Mineração de Dados na Construção de Modelo de Predição de Acidentes com Vítimas em Recife

Manuscript template for the REPA Journal

Adriano de Melo Costa [1] orcid.org/0000-0003-2964-1779

Arthur Guilherme Oliveira de Freitas [1] orcid.org/0000-0002-6623-2280

Ricardo Paranhos Pinheiro [1] orcid.org/0000-0003-4131-7744

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil.

E-mail do autor principal: Ricardo Paranhos Pinheiro rpp3@ecomp.poli.br

Resumo

Um dos problemas mais relevantes na cidade do Recife é o seu trânsito. Aqui pretende-se auxiliar na atenuação desta questão, atuando na previsão dos acidentes com vítimas. Foram usadas bases de trânsito disponibilizadas pela autarquia de trânsito e transporte urbano do Recife - CTTU no portal de dados abertos da prefeitura de Recife. Estas bases compreendem os acidentes ocorridos entre junho de 2015 e fevereiro de 2020. Então, estas informações passaram por um processo de mineração de dados. Daí, criou-se um modelo de predição dos acidentes de trânsito com vítima na cidade do Recife nos doze últimos meses avaliados, por meio do uso de técnicas de aprendizagem de máquina. O combate à pandemia de COVID-19 ocasionou restrições de movimentação que mudaram o perfil do trânsito esperado na cidade a partir de março de 2020, causa da exclusão do período neste modelo. Foram propostos quatro modelos de predição e, no de melhor acurácia, a taxa de erro média foi de 13 acidentes/mês.

Palavras-Chave: Mineração de Dados; Trânsito; Aprendizagem de Máquina; Predição.

Abstract

One of the prominent urban questions in Recife is traffic. Here we intend to aid mitigating this issue, predicting accidents with victims. To achieve this goal, data from CTTU, available in Portal de Dados Abertos da Prefeitura do Recife were used. These data cover the accidents that happened between June 2015 and February 2020. So, this information went through a data mining process. Hence, a machine learning prediction model for traffic accidents with victims in last 12 months in Recife was created. Combating the COVID-19 caused movement restrictions that changed the expected traffic profile in the city starting in March 2020, reason for the exclusion of the period in this model. We proposed four prediction models and, for the model with the best accuracy, the average error rate was 13 accidents per month.

Key-words: *Data Mining; Traffic; Machine Learning; Prediction.*

DOI: 10.25286/repa.v6i3.1707

1 Introdução

1.1 Contextualização

Um estudo realizado em 2019 pelo *Traffic Index Ranking* **[1]** coloca o trânsito da cidade do Recife como o 15º pior do mundo, e pior no Brasil. Por meio da mineração dos dados de acidentes ocorridos na cidade, pretende-se chegar a uma solução que contribua na criação de políticas públicas, a fim de auxiliar na melhoria deste quadro.

1.2 Descrição do Problema

A prefeitura da cidade do Recife, por meio de seu portal de dados abertos **[2]**, disponibiliza diversas bases relacionadas aos acidentes de trânsito, mas estas encontram-se desconexas entre si. Caso estas informações se combinem adequadamente, podem servir de insumo aos órgãos competentes na manutenção de providências benéficas ao trânsito da cidade, como também na substituição daquelas que se mostrarem inadequadas.

1.3 Objetivo

Este trabalho tem como objetivo o desenvolvimento de um modelo de predição da taxa de acidentes de trânsito com vítimas na cidade do Recife. A utilização deste modelo pode desempenhar um papel importante em diversos aspectos do desenvolvimento social da cidade, que vão desde o replanejamento urbano, com a criação de estratégias que ajudem a mitigar os crescentes números de acidentes de trânsito com vítimas, passando pela criação de políticas públicas que tenham por finalidade deixar o trânsito da cidade mais seguro, e até mesmo tentando abrandar os prejuízos econômicos causados pelo afastamento do trabalho dos envolvidos em recuperação, e outras questões previdenciárias decorrentes de um acidente de trânsito com vítimas.

1.4 Justificativa

Por meio da criação destas políticas, pretende-se alcançar objetivos concretos tanto na diminuição dos acidentes em geral, que contribuem com o posicionamento do recife no ranking dos piores trânsitos do mundo; como também nos acidentes com vítimas, que causam toda gama de problemas, que vão desde questões econômicas, por causa do tempo de convalescência dos envolvidos afastados do trabalho, até catástrofes irreversíveis, com danos irreparáveis à saúde dos acidentados, levando no extremo à morte das pessoas.

1.5 Escopo Negativo

Não está entre os objetivos desta proposta a análise dos dados de infrações de trânsito, apenas dos acidentes com vítimas ocorridos entre junho de 2015 até fevereiro de 2020, excluindo-se então os meses de trânsito atípico observados a partir de março de 2020, por causa das restrições de mobilidade causadas ela COVID-19. Tampouco pretende-se fazer o *deploy* de uma solução em produção das conclusões aqui observadas. Esta realização fugiria tanto do escopo do projeto quanto do prazo disponível para a realização desta atividade. Também não será feita, neste momento, nenhuma distinção entre a quantidade de vítimas nos acidentes, os tipos ou localizações das ocorrências. A predição será limitada a um mês a frente dos dados observados, e não serão propostos novos modelos matemáticos estatísticos para tratamento ou modelagem dos dados, serão utilizados métodos já amplamente divulgados na literatura.

