIDADE DE SÃO PAULO - SP

Trabalho de Conclusão de Curso II apresentado como requisito parcial para a conclusão da disciplina de Trabalho de Conclusão de Curso II do curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – UniEVANGÉLICA.

Orientador(a): Prof. Aline Dayany de Lemos.

ANÁPOLIS
2020-01

GUILHERME MACHADO DE OLIVEIRA
LUCAS DA SILVA PEREIRA

APLICAÇÃO DE MINERAÇÃO DE DADOS NA DESCOBERTA DE PADRÕES EM CRIMES DE FURTO E ROUBO DE VEÍCULOS NA CIDADE DE SÃO PAULO - SP

Trabalho de Conclusão de Curso II apresentado como requisito parcial para a obtenção de grau do curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – UniEVANGÉLICA.

Aprovado(a) pela banca examinadora em ___ de _____ de 2020, composta por:

_____
Prof. Aline Dayany de Lemos
Orientador

_____
Prof. [nome do professor]

_____
Prof. [nome do professor]

RESUMO

Quanto maior o nível de urbanização de um local maior o risco de ocorrência de crimes contra o patrimônio. O objetivo deste trabalho foi aplicar técnicas de mineração de dados, para descoberta de conhecimentos sobre furtos e roubos de veículos na cidade de São Paulo, no período de 2016 a 2019, disponíveis na base de dados da Secretaria de Segurança Pública do estado de São Paulo. O desenvolvimento foi realizado com a utilização da ferramenta *Weka* seguindo a metodologia DCBD juntamente com a escolha de tarefas, técnicas e o algoritmo *Apriori*. Com a realização deste trabalho foi possível obter a compreensão de novas informações consolidadas sobre Segurança Pública na cidade de São Paulo, encontrando associações que podem ser utilizadas na solução do problema de furto e roubo de veículos. Espera-se que este trabalho possa auxiliar na tomada de decisões visando o bem-estar da população.

Palavras-chave: Banco de dados, Mineração de dados, Segurança pública, *Weka*.

ABSTRACT

The higher the level of urbanization in a place, the greater the risk of crimes against property. The objective of this work was to apply data mining techniques to discover knowledge about burglaries and robberies of vehicles in the city of São Paulo, in the period from 2016 to 2019, available in the database of the Secretariat of Public Security of the State of São Paulo. The development carried out using the Weka tool following the DCBD methodology together with the choice of tasks, techniques and the Apriori algorithm. With this work, it was possible to gain an understanding of new consolidated information on Public Security in the city of São Paulo, finding associations that could be use to solve the problem of burglaries and robberies of vehicles. It is hope that this work can assist in making decisions aimed at the well-being of the population.

Keywords*: Database, Data mining, public security, Weka.*

LISTA DE FIGURAS

Página

LISTA DE QUADROS

Página

LISTA DE ABREVIATURAS E SIGLAS

AIS	Área Integrada de Segurança
DCBD	Descoberta de Conhecimento em Bases de Dados
KDD	*Knowledge Discovery in Databases*
SSP	Secretaria de Segurança Pública
WEKA	*Waikato Environment for Knowledge Analysis*
JDBC	*Java Database Connectivity*
DDL	*Data Definition Language*
DML	*Data Manipulation Language*
SQL	*Structured Query Language*
SSP	Secretária de Segurança Pública
SP	São Paulo
B.O	Boletim de Ocorrência

SUMÁRIO

Página

1. INTRODUÇÃO

A legislação brasileira prevê nos artigos 155 à 157 do Código Penal (Decreto-Lei n° 2.848, de 7 de dezembro de 1940) os atos de furto e de roubo, sendo que ambos tratam-se, em síntese, na subtração de objeto móvel de outra pessoa sem o consentimento. A diferença entre essas modalidades criminosas encontra-se no fato de que na primeira inexiste o contato do agressor com a vítima (o proprietário do bem), enquanto que na segunda, além da subtração do bem, o agressor estabelece contato com a vítima, podendo utilizar de violência ou ameaça (LUZ, 2017).

Esses tipos de delitos estão diretamente relacionados ao patrimônio de caráter privado, e associados as características socioeconômicas de cada região. Conforme ressalta Souza (2017, p. 3) "[...] os crimes contra o patrimônio estão positivamente associados às variáveis socioeconômicas, uma vez que a sua mobilidade e a sua incidência aparecem em paralelo com o crescimento da renda per capita e com o nível de urbanização [...]".

Assim, quanto maior o nível de urbanização de um local maior o risco de ocorrência de crimes contra o patrimônio, sendo destacado por Souza (2017, p. 3) que "[...] os locais com o maior retorno para esse tipo de empreendimento delitivo – portanto, as áreas mais beneficiadas economicamente falando – são também as regiões de maior risco, uma vez considerados os crimes contra o patrimônio".

Na cidade de São Paulo, as estatísticas de dados relativos ao roubo de veículos apontam que no mês de agosto de 2017 foram registradas 5.465 ocorrências, em agosto de 2018 caíram para 4.822 e em agosto de 2019 registraram-se um total de 3.554 ocorrências. Essa diminuição no número deste tipo de crime deve-se a organização da polícia Militar, Civil e Técnico-Científica que resultou em 17.576 prisões, e a retirada de 1.034 armas de fogo (SÃO PAULO, 2019).

Ainda que os números apresentados pela Secretaria de Segurança Pública do Estado de São Paulo (SSP/SP) mostrem uma diminuição dos casos de furto e roubos de veículos no período de 2015 a 2017, há a necessidade de ferramentas de prevenção à esses tipos de crimes. Conforme a pesquisa nacional de vitimização realizada pelo Datafolha (2013, p. 65) o "Furto de veículos [...] tende a ser mais frequente, de um modo geral, entre os moradores das regiões Sudeste (10,4%), Sul (7,1%) e Centro-Oeste (7,0%), especialmente entre os que residem nos estados de São Paulo (10,7%), [...] e em Goiás (8,1%)".

A SSP/SP disponibiliza informações sobre furto e roubo de veículos em seu sistema web. Sendo assim, este trabalho trata da seguinte questão: A partir das bases de dados é possível identificar padrões de furtos e roubos de veículos na cidade de São Paulo?

1.1. Objetivos

1.1.1. Objetivo Geral

Identificar, através de mineração de dados, padrões de furtos e roubos de veículos na cidade de São Paulo, no período de 2016 a 2019.

1.1.2. Objetivos Específicos

- Obter base de dados de furto e roubo de veículos na cidade de São Paulo.
- Analisar técnicas para mineração de dados.
- Definir melhor técnica de mineração de dados para os dados obtidos.
- Tabular os dados de furtos e roubos de veículos da cidade de São Paulo.
- Estabelecer padrão de furtos e roubos de veículos da Cidade de São Paulo no período de 2016 a 2019.

1.2. Justificativa

Os crimes de furto e roubo tem um impacto social significativo nas vítimas. De acordo com Fraga (2015, p. 4) "[...]os crimes contra o patrimônio podem estimular um ônus social significativo, o indivíduo passa a ter o seu direito de ir e vir comprometido pelo risco de vitimização, altera hábitos rotineiros, que por vezes, podem limitar ou redefinir seu contato social [...]".

Outro ponto relevante é que o crime (ou a possibilidade de ocorrência do mesmo) é um fator de privação da liberdade da vítima (ou potencial vítima). Fraga (2015, p. 19) coloca que "[...] o crime afeta a liberdade do indivíduo de formas diversas: mitiga o direito de ir e vir do indivíduo, compromete o relacionamento humano, cercando as relações de desconfiança, o que afeta substancialmente a vida em comunidade [...]".

Logo, garantir a segurança em uma sociedade é promover a liberdade individual e assegurar os direitos básicos dos indivíduos. Segundo Fraga (2015, p. 20) “A segurança, além de um direito humano é também um exemplo de liberdade. Exercer amplamente a segurança ainda é um desafio comum aos cidadãos, dado o nível de criminalidade e violação de demais regras constitucionais [...]”.

Nesse sentido, utilizar dados para melhorar a gestão de segurança pública, garantir a segurança à população e prevenir a ocorrência de crimes contribui para uma sociedade que assegura os direitos fundamentais dos indivíduos. Perón (2016, p. 8-9) destaca que

> Conforme os departamentos de polícia adotam sistemas de gestão massiva de dados e registros, a sua capacidade de agrupar e analisar dados sobre crime e desordem se amplia muito. Ainda que esse movimento possa ser entendido como uma mera forma de “mapeamento”, muito comum em investigações criminais ao longo dos últimos anos, a sua particularidade reside na possibilidade de construir modelos estatísticos e de geo-referenciamento, a partir de análise massiva de dados públicos, e seu cruzamento com plataformas de dados criminais, capazes de classificar grupos de indivíduos, e apontar padrões de criminalidade futura [...].

Assim, a partir de uma base de dados é possível desenvolver padrões de ocorrência (e da probabilidade de ocorrer) crimes, além de mostrar as características comuns a cada tipo de crime ocorrido. Auxiliando, portanto, no desenvolvimento de políticas públicas voltadas a segurança, bem como a gestão de segurança em determinada localidade. Neste ponto, a informatização desse processo contribui para gestão da informação e assertividade das políticas de segurança. Como ressalta Araujo e Maciel (2018, p. 160)

> [...] as regras de associação extraídas dos algoritmos aplicados às bases de dados, podem ser relevantes para pesquisa e análise em relação ao combate a crimes. Os resultados gerados dificilmente seriam encontrados sem mineração de dados, por conta do tamanho e diversidade dos registros de ocorrências criminais [...].

Deste modo, a mineração de dados proporciona aos gestores de segurança pública informações privilegiadas, que haveria uma dificuldade maior de serem geradas sem a utilização desse recurso. Araujo e Maciel (2018, p. 166-167) expõe que os resultados da mineração de dados

> [...] mostraram que as predições geradas tiveram um valor de Confiança médio maior que 57%, confirmando a relevância das informações, sendo possível através delas, que a Polícia atue de forma mais efetiva no combate e prevenção de crimes, pois eles saberão por exemplo quais dias ou períodos precisarão intensificar a frota de policiais nas ruas, assim como determinada AIS [Área Integrada de Segurança] de que apresenta maior ocorrência de crimes precisa de maior prioridade nas investigações.

Portanto, o uso de mineração de dados neste trabalho procurou estabelecer os padrões de furto e roubo de veículos na cidade de São Paulo, utilizando a base de dados disponibilizada pela SSP/SP do período de 2016-2019, foram definidas as características, locais, períodos, e tipos de veículos suscetíveis a maior ocorrência desses tipos de crimes.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Furto e Roubo

Os crimes de furto e roubo estão previstos no código penal brasileiro, onde são caracterizadas, tipificadas e estabelecidas as penas aos indivíduos que praticam tais crimes. O Decreto-Lei nº 2.848, de 7 de Dezembro de 1940 (BRASIL, 1940) descreve nos artigos 155 a 156 o crime de furto, colocando que:

> Furto
> Art. 155 - Subtrair, para si ou para outrem, coisa alheia móvel:
> Pena - reclusão, de um a quatro anos, e multa.
> § 1º - A pena aumenta-se de um terço, se o crime é praticado durante o repouso noturno.
> § 2º - Se o criminoso é primário, e é de pequeno valor a coisa furtada, o juiz pode substituir a pena de reclusão pela de detenção, diminuí-la de um a dois terços, ou aplicar somente a pena de multa.
> § 3º - Equipara-se à coisa móvel a energia elétrica ou qualquer outra que tenha valor econômico.
> Furto qualificado
> § 4º - A pena é de reclusão de dois a oito anos, e multa, se o crime é cometido:
> I - com destruição ou rompimento de obstáculo à subtração da coisa;
> II - com abuso de confiança, ou mediante fraude, escalada ou destreza;
> III - com emprego de chave falsa;
> IV - mediante concurso de duas ou mais pessoas.
> § 4º- A pena é de reclusão de 4 (quatro) a 10 (dez) anos e multa, se houver emprego de explosivo ou de artefato análogo que cause perigo comum
> § 5º - A pena é de reclusão de três a oito anos, se a subtração for de veículo automotor que venha a ser transportado para outro Estado ou para o exterior.
> § 6º - A pena é de reclusão de 2 (dois) a 5 (cinco) anos se a subtração for de semovente domesticável de produção, ainda que abatido ou dividido em partes no local da subtração.
> § 7º - A pena é de reclusão de 4 (quatro) a 10 (dez) anos e multa, se a subtração for de substâncias explosivas ou de acessórios que, conjunta ou isoladamente, possibilitem sua fabricação, montagem ou emprego.
> Furto de coisa comum
> Art. 156 - Subtrair o condômino, co-herdeiro ou sócio, para si ou para outrem, a quem legitimamente a detém, a coisa comum:
> Pena - detenção, de seis meses a dois anos, ou multa.
> § 1º - Somente se procede mediante representação.
> § 2º - Não é punível a subtração de coisa comum fungível, cujo valor não excede a quota a que tem direito o agente.

Assim, o furto se caracteriza pela subtração de patrimônio móvel que pertence a outro indivíduo. Esse tipo de crime pode ainda ser qualificado, e ter a pena aumentada, em casos de

agravantes. Como no objeto deste estudo, caso o veículo venha a ser transportado para outro Estado, ou mesmo país, a pena é aumentada.

