PREVISÃO DA OCORRÊNCIA DE ATRASOS EM DECOLAGENS DO AEROPORTO DE GUARULHOS

Hélio da Silva Queiroz Júnior*[1], Viviane Adriano Falcão[1], Francisco Gildemir Ferreira da Silva[2]
[1]Universidade Federal de Pernambuco, [2]Universidade Federal do Ceará

*** Corresponding author e-mail address**: helio.junior@ufpe.br

PAPER ID: SIT2312

ABSTRACT

Flight delays in commercial aviation have become a common problem worldwide due to increasing air transportation demands, resulting in higher costs and continuous adjustments in flight management. Efficiently predicting these delays is a recurring research topic in air traffic operations. Machine learning has emerged as a robust alternative to classical statistical methods, particularly Artificial Neural Networks (ANNs), known for their modular adaptability. However, varying study scenarios yield different accuracy outcomes. This study aims to define an appropriate prediction and classification scenario for regular departures at an airport, considering influencing variables, acceptable accuracy levels, and the most effective prediction model. It proposes a class-based predictive system applicable to São Paulo International Airport (SBGR) and the national airline Gol Linhas Aéreas (GLO). Bibliometric review and meta-analysis were employed to establish a baseline reference for method accuracy evaluation. Furthermore, studies were categorized using Data Envelopment Analysis by metafrontier to identify ideal prediction scenarios. The analyses reveal that the MultiLayer Perceptron (MLP) Neural Network demonstrates better predictive efficacy for route or airline-specific analyses, irrespective of delay reasons. A comparative case study was conducted for the defined prediction scenario, confirming that classifying delay time yielded the highest accuracy regardless of delay reasons. It was found that the specificity of scenarios and delay reasons within the chosen analysis area had a stronger relationship with predictive accuracy than the volume of data obtained.

Keywords: Flight delays, Artificial Neural Networks, Prediction, Classification, Air transportation.

ACKNOWLEGDEMENTS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

1. INTRODUCTION

According to EUROCONTROL's annual report in 2019, 24.2% of flights in Europe experience delays. In the United States, the cost per minute of delay for American airlines was estimated to be $68.48 in 2017 (A4A, 2017).

In Brazil, approximately half of all flights were delayed in 2019, as reported by the National Civil Aviation Agency (ANAC, 2019). ANAC provides data on flight delays and cancellations, showing that Brazil has an average annual delay rate of 6.97% for delays exceeding thirty minutes and 2.49% for delays exceeding one hour across approximately fifty-four scheduled flight legs. This corresponds to around 10 out of every 100 flights experiencing delays over thirty minutes (ANAC, 2021). It is estimated that each flight with a delay of over thirty minutes incurs a loss of approximately U$2,05 for Airlines.

São Paulo International Airport (SBGR), which has the highest passenger and cargo traffic in Brazil, experiences an average monthly delay rate of 14% for both arrivals and departures (ANAC, 2022). The availability of new and accessible technologies has facilitated data acquisition, enabling the application of aviation studies in critical areas such as big data analysis, air transport network management, and machine learning-based flight delay prediction (CHU et al., 2020).

To address these challenges, this study focuses on proposing a classification-based predictive system tailored to SBGR and the national airline Gol Linhas Aéreas (GLO). The objective is to implement a predictive system that accurately forecasts delays and aligns with the operational reality of the airport and airline, allowing for better planning and decision-making.

2. MAIN CAUSES OF FLIGHT DELAYS

Understanding the current air market situation is crucial to characterize the cause of delays at an airport, including capacity in relation to existing traffic, expected demand growth, and potential consequences such as increased flight delays. Chen et al. (2021) emphasized the incorporation of specific airport functionality information and other key factors like queue and en-route delays, once available.

Schultz & Reitmann (2019) highlighted ground operations as a significant contributor to delays, as uncertainties during the airborne phase had a lower impact compared to ground operations deviations. Flight operations planning favors certain delay conditions depending on the flight type. International flights, with their additional requirements and potential complexities in-ground and en-route operations, pose particular challenges (KHAN et al., 2021). Additionally, weather and seasonal conditions can disrupt scheduled departures at airports.

For Chinese airports, Chen et al. (2021) recommended considering either the current weather or then weather conditions in the preceding two hours. Holidays and peak demand periods should also be considered due to observed variations in aircraft and passenger movements (CHEN et al., 2021). Beijing International Airport (ZZBA) has conducted delay studies utilizing real-time data collection and macro variables, including weather, seasonal effects, delay propagation, and air traffic control (Yu et al., 2019).

Prediction studies highlight the need for diverse information to account for the various classifiable reasons for flight delays. To address this, the International Air Transport Association (IATA) has categorized the main causes of aircraft delays into eight major categories, facilitating communication of delay reasons and related messages (IATA, 2016).

For Brazil, ANAC (National Civil Aviation Agency) uses the same classification for delays as described by IATA, with codes assigned to each justification identified after the occurrence of the delay. Then, these codes are distributed among reasons related to passengers and baggage, technical and aeronautical equipment, or even weather conditions (Table 1).

Table 1: Codes to be used in aircraft movement and deviation messages. Source: Adapted from IATA (2016), ANAC (2000).

