Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/84/2025.21196

Research on Data Analysis of Factors Influencing Air Crashes Based on Machine Learning

Yuchen Zhang[1,a,*]

[1]*The High School Affiliated to Renmin University of China, Beijing, 100000, China*
a. zhangmiqiu@126.com
**corresponding author*

Abstract: Accompanying the establishment of intimate economic and cultural connections between different countries and regions, people are required to travel through different places rapidly, which fosters the prosperity of the aviation industry and makes airplanes a crucial means of transportation. In 2024, however, plane crashes such as Azerbaijan Airlines Flight 8243 and Jeju Air Flight 2216 posed people's concerns about the security of airplanes and the effectiveness of aviation systems. To that end, this study focuses on factors that influence plane crashes' mortality and injury rate in order to find the most influential factor associated with plane crashes, aiming to be referenced to make rules protecting passengers' safety during flights. At the same time, analyzing data on plane crashes can also avoid people's unnecessary worries about plane crashes. This study uses random forest machine learning model to analyze plane crash data, which can effectively find the most important factor that will influence mortality and injury rates.

Keywords: Machine Learning, Air Crash, Random Forest Machine Learning Model, Mortality and Injury Rate

1. Introduction

Air crashes refer to events in which aircraft crash due to malfunctions, natural disasters, or unexpected events during flights, resulting in property damage and casualties. Recently, due to the number of air crashes being on the rise, air crash research has been achieved a lot. In 1998, some scientists analyzed factors that influence pilots' fatality during plane crashes, concluding that using lap and shoulder restraints during air crashes reduced pilots' risks of being killed during plane crashes [1]. In 2015, *TIME* raised another research about the mortality of passengers in different seats during air crashes from 1985 to 2000. It suggested that the mortality rate in the back three rows of seats on the plane was the lowest, at 32%, lower than the mortality rate in the middle three rows, which was 39%, and the mortality rate in the front three rows, which was 38%. At the same time, the middle seat at the rear of the aircraft has the best effect (28% mortality rate), and the seat with the worst situation is located in the middle one-third aisle of the cabin (with a mortality rate of 44%) [2]. Later, in 2017, a passage showed that the main reasons for plane crashes include communication issues between crew members and air traffic controllers. However, there is still a lack of a whole-sided analysis and investigation of factors that may influence passengers' injury rates and mortality [3]. Therefore, this study analyzed numerous factors that may affect passenger mortality and injury rates in air crashes, including crash altitude, pilot experience, weather conditions, and aircraft conditions. This study used

DOI: 10.54254/2753-8818/84/2025.21196

linear regression to conduct regression analysis on the above factors and employed a random forest machine learning model to identify the factors that have the greatest impact on passenger mortality and injury rates in air crashes. This study can provide targeted improvements to future air disaster prevention measures to reduce the losses caused by air disasters, which will contribute to future research in related fields.

2. Factors influencing air accidents

When an accident occurs, the aircraft may be in different situations, and the pilot's ability to handle abnormal situations may also vary. Therefore, these accidents will lead to different consequences, namely differences in mortality and injury rates. In this study, SRS was used to randomly select 101 flights with accidents from 1987 to 2023 on the National Transportation Safety Board (NTSB) website in the United States, and a dataset was established to collect mortality and injury rates for each flight. The mortality rate is the total number of deaths on the flight divided by the total number of passengers, while the injury rate is the sum of the number of injuries and deaths on the flight divided by the total number of passengers. At the same time, this study investigated factors that may affect mortality and injury rates, including flight altitude (feet) at the time of the accident, captain's age, captain's total flight time (hours), aircraft type captain's flight time (hours), first officer's age, first officer's total flight time (hours), aircraft type first officer's flight time (hours), aircraft age, visibility (feet), and wind speed (knots). The missing data in the dataset will be replaced with the average value of the data.

The different flight altitudes at the time of the incident will result in varying lengths of time for pilots to handle the accident, and the magnitude of injuries caused by flight accidents will also differ, thus affecting the mortality and injury rates of the flight.

The age, flight duration, and other data of the captain and first officer reflect their familiarity with the aircraft and their experience and reaction ability in handling emergencies. The differences in these abilities can also affect the size of flight accidents, resulting in different mortality and injury rates.

The age of an aircraft reflects its durability, while newer aircraft may not experience wear and tear and have more advanced systems that are easier to operate. However, there may also be situations where the system is immature and leads to human-machine confrontation. Therefore, different ages of aircraft have varying degrees of impact on the size and severity of unexpected situations that may occur, resulting in different mortality and injury rates caused by accidents.

Visibility and wind speed represent the weather conditions at the time of an accident, and different weather conditions make it difficult for pilots to handle accidents. More severe weather conditions can make the situation more urgent and increase the difficulty for pilots to handle. Therefore, the mortality and injury rates caused by accidents vary under different weather conditions.

3. Methodology

All data for this survey comes from official investigation reports released by the National Transportation Safety Board (NTSB) in the United States. The SRS method was used to extract investigation reports from 101 air accidents and regression and random forest analysis were performed to identify factors related to air crash mortality and injury rates.

In this survey, the impact of various factors on air crash mortality and injury rates was first identified through scatter plots(see Figure 1-20). It was found that flight altitude, pilot age and flight duration, and wind speed were all negatively correlated with air crash mortality and injury rates, while aircraft age and visibility, as well as air crash mortality and injury rates, were positively correlated.

