INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Ygor Rodrigo de Melo Fontes Santos

ANÁLISE E PREVISÃO DE ATRASOS EM VOOS NO SISTEMA DE TRANSPORTE AÉREO DO AEROPORTO INTERNACIONAL DE GUARULHOS

Trabalho de Graduação
2022

Curso de Engenharia Civil-Aeronáutica

CDU 656.71

Ygor Rodrigo de Melo Fontes Santos

ANÁLISE E PREVISÃO DE ATRASOS EM VOOS NO SISTEMA DE TRANSPORTE AÉREO DO AEROPORTO INTERNACIONAL DE GUARULHOS

Orientador

Prof. Dr. Alessandro Vinícius Marques de Oliveira (ITA)

ENGENHARIA CIVIL-AERONÁUTICA

SÃO JOSÉ DOS CAMPOS
INSTITUTO TECNOLÓGICO DE AERONÁUTICA

2022

Dados Internacionais de Catalogação-na-Publicação (CIP)

Divisão de Informação e Documentação

Santos, Ygor Rodrigo de Melo Fontes
Análise e previsão de atrasos em voos no sistema de transporte aéreo do Aeroporto Internacional de Guarulhos / Ygor Rodrigo de Melo Fontes Santos
São José dos Campos, 2022.
104f.

Trabalho de Graduação – Engenharia Civil-Aeronáutica – Instituto Tecnológico de Aeronáutica, 2022. Orientador: Prof. Dr. Alessandro Vinícius Marques de Oliveira.

1. Transporte aéreo. 2. Atraso. 3. Voo. 4. Árvore de decisão. 5. Aprendizagem (inteligência artificial). 6. Operações de linha aéreas. 7. Análise de fatores. 8. Transportes. I. Instituto Tecnológico de Aeronáutica. II.Análise e previsão de atrasos em voos no sistema de transporte aéreo do Aeroporto Internacional de Guarulhos.

REFERÊNCIA BIBLIOGRÁFICA

SANTOS, Ygor Rodrigo de Melo Fontes. **Análise e previsão de atrasos em voos no sistema de transporte aéreo do Aeroporto Internacional de Guarulhos.** 2022. 104f. Trabalho de Conclusão de Curso. (Graduação em Engenharia Civil-Aeronáutica) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Ygor Rodrigo de Melo Fontes Santos
TÍTULO DO TRABALHO: Análise e previsão de atrasos em voos no sistema de transporte aéreo do Aeroporto Internacional de Guarulhos.
TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2022

Ygor Rodrigo de Melo Fontes Santos
Rua H8E, 105
12.228-461 – São José dos Campos – SP

ANÁLISE E PREVISÃO DE ATRASOS EM VOOS NO SISTEMA DE TRANSPORTE AÉREO DO AEROPORTO INTERNACIONAL DE GUARULHOS

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

Ygor Rodrigo de Melo Fontes Santos

Autor

Prof. Dr. Alessandro Vinícius Marques de Oliveira (ITA)

Orientador

Prof. Dr. Evandro José da Silva

Coordenador do Curso de Engenharia Civil-Aeronáutica

São José dos Campos, 23 de novembro de 2022

Resumo

Atrasos em voos são inevitáveis e causam muitos prejuízos às companhias aéreas e à sociedade como um todo. Além do impacto na economia, eles exercem influência sobre a satisfação dos consumidores. Diante desse contexto, o presente estudo faz uma análise dos atrasos em voos do aeroporto mais movimentado do Brasil, o Aeroporto Internacional de Guarulhos. Assim, pelo motivo referido, esse aeroporto representa o que mais sofre em virtude da concentração e do congestionamento por atraso de voos. Nesse aspecto, a motivação da pesquisa se justifica pela relevância do Aeroporto assinalado. Além disso, o problema de pesquisa a ser desenvovido é o seguinte: Qual algoritmo de *machine learning* tem melhor desempenho ao prever atrasos de voo no aeroporto de Guarulhos?. Ademais, com relação à hipótese de pesquisa, configurou-se a seguinte hipótese: O algoritmo de *machine learning* que apresenta melhor desempenho ao prever atrasos de voo no aeroporto de Guarulhos é o de redes neurais artificiais. Dessa forma, este estudo utiliza 4 algoritmos de machine learning, quais sejam, KNN, árvore de decisão, regressão logística e rede neural artificial, com o intuito de elaborar um modelo preditivo para os atrasos no aeroporto em análise. Nesse sentido, os algoritmos são todos aplicados a 2 modelos de classificação e 1 de regressão. Quanto às variáveis eleitas para a pesquisa, são consideradas as variáveis referentes à operação dos voos e as variáveis metereológicas. Enfim, com relação à conclusão da pesquisa, compreendeu-se que tanto os modelos de classificação, quanto o modelo de regressão propostos apresentaram erros bastante elevados. Por fim, os resultados preliminares levaram ao teste de dois estudos de caso, a fim de verificar uma aplicação verossímel dos algoritmos analisados, mitigando-se as condições preliminarmente estabelecidas.

Abstract

Flight delays are inevitable and cause a lot of damage to airlines and society as a whole. Besides the impact on the economy, they have an influence on consumer satisfaction. Given this context, this study analyzes flight delays at the busiest airport in Brazil, Guarulhos International Airport. Thus, for the aforementioned reason, this airport represents the one that suffers the most from concentration and congestion due to delayed flights. In this aspect, the motivation of the research is justified by the relevance of the Airport pointed out. Furthermore, the research problem to be developed is as follows: Which machine learning algorithm performs best in predicting flight delays at Guarulhos Airport? Furthermore, regarding the research hypothesis, the following hypothesis was configured: The machine learning algorithm which presents the best performance when predicting flight delays at the Guarulhos airport is the artificial neural network algorithm. Thus, this study uses 4 machine learning algorithms, namely, KNN, decision tree, logistic regression and artificial neural network, in order to develop a predictive model for delays at the airport under analysis. In this sense, the algorithms are all applied to 2 classification models and 1 regression model. As for the variables chosen for the research, the variables related to the operation of flights and the meteorological variables are considered. Finally, regarding the conclusion of the research, it is understood that both the classification models and the proposed regression model presented quite high errors. Finally, the preliminary results led to the testing of two case studies in order to verify a credible application of the algorithms analyzed, mitigating the conditions preliminarily established.

Lista de Figuras

FIGURA 1 – Principais fatores que influenciam na OTP de companhias aéreas - Fonte: (WU, 2005, p. 274).....13
FIGURA 2 – Esquema ilustrativo dos algoritmos de aprendizado de máquina supervisionado - Fonte: (MCDONALD, 2021).....18
FIGURA 3 – Os processos de aprendizado de máquina de classificação buscam encontrar fronteiras – lineares ou não – entre os dados - Fonte: (SULLIVAN, 2018).....19
FIGURA 4 –Situação dos voos da base VRA da ANAC.....21
FIGURA 5 – Classificação das rotas dos voos da base VRA da ANAC.....22
FIGURA 6 – Aeródromos brasileiros contemplados pela base de dados da Universidade Estadual de Iowa.....23
FIGURA 7 – Porcentagem dos voos operados pelas companhias aéreas.....26
FIGURA 8 – Gráfico das taxas de atraso de acordo com as classes de número de assentos....31
FIGURA 9 – Aeroportos de origem dos voos da base de dados da ANAC.....34
FIGURA 10 – Gráfico das taxas de atraso das classes de distância.....39
FIGURA 11 – Gráfico das taxas de atraso das classes de *sch_d*.....40
FIGURA 12 – Gráfico das taxas de atraso das classes de *sch_h*.....41
FIGURA 13 – Número de voos operados (em porcentagem) de acordo com o horário previsto de chegada.....43
FIGURA 14 – Gráfico das taxas de atraso ao longo dos horários do dia.....43
FIGURA 15 – Número de voos operados (em porcentagem) ao longo dos dias da semana....44
FIGURA 16 – Taxas de atraso dos dias da semana.....44
FIGURA 17 – Número de voos operados (em porcentagem) ao longo dos dias do mês.....46
FIGURA 18 – Taxas de atraso ao longo dos dias do mês.....46
FIGURA 19 – Número de voos operados (em porcentagem) ao longo dos meses do ano.....47
FIGURA 20 – Taxas de atraso de acordo com as classes de temperatura.....50
FIGURA 21 – Taxas de atraso de acordo com as classes de pressão.....52
FIGURA 22 – Taxas de atraso de acordo com a classe de visibilidade horizontal.....54
FIGURA 23 – Taxas de atraso de acordo com a classe de visibilidade vertical.....55
FIGURA 24 – Número de voos (em porcentagem) de acordo com a velocidade do vento.....57
FIGURA 25 – Taxas de atraso de acordo com a velocidade do vento.....58
FIGURA 26 – Número de voos (em porcentagem) de acordo com a direção do vento.....59
FIGURA 27 – Taxas de atraso de acordo com a direção do vento.....59

FIGURA 28 – As árvores são construídas por meio de partição recursiva.....62
FIGURA 29 – Modelo de um neurônio artificial. Fonte: (MARTÍNEZ-ÁLVAREZ et al., 2015).....63
FIGURA 30 – Ilustração das camadas de uma rede neural artificial.....64
FIGURA 31 – Matriz confusão de um problema de classificação binário.....65
FIGURA 32 – Matriz confusão utilizada no cálculo da sensibilidade e da precisão.....65
FIGURA 33 – Acurácia dos modelos da validação cruzada.....70
FIGURA 34 – *F1 score* dos modelos da validação cruzada.....71
FIGURA 35 – Acurácia das árvores de decisão elaboradas.....72
FIGURA 36 – Valores de *F1 score* das árvores de decisão elaboradas.....72
FIGURA 37 – Árvore com maior acurácia.....73
FIGURA 38 – Coeficiente e estatística t de Student das variáveis de input.....75
FIGURA 39 – Acurácia das redes neurais avaliadas.....76
FIGURA 40 – *F1 score* das redes neurais avaliadas.....77
FIGURA 41 – Rede neural selecionada a partir da validação cruzada.....77
FIGURA 42 – Acurácia dos modelos da validação cruzada.....79
FIGURA 43 – F1 score dos modelos da validação cruzada.....79
FIGURA 44 – Acurácia das árvores de decisão da validação cruzada.....80
FIGURA 45 – *F1 score* das árvores de decisão da validação cruzada.....81
FIGURA 46 – Árvore de decisão com melhor acurácia.....81
FIGURA 47 – Árvore com apenas um nó.....82
FIGURA 48 – Coeficiente e estatística t de Student das variáveis de input.....84
FIGURA 49 – Acurácia das redes neurais da validação cruzada.....85
FIGURA 50 – *F1 score* das redes neurais da validação cruzada.....86
FIGURA 51 – Valor de MAE dos modelos avaliados pela validação cruzada.....88
FIGURA 52 – Valor de RMSE dos modelos avaliados pela validação cruzada.....88
FIGURA 53 – Erro das árvores de decisão avaliado em relação aos dados de validação.....89
FIGURA 54 – Coeficientes e valores da estatística t de Student das variáveis de input.....91
FIGURA 55 – MAE e RMSE das redes neurais da validação cruzada.....92
FIGURA 56 – Rota aérea Porto Alegre – Guarulhos.....96
FIGURA 57 – Modelo de classificação de melhor acurácia.....98
FIGURA 58 – Árvore de decisão com apenas um nó.....98

Lista de Tabelas

TABELA 1 – Impacto financeiro dos atrasos na economia americana em 2007 – Fonte: NEXTOR, 2010 14
TABELA 2 – Número de passageiros afetados por cancelamentos ou atrasos de voos em 2021 - Fonte: Redação Guarulhos Hoje (2022) 15
TABELA 3 – Descrições dos tipos de linha apresentados na base de dados da ANAC 21
TABELA 4 – Frequência relativa das classes de atraso na chegada 25
TABELA 5 – Frequência relativa das classes de atraso na partida 25
TABELA 6 – Companhias aéreas que operaram os voos analisados 25
TABELA 7 – Taxas de atraso das companhias aéreas 27
TABELA 8 – Número de voos operados (em porcentagem) nos diversos modelos de aeronave 28
TABELA 9 – Taxas de atraso dos diversos modelos de aeronave 29
TABELA 10 – Número de voos (em porcentagem) de cada classe de número de assentos 30
TABELA 11 – Aeroportos de origem dos voos analisados 32
TABELA 12 – Taxas de atraso dos aeroportos de origem dos voos analisados 35
TABELA 13 – Número de voos (em porcentagem) de cada uma das classes de distância 38
TABELA 14 – Número de voos operados (em porcentagem) em cada uma das classes de *sch_d* 39
TABELA 15 – Número de voos (em porcentagem) das classes de *sch_h* 40
TABELA 16 – Taxas de atraso ao longo dos meses do ano 47
TABELA 17 – Número de voos, número de voos atrasados e taxa de atraso de cada ano 48
TABELA 18 – Número de voos de acordo com a classe de temperatura 49
TABELA 19 – Número de voos de acordo com a classe de pressão 50
TABELA 20 – Tipos de precipitação decodificados a partir do METAR 52
TABELA 21 – Número de voos (em porcentagem) de acordo com a classe de visibilidade horizontal 52
TABELA 22 – Número de voos (em porcentagem) de acordo com as classes de visibilidade vertical 54
TABELA 23 – Modelos propostos 66
TABELA 24 – Partição dos dados para a validação cruzada 68
TABELA 25 – Matriz confusão do algoritmo KNN 70
TABELA 26 – Métricas que buscam determinar a qualidade do modelo 70

TABELA 27– Matriz confusão da árvore com maior acurácia 72
TABELA 28 – Métricas que buscam determinar a qualidade da árvore de decisão 72
TABELA 29 – Valores de acurácia e de *F1 score* dos modelos analisados 73
TABELA 30 – Matriz confusão do modelo de regressão logística 74
TABELA 31 – Métricas que determinam a qualidade do modelo de regressão logística 74
TABELA 32 – Matriz confusão do modelo de rede neural selecionado 76
TABELA 33 – Métricas que determinam a qualidade geral do modelo selecionado 77
TABELA 34 – Matriz confusão do algoritmo KNN 78
TABELA 35 – Métricas da qualidade do modelo selecionado 79
TABELA 36 – Matriz confusão do modelo com maior acurácia 81
TABELA 37 – Medidas da árvore com 1 nó e da árvore com 36 nós 81
TABELA 38 – Valores de acurácia e de *F1 score* dos modelos analisados 82
TABELA 39 – Matriz confusão do modelo e regressão logística com família binomial 83
TABELA 40 – Métricas que determinam a qualidade do modelo de regressão logística 84
TABELA 41 – Matriz confusão da rede neural selecionada 86
TABELA 42 – Métricas que determinam a qualidade geral do modelo 86
TABELA 43 – MAE e RMSE do modelo com menor erro, avaliados em relação aos dados de teste 87
TABELA 44 – MAE e RMSE da árvore selecionada, avaliados em relação aos dados e teste 88
TABELA 45 – MAE e RMSE dos modelos analisados em relação aos dados de validação 89
TABELA 46 – Valores de MAE e de RMSE do modelo gaussiano em relação aos dados de teste 89
TABELA 47 – MAE e RMSE do modelo selecionado a partir da validação cruzada 92
TABELA 48 – Resultados do modelo 1 92
TABELA 49 – Resultados do modelo 2 92
TABELA 50 – Resultados do modelo 3 93
TABELA 51 – Métricas de qualidade do modelo com melhor acurácia e do modelo de árvore com apenas um nó 93
TABELA 52 – Resultados obtidos para o modelo 1 95
TABELA 53 – Resultados obtidos para o modelo 2 96
TABELA 54 – Resultados obtidos para o modelo 3 96
TABELA 55 – Resultados obtidos para o caso 2 99

Lista de Abreviaturas e Siglas

ANAC	Agência Nacional de Aviação Civil
FAA	*Federal Aviation Administration*
METAR	*Meteorological Aerodrome Report*
NAS	*National Aviation System*
NEXTOR	*National Center of Excellence for Aviation Operations Research*
OTP	*On-Time Performance*

Sumário

1 INTRODUÇÃO.....13
1.1 Relevância do tema.....15
1.2 Delimitação do problema de pesquisa e formulação da hipótese.....15
1.3 Objetivo principal da pesquisa.....16
1.4 Objetivos específicos da pesquisa.....17
2 REVISÃO BIBLIOGRÁFICA.....17
3 METODOLOGIA.....18
3.1 Coleta de Dados.....20
3.2 Análise das Condições Concomitantes dos Pousos dos Voos.....24
3.3 Variáveis referentes à operação dos voos.....24
3.3.1 Atraso.....24
3.3.2 Companhia aérea.....28
3.3.3 Modelo de aeronave.....28
3.3.4 Número de assentos.....30
3.3.5 Aeroporto de origem.....32
3.3.6 Distância.....38
3.3.7 Variáveis de congestionamento aéreo.....39
3.3.8 Horário previsto de chegada..... 42
3.3.9 Dia da semana.....44
3.3.10 Dia do mês.....45
3.3.11 Mês.....46
3.3.12 Ano.....48
3.4 Variáveis metereológicas.....49
3.4.1 Temperatura.....49
3.4.2 Pressão.....50
3.4.3 Precipitação.....52
3.4.4 Visibilidade horizontal.....52
3.4.5 Visibilidade vertical.....54
3.4.6 Vento.....55
3.4.7 Velocidade do vento.....56
3.4.8 Rajadas de vento.....58

3.4.9 Direção do vento....58
3.5 Teste dos Algoritmos de *Machine Learning*....59
3.5.1 KNN....59
3.5.2 Árvore de decisão....61
3.5.3 Regressão logística....62
3.5.4 Redes neurais artificiais....63
3.6 Avaliação da Performance Preditiva: Problema de Classificação e Problema de Regressão....64
3.6.1 Modelos preditivos de atraso....66
4 RESULTADOS: TESTE DOS ALGORITMOS 69
4.1 Modelo 1 de classificação....70
4.1.1 KNN: Modelo 1 de classificação 70
4.1.2 Árvore de decisão: Modelo 1 de classificação....71
4.1.3 Regressão logística: Modelo 1 de classificação....74
4.1.4 Redes neurais artificiais: Modelo 1 de classificação....76
4.2 Modelo 2 de classificação....78
4.2.1 KNN: Modelo 2 de classificação 78
4.2.2 Árvore de decisão: Modelo 2 de classificação....80
4.2.3 Regressão logística: Modelo 2 de classificação....83
4.2.4 Redes neurais artificiais: Modelo 2 de classificação....85
4.3 Modelo de regressão testado em cada algoritmo....87
4.3.1 KNN: Modelo regressão....87
4.3.2 Árvore de decisão: Modelo de regressão....88
4.3.3 Regressão logística: Modelo de regressão....89
4.3.4 Redes neurais artificiais: Modelo de regressão....91
5 CONCLUSÕES....93
5.1 Limitações da pesquisa e nova proposta: Breve estudo de caso....94
5.1.1 Caso 1....95
5.1.2 Caso 2....98
REFERÊNCIAS....101

1 Introdução

Atrasos em voos são inevitáveis e são um dos problemas mais difíceis do controle de aviação, exercendo uma forte influência sobre os lucros e os prejuízos das companhias aéreas. Nessa perspectiva, a *On-Time Performance* (OTP) é um método que visa a compreender a pontualidade de diferentes modais de transporte, não somente da aviação.

Voltando-se para a questão da OTP, ela provê um meio padronizado de comparar como uma companhia aérea opera de acordo com os seus horários programados em relação a outras companhias. Wu (2005, p. 274) examinou os principais fatores que influenciam na OTP das companhias aéreas, os quais se encontram esquematizados na figura a seguir:

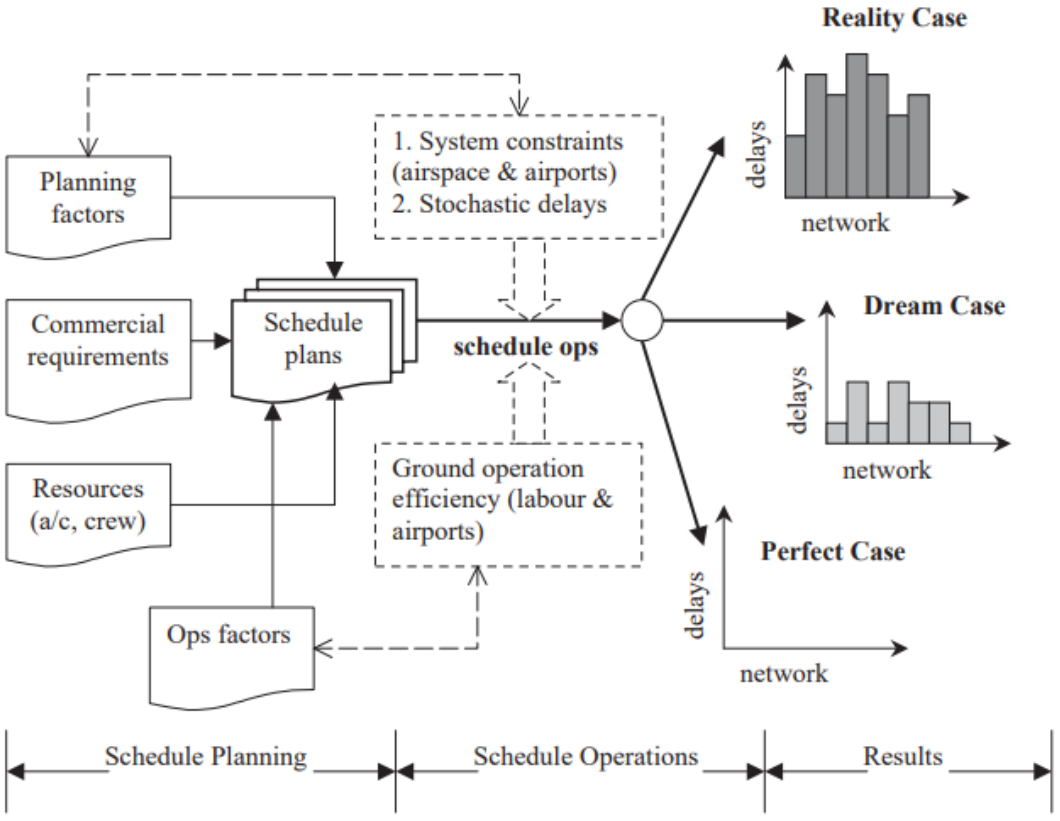

FIGURA 1 – Principais fatores que influenciam na OTP de companhias aéreas - Fonte: (WU, 2005, p. 274)

A OTP exerce papel fundamental na administração das operações das companhias aéreas e dos aeroportos, os quais podem utilizar essa medida para avaliar a sua eficiência e a qualidade do seu serviço. A correlação entre a OTP e a eficiência se deve ao fato de que atrasos afetam a produtividade e causam prejuízos consideráveis, influenciando a estrutura de

custos do setor de transporte aéreo (ON-TIME, 2022).

Nessa perspectiva, estudos indicam que os custos dos atrasos aéreos atingem bilhões de dólares e afetam não somente as companhias aéreas e os passageiros, mas a sociedade como um todo. Em 2010, um estudo foi comissionado pela *Federal Aviation Administration* (FAA) a respeito do impacto financeiro dos atrasos na economia americana no ano de 2007 (NATIONAL CENTER OF EXCELLENCE FOR AVIATION OPERATIONS RESEARCH (NEXTOR), 2010). O estudo, cujos resultados são apresentados na TABELA 1, concluiu que atrasos em voos causaram um prejuízo de 8,3 bilhões de dólares para as companhias aéreas e 16,7 bilhões para os passageiros, além de acarretarem um impacto indireto na economia de 4 bilhões e de causarem um custo de perda de demanda de 3,9 bilhões de dólares (GUY, 2010).

TABELA 1 – Impacto financeiro dos atrasos na economia americana em 2007 – Fonte: NEXTOR, 2010

Custo	Valor (em bilhões de dólares)
Custo para as companhias aéreas	8,3
Custo para os passageiros	16,7
Impacto indireto na economia	4,0
Custo da perda de demanda	3,9
Custo total	32,9

Além do impacto na economia, os atrasos influenciam a escolha da companhia aérea por parte dos consumidores (Gayle e Yimga, 2018). Ademais, de acordo com Efthymiou *et al.* (2018), em estudo que analisou a OTP da *British Airlines* no aeroporto de Heathrow, os atrasos exercem influência na satisfação dos consumidores. Nessa perspectiva, a OTP tem se tornado uma vantagem competitiva à medida em que as expectativas dos passageiros referentes à pontualidade das companhias têm aumentado nos últimos anos.

Dessa forma, eventos como condições meteorológicas adversas, congestionamentos aéreos e incidentes podem causar atrasos nos voos. Yu *et al.* (2019) examinaram o transporte aéreo na China e chegaram à conclusão de que a OTP é significativamente impactada pelas condições meteorológicas no país. Nos EUA, por exemplo, o tempo foi responsável, de acordo com Oliveira *et al.* (2021), por 57% dos atrasos no *National Aviation System* (NAS) em 2019.

Em virtude de seu impacto na economia, na decisão e na satisfação dos consumidores, muitos pesquisadores têm se dedicado a analisar os atrasos e a desenvolver estratégias com o

intuito de reduzi-los. Muitos estudos têm utilizado técnicas de *machine learning* para examinar a relação entre as condições meteorológicas e os atrasos aéreos, produzindo um conhecimento que pode ser utilizado para a tomada de decisões na administração de companhias aéreas e de aeroportos.

1.1 Relevância do tema

Nesse sentido, o foco deste estudo quanto à questão do atraso nos voos é o Aeroporto Internacional de Guarulhos, tendo em vista que, atualmente, ele representa o maior *hub* dos aeroportos brasileiros, sendo também o que mais sofre em virtude da concentração e do congestionamento por atraso de voos (SCARPEL; PELICIONI, 2018, p. 1). Assim, a motivação da pesquisa se justifica pela relevância do Aeroporto assinalado.

1.2 Delimitação do problema de pesquisa e formulação da hipótese

O cenário apresentado indica a necessidade de delimitação do problema de pesquisa. Nessa seara, a relevância de se analisarem os atrasos dos voos e os fatores que exercem impacto sobre tais atrasos, como as condições meteorológicas, provém do fato de que atrasos causam um impacto negativo na economia e na satisfação dos consumidores em relação ao serviço prestado pelas companhias aéreas e pelos aeroportos.

Assim, o problema de pesquisa a ser desenvovido é o seguinte: Qual algoritmo de *machine learning* tem melhor desempenho ao prever atrasos de voo no aeroporto de Guarulhos?. Dessa forma, a fim de demonstrar a escolha pela delimitação ao aeroporto de Guarulhos, a TABELA 2 apresenta os 10 aeroportos do país com os maiores números de passageiros afetados por cancelamentos ou atrasos de voos, de acordo com os resultados do levantamento realizado pela AirHelp.

TABELA 2 – Número de passageiros afetados por cancelamentos ou atrasos de voos em 2021 - Fonte: Redação Guarulhos Hoje (2022)

Aeroporto	Número de passageiros
Aeroporto Internacional de Guarulhos	988.676
Aeroporto Internacional de Viracopos	321.269
Aeroporto Internacional de Recife/	317.270

Guararapes	
Aeroporto do Rio de Janeiro / Santos Dumont	304.904
Aeroporto Internacional de Brasília	249.815
Aeroporto de São Paulo / Congonhas	246.413
Aeroporto Internacional de Belo Horizonte / Confins	235.128
Aeroporto Internacional de Porto Alegre	151.077
Aeroporto Internacional Tom Jobim / Rio de Janeiro	147.632
Aeroporto Internacional de Salvador	145.686

Então, de acordo com o levantamento referido, que analisou voos regulares brasileiros contidos em seu banco de dados global, o Aeroporto Internacional de Guarulhos foi o que mais teve passageiros afetados por cancelamentos e atrasos de voos no ano de 2021 no Brasil. Esse fato justifica a necessidade de um estudo a respeito dos atrasos e dos fatores que os influenciam no aeroporto supracitado.

Ademais, com relação à hipótese de pesquisa, cuja formulação depende, conforme Toledo, Flikkema e Toledo-Pereyra (2011), do fato de haver informações suficientes para testar o problema de pesquisa e de suas respectivas variáveis, é necessário aduzir que ela representa um tópico frasal que dialoga com o problema, caracterizando as questões colocadas e definindo as conclusões alcançadas.

Enfim, dada a concepção de hipótese assinalada, bem como a possibilidade de testar o problema de pesquisa, configurou-se a seguinte hipótese: O algoritmo de *machine learning* que apresenta melhor desempenho ao prever atrasos de voo no aeroporto de Guarulhos é o de redes neurais artificiais. Tal hipótese, pois, será testada, de modo que haverá na conclusão do trabalho a comparação entre os algoritmos utilizados, a fim de confirmar ou não a hipótese.

1.3 Objetivo principal da pesquisa

Diante desse cenário, o objetivo geral deste estudo é elaborar, por meio de técnicas de *machine learning*, um modelo matemático para a previsão de atrasos em voos domésticos no Aeroporto Internacional de Guarulhos, o maior e mais movimentado do país, conforme já apontado, e avaliar, dentre os algoritmos de *machine learning* testados, qual seria o que

apresenta melhor desempenho quanto à previsão de atrasos de voo no aeroporto em análise.

