Previsão do Tempo de Viagem com Aprendizado de Máquina para o Transporte Público por Ônibus em Florianópolis

Leonardo Villamarin de Souza[1], **Thiago Rodrigues da Motta Fagundes**[1], **Tielle da Silva Alexandre**[1], **Flavia Bernardini**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói, RJ – Brasil

{lvillamarin,trmfagundes,tiellesa}@id.uff.br, fcbernardini@ic.uff.br

Abstract. *In the Smart Cities context, planning public transport services is essential in cities with high population density. To this end, having access to Travel Time Prediction (TTP) for bus lines is important for identifying bottlenecks in the city. The TTP has been explored in the literature over the last 2 decades, but there is a significant gap in tools to support scientists to process data from this domain and to build predictive models with Machine Learning (ML). This work presents a simplified process for constructing a travel time predictor, which we model, exploring possible strategies that small city governments can adopt. The process is evaluated using real data from the city of Florianópolis.*

Resumo. *No contexto de cidades inteligentes, o planejamento dos serviços de transportes públicos é fundamental em cidades com elevada densidade demográfica. Para isso, ter acesso à Previsão de Tempo de Viagem (PTV) para linhas de ônibus é importante para a identificação de gargalos na cidade. PTV tem sido explorada na literatura ao longo das últimas 2 décadas, mas há uma lacuna importante de ferramental para apoiar os cientistas de dados a processar dados desse domínio e para a construção de modelos preditores com Aprendizado de Máquina (AM). O objetivo deste trabalho é apresentar um processo simplificado para a construção de um preditor de tempo de viagem, o qual modelamos, explorando possíveis estratégias que podem ser adotadas por prefeituras de pequeno porte. O processo é avaliado por meio de dados reais da cidade de Florianópolis.*

1. Introdução

Atualmente, o transporte público por ônibus desempenha um papel significativo na mobilidade urbana, podendo impactar a qualidade de vida cotidiana, prestando serviços a um número cada vez maior de pessoas [Bahuleyan and Vanajakshi 2016]. A Previsão por Tempo de Viagem (PTV) em rotas de ônibus do transporte público pode ajudar na eficiência e na confiabilidade do sistema de transporte nas cidades. A capacidade de estimar com precisão o tempo de viagem permite que os passageiros planejem suas rotas de forma mais eficiente, reduzindo o tempo de espera e melhorando a experiência geral do usuário [Yamaguchi et al. 2018]. Além disso, as empresas de transporte público podem utilizar essas previsões para otimizar a programação dos ônibus, melhorando sua eficiência operacional e oferecendo viagens de maneira mais adequada às necessidades dos usuários. Transportes públicos de melhor qualidade incentivam

o uso pela população e, consequentemente, diminuem o número de veículos circulando [Agafonov and Yumaganov 2019].

A PTV em rotas de ônibus apresenta desafios devido à sua natureza complexa e à dinâmica do tráfego urbano. Diversos fatores, como congestionamentos, condições climáticas adversas, eventos e acidentes de trânsito, podem afetar os tempos de deslocamento [Bahuleyan and Vanajakshi 2016]. Para construir preditores mais precisos, pode ser necessário integrar múltiplas fontes de dados, realizar uma preparação de dados mais robusta e criar *features* relevantes. Contudo, isso exige tanto mão de obra especializada quanto infraestrutura tecnológica com alto poder de processamento, o que pode estar fora do alcance de prefeituras de menor porte. Nesse cenário, a adoção de estratégias mais simples e compatíveis com máquinas convencionais para a construção de modelos de PTV pode representar uma alternativa promissora, especialmente para municípios que enfrentam restrições orçamentárias e carência de mão de obra qualificada.

Diante disso, este trabalho propõe um processo simplificado que detalha as tarefas necessárias para o processamento de dados e a construção dos modelos de PTV, com foco especial em prefeituras de menor porte. Esse processo contempla estratégias viáveis relacionadas à escolha de algoritmos de Aprendizado de Máquina AM, à otimização dos hiperparâmetros desses algoritmos, à integração de uma fonte adicional de dados e à análise do impacto do uso de tempos de viagem anteriores, que podem contribuir para a criação de modelos de predição mais precisos. Por exemplo, a escolha de algoritmos tradicionais como XGBoost (XGB) e Random Forest (RF) pode ser uma alternativa eficaz às redes neurais profundas, que requerem mais recursos computacionais para treinar os modelos de PTV.

Diversos cenários experimentais são definidos para testar tais estratégias, utilizando a base de dados do sistema de transporte de Florianópolis. Essa base de dados foi escolhida por sua disponibilidade e pela complexidade do tráfego urbano, de modo que, se as estratégias propostas forem eficazes nesse contexto mais desafiador, é possível supor que também serão adequadas em municípios com menor volume de tráfego. Por fim, com base na condução do processo proposto, esperamos responder à seguinte questão de pesquisa: “Quais estratégias (i.e., algoritmos, seleção de atributos) são eficazes para a construção de modelos de PTV especialmente em contextos de prefeituras de pequeno porte ?”. Este trabalho está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica sobre ciência de dados e algoritmos de AM; a Seção 3 discute os trabalhos relacionados sobre PTV; a Seção 4 descreve o processo de construção do modelo; a Seção 5 aplica esse processo a dados reais e analisa os resultados; por fim, a Seção 6 traz as conclusões e sugestões para trabalhos futuros.

