Exploratory Data Analysis of Titanic Dataset

1. Introduction

The Titanic dataset provides insights into survival rates based on passenger characteristics. I performed an exploratory data analysis to understand missing values, survival rates, outliers, correlations, performance metrics, and model optimization.

---

2. Handling Missing Values

I identified missing values in **Age, Embarked, and Cabin** columns. To handle them:

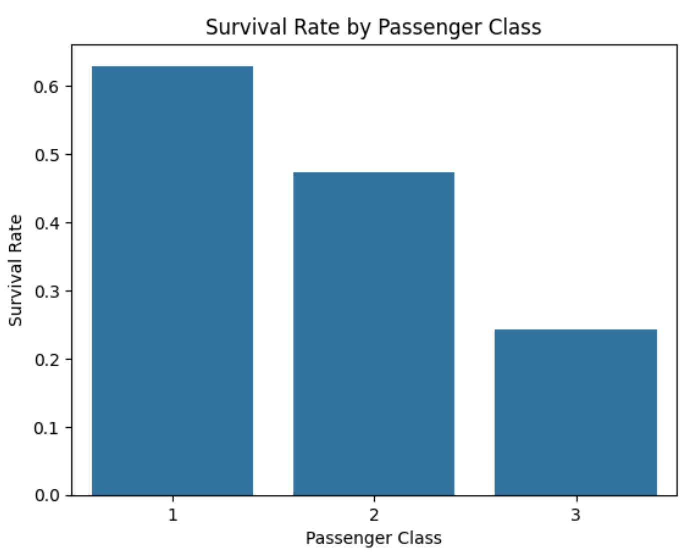
- **Age**: I replaced missing values with the median to prevent extreme values from affecting the data.
- **Embarked**: I filled missing values with the most common category to maintain consistency.
- **Cabin**: I dropped this column because **77%** of its values were missing, making it unreliable.

---

3. Relationship Between Ticket Class and Survival

I analyzed how ticket class affected survival rates. Here's what I found:

- **1st Class**: **62.9%** survived.
- **2nd Class**: **47.3%** survived.
- **3rd Class**: Only **24.2%** survived.



- Higher-class passengers had a much better chance of survival compared to lower-class passengers.

---

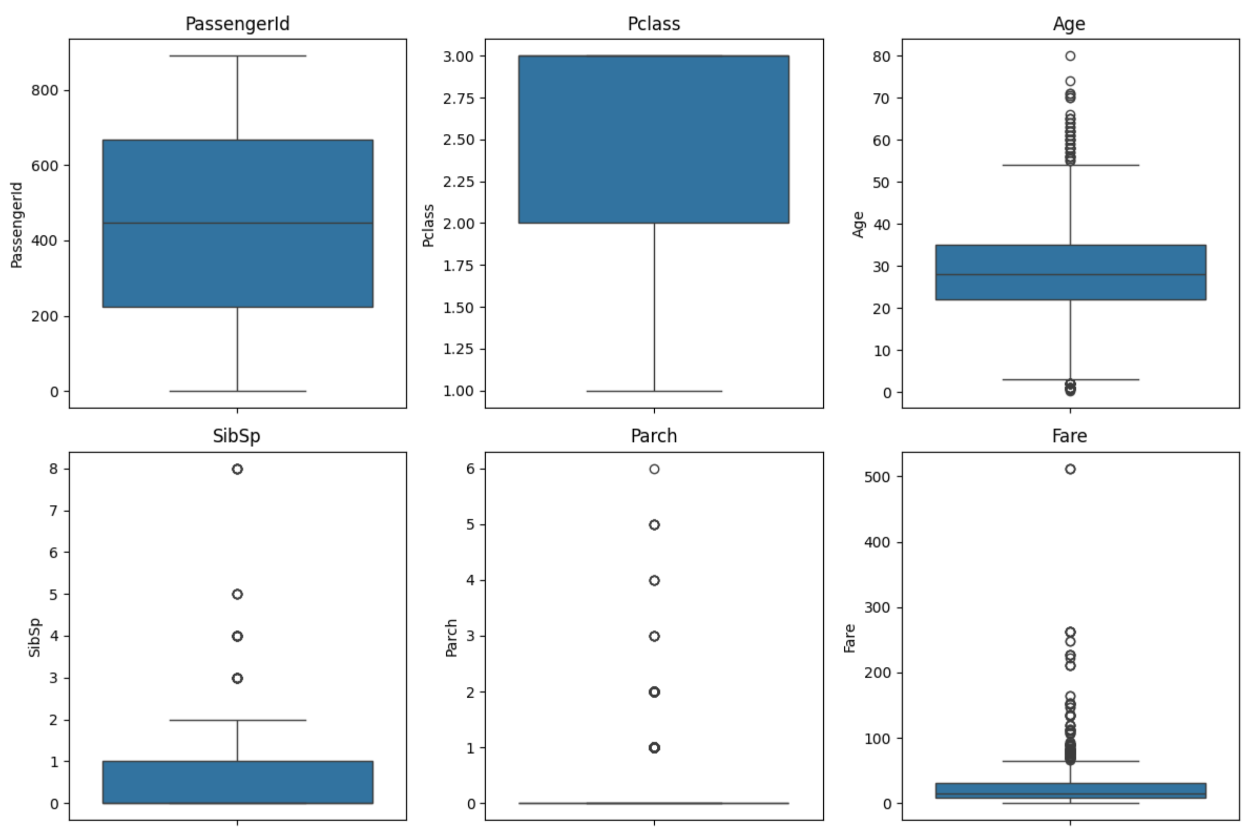
4. Identifying Outliers

I checked for outliers in numerical features using boxplots:

- **Age**: Some passengers were **older than 60**.
- **SibSp & Parch**: Some passengers had unusually large families.
- **Fare**: High ticket prices for first-class passengers created extreme values.