Quanto ao crime de roubo, o Decreto-Lei nº 2.848, de 7 de Dezembro de 1940 (BRASIL, 1940) prevê no artigo 157 que:

> Roubo
> Art. 157 - Subtrair coisa móvel alheia, para si ou para outrem, mediante grave ameaça ou violência a pessoa, ou depois de havê-la, por qualquer meio, reduzido à impossibilidade de resistência:
> Pena - reclusão, de quatro a dez anos, e multa.
> § 1º - Na mesma pena incorre quem, logo depois de subtraída a coisa, emprega violência contra pessoa ou grave ameaça, a fim de assegurar a impunidade do crime ou a detenção da coisa para si ou para terceiro.
> § 2º - A pena aumenta-se de 1/3 (um terço) até metade:
> I – (revogado);
> II - se há o concurso de duas ou mais pessoas;
> III - se a vítima está em serviço de transporte de valores e o agente conhece tal circunstância.
> IV - se a subtração for de veículo automotor que venha a ser transportado para outro Estado ou para o exterior;
> V - se o agente mantém a vítima em seu poder, restringindo sua liberdade.
> VI – se a subtração for de substâncias explosivas ou de acessórios que, conjunta ou isoladamente, possibilitem sua fabricação, montagem ou emprego.
> § 2º- A pena aumenta-se de 2/3 (dois terços):
> I – se a violência ou ameaça é exercida com emprego de arma de fogo;
> II – se há destruição ou rompimento de obstáculo mediante o emprego de explosivo ou de artefato análogo que cause perigo comum.
> § 3º - Se da violência resulta:
> I – lesão corporal grave, a pena é de reclusão de 7 (sete) a 18 (dezoito) anos, e multa;
> II – morte, a pena é de reclusão de 20 (vinte) a 30 (trinta) anos, e multa.

Diferentemente do crime de furto, no crime de roubo existe o contato do agressor com a vítima, onde o mesmo, além de subtrair o bem móvel, utiliza de agressão, violência, ou ameaça à vítima. Assim como o crime de furto, o crime de roubo possui agravantes que aumentam a pena. No caso de roubo de veículos, a pena é aumentada em 1/3 caso este seja transportado para outro Estado ou país. Outros pontos a serem observados são: o parágrafo 2º inciso V e o parágrafo 3º, onde os agravantes estão relacionados a violência exercida contra a vítima.

Com dados e informações a respeito tanto do crime de roubo como do crime de furto é possível elaborar estratégias para a gestão de segurança, com a finalidade de proteger um potencial vítima de um agressor, bem como assegurar o direito de propriedade aos indivíduos.

2.2. Dado, Informação e conhecimento

Os termos que designam os dados, mesmo que pareçam similares, não o são, tantos os dados, quanto a informação e o conhecimento podem ser utilizados no mesmo contexto de aplicabilidade. Deve-se ter o cuidado em distinguir a diferença entre eles, principalmente quando se tratar do uso entre dado e informação, ou informação e conhecimento (STEZER, 2015).

Dados são um conjunto de informações sintáticas ainda sem um significado lógico. Mas, a partir do momento em que se atribuí um significado que lhe confira organização, passam a ter valor a quem os recebe (SILVA; PERES; BOSCARIOLI, 2016).

É importante conhecer o tipo dos dados com o quais se pretende trabalhar, somente assim é possível planejar o caminho metodológico que será aplicado a cada formato do projeto. Sendo que, pode-se escolher entre dois tipos de categorização de dados: quantitativos e qualitativos. Os dados quantitativos são a representação em forma de dados numéricos, podendo ser diferenciados em discretos e contínuos. Tratando-se de dados qualitativos esses são expressos de forma nominal (CAMILO; SILVA, 2009).

Ressalta-se que existe uma diferenciação básica entre os conceitos de dados e informação. Os dados são os conteúdos armazenados ainda brutos sem sentido quando em estado de isolamento. As informações são construídas através do agrupamento dos dados de forma organizada e que passam a fazer sentido e sirvam como fonte de conhecimento (SILVA, 2015).

Ao atribuir-se uma semântica, esses dados passam a ser entendidos para a quem deles esteja fazendo uso. A partir desse momento o processo de transformação dos dados passa a ser conhecido por informação. O conhecimento é então concebido no momento em que o usuário passa a ter noção das informações, permitindo assim que sejam tomadas decisões baseadas nas análises desenvolvidas pelos homens (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

O processo de descoberta de conhecimento em bases de dados pode ser definida como: “o processo não trivial de identificar nos dados, padrões válidos, novos, potencialmente utilizáveis e compreensíveis” (FAYYAD et al., 1996, p. 40). Logo, o conhecimento pode ser construído a partir das ações de padronizar e identificar dados e informações em um banco de dados.

2.3. Banco de dados

Banco de dados é uma coleção de informações variadas formando um sentido lógico, que possua uma estrutura organizada de dados que permitam aos usuários o acesso e extração dos conteúdos (SILVA, 2015). Na visão de Kuhnen (2016, p. 10):

> [...] um banco de dados é uma coleção de dados inter-relacionados, representando informações sobre um domínio específico, ou seja, sempre que for possível agrupar informações que se relacionam e tratam de um mesmo assunto, posso dizer que tenho um banco de dados.

Os bancos de dados têm a capacidade de armazenar vários tipos de informações. Para que essas informações tenham o tratamento adequado e possam ser utilizadas da maneira necessária é preciso fazer a coleta digital e automática de dados. Desta forma pode-se entender que basicamente um banco de dados é um sistema computadorizado de armazenagem de registros (ELMASRI; NAVATHE, 2011).

Um sistema de banco de dados se atualiza à medida que os acessos de usuários vão acontecendo e novas informações são adicionadas, desde que estas sejam condizentes com o significado ao indivíduo ou a uma organização no qual o sistema esteja sendo utilizado dentro do processo de troca de informações (ELMASRI; NAVATHE, 2011).

2.4. Mineração de Dados

A Mineração de Dados é um processo que faz a análise de um banco de dados, e pode ser feito de forma semiautomática para obter padrões em seus dados. Com a descoberta de conhecimento, utilizando a inteligência artificial ou a análise estática, a mineração de dados tenta descobrir padrões e regras a partir desses dados (CASTRO; FERRARI, 2016).

A Mineração de Dados pode ser dividida de três maneiras: a estatística, o aprendizado de máquina e o banco de dados. A estatística é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam úteis e compreensíveis ao dono dos dados. O aprendizado de máquina é um passo no processo de Descoberta de Conhecimento, em que é feito a análise dos dados e a aplicação de algoritmos, produzindo um conjunto de certos padrões de dados. A Mineração de dados em banco de dados une técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de

dados e visualização, para extrair informações de grandes bases de dados. (CAMILO; SILVA, 2009).

Sendo assim, a mineração de dados visa desenvolver um processo de busca por anomalias, padrões e correlações dentro dos conjuntos de dados prevendo desta forma os resultados. Algumas estruturas que compõe as formas da mineração de dados são apresentadas no Quadro 1 (HAN; KAMBER, 2006):

Quadro 1 – Diversas Formas de Mineração de Dados

Tipo de mineração	Descrição
Fluxo de Dados	Algumas aplicações trafegam um volume alto de dados, temporalmente ordenados, voláteis e potencialmente infinitos. Minerar estas informações após terem sido armazenadas é uma tarefa inviável. Ao invés disso, a mineração ocorre na medida em que os dados são lidos.
Séries Temporais	Armazenam informações de certo evento em um intervalo de tempo definido.
Grafos	São muito importantes na modelagem de estruturas complexas, como circuitos, imagens, proteínas, redes biológicas e redes sociais. Variações de algoritmos tradicionais e novos algoritmos têm sido desenvolvidos para esse fim.
Relacionamentos	As redes sociais representam o relacionamento (link) entre as entidades envolvidas (similar a uma estrutura de grafos). Nas últimas décadas elas têm chamado muita atenção pela riqueza de padrões que podem ser extraídos.
Dados Multirelacionais	A Mineração de Dados Multirelacionais visa aplicar algoritmos que utilizam as estruturas originais das bases, sem a necessidade de uma conversão.
Objetos	Guardam os dados em forma de objetos (formados por um identificador, atributos e métodos).
Dados Espaciais	A mineração espacial visa identificar os padrões armazenados em dados de uma forma implícita.
Dados Multimídia	Armazenam dados em formato de áudio, vídeo, imagens, gráficos, texto. Servindo como um padrão de reconhecimento facial em imagens, ou apresentando uma proposta na geração de regras associativas de documentação textual escaneados.
Textos	Grande parte dos dados de uma instituição é armazenada de forma semiestruturada e não-estruturada, através de textos, e-mail, artigos, documentos (atas, memorandos, ofícios), etc. Mas por problemas de acúmulo, informações não confiáveis e baixa procura podem ocasionar um volume alto e dispensável, sendo resolvido pelos atos da mineração de dados.
Internet	A Mineração da Internet (ou Web Mining), consiste em minerar as estruturas de ligação, o conteúdo, os padrões de acesso, classificação de documentos, entre outras.

Fonte: Adaptado de Camilo e Silva (2009)

A mineração de dados é parte integrante de um processo mais amplo. O termo KDD, iniciais de *Knowledge Discovery in Databases* ou em português Descoberta de Conhecimento

em Bases de Dados (DCBD), refere-se a um conceito geral de processo de extração de conhecimento a partir de bases de dados, criado em 1989.

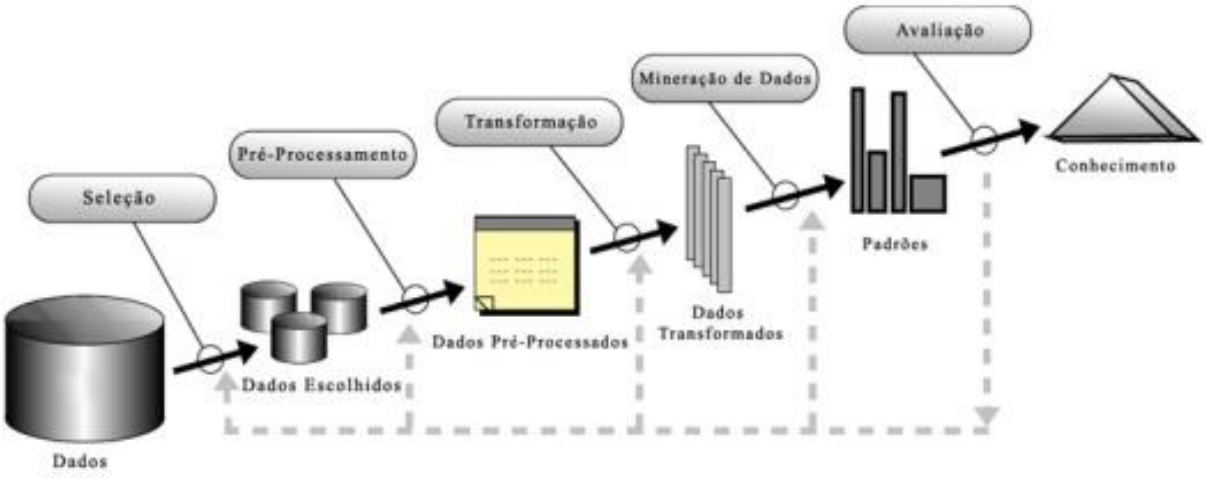
2.5. Metodologia DCBD

O termo KDD, ou DCBD, refere-se a um conceito geral de processo de extração de conhecimento a partir de bases de dados, criado em 1989. O DCBD é utilizado na extração de conhecimento em base de dados para encontrar padrões que tenha potencial utilidade e compreensão. Logo o DCBD é um processo de extração de conhecimento em base de dados, sejam elas convencionais, ou não (FRACALANZA, 2009).

O DCBD é composto de várias etapas, e é caracterizada como não trivial, interativo e iterativo, que identifica padrões válidos, úteis, compreensíveis e novos a partir de um conjunto macro de dados. Pode-se conceituar conhecimento útil como sendo a extração e derivação de dados que quando inter-relacionados em tarefas funcionais auxiliam em um processo de tomada de decisão. A possibilidade de controle por ação humana, onde se analisa e interpreta os dados, caracteriza-se esse método como interativo. Quanto a ser iterativo o método possibilita repetições em sua totalidade ou parcialmente dentro do conjunto do processo, podendo desta forma aperfeiçoar e chegar a resultados positivos (FAYYAD et al., 1996).

Esse processo é composto por etapas operacionais, conforme é ilustrado na Figura 1

Figura 1 – Etapas do processo de KDD/DCBD.



Fonte: Fayyad et al (1996) *apud* Camilo e Silva (2009).

O processo descrito na Figura 1 inicia ao estabelecer os objetivos e metas a serem seguidos no processo de busca do conhecimento, sendo preciso identificar os conhecimentos relevantes a serem tratados. Na sequência a etapa de seleção de dados é utilizada para identificar-se quais as informações constantes do banco de dados serão úteis para a análise e busca do conhecimento desejado (CARVALHO et al., 2009; FRACALANZA, 2009; CAMILO e SILVA, 2009).

Na etapa de pré-processamento de dados são realizadas as operações que visam modificar, corrigir ou retirar dados inconsistentes ou errados. Nessa etapa é realizada a coleta de informações necessárias para a modelagem e decisão das estratégias a serem aplicadas na mineração (CARVALHO et al., 2009; FRACALANZA, 2009; CAMILO e SILVA, 2009).

A fase de transformação dos dados tem como objetivo buscar por maneiras práticas de representação de dados e métodos reduzindo desta forma o número de variáveis importantes dentro do processo de coleta que esteja sendo realizado (CARVALHO et al., 2009; FRACALANZA, 2009; CAMILO e SILVA, 2009).

A próxima etapa é a mineração de dados, sendo considerado o principal processo de DCBD, pois é nesta etapa que se aplicam as técnicas e algoritmos específicos para extração do padrão e modelos. Ressalta-se que a técnica escolhida depende da tarefa a ser efetivada, estas tarefas podem ser: associação, classificação, regressão, *clusterização*, sumarização, detecção de desvios e sequências (CARVALHO et al., 2009; FRACALANZA, 2009; CAMILO e SILVA, 2009).

Por fim, a última etapa é a interpretação e avaliação do conhecimento descoberto, sendo realizado o processo no qual é verificado se os dados gerados são válidos na resolução do problema proposto (CARVALHO et al., 2009; FRACALANZA, 2009; CAMILO e SILVA, 2009).