2. Ciência de Dados e Modelos de PTV

O processo de ciência de dados é dividido em seis etapas: definição do projeto, recuperação dos dados, preparação dos dados, exploração dos dados, modelagem dos dados e apresentação e automações [Cielen and Meysman 2016]. Este trabalho se concentra nas etapas de preparação, exploração e modelagem dos dados. A preparação dos dados se divide em três subprocessos: limpeza, integração e transformação dos dados. A limpeza dos dados é o foco deste trabalho e envolve a identificação, a correção ou remoção de dados inconsistentes, redundantes, incompletos e ruidosos nos conjuntos de

dados [Faceli et al. 2021]. A integração visa definir estratégias para integrar dados em um único conjunto de dados, considerando as características inerentes de diferentes fontes de dados. A modelagem dos dados é iterativa, que envolve selecionar os atributos para treinamento do modelo, escolher o algoritmo de AM, definir os parâmetros do algoritmo escolhido, treinar e avaliar o modelo [Cielen and Meysman 2016].

Para a análise dos modelos, a técnica *Hold Out* foi adotada na construção dos modelos de PTV, que separa os dados em duas partes: uma de treino e outra de teste [Hastie et al. 2009]. Essa técnica é ideal para o treinamento de dados do tipo séries temporais. Uma série temporal é formada por registros definidos por $x_i, \{i = 1, ...N\}$, sendo N o número total de registros da base de dados da série, e cada valor x_i foi coletado em um tempo t_i. Daí, um registro observado em um tempo t_j não pode ser utilizado para prever valores da série coletados em $t_k, k < j$. Assim, as técnicas de validação cruzada ou de embaralhamento da base de dados original não podem ser utilizadas.

Os modelos de regressão têm como objetivo prever um valor numérico, e uma das formas é encontrar modelos que reduzam o erro médio [Aurélien 2017]. Para o treinamento dos modelos, os seguintes algoritmos foram escolhidos: um algoritmo para construção de um modelo de Regressão Linear (RL), RF e XGB. Um modelo de RL é frequentemente usado como base de comparação com outros modelos de regressão mais robustos. De forma mais geral, um modelo linear faz uma previsão calculando uma soma ponderada dos atributos de entrada, mais uma constante chamada termo de polarização [Aurélien 2017]. Já os algoritmos RF e XGB são da categoria de algoritmos do tipo *ensemble*, que combinam a saída de diferentes estimadores para tomar decisões mais confiáveis, reunindo as saídas obtidas por esses estimadores em uma única previsão [Zhou 2012].

A métrica de avaliação de regressores utilizada neste trabalho foi a medida MAPE (*Mean Absolute Percentage Error*), por se tratar de uma métrica que normaliza os valores preditos com base no valor real. Tal métrica é particularmente interessante para os casos em que o valor a ser predito está em intervalos de valores muito maiores que o intervalo $[0, 1] \in \mathbb{R}$. Considere para cálculo da métrica $MAPE(h, S_t) = \frac{\sum_{i=1}^{N_t} \frac{|y_i - \hat{y}_i|}{y_i}}{N_t}$, um conjunto de dados de teste $S_t = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{N_t}, y_{N_t})\}$ com N_t registros, sendo cada vetor $\mathbf{x}_i$ uma sequência de valores relativos ao conjunto de atributos do conjunto de dados e y_i o valor do atributo alvo (no caso deste trabalho, PTV) associado ao vetor $\mathbf{x}_i$. O previsor ou estimador h a ser avaliado estima cada vetor $\mathbf{x}_i$ com um valor estimado $\hat{y}_i$.

3. Revisão da Literatura

Uma revisão sistemática de literatura foi publicada com o objetivo de apresentar as soluções baseadas em AM propostas especificamente para o transporte público por ônibus, assim como detalhar a modelagem dessas soluções, incluindo tipos de dados e algoritmos de AM utilizados [Alexandre et al. 2023]. As soluções foram agrupadas em quatro categorias: licitação, planejamento operacional, controle operacional e a demanda de ônibus-passageiros. A PTV está contida na categoria de planejamento operacional e na demanda de ônibus-passageiros. A construção de *features* relevantes, a integração de dados e a escolha dos algoritmos de AM usados neste trabalho foram baseadas nos levantamentos feitos por esta revisão da literatura.

A RSL identificou que os algoritmos de AM tradicionais mais usados para construção de modelos de predição são: RF, XGB, *k-Nearest Neighbors*, *Support Vector Regressors* (SVR), as Redes Bayesianas, e outras diversas variações de algoritmos *ensembles*. A integração dos dados climáticos em bases reais de geolocalização, explorada neste trabalho, é considerada uma das abordagens mais comuns encontradas na literatura [Alexandre et al. 2023], além de ser disponibilizada em portais de dados abertos e organizações sem fins lucrativos. O uso de *features temporais* e climáticas usadas neste artigo também emergiu dessa RSL. A seguir, são apresentados outros trabalhos que forneceram suporte às decisões tomadas durante o processo e na condução dos experimentos.

Yamaguchi, Mansur e Tsunenori propuseram comparação de modelos RNA, SVR, XGB, RL e RF para previsão de atraso de ônibus entre paradas consecutivas [Yamaguchi et al. 2018]. Os autores concluíram que os modelos XGB apresentaram os melhores resultados e observaram que, para aprimorar o desempenho de seus modelos, são necessários dados de viagens anteriores de, no mínimo, três semanas. Os autores de [Bahuleyan and Vanajakshi 2016] construíram modelos para PTV usando dados de GPS, para qualquer rota feita dentro de uma malha rodoviária, e dividiram a malha em diversos nós e links. Utilizaram os algoritmos k-NN e RF e uma combinação dos dois. O método que combinou ambos obteve melhor desempenho em relação ao K-NN. Como em diversas prefeituras o recurso computacional pode ser escasso, optamos por algoritmos que consomem menos recursos computacionais e que apresentaram bons resultados, como descrito nos dois trabalhos anteriores. Portanto, os algoritmos RF e XGB são selecionados, juntamente com o RL (base), como os modelos de regressão mais simples.

