UNIVERSIDADE FEDERAL DO ACRE

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

PREVISÃO DA NATUREZA DE OCORRÊNCIAS POLICIAIS NA CIDADE DE RIO BRANCO

RIO BRANCO

2018

ALAN CORDEIRO RODRIGUES

PREVISÃO DA NATUREZA DE OCORRÊNCIAS POLICIAIS NA CIDADE DE RIO BRANCO

Projeto de TCC apresentado como exigência parcial para obtenção do grau de bacharelado em Sistemas de Informações da Universidade Federal do Acre.

Prof. Orientador: Manoel Limeira de Lima Júnior Almeida.

RIO BRANCO

2018

TERMO DE APROVAÇÃO

ALAN CORDEIRO RODRIGUES

PREVISÃO DA NATUREZA DE OCORRÊNCIAS POLICIAIS NA CIDADE DE RIO BRANCO

Esta monografia foi apresentada como trabalho de conclusão de Curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre, sendo aprovado pela banca constituída pelo professor orientador e membros abaixo mencionados.

Compuseram a banca:

_____

Prof. Manoel Limeira de Lima Júnior Almeida, Dr.
Curso de Bacharelado em Sistemas de Informação

_____

Prof. Daricélio Moreira Soares, Dr.
Curso de Bacharelado em Sistemas de Informação

_____

Prof. Laura Costa Sarkis, Dra.
Curso de Bacharelado em Sistemas de Informação

Rio Branco, 24 de agosto de 2018.

Dedico esta monografia aos meus pais, por sempre colocarem a educação da família em primeiro plano.

AGRADECIMENTOS

Em primeiro lugar a Deus, por tudo. À minha família que ofereceu o suporte para que eu alcançasse os meus objetivos, aos colegas de curso que me motivaram e cooperaram durante toda a jornada, ao meu orientador, aos professores e a equipe gestora da Universidade Federal do Acre por manterem a paciência e dedicação com a educação do corpo discente desta Universidade.

”Não há fatos eternos, como não há verdades absolutas" – Friedrich Nietzsche

RESUMO

Apesar de a segurança ser um direito estabelecido em lei no Brasil, cidadãos da cidade de Rio Branco, capital do Estado do Acre, convivem com cada vez mais insegurança. A Polícia Militar do Estado do Acre, uma das forças policiais que tem a função de prover a segurança à população do Acre, utiliza-se de mapas térmicos em seus planejamentos para priorizar o empenho de policiais em regiões com maior incidência de crimes. Porém vê-se no uso de técnicas de Mineração de Dados um meio para otimizar esses planejamentos, de forma a procurar prever com base no histórico de registros de ocorrências, quais naturezas de ocorrências podem ocorrer em determinadas regiões. Por fim, com um estudo de caso obteve-se resultados satisfatórios, com uma taxa de acerto de até 71,41% na previsão das 5 principais naturezas de ocorrências policiais da cidade de Rio Branco.

Palavras-chave: Mineração de Dados, Segurança Pública, Prevenção de Crimes, Aprendizado de Máquina

ABSTRACT

Although the personal security is a Brazilian right in constitution, Capital citizens from Rio Branco city, has to deal with growing crimes and insecurity feel on the streets. In response, the state Police (Polícia Militar do Acre) is in charge to promote security to the people. Today, they are using heat maps as a tool to identify regions who demands more attention from the police. To obtain best results and more precise data where the crime can and could occur, it is proposed the utilization of Data-Mining to optimize the task of the intelligence agencie. Using historical, records of the events it's possible to obtain this kind of information. Finally, with of a case study we obtained satisfactory results, with a hit rate of up to 71.41% in the prediction of the 5 main types of police occurrences in the city of Rio Branco.

Key-words: Data Mining, Public Security, Crime Forecasting, Machine Learning

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Segundo a Constituição Federativa do Brasil de 1988, a segurança é um direito fundamental do ser humano, e é dever do Estado provê-la aos seus cidadãos (BRASIL, 1988). Entretanto essa não é a realidade da população brasileira, visto que o Brasil ocupa a 121ª posição no ranking mundial no quesito Segurança Individual (SPI, 2017).

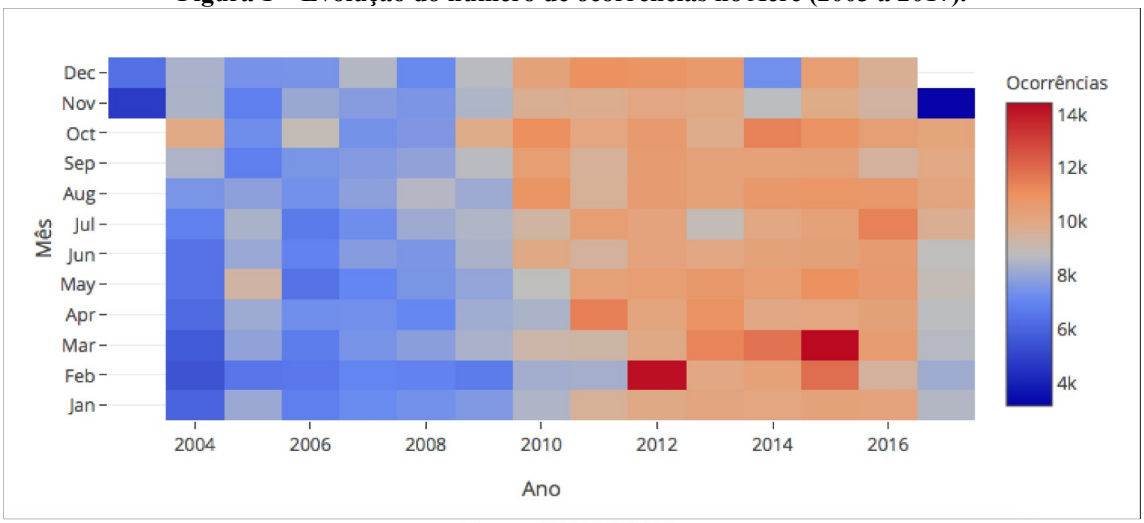
Na cidade de Rio Branco, capital do Estado do Acre, uma constante observada tem sido a tendência de crescimento no aumento dos índices de violência. Entre os anos de 2013 e 2014 foi registrado um crescimento de 138,6% no número de homicídios por arma de fogo, sendo então classificada como a capital brasileira com maior variação do índice no período (WAISELFISZ, 2016).

O aumento da violência no Estado do Acre também pode ser constatado ao observar o quantitativo mensal de ocorrências policiais entre os anos de 2003 e 2017. Nesse período, o Centro Integrado de Operações em Segurança Pública (CIOSP) registrou um aumento gradativo do número de ocorrências mensais na cidade de Rio Branco.

Essa evolução pode ser visualizada na Figura 1, onde o eixo x dispõe os anos do período de 2003 a 2017 e o eixo y representa os meses de cada ano. O número de ocorrências de cada mês é representado pelos tons de cores entre Azul e Vermelho, meses com menos de 4.000 são representados pela cor Azul e meses com mais de 14.000 ocorrências são representados pela cor Vermelha. Com isso, é possível observar um avermelhamento gradual

do gráfico no período observado.

Figura 1 – Evolução do número de ocorrências no Acre (2003 a 2017).



Fonte: SIAP (2017).

Para planejar as ações de enfrentamento de crimes e como a força policial deve ser empenhada, a Polícia Militar do Estado do Acre (PMAC), utiliza-se de mapas térmicos para visualizar como os registros de ocorrências se intensificam em determinadas regiões. No entanto, técnicas específicas e mais aprimoradas podem ser utilizadas para extrair conhecimento sobre os dados das ocorrências registradas pelo CIOSP e auxiliar no planejamento das ações da PMAC.

Técnicas de Mineração de Dados obtiveram resultados positivos nas mais variadas áreas, como na Medicina (BELLAZZI, 2008), Marketing (MITIK, 2017), Educação (SIN, 2015) e inclusive na Segurança Pública (BRAZ, 2009). Na PMAC, a técnica de classificação da mineração de dados pode ser utilizada para disponibilizar conhecimentos que não podem ser observados apenas com a leitura de mapas térmicos, como por exemplo, prever a natureza dos crimes que podem ocorrer em uma determinada região da cidade de Rio Branco.

1.1 PROBLEMA DA PESQUISA

O uso de tecnologias da informação é algo vantajoso para o planejamento policial, tendo como resultado, policiais desempenhando suas funções com informações mais precisas

e concretas, aumentando a satisfação no trabalho e sua capacidade em prover segurança (AGRAWAL, 2003).

Entretanto, apesar da PMAC possuir mais de 1.500.000 registros de ocorrências armazenados em plataformas digitais, o potencial desses dados para prever a natureza das próximas ocorrências e auxiliar no planejamento policial não é explorado.

Todas as ocorrências atendidas no CIOSP são classificadas pelos atendentes com as suas respectivas naturezas de crime. Estas ocorrências já classificadas podem ser utilizadas para ensinar algoritmos de classificação a prever a natureza de ocorrências futuras. E com base nas naturezas previstas (e portanto com maior probabilidade de ocorrerem), a PMAC pode desenvolver ações para prevenção de crimes.

Nesse contexto, esta pesquisa justifica-se pela necessidade de explorar a técnica de Mineração de Dados denominada classificação para prever a natureza de ocorrências policiais na cidade de Rio Branco com enfoque na prevenção de crimes. Neste sentido, esta pesquisa apresenta o seguinte questionamento: como usar a classificação para a prever a natureza de crimes na cidade de Rio Branco?

1.2 OBJETIVO GERAL

Prever a natureza das ocorrências policiais na cidade de Rio Branco utilizando a técnica de classificação da mineração de dados.

1.3 OBJETIVOS ESPECÍFICOS

Os objetivos específicos desta pesquisa são:

a. Obter dados dos Sistemas de registro de ocorrências policiais de Rio Branco;
b. Processar e estruturar um repositório de dados de ocorrências;

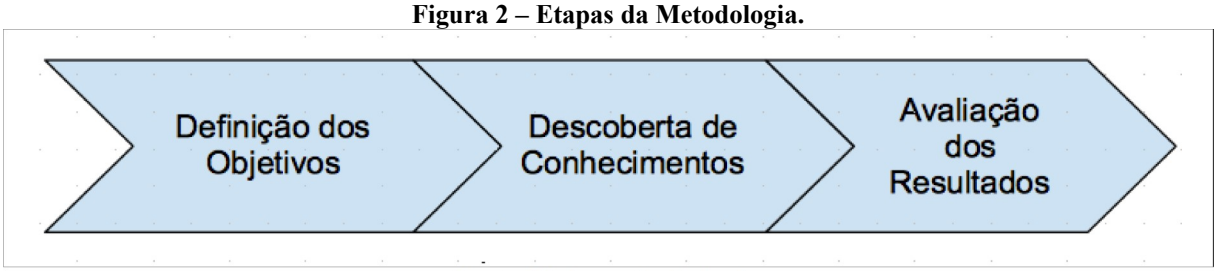
c. Treinar e avaliar modelos de previsão sobre a natureza das ocorrências;
d. Analisar o conhecimento dos modelos de previsão.

1.4 METODOLOGIA

Quanto a natureza, esta pesquisa define-se como um resumo de assunto, por sistematizar uma área de conhecimento já explorada por outros autores. Quanto aos objetivos, a classifica-se como exploratória, por buscar anomalias ainda desconhecidas e que possam ser exploradas. Sobre os procedimentos técnicos, trata-se de uma pesquisa documental por examinar relatórios e bancos de dados em busca de informações e padrões ainda não tratados sistematicamente (WAZLAWICK, 2014).