1.4 Objetivos específicos da pesquisa

Com a finalidade de conduzir o estudo para atingir o seu objetivo geral, é necessário subdividi-lo nos seguintes objetivos específicos:

- Coletar o histórico de dados dos voos que desembarcaram no aeroporto de Guarulhos no período de janeiro de 2019 até dezembro de 2021;
- Analisar as variáveis referentes às operações dos voos e as variáveis metereológicas;
- Utilizar 4 algoritmos de machine learning – KNN, árvore de decisão, regressão logística e rede neural artificial – a fim de elaborar um modelo preditivo para os atrasos no aeroporto em análise;
- Escolher o modelo mais adequado aos dados coletados e analisados.

2 Revisão Bibliográfica

Na literatura, há diversos estudos acerca da influência das condições meteorológicas nas operações aéreas. Pitfield e Jerrard (1999) utilizaram simulação de Monte Carlo para avaliar o impacto do tempo na capacidade da pista do Aeroporto Internacional de Roma – Fiumicino. Markovic *et al.* (2008) utilizaram um modelo de regressão para avaliar o impacto de diversas variáveis meteorológicas na OTP de decolagens e pousos no Aeroporto Internacional de Frankfurt. Santos *et al.* (2018) correlacionaram taxas pluviométricas com atrasos e cancelamentos no Aeroporto Internacional de Guarulhos. Por fim, Steinheimer (2019) chegou à conclusão em seu estudo de que o tempo meteorológico é a principal causa de atrasos nos voos que pousam no Aeroporto Internacional de Viena.

Outros estudos têm feito análises similares para múltiplos aeroportos. Borsky e Unterberger (2019) avaliaram o impacto do tempo em atrasos de voos nos 10 aeroportos mais movimentados dos EUA e chegaram à conclusão de que condições meteorológicas adversas aumentavam o atraso em 23 minutos em média. Além disso, um estudo realizado por OLIVEIRA *et al.* (2021) desenvolveu um modelo estatístico para avaliar o impacto de diversas variáveis meteorológicas nos atrasos em 79 aeródromos do Brasil. O estudo concluiu

que variáveis como visibilidade horizontal, visibilidade vertical, chuva e vento eram estatisticamente significativas no modelo de regressão desenvolvido.

Na literatura, há também estudos que visam a elaborar modelos preditivos para atrasos por meio de técnicas de *machine learning*. Priyanka (2018) desenvolveu um modelo utilizando algoritmo KNN para prever atrasos da companhia aérea Jetblue Airways com voos com origem no Aeroporto Internacional de Boston e destino no Aeroporto Internacional de Los Angeles. Zoutendijk e Mihaela Mitici (2021) utilizaram diversas variáveis – incluindo meteorológicas – para o desenvolvimento de um modelo preditivo de atrasos em voos no Aeroporto de Roterdã-Haia. Alonso e Loureiro (2015) utilizaram uma rede neural artificial para elaborar um modelo preditivo de atrasos no Aeroporto de Porto. Khanmohammadi, Tutun e Kucuk (2016) utilizaram uma rede neural com diversas camadas para elaborar um modelo preditivo de atrasos no Aeroporto Internacional de John F. Kennedy.

Nessa perspectiva, este trabalho se soma à literatura já existente, ao tentar elaborar um modelo matemático para prever atrasos em voos no aeroporto em análise.

3 Metodologia

Neste trabalho, para a elaboração dos modelos de previsão de atraso para o Aeroporto Internacional de Guarulhos, foram utilizados algoritmos de aprendizado de máquina supervisionado. Dadas n observações $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, onde x denota um *input* e y denota um *output*, o objetivo desses algoritmos é aprender uma função de mapeamento f que preveja y para um novo exemplo x, de acordo com a FIGURA 2.

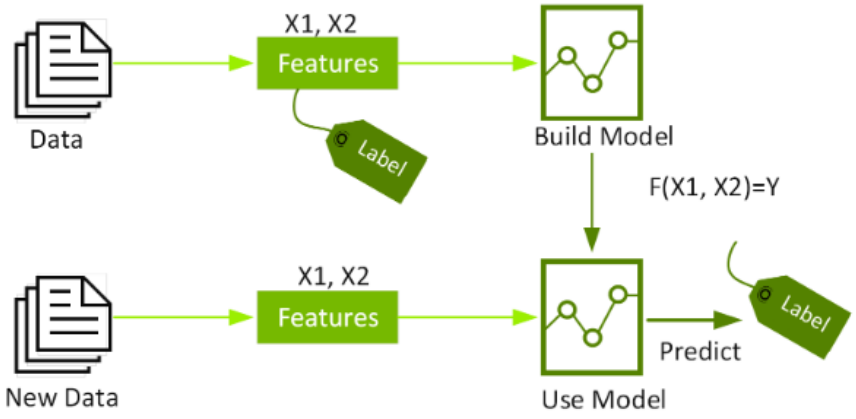

FIGURA 2 – Esquema ilustrativo dos algoritmos de aprendizado de máquina supervisionado -

Fonte: (MCDONALD, 2021)

Os algoritmos de aprendizado supervisionado podem ser de classificação, quando *y* é uma variável categórica, ou de regressão, quando *y* é uma variável real. Neste estudo, ambas as abordagens foram consideradas. No primeiro caso, a variável *y* representa o atraso dos voos na chegada, sendo uma variável categórica binária. Os processos de aprendizado de máquina abordados nesse caso são, portanto, de classificação, e visam encontrar fronteiras de separação – lineares ou não – entre os dados, conforme indica a FIGURA 3. No segundo caso, a variável *y* é uma variável real e representa o atraso na chegada em minutos.

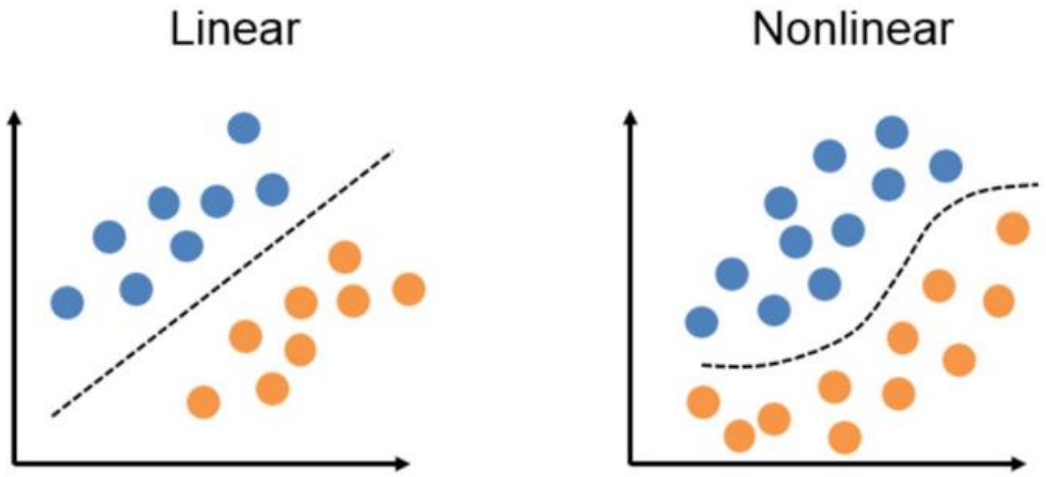

FIGURA 3 – Os processos de aprendizado de máquina de classificação buscam encontrar fronteiras – lineares ou não – entre os dados - Fonte: (SULLIVAN, 2018)

Para se evitar uma péssima generalização dos modelos, pode-se recorrer à validação cruzada, ilustrada na FIGURA 4. Para realizar esse processo, devem ser seguidos os passos descritos a seguir:

- Treinar *n* modelos com diferentes parâmetros de configuração com os dados de treinamento;
- Avaliar o erro com os dados de validação;
- Escolher o modelo com o menor erro de validação;
- Avaliar o erro com os dados de teste para obter a performance final do modelo escolhido.

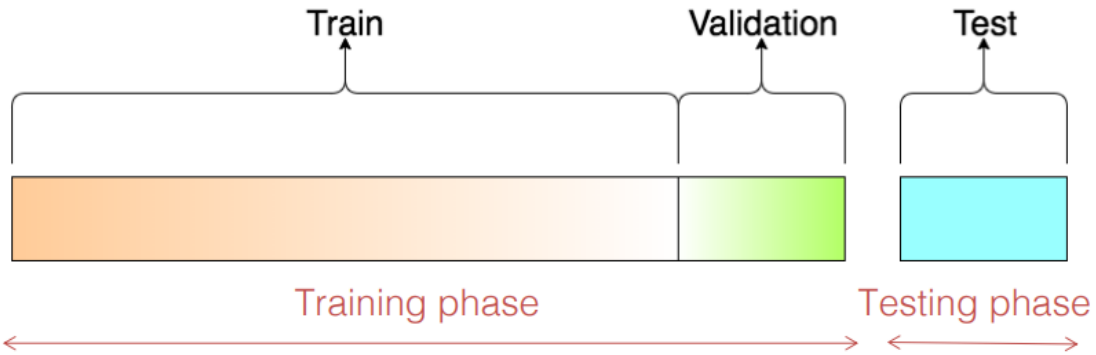

Figura 4 – Esquema ilustrativo da validação cruzada. Fonte: (SHAH, 2017)

Então, dada a metodologia adotada na presente pesquisa, os capítulos a seguir, bem como suas respectivas seções e subseções, estão associados, respectivamente, a cada objetivo específico levantado, com a devida descrição dos algoritmos de aprendizado de máquina supervisionado utilizados neste estudo.

3.1 Coleta de dados

Os dados referentes aos voos neste estudo foram coletados na seção de consulta de voos passados no site da Agência Nacional de Aviação Civil (ANAC). A ANAC mantém na sua base de dados informações a respeito de voos realizados a partir de 01/01/2000 e atualiza os dados até o fim de cada mês com informações acerca de voos referentes ao mês anterior.

As variáveis originais na base VRA da ANAC (2022b) são:

- Empresa aérea: sigla ICAO e nome da companhia aérea que realizou o voo.
- Número Voo: identificador do voo.
- Código DI: classificação interna do voo.
- Código Tipo Linha: classificação da rota.
- Modelo Equipamento: modelo da aeronave que realizou o voo.
- Número de Assentos: número de assentos disponíveis na aeronave que realizou o voo.
- Descrição Aeroporto Origem: sigla ICAO e nome do aeroporto de origem da aeronave.

- Descrição Aeroporto Destino: sigla ICAO e nome do aeroporto de destino da aeronave.
- Partida Prevista: horário previsto para a partida.
- Partida Real: horário real de partida.
- Chegada Prevista: horário previsto para a chegada.
- Chegada real: horário real de chegada.
- Situação Voo: informa se o voo foi realizado, cancelado ou não-informado.

Os códigos de tipo de linha, que indicam as classificações da rota, são indicados na TABELA 3.

TABELA 3 – Descrições dos tipos de linha apresentados na base de dados da ANAC

Sigla Tipo Linha	Descrição Tipo Linha
C	Cargueiro
E	Especial
G	Cargueiro internacional
I	Internacional
L	Rede postal
N	Nacional
R	Regional

Neste estudo, foram considerados apenas os voos domésticos realizados que tiveram como aeroporto de destino o Aeroporto Internacional de Guarulhos, no período de 01/01/2019 até 31/12/2021. Os voos realizados correspondem a 97,42% dos voos da base VRA da ANAC (2022b), ao passo que os voos cancelados e os não informados, juntos, somam apenas 2,58% do total de voos.

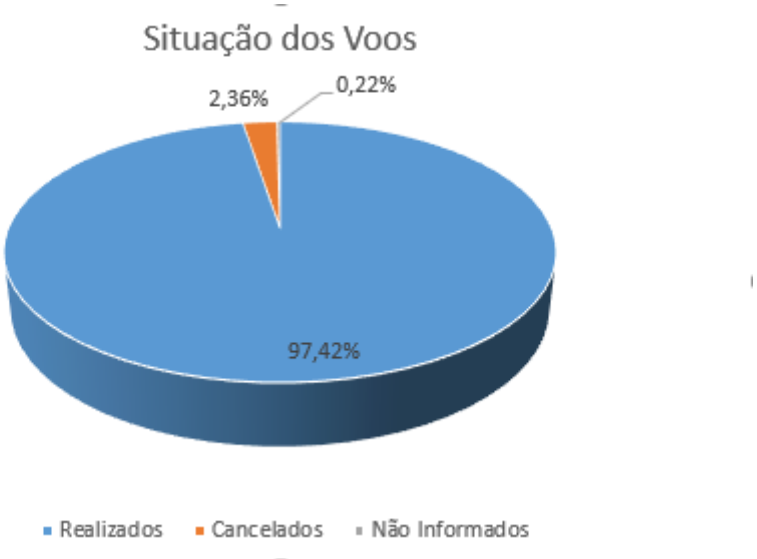

FIGURA 4 –Situação dos voos da base VRA da ANAC

Neste estudo, foram considerados apenas os voos nacionais e regionais, ou seja, com código de linha “N” e “R”, que equivalem a 75,40% do total de voos na base da ANAC.

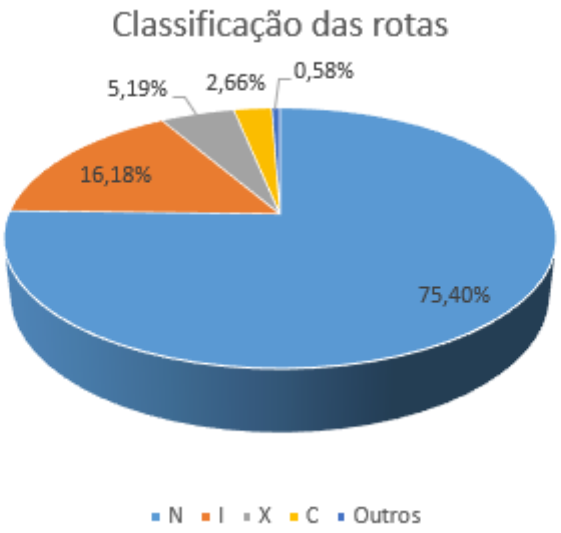

FIGURA 5 – Classificação das rotas dos voos da base VRA da ANAC

Para o atraso, foi adotado o seguinte critério: um voo é considerado atrasado na chegada se ele chega ao destino final 15 minutos ou mais além do horário previsto para a sua chegada. Analogamente, um voo é considerado atrasado na partida se ele parte de sua origem 15 minutos ou mais além do horário previsto para a sua partida.

Para mitigar os efeitos de erros na base de dados original, foram excluídos da amostra os voos que sofreram atraso na chegada ou na partida superior a 5 horas, que representam apenas 0,3% do total de voos realizados.

- Dados referentes às condições meteorológicas

Os dados referentes às condições meteorológicas foram obtidos no endereço eletrônico da Universidade Estadual de Iowa (IOWA, 2022). A base de dados contém informações a respeito das condições meteorológicas em diversos aeroportos ao redor do mundo, incluindo vários aeródromos brasileiros, conforme mostra a FIGURA 6.

FIGURA 6 – Aeródromos brasileiros contemplados pela base de dados da Universidade Estadual de Iowa

Os dados obtidos da base de dados da Universidade Estadual de Iowa que foram analisados neste estudo são:

- Valid: data e horário da observação meteorológica em relação ao Tempo Universal Coordenado.
- Drct: direção do vento (em graus em relação ao norte verdadeiro).
- Sknt: velocidade do vento (em nós).
- Vsby: visibilidade horizontal (em milhas).
- Gust: rajada de vento (em nós).
- Skyc1: cobertura do céu no nível 1.
- Skyc2: cobertura do céu no nível 2.
- Skyc3: cobertura do céu no nível 3.
- Skyc4: cobertura do céu no nível 4.
- Skyl1: altitude do céu no nível 1 (em pés).
- Skyl2: altitude do céu no nível 2 (em pés).
- Skyl3: altitude do céu no nível 3 (em pés).
- Skyl4: altitude do céu no nível 4 (em pés).

- Metar: observação meteorológica no formato METAR.

3.2 Análise das Condições Concomitantes aos Pousos dos Voos

O presente capítulo traz explicações relevantes a respeito de definições primordiais para a pesquisa, como é o caso da definição referente ao atraso de voo, no sentido de quando um voo é considerado atrasado, além de relacionar condições concomitantes aos pousos dos voos coletados no capítulo anterior, subdivididas em variáveis referentes à operação dos voos e variáveis metereológicas.

3.3 Variáveis referentes à operação dos voos

Quanto às variáveis de que trata esta seção, há as seguintes subseções: 3.3.1 Atraso; 3.3.2 Companhia aérea; 3.3.3 Modelo de aeronave; 3.3.4 Número de assentos; 3.3.5 Aeroporto de origem; 3.3.6 Distância; 3.3.7 Variáveis de congestionamento aéreo; 3.3.8 Horário previsto de chegada; 3.3.9 Dia da semana; 3.3.10 Dia do mês; 3.3.11 Mês e, finalmente, 3.3.12 Ano.

3.3.1 Atraso

Um voo é considerado atrasado na partida se parte com um atraso igual ou superior a 15 minutos. A definição é análoga para voos atrasados na chegada. Dos 222.814 voos analisados, 27.896 sofreram atraso na chegada, totalizando cerca de 12,52% do total de voos. Em 26.684 voos houve atraso na partida, totalizando aproximadamente 11,98% dos voos.

Definindo-se o atraso como a diferença em minutos entre o horário real e o horário previsto de chegada ou de partida, pode-se subdividi-lo nas classes indicadas na TABELA 4 e na TABELA 5. Os valores variam no intervalo $[-300,300]$ porque os voos com atrasos e adiantamentos superiores a 5 horas foram retirados da amostra, com o intuito de mitigar erros na base de dados da ANAC. A TABELA 4 e a TABELA 5 apresentam as frequências relativas de cada classe de atraso na chegada e na partida, respectivamente.

TABELA 4 – Frequência relativa das classes de atraso na chegada

Classe de atraso (em minutos)	Frequência relativa
$[-300,0]$	73,46%
$]0,15]$	14,47%
$]15,30]$	5,93%
$]30,60]$	3,83%
$]60,300]$	2,30%

TABELA 5 – Frequência relativa das classes de atraso na partida

Classe de atraso (em minutos)	Frequência relativa
$[-300,0]$	73,08%
$]0,15]$	15,36%
$]15,30]$	5,73%
$]30,60]$	3,62%
$]60,300]$	2,22%

Em ambas as análises – chegada e partida – a frequência relativa de atrasos cai à medida que o tempo de atraso aumenta. Atrasos de mais de 60 minutos são muito raros em ambos os casos – apenas 2,30% dos atrasos na chegada e 2,22% dos atrasos na partida.

3.3.2 Companhia aérea

Considerando-se apenas os voos com código de tipo de linha “N” e “R” na base de dados da ANAC, ao todo 10 companhias aéreas realizaram voos domésticos de 01/01/2019 a 31/12/2021 com passageiros desembarcando no Aeroporto Internacional de Guarulhos. A TABELA 6 apresenta a sigla ICAO e o nome das companhias que realizaram os voos analisados.

TABELA 6 – Companhias aéreas que operaram os voos analisados

Sigla ICAO	Nome da companhia
AZU	Azul
GLO	Gol
ONE	Avianca

PTB	Passaredo
TAM	Latam
TTL	Total
OWT	Azul Conecta
PAM	Map
SID	Sideral
IPM	Itapemirim

Na FIGURA 7, é apresentado um gráfico com os valores percentuais dos números de voos operados pelas companhias aéreas em relação ao número total de voos. Conforme mostra o gráfico, a Latam é a companhia que operou o maior número de voos domésticos, com 45,38% do total, seguida da Gol e da Azul, com 37,27% e 13,12%, respectivamente.

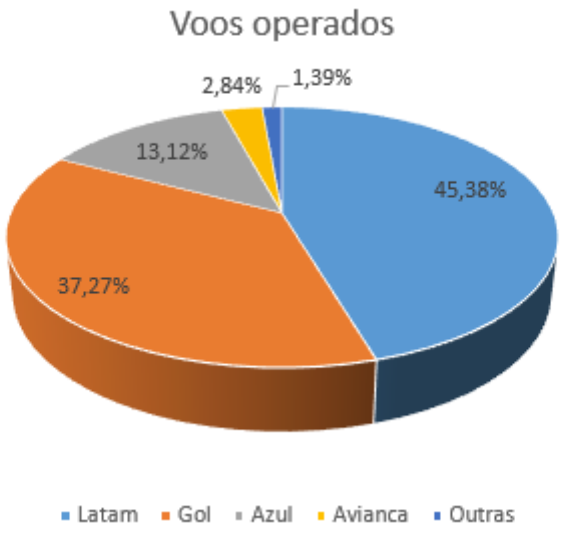

FIGURA 7 – Porcentagem dos voos operados pelas companhias aéreas

A TABELA 7 apresenta as taxas de atraso das companhias aéreas em porcentagem, definidas como as razões entre o número de atrasos na chegada das companhias aéreas e o número total de voos realizados pelas respectivas companhias.

TABELA 7 – Taxas de atraso das companhias aéreas

Companhia aérea	Taxa de atraso
Total	70,59%
Itapemirim	47,34%

Azul Conecta	38,89%
Sideral	33,33%
Passaredo	31,39%
Avianca	24,71%
Gol	14,91%
Azul	14,74%
Latam	8,47%
Map	0,00%

A variável *airline*, a qual indica a companhia aérea referente a cada voo da base de dados analisada, é definida nos modelos por meio de *target encoding*, ou seja, por meio da taxa de atraso. Ou seja, se um voo é operado, por exemplo, pela companhia aérea Total, o seu valor para a variável *airline* é de aproximadamente 0,7059, de acordo com a TABELA 7. Zoutendijk e Mihaela Mitici (2021) utilizaram *target encoding* para definir variáveis categóricas e obtiveram excelentes resultados na predição de atrasos em voos no Aeroporto de Roterdã-Haia.

Dentre os 3 maiores *players* do mercado de voos domésticos no Aeroporto Internacional de Guarulhos, a Latam é a companhia com a melhor OTP, com apenas 8,47% de voos atrasados. A Latam é seguida pela Azul e pela Gol, com 14,74% e 14,91% de atrasos, respectivamente. As companhias aéreas com as piores taxas de atraso são a Total, a Itapemirim e a Azul Conecta, com 70,59%, 47,34% e 38,89% de atrasos, respectivamente.

A taxa de atraso sofre uma ampla variação entre as empresas, variando de 0,00% a 70,59%. Essa grande diferença entre as companhias aéreas com relação à OTP indica que a variável *airline* pode exercer influência considerável sobre os atrasos. Portanto, ela foi incluída nos modelos de previsão de atraso do aeroporto em análise.

3.3.3 Modelo de aeronave

Ao todo, os voos domésticos da base de dados analisada foram realizados em 23 modelos de aeronave distintos. Os modelos com o maior número de voos são o A320, B738 e A321, com 30,93%, 27,30% e 15,52% do total de voos realizados, respectivamente. A TABELA 8 apresenta os modelos e os seus respectivos números de voos operados, dados em porcentagem em relação ao número total de voos.

TABELA 8 – Número de voos operados (em porcentagem) nos diversos modelos de aeronave

Modelo de aeronave	Voos operados (em %)
A320	30,93%
B738	27,30%
A321	15,52%
B737	7,36%
A319	5,46%
E195	5,28%
B38M	2,60%
A20N	2,11%
B763	1,39%
AT72	1,32%
E190	0,31%
A318	0,12%
E295	0,09%
B77W	0,06%
B773	0,04%
A359	0,04%
B789	0,02%
ATP	0,01%
B733	0,01%
B735	0,01%
C208	0,01%
B722	0,01%
B734	0,00%

A variável *aircraft*, a qual indica o modelo de aeronave referente a cada voo da base de dados analisada, é definida nos modelos por meio de *target encoding*. A TABELA 9 apresenta as taxas de atraso em porcentagem, definidas como as razões entre o número de atrasos na chegada em voos operados pelos modelos de aeronave e o número total de voos realizados pelos respectivos modelos.

TABELA 9 – Taxas de atraso dos diversos modelos de aeronave

Modelo de aeronave	Taxa de atraso
B722	70,59%
ATP	55,56%
C208	38,89%
B733	34,62%
B789	31,71%
B735	31,58%
AT72	29,76%
A318	20,16%
B77W	19,15%
B773	18,75%
B38M	16,03%
E190	15,93%
B738	15,66%
A359	14,46%
E195	13,77%
B737	11,70%
A320	11,41%
A20N	11,23%
A321	9,49%
E295	8,17%
A319	7,12%
B763	6,19%
B734	0,00%

O modelo de aeronave com pior OTP é o B722, com taxa de atraso de 70,59%. Os modelos B734 e B763, por sua vez, tiveram 0,00% e 6,19% de atrasos, respectivamente, sendo os modelos com mais pontualidade.

A grande variação das taxas de atraso entre os modelos de aeronave, indicada na TABELA 9, indica que a variável *aircraft* pode exercer influência considerável sobre os atrasos. Portanto, ela foi incluída nos modelos de previsão de atraso do aeroporto em análise.

3.3.4 Número de assentos

A informação referente ao número de assentos disponíveis em cada voo é extraída da base de dados da ANAC. Conforme apresenta a TABELA 10, pode-se subdividir o número de assentos em 7 classes distintas para analisar os dados. A TABELA 10 indica o número de voos operados em cada uma das 7 classes, dado em porcentagem em relação ao número total de voos realizados.

TABELA 10 – Número de voos (em porcentagem) de cada classe de número de assentos

Número da classe	Classe de assentos	Número de voos
1	[0,50[	0,26%
2	[50,100[	1,34%
3	[100,150[	18,73%
4	[150,200[	62,71%
5	[200,250]	16,83%
6	[250,350]	0,05%
7	[350,410]	0,07%

A TABELA 10 indica que a classe com maior número de voos é a [150,200[, com 62,71% dos voos operados; ao passo que a classe com o menor número é a [250,350[,com apenas 0,05% dos voos realizados.

O gráfico da FIGURA 8 apresenta a taxa de atraso de cada uma das classes de assentos, ou seja, a razão em percentagem entre o número de voos atrasados em cada classe e o número de voos operados na classe.

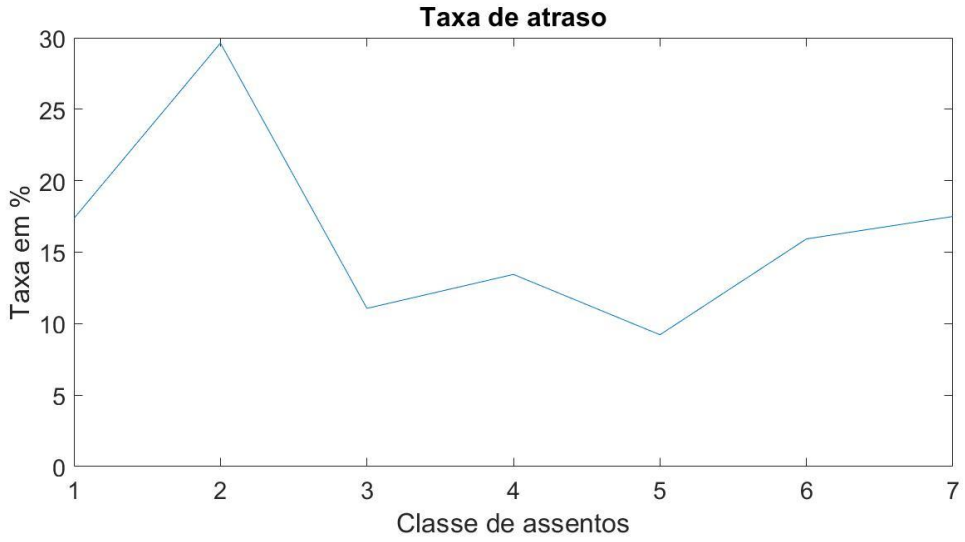

FIGURA 8 – Gráfico das taxas de atraso de acordo com as classes de número de assentos

Esperava-se que a taxa de atrasos fosse maior nas classes com maior número de assentos, devido ao tempo maior para o embarque e o desembarque de passageiros. Entretanto, a FIGURA 8 não ratifica o comportamento esperado: a classe com maior taxa de atrasos é a 2, com taxa de 29,63%, ao passo que a classe com a menor taxa é a 5, com apenas 9,22%. Diversas hipóteses podem explicar esse comportamento, tais como a alocação de aeronaves menores a terminais com desvantagens operacionais e uma eventual prioridade dada pelos controladores de voo para o pouso de aeronaves maiores em situações críticas.

A grande variação das taxas de atraso entre as classes de número de assentos, indicada no gráfico da FIGURA 8, indica que a variável numérica *seats*, que indica o número de assentos disponíveis em cada voo, pode exercer influência considerável sobre os atrasos. Portanto, ela foi incluída nos modelos de previsão de atraso do aeroporto em análise.

3.3.5 Aeroporto de origem

Ao todo, os voos domésticos que desembarcaram no Aeroporto Internacional de Guarulhos no período analisado partiram de 66 aeródromos distintos. A TABELA 11 apresenta a lista dos aeroportos de origem dos voos analisados, com as siglas ICAO dos aeródromos, seus nomes e o número de voos que partiram de cada um dos aeroportos, dado em porcentagem em relação ao número total de voos.

