Um Sistema para Predição de Risco de Acidentes em Rodovias de Pernambuco

A System for Predicting Accident Risk on Highways of Pernambuco

Rodrigo S. Sousa [1] orcid.org/0000-0002-2471-2446

Danilo R. B. de Araújo [2] orcid.org/0000-0002-4822-0390

Victor Mendonça de Azevedo [3] orcid.org/0000-0003-2943-4622

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Departamento de Computação, Universidade Federal Rural de Pernambuco, Pernambuco, Brasil,
[3] Fundação para Inovações Tecnológicas, Pernambuco, Brasil.

E-mail do autor principal: Rodrigo Sousa rss10@poli.br

Resumo

As estatísticas dos acidentes de trânsito são uma preocupação mundial e trazem grandes prejuízos à sociedade, tanto econômicos quanto sentimentais. Modelos de aprendizado de máquina aplicados à predição de acidentes tem o potencial de servir como ferramenta no auxílio de decisão e melhorar a precisão e o impacto de medidas para a redução de acidentes. Este trabalho tem como proposta desenvolver um sistema visual e interativo de predição de acidentes com o objetivo de auxiliar o processo de decisão de agentes rodoviários federais em Pernambuco. O sistema é composto por um modelo de aprendizagem de máquina treinado com dados de ocorrências de acidentes disponibilizados pela Polícia Rodoviária Federal e uma ferramenta interativa de visualização dos pontos com maiores riscos no mapa de Pernambuco. Foram aplicados modelos de regressão para a predição do número de acidentes dado a identificação da rodovia, altura do trecho, ano, mês, dia da semana e condições meteorológicas. O modelo Random Forest apresentou os melhores resultados de acordo com as métricas de avaliação consideradas no trabalho.

Palavras-Chave: Predição de acidentes; Generalized Linear Model; Random Forest.

Abstract

Traffic accident statistics are a worldwide concern and bring great damage to society, both economic and sentimental. Machine learning models applied to accident prediction have the potential to serve as a tool in decision making and to improve the accuracy and impact of accident reduction measures. This paper aims to develop a visual and interactive accident prediction system to help the decision-making process of federal road agents in Pernambuco. The system consists of a machine learning model trained with accident data provided by Brazil's Federal Highway Police and an interactive tool for viewing the riskiest points on the map of Pernambuco. Regression models were applied to predict the number of accidents given the identification of the highway, section, year, month, day of the week and weather conditions. The Random Forest model presented the best results according to the evaluation metrics considered in the study.

Key-words: *Accident prediction; Generalized Linear Model; Random Forest.*

DOI: 10.25286/repa.v5i2.1328

1 INTRODUÇÃO

Aproximadamente 1,3 milhões de pessoas morreram em estradas no mundo todo apenas em 2016 [1]. No mesmo ano, ocorreram 96 mil acidentes com mais de 6 mil vítimas fatais nas rodovias federais brasileiras. No ano seguinte, em 2017, 20% dos acidentes registrados no Brasil ocorreram na região nordeste, destes, 18% ocorreram em estradas do estado de Pernambuco, o que fez o estado ficar atrás apenas da Bahia em número de acidentes na região [2].

O Brasil possui uma extensa malha rodoviária, aproximadamente 120.539 quilômetros de rodovias federais segundo dados de 2018. O país depende majoritariamente do transporte rodoviário para operações logísticas, sendo a malha ferroviária ainda muito pouco desenvolvida. Uma pesquisa da Fundação Dom Cabral [3] mostra que, em 2017, cerca de 76% do escoamento da produção brasileira foi realizado por meio de caminhões cruzando grandes extensões de rodovias federais no Brasil. Apenas 24% da produção é transportada por outros meios de transporte como trens, navios e aviões.

A estagnação do transporte ferroviário no Brasil [4] contribui para que um grande número de veículos pesados eleve o risco nas estradas quando estes compartilham as mesmas vias que outros atores muito mais vulneráveis, como é o caso de pedestres, ciclistas e motociclistas, que representam mais da metade das vítimas fatais no mundo [1].

O alto número de acidentes nas estradas é naturalmente convertido em um alto custo econômico para a sociedade. Em 2017, o prejuízo foi estimado em mais de 11 bilhões de reais. Sem contar os danos traumáticos irreparáveis que acometem as vítimas e familiares. O número de acidentes e os impactos decorrentes são vastos e especialmente preocupantes. Tal preocupação levou a Organização Mundial da Saúde (OMS) a elaborar um plano de ação para melhorar a segurança nas estradas no mundo todo [5], com atenção especial destinada às nações mais pobres, que possuem apenas 1% dos veículos do mundo, mas contribuem para 13% do total global de mortes.