A metodologia desta pesquisa foi dividida em três fases principais, ilustradas na Figura 2. A primeira fase designa-se à definição dos objetivos que norteiam a pesquisa. A segunda fase, inclui a definição do Processo de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery from Data* – KDD) no Capítulo 2 e a sua execução no Capítulo 3. A terceira e última fase, engloba a coleta de dados e análise dos resultados:

Figura 2 – Etapas da Metodologia.



Fonte: Elaboração Própria.

Acrescenta-se que os procedimentos metodológicos utilizados para a execução desta pesquisa compreende um conjunto de técnicas definidos por outros autores, que foram adaptadas e aplicadas em um cenário específico. Podendo ainda, ser reutilizados como uma abordagem para execução de outras pesquisas para previsão da natureza de ocorrências policiais.

1.5 ORGANIZAÇÃO DA PESQUISA

No Capítulo 2 deste estudo, são abordados os principais conteúdos que servem como base para o trabalho desenvolvido referente ao processo de descoberta de conhecimento em bancos de dados.

No Capítulo 3, são apresentados os passos, etapas, métodos e resultados utilizados e obtidos para o desenvolvimento deste trabalho.

Por fim, o Capítulo 4 conclui este estudo apresentando as considerações finais e as recomendações para trabalhos futuros.

2 DESCOBERTA DE CONHECIMENTO EM BASE DADOS

Assim como ocorre em um livro, em um banco de dados é necessário que se leia os dados para entender quais informações ele contém. Porém, diferente de um livro, ler e compreender os dados armazenados em um banco de dados é uma tarefa bastante dispendiosa para humanos. Por exemplo, um único supermercado de uma rede de supermercados é capaz de registrar uma grande quantidade de registros de vendas de inúmeros produtos em um único dia. Seria complicada a tarefa do gerente interpretar os registros de vendas no final do dia sem o uso de recursos computacionais. O problema se agravaria caso fosse necessário extrair conhecimentos das vendas da rede de supermercados do período de um ano para planejar o próximo ano. Dessa forma, um livro é apenas um conjunto de caracteres antes que alguém o leia e o interprete. Assim como, um banco de dados nada mais é que, uma coleção de dados inter-relacionados, representando informações sobre um domínio específico (KORTH, 1994).

Observa-se que o principal problema para se interpretar os Bancos de Dados está no constante aumento do volume dos seus dados. Dentre os anos de 2006 a 2010, o volume de dados digitais passou de 166 Exabytes para 988 Exabytes (GANTZ, 2012) e espera-se que até 2020 os dados digitais somem mais de 40.000 Exabytes (MANYIKA, 2011).

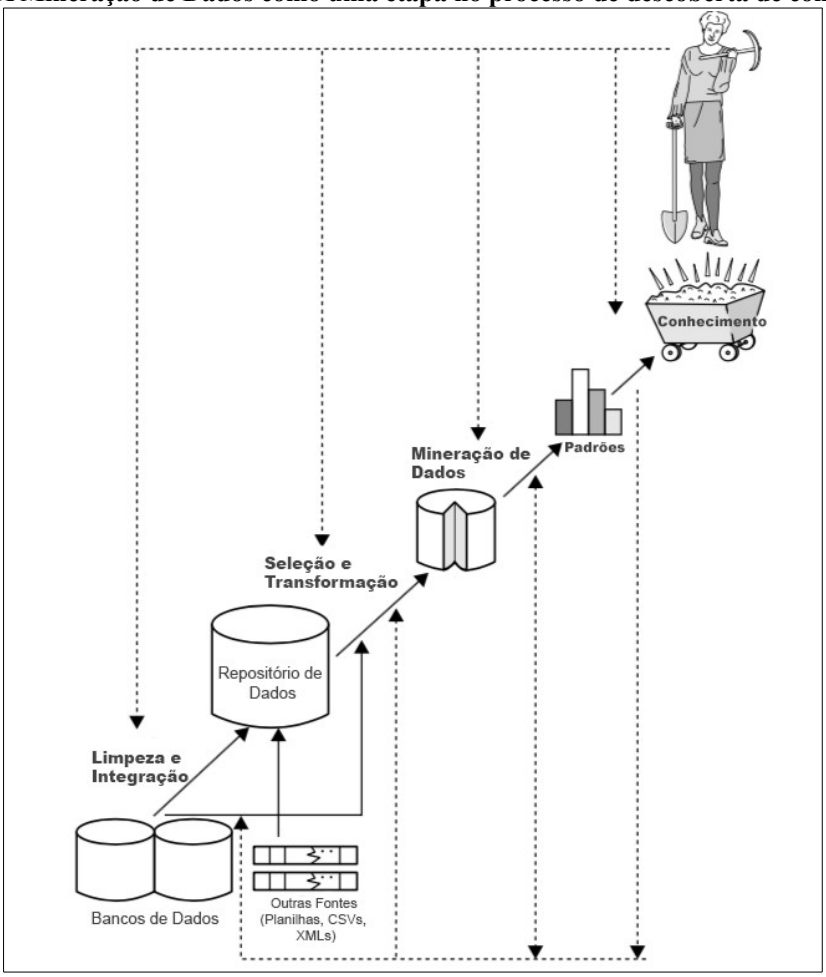
Diante do problema, de se obter conhecimento a partir de volumes de dados cada vez maiores, armazenados em diversas fontes de dados, como por exemplo, os Sistemas de Gerenciamento de Bancos de Dados (SGBDs), desenvolveu-se o processo de Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery from Data* – KDD) (HAN *et al*.,

2012).

De acordo com Fayyad *et al*. (1996) “KDD é um processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir dos dados armazenados em um banco de dados”.

Segundo Han *et al*. (2012), o processo de KDD é composto pelas seguintes etapas iterativas ilustrados pela Figura 3: inicia-se com a Limpeza e Integração de bancos de dados e arquivos de texto simples, segue-se com a Seleção e Transformação onde são selecionados dados de interesse e transformados para então proceder a Mineração de Dados onde são extraídos padrões que podem ser convertidos em conhecimentos. Fayyad, *et al*. (1996) ainda defende que por ser um processo iterativo, o KDD permite o retorno para qualquer passo e a repetição dos mesmos com o objetivo de aprimorar o conhecimento extraído.

Figura 3 – A Mineração de Dados como uma etapa no processo de descoberta de conhecimento.



Fonte: Adaptado de Han *et al*. (2012).

Na literatura, existem outras classificações para os passos do processo de KKD que geralmente envolvem as mesmas etapas citadas na Figura 3. Camilo e Silva (2009) descrevem o KDD dividido nos seguintes passos: Seleção, Pré-processamento, Transformação, Mineração de Dados e Avaliação.

Na sequência, da seção 2.1 até a seção 2.5 deste estudo, serão detalhados os conhecimentos referentes a cada etapa do processo de descoberta de conhecimento exibidos na Figura 3.

2.1 LIMPEZA E INTEGRAÇÃO

No mundo real, os dados tendem a serem incompletos, terem ruídos e ser inconsistentes. A limpeza se faz necessária para evitar que ruídos existentes nos dados prejudiquem a qualidade dos resultados, pois dados de baixa qualidade produzem resultados de baixa qualidade (HAN *et al*., 2012).

Goldschmidt e Passos (2015), defendem que a limpeza dos dados compreende "qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos representados". Um exemplo simples da execução da limpeza de dados é o processo de definição de um grupo de valores não aceitos (valores não definidos, números iguais a zero, valores inválidos no contexto ou valores ausentes) para um atributo e caso algum registro com atributo dentro dos valores não aceitos surgisse seria removido.

Após a limpeza, sejam os dados de SGBDs ou de outras fontes de dados mais simples, como arquivos de texto, planilhas, CSVs (*Comma Separated Values*) ou XMLs (*Extensible Markup Language*) podem ser integrados. Por exemplo, os Bancos de Dados de dois sistemas de cadastro de clientes de uma rede de supermercados, podem ser integrados em um só, atentando-se para não gerar duplicidade nos cadastros dos clientes. O resultado da integração deve ser uma fonte dados unificada, com registros mais consistentes e com menos redundâncias para facilitar a execução das outras tarefas do processo de KDD (HAN *et al*.,

2012).

2.2 SELEÇÃO E TRANSFORMAÇÃO

A seleção e a transformação dos dados compreendem a etapa do KDD que precede a Mineração de Dados. Nessa etapa, os dados são selecionados, e transformados para que possam ser compreendidos pelos algoritmos de mineração de dados.

A seleção tem o objetivo de selecionar os dados que serão efetivamente utilizados no processo de KDD. Registros (instâncias) e atributos relevantes são selecionadas para então serem transformados (GOLDSCHMIDT e PASSOS, 2015). Por exemplo, em uma situação hipotética onde se deseja descobrir um padrão ou conhecimento sobre os interesses musicais das pessoas, com base nas músicas que escutam, registros coletados há mais de 5 anos podem não ser mais relevantes.

Em geral, após a seleção pode ser necessário proceder algum tipo de transformação nos dados, para permitir que os mesmos sejam utilizados pelos algoritmos de mineração de dados. A transformação é necessária, pois existem algoritmos que não funcionam com determinados tipos de dados, como por exemplo, cadeia de caracteres. Além disso, alguns algoritmos podem oferecer melhores resultados dependendo do tipo de dado armazenado nos atributos. Dentre as diversas transformações possíveis, dois tipos tradicionais são: Numérica – Categórica, quando se transforma valores reais em categorias ou intervalos; ou Categórica – Numérica, quando se representa numericamente valores de atributos categóricos (BOENTE, 2008).

Han *et al*. (2012) definem outras técnicas de transformação de dados:

a. **Suavização:** remove-se outros ruídos dos dados utilizando técnicas como clusterização ou regressão.
b. **Agregação:** dados são agregados ou sintetizados para facilitar a tarefa de mineração. Como exemplo, as vendas diárias podem ser agregadas para se ter as vendas do mês.

c. **Normalização:** os dados de atributos são normalizados de modo a cair dentro de um intervalo menor, como -1,0 a 1,0, ou 0,0 a 1,0.

d. **Discretização:** atributos numéricos são convertidos em intervalos, como: 10-20, 25-37.

e. **Geração do conceito hierárquico de dado nominal:** dados como nome de rua de um endereço podem ser atribuído a níveis mais altos como nome do Bairro, Cidade ou País. Nesta etapa, também é possível obter dados ausentes, através da transformação ou combinação de outros dados, estes são os chamados de "dados derivados". Um exemplo de um dado que pode ser calculado a partir de outro é a idade de um indivíduo, que pode ser obtida a partir de sua data de nascimento. Outro exemplo é nome da rua de um endereço que pode ser obtidos a partir de coordenadas geográficas (PRASS, 2012).

2.3 MINERAÇÃO DE DADOS

Segundo Han *et al*. (2012), a etapa de mineração de dados é definida como o processo de extrair informação implícita, previamente desconhecida e potencialmente útil de dados de forma automática ou semiautomática.

A etapa de mineração de dados é por muitas vezes, citada como sinônimo de descoberta de conhecimento em base dados. Isso ocorre dada a sua importância no processo de KDD e porque apesar de serem necessárias várias etapas para se chegar nesta etapa, é justamente na mineração de dados que se utiliza de algoritmos de mineração para extrair o conhecimento das bases de dados (HAN *et al*., 2012).

O conhecimento extraído de bases de dados pelo processo de KDD também é chamado de Modelo de Conhecimento e a forma com que esses modelos são representados depende diretamente do algoritmo utilizado (GOLDSCHIMIT, 2015). Estes algoritmos geralmente são desenvolvidos com base em técnicas das áreas de Aprendizado de Máquina, Reconhecimento de Padrões e Estatística. Essas técnicas também podem ser combinadas para

se obter melhores resultados.

