XIII SIMMEC
Simpósio de Mecânica Computacional
29 de Outubro a 1º de Novembro de 2018
UFES - Campus Goiabeiras - Vitória, ES

USO DE MINERAÇÃO DE DADOS E TECNOLOGIA PREDITIVA NA PREVENÇÃO DE ACIDENTES DE TRÂNSITO NO BRASIL

Use of data mining and predictive technology in the prevention of traffic accidents in Brazil

Maxuel Pereira de Oliveira (1); Jheison Maciel Inês (2); Alan da Silva Lopes (3); Saulo Henrique Gomes Castro (4); Willyan Michel Ferreira (5)
(1)(2)(3)(4) Bacharelandos em Ciência da Computação, Centro Universitário de Formiga – UNIFOR-MG, Formiga-MG, Brasil.
(5) Prof. Me. do Curso Ciência da Computação, Centro Universitário de Formiga – UNIFOR-MG, Formiga-MG, Brasil.
Email para correspondencia: maxuelpoliveira@gmail.com; (P) Maxuel Pereira de Oliveira

Resumo: O Brasil atualmente conta com uma grande frota de veículos em circulação, estima-se aproximadamente 91,1 milhões de veículos. Com isso tem se observado uma grande preocupação de como reduzir o elevado índice de acidentes de trânsito que vem ocorrendo, esses causando traumas, lesões, danos materiais ou óbitos. Os órgãos de fiscalização fazem balanços sobre os acidentes que ocorrem no Brasil, fornecendo dados que podem ser estudados. Pesquisas estão voltadas atualmente para tecnologia preditiva, mediante técnicas de análise de dados e estatísticas, é possível encontrar correlações, padrões, extrair conhecimentos e encontrar tendências que levam a ocorrências destes acidentes. Esta tecnologia está sendo usada em vários segmentos da sociedade, por exemplo, buscar *insights* na área de vendas, *marketing*, instituições financeiras, universidades e principalmente na área de saúde. Anualmente são feitos balanços sobre os acidentes, com isso podemos adotar políticas de controle e prever as características de onde tem maior tendência de ocorrer estes incidentes. Neste artigo é proposto a utilização de algoritmos de árvores de decisão utilizado para classificação de dados. Utilizando o *software WEKA*, no qual, possui algoritmos como o de *RandomTree* e *J48*, o objetivo é aplicar esses na base de dados fornecida pela Polícia Rodoviária Federal do ano de

2007 a 2015 para treinamento e no 2017 para predição. Com os resultados é proposto definir padrões, correlações de forma a contribuir com a prevenção e diminuição destes acidentes.

Palavras chaves: Mineração de dados; Tecnologia preditiva; Árvore de decisão; Árvore Aleatória; Algoritmo J48.

***Abstract*:** *Brazil has a large number of vehicles in circulation, estimated at 91.1 million vehicles. With this, you can get a large number of cases, injuries, property damage or death. The databases on the events that occur in Brazil, providing data that can be studied. Researchers are currently focused on information technology, through data analysis techniques and statistics, are available, correlate with patterns, extra features and find trends that lead to an occurrence of accidents. This technology is being used in various segments of society, for example, seeking insights in sales, marketing, institutions, universities and especially in the health area. The balance sheets on accidents can be included, thus enabling control policies and predicting the characteristics of greater tendency to exercise these incidents. This article is presented for use with images. Using the software WEKA, in which configuration algorithms such as RandomTree and J48 are applied to the database by the Federal Highway Police from year 2007 to 2015 for training and in 2017 for predction. The results were compared to patterns, correlations in order to contribute to the prevention and cancellation of the same.*
Keywords: *Data Mining; Predictive Technology; Decision tree; RandomTree; Algorithm J48.*

1 INTRODUÇÃO

Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE, 2016), o Brasil possui uma população de aproximadamente 207,9 milhões de habitantes, possui também uma frota de 91,1 milhões de veículos em circulação, em média um veículo para 2,28 habitantes. Com isso tem-se uma grande preocupação com a quantidade de acidentes que ocorrem e que possuam vítimas envolvidas. O relatório *Global status report on road safety* da Organização Mundial da Saúde (OMS, 2015), exibe informações de 180 países, estima-se que mundialmente a quantidade de mortes no trânsito atinge cerca de 1,25 milhões pessoas por ano, com uma taxa de mortalidade concentrada em países de baixa renda. Este relatório apresenta também um estudo, que aponta no ano de 2030 este número chegará a aproximadamente 3,6 milhões de

vítimas. Outro fator importante que é levado em conta neste relatório é que a faixa etária das vítimas nos acidentes está entre 15 a 29 anos de idade.

