Aplicando *Random Forest* às vítimas de Crimes Violentos Intencionais atendidas pelo Corpo de Bombeiros Militar de Pernambuco

Trabalho de Conclusão de Curso

Engenharia da Computação

Gustavo Coutinho de Amorim Damasceno
Orientador: Prof. Mêuser Jorge Silva Valença

UNIVERSIDADE DE PERNAMBUCO

Universidade de Pernambuco
Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação

Gustavo Coutinho de Amorim Damasceno

Aplicando *Random Forest* às vítimas de Crimes Violentos Intencionais atendidas pelo Corpo de Bombeiros Militar de Pernambuco

Artigo apresentado como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, Junho de 2019.

ESCOLA POLITÉCNICA
DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 02/07/2019, às 8h, reuniu-se para deliberar sobre a defesa da monografia de conclusão de curso do(a) discente **GUSTAVO COUTINHO DE AMORIM DAMASCENO**, orientado(a) pelo(a) professor(a) **MÊUSER JORGE SILVA VALENÇA**, sob título Aplicando Random Forest as vítimas de Crimes Violentos Intencionais atendidas pelo Corpo de Bombeiros Militar de Pernambuco, a banca composta pelos professores:

SERGIO MARIO LINS GALDINO (PRESIDENTE)
MÊUSER JORGE SILVA VALENÇA (ORIENTADOR)

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

☒ Aprovada ☐ Aprovada com Restrições* ☐ Reprovada

e foi-lhe atribuída nota: 10,0 (dez)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O(A) discente terá 7 dias para entrega da versão final da monografia a contar da data deste documento.

AVALIADOR 1: Prof (a) **SERGIO MARIO LINS GALDINO**

AVALIADOR 2: Prof (a) **MÊUSER JORGE SILVA VALENÇA**

AVALIADOR 3: Prof (a)

* Este documento deverá ser encadernado juntamente com a monografia em versão final.

ESCOLA POLITÉCNICA
DE PERNAMBUCO

Autorização de publicação de PFC

Eu, **Gustavo Coutinho de Amorim Damasceno** autor(a) do projeto de final de curso intitulado: **Aplicando Random Forest as vítimas de Crimes Violentos Intencionais atendidas pelo Corpo de Bombeiros Militar de Pernambuco;** autorizo a publicação de seu conteúdo na internet nos portais da Escola Politécnica de Pernambuco e Universidade de Pernambuco.

O conteúdo do projeto de final de curso é de responsabilidade do autor.

Gustavo Coutinho de Amorim Damasceno

Orientador(a): **Mêuser Jorge Silva Valença**

Coorientador(a):

Prof, de TCC: **Daniel Augusto Ribeiro Chaves**

Data: 02/07/2019

Aplicando *Random Forest* às vítimas de Crimes Violentos Intencionais atendidas pelo Corpo de Bombeiros Militar de Pernambuco

Mêuser Jorge Silva Valença*

Gustavo Coutinho de Amorim Damasceno**

RESUMO

O artigo vem aplicar aprendizado de máquina, com o intuito de extrair padrões e identificar fatores importantes, como resultante da aplicação de algoritmos nos dados registrados nos atendimentos às vítimas de crime violento intencional pelo Grupamento de Bombeiros de Atendimento Pré-Hospitalar do Corpo de Bombeiros Militar de Pernambuco, no período de janeiro de 2012 a março de 2018. O objetivo desse estudo é construir um modelo de classificação, utilizando o algoritmo *Random Forest*, identificando as vítimas como um "destaque" ou "não destaque" em função das características registradas em cada ocorrência atendida e que proporciona a chegada das vítimas com vida ao hospital pelas equipes dos Bombeiros. Os resultados apontaram que o modelo construído se mostrou bastante eficaz na classificação das variáveis presentes nos dados analisados, porém se faz necessário aprofundar o estudo, levando em consideração outros aspectos importantes na dinâmica do atendimento ao vitimado, como os registros quantitativos de dados da própria definição do termo "destaque", relacionados aos sinais vitais (frequências respiratória e cardíaca, pulso, pressão arterial, temperatura, saturação de oxigênio) e avaliação do nível de consciência e responsividade, através da Escala de Coma de Glasgow registradas nas fichas de ocorrência no ato do atendimento das equipes do Corpo de Bombeiros.

Palavras-chaves: *Random Forest*, Aprendizado de Máquina, Atendimento Pré-Hospitalar, Crime Violento Intencional.

* Escola Politécnica de Pernambuco – Universidade de Pernambuco – UPE – Professor do curso de Engenharia da computação da Escola Politécnica de Pernambuco. meuser@ecomp.poli.br

** Escola Politécnica de Pernambuco – Universidade de Pernambuco – UPE – Aluno do curso de Engenharia da computação da Escola Politécnica de Pernambuco. gcad@ecomp.poli.br

ABSTRACT

The article comes to apply machine learning, with the purpose of extracting patterns and identifying important factors, as a result of the application of algorithms in the data recorded in the care of the victims of intentional violent crime by the Predental Fire Brigade Group of the Military Fire Brigade of Pernambuco from January 2012 to March 2018. The objective of this study is to construct a classification model, using the Random Forest algorithm, identifying the victims as a "highlight" or "not highlighted" in function of the characteristics registered in each occurrence attended and that provides the arrival of the victims alive to the hospital by the Fire Brigade teams. The results showed that the constructed model proved to be quite effective in the classification of the variables present in the data analyzed, but it is necessary to deepen the study, taking into account other important aspects in the dynamics of victim care, such as the quantitative data records of the definition itself (respiratory and heart rates, pulse, blood pressure, temperature, oxygen saturation) and assessment of level of consciousness and responsiveness, through the Glasgow Coma Scale recorded in the event reports of the attendance of the teams of the Fire Department.

Keywords: Random Forest, Machine learning, Pre-Hospital Care, Intentional Violent Crime.

1 INTRODUÇÃO

Segundo o IBGE, em 2016 o Brasil, tinha uma população estimada em 206 Milhões de pessoas [1]. O Atlas da violência 2018 [2], indica que nosso país conseguiu chegar no índice de 62.517 homicídios neste mesmo ano. Esses números representa uma taxa de 30,3 mortes para cada 100 mil habitantes que, ao comparar com a taxa europeia, corresponde a 30 vezes mais, no mesmo período. Nos últimos dez anos, 553 mil pessoas perderam suas vidas devido à violência intencional no Brasil.

Ainda com referência aos dados do Atlas da violência 2018 [2] Pernambuco apresentou uma taxa média de 47,3 homicídios por 100.000 habitantes em 2016, tendo uma população de aproximadamente de 8,5 milhões de pessoas neste mesmo ano [3], sendo o núcleo de densidade populacional do estado a chamada Região Metropolitana do Recife – RMR, conjunto de 15 municípios que circundam a capital compondo, ao todo, uma população com mais de 3,7 milhões de pessoas [4], o que representa, percentualmente, mais de 45% de toda a população de Pernambuco, vivendo em um território que não chega a 3% da área do Estado.

