UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
CENTRO MULTIDISCIPLINAR UFRJ-MACAÉ
INSTITUTO POLITÉCNICO
CURSO DE ENGENHARIA CIVIL

ALLAN SILVA CAMPOS

APLICAÇÃO DE FLORESTAS ALEATÓRIAS PARA PREVISÃO DA DEMANDA DE PASSAGEIROS NO TRANSPORTE PÚBLICO

Macaé
2022

ALLAN SILVA CAMPOS

APLICAÇÃO DE FLORESTAS ALEATÓRIAS PARA PREVISÃO DA DEMANDA DE PASSAGEIROS NO TRANSPORTE PÚBLICO

Trabalho de Conclusão de Curso de graduação submetida ao Instituto Politécnico do CM UFRJ-Macaé como parte dos requisitos necessários à obtenção do grau de bacharel em Engenharia Civil.

Orientadores:

Prof. Conrado Vidotte Plaza

Profa. Janaína Sant'Anna Gomide Gomes

Macaé

2022

CIP - Catalogação na Publicação

C198
Campos, Allan Silva
Aplicação de floresta aleatórias para a previsão de demanda de passageiros no transporte público / Allan Silva Campos - Macaé, 2022.
64 f.

Orientador(a): Conrado Vidotte Plaza.
Coorientador(a): Janaina Sant'anna Gomide Gomes.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto Politécnico, Bacharel em Engenharia Civil, 2022.

1. Transporte coletivo. 2. Aprendizado de máquina. 3. Floresta aleatória. 4. Previsão de demanda. I. Plaza, Conrado Vidotte, orient. II. Gomes, Janaina Sant'anna Gomide, coorint. III. Título.

CDD 624

Ficha catalográfica elaborada pela Biblioteca com os
dados fornecidos pelo(a) autor(a)
Biblioteca Central do Centro Multidisciplinar UFRJ-Macaé
Bibliotecário: Anderson dos Santos Guarino CRB7 – 5280

ALLAN SILVA CAMPOS

APLICAÇÃO DE FLORESTAS ALEATÓRIAS PARA PREVISÃO DA DEMANDA DE PASSAGEIROS NO TRANSPORTE PÚBLICO

Trabalho de Conclusão de Curso de graduação submetida ao Instituto Politécnico do CM UFRJ-Macaé como parte dos requisitos necessários à obtenção do grau de bacharel em Engenharia Civil.

Aprovado em Macaé, 30 de junho de 2022.

BANCA EXAMINADORA:

Profa. Janaína Sant'Anna Gomide Gomes, D.Sc. (UFRJ)

Prof. Conrado Vidotte Plaza, M.Sc. (UFRJ)

Profa. Laura Emmanuella Alves dos Santos Santana, D.Sc. (UFRJ)

Prof. Glaydston Ribeiro, Ph.D (UFRJ)

AGRADECIMENTOS

Agradeço aos meus pais, pelo suporte incondicional desde sempre, e não só por aguentarem as minhas chatices, mas principalmente as chatices do meu irmão. Que eu agradeço pelos dotes culinários diários, e percepção mais que aguçada da realidade.

Agradeço ao meu orientador Conrado, e orientadora Janaina, pela orientação na confecção desse trabalho, e também pelas aulas de suas respectivas disciplinas no decorrer da graduação, sempre com uma didática impecável e bom humor.

“And life flows on within you and
without you”

George Harrison

RESUMO

O planejamento de um sistema de transporte público passa necessariamente pela previsão da demanda de passageiros, visto que todo um equilíbrio sistemático depende dessa noção, que se mal dimensionada, pode causar uma miríade de situações indesejadas para as operadoras de transporte, para os usuários do sistema, para o meio ambiente e consequentemente para todo restante da sociedade. Uma das soluções possíveis para realizar tal previsão é através do auxílio de Aprendizado de Máquina, materializado através do método de Florestas Aleatórias. Após uma revisão de conceitos do transporte, como oferta, demanda e alocação de recursos, e de conceitos de algoritmo, como métricas de desempenho e divisão treino e teste, foram conduzidos uma série de experimentos envolvendo a base de dados do transporte coletivo de ônibus de Belo Horizonte, entre 2016 e 2021. Experimentos estes, realizados com modelos de Floresta Aleatória, que variam em características, ou hiperparâmetros, a fim de avaliar a melhor estratégia de previsão, e seguindo a técnica "Janela Crescente com Validação Adiante" para percorrer o conjunto de dados com divisões de treino e teste. Além da repetição dos experimentos em diferentes divisões do conjunto de dados, permitindo uma melhor demonstração do Aprendizado de Máquina e das métricas de desempenho como R^2 e Erro Médio Absoluto. Ao fim do estudo, chegou-se à conclusão que os modelos geravam resultados interessantes na capacidade preditiva a partir da variação do hiperparâmetro "Profundidade Máxima da Árvore", sem necessariamente aumentar o tempo de execução do modelo. Onde, por sua vez a variação do hiperparâmetro "Quantidade de Árvores" não causava tanta mudança no desempenho, com um aumento no tempo de execução do modelo.

Palavras-chave: Aprendizado de Máquina. Previsão de Demanda de Passageiros. Florestas Aleatórias. Transporte Público Urbano.

ABSTRACT

The planning of a public transport system necessarily involves the forecasting of passenger demand, since a whole systematic balance depends on this notion, which, if poorly dimensioned, can cause a myriad of undesired situations for transport operators, for system users, for the environment and consequently for the rest of society. One of the possible solutions to carry out such a prediction is through the aid of Machine Learning, materialized through the Random Forests method. After a review of transport concepts, such as supply, demand and resource allocation, and algorithm concepts, such as performance metrics and training and testing division, a series of experiments were led involving the public transport database of Belo Horizonte buses, between 2016 and 2021. These experiments were carried out with Random Forest models, which vary in characteristics, or hyperparameters, in order to evaluate the best forecasting strategy, and following the "Growing Window with Forward Validation" technique to traverse the dataset with training and testing divisions. In addition to repeating the experiments in different divisions of the dataset, allowing a better demonstration of Machine Learning and performance metrics such as R-squared and Mean Absolute Error. At the end of the study, it was concluded that the models generated interesting results in the predictive capacity from the variation of the hyperparameter "Maximum Tree Depth", without necessarily increasing the execution time of the model. Where, in turn, the variation of the "Number of Trees" hyperparameter didn't caused much change in performance with an increase in model execution time.

Key words: Machine Learning. Passenger Demand Forecasting. Random Forests. Urban Public Transport.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

A 100 N 100 – Modelo de Floresta Aleatória com hiperparâmetros igual a 100 tanto para "Quantidade de Árvores" e "Profundidade Máxima da Árvore" respectivamente "A" e "N".

ANTP - Associação Nacional de Transportes Públicos Urbanos

CART – Árvore de Classificação e Regressão

INMET - Instituto Nacional de Meteorologia

MAE – Erro médio absoluto

MSE – Erro médio quadrático

N – Número de instâncias da base de dados.

NTU - Associação Nacional das Empresas de Transportes Urbanos

R^2 ou R^2 – Coeficiente de determinação

RMSE – Raiz do erro médio quadrático

UFMG – Universidade Federal de Minas Gerais

yj – Valores advindos da previsão do algoritmo, na j-ésima instância.

$\hat{y}$j – Valores alvos da base de dados, para j-ésima instância.

$\bar{y}$j – Média dos dados de valores alvos.

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Países em desenvolvimento, como o Brasil, possuem sérios problemas relacionados ao transporte, não só pela falta de estradas e infraestruturas em regiões ermas, mas também situações características de uma sociedade pós-industrializada, como o congestionamento e poluição.

Estes problemas se tornam ainda mais críticos em decorrência às especificidades socioeconômicas e de desenvolvimento, como baixos salários relativos ao mundo globalizado, urbanização acelerada e alta demanda do transporte público, além da escassez de recursos, como capital, mão de obra especializada e coleta de dados sólidos (ORTUZAR e WILLUMSEN, 2011).

Nesse contexto, quanto maior o desenvolvimento da sociedade, maior a atividade econômica e, com isso, maior a necessidade por deslocamentos para atendimento das necessidades sociais e econômicas. Um crescimento urbano desordenado em relação ao uso do solo urbano e consequente especialização das zonas da cidade entre áreas residenciais, comerciais e industriais, pode resultar em maior necessidade do transporte motorizado para cobrir as grandes distâncias das cidades.

Os deslocamentos urbanos e a consequente necessidade de transporte ocorrem majoritariamente com uma periodicidade. Isto é, com picos de demanda em determinados horários ao longo do dia. Tradicionalmente, o primeiro desses picos ocorre na manhã, gerados justamente pelas viagens por motivo de trabalho (casa-trabalho), enquanto no fim da tarde ocorre outro pico em consequência do percurso reverso (movimentos pendulares).

Segundo a Associação Nacional de Transportes Públicos Urbanos (ANTP, 2018), a partir da análise de dados de mobilidade de 533 municípios com população acima de 60 mil habitantes, aproximadamente 28% dos deslocamentos são realizados em Transporte Coletivo. Estes deslocamentos representam cerca de 53% das distâncias percorridas pelos deslocamentos. Cerca de 85% dessas viagens ocorrem em ônibus, que representam mais de 16 bilhões de deslocamentos por ano.

No entanto, segundo informações da Associação Nacional das Empresas de Transportes Urbanos (NTU, 2019), 12,5 milhões de brasileiros deixaram de usar ônibus urbano entre 2018 e 2019, uma redução de 4,3% na demanda. Fenômeno agravado ano

após ano com o surgimento de novas tecnologias, como serviços de transporte por aplicativo.

Diante do relevante uso dos transportes por ônibus e dessas incertezas na demanda, faz-se importante avaliar meios para previsão de demanda. Visando encorpar o planejamento do transporte coletivo com estimativas confiáveis, que serviriam para melhorar a estratégia da oferta e alocação dos recursos.

Dentre os meios disponíveis para previsão de demanda, encontram-se as técnicas de Aprendizado de Máquina, que utilizam de conjuntos de operações, ou algoritmos, para compreender padrões em uma base de dados, e a partir do aprendizado gerado, obter uma habilidade de generalização para outros casos, e com isso, realizar previsões.

1.2 MOTIVAÇÃO E JUSTIFICATIVA

Nesse cenário e notando a tendência mundial com objetivos nos planejamentos de transporte moderno, de incrementar a participação do transporte coletivo (FERRONATTO, 2002), precisa-se pensar em soluções inteligentes para poder melhorar o planejamento de curto prazo do transporte coletivo, e mais especificamente o transporte por ônibus.

Com isso, o trabalho a seguir, pretende discutir e apresentar um estudo baseado na utilização dos métodos de Aprendizado de Máquina, voltados para a previsão de demanda de passageiros no sistema de ônibus da cidade de Belo Horizonte. Cidade esta que foi escolhida por conta da disponibilidade de um conjunto sólido de dados do sistema de transporte.

1.3 OBJETIVOS

Esse trabalho tem como objetivo avaliar modelos de aprendizado de máquina, utilizando do algoritmo de Florestas Aleatórias para realizar a previsão da demanda de passageiros no transporte público, utilizando da base de dados de transporte público urbano da cidade de Belo Horizonte entre 2016 e 2021.

1.3.1 Objetivos Específicos

- Avaliar o desempenho de diferentes combinações de hiperparâmetros das Florestas Aleatórias.
- Refazer os testes com Aprendizado de Máquina em intervalos de dados referentes ao período da pandemia e ao período antes desta.

1.4 ESTRUTURA DO TRABALHO

No segundo capítulo desse trabalho é feita a Revisão Bibliográfica, onde são abordadas questões de Transporte como demanda, oferta e alocação de recursos, além de questões do Aprendizado de Máquina, como métricas, divisão treino e teste, e algoritmos. Finalizando com um panorama da pandemia e referências a trabalhos relacionados.

Já no terceiro capítulo encontra-se a Metodologia do trabalho, que trata das ferramentas e métodos utilizados, durante todo o processo de confecção dos modelos e experimentos.

O quarto capítulo trata dos resultados desses experimentos, divididos por subconjuntos da base de dados, sendo estes: o conjunto inteiro, o período de normalidade antes da pandemia, e o período da pandemia.

No quinto capítulo chega-se à conclusão do trabalho, com um fechamento de ideias e raciocínios após a realização do estudo.

2 REVISÃO BIBLIOGRÁFICA

2.1 TRANSPORTE COLETIVO

O transporte coletivo é o único meio de transporte motorizado para grande parte das pessoas, além de existir uma tendência ao redor do mundo de aumentar sua participação nos planejamentos de transporte. Isso ocorre principalmente pela maior eficácia do transporte coletivo frente ao automóvel individual, levando em conta pontos como gasto enérgico, poluição e quantidade de pessoas transportadas (FERRONATTO, 2002).

Nesse sentido, nota-se que os estudos de mobilidade devem cada vez mais não serem restrito apenas às questões econômicas, mas também considerando aspectos ambientais e sociais, em direção ao desenvolvimento sustentável dos transportes.

Contudo, é preciso analisar além do importante prisma ambiental, através de questões como gerenciamento da demanda nos transportes, ou ainda o serviço de oferta do transporte e alocação dos recursos. Visto que, segundo Ferronatto (2002), problemas como congestionamentos de tráfego, superlotação de transportes coletivos e novamente a saturação da capacidade ambiental, são causados principalmente pelo desequilíbrio entre oferta e demanda dos transportes. Provocado pela capacidade limitada de investimento em infraestrutura dos países que como o Brasil, estão em desenvolvimento.

2.1.1 Demanda

Para entender a demanda, é preciso ver tal carecimento como uma derivação de algo, e não algo por si só, visto que as pessoas precisam do transporte e consequente locomoção de motorizados para satisfazer necessidades como trabalho, lazer e saúde. Logo, para entender os princípios da demanda, é necessário entender a distribuição de atividades no espaço das cidades.

O que torna tão complicado analisar e prever a demanda dos serviços de transporte, é a característica desse modelo de transmutação em possuir uma variedade grande de especificidades, que variam com a hora do dia, com os dias da semana, com o propósito da jornada e a importância da rapidez ou frequência dela, dentre outros fatores.

Ainda assim, o meio mais comum de tratar dessa situação de acordo com Ortuzar e Willumsen (2011), é dividir os espaços de solo urbano em zonas, e transformar isso em código de computador, junto com as redes de transporte. Onde deve-se levar em conta as

distâncias, além da alocação dos pontos de origem e de destinação das viagens, com seus devidos atributos e particularidades.

Inclusive, nos sistemas com demanda dimensionada de forma reduzida, identifica-se viagens com a ocupação do veículo constantemente superior a lotação máxima. Entretanto, não é tão simples resolver o subdimensionamento utilizando somente como solução a mobilização de viagens adicionais nas regiões de lotação máxima. Visto que tal prática se mostraria inadequada quando a criação de viagens adicionais causasse uma redistribuição de horários de viagens anteriores e posteriores, mexendo em todo um sistema e criando outras disrupções.

Por outro lado, é comum constatar a existência de veículos subutilizados em períodos com menor movimento, gerando um incremento desnecessário dos custos. Ao mesmo tempo que as operadoras, segundo Cruz (1991), ainda consideram um exagero utilizar metodologias mais sofisticadas para o sistema de transporte coletivo, justificando tal atitude com sua preocupação principal de geração de recursos, ignorando necessidades do usuário e possibilidades de otimização do sistema.

Sem contar que, é preciso entender o efeito de eventos extraordinários, nos modelos da previsão de demanda, como por exemplo a pandemia do novo coronavírus, que em diferentes regiões do mundo, o assunto da pandemia e seus impactos nas populações é dominante. Em janeiro de 2020, a Organização Mundial da Saúde (OMS) sinalizou um surto de um novo coronavírus na China. Para em março do mesmo ano, após debates e busca por evidências, declarar situação de Emergência de Saúde Pública de Interesse Internacional. Tendo em vista a proliferação em escopo planetário, a doença batizada como COVID-19 foi caracterizada como uma pandemia (CRUZ, 2020).

2.1.2 Oferta e Alocação de Recursos

Como exposto por Ortuzar e Willumsen (2011), antes de tudo é preciso definir que a oferta de transporte é um serviço, e não um bem, ou seja, não é possível estocar para depois utilizar nos picos de demanda. Com isso em mente, é necessário perceber que um serviço de transporte deve ser utilizado na hora e local em que este foi criado, caso contrário seu benefício é perdido. Esta é inclusive mais uma das razões pela qual a previsão de demanda com alta precisão ser extremamente importante dentro de um sistema.

Sistema este que, em termos gerais, possui um número fixo de bens, ou infraestrutura, e um número de unidades móveis, ou veículos. A partir da combinação

desses, com um conjunto de regras para a operação dos próprios, torna-se possível a movimentação de pessoas e cargas.

Pode-se olhar para o equilíbrio entre oferta e demanda, e subsequente alocação de recursos como dependendo de sete fatores expostos por Ferronatto (2002), sendo o primeiro destes o uso do solo, já citado anteriormente e apontado como motivador do transporte motorizado, visto que o mesmo divide áreas urbanas em zonas. Tal fator torna-se insustentável na medida que gera condições de congestionamento e poluição.

Em seguida vem a renda, tendo efeito direto no aumento da motorização da população e consequente queda de demanda por transporte coletivo. Na mesma linha econômica segue-se a tarifa do transporte coletivo, habitando na beira de um ciclo vicioso onde o aumento da tarifa pode reduzir a demanda, e pode acabar por provocar um aumento de custos que se refletem na tarifa. Entretanto, zerar a tarifa não necessariamente significaria aumentar a demanda. Tal sequência perigosa poderia ser quebrada com auxílio de subsídios, restrições para automóveis particulares e outras medidas complementares visando manter níveis estáveis de demanda.

De forma mais ampla, pode-se imaginar a tarifa como uma das partes do custo total do transporte para o usuário. Tendo entre as partes, o tempo gasto para se transportar e o esforço físico para alcançar o transporte. Por outra ótica, o custo de oferta dos transportes é apontado como quarto fator da lista de sete. Isto é, o custo do ponto de vista das operadoras de sistemas de transporte coletivo, que acabam por ser elevados por conta dos picos de serviço e seus requisitos acima da média.

Também é necessário comentar sobre o fator qualitativo do serviço, onde o efeito de mudanças nos serviços frequentes e confiáveis é menos significativa, do que nos serviços menos confiáveis e de baixa frequência. Mudanças estas que podem ocorrer no conforto e conveniência destes serviços, que vão desde viajar sentado, até confiabilidade, competência, acessibilidade, cortesia dos funcionários, credibilidade e segurança. Nesse panorama, mudanças no veículo pesam mais do que mudanças físicas na estação.

O encadeamento de viagens é um dos últimos fatores, visto que a maioria das viagens a trabalho ocorre nos períodos de pico e não tem múltiplos propósitos, ou seja, não possuem uma variabilidade de comportamentos complexos, como mais paradas ou destinos alternativos, e o reescalonamento das viagens com múltiplos propósitos poderia reduzir o congestionamento.