Os acidentes de trânsito no Brasil de uma maneira geral deixam muitas pessoas feridas que na maioria das vezes ficam com sequelas provisórias ou por toda a sua vida, gerando elevados custos com procedimentos hospitalares e gastos com a previdência (Oliveira; Souza, 2012). Em 2012, a Previdência Social gastou um valor acima de R$ 12 bilhões em vantagens relacionadas aos acidentes de trânsito. Essa é uma questão que deixa muita preocupação, pois a maioria dessas pessoas estão em idade funcional (MPS, 2013). Neste contexto, o Brasil encontra-se entre os países com os valores mais elevados de mortes no trânsito (DNIT, 2010), como pode ser observado na **Tabela 1** abaixo.

Tabela 1 - Quantidade de Acidentes X Vítimas Fatais

Ano	Quantidade de acidentes	Quantidade de vítimas fatais
1962	3.486	668
1972	27.114	2.673
1982	42.090	4.056
1992	67.021	5.756
2002	109.025	6.312
2010	182.900	8.616

Fonte: (DNIT, 2010)

Uma proposta apresentada no relatório é aumentar a sinalização nas vias onde concentra os maiores índices de acidentes (OMS, 2015).

A “Era da informação” tem como características o grande volume de dados que são encontrados e armazenados, muitas das vezes inexplorados. Com este conteúdo de dados faz necessário a utilização de procedimentos eficientes, que consigam fazer uma análise mais detalhada e que processe um volume grande de dados, para que a informação gerada seja útil, viabilizando sua efetiva utilização (Berthold; Hand, 2007). Com isso faz necessário o uso da mineração de dados que utiliza técnicas para descoberta de padrões, onde não se faz a retirada de dados e sim a retirada de conhecimento (Han; Kamber, 2011). A Inteligência Artificial (IA) é um dos assuntos com discussões marcantes na comunidade científica atualmente. A procura por técnicas ou ferramentas capazes de simular o funcionamento do raciocínio humano vem

sendo o grande marco desta área a bastante tempo. Desde que começou no início da de década de 50, a Inteligência Artificial vem se aprimorando em muitos setores da ciência e com isso muitos pesquisadores vêm seguindo estas linhas de pesquisa com o intuito de oferecer ao computador as aptidões de efetuar funções que somente o cérebro humano tem a capacidade de resolver. Uma das áreas das IA são os sistemas que apoiam a tomada de decisões (Gomes, 2011).

A Tecnologia Preditiva possui uma enorme abrangência de possibilidades de aplicações. Ferramentas atualmente tratam com facilidade os problemas de análise de dados que uma pessoa demoraria muito tempo para analisar manualmente devido ao uso de técnicas e com a ajuda de máquinas cada dia mais poderosas, podendo ser introduzida no trabalho do dia a dia. Com a ajuda da Inteligência Artificial (IA) que nada mais é que modelos matemáticos e tem como inspiração imitar o funcionamento do cérebro humano (Russell; Norvig, 2004).

Pesquisas estão voltadas para tecnologia preditiva onde se utiliza de técnicas de mineração de dados, descoberta de conhecimento, aprendizado de máquina, onde por meio de análise de dados e estatísticas é possível extrair conhecimentos. Esta tecnologia está sendo utilizada em vários segmentos da sociedade, por exemplo, buscar previsões na área de vendas, marketing, instituições financeiras, universidades e principalmente na área de saúde. Com isso podemos utilizá-la também na área destes acidentes de trânsito. A tecnologia preditiva procura fornecer uma dinâmica mais eficiente de saber qual medida a ser tomada com menor impacto e obter um resultado satisfatório (Galvão; Marin, 2010).

