Revista de Sistemas de Informação da FSMA
n. 16 (2015) pp. 16-25

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA ESTUDANTIL

Prevaba: Um Modelo Bayesiano para Predição da Existência de Vítimas em Acidentes de Trânsito

Marcelo Josué Telles, *Mestrando em Computação Aplicada, UNISINOS*
Paulo Henrique Santini, *Mestrando em Computação Aplicada, UNISINOS*,
José Vicente Canto dos Santos, *Professor, Doutor em Engenharia Elétrica, UNISINOS* e
Jorge Luis Victória Barbosa, *Professor, Doutor em Ciência da Computação, UNISINOS*.

***Resumo*—A segurança no trânsito é uma área que se preocupa tanto com a redução dos acidentes quanto com o atendimento prestado às vítimas. Diversas iniciativas são propostas para colaborar com a redução dos acidentes, tais como: fiscalização, campanhas de conscientização e equipamentos de auxílio aos condutores. Outras iniciativas para prevenção e proteção são propostas pelos fabricantes de veículos em função de exigências dos órgãos públicos. Como recurso final, isto é, caso ocorra o acidente e a vítima necessite de atendimento médico, este deve ser feito o mais rápido possível. Para auxiliar na identificação da existência de vítima e a necessidade de atendimento médico, é proposto um Modelo Bayesiano, chamado Prevaba. O modelo utiliza Redes Bayesianas (RBs) e tem por finalidade predizer a existência de vítimas em acidentes de trânsito. Para validar o modelo, foi desenvolvido um protótipo que realizou a classificação de dados reais da cidade de Porto Alegre - RS relativos aos acidentes do ano de 2013. O protótipo realizou a classificação com base nos dados do ano anterior (2012), demonstrando um índice de acerto superior a 90%, levando em consideração que as classificações incorretas são apenas as classificadas como sem vítimas, mas na verdade havia vítima.**

***Palavras-chave*—Inferência estatística, classificadores, mineração de dados, tomada de decisão.**

Prevaba: a Bayesian Model to Predict the Existence of Victims in car accidents

***Abstract*—Road safety is an area which is concerned with both the reduction of accidents as with the care provided to the victims. Several initiatives are proposed to assist with reducing the number of accidents, such as surveillance, awareness campaigns and support equipment to drivers. Other initiatives for prevention and protection are proposed by vehicle manufacturers in terms of requirements of governament entities. As a final resort, that is, in the event of the accident and the victim needs medical attention, this should be done as quickly as possible. To assist in identifying the existence of the victim and the need for medical care, we propose a Bayesian model, called Prevaba, which uses Bayesian Networks (BN), which aims to predict the existence of victims in traffic accidents. In order to validate the model, we developed a prototype that performed the actual data classification in Porto Alegre - RS for the year 2013. The prototype made the classification based on the previous year's data (2012), showing an index above 90% accuracy, taking into account the incorrect classifications are only classified as victimless, but actually was has a victim.**

***Index Terms*—Statistical inference, classifiers, data mining, decision making.**

Autor correspondente: primeiro Autor, *marcelojtelles@gmail.com*

I. Introdução

Com a frequente evolução tecnológica relacionada aos dispositivos móveis e computacionais, torna-se possível uma maior interação, aplicação e inserção destes mecanismos no ambiente pessoal do usuário. Essa realidade vem concretizando uma previsão clássica relacionada à possibilidade de monitorar as ações humanas, e receber informações captadas a qualquer momento e em qualquer lugar [1]. Uma possibilidade que surge com a implantação do monitoramento, aliado ao advento da conectividade é o controle do trânsito e a identificação de locais onde a probabilidade de ocorrer um acidente é maior.

É consenso no Brasil a necessidade de reduzir as taxas de acidentes no trânsito. Neste contexto, um modelo de Plano Nacional de Ação (PNA) para a realidade brasileira foi sugerido pela Organização das Nações Unidas (ONU) [2]. Inicialmente o plano teve uma boa aceitação e foram adotadas as medidas necessárias, sendo ainda instituída a nomenclatura "Década de Ações de Segurança do Trânsito (2011-2020)", porém até junho de 2013 nenhum plano havia sido publicado. Mesmo sem uma política nacional definida, existe uma proposta preliminar [3], composta por um conjunto de medidas que visam contribuir para a redução das taxas de mortalidade e lesões por acidentes de trânsito.

Órgãos brasileiros encaram constantemente o desafio de um trânsito mais seguro, como é o caso da Empresa Pública de Transporte e Circulação (EPTC) de Porto Alegre-RS. As ações tomadas pela EPTC impactam positivamente na promoção da segurança no trânsito, uma vez que ela analisa todos os projetos relacionados à mobilidade urbana com ênfase na segurança viária [4]. Já o Instituto Nacional de Metrologia, Qualidade e Tecnologia (Inmetro) [5], define itens de segurança para os meios de transporte.

Visando colaborar com o PNA e também com a preocupação ao atendimento rápido para as vítimas de acidentes no trânsito, é proposto um modelo Bayesiano, que utiliza informações coletadas em eventos ocorridos anteriormente, sobre acidentes de trânsito com feridos ou sem feridos, a fim de predizer se há ou não vítimas em um acidente que ocorreu em um momento específico, sendo que este momento específico pode ser o momento corrente.

O modelo realiza um cálculo probabilístico para estimar a probabilidade da existência de vítimas, com base nessa informação as autoridades locais podem tomar ações adequadas para cada situação, seja no deslocamento de equipes para socorro ou implantação de outros recursos para segurança.

Este artigo está organizado da seguinte forma. A seção II contém uma descrição de conceitos, entre eles: Teoria da Probabilidade, Lei da Probabilidade Total e Teorema de Bayes e Redes Bayesianas. A seção III é dedicada ao modelo proposto, incluindo metodologia e implementação. A seção IV apresenta os testes e resultados obtidos. Finalmente, a seção V apresenta as considerações finais e possíveis trabalhos futuros.

II. Redes Bayesianas

Em 1763 o matemático Thomas Bayes [6] publicou o seu teorema que defendia a ideia de representação do conhecimento que trabalha com as incertezas através da teoria de probabilidade. As RBs são constituídas de nós e arcos, cada nó é interpretado como sendo um atributo que pode receber uma quantidade determinada de valores, tais valores devem ser nominais. De acordo com o valor que determinado nó assume, este pode estabelecer uma ligação, através de um arco, com outro nó. Um nó da rede (geralmente o último) pode ser definido como o nó que identifica a classe de uma instância. Neste trabalho identificar a classe, consiste em atribuir para um acidente, a existência ou não de feridos.

