Improving the Classification of Airplane Accidents Severity using Feature Selection, Extraction and Machine Learning Models

Rachid KAIDI[1], Mohammed AL ACHHAB[2], Mohamed LAZAAR[3], Hicham OMARA[4]
ENSA, Abdelmalek Essaadi University, Tetouan, Morocco[1,2]
ENSIAS, Mohammed V University, in Rabat, Morocco[3]
FS, Abdelmalek Essaadi University, Tetouan, Morocco[4]

***Abstract*—Airplane mode of transportation is statistically the most secure means of travel. This is due to the fact that flights require several conditions and precautions because aviation accidents are most of the time fatal and have disastrous consequences. For this purpose, in this paper, the mean goal is to study the different levels of fatality of airplane accidents using machine learning models. The study rely on airplane accident severity dataset to implement three machine learning models: KNN, Decision Tree and Random Forest. This study began with implementing two features selection and extraction methods, PCA and RFE in order to reduce dataset dimensionality and complexity of models and reduce training time by implementing machine learning models on dataset and measuring their performance. Results show that KNN and Decision Tree demonstrates high levels of performances by achieving 100% of accuracy and f1-score metrics; while Random Forest achieves its best performances after application of PCA when it reaches an accuracy equal to 97.83% and f1-score equal to 97.82%.**

***Keywords*—Airplane accident; severity; flights safety; machine learning; KNN; Random Forest (RF); Decision Tree (DT)**

I. Introduction

It is well known that the plane is the safest mode of transportation. In 2022, there have been only two fatal plane crashes without counting small and helicopter crashes [1]. In 2007, The National Transportation Safety Board known as NTSB, claimed that 24 million hours of plane travels had occurred, only 6.84 of every 100.000 flights hours had a plane accident and 1.19 of every 100.000 plane flights are fatal crashes. The aviation industry contribute significantly to the national economic of each country if they have robust and strong aviation policies and technologies. Thousands of incomes in this field can be reached every year; for this purpose, this industry is well developed and controlled by international standards. Among the most important requirements in aviation industry is safety. That is why different measures of security are taken into consideration before and after any flight. These requirements include:

- A strict safety requirements based of international standards in order to establish a base to rate degree of safety in any flight.
- Collecting data in order to perform data analytics to find out any shortcoming and to perform safety improvements.
- Continuous extensive training for pilots to update their knowledge and skills.
- Safety Management System (SMS) should be implemented in any plane in order to have synchronous state of the plane.
- Auditing safety measures by investigating incidents, analysing performances indicators, etc.

The most important requirement for us in this study is the use of data in order to perform safety audit, prevent accidents and rectify any breaches of policies or technologies misconfigurations or malfunctions. As we can say, safety requirements are very hard and complex to implement because any small error or misconfigurations may lead to fatal consequences. According to [13] airplane accidents can be caused by so many factors including: Pilot error due to miscommunication, distraction, exhaustion, drainage, etc. mechanical error, bad weather conditions, sabotage and human errors. So many scientific contributions had tried to implement data0based approaches using machine learning (ML) models to deal with these safety issues specially in the context of our study which is the prediction of severity of airplane incidents. [4,6,7,12,13] propose ML-based models, deep learning are used and implemented on complex dataset that need deep models such as in [11,14]. Authors using machines learning (ML) models achieved promising results but there is always some data and implementations constraints including limited resources and information about fatal accidents because there are very rare to occur; that is why it is difficult to collect enough data to establish meaningful statistical analysis. The factors that control the operations of an airplane are numerous and complex; they include environmental factors, techniques, human resources factors. Data may be biased toward the condition and purposes so it can misrepresent some conditions and important factors. Mathematical representation of severity may be very difficult for modelisation using classical approaches such as text mining [2]. Hence the need for machine learning solutions. Emre Kuşkapan et al. [8] propose an approach for aviation accidents classification using data mining algorithms. They collect data worldwide from 2000 to 2019. They implement J48, Naïve Bayes and Sequential Minimal Optimization methods. J48 outperforms all methods based on Precision, Recall, F-measure and ROC Area. L. J. Raikar et al. [3] implement SVM, KNN, Adaboost, XGboost to analyse airplane crash. They include feature selection and scaling methods in order to reduce di-

