Application of learning methods for predictive modelling with human factors in aviation industry

Rui Nogueira[1]

Abstract

Aviation demands has increased over the years, and its safety standards are the most rigorous, by managing risk and preventing failures from human factors. However, there is more need to build models capable of predicting potential failures or risky situations in order to improve safety standards. The aim of this dissertation is to propose a model capable of predicting fatal occurrences and the degree of mortality, taking into account the human factors that contributed for the incident and information about the flight. The database was provided by the Aviation Safety Network (ASN), an organization which gathers reports about aviation occurrences, consisting of 1105 reports between 2007 and 2017.

Correlations between leading causes of incident and the human element are proposed, thanks to the Human Factors Analysis Classification System (HFACS). A classification model system is proposed, with the database preprocessed for the use of machine learning techniques. For modelling, supervised learning algorithms Random Forest (RF) and Artifical Neural Networks (ANN) and semi-supervised learning Active Learning (AL) are considered. For optimizing their respective structure, optimization methods are applied for hyperparameter analysis to improve the model. The performance is measure with precision, recall, accuracy and F1 score.

The best predictive model (Model 1), with use of RF, was able to achieve an accuracy of 90%, macro F1 87% and recall 86%, whereas ANN had worse results due to less ability for predicting fatal accidents. For Model 2, only AL had promising results after considerable training due to lower data sets.

Keywords: aviation safety, predictive modelling, human factors, supervised learning, machine learning

1. Introduction

Air transportation has developed into a crucial method of long-distance travel, with widely known contributions for economic and social development in a global capacity. Technological and management systems in air travel have facilitated by a close relationship towards safety improvement by aviation manufacturers and regulators leading to one of the safest transportation methods (figure 1) (ICAO, 2013).

However, with the sharp decrease of the accident rate, not only has the air traffic considerably increased but also the absolute number of accidents (Wiegmann and Shappell, 2003; Mazareanu, 2021), and market demands are such that professionals are required to work through large stretches of the day and/or night (Santos and Melicio, 2019; Dias et al., 2019). Approximately between 1975-1990, 70-80% of the accidents had human error causation, and studies within the

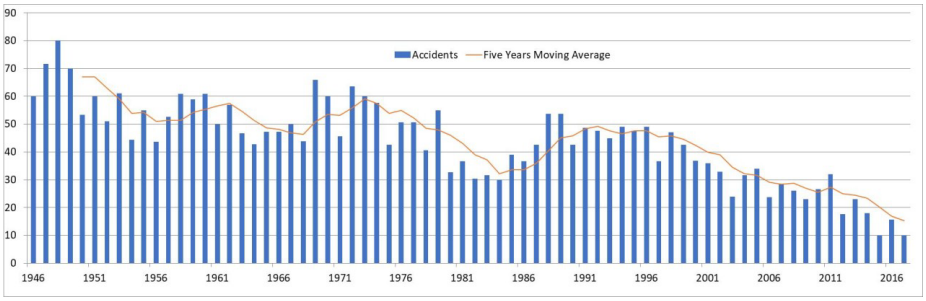


Figure 1: Number of fatal accidents between 1947 and 2016 (from Santos and Melicio (2019))

industry helped prove a link with the causation of human error in aviation and the added work effort requirements (Santos and Melicio, 2019; Wiegmann and Shappell, 2003; Dias et al., 2019).

Safety in aircraft industry is currently defined as the risk in which all aviation activities, either directly supports or are related to the operation of the aircraft, are reduced and controlled to an acceptable level ICAO (2013).

The Reason Model, as designed for information processing related to the type of errors. It defines two types of approaches regarding errors: the person approach, where the focus is solely on the individuality, such as personal moral weakness, forgetfulness and distraction, and the system approach, where the focus lies on the conditions which promotes human error, with the intention of building layers of defense to manage risk and mitigate hazard, so it is a safety management model Reason (1990).

Considering as cited that most accidents in aviation can be blamed at least in part on human factors, the Human Factors Analysis and Classification System (HFACS) tries to provide a framework for human error so that its causality in accidents can be measured and assessed Shappell and Wiegmann (2000, 1996). The taxonomy compiles the relations between human interactions and the possibility of error through a sequential framework, achieving three levels of potential error , with each level being increasingly specific and descriptive, from supervisory practices to operators actions, because its failures might lead to an accident.

