Titanic Survival Prediction: Model Development Report

1. Introduction

The objective of this analysis is to develop a predictive model to determine whether a passenger survived the sinking of the Titanic based on various features such as age, gender, ticket class, and family relations. The dataset consists of a training set with survival outcomes and a test set for evaluation.

2. Data Loading and Exploration

2.1 Data Import

The training and test datasets were loaded using "pandas.read_csv()". An initial inspection using "info()" and "describe()" provided an overview of missing values and data distributions.

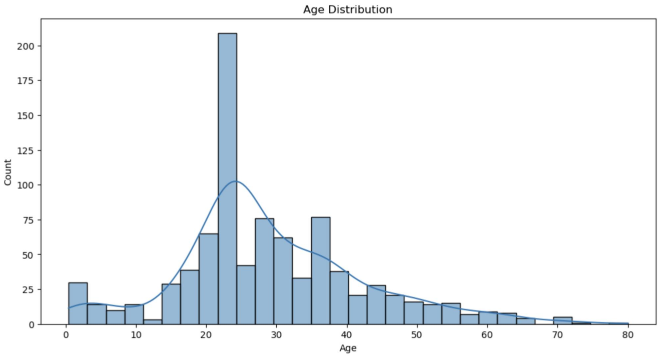
2.2 Data Overview

- The dataset includes features like "Pclass" (ticket class), "Sex", "Age", "SibSp" (siblings/spouses aboard), "Parch" (parents/children aboard), "Ticket", "Fare", "Cabin", and "Embarked" (port of embarkation).

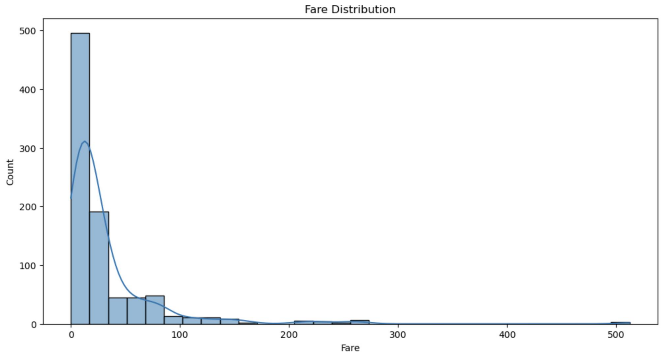
- "Survived" is the target variable (0 = No, 1 = Yes).

- Visualizations:

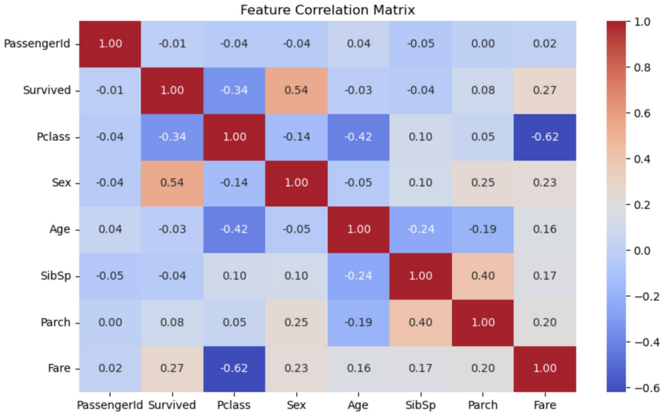
  - Age Distribution: The age distribution is right-skewed, with a peak around 25-30 years old. A noticeable number of passengers were children below 10, and fewer passengers were older than 60. This visualization helps in understanding potential missing age values and its impact on survival.



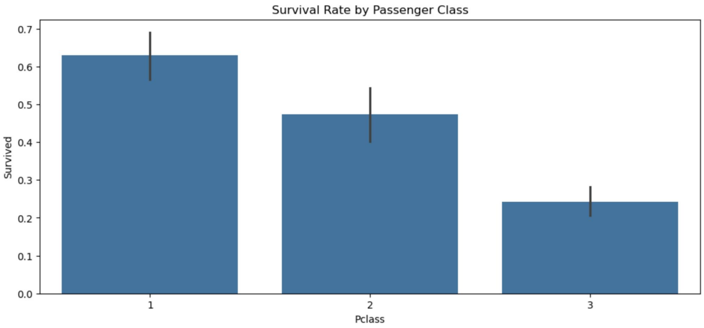
  - Fare Distribution: The fare distribution is highly skewed with most passengers paying low fares, while a few paid significantly higher fares exceeding 500. The skewness indicates that fare transformation might be necessary to improve model performance.

