National College of Ireland

Using Machine Learning Models to Study Human Error Related Factors in Aviation Accidents and Incidents

MSc Research Project
Data Analytics

Naumaan Mohameed Saeed Kazi
Student ID: x18130208

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Naumaan Mohammed Saeed Kazi
Student ID:	x18130208
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	12/12/2020
Project Title:	Using Machine Learning Models to Study HumanError Related Factors in Aviation Accidents and Incidents
Word Count:	7357
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	30th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Using Machine Learning Models to Study Human Error Related Factors in Aviation Accidents and Incidents

Naumaan Mohammed Saeed Kazi
18130208

Abstract

The importance of Human Factor (HF) is long been recognized in aviation industry, in order to deeply understand and prevent the errors caused by humans was the foremost challenge for the safety board of aviation. The focus of this study is to identify the characteristic of human error causing aviation accidents and incidents, with the presence of these attributes in a large sample of aviation crashes. Archaeological data was collected from 1971 to 2018, which is of 47 years as it was used to identify the presence of HF was thoroughly analyzed in correlation to attributes indicating pilot features, crash conditions, and aircraft features. Models Gaussian Naïve Bayes, Random Forest, Logistic Regression, XGBoost classifier, SVM and Artificial Neural Network (ANN) modeling was performed to evaluate the associations of individual attributes with the probability of HF given a crash. Through this study we found accuracy to give the accurate evaluation for every classifier. In comparison between top three models, SVM with cross validation managed to give highest accuracy of 96%. The result of 93.19% in ANN model was improved using Hyper-Parameter tuning which gave an accuracy of 93.29%. During the evaluation of this study we would demonstrate to yield meaningful information using machine learning models.

1 Introduction

The importance of human factors has been known since a long time, in particular Human Factor (HF). In account 50% of the incidents and 80% of the aviation accidents are caused due to HF.Li et al. (2001) The manufacturing and designing of aircraft has become very reliable as the technology has been developing, this has considerably improved the safety in aviation industry in the past 3 decades. Majorly the accidents and incidents which has been recorded are only of negligence caused by human. The root causes and the accidents should be very carefully be analysed and inspected in-order to improve the aviation safety.Mathur et al. (2017) Looking at the overall accidents taken place in USA of all the transport 0.03% of these accidents were accounted of aviation ranging from 1990 to 2010, highways were at 99.86% and the railways second at 0.11%. Noting that at the same time range 1.98% of these fatalities were caused by airways, 2.46% of the fatalities by railways and the highest recorded figure was at 95.56% by highways. Bazargan and Guzhva (2011) Inferring from the Figure 1 if we have to classify the accidents in aviation

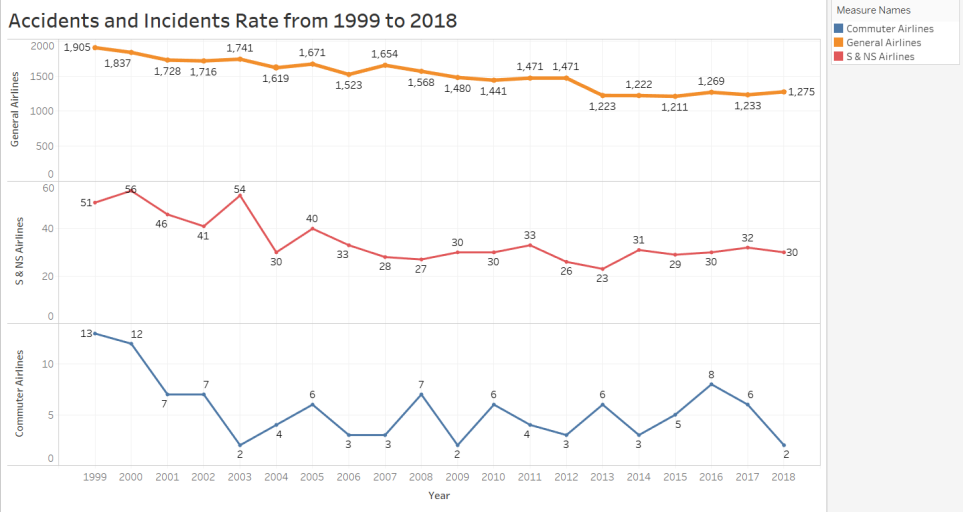

Figure 1: Aviation Accidents and Incidents

in all of U.S. in general aviation and the commercial aviation then one would easily understand that there has been an uneven distribution among the subgroups.

Here General aviation is accountable for a total of 82% total airways related accidents and incidents. And commercial aviation is accounted for 83% in total.[1] The General aviation has been classified flight for civilians other than that of business, personal or instructional flying's. With analyzing the historic data, it will be very informative and helpful for the industry of aviation to find the root cause of the accidents and incidents. There has been very less verifiable research conducted on the HF and it has been restricted to only discussing the behavioral and the operational events then to classify these into many glossaries. Nevertheless there has been progress and causes identified with these restricted work to reduce the pilot errors, where these information are illustrative in nature and also gives very less knowledge. There has been a lack of research done on the archaeological data and factors to correctly identify the HF errors has been difficult.

Dedicated to this research study and all over this paper these definitions will be used which are defined as follows:

Factors- It could be predefined condition, an mistake or occurrence that would lead up to an incidents or an accident.

Accidents- It is a condition which is related to airplane in which humans suffer injury or death. In other case the airplane has been caused huge damage.

Incidents- It is a condition which is related to airplane in which it would not be an accidents but would be some addition of 1 or many factors, which would have an outcome of fatalities, damage to the airplane or injuries to the human.

1.1 Research Objectives

Safety of passenger is the utmost priority in any type of transport be it roadways, airways or railways. And as technology is improving day by day there are very few incidents happening because of technical failure or engine malfunction. The highest percentage of

[1] `https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/2000s/media/200618.pdf`

accidents and incidents happening these days are because of human error which accounts for 2/3rd of the total aviation accidents.

The sole purpose of this research is to study the importance of HF in general aviation transport accidents and incidents from 1986 to 2018, which is of 32 years epidemiological data which is novel in its way as there has not been and significant research conducted on this archaeological data. Here the term HF is used, rather than other many types of factors in aviation. There are other factors such as cabin-crew, ground staff, air traffic controller and so on, which also plays a very important role in these accidents and incidents. The aim of this research would be to provide aviation federation with meaningful analytical insights to make improvements to the safety of aircrafts handling and also the bring down the percentage of general aviation accidents and incidents.

In order to give granularity, The following research question and research objectives are discussed below.

"Can related human factor error attributes like age, certificate of pilot, flight hours, total hours of pilot and operation type help to improve in predicting the fatalities and injuries because of accidents in aviation using machine learning techniques?"

The primary focus of the research is to propose a prediction model which would be very efficient in giving prediction of aviation accidents and incidents for the country U.S. With the question the research would also be addressing the below mentioned objectives:

- Designing a model that would be identifying the variables which will be best fit for prediction of Human Factors in aviation accidents and incidents
- To explore and to observe the data & implementation of the proposed machine learning models to get correct solutions.
- Comparing and evaluating the best machine learning algorithm for prediction.
- Implementation of methodology which is very robust in nature and on running many experiments which will improve the models performance.

The research paper has been broken down in many section which are as follows: Related work is discussed in Section 2. Section 3 is explanation of CRISP-DM methodology which is implemented in this research. Implementation is discussed in Section 4. Results are evaluated in Section 5. And lastly, conclusion is in Section 6.