2.6. Tarefas

A mineração de dados pode ser classificada pelas tarefas que consegue realizar, além disso essas tarefas, bem como os algoritmos utilizados, são definidas com base nos objetivos do estudo, a fim de obter uma resposta ao problema (CAMILO e SILVA, 2009; GALVAO e MARIN, 2009).

As tarefas mais comuns são descritas no Quadro 2, e apresentada resumidamente as principais características de cada tipo de tarefa, bem como sua utilização para mineração de dados de acordo com os objetivos a serem alcançados.

Quadro 2 – Tipos de tarefas

Tipo de tarefa	Característica
Classificação	Visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de 'aprender' como classificar um novo registro (aprendizado supervisionado).
Estimativa	É usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um. Após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por um novo consumidor
Associação	Consiste em identificar quais atributos estão relacionados. Apresentam a forma: SE atributo X ENTÃO atributo Y.
Predição	A tarefa de predição é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo.
Agrupamento	Visa identificar e aproximar os registros similares. Um agrupamento (ou *cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos.

Fonte: adaptado de Camilo e Silva (2009)

Algumas tarefas possuem atividades semelhantes como a classificação e o agrupamento, porém o método de cada tarefa difere-se, a classificação utiliza um aprendizado supervisionado (necessitando que cada registro seja categorizado), enquanto que o agrupamento utiliza um aprendizado não supervisionado (pois não necessita que os registros sejam previamente categorizados) (CAMILO e SILVA, 2009).

Outra tarefa similar é a predição que se assemelha tanto a classificação como a estimativa, porém a tarefa de predição tem como objetivo predizer um valor futuro, enquanto a estimativa determina o valor da próxima variável, e a classificação apenas categoriza um próximo registro (CAMILO e SILVA, 2009).

Já a associação é uma tarefa que identifica os atributos de acordo com o relacionamento com outros atributos, ou seja, na tarefa de associação são realizadas a identificação e o relacionamento dos atributos visando determinar e estabelecer padrões com base nessas relações (CAMILO e SILVA, 2009).

Essas tarefas são realizadas de acordo com o método (ou técnicas) de mineração de dados, buscando transformar os dados em conhecimento útil. As técnicas podem ser combinadas a fim de se encontrar um melhor resultado, (CAMILO e SILVA,2009; GALVAO e MARIN, 2009).

As tarefas podem ainda ser agrupadas de acordo com as técnicas que utilizam, sendo essas técnicas aprendizado supervisionado (preditivo) e não supervisionado (descritivo). As tarefas que utilizam a técnica de aprendizado supervisionado são: classificação e predição.

Enquanto que as tarefas que utilizam a técnica de aprendizado não supervisionado são: agrupamento, associação, estimativa (FACELI et. al., 2011).

2.7. Técnicas

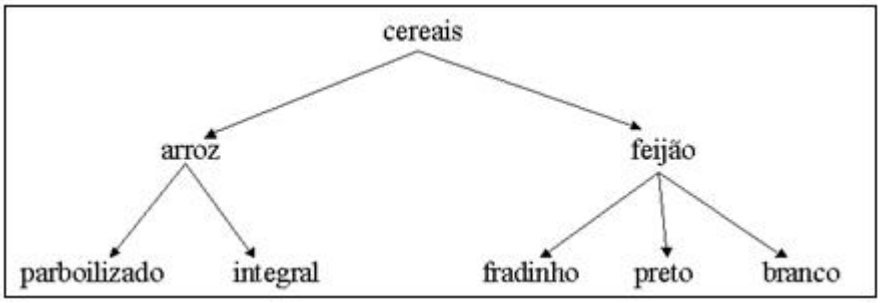
Habitualmente, os métodos de mineração de dados são divididos por duas técnicas, sendo elas, aprendizado supervisionado (preditivo) e não supervisionado (descritivo) (HAN; KAMBER, 2006).

A diferença entre os dois métodos concentra-se no fato de que os métodos de aprendizado não supervisionados não existem a necessidade de uma pré-categorização para os registros, ou seja, não é necessário um atributo alvo, enquanto os métodos supervisionados precisam dessa categorização (SELIYA; KHOSHGOFTAAR, 2007).

Os modelos supervisionados e não supervisionados, normalmente usam uma medida variável de similaridade entre os atributos. Durante o processo de mineração, diversas técnicas devem ser testadas e combinadas a fim de que comparações possam ser feitas e então a melhor técnica, ou se necessário uma combinação de técnicas, seja utilizada (MCCUE, 2007).

Dentre os diversos algoritmos que poderiam ser aplicados à este estudo, como as regras de associação, padrões sequenciais e os segmentos de dados (*clusters*), destacam-se, por sua aplicabilidade, as regras de associação. A Figura 2 apresenta um exemplo de regras de associação, trata-se da mineração de dados relativos à venda de cereais.

Figura 2 – Hierarquia de associação



Fonte: Gonçalves (2007).

Conforme o exposto na figura 2, ao se aplicar técnicas de mineração de dados com as regras de associação é possível identificar os itens de acordo com as características (classificação) de um item (ou produto) principal. Dessa forma, é possível, por exemplo,

associar as vendas de um determinado cereal de acordo com a característica específica deste item, com a finalidade de obter conhecimento de qual cereal tem maior saída.

As regras de associação têm como princípio encontrar elementos que se inter-relacionam com outros elementos em uma só transação. O termo transação serve para indicar quais os itens mais consultados em um processo de busca, assim se torna possível encontrar padrões de dados frequentes entre os vários conjuntos de dados. (VASCONCELOS; CARVALHO, 2004).

As regras de associação são caracterizadas por serem padrões descritivos que apresentam a probabilidade de itens serem encontrados dentro de uma transação que inclua outro conjunto de item (VASCONCELOS; CARVALHO, 2004). Para Mendes (2011) a regra de associação é muito útil para descobrir relacionamentos em bases de dados grandes, porém o processo pode gerar muito custo computacional e alguns padrões podem surgir por acaso.

Para se obter certo grau de confiança sobre os padrões descobertos, Tan e Kumar (2009) destacam dois fatores denominados suporte e confiança, onde:

- Suporte define a frequência da regra aparecer na base de dados analisada. Sua importância se obtém pelo fato de mostrar as regras sem interesse e identificar regras que oferecem pouca similaridade.
- Confiança define a frequência que um objeto Y surge em transações que possuem X. Mede a precisão da regra de associação feita, como pode ser entendido a regra de associação compreende a expressão se X então Y. Quanto maior a confiança melhor a regra.

A forma de medida do suporte e confiança segundo Itakura (2004, n. p.) é definida como:

- Suporte = Número de registros com X e Y / número total de registros.
- Confiança = Número de registros com X e Y / número de registros com X.

2.8. Algoritmos

Ao utilizar a mineração de dados são empregados algoritmos, ou seja, um número finito de etapas para a descoberta de conhecimento. Algoritmo é uma sequência de passos a serem seguidos na execução de um trabalho. Especificamente em informática a definição de algoritmo seria a de um conjunto de regras e procedimentos lógicos, que alcançam uma solução de uma

problemática através de procedimentos definidos, tendo como parâmetro um número finito de etapas (CAMPOS; ASCÊNSIO, 2003).

Na implementação das técnicas de mineração de dados os algoritmos recebem uma divisão quanto ao alcance do aprendizado que podem oferecer. Sendo que essa divisão se configura em quatro técnicas: Redes Neurais Artificiais; Algoritmos Genéticos; Árvores de Decisão e Descoberta de Regras de Associação (SILVA; PERES; BOSCARIOLI, 2016).

O Quadro 3 expõe as técnicas de algoritmos e exemplos de cada técnica:

Quadro 3 – Técnicas de Algoritmos

Técnica	Exemplos de Algoritmo
Redes Neurais Artificiais	*Back-Propagation* e *Multilayer Perceptron*
Algoritmos Genéticos	*Rule Envolver*, Algoritmo de *Hillis* e GA-*Nuggets*
Árvores de Decisão	*Classification and Regression Trees* (CART), *C4.5* e *ID-3*
Descoberta de Regras de Associação	*Apriori* e *FP-Growth*

Fonte: Vieira (2018, p. 38)

Dentre os modelos apresentados no Quadro 3 entre os algoritmos na descoberta de regras de associação o *Apriori* caracteriza-se por ser dividido em duas partes: a primeira seleciona todos os subconjuntos que podem ser utilizados em algumas das regras devendo ter um suporte acima do exigido pela análise em questão. A segunda fase do algoritmo *Apriori* gera as regras usando os resultados da primeira parte e tem por finalidade conceder maior segurança em relação aos resultados obtidos (SILVA FILHO; SILVA, 2013).

O *Apriori* realiza uma busca aprofundada nos dados e gera conjuntos padronizados, em que se mantêm os mais frequentes e elimina os dados pouco usados. Este algoritmo é utilizado por ser compatível com a ferramenta *Weka* que é apresentada a seguir.

2.9. *Software* de Mineração de Dados - *Orange* e *Weka*

O *Orange* é uma ferramenta da informática que possui código aberto e realiza a classificação, regressão e tarefas descritivas de dados e mineração visual de dados. Através de estruturas de nodos adicionados ao campo de fluxo executando cada um deles uma determinada tarefa previamente agendada. Também possível oferecer ao usuário o *feedback* para cada etapa, retornando-lhe os dados de entrada e de saída de cada uma (VITERBO et al., 2016).

No *Orange* é possível realizar a extração de dados através da programação visual ou *scripts Python*, como também é possível: explorar estatísticas, realizar *box plots* ou *scatter plots* e aprofundar dados com árvores de decisão, agrupamento hierárquico, *heatmaps* e projeções lineares. A interface se concentra na análise exploratória de dados e não na codificação dos mesmos. *Orange* possui interface *Machine Learning* e complementos de mineração de dados de fontes externas para execução de processamento de linguagem natural, mineração de texto, bioinformática, análise de rede e mineração de regras de associação (DALLAGASSA, 2014).

O *Weka* (*Waikato Environment for Knowledge Analysis*) foi criado pela universidade de *Waikato* da Nova Zelândia e se tornou uma da ferramenta muito utilizada na mineração de dados. O *Weka* é uma ferramenta que pode acessar diretamente a base de dados via JDBC (*Java Database Connectivity*) ou na própria ferramenta com arquivos específicos. Por ser extensível permite adicionar novos algoritmos ou aperfeiçoar suas funcionalidades (VALENTIN et al., 2009).

O *Weka* é uma ferramenta que pode ser utilizada para comparar o desempenho de vários algoritmos de mineração de dados dentro de um conjunto de algoritmos de classificação, regras de associação, pré-processamento e *clustering* (VALENTIN et al., 2009; MANHÃES et al., 2011).

Contendo maior número de referências disponíveis na *web*, sendo uma ferramenta utilizada em diferentes pesquisas acadêmicas desde cursos de graduação até doutorado, esta trabalha com o algoritmo *Apriori* e é disponibilizada de forma *open source*. A ferramenta *Weka* foi escolhida, pelos motivos mencionados, para o desenvolvimento do trabalho.

3. DESENVOLVIMENTO

Este trabalho utilizou como fonte de dados, a base de dados disponibilizada no sistema da SSP/SP. Após o levantamento bibliográfico sobre técnicas e tarefas de mineração de dados, foi escolhido a tarefa de Regra de Associação juntamente com a técnica *Apriori*, possibilitado a verificação de resultados em comparação com alguns algoritmos diferentes, como o *cobweb* de clusterização.

Esses resultados não se mostraram tão eficazes para este trabalho quanto resultados do algoritmo *Apriori*. Na realização da mineração de dados foram trabalhados diferentes relacionamentos entre os furtos e roubos, como, dia da semana, período do dia, bairro, a marca do veículo, etc. Na utilização da regra de associação dois conceitos foram utilizados, o suporte e a confiança, e com mudanças de valores, foram obtidos os resultados.

3.1. Seleção dos Dados

Ocorreu a coleta dos dados em formato planilha eletrônica em 48 arquivos, pois eram divididas entre furto e roubo para cada mês do ano em uma planilha. Contudo, foi feita a junção desses dados copiando os dados de cada planilha e colando no final de uma única planilha, totalizando 490.365 registros, ilustrados na Figura 3.

