



Uso de Dados Provenientes de Rede Social e Técnica de Mineração de Dados para Classificar Crimes em Belém-PA

Furtado LS*, Souza AMF*

**Universidade da Amazônia (UNAMA) – Belém, Pará, BR*

Resumo. A Grande Belém enfrenta o desafio de reduzir a taxa de criminalidade na capital e nos municípios. O ano de 2017 foi o mais violento da história: a recordista marca de 4.416 mortes violentas foi registrada no referido estado, com a média diária de 12 mortes. Inserido neste contexto, este trabalho tem como objetivo identificar e classificar crimes realizados em determinados bairros de Belém utilizando a técnica de mineração de dados. Na primeira etapa foi realizada a extração e seleção dos dados usando a ferramenta computacional python e armazenando-os em uma planilha eletrônica (Excel). Na segunda etapa, aconteceu o pré-processamento, a organização dos dados, ou seja, realizamos a limpeza, integração, transformação e redução do conjunto de dados. No penúltimo momento, foi utilizado o algoritmo chamado Árvore de Classificação e Regressão nos dados coletados utilizando a biblioteca SciKit-Learning. A última etapa do processo foi a avaliação dos dados das quatro métricas de desempenho estabelecidas: acurácia, taxa de erro, sensibilidade e especificidade. Através da Utilização da seleção dos dados foi possível sumarizá-los, assim como, dos testes e experimentos realizados com a técnica de mineração de dados, foi possível identificar padrões escondidos nas postagens em redes sociais, facilitando a identificação das ocorrências criminosas em Belém-PA. Dessa forma, a sociedade tomará conhecimento da situação que se encontra os crimes na cidade nos seus respectivos meses, dias da semana, períodos do dia e bairros no intuito de se proteger ou ficar em alerta ao andar por determinados locais onde a incidência de crimes é grande.

Palavras-Chave: *Mineração de Dados, Árvore de Decisão, Crimes em Belém do Pará*

Abstract. Greater Belém faces the challenge of reducing crime rates in the capital and in municipalities. The year 2017 was the most violent in history: the record mark of 4,416 violent deaths was registered in that state, with a daily average of 12 deaths. In this context, this work aims to identify and classify crimes carried out in certain neighborhoods of Belém using the technique of data mining. In the first step the data extraction and selection was performed using the python computational tool and stored in a spreadsheet (Excel). In the second stage, the preprocessing, the organization of the data was performed, that is, we performed the cleaning, integration, transformation and reduction of the data set. In the penultimate moment, the algorithm called Classification and Regression Tree was used in the data collected using the SciKit-Learning library. The last step of the process was the evaluation of the data of the four performance metrics established: accuracy, error rate, sensitivity and specificity. Through the use of data selection it was possible to summarize them, as well as the tests and experiments performed with the data mining technique, it was possible to identify hidden patterns in social network posts, facilitating the identification of criminal occurrences in Belém-PA. In this way, society will be aware of the

situation of crimes in the city in their respective months, days of the week, times of day and neighborhoods in order to protect themselves or to be alert when walking through certain places where the incidence of crimes is big.

Keywords: Data Mining, Decision Tree, Crimes in Belém *do Pará*

Introdução. Mesmo com os investimentos divulgado pelo governo do estado do Pará, a violência continua elevada, o Pará fechou o ano de 2017 entrando para história com elevados índices de assassinatos já catalogados de criminalidade no estado, demonstrando assim, o real insucesso na gestão das políticas públicas atuais.

De acordo com dados obtidos pelo Diário do Pará (2018)¹, foram registradas, em todo território paraense, em 2017, a recordista marca de 4.416 mortes violentas, com a média diária de 12 mortes, ou seja, a cada duas horas uma pessoa é morta no Pará. Os dados indicam que as atuais políticas de segurança pública estão sendo ineficazes e as perspectivas/projeções quanto aos índices de violência para 2018 seguirão a mesma tendência com recordes de violência. Segundo outra pesquisa mais recente do Diário do Pará², o índice de violência aumentou em aproximadamente em 20% no Pará em relação ao ano anterior, fato que confirma a calamidade pública e crise de segurança do estado. Conforme detalhado por Cerqueira e Soares (2016), o problema da violência letal afeta não apenas as gerações presentes, mas, também as futuras, cujo montante os economistas chamam de custo de bem-estar econômico devido aos homicídios.

Segundo o G1 Pará (2018), a Secretaria de Segurança Pública reconhece que o combate à violência feito pela polícia não é suficiente para resolver o problema de vez. De acordo com o especialista em segurança pública Miranda (2018)³ informou ao jornal que:

“Demonstra que nós estamos perdendo o controle. A polícia não está conseguindo agir de maneira preventiva buscando a identificação, ela não está conseguindo atuar principalmente no enfrentamento do tráfico e do crime organizado dentro do estado. Outro fator é: a justiça não está conseguindo dar uma resposta efetiva” (MIRANDA, 2018).

A partir da análise acima, junto ao estudo realizado pelo Instituto de Pesquisa Econômica e Aplicada (IPEA) onde fez um mapa da violência no Brasil com base nos dados do Sistema de informações sobre Mortalidade (SIM) que é um sistema federal desenvolvido pelo Ministério da Saúde através do Datasus. Esse estudo mostrou que entre 2005 e 2015 houve um aumento de 62%

¹ Diário do Pará. (2018). “**Mais de 4.400 pessoas foram assassinadas no Pará em 2017**”. Disponível em: <<https://www.diarioonline.com.br/noticias/para/noticia-478152-.html>>. Acessado em: março de 2018.