IATA Group	Justifications	IATA/ANAC Codes	Source
1. Passenger and Baggage (Cargo and Mail)	Airport facilities - support restrictions	AF	IATA AHM-730
	Migration/customs/health	AG	IATA AHM-730
	Security/pax/cargo/alarm	AS	IATA AHM-730
	Release of air traffic services/anticipation	AT	IATA AHM-730
2. Aircraft Handling and Ramps	Fueling/defueling	GF	IATA AHM-730
3. Technical and Aeronautical Equipment	Aircraft replacement	TC	IATA AHM-730
	Aircraft defects	TD	IATA AHM-730
4. Aircraft Damage or Automated Equipment	Failure Damage during flight operations	DF	IATA AHM-730
	Damage during ground operations	DG	IATA AHM-730
	Failure of automotive and passenger service equipment	MA	AIC1504 A
5. Flight Operations and Crew	Authorized	AO	IATA AHM-730
	Flight plan - approval	FP	IATA AHM-730
6. Weather	Alternative below limits	WA	AIC1504 A
	Deicing and removal of snow and/or mud on aircraft	WI	IATA AHM-730
	Origin airport below limits	WO	IATA AHM-730
	Delay due to return - weather conditions	WR	IATA AHM-730
	Removal of ice/water/mud/sand at airport	WS	IATA AHM-730
	Delay at alternative airport - weather conditions	AM	IATA AHM-730
	Destination airport below limits	WT	IATA AHM-730
7. Air Traffic Flow Restriction and Government Authorities	Origin airport closed	AI	AIC1504 A
	Destination airport closed	AJ	AIC1504 A
	Airport with operational restrictions	AR	AIC1504 A
8. Reactive and Miscellaneous	Non-specific delays, others	MX	IATA AHM-730
	Aircraft connection	RA	IATA AHM-730
	Aircraft connection/return - outbound flight not penalized airport closed	RI	AIC1504 A
	Aircraft connection/return - outbound flight not penalized weather conditions	RM	AIC1504 A

2.1. Predictability and Flight Information

According to Khan et al. (2021), flight plans are meticulously prepared before each flight, detailing various operational parameters essential for the aircraft's smooth operation. For international flights, airlines develop the flight plan four hours in advance. These plans enable controllers to monitor the projected flight stages and assess the deviations in estimated time across different functional departments, including weather conditions, air traffic flow, aircraft performance, and other factors crucial for determining the optimal flight trajectory and ensuring safe operations.

In addition to the intricate nature of the required procedures, the overall air demand at analyzed airports significantly impacts flight punctuality. Yu et al. (2019) highlighted one of the challenges in delay analysis, which is the vast amount of information collected from flight plans at both macro and micro levels. This abundance of data can impede the system's ability to predict delays accurately or introduce collinearity issues among redundant factors. Consequently, the lack of generalizability presents a second challenge, as a large dataset often contains redundant information.

It is crucial to understand the current situation of the analyzed area, such as airports, airlines, or routes, in terms of existing capacity and expected demand growth to characterize the effects of delays at an airport. Apart from the previously mentioned factors related to airflow and control, more specific and time-specific factors like queue delays, en-route delays, holidays, and peak demand periods are considered whenever data is available (CHEN et al., 2021).

In light of these considerations, Khan et al. (2021) emphasized the need for developing a predictive system that accurately anticipates delays, surpassing the mere indication of their occurrence or ambiguities in categorization. This viewpoint aligns with Truong's (2021) perspective, which underscores the importance of analyzing the relationship between selected input variables for delay prediction systems and their current occurrence. Many current studies focus on employing various machine learning methods or their combinations to predict delays, but they often lack in-depth exploration of the data or consideration of real-world applicability of the models in establishing a correlation between input variables and delays.

Thus, the complexity of defining flight delay prediction models entails selecting influential variables, possessing prior knowledge of the required data, choosing the appropriate prediction model, and ensuring a significant accuracy margin in correlating causes and effects.

2.2. Variables that influence flight delays

Several studies were reviewed to identify commonly applied scenarios in predictions, allowing a comparison between methods, their accuracy, and the variables used.

Each scenario, set of variables, and selected models aimed to identify and describe distinct reasons for delays, whether due to crew and cabin boarding (WEI & VAZEA (2018), SCHULTZ & REITMANN (2019)) severe or unexpected weather impacts (CHEN & WANG (2019), LIU et al. (2019) and OLIVEIRA et al. (2021)) related to cargo deliveries and boarding (LIU et al. (2019)b); due to airline misconduct (ARORA & MATHUR (2020)); capacity issues on congested airport days (SCARPEL & PELICIONI (2018)); or even models that encompass the universality of reasons (LAMBELHO et al. (2020), KHAN et al. (2021), YU et al. (2019), CHEN & LIN (2021), TRUONG (2021)).

In the analyzed works, machine learning models stood out as the majority compared to classical forecasting methods, such as logit regression, panel regression, and multinomial logistics used by Oliveira et al. (2021), Chen & Lin (2021), and Arora & Mathur (2020), respectively.

This situation aligns with the findings presented by Karlaftis & Vlahogianni (2011) when comparing statistical data analysis and the use of artificial intelligence, specifically neural networks, for transportation research. They emphasized that, despite the similarities, when the emphasis of information processing is on obtaining accurate predictions, machine learning is suggested as the most suitable method.

This principle is consistent with the findings of Chen et al. (2021), who defined linear regression as an effective and applicable forecasting method due to the smaller size of the database used.

3. MACHINE LEARNING METHODS

Simon (1983) defined machine learning as improving a system's performance when it repeats the same task or another task from the same population. Regarding the prediction of flight delays, supervised data generalization models are commonly used due to the complexity involved in analyzing unsupervised learning responses. Khan et al. (2021) highlighted two commonly applied estimation mechanisms in supervised learning: regression and classification. Regression predicts flight delay as a continuous output, such as expected delay time or delay groups, while classification categorizes flight delays into binary form (1 = delayed flight, 0 = on-time flight).

Truong (2021) emphasized the necessity of a supervised causal network structure for prediction purposes. When analyzing risk prediction as the target variable, the network structure can be used to predict the risk (probability of an incident) based on other variables (predictors) in the network. If the risk is a discrete variable, the prediction model becomes a classification model, and the target variable is referred to as a classifier.

Considering the uncertainty inherent in-flight delay data, Khan et al. (2021) argued against assuming normality in regression and class balance in classification. Pre-processing and transformation techniques aimed at improving the distribution of highly skewed and scattered data offer limited benefits for regression models. However, the application of various sampling techniques improves classifier performance by balancing the classes and smoothing decision boundaries.