As duas maiores causas de acidentes e mortes em estradas no Brasil identificadas entre os anos de 2007 e 2016 foram decorrentes do desrespeito às normas de trânsito e falta de atenção ao volante, com percentuais de 30,3% e 23,4%, respectivamente [6]. O fato de maior parte das causas ter relação com atitudes do condutor mostra como o reforço e fiscalização das leis de trânsito é importante para garantir a segurança nas estradas brasileiras.

A Policia Rodoviária Federal (PRF) é responsável por realizar o patrulhamento de grande parte da malha rodoviária brasileira e promove diversas ações educativas e de fiscalização nas estradas. A atuação do agente rodoviário é muito importante para a manutenção da segurança nas estradas. É observado que a atuação do agente de segurança pública apresenta impacto positivo relevante na redução de acidentes e fatalidades no trânsito [7, 8]. Não apenas fiscalizações diretas, mas a simples presença ostensiva do agente de segurança já é capaz de oferecer uma ajuda significativa e colaborar para a diminuição de condutas inapropriadas por parte dos condutores [9]. O trabalho da PRF é significativo, mas os recursos destinados à instituição são limitados. É muito importante garantir uma aplicação de recursos otimizada, priorizando os pontos de maior risco para proporcionar o maior retorno positivo possível.

No momento, o operacional diário da PRF na escolha dos pontos de atuação e fiscalização é realizado manualmente pelos agentes rodoviários. Todos os dias um policial rodoviário federal é responsável por decidir as rodovias e trechos dentro da sua jurisdição que serão destinados às ações de fiscalização e gerar um documento com o roteiro do dia, contendo pontos e horários. Um processo de decisão que é guiado primariamente pela experiência do policial.

Técnicas de aprendizado de máquina já foram aplicadas a problemas relacionados à predição de acidentes e demonstraram resultados promissores. Um modelo baseado em GLM foi aplicado na previsão de acidentes em vias urbanas de três tipos: retas, junções e rotatórias [10]. Outra aplicação de GLM foi usada para prever o número de acidentes por ano em estradas da Itália, utilizando dados do fluxo de tráfico anual, comprimento do segmento da estrada e um conjunto de fatores relacionados ao acidente [11]. Diferentes modelos de regressão foram aplicados em [12] utilizando uma quantidade restrita de variáveis independentes. Uma rede neural artificial foi modelada para a predição de acidentes em um estudo de caso de estradas em uma região limitada da Turquia [13].

Tendo em vista a situação da segurança nas estradas brasileiras e o impacto positivo da PRF, o

objetivo deste trabalho é desenvolver um sistema para predição do risco de acidentes em trechos de rodovias federais brasileiras, mais especificamente limitando-se às rodovias federais do estado de Pernambuco como estudo de caso. A PRF vem coletando dados de ocorrências de acidentes em rodovias federais desde 2007 e já possui um vasto banco de dados com mais de 1,7 milhões de ocorrências. Todas coletadas entre 2007 e setembro de 2019 no Brasil inteiro e publicamente disponibilizadas no portal de dados abertos da PRF.

A ideia é ter um sistema que utiliza modelos de aprendizado de máquina alimentados por dados de acidentes ocorridos em anos anteriores capaz de prever a frequência de acidentes ou risco potencial de acidentes para um dado trecho de uma rodovia federal: Produzindo assim um ranqueamento de locais, priorizando aqueles com maior potencial de risco. O sistema proposto tem por finalidade contribuir com o processo de decisão dos agentes da PRF, fornecendo predições em uma ferramenta visual e interativa.

O restante deste artigo está organizado da seguinte forma: A Seção 2 apresenta o referencial teórico sobre o problema de predição de acidentes e modelos usados na literatura. A Seção 3 apresenta a solução proposta e discute o tratamento dos dados e construção do modelo. A Seção 4 apresenta os experimentos realizados e resultados obtidos. A ferramenta desenvolvida é descrita na Seção 5. Por último, a Seção 6 apresenta a conclusão e os trabalhos futuros.