² Diário do Pará. (2018). “**Violência aumenta quase 20% no Pará e polícia não dá conta**”. Disponível em: <<https://www.diarioonline.com.br/noticias/policia/noticia-481110-.html>>. Acessado em: março de 2018.

³ MIRANDA, W. “**No Pará, um assassinato foi registrado a cada duas horas em Janeiro de 2018**” Disponível em: <<https://g1.globo.com/pa/para/noticia/no-para-um-assassinato-foi-registrado-a-cada-duas-horas-em-janeiro-de-2018.ghtml>>. Acessado em: janeiro de 2018.

na taxa de homicídios no Pará. De acordo com o 11º Anuário Brasileiro de Segurança Pública, viver em Belém é três vezes mais perigoso do que no Rio de Janeiro. Belém é a segunda capital mais perigosa do Brasil, com taxa de 64,9 crimes violentos a cada 100 mil habitantes.

Este artigo utiliza dados provenientes da rede social Twitter, para analisar a violência nos bairros da Região Metropolitana de Belém (RMB). Extração, filtragem e análise estatística dos dados foram realizadas. Depois disso, usou-se técnica de Mineração de Dados (MD) conhecida, como, Árvore de Decisão que tem a capacidade de criar um modelo viável que preveja o valor de uma variável de destino com base no conjunto de variáveis de entrada (BELL, 2014). As árvores de decisão podem ser aplicadas a problemas de regressão e de classificação (JAMES et al., 2013; GOLLAPUDI, 2016).

Para verificar a qualidade dos resultados encontrados pela Árvore de Decisão, as seguintes métricas de desempenho foram usadas: acurácia, taxa de erro, sensibilidade e especificidade.

Portanto, buscou-se reunir dados/informações da rede social Twitter com o propósito de responder ao seguinte problema de pesquisa: Que variáveis impactam na ocorrência de crimes na Região Metropolitana de Belém (RMB)?

Objetivo geral. Identificar e classificar crimes realizados em determinados bairros de Belém utilizando dados provenientes de redes sociais e técnicas de mineração de dados.

Objetivos específicos. Como objetivos específicos, pretende-se:

- Utilizar dados provenientes de rede social para analisar a violência nos bairros de Belém-PA;
- Usar script Python para extrair e filtrar dados (raspagem de dados);
- Utilizar ferramentas de planilha eletrônica para fazer a análise estatística e gerar gráficos para visualizar os resultados;
- Usar o pacote SciKit Learning (Python) para aplicar o algoritmo de Árvore de Decisão (ou outros algoritmos) nos dados extraídos;
- Verificar a qualidade dos resultados encontrados pela Árvore de Decisão.

Materiais e métodos. Neste tópico serão apresentadas as informações metodológicas de como foi desenvolvida a pesquisa qualitativa, que se constitui por métodos, formas e objetivos na busca da qualificação dos dados coletados para que se possa traduzir e expressar o sentido dos fenômenos sociais em estudo, apresentando como principal horizonte e base às postagens na rede social Twitter relacionadas a crimes para a realização e efetivação do trabalho e o público alvo. A descrição será exposta nas fases do processo de KDD que foram necessárias ao nosso trabalho, sendo: extração, seleção, pré-processamento, mineração e avaliação dos dados.

Extração e Seleção dos dados. Foi realizada a extração dos dados, onde usou-se, a API do Twitter através da ferramenta computacional python IDLE na versão 2.7.14, usando o script de Masanori (2018), tendo início e término no dia 25 de Janeiro, sendo projetado para coletar dados apenas dos últimos doze meses a contar da raspagem de dados, inserindo-os em uma planilha eletrônica (Excel) no formato CSV (Comma-separated values ou Valores Separados por Vírgula).

Para Weiss (2005), entender os dados que serão utilizados é crucial para o sucesso no desenvolvimento das aplicações que serão desenvolvidas.

Pré-processamento. A etapa de pré-processamento é essencial para se adquirir um resultado satisfatório e requer grande parte de dedicação do tempo nessa fase. Segundo Olson e Delen (2008), a referida fase pode compreender mais de 50% do tempo em projetos de mineração de dados. Portanto, nessa etapa, aconteceu a organização dos dados, seguindo a visão de Han et al. (2006), realizamos a limpeza, integração, transformação e redução do conjunto de dados.

