



Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP

IX SAEPRO – 25 e 26 de novembro de 2025

Análise de Acidentes de Trânsito usando Machine Learning

Carlos Eduardo Celestino de Oliveira

Universidade Federal do Vale do São Francisco

(carlos.celestino@discente.univasf)

Thiago Magalhães Amaral

Universidade Federal do Vale do São Francisco

(thiago.magalhaes@univasf.edu.br)

CONTEXTUALIZAÇÃO: De acordo com Azambuja (2022), o avanço das tecnologias e a ampliação do acesso a dados de pesquisa no Brasil têm estimulado o uso de ferramentas computacionais para apoiar a tomada de decisão em diversas áreas, incluindo o setor público. O projeto traz um uso importante de dados colhidos e disponíveis na plataforma Kaggle, mostrando como o estado pode encontrar uma importante utilidade para esse dados, tendo em vista que existem ferramentas como o próprio Orange que permite para eles fazerem uma análise de maneira rápida e intuitiva, e por consequência tomar conhecimento amplo das casualidades que vem a causar determinados problemas de interesse público, tornando possível aplicar tentativas de soluções para diminuir acidentes de trânsito, e ou pelo menos a ocorrência de acidentes graves, tendo como base esse trabalho que vem com foco em prever a severidade de tais acidentes. O trabalho demonstra como o uso de algoritmos de aprendizado de máquina pode auxiliar na previsão da gravidade de acidentes a partir de variáveis descritivas como idade, sexo e tipo de habilitação dos envolvidos, mostrando assim para entidades públicas maneiras de ter diagnósticos eficientes dos problemas de trânsito.



Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP

IX SAEPRO – 25 e 26 de novembro de 2025

OBJETIVO: Identificar padrões capazes de prever a severidade de acidentes de trânsito com pessoas, utilizando algoritmos de aprendizado supervisionado implementados na ferramenta Orange. A análise visa demonstrar a aplicabilidade dessas técnicas para auxiliar órgãos públicos e empresas privadas no entendimento dos fatores que influenciam a gravidade dos acidentes.

MÉTODO: O estudo utilizou o dataset ‘Relação de Acidentes de Trânsito 2011–2021’, disponível no Kaggle, com enfoque no arquivo CSV referente ao ano de 2021. Após a importação dos dados, realizou-se o pré-processamento que inclui normalização e codificação de variáveis categóricas, com foco na variável-alvo ‘desc_severidade’. Foram aplicados os algoritmos k-Nearest Neighbors (kNN) e Random Forest, sobre o kNN foi utilizado a estratégia de usar vários valores para k e por sua consequência testar a variação da eficiência entre cada um, sendo esses valores para k de 3, 5, 7, 9, visualizando então se os valores baixos eram mais ou menos eficientes, porém, também para fins comparativos, usou-se o próprio Random Forest configurado entre 100 a 150 árvores visando uma maior eficiência perante a alta quantidade de linhas e colunas do dataset, ou seja, uma grande quantidade de informações. Os modelos foram avaliados com base em métricas de acurácia, precisão, AUC, F1-score e matriz de confusão, por meio dos módulos Test & Score e Confusion Matrix do Orange.

RESULTADOS E DISCUSSÕES: Os resultados indicaram que o modelo kNN apresentou desempenho superior ao Random Forest, alcançando acurácias entre 83% e 87%, e não 81.4% que é o caso do Random Forest. Nota-se que nos kNN com o k mais baixo tinham um resultado muito inferior como o próprio k=3(menor), e a partir de k=5 tinha-se uma melhora significativa em todos os seus dados, mas, no k=9 observa-se uma estabilidade nas métricas, além de uma melhora significativa na qualidade de resultados, o qual todos(incluindo o Random Forest) tinham os seguintes resultados: AUC: entre 87% e 90%; Acurácia: entre 81% e 87%; F1: entre 83% e 85,8%; Precisão: entre 83% e 85,5%;



Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP

IX SAEPRO – 25 e 26 de novembro de 2025

Recall: entre 81,4% e 86%. A priori, todos os resultados estão bons, com margem acima de 80%, apesar de nenhum acima de 90%, mas, são métricas aceitáveis tendo em vista que o MCC indica que a aleatoriedade dos resultados está no limite aceitável, pois, o MCC está entre 66% e 73%, todavia, importante ressaltar que de todos os resultados citados, os melhores, ou seja, os maiores vinham de $k=9$.