Anualmente os órgãos fazem balanços sobre os acidentes que ocorrem no Brasil, neste trabalho será utilizado os dados fornecidos pela Polícia Rodoviária Federal (PRF, 2017) do ano de 2007 a 2015 para treinamento e do ano de 2017 para predição, com estes balanços podemos fazer uma análise detalhada e utilizar as ferramentas e técnicas utilizadas na predição de dados. Com resultados obtidos podemos prever locais com maior incidência, propor políticas de controle e tomar as medidas corretas para diminuição e prevenção destes acidentes. Devido ao aumento do índice de acidentes no Brasil e a divulgação pelos órgãos de fiscalização de registros dos acidentes, este artigo tem como motivação contribuir nas políticas de controle e prevenção dos acidentes de trânsito e diminuir a quantidade de vítimas envolvidas. Com isso acontece consequentemente a redução dos gastos governamentais com indenizações, podendo estes recursos financeiros serem investidos em outras áreas da sociedade.

2 METODOLOGIA E DESENVOLVIMENTO

2.1 ACIDENTES DE TRÂNSITO

A terminologia acidente de trânsito, pode-se tirar uma conclusão que é um fato acidental, ou não acidental, por envolver várias variáveis distintas que possa ocasionar o fato, dentre estes fatores podemos citar o homem, a rodovia, condições do veículo e várias outras variáveis, onde quando ocorre o fato tem-se vários danos que são os resultados que estes acidentes causam (Reis, 2014). Estes acidentes possuem uma frequência, de acordo com a modo de pensar das pessoas, mais para outro lado, de acordo com o modelo científico, o acidente de trânsito suporta a consideração de ser eventos arbitrários no tempo e no espaço. Com isso faz necessário um estudo minucioso afim de obter a dinâmica dos fatos, através de suas variáveis e características, com o intuito de descobrir as causas mais frequentes destes acidentes de trânsito e com isso adotar políticas de controle e prevenção (Oliveira et al., 2008).

Para se ter um conhecimento e um entendimento dos incidentes que possam levar a um acidente de trânsito, suas variáveis e seus padrões precisam ser estudados. Devido a isso ele reitera que, para o acontecimento de um acidente de trânsito é exigida um distúrbio na dinâmica de uma ou mais variáveis que fazem parte do sistema, no qual ele cita: o condutor, rodovia e condições dos veículos, onde são as principais causas que auxiliam para os acontecimentos dos acidentes de trânsito. Uma variável importante que ele demonstra é quanto a situação do movimento, onde determina que algum dos elementos esteja em deslocamento (Coelho, 1999).

2.2 FATORES CONTRIBUINTES

A incidência de ocorrer um acidente de trânsito quase sempre é causada pela negligência, imperícia ou imprudência humana. Segundo o autor é demonstrado ainda os principais resultados que favorecem ou são causa dos acidentes de trânsito, são eles: Resultados humanos; Resultados com relação as condições do veículo; Resultados de acordo com a rodovia/natureza e Resultados com relação a fatos institucionais ou sociais (Reis, 2014). As variáveis do sistema os condutores, a rodovia e as condições do veículo, demonstra que estes necessitam de um estudo bem elaborado e que cheguem a um entendimento quanto a dinâmica de acontecimentos destas variáveis juntas, para que seja produzido um conhecimento quanto as causas mais importantes de ocorrência dos acidentes de trânsito, e com isso adotar políticas eficientes de controle e prevenção (Coelho, 1999).

Segundo Sivak e Bao (2012), em uma pesquisa realizada nas cidades de Nova York e Los Angeles, foram encontrados padrões dos acidentes com relação ao restante do país e comparar os índices encontrados. Os resultados foram colocados em conjunto de acordo com as características do acidente, local do evento, as condições climáticas, a fase do dia, noturno ou diurno e características dos condutores. Os resultados obtidos, mostraram que as duas cidades pesquisadas possuem uma quantidade mais elevada de acidentes em comparação com o restante do país. Foram encontrados fatores em comum, que contribuíram para ocorrência dos acidentes. Um estudo desenvolvido pelo Instituto de Pesquisa Econômica Aplicada (IPEA, 2006), com o intuito de adotar políticas públicas de controle eficientes, programas de prevenção e coibir a evolução dos acidentes de trânsito, foi gerado resultados que auxiliam no entendimento do problema em questão. Dos incidentes, 60% ocorreram em rodovias federais no período diurno, em condições climáticas boas e o tipo de pista simples, 70% destes incidentes ocorrem em linha reta. A maior quantidade destes incidentes foi ocorrida nos finais de semana e nos meses de janeiro, dezembro e julho, todos meses que possuem um maior fluxo de veículos em circulação. Quanto a gravidade destes incidentes, foi destacada as colisões frontais (33 mortes a cada 100 incidentes) e atropelamentos (29 mortes a cada 100 incidentes). Entender sobre a ocorrências dos acidentes de trânsito é necessário um estudo das variáveis que possam levar ao acontecimento de um acidente de trânsito, levando em consideração o estudo das vias onde estes veículos circulam e os trechos que possuem maior incidência. Estes locais são denominados de locais ou segmentos críticos (Reis, 2014).