Por fim, políticas públicas podem incentivar ou reprimir demanda por transportes, ou até uma migração entre modais, através de restrições ao uso de automóveis, prioridade de circulação para veículos coletivos, escalonamento de horários e subsídios ao transporte coletivo.

Com isso, precisa-se observar que toda essa análise de predição é de fundamental importância para o planejamento de investimentos, que seriam destinados para garantir a mobilidade e consequentemente o desenvolvimento de atividades produtivas em uma região. Ainda mais no Brasil, onde o meio de transporte coletivo predominante nas cidades é o ônibus, e por isso, o gerenciamento destes depende diretamente de análises de demanda e capacidade do modal.

De acordo com Murça e Müller (2014), também é necessário realizar que no contexto brasileiro, a pequena diversificação dos tipos de opções para transporte, ou matriz modal, e restrições de capacidade dos meios de transporte públicos mostram a recorrente insuficiência de planejamento voltado para priorização e alocação de investimentos.

2.2 APRENDIZADO DE MÁQUINA

Segundo Géron (2019), Aprendizado de Máquina pode ser definido como a ciência da programação que torna possível o aprendizado de computadores a partir de dados. Assim, com auxílio de sistemas de aprendizado, ou programas de computador, ou algoritmos de Aprendizado de Máquina, é possível criar um tomador de decisões que se baseia em experiência acumulada das soluções bem-sucedidas de problemas anteriores. Este tomador de decisão possui características que o diferencia dos outros tomadores de decisão, como forma de aprendizado utilizado, modo, paradigma e conceitos para aprender (MONARD e BARANAUSKAS, 2003). A partir de tal habilidade, abre-se um campo imenso de possibilidades para aplicação de tal conhecimento, seja através de um filtro de e-mails do tipo spam, ou uma máquina capaz de reconhecer dígitos manuscritos com uma precisão acima de 98%.

Dentro dos tipos de aprendizado indutivo demonstrados por Monard e Baranauskas (2003), isto é, do aprendizado que se origina de uma parte e é generalizado para um todo, estão contidos o Aprendizado Supervisionado e o Aprendizado Não-Supervisionado. No primeiro, é preciso fornecer ao programa computacional um conjunto de exemplos para treinamento e teste, no qual se tem conhecimento do rótulo desse conjunto de exemplos. Como por exemplo, um conjunto de textos de e-mail com um

rótulo indicando qual texto é referente ao e-mail que não é spam, e qual texto se refere ao e-mail do tipo spam, portanto descartável.

Cada instância nos conjuntos de dados desse tipo de aprendizado, pode ser descrita como um vetor de valores de características, ou atributos, acompanhadas do rótulo. Este podendo ser uma classe como “Spam” e “Não Spam”, para problemas conhecidos como classificatórios, ou pode-se ter valores contínuos como rótulos, em problemas conhecidos como regressão.

O objetivo nesse tipo de aprendizado é a construção de um classificador ou regressor, para determinar corretamente o rótulo de novos exemplos ainda não rotulados, ou seja, determinação do rótulo a partir de vetores de atributos, com base no treinamento e teste realizados previamente com os dados rotulados.

No Aprendizado Não-Supervisionado, o programa analisa exemplos fornecidos não rotulados, e partir destes, tenta construir agrupamentos destes exemplos, podendo buscar semelhanças entre eles. Após essa etapa, se faz necessário uma análise que visa determinar o significado de cada um dos agrupamentos no contexto do problema analisado.

De acordo com Géron (2019), em ambos tipos de aprendizado, nota-se que estes são boas soluções para problemas que exigem muita configuração manual, ou uma longa lista de regras, visto que a prática do Aprendizado de Máquina geralmente simplifica o código. Ou então problemas complexos, para os quais não existe uma boa solução por meio de uma abordagem tradicional. Além do aprendizado se adaptar bem a novos dados em um ambiente flutuante e volátil. Sem contar com a compreensão dos problemas envolvendo uma grande quantidade de dados.

Dois dos principais problemas do Aprendizado de Máquina são o sobreajuste e o subajuste de dados. No primeiro, o modelo funciona bem no treinamento, porém não consegue ter a capacidade de generalizar, e com isso tem um desempenho ruim na rotulação dos valores sem rótulo.

Tal ocorrência pode ser solucionada, segundo Géron (2019), com a simplificação do modelo, ou a redução do número de atributos, ou na coleta de mais dados de treinamento, ou ainda na redução de ruído nos dados de treinamento, como a remoção de erros nos dados ou pontos discrepantes.

Enquanto o subajuste ocorre quando o modelo é simples demais para o problema em questão, gerando previsões imprecisas mesmo no campo do treinamento. A solução para tal problema seria a seleção de um modelo mais poderoso, ou alimentar o algoritmo com melhores características, ou reduzir as restrições de hiperparâmetro do modelo.

2.2.1 Métricas de Desempenho

As métricas de desempenho fazem parte de forma integral do aprendizado de máquina, visto que tais cálculos fornecem quantitativos dos erros gerados pelos algoritmos, tornando possível a verificação de parâmetros que fazem o modelo melhorar ou piorar em desempenho. Estas métricas foram abordadas baseadas em Géron (2019).

Nos problemas de regressão, geralmente se estabelece uma função de custo, ou seja, uma medição do erro do modelo. Medindo a distância das previsões para os valores de treinamento, a partir disso, tem-se como objetivo minimizar essa distância, e com isso aprimorar o modelo.

Uma das métricas usadas em problemas de regressão é a Raiz do Erro Quadrático Médio (RMSE), descrita na Equação 1. Tal medida quantifica os erros gerados pelas previsões do sistema, dando um peso maior para grandes erros. O quadrado do RMSE também pode ser utilizado, e é chamado de Erro Médio Quadrático (MSE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \qquad (1)$$

Onde n se refere ao número de instâncias, yj aos valores preditos para o j-ésimo exemplo, e $\hat{y}$j aos valores alvos para o j-ésimo exemplo.

Em outros contextos, pode-se utilizar do Erro Médio Absoluto, descrito na Equação 2. Métrica esta que é menos sensível aos valores discrepantes como a RMSE, visto que os erros não são levados ao quadrado, ou seja, valores distantes tem um peso menor no cálculo do desempenho.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad (2)$$

Ao contrário das funções de custo mostradas acima, a métrica R^2 é uma função de utilidade e busca medir a qualidade das previsões, ou seja, uma métrica que quanto maior, melhor estarão sendo as previsões realizadas pelo algoritmo. A Equação 3 representa tal métrica.

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(\hat{y}_j - \bar{y}_j)^2} \quad (3)$$

Onde $\bar{y}$j corresponde à média dos dados de valores alvos $\hat{y}$j. Subtraindo de 1 a divisão do quadrado da diferença entre valores preditos e valores alvos, pelo quadrado da diferença entre valores alvos e a média dos valores alvos.

2.2.2 Divisão Treino e Teste

Para trabalhar com um conjunto de dados em um modelo de Aprendizado de Máquina é preciso pensar em duas etapas: Treino e Teste. Como diz o nome, a primeira etapa utiliza da base de dados para treinar o algoritmo, enquanto a segunda etapa testa o modelo utilizando da base de dados. Para as métricas de desempenho do teste representarem de fato o funcionamento do modelo com instâncias inéditas, ou vetores de atributos não rotulados, é preciso treinar e testar o modelo com diferentes partes do conjunto de dados.

É bastante comum utilizar-se de 80% do conjunto de dados para treinar o algoritmo, e usar os 20% restantes para realizar o teste do mesmo, essas divisões de dados são selecionadas de forma aleatória (GÉRON, 2019). Porém, só se pode realizar a divisão dessa forma se as instâncias de dados são independentes entre si, principio este que é quebrado em séries temporais, onde observações dependem de valores anteriores, além da dinâmica de geração de dados poderem mudar com o passar do tempo.

Com isso, é preciso definir um conjunto de dados para treino anteriores temporalmente ao conjunto de teste. Uma das formas de divisão de dados, que obtêm bom desempenho em séries temporais não-estacionárias, ou séries temporais com tendências e efeitos sazonais, é o modelo de “Janela Crescente com Validação Adiante”. (SCHNAUBELT, 2019)

No conceito de “Janela Crescente com Validação Adiante” para treinamento e validação do modelo, define-se a porção inicial da série temporal para treinar o algoritmo, usando a porção imediatamente posterior a essa para testar o modelo.

Já na segunda iteração, a porção inicial de treino se expande em tamanho e passa a compreender tanto a porção de treino como a porção de teste da primeira iteração, enquanto isso a nova porção de teste se torna a porção imediatamente posterior a essa

nova porção expandida de treino, mantendo o mesmo tamanho da primeira iteração. Fazendo com que com o passar das iterações, a razão de tamanho entre a porção de treino para a porção de teste fique cada vez maior, tendo em vista que o teste se mantém no tamanho inicial, apenas selecionando dados cada vez mais a frente nessa sequência. A Figura 1 ilustra esse processo.

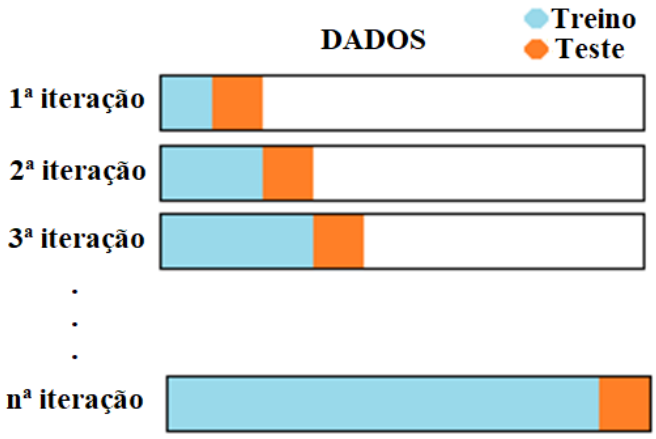


Figura 1: Ilustração do conceito de "Janela Crescente com Validação Adiante". Adaptado de Schnaubelt (2019).

2.2.2.1 Séries Temporais

Dentro dos tipos de dados que podem ser analisados estão as séries temporais, que são conjuntos de observações dispostos de maneira sequenciada pelo período de tempo. Como por exemplo um conjunto de dados que mostre a temperatura média diária de uma região ao longo de um ano.

Uma série temporal é estacionária quando fatores como média e variância não mudam com o passar do tempo, ou seja, a série teve seus movimentos de tendência e sazonalidade ajustados. É necessário realizar processos matemáticos para o ajuste dos dados com o fim de transformar uma série temporal não-estacionária em estacionária, visto que tal propriedade é um requerimento para algumas análises estatísticas. Porém, a estratégia "Janela Crescente com Validação Adiante" consegue obter bom desempenho em séries não-estacionárias. (SCHNAUBELT, 2019)

2.2.3 Algoritmos

Ainda não se sabe como fazer computadores aprenderem tão bem como as pessoas, mas foram inventados algoritmos, ou sequências de operações, que são efetivos para certas tarefas de aprendizado, e, com isso, um entendimento sobre o aprendizado começa a surgir.

2.2.3.1 Árvores de Decisão

Segundo Mitchell (1997), o aprendizado por meio das Árvores de Decisão pode ser simplificado para melhor entendimento, como um conjunto de perguntas condicionais “se” e “então”. Esse método de aprendizado classifica as instâncias do conjunto de dados, por meio de testes dos atributos dessa instância. Tais testes podem ser representados por nós da árvore, e se iniciam no nó raiz, enquanto os galhos descendentes dos nós correspondem aos possíveis valores do atributo.

Conforme os testes ocorrem, as instâncias percorrem as condições dos nós, descendo pelos galhos de acordo com seus atributos, e com isso chega-se ao galho final responsável por sua classificação no algoritmo. A Figura 2 representa um modelo de Árvore de Decisão para classificar se sábados de manhã são adequados para jogar Tênis, por meio de tipos de atributos como “Panorama”, “Umidade” e “Vento”.

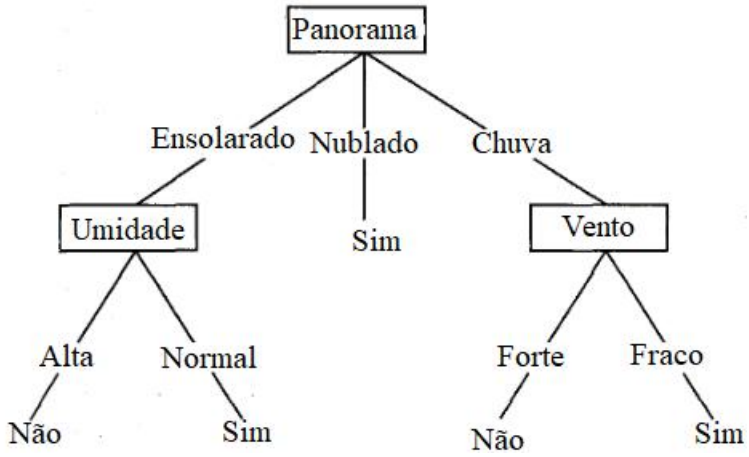


Figura 2: Modelo de Árvore de Decisão para classificar dias adequados para jogar Tênis. Traduzido de Mitchell (1997).

Imagina-se uma instância de uma base de dados com atributos tais como “Ensolarado”, “Alta Umidade” e “Ventos Fortes”. Tal instância seria classificada como um sábado não adequado para jogar Tênis, conforme o modelo de árvore anterior, que representa uma Árvore de Decisão Classificadora, dando uma resposta de “Sim” ou “Não” ao fim da sua execução.

Também é possível utilizar tal modelo de algoritmo com a função de Regressão, ao invés de Classificação. Isto é, previsão de valores ao invés de classes.

Com isso, pode-se pensar no algoritmo da Árvore de Classificação e Regressão (*Classification and Regression Tree* - CART). Que se inicia dividindo o conjunto de treinamento em dois subconjuntos, utilizando de um atributo “k” e um valor limite “tk”. (GÉRON, 2019)

É possível imaginar o par (k,tk) para o exemplo da Figura 2, onde k poderia ser Panorama, Umidade ou Vento, enquanto tk seria o valor que divide o conjunto em um primeiro subconjunto de dados em que o valor do atributo k é maior que tk e outro subconjunto em que o valor do atributo k é menor que tk.

O algoritmo então, busca pelo par (k, tk) que divide o conjunto de treinamento nos dois subconjuntos mais puros, ou seja, que minimizam a função de custo da Equação 4 a seguir.

$$J(k,t_k) = \frac{m_{\text{esquerda}}}{m} G_{\text{esquerda}} + \frac{m_{\text{direita}}}{m} G_{\text{direita}} \qquad (4)$$

Onde, “m” é o número de instâncias do conjunto, “$m_{esquerda}$” e “$m_{direita}$” do número de instâncias do subconjunto da esquerda e da direita respectivamente, “$G_{esquerda}$” e “$G_{direita}$” a medição da impureza (MAE, RMSE ou MSE para modelos de regressão) dos subconjuntos da esquerda e direita respectivamente.

Após realizar essa divisão inicial, o algoritmo busca dividir os subconjuntos por meio da mesma lógica, e assim por diante, em um processo recursivo. Parando o processo em uma profundidade máxima pré-determinada, ou se não encontrar uma divisão para reduzir a impureza. E com isso, construindo uma Árvore de Decisão, com seus nós e galhos.

Tais condições pré-determinadas, como a profundidade máxima da árvore, citada no parágrafo anterior, são denominadas de hiperparâmetros, nada mais do que opções de utilização do algoritmo, e de acordo com Géron (2019), podendo restringir o modelo conforme a necessidade do usuário, como uma tentativa de minimizar o sobreajuste, e melhorar a capacidade de generalização do modelo.

O modelo de Árvore de Decisão de Classificação também possui parâmetros similares, como o número mínimo de instâncias que um nó deve ter, ou número máximo de instâncias que um nó pode ter, sem contar com a restrição do número máximo de atributos que são avaliadas para divisão em cada nó.

2.2.3.2 Florestas Aleatórias

Por sua vez, segundo Géron (2019), o método de Florestas Aleatórias é um algoritmo de aprendizado em comitê, isto é, gera-se um resultado advindo de um conjunto de previsores (sejam estes classificadores ou regressores). Previsores estes que são as Árvores de Decisão da seção anterior.

Uma abordagem possível de Floresta Aleatória envolve a prática de *bagging*, utilizando do mesmo algoritmo de treinamento para cada previsor do conjunto, ou seja, várias Árvores de Decisão iguais, porém os previsores são treinados com diferentes subconjuntos aleatórios do conjunto total de treinamento.

Portanto, tal prática permite que o conjunto de treinamento possa ser amostrado várias vezes pelo mesmo previsor. Tal abordagem permite o uso de um modelo versátil como Árvores de Decisão, mas com uma variância menor, justamente por introduzir uma aleatoriedade extra para o processo, ao contar com a previsão de um sistema detentor de uma grande diversidade de previsões provenientes das árvores.

Na fase final de previsão nas Florestas Aleatórias, leva-se em conta em um modelo classificador, aquela previsão de classe que aparece de forma majoritária no conjunto de Árvores de Decisão. Enquanto num modelo de regressão, calcula-se a média do conjunto de resultados gerados pelo conjunto de previsores.

Além dos hiperparâmetros de Árvores de Decisão, as Florestas Aleatórias também contam com os hiperparâmetros da técnica de aprendizado baseado em comitês, como o controle da quantidade de Árvores de Decisão presentes no modelo.

2.3 TRABALHOS RELACIONADOS

Em Tiburcio (2018) foram aplicados algoritmos de Redes Neurais Artificiais para previsão de demanda de passageiros por dia, em uma linha específica do transporte público por ônibus da cidade de Joiville – SC, obtendo um Erro Médio Absoluto em percentual de 11% nos melhores modelos.

Já em Bezerra (2021) utilizou-se de 5 métodos diferentes do Aprendizado de Máquina para avaliar e comparar quantitativamente entre os algoritmos, os melhores resultados para previsão de demanda de passageiros, a partir de dados do sistema de transporte público urbano da cidade de Joinville – SC. O trabalho também investigou o efeito da quantidade de dados utilizado no processo de treinamento dos modelos.

Os melhores resultados de previsão foram obtidos a partir do modelo de Árvore de Decisão com regressão, obtendo um valor de 64% para R^2, e os piores no modelo de Regressão Linear, obtendo cerca de 62% para o R^2. (BEZERRA, 2021)

Em relação a quantidade de dados, viu-se para conjuntos com uma tendência de queda ou aumento da demanda, em outras palavras séries temporais não estacionárias, que os melhores resultados foram atingidos com as janelas deslizantes de 30 dias, que

percorrem a série temporal de 30 em 30 dias, diminuindo o efeito de dados antigos, visto que segundo a autora, dados de treinamento mais antigos podem levar a maiores erros de previsão. Uma visão contrária ao que foi estabelecido nesse trabalho, onde procurou-se utilizar da “Janela Crescente com Validação Adiante”, que usa cada vez mais dados de treino antigos, com base em Schnaubelt (2019) e conforme explicado na seção 2.2.2.

Em Pianucci et al. (2019), utilizou-se de Redes Neuras Artificias para modelagem da demanda por transporte e de dados sintéticos, através da geração de uma população sintética visando gerar uma base de dados sólida para análise. Os resultados atingidos nas Redes Neurais foram similares aos resultados obtidos por meio do método de Regressão Linear Múltipla.

