INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Jonathans Schaffer Torres

PREDIÇÃO DAS ROTAS DA AZUL LINHAS AÉREAS USANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Trabalho de Graduação
2022

Curso de Engenharia Civil-Aeronáutica

CDU 656.7

Jonathans Schaffer Torres

PREDIÇÃO DAS ROTAS DA AZUL LINHAS AÉREAS USANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

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

Schaffer Torres, Jonathans
Predição das rotas da Azul Linhas Aéreas usando algoritmos de aprendizado de máquina / Jonathans Schaffer Torres.
São José dos Campos, 2022.
75f.

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

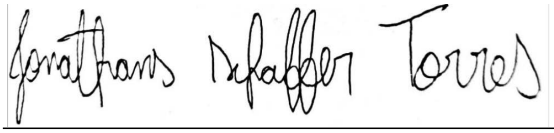
1. Aprendizado de máquina. 2. Mercado aéreo. 3. Previsão de preço. I. Instituto Tecnológico de Aeronáutica. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

SCHAFFER TORRES, Jonathans. **Predição das rotas da Azul Linhas Aéreas usando algoritmos de aprendizado de máquina**. 2022. 75f. Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Jonathans Schaffer Torres
TITULO DO TRABALHO: Predição das rotas da Azul Linhas Aéreas usando algoritmos de aprendizado de máquina.
TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2022



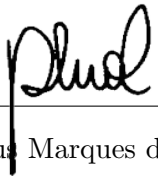
Jonathans Schaffer Torres
Rua H8B, 236
12228-461 – São José dos Campos–SP

PREDIÇÃO DAS ROTAS DA AZUL LINHAS AÉREAS USANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

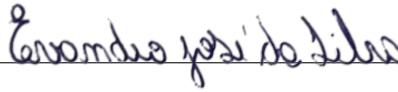
Jonathans Schaffer Torres

Autor



Alessandro Vinícius Marques de Oliveira (ITA)

Orientador



Prof. Dr. Evandro José da Silva
Coordenador do Curso de Engenharia Civil-Aeronáutica

São José dos Campos, 18 de novembro de 2022.

Dedico este trabalho ao meu pai, Giancarlo, à minha mãe, Vânia, e ao meu irmão, Giancarlo Júnior, por terem me dado força ao longo desta longa e árdua jornada.

Agradecimentos

Inicialmente agradeço ao meu pai, Giancarlo, à minha mãe, Vânia, e ao meu irmão, Giancarlo Júnior, por sempre terem acreditado em mim, dando-me forças para superar todas as dificuldades.

Agradeço aos mestres que tive ao longo da minha formação, desde os meus professores na Babylândia, no início do meu ensino básico, até os que me acompanharam no curso preparatório, em especial a todos do Colégio Farias Brito, e principalmente ao coordenador Airton, que além de me dar a oportunidade, despertou em mim a vontade e o sonho de ingressar neste instituto.

Agradeço às amizades que fiz e às pessoas que conheci durante esses anos, sem vocês o trajeto teria sido muito mais árduo. Obrigado por transformarem os difíceis anos de ITA nos mais memoráveis da minha vida até hoje.

Agradeço ao meu professor orientador, Alessandro, por toda a confiança depositada em mim e pelo suporte ao longo deste trabalho. O senhor é um excelente professor, exemplo de dedicação e compromisso com o aluno, meu muito obrigado!

"A persistência é o caminho do êxito."
— Charlie Chaplin.

Resumo

As consequências das entradas em novas rotas e o que leva as companhias aéreas a optarem pela adição de novos destinos são aspectos importantes e de extremo interesse da indústria, sendo considerado um dos elementos cruciais no planejamento estratégico das empresas. O presente trabalho tem como objetivo desenvolver modelagens baseadas em aprendizado de máquina para realizar a predição de rotas da Azul Linhas Aéreas, tendo como referência os dados de passageiros e de tráfego no período de 2008 a 2018. A base de dados utilizada possui um desequilíbrio de classes, evidenciando a necessidade da utilização de métodos de pré-processamento de dados, que foram implementados, majoritariamente, através da reamostragem do conjunto. Aplicando as técnicas subamostragem, sobreamostragem e a combinação destas aliadas às estratégias de análise de correlações e significância, seleção de variáveis e validação da modelagem com *K-fold cross validation*, comparam-se os métodos de redes neurais, regressão logística, KNN e árvore de decisão para o modelo preditivo, chegando a um melhor desempenho para as redes neurais artificiais. Como métricas de desempenho, consideram-se a acurácia, sensibilidade, especificidade, precisão, *F1-score*, *G-Mean* e tempo de treinamento do modelo. Para o modelo de redes neurais, realizam-se predições de rotas por ano, por região do território brasileiro e por proporção de balanceamento do conjunto após a reamostragem. Compara-se ainda o método de sobreamostragem e a abordagem híbrida na previsão de rotas de 2018 em todo o território nacional, concluindo com um melhor desempenho da abordagem híbrida em todas as métricas.

Palavras-chave: Transporte aéreo; Predição de rotas; Desequilíbrio de classes; Pré-processamento de dados; Reamostragem; Aprendizado de Máquina; Rede neural.

Abstract

The consequences of entering new routes and what leads airlines to choose to add new destinations are important aspects and are part of the interest of the industry, being considered one of the crucial elements in the strategic planning of companies. The present work aims to develop models based on machine learning to predict routes for Azul Linhas Aéreas, using passenger and traffic data from 2008 to 2018 as a reference. The database used has a class imbalance, highlighting the need to use data pre-processing methods, which were mostly implemented through resampling of the set. Applying undersampling, oversampling and combination techniques, together with correlation and significance analysis strategies, variable selection and modeling validation with K-fold cross validation, the methods of neural networks, logistic regression, KNN and decision tree methods are compared for the predictive model, reaching a better performance for the artificial neural networks. As performance metrics, accuracy, sensitivity, specificity, precision, F1-score, G-Mean and model training time are considered. For the neural network model, route predictions are made per year, by region of the brazilian territory and by balancing proportion of the set after resampling. It also compares the oversampling method and the hybrid approach in the 2018 route forecast across the national territory, concluding with a better performance of the hybrid approach in all metrics.

Keywords: Air transport; Route forecast; Class imbalance; Data pre-processing; Resampling; Machine learning; Neural network.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ANAC	Agência Nacional de Aviação Civil
RPK	*Revenue Passenger Kilometers*
RUS	*Random Undersampling*
ROS	*Random Oversampling*
SVM	*Support Vector Machine*
KNN	*k-Nearest Neighbor*
IBGE	Instituto Brasileiro de Geografia e Estatística
PIB	Produto Interno Bruto
LCC	*Low Cost Carrier*
FSC	*Full Service Carrier*
PCC	*Pearson Correlation Coefficient*

Sumário

1 Introdução

Em diversos setores, o sucesso e a sustentabilidade da estratégia de negócios de uma empresa está ligada às decisões de entrar em novos mercados. Os exemplos mais evidentes incluem redes de restaurantes, supermercados, bancos e serviços de transporte, cujas empresas do ramo enfrentam o grande desafio de otimizar a sequência de entradas em mercados diferentes - levando em conta possíveis restrições de recursos e barreiras de entrada – para operar de forma lucrativa e construir uma presença sustentável no mercado (MULLER *et al.*, 2012).

No aspecto do transporte aéreo, Oliveira e Oliveira (2022) introduzem que as consequências das entradas em novas rotas e o que leva as companhias aéreas a optarem pela adição de novos destinos são aspectos importantes e de extremo interesse da indústria, sendo considerado um dos elementos cruciais no planejamento estratégico das empresas. Sob outra perspectiva, a entrada no mercado de uma certa rota é um tema que relaciona diversas áreas do conhecimento, dentre elas economia, administração e marketing (DIXIT; CHINTAGUNTA, 2007), de modo que além de estimular a concorrência também induz a inovação (OLIVEIRA; OLIVEIRA, 2022).

Conforme afirmam Dixit e Chintagunta (2007), os gerentes de marketing decidem fazer a entrada no mercado mesmo quando as informações disponíveis são limitadas ou propensas a mudanças. Além disso, devido à natureza incerta das informações iniciais de atratividade do mercado e à experiência limitada com novos negócios, os resultados reais não são claros, exigindo dos gerentes de marketing a necessidade de reavaliar a atratividade do mercado e decidir sobre a permanência ou entrada no ramo.

Em vista disso, Muller et al. (2012) categorizam que a decisão de entrar em uma nova rota costuma conciliar da forma mais eficaz duas estratégias: entrar em mercados existentes, enfrentando a concorrência das empresas consolidadas, porém com maior clareza dos desafios, e entrar em novos mercados, que embora precise lidar com as barreiras de entrada, pode conferir maior lucratividade e o sucesso da empresa. Nas fases iniciais de suas operações, é comum as companhias aéreas do setor de baixo custo buscarem a entrada em novas rotas. No entanto, a disponibilidade de tais rotas pode ter se tornado mais escassa ao longo tempo, obrigando os novos entrantes a operarem em mercados com

a presença de concorrência ativa (GIL-MOLTO; PIGA, 2008).

Bogulaski et al. (2004) afirmam que a densidade de passageiros, as distâncias entre aeroportos, os *hubs* de seus concorrentes e a receita na operação têm sido fatores importantes para determinar a malha de operação de algumas companhias do setor *low-cost* norte-americano. Além disso, Sinclayr (1995) também defende a hipótese de que as decisões de entrada e saída do mercado são afetadas pelas características do sistema de *hubs* que operam nos aeroportos.

Este trabalho contempla o caso da Azul Linhas Aéreas no setor do transporte aéreo brasileiro. Estabelecida em 2008 por David Neeleman, fundador da JetBlue (uma das maiores operadoras de baixo custo do mercado estadunidense), a empresa brasileira que iniciou as atividades com apenas 3 destinos, atualmente conta com cerca de 151 destinos e mais de 900 voos diários.

Consolidando os levantamentos da Base de Dados Estatísticos de Transporte Aéreo da ANAC (Agência Nacional de Aviação Civil), é possível obter o número de entradas da Azul em novas rotas por ano, ao longo dos anos de 2008 a 2018, cujo valores estão apresentados na tabela 1.1.

TABELA 1.1 – Variação anual do número de entradas em novas rotas da Azul Linhas Aéreas.
Fonte: ANAC, 2022.

Ano	**Novas rotas**
2008	4
2009	36
2010	70
2011	85
2012	589
2013	182
2014	165
2015	179
2016	251
2017	130
2018	165

Um marco importante na trajetória da Azul foi sua fusão com a Trip Airlines em maio de 2012, fato evidenciado pelo elevado número de novas rotas no respectivo ano, cuja inclusão de trajetos foi a maior dentre todos os anos analisados. A Azul Linhas Aéreas costuma explorar o mercado de curta e média distância em pequenos aeroportos,

característica também observada na estratégia de uma gigante dos Estados Unidos: a Southwest Airlines (BOGUSLASKI *et al.*, 2004).

Entretanto, em contraste a grandes companhias do setor *low-cost* norte-americano, a Azul não mantém uma frota padronizada, sendo detentora de um portfólio bastante diversificado e que abrange aeronaves de baixa, média e alta densidade de passageiros (OLIVEIRA; OLIVEIRA, 2022), permitindo a consolidação da empresa no mercado nacional. Conforme dados da ANAC (Agência Nacional de Aviação Civil) apresentados na figura 1.1, referentes a agosto de 2022, nos últimos 12 meses a participação da Azul no mercado doméstico brasileiro atinge 30,6% em RPK, que representa a demanda medida por passageiros por km pago transportado.

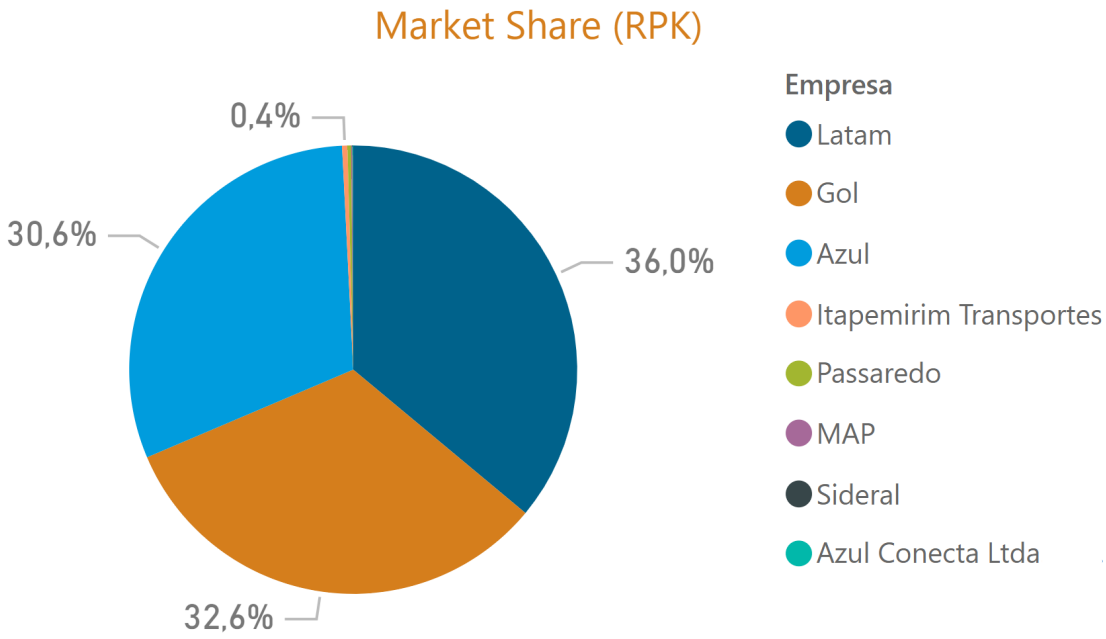


FIGURA 1.1 – Gráfico da participação de mercado RPK (set-2021 a ago-2022).
Fonte: ANAC, 2022.

Resultado das decisões estratégicas da empresa, englobando a constante inclusão de novos destinos, o notável crescimento da Azul nos últimos anos e seu impacto recente na evolução do setor de transporte aéreo brasileiro serviram de motivação para este trabalho que visa prever as rotas da companhia. Mediante o exposto, o presente trabalho de graduação tem como objetivo desenvolver modelagens baseadas em aprendizado de máquina para realizar a predição de rotas da Azul Linhas Aéreas, tendo como referência os dados de passageiros e de tráfego no período de 2008 a 2018, compilados e tratados por Oliveira e Oliveira (2022).

A análise inicial do conjunto de dados permitiu identificar o desequilíbrio de classes em relação à variável que se deseja prever. Segundo Bach et al. (2019), o desbalanceamento de classes reflete em modelos fortemente preditivos para a classe em maior quantidade e

pouco preditivos para a classe em menor quantidade, de modo que o custo da classificação errônea de exemplos minoritários é tipicamente maior do que o custo de classificação errônea dos majoritários. Nesta situação, embora possa ser obtida uma boa acurácia, ao analisar outras métricas de desempenho calculadas a partir da matriz de confusão, é possível identificar a ineficácia preditiva (MOHAMMED *et al.*, 2020). Em consideração a isso, faz-se necessária a adoção de técnicas para lidar com o desequilíbrio de classes, dado que a nossa classe de interesse é a classe minoritária. Ou seja, o pré-processamento dos dados incluiu a utilização de estratégias para o balanceamento das classes.

Portanto, após obter as bases balanceadas pela redução da classe majoritária, realizam-se as predições de rotas para cada ano da base de dados, comparando os resultados obtidos para as diferentes proporções de balanceamento. Em seguida, implementam-se os modelos com as bases balanceadas obtidas pelo aumento da classe minoritária para prever rotas em cada região do Brasil, comparando outra vez os resultados obtidos para as diferentes proporções de balanceamento. Além disso, utilizam-se as bases balanceadas pela combinação do aumento da classe minoritária com a redução da classe majoritária para prever novamente as rotas por região, desta vez comparando os resultados da predição de cada ano. Por fim, comparam-se os resultados das diferentes formas de reamostragem implementadas através da predição das rotas em todo o território nacional, cuja modelagem foi desenvolvida para prever as rotas do ano mais recente do conjunto de dados.

Os próximos capítulos deste trabalho estão divididos da seguinte forma: o capítulo 2 apresenta a revisão de literatura sobre a análise e o pré-processamento dos dados, além de discorrer sobre os métodos de modelagem preditiva de rotas. O capítulo 3 apresenta a metodologia empregada, abrangendo os recursos tecnológicos utilizados, descrição e análise dos dados, aplicação das estratégias de pré-processamento e a caracterização do modelo preditivo desenvolvido. O capítulo 4 apresenta as métricas de desempenho adotadas e a análise dos resultados das predições para os diferentes modelos implementados. Por fim, o capítulo 5 finaliza o trabalho resumindo os principais resultados e sintetizando conclusões.

2 Revisão de literatura

2.1 Análise e pré-processamento de dados

Hasanin et al. (2019) definem que um *big data* é caracterizado por um conjunto de propriedades relacionadas, incluindo volume, variedade, velocidade, variabilidade, valor e complexidade. Adicionalmente, conjecturam que o volume de dados é provavelmente a característica mais evidenciada de um *big data*, especialmente se os conjuntos de dados excederem 1 milhão de observações, tal como o nosso estudo de caso.

Embora seja verdade que o desequilíbrio de classes afete conjuntos de dados grandes ou pequenos, os efeitos adversos são geralmente mais pronunciados no primeiro. Ou seja, graus extremos de desbalanceamento de classes podem existir dentro de um *big data* devido à representação massiva de uma das classes nos conjuntos de dados. A classe minoritária, que compõe um grupo menor parte do conjunto de dados, costuma ser a classe de interesse. Ademais, a classe majoritária compõe a maior parte do conjunto de dados (GOSAIN; SARDANA, 2017).