Na etapa de mineração de dados, diversas técnicas podem ser empregadas para se extrair o conhecimento em bases de dados. A seguir são abordados alguns conceitos das principais tarefas de mineração:

a**) Clusterização:** segundo Fayyad *et al*. (1996), a Clusterização que também é chamada de Segmentação ou Agrupamento, é utilizada para separar os registros de uma base de dados em grupos (*clusters*) de forma que os registros de um grupo possuam características comuns que os distingam de registros de outros grupos. Como exemplo, agrupar clientes por região do país; agrupar clientes com comportamento de compra similar; agrupar seções de usuários *Web* para prever comportamento futuro de usuário (DIAS, 2001).

b) Sumarização: também conhecida como Descrição de Conceitos, esta tarefa consiste em procurar identificar e apresentar as principais características comuns entre um conjunto de dados. Essa tarefa é aplicada nos resultados de agrupamentos obtidos com a tarefa de clusterização. Como exemplo, a descoberta da faixa etária do grupo (GOLDSCHIMIT e PASSOS, 2015).

c) Estimativa (ou Regressão): basicamente busca-se por funções, lineares ou não que mapeiem os registros de um banco de dados. Como exemplo de aplicação pode-se: estimar a probabilidade de um paciente sobreviver dado o resultado de um conjunto de diagnóstico de exames ou a definição do limite do cartão de crédito para cada cliente em um banco (DIAS, 2001); estimar o número de filhos ou a renda total de uma família ou prever a demanda de um consumidor para um novo produto (GOLDSCHIMIT e PASSOS, 2015).

d) Regras de Associação: dado um conjunto de transações (ou registros) onde cada transação contém um conjunto de itens (ou atributos). Uma regra de associação é expressa na forma $X \rightarrow Y$ (X, então Y) onde X e Y são itens (SRIKANT, 1997). Por exemplo, a técnica poderia ser aplicada nos registros de compras de um supermercado para descobrir a associação do final de semana, com a compra de carne para churrasco, ‘Final de semana, Então Carne para churrasco’.

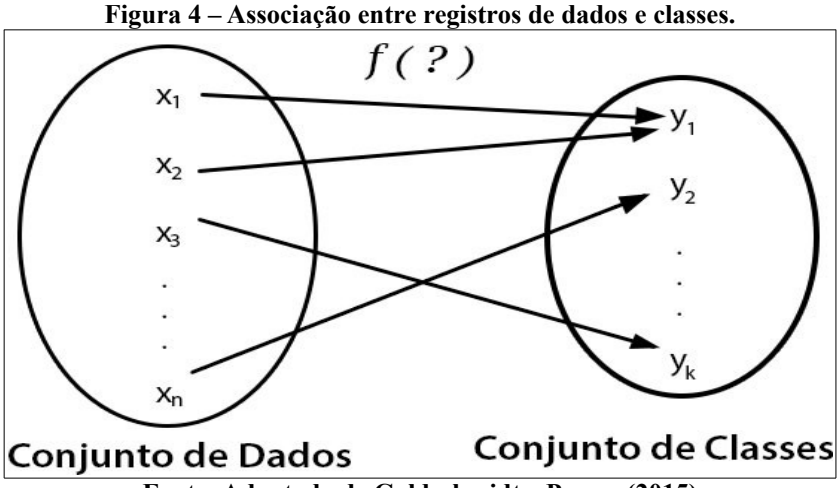
Na Mineração de Dados definem-se dois principais parâmetros para as regras de associação, o suporte e a confiança. O suporte representa a porcentagem de registros que contêm os atributos de X e Y. Enquanto a confiança avalia a porcentagem de registros que

possuem os atributos de *X*, e que também possuem os itens *Y* (AGRAWAL *et al*., 1997).

e) Classificação: considerada uma das tarefas mais importantes do processo de KDD e a mais popular, esta tarefa utiliza-se do aprendizado supervisionado, quando se disponibiliza para aprendizado valores de *entrada* e *saída* desejados para identificar a qual classe um registro pertence. Na sua execução, os atributos da base de dados são divididos em dois grupos. O primeiro é chamado de Classes, estes para qual será feita a previsão de valor e o segundo o grupo é chamado de Atributos Preditivos, este segundo grupo contém os atributos a serem utilizados na previsão da classe (GOLDSCHMIDT e PASSOS, 2015).

Conforme apresentado na Figura 4, a classificação é utilizada para buscar uma função que permita associar corretamente cada registro de um conjunto de dados a uma classe. Uma vez identificada, essa função pode ser aplicada a novos registros para prever em quais classes esses novos registros se enquadram (GOLDSCHMIDT e PASSOS, 2015). Por exemplo, baseando-se na idade e no histórico financeiro dos clientes de um banco, a função de classificação poderia associar um novo cliente a uma classe de risco, bastando que este informe a sua idade.

Figura 4 – Associação entre registros de dados e classes.



Fonte: Adaptado de Goldschmidt e Passos (2015)

Para construir e avaliar o modelo de classificação para predição, utilizam-se vários métodos de amostragem que buscam simular o processo de amostragem que ocorre no mundo real. Divide-se o registros em dois conjuntos: treino e teste. O conjunto de treino é usado para a criação do modelo de classificação e o de teste para avaliação da eficiência do modelo em executar a predição.

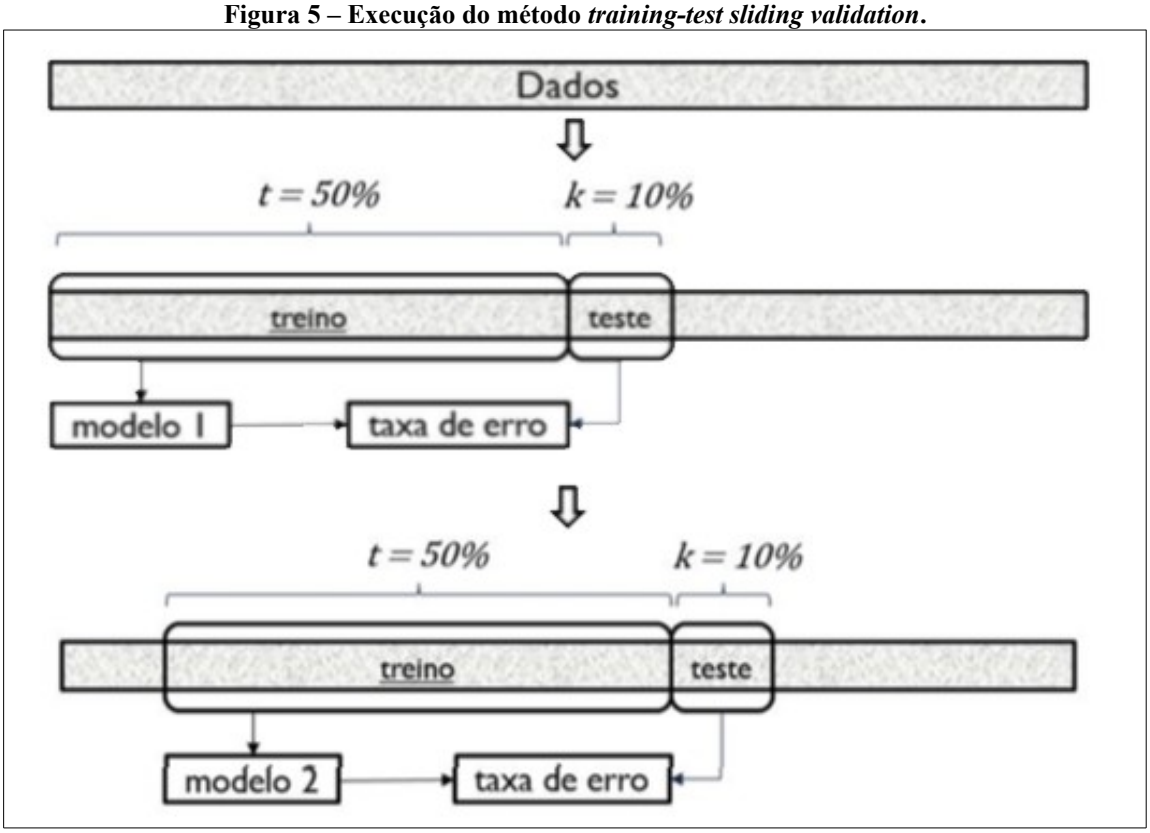
Os métodos de amostragem (ou de avaliação) definem como se deve dividir os conjuntos de treino e teste. Um dos métodos mais tradicionais é o *hold out* em que se divide os registros aleatoriamente em dois conjuntos, geralmente o primeiro com 70% do registros para treino e o segundo com os 30% registros restantes para teste. Porém por utilizar apenas o registros iniciais ou aleatórios para treino, um modelo criado usando este método pode não representar bem um conjunto de dados. Um método de amostragem que procura proporcionar um melhor representatividade do modelo é o *cross-validation*, este divide os dados aleatoriamente em *k* conjuntos (*D1*, *D2*, *D3*. …, *Dk*), com aproximadamente o mesmo tamanho, para em seguida realizar *k* interações de treino e teste. Em cada interação *i* o conjunto *Di* é reservado para teste e os conjuntos restantes são usados para treino (HAN *et al*., 2012).

Porém os métodos de amostragem citados anteriormente podem não ser tão eficientes na previsão de dados cronologicamente ordenados. Ocorre que registros futuros podem interferir na previsão de registros do passado. O método chamado *training-test sliding validation*, têm uma abordagem diferente para contornar o problema, para isso o método mantém a ordem cronológica dos registros, dividindo as amostragens em várias janelas para então dividir cada janela em conjuntos de treino e teste (LIMA JÚNIOR, 2017).

Lima (2017) desenvolveu um complemento para incorporar esse recurso na ferramenta de descoberta de conhecimento Weka e explica que a execução do método consiste em inicialmente definir dois valores *t* e *k* para o treino e para teste. Esses valores podem ser relacionados a porcentagem da base de dados, número de instâncias ou intervalo de datas.

A primeira interação do método é executada com os registros localizados no inicio da base de dados ordenada cronologicamente, utilizando um conjunto de treino de tamanho *t* e um conjunto de teste de tamanho *k.* Após a primeira interação os conjuntos de treino e teste são movidos para direita, ou seja, dos registros mais antigos para os mais atuais, selecionando para teste um conjunto de tamanho *k* logo após o conjunto de teste anterior. Esse processo se repete até que o conjunto de teste chegue ao fim da base. A Figura 5 ilustra as duas primeira interações do método com *t = 50%* e *k* = 10% (LIMA, 2017).

Figura 5 – Execução do método *training-test sliding validation*.



Fonte: Adaptado de Lima (2017).

Após a execução da classificação, a Matriz de Confusão pode ser analisada para visualizar um detalhamento do desempenho do modelo de classificação. Para cada classe, os resultados são apresentados em duas dimensões: **classes verdadeiras** e **classes preditas**. Então, cada elemento da matriz representa o número de elementos da classe verdadeira classificados em uma das classes preditas. Quando a matriz é referente a um problema com apenas duas classes, as suas predições recebem denominações especiais, Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN) (GOLDSCHMIDT e PASSOS, 2015). Conforme o Quadro 1 ilustra:

Quadro 1 – Exemplo de Matriz de Confusão

Classe Verdadeira	**Classe Predita**	
	Verdadeira	**Negativa**
Verdadeira	VP	FN
Negativa	FP	VN

Fonte: Elaboração Própria

É também com base nessas denominações especiais que algumas medidas de qualidade da classificação podem ser extraídas. Segundo Medeiros (2004), três medidas de

qualidade podem verificadas para avaliar um modelo de classificação:

• A porcentagem dos registros da classe em questão que conseguiram ser recuperados é definida como cobertura (*recall*), representada pela fórmula:

$$recall = \frac{VP}{VP + FN}$$

• A porcentagem dos registros que foram corretamente classificados como pertencentes à classe é a medida de precisão (*precision*), de acordo com a fórmula:

$$precision = \frac{VP}{VP + FP}$$

• A porcentagem dos registros que foram corretamente classificados corresponde à medida denominada acurácia (*accuracy*), definido na fórmula:

$$accuracy = \frac{VN + VP}{FN + FP + VN + VP}$$

Quão maior o valor dessas três medidas citadas, pode-se dizer que melhor é o modelo de classificação.

