INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Ícaro de Almeida Varão

Análise preditiva de atraso nos principais aeroportos brasileiros

Trabalho de Graduação 2021

Curso de Engenharia Civil Aeronáutica

CDU 656.71

Icaro de Almeida Varão

Análise preditiva de atraso nos principais aeroportos brasileiros

Orientador

Prof. Dr Marcelo Xavier Guterres (ITA)

Relator

Prof. Dr Alessandro Vinícius Marques de Oliveira (ITA)

ENGENHARIA Civil Aeronáutica

SÃO JOSÉ DOS CAMPOS

INSTITUTO TECNOLÓGICO DE AERONÁUTICA

2021

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Varão, Ícaro de Almeida
Análise preditiva de atraso nos principais aeroportos brasileiros / Ícaro de Almeida Varão São José dos Campos, 2021. 51f.

Trabalho de Graduação – Cursos de Engenharia Civil Aeronáutica– Instituto Tecnológico de Aeronáutica, 2021. Orientador: Prof. Dr. Marcelo Xavier Guterres. Relator: Prof. Dr. Alessandro Vinícius Marques de Oliveira.

1. . 2. . 3. . I. Ícaro de Almeida Varão II. Instituto Tecnológico de Aeronáutica. III.Título.

REFERENCIA BIBLIOGRAFICA

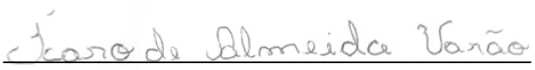
Varão, Ícaro de Almeida . **Análise preditiva de atraso nos principais aeroportos brasileiros**. 2021. 51f.. Trabalho de Conclusão de Curso. (Graduação em Engenharia Civil Aeronáutica) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DOS AUTORES: Ícaro de Almeida Varão

TÍTULO DO TRABALHO: Análise preditiva de atraso nos principais aeroportos brasileiros

TIPO DO TRABALHO/ANO: Graduação / 2021



Ícaro de Almeida Varão
Rua H8B-213
12.228-461, São José Dos Campos - SP

Análise preditiva de atrasos nos principais aeroportos brasileiros

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

Ícaro de Almeida Varão

Autor

Prof. Dr. Marcelo Xavier Guterres

Orientador

Prof. Dr. Alessandro Vinícius Marques de Oliveira

Relator

Prof. Dr. João Claudio Bassan de Moraes

Coordenadora do Curso de Engenharia Civil Aeronáutica

São José dos Campos, 1 de Julho de 2021

Dedico a todos que fizeram parte dessa odisséia. Àqueles que mesmo sabendo que minhas asas eram de cera, nunca duvidaram que eu poderia voar

Agradecimentos

Ao Prof. Dr. Marcelo Xavier Guterres por possuir um grande coração e tentar me ajudar de todas as formas possíveis a concluir essa pesquisa. Serei eternamente grato por todos os ensinamentos passados.

Ao prof. Dr. Alessandro Vinícius Marques de Oliveira por ser uma pessoa extremamente solícita e transmitir confiança e tranquilidade em todas as situações, por mais difíceis que elas pareçam.

À profa. Dr. Juliana Meirelles de Lima por me ajudar a montar o grande quebra-cabeça chamado Ícaro, que as vezes parece não ter solução, mas você me ajuda a achar algumas peças perdidas nos vazios dos meus passos.

"E custou caro, meu caro?

Custou. Tentando ir depressa, acabou chegando depressão.

Mas nem por isso caiu. Tinha asas e por isso devia voar."

— Autor Desconhecido

Resumo

A pesquisa tem como objetivo a análise preditiva de atrasos nos 29 principais aeroportos brasileiros, utilizando uma árvore de decisão. No cenário de crise gerada pela pandemia de Sars-Cov-19, foi avaliado o impacto no setor aeroportuário e a influência na pontualidade dos voos. Desse modo, era esperado que as medidas sanitárias, dentro dos aeroportos em análise, impactassem negativamente na pontualidade desses voos.

Foram utilizadas bases de dados com registros dos horários da movimentação das aeronaves durante o percurso de pouso e decolagem, foram utilizadas bases com registros horários das condições climáticas em cada um dos aeroportos da pesquisa e foram utilizadas bases que possuíam as características de demanda de todos os voos que ocorreram nesses aeroportos. Com essas, foram definidas variáveis classificatórias para a aplicação como *input* no algoritmo da árvore de decisão.

Para a realização da análise por meio da árvore de decisão, foi utilizada a biblioteca *scikit* do *python* a fim de agilizar a pesquisa e possuir boas ferramentas de avaliação da árvore final. Para evitar *overfitting* ao longo do treinamento da árvore de decisão, foram feitas análises para o processo de poda dessa, utilizando o algoritmo de poda por custo de complexidade, em que se usou o parâmetro Alpha para determinar os elos mais frágeis da árvore obtida inicialmente, e qual o comportamento dela quando aplicada para as bases de treino e de teste vaiando o valor do parâmetro Alpha.

Abstract

The research aims at the predictive analysis of delays in the 29 main Brazilian airports, using a decision tree. In the crisis scenario generated by the Sars-Cov-19 pandemic, the impact on the airport sector and the influence on the punctuality of flights were evaluated. Thus, it was expected that sanitary measures within the airports under analysis would negatively impact the punctuality of these flights.

Databases with records of aircraft movement times during the landing and take-off route were used, bases were used with hourly records of weather conditions at each of the airports in the study, and bases that had the demand characteristics of all flights that took place at these airports were used. With these, classification variables were defined for the application as input in the decision tree algorithm.

To perform the analysis through the decision tree, python's scikit library was used in order to streamline the search and have good tools of evaluation of the final tree. To avoid overfitting during the decision tree training, analyzes were performed for its pruning process, using the complexity cost pruning algorithm, in which the Alpha parameter was used to determine the weakest links of the tree initially obtained, and what is its behavior when applied to the training and test bases by using the value of the Alpha parameter.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ANAC	Agência Nacional de Aviação Civil
BBC	British Broadcasting Corporation
RPK	Revenue Passenger Kilometers
TATIC	*Total Air Traffic Information Control*
CTA	Controle de Tráfego Aéreo
GEA	Gestão do espaço aéreo
GFTA	Gerenciamento do Fluxo de Tráfego Aéreo
SIV	Serviço de informação de voo
IGT	Iniciativas de gerenciamento de tráfego

Sumário

1 Introdução

1.1 Motivação

1.1.1 Pandemia do Sars-Cov-19

O período de análise da pesquisa foi marcado pela pandemia do Sars-Covid-19, popularmente chamado de Covid-19. Sendo assim, será feito uma breve análise da progressão desse cenário.

Os primeiros indícios da nova doença são datados do segundo semestre de 2019 na província de Wuhan, China. Nessa fase, os casos eram tratados como um leve surto de pneumonia, embora Li Wenliang, médico oftamologista do hospital cental de Wuhan, tentou alertar sobre a possível nova doença que estaria se alastrando pela província(BBC, 2020). O governo tentou conter essa informação, acusou o médico de "disseminar boatos"e o obrigou a voltar atrás em seu posicionamento.

No primeiro semestre de 2020, a doença começou a tomar escalas globais, se destacando o grande número de baixas em países europeus. Alguns países do continente europeu possuem boa parte da população composta por idosos, esses estavam no grupo de risco para complicações graves da doença, o que fez esses países viverem períodos sombrios e foram obrigados a manter um severo regime de isolamento social, popularmente conhecido como *Lock Down*. A seguir tem-se os dados do início da pandemia no continente europeu referentes ao número de casos por milhões de habitantes e ao número de óbitos por milhões de habitantes, respectivamente.

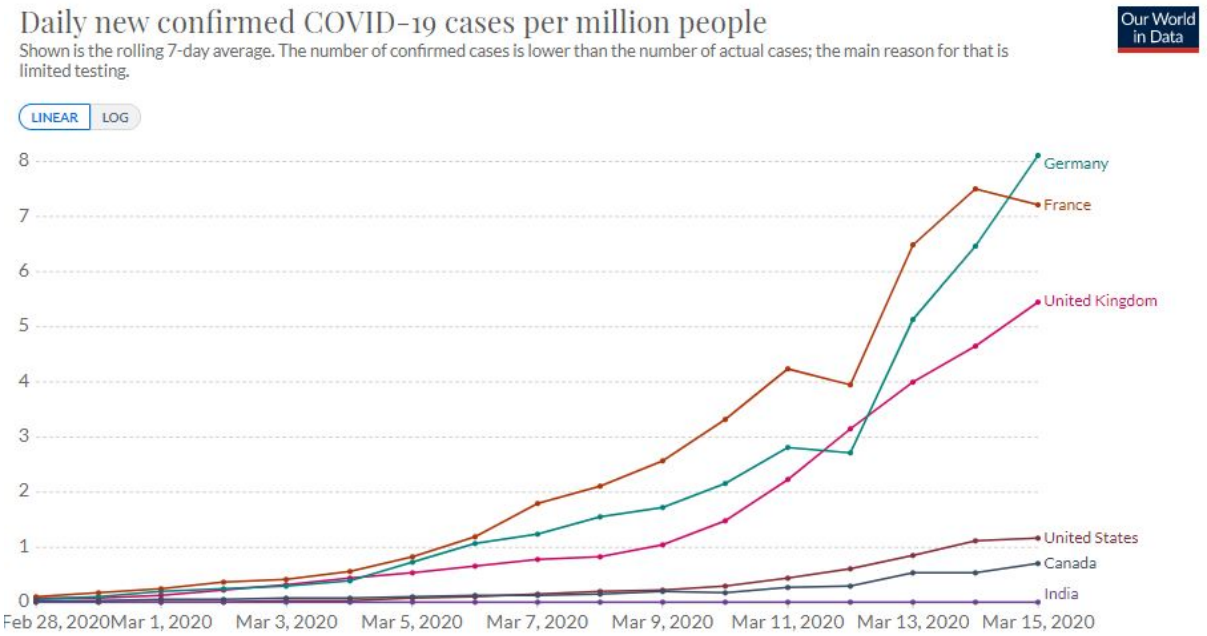


FIGURA 1.1 – Número de casos de covid-19 na Europa no início da pandemia. (OURWORLD, 2020)

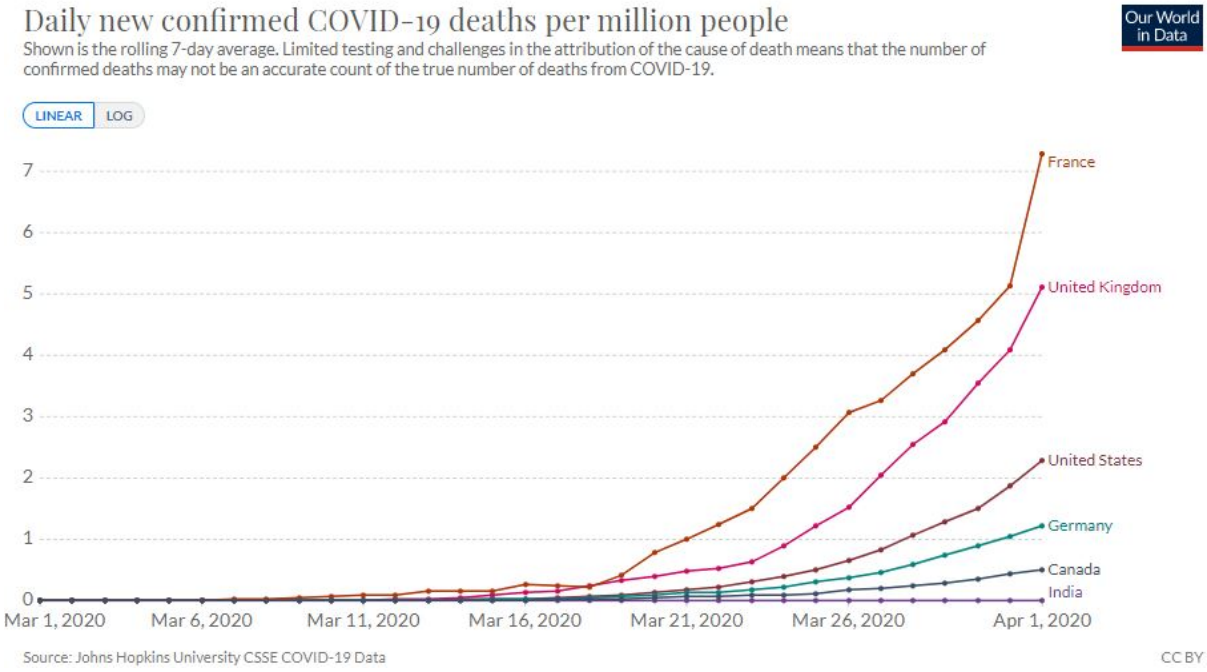


FIGURA 1.2 – Número de óbitos por covid-19 na Europa no início da pandemia. (OURWORLD, 2020)

Ainda no primeiro semestre de 2020, o covid-19 chega ao Brasil. O primeiro caso de covid foi registrado no dia 26 de fevereiro de 2020 no estado de São Paulo, um cidadão que havia voltado da Itália(AGENCIALBRASIL, 2020). No dia 17 de março é registrado o primeiro óbito decorrente da contaminação pelo Covid-19, no dia 20 de Março o ministério da saúde publica uma portaria confirmando a transmissão comunitária por todo o

território Brasileiro. A situação piorou drasticamente e o houve uma alta rotatividade do ministério da saúde. Nesse cenário, o presidente do Brasil, Jair Mesias Bolsonaro, estimulava, sem fundamenos científicos, o tratamento precoce com a utilização do fármaco hidroxicloroquina, além de incentiar a população a andar pelas ruas sem o uso de máscara. As consequências para a população brasileira foram um desastre, o número de óbitos cresceu drasticamente, uma nova variante do covid-19 surgiu na região de Manaus e duas variantes de outros países, ìndia e Inglaterra, foram detectadas no território nacional.