mensionality of data by removing unnecessary characteristics. In this study, we will implement machine learning models on airplane accident severity dataset to predict severity of airplane accidents. Before that, we established a robust phase of feature selection and extraction in order to get the most relevant and important features. These features will be the focus of future work. We get interesting results. Some ML models reached 100% of accuracy using KNN and Decision Tree. The rest of this paper is organised as follows: In Section 2, we will discuss the related work, in Section 3, we will present the background of this study and the followed methodology. In Section 4, we will present our results, discuss and criticise them and finally a conclusion where we will mention the relevant results, limits of this study and its perspectives in future work.

II. Related Work

R. A. Burnett et al. [6] implement machine learning models in order to predict the injuries and fatalities in aviation accidents. They face so many problems in data processing. First of all, they needed to deal with redundant fields, missing values, lack of generalisation which means that there are lot of changes of conditions over years, so implementing machine learning models on old statistics may lead to misrepresent the results. Another raised problem is that they needed to deal with it is imbalanced data which is a very complex task in machine learning context. They used six Federal Aviation Administration Aviation Incidents and accidents records in range from 1975-2002. They Implement KNN, SVM and KNN models to predict the rate of aviation accidents. Results show that ANN gives a promising results and its obvious because ANN models can generalise and analyse internal relationships between features in order to extract pattern more better than regular machine learning models.

It is well know that in any information system, the human being is the most vulnerable asset. Human factor contribute to approximatively 75% of aircraft accidents and incidents [5]. In the paper by M. Bagarzan et al. [7], they conducted an interesting study in order to analyse the impact of the age, experience and gender of pilots on aviation accidents. They specify six categories of age: “less than 20”, “20-29”, “30-39”, “40-49”, “50-59” and “more than 60”. “Experience” in this study is based on the number of hours for each pilot involved in an accident. the records belong to the NTSB database. They implement chi-square and logistic regression models in their study in order to figure out if there is any relationships between pilot characteristics and how much these characteristics can contribute in causing aviation accident that lead to serious consequences. Results show that the gender had no great impact of the pilot error but females make fatal errors less than men. Pilots that are older than 60 years old can make pilot error. Experienced pilots can easily get involved in fatal accidents because due to their experiences they can fly in conditions that non experienced pilots cannot do but these experienced pilots are less likely to make pilot errors. Authors suggest that there are environment conditions that can affect the performances of a pilot and they suggest to maintain training for pilots and improve performances of technologies used in this mode of transportation.

In the study by N. Pande et al. [13], they conducted a study about the prediction of fatal aviation accidents. They used Random Forest, XGBoost, Neural Network, multiple Linear Regression, chi-square, linear regression, ensemble model that combine so many machine learning models and logistic regression. This study is based on a data that is collected since 1908. 4700 data points of it were used in this study. What is interesting about this study is that it is based on simple classical machine learning models like RF, XGBoost and complex machine learning models that refers to the combination of different machine learning models in order to predict the dependant feature. The evaluation metrics in this study include minimum error, maximum error, mean absolute error, linear correlation and standard deviation. Results show that Neural Network model outperform all of the implemented models reaching an accuracy equal to 90.6% which is the case for [13].

We can say that machine learning models are widely used in the field of severity and fatalities of flights accidents. The usefulness of machine learning models in this field are related to the complexity of data. In order words, same conditions of weather and other plane characteristics that we can gather using sensors, the same data can lead to different results. We can never be sure of these characteristics because even a bird strike [9] can cause fatal damage.

III. Theoretical Background

In order to implement our ML models, we will rely on two models for data processing: PCA for feature extraction and RFE for features selection. Then we will implement our ML models that are: RF, DT and KNN.

A. Feature Extraction: PCA

It is an abbreviation for Principal Component Analysis [19]. It is an unsupervised ML model that is used to reduce data dimensionality based on statistical measurements, and generates new components to contain the most significant feature data by capturing a large amount of variance[20]. PCA is used to transform a large set of data into a small one but keeping relevant information about data. The principal components of PCA are orthogonal. PCA is useful to reduce the noise in data, to compress data and helps in visualising data with high dimensions and to detect any relationships between features mainly correlation and other relationships in order to gather correlated features with each others. Given a dataset X with n observations and p variables, we can perform PCA by following these steps:

Center the data by subtracting the mean $\bar{x}$ from each variable:

$$\tilde{x}ij = xij - \bar{x}j \tag{1}$$

where $\tilde{x}ij$ is the centered value of variable j in observation i.