The problem addressed in this article is the modelling of a classification system that encompasses human factors with the circumstances of an aviation incident or accident, with the intent of building a predictive system that helps with increasing safety standards within the industry. In order to build that model, a database provided by the Aviation Safety Network (ASN). There have been usages of predictive models using neural networks, to build a model for human factors evaluation in maritime accidents Qiao et al. (2020). The HFACS taxonomy was re-designed such that the factors were broke down into basic, intermediate and top events, because this helped to develop the structure of the neural network, with satisfactory results in terms of dealing with uncertainties and dynamics of the problem being studied and the models developed. There have been also approaches in developing new models, most notably semi-supervised learning algorithms. For example, in studying human factors in the aviation industry, an approach was made by extracting text data from reports using text related methods and modeling it using semi-supervised methods Madeira et al. (2021).

The paper is organized as follows. Section 2, it is described the algorithms used to implement the predictive models. There is also a description of criteria commonly used to evaluate the performance of classification models. Section 3 has a detailed explanation of how the ASN's

database was expanded and the pre-processing of data regarding human factors. Section 4 will have the results of models applied and their validity is determined, and the optimization of the algorithms considering the database is also achieved, and finally the last one (chapter) will conclude if the aims that were proposed for this thesis were achieved and future references to build upon the results and findings of this thesis.

2. Data Implementation

2.1. Random Forest

Random Forest is a supervised learning algorithm made up of a collection of tree-structured classifiers, which are defined as decision tree, applied throughout a given dataset on multiple sub-samples Breiman (2001). The decision tree is built up from a number of nodes, connected by branches, descending from the root node, placed at the top by convention, to the leaf nodes Larose and Larose (2014). Features are tested at the decision nodes, leading into a branch. Those branches can lead up onto another decision node or concluding in a leaf node. The process to generate a decision tree starts with splitting the root node into binary pieces. TThe splitting procedure is as follows:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{1}$$

where s is the candidate split at node t, $\Delta i(s,t)$ is a measure of impurity reduction from split s, $i(t)$ represents the impurity before splitting, and $i(t_L)$ and $i(t_R)$ show the impurity of the left child node T_L and of the right child node T_R after halving node t by split s.

A random forest can be defined as a "combination of tree predictors such that each tree depends on the values of a random vector sampled independently, with the same distribution for all trees in the forest", and each tree votes for the most popular class at a given input (figure 2) Breiman (2001). The procedure goes as follows: a random vector θ_k is generated, independent of the past random vectors $\theta_1, ..., \theta_{k-1}$ but with the same distribution; and a tree is grown using the training set and θ_k, resulting in a classifier. The number of trees that can be added to a random forest can increase without a limit, according to Breiman, and will not cause overfitting problems on the model Breiman (2001).

2.2. Artificial Neural Network

The artificial neural network is a widely used model, capable of performing tasks such as classification, pattern recognition and knowledge that is "acquired from its environment through a learning proces´´ and stored in synapses Larose and Larose (2014); Haykin (1999). A multilayered perceptron (figure 3, a type of neural network, is made of at least three layers: an input layer, one or more hidden layers, and an output layer. The input signal goes through the network in a forward direction, on a layer-by-layer basis. An output y from neuron k can be defined as:

$$y_k = \varphi(u_k + b_k) \quad \text{and} \quad u_k = \sum_{j=0}^{m} w_{kj} x_j \tag{2}$$

where w_{kj} is the synaptic weight from input x_j to k, b_k is the bias that influences the input of the activation function φ. The computational power of a multilayer perceptron is due to the fact that there are more than one layer of hidden neurons, with significant connectivity between them through their synapses, facilitating pattern recognition in the network, a crucial component in solving complex problems Haykin (1999).

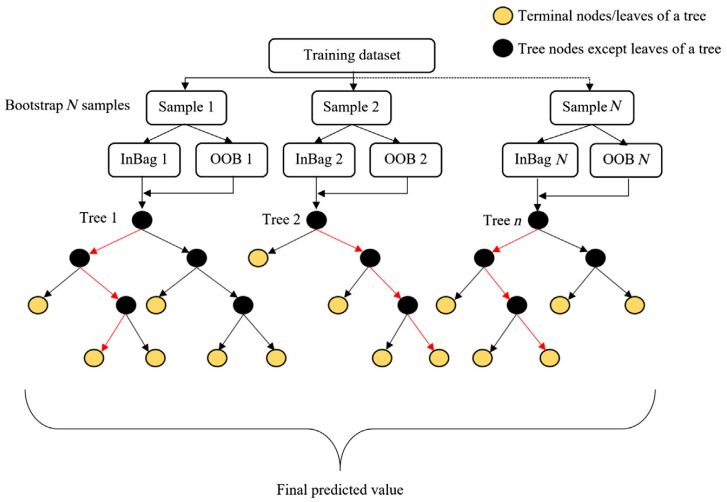


Figure 2: Structure of a random forest Hemmati-Sarapardeh et al. (2020)