CONSIDERAÇÕES FINAIS: A aplicação de ferramentas computacionais para análise de dados mostrou-se eficiente na identificação de padrões em acidentes de trânsito. O modelo kNN apresentou maior estabilidade e desempenho satisfatório, enquanto o Random Forest evidenciou limitações de precisão. Conclui-se que o uso do Orange facilita o entendimento de dados complexos e pode ser expandido para estudos que relacionem variáveis adicionais, como histórico de infrações ou problema de alcoolismo na família, aprimorando o potencial preditivo para facilitar a busca pelo público alvo e as aplicações práticas, mas uma grande variação do uso desses dados, seria fazer uma com dataset de acidentes relacionado a zonas de tráfegos para se criar um fator de periculosidade e ao cruzar com dados deste trabalho, aumentar a fiscalização de embriaguez e CNH apropriada e determinados locais.

PALAVRAS-CHAVE: Aprendizado de Máquina; KNN; Árvore de Decisão; Tráfego; Acidentes.

REFERÊNCIAS:

AZAMBUJA, L. A. *Acesso aberto a dados de pesquisa no Brasil: mapeamento de repositórios, práticas e percepções dos pesquisadores e tecnologias. Ciência da Informação*, Brasília, v. 51, n. 3, 2022.

Disponível em: <https://revista.ibict.br/ciinf/article/view/4958>. Acesso em: 23 out. 2025.



Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP

IX SAEPRO – 25 e 26 de novembro de 2025

BRASIL. Ministério da Infraestrutura. *Política Nacional de Trânsito: diagnóstico e metas*. Brasília, 2021.

Disponível em: <https://www.gov.br/transportes/pt-br/assuntos/transito/arquivos-senatran/educacao/publicacoes/pnt.pdf>. Acesso em: 16 out. 2025.

CAPOMACCIO, S. *Alcool e direção são responsáveis por mais de 10 mil mortes no trânsito no País*. *Jornal da USP*, São Paulo, 5 dez. 2024.

Disponível em: <https://jornal.usp.br/radio-usp/alcool-e-direcao-sao-responsaveis-por-mais-de-10-mortes-no-transito-no-pais/>. Acesso em: 16 out. 2025.

DEPARTAMENTO NACIONAL DE TRÂNSITO (DENATRAN). *Anuário Estatístico de Acidentes de Trânsito*. Brasília: DENATRAN, [s.d.].

Disponível em: https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/diest-arquivos/anuario-2024_final.html. Acesso em: 16 out. 2025.

JORDAN, M. I.; MITCHELL, T. M. *Machine learning: trends, perspectives, and prospects*. *Science*, v. 349, n. 6245, p. 255–260, 2015.

Disponível em: <https://www.science.org/doi/10.1126/science.aaa8415>. Acesso em: 16 out. 2025.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.

Disponível em: <https://www.cs.cmu.edu/~tom/mlbook.html>. Acesso em: 16 out. 2025.



Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP

IX SAEPRO – 25 e 26 de novembro de 2025

PORTAL DATA SCIENCE. *O algoritmo k-Nearest Neighbors (kNN) em Machine Learning.* [S.l.]: Portal Data Science, [s.d.].

Disponível em: <https://portaldatascience.com/o-algoritmo-k-nearest-neighbors-knn-em-machine-learning/>. Acesso em: 15 out. 2025.

PORTAL EBAC. *Random Forest: o que é, onde é utilizado e por que é importante.* [S.l.]: EBAC Online, [s.d.].

Disponível em: <https://ebaonline.com.br/blog/random-forest-seo>. Acesso em: 16 out. 2025.

RAMOS, R. *Relação de acidentes de trânsito 2011–2021* [base de dados]. [S.l.]: Kaggle, 2021.

Disponível em: <https://www.kaggle.com/datasets/ricsonramos/relao-de-acidentes-de-transito-2011-2021>. Acesso em: 27 set. 2025.

WALKER, S. M., II. *What are Accuracy, Precision, Recall, and F1 Score?* [S.l.]: Klu AI, [s.d.].

Disponível em: <https://klu.ai/glossary/accuracy-precision-recall-f1>. Acesso em: 16 out. 2025.