- **Limpeza dos dados.** A limpeza dos dados consiste em eliminar as inconsistências encontradas: atributos incompletos, valores errados, valores nulos, etc. Neste caso, optou-se por remover os registros vazios, seis colunas que não seriam úteis a priori (como código, data da postagem, tempo, texto, n° de retweets, n° de favoritos) e aglutinou-se palavras similares de tipos de crimes tais como: (Roubo é composto por arrastão + arrombaram + assalto + assaltado + Roubado); (Acidente é composto por ferido + caiu + queda + capota + capotamento + colisão + atropelado + atropelamento + batida + incêndio + explosão + queimadura + afogado + afogamento); (Esfaqueado é composto por esfaqueamento + esfaqueou); (Estuprado é composto por Estupro); (Briga é composto por confusão).
- **Integração dos dados.** A integração dos dados consiste em uma análise aprofundada dos dados observando redundância, categorias diferentes para os mesmos valores, chaves divergentes, mesmos dados, entre outros. No conjunto de dados utilizado, grande parte dos atributos foram desconsiderados por serem redundantes.
- **Transformação dos dados.** Na etapa de transformação dos dados, foram feitas algumas conversões de atributos com sequências de caracteres em dados quantitativos. No Python, as bibliotecas utilizadas para criação dos modelos já aplicam algumas técnicas de transformação dos dados internamente de acordo com o algoritmo a ser utilizado.
- **Redução dos dados.** A redução do conjunto de dados é importante, pois, em alguns casos, o volume de dados usado no processo de mineração de dados é considerado muito grande a ponto de tornar-se impraticável a aplicação das técnicas e até mesmo a análise dos dados (HAN E KAMBER, 2006). Desta forma, somente os crimes de Roubo, Acidente e Homicídio da cidade de Belém-PA foram selecionados resultando em 547 registros.

Mineração de dados. No penúltimo momento, foi aplicado a mineração de dados para exploração de informações escondidas nos dados e descobertas de padrões, pois, aplicou-se o algoritmo de



árvore de decisão nos dados coletados utilizando a biblioteca SciKit-Learning, a técnica foi útil para modelar a árvore de decisão.

Avaliação dos dados. E o processo foi finalizado com avaliação dos dados que, para verificar a qualidade dos resultados encontrados pela árvore de decisão, as seguintes métricas de desempenho foram usadas:

- ✓ **Acurácia:** a proporção de predições corretas, sem levar em consideração o que é positivo e negativo, ou seja, é a capacidade do método de acertar o diagnóstico. Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho do sistema;
- ✓ **Taxa de erro:** é a proporção de predições incorretas de blocos recebidos do número total de blocos enviados para o diagnóstico, ou seja, é usado para expressar a relação entre o número de blocos de informações recebidos com erro na recepção a partir do número total de blocos de informações enviados na transmissão;
- ✓ **Sensibilidade:** a proporção de verdadeiros positivos, ou seja, é a capacidade do sistema em prever corretamente a condição para casos que realmente a têm; e
- ✓ **Especificidade:** a proporção de verdadeiros negativos, ou seja, é a capacidade do sistema em prever corretamente a ausência da condição para casos que realmente não a têm.

Sumário dos Dados

Com base nos dados citados na figura 1, pôde-se analisar que o Roubo foi apontado 409 vezes, Acidente com 347 e Homicídio com 251.

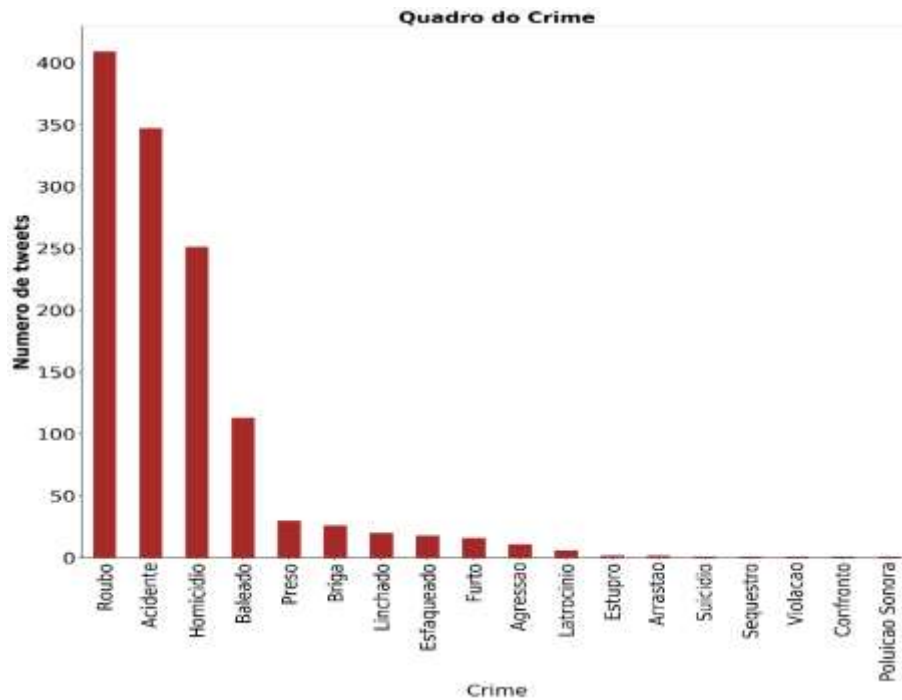


Figura 1. Crimes em Belém do Pará entre os Anos de 2017 – 2018

Com a exposição dos dados na figura 2, é possível analisar que do total de crimes, 563 vezes aconteceram no período noturno (18 <= hour <= 23), em seguida com 384 ficou o Vespertino (12 <= hour < 18), Matutino (6 <= hour < 12) com 187 e Madrugada (0 <= hour < 5) com 122.