2 Related Work

A very popular term in aviation is Human Factor Analysis and Classification System as this was originated based on more aviation accidents and incidents happens because of human error then, that caused because of technical failure. HF (Human Factor) is a very broad topic in relation to the scope it has and the base it holds. This involves a huge collection of attributes such as humans' abilities, task, machine, an envelope which creates a comfortable, effective and secure process to be used by the humans.

The following research has been divided into three sections namely Human factors, Environment Factor and Others which consist of technical aspects. In Table 1 a general comparison of all previous algorithms used and the outcomes achieved are mentioned .

2.1 Human Factors

Using data mining approach the author Burnett and Si (2017) builds a classification techniques for prediction of aviation accidents and fatalities. For this research the data used was of 27 years which is for the period of 1975-2002 from FAA. Predictions of those situations which might have the high probability of aviation accidents which would be resulting in fatalities or injuries. There is study stating that 80% of accidents and injuries in aviation occurs because of pilot error Billings and Reynard (1984) & Li et al. (2001). Various models such as Decision Tree, ANN (Artificial Neural Network), K-Nearest Neighbors and Support Vector Machine were implemented in this research. MATLAB which is a statistical tool was used by the Burnett and Si (2017) for the purpose of research,which had used for different function which has been used for implementation of these algorithm. LogitBoost function on MATLAB is used for Decision Tree classification model. Similarly, fitcknn function was used for KNN classifier, fitcecoc function was used for SVM classifier and patternnet function was used to deploy ANN. The study proposed that ANN algorithm has given the highest prediction in aviation accidents and injuries

A very strong theory was published by Walton and Politano (2016) and by Bazargan and Guzhva (2011) stating about gender and experience. Where the research states the female pilots makes errors at training phase and while there is higher probability of accidents by male on the actual flight with passengers. With the experience high the pilot gets older and there are considerable number fatigue and restlessness in the pilots. Both the paper uses the same approach of Chi-Squared & Logistic Regression. Understanding of both the study is that pilot above age of 60 are more likely to be in airplane accidents.

$$\begin{aligned}\text{pilotErrori} \; - a + b_1 \; \text{Gender}_{\;i} + b_2 Age1_i + b_3 Age2_i + b_4 Age4 + b_5 Age5_i + b_5 Age6_i \\ + b_7 \, \text{Exp}\, 1_i + b_8 Exp2_i + b_9 Exp3_i + b_{10} ExpS_i + e_i\end{aligned}$$

Feng and Li (2010) and Bazargan and Guzhva (2007) both proposes a very similar research stating very similar human factors which are responsible for the accidents and incidents in aviation. Feng and Li (2010) has used data on 14 years and discusses about the events, condition and errors made by the pilot during the course of the journey. when compared these two researches the author has used very different algorithms to support its objectives by use of Contrast-Set Mining and Attribute Focusing Algorithm and the later researchers uses statistical approach of Logistic Regression. Having used data of 20 years generated better results. with having overall 89.9% accuracy.

Pilot error is the accounted to 80% of the the total aviation accidents the author Management et al. (2014) and Kharoufah et al. (2018) uses pilot age and experience as the two attributes for the analysis. The techniques used here are statistical method of chi square test and the logistic regression. Here the data used is of 14 years which was taken NTSB. Kharoufah et al. (2018) has taken it further by categorization of age groups and flight experience. This categorization technique was later followed as benchmark in many researches.

Management et al. (2014) commented that improved SVM is required to make better prediction in stating that human factors are the largest contributors in aviation accidents and for this it takes all the general attributes of the pilot andand the tweak the SVM to get better outcomes where it manages to get gamma = 0.06. It also discusses that for the improved SVM they tuned the kernel density estimation, and this also helps in knowing the probability when the data is not correctly distributed.

In comparison of two study by Kharoufah et al. (2018) & Mathur et al. (2017), the author Mathur et al. (2017) strongly suggest that the only cause of aircraft accidents are human errors. It states that aircraft technology is highly reliable and advance as there is next to none chance of accidents taking place of technical issues. By the use of Logistic Regression, it proves so. For this research there has been use of three dataset and the total record to support his theory were 7415. While Kharoufah et al. (2018) states that there is 75% chance of Human Error. For which it uses data of 16 years of over 200 air transport commercial records. The method used is similar with logistic regression it also uses Chi-Squared method.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Human factor causal are briefly discussed by Erjavac et al. (2018), Li (2014) and Loughney and Wang (2018). Here the first author suggests that there should be two class of mishaps. one is the latent and the other would be symptomatic factors. For this research it used Multiple Variable Logistic Regression in finding the relationships between them. As the Latent based factors would have the higher probability of accidents then the later. Using statistical approaches Li (2014) states that many factors are unnoticed while finding the root cause of the accidents similar to Erjavac et al. (2018) the author has divided it into two categories which are experimental and observational studies. Using the statistical approach outcomes were recorded. So was the work of Loughney and Wang (2018) human factors were in depth defined and the classified in many different categories. The author then develops a framework of the human factors based on 517 factors.

Table 1: Algorithms and results from previous study.

References	**Algorithm**	**Accuracy%**
Burnett and Si (2017)	K-Nearest Neighbor	90.41
	Decision Tree Classifiers	90.35
	Support Vector Machine Classifiers	88.76
	Artificial Neural Network Classifiers	91.84
Bazargan and Guzhva (2007)	Logistic Regression	88
Li (2014)	Decision Tree Classifiers	87.39
Lukáčová et al. (2014)	C5.0	98.63
Christopher and Appavu (2013)	Multiple Logistic Regression	56.89
Hofmann (2019)	K-Nearest Neighbor	41.71
	Naïve Bayes	45.47
	Decision Tree Classifiers	59.90
	Gradient Boosted Tree	52.68
	Deep Learning	63.97

Controlled flight in terrain has been accounted for a very high number of accidents and fatalities when compared to other types of accidents. For this proposed research

authorKelly (2019) had taken 50 such accidents from 10 years period of 24 countries. as the results showed that 1289 were the number of causal and contributors for the accidents. and that 44% accidents occurred because of pilot error. To justify this claim one such study by Kumar et al. (2016) states that human error needs to be categorized and standardized. This is classified into error rate and critical index and consequences. Using MATLAB the researcher had applied SVM.

On approaches of Human Factor, a study done on Australian aviation accidents data. The author Olsen and Williamson (2017) makes a contrast comparison of the base paper where its finding of 90.7% on taxonomy and 79.5% on sub category which were relatively low. When compared by O'Hare (2000) it makes similar claims with two surveys on the human factor 1st is flight time entries in the form of long book where there are actual hours flown in various other categories of flights. and the 2nd is the information of accidents and incidents.

Li et al. (2001) proposes Multivariate Logistic Regression and Mcfadden (1997) had proposed Logistic Regression which are a direct comparison as both uses the same pilot error objectives for the same dataset ranging from 1986-1992. where the Multivariate Logistic Regression performs better than the later. But to this adverse weather conditions are very closely related in both the findings.

2.2 Weather Factors

Using Decision Tree Classifier in data mining approach the author Christopher and Appavu (2013) discusses the accidents of Turkey Airline where the data is use from FAA database for 41 years. With an accomplished accuracy of 87.39%. As the dependent variable used was wind. For this test the tool used was WEKA.

Meteorological phenomena are probably consisting of 40% of the aircraft accidents attributes such as low clouds or fog. Lukáčová et al. (2014) published a research in which a prediction model is build using CRISP-DM methodology where models used are C5.0, CHAID and CART. The testing was carried on two test models where the accuracy was close to 98.63% for the first model and 97.05% for the second model. On the basis of this research a very similar work was published by Fultz and Ashley (2016) where it goes on in suggesting 60% of the accidents were caused because of weather conditions. It makes prediction stating that mostly accidents take place from October to April and on weekends. This data was collected for USA for 31 years and factors used for weather were 53. The method used here was statistical technique.

