Mineração de Dados para Análise e Predição das Infrações de Trânsito na Cidade do Recife

Data mining for analysis and prediction of traffic violations in the city of Recife

Arianne Sarmento Torcate [1] orcid.org/0000-0003-2779-873X

Maicon H. L. F. da S. Barros [1] orcid.org/0000-0002-0275-3298

Flávio Secco Fonseca [1] orcid.org/0000-0003-4956-1135

Marcos André Santos Galindo [1] orcid.org/0000-0002-1121-0084

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

E-mail do autor principal: Arianne Sarmento Torcate ast@ecomp.poli.br

Resumo

O aumento significativo de infrações de trânsito tem se tornado algo casual na vida dos brasileiros. A cidade do Recife, estado de Pernambuco, segundo a empresa Holandesa TomTom Traffic, no ano de 2018, ocupava a 10ª posição entre as cidades com o pior tráfego de automóveis no mundo. Em 2019 passou a ocupar a 15ª posição. Diante disto, esta pesquisa tem como intuito investigar fatores relacionados ao aumento da quantidade de equipamentos de aferição automática e de agentes de trânsito. O objetivo é criar um modelo de predição de delitos no trânsito por turnos, testado com a base real referente ao ano de 2019. Para guiar o processo de mineração e análise de dados, a metodologia CRISP-DM foi aplicada. Os resultados obtidos apontam que há um aumento significativo de infrações em feriados, principalmente no Corpus Christi. Além disso, as predições mensais apresentam bons resultados quando comparados aos números reais de infrações.

Palavras-Chave: Infrações; Mineração de Dados, Previsões; Visualizações; Análises.

Abstract

The significant increase in traffic violations has become somewhat casual in the lives of Brazilians. The city of Recife, state of Pernambuco, according to the Dutch company TomTom Traffic, in 2018, ranked 10th among the cities with the worst car traffic in the world. In 2019 it ranked 15th. Given this, this research aims to investigate factors related to the increase in the amount of automatic measurement equipment and traffic agents. The objective is to create a model of prediction of violations in traffic by turns, tested with the real basis for the year 2019. To guide the mining and data analysis process, the CRISP-DM methodology was applied. The results obtained indicate that there is a significant increase in infractions on holidays, mainly in Corpus Christi. In addition, monthly predictions show good results when compared to the actual numbers of infractions.

Key-words: *Infractions; Data Mining, Forecasting; Views; Reviews.*

DOI: 10.25286/repa.v6i3.1679

1 Introdução

As infrações de trânsito tem sido um fator presente na vida das pessoas. As estatísticas apresentadas pelos órgãos de trânsito indicam que cometer infrações e desrespeitar os preceitos da legislação são práticas muito frequentes entre os motoristas brasileiros [1]. O artigo 161 do Código de Trânsito Brasileiro [2] define infrações como a inobservância de qualquer preceito deste código, da legislação complementar ou das resoluções do Conselho Nacional de Trânsito (CONTRAN), sendo o infrator sujeito às penalidades e medidas administrativas. Os artigos 162 a 255 destacam quais são as gravidades e penalidades de acordo com cada infração.

No estudo realizado por Rocha [3], é exposto que as infrações cometidas no trânsito do Brasil resultam em milhares de mortes, feridos e prejuízos materiais. É diante desse cenário que o Brasil vem se destacando negativamente no ranking mundial de acidentes de trânsito. De acordo com os dados mostrados, há em média 6,8 mortes para cada 10 mil veículos [4].

Nesse contexto, pesquisas [5, 6] na literatura defendem que a educação é a principal ferramenta para alcançar soluções referentes ao trânsito brasileiro. Disseminar a educação no trânsito, segundo Macêdo e Souza [4], refere-se a ensinar que evitar infrações, por exemplo, deve ser realizado devido aos benefícios alcançados e não por ser uma forma ou instrumento para evitar possíveis punições.

Diante desse cenário, somente nos últimos anos, surgiu um crescente interesse entre estudiosos para entendimento deste contexto envolto às infrações [3]. Portanto, a presente pesquisa tem como foco o entendimento das infrações referentes a cidade de Recife no estado de Pernambuco. De acordo com os dados apresentados pela empresa holandesa TomTom Traffic [7] em relação ao ano de 2019, a cidade do Recife ocupava a 15ª posição entre as cidades com pior tráfego de automóveis do mundo. A mesma fonte aponta que em 2018 a situação era ainda mais preocupante, onde Recife ocupava a 10ª posição. A fim de entender esse contexto e correlacionar com o índice crescente de infrações, as bases de dados referentes aos anos de 2017 e 2018 disponibilizadas pela Autarquia de Trânsito e Transporte Urbano do Recife (CTTU) [8] foram analisadas.

O objetivo é aplicar técnicas e tarefas de Mineração de Dados (MD) com o auxílio de tecnologias correlacionadas para a análise, visualização e previsão das infrações referentes ao ano de 2019. Segundo Cardoso e Machado [9], mineração de dados é uma técnica ou processo que extrai conhecimento de conjuntos de dados, capaz de revelar conhecimentos implícitos pela grande quantidade de informações armazenadas. Além disso, os autores apontam que as técnicas de MD possibilitam análise de eventos, previsão de tendências e comportamentos futuros.

A metodologia de Processo Padrão Inter-Indústrias para Mineração de Dados (CRISP-DM) [10] foi utilizada para guiar a mineração e análise dos dados desta pesquisa. Isto foi possível com o auxílio de softwares como o Orange Data Mining, GretL, Weka, Pentaho PDI e Python, que foram utilizados em paralelo, tanto para análise, como para mineração dos dados e previsões. O intuito é criar um modelo que possa ser adaptado conforme o passar dos anos, através de previsões a respeito das multas por períodos do dia e mês de cada ano. Através do modelo espera-se ofertar um feedback assertivo e útil para que as autoridades responsáveis tomem as medidas de prevenção e conscientização cabíveis para evitar tais infrações e prejuízos ocasionados.