Therefore, when conducting analyses involving diverse categories, it is essential to implement data sampling adjustment mechanisms. Among the sampling techniques addressing class imbalance and overlapping decision boundaries, Khan et al. (2021) identified the Synthetic Minority Over-sampling Technique (SMOTE) as the most effective approach for flight delay prediction studies. This technique oversamples the minority class. Subsequently, it removes the noisy majority class, borderline majority, and minority classes using under-sampling and data cleaning methods like TOMEK, resulting in smoother decision boundaries.

As a result, the most effective method for flight delay prediction using machine learning should achieve high accuracy values defined by the F1-Score and demonstrate robust performance when exposed to sampling techniques for large datasets. Such a method's capability would enable its practical application in real-world scenarios, benefiting air traffic control and management in airports or airlines.

3.1. Artificial Neural Networks (ANNs) applied in flight delay prediction studies

Artificial Neural Networks (ANNs), in their various form, have been gaining prominence in more complex data analyses among the set of artificial intelligence methods classified as machine learning. ANNs are defined as parallel and distributed systems composed of interconnected simple processing units called artificial neurons. They can store information, recognize and classify patterns, and make predictions (MARCY, 2017). As characteristics of supervised prediction models, the generalization ability of a neural network comes from training. During training, known input values and their corresponding expected outputs are fed into the network to calibrate the network and the weights assigned to each interaction, aiming to obtain responses that are closest to the initially known value θ.

However, searching for an ideal set of weights, which can make an ANN generalizable, involves distinct processes. One of these processes is called the backpropagation algorithm, which, according to Haykin (2001), determines the network's learning behavior. Back propagation aims to reduce the error between the threshold and the predicted value in two stages during training. Some studies implement ANNs with the multilayer perceptron configuration using backpropagation, like Khan et al. (2021), who predicted flight delays and departure duration, achieving an average prediction accuracy of 65.5%, 61.5%, and 59% for classifying delay status and predicting delay duration within 60 min and 30 min limits, respectively.

The system proposed by Yu et al. (2019) using deep learning neural networks, which are robust learning models due to their ability to handle large datasets, demonstrated high capability in dealing with the challenges posed by large datasets and can capture key factors influencing delays. New influential factors are identified, in addition to the commonly employed ones such as previous flight delays, number of passengers per flight, air traffic control, and so forth, including the route situation and airport congestion level (based on the number of flights and passenger flows).

Another study utilizing ANNs was conducted by Truong (2021), who proposes the use of the USELEI neural network methodology, corresponding to the steps of understanding (U), sampling (S), exploration (E), learning (L), evaluation (E), and inference (I), to predict the risk of delay and non-delay cases in the U.S. National Airspace System (NAS). The study achieved positive and negative predictive values of 91.56% and 95.45%, respectively.

Therefore, applying ANNs to predict the occurrence of delays is a reality in various scenarios. The considerable volume of data and the multitude of possible reasons for delay classification require models with higher computational capacity and generalization of responses.

2. METHODOLOGY

A case study was conducted to compare the efficiency of flight delay prediction in the operation of a Brazilian airport and an airline based on a proposed prediction scenario. This comparison aims to indicate the predictive capacity of the identified scenario, serving as a baseline parameter for an approximation of the current analysis of traffic and control used in Brazil.

The model accuracy derives from the machine learning method known as Artificial Neural Networks (ANNs), with the obtained responses evaluated for the two specified analysis areas (universal delay prediction for an airport and an airline). In both situations, the output of the ANN was the categorization of the occurrence of a possible delay, including the type of delay or the expected delay time interval for the flight.

Thus, the methodology applied in the study is divided into six execution stages, in addition to an initial stage of delimiting the analyzed areas, described as follows.

3.1.1. Preliminary Stage (Stage 0): Definition of the analysis áreas.

The dimensions of the analyzed areas in the case study indicate the predictive capacity of the model applied to the airport and the airline, distinguishing the level of competence of the operational flight logic applied by each entity in the occurrence of delays.

Despite the movements and the size presented for Guarulhos-SP Airport, in Chapter 5, the justification for applying the scenario definition responses in a Brazilian case leads to the verification of the most prominent national airlines operating at SBGR and the annual delay rate. Among the airlines operating at SBGR in 2019, three national airlines are listed with the highest number of passengers and aircraft movements: Latam (TAM), Gol (GLO), and Azul (AZU).

Thus, analyses of delay occurrences were conducted for these three national airlines operating at SBGR (TAM, GLO, and AZU), selecting the company with the highest occuring delays exceeding 15 minutes. The collected data are from regular departures at SBGR for domestic and international aviation, regardless of whether they were passenger or cargo flights in the year 2019.

1st Stage: Definition of the prediction logic

According to the conducted theoretical review, the multilayer perceptron method known as Artificial Neural Networks (ANNs) with the backpropagation algorithm was chosen for the prediction task.

The types of classes assigned to delays will be two-fold, with the desired output of the network being the classification of possible flight delays at SBGR. The first logic employed will classify delays based on the IATA/ANAC groups of possible causes for these delays.

Thus, the predicted output of the ANN for the first logic corresponds to an integer corresponding to the numbers assigned to each group. However, it was necessary to include a tenth group with an output of 10, representing non-delayed flights, to account for these cases when flights depart on the previously scheduled time.

However, although a regression analysis by group pf delays is applied to a forecast situation for only one airport and one airline, the greater range of cases may result in a more comlex definition of delay due to different causes.