TABELA 11 – Aeroportos de origem dos voos analisados

Sigla ICAO	Aeroporto de origem	Número de voos
SBPA	Aeroporto Internacional de Porto Alegre	7,42%
SBRF	Aeroporto Internacional do Recife / Guararapes	6,22%
SBCF	Aeroporto Internacional de Belo Horizonte / Confins	5,86%
SBSV	Aeroporto Internacional de Salvador	5,74%
SBCT	Aeroporto Internacional de Curitiba / Afonso Pena	5,38%
SBBR	Aeroporto Internacional de Brasília	4,91%
SBFZ	Aeroporto Internacional de Fortaleza	4,85%
SBFL	Aeroporto Internacional de Florianópolis	4,57%
SBRJ	Aeroporto do Rio de Janeiro / Santos Dumont	4,41%
SBGL	Aeroporto Internacional Tom Jobim / Rio de Janeiro	3,81%
SBVT	Aeroporto Internacional de Vitória	3,07%
SBGO	Aeroporto Internacional de Goiânia	3,04%
SBCY	Aeroporto Internacional de Cuiabá	2,95%
SBSG	Aeroporto Internacional de Natal	2,87%
SBMO	Aeroporto Internacional de Maceió	2,87%
SBFI	Aeroporto Internacional de Foz do Iguaçu	2,72%
SBEG	Aeroporto Internacional de Manaus	2,29%
SBPS	Aeroporto Internacional de Porto Seguro	2,25%
SBNF	Aeroporto Internacional de Navegantes	2,18%
SBCG	Aeroporto Internacional de Campo Grande	2,11%
SBBE	Aeroporto Internacional de Belém	1,96%
SBJP	Aeroporto Internacional de João Pessoa	1,65%
SBAR	Aeroporto Internacional de Aracaju	1,51%
SBSL	Aeroporto Internacional de São Luís	1,21%
SBJU	Aeroporto de Juazeiro do Norte	0,95%
SBTE	Aeroporto de Teresina	0,94%
SBMG	Aeroporto Regional de Maringá	0,93%

SBLO	Aeroporto de Londrina	0,92%
SBUL	Aeroporto de Uberlândia	0,92%
SBIL	Aeroporto Jorge Amado / Ilhéus	0,89%
SBRP	Aeroporto Estadual de Ribeirão Preto	0,88%
SBCH	Aeroporto de Chapecó	0,86%
SBPJ	Aeroporto de Palmas	0,64%
SBSR	Aeroporto Estadual de São José do Rio Preto	0,60%
SBPL	Aeroporto Internacional de Petrolina	0,52%
SBVC	Aeroporto de Vitória da Conquista	0,46%
SBJV	Aeroporto de Joinville	0,45%
SBIZ	Aeroporto de Imperatriz	0,45%
SBPV	Aeroporto Internacional de Porto Velho	0,42%
SBRB	Aeroporto Internacional de Rio Branco	0,35%
SBCA	Aeroporto Municipal de Cascavel	0,33%
SBKG	Aeroporto de Campina Grande	0,30%
SBJE	Aeroporto de Jericoacoara	0,25%
SBMK	Aeroporto de Montes Claros	0,24%
SBAE	Aeroporto Estadual de Bauru Arealva	0,20%
SBPF	Aeroporto de Passo Fundo	0,19%
SBSI	Aeroporto de Sinop	0,18%
SBCX	Aeroporto Regional de Caxias do Sul	0,17%
SBAU	Aeroporto Estadual de Araçatuba	0,17%
SBDN	Aeroporto Estadual de Presidente Prudente	0,15%
SBQV	Aeroporto Pedro Otacílio Figueiredo	0,13%
SBGR	Aeroporto Internacional de São Paulo / Guarulhos	0,12%
SBCN	Aeroporto de Caldas Novas	0,12%
SBKP	Aeroporto Internacional de Viracopos / Campinas	0,10%
SBDO	Aeroporto Regional de Dourados	0,07%
SBJA	Aeroporto de Jaguaruna	0,07%
SBSP	Aeroporto de São Paulo / Congonhas	0,06%
SBZM	Aeroporto Regional da Zona da Mata	0,05%
SBCB	Aeroporto Internacional de Cabo Frio	0,04%
SBTC	Aeroporto de Una-Comandatuba	0,02%

SBBV	Aeroporto Internacional de Boa Vista	0,01%
SBSJ	Aeroporto Internacional de São José dos Campos	0,00%
SBJD	Aeroporto Estadual de Jundiaí	0,00%
SBBH	Aeroporto de Belo Horizonte / Pampulha	0,00%
SBCZ	Aeroporto Internacional de Cruzeiro do Sul	0,00%
SBSN	Aeroporto Internacional de Santarém	0,00%

De acordo com a TABELA 11, os aeroportos com o maior número de voos são o Aeroporto Internacional de Porto Alegre, o Aeroporto Internacional do Recife, o Aeroporto Internacional de Belo Horizonte, o Aeroporto Internacional de Salvador e o Aeroporto Internacional de Curitiba. O mapa da FIGURA 9 apresenta os aeroportos de origem da base de dados.

FIGURA 9 – Aeroportos de origem dos voos da base de dados da ANAC

A variável *airport*, a qual indica o aeroporto de origem referente a cada voo da base de dados utilizada, é definida nos modelos por meio de *target encoding*. A TABELA 12 apresenta as taxas de atraso em porcentagem, definidas como as razões entre o número de voos atrasados provenientes de cada aeroporto e o número total de voos provenientes do aeródromo.

TABELA 12 – Taxas de atraso dos aeroportos de origem dos voos analisados

Sigla ICAO	Aeroporto de origem	Taxa de atraso
SBQV	Aeroporto Pedro Otacílio Figueiredo	64,21%
SBDO	Aeroporto Regional de Dourados	43,37%
SBBV	Aeroporto Internacional de Boa Vista	32,00%
SBRP	Aeroporto Estadual de Ribeirão Preto	26,47%
SBJV	Aeroporto de Joinville	17,38%
SBJD	Aeroporto Estadual de Jundiaí	16,67%
SBBE	Aeroporto Internacional de Belém	16,45%
SBPA	Aeroporto Internacional de Porto Alegre	16,06%
SBCH	Aeroporto de Chapecó	16,06%
SBAE	Aeroporto Estadual de Bauru Arealva	15,77%
SBCF	Aeroporto Internacional de Belo Horizonte / Confins	14,59%
SBEG	Aeroporto Internacional de Manaus	14,57%
SBKG	Aeroporto de Campina Grande	14,55%
SBCB	Aeroporto Internacional de Cabo Frio	14,29%
SBRF	Aeroporto Internacional do Recife / Guararapes	14,10%
SBNF	Aeroporto Internacional de Navegantes	14,04%
SBPF	Aeroporto de Passo Fundo	13,80%
SBDN	Aeroporto Estadual de Presidente Prudente	13,72%
SBFL	Aeroporto Internacional de Florianópolis	13,26%
SBFZ	Aeroporto Internacional de Fortaleza	12,96%
SBCT	Aeroporto Internacional de Curitiba / Afonso Pena	12,88%

SBIL	Aeroporto Jorge Amado / Ilhéus	12,64%
SBSV	Aeroporto Internacional de Salvador	12,54%
SBTC	Aeroporto de Una-Comandatuba	12,50%
SBFI	Aeroporto Internacional de Foz do Iguaçu	12,41%
SBAR	Aeroporto Internacional de Aracaju	12,37%
SBJE	Aeroporto de Jericoacoara	12,23%
SBPV	Aeroporto Internacional de Porto Velho	12,20%
SBSG	Aeroporto Internacional de Natal	12,17%
SBJP	Aeroporto Internacional de João Pessoa	12,10%
SBJU	Aeroporto de Juazeiro do Norte	11,90%
SBBR	Aeroporto Internacional de Brasília	11,50%
SBGL	Aeroporto Internacional Tom Jobim / Rio de Janeiro	11,48%
SBGO	Aeroporto Internacional de Goiânia	11,45%
SBCY	Aeroporto Internacional de Cuiabá	11,21%
SBPS	Aeroporto Internacional de Porto Seguro	11,20%
SBVT	Aeroporto Internacional de Vitória	11,17%
SBCX	Aeroporto Regional de Caxias do Sul	11,08%
SBRB	Aeroporto Internacional de Rio Branco	10,94%
SBRJ	Aeroporto do Rio de Janeiro / Santos Dumont	10,67%
SBMK	Aeroporto de Montes Claros	10,53%
SBCN	Aeroporto de Caldas Novas	10,38%
SBMG	Aeroporto Regional de Maringá	9,99%
SBMO	Aeroporto Internacional de Maceió	9,89%
SBSL	Aeroporto Internacional de São Luís	9,53%
SBZM	Aeroporto Regional da Zona da Mata	8,93%
SBSI	Aeroporto de Sinop	8,46%
SBCA	Aeroporto Municipal de Cascavel	8,22%
SBPL	Aeroporto Internacional de Petrolina	8,05%
SBSR	Aeroporto Estadual de São José do Rio Preto	7,74%
SBCG	Aeroporto Internacional de Campo Grande	7,37%

SBTE	Aeroporto de Teresina	6,82%
SBPJ	Aeroporto de Palmas	6,79%
SBVC	Aeroporto de Vitória da Conquista	6,42%
SBJA	Aeroporto de Jaguaruna	6,33%
SBIZ	Aeroporto de Imperatriz	6,19%
SBAU	Aeroporto Estadual de Araçatuba	5,79%
SBUL	Aeroporto de Uberlândia	5,73%
SBLO	Aeroporto de Londrina	5,21%
SBKP	Aeroporto Internacional de Viracopos / Campinas	1,38%
SBGR	Aeroporto Internacional de São Paulo / Guarulhos	0,00%
SBSP	Aeroporto de São Paulo / Congonhas	0,00%
SBSJ	Aeroporto Internacional de São José dos Campos	0,00%
SBCZ	Aeroporto Internacional de Cruzeiro do Sul	0,00%
SBSN	Aeroporto Internacional de Santarém	0,00%
SBBH	Aeroporto de Belo Horizonte / Pampulha	0,00%

A TABELA 12 indica que há uma variação de 0,00% a 64,21% nas taxas de atraso dos aeroportos de origem da base de dados utilizada. Essa ampla variação indica que a variável *airport* pode exercer influência considerável sobre os atrasos. Portanto, ela foi incluída nos modelos de previsão de atraso no aeroporto em análise.

3.3.6 Distância

A distância *d* entre o Aeroporto Internacional de Guarulhos e os aeroportos de origem dos voos pode ser calculada por meio da fórmula de Haversine em função das coordenadas geográficas dos aeroportos:

$$d = 2r * arcsen\left(\sqrt{sen^2\left(\frac{\varphi_i - \varphi_{gru}}{2}\right) + \cos(\varphi_{gru}) * \cos(\varphi_i) * sen^2(\frac{\lambda_i - \lambda_{gru}}{2})}\right)$$

Onde:

- r: raio da Terra.
- φ_{gru}: latitude do Aeroporto Internacional de Guarulhos.
- φ_i: latitude do aeroporto i.
- λ_{gru}: longitude do Aeroporto Internacional de Guarulhos.
- λ_i: longitude do aeroporto i.

A distância pode ser subdividida em 6 classes distintas, indicadas na TABELA 13, que apresenta o número de voos realizados em cada uma das classes, dado em porcentagem em relação ao número total de voos operados.

TABELA 13 – Número de voos (em porcentagem) de cada uma das classes de distância

Classe de distância (em km)	Número de voos
[0,500[	25,24%
[500,1000[	31,87%
[1000,1500[	13,10%
[1500,2000[	6,48%
[2000,2500[	20,67%
[2500,3310]	2,65%

O gráfico da FIGURA 10 apresenta a taxa de atraso de cada uma das classes de distância, definida como a razão entre o número de atrasos de cada classe e o número de voos operados na classe.

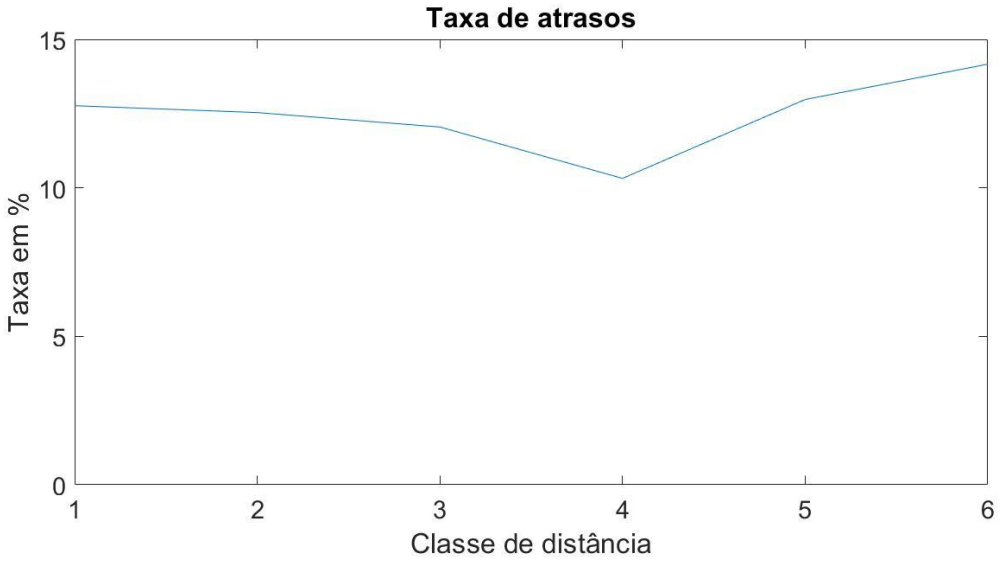

FIGURA 10 – Gráfico das taxas de atraso das classes de distância

O gráfico da FIGURA 10**Erro! Fonte de referência não encontrada.** não apresenta uma grande variação nas taxas de atraso das classes de distância: a taxa varia de 10,32% na classe 4 a 14,16% na classe 6. Contudo, a variável *distance*, que indica a distância do aeroporto de origem de cada voo em relação ao Aeroporto Internacional de Guarulhos, será incluída nos modelos de previsão de atraso no aeroporto em análise, pois está presente em muitos modelos na literatura.

3.3.7 Variáveis de congestionamento aéreo

Diversos fatores exercem impacto sobre atrasos em voos. Entre esses fatores, podem ser citadas as condições meteorológicas e o congestionamento aéreo. Quanto mais aeronaves pousam e decolam em um determinado período de tempo em um aeroporto, maior é a tendência de ocorrência de atrasos.

Duas variáveis presentes nos modelos a serem desenvolvidos, a *sch_d* e a *sch_h*, estão relacionadas com o congestionamento aéreo. A variável *sch_d* indica o número de pousos e de decolagens no mesmo dia que o pouso do voo correspondente. A *sch_h*, por sua vez, indica o número de pousos e de decolagens no intervalo de tempo entre 1 hora antes e 1 hora depois do horário previsto para o pouso do voo correspondente.

A variável *sch_d* pode ser subdividida em 9 classes distintas, indicadas na TABELA 14, que apresenta o número de voos realizados em cada uma das classes, dado em porcentagem em relação ao número total de voos operados.

TABELA 14 – Número de voos operados (em porcentagem) em cada uma das classes de *sch_d*

Classe de *sch_d*	Número de voos
[0,100[	0,61%
[100,200[	1,26%
[200,300[	3,17%
[300,400[	7,91%
[400,500[	12,99%
[500,600[	16,11%
[600,700[	11,99%
[700,800[	28,05%
[800,898]	17,92%

O gráfico da FIGURA 11 apresenta as taxas de atraso em porcentagem de cada uma das classes de *sch_d*, definidas como a razão entre o número de voos atrasados de cada classe e o número de voos operados dentro da classe.

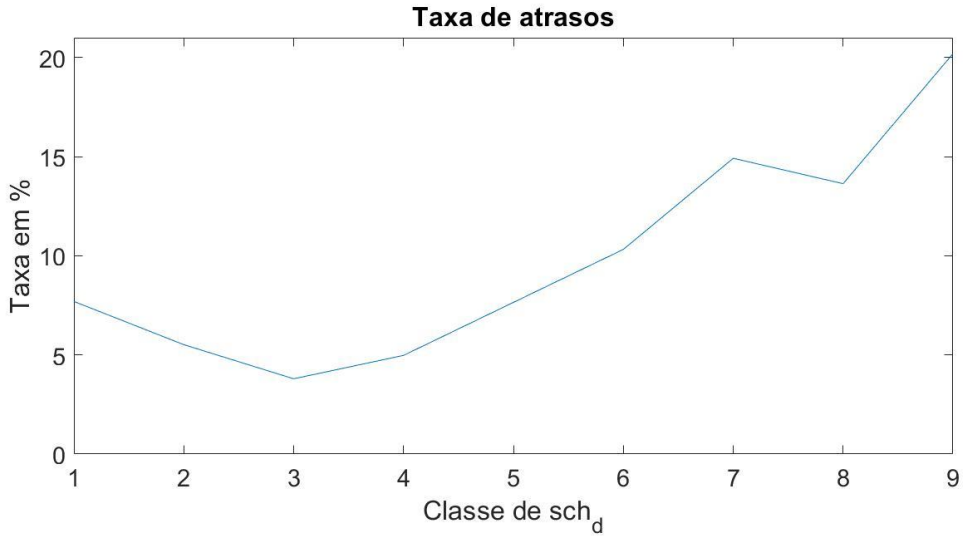

FIGURA 11 – Gráfico das taxas de atraso das classes de *sch_d*

O gráfico da FIGURA 11 indica uma tendência de haver mais atrasos quando a *sch_d* é maior, com uma variação na taxa de atrasos de 3,80% na classe 3 a 20,20% na classe 9. Esse comportamento era esperado, visto que quanto maior é o número de pousos e decolagens em um dia, maior tende a ser o número de atrasos.

A grande variação das taxas de atraso entre as classes de *sch_d* indica que essa variável pode exercer uma influência considerável sobre os atrasos. Portanto, a variável *sch_d* será incluída nos modelos de previsão de atraso do aeroporto em análise.

A variável *sch_h*, por sua vez, pode ser subdividida em 6 classes distintas, indicadas na TABELA 15, que apresenta o número de voos realizados em cada uma das classes, dado em porcentagem em relação ao número total de voos operados.

TABELA 15 – Número de voos (em porcentagem) das classes de *sch_h*

Classe de *sch_h*	Número de voos
[0,25[	1,87%
[25,50[	7,63%
[50,75[	19,47%
[75,100]	28,46%
[100,125[	27,07%
[125,168]	15,51%

O gráfico da FIGURA 12 apresenta as taxas de atraso em porcentagem de cada uma das classes de *sch_h*, definidas como a razão entre o número de voos atrasados de cada classe e o número de voos operados dentro da classe.

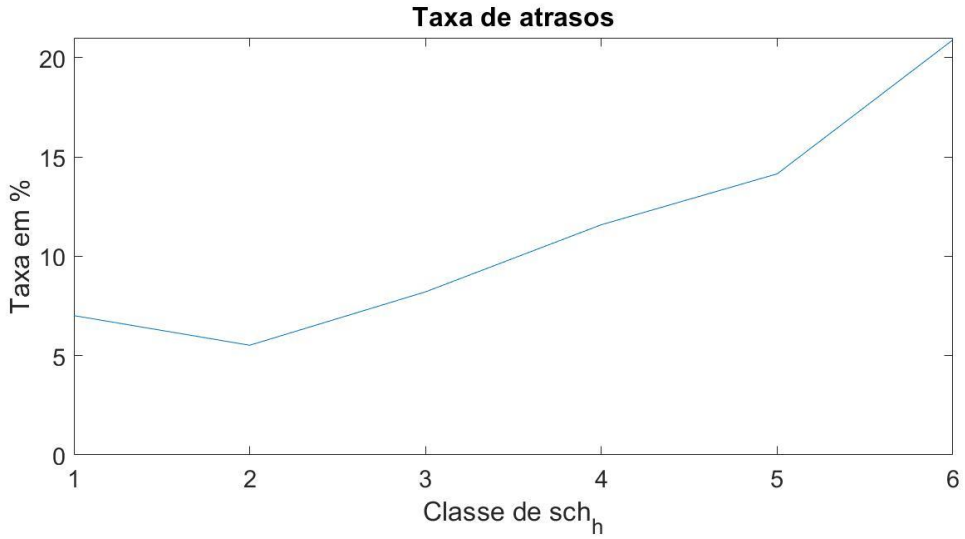

FIGURA 12 – Gráfico das taxas de atraso das classes de *sch_h*

A análise do gráfico indica que há uma tendência de aumento no número de atrasos à

medida que se aumenta a variável *sch_h*, sendo que a taxa de atraso varia de 5,52% na classe 2 a 20,90% na classe 6. Esse comportamento já era esperado, pois quanto maior é o número de pousos e decolagens em um determinado horário, maior tende a ser o número de atrasos.

A grande variação das taxas de atraso entre as classes de *sch_h* indica que essa variável pode exercer uma influência considerável sobre os atrasos. Portanto, a variável *sch_h* será incluída nos modelos de previsão de atraso do aeroporto em análise.

3.3.8 Horário previsto de chegada

Na base de dados da ANAC, o horário previsto de chegada dos voos é apresentado com precisão de minutos, com o formato $hh{:}\,mm{:}\,00$.

Para obter o horário previsto de chegada considerado nos modelos, foram desprezados os minutos. Assim, um voo com horário 0 pode ter ocorrido de 00h até 00:59h, um voo com horário 1 pode ter ocorrido de 01h até 01:59h, e assim por diante. Dessa forma, os horários assumem números inteiros no intervalo $[0{,}23]$.

Para preservar a periodicidade do horário previsto de chegada, foram definidas duas variáveis trigonométricas para o horário, definidas da seguinte maneira:

$$hour_{cos} = \cos\left(\frac{2\pi * hor\acute{a}rio}{24}\right)$$

$$hour_{sin} = sen\left(\frac{2\pi * hor\acute{a}rio}{24}\right)$$

O gráfico da FIGURA 13 apresenta o número de voos por horário do dia ao longo dos 3 anos de análise, dado em porcentagem em relação ao número total de voos. A partir do gráfico, é possível concluir que há horários-pico, com uma quantidade grande de voos: o horário 7, o 12, o 16 e o 20; com 8,16%, 7,42%, 8,33% e 8,40% do total de voos, respectivamente.

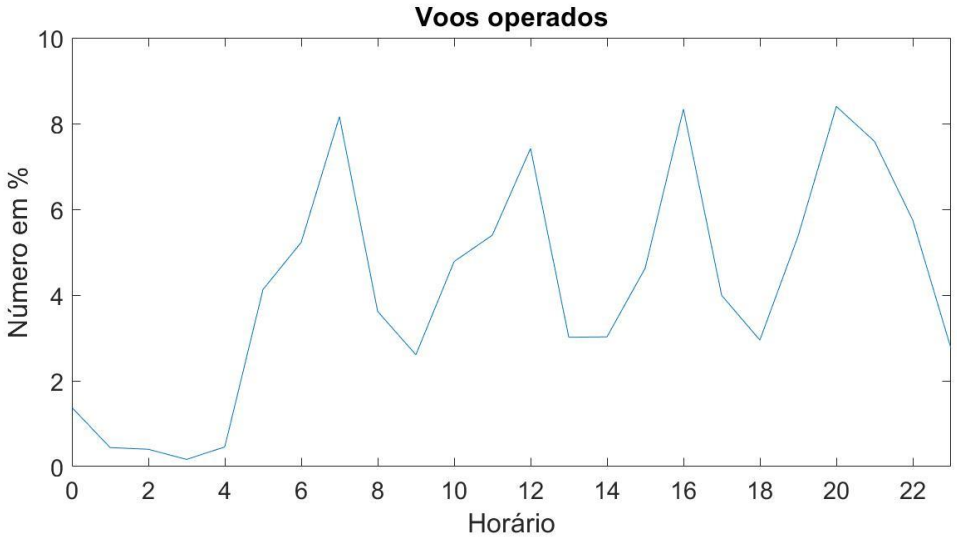

FIGURA 13 – Número de voos operados (em porcentagem) de acordo com o horário previsto de chegada

O gráfico da FIGURA 14 indica que há uma variação considerável na taxa de atrasos ao longo dos horários do dia. A taxa mínima ocorre no horário das 4:00 às 4:59 e equivale a cerca de 6,46%, ao passo que a taxa máxima ocorre das 19:00 às 19:59 e equivale a cerca de 19,99% do total de voos. Essa variação nas taxas de atraso indica que as variáveis relacionadas ao horário previsto de chegada podem exercer uma influência considerável sobre os atrasos, logo serão consideradas nos modelos de previsão de atrasos do aeroporto em análise.

FIGURA 14 – Gráfico das taxas de atraso ao longo dos horários do dia

3.3.9 Dia da semana

Neste estudo, foi adotada a seguinte convenção: o domingo é considerado como o dia 1; a segunda-feira, como o dia 2; a terça-feira, como o dia 3, e assim por diante. Foi considerado o dia da semana do horário previsto de chegada dos voos.

De forma análoga ao horário do dia, para preservar a periodicidade dos dias da semana, foram definidas duas variáveis trigonométricas, definidas da seguinte maneira:

$$day_{week_{cos}} = \cos\left(\frac{2\pi * dia}{7}\right)$$

$$day_{week_{sin}} = sen\left(\frac{2\pi * dia}{7}\right)$$

O gráfico da FIGURA 15 indica que há um maior número de voos nas sextas-feiras e nas segundas-feiras, quando ocorrem 15,32% e 15,00¨% dos voos, respectivamente. Entretanto, não há uma grande variação no número de voos ao longo da semana: o dia com menos voos é domingo, com cerca de 12,68% do total de voos operados.

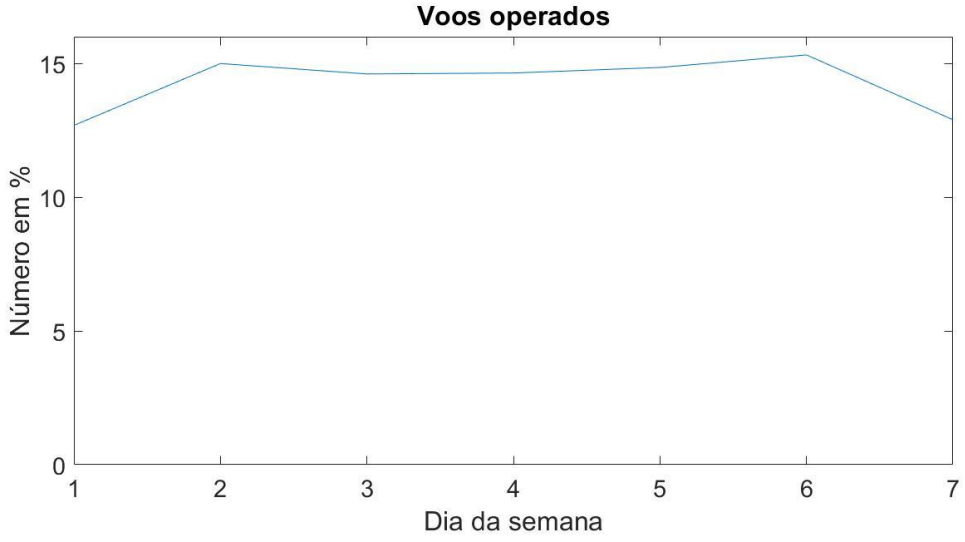

FIGURA 15 – Número de voos operados (em porcentagem) ao longo dos dias da semana

O gráfico da FIGURA 16 apresenta as taxas de atraso de cada um dos dias da semana. Os dias com as maiores taxas de atraso são também os dias com maiores volumes de voos: a sexta-feira e a segunda-feira, com taxas de 14,69% e 13,51%, respectivamente. Os dias com as menores taxas de atraso são os dias com menores volumes de voos: o sábado e o domingo,

com taxas de 10,27% e 10,34%, respectivamente. Embora não haja uma grande variação nas taxas de atraso ao longo dos dias da semana, as variáveis *day_week_cos* e *day_week_sien* foram incluídas nos modelos de previsão de atraso do aeroporto em análise, pois essas variáveis são consideradas em muitos modelos pesquisados na literatura.