Em comparação com técnicas estatísticas tradicionais, algoritmos de aprendizado de máquina produzem melhores resultados de classificação; entretanto, tratando-se de dados com desequilíbrio de classes, o algoritmo classificador pode não conseguir discriminar efetivamente entre as classes minoritária e majoritária (HASANIN *et al.*, 2019).

Murphey et al. (2004) afirmam que, sem um método de aprendizado adequado, as observações que representam as classes minoritárias no conjunto de treinamento podem ser ignoradas pelo aprendizado de máquina. Este problema é causado pelo grande número de exemplos de aprendizado da classe majoritária, afetando parcialmente o efeito do treinamento através do pequeno número de exemplos de aprendizado da classe minoritária. Sob outra perspectiva, Bach et al. (2019) defendem que os modelos preditivos desenvolvidos usando algoritmos convencionais de aprendizado de máquina podem ser tendenciosos e imprecisos, uma vez que os algoritmos de aprendizado de máquina geralmente são projetados para melhorar a precisão reduzindo o erro, não levando em conta a distribuição de classes. Segundo Leevy et al. (2018), diversos métodos para lidar com o problema de desbalanceamento de classes têm sido relatados na literatura, permitindo segmentá-los

em três grandes grupos:

- *Data-level*:
  Abordagem em que as observações de treinamento são modificadas adicionando ou removendo observações para alcançar uma distribuição de classes mais equilibrada, implementada através da reamostragem de dados e seleção de variáveis.

- *Algorithmic-level*:
  Consiste em criar novos algoritmos ou modificar os existentes para serem mais sintonizados com os problemas de desbalanceamento de classes, abrangendo os algoritmos sensíveis ao custo, em que custos diferentes são atribuídos para a classificação de cada classe de acordo com a sua distribuição, e a aprendizagem em conjunto, em que o aprendizado é feito simultaneamente com diferentes classificadores e ao final é unificado em somente um modelo.

- Combinação de *Data-level* e *Algorithmic-level*:
  Métodos que consolidam abordagens híbridas de *Data-level* e *Algorithmic-level*, visando produzir uma melhor solução para resolver o problema de desequilíbrio de classes.

Neste cenário, enquadrando-se como uma abordagem em *Data-level*, a aplicação de métodos que alteram a distribuição dos dados para obter uma base equilibrada é uma solução eficaz para o problema de desbalanceamento (BRANCO *et al.*, 2015). Mohammed et al. (2020) esclarecem que as abordagens comuns para o problema de desequilíbrio de classes consistem em usar técnicas de reamostragem para balancear o conjunto de dados. Estas técnicas de reamostragem podem ser aplicadas por sub ou sobreamostragem do conjunto de dados, de forma que a subamostragem é o processo de diminuir a quantidade de observações ou amostras da classe majoritária, enquanto a sobreamostragem pode ser realizada aumentando a quantidade de observações da classe minoritária, implementada através da produção de novas observações ou replicação de algumas. A subamostragem e a sobreamostragem, quando feitas de forma aleatória, são definidas respectivamente como *Random Undersampling* (RUS) e *Random Oversampling* (ROS) (FERNáNDEZ *et al.*, 2018). Além disso, Hasanin et al. (2019) citam uma abordagem utilizando a seleção de variáveis para lidar com este desequilíbrio, que consiste em selecionar os atributos mais influentes que podem fornecer informações sobre a diferenciação entre classes.

Segundo Hasanin et al. (2019), a técnica de *Random Undersampling* tem sido recomendada em detrimento da *Random Oversampling*, pois a subamostragem impõe uma menor carga computacional e resulta em um tempo de treinamento mais rápido, o que é benéfico para análises de dados em geral. Em contrapartida, as análises de Mohammed et al. (2020) e de Leevy et al. (2018) evidenciaram a sobreamostragem desempenhando

melhor do que os métodos de subamostragem para diferentes modelos classificadores de aprendizado de máquina.

Aliado a isso, é importante ressaltar que as alterações no conjunto de dados inicial por sua sub ou sobreamostragem podem causar a distorção dos dados. No caso de métodos de subamostragem, existe a possibilidade de descartar dados potencialmente úteis relacionados à classe majoritária. Por outro lado, ao adicionar observações minoritárias, alerta-se sobre o risco de *overfitting*, em que o classificador pode construir modelos que parecem ser precisos, mas na verdade cobrem apenas os exemplos replicados (BACH *et al.*, 2019).

Fernández et al. (2018) atestam, portanto, que as hibridizações de subamostragem e sobreamostragem surgem como alternativa às desvantagens mencionadas associadas a cada família de métodos. Com a combinação de métodos, busca-se o equilíbrio ideal entre remover exemplos majoritários e gerar novos exemplos minoritários para obter o melhor desempenho possível. Em outras palavras, esta combinação de diferentes estratégias de pré-processamento de dados, como as técnicas de subamostragem e sobreamostragem, ou até mesmo a seleção de atributos, podem levar à diminuição dos impactos negativos introduzidos pelos métodos aplicados individualmente.

Adicionalmente, visando melhorar os resultados da aplicação de modelos clássicos de aprendizado de máquina, Ferreira et al. (2019) sugerem a adoção de métodos para normalização dos dados antes do treinamento do modelo. Para a aplicação de aprendizado de máquina em conjuntos de informações multidimensionais, a normalização de recursos é um passo importante no pré-processamento de dados para os casos em que o objetivo é empregar um algoritmo baseado na minimização de erros.

2.2 Modelos preditivos

As rotas entre aeroportos podem ser classificadas como existentes ou inexistentes, configurando, portanto, um problema de classificação binária em aprendizado de máquina. Entretanto, a literatura atual é escassa no que diz respeito à temática de predição de malhas aéreas através dos modelos de aprendizagem.

No contexto de transporte aéreo, Xie et al. (2014) aplicaram a combinação de máquina de vetores de suporte (SVM) e regressão para prever a quantidade de passageiros aéreos no Aeroporto Internacional de Hong Kong. Alekseev e Seixas (2009) desenvolveram um modelo de previsão neural artificial com pré-processamento de dados por decomposição, visando também a análise de passageiros do transporte aéreo. Srisaeng et al. (2015) compararam os métodos de rede neural artificial e regressão linear múltipla, aplicados à previsão da demanda por passageiros e da receita de passageiros por quilômetros percorridos no transporte aéreo de baixo custo da Austrália, cujos resultados evidenciaram uma

melhor predição segundo a modelagem da rede neural.

Acerca do tema de predição de malhas aéreas, Sriratanawilai e Erjongmanee (2018) desenvolveram um modelo preditivo de rotas existentes ou inexistentes nos Estados Unidos, comparando três métodos de aprendizagem: regressão logística, SVM e redes neurais. Novamente os resultados apontaram a predição do modelo de rede neural artificial desempenhando melhor em relação aos outros dois métodos.

Por outro lado, para problemas de classificação de dados em geral, a maioria dos classificadores, incluindo regressão logística, árvore de decisão e rede neural, funcionam bem quando a distribuição de classe do alvo no conjunto de dados é balanceada (YAP *et al.*, 2013).

Dessa forma, acerca do tópico de reamostragem de dados em desbalanceamento de classes e modelagem preditiva classificatória, Hasanin et al. (2019) mensuraram a performance de classificação mediante os métodos de *random forest*, *gradient-boosted trees* e regressão logística. Através das análises dos desempenhos na modelagem preditiva, Mohammed et al. (2020) compararam, dentre outros algoritmos, as técnicas de SVM, regressão logística e árvore de decisão. Bach et al. (2019) mediram a acurácia de classificação em dados desbalanceados com diversas formas de aprendizagem, dentre elas a rede neural artificial, SVM e *k-Nearest Neighbor* (KNN). Leevy et al. (2018) e Bach et al. (2019) aplicaram, por fim, uma etapa de validação da modelagem preditiva com a metodologia de *K-Fold Cross Validation*.

3 Metodologia

3.1 Recursos tecnológicos

3.1.1 Linguagem de programação

A linguagem de programação *R* foi a adotada para a implementação dos modelos. Trata-se de uma linguagem de programação multi-paradigma orientada a objetos, funcional, dinâmica, voltada à manipulação, análise e visualização de dados.

3.1.2 Ferramentas

O ambiente de desenvolvimento utilizado foi o *RStudio*, cujas implementações dos modelos foram suportadas principalmente pelas seguintes bibliotecas e suas respectivas funções desempenhadas:

- *readr*:
  Importa os dados em CSV e converte-os para o formato de *data frame*.
- *scales*:
  Permite a confecção dos histogramas em porcentagens para as distribuições de frequência das variáveis.
- *ggplot2*:
  Elabora gráficos personalizados através do método de construção gráfica por camadas.
- *stringr*:
  Facilita o tratamento dos dados com formatação de texto.
- *ROSE*:
  Cria amostras possivelmente balanceadas por exemplos minoritários de sobreamostragem aleatória, exemplos majoritários de subamostragem ou combinação de sobreamostragem e subamostragem.

- *Hmisc*:
  Calcula os coeficientes de correlação e agrupa em uma matriz.

- *corrplot*:
  Interpola a matriz de correlação com as devidas formatações e restrições.

- *e1071*:
  Realiza o treinamento do modelo a partir do método de SVM.

- *caret*:
  Calcula a matriz de confusão das classes observadas e previstas e implementa o método de KNN.

- *neuralnet*:
  Treina redes neurais através de configurações flexíveis, permitindo escolha personalizada de erro e função de ativação.

3.2 Descrição do conjunto de dados

A base de dados utilizada consiste em informações de passageiros e de tráfego da Azul Linhas Aéreas no período de 2008 a 2018, extraídos da Base de Dados Estatísticos de Transporte Aéreo da ANAC, do Instituto Brasileiro de Geografia e Estatística (IBGE) e do Banco Central do Brasil. As informações foram compiladas e tratadas por Oliveira e Oliveira (2022).

O conjunto de dados possui um total de 1.052.678 observações (382.792 observações antes da fusão (2008-2011) e 669.886 observações no período após a fusão (2012-2018)). Estas observações contemplam um total de 95.698 rotas e 312 aeroportos do território brasileiro.

As 39 variáveis existentes na base de dados estão descritas na tabela 3.1.

TABELA 3.1 – Descrições das variáveis presentes no conjunto de dados.
Fonte: autoria própria, 2022.

Variável	Descrição
k	Representação do par de aeroportos e da rota
ao	Aeroporto de origem
ad	Aeroporto de destino
year	Ano
AZ_FULL	Decisão da Azul em relação à entrada sem considerar a data de fusão
AZ_BEF	Decisão da Azul em relação à entrada considerando o período antes da fusão
AZ_AFT	Decisão da Azul em relação à entrada considerando o período depois da fusão
PAX	Passageiros pagos
DIST_300	Distância maior que 300 milhas
DIST_600	Distância maior que 3600 milhas
DIST_900	Distância maior que 900 milhas
DIST_1200	Distância maior que 1200 milhas
DIST_1500	Distância maior que 1500 milhas
POP	População média das cidades de origem e destino
INC	PIB per capita
UNEMPL	Taxa de desemprego média das cidades de origem e destino
VACATION	Receita do turismo sobre o PIB média das cidades de origem e destino
SECND	Presença de aeroporto secundário
SLOT	Presença de slot aeroportuário
FEE	Taxas de pouso no aeroporto média das cidades de origem e destino
NETWEC	Total de cidades atendidas pela Azul nas cidades de origem e destino
ZERAZCIT	Terminal não atendido pela Azul
AZSHCON	Participação da Azul na conexão de passageiros
HUBOTH	Terminal aeroportuário é hub de uma rival da Azul
NONHUB	Terminal aeroportuário não é um hub principal
HHI	Concentração de rota
MAXHHI	Concentração de rota máxima
MAXHHI_NONHUB	Concentração de rota máxima por terminais que não são hubs principais
FSCMAJ	Presença de grandes FSCs
LCCMAJ	Presença de grandes LCCs
LCCCOMP	Presença de LCCs
BANKR	Presença de falência
REGSMA	Presença de pequenos regionais
NEW	Nova rota
TREND	Tendência de tempo
TREND_DIST	Tendência de tempo por distância
TREND_HUB	Tendência de tempo por hub
TREND_SECND	Tendência de tempo por aeroporto secundário
TREND_NEW	Tendência de tempo por nova rota

3.3 Análise exploratória dos dados

3.3.1 Classificação dos dados

As métricas de cada variável estão apresentadas na tabela 3.2.

TABELA 3.2 – Métricas das variáveis presentes no conjunto de dados.
Fonte: autoria própria, 2022.

Variável	**Métrica**
k	Valor inteiro
ao	Texto
ad	Texto
year	Inteiro no intervalo de 2008 a 2018
AZ_FULL	*Dummy*
AZ_BEF	*Dummy*
AZ_AFT	*Dummy*
PAX	Logaritmo da contagem
DIST_300, *DIST_600*, etc	*Dummies* mutuamente exclusivos
POP	Logaritmo da contagem
INC	Logaritmo do valor
UNEMPL	Valor no intervalo de 0 a 100
VACATION	Proporção
SECND	*Dummy*
SLOT	*Dummy*
FEE	Logaritmo do valor
NETWEC	Contagem
ZERAZCIT	*Dummy*
AZSHCON	Proporção
HUBOTH	*Dummy*
NONHUB	*Dummy*
HHI	Valor no intervalo de 0 a 1
MAXHHI	Valor no intervalo de 0 a 1
MAXHHI_NONHUB	Valor no intervalo de 0 a 1
FSCMAJ	*Dummy*
LCCMAJ	*Dummy*
LCCCOMP	*Dummy*
BANKR	*Dummy*
REGSMA	*Dummy*
NEW	*Dummy*
TREND	Valor no intervalo de 1 a 11
TREND_DIST	Valor no intervalo de 1 a 306
TREND_HUB	Valor no intervalo de 0 a 11
TREND_SECND	Valor no intervalo de 0 a 11
TREND_NEW	Valor no intervalo de 0 a 11

3.3.2 Mapa de rotas

As rotas em que a Azul Linhas Aéreas decidiu iniciar ou retomar operações no ano de 2018 estão apresentadas na figura 3.1.

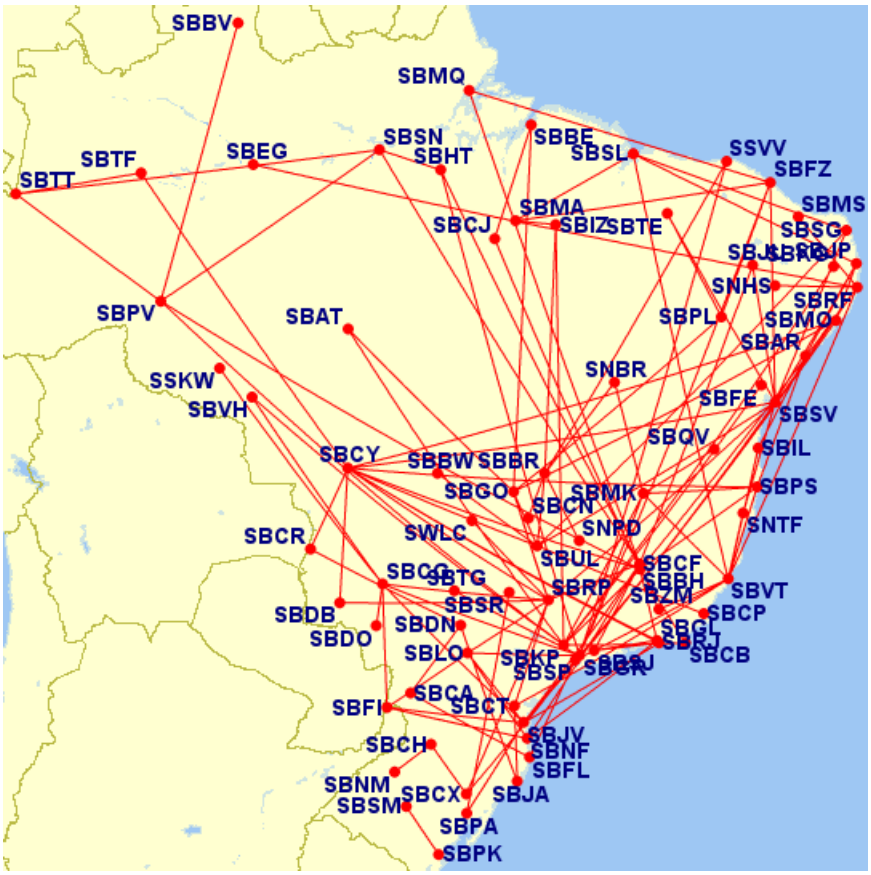


FIGURA 3.1 – Mapa de novas rotas da Azul em 2018.
Fonte: autoria própria, 2022.
Mapas gerados pelo Great Circle Mapper.

3.3.3 Distribuição de frequências

Conforme apresentado na figura 3.2, identifica-se um evidente desbalanceamento na distribuição de frequências da variável que se deseja prever, *'AZ_FULL'*, correspondente à decisão da Azul em relação à entrada na dada rota e ano.

