**Final Project\_ Airliner\_Safety\_Analysis**

1. **Data Descrption**

The "Airline Safety" dataset provides a comprehensive overview of various safety metrics related to different airlines, with a particular focus on incidents, fatal accidents, and fatalities. The data spans over three decades, from 1985 to 2014, and is intended to facilitate an exploration into the historical safety records of airlines. This dataset was utilized in the analysis presented in the story "Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?".

Source:

https://www.kaggle.com/datasets/danoozy44/airline-safetyKey Features:

|  |  |
| --- | --- |
| airline | avail\_seat\_km\_per\_week |
| incidents\_85\_99 | fatal\_accidents\_85\_99 |
| fatalities\_85\_99 | incidents\_00\_14 |
| fatal\_accidents\_00\_14 | fatalities\_00\_14 |
| fatalities\_per\_fatal\_accidents\_85\_99 | fatalities\_per\_fatal\_accidents\_00\_14 |

Potential Use-Cases:

Safety Analysis: Analyzing the safety performance of different airlines over the specified periods.

Risk Assessment: Evaluating the risk factors associated with various airlines based on historical incidents and fatalities.

Trend Analysis: Identifying trends in airline safety and incidents over time.

Comparative Analysis: Comparing the safety metrics of different airlines to determine relative safety standings.

1. **Research Question**

How do historical incidents, fatal accidents, and fatalities correlate with the size and extent of airline operations, and can we predict future safety metrics based on this historical data?

1. **Machine Learning Model**

Regression Analysis: Given that the dataset contains numerical and time-series data, a regression model will be apt to analyze the relationships between different safety metrics and the operational size of airlines. Additionally, regression analysis can help predict future safety metrics based on historical data, providing valuable insights into potential future trends or risks in airline safety.

1. **Model Performance and Results**

Period: 1985-1999

|  |  |  |
| --- | --- | --- |
| 1: Logistic Regression  - Accuracy: 93.33%  - F1 Score: 93.54%  - Precision: 94.67%  - Recall: 93.33%  - AUC: 90.91%  - TP: 10, TN: 4, FP: 0, FN: 1 | 2: Decision Tree Classifier  - Accuracy: 100%  - F1 Score: 100%  - Precision: 100%  - Recall: 100%  - AUC: 100%  - TP: 11, TN: 4, FP: 0, FN: 0 | 3: Random Forest Classifier  - Accuracy: 100%  - F1 Score: 100%  - Precision: 100%  - Recall: 100%  - AUC: 100%  - TP: 11, TN: 4, FP: 0, FN: 0 |
| **confusion_matrix_LogisticRegression_85_99** | **confusion_matrix_DecisionTreeClassifier_85_99** | **confusion_matrix_RandomForestClassifier_85_99** |

4: Baseline Model (Predict all as 1)

Accuracy: 0.7333333333333333

5: Regression Model: Linear Regression

R-Squared (R2) on test data = 1.0

Root Mean Squared Error (RMSE) on test data = 6.704864714241624e-15

|  |
| --- |
| Precision-Recall Curve |
| precision_recall_curve_85_99 |

|  |
| --- |
| Residuals Plot |
| residuals_vs_predicted_85_99 |

|  |
| --- |
| ROC Curve |
| roc_curve_logreg_85_99 |

Period: 2000-2014

|  |  |  |
| --- | --- | --- |
| 1:Logistic Regression  - Accuracy: 100%  - F1 Score: 100%  - Precision: 100%  - Recall: 100%  - AUC: 100%  - TP: 5, TN: 10, FP: 0, FN: 0 | 2:Decision Tree Classifier  - Accuracy: 100%  - F1 Score: 100%  - Precision: 100%  - Recall: 100%  - AUC: 100%  - TP: 5, TN: 10, FP: 0, FN: 0 | 3:Random Forest Classifier  - Accuracy: 100%  - F1 Score: 100%  - Precision: 100%  - Recall: 100%  - AUC: 100%  - TP: 5, TN: 10, FP: 0, FN: 0 |
| **confusion_matrix_LogisticRegression_00_14** | **confusion_matrix_DecisionTreeClassifier_00_14** | **confusion_matrix_RandomForestClassifier_00_14** |

4: Baseline Model (Predict all as 0)

Accuracy: 0.6666666666666666

5: Regression Model: Linear Regression

R-Squared (R2) on test data = 1.0

Root Mean Squared Error (RMSE) on test data = 1.357864494910396e-15

|  |
| --- |
| Precision-Recall Curve |
| precision_recall_curve_00_14 |

|  |
| --- |
| Residuals Plot |
| residuals_vs_predicted_00_14 |

|  |
| --- |
| ROC Curve |
| roc_curve_logreg_00_14 |

|  |
| --- |
| fatalities\_over\_years |
| **fatalities_over_years** |

|  |
| --- |
| incident\_counts\_per\_airline |
| **incident_counts_per_airline** |

The models demonstrated high accuracy and precision in predicting the safety metrics of airlines for both periods. Particularly noteworthy is the 100% accuracy achieved by the Decision Tree Classifier, Random Forest Classifier, and Linear Regression models, indicating a perfect fit to the test data. However, it's crucial to approach these results with caution as such high accuracy might indicate overfitting, especially in the context of the Decision Tree and Random Forest models.