Além disso, esses estudos relacionados propõem diversas estratégias utilizadas para a construção de modelos para PTV. Neste trabalho, buscamos integrar duas dessas estratégias com o objetivo de realizar um estudo exploratório, visando à construção de bons modelos de PTV para as prefeituras. A primeira estratégia que observamos é a integração de dados climáticos para o treinamento dos modelos de PTV [Agafonov and Yumaganov 2019, Samaras et al. 2015]. Isso nos inspirou a buscar dados climáticos do Instituto Nacional de Meteorologia (INMET) para serem usados no processo de construção de modelos. A segunda estratégia que identificamos é a adoção do tempo das viagens realizadas anteriormente ao momento da predição [Alexandre et al. 2023]. Essas estratégias são exploradas neste trabalho e os resultados são apresentados.

4. Processo de Construção de Modelo para PTV

A Figura 1 apresenta o processo que construímos[1], com todas as tarefas para construção e avaliação de modelos para PTV utilizando (i) dados de GPS de ônibus da prefeitura de Florianópolis[2], relativos às linhas de ônibus da cidade; e (ii) dados de clima disponibilizados pelo INMET[3] (Instituto Nacional de Meteorologia). Os passos utilizados para

[1]Utilizamos a linguagem de modelagem BPMN (*Business Process Modeling Notation*) e a ferramenta draw.io para o desenho do modelo. BPMN: `https://www.omg.org/spec/BPMN/`. draw.io: `https://www.drawio.com/`. Acesso em: 13 maio 2025.

[2]Dados históricos disponíveis em: `https://github.com/TielleAlexandre/datasetFlorianopolis`. Acesso em: 13 maio 2025.

[3]Dados históricos disponíveis em: `https://portal.inmet.gov.br/dadoshistoricos`. Acesso em: 13 maio 2025.

a construção dos modelos de PTV também são apresentados. O modelo de processo foi concebido considerando 2 etapas de tarefas: **ET1. Pré-Processamento dos Dados** e **ET2. Construção e Análise de Modelo**. As estratégias (i.e., P1, C1, C2, C3) usadas para melhorar a qualidade dos modelos emergiram da revisão sistemática da literatura realizada por artigo [Alexandre et al. 2023]