Figura 3 – Arquivo com junção dos dados

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
490347	2019	4047	4047/2019	30/09/2019 17:56:56	30/09/2019 18:	29/09/201	22:20	A NOITE	30/09/201	30/09/201	Desconhe	Não		AVENIDA	1000	JD SÃO LU	S.PAULO	SP	########
490348	2019	4926	4926/2019	30/09/2019 17:38:29	30/09/2019 18:	30/09/201	14:50	A TARDE	30/09/201	30/09/201	Desconhe	Não		RUA GENA	70	IGUATEMI	S.PAULO	SP	########
490349	2019	4909	4909/2019	30/09/2019 18:45:42	30/09/2019 18:	25/09/201	16:10	A TARDE	30/09/201	30/09/201	Desconhe	Não	4843/2019	RUA MITIN	126	CAMPO LI	S.PAULO	SP	########
490350	2019	6276	6276/2019	30/09/2019 20:04:00	30/09/2019 20:	26/09/201	21:30	A NOITE	30/09/201	30/09/201	Desconhe	Não	8579/2019	AVENIDA	900	SAUDE	S.PAULO	SP	########
490351	2019	7490	7490/2019	30/09/2019 20:38:49	30/09/2019 20:	28/09/201	19:30	A NOITE	30/09/201	30/09/201	Desconhe	Não		RUA MARI	74	SAO MATE	S.PAULO	SP	########
490352	2019	1233137	1233137/2	30/09/2019 20:52:32	30/09/2019 20:	30/09/201	18:40	A NOITE	30/09/201	30/09/201	Desconhe	Não		RUA ANTC	731	ITAQUERA	S.PAULO	SP	########
490353	2019	2925	2925/2019	30/09/2019 20:46:22	30/09/2019 21:	30/09/201	18:40	A NOITE	30/09/201	30/09/201	Desconhe	Não		RUA ANA	175	SAO LUCA	S.PAULO	SP	########
490354	2019	9285	9285/2019	30/09/2019 20:51:17	30/09/2019 21:	28/09/201	21:30	A NOITE	30/09/201	30/09/201	Desconhe	Não		AVENIDA	128	VILA CURU	S.PAULO	SP	########
490355	2019	9285	9285/2019	30/09/2019 20:51:17	30/09/2019 21:	28/09/201	21:30	A NOITE	30/09/201	30/09/201	Desconhe	Não		AVENIDA	128	VILA CURU	S.PAULO	SP	########
490356	2019	9285	9285/2019	30/09/2019 20:51:17	30/09/2019 21:	28/09/201	21:30	A NOITE	30/09/201	30/09/201	Desconhe	Não		AVENIDA	128	VILA CURU	S.PAULO	SP	########
490357	2019	9285	9285/2019	30/09/2019 20:51:17	30/09/2019 21:	28/09/201	21:30	A NOITE	30/09/201	30/09/201	Desconhe	Não		AVENIDA	128	VILA CURU	S.PAULO	SP	########
490358	2019	9287	9287/2019	30/09/2019 21:51:22	30/09/2019 21:	30/09/201	21:00	A NOITE	30/09/201	30/09/201	Desconhe	Não		RUA SARD	8	ITAIM PAU	S.PAULO	SP	########
490359	2019	5082	5082/2019	30/09/2019 21:59:08	30/09/2019 22:	30/09/2019		A NOITE	30/09/201	30/09/201	Desconhe	Não		RUA BOIC	168	ARTUR AL	S.PAULO	SP	########
490360	2019	7282	7282/2019	30/09/2019 21:12:00	30/09/2019 22:	30/09/201	08:40	PELA MAN	30/09/201	30/09/201	Desconhe	Não		AVENIDA	350	VILA MAR	S.PAULO	SP	########
490361	2019	7282	7282/2019	30/09/2019 21:12:00	30/09/2019 22:	30/09/201	08:40	PELA MAN	30/09/201	30/09/201	Desconhe	Não		AVENIDA	350	VILA MAR	S.PAULO	SP	########
490362	2019	7282	7282/2019	30/09/2019 21:12:00	30/09/2019 22:	30/09/201	08:40	PELA MAN	30/09/201	30/09/201	Desconhe	Não		AVENIDA	350	VILA MAR	S.PAULO	SP	########
490363	2019	7282	7282/2019	30/09/2019 21:12:00	30/09/2019 22:	30/09/201	08:40	PELA MAN	30/09/201	30/09/201	Desconhe	Não		AVENIDA	350	VILA MAR	S.PAULO	SP	########
490364	2019	7906	7906/2019	30/09/2019 22:14:52	30/09/2019 22:	21/09/201	16:00	A TARDE	30/09/201	30/09/201	Desconhe	Não	7652/2019	RUA PAPC	42	SAO RAFA	S.PAULO	SP	########
490365	2019	7907	7907/2019	30/09/2019 22:39:45	30/09/2019 23:	30/09/201	20:00	A NOITE	30/09/201	30/09/201	Desconhe	Não		AVENIDA	4285	IGUATEMI	S.PAULO	SP	########
490366	2019	7285	7285/2019	30/09/2019 23:33:50	30/09/2019 23:	29/09/201	13:20	A TARDE	30/09/201	30/09/201	Desconhe	Não		AVENIDA	3000	ITAIM BIB	S.PAULO	SP	########
490367																			
490368																			
490369																			

Fonte: OLIVEIRA; SILVA (2020).

Alguns registros obtidos apresentaram inconsistências em campos, que indicou-se, pelo valor nele contido, ser de preenchimento manual, possivelmente algum erro de digitação ao fazer o cadastro. No Quadro 4 contém o dicionário de dados das informações obtidas no sistema SSP/SP com suas respectivas variáveis, descrições e a categoria a qual pertence.

Quadro 4 – Dicionário de dados do sistema SSP/SP

Variável	Descrição	Categoria
ANO_BO	Ano do BO	Data
NUM_BO	Número do BO	Numérico
NUMERO_BOLETIM	Número do Boletim	Numérico
BO_INICIADO	Data e Hora do BO Iniciado	Data
BO_EMITIDO	Data e Hora do BO Emitido	Data
DATAOCORRENCIA	Data da Ocorrência	Data
HORAOCORRENCIA	Hora da Ocorrência	Data
PERIDOOCORRENCIA	Período da Ocorrência	da manhã, tarde, noite, madrugada e hora incerta
DATACOMUNICACAO	Data da Comunicação	Data
DATAELABORACAO	Data da Elaboração	Data
BO_AUTORIA	Autoria do BO	TEXTO
FLAGRANTE	Indica se Houve Flagrante	Sim, Não
NUMERO_BOLETIM_PRINCIPAL	Número do Boletim Principal	TEXTO
LOGRADOURO	Logradouro dos fatos	TEXTO
NUMERO	Número dos fatos	TEXTO
BAIRRO	Bairro da Ocorrência	TEXTO
CIDADE	Cidade do Fato	TEXTO
UF	Estado do Fato	TEXTO
LATITUDE	Latitude da Ocorrência	TEXTO
LONGITUDE	Longitude da Ocorrência	TEXTO
DESCRICAOLOCAL	Descrição do Local do Fato	TEXTO
EXAME	Exame	TEXTO
SOLUCAO	Solução	TEXTO
DELEGACIA_NOME	Delegacia responsável pelo registro	TEXTO
DELEGACIA_CIRCUNSCRICAO	Delegacia de Circunscrição	TEXTO
ESPECIE	Espécie	TEXTO
RUBRICA	Natureza jurídica da ocorrência	TEXTO
DESDOBRAMENTO	Desdobramentos jurídicos envolvidos na ocorrência	TEXTO
STATUS	Indica se é crime consumado ou tentado	Consumado, Tentado
NOMEPESSOA	Nome da Pessoa	TEXTO
TIPOPESSOA	Tipo da Pessoa	TEXTO
VITIMAFATAL	Indica se a Pessoa Relacionada é Vítima Fatal	TEXTO
RG	RG da Vítima	TEXTO
RG_UF	Estado do RG da Vítima	TEXTO
NATURALIDADE	Naturalidade da Vítima	TEXTO
NACIONALIDADE	Nacionalidade da Vítima	TEXTO
SEXO	Sexo da Vítima	Masculino, Feminino
DATANASCIMENTO	Data de Nascimento da Vítima	Data
IDADE	Idade da Vítima	TEXTO
ESTADOCIVIL	Estado Civil da Vítima	TEXTO
PROFISSAO	Profissão da Vítima	TEXTO

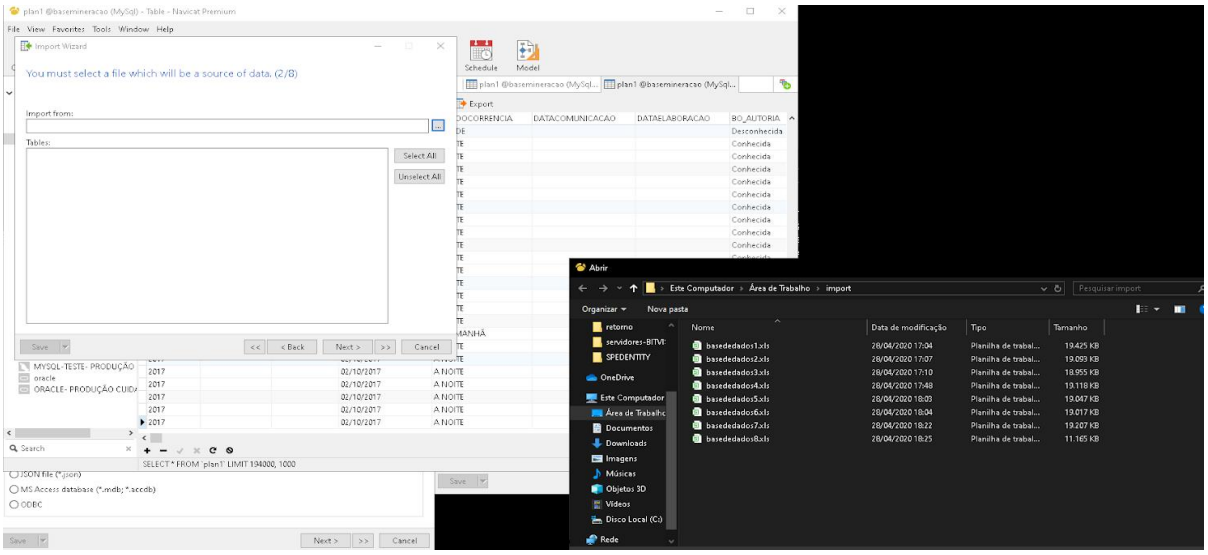
GRAUINSTRUCAO	Grau de Escolaridade do Envolvido	TEXTO
CORCUTIS	Cor da Pele	TEXTO
NATUREZAVINCULADA	Natureza Vinculada	TEXTO
TIPOVINCULO	Tipo de Vínculo	TEXTO
RELACIONAMENTO	Relacionamento Entre a Vítima e o Autor	TEXTO
PARENTESCO	Parentesco Entre a Vítima e o Autor	TEXTO
PLACA_VEICULO	Placa do Veículo	TEXTO
DESCR_COR_VEICULO	Cor do Veículo	TEXTO
DESCR_MARCA_VEICULO	Marca/Modelo	TEXTO
ANO_MODELO	Ano do Modelo	TEXTO
UF_VEICULO	Estado do Registro do Veículo	TEXTO
CIDADE_VEICULO	Cidade do Registro do Veículo	TEXTO
ANO_FABRICACAO	Ano de Fabricação	TEXTO
DESCR_TIPO_VEICULO	Tipo de Veículo	TEXTO
QUANT_CELULAR	Quantidade de Celular	TEXTO
MARCA_CELULAR	Marca do Celular	TEXTO

Fonte: OLIVEIRA; SILVA (2020).

3.2. Pré-Processamento dos Dados

O pré-processamento dos dados, foi feito no banco de dados para melhor performance ao invés da planilha. Para importar a quantidade de 490.365 registros da planilha para o banco de dados, foi necessário separar os dados em 8 arquivos distintos contendo 65.000 registros cada, ilustrados na Figura 4.

Figura 4 – Arquivos para importação na base de dados Mysql



Fonte: OLIVEIRA; SILVA, (2020).

Os 8 arquivos estavam no formato Pasta de Trabalho do Excel 97-2003(.xls), com os seguintes nomes: basededados1.xls, basededados2.xls, basededados3.xls, basededados4.xls, basededados5.xls, basededados6.xls, basededados7.xls, basededados8.xls, pois o formato Pasta de Trabalho do Excel(.xlsx) que pode conter mais dados em uma planilha, era corrompido ao importar.

Com a importação dos registros pronta, foi iniciado o pré-processamento dos dados, onde a limpeza ocorreu. Com a análise dos dados foram encontrados campos nulos e dados insuficientes, os quais foram removidos com o objetivo de aprimorar a eficiência do algoritmo de mineração, e posteriormente obter um resultado final mais eficiente. Os campos removidos encontram-se no Quadro 5, juntamente com o motivo.

Quadro 5 – Dados desconsiderados

Variáveis	Motivo
NUM_BO, NUMERO_BOLETIM, BO_INICIADO, BO_EMITIDO, DATACOMUNICACAO, DATAELABORACAO, BO_AUTORIA, NUMERO_BOLETIM_PRINCIPAL	Por se tratar de dados da criação do boletim, não interfere nos resultados da pesquisa.
LOGRADOURO, LATITUDE, LONGITUDE, NUMERO	Dados referentes à esses campos estavam incompletos e inconsistentes. Foram utilizados BAIRRO e DESCRICAOLOCAL para obter melhor resultados na pesquisa.
EXAME, SOLUCAO	Saber se foi feito algum exame na ocorrência e qual é a solução do boletim de ocorrência não se trata de dados que possam interferir nos resultados da pesquisa.
DELEGACIA_NOME, DELEGACIA_CIRCUNSCRICAO, ESPECIE, DESDOBRAMENTO	Saber dados da delegacia, espécie e desdobramento do crime não interfere nos resultados da pesquisa.
STATUS, NOMEPESSOA, TIPOPESSOA, VITIMAFATAL, RG, RG_UF, NATURALIDADE, NACIONALIDADE, SEXO, DATANASCIMENTO, IDADE, ESTADOCIVEL, PROFISSAO, GRAUINSTRUCAO, CORCUTIS, NATUREZAVINCULADA, TIPOVINCULO, RELACIONAMENTO, PARENTESCO	Campos referentes à dados pessoais das vítimas e agressores. A base de dados obtidas veio com esses campos vazios, e por questão de proteção de dados pessoais a pesquisa não pode trabalhar com esses dados.
PLACA_VEICULO	Dados deste campo é unitário para cada veículo, poderia gerar inconsistências no resultado, e portanto não foi utilizado na pesquisa.
UF_VEICULO, DESCR_TIPO_VEICULO, ANO_FABRICACAO	Dados referentes à esses campos estavam incompletos e inconsistentes.
QUANT_CELULAR, MARCA_CELULAR	Dados relativos ao patrimônio da vítima. Dados referentes a esses campos estavam vazios. Dados não interfeririam no resultado da pesquisa.

Fonte: OLIVEIRA; SILVA (2020)

Foram encontrados registros duplicados da coluna RUBRICA contendo os seguintes dados: furto qualificado (art. 155, §4o.) - residência, localização/apreensão e entrega de veículo, lesão corporal culposa na direção de veículo automotor, furto (art. 155) - residência, Entrega de veículo localizado/apreendido, Roubo (art. 157) - carga. Mas apenas os dados: furto (art. 155) - veículo e roubo (art. 157) - veículo, eram referentes a este trabalho científico, estes registros se encontravam de forma duplicada, pois veículos que eram recuperados pela polícia tinha um novo registro, com dados da ocorrência igual mudando apenas a coluna RUBRICA. Posteriormente foi desenvolvido um *script* utilizando DML (*Data Manipulation Language*), para remoção desses registros (*script* no Apêndice A, figura 15).