Portanto, convivendo com os altos índices de violência apresentados historicamente, o governo do estado de Pernambuco, em 2007, inicia o Programa Pacto Pela Vida (PPV) [5], com o intuito de diminuir a criminalidade e controlar a Violência. O programa veio com o foco principal na redução dos crimes contra a vida, regulada pela meta de diminuição em 12% ao ano na taxa de Crimes Violentos Letais Intencionais - CVLI) para todo o estado de Pernambuco e no intenso debate com a sociedade civil. A meta na redução da taxa dos CVLIs converteu-se em um elemento regulador da gestão do Programa [6].

Diante do exposto, o Corpo de Bombeiros Militar de Pernambuco (CBMPE), concebe o projeto Resgate de Vidas, executado pelo Grupamento de Bombeiros de Atendimento Pré-Hospitalar (GBAPH) [8] objetivando ampliar sua capacidade de atendimento às vítimas de crimes violentos intencionais (CVI) registradas como "destaque" [7], possibilitando que estas vítimas cheguem ao sistema de atendimento médico de emergência ainda com vida, decorrente do tempo e atendimento, evitando a morte do vitimado.

Neste trabalho, nosso interesse se concentra em criar um modelo de aprendizado de máquina considerando a capacidade de classificação que as *Random Forest* possuem [10], identificando as vítimas como um "destaque" ou "não destaque", através dos atributos independentes registrados no ato dos atendimentos às vítimas de CVI, pelas equipes de bombeiros do GBAPH, analisando sua acurácia e levando em consideração a simplicidade de entendimento do modelo de classificação.

O trabalho está organizado da seguinte forma. Na seção 2 apresentamos a fundamentação teórica, sendo na seção 2.1 falaremos brevemente sobre o projeto Resgate de Vidas; na seção 2.2 teremos as definições de *Random Forest* e seus métodos básicos de construção. A seção 3 apresenta a metodologia utilizada nos dados para serem usados nos experimentos. Os resultados e suas análises são descritos na seção 4. Na seção 5, são apresentadas as considerações finais.

2 Fundamentação Teórica

2.1. Projeto Resgate de Vidas

O Projeto Resgate de Vidas foi idealizado pelo CBMPE em 2007, em estudo realizado pelo Comando Geral à época, através da análise dos números levantados junto às emergências dos hospitais de referência da RMR do Recife, no qual uma grande quantidade dos atendimentos realizados, seria decorrente de CVI [8].

O objetivo principal visava o atendimento prioritário as vítimas decorrentes de CVI, de modo a garantir que estas pessoas pudessem chegar ao sistema de atendimento médico de emergência, quer fosse em hospital ou em sistemas de atendimento pré-hospitalar fixo, em intervalo de tempo e condição clínica suficientes para garantir a sobrevida e, consequentemente, evitar a letalidade contribuindo para a obtenção das metas estipuladas pelo governo do Estado, sendo estas equivalentes à redução de 12% do total de CVLIs [8].

A execução do projeto se deu, conforme explica Alves [8], de forma bastante sucinta:

> Em 22 de agosto de 2010 o projeto foi ativado com 459 (quatro centos e cinquenta e nove) bombeiros capacitados, 26 (vinte e seis) bases operacionais funcionando 10 (dez) em estruturas partilhadas com Policia Militar, 3 (três) em delegacias da Policia Civil e 2 (duas) com apoio de prefeituras da RMR [8].

Dentro das metas estabelecidas para o CBMPE, na aferição de resultados realizada semanalmente no âmbito do comitê gestor do Programa Pacto pela Vida, são especialmente destacados e contabilizados os atendimentos intitulados "destaque" [7].

Para a definição do termo "destaque" Correa [7] descreve:

> Temos o pressuposto, para esta definição, que os vitimados estejam em iminente risco de morte, sendo, para tanto, catalogados e registrados dados como tipo de agressão (física, perfuração por arma branca – PAB - ou perfuração por arma de fogo – PAF), características dos ferimentos, local da lesão (crânio, face, pescoço, ombro, tórax, abdome, membros superiores e inferiores), quantificação dos sinais vitais (frequências respiratória e cardíaca, pulso, pressão arterial, temperatura, saturação de oxigênio), estado clínico (sinais indicativos de choque e avaliação do nível de consciência e responsividade, através da Escala de Coma de Glasgow) e demais características que indiquem a gravidade da lesão.

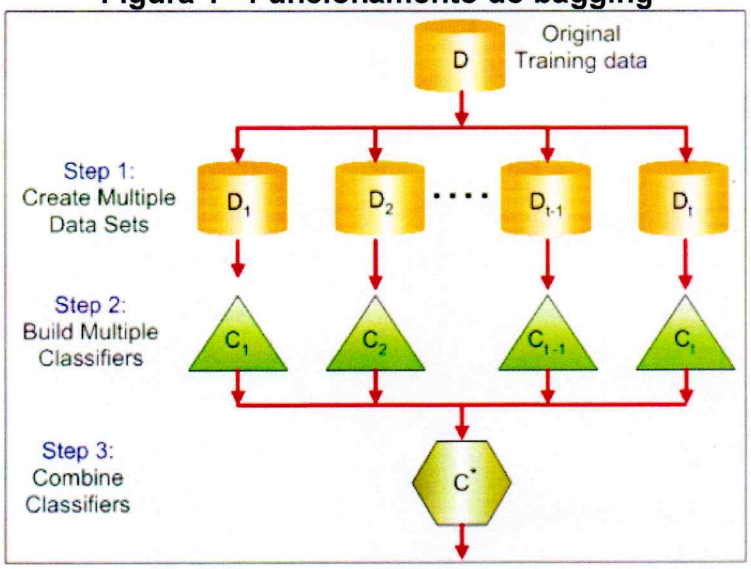
As informações de registro das ocorrências devem constar nos boletins de ocorrência específicos, que são preenchidos pelos comandantes das guarnições de resgate, para aquelas que têm a agressão como natureza. A divisão de operações do GBAPH, transcreve os registros dos boletins para arquivos em planilhas, para definir se determinada ocorrência será classificada como um "destaque" ou não. Posteriormente, as informações são computadas em planilhas específicas, que são encaminhadas diária e semanalmente ao Comitê Gestor do Pacto pela Vida, que irá analisar a descrição dos atendimentos e contabilizá-los [9].