Weather can be very unpredictable in high elevation and mountainous regions. Based on the factors such as wind-shear, mountain obscuration, gusting winds & whiteouts and so on the researcher Aguiar et al. (2017) proposed a Chi-Squared test. The data was collected for 14years and this data was grouped of years and the results was obtained.

In order to reduce the weather related aviation accidents a researcher from NASA Schaffner (2019) suggested a highly distributed prediction case. This will very accurately make prediction on the weather and make an intuitive message to the pilot and air traffic controller. Such research has significantly dropped the rate of accidents

2.3 Technical Factors

Major causes of accidents in twin engine aircraft has been the the fractured pipeline carrying the fuel. on comparison of research by Sujata et al. (2019) & Boyd (2015) it was

found out using the the Logistic Regression and Chi-Squared test for the both results. The major buyers of these twin-engine aircraft were the 4-8 capacity Cessna airplanes. running the test on 376 accidents the research was published. There finding suggested that 27% fatal accidents were caused because of technical errors. Logistic Regression performed the best in this comparison.

On using of Multiple Logistic Regression model for a similar study Handel and Yackel (2011) takes incidents and accidents caused in commercial, medical and on ground flights. Making predictions on fixed wing 2-4 capacity aircrafts for the data collected from 1984 to 2009. On model execution gets 56.89% on Confidence Interval of 95%.

Ensemble model in machine learning was used for predicting the incidents.Zhang and Mahadevan (2019) This ensemble was of neural network and hybrid SVM developed to causes and risk associated with technical failure in aircrafts. The implementation was done with 10-fold cross validation on both the mentioned models.

Weight stabilization is a known factor in airplanes. One such research Boyd (2016) suggest that exceedance in weight would result in accidents. The gravity limits the gradient climb of a airplane. Which is because the airframe is designed in a different approach. And the method used here was statistical in calculating the Poisson Distribution, T-test and Proportion test. Getting the P value of 0.072 which is not less than 0.001.

A comprehensive study was done on Loss-of-Control on an aircraft. Ancel et al. (2015) a very simple model was build which had enabled in checking many impacts on such accidents by use of NASA AvSP. The study was on the basis of data of 22 years which was collected from NTSB. This will the airlines hierarchy to fix the issue by training the ground staff and cabin-crew.

The two researchers Diamoutene et al. (2018) & Hofmann (2019) proposes an analyzes on accidents with the same data and different method while one uses the statistical approach the other with data mining methods. The attributes under consideration were the purpose of the flight, aircraft attributes and geographical. There were many different models used under which few are Gaussian Naïve Bayes, Decision tree and KNN classifier.

3 Methodology

Research study would primarily be focused on predicting the Human Factors(HF) for the aviation accidents and incidents caused right from the year 1986 to 2018, this an epidemiological data. As studied from the previous section CRISP-DM has been used to get the expected result. Here we would be establishing a classification model for categorization of HF attributes and then would predict the results. The method that has been selected is CRISP-DM which was proposed by Wirth (2000) where the research has been divided in to 5 sections as per shown in the Figure 2. This research would be constructed in such a manner that it would be able to answer all the objectives and the questions mentioned in Section 1.1 in a very accurate manner. The following are the steps of CRISP-DM which are followed in this research:

3.1 Business Understanding

With the improving technology the demand for safer airway transports is ever increasing. Aircraft's were first invented in 1800's century and since then lot of improvement and purpose of the flights have been changed. In today's age aircraft for passenger flights our increasing every single quarter. In accounted to the statistic of end of 2018 there

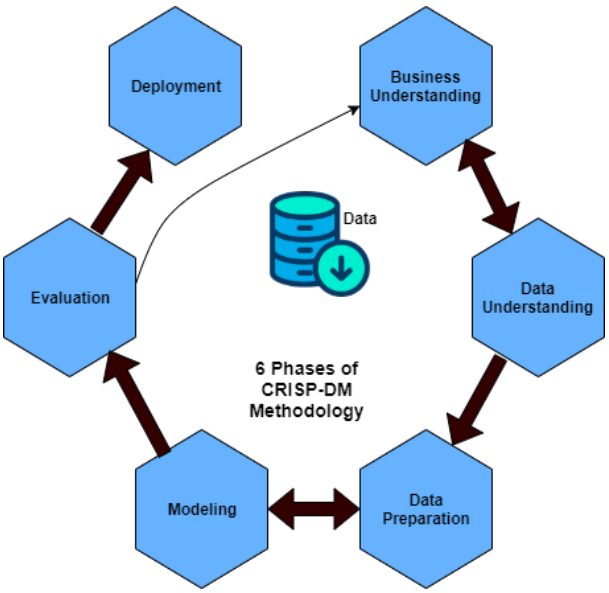

Figure 2: CRISP-DM Methodology

were 4378 million passenger travelling through aircrafts around the globe[2]. Let alone there were over 1,011.million passenger which was increase by 4.8% from the previous year[3]. However, there has been a lot of research conducted to improve the safety of the passenger-general aircraft, there has been not a convincing methods used to get to the root cause. Most of the research has been conducted using the age old statistical methods, with the advancement of machine learning and new improved tools these prediction can more accurately be find-out and the would help the aviation industry in finding out the correct cause. Burnett and Si (2017) had implemented the approach of HF but they are still not satisfied with their outcomes and have not covered attributes which would be very vital in these types of researches. These new attributes needs to be considered in finding more accurate predictions.

The loop holes with previous research was that the models and the attributes used for the study were not accurate. Using classification method highly accurate and efficient approach would be be taken. The two principle ways to solve this issue are as follows:

1. Classification of these HF related attributes to generate a highly accurate predicting models.

2. Using machine learning models and using archaeological data for better analysis.

3.2 Data Acquisition

The first step of the implementation is gathering data, as the data for this research has been downloaded from National Transportation Safety Board (NTSB) [4]. As the data was originally in database structure which the file was had to be converted to CSV format. This file was exported from MS Access to MS Excel. After gathering the data and then

[2]https://www.statista.com/statistics/564717/airline-industry-passenger-traffic-globally/
[3]https://www.bts.dot.gov/newsroom/2018-traffic-data-us-airlines-and-foreign-airlines-us-flights
[4]https://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx

exporting in the file format required the was of 90000 records from 1971 to 2018. And attributes consisting of 203. This data used modelling has to be divided in two class one is the training and the other is testing. Splitting the manually is more inclined to having inaccuracy in results and errors. Using Python scikit learn libraries the splitting of data was achieved which would get us accurate results.The role of data is very important in getting the target accuracy.In order to achieved the required results huge dataset is required for such analysis, it also plays a very important role as there is an increase in volume of data to get correct accuracy Wirth (2000). As the idea is basically universal and it could be applied on any other transport like railways or roadways if at all needed in future.

3.3 Data Preparation

The data was downloaded in a MS Access database file format which is MDB. Which to be used for a analysis would not be appropriate, data file had 4 dataset's which were then extracted to a MS Excel in CSV file format.

The data has various attributes like TIME_OCCUR,ACC_INC_CLASS, HOURS _ TOTAL_PILOT1, AGE_PILOT1,RATING_1, CERTIFICATE_PILOT1, RATED_ AIRCRAFT_1, PASSENGERS_TOTALS, OPERATION_TYPE, PILOT_FATAL, TOTAL_ ABRD_TOTALS. Null values were handled in this by taking mean and mode of the attributes and deleting the null value only of age attributes as it cannot be substituted by any other way. The date is considered from 1-01-1971 to 31-12-2018, age of pilot from 21 years to 65 years. The certification of pilot, hours the pilot have flown the aircraft in total. Moreover the rating of each pilot based on his performance of flying aircraft. Each an every attributes of pilot has been recorded, while making prediction the data is divided into two section as training model is given on 80% data and tested on 20% data.