A seguir, tem-se os gráficos dos números totais confirmados no Brasil, do início da pandemia até o dia 22/06/2021 de casos e de óbitos respectivamente.

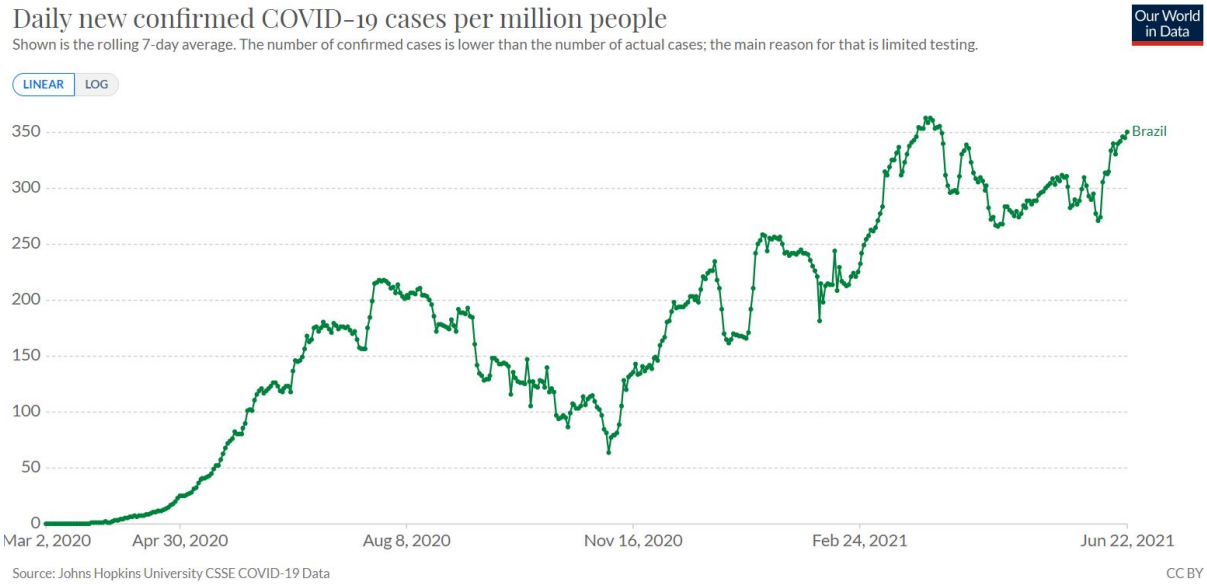


FIGURA 1.3 – Número de casos de covid-19 no Brasil até o dia 22/06/2021. (OURWORLD, 2020)

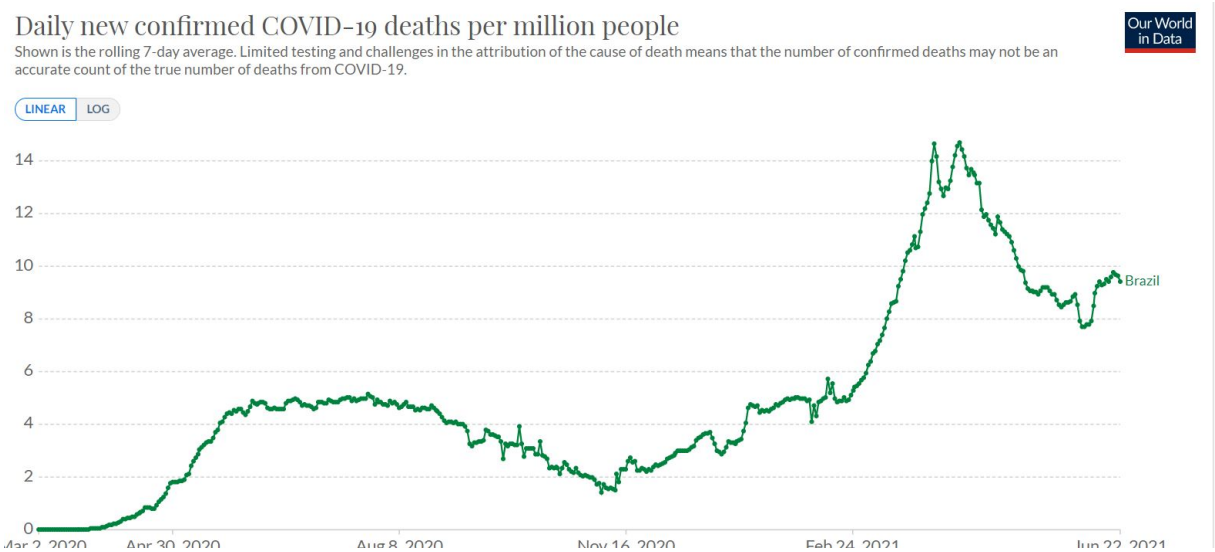


FIGURA 1.4 – Número de óbitos por covid-19 no Brasil até o dia 22/06/2021. (OURWORLD, 2020)

1.1.2 Impactos da pandemia no transporte aéreo

Durante o período de análise da pesquisa, o ano de 2020 foi marcado por uma grande retração na demanda por voos no mercado doméstico, 29,5% na demanda de passageiros pagos transportados(RPK) e de 27,5% na oferta de assentos-quilômetros(ANAC, 2020), a quantidade de passageiros pagos por vôo teve uma queda acumulada de 52,5% e o percentual de ocupação médio das aeronaves sofreu retração de 3,2%.

Já no mercado internacional, a retração foi ainda mais significante. As quedas de demanda e oferta foram de 71% e 62,6%, respectivamente, comparando 2020 com 2019. A seguir, tem-se gráfico comparativo de demanda por vôos, do número total de passageiros e de carga e correio, respectivamente.

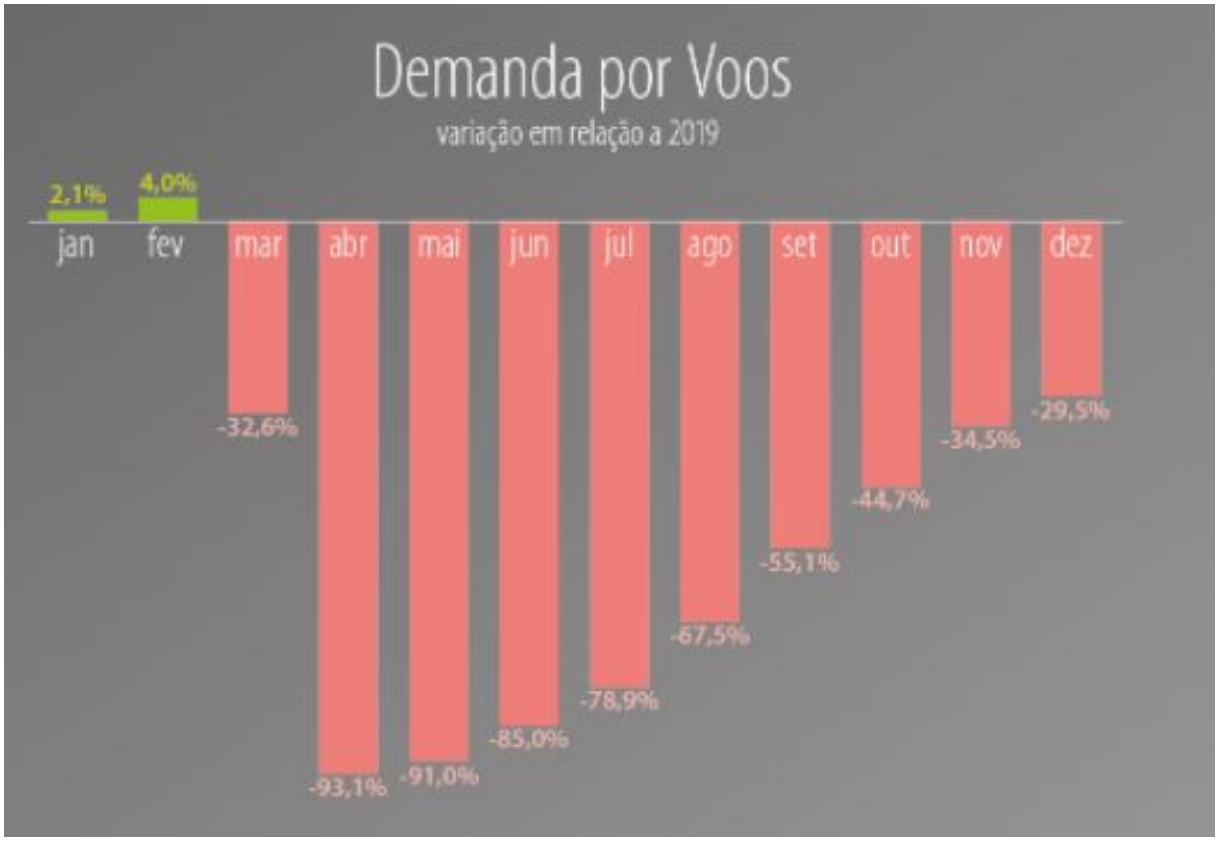
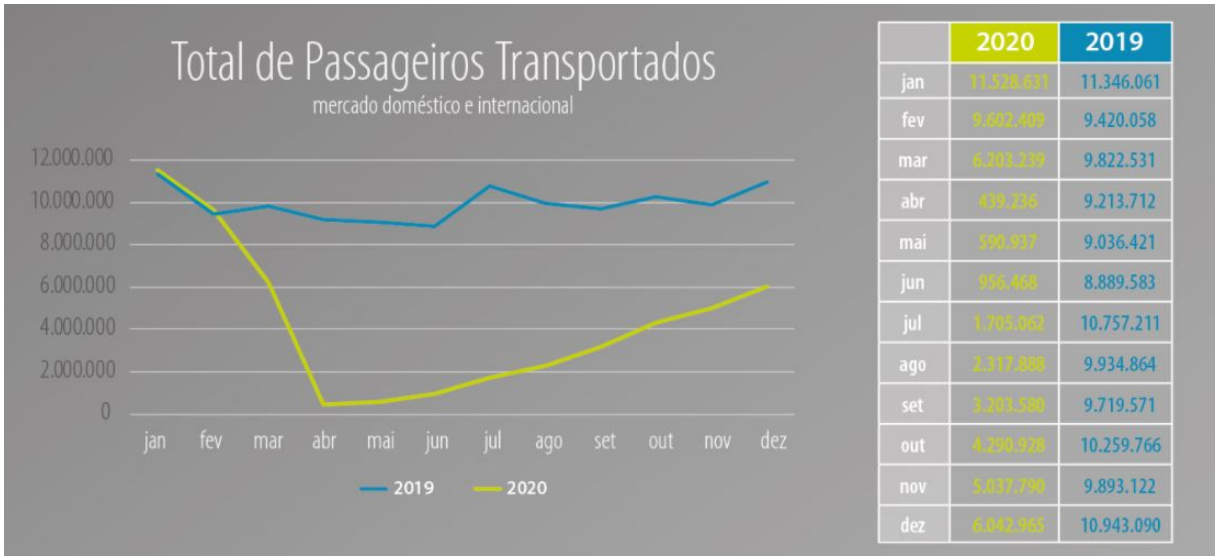


FIGURA 1.5 – Variação na demanda por voos, comparativo entre 2020 e 2019 (ANAC, 2020)



	2020	2019
jan	11.528.631	11.346.061
fev	9.602.409	9.420.058
mar	6.203.239	9.822.531
abr	439.236	9.213.712
mai	590.937	9.036.421
jun	956.468	8.889.583
jul	1.705.062	10.757.211
ago	2.317.888	9.934.864
set	3.203.580	9.719.571
out	4.290.928	10.259.766
nov	5.037.790	9.893.122
dez	6.042.965	10.943.090

FIGURA 1.6 – Número total de passageiros, comparativo entre 2020 e 2019. (ANAC, 2020)