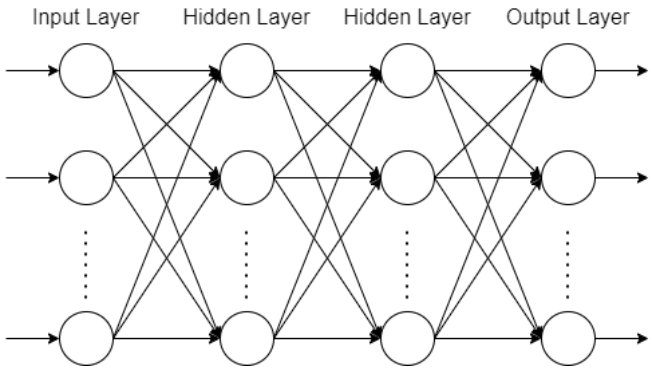


Figure 3: Architectural graph of a MLP

Every neuron located in the network performs two types of functions: the first one is a function signal that propagates forward from the input onto the output, which has had a previous and brief explanation. The second is an error function, that begins in the output neurons, and propagates, layer by layer, in a backward course through the network Haykin (1999). The back-propagation (BP) algorithm represents mathematically how a neural network learns as model of supervised learning, with the propagation of errors in the opposite direction of the output layer, and corrections are made with the synaptic weights, with the delta rule. The delta rule for output layer is computed:

$$\Delta w_{ij} = \eta \delta_i y_i \tag{3}$$

where η is the learning rate and δ_i is the local gradient, a component that indicates to where changes need to happen in the synaptic weight. For a hidden layer, the local gradient has to adapt the corrections made for its synaptic weights and the ones made by the earlier layer. The

correction of the synaptic weights is a recursive computation.

$$w_{ij}^{*} = w_{ij} + \Delta w_{ij} \tag{4}$$

where w_{ij}^{*} is the corrected weight. Depending on the result of the variation of error, the synaptic weights can increase or decrease in order to minimize the error function, with the learning rate helping to adjust the error Larose and Larose (2014).

2.2.1. Hyperparameter Tuning

Hyperparameter tuning can be defined as an optimization problem with the intent of determining those parameters that lead into an optimal value. This pursuit can be computationally expensive and time consuming, specially if a brute force type of search is performed, where all possible data points are verified. Therefore, this hyperparameter tuning problem can be solved through an algorithm designed for its optimization, for example the Bayesian optimization, where it has been tested for machine learning algorithms such as RF and NN with success Wu et al. (2019).

2.3. Active Learning

Active Learning (AL) is a particular sub-environment of machine learning in which an algorithm can choose the data from which it will learn, therefore performing better with less training and less data than a supervised learning algorithm Settles (2009). In practice, from a small and labelled testing data, an AL system will add more data by asking queries from an oracle to label specific data. The goal is to achieve high accuracy from sparse labelled data, minimizing the expense of obtaining these type of data Settles (2009).

One way the learner can ask queries is through a Pool-Based Sampling where the input has a small labelled data, and a larger pool of unlabelled data set is available. Queries are then drawn from the pool in a greedy way, by selecting the best data point from the entire pool. Querying strategy could be uncertainty sampling, a simple framework in which the learner queries data with the least certainty on how to label, or entropy sampling, a more general strategy, in which it tries to map the distribution of probabilities with the information given Settles (2009).

3. Database Modelling

The data used was provided by the Aviation Safety Network (ASN), which is ”a private, independent initiative created in 1996”, that ”covers accidents and safety issues with regards to airliners, military transport planes and corporate jets”, (Ranter and Lujan, 2016). The database given by the ASN has 3242 data points, which are the occurrences suffered by a given aircraft between 2000 and 2020. Each data point has the narrative, causes, contributory factors precluding the occurrence and outcome on the aircraft produced by those failures. In addition to the information provided in the database, there is also other components of particular relevance, for example, damage sustained by the aircraft, phase of flight, if the occurrence was fatal and the degree of mortality. To extract that information, it is required to search in each data point specifically, requiring building an independent database, which is a time consuming process, so a 10 year gap was selected, specifically between 2007 and 2017, and 1105 occurrences were extracted.

