

# Análise de Dados de Tráfego Urbano com ML

Luis Henrique Turra Ramos

18 de grafo de 2025

**Objetivo:** O objetivo central deste projeto foi desenvolver um modelo preditivo capaz de estimar o percentual de lentidão no trânsito da cidade de São Paulo a partir de variáveis operacionais e incidentes registrados em tempo real. Além da modelagem preditiva, buscou-se criar um dashboard interativo que permitisse visualizar tanto os dados reais quanto as previsões do modelo em um mapa e em séries temporais, oferecendo uma ferramenta prática para análise de mobilidade urbana e suporte à tomada de decisão.

**Introdução:** O trânsito de São Paulo é um dos mais congestionados do mundo e sofre influência direta de eventos pontuais como acidentes, veículos quebrados, manifestações e condições climáticas. Compreender como esses fatores impactam a lentidão e antecipar seu comportamento é essencial para planejamento urbano, gestão de frotas e políticas públicas. Este trabalho propõe uma solução completa de ponta a ponta: desde a extração e tratamento dos dados com PySpark até a entrega de um dashboard online 24h com previsões geradas por um modelo de Random Forest otimizado.

### **Metodologia:**

Fontes de Dados:  
<https://archive.ics.uci.edu/dataset/483/behavior+of+the+urban+traffic+of+the+city+of+sao+paulo+in+brazil>

Técnicas de Análise: O dashboard foi construído com Dash e Dash Leaflet, permitindo a visualização temporal da lentidão, comparação entre valores reais e previstos e um heatmap georreferenciado que destaca as regiões de maior intensidade de congestionamento. A solução foi implantada no Render como um serviço web gratuito, com arquitetura híbrida, treino pesado executado

offline com PySpark e inferência leve no ambiente de produção utilizando apenas Pandas e o modelo já treinado.

**Limitações:** O período dos dados é relativamente antigo (2006–2007), o que pode não refletir completamente a dinâmica atual da cidade, as variáveis são majoritariamente binárias e não incluem informações de volume de tráfego ou condições meteorológicas detalhadas, e o modelo foi treinado apenas em dados históricos de São Paulo, o que restringe sua generalização direta para outras cidades.

**Resultados:** O modelo final de Random Forest, com 50 árvores e profundidade máxima de 8, alcançou um RMSE de 3,15 pontos percentuais, MAE de 2,45 e  $R^2$  de 0,55 no conjunto de teste. Esses valores indicam capacidade moderada a boa de prever a lentidão. A análise temporal mostrou picos claros de lentidão nos horários de rush da manhã (7h30–9h) e, principalmente, da tarde/noite (17h–20h), comportamento que o modelo conseguiu reproduzir com precisão razoável. O heatmap gerado a partir das previsões evidencia concentração de lentidão na região central e em corredores principais.

**Discussão:** Embora o  $R^2$  de 0,55 indique que ainda há variância não explicada, o erro médio absoluto inferior a 2,5 pontos percentuais é operacionalmente relevante, em um cenário onde a lentidão média fica em torno de 12–15%, prever com precisão de  $\pm 2,5\%$  já permite antecipar situações críticas e acionar medidas preventivas. A arquitetura híbrida implementada (treino offline com PySpark + inferência leve com Pandas) provou ser robusta e econômica, permitindo que o dashboard rode 24h em um plano gratuito do Render sem qualquer dependência de Java ou Spark em produção.

**Conclusões:** Foi possível construir e colocar em produção um sistema completo de previsão de lentidão no trânsito de São Paulo capaz de operar com dados reais, apresentar resultados visualmente claros e manter disponibilidade contínua na web. O modelo desenvolvido demonstra viabilidade técnica para aplicação prática em monitoramento urbano e pode servir como base para evoluções futuras que incorporem dados mais recentes, variáveis de volume de tráfego e integração com APIs em tempo real. O projeto entregue cumpre

plenamente o objetivo inicial de transformar dados brutos de incidentes em inteligência acionável para a mobilidade urbana.