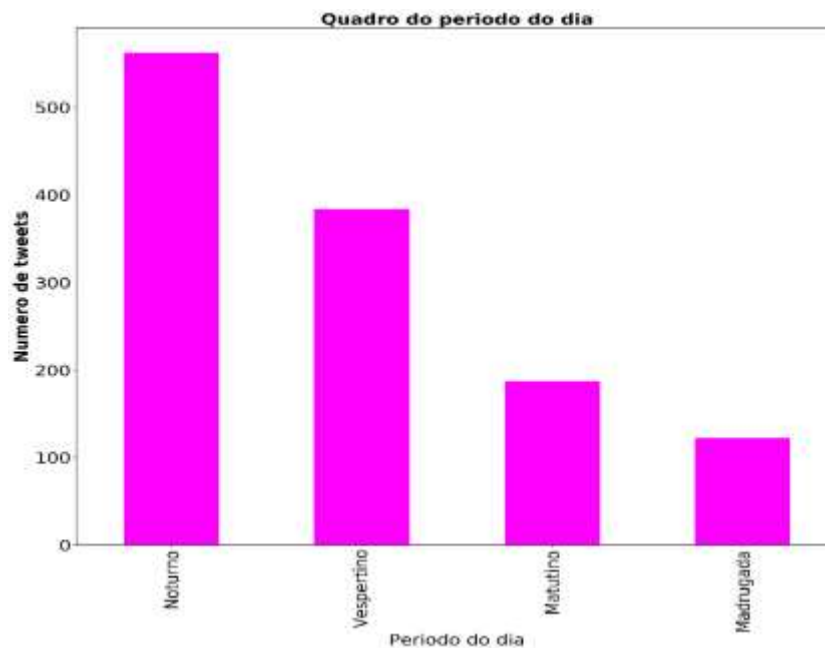


Figura 2. Períodos do dia dos crimes em Belém do Pará entre os Anos de 2017 – 2018

A partir dos dados apresentados na figura 3, elaborado de acordo com os dados coletados pela rede social Twitter, os dias da semana com ocorrências de crimes em Belém do Pará ficaram assim distribuídos: Sábado registrou 246 vezes, Domingo com 222, Sexta-feira com 200, Quarta-feira com 162, Terça-feira com 154, Quinta-feira com 141 e Segunda-feira com 131.

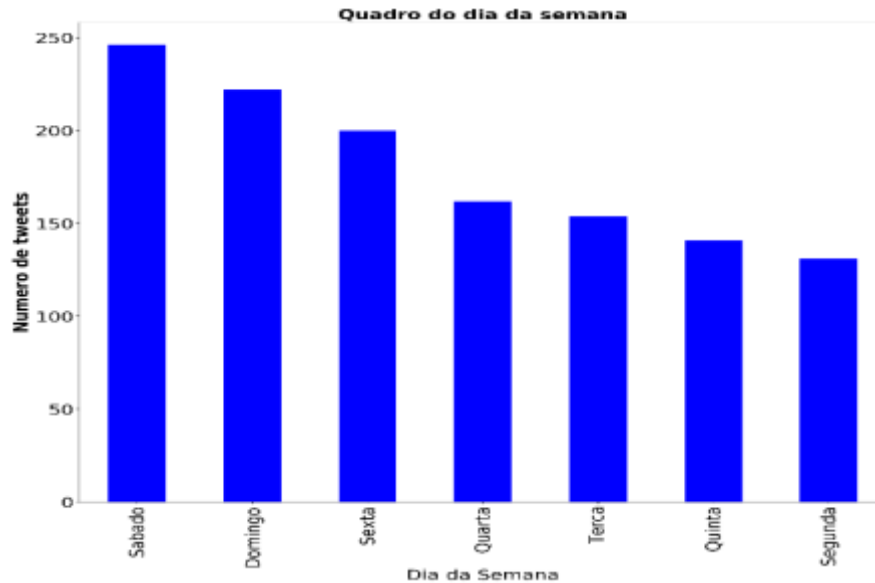


Figura 3. Dias da semana dos crimes em Belém do Pará nos últimos doze meses

Com a exposição dos dados na figura 4, pode-se analisar que o mês de Agosto foi apontado 156 vezes, Setembro e Abril, ambos com 147, Maio com 140, o mês de Junho com 135, Outubro com 132, Dezembro com 123, Novembro com 96, Julho com 89, Janeiro com 51, Fevereiro com 24 e Março com 16. Logo identificou-se que o índice de crimes em certos meses foram consideravelmente acentuados.

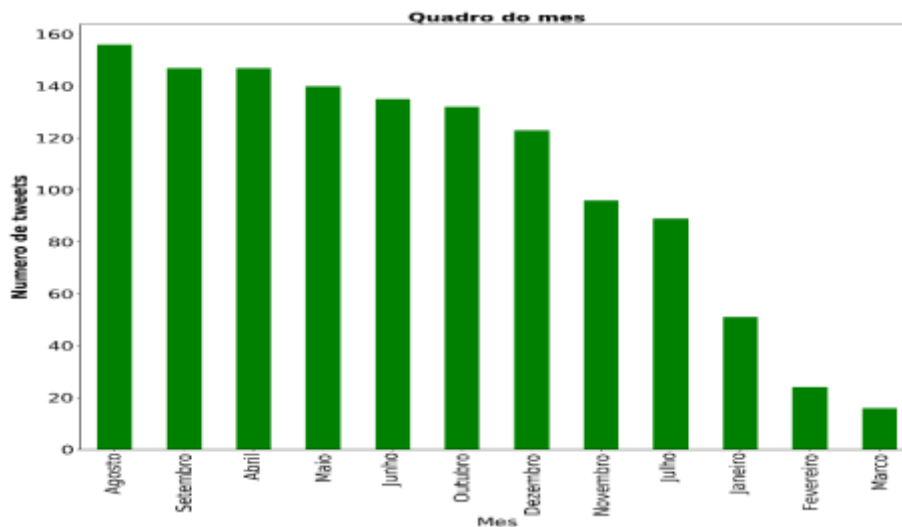


Figura 4. Meses dos crimes ocorridos em Belém do Pará nos últimos doze meses

Com a exibição dos dados na figura 5, pode-se identificar que os três bairros que tiveram mais incidências de crimes foram da Pedreira registradas 111 vezes, Marco com 93, Marambaia com 75 e assim sucessivamente como demonstra a figura abaixo.