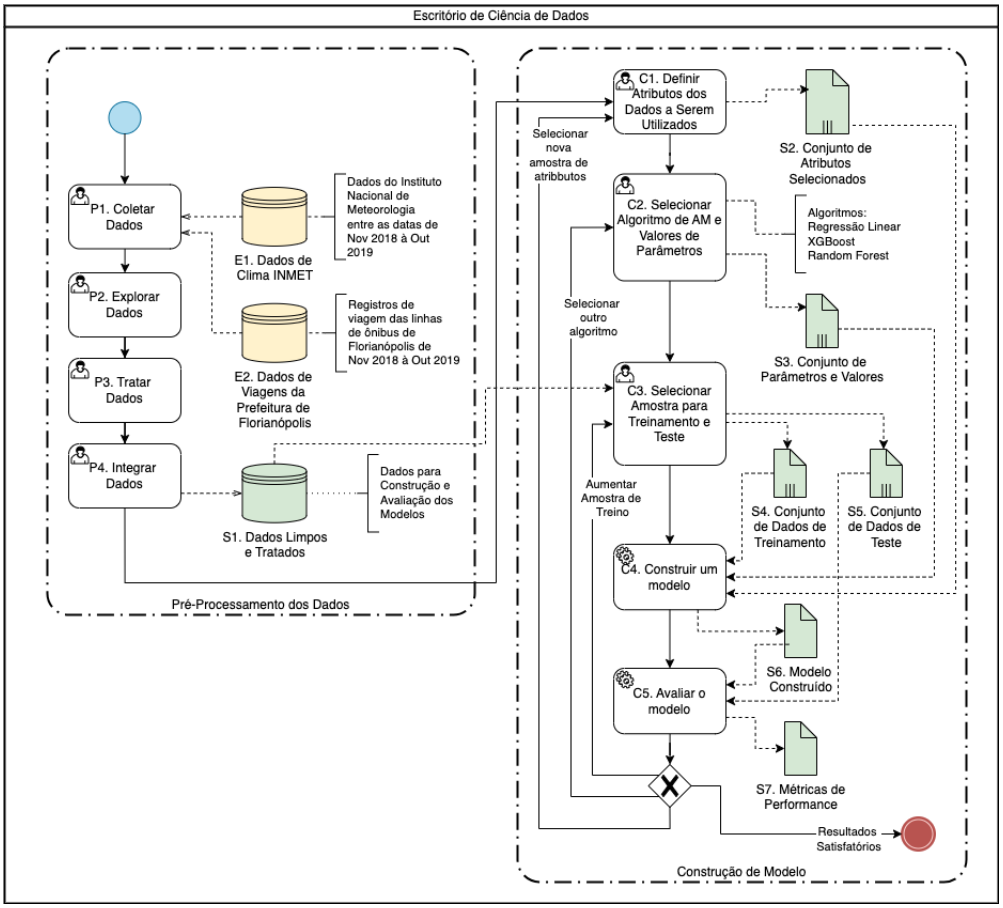


Figura 1. Processo para construção dos modelos para PTV

Na etapa **ET1**, os dados brutos disponibilizados pela prefeitura de Florianópolis são coletados, tratados e integrados, para garantir a qualidade dos dados para a etapa seguinte. Considerando sua população e infraestrutura, Florianópolis pode ser caracterizada como uma cidade de grande porte. Ainda assim, optamos por utilizar essa base de dados devido à sua disponibilidade e também para demonstrar que as estratégias previstas no processo são capazes de lidar com a previsibilidade do tempo de viagens mesmo em contextos de trânsito denso e caótico. Dessa forma, em municípios de menor porte, com menor volume de veículos e ocorrências, tais estratégias tendem a ser mais do que suficientes.

A etapa de Coleta de Dados é composta por 4 tarefas, que são atualmente realizadas de forma manual, por um cientista de dados, detalhadas a seguir. *P1. Coletar Dados* realiza a coleta dos dados de GPS das linhas de ônibus de Florianópolis e dos dados de clima do INMET, e que usa como entrada `E1. Dados de Clima INMET` e `E2. Dados de Viagens da Prefeitura de Florianópolis`. As tarefas *P2* e

P3 foram concebidas para garantir a consistência dos dados e para reduzir o número de linhas de ônibus. *P2. Explorar Dados* explora e analisa os dados para melhor compreensão e busca de inconsistências e ruídos. Já *P3. Tratar Dados* trata a inconsistência dos dados identificada na tarefa anterior.

Na etapa **ET2**, são realizadas 5 tarefas para construção e análise de um modelo de PTV, permitindo ao cientista de dados retornar a tarefas anteriores caso os resultados não sejam satisfatórios, gerando novos artefatos. As tarefas são: *C1. Definir Atributos*, escolhendo atributos da base `S1` para gerar `S2. Conjunto de Atributos Selecionados`; *C2. Selecionar Algoritmo e Parâmetros*, escolhendo entre RL, RF ou XGBoost, definindo parâmetros e gerando `S3. Conjunto de Parâmetros e Valores`; *C3. Selecionar Amostra*, definindo percentuais para gerar `S4. Dados de Treinamento` e `S5. Dados de Teste`; *C4. Construir Modelo*, com `S2`, `S3` e `S4` como entrada e `S6. Modelo Construído` como saída; e *C5. Avaliar Modelo*, utilizando `S6` para gerar `S7. Valor do MAPE`. As tarefas *C1* a *C3* são realizadas manualmente, enquanto *C4* e *C5* são automatizadas. O processo termina quando o cientista de dados considera os resultados satisfatórios ou esgota as estratégias disponíveis.

5. Execução do Processo

Etapa ET1: Em *P1*, analisamos inicialmente a base de dados da cidade de Florianópolis, com dados estruturados e com vários atributos disponíveis. Os dados são compostos por registros de viagem [4] das linhas de ônibus que circulam na cidade, de novembro de 2018 a novembro de 2019, sendo um arquivo para cada mês (12 arquivos). Após a concatenação de todos os 12 arquivos, foram obtidos 2.694.360 registros de 230 linhas de ônibus. Os atributos presentes no arquivo são (A_1) Data de início da rota, (A_2) Hora de início da rota, (A_3) Data de término da rota, (A_4) Hora de término da rota, (A_5) Sentido da rota (ida ou volta), (A_6) Identificador da linha, (A_7) Identificador do veículo, (A_8) Total de giros na roleta durante a viagem, (A_9) Distância da rota em Km e (A_{10}) Duração de tempo da viagem, que é o atributo alvo.

Em *P2 e P3*, os atributos A_1 e A_2 foram concatenados e transformados em um único atributo A_{11} para indicar o tempo t_{ini} de início da viagem, assim como A_3 e A_4, para um único atributo A_{12} de indicação do tempo t_{fim} término da rota. Daí, outros atributos foram construídos a partir de t_{ini} e t_{fim}, que foram dia da semana (A_{13}), hora do dia (A_{14}), dia do ano (A_{15}), dia do mês (A_{16}) e o turno do dia, com os valores madrugada, manhã, tarde e noite (A_{17}). Quanto a A_{10} (alvo), registros com valores $A_{10} = 0$ foram eliminados por serem considerados erro de medição. Construímos atributos temporais $\{A_{10}^{-1}(A_{18}), ..., A_{10}^{-5}(A_{22})\}$ para a construção dos modelos. Para isso, considere um registro $\mathbf{x}_i$, com $i = 6, ..., N$ (eliminados os 5 primeiros registros para considerar uma janela temporal de 5 valores), que possui um valor de tempo de viagem coletado no tempo t_i. Daí, considerando o novo conjunto de atributos $\{A_{10}^{-1}(A_{13}), ..., A_{10}^{-5}(A_{17})\}$, a cada registro $\mathbf{x}_i$ são adicionados, respectivamente, os valores de A_{10} coletados em $t^{-1}, ..., t^{-5}$.