DOI: 10.54254/2753-8818/84/2025.21196

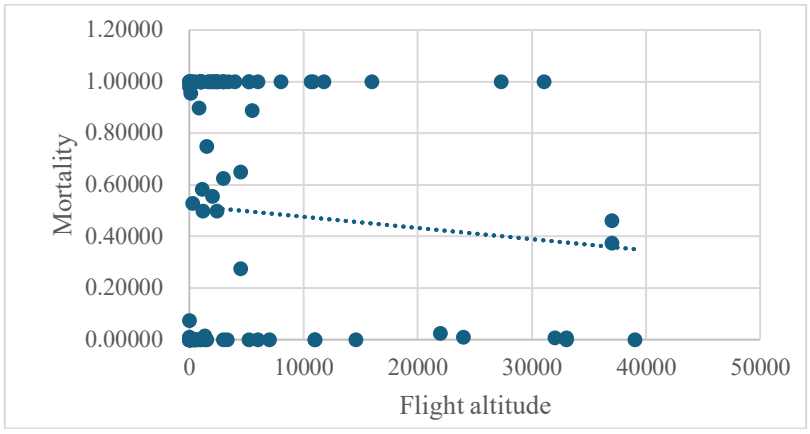

Figure 1: Relationship between air crash mortality rate and aircraft flight altitude

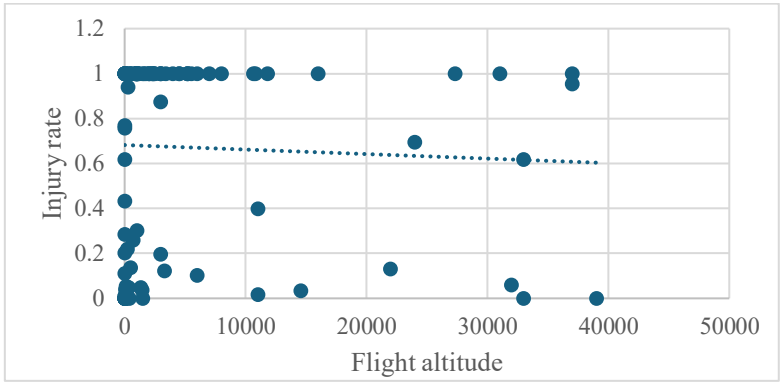

Figure 2: Relationship between air crash injury rate and aircraft flight altitude

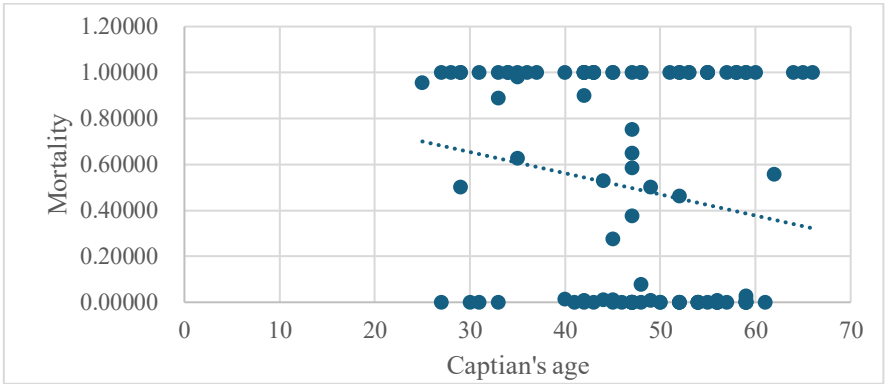

Figure 3: Relationship between mortality rate and captain age

DOI: 10.54254/2753-8818/84/2025.21196

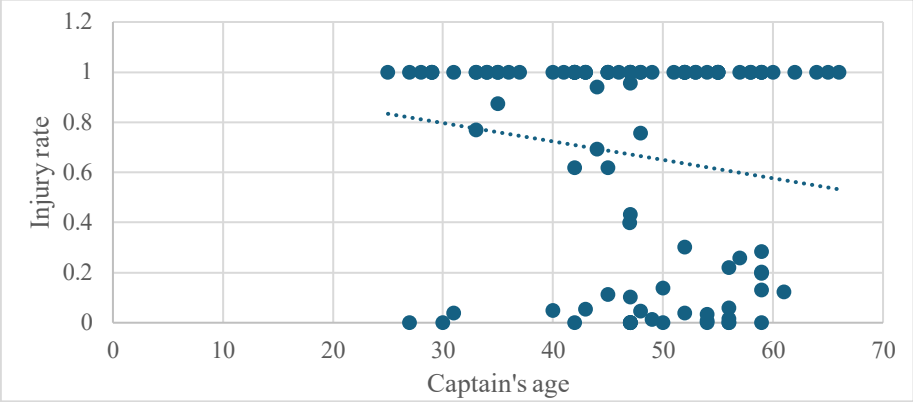

Figure 4: The relationship between injury rate and captain's age

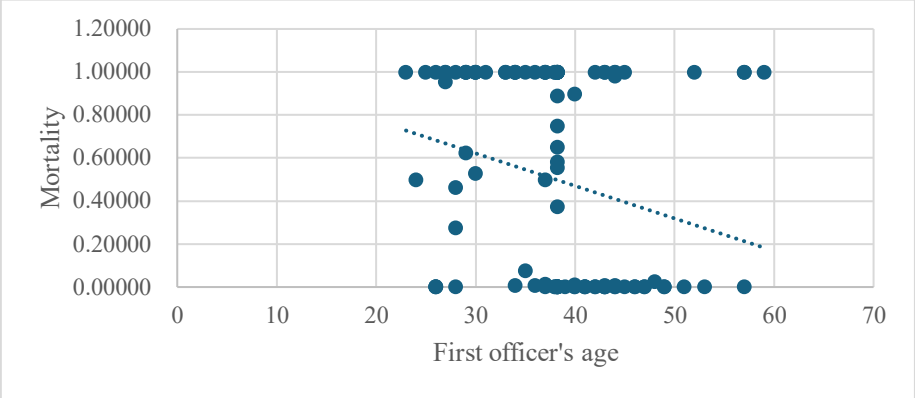

Figure 5: The relationship between mortality rate and the age of the first officer