3.4 Class Imbalance

On performing under sampling we lose out on very important information. Class imbalance has been handled in this study, there is comparison between the two and of which the best performing approach has been chosen. For imbalance Binning and SMOTE technique has been implemented. For this both techniques has been performed on model Logistic regression. The was was sorted in two categorizes "low" and "mid", as the deaths less then or equal to '9' was binned as low and the deaths above 9 would be mid. The number of death happening in aircraft after the technological advancement has been very less Kharoufah et al. (2018).This helped in solving the heavy loss of data. The minority class was 39459 and the majority class was 45866. As binning is used for smoothing the data and at times handle noise. First the data will be sorted and stored values would be divided into many bins. SMOTE will synthesis the data and perform over sampling. It is a very biased prediction. And we decided to working with binning and keep the data rather from getting lose of information or random synthesis data.

3.5 Feature Engineering

There were approaches used for feature selection one was the correlation matrix and the other was the Carmer's V rule. As firstly the correlation matrix was plotted using Seaborn and Matplot library for the numerical data, as these comes under the Scipy.stats

which is statistical libraries of Python. Secondly Carmer's V rule was applied on the data set. As Carmer's V rule is only capable of finding relations for categorical data. Both the approaches were used in the research. And then depending on the highly co-related attributes the model was constructed. We had got attributes which highly co-related and out of the bunch we selected 31 attributes for the next step. For better and accurately predicting we had applied random Forest Classifier for feature selection as this a type of ensemble decision tree which works on a randomly generated data. As shown Figure 3 in this type of classification each of the tree would vote and the most voted class would be selected. This approach was chosen as it is very powerful when it is compared to classification algorithms which are non-linear.

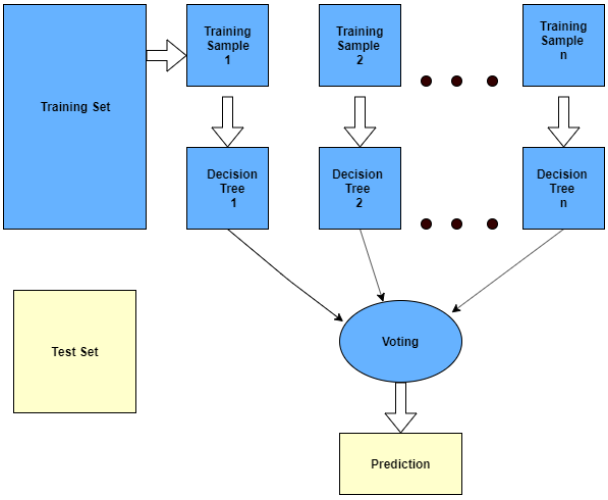

Figure 3: Random Forest Classifier for Feature Selection

3.6 Modelling

Through this research there has been several of classification algorithms which has been implemented namely Logistic Regression, Gaussian Naive Bayes, Random Forest, XG-Boost, Support Vector Machine and Artificial Neural Network in order to predict the Human Factors in aviation industry.

3.6.1 Random Forest

Random Forest is a type of estimator which would put many number of decision trees classifiers on various small samples of a given dataset and would be averaging the them in order to improve the prediction on control on over-fitting and the accuracy Aliwy and Ameer (2017). the various small samples would be the same as the one from the original sample.As shown in the Figure 4 RFclassifer function is used in thi study. The learning would be smooth within the interactions of these predictors of data which would have no restrictions of scaling, outliers or missing values. This model is very well known for its efficiency and the speed for performance Sousa et al. (2019).

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Figure 4: RFClassifier Function for Random Forest

3.6.2 Gaussian Naïve Bayes

One of the most feasible and simple model for the classification is the Gaussian Naïve Bayes. This model is very much suitable for large data as it gives very efficient results. Adding to it is well known for it exceptional speed and accuracy in prediction Sousa et al. (2019). Naïve Bayes is very well known for handling missing values, low level of variance and noise.

3.6.3 Support Vector Machine

The most preferred model for classification issue in supervised type machine learning is the SVM. As this model is a kernel based algorithm it deals with a fraction of training samples, where the primary objective is to increase the margins of these hyperplane which are used in different classes Burnett and Si (2017). When the number of samples are increased it is highly effective and the efficient with memory.

3.6.4 Logistic Regression

For the use of prediction Logistic Regression is very simple algorithm and it also requires the least computational time. It gives knowledge about the input which has a huge impact on the output.

$$y = e\hat{} \left(b0 + b1^{\star}x\right) / \left(1 + e\hat{} \left(b0 + b1^{\star}x\right)\right)$$

3.6.5 XGBoost

The most popular and hyped about model is XGBoost, it as integration of a tree type learning along with liner models and also it can be executed for parallel computing. It was developed by Chen and Guestrin (2016). The use of XGBoost is made possible in Python is by XGBClassifier function as seen in the Figure 5.

