Feature Selection and Correlation Analysis Report

Possible Target Columns for Prediction

Based on the dataset, the most obvious target column for prediction is `Survived`, which indicates whether a passenger survived the Titanic disaster.

Selected Features for Predicting `Survived`

Using the SelectKBest method with ANOVA F-value as the score function, the top 5 features selected for predicting `Survived` are:

1. **Pclass**

2. **Parch**

3. **Fare**

4. **Sex_male**

5. **Embarked_S**

Correlation Matrix Analysis

The correlation matrix provides insights into the relationships between different features. Here are some key observations:

- **Survived**:

- Positively correlated with `Fare` (0.257)

- Negatively correlated with `Pclass` (-0.338)

- Weak correlations with other features

- **Pclass**:

- Negatively correlated with `Fare` (-0.549)

- Negatively correlated with `Age` (-0.340)

- Weak correlations with other features

- **Fare**:

- Positively correlated with `Parch` (0.216)

- Positively correlated with `SibSp` (0.160)

- Weak correlations with other features

Interpretation

Based on the feature selection and correlation analysis, the most relevant features for predicting whether a passenger survived (`Survived`) are:

- **Pclass**: The class of the ticket purchased (1st, 2nd, or 3rd class).

- **Parch**: The number of parents/children aboard the Titanic.

- **Fare**: The fare paid for the ticket.

- **Sex_male**: The gender of the passenger (male).

- **Embarked_S**: The port of embarkation (Southampton).

These features have been identified as the most significant predictors of survival in the Titanic dataset. The correlation matrix further supports the importance of these features, particularly `Pclass` and `Fare`, which show significant correlations with the target variable `Survived`.

Conclusion

For building a machine learning model to predict the survival of passengers on the Titanic, the target column should be `Survived`. The most relevant features to include in the model are `Pclass`, `Parch`, `Fare`, `Sex_male`, and `Embarked_S`. These features have been selected based on their statistical significance and their correlations with the target variable.