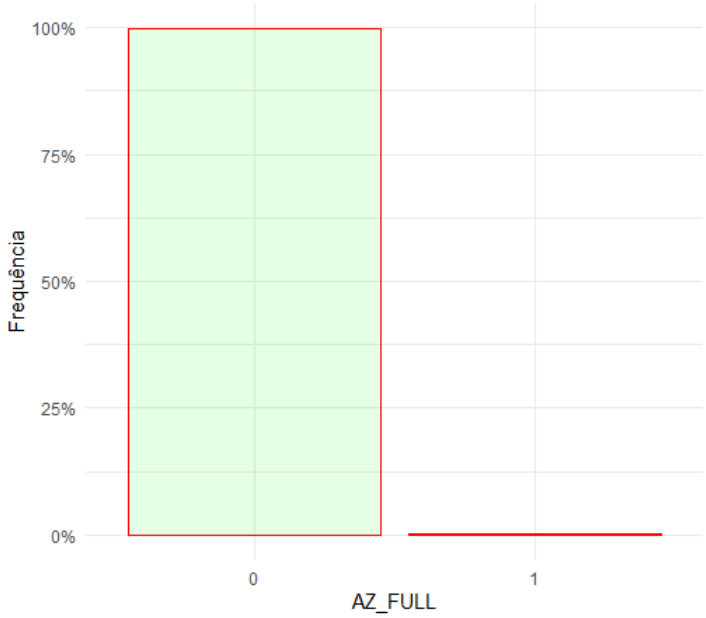


FIGURA 3.2 – Histograma da variável *'AZ_FULL'*.
Fonte: autoria própria, 2022.

Dado isso, analisa-se as distribuições de frequências de todas as variáveis para encontrar padrões que nos permitam identificar segmentações de base com melhor proporção de balanceamento. Dentre todas segmentações analisadas, as amostras minimamente balanceadas correspondem aos filtros *LCCCOMP=1* e *NEW=1*, cujos histogramas para a variável *'AZ_FULL'* estão apresentados nas figuras 3.3 e 3.4.

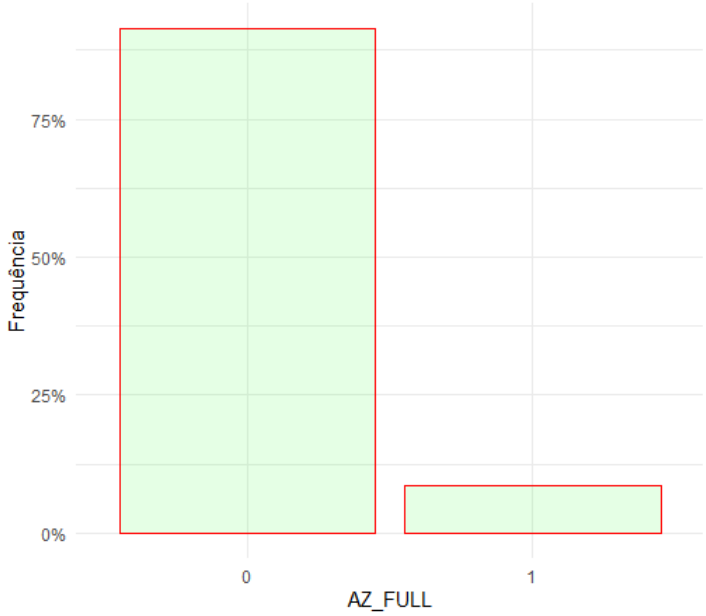


FIGURA 3.3 – Histograma da variável *'AZ_FULL'* para a segmentação *LCCCOMP=1*.
Fonte: autoria própria, 2022.

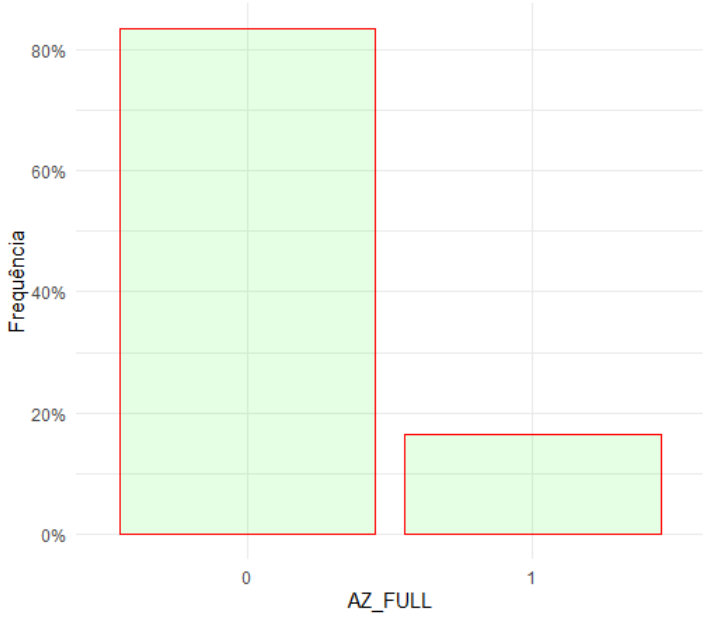


FIGURA 3.4 – Histograma da variável *'AZ_FULL'* para a segmentação *NEW=1*.
Fonte: autoria própria, 2022.

Apesar de nenhum destes recortes corresponder de fato a uma base balanceada, é perceptível uma discrepância muito menor nas frequências de aparição das classes, se comparadas ao histograma da base completa. Entretanto, como não existe segmentação do conjunto que seja naturalmente balanceada, precisaremos obter os dados balanceados por meio da reamostragem.

3.3.4 Análise dos atributos

Fernández et al. (2018) afirmam que a remoção de variáveis irrelevantes e/ou redundantes pode diminuir o ruído nos dados de entrada, especialmente para as observações de classe minoritária. Dessa forma, a análise inicial dos atributos consiste em identificar as correlações entre as variáveis do conjunto de dados, eliminando as redundâncias e aprimorando a performance preditiva.

A metodologia aplicada para a análise de correlações é baseada no Coeficiente de Correlação de Pearson (PCC), que representa a métrica da relação estatística entre duas variáveis contínuas. Esta famosa estratégia de mensuração da associação entre variáveis se baseia no método de covariância, fornecendo informações sobre a magnitude da correlação, bem como a direção do relacionamento ((SCHOBER *et al.*, 2018)). Aliado a isso, a análise de correlações levou em consideração a significância da correção através da análise do *p-value*.

Para a análise do coeficiente de correlação de Pearson, Schober et al. (2018) sugerem que módulos dos coeficientes maiores que 0,70 atestam uma forte correlação. Além disso,

Andrade (2019) considera como estatisticamente significantes as correlações com p $< 5\%$.

A matriz de correlações com as devidas associações não significativas, representadas pelo *'X'*, está apresentada na figura 3.5. O diâmetro das circunferências representa a magnitude da correlação.

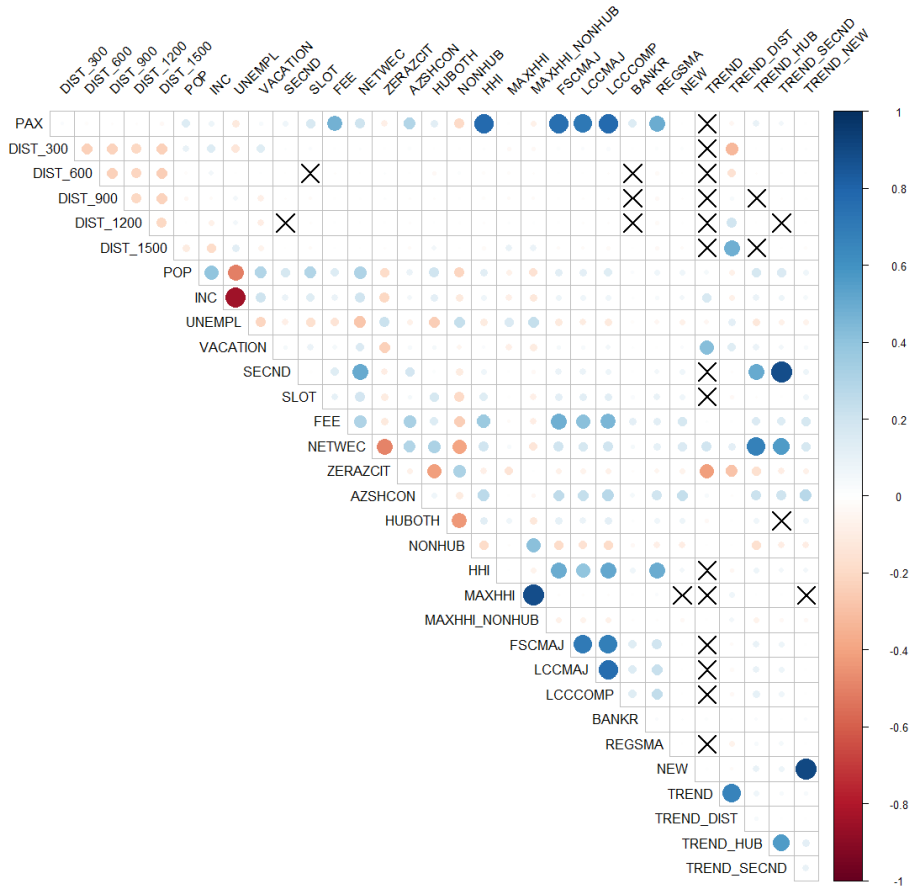


FIGURA 3.5 – Matriz de correlações considerando significâncias das associações. Fonte: autoria própria, 2022.

Observa-se, portanto, fortes correlações entre as variáveis:

- *'PAX'* e *'HHI'*;
- *'PAX'* e *'FSCMAJ'*;
- *'PAX'* e *'LCCMAJ'*;
- *'PAX'* e *'LCCCOMP'*;
- *'INC'* e *'UNEMPL'*;
- *'SECND'* e *'TREND_SECND'*;
- *'MAXHHI'* e *'MAXHHI_NONHUB'*;

- *'FSCMAJ'* e *'LCCMAJ'*;
- *'LCCMAJ'* e *'LCCCOMP'*;
- *'NEW'* e *'TREND_NEW'*.

3.3.5 Dados atípicos

Por fim, a análise exploratória realiza a busca por dados atípicos, como por exemplo *missing values* e *outliers*. O conjunto de dados utilizado neste trabalho não possui valores faltantes. As análises de *outliers* são aplicadas somente em variáveis numéricas e contínuas. Desse modo, considerando apenas as variáveis que se enquadram nas condições anteriores, a identificação de possíveis valores anormais é feita através das visualizações de *box-plot* para cada um desses atributos, conforme constam na figura 3.6.

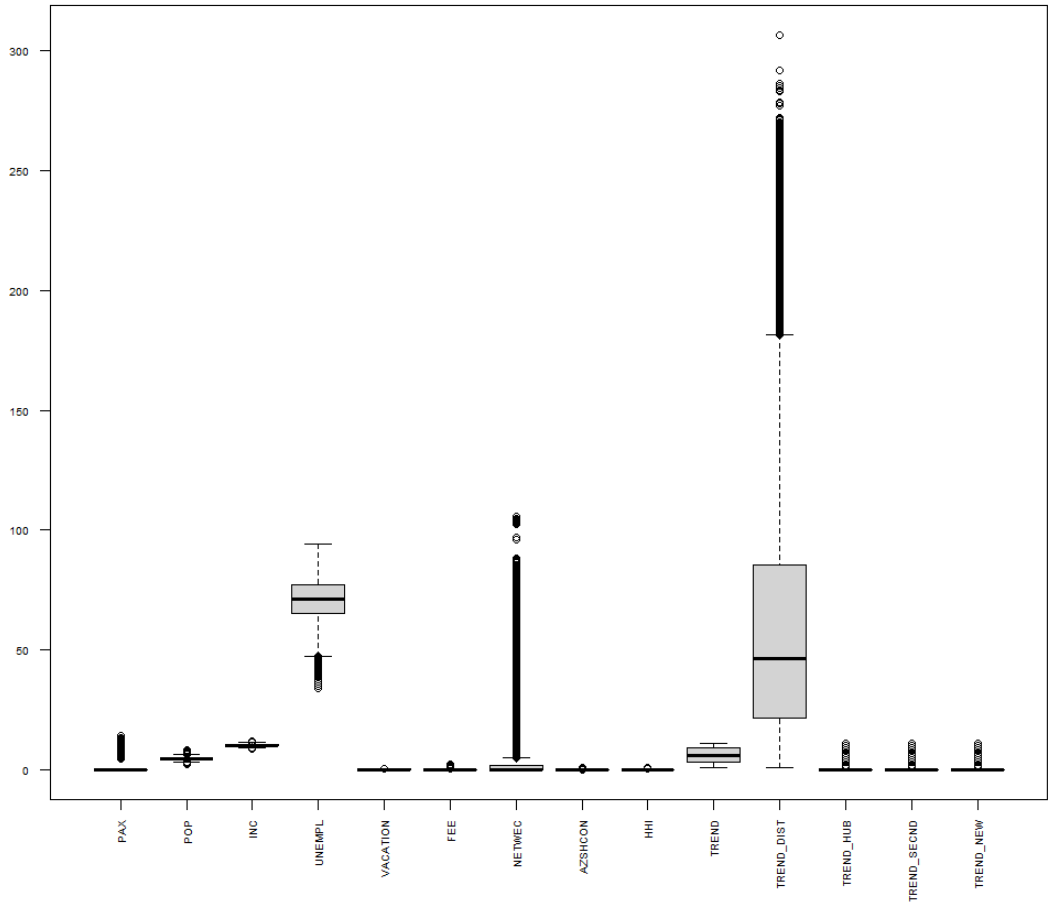


FIGURA 3.6 – *Box-plot* das variáveis numéricas e contínuas.
Fonte: autoria própria, 2022.

Os valores que estão presentes acima do quartil superior e abaixo do quartil inferior

podem ser considerados como dados discrepantes. Entretanto, Carreño et al. (2019) analisaram eventos raros, anomalias, novidades e *outliers* e categorizam cada um destes da seguinte maneira:

- Raro:
  Dados temporais com todas as classes representadas no conjunto de treino. Costuma ocorrer em classificações supervisionadas de dados com desequilíbrio de classes.
- Anomalia:
  Dados estáticos com todas as classes representadas no conjunto de treino. Costuma ocorrer em classificações supervisionadas de dados com desequilíbrio de classes.
- Novidade:
  Em geral surge nos casos de classificação supervisionada de dados possivelmente desbalanceados, cujo conjunto de treinamento contém apenas uma classe.
- *Outlier*:
  Normalmente associado a casos de classificação não-supervisionada.

Portanto, no contexto do presente trabalho, os dados com valores anormais enquadram-se na categoria de eventos raros, descartando a necessidade de remoção destas observações.

3.4 Pré-processamento dos dados

3.4.1 Transformação de dados

O tratamento inicial que foi desenvolvido na base de dados consiste na transformação das informações que contêm os aeroportos de origem e destino, atribuindo-lhes a sigla corresponde ao estado em que o respectivo aeroporto se encontra, uma vez que também foram analisadas as segmentações regionais.

3.4.2 Seleção de atributos

Diante das correlações identificadas através da análise dos coeficientes de Pearson considerando as significâncias das associações, as variáveis redundantes e que, por conseguinte, devem ser removidas da análise são:

- *'PAX'*;
- *'UNEMPL'*;

- *'TREND_SECND'*;
- *'MAXHHI_NONHUB'*;
- *'LCCMAJ'*;
- *'TREND_NEW'*.

Por outro lado, a modelagem não utiliza as variáveis *'k'* e *'year'* da base, que representam o par de aeroportos e o ano das informações, respectivamente. Além disso, não utilizaremos também as variáveis *'AZ_BEF'* e *'AZ_AFT'*, que informam se a operação naquela rota foi iniciada no dado ano considerando o período antes da fusão Azul-Trip (2008-2011) ou depois deste evento de fusão (2012-2018).

Por fim, as reamostragens implementadas podem fornecer conjuntos de dados que exijam a retirada pontual de alguma variável, fato que surge como consequência da normalização de possíveis atributos constantes na amostra avaliada.

3.4.3 Reamostragem dos dados

Visando minimizar as discrepâncias entre distribuições de classes e construir bases balanceadas segundo variadas proporções, utilizam-se métodos de reamostragem através da aplicação das técnicas de pré-processamento (KRAWCZYK, 2016).

Levando em consideração as informações levantadas na revisão de literatura, os métodos adotados consistem em subamostragem, sobreamostragem e a combinação destes, respectivamente descritos abaixo.

- Subamostragem:
  Processo de diminuir a quantidade de observações ou amostras da classe majoritária. Adota-se a estratégia de remoção aleatória de observações, definindo o método utilizado como *Random Undersampling*.

- Sobreamostragem:
  Estratégia de aumento da quantidade de observações da classe minoritária, implementada através da produção de novas observações ou replicação de algumas. Adotando novamente o caráter de aleatoriedade, o método escolhido foi o de *Random Oversampling*.

- Abordagem híbrida:
  Visando amenizar as desvantagens associadas a cada um dos métodos anteriores, implementa-se ainda a análise da combinação dos métodos de *Random Oversampling*

e *Random Undersampling*, por meio da qual ocorre, conjuntamente, a remoção de exemplos majoritários e adição de exemplos minoritários.

Para as análises das aplicações individuais dos métodos de sub e sobreamostragem, comparou-se os dados com as classes minoritária e majoritária equilibradas nas proporções 1:1, 1:2, 1:3 e 1:4, enquanto na abordagem híbrido utilizou-se a proporção 1:1.

3.4.4 Separação em treinamento e teste

Diferentes separações de treinamento e teste foram implementadas, de forma que o conjunto de teste realizou a previsão de rotas para todos os anos da base de dados. Para cada ano de teste utilizado, foram implementados como conjunto de treino os dados referentes aos anos anteriores. Por exemplo, para uma previsão de rotas de 2018, utiliza-se como base de treinamento os dados de 2008 a 2017.