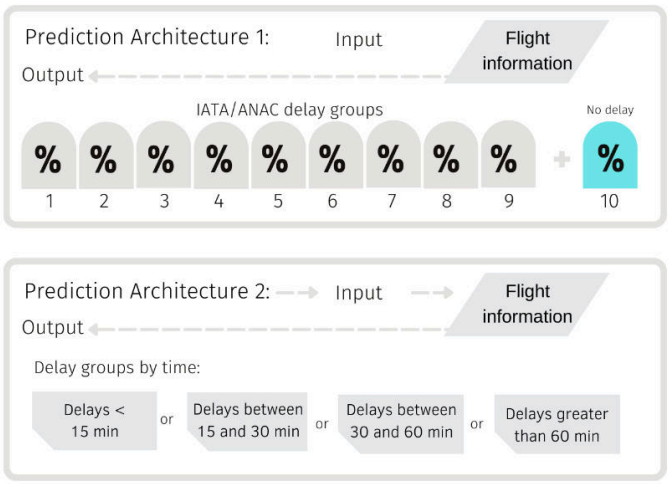
Therefore, a second prediction logic with a single dimension of classification (in this case, time) was necessary. Thus, the expected response by the network in this second logic was the time interval in which the delay falls. The classification using time intervals by Khan et al. was employed (2021) as part of the so-called "hierarchical analysis" proposed in their study. In this classification, the authors partitioned the database so that the flight prediction occurred according to the universe that favored its accuracy the most.

In this case, the predictive system of delays groups them according to their interval, which can be delays up to 15 minutes from the scheduled time (interval 1), between 15 and 30 minutes (interval 2), between 30 and 60 minutes (interval 3), or delays over 60 minutes (interval 4).

Thus, the logical prediction models can be summarized according to Figure 1.

.

Figure 1: Organization of the prediction logics applied in the study.



3.1.2. 2nd Stage: Selection of variables for the model

The first issue encountered when defining the explanatory variables in a forecasting model is the availability of compatible data before takeoff. In a way, the values predicted for operations, weather, load, and expected flight passengers, are used. Using data already predicted adds more uncertainties to the model. However, the information that will feed the system must be capable of approximating the estimated responses to the daily reality of operations, whether they come from the traffic operator, the airline, or the airport itself, while ensuring the effectiveness of the proposed model.

Thus, based on the groups of influential variables already used in studies of delay prediction in commercial flights, the explanatory variables were selected for the network-defined output.

Data from reports, studies, and information on the Brazilian air market conditions made available by ANAC were collected. The historical series of statistical data on air transportation in Brazil, with a high degree of detail, allowed for capturing information for each flight performed at the analyzed airport.

A database of aerial meteorological information from Iowa State University in the United States was collected from the Iowa Environmental Mesonet to complement the data set required for categorizing delays by local weather conditions, regular meteorological reports (METeorological Aerodrome Report - METAR) for SBGR (São Paulo-Guarulhos International Airport)

Therefore, the set of available and calculated variables that will compose the study is presented in Table 3, along with the description and type of each applied variable.

3.1.3. 3rd Stage: Construction of the database

After collecting data regarding the set of selected variables for the prediction model, Table x, it was necessary to verify the occurrences of delays throughout the year 2019.

Thus, the data were processed every month, aiming to establish a relationship between the verified delay occurrences, the IATA/ANAC delay class, and the observed time interval for the delay from the perspective of SBGR (São Paulo-Guarulhos International Airport) and the selected airline, GLO.

.

Table 3: Set of variables applied in the forecasting study.

Group	Author	Variable description	Unity	Source
Flight Plan	Lambelho et al. (2018); Schultz & Reitman (2019);	Estimated departure month	-	ANAC Microdata
		Estimated departure day	-	
		Estimated departure hour	-	
		Estimated departure minutes	-	
		Estimated arrival month	-	

	Khan et al. (2021); Yua et al. (2019); Arora & Mathur (2020); Truong (2021)	Estimated arrival day	-	
		Estimated arrival hour	-	
		Estimated arrival minutes	-	
		Day of the week the flight occurred	-	
		Distance to be flown	Km	
		Airline responsible for the flight	-	
		Number of seats offered for the flight	-	
Air Traffic	Yua et al. (2019); Chen et al. (2021); Liu et al. (2019); Chen & Lin (2021); Truong (2021)	Simultaneous takeoffs predicted to occur in the whole hour interval in which the flight is scheduled to depart	-	
		Simultaneous landings predicted to occur in the whole hour interval in which the flight is scheduled to depart	-	
Flight Operations	Chen et al. (2021); Liu et al. (2019); Yua et al. (2019); Truong (2021)	Aircraft model payload to be used in the flight	Kg	
		Estimated maximum takeoff weight	Kg	
		Average operating speed of the flight	Km/s	
		Estimated flight duration	-	
Meteorological	Chen & Wang (2019); Liu et al. (2019); Oliveira et al. (2021)	Air temperature at the time of the flight, measured at 2 meters above the ground (°F)	F	Iowa State University: Iowa Environmental Mesonet - Iowa State University
		Wind Chill or Heat Index in Fahrenheit	F	
		Relative humidity at the time of the flight	%	
		Wind direction in degrees relative to true north	º	
		Wind speed	Knots	
		Atmospheric pressure	Inches	
		Visibility at the time of the flight	Miles	

However, for calculating the delay time, ANAC (National Civil Aviation Agency) provides data obtained from the history of active regular flights - VRA, including the estimated time for takeoffs at SBGR and the IATA/ANAC delay class. On the other hand, the historical series of flights provided by the airline contain the actual departure times for each flight. Similar to the case of meteorological data, which requires a connection with the historical series of flights through the date and time of flight execution.

Therefore, the database construction resulted from the combination of information contained in the three original datasets collected, with the similarity between them applied as the key factor for data integration.

3.1.4. 4th Stage: Data preprocessing

At this stage, the database previously formed was preprocessed by removing null values (N/A) and outliers identified in the variables and arranging them in a matrix format, ensuring greater efficiency in data computation.

The outliers were defined following the rules set by Yu et al. (2019), where data entries with extreme delay values at the top and bottom, 1% are considered outliers, thus being eliminated from the dataset. The remaining 98% of instances were directly used for analysis.