2.4 ALGORITMOS DE MINERAÇÃO

Diversos algoritmos podem ser utilizados para executar as principais tarefas de mineração, tais como: classificação, regras de associação e clusterização. Esses algoritmos implementam abordagens tradicionais para a execução das tarefas de mineração. Uma breve descrição de alguns algoritmos de clusterização, regras de associação e classificação:

a) **k-means:** é um algoritmo de clusterização em que cada *cluster* é representado pelo valor médio dos objetos no *cluster*. O algoritmo atribui cada registro ao *cluster* a qual o valor médio seja mais semelhante e em seguida atualiza o valor médio com base nessa atribuição do registro (HAN *et al*. 2012).

b) **Apriori:** é um algoritmo de Regras de Associação que busca um conjunto de itens

frequentes chamado de *itemset*, que satisfaçam uma confiança e um suporte mínimo. Um conjunto quem contem *k* itens é um *k-itemset*. Então à partir do *k-itemset* e as medidas de suporte e confiança o algoritmo produz regras de associação (SRIKANT, 1997).

c) **NB (*Naive Bayes*):** é um classificador bayesiano, que pressupõe que os atributos são interdependentes, portanto não têm relação entre si (GOLDSCHIMIT e PASSOS, 2015). Este, usando o Teorema de Bayes, calcula a probabilidade posteriori, a probabilidade de um registro *X* pertencer a alguma classe *Y* (LIMA JÚNIOR, 2017).

d) **J48:** É um algoritmo classificador que executa a classificação por meio de Árvore de Decisão, onde cada nó não folha representa uma condição do tipo SE <condição>, ENTÃO <conclusão> entre um atributo e um conjunto de valores (GOLDSCHIMIT e PASSOS, 2015).

e) **IBk (*Instance Based Learner*):** algoritmo classificador que corresponde à uma implementação do algoritmo *k-Nearest Neighbors* (*k*-NN). Este calcula a distância de um novo registro para os registrados já existentes. Usa-se a classe mais comum dos k registros mais semelhantes (com menor distância) para classificar o novo registro (GOLDSCHIMIT e PASSOS, 2015).

Na Mineração de Dados, também é possível avaliar a relevância dos atributos envolvidos na tarefa. Técnicas de Seleção de Atributos, podem ser aplicadas com utilização de critérios de seleção e algoritmos distintos para avaliar e encontrar de forma heurística os atributos (LEE, 2005). Dentre os algoritmos para Seleção de Atributos temos:

a) ***CorrelationAttributeEval*:** Avalia a relevância de um atributo medindo a correlação entre ele e a classe.

b) ***InfoGainAttributeEval:*** Avalia a relevância de um atributo medindo o ganho de informação em relação à classe.

c) ***ReliefFAttributeEval*:** Avalia a relevância de um atributo por meio de amostragem de exemplos e considerando o valor do atributo para a instância mais próxima da mesma e de outra classe.

Dentre os algoritmos de Mineração de Dados e Seleção de Atributos descritos anteriormente, foram utilizados nos experimentos do Capítulo 3, que contém o estudo de caso, apenas os algoritmos de classificação NB, IBK e J48. Visto que estes algoritmos são aqueles que trabalham com o objetivo desta pesquisa que visa predizer as naturezas das ocorrências

policiais na cidade de Rio Branco – Acre. Além disso, todos os algoritmos de Seleção de Atributos descritos anteriormente foram utilizados no Capítulo 3.

2.5 INTERPRETAÇÃO

Após a etapa de Mineração de Dados e com a obtenção dos padrões, é necessário que se proceda uma interpretação destes padrões pelo cientista de dados. É também nessa etapa que geralmente o especialista no domínio da aplicação avalia os resultados obtidos (GOLDSCHIMIT e PASSOS, 2015). Como por exemplo, após proceder KDD nos registros de compra de um supermercado o cientista de dados pode perceber que clientes do sexo masculino estão comprando mais fraldas em finais de semana e com isso interpretar que essas compras remetem aos pais divorciados ou não que estão cuidando dos filhos.

Já que o KDD é um processo iterativo e garante o retorno a etapas anteriores, caso os resultados observados não sejam satisfatórios o analista pode retornar a qualquer das etapas anteriores ou até julgar necessário refazer todo o processo KDD.

2.6 FERRAMENTAS DE DESCOBERTA DE CONHECIMENTO

Existem várias ferramentas que oferecem um conjunto de recursos para a aplicação do KDD. Estas ferramentas permitem que o cientista de dados execute todas etapas do KDD, por meio de interfaces gráficas, que contêm representações visuais dos recursos. Também incluem um conjunto variado de algoritmos que podem ser executados na Mineração de Dados e Seleção de Atributos.

Alguns exemplos dessas ferramentas são: Orange[1], RStudio[2], RapidMiner[3],

1 https://orange.biolab.si
2 https://www.rstudio.com
3 https://rapidminer.com

Anaconda[4] e Weka[5].

Dentre as ferramentas disponíveis, escolheu-se a Weka para desenvolver o estudo de caso deste trabalho. A Weka possui um conjunto de implementações de algoritmos de diversas técnicas de Mineração de Dados e contêm ferramentas para pré-processamento, classificação, regressão, clusterização, regras de associação e visualização (UNIVERSITY OF WAIKATO, 2018).

Além disso, os fatores que consolidaram a escolha da Weka dentre as opções disponíveis, foi a sua facilidade de uso e a existência do complemento *Chronological Classify*, que incorpora na Weka o método *training-test sliding validation*, utilizado para avaliação de classificadores em dados cronológicos (como por exemplo, os registros de ocorrências policiais).

2.7 TRABALHOS RELACIONADOS

Tem sido recorrente o uso do KDD para extração de conhecimentos de bancos de dados policiais. O programa Policiamento Preditivo de Santa Cruz, utilizando como base, modelos de previsão de tremores de terremotos e dados fornecidos pelo Departamento de Polícia da cidade de Los Angeles, mostrou-se capaz de indicar pontos geográficos propensos a terem novos crimes (THE NEW YORK TIMES, 2011).

No Brasil, Braz *et al*. (2009) aplicou o KDD em registros de ocorrências da Polícia Militar do Alagoas e mostrou que criminosos agem com conduta semelhante a ponto de formarem-se grupos de ocorrências com horário, dia da semana, bairro, cidade e natureza de crimes semelhantes. Em Braz *et al*. (2012) além de aplicar o KDD, desenvolveu-se um sistema de apoio à decisão que demonstra associações entre naturezas de ocorrências.

Esta pesquisa se difere das demais citadas anteriormente por tratar em específico da previsão da natureza das ocorrências policiais com base no logradouro, data e hora das

4 https://anaconda.org
5 https://www.cs.waikato.ac.nz/ml/weka/

ocorrências registradas na cidade de Rio Branco – Acre e o uso de um método de amostragem adequado para bases de dados onde os registros são organizados de forma cronológica, denominada *training-test sliding validation*.

3 ESTUDO DE CASO: PREVISÃO DA NATUREZA DE OCORRÊNCIAS POLICIAIS NA CIDADE DE RIO BRANCO

Este capítulo apresenta os resultados obtidos após a realização das etapas necessárias para o desenvolvimento deste trabalho. O processo de KDD configurou-se como base para execução das tarefas que se sucedem utilizando os dados disponibilizadas pela Secretaria de Estado de Segurança Pública (SESP), oriundos dos registros de ocorrências do Centro Integrado de Segurança Pública (CIOSP).

Durante o período de novembro de 2003 a novembro de 2017 o CIOSP utilizou o sistema SIAP (Sistema Integrado de Atendimento ao Público) para registrar 1.509.226 ocorrências. Em novembro de 2017, visando redução de gastos e integração com o Sistema Nacional de Segurança Pública (SINESP), o SIAP foi substituído por um novo sistema, desenvolvido pelo SINESP, o SINESP CAD (Central de Atendimento) que nos primeiros 8 meses de funcionamento foi utilizado no registro mais de 50.000 ocorrências nos Estado do Acre. Portanto, os dados utilizados nos experimentos compreendem o período de novembro de 2003 a julho de 2018, com um total de 1.563.870 registros, contabilizando uma média mensal de 8.920 ocorrências no Estado do Acre.

Por ser Policial Militar e desempenhar a função de desenvolvedor de *softwares* na Divisão de Tecnologia e Informação da PMAC, o pesquisador teve acesso facilitado aos dados de ocorrências da PMAC para a execução deste estudo de caso. E por possuir conhecimento prévio dos dados, configura-se também como especialista de domínio no processo de KDD.

Apresenta-se neste capítulo as medidas de segurança com os dados utilizados na

pesquisa na seção 3.1. Além disso, a seção 3.2 contextualiza os procedimentos a serem executados para a descoberta de conhecimentos focados na previsão da natureza das ocorrências policias. As subseções 3.2.1 à 3.2.3 detalham cada fase e os resultados do KDD aplicado neste estudo de caso.

3.1 SEGURANÇA DA INFORMAÇÃO

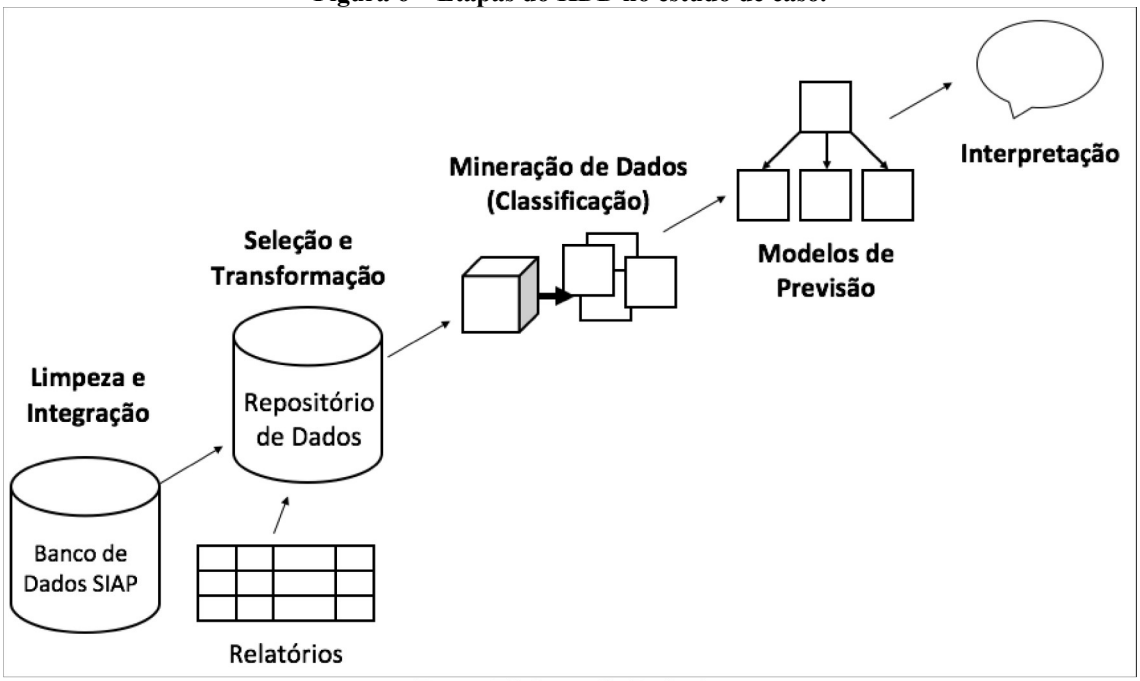
A fonte de dados deste trabalho contém dados sensíveis de cidadãos, por esse motivo, após os dados terem sido cedidos e os experimentos autorizados pela Secretaria de Estado de Segurança Pública, medidas que garantissem a confidencialidade dos dados, como controle de acesso e criptografia foram adotadas (COELHO *et al.*, 2014).