2.3 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS

O avanço tecnológico cada mais acelerado e com o crescimento exponencial do volume de dados produzida pelos sistemas, os modelos, as planilhas ou as técnicas que são utilizados convencionalmente para estudar os dados tem se tornado ineficientes, com isso tem-se a necessidade de estudar modelos mais eficientes para a descoberta de conhecimento importantes, pois os modelos convencionais tem a capacidade de gerar informação informativas, mais não geram conhecimentos. Esta necessidade busca encontrar métodos, técnicas, procedimentos para gerar informações inteligentes que possam contribuir os usuários nas tomadas de decisões e encontrar tendências em um volume de dados mais elevado. A Descoberta de Conhecimento em Banco de Dados é uma área de estudo científico em ascensão que vem da terminologia em inglês *KDD* (*Knowledge Discovery in Databases*) (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

Para Norton (1999), o processo de *KDD* é demonstrado com o objetivo de encontrar novos padrões, correlações e tendências com resultados satisfatórios através de um estudo detalhado em um largo volume de dados em conjunto armazenados. O processo de descoberta de conhecimento utiliza-se de métodos de encontrar padrões e tecnologias que utilizam modelos matemáticos e estatísticos. O *Data Mining* (Mineração de Dados) é um dos procedimentos mais utilizados no processo de KDD. No processo de *KDD* é necessário que três dispositivos estejam em sincronia para se obter um sucesso do processo, são os especialistas, analistas e os usuários. Os especialistas são os que possuem o conhecimento a ser desenvolvido, os analistas são os o que entendem do processo de *KDD* e os usuários serão os responsáveis por utilizar o conhecimento extraído do processo descoberta de conhecimento. Esta equipe e essencial para definir as estratégias que serão utilizadas no processo e a avaliação dos resultados gerados (Sousa, 2009). Segundo Fayyad; Piatetsky-Shapiro; Smyth (1996), o processo de *KDD* possui algumas etapas que devem ser seguidas para gerar a inteligência esperada, com isso é importante que respeite cada etapa para não gerar resultados negativos, pode-se observar na Figura 1, abaixo.

Figura 1 - Modelo das etapas de KDD

Fonte: (Fayyad; Piatetsky-Shapiro; Smyth, 1996)

Seguindo ainda o modelo acima de Fayyad; Piatetsky-Shapiro; Smyth (1996), estas etapas podem ser explicadas abaixo:

- Seleção: Nesta etapa, busca-se encontrar qual as necessidades dos usuários e implementar um controle dos conjuntos de dados que farão parte da aplicação, possuindo as variáveis e os relatórios que serão analisados;

- Dados alvo: Nesta etapa, é feito um filtro dos dados relevantes que serão utilizados no desenvolvimento do processo de *KDD*;
- Pré-processamento: Nesta fase, é feito o tratamento dos dados que foram selecionados, onde é feita a retirada dos dados que estão em duplicidade e a execução de funções e qual será a medida adotada com os campos que estão em branco ou nulos. Outras funções que são importantes e que tem impacto nos resultados são efetuadas nesta fase;
- Transformação: Nesta etapa, tem o objetivo de diminuir as variáveis geradas e busca propriedades que possam abaixar variações dos dados;
- Mineração de dados: Nesta etapa, dentro do processo de descobrimento de conhecimento é a mais valiosa pois é necessário identificar os objetivos relevantes para adotar a medida certa de mineração de dados. É selecionado nesta fase também o algoritmo ideal para a solução do objetivo no qual poderá ser utilizado de árvores de decisão, classificação, associação, cluster, entre outros;
- Interpretação/avaliação: Nesta fase, após os resultados serem encontrados busca-se interpretar os dados e identificar os padrões encontrados, com isso avaliar se os resultados gerados são satisfatórios, se não, pode se voltar as fases anteriores para adequação até que seja produzido os dados esperados.

