AUTUMN INTERNSHIP PROJECT REPORT

Titanic Dataset Analysis and Survival Prediction using Machine Learning

Somoshree Sadhukhan

B.Sc. Statistics 2024-28

Sister Nivedita Government General Degree College for Girls

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project, **Titanic Dataset Analysis and Survival Prediction using Machine Learning** analyses the Titanic passenger dataset to predict survival outcomes using machine learning techniques. The dataset contains demographic and travel-related information such as age, gender, class, and fare. Various preprocessing steps, including handling missing values and encoding categorical variables, were performed. Multiple models, such as Logistic Regression, KNN, Support Vector Machine, were applied and evaluated. Performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC were used to compare the models which one performed the best. The study highlights key factors affecting survival, such as gender and passenger class. The project demonstrates practical applications of Python, pandas, scikit-learn, and data visualization libraries. The workflow emphasizes data cleaning, feature engineering, model building, and evaluation. It provides insights into predictive analytics for real-world datasets. This project demonstrated how data science techniques can solve classification problems.

2. Introduction

The Titanic dataset project aims to understand survival patterns of passengers aboard the Titanic and predict outcomes using machine learning. This project introduces fundamental concepts of data analysis, predictive modelling, and feature importance. The technologies used include Python, Google Colab/Jupyter Notebook, pandas, NumPy, Matplotlib, Seaborn, and scikit-learn. The workflow involved:

1. Data cleaning and preprocessing
2. Exploratory data analysis (EDA)
3. Feature engineering and encoding
4. Model training and evaluation
5. Interpretation of results

The purpose of the project is to illustrate how factors like age, gender, and passenger class influence survival probability and to compare the effectiveness of different machine learning models.

The topics which were covered during the first two weeks of internship classes, under the guidance of eminent professors, are Python basics including: -

- Data, Variables, Lists, Loops
- Data Structures
- Class, Functions, OOPS concepts
- NumPy, Pandas
- Machine Learning Overview
- Regression Lab
- Classification lab
- LLM Fundamentals
- Communication skills

3. Project Objective

- Predicting the survival of Titanic passengers based on demographic and travel-related features.
- Identifying key factors (e.g., age, gender, class) that influenced survival.
- Applying and compare multiple machine learning models for classification accuracy.
- Demonstrating data preprocessing techniques, including missing value imputation and feature encoding.
- Conducting exploratory data analysis to find patterns and insights.

4. Methodology

- ## Data Handling: -
    - **:** pandas, NumPy
    - **Visualization:** Matplotlib, Seaborn
    - **Machine Learning & Modelling:** scikit-learn
    - **Evaluation Metrics:** accuracy score, precision score, recall score, f1_score, roc-auc-score, confusion matrix

- ## Data Preprocessing Steps

1. **Data Cleaning:**
    - Missing values in the **Age** column were replaced with the median age.
    - Missing values in **Embarked** were replaced with the mode of the column.
    - Unnecessary columns such as PassengerId, Name, Ticket, and Cabin were removed for modelling.
2. **Encoding Categorical Variables:**
    - Sex and Embarked were encoded into numeric values for modelling.
    - Example: Male = 0, Female = 1; Embarked: S = 0, C = 1, Q = 2.
3. **Feature Selection & Engineering:**
    - Created new features if necessary (e.g., family size = SibSp + Parch).
    - Selected relevant features affecting survival: Pclass, Sex, Age, Sib/Sp, Parch, Fare, Embarked.

- ## Exploratory Data Analysis (EDA)

1. **Descriptive statistics** to understand distributions, mean, median, outliers.
2. **Visualizations**:
    - Histograms for Age and Fare
    - Bar charts for Survival by Sex, Pclass, Embarkation
    - Correlation heatmap for numeric features

• Machine Learning Modelling

1. **Data Splitting:**
    - Dataset split into training and testing sets using **train_test_split** (80% training, 20% testing).
    - Features (X) and target (y = Survived) separated before splitting.
2. **Model Selection:**
    - Evaluated multiple classification models:
        - Logistic Regression
        - K-Nearest Neighbors (KNN)
    - Selection based on predictive performance and interpretability.
3. **Model Training & Validation:**
    - Models trained on the training set.
    - Hyperparameters tuned using default settings initially.
    - Performance evaluated on the testing set using: Accuracy, Precision, Recall, F1-Score, AUC.
4. **Model Comparison:**
    - Confusion matrices and ROC curves generated for visualization of performance.

5. Data Analysis and Results

Descriptive Analysis

Summary Statistics:

Feature	Mean / Mode / Median	Notes
Age	29.7 (mean)	Missing values filled with median
Fare	32.2 (mean)	Outliers handled
Pclass	2 (mode)	Categorical (1st, 2nd, 3rd)
Sex	Male/Female	Encoded as 0 (male), 1 (female)
Survived	0.38 (mean)	38% survived, 62% did not

Correlation Heatmap of Numeric Features

	Survived	**Pclass**	**Age**	**Sib/Sp**	**Parch**	**Fare**
Survived	1.00	-0.34	-0.08	0.08	0.08	0.26
Pclass	-0.34	1.00	0.00	-0.06	-0.04	-0.55
Age	-0.08	0.00	1.00	0.14	0.07	0.09
Sib/Sp	0.08	-0.06	0.14	1.00	0.41	0.16
Parch	0.08	-0.04	0.07	0.41	1.00	0.22
Fare	0.26	-0.55	0.09	0.16	0.22	1.00

Machine Learning Model Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Logistic Regression	0.81	0.77	0.73	0.75	0.87
K-Nearest Neighbours	0.76	0.72	0.70	0.71	0.78

6. Conclusion

The Titanic dataset project provided a comprehensive experience in applying data science techniques to a real-world problem. I successfully cleaned and pre-processed the data, performed exploratory data analysis, and built multiple machine learning models to predict passenger survival. Logistic Regression performed well, demonstrating the importance of simpler models. Key factors such as gender, passenger class, and fare were identified as significant predictors of survival. The project strengthened my skills in Python programming, data visualization, and model evaluation. It also gave me a clear understanding of the end-to-end workflow of a data science project. I am grateful for the opportunity to undertake this internship at such a reputed institute, which provided excellent guidance and learning resources. This experience has been highly motivating and has increased my confidence in handling real-world datasets. I look forward to applying these skills to more advanced analytical projects in the future.