Em Vasconcelos et al. (2021), fez-se uso das Redes Neuras Artificiais em configurações diferentes, com objetivo de modelar uma previsão de demanda para o sistema de estações de metrô e de trem da Região Metropolitana de São Paulo. Os melhores resultados foram obtidos pelo modelo que alcançou um erro médio quadrático de 0,045%.

Por fim, Silva (2019) investiga a aplicação de métodos baseados em Aprendizado de Máquina e Redes Neurais Artificias, para prever a demanda diária de refeições, do restaurante universitário da Universidade Federal de Uberlândia. Um conjunto de dados com valores bastante voláteis, influenciados por diversas características. Como sugestão de aprimoramentos, o autor sugere que se utilize métodos de aprendizado baseados em comitês, como as Florestas Aleatórias.

3 METODOLOGIA

As etapas da metodologia desse trabalho estão organizadas conforme o fluxograma da Figura 3.

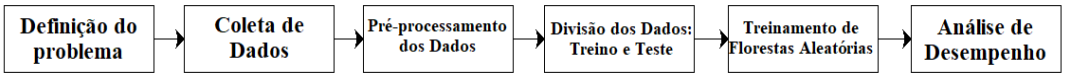


Figura 3: Fluxograma das etapas do projeto.

3.1 DEFINIÇÃO DO PROBLEMA

O problema consiste em definir um modelo preditivo envolvendo múltiplas variáveis, para estimar a demanda de passageiros do sistema de transporte coletivo. Isso será feito através de conjuntos de dados disponibilizados *online*, que devem ser tratados a fim de fornecer informações para um algoritmo de Aprendizado de Máquina treinar a capacidade de predição.

Através de experimentos com diferentes parâmetros, referentes a construção desse modelo, como quantidade de árvores na Floresta Aleatória e profundidade máxima dessas árvores, foi possível avaliar o desempenho do algoritmo com diferentes estruturas e diferentes entradas de dados, considerando períodos de pandemia e períodos de normalidade.

3.2 COLETA DE DADOS

Nesta seção, explica-se como os dados do trabalho foram obtidos.

3.2.1 Dados de Transporte

Foram utilizados dados do sistema público de ônibus referente à cidade de Belo Horizonte – MG, disponíveis dentro do Portal de Dados Abertos da Prefeitura de Belo Horizonte e na seção “Mapa de Controle Operacional Consolidado” (PREFEITURA DE BELO HORIZONTE, 2022). A escolha deste objeto de estudo foi motivada pela disponibilidade de dados.

Os arquivos possuem informações de demanda e oferta do sistema de ônibus de Belo Horizonte desde janeiro de 2016, sendo atualizados de forma mensal. Estes contêm conjuntos de informações referentes às viagens feitas por passageiros, dentro do município de Belo Horizonte, detalhadas para as diversas linhas e horários de partida. Contém informações da linha de ônibus da viagem, número do veículo, registro da catraca no início e ao fim da viagem, dentre outros atributos. Todos os atributos informados são dispostos na Tabela 1.

Tabela 1: Dicionário de Dados.

Nome do Atributo	Tipo	Descrição
VIAGEM	ALFANUMÉRICO	Data em que a viagem foi realizada
LINHA	ALFANUMÉRICO	Número da linha em que a viagem foi realizada
SUBLINHA	NUMÉRICO	Número da sublinha em que a viagem foi realizada
PC	NUMÉRICO	Número do Ponto de Controle de origem em que a viagem foi iniciada
CONCESSIONÁRIA	NUMÉRICO	Número da concessionária ao qual a linha está vinculada
SAÍDA	ALFANUMÉRICO	Hora de saída da viagem do PC de origem
VEÍCULO	NUMÉRICO	Número de ordem do veículo
CHEGADA	ALFANUMÉRICO	Hora de chegada da viagem do PC de destino
CATRACA SAÍDA	NUMÉRICO	Catraca registrada no início da viagem com 5 dígitos
CATRACA CHEGADA	NUMÉRICO	Catraca registrada no fim da viagem com 5 dígitos
OCORRÊNCIA	ALFANUMÉRICO	Indicador se houver interrupção de viagem
JUSTIFICATIVA	ALFANUMÉRICO	Indicador do tipo de justificativa da ocorrência
TIPO DIA	NUMÉRICO	Tipo do dia em que a viagem foi realizada
EXTENSÃO	NUMÉRICO	Extensão da viagem realizada, em metros
FALHA MECÂNICA	ALFANUMÉRICO	Indicador se houve falha mecânica durante a viagem
EVENTO INSEGURO	ALFANUMÉRICO	Indicador se houve evento inseguro durante a viagem
INDICADOR FECHAMENTO	ALFANUMÉRICO	Indicador se a viagem foi fechada
DATA FECHAMENTO	ALFANUMÉRICO	Data e hora do fechamento da viagem

Fonte: Portal de Dados Abertos da Prefeitura de Belo Horizonte.

Dependendo do número de viagens realizadas no mês, o arquivo contendo a planilha eletrônica pode conter até 990 mil linhas de informações, totalizando 4,77 GB de informações entre janeiro de 2016 e dezembro de 2021. Na Figura 4 ilustra-se o formato da base de dados a partir de um exemplar de parte da planilha de janeiro de 2016.

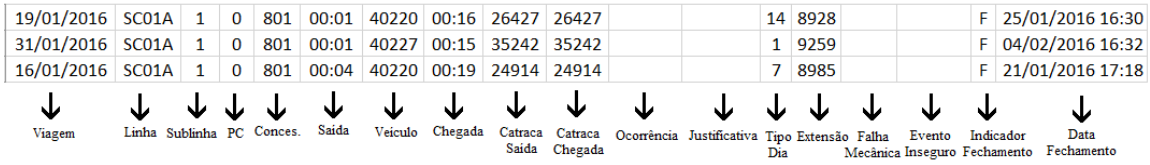


Figura 4: Exemplo do arquivo Mapa de Controle Operacional de janeiro-2016.
Fonte: Autor.

3.2.2 Dados de Clima

Com o objetivo de avaliar o efeito da condição climática na variação da demanda por transporte coletivo, extraiu-se das bases de dados históricas do Instituto Nacional de Meteorologia - INMET (INSTITUTO NACIONAL DE METEOROLOGIA, 2022), o conjunto de dados climáticos da cidade de Belo Horizonte de 2016 até os dias atuais, mesmo período temporal da base de dados do transporte coletivo.

Dentre as estações meteorológicas disponíveis, escolheu-se a Estação Meteorológica de Pampulha, localizada no campus da Universidade Federal de Minas Gerais (UFMG) e no meio da cidade de Belo Horizonte. Cada planilha eletrônica representa um ano de coleta de dados. Cada linha destas planilhas apresenta as condições climáticas horárias de cada dia do ano referente. Optou-se por utilizar, respectivamente, dos atributos “Precipitação Total (mm)” e “Temperatura do Ar (ºC)” para os índices de pluviometria e temperatura para o modelo. A Tabela 2 demonstra uma parte da planilha referente ao ano de 2016.

Tabela 2: Exemplo dos dados climáticos.

Data	Hora	Precipitação total (mm)	Temperatura do ar (ªC)
01/01/2021	0000 UTC	0,6	18,9
01/01/2021	0100 UTC	1,6	18,8
01/01/2021	0200 UTC	0,2	18,7
01/01/2021	0300 UTC	0,2	18,8

3.3 PRÉ-PROCESSAMENTO DOS DADOS

Para realizar o pré-processamento dos dados, usou-se da linguagem de programação Python 3.10.2, juntamente com as bibliotecas pandas (1.4.0), datetime (3.10.2) e numpy (1.22.2), para leitura e formatação de dados, e matplotlib (3.5.1) e seaborn (0.11.2), para geração de gráficos representativos do problema.

Essa etapa tem início na análise dos dados, para verificar possíveis erros e realizar tratamentos adequados para contorná-los. Após isso é preciso expandir os dados existentes em outros atributos, indiretamente conectado a essas informações, como, por exemplo, definir o dia da semana a partir de uma data. Por fim, separou-se os dados em 3 conjuntos de teste distintos em tamanho para observar o desempenho destes no algoritmo. Estes processos são descritos a seguir.

3.3.1 Análise dos Dados

Inicia-se o pré-processamento de dados com uma análise aparente, visando conhecer melhor a base de informações que está sendo trabalhada, verificar a existência de erros, revelar quantitativos e, também, guiar o processo de criação do algoritmo que melhor se adeque a esse aglomerado de elementos. Os erros identificados nessa análise serão enumerados com “PX”, onde X corresponde ao número identificador desse problema.

Para os dados de transporte separou-se as seguintes informações: data de ocorrência da viagem; código da linha de ônibus de cada viagem realizada, e; valor registrado na catraca no momento de saída e chegada dos veículos. A demanda de cada linha para cada horário foi calculada a partir da subtração do valor de passageiros registrados na chegada do veículo (fim da viagem) do valor registrado na saída do veículo (início da viagem).

Com isso, foi gerado um conjunto de dados com 57.627.733 linhas, que correspondem à cada viagem de cada linha de ônibus ocorrida dentro do período entre 1 de janeiro de 2016 e 31 de dezembro de 2021.

Na Figura 5, apresenta-se um gráfico de dispersão onde cada ponto representa uma linha de ônibus com operação, no período analisado neste trabalho (abscissa) e respectivo número de passageiros registrados na viagem (ordenada). Percebe-se, a princípio, o efeito da pandemia no número de passageiros por volta de março de 2020, coincidindo com as medidas de restrição impostas para controle da pandemia da COVID-19. Observa-se, ainda, algumas anomalias na base de dados de demanda, como valores zerados **(P1)** e valores muito altos, tanto no espectro dos números positivos **(P2)** como dos negativos **(P3)**. A Figura 6 expande a escala do eixo das ordenadas do gráfico ilustrado na Figura 5, para deixar ainda mais evidente estes valores discrepantes.

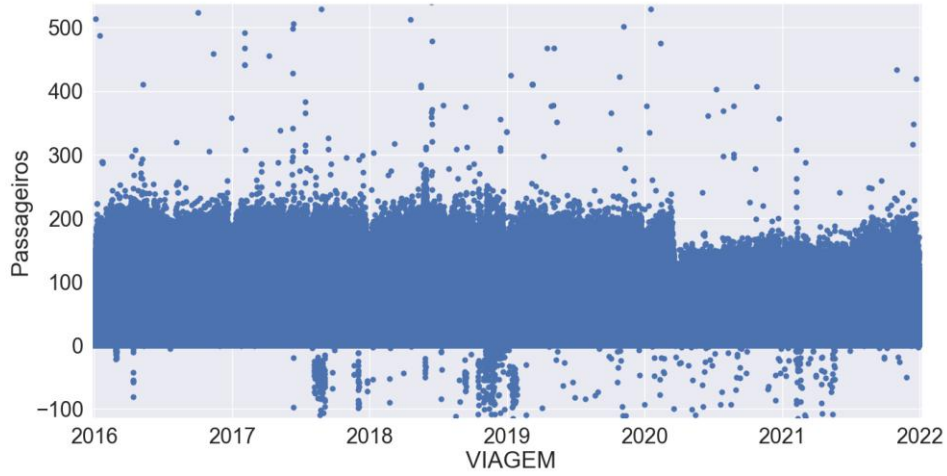


Figura 5: Número de passageiros do conjunto dos dados de transporte.

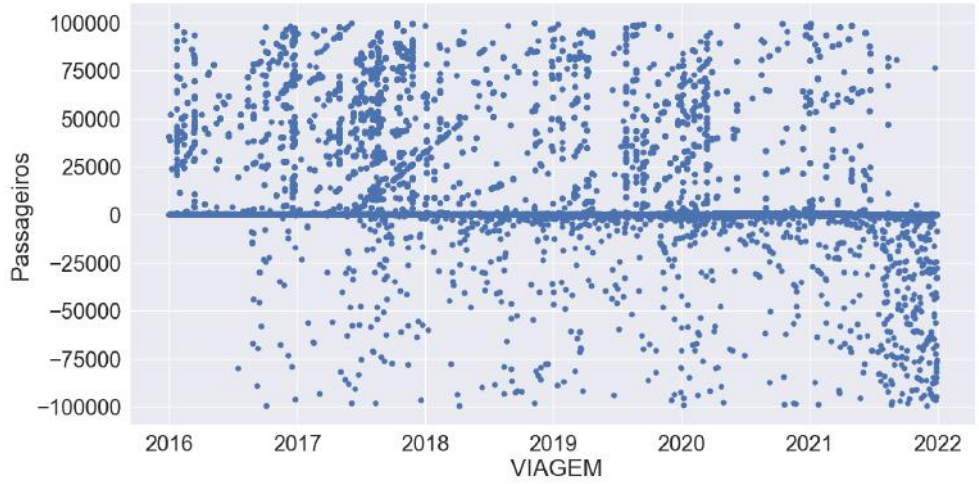


Figura 6: Anomalias do conjunto dos dados de transporte.

Este conjunto apresenta uma média de 34,25 passageiros por viagem, mesmo possuindo claras anomalias (que serão tratadas nas seções seguintes). As linhas que possuem valores negativos **(P3) e** altamente positivos **(P2)** representam, respectivamente, 0,027% e 0,019% do total do conjunto de dados. Já as linhas com valor de passageiro igual a 0 **(P1)** representam cerca de 30% do conjunto de dados. Acredita-se que a ausência de passageiros justifica-se, principalmente, à erros de registro do valor das catracas ou viagens de fato sem passageiros.

Igualmente, para a base de dados climática separou-se as colunas referentes à data e hora da medição, além dos índices de temperatura e precipitação. Essas medições são dispostas a cada hora de cada dia.

Em seguida, gerou-se um conjunto de dados com 52.608 linhas. Cada uma dessas linhas do conjunto representam uma hora dentro do período compreendido entre 1 de

janeiro de 2016 até 31 de dezembro de 2021, acompanhada das medições de temperatura e precipitação. Na Figura 7 ilustra-se a distribuição dos valores de temperatura, e na Figura 8 a distribuição dos valores de precipitação. Ambos os conjuntos apresentam valores anômalos muito negativos **(P4)** e outros vazios **(P5)**, como destacado na Figura 9. Estas divergências serão corrigidas nas seções adiante.

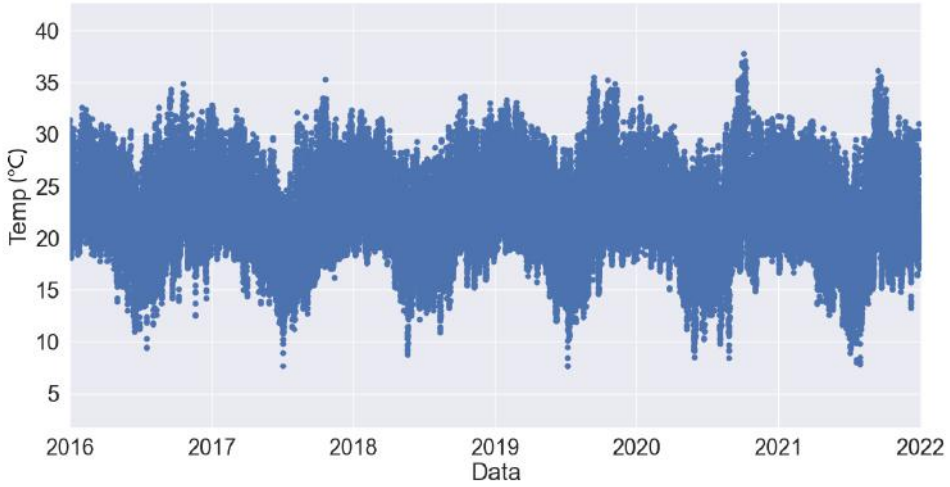


Figura 7: Distribuição dos valores de temperatura.

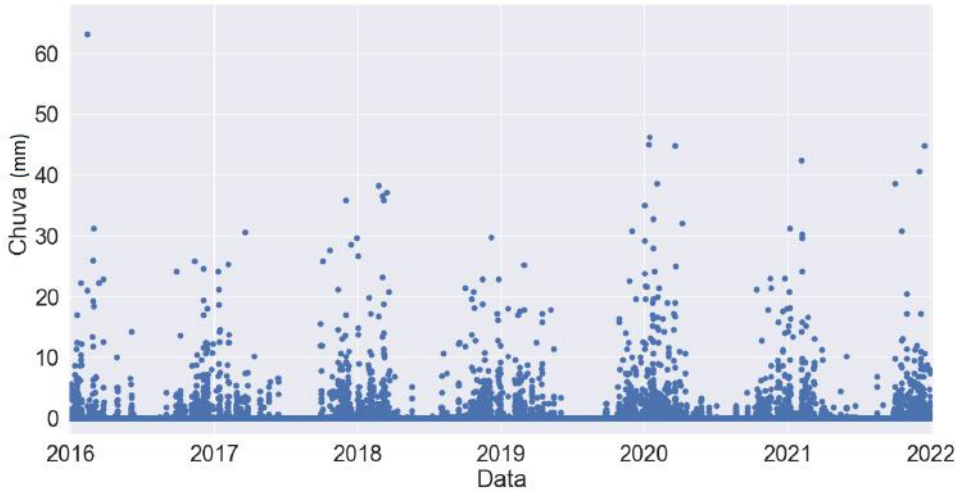


Figura 8: Distribuição dos valores de precipitação.

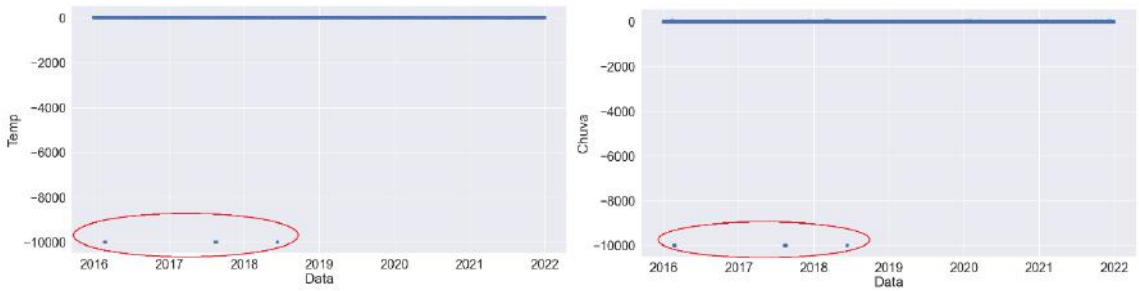


Figura 9: Exemplificação das anomalias nas séries de Temperatura e Precipitação.

Estes valores altamente negativos **(P4),** somados dos valores vazios **(P5)** que os dois conjuntos possuem, representam 0,26% de valores anômalos no conjunto de temperatura e 0,22% no conjunto de precipitação. Os valores de média por hora antes do tratamento dos dados são respectivamente 12,47 ºC e -5,3mm. Demonstrando que, apesar dos conjuntos terem baixa quantidade de valores anômalos, estes afetam certas aferições realizadas, como a média. Isto é ainda mais perceptível nas estatísticas de precipitação, pois assumem uma quantia expressiva negativa de precipitação.