2.2. Random Forest

As *Random Forests* (RFs) são obtidas através do processo chamado de *bootstrapping aggregating (bagging)* [10]. Um método que combina o resultado de vários classificadores, modelados em diferentes reamostras do mesmo conjunto de dados.

A figura 1 traz um exemplo para ilustrar a estrutura de uma *bagging*:

Figura 1 - Funcionamento do bagging



Fonte: SITE ANALYTICS VIDHYA CONTENT TEAM, 2016. Disponível em:< https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/>

Considerando que o nosso interesse se dá nos métodos de seleção de características RFs propostas por Breinman [10]. A classificação final será dada por meio de um processo de votação entre as árvores que compõem a floresta, onde a característica mais votada será a escolhida.

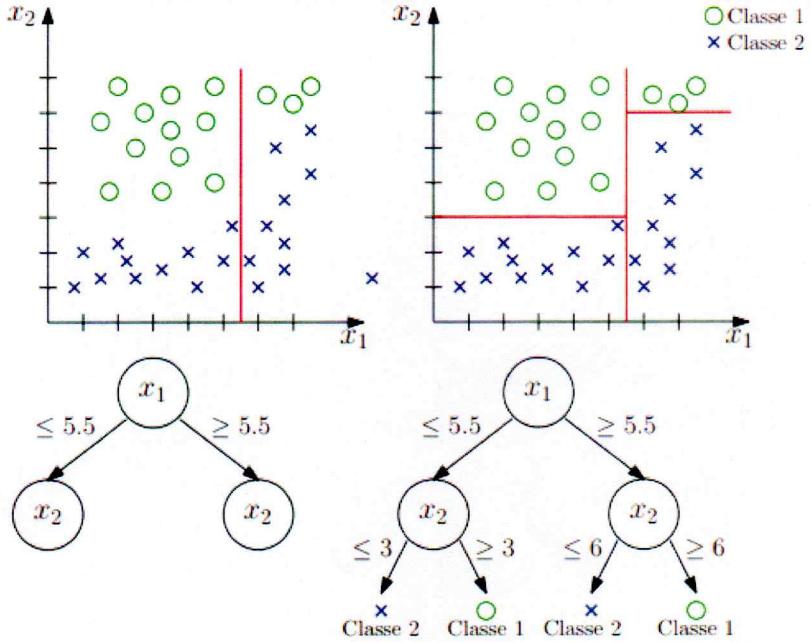
2.3 Árvores de Decisão

Para entendermos como as RFs funcionam, precisamos entender a Árvore de Decisão, que funciona tanto para as variáveis de entrada categóricas (quando os atributos são variáveis observáveis e independentes e que possuem um conjunto de valores finitos) como para as contínuas (os atributos observáveis e independentes e assumem valores numéricos em intervalos no eixo dos números reais) e sua saída, uma variável dependente e determinada em função dos atributos de entrada.

Árvores de decisão definem um fluxo organizado de representar o conhecimento. Elas funcionam com sucessivas divisões da população utilizada, em vários outros subconjuntos, até que cada um destes, pertença a uma mesma classe, não havendo necessidade de novas divisões.

Os resultados dos vários subconjuntos obtidos com a construção de uma árvore de decisão são dados organizados de maneira compacta, utilizados para classificar novos exemplos. A figura 2 exemplifica esse conceito.

Figura 2-Exemplo de árvore de decisão com dois parâmetros



Fonte: Marins, Matheus Araújo, Classificação De Falhas Em Máquinas Rotativas Utilizando Métodos De Similaridade E *Random Forest*, p25.

O algoritmo a seguir mostra um algoritmo genérico para construir árvores de decisão, em que S representa o conjunto de exemplos aplicado à árvore, sendo que inicialmente S contém todos os exemplos de treino.

ALGORITMO 1 – Algoritmo genérico para construção de árvores de decisão

---

I. Se todos os exemplos no atual conjunto de exemplos S satisfazem um critério de parada;

Então

II. cria um nó folha com algum nome da classe e para;

senão

III. seleciona um atributo A para ser utilizado como um atributo de particionamento e cria um nó com o nome do atributo de particionamento;

IV. escolhe um teste sobre os valores de A, com resultados mutualmente exclusivos e coletivamente exaustivos $R_1, \cdots, R_k$, e cria um ramo, a partir do nó recentemente criado, para cada teste;

V. particiona S nos subconjuntos $S_1, \cdots, S_k$, tal que cada S_i, $i = 1, \cdots, k$, contenha todos os exemplos em S com resultado R_i do teste escolhido;

VI. aplica este algoritmo recursivamente para cada subconjunto S_i, $i = 1, \cdots, k$;

fim_senão

fim_se

Como forma de melhor explicar o funcionamento do algoritmo 1, cada passo do algoritmo é identificado por um número sequencial entre parênteses à sua esquerda.

Os passos (I) e (III) são os passos mais relevantes do algoritmo, por requererem acesso aos dados para avaliar uma regra candidata. Estes são os passos que consomem mais tempo do algoritmo quando a descoberta de conhecimento se dá em grandes bases de dados.

O passo (I) consiste em decidir o momento de parar o particionamento recursivo. O processo é parado se todos os exemplos no nó atual possuem a mesma classe. No entanto, se esta condição não for verdadeira, pode ser interessante parar o particionamento no atual nó, para evitar que a árvore se expanda muito.

No passo (II), se todos os exemplos no recém-criado nó folha, têm a mesma classe, o algoritmo dá ao nó o nome da classe. Contudo, o algoritmo pode dar ao nó folha o nome da classe mais frequente ocorrida neste nó.

No passo (III) é computado o melhor atributo candidato à partição, avaliando como efetivamente, os valores do atributo candidato discriminam as classes dos exemplos.

No passo (IV) é criado um arco para cada valor distinto, resultante do particionamento do atributo selecionado no passo (III).

O passo (V) consiste em designar cada exemplo a um dos arcos criados, de acordo com o valor do atributo de particionamento.

Finalmente, no passo (VI), o algoritmo é aplicado recursivamente para cada subconjunto do conjunto de exemplos S.

2.4 **Critérios de seleção de atributos**

Os critérios para a seleção de atributos são utilizados para determinar qual atributo pertencente ao conjunto de exemplos será utilizado em cada nó da árvore e existem diversos tipos de critérios para a seleção, durante a construção da árvore de decisão. Deste modo avalia-se a melhor partição a ser realizada, de acordo com a capacidade informativa do atributo [18].

2.5 **Ganho de Informação**

Um dos mais antigos e conhecidos critérios de seleção de atributos é o ganho de informação (*information gain*), utilizado pelos também conhecidos algoritmos de indução para árvores de decisão ID3 [17], C4.5 [16] e CART [19].

O ganho de informação tem, como base geradora, uma medida conhecida como entropia [11]. O ganho pode ser conceituado como a redução esperada da entropia, tendo, como função, a seleção de atributos utilizados no particionamento de um conjunto de dados.