1. **Interpretation of Finding**

The models, especially the Decision Tree and Random Forest classifiers, demonstrated impeccable performance across all metrics. However, the Logistic Regression model also showcased commendable predictive capabilities, particularly in the 1985-1999 period, with an accuracy of 93.33%. The Linear Regression model achieved an R2 value of 100% for both periods, indicating that the model explains all the variability of the response data around its mean.

1. **Implications**

The findings suggest that historical safety metrics can be predictive of future metrics, providing a valuable tool for regulatory bodies and airline operators in assessing and enhancing airline safety. However, the potential for overfitting in the models, especially those with 100% accuracy, necessitates further validation and refinement to ensure the models generalize well to new data.

1. **Limitations and Challenges**

The primary limitation is the potential overfitting of models, given the 100% accuracy in several instances. Additionally, the models were tested on historical data, and their predictive power for future metrics needs to be validated with more recent data.

1. **Answer for Research Questions**

Correlation: The models show that there is a strong correlation between historical incidents, fatal accidents, fatalities, and future safety metrics, as evidenced by the high accuracy and F1 score in predictions.

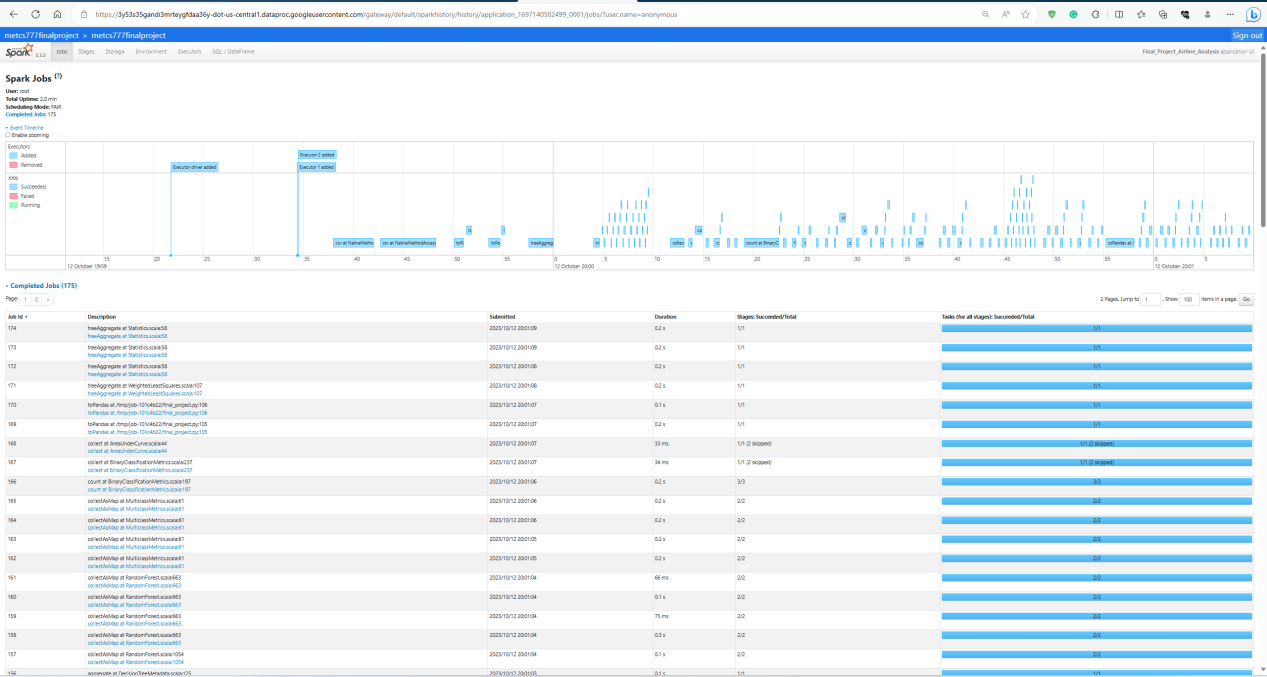
Predicting Future Safety Metrics: Yes, we can predict future safety metrics based on this historical data. The models, especially the Decision Tree and Random Forest classifiers, have demonstrated high predictive power.

In conclusion, while the models show promising results, it's essential to approach the deployment of such models with caution, ensuring thorough validation and considering the ethical implications of their use.

1. **Conclusion**

The study successfully developed models to analyze and predict airline safety metrics, providing valuable insights into the correlations and trends within the data. While the models demonstrated high accuracy, further validation and refinement are necessary to ensure their reliability and applicability to future safety metrics prediction.

1. **Spark History**

****