3.3.2 Tratamento de Problemas nos Dados

Visando construir um modelo computacional mais fidedigno à realidade, deve-se tratar ou corrigir problemas dos conjuntos de dados, como aqueles descritos na seção anterior e enumerados a seguir.

O tratamento da base de dados de transporte teve início abordando-se o problema dos números negativos (**P3)**. Percebeu-se que cerca de 90% dos valores negativos de passageiros aconteciam por causa do funcionamento da catraca, que registra números de até 5 dígitos (00000 a 99999). Assim, quando o contador está registrando o passageiro nº 99999 e um novo passageiro embarca, a catraca reinicia a contagem, voltando para o nº 00000 . Portanto, mudou-se a lógica da construção matemática da coluna de passageiros no conjunto de dados, que antes só subtraía um valor do outro, e, que nesse caso, gerava um valor negativo, pois o valor da catraca ao fim da viagem era menor do que no início. Criou-se, então, uma exceção visando corrigir esse problema, avaliando que quando tal subtração fosse negativa e o valor da catraca no início da viagem for de pelo menos 99800, iria-se subtrair o valor de catraca do início da viagem de 100000, para obter quantos passageiros entraram até a catraca atingir seu registro máximo, além de somar o valor da catraca ao fim da viagem, para obter ainda quantos passageiros entraram após a catraca ter atingido o seu registro máximo.

Nesta exceção, foi estabelecido esse valor mínimo de 99800 para o valor da catraca no início da viagem, a fim de evitar a correção de valores negativos de passageiros que decorrem de erros de registro numérico (como de anotação e digitação). Como, por exemplo, valores de catraca ao fim da viagem menores do que o valor de início, porém o valor de início distante do valor máximo da catraca.

Após esta correção, os valores negativos de passageiros, que representavam aproximadamente 0,02% do total, agora representam 0,002% da base de dados e foram descartados.

Visto que o propósito do modelo é criar um algoritmo de predição de demanda total para o sistema de transporte coletivo e, para isso, vai-se somar a quantidade de passageiros transportadas por cada linha em cada dia, as viagens com número de passageiros igual a 0 **(P1)** tornam-se passíveis de serem desconsideradas. Isto é, as viagens com valor de catraca igual no início e fim da viagem não contribuem em nada para a soma final de cada dia de cada linha de ônibus, mas demandam maior poder de processamento na etapa de teste o modelo.

Para os valores altamente positivos **(P2)**, precisou-se, inicialmente, definir o que seria um valor limitante superior ("alto demais") de passageiros transportados em uma viagem. Este limitante superior passa a ser utilizado como referência de corte: são assumidos como erro valores acima desse limite e, então, passam a ser desconsiderados. Recorreu-se a essa solução pois, de um lado, há valores absurdos que devem ser desconsiderados, por se tratarem de erros. Porém, há a necessidade de identificar e incluir no conjunto final os valores de alta demanda, possíveis de acontecer no cenário do transporte público de Belo Horizonte.

Visando realizar tal estimativa, encontrou-se no Portal de Dados Abertos da Prefeitura de Belo Horizonte, na seção de "Números do Transporte Coletivo", a média de passageiros registrados por veículo, por mês e por consórcio no período de 2021 até 2017, como apresentado na Tabela 3, um exemplo desta média do ano de 2017. Esse formato de tabela foi a única referência encontrada que proporcionava uma ordem de grandeza para a quantidade de passageiros separadas por veículo, e não foram encontrados dados para o ano de 2016. O motivo pela escolha da tabela de 2017 será esclarecido a seguir.

Tabela 3: Média de passageiros registrados por veículo/mês/consórcio.

Mês/Ano	Consórcio Pampulha	Consórcio BH Leste	Consórcio Dez	Consórcio D. Pedro II	Média do Sistema
Fev/17	11356	9905	10024	6948	9706
Mar/17	13792	11725	11962	8443	11655
Abr/17	11744	9924	10150	7142	9883
Mai/17	13962	11804	11957	8469	11718
Jun/17	13153	11186	11336	7965	11072
Jul/17	12609	10863	10948	7690	10687
Ago/17	**13959**	11799	11954	8486	11719
Set/17	13119	11101	11247	7936	11012
Out/17	13083	11301	11254	8153	11123

A partir dos dados disponíveis, pode-se concluir que a maior média de passageiros por veículo e por mês ocorreu no mês de agosto de 2017, no consórcio da Pampulha, com 13959 passageiros. Nota-se que os valores médios tendem a aumentar conforme os anos regridem no período de 2021 até 2017, por isso escolheu-se apresentar a Tabela 3 referente a 2017, para realizar a estimativa com os maiores valores registrados no período.

O mês de agosto de 2017 possuiu 23 dias úteis, 4 sábados e 4 domingos. Sabe-se que a demanda dos sistemas de transportes aos domingos tende a ser menor do que aos sábados e que, por sua vez, tende a ser menor do que nos dias úteis. Assim, considerou-se que a demanda aos sábados e domingos correspondem, respectivamente, à 75% e 50% dos dias úteis. Por tanto, multiplicando-se essas porcentagens à quantidade de dias, obteve-se que este mês possuiu 28 dias equivalentes.

A divisão do maior valor de média de passageiros por veículo e por mês (13.959 passageiros) por 28 dias equivalentes resulta em um valor limite conservador por viagem de 500 passageiros. Embora a capacidade típica dos ônibus varie entre 80 e 160 passageiros, dependendo as características físicas do veículo, há, no sistema de Belo Horizonte, linhas bastante extensas, que cortam a cidade de uma ponta à outra. Assim, com a renovação de passageiros (desembarques e embarques ao longo do percurso), o registro de passageiros por itinerário pode ser superior à capacidade dos veículos. Com isso, descartou-se, da base de dados, todos os registros que apresentavam viagens acima desse valor limite, consideradas aqui como valores irreais, podendo ser decorrentes de erros na hora de registrar. Estes registros representavam apenas 0,019% do conjunto inteiro de dados.

Para o conjunto de dados de precipitação, substituiu-se os valores inexistentes **(P5)** e negativos **(P4)** por 0. A decisão de preencher com zero justifica-se por ser um valor aceitável para precipitação, visto que mais de 90% dos registros de chuva são iguais a 0mm. Com essa substituição o valor médio de precipitação por hora resulta em 0.17mm.

No conjunto de temperatura, substitui-se os valores altamente negativos **(P4)** por valores inexistentes **(P5)**, ou seja, vazios de informação. Acabando com o problema **(P4)** e aumentando o **(P5)**, seguiu-se tal estratégia pois estes valores inexistentes não influenciam tanto a média de temperatura dos valores da base de dados, como os valores altamente negativos fazem mesmo sendo minoria absoluta. Por conta dessa modificação, a média dos valores de temperatura foi de 12,47ºC para cerca de 22ºC, pois a medição anterior contava com alguns valores altamente negativos tais quais “-3000ºC” **(P4)**. A

partir dessa mudança, substitui-se todos valores inexistentes por essa média de temperatura igual a 22ºC, visando completar a base de dados climática com o tratamento de **(P5)**.

3.3.3 Desmembramento de Atributos

Feito o tratamento dos problemas encontrados nas bases de dados, procurou-se aumentar a quantidade de atributos e informações através de dois processos: i) desmembramento dos dados já existentes, e; ii) agrupamento de viagens ocorridas no mesmo dia na mesma linha de ônibus, visando obter uma soma diária de passageiros transportados para cada linha por dia. O resultado da aplicação desses processos foi uma base de dados com 595.979 linhas de dados. Esses processos são detalhados a seguir, enquanto, na Tabela 4, pode-se verificar um recorte de como ficou a base dados formatada e agrupada, que será utilizada nos treinamentos.

No processo de desmembramento, inicialmente dividiu-se a informação de data, antes apresentada em apenas uma coluna com a data completa, em três novas colunas: ”Ano”, ”Mês” e ”Dia”. Adicionou-se, ainda, uma quarta coluna contendo o dia da semana (“Dia da Semana”), que, a partir da data, estabelece valores de 0 até 6, onde 0 representa segunda-feira e 6 o domingo. Por fim, criou-se a coluna para indicar a numeração da semana do ano.

Também foram utilizados calendários de Belo Horizonte para compor uma variável binária que indique as datas que são feriados. Com isso criou-se a coluna “Feriado”, onde todo dia de feriado recebe o valor 1 e o restante recebe 0. Criou-se, ainda, as colunas ”Pré-feriado” e ”Pós-feriado”, também com variáveis binárias, visando gerar atributos que podem influenciar na demanda do sistema de ônibus. Essas recebem o valor de 1 quando atendem a condição estabelecida por seu nome e 0 quando não atendem.

O mesmo processo de agrupamento foi realizado na base de dados climática, que possui dados a cada hora de cada dia do ano. Agrupou-se os valores por dia e, a partir desse agrupamento, foram gerados novos atributos, como: Temperatura média do dia, Temperatura máxima do dia, Temperatura mínima do dia e Precipitação total do dia.

Ao fim desse processo, gerou-se a coluna “Pandemia”, que fornece o valor 1 para todas as datas a partir de 15 de março de 2021, para determinar o início das possíveis consequências no transporte decorrentes pandemia no Brasil, e 0 para o restante.

Tabela 4: Recorte da base de dados final.

Ano	Mês	Dia	Linha	Passageiros	Dia da Semana	Semana do Ano	Pré-Feriado	Feriado	Pós-Feriado	Pandemia	Chuva (mm)	Temp max.	Temp min.	Temp med.
2016	01	04	125	3993	0	1	0	0	0	0	12	25,4	19,7	21,17
2018	09	09	195	1531	6	36	0	0	0	0	0	27,7	15,13	20,76
2020	09	19	140	405	5	38	0	0	0	1	0	35,1	21,1	27,64

A Tabela 5 apresenta, a partir da análise das bases de dados, os valores de máximo, mínimo, desvio padrão e média das variáveis de passageiros, precipitação, temperatura máxima, temperatura mínima e temperatura média.

Tabela 5: Métricas para os atributos da base de dados final.

Coluna	Valor Máximo	Valor Mínimo	Desvio Padrão	Média
Passageiros	23.753	1	2.669,91	2.596,81
Chuva	152	0	11,94	4,16
Temperatura Máxima	37,08	15,5	2,99	27,47
Temperatura Mínima	23,90	7,70	2,70	17,78
Temperatura Média	29,69	13,62	2,47	22,01

3.3.4 Seccionando os Dados

A principal motivação para a divisão da base de dados foi a pandemia, que afetou a demanda de passageiros do sistema, como evidenciado na Figura 5. Dividiu-se, então, o conjunto de dados em três instâncias para serem testadas separadamente e verificar o desempenho de aprendizado do algoritmo para cada uma delas: i) período completo, ii) anos sem pandemia e iii) anos com pandemia.

Escolheu-se por testar, inicialmente, o conjunto inteiro de dados, ou seja, para todo o período de 1 de janeiro de 2016 até 31 de dezembro de 2021. O segundo conjunto a ser testado será para o conjunto que contém os anos sem pandemia, compreendido entre 1 de janeiro de 2016 e 31 de dezembro de 2019, equivalente a cerca de 67% do conjunto completo. Por fim, testa-se o conjunto de dados para os anos com pandemia, que compreende o período entre 1 de janeiro de 2020 até 31 de dezembro de 2021 (os dois anos pandêmicos, contando com os primeiros meses de 2020 considerados como período de normalidade), equivalente a cerca de 33% do conjunto completo.

3.4 DIVISÃO DE DADOS: TREINO E TESTE

Nesse trabalho, se fez uso de séries temporais com tendências e efeitos sazonais como dados de entrada. Isto é, um tipo de informação sequencial, característica primordial que impede uma seleção aleatória para treino e teste, visto que as informações não são

independentes entre si e não faz sentido treinar o modelo usando valores do futuro para prever valores do passado. Como apresentado anteriormente, o método de “Janela Crescente com Validação Adiante” possui bom desempenho para a divisão das bases de dados entre treino e teste em um cenário de séries temporais. No conceito de “Janela Crescente com Validação Adiante” para treinamento e validação do modelo, deve-se, inicialmente, definir uma porção inicial da série temporal para treinar o algoritmo. Posteriormente, faz-se uso da porção imediatamente posterior à de treino para testar o desempenho do modelo.

Tendo em vista que cada instância de dados, descritas na Seção 3.3.4, compreendem períodos com durações variadas entre si (período completo, anos sem pandemia e anos com pandemia), estas acabam possuindo quantidades diferentes de dados. Por isso, para que pudessem ser divididas em porções com quantidades aproximadamente iguais de dados a serem testados, tiveram que ser divididas por denominadores diferentes.

O tamanho das porções foi determinado para ser equivalente a cerca de 3 meses ou 90 dias. Por exemplo, a 1ª iteração utiliza como dados de treino os 3 primeiros meses do período e, então, a validação ocorre com os 3 meses posteriores. Na 2ª iteração, o treinamento ocorre utilizando-se os 6 primeiros meses do período, enquanto a validação utiliza os próximos 3 meses e assim sucessivamente até a última iteração, conforme ilustrado na Figura 1 da seção 2.2.2.

Porém, como cada linha do conjunto de dados representa um agrupamento por linha de ônibus e por dia, existem dias com mais e outros com menos linhas de dados. Por exemplo, o dia 1 de janeiro de 2016 possui 247 linhas de dados, cada uma representando uma linha de ônibus no dia em questão. Enquanto o dia 2 de janeiro de 2016 possui 272 linhas de dados, visto que em tal dia houve o registro de passageiros em 25 linhas de ônibus a mais do que no dia anterior.

Como o algoritmo interpreta apenas as linhas de dados, para fazer a divisão dos conjuntos de treino e teste, este divide o conjunto inteiro em partes iguais de dados, mas não necessariamente em partes iguais de períodos temporais, visto que um dia pode ter registros de passageiros em um número diferente de linhas de ônibus do que o próximo dia. Ou seja, quantidades iguais de dados que compreendem períodos de tempo diferentes.

Inclusive, esse foi um dos motivos por trás da escolha de prosseguir a divisão treino e teste seguindo a ideia da “Janela Crescente com Validação Adiante”, visto que nenhum dado para treino é descartado. Logo, por mais que a validação venha a compreender

períodos ligeiramente diferentes de tempo em cada iteração, mesmo assim existirão dados no conjunto de treino para tais períodos.

A partir disso, separou-se na Tabela 6, Tabela 7 e Tabela 8 respectivamente os períodos compreendidos em cada iteração de teste para as instâncias de: período completo, anos sem pandemia e anos com pandemia. Na primeira linha das tabelas está o período referente ao primeiro conjunto de treino. Estas foram as melhores divisões possíveis visando manter um período aproximadamente uniforme de 90 dias.

Tabela 6: Divisões de teste para a instância de período completo.

Iteração	Data inicial	Data final	Período (dias)
Primeiro conjunto de treino	01/01/2016	28/03/2016	87
0	28/03/2016	22/06/2016	86
1	22/06/2016	17/09/2016	87
2	17/09/2016	13/12/2016	87
3	13/12/2016	10/03/2017	87
4	10/03/2017	04/06/2017	86
5	04/06/2017	30/08/2017	87
6	30/08/2017	24/11/2017	87
7	24/11/2017	19/02/2018	87
8	19/02/2018	16/05/2018	86
9	16/05/2018	11/08/2018	87
10	11/08/2018	16/11/2018	97
11	16/11/2018	10/02/2019	86
12	10/02/2019	07/05/2019	86
13	07/05/2019	01/08/2019	86
14	01/08/2019	25/10/2019	85
15	25/10/2019	20/01/2020	87
16	20/01/2020	16/04/2020	87
17	16/04/2020	16/07/2020	91
18	16/07/2020	15/10/2020	91
19	15/10/2020	13/01/2021	90
20	13/01/2021	12/04/2021	89
21	12/04/2021	09/07/2021	88
22	09/07/2021	03/10/2021	86
23	03/10/2021	31/12/2021	89
Média			87,70

Tabela 7: Divisões de teste para a instância de anos sem pandemia.

Iteração	Data inicial	Data final	Período (dias)
Primeiro conjunto de treino	01/01/2016	26/03/2016	87
0	26/03/2016	20/06/2016	86
1	20/06/2016	14/09/2016	86
2	14/09/2016	09/12/2016	86
3	09/12/2016	04/03/2017	85
4	04/03/2017	29/05/2017	86
5	29/05/2017	22/08/2017	85
6	22/08/2017	15/11/2017	85
7	15/11/2017	09/02/2018	86
8	09/02/2018	05/05/2018	85
9	05/05/2018	30/07/2018	86
10	30/07/2018	03/11/2018	96
11	03/11/2018	28/01/2019	86
12	28/01/2019	22/04/2019	96
13	22/04/2019	16/07/2019	86
14	16/07/2019	08/10/2019	84
15	08/10/2019	01/01/2020	85
Média			86,05

Tabela 8: Divisões de teste para a instância de anos com pandemia.

Iteração	Data inicial	Data final	Período (dias)
Primeiro conjunto de treino	01/01/2020	20/03/2020	79
0	20/03/2020	11/06/2020	83
1	11/06/2020	03/09/2020	84
2	03/09/2020	24/11/2020	82
3	24/11/2020	14/02/2021	82
4	14/02/2021	06/05/2021	81
5	06/05/2021	25/07/2021	80
6	25/07/2021	11/10/2021	78
7	11/10/2021	31/12/2021	81
Média			81,1

3.5 TREINAMENTO DE FLORESTAS ALEATÓRIAS

Com auxílio da biblioteca *scikit-learn*, além das bibliotecas já comentadas na seção 3.3, que auxiliam na montagem de gráficos e formatação de informações, foi possível treinar o modelo descrito através do método das Florestas Aleatórias.

O treinamento partiu de um conjunto inicial de hiperparâmetros de Floresta Aleatória, utilizando-se as instâncias apresentadas. Visando buscar o modelo mais efetivo para o problema em estudo, considerou-se os valores atribuídos aos hiperparâmetros mudando com o passar dos experimentos, e com isso gerando diferentes modelos de Floresta Aleatória, resultando em gráficos e medições estatísticas para os múltiplos testes realizados. As variações utilizadas nos experimentos estão definidas na Tabela 9.

Tabela 9: Variações de hiperparâmetros para o treinamento de Florestas Aleatórias.

Modelo	**Hiperparâmetros**	
	Quantidade de árvores	**Profundidade máxima da árvore**
A2 N5	2	5
A2 N10	2	10
A2 N20	2	20
A2 N50	2	50
A2 N100	2	100
A10 N5	10	5
A10 N10	10	10
A10 N20	10	20
A10 N50	10	50
A10 N100	10	100
A20 N5	20	5
A20 N10	20	10
A20 N20	20	20
A20 N50	20	50
A20 N100	20	100
A50 N5	50	5
A50 N10	50	10
A50 N20	50	20
A50 N50	50	50
A50 N100	50	100
A100 N5	100	5
A100 N10	100	10
A100 N20	100	20
A100 N50	100	50
A100 N100	100	100

3.6 ANÁLISE DE DESEMPENHO

Para analisar o desempenho dos modelos, considerou-se métricas estatísticas como R^2 e Erro Médio Absoluto, calculadas no decorrer das iterações de aprendizado. Para a escolha do modelo mais bem ajustado, observou-se aquele que apresenta as melhores métricas de desempenho em termos absolutos, ou seja, maiores valores de R^2 (tendendo a 1) e menores valores de Erro Médio Absoluto. Calculou-se, ainda, o tempo de processamento decorrido por cada modelo, além de outras métricas estatísticas como valor máximo, valor mínimo, média e desvio padrão.