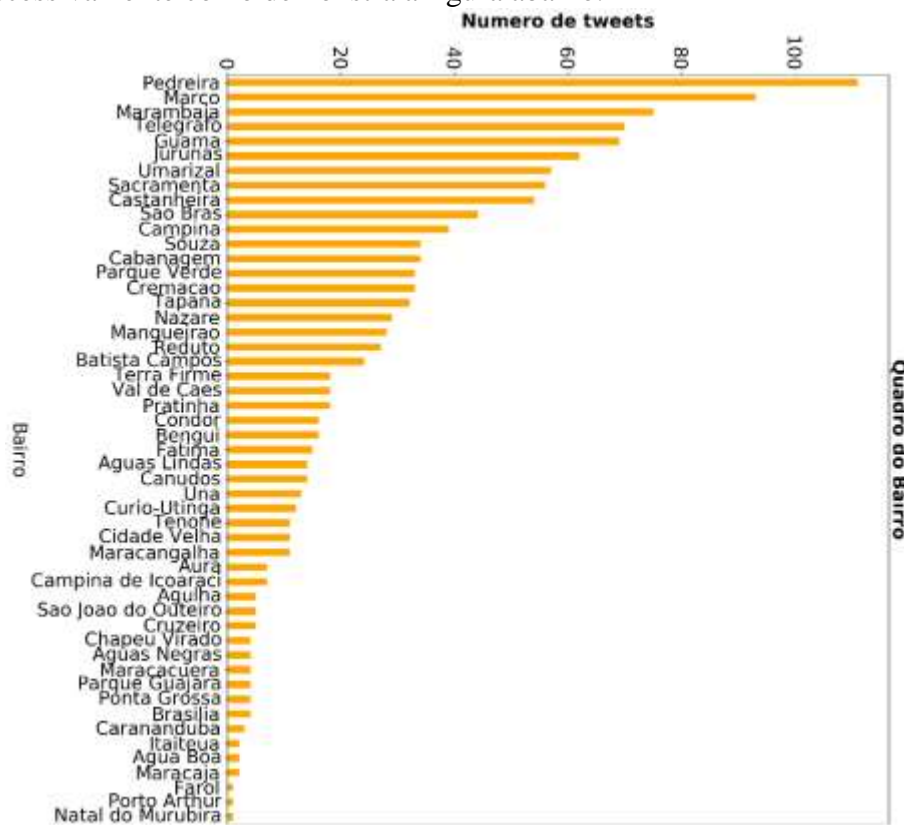


Figura 5. Bairros dos crimes ocorridos em Belém do Pará no intervalo de 2017-2018

Experimentos. Esta seção apresenta os experimentos desenvolvidos com os microdados fornecidos pelo Twitter que tem o objetivo de gerar resultados para análise e chegar nas variáveis que mais impactam na permanência ou não da criminalidade em Belém do Pará. Para os experimentos foram utilizados quatro entradas e uma saída.

Sendo assim, foi selecionado 10 Bairros com maiores incidências de crimes e extraiu-se três classes alvos – Roubo, Acidente e Homicídio – sendo 252 Roubos, 178 Acidentes e 117 Homicídios como exposto no quadro abaixo.

Entradas	Saída
Mês	Crimes ↓ Classes { - Roubo - Acidente - Homicídio
Dia_da_semana	
Período_do_dia	
Bairro	

Figura 6. Features (Variáveis)

Durante os experimentos, os dados foram submetidos a execução do algoritmo de classificação que prever resposta qualitativa. Os dados foram divididos em 80% para treino e 20% para teste mantendo a proporção de quantidades de registros em cada classe (Roubos, Acidentes e Homicídios).

Segundo Grupta (2017), ao reduzir os dados de treinamento, corre-se o risco de perder padrões ou tendências importantes para o modelo, o que faz aumentar o erro do modelo na fase de teste. Sendo assim, precisamos de um método que forneça dados amplos para treinamento do modelo e também para a validação.

Desta forma, devido a sua ampla utilização para resolver esse problema foi utilizado a técnica de validação cruzada (2008)⁴, que divide o conjunto em “n” partes iguais, onde “n-1” partes são utilizadas para treino e a parte restante é utilizada para teste, repetindo até que todas as “n” partes tenham sido utilizadas para teste. Em cada teste é calculada a acurácia a fim de obter o melhor modelo a ser utilizado com os dados de teste.

Resultados. Foi aplicado uma técnica de mineração de dados, chamado Classification And Regression Tree (CART) ou Árvore de Classificação e Regressão e gerado a matriz de confusão para o critério com os dados de teste. Foram calculadas quatro métricas, sendo elas: Acurácia (total de acertos / total de dados no conjunto), Taxa de erro (1-Acurácia), Sensibilidade (acertos positivos / total de positivos) e Especificidade (acertos negativos / total de negativos).

