Titanic Dataset – Data Cleaning & Exploratory Data Analysis

Internship: Prodigy Infotech (Task 01 – Data Science)

1. Data Cleaning

Missing values in the Age column were replaced using the median value to maintain data consistency. The Embarked column missing values were filled using the mode. The Cabin column was dropped due to a high number of missing entries.

2. Survival Distribution

The survival distribution shows that more passengers did not survive compared to those who survived the disaster.

3. Survival by Gender

Female passengers had a significantly higher survival rate compared to male passengers, indicating gender played an important role in survival.

4. Survival by Passenger Class

Passengers in first class had a higher survival rate, followed by second class, while third class passengers had the lowest survival rate.

5. Age Distribution

The age distribution shows that most passengers were young adults, with a higher concentration between 20 and 40 years.

6. Fare vs Survival

Passengers who paid higher fares generally had better survival chances, suggesting a relationship between socioeconomic status and survival.

7. Correlation Analysis

The correlation heatmap indicates that passenger class and fare show notable correlations with survival, while age has a weaker relationship.

Conclusion

Exploratory Data Analysis of the Titanic dataset reveals that gender, passenger class, and fare were key factors influencing survival. This analysis provides valuable insights for predictive

modeling and further machine learning applications.