Outros 2 atributos temporais também foram construídos, nomeados $\{A_{10}^{D-1}(A_{23}), ..., A_{10}^{D-7}(A_{24})\}$, que são preenchidos em cada registro $\mathbf{x}_i$ com o valor de A_{10} respectivamente no mesmo horário do dia anterior e da semana anterior (exatos 7 dias atrás). O atributo A_7 foi removido por se tratar de um identificador que não tinha

[4] Corresponde a uma viagem realizada por um ônibus entre o ponto inicial e final das linhas de ônibus

relevância para o problema. Foi criado também o atributo A_{25}, velocidade média da viagem (km/h), por meio de A_9 e A_{10}. Registros cujo valor em A_{18} estavam acima de 80km/h e abaixo de 3km/h foram eliminados, já que 80km/h é o limite de velocidade para veículos pesados em estradas e 3km/h é a velocidade média de um humano andando. Como o atributo A_8 (total de giros) também pode ser utilizado como um atributo alvo, cujo valor somente é preenchido ao final da viagem, foi criado um atributo A_{26} com o total de giros que ocorreu na mesma linha e no mesmo horário no dia anterior.

Observamos ainda um total de 223 linhas diferentes. A nossa proposta é que cada modelo de PTV seja construído para uma única linha, já que os dados de uma linha podem ter uma distribuição distinta de outras linhas. Além disso, reconstruir ou adaptar modelos para uma única linha, em casos de mudança de distribuição de probabilidade conjunta dos atributos de entrada, é mais simples do que reconstruir um modelo único para todas as linhas, caso a performance também seja adequada. Devido a isso, buscamos linhas com número de registros uniforme ao longo do período analisado. Para isso, usamos como critérios de seleção (i) linha com registros em todos os meses; (ii) linha com registros em mais de 21 dias de cada mês; e (iii) o desvio padrão do número de registros dentre os meses deve ser menor que o valor 3, visto que queremos linhas o mais próximas possível de uma distribuição constante, e que o número de registros por mês pode ser muito grande.

O número 3 foi selecionado após testes de diferentes valores de desvio padrão. Com essas remoções, o número de linhas de ônibus foi diminuído para 185. Também queremos linhas com uma quantidade de observações suficientes, pois com as premissas de seleção poderiam ter passado alguma rota que continha apenas uma observação por dia, o que seria insuficiente para a construção de um modelo. Então, construímos um gráfico de frequência, considerando intervalos de quantidade de registros para o eixo X e a quantidade de linhas por intervalo no eixo Y. Selecionamos as linhas situadas entre o primeiro e o terceiro quartil, com cerca de 3.900 a 16.900 observações, resultando em 94 linhas de ônibus.

Por fim, foram também removidos os registros com *outliers* no atributo A_{10} utilizando o Intervalo Interquartil (IQR), ou seja, registros cujo valor em A_{10} que estão acima do Limite Superior LS ou abaixo do limite inferior LI de IQR. Para cálculo desses limites, considere o valor $IQR = 1,5 * (Q_3 - Q_1)$. Daí, $LS = Q_3 + IQR$ e $LI = Q_1 - IQR$, tal que Q_1 é o primeiro quartil e Q_3 é o terceiro quartil de A_{10}. Aproximadamente 3,96% dos registros foram removidos do conjunto de dados. Para avaliação do processo construído, foram escolhidas seis linhas utilizando o desvio padrão do tempo de viagem, sendo duas com o desvio mediano (Linha 1 – 605 e Linha 2 – 1120), duas com menor desvio (Linha 3 – 271 e Linha 4 – 362) e duas com maior desvio (Linha 5 – 235 e Linha 6 – 4122).

Na base de dados do INMET, selecionamos dados de clima relativos ao mesmo período dos dados de tempo de viagem de Florianópolis, ou seja, de Novembro de 2018 a Novembro de 2019. Todos os arquivos foram unificados em um único arquivo. Cada registro dos dados do INMET é relativo a um determinado dia e hora do ano, contendo diversas informações de precipitação, pressão atmosférica, radiação, temperatura, vento e umidade. Neste trabalho, utilizamos os atributos de data e hora para integrar com o conjunto de dados das viagens de Florianópolis, e os atributos de precipitação (A_{27}) e temperatura máxima e mínima. As temperaturas máximas e mínimas também foram re-

duzidas para um único atributo chamado temperatura média.

Construímos um novo atributo A_{27} resultante da categorização do atributo precipitação, com as categorias `Nula` (valor igual a 0), `Fraca` (valor entre 0 e 2.5), `Moderada` (valor entre 2.5 e 10) e `Forte` (valor maior que 10), com base na literatura [World Meteorological Organization 2021]. Construímos também um novo atributo A_{28} resultante da categorização do atributo temperatura média, com as categorias `Frio` (valor menor que 17°C), `Normal` (valor entre 17°C e 25°C) ou `Quente` (valor maior que 25°C). Esses intervalos foram calculados a partir da temperatura média do ano (21,8°C) e consideramos intervalo de 4°C para a categoria `Normal`. Na tarefa **P4**, o conjunto de dados de clima foi integrado à base de dados contendo as viagens de ônibus de Florianópolis através dos atributos de data de ambos os conjuntos.

Para a etapa *ET2*, a biblioteca ScikitLearn foi utilizada. Consideramos diferentes cenários de experimentação, que são resultantes das decisões em cada uma das tarefas *C1* a *C3*. Lembrando que, na nossa proposta, cada modelo é construído para uma única linha de ônibus. Na tarefa *C3*, os dados da linha foram divididos em 70% para treino e 30% para teste, para todos os modelos construídos. Cada combinação das tarefas *C1* a *C3* executadas geraram o que chamamos de diferentes configurações Co de experimentos, conforme mostrado na Tabela 1. Como nosso objetivo é não somente avaliar os modelos mas também o impacto dos dados de clima e de dados de temporalidade da série, dividimos as configurações em 4 grupos. O Grupo $Co_1 = \{Co_{11}, ..., Co_{15}\}$ utilizou somente os atributos da base de dados de ônibus de Florianópolis (coluna Ônibus) e considerou diferentes subconjuntos de atributos do conjunto $\{A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}$, tendo sempre como base o subconjunto A_{13}, A_{14}, A_{15} (respectivamente dia da semana, hora do dia e dia do ano).