O modo de representação dessas medições foi através de diagramas de caixa, ou *box plot*, que permitem uma visualização da distribuição dos dados por meio de uma simbologia que inclui o 1º e 3º quartis, mediana, limite inferior e limite superior. Esses dois últimos calculados através da diferença do 3º e 1º quartil, que é então multiplicada por 1,5. Somando esse valor com o 3º quartil, se obtêm o limite superior e, subtraindo esse mesmo valor do 1º quartil, se obtêm o limite inferior. Dados que se encontram além desses limites são pontos considerados discrepantes da distribuição dos dados. Enquanto a diferença entre o 3º e 1º quartil representado pela caixa, é chamado de intervalo interquartil.

4 RESULTADOS E DISCUSSÕES

Nessa seção, apresenta-se os resultados desse trabalho, acompanhado de discussões referente aos experimentos realizados com diferentes hiperparâmetros de Florestas Aleatórias. Discute-se, ainda, a comparação de desempenho do algoritmo para as três instâncias, explicadas na Seção 3.3.4, sendo estas: i) período completo, ii) anos sem pandemia e iii) anos com pandemia. Destaca-se, ao final desta, o modelo que apresentou melhor ajuste considerando-se as métricas estatísticas elencadas.

4.1 INSTÂNCIA: PERÍODO COMPLETO (PC)

Os primeiros experimentos foram realizados com o conjunto completo de dados, compreendendo o período entre 1 de janeiro de 2016 até 31 de dezembro de 2021. A Figura 10, apresenta o diagrama de caixa do R^2 para todas as iterações do processo de janela crescente com validação adiante por modelo de Floresta Aleatória, conforme descrito na seção 3.5. O eixo das abscissas representa os modelos, que possuem hiperparâmetros variados conforme descrito na Tabela 9. Os valores que pospõem-se as letras A e N neste eixo referem-se à quantidade de árvores do modelo e à profundidade máxima da árvore (conforme Tabela 9).

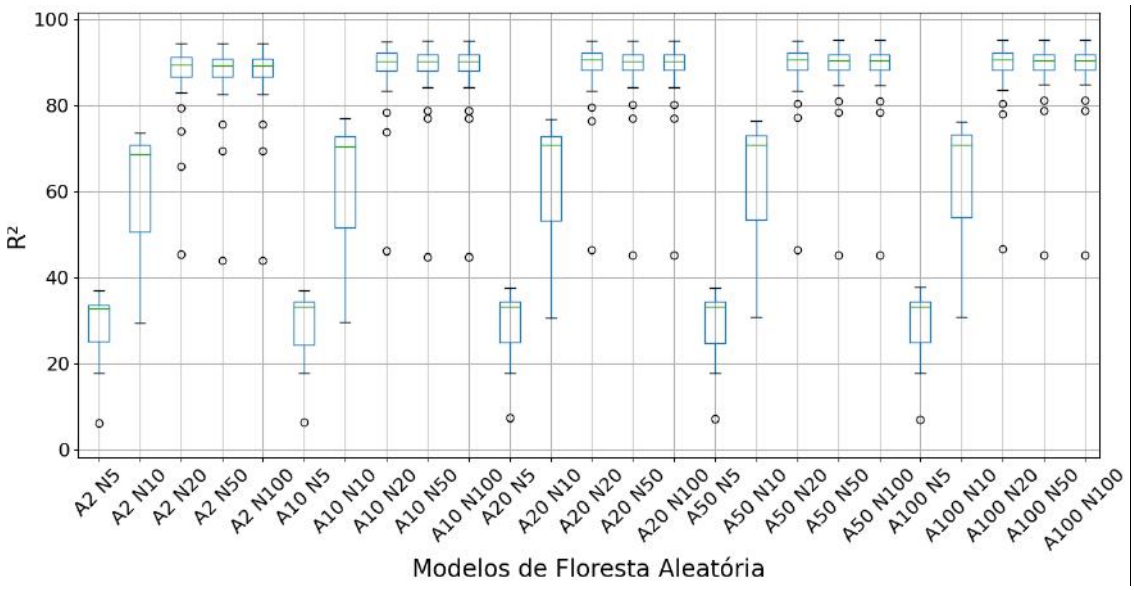


Figura 10: Diagrama caixa de R^2 para todos os modelos de Floresta Aleatória no conjunto de dados do período completo.

Valores abaixo de zero e muito discrepantes surgem apenas nos dois primeiros modelos, com uma profundidade máxima da árvore de até 10 e quantidade de árvores igual a 2, estes valores foram cortados da figura para melhor visualização da mesma, além dos modelos com profundidade máxima da árvore de até 10 apresentarem valores de R^2 abaixo do restante dos modelos, e a mediana se encontra deslocada para próximo do 3º

quartil, indiciando uma assimetria negativa, ou seja, mais resultados próximos aos valores mais altos, além de possuírem uma maior dispersão de dados que pode ser observado na Figura 10 pela diferença em tamanho do intervalo interquartil. Comparando os modelos “A2 N5” e “A2 N100”, nota-se um aumento de 63% no valor de mediana. A partir dos valores de profundidade 20, os modelos se tornam bem similares.

Com o aumento da quantidade de árvores, também se torna possível ver uma ligeira diminuição dos valores discrepantes, onde o modelo “A2 N20” possui 4 pontos discrepantes (45.3, 65.6, 73.9, 79.4), e o modelo “A100 N20” apenas 3 pontos discrepantes (46.54, 77.92, 80.44). Também houve um aumento de 1% no R^2, na comparação dos mesmos modelos. Demonstrando que o efeito no aumento da profundidade máxima da árvore tem um impacto maior nas métricas de desempenho, do que alterações na quantidade de árvores.

Nota-se na Figura 11, que remete ao diagrama caixa do Erro Médio Absoluto, a repetição de um padrão similar de desempenho dos modelos a partir de um valor de profundidade máxima da árvore igual a 20. Os modelos acima desse valor apresentam uma mediana por volta de 360 passageiros. Mostrando apenas uma diferença de 7,5% com o aumento da quantidade de árvores, como visto entre o modelo “A2 N20” que tem mediana de 372 passageiros, e o modelo “A100 N20” que tem mediana igual a 344 passageiros. Enquanto na comparação com a mudança da profundidade máxima da árvore, entre “A2 N5” e “A2 N100”, mostra uma diminuição de 76% no valor de Erro Médio Absoluto.

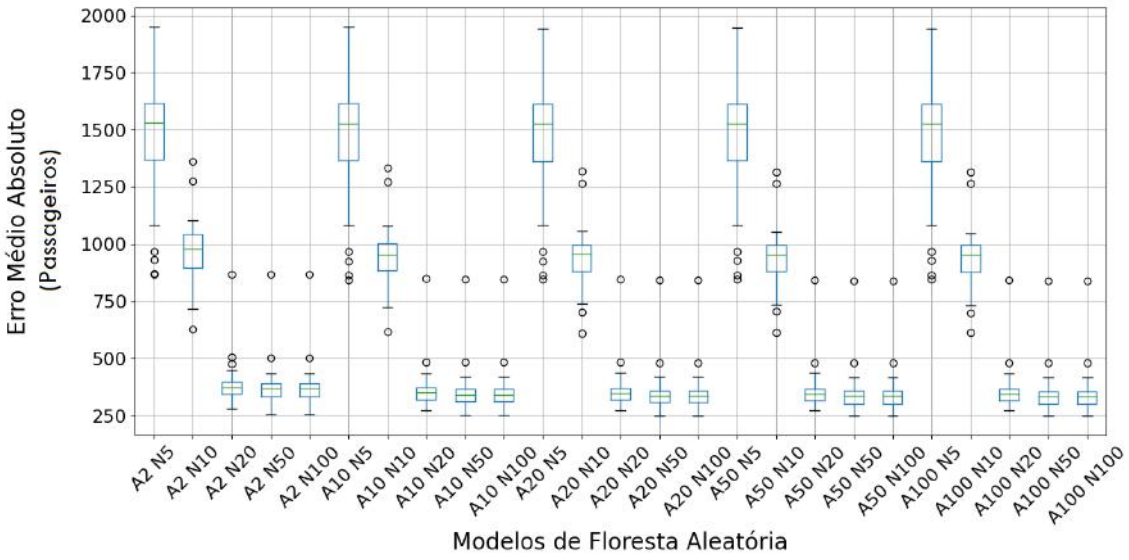


Figura 11: Diagrama caixa da Erro Médio Absoluto para todos os modelos de Floresta Aleatória, no conjunto de dados do período completo.

Os modelos de Floresta Aleatória que obtiveram os maiores valores de R^2, ao mesmo tempo, possuem os menores valores de Erro Médio Absoluto, mostrando que as métricas estão condizentes entre si. Isto é, pela Figura 11, observa-se que o menor resultado para Erro Médio Absoluto tem a mediana de 334 passageiros, modelo este que será abordado a seguir e possui o R^2 mais alto dentre os outros, com uma mediana de 90,3.

Para aprofundar a investigação tanto dos resultados de R^2 como de Erro Médio Absoluto, é preciso analisar as iterações da janela crescente com validação adiante do modelo mais ajustado em números absolutos para esse conjunto de dados. Para isto, selecionou-se o modelo "A100 N100", que possui todos hiperparâmetros em seu valor máximo e que apresentou melhor ajuste. Na Figura 12, pode-se observar, respectivamente, a projeção do R^2 e do Erro Médio Absoluto por iteração da janela.

Percebe-se que, justamente na 16ª iteração da janela deslizante, as métricas atingem seus piores valores e, no caso do R^2, ainda demoram algumas iterações para retornar ao desempenho até então obtido. Como explicado na seção 3.3.4, cada iteração da janela deslizante corresponde a aproximadamente 3 meses de dados. Portanto, a 16ª iteração é justamente a parte do algoritmo em que se inclui o período pandêmico de 2020.

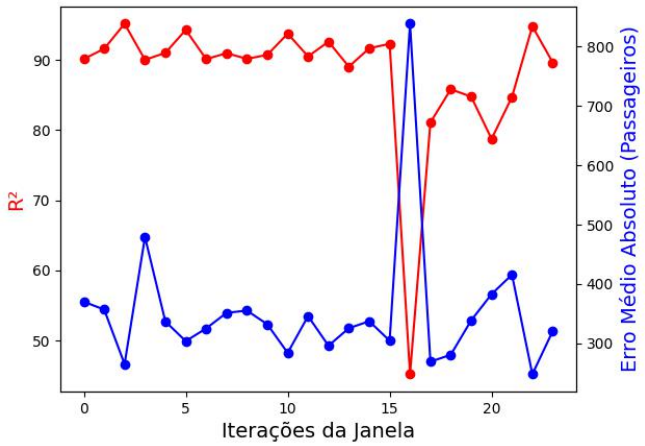


Figura 12: Valor de R^2 e Erro Médio Absoluto por iteração de janela do modelo "A100 N100", no conjunto do período completo.

Também foi selecionado o modelo "A10 N50" para uma comparação com o modelo mais ajustado em números absolutos, visto que o tempo decorrido para execução dos algoritmos é diferente e será analisando adiante. Com isso, a Tabela 10 e Tabela 11 mostram métricas de valor máximo, mínimo, desvio padrão e média para o modelo mais ajustado, "A100 N100", e o seu comparativo "A10 N50". Demonstrando modelos com desempenho similar.

Tabela 10: Resultados de R^2, no conjunto de dados do período completo.

Modelo	Valor máximo	Valor mínimo	Desvio Padrão	Média
A100 N100	95,15	45,22	9,94	87,85
A10 N50	94,90	44,83	10,13	87,51

Tabela 11: Resultados de Erro Médio Absoluto, no conjunto do período completo.

Modelo	Valor máximo	Valor mínimo	Desvio Padrão	Média
A100 N100	838,51	248,10	115,01	352,73
A10 N50	848,07	250,60	114,63	360,65

A Figura 13 apresenta a distribuição das previsões realizadas utilizando-se o modelo que apresentou melhor ajuste ao fim das iterações de janela. A linha vermelha representa a reta de tendência que melhor se ajusta aos pontos, obtida por meio de regressão dos pontos.

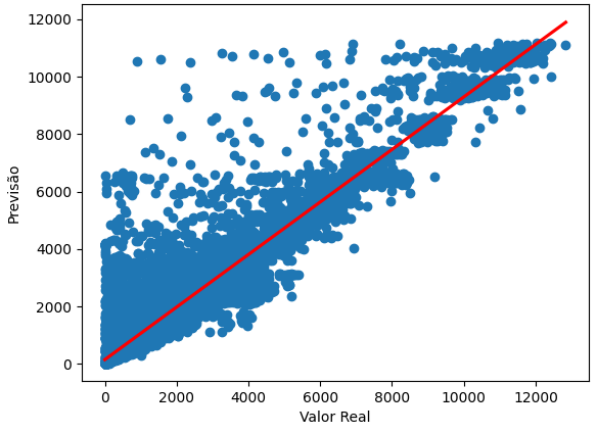


Figura 13: Distribuição das previsões realizadas pelo modelo “A100 N100” na última iteração da janela, no conjunto do período completo.

Já no Figura 14, apresenta-se a distribuição de todas as previsões realizadas por iteração de janela, com eixo x “Valor Real” e eixo y “Previsão”, onde a linha vermelha representa a reta que melhor se ajusta aos pontos, enquanto a linha verde representa a reta x=y, correspondente a um cenário com 100% de acerto nas previsões. Em instâncias onde a reta vermelha se encontra abaixo da verde, significa que as previsões geralmente estão sendo abaixo dos valores reais, enquanto se a reta vermelha estiver acima da reta verde, significa que as previsões estão geralmente acima dos valores reais. No Anexo A, a Figura 14 se encontra em maior resolução.

A linha com as datas no topo da figura se referem a porção do conjunto de dados que está sendo testada em cada iteração.

Nota-se, a princípio, uma certa regularidade na distribuição de previsões no período de normalidade, isto é, até a 16ª iteração, ou o gráfico referente ao período 20/01/2020 a 16/04/2020 na Figura 14. Esta regularidade é expressa pelas previsões localizadas arranjadas bem próximo da linha de perfeição (verde). Tal queda também pode ser vista na Figura 12.

Quando se entra no período de pandemia, há uma queda de 51% no R^2 e aumento de 64% do Erro Médio Absoluto e, com isso, a distribuição das previsões começa a se afastar da linha verde, principalmente na parte inferior do gráfico. Visto que na próxima iteração (16/04/2020 – 16/07/2020), apesar de não ter métricas tão ruins como a 16ª iteração, conta com uma distribuição de pontos que conseguiu se arranjar na base da reta verde, porém acaba se separando bastante nos valores altos de passageiros. Inclusive quando se entra nesse período, o eixo das ordenadas muda a escala, devido aos baixos valores de passageiros na pandemia em comparação ao período passado.

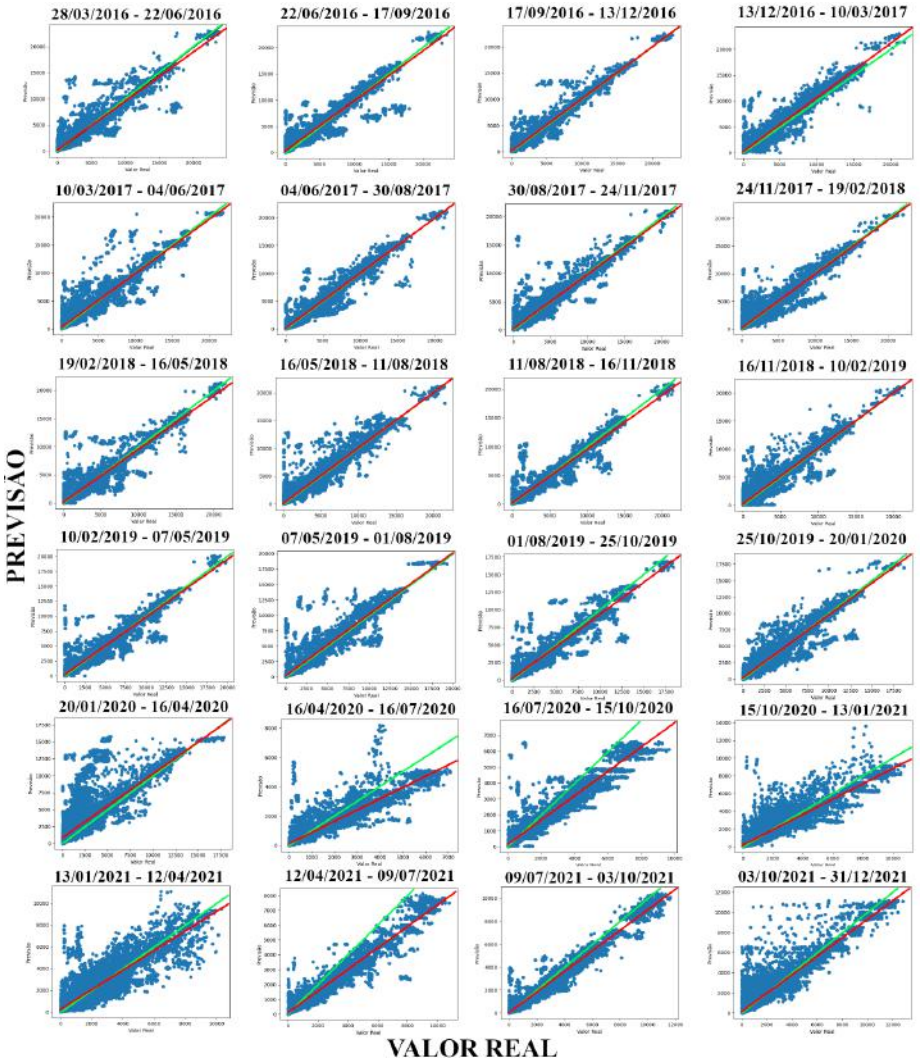


Figura 14: Distribuição de todas as previsões realizadas por iteração de janela no modelo “A100 N100”, no conjunto de dados do período completo. Anexo A apresenta a figura com melhor resolução.

Como valores maiores de passageiros são minoria na base de dados (pode ser verificado pelos valores médios e máximos de passageiros na Tabela 5), entende-se como a 16ª iteração têm as piores métricas, mesmo com a distribuição de previsões não parecendo tão distante da reta verde como outras iterações. Isto pode ser conferido na Figura 15, visto que na 16ª iteração (20/01/2020 – 16/04/2020) a separação na base da distribuição ideal (reta verde), indica um modelo que erra os valores baixos de passageiro, ou seja, erra a maioria dos valores do conjunto, visto que valores baixos compõe a maior parte da base de dados.

