VILNIUS TECH
Vilnius Gediminas Technical University

AVIATION
ISSN: 1648-7788 / eISSN: 1822-4180
2023 Volume 27 Issue 1: 47–56
https://doi.org/10.3846/aviation.2023.18641

AVIATION ACCIDENT AND INCIDENT FORECASTING COMBINING OCCURRENCE INVESTIGATION AND METEOROLOGICAL DATA USING MACHINE LEARNING

Mauro CAETANO *

Aeronautics Institute of Technology (ITA), São José dos Campos/SP, Brazil

Received 12 April 2022; accepted 12 October 2022

Abstract. Studies on safety in aviation are necessary for the development of new technologies to forecast and prevent aeronautical accidents and incidents. When predicting these occurrences, the literature frequently considers the internal characteristics of aeronautical operations, such as aircraft telemetry and flight procedures, or external characteristics, such as meteorological conditions, with only few relationships being identified between the two. In this study, data from 6,188 aeronautical occurrences involving accidents, incidents, and serious incidents, in Brazil between January 2010 and October 2021, as well as meteorological data from two automatic weather stations, totaling more than 2.8 million observations, were investigated using machine learning tools. For data analysis, decision tree, extra trees, Gaussian naive Bayes, gradient boosting, and k-nearest neighbor classifiers with a high identification accuracy of 96.20% were used. Consequently, the developed algorithm can predict occurrences as functions of operational and meteorological patterns. Variables such as maximum take-off weight, aircraft registration and model, and wind direction are among the main forecasters of aeronautical accidents or incidents. This study provides insight into the development of new technologies and measures to prevent such occurrences.

Keywords: air transport, artificial intelligence, aviation accident, aviation incident, innovation, machine learning, safety.

Introduction

The use of computational tools in the analyses of large volumes of data, particularly those incorporating the principles of artificial intelligence (AI), can be considered a powerful instrument for forecasting and preventing aeronautical accidents. Since the first studies on AI were conducted by John McCarthy in the 1950s, the technology has been applied in many areas for different purposes. Its main applications in aviation can be identified from the studies conducted by Gosling (1987) and include the strategic management and tactical control of air traffic flow, improved simulation techniques, aircraft on-board equipment development, and the development of Next-Generation Air Transportation System (NextGen) (Post, 2021), among others involving high levels of complexity in decision-making. Such decision-making can be supported by identifying patterns in the behavior of input data and their effects on output data.

Several examples of AI applications in aviation can be identified from recent literature in which machine learning techniques have been used to develop algorithms for forecasting and preventing aeronautical accidents (Patriarca et al., 2022); detecting normality or anomalies in operations from flight data (Puranik & Mavris, 2020; Stogsdill et al., 2021; Xu et al., 2020); providing support for airport pavement maintenance (Barua & Zou, 2021); forecasting take-off times (Dalmau et al., 2021); predicting the true air and ground speeds during aircraft touchdown (Puranik et al., 2020); and defining airport capacity (Choi & Kim, 2021), airport congestion, and arrival delays (Rodríguez-Sanz et al., 2019). However, it is noted that such studies predict occurrences based on target variables that may influence operational flight safety and do not consider real data from accidents or incidents, such as meteorological conditions that can affect such accidents or incidents.

To fill this theoretical gap, this paper proposes forecast procedures for aeronautical occurrences, which are a function of parameters considered in the investigation of accidents and meteorological conditions. Machine learning techniques are used to support the collection, processing, and treatment of data using Brazilian cases as a reference.

*Corresponding author. E-mail: *caetano@ita.br*

1. Artificial intelligence and aviation safety

Three different machine learning paradigms in AI have been proposed by Sutton and Barto (2018): (a) supervised learning, i.e., the processing of data from datasets and external supervisor knowledge; (b) unsupervised learning, i.e., the identification of mathematical structures in an unlabeled dataset; and (c) reinforcement learning, i.e., the constant interaction in the analysis of a given problem, without prior definition of the actions to be taken, demonstrating the consequences of certain actions over time. In these paradigms, different techniques and algorithms, such as the Bayesian network, decision tree, gradient boosting, neural network, and random forest, which have different advantages according to the type and amount of data analyzed (Truong & Choi, 2020), can be used to design predictive models in relation to a given object.

For operational safety in aviation, considering the approach and landing phases of general aviation, Puranik and Mavris (2018) identified flight-level anomalies from procedures such as one-class support vector machine (SVM) and density-based clustering algorithm (DBSCAN) in the grouping of the analyzed data in clusters in such a way that the outliers were highlighted, thus indicating possible anomalies associated with the considered parameter. The relevance of the computational tool used in data analysis is not only in the identification of patterns, but also in the identification of outliers that may influence the safety of operations.

Meanwhile, Xu et al. (2020) used a deep neural network, among other different classifiers (k-nearest-neighbor, decision tree, adaboosted decision tree, random forest, naive Bayes), to analyze 1,572 accident data from January 2005 to December 2015. The study considered this method to be the best for accident prediction. Based on this, the authors used data from 825 accidents involving only helicopters.

A review of the literature indicates that a significant volume of historical data is commonly used to predict future recurrences. However, in the case presented by Patriarca et al. (2022), the authors recommend the internal adoption of business intelligence and machine learning solutions by air navigation service providers (ANSPs), and using available data in a self-service safety intelligence approach, which the authors consider as democratizing safety intelligence in aviation. According to the authors, ANSPs must continually develop business intelligence and machine learning applications from traditional databases, which can be improved from the identified solutions. The four different phases proposed by the authors refer to the analysis and collection of information needs, planning of the architecture, development and practical application of the solution. Table 1 presents some of the main studies on the use of AI in aviation safety.