Os atributos derivados de data e hora foram inicialmente avaliados quanto à correlação, mas foram descartados por alta correlação com outros atributos. O Grupo $Co_2 = Co_{21}, ..., Co_{26}$ utilizou os atributos de Co_{15} — melhor configuração de Co_1 — combinados com amostras dos atributos temporais $A_{18}, ..., A_{24}$ (coluna Temporal). A configuração Co_{26} foi definida com base na observação dos melhores resultados anteriores. Em seguida, o Grupo $Co_3 = Co_{31}, ..., Co_{33}$ adotou os atributos de Co_{26}, acrescidos de variáveis climáticas A_{27}, A_{28} (coluna Clima). Já o Grupo $Co_4 = Co_{41}$ teve como foco avaliar se a otimização dos hiperparâmetros dos algoritmos RF e XGB melhora o desempenho, mantendo os atributos de Co_{26}, dado o impacto limitado da adição de variáveis climáticas. Na coluna Algoritmos, indicamos os modelos de AM utilizados: RL, RF e XGB com hiperparâmetros padrão; e RF* e XGB* com hiperparâmetros ajustados via RandomizedSearchCV[5].

Executando as 15 configurações de experimentos para cada uma das 6 linhas de ônibus, foram construídos 264 modelos (cada algoritmo de AM gera um modelo). A Figura 2 mostra os valores de MAPE obtidos para cada uma das linhas. É importante lembrar que como o MAPE é uma métrica de erro, *quanto menor o valor, melhor o modelo*. Podemos observar que a diferença maior entre os modelos foi para as linhas 2

[5]Para XGBoost: número de estimadores (100–1000, passo 100), profundidade máxima (2–15), peso mínimo (1–11), taxa de aprendizado (0.3 a 0.005) e gamma (0–0.7, passo 0.1). Para RF: número de estimadores (100–1000), profundidade máxima (2–15), bootstrap (true/false), número máximo de atributos (N, $\sqrt{N}$, $\log_2 N$), mínimo de amostras por divisão e por folha (2–11).

Tabela 1. Configurações para construção dos modelos

ID	Ônibus	Clima	Temporal	Algoritmos
Grupo Co_1 – somente dados de ônibus				
Co_{11}	A_{13}, A_{14}, A_{15}	–	–	RL, RF, XGB
Co_{12}	A_{13}, A_{14}, A_{15}, A_5	–	–	RL, RF, XGB
Co_{13}	A_{13}, A_{14}, A_{15}, A_{26}	–	–	RL, RF, XGB
Co_{14}	A_{13}, A_{14}, A_{15}, A_5, A_{26}	–	–	RL, RF, XGB
Co_{15}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	–	–	RL, RF, XGB
Grupo Co_2: atributos de Co_{15} e atributos temporais				
Co_{21}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	–	A_{18}	RL, RF, XGB
Co_{22}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	–	A_{18}, A_{19}	RL, RF, XGB
Co_{23}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	–	A_{18}, A_{19}, A_{20}	RL, RF, XGB
Co_{24}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	–	A_{18}, A_{19}, A_{20}, A_{21}	RL, RF, XGB
Co_{25}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	–	A_{18}, A_{19}, A_{20}, A_{21}, A_{22}	RL, RF, XGB
Co_{26}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	–	A_{18}, A_{23}, A_{24}	RL, RF, XGB
Grupo Co_3: atributos de Co_{26} e atributos de clima				
Co_{31}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	A_{27}	A_{18}, A_{23}, A_{24}	RL, RF, XGB
Co_{32}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	A_{27}, A_{28}	A_{18}, A_{23}, A_{24}	RL, RF, XGB
Co_{33}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	A_{28}	A_{18}, A_{23}, A_{24}	RL, RF, XGB
Grupo Co_4: atributos de Co_{2_6} com otimização de parâmetros de RF e XGB				
Co_{41}	A_{13}, A_{14}, A_{15}, A_5, A_9, A_{26}	–	A_{18}, A_{23}, A_{24}	RF*, XGB*