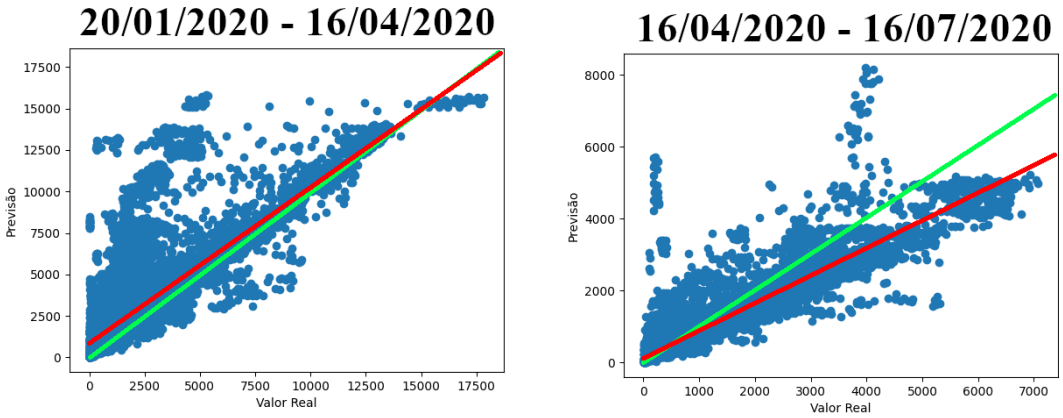


Figura 15: 16ª e 17ª iteração da janela

Percebe-se nessa sequência de passos um pouco do processo de aprendizado, já que depois do descolamento no topo da reta na 17ª iteração (16/04/2020 – 16/07/2020), nos próximos passos as distribuições acabam cada vez mais tendendo para sobreposição das linhas, através da aproximação do topo de ambas retas, isso pode ser verificado na Figura 16, com a 20ª iteração (13/01/2021 – 12/04/2021).

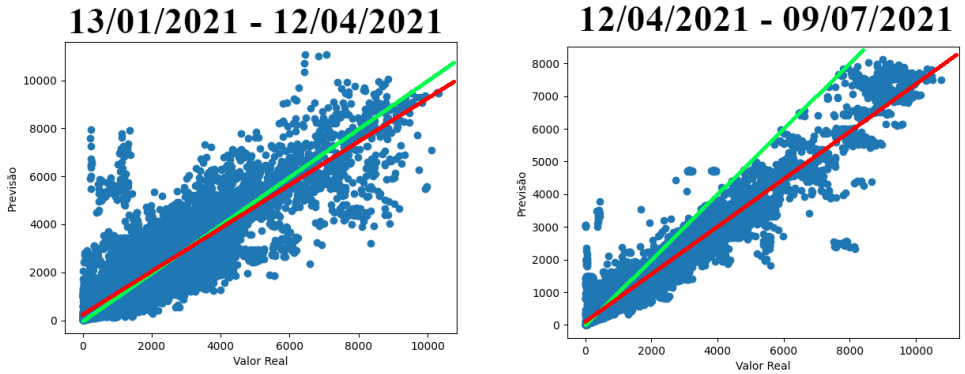


Figura 16: 20ª e 21ª iteração da janela

Novamente, quando se atinge o período entre abril e julho como na 17ª iteração, dessa vez em 2021 na 21ª iteração (12/04/2021 – 09/07/2021), a distribuição das previsões tende a se separar da reta ideal para valores maiores, quase um reflexo do mesmo processo que ocorre após a 16ª iteração. Para nos passos seguintes, as previsões irem se

aproximando de uma distribuição ideal novamente pela aproximação do topo das retas. Essa normalização pode ser verificada no Anexo A.

Em modelos anteriores ao “A100 N100”, principalmente nos modelos com número máximo de nós abaixo de 20, também é perceptível uma queda das métricas na 16ª iteração. Inclusive, nos modelos “A2 N5” e “A2 N10” nesse período da 17ª iteração (16/04/2020 – 16/07/2020), o desempenho do algoritmo atinge valores de R^2 abaixo de zero, justamente os valores discrepantes que foram cortados nas Figura 10.

Por fim, a Figura 17 apresenta a evolução do tempo em minutos decorrido em cada variação de Floresta Aleatória. A contagem de tempo foi realizada uma vez em um computador com processador AMD Ryzen 5 5500U e Memória RAM instalada de 8 GB. Observa-se que o tempo aumenta de forma expressiva nos últimos modelos, que possuem quantidade de árvores igual a 100. Nesse sentido, os modelos com quantidade de árvores menores, desde que o parâmetro de profundidade máxima da árvore sejam de pelo menos 20, não apresentam resultados tão distantes do modelo mais ajustado. O modelo “A10 N50”, por exemplo conforme demonstrado na Tabela 10 e Tabela 11, possui métricas ligeiramente inferiores ao modelo mais ajustado em números absolutos, enquanto demora 10% do tempo deste melhor algoritmo, conforme demonstrado na Figura 17.

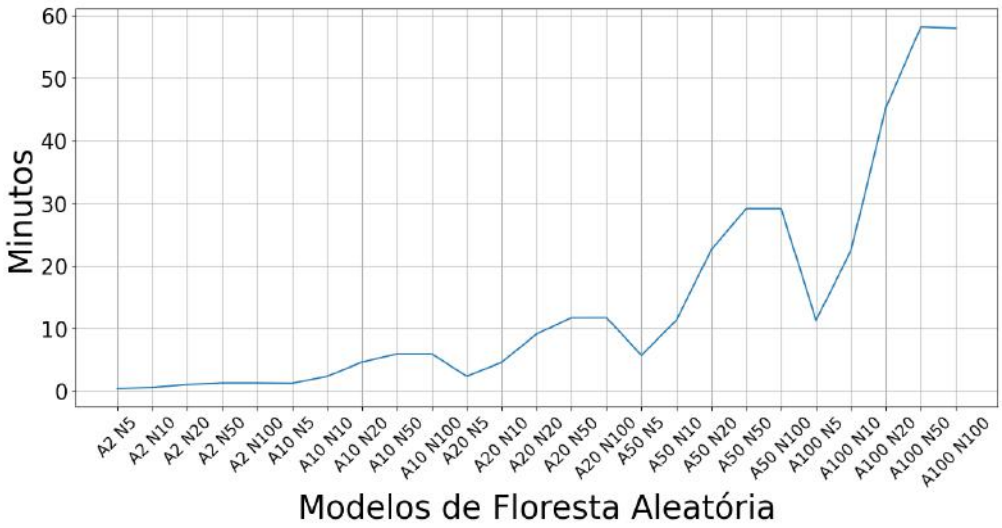


Figura 17: Evolução do tempo decorrido na execução de cada modelo de Floresta Aleatória, para o conjunto do período completo.

4.2 INSTÂNCIA: ANOS SEM PANDEMIA (ASP)

Na segunda subseção de dados, foram realizados experimentos no conjunto que compreende o período entre 1 de janeiro de 2016 até 31 de dezembro de 2019. Na Figura 18, encontra-se o diagrama de caixa do R^2 para as iterações de janela para todos os modelos de Floresta Aleatória apresentadas na seção 3.5. Já a Figura 19 apresenta o

diagrama de caixa do Erro Médio Absoluto para as mesmas iterações, de todos os modelos.

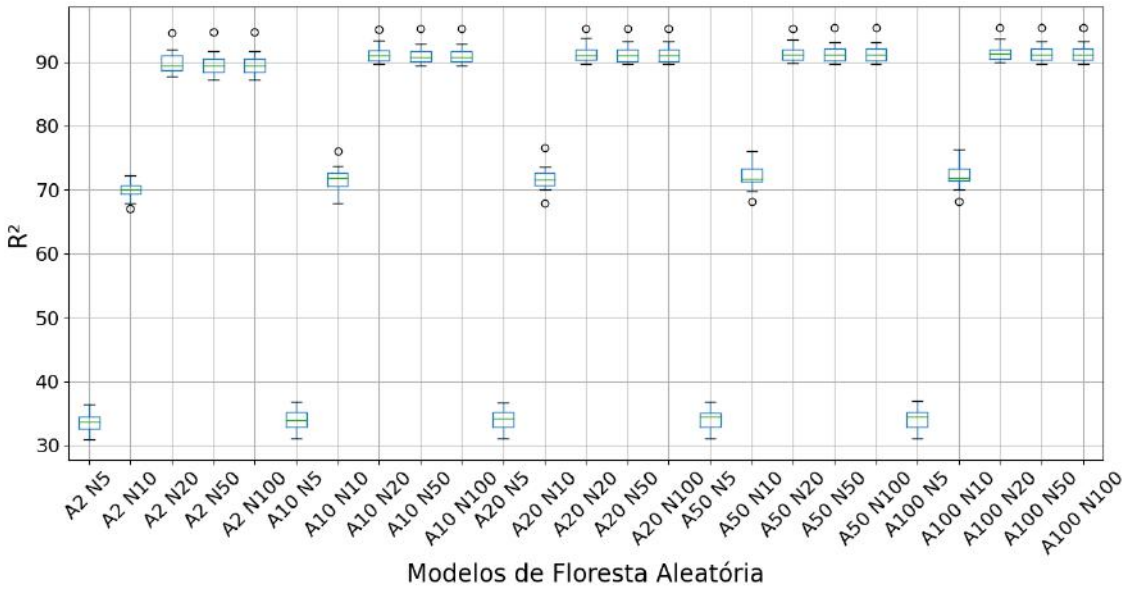


Figura 18: Diagrama caixa de R² para todos modelos de Floresta Aleatória, no conjunto de dados dos anos sem pandemia.

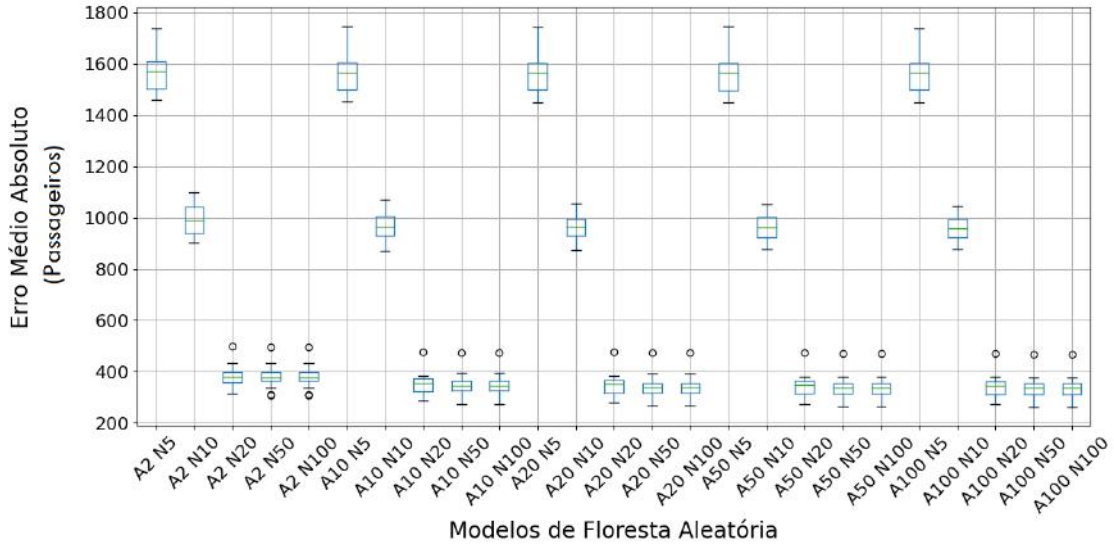


Figura 19: Diagrama caixa da Erro Médio Absoluto para todos modelos de Floresta Aleatória, no conjunto de dados dos anos sem pandemia.

Observa-se os mesmos padrões de comportamento com as alterações de hiperparâmetros da Floresta Aleatória verificados para a instância de período completo, como apresentado na seção anterior. Porém, agora os resultados possuem menor dispersão, o que pode se ver pelo tamanho do intervalo interquartil e menos valores discrepantes. O modelo “A10 N20” na seção anterior tinha limites superiores de 86,6 e 333 e inferiores 82,6 e 255,6 para R² e Erro Médio Absoluto respectivamente, enquanto nessa seção o mesmo modelo tem limites superiores de 88.5 e 364,9 e inferiores de 87,2 e 334,8. Uma redução da distância entre limites de 67,5% para R² e 61,1% para o Erro Médio Absoluto.

O R^2 teve uma pequena melhora em relação à instância anterior, a mediana do modelo “A100 N100” foi de 90,3 para 91,2, enquanto o Erro Médio Absoluto se manteve com a mesma mediana. Já a média do mesmo modelo, saltou de 87,8 para 91,4 no R^2 e 352 para 335 no Erro Médio Absoluto. Tais comparativos encontram-se na Tabela 12 e Tabela 13, referentes ao desempenho do modelo “A100 N100” em números absolutos comparado com os resultados da seção anterior, junto do “A10 N50” para comparação de um modelo com desempenho menos ajustado e tempo de execução menor, como será explicitado a diante. Além da Figura 20, com a evolução das métricas em cada iteração da janela deslizante.

Tabela 12: Resultados de R^2, no conjunto de dados dos anos sem pandemia em comparação ao conjunto do período completo.

Modelo	Valor máximo	Valor mínimo	Desvio Padrão	Média
A100 N100 ASP	95,45	89,66	1,52	91,45
A10 N50 ASP	95,25	89,43	1,47	91,17
A100 N100 PC	95,15	45,22	9,94	87,85
A10 N50 PC	94,90	44,83	10,13	87,51

Tabela 13: Resultados de Erro Médio Absoluto, no conjunto de dados dos anos sem pandemia em comparação ao conjunto do período completo.

Modelo	Valor máximo	Valor mínimo	Desvio Padrão	Média
A100 N100 ASP	467,93	261,13	48,03	335,71
A10 N50 ASP	472,72	272,25	47,21	346,59
A100 N100 PC	838,51	248,10	115,01	352,73
A10 N50 PC	848,07	250,60	114,63	360,65

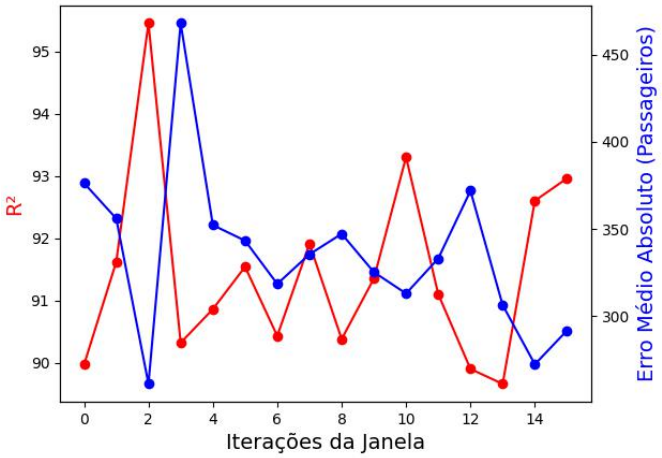


Figura 20: Valor de R^2 e Erro Médio Absoluto por iteração de janela do modelo “A100 N100”, no conjunto de dados do período de normalidade.

Através da Tabela 12 e Tabela 13, é possível verificar uma variação menor em todas as métricas, ou seja, menor desvio padrão (cerca de 86% menor para o modelo “A10 N50” no R^2), além de uma melhora frente aos valores da instância anterior (redução de cerca de 5% na média de Erro Médio Absoluto para o modelo “A100N100”). Apenas em relação ao valor mínimo do Erro Médio Absoluto, que a instância anterior conseguiu um valor 5% inferior ao conjunto de dados analisado aqui.

Na Figura 21, apresenta-se a distribuição de previsões realizadas na última iteração do modelo “A100 N100” e, na Figura 22, uma demonstração da esperada redução de tempo para cada modelo de Floresta Aleatória em relação a instância anterior. Visto que aqui se trabalha com uma instância com 67% do tamanho do conjunto inteiro (período completo), verificou-se que o tempo do melhor modelo diminuiu em cerca de 60%, ou decorreu em 40% do tempo do modelo que trabalha com o conjunto inteiro. Comparando os modelos dessa subseção, pode-se ver novamente que o modelo “A10 N50”, apesar das métricas menos ajustadas ao “A100 N100”, continua desempenhando em 10 % do tempo do melhor ajustado em números absolutos.

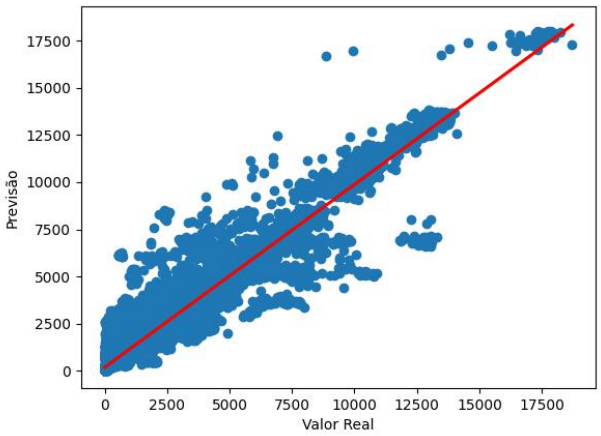


Figura 21: Distribuição das previsões realizadas pelo modelo “A100 N100” na última iteração da janela deslizante, no conjunto de dados do período de normalidade.

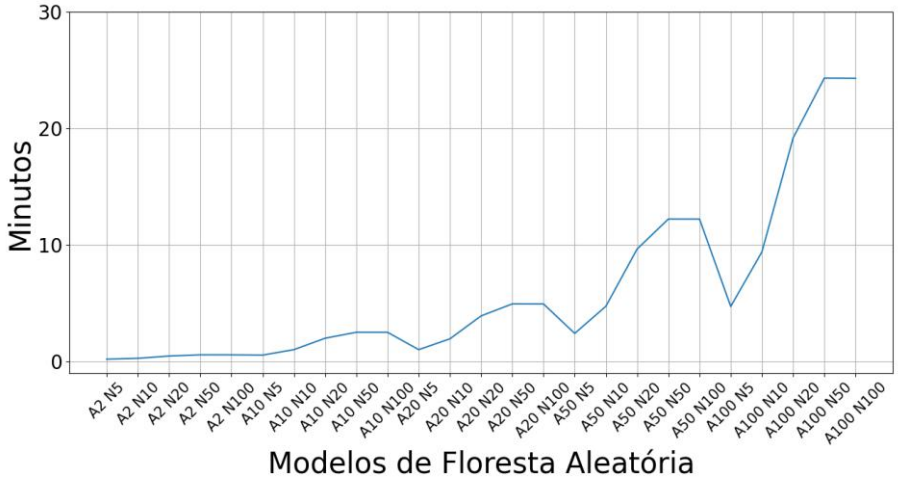


Figura 22: Evolução do tempo decorrido na execução de cada modelo de Floresta Aleatória, no conjunto de dados do período de normalidade.

4.3 INSTÂNCIA: ANOS COM PANDEMIA (ACP)

Por fim, na última instância de dados, realizou-se experimentos nos anos com pandemia, que compreendem o período entre 1 de janeiro de 2020 até 31 de dezembro de 2021. A Figura 23 e Figura 24 apresentam, respectivamente, o diagrama de caixa de todas iterações da janela crescente com validação adiante do R^2 e Erro Médio Absoluto nas diferentes configurações de Floresta Aleatória.