2 Fundamentação Teórica

2.1 Área do Negócio

A criação dos veículos foi uma das maiores conquistas do ser humano moderno e trouxe grande revolução no cenário mundial. Com o passar do tempo, a fabricação de novos veículos cresceu absurdamente. Enquanto esse crescimento seguia disparado, como consequência também foi crescendo a quantidade de incidentes no trânsito, principalmente nos grandes centros urbanos.

Segundo a Organização Pan-Americana de Saúde (OPAS) e a Organização Mundial de Saúde (OMS) **[3]**, cerca de 1,35 milhão de pessoas, por ano, perdem a vida em decorrência de acidentes de trânsito, sendo esta a principal causa de morte entre crianças e jovens de 5 a 29 anos.

Ainda segundo a OPAS/OMS, os acidentes de trânsito custam à maioria dos países 3% de seu produto interno bruto (PIB). Esses custos vão desde os valores gastos com tratamento das lesões e reabilitações, bem como com as investigações dos acidentes por parte das autoridades, até a redução e perda de produtividade dos indivíduos na vida profissional e econômica.

As formas de redução desse quadro são temas de debates e estudos nas mais diversas áreas de conhecimento. Pelo fato de os acidentes serem eventos imprevisíveis com circunstâncias aleatórias e

de natureza e causa relacionados a diversos fatores, torna-se muito complexo o estudo e a elaboração de modelos com medidas realmente eficazes para redução dos acidentes e dos seus danos (Brandão, 2006) **[4]**.

Vários órgãos costumam disponibilizar relatórios e dados abertos sobre os acidentes, sejam eles a nível municipal, estadual ou federal. Em alguns países esses dados são extraídos diretamente de sistemas de registros policiais, como é o caso do Portal *Town of Cary* **[5]**, um portal de dados abertos da cidade de *Cary*, na Carolina do Norte, nos Estados Unidos da América. É graças a disponibilização desses tipos de dados que é possível realizar estudos e trabalhos a fim de apoiar decisões governamentais com medidas eficazes capaz de reduzir os acidentes e seus impactos.

2.2 Mineração de Dados

Um dos maiores desafios das empresas é o de tentar expandir, além de manter, grande parte de seus clientes. Com o competitivo mercado global de hoje, possuir uma boa organização da produção, redução de custos, bons atendimentos e excelência na qualidade de produtos e serviços, além de muitas outras características, passaram a ser insuficientes para vencer e se destacar dentre a disputada concorrência. Grande parte das organizações hoje enxergam que é preciso entender melhor seu cliente, traçando perfis, interpretando seus objetivos, expectativas e desejos. Muitas vezes esses conhecimentos estão em meio à grande massa bruta de dados armazenados. É nesse processo de extração de informações, na maioria das vezes de muita valia, que surge a necessidade de mineração de dados (ou *data mining*, em inglês).

Segundo Braga (2005) **[6]**, "a mineração de dados compreende um conjunto de técnicas para descrição e predição a partir de grandes massas de dados, provendo um método automático para descobrir padrões em dados, sem a tendenciosidade e a limitação de uma análise baseada meramente na intuição humana". Para Castro e Ferrari (2016) **[7]**, a mineração de dados corresponde à aplicação de algoritmos capazes de extrair conhecimentos a partir dos dados pré-processados.

Esse processo de mineração pode ser feito tanto de forma descritiva (também chamada de exploratória), quando são utilizadas métricas e técnicas estatísticas para entender e explicar como os dados são classificados e/ou agrupados, bem como de forma preditiva, quando se há a intenção de apontar como serão os dados no futuro e como se comportam dadas diversas condições. Para esses modelos preditivos são usadas, além de modelos estatísticos avançados, modelos de *machine learning*, inteligência artificial, algoritmos complexos, dentre outras técnicas.

"A mineração de dados está inserida em um processo maior denominado descoberta de conhecimento em banco de dados, Knowledge Discovery in Database (KDD). Rigorosamente o DM se restringe à obtenção de modelos, ficando as etapas anteriores e o próprio DM como instâncias do KDD." (BRAGA, 2005, p. 15) **[6]**.

2.3 Trabalhos Relacionados

A seguir serão descritos seis trabalhos que possuem semelhanças temáticas com o aqui proposto, seja na metodologia similar que utilizam, seja pelo problema em comum.

Em Nonparametric Regression for the Short-term Traffic Flow Forecasting **[8]**, os autores usam como base de treinamento 12 meses do trânsito da cidade de Duluth, nos Estados Unidos, entre os meses de janeiro e dezembro de 2006, a fim de preverem o fluxo de tráfego de curto prazo, no caso o mês seguinte. Com o uso de knn baseado em regressão não paramétrica, conseguem um resultado melhor que com o uso de redes neurais. Em A Time Series Model for Assessing the Trend and Forecasting the Road Traffic Accident Mortality **[9]**, os autores usam como base de treinamento 7 anos de trânsito da cidade de Zanjan, no Irã, entre 2007 e 2013, a fim de preverem os acidentes dos próximos 4 anos. Para isso usam series temporais, e o melhor resultado é alcançado por meio do uso do modelo SARIMA. Estes dois trabalhos possuem abordagem semelhante ao usado aqui.