O restante do artigo está organizado da seguinte forma: A Seção 2 apresenta o referencial teórico utilizado nesta pesquisa, a Seção 3 discorre sobre os materiais e métodos utilizados e experimentos, na seção 4 são apresentados os resultados obtidos e suas respectivas discussões e, por fim, na seção 5, relatamos as considerações finais juntamente com as perspectivas de trabalhos futuros.

1.1. Escopo Negativo

A previsão desenvolvida na presente pesquisa será referente ao ano de 2019, a ideia inicial do projeto era desenvolver referente ao ano atual, 2020. Entretanto, diante dos impasses provocados e ocasionados pela pandemia de Covid-19, os pesquisadores acreditam que as infrações de 2020 podem ser influenciadas, o que afetaria diretamente nos resultados e na previsão realizada neste estudo. Além disso, como é exposto na subseção 2.2.1, dentre as etapas do CRISP-DM, esta pesquisa não contempla a 6ª etapa, pois diante dos impasses e do contexto remoto, fica inviável aplicar a solução neste momento em um contexto real.

2 Fundamentação Teórica

Nesta seção, apresentamos de forma objetiva as principais tecnologias e abordagens para realização deste projeto, com ênfase em mineração de dados e na metodologia CRISP-DM, além dos trabalhos motivadores desta pesquisa.

2.1 Mineração de Dados

A área de mineração de dados (do inglês, data mining) surge como um processo não trivial e com o objetivo de identificar em conjuntos de dados (do inglês, dataset) padrões válidos, novos, compreensíveis e potencialmente úteis **[11]** para geração de conhecimentos e auxílio na tomada de decisões. Resumidamente, mineração de dados é um processo altamente cooperativo entre seres humanos e máquinas que visa a exploração de grandes bancos de dados com o objetivo de extrair conhecimentos através de reconhecimento de padrões e relacionamento entre variáveis **[12]**.

Dessa maneira, para gerar conhecimentos úteis, utiliza-se técnicas de inteligência artificial, métodos estatísticos e ferramentas de visualização e mineração **[11]** (Exemplo: Weka, RapidMiner **[13]**, Orange canvas **[14]**, R e Tanagra **[15]**). É válido mencionar que a mineração de dados é uma área abrangente e possui aplicação em diversos contextos, como educação, indústrias, empresas, medicina, finanças e dentre outras **[16]**.

É importante observar que para alcançar o objetivo desta pesquisa, séries temporais foram utilizadas para realizar as previsões. Segundo Antunes e Cardoso **[17]**, as séries temporais podem ser compreendidas como uma sequência de dados quantitativos, relativos a momentos ou períodos específicos que estão organizados e distribuídos no tempo. Já as previsões, podem ser encaradas como informações críticas que auxiliam na tomada de decisão de negócios **[18]**. Ambas, utilizadas em conjunto, possuem aplicabilidade em diversos contextos e áreas do conhecimento.

Para guiar o processo de mineração de dados de forma organizacional, metodologias foram desenvolvidas, exemplos disso é a metodologia KDD (Knowledge Discovery in Databases) e SEMMA (Sample, Explore, Modify, Model, Assess) **[19]**. A subseção a seguir apresenta a metodologia utilizada na presente pesquisa.

2.1.1 CRISP-DM

Dentre as metodologias disponíveis para guiar o processo de mineração de dados em projetos, destaca-se a Cross Industry Standard Process for Data Mining (CRISP-DM), apontada por Martínez-Plumed et al. **[10]** como a metodologia analítica mais completa e utilizada em projetos industriais que visam a descoberta de conhecimentos ou informações úteis. O CRISP-DM é uma metodologia que foca nas necessidades dos gestores e na resolução dos seus problemas de gestão **[20]**.

A metodologia CRISP-DM tem adesão e preferência de 42% dos profissionais **[21]**. Isso ocorre pelo fato desta metodologia proporcionar ao pesquisador, através de sua estrutura organizacional, uma visão ampla que contempla desde o entendimento do negócio até a apresentação dos resultados obtidos pela mineração dos dados **[10]**. Segundo Mattozo **[19]**, isto é possível através das seis etapas que compõem o ciclo de vida deste modelo, conforme apresenta o Quadro 1, baseado no estudo de Chapman *et al*. **[22]**.

Quadro 1: Etapas do CRISP-DM.

Etapas	Definição
Business Understanding	Responsável pelo entendimento dos objetivos, requisitos, definição de problema e elaboração do planejamento.
Data Understanding	Fase de coleta inicial dos dados para entendimento e familiarização com os mesmos a fim de identificar problemas e níveis de qualidade.
Data Preparation	Abrange as atividades necessárias para construir o conjunto de dados final. Ou seja, são realizadas as tarefas de seleção, normalização, integração e limpeza.
Modeling	Seleção e aplicação de técnicas de modelagem, além da calibração e configuração dos parâmetros para valores ideais e aplicáveis.
Evaluation	Avaliar e revisar as etapas executadas do modelo para garantir que o mesmo atinja os objetivos.
Deployment	Aplicar os modelos criados em contexto real.

Por fim, esta metodologia é considerada mais completa e documentada em relação a outras, como KKD e SEMMA **[17]**, pelo fato de concentrar benefícios na avaliação da qualidade dos dados, bem como na sua interpretação dos achados no processo de Mineração de dados, além de contribuir fortemente no entendimento do negócio **[21]**.