2.4 MINERAÇÃO DE DADOS

O que se procura encontrar com o processo de mineração de dados é um conjunto de dados agrupados em uma determinada base de dados e que possuam um padrão de acontecimentos correspondente entre si, e que tenham sentido. Através das análises feitas no processo em dados passados espera-se que gerem previsões para tomadas de decisões futuras. As tendências encontradas no processo de mineração de dados podem ser incorporadas a vários processos decisórios, gerenciamento de informações e outras aplicações nas mais distintas áreas de negócio (Dias et al., 2001).

O processo de mineração de dados pode ser conceituado também da seguinte maneira, um processo automático de previsão de novas tendências ou informações, aplicado em um volume de dados elevado, sendo um procedimento indispensável no processo de produção de inteligência. Durante este processo de descoberta de conhecimento pode ser utilizado várias ferramentas, técnicas e algoritmos baseados em modelos matemáticos e estatísticos (Rud,

2001). O processo de mineração de dados pode ser definido através da remoção de métodos de associação, classificação, agrupamento, regressão, entre outros métodos. O processo de mineração de dados pode ainda ser utilizado sem ter o conhecimento dos métodos citados, pois podem gerar a descoberta de dados sem sentido. Devido a isso, a atividade de encontrar qual a metodologia correta para cada aplicação a ser desenvolvida e que gere resultados satisfatórios é indispensável e possui complexidades (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

Segundo Goebell; Gruenwald (1999), os mecanismos utilizados no processo de mineração dados para extração de conhecimento são utilizados em várias aplicações, e tem como aspiração obter informações que estão ocultas no grande volume de dados, com isso é utilizada várias técnicas de pesquisa de conhecimento, por exemplo, busca encontrar padrões, classificações e associações, dentro do volume de dados.

2.5 CLASSIFICAÇÃO DE DADOS

Para Fayyad; Piatetsky-Shapiro; Smyth (1996), o método de classificação é um procedimento de predição que se baseia no estudo de uma elevada massa de dados em busca encontrar padrões que expliquem tendências futurística desses dados. A metodologia de classificação é a técnica mais utilizada dentro do *Data Mining*, e se objetivo é encontrar propriedades, características comuns entre variáveis de uma base de dados determinada.

Alguns exemplos que o autor cita de classificação de dados são: classificar transações financeiras de créditos como risco baixo, médio ou alto; esclarecer solicitações de seguro podendo ser fraudulentas ou não, ver se determinado e-mail é spam ou não, entre outros (Bartolomeu et al., 2002). A classificação como um procedimento de buscar uma coleção de métodos que explicam e diferenciam classes, com o objetivo de utilizar o método final, para encontrar a classe de objetos que ainda não classificados. O método desenvolvido tem como teoria a análise antecipada de uma coleção de dados de amostra ou dados de treino, contendo características certamente classificadas. A técnica de classificação é baseada na predição de um valor conclusivo, como, por exemplo, encontrar a cobertura ou não de uma classe de defeitos (Weiss, 1998).

2.6 ÁRVORES DE DECISÃO

As Árvores de Decisão são modelos de classificação de dados no ramo da chamada Mineração de Dados (*Data Mining*). Podem ser utilizados juntamente com a tecnologia de indução de regras, mas são as únicas a apresentar os resultados de forma hierárquica. Na Árvore,

a variável maior relevância é apresentada na árvore como o primeiro nó, e as variáveis de menor relevância são mostradas nos nós subsequentes. A vantagem principal das Árvores de Decisão é a tomada de decisões levando em consideração as variáveis mais relevantes, além de uma melhor compreensão para a maioria das pessoas. Ao escolher e apresentar as variáveis em ordem de relevância, as Árvores de Decisão permitem aos usuários conhecer quais características mais influenciam os seus trabalhos (Lemos; Nievola; Steiner, 2005). As Árvores de Decisão são representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados. Árvores de Decisão consistem em nós, que demonstram as variáveis, e de ramos, provenientes desses nós e que recebem os valores possíveis para essas variáveis. Nas árvores existem nós folha, que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está relacionada a uma classe. Cada trajeto na árvore, da raiz à folha, corresponde a uma regra de classificação (Garcia, 2003). Segundo Lemos; Nievola; Steiner (2005), um exemplo de Árvore de Decisão, na qual constam informações que relatam as condições para uma pessoa receber um empréstimo. Nesse contexto existem duas possíveis possibilidades: SIM (receber empréstimo) e NÃO (não receber empréstimo). As Variáveis são montante, salário e conta. A variável MONTANTE pode assumir os valores de médio, alto ou baixo; a variável SALÁRIO pode assumir valores baixo ou Alto; e a variável CONTA pode ser sim ou não. Alguns dados são exemplos da possibilidade SIM, ou seja, os requisitos que são exigência de um banco a uma pessoa para ser concedido um empréstimo são satisfatoriamente preenchidos. Outros são da possibilidade NÃO, isto é, os requisitos que são exigência de um banco não são plenamente satisfeitos. A classificação, nesse contexto, e baseada em uma estrutura de árvore, que pode ser utilizada para todos os objetos do conjunto, pode-se observar na Figura 2, abaixo.

