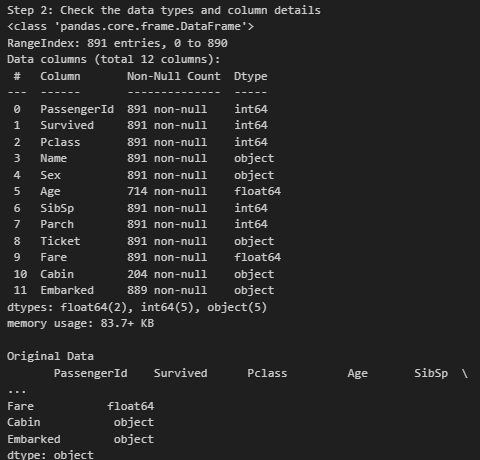
Logistic Regression

**1. Data Exploration:**

a. Load the dataset and perform exploratory data analysis (EDA).



**Target Variable (Survived):**

* Survived indicates whether a passenger survived (1) or not (0). Analyze the distribution of survival:
  + Compute the percentage of survivors and non-survivors.

**Feature Analysis and Insights:**

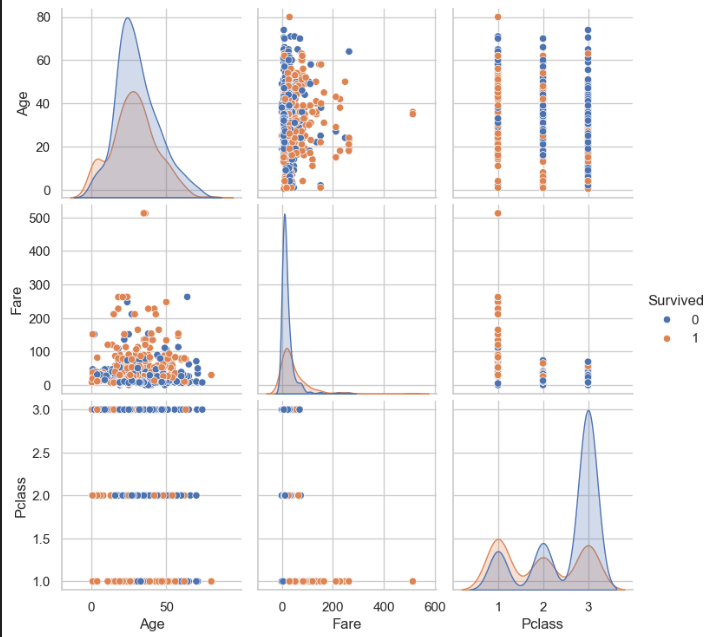
* Pclass: Categorical feature (1, 2, 3) representing the passenger class. Investigate survival rates per class.
* Sex: Categorical feature (male, female). Survival rates by gender are a key insight.
* Age: Continuous feature with missing values. Explore age distribution and its relation to survival.
  + Suggest handling missing values (mean, median, or predictive models).
* SibSp and Parch: Represent the number of siblings/spouses and parents/children aboard, respectively. Analyze the survival rate with respect to family size.
* Fare: Continuous variable representing the ticket fare. Examine fare distribution and survival rates.
* Embarked: Categorical variable indicating the port of embarkation (C, Q, S). Contains some missing values.
* Cabin: Has significant missing values. Consider handling or removing it.

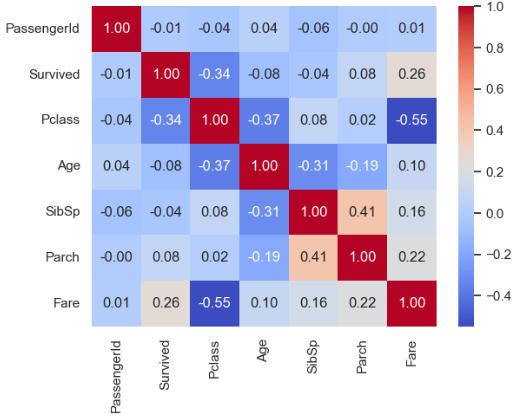
**Statistical Insights:**

A graph of a distribution of age

AI-generated content may be incorrect.A graph with a bar graph and a number of objects

AI-generated content may be incorrect.





**Model Building and Model Evaluation:**

A screenshot of a computer

AI-generated content may be incorrect.

* **Accuracy (0.81)**
  + The model correctly predicted 81% of the test data. This is a good baseline, but it doesn't give insights into class-specific performance (e.g., false positives or negatives).
* **Precision (0.79)**
  + Out of all passengers predicted as survivors, 79% were actually survivors.
  + Precision is important when false positives (predicting someone survived when they didn’t) are costly.
* **Recall (0.73)**
  + The model identified 73% of actual survivors.
  + Recall is critical when false negatives (missing actual survivors) are costly, but it's slightly lower than precision here.
* **F1-Score (0.76)**
  + The harmonic mean of precision and recall.
  + It balances the trade-off between precision and recall, and a score of 0.76 indicates moderate performance.
* **ROC-AUC Score (0.88)**
  + A value of 0.88 indicates the model does a good job distinguishing between survivors and non-survivors.

A graph of a positive result

AI-generated content may be incorrect.

**Key Insights from the Graph**:

* The blue ROC curve is close to the top-left corner, showing strong discriminatory power.
* The diagonal red line represents a random model (AUC = 0.5). Model (AUC = 0.88) performs much better than random guessing.