DOI: 10.25286/repa.v6i3.1679

2.2 Trabalhos Relacionados

A pesquisa realizada por Amiruzzaman **[24]** teve como objetivo prever violações de tráfego com base em incidentes passados. Para isso, algoritmos (J48, Árvore de decisão e Naive Bayes) de mineração de dados foram utilizados com auxílio de ferramentas (SPSS e WEKA). Os resultados da mineração de dados apontam que é perigoso sair por volta de uma da manhã, pois a maioria dos danos materiais e lesões corporais ocorreu devido a motoristas bêbados que circulavam entre as onze horas e uma da manhã. O trabalho possui potencial para ajudar a evitar violações de tráfego ou reduzir a chance de ocorrência, além de ofertar orientações de cuidados.

No trabalho de Shiau *et al.* **[25]** é apresentada uma aplicação de mineração de dados para prever acidentes de trânsitos e quais causas/infrações cometidas. Os pesquisadores analisaram 2.471 acidentes de trânsito no centro de Taiwan utilizando variados métodos (Fuzzy Robust, FRPCA, Neural Network e LR). Os resultados apontam taxas de acurácias acima de 84,37% em todos os classificadores testados. A análise estatística reforça que o modelo criado pode ser utilizado pela polícia ou autoridades reguladoras para projetar, planejar e melhorar a segurança do tráfego, diminuindo os acidentes e perda de vidas naquela região.

Cuenca *et al.* **[26]** apresenta um protótipo de um sistema de visualização e avisos de pontos de acesso de tráfego através de informações geográficas da área metropolitana de Madri. O objetivo é alertar os motoristas ao se aproximarem de pontos de situações incomuns. Para isso, o processo de mineração de dados é utilizado para geração do conjunto de dados necessários e assertivos para a alimentação do sistema. Isso foi possível com o auxílio da ferramenta Pentaho Data Integration (PDI) e Carto-API.

Diante dos trabalhos apresentados nesta seção, fica evidente o potencial da área de mineração de dados aplicada a contextos com objetivos diversos. Diferentemente dos estudos apresentados, esta pesquisa tem como diferencial prever as infrações de trânsito na cidade do Recife/PE, utilizando em conjunto tecnologias como Orange Data Mining, WEKA, Python e Pentaho PDI. Além disso, propomos visualizações que sejam entendíveis para quaisquer públicos que tenham interesse em acessar e entender as infrações e suas respectivas correlações.

3 Materiais e Métodos

Esta seção com suas respectivas subseções é responsável por apresentar os materiais, ferramentas e métodos utilizados no decorrer deste projeto, bem como a aplicação da metodologia CRISP-DM.

3.1 Descrição da Base de Dados

A base de dados utilizada para realização deste projeto foi adquirida através de duas fontes. A primeira refere-se ao portal de dados abertos da CTTU **[8]** da cidade do Recife, onde o dataset obtido faz referência ao ano de 2017 e 2018 e, possuem oito atributos. A segunda fonte foi adquirida pela Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais (ANBIMA) **[23]** e pela prefeitura da cidade do Recife **[27]** e, possui três atributos. O Quadro 2 apresenta os onze atributos.

Quadro 2: Atributos do dataset.

Atributos	Descrição
Data de Infração	Este atributo é do tipo *timestamp* e contém as datas que as infrações foram cometidas.
Hora da Infração	Atributo do tipo *text* e possui informações sobre o horário que aconteceu as infrações.
Data de Implementação	Atributo *timestamp,* contém as datas que as infrações foram realmente implementadas.
Agente Equipamento	Atributo do tipo *text*, contém informações sobre qual equipamento registrou a infração.
Infração	Atributo *numeric*, composto pelos códigos referentes a cada infração cometida.
Descrição da infração	Atributo do tipo *text*, composto pela descrição dos detalhes sobre a infração.
Amparo legal	Atributo *text*, contém os artigos do Código de Trânsito Brasileiro violado.
Local	Atributo do tipo *text*, contém detalhes sobre o local que a infração ocorreu.
Data de evento cultural	Este atributo é do tipo *text* e contém datas de eventos culturais e feriados.
Desc. de evento cultural	Atributo *text* com detalhes sobre o feriado ou evento cultural.
Dia de evento cultural	Este atributo é *numeric* e identifica se o dia que a infração foi cometida é um dia de evento cultural (0) ou, um dia comum (1).

É válido destacar que foi realizada a junção de atributos de fontes distintas para encontrar relação entre o aumento de infrações com o período de eventos culturais (ex: Carnaval, São João, Feriados e dentre outros). É importante enfatizar que na etapa de pré-processamento dos dados (Subseção 3.2) estes onze atributos foram adaptados para atender os objetivos do projeto.

3.2 Pré-processamento dos Dados

A ferramenta utilizada para realizar o pré-processamento dos dados foi o Pentaho PDI. Com isso, inicialmente, foi realizada a junção da base de dados referente ao ano de 2017 com a de 2018.

Após realizar a junção de bases, foi realizada a leitura do arquivo CSV. Em seguida, foi feito a substituição de "/" por "-" no atributo "datainfracao", pois foi identificado que existiam períodos com formatações divergentes, ou seja, algumas estavam com padrões de "dd/mm/aaaa" e outras com "dd-mm-aaaa". Tendo em vista que a maioria dos dados estavam com o padrão "dd-mm-aaaa", optou-se por substituir apenas as datas que estavam no padrão "dd/mm/aaaa", visando a economia de processamento computacional.