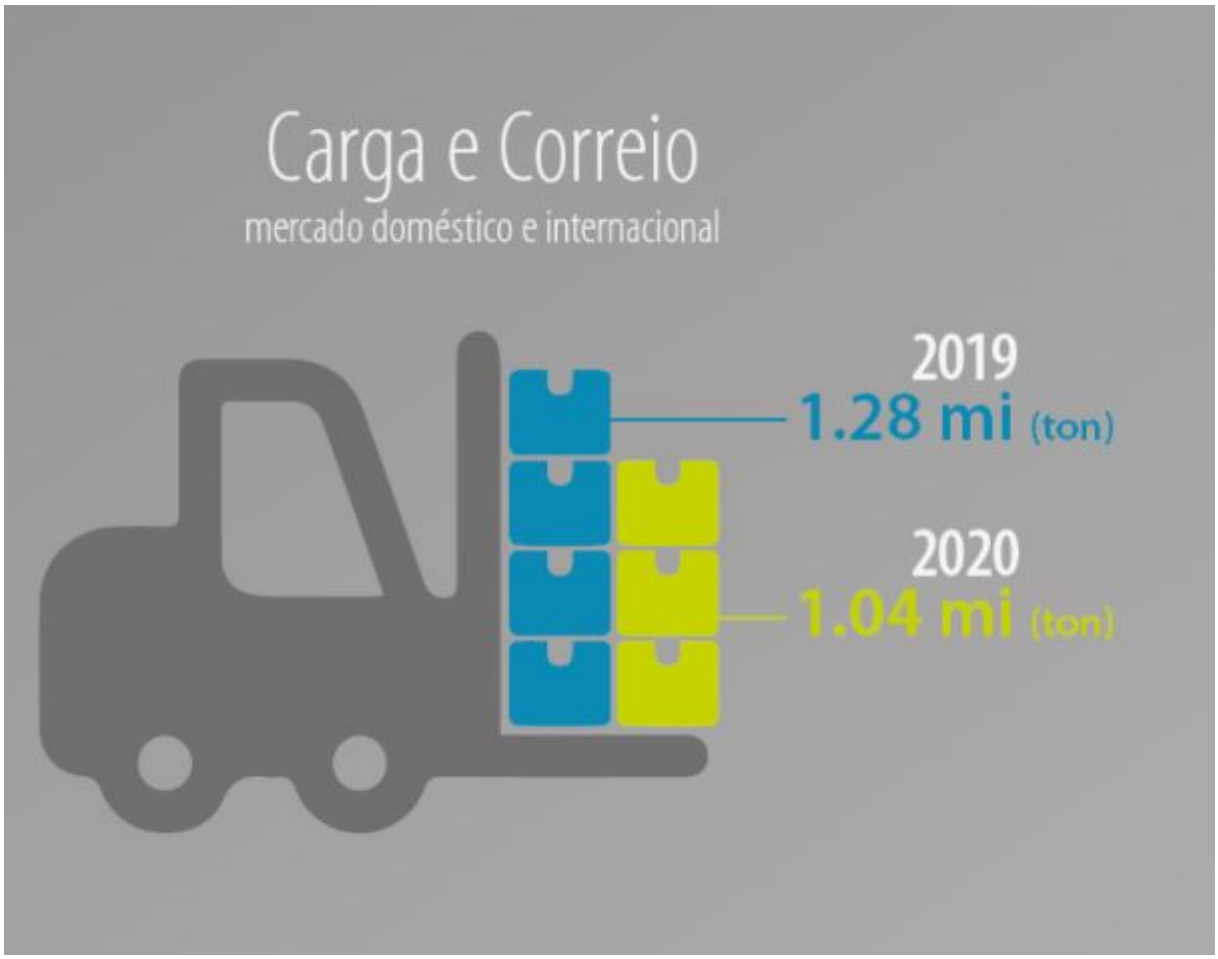


FIGURA 1.7 – Carga e Correio, comparativo entre 2020 e 2019. (ANAC, 2020)

Embora se tenha tal cenário de demanda, as medidas restritivas tomadas para conter a disseminação do vírus (distanciamento,aferição de temperatura, higienização das mãos

entre outras) acabaram gerando um maior tempo de embarque e desembarque, propiciando o atraso das programações de voos.

O seguinte estudo visa a análise e o entendimento dos atrasos ocorridos nos principais aeroportos brasileiros durante o período da pandemia do sars-cov-19.

1.2 Objetivo

1.2.1 Objetivo Geral

O objetivo geral dessa pesquisa é fazer uma análise preditiva de atrasos nos principais aeroportos brasileiros, utilizando para esse fim uma árvore de decisão composta por variáveis classificatórias obtidas das bases de dados TATIC, de dados das condições climáticas dos aeroportos e de dados de demanda disponibilizados pela ANAC.

1.2.2 Objetivos específicos:

- Tratar as bases de dados disponíveis e uni-las.
- Criar variáveis classificatórias..
- Desenvolver o código da árvore de decisão
- aplicar a base de dados na árvore de decisão criada.
- Fazer a análise dos resultados obtidos e, a análise conclusão da pesquisa.

1.3 Organização do trabalho

O trabalho foi organizado em 4 capítulos e a conclusão, sendo eles:

1.3.1 Capítulo 1

O capítulo 1 retrata a contextualização do tema da pesquisa e o objetivo geral e os objetivos específicos.

1.3.2 Capítulo 2

O capítulo 2 contém a revisão bibliográfica da pesquisa.

1.3.3 Capítulo 3

O capítulo 3 possui a metodologia utilizada no trabalho.

1.3.4 Capítulo 4

No capítulo 4 é feita a discussão dos resultados obtidos durante a pesquisa.

2 Revisão Bibliográfica

2.1 Análise de tráfego aéreo

Além de melhorias tecnológicas e operacionais, o uso eficiente de os dados operacionais também são essenciais para melhorar o CTA e trazer ganhos de eficiência para o tráfego aéreo . A cada minuto, um grande volume de dados é produzido a partir do planejamento de voo para a execução da operação de voo e ficará cada vez mais disponível e acessível à medida que ocorre a transformação digital do sistema de CTA. Explorando esses grandes dados de aviação com técnicas analíticas avançadas são uma forma promissora para melhorar avaliar e prever o desempenho operacional do tráfego aéreo e desenvolver novos ferramentas de apoio à decisão para CTA. Os dados da trajetória da aeronave são um exemplo de dados operacionais que se tornaram cada vez mais acessível com novas tecnologias de vigilância. Com *Automatic Dependent Surveillance-Broadcast* (ADS-B), dados de trajetória de aeronaves de alta fidelidade, incluindo dados espaciais e informações temporais, são registradas e disponibilizadas ao público em geral por meio serviços de rastreamento online, como (FLIGHRADAR, 2019) e (FLIGHAWARE, 2019).

2.1.1 Conceitos básicos e características do controle de tráfego aéreo

CTA é um elemento chave do transporte aéreo e tem como função principal garantir um fluxo seguro e ordenado de aeronaves através do espaço aéreo, tornando-se cada vez mais significativo com o crescimento das operações aéreas. CTA pode ser geralmente subdividido em três funções distintas:

- Controle de Tráfego Aéreo (CTA);
- Gestão do espaço aéreo (GEA); e
- Gerenciamento do Fluxo de Tráfego Aéreo (GFTA).

O objetivo principal do CTA é garantir que as separações padronizadas sejam mantidas entre a aeronave e entre a aeronave e o solo, evitando assim a ocorrência de acidentes e

incidentes aeronáuticos. No serviço CTA, cada instalação de controle é responsável pela segurança das operações aéreas em seu volume de espaço aéreo alocado, e o piloto deve cumprir as autorizações de tráfego aéreo emitidas pelos controladores de tráfego aéreo. As unidades de controle são organizadas em uma hierarquia operacional para gerenciar o tráfego nas diferentes fases do voo, de porta a porta. A Torre de Controle é responsável por controlar a aeronave no solo e nas proximidades dos aeroportos. O Controle de Aproximação apóia operações aéreas nos terminais, na transição entre o aeroporto e as aerovias durante as fases de chegada e partida. Por fim, o Centro de Controle é responsável por controlar aeronaves no espaço aéreo em rota. A figura a seguir representa as instalações que controlam o voo ao longo do seu ciclo de vida, desde a porta de embarque até à porta de desembarque.(OLIVEIRA, 2021)

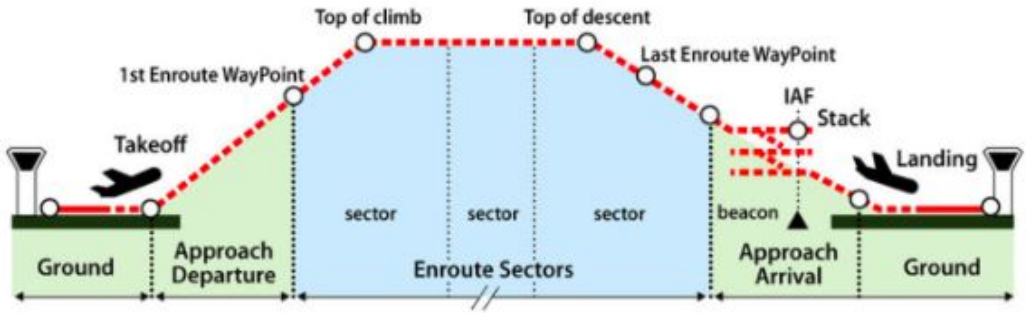


FIGURA 2.1 – Ilustração de um voo controlado ao longo do seu ciclo de vida. (OLIVEIRA, 2021)

As unidades de controle de tráfego aéreo também são responsáveis por garantir os serviços de informação de voo aos usuários, fornecendo informações do espaço aéreo necessárias para o planejamento e executando suas operações, tais como clima aeronáutico, status de equipamentos, manutenção do aeroporto, ou quaisquer outras notificações relacionadas a restrições que podem afetar a segurança das operações. O serviço de informação de voo também auxilia nas operações de busca e resgate de aeronaves. Em regiões onde apenas serviço de informação de voo são fornecidos, os pilotos são responsáveis pela manutenção das separações de segurança. Os serviços de tráfego aéreo (CTA ou SIV) são estabelecidos com base na categoria definida para cada espaço aéreo, que pode variar de acordo com sua complexidade. A função principal do controle aéreo é estruturar e alocar o espaço aéreo para acomodar diferentes volumes de tráfego, diferentes tipos de operação e atender diferentes objetivos, com foco no planejamento das operações no horizonte de longo prazo. O controle aéreo requer coordenação intensa entre várias partes interessadas e deve levar em consideração vários fatores como demanda, capacidade, comunicação, navegação e vigilância infraestrutura, restrições ambientais etc. Finalmente, a função principal do CFTA é ajustar os fluxos de tráfego em resposta a incompatibilidades entre

demanda e capacidade para mitigar atrasos. CFTA aplica iniciativas de gerenciamento de tráfego em prazos estratégicos e táticos para equilibrar a demanda e capacidade em recursos aeroportuários e do espaço aéreo. Iniciativas de gerenciamento de tráfego estratégicos são aplicados em uma região ou escala nacional em horizontes de planejamento mais longos (2h - 8 h), enquanto iniciativas de gerenciamento de tráfego táticos são aplicados em áreas localizadas em horizontes de planejamento mais curtos ($<$ 2 h). IGT diferem com base no características do recurso restrito, o mecanismo de controle e o horizonte de tempo.(OLIVEIRA, 2021)

2.1.2 Sistema de controle do espaço aéreo brasileiro

O Sistema de Controle do Espaço Aéreo Brasileiro (SISCEAB), é administrado pelo Departamento de Controle do Espaço Aéreo (DECEA). O SISCEAB contabiliza 22 km^2, sendo divididos em 5 Regiões de Informação de Voo. A estrutura ainda é composta por 42 estações de controle de aproximação e 59 torres de comando. DECEA também gerencia a infraestrutura de Comunicação, Navegação e Vigilância para ATM, que é composto por cerca de 90 estações de telecomunicações, 170 radares e 50 *Instrument Landing Systems* (ILS).

Como outros países, o Brasil implementou um programa específico (SIRIUS) para abordar melhorias no SISCEAB e se adequar ao crescimento do voo projetado demanda, com base na metodologia ASBU. As principais tecnologias e conceitos são descritas abaixo:

- *Performance-Based Navigation (PBN)*:O conceito visa melhorar a estrutura do espaço aéreo baseada em tecnologias de navegação, com GNSS como sua principal fonte. Com base nesse conceito, é possível flexibilizar as trajetórias de vôo, permitindo ganhos na capacidade do espaço aéreo. Atualmente, as rotas PBN representam 82% da movimentação dos voos.

- *Automatic Dependent Surveillance-Broadcast (ADS-B)* : a tecnologia permite um rastreamento mais preciso da aeronave por meio da transmissão de informações derivadas do sistema de navegação da aeronave com taxas de atualização mais altas, aumentando a consciência situacional dos controladores de tráfego aéreo. ADS-B fornece a oportunidade de reduzir os padrões de separação de segurança, impactando diretamente na capacidade do espaço aéreo. A tecnologia foi implantada na Bacia do Petróleo de Campos e atualmente encontra-se em fase de testes para implantação em todo espaço aéreo brasileiro.

- *Digital Data Link Communications (CPDLC)* : consiste em um sistema digital de intercâmbio entre controladores de tráfego aéreo e pilotos. Interface CPDLC permite uma troca mais rápida de informações e espera-se que reduza os erros que surgem

da comunicação de voz. CPDL é usado na Oceanic FIR para voos transcontinentais e também nas torres de comando centrais brasileiras. A tecnologia encontra-se em fase de testes para implementação no espaço aéreo continental associado à ADS-B.

- *System-Wide Information Management (SWIM)* : O conceito consiste em padrões, infraestrutura e governança que permite o gerenciamento do controle de tráfego aéreo- informações relacionadas e sua troca entre partes qualificadas por meio serviços interoperáveis. Infraestrutura SWIM permitirá diferentes transportes aéreos usuários compartilhem informações para aumentar a consciência situacional do sistema e eficiência.

- *Collaborative Decision-Making (CDM)* : o conceito visa aumentar a participação dos usuários nas decisões de CTA, com base em dados e procedimentos compartilhados definido entre as partes. Por exemplo, espera-se que contabilize melhor preferências do usuário em decisões de gerenciamento de fluxo de tráfego, como escolhas de rota.

- *Automation Support (AS)* : O desenvolvimento de ferramentas automatizadas visa auxiliar tomada de decisão em CTA, especialmente no equilíbrio de demanda e capacidade. Atualmente, muitas decisões são tomadas com base na experiência de profissionais de CTA, que pode resultar em desempenho abaixo do ideal. O DECEA usa algumas ferramentas, como ,para sequenciamento tático e programação de voos de chegada nos CTAs mais movimentados, (AMAN - Gerente de Chegada) e para gestão de fluxo nos setores FIR (SIGMA), mas ainda há uma grande oportunidade de desenvolver novas ferramentas para melhor prever e otimizar o desempenho do sistema.