2 REFERENCIAL TEÓRICO

A análise de acidentes é um tema bastante investigado na literatura. Na maioria dos trabalhos relacionados as variáveis independentes mais comumente consideradas para a construção dos modelos são o fluxo de veículos e as características geométricas da estrada **[10, 14, 15, 16]**. As características geométricas das rodovias têm uma importância especial em trabalhos onde o objetivo é o de identificar potenciais melhorias para o desenho e construção de estradas mais seguras **[16]**. Já no contexto da predição do risco de acidentes, dados sobre o fluxo de veículos são considerados como um componente de vital importância para os modelos **[15]**. Os dados de sensores no veículo e dados comportamentais do motorista são utilizados em previsões de risco de acidentes em tempo real e na previsão da gravidade de acidentes **[17, 18, 19]**.

Considerando o problema de predição do número de acidentes em específico, geralmente é esperado que a contagem de acidentes siga algum tipo de distribuição de Poisson, uma distribuição discreta e não negativa; este costuma ser um problema para um modelo linear comum, o qual assume que os dados seguem uma distribuição normal. Dessa forma, muitos trabalhos focam na aplicação de algum modelo linear generalizado (*Generalized Linear Model* - GLM) baseado nas distribuições de Poisson ou na distribuição negativa binomial **[10, 11]**.

Uma desvantagem desse tipo de abordagem é também a grande dependência no tipo de distribuição dos dados; caso a distribuição assumida venha a ser violada o modelo pode levar a previsões errôneas. Dessa forma, modelos não lineares e que não assumem uma distribuição específica podem ser usados como alternativa e apresentam resultados satisfatórios **[13, 20, 21]**.

3 PREDIÇÃO DE RISCO DE ACIDENTES

O objetivo deste trabalho é desenvolver um sistema visual e interativo para predição de risco de acidentes em rodovias federais do estado de Pernambuco usando dados abertos oficiais disponibilizados pela PRF. Como observado na Seção 2, os dados sobre o fluxo de veículos nas rodovias é uma variável usada amplamente nos modelos para predição da frequência de acidentes. No entanto, informações do volume de fluxo de rodovias são raras e esparsas **[15]**. O sistema proposto neste trabalho deve atender a um amplo número de rodovias para que venha a ser utilizado de maneira efetiva na redução de acidentes. Dessa forma, o modelo desenvolvido deve abordar o problema utilizando apenas dados de fácil obtenção.

A ideia é que o sistema seja utilizado para identificar *blackspots* de acidentes dado um conjunto de fatores fornecidos pelo usuário e servir de suporte para a definição de operações futuras da PRF nas rodovias do estado.

3.1 Descrição dos dados

Todos os dados utilizados neste trabalho foram obtidos a partir do portal de dados abertos da PRF[1]. No portal da PRF é possível fazer o download do conjunto de dados de ocorrências de acidentes em

[1]https://portal.prf.gov.br/dados-abertos-acidentes

DOI: 10.25286/repa.v5i2.1328

rodovias federais separados por ano. Estão disponíveis dados dos anos de 2007 até 2019, dada a data que este trabalho foi desenvolvido o conjunto de dados do ano de 2019 não se encontra completo, contendo ocorrências apenas até o mês de setembro do ano vigente.

Os dados são fornecidos no formato *comma-separated values* (CSV), com colunas separadas por ponto e vírgula (;). Cada linha do arquivo corresponde a uma ocorrência de acidente documentada pela PRF. Em cada ocorrência tem-se um relatório com informações locais da rodovia, trecho em km, clima, data e hora, fase do dia, número de pessoas e veículos envolvidos no acidente, número de vítimas mortas, feridas e ilesas, a causa, o tipo e a classificação do acidente, entre outras informações. Também é importante notar que a partir de 2017 informações da geolocalização dos acidentes passaram a ser incluídas nos conjuntos de dados disponibilizados.

No mesmo portal é possível consultar dicionários de dados com mais informações dos campos e variáveis. No entanto, as informações de alguns campos estão desatualizadas ou inconsistentes, esses problemas são melhor discutidos na seção seguinte.