Os dados foram postos em um ambiente seguro e dados sensíveis, como telefones e nomes dos solicitantes foram descartados. O ambiente de teste foi um Macbook Pro protegido com senha de acesso, com 8GB de memória RAM (*Random Access Memory*), 2 unidades de armazenamento SDD (*Solid State Drive*) com 120GB cada, ambas com dados criptografados, processador Intel Core 2 Duo 2,66Ghz e Sistema Operacional macOS High Sierra, com todas as atualizações de segurança instaladas.

3.2 DESCOBERTA DE CONHECIMENTO

As fontes de dados utilizadas para execução do processo de KDD foram: os Bancos de Dados dos sistemas SIAP e os relatórios mensais do SINESP CAD. Os dados de ambas fontes foram limpos, integrados, selecionados e transformados para que então se procedesse a Mineração de Dados em busca de se identificar padrões e gerar modelos de previsão para então serem interpretados, possibilitando a extração do conhecimento. A Figura 5 ilustra as etapas do KDD aplicadas a este estudo de caso.

Figura 6 – Etapas do KDD no estudo de caso.



Fonte: Elaboração Própria.

Os procedimentos executados neste capítulo buscaram prever a natureza das ocorrências policiais da cidade de Rio Branco. Seguiu-se todas as etapas do KDD usando o atributo natureza como atributo alvo na tarefa de classificação.

Destaca-se que o registros de ocorrências policiais são cronologicamente ordenados, portanto, usou-se o complemento *Chronological Classify* na ferramenta Weka para executar o método de avaliação *training-test sliding validation* na tarefa de classificação.

Complementa-se que foram executados experimentos de classificação usando os algoritmos NB (*Naive Bayes*), IBK (*K-nearest neighbours classifier*) e J48 (*Decision Tree*). As medidas de acurácia foram usadas para avaliar o desempenho dos modelos de previsão. Enquanto, a análise da Matriz de Confusão foi usada para avaliar a previsibilidade das naturezas.

Em seguida, um ramo da árvore de decisão foi usado para avaliar de forma intuitiva a previsão das naturezas. Por fim, algoritmos de seleção de atributos, *CorrelationAttributeEval*, *InfoGainAttributeEval* e *ReliefFAttributeEval* foram aplicados para verificar a importância dos atributos utilizados para prever as naturezas de ocorrências.

3.2.1 LIMPEZA E INTEGRAÇÃO

Os registros de ocorrências possuem alguns dados que não são relevantes ao processo de KDD. Número de protocolo, nome e telefones dos solicitantes e identificação do atendente são exemplos de dados que não interferem de forma positiva na Mineração de Dados quando se busca prever a natureza das ocorrências, portanto foram descartados.

Quanto a qualidade dos dados presentes, destaca-se que, os sistemas de registro de ocorrências têm naturezas de ocorrências possíveis (roubo e furto como exemplo), bairros e ruas existentes no Estado do Acre pré estabelecidos e de preenchimento obrigatório. Isso evitou que erros de digitação ocorressem no preenchimento desses atributos por parte dos usuários. Em apenas em 0,11% dos registros não foram informados o bairro ou o logradouro e em apenas 0,03% dos registros estão ausentes os atributos bairro e logradouro em conjunto, provavelmente por não serem de preenchimento obrigatório em versões anteriores do sistema. Em nenhum registro está ausente o atributo da natureza da ocorrência. Desta forma, apenas os registros que não possuíam o atributo logradouro e bairro em conjunto foram descartados por não ser possível a partir deles extrair informações referentes ao local da ocorrência.

Os atributos que não estavam preenchidos em nenhum dos registros, ‘cod_evento’, ‘ocor_num_digitado’ e ‘cod_subtiponat’, também foram descartados pela ausência de valores nos atributos de todos registros ser totalmente irrelevante para Mineração de Dados.

Para realizar a integração dos dados oriundos dos sistemas SIAP e SINESP CAD, um repositório de dados utilizando o SGBD MariaDB versão 14.14 foi construído contendo os principais atributos dos registros de ocorrências das principais naturezas. Os dados do SIAP correspondem 1.509.226 ocorrências, do período de novembro de 2003 a novembro de 2017, enquanto do SINESP CAD, um total de 54.644 ocorrências do período de novembro de 2017 a julho de 2018 da cidade de Rio Branco. Portanto, o repositório de dados incluiu os registros de ocorrências do Estado do Acre do período de novembro de 2003 a julho de 2018, com um total de 1.563.870 registros de ocorrências.

O principal desafio para a tarefa de integração dos dados se deu no mapeamento das naturezas das ocorrências. Visto que, ambos sistemas utilizam naturezas distintas ou com escrita diferente para referenciar um mesmo tipo de crime. Além disso, enquanto o SIAP possuía 420 naturezas distintas, o SINESP CAD dispõe de apenas 129 tipos de natureza.

Para contornar esse problema, a Tabela 1 foi criada para compatibilizar a equivalência das principais naturezas entre os sistemas SINESP CAD e SIAP. Com a compatibilização das principais naturezas e a integração dos dados, o Repositório de Dados passou a possuir 458 naturezas distintas.

Tabela 1 – Equivalência de Naturezas do SINESP CAD com SIAP.

SINESP CAD	SIAP
Abordagem a pessoa em atitude suspeita	Abordagem atitude suspeita
Acidente de trânsito com vítima	Acidente(s) de trânsito
Acidente de trânsito sem vítima	Acidente(s) de trânsito
Ameaça	Ameaça
Dano	Dano
Direção perigosa de veículo em via pública	Direção perigosa
Dirigir veículo sem a devida CNH gerando perigo de dano	Dirigir sem a devida permissão ou habilitação
Disparo de arma de fogo	Disparo de arma de fogo
Embriaguez	Embriaguez
Conduzir veículo sob efeito de álcool ou de drogas	Embriaguez ao volante
Furto	Furto
Lesão corporal	Lesões corporais
Desobstrução de via pública	Obstrução de via pública
Animal em via pública	Omissão de cautela na guarda ou condução de animais
Perturbação da tranquilidade	Perturbação da tranquilidade
Porte ilegal de arma de fogo	Porte ilegal de arma
Porte de arma branca	Porte ilegal de arma branca
Rixa	Rixa
Roubo	Roubo
Tentativa de homicídio	Tentativa de homicídio
Tráfico de drogas	Tráfico de entorpecentes
Porte de drogas para consumo pessoal	Usuário de entorpecentes
Vadiagem	Vadiagem
Vias de fato	Vias de fato
Comunicação de violação	Violação
Violência doméstica	Violência doméstica e familiar contra a mulher

Fonte: Elaboração Própria.

Como resultado da seleção de atributos, um total de quatro atributos foram selecionados: ‘data_ocorrencia’, natureza, logradouro e bairro.

3.2.2 SELEÇÃO E TRANSFORMAÇÃO

Após a seleção dos atributos, selecionou-se os registros incluídos nas principais naturezas de ocorrências e classificou-se as não principais como **outras**. A seleção da principais naturezas, objetiva reduzir o número de naturezas a se trabalhar, o uso de um número tão variado de naturezas aumentaria ainda mais a complexidade da Mineração de Dados e Interpretação. Portanto, definiu-se como principais naturezas, as 30 naturezas com maior reincidência, de um total de 458 naturezas. As principais naturezas somaram 67,33% das ocorrências policiais observadas do período de novembro de 2003 a julho de 2018 no Estado do Acre, de acordo com a Tabela 2.

Tabela 2 – Principais Naturezas de Ocorrências Policiais em Rio Branco.

#	Natureza	Total	Porcentagem
01	Perturbação da tranquilidade	125.718	8,10%
02	Ameaça	106.705	6,88%
03	Acidente(s) de trânsito	81.481	5,25%
04	Violência doméstica e familiar contra a mulher	78.543	5,06%
05	Roubo	67.464	4,35%
06	Furto	66.107	4,26%
07	Usuário de entorpecentes	56.280	3,63%
08	Violação	54.860	3,54%
09	Abordagem atitude suspeita	54.094	3,49%
10	Agressão física	48.646	3,14%
11	Vadiagem	35.140	2,26%
12	Vias de fato	33.529	2,16%
13	Ocorrência com menor	26.671	1,72%
14	Tráfico de entorpecentes	26.166	1,69%
15	Lesões corporais	24.200	1,56%
16	Embriaguez	23.520	1,52%
17	Porte ilegal de arma	20.418	1,32%
18	Rixa	19.067	1,23%
19	Porte Ilegal de arma branca	18.658	1,20%
20	Direção perigosa	14.308	0,92%
21	Disparo de arma de fogo	10.297	0,66%
22	Dano	8.632	0,56%
23	Recuperação de bens móveis/outros	7.236	0,47%
24	Embriaguez ao volante	6.435	0,41%
25	Obstrução de via pública	5.997	0,39%
26	Disparo de alarme	5.254	0,34%
27	Dirigir sem a devida permissão ou habilitação	5.228	0,34%
28	Tentativa de homicídio	5.066	0,33%
29	Outras Fraudes (negar-se a saldar despesa)	5.002	0,32%
30	Omissão de cautela na guarda ou condução de animais	3.846	0,25%
PRINCIPAIS NATUREZAS		**1.044.568**	**67,33%**
OUTRAS NATUREZAS		**506.954**	**32,67**
TOTAL GERAL		**1.563.007**	**100%**

Fonte: Elaboração Própria.

Após a criação do repositório de dados, separou-se as ocorrências em 10 grupos, de diferentes períodos para experimentos na etapa próxima etapa que corresponde à Mineração de Dados. A seleção de diferentes períodos serve para avaliar como a redução ou aumento do intervalo de tempo de ocorrências interfere nos resultados da Mineração de Dados.

Os períodos selecionados estão representados no Quadro 2. Na coluna **Ocorrências**, consta o número de registros de cada período, na coluna **Intervalo** contém as datas aproximadas correspondentes ao período selecionado e na coluna **Período** consta o total de dias do período.

Quadro 2 – Seleção de períodos de ocorrências.

#	Ocorrências	Intervalo	Período
01	1.000	14/07/2018 a 19/07/2018	5 dias
02	2.000	08/07/2018 a 19/07/2018	11 dias
03	3.000	03/07/2018 a 19/07/2018	16 dias
04	4.000	27/06/2018 a 19/07/2018	22 dias
05	5.000	22/06/2018 a 19/07/2018	27 dias
06	10.000	28/05/2018 a 19/07/2018	≈ 2 meses (52 dias)
07	50.000	20/11/2017 a 19/07/2018	≈ 8 meses (241 dias)
08	100.000	16/06/2017 a 19/07/2018	≈ 1 ano e 1 mês (398 dias)
09	200.000	23/07/2016 a 19/07/2018	≈ 1 ano e 12 meses (726 dias)
10	300.000	04/10/2015 a 19/07/2018	≈ 2 anos e 9 meses (1019 dias)

Fonte: Elaboração Própria.