- *Trajectory-Based Operations (TBO)* : Como mencionado antes, o futuro ATM conceito operacional irá alavancar todos os avanços tecnológicos em sistemas de comunicação, navegação, vigilância, informação e automação para entregar trajetórias mais eficientes e previsíveis que satisfaçam os usuários do espaço aéreo. O conceito procura substituir a gestão do fluxo de tráfego local por porta a porta gerenciamento de trajetória que considera as quatro dimensões da trajetória de voo (latitude, longitude, altitude e hora). O conceito enfatiza a importância de informações de rastreamento de aeronaves para gerenciamento preciso.(OLIVEIRA, 2021)

2.1.3 Análise da programação de pousos

Ao se aproximar do aeroporto, o piloto entra em contato com a torre de controle que faz uma estimativa do horário de pouso, quanto mais próximo desse horário inicial, para um cenário estático, desconsiderando os efeitos dinâmicos de preferência de pouso para diferentes aeronaves, maior será a eficiência da ordenação no controle do tráfego

aeroportuário. Cada aeronave possui uma velocidade ideal para chegar ao aeroporto o mais rápido possível e é chamada de velocidade de cruzeiro. Toda vez que a aeronave é forçada a reduzir a velocidade, acelerar ou circular o aeroporto, aguardando a autorização de pouso da torre de comando, há um acréscimo de gasto de combustível, o que em grande volume pode gerar gasto monetário excessivo com baixa eficiência de alocação monetária. Além desses fatores, é interessante notar que o intervalo de tempo entre dois voos deve ser sempre maior que um intervalo mínimo, que é chamado de tempo de separação. Essa janela de tempo se faz necessária, já que o pouso de uma aeronave provoca turbulência e um pouso acontecendo seguidamente de outro pode ocasionar perda da estabilidade do avião e, consequentemente, um possível acidente (Mullins,1996).

2.2 Análise de dados com uma Árvore de decisão

Uma árvore de decisão é uma forma de classificação de dados que possui regras claras para a subpartição dos dados em análise. Normalmente, uma árvore de decisão começa com um único nó e esse se divide de acordo com as regras definidas para a classificação, dando origem a uma camada de nós e esses, sequencialmente, são testados com um novo conjunto de regras, gerando uma outra camada de nós. O processo continua iterando até que o input em questão seja classificado.

A seguir, tem-se um exemplo de uma árvore de decisão para classificação de clientes de um banco que recebem oferta de empréstimo entre os casos de sucesso e fracaso, com algumas variáveis classificatórias como renda, nível educacional e média de gastos no cartão de crédito.

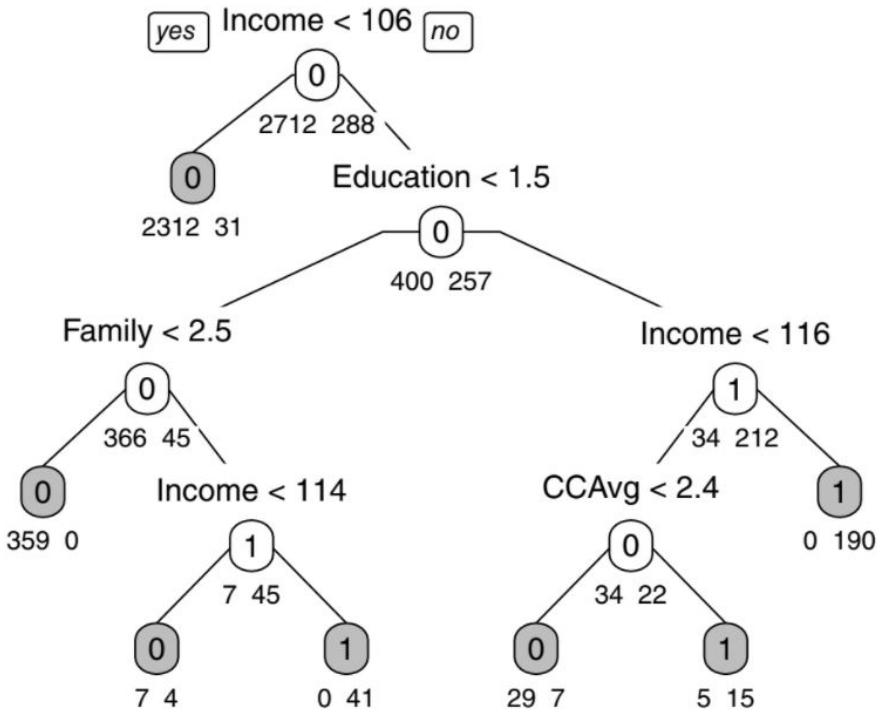


FIGURA 2.2 – Árvore de classificação para clientes de um banco com relação a oferta de empréstimo. (SHUMELI, 2018)

Os números internos aos nós representam 0 para fracasso e 1 para sucesso. Os valores a cima dos nós são as regras que classificam os dados e os valores a baixo dos nós representam o número de dados que passaram pela regra.

Dois conceitos muito importantes para o entendimento das árvores de decisão são a partição recursiva e a poda da árvore.

2.2.1 Partição Recursiva

Considerando Y como o output da árvore de decião e X1,X2,X3...Xp como sendo as variáveis classificatórias. A partição recursiva dividiria o espaço p dimensional das variáveis de entrada em retângulos sem interseção. Essa divisão seria obtida recursivamente, ou seja, a partir de uma divisão inicial do espaço amostral em retângulos menores, outra partição seria feita originando novos retângulos até que, idealmente, todos os dados fiquem classificados em um único retângulo. A seguir um exemplo demonstrativo.

Uma fábriaca de cortadores de grama gostaria de descbrir um meio de classificar famílias de uma região entre compradoras e não compradoras do seu produto. Uma amostra aleatória de 12 famílias compradoras e 12 não compradoras é obtida na região em questão, dados na tabela a seguir:

TABELA 2.1 – Amostra aleatória das famílias da região.

Número da família	**Renda(000USD)**	**Tamanho da propriedade(000ft²)**	**Possui cortador de grama?**
1	60,0	18,4	Sim
2	85,5	16,8	Sim
3	64,8	21,6	Sim
4	61,5	20,8	Sim
5	87,0	23,6	Sim
6	110,1	19,2	Sim
7	108,0	17,6	Sim
8	82,8	22,4	Sim
9	69,0	20,0	Sim
10	93,0	20,8	Sim
11	51,0	22,0	Sim
12	81,0	20,0	Sim
13	75,0	19,6	Não
14	52,8	20,8	Não
15	64,8	17,2	Não
16	43,2	20,4	Não
17	84,0	17,6	Não
18	49,2	17,6	Não
19	59,4	16,0	Não
20	66,0	18,4	Não
21	47,4	16,4	Não
22	33,0	18,8	Não
23	51,0	14,0	Não
24	63,0	14,8	Não

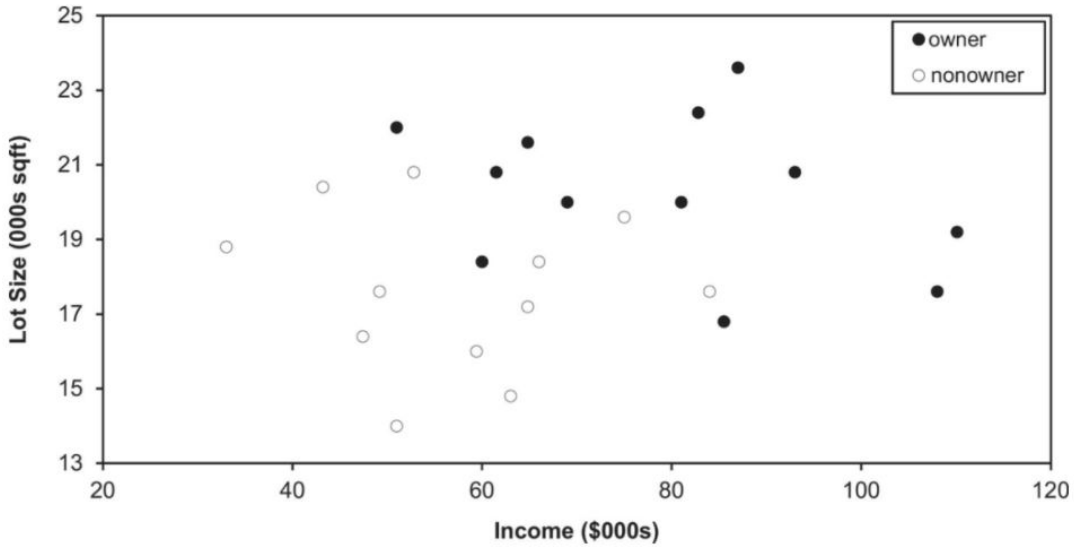


FIGURA 2.3 – Donos e não donos de cortadores de grama de acordo com renda e tamanho da propriedade. (SHUMELI, 2018)

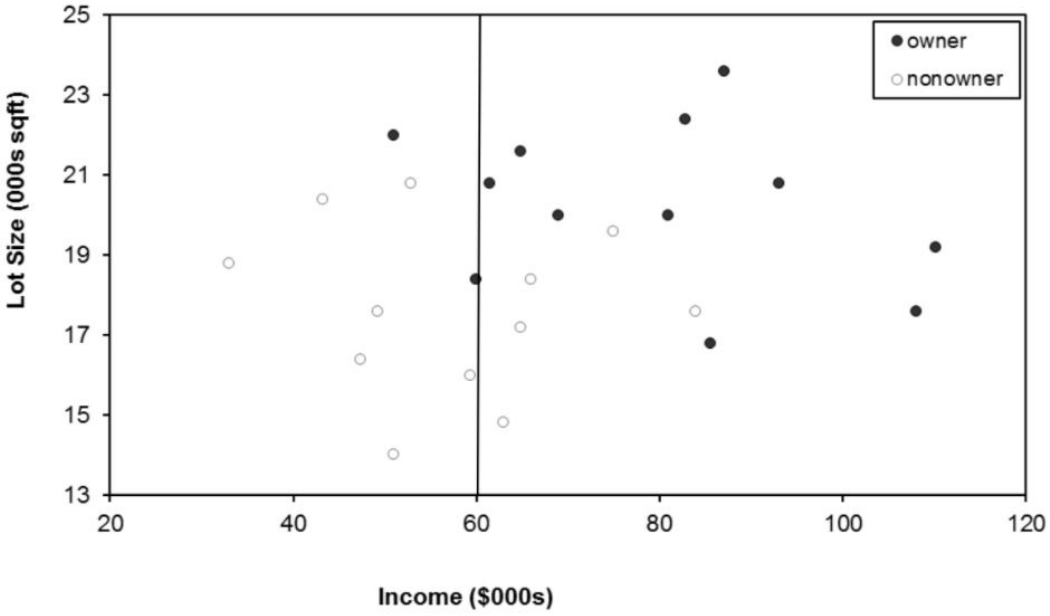


FIGURA 2.4 – Tentativa de divisão do espaço amostral entre dois grupos de donos e não donos de cortadores motorizados. (SHUMELI, 2018).

O algoritmo tentou selecionar dois retângulos de modo a ter o mínimo de impurezas posssíveis, ou seja, tentou fazer que cada retângulo da divisão possuísse uma única classe. O eixo X variava de uma renda de 20 a 120 mil dólares enquanto o eixo Y variava de 13 a 25 mil pés quadrados e em cada um desses valores entre as duas extremidades, de forma contínua, poderia ser dividido o retângulo maior em dois outros retângulos menores de modo a gerar o mínimo de impureza, assim funciona o algoritmo de uma árvore de decisão

2.2.2 Medidas de impurezas

A vários índices para apuração de medidas de impurezas. Os dois mais utilizados são os índice de Gini e a medida de entropia.

O índice de Gini é definido da seguinte maneira:

Índice de Gini = 1- $\sum_1^k$ Pi Em que Pi representa a frequência relativa de cada classe no retângulo em que se encontra e k é o número de classes. Quanto mais próximo de zero é o índice, maior será a pureza da classificação e quanto mais próximo de um maior será a impureza. A seguir temos uma análise gráfica do índice de Gini com a variação dos Pi, para um problema de duas classes:

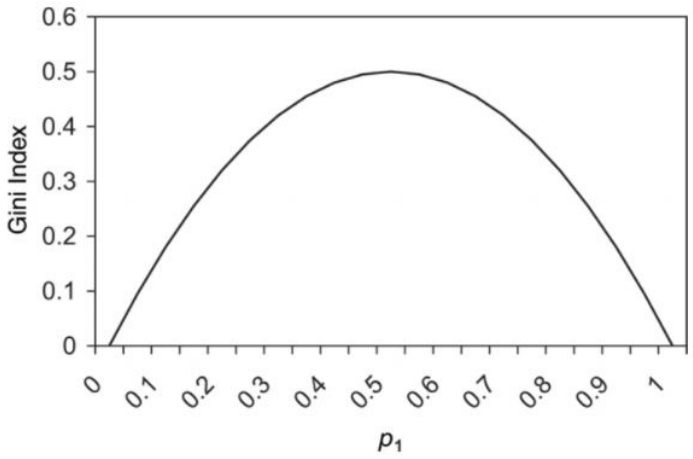


FIGURA 2.5 – Índice de Gini em função de Pi em um problema de duas classes. (SHUMELI, 2018).

A segunda medida de impureza que foi citada é a entropia. A entropia é definida da seguinte maneira para um retângulo x qualquer:

Entropia de x = - $\sum_1^k$ Pi*log2(Pi)

Seguindo a mesma notação definida para o índice de Gini. A entropia varia de 0, em que estaria com a maior pureza possível, a log2(k) situação de maior impureza.

Voltando ao exemplo anterior citado dos cortadores de grama motorizados, há os donos e os não donos de cortadores, logo trata-se de um problema de duas classes. Para o primeiro retângulo temos 12 donos e 12 não donos, o que gera tanto para o índice de Gini, 0,5, quanto para a entropia, 1, o valor de maior impureza possível para o problema. Após a primeira divisão, calcula-se o índice de Gini de 0,359 e a entropia de 0,779. Desse modo, percebe-se que a impureza do sistema foi reduzida. E assim o algoritmo de uma árvore de decisão continua fazendo a partição dos retângulos de modo a gerar a menor

entropia possível. Para o problema citado, será apresentado a evolução do esquema de retângulos e da árvore de decisão.