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=3,
              min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
              nthread=None, objective='binary:logistic', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
```

Figure 5: XGBClassifier function for XGBoost Model

3.6.6 Artificial Neural Network

This is a supervised learning algorithm that learns from the function $f(\cdot) : R^m \rightarrow R^o$ which is through training on the given dataset, here the "m" denotes the number of inputs

dimensions and the "o" denotes the number of output dimensions Brownlee (2018). On giving the a fix number of features X = x1, x2,...xm and y as the target, it will be learning the non-linear type functions approximator for both Regression and Classification.

As for our research ANN Classifier has been used where the class would implement the multi-layer perceptron (MLP) which is trained in backpropagation fashion. The library used is Keras and the function is classifier.

3.6.7 Evaluation

There are various methods to perform evaluation on a research outcome namely F1-Score, Accuracy, Precision, Recall. Using F1-Score we could have a balance in between Recall and Precision.Cross Validation and Stratified K-fold are performed for the robust machine learning algorithms for the evaluating. The proposed research is an improvised version of Burnett and Si (2017) and Kumar et al. (2016). Precision is the percentage of the relevant total results which is correctly classified by the given model[5]. For this research Recall is one of the metrics where recall is the proportion of predicting in the actual class all the observation positively[6]

4 Implementation

The Figure shown below gives us an illustrative view of the research which is implemented. The very first step is the collection of data which starts from downloading the zip file from the NTSB website which on extraction is in MS Access database format. Here after the file which needed for our research must be in CSV file format has been extracted and then loaded in Python (Juypter Notebook) for further analysis. To perform classification the very next step is pre-processing of data. And here after the results of classification are compared and to identify which of them is the best and these results are then represented using the graphs.Figure 6

4.1 Data Collection

The data was gathered from National Transport (NTSB) which the U.S. government which keeps track of all the safety statistics of all modes of transportation in the country. The data gathered is from 1971 to 2018. The file was in a zip folder which on downloading was MS Access database which was then extracted to CSV.

4.2 Data Pre-Processing

The data selected for this research had 90000 records which here means number of accidents and incidents in U.S. for aviation industry, in which there were 2023 attributes. These attributes covered from aircraft make, model and engine to pilot grade, age, hours of flight, experience and certificate to cabin crew information and technical attributes to weather attributes like wind speed, snow or raining and many more. Particular for this study we needed HF which reduced down to 110 attributes. Many anomalies such as characters in a numerical attribute or vice versa were removed using the replace function.

[5] https://towardsdatascience.com/precision-vs-recall-386cf9f89488

[6] https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-me

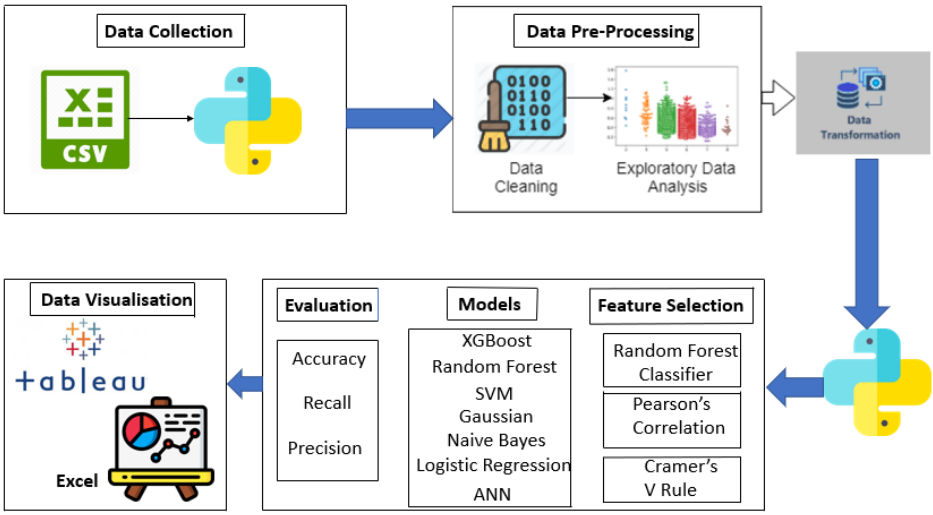

Figure 6: Flow Diagram

These anomalies were present in all attributes as this data was a extraction from different file format. This was implemented using Python. The major part of pre-processing data is handling of missing values in data.The library "missingno" missing values were identified. There were three ways how the missing values were handled which were Mean, Median and Drop. These handling of data was done after researching in detail about the attributes and events and the filling the data. Like example the age of the pilot cannot be replaced by any other value as there cannot be a substitution, so the values were dropped.

4.3 Feature Extraction

Feature extraction is done on pre- processed data. In this research there are has been used of three types of feature selection these were compared against each-other and the best fitting one was selected for model implementation. Pearson Correlation Matrix, Carmer's V Rule and random Forest Classifier (RFC) for selection. Each applied on XGBoost. Of these RFC was the best performing feature selection model followed by Pearson Correlation and lastly Carmer's V Rule. As we clearly know Carmer's V Rule is used for categorical data. The library used for RFC is sklearn.ensemble.forest. For classifier it would measure those impurity rather then the information gained.

4.4 Data Preparation

As many times the data that we have is in the form of qualitative nature, which means there are text in our data. Rather than the numbers, if the data is in text format and difficult to process them as these models are in mathematics form i.e. calculations and equations. That's the reason it is necessary to encode these categorical data. For this research we are making use of LabelEncoder library from the Scikit library. The data set was biased as death in very few accidents would be extremely high but in nature of these improved technology over the past three decades the death rate has been less. So, this lead to class imbalance, So to handle this there were two methods applied and one