3.4.5 Normalização

A normalização de dados é uma etapa essencial no pré-processamento de dados para os casos em que o objetivo é empregar um algoritmo baseado na minimização de erros. O método aplicado na normalização foi o de *Standard Scaler*, que atua padronizando os dados pela remoção da média, anulando seu valor, e escalando a variância a uma unidade. O cálculo utilizado está apresentado na equação 3.1.

$$X_{norm,(i,j)} = \frac{X_{(i,j)} - mean(X_j)}{std(X_j)} \tag{3.1}$$

Segundo Ferreira et al. (2019), como a determinação do valor normalizado depende apenas da média e da variância, tal método apresenta vantagens como ser linear, reversível, rápido e altamente escalável. Por outro lado, o *Standard Scaler* também apresenta algumas desvantagens, pois manifesta uma alta sensibilidade a *outliers* e é mais adequado para dados normalmente distribuídos.

3.5 Modelagem preditiva

3.5.1 Métodos

Com base nos resultados obtidos pelos trabalhos investigados na revisão de literatura, foram testados alguns algoritmos para a análise da predição. Como o desempenho predi-

tivo está associado às características da base de dados, é importante encontrar um modelo que melhor se adequa ao conjunto de informações e gera bons resultados.

3.5.1.1 Rede neural

Para resolver problemas de classificação, a rede neural artificial é um dos principais métodos de aprendizagem que pode ser utilizado. A estrutura da rede neural contém camadas de entrada e saída, com possíveis múltiplas camadas ocultas. Cada nó na rede neural é chamado de neurônio, e as sinapses que conectam os neurônios contêm pesos. A retropropagação pode ser aplicada para distribuir erros de volta à rede e ajustar os pesos entre os neurônios. Isso resulta em um modelo de aprendizado iterativo e mais preciso (SRIRATANAWILAI; ERJONGMANEE, 2018).

3.5.1.2 Regressão logística

Uma regressão logística pode ser vista como um classificador linear para uma variável dependente binária. Consiste em encontrar o vetor de pesos, w, que minimize, sobre todas as amostras, a soma dos quadrados dos erros entre a função de predição $\sigma(w^T x)$ e o valor verdadeiro para cada amostra x. No entanto, a regressão logística é um modelo simples, produzindo um limite de decisão linear. Como suas saídas podem ser facilmente descritas em termos de suas entradas, este método nos fornece um modelo altamente interpretável (FERREIRA *et al.*, 2019).

3.5.1.3 KNN

O algoritmo *k-nearest neighbors* (KNN) é um método de aprendizado supervisionado usado para classificação e regressão. Em ambos os casos, a entrada consiste nos k exemplos de treinamento mais próximos em um conjunto de dados. Na classificação, o modelo de aprendizagem consiste em classificar um objeto por seus vizinhos, com o objeto sendo atribuído à classe mais comum entre seus k vizinhos mais próximos (k é um inteiro positivo, tipicamente pequeno).

3.5.1.4 Árvore de decisão

Uma árvore de decisão é um classificador que é modelado em uma estrutura semelhante a uma árvore contendo nós internos, ramos e nós terminais. Pode ser interpretado como um classificador não-linear que divide a entrada em diferentes regiões visando maximizar a informação obtida em cada divisão. Árvores de Decisão podem desenhar complexos

limites de decisões e também fornecem resultados muito interpretáveis (FERREIRA *et al.*, 2019).

3.5.1.5 Método adotado

Visando comparar os métodos acima, realiza-se a predição de rotas para o ano de 2018 através da base obtida pelo *Random Undersampling* na proporção de 1:1, pois esta razão entre classes gera um menor conjunto de treino e, consequentemente, um tempo de execução hábil para esta análise comparativa.

A análise preliminar dos resultados da aplicação dos algoritmos descritos confirma o que foi observado em trabalhos de outros autores (Srisaeng et al. em 2015, Sriratanawilai e Erjongmanee em 2018), com o método de redes neurais desempenhando melhor em todas as métricas, exceto tempo de treinamento, conforme observamos na tabela 3.3.

TABELA 3.3 – Métricas das variáveis presentes no conjunto de dados.
Fonte: autoria própria, 2022.

Método	**Acurácia**	**Sensib.**	**Especif.**	**Precisão**	***F1-score***	***G-Mean***	**Tempo (s)**
Redes neurais	98,75%	98,79%	98,72%	98,79%	98,79%	98,75%	0,582
Regressão logística	95,33%	95,15%	95,51%	95,73%	95,44%	95,33%	0,054
KNN	89,72%	87,27%	92,31%	92,31%	89,72%	89,75%	0,011
Árvore de decisão	91,28%	93,33%	89,10%	90,06%	91,67%	91,19%	0,047

Portanto, a modelagem preditiva adotada na previsão das rotas daqui em diante será baseada em redes neurais artificiais.

3.5.2 Parâmetros do modelo

A implementação do treinamento do modelo de rede neural recebe como parâmetros os atributos e observações a partir das quais o aprendizado de máquina deve atuar, a configuração de camadas escondidas e quantidade de neurônios por camada, além da função de ativação utilizada na modelagem.

Os atributos restantes, após a seleção de variáveis, totalizam 29 variáveis. Tomando como base essa quantidade de variáveis de *input* e sabendo que nosso problema de classificação gera como *output* uma classificação binária, podemos estimar as quantidades ideais de neurônios por camada através das seguintes regras propostas por Heaton (2015):

- O número de neurônios ocultos deve estar entre o tamanho da camada de entrada e o tamanho da camada de saída.

- O número de neurônios ocultos deve ser inferior a 2/3 do tamanho da camada de entrada, mais o tamanho da camada de saída.
- O número de neurônios ocultos deve ser menor que o dobro do tamanho da camada de entrada.

Dessa forma, a estrutura de camadas e neurônios escondidos consiste em 5 camadas, contendo, respectivamente, 16, 9, 6, 4 e 3 neurônios escondidos. Por fim, a função de ativação adotada foi a logística, devido ao seu bom desempenho em problemas de classificação (HEATON, 2015).

3.5.3 Validação

A estratégia de validação adotada é o método de *K-fold cross validation*, que consiste em um procedimento de reamostragem usado para avaliar modelos de aprendizado de máquina em uma amostra de dados, identificando possíveis problemas de ajuste como *overfitting* ou *underfitting*.

É um método popular por ser de fácil entendimento e por geralmente resultar em uma estimativa menos tendenciosa ou menos otimista da performance do modelo do que outros métodos, como uma simples divisão em treino e teste.

O procedimento possui um único parâmetro chamado k que se refere ao número de grupos em que uma determinada amostra de dados deve ser dividida. Quando um valor específico para k é escolhido, ele pode ser usado no lugar de k na referência ao modelo, como k=10 resultando em *10-fold cross validation*. O processo geral é:

1. Embaralhar o conjunto de dados aleatoriamente;
2. Dividir o conjunto de dados em k grupos;
3. Para cada grupo, utiliza-se o grupo como conjunto de teste, considerando o restante como conjunto de treino;
4. Ajusta-se, por grupo, um modelo de aprendizagem no conjunto de treinamento e avalia a predição do conjunto de teste;
5. Resume o desempenho do modelo através da análise conjunta dos resultados de cada grupo.

É importante ressaltar que cada observação na amostra de dados é atribuída a um grupo individual e permanece nesse grupo durante o procedimento. Isso significa que cada amostra tem a oportunidade de ser usada como conjunto de teste 1 vez e para treinar o modelo k-1 vezes.

3.5.4 Métricas de desempenho

Para atribuir as métricas de desempenho, considerou-se a seguinte nomenclatura:

- TN (*True Negatives*) é o número de exemplos 0 corretamente classificados;
- FN (*False Negatives*) é o número de 1's incorretamente classificados como 0's;
- FP (*False Positives*) é o número de 0's incorretamente classificados como 1's;
- TP (*True Positives*) é o número de exemplos 1 corretamente classificados.

Desse modo, as métricas de desempenho consideradas na análise dos resultados e seus respectivos cálculos estão listados nas equações 3.2, 3.3, 3.4, 3.5, 3.6 e 3.7.

- Acurácia:

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP} \tag{3.2}$$

- Sensibilidade ou *Recall*:

$$Sensitivity = \frac{TP}{FN + TP} \tag{3.3}$$

- Especificidade:

$$Specificity = \frac{TN}{TN + FP} \tag{3.4}$$

- Precisão:

$$Precision = \frac{TP}{FP + TP} \tag{3.5}$$

- *F1-score* ou *F-Measure*:

$$F1 - score = 2 \cdot \frac{Sensitivity \cdot Precision}{Sensitivity + Precision} \tag{3.6}$$

- *G-Mean* ou *G-Measure*:

$$G - Mean = \sqrt{Sensitivity \cdot Specificity} \tag{3.7}$$

- Tempo de treinamento:
  Tempo de execução medido somente para o treinamento do modelo de aprendizagem, que é a etapa lenta do processo.

4 Análise de resultados

A análise dos resultados utilizando as estratégias de reamostragem segmenta-se em 4 seções, pois em cada abordagem foram utilizadas estratégias comparativas diferentes em busca do modelo preditivo mais eficaz para cada caso. As diferentes técnicas aplicadas foram:

1. Subamostragem da base de dados variando a proporção de balanceamento e o ano usado como conjunto de teste;
2. Sobreamostragem da base de dados segmentada por região, variando a proporção de balanceamento e utilizando os dados de 2018 como conjunto de teste;
3. Combinação de subamostragem e sobreamostragem da base de dados segmentada por região, variando o ano usado como conjunto de teste e utilizando a proporção de balanceamento 1:1;
4. Comparação dos métodos de sobreamostragem e a abordagem híbrida, realizando a previsão de rotas para o ano de 2018 em todo o território nacional, totalizando um conjunto de treinamento de aproximadamente 2.000.000 de observações.

4.1 RUS por ano e por balanceamento

Utilizando as proporções de balanceamento das classes 1:1, 1:2, 1:3 e 1:4, realiza-se a modelagem preditiva no conjunto obtido pelo método de *Random Undersampling* para todos os anos de 2009 a 2018. Os resultados obtidos estão apresentados na tabelas 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 e 4.10.

TABELA 4.1 – Resultados do conjunto de teste com dados de 2018.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	98,75%	98,79%	98,72%	98,79%	98,79%	98,75%	0,58
1:2	97,98%	98,79%	97,58%	95,32%	97,02%	98,18%	2,12
1:3	98,63%	98,79%	98,57%	95,88%	97,31%	98,68%	1,62
1:4	97,72%	98,18%	97,61%	91,01%	94,46%	97,89%	9,23

TABELA 4.2 – Resultados do conjunto de teste com dados de 2017.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	97,96%	99,23%	96,95%	96,27%	97,73%	98,08%	0,48
1:2	96,20%	97,69%	95,64%	89,44%	93,38%	96,66%	1,13
1:3	97,81%	97,69%	97,84%	92,03%	94,78%	97,77%	1,87
1:4	97,32%	94,62%	97,83%	89,13%	91,79%	96,21%	42,07

TABELA 4.3 – Resultados do conjunto de teste com dados de 2016.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	98,86%	99,20%	98,40%	98,81%	99,01%	98,80%	0,46
1:2	96,40%	94,82%	97,60%	96,75%	95,77%	96,20%	0,87
1:3	97,06%	93,23%	99,00%	97,91%	95,51%	96,07%	1,23
1:4	91,67%	74,10%	98,34%	94,42%	83,04%	85,36%	24,77

TABELA 4.4 – Resultados do conjunto de teste com dados de 2015.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	95,89%	96,65%	95,06%	95,58%	96,11%	95,85%	0,40
1:2	97,11%	97,21%	97,06%	94,57%	95,87%	97,13%	0,82
1:3	97,85%	96,65%	98,27%	95,05%	95,84%	97,46%	0,77
1:4	96,97%	91,06%	98,53%	94,22%	92,61%	94,72%	10,64

TABELA 4.5 – Resultados do conjunto de teste com dados de 2014.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	97,25%	99,39%	95,06%	95,35%	97,33%	97,20%	0,36
1:2	97,54%	98,79%	96,90%	94,22%	96,45%	97,84%	0,75
1:3	97,72%	98,79%	97,36%	92,61%	95,60%	98,07%	0,70
1:4	96,47%	91,52%	97,66%	90,42%	90,96%	94,54%	5,94

TABELA 4.6 – Resultados do conjunto de teste com dados de 2013.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	99,18%	100,00%	98,35%	98,38%	99,18%	99,17%	0,30
1:2	98,16%	98,90%	97,78%	95,74%	97,30%	98,34%	0,51
1:3	99,20%	100,00%	98,94%	96,81%	98,38%	99,47%	0,58
1:4	96,90%	90,11%	98,61%	94,25%	92,13%	94,27%	4,06

TABELA 4.7 – Resultados do conjunto de teste com dados de 2012.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	53,52%	39,56%	100,00%	100,00%	56,69%	62,90%	0,30
1:2	59,51%	35,82%	99,71%	99,53%	52,68%	59,77%	0,34
1:3	64,91%	34,80%	99,61%	99,03%	51,51%	58,88%	0,79
1:4	70,39%	36,50%	99,71%	99,08%	53,35%	60,33%	1,35

TABELA 4.8 – Resultados do conjunto de teste com dados de 2011.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	97,59%	95,29%	98,78%	97,59%	96,43%	97,02%	0,20
1:2	98,39%	94,12%	99,43%	97,56%	95,81%	96,74%	0,28
1:3	96,60%	77,65%	99,80%	98,51%	86,84%	88,03%	0,41
1:4	97,63%	83,53%	99,41%	94,67%	88,75%	91,12%	0,51

TABELA 4.9 – Resultados do conjunto de teste com dados de 2010.
Fonte: autoria própria, 2022.

Razão	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
1:1	94,30%	85,71%	98,10%	95,24%	90,23%	91,70%	0,14
1:2	92,70%	62,86%	99,08%	93,62%	75,21%	78,92%	0,19
1:3	97,59%	82,86%	99,61%	96,67%	89,23%	90,85%	0,31
1:4	96,60%	65,71%	99,85%	97,87%	78,63%	81,00%	0,34

TABELA 4.10 – Resultados do conjunto de teste com dados de 2009.
Fonte: autoria própria, 2022.

Razão	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
1:1	97,09%	86,11%	99,41%	96,88%	91,18%	92,52%	0,60
1:2	97,94%	83,33%	99,67%	96,77%	89,55%	91,14%	0,76
1:3	97,21%	72,22%	99,14%	86,67%	78,79%	84,62%	0,15
1:4	97,87%	61,11%	100,00%	100,00%	75,86%	78,17%	0,15

Calculando as médias de cada uma das métricas por ano usado como conjunto de teste, temos os resultados apresentados na tabela 4.11.

TABELA 4.11 – Médias dos resultados por ano.
Fonte: autoria própria, 2022.

Ano	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
2018	98,27%	98,64%	98,12%	95,25%	96,90%	98,38%	3,39
2017	97,32%	97,31%	97,06%	91,72%	94,42%	97,18%	11,39
2016	96,00%	90,34%	98,33%	96,97%	93,33%	94,11%	6,83
2015	96,96%	95,39%	97,23%	94,85%	95,11%	96,29%	3,16
2014	97,24%	97,12%	96,75%	93,15%	95,09%	96,91%	1,94
2013	98,36%	97,25%	98,42%	96,30%	96,75%	97,81%	1,36
2012	62,08%	36,67%	99,76%	99,41%	53,56%	60,47%	0,69
2011	97,55%	87,65%	99,35%	97,08%	91,96%	93,23%	0,35
2010	95,30%	74,29%	99,16%	95,85%	83,33%	85,62%	0,25
2009	97,53%	75,69%	99,56%	95,08%	83,84%	86,61%	0,41

As variações temporais destes resultados acima estão expressas nas figuras 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 e 4.7.

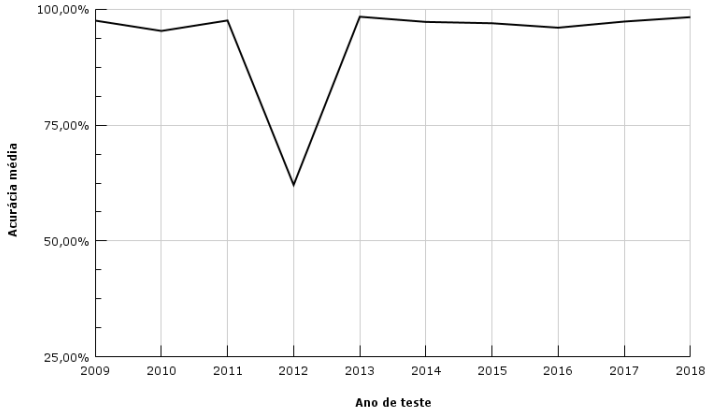


FIGURA 4.1 – Acurácia média por ano de teste.
Fonte: autoria própria, 2022.

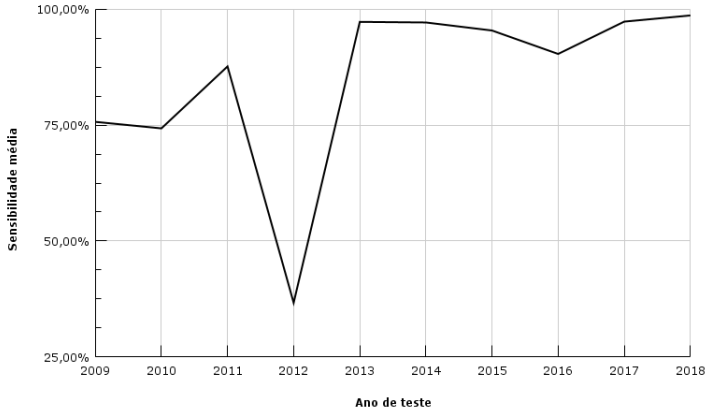


FIGURA 4.2 – Sensibilidade média por ano de teste.
Fonte: autoria própria, 2022.