Diante da análise da figura 7, é possível notar que na base de teste, 43,23% é para acurácia, 56,77% é para taxa de erro, sensibilidade com 40,43% e especificidade com 71,03%.

⁴ PAYAM REFAEILZADEH, LEI TANG, HUAN LIU. “Cross-Validation”. Arizona State University. 2008. Disponível em: < <http://leitang.net/papers/ency-cross-validation.pdf> >. Acessado em: julho de 2018.

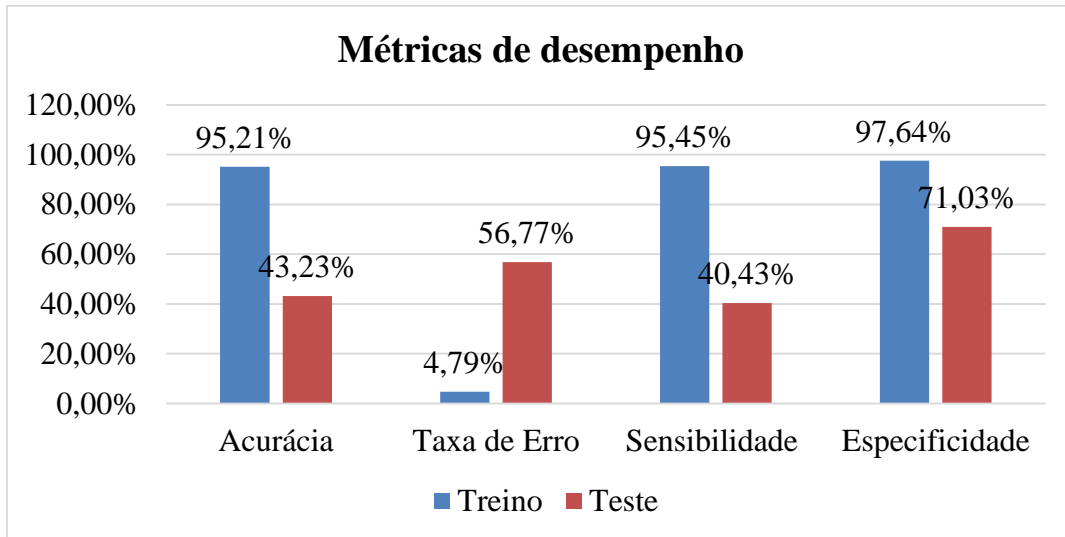


Figura 7. Métricas de desempenho

Com a exposição da árvore na figura 8, pode-se dizer que trouxe facilidade na interpretação, análise e apresentação do modelo para terceiros, assim como o entendimento de como a mesma está sendo modelada, já que entre um nó e outro existe uma tomada de decisão e cada nó da árvore denota um teste de um atributo, e os ramos denotam os possíveis valores que o atributo pode tomar.