3.3. Transformação dos Dados

Ao concluir as etapas de seleção e pré-processamento dos dados, observou-se que foram recebidos um total de 290.029 mil registros, com as informações relevantes para a pesquisa. Em seguida foi realizada a transformação de dados para obter os resultados da pesquisa. Utilizando o MySQL *Workbench* foi possível fazer as alterações necessárias, e também analisar os campos que iriam ser utilizados na pesquisa, expostos no Quadro 6.

Quadro 6 – Dados considerados

Variável	Descrição
ANO_BO	Ano da ocorrência.
DATAOCORRENCIA	Data da ocorrência.
PERIODOOCORRENCIA	Período do dia da ocorrência.
FLAGRANTE	Flagrante.
BAIRRO	Bairro da ocorrência.
DESCRICAOLOCAL	Descrição do local.
RUBRICA	Descrição ao ato infracional da ocorrência.
CIDADE_VEICULO	Cidade do veículo da ocorrência.
DESCR_COR_VEICULO	Cor do veículo da ocorrência.
DESCR_MARCA_VEICULO	Modelo do veículo da ocorrência.
ANO_MODELO	Ano do veículo da ocorrência.

Fonte: OLIVEIRA; SILVA, 2020

Com a análise das colunas CIDADE_VEICULO e ANO_MODELO, foram encontrados registros incorretos como por exemplo: ANO_MODELO contendo data menor que 1940 e CIDADE_VEICULO com valor nulo, que não poderiam simplesmente serem deletados, pois se deletado a pesquisa se tornaria inviável pelo número de dados removidos. Sendo assim para

resolver esse problema, foi desenvolvido um *script* utilizando DML (*Data Manipulation Language*) para substituição destes registros para sinais de interrogação (?). Dessa forma o algoritmo *Apriori* pode ler os campos com esse sinal e ignorar durante a mineração dos dados, para não haver interferência de dados inválidos nas associações, esses dados são considerados faltante, encontra-se no Apêndice A o *script* para a resolução desse problema (figura 14).

Também foram identificados registros com acentuação, sendo assim, para a execução da mineração de dados houve a necessidade de remoção dos acentos. Utilizando o MySQL *Workbench* e DML (*Data Manipulation Language*), foi realizado o *script* para a remoção, disponível no Apêndice A (figura 13).

Após realizar essas alterações, os arquivos apresentava-se no formato .sql, no entanto o formato do arquivo para a manipulação no *Weka* é no formato .arff. Para a transformação do arquivo foram efetuadas as seguintes etapas: com o MySQL *Workbench* gerou-se o arquivo .csv, que foi importado para o *Weka*; logo em seguida gerou-se o arquivo .arff na própria ferramenta *Weka*, possibilitando que os dados fossem manipulados.

O arquivo .arff possui 3 anotações: @relation (nome do conjunto de dados), @attribute (nome do atributo seguido do seu tipo) e @date(logo em seguida é a inserção dos dados). Considerando essa característica do formato, foram realizadas algumas alterações do tipo de dados nos atributos: ANO_BO, DATAOCORRENCIA e ANO_MODELO que estava respectivamente como numérico, date e numérico como mostrado na figura 5.

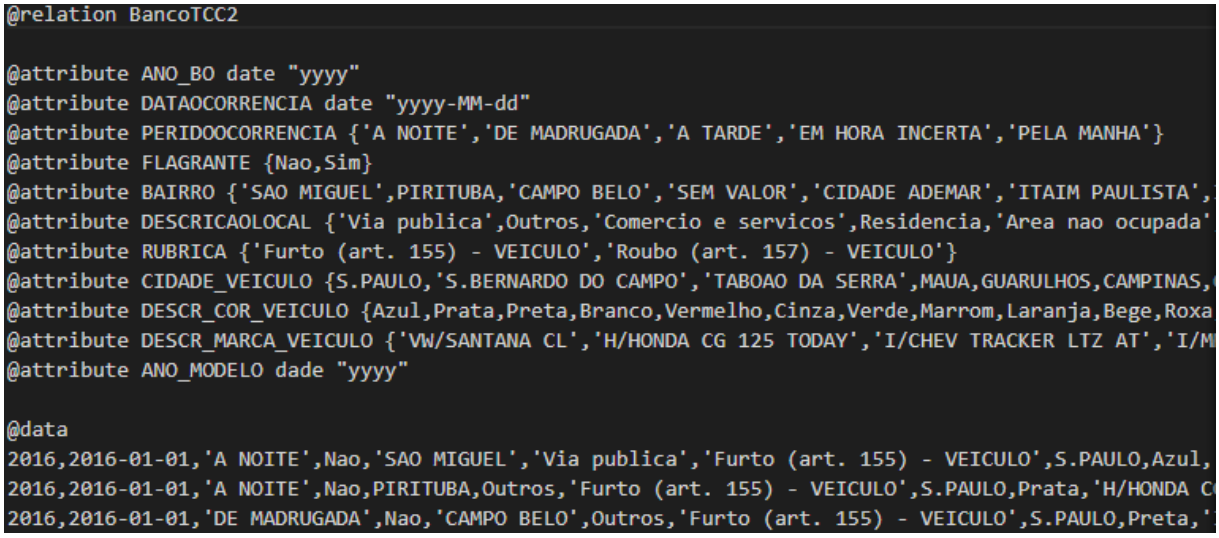
Figura 5 – Arquivo no formato .arff

```
@relation BancoTCC2

@attribute ANO_BO numeric
@attribute DATAOCORRENCIA {2016-01-01,2016-02-01,2016-03-01,2016-04-01,2016-05-01,2015-05-12,2016-06
@attribute PERIDOOCORRENCIA {'a noite','de madrugada','a tarde','em hora incerta','pela manha'}
@attribute FLAGRANTE {Sim}
@attribute BAIRRO {'sao miguel',pirituba,'campo belo','sem valor','cidade ademar','itaim paulista',i
@attribute DESCRICAOLOCAL {'Via publica','Comercio e servicos',Residencia,'Area nao ocupada','Lazer
@attribute RUBRICA {'Furto (art. 155) - VEICULO','Roubo (art. 157) - VEICULO'}
@attribute CIDADE_VEICULO {s.paulo,'s.bernardo do campo','taboao da serra',maua,guarulhos,campinas,c
@attribute DESCR_COR_VEICULO {Azul,Prata,Preta,Branco,Vermelho,Cinza,Verde,Marrom,Laranja,Bege,Roxa,
@attribute DESCR_MARCA_VEICULO {'vw/santana cl','h/honda cg 125 today','i/chev tracker ltz at','i/mm
@attribute ANO_MODELO numeric

@data
2016,2016-01-01,'a noite',?,'sao miguel','Via publica','Furto (art. 155) - VEICULO',s.paulo,Azul,'vw
2016,2016-01-01,'a noite',?,pirituba,?,'Furto (art. 155) - VEICULO',s.paulo,Prata,'h/honda cg 125 to
2016,2016-01-01,'de madrugada',?,'campo belo',?,'Furto (art. 155) - VEICULO',s.paulo,Preta,'i/chev t
```

Fonte: OLIVEIRA; SILVA (2020)

Após feitas as alterações dos tipos de dado nos atributos ANO_BO, DATAOCORRENCIA e ANO_MODELO para tipo date o arquivo ficou como o ilustrado na figura 6.

Figura 6 – Alterações do tipo de dado no arquivo .arff



```
@relation BancoTCC2

@attribute ANO_BO date "yyyy"
@attribute DATAOCORRENCIA date "yyyy-MM-dd"
@attribute PERIDOOCORRENCIA {'A NOITE','DE MADRUGADA','A TARDE','EM HORA INCERTA','PELA MANHA'}
@attribute FLAGRANTE {Nao,Sim}
@attribute BAIRRO {'SAO MIGUEL',PIRITUBA,'CAMPO BELO','SEM VALOR','CIDADE ADEMAR','ITAIM PAULISTA',
@attribute DESCRICAOLOCAL {'Via publica',Outros,'Comercio e servicos',Residencia,'Area nao ocupada'
@attribute RUBRICA {'Furto (art. 155) - VEICULO','Roubo (art. 157) - VEICULO'}
@attribute CIDADE_VEICULO {S.PAULO,'S.BERNARDO DO CAMPO','TABOAO DA SERRA',MAUA,GUARULHOS,CAMPINAS,
@attribute DESCR_COR_VEICULO {Azul,Prata,Preta,Branco,Vermelho,Cinza,Verde,Marrom,Laranja,Bege,Roxa
@attribute DESCR_MARCA_VEICULO {'VW/SANTANA CL','H/HONDA CG 125 TODAY','I/CHEV TRACKER LTZ AT','I/M
@attribute ANO_MODELO dade "yyyy"

@data
2016,2016-01-01,'A NOITE',Nao,'SAO MIGUEL','Via publica','Furto (art. 155) - VEICULO',S.PAULO,Azul,
2016,2016-01-01,'A NOITE',Nao,PIRITUBA,Outros,'Furto (art. 155) - VEICULO',S.PAULO,Prata,'H/HONDA C
2016,2016-01-01,'DE MADRUGADA',Nao,'CAMPO BELO',Outros,'Furto (art. 155) - VEICULO',S.PAULO,Preta,'
```

Fonte: OLIVEIRA; SILVA, 2020

3.4. Mineração de Dados

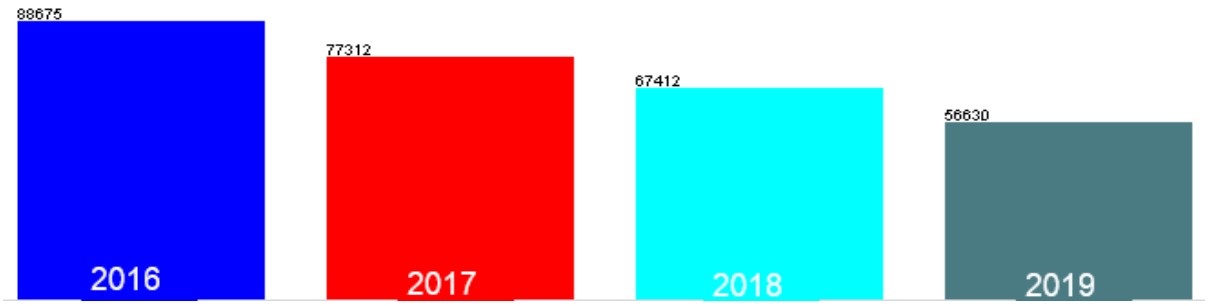
3.4.1. Análise superficial dos resultados

Após inserir os dados na ferramenta *Weka* foi possível analisar alguns gráficos que foram gerados automaticamente. O primeiro gráfico analisado foram as ocorrências por ano no período de 2016 a 2019, ilustrado na figura 7.

No período de 2016 a 2019 o índice de furto e roubos de veículos teve uma redução, sendo que em 2016 teve 88.675 registros e 2019 56.630 registros, havendo uma redução de 36% nesses tipos de crimes no período analisado. A média de redução dos crimes no período foi de 13,8% para cada ano.

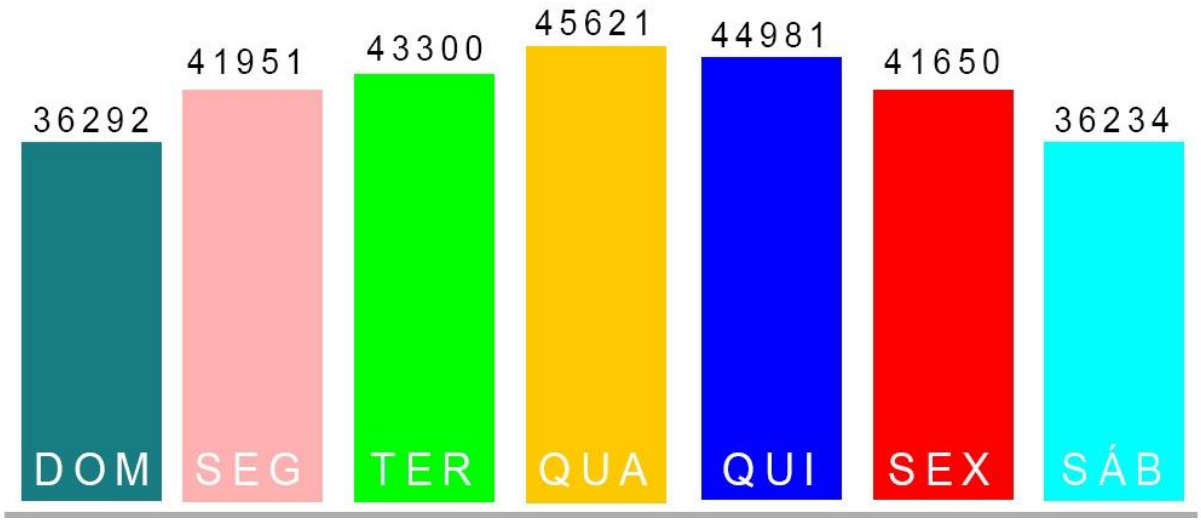
Foram analisados, também, os números de ocorrências dos crimes por dia da semana, sendo que, quarta-feira obteve o maior número de furtos e roubos, enquanto sábado registra o menor número. A diferença entre o dia com maior número de ocorrências e o menor é 20,6% de registros. O número de ocorrências por dia da semana pode ser observado na figura 8.