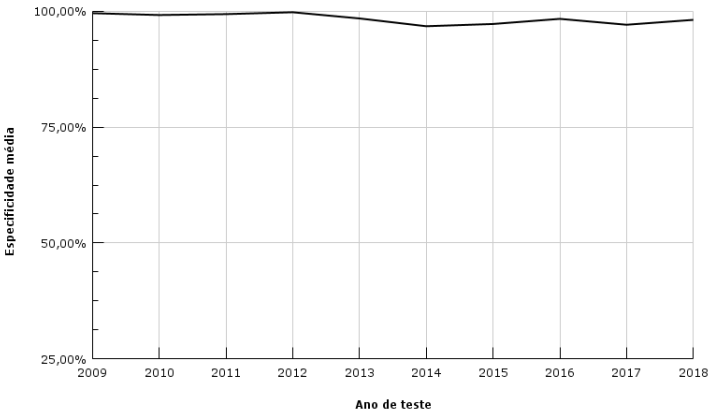


FIGURA 4.3 – Especificidade média por ano de teste.
Fonte: autoria própria, 2022.

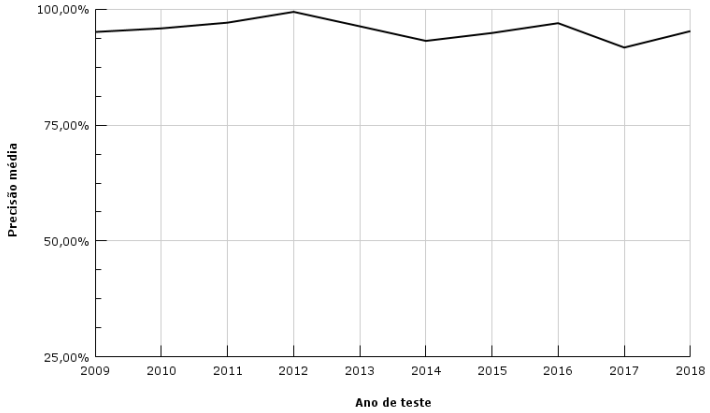


FIGURA 4.4 – Precisão média por ano de teste.
Fonte: autoria própria, 2022.

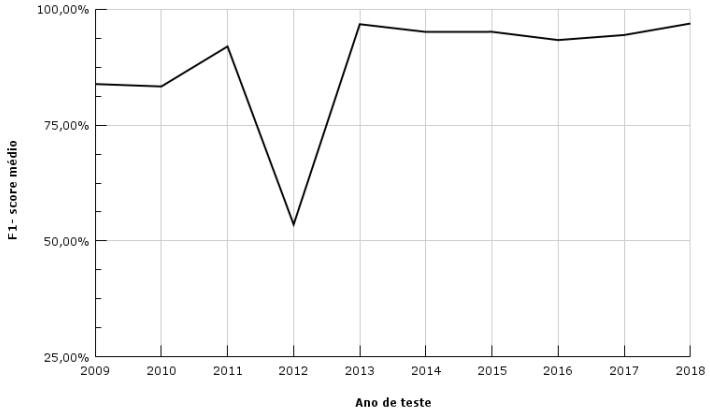


FIGURA 4.5 – *F1-score* médio por ano de teste.
Fonte: autoria própria, 2022.

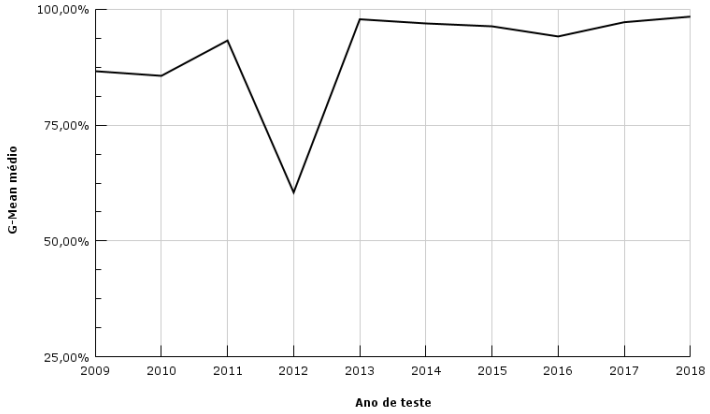


FIGURA 4.6 – *G-Mean* médio por ano de teste.
Fonte: autoria própria, 2022.

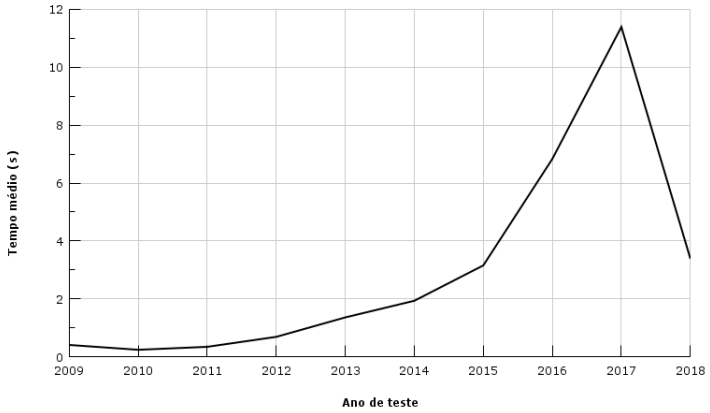


FIGURA 4.7 – Tempo médio por ano de teste.
Fonte: autoria própria, 2022.

Percebe-se que os gráficos das médias da acurácia, sensibilidade, *F1-score* e *G-Mean* seguem a mesma tendência, o mesmo ocorrendo para os gráficos das médias da especificidade e precisão. Era esperado o decréscimo do tempo com a utilização de um conjunto de treinamento menor (quanto menor o ano do conjunto de teste, menor o tamanho do conjunto de treinamento, pois treina-se o modelo com os anos anteriores ao de teste). Observa-se ainda que as previsões com anos de teste 2009, 2010 e 2011 possuem tempo médio de execução bem abaixo dos demais devido ao menor tamanho da base de treinamento, além do tempo médio para a base de teste de 2017 que foi consideravelmente acima dos demais. Em 2012 ocorreu a fusão da Azul com a Trip Airlines, consequência do pior desempenho na modelagem preditiva de rotas para este ano. O modelo utilizando o ano de teste de 2018, além de possuir tempo médio de treinamento aceitável, teve o melhor desempenho nas métricas de sensibilidade, *F1-score* e *G-Mean*, sendo, portanto, o conjunto de teste com os melhores resultados preditivos para a abordagem de subamostragem.

Por outro lado, calculando as médias das métricas de previsão por razão de balanceamento, temos os resultados apresentados na tabela 4.12.

TABELA 4.12 – Médias dos resultados por razão de balanceamento.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	97,43%	95,60%	97,65%	96,99%	96,22%	96,57%	0,39
1:2	96,94%	91,83%	97,86%	94,89%	92,93%	94,57%	0,83
1:3	97,74%	90,87%	98,72%	94,68%	92,48%	94,56%	0,85
1:4	96,57%	83,33%	98,65%	94,00%	87,58%	90,37%	10,86

As visualizações gráficas dos resultados por razão de balanceamento estão expressas

nas figuras 4.8, 4.9, 4.10, 4.11, 4.12, 4.13 e 4.14.

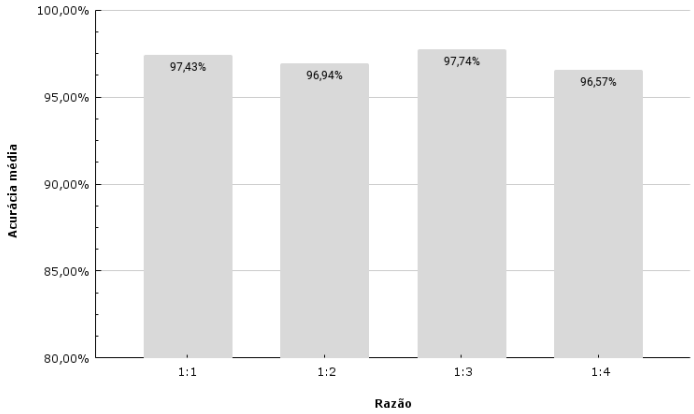


FIGURA 4.8 – Acurácia média por razão de balanceamento.
Fonte: autoria própria, 2022.

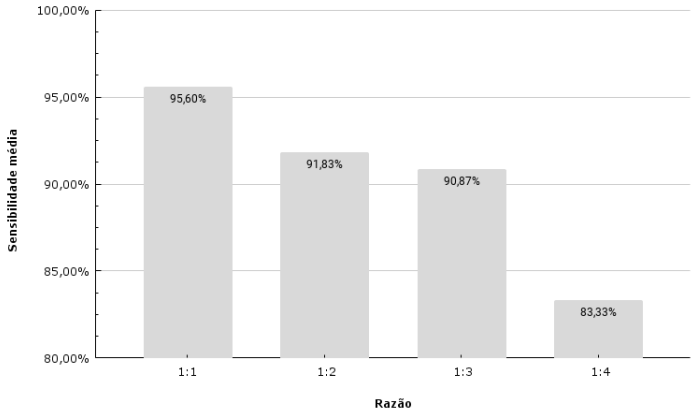


FIGURA 4.9 – Sensibilidade média por razão de balanceamento.
Fonte: autoria própria, 2022.

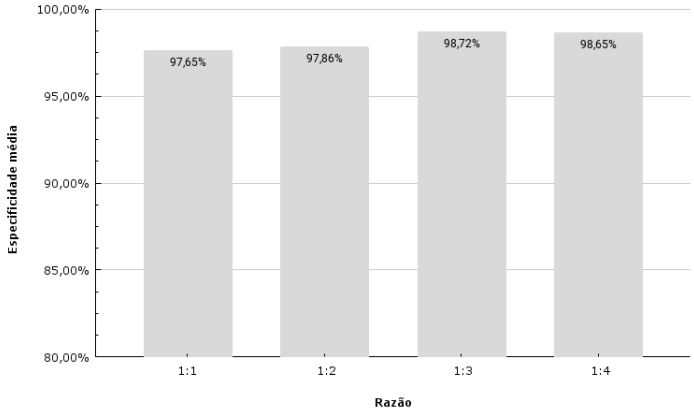


FIGURA 4.10 – Especificidade média por razão de balanceamento.
Fonte: autoria própria, 2022.

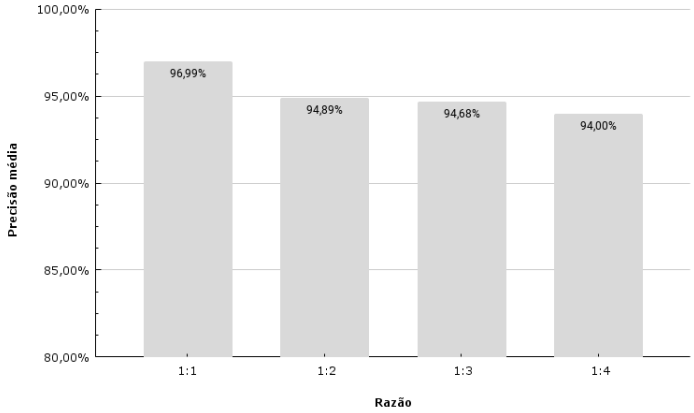


FIGURA 4.11 – Precisão média por razão de balanceamento.
Fonte: autoria própria, 2022.

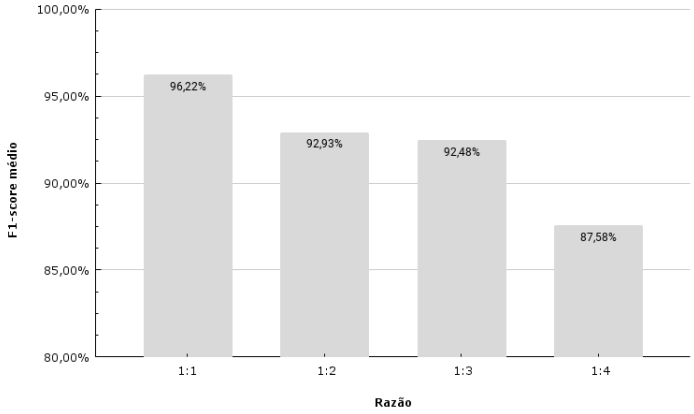


FIGURA 4.12 – *F1-score* médio por razão de balanceamento.
Fonte: autoria própria, 2022.

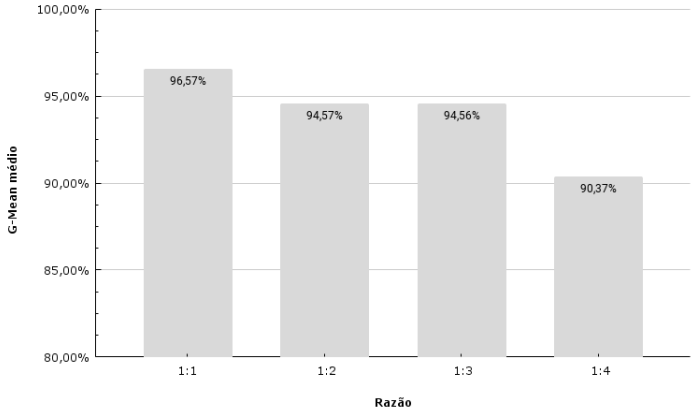


FIGURA 4.13 – *G-Mean* médio por razão de balanceamento.
Fonte: autoria própria, 2022.

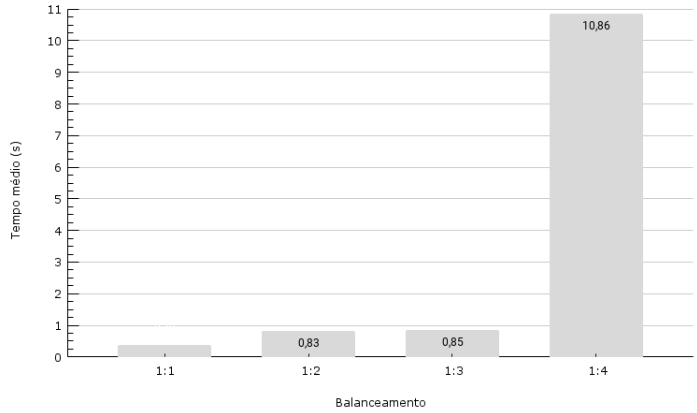


FIGURA 4.14 – Tempo médio por razão de balanceamento.
Fonte: autoria própria, 2022.

Observa-se os valores médios da acurácia e da especificidade variando pouco para as proporções de balanceamento analisadas. Por outro lado, nota-se uma diminuição dos valores médios da sensibilidade, precisão, *F1-score* e *G-Mean* com o aumento da desproporção. Portanto, percebe-se que a razão de balanceamento 1:4 fornece os piores desempenhos preditivos, além do maior tempo de treinamento do modelo.

Mediante o exposto, não utilizaremos o conjunto com proporção 1:4 para a validação do modelo. Portanto, iremos aplicar o métodos de validação *K-fold cross validation* nos dados com razões de balanceamento 1:1, 1:2 e 1:3. Para as três razões de balanceamento aplicamos o método *8-fold cross validation*, cujos resultados de acurácia e tempo de treinamento estão apresentados nas tabelas 4.13, 4.14 e 4.15.

TABELA 4.13 – *8-fold cross validation* para o conjunto na razão de balanceamento 1:1.
Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	97,63%	0,86
2	98,49%	2,38
3	97,41%	0,66
4	98,92%	1,00
5	98,49%	1,09
6	98,71%	0,72
7	98,28%	1,01
8	98,71%	0,70
Acurácia média		**98,33%**
Desvio padrão da acurácia		**0,54%**
Tempo médio (s)		**1,05**

TABELA 4.14 – *8-fold cross validation* para o conjunto na razão de balanceamento 1:2. Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	98,99%	0,80
2	98,56%	2,50
3	96,41%	7,03
4	97,99%	3,69
5	98,13%	2,40
6	98,85%	2,27
7	97,99%	1,29
8	98,85%	0,85
Acurácia média		**98,22%**
Desvio padrão da acurácia		**0,84%**
Tempo médio (s)		**2,60**

TABELA 4.15 – *8-fold cross validation* para o conjunto na razão de balanceamento 1:3. Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	97,41%	30,10
2	98,92%	6,15
3	98,92%	7,80
4	97,31%	12,93
5	98,92%	6,49
6	98,06%	2,04
7	98,81%	2,92
8	97,63%	4,33
Acurácia média		**98,25%**
Desvio padrão da acurácia		**0,73%**
Tempo médio (s)		**9,09**

Portanto, com base nos resultados da validação do modelo aplicando a estratégia de subamostragem, percebe-se a modelagem com dados na proporção de balanceamento 1:1 desempenhando melhor, visto que possui a maior acurácia dentre os modelos analisados e um tempo médio de treinamento da rede neural consideravelmente menor que os demais.

4.2 ROS por região e por balanceamento

Utilizando as proporções de balanceamento das classes 1:1, 1:2, 1:3 e 1:4, realiza-se a modelagem preditiva no conjunto obtido pelo método de *Random Oversampling* para o ano de 2018 segmentada por região do território brasileiro. Os resultados obtidos estão apresentados nas tabelas 4.16, 4.17, 4.18, 4.19 e 4.20.