As shown in Table 1, studies commonly use data from aeronautical operations, such as vertical speed, angle of attack, altitude, flight duration, and number of seats (Dalmau et al., 2021; Puranik & Mavris, 2018; Rodríguez-Sanz et al., 2021), or meteorological conditions, such as wind direction, wind speed, visibility, and temperature (Choi & Kim, 2021; Schultz et al., 2021), and airport operating conditions, such as capacity and infrastructure management (Barua & Zou, 2021; Rodríguez-Sanz et al., 2019). A significant gap can be identified in the literature when considering the actual data on aeronautical accidents and incidents and their correlation with climatic variables, which is the main contribution of this study to the state of the art.

2. Methodology

This study used a process of three steps to forecast certain phenomena from a set of operational and meteorological data, starting with the identification, collection, and selection of data, identified here as feature engineering, followed by the identification and application of possible machine learning classifiers, and finally, ending with the development of the forecasting algorithm based on feature importance.

Data from 6,188 aeronautical accidents, incidents, and serious incidents that occurred in Brazil between January 03, 2010, and August 08, 2021, investigated by the Center for Investigation and Prevention of Aeronautical Accidents (CENIPA), as well as meteorological data from approximately 2.6 million measurements collected during the same time period in two different automatic weather stations, were analyzed. Based on the analysis, machine learning and data analytics were used to identify patterns in occurrences and their relationship with meteorological conditions. An algorithm capable of predicting certain occurrences as a function of operational and meteorological conditions was proposed.

Different classifiers, such as decision trees, extra trees, Gaussian naive Bayes, gradient boosting, and k-nearest neighbor have been used. However, the random forest classifier was mainly adopted for prediction as it presented the best accuracy in the prediction model, as discussed in the results and was also adopted by Puranik et al. (2020), Rodríguez-Sanz et al. (2021), and Xu et al. (2020). Given that the random forest classifier presented the best accuracy in relation to the other classifiers adopted in various analyses, it was chosen for the development of the forecasting algorithm.

For different classes of aeronautical occurrences (Y), with y labeled between accident, incident, and serious incident, and X is a set of predictive data, such as the type of aircraft and maximum take-off weight (MTOW) from the occurrence dataset, and the wind speed and temperature from the meteorological dataset, different probabilities (P) can be identified for y as a function of the behavior of (x), such that $\{p(j, m)\}$ is given by j varying between 1 and 3, and m between 1 and 22, which refers to the number of inputs adopted for the classification of y.

From the set of classifiers $h1(x)$, $h2(x)$, ..., $hk(x)$, considered in the implication of x in y, several combinations

Table 1. Studies using machine learning in the analysis of aviation safety

Authors	Focus of study	Main methods or machine learning algorithms	Case study	Features
Barua and Zou (2021)	Cost reduction on maintenance and rehabilitation of airport pavement.	Q-learning and gradient boosting machine.	Chicago O'Hare International Airport (ORD).	PCI, time elapsed, pavement age, pavement type, number of minor M&R actions, occurrence of a major M&R, location for runways only, rainfall, freeze-thaw cycles, and traffic loading.
Choi and Kim (2021)	Influence of meteorological conditions on airport capacity.	Artificial neural network, multilayer perceptron, recurrent neural networks, and long short-term memory.	Hartsfield–Jackson Atlanta International Airport (ATL).	Date, time, and terminal weather (wind speed and direction, temperature, cloud height, sea level pressure, visibility, rainfall, and snow depth).
Dalmau et al. (2021)	Takeoff time predictions from the submission of the initial flight plan (IFP) to the actual take-off time (ATOT).	Enhanced tactical flow management system and gradient boosted decision trees.	EUROCONTROL Maastricht Upper Area Control Centre (MUAC).	Available turn-around time, flight duration, taxi time, previous flight leg, time from last message, time to destination, and ATFM regulations.
Herrema et al. (2021)	Analysis of runway exits.	Gradient boosting.	Vienna International Airport (VIE).	SODAR velocity, SODAR direction, wind speed, ground speed 5NM, height 5NM, cloud, visibility, height 2NM, ground speed 2NM, ICAO cat, and ACType.
Puranik and Mavris (2018)	Identification of outliers in General Aviation approach and landing operations.	One-class support vector machine and density-based clustering algorithm.	Garmin G1000, from training flights on a Cessna 172S.	Flight parameters collected from flight data (vertical speed, altitude above ground level, etc.) and energy metric (specific potential energy, kinetic energy, etc.).
Puranik et al. (2020)	Predict the approach speed and touch the runway.	Random forest.	Aircraft groups: B737, B777, B757, MD80, A320, and A330, from operations in 70 airports.	True airspeed, angle of attack, wind direction, vertical speed, location of touchdown point on runway, etc.
Rodríguez-Sanz et al. (2019)	Airport arrival congestion and delay.	Bayesian network approach and multi-state system structure with Markov process technique.	Adolfo Suárez Madrid-Barajas Airport (MAD).	Airport infrastructure (runway configuration and arrival declared capacity), airline, aircraft and route (operator type, flight origin), operational times (time frame, taxi-in delay), arrival congestion index, arrival throughput, and meteorology (wind, clouds).
Rodríguez-Sanz et al. (2021)	Patterns of passenger behavior in check-in lines and airport security control.	Random forest.	Palma de Mallorca Airport (PMI).	Number of inter-island flights scheduled, Schengen flights scheduled, international flights, mean number of seats on inter-island flights, maximum number of seats on inter-island flights, minimum number of seats on inter-island flights scheduled within the next 90 minutes.
Schultz et al. (2021)	Meteorological conditions and airport performance.	Artificial neural network, convolutional neural network, and recurrent neural network (long short-term memory).	London–Gatwick Airport (LGW).	METAR data (wind direction and speed, visibility, temperature, and rain), and ATFM data (traffic demand, mix of ARR-/DEP movements).
Truong and Choi (2020)	Operational security in violation of airport airspace by a small unmanned aircraft system (sUAS).	Classification regression, decision tree, neural network, gradient boosting, random forest, Bayesian networks, and memory-based reasoning.	FAA's UAS sightings report sets between October 2016 and September 2017 (sample size of 2,088).	Event date and time, city, state, and event report narrative such as incident time, location, violation type, etc.
Xu et al. (2020)	Analysis of helicopter accidents.	K-nearest-neighbor, decision tree, adaboosted DT, random forest, naive Bayes, and deep neural network (DNN).	Data of civil helicopters and accidents from FAA, NTSB, and helicopter manufacturers.	Number of main rotor blades, number of engines, rotor diameter, and weight.