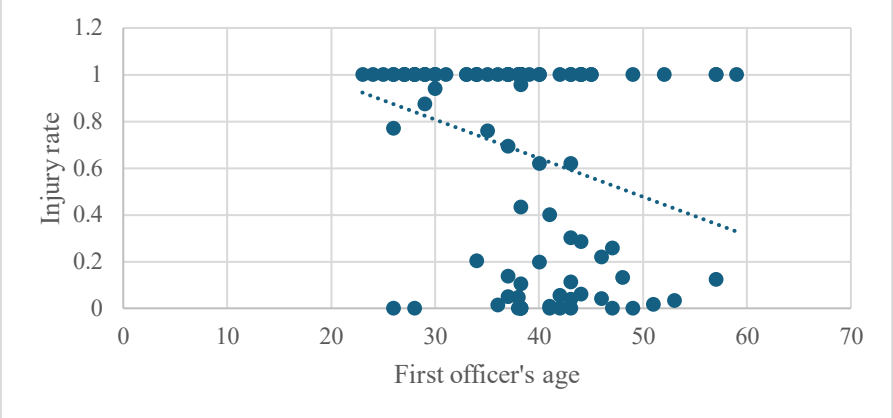

Figure 6: The relationship between injury rate and the age of the first officer

DOI: 10.54254/2753-8818/84/2025.21196

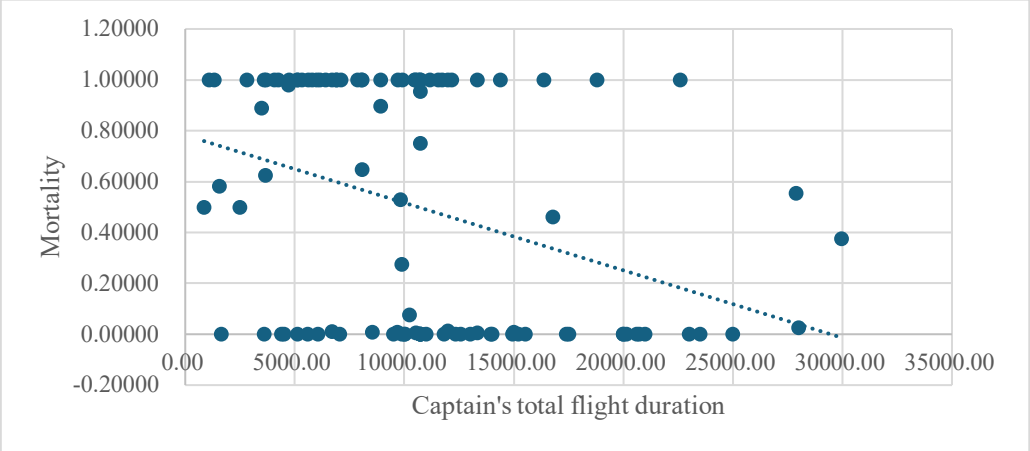

Figure 7: The relationship between mortality rate and total flight time of the captain

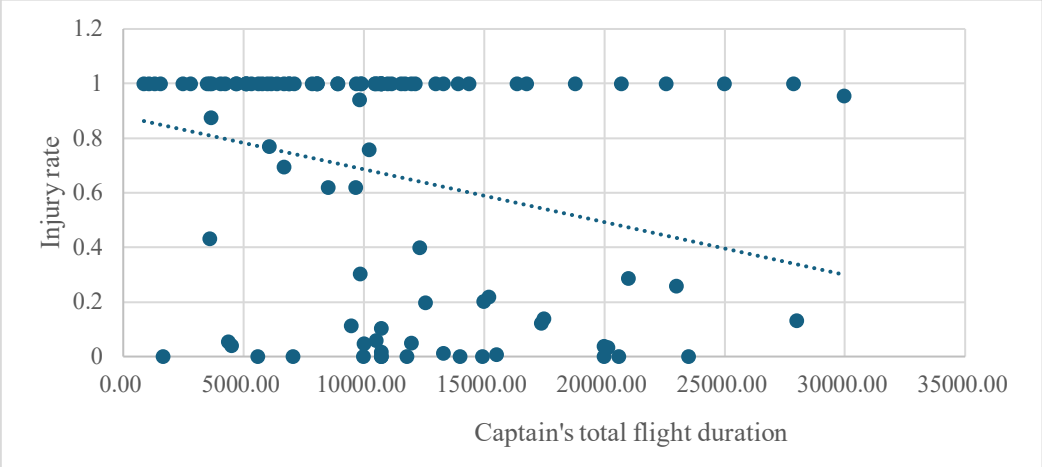

Figure 8: The relationship between injury rate and total flight duration of the captain