TABELA 4.16 – Resultados para a segmentação da região Norte.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	88,19%	71,45%	97,27%	93,42%	80,98%	83,37%	270,01
1:2	98,56%	100,00%	98,18%	93,47%	96,63%	99,09%	110,63
1:3	96,40%	86,86%	98,06%	88,62%	87,73%	92,29%	250,40
1:4	97,53%	87,13%	98,84%	90,49%	88,78%	92,80%	185,79

TABELA 4.17 – Resultados para a segmentação da região Nordeste.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	66,34%	41,53%	98,78%	97,80%	58,30%	64,05%	324,97
1:2	84,45%	62,74%	98,80%	97,20%	76,26%	78,73%	155,82
1:3	89,26%	68,45%	98,25%	94,40%	79,36%	82,01%	115,44
1:4	84,46%	43,16%	98,22%	88,98%	58,12%	65,11%	245,62

TABELA 4.18 – Resultados para a segmentação da região Centro-Oeste.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	76,61%	60,80%	97,10%	96,45%	74,58%	76,83%	23,07
1:2	83,33%	60,14%	98,26%	95,70%	73,86%	76,87%	84,20
1:3	87,73%	63,54%	98,20%	93,86%	75,78%	78,99%	10,17
1:4	84,62%	41,52%	99,07%	93,75%	57,55%	64,14%	35,95

TABELA 4.19 – Resultados para a segmentação da região Sudeste.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	85,37%	68,41%	96,02%	91,53%	78,30%	81,05%	459,73
1:2	91,77%	74,42%	97,17%	89,10%	81,10%	85,04%	426,51
1:3	91,93%	62,52%	97,98%	86,44%	72,56%	78,27%	931,33
1:4	92,42%	51,86%	98,83%	87,46%	65,11%	71,59%	4.799,35

TABELA 4.20 – Resultados para a segmentação da região Sul.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	69,01%	47,12%	98,04%	96,96%	63,42%	67,97%	76,43
1:2	89,15%	76,26%	97,83%	95,94%	84,98%	86,37%	77,33
1:3	86,66%	60,07%	98,30%	93,91%	73,27%	76,84%	94,31
1:4	92,09%	74,01%	98,04%	92,56%	82,25%	85,18%	49,70

Calculando as médias das métricas por região, temos os resultados apresentados na tabela 4.21.

TABELA 4.21 – Médias dos resultados por região.
Fonte: autoria própria, 2022.

Região	***Acc.***	***Sens.***	***Spec.***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
Norte	95,17%	86,36%	98,09%	91,50%	88,53%	91,89%	204,21
Nordeste	81,13%	53,97%	98,51%	94,60%	68,01%	72,47%	210,46
Centro-Oeste	83,07%	56,50%	98,16%	94,94%	70,44%	74,21%	38,35
Sudeste	90,37%	64,30%	97,50%	88,63%	74,27%	78,99%	1.654,23
Sul	84,23%	64,36%	98,05%	94,84%	75,98%	79,09%	74,44

As visualizações gráficas dos resultados por região estão expressas nas figuras 4.15, 4.16, 4.17, 4.18, 4.19, 4.20 e 4.21.

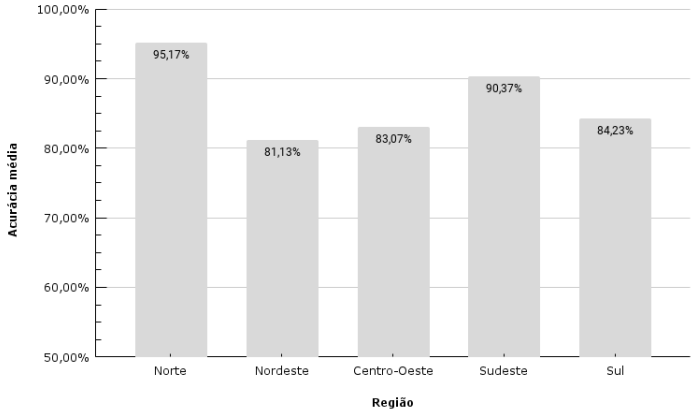


FIGURA 4.15 – Acurácia média por região.
Fonte: autoria própria, 2022.

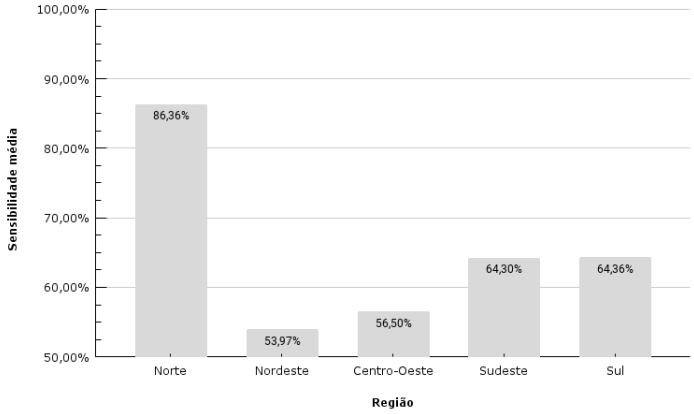


FIGURA 4.16 – Sensibilidade média por região.
Fonte: autoria própria, 2022.

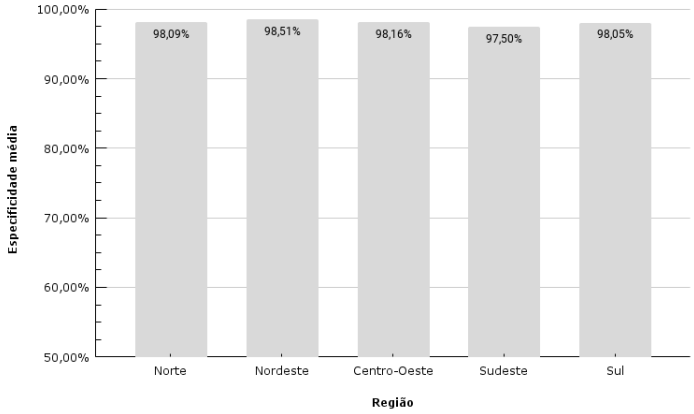


FIGURA 4.17 – Especificidade média por região.
Fonte: autoria própria, 2022.

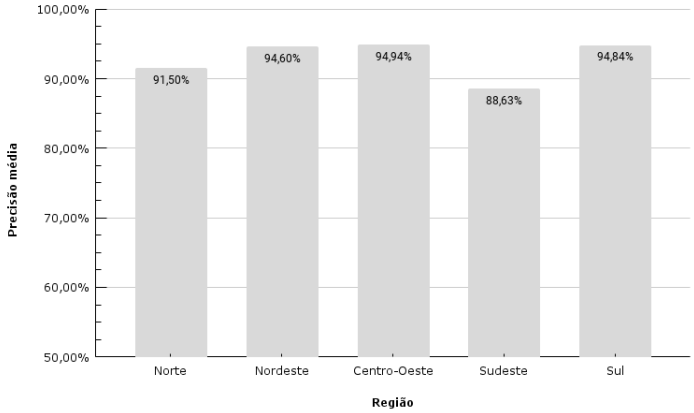


FIGURA 4.18 – Precisão média por região.
Fonte: autoria própria, 2022.

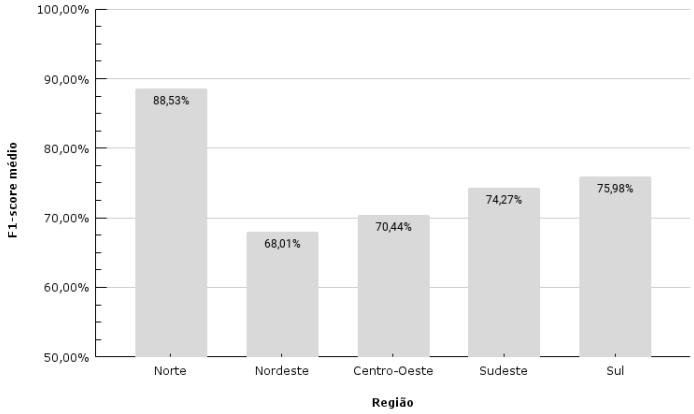


FIGURA 4.19 – *F1-score* médio por região.
Fonte: autoria própria, 2022.

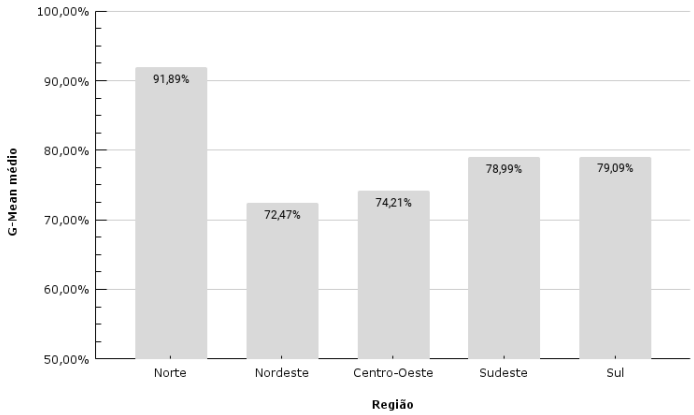


FIGURA 4.20 – *G-Mean* médio por região.
Fonte: autoria própria, 2022.

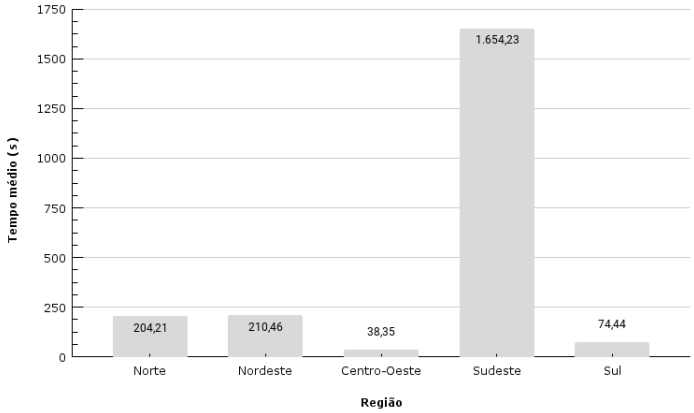


FIGURA 4.21 – Tempo médio por região.
Fonte: autoria própria, 2022.

Percebe-se que os gráficos das médias da acurácia, sensibilidade (em menor grau), *F1-score* e *G-Mean* seguem a mesma tendência, o mesmo ocorrendo para os gráficos das médias da especificidade e precisão. Era esperado o decréscimo do tempo de treinamento com a utilização de um conjunto de treino menor, sendo esta a justificativa para o menor tempo médio nos dados da região Centro-Oeste, visto que esta região possui menos observações. Portanto, nesta abordagem de sobreamostragem, a previsão para a região Norte possui o melhor desempenho dentre as métricas de acurácia, sensibilidade, *F1-score* e *G-Mean*, além de um tempo de execução pequeno se comparado ao tempo médio de treinamento para os dados da região Sudeste, cuja performance preditiva foi a segunda melhor.

Por outro lado, calculando as médias das métricas de previsão por razão de balanceamento, temos os resultados apresentados na tabela 4.22.

TABELA 4.22 – Médias dos resultados por razão de balanceamento.
Fonte: autoria própria, 2022.

Razão	***Accuracy***	***Sensibility***	***Specificity***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
1:1	77,10%	57,86%	97,44%	95,23%	71,11%	74,65%	230,84
1:2	89,45%	74,71%	98,05%	94,28%	82,57%	85,22%	170,90
1:3	90,40%	68,29%	98,16%	91,45%	77,74%	81,68%	280,33
1:4	90,22%	59,54%	98,60%	90,65%	70,36%	75,76%	1.063,28

As visualizações gráficas dos resultados por razão de balanceamento estão expressas nas figuras 4.22, 4.23, 4.24, 4.25, 4.26, 4.27 e 4.28.

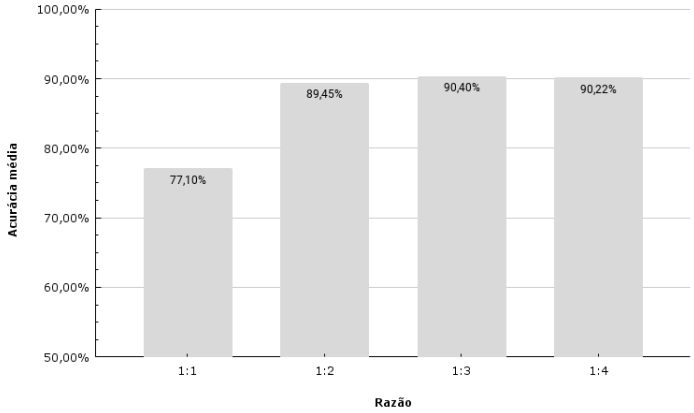


FIGURA 4.22 – Acurácia média por razão de balanceamento.
Fonte: autoria própria, 2022.

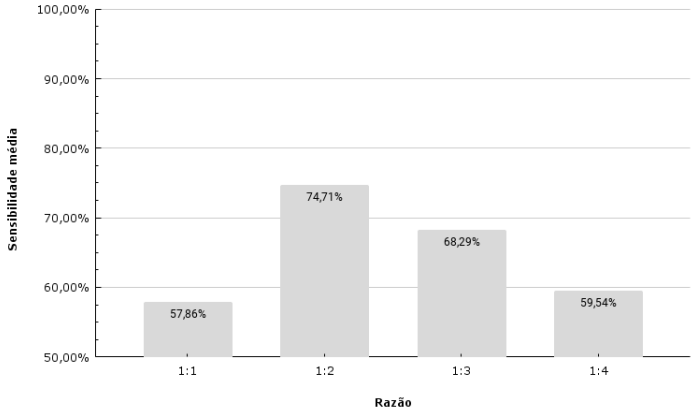


FIGURA 4.23 – Sensibilidade média por razão de balanceamento.
Fonte: autoria própria, 2022.

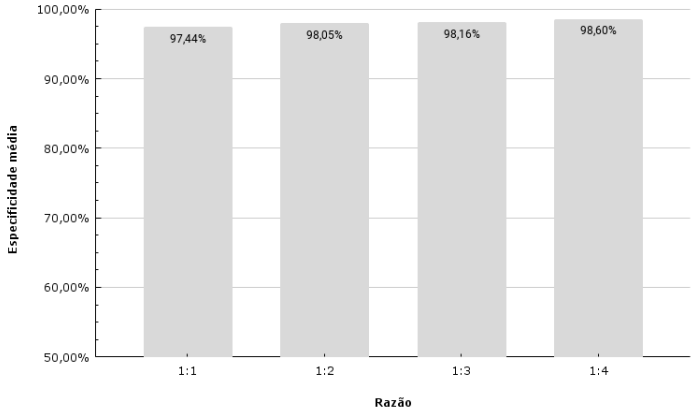


FIGURA 4.24 – Especificidade média por razão de balanceamento.
Fonte: autoria própria, 2022.

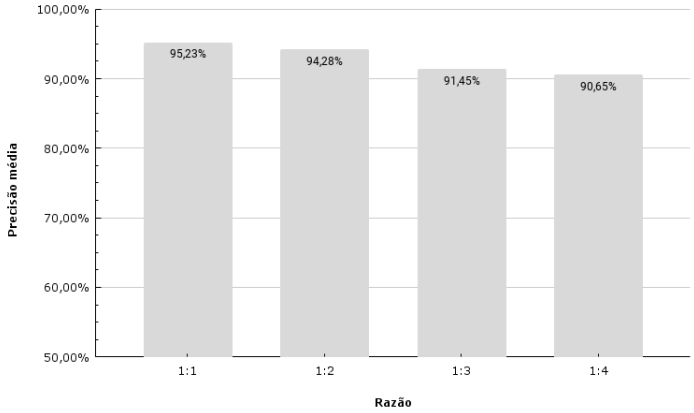


FIGURA 4.25 – Precisão média por razão de balanceamento.
Fonte: autoria própria, 2022.

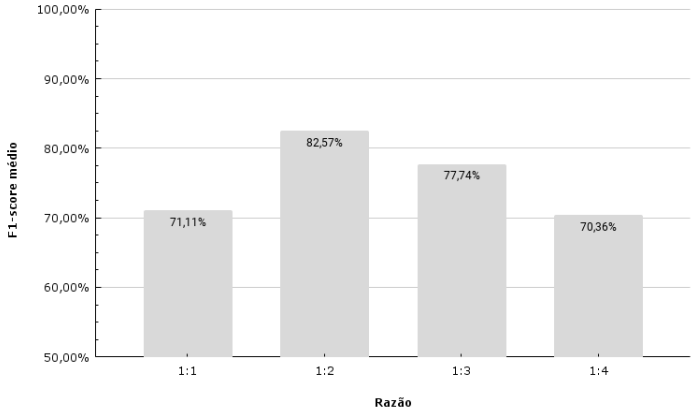


FIGURA 4.26 – *F1-score* médio por razão de balanceamento.
Fonte: autoria própria, 2022.

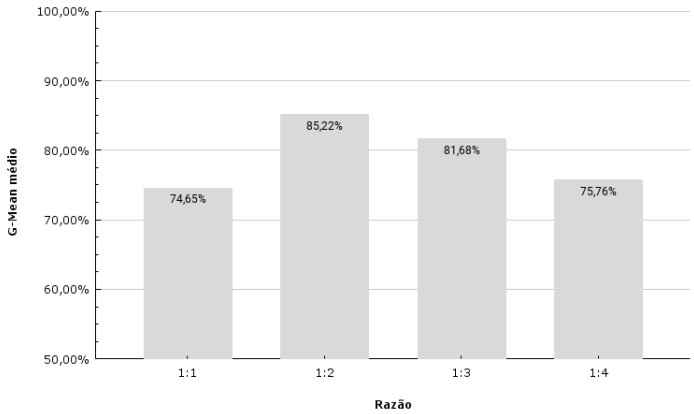


FIGURA 4.27 – *G-Mean* médio por razão de balanceamento.
Fonte: autoria própria, 2022.