Figura 7 – Ocorrências de furto e roubo de veículos em São Paulo-SP 2016-2019



Fonte: OLIVEIRA; SILVA (2020)

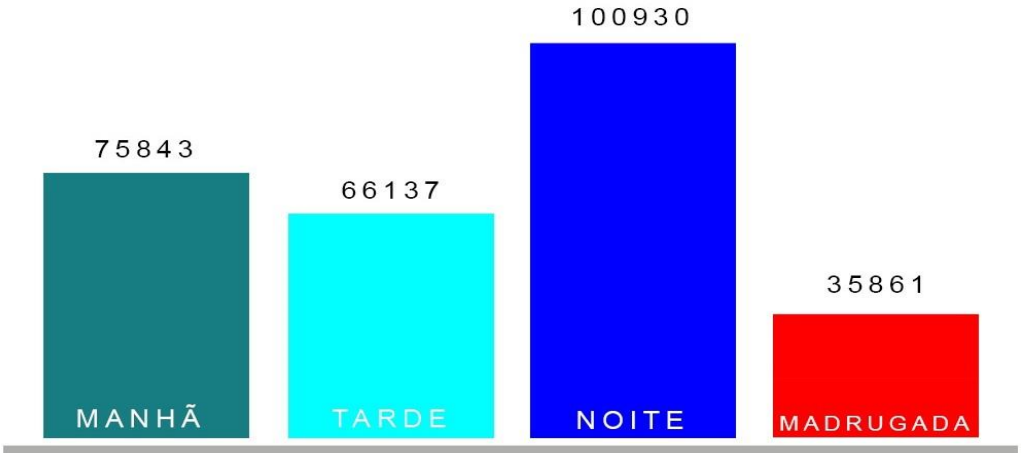
Figura 8 – Ocorrências de furto e roubo de veículos em São Paulo-SP por dia da semana 2016-2019



Fonte: OLIVEIRA; SILVA (2020)

Analisou-se, também, o período ocorrência dos crimes, sendo eles: manhã, tarde, noite, madrugada e hora incerta. Havia um total de 11.258 registros definidos como hora incerta, como esses dados impossibilitariam a obtenção de resultados mais preciso foi utilizado o filtro *RemoveFrequentValues* para a remoção deste dado. Após a remoção dos dados referentes a hora incerta obteve-se os números para os demais períodos, ilustrados na figura 9. O período da noite obteve o maior número de registro, enquanto que o menor número de registros foi encontrado no período da madrugada.

Figura 9 – Ocorrências de furto e roubo de veículos em São Paulo-SP por período do dia 2016-2019

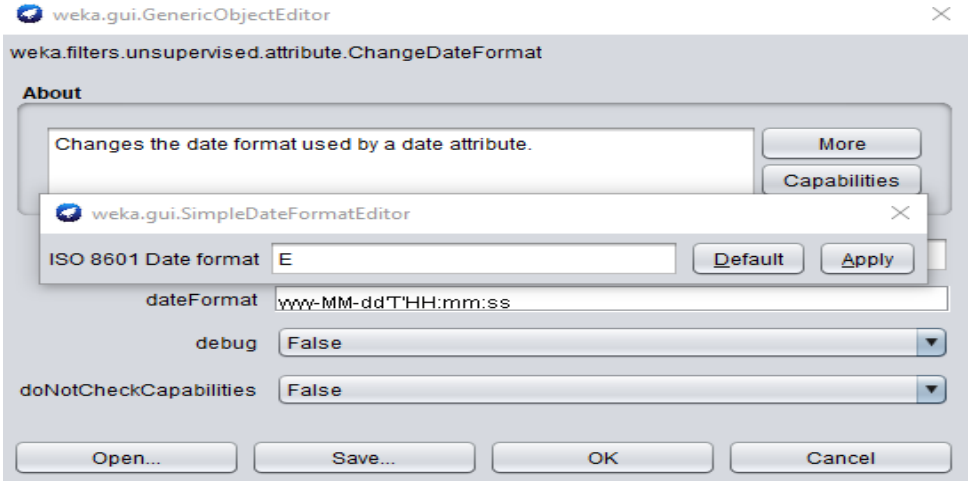


Fonte: OLIVEIRA; SILVA (2020)

3.4.2. Análise dos dados com aplicação do algoritmo *Apriori*

Após a abertura do arquivo no formato .arff na ferramenta *Weka*, foram utilizados filtros para transformação dos dados. Utilizando o filtro *ChangeDateFormat* foi possível mudar o formato dos atributos: ANO_BO e ANO_MODELO para o formato "Y", mostrando, a partir da alteração, apenas o ano do registro. Utilizando o mesmo filtro foi alterado o atributo DATAOCORRENCIA para o formato "E", com essa mudança as datas eram formatadas para seus respectivos dias da semana, conforme ilustrado pela figura 10.

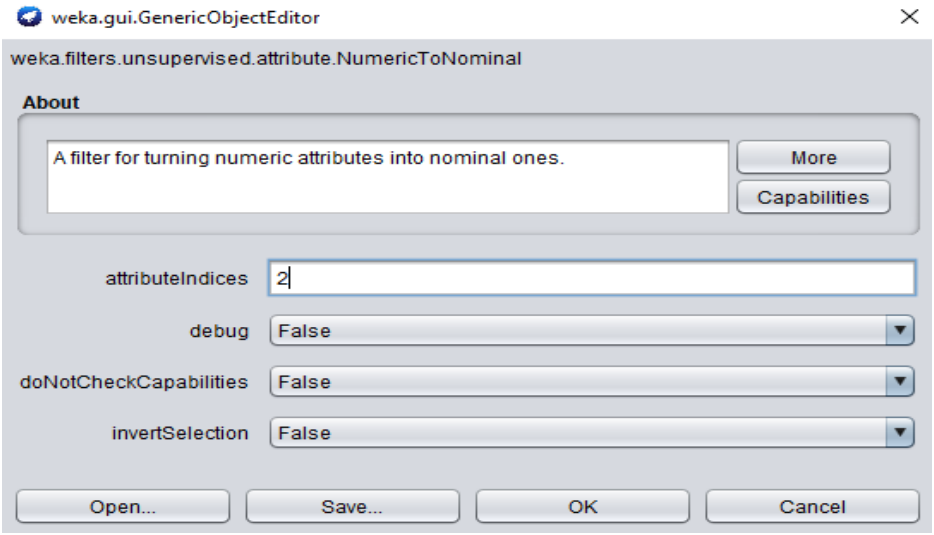
Figura 10 – Configurando o filtro *ChangeDateFormat* para dias da semana



Fonte: OLIVEIRA; SILVA (2020)

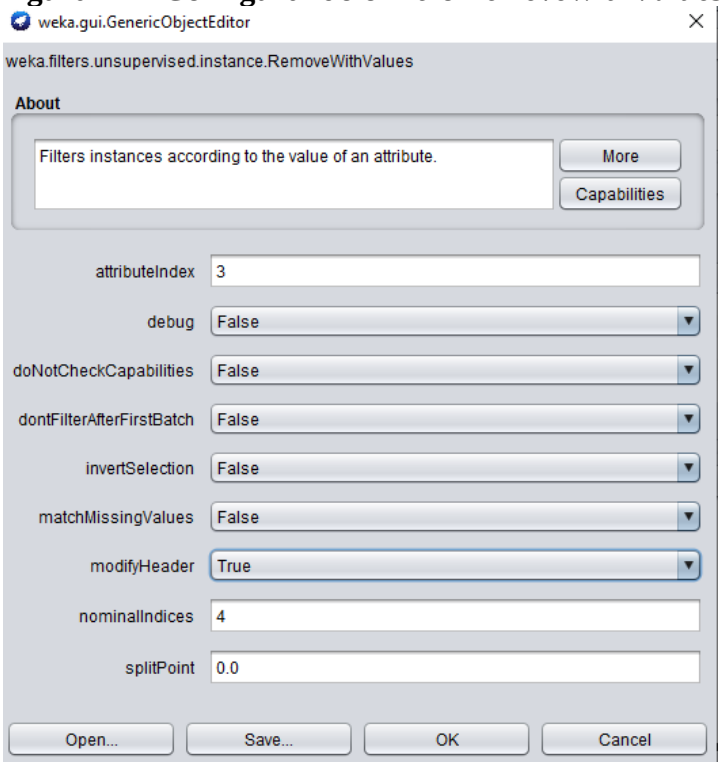
Após aplicar o filtro para as datas foi utilizado o filtro *NumericToNominal*, para os atributos citados anteriormente, transformando os dados desses atributos para o tipo nominal, como o ilustrado na Figura 11.

Figura 11 – Configurando o filtro *NumericToNominal*



Fonte: OLIVEIRA; SILVA (2020)

Figura 12 – Configurando o filtro *RemoveWithValues*



Fonte: OLIVEIRA; SILVA (2020)

Após a transformação dos dados de numéricos para nominal foram removidas instâncias de atributos que não continham dados satisfatórios. Por exemplo o atributo PERIDOOCORRENCIA possuía as respectivas instâncias: a noite, de madrugada, a tarde, pela manhã e hora incerta, os dados referentes a hora incerta não eram adequados para a pesquisa, pois esses dados não forneciam um período do dia certo, comprometendo, assim, o resultado final. Portanto, para obter melhores resultados foi utilizado o filtro *RemoveWithValues*, para a remoção dessa instância, como mostra a Figura 12.

Com os dados configurados, foram realizados testes e foram obtidos os seguintes cenários:

Cenário 1: na realização desse cenário foram desconsiderados os atributos ANO_BO, FLAGRANTE, DESCRICAOLOCAL, CIDADE_VEICULO e ANO_MODELO. Os resultados obtidos nesse cenário estão listados no Quadro 7, os resultados em negrito satisfazem ao objetivo da associação. A escolha desse cenário justifica-se pela porcentagem de confiança alcançada, atingiu-se 94% na melhor regra apresentada, e obteve-se resultados acima de 82%. Abaixo estão listados os números de instâncias, atributos (quais atributos), e o número de ciclos executados nesse cenário.

- *Instances*: 72860
- *Attributes*: 6
  - DATAOCORRENCIA
  - PERIDOOCORRENCIA
  - BAIRRO
  - RUBRICA
  - DESCR_COR_VEICULO
  - DESCR_MARCA_VEICULO
- *Minimum support*: 0.01 (729 *instances*)
- *Minimum metric <confidence>*: 0.5
- *Number of cycles performed*: 20

Quadro 7 – Resultado aplicação do algoritmo *Apriori* - Cenário 1

Nº	RESULTADO	CONF
1	BAIRRO=tucuruvi 1059 ==> RUBRICA=furto 960	94%

2	DESCR_MARCA_VEICULO=chevrolet/celta 1.0l lt 859 ==> RUBRICA=furto 765	89%
3	BAIRRO=vila mariana 1249 ==> RUBRICA=furto 1076	86%
4	BAIRRO=perdizes 895 ==> RUBRICA=furto 762	85%
5	PERIDOOCORRENCIA=pela manha DESCR_COR_VEICULO=Azul 876 ==> RUBRICA=furto 744	85%
6	DATAOCORRENCIA=qua PERIDOOCORRENCIA=pela manha DESCR_COR_VEICULO=Preta 920 ==> RUBRICA=furto 777	84%
7	PERIDOOCORRENCIA=pela manha DESCR_COR_VEICULO=Vermelho 2811 ==> RUBRICA=furto 2374	84%
8	BAIRRO=santo amaro 906 ==> RUBRICA=furto 756	83%
9	DATAOCORRENCIA=qua PERIDOOCORRENCIA=pela manha 3589 ==> RUBRICA=furto 2991	83%
10	BAIRRO=lapa 965 ==> RUBRICA=furto 800	83%
11	BAIRRO=pinheiros 1070 ==> RUBRICA=furto 887	83%
12	DATAOCORRENCIA=ter PERIDOOCORRENCIA=pela manha DESCR_COR_VEICULO=Preta 911 ==> RUBRICA=furto 753	83%
13	BAIRRO=tatuape 1103 ==> RUBRICA=furto 911	83%
14	**DATAOCORRENCIA=qui PERIDOOCORRENCIA=pela manha DESCR_COR_VEICULO=Preta 890 ==> RUBRICA=furto 732**	**82%**
15	**DATAOCORRENCIA=qua PERIDOOCORRENCIA=pela manha DESCR_COR_VEICULO=Prata 888 ==> RUBRICA=furto 730**	**82%**

Fonte: OLIVEIRA; SILVA (2020)

Cenário 2: para o segundo cenário, na etapa de pré-processamento, foram aplicados filtros removendo os atributos ANO_BO, FLAGRANTE, DESCRICAOLOCAL, CIDADE_VEICULO, DESCR_COR_VEICUL, RUBRICA e ANO_MODELO. O cenário 2 foi selecionado por apresentar resultados em relação a bairros e períodos em que os crimes ocorreram. Os resultados encontrados desse cenário encontram-se no Quadro 8. Esse cenário apresentou um nível de confiança máximo de 60%, os resultados destacados em negrito satisfazem ao objetivo da associação. Abaixo estão listados os números de instâncias, atributos (quais atributos), e o número de ciclos executados nesse cenário.