Para a implementação de uma RB são necessários dois conceitos: evidência e inferência. As evidências são novas informações disponíveis que surgem em um dado evento. Essa informação faz com que aconteça uma possível mudança no estado de um nó, assim, inicia-se o processo de propagação das probabilidades. Por exemplo, caso um nó da rede assuma um valor x, pode-se afirmar que essa situação recebe o nome de instanciação [7]. Uma consideração importante é a não obrigatoriedade de que todos os nós tenham seus valores preenchidos. O processo de inferência ocorre quando as probabilidades de uma rede são calculadas, dada uma ou mais evidências. O cálculo das probabilidades de um nó fará com que seja criado um fluxo de informação, que se propagará por toda a rede. A inferência, também pode ser chamada de propagação das probabilidades.

O cálculo de probabilidades é um campo da matemática que permite o tratamento de erros e da incerteza, além disso, é um ferramental estatístico destinado ao estudo de fenômenos probabilísticos. Em tais fenômenos o resultado de um experimento não pode ser previsto com certeza, mas em geral é possível relacionar todos os resultados possíveis de ocorrer. Este conjunto de resultados é chamado de espaço amostral denotado por S ou espaço das probabilidades. Outro fundamento presente nas RBs é o conhecimento de um especialista, o qual deve estabelecer a estrutura das probabilidades.

A. Teoria da Probabilidade

A tomada de decisão é uma teoria prescritiva ou normativa [8] com o objetivo de ajudar as pessoas a tomarem decisões mais adequadas às suas preferências. A probabilidade é definida por uma base matemática firme, a qual é constituída de axiomas [9].

Uma probabilidade geralmente está condicionada com o estado de uma informação disponível [10]. É bastante frequente o caso em que o estado de uma informação é modificado pela ocorrência de algum outro evento relacionado com o experimento em questão. Portanto, é possível usar a definição clássica sobre probabilidade [11], quando cada resultado no espaço amostral tem a mesma probabilidade de ocorrer. Desta forma, a probabilidade de um evento A ocorrer é representada pela equação (1):

$$P(A) = \frac{n_A}{n} \tag{1}$$

Onde n_A é o número de resultados favoráveis ao evento A, e o n é o número de resultados possíveis.

Já se $\mathrm{P}(A)$ for condicionada ao evento B, a probabilidade do evento A será denotada por $\mathrm{P}(A|B)$, para $\mathrm{P}(B)>0$ e $\mathrm{P}(A)>0$. Desta forma, temos as seguintes equações (2 e 3):

$$P(A|B) = P\left(A \bigcap B\right) / P(B). \tag{2}$$

$$P(B|A) = P\left(A \bigcap B\right) / P(A). \tag{3}$$

Com isso, obtém-se a regra do produto, dada pela equação (4).

$$P\left(A \bigcap B\right) = P(A) * P(B|A) = P(B) * P(A|B). \tag{4}$$

B. Lei da Probabilidade Total e Teorema de Bayes

Um dado espaço amostral S, juntamente com os eventos A_1, A_2, A_3, A_4, ..., A_n que são mutuamente exclusivos e exaustivos e um evento B podem ser simbolicamente representados por um Diagrama de Venn, conforme mostra a Figura 1 [8].

A coleção de eventos do $A_1,\ldots,A_n$ define uma distribuição de probabilidade, significando que um e somente um desses eventos irá ocorrer, respectivamente, com probabilidades $\mathrm{P}(A_1),\ldots,\mathrm{P}(A_n)$ cuja soma é unitária.

O Teorema da Probabilidade Total é dado pela equação (5):

$$P(B) = \sum_i P(A_i) P(B|A_i) \tag{5}$$

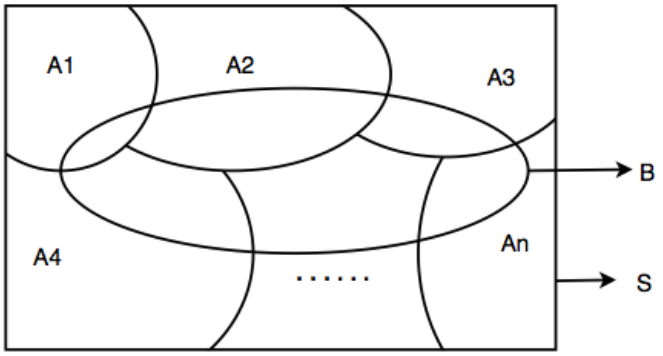

Figura 1: Diagrama de Venn adaptado de [8].

Onde i, pode ser qualquer valor de 1 até n. Usando este teorema podemos calcular a probabilidade de A_i dada a ocorrência de B, conforme equação 6:

$$P(A_i|B) = \frac{P(A_i \bigcap B)}{P(B)} \Longrightarrow$$

$$P(A_i|B) = \frac{P(A_i)\,P(B|A_i)}{\sum_j P(A_j)\,P(B|A_j)} \quad (6)$$

Esta é a fórmula de Bayes, útil quando conhecemos as probabilidades de A e a probabilidade condicional de B dado A_i, mas não conhecemos diretamente a probabilidade de B [9].

Na área de segurança no trânsito, a tomada de decisão pode ser auxiliada por modelos probabilísticos [12]. Os modelos descritos em [13], [14], [15], [16], [17], [18] e [19] visam predizer os possíveis locais de acidentes. O modelo proposto neste trabalho é uma extensão dos supra citados, uma vez que objetiva-se predizer a ocorrência de vítimas em um acidente.

C. Redes Bayesianas

Uma RB pode ser representada por um grafo acíclico dirigido, onde cada nó terá informações probabilísticas daquele evento e com arcos representando dependências sobre estes eventos. Os possíves grafos podem ser seriais, divergentes ou convergentes [20]. A Figura 2 mostra a representação destes 3 tipos de grafos.