A Figura 7 mostra o código utilizado para consultar no repositório de dados o grupo de dados descrito na linha 07 do Quadro 2. Em resumo, a consulta escreve em um arquivo CSV os atributos de 1.000 registros do repositório de dados que sejam da principais naturezas de ocorrências, ordenados do mais recente para o mais antigo, para então ordenar do registro mais antigo para o mais recente. As ocorrências que não pertencem às principais naturezas tiveram suas naturezas renomeadas para a categoria ‘OUTRAS’:

Figura 7 – Consulta para Seleção e Transformação no Repositório de Dados.

```sql
SELECT 'dia_semana', 'horario', 'bairro', 'logradouro', 'natureza'
UNION ALL
SELECT * FROM
(SELECT dia_semana,horario,bairro,logradouro,natureza FROM
    (
    SELECT
        data_ocorrencia,
        DAYOFWEEK(data_ocorrencia) dia_semana,
        HOUR(data_ocorrencia) horario,
        bairro,
        logradouro,
        COALESCE(pn.natureza, 'OUTRAS') natureza
    FROM ocorrencias oco
    LEFT JOIN principais_naturezas pn ON pn.natureza = oco.natureza
    ORDER BY data_ocorrencia DESC
    LIMIT 1000
    )  selecao  ORDER BY data_ocorrencia ASC
) selecao
INTO OUTFILE '/Users/Alan/Selecao/455-50000.csv'
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n';
```

Fonte: Elaboração Própria.

O SINESP CAD disponibiliza para Policiais Militares autorizados e funcionários do CIOSP, relatórios mensais no formato de arquivo CSV. Os relatórios foram baixados e combinados em um único arquivo CSV para posteriormente serem inseridos no Repositório de Dados.

As datas e horários dos relatórios do SINESP CAD são expressas no formato brasileiro: ‘Dia/Mês/Ano Hora:Minuto’ (IBM, 2018). Porém o Bancos de Dados MariaDB, utilizado no repositório de dados trabalha com datas no formato definido pela norma internacional ISO 8601: ‘Ano-Mês-Dia Hora:Minuto’. Para converter as datas para o formato aceito no Banco de Dados, um *script* de conversão foi escrito utilizando a linguagem de programação PHP (*Hypertext Preprocessor*) ilustrado no trecho de código da Figura 8:

Figura 8 – Conversão de Datas.

```php
// Prepara a atualização
$query = 'UPDATE cad_ocorrencias SET data_hora = ? WHERE id = ? LIMIT 1'
$atualizar = $banco_de_dados->prepare($query);

// Para cada registro de ocorrência do SINESP CAD
foreach($banco_de_dados->query('SELECT * FROM cad_ocorrencias') as $linha) {

    // Converte-se a data do formato brasileiro para o padrão internacional
    $data = date('Y-m-d H:i:s',strtotime(str_replace('/', '-',$linha['data_hora'])));

    // Executa-se a atualização
    $atualizar->execute(array($data, $linha[0]));
}
```

Fonte: Elaboração Própria.

O Banco de Dados do SIAP é do tipo relacional e estruturado e está hospedado em um SGBD Microsoft SQL Server (MS SQL) 2008. Pelo fato do Repositório de Dados utilizar o SGBD MariaDB que utiliza Linguagem de Consulta Estruturada (*Structured Query Language* — SQL) diferente da utilizada pelo MS SQL, foi necessário proceder a exportação dos dados na linguagem compatível com o Repositório de Dados que utiliza o SGBD MariaDB. Para proceder a tarefa utilizou-se a função de exportação de dados do *Software* Navicat 11.

A Figura 9 contém uma captura de tela de alguns registros presentes no Repositório de Dados, resultante da Limpeza e Integração. Observa-se que o atributo 'data_ocorrencia' foi mantido e transformado nos atributos 'dia_semana' e 'horario' apenas durante a execução da Mineração de Dados.

Figura 9 – Repositório de Dados.

data_ocorrencia	natureza	bairro	logradouro
2018-06-	PERTURBAÇÃO DA TR...	CONJUNTO MANO...	RUA
2018-06-	AVERIGUAÇÕES	MOCINHA MAGAL...	RUA
2018-06-	AVERIGUAÇÕES	BOSQUE	AVENIDA NAÇÕE
2018-06-	PORTE ILEGAL DE ARM...	LOTEAMENTO PR...	AVENIDA PONTA
2018-06-	APOIO A OUTROS ORG...	SANTA INES	RUA DA
2018-06-	VIOLÊNCIA DOMÉSTIC...	VILA ACRE	AC-040

Fonte: Elaboração Própria.

3.2.3 MINERAÇÃO DE DADOS

Com as etapas anteriores foi possível limpar, integrar, selecionar e transformar os dados para a aplicação da etapa de Mineração de Dados. Para esta, foram utilizadas técnicas de classificação para criar um modelo de previsão da natureza de ocorrências usando o atributo alvo 'natureza' e os atributos preditivos 'horario' e 'dia_semana'. Os atributos 'horario' e 'dia_semana' foram utilizado com a hipótese de que criminosos agem com conduta semelhante em relação ao horário e dia da semana. Já o atributo 'logradouro' foi usado com base na hipótese de que determinadas regiões têm maior tendência a determinadas naturezas

de ocorrências. O Quadro 3 contêm exemplos dos dados presentes nos atributos utilizados.

Quadro 3 – Atributos usados para previsão da natureza das ocorrências.

#	horario	dia_semana	logradouro	natureza
01	12	7	TRAVESSA TANCREDO	OUTRAS
02	1	6	RUA DA AMIZADE	AMEAÇA
03	18	1	AVENIDA BRASIL	VIOLAÇÃO
04	2	5	RUA DO PASSEIO	ROUBO
05	17	4	AVENIDA CEARA	ACIDENTE
06	18	5	RUA MANAUS	AMEAÇA

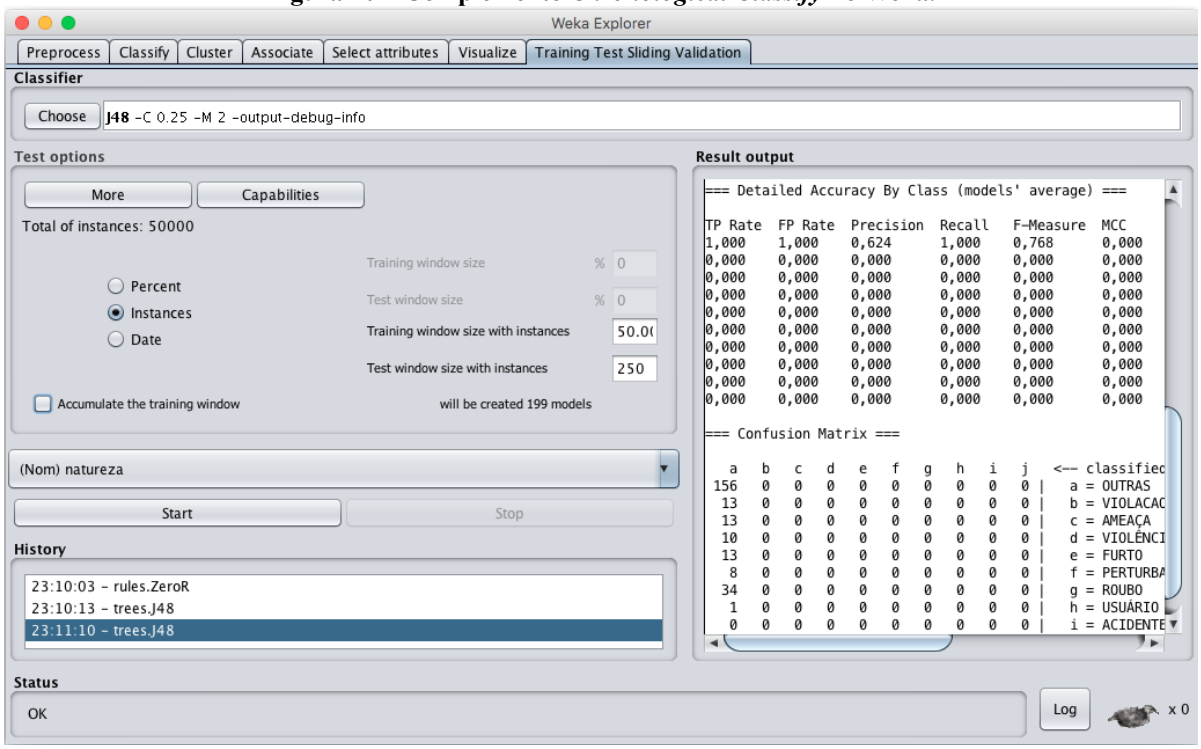
Fonte: Elaboração Própria.

Por se tratar de uma base com dados cronologicamente ordenados, foi utilizado o método de avaliação de algoritmos denominado *training-test sliding validation*. Para a execução dessa tarefa foi utilizada a ferramenta Weka, com o complemento *Chronological Classify*, desenvolvido por Lima (2017) e posteriormente melhorado por Maia (2018) para permitir ao cientista de dados customizar o tamanho dos conjuntos de treino e teste informando um número de registros ou um intervalo de datas para cada modelo de avaliação.

A Figura 10 contém a captura de tela do uso do complemento *Chronological Classify* na ferramenta Weka. Na aba *Training Test Sliding Validation* o cientista de dados seleciona o algoritmo de classificação pressionando o botão *Choose.* Em seguida, no formulário *Test options,* configura-se o método de amostragem a ser utilizado para classificação. Logo abaixo do botão, uma entre as três opções de configurações disponíveis (*Percent, Instances* e *Date*) pode ser selecionada e no lado direito do formulário as opções mudam de acordo com a opção selecionada.

Os experimentos desse estudo de caso foram configurados de acordo com números de instâncias, portanto a opção *Instances* foi selecionada e no lado direito do formulário, os campos para configuração dos conjuntos de treino e teste com base no número de instâncias foram ajustados. Abaixo do formulário seleciona-se o atributo alvo para classificação, neste estudo, o atributo natureza. Por fim, a classificação é iniciada pressionado o botão *Start.* O resultado da classificação é exibido no quadro *Result output* no lado direito da janela.

Figura 10 – Complemento *Chronological Classify* no Weka.



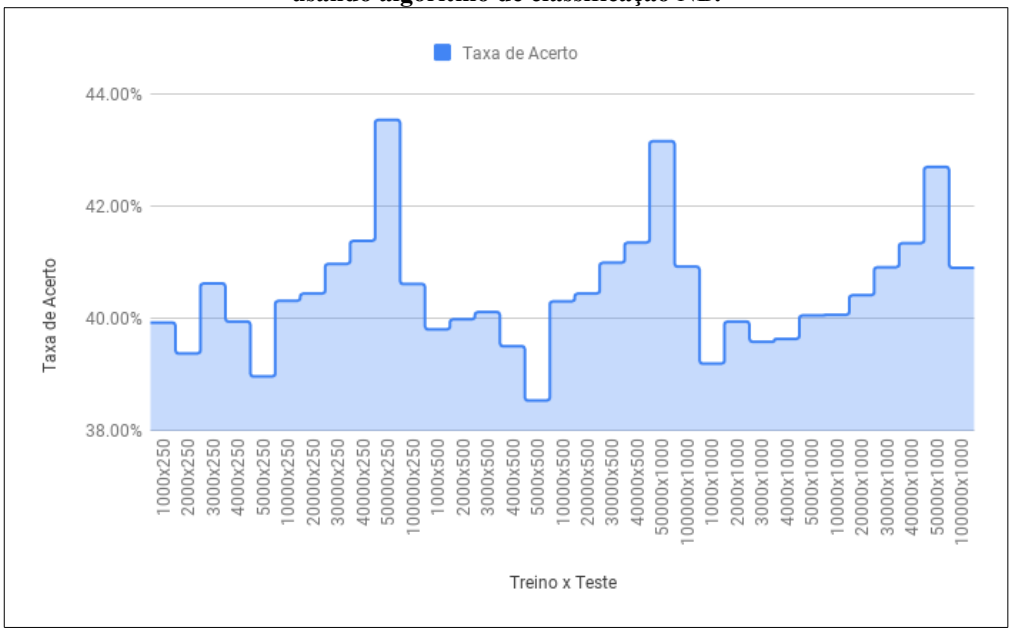
Fonte: Elaboração Própria.