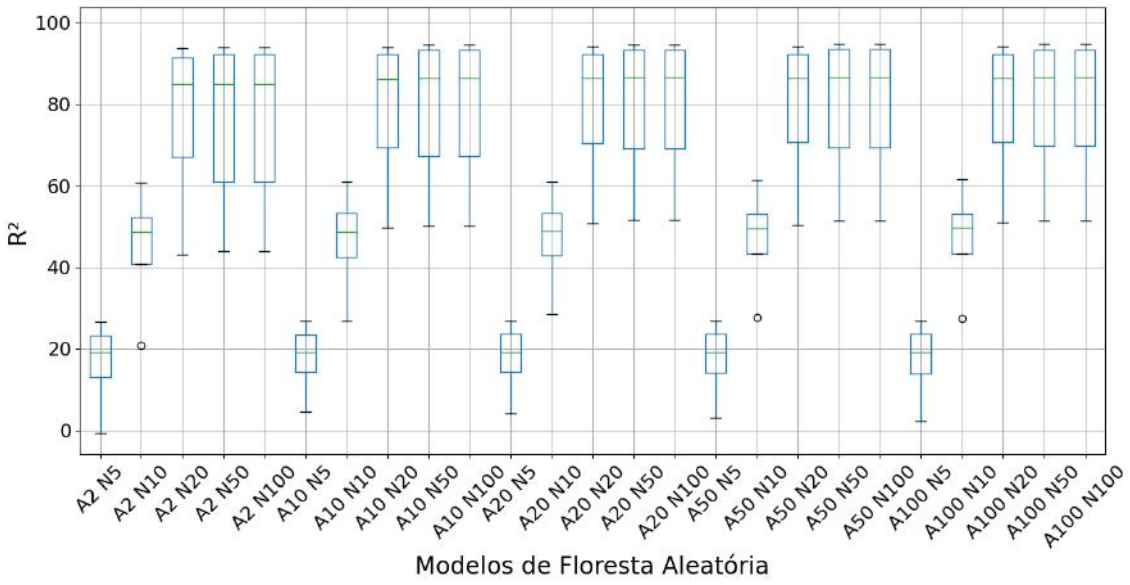


Figura 23: Diagrama caixa de R^2 para todas os modelos de Floresta Aleatória, no conjunto de dados dos anos com pandemia.

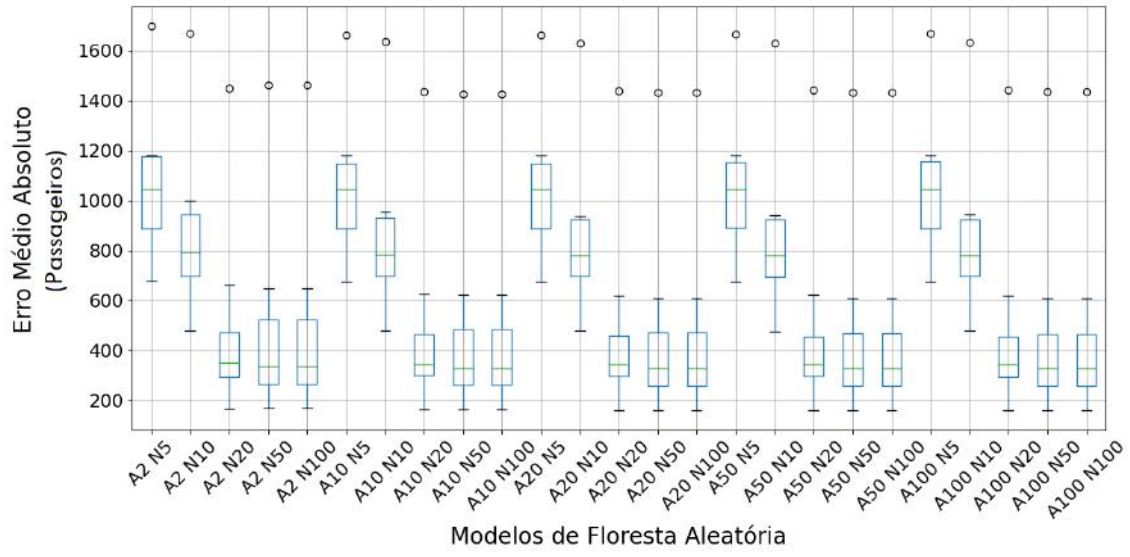


Figura 24: Diagrama caixa do Erro Médio Absoluto para todos os modelos de Floresta Aleatória, no conjunto de dados dos anos com pandemia.

Novamente, os padrões de comportamento do desempenho com as alterações de hiperparâmetros de Floresta Aleatória se mostram similares, obtendo modelos mais bem ajustados, depois que o hiperparâmetro profundidade máxima da árvore passa de 10, como por exemplo no modelo “A2 N20” que possui mediana do R^2 igual 84,8 frente a

48,5 do modelo “A2 N10”, o mesmo padrão se repete para o Erro Médio Absoluto. Porém, dessa vez as métricas de R^2 apresentam uma maior dispersão comparado ao conjunto dos anos sem pandemia, que pode ser observado pelo tamanho do intervalo interquartil, e também comparado ao período completo, principalmente nos modelos com profundidade máxima da árvore acima de 10. O modelo “A20 N100” da instância dos anos sem pandemia, possui um intervalo entre limites superiores e inferiores de R^2 equivalente a 2,3% do tamanho desse mesmo intervalo na instância dos anos com pandemia, ou seja, a instância dos anos de pandemia apresenta uma maior dispersão frente a instância previamente analisada.

Comportamento contrário ao que tinha sido estabelecido até aqui, onde modelos com profundidade máxima da árvore acima de 10 acabavam tendo um desempenho cada vez mais com menor dispersão, além do melhor ajuste nas métricas. Dessa vez as métricas continuaram sendo melhores ajustadas com aumento da profundida máxima da árvore, como demonstrado no parágrafo anterior, porém os valores se distribuíram em um intervalo maior e mais discrepante.

A Figura 25 mostra a evolução das métricas do modelo “A100 N100”, conforme a passagem de iterações da janela crescente com validação adiante. No eixo de R^2, precisou-se cortar o valor da iteração 0, pois esse apresentava um valor negativo, e com isso dificultando a visualização do resto das iterações.

Esse valor negativo no R^2, além da explosão do Erro Médio Absoluto na iteração 0 ocorreu em todos modelos de Floresta Aleatória dessa instância, e acontece primariamente pelo que foi exposto na Tabela 8, da seção 3.4. Em que a iteração de número 0, utiliza dos meses de janeiro até março de 2020, ou seja, meses iniciais da pandemia, como mostrado no fim da seção 2.2.1, para testar os valores dos meses de abril até junho, meses de pandemia agravada.

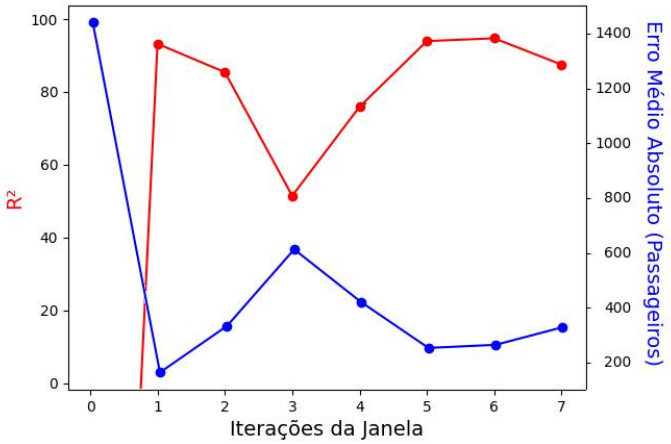


Figura 25: Valor de R² e Erro Médio Absoluto por iteração de janela no modelo “A100 N100”, no conjunto de dados dos anos com pandemia.

A seguir, na Figura 26 encontra-se a distribuição das previsões realizadas pelo modelo “A100 N100”. As Tabelas 14 e 15 dispõe o comparativo final entre o modelo “A100 N100” em números absolutos e o modelo “A10 N50”, das três divisões de dados realizadas.

Os valores de tempo decorridos no conjunto de valores no período dos anos com pandemia encontram-se na Figura 27, onde novamente o modelo “A10 N50” teve cerca 10% do tempo do modelo “A100 N100”. Que por sua vez tem 8% do tempo decorrido no mesmo modelo com o conjunto de dados do período completo, ou uma redução de 92% em tempo, mesmo trabalhando com 33% dos dados deste.

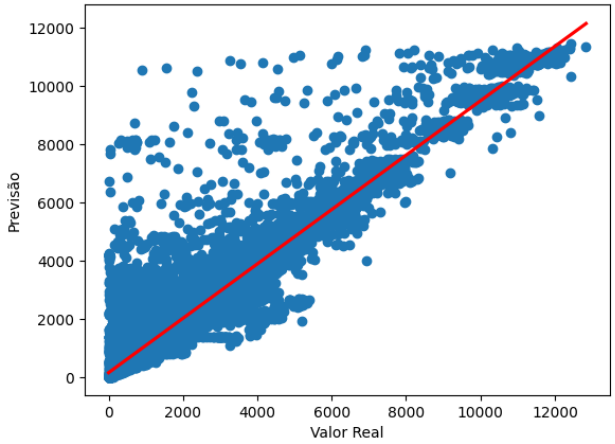


Figura 26: Distribuição das previsões realizadas pelo modelo “A100 N100” na última iteração da janela, no conjunto de dados dos anos com pandemia.

Tabela 14: Resultados de R², comparação entre as três subseções.

Modelo	Valor máximo	Valor mínimo	Desvio Padrão	Média
A100 N100 ASP	95,45	89,66	1,52	91,45
A10 N50 ASP	95,25	89,43	1,47	91,17
A100 N100 PC	95,15	45,22	9,94	87,85

A10 N50 PC	94,90	44,83	10,13	87,51
A100 N100 ACP	94,72	-288,07	132,02	36,75
A10 N50 ACP	94,57	-289,87	132,49	35,94

Tabela 15: Resultados de Erro Médio Absoluto, comparação entre as três subseções.

Modelo	Valor máximo	Valor mínimo	Desvio Padrão	Média
A100 N100 ASP	467,93	261,13	48,03	335,71
A10 N50 ASP	472,72	272,25	47,21	346,59
A100 N100 PC	838,51	248,10	115,01	352,73
A10 N50 PC	848,07	250,60	114,63	360,65
A100 N100 ACP	1436,09	160,12	411,34	474,38
A10 N50 ACP	1427,72	163,04	408,26	477,86

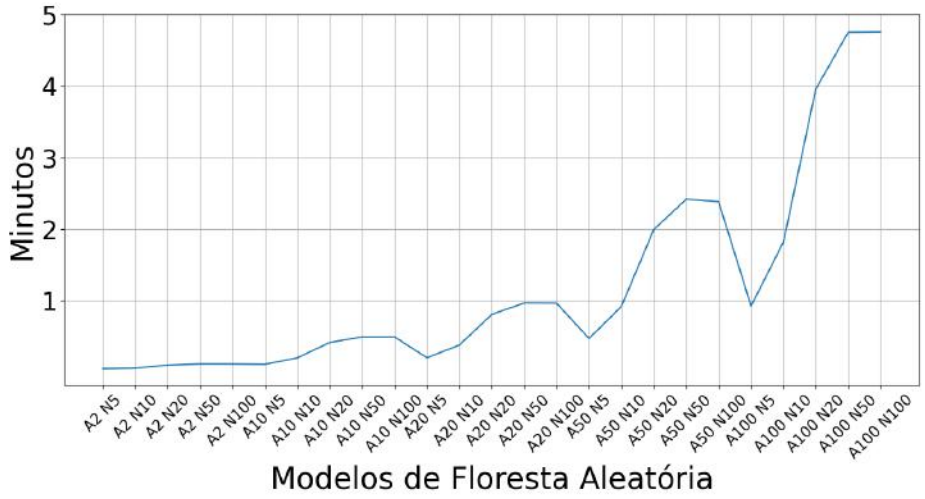


Figura 27: Evolução do tempo decorrido na execução de cada modelo de Floresta Aleatória, no conjunto de dados dos anos com pandemia.

5 CONCLUSÕES

Esse trabalho tem como objetivo avaliar modelos de aprendizado de máquina, utilizando do algoritmo de Florestas Aleatórias, para realizar a previsão da demanda de passageiros no transporte público. Para isso, foi utilizada a base de dados de transporte público urbano da cidade de Belo Horizonte entre 2016 e 2021.

Para atingir o objetivo, utilizou-se do arcabouço teórico de Transporte coletivo, focado em conceitos de demanda, previsão de demanda, oferta e alocação de recursos, para contextualizar o trabalho e seus objetivos. Além disso, foi utilizada a teoria por detrás do Aprendizado de Máquina, para colocar em prática e realizar tais previsões de demanda, utilizando os algoritmos de Árvores de Decisão e Florestas Aleatória, além do conceito de janela crescente com validação adiante, usado para divisão dos dados de série temporal.

Ao fim dos experimentos, esse estudo chegou à conclusão de que o mais ajustado de Floresta Aleatória para as previsões de demanda realizadas é o que possui a maior profundidade (dentre os valores estudados nesse experimento) nas Árvores de Decisão e não necessariamente aquele com a maior quantidade de árvores. Pode-se obter uma melhora de 1% no R^2 e 7,5% no Erro Médio Absoluto com o aumento da quantidade de árvores, como demonstrado na seção 4.1 no caso do modelo “A2 N20” e “A100 N20”, porém tem-se como consequência um tempo de execução do modelo 44 vezes maior. Tal comportamento é observado em inúmeras outras comparações como mostrado na seção anterior.

Além disso, é importante ressaltar que mesmo fora dos modelos com melhores métricas absolutas, os resultados como um todo se mostraram promissores, visto que modelos como o “A10 N50” da Tabela 10 e Tabela 11 alcançaram valores de desempenho similares aos modelos com maiores hiperparâmetros como o “A100 N100”, enquanto é executado em 10% do tempo deste. Demonstrando que a abordagem de janela crescente com validação adiante de Schnaubelt (2019) também obteve resultados promissores para essa base de dados, que não foi tratada visando obter uma série temporal estacionária. E mesmo sob influência de dados mais antigos, os modelos conseguiram obter uma habilidade de predição considerável, obtendo R^2 de 95,15 e Erro Médio Absoluto de 248 passageiros no modelo “A100 N100”.

Os experimentos realizados a partir do período de pandemia se mostraram tentativas de montar uma série temporal com diferenças menores ao longo do tempo,

porém o que se viu é que sem os dados da pandemia a predição começou a atuar numa faixa menor de valores, como demonstrado na seção 4.2, sem necessariamente alcançar valores maiores. Sem contar que os experimentos conduzidos no conjunto de dados inteiro mostraram-se oportunidades de observar iteração por iteração da janela crescente com validação adiante, para ter uma percepção melhor dos ajustes realizados no decorrer do aprendizado para com a base de dados utilizada, e o que exatamente significavam valores maiores ou menores de R^2 ou MAE no sentido macro do modelo, ou seja, uma visão global da distribuição de suas predições em comparações aos valores reais.

Já no conjunto de pandemia, mostrou-se como prejudicial para as métricas de desempenho, o treinamento dos três primeiros meses de 2020 para testar com os três meses seguintes que foram altamente afetados pela pandemia, resultando em valores de R^2 negativos e Erro Médio Absoluto indo para 1436 passageiros, como ilustrado na Figura 25. Apenas com a evolução das iterações que o programa mostrou sinais de melhor ajuste e estabilidade do desempenho, e na reaprendizagem de um novo normal, retornando para valores de R^2 iguais a 87,5 e Erro Médio Absoluto de 325.

Por fim, o Aprendizado de Máquina e principalmente as Florestas Aleatórias, mostraram-se como uma alternativa interessante para a modelagem de sistemas capazes de realizar previsões de curto prazo, em bases de dado de transporte na cidade de Belo Horizonte. Isto visa equilibrar a oferta e demanda dos sistemas de transporte, buscando a sua otimização, diminuição de incertezas e seu subsequente desenvolvimento.

Uma abordagem sugerida para trabalhos futuros seria realizar o mesmo experimento em bases de dados de outras cidades, com diferentes escopos do sistema de transporte coletivo. Visto que uma das maiores dificuldades na condução de um trabalho como esse, é a obtenção de um conjunto sólido de dados. Outra sugestão seria encontrar uma forma melhor de lidar com as variações de demanda causadas pelo período de pandemia, talvez abordando o problema de forma diferente, através de análises mais profundas de estacionariedade da série temporal.

REFERÊNCIAS BIBLIOGRÁFICAS

ANTP. **Sistema de Informação da Mobilidade Urbana: Relatório Geral 2018**. São Paulo, 2018. Disponível em <http://files.antp.org.br/simob/sistema-de-informacoes-da-mobilidade--simob--2018.pdf>. Acessos em 30 mar. 2022.

BEZERRA, L. M. **Comparação de Métodos de Aprendizado de Máquina Para Previsão de Demanda no Transporte Público Urbano**. Trabalho de Conclusão de Curso (Bacharelado) — Universidade Federal de Santa Catarina, 2021.

CRUZ, J. A. **Modelo de Demanda Variável para a Determinação da Oferta de Transporte Coletivo Urbano por Ônibus**. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 1991.

CRUZ, R. M. et al . COVID-19: emergência e impactos na saúde e no trabalho. **Rev. Psicol., Organ. Trab.**, Brasília, v.20, n.2, p.I-III, jun.2020. Disponível em <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1984-66572020000200001&lng=pt&nrm=iso>. Acessos em 30 mar. 2022. http://dx.doi.org/10.17652/rpot/2020.2.editorial.

FERRONATTO, L. G. **Potencial de Medidas de Gerenciamento da Demanda no Transporte Público Urbano por Ônibus**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, 2002.

GÉRON, A. **Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow**. [S.l.]: Alta Books, 2019. ISBN 978-8550803814.

INSTITUTO NACIONAL DE METEOROLOGIA (comp.). **Banco de Dados Meteorológicos do INMET**. Disponível em: https://bdmep.inmet.gov.br/. Acessos em: 12 fev. 2022.

MITCHELL, T. M. **Machine learning**. [S.l.]: McGraw-Hill, 1997. ISBN 978-0070428072.

MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos Sobre Aprendizado de Máquina. Sistemas Inteligentes Fundamentos e Aplicações**. 1 ed. Barueri-SP: Manole Ltda, 2003. ISBN 85-204-168.

MURÇA, M. e MÜLLER, C. Transporte coletivo urbano: uma análise de demanda para a cidade de Salvador. **Journal of Transport Literature**, vol. 8, n. 1, pp. 265-284. Salvador, 2014.

NTU. **NTUrbano,** Brasília, v.41, set./out.2019. Disponível em: <www.ntu.org.br/novo/upload/Publicacao/Pub637110488381579841.pdf>. Acessos em 30 mar. 2022.