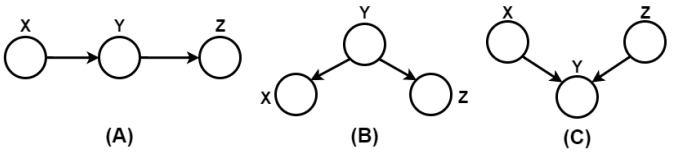

Figura 2: Possíveis grafos das RBs [20].

Na Figura 2, o item (A) mostra um grafo serial, o item (B) um grafo divergente e o item (C) um grafo convergente. O modelo proposto usa o classificador *Naive Bayes* [21], o qual adota um grafo divergente. Este classificador foi escolhido, pois é possível construir modelos probabílisticos, com as informações coletadas automaticamente por sensores.

O classificador *Naive Bayes* assume que os elementos são independentes, não havendo uma busca da hipótese para estimar a classe do novo elemento. A hipótese ou a classificação é formada simplesmente pela contagem de frequências [22].

III. Modelo Proposto

Atualmente existem diversas ferramentas disponíveis para o desenvolvimento de modelos Bayesianos [23], [24] e [25]. Uma lista completa de ferramentas para desenvolvimento de RBs, pode ser obtida em [23], que é um site onde encontram-se referências para desenvolvimento de RBs baseadas em diversas linguagens de programação, tais como C, C++, LISP, Python, C#, Java, Scala e Matlab. Algumas ferramentas apenas realizam a classificação dos dados de entrada, outras geram a RB. JBNC [24] é utilizada em várias aplicações de inteligência artificial [7] e [26], aprendizagem de máquina [20] e mineração de dados [27]. Uma das desvantagens do JBNC, é que todos os classificadores disponibilizados por esta ferramenta, se baseiam no algoritmo *Naive Bayes*. Por fim, a ferramenta Weka, oferece diversos classificadores, entre eles, o BayesNet [28].

Neste trabalho foi adotado o ferramental estatístico disponibilizado pela Universidade de Waikato, denominado Weka [25]. A ferramenta permite explorar recursos de forma gráfica ou ainda adicionar as suas funcionalidades em programas externos desenvolvidos na linguagem de programação Java, através de uma biblioteca. A escolha desta ferramenta foi a facilidade de encontrar documentação e a possibilidade de oferecer dados para classificação separadamente dos dados de treinamento.

A. Metodologia

Para modelar a RB, foi utilizado um conjunto de dados para sua configuração. Tal conjunto vem de uma fonte real, o DataPoa [29], disponibilizado pela prefeitura municipal da cidade de Porto Alegre-RS, com o objetivo de proporcionar à comunidade o desenvolvimento de soluções inteligentes utilizando dados abertos.

A estrutura da RB é definida pelo algoritmo BayesNet [28], este algoritmo não atribui relação entre os valores dos atributos, desta forma as probabilidades são calculadas sem levar em consideração a dependência dos possíveis valores que um atributo pode assumir.

O algoritmo BayesNet é adotado em situações onde não existem informações suficientes para estabelecer dependência entre os atributos, ou seja, o BayesNet classifica o elemento de uma RB com atributos independentes. Na Figura 3 é apresentada a estrutura da RB.

As informações utilizadas são restritas aquelas que podem ser coletadas de forma automática através de sensores. Algumas, disponíveis pelo DataPoa [29] não foram consideradas, uma vez que estas informações, necessitam

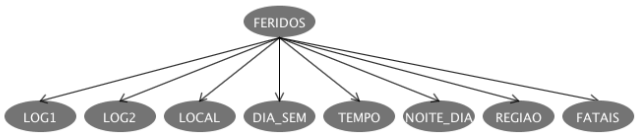

Figura 3: RB usando algoritmo BayesNet.

de interpretação humana. Por exemplo: o tipo do acidente, se existe feridos ou mortos.

A abordagem Bayesiana do modelo proposto, adota as informações a priori, que formam a base para os cálculos. Estas informações formam a base para os cálculos dos percentuais de ocorrência para cada evento.

Os dados utilizados para configuração da rede são do ano de 2012 [30], que totalizam 20202 acidentes registrados na cidade de Porto Alegre - RS. Destes acidentes, houve um total de 6122 com feridos e 97 com vítimas fatais.

Para realizar a validação do modelo foram classificados dados do ano de 2013 [31]. A classificação consistiu em atribuir o valor *YES* ou *NO* para o atributo feridos (uma variável dicotômica). Cada instância (acidente) tanto de 2012 quanto de 2013, possui um atributo que o classifica, indicando se houve vítima no acidente ou não. Desta forma, cada instância dos dados de 2013 teve o atributo classe removido, em seguida os dados foram classificados com base em informações de 2012, atribuindo a classe probabilisticamente. Por fim, cada instância foi validada comparando o atributo classe gerado pelo modelo com o atributo que de fato foi obtido nos dados coletados.

Os dados coletados são apresentados na Tabela I, onde a primeira coluna se refere ao atributo e a segunda coluna detalha cada atributo (apêndice A).

A RB desenvolvida é constituída por um algoritmo que consegue realizar a classificação de 1 até n instâncias, desde que tenha recebido informações anteriores com a quantidade de instâncias maior ou igual a n. Essa classificação consiste em predizer se em um acidente existe(m) ou não vítima(s).

Um sistema de trânsito poderia registrar um acidente com sensores específicos, capazes de identificar automaticamente os itens citados na Tabela I. Existem outros dados importantes no desenvolvimento de modelos de previsão de acidentes [32], que não foram contemplados, pois não podem ser coletados automaticamente. Ao identificar um acidente, os dados são enviados ao modelo Bayesiano e este infere a probabilidade de ocorrência de vítima(s). Com a inferência realizada é possível executar ações de prestação de socorro, uma vez que o modelo informa instantaneamente se em um acidente existe(m) vítima(s) que necessita(m) de atendimento. Identificada a necessidade de atendimento, é possível emitir um alarme em uma central, a fim de deslocar atendimento para prestação de socorro.

B. Implementação

Para realizar a implementação, os dados originais do DataPoa [29] foram tratados a fim de obter apenas dados nominais, isto é, apenas dados que tivessem valores previamente identificados, por exemplo “segunda-feira”, “terça-feira”. Alguns dados foram desconsiderados, sendo que os dados utilizados foram os seguintes: Log1, Log2, Local, Dia Semana, Tempo, Noite Dia, Região, Feridos e Fatais, detalhados na Tabela I.

