# Kaggle Titanic ML Submission Report

A screenshot of a chat

AI-generated content may be incorrect.

## 1. What data preprocessing methods did you apply?

To prepare the data for modeling, I applied several preprocessing steps to ensure the features were clean, consistent, and meaningful:

- Extracted the 'Title' from the 'Name' column to capture social status and grouped rare titles into a single category called 'Rare'. This was then encoded into numeric values for model compatibility.  
- Converted the 'Sex' column into binary form (0 for male, 1 for female) and handled missing values in the 'Embarked' column by imputing the most common value ('S'), followed by mapping it to numerical categories (S=0, C=1, Q=2).  
- Missing 'Age' values were filled using median values based on gender and passenger class, which were then binned into five age groups to capture age distribution patterns.  
- Missing values in the 'Fare' column were filled with the median fare. The 'Fare' feature was then discretized into four categories using quartile-based binning.  
- Created new features such as 'FamilySize' (by combining 'SibSp' and 'Parch') and 'IsAlone' (a binary feature indicating if a passenger was alone).  
- Introduced an interaction term 'Age\*Class' to capture any compounded effect of age group and passenger class.  
- Finally, I removed irrelevant columns like 'Name', 'Ticket', 'Cabin', and 'PassengerId' to reduce noise and dimensionality.

## 2. What machine learning model(s) did you apply?

I used an ensemble learning approach to improve prediction accuracy by combining the strengths of multiple classifiers. Specifically, I employed the VotingClassifier from scikit-learn with soft voting, which considers predicted probabilities rather than hard class labels. The ensemble consisted of:  
- Logistic Regression (with increased max\_iter for convergence)  
- Support Vector Classifier (SVC) with probability enabled  
- Random Forest Classifier with 100 estimators and a fixed random seed for reproducibility  
This combination was chosen to leverage both linear and non-linear patterns in the data.

## 3. What was your best score and ranking on the Kaggle leaderboard?

My best public leaderboard score for the Titanic competition was 0.78468. This score placed me at position 2677 on the leaderboard. Although this score indicates a solid understanding of the basics of classification and ensemble techniques, I am continuing to iterate and explore advanced feature engineering and model tuning to push the score even higher.

## 4. Model Evaluation on Training Data

To assess the model's performance on the training set, I evaluated several metrics:  
- Accuracy: 84.29%  
- F1 Score: 0.7785  
- Confusion Matrix:  
 [[505 44]  
 [ 96 246]]

The confusion matrix indicates that the model correctly predicted 505 true negatives and 246 true positives. There were 44 false positives (predicted survived, but actually did not survive) and 96 false negatives (predicted did not survive, but actually survived). This suggests the model performs well but has room for improvement, particularly in reducing false negatives. The F1 score of 0.7785 balances precision and recall, making it a good indicator of model performance given class imbalance.