As required by the selected method, in supervised training for neural networks, the data needs to be separated into training and testing sets. For this, the proportion used by Chen et al. (2021) was replicated, where 80% of the data is allocated for the network's training phase and 20% for testing.

However, due to the high dispersion of the data resulting from the size of the formed database, it became evident the need for redistributing the classes for the first and second prediction logics.

Therefore, the SMOTE-TOMEK balancing method was applied to the formed

database to distribute the IATA/ANAC delay classes, which balances the first logic, and for the delay interval classes, which balances the second logic.

3.1.5. 5th Stage: Training and definition of architectures

After preprocessing the formed database, the training stage of the network is responsible for generalizing known results so that the causal factors approximate the expected output responses.

For this purpose, the architectures applied to the neural networks should be capable of expressing the highest accuracy in prediction, considering the following factors: the number of hidden layers, the number of neurons per layer, activation functions applied, and the programming algorithm. These factors should mathematically replicate the behavior of the data.

Thus, for this research, network architectures were defined based on the selected analysis areas, as shown in Table 4. The training was then performed for each architecture, assuming a maximum of 500 epochs for each training.

Table 4: Architectures applied in the training of the neural networks.

AREA	ANN ARCHITECTURE			
	Name	Layers	Neurons per layer	Activation Function
A (Airport - SBGR)	RNA - A1	3	20	Hyperbolic tangent sigmoid
A (Airport - SBGR)	RNA - A2	3	20	Log-sigmoid
B (Airline - GLO)	RNA - B1	3	20	Hyperbolic tangent sigmoid
B (Airline - GLO)	RNA - B2	3	20	Log-sigmoid Log-sigmoid

3.1.6. 6th Stage: Output testing and analysis of results

After training the network, the final stage of neural network modeling is testing the system with data not used in the training phase. This stage is responsible for evaluating the predictive capability of the created neural network based on the constructed architectures.

The network's accuracy results are measured using the F1-Score. For the first logic, correct predictions of delay groups (groups 1 to 9) are considered true positives (TP), while correct predictions for group 10 are considered true negatives (TN).

In the second logic, all delays with less than 15 minutes are classified as non-delays due to the small delay duration and ANAC's considerations in classifying a flight delay only for cases exceeding this time.

The neural networks were developed using the Python programming language, with the assistance of the PyCharm algorithm development environment. The languages used were Numpy for numerical data processing and TensorFlow for building the architectures and testing the networks.

4. RESULTS AND DISCUSSIONS

Based on the defined methodology, the results are presented according to the planned steps.

4.1. *Preliminary Step 0 and 1st Step: Definition of analysis areas and prediction logic determination*

In the preliminary step, the aim was to select one of the active Brazilian airlines at the airport to apply the analysis scenario. According to data provided by ANAC (2022), for the year with the highest number of passenger and aircraft movements at SBGR, 2019, the three main airlines, Azul, Latam, and Gol, accounted for over 35,000 delayed flights. Of these, 50.22% were TAM, 37.73% were GLO, and 12.05% were AZU.

However, despite TAM having the highest percentage of delays in 2019, it recorded a lower total number of flights with delays exceeding 15 minutes compared to GLO, as shown in Table 5.

.

Table 5: Occurrence of flight delays by airline at SBGR, year 2019.

Airline Code (ANAC)	*ICAO Code*	*On-time*	*Delay Group 1*	*Delay Group 2*	*Delay Group 3*	*Delay Group 4*
1000909	AZU	64,07%	17,04%	9,17%	6,62%	3,09%
1001093	TAM	59,11%	28,72%	6,18%	3,60%	2,39%
1000854	**GLO**	**47,49%**	**30,02%**	**10,57%**	**7,64%**	**4,29%**

Source: Adapted from ANAC (2022).

In the classification of delays observed in 2019 for GLO at SBGR, the predominant IATA/ANAC groups are related to logistical and fleet issues of the airline. Cases related to miscellaneous and reactive delays, group 8, specify situations with unclear reasons for the delay but also issues with flight connections and management. Group 4 refers to aircraft damage on the ground or during the flight that leads to the occurring delays.

Therefore, in addition to SBGR, the magnitude and consequences of the delays make GLO the selected airline for analysis.

4.2. *2nd and 3rd Stages: Selection of variables for the model and Formation of the database*

With the formed database, the statistical distribution of departure information at SBGR demonstrates that simultaneous movements during takeoff can indicate a capacity problem as a cause of delays.

The relationship between capacity and departure delays at Guarulhos can be observed by correlating the recorded delay times in the database with the expected number of simultaneous flights at the time of occurrence.

Therefore, due to the longer operation time, the higher the number of simultaneous departures (Figure 4 a), the expected trend is an increase in the observed delay time. The opposite occurs when comparing landings (Figure 4 b), in which the trend is decreasing.

Despite the predominance of unexplained occurrences in Group 9, the causes described as connection issues, passenger and baggage problems, and mechanical issues with aircraft indicate that the hypothesis of airport capacity failures for aircraft movement may not be significant based on the distribution into delay groups

As a result of predominant domestic operations, the use of aircraft with lower seating capacity is observed in the data. However, the fleet appears as one of the groups with the highest delays recorded for GLO in SBGR.