Calculate the sample covariance matrix S:

$$S = \frac{1}{n-1}\sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T \tag{2}$$

where $\tilde{x}_i$ is the centered vector of observation i.

Compute the eigenvectors $v_1, v_2, ..., v_p$ and corresponding eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ of S.

Select the k eigenvectors with the highest eigenvalues, where k is the number of dimensions in the reduced feature space.

Project the centered data onto the k selected eigenvectors to obtain the reduced feature space:

$$Y = XV_k \tag{3}$$

where V_k is a matrix containing the top k eigenvectors as columns.

The final output Y will have dimensions $n \times k$, where each row represents an observation and each column represents a principal component.

B. Feature Selection: RFE

Recursive Feature Elimination is a feature selection method used to get the most important features that contribute to improve ML models performances. RFE is a recursive method that removes in each iteration the worst features. We first provide RFE with all features, RFE run on ML models on the dataset for instance Random Forest. After training the model -ML model- RFE measures the contribution and importance of each feature. Less relevant features will be removed and re-run the model until we get the best number of features that contribute the most to the model. The process of Recursive Feature Elimination (RFE) involves assigning weights to different features in order to identify which ones contribute the most towards predicting the target variable. This is done by ranking the features based on their relative importance [21], which will help to decrease complexity of the model, minimising time of training and increasing model performances. The RFE algorithm can be mathematically represented as follows: Let X be the feature matrix and y be the target vector. Let n be the total number of features and k be the desired number of features.

1) Initialize $X_{\text{RFE}} = X$ and $k_{\text{RFE}} = n$.
2) Train a model on X_{RFE} and y to obtain coefficients or feature importance.
3) Calculate the importance of each feature
4) Remove the least important feature from X_{RFE} to obtain X'_{RFE} and decrement k_{RFE}.
5) we repeat the same steps until $k_{\text{RFE}} = k$.

C. Decision Tree

It is a supervised ML model used for classification and regression.DT are not only highly useful in various applications but also renowned for their interpretability and resilience [16] . DT split the data based on features with most importance. These features importance is measured using Gini index or Entropy. The final structure of a DT model is a tree, where nodes represent features and leafs represent the dependant feature. The process of a DT model is as follow: The dataset went through DT from input node to the leafs where each leaf refer to a value of the dependant feature. Unlike RF, DT may have bad performances in front of high dimensionality datasets or noisy datasets.

To construct a Decision Tree, the ID3[15] algorithm is utilized, which involves several steps. The initial stage involves computing the entropy or Gini impurity of the target class in order to assess the data's impurity.

$$Gini(S) = 1 - \sum_{i=1}^{k} p_i^2 \tag{4}$$

Where k represents the total number of classes, while p_i denotes the proportion of instances that belong to the i-th class.

Afterward, we calculate the Gini gain for each attribute in our dataset and select the attribute that provides the greatest value and create a node for that attribute. The formula used to calculate the Gini gain for each attribute is:

$$Gini\_Gain(S, A) = Gini(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Gini(S_v) \tag{5}$$

Where A represents an attribute, while S refers to the dataset. S_v represents the subset of instances in S where attribute A has a value of v. Repeat steps 2 recursively for every subset of data generated by the split until all instances within a subset are categorized under the same class or there are no remaining attributes left to split the data.

D. Random Forest

Random forest [17] is a supervised machine learning model used for many purposes including classification and regression. It is based on building Decision trees on subset of the samples and features. RF contains n DT. The choice of n depends on the task. Each DT is trained on a random part of the data using random partitions that helps to decrease the model's complexity and to prevent it from overfitting. Each DT makes a prediction and then the majority vote will be considered for prediction. RF gives best results on large data of high dimensions and RF is very useful in case of noisy data.