3.2 Tratamento dos dados

Após uma exploração inicial dos dados uma quantidade considerável de problemas e inconsistências foram encontradas. Como os conjuntos de dados são disponibilizados separadamente por ano de coleta e o esquema dos campos e variáveis sofreu algumas alterações entre os anos, a maioria dos passos de limpeza foram realizados separadamente por ano antes de agrupar todos os dados em um único conjunto contendo todas as ocorrências de 2007 até setembro de 2019. Todos os passos de limpeza realizados neste trabalho são listados e descritos a seguir:

1 Remoção de linhas com valores faltantes;
2 Remoção de ocorrências com identificadores duplicados;
3 Remoção de colunas redundantes;
4 Verificação da consistência dos campos referentes ao número de vítimas;
5 Remoção de espaços em branco nos valores de colunas categóricas;
6 Padronização dos valores categóricos;
7 Correção do ponto decimal e valores do campo "km";
8 Inferir classificação do acidente a partir do total de vítimas ou feridos.

No primeiro passo foram removidas linhas com valores faltantes que não poderiam ser facilmente inferidos a partir das demais informações presentes em outros campos ou linhas. No passo 2 foram removidas as linhas cujo campo identificador estava duplicado dentro do mesmo ano. Ainda que a linha completa não estivesse duplicada, foi observado que elas possuíam a maioria das informações repetidas, portanto apenas a primeira ocorrência foi mantida. Além da repetição dentro do mesmo ano, ocorriam repetições de identificadores com amostras de anos diferentes, assim, no passo 3 foram removidos o campo de identificação juntamente com outros campos com informações consideradas redundantes ou não aproveitáveis para o modelo e ferramenta propostos neste trabalho, como é o caso dos campos de data e horário que foram combinados em um único campo, e os campos com informações consideradas posteriores ao acidente, como é o caso do número de veículos envolvidos, causa do acidente, etc.

No passo 4 foram removidas todas as colunas com inconsistências nos valores de pessoas envolvidas no acidente, número de mortos, feridos graves, feridos leves e ilesos. A coluna de pessoas deve ter como valor o somatório das colunas de feridos, mortos, ilesos e ignorados; e a coluna de feridos deve ser o somatório do número de feridos graves e feridos leves. Todas as linhas com valores inconsistentes foram removidas. Os passos 5 e 6 tratam das variáveis categóricas, removendo espaços em branco nas *strings* e padronizando o uso de acentuação e letras maiúsculas nos valores assumidos por cada campo.

No passo 7 foram tratadas todas as entradas com vírgula no lugar do ponto decimal para os valores numéricos. Por último, no passo 8 os valores faltantes do campo de classificação do acidente foram inferidos a partir das informações nos campos com os números de vítimas, sempre que possível.

3.3 Preparação dos dados

Os dados são agrupados visando contabilizar a frequência de ocorrências de acidentes dado um determinado nível de granularidade espacial, na forma da identificação da BR e o tamanho do trecho em quilômetros; e granularidade temporal, na separação de anos, meses e dias. Além da combinação de outros fatores

relacionados ao acidente, como é o caso dos fatores meteorológicos. A respectiva contabilização total das ocorrências dado o agrupamento escolhido é atribuída a cada uma das amostras pertencentes aos grupos e o valor é usado como alvo, ou variável dependente, do modelo de aprendizado. Dessa forma, o modelo pode ter sua precisão espacial ou temporal calibrada a partir das escolhas de variáveis dependentes durante a preparação dos dados.

Para este trabalho foi considerado a repartição de cada BR em trechos de aproximadamente 37km, as ocorrências foram então agrupadas espacialmente dentro de seus respectivos trechos, além de serem agrupadas por ano, mês, dia da semana e informações da condição meteorológica relatadas na ocorrência do acidente. Dessa forma tem-se uma distribuição de frequências de acidentes espalhados pela extensão das BRs do estado, dada uma certa combinação de fatores. A distribuição aqui discutida, considerando o conjunto de dados completo, pode ser vista na Figura 2.

Para a geração do modelo o número de intervalos de kms pode ser facilmente configurado, aumentando ou diminuindo a granularidade espacial da distribuição da frequência dos acidentes.

Por fim, as variáveis categóricas, dia da semana e condição meteorológica, são transformadas em vetores binários através de *one-hot-encoding*.

Figura 1: Visualização da geolocalização dos *hotspots* de risco do conjunto de dados completo.

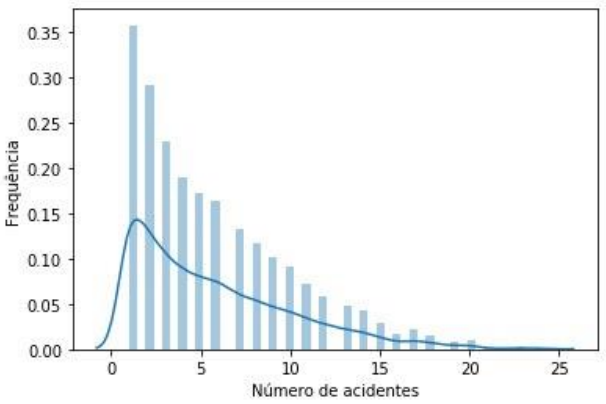

Figura 2: Distribuição da frequência de acidentes para o conjunto de dados completo.