- *Instances*: 72860
- *Attributes*: 6
  DATAOCORRENCIA
  PERIDOOCORRENCIA
  BAIRRO
  DESCR_MARCA_VEICULO
- *Minimum support*: 0.02 (5575 *instances*)

- *Minimum metric <confidence>*: 0.5
- *Number of cycles performed*: 98

Quadro 8 – Resultado aplicação do algoritmo *Apriori* - Cenário 2

Nº	RESULTADO	CONF
1	**BAIRRO=ipiranga 3356 ==> PERIDOOCORRENCIA=pela manha 2007**	**60%**
2	**BAIRRO=vila mariana 2474 ==> PERIDOOCORRENCIA=pela manha 1438**	**58%**
3	BAIRRO=sao mateus 4795 ==> PERIDOOCORRENCIA=a tarde 2674	56%
4	BAIRRO=sapopemba 3525 ==> PERIDOOCORRENCIA=a tarde 1883	53%
5	BAIRRO=vila prudente 2833 ==> PERIDOOCORRENCIA=pela manha 1487	52%
6	BAIRRO=sao lucas 3029 ==> PERIDOOCORRENCIA=a tarde 1578	52%
7	BAIRRO=jabaquara 3276 ==> PERIDOOCORRENCIA=a tarde 1640	50%
8	BAIRRO=jabaquara 3276 ==> PERIDOOCORRENCIA=pela manha 1636	50%
9	BAIRRO=sao lucas 3029 ==> PERIDOOCORRENCIA=pela manha 1451	49%
10	BAIRRO=sapopemba 3525 ==> PERIDOOCORRENCIA=pela manha 1642	48%

Fonte: OLIVEIRA; SILVA (2020)

Cenário 3: foram utilizados nesse cenário filtros para remoção dos atributos ANO_BO, FLAGRANTE, DESCRICAOLOCAL, CIDADE_VEICULO, ANO_MODELO, DESCR_COR_VEICUL, BAIRRO e DESCR_COR_VEICU. Esse cenário foi escolhido por apresentar ótimos resultados em relação a modelos e tipos de veículos, e também por mostrar qual período contém o maior número de ocorrências para cada modelo e tipo de veículo. Os resultados do cenário 3 encontram-se no Quadro 9, foram destacados em negrito os resultados relativos ao tipo de veículo motocicleta. Esse cenário apresentou um nível de confiança máximo de 61%. Abaixo estão listados os números de instâncias, atributos (quais atributos), e o número de ciclos executados nesse cenário.

- *Instances*: 189099
- *Attributes*: 3

  DATAOCORRENCIA

  PERIDOOCORRENCIA

  DESCR_MARCA_VEICULO

- *Minimum support*: 0 (378 *instances*)
- *Minimum metric <confidence>*: 0.3
- *Number of cycles performed*: 998

Quadro 9 – Resultado aplicação do algoritmo *Apriori* - Cenário 3

Nº	RESULTADO	CONF
1	**DESCR_MARCA_VEICULO=honda/pcx 150 2042 ==> PERIDOOCORRENCIA=a noite 1033**	**61%**
2	**DESCR_MARCA_VEICULO=honda/cg 160 fan esdi 1252 ==> PERIDOOCORRENCIA=a noite 559**	**60%**
3	DESCR_MARCA_VEICULO=i/chevrolet agile ltz 1122 ==> PERIDOOCORRENCIA=pela manha 560	50%
4	DESCR_MARCA_VEICULO=ford/ka flex 979 ==> PERIDOOCORRENCIA=pela manha 478	49%
5	DESCR_MARCA_VEICULO=ford/fiesta flex 1531 ==> PERIDOOCORRENCIA=pela manha 722	47%
6	DESCR_MARCA_VEICULO=vw/fox 1.6 gii 1062 ==> PERIDOOCORRENCIA=pela manha 499	47%
7	DESCR_MARCA_VEICULO=fiat/palio fire economy 1661 ==> PERIDOOCORRENCIA=pela manha 778	47%
8	DESCR_MARCA_VEICULO=fiat/palio elx flex 967 ==> PERIDOOCORRENCIA=pela manha 443	46%
9	**DESCR_MARCA_VEICULO=honda/cg 150 titan ks 906 ==> PERIDOOCORRENCIA=de madrugada 410**	**45%**
10	DESCR_MARCA_VEICULO=vw/fox 1.0 gii 1298 ==> PERIDOOCORRENCIA=pela manha 587	45%
11	DESCR_MARCA_VEICULO=fiat/fiorino flex 948 ==> PERIDOOCORRENCIA=pela manha 526	55%
12	DESCR_MARCA_VEICULO=ford/fiesta 1.6 flex 1342 ==> PERIDOOCORRENCIA=pela manha 674	44%
13	**DESCR_MARCA_VEICULO=honda/cb 300r 1750 ==> PERIDOOCORRENCIA=pela manha 768**	**44%**
14	**DESCR_MARCA_VEICULO=honda/cg150 fan esdi 1491 ==> PERIDOOCORRENCIA=pela manha 644**	**43%**
15	**DESCR_MARCA_VEICULO=honda/xre 300 1867 ==> PERIDOOCORRENCIA=pela manha 799**	**43%**

Fonte: OLIVEIRA; SILVA (2020)

3.5. Resultados Obtidos

No início desta pesquisa foi obtida base de dados sobre os crimes de furto e roubos de veículos na cidade de São Paulo-SP, referente ao período de 2016 a 2019, no site da SSP-SP.

Os dados estavam divididos em ano e mês, sendo cada ano e mês em um arquivo diferente, logo foi necessário transformar os arquivos obtidos no formato Pasta de Trabalho do Excel 97-2003(.xls), para que posteriormente ocorresse a importação no banco de dados.

Após importados para o banco de dados, os dados foram exportados no formato .csv e importados para ferramenta *Weka* onde gerou-se o arquivo .arff, que possibilitou a leitura dos dados. Com a leitura dos dados na ferramenta *Weka*, foram gerados alguns gráficos automaticamente, sendo possível analisar parcialmente os resultados sobre o tema. Através do processo KDD para a mineração dos dados foram obtidos os seguintes resultados:

- A análise dos furtos e roubos de veículos na cidade de São Paulo-SP, no período de 2016 a 2019, mostra uma redução de 36% no número de ocorrências. Havendo uma redução média de 13,8% para cada ano.
- Para o conjunto semanal de registros, a quarta-feira contém o maior número de furtos e roubos de veículos, enquanto sábado registra menor número. Ocorrendo uma redução de 20,6% de registros do dia com mais casos para o com menos casos.
- Na divisão dos resultados por períodos (manhã, tarde, noite e madrugada) observou-se que a maioria das ocorrências aconteceram no período da noite, com 35% do total.

Após a utilização do algoritmo *Apriori* na base de dados, foram selecionados 3 cenários para serem analisados, esses cenários e os resultados obtidos são:

- No cenário 1 foram obtidos resultados significativos para a pesquisa. Destacando a seguinte associação, quinta-feira pela manhã com a cor do veículo sendo preta as chances que o veículo será furtado é de 82%. Outra associação destacada é na quarta-feira pela manhã a cor do veículo sendo prata as chances de o veículo ser furtado é de 82%. Ressaltasse que a maior parte dos veículos são fabricados nas cores preta e prata, esse fato pode ter influenciado no valor de confiabilidade do resultado, gerando, assim, uma porcentagem maior.
- No cenário 2 foram identificadas 10 regras, sendo geradas com 3 atributos, com confiança máxima 60%, no qual obteve-se os bairros e período das ocorrências dos crimes de furto e roubo de veículos. Destacam-se os seguintes resultados:

no bairro Ipiranga há 60% de confiança de haver ocorrência no período da manhã, e no bairro Vila Mariana há 58% de confiança de haver ocorrência no período da manhã.

- No cenário 3 os resultados obtidos foram voltados aos modelos e tipos de veículos. Destacou-se os resultados obtidos para o tipo de veículo motocicleta e os período com mais ocorrências. Os resultados que apresentaram maior confiabilidade foram: modelo honda/pcx 150, possui a confiança de 61% para o período a noite; modelo honda cg 160 fan esdi, possui a confiança de 60% de ser furtada ou roubada a noite; e modelo honda/cg 150 titan ks, possui uma confiança de 45% para ocorrências no período de madrugada.

3.6. Trabalhos Futuros

Para trabalhos futuros, sugere-se:

- Desenvolver uma aplicação *web* para disponibilização dos conhecimentos extraídos desse trabalho à população de São Paulo.
- Extrair os poucos dados contidos na coluna sexo das ocorrências, para descobrir padrões de roubos de veículos voltados ao público feminino com um desvio padrão.
- Aplicação do método de pesquisa deste trabalho em bases de dados disponibilizadas por SSP de outras localidades do Brasil.

4. CONSIDERAÇÕES FINAIS

Com a utilização da metodologia DCBD obteve-se resultados sobre associações feitas a respeito dos crimes de furto e roubo de veículos em São Paulo-SP, no período de 2016 a 2019, com base nos dados coletados da SSP/SP. Os resultados do trabalho mostraram o impacto positivo que a tecnologia pode ter na sociedade. A partir da compreensão dos dados, no qual o uso de ferramentas e processos de mineração de dados são úteis para auxiliar os órgãos públicos, é possível melhorar a gestão da segurança pública.

Utilizando como referência as associações feitas a partir da mineração dos dados é possível que a SSP/SP tome decisões que visem a segurança da população, e assim, reduza ainda mais os casos de furto e roubo de veículos na cidade de São Paulo-SP. A partir dos dados obtidos os órgãos públicos podem, por exemplo, construir campanhas de conscientização para que a população evite circular com veículos em zonas de risco, onde há altos índices de furto e roubo de veículos, bem como podem ser realizadas propagandas em diversas mídias, para informar os condutores de veículos sobre os dados de furto e roubo, buscando que os cidadãos previnam-se e orientem-se.

Além disso, os dados levantados podem cooperar para reforçar o patrulhamento nos bairros e períodos do dia no qual existem altos índices de ocorrências. Contudo, verificou-se a existência de dados que não apresentaram qualidade suficiente para serem introduzidos a este trabalho, variáveis como sexo, idade e etnia da vítima apresentavam falta de preenchimento em muitos dados. Para uma análise mais detalhada, e para políticas públicas de segurança voltadas a grupos populacionais seria necessário obter esses dados.

Com a finalização deste trabalho espera-se que as informações e conhecimento gerados auxiliem nas políticas de segurança pública da cidade de São Paulo-SP. Espera-se também que a metodologia e a utilização da ferramenta de mineração de dados nesse trabalho possa ajudar estudos futuros, bem como na implementação de tecnologia para obtenção de informações por órgãos públicos. Possibilitando, assim ampliar a eficiência das ações e políticas no combate aos crimes de furto e roubo de veículos.

REFERÊNCIAS BIBLIOGRÁFICAS

ARAÚJO, B.; MACIEL, A. **Aplicação de Regras de Associação em Dados da Criminalidade da Cidade do Recife**. Revista De Engenharia E Pesquisa Aplicada, 3(3), (2018). Disponível em: <https://doi.org/10.25286/repa.v3i3.974>. Acesso em: 16 Dez. 2019

BRASIL. **Decreto-Lei Nº 2.848, de 7 de Dezembro de 1940**. Código Penal, Brasília, DF, dez. 1940. Disponivel em: <http://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm>. Acesso em: 16 Dez. 2019

CAMILO, C. O.; SILVA, J. C. **Mineração de Dados: conceitos, tarefas, métodos e ferramentas**. 2009. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 15 set. 2019.

CAMPOS, E. A. V.; ASCENCIO, A. F. G. **Fundamentos da programação de computadores**. Prentice Hall, 2003.

CARVALHO, I. M. et al. **Contribuições das tecnologias KDD e DW como ferramentas de gestão do conhecimento aplicadas ao processo de compras do governo eletrônico**. Anais da V Conferência Sul-Americana em Ciência e Tecnologia aplicada ao Governo Eletrônico - CONEGOV. Florianópolis: Digital Ijúris, 2009, p. 95-113.

CASTRO, L. **Legislação comentada: furto art. 155 do CP. 2014**. Disponível em: <https://leonardocastro2.jusbrasil.com.br/artigos/136366573/legislacao-comentada-furto-art-155-do-cp>. Acesso em: 30/09/2019.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações**. São Paulo: Saraiva, 2016.

CURA, A. A. V. **Crimes, delitos e penas no Direito Romano Clássico**. Aveiro: Universidade de Aveiro, 2005.

DALLAGASSA, M. R. **Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas**. Ato Z: novas práticas em informação e conhecimento, v. 3, n. 2, p. 82-86, dez. 2014.

DATAFOLHA. **Pesquisa Nacional de Vitimização: Questionário SENASP**. Maio, 2013. Disponível em: <http://www.crisp.ufmg.br/wp-content/uploads/2013/10/Relat%C3%B3rio-PNV-Senasp_final.pdf>. Acesso em: 16 Dez. 2019.

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

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André C.P.L.F. **Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina**. Rio de Janeiro, LTC, 2011.

FAYYAD, U. M. et al. ***From data mining to knowledge discovery: an overview***. *knowledge discovery and data mining*, *Menlo Park : AAAI Press*, 1996.

FRACALANZA, Livia Fonseca. **Mineração de dados voltada para recomendação no âmbito de marketing de relacionamento**. Dissertação (Mestrado em Informática) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009. Disponível em: <https://web.tecgraf.puc-rio.br/press/publication/Fracalanza2009/Fracalanza2009.pdf>.

FRAGA, Thaís Lima. **Qual o impacto do crime para as vítimas? Uma análise considerando a influência dos roubos e furtos na percepção de segurança e migração no Brasil**. Programa de Pós-graduação em Economia, Universidade Federal de Viçosa, Viçosa, MG, 2015. Disponível em: <https://www.locus.ufv.br/bitstream/handle/123456789/9641/texto%20completo.pdf?sequence=3&isAllowed=y>. Acesso em: 16 Dez. 2019.

GALVAO, Noemi Dreyer; MARIN, Heimar de Fátima. **Técnica de mineração de dados: uma revisão da literatura**. Acta paul. enferm., São Paulo , v. 22, n. 5, p. 686-690, Oct. 2009 . Disponível em: <http://dx.doi.org/10.1590/S0103-21002009000500014>. Acesso em: 16 Dez. 2019.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. ***Data mining*: conceitos, técnicas, algoritmos, orientações e aplicações**. 2. ed. Rio de Janeiro: Elsevier, 2015.

GONÇALVES, E. C. **Data mining de regras de associação – Parte 2**. 2007. Disponível em: <https://www.devmedia.com.br/data-mining-de-regras-de-associacao-parte-2/6941>. Acesso em: 08 nov. 2019.

HAN, J; KAMBER, M. **Mineração de dados: conceitos e técnicas**. São Paulo: Elsevier, 2006.

IGLESIAS, J. ***Derecho Romano: historia e instituciones***. Sello Editorial, Madrid, 2010.