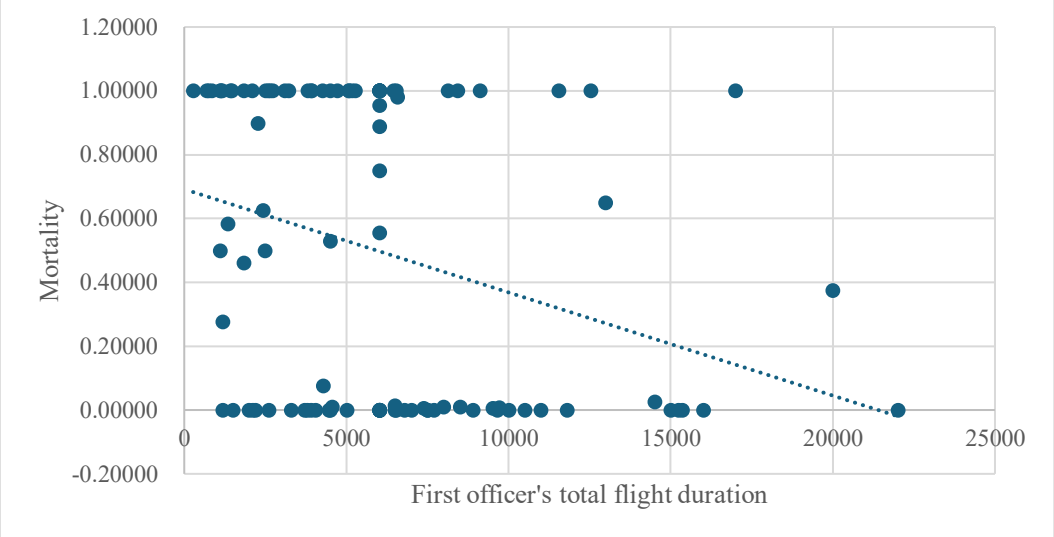

Figure 9: The relationship between mortality rate and total flight time of first officer

DOI: 10.54254/2753-8818/84/2025.21196

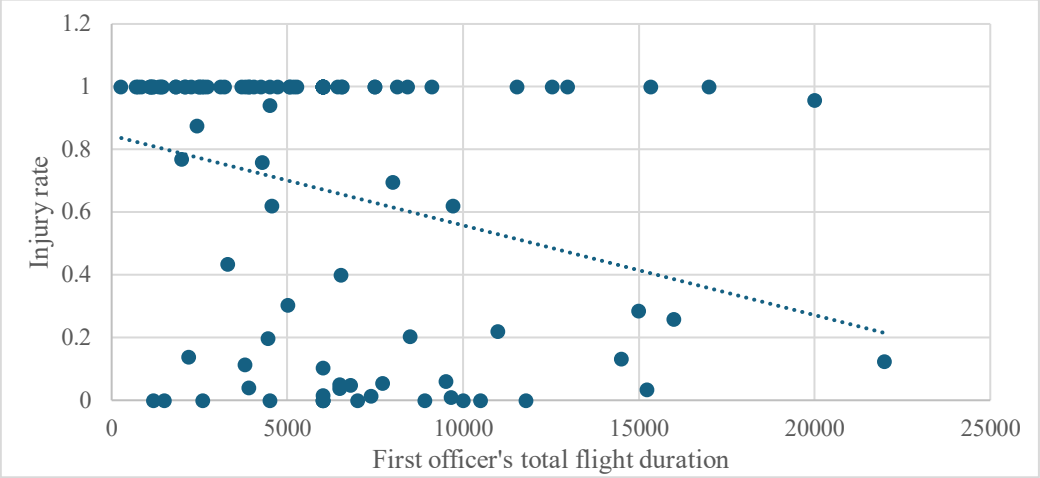

Figure 10: The relationship between injury rate and total flight time of first officer

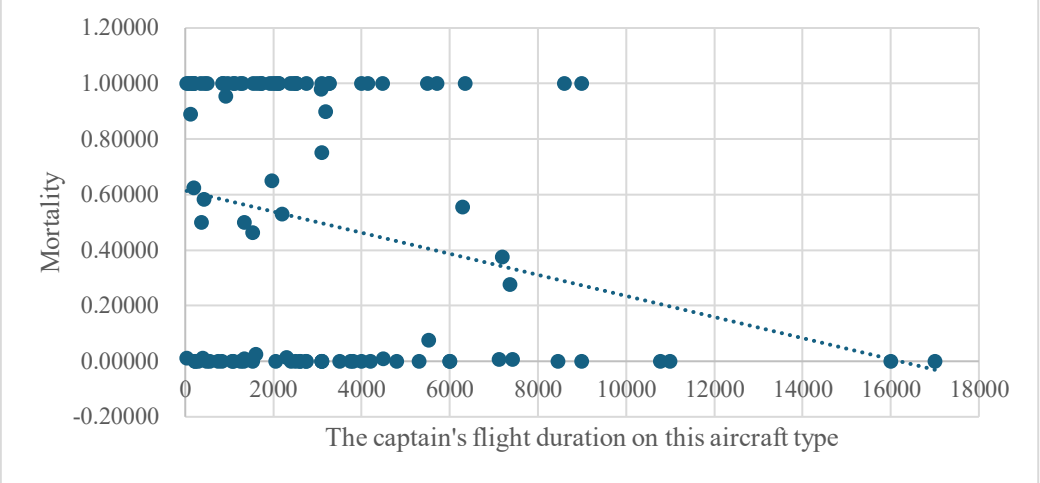

Figure 11: the relationship between mortality rate and captain's flight duration of this aircraft type