In order to proceed with the analysis proposed and to factor the human factors in aviation safety, the taxonomy HFACS needs to be utilized, as the framework proved to be reliable in identifying human factors issues that were hidden, highlight critical parts of human factor failure that required intervention and improved data quality and quantity Wiegmann and Shappell (2001). from the contributory factors prior to the accident a human factors analysis can be performed, relating them with either underlying conditions or probable causes that triggered or may trigger those events Reason (1990). Considering the objectives describe above, it was possible to create two models for analysis considering the database features, as illustrated in figure 4. They try to answer two questions:

1. Is it possible to predict whether an incident or accident produced any fatality? (Model 1)
2. If an occurrence was fatal, is it possible to estimate the percentage of people killed? (Model 2)

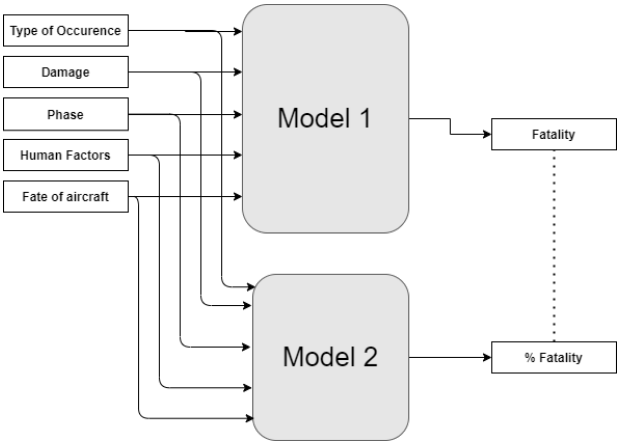


Figure 4: Schema of the database modelling

4. Results

4.1. Performance Criteria

Classification models requires certain parameters to evaluate its validity, and a classifier is only valid if it can predict correctly a label when information is provided. For binary classification problem, the prediction made will belong to one element in a set $\{0, 1\}$ where 0 indicates the negative class, and 1 the positive class. The results can be presented in a confusion matrix, such as in table 1

		True Class	
		0	1
Predicted Class	0	True Negative	False Negative
	1	False Positive	True Positive

Table 1: Confusion Matrix for binary classification

With this confusion matrix, several quality parameters can be determined, such as accuracy, precision, recall, and F1-score:

$$Precision = \frac{TP}{TP + FP} \quad \text{and} \quad Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Accuracy = \frac{TP + TN}{Total} \quad \text{and} \quad F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{6}$$

where where β is a positive real factor. β can be translated as the amount of times recall is more important than precision. If $\beta = 1$ then both criteria achieve a harmonic mean and the F-score is deemed balanced. If a model has class imbalances, the Receiver Operating Characteristic (ROC) and Precision-Recall graphs can be used for criteria. ROC curve is defined as a plot of the False Positive Rate (FPR) on the x-axis and True Positive Rate (TPR) on the y-axis.

4.2. Models Results

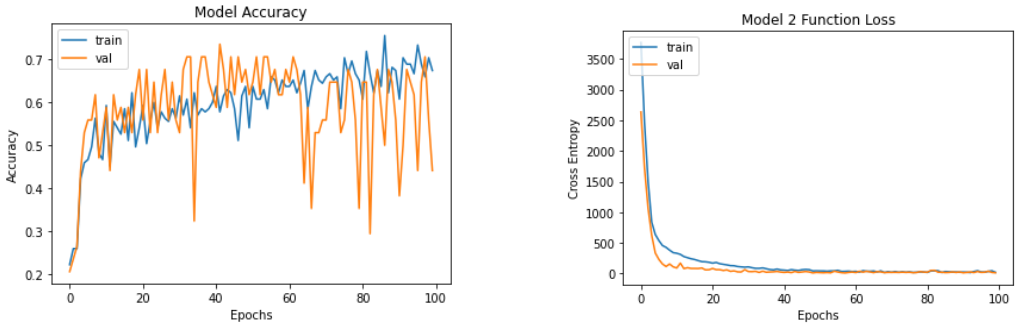
For Model 1, a single-class binary output type of model was considered. To apply the supervised learning algorithms, 75% of the data set was used for training, and 25% for testing to validate them. After both algorithms are trained, a confusion matrix with the validation set is obtained for the respective algorithm implementation, and the respective criteria performances are determined.

