TITANIC SURVIVAL PREDICTION

Welcome to my Machine Learning project where I tackle one of the most iconic datasets in data science history: the **Titanic Survival Prediction** dataset. In this notebook-based project, I analyze passenger data to build a predictive model that estimates whether a given individual would have survived the Titanic disaster.

🎯 OBJECTIVE

The aim of this project is to:

- Understand the real-world process of data science, from cleaning raw data to deploying predictive models.

- Build a classification model to **predict survival** based on passenger information.

- Gain hands-on experience with **exploratory data analysis (EDA)** and **model evaluation.**

- Strengthen my knowledge of machine learning using an approachable and historical dataset.

📊 Dataset Description

- **Source:** [Kaggle - Titanic: Machine Learning from Disaster](https://www.kaggle.com/c/titanic)

- **File Used**: `tested.csv`

Structure:

Column Name	Description
PassengerId	Unique ID assigned to each passenger
Survived	Survival status (0 = No, 1 = Yes)
Pclass	Passenger's ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
Name	Full name of the passenger
Sex	Gender of the passenger (male/female)
Age	Age of the passenger in years
SibSp	Number of siblings or spouses aboard
Parch	Number of parents or children aboard
Ticket	Ticket number
Fare	Amount of fare paid by the passenger
Cabin	Cabin number (if available)
Embarked	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

🧹 Data Cleaning and Preprocessing

Steps taken to clean and prepare the data:

- Handled missing values in `Age`, `Embarked`, and `Cabin`.

- Converted categorical variables (`Sex`, `Embarked`) into numeric using encoding.

- Removed irrelevant columns (`Ticket`, `Name`, `Cabin`) to simplify the model.

- Scaled numerical features where necessary to improve model performance.

python

Example transformation

dataset['Sex'] = dataset['Sex'].replace({'male': 0, 'female': 1}).astype(int)

📈 Exploratory Data Analysis (EDA)

Conducted in-depth visual and statistical analysis:

- **Survival Rate by Gender:** Females had higher survival rates.
- **Impact of Class:** 1st class passengers were more likely to survive.
- **Age Distribution:** Children had a slightly higher survival chance.
- **Fare Distribution:** Higher fare correlated with higher survival probability.

📊 Visualizations included:

- Count plots for categorical features.

🤖 Model Selection and Training

- **Model Used:** Logistic Regression (Binary Classification)
- **Split Ratio:** 80% training, 20% testing
- **Evaluation Metrics:** Accuracy, Confusion Matrix, Classification Report

📊 Model Evaluation

- **Accuracy Score:** ~81% (example)
- **Confusion Matrix:** Showed good balance between precision and recall
- **Classification Report:**
  - Precision, Recall, F1-score for each class
  - Indicates strong performance in predicting survivors

✅ Results and Key Findings

- Gender and class are the most influential features.
- Simpler models like logistic regression performed surprisingly well.
- Feature engineering had a significant impact on improving model accuracy.

⚠️ Challenges Faced and Solutions

Challenge	Solution Implemented
Missing age values	Filled with median based on Pclass and Sex grouping
Imbalanced feature impact	Performed normalization and removed irrelevant features
Overfitting with too many features	Kept the model lean and used regularization

🛠️ Tools, Technologies & Libraries Used

- **Language:** Python
- **IDE:** Google Colab / Jupyter Notebook
- **Libraries:**
  - pandas, numpy – data manipulation
  - matplotlib, seaborn – visualization
  - scikit-learn – model building and evaluation

📝 Conclusion

This project provided an end-to-end experience in solving a supervised classification problem using the Titanic dataset. From cleaning and analyzing data to training a predictive model, I gained deeper insights into model interpretability and real-world data quirks.

🙏 Acknowledgements

- Kaggle for the dataset.
- RD INFRO TECHNOLOGY
- Scikit-learn documentation and the broader data science community for guidance and code snippets.

📬 Contact & Portfolio

👤 **Saran S**
✉️ **Email:** [saranselvaraj2401@gmail.com]
💼 **LinkedIn:** [linkedin.com/in/saranselvaraj2401]
📂 **GitHub:** [github.com/saran2007s]