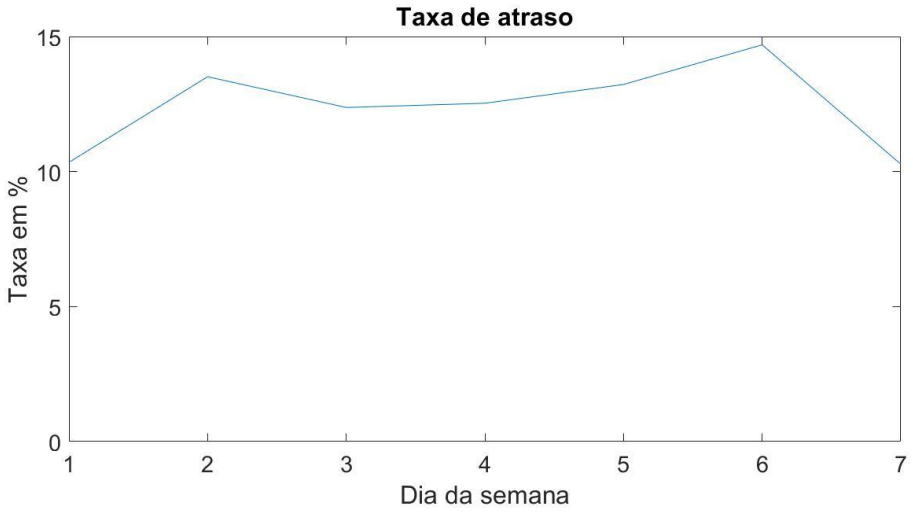

FIGURA 16 – Taxas de atraso dos dias da semana

3.3.10 Dia do mês

Para o dia do mês, foram consideradas 3 variáveis numéricas, sendo 2 trigonométricas. A variável numérica *day_month* assume valores inteiros de 1 a 31 e indica o dia do mês do horário previsto para o pouso do voo correspondente. As variáveis trigonométricas preservam a periodicidade do dia do mês e são definidas da seguinte maneira:

$$day_{month_{cos}} = \cos\left(\frac{2\pi * dia}{31}\right)$$

$$day_{month_{sin}} = sen\left(\frac{2\pi * dia}{31}\right)$$

O gráfico da FIGURA 17 indica que não há uma grande variação no número de voos ao longo dos dias do mês. A exceção é o dia 31, no qual é operado um número menor de voos em relação aos demais dias, porque nem todos os meses têm 31 dias.

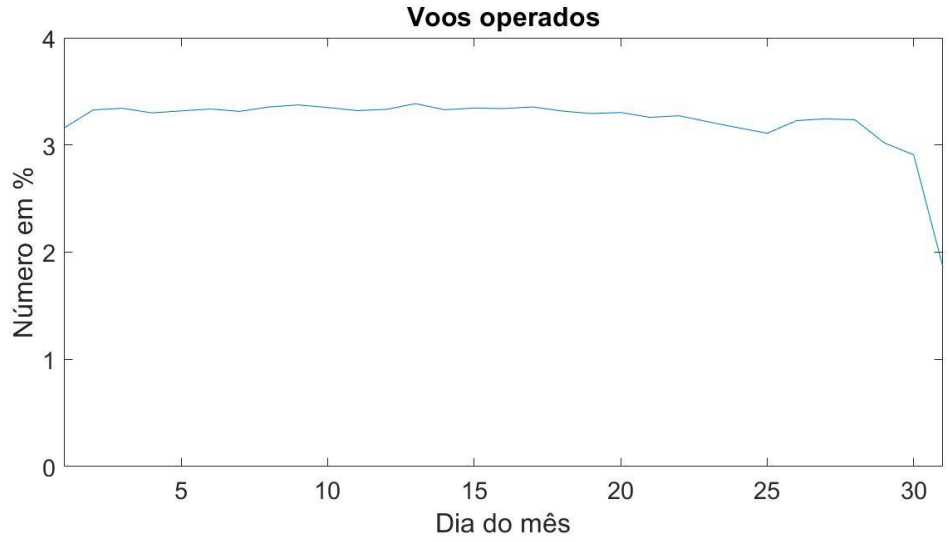

FIGURA 17 – Número de voos operados (em porcentagem) ao longo dos dias do mês

O gráfico da FIGURA 18 apresenta a variação das taxas de atraso ao longo dos dias do mês. As taxas variam de 10,15% no dia 25 a 15,98% no dia 17. Embora não haja uma grande variação nas taxas de atraso ao longo dos dias do mês, as variáveis referentes ao dia do mês foram incluídas nos modelos de previsão de atraso no aeroporto em análise porque elas são consideradas em muitos modelos pesquisados na literatura.

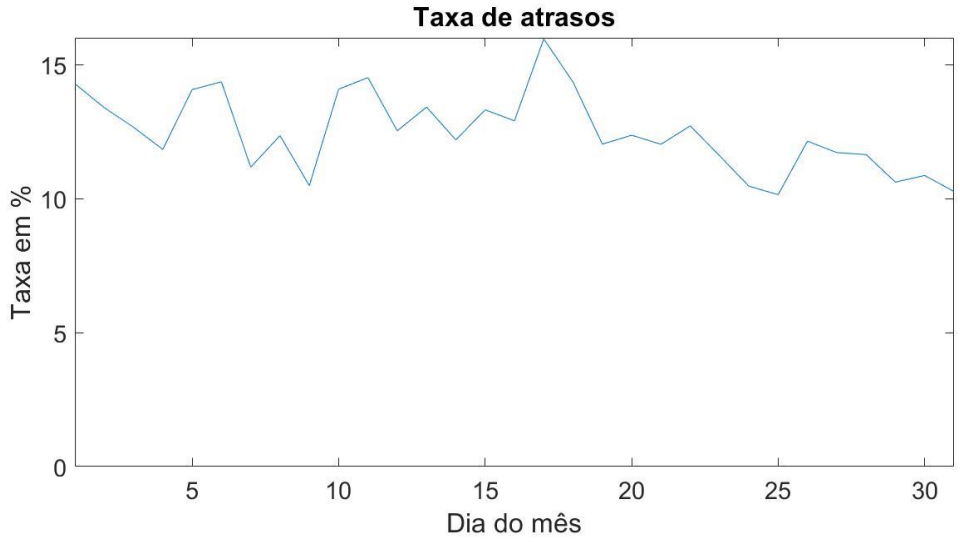

FIGURA 18 – Taxas de atraso ao longo dos dias do mês

3.3.11 Mês

Neste estudo, o mês de janeiro é considerado o mês 1; o mês de fevereiro, o mês 2, e assim por diante. Foi considerado o mês do horário previsto de chegada dos voos.

De forma análoga ao horário do dia, para preservar a periodicidade dos meses, foram definidas duas variáveis trigonométricas, definidas da seguinte maneira:

$$month\_cos = \cos\left(\frac{2\pi * mês}{12}\right)$$

$$month\_sin = sen\left(\frac{2\pi * mês}{12}\right)$$

De acordo com o gráfico da FIGURA 19, os meses com os maiores números de voos são janeiro e dezembro, com 11,41% e 10,71% dos voos, respectivamente. Esse resultado era esperado, visto que são os meses de férias e das festas de fim de ano, quando os aeroportos costumam ficar mais lotados. O mês de abril, por sua vez, é o mês com menor volume de voos, com apenas 5,35% do total de voos.

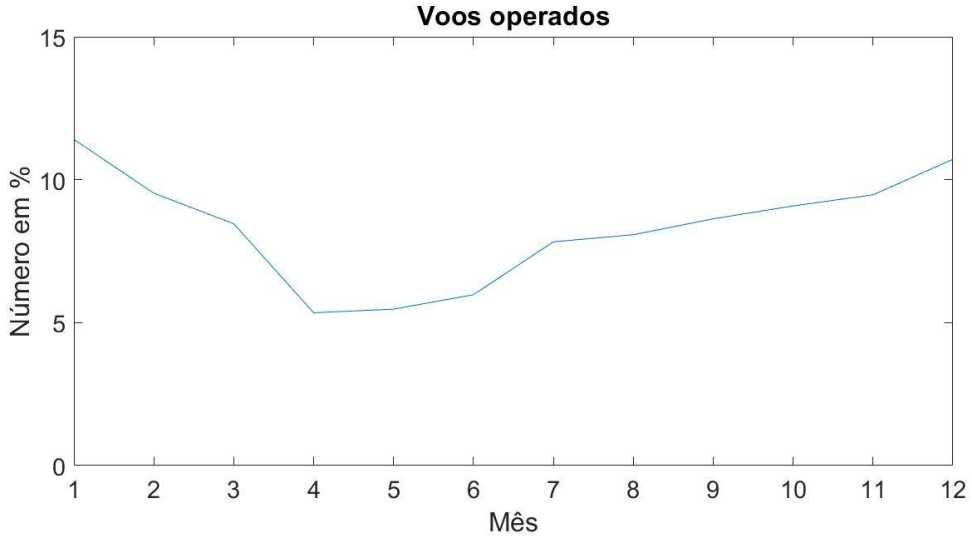

FIGURA 19 – Número de voos operados (em porcentagem) ao longo dos meses do ano

Além das 2 variáveis trigonométricas supracitadas, há mais uma variável relacionada com o mês referente ao horário previsto de pouso de cada voo: a variável *season*, a qual é definida por *target encoding*. A TABELA 16 apresenta as taxas de atraso em porcentagem, definidas como as razões entre o número de voos atrasados de cada mês e o número total de voos realizados por mês.

TABELA 16 – Taxas de atraso ao longo dos meses do ano

Mês	Taxa de atraso
Janeiro	16,36%
Fevereiro	15,51%
Março	11,42%
Abril	11,84%
Maio	7,36%
Junho	6,61%
Julho	9,54%
Agosto	6,79%
Setembro	9,57%
Outubro	9,67%
Novembro	16,81%
Dezembro	20,42%

A análise da TABELA 16 indica que há uma grande variação nas taxas de atraso ao longo dos meses do ano e que as taxas são maiores nos meses em que há mais voos. A taxa varia de 6,61% no mês de junho até 20,42% do mês de dezembro. Essa ampla variação das taxas de atraso ao longo dos meses do ano indica que as 3 variáveis referentes ao mês do ano podem exercer influência considerável sobre os atrasos e, portanto, serão incluídas nos modelos de previsão de atraso do aeroporto em análise.

3.3.12 Ano

Foram coletadas amostras ao longo de 3 anos: 2019, 2020 e 2021. A variável *year* indica o ano correspondente ao horário previsto de chegada de cada voo.

A TABELA 17 apresenta o número de voos e a taxa de atraso dos 3 anos analisados.

TABELA 17 – Número de voos, número de voos atrasados e taxa de atraso de cada ano

Ano	Número de voos	Número de voos atrasados	Taxa de atraso
2019	97.682	15.126	15,48%

2020	56.407	5.913	10,48%
2021	68.725	6.857	9,98%

Houve uma queda no número de voos nos anos de 2020 e 2021 em relação a 2019 devido à pandemia de COVID19, que causou uma crise sanitária que afetou o transporte aéreo em todo o mundo. Em contrapartida, houve uma queda na taxa de atraso ao longo dos 3 anos. Embora essa variação na taxa de atraso não seja alta, a variável *year* foi adotada nos modelos de previsão de atraso do aeroporto em análise porque foi considerada em muitos modelos pesquisados na literatura.

3.4 Variáveis metereológicas

Nas subseções que seguem serão destacadas as variáveis referentes às questões metereológicas abordadas no presente trabalho.

3.4.1 Temperatura

A variável *temperature* indica a temperatura em graus Fahrenheit medida na hora correspondente ao horário previsto para a chegada dos voos da base de dados da ANAC. Os dados referentes a temperatura e a demais condições meteorológicas foram extraídos da base de dados da Universidade Estadual de Iowa.

Pode-se dividir a temperatura em 6 classes distintas, conforme indica a TABELA 18, que apresenta o número de voos operados em cada uma das classes, dado em porcentagem em relação ao número total de voos.

TABELA 18 – Número de voos de acordo com a classe de temperatura

Classe de temperatura em Fahrenheit	Número de voos (em %)
$[35{,}6, 45[$	0,34%
$[45, 55[$	3,14%

[55, 65[	27,23%
[65, 75[	42,03%
[75, 85[	21,80%
[85, 98,6]	5,46%

A partir da TABELA 18, conclui-se que a classe de temperatura mais comum é a [65, 75] graus Fahrenheit, que corresponde a cerca de 42,03% dos voos. Essa faixa equivale a [18,3, 23,9] graus Celsius, a qual inclui a temperatura média de Guarulhos, 19,2 graus Celsius.

O gráfico da FIGURA 20 apresenta a variação da taxa de atraso ao longo das classes de temperatura. A classe 4 é a classe com maior taxa de atraso, com 15,10% de voos atrasados, enquanto a classe 1 é a classe com menor taxa de atraso, com apenas 7,70%. Da classe 1 à classe 4, a taxa de atraso aumenta à medida que a temperatura aumenta; ao passo que, da classe 4 à classe 6, a taxa diminui à medida que a temperatura aumenta.

Embora não haja uma ampla variação das taxas de atraso ao longo das classes de temperatura, a variável *temperature* será incluída nos modelos de previsão de atraso do aeroporto em análise, pois foi incluída em muitos modelos pesquisados na literatura.

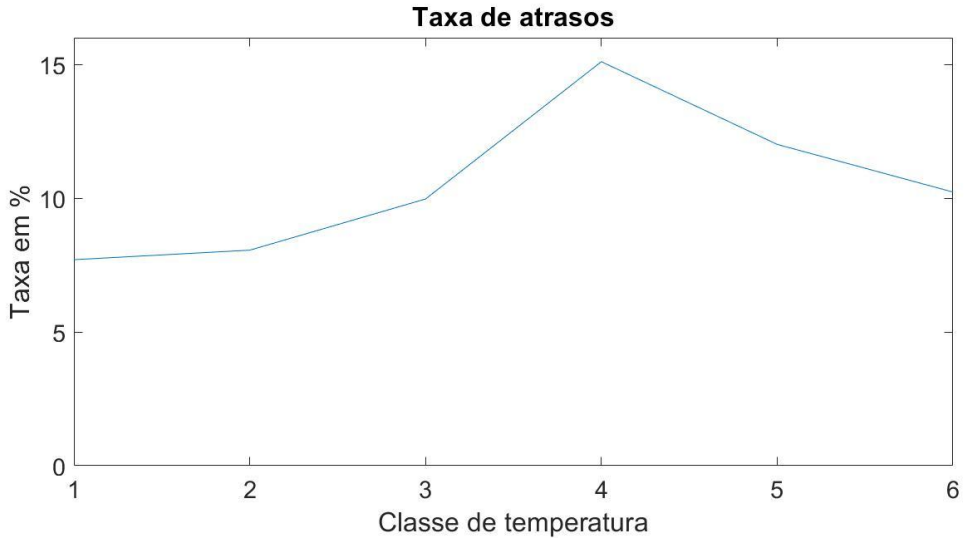

FIGURA 20 – Taxas de atraso de acordo com as classes de temperatura

3.4.2 Pressão

A variável *pressure* indica a pressão atmosférica em polegadas medida na hora

correspondente ao horário previsto para a chegada dos voos da base de dados da ANAC.

Pode-se dividir a pressão em 4 classes distintas, conforme indica a TABELA 19, que apresenta o número de voos operados em cada uma das classes, dado em porcentagem em relação ao número total de voos.

TABELA 19 – Número de voos de acordo com a classe de pressão

Classe de pressão em polegadas	Número de voos (em %)
[27,23, 29,75[	0,08%
[29,75, 30[	24,33%
[30, 30,25[	67,63%
[30,25, 30,47]	7,96%

A partir da TABELA 19, conclui-se que a classe mais comum é a [30, 30,25] polegadas, com 67,63% dos voos. Em apenas 0,08% dos voos a pressão atmosférica estava abaixo das 29,75 polegadas.

O gráfico da FIGURA 21 apresenta a taxa de atraso em função das classes de pressão. Excluindo-se a classe 1, que corresponde a apenas 0,08% dos voos, a taxa de atraso cai à medida que se aumenta a pressão. A taxa varia de 8,39% na classe 4 a 15,57% na classe 2.

Embora não haja uma ampla variação nas taxas de atraso de acordo com as classes de pressão, a variável *pressure* será incluída nos modelos de previsão de atraso do aeroporto em análise, pois foi incluída em muitos modelos pesquisados na literatura.

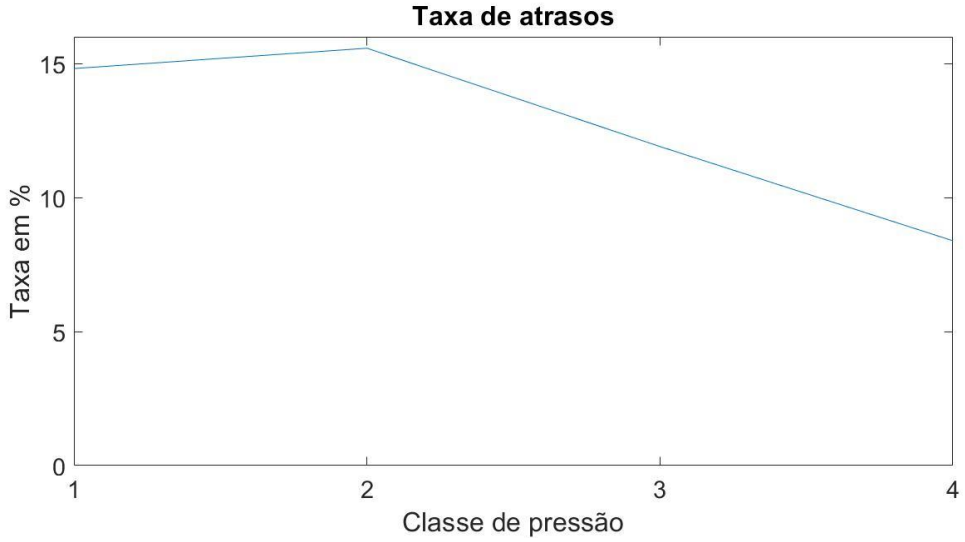

FIGURA 21 – Taxas de atraso de acordo com as classes de pressão

3.4.3 Precipitação

Um *Meteorological Aerodrome Report* (Metar) é uma observação meteorológica codificada, que fornece informações sobre o tempo atual em algum aeródromo. A base de dados da Universidade Estadual de Iowa traz o Metar do Aeroporto Internacional de Guarulhos (METAR, 2022). A partir do Metar, é possível concluir se houve ou não precipitação em determinado horário na região ao redor do aeroporto. Os diversos tipos de precipitação são indicados pelos códigos apresentados na TABELA 20 (DEPARTAMENTO DE CONTROLE DO ESPAÇO AÉREO, 2022).

TABELA 20 – Tipos de precipitação decodificados a partir do METAR

Código	Tipo de Precipitação
RA	Chuva *(rain)*
SH	Pancada de chuva *(rain shower)*
TS	Trovoada, raios e relâmpagos *(thunderstorm)*
GR	granizo

A variável *rain* é binária e indica se houve algum dos tipos de precipitação indicados na TABELA 20 no horário previsto para o pouso do voo.

Dos 222.814 voos analisados, em 21.574 deles, ou aproximadamente 9,68% do total de voos, ocorreu precipitação durante o horário previsto para a chegada. Houve atraso em 11,29% dos voos onde não houve precipitação e em 24,04% dos voos onde houve precipitação. Essa ampla diferença indica que a precipitação pode exercer influência considerável sobre os atrasos, logo a variável *rain* foi incluída nos modelos de previsão de atraso do aeroporto em análise.

3.4.4 Visibilidade horizontal

De acordo com a ANAC (2022a), visibilidade horizontal é a distância máxima na qual um objeto pode ser visto e identificado, quando situado nas proximidades do grande plano horizontal onde ele próprio se encontra. A base de dados da Universidade Estadual de Iowa informa diretamente a visibilidade horizontal em milhas. A variável *v_h* indica a visibilidade

horizontal do horário previsto de chegada do voo correspondente na base de dados.

Pode-se dividir a visibilidade horizontal nas 7 classes indicadas na TABELA 21, que apresenta o número de voos operados em cada uma das classes.

TABELA 21 – Número de voos (em porcentagem) de acordo com a classe de visibilidade horizontal

Classe de visibilidade horizontal (em milhas)	Número de voos (em %)
[0,06, 1[	0,86%
[1, 2[	1,03%
[2, 3[	2,27%
[3, 4[	5,28%
[4, 5[	9,03%
[5, 6[	4,44%
[6, 6,21]	77,09%

A TABELA 21 indica que a maior parte dos voos – cerca de 77,09% - tiveram horário previsto para pouso em horários com alta visibilidade horizontal – acima de 6 milhas. O gráfico da FIGURA 22**Erro! Fonte de referência não encontrada.** apresenta as taxas de atraso de acordo com a classe de visibilidade horizontal.

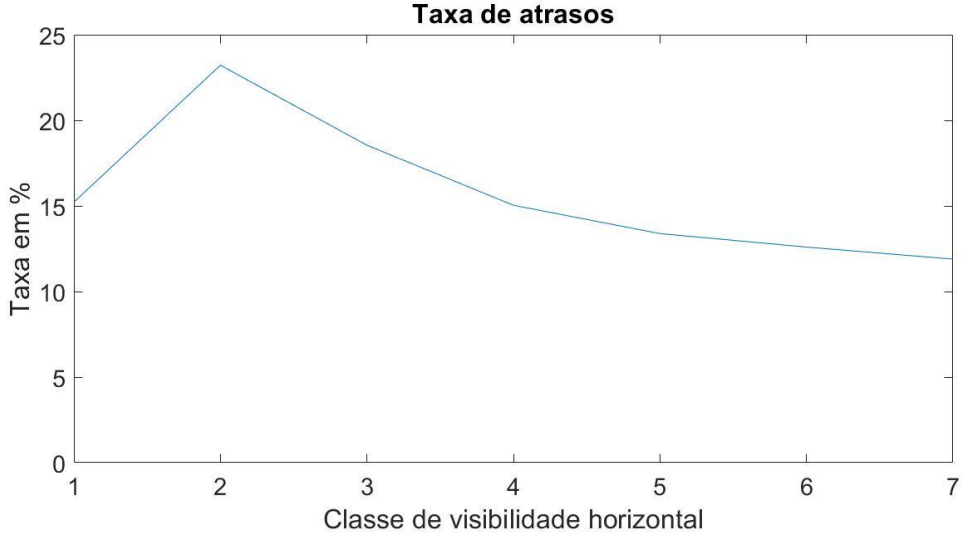

FIGURA 22 – Taxas de atraso de acordo com a classe de visibilidade horizontal

A análise do gráfico da FIGURA 22 permite concluir que há uma tendência de

aumento de atrasos à medida que se diminui a visibilidade horizontal, o que era esperado, visto que a baixa visibilidade horizontal prejudica as operações das aeronaves. A classe com maior taxa de atraso é a 2, com 23,22%, e a com menor taxa de atraso é a 7, com apenas 11,89% de atrasos.

A ampla variação das taxas de atraso de acordo com as classes de visibilidade horizontal indica que a variável *v_h* pode exercer influência considerável sobre os atrasos, logo ela foi incluída nos modelos de previsão de atraso do aeroporto em análise.

3.4.5 Visibilidade vertical

A variável *v_v* indica a visibilidade vertical do horário previsto para a chegada do voo correspondente na base de dados. De acordo com a ANAC, essa grandeza pode ser definida como a distância máxima em que um observador pode ver e identificar um objeto na mesma vertical que ele próprio, tanto acima como abaixo.

A base de dados analisada não informa diretamente a visibilidade vertical, entretanto ela pode ser definida como a mínima altitude da base de nuvens nas condições BKN *(broken)* e OVC *(overcast)*, podendo chegar ao limite de 10.000 pés. Nas condições FEW *(few clouds)*, SCT *(scattered clouds)*, CAVOK e NCD, o valor máximo de visibilidade vertical foi adotado.

A visibilidade vertical pode ser dividida em 10 classes, conforme indica a TABELA 22, que apresenta o número de voos em porcentagem de cada uma das classes.

TABELA 22 – Número de voos (em porcentagem) de acordo com as classes de visibilidade vertical

Classe de visibilidade vertical (em pés)	Número de voos (em %)
$[0, 1000[$	11,14%
$[1000, 2000[$	18,94%
$[2000, 3000[$	7,20%
$[3000, 4000[$	4,15%
$[4000, 5000[$	3,71%
$[5000, 6000[$	0,52%
$[6000, 7000[$	0,17%
$[7000, 8000[$	0,61%

[8000, 9000[	1,18%
[9000, 10000[	52,37%

A TABELA 22 indica que a maior parte dos voos – cerca de 52,37% - teve horário previsto de chegada com alta visibilidade – acima de 9 mil pés.

O gráfico da FIGURA 23 apresenta as taxas de atraso segundo a classe de visibilidade vertical.

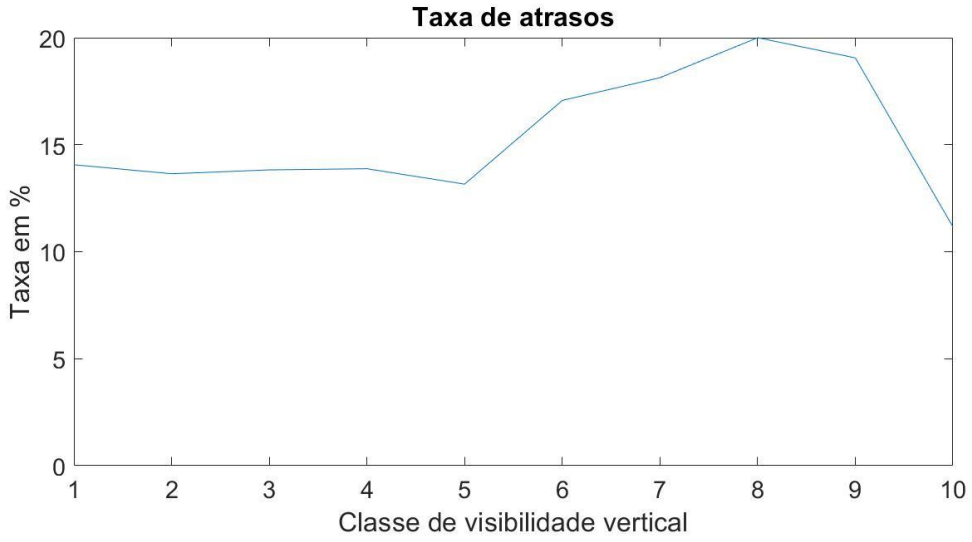

FIGURA 23 – Taxas de atraso de acordo com a classe de visibilidade vertical

A taxa de atraso varia de 11,15% na classe 10 a 20,00% na classe 8. Esperava-se que houvesse queda das taxas de atraso à medida que aumentasse a visibilidade vertical, contudo o gráfico da FIGURA 23 não ratifica o comportamento esperado.

A ampla variação das taxas de atraso de acordo com as classes de visibilidade vertical indica que a variável *v_v* pode exercer influência considerável sobre os atrasos, logo ela foi incluída nos modelos de previsão de atraso do aeroporto em análise.

3.4.6 Vento

De acordo com o Regulamento Brasileiro da Aviação Civil, deve ser assumido que o pouso ou decolagem de aeronaves são, em circunstâncias normais, comprometidos, quando o componente de vento de través – componente de vento de superfície em ângulo reto ao eixo da pista – exceder (BRASIL, 2021):

- 37 km/h (20 kt), no caso de aeronaves cujo comprimento básico de pista é maior ou igual a 1.500 m, exceto quando houver, com certa frequência, uma baixa ação de frenagem na pista devido a um coeficiente de atrito longitudinal insuficiente, quando, então, deve ser assumido um componente de vento de través que não exceda 24 km/h (13 kt);
- 24 km/h (13 kt), no caso de aeronaves cujo comprimento básico de pista é maior ou igual a 1.200 m e menor que 1.500 m;
- 19 km/h (10 kt), no caso de aeronaves cujo comprimento básico de pista for menor que 1.200 m.

Devido ao comprometimento das operações de pouso e de decolagem que ocorrem quando há ventos de través com velocidades muito elevadas, os ventos podem ocasionar atrasos e inclusive cancelamentos de voos. Não só a magnitude da velocidade do vento importa, mas também a sua direção, visto que é o vento de través que pode comprometer as operações das aeronaves. 3 variáveis relacionadas ao vento foram consideradas no modelo de previsão de atraso:

- *speed*, que indica a velocidade do vento em nós.
- *direction,* que indica a direção do vento em graus em relação ao norte verdadeiro.
- *gust*, que informa as rajadas de vento em nós. Se o Metar não menciona rajadas, a variável assume valor nulo.

3.4.7 Velocidade do vento

A variável *speed* indica a velocidade do vento em nós e assume números inteiros no intervalo $[0, 24]$. O gráfico da FIGURA 24 apresenta o número de voos operados em porcentagem em relação ao número total de voos, para cada valor de velocidade de vento.

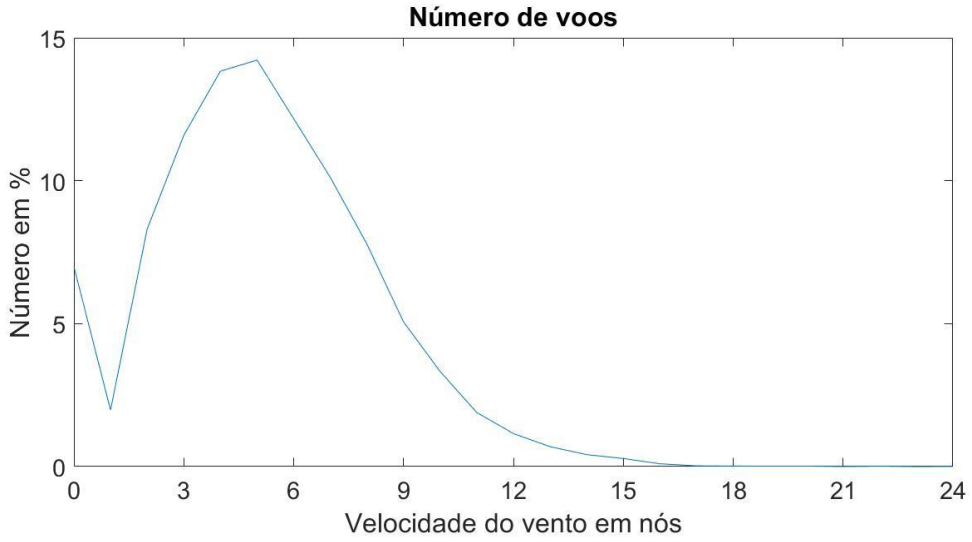

FIGURA 24 – Número de voos (em porcentagem) de acordo com a velocidade do vento

Cerca de 14,22% dos voos tiveram como horário previsto de chegada um horário com velocidade de vento igual a 5 nós, ao passo que em apenas 2,74% dos voos a velocidade do vento foi igual ou superior a 12 nós.