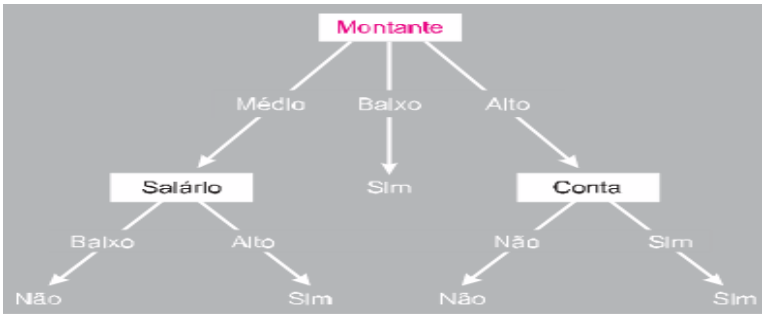

Figura 2 - Exemplo de uma árvore de decisão de empréstimo bancário

Fonte: (Lemos; Nievola; Steiner, 2005)

2.7 WEKA

A ferramenta *WEKA* possui uma variedade de algoritmos de regras de associação, classificação, clusterização, pré-processamento e de regressão, utilizados no processo de mineração de dados, todos os algoritmos foram desenvolvidos na linguagem *JAVA* e o *software* possui uma aparência bem simples e intuitiva. O *WEKA* consegue utilizar de dados que provém de banco de dados (*JDBC*), CSV ou também de arquivos de dados que a própria ferramenta utiliza, os dados na extensão *ARRF*. O *software* foi implementado na Universidade de *Waikato* na Nova Zelândia, e possui o código fonte aberto além de possuir uma facilidade de ser executada em diversas plataformas de Sistemas Operacionais. Pode ser caracterizada, através de um conjunto de algoritmos de classificação, regressão, associação, entre outros, que são utilizados como aprendizado de máquina no processo de mineração de dados. A ferramenta foi projetada através de uma iniciativa da Fundação de Pesquisa, Ciência e Tecnologia do governo da Nova Zelândia, e com os avanços tecnológicos cada mais acelerado e uma evolução na quantidade de dados fornecidos pelas mais diversas áreas, estes dados necessitam de um tratamento especial. O *WEKA* possui algoritmos para *Data Mining*, muito didáticos, sua utilização é parcialmente simples, devido a possuir recursos maleáveis para experiências e sendo estes recursos progressivamente atualizados com melhorias e inserção de novos modelos de algoritmos, sendo muito destes algoritmos possuírem definições na área de inteligência artificial (Markov, 2017).

2.7.1 OPÇÕES DE CLASSIFICAÇÃO

Após o tratamento da base, discretização e agrupamento dos dados que serão utilizados para treinamento e predição entra então na fase das opções de classificação de dados. Existe um painel chamado *Test options* (Opções de Teste), onde é possível escolher algumas configurações para o classificador que será utilizado. Estas opções determinam pontos importantes de como será o comportamento do algoritmo e de como a base de dados será testada. Existem quatro modos de teste, porém neste artigo será utilizado dois:

- *Use training set* (Use conjunto de treinamento): Nesta opção o classificador é avaliado em como ele consegue predizer a classe das ocorrências que ela foi treinada. Ou seja, o algoritmo tenta encontrar padrões entre os dados informados.
- *Supplied teste set* (Conjunto de testes fornecidos): O classificador é avaliado de quão bem ele consegue prever a classe em um conjunto de instâncias carregadas a partir de

um arquivo. Ao clicar no botão "*Set* ..." abrirá uma janela que lhe permite escolher o arquivo para testar.