Dando continuidade, substituiu-se o apóstrofo nos atributos "agenteequipamento" e "localcometimento" por um caractere vazio "", pois os arquivos do tipo Comma-Separated Values (CSV) para leitura em softwares de predição e visualização dos dados possuíam erros devido ao padrão do CSV e delimitadores utilizados.

Posteriormente, foi realizada também a aplicação de expressão regular para extração do código do equipamento da infração na variável "agenteequipamento", pois ora estava como um código categórico, ora estava com um código categórico concatenado com um texto (exemplo: "8" e "8 - Talão").

Em seguida, fez-se a junção da base de dados da primeira fonte com a segunda (referente aos três atributos de eventos culturais). A Figura 1 ilustra a estrutura de pré-processamento no Pentaho PDI.

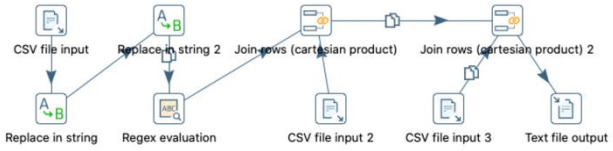

Figura 1. Primeira etapa Pré-Processamento.

A segunda etapa do pré-processamento teve como objetivo preparar a base de dados para predição. Para isso, o dia, mês e ano foram separados em campos distintos. Em seguida, a hora foi segregada para captar apenas o momento exato do acontecimento da infração. Para isso, 4 variáveis de hora foram criadas juntamente com 4 expressões regulares para que cada variável (coluna) fosse preenchida com a finalidade de visualização dos dados, sendo: i) Maturno (de 00:00 às 05:59); ii) Matutino (de 06:00 às 11:59); iii) Vespertino (de 12:00 às 17:59) e, por fim, iv) Noturno (de 18:00 às 23:59).

Foram concatenadas, ainda, as quatro colunas de turno, bem como o ano e o mês. Com a concatenação das variáveis dos turnos, ficaram misturadas as informações de horas com os turnos. Por isso, foi criada mais uma expressão regular que extraísse dessa coluna os nomes dos turnos da infração mencionados anteriormente. Adiante, foi realizada uma operação de agrupamento com contagem de valores, ou seja, o número de infrações por mês/ano e por turno. Foram ordenados os valores por mês/ano, turno e quantidade de infrações de maneira ascendente. Em seguida, substituiu-se os turnos maturno, matutino, vespertino e noturno por códigos, sendo respectivamente 0, 1, 2, 3 com a finalidade de predição. A Figura 2 ilustra a dinâmica utilizada no Pentaho para a segunda etapa do pré-processamento.

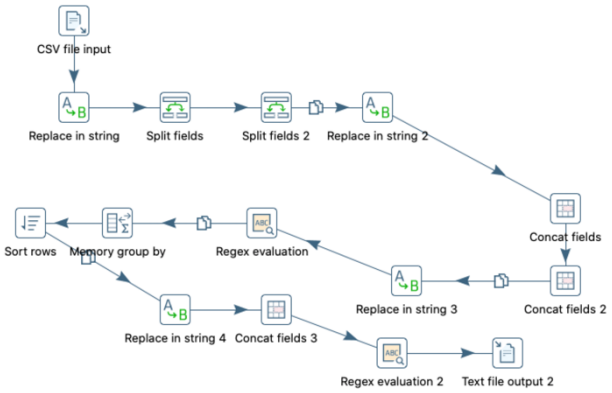

Figura 2. Segunda etapa do Pré-Processamento.

A expressão regular foi executada para extrair apenas os números categóricos dos turnos e, por fim, foi escrito em um arquivo do tipo CSV os resultados da base para predição.

3.3 Metodologia Experimental

Para a metodologia do experimento, foram utilizadas ferramentas de software e algoritmos desenvolvidos para analisar e verificar qual o modelo com melhor desempenho para a predição das séries temporais utilizando a base de infrações de 2017 e 2018. Inicialmente, a hipótese deste trabalho foi

DOI: 10.25286/repa.v6i3.1679

realizar a predição referente aos doze meses de 2019. Porém, com o treinamento dos modelos computacionais, os pesquisadores verificaram que alguns modelos apresentavam overfitting. Por este motivo, foram também testadas predições utilizando a frequência diária dos dados, frequência mensal e frequência anual para prever um dia e um mês referente ao ano de 2019 e prever com base nos dados do ano de 2017 o ano de 2018.

Com isso, as bases foram divididas em oito datasets de acordo com cada frequência da série temporal e considerando seus respectivos turnos (Conforme apresenta a Quadro 3).

Quadro 3: Informações sobre a divisão dos datasets.

Datasets	Descrição
dataset_2017 _mensal	Quatro datasets referente aos turnos (maturno, matutino, vespertino e noturno) foram criados separadamente para predição mensal do ano de 2018.
dataset_20172018 _mensal	Quatro datasets referente aos turnos (maturno, matutino, vespertino e noturno) foram criados separadamente para predição mensal.
dataset_20172018 _diario	Quatro *datasets* referente aos turnos (maturno, matutino, vespertino e noturno) foram criados separadamente para predição diária, referente ao dia 01 de Janeiro.

A pesquisa foi dividida em quatro experimentos visando obter o melhor resultado para o modelo e a melhor configuração através da avaliação do acerto da predição com relação aos dados reais já existentes referentes ao primeiro dia de 2019 e ao mês de janeiro de 2019, bem como referente ao ano de 2018 através dos dados de 2017 para fins de definição do melhor modelo computacional a ser utilizado para predições de infrações usando estas bases de dados. Destacam-se os três experimentos:

1. **Primeiro experimento:** Foram utilizados os *datasets* por periodicidade diária mencionados no Quadro 3, com o software Weka. Com isso, os dados foram carregados na ferramenta e, em seguida, foi realizado o teste dos seguintes modelos computacionais para predição: *Linear Regression* (LR), *Multi Layer Perceptron* (MLP), *HoltWinters* (HW), *Random Forest* (RF) e um modelo *Ensemble* com a combinação de todos os modelos anteriormente mencionados.