Já em A Comparison of Multivariate SARIMA and SVM Models for Emergency Department Admission Prediction **[10]**, os autores usam dados de janeiro de 2009 a agosto de 2012 para prever os acessos a um hospital de Madri, Espanha, em setembro de 2012. Utilizaram SARIMA e SVM, e o segundo apresentou resultados superiores. Em the prediction of traffic flow with regression analysis **[11]**, os autores usam dados de tráfego da cidade do Porto, Portugal, entre os anos de 2013 a 2015, sempre usando as três últimas semanas como treino, a fim de prever os dados da próxima. Foram usados cinco modelos, e o que apresentou melhor resultado foi árvore de regressão. Estes dois trabalhos, apesar de não tratarem

DOI: 10.25286/repa.v6i3.1707

especificamente de acidentes de trânsito, usam técnicas e metodologias semelhantes às observadas neste artigo.

Os próximos dois trabalhos também tratam de acidentes de trânsito, e usam técnicas que não se distanciam muito das vistas aqui, mas os objetivos diferem dos quatro anteriores, e deste. Em Early Warning of Traffic Accident in Shanghai Based on Large Data set Mining **[12]**, são usados dados de trânsito da cidade de Xangai, China, entre julho de 2014 e abril de 2015, como entradas de algoritmos de classificação e análise de regressão. De acordo com os dados observados, foram criados níveis de segurança e oferecidas medidas a fim de diminuir os acidentes. E finalmente, em Data mining of tree-based models to analyze freeway accident frequency **[13]**, os autores têm como objetivo determinar os motivos dos acidentes, e para tal analisam os dados da National Freeway 1, em Taiwan, entre 2001 e 2002. Descobriram que o volume médio do tráfego e variáveis relacionadas a chuva foram determinantes na frequência de acidentes na estrada.

2.4 Técnicas Utilizadas

Regressão Linear: A análise da regressão consiste na realização de uma análise estatística com o objetivo de verificar a existência de uma relação entre uma variável dependente com uma ou mais variáveis independentes **[17]**. Esta relação é representada por meio do uso de um modelo matemático, que associa estas variáveis. A regressão linear simples pode ser definida como uma relação linear entre a variável dependente e uma variável independente. Se forem usadas múltiplas variáveis independentes, caracteriza-se uma regressão linear múltipla.

MLP: Um perceptron pode ser definido como uma rede neural composta por uma camada de entrada onde cada neurônio representa uma variável considerada como entrada para o problema, sendo então as informações que alimentam a rede; uma função que pondera estas entradas através de pesos sinápticos, e uma função de saída que fornece o sinal emitido pelos neurônios da camada de saída. Este tipo de rede neural não é capaz de resolver problemas reais, pois classifica apenas padrões linearmente separáveis **[18]**.

A fim de resolver esta limitação, foi implementada a rede perceptron de múltiplas camadas, ou multi-layer perceptron (MLP), que se diferencia da rede perceptron tradicional pela adição de pelo menos uma camada de neurônios intermediária, que possuem funções de ativação.

Algumas das principais preocupações existentes ao se desenvolver redes MLP estão relacionadas com a definição da quantidade ideal de neurônios na camada escondida, e com a definição do critério de parada que será usado na fase de treinamento. A preocupação com estes fatores está ligada com a capacidade de generalização das redes: se for usada uma quantidade muito grande de neurônios na camada escondida, a rede resultante pode apenas memorizar os dados de treinamento, piorando sua capacidade de generalização. E se a rede for treinada objetivando-se apenas a diminuição no erro de treinamento, pode-se chegar no mesmo resultado indesejado da baixa capacidade de generalização, pela ocorrência do superajustamento, ou overfitting.

SVM: As máquinas de vetores de suporte, ou support vector machines (SVM) são redes neurais artificiais que foram desenvolvidas tendo por base a teoria do aprendizado estatístico. O desenvolvimento destas redes teve como objetivo a obtenção de redes neurais com alta capacidade de generalização, uma vez que durante o treinamento supervisionado destas redes se busca não só minimizar o erro de treinamento, como também a complexidade da rede obtida. Foram propostas inicialmente por Vapnik em 1971 **[19]**, e detalhadas pelo mesmo pesquisador em 1995 e 1998.

Random Forest: Árvores de decisão são modelos estatísticos que também usam treinamento supervisionado, com entradas e saídas, na classificação e previsão de dados. seu modus operandi envolve a divisão recursiva do problema em subproblemas mais simples. Pode ser representada como uma sequência de nós, onde cada nó contém um teste para algum atributo, e cada ramo descendente corresponde a um possível valor deste atributo **[20]**. Assim como as técnicas observadas anteriormente, pode sofrer de overfitting, não generalizando bem, caso o número de nós seja muito alto ou muito baixo.

Random Forest, ou floresta aleatória, é um método de aprendizado conjunto, onde são treinadas várias árvores de decisão não relacionadas a partir de amostras dos dados, a fim de se resolver problemas de classificação, por meio do agrupamento de resultados semelhantes, ou problemas de predição, utilizando a média dos valores obtidos. São usadas técnicas estatísticas para que a partir de amostras da base a ser analisada, sejam recriadas cópias únicas da população, mas com compatíveis com as da amostra. Em seguida são selecionadas características

aleatórias das árvores, a fim de garantir uma baixa correlação entre as árvores. Uma boa prática para aumentar a capacidade de generalização envolve a utilização de árvores com alturas intermediárias, e a combinação de modelos, sendo estas possibilidades de tratamento do superajustamento uma das razões de dar preferência ao uso da técnica da floresta aleatória e não de apenas uma árvore de decisão **[21]**.

Os experimentos foram realizados por meio do uso da ferramenta de data mining Orange **[22]**, um projeto open source desenvolvido em Python pelo Bioinformatics Lab da universidade de Ljubljana, na Slovenia.