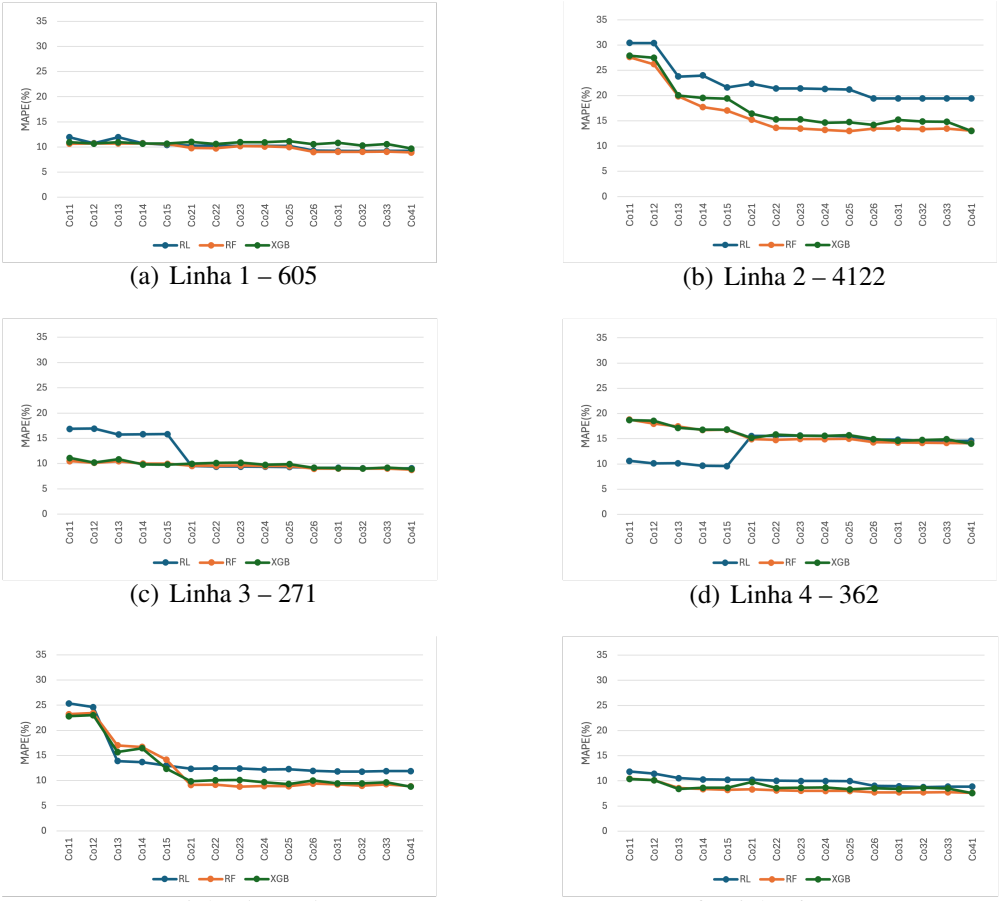
(Figura 2(b)) e 5 (Figura 2(e)), que têm desvio de tempo de viagem mediano e maior, respectivamente. É interessante observar que, com a adição dos atributos temporais (Grupo Co_2), (i) os modelos RL melhoram para a linha 3 (Figura 2(c)) e pioram para a linha 4 (Figura 2(d)); e (ii) todos os modelos melhoram para as linhas 2 (Figura 2(b)) e 5 (Figura 2(e)).

Para avaliar se há melhora dos modelos segundo a medida MAPE com significância estatística, construímos 4 hipóteses nulas quanto aos subconjuntos de atributos de ônibus utilizados (Co_1), ao efeito dos atributos temporais (Co_2), ao efeito dos atributos de clima (Co_3), e à otimização dos hiperparâmetros (Co_4), que são: H_0^1: os diferentes subconjuntos de atributos de ônibus não trazem diferença para os modelos; H_0^2: os atributos temporais não trazem diferença para os modelos quando comparado a Co_{15}; H_0^3: os atributos de clima não trazem diferença para os modelos quando comparado Co_{26}; e H_0^4: a otimização dos parâmetros não traz diferença para os modelos RF e XGB quando comparado a Co_{26}. O teste de Friedman[6] [Demšar 2006] foi utilizado para testar as hipóteses nulas H_0^1 a H_0^3. Em todos os 3 casos, as hipóteses nulas foram rejeitadas com 95% de confiança.

A hipótese nula H_{04} foi testada com o teste de Wilcoxon [7], que também rejeitou a hipótese nula com 95% de confiança, indicando que a otimização tem impacto positivo para os conjuntos de dados utilizados. Aplicamos pós-testes para H_0^1 a H_0^3 para verificar se há diferença crítica entre cada configuração para cada teste. O teste de Nemenyi foi aplicado para a hipótese H_{01}, pois não há uma configuração de base de comparação, e

[6]O teste de Friedman é não paramétrico e adequado para cenários em que mais que 2 configurações são comparadas.

[7]O teste de Wilcoxon é não paramétrico e adequado para cenários em que somente 2 configurações são comparadas.



(a) Linha 1 – 605 (b) Linha 2 – 4122

(c) Linha 3 – 271 (d) Linha 4 – 362

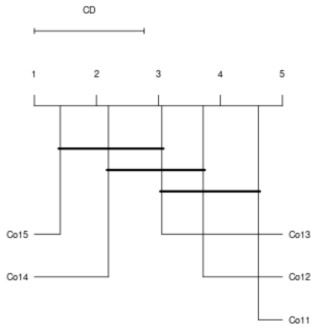
(e) Linha 5 – 235 (f) Linha 6 – 1120

Figura 2. Resultados obtidos em todas as configurações de experimentos para as 6 linhas de ônibus de Florianópolis selecionadas

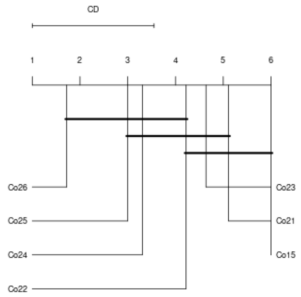
o Bonferroni-Dunn, para H_{02} e H_{03}, devido às configurações de base de comparação. Nas Figuras 3(a) a 3(c) são mostrados os gráficos de diferença crítica construídos, que apresentam a ordem da qualidade de cada configuração (nesse tipo de visualização, o mais próximo de 1 é o melhor segundo a métrica MAPE). Podemos observar que, para H_0^1 rejeitada, a melhor configuração é a Co_{15} (empatada estatisticamente com Co_{14} e Co_{13}) e, para H_0^2 rejeitada, a melhor configuração é a Co_26 (empatada estatisticamente com Co_{25}, Co_{24} e Co_{22}), o que confirmou as decisões anteriores no estabelecimento de conjuntos de atributos. Por outro lado, para H_0^3, ainda que tenha sido rejeitada, todas as configurações ficam empatadas, o que confirma nossa observação pelos gráficos de que os atributos de clima não trazem diferença significativa.