   As configurações utilizadas no experimento para cada modelo foram as seguintes:

   - LR: *batchSize*: 100; método M5 de seleção de atributos;
   - MLP: *batchSize*: 100; camadas ocultas automáticas, ou seja, o Weka encontra qual a melhor configuração de camadas; taxa de aprendizagem: 0,3;
   - HW: *batchSize*: 100; duração do ciclo da temporada 12; duração fator de suavização sazonal (smoothing) 0,2;
   - RF: *batchSize*: 100; percentual do tamanho do *batchSize* 100;
   - *Ensemble*: modelo meta *Stacking* combinando LR + MLP + HW + RF.

2. **Segundo experimento:** Foram utilizados os *datasets* por periodicidade mensal mencionados no Quadro 3, com o software Weka. Com isso, foi realizado o mesmo que no primeiro experimento. Entretanto, a periodicidade da série temporal foi mensal, alterando apenas o arquivo importado para que o software Weka pudesse considerar os *datasets* mensais com as mesmas configurações do experimento anterior;

3. **Terceiro experimento:** Para o terceiro experimento, foi utilizado o software Gretl **[28]**, onde foram importados somente os datasets mensais para se obter a predição do modelo ARIMA **[29]**. Neste modelo, as séries temporais univariadas foram utilizadas, considerando como variável dependente a quantidade de infrações, sem regressores, com a série temporal não-sazonal, utilizando a matriz de covariância por Hessiana, e com Máxima Verossimilhança Exata;

4. **Quarto experimento:** Foram utilizados os datasets por periodicidade anual mencionados no Quadro 3, com o software Weka. Com isso, foi realizado o mesmo que no primeiro experimento. A periodicidade da série temporal foi mensal, prevendo os 12 meses de 2018 através dos dados de 2017, alterando apenas o arquivo importado para que o software Weka pudesse considerar os datasets mensais com as mesmas configurações do experimento anterior.

4 Análise e Discussão dos Resultados

Nesta seção, apresentamos os resultados obtidos através dos softwares e algoritmos que foram utilizados em conjunto e aplicados às bases de dados apresentadas na subseção 3.3.

4.1 Resultados

A fase relacionada à etapa de mineração de dados foi de grande relevância, pois este processo permitiu a extração e melhor visualização das informações úteis e não triviais relacionadas aos aspectos que envolvem as infrações. Um exemplo dos resultados desta etapa pode ser visualizado na Figura 3, através do gráfico *box plot*, onde o atributo de feriados foi criado e associado com a frequência relativa dos agentes e equipamentos, permitindo a observação de informações previamente desconhecidas vinculadas a estes eventos culturais. Os agentes/equipamentos associados a cada cor são: Azul - lombada eletrônica; Vermelho - fotossensor; Verde - Zona Azul; Laranja - Talão eletrônico e Amarelo - Faixa Azul.

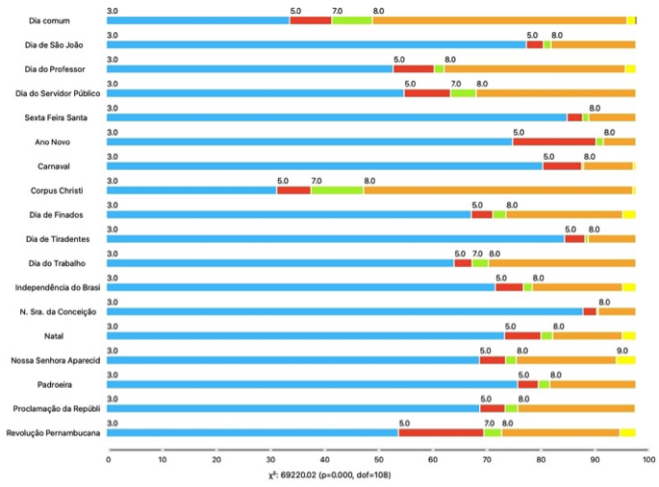

Figura 3. Frequência de infrações por agente/equipamento e feriados.

Com a finalidade de atender aos *stakeholders* do projeto foi gerado o gráfico de dispersão apresentado na Figura 4, contendo as quantidades de infrações dos anos de 2017 e 2018 por turnos do dia. Sendo codificado da seguinte forma: 1 – domingo, 2 – segunda, 3 – terça, 4 – quarta, 5 – quinta, 6 – sexta e 7 – sábado.

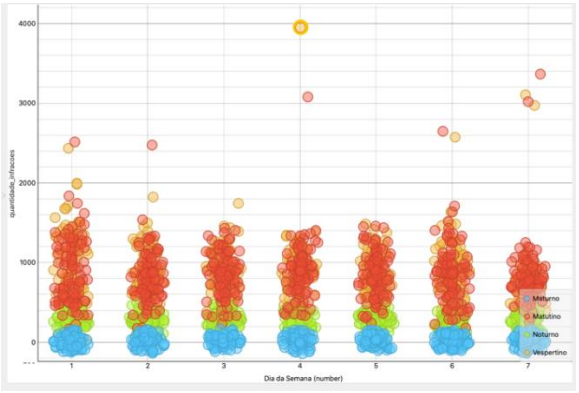

Figura 4. Dispersão – Quantidade de infrações por dias da semana dos anos de 2012 e 2018.