are identified in the construction of cause-effect trees, such that the random vector Θ is generated for the *kth* tree. Thus, classifier *h(X, Θk)* is identified from input vector *x*. As several trees are built, the most common class *(j)* is presented with a certain probability of success, identified as accuracy, thus defining a random forest (Breiman, 2001). Using random forest, the random distribution of the vector *Y*, *X* is defined by the generalization of the error *PE*, presented in Equation (1), based on the margin function *mg(X, Y)*, presented in Equation (2).

$$PE = P_{X,Y}\left(mg\left(X, Y\right) < 0\right); \quad (1)$$

$$mg\left(X, Y\right) = av_k I\left(h_k\left(X\right) = Y\right) - \max_{j \neq Y} av_k I\left(h_k\left(X\right) = j\right). \quad (2)$$

Because *I* is an indicator function, as the number of test trees increases, considering *hk(X)=h(X,Θk)*, the probability of *X* in *Y*, *P(X,Y)* is expressed by Equation (3).

$$P_{X,Y}\left(P_\Theta\left(h(X,\Theta) = Y\right) - \max_{j \neq Y} P_\Theta\left(h\left(X,\Theta\right) = j\right) < 0. \quad (3)$$

Finally, the result of applying the random forest consists of the presentation of the most common class identified in the classifiers, as shown in Figure 1.

The different classes obtained from Figure 1 follow the taxonomy proposed by the International Civil Aviation Organization (International Civil Aviation Organization [ICAO], 2020), which presents three classes of occurrence: incident, serious incident, and accident, named according to their degrees of severity. An incident is associated with the operation of an aircraft, in which a certain action may compromise the safety of people or the aircraft itself. A serious incident, on the other hand, is related to a critical situation in which a certain incident can lead to an accident. The accident, in turn, is associated with the operation of an aircraft in which there is significant damage to the aircraft or a certain person, on board or outside the aircraft, suffers serious injury or death. To illustrate this taxonomy, Figures 2, 3, and 4 show examples of incidents, serious incidents, and accidents, respectively.

Figure 2 shows an incident, where a B777 aircraft taxiing at the Antonio Carlos Jobim Airport (GIG) collided its wingtip with the vertical stabilizer of a parked B737 aircraft. Only these aircraft components showed significant damage, and none of these aircraft were injured. Figure 3 shows a serious incident, where the aircraft, a helicopter, lost the tip of one of the tail rotor blades (tipcap) and had

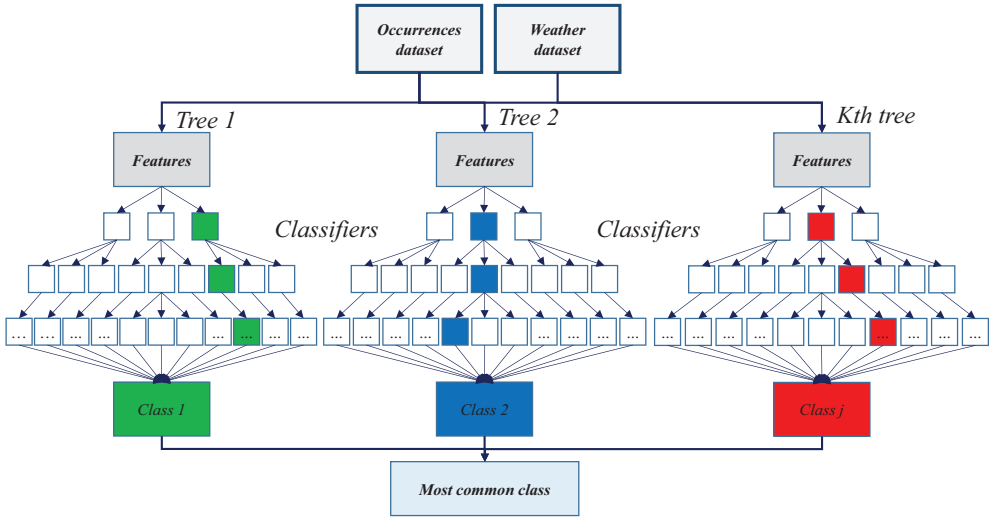

Figure 1. Random forest built from different datasets

Figure 2. Incident (wingtip ground collision of the B777 with the vertical stabilizer of the B737) (source: Aeronautical Accidents Investigation and Prevention Center [CENIPA], 2019)

Figure 3. Serious incident (component loss in flight) (source: CENIPA, 2015)

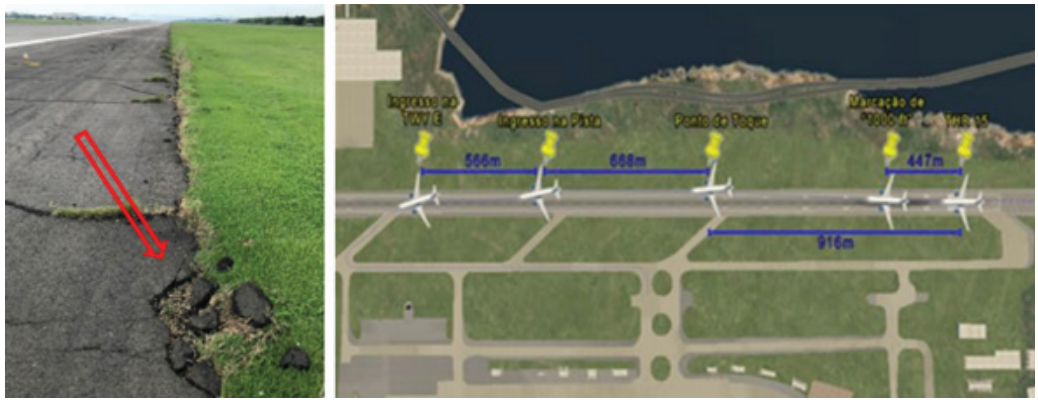

Figure 4. Accident (point of first contact of the right main landing gear – left, and the aircraft's ground trajectory – right) (source: CENIPA, 2013)

to perform an emergency landing. The two crew members on board were unharmed, and the aircraft sustained minor damage. Finally, in Figure 4, during the landing procedure of an A320 aircraft at Antonio Carlos Jobim Airport (GIG), the aircraft touched down outside the limits of the runway, which caused serious damage to the landing gear with no injuries sustained by the crew or passengers.