Type of Learning			Precision	Recall	F1-score	Accuracy
Random Forest	Class	No Fatality	0.92	0.94	0.93	0.90
		Fatality	0.84	0.77	0.80	
	Averages	Macro	0.88	0.86	0.87	
		Weighted	0.89	0.90	0.89	
Multilayer Perceptron	Class	No Fatality	0.89	0.92	0.90	0.83
		Fatality	0.75	0.59	0.66	
	Averages	Macro	0.80	0.76	0.77	
		Weighted	0.83	0.83	0.83	

Table 2: Classification Report for Model 1

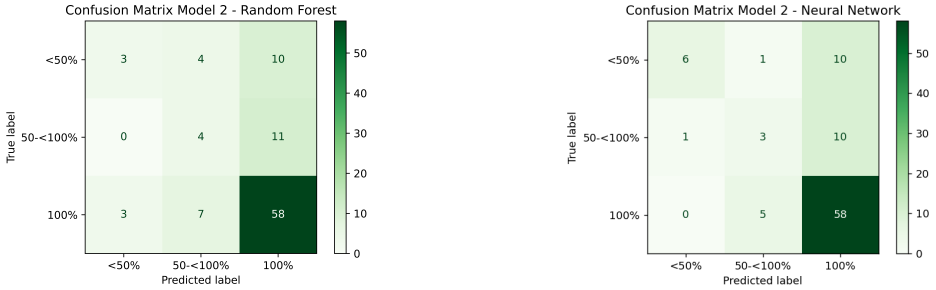
The validity of the model can be confirmed by the MLP's binary cross entropy function, which is computed between true and predicted labels. Its behavior throughout the epochs can help determine the fit of data set, which helps shaping its structure. Both training and validation curves have an exponential decay over the first epochs, which means the MLP finds quickly a good fit and does not seem to overfit, as the validation cross entropy curve does not seem to increase and the training curve decreases. From table 2, the RF is better at correctly predicting the positive class the MLP. The semi-supervised was used, with a pool-based strategy, but did not improve from the other algorithms.

For Model 2, there is less quantity of data since only the accidents that produced fatalities are considered, which means for an algorithm such as MLP, there are limitations with quantity of data that error function for first epochs is very high (figure 5(b)). Despite the good shape of the loss function, the loss decay of the training data set is severe, which makes it difficult to judge the behavior of both curves as the epochs progress, regarding overfit or underfit of the model. With the accuracy function, there can be more understanding on the evolution of MLP with more training (figure 5(a))



(a) Accuracy (b) Loss

Figure 5: Confusion Matrix for Model 1 with RF algorithm (a) and MLP algorithm(b)



(a) Random Forest (b) Neural Network

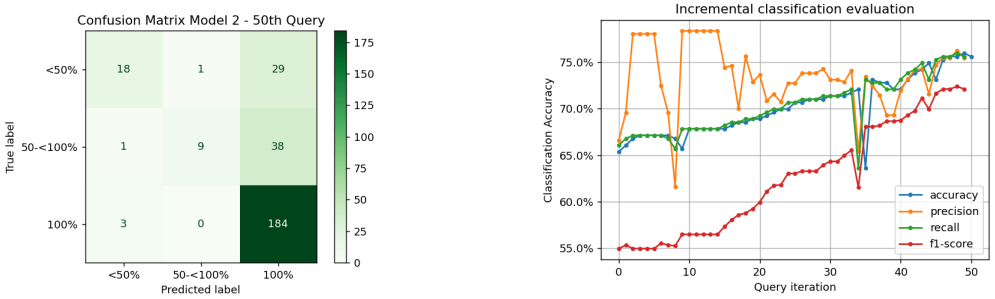
Figure 6: Confusion Matrix for Model 2 with RF algorithm (a) and MLP algorithm(b)

It was also used the semi-supervised learning algorithm Active Learning (AL) for Model 2, with the pool-based strategy. The labelled data selected was 1% of the total data available, with the rest being used as pooling data. The query selection was the entropy one because of the labels skewed distribution towards one class. After the last query, the confusion matrix can be computed. With the constant retraining of data where new testing data is added for the AL algorithm, the prediction can be improved (figure 7).

