Air Crash Data Analysis and Prediction

Nandana Manoj[#1], Preethika Ajay Kumar[#2] , R Harshita[#3]

Department of CSE, PES University

[1]nandana.v.manoj@gmail.com

[2]preethikaajay@gmail.com

[3]harshitar1@gmail.com

***Abstract*— Since the beginning of air travel, aviation accidents have always been a huge cause of concern for people. Crashes could occur due to multiple factors like mechanical failure, human error, geographical location etc. Thus, identifying the factors that could have influenced the crash becomes vital in averting such occurrences in the future. In this project, we aim to analyze the data from multiple air crashes to understand the possible factors that could have impacted them. The factors could change with each accident, so we propose a model that can correctly identify the factors that could have caused a crash based on the data provided.**

***Keywords*— Crash, factors, influence, model, analyze**

Github Repository Link:
https://github.com/preethika-ajay/UE20CS312-DA-PROJECT

I. Introduction

The aircraft industry is a multimillion dollar industry, with thousands of planes flying through the skies everyday. Aircrafts are used to transport various good, people across different destinations. Over the years, the number of airplanes taking off also have steadily increased. According to the FAA, around 45,000 flights are handled by them on a daily basis. The sheer number of planes flying showcase the importance they bring to the economy. Hence, any crash can bring about significant financial and human loss.

Airline safety has been the major concern of engineers in the past years. Safety has significantly improved over time, gradually resulting in less disasters. Between 2016 and 2020, there has been only 1.38 or 1 accident for every 0.75 million flights. According to the National Transport Safety Board, the fatal accident rate for U.S. passenger and cargo airlines has declined significantly since 2001, down to 0.006 accidents per 100,000 flight hours. Even though airline safety has improved remarkably, air crashes still possess risks.

Air crashes can be due to a multitude of reasons, few of them being mechanical failure, air collision, ground collision, bird strike, pilot error, etc. On analysis of the past trends of these crashes, such mechanical failures can be prevented or pilots can be given better training to handle a larger variety of situations. For example, the Tenerife Airport Disaster was a major crash that occurred in 1977 involving two aircrafts - KLM Flight 4805 and Pan Am Flight 1736. The two aircrafts collided on the runway as Flight 4808 initiated takeoff while Flight 1736 was still on the runway. After investigation, it was speculated that one major cause of the accident was lack of proper communication between the pilots. An aftermath of this accident saw the aviation board enforce standard phrases for acknowledgement among pilots, and an emphasis of English as the working language.

It can be seen that analysis of probable causes of air crashes, can help rectify problems in the industry and make aviation safer still. Taking all these factors into consideration, this project looks to analyze various factors of air crashes based on location, type of flight, phase of flight, aircraft type, etc. and infer the correlation between them. The project also involves prediction of air crashes based on various factors. In doing so, a better assessment of the risk and safety standards of the aviation industry can be understood.

II. Related Work

In this section, we summarize the research done on similar problem statements in other research papers. A summary of the paper, method to build the problem statement and results are mentioned

below. Further, future scope of the work has also been talked about.

Researchers at Federal University of Technology[1] have compared the air crashes that happened in every decade from 1920 to 2011. The dataset used has been resourced from multiple international bodies for the years 1920 to 2011. Air crashes are grouped by decade, and the major causes of the crash in each decade has been inferred through data visualization. The paper goes on to analyze the distribution of air crashes based on its location due to reasons such as hijacking, sabotage, lightning strike, etc.

Researchers in Vietnam have made use of IBM Watson and Cognac Analytics to determine airline crash causes[2]. The main aim of the research was to find the cause of a crash, analyze crash patterns from data throughout the world, and find solutions to prevent such crashes. The dataset used contained around 73,000 data points from two different data sources. One dataset that collects major accidents data—more specifically commercial airplane and military aircraft crashes—has 5,514 data points from 17 September 1908 to 06 November 2018 , while the other dataset that collects minor accidents data has 67,640 data points from 01 January 1987 to 31 December 2019 . The paper is divided into two parts - analysis of major accidents and analysis of minor accidents. Data visualization based on location, number of fatalities, engine type, aircraft type, yearly, monthly and daily trend is done for both major and minor crashes separately. A limitation of this project is that a large amount of data is missing from the dataset. Also, the dataset contained typographical errors.