E. KNN

Abbreviation of K-Nearest Neighbors. It is a non-parametric supervised machine learning model [18], mainly used for classification and regression tasks. The purpose of KNN is to find the K nearest data points of data in order to make a prediction. KNN measures the distance between data points using Euclidean distance:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{6}$$

Or Manhattan distance:

$$d(X, Y) = \sum_{i=1}^{n} |X_i - Y_i| \tag{7}$$

These distances are used to group subsets of data in order to measure the dependant variable in the training phase. In case of large datasets, KNN may need additional computational resources. The choice of the k is not trivial, choosing a wrong k will lead to performances degradation.

F. Accuracy

Accuracy is a commonly used metric in machine learning to evaluate the performance of a model. It involves counting the number of true positive (TP) and true negative (TN) samples in a given dataset, and dividing it by the total number of samples including false positive (FP) and false negative (FN) samples. In other words, accuracy measures how many samples are correctly predicted out of all samples in the dataset.

$$Accuracy = \frac{TP + TN}{TP + TF + FP + FN)} \tag{8}$$

G. F1-Score

The F1-score is a metric used to measure the classification performance of machine learning models. It combines two different metrics, precision and recall, to provide a single score that reflects the overall performance of the model. A good F1-score requires good results for both precision and recall, or high results for one metric if the other metric has low results, in order to balance the results.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{9}$$

IV. Results and Discussion

Table I shows the results of KNN, DT and Random Forest using all features. Results show that KNN and DT reached 100% of performances for both metrics: accuracy and f1-score. RF reaches an accuracy equal to 95.48% and f1-score equal to 95.49%.

TABLE I. KNN, DT and RF ML Models Performances using All Features based on Accuracy and f1-score

	Accuracy	F1-Score
Random Forest	95.48%	95.49%
Decision Tree	100%	100%
KNN	100%	100%

In order to reduce the dimensionality of the dataset, we used PCA with Principal Components equal to 8. Results in Table II show that results remain the same for KNN and DT, which means that PCA preserved relevant inertia of original dataset while reducing complexity of the models and training time. For RF model, we find out that RF performances increase by 2.35% to reach 97.83% of accuracy. Same remark for f1-score metric, it increase by 2.33% to reach 97.82%.

TABLE II. KNN, DT and RF ML Models Performances after Applying PCA based on Accuracy and f1-score

	Accuracy	F1-Score
Random Forest	97.83%%	97.83%
Decision Tree	100%	100%
KNN	100%	100%

Table III shows the performances of ML models after selecting the most important features based on RFE feature selection metric. As for PCA metric, the performances of KNN and DT remain the same (100% of accuracy and f1-score for both models). For RF model, accuracy decreased by 0.71% to become 94.77%. Same behaviour for f1-score, it decreased by 0.73% to become 94.76%.

TABLE III. KNN, DT and RF ML Models Performances after Applying RFE based on Accuracy and f1-score

	Accuracy	F1-Score
Random Forest	94.77%	94.76%
Decision Tree	100%	100%
KNN	100%	100%

Fig. 1 shows the confusion matrix of RF model using all features of the dataset. It shows that the accuracy of multi-classification is in the range [94%-97%]. Fig. 2 and Fig. 3 show respectively the multi-classification for both KNN and DT, each class from 0 to 3 are well classified which means that these two models can be a good choice to deploy them as real classifiers for predicting severity of aviation accidents.

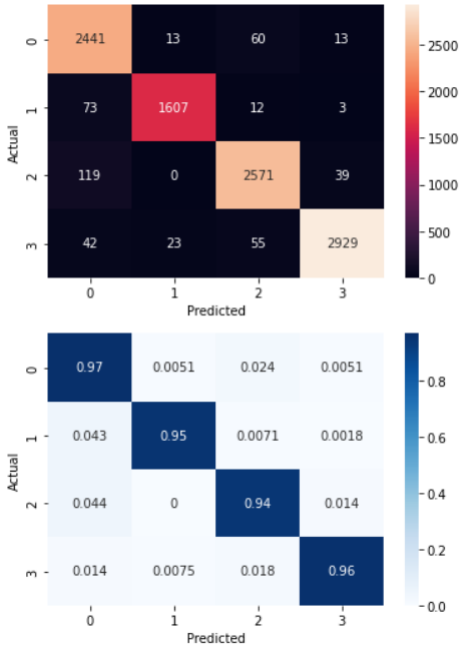

Fig. 1. Multi-classification results of RF model using all features.