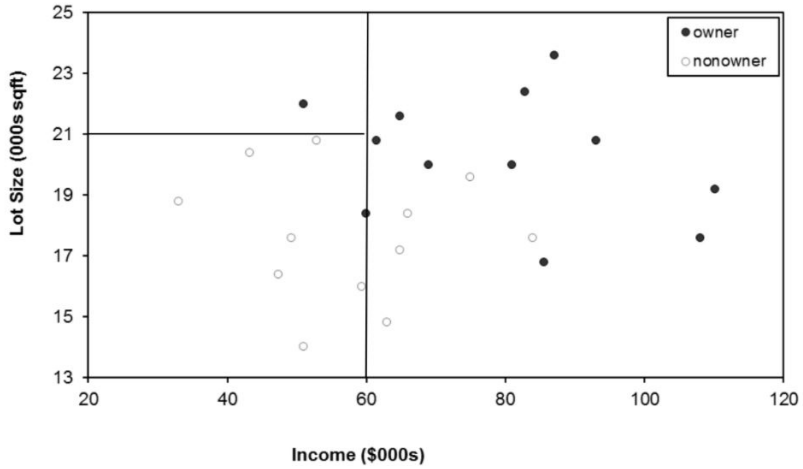


FIGURA 2.6 – .Segunda divisão do retângulo inicial para o exemplo dos cortadores de grama motorizados (SHUMELI, 2018).

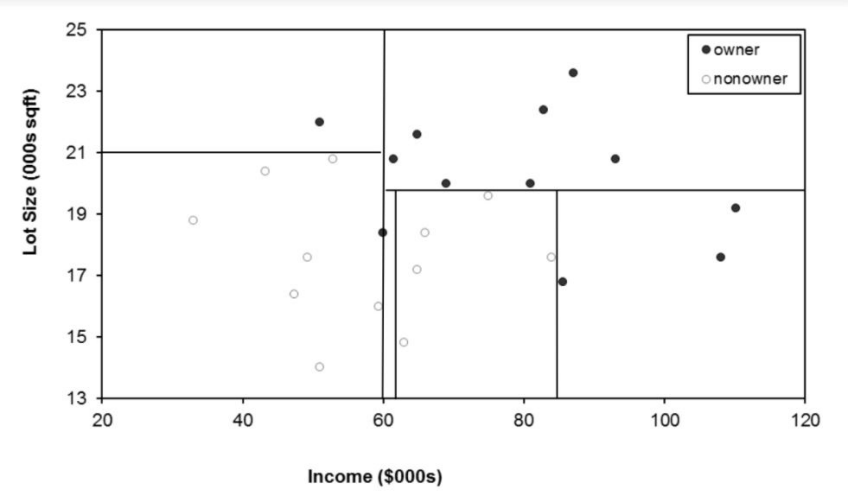


FIGURA 2.7 – .Resultado da divisão do retângulo para o exemplo dos cortadores de grama motorizados (SHUMELI, 2018).

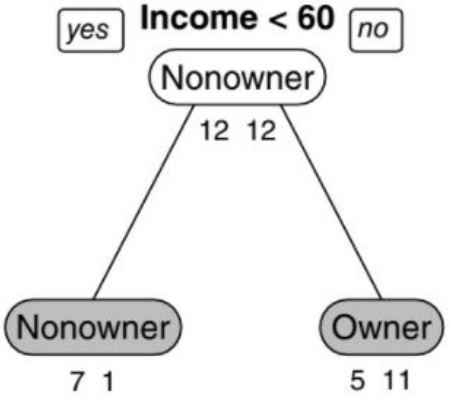


FIGURA 2.8 – .Primeira divisão da árvore para o exemplo dos cortadores de grama motorizados (SHUMELI, 2018).

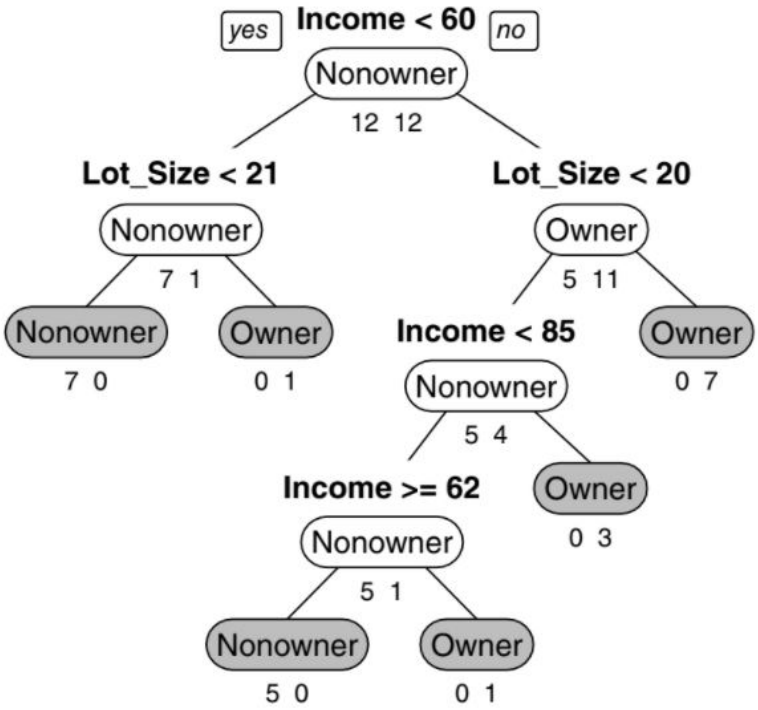


FIGURA 2.9 – .Arvore final para o exemplo dos cortadores de grama motorizados (SHUMELI, 2018).

2.2.3 Avaliando a performance da árvore de decisão

Para um árvore de decisão, a adequação ao modelo da base de treino não é suficiente para garantir que se terá uma boa classificação final, faz-se necessário o teste da árvore com dados fora da amostra do problema para se garantir uma melhor solução do problema. Há duas grandes razões para que isso seja necessário:

- A estrutura da árvore de decisão pode ser instável, mudando drasticamente dependendo dos *inputs* colocados
- Uma árvore que se adequa totalmente a base de dados de treino, provavelmente levará ao problema de *overfitting*

2.2.3.1 Evitando o *overfitting*

Para estruturas de árvores muito grandes, é comum se chegar ao problema de *overfitting*. Avaliando um número grande de árvores de todos os tamanhos, é possível perceber que o erro decresce até o ponto de *overfitting*. Desse modo, faz-se necessário a avaliação dos tamanhos das estruturas das árvores a fim de evitar esse problema. Uma das formas mais comuns e eficientes de contornar esse impasse é a poda da árvore que será abordada a seguir.

2.2.3.2 Poda de árvore

Em "Uma comparação empírica entre métodos de podas de árvores de decisão"(MINGUERS, 1989) é feita uma análise entre cinco métodos de poda de árvores:

- Poda Erro-Complexidade
- Poda por valor crítico
- Podando pelo mínimo erro
- podando pelo erro reduzido
- Podando pelo erro pessimista

Os cinco métodos foram comparados usando base de dados com grande domínio, possuindo muitas características e possibilitando várias classificações distintas. 60% dos dados foram selecionados aleatoramente para ser a base de treino e os outros 40% restantes foram divididos em duas base de testes. É interessante notar que, além dos métodos de poda, também foram utilizadas bases de dados para o processo de validação cruzada a fim de reduzir o dilema do *overfitting*. Dois dos principais critérios para avaliação das árvores de decisão foram:

- Tamanho: A complexidade da árvore de decisão, estatisticamente, melhora o poder explicatório da base de treino, entretanto, piora o , preditivo em um teste independente. Desse modo, o algoritmo da árvore de decisão será melhor quando minimizar o tamanho da árvore. Nessa pesquisa, as folhas foram selecionadas para medir o tamanho da árvore.

- Acurácia: Se refere a habilidade da árvore de decisão de classificar uma base de dados de teste independente. Ela é medida pela taxa de erro, isto é, a proporção de predições erradas que a árvore faz em seu espaço amostral de predição.

A pesquisa concluiu que todos os cinco métodos de poda de árvore são eficientes, sendo que a poda da árvore aumentou de 20 a 25% a acurácia das predições.

3 Metodologia

3.1 Obtenção e tratamento da base de dados

Três bases de dados foram selecionadas para a seguinte pesquisa.

- TATIC: É oriunda do TATIC TWR, sistema de operação e gerenciamento para torres de controle baseado no uso de *strips* eletrônicas, tendo como principais finalidades o auxílio ao controlador de tráfego aéreo na organização do tráfego aéreo local, o armazenamento de informações operacionais relevantes para uso em nível local ou global e, por fim, a integração de dados com sistemas próprios ou de terceiros com vistas a auxiliar na tomada de decisões colaborativas. Essa base de dados possui as datas e horários das movimentações das aeronaves, tanto dentro do aeroporto quanto os registros de solicitações de autorizações na torre de controle

- Wheather: Foi obtida do *Iowa enviromental Mesonet* da universidade de Iowa(IOWA, 2020). Base de dados das condições climátias dos aeroportos brasileiros. Seus registros abrangem toda a janela de tempo da pesquisa, possuindo uma hora de frequência nas marcações

- Microdados de demanda e oferta: Com o intuito de ampliar o conhecimento da sociedade brasileira e de subsidiar a realização de pesquisas, estudos e análises mais abrangentes sobre o setor, a ANAC tem disponibilizado, na seção "Dados e Estatísticas" do seu portal na internet, relatórios, estudos e informações sobre as condições de mercado. Nesse sentido, a Agência agrega, para a livre consulta de qualquer interessado, a série histórica dos dados estatísticos do transporte aéreo do Brasil, com elevado grau de detalhamento. Estão contempladas informações sobre as quantidades de passageiros, carga e mala postal transportados, distância voada, combustível consumido, entre outras, por etapa de voo e por empresa aérea. Essas informações podem ser utilizadas para apurar importantes indicadores do setor, como demanda (RPK, RTK), oferta (ASK, ATK), participação de mercado, taxa de ocupação das aeronaves (Load Factor), entre outros. Os dados são retroativos até o ano de 2000 e estão disponíveis com elevado grau de detalhamento.

O processo de tratamento de dados foi iniciado com a base de dados TATIC. Primeiramente, filtrou-se os 29 principais aeroportos nacionais, tanto para pouso quanto para decolagem,utilizou-se o seguinte código na linguagem Python:

```
# filtrando a TATIC para os 29 aeroportos brasileiros de maior movimentacao
df = df[(df['ADep'] == 'SBAR') | (df['ADep'] == 'SBBE') |
        (df['ADep'] == 'SBBR') | (df['ADep'] == 'SBBV') |
        (df['ADep'] == 'SBCF') | (df['ADep'] == 'SBCG') |
        (df['ADep'] == 'SBCT') | (df['ADep'] == 'SBCY') |
        (df['ADep'] == 'SBEG') | (df['ADep'] == 'SBFL') |
        (df['ADep'] == 'SBGL') | (df['ADep'] == 'SBGO') |
        (df['ADep'] == 'SBGR') | (df['ADep'] == 'SBJP') |
        (df['ADep'] == 'SBKP') | (df['ADep'] == 'SBMO') |
        (df['ADep'] == 'SBMQ') | (df['ADep'] == 'SBPA') |
        (df['ADep'] == 'SBPJ') | (df['ADep'] == 'SBPV') |
        (df['ADep'] == 'SBRB') | (df['ADep'] == 'SBRF') |
        (df['ADep'] == 'SBRJ') | (df['ADep'] == 'SBSG') |
        (df['ADep'] == 'SBSL') | (df['ADep'] == 'SBSP') |
        (df['ADep'] == 'SBSV') | (df['ADep'] == 'SBTE') |
        (df['ADep'] == 'SBVT')]
```

FIGURA 3.1 – .Código que filtra as decolagens para os 29 principais aeroportos .

```
df = df[(df['ADes'] == 'SBAR') | (df['ADes'] == 'SBBE') |
        (df['ADes'] == 'SBBR') | (df['ADes'] == 'SBBV') |
        (df['ADes'] == 'SBCF') | (df['ADes'] == 'SBCG') |
        (df['ADes'] == 'SBCT') | (df['ADes'] == 'SBCY') |
        (df['ADes'] == 'SBEG') | (df['ADes'] == 'SBFL') |
        (df['ADes'] == 'SBGL') | (df['ADes'] == 'SBGO') |
        (df['ADes'] == 'SBGR') | (df['ADes'] == 'SBJP') |
        (df['ADes'] == 'SBKP') | (df['ADes'] == 'SBMO') |
        (df['ADes'] == 'SBMQ') | (df['ADes'] == 'SBPA') |
        (df['ADes'] == 'SBPJ') | (df['ADes'] == 'SBPV') |
        (df['ADes'] == 'SBRB') | (df['ADes'] == 'SBRF') |
        (df['ADes'] == 'SBRJ') | (df['ADes'] == 'SBSG') |
        (df['ADes'] == 'SBSL') | (df['ADes'] == 'SBSP') |
        (df['ADes'] == 'SBSV') | (df['ADes'] == 'SBTE') |
        (df['ADes'] == 'SBVT')]
```

FIGURA 3.2 – .Código que filtra os pousos para os 29 principais aeroportos .