A entropia empregada para a obtenção do ganho de informação tem sua origem na teoria da informação e baseia-se no trabalho realizado por Claude Shannon e Warren Weaver, em 1949 [12] entropia usa, como estratégia, a redução da impuridade, ou seja, mede a quantidade de informação necessária para codificar uma situação encontrada em um nó [13]. A impuridade é máxima se todas as classes de um nó têm igual prioridade e mínima quando existe apenas uma classe. Na teoria da informação, a informação é medida em *bits*.

A entropia é dada pela equação (1), que determina o número de exemplos de S pertencentes à classe C_j, podendo o atributo ter m possíveis valores:

$$Entropia(S) = \sum_{j=1}^{m} -p_j \log_2 p_j \qquad (1)$$

Onde: S é o conjunto de exemplos;

m é o número de classes;

p_j é a proporção de S pertencer à classe j, tendo então a equação (2):

$$P_j = \frac{|S_j|}{|S|} \qquad (2)$$

Onde: $|S_j|$ é o número de exemplos classificados na j-ésima partição;

$|S|$ é o número total de exemplos do conjunto S

Com base na medida da entropia, uma classificação é considerada perfeita, se todos os membros de um conjunto S pertencem a uma mesma classe, sendo a entropia igual a zero. Por exemplo, se todos os membros são positivos, p+ = 1, então p- = 0 (zero). Quando a entropia é igual a um, é dito que os membros de um conjunto foram classificados ao acaso, pois o conjunto possui número igual de exemplos positivos e negativos. Se o conjunto contiver números diferentes de exemplos positivos e negativos, a entropia estará entre 0 (zero) e 1 (um).

2.6 **Gini**

O Gini é um critério utilizado na seleção de atributos, sendo aplicável a árvores m-árias, cujo objetivo é a minimização da impuridade [14].

Para um dado conjunto de dados S contendo exemplos de m classes, $Gini(S)$ é definido pela equação (3):

$$Gini(S) = 1 - \sum_{j=1}^{m} p_j^2 \qquad (3)$$

Onde: p_j é a frequência relativa da classe j em S

Se dividirmos S em dois subconjuntos, S_1 e S_2, com n_1 e n_2 exemplos respectivamente, o novo índice de divisão dos dados, $Gini_{split}(S)$, é dado pela equação (4):

$$Gini_{split}(S) = \frac{n_1}{n} Gini(S_1) + \frac{n_2}{n} Gini(S_2) \qquad (4)$$

2.7 Algoritmo de indução de Árvore de Decisão

Esta seção irá se ater apenas ao algoritmo CART, mesmo sabendo da existência de outros, o nosso trabalho o utiliza para a construção do modelo de *Random Forest*, portanto vamos descrever alguns pontos fundamentais, como: sua origem, características e as metodologias aplicadas à construção de árvores de decisão.

2.8 CART

O algoritmo CART (*Classification and Regression Trees*) foi desenvolvido por Leo Breiman, Jerone Friedman, Richard Oslen e Charles Stone e publicado em 1984.

Tem como principal característica a capacidade de pesquisa de relação entre os dados, mesmo quando elas não são evidentes e com a produção de árvores de grande simplicidade e legibilidade [20].

O método que o algoritmo CART utiliza para a construção de árvores de decisão e possibilita sua utilização tanto para atributos previsores categóricos ou quantitativos. No particionamento baseado em atributos categóricos são testadas todas as possibilidades de formação de dois subconjuntos com os possíveis valores.

Sua árvore grada é baseada na técnica recursiva de divisão binária. O processo é binário porque cada nodo é separado sempre em exatamente dois subconjuntos e à medida que se percorre a árvore, da raiz às folhas são respondidas questões simples do tipo sim/não.

A recursividade se dá a cada subconjunto gerado, até que não seja possível ou não seja necessário mais efetuar partições na árvore.

A eleição da melhor característica é efetuada geralmente com base num dos dois critérios: Gini ou Entropia, e a atribuição de uma classe a cada folha é realizada com base no critério da classe mais provável ou da minimização dos custos.

2.9 Definindo uma *Random Forest*

Neste trabalho, estamos interessados nos métodos de classificação de atributos categóricos, baseados no algoritmo *Random Forest* (RF), proposto por Breiman [10]. Portanto a RF funcionará como um classificador formado por uma coleção de árvores de decisão, cada qual construída a partir de uma reamostra aleatória do conjunto de treinamento original. A RF utiliza o voto de cada árvore gerada no seu modelo, classificando a mais votada entre todas.

As RF suprem uma lacuna importante em relação às árvores isoladas. É sabido que algoritmos de construção de árvores de classificação são instáveis em relação ao conjunto de treinamento, no sentido de que perturbações nos atributos de entrada ou a inclusão de novos exemplos podem resultar em árvores consideravelmente diferentes, com diferentes erros classificação [15].

Por outro lado, RF não possuem uma interpretação direta como ocorre com as árvores de decisão individuais, pela dificuldade prática de se analisar e comparar as centenas de árvores constituintes da floresta. Outro aspecto é que, para garantir uma baixa correlação entre as classificações das árvores individuais (condição necessária para a obtenção de florestas com baixo erro de generalização), é adotada uma seleção aleatória dos atributos candidatos para partição de cada nó. Isso pode implicar na eventual seleção de atributos com baixo poder preditivo.

3 METODOLOGIA

3.1 Base de dados

O trabalho proposto foi realizado com os dados disponibilizados pela Divisão de Operações do GBAPH, juntamente com os dados do Centro Integrado de Defesa Social (CIODS), relativos ao período de janeiro de 2012 a março de 2018. A partir desses dados foi criado uma base com 4462 observações de registro de ocorrências especificamente de vítimas de CVI sendo computadas as seguintes informações:

- CIDADE (Município da RMR onde a ocorrência foi registrada);