6. Conclusão

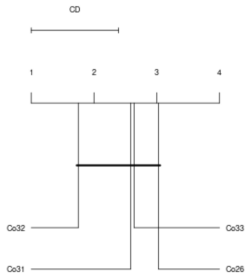
Com base nos experimentos realizados, observamos que o processo proposto atende às necessidades de um cientista de dados para a construção de modelos de PTV com AM, ainda que cada tarefa possa ser bastante complexa. Com relação à questão de pesquisa: "Quais estratégias (i.e., algoritmos, seleção de atributos) são eficazes para a construção de modelos de PTV especialmente em contextos de prefeituras de pequeno porte ?" observamos que (i) o processo auxilia a atacar os desafios de forma incremental, mas o



(a) Pós-teste Nemenyi – H_0^1: Configurações de Experimentos Co_1



(b) Pós-teste Bonferroni-Dunn – H_0^2: Configurações de Experimentos Co_2 comparado ao experimento Co_{15}



(c) Pós-teste Bonferroni-Dunn – H_0^3: Configurações de Experimentos Co_3 comparado ao experimento Co_{26}

Figura 3. Visualização gráfica do resultado dos pós-testes de Friedman

cientista de dados ainda precisa investir bastante tempo no processamento dos dados; (ii) há muitas configurações de experimentos que precisam ser executadas, tanto para os dados de geolocalização da cidade de Florianópolis quanto para outras cidades; (iii) os algoritmos de AM mais simples mostram resultados promissores; e (iv) a otimização dos hiperparâmetros do RF e XGB gera resultados significativos, mas implica em um maior tempo de construção para esses modelos.

Quanto às demais estratégias discutidas como o uso de atributos climáticos e de atributos temporais das séries de dados, os valores de MAPE e os testes estatísticos mostraram que (i) os atributos temporais melhoraram consideravelmente os modelos construídos e (ii) os atributos climáticos não tiveram impacto significativo nos experimentos realizados. No entanto, é importante observar que os achados são específicos para as bases de dados construídas. Para novos conjuntos de dados, todo o processo deve ser reexecutado, incluindo as avaliações dos modelos. Além disso, outros atributos climáticos podem ser explorados com o auxílio de especialistas.

Este trabalho destaca que é possível construir modelos com baixos valores de MAPE usando algoritmos de AM mais simples, tornando essas soluções mais acessíveis às prefeituras. Como os resultados foram satisfatórios na previsão de rotas de ônibus em Florianópolis, uma cidade de grande porte, é razoável supor que as estratégias previstas no processo sejam suficientes — ou até mais adequadas — em municípios menores, com menor complexidade de tráfego. No entanto, para algumas rotas de ônibus, outros algoritmos precisam ser explorados, como o uso de redes neurais recorrentes ou de algoritmos de aprendizado em fluxo de dados. Logo, o processo apresentado pode evoluir para tra-

tar casos em que as avaliações não apresentam resultados tão satisfatórios. Além disso, a automação desses processos por meio de workflows científicos pode ser interessante, bem como o desenvolvimento de outros frameworks computacionais para apoiar os cientistas de dados na construção de modelos de PTV.

Agradecimentos

Agradecemos aos revisores do comitê de programa pela revisão detalhada, que nos permitiu evoluir o trabalho. Agradecemos também ao CNPq pelas discussões realizadas com colegas via projeto financiado pela instituição, Proc. n. 441713/2023-8. Por fim, agradecemos também à prefeitura de Florianópolis por ter nos disponibilizado todos os dados utilizados neste trabalho, via e-SIC. Sem tal base de dados, esse trabalho não poderia ter sido executado.

Referências

Agafonov, A. and Yumaganov, A. (2019). Performance comparison of machine learning methods in the bus arrival time prediction problem. *CEUR Workshop Proceedings*.

Alexandre, T., Bernardini, F., Viterbo, J., and Pantoja, C. E. (2023). Machine learning applied to public transportation by bus: A systematic literature review. *Transportation Research Record*, 2677(7):639–660.

Aurélien, G. (2017). *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. O'Reilly Media.

Bahuleyan, H. and Vanajakshi, L. (2016). Arterial path-level travel-time estimation using machine-learning techniques. *Journal of Computing in Civil Engineering 31, 3*.

Cielen, D. and Meysman, A. (2016). *Introducing data science: big data, machine learning, and more, using Python tools*. Manning Publications.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30.

Faceli, K., Lorena, A. C., Gama, J., de Almeida, T., and de Carva, A. (2021). *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. LTC.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.

Samaras, P., Fachantidis, A., Tsoumakas, G., and Vlahavas, I. (2015). A prediction model of passenger demand using avl and apc data from a bus fleet. *Proceedings of the 19th Panhellenic Conference on Informatics*.

World Meteorological Organization (2021). Guide to instruments and methods of observation wmo-no. 8. Technical Report WMO-No. 8, World Meteorological Organization (WMO). Acessado em: 19/08/2023.

Yamaguchi, T., A.S., M., and Mine, T. (2018). Prediction of bus delay over intervals on various kinds of routes using bus probe data. *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.