Exploratory Data Analysis on Titanic Dataset

1. Dataset Overview:

- Total Passengers: 891

- Features: Survived, Pclass, Name, Sex, Age, SibSp, Parch, Fare, Cabin, Embarked

- Missing Values: Age (177), Cabin (687), Embarked (2)

- Removed: PassengerId and Ticket

2. Survival Rate:

- 38.3% people survived.

- Accuracy from naive model (everyone perished): 62%. Accuracy alone is not a good metric.

3. Gender and Survival:

- Female survival rate: ~74%

- Male survival rate: ~19%

- Submission with only 'Sex' gives ~78.67% accuracy.

4. Class and Survival:

- 1st Class: 62.96% survived

- 2nd Class: 47.28% survived

- 3rd Class: 24.24% survived

5. Family (SibSp, Parch) Analysis:

- Majority had no siblings or parents onboard.

- No strong priority was given to passengers with family.

6. Fare Analysis:

- Clear distinction in Fare across Pclass.

- Positive correlation (0.257) between Fare and survival.

7. Correlation Insights:

- Weak correlation: Age, SibSp, Parch

- Moderate correlation: Fare, Pclass

8. Age Imputation:

- Categories extracted from 'Name' field (Mr, Mrs, Miss, Master, Dr)

- Used average age per category for imputation:

Mr: 32.37, Mrs: 35.89, Miss: 21.77, Master: 4.57, Dr: 42.0

Conclusion:

- Strongest predictors of survival: Sex, Pclass, Fare.

- Smart feature engineering (like 'Title') can help improve model performance.