Para classificação "*Use training set"* será utilizada a base do ano de 2007 a 2015, e para classificação de previsão "*Supplied teste set"* será utilizada a base do ano de 2017.

2.7.2 ATRIBUTO CLASSE

Os classificadores em *WEKA* foram desenvolvidos para serem treinados para prever uma única "classe" atributo, que é utilizada para a predição. Alguns classificadores só podem aprender classes nominais, outros só pode aprender classes numéricas. Outros ainda pode aprender ambos. Por padrão, o atributo classe é considerado como sendo o último atributo nos dados. Caso seja treinado um classificador para prever um atributo diferente, na caixa abaixo da caixa de opções de teste para abrir uma lista suspensa de atributos para escolher. No caso do problema proposto por este artigo, o atributo classe utilizado para classificação é justamente causa do acidente.

2.7.3 LISTAGEM DE RESULTADOS

O que será analisado nesta etapa de verificação a partir do problema de classificação dos dados de treinamento e predição será os seguintes resultados:

- *Correctly Classified Instances*. Demonstra a quantidade de instâncias que foram classificadas corretamente após o treinamento da árvore. Este valor depende diretamente da quantidade de instâncias utilizadas na base de treinamento.
- *Incorrectly Classified Instances*. Neste tópico são descritas as instâncias que foram classificadas incorretamente.

Pode ser visto também a árvore de decisão no campo "*Visualize tree or Visualize graph"*, gerada a partir da classificação dos dados que demonstra uma melhor visualização e entendimento do problema proposto neste artigo.

- *Visualize tree or Visualize graph*. É obtida uma demonstração gráfica da estrutura do modelo de classificação proposto. Na tela de visualização da árvore, é possível ver as instâncias de formação em cada nó. Teclando *CTRL* + clique amplia a visão da árvore para fora, enquanto *SHIFT* amplia a visão par dentro.

3 RESULTADOS

Os algoritmos *J48* e o *RandomTree* utilizados para classificação das bases, apresentaram resultados muito similares. O que pode ser observado é que o algoritmo *J48* é melhor em questão de performance em velocidade de execução, pois apresentou um tempo mais baixo durante a etapa de análise de dados, contudo, o algoritmo *RandomTree* apresentou resultados mais eficientes na questão de desempenho pois encontrou padrões com melhores resultados de classificação. Após execução dos algoritmos de classificação (*J48* e *RandomTree*), propostos neste artigo, os resultados obtidos nas bases de treinamento e predição foi o seguinte:

Tabela 2 – Resultados da base de treinamento (Ano 2007 a 2015) – (585.257 Instâncias)

Algoritmo	*Correctly Classified Instances*	*Incorrectly Classified Instances*
J48	67,0032%	32,9968%
RandomTree	67,104%	32,896%

Fonte: Autoria própria

Tabela 3 – Resultados da base de predição (Ano 2017) – (40.448 Instâncias)

Algoritmo	*Correctly Classified Instances*	*Incorrectly Classified Instances*
J48	82,229%	17,771%
RandomTree	82,5381%	17,4619%

Fonte: Autoria própria

Devido a classificação correta obter uma taxa de 82%, o que deve ser analisado quanto a ocorrência dos acidentes e os resultados encontrados na classificação das bases, deve ser os CONDUTORES, por apresentar um elevado índice de ocorrência de eventos. Com isso observa-se que independentemente da pista e região e as outras classes infelizmente o erro ainda é humano. Políticas de conscientização devem ser feitas com mais frequência e com uma cobrança mais rigorosa dos condutores.

4 CONCLUSÃO

Ao término deste é possível notar o quanto a ciência de dados está contribuindo aos mais diversos setores do país, e com o auxílio da computação é possível obter resultados mais

rápidos, satisfatórios e eficientes. Problemas acontecem com um volume muito grande destes dados que para ser feita uma análise manual levaria muito tempo, recursos humanos e dificuldades para se encontrar os padrões para identificar as possíveis correlações entre estes dados. Para tratar a base utilizada neste artigo foi aplicado algoritmos de classificação utilizando a ferramenta *WEKA*, que gera de forma automática as correlações destes dados de maneira rápida e eficiente e com uma taxa de acerto satisfatória, com isso utilizando o conhecimento gerado pelas árvores de decisão pode-se desenvolver e adotar políticas para prevenção de acidentes nas vias. Com o avanço tecnológico cada vez mais acelerado e com o surgimento de novas tendências nas mais diversas áreas da sociedade, várias possibilidades surgem com pesquisas voltadas para área da tecnologia preditiva. Com os resultados encontrados será divulgado para os órgãos controladores para que seja adotada medidas de segurança e controle para redução dos índices de incidente. Enfim, pode-se concluir que o uso da computação utilizando tecnologia preditiva e dados pouco inexplorados, contribui com políticas que podem ser adotadas para redução de acidentes de trânsito no país, consequentemente reduzindo recursos financeiros, danos físicos, materiais e gerando mais segurança para quem utiliza as estradas do nosso país.