3.4 Modelos

Para os dados e o problema tratados neste trabalho foram considerados exclusivamente modelos de regressão. Primeiro foi considerada a aplicação de um modelo linear generalizado (*Generalized Linear Model* - GLM), que é um modelo simples e bastante utilizado em aplicações dentro do mesmo contexto **[10, 11]**.

Assim como o nome sugere, GLM **[22]** é uma generalização de um modelo de regressão linear comum. O GLM é mais flexível no sentido de possibilitar a definição de distribuições arbitrárias para a modelagem do erro. A distribuição de Poisson foi utilizada para o problema pois é normalmente recomendada para dados onde a variável dependente é uma distribuição de contagens. A implementação do GLM utilizada foi a disponibilizada pelo pacote statsmodels **[23]**, versão 0.10.1.

Posteriormente foram aplicados modelos mais complexos, que também são bastante utilizados em diversos problemas de aprendizado de máquina e que apresentam bons resultados, como é o caso dos modelos baseados em *ensembles* de árvores de decisão **[24]**, Random Forests (RF) **[25]** e XGBoost **[26]**.

Os modelos de árvore de decisão constroem uma estrutura em forma de árvore com base em testes realizados com as observações de treinamento. Modelos baseados em árvore são modelos relativamente simples e que costumam conseguir bons resultados. A árvore de decisão utilizada neste trabalho está inclusa no pacote scikit-learn **[27]**, versão 0.21.

Um ensemble **[28]** é definido como um conjunto de modelos simples. O ensemble pode ser homogêneo, composto apenas por modelos do mesmo tipo, ou heterogêneo, composto por diferentes tipos de modelos. Em um ensemble os dados de treinamento podem ser divididos entre

DOI: 10.25286/repa.v5i2.1328

os modelos individuais com o intuito de gerar modelos especialistas em diferentes subconjuntos dos dados. Durante a predição, todos os modelos individuais são consultados e a predição final é obtida levando em conta alguma política de agrupamento, alguns exemplos são o voto majoritário e o voto ponderado de acordo com as probabilidades da predição de cada modelo individual pertencente ao ensemble. Os ensembles testados neste trabalho foram Random Forest, disponível no pacote scikit-learn **[27]**, versão 0.21; e XGBoost, disponível no pacote xgboost **[26]**, versão 0.9.

XGBoost é um modelo que utiliza um ensemble de árvores de decisão treinadas através de uma técnica chamada gradient boosting. A técnica de boosting é bastante utilizada em ensembles e consiste em criar uma sequência de modelos que são alimentados com os erros dos modelos precedentes com o objetivo de minimizar o erro total do conjunto de modelos. O gradient boosting segue o mesmo princípio do boosting com a adição de um algoritmo de gradiente descendente para realizar a otimização do erro.

4 EXPERIMENTOS E RESULTADOS

Os experimentos foram realizados a partir de um conjunto de dados com ocorrências de acidentes apenas em rodovias federais do estado de Pernambuco. O conjunto de dados de Pernambuco foi dividido temporalmente em um conjunto de treinamento com dados dos anos

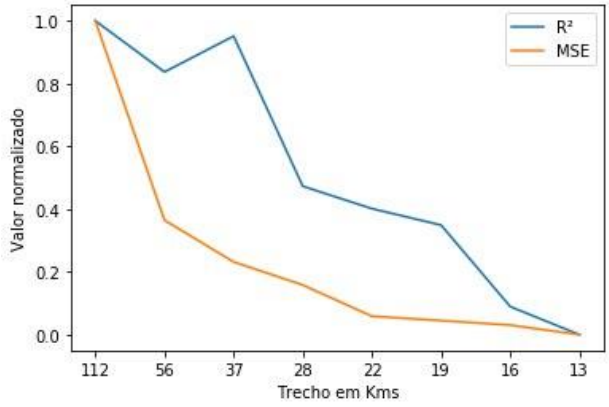

Figura 3: Evolução das métricas R2 e MSE para diferentes trechos.

