■ Titanic: Machine Learning from Disaster – Project Report

This report documents the process of solving the Kaggle competition 'Titanic: Machine Learning from Disaster'. The task is to predict passenger survival based on socio-economic status, gender, age, and other features. The dataset was split into training and test sets, and a predictive model was developed using the training data.

■ Dataset Overview

The dataset includes details about passengers such as ticket class, sex, age, family size, fare, cabin, and port of embarkation. The target variable is 'Survived', where 0 = No and 1 = Yes.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
age	Age in years	
sibsp	# of siblings/spouses aboard	
parch	# of parents/children aboard	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southamp

■ Exploratory Data Analysis

- Females had a significantly higher survival rate than males. - First-class passengers had a higher survival rate compared to second and third class. - Younger passengers had better chances of survival. - Higher fare was associated with a higher survival probability. - Missing values were present in Age, Cabin, and Embarked columns.

■ Feature Engineering

- Extracted passenger title from 'Name'. - Created 'FamilySize' from SibSp + Parch + 1. - Created 'IsAlone' indicator. - Extracted deck from 'Cabin' and grouped missing as 'Unknown'. - Binned 'Age' and 'Fare' into categories. - One-hot encoded categorical variables.

■ Model Building

A Random Forest Classifier was chosen for its ability to handle mixed data types and non-linear relationships. A preprocessing pipeline was used to impute missing values, encode categorical features, and standardize numeric ones.

■ Results

- Cross-validation accuracy: ~0.82 (example result) - Test set accuracy (public leaderboard): ~0.78 (example result) - Feature importance: Sex, Pclass, Fare, Age, and Title were most important.

■ Submission

The final model was trained on the entire training dataset and predictions were made on the test dataset. The submission file contained PassengerId and the predicted Survived column.

■ Conclusion

The Titanic competition provided an opportunity to practice end-to-end machine learning: from data cleaning and feature engineering to model training, evaluation, and submission creation. Further improvements could be achieved through hyperparameter tuning, ensembling, and more advanced feature extraction.