Then, data from 6,188 occurrences of aeronautical accidents, incidents, and serious incidents that occurred in Brazil between January 3, 2010, and August 8, 2021, recorded by the Center for Investigation and Prevention of Aeronautical Accidents (CENIPA, 2021), and available as open public data were considered. It is necessary to clarify that, although this value of 6,188 occurrences is a high value for the approximately 12 years considered, it represents only about 0.03% of all 18,206,015 aircraft movements over this period (CGNA, 2022). The combination of data referring to the aircraft involved in these occurrences and the type of occurrence, such as occurrence code, aircraft registration number, aircraft model, and seats, totaled 31 variables, corresponding to 191,828 observations. The data frame contained common variables such as the country of occurrence, all of them in Brazil, aircraft model, already represented by the ICAO model, as well as seats in the aircraft involved in the occurrences, aircraft year of manufacture, and status of investigation. Other non-representative variables in the model were removed from the analysis as they presented many empty lines. In addition, variables such as damage level and number of fatalities were removed from the mathematical analysis because such variables are presented as consequences of occurrences, and not necessarily conditioning factors. In addition to the occurrence class variable, 14 more relevant variables referring to aeronautical occurrences are listed in Table 2.

Table 2. Characterization of variables related to aeronautical occurrences

Variable	Description	Measure	Coding format
Aerodrome	Aerodrome of occurrence	Nominal	1 = Out of aerodrome, 2 = SIBH, ..., 19 = SDPG
Class	Classification of the type of occurrence	Nominal	1 = accident, 2 = incident, 3 = serious incident
City	City where the occurrence was registered	Nominal	1 = Guarulhos/SP, 2 = Rio de Janeiro/RJ, 3 = São Paulo/SP
Day	Day of occurrence	Nominal	1 = Jan/03/2010, 2 = Jan/10/2010, ..., 696 = July/27/2021
Hour	Hour of occurrence	Nominal	1 = 0:00:00, 2 = 1:00:00, ..., 24 = 23:00:00
Destination	Flight destination	Nominal	1 = not defined, 2 = Adalberto Mendes da Silva, ..., 75 = Zumbi dos Palmares
EngineType	Aircraft engine type	Nominal	1 = jet, 2 = piston, 3 = turboshaft, 4 = turboprop
EngineNumber	Number of aircraft engines	Nominal	1 = not defined, 2 = twin engine, 3 = single engine, 4 = no traction, 5 = tri engine
ICAOModel	ICAO model of the main aircraft involved in the occurrence	Nominal	1 = A109, 2 = A119, ..., 112 = not defined
MTOW	Maximum Takeoff Weight	Continuous	Weight in tonne
OperPhase	Operation phase	Nominal	1 = not defined, 2 = final approach, ..., 31 = low flight
Origin	Origin of the flight	Nominal	1 = not defined, 2 = Campinas – Amarais State Aerodrome, ..., 82 = Campinas – Viracopos Internacional Airport
Registry	Aircraft registry	Nominal	1 = 9VSWS, 2 = AF443, ..., 569 = PUVSM
Segment	Purpose of the flight	Nominal	1 = not defined, 2 = direct administration, … 11 = air taxi
Type	Type of aircraft	Nominal	1 = airplane, 2 = helicopter, 3 = ultralight

Approximately 40% of the occurrences occurred outside the limits of the aerodromes. Once the main characteristics of the occurrences were identified, the cities with the highest rates of occurrence were defined for the analysis of meteorological data, such that three of the cities with the highest number of occurrences (Rio de Janeiro/RJ, São Paulo/SP, and Guarulhos/SP), corresponding to 13% of the total, were used as a reference in relation to the other 1,130 cities in the occurrence data frame.

Regarding the meteorological data of these cities, open access data, available from the National Institute of Meteorology (Instituto Nacional de Meteorologia [INMET], 2021), was used, and measurements were carried out at 588 automatic weather stations active in 2021, distributed throughout the national territory and the Atlantic Ocean. Hourly measurements between January 01, 2010, and August 08, 2021, were considered, accounting for 8,760 annual measurements at each automatic meteorological station. Thus, for the two automatic stations considered in the study period, 201,096 measurements were identified for 13 meteorological measurement parameters, totaling 2,614,248 observations.

As two of these cities are neighbors, Guarulhos/SP and São Paulo/SP, data from a single weather station, the São Paulo – Mirante Santana (SPMS) Automatic Station (Lat: –23.495589696406164, Long: –46.61985646210017, Alt: 785.16 m) was adopted. For Rio de Janeiro, data from the Rio de Janeiro – Forte de Copacabana (RJCP) Automatic Station (Lat: –22.98807188482133, Long: –43.19044887942751, Alt: 25.59 m) was used. The two weather stations had the same measurement variables.