2007 até 2017 e um outro conjunto de teste com ocorrências a partir do ano 2018 até o final do mês de setembro de 2019. Em termos de proporção, o conjunto de treinamento tem um total de 65961 ocorrências, ~93% do total, enquanto o conjunto de teste possui 4712 amostras, aproximadamente 7% do total.

A escolha dessa divisão de conjuntos de treino e teste em especifico é justificada pelo funcionamento da ferramenta proposta. O objetivo é desenvolver um modelo com treinamento *online* que realize previsões fazendo uso da maior quantidade de dados atualizados possível, portanto, a divisão foi feita de forma que houvessem dados de ao menos um ano completo para testes.

Os modelos testados neste trabalho foram os modelos: GLM, baseado na distribuição de Poisson; árvore de decisão, XGBoost e Random Forest. Todos os modelos baseados em árvores, tiveram a profundidade máxima das árvores limitada a 10. Os modelos de ensemble foram configurados para usar 100 estimadores. Além dos modelos citados, um modelo simples baseado apenas na média da distribuição do conjunto de treinamento foi adicionado para fins de comparação.

A avaliação dos modelos é feita a partir de duas métricas principais, o coeficiente de determinação (*R-squared* - R2) e o erro médio quadrático (Mean Squared Error - MSE). O R2 é uma métrica que indica o nível de ajuste das predições do modelo aos valores reais do alvo. O R2 pode assumir valores menores que zero, para modelos arbitrariamente ruins; e valor máximo igual a 1, que indica que o modelo possui um ajuste perfeito.

Todos os modelos foram testados com diferentes valores de trecho da rodovia. A Figura 2 mostra um gráfico com a evolução dos valores normalizados, R^2 e MSE, do modelo Random Forest usando dados preparados com diferentes comprimentos de trechos. É possível identificar quedas acentuadas da métrica R^2, desproporcionais às variações do tamanho do trecho. Por outro lado, os valores de MSE apresentam uma redução suave e quase contínua. Embora a correlação de R2 com o tamanho do trecho não seja linear, é fácil perceber o compromisso entre a métrica R2 e a escolha de usar trechos menores para realizar previsões mais localizadas. Considerando os valores de R2 e MSE da Figura 2, os experimentos deste trabalho usaram dados com trechos de aproximadamente 37km pois fornecem uma precisão espacial razoável com valores satisfatórios para as métricas.

Tabela 1: Resultados dos experimentos.

Modelos	R^2	MSE	MAE
Modelo base	-1.59	5.09	2.09
GLM(Poisson)	-1.74	5.39	1.76
Árvore de decis	0.34	1.28	0.78
XGB	0.23	1.50	0.81
Random Forest	0.41	1.15	0.76

Os resultados dos experimentos foram compilados na Tabela 1. É possível perceber que os modelos base e GLM(Poisson) tiveram valores de R2 menores que 0, isso indica que ambos modelos tiveram um ajuste pior que o ajuste proporcionado por uma simples linha horizontal traçada na média da distribuição das variáveis dependentes do conjunto de teste. Dos modelos simples apenas a árvore de decisão apresentou um bom desempenho, superando até mesmo o ensemble XGBoost tanto na métrica R2 quanto nas medidas de erro.

Por fim, o modelo com os melhores resultados em todas as métricas foi o ensemble Random Forest, ele obteve o melhor ajuste ao conjunto de teste e também apresentou o menor erro. Na Figura 4 tem-se o histograma dos resíduos das predições do modelo Random Forest. A distribuição se encontra bem agrupada em torno de zero, confirmando o erro baixo; porém levemente deslocada para os valores positivos, o que indica que o modelo tende a superestimar as predições.

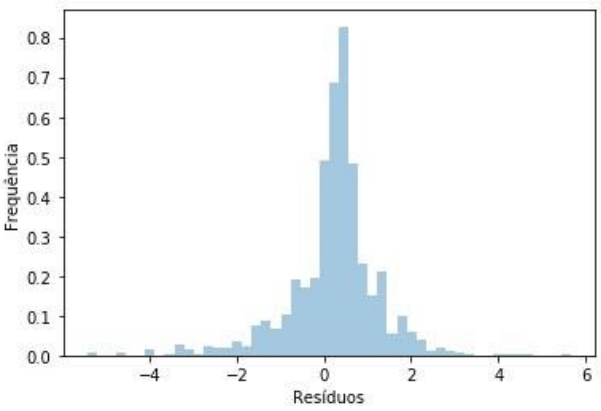

Figura 4: Histograma dos resíduos do modelo Random Forest.