Tabela I: *Layout* dos dados utilizados no modelo

Atributo	Significado do atributo
LOG1	Rua na qual o veículo que provocou o acidente se localizava.
LOG2	Indica a próxima rua no trajeto do veículo que provocou o acidente.
LOCAL	Indica se foi em um segmento contínuo da rua ou em um cruzamento.
DIA_SEM	Indica o dia da semana em que ocorreu o acidente.
TEMPO	Indica as condições do clima quando ocorreu o acidente: Bom, Nublado, Chuvoso e Não cadastrado.
NOITE_DIA	Indica se estava noite ou dia quando o acidente ocorreu.
REGIÃO	Indica a região da cidade que ocorreu o acidente: Leste, Norte, Sul, Centro e Não Identificado.
FERIDOS	Indica se houve feridos ou não. Os possíveis valores são: Sim ou Não. Este atributo é considerado como atributo classe.
FATAIS	Indica se houve vítima fatal no acidente. Os possíveis valores são: Sim ou Não. Este atributo também poderia ser considerado como um atributo classe.

Os dados disponibilizados pelo DataPoa encontram-se no formato *Comma Separated Values (CSV)*, no entanto para facilitar a leitura dos arquivos, estes foram convertidos para o formato *Attribute Relation File Format (ARFF)*. Os dados relativos a posição geográfica do acidente (latitude e longitude) foram utilizados apenas para apresentação do mapa com os locais dos acidentes. O classificador apenas utilizou o atributo região, pois Porto Alegre se divide em: Leste, Norte, Sul e Centro. A latitude e longitude, não foram utilizados pois se tratam de valores distribuídos em um intervalo que contempla muitos valores, os quais teriam probabilidades dispersas.

Durante a implementação foi desenvolvido um sistema para *desktop* utilizando a linguagem de programação Java. Na Figura 4 é apresentada a tela principal (*Classifiers*) do sistema.

Na tela principal é possível selecionar qual é a base de treinamento e definir a(s) instância(s) a ser(em) classificadas do ano de 2013. É possível realizar a classificação de acidentes de outros anos, no entanto nesta validação foram adotados dados de 2013. Pode ser classificada apenas uma instância específica ou ainda uma quantidade qualquer desde que menor que 20202, pois a base de treinamento possui tal quantidade de instâncias. Também é possível classificar grupos pré-definidos de instâncias, sendo que tais grupos foram definidos apenas para fins de testes.

Figura 4: Tela principal (*Classifiers*) do sistema.

O sistema também possui uma tela secundária chamada (*Geo Statistics*), onde é possível visualizar graficamente em um mapa da cidade de Porto Alegre - RS os locais onde ocorreram acidente no ano de 2012, 2013, acidentes com vítimas fatais em 2012 e acidentes com vítimas fatais em 2013. Na Figura 5 é apresentada a tela secundária do sistema.

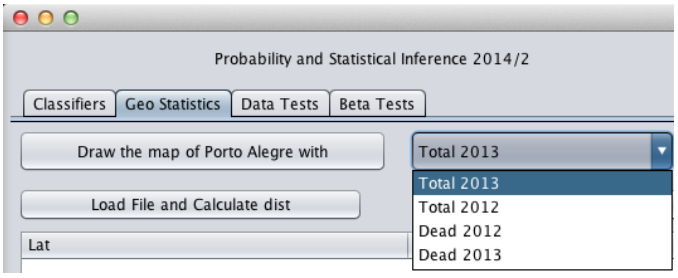

Figura 5: Tela secundária (*Geo Statistics*) do sistema.

Na Figura 6 é apresentado o mapa com os acidentes do ano de 2012, utilizando a latitude e longitude dos acidentes. Com a visualização dos acidentes no mapa é possível identificar os locais onde existe maior concentração de acidentes. Os locais onde poderiam ser instalados os equipamentos de monitoramento de trânsito para coleta automática de dados, poderiam ser definidos com o auxílio deste mapa. Cabe salientar, que é possível aproximar os marcadores, aplicando um *zoom*, porém não seria possível identificar que se trata de Porto Alegre-RS.

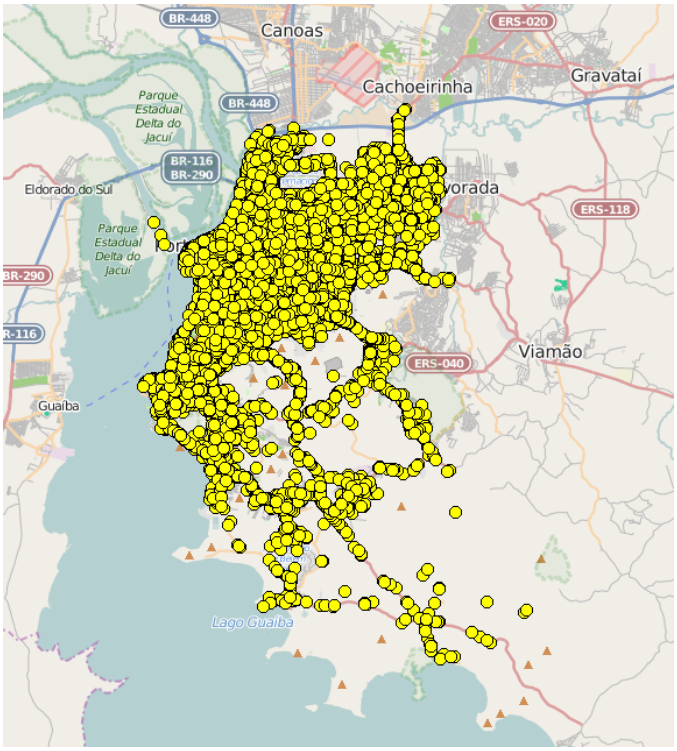

Figura 6: Mapa com os locais dos acidentes.

C. Funcionamento

Para testar o funcionamento da ferramenta Weka, foram testados exemplos disponibilizados no pacote da mesma, além disso, foi montada uma RB que leva em consideração 9 atributos para definir a classe de uma instância.

Para configuração da RB e criação da tabela de probabilidades foram inseridas duas instâncias de teste, desta forma a tabela de probabilidades pode ser testada (Figura 7), mostrando 50% de chance para *YES* e 50% de chance para *NO*.