Com a utilização do complemento *Chronological Classify* no Weka, iniciou-se a etapa de Mineração de Dados, buscando em um primeiro momento, descobrir a melhor configuração para o tamanho dos conjuntos de treino e teste. Utilizou-se o algoritmo de classificação NB para a execução dos experimentos, considerando sua performance e sua comprovada eficiência na classificação de ocorrências policiais (YU, 2011).

Os experimentos foram executados em 7 grupos (3.000 a 300.000) de ocorrências, especificados no Quadro 2, com 11 configurações de treino (1.000 a 100.000) e 3 de teste (250 a 1.000). Um total de 170 experimentos.

Após os experimentos iniciais, com a taxa média de acerto para cada configuração elaborou-se o gráfico da Figura 11. No eixo *x,* estão as configurações de treino-teste e em seu eixo *y* estão as médias das taxas de acerto (acurácia).

Figura 11 – Comparativo entre configurações de treino e teste do método *training-test sliding validation* usando algoritmo de classificação NB.



Fonte: Elaboração Própria.

Com a análise do gráfico, foi possível perceber que a configuração de treino com 50.000 ocorrências, o equivalente a aproximadamente 9 meses de ocorrências, produziu melhores resultados. Quanto a configuração de teste, a configuração com 250 ocorrências que equivale a aproximadamente 6 horas de ocorrências se sobressaiu em relação as demais.

Portanto, os experimentos seguintes tiveram como base uma configuração de 50.000 ocorrências para treino e 250 ocorrências para teste, o que representa em unidade de medida de tempo: aproximadamente 9 meses para treino e 6 horas para teste.

Com uma configuração de treino e teste definida, buscou-se avaliar e comparar os resultados de outros algoritmos de classificação. Os experimentos foram executados usando os algoritmos de classificação NB, J48 e IBK. Os resultados da acurácia podem ser observados na Tabela 3.

Tabela 3 – Acurácia dos algoritmos de classificação.

	Acurácia		
Registros	**NB**	**J48**	**IBK**
50.000	47,20%	50,03%	33,60%
100.000	47,12%	34,04%	34,14%
200.000	41,78%	42,61%	21,26%
300.000	37,63%	31,19%	27,92%
Média	**43,43%**	**39.46%**	**29,23%**

Fonte: Elaboração Própria.

No comparativo foi possível notar que usando dados de treino mais próximos dos de teste, obtêm-se uma melhor taxa de acerto. Ocorrências de até 9 nove meses antes (50.000 registros) para treino trouxeram melhores resultados. Observou-se também que o algoritmo NB obteve, em média, uma taxa de acerto ligeiramente maior que outros algoritmos na tarefa de previsão da natureza das ocorrências. Por esse motivo decidiu-se utilizá-lo como algoritmo base para a execução dos próximos experimentos.

Com o objetivo de melhorar o resultado, é importante destacar que o algoritmo Naive Bayes (NB) foi reavaliado com ajustes nos parâmetros de entrada para verificar a influência no aumento da sua taxa de acerto. Para isso, os experimentos foram reexecutados alterando os valores dos parâmetros *useSupervisedDiscretization* para que o algoritmo faça uso de discretização supervisionada para converter atributos numéricos em discretos e o *useKernelEstimator* para que se estime um centro para atributos numéricos em vez de uma distribuição normal. Porém alterações nesses parâmetros, pouco influenciaram na eficácia da classificação, com variação da taxa de acerto em 0,04% para mais ou para menos.

Em seguida, buscou-se investigar a acurácia do melhor algoritmo de classificação em relação ao número de naturezas possíveis e como isso interfere na previsão.

A Tabela 4 contém os resultados das taxas de acerto para cada cenário de número de naturezas e de quantidade de registros usando o algoritmo de classificação NB e o mesmo método de avaliação usado nos experimentos anteriores.

Tabela 4 – Taxa de acerto entre diferentes quantidades de naturezas.

Registros	**Quantidade de Naturezas**						
	5	**10**	**15**	**20**	**25**	**30**	**455**
50.000	70,00%	59,20%	51,60%	49,20%	48,00%	47,20%	31,45%
100.000	71,41%	61,01%	56,02%	51,83%	48,54%	47,12%	17,89%
200.000	67,67%	56,00%	50,78%	46,64%	43,67%	41,78%	18,99%
300.000	66,05%	51,41%	45,76%	41,72%	38,99%	37,63%	17,66%
Média	**68,78%**	**56,90%**	**51,04%**	**47.34%**	**44,80%**	**43,43%**	**21,49%**

Fonte: Elaboração Própria.

Por fim, buscou-se saber como o modelo de previsão se comportou na previsão das principais naturezas de ocorrências. Para isso, analisou-se a Matriz de Confusão do experimento com melhor resultado, 71,41% usando as 5 principais naturezas e 100.000 registros, descrito na Tabela 4. Porém, pelo fato das ocorrências de natureza ‘outras’ representarem uma grande percentagem das ocorrências, o algoritmo de classificação direcionou em grande maioria as previsões para esta natureza, o que resultou na prevalência

de números iguais a zero na matriz. Então, o procedimento adotado para este experimento em específico para interpretação foi o de remover da seleção as naturezas de natureza ‘outras’ e reexecutar o experimento.

Para melhor compreensão, as naturezas foram abreviadas em três caracteres: ROU: ‘Roubo’, PER: ‘Perturbação da Tranquilidade’, AME: ‘Ameaça’, VIO: ‘Violação’ e ACT: ‘Acidente de Trânsito’. Na matriz, as naturezas reais estão nas linhas e as naturezas previstas nas colunas. Portanto, para a leitura dos acertos de previsão de cada natureza, observa-se os valores da diagonal principal da matriz. Os valores fora da diagonal principal correspondem aos erros de previsão. A matriz de confusão do experimento está ilustradas na Figura 12, para melhor visualização, os elementos da diagonal principal foram destacados com retângulos:

Figura 12 – Matriz de Confusão.

	ROU	PER	AME	VIO	ACT
ROU	5918	2420	1709	2720	1502
PER	2123	4742	614	1598	130
AME	2608	1813	1318	2212	803
VIO	3434	3145	1577	3786	520
ACT	1019	495	451	594	2749

Fonte: Elaboração Própria.

Analisando a Matriz de Confusão, percebeu-se que por ser a classe com maior número de registros, a classe ‘Roubo’**,** foi a principal responsável nos erros de previsão. Outras naturezas com maior número de ocorrências também tiveram influência nas outras, não por ter características similares e sim por terem maior presença nos registros. Porém ainda foi possível fazer algumas inferências analisando a Matriz de Confusão. Nota-se que ocorrências de ‘Violência doméstica’ foram confundidas 3145 vezes com ‘Perturbação da tranquilidade’, é provável que ambas tenham características semelhantes. Observa-se também um grande dificuldade em prever ocorrências de ‘Ameaça’. Em contrapartida, ‘Acidente de Trânsito’ apresenta-se como a com menor número de erros na previsão.

Com o objetivo de verificar de forma mais intuitiva a ocorrência das naturezas, o experimento anterior foi reexecutado, porém utilizando o algoritmo de classificação baseado em indução de árvores de decisão J48. A árvore mostra as regras de cada nó folha utilizadas para realizar a previsão. Com isso, foi possível visualizar como o algoritmo utilizou os

atributos e decidiu a previsão das ocorrências. Não foi possível usar o complemento *Chronological Classify* nesse experimento. Por um motivo desconhecido este não exibe a árvore de decisão do algoritmo J48. A alternativa encontrada foi executar a classificação com o método de amostragem *hold out* com as 5 principais naturezas e 50.000 registros, usando 90% para treino, 45.000 ocorrências e 10% para teste, 5.000 ocorrências. Foi também ativada a opção *Preserve order for % split,* para que o algoritmo não selecionasse as ocorrências aleatoriamente, preservando a ordem cronológica dos registros.

A Figura 13 contêm uma seção da árvore de decisão para previsão de ocorrências no 'logradouro = ESTRADA DA SOBRAL'. Os valores entre parênteses correspondem aos acertos e erros de previsão ocorridos na fase de teste. Então é possível perceber na linha 6 que, nos domingos após as 01:00, previu-se ocorrências como 'Perturbação da Tranquilidade' com aproximadamente 10 acertos e em torno de 6 erros:

Figura 13 – Árvore de Decisão da Estrada da Sobral com algoritmo J48.

```
1  logradouro = ESTRADA DA SOBRAL
2  |   horario <= 4
3  |   |   dia_semana <= 1
4  |   |   |   horario <= 1: OUTRAS (16.59/10.33)
5  |   |   |   horario > 1
6  |   |   |   |   horario <= 3: PERTURBAÇÃO DA TRANQUILIDADE (10.33/6.2)
7  |   |   |   |   horario > 3: OUTRAS (7.19/3.09)
8  |   |   dia_semana > 1
9  |   |   |   dia_semana <= 2
10 |   |   |   |   horario <= 3: OUTRAS (31.43/9.15)
11 |   |   |   |   horario > 3: AMEAÇA (5.11/2.11)
12 |   |   |   dia_semana > 2
13 |   |   |   |   horario <= 0
14 |   |   |   |   |   dia_semana <= 6: OUTRAS (11.54/2.26)
15 |   |   |   |   |   dia_semana > 6: VIOLÊNCIA DOMÉSTICA E FAMILIAR CONTRA A MULHER (6.2/3.2)
16 |   |   |   |   horario > 0
17 |   |   |   |   |   dia_semana <= 6
18 |   |   |   |   |   |   horario <= 2
19 |   |   |   |   |   |   |   dia_semana <= 4
20 |   |   |   |   |   |   |   |   horario <= 1: VIOLACAO (8.26/4.22)
21 |   |   |   |   |   |   |   |   horario > 1: OUTRAS (14.22/6.11)
22 |   |   |   |   |   |   |   dia_semana > 4: OUTRAS (22.46/8.17)
23 |   |   |   |   |   |   horario > 2
24 |   |   |   |   |   |   |   horario <= 3: OUTRAS (11.33/6.19)
25 |   |   |   |   |   |   |   horario > 3: VIOLACAO (7.22/5.17)
26 |   |   |   |   |   dia_semana > 6: OUTRAS (35.56/11.2)
```

Fonte: Elaboração Própria.

Por fim, objetivando saber quais atributos são mais relevantes para prever a natureza das ocorrências, foi avaliado como os atributos utilizados influenciaram na previsão da natureza das ocorrências. Para isso, utilizou-se os algoritmos de Seleção de Atributos *CorrelationAttributeEval, InfoGainAttributeEval* e *ReliefFAttributeEval*.

Os resultados da avaliação dos atributos com o uso dos algoritmos de seleção de

atributos estão presentes no Quadro 4. Os valores em parênteses representam a classificação do atributo em relação aos outros, quanto menor o valor entre parênteses, melhor a classificação portanto.

Quadro 4 – Avaliação dos atributos.