Variables that presented inconsistent data or a lack of data were removed to adopt the most representative variables in the forecast model. As a result, eight meteorological variables were used, as listed in Table 3. Some variables such as global radiation, maximum dew temperature in

Table 3. Characterization of meteorological variables

Variable	Description	Measure
AtmPres	Atmospheric pressure hourly at station level	mb
AtmPresMax	Maximum atmospheric pressure in old hour	mb
AtmPresMin	Minimum atmospheric pressure at the previous hour	mb
Precipitation	Total precipitation hourly	mm
Temperature	Air temperature hourly (dry bulb)	°C
WindDirection	Wind direction hourly	°gr
WindBlast	Maximum wind blast hourly	m/s
WindSpeed	Wind speed hourly	m/s

Table 4. Examples of adopted codes

Input: city, aerodrome, day, hour, matr, type, ICAOModel, eng, QEng, MTOW, segm, origin, destin, operphase, precipitation, atmpres, atmpresmax, atmpresmin, temperature, winddirection, and windblast. *Classifier: random forest.* *Output: y_pred.*	
1:	*import['all necessary libraries']*
2:	*while True:*
3:	*print("y_pred", result)*
4:	*df = pd.read_csv("http://landpage-h.cgu.gov.br/dadosabertos/index.php?url=http://sistema.cenipa.aer.mil.br/cenipa/media/opendata/ocorrencia.csv")*
…	
22:	*y = df['Class']*
23:	*x = df.drop('Class', axis = 1)*
24:	*x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2)*
25:	*model = RandomForestClassifier(n_estimators=100)*
26:	*model.fit(x_train, y_train)*
27:	*importances = model.feature_importances_*
28:	*indices = np.argsort(importances)*
29:	*result = model.score(x_test, y_test)*
30:	*preview = model.predict(x_test)*
31:	*y_pred = model.predict(x)*
32:	*time.sleep(5)*

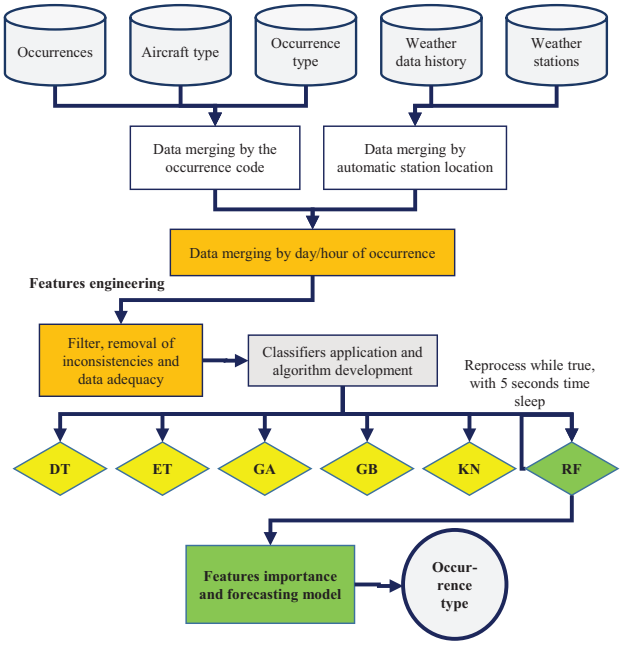

Figure 5. Main research procedures

the previous hour, and hourly air humidity were identified, but not considered for analysis as there were many empty observations in the database.

In the development of the model, after choosing the classifier with the best accuracy, reprocessing at 5-second intervals was obtained from the commands while true and sleep time, using 80% of the data for training and 20% for testing. Some of the primary codes are listed in Table 4.

Despite the codes being presented in a simple form with only 32 lines of codes in Table 4, significant efforts were required in the standardization of variables, elimination of redundancies, and removal of inconsistent data, amongst others. Figure 5 shows the main research procedure adopted in this study.

According to Figure 5, the data collection and processing involved machine learning tools, such as the assignment of automatic procedures for reading data in the CENIPA databases and meteorological stations, identified as robots for locating and downloading data (code from line 4 of Table 4, for example), and analyses in small time intervals, in this case, five seconds, to improve machine learning in the demonstration of results.

3. Results

In the analysis of the results, descriptive statistics were initially presented with some of the main characteristics of the data for all identified aeronautical occurrences. In the data analysis, only the occurrences and meteorological issues related to the reference cities of the study were considered.

Of the 6,188 aeronautical occurrences registered between January 2010 and August 2021, incidents, accidents, and serious incidents accounted for 56%, 31%, and 13 % of the occurrences, respectively (Figure 6). Among the accidents, 93% had no fatalities, however, 3.8% had one fatality, and 2% had two fatalities. The vast majority of occurrences (80%) refer to occurrences where the main aircraft is an airplane, followed by helicopter (11%), ultralight (6%), and others (3%), as shown in Figure 7. The cities of Rio de Janeiro/RJ, São Paulo/SP and Guarulhos/SP, analyzed in this study, together account for 12.86% of the total occurrences.

The aircraft registry includes aircraft without registrations, with 14 repetitions, and registrations repeated up to 10 times, as is the case with registrations PRTTK (ATR-42-500), PRTTP (B727-2M7), and PPGMA (AB-115). From the data, the models that appear the most among the occurrences are aircraft ATR-72-212A (3.31%), ERJ 190-200 IGW (2.95%), and AB-115 (2.94%). Approximately 60% of the aircraft involved in the incidents contained up to 7 seats, with the majority (17%) having 6 seats. There was no common hour for the occurrence, however, the highest frequencies, in percentage, were at 20:00:00 (2.16%), 13:00:00 (1.64%), 19:00:00 (1.63%), 13:30:00 (1.58%), and 20:30:00 (1.48%). Note that the end of the day, when the crew is most possibly fatigued, and close to midday, where there are intense thermal activities, are among the main times of occurrence.