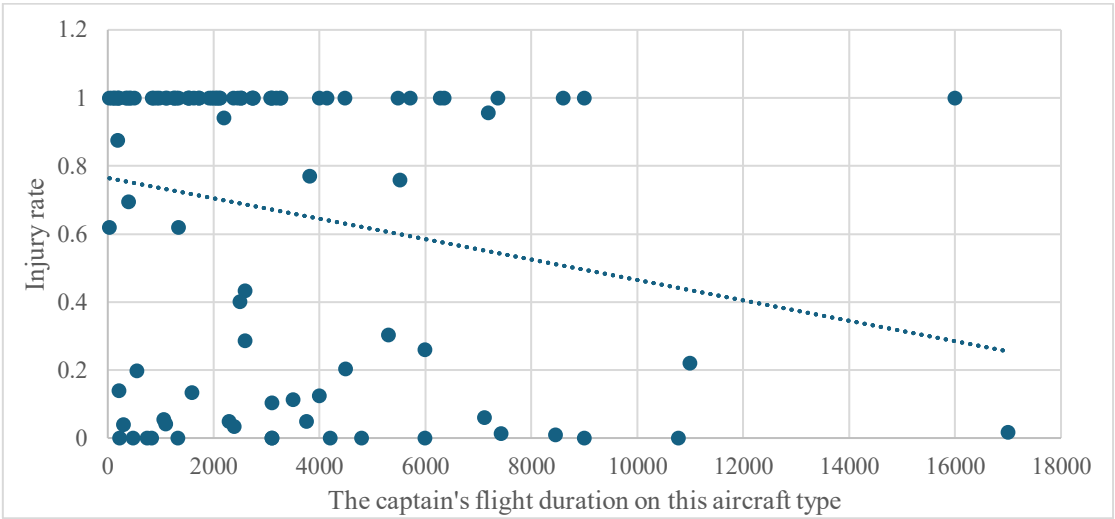

Figure 12: the relationship between Injury rate and captain's flight duration of this aircraft type

DOI: 10.54254/2753-8818/84/2025.21196

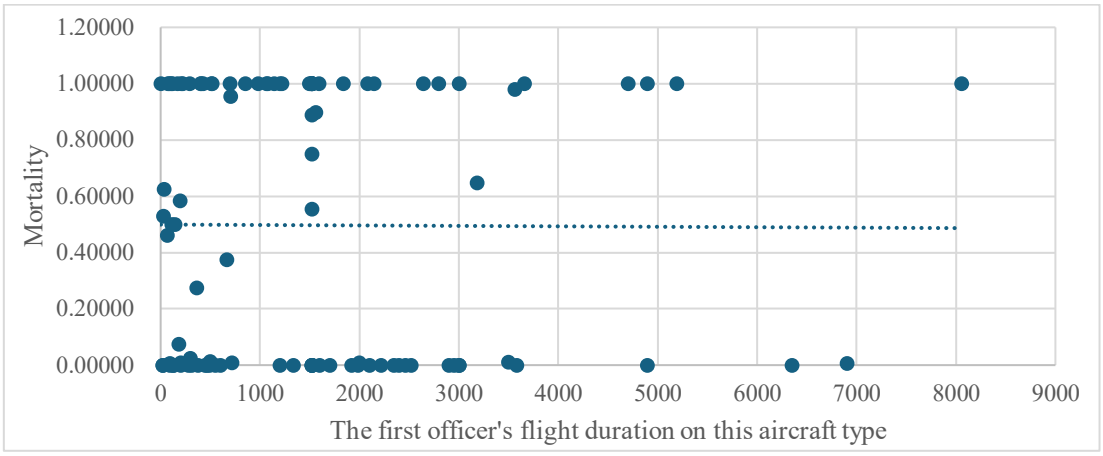

Figure 13: the relationship between mortality rate and first officer's flight duration of this aircraft type

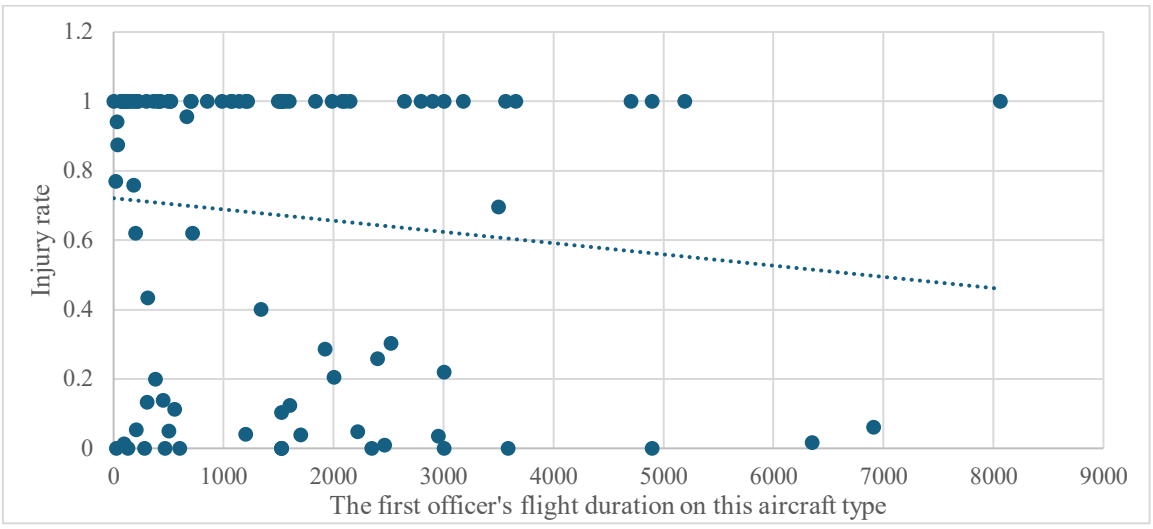

Figure 14: the relationship between Injury rate and first officer's flight duration of this aircraft type