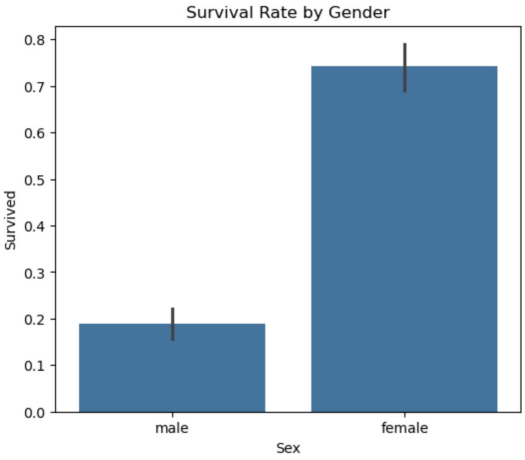


- Feature Correlation Matrix: A heatmap was generated to examine feature correlations. Sex (being female) has a strong positive correlation with survival, while Pclass and Fare also exhibit significant relationships. Higher-class passengers had better survival rates, while lower-class passengers faced higher risks.



- Survival Rate by Passenger Class & Gender: Higher survival rates in first-class passengers, and women had significantly higher survival rates.

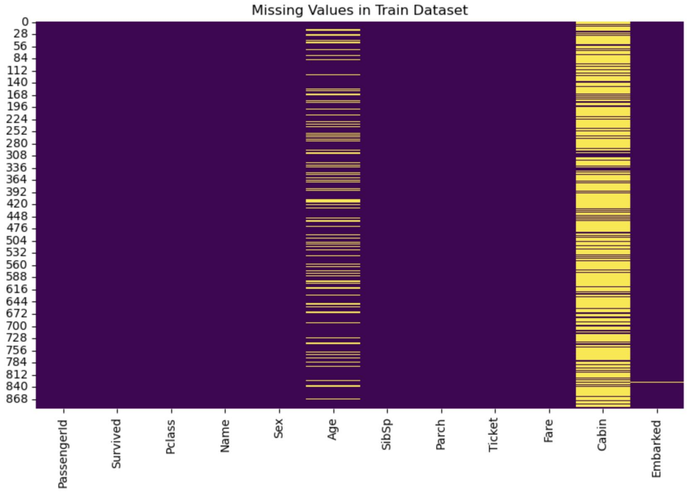


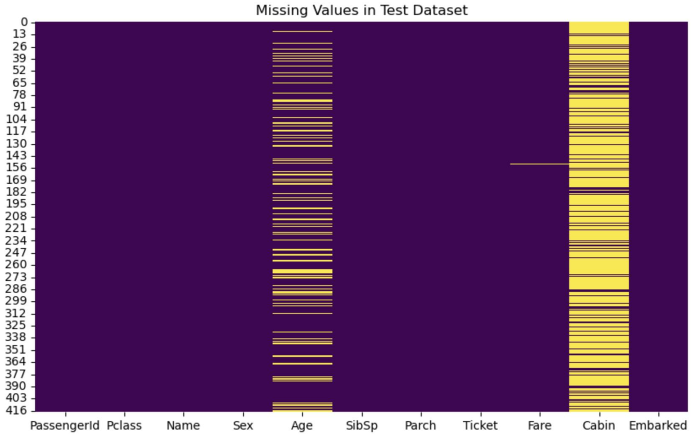


3. Data Preprocessing and Feature Engineering

3.1 Handling Missing Values

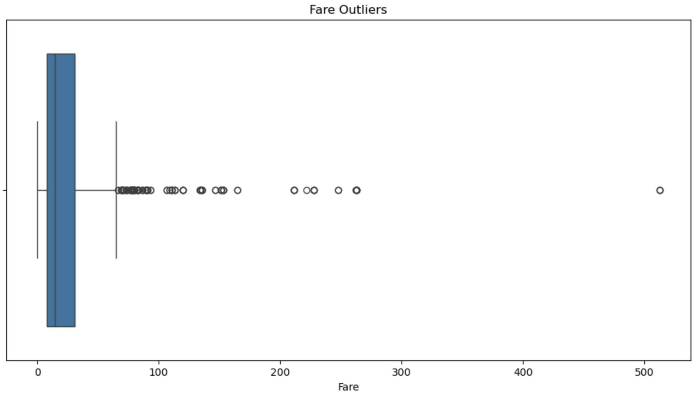
Missing Values Heatmap: Showed significant missing values in "Age", "Fare", "Cabin", and "Embarked".

