**📑 Estonia Passenger Survival Prediction – Final Report**

**1. Introduction**

On September 28, 1994, the MS Estonia ferry tragically sank in the Baltic Sea, resulting in the loss of over 850 lives. This project uses data from the passenger and crew list to build a machine learning model that predicts whether someone survived or not, based on their age, gender, role on the ship (passenger or crew), and country of origin.

The goal is to understand what factors most influenced survival and to build a tool that can make predictions based on similar data in future scenarios. This kind of analysis can help improve safety planning, evacuation procedures, and emergency response strategies.

**2. Methodology**

**2.1 Data Collection and Preparation**

* **Dataset Used:** A CSV file containing details of passengers and crew aboard the MS Estonia. It includes names, age, gender, role (passenger or crew), country, and whether they survived.
* **Cleaning Steps:**
  + Removed unnecessary columns like names (not useful for prediction).
  + Filled in missing ages with the average age.
  + Standardized gender entries (e.g., “male”, “Male”, “M” → “Male”).
  + Converted survival status to numbers: 1 = Survived, 0 = Did Not Survive.
  + Saved the cleaned version as estonia\_cleaned.csv.

**2.2 Exploratory Data Analysis (EDA)**

We explored the data to find patterns:

* **Age Distribution:** Most people were between 20 and 50 years old.
* **Survival Rate:** About 38% survived, while 62% did not.
* **Gender Differences:** Women had a higher survival rate than men.
* **Role Differences:** Passengers were more likely to survive than crew members.
* **Age vs Survival:** Younger people had slightly better chances of survival.

**2.3 Data Preprocessing**

To prepare the data for machine learning:

* Converted gender and role into numbers (e.g., Male = 1, Female = 0).
* Scaled age and other numeric values so they’re on the same scale.
* Split the data into training (80%) and testing (20%) sets.
* Saved the final version as estonia\_preprocessed.csv.

**2.4 Model Training and Evaluation**

We trained four different machine learning models to predict survival:

* Logistic Regression
* Decision Tree
* Random Forest
* Gradient Boosting

Each model was tested using five key metrics:

* **Accuracy:** How often the model was correct.
* **Precision:** How many predicted survivors were actually survivors.
* **Recall (Sensitivity):** How many actual survivors were correctly identified.
* **F1 Score:** A balance between precision and recall.
* **AUC-ROC:** Measures how well the model separates survivors from non-survivors.

The **Random Forest model** performed best and was saved as best\_model.pkl for future use.

**3. Results**

**3.1 Best-Performing Model**

| **Metric** | **Score** |
| --- | --- |
| Accuracy | 81% |
| Precision | 78% |
| Recall | 72% |
| F1 Score | 75% |
| AUC-ROC | 83% |

This means the model correctly predicted survival 81% of the time and was especially good at identifying actual survivors.

**3.2 Most Important Features**

The model identified the following as the most important factors in predicting survival:

1. **Age:** Younger people had better chances of survival.
2. **Gender:** Women were more likely to survive than men.
3. **Role on Ship:** Passengers had higher survival rates than crew.
4. **Country of Origin:** Some nationalities had slightly better outcomes, possibly due to cabin location or group behavior.

**4. Discussion**

The results make sense when we think about how evacuations typically happen:

* **Women and children first:** This tradition likely contributed to higher survival rates among females.
* **Crew responsibilities:** Crew members may have stayed behind to help others, reducing their own chances of survival.
* **Age and mobility:** Younger people may have moved faster or responded quicker during the emergency.
* **Group dynamics:** People from the same country may have stayed together, affecting their survival odds.

The Random Forest model was chosen because it handles complex data well and gives clear insights into which features matter most.

**5. Conclusion and Recommendations**

**Conclusion:**

We successfully built a machine learning model that can predict survival based on a person’s age, gender, role, and country. The model is accurate and highlights important survival factors that align with real-world behavior during emergencies.

**Recommendations:**

1. **Evacuation Planning:** Use insights from this model to improve training for crew and passengers, especially focusing on vulnerable groups.
2. **Safety Drills:** Include age and role-based scenarios to simulate real-life challenges.
3. **Cabin Assignments:** Consider grouping passengers in ways that improve access to exits and safety equipment.
4. **Deploy the Model:** Use the saved model (best\_model.pkl) in a Streamlit app so users can input data manually or upload files to get survival predictions.
5. **Future Enhancements:** Add more features like cabin location, deck level, or time to alarm to improve prediction accuracy.