4.3. *3rd and 4th Steps: Database Formation and Data Preprocessing*

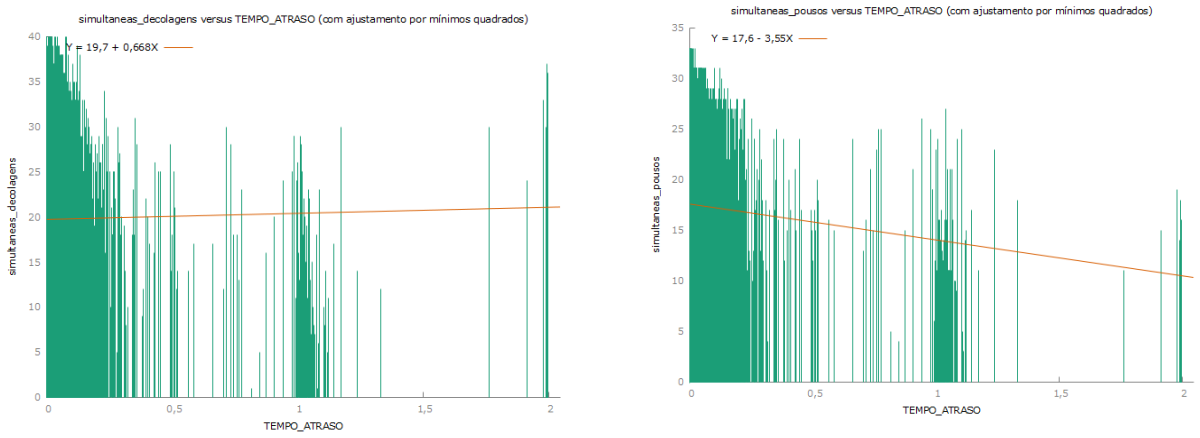
The database formed by the set of selected descriptive variables along the defined outputs for the prediction logics in constructed classes initially consisted of 189,547 flights. Through the preprocessing process involving the outlier's elimination, as Yu et al. (2019), there was a reduction in the number of instances in the final database to 94,381 departures.

After removing the outliers, the data maintained the distribution relationship between classes observed for SBGR. Thus, a problematic issue raised by Khan et al. (2021) became evident; the generalizability of the predictions for class-based studies being influenced by the balanced distribution of possible outputs.

In this regard, the database distribution highlights a predominance of lower magnitude delays, equal to or less than 15 minutes. Therefore, the mixed balancing method SMOTE-TOMEK, indicated as the most suitable in-flight delay prediction studies by Khan et al. (2021), was applied.

As a result, the balancing process assigned an equal number of instances to each class so that the predominant delay class ($\leq$ 15 minutes) had a reduction in the departure numbers for the second programming logic. However, the other classes had virtual value gains to compensate for the preliminary outcome.

Figure 4: Relationship between recorded delay time and simultaneous takeoffs, a), and recorded simultaneous landings within the hour interval, b).



a) b)

Table 6: Division of instances for training and testing.

Area of Analysis	***Prediction Logic***	***Instances Used***	
		Training (80%)	***Testing (20%)***
A (Airport - SBGR)	1st (IATA/ANAC Groups)	396192	99048
	2nd (Delay Interval)	233790	58448
B (Airline - GLO)	1st (IATA/ANAC Groups)	63981	15995
	2nd (Delay Interval)	63919	15980

Predicting logic based on the delay classes according to the IATA/ANAC groups requires balancing as the goal of the analysis at an airport, where it is possible to differentiate delays based on their possible causes.

Therefore, the result for this group shows considerable balancing with the generation of virtual instances for the less frequent classes The on-time departure group (no delay), group 10, is the only one that experiences a reduction in the number of instances.

Classifying on-time departures equally to the delay groups can enable the network to attribute the cause in flight analysis correctly. Still, it may introduce a false positive bias for delays, considering the common proportion between delays and on-time flights.

Therefore, the balanced databases for each prediction logic were applied to Artificial Neural Network models, with the datasets divided into training and testing sets according to the previously defined proportions. Thus, the set of instances used for training and testing in both analysis areas is presented in Table 6.

4.4. *5th and 6th Steps: Training and architecture definition, Testing the output and analysis of results*

The networks were modeled in Python using the PyCharm development environment. The dataset, divided into test and train sets, was applied to the defined architectures, allowing the observation of training responses.

Next, the test results were categorized based on correctness and error considerations to form the correlation matrix. Thus, the summary of results for training and testing, according to the analyzed areas and logic, is presented in Table 7.

Among the tested architectures, the activation function that best describes the relationships between the selected variables and the occurrence of delays is the hyperbolic tangent sigmoid (TANSIG) function. The numerical variation provided by the sigmoid function allows the system to assign more response classes to the network due to the positive and negative amplitude of the function.

Since the proposed model assigns four to ten distinct classes, the larger the numerical range allowed by the generated wave's amplitude, the greater the possibility of distinction between classes.

This hypothesis is supported by comparing it with the responses of the log-sigmoid model, which has a negatively limited amplitude range. The tests showed not only a lower accuracy, according to the calculated F1-Score in a specific analysis of the results but also indicated biases for the delay groups in both logics.

Another point to consider is the division of prediction logics. As indicated in the first part, more specific analysis groups can achieve higher accuracy in the system's response.

When both airport and airline areas are analyzed, the results indicated two opposite situations for each logic. The first logic is based on grouping flights into delay categories. This division, besides having the largest number of groups, incorporates specific variables and occurrence conditions for each situation. In summary, it can be concluded that the first prediction logic has a variety of units considered for each possibility of delay.

On the other hand, the second prediction logic uses only one unit to classify the system inputs, which is the delay time. Combined with a smaller number of possible response groups, four instead of ten, the results in Table L demonstrate the higher predictive capability of the model according to this logic.

However, even though the second prediction logic proves to be more suitable due to the observed accuracy in the tests, the area of analysis with the highest accuracy is the airport analysis, not the airline analysis, as indicated by the results in the first part.

This result supports the proposition made in the first part when defining the ideal prediction scenario, as characterizing the scenario in terms of the studied area is more related to the availability and capacity of the obtained data to describe the phenomenon, in this case, the delay in a specific flight rather than the sample size or the investigation region.

However, the set of variables selected provides more information about the airport than the specific airline. The gap in terms of variable specificity is a result of the lack of availability of information about the operations carried out at SBGR or by GLO. The use of data from ANAC or the University of Iowa provides a macro view of each flight. Moreover, with a difference of only 4% in accuracy, it can be said that the models are statistically nearly tied.