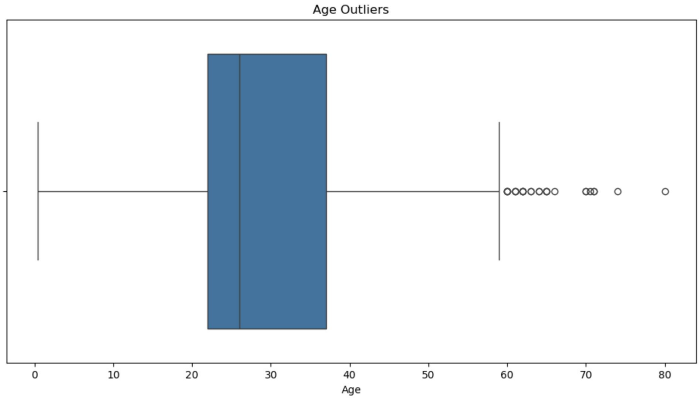




- "Age": Missing values were imputed with the median age within each "Pclass" group.
- "Embarked": Missing values were filled with the most frequent embarkation point ("S" for Southampton).
- "Fare": The single missing value was imputed with the median fare.

3.2 Handling Outliers





- Boxplots revealed extreme values in "Fare" and "Age".

- Outliers were removed using the 99th percentile for both "Fare" and "Age" to ensure robust model training.

3.3 Feature Engineering

- Encoding Categorical Variables:

- "Sex" was converted to numeric (0 = Male, 1 = Female).
- "Embarked" and "Pclass" were one-hot encoded.

- Family Features:

- "FamilySize" was created as "SibSp + Parch + 1".
- "IsAlone" was introduced, indicating whether the passenger traveled alone.

- Title Extraction:

- Titles (e.g., Mr, Mrs, Miss) were extracted from names and grouped into meaningful categories.
- Unique titles in the dataset included 'Mr', 'Mrs', 'Miss', 'Master', 'Noble', 'Officer', and 'Other'.

3.4 Feature Scaling

- "Age", "Fare", and "FamilySize" were standardized using "StandardScaler" for improved model convergence.

4. Model Training and Evaluation

4.1 Model Selection

The following models were trained and evaluated:

- Logistic Regression: Interpretable linear classification model.

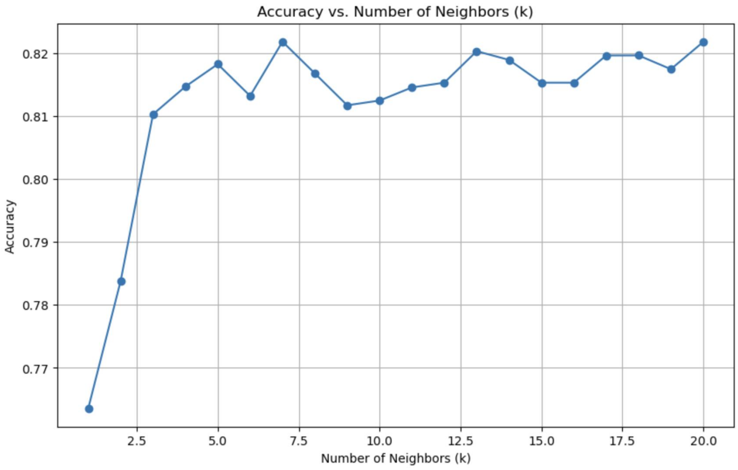
- Linear Discriminant Analysis (LDA): A dimensionality reduction technique for classification.

- K-Nearest Neighbors (KNN): A non-parametric classifier optimized with Grid Search.

4.2 Hyperparameter Tuning

- "GridSearchCV" was used to optimize KNN parameters ("k", "weights", and distance metric).

- From the line chart and GridSearch output, the best KNN model had "k=13", "p=1" (Manhattan distance), and "weights=uniform".