Researchers at the Institute of Technology Blanchardstown, Ireland have performed analysis on the same dataset. They have implemented the CRISP-DM methodology to create a classification model. The maximization of relevance and the minimization of redundancy are considered for feature selection. Various feature selection methods were performed and Principal Component Analysis was found to be the most effective. On the creation of the correlation matrix, no significant correlation was found between the attributes. The algorithms that were used as part of the modeling technique were K-NN, Naïve Bayes, Gradient Boosted Trees and Deep Learning. The analysis performed was ultimately inconclusive with poor accuracy. Also, the number of attributes used to generate the most accurate results was not constant.

Researchers at the Sardar Patel Institute of Technology, India have performed Airplane Crash Severity Prediction using Machine Learning. The same dataset has been used. The data preprocessing techniques have been described in detail. Various machine learning algorithms like Support Vector Machine, Random Forest Classifier, Artificial Neural Network, etc., have been used to determine the severity of an air crash. The preprocessing techniques used in this paper have been referred to for the purpose of this project.

Researchers at Amity University have conducted a study on Prediction of Aviation Accidents using Logistic Regression Model. In this type of model, the outcome variable is categorical and can be used to predict binary outcomes. This model basically predicts whether the flights are accident prone or not and under what conditions the accident will occur. This model provides satisfactory results with good accuracy.

III. Problem Statement

This project aims to create a model that can detect the causes of an air crash. We also plan to predict the possibility of occurrence of an air crash based on the different parameters. A dataset containing multiple features is analyzed to identify the attributes that are the most influential causes of an air crash and this is reported.

A. Dataset

The dataset used is from the US National Transportation Safety Board. The data is a collation of air crashes from 2010-2019. The dataset has multiple sheets with different sets of parameters that can be useful for data analysis and prediction.

Datasets from other sources have been referred to and could be used to train and test the model.

IV. Exploratory Data Analysis and Data Visualization

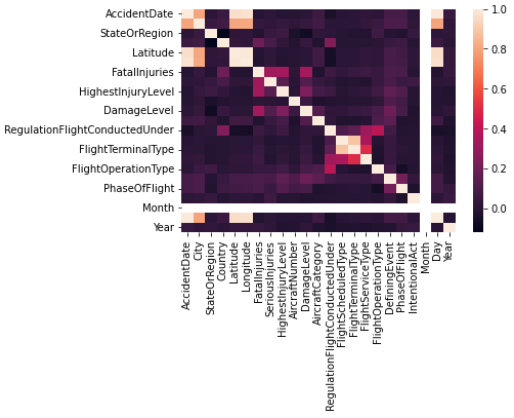
On performing exploratory data analysis, it was found that the dataset contains 22 columns. Out of

the 22 columns, 4 of them are of type float64, 1 of int64 and the rest 17 of type object.

Few attributes such as AcciddentReport and NTSBNumber were considered invalid for analysis and hence removed. On checking for duplicate tuples, it was found that the dataset does not contain any duplicate values. On searching for null values the following attributes were found to have null values. The column on the right indicates the number of null values found.

