# **Titanic Dataset Analysis using Machine Learning**

# **Introduction**

The Titanic dataset is a classic example of data science, providing information on passenger demographics and survival rates from the terrible journey. This dataset offers a rich tapestry of information for research since it includes passenger details like age, gender, class, and ticket price. To extract significant patterns and trends, nevertheless, requires skilful analytical techniques to navigate through its complexities.

Using the Titanic dataset, our study builds a comprehensive predictive modelling system to take on this challenge. Predicting passenger survival status using different features from the dataset is the main goal. Additionally, by using exploratory data analysis (EDA) to identify underlying trends and connections between the variables, the system hopes to increase its usefulness.

In order to do this, we use machine learning (ML) approaches, which use algorithms to find complex patterns in the data and provide predictions. This project requires careful coordination of feature engineering, evaluation, model selection, and data preprocessing.

We prioritize the prediction models' accuracy, interpretability, and generalizability by approaching this complex problem with a methodical technique. Our system provides a solid solution that helps decision-makers comprehend passenger survival dynamics and factors impacting outcomes by converting raw datasets into meaningful insights.

The project report outlines how this predictive modelling system was developed methodically, emphasizing how machine learning, feature engineering, and exploratory data analysis were carefully combined. We clarify the meticulous steps involved in selecting, preparing, training, evaluating, and exploring data in order to create a survival prediction model that works.

The article goes on to explain in detail the technical foundations and techniques that allow the system to handle the Titanic dataset with accuracy and context awareness. This document provides a proof of the project's accomplishments as well as a guide for next developments in predictive modelling and analysis. The system's gradual evolution proceeds as follows:

# **1. Exploratory Data Analysis and Pre-Processing:**

The Titanic dataset was put into the analysis environment during this phase. In order to understand the structure, substance, and intrinsic properties of the dataset, the project carried out an exploratory data analysis, or EDA. This involved producing summary statistics and displaying the distributions of important data, especially the status of passenger survival and characteristics like age, class, and fare.

Depending on how missing values would affect the integrity of the dataset, they were recognized and handled accordingly through imputation or exclusion. Data entry errors were fixed, duplicates were eliminated, and data was standardized as part of the data cleansing process. For instance, the most prevalent cabin category was used to fill in the missing values for cabin, while median values were used to impute missing values for age. With "0" for every possible value, Figure 1 shows that the dataset is devoid of missing values.

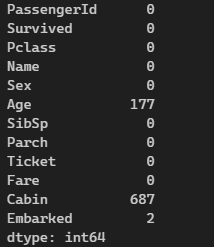


Figure 1: Showing Missing Values

Histograms, box plots, and scatter plots were among the visualization approaches used to analyse the correlations between the variables and find any discernible patterns. The distribution of passenger survival status in the dataset is shown in Figure 2, which also shows the percentage of survivors compared to non-survivors.

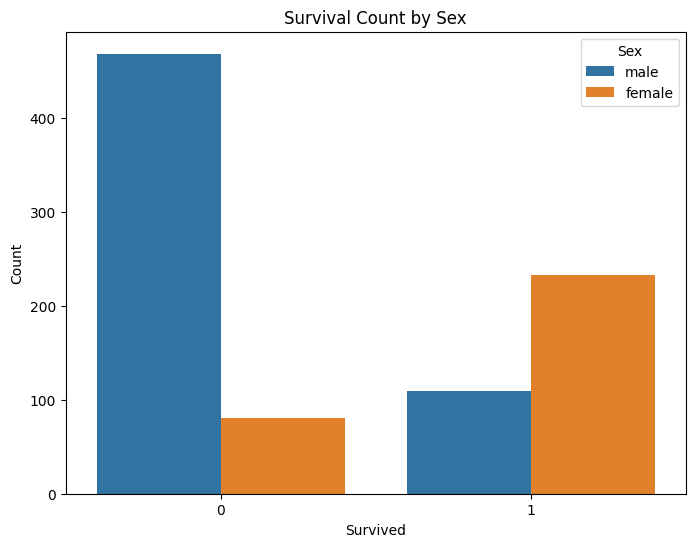


Figure 2: Survivors Gender Wise

The distribution of passenger ages is shown in Figure 3, which sheds light on the age makeup of the Titanic's passenger population. Comprehending the age distribution facilitates evaluating its influence on survival results.

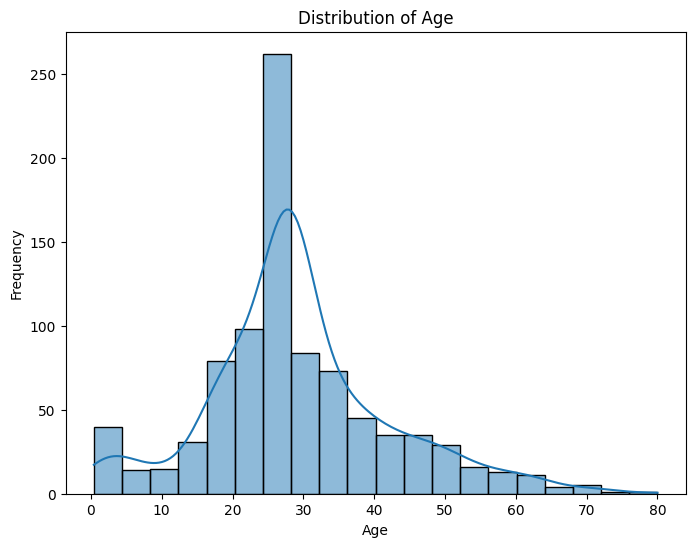


Figure 3: Passengers Ages Frequency

# **2. Feature Engineering:**

It was determined which characteristics, such as passenger class, age, gender, and family size, were pertinent for survival prediction. One-hot encoding and other approaches were used to encode numerical formatted categorical data, such as gender and embarkation port.