O segundo passo foi transformar os horários da base de dados, que inicialmente eram *strings*, em um formato que permite comparação entre variáveis que possuem datas e

horas. Para isso foi utilizado a biblioteca *pandas* do Python, utilizando o seguinte código:

```
df['EOBT Previsto'] = pd.to_datetime(df['EOBT Previsto'],
                                     format='%Y/%m/%d %H:%M:%S', errors='coerce')

df['Autorizado Push-Back'] = pd.to_datetime(df['Autorizado Push-Back'],
                                            format='%d/%m/%Y %H:%M:%S', errors='coerce')

df['Acionamento Push-Back'] = pd.to_datetime(df['Acionamento Push-Back'],
                                             format='%d/%m/%Y %H:%M:%S', errors='coerce')

df['Autorizado Decolagem'] = pd.to_datetime(df['Autorizado Decolagem'],
                                            format='%d/%m/%Y %H:%M:%S', errors='coerce')

df['Decolagem'] = pd.to_datetime(df['Decolagem'],
                                 format='%d/%m/%Y %H:%M:%S', errors='coerce')

df['Pouso Real'] = pd.to_datetime(df['Pouso Real'],
                                  format='%d/%m/%Y %H:%M:%S', errors='coerce')

df['data'] = df['Acionamento Push-Back'].fillna(df['Pouso Real'])

df['ETA previsto'] = pd.to_datetime(df['ETA previsto'],
```

FIGURA 3.3 – .Código que transforma as *strings* em variáveis de data e hora comparáveis .

Como a base de dados Weather possui frequência de marcação de uma hora, os horários da base TATIC foram truncados para que seja possível a junção entre as duas bases. Dessa forma, foi criada a função *pegaHoras* e aplicando na variável para comparação entre as bases chamada de ”data”, utilizando o código a seguir:

```
#função para trucar a hora
def pegaHoras(data: pd.Timestamp):
    return data.replace(minute=0, second=0)


#definição da data base para TATIC
    df['data'] = df['data'].map(pegaHoras)
```

FIGURA 3.4 – .Código para truncar os horários da variável data para efeito de comparação entre as bases Weather e TATIC .

Após o processo de da função ”pegaHoras”, foi criada uma variável ”comparação”e nela foi colocados os seguintes dados: aeroportos de partida, aeroportos de chegada, data do vôo e hora da partida. Em seguida, foi feito o processo da junção das base de dados, utilizando o seguinte código:

```
dfm = df.merge(dfw, how='inner', on='comparativo')
```

FIGURA 3.5 – Junção das bases de dados .

O processo foi análogo para a junção com a base de dados de demanda.

3.2 Definição das variáveis de classificação

3.2.1 Magnitude de tempo em solo

Foram definidas duas variáveis auxiliares:

- Tempo de *taxi in*: horário em que a aeronave está completamente estacionada menos horário do pouso real
- Tempo de *taxi out*: horário da decolagem menos horário da autorização do *push back*

A variável magnitude de tempo em solo é representada pelo tempo de *taxi in* quando o processo é de pouso e pelo tempo de *taxi out*, quando de decolagem. O código seguinte foi utilizado para a formulação da variável:

```
df['TempoTaxiOut'] = df['Decolagem'] - df['Autorizado Push-Back']

df['TempoTaxiIn'] = df['Aeronave Estacionada'] - df['Pouso Real']

df['MagnitudeTempoEmSolo'] = df['TempoTaxiOut'].fillna(df['TempoTaxiIn'])
```

FIGURA 3.6 – .Definição da magnitude de tempo em solo

3.2.2 *Taxi in* e *taxi out* adicionais

Foram definidos os tempos adicionais de acordo com os valores iniciais estimados para cada processo. Dessa forma, todo resgistro real foi comparado com o teórico para extrair os intervalos de tempo adicionais.

```
df['TAdicionalTaxiOut'] = df['Autorizado Push-Back'] - df['EOBT Previsto'] + df['Decolagem'] - df[
    'Autorizado Decolagem']

df['TAdicionalTaxiIn'] = df['Pouso Real'] - df['ETA previsto'] + df['cArr'] - df['wArr'] + df[
    'Aeronave Estacionada'] - df['wPos']
```

FIGURA 3.7 – .Definição dos tempos de *Taxi in* e *taxi out* adicionais.

3.2.3 Orvalho

A base *Weather* possui os valores da temperatura registrada no momento e da temperatura de orvalho. Dessa forma, foi feita a análise se tinha orvalho ou não nos horários das operações.

```
df['orvalho'] = df['temperatura celcius'] <= df['ponto de orvalho celcius']
df['orvalho'] = df['orvalho'].astype(int)
```

FIGURA 3.8 – .Definição da variável de orvalho.

3.2.4 visibilidade

A base *Weather* também possui dados de visibilidade em milhas. Sendo assim, foi definida a variável "visibilidade"considerando o valor de 4 milhas como um limiar para dificuldade no processo de operação.

```
#Analise da visibilidade em milhas
df['visibilidadeRuim'] = df['visibilidade em milhas'] < 4 ##é uma boa métrica
df['visibilidadeRuim'] = df['visibilidadeRuim'].astype(int)
```

FIGURA 3.9 – .Definição da variável de visibilidade.

3.2.5 Variáveis de demanda

A base de dados de demanda fornecida pela ANAC possui variáveis já definidas que foram utilizadas para a análise da pesquisa, enumeradas a seguir:

- Carga paga
- Distancia percorrida

- Número de passageiros
- número de horas voadas

3.3 Análise com a Árvore de decisão

Primeiramente, para que a análise pudesse ser feita foi utilizada a biblioteca *Scikit*, biblioteca gratuita de *python* que possui vários algoritmos de grande utilidade para abordagem de análise utilizando *machine learning*. Após importar a biblioteca foi definido o vetor *input* com as variáveis classificatórias e também o vetor com as variáveis *booleanas* que definem se o vôo atrasou ou não. Os vetores foram divididos para que parte dos dados sejam usados para treino e parte para teste. Após as preparações iniciais, os vetores foram inseridos para a construção da árvore de decisão e testados para uma análise preditiva. Segue o código com esses procedimentos:

```
X = pd.DataFrame()
X = df[['MagnitudeTempoEmSolo', 'orvalho',
        'visibilidadeRuim', 'kg_payload', 'km_distancia',
        'nr_passag_pagos', 'nr_horas_voadas']]
X = X.fillna(0)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
                test_size=0.4, random_state=1) # 60% training and 40% test
Y = df['atraso']
clf = DecisionTreeClassifier()
clf = clf.fit(X_train,Y_train)
Y_pred = clf.predict(X_test)
```

FIGURA 3.10 – .Definição dos *inputs* e aplicação na árvore de decisão

A fim de otimizar a acurácia da árvore de decisão será realizado o processo de poda. Para isso será utilizado o algoritmo de poda por custo de complexidade.

3.3.1 Algoritmo de poda por custo de complexidade

Desenvovido por (BREIMAN L.; STONE, 1984), é um método de dois estágios. Primeiro gerando uma série de árvores podadas em diferentes quantidades e, em seguida, examinando o número de erros de classificação que cada uma delas comete com um conjunto de dados independente. Na poda, o método de custo de complexidade leva em consideração tanto o número de erros quanto a complexidade (tamanho) da árvore.

Cada nó da árvore é o ponto de partida para uma subárvore que terminará com várias folhas. Antes da poda, as folhas conterão exemplos pertencentes a apenas uma classe, mas, à medida que a poda avança, as folhas restantes incluirão exemplos de várias classes diferentes. Quando isso acontece, os exemplos na folha são examinados e a folha é alocada à classe que ocorre com mais frequência. A taxa de erro de um folha é então a proporção de exemplos de treinamento que não pertencem a essa classe. Se a subárvore é podada, então a taxa de erro esperada é a do nó inicial, que se torna uma folha. Se a subárvore não for podada, a taxa de erro é a média das taxas de erro, sendo que as folhas são ponderadas pelo número de exemplares em cada folha. Com os dados de treinamento, a poda sempre levará a um aumento na taxa de erro, e esse aumento é uma medida do valor da subárvore. Dividindo este aumento pelo número de folhas na subárvore dá uma medida da redução do erro por folha para essa subárvore. Esta é a medição da complexidade do erro.

Analisando o algortimo do ponto de vista numérico, é necessário otimizar a função de custo de complexidade: Rα(T)=R(T)+α|f(T)| em que:

- R(T) é o erro de aprendizado
- f(T) é a função que retorna o número de folhas da árvore T
- α é o parâmetro de regularização

Para o erro de aprendizado: R(T)=$\sum$tϵf(T)r(t)p(t)=$\sum$tϵf(T)R(t), em que:

- $\sum$tϵf(T)R(t) é a soma do erro de classificação para cada folha
- r(t)=1-maxkp(Ck - t) é a frequência do erro de classificação
- p(t)=n(t)/n em que n(t) é a frequência no nó t e n é a frequência total

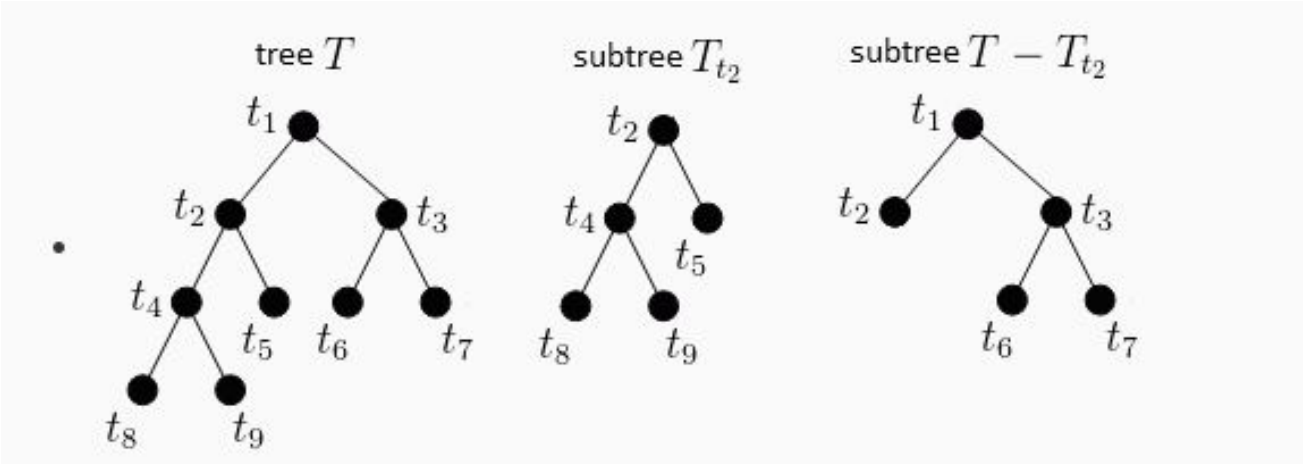


FIGURA 3.11 – .Exemplo de poda para o algoritmo

Fazendo uma poda de uma subárvore Tt:

- Rα (T - Tt) - Rα(T) variação na função de custo de complexidade
- Rα(T - Tt) - Rα(T) = R(T - Tt) - R(T)+α(|f(T-Tt)|-|f(T)|)=R(t)-R(Tt)+α(1-|f(Tt)|)
- α' = R(t) - R(Tt) / |f(Tt)| - 1
- Avaliando as variações

  nula se $\alpha = \alpha$'

  negativa se $\alpha < \alpha$'

  positiva se $\alpha > \alpha$'

Pseudocódigo:

- Inicialização:

  T1 é a árvore obtida com α1=0

  para minimizar R(T)

- Passo 1:

  Selecionar o nó t ϵ T^1 que minimiza a expressão:

  g1(t) = R(t) - R(T1t) / |f(T1t)| - 1

  Seja t1 este nó

  Seja α^2= g1(t1) e $T^2 = T^1 - T^1t$

- Passo i:

  Selecionar o nó t ϵ Ti que minimiza a expressão:

  gi(t) = R(t) - R(Tit) / |f(Tit)| - 1

  Seja ti este nó

  Seja αi+1= gi(ti) e Ti+1 = Ti - Titi

- Output

  Uma sequência de árvores T1$\supseteq$T2$\supseteq$ ... $\supseteq$Tk$\supseteq$ ...$\supseteq$ raiz

  Uma sequência de parâmetros α1 < α2 ... < αk <

- Escolhendo α

  Usando o cross-validation

  É o parâmetro que minimiza o erro de validação

O parâmetro para essa técnica de poda é o ccp_alpha, quanto maiores os valores desse mais folhas serão podadas. Esse método permite avaliar os *links* mais frágeis na árvore de decisão, de modo que esses últimos são eliminados primeiro ao passo que se aumenta o alpha. Partindo desse princípio, analisaremos até que pondo o aumento do alpha realmente retira os elos frágeis da árvore. O seguinte código foi usado para esta análise:

```
print("Accuracy:", metrics.accuracy_score(Y_test, Y_pred))

clf = DecisionTreeClassifier(random_state=0)
path = clf.cost_complexity_pruning_path(X_train, Y_train)
ccp_alphas, impurities = path.ccp_alphas, path.impurities
fig, ax = plt.subplots()
ax.plot(ccp_alphas[:-1], impurities[:-1], marker='o', drawstyle="steps-post")
ax.set_xlabel("Alpha efetivo")
ax.set_ylabel("Impureza total das folhas")
ax.set_title("Impureza total vs Alpha efetivo")
```