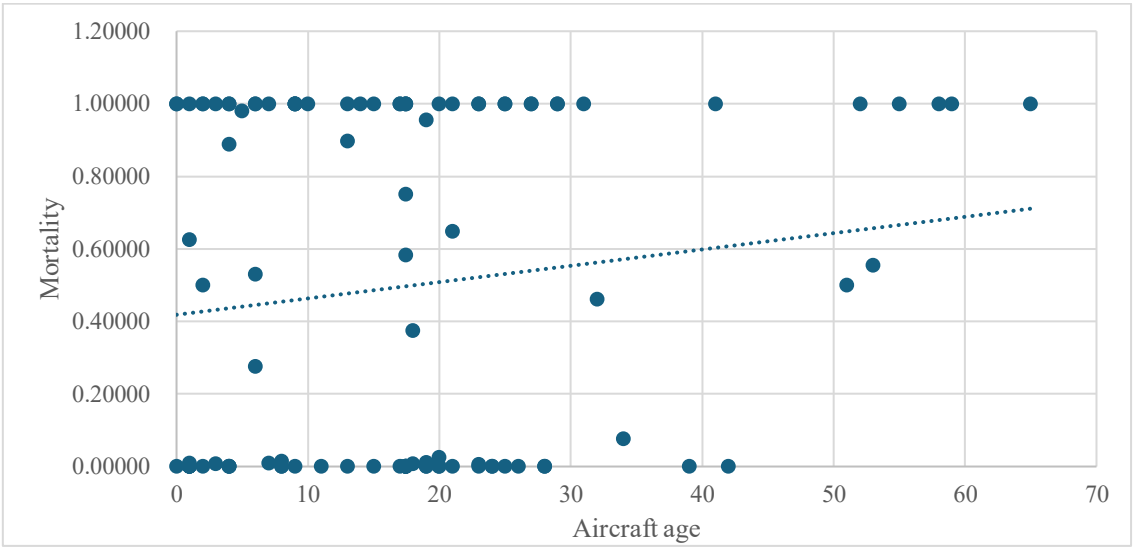

Figure 15: Relationship between mortality rate and aircraft age

DOI: 10.54254/2753-8818/84/2025.21196

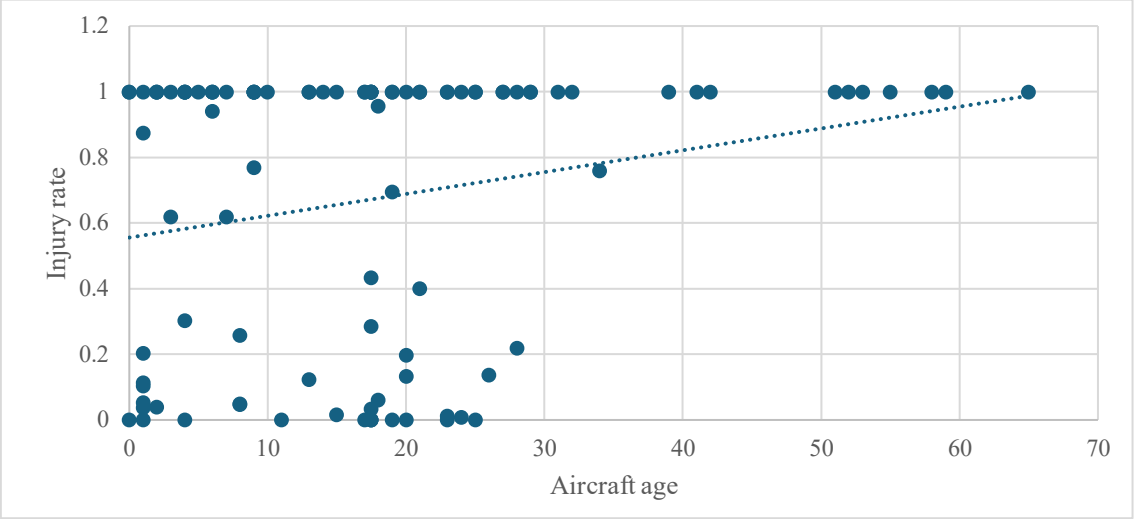

Figure 16: Relationship between injury rate and aircraft age

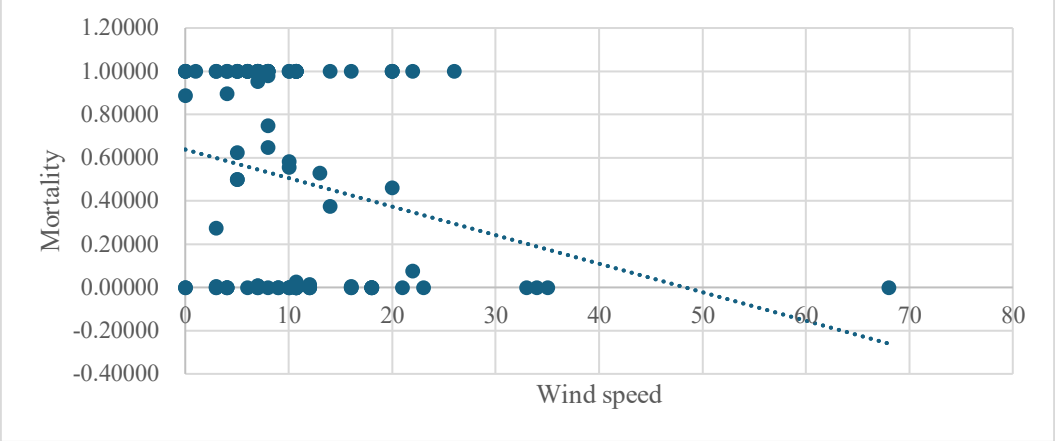

Figure 17: Relationship between mortality rate and wind speed

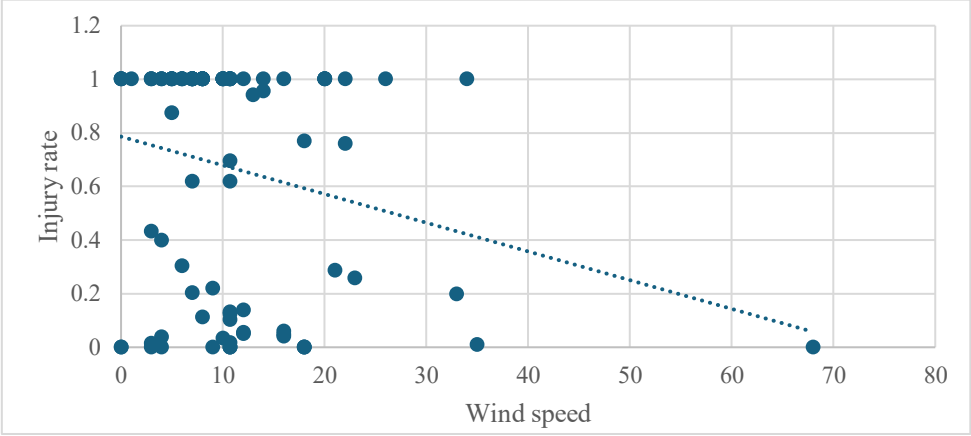

Figure 18: Relationship between injury rate and wind speed

It should be noted that as visibility increases, the mortality and injury rates of air accidents actually increase. This may be due to the fact that there is almost no difference in flight when visibility is 10

DOI: 10.54254/2753-8818/84/2025.21196

or above, and only when visibility is very low will it affect the aircraft. Therefore, this upward trend does not indicate the relationship between visibility and air crash mortality and injury rates.

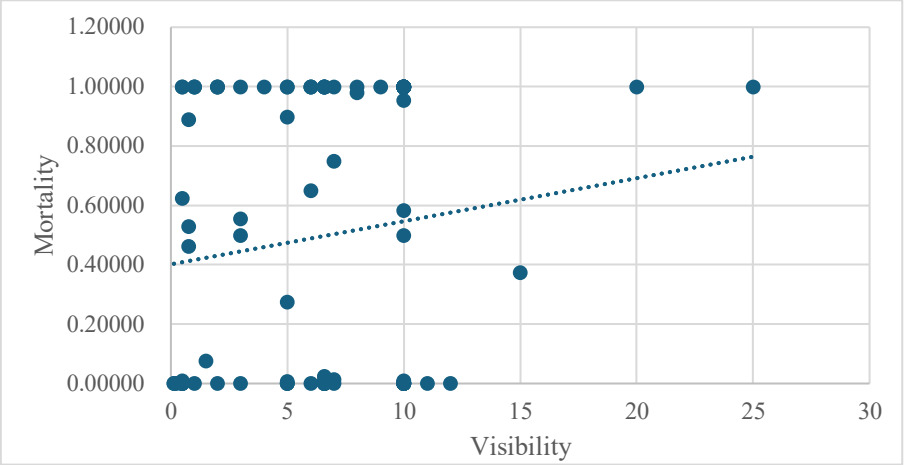

Figure 19: Relationship between injury rate and visibility

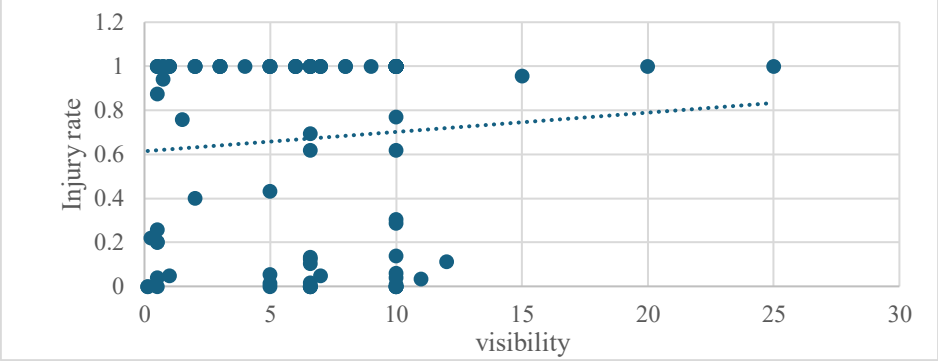

Figure 20: Relationship between injury rate and visibility

Then, this study analyzed air crash data using a random forest machine learning model to identify the importance of influencing factors related to air crash mortality and injury rates. Random forest is a powerful machine learning algorithm that improves the accuracy and robustness of a model by constructing multiple decision trees and combining their predictions. By randomly selecting features and samples, random forests can effectively process high-dimensional data and reduce the model's sensitivity to noise [4]. In addition, it can also provide internal estimates about model performance and feature importance, which is very helpful for model tuning and interpreting model predictions [5].

Finally, this study established a linear regression model on the influencing factors of air accidents and the mortality and injury rates of air accidents, which is beneficial for predicting the mortality and injury rates of air accidents in relevant situations.