Para verificar a operação da tabela de probabilidades, foi inserida mais uma instância, com valor *YES*, sendo que esta provocou um aumento nas chances do atributo classe receber o valor *YES*.

Conforme pode ser observado na Figura 8 a tabela de probabilidades possui um percentual maior de chances (62,50%) para o valor *YES*, já que mais uma instância foi adicionada na base de treinamento com o valor *YES*.

IV. Testes e Resultados

Primeiramente, foi feita a classificação de 19000 instâncias dos dados de 2013, usando como base de treinamento os próprios dados de 2013. Nestes dados houve um total de 5870 acidentes com feridos. Na tabela II, é apresentada a matriz de confusão gerada pelo classificador. A classificação foi feita com *cross-validation* [33] *folds*=10, isto é, para cada 10 instâncias foi feito o treinamento e classificação das demais, com subsequentes trocas pelas próximas 10 instâncias para treinamento e validação dos remanescentes.

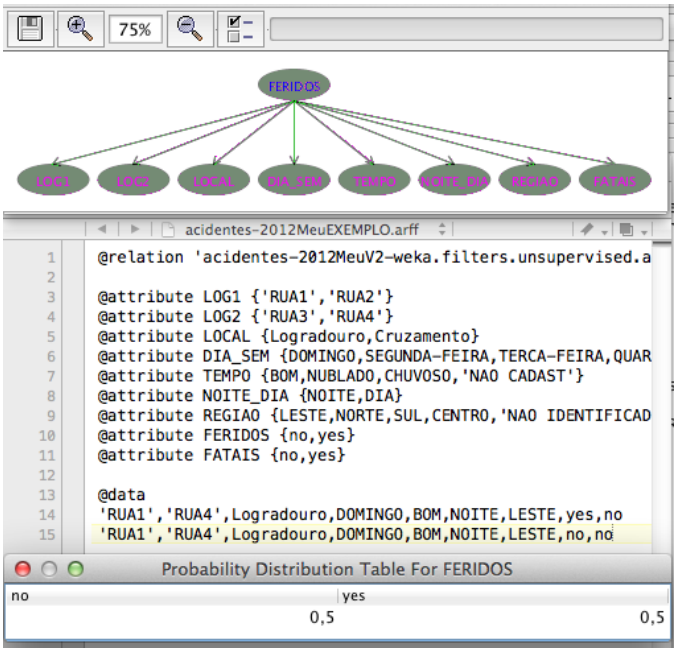

Figura 7: RB e tabela de probabilidade gerada pela ferramenta Weka, para teste de probabilidade.

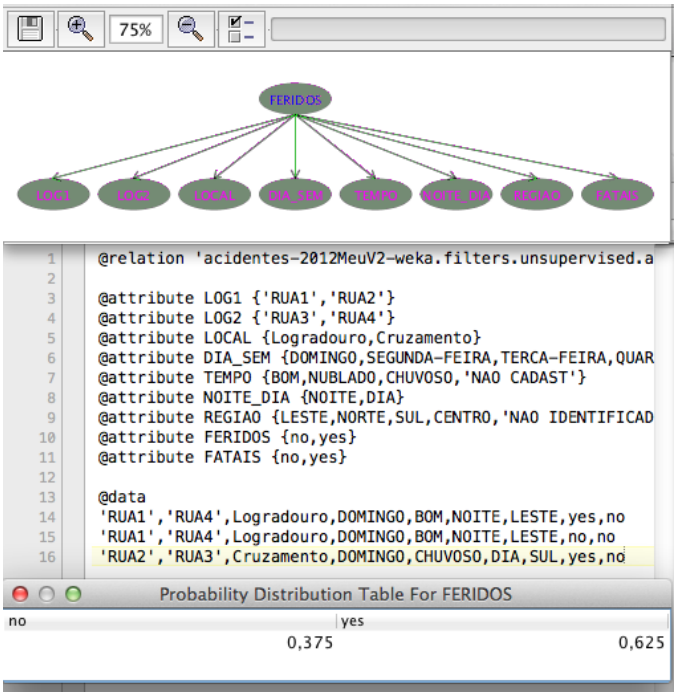

Figura 8: Tabela de probabilidade com valores alterados.

A matriz de confusão apresenta as classificações corretas na diagonal principal. O classificador classificou corretamente 13676 instâncias (10323+3353), sendo que 10323 instâncias com o atributo *NO* foram classificadas corretamente com o valor *NO*. Em 2517 classificações o valor *NO* foi atribuído, no entanto se tratavam de instâncias com valor *YES*. Isso indica que 2517 instâncias foram classificadas como acidentes sem feridos, mas na verdade se tratavam de acidentes com feridos. O percentual de

Tabela II: Matriz de Confusão com dados *in sample*

Frequências Observadas	Frequências Estimadas		Total
	No	Yes	
No	10323	2807	13130
Yes	2517	3353	5870
Total	12840	6160	19000

acerto do classificador foi de 71,97% (13676 de 19000). Já o percentual de acerto para os acidentes com feridos foi de 57,12% (3353 de 5870).

O percentual de 57,12% é considerado como um percentual de acerto baixo, pois praticamente metade das classificações são feitas incorretamente. Na tabela V será apresentado o percentual obtido com o modelo.

Os resultados *in sample* consideraram a classificação dos acidentes de 2013 com base nas respostas da rede treinada com dados de 2013. Por sua vez, os resultados *out sample* consideraram a classificação dos acidentes de 2013 com base nas respostas da rede treinada com dados de 2012.

Na sequência os testes foram realizados passando os dados de 2012 para o modelo, desta forma, a RB foi configurada gerando a tabela de probabilidades, a fim de predizer se as instâncias de 2013 tiveram feridos nos acidentes. É válido destacar que os dados de 2013 tiveram o atributo feridos (que define a classe) removido. Com a finalidade de aumentar o percentual de acerto (para *YES*), os dados de 2012 foram alterados, adicionando-se uma nova instância com o valor *YES*. Tal adição foi realizada utilizando-se como métrica o valor de ocorrências de YES a cada 10 amostras, isto é, para cada 10 acidentes, foi realizada a verificação se já havia 4 amostras com valor YES, caso não tivesse, foi adicionada uma amostra com YES, visando aumentar o percentual de probabilidade para tal classificação. Nesse e em outros casos, no final, foi realizado um teste para verificar se o padrão de 40% foi obedecido ou não.