Specifically for accuracy, both prediction logics in classes do not reach the acceptable accuracy defined by the meta-analysis. With a maximum of 44.46% accuracy according to the F1-Score, the regression-type analysis is still below the minimum of 77.22% determined in the first part.

When trying to find a predictive system applicable to the operational reality experienced in an airport and an airline, the variables that were known in advance and made available for study were the main limitations observed.

In comparison to the specificity of the assumed variables, for example, Wei & Vazea (2018) used the times verified for each crew boarding stage in an airline, while Scarpel & Pelicioni (2018) used the estimated times for each flight operation and the airport's capacity at the time of takeoff.

Thus, a hypothesis still relies on the definitions made in the first part, that the specificity of the data has a greater influence on the response than the sample size.

The distribution of results in the confusion matrices also shows that, for the first prediction logic, the predominance of false negatives (delayed flights classified as non-delayed) occurs due to the balance between the adopted classification groups. In general, the class distribution ends up prioritizing the delays when nine delay groups are defined, while there is only one group for non-delay cases.

Table 7: Summary of results obtained by the tested neural networks for training performance and test accuracy.

PREDICTION LOGIC	*AREA*	*ANN ARCHITECTURE*				*TRAINING (80%)*			*TEST (20%)*			*ACCURACY*
		Name	*Layers*	*Neurons per Layer*	*Activation Function*	*Data*	*Epochs*	*Performance*	*Dados*	*Matriz de Confusão*		*F1-Score*
1st: IATA/ANAC Delay Group	A (Airport - SBGR)	ANN - A1	3	20	TANSIG	396192	785	54%	99048	12918 4820	80897 413	0,134
	A (Airport - SBGR)	ANN - A2	3	20	LOSIG	396192	500	54%	99048	11733 4594	82082 639	0,124
	B (Airline - GLO)	**ANN - B1**	**3**	**20**	**TANSIG**	**63981**	**554**	**54%**	**15995**	**2227** **1299**	**12155** **314**	**0,158**
	B (Airline - GLO)	ANN - B2	3	20	LOSIG	63981	500	54%	15995	1864 1182	12518 431	0,143
2nd: Delay Interval	**A (Airport - SBGR)**	**ANN - A1**	**3**	**20**	**TANSIG**	**233790**	**500**	**69%**	**58448**	**11905** **18011**	**14453** **14079**	**0,444**
	A (Airport - SBGR)	ANN - A2	3	20	LOSIG	233790	500	56%	58448	11716 32090	14642 0	0,200
	B (Airline - GLO)	ANN - B1	3	20	TANSIG	63919	500	64%	15980	4037 5119	4190 2634	0,417
	B (Airline - GLO)	ANN - B2	3	20	LOSIG	63919	500	58%	15980	4606 7753	3621 0	0,2885

Although still significant for the second prediction logic, the reduction of false negatives represents a better balance of the classes adopted in the prediction. However, even with only one dimension applied in the adopted classes (delay time), the groups assigned to the "delay" situation are larger than those for non-delay cases.

Thus, although the selected influential variables are a possible cause for the models not reaching the acceptable ranges of accuracy obtained in the first part of the study, the distribution of the adopted classes in both logics can also be pointed out as a consequence of the observed responses, as they prioritize delays over non-delay cases.

5. CONCLUSIONS

Based on the discussions presented, there is a clear trend towards applying more robust methods of data analysis and regression for predicting delays in global aviation. Given the multitude of machine learning methods available, some stand out in terms of complexity and availability of the required information for conducting these predictions.

In the conducted case study, the analyses for SBGR (airport) proved to be more efficient than for the selected airline, GLO. However, the accuracy obtained is outside acceptable limits. It is assumed that the required data lead time, with the availability of less specific data, contributed to this result.

This situation highlights the need for a specific analysis of variables since, enumerating the causes, specifying the

responses by delay group can be an alternative to the first prediction logic used.

Despite the considerable amount of data used, defining isolated prediction models for each type of delay, by using specific variables for each expected output, the classification logic based on delay groups and causes may achieve higher accuracies both in airport operations and airlines.

Additionally, another limiting factor to consider in the study is the use of only one prediction method. Even with the adaptive capacity of ANN architectures, other models defined in the first part can be tested for the range of accuracy for the applied set of variables.

Therefore, the study manages to prove that the accuracy of the methods is linked to the specific definition of the area to be analyzed and the main factors that generate these occurrences.

Considering the applied methodology, some recommendations for further work can be enumerated to enhance the solution's effectiveness, such as obtaining specific technical information about the analyzed airport, including knowledge about the preceding and subsequent stages of each flight or the instrumentation used for takeoff and the runway and threshold used.

The use of network variables to improve model estimation, and defining the interactions between these variables and the influencing factor, can also be explored. Additionally, the use of time series in high-dimensional and high-frequency data can aim for specificity for each flight.

Lastly, for assigning prediction classes, an alternative to be tested is the serial comparison, where each delay group is individually tested against the non-delay group to minimize the dimensionality issue.

REFERENCES

AENA. (2016). *AENA BRASIL*. Fonte: https://www.aenabrasil.com.br/pt/aeroportos/aeroporto-internacional-do-recife-guararapes-gilberto-freyre/Caracteristicas.html

ANAC. (04 de 02 de 2021). *Metadados do conjunto de dados: Percentuais de atrasos e cancelamentos.* Fonte: Agência Nacional de Aviação Civil: https://www.anac.gov.br/acesso-a-informacao/dados-abertos/areas-de-atuacao/voos-e-operacoes-aereas/percentuais-de-atrasos-e-cancelamentos

Arjomandi, A., Dakpo, K., & Seufert, J. (2018). Have Asian airlines caught up with European Airlines? A by-production efficiency analysis. *Transportation Research Part A*, pp. 389-403.

Arora, S. D., & Mathur, S. (2020). Effect of airline choice and temporality on flight delays. *Journal of Air Transportation Managemant*.

Button, K. (2019). Applied economics and understanding trends in air transportation policy. *Transport Policy*, *80*, 78-85.

