Titanic Dataset Analysis Report

Data Cleaning:

Handling Missing Values:

- **Age:** Missing values were filled with the median age of passengers.
- **Embarked:** Missing values were filled with the most frequent port of embarkation ('S').
- **Cabin:** The 'Cabin' column was dropped due to a large number of missing values.

Data Exploration (EDA) and Visualization:

1. Survival Rate by Passenger Class:

- Visualized using a bar plot to show how survival rates varied across different passenger classes. It was observed that passengers in higher classes (1st class) had higher survival rates compared to lower classes.

2. Survival Rate by Sex:

- A bar plot was used to compare survival rates between male and female passengers. It showed that females had significantly higher survival rates compared to males.

3. Age Distribution:

- Histograms and KDE plots were used to visualize the age distribution of passengers. Most passengers were in the 20–30 age range.

4. Fare Distribution:

- histogram with KDE was used to visualize the fare distribution paid by passengers, showing a skewed distribution with some outliers.

5. Survival Rate by Embarkation Port:

- bar plot showed survival rates based on the port of embarkation ('S', 'C', 'Q'). Passengers embarking from port 'C' had higher survival rates.

6. Survival Rate by Family Size:

- bar plot depicted how survival rates varied with family size (combining siblings, spouses, and parents/children). Small families tended to have higher survival rates.

7. Correlation Heatmap:

- Analyzed correlations between numerical features (e.g., age, fare, family size) using a heatmap. It showed correlations that could impact survival predictions.

8. Survival Rate by Age and Gender:

- The violin plot displayed the age distribution by survival status and gender. It indicated that younger females had higher survival rates.

9. Passenger Count by Embarkation Port:

- The count plot illustrated the number of passengers embarking from each port, providing insights into the dataset's distribution.

10. Survival Rate by Age:

- KDE plot demonstrated survival rates by age, highlighting how survival varied across different age groups.

11. Fare Distribution by Survival:

- box plot illustrated the fare distribution for survivors and non-survivors, revealing potential differences in fare paid based on survival.

12. Survival Rate by Passenger Class and Gender:

- bar plot depicts survival rates categorized by both passenger class and gender, showing survival patterns across different demographics.

13. Survival Rate by Age Group:

- Bar plots categorized survival rates into age groups (e.g., children, teenagers, and adults), offering insights into age-based survival trends.

14. FacetGrid for Survival Rate by Age, Passenger Class, and Gender:

- I used FacetGrid to visualize survival rates across different combinations of age, passenger class, and gender, providing a comprehensive view of survival patterns.

15. Pair Plot of Numerical Features and Survival Outcome:

- Pair plots showed relationships between numerical features (e.g., age, fare) and survival outcome, with KDE plots highlighting distribution differences between survivors and non-survivors.

Machine Learning Model Building and Evaluation:

- **Data Preprocessing:**
  - Created age groups and dropped irrelevant columns.
  - Handled missing values for age and fare, encoded categorical variables ('Sex', 'AgeGroup').
  - Split data into training and testing sets and scale numerical features using StandardScaler.
- **Model Evaluation:**
  - Trained multiple classification models (logistic regression, decision tree, random forest, SVM).
  - Evaluated models using accuracy scores and classification report metrics, highlighting model performance in predicting survival.
- **Hyperparameter Tuning:**
  - Conducted GridSearchCV to optimize model performance by tuning hyperparameters.
  - Identified best parameters for each model to maximize accuracy and cross-validation scores.

Analysis and Recommendations:

- **Model Performance:**
  - **Random Forest** emerged as the top-performing model after tuning, achieving the highest accuracy (~82%) and balanced performance metrics.
  - **Logistic Regression** and **Decision Tree** also performed well, with accuracies around 80%, demonstrating robustness in predicting survival.
- **Recommendations:**
  - Based on the analysis, **Random Forest** is recommended for predicting survival on the Titanic dataset due to its high accuracy and balanced performance across metrics.

- Further model refinement and ensemble techniques could potentially enhance predictive capabilities, especially for scenarios with complex interactions between features.

Conclusion:

The detailed analysis and visualization of the Titanic dataset provide valuable insights into factors influencing survival rates among passengers. By leveraging exploratory data analysis and machine learning techniques, I've uncovered patterns and relationships within the data, ultimately leading to actionable recommendations for predictive modeling. This approach not only enhances understanding of historical events but also showcases the application of data science methodologies in real-world scenarios.