Fig. 4 shows the multi-classification performances of RF after application of PCA. We can remark that the range of accuracy for the four classes is [97%-99%]. We also remark that comparing with results using all features, we can say that accuracy augmented for all classes after applying PCA by 2%, 4%, 3% and 1% for respectively class 0, class 1, class 2 and class 3.

Fig. 5 and Fig. 6 show the performance of both KNN and DT. Results remain the same, each class reach 100% of accuracy for the four classes for both KNN and DT; results same the same comparing them with results obtained after application of PCA. Fig. 7 shows the performances of RF model after application of RFE feature selection based method. Results show a decrease of performances by 1%, 1%, 1% for respectively class 0, class 1 and class 2, while the accuracy of class 3 remain the same. Fig. 8 and Fig. 9 show performances of both KNN and DT after application of RFE. Results remain the same such in the two cases (after application of PCA & using all features).

After discussing all these results, we can say that RF,

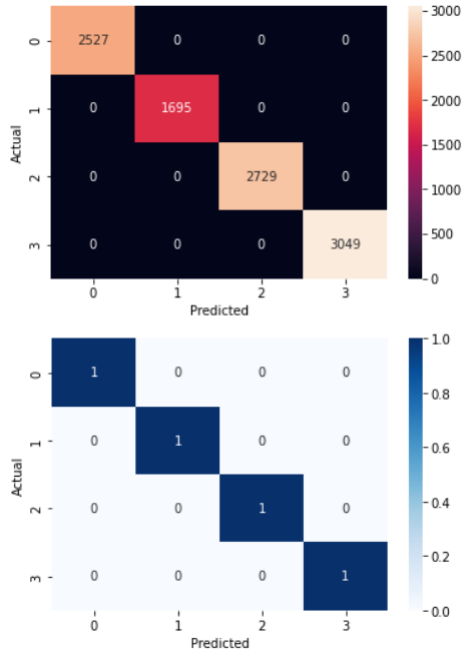

Fig. 2. Multi-classification results of KNN model using all features.

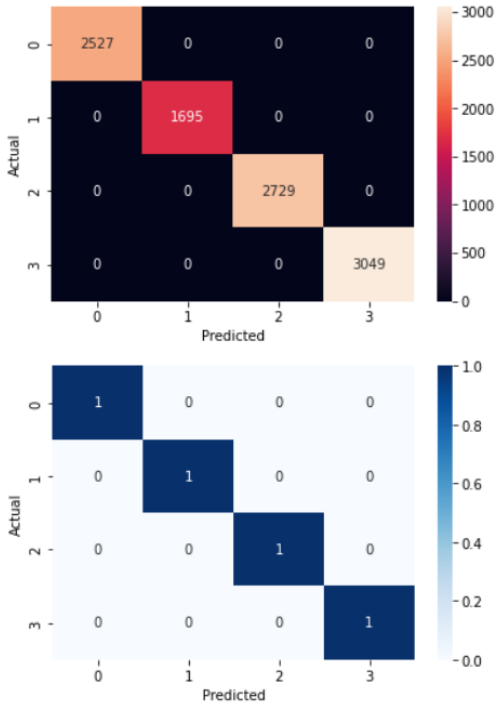

Fig. 3. Multi-classification results of DT model using all features.

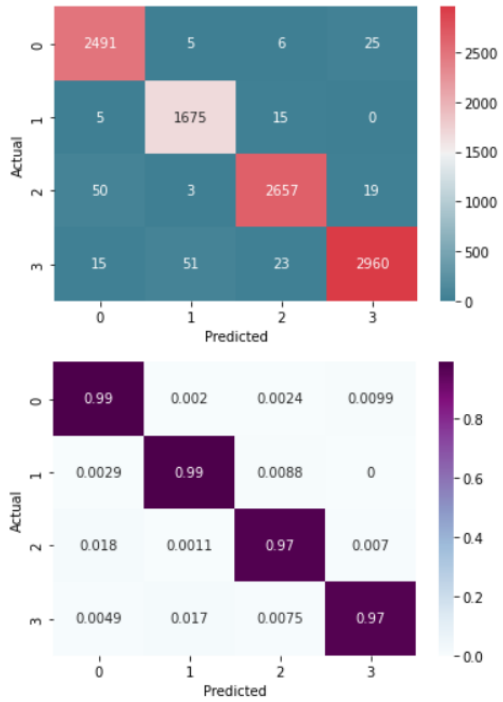

Fig. 4. Multi-classification results of RF model after application of PCA.