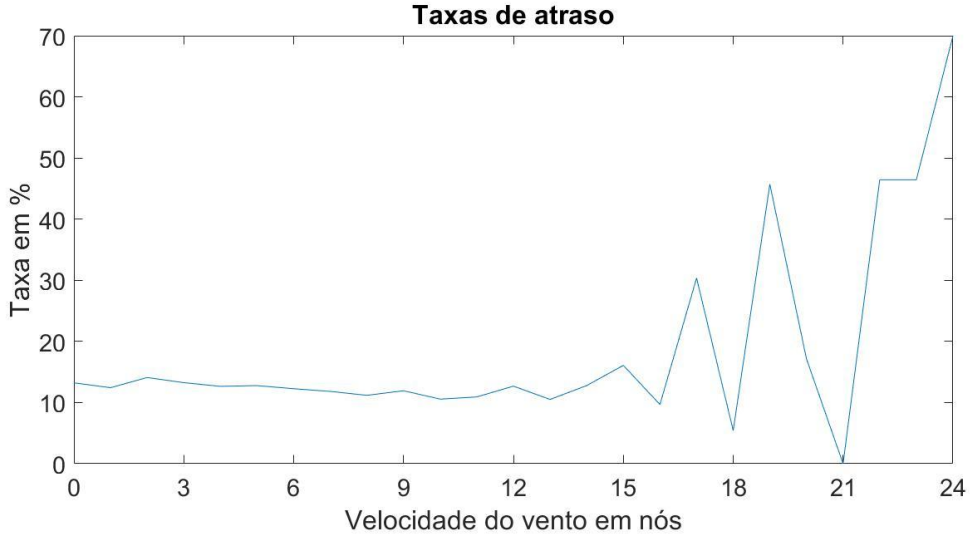

FIGURA 25 – Taxas de atraso de acordo com a velocidade do vento

O gráfico da FIGURA 25 indica que a partir de 16 nós, a taxa de atraso oscila até atingir um valor máximo de 70,00% à velocidade de 24 nós. A ampla variação das taxas de atraso em função da velocidade do vento indica que a variável *speed* pode exercer influência considerável sobre os atrasos, portanto essa variável será considerada nos modelos de

previsão de atraso do aeroporto em análise.

3.4.8 Rajadas de vento

Dos 222.814 voos em análise, em apenas 1971 deles, ou cerca de 0,89% do total, houve rajadas de vento no horário previsto de chegada do voo. Os valores das rajadas variam de 15 a 39 nós.

Houve atraso em 19,48% dos voos em que houve rajadas e em 12,46% dos voos em que não houve rajadas de vento. Essa diferença indica que as rajadas de vento podem exercer influência considerável sobre os atrasos no Aeroporto Internacional de Guarulhos, logo a variável *gust* foi incluída nos modelos de previsão de atraso do aeroporto em análise.

3.4.9 Direção do vento

A variável *direction* indica a direção de onde o vento sopra em graus em relação ao norte verdadeiro e assume valores inteiros múltiplos de 10 no intervalo $[0, 360]$. O gráfico da FIGURA 26 indica o número de voos operados para cada valor de direção do vento.

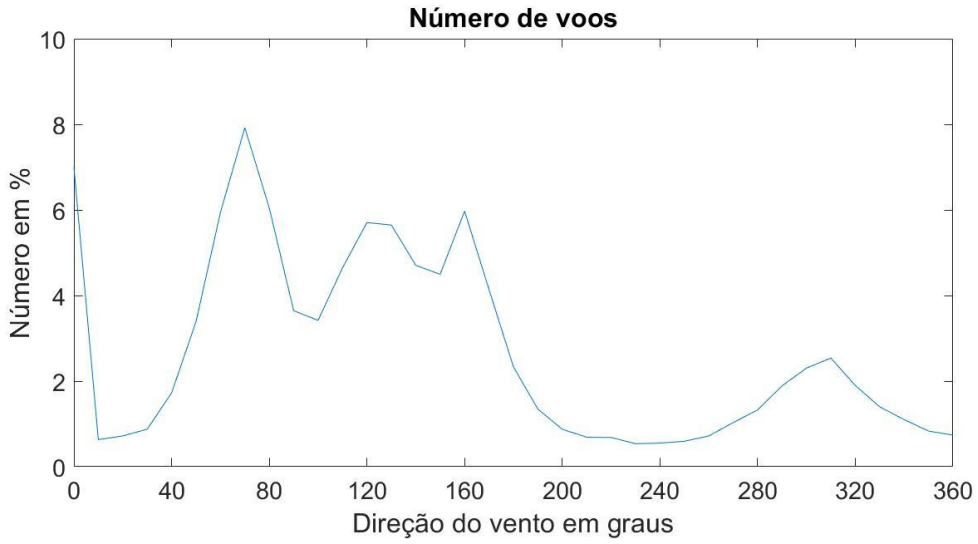

FIGURA 26 – Número de voos (em porcentagem) de acordo com a direção do vento

O gráfico da FIGURA 26 indica que a maior parte dos voos tiveram horário previsto de chegada com o vento soprando das direções de 50 a 170 graus. De fato, em 65,67% dos voos, o vento soprava desse intervalo de direção.

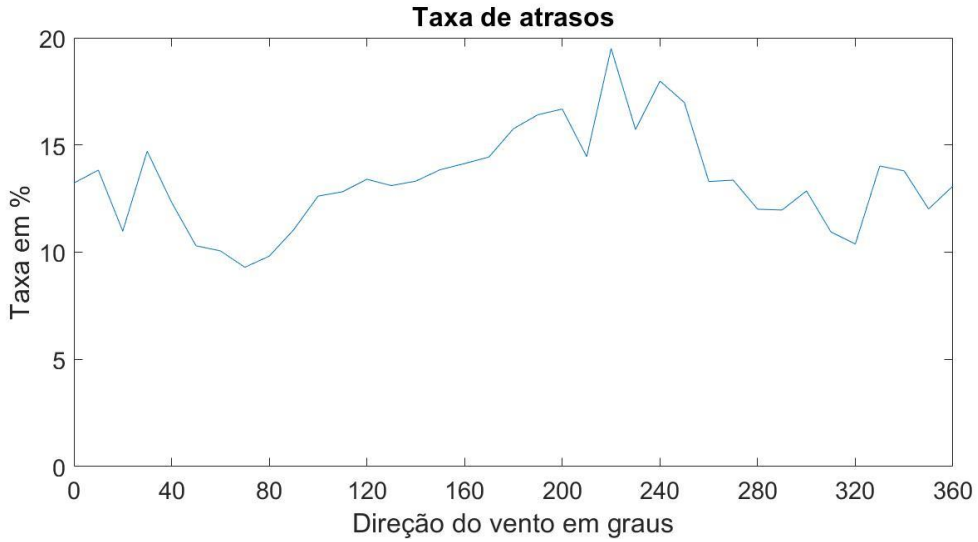

FIGURA 27 – Taxas de atraso de acordo com a direção do vento.

O gráfico da FIGURA 27 apresenta as taxas de atraso em função da direção do vento. As maiores taxas de atraso ocorrem nas direções de 200 a 250 graus, com o valor máximo ocorrendo para 220 graus e sendo igual a 19,50%. As menores taxas de atraso, por sua vez, ocorrem nas direções de 40 a 80 graus, com valor mínimo ocorrendo para 70 graus e sendo igual a 9,30%.

A ampla variação das taxas de atraso em função da direção do vento indica que a variável *direction* pode exercer influência considerável sobre os atrasos, logo essa variável foi considerada nos modelos de previsão de atraso do aeroporto em análise.

3.5 Teste dos algoritmos de *machine learning*

Os algoritmos de aprendizado de máquina supervisionado utilizados neste estudo são descritos nas seções a seguir, de modo que eles totalizam 4 algoritmos de machine learning, quais sejam: KNN, árvore de decisão, regressão logística e rede neural artificial. A utilização desses algoritmos teve o intuito de elaborar um modelo preditivo para os atrasos no aeroporto de Guarulhos.

3.5.1 KNN

O KNN é um algoritmo de aprendizado de máquina supervisionado que requer a escolha de uma função de distância e de um parâmetro natural não-nulo *k*. É um algoritmo

que gera fronteiras de decisão com formatos arbitrários.

Qualquer função de distância que satisfaça às condições descritas a seguir pode ser chamada de métrica de distância e pode ser utilizada para a aplicação do algoritmo KNN. Dadas *x*, *y* e *z* observações $\in \mathbb{R}^p$, tem-se que:

a) A distância entre *x* e *y* é sempre um valor maior ou igual a zero:

$$d(x, y) \geq 0$$

b) A distância entre *x* e *y* é nula se e somente se *x* for igual a *y*:

$$d(x, y) = 0 \leftrightarrow x = y$$

c) A distância entre *x* e *y* é igual à distância entre *y* e *x*:

$$d(x, y) = d(y, x)$$

d) A desigualdade triangular deve ser satisfeita:

$$d(x, y) \leq d(x, z) + d(z, y)$$

Três das métricas de distância mais utilizadas são a distância de Manhattan, a distância euclidiana e a distância de Minkowski. A distância de Manhattan é calculada de acordo com a seguinte equação:

$$d(x, y) = \sum_{i=1}^{p} |x_i - y_i|$$

A distância euclidiana, por sua vez, é calculada de acordo com a equação a seguir:

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

A distância de Manhattan e a distância euclidiana são casos particulares da distância de Minkowski. Sendo *c* um número real maior ou igual a 1, essa métrica pode ser calculada por meio da seguinte equação:

$$d(x, y) = (\sum_{i=1}^{p} |x_i - y_i|^c)^{1/c}$$

Sendo adotada uma métrica de distância e sendo dada uma nova observação $x^* \in \mathbb{R}^p$, o algoritmo KNN funciona da seguinte maneira:

- É computada a distância entre x^* e todas as observações de treinamento;
- As distâncias são dispostas em ordem crescente;
- Selecionam-se os *k* vizinhos mais próximos com base nas distâncias em ordem crescente;

- Prevê-se y^*como a classe que mais se repete entre os *k* vizinhos mais próximos, de acordo com a regra da maioria.

É importante normalizar os dados se as variáveis independentes estiverem em diferentes escalas, com o intuito de prevenir que as medidas de distância sejam dominadas por uma das variáveis.

O parâmetro *k* e a métrica de distância são tipicamente escolhidos por meio de validação cruzada.

3.5.2 Árvores de decisão

Árvores de decisão são estruturas hierárquicas formadas por uma série de regras usadas para tomar uma decisão. O algoritmo de aprendizado das árvores de decisão é comumente chamado de CART. Ele utiliza os dados de treinamento para construir a árvore, a qual é utilizada para prever o output *y* das observações teste *x*.

As árvores são construídas por meio de partição recursiva: o conjunto de treinamento é dividido repetidamente em 2 partes com o intuito de obter a máxima homogeneidade das novas partes. Para isso, é necessário utilizar uma métrica para avaliar a qualidade de cada divisão. Essa métrica é o índice Gini, que mede a impureza de um nó. Dado um nó com *n* observações, em que n_c observações pertencem à classe C, o índice de Gini para esse nó é dado pela expressão:

$$G = 1 - \sum_{C} (\frac{n_c}{n})^2$$

A árvore é, portanto, construída de tal forma que seja minimizada a impureza das partições em cada divisão.

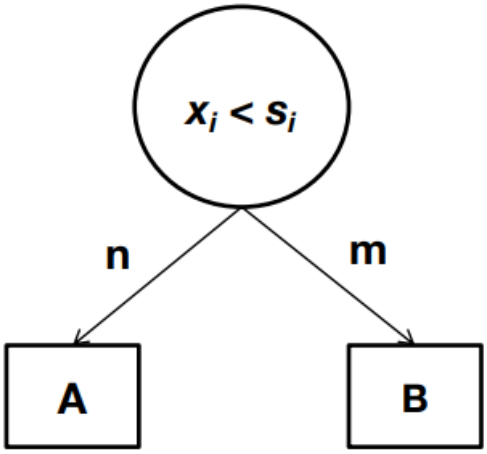

FIGURA 28 – As árvores são construídas por meio de partição recursiva

Na partição indicada na FIGURA 28, deve-se minimizar a expressão $\frac{n}{n+m}G(A) + \frac{m}{n+m}G(B)$.

O fim natural da partição recursiva ocorre quando se atinge 100% de pureza em cada nó. Entretanto, esse processo pode gerar uma árvore muito complexa, com uma péssima generalização, enquanto a árvore ideal não deve ser grande demais e deve se ajustar aos dados de treinamento razoavelmente. Para evitar árvores muito complexas, podem ser utilizados os dados de validação para podar a árvore. A poda pode ocorrer durante a construção da árvore, determinando-se que cada nó deve ter um número mínimo de observações, ou após a sua construção, combinando-se os nós para melhorar a performance na validação.

3.5.3 Regressão logística

A regressão logística modela a probabilidade de que uma observação pertença a uma determinada classe usando uma função logística. Dado um problema de classificação binário, onde $x \in \mathbb{R}^p$ e $y \in \{0,1\}$, a probabilidade condicional de que a observação pertença à classe 1 dado *x* é escrita da seguinte maneira:

$$p = P(y = 1|x) = sigmoid(\beta^T x + \beta_0) = \frac{1}{1 + e^{-(\beta^T x + \beta_0)}}$$

Os parâmetros β podem ser aprendidos a partir dos dados de treinamento por meio do método da estimativa da probabilidade máxima (MLE):

$$L = \prod_{i=1}^{n} (P(y_i = 1|x_i))^{y_i} (P(y_i = 0|x_i))^{(1-y_i)}$$

Uma nova observação x^* pertencerá à classe 1 se:

$$y^* = 1 \leftrightarrow \frac{P(y^* = 1 \mid x^*)}{P(y^* = 0 \mid x^*)} > 1 \leftrightarrow \log\left(\frac{P(y^* = 1|x^*)}{P(y^* = 0|x^*)}\right) > 0 \leftrightarrow \log\left(\frac{\frac{1}{1 + e^{-(\beta^T x^* + \beta_0)}}}{\frac{e^{-(\beta^T x^* + \beta_0)}}{1 + e^{-(\beta^T x^* + \beta_0)}}}\right) > 0$$

$$\leftrightarrow \beta^T x^* + \beta_0 > 0$$

A regressão logística gera, portanto, fronteiras de decisão lineares.

3.5.4 Redes neurais artificiais

As redes neurais artificiais são inspiradas no comportamento do cérebro como uma rede de unidades chamadas de neurônios. Estima-se que o cérebro humano tenha mais de 10 bilhões de neurônios, cada um deles conectado em média com 10 mil outros.

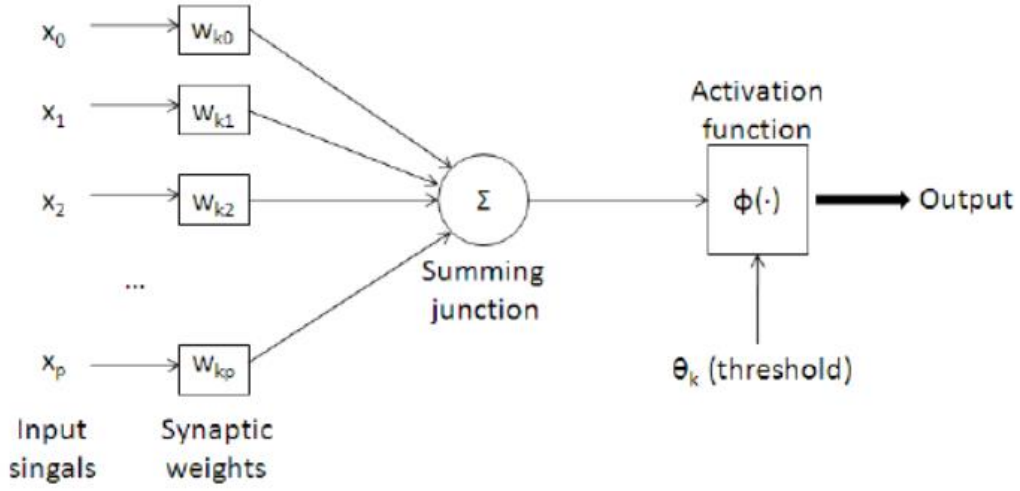

FIGURA 29 – Modelo de um neurônio artificial. Fonte: (MARTÍNEZ-ÁLVAREZ *et al.*, 2015)

A FIGURA 29 apresenta um modelo matemático de um neurônio artificial. x_1, x_2, ..., x_p são os inputs; w_{k1}, w_{k2}, ..., w_{kp}, os pesos e o w_{k0}, o viés. O neurônio computa a seguinte soma:

$$z_j = \sum_{i=1}^{p} w_{ki} x_i + w_{k0}$$

O output é o resultado da aplicação da função de ativação à soma z_j:

$$output = \emptyset(z_j)$$

Redes neurais artificiais são compostas por uma sequência de camadas, com cada camada contendo neurônios artificiais. Na camada de input, o número de neurônios é tipicamente o número de variáveis no input. Cada neurônio transfere o valor da variável de input para cada neurônio na próxima camada, tipicamente a camada escondida.

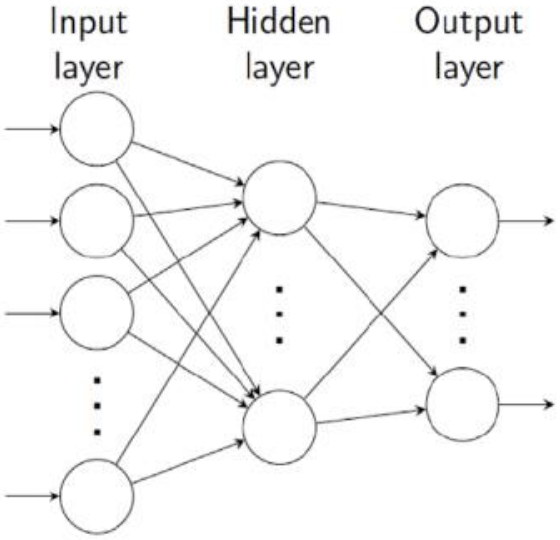

FIGURA 30 – Ilustração das camadas de uma rede neural artificial

Na camada escondida, o número de neurônios é baseado na validação cruzada. Cada neurônio nessa camada transfere a soma ponderada dos inputs, transformada com uma função de ativação, para cada neurônio na próxima camada (camada escondida ou de output).

A camada de output, por sua vez, tem apenas um neurônio nos casos de classificação binária. Funciona da mesma forma que a camada escondida e o seu output é o output previsto pela rede.

Várias funções de ativação podem ser utilizadas, permitindo capturar relações não-lineares entre as variáveis. Neste estudo, foi utilizada a função tangente hiperbólica:

$$\tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

3.6 Avaliação da performance preditiva: Problema de classificação e Problema de regressão

O objetivo dos algoritmos de aprendizado de máquina supervisionado é fazer previsões. Consequentemente, são necessárias métricas para avaliar a performance preditiva desses algoritmos. Nesse aspecto, tratando-se do problema de classificação, dado um

problema de classificação binária, em que as observações são classificadas em duas classes - C_1 e C_2 - a matriz confusão $N = (n_{ij})_{2x2}$ é definida da seguinte forma:

		Predicted class	
		C_1	C_2
Actual class	C_1	n_{11}	n_{12}
	C_2	n_{21}	n_{22}

FIGURA 31 – Matriz confusão de um problema de classificação binário

- n_{11}: Número de casos C_1 corretamente classificados;
- n_{12}: Número de casos C_1 incorretamente classificados como C_2;
- n_{21}: Número de casos C_2 incorretamente classificados como C_1;
- n_{22}: Número de casos C_2 corretamente classificados.

A acurácia é definida como:

$$acurácia = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}}$$

No caso deste estudo, os dados são desbalanceados, ou seja, há muito mais voos sem atraso do que voos com atraso. Consequentemente, além da medida de acurácia, são necessárias outras medidas. Sendo a classe 1 a classe dos voos com atraso e a classe 0 a classe dos voos sem atraso, a matriz confusão pode ser representada da seguinte maneira:

		Predicted class	
		1	0
Actual class	1	True Positive (tp)	False Negative (fn)
	0	False Positive (fp)	True Negative (tn)

FIGURA 32 – Matriz confusão utilizada no cálculo da sensibilidade e da precisão

A sensibilidade ou *recall* é dada por:

$$sensibilidade = \frac{tp}{tp + fn}$$

A precisão, por sua vez, é calculada por meio da expressão a seguir:

$$precisão = \frac{tp}{tp + fp}$$

O *F1 Score* é uma métrica calculada por meio da seguinte equação:

$$F_1 = \frac{1}{sensibilidade^{-1} + precisão^{-1}}$$

Ademais, quanto ao problema de regressão, dadas *n* observações $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, onde *x* denota um *input* e *y* denota um *output*, o objetivo dos algoritmos de regressão é aprender uma função de mapeamento *f* que preveja a variável real *y* para um novo exemplo *x*. Para avaliar a performance preditiva desses algoritmos, podem-se computar algumas métricas de erro, como RMSE (*root mean squared error*), MAE (*mean absolut* error) e MAPE (*mean absolute percentage error*).

Sendo $\widehat{y_j}$ o valor previsto; y_j, o valor real e n, o número de dados da base de teste, as métricas de erro podem ser calculadas por meio das seguintes expressões:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \widehat{y_j})^2}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \widehat{y_j}|$$

$$MAPE = \frac{1}{n}\sum_{j=1}^{n}\left|\frac{y_j - \widehat{y_j}}{y_j}\right|$$

O erro MAPE não pode ser aplicado a todos os dados da base de teste, visto que há muitos dados com valor real nulo. Portanto, ele será aplicado apenas aos dados referentes a voos que sofreram atraso.

3.6.1 Modelos preditivos de atraso

Neste trabalho, os algoritmos de aprendizado de máquina supervisionado foram utilizados para elaborar modelos preditivos de atraso para o Aeroporto Internacional de Guarulhos.

Foram analisados 3 tipos de modelos distintos, com base no tipo da variável *delay_a_bin* (atraso na chegada*)* e *delay_d_bin* (atraso na partida). O modelo 1 é de classificação, com as variáveis de atraso binárias. A variável *delay_a_bin* é definida da seguinte maneira:

$$delay_{a_{bin}} = 1 \leftrightarrow houve\ atraso\ na\ chegada$$

$$delay_{a_{bin}} = 0 \leftrightarrow não\ houve\ atraso\ na\ chegada$$

A variável *delay_d_bin*, por sua vez, é definida de maneira análoga:

$$delay_{d_{bin}} = 1 \leftrightarrow houve\ atraso\ na\ partida$$

$$delay_{d_{bin}} = 0 \leftrightarrow não\ houve\ atraso\ na\ partida$$

O modelo 2 é de classificação, com *delay_a_bin* binária e *delay_d_bin* real. O modelo 3, por sua vez, é de regressão, com ambas as variáveis reais.

TABELA 23 – Modelos propostos

Variável	Modelo 1	Modelo 2	Modelo 3
delay_a_bin	Binária	Binária	Real
delay_d_bin	Binária	Real	Real

Além da variável *delay_d_bin*, os modelos propostos têm mais 26 variáveis de *input*, apresentadas a seguir.

a) *airline:* Variável numérica (definida por meio de *target encoding*) que indica a companhia aérea que operou o voo.

b) *aircraft*: Variável numérica (definida por meio de *target encoding*) que indica o modelo da aeronave com a qual foi operado o voo.

c) *seats:* Variável numérica que indica o número de assentos disponíveis na aeronave com a qual foi operado o voo.

d) *airport:* Variável numérica (definida por meio de *target encoding*) que indica o aeroporto de origem do voo.

e) *distance:* Variável numérica que indica a distância em km do aeroporto de origem do voo em relação ao aeroporto de destino (o Aeroporto Internacional de Guarulhos).

f) *sch_d:* Variável numérica que indica o número de pousos e decolagens no Aeroporto Internacional de Guarulhos no mesmo dia que o pouso do voo correspondente.

g) *sch_h:* Variável numérica que indica o número de pousos e decolagens no Aeroporto Internacional de Guarulhos no intervalo de tempo compreendido entre 1 hora antes e 1 hora depois do horário previsto para o pouso do voo correspondente.

h) *hour_cos*: Variável trigonométrica associada ao horário previsto de chegada do voo correspondente no Aeroporto Internacional de Guarulhos.

i) *hour_sin:* Variável trigonométrica associada ao horário previsto de chegada do voo correspondente no Aeroporto Internacional de Guarulhos.

j) *day_week_cos:* Variável trigonométrica associada ao dia da semana do horário previsto de chegada do voo correspondente no Aeroporto Internacional de Guarulhos.

k) *day_week_sin:* Variável trigonométrica associada ao dia da semana do horário previsto de chegada do voo correspondente no Aeroporto Internacional de Guarulhos.

l) *day_month:* Variável numérica que indica o dia do mês do horário previsto de chegada do voo correspondente.

m) *day_month_cos:* Variável trigonométrica associada ao dia do mês do horário previsto de chegada do voo correspondente.

n) *day_month_sin:* Variável trigonométrica associada ao dia do mês do horário previsto de chegada do voo correspondente.

o) *season:* Variável numérica (definida por meio de *target encoding*) que indica o mês associado ao horário previsto de pouso do voo correspondente.

p) *month_cos:* Variável trigonométrica associada ao mês referente ao horário previsto de pouso do voo correspondente.

q) *month_sin:* Variável trigonométrica associada ao mês referente ao horário previsto de pouso do voo correspondente.

r) *year:* Variável numérica que indica o ano referente ao horário previsto de pouso do voo correspondente.

s) *temperature:* Variável numérica que indica a temperatura em graus Fahrenheit no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

t) *pressure:* Variável numérica que indica a pressão atmosférica em polegadas no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

u) *rain:* Variável categórica binária que indica se houve precipitação atmosférica no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

$$rain = 1 \leftrightarrow houve\ precipitação$$

$$rain = 0 \leftrightarrow não\ houve\ precipitação$$

v) *v_h:* Variável numérica que indica a visibilidade horizontal em milhas no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

w) *v_v:* Variável numérica que indica a visibilidade vertical em pés no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

x) *speed:* Variável numérica que indica a velocidade do vento em nós no horário previsto de pouso do voo no Aeroporto Internacional de Gaurulhos.

y) *gust:* Variável numérica que indica valores de rajadas de vento em nós no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos. Se não foram registradas rajadas, seu valor é nulo.

z) *direction:* Variável numérica que indica o valor da direção do vento em graus em relação ao norte verdadeiro no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

No caso deste estudo, os dados são desbalanceados, ou seja, há muito mais voos sem atraso do que voos com atraso, sendo estes classificados como eventos raros. Dos 222.814 voos analisados, 27.896 sofreram atraso na chegada, totalizando apenas cerca de 12,52% do total de voos.

Para mitigar o problema do desbalanceamento dos dados, foram introduzidos pesos associados às observações, definidos a partir das proporções das classes da variável *delay_a_bin*: os pesos são numericamente iguais ao inverso das proporções das classes da variável de *output*.

Com o intuito de evitar uma péssima generalização dos modelos, foram efetuados processos de validação cruzada na elaboração de todos os modelos analisados, sempre com os dados sendo particionados com as proporções indicadas na TABELA 24.

TABELA 24 – Partição dos dados para a validação cruzada

Dados	Porcentagem em relação ao total
Treinamento	56,25%
Validação	18,75%
Teste	25%

4 RESULTADOS: TESTE DOS ALGORITMOS

Nas seções do presente capítulo há uma descrição minuciosa dos quatro algoritmos testados, conforme o proposto na introdução do trabalho.

4.1 Modelo 1 de classificação

Nas subseções que seguem, será demonstrado o teste do Modelo 1 em cada um dos 4 algoritmos analisados na pesquisa.

4.1.1 KNN: Modelo 1 de classificação

Para a implementação do método KNN, foi utilizada a função *fitcknn* do *software* Matlab. Para a validação cruzada, foram considerados os valores de *k* pertencentes ao conjunto {1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50} e duas métricas de distância: a euclidiana e a de Manhattan. A FIGURA 33 e a FIGURA 34 apresentam os valores de acurácia e de *F1 Score* dos diferentes modelos testados com os dados de validação.