To develop the forecast model based on occurrences and weather data, only the data referring to 796 occurrences registered in the cities of Rio de Janeiro/RJ, São Paulo/SP, and Guarulhos/SP were analyzed in the next stage of the study, which represents a combined total of 12.86%. In the analysis of data referring to the occurrences registered in these cities, the vast majority (83.44%) refers to incidents, followed by accidents (9.46%), and serious incidents (7.10%). Excluding the lines with empty fields, 786 occurrences in 22 variables were considered, totaling 17,292 observations. In the definition of the occurrence class, the forecasting was made from six different

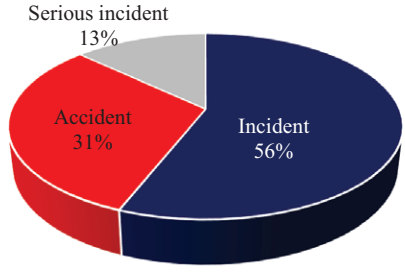

Figure 6. Total occurrences by classes

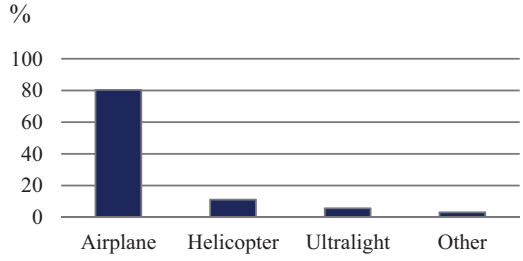

Figure 7. Occurrences by aircraft type

classifiers, for which the accuracies, in percentage, were obtained for DT (85.21%), ET (90.14%), GNB (70.25%), GB (85.91%), KN (86.07%), and RF (96.20%) classifiers.

These results, particularly the RF, are satisfactory both in relation to the problem studied, and superior compared to other studies that use similar techniques, as in Herrema et al. (2019), with an accuracy of 79%, or Rodríguez-Sanz et al. (2021), with accuracies ranging between 66 and 72% for the analysis of check-in units, and between 69 and 74% for security control units at airports. The results are also superior to the analysis of the impact of meteorological conditions on airport performance by Schultz et al. (2021), which generated an accuracy of 95.3%. The feature importance identified in this study is shown in Figure 8, in which the features considered are presented on the y-axis and the percentage of relative importance of each feature in the model composition on the x-axis.

From the analysis of the features presented in Figure 8, it is noted that the MTOW presents the highest level of importance among the features considered. In this case, the highest percentage of aircraft (approximately 10%) involved in the occurrences have MTOW above 70 tons with the ICAO models A320 and B738 being the main aircraft. This demonstrates that the highest number of occurrences is among commercial aviation aircraft, even though such aircraft represent less than 5% of all aircraft registered in Brazil – 640 of 16,213 aircraft (Caetano et al., 2022).

From the analyzed data, it is noted that even if the main segment identified among the occurrences was regular aviation, the same aircraft, registered PRSAU, or PRSAO, used for flight instructions, was identified in seven different occurrences, classified as incidents. This demonstrates the need for an intense and exhaustive series of pilot training on the ground using simulators that consider meteorological conditions (Ahmadi et al., 2022), before entering the command of the aircraft itself in a real operation. Additionally, the feature importance if the

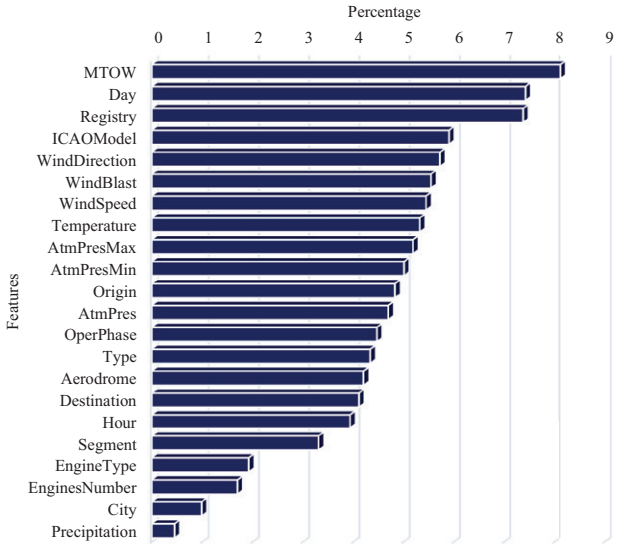

Figure 8. Feature importance

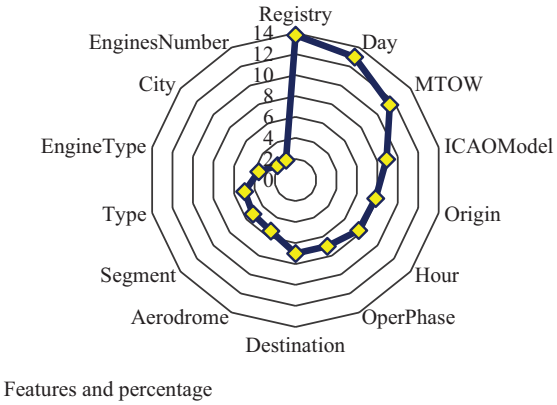

Figure 9. Feature importance of only occurrence characteristics/variables

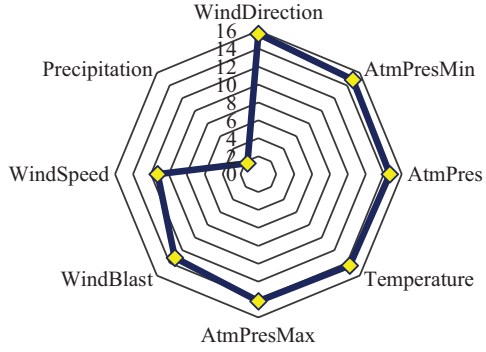

Figure 10. Feature importance of only meteorological characteristics/variables

classes are analyzed separately as a function of variables related to occurrences and meteorological conditions is presented in Figures 9 and 10, respectively.

According to Figures 9 and 10, only for the case of occurrence characteristics are there small changes in the ordering of features, especially among the first three, reaffirming the greater relative importance of these features in relation to the others. Note that Figure 10 shows that precipitation has the lowest percentage of importance in the classification of the occurrence, which contradicts the results of many studies conducted on the subject, such as those by Bandeira et al. (2018) and Pacheco Jr. et al. (2020). This makes it clear that when considering that the type of aircraft involved in the occurrence, in this case, identified mainly as high MTOW, as well as the type of operation, such as pilot training, should receive relevant attention regarding meteorological issues in the management and prevention of aeronautical occurrences. Among the variables considered, there are some that can be managed, such as profiles of aircraft operating under certain meteorological conditions and definition of an airport site in locations with less susceptibility to meteorological variables (Alves et al., 2020), such as regular relief, which can interfere with the safety of the operation due to the

movement of air masses and aircraft instability in the final approach. In addition, other theoretical contributions can be identified in collaboration with the findings of Schultz et al. (2021) by adding variables on accidents and aeronautical incidents to some of the meteorological variables analyzed by the authors. The results presented here also complement the proposal presented by Patriarca et al. (2022) for the development of solutions based on business intelligence and machine learning from real data. As there is a continuous update of data, it is possible to provide continuous and readjusted forecasts, complementing the studies by Sineglazov et al. (2013) and Stogsdill (2022), for different new circumstances involving both the characteristics of the occurrences and meteorological conditions.