Diante do que já foi descrito na subseção 3.4 sobre os experimentos referente às etapas de previsões, a Figura 5 apresenta um comparativo entre os resultados obtidos com os modelos *Linear Regression* (LR), *Multi Layer Perceptron* (MLP), *HoltWinters* (HW), *Random Forest* (RF) e ARIMA combinados no *Ensemble* e os dados reais por turnos no mês de janeiro do ano de 2019.

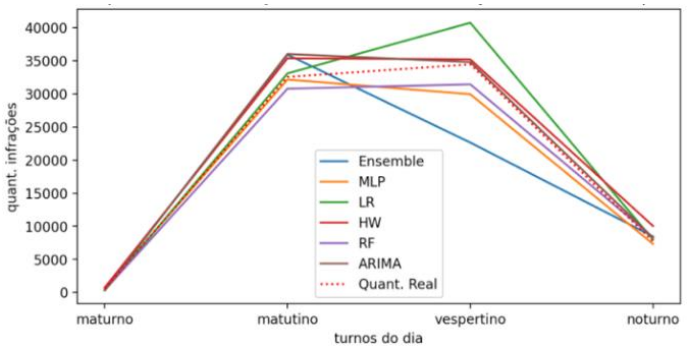

Figura 5. Previsão da qualidade de infrações para janeiro de 2019.

Para melhor visualização e entendimento numérico, a Tabela 1 expõe de forma organizada e quantificável os resultados obtidos.

Tabela 1: Resultado da predição de infrações para janeiro de 2019 na cidade do Recife.

Turno	Ens	LR	MLP	HW	RF	ARIMA	Qtd.Real
Maturno	428	313	608	753	376	350	542
Matutino	35.937	33.100	32.182	35.389	30.795	36.038	23.569
Vespertino	22.656	40.755	29.955	35.223	31.458	34.815	34.474
Noturno	8.464	7.895	7.332	10.040	8.343	8.273	7.739

DOI: 10.25286/repa.v6i3.1679

O mesmo procedimento de previsão foi realizado com uma menor granularidade, prevendo o número de infrações apenas para o primeiro dia de janeiro de 2019, conforme apresenta a Figura 6.

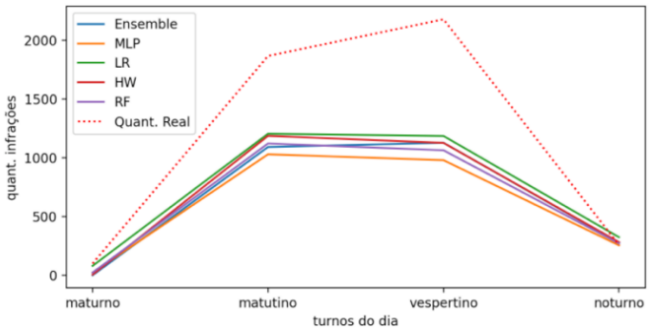

Figura 6. Previsão da quantidade de infrações para 01 de janeiro de 2019.

É possível identificar pela análise da Figura 6, que os resultados apresentam uma distância maior em relação a quantidade real de infrações registradas nesta data. Este fato fica ainda mais evidente com os dados numéricos apresentados pela Tabela 2.

Tabela 2: Resultado da predição de infrações para 01 de janeiro de 2019 na cidade do Recife.

Turno	Ens	LR	MLP	HW	RF	Qtd.Real
Maturno	0	80	22	5	23	100
Matutino	1.092	1.205	1.030	1.187	1.122	1.867
Vespertino	1.128	1.186	980	1.127	1.064	2.179
Noturno	278	324	255	282	271	258

É válido enfatizar que essa mesma previsão foi realizada referente a outros dias de janeiro de 2019, mas, mesmo assim, a taxa de *overfitting* é significativamente alta. Esse teste também foi realizado com um espaço temporal maior, considerando os dados de 2006 a 2018, mas a previsão diária continuou obtendo resultados discrepantes.

Para avaliar o resultado das predições, os autores utilizaram a taxa de erro real, ou seja, o percentual de variação entre o número previsto e o número real que ocorreu no mês em questão, tendo em vista que o objetivo do trabalho era predizer a quantidade de infrações por turno. A Tabela 3 apresenta os resultados obtidos.

Tabela 3: Taxa de erro real dos modelos por turno.

Turno	Ens	LR	MLP	HW	RF	ARIMA
Maturno	27%	73%	11%	28%	44%	55%
Matutino	9%	2%	1%	8%	6%	10%
Vespertino	52%	15%	15%	2%	10%	1%
Noturno	9%	3%	6%	23%	7%	6%

Com a finalidade de validar qual o melhor modelo indicado para previsão de séries temporais com relação a base da CTTU, o experimento 4 foi realizado utilizando a base de dados de 2017 para prever os próximos dozes meses de 2018. Os resultados estão apresentados na Figura 7.

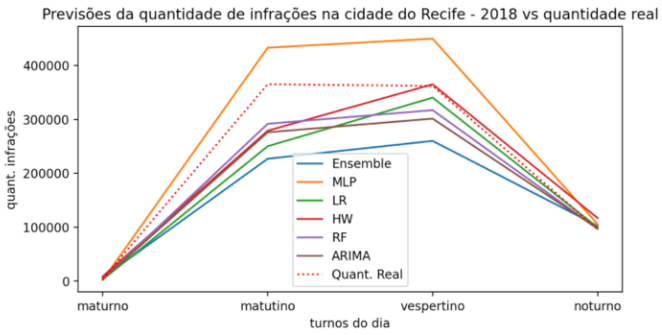

Figura 6. Previsão da quantidade de infrações para o ano de 2018 utilizando o ano de 2017.