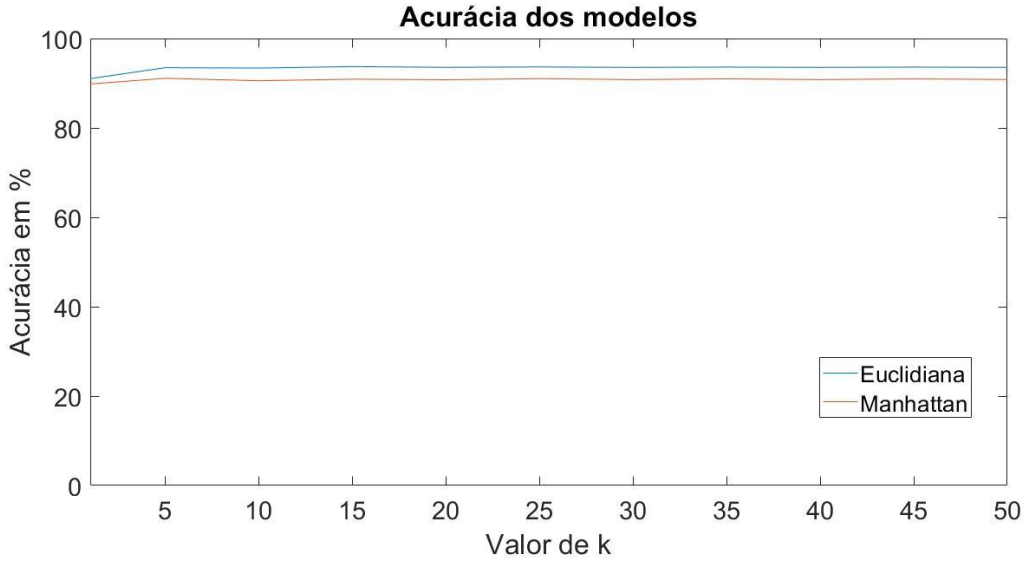

FIGURA 33 – Acurácia dos modelos da validação cruzada

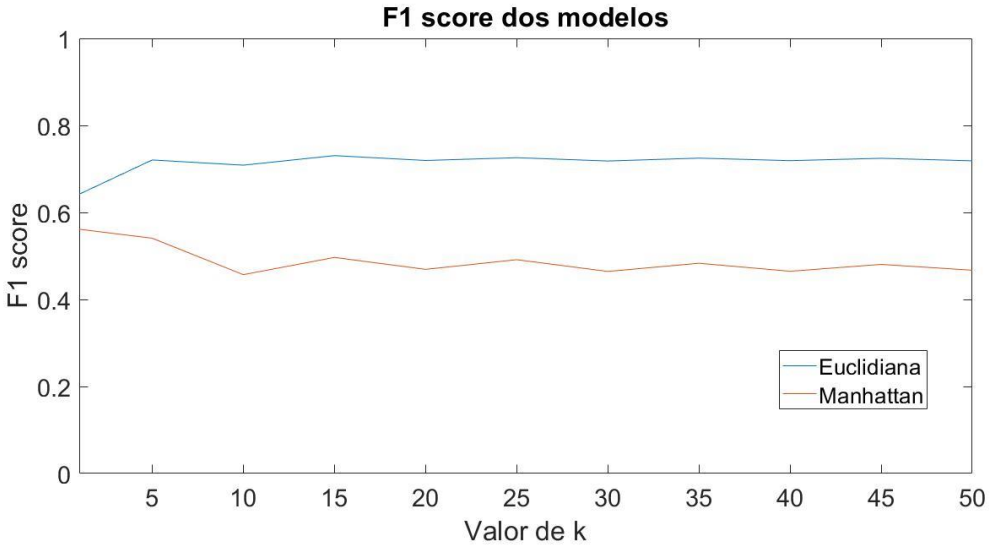

FIGURA 34 – *F1 score* dos modelos da validação cruzada

O modelo com o menor erro avaliado com os dados de validação é aquele que adota k igual a 15 e a métrica de distância euclidiana. A matriz confusão obtida a partir dos dados de teste é apresentada na TABELA 25.

TABELA 25 – Matriz confusão do algoritmo KNN

	1	0
1	4.844	2.209
0	1.380	47.271

A TABELA 26 apresenta as métricas que buscam determinar a qualidade geral do modelo selecionado.

TABELA 26 – Métricas que buscam determinar a qualidade do modelo

Medida	Valor
Acurácia	93,56%
Sensibilidade	68,68%
Precisão	77,83%
F1	0,7296

4.1.2 Árvore de decisão: Modelo 1 de classificação

Para a implementação da árvore de decisão, foi utilizada a função *fitctree* do *software* Matlab. Variando-se o parâmetro *MaxNumSplits* da função supracitada, o qual indica o número máximo de nós, foram elaboradas 50 árvores. As acurácias e os valores de *F1 score* dessas árvores, avaliados em relação aos dados de validação, são indicados na FIGURA 35 e na FIGURA 36, respectivamente.

FIGURA 35 – Acurácia das árvores de decisão elaboradas

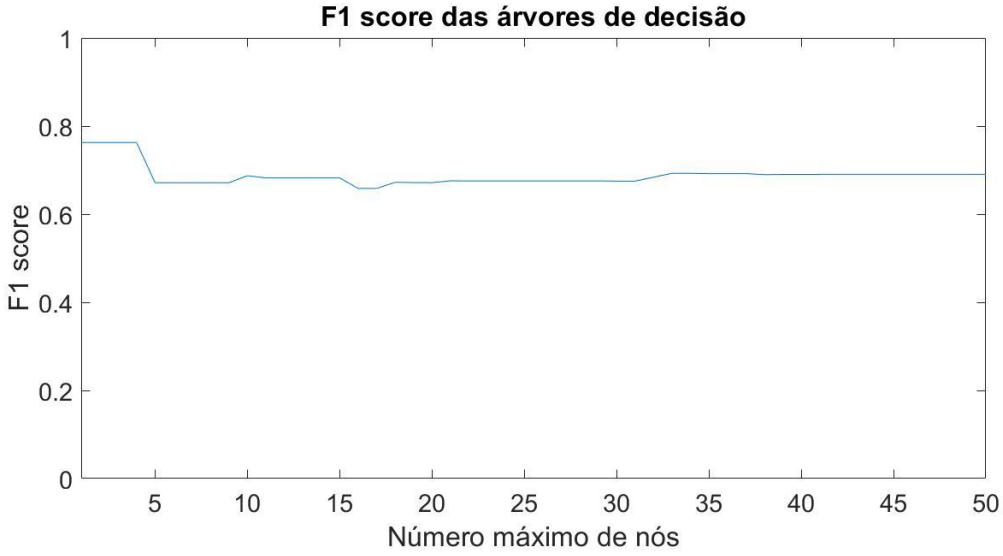

FIGURA 36 – Valores de *F1 score* das árvores de decisão elaboradas

A análise da FIGURA 35 indica que não há uma grande variação na acurácia com a variação do número máximo de nós e que a árvore com a maior acurácia é aquela com o máximo número de nós igual 1. Essa árvore é indicada na FIGURA 37.

FIGURA 37 – Árvore com maior acurácia

A matriz confusão do modelo associado à árvore com maior acurácia, avaliada em relação aos dados de teste, é indicada na TABELA 27.

TABELA 27– Matriz confusão da árvore com maior acurácia

	1	0
1	5.060	1.790
0	1.528	47.326

A TABELA 28 apresenta as métricas que buscam determinar a qualidade geral do modelo selecionado, avaliadas em relação aos dados de teste.

TABELA 28 – Métricas que buscam determinar a qualidade da árvore de decisão

Medida	Valor
Acurácia	94,04%
Sensibilidade	73,87%
Precisão	76,81%
F1	0,7530

O modelo associado à árvore com acurácia máxima, representada na FIGURA 37, possui uma variável dominante: o atraso na partida. A árvore simplesmente prevê que, se há atraso na partida, há atraso na chegada; se não há atraso na partida, não há atraso na chegada.

4.1.3 Regressão logística: Modelo 1 de classificação

Para a implementação da regressão logística, foi utilizada a função *glm* do *software* R. A validação cruzada foi efetuada variando os objetos de família dos modelos. As acurácias e os valores de *F1 score* dos diferentes modelos analisados, avaliados em relação aos dados de validação, são indicados na TABELA 29.

TABELA 29 – Valores de acurácia e de *F1 score* dos modelos analisados

Objeto de família	Acurácia	*F1 score*
Binomial	94,08%	0,7518
Gaussian	94,18%	0,7588
Poisson	93,83%	0,7336

De acordo com a TABELA 29, o modelo com melhor acurácia e melhor *F1 score* é aquele que adota a família gaussiana. A FIGURA 38 apresenta os coeficientes e os valores da estatística *t* de *Student* obtidos para as variáveis de *input* por meio do algoritmo de regressão logística com família gaussiana. A estatística *t* de *Student* é adimensional e corresponde à razão entre o valor do coeficiente e o seu valor de desvio-padrão. Pode-se fazer teste de hipótese a partir do seu valor: as variáveis estatisticamente significativas podem ser adotadas como sendo aquelas que possuem módulo de *t* superior a 2.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.747e+00  2.123e+00  -2.235 0.025387 *
delay_d_bin    7.171e-01  2.001e-03 358.440  < 2e-16 ***
airport        2.776e-01  1.908e-02  14.549  < 2e-16 ***
airline        3.372e-01  1.855e-02  18.173  < 2e-16 ***
aircraft      -8.316e-02  2.407e-02  -3.455 0.000550 ***
season         2.828e-01  3.347e-02   8.449  < 2e-16 ***
hour_cos       3.200e-03  1.355e-03   2.362 0.018165 *
hour_sin      -1.323e-02  1.085e-03 -12.193  < 2e-16 ***
week_cos      -3.283e-03  9.110e-04  -3.603 0.000314 ***
week_sin      -1.112e-03  8.993e-04  -1.237 0.216060
day_month     -3.762e-04  1.209e-04  -3.112 0.001856 **
day_month_cos -1.394e-03  9.079e-04  -1.536 0.124613
day_month_sin -1.875e-03  1.483e-03  -1.265 0.206023
month_cos      1.962e-03  2.278e-03   0.861 0.389105
month_sin     -9.494e-04  1.168e-03  -0.813 0.416186
year           2.382e-03  1.034e-03   2.304 0.021220 *
distance       6.036e-07  9.025e-07   0.669 0.503662
seats         -6.875e-05  2.259e-05  -3.044 0.002335 **
sch_d         -1.697e-05  6.900e-06  -2.459 0.013934 *
sch_h          6.157e-04  3.526e-05  17.461  < 2e-16 ***
temp          -4.801e-04  1.275e-04  -3.767 0.000166 ***
direction     -1.329e-05  8.152e-06  -1.631 0.102943
speed         -2.766e-05  2.522e-04  -0.110 0.912676
pressure      -2.491e-03  7.366e-03  -0.338 0.735276
vh            -9.496e-03  6.558e-04 -14.481  < 2e-16 ***
gust           2.556e-03  3.015e-04   8.477  < 2e-16 ***
vv             2.126e-07  1.791e-07   1.187 0.235113
rain           5.112e-02  2.377e-03  21.509  < 2e-16 ***
```

FIGURA 38 – Coeficiente e estatística t de Student das variáveis de input

A matriz confusão do modelo de regressão logística com família gaussiana, avaliado com os dados de teste, é apresentada na TABELA 30.

TABELA 30 – Matriz confusão do modelo de regressão logística

	1	0
1	5.120	1.823
0	1.539	47.222

A TABELA 31 apresenta as métricas que buscam determinar a qualidade geral do modelo.

TABELA 31 – Métricas que determinam a qualidade do modelo de regressão logística

Medida	Valor
Acurácia	93,96%

Sensibilidade	73,74%
Precisão	76,89%
F1	0,7528

4.1.4 Redes neurais artificiais: Modelo 1 de classificação

Para a implementação da rede neural artificial, foi utilizada a ferramenta *nntool* do *software* Matlab. Para a validação cruzada, foram consideradas apenas redes com 3 camadas: a camada de *input*, com 27 neurônios; a camada oculta, com um número variável de neurônios e a camada de *output*, com apenas um neurônio. A validação cruzada foi efetuada variando-se o número de neurônios da camada oculta de 1 a 10.

As acurácias e os valores de *F1 score* das redes neurais elaboradas durante a validação cruzada, avaliados em relação aos dados de validação, são indicados na FIGURA 39 e na FIGURA 40, respectivamente.

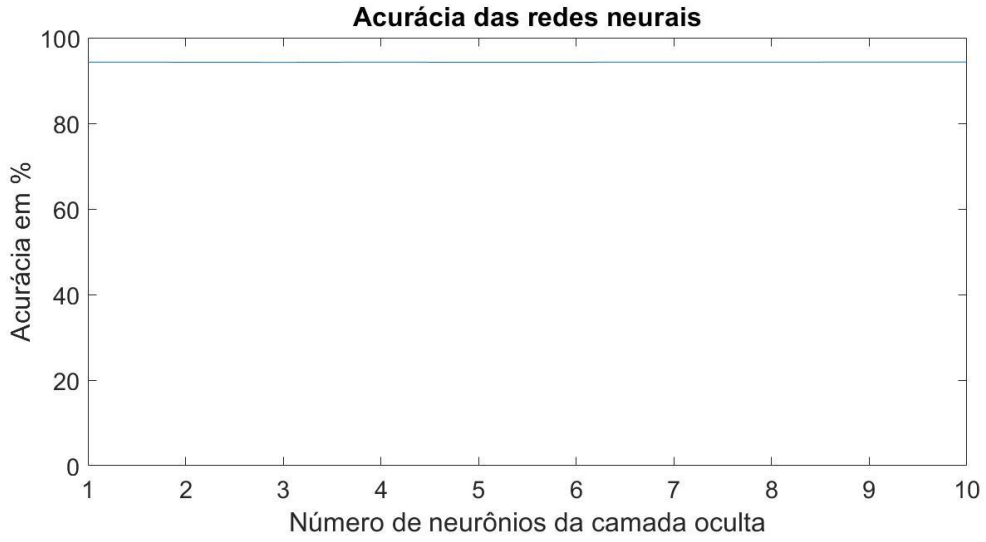

FIGURA 39 – Acurácia das redes neurais avaliadas

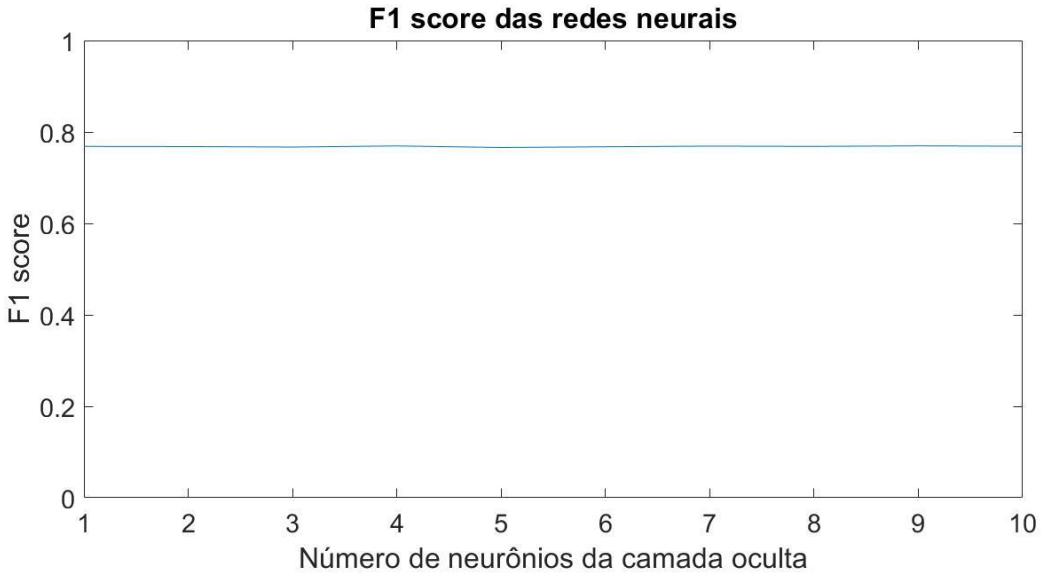

FIGURA 40 – *F1 score* das redes neurais avaliadas

A análise da FIGURA 39 e da FIGURA 40 indica que a variação da acurácia e do *F1 score* em função do número de neurônios da camada oculta é bastante pequena no intervalo analisado. Para o número de neurônios igual a 9, a acurácia e o *F1 score* assume valores máximos, iguais a 94,32% e a 0,3848, respectivamente. Essa é a rede neural selecionada a partir da validação cruzada.

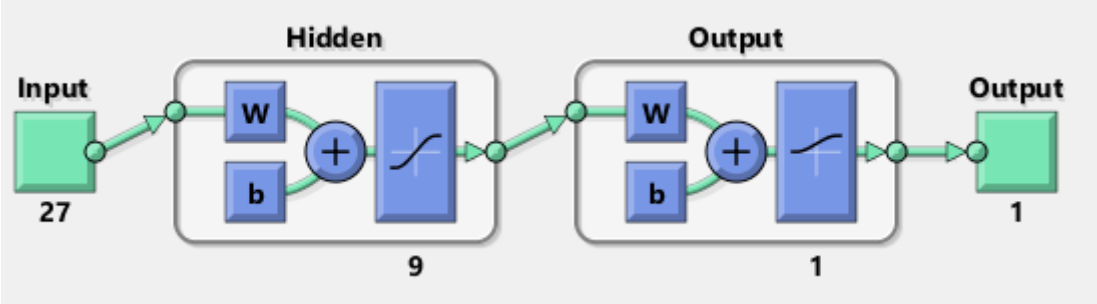

FIGURA 41 – Rede neural selecionada a partir da validação cruzada

A matriz confusão do modelo de rede neural selecionado, avaliado com os dados de teste, é apresentada na TABELA 32.

TABELA 32 – Matriz confusão do modelo de rede neural selecionado

	1	0
1	5.239	1.786
0	1.513	47.166

A TABELA 33 apresenta as métricas que buscam determinar a qualidade geral do modelo.

TABELA 33 – Métricas que determinam a qualidade geral do modelo selecionado

Métrica	Medida
Acurácia	94,08%
Sensibilidade	74,58%
Precisão	77,59%
F1 score	0,7606

4.2 Modelo 2 de classificação

Nas subseções que seguem, será demonstrado o teste do Modelo 2 em cada um dos 4 algoritmos analisados na pesquisa.

4.2.1 KNN: Modelo 2 de classificação

Para a implementação do método KNN, foi utilizada a função *fitcknn* do *software* Matlab. Para a validação cruzada, foram considerados os valores de *k* pertencentes ao conjunto {1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50} e duas métricas de distância: a euclidiana e a de Manhattan.

A FIGURA 42 e a FIGURA 43 apresentam os valores de acurácia e de *F1 Score* dos diferentes modelos testados com os dados de validação.

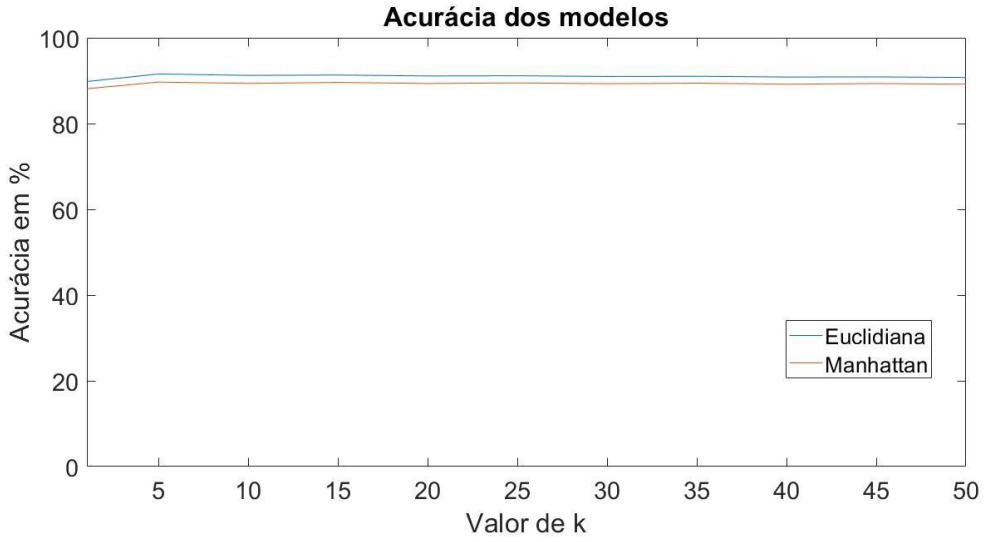

FIGURA 42 – Acurácia dos modelos da validação cruzada

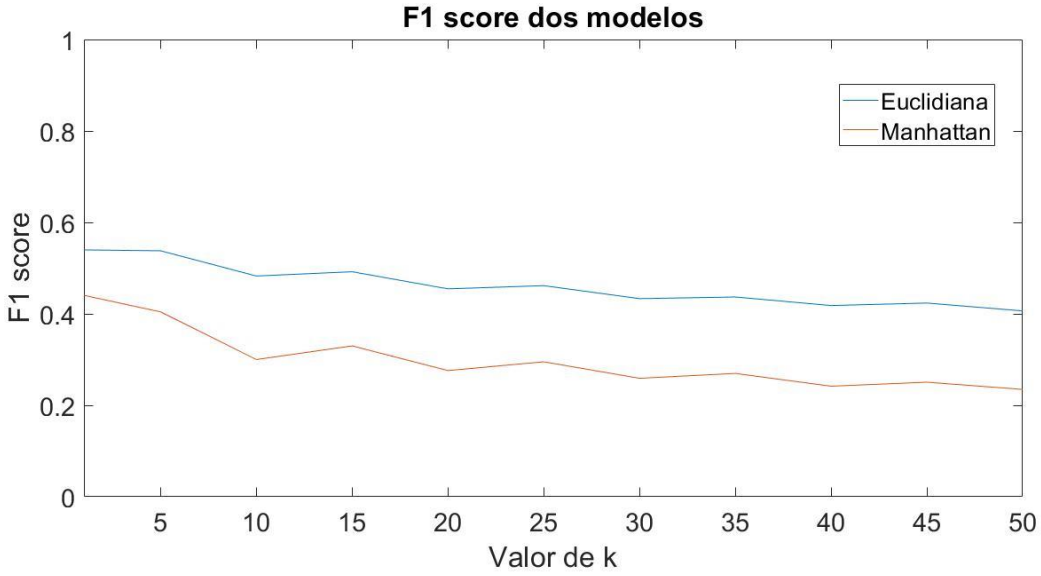

FIGURA 43 – F1 score dos modelos da validação cruzada

O modelo com o menor erro avaliado com os dados de validação é aquele que adota k igual a 5 e a métrica de distância euclidiana. A matriz confusão obtida a partir dos dados de teste é apresentada na TABELA 34.

TABELA 34 – Matriz confusão do algoritmo KNN

	1	0
1	2.868	4.131
0	625	48.067

A TABELA 35 apresenta as métricas que buscam determinar a qualidade geral do modelo selecionado.

TABELA 35 – Métricas da qualidade do modelo selecionado

Medida	Valor
Acurácia	91,46%
Sensibilidade	40,98%
Precisão	82,11%
F1	0,5468

4.2.2 Árvore de decisão: Modelo 2 de classificação

Para a implementação da árvore de decisão, foi utilizada a função *fitctree* do *software* Matlab. Variando-se o parâmetro *MaxNumSplits* da função supracitada, o qual indica o número máximo de nós, foram elaboradas 50 árvores para a validação cruzada. As acurácias e os valores de *F1 score* dessas árvores, avaliados em relação aos dados de validação, são indicados na FIGURA 44 e na FIGURA 45.

FIGURA 44 – Acurácia das árvores de decisão da validação cruzada

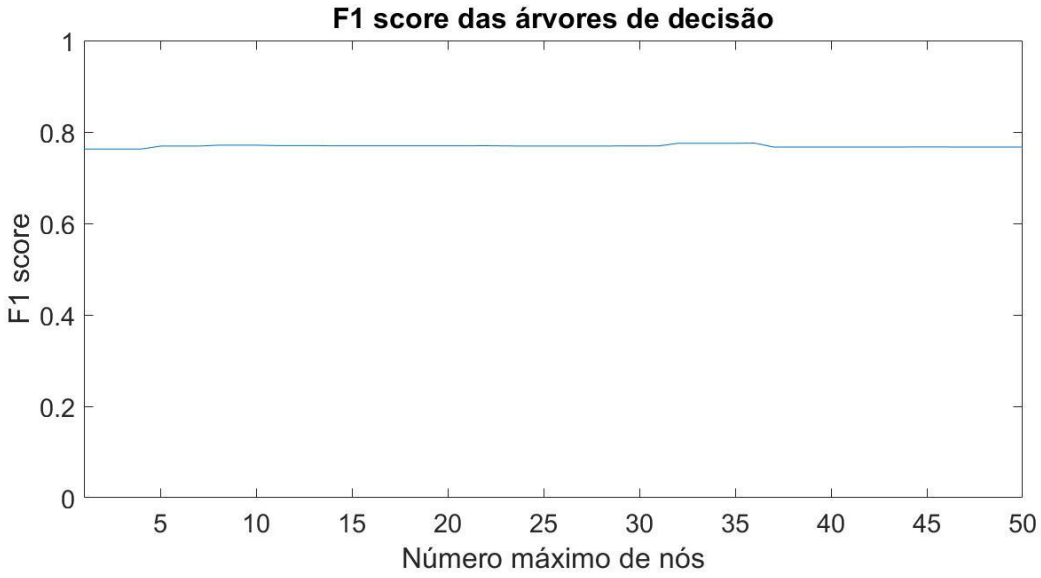

FIGURA 45 – *F1 score* das árvores de decisão da validação cruzada

A análise da FIGURA 44e da FIGURA 45 indica que não há uma grande variação na acurácia e do *F1 score* com a variação do número máximo de nós. A árvore com a maior acurácia é aquela com o máximo número de nós igual a 36. Essa árvore é indicada na FIGURA 46.

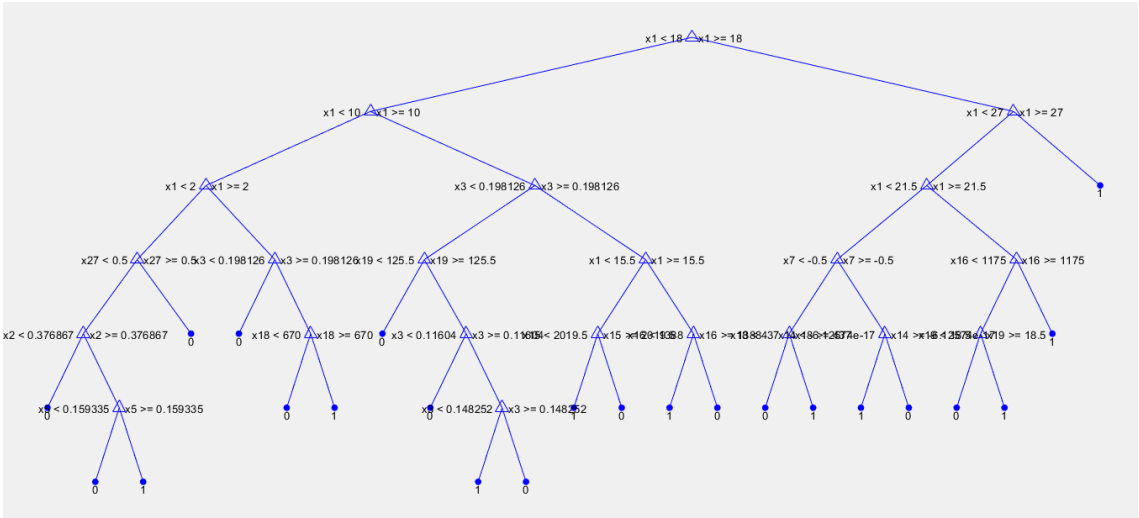

FIGURA 46 – Árvore de decisão com melhor acurácia

A árvore com maior acurácia, com número máximo de nós igual a 36, tem praticamente os mesmos valores de acurácia e de *F1 score* da árvore com apenas um nó, apresentada na FIGURA 47.

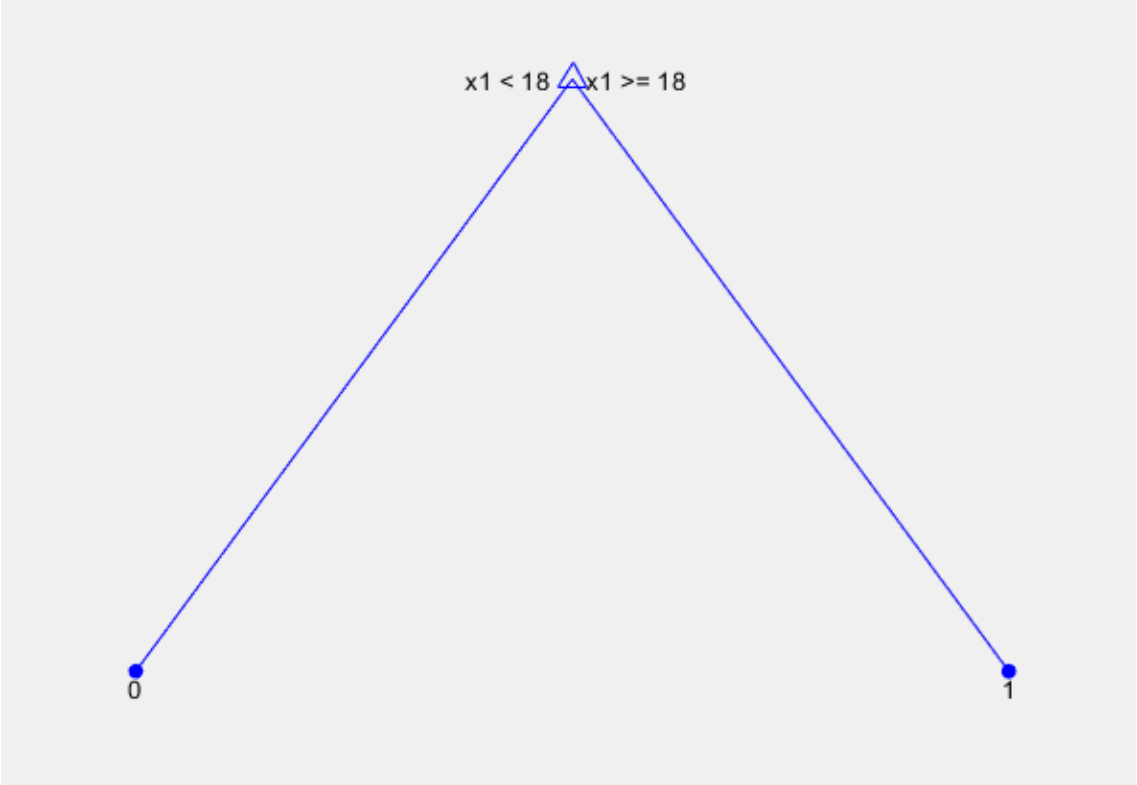

FIGURA 47 – Árvore com apenas um nó

A matriz confusão do modelo associado à árvore com maior acurácia, avaliada em relação aos dados de teste, é indicada na TABELA 36.

TABELA 36 – Matriz confusão do modelo com maior acurácia

	1	0
1	4.868	2.185
0	747	47.904

A TABELA 37 apresenta as métricas que buscam determinar a qualidade geral do modelo selecionado e da árvore com apenas um nó, avaliadas em relação aos dados de teste.

TABELA 37 – Medidas da árvore com 1 nó e da árvore com 36 nós

Medida	Árvore com 1 nó	Árvore com 36 nós
Acurácia	94,49%	94,74%
Sensibilidade	68,59%	69,02%
Precisão	85,03%	86,70%

F1	0,7594	0,7686

A árvore com número máximo de nós igual a 36 tem acurácia apenas 0,25% maior do que a árvore com apenas um nó. Portanto, neste caso, não compensa selecionar o modelo com 36 nós, visto que ele é muito mais complexo e gera um aumento ínfimo de acurácia em relação ao modelo mais simples.

4.2.3 Regressão logística: Modelo 2 de classificação

Para a implementação da regressão logística, foi utilizada a função *glm* do *software* R. A validação cruzada foi efetuada variando os objetos de família dos modelos. As acurácias e os valores de *F1 score* dos diferentes modelos analisados, avaliados em relação aos dados de validação, são indicados na TABELA 38.

TABELA 38 – Valores de acurácia e de *F1 score* dos modelos analisados

Objeto de família	Acurácia	*F1 score*
Binomial	94,88%	0,7692
Gaussian	90,20%	0,3530
Poisson	89,13%	0,2230

De acordo com a TABELA 38, o modelo com melhor acurácia e melhor *F1 score* é aquele que adota a família binomial. A FIGURA 48 apresenta os coeficientes e os valores da estatística *t* de *Student* obtidos para as variáveis de *input* por meio do algoritmo de regressão logística com família binomial.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.297e+02  4.942e+01  -4.647 3.37e-06 ***
delay_d_bin    1.772e-01  1.368e-03 129.566  < 2e-16 ***
airport        3.119e+00  3.763e-01   8.290  < 2e-16 ***
airline        4.711e+00  3.705e-01  12.718  < 2e-16 ***
aircraft      -1.503e+00  4.857e-01  -3.095  0.00197 **
season         3.392e+00  7.408e-01   4.579 4.68e-06 ***
hour_cos       2.289e-02  3.002e-02   0.763  0.44571
hour_sin      -1.310e-01  2.469e-02  -5.305 1.12e-07 ***
week_cos      -1.025e-01  2.064e-02  -4.966 6.83e-07 ***
week_sin      -1.624e-02  1.978e-02  -0.821  0.41162
day_month     -6.271e-03  2.706e-03  -2.317  0.02049 *
day_month_cos -1.154e-02  2.014e-02  -0.573  0.56657
day_month_sin -1.930e-02  3.300e-02  -0.585  0.55864
month_cos      1.068e-01  5.215e-02   2.048  0.04052 *
month_sin      1.196e-02  2.684e-02   0.446  0.65585
year           1.121e-01  2.404e-02   4.663 3.12e-06 ***
distance       6.214e-05  1.967e-05   3.159  0.00158 **
seats         -3.247e-03  5.125e-04  -6.335 2.37e-10 ***
sch_d         -1.006e-03  1.703e-04  -5.910 3.42e-09 ***
sch_h          1.715e-02  8.220e-04  20.869  < 2e-16 ***
temp          -8.062e-03  2.929e-03  -2.752  0.00592 **
direction     -3.486e-04  1.826e-04  -1.909  0.05630 .
speed          4.282e-03  5.719e-03   0.749  0.45406
pressure      -3.742e-03  1.653e-01  -0.023  0.98194
vh            -2.040e-01  1.344e-02 -15.184  < 2e-16 ***
gust           3.844e-02  5.013e-03   7.668 1.75e-14 ***
vv             6.905e-06  4.015e-06   1.720  0.08549 .
rain           7.134e-01  4.361e-02  16.359  < 2e-16 ***
```

