PDA-3

Name: Ahmed Khawar

Azka Falak Sher

Reg No#: L1F20BSSE0558

L1F20BSSE0624

Subject : BDA(Big Data Analytics)

Submitted to: Ma’am Mubashera anwar

# **Analysis Techniques:**

## In this particular scenario, I opted for a classification technique over regression. The rationale behind this decision lies in the nature of regression, which is typically employed for predicting continuous values. Regression assumes a linear relationship between variables, and while it might appear plausible for the task at hand, there is a concern that it could overlook other independent variables if not applied judiciously. Striking the right balance in regression might be challenging, and using a classification approach provides a more intuitive and robust solution, mitigating the risk of neglecting crucial independent variables.

## **The reason behind using the classification**

The rationale for choosing classification in this context is that we are examining a variable to determine whether the airline is likely to crash or not. The task inherently involves making a binary judgment, and using a classification approach aligns more logically with the nature of this specific prediction task.

# **Model Selection**

The Decision Tree Classifier was chosen for its interpretability, ability to handle diverse data types, and quick training times. Its transparent decision rules make it suitable for tasks where interpretability is crucial. With the capability to capture non-linear relationships and provide feature importance, the model serves as a baseline for comparison with more complex algorithms. This choice aligns with task requirements, emphasizing transparency and insights into feature contributions.

# **Parameter Tuning**

Setting constraints on parameters such as the maximum depth of the tree and the minimum number of leaf nodes is crucial as it helps prevent potential issues of overfitting or underfitting in the model. These constraints serve to control the complexity of the Decision Tree, ensuring that it does not become overly intricate (leading to overfitting) or excessively simple (resulting in underfitting). By defining limits on parameters, we aim to strike a balance that allows the model to generalize well to unseen data, enhancing its overall robustness and performance.

# **Results**

The Decision Tree model achieved an overall accuracy of approximately 30.16%. The weighted precision, reflecting the accuracy of positive predictions, is 58.91%, and the weighted recall is 30.16%. The confusion matrix provides detailed insights into true positive, true negative, false positive, and false negative predictions for each class. The classification report presents precision, recall, and F1-score for individual classes, indicating varying performance across different categories. Further investigation and potential model improvement strategies, such as hyperparameter tuning, may be considered, especially for classes with lower precision and recall values.

# **Visualization Choices**

The model's decision-making process, feature importance, and class distribution were all clearly discernible from the decision tree visualization, which was selected for its interpretability.

# **Conclusion**

In conclusion, the analysis involved cleaning and transforming dataset followed by building and evaluating a decision tree classifier for multiclass classification. The decision tree model demonstrated an accuracy of approximately 29%, with varied precision.Visualization of the decision tree aided in interpreting the model's decision logic.

# **Limitation**

The analysis does, however, have several shortcomings. The models' performance and generalization are impacted by the imbalanced nature of the dataset, where certain classes have few examples. With several classes, the decision tree visualization could get complicated and difficult to understand.

# **Future work**

Future research could enhance model performance by investigating more sophisticated algorithms or resampling strategies to overcome class imbalance. Furthermore, a better comprehension of the data and feature engineering may improve prediction abilities. In this case, additional research into ensemble techniques or neural networks may provide different strategies for multiclass categorization.