ORTUZAR, J. D.; WILLUMSEN, L. G. **Modelling Transport**. [S.l]: John Wiley & Sons, 2011. ISBN 9780470760390.

PIANUCCI, M. N.; PITOMBO, C. S.; CUNHA, A. L.; LIMA SEGANTINE, P. C. Previsão da demanda por viagens domiciliares através de método sequencial baseado em população sintética e redes neurais artificiais. **TRANSPORTES**, [S. l.], v. 27, n. 4, p. 1–23, 2019. Disponível em: https://revistatransportes.org.br/anpet/article/view/1409. Acesso em: 6 jul. 2022.

PREFEITURA DE BELO HORIZONTE (Belo Horizonte). Portal de Dados Abertos (comp.). **Mapa de Controle Operacional Consolidado**. Disponível em: https://dados.pbh.gov.br/dataset/mapa-de-controle-operacional-mco-consolidado. Acessos em: 11 fev. 2022.

SCHNAUBELT, M. A comparison of machine learning model validation schemes for non-stationary time series data. **FAU Discussion Papers in Economics**, Nürnberg, 2019.

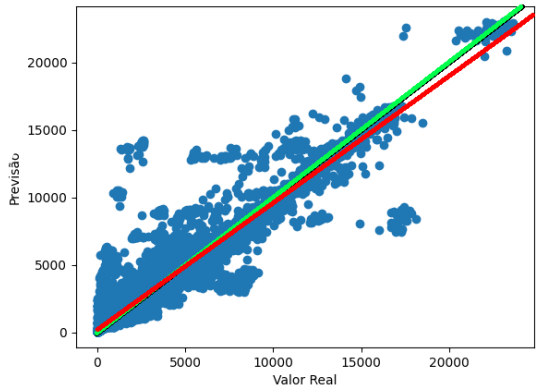
SILVA, L. C. **Aprendizado de Máquina com Treinamento Continuado Aplicado à Previsão de Demanda de Curto Prazo: O Caso do Restaurante Universitário da Universidade Federal de Uberlândia**. Dissertação (Mestrado) — Universidade Federal de Uberlândia, 2019.

TIBURCIO, F. S. **Uma Aplicação de Redes Neurais Artificias Para previsão da Demanda de Passageiros no Transporte Público da Cidade de Joinville**. Trabalho de Conclusão de Curso (Bacharelado) — Universidade Federal de Santa Catarina, 2018.

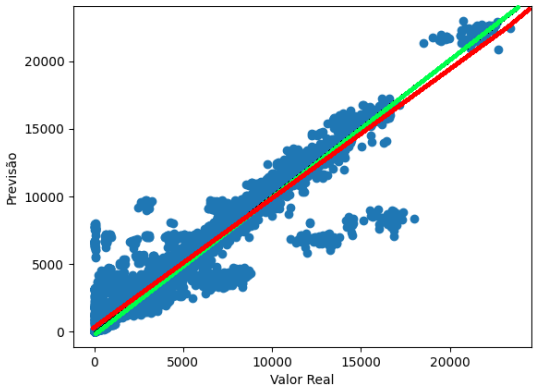
VASCONCELOS, V. S.; QUEVEDO-SILVA, F.; ROVAI, R. L. Modelo de previsão de demanda baseado em redes neurais artificiais para projetos de transporte de passageiros. **REVISTA BRASILEIRA DE GESTÃO URBANA** [online]. 2021, v. 13. Disponível em: <https://doi.org/10.1590/2175-3369.013.e20200160>. Acesso em: 6 jul. 2022.

ANEXO A

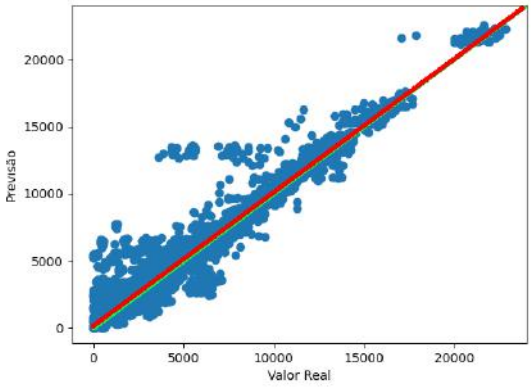
28/03/2016 - 22/06/2016



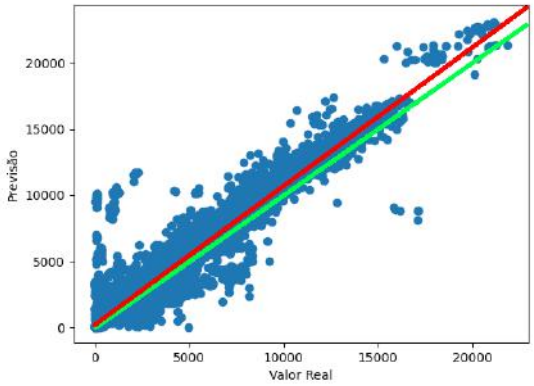
22/06/2016 - 17/09/2016



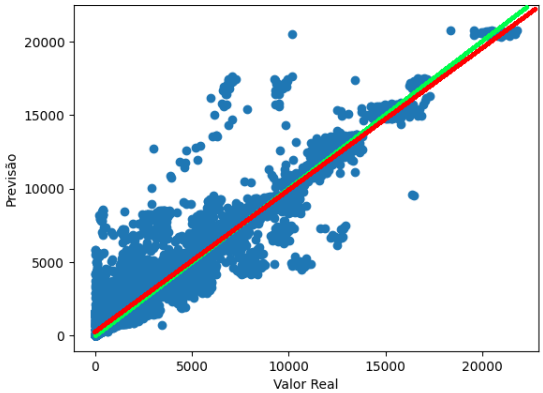
17/09/2016 - 13/12/2016



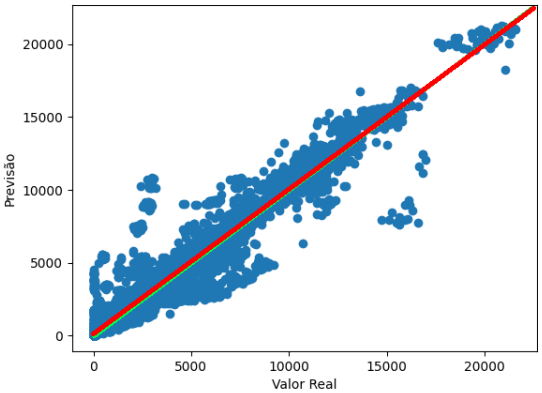
13/12/2016 - 10/03/2017



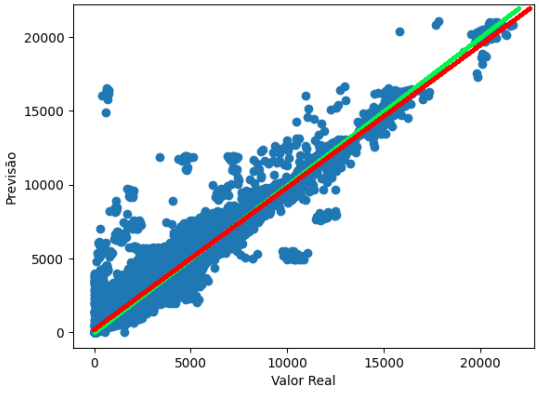
10/03/2017 - 04/06/2017



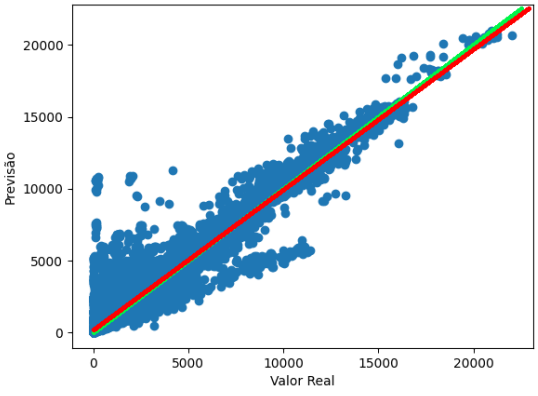
04/06/2017 - 30/08/2017



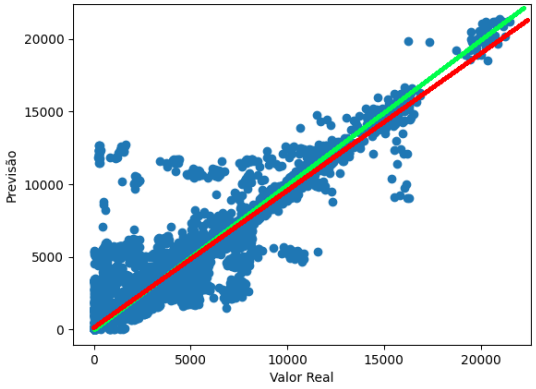
30/08/2017 - 24/11/2017



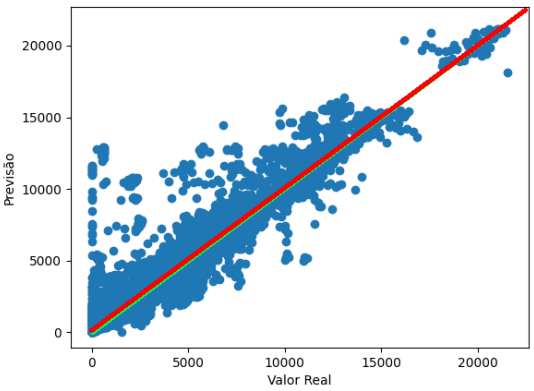
24/11/2017 - 19/02/2018



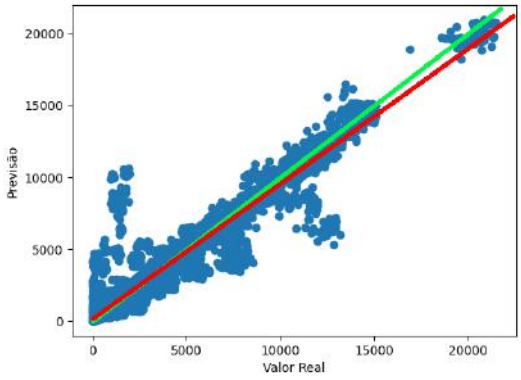
19/02/2018 - 16/05/2018



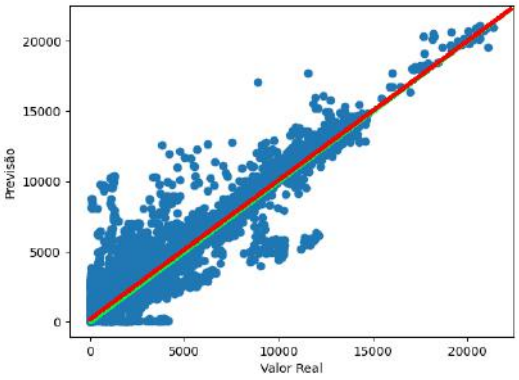
16/05/2018 - 11/08/2018



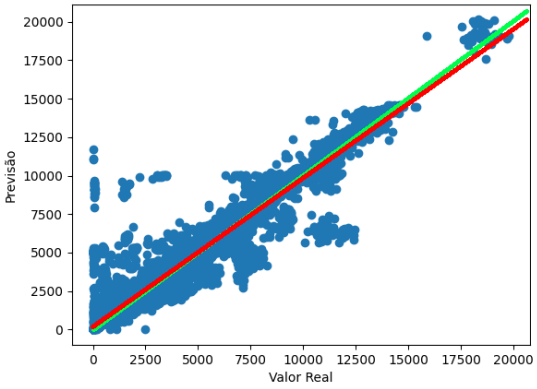
11/08/2018 - 16/11/2018



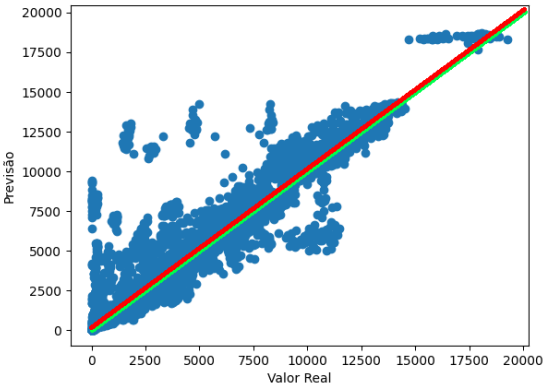
16/11/2018 - 10/02/2019



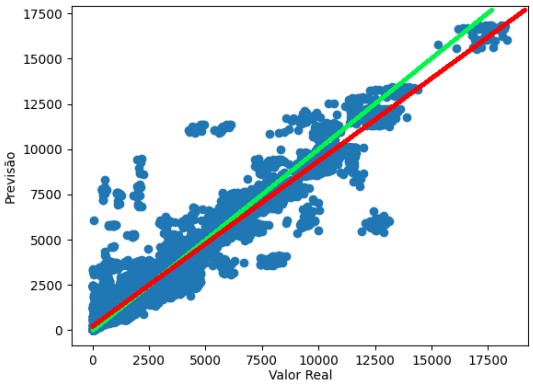
10/02/2019 - 07/05/2019



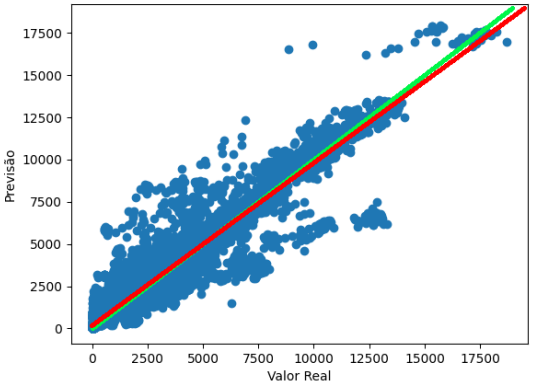
07/05/2019 - 01/08/2019



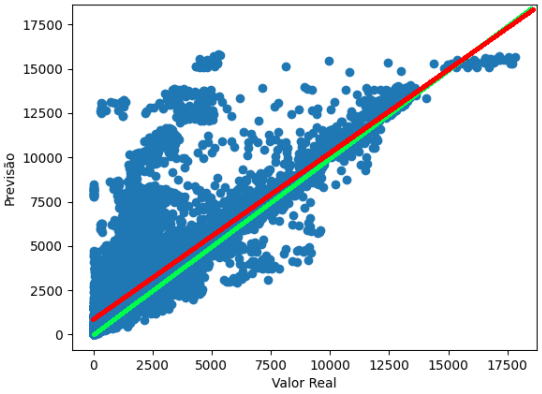
01/08/2019 - 25/10/2019



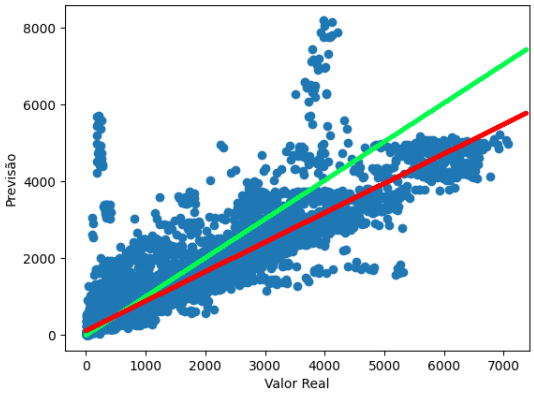
25/10/2019 - 20/01/2020



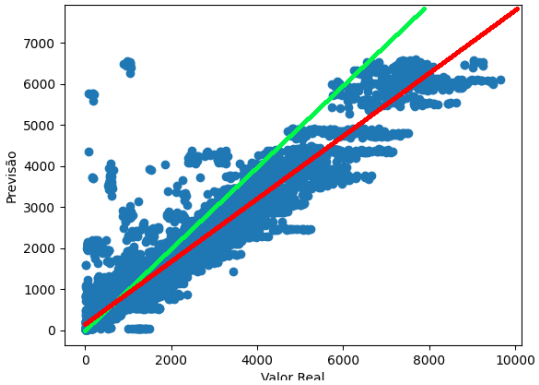
20/01/2020 - 16/04/2020



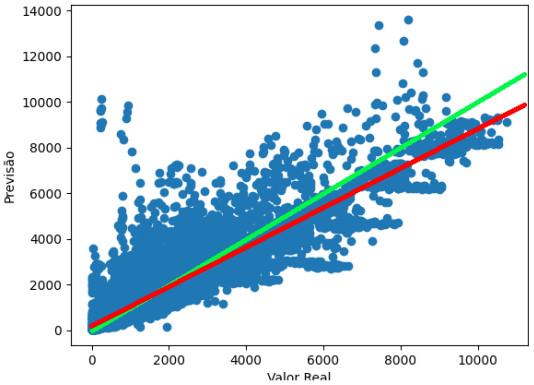
16/04/2020 - 16/07/2020



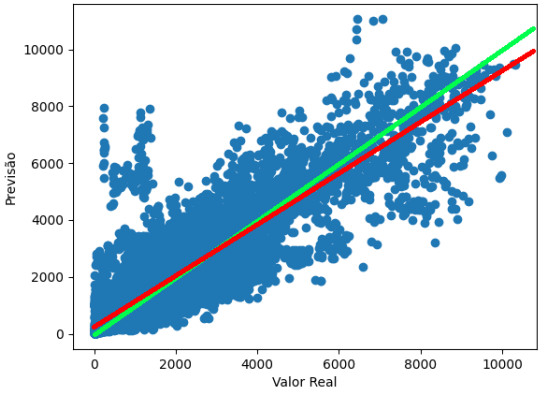
16/07/2020 - 15/10/2020



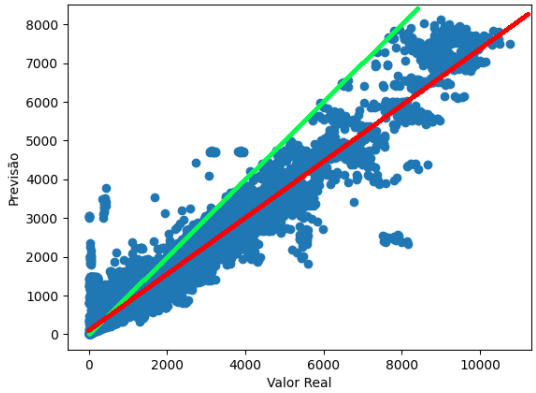
15/10/2020 - 13/01/2021



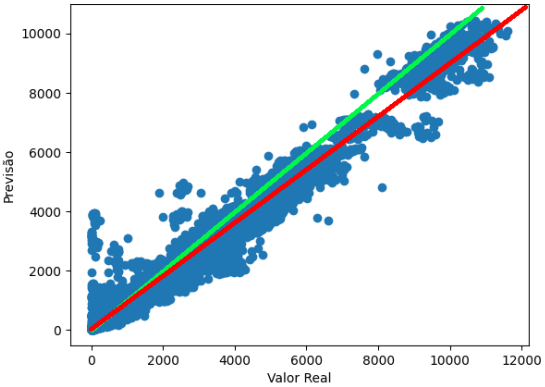
13/01/2021 - 12/04/2021



12/04/2021 - 09/07/2021



09/07/2021 - 03/10/2021



03/10/2021 - 31/12/2021