was selected for the research. One is the under sampling, at first the data was for under sampling was divided in to three class upon execution of this the model was over fitting and the accuracy was close to 99% and above.And the data loss was very high. The data after this three-class categorization was 9581. To handle this, we read few more research papers and deeper understanding the data it was decided to categories this data in to two classes i.e. "low" and "high". Upon this the data for our analysis was 85325. Second method SMOTE it would create synthetic samples and not the duplicate sample of the minority class. it would select the same or similar records and twist the record 1 column at a given time with a different amount. That's how the minority class was be equal to majority class. But we selected the under sampling over SMOTE as we wanted true records and the data to be high in volume and lastly high accuracy.

4.5 Dataset Splitting

For model implementation the data would be split in training data and testing data. The machine learning would learn about correlation from the training data and would test model on the training data in order to inspect the prediction of how accurately it has been done. In this research 80% data is training and 20% data is split to testing. From scikit library test_train_split was used.

5 Evaluation

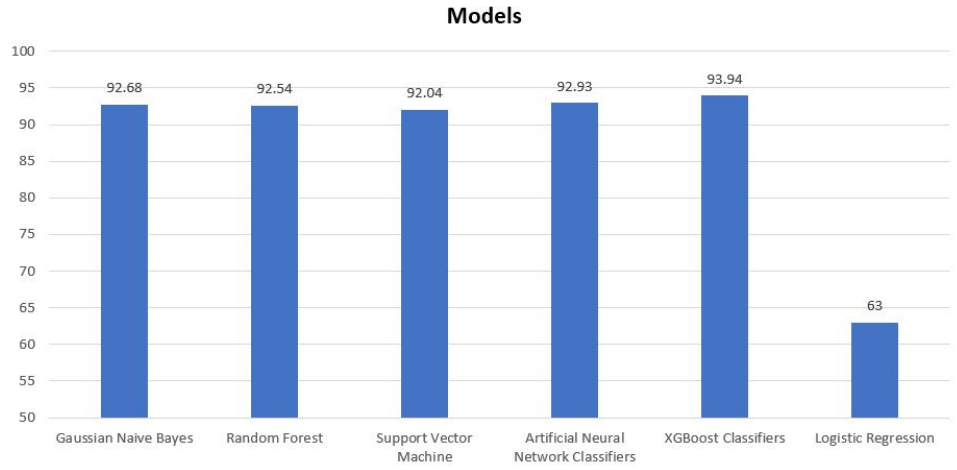

Figure 7: Results Before Hyper-Parameter Tuning.

Evaluation of these six classifier is performed using accuracy. In Figure 7, six classifier has been mentioned with the accuracy. Amongst all the models for classification, XGBoost and ANN algorithms gives the highest accuracy of 93.94% and 92.93%. Using hyper parameter tuning ANN algorithm has been improved. Gaussian Naïve Bayes gives accuracy of 92.68%. Logistic Regression model gives the lowest accuracy of 63% which was improved using stratified k-fold and 10-fold cross validation. SVM algorithm gives accuracy of 92.04%. The accuracy obtained from Random Forest Classifier is 92.54%.

5.1 Experiment 1: Cross Validation

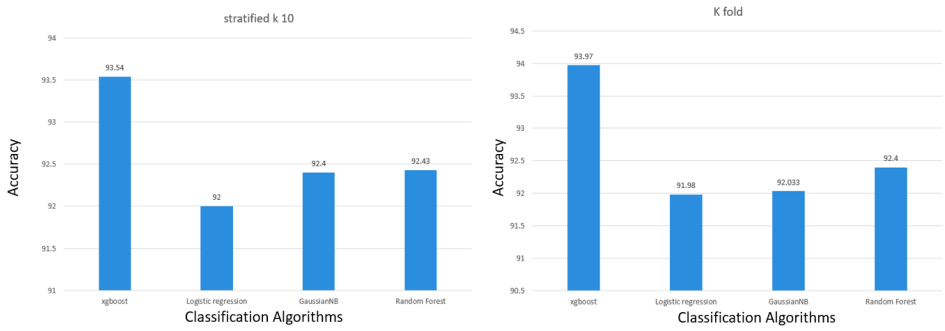

Figure 8: Comparison of Stratified Fold and 10-Fold Cross validation

In Figure 8 there are two approaches chosen for validation one Stratified K-Fold and 10-Fold cross validation and a comparison is done between both. For models XGBoost and Random Forest the models have given better accuracy in Stratified K-10 which is 93.54% and 92.43% respectively and for models Logistic Regression and Gaussian Naïve Bayes we managed to get 92% and 92.49% respectively. These are improvement over the initial accuracy**??**.We had implemented K-folds for 3, 5 and 7 but wasn't that satisfactory.The stratified K-fold makes each means of the given response value a approx equal to the mentioned folds.

5.2 Experiment 2: Feature Selection

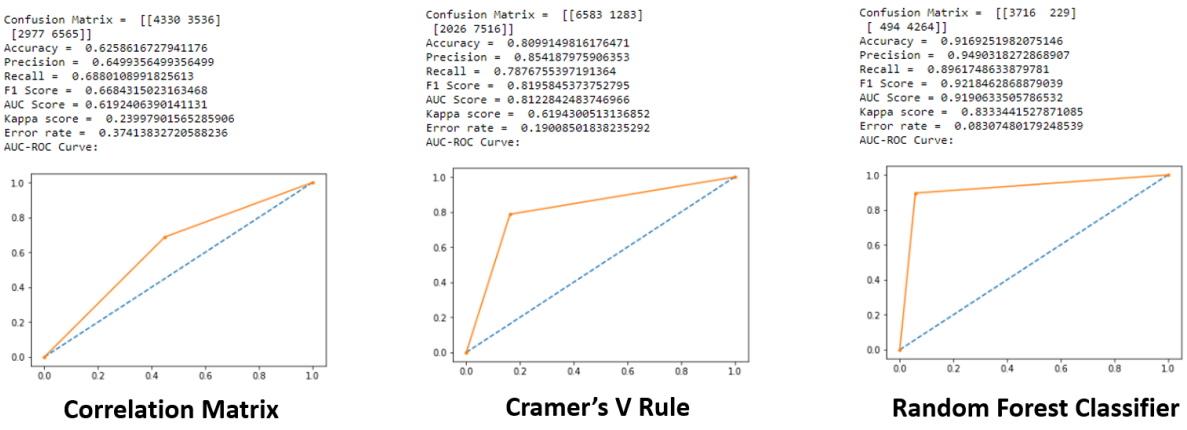

Figure 9: Comparison of Feature Selection

This study there are three methods were implemented. First Correlation matrix, Carmer's V rule and Random Forest Classifier for selection, all these were implemented first on logistic regression compared against each other to get the best result. From Figure 9 we can infer that upon using random Forest Selection Classifier we get the best result i.e. 91.69%. whereas in the case of Correlation matrix the accuracy was 63% and Carmer's V Rule the accuracy was 80.91%. As we know Carmer's V Rule is used

for finding relation between which categorical in nature. So, for implementation of our model we selected random forest selection classifier.

5.3 Experiment 3: Support Vector Machine

The results of SVM classifier without CV is 92.04%. It would separate the points of data with the help of hyperplane which having the highest margins. It is known for discovering new points in data. It also deals well with planes which are inseparable and non-linear in manner. As seen from the Figure 10 the confusion matrix obtained is the best performer. This is binary classifier in 0's and 1's, model performs best when it accurately predicts True positive and True Negative. In our case the SVM model predicts 7694 correct values for 0's and 1190 incorrect, similarly 8330 correct 1's and 194 incorrect.

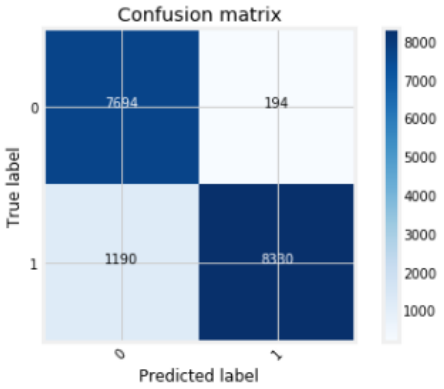

Figure 10: Confusion Matrix for SVM

SVM is known for offering high accuracy in machine learning as well as t is a slow learner. Cross validation with 10-folds are performed and the outcomes is shown in Figure 11 this models gives the accuracy of all the other models which is 96.93%. It performs this well is because of up sampling of data, the data was first down sampled and because of that the accuracy was unsatisfactory. Another reason is because of feature selection in which the we selected the best technique i.e. Random Forest selector. As using binning technique was used to up-sample the data we managed to get the most accurate results. Interpreting the confusion matrix has predicted 7728 values for True Positive and 373 values for false Negatives. And 160 False Positive and 9147 as the True Negatives.

5.4 Experiment 4: ANN

Artificial Neural Network is the first preference when there is any relationships,that needs to be incorporate. When the non-linear relationships are followed for prediction. Using Keras library in Python ANN was implemented and the parameter for activation is Relu which is because of classifier nature of the research Brownlee (2018). And the optimizer Adam is used. In neural network the optimization can be achieved with epoch, hidden layer and node. The ANN model is run for multiple epochs to get the desirable outcome.

On running for 10 epochs the accuracy for 50 epochs is 92.93, on 100 epochs was 93.19 and on 150 epochs was 92.75 from 2, as the on the epoch 150 the model was over fitting