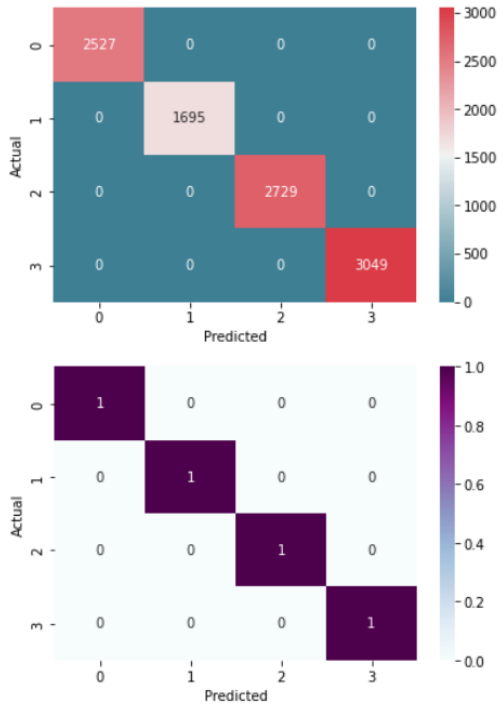

Fig. 5. Multi-classification results of KNN model after application of PCA.

KNN and DT are promising models that can deployed in real situations for predicting severity of aviation accidents. KNN and DT give an accuracy and f1-score equal to 100% which means that all classes of severity have been adequately classified so we can say that intelligent solutions based on ML or Deep Learning models can be relevant alternatives to overcome the limits of classical solutions such as statistical analysis, solutions based on expertise, solutions that require sometimes advanced mathematical modelisation that are complex and may lead sometimes to misrepresent real conditions in the phase of abstraction.

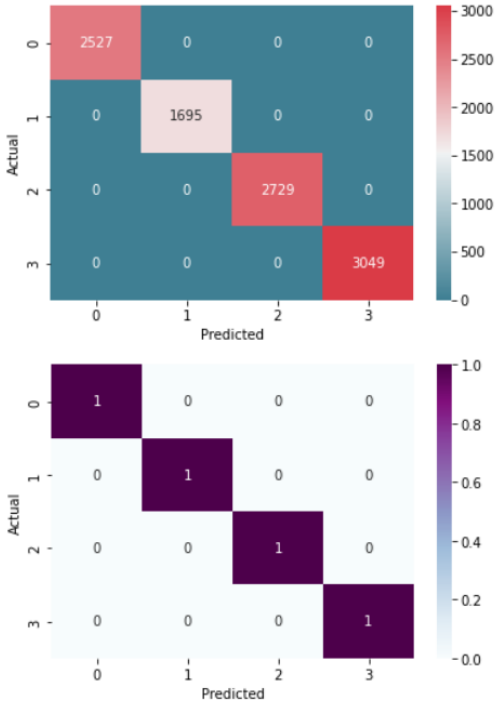

Fig. 6. Multi-classification results of DT model after application of PCA.

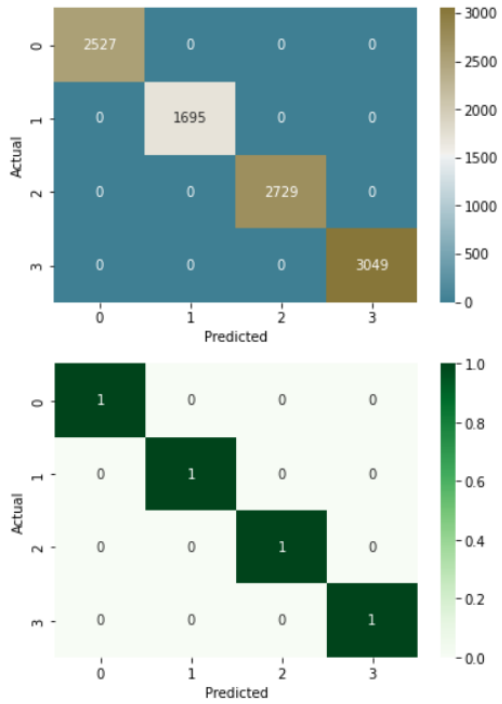

Fig. 8. Multi-classification results of KNN model after application of RFE.