To capture the contextual subtleties required for survival prediction, new characteristics were developed. For instance, to reflect the total number of family members onboard, the 'SibSp' and 'Parch' features were combined to produce the 'FamilySize' feature.

In order to handle missing values and standardize numerical features, the dataset underwent preprocessing. The median age was used to impute missing values for "Age," while the mean fare was used to fill in missing values for "Fare."

Furthermore, features were designed to extract valuable information. To illustrate the impact of different age groups on survival, the variable "Age" was, for example, binned into categories like "child," "adult," and "elderly."

# **3. Model Selection and Training:**

To enable robust model training while keeping an impartial subset for evaluation, we split the Titanic dataset into training and testing sets using an 80-20% split.   
For the survival prediction task, a number of machine learning models were taken into consideration: decision trees and random forests were chosen for their hierarchical decision-making, support vector machines (SVM) and logistic regression for their interpretability, and ensemble methods like these were chosen for their resilience to outliers and high-dimensional data.   
Using the training data, each model was trained to identify the survival labels that corresponded to input parameters including passenger class, age, gender, and family size.

# **4. Hyperparameter Tuning:**

We adjusted the model hyperparameters to maximize performance using methods like grid search and cross-validation. In order to ensure that our models performed well in terms of generalizing to new data while reducing the possibility of overfitting, this step was critical in determining the ideal ratio between bias and variance.   
Each algorithm's hyperparameters were adjusted as follows:   
 **1.** Logistic Regression: Two values were used to adjust the 'C' and 'penalty' parameters.   
Random Forest: 'n\_estimators', 'criterion' (using 'gini' and 'entropy'),'max\_depth', and'min\_samples\_split' were adjusted.   
**2.** SVM: We adjusted the 'C' parameter, 'degree', 'gamma' (with'scale', 'auto'), 'kernel' (with 'linear', 'rbf', 'poly', and'sigmoid').

# **5. Model Evaluation**

A variety of measures were employed to assess the models. To comprehend true positives, false positives, true negatives, and false negatives, a confusion matrix was created for every model.   
 We prioritized precision, recall, and the F1-score over accuracy in our survival prediction job because of the potential for class imbalance.

The following are the outcomes for every algorithm:

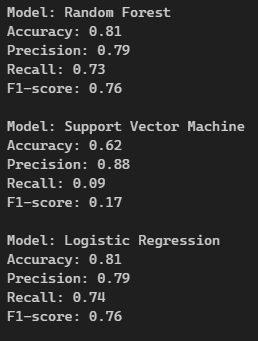


Figure 4: Results

# **6. Conclusion and Recommendations**

A summary of our findings is provided in the report's conclusion, highlighting the effectiveness and usefulness of the predictive modelling approach in correctly projecting survival outcomes based on passenger characteristics. We emphasized how better resource allocation and decision-making processes could result from these projections in emergency situations.   
We also talked about other possible enhancements, like adding more features to the training dataset or using more complex machine learning algorithms, like neural networks—especially those with architectures like Long Short-Term Memory (LSTM)—to detect temporal dependencies in the data.

The recommendations included considering future feature enhancements based on emerging trends in passenger data and evolving factors influencing survival probabilities, such as changes in passenger demographics or advancements in maritime safety protocols, as well as continuous model monitoring to ensure its predictive performance remains robust over time.

**References**

Ekinci, E., Omurca, S.İ. and Acun, N., 2018, December. A comparative study on machine learning techniques using Titanic dataset. In *7th international conference on advanced technologies* (pp. 411-416).

Singh, A., Saraswat, S. and Faujdar, N., 2017, May. Analyzing Titanic disaster using machine learning algorithms. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 406-411). IEEE.

Kakde, Y. and Agrawal, S., 2018. Predicting survival on Titanic by applying exploratory data analytics and machine learning techniques. *International Journal of Computer Applications*, *179*(44), pp.32-38.

Farag, N. and Hassan, G., 2018, May. Predicting the survivors of the Titanic Kaggle, machine learning From disaster. In *Proceedings of the 7th international conference on software and information engineering* (pp. 32-37).

Ibrahim, A.A. and Abdulaziz, R.O., 2020. Analysis of Titanic Disaster using Machine Learning Algorithms. *Engineering Letters*, *28*(4).

Tabbakh, A., Rout, J.K. and Rout, M., 2021. Analysis and prediction of the survival of titanic passengers using machine learning. In *Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2020* (pp. 297-304). Springer Singapore.

Cao, E.Y., Xie, W., Dong, C. and Qiu, J., 2020. Titanic Machine Learning Study from Disaster.

Dasgupta, A., Mishra, V.P., Jha, S., Singh, B. and Shukla, V.K., 2021, March. Predicting the Likelihood of Survival of Titanic’s Passengers by Machine Learning. In *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)* (pp. 52-57). IEEE.

Durmuş, B. and Güneri, Ö.İ., 2020. Analysis and detection of Titanic survivors using generalized linear models and decision tree algorithm. *International Journal of Applied Mathematics Electronics and Computers*, *8*(4), pp.109-114.