5 FERRAMENTA

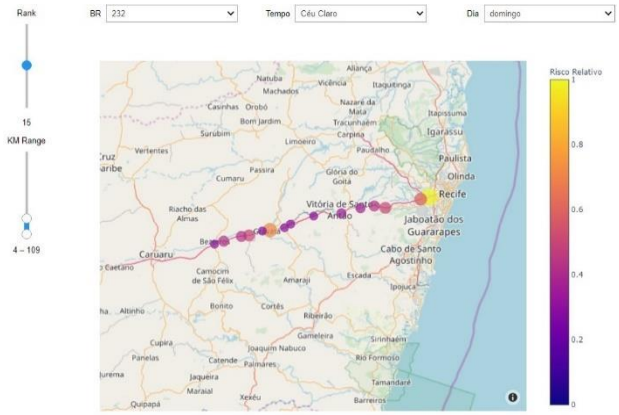

Figura 5: Interface gráfica desenvolvida para facilitar a análise por parte dos agentes públicos responsáveis.

A ferramenta desenvolvida tem foco na visualização intuitiva e objetiva do risco previsto pelo modelo de aprendizado de máquina. Como pode ser visto na Figura 5, a interface permite que o usuário escolha uma combinação de fatores e o modelo treinado retorna as previsões de acidente para as entradas passadas. Os resultados são exibidos no mapa de acordo com a geolocalização mediana de cada trecho dentro da rodovia. Os trechos podem ser facilmente comparados e priorizados de acordo com a forma que são plotados no mapa: cada ponto tem seu nível de risco relativo representado visualmente por meio do tamanho do ponto e cor na escala da barra de *heatmap*. Os pontos com maior tamanho e cor mais clara são os pontos com maior risco dentro de uma rodovia.

Além do mapa visual, o usuário pode obter as informações dos pontos retornados clicando nos mesmos ou checando um painel que lista todos os pontos já ranqueados para rápida visualização. O objetivo final é ter um sistema leve e de fácil utilização por parte dos agentes rodoviários; com a visualização dos dados feita através de um *frontend* capaz de ser acessado em qualquer aparelho *smartphone*, enquanto o modelo reside em um servidor central onde é treinado de maneira online e atualizado em tempo real com dados de novas ocorrências.

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foi relatada a aplicação de modelos de aprendizado de máquina para desenvolvimento de um sistema de predição de risco de acidentes em rodovias de Pernambuco. O modelo segue uma preparação dos dados que permite a configuração da granularidade espacial

DOI: 10.25286/repa.v5i2.1328

das predições de acordo com o tamanho de trecho de rodovia considerado. Além disso, os modelos analisados foram treinados com dados abertos de fácil acesso. Embora exista uma tendência dos modelos em superestimar as previsões, os experimentos mostraram resultados satisfatórios para o problema, especialmente considerando o erro obtido.

Para trabalhos futuros seria interessante a adição de mais dados com informações locais dos trechos, como por exemplo o limite de velocidade, condição da estrada, proximidade de área de fiscalização ou pedágio, etc. Além disso, um estudo da influência das operações da PRF baseadas nas previsões do modelo poderia fornecer um método de auto-ajuste para o sistema.

REFERÊNCIAS

[1] **Global status report on road safety 2018: summary.** Geneva: World Health Organization; 2018 (WHO/NMH/NVI/18.20). Licence: CC BY-NC-SA 3.0 IGO).

[2] **Ministério dos Transportes, Anuário Estatístico de Segurança Rodoviária**, Ministério dos Transportes, 2018. Disponível em: <http://portaldaestrategia.infraestrutura.gov.br/publicacoes/item/42-anuário-estatístico-2010-2017.html>. Último acesso em 23/11/2019.

[3] Resende, P. **Custos logísticos no Brasil 2017,** Fundação Dom Cabral, 2018. Disponível em: <https://www.fdc.org.br/conhecimento/publicacoes/relatorio-de-pesquisa-33324>. Último acesso em 23/11/2019.

[4] DA SILVA, B. L., SARMENTO, T. A., DA SILVA SANTOS, V. É. e TAVARES, F. B. R., **CRISE PETROLÍFERA E O DESCASO FERROVIÁRIO: DA DEPENDÊNCIA AO COLAPSO.** Revista da Universidade Vale do Rio Verde, 17(1), 2019.