5 REFERÊNCIAS

Bartolomeu, Tereza Angélica et al. Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento. 2002.

Berthold, Michael R.; Hand, David J. (Ed.). Intelligent data analysis: an introduction. Springer, 2007.

Brasil, I. P. E. A. Instituto de Pesquisa Econômica Aplicada (IPEA). 2006.

Cidades, I., 2015. Instituto brasileiro de geografia e estatística. Frota de veículos, 2016.

Coelho, Helena da Silva. Análise da Influência das Características Física-Operacionais das Vias na Ocorrência de Acidentes de Trânsito nas Rodovias Federais. 1999. Tese de Doutorado. Dissertação (Mestrado em Transportes) – Departamento de Engenharia Civil. Brasília: Universidade de Brasília.

De Oliveira, Nelson Luiz Batista; de Sousa, Regina Marcia Cardoso. Fatores associados ao óbito de motociclistas nas ocorrências de trânsito. Revista da Escola de Enfermagem da USP, v. 46, n. 6, p. 1379-1386, 2012.

Dias, Maria Madalena et al. Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados. 2001.

DNIT, M. T. Manual de estudos de tráfego. 2010.

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37, 1996.

Federal, Polícia Rodoviária, 2017. Ministério da defesa. Ministério da Industria e Comércio.

Galvão, Noemi Drever; de Fátima Marin, Heimar. Características das vítimas de acidente de trânsito por meio da técnica da mineração de dados. Journal of Health Informatics, v. 2, n. 4, 2010.

Garcia, Simone C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. 2003.

Goebel, Michael; Gruenwald, Le. A survey of data mining and knowledge discovery software tools. ACM SIGKDD explorations newsletter, v. 1, n. 1, p. 20-33, 1999.

Gomes, Dennis dos Santos. Inteligência Artificial: Conceitos e Aplicações. Olhar Científico, v. 1, n. 2, p. 234-246, 2011.

Han, Jiawei; Pei, Jian; Kamber, Micheline. Data mining: concepts and techniques. Elsevier, 2011.

Norton, Melanie J. Knowledge discovery in databases. Library Trends, v. 48, n. 1, p. 9, 1999.

Oliveira, Marcos Pimentel de et al. O impacto da utilizacão de medidores eletrônicos de velocidade na redução de acidentes de trânsito em área urbana. 2008.

Prezepiorski Lemos, Eliane; Arns Steiner, Maria Teresinha; Nievola, Julio César. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. Revista de Administração-RAUSP, v. 40, n. 3, 2005.

Reis, Cristian Virgílio Roque. O uso da descoberta de conhecimento em Banco de Dados nos acidentes da BR-381. Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento, v. 3, n. 1, 2014.

Rud, Olivia Parr. Data mining cookbook: modeling data for marketing, risk, and customer relationship management. John Wiley & Sons, 2001.

Russell, Ingrid; Markov, Zdravko. An Introduction to the Weka Data Mining System. In: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education. ACM, 2017. p. 742-742.

Russell, Stuart; Norvig, Peter. Inteligência Artificial: tradução da segunda edição. Elsevier, 2004.

Sivak, Michael; Bao, Shan. Road safety in New York and Los Angeles: US megacities compared with the nation. 2012.

Social, Previdência. Ministério da Previdência Social. v. 3, 2013.

Sousa, Ricardo Miguel Oliveira Pires de. Extracção de regras de associação com itens raros e frequentes. 2009. Tese de Doutorado. Instituto Politécnico do Porto. Instituto Superior de Engenharia do Porto.

Weiss, Sholom M.; Indurkhya, Nitin. Predictive data mining: a practical guide. Morgan Kaufmann, 1998.

World Health Organization. Global status report on road safety 2015. World Health Organization, 2015.