Button, K. (2019). Applied economics and understanding trends in air transportation policy. *Transport Policy*.

Chen, Y.-H., & Lin, J.-J. (2021). Determinants of flight delays at East Asian airports from an airport, route and network perspective . *Journal of Air Transport Management*.

Chen, Z., & Wang, Y. (2019). Impacts of severe weather events on high-speed rail and aviation delays. *Transportation Research Part D, 69*, 168-183.

Chen, Z., Wang, Y., & Zhou, L. (2021). Predicting weather-induced delays of high-speed rail and aviation in China. *Transport Policy*, *101*, 1-13.

Chen, Z., Wang, Y., & Zhou, L. (2021). Predicting weather-induced delays of high-speed rail and aviation in China. *Transport Policy*, *101*, 1-13.

Chicco, D., & Jurman, G. (21 de Junho de 2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*.

Christopher J. , O., D. S. Prasada , R., & George E., B. (2008). Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. *Empirical Economics*, pp. 231-255.

Chung, H. M., Kwon, O. K., Han, O. S., & Kim, H.-J. (2020). Evolving network characteristics if the asian international aviation market: A weighted network approach. *Transport Policy*.

Chung, S.-H., Ma, H.-L., Hansen, M., & Choi, T.-M. (2020). Data science and analytics in aviation. *Transportation Research Part E, 134*.

Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health. 2009;12(4):409–18.

Garrow, L., German, B., & Leonard, C. (2021). Urban air mobility: A comprehensive review and comparative analysis with autonomous and electric ground transportation for informing future research. *Transportation Research Part A*.

IBGE, I. (2020). *Censo 2010: Índice de desenvolvimento humano municipal*.

Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation*

Research Part C: Emerging Technologies, 387-399.

Kaye, S.-A., Somoray, K., Rodwell, D., & Lewis, I. (2021). Users' acceptance of private automated vehicles: A systematic review and meta-analysis. *Journal of Safety Research*, pp. 352-367.

Khan, W. A., Ma, H.-L., Chung, S.-H., & Wen, X. (2021). Hierarchical integrated machine learning model for predicting flight departure delays and duration in series. *Transportation Research Part C, 129*.

Lambelho, M., Mitici, M., Pickup, S., & Marsden, A. (2020). Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *Journal of Air Transport Management*.

Liu, Y., Liu, Y., Hansen, M., Pozdnukhov, A., & Zhang, D. (2019). Using machine learning to analyze air traffic management actions: Ground delay program case study. *Transportation Research Part E, 131*, 80-95.

Liu, Y., Yin, M., & Hansen, M. (2019). Economic costs of air cargo flight delays related to late package deliveries. *Transportation Research Part E, 125*, 388-401.

Luger, G. F. (2004). *Inteligência Artificial: Estruturas e Estratégias para a resolução de problemas complexos.* Porto Alegre: Bookmann.

Marcy, M. C. (2017). *METODOLOGIA BASEADA EM REDES NEURAIS ARTIFICIAIS PARA A DETECÇÃO DE DANOS ESTRUTURAIS.* Brasília: Tese de Doutorado apresentada a Universidade de Brasília - UnB.

Odeck, J., & Bråthen, S. (2012). A meta-analysis of DEA and SFA studies of the technical efficiency of seaports: A comparison of fixed and random-effects regression models. *Transportation Research Part A*, pp. 1574-1585.

Oliveira Filho, P. (2015). *Epidemologia e bioestatística: Fundamentos para a leitura crítica.* Rio de Janeiro: Editora Rubio.

Oliveira, M., Eufrásio, A. R., Guterres, M. X., Murça, M. C., & Gomes, R. (2021). Analysis of airport weather impact on on-time performance of arrival flights for the Brazilian domestic air transportation system. *Journal of Air Transport Management*.

Rennert, L. (2022). meta-analysis of the impact of rail stations on property values: Applying a transit planning lens. *Transportation Research Part A*, pp. 165-180.

Scarpel, R., & Pelicioni, L. (2018). A data analytics approach for anticipating congested days at the São Paulo International Airport. *Journal of Air Transport Management*.

Schultz, M., & Reitmann, S. (2019). Machine learning approach to predict aircraft boarding. *Transportation Research Part C, 98*, 391–408.

Simon, H. (1983). Why should machines learn? Em R. S. Michalski, J. G. Carbonell, & T. M. Mitchell, *Machine Learning: An Artificial Intelligence Approach* (pp. 25-37). Berlin : Springer-Verlag.

Singh, V., Singh, V., & Vaibhav, S. (2020). A review and simple meta-analysis of factors influencing adoption of electric vehicles. *Transportation Research Part D*.

Stivali, M. (2022). *Valor de uma vida estatística: uma revisão da literatura empírica para o Brasil.* Brasília - DF. Brasil.: IPEA – Instituto de Pesquisa Econômica Aplicada. Programa de Pesquisa para o Desenvolvimento Nacional - PNPD.

Truong, D. (2021). Using causal machine learning for predicting the risk of flight delays in air transportation. *Journal of Air Transport Management, 91*.

Wang, A., Xu, J., Tu, R., Saleh, M., & Hatzopoulou, M. (2020). Potential of machine learning for prediction of traffic related air pollution. *Transportation Research Part D*.

Wang, C., & Wang, X. (2019). Airport congestion delays and airline networks. *Transportation Research Part E, 122*, 328–349.

Wardman, M. (2022). Meta-analysis of price elasticities of travel demand in great britain: Update and extension. *Transportation Research Part A*, pp. 1-18.

Wei, K., & Vazea, V. (2018). Modeling Crew Itineraries and Delays in the National Air Transportation System. *Transportation Science*.

Yen, B., & Li, J.-S. (2022). Route-based performance evaluation for airlines – A metafrontier data envelopment analysis approach. *Transportation Research Part E*.

Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E, 125*, 203-221.

.