As demais métricas que os autores deste trabalho utilizaram foram as técnicas do Erro Médio Absoluto (do inglês, MAE), e a técnica raiz média do erro médio quadrático (do inglês, RMSE).

4.2 Discussões

Na análise dos dados exposto na Figura 3, é possível verificar que há um aumento das infrações registradas por lombadas eletrônicas de trânsito em dias de feriados com relação aos dias comuns. Também é notório que o maior número de infrações está concentrado nos agentes de talão eletrônico e nas lombadas eletrônicas. O feriado com o maior número de infrações registradas é a Sexta-feira Santa, com 87% das infrações de trânsito registradas nesse dia no agente denominado Lombada Eletrônica. Já o feriado com o maior número de infrações no agente de Talão Eletrônico, é no Corpus Christi, com 50% das infrações registradas neste equipamento. Com relação ao agente Fotossensor, teve o número mais elevado de

infrações nos feriados de Ano Novo e na Revolução Pernambucana. O agente Zona Azul representou 10% no feriado de Corpus Christi, sendo o seu maior pico. Em continuidade, o agente Faixa Azul apresentou 4% no feriado de Nossa Senhora Aparecida.

Com relação a quantidade de infrações, registradas por turno, referentes aos dias da semana apresentados na Figura 4 é possível notar que o dia com maior registro de infrações é o domingo, seguido da sexta-feira. E o dia com menor número de infrações é o sábado. Todos os dias apresentam um padrão de comportamento com relação ao número de infrações, sendo um número menor com relação ao turno maturno e noturno, e registrando picos de infrações nos períodos matutino e vespertino.

A predição das séries temporais ocorreu através dos experimentos mencionados na seção 3.4 e 4.1 deste trabalho. De acordo com a Figura 5, percebe-se que os modelos Holt Winters e o ARIMA foram os que mais se aproximaram com relação aos valores previstos para o mês de janeiro de 2019, por turno. É importante lembrar, que as bases de dados utilizadas na predição das séries temporais foram segregadas por turnos. Por este motivo, é possível indicar mais de um modelo de predição, de acordo com cada período do dia em que o modelo representou um melhor acerto com relação ao número de infrações.

Desta forma, é possível indicar um modelo para cada turno conforme a taxa de acerto real da predição. Com relação ao turno maturno, a técnica com menor taxa de erro real é de 11% para a MLP. Já com relação ao turno matutino o melhor modelo seria o MLP com 1% de taxa de erro real. O turno vespertino apresenta um melhor desempenho no modelo ARIMA com 1% de taxa de erro real. Por conseguinte, o turno noturno apresenta um melhor resultado no modelo MLP. Então, em caso de escolha de um modelo pela taxa de assertividade das previsões reais, ou seja, que ocorreram de fato em janeiro de 2019, o melhor modelo para predição pela média geral seria o modelo de MLP.

Como apresenta a Figura 6 em relação a periodicidade diária, fica nítido que a previsão de um dia de infração, tem valores muito distantes dos valores reais registrados em 01 janeiro de 2019. Isso pode ter ocorrido, devido a um aumento significativo de infrações registrados neste dia, por ser um feriado prolongado. Por este motivo, o experimento 2 foi realizado, utilizando a granularidade mensal ao invés de diária, e este obteve um melhor resultado conforme já descrito.

Por fim, diante do percentual exposto na Tabela 3, é possível identificar que o modelo que mais obteve destaque para predizer os turnos, de forma geral, foi o MLP, seguido pelo HW. Entretanto, para o experimento 4 apresentado nos resultados da Figura 7, é possível notar que os modelos que mais se aproximaram das quantidades reais para o ano de 2018 foram os RF seguido pelo modelo ARIMA.

É válido enfatizar que utilizar em conjunto softwares como Weka, Orange data Mining, Python e GretL foi de grande importância para identificar a relação entre os atributos de forma abrangente e assertiva. Assim como também foi de grande importância utilizar figuras e tabelas em conjunto para verificar e analisar de forma visual e quantificável os resultados obtidos em cada modelo e situação pesquisada.

5. Conclusões e Trabalhos Futuros

O principal objetivo deste trabalho foi aplicar técnicas de predição para a extração de conhecimento de uma base de dados bruta relacionada as infrações de trânsito da cidade do Recife. Em continuidade, foi importante calibrar os modelos de séries temporais para que os mesmos possam ser utilizados futuramente, por exemplo, para investir um maior número de determinados agentes em alguns períodos do dia.

Os modelos criados podem ser utilizados pelos órgãos responsáveis de forma adaptada, com o passar dos anos, para previsões a respeito das multas por períodos do dia e mês de cada ano. Com isso, é possível que estes órgãos tenham feedbacks assertivos para, por exemplo, tomar medidas de prevenção e conscientização cabíveis para evitar tais infrações e respectivos prejuízos. Entretanto, esta pesquisa a concluir também que o melhor modelo para predição de séries temporais de infrações de trânsito da base de dados CTTU, Recife, depende da granularidade da série temporal, ou seja, a depender do objetivo da predição, seja ela diária, mensal ou anual, cada modelo apresenta melhores repostas para cada situação em particular.

DOI: 10.25286/repa.v6i3.1679

Como perspectivas de trabalhos futuros, pretende-se analisar e prever o número de infrações considerando os bairros. Além disso, outro objetivo é analisar e identificar possíveis influências e relações de fatores temporais, como por exemplo, dias chuvosos no aumento de infrações. Também pretende-se treinar os modelos com um período maior de dados, entre 2006 e 2019.