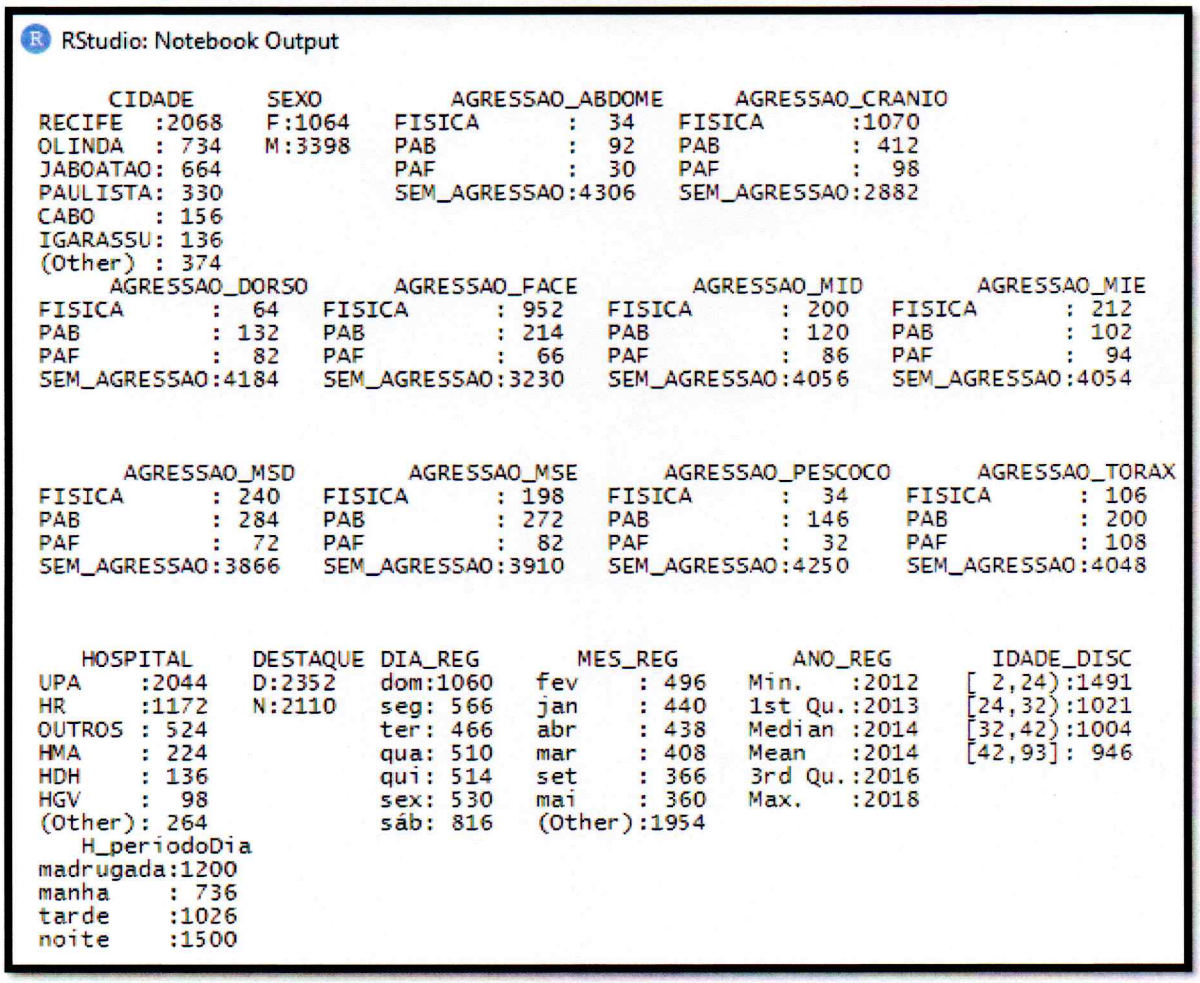
- SEXO (masculino - M, feminino - F);
- AGRESSAO_ABDOME (local da agressão no abdome, categorizado com o tipo da agressão deferida [FISICA, PAF, PAB, SEM_AGRESSAO]);
- AGRESSAO_CRANIO (local da agressão no crânio, categorizado com o tipo da agressão deferida [FISICA, PAF, PAB, SEM_AGRESSAO]);
- AGRESSAO_ FACE (local da agressão na face, categorizado com o tipo da agressão deferida [FISICA, PAF, PAB, SEM_AGRESSAO]);
- AGRESSAO_ DORSO (local da agressão no dorso, categorizado com o tipo da agressão deferida [FISICA, PAF, PAB, SEM_AGRESSAO]);
- AGRESSAO_ MID (local da agressão no membro inferior direito, categorizado com o tipo da agressão deferida [FISICA, PAF, PAB, SEM_AGRESSAO]);
- AGRESSAO_ MIE (local da agressão no membro inferior esquerdo, categorizado com o tipo da agressão deferida [FISICA, PAF, PAB, SEM_AGRESSAO]);
- AGRESSAO_ MSD (local da agressão no membro superior direito, categorizado com o tipo da agressão deferida [FISICA, PAF, PAB, SEM_AGRESSAO]);
- AGRESSAO_ MSE (local da agressão no membro superior esquerdo, categorizado com o tipo da agressão deferida [FISICA, PAF, PAB, SEM_AGRESSAO]);
- HOSPITAL (unidade de saúde para o qual a vítima foi levada);
- DIA_REG (o dia da semana do registro);
- MES_REG (o mês do registro);
- ANO_REG (o ano do registro);
- IDADE_DISC (idade das vítimas discretizadas em 04 faixas: [02 a 24 anos], [25 a 32 anos], [33 a 42 anos] e [43 a 93 anos]);
- H_periodoDia (período do dia discretizado como: manhã [06:00:00 ás 11:59:59], tarde [12:00:00 ás 17:59:59], noite [18:00:00 ás 23:59:59] e madrugada [00:00:00 as 05:59:59]);
- A classe alvo designada DESTAQUE que classifica a vítima como "S" ou "N", pela Divisão de Operações do GBAPH de forma manual, após análise das fichas de ocorrências geradas nos atendimentos.

3.2 Extração de informações

Foi utilizado, para a análise dos dados, a linguagem R através da ferramenta RSTUDIO, versão 1.2.1335, na plataforma Windows 10.

Considerando a necessidade de iniciar com uma visualização geral das informações de cada variável, extraímos um sumário, conforme mostra a figura 3 a seguir:

Figura 3-Sumario das variaveis disponiveis na base de dados



RStudio: Notebook Output

```
     CIDADE      SEXO          AGRESSAO_ABDOME     AGRESSAO_CRANIO
RECIFE  :2068   F:1064   FISICA       :  34   FISICA       :1070
OLINDA  : 734   M:3398   PAB          :  92   PAB          : 412
JABOATAO: 664            PAF          :  30   PAF          :  98
PAULISTA: 330            SEM_AGRESSAO:4306   SEM_AGRESSAO:2882
CABO    : 156
IGARASSU: 136
(Other) : 374
     AGRESSAO_DORSO      AGRESSAO_FACE        AGRESSAO_MID        AGRESSAO_MIE
FISICA       :  64   FISICA       : 952   FISICA       : 200   FISICA       : 212
PAB          : 132   PAB          : 214   PAB          : 120   PAB          : 102
PAF          :  82   PAF          :  66   PAF          :  86   PAF          :  94
SEM_AGRESSAO:4184   SEM_AGRESSAO:3230   SEM_AGRESSAO:4056   SEM_AGRESSAO:4054


      AGRESSAO_MSD        AGRESSAO_MSE      AGRESSAO_PESCOCO      AGRESSAO_TORAX
FISICA       : 240   FISICA       : 198   FISICA       :  34   FISICA       : 106
PAB          : 284   PAB          : 272   PAB          : 146   PAB          : 200
PAF          :  72   PAF          :  82   PAF          :  32   PAF          : 108
SEM_AGRESSAO:3866   SEM_AGRESSAO:3910   SEM_AGRESSAO:4250   SEM_AGRESSAO:4048


   HOSPITAL     DESTAQUE DIA_REG       MES_REG         ANO_REG        IDADE_DISC
UPA     :2044   D:2352   dom:1060   fev     : 496   Min.   :2012   [ 2,24):1491
HR      :1172   N:2110   seg: 566   jan     : 440   1st Qu.:2013   [24,32):1021
OUTROS  : 524            ter: 466   abr     : 438   Median :2014   [32,42):1004
HMA     : 224            qua: 510   mar     : 408   Mean   :2014   [42,93]: 946
HDH     : 136            qui: 514   set     : 366   3rd Qu.:2016
HGV     :  98            sex: 530   mai     : 360   Max.   :2018
(Other): 264             sáb: 816   (Other):1954
   H_periodoDia
madrugada:1200
manha    : 736
tarde    :1026
noite    :1500
```

Fonte: Produzido pelo Autor

Considerando o sumário gerado da figura 3, lembrando que estes representam dados coletados e analisados no período de janeiro de 2012 a março de 2018, podemos identificar alguns padrões importantes, que passamos a descrever:

I. Em relação ao atributo CIDADE, observamos que 46,3% de todas as ocorrências registradas como a natureza Agressão, aconteceram na cidade do Recife;

II. Em relação ao atributo SEXO, observamos que 76,1% das vítimas são do sexo masculino;

III. Considerando a agressão do tipo FISICA, observamos que a região do crânio foi a que teve maior registro com 1070 valores e a região com menor número foram a do abdome juntamente com o pescoço com 34 valores cada;

IV. Considerando a agressão do tipo arma de branca - PAB, observamos que novamente a região do crânio foi a que teve maior registro com 412 valores e a região com menor registro foi novamente a do abdome com 92 valores;

V. Considerando a agressão do tipo arma de fogo - PAF, observamos que a região do tórax foi a que teve maior registro com 108 valores e a região com menor registro foi novamente o abdome com 30 valores;

VI. Considerando a data/hora/período do registro da ocorrência, observamos que o dia de domingo, foi o dia da semana com maior números de registros; que o mês de fevereiro foi também o de maior número;

VII. Em relação aos hospitais para o qual as vítimas foram conduzidas, a variável UPA, contempla todas as UPAs da região metropolitana do Recife, portanto sendo um registro de várias unidades separadas para o qual as vítimas tiveram seu atendimento, consequentemente o Hospital da Restauração vem como a unidade hospitalar com o maior número de atendimentos com 26,3% dos atendimentos realizados;

3.3 Aplicando o modelo Random Forest

Para a aplicação do modelo, utilizamos o pacote randomForest do R para executar a avaliação do modelo.

O Próximo passo foi criar a uma base de teste com 70% das amostras e outra base de validação com 30% restantes.

Com as bases de teste e de validação criada, criamos um primeiro modelo com os parâmetros padrão, conforme mostra a figura 4:

Figura 4 -criação do primero modelo random forest com parametros padrão

```
> # Create a Random Forest model with default parameters
> modelo_rf <- randomForest(DESTAQUE ~ ., data = TrainSet, importance = TRUE, do.trace=100)
ntree      OOB      1      2
  100:   8.17%  9.29%  6.93%
  200:   7.17%  8.74%  5.45%
  300:   7.49%  9.29%  5.51%
  400:   7.36%  9.17%  5.38%
  500:   7.43%  9.05%  5.65%
> table(predict(modelo_rf), TrainSet$DESTAQUE)

      D    N
  D 1488   84
  N  148 1403
> print(modelo_rf)

Call:
 randomForest(formula = DESTAQUE ~ ., data = TrainSet, importance = TRUE,       do.trace = 100)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 7.43%
Confusion matrix:
     D    N class.error
D 1488  148  0.09046455
N   84 1403  0.05648958
> round(importance(modelo_rf), 2)
                      D      N MeanDecreaseAccuracy MeanDecreaseGini
CIDADE            71.57  55.99                80.58           121.15
SEXO              37.16  33.29                40.92            33.57
AGRESSAO_ABDOME   35.47  32.49                41.05            25.35
AGRESSAO_CRANIO   70.92  65.97                81.47            86.78
AGRESSAO_DORSO    29.36  33.55                39.24            31.16
AGRESSAO_FACE     45.34  37.46                53.24            50.34
AGRESSAO_MID      24.35  20.70                28.43            19.97
AGRESSAO_MIE      23.22  24.98                32.30            23.04
AGRESSAO_MSD      30.82  28.17                37.05            28.14
AGRESSAO_MSE      25.70  28.03                36.33            24.63
AGRESSAO_PESCOCO  35.97  28.42                40.51            25.30
AGRESSAO_TORAX    49.57  47.80                58.82            54.35
HOSPITAL         131.80 126.38               147.09           430.00
DIA_REG           54.10  51.57                69.49            94.87
MES_REG           63.76  56.23                76.48           121.79
ANO_REG           72.91  66.19                80.34           120.28
IDADE_DISC        52.03  49.06                67.32            80.73
H_periodoDia      48.17  47.78                61.49            74.34
```

Fonte: Produzido pelo Autor

Analisando as informações da figura 4, identificamos que a taxa de erro foi pequena, cerca de 7,43% utilizando 500 árvores e 4 variáveis em cada divisão, classificando erroneamente 148 vítimas que foram intituladas destaques, como não destaque e 84 vítimas que não foram destaques como destaque.

Observamos ainda que em relação a importância relativa de cada variável para a precisão da identificação, as 05 mais importantes são (HOSPITAL, ANO_REG, CIDADE, AGRESSAO_CRANIO, MÊS_REG).