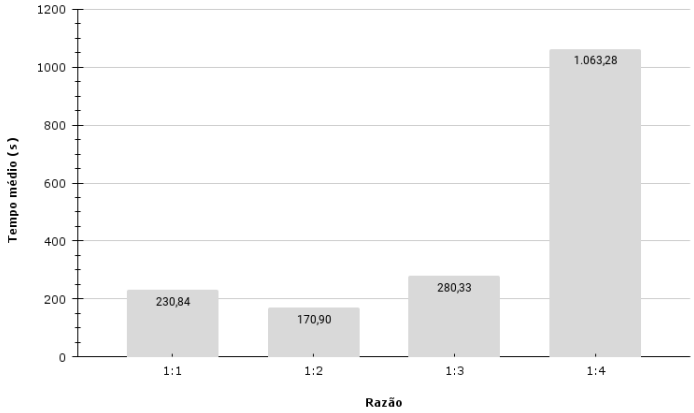


FIGURA 4.28 – Tempo médio por razão de balanceamento.
Fonte: autoria própria, 2022.

Observa-se os valores médios da especificidade e da precisão variando pouco para as proporções de balanceamento analisadas. Por outro lado, percebe-se que os menores valores médios da acurácia, sensibilidade e *G-Mean* ocorrem na proporção de balanceamento 1:1.

Mediante o exposto, não utilizaremos o conjunto com proporção 1:1 para a validação do modelo. Portanto, iremos aplicar o métodos de validação *K-fold cross validation* nos conjunto de dados da região Norte com razões de balanceamento 1:2, 1:3 e 1:4. Para as razões de balanceamento 1:2 e 1:4 aplicamos o método *8-fold cross validation*, enquanto para a proporção 1:3 foi aplicado o *6-fold cross validation*, cujos resultados de acurácia e tempo de treinamento estão apresentados nas tabelas 4.23, 4.24 e 4.25.

TABELA 4.23 – *8-fold cross validation* para o conjunto na razão de balanceamento 1:2.
Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	99,24%	91,13
2	99,20%	170,98
3	99,46%	154,67
4	99,04%	202,01
5	99,22%	180,02
6	99,44%	91,69
7	98,64%	115,46
8	99,22%	171,07
Acurácia média		**99,18%**
Desvio padrão da acurácia		**0,26%**
Tempo médio (s)		**147,13**

TABELA 4.24 – *6-fold cross validation* para o conjunto na razão de balanceamento 1:3. Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	99,25%	113,72
2	99,35%	80,08
3	98,98%	107,39
4	99,07%	120,58
5	98,63%	73,78
6	99,18%	182,82
Acurácia média		**99,08%**
Desvio padrão da acurácia		**0,25%**
Tempo médio (s)		**113,06**

TABELA 4.25 – *8-fold cross validation* para o conjunto na razão de balanceamento 1:4. Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	98,94%	110,66
2	97,55%	305,52
3	97,77%	481,26
4	99,48%	224,83
5	99,04%	157,65
6	97,21%	247,26
7	99,30%	339,27
8	98,90%	214,63
Acurácia média		**98,52%**
Desvio padrão da acurácia		**0,88%**
Tempo médio (s)		**260,14**

Portanto, com base nos resultados da validação do modelo aplicando a estratégia de sobreamostragem nos dados da região Norte, percebe-se a modelagem com dados na proporção de balanceamento 1:3 desempenhando melhor, visto que possui o menor tempo médio de treinamento da rede neural dentre os modelos analisados e uma acurácia aproximadamente igual à maior acurácia média obtida nesta validação.

4.3 RUS+ROS por região e por ano

Segmentando novamente a base por regiões, aplica-se a combinação dos métodos de *Random Undersampling* e *Random Oversampling* para obter um conjunto equilibrado na

proporção 1:1 e realizar a predição de rotas para os anos de 2012 a 2018. Conforme observado anteriormente, em 2012 ocorreu a fusão da Azul com a Trip Airlines, resultando em números atípicos para a entrada em rotas e consequentemente prejudicando a modelagem. Para os anos anteriores a 2012, não foi possível realizar a predição. Os resultados obtidos estão apresentados nas tabelas 4.26, 4.27, 4.28, 4.29 e 4.30.

TABELA 4.26 – Resultados para a segmentação da região Norte.
Fonte: autoria própria, 2022.

Ano	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
2018	98,31%	100,00%	97,34%	95,59%	97,75%	98,66%	81,54
2017	89,47%	100,00%	88,69%	39,55%	56,68%	94,18%	110,87
2016	82,93%	69,20%	96,69%	95,45%	80,23%	81,80%	88,19
2015	88,73%	80,38%	95,24%	92,95%	86,21%	87,50%	57,98
2014	94,23%	100,00%	92,55%	79,69%	88,70%	96,20%	37,19
2013	85,78%	77,71%	97,03%	97,34%	86,42%	86,84%	15,98
2012	18,43%	6,95%	99,95%	99,90%	13,00%	26,36%	2,11

TABELA 4.27 – Resultados para a segmentação da região Nordeste.
Fonte: autoria própria, 2022.

Ano	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
2018	65,88%	39,74%	99,18%	98,41%	56,62%	62,78%	163,78
2017	90,35%	80,73%	97,46%	95,92%	87,67%	88,70%	79,06
2016	59,71%	34,22%	99,18%	98,47%	50,79%	58,26%	246,79
2015	82,46%	59,36%	97,92%	95,03%	73,07%	76,24%	76,58
2014	64,42%	41,69%	98,64%	97,89%	58,48%	64,13%	28,58
2013	70,97%	50,63%	98,51%	97,87%	66,74%	70,62%	8,38
2012	53,26%	33,11%	99,72%	99,64%	49,70%	57,46%	2,57

TABELA 4.28 – Resultados para a segmentação da região Centro-Oeste.
Fonte: autoria própria, 2022.

Ano	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
2018	87,99%	81,07%	97,17%	97,44%	88,51%	88,76%	4,07
2017	80,65%	70,72%	98,71%	99,00%	82,50%	83,55%	2,85
2016	99,56%	100,00%	99,12%	99,12%	99,56%	99,56%	2,53
2015	76,48%	59,77%	99,16%	98,98%	74,53%	76,99%	2,46
2014	Não foi possível prever						
2013	Não foi possível prever						
2012	Não foi possível prever						

TABELA 4.29 – Resultados para a segmentação da região Sudeste.
Fonte: autoria própria, 2022.

Ano	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
2018	88,96%	74,83%	97,89%	95,73%	84,00%	85,59%	278,77
2017	90,90%	85,84%	94,67%	92,30%	88,95%	90,15%	195,94
2016	90,15%	81,82%	97,35%	96,40%	88,51%	89,25%	155,61
2015	80,02%	63,61%	98,22%	97,54%	77,01%	79,05%	173,72
2014	92,44%	85,50%	97,01%	94,97%	89,99%	91,08%	58,45
2013	82,71%	68,32%	98,63%	98,22%	80,58%	82,09%	61,33
2012	23,37%	8,28%	99,56%	98,96%	15,28%	28,71%	6,90

TABELA 4.30 – Resultados para a segmentação da região Sul.
Fonte: autoria própria, 2022.

Ano	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
2018	72,39%	53,91%	96,80%	95,70%	68,97%	72,24%	35,15
2017	59,57%	30,80%	98,49%	96,51%	46,70%	55,08%	28,16
2016	54,64%	23,58%	95,94%	88,52%	37,24%	47,56%	45,82
2015	64,19%	41,39%	97,41%	95,88%	57,82%	63,50%	11,25
2014	91,10%	100,00%	89,28%	65,60%	79,23%	94,49%	7,70
2013	91,65%	85,50%	96,71%	95,53%	90,23%	90,93%	1,82
2012	25,32%	5,05%	100,00%	100,00%	9,61%	22,47%	1,06

Calculando as médias das métricas para cada ano usado como conjunto de teste, temos os resultados apresentados na tabela 4.31.

TABELA 4.31 – Médias dos resultados por ano.
Fonte: autoria própria, 2022.

Ano	*Accuracy*	*Sensibility*	*Specificity*	*Precision*	*F1-score*	*G-Mean*	Tempo (s)
2018	82,71%	69,91%	97,68%	96,57%	79,17%	81,61%	112,66
2017	82,19%	73,62%	95,60%	84,66%	72,50%	82,33%	83,38
2016	77,40%	61,76%	97,65%	95,59%	71,27%	75,28%	107,79
2015	78,38%	60,90%	97,59%	96,08%	73,73%	76,65%	64,40
2014	85,55%	81,80%	94,37%	84,54%	79,10%	86,47%	32,98
2013	82,78%	70,54%	97,72%	97,24%	80,99%	82,62%	21,88
2012	30,09%	13,35%	99,81%	99,62%	21,90%	33,75%	3,16

As variações temporais destes resultados acima estão expressas nas figuras 4.29, 4.30, 4.31, 4.32, 4.33, 4.34 e 4.35.

FIGURA 4.29 – Acurácia média por ano de teste.
Fonte: autoria própria, 2022.

FIGURA 4.30 – Sensibilidade média por ano de teste.
Fonte: autoria própria, 2022.

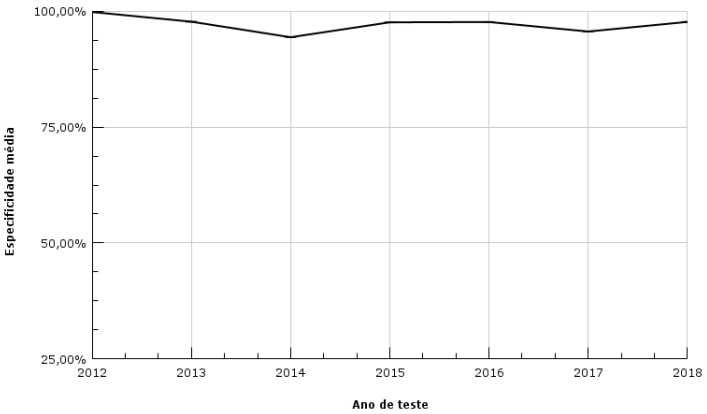


FIGURA 4.31 – Especificidade média por ano de teste.
Fonte: autoria própria, 2022.

FIGURA 4.32 – Precisão média por ano de teste.
Fonte: autoria própria, 2022.

FIGURA 4.33 – *F1-score* médio por ano de teste.
Fonte: autoria própria, 2022.

FIGURA 4.34 – *G-Mean* médio por ano de teste.
Fonte: autoria própria, 2022.

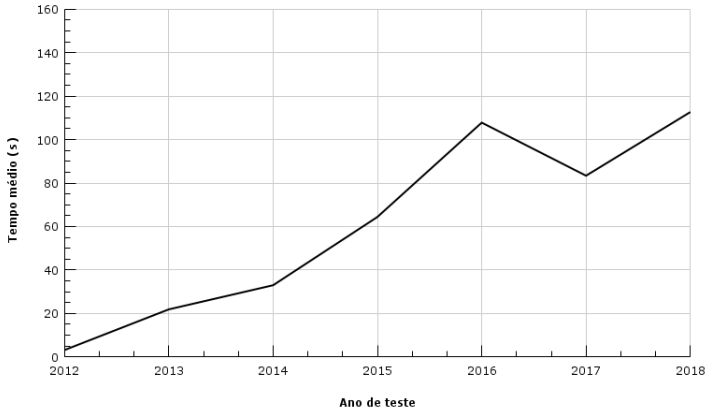


FIGURA 4.35 – Tempo médio por ano de teste.
Fonte: autoria própria, 2022.

Percebe-se que os gráficos das médias da acurácia, sensibilidade, *F1-score* e *G-Mean* seguem a mesma tendência, o mesmo ocorrendo para os gráficos das médias da especificidade e precisão. Era esperado o decréscimo do tempo com a utilização de um conjunto de treinamento menor (quanto menor o ano do conjunto de teste, menor o tamanho do conjunto de treinamento, pois treina-se o modelo com os anos anteriores ao de teste). Novamente observa-se para o ano de 2012 o pior desempenho na modelagem preditiva. Além disso, o modelo utilizando o ano de teste de 2014, além de possuir tempo médio de treinamento abaixo da média, teve o melhor desempenho nas métricas de acurácia, sensibilidade e *G-Mean*, sendo, portanto, o conjunto de teste com os melhores resultados preditivos para a abordagem híbrida.

Por outro lado, calculando as médias das métricas de previsão por região, temos os resultados apresentados na tabela 4.32.

TABELA 4.32 – Médias dos resultados por região.
Fonte: autoria própria, 2022.

Região	***Acc.***	***Sens.***	***Spec.***	***Precision***	***F1-score***	***G-Mean***	**Tempo (s)**
Norte	89,91%	51,86%	98,83%	87,46%	65,11%	71,59%	65,29
Nordeste	72,30%	74,01%	98,04%	92,56%	82,25%	85,18%	100,53
Centro-Oeste	89,40%	83,93%	98,33%	98,52%	90,19%	90,62%	2,98
Sudeste	87,53%	76,65%	97,30%	95,86%	84,84%	86,20%	153,97
Sul	72,26%	55,86%	95,77%	89,62%	63,37%	70,63%	21,65

As visualizações gráficas dos resultados por região estão expressas nas figuras 4.36, 4.37, 4.38, 4.39, 4.40, 4.41 e 4.42.

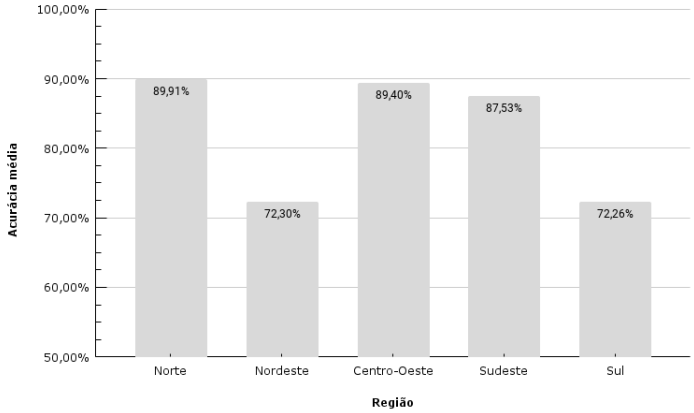


FIGURA 4.36 – Acurácia média por região.
Fonte: autoria própria, 2022.

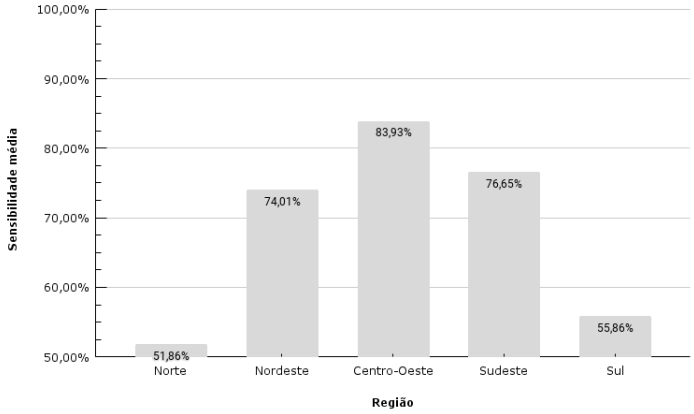


FIGURA 4.37 – Sensibilidade média por região.
Fonte: autoria própria, 2022.

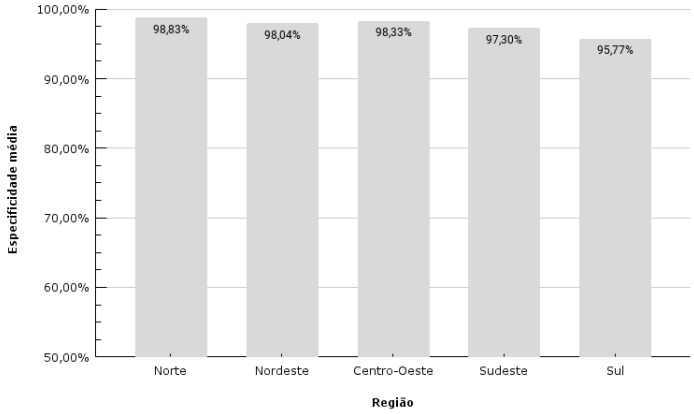


FIGURA 4.38 – Especificidade média por região.
Fonte: autoria própria, 2022.

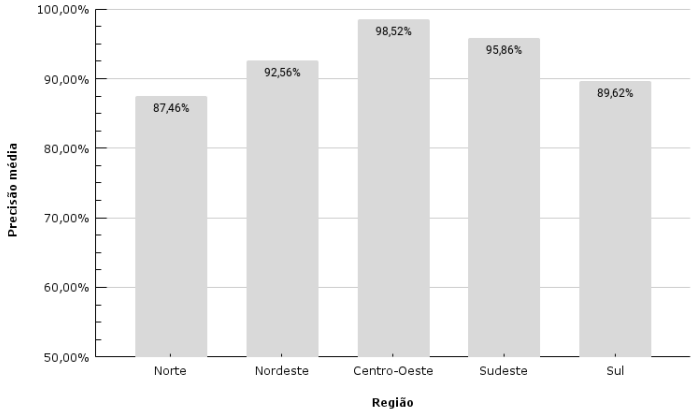


FIGURA 4.39 – Precisão média por região.
Fonte: autoria própria, 2022.