FIGURA 3.12 – .Código da análise Impureza total VS Alpha efetivo

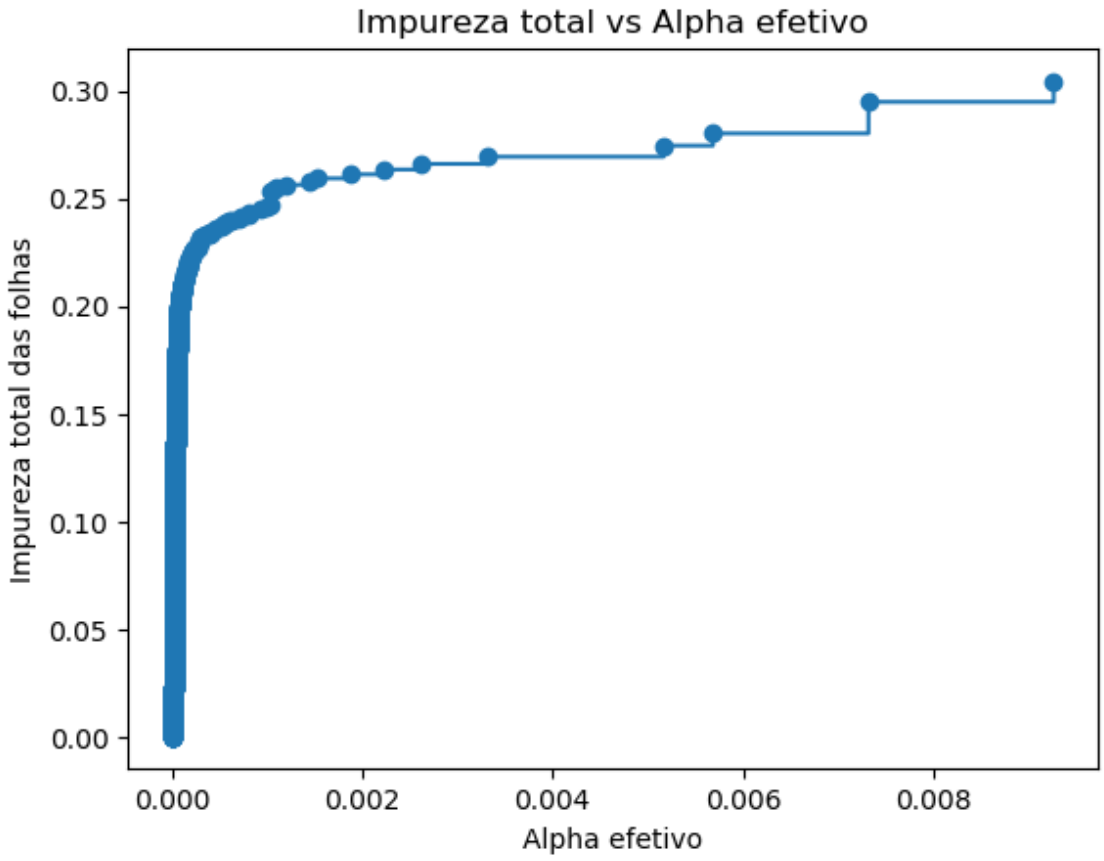


FIGURA 3.13 – Impureza total VS Alpha efetivo

Após calculado o Alpha efetivo, foi feita uma análise entre Número de nós x Alpha e Profundidade x Alpha, seguindo o código a seguir:

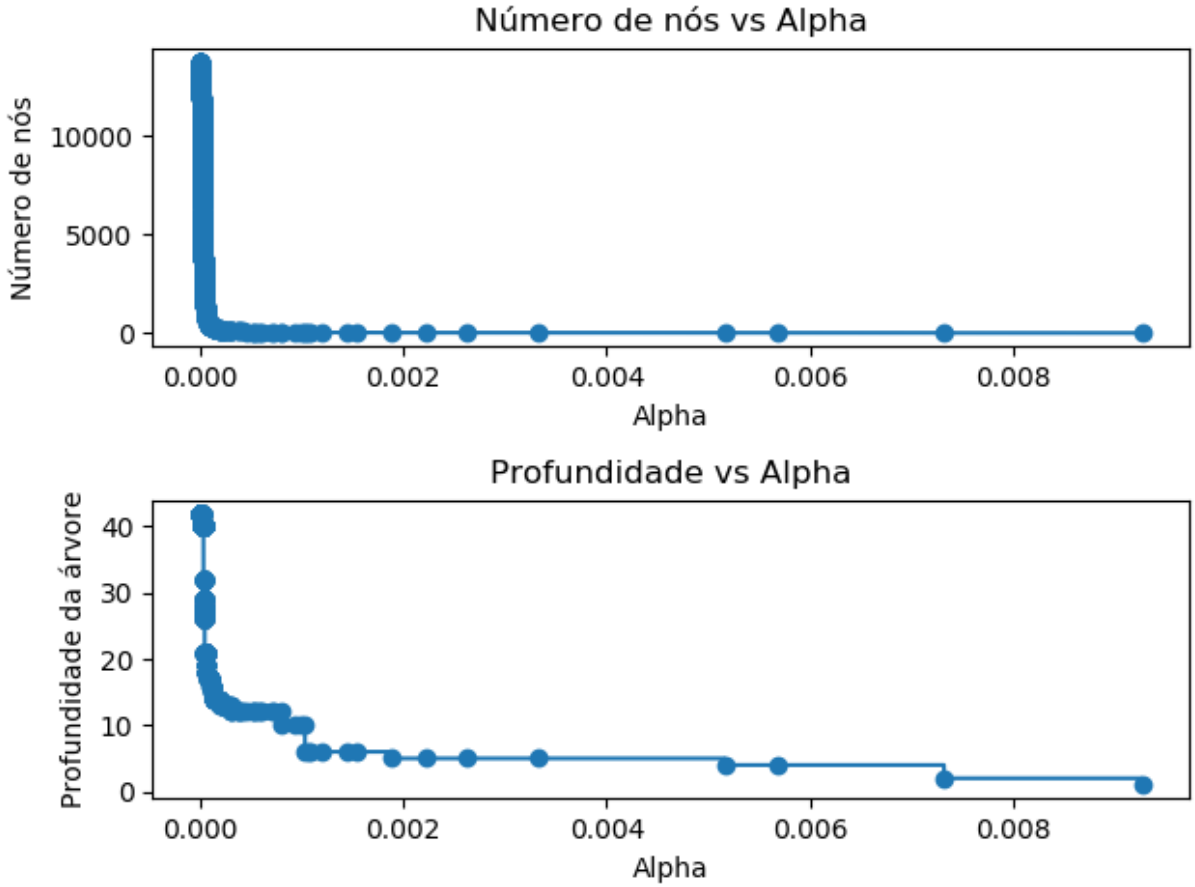


FIGURA 3.14 – Análises para o Alpha efetivo

A próxima análise foi feita avaliando a variação da acurácia para a árvore aplicada aos dados de treino e de teste, varaindo o valor de alpha.

```
train_scores = [clf.score(X_train, Y_train) for clf in clfs]
test_scores = [clf.score(X_test, Y_test) for clf in clfs]

fig, ax = plt.subplots()
ax.set_xlabel("Alpha")
ax.set_ylabel("Acurácia")
ax.set_title("Acurácia vs Alpha para os dados de treino e testes")
ax.plot(ccp_alphas, train_scores, marker='o', label="Treino",
        drawstyle="steps-post")
ax.plot(ccp_alphas, test_scores, marker='o', label="Testes",
        drawstyle="steps-post")
ax.legend()
plt.show()
```

FIGURA 3.15 – Código para Acurácia para variações de Alpha, para árvore aplicada aos dados de treino e aos de teste

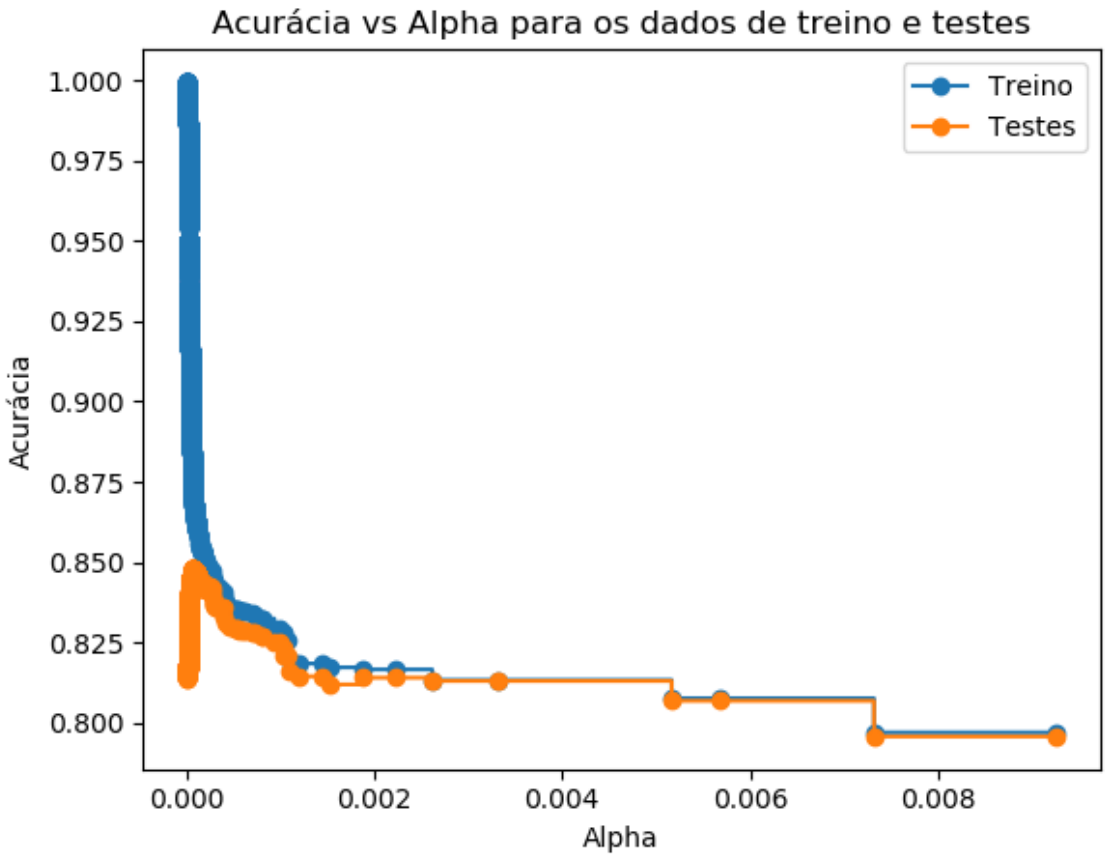


FIGURA 3.16 – Acurácia para variações de Alpha, para árvore aplicada aos dados de treino e aos de teste

4 Discussão

4.1 Análise das variáveis

- Magnitude de tempo em solo Essa variável foi selecionada, pois é esperado que quanto mais tempo a aeronave passe em trânsito maior a chance dela atrasar, dado que vôos que estão muito além do desvio padrão com relação a média,para os cenários superiores, provavelmente atrasam.

- *Taxi in* e *Taxi out* adicionais Essa variável está diretamente relacionada com o atraso, já que os intervalos de tempos adicionais culminam no atraso do voo, considerando que a velocidade de cruzeiro permanece a mesma com tempos adicionais ou não.

- Orvalho e Visibilidade As variáveis climáticas são estáveis e amenas para a maioria dos aeroportos a maior parte do ano. Ainda assim, essas duas variáveis foram consideradas para avaliar se a visibilidade tanto quantificada quanto pela presença ou não de orvalho geram atraso nos voos em questão

- Variáveis de demanda Dado o contexto da pandemia de sars-cov-19, as variáveis de demanda foram incluídas considerando que as medidas sanitárias tomadas poderiam se tornar relevantes para um cenário de atraso.

4.2 Análise da árvore de decisão

4.2.1 Impureza total x Alpha

Avaliando ,primeiramente, a impureza total presente na árvore resultante da pesquisa, percebe-se que o valor da impureza aumenta sensivelmente com o Aplha para valores bem próximo de zero, o que indica uma porção considerável de elos frágeis para a estrtutura da árvore de decisão para variação de Alpha nessa região.

4.2.2 Número de nós e Profundidade em função do Alpha

Seguindo a análise de impureza total, a quebra dos elos frágeis resulta em uma variação brusca do número de nós e da profundidade para valores de Alpha próximos de zero. Após a queda abrupta, o número de nós fica mais estável, entretanto, é interessante notar que para a profundidade da árvore ainda há uma queda considerável próximo a alpha=0.01, o que mostra a existência de mais alguns elos frágeis antes de se atingir a estabilidade.

4.2.3 Acurácia x Alpha, um comparativo para a árvore de decisão aplicada a base de treino e a de teste

Seguindo o padrão das análises anteriores, a variação de acurácia para a árvore aplicada a base de treino mostrou-se mais sensível para valores de alpha próximos de zero. Após essa disparidade inicial, ambas as curvas passam a ter comportamentos semelhantes. Desse modo, para termos uma estrutura de árvore de decisão que seja mais abrangente e possua maior probabilidade de sucesso com uma amostra aleatória, seria interessante adotar-se um valor de alpha próximo a 0.001, o que foi feito utilizando o código a seguir:

```
tree = DecisionTreeClassifier(ccp_alpha=0.001, random_state = 40)
tree.fit(X_train,Y_train)
Y_train_pred = tree.predict(X_train)
Y_test_pred = tree.predict(X_test)
print(accuracy_score(Y_train, Y_train_pred)), accuracy_score(Y_test, Y_test_pred)
```

FIGURA 4.1 – Código para árvore de decisão mais abrandente.