Após uma primeira análise, executamos alguns ajustes nos parâmetros do algoritmo de RF com o intuito de buscar uma melhora na classificação final do modelo, conforme mostra a figura 5:

Figura 5- Criação do segundo modelo random forest com ajustes na quantidade de árvores

```
> # Fine tuning parameters of Random Forest model
> modelo_rf_fine <- randomForest(DESTAQUE ~ ., data = TrainSet, ntree = 900, mtry = 5, importance = TRUE, do.trace=100)
ntree      OOB      1      2
  100:   7.36%  8.50%  6.12%
  200:   7.30%  8.44%  6.05%
  300:   7.08%  8.44%  5.58%
  400:   7.14%  8.44%  5.72%
  500:   7.04%  8.37%  5.58%
  600:   7.08%  8.44%  5.58%
  700:   7.17%  8.62%  5.58%
  800:   7.27%  8.62%  5.78%
  900:   7.24%  8.62%  5.72%
> print (modelo_rf_fine)

Call:
 randomForest(formula = DESTAQUE ~ ., data = TrainSet, ntree = 900,      mtry = 5, importance = TRUE, do.trace = 100)
               Type of random forest: classification
                     Number of trees: 900
No. of variables tried at each split: 5

        OOB estimate of  error rate: 7.24%
Confusion matrix:
     D    N class.error
D 1495  141  0.08618582
N   85 1402  0.05716207
> round(importance(modelo_rf_fine), 2)
                     D      N MeanDecreaseAccuracy MeanDecreaseGini
CIDADE          105.93  80.33               122.54           127.75
SEXO             46.70  48.16                53.84            33.25
AGRESSAO_ABDOME  51.98  45.99                60.61            26.17
AGRESSAO_CRANIO 102.46  91.41               117.35            87.39
AGRESSAO_DORSO   43.11  45.01                55.72            29.33
AGRESSAO_FACE    62.92  50.75                71.54            49.73
AGRESSAO_MID     31.66  32.73                41.52            20.74
AGRESSAO_MIE     31.19  36.38                45.30            24.11
AGRESSAO_MSD     40.06  38.89                51.27            27.69
AGRESSAO_MSE     35.66  39.20                49.46            24.59
AGRESSAO_PESCOCO 47.90  44.39                58.11            25.67
AGRESSAO_TORAX   65.91  66.46                81.03            53.37
HOSPITAL        211.93 189.44               243.55           444.77
DIA_REG          78.28  77.64               100.94           102.81
MES_REG          84.61  82.92               109.11           131.03
ANO_REG         107.97 106.26               130.32           126.06
IDADE_DISC       73.32  73.31                96.08            84.33
H_periodoDia     69.38  70.79                91.55            78.19
```

Fonte: Produzido pelo Autor

Analisando as informações do mesmo modelo sendo com ajustes na quantidade de árvores criadas que passou para 900 e na seleção de 5 variáveis para cada divisão, conforme informações da figura 5, identificamos que a taxa de erro teve uma diminuição de 7,43% para 7,24%, classificando erroneamente 141 vítimas que foram intituladas destaques, como não destaque 85 vítimas classificadas. Neste modelo o erro na classificação da vítima como destaque caiu melhorou, acertando em mais 7 vítimas, porém, na classificação como não destaque errou em 01 vítima a mais.

Observamos ainda que em relação a importância relativa de cada variável para a precisão da identificação, as 05 mais importantes continuaram as mesmas (HOSPITAL, ANO_REG, CIDADE, AGRESSAO_CRANIO, MÊS_REG).

4 Resultados

Os resultados dos experimentos foram os seguintes.

Apresentaremos os resultados tanto para os dados de treinamento como para os dados de validação, conforme mostra a figura 6:

Figura 6-apresentando o modelo random forest para os ados de treinamento e validação

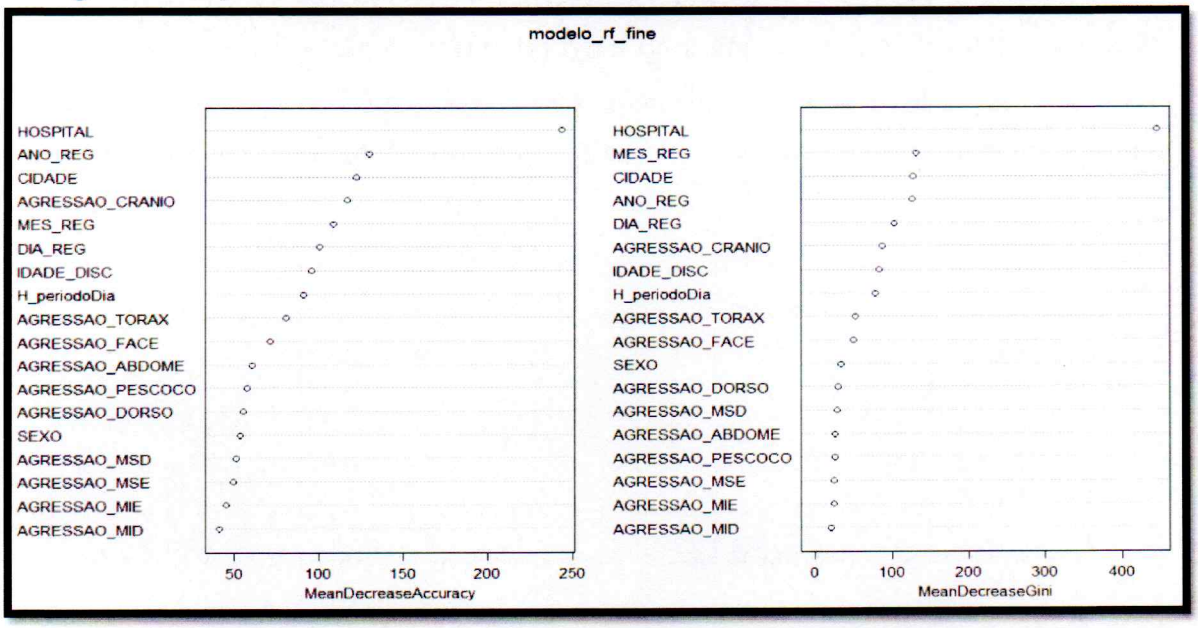
```
> # Predicting on train set
> predTrain <- predict(modelo_rf_fine, TrainSet, type = "class")
>
> # Checking classification accuracy
> table(predTrain, TrainSet$DESTAQUE)

predTrain    D    N
        D 1633    1
        N    3 1486
>
>
> # Predicting on Validation set
> predValid <- predict(modelo_rf_fine, ValidSet, type = "class")
> # Checking classification accuracy
> mean(predValid == ValidSet$DESTAQUE)
[1] 0.9536968
> table(predValid,ValidSet$DESTAQUE)

predValid   D   N
        D 681  27
        N  35 596
```

Fonte: Produzido pelo próprio autor

Analisando os resultados do modelo, observamos que a acurácia foi de aproximadamente 95%, ainda assim, na base de treino, errou na classificação de 01 vítima como destaque e 03 como não destaque. Já na base de validação, o modelo errou na classificação de 27 vítimas intituladas destaque e 35 intituladas não destaque.