```
10-Fold Cross  Validation Score of SVC 0.8952190962201204
Time Elapsed: 25402.251153469086
Confusion Matrix =  [[7728  160]
 [ 373 9147]]
Accuracy =  0.9693818933823529
Precision =  0.9828086386590739
Recall =  0.9608193277310925
F1 Score =  0.9716895947309715
AUC Score = 0.9702676760359316
Kappa score =  0.9383634661877212
Error rate =  0.03061810661764708
AUC-ROC Curve:
```

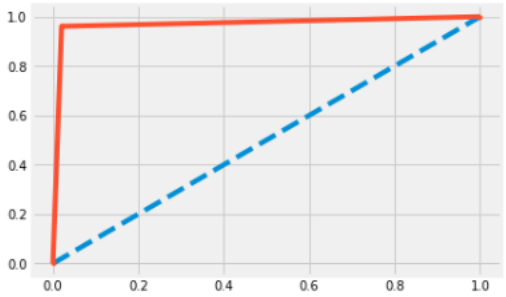

Figure 11: Implementation of SVM Using Cross Validation

Table 2: ANN Implementation Results of Epochs.

Epochs	**Accuracy**
50	92.93%
100	93.19%
150	92.75%

and the ideal epoch to run and get the desired outcome was for 100 epochs. As inferred from **??** ANN gives second best results so to get better results we have applied Hyper parameter tuning. In the Figure 12 below we can see that the model is implemented for 50 epochs in which it gives the most accurate results. we can observe that the model has very good performance on the train data-set. Also to take a note at epoch 50 is ideal for the model to learn. With implementation of hyper parameter tuning the results were improved as shown the figure below.

Table 3: ANN Implementation Results of Epochs with Hyper Parameter Tuning.

Epochs	**Accuracy**
50	92.21%
100	92.29%
150	93.29%

The accuracy without tuning was 93.19% and accuracy with hyper parameter tuning is 93.29% 3. An epoch is a complete pass over a given training data-set. whereas the loss is a value that needs to be minimize in our model at the time of training. In order to

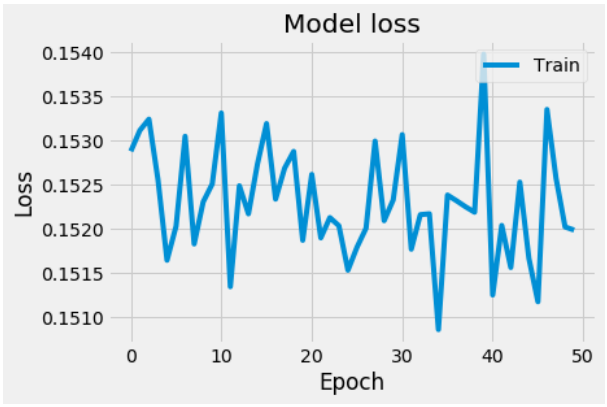

Figure 12: Plot of Model Accuracy on Training data-set in ANN

get predictions close to true labels we need to get the loss as less as possible[7]. The epoch running at 50epochs gives a less efficient model. So we ran an experiment on 50epochs which gave us an accuracy of 92.12%, 10epochs on 92.29%, 150epochs which gave the best accuracy of 93.29% and 200epochs which is decaying and gives less accuracy which is 92.98%. As seen in the Figure 13 the model running at 50epochs is not the suitable state and must be trained more at higher epochs to meet the accurate results.

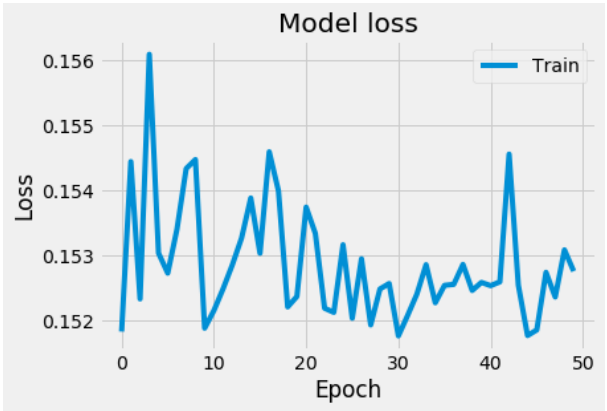

Figure 13: Plot of Model Accuracy on Training data-set in ANN using Hyper-Parameter Tuning

5.5 Discussion

In this study, we have achieved the objectives and the models have outperformed the bench mark set by the previous research. As seen the Figure 14 with applying hyper-parameter tuning to ANN it gives better results than the one achieved in the initial stages as discussed in Figure 7.

Burnett and Si (2017) had build a prediction model for aviation accidents and incidents, where it predicts the accidents caused by humans are highest. In this researcher had applied SVM which achieved an accuracy of 88.76% and in the implementation in

[7] https://stackoverflow.com/questions/34673396/what-does-the-standard-keras-model-output-mean-what

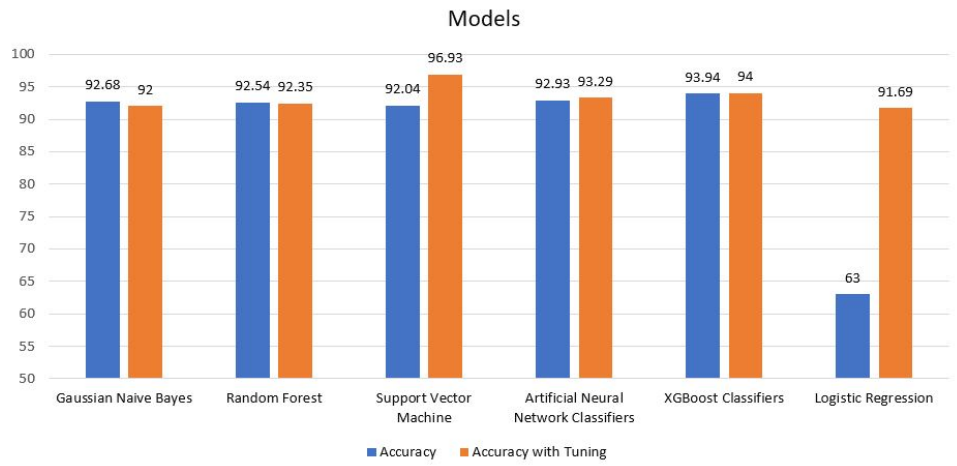

Figure 14: Comparison of Classifier Algorithms

this research got and accuracy of 96.93%. Comparatively, ANN gained 91.4% and with this research we obtained 93.29%. The models applied in this research when compared to the Burnett and Si (2017) has out performed the previous study. As particular to this comparison we have used data of 47 years for general aviation's and managed these significantly improved outcomes with better handling of data and with applying cross validation and hyper-parameter tuning.

As from this research Bazargan and Guzhva (2007) the attributes taken under consideration for the implementation were very limited and the models used here was Logistic Regression which was applied using the statistical analysis tool. The author achieved an accuracy of 88%. We managed to get an accuracy of 91.69%. As the number of attributes were high in this research and models used were machine learning.

Hofmann (2019) focused of using machine learning models for the implementation, where it used data from the same source but didn't used right attributes for the research. There were loops loses in handling of data. The models used by Hofmann (2019) were Naïve Bayes which achieved 41.71%, in our research the Naïve Bayes managed to achieve 92%. Then second algorithm applied by the author was XGBoost which managed to get 52.6%, in our research we managed to get 94%. And lastly Artificial Neural Network gave 63.97% for the researcher, in the implementation proposed by us we managed to get 93.29% with applying hyper-parameter tuning to the model which before hyper-parameter gave result i.e. 93.19%.

Using stratified K-fold and cross validation the accurate results were obtained. Through this research we implemented three types of feature selection. Which were Correlation matrix, Carmer's V rule, Random forest classifier. With the feature selection methods of random forest classifier the best features were then selected to make the model much effective and in giving better predictions. The model ANN was first implemented without hyper-parameter tuning which was 93.29% and after hyper-parameter tuning has increased to 93.97% at an epoch of 150. The implementation has been run on epochs 50, 100 and 100 to check the results for both with tuning and without tuning and the best fitting epoch and results were considered for this study. To better understand the data and selection of which feature selection model to select or best K-Fold methods to opt for we ran a continuous experiments on Logistic Regressions (LR). Binning and SMOTE has

been applied on LR in which we got an accuracy of 92% and on using binning technique to handle class imbalance we did under-sampling by implementation we got an accuracy of 91.98%. After applying cross validation on SVM we managed get an accuracy of 96.93% which is the best when compared to all the other models applied during this research, as this was achieved with smooth cleaning of data, feature selection where choosing the RF Selectors and handling the class imbalance with over sampling and the amount of data work the model to learning better and on that base making accurate predictions.

6 Conclusion and Future Work

As the research objectives and question stated in above section 1.1 , The implementation of this study was on classification models, which to understand the HF and compare & contrast the obtained outcomes from the previous researches and to considerably performed better. It was observed that XGBoost and ANN performed the best and later ANN performance was improved using hyper-parameter tuning with GridSearchCV function in Python. With cross validation of 10-Fold, which was implemented to achieve accurate results. From this research we achieved the highest accuracy for SVM with Cross Validation of 10-fold that was 96.93%, on comparison beats the benchmark of 88.76%. The objective was to build a robust machine learning models to execute many experiments and to find a very improve performance model. Moreover, the accuracy and efficiency of the matrix for evaluation has been achieved.