5 Conclusão

Avaliando a quantidade de atrasos para os aeroportos chegamos ao seguinte gráfico:

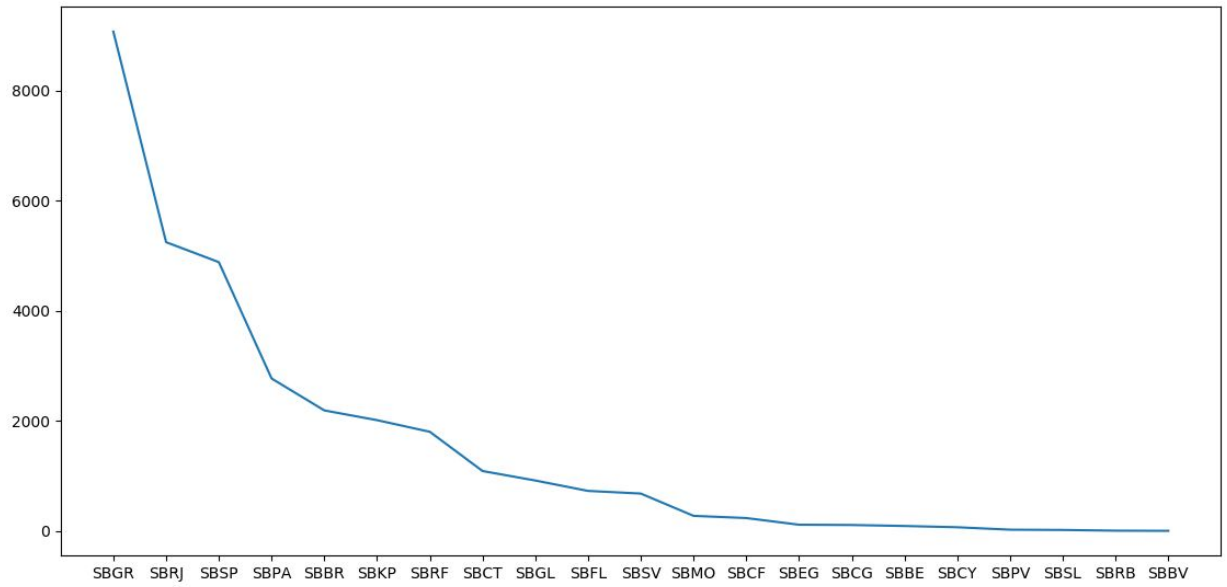


FIGURA 5.1 – Número de atrasos por aeroporto

Nota-se que para o período avaliado, a maioria dos atrasos ocorreu nos aeroportos de Guarulhos, Santos Dumont,Congonhas, Porto Alegre e Brasília, como retratado no gráfico a seguir para a porcentagem de voos atrasados:

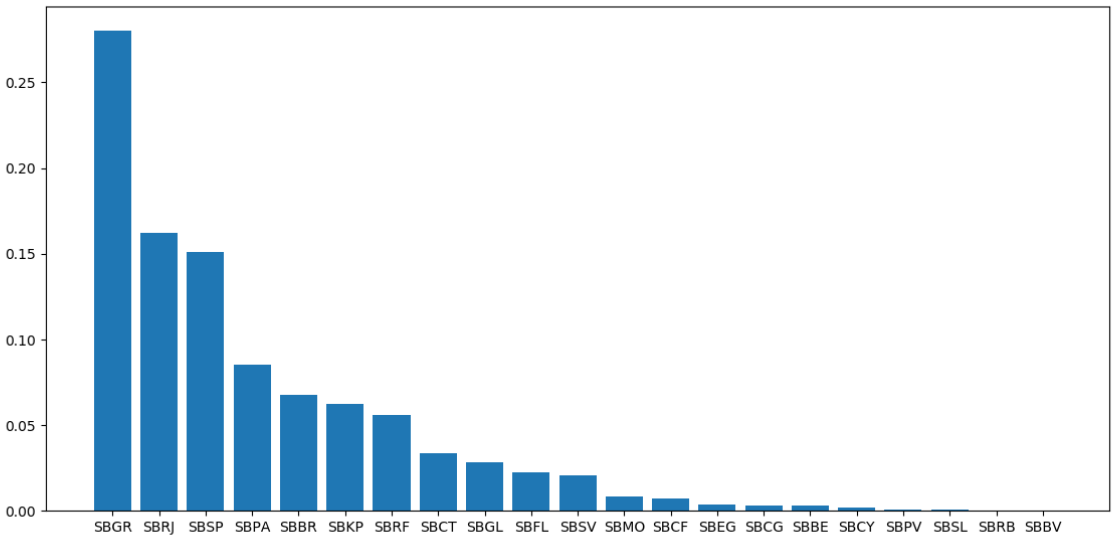


FIGURA 5.2 – Porcentagem de atrasos por aeroporto

Ao se avaliar os meses que possuíram maior número de atrasos, chegamos ao seguinte gráfico:

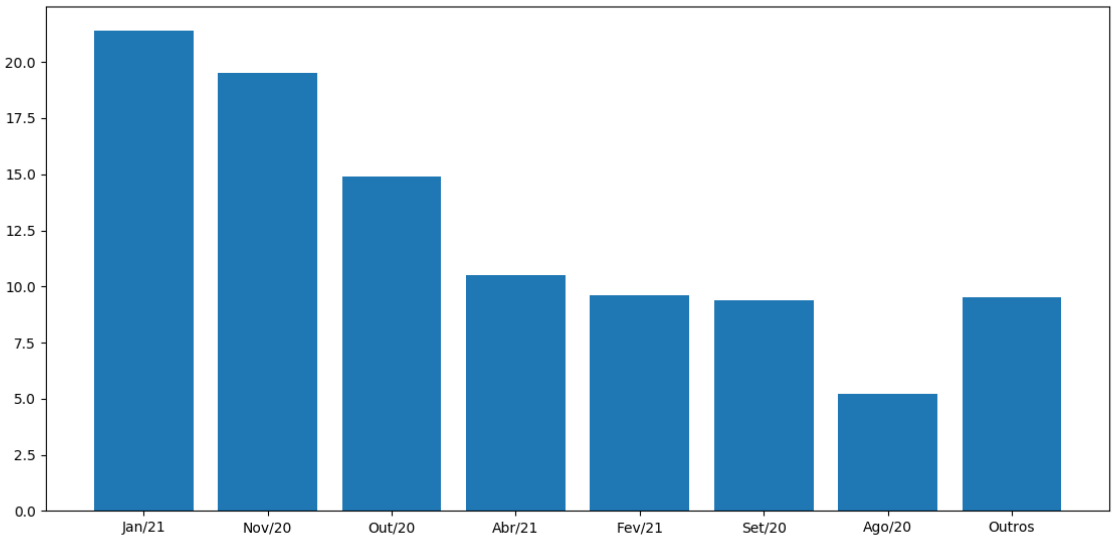


FIGURA 5.3 – Porcentagem de atrasos para os meses em análise

É interessante notar que o período do fim do ano de 2020 e começo do ano de 2021, que é marcado pelas comemorações festivas de Natal e *Réveillon*, foram os que tiveram maior número de atrasos, o que corresponde as expectativas, haja vista que é um janela de muita demanda por voos e a oferta ainda estava afetada pela pandemia de covid-19.

Fazendo a análise para os horários de atraso, tem-se:

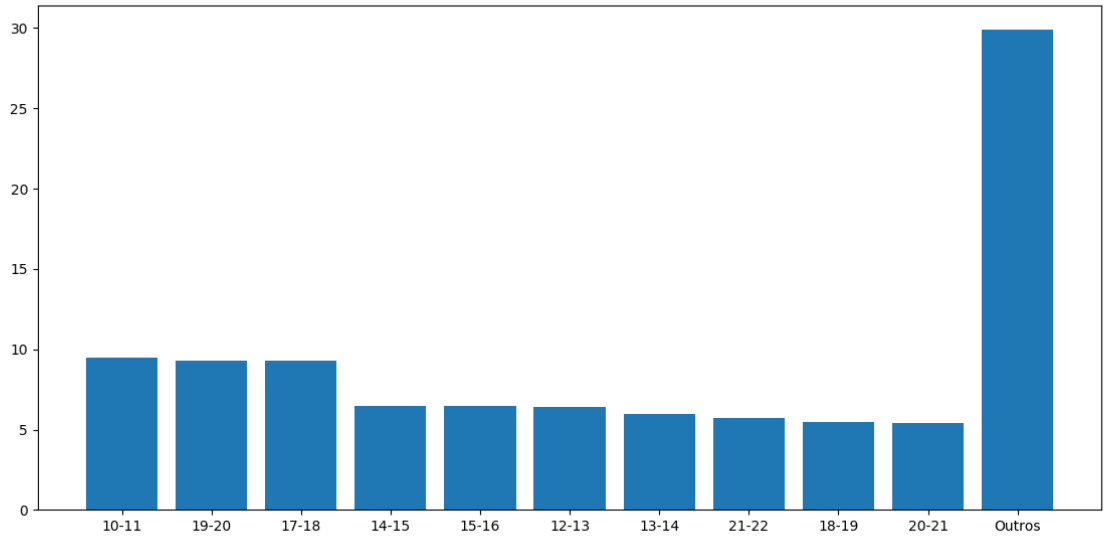


FIGURA 5.4 – Porcentagem de atrasos para intervalor em horas ao longo do dia

Observa-se que, diferente da avaliação para os meses, os atrasos não possuem um padrão bem definido ao longo das horas do dia.

Chegando ao fim da pesquisa, constata-se que o processo de obtenção e tratamento da base de dados foi bem sucedido, possibilitando dados suficientes para que fosse montada uma estrutura de árvore de decisão razoável. Assim aconteceu com as variáveis classificatórias, que possibilitaram uma boa distinção das variáveis de entrada, sendo a árvore que possuiu resultado mais abrangente quando aplicada tanto para base de treino, quanto para a base de teste, possui uma acurácia em sua classificação de 82%, em que foi considerado Alpha de 0.001.

Para uma abordagem futura, seria possível a abordagem de novas variáveis qualificatórias e de outros métodos de poda da árvore de decisão, assim como foi feito no artigo (MINGUERS, 1989). Além disso, outros métodos de classificação poderiam ser utilizados. Métodos como *Random Forrest* ou *Gradient boosted tree* são alternativas próximas a utilizada nessa pesquisa e também podem gerar bons resultados.

Referências

AGENCIALBRASIL. 2020. Disponível em: <https://agenciabrasil.ebc.com.br/saude-/noticia/2021-02/primeiro-caso-de-covid-19-no-brasil-completa-um-ano>.

ANAC. 2020. Disponível em: <https://www.anac.gov.br/noticias/2021/com-pandemia-indicadores-do-setor-aereo-reduzem-50-em-2020-1>.

BBC. 2020. Disponível em: <https://www.bbc.com/portuguese/internacional-51369300>.

BREIMAN L., F. J. O. R.; STONE, C. **Classification and Regression Trees**. [S.l.: s.n.], 1984.

FLIGHAWARE. 2019. Disponível em: <https://pt.flightaware.com/>.

FLIGHRADAR. 2019. Disponível em: <https://www.flightradar24.com/>.

IOWA. 2020. Disponível em: <https://mesonet.agron.iastate.edu/request/download.phtml?network=BR_ASOS>.

MINGUERS, J. **An Empirical Comparison of Pruning Methods for Decision Tree Induction**. — School of Industrial and Business Studies, University of Warwick, England, 1989.

OLIVEIRA, M. de. **A data-driven approach for air traffic operational performance characterization and prediction**. — Instituto Tecnológico de Aeronáutica, Brasil, 2021.

OURWORLD. 2020. Disponível em: <https://ourworldindata.org/coronavirus>.

SHUMELI, G. **Data Mining for Business Analytics**. [S.l.: s.n.], 2018.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. REGISTRO N°	4. N° DE PÁGINAS
TC	27 de outubro de 2022	DCTA/ITA/TC-155/2021	54

5. TÍTULO E SUBTÍTULO:

Análise preditiva de atraso nos principais aeroportos brasileiros.

6. AUTOR(ES):

Ícaro de Almeida Varão

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Árvore de decisão; *Scikit* ; *Pyhton*

9.PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Transporte aéreo; Atraso; Voo; Árvores de decisão; Análise de fatores; Aeroportos; Infraestrutura (transporte); Transportes.

10. APRESENTAÇÃO: **(X) Nacional ( ) Internacional**

ITA, São José dos Campos. Curso de Graduação em Engenharia Civil-Aeronáutica. Orientador: Prof. Dr. Marcelo Xavier Guterres; coorientador: Prof. Dr. Alessandro Vinícius Marques de Oliveira. Publicado em 2021.

11. RESUMO:

A pesquisa tem como objetivo a análise preditiva de atrasos nos 29 principais aeroportos brasileiros, utilizando uma árvore de decisão. No cenário de crise gerada pela pandemia de Sars-Cov-19, foi avaliado o impacto no setor aeroportuário e a influência na pontualidade dos voos. Desse modo, era esperado que as medidas sanitárias, dentro dos aeroportos em análise, impactassem negativamente na pontualidade desses voos. Foram utilizadas bases de dados com registros dos horários da movimentação das aeronaves durante o percurso de pouso e decolagem, foram utilizadas bases com registros horários das condições climáticas em cada um dos aeroportos da pesquisa e foram utilizadas bases que possuíam as características de demanda de todos os voos que ocorreram nesses aeroportos. Com essas, foram definidas variáveis classificatórias para a aplicação como *input* no algoritmo da árvore de decisão. Para a realização da análise por meio da árvore de decisão, foi utilizada a biblioteca scikit do python a fim de agilizar a pesquisa e possuir boas ferramentas de avaliação da árvore final. Para evitar overfitting ao longo do treinamento da árvore de decisão, foram feitas análises para o processo de poda dessa, utilizando o algoritmo de poda por custo de complexidade, em que se usou o parâmetro Alpha para determinar os elos mais frágeis da árvore obtida inicialmente, e qual o comportamento dela quando aplicada para as bases de treino e de teste vaiando o valor do parâmetro Alpha.

12. GRAU DE SIGILO:

(X) OSTENSIVO ( ) RESERVADO ( ) SECRETO