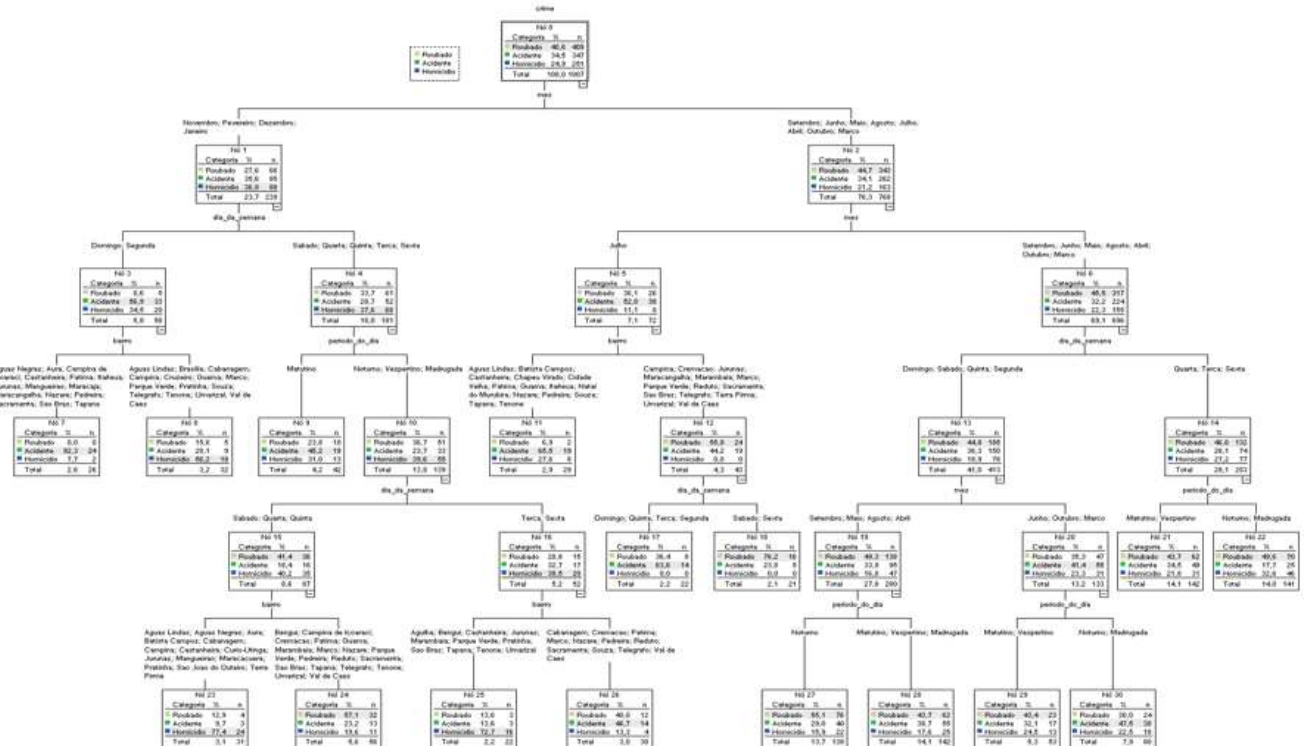


Figura 8. Árvore de Decisão – Validação Cruzada

Pois na figura 9, mostra-se algumas regras extraídas da árvore de decisão citada anteriormente.



```
/* Node 1 */.
```

```
IF (dia_da_semana = "Domingo" OR dia_da_semana = "Segunda")
```

```
THEN
```

```
Node = 1
```

```
Prediction = 2
```

```
Probability = 0.375887
```

```
/* Node 9 */.
```

```
IF (dia_da_semana != "Domingo" AND dia_da_semana != "Segunda") AND (mes != "Setembro" AND mes != "Maio" AND mes != "Julho" AND mes != "Marco") AND (((periodo_do_dia = "Matutino" OR periodo_do_dia = "Vespertino") OR (periodo_do_dia != "Noturno" AND periodo_do_dia != "Madrugada") AND ((mes = "Novembro" OR mes = "Fevereiro" OR mes = "Outubro" OR mes = "Janeiro") OR (mes != "Junho" AND mes != "Agosto" AND mes != "Dezembro" AND mes != "Abril") AND (dia_da_semana = "Terca")))) AND (mes != "Novembro" AND mes != "Fevereiro" AND mes != "Janeiro")
```

```
THEN
```

```
Node = 9
```

```
Prediction = 2
```

```
Probability = 0.309783
```

```
/* Node 10 */.
```

```
IF (dia_da_semana != "Domingo" AND dia_da_semana != "Segunda") AND (mes != "Setembro" AND mes != "Maio" AND mes != "Julho" AND mes != "Marco") AND (((periodo_do_dia = "Matutino" OR periodo_do_dia = "Vespertino") OR (periodo_do_dia != "Noturno" AND periodo_do_dia != "Madrugada") AND ((mes = "Novembro" OR mes = "Fevereiro" OR mes = "Outubro" OR mes = "Janeiro") OR (mes != "Junho" AND mes != "Agosto" AND mes != "Dezembro" AND mes != "Abril") AND (dia_da_semana = "Terca")))) AND (mes = "Novembro" OR mes = "Fevereiro" OR mes = "Janeiro")
```

```
THEN
```

```
Node = 10
```

```
Prediction = 1
```

```
Probability = 0.285714
```

```
/* Node 11 */.
```

```
IF (dia_da_semana != "Domingo" AND dia_da_semana != "Segunda") AND (mes != "Setembro" AND mes != "Maio" AND mes != "Julho" AND mes != "Marco") AND (((periodo_do_dia = "Noturno" OR periodo_do_dia = "Madrugada") OR (periodo_do_dia != "Matutino" AND periodo_do_dia != "Vespertino") AND ((mes = "Junho" OR mes = "Agosto" OR mes = "Dezembro" OR mes = "Abril") OR (mes != "Novembro" AND mes != "Fevereiro" AND mes != "Outubro" AND mes != "Janeiro") AND (dia_da_semana != "Terca")))) AND (mes = "Fevereiro" OR mes = "Dezembro" OR mes = "Outubro")
```

```
THEN
```

```
Node = 11
```

```
Prediction = 3
```

```
Probability = 0.333333
```

```
/* Node 12 */.
```

```
IF (dia_da_semana != "Domingo" AND dia_da_semana != "Segunda") AND (mes != "Setembro"
AND mes != "Maio" AND mes != "Julho" AND mes != "Marco") AND (((periodo_do_dia =
"Noturno" OR periodo_do_dia = "Madrugada") OR (periodo_do_dia != "Matutino" AND
periodo_do_dia != "Vespertino") AND ((mes = "Junho" OR mes = "Agosto" OR mes = "Dezembro"
OR mes = "Abril") OR (mes != "Novembro" AND mes != "Fevereiro" AND mes != "Outubro" AND
mes != "Janeiro") AND (dia_da_semana != "Terca")))) AND (mes != "Fevereiro" AND mes !=
"Dezembro" AND mes != "Outubro")
```

```
THEN
```

```
Node = 12
```

```
Prediction = 1
```

```
Probability = 0.415301
```

```
/* Node 7 */.
```

```
IF (dia_da_semana != "Domingo" AND dia_da_semana != "Segunda") AND (mes = "Setembro"
OR mes = "Maio" OR mes = "Julho" OR mes = "Marco") AND (((periodo_do_dia = "Matutino" OR
periodo_do_dia = "Noturno") OR (periodo_do_dia != "Vespertino" AND periodo_do_dia !=
"Madrugada") AND ((dia_da_semana = "Quarta" OR dia_da_semana = "Quinta" OR
dia_da_semana = "Sabado" OR dia_da_semana = "Sexta") OR (dia_da_semana != "Terca") AND
(mes != "Marco"))))
```

```
THEN
```

```
Node = 7
```

```
Prediction = 1
```

```
Probability = 0.459016
```

```
/* Node 8 */.
```

```
IF (dia_da_semana != "Domingo" AND dia_da_semana != "Segunda") AND (mes = "Setembro" OR
mes = "Maio" OR mes = "Julho" OR mes = "Marco") AND (((periodo_do_dia = "Vespertino" OR
periodo_do_dia = "Madrugada") OR (periodo_do_dia != "Matutino" AND periodo_do_dia !=
"Noturno") AND ((dia_da_semana = "Terca") OR (dia_da_semana != "Quarta" AND
dia_da_semana != "Quinta" AND dia_da_semana != "Sabado" AND dia_da_semana != "Sexta")
AND (mes = "Marco"))))
```

```
THEN
```

```
Node = 8
```

```
Prediction = 2
```

```
Probability = 0.385417
```