3 Materiais e Métodos

3.1 Descrição da Base de Dados

Como resultado do mapeamento dos dados obtida do Portal de Dados Abertos da Prefeitura de Recife, foram elencados um conjunto de bases para comporem o Datalake do projeto.

Os dados foram exportados do portal e importados para um banco de dados SQL SERVER, em um sistema operacional Microsoft Windows 10, por meio dos softwares de ETL Pentaho Data Integration e FME. A análise dos dados foi feita com o uso do Orange, e do Microsoft Excel. Durante o processo de carga de dados, foi realizada a seleção dos campos que são de interesse para o projeto: os acidentes registrados pela prefeitura desde junho de 2015 até fevereiro de 2020.

A base possui 56.377 registros, e desta foram selecionados os registros que continham acidentes com vítimas, e excluídos os acidentes que aconteceram a partir de março de 2020, por causa da mudança na característica do trânsito ocasionada pelo COVID-19.

São descritos a seguir os principais atributos das tabelas:

- Tipo: descrição do acidente em texto livre.
- data: data da ocorrência do acidente.
- Hora: horário no qual o acidente ocorreu.
- bairro: nome do bairro onde o acidente ocorreu em texto livre.
- endereco: nome do logradouro onde o acidente ocorreu em texto livre.
- auto: quantidade de automóveis de pequeno e médio porte envolvidos.
- moto: quantidade de motocicletas envolvidas no acidente.
- vítimas: quantidade de vítimas envolvidas no acidente.
- Vítimas fatais: quantidade de vítimas fatais envolvidas no acidente.

3.2 Análise Descritiva dos Dados

Uma análise descritiva ajuda a ter um melhor entendimento dos fatos descritos pelos dados. Usando simples consultas, agrupamentos e ordenações foi possível elaborar gráficos e extrair algumas informações que contribuíram para as próximas etapas do processo, como por exemplo as médias anuais e mensais de acidentes, bem como a evolução dos acidentes com vítimas.

O Gráfico 1 ilustra a quantidade de acidentes por ano dentro do intervalo analisado neste trabalho. Ao observar o Gráfico 1 percebe-se uma maior divergência nos anos de 2015 e 2020. Essa baixa quantidade é devido ao ano não estar "completo" com todos os meses, pois, como citado anteriormente, o intervalo analisado se inicia no mês de junho de 2015 indo até março de 2020. Portanto, para o ano de 2015 estão ausentes dados dos meses de janeiro à maio, e de março à dezembro para o ano de 2020.

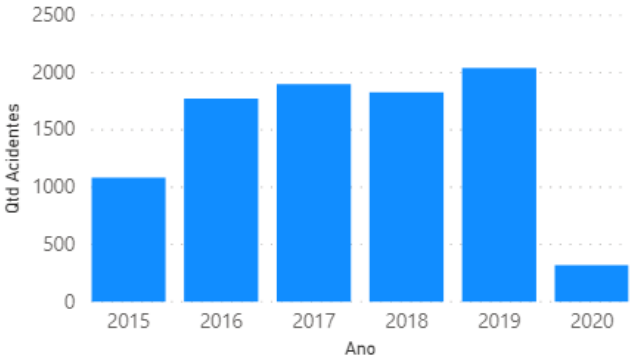

Gráfico 1: Quantidade total de acidentes com vítimas por ano. Fonte: Própria, 2020.

Detalhando um pouco mais, o Gráfico 2 ilustra os acidentes distribuídos mensalmente. Desta forma, é possível ter uma noção de quais os meses que possuem as maiores e as menores médias de acidentes.

De forma ainda mais detalhada e alinhada com o objetivo deste trabalho, o Gráfico 3 ilustra a evolução da quantidade de acidentes com vítimas no decorrer dos anos. Percebe-se que não há indícios de

DOI: 10.25286/repa.v6i3.1707

estabilidade e a quantidade segue uma forte variação, tanto para mais como para menos.

Com o Gráfico 4 é possível visualizar a média e a mediana da quantidade de acidentes com vítimas por mês através do Violin plot.

Utilizando as coordenadas dos acidentes, foi possível realizar uma plotagem no mapa e visualizar um mapa de calor, conforme é ilustrado no Gráfico 4. Com essa imagem é possível ter uma maior noção das regiões/bairros com maior concentração dos acidentes.

As informações obtidas até aqui e ilustradas nos gráficos indicam uma grande variação na quantidade de ocorrências ao longo dos anos e meses. A fim de descobrir se houve quedas ou aumentos significativos após alguns marcos relevantes relacionados ao trânsito, foi montado o Gráfico 3, que ilustra o comportamento dos acidentes juntamente com os pontos dos momentos dos marcos.

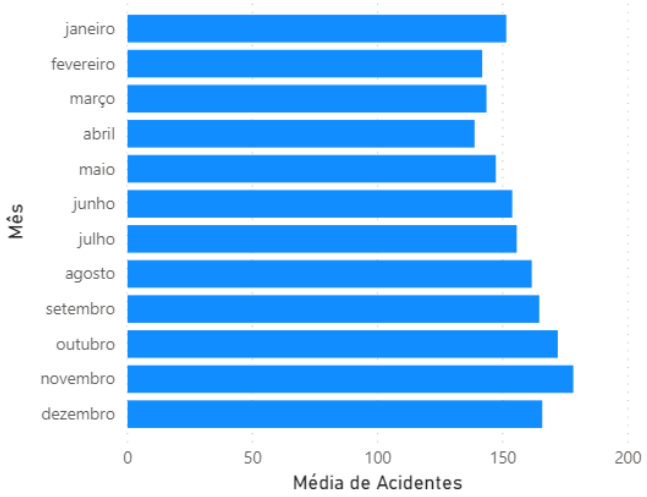

Gráfico 2: Média da quantidade de acidentes com vítimas por mês. Fonte: Própria, 2020.