ITAKURA, Fernando Takashi. **Inteligência Artificial**. Guarapuava: Escola Regional de Informática, 2004. 230 p.

LUZ, Rafael Costa Da. **Crimes contra o Patrimônio**. 2017. Disponível em: <https://www.conteudojuridico.com.br/consulta/Artigos/50273/crimes-contra-o-patrimonio>. Acesso em: 16 Dez. 2019.

MANHÃES, L. M. B. et al. **Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados**. Anais do XXII SBIE - XVII WIE, Aracaju, 21 a 25 de novembro de 2011.

MCCUE, C. **Mineração de dados e análise preditiva: coleta de informações e análise de crimes**. São Paulo: Elsevier, 2007.

MENDES, Luciana. ***Data Mining*: – Estudo de Técnicas e Aplicações na Área Bancária**. 2011. Disponível em: <https://goo.gl/GQCfZQ>. Acesso em: 15 abr. 2020.

PERON, Alcides Eduardo Dos Reis. **Segurança Preditiva? A incorporação De Técnicas De Mineração De Dados E Perfilização Em Conflitos Internacionais Com Drones Pelos Eua E Em Práticas De Vigilância Pela Policia Militar Do Estado De São Paulo**. IV Simposio Internacional LAVITS, Buenos Aires - 2016. Disponível em: http://lavits.org/wp-content/uploads/2017/08/P4_Peron.pdf . Acesso em: 16 Dez. 2019.

SÃO PAULO (Estado). **Estado de SP reduz homicídios, latrocínios e roubos e furtos de veículos em agosto**. 2019. Disponível em:< http://www.saopaulo.sp.gov.br/spnoticias/estado-de-sp-reduz-homicidios-latrocinios-e-roubos-e-furtos-de-veiculos-em-agosto/>. Acesso em: 02 Fev. 2020.

SELIYA, N; KHOSHGOFTAAR, T. M. **Modelagem de qualidade de *software* com dados limitados de defeitos *Apriori***. Chapter, v. 1, p. 1–16. *Idea Group Publishing*, 2007.

SILVA FILHO, A. P.; SILVA, S. B. ***Data mining* através da regra de associação *apriori***. 2013. Disponível em: <https://www.profissionaisti.com.br/2013/11/data-mining-atraves-da-regra-de-associacao-apriori/>. Acesso em: 07 nov. 2019.

SILVA, D. **Banco de dados**. 2015. Disponível em: https://www.estudopratico.com.br/banco-de-dados/. Acesso em: 15/09/2019.

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

SOUZA, Marcelo Rodrigues De. **O Impacto Dos Principais Crimes Contra O Patrimônio Na Segurança Pública No Ano De 2016, Em Goiânia**. Curso De Especialização Em Gerenciamento De Segurança, Universidade Estadual De Goiás, Goiânia, 2017. Disponível em:

<https://acervodigital.ssp.go.gov.br/pmgo/bitstream/123456789/431/53/O%20Impacto%20dos%20Principais%20Crimes%20Contra%20o%20Patrim%C3%B4nio%20na%20Seguran%C3%A7a%20P%C3%BAblica%20no%20Ano%20de%202016%2C%20em%20Goi%C3%A2nia%20-%20Marcelo%20Rodrigues%20de%20Souza.pdf>. Acesso em: 16 Dez. 2019.

STEZER, V. W. **Dado, Informação, Conhecimento e Competência**. In: SETZER, V. W. Meios Eletrônicos e Educação: uma visão alternativa. São Paulo: Escrituras, 2015. Cap. 11, p. Não paginado. Disponível em: https://www.ime.usp.br/~vwsetzer/dado-info.html. Acesso em: 19/10/2019.

TAN, Pang Ning; STEINBACH, Michael; KUMAR, Vipin. ***Data Mining*: Mineração de dados**. Rio de Janeiro: Ciência Moderna, 2009. 900 p. Tradução de: Acauan P. Fernandes.

VALENTIN, A. P. P. et al. ***WEKA*-G: mineração de dados paralela em grades computacionais**. Rev. de Sistemas de Informação da FSMA, n. 4, p. 41-50, 2009.

VASCONCELOS, L. M. R.; CARVALHO, C. L. **Aplicação de Regras de Associação para Mineração de Dados na Web**. 2004. Disponível em: <http://www.portal. inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-04.pdf>. Acesso em: 08 nov. 2019.

VIEIRA, R. S. G. **Descoberta de conhecimento na relação entre acidentes de trânsito rodoviário e fatores climáticos, no eixo Goiânia-Distrito Federal**. TCC II Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – UniEVANGÉLICA. 2018, p. 128.

VITERBO, J. **Avaliação de ferramentas de apoio ao ensino de técnicas de mineração de dados em cursos de graduação**. 2016. Disponível em: <http://editora.pucrs.br/anais/ csbc/assets/2016/wei/02.pdf>. Acesso em: 20/10/2019.

APÊNDICE

APÊNDICE A - *Scripts* DDL e SQL - transformação de dados

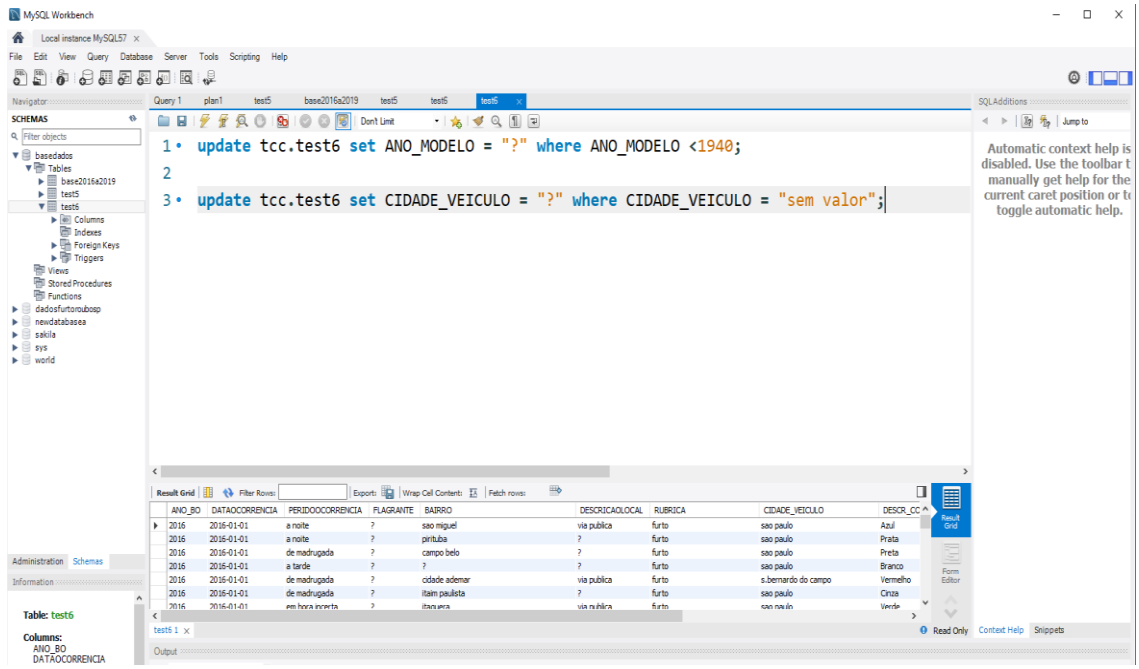
Os *scripts* ilustrados nas figuras 13 a 16 foram utilizados para auxiliar nas etapas de pré-processamento e transformação dos dados durante o trabalho.

Figura 13 – *Script* para remoção de acentos

```
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'À','A');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Á','A');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Â','A');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ã','A');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ä','A');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Å','A');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ç','C');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'È','E');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'É','E');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ê','E');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ë','E');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ñ','N');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ò','O');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ó','O');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ô','O');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Õ','O');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ö','O');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ù','U');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ú','U');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Û','U');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'Ü','U');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'à','a');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'á','a');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'â','a');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ã','a');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ä','a');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ç','c');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'è','e');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'é','e');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ê','e');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ë','e');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ì','i');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'í','i');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'î','i');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ï','i');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ð','o');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ñ','n');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ò','o');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ó','o');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ô','o');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'õ','o');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ö','o');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ø','o');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ù','u');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ú','u');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'û','u');
UPDATE planl SET BAIRRO = REPLACE(BAIRRO,'ê','e');
```

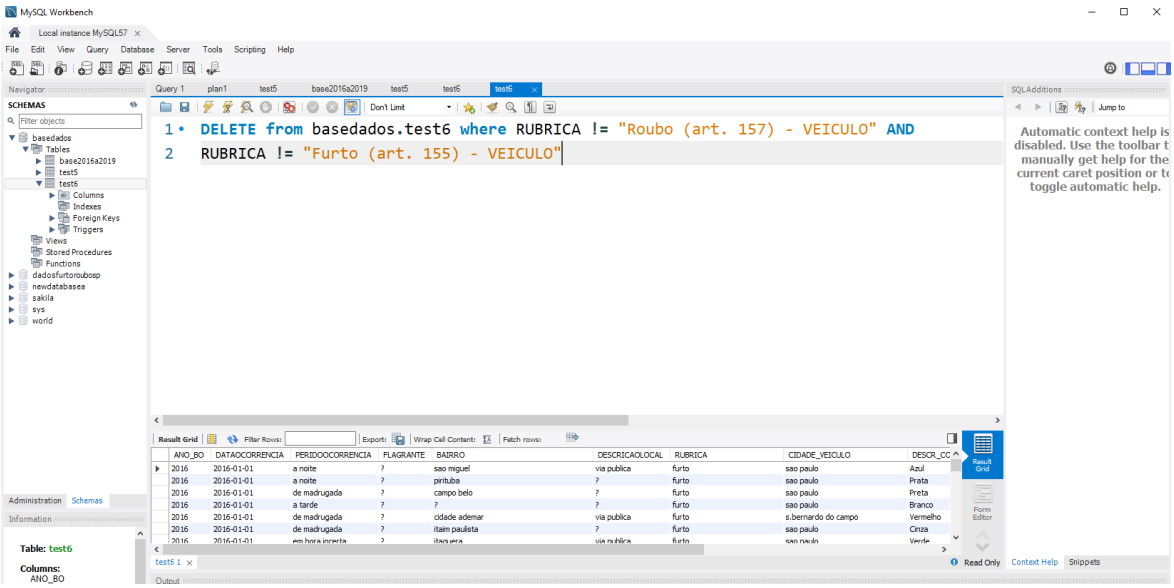
Fonte: OLIVEIRA; SILVA (2020)

Figura 14 – *Script* para remoção de valores inválidos



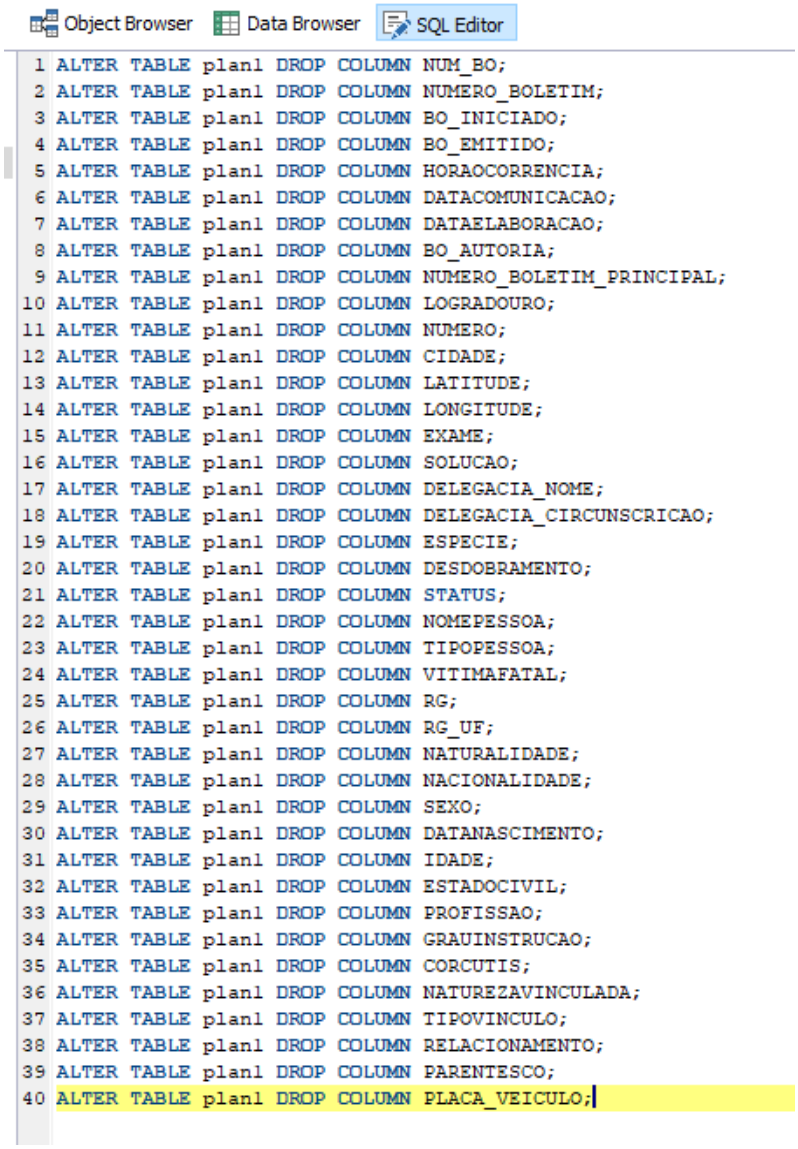
Fonte: OLIVEIRA; SILVA (2020)

Figura 15 – *Script* Sql para remover dados duplicados



Fonte: OLIVEIRA; SILVA (2020)

Figura 16 – *Script* para exclusão de colunas



Fonte: OLIVEIRA; SILVA (2020)