Although not considered in this study, different factors may also have caused such occurrences, such as cognitive factors of the crew-stress level, situational awareness, dependencies on automated systems (Martins, 2016), operational failures or violation of procedures (Medvedev, 2013), and even other human factors related to aviation operations (Wan et al., 2021).

Conclusions

Through the use of machine learning techniques, the study demonstrated that different variables can be used in the forecasting of occurrences of accidents, incidents, or serious aeronautical incidents, with more than 96% accuracy. The main theoretical contribution of this study lies in the combination of the factors associated with the occurrences, recorded by an accident investigation agency, with the meteorological conditions identified moments before the occurrences so that, from the identified patterns, predictions from machine learning tools can be made to avoid such future occurrences.

Although meteorological-related variables are not manageable, the results of this study can be used by airport managers, airlines, and pilots, combined with meteorological forecast data, such as those made available by the European Center for Medium-Range Weather Forecasts (ECMWF), with a range of 9 km, Icosahedral Nonhydrostatic Model (ICON)/Deutscher Wetterdienst (DWD), with a range of 13 km; and Global Forecast System (GFS)/National Centers for Environmental Prediction (NCEP), with a range of 22 km, among others, in order to identify probabilities of occurrence of a certain accident or incident and, with that, take the necessary measures in advance. In addition, new technologies can be developed and incorporated into avionics to alert crew members about the possible risks associated with the safety of the operation.

Thus, this study can be used as a reference in the identification of target factors, such as aircraft models and wind direction/intensity in relation to the central axis of the runway in operation, among others, to carry out applied studies, which are also possible practical implications of the study. In addition, statistical inferences can be made with the number of flights performed involving occurrences of accidents and incidents, and flights performed safely, in such a way that new classes can be incorporated into the model, such as flights performed safely on time, out of time, more efficient routes, among other complementary studies. The results of this study provide significant guidance for the prevention and even avoidance of such occurrences.

As future research challenges, the greater agility in the operationalization of this type of data analysis demands a better standardization of the presented data from official agents, thus minimizing possible errors of tokenizing data via the automatic analysis of both occurrence and meteorological data.

Acknowledgements

Research Group in Air Transport Innovations (MTOW), Air Transport Laboratory (LABTAR) / Aeronautics Institute of Technology (ITA), National Council for Scientific and Technological Development (CNPq), Graduate Support Program (PROAP) / Coordination for the Improvement of Higher Education Personnel (CAPES) / Federal University of Goiás (UFG), Brazil, to the Aviation's Editor, editorial office, and Reviewers #1, #2, and #3, who contributed significantly to the improvement of the paper.

Data availability statement

All input data used and analyzed in the calculations are freely available, and their sources are listed below.

The dataset referring to aeronautical occurrences is available in the CENIPA (2021) repository (https://www2.fab.mil.br/cenipa/index.php/estatisticas).

The dataset referring to meteorological conditions is available in the INMET (2021) repository (https://portal.inmet.gov.br).

Finally, regarding the datasets generated as results of the analyses, if needed, the author will provide them to anyone.

Disclosure statement

The Author declare no conflict of interest.

References

Aeronautical Accident Investigation and Prevention Center. (2015). *Comando da Aeronáutica Centro de Investigação e Prevenção de Acidentes Aeronáuticos*. CENIPA. http://sistema.cenipa.aer.mil.br/cenipa/paginas/relatorios/rf/pt/SUMA_IG-035CENIPA2015_PT-YPB.pdf

Aeronautical Accident Investigation and Prevention Center. (2019). *Final Report A – 036/CENIPA/2019*. http://sistema.cenipa.aer.mil.br/cenipa/paginas/relatorios/rf/en/PROCW_03MAR2019_AC.ING..pdf

Aeronautical Accident Investigation and Prevention Center. (2013). *Final Report I – 235/CENIPA/2013*. http://sistema.cenipa.aer.mil.br/cenipa/paginas/relatorios/rf/en/RF_I-235CENIPA2013_A6EWI_Englis_Version.pdf

Aeronautical Accident Investigation and Prevention Center. (2021). *Ocorrências aeronáuticas na aviação civil brasileira*. https://www2.fab.mil.br/cenipa/index.php/estatisticas

Air Navigation Management Center. (2022) *Portal Operacional*. CGNA. http://portal.cgna.decea.mil.br/

Ahmadi, N., Romoser, M., & Salmon, C. (2022). Improving the tactical scanning of student pilots: A gaze-based training intervention for transition from visual flight into instrument meteorological conditions. *Applied Ergonomics, 100*, 103642. https://doi.org/10.1016/j.apergo.2021.103642

Alves, C. J. P., Silva, E. J., Müller, C., Borille, G. M. R., Guterres, M. X., Arraut, E. M., Peres, M. S., & Santos, R. J. (2020). Towards an objective decision-making framework for regional airport site selection. *Journal of Air Transport Management*, *89*, 101888. https://doi.org/10.1016/j.jairtraman.2020.101888

Bandeira, M. C. G. S. P., Correia, A. R., & Martins, M. R. (2018). General model analysis of aeronautical accidents involving human and organizational factors. *Journal of Air Transport Management*, *69*, 137–146. https://doi.org/10.1016/j.jairtraman.2018.01.007