FIGURA 48 – Coeficiente e estatística t de Student das variáveis de input

A análise da FIGURA 48 indica que as variáveis estatisticamente significativas são: *delay_d_bin*; *airport; airline; aircraft; season; hour_sin; week_cos; day_month; month_cos; year; distance; seats; sch_d; sch_h; temp; vh; gust* e *rain*.

A matriz confusão do modelo de regressão logística com família binomial, avaliado com os dados de teste, é apresentada na TABELA 39.

TABELA 39 – Matriz confusão do modelo e regressão logística com família binomial

	1	0
1	3562	1624
0	513	36079

A TABELA 40 apresenta as métricas que buscam determinar a qualidade geral do modelo.

TABELA 40 – Métricas que determinam a qualidade do modelo de regressão logística

Medida	Valor
Acurácia	94,62%
Sensibilidade	68,04%
Precisão	85,83%
F1	0,7590

4.2.4 Redes neurais artificiais: Modelo 2 de classificação

Para a implementação da rede neural artificial, foi utilizada a ferramenta *nntool* do *software* Matlab. Para a validação cruzada, foram consideradas apenas redes com 3 camadas: a camada de *input*, com 27 neurônios; a camada oculta, com um número variável de neurônios e a camada de *output*, com apenas um neurônio. A validação cruzada foi efetuada variando-se o número de neurônios da camada oculta de 1 a 10.

As acurácias e os valores de *F1 score* das redes neurais elaboradas durante a validação cruzada, avaliados em relação aos dados de validação, são indicados na FIGURA 49 e na FIGURA 50, respectivamente.

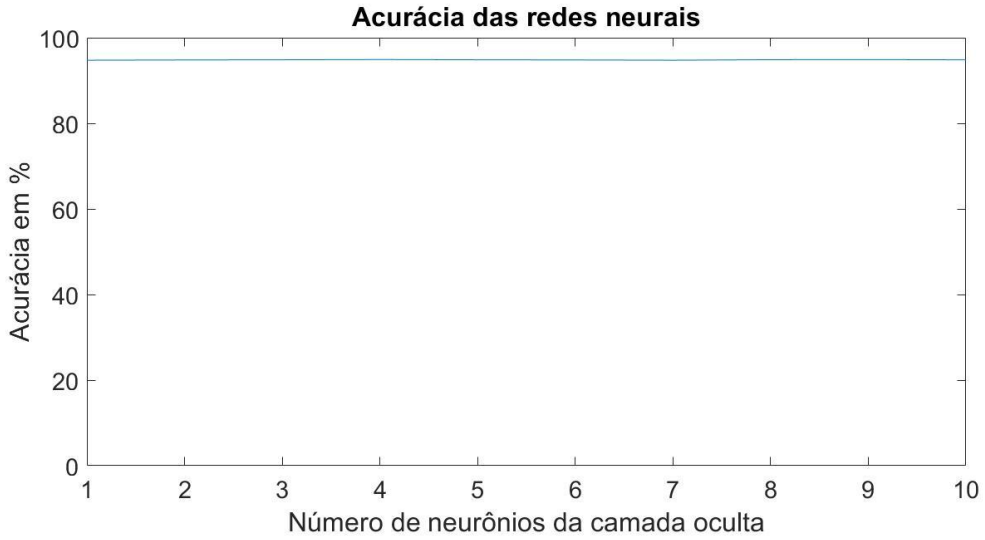

FIGURA 49 – Acurácia das redes neurais da validação cruzada

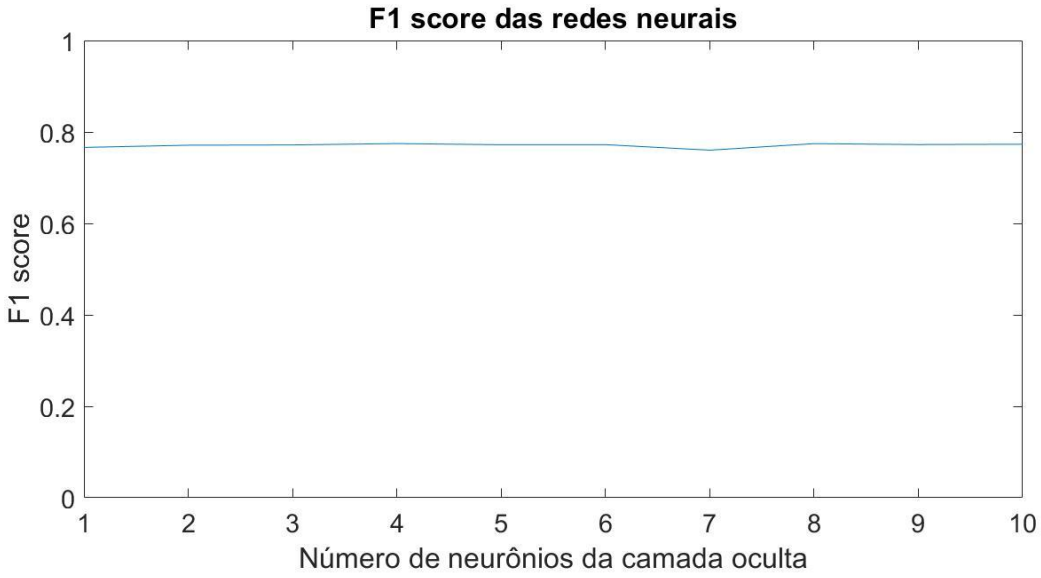

FIGURA 50 – *F1 score* das redes neurais da validação cruzada

A análise da FIGURA 49 e da FIGURA 50 indica que a variação da acurácia e do *F1 score* em função do número de neurônios da camada oculta é bastante pequena no intervalo analisado. Para o número de neurônios igual a 4, a acurácia e o *F1 score* assume valores máximos, iguais a 94,90% e a 0,7746, respectivamente. Essa é a rede neural selecionada a partir da validação cruzada.

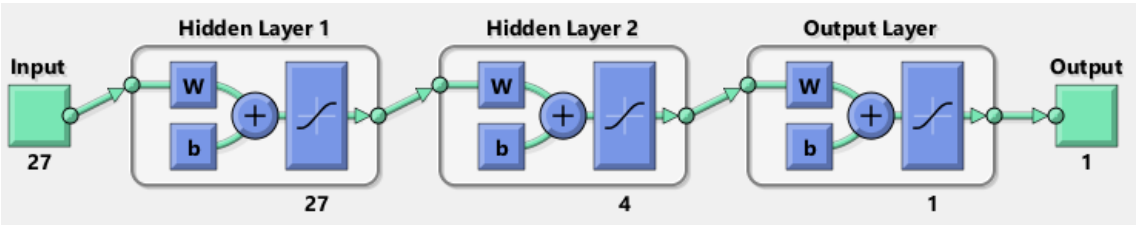

Figura 5: Rede neural selecionada a partir da validação cruzada.

A matriz confusão do modelo de rede neural selecionado, avaliado com os dados de teste, é apresentada na TABELA 41.

TABELA 41 – Matriz confusão da rede neural selecionada

	1	0
1	4.928	2.071
0	835	47.857

A TABELA 42 apresenta as métricas que buscam determinar a qualidade geral do

modelo.

TABELA 42 – Métricas que determinam a qualidade geral do modelo

Métrica	Medida
Acurácia	94,78%
Sensibilidade	70,41%
Precisão	85,51%
F1 score	0,7722

4.3 Modelo de regressão testado em cada algoritmo

Nas subseções que seguem, será demonstrado o teste do Modelo de regressão em cada um dos 4 algoritmos analisados na pesquisa.

4.3.1 KNN: Modelo de regressão

Para a implementação do método KNN, foi utilizada a função *fitcknn* do *software* Matlab. Para a validação cruzada, foram considerados os valores de *k* pertencentes ao conjunto {1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50} e duas métricas de distância: a euclidiana e a de Manhattan. A FIGURA 51 e a FIGURA 52 apresentam, respectivamente, gráficos do MAE e do RMSE dos modelos da validação cruzada, avaliados em relação aos dados de validação.

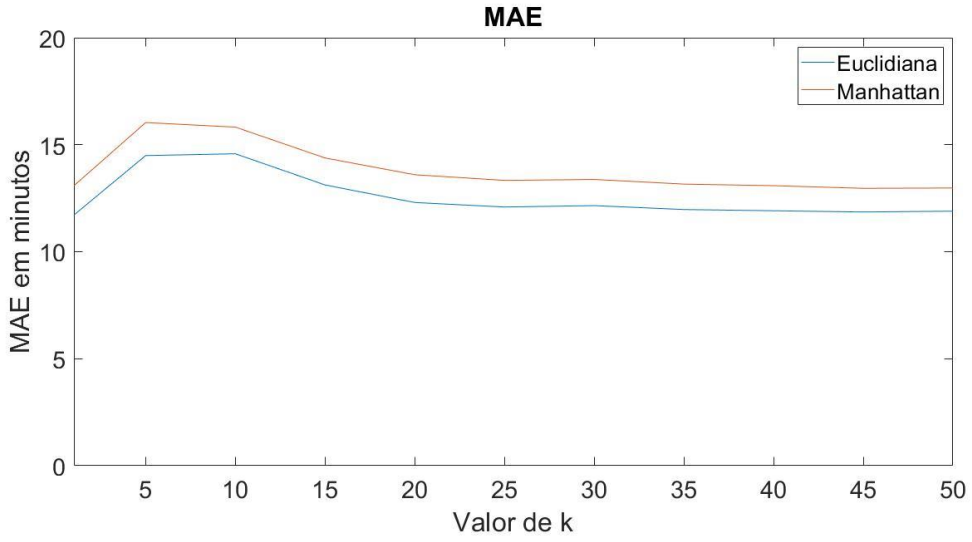

FIGURA 51 – Valor de MAE dos modelos avaliados pela validação cruzada

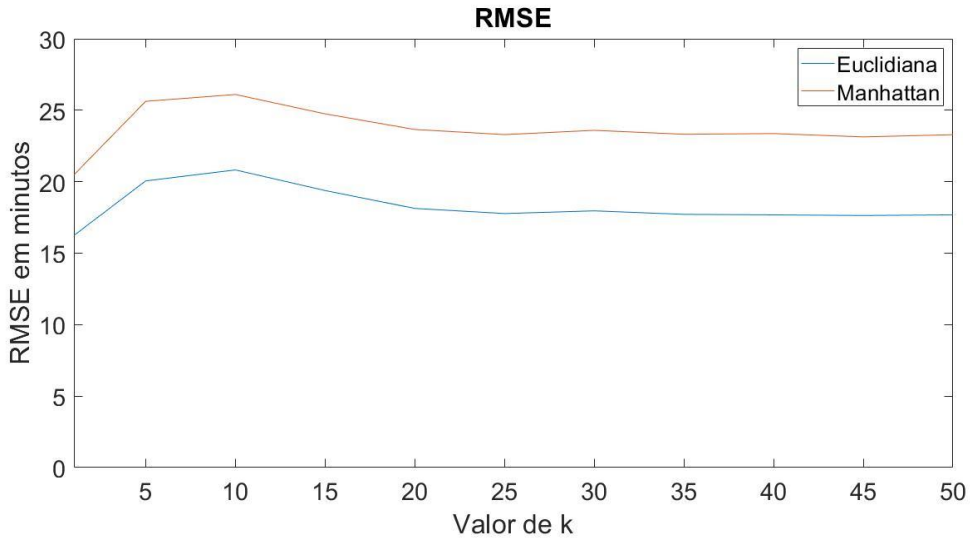

FIGURA 52 – Valor de RMSE dos modelos avaliados pela validação cruzada

Os gráficos da FIGURA 51 e da FIGURA 52 indicam que o modelo com os menores erros é aquele com valor de *k* igual a 1 e métrica de distância euclidiana. Esse modelo apresenta MAE e RMSE de 11,71 minutos e 16,24 minutos, respectivamente. Avaliando-o com os dados de teste, os valores de MAE e de RMSE são apresentados na TABELA 43.

TABELA 43 – MAE e RMSE do modelo com menor erro, avaliados em relação aos dados de teste

Erro	Valor (em minutos)
MAE	13,82
RMSE	22,51

4.3.2 Árvores de decisão: Modelo de regressão

Para a implementação da árvore de decisão, foi utilizada a função *fitctree* do *software* Matlab. Variando-se o parâmetro *MaxNumSplits* da função supracitada, o qual indica o número máximo de nós, foram elaboradas 500 árvores no processo de validação cruzada. O MAE e o RMSE das árvores, avaliados em relação aos dados de validação, são apresentados no gráfico da FIGURA 53.

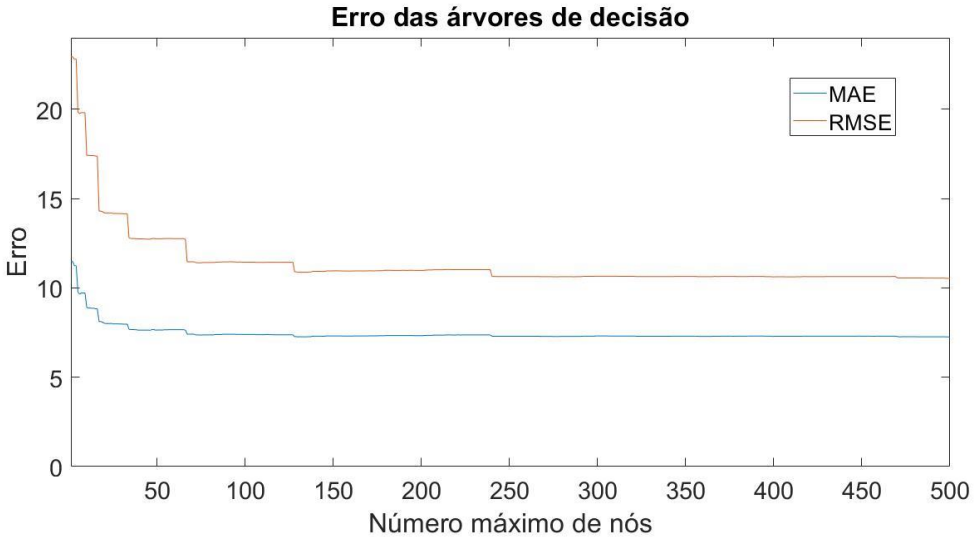

FIGURA 53 – Erro das árvores de decisão avaliado em relação aos dados de validação

A árvore com menor erro é aquela com número máximo de nós igual a 498, com MAE de aproximadamente 7,26 minutos e RMSE de 10,53 minutos, avaliados em relação aos dados de validação. Avaliando essa árvore com os dados de teste, os valores de MAE e de RMSE são apresentados na TABELA 44.

TABELA 44 – MAE e RMSE da árvore selecionada, avaliados em relação aos dados e teste

Erro	Valor (em minutos)
MAE	7,20
RMSE	10,46

4.3.3 Regressão linear: Modelo de regressão

A regressão logística é apropriada para problemas de classificação. A função *glm* do *software* R cria modelos de regressão linear quando aplicada a problemas de regressão.

Adotando-se o modelo de regressão linear, os valores de MAE e de RMSE em relação aos dados de teste são indicados na seguinte tabela:

TABELA 45 – Valores de MAE e de RMSE do modelo gaussiano em relação aos dados de teste

Erro	Valor (em minutos)