Entre o intervalo de datas desse estudo, alguns acontecimentos relevantes que influenciaram diretamente no trânsito foram a chegada do aplicativo Uber e os aumentos das penas da Lei Seca e da prisão de condutores embriagados. Com exceção do primeiro, que fez com que houvesse um aumento na quantidade de acidentes, os outros dois marcos, segundo os dados, não trouxeram significativas reduções. Pode-se supor, por exemplo, que o aumento de acidentes após a chegada do Uber está diretamente ligado ao aumento do fluxo de veículos nas ruas.

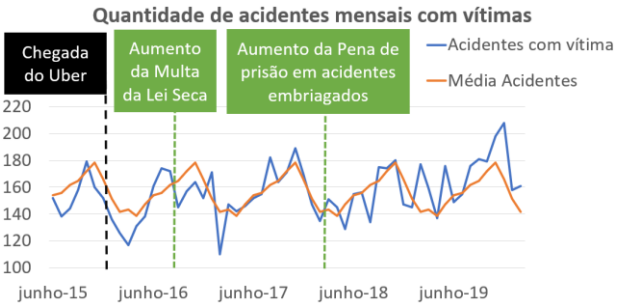

Gráfico 3: Distribuição dos acidentes, média mensal e marcos no trânsito. Fonte: Própria, 2020.

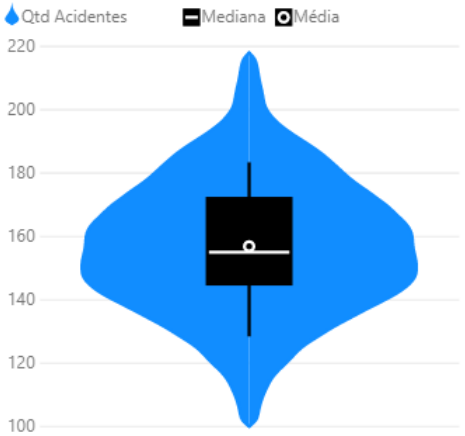

Gráfico 4: Violin Plot da quantidade de acidentes com vítimas por mês. Fonte: Própria, 2020.

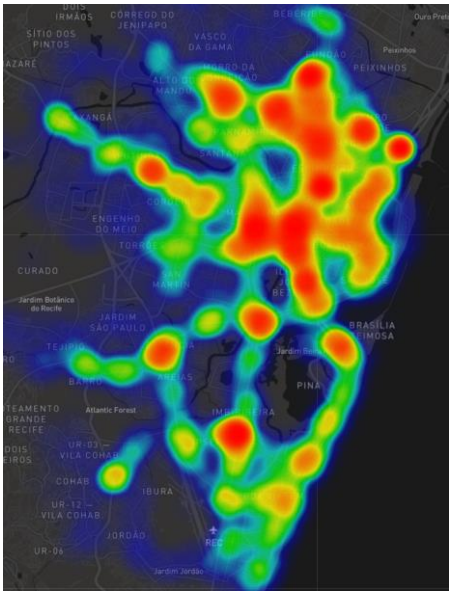

Gráfico 5: Mapa de calor da quantidade de acidentes com vítimas. Fonte: Própria, 2020.

3.3 Pré-processamento dos Dados

3.3.1 Integração dos dados

O primeiro passo foi agrupar os dados em uma só tabela do banco de dados. Eles são originalmente

disponibilizados separados, em arquivos divididos por ano. Entre esses arquivos foram encontradas divergências de quantidade de colunas e de tipos de dados diferentes: um exemplo disso é uma coluna ter sido encontrada com dados numéricos em um conjunto de um determinado ano, e em outro, essa mesma coluna apresentava os dados de forma textual. Além disso, houve também divergências nos formatos de datas, variando, como por exemplo, entre os formatos "dd/mm/aaaa" e "aaaa-mm-dd".

3.3.2 Limpeza dos dados

A limpeza dos dados envolveu o tratamento dos problemas encontrados para a integração. As datas foram padronizadas para um mesmo formato; os campos numéricos que estavam como texto foram convertidos; o tamanho dos campos foi limitado e ajustados a sua codificação para UTF-8; os campos vazios foram marcados como nulos.

Durante a primeira etapa da análise descritiva foi identificada uma possível anomalia nos registros de 2016. O caso foi apresentado ao stakeholder e ele prontamente disponibilizou um reprocessamento dos dados de 2015 a 2018.

3.3.3 Transformação dos dados

Até então, a base era composta pelos dados analíticos dos acidentes. A fim de delimitar o escopo, foi aplicado um filtro das ocorrências com vítimas e a totalização de ocorrências por mês e ano. Com isso, foram criados 5 atributos adicionais: Número do mês no qual a taxa foi medida, taxas de acidentes para cada um dos últimos 4 meses imediatamente anteriores.

3.3.4 Redução dos dados

Foram removidos os dados dos meses de março a maio de 2020 por serem outliers influenciados pelas medidas de distanciamento social impostas durante a Pandemia de COVID-19.