A primeira coluna da Tabela III mostra o número que indica qual é o acidente que está sendo classificado, a segunda e terceira indicam o total acumulado para o valor *YES* e *NO* no atributo FERIDOS, a quarta e a quinta apresentam o valor estimado acumulado (classificado pelo modelo) para o atributo FERIDOS.

É possível perceber que o valor de *YES* aumenta consideravelmente, fazendo com que a taxa de acerto seja reduzida, pois os valores são predominantemente *YES*. Isso ocorreu pois o objetivo é predizer corretamente os acidentes com *YES*. Geralmente esse valor deve ficar em torno de 22,50%, conforme pode ser observado na Tabela IV. Mesmo com valores diferentes, em cada ano, é possível estabelecer uma estimativa para o próximo ano, tomando como base a média e o desvio padrão.

Na Tabela IV é apresentado o total de acidentes (coluna "Acidentes"), para cada ano, desde o ano 2000 até 2013. Adicionalmente é apresentada a quantidade de acidentes com feridos (coluna "Com Feridos"). Na coluna "Percentual" é calculado o percentual dos acidentes com feridos. A média de acidentes com feridos levando-se em consideração

Tabela III: Tabela de instâncias – acidentes ocorridos em janeiro de 2013

	Dados 2013		Dados Estimados		Dados 2012	
Acidente	YES	NO	YES	NO	YES	NO
1	0	1	0	1	1	0
2	0	2	0	2	2	0
3	0	3	1	2	2	1
4	1	3	2	2	2	2
5	1	4	3	2	2	3
6	1	5	4	2	3	3
7	2	5	5	2	3	4
8	2	6	5	3	4	4
9	3	6	5	4	5	4
10	4	6	5	5	6	4
11	4	7	6	5	7	4
12	4	8	7	5	7	5
13	4	9	8	5	8	5
14	4	10	9	4	8	6
15	4	11	10	5	9	6
16	5	11	10	6	10	6
17	6	11	10	7	11	6
18	6	12	11	7	11	7
19	6	13	11	8	11	8
20	7	13	12	8	11	9
21	8	13	12	9	11	10
22	8	14	13	9	11	11
23	8	15	14	9	12	11
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1550	436	1113	1224	326	651	899

do ano 2000 até 2013 foi calculada e apresentada na última linha da tabela (22,50%). Nesta tabela é apresentado o total de acidentes com feridos no ano de 2013. É possível verificar que o ano de 2013 apresenta um percentual de acidentes com feridos, maior do que a média calculada, pois foram registrados 6456 acidentes com feridos, chegando a média de 31,04%. No ano de 2012, a média foi de 30,30%, levando-se em consideração a alteração proposta, a média tende chegar em 40%. Esta alteração foi proposta para garantir que o maior número de acidentes observado, tivesse como ser alcançado, impossibilitando uma classificação menor do que qualquer uma já observada.

O percentual de acidentes com feridos trata-se de uma informação aleatória, desta forma não é possível afirmar com certeza, que o percentual 22,50, será observado sempre. Na Tabela III é apresentada a classificação dos dados de 2013, com treinamento nos dados de 2012, já alterados. A alteração foi realizada identificando se a cada 10 acidentes já haviam ocorrido pelo menos 3 acidentes com feridos (aproximadamente 22,50%), caso não tivesse, foi gerado uma instância com o valor *YES*. A instância foi definida utilizando valores existentes na base de treinamento da rede, ou seja, os registros de 2012.

Tabela IV: Tabela do percentual de acidentes com vítima

Ano	Acidentes	Com Feridos	Percentual
2000	18154	4232	23,31
2001	21139	3870	18,31
2002	22146	4157	18,77
2003	20937	4836	23,10
2004	20663	4727	22,88
2005	20759	4679	22,54
2006	20334	4641	22,82
2007	22245	4828	21,70
2008	22290	4232	18,99
2009	22128	4587	20,73
2010	25475	4879	19,15
2011	23580	5025	21,31
2012	20202	6122	30,30
2013	20799	6456	31,04
		desvio padrão	3,73
		média	**22,50**

Após a classificação das 19000 instâncias de 2013 foi gerada a matriz de confusão com a base alterada, apresentada na Tabela V.

Tabela V: Matriz de Confusão *out sample* com base alterada

Frequências Observadas	Frequências Estimadas		Total
	No	Yes	
No	2860	10270	13130
Yes	1341	4529	5870
Total	4201	14799	19000

A matriz de confusão indicada pela Tabela V demonstra que o modelo classificou corretamente 7389 instâncias (2860+4529), sendo que 4529 instâncias com o atributo *YES* foram classificadas corretamente com o valor *YES*. Em apenas 1341 classificações o valor *NO* foi atribuído, no entanto se tratavam de instâncias com valor *YES*. Isso indica que 1341 instâncias foram classificadas como acidentes sem feridos, mas na verdade se tratavam de acidentes com feridos, um erro bem inferior considerando a primeira classificação, indicada pela Tabela II.

O percentual de acerto do classificador com a base alterada, foi menor, 38,88% (7389 de 19000). Já o percentual de acerto para os acidentes com feridos foi maior, assumindo um valor de 77,15% (4529 de 5870).

Na prática a ambulância deve ser acionada para todos acidentes. No entanto em alguns acidentes não será necessária a presença da mesma, visto que o classificador irá estimar a não existência de vítimas. Como o objetivo é classificar corretamente os acidentes com vítima e não os acidentes sem vítima este inconveniente passa a ser uma preocupação de implementação, não limitando a viabilidade do classificador.

Considerando como classificações erradas apenas os falsos negativos, pois são acidentes fatais classificados como

acidentes sem vítimas fatais, o classificador errou apenas 1341 casos entre 19.000, ou seja, uma taxa de acerto de 92,94%. Nesta mesma tabela são apresentados os totalizadores da classificação das instâncias. Das 19000 instâncias, 5870 eram originalmente acidentes com vítimas. O classificador atribuiu para 14799 (10270+4529) instâncias o valor *YES* e 4201 o valor *NO*. Das classificações realizadas, 1341 foram classificadas como não sendo acidente com vítima (coluna C), no entanto se tratavam de acidentes com vítimas, desta forma foram classificadas incorretamente 1341 instâncias.