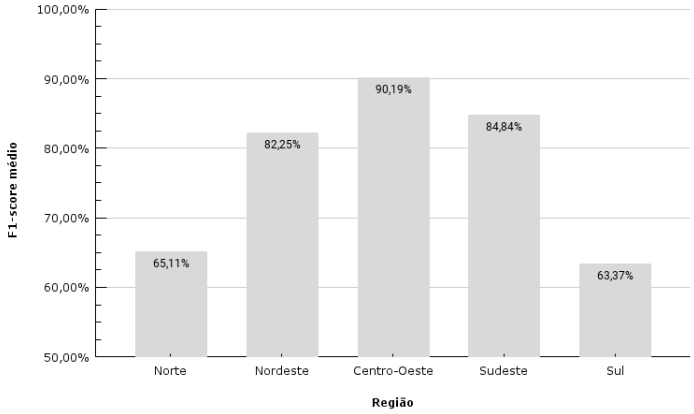


FIGURA 4.40 – *F1-score* médio por região.
Fonte: autoria própria, 2022.

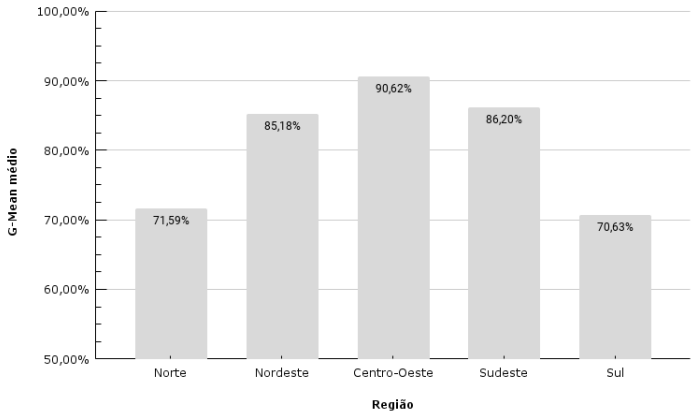


FIGURA 4.41 – *G-Mean* médio por região.
Fonte: autoria própria, 2022.

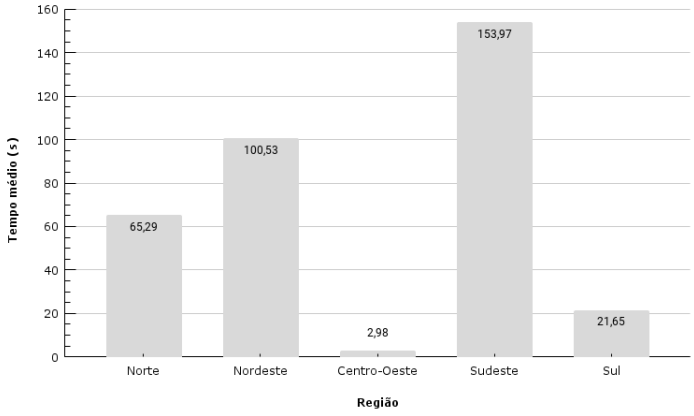


FIGURA 4.42 – Tempo médio por região.
Fonte: autoria própria, 2022.

Observa-se os valores médios da especificidade e da precisão variando pouco para as diferentes regiões. Por outro lado, percebe-se que melhores indicadores de performance foram obtidos para as regiões Norte, Sudeste ou Centro-Oeste, conforme análise dos valores médios da acurácia, sensibilidade, especificidade, precisão, *F1-score* e *G-Mean*, além de ambos possuírem tempos de execução aceitáveis.

Mediante o exposto, não utilizaremos os conjuntos de dados das regiões Nordeste e Sul para a validação do modelo. Portanto, iremos aplicar o métodos de validação *K-fold cross validation* nos dados das regiões Norte, Centro-Oeste e Sudeste. Para as três razões de balanceamento aplicamos o método *6-fold cross validation*, cujos resultados de acurácia e tempo de treinamento estão apresentados nas tabelas 4.33, 4.34 e 4.35.

TABELA 4.33 – *6-fold cross validation* para os dados do Norte.
Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	99,50%	48,88
2	99,72%	61,43
3	99,08%	60,43
4	98,73%	139,05
5	99,40%	126,63
6	99,36%	51,02
Acurácia média		**99,30%**
Desvio padrão da acurácia		**0,35%**
Tempo médio (s)		**81,24**

TABELA 4.34 – *6-fold cross validation* para os dados do Centro-Oeste.
Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	99,68%	3,35
2	99,65%	2,73
3	99,65%	2,26
4	99,62%	5,14
5	99,62%	5,71
6	99,68%	3,00
Acurácia média		**99,65%**
Desvio padrão da acurácia		**0,03%**
Tempo médio (s)		**3,70**

TABELA 4.35 – *6-fold cross validation* para os dados do Sudeste.
Fonte: autoria própria, 2022.

Treinamento	***Accuracy***	***Tempo (s)***
1	98,60%	166,15
2	99,07%	765,48
3	98,75%	253,60
4	95,88%	442,09
5	98,56%	171,01
6	98,43%	138,48
Acurácia média		**98,21%**
Desvio padrão da acurácia		**1,16%**
Tempo médio (s)		**322,80**

Portanto, com base nos resultados da validação do modelo aplicando combinação das estratégias de sub e sobreamostragem por região, percebe-se a modelagem com dados da região Centro-Oeste desempenhando melhor, visto que possui o menor tempo médio de treinamento da rede neural dentre os modelos analisados e a maior das acurácias obtidas nesta validação.

4.4 ROS vs RUS+ROS na predição das rotas de 2018

A análise comparativa dos métodos de *Random Oversampling* e a abordagem associando *Random Undersampling* e *Random Oversampling*, em uma razão de balanceamento 1:1, consistiu em prever as rotas em todo o território nacional para o ano de 2018, o que

exige um conjunto de treinamento com grande quantidade de observações e um tempo de execução consideravelmente mais elevado se comparado às estratégias anteriores. Os resultados obtidos estão apresentados na tabela 4.36.

TABELA 4.36 – Resultados para a previsão completa.
Fonte: autoria própria, 2022.

Método	***Acc.***	***Sens.***	***Spec.***	***Precision***	***F1-score***	***G-Mean***	**Tempo (h)**
ROS	91,25%	83,59%	98,73%	98,47%	90,42%	90,84%	47,24
RUS+ROS	93,97%	89,10%	98,78%	98,63%	93,63%	93,82%	9,67

Portanto, observa-se a abordagem híbrida, combinando sub e sobreamostragem, desempenhando melhor em todas as métricas de performance na predição de rotas de 2018. O tempo de execução de horas ou até dias impediu que mais análises fossem desenvolvidas, em tempo hábil, analisando outros conjuntos de treino e teste.

5 Conclusão

O presente trabalho teve por objetivo desenvolver modelagens baseadas em aprendizado de máquina para realizar a predição de rotas da Azul Linhas Aéreas, cujo desenvolvimento evidenciou a necessidade de implementar estratégias de pré-processamento de dados para lidar com o problema de desequilíbrio de classes. Nesse contexto, após a análise exploratória, comparou-se diferentes abordagens de reamostragem e modelos de aprendizagem.

Dessa forma, após obter as bases balanceadas pela redução da classe majoritária, realizaram-se as predições de rotas para cada ano da base de dados, comparando os resultados obtidos para as diferentes proporções de balanceamento e concluindo que o ano de teste de 2018, além de possuir tempo médio de treinamento aceitável, teve o melhor desempenho nas métricas de sensibilidade, *F1-score* e *G-Mean*, sendo, portanto, o conjunto de teste com os melhores resultados preditivos para a abordagem de subamostragem. Aliado a isso, com base nos resultados da validação do modelo aplicando a estratégia de subamostragem, percebe-se a modelagem com dados na proporção de balanceamento 1:1 desempenhando melhor, visto que possui a maior acurácia dentre os modelos analisados e um tempo médio de treinamento da rede neural consideravelmente menor que os demais.

Em seguida, implementam-se os modelos com as bases balanceadas obtidas pelo aumento da classe minoritária para prever rotas em cada região do Brasil, comparando outra vez os resultados obtidos para as diferentes proporções de balanceamento e concluindo que a região Norte possui o melhor desempenho dentre as métricas de acurácia, sensibilidade, *F1-score* e *G-Mean*, além de um tempo de execução pequeno se comparado ao tempo médio de treinamento para os dados da região Sudeste, cuja performance preditiva foi a segunda melhor. Por outro lado, com base nos resultados da validação do modelo aplicando a estratégia de sobreamostragem nos dados da região Norte, percebe-se a modelagem com dados na proporção de balanceamento 1:3 desempenhando melhor, visto que possui o menor tempo médio de treinamento da rede neural dentre os modelos analisados e uma acurácia aproximadamente igual à maior acurácia média obtida nesta validação.

Além disso, utilizam-se as bases balanceadas pela combinação do aumento da classe

minoritária com a redução da classe majoritária para prever novamente as rotas por região, desta vez comparando os resultados da predição de cada ano e concluindo que o ano de teste de 2014, além de possuir tempo médio de treinamento abaixo da média, teve o melhor desempenho nas métricas de acurácia, sensibilidade e *G-Mean*, sendo, portanto, o conjunto de teste com os melhores resultados preditivos para a abordagem híbrida. Ademais, com base nos resultados da validação do modelo aplicando combinação das estratégias de sub e sobreamostragem por região, percebe-se a modelagem com dados da região Centro-Oeste desempenhando melhor, visto que possui o menor tempo médio de treinamento da rede neural dentre os modelos analisados e a maior das acurácias obtidas nesta validação.

Por fim, realizou-se a análise comparativa dos métodos de *Random Oversampling* e a abordagem associando *Random Undersampling* e *Random Oversampling*, gerando conjuntos com razão de balanceamento 1:1. Os algoritmos foram implementados para realizar a predição das rotas em todo o território nacional, cuja modelagem foi desenvolvida para prever o ano mais recente do conjunto de dados. Observou-se, portanto, a abordagem híbrida, combinando sub e sobreamostragem, desempenhando melhor em todas as métricas de performance na predição de rotas de 2018. O tempo de execução elevado impediu que mais análises fossem desenvolvidas em tempo hábil.

Como possibilidades de trabalhos futuros, sugere-se a implementação de outros algoritmos de seleção de atributos, como por exemplo a análise de *feature importance*. Além disso, implementar outros métodos de pré-processamento em *Algorithmic-level*, ou combiná-los com os atuais, pode melhorar o desempenho preditivo. Comparar outros modelos de aprendizagem também pode ser explorado nesse contexto. Por fim, sugiro a realização da predição de rotas para ano posteriores a 2019, o que não foi possível por falta de dados no momento do desenvolvimento deste trabalho.

Referências

ALEKSEEV, K. P. G.; SEIXAS, J. M. A multivariate neural forecasting modeling for air transport – preprocessed by decomposition: A brazilian application. **Journal of Air Transport Management**, v. 15, n. 5, p. 212–216, 2009.

ANDRADE, C. The p value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. **Indian Journal of Psychological Medicine**, v. 41, n. 3, p. 210–215, 2019.

BACH, M. *et al.* The proposal of undersampling method for learning from imbalanced datasets. **Procedia Computer Science**, v. 159, p. 125–134, 2019.

BOGUSLASKI, C. *et al.* Entry patterns in the southwest airlines route system. **Review of Industrial Organization**, v. 25, p. 317–350, 2004.

BRANCO, P. *et al.* A survey of predictive modelling under imbalanced distributions. **Review of Industrial Organization**, v. 49, n. 2, p. 31–48, 2015.

CARREÑO, A. *et al.* Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. **Artificial Intelligence Review**, v. 53, p. 3575–3594, 2019.

DIXIT, A.; CHINTAGUNTA, P. K. Learning and exit behavior of new entrant discount airlines from city-pair markets. **Journal of Marketing**, v. 71, n. 2, p. 150–168, 2007.

FERNáNDEZ, A. *et al.* **Learning from Imbalanced Data Sets**. [S.l.]: Springer International Publishing, 2018.

FERREIRA, P. *et al.* Exploring feature normalization and temporal information for machine learning based insider threat detection. In: INTERNATIONAL CONFERENCE ON NETWORK AND SERVICE MANAGEMENT, 15TH, 2019. **Proceedings**. [S.l.]: IEEE, 2019. p. 1–7.

GIL-MOLTO, M. J.; PIGA, C. A. Entry and exit by european low-cost and traditional carriers. **Tourism Economics**, v. 14, p. 577–598, 2008.

GOSAIN, A.; SARDANA, S. Handling class imbalance problem using oversampling techniques: A review. In: INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, COMMUNICATIONS AND INFORMATICS, 2017. **Proceedings**. [S.l.]: IEEE, 2017. p. 79–85.

HASANIN, T. *et al.* Examining characteristics of predictive models with imbalanced big data. **Journal of Big Data**, v. 6, n. 69, 2019.

HEATON, J. **Artificial Intelligence for Humans: Deep learning and neural networks**. [S.l.]: Heaton Research, Incorporated., 2015. (Artificial Intelligence for Humans).

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. **Progress in Artificial Intelligence**, v. 5, n. 4, p. 221–232, 2016.

LEEVY, J. L. *et al.* A survey on addressing high-class imbalance in big data. **Journal of Big Data**, v. 5, n. 42, 2018.

MOHAMMED, R. *et al.* Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In: INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION SYSTEMS, 11 TH, 2020. **Proceedings**. [S.l.]: IEEE, 2020. p. 243–248.

MULLER, K. *et al.* The construction of a low-cost airline network – facing competition and exploring new markets. **Managerial and decision economics**, v. 33, p. 485–499, 2012.

MURPHEY, Y. L. *et al.* Neural learning from unbalanced data. **Applied Intelligence volume**, v. 21, n. 2 spec., p. 117–128, 2004.

OLIVEIRA, B. F.; OLIVEIRA, A. V. An empirical analysis of the determinants of network construction for azul airlines. **Journal of Air Transport Management**, v. 101, n. 102207, 2022.

SCHOBER, P. *et al.* Correlation coefficients: Appropriate use and interpretation. **Anesthesia & Analgesia**, v. 126, n. 5, p. 1763–1768, 2018.

SINCLAIR, R. A. An empirical model of entry and exit in airline markets. **Review of Industrial Organization**, v. 10, p. 541–557, 1995.

SRIRATANAWILAI, S.; ERJONGMANEE, S. Route prediction in air travel network using socio-economic factors and learning models. In: INTERNATIONAL CONFERENCE ON BUSINESS AND INDUSTRIAL RESEARCH, 5 TH, 2018. **Proceedings**. [S.l.]: IEEE, 2018. p. 100–105.

SRISAENG, P. *et al.* Forecasting demand for low cost carriers in australia using an artificial neural network approach. **Aviation**, v. 19, n. 2, p. 90–103, 2015.

XIE, G. *et al.* Short-term forecasting of air passenger by using hybrid seasonal decomposition and least squares support vector regression approaches. **Journal of Air Transport Management**, v. 37, p. 20–26, 2014.

YAP, B. W. *et al.* An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. **Lecture Notes in Electrical Engineering**, v. 285, p. 13–22, 2013.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
TC	23 de novembro de 2022	DCTA/ITA/TC-088/2022	75

5. TÍTULO E SUBTÍTULO:

Predição das rotas da Azul Linhas Aéreas usando algoritmos de aprendizado de máquina

6. AUTOR(ES):

Jonathans Schaffer Torres

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

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Transporte aéreo; Predição de rotas; Desequilíbrio de classes; Pré-processamento de dados; Reamostragem; Aprendizado de Máquina; Rede neural.

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Operações de linhas aéreas; Rotas aéreas; Aprendizagem (inteligência artificial); Redes neurais; Planejamento estratégico; Transporte aéreo; Transportes.

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

ITA, São José dos Campos. Curso de Graduação em Engenharia Civil-Aeronáutica. Orientador: Alessandro Vinícius Marques de Oliveira. Publicado em 2022.

11. RESUMO:

As consequências das entradas em novas rotas e o que leva as companhias aéreas a optarem pela adição de novos destinos são aspectos importantes e de extremo interesse da indústria, sendo considerado um dos elementos cruciais no planejamento estratégico das empresas. O presente trabalho tem como objetivo desenvolver modelagens baseadas em aprendizado de máquina para realizar a predição de rotas da Azul Linhas Aéreas, tendo como referência os dados de passageiros e de tráfego no período de 2008 a 2018. A base de dados utilizada possui um desequilíbrio de classes, evidenciando a necessidade da utilização de métodos de pré-processamento de dados, que foram implementados, majoritariamente, através da reamostragem do conjunto. Aplicando as técnicas subamostragem, sobreamostragem e a combinação destas aliadas às estratégias de análise de correlações e significância, seleção de variáveis e validação da modelagem com *K-fold cross validation*, comparam-se os métodos de redes neurais, regressão logística, KNN e árvore de decisão para o modelo preditivo, chegando a um melhor desempenho para as redes neurais artificiais. Como métricas de desempenho, consideram-se a acurácia, sensibilidade, especificidade, precisão, *F1-score*, *G-Mean* e tempo de treinamento do modelo. Para o modelo de redes neurais, realizam-se predições de rotas por ano, por região do território brasileiro e por proporção de balanceamento do conjunto após a reamostragem. Compara-se ainda o método de sobreamostragem e a abordagem híbrida na previsão de rotas de 2018 em todo o território nacional, concluindo com um melhor desempenho da abordagem híbrida em todas as métricas.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**