Algoritmo	dia_semana	horario	logradouro
CorrelationAttributeEval	0,01752 (1)	0,0469 (2)	0,00695 (3)
InfoGainAttributeEval	0,0689 (3)	0,2032 (2)	0,8055 (1)
ReliefFAttributeEval	0,00358 (3)	0,00626 (2)	0,09733 (1)

Fonte: Elaboração Própria.

Os resultados mostram que o atributo logradouro tem grande relevância na previsão da natureza da ocorrência, reforçando a hipótese da natureza da ocorrência estar associada com a região. Também, conclui-se que a ordem de relevância dos atributos para previsão da natureza das ocorrências em Rio Branco é: ‘logradouro ‘, ‘horario’ e ‘dia_semana’.

Com a conclusão da Mineração de Dados, procedeu-se a interpretação dos resultados obtidos nesta etapa do KDD, presente na subseção 3.2.4.

3.2.4 INTERPRETAÇÃO

Os modelos de previsão da Mineração de Dados, usando o algoritmo de Naive Bayes, obtiveram uma taxa de acerto média de 43,43% no cenário com as 30 principais naturezas e média de 68,78% no cenário com 5 principais naturezas. Observou-se que os 9 meses anteriores correspondem a melhor seleção de ocorrências a ser usada na busca de prever a natureza das próximas ocorrências. Isso demonstra que locais e períodos mais propícios a ter crimes de uma natureza podem ser diferentes com o passar do tempo.

Analisando a Matriz de Confusão resultante da classificação das 5 principais ocorrências, foi possível notar que as ocorrências de violência doméstica, em um maior número de vezes foram classificadas como de perturbação da tranquilidade. Inferiu-se que ambas têm características de dias da semana e horários semelhantes. É nos finais de semana que amigos e familiares se reúnem para festejar e o som alto nos períodos noturnos acaba

resultando em vizinhos denunciando perturbação da tranquilidade. É também nos finais de semana que cônjuges abusam do consumo de bebidas alcoólicas e agridem suas esposas ao retornarem para casa.

Percebeu-se também uma maior dificuldade em prever ocorrência de ameaça, esta, por ter características e motivações bastante diversificadas, dificulta a sua associação com os parâmetros disponíveis. Em contrapartida, viu-se as ocorrências de acidente de trânsito como as mais previsíveis. Ocorre que acidentes de trânsito têm dias da semana e horários bem definidos. O maior número ocorre nos horários que as pessoas estão indo ou voltando do trabalho, nos dias úteis da semana. São nesses dias e horários que se têm maior tráfego de veículos nas ruas, portanto mais acidentes.

Ao analisar a árvore de decisão foi possível fazer inferências bem específicas em relação aos inúmeros logradouros da cidade de Rio Branco. Analisando o ramo da árvore que se refere ao logradouro Estrada da Sobral. Percebeu-se que entre as 00:00 e 03:00 dos domingos aumenta a probabilidade de ocorrências de Perturbação da tranquilidade. De segunda-feira a terça-feira, após as 03:00, têm-se mais chances de ocorrências de ameaça. De quarta-feira a sábado têm-se uma concentração maior de ocorrências de violência doméstica e de violação de domicílio. Porém com maior probabilidade para violência doméstica após os sábados.

Com a execução da Mineração Dados, pôde se constatar a alta complexidade em prever a natureza de crimes na cidade de Rio Branco. Porém, as técnicas utilizadas para Mineração de Dados mostraram potencial para apresentar resultados que podem ser utilizados para melhorar o planejamento policial. Ao observar as prováveis ocorrências que podem ocorrer em determinado horário e logradouro, ações preventivas focadas nas naturezas de ocorrência previstas podem influenciar na redução de crimes.

4 CONSIDERAÇÕES FINAIS

Nesta seção são apresentadas as considerações finais na subseção 4.1 e as recomendações para trabalhos futuros na subseção 4.2.

4.1 CONSIDERAÇÕES FINAIS

Este trabalho apresentou o problema do aumento da criminalidade no Estado do Acre, propondo o uso de Técnicas de Mineração de Dados para disponibilizar conhecimentos que pudessem ser usados pela gestão da PMAC em seus planejamentos.

Realizou-se um estudo do processo de descoberta de conhecimento em base de dados, conhecido como KDD, em que se apresentou a Mineração de Dados como uma fase do processo. Apresentou-se também algumas ferramentas a serem usadas para a execução do processo. Após o estudo do KDD, este foi aplicado nos registros de ocorrências policiais da cidade de Rio Branco, disponibilizados pela SESP AC. A execução de todos as fases do processo foram detalhadas no estudo de caso.

O estudo de caso iniciou-se com a Limpeza e Integração de Dados, onde dados de dois sistemas de registro ocorrências foram limpos e integrados. Prosseguiu com a Seleção e Transformação, em que selecionou e transformou os atributos relevantes e selecionou grupos de registros das principais naturezas de ocorrências, para serem utilizados na Mineração de Dados.

A Mineração de Dados, compreendeu uma série de experimentos e análises que foram executadas buscando verificar a taxa de acerto dos modelos de previsão e a produção de artefatos que pudessem ser utilizados para fazer inferências relacionadas a previsão da natureza de ocorrências policiais na última etapa do processo de KDD, a Interpretação.

Apesar da complexidade do problema e recursos computacionais limitados para execução dos algoritmos de Mineração de Dados, foi possível obter uma taxa de acerto de até 71,41% na previsão das 5 principais ocorrências policiais da cidade de Rio Branco.

4.2 RECOMENDAÇÕES

Sugere-se para trabalhos futuros a integração de outros atributos preditivos ou fontes de dados para aplicação de outras técnicas de mineração de dados, como por exemplo: a clusterização de ocorrências policiais em busca de grupos com características semelhantes para então executar experimentos de classificação usando estes grupos como atributos preditivos; e a extração de regras de associação para verificar a relação entre as características das ocorrências policiais.

Recomenda-se também o uso de Técnicas de Mineração de Textos nos históricos de ocorrências produzidos pelos atendentes, para extrair outros atributos a serem usados na previsão da natureza.

Outra sugestão para trabalho futuro é que se desenvolva um Sistema de Apoio à Decisão que possa ser utilizado por gestores de forças policiais a fim de disponibilizar de forma facilitada o conhecimento obtido na previsão da natureza das ocorrências policiais e de outros tipos de previsão.

REFERÊNCIAS

AGRAWAL, M. **Impact of Mobile Computing Terminals in Police Work**, 2003.

AGRAWAL, R.; IMIELINSKI T.; SWAMI A. **Mining Association Rules between Sets of Items in Large Databases**, 1997.

BELLAZZI, R; ZUPAN, B. **Predictive data mining in clinical medicine:** Current issues and guidelines. International Journal of Medical Informatics v. 77, n. 2, p. 81–97 , 2008.

BRAZ, L. M.; DERMEVAL, D.; VÉRAS D.; LIMA, M.; TIENGO W. **Aplicando Mineração de Dados para Apoiar a Tomada de Decisão na Segurança Pública do Alagoas**, 2009.

BRAZ, F. J.; COAN W. S.; ROSSETI A. **Uma proposta de Solução de Mineração de Dados aplicada à Segurança Pública, Simpósio Brasileiro de Sistemas de Informação**, 2012.

BOENTE, A. N. P.; GOLDSCHMIDT R. R.; ESTRELA, V. V. **Uma Metodologia de Suporte ao Processo de Descoberta de Conhecimento em Bases de Dados**. Rio de Janeiro: V Simpósio de Excelência em Gestão e Tecnologia, 2008, Resende - RJ. V SEGeT, 2008.

CAMILO, C. O.; SILVA, J. C.. **Mineração de Dados**: Conceitos, Tarefas, Métodos e Ferramentas. Goiás, Brasil: Instituto de Informática Universidade Federal de Goiás, 2009.

COELHO, F., E. S.; ARAÚJO, L., G., S, BEZERRA, E. K. **Gestão da Segurança da Informação**: NBR 27001 e NBR 27002, Rio de Janeiro, 2014

DIAS, M. M. **Um modelo de formalização do processo de sistema de descoberta de Conhecimento em banco de dados**. 2001. Tese (Doutorado)-Pós Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina. Florianópolis, Santa Catarina, 2001.

FAYYAD, U.; PIATETSKI-SHAPIRO, G.; SMYTH, P. **The KDD**: Process for Extracting Useful Knowledge from Volumes of Data. 1996.

GANTZ, J; REINSEL D. **The Digital Universe In 2020**: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. (2012). <http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf> Acesso em: 1 de Ago. 2018.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining**: um guia prático. 2. ed. Rio de Janeiro/RJ: Elsevier, 2005.

HAN, J.; KAMBER, M.; PEI J. **Data Mining - Concepts and Techniques**, Morgan Kaufmann Publishers, Inc, 2012.

IBM. IBM Knowledge Center: **Portuguese (Brazil) (pt-BR)**: Formats for Date Outputs, 2018. <https://www.ibm.com/support/knowledgecenter/en/SSS28S_3.0.0/com.ibm.help.forms.doc/locale_spec/i_xfdl_r_formats_pt_BR.html> Acesso em: 02 de ago. 2018.

LEE H. D. **Seleção de atributos importantes para a extração de conhecimento de bases de dados**, São Carlos, 2005

LIMA JÚNIOR, MANOEL LIMEIRA DE. **Previsão de Integradores e Tempo de Vida de Pull Requests**, Niterói, 2017.

LIMA, M. W. S., **Uma Extensão da Ferramenta Weka para a Avaliação de Tarefas Preditiva**, 2017.

MAIA B., **Análise de Variações do Método de Avaliação Janela Deslizante em Modelos Preditivos:** Um Estudo De Caso No Contexto De Pull Requests, Rio Branco, 2018.

KORTH, H.F. e SILBERSCHATZ, A. **Sistemas de Bancos de Dados**, Makron Books, 2a. edição revisada, 1994.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C.; BYERS, A. H. **Big data**: The Next Frontier For Innovation, Competition, And Productivity, 2011. Disponível em: <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>. Acesso em: 01 de Ago. 2018.

MEDEIROS, E. A. **Técnica de Aprendizagem de Máquina para Categorização de Textos**, Recife, 2004.

MITIK, **Merve et al. Data Mining Approach for Direct Marketing of Banking Products with Profit/Cost Analysis**. The Review of Socionetwork Strategies v. 11, n. 1, p. 17–31 , 2017

PRASS, **Fernando Sarturi. Estudo Comparativo entre Algoritmos de Análise de Agrupamento em Data Mining**. Santa Catarina, 2012.

SIN, K; MUTHU L. **Application of Big Data in Education Data Mining and Learning**

Analytics: A Literature Review, 2015. <http://ictactjournals.in/paper/IJSC_V5_I4_paper6_1035_1049.pdf> Acesso em: 31 de jul. 2018.

SPI. **Social Progress Index**. 2017. Disponível em: <http://www.socialprogressindex.com/?tab=2&code=BRA> Acesso em: 25 de jul. 2018.

SRIKANT R.; VU Q.; AGRAWAL R. **Mining Association Rules with Item Constraints**, 1997.

STAIR, R. M; REYNOLDS, G. W. **Princípios de Sistemas de Informação**. 9. ed. São Paulo: Cengage Learning, 2012.

YU, C.; WARD, M. W.; MORABITO M.; DING, W. **Crime Forecasting Using Data Mining Techniques**, 2011.

WAISELFISZ, J. J. **Mapa da Violência 2016**: Homicídios por armas de fogo no Brasil, 2016.

WAZLAWICK, R. S. **Metodologia de Pesquisa para Ciência da Computação.** Elsevier, 2014.