Referências

[1] NETO, I. L., IGLESIAS, F. & GÜNTHER, H. Uma Medida de Justificativas de Motoristas para Infrações de Trânsito, Psico, Porto Alegre, PUCRS, v. 43, n. 1, pp. 7-13, 2012.

[2] BRASIL. Código de trânsito brasileiro. Brasília: Senado Federa. 1998.

[3] ROCHA, J. B. A. Infrações no trânsito: uma necessária distinção entre erros e violações. Interação em Psicologia, 9 (1), p. 177-184 1; 2005.

[4] MACÊDO, A. P. B.; SOUZA, P. R. P. Traffic education: a study on human behavior. Brazilian Journal of Development Braz. J. of Develop., Curitiba, v. 6, n. 7, p. 44548-44566, Jul. 2020.

[5] PEREIRA, L. B. F. *ET AL.* Educação para o trânsito no ensino básico. Congresso de Pesquisa e Ensino em Transporte da ANPET, 2019.

[6] JUNIOR, D. G. A. Educação de trânsito: a necessidade premente de um trânsito mais altruísta. Gestão de Trânsito - Unisul Virtual, 2019.

[7] TOMTOM TRAFFIC. Full Ranking, 2019.

[8] AUTARQUIA DE TRÂNSITO E TRANSPORTE URBANO DO RECIFE (CTTU). Registro de Infrações Trânsito, 2020.

[9] CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. RAP – Rio de Janeiro 42(3):495-528; 2008.

[10] MARTINEZ-PLUMED *ET AL*. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 1–1; 2019.

[11] GALVÃO, N. D.; MARIN, H. F. Técnica de mineração de dados: uma revisão da literatura. Acta Paul Enferm - 22(5):686-90; 2009.

[12] CÔRTES, S. C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de dados - funcionalidades, técnicas e abordagens. PUC – RioInf.MCC10/02; 2002.

[13] ROSSINI, L. A. S. *ET AL.* DATA MINING: conceitos e consequências. Revista Interface Tecnológica, V. 15 N. 2; 2018.

[14] ANTUNES, R. R.; BIAS, E. D.; BRITES, R. S.; COSTA, G. A. O. P. Análise de Integração de Mineradores de Dados com a Plataforma InterIMAGE – Qual a Melhor Solução?. Rev. Bras. Cartogr. v. 70 n. 4; 2018.

[15] KUCHINISKI, B. C. T. Aplicação de Métodos de Mineração de Dados em Bases de Dados de Crédito e Seguro de Clientes. 57p. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Produção) - Universidade Tecnológica Federal do Paraná; 2018.

[16] AMARAL, F. Introdução à Ciência de Dados: mineração de dados e big data. Rio de Janeiro: Alta Books; 2016.

[17] ANTUNES, J. L. F.; CARDOSO, M. R. A. Uso da análise de séries temporais em estudos epidemiológicos. Epidemiol. Serv. Saúde, Brasília, 24(3):565-576, 2015.

[18] SCHMIDT, C. A. P.; TAYANO, P. D.; SANTOS, J. A. A.; MARUJO, L.; PROENÇA, G. G. Previsões Estatísticas com base em Séries Temporais da Cultura da Soja no Brasil. Revista Técnico-Científica do CREA-PR - 24ª edição, 2020.

[19] MATTOZO, T. C. Análise de desempenho de vendas em telecomunicações utilizando técnicas de mineração de dados. Dissertação de Mestrado. UFRN, Natal; 2007.

[20] LAUREANO, R. M. S.; CAETANO, M.; CORTEZ, P. Previsão de tempos de internamento num hospital português: aplicação da metodologia CRISP-DM. RISTI, Nº 13, 06; 2014.

[21] PÁDUA, A. F. L. O.; SOUSA, F. A. Metodologia CRISP-DM: Potencialidades na Descoberta do Conhecimento em Dados Educacionais. XVI Congresso Internacional de Tecnologia na Educação; 2018.

[22] CHAPMAN, P. *ET AL*. CRISP-DM 1.0: Step - by-step data mining guide; 2000.

[23] AMBIMA. Feriados nacionais. Acesso em 20 de julho de 2020. Disponível em: https://www.anbima.com.br/feriados/fer_nacionais/2018.asp.

[24] AMIRUZZAMAN, M. Prediction of Traffic-Violation Using Data Mining Techniques. Advances in Intelligent Systems and Computing, 283–297; 2018.

[25] SHIAU *ET AL*. The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents. Journal Mathematical Problems in Engineering, 2015.

[26] CUENCA, L. G., ALIANE, N., PUERTAS, E., & ANDRES, J. F. Traffic Hotspots Visualization and Warning System. IEEE International Conference on Vehicular Electronics and Safety (ICVES), 2018.

[27] PREFEITURA DO RECIFE. Feriados municipais. Acesso em 20 de julho de 2020. Disponível:www2.recife.pe.gov.br/servico/feriados municipais.

[28] GRETL. Gnu Regression, Econometrics And Time-Series Library. Acesso em 15 de ago. 2020. Disponível em: http://gretl.sourceforge.net.

[29] SATO, RENATO CESAR. Gerenciamento de doenças utilizando séries temporais com o modelo ARIMA. Einstein (São Paulo), 11(1), 128-131; 2013.

[30] WIEMER, H.; DROWATZKY, L.; IHLENFELDT, S. Data Mining Methodology for Engineering Applications (DMME) — A Holistic Extension to the CRISP-DM Model. Appl. Sci. 9(12), 2407; 2019.

DOI: 10.25286/repa.v6i3.1679