Few limitations that was observed during the implementation of was the limited data, as in this case it is only restricted to general aviation. As in the real world scenario there are accidents happening due to multiple categorise of aviation transport like commercial flights, civil aircrafts, air transport and military aviation, which needs to be considered to make a better model for understanding the causes much clearly and to achieve the maximum efficiency which can be implemented in the future work. And using advance ANN like Artificial Neural Network using Tensorflow can be used to explore hidden layers of the models. So can merging of data from FAA database repository and NTSB repository be done in the future study as it couldn't not be achieved due to time constraint. Apart from this, there has been a very genuine and honest efforts been taken to successfully achieve the research question and objectives, in order it to be useful for the aviation industry to take better strategical actions to improve the safety of passengers and bring further down the percentage of HF in aviation.

Acknowledgement

Firstly, I would sincerely like to share my gratitude for my mentor and supervisor Dr. Muhammad Iqbal for guiding and supporting me right from the very first day and have patience. It was all because of his constant feedback and guidance in the right path that helped me achieve these results. I was able to present these results in a very well mannered and meaningful way, is because of his inputs. I would like to thank my parents to always support me and have that constant believe in me. And lastly, I would also thank all my friends to push me to work hard.

References

Aguiar, M., Stolzer, A. and Boyd, D. D. (2017). Rates and causes of accidents for general aviation aircraft operating in a mountainous and high elevation terrain environment, *Accident Analysis and Prevention* **107**(November 2016): 195–201.

Aliwy, A. H. and Ameer, E. H. A. (2017). Comparative Study of Five Text Classification Algorithms with their Improvements, **12**(14): 4309–4319.

Ancel, E., Shih, A. T., Jones, S. M. and Reveley, M. S. (2015). Predictive safety analytics : inferring aviation accident shaping factors and causation, **18**(4): 428–451.

Bazargan, M. and Guzhva, V. S. (2007). Factors contributing to fatalities in General Aviation accidents, *World Review of Intermodal Transportation Research* **1**(2): 170.

Bazargan, M. and Guzhva, V. S. (2011). Impact of gender , age and experience of pilots on general aviation accidents, *Accident Analysis and Prevention* **43**(3): 962–970.
URL: *http://dx.doi.org/10.1016/j.aap.2010.11.023*

Billings, C. E. and Reynard, W. D. (1984). Human factors in aircraft incidents - Results of a 7-year study (Andre Allard Memorial Lecture).

Boyd, D. D. (2015). Causes and risk factors for fatal accidents in non-commercial twin engine piston general aviation aircraft, *Accident Analysis and Prevention* **77**: 113–119.

Boyd, D. D. (2016). General aviation accidents related to exceedance of airplane weight/center of gravity limits, *Accident Analysis and Prevention* **91**: 19–23.
URL: *http://dx.doi.org/10.1016/j.aap.2016.02.019*

Brownlee, J. (2018). Deep Learning for Natural Language Processing Develop Deep Learning Models for Natural Language in Python.

Burnett, R. A. and Si, D. (2017). Prediction of Injuries and Fatalities in Aviation Accidents through Machine Learning, pp. 60–68.

Chen, T. and Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System, pp. 785–794.

Christopher, A. B. and Appavu, S. (2013). Data mining approaches for aircraft accidents prediction: An empirical study on Turkey airline, *2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, ICE-CCN 2013* (Iceccn): 739–745.

Diamoutene, A., Kamsu-Foguem, B., Noureddine, F. and Barro, D. (2018). Prediction of U.S. General Aviation fatalities from extreme value approach, *Transportation Research Part A: Policy and Practice* **109**(October 2017): 65–75.
URL: *https://doi.org/10.1016/j.tra.2018.01.022*

Erjavac, A. J., Iammartino, R. and Fossaceca, J. M. (2018). Evaluation of preconditions a ff ecting symptomatic human error in general aviation and air carrier aviation accidents, *Reliability Engineering and System Safety* **178**(June): 156–163.
URL: *https://doi.org/10.1016/j.ress.2018.05.021*

Feng, X. and Li, J. (2010). Analyzing pilot-related accidents and incidents by data mining, *ICCASM 2010 - 2010 International Conference on Computer Application and System Modeling, Proceedings* **14**(Iccasm): V14–325–V14–327.

Fultz, A. J. and Ashley, W. S. (2016). Fatal weather-related general aviation accidents in the United States, *Physical Geography* **37**(5): 291–312.

Handel, D. A. and Yackel, T. R. (2011). Fixed-wing medical transport crashes: Characteristics associated with fatal outcomes, *Air Medical Journal* **30**(3): 149–152.

Hofmann, M. (2019). Analysis of aviation accidents data, (October 2018).

Kelly, D. (2019). An analysis of human factors in fi fty controlled fl ight into terrain aviation accidents from 2007 to 2017, *Journal of Safety Research* **69**: 155–165.
URL: *https://doi.org/10.1016/j.jsr.2019.03.009*

Kharoufah, H., Murray, J., Baxter, G. and Wild, G. (2018). Progress in Aerospace Sciences A review of human factors causations in commercial air transport accidents and incidents : From to 2000 – 2016, **99**(March): 1–13.

Kumar, P., Gupta, S., Agarwal, M. and Singh, U. (2016). Categorization and standardization of accidental risk-criticality levels of human error to develop risk and safety management policy, *Safety Science* **85**: 88–98.
URL: *http://dx.doi.org/10.1016/j.ssci.2016.01.007*

Li, G. (2014). Pilot-Related Factors in Aircraft Crashes : A Review of Epidemiologic Studies, (November 1994).

Li, G., Baker, S. P., Grabowski, J. G. and Rebok, G. W. (2001). Factors Associated with Pilot Error in Aviation Crashes, (February).

Loughney, S. and Wang, J. (2018). Modi fi ed human factor analysis and classi fi cation system for passenger vessel accidents (HFACS-PV), **161**(August 2017): 47–61.

Lukáčová, A., Babič, F. and Paralič, J. (2014). Building the prediction model from the aviation incident data, *SAMI 2014 - IEEE 12th International Symposium on Applied Machine Intelligence and Informatics, Proceedings* pp. 365–369.

Management, E., College, S. E. and Force, A. (2014). Study on the Aviation Accidents Due to Human Factors Based on Improved Support Vector Machine 2 The selection of key indicator based on Analytic Hierarchy Process (AHP), pp. 278–283.

Mathur, P., Khatri, S. K. and Sharma, M. (2017). Prediction of Aviation Accidents using Logistic Regression Model, pp. 1–4.

Mcfadden, K. L. (1997). Predicting pilot-error incidents of US airline pilots using logistic regression, **28**(3).

O'Hare, D. (2000). Copyright ©2000. All Rights Reserved.

Olsen, N. and Williamson, A. (2017). Application of classi fi cation principles to improve the reliability of incident classi fi cation systems : A test case using HFACS-ADF, *Applied Ergonomics* **63**: 31–40.
URL: *http://dx.doi.org/10.1016/j.apergo.2017.03.014*

Schaffner, P. R. (2019). REDUCING AVIATION WEATHER-RELATED ACCIDENTS THROUGH HIGH-FIDELITY WEATHER INFORMATION DISTRIBUTION AND PRESENTATION.

Sousa, A. L., Ribeiro, T. P., Relvas, S. and Barbosa-p, A. (2019). Using Machine Learning for Enhancing the Understanding of Bullwhip E ff ect in the Oil and Gas Industry, pp. 994–1012.

Sujata, M., Madan, M., Raghavendra, K., Jagannathan, N. and Bhaumik, S. K. (2019). Unraveling the cause of an aircraft accident, *Engineering Failure Analysis* **97**(January): 740–758.
URL: *https://doi.org/10.1016/j.engfailanal.2019.01.065*

Walton, R. O. and Politano, P. M. (2016). Characteristics of General Aviation Accidents Involving Male and Female Pilots, **6**: 39–44.

Wirth, R. (2000). Crisp-dm: Towards a standard process model for data mining, pp. 29–39.

Zhang, X. and Mahadevan, S. (2019). Ensemble machine learning models for aviation incident risk prediction, *Decision Support Systems* **116**(October 2018): 48–63.
URL: *https://doi.org/10.1016/j.dss.2018.10.009*