**Project Overview**In this project, you will apply your data science and machine learning skills to develop a predictive model for the Estonia Disaster Dataset. The project involves data cleaning, exploratory data analysis (EDA), feature engineering, model training, hyperparameter tuning, and evaluation to understand the factors influencing survival during the MS Estonia ferry disaster.

**Project Introduction** Maritime disasters not only mark critical historical events but also offer insights into safety protocols and emergency response strategies. The MS Estonia ferry disaster, one of the worst maritime tragedies in recent history, resulted in a significant loss of life. Analyzing this dataset enables us to explore patterns related to survival and risk factors among the passengers. This study is essential for historians, safety regulators, and policymakers who aim to learn from past incidents and improve future maritime safety standards.

In 1994, the MS Estonia sank in the Baltic Sea under harsh weather conditions and operational challenges. By examining passenger demographics, ticketing information, and other related features, we can shed light on what factors contributed to survival or loss, thereby informing future safety measures and emergency protocols

**Project Objective** Clean and analyze the dataset to uncover insights that might have influenced survival outcomes for passengers aboard the MS Estonia.

Specifically, the objectives include:

* **Analyzing the data** to uncover significant trends and relationships between passenger characteristics and survival.
* **Identifying key factors** that influenced survival during the disaster.
* **Providing actionable insights** that could contribute to improved safety measures and risk assessment in maritime travel.
* **Automating the prediction process** through a robust machine learning pipeline that includes model training, evaluation, and interpretation.

**Project Phases**

**Phase 1: Data Collection and Preparation**

**Task 1.1:** Load the [Estonia Disaster Dataset](https://github.com/ek-chris/Practice_datasets/blob/main/estonia-passenger-list.csv) from the provided source.

**Task 1.2:** Import the dataset into a Pandas DataFrame.

**Task 1.3:** Inspect the dataset for missing values and outliers and address them appropriately.

**Task 1.4:** Perform initial data cleaning to prepare the dataset for further analysis.

**Phase 2: Exploratory Data Analysis (EDA)**

**Task 2.1:** Conduct an exploratory data analysis to understand the distribution of features and the target variable (survival).

**Task 2.2:** Visualize relationships between key features and survival using histograms, scatter plots, and box plots.

**Task 2.3:** Identify and manage potential outliers and anomalies within the dataset.

**Phase 3: Data Preprocessing**

**Task 3.1:** Engineer new features (e.g., age groups, cabin segments) that may enhance model performance.

**Task 3.2:** Encode categorical variables using one-hot or label encoding.

**Task 3.3:** Normalize or standardize numerical features where appropriate.

**Phase 4: Model Training and Evaluation**

**Task 4.1:** Split the dataset into training and testing subsets.

**Task 4.2:** Select and justify appropriate machine learning algorithms (e.g., Logistic Regression, Decision Trees, Random Forests, Gradient Boosting).

**Task 4.3:** Train multiple models and evaluate their performance using metrics such as precision, recall, accuracy, F1-score, and AUC-ROC.

**Task 4.4:** Conduct hyperparameter tuning to optimize the best-performing models.

**Task 4.5:** Evaluate the selected model on the testing set and document its predictive performance.

**Phase 5: Model Interpretation and Reporting**

**Task 5.1:** Interpret the results of the best-performing model, highlighting the importance of critical features affecting survival.

**Task 5.2:** Create comprehensive visualizations to support your analysis and conclusions.

**Task 5.3:** Write a detailed report summarizing your methodology, findings, and recommendations.

**Deliverables**

* **Code:** Submit the complete Jupyter Notebook or codebase used for data preparation and EDA.
* **Report:** Provide a detailed report (PDF or Google Doc) summarizing the project, including methodologies, visualizations, key findings, and conclusions.

**Data Description** The Estonia Disaster Dataset contains detailed information on the passengers aboard the MS Estonia ferry. The data dictionary is as follows:

**passenger\_id:** Unique identifier for each passenger.

**age:** Age of the passenger at the time of the disaster.

**gender:** Gender of the passenger (Male/Female).

**class:** Ticket class or cabin category indicating the accommodation level (e.g., First, Second, Third).

**ticket\_number:** Identifier for the passenger’s ticket.

**fare:** The fare paid by the passenger, if applicable.

**embarked:** Port of embarkation or point of departure (e.g., Tallinn, Helsinki).

**cabin:** Cabin number or section (if available).

**nationality:** Nationality or residency information of the passenger.

**survival:** Binary indicator showing whether the passenger survived (1) or not (0).

**Dataset:** [The Estonia Disaster Dataset](https://github.com/ek-chris/Practice_datasets/blob/main/estonia-passenger-list.csv) along with any additional documentation or data dictionaries.