3.4 Metodologia experimental

A atividade de mineração de dados engloba uma ampla gama de ações, gerando a possibilidade de o pesquisador perder-se diante da tarefa de processar e extrair conhecimento de grandes volumes de dados. Demandando assim uma abordagem estruturada para a execução eficiente do processo de mineração. Segundo WIRTH **[14]**, "o modelo de referência **CRISP-DM** para mineração de dados fornece uma visão geral do ciclo de vida de um projeto de mineração de dados. Ele contém as fases de um projeto, suas respectivas tarefas e seus resultados".

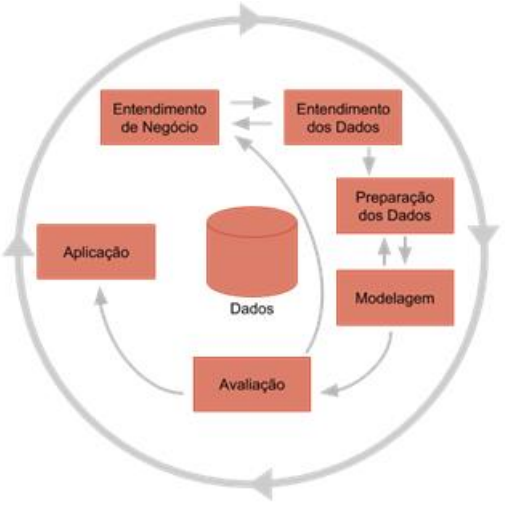

Figura 1: Fases do modelo CRISP-DM para mineração de dados. Fonte: Wirth, 2000.

O CRISP-DM segue uma abordagem iterativa, onde a cada ciclo os resultados de cada etapa são avaliados e fornecem insights para uma nova execução do processo. Observa-se na Figura 1 uma representação resumida do processo.

Neste trabalho foram seguidos os passos do processo CRISP-DM como forma de obter uma melhor gestão e controle do processo de mineração.

Após um primeiro ciclo onde foi realizada a análise descritiva, preparou-se um conjunto de dados para servir de entrada para a construção do modelo de predição. O conjunto de dados que inicialmente era formado pelo registro analítico das ocorrências foi totalizado por mês. De forma a obter uma granularidade que compensasse a aleatoriedade natural dos acidentes de trânsito e viabiliza-se a uma predição mais acurada. A tabela de entrada para o processo foi definida conforme segue:

- [DH_REF]: meta dado que descreve no formato de data o mês e o ano, foi utilizado apenas para ordenação das instâncias.
- [NU_MES]: número do mês.
- [QTD_ACIDENTES_MES-0]: quantidade de acidentes com vítimas ocorridos no mês. Esse é o atributo alvo do modelo.
- [QTD_ACIDENTES_MES-N] (com N de 1 a 4): quantidade de acidentes com vítimas

DOI: 10.25286/repa.v6i3.1707

ocorridos em cada um dos últimos 4 meses imediatamente anteriores.

Para os experimentos foi utilizado o programa Orange na versão 3.26.0 instalado a partir da plataforma Anaconda 3, versão 2020.07. O ambiente operacional foi um computador Intel Core i7 com S.O. Windows 10.

O programa Orange foi escolhido devido a sua interface intuitiva e versátil para a modelagem de processos de machine learning.

O experimento no Orange foi configurado da seguinte forma:

Os dados foram ordenados pelo atributo DH_REF;

1. O atributo [QTD_ACIDENTES_MES-0] foi definido como atributo alvo.
2. Os atributos de entrada [NU_MES] e [QTD_ACIDENTES_MES-N] (sendo N de 1 a 4) tiveram seus valores normalizados para o intervalo 0-1. Após isso foi aplicado o método PCA, do qual foram extraídos os 5 componentes criados (ou seja, sem perda de informação).
3. Foram criados 12 experimentos. Para cada um foi tomado um intervalo contínuo de meses e o último mês foi reservado para teste. Ou seja, o primeiro experimento foi treinado com o período de junho de 2015 a fevereiro de 2019 e testado com o mês de março de 2019. Essa modelagem visa aproximar o teste de uma aplicação no mundo real, onde seriam usados todo o histórico disponível com obtivo de predizer o mês seguinte.
4. Ao final do passo anterior, os 12 resultados de cada experimento foram agrupados e as métricas de erro foram calculadas e comparadas.

Para cada algoritmo escolhido, foram feitos alguns testes amostrais onde seus parâmetros de ajuste foram configurados. O experimento descrito acima foi feito com aplicação dos parâmetros que obtiveram o melhor desempenho nestes testes amostrais.

Para a regressão linear, o Orange disponibiliza opções de regularização, com a finalidade de diminuir o overfitting, mas dado que o cenário proposto analisa a relação entre apenas duas variáveis, não foram encontradas diferenças relevantes nos resultados a ponto de justificar a configuração de tais parâmetros, então eles permaneceram em suas configurações padrão.

Quanto às configurações disponíveis para a rede neural MLP, os parâmetros que apresentaram melhores resultados foram 100 neurônios em uma camada escondida, cálculo da tangente hiperbólica como função de ativação da camada escondida, gradiente-descendente estocástico como solucionador para otimização dos pesos, a regularização para tratamento do overfitting foi deixada no valor padrão, e o número máximo de iterações foi definido em 200.