Barua, L., & Zou, B. (2021). Planning maintenance and rehabilitation activities for airport pavements: A combined supervised machine learning and reinforcement learning approach. *International Journal of Transportation Science and Technology*, *11*(2), 423–435. https://doi.org/10.1016/j.ijtst.2021.05.006

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Caetano, M., Silva, E. J., Vieira, D. J., Alves, C. J. P., & Müller, C. (2022). Criteria prioritization for investment policies in General Aviation aerodromes. *Regional Science Policy & Practice*, *14*(6), 211–233. https://doi.org/10.1111/rsp3.12538

Choi, S., & Kim, Y. J. (2021). Artificial neural network models for airport capacity prediction. *Journal of Air Transport Management, 97*, 102146. https://doi.org/10.1016/j.jairtraman.2021.102146

Dalmau, R., Ballerini, F., Naessens, H., Belkoura, S., & Wangnick, S. (2021). An explainable machine learning approach to improve take-off time predictions. *Journal of Air Transport Management, 95*, 102090. https://doi.org/10.1016/j.jairtraman.2021.102090

Gosling, G. D. (1987). Identification of artificial intelligence applications in air traffic control. *Transportation Research Part A: General, 21*(1), 27–38. https://doi.org/10.1016/0191-2607(87)90021-5

Herrema, F., Curran, R., Hartjes, S., Ellejmi, M., Bancroft, S., & Schultz, M. (2019). A machine learning model to predict runway exit at Vienna airport. *Transportation Research Part E: Logistics and Transportation Review*, *131*, 329–342. https://doi.org/10.1016/j.tre.2019.10.002

International Civil Aviation Organization. (2020). *Annex 13 – Aircraft Accident and Incident Investigation* (12th ed.). ICAO.

Instituto Nacional de Meteorologia. (2021). *Histórico de dados meteorológico*. https://portal.inmet.gov.br/

Martins, A. P. G. (2016). A review of important cognitive concepts in aviation. *Aviation, 20*(2), 65–84. https://doi.org/10.3846/16487788.2016.1196559

Medvedev, A. (2013). Airplane catastrophe as a result of operational errors and violations. *Aviation, 17*(2), 70–75. https://doi.org/10.3846/16487788.2013.805866

Pacheco, Jr., G., Camargo, M., & Halawi, L. (2020). An evaluation of the operational restrictions imposed to Congonhas airport by civil aviation instruction 121-1013. *International Journal of Aviation, Aeronautics, and Aerospace*, *7*(2).

Patriarca, R., Di Gravio, G., Cioponea, R., & Licu, A. (2022). Democratizing business intelligence and machine learning for air traffic management safety. *Safety Science*, *146*, 105530. https://doi.org/10.1016/j.ssci.2021.105530

Post, J. (2021). The next generation air transportation system of the United States: Vision, accomplishments, and future directions. *Engineering, 7*(4), 427–430. https://doi.org/10.1016/j.eng.2020.05.026

Puranik, T. G., & Mavris, N. (2018). Anomaly detection in general-aviation operations using energy metrics and flight-data records. *Journal of Aerospace Information Systems*, *15*(1), 22–35. https://doi.org/10.2514/1.I010582

Puranik, T. G., & Mavris, N. (2020). Identification of instantaneous anomalies in general aviation operations using energy metrics. *Journal of Aerospace Information Systems*, *17*(1), 51–65. https://doi.org/10.2514/1.I010772

Puranik, T. G., Rodriguez, N., & Mavris, N. (2020). Towards online prediction of safety-critical landing metrics in aviation using supervised machine learning. *Transportation Research Part C: Emerging Technologies, 120*, 102819. https://doi.org/10.1016/j.trc.2020.102819

Rodríguez-Sanz, A., Comendador, F. G., Valdés, R. A., Pérez-Castán, J., Montes, R. B., & Serrano, S. C. (2019). Assessment of airport arrival congestion and delay: Prediction and reliability. *Transportation Research Part C: Emerging Technologies*, *98*, 255–283. https://doi.org/10.1016/j.trc.2018.11.015

Rodríguez-Sanz, A., Marcos, A. F., Pérez-Castán, J. A., Comendador, F. G., Valdés, R. A., & Loreiro, A. P. (2021). Queue behavioural patterns for passengers at airport terminals: A machine learning approach. *Journal of Air Transport Management, 90*, 101940. https://doi.org/10.1016/j.jairtraman.2020.101940

Schultz, M., Reitmann, S., & Alam, S. (2021). Predictive classification and understanding of weather impact on airport performance through machine learning. *Transportation Research Part C: Emerging Technologies*, *131*, 103119. https://doi.org/10.1016/j.trc.2021.103119

Sineglazov, V., Chumachenko, E., & Gorbatyuk, V. (2013). An algorithm for solving the problem of forecasting. *Aviation*, *17*(1), 9–13. https://doi.org/10.3846/16487788.2013.777219

Stogsdill, M. (2022). When outcomes are not enough: An examination of abductive and deductive logical approaches to risk analysis in aviation. *Risk Analysis, 42*(8), 1806–1814. https://doi.org/10.1111/risa.13681

Stogsdill, M., Baranzini, D., & Ulfvengren, P. (2021). Development of a metric concept that differentiates between normal and abnormal operational aviation data. *Risk Analysis*, *42*(8), 1815–1833. https://doi.org/10.1111/risa.13680

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. The MIT Press.

Truong, D., & Choi, W. (2020). Using machine learning algorithms to predict the risk of small Unmanned Aircraft System violations in the National Airspace System. *Journal of Air Transport Management, 86*, 101822. https://doi.org/10.1016/j.jairtraman.2020.101822

Xu, Z., Saleh, J. H., & Subagia, R. (2020). Machine learning for helicopter accident analysis using supervised classification: Inference, prediction, and implications. *Reliability Engineering and System Safety, 204*, 107210. https://doi.org/10.1016/j.ress.2020.107210

Wan, M., Liang, Y., Yan, L., & Zhou, T. (2021). Bibliometric analysis of human factors in aviation accident using MKD. *IET Image Processing*, 1–9. https://doi.org/10.1049/ipr2.12167