Titanic Dataset - EDA Summary

Exploratory Data Analysis (EDA) Report

Dataset: Titanic (from Kaggle)

Objective:

The goal is to explore the Titanic dataset using visual and statistical tools, and draw insights related to passenger survival.

Tools Used:

- Python
- Pandas
- Matplotlib
- Seaborn
- Jupyter Notebook

1. Data Overview:

- 891 passenger records
- Key features: Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked, Survived
- Target Variable: Survived (0 = No, 1 = Yes)

2. Preprocessing:

Titanic Dataset - EDA Summary

- Handled missing values: 'Age' filled with median, 'Embarked' with mode.

- Converted categorical variables (e.g., Sex, Embarked) to numeric using Label Encoding.

- Removed irrelevant features (e.g., Name, Ticket, Cabin) for modeling.

3. Univariate Analysis:

- Age: Majority of passengers were 20-40 years old.

- Sex: About 65% male, 35% female.

- Pclass: Most passengers were in 3rd class.

- Survived: 38.4% survived, 61.6% did not.

4. Bivariate Analysis:

- Gender vs Survival: Females had a significantly higher survival rate.

- Pclass vs Survival: 1st class had highest survival rate.

- Age vs Survival: Children and young adults had better chances.

5. Multivariate Analysis:

- Pairplot to visualize combinations of features.

- Heatmap used to examine correlation. Sex and Pclass had strong correlation with survival.

6. Feature Importance:

- Most important: Sex, Pclass, Age

Titanic Dataset - EDA Summary

7. Model Summary:

Model	Accuracy (%)
Logistic Regression	81.36
K-Nearest Neighbors (K=3)	81.03
Decision Tree	92.59
Random Forest	92.59
Tuned Decision Tree	86.86
SVM	81.36
SGD	66.32
Perceptron	81.59
Naive Bayes	77.20
Stacking (Proposed)	**94.17**

8. Conclusions:

- Gender and class played a major role in survival.

- Ensemble models (stacking) performed best.

- Visualization helps identify patterns in survival data effectively.

9. Future Work:

- Use of XGBoost, AdaBoost, deep neural networks.

- Try advanced preprocessing (e.g., feature engineering).

- Validate with cross-validation and more robust tuning.

Prepared by:

Group IT-21