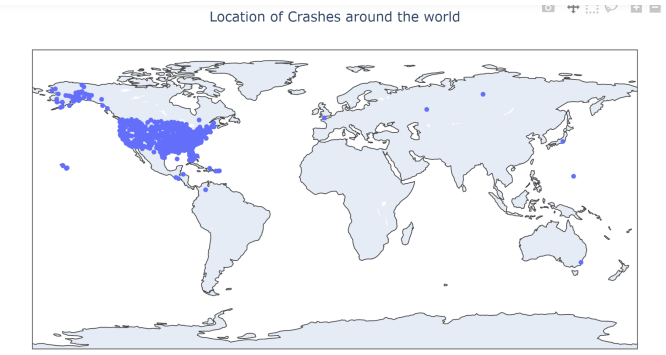
StateOrRegion	20
FatalInjuries	945
SeriousInjuries	1003
FlightScheduledType	1229
FlightTerminalType	1230
FlightServiceType	1233
FlightOperationType	10
DefiningEvent	4
IntentionalAct	1232

On plotting the correlation matrix for the datasheet containing all the attributes, one of the data sheets, it was inferred that no valid correlation exists among the variables.

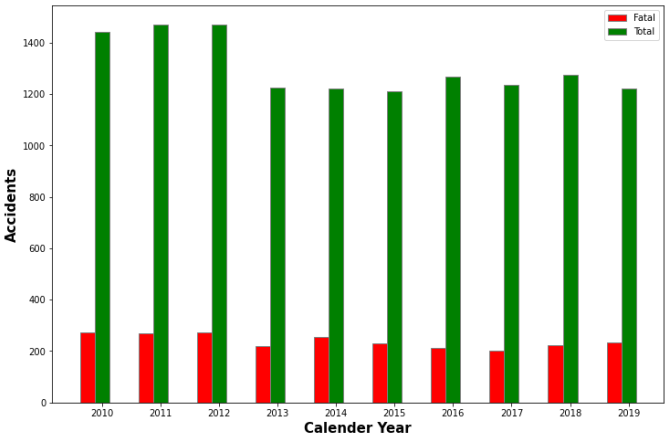


The Dataset contains multiple sheets. Although the first sheet of the dataset does not have any correalatble features, the proceeding sheets have attributes that show strong correlation.

Graphical visualizations for few attributes were done. One of the visualizations plotted was to understand the geographical location of crashes. The below plot shows the various locations of crashes in the year 2019.



On plotting a graph to visualize the total number of accidents along with accidents that were classified as fatal, the following graph was obtained.



V. Proposed Methodology

The proposed method for carrying on this project is to further analyze the data, so significant features can be gathered, and correlated attributes can be visualized. Further, an intended approach to predicting severity of a crash is using a logistic regression model. Another approach is to use a

random classifier on the dataset. Further models will be built and their accuracies will be tested on the dataset. Results from the different models will be compared to understand the best model for this problem statement.

References

[1] Stephens, Mobolaji & Ukpere, Wilfred. (2014). An Empirical Analysis of the Causes of Air Crashes from a Transport Management Perspective. Mediterranean Journal of Social Sciences. 5. 699. 10.5901/mjss.2014.v5n2p699.

[2] Taiyo Miyamoto, Neil Whitehead, and Emanuel Santos. 2020. Investigating Airplane Crash Data with Watson Analytics and Cognos Analytics. In 2020 the 3rd International Conference on Computing and Big Data (ICCBD '20). Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3418688.3418689

[3] P. Mathur, S. K. Khatri and M. Sharma, "Prediction of aviation accidents using logistic regression model," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), 2017, pp. 725-728, doi: 10.1109/ICTUS.2017.8286102.

[4] J. Mehta, V. Vatsaraj, J. Shah and A. Godbole, "Airplane Crash Severity Prediction Using Machine Learning," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579711.

[5] Kerfoot, Darren & Hofmann, Markus. (2018). ANALYSIS OF AVIATION ACCIDENTS DATA.

[6] ANALYSIS OF AVIATION ACCIDENTS DATA
Darren Kerfoot1, Dr. Markus Hofmann2
1Institute of Technology Blanchardstown, Blanchardstown Road North, Dublin 15, Ireland
2 Department of Informatics, Institute of Technology Blanchardstown, Blanchardstown Road North, Dublin 15, Ireland
email: darrenkerfoot@gmail,com, Markus.Hofmann@itb.ie