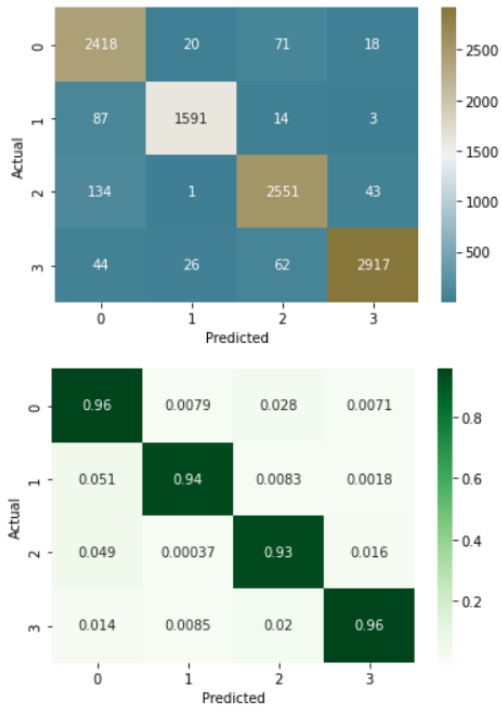

Fig. 7. Multi-classification results of RF model after application of RFE.

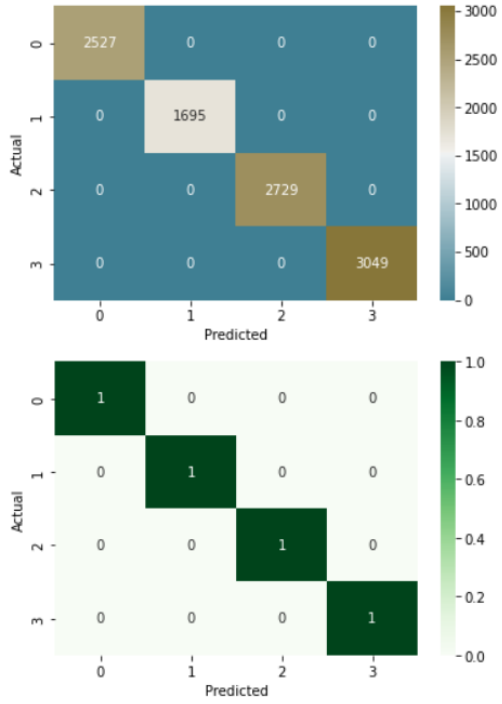

Fig. 9. Multi-classification results of DT model after application of RFE.

V. Conclusion and Future Work

This paper shows the robustness of machine learning models for predicting severity of aviation accidents. KNN and DT achieved high level of performances based on accuracy and f1-score metrics while RF gives respectful results but never achieved 100% of accuracy or f1-score but we should pay attention that these models should be tested in real situations to test its stability. Furthermore, as we had discussed earlier, Neural Network and Deep Learning models can be a good alternatives of the models we had implemented in this study.

This proposition must be the perspective of this data and test our models on other datasets that contain different information and then different conditions. Predictions models in field of aviation are a very complex tasks that require gathering the maximum possible of information about environment plane, human, technologies, etc. in order to improve intelligent solutions like ML and DL models to overcome limits of classical solutions.

REFERENCES

[1] Aviation and Plane Crash Statistics (Updated 2022). (n.d.). Panish — Shea — Boyle — Ravipudi LLP. Retrieved April 8, 2023, from https://www.psbr.law/aviation_accident_statistics.html.

[2] BAUGH, Bradley S. Predicting general aviation accidents using machine learning Algorithms. Embry-Riddle Aeronautical University, 2020.