The AL is capable to perform better than the RF when it comes to predicted correctly labels, as seen with table 3 where AL has better performance across all criteria.

Type of Learning	**Classes**	Precision	Recall	Macro f1-score
Random Forest	Below 50%	0.43	0.18	0.41
	50% -<50%	0.21	0.21	
Active Learning	Below 50%	0.75	0.38	0.72
	50% -<50%	0.75	0.31	

Table 3: Comparison between RF and AL for predicting sparse labels for Model 2



(a) Confusion Matrix

(b) Incremental Evaluation

Figure 7: Results for Model 2 with AL after 50 queries performed with confusion matrix (a) and MLP algorithm(b)

5. Conclusions

In this article, a modelling system for the aviation industry based with data from incident and accident reports, was proposed. he main objective in most systems of this industry area is to improve the safety standards of a highly regulated industry. Since there is more work done regarding the role of human factors in safety procedures for aviation, a correlation was made with those contributory causes and human factors by applying the HFACS taxonomy. After those models were created, three algorithms (MLP,RF and AL) were proposed based upon previous work with the ASN database.

With the use of all techniques, it was proven that both models were capable to learn from the data given, especially for Model 2, where there was a less quantity of data. For Model 1, RF was capable of producing better results than MLP across all metrics, especially when predicting correctly a fatal accident instead of non-fatal. For Model 2, with imbalanced classes and lower class, the RF and MLP struggled with predicting those classes however RF did perform better. With AL though, after a run of querying, the algorithm was able to improve significantly from the supervised learning methods.

References

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Dias, N.G., Santos, L.F., Melicio, R., 2019. Aircraft maintenance professionals: Stress, pressure and fatigue, in: MATEC Web of Conferences, EDP Sciences. p. 06001.

Haykin, S., 1999. Neural networks: a comprehensive foundation. Prentice hall. chapter 1,4. pp. 23–,178–278.

Hemmati-Sarapardeh, A., Larestani, A., Nait Amar, M., Hajirezaie, S., 2020. Intelligent models, in: Applications of Artificial Intelligence Techniques in the Petroleum Industry. Gulf Professional Publishing. chapter 2, pp. 23–50.

ICAO, 2013. Safety Management Manual. third ed. Montreal. DOC 9859 AN/474.

Larose, D.T., Larose, C.D., 2014. Discovering knowledge in data: an introduction to data mining. 2nd ed.. John Wiley & Sons. chapter 8–9. pp. 165–208.

Madeira, T., Melício, R., Valério, D., Santos, L., 2021. Machine learning and natural language processing for prediction of human factors in aviation incident reports. Aerospace 8, 47.

Mazareanu, E., 2021. Passenger air traffic each year. URL: https://www.statista.com/statistics/564717/airline-industry-passenger-traffic-globally/.

Qiao, W., Liu, Y., Ma, X., Liu, Y., 2020. A methodology to evaluate human factors contributed to maritime accident by mapping fuzzy FT into ANN based on HFACS. Ocean Engineering 197, 106892.

Ranter, H., Lujan, F.I., 2016. Aviation Safety Network. URL: https://aviation-safety.net/about/. (accessed: 03.02.2021).

Reason, J., 1990. Human error. Cambridge University Press.
Santos, L.F., Melicio, R., 2019. Stress, pressure and fatigue on aircraft maintenance personal. International Review of Aerospace Engineering .
Settles, B., 2009. Active Learning literature survey .
Shappell, S.A., Wiegmann, D.A., 1996. US naval aviation mishaps, 1977-92: differences between single-and dual-piloted aircraft. Aviation, space, and environmental medicine 67, 65–69.
Shappell, S.A., Wiegmann, D.A., 2000. The human factors analysis and classification system–HFACS. Embry-Riddle .
Wiegmann, D.A., Shappell, S.A., 2001. A human error analysis of commercial aviation accidents using the human factors analysis and classification system (HFACS). Technical Report. Office of Aviation Medicine, FAA.
Wiegmann, D.A., Shappell, S.A., 2003. A Human Error Approach to Aviation Accident Analysis. Ashgate Publishing Ltd.
Wu, J., Chen, X.Y., Zhang, H., Xiong, L.D., Lei, H., Deng, S.H., 2019. Hyperparameter optimization for machine learning models based on bayesian optimization. Journal of Electronic Science and Technology 17, 26–40.