Já os parâmetros disponíveis na configuração da rede neural SVM envolvem a seleção entre o algoritmo v-SVM, aplicável apenas em regressão, e com a definição do limite de complexidade aplicável tanto para classificação quanto para regressão; deu-se preferência pelo algoritmo SVM, com o parâmetro custo aplicável tanto para regressão quanto para classificação, mas com a definição do parâmetro épsilon, exclusivo de regressão: ele define até que distância dos valores reais nenhuma penalidade será aplicada aos valores previstos. Para manter a complexidade baixa, foi escolhido o kernel linear, a tolerância numérica (o valor de desvio máximo permitido do valor esperado) foi definida em 0,0010, e o número máximo de iterações em 100.

Finalmente, os melhores parâmetros observados da Random Forest: o número de árvores foi definido em 10, poderia variar de 1 a 10000; treinamento replicável, mantendo fixa a geração de árvores; o número de atributos é a raiz quadrada do número de atributos da base; para controle de crescimento, a profundidade limite de árvores individuais, que poderia variar entre 1 e 50, foi definida em 3 (para que fosse mantida a complexidade baixa); e com o mesmo propósito de manter a complexidade baixa, subgrupos com profundidade menor que 2 não podem ser subdivididos.

4 Análise e Discussão dos Resultados

4.1 Resultados

Na Tabela 1 e no Gráfico 6 estão os resultados obtidos com os algoritmos de Machine Learning.

Algoritmo Testado	Erro Quadrático Médio (MSE)	Raiz do Erro Quadrático Médio (RMSE)	Erro Absoluto Médio (MAE)
Random Forest	245,05	15,65	13,23
Regressão Linear	312,69	17,68	15,06
SVM	357,28	18,90	16,17
Rede Neural	372,26	19,29	16,88

Tabela 1: Resultados dos algoritmos de machine learning testados. Fonte: Própria, 2020.

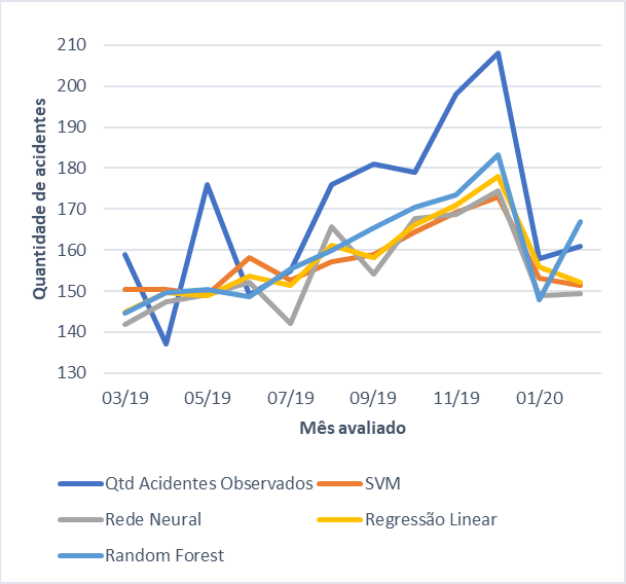

Gráfico 6: Comparativo das predições dos algoritmos de machine learning com as quantidades de acidentes com vítimas ocorridos no mês. Fonte: Própria, 2020.

O algoritmo Random Forest é uma técnica de regressão que combina o desempenho de vários algoritmos de árvores de decisão para classificar ou prever o valor de uma variável **[15]**. Pela combinação de um conjunto de classificadores este método é classificado como um ensembler. E tem as vantagens dessa classe de algoritmos, ele pode aplicar simultaneamente mais de um modelo aos dados. Este algoritmo apresentou o melhor resultado, atingindo um erro absoluto médio de 13 casos para mais ou para menos.

Outro algoritmo cujo resultado testado foi a regressão linear. Este algoritmo busca encontrar uma função polinomial que descreva o comportamento dos dados. Comparado a outras técnicas usadas neste trabalho, esta é uma das mais simples.

Os algoritmos de Redes Neurais (MLP) e Suporte Vector Machine (SVM) foram incluídos nos testes devido a sua predominância nos trabalhos relacionados que foram encontrados na literatura. Porém, para as condições do experimento, tiveram resultado inferior ao apresentado pela regressão linear.

Avaliou-se a relevância estatística dos resultados obtidos por meio de um teste de hipótese. Para este, foi assumido que os resultados seguem uma distribuição normal e foi tomado um grau de confiança de 95%. O teste pode ser descrito conforme equações na Figura 2. Nela A_s representa a o erro quadrático médio do algoritmo comparado e A_p representa o mesmo indicador para o algoritmo base da comparação. A operação *dp* representa o desvio padrão e *ad* representa o resultado do teste de hipótese **[16]**.

$$media(A_S - A_P) = media(A_S) - media(A_P)$$

$$dp(A_S - A_P) = \sqrt{\frac{dp(A_S)^2 + dp(A_P)^2}{2}}$$

$$ad(A_S - A_P) = \frac{media(A_S - A_P)}{dp(A_S - A_P)}$$

Figura 2: Fórmulas para cálculo de um teste de hipótese. Fonte: MONARD e BARANAUSKAS, 2003.

Comparação		Algoritmo Comparado			
		Regressão Linear	SVM	Rede Neural	Random Forest
Algoritmo Base	Regressão Linear	-	0,12	0,20	-0,21
	SVM	-0,12	-	0,07	-0,32
	Rede Neural	-0,20	-0,07	-	-0,41
	Random Forest	0,21	0,32	0,41	-

Tabela 2: Resultados do teste de hipótese comparando os algoritmos de machine learning testados. Fonte: Própria, 2020.