Nas tabelas VII e VIII são apresentados os resultados *in sample* e *out sample*. Por sua vez, a tabela IX mostra uma visão geral dos resultados, composta por especificidade e sensibilidade. A especificidade é a parcela (em percentual) de acidentes sem vítimas classificados corretamente entre o total de acidentes sem vítimas. Já a sensibilidade, é a parcela (em percentual) de acidentes com vítimas classificados corretamente sobre o total de acidentes com vítimas e por fim, a taxa de acerto torna-se visível após o conhecimento da especificidade e sensibilidade.

Tabela VII: Especificidade e sensibilidade com dados *in sample*

Especificidade	Sensibilidade
Acertos com Yes	Acertos com No
57,12%	78,62%

Tabela VIII: Especificidade e sensibilidade com dados *out sample*

Especificidade	Sensibilidade
Acertos com Yes	Acertos com No
32,31%	69,33%

Na Tabela VI são apresentadas as instâncias classificadas. O total de classificações com valor *YES*, foi de 14799 (coluna B), no entanto haviam 5870 (coluna A) instâncias que deveriam ser classificadas com tal valor. Para as 13130 instâncias em que não havia ferido, foram classificadas corretamente 2860. Considerando que apenas as classificações indicadas na coluna C são classificações que conduzem para uma tomada de decisão incorreta, o classificador estimou corretamente 17659 instâncias.

Tabela IX: Especificidade e sensibilidade com dados *out sample* (base alterada)

Especificidade	Sensibilidade
Acertos com Yes	Acertos com No
77,15%	21,78%

V. Conclusões e Trabalhos Futuros

Prever a existência de feridos em acidentes de trânsito é de fundamental importância, visto que em muitos acidentes (77,50%) não há necessidade de deslocamentos de equipe de socorro, conforme observado no complemento para 100% da Tabela IV.

Este trabalho contribui para as áreas de segurança de trânsito, gestão de recursos da saúde pública e prevenção de vítimas fatais (por demora no atendimento) tornando-se um aliado na tomada de decisão destes três segmentos. A abordagem Bayesiana mostrou-se adequada, visto que realizou 4529 inferências corretas, levando-se em consideração que foram classificados dados de 2013, com base em tabelas de probabilidades estatísticas geradas com dados de 2012, durante a validação realizada na fase de testes.

Quanto a capacidade de classificação do modelo é importante destacar que existe uma exigência de informações anteriores, que devem ser coletadas, para a alimentação do sistema. Para estender o modelo a fim de abranger outras cidades, é necessária a coleta de dados reais, assim como foi feito na cidade de Porto Alegre - RS. Outras questões relacionadas à modelagem que considerem a dependência entre os atributos podem ser implementadas para este tipo de situação, desta forma seria possível desenvolver um outro modelo que contemple a relação entre os atributos.

Outro aspecto importante é com relação à validação, pois isso possibilita consolidação do modelo que foi proposto, além da verificação de sua confiabilidade, para possíveis usos práticos. Sendo assim, como trabalho futuro se vislumbra uma validação usando outros dados de cidades distintas.

Como trabalhos futuros considera-se a necessidade de testar o classificador com os dados de cada ano, fazendo um registro das classificações corretas e incorretas de cada ano com base nos dados de anos anteriores, um a um, ou seja classificar os dados de 2013 com base nos dados de 2012, 2011, 2010 e assim sucessivamente. Possivelmente a melhor base de classificação para cada ano, será o ano anterior, com excessão de 2011 que diverge bastante dos dados observados em 2012.

Para melhorar o classificador cogita-se a possibilidade de adicionar mais atributos relevantes. Para isso, deve ser definido um novo conjunto de atributos, a fim de descobrir qual o mais adequado, um atributo a ser considerado é o tipo de acidente. Acredita-se que a classificação baseada em dados de 2014 tenha resultados diferentes, pois a partir deste ano o atributo feridos se separa em dois tipos, feridos e feridos graves.

Finalmente, a modelagem desenvolvida no presente trabalho indica que é possível modelar uma RB que preveja os acidentes com vítimas, desde que se saiba as reais características do fluxo de veículos nas vias e também sobre os acidentes ocorridos.

Referências

[1] M. Weiser, "The computer for the 21st century," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 3, no. 3, pp. 3–11, Jul. 1999. [Online]. Disponível em: http://doi.acm.org/10.1145/329124.329126

[2] Vias-Seguras, "Plano nacional de redução de acidentes, 2011 - 2020," 2015. [Online]. Disponível em: http://www.vias-seguras.com/a_prevencao/a_decada_de_acoes_de_seguranca_do_transito_2011_2020/plano_nacional_de_reducao_de_acidentes_2011_2020

Tabela VI: Tabela do percentual de acidentes com vítima

	A	B	C	D	E	
	Dados Reais	Dados Estimados	Estimação	Estimação	Estimação	
	YES	YES	Incorreta	Correta	considerada correta	
Instâncias			Gerou NO era YES	Gerou YES era YES		Total (C+E)
19000	5870	14799	1341	4529	17659	19000

[3] Daer, "Estudos estatísticos de acidentes de trânsito," 2012. [Online]. Disponível em: http://www.daer.rs.gov.br/site/controle_estudos_estatisticos_acidentes_transito.php

[4] EPTC, "Empresa pública de transporte e circulação (eptc)," 2015. [Online]. Disponível em: http://www2.portoalegre.rs.gov.br/eptc/default.php?p_secao=203

[5] Inmetro, "Rtq - inspeção de segurança veicular de motocicletas a assemelhados - modificação ou fabricação artesanal," 2014. [Online]. Disponível em: http://www.inmetro.gov.br/rtac/pdf/RTAC0008814.pdf

[6] ISBA, "International Society for Bayesian Analysis ISBA," 2009. [Online]. Disponível em: http://bayesian.org/

[7] K. B. Korb A. E. Nicholson, *Bayesian Artificial Intelligence, Second Edition*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2010.

[8] O. R. Bekman, *Análise estatística da decisão*. Edgard Blucher, 1980.

[9] B. R. James, *Probabilidade: um curso em nível intermediário*, ser. Projeto Euclides, C. Instituto de Matemática Pura e Aplicada, Ed., 1996.