4. Data modeling

4.1. Random forest

In the random forest model, we first used the pandas database to extract relevant data from our self-built dataset, where the independent variables included aircraft altitude (feet), pilot age, total flight hours of the pilot (hours), flight hours of the pilot on this aircraft type (hours), first officer age, total

DOI: 10.54254/2753-8818/84/2025.21196

flight hours of the first officer (hours), flight hours of the first officer on this aircraft type (hours), aircraft age (as of the accident date), visibility (miles), and wind speed (knots), while the dependent variable was the mortality and injury rate of air crashes. By using the train_test_split function, we randomly divided the dataset into training and testing sets, with the testing set accounting for 40%, and the random seed was set to 42 to ensure the reproducibility of the results. Through GridSearchCV for hyperparameter tuning, we tested different parameter combinations, including:

- The number of decision trees, with options including 25, 50, 75, 100, 125, 150, 175, 200.
- The maximum depth of the tree, with options including None (unlimited) and integers from 5 to 30.
- The minimum number of samples required to split an internal node, with options including 2, 5, 10.
- The minimum number of samples required at a leaf node, with options including 1, 2, 4.

In the process of model selection, we utilized 5-fold cross-validation and negative mean squared error as the scoring criterion. Through grid search, we obtained the best parameter combinations and the corresponding best models. For the mortality rate, the best parameters were: `max_depth` set to None, `min_samples_leaf` set to 1, `min_samples_split` set to 5, and `n_estimators` set to 50. For the injury rate, the best parameters were: `max_depth` set to None, `min_samples_leaf` set to 4, `min_samples_split` set to 10, and `n_estimators` set to 200. We then used the best models to make predictions on the test set and evaluated the model performance using mean squared error (MSE). For the mortality rate, the MSE was 0.22, and for the injury rate, the MSE was also 0.22.

Finally, we identified the importance of each factor in air crash mortality and injury rates. For the mortality rate, the factor importances were as follows: Altitude (feet) at 0.3475, Pilot age at 0.0965, Total flight hours of the pilot (hours) at 0.1210, Flight hours of the pilot on this aircraft type (hours) at 0.0549, First officer age at 0.0632, Total flight hours of the first officer (hours) at 0.0986, Flight hours of the first officer on this aircraft type (hours) at 0.1004, Aircraft age (as of the accident date) at 0.0303, Visibility (miles) at 0.0349, and Wind speed (knots) at 0.0527. For the injury rate, the factor importances were: Altitude (feet) at 0.6092, Pilot age at 0.0088, Total flight hours of the pilot (hours) at 0.0557, Flight hours of the pilot on this aircraft type (hours) at 0.0487, First officer age at 0.0317, Total flight hours of the first officer (hours) at 0.1019, Flight hours of the first officer on this aircraft type (hours) at 0.0847, Aircraft age (as of the accident date) at 0.0346, Visibility (miles) at 0.0117, and Wind speed (knots) at 0.0130.

4.2. Regression line

In the regression model, we first used the pandas database to extract relevant data from our self-built dataset, where the independent variables included aircraft altitude (feet), pilot age, total flight hours of the pilot (hours), flight hours of the pilot on this aircraft type (hours), first officer age, total flight hours of the first officer (hours), flight hours of the first officer on this aircraft type (hours), aircraft age (as of the accident date), visibility (miles), and wind speed (knots), while the dependent variable was the mortality and injury rate of air crashes. We created a Linear Regression object and used the training dataset (features and target variables) to fit this linear regression model. To assess the model's performance, researchers calculated the coefficient of determination (R^2 score) for the model on the training data. This statistical measure indicates how well the model fits the data, with values closer to 1 signifying a better fit. Specifically, the R^2 score for the mortality rate was 0.2838, and for the injury rate, it was 0.2583.

Additionally, the coefficients and intercepts of the linear regression model were extracted. These values describe the relationship between the independent variables and the dependent variable. For the mortality rate, the coefficients were as follows: [1.44377558e-06, -2.60914016e-04, -9.00690496e-06, -3.41774429e-05, -9.12491606e-03, -2.03363468e-05, 6.69021160e-05,

DOI: 10.54254/2753-8818/84/2025.21196

6.05035944e-03, 5.55255676e-03, -1.12379643e-02]. The intercept was 1.0438.For the injury rate, the coefficients were: [5.27209312e-06, -2.54350906e-04, -6.11392347e-06, -2.18760697e-05, -1.16247572e-02, -1.33434173e-05, 1.69910447e-05, 7.97224743e-03, 2.46380127e-03, -9.79986272e-03]. The intercept was 1.2343.5.

5. Conclusion

Through random forests, we have identified that flight altitude is the most significant factor in both the mortality and injury rates of airplane accidents. For mortality rates following airplane accidents, the ages of the two pilots, total flight hours, flight hours on the accident aircraft model, and wind speed all have a certain impact. Regarding injury rates after airplane accidents, the total flight hours of the two pilots and flight hours on the accident aircraft model also have an impact. Therefore, it is recommended that airplanes take protective measures against accidents, especially during takeoff and landing when flight altitude is lower. Pilots should enhance their ability to handle emergencies at low altitudes. Additionally, pilots should avoid switching between different aircraft models and instead focus on one model to increase familiarity with the aircraft, which can enhance their ability to respond to emergencies and reduce the losses caused by accidents.

References

[1] *Rostykus, P. S., Cummings, P., & Mueller, B. A. (1998). Risk factors for pilot fatalities in general aviation airplane crash landings. JAMA, 280(11), 997-999.*
[2] *Emily Barone. This Is the Safest Place to Sit on a Plane, TIME, 2015, https://time.com/3934663/safest-seat-airplane/.*
[3] *Enomoto, C. E., & Geisler, K. R. (2017). Culture and plane crashes: A cross-country test of the Gladwell hypothesis. Economics & Sociology, 10(3), 281-293.*
[4] *Y. Freund & R. Schapire, Machine Learning: Proceedings of the Thirteenth International conference, 148-156*
[5] *Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.*