Figura 9. Regras da Árvore de Decisão

Conclusão e Trabalhos Futuros. O presente trabalho buscou analisar a violência nos bairros de Belém do Pará, através da extração, filtragem e análise dos dados do aplicativo Twitter. A partir disso, no intuito de descobrir conhecimento nos dados coletados, foi utilizado a técnica de mineração de dados.

Portanto, através dos testes e experimentos realizados com a técnica de mineração de dados, chamada CART, é possível identificar padrões escondidos nas postagens em redes sociais, facilitando a identificação das ocorrências criminosas em Belém-PA.

Como trabalhos futuros, almeja-se:

- Expandir o conjunto de dados utilizados, considerando outros tipos de crimes e de outras regiões do Pará e do Brasil;
- Criar um aplicativo para dispositivos móveis, que possibilite mostrar um panorama atualizado dos crimes que vem acontecendo na região, permitindo que o usuário evite regiões perigosas;
- Realizar um levantamento dos fatores que mais contribuem para os crimes em Belém do Pará;
- Consultar um especialista em segurança pública para realizar uma validação do conhecimento descoberto.

Divulgação. Os autores relatam não haver conflitos de interesse neste trabalho.

Referências.

- (1) Diário do Pará. (2018). “Mais de 4.400 pessoas foram assassinadas no Pará em 2017” - <https://www.diarioonline.com.br/noticias/para/noticia-478152-.html>. Acessado em: março de 2018.
- (2) Diário do Pará. (2018). “Violência aumenta quase 20% no Pará e polícia não dá conta” - <https://www.diarioonline.com.br/noticias/policia/noticia-481110-.html>. Acessado em: março de 2018.
- (3) Cerqueira, Daniel & Soares, Rodrigo (2016). “The welfare cost of homicides in brazil: accounting for heterogeneity in the willingness to pay for mortality reductions”. *Health Economics*. 25: 259–276 (2016).
- (4) G1 Pará. (2018). “No Pará, um assassinato foi registrado a cada duas horas em Janeiro de 2018” - <https://g1.globo.com/pa/para/noticia/no-para-um-assassinato-foi-registrado-a-cada-duas-horas-em-janeiro-de-2018.ghtml>. Acessado em: março de 2018.
- (5) Miranda W. “No Pará, um assassinato foi registrado a cada duas horas em Janeiro de 2018” - <https://g1.globo.com/pa/para/noticia/no-para-um-assassinato-foi-registrado-a-cada-duas-horas-em-janeiro-de-2018.ghtml>. Acessado em: janeiro de 2018.
- (6) IPEA - Instituto de Pesquisa Econômica e Aplicada. (2017). “Atlas da Violência” - http://www.ipea.gov.br/portal/images/170602_atlas_da_violencia_2017.pdf. Acessado em: março de 2018.
- (7) FBSP – Fórum Brasileiro de Segurança Pública. (2017). “Anuário Brasileiro de Segurança Pública”. São Paulo, ano 11, 2017 - http://www.forumseguranca.org.br/wp-content/uploads/2017/12/ANUARIO_11_2017.pdf . Acessado em: janeiro de 2018.
- (8) Bell J. Machine Learning: Hands-on for developers and technical professionals. [S.l.]: John Wiley & Sons, 2014.
- (9) James G, et al. An introduction to statistical learning. [S.l.]: Springer, 2013. v. 6.
- (10) Gollapudi S. Practical Machine Learning. [S.l.]: Packt Publishing Ltd, 2016.



- (11) Masanori F. “Curso de Raspagem de Dados com Python 3” - <https://github.com/fmasanori/treinamento>. Acessado em: janeiro de 2018.
- (12) Weiss GM. “Data Mining in Telecommunications. Data Mining and Knowledge Discovery Handbook”, pages 1189-1201. Springer. 2005.
- (13) Olson DL, Delen D. “Advanced Data Mining Techniques”. Springer. 2008.
- (14) Han J, Kamber M. “Data Mining: Concepts and Techniques”. Elsevier. 2006.
- (15) Gupta P. (2017). “Cross-Validation in Machine Learning” - <https://medium.com/towards-data-science/cross-validation-in-machine-learning-72924a69872f>. Acessado em: fevereiro de 2018.
- (16) Payam Refaeilzadeh, Lei Tang, Huan Liu. “Cross-Validation”. Arizona State University. 2008 - <http://leitang.net/papers/ency-cross-validation.pdf>. Acessado em: julho de 2018.