[5] SMINKEY, L. Global **Plan for the Decade of Action for Road Safety 2011-2020.** World Health Organization www who int/roadsafety/decade_of_action (2011).

[6] Ministério dos Transportes, **Avaliação das Políticas Públicas de Transportes**, Ministério dos Transportes, 2018. Disponível em: <https://www.infraestrutura.gov.br/component/content/article/113-politica-e-planejamento-de-transportes/7385-apt.html>. Último acesso em 23/11/2019.

[7] GONÇALVES, D.P.; SANCHES, C.P. **A Educação e O Policiamento no Trânsito de Goiânia.** 2019.

[8] PACHECO, E.L. **A redução de acidentes por meio da educação para o trânsito.** 2017.

[9] CASTILLO-MANZANO, J. I. et al. **From legislation to compliance: The power of traffic law enforcement for the case study of Spain.** Transport policy 75 (2019): 1-9.

[10] GIANFRANCO, F.; SODDU, S.; FADDA, P. **An accident prediction model for urban road networks.** Journal of Transportation Safety & Security, v. 10, n. 4, p. 387-405, 2018.

[11] LA TORRE, F. et al. **Development of an accident prediction model for Italian freeways.** Accident Analysis & Prevention, v. 124, p. 1-11, 2019.

[12] SHARMA, B.; KATIYAR, V. K.; KUMAR, K**. Traffic accident prediction model using support vector machines with Gaussian kernel.** Proceedings of fifth international conference on soft computing for problem solving. Springer, Singapore, 2016. p. 1-10.

[13] YASIN ÇODUR, M.; TORTUM, A. **An artificial neural network model for highway accident prediction: A case study of Erzurum, Turkey**. PROMET-Traffic&Transportation, v. 27, n. 3, p. 217-225, 2015.

[14] ABDEL-ATY, M. A.; RADWAN, A. E. **Modeling traffic accident occurrence and involvement.** Accident Analysis & Prevention, v. 32, n. 5, p. 633-642, 2000.

[15] RYDER, B. et al. **Spatial prediction of traffic accidents with critical driving events–Insights from a nationwide field study.** Transportation research part A: policy and practice, v. 124, p. 611-626, 2019.

[16] CHEN, S. et al. **Safety sensitivity to roadway characteristics: a comparison across highway classes.** Accident Analysis & Prevention, v. 123, p. 39-50, 2019.

[17] ZHAO, H. et al. **Research on Traffic Accident Prediction Model Based on Convolutional Neural Networks in VANET.** 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE, 2019. p. 79-84.

[18] RYDER, B. et al. **Spatial prediction of traffic accidents with critical driving events–Insights**

from a nationwide field study. Transportation research part A: policy and practice, v. 124, p. 611-626, 2019.

[19] SHAON, M. R. R. et al. **Incorporating behavioral variables into crash count prediction by severity: A multivariate multiple risk source approach.** Accident Analysis & Prevention, v. 129, p. 277-288, 2019.

[20] CHANG, L.-Y. **Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network.** Safety science, v. 43, n. 8, p. 541-557, 2005.

[21] CHANG, L.-Y.; CHEN, W.-C. **Data mining of tree-based models to analyze freeway accident frequency.** Journal of safety research, v. 36, n. 4, p. 365-375, 2005.

[22] NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized linear models.** Journal of the Royal Statistical Society: Series A (General), v. 135, n. 3, p. 370-384, 1972.

[23] SEABOLD, S.; PERKTOLD, J. **Statsmodels: Econometric and statistical modeling with python.** Proceedings of the 9th Python in Science Conference. Scipy, 2010. p. 61.

[24] ROKACH, L.; MAIMON, O. Z. **Data mining with decision trees: theory and applications.** World scientific, 2008.

[25] BREIMAN, L. **Random forests.** Machine learning, v. 45, n. 1, p. 5-32, 2001.

[26] CHEN, T.; GUESTRIN, C. **Xgboost: A scalable tree boosting system.** Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016. p. 785-794.

[27] PEDREGOSA, F. et al. **Scikit-learn: Machine learning in Python.** Journal of machine learning research, v. 12, n. Oct, p. 2825-2830, 2011.

[28] DIETTERICH, T. G. et al. **Ensemble learning.** The handbook of brain theory and neural networks, v. 2, p. 110-125, 2002.

DOI: 10.25286/repa.v5i2.1328