To handle these outliers:

- **Capped values** for Age (60), SibSp (3), and Parch (3) to keep data meaningful.
- **Applied log transformation** on Fare to reduce extreme variations.

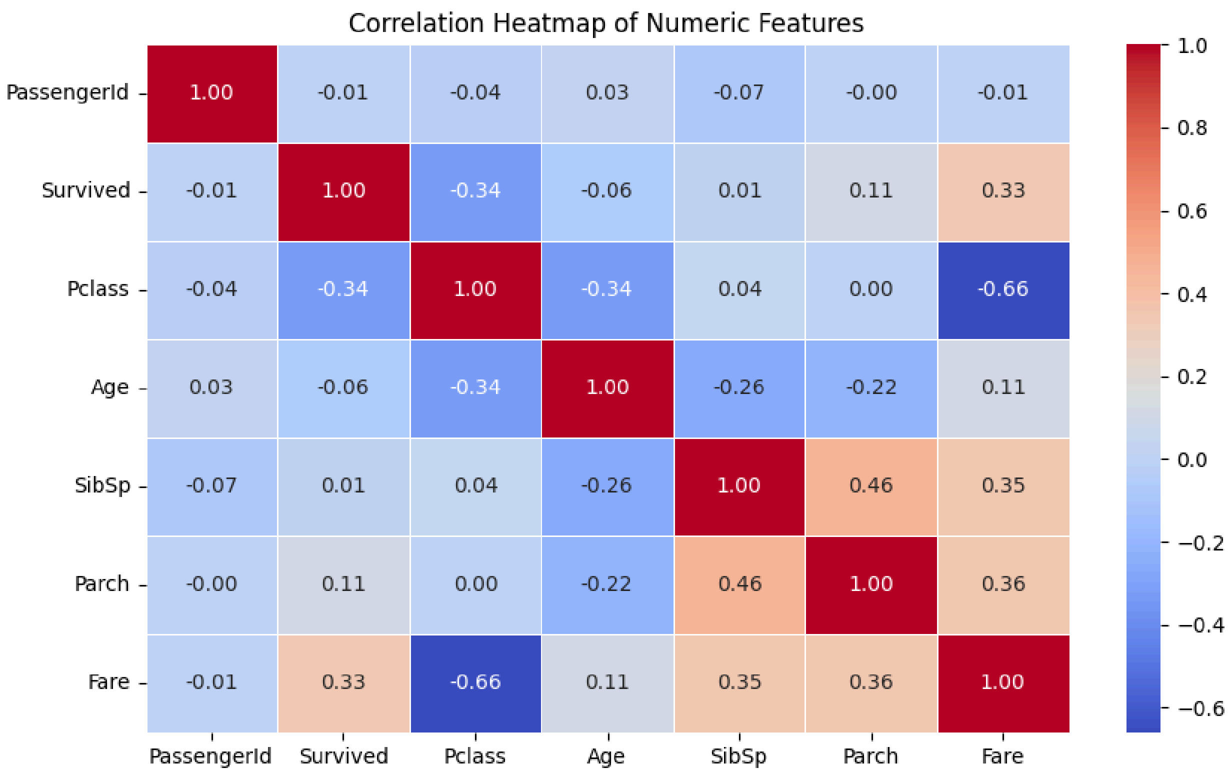


---

5. Correlation Analysis

I created a heatmap to analyze relationships between numerical features. Key findings:

- **Pclass & Fare (-0.66)**: Higher-class passengers paid more.
- **Pclass & Survived (-0.34)**: Higher-class passengers had a better survival rate.
- **Fare & Survived (0.33)**: Passengers who paid higher fares had a higher survival chance.



---

6. Measuring Model Performance

To evaluate my model, I used these metrics:

- **Accuracy**: Measures correct predictions.
- **ROC AUC Score**: Measures how well the model separates survivors from non-survivors.
- **Precision & Recall**: Evaluates how well true survivors are identified.
- **F1-Score**: Balances precision and recall.

- **Confusion Matrix**: Shows misclassified prediction

7. Model Optimization and Results

I trained a **Random Forest Classifier** and initially achieved:

- **Accuracy: 81.56%**
- **ROC AUC Score: 0.8920**

To improve performance, I performed **hyperparameter tuning** using GridSearchCV. After optimization, I achieved:

- **Final Accuracy: 83.80%** (Improved from 81.56%)
- **Final ROC AUC Score: 0.8981** (Better class separation)
- **Fewer misclassifications**: More passengers were correctly classified.

---

8. Execution Performance Enhancements

I improved the execution of my model by:

- **Using pipelines** to automate data preprocessing and model training.
- **Hyperparameter tuning** to find the best model settings.
- **Feature importance analysis** to identify key survival factors.
- **Saving the final model** using Joblib for future use.

---

9. Conclusion

Through this analysis, I identified key survival factors, optimized model performance, and improved execution efficiency. The final model is more accurate and provides valuable insights into survival patterns on the Titanic.

CT/DT Number: DT20234683428
Name: Vallabhareddy Datha Paneswara Ganesh