MAE	7,00
RMSE	9,67

A FIGURA 54 apresenta o valor dos coeficientes e os valores da estatística *t* de *Student* obtidos para as variáveis de *input* por meio do algoritmo de regressão linear.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.796e+03  9.051e+01 -19.844  < 2e-16 ***
delay_d_bin    9.917e-01  1.271e-03 780.533  < 2e-16 ***
airport        1.285e+01  8.100e-01  15.865  < 2e-16 ***
airline        2.876e+01  8.050e-01  35.725  < 2e-16 ***
aircraft      -8.216e+00  1.026e+00  -8.008 1.18e-15 ***
season         2.841e+00  1.426e+00   1.992  0.04633 *
hour_cos      -9.662e-01  5.782e-02 -16.710  < 2e-16 ***
hour_sin      -5.625e-01  4.612e-02 -12.196  < 2e-16 ***
week_cos      -3.522e-01  3.887e-02  -9.060  < 2e-16 ***
week_sin      -7.021e-02  3.827e-02  -1.834  0.06659 .
day_month     -1.171e-02  5.140e-03  -2.279  0.02267 *
day_month_cos  2.182e-02  3.871e-02   0.564  0.57295
day_month_sin  2.198e-02  6.319e-02   0.348  0.72795
month_cos      7.270e-01  9.713e-02   7.485 7.20e-14 ***
month_sin     -3.679e-01  4.977e-02  -7.391 1.46e-13 ***
year           8.778e-01  4.407e-02  19.917  < 2e-16 ***
distance      -1.000e-04  3.856e-05  -2.594  0.00949 **
seats         -8.004e-03  9.581e-04  -8.354  < 2e-16 ***
sch_d         -4.410e-04  2.945e-04  -1.497  0.13430
sch_h          6.358e-02  1.500e-03  42.382  < 2e-16 ***
temp          -3.701e-02  5.440e-03  -6.803 1.03e-11 ***
direction     -3.671e-03  3.480e-04 -10.548  < 2e-16 ***
speed          3.029e-02  1.075e-02   2.817  0.00486 **
pressure       5.443e-01  3.159e-01   1.723  0.08488 .
vh            -5.464e-01  2.788e-02 -19.597  < 2e-16 ***
gust           1.105e-01  1.278e-02   8.647  < 2e-16 ***
vv            -9.654e-07  7.636e-06  -0.126  0.89940
rain           2.946e+00  1.010e-01  29.178  < 2e-16 ***
```

FIGURA 54 – Coeficientes e valores da estatística t de Student das variáveis de input

A variável *day_month* está negativamente relacionada com a variável *delay_a_bin*. De fato, a FIGURA 18 indica uma tendência sutil de queda na taxa de atrasos à medida que se aproxima dos últimos dias do mês.

A variável *distance* está negativamente relacionada com a variável de *output*. De fato, a FIGURA 10 indica que até a classe 4, há uma tendência de queda na taxa de atrasos com o aumento da distância.

A variável *seats* está negativamente relacionada com a variável *delay_a_bin*. Esse resultado ratifica a conclusão feita a partir do gráfico da FIGURA 8 de que há uma maior tendência de atrasos em voos com um menor número de assentos.

A variável *sch_h* está positivamente relacionada com a variável *delay_a_bin*. Isso

ocorre porque quanto maior é o número de aeronaves pousando e decolando num determinado horário, maior é o congestionamento aéreo do aeroporto e, consequentemente, maior é a tendência de atraso de voos.

A variável *direction* está negativamente relacionada com a variável de *output*. Já a variável *speed* está positivamente relacionada com a variável *delay_a_bin*. Esse comportamento já era esperado, visto que há uma maior tendência de atrasos quando o vento tem velocidades mais altas.

A variável *temp* também está negativamente relacionada com a variável de *output*, indicando que há um aumento na tendência de ocorrência de atrasos à medida que a temperatura cai.

A variável *v_h* está negativamente relacionada com a variável de *output*, ratificando a maior tendência de atraso nos voos quando a visibilidade horizontal é menor.

As variáveis *gust* e *rain* estão positivamente relacionadas com a variável *delay_a_bin*, indicando a maior tendência de atrasos em condições meteorológicas adversas.

4.3.4. Redes neurais artificiais: Modelo de regressão

Para a implementação da rede neural artificial, foi utilizada a ferramenta *nntool* do *software* Matlab. Para a validação cruzada, foram consideradas apenas redes com 3 camadas: a camada de *input*, com 27 neurônios; a camada oculta, com um número variável de neurônios e a camada de *output*, com apenas um neurônio. A validação cruzada foi efetuada variando-se o número de neurônios da camada oculta de 1 a 10.

Os valores de MAE e de RMSE das redes neurais elaboradas durante a validação cruzada, avaliadas em relação aos dados de validação, são indicados na FIGURA 55.

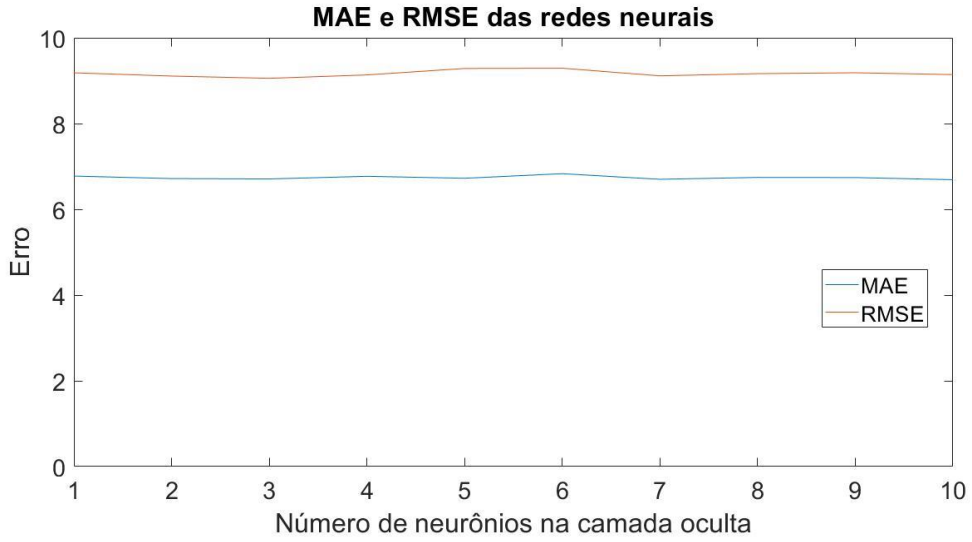

FIGURA 55 – MAE e RMSE das redes neurais da validação cruzada

A FIGURA 55 indica que não há uma grande variação do erro com a variação do número de neurônios na camada oculta. O modelo com menor erro MAE é aquele com 10 neurônios, que possui MAE e RMSE, avaliados em relação aos dados de validação, de 6,69 minutos e 9,14 minutos, respectivamente. Avaliando esse modelo com os dados de teste, os erros são apresentados na tabela a seguir:

TABELA 46 – MAE e RMSE do modelo selecionado a partir da validação cruzada

Erro	Valor (em minutos)
MAE	6,69
RMSE	9,17

5 CONCLUSÕES

A TABELA 47, a TABELA 48 e a TABELA 49 sintetizam os resultados obtidos para o modelo 1, para o modelo 2 e para o modelo 3, respectivamente.

TABELA 47 – Resultados do modelo 1

Algoritmo	Acurácia	*F1 score*
KNN	93,56%	0,7296
Árvore	94,04%	0,7530
Regressão	93,96%	0,7528

Rede neural	94,08%	0,7606

TABELA 48 – Resultados do modelo 2

Algoritmo	Acurácia	*F1 score*
KNN	91,46%	0,5468
Árvore	94,74%	0,7686
Regressão	94,62%	0,7590
Rede neural	94,78%	0,7722

TABELA 49 – Resultados do modelo 3

Algoritmo	MAE (em minutos)	RMSE (em minutos)
KNN	13,82	22,51
Árvore	7,20	10,46
Regressão	7,00	9,67
Rede neural	6,69	9,17

O modelo de classificação com melhor acurácia e mais alto *F1 score* é o modelo 2 implementado com o algoritmo de rede neural. A TABELA 51 compara as métricas desse modelo com as métricas do modelo 2 implementado com a árvore de decisão de apenas um nó, apresentada na FIGURA 47. A partir da TABELA 51, conclui-se que a acurácia do modelo de rede neural é apenas 0,29% maior do que a acurácia do modelo de árvore, ao passo que o *F1 score* é apenas 1,69% maior. Ocorre, portanto, uma degeneração do problema proposto: todos os modelos têm métricas de qualidade muito próximas das de um modelo bastante simples, que consiste de uma árvore com apenas um nó.

TABELA 50 – Métricas de qualidade do modelo com melhor acurácia e do modelo de árvore com apenas um nó

Métricas do modelo	Rede neural do modelo 2	Árvore c/ um nó do modelo 2
Acurácia	94,78%	94,49%

Sensibilidade	70,41%	68,59%
Precisão	85,51%	85,03%
F1 score	0,7722	0,7594

A árvore de decisão com apenas um nó da FIGURA 47, cujas métricas de qualidade são apresentadas na TABELA 51, simplesmente prevê que se o atraso na partida for superior ou igual a 18 minutos, ocorre atraso na chegada; caso contrário, não ocorre atraso na chegada. Conclui-se, portanto, que há uma variável dominante nos modelos analisados: *delay_d_bin*.

O modelo com melhores métricas apresenta precisão de 85,51% e sensibilidade de 70,41%. Ou seja, apenas 70,41% dos atrasos foram classificados corretamente e apenas 85,51% dos voos classificados como atrasados são, de fato, atrasados.

O *F1 score* do melhor modelo, cujo valor é igual a 0,7722, é inferior aos valores encontrados na literatura. Priyanka (2018) obteve *F1 score* igual a 0,92, ao passo que Zoutendijk e Mihaela Mitici (2021) obtiveram o valor de 0,87.

O modelo de regressão com menores erros é o de rede neural, apresentando MAE igual a 6,69 minutos, valor inferior ao obtido por Zoutendijk e Mihaela Mitici (2021), que foi igual a 14,99 minutos. O MAPE calculado apenas com os voos que atrasam (*delay_a_bin* igual ou superior a 15 minutos) é igual a 36,29%, valor bastante elevado. Logo, assim como os modelos de classificação, os modelos de regressão não trouxeram resultados satisfatórios.

Portanto, os modelos de classificação e de regressão não proporcionaram resultados satisfatórios, pois apresentaram erros bastante elevados.

5.1 Limitações da pesquisa e nova proposta: Breve estudo de caso

Não foram obtidos resultados satisfatórios com a base de dados analisada, visto que todos os modelos elaborados apresentam erros bastante elevados. Nessa perspectiva, serão apresentados dois casos em que é aplicada uma metodologia similar a bases de dados significativamente menores e mais simples, com o intuito de obter erros menores.

Serão analisadas bases de dados obtidas a partir da base anterior, considerando apenas voos partindo do Aeroporto Internacional de Porto Alegre (IATA: POA, ICAO: SBPA) operados pela companhia aérea LATAM.

A rota Porto Alegre – Guarulhos é a rota com o maior número de voos da base de dados original, com cerca de 7,42% do total de voos, de acordo com a TABELA 11. Dos

6.933 voos analisados, 764 sofreram atraso na chegada, correspondendo a 11,02% dos voos.

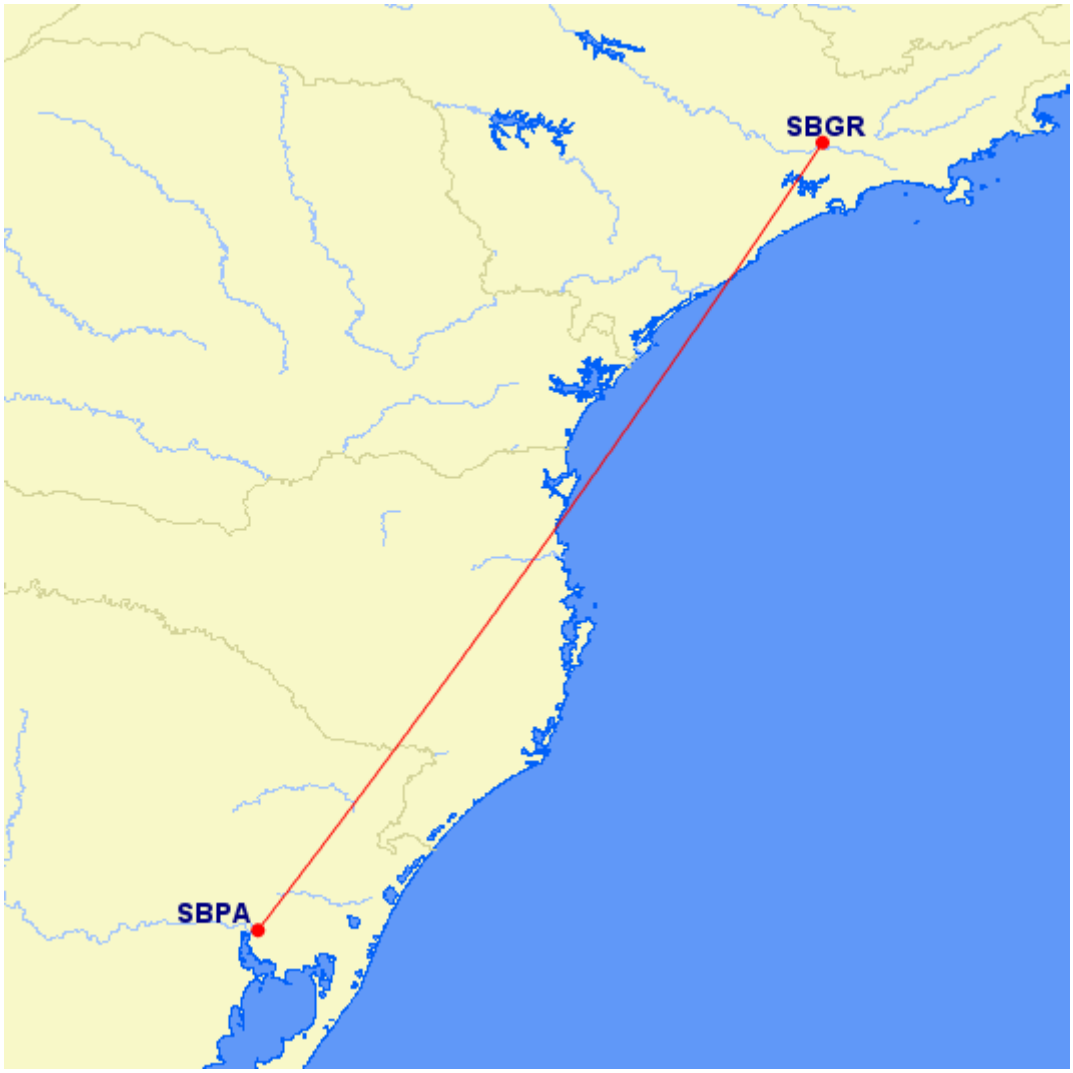

FIGURA 56 – Rota aérea Porto Alegre – Guarulhos

5.1.1 Caso 1

No primeiro caso, foi adotada a mesma metodologia da seção anterior à base de dados considerando apenas voos partindo do Aeroporto Internacional de Porto Alegre operados pela companhia aérea Latam. Os resultados obtidos para o modelo 1, para o modelo 2 e para o modelo 3 são apresentados na TABELA 52, na TABELA 53 e na TABELA 54, respectivamente.

TABELA 51 – Resultados obtidos para o modelo 1

Algoritmo	Acurácia	*F1 score*	Parâmetros
KNN	93,48%	0,6606	k = 10 e

			métrica euclidiana
Árvore	94,77%	0,7702	1 nó
Regressão	94,29%	0,7280	Família gaussiana
Rede neural	94,46%	0,7000	3 neurônios na camada oculta

TABELA 52 – Resultados obtidos para o modelo 2

Algoritmo	Acurácia	*F1 score*	Parâmetros
KNN	91,87%	0,4924	$k = 5$ e métrica euclidiana
Árvore	95,39%	0,7576	Máximo número de nós igual a 17
Regressão	94,69%	0,7458	Família binomial
Rede neural	94,52%	0,7354	5 neurônios na camada oculta

TABELA 53 – Resultados obtidos para o modelo 3

Algoritmo	MAE (em minutos)	RMSE (em minutos)	Parâmetros
KNN	12,46	19,19	$k = 1$ e métrica euclidiana
Árvore	8,18	13,07	Máximo número de nós igual a 25
Regressão	6,66	9,38	Família gaussiana
Rede neural	7,04	9,49	7 neurônios na camada oculta

Entre os modelos de classificação, o que obteve melhor acurácia e melhor *F1 score* foi o modelo 2 com o algoritmo da árvore de decisão. A acurácia obtida com esse modelo foi de 95,39%, apenas um pouco maior do que melhor acurácia obtida na seção anterior, que foi de 94,78%. A árvore é apresentada na FIGURA 57.

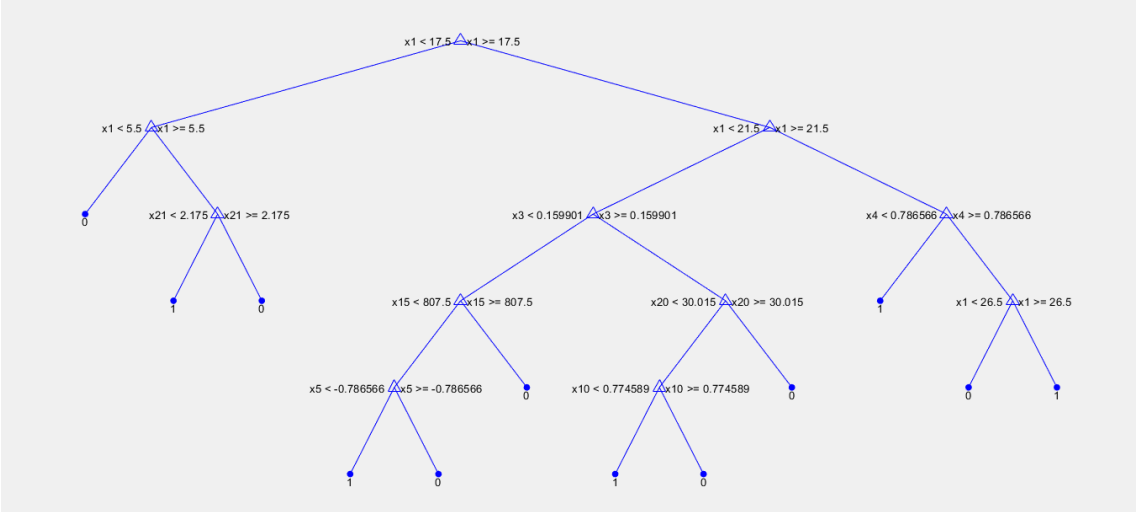

FIGURA 57 – Modelo de classificação de melhor acurácia

A árvore de decisão do modelo 2 com apenas um nó tem acurácia aproximadamente igual à da árvore com melhor acurácia e é apresentada na FIGURA 58. Assim como na seção anterior, o melhor modelo tem acurácia muito próxima da acurácia de um modelo bastante simples – o de árvore com apenas um nó.

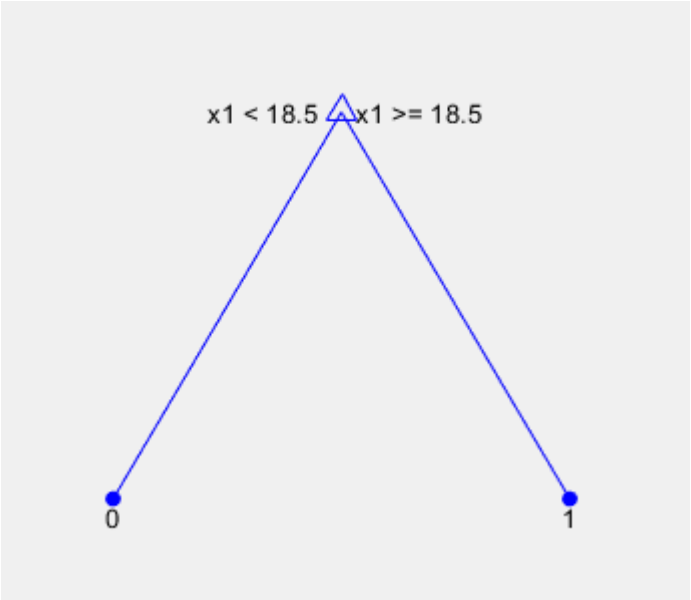

FIGURA 58 – Árvore de decisão com apenas um nó

Entre os modelos de regressão, o que proporcionou menores erros foi o modelo de regressão logística, com MAE e RMSE de aproximadamente 6,66 e 9,38 minutos,

respectivamente. Esses valores são muito próximos dos obtidos na seção anterior – 6,69 e 9,17 minutos. O MAPE aplicado aos dados de teste referentes a voos atrasados foi igual a 39,61%, valor bastante elevado.

Portanto, mesmo com a redução da base de dados, não houve uma melhora significativa nos modelos obtidos.

5.1.2 Caso 2

Neste caso, será analisado um modelo bastante similar ao do estudo de Priyanka (2018), no qual foi aplicado apenas o algoritmo KNN. Nesse estudo, os dados foram coletados para uma companhia específica – a Jetblue Airways – e para uma rota específica, com origem do Aeroporto Internacional de Boston e destino no Aeroporto Internacional de Los Angeles.

No modelo considerado, a variável de *output* é binária e refere-se ao atraso na chegada:

$$delay_{a_{bin}} = 1 \leftrightarrow houve\ atraso\ na\ chegada$$

$$delay_{a_{bin}} = 0 \leftrightarrow não\ houve\ atraso\ na\ chegada$$

As variáveis de *input* são as seguintes:

a) *delay_d_bin:* variável binária que se refere ao atraso na partida:

$$delay_{d_{bin}} = 1 \leftrightarrow houve\ atraso\ na\ partida$$

$$delay_{d_{bin}} = 0 \leftrightarrow não\ houve\ atraso\ na\ partida$$

b) *year*: Variável numérica que indica o ano referente ao horário previsto de pouso do voo correspondente.

c) *month:* Variável numérica que indica o mês associado ao horário previsto de pouso do voo correspondente. Ao contrário dos modelos da seção anterior, *month* não é definida por *target encoding*, assumindo valores inteiros de 1 a 12.

d) *day_month:* Variável numérica que indica o dia do mês do horário previsto de chegada do voo correspondente. Assume valores inteiros de 1 a 31.

e) *hour_d*: horário previsto de partida do Aeroporto Internacional de Porto Alegre. Assume valores inteiros de 0 a 23.

f) *hour_a*: horário previsto de chegada ao aeroporto Internacional de Guarulhos. Assume valores inteiros de 0 a 23.

g) *Speed*: Variável numérica que indica a velocidade do vento em nós no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

h) *rain*: Variável categórica binária que indica se houve precipitação atmosférica no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

$$rain = 1 \leftrightarrow houve\ precipitação$$

$$rain = 0 \leftrightarrow não\ houve\ precipitação$$

i) *temperature:* Variável numérica que indica a temperatura em graus Fahrenheit no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos.

j) *v_h*: Variável numérica que indica a visibilidade horizontal em milhas no horário previsto de pouso do voo no Aeroporto Internacional de Guarulhos. Essa variável não está presente em Priyanka (2018), entretanto foi incluída para substituir a variável relacionada com neve, incluída no artigo. Como não há precipitações de neve em Guarulhos, foi considerada a variável *v_h*.

Os resultados obtidos são apresentados na TABELA 55.

TABELA 54 – Resultados obtidos para o caso 2

Algoritmo	Acurácia	*F1 score*	Parâmetros
KNN	85,41%	-	Distância de Manhattan
Árvore de decisão	93,60%	0,6726	Máximo número de nós igual a 4
Regressão	94,87%	0,7688	Família gaussiana
Rede neural	94,00%	0,7204	10 neurônios na camada oculta

No caso do algoritmo de KNN, os modelos com todos os valores de *k* testados apresentam sensibilidade nula, ou seja, não preveem corretamente um atraso sequer. Portanto, o *F1 score* não está definido para esses modelos. Esse resultado é bem diferente do obtido em Priyanka (2018), que obteve sensibilidade de 95%, precisão de 90% e acurácia de 90% com o algoritmo KNN.

O melhor modelo é o de regressão logística e apresenta acurácia e *F1 score* de aproximadamente 94,87% e 0,7688, aproximadamente. Esses valores são muito próximos dos obtidos na seção anterior, os quais não são satisfatórios. Logo, mesmo com a redução da base de dados e a simplificação do modelo, não houve uma melhora significativa nos modelos obtidos.

REFERÊNCIAS

Alonso, Hugo; Loureiro, António. Predicting Flight Departure Delay at Porto Airport: A Preliminary Study. **Proceedings of the 7th International Joint Conference on Computational Intelligence - NCTA**, *Lisboa*, p. 93-98, 2015. DOI: 10.5220/0005587700930098. Disponível em: https://www.scitepress.org/PublicationsDetail.aspx?ID=mvP5LXPSXFU=&t=1. Acesso em: 19 abr. 2022.

AGÊNCIA NACIONAL DE AVIAÇÃO CIVIL (ANAC). **ANACpédia**, 2022a**.** Disponível em: https://www2.anac.gov.br/anacpedia/por_ing/tr5125.htm. Acesso em: 2 maio 2022a.

AGÊNCIA NACIONAL DE AVIAÇÃO CIVIL (ANAC). **Consulta de Voos Passados – VRA**, 2022b. Disponível em: https://sas.anac.gov.br/sas/bav/view/frmConsultaVRA. Acesso em: 2 maio 2022.

BORSKY, Stefan; UNTERBERGER, Christian. Bad weather and flight delays: the impact of sudden and slow onset weather events. **Economics Of Transportation**, *[S.I.]*, v. 18, p. 10-26, jun. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.ecotra.2019.02.002. Disponível em: https://www.sciencedirect.com/science/article/pii/S2212012218300753. Acesso em: 19 abr. 2022.

BRASIL. **Regulamento Brasileiro da Aviação Civil (RBAC) Nº 154**. Disponível em: https://www.anac.gov.br/assuntos/legislacao/legislacao-1/rbha-e-rbac/rbac/rbac-154/@@display-file/arquivo_norma/RBAC154EMD07%20-%20retificado.pdf. Acesso em: 2 maio 2022.

Departamento de Controle do Espaço Aéreo. **Como decodificar o METAR e o SPECI?** Disponível em: https://ajuda.decea.mil.br/base-de-conhecimento/como-decodificar-o-metar-e-o-speci/. Acesso em: 2 maio 2022.

EFTHYMIOU, Marina *et al*. The Impact of Delays on Customers' Satisfaction: an empirical analysis of the british airways on-time performance at heathrow airport. **Journal Of Aerospace Technology And Management**, São José dos Campos, v. 11, n. 0219, p. 1-13, 12 dez. 2018. FapUNIFESP (SciELO). http://dx.doi.org/10.5028/jatm.v11.977. Disponível em: https://www.scielo.br/j/jatm/a/8B9fSXxbstXNgHHnQGdgtPS/?lang=en. Acesso em: 16 abr. 2022.

GUY, Ann Brody. **Flight delays cost $32.9 billion, passengers foot half the bill**. 2010. Disponível em: https://news.berkeley.edu/2010/10/18/flight_delays/#:~:text=The%20cost%20of%20domestic%20flight,the%20University%20of%20California%2C%20Berkeley. Acesso em: 19 abr. 2022.

IOWA State University: Iowa Environmental Mesonet. Iowa Environmental Mesonet. Disponível em: https://mesonet.agron.iastate.edu/request/download.phtml?network=BR__ASOS. Acesso em: 2 maio 2022.

KHANMOHAMMADI, Sina; TUTUN, Salih; KUCUK, Yunus. A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport. **Procedia Computer Science**, *[S.I.]*, v. 95, p. 237-244, 2016. Elsevier BV. http://dx.doi.org/10.1016/j.procs.2016.09.321. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050916324942. Acesso em: 22 abr. 2022.

LIANG, Ja-Der *et al*. Recurrence predictive models for patients with hepatocellular carcinoma after radiofrequency ablation using support vector machines with feature selection methods. **Computer Methods And Programs In Biomedicine**, *[S.I.]*, v. 117, n. 3, p. 425-434, dez. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.cmpb.2014.09.001. Disponível em: https://www.researchgate.net/publication/265603987_Recurrence_Predictive_Models_for_Patients_with_Hepatocellular_Carcinoma_after_Radiofrequency_Ablation_Using_Support_Vector_Machines_with_Feature_Selection_Methods. Acesso em: 1 set. 2022.

OLIVEIRA, McWillian de *et al*. Analysis of airport weather impact on on-time performance of arrival flights for the Brazilian domestic air transportation system. **Journal Of Air Transport Management**, [S.L.], v. 91, p. 1-5, mar. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.jairtraman.2020.101974. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0969699720305573. Acesso em: 22 abr. 2022.

ON-TIME Performance. 2022. Disponível em: https://www.oag.com/on-time-performance-airlines-airports. Acesso em: 16 abr. 2022.

MARKOVIC, D. *et al*. A statistical study of the weather impact on punctuality at Frankfurt Airport. **Meteorological Applications**, Hannover, v. 15, n. 2, p. 293-303, 2008. Wiley. http://dx.doi.org/10.1002/met.74. Disponível em: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.74. Acesso em: 19 abr. 2022.

MARTÍNEZ-ÁLVAREZ, Francisco *et al*. A Survey on Data Mining Techniques Applied to Electricity-Related Time Series Forecasting. **Energies**, *[S.I.]*, v. 8, n. 11, p. 13162-13193, 19 nov. 2015. MDPI AG. http://dx.doi.org/10.3390/en81112361. Disponível em: https://www.researchgate.net/publication/284516167_A_Survey_on_Data_Mining_Techniques_Applied_to_Electricity-Related_Time_Series_Forecasting. Acesso em: 10 maio 2022.

MCDONALD, Carol. **Accelerating Apache Spark 3**. 2021. Disponível em: https://www.nvidia.com/es-la/ai-data-science/spark-ebook/predictive-analytics-spark-machine-learning/. Acesso em: 10 maio 2022.

METAR & TAF: Decodificador visual. Decodificador visual. Disponível em: https://metar-taf.com/pt. Acesso em: 2 maio 2022.

NATIONAL CENTER OF EXCELLENCE FOR AVIATION OPERATIONS RESEARCH (NEXTOR). **Total Delay Impact Study**. 2010. Disponível em: https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf. Acesso em: 2 maio 2022.

PITFIELD, D.E; JERRARD, E.A. Monte Carlo comes to Rome: a note on the estimation of unconstrained runway capacity at rome fiumucino international airport. **Journal Of Air**

Transport Management, Brussel, v. 5, n. 4, p. 185-192, out. 1999. Elsevier BV. http://dx.doi.org/10.1016/s0969-6997(99)00012-5.

PRIYANKA, G.. Prediction of Airline Delays Using K-Nearest Neighbor Algorithm. **International Journal Of Emerging Technology And Innovative Engineering**, *[S.I.]*, v. 4, n. 5, p. 87-90, ago. 2018. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3340771. Acesso em: 28 abr. 2022.

REDAÇÃO GUARULHOS HOJE. **Guarulhos lidera atrasos e cancelamentos de voos em 2021, revela pesquisa**. 2022. Disponível em: https://www.guarulhoshoje.com.br/2022/02/04/guarulhos-lidera-atrasos-e-cancelamentos-de-voos-em-2021-revela-pesquisa/#:~:text=De%20acordo%20com%20a%20AirHelp,tamb%C3%A9m%20s%C3%A3o%20motivo%20de%20indeniza%C3%A7%C3%A3o.. Acesso em: 19 abr. 2022.

SANTOS, Thiago Adriano dos *et al*. Modelo de identificação do impacto futuro de chuvas extremas nos atrasos/cancelamentos de voos. **Transportes**, *[S.I.]*, v. 26, n. 2, p. 44-53, 31 ago. 2018. Lepidus Tecnologia. http://dx.doi.org/10.14295/transportes.v26i2.1379. Disponível em: https://revistatransportes.org.br/anpet/article/view/1379. Acesso em: 19 abr. 2022.

SCARPEL, Rodrigo Arnaldo; PELICIONI, Luciele Cristina. A data analytics approach for anticipating congested days at the São Paulo International Airport. **Journal Of Air Transport Management**, São José dos Campos, v. 72, p. 1-10, set. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.jairtraman.2018.07.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0969699717300777. Acesso em: 19 abr. 2022.

SHAH, Tarang. **About Train, Validation and Test Sets in Machine Learning**. 2017. Disponível em: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7. Acesso em: 26 ago. 2022.

STEINHEIMER, Martin. Decision support for air traffic management based on probabilistic weather forecasts. **Tu Wien**, *[S.I.]*, p. 1-73, 2019. TU Wien. http://dx.doi.org/10.34726/HSS.2019.68530. Disponível em: https://repositum.tuwien.at/handle/20.500.12708/6488. Acesso em: 19 abr. 2022.

SULLIVAN, John. **6 Steps To Write Any Machine Learning Algorithm From Scratch: Perceptron Case Study**. 2018. Disponível em: https://towardsdatascience.com/6-steps-to-write-any-machine-learning-algorithm-from-scratch-perceptron-case-study-335f638a70f3. Acesso em: 10 maio 2022.

TOLEDO, Alexander H.; FLIKKEMA, Robert; TOLEDO-PEREYRA, Luis H.. Developing the Research Hypothesis. **Journal Of Investigative Surgery**, *[S.I.]*, v. 24, n. 5, p. 191-194, 25 ago. 2011. Informa UK Limited. http://dx.doi.org/10.3109/08941939.2011.609449. Disponível em: https://www.tandfonline.com/doi/abs/10.3109/08941939.2011.609449. Acesso em: 22 abr. 2022.

YU, Bin *et al*. Flight delay prediction for commercial air transport: a deep learning approach. **Transportation Research Part e**: Logistics and Transportation Review, *[S.I.]*, v. 125, p. 203-221, maio 2019. Elsevier BV. http://dx.doi.org/10.1016/j.tre.2019.03.013.

Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1366554518311979. Acesso em: 19 abr. 2022.

WU, Cheng-Lung. Inherent delays and operational reliability of airline schedules. **Journal Of Air Transport Management**, New South Wales, v. 11, n. 4, p. 273-282, jul. 2005. Elsevier BV. http://dx.doi.org/10.1016/j.jairtraman.2005.01.005. Disponível em: https://www.researchgate.net/publication/200035418_Inherent_delays_and_operational_reliability_of_airline_schedules. Acesso em: 16 abr. 2022.

ZOUTENDIJK, Micha; MITICI, Mihaela. Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem. **Aerospace**, *[S.I.]*, v. 8, n. 6, p. 152, 28 maio 2021. MDPI AG. http://dx.doi.org/10.3390/aerospace8060152. Disponível em: https://www.mdpi.com/2226-4310/8/6/152/htm. Acesso em: 22 abr. 2022.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. REGISTRO N°	4. N° DE PÁGINAS
TC	23 de novembro de 2022	DCTA/ITA/TC-089/2022	103

5. TÍTULO E SUBTÍTULO:

Análise e previsão de atrasos em voos no sistema de transporte aéreo do Aeroporto Internacional de Guarulhos

6. AUTOR(ES):

Ygor Rodrigo de Melo Fontes Santos

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

Instituto Tecnológico de Aeronáutica - ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Análise, Previsão, Atrasos, Voos, Aeroporto, Guarulhos, Machine learning

9.PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Transporte aéreo; Atraso, Voo; Árvore de decisão; Aprendizagem (inteligência artificial); Operações de linha aéreas; Análise de fatores; Transportes.

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

ITA, São José dos Campos. Curso de Graduação em Engenharia Civil-Aeronáutica. Orientador: Alessandro Vinícius Marques de Oliveira; coorientadora: Mayara Condé Rocha Murça. Publicado em 2022.

11. RESUMO:

Atrasos em voos são inevitáveis e causam muitos prejuízos às companhias aéreas e à sociedade como um todo. Além do impacto na economia, eles exercem influência sobre a satisfação dos consumidores. Diante desse contexto, o presente estudo faz uma análise dos atrasos em voos do aeroporto mais movimentado do Brasil, o Aeroporto Internacional de Guarulhos. Assim, pelo motivo referido, esse aeroporto representa o que mais sofre em virtude da concentração e do congestionamento por atraso de voos. Nesse aspecto, a motivação da pesquisa se justifica pela relevância do Aeroporto assinalado. Além disso, o problema de pesquisa a ser desenvovido é o seguinte: Qual algoritmo de machine learning tem melhor desempenho ao prever atrasos de voo no aeroporto de Guarulhos?. Ademais, com relação à hipótese de pesquisa, configurou-se a seguinte hipótese: O algoritmo de machine learning que apresenta melhor desempenho ao prever atrasos de voo no aeroporto de Guarulhos é o de redes neurais artificiais. Dessa forma, este estudo utiliza 4 algoritmos de machine learning, quais sejam, KNN, árvore de decisão, regressão logística e rede neural artificial, com o intuito de elaborar um modelo preditivo para os atrasos no aeroporto em análise. Nesse sentido, os algoritmos são todos aplicados a 2 modelos de classificação e 1 de regressão. Quanto às variáveis eleitas para a pesquisa, são consideradas as variáveis referentes à operação dos voos e as variáveis metereológicas. Enfim, com relação à conclusão da pesquisa, compreendeu-se que tanto os modelos de classificação, quanto o modelo de regressão propostos apresentaram erros bastante elevados. Por fim, os resultados preliminares levaram ao teste de dois estudos de caso, a fim de verificar uma aplicação verossímel dos algoritmos analisados, mitigando-se as condições preliminarmente estabelecidas.

12. GRAU DE SIGILO:

(X) OSTENSIVO ( ) RESERVADO ( ) SECRETO