Em relação a importância das variáveis para a realização da classificação dos dados, a figura 7 nos traz detalhes conforme abaixo descrito:

Figura 7- importancia das variáveis baseadas na classificação dos dados



Fonte: Produzido pelo próprio autor

Ao analisar o gráfico acima, ele nos mostra que a grande maioria das variáveis analisadas não agregam muito valor na classificação final do vitimado, dentre todas observamos que o hospital para o qual a vítima é conduzida se mostra um fator de grande valor na classificação da vítima de agressão como destaque ou não e características registradas como a agressão aos membros (braços direito e esquerdo, pernas direita ou esquerda) não agregam quase nenhum valor no objetivo de classificação da vítima.

5 Conclusão

Neste trabalho foram feitas classificações de vítimas de agressão catalogadas e disponibilizadas pelo CBMPE, utilizando *Random Forest* buscando a identificar se as vítimas eram Destaques ou não.

Foi observado que o modelo desenvolvido, baseado nos dados que foram disponibilizados, teve um excelente desempenho, com uma acurácia de aproximadamente 95%, mesmo sem que fosse necessário muito ajuste tanto nos dados como nos parâmetros do modelo.

Vale destacar a análise de importância dos atributos, através do algoritmo *random Forest,* em função do resultado apresentado na tabela 7, no qual indica que a maioria dos atributos analisados não agregam valor para a caracterização da vítima como "destaque" ou "não destaque".

Por fim, concluímos que o modelo construído se mostrou bastante eficaz na classificação das variáveis presentes nos dados analisados, porém se faz necessário aprofundar o estudo, levando em consideração outros aspectos importantes na dinâmica do atendimento ao vitimado, como os registros quantitativos de dados da própria definição do termo "destaque", relacionados aos sinais vitais (frequências respiratória e cardíaca, pulso, pressão arterial, temperatura, saturação de oxigênio) e avaliação do nível de consciência e responsividade, através da Escala de Coma de Glasgow registradas nas fichas de ocorrência no ato do atendimento das equipes do Corpo de Bombeiros, comparando o grau de importância dessas novas variáveis apresentadas, com os resultados atuais apresentados.

6 Referências

[1] IBGE, Instituto Brasileiro de Geografia e Estatística, **Perfil das Cidades – Pernambuco**. Disponível em: <https://cidades.ibge.gov.br/brasil/panorama>. Acesso em 23 mai. 2019, 23:15:00

[2] IPEA, Instituto de pesquisa Econômicas Aplicada, **Relatório do Atlas da Violência 2018**, Rio de janeiro: Ipea, jun. 2018. Disponível em <http://www.forumseguranca.org.br/wp-content/uploads/2018/06/FBSP_Atlas_da_Violencia_2018_Relatorio.pdf>. Acesso em 23 mai. 2019, 23:18:00

[3] IBGE, Instituto Brasileiro de Geografia e Estatística, **Contagem da População 2007**, Disponível em: <https://ww2.ibge.gov.br/home/estatistica/populacao/contagem2007/contagem.pdf > Acesso em 23 mai. 2019, 23:22:00

[4] População recenseada, por situação do domicílio, **Base de Dados do Estado de PE, 2007,** Disponível em: <http://www.bde.pe.gov.br/visualizacao/Visualizacao_formato2.aspx?CodInformacao=942&Cod=3> Acesso em 23 mai. 2019, 23:30:00

[5] PERNAMBUCO 2014 **Pacto pela Vida: Plano Estadual de Segurança Pública, 2007**. Disponível em: < http://www.seres.pe.gov.br/index/pacto_pela_vida.pdf> Acesso em: 23 mai. 19. 23:35:00.

[6] RATTON, José Luiz; GALVÃO, Clarissa; FERNANDEZ, Michelle. O Pacto Pela Vida e a Redução de Homicídios em Pernambuco. **Artigo Estratégico – Instituto Igarapé**, 2014. Disponível em: http://igarape.org.br/wp-content/uploads/2014/07/artigo-8-p2.pdf. Acesso em: 23 mai. 19, 23:37:00

[7] CORREA, C. et al. Atendimento pré-hospitalar a vítimas de crime violento intencional: efetividade do Corpo de Bombeiros Militar de Pernambuco. **Revista Flammae**, Recife, v.2 n.5, p. 106-121, edição especial 2016, Disponível em: < https://www.revistaflammae.com/copia-edicao-atual>. Acesso em: 23 mai. 19, 23:43:00

[8] ALVES, F. A. C. Planejamento estratégico como instrumento de gestão pública: uma análise do Projeto Resgate de Vidas do Corpo De Bombeiros Militar de Pernambuco. **Revista Flammae**, Recife, v.1 n.1, p. 331-335, jan./jun. 2015, Disponível em: < https://www.revistaflammae.com/edio-atual>. Acesso em: 23 mai. 19, 23:47:00

[9] DUTRA, K. L.C. Dos destaques às vidas salvas: as milhares de vítimas de crime violento intencional atendidas pelo Corpo De Bombeiros Militar De Pernambuco. **Revista Flammae**, Recife, v.3 n8, p. 331-335, edição extra 2017, Disponível em: < https://www.revistaflammae.com/copia-vol-3-numero-7-1>. Acesso em: 23 mai. 19, 23:55:00

[10] BREIMAN, L. **Random forests**. Machine Learning, 45:5–32. 2001

[11] - QUINLAN, J. **C4.5: Programs for machine learning**. San Mateo: Morgan Kaufmann, 1993. 302p.

[12] = SHANNON Claude E; Weaver,Warren. **The mathematical theory Of communication**, the university of illinois press . Urbana· 1964

[13] CAMPANI, C.; Oliveira, S.; Rodrigues, A. **Entropia e Teoria da Informação**.

FONTE FIGURA 01:Marins, Matheus Araújo, Classificação De Falhas Em Máquinas Rotativas Utilizando Métodos De Similaridade E Random Forest, p25

[14] AGRAWAL, R. **Data Mining**. Tutorial apresentado no 12. Simpósio Brasileiro de Banco de Dados. Fortaleza: Ufc, 1997

[15]S Ali, LC Briand, H Hemmati, RK Panesar-Walawege - IEEE **Transactions on Software Engineering**, 2009

[16] QUINLAN, J. R. **C4.5: programs for machine learning**. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 1993

[17] QUINLAN, J. R. **Induction of decision trees. Machine Learning**, 1:81-106. 1986

[18] TAN, P.-N., Steinbach, M., KUMAR, V. **Introduction to Data Mining**, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 2005

[19] BREIMAN, L., Friedman, J. H., Olshen, R. A., & Stone, C. J.. **Classification and Regression Trees**. Wadsworth. 1984

[20] FONSECA, J. M. M. R. da (1994), **Indução de Árvores de Decisão** – HistClass - Proposta de um algoritmo não paramétrico. Departamento de Informática, Universidade Nova de Lisboa.