KAGGLE-AIRPLANE CRASHES:

Exploring Patterns and Risk Factors in Aviation Accidents Data

Project repository: https://github.com/Janspitoy/AirplaneCrashes

Team Members: Ivan Kliuchyshche, Anna Borosh, Hans Eduard Säre

1. Business understanding

1.1 Identifying Your Business Goals

Background:

Aviation accidents have a profound impact on public safety, airlines' reputations, and operational costs. By understanding the risk factors leading to airplane crashes, we can help enhance aviation safety. This project analyzes data on airplane crashes and flight operations, focusing on scheduled flights from 1990 to 2023, to provide insights into crash risks and their mitigation.

Business Goals:

1. Create a predictive model to classify flights as high-risk or low-risk.
2. Identify critical factors contributing to crash risks, such as weather conditions, mechanical issues, or operational errors.
3. Provide actionable insights for stakeholders to improve safety measures and minimize accidents.

Business Success Criteria:

- **Quantitative:** Achieve at least 80% accuracy in flight risk classification.
- **Qualitative:** Deliver a clear set of actionable insights, supported by data-driven analysis, to aviation stakeholders.

1.2 Assessing Your Situation

Inventory of Resources:

- **Datasets:**
  - Dataset 1: Crash history (1908–2023) with details on causes and fatalities.
  - Dataset 2: European flight metadata.
  - Dataset 3: US international traffic data.
- **Tools:** Python libraries (`pandas`, `scikit-learn`, `matplotlib`), Tableau, and GitHub for collaboration.
- **Human Resources:** A team of three members with skills in machine learning, data visualization, and aviation analytics.

Requirements, Assumptions, and Constraints:

- **Requirements:** Clean, consistent data for accurate modeling.
- **Assumptions:**
  - Historical crash patterns are relevant to current risks.
  - Features like weather and aircraft type are consistent across datasets.
- **Constraints:**

    - Data quality issues, such as missing values or irrelevant records, may limit the analysis.
    - Limited time to develop and test the model.

Risks and Contingencies:

- **Risk:** Missing or incomplete data for critical variables.
    - **Contingency:** Use imputation techniques or exclude such records to maintain data integrity.
- **Risk:** Model performance below expectations.
    - **Contingency:** Experiment with different algorithms and feature engineering to optimize performance.

Terminology:

- **High-risk flights:** Flights with an elevated probability of accidents based on risk factors.
- **Crash causes:** Factors contributing to accidents, such as mechanical failure or weather.

Costs and Benefits:

- **Costs:** Time and computational resources required for data cleaning, modeling, and analysis.
- **Benefits:**
    - Improved understanding of aviation safety risks.
    - Potential to save lives and reduce operational losses.

1.3 Defining Your Data-Mining Goals

Data-Mining Goals:

1. Build a predictive model for flight risk classification using machine learning.
2. Analyze the contribution of each variable (e.g., weather, operator, aircraft type) to crash risks.

Data-Mining Success Criteria:

- Achieve >80% accuracy for the risk classification model.

2. Data Understanding

2.1. Gathering Data

Outline Data Requirements:

- Key variables such as flight details, crash causes, fatalities, weather conditions, and traffic patterns.

Verify Data Availability:

- Dataset 1 provides crash history and fatalities from 1908–2023.
- Dataset 2 includes European flight metadata.
- Dataset 3 contains US international traffic data.

Define Selection Criteria:

- Use data entries with complete records for critical variables (e.g., date, location, crash cause).
- Focus on flights from the last 30-35 years for relevance to modern aviation.

2.2. Describing Data

Crash History Dataset:

- Fields: Date, Location, Operator, Aircraft Type, Fatalities, Crash Cause.
- Summary: Contains over 5,000 records of crashes worldwide.

European Flights Dataset:

- Fields: Flight ID, Origin, Destination, Flight Duration, etc.
- Summary: Details millions of flights within Europe.

US International Traffic Dataset:

- Fields: Airport, Region, Passengers, Freight, etc.
- Summary: Summarizes international air traffic trends to/from the US.

2.3. Exploring Data

Examples of Trends:

- Crashes over time: Visualize trends by decade.
- Crash factors: Examine correlations between weather and crash rates.

Verifying Data Quality

- **Missing Values:** Identify and impute missing entries for critical fields.
- **Outliers:** Use visualization techniques to identify and handle outliers.
- **Consistency:** Standardize units, dates, and terminology across datasets.

3. Planning The Project

3.1. Detailed Task List:

1. **Data Gathering and Preprocessing (6 hours per member):**
    - Collect datasets, clean missing values, and ensure consistency.
2. **Exploratory Data Analysis (6 hours per member):**
    - Perform statistical analysis and create visualizations to understand patterns.
3. **Feature Engineering and Model Development (8 hours per member):**
    - Engineer features and build a machine learning model to classify flights.
4. **Risk Analysis and Insights (4 hours per member):**
    - Analyze risk factors and create dashboards to present insights.
5. **Final Report and Presentation (6 hours per member):**

    - Summarize findings, prepare visualizations, and finalize the PDF report.

Methods and Tools:

- **Methods:** CRISP-DM framework, machine learning (e.g., logistic regression, decision trees), exploratory analysis.
- **Tools:** Python (pandas, scikit-learn, seaborn, matplotlib), Tableau, GitHub.

Comments:

- Allocate extra time for iterative model improvement.
- Schedule regular team meetings to ensure alignment across tasks.