[3] RAIKAR, Likita J., PARDESHI, Sayali, et SAWALE, Pritam. Airplane Crash Analysis and Prediction using Machine Learning. in International Research Journal of Engineering and Technology (IRJET), vol. 7, no 03,2020.

[4] A.O. Alkhamisi, R. Mehmood, "An ensemble machine and deep learning model for risk prediction in aviation systems", Conf. Data Sci. and Mach. Learn. Appl., Vol. 2020, No. 6, pp. 54-59, Mar. 2020.

[5] H. Kharoufah, J. Murray, G. Baxter, G. Wild, "A review of human factors causations in commercial air transport accidents and incidents: From 2000-2016", Prog. in Aerospace Sci., Vol. 99, pp. 1-13, 2018.

[6] BURNETT, R. Alan et SI, Dong. Prediction of injuries and fatalities in aviation accidents through machine learning. In : Proceedings of the International Conference on Compute and Data Analysis. 2017. p. 60-68.

[7] M. Bazargan, V.S. Guzhva, "Impact of gender, age and experience of pilots on general aviation accidents", Accid. Anal. & Prev., Vol. 43, No. 3, pp. 962-970, 2011.

[8] Kuşkapan, Emre, SAHRAEİ, Mohammad Ali, et Çodur, Muhammed Yasin. Classification of aviation accidents using data mining algorithms. Balkan Journal of Electrical and Computer Engineering, vol. 10, no 1, p. 10-15, 2021.

[9] NIMMAGADDA, SreeRam, SIVAKUMAR, Soubraylu, KUMAR, Naveen, et al. Predicting airline crash due to birds strike using machine learning. In : 2020 7th international conference on smart structures and systems (ICSSS). IEEE. p. 1-4, 2022.

[10] Airplane Accidents Severity Dataset https://www.kaggle.com/datasets/kaushal2896/airplane-accidents-severity-dataset.

[11] Y. Guo, Y. Sun, Y. He, F. Du, S. Su, C. Peng, "A Data-driven Integrated Safety Risk Warning Model based on Deep Learning for Civil Aircraft", IEEE Trans. on Aerospace and Electronic Systems, 2022, pp. 1-14.

[12] ZHANG, Xiaoge et MAHADEVAN, Sankaran. Ensemble machine learning models for aviation incident risk prediction. Decision Support Systems, vol. 116, p. 48-63, 2019.

[13] PANDE, Nikita, GUPTA, Devyani, SHREEMALI, Jitendra and CHAKRABARTI, Prasun. Predicting Fatalities in Air Accidents using CHAID XG Boost Generalized Linear Model Neural Network and Ensemble Models of Machine Learning. Vol. 9 , 30 March 2020.

[14] ZHANG, Xiaoge, SRINIVASAN, Prabhakar, et MAHADEVAN, Sankaran. Sequential deep learning from NTSB reports for aviation safety prognosis. Safety science, vol. 142, p. 105390, 2021.

[15] QUINLAN, J.. Ross . Induction of decision trees. Machine learning, vol. 1, p. 81-106, 1986 .

[16] COSTA, Vinícius G. et PEDREIRA, Carlos E. Recent advances in decision trees: An updated survey. Artificial Intelligence Review, vol. 56, no 5, p. 4765-4800, 2023.

[17] BREIMAN, Leo. Random forests. Machine learning, vol. 45, p. 5-32, 2001.

[18] COVER, Thomas et HART, Peter. Nearest neighbor pattern classification. IEEE transactions on information theory, vol. 13, no 1, p. 21-27, 1967.

[19] HOTELLING, Harold. Analysis of a complex of statistical variables into principal components. Journal of educational psychology, vol. 24, no 6, p. 417, 1933.

[20] GÁRATE-ESCAMILA, Anna Karen, EL HASSANI, Amir Hajjam, et ANDRÈS, Emmanuel. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, vol. 19, p. 100330, 2020.

[21] KANNARI, Phanindra Reddy, CHOWDARY, Noorullah Shariff, et BIRADAR, Rajkumar Laxmikanth. An anomaly-based intrusion detection system using recursive feature elimination technique for improved attack detection. Theoretical Computer Science, vol. 931, p. 56-64, 2022.