Pelo teste de hipótese (Tabela 2), tem-se que os valores acima de 0 para o algoritmo base Random Forest apontam que ele supera os demais. Porém seria necessário um valor acima de 2 para que a diferença tivesse um grau de confiança de 95%. Assim, percebe-se que a diferença entre os resultados dos algoritmos é marginal.

5. Conclusões

Através das aplicações do processo CRISP-DM conseguiu-se extrair um conjunto de informações de grande utilidade para o stakeholder. Com saída do processo de modelagem foi selecionado o algoritmo Random Forest como técnica a ser usada na predição da taxa futura de acidentes.

DOI: 10.25286/repa.v6i3.1707

5.1 Discussões

A taxa de erro relativamente baixa apresentada pela regressão quando comparada com outras técnicas mais sofisticadas destacam o caráter linear apresentado pelo problema.

Um motivo para que os algoritmos de SVM e MLP tenham atingido baixos desempenhos pode residir no fato de terem sido usados poucos exemplos para treino e para testes. uma forma de tentar atenuar esta situação envolve a mudança da granularidade da base observada, avaliando não a quantidade de acidentes por mês, mas a quantidade de acidentes por semana, aumentando a quantidade de amostras.

Foram realizados doze testes, e não um número mais adequado aos parâmetros da significância estatística, por causa de como a base e a metodologia de testes foi imaginada: não é usado um percentual aleatório da base para treino e o restante para testes, pois se fosse desta forma, poderia ser usado um mês adiante no treinamento, para prever um período passado, o que não teria utilidade prática; por isso, são selecionados os primeiros períodos para treino, e o seguinte para teste, na primeira rodada; na seguinte, novamente são usados os primeiros períodos, acrescido do período utilizado na rodada anterior, para prever o próximo, e assim por diante.

5.2 Trabalhos Futuros

Como proposta de trabalhos futuros, sugere-se:

- A expansão dos modelos de predição para períodos maiores, até 12 meses à frente;
- A mudança da granularidade da predição, por semana e por dia;
- A análise do impacto dos feriados na quantidade de acidentes;
- E o uso de informações georreferenciadas para predição da distribuição dos acidentes pela cidade.

Serão incluídas também outras técnicas de predição com o intuito de inserir a sazonalidade na análise, por meio da aplicação do modelo SARIMA, a fim de avaliar se com esta técnica obtém-se um resultado melhor que os vistos até então.

Referências

[1] Traffic Index 2019, disponível em https://www.tomtom.com/en_gb/traffic-index/ranking/. Visitado em 16/07/2020.

[2] Portal de dados abertos da prefeitura do recife, disponível em http://dados.recife.pe.gov.br/. Visitado em 16/07/2020.

[3] YONEZAWA, A; SHIBAYAMA, E.; TAKADA, T.; et al. Modeling and Programming in an Object-Oriented Concurrent Language. In A. Yonezawa, M. Tokoro, (eds.) Object-Oriented Concurrent Programming. MIT Press. páginas 55-90, 1991.

[4] Brandão, Lúcia Maria. Medidores eletrônicos de velocidade. Ponta Grossa. Paraná: 2006.

[5] Portal Town of Cary https://www.townofcary.org/home. Visitado em 31/07/2020.

[6] BRAGA, Luis Paulo Vieira. Introdução à Mineração de Dados. 2ª edição. Rio de Janeiro: E-papers, 2005.

[7] CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. Introdução à Mineração de Dados: Conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

[8] ZHANG, Tao et al. Nonparametric regression for the short-term traffic flow forecasting. In: 2010 International conference on mechanic automation and control engineering. IEEE, 2010. p. 2850-2853.

[9] YOUSEFZADEH-CHABOK, Shahrokh et al. A time series model for assessing the trend and forecasting the road traffic accident mortality. Archives of trauma research, v. 5, n. 3, 2016.

[10] ZLOTNIK, Alexander; MONTERO-MARTÍNEZ, Juan Manuel; GALLARDO-ANTOLÍN, Ascensión. A Comparison of Multivariate SARIMA and SVM Models for Emergency Department Admission Prediction. In: HEALTHINF. 2013. p. 245-249.

[11] ALAM, Ishteaque; FARID, Dewan Md; ROSSETTI, Rosaldo JF. The prediction of traffic flow with regression analysis. In: Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. p. 661-671.

[12] YANBIN, Yang et al. Early Warning of Traffic Accident in Shanghai Based on Large Data Set Mining. In: 2016 International Conference on

Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE, 2016. p. 18-21.

[13] CHANG, Li-Yen; CHEN, Wen-Chieh. Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research, v. 36, n. 4, p. 365-375, 2005.

[14] WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. London, UK: Springer-Verlag, 2000. p. 29-39.

[15] RODRIGUEZ-GALIANO, V. et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, v. 71, p. 804-818, 2015.

[16] MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003.

[17] HOFFMAN, R. e VIEIRA, S. Análise de Regressão. Uma introdução à econometria. 2ed. HUCITEC, São Paulo, 1983.

[18] VALENÇA, M. Fundamentos das Redes Neurais: Exemplos em Java. Livro Rápido, 2010.

[19] VAPNIK, V. Statistical Learning Theory. Wiley, 1998.

[20] GAMA, J. a. Functional trees. Machine Learning, v. 55, p. 219–250, 2004.

[21] HARTSHORN, S. Machine Learning with Random Forests and Decision Trees: A Visual Guide for Beginners, 2017.

[22] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research* 14(Aug): 2349−2353.

DOI: 10.25286/repa.v6i3.1707