[10] M. J. Zaki J. Wagner Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York, NY, USA: Cambridge University Press, 2014.

[11] R. Larson, B. Farber, C. tradução técnica Patarra, *Estatística aplicada*. Prentice Hall, 2004.

[12] G. F. d. M. Claude, "Previsão da ocorrência de acidentes de trânsito em interseções de vias arteriais urbanas: o caso de Taguatinga – Distrito Federal," Dissertação, Mestrado em Transportes, Universidade de Brasília, Brasília, Maio 2012.

[13] H. Barbosa, F. Cunto, B. Bezerra, C. Nodari, M. A. Jacques, "Safety performance models for urban intersections in Brazil," *Accident Analysis & Prevention*, vol. 70, pp. 258–266, 2014.

[14] G. Cardoso L. G. Goldner, "Desenvolvimento e aplicação de modelos para previsão de acidentes de trânsito," *ANPET - Associação Nacional de Pesquisa e Ensino em Transportes*, vol. 15, no. 2, pp. 43–51, 2007. [Online]. Disponível em: http://www.anpet.org.br

[15] F. J. C. Cunto, M. M. C. Neto, D. S. Barreira, "Modelos de previsão de acidentes de trânsito em interseções semaforizadas de Fortaleza," *ANPET - Associação Nacional de Pesquisa e Ensino em Transportes*, vol. 20, no. 2, pp. 55–62, 2012. [Online]. Disponível em: http://www.anpet.org.br

[16] A. Ferreira, Sara e Couto, "Método probabilístico para identificação de zonas de acumulação de acidentes," *ANPET - Associação Nacional de Pesquisa e Ensino em Transportes*, vol. 21, no. 3, pp. 48–55, 2013. [Online]. Disponível em: http://www.anpet.org.br

[17] U. Brüde J. Larsson, "Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit?" *Accident Analysis & Prevention*, vol. 25, no. 5, pp. 499–509, P1993. [Online]. Disponível em: http://www.sciencedirect.com/science/article/pii/000145759390001D

[18] P. Greibe, "Accident prediction models for urban roads," *Accident Analysis & Prevention*, vol. 35, no. 2, pp. 273–285, 2003. [Online]. Disponível em: http://www.sciencedirect.com/science/article/pii/S0001457502000052

[19] E. Hauer, "Overdispersion in modelling accidents on road sections and in empirical bayes estimation," *Accident Analysis & Prevention*, vol. 33, no. 6, pp. 799–808, 2001. [Online]. Disponível em: http://www.sciencedirect.com/science/article/pii/S0001457500000944

[20] S. Theodoridis, *Machine Learning, A Bayesian and Optimization Perspective, First Edition*, 1st ed. Boca Raton, FL, USA: Elsevier, 2015.

[21] G. H. John P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[22] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

[23] K. Murphy, "Ferramentas para desenvolver uma rede bayesiana," 2014. [Online]. Disponível em: http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html

[24] Jbnc, "Modelos bayesianos," 2014. [Online]. Disponível em: http://jbnc.sourceforge.net/

[25] D. of Computer Science Tari Rorohiko, "Weka - machine learning algorithms in java," 2014. [Online]. Disponível em: http://www.cs.waikato.ac.nz/

[26] P. Dagum M. Luby, "Approximating probabilistic inference in bayesian belief networks is np-hard," *Artif. Intell.*, vol. 60, no. 1, pp. 141–153, Mar. 1993. [Online]. Disponível em: http://dx.doi.org/10.1016/0004-3702(93)90036-B

[27] D. Heckerman, "Bayesian networks for data mining," *Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 79–119, Jan. 1997. [Online]. Disponível em: http://dx.doi.org/10.1023/A:1009730122752

[28] BayesNet, "Classificador bayesnet," 2014. [Online]. Disponível em: http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/BayesNet.html

[29] DataPoa, "Dados coletados referentes acidentes de trânsito," 2014. [Online]. Disponível em: http://www.datapoa.com.br

[30] ——, "Dados coletados referentes acidentes de trânsito," 2012. [Online]. Disponível em: http://www.datapoa.com.br/dataset/acidentes-de-transito/resource/d9add4bb-00a7-4205-bd36-7bd89439a09a

[31] ——, "Dados coletados referentes acidentes de trânsito," 2013. [Online]. Disponível em: http://www.datapoa.com.br/dataset/acidentes-de-transito/resource/a027ffcb-dcbe-46df-b6b7-b8ba69ee1559

[32] G. Cardoso, "Modelos para previsão de acidentes de trânsito em vias arteriais urbanas," Tese de Doutorado, Engenharia de Produção, Universidade Federal do Rio Rande do Sul, Escola de Engenharia, Porto Alegre, Agosto 2006.

[33] K. Yang, H. Wang, G. Dai, S. Hu, Y. Zhang, J. Xu, "Determining the repeat number of cross-validation," in *Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on*, vol. 3, Oct 2011, pp. 1706–1710.

Apêndice

Apêndice A:

Tabela X: Convenção dos atributos

Atributos Originais	Convenção
ID	ignorado
LOG1	utilizado na classificação
LOG2	utilizado na classificação
PREDIAL1	ignorado
LOCAL	utilizado na classificação
TIPO_ACID	ignorado
LOCAL_VIA	ignorado
DATA_HORA	ignorado
DIA_SEM	utilizado na classificação
FERIDOS	utilizado na classificação (YES ou NO)
MORTES	ignorado
MORTE_POST	ignorado
FATAIS	utilizado na classificação (YES ou NO)
AUTO	ignorado
TAXI	ignorado
LOTACAO	ignorado
ONIBUS_URB	ignorado
ONIBUS_INT	ignorado
CAMINHAO	ignorado
MOTO	ignorado
CARROCA	ignorado
BICICLETA	ignorado
OUTRO	ignorado
TEMPO	utilizado na classificação
NOITE_DIA	utilizado na classificação
FONTE	ignorado
BOLETIM	ignorado
REGIAO	utilizado na classificação
DIA	ignorado
MES	ignorado
ANO	ignorado
FX_HORA	ignorado
CONT_ACID	ignorado
CONT_VIT	ignorado
UPS	ignorado
LATITUDE	utilizado apenas para identificar locais para instalação dos sensores
LONGITUDE	utilizado apenas para identificar locais para instalação dos sensores