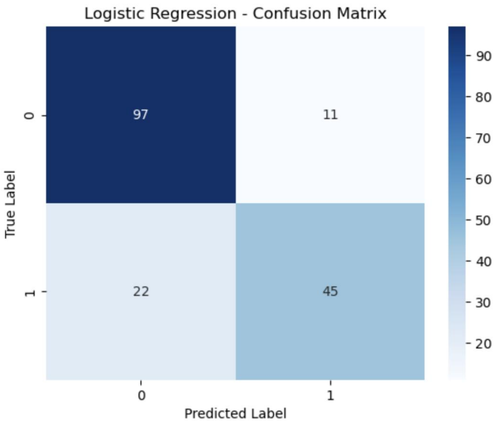
4.3 Model Performance Metrics

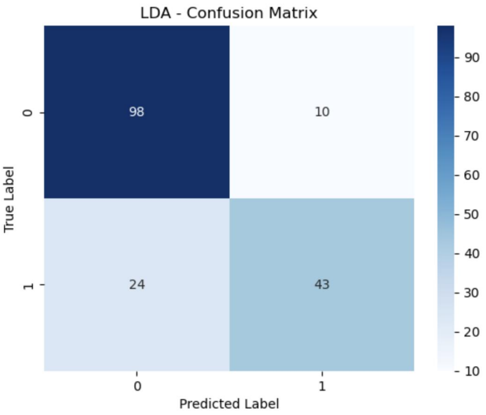
- Accuracy Scores (Cross-Validation Results):

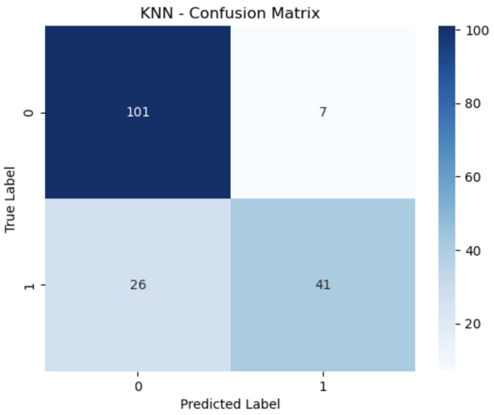
- Logistic Regression: 81%
- LDA: 81%
- KNN: 81% (Best parameters from Grid Search)

- Confusion Matrix Analysis:

- True positives and false negatives were visualized for each model.
- KNN performed best for classifying survivors correctly but had more false positives.





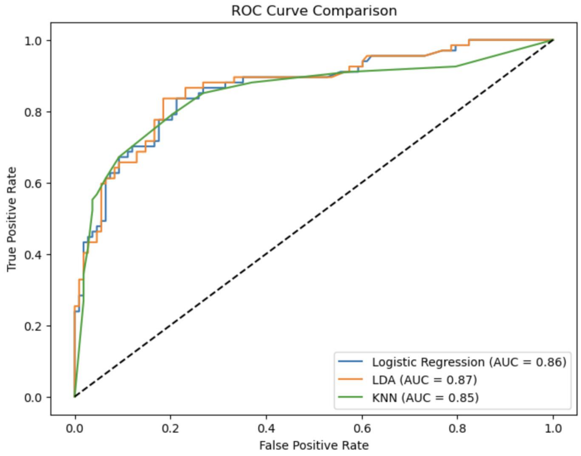


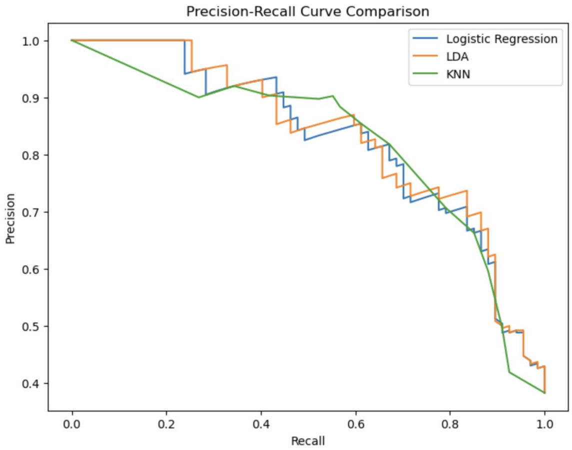
- ROC Curve & AUC Comparison:

- Logistic Regression achieved the highest AUC.
- Visual comparison of model performances showed a clear ranking.

- Precision-Recall Curve:

- Precision and recall trade-offs analyzed for imbalanced class distribution.





5. Results and Interpretation

- Logistic Regression provided the best balance between accuracy and interpretability.

- LDA performed similarly but was more sensitive to assumptions about data distribution.

- KNN was highly dependent on parameter tuning, with "k=13" being optimal.

- The ROC and Precision-Recall curves highlighted Logistic Regression as the most robust model.

6. Conclusion and Next Steps

This analysis successfully developed a predictive model for Titanic survival. Future improvements include:

- Exploring ensemble models like Random Forest and Gradient Boosting.

- Enhancing feature engineering with additional text analysis on names and tickets.