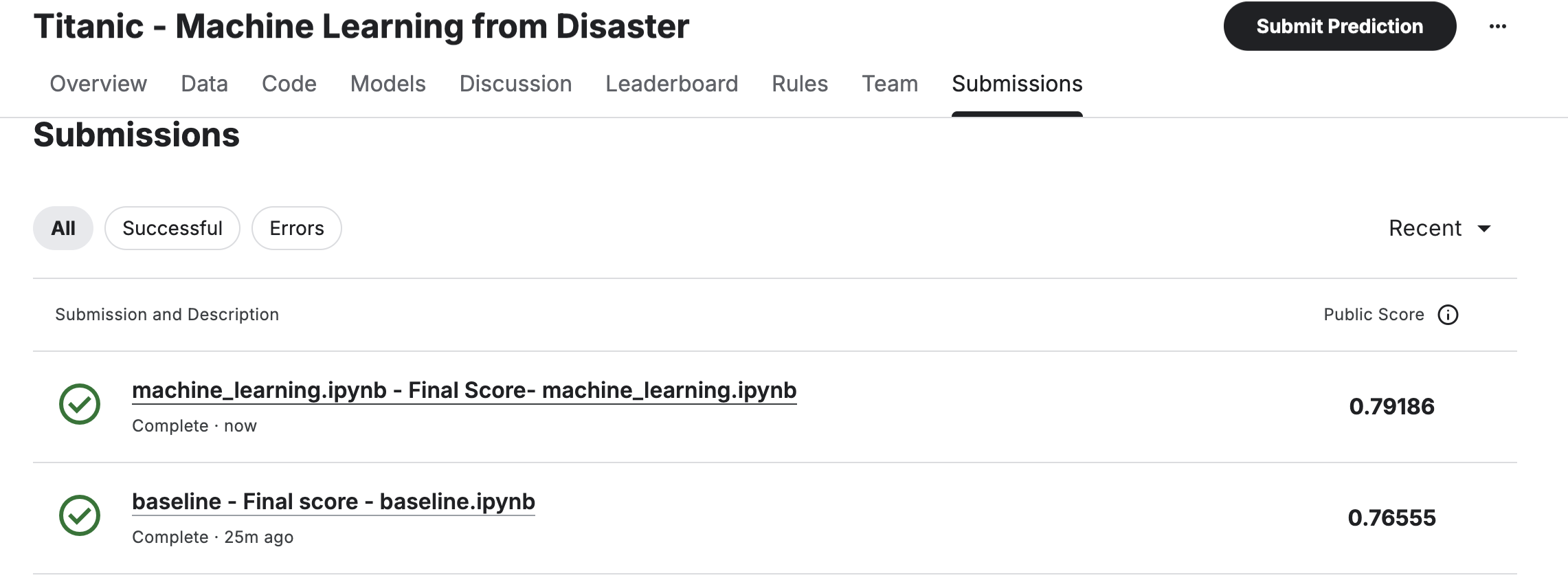
**HW-5**

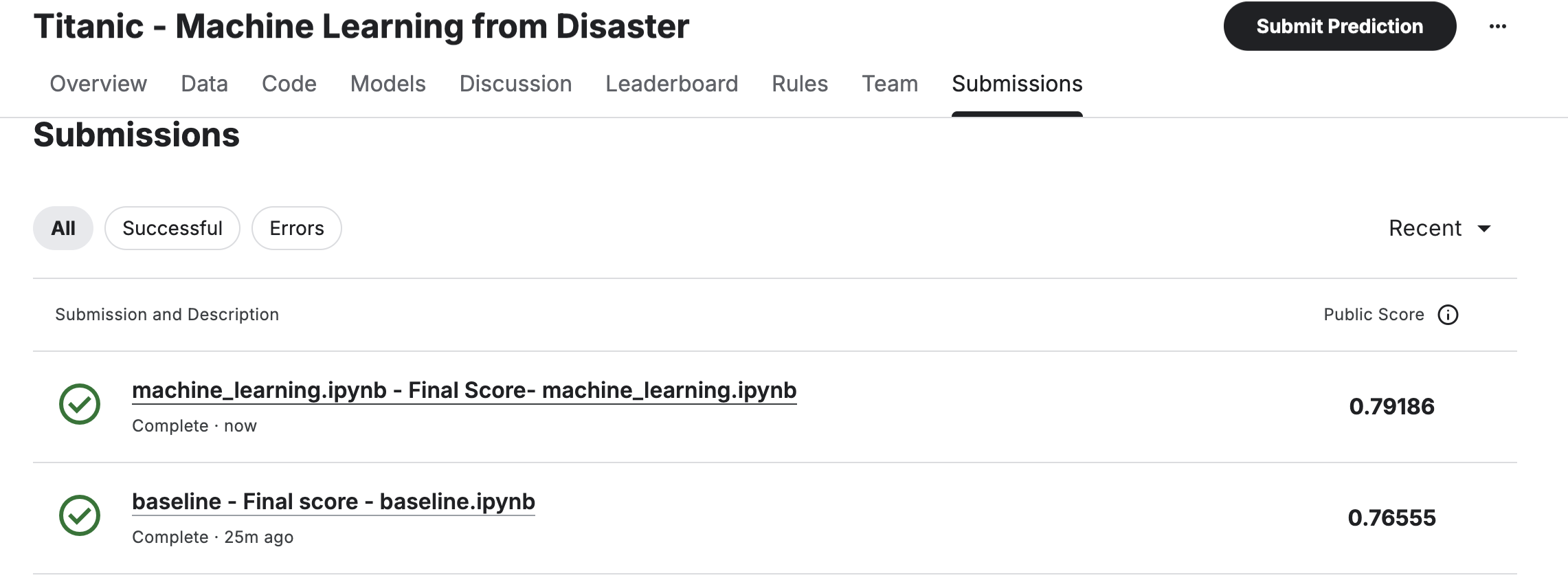
**Kaggel Machine Learning Contest**

**Titanic – Machine Learning from Disaster**

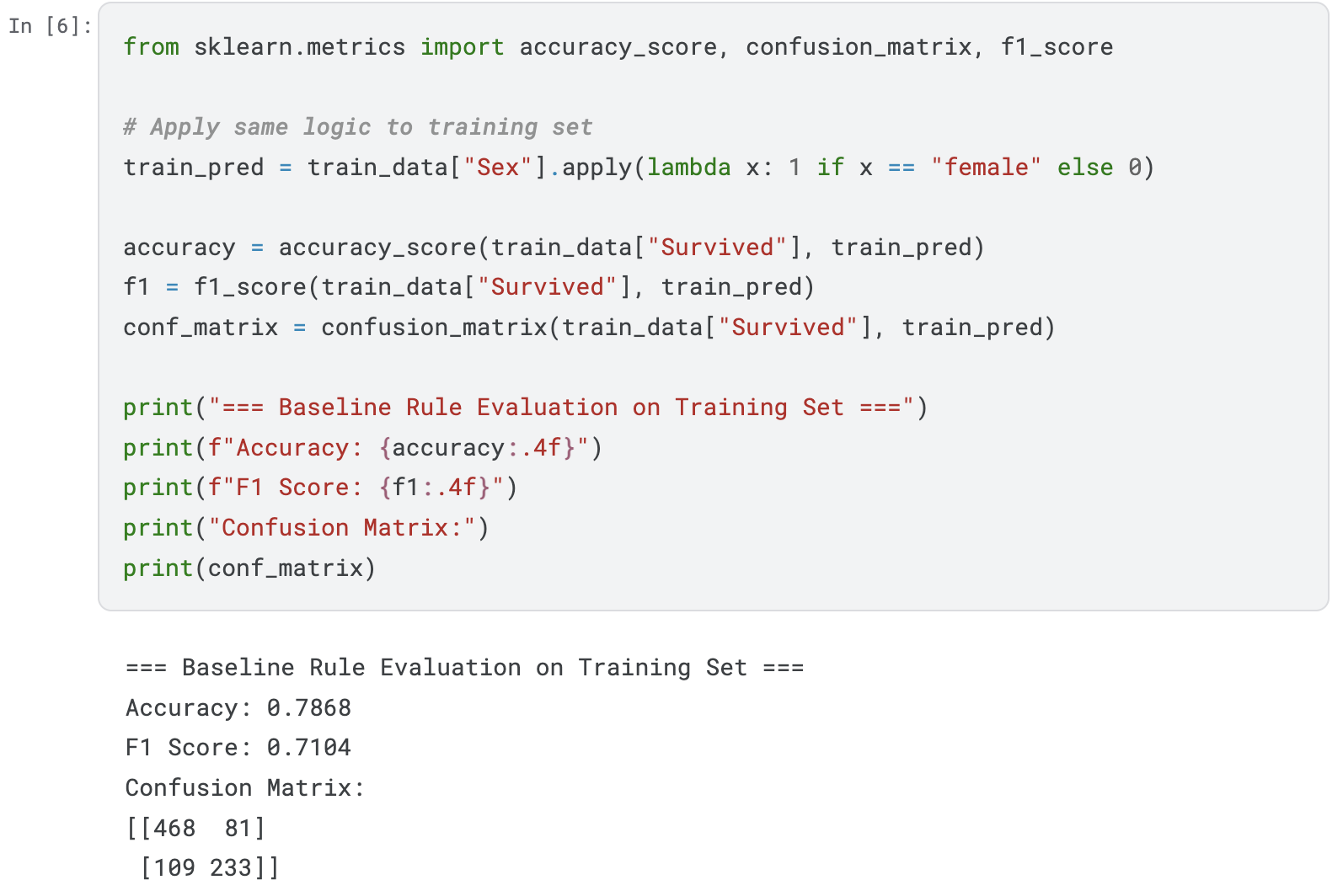
1. **baseline.ipynb**

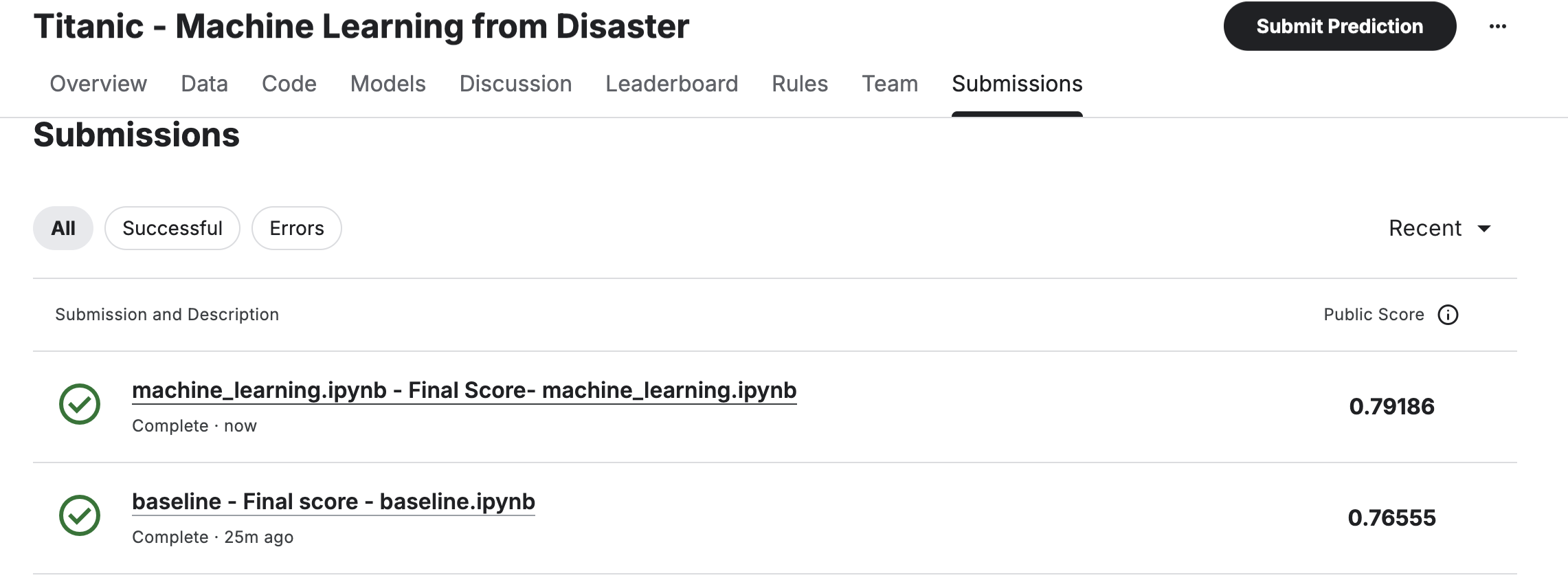
**Leaderboard score**





**Evaluation score:**



1. **Machine\_learning.ipynb**

**Questions:**

1. **What data preprocessing methods did you apply, such as feature scaling with a min- max method or one-hot encoding for a feature?**

I applied a combination of **imputation, feature scaling, and categorical encoding** using a **scikit-learn preprocessing pipeline**:

* **Missing Value Imputation:**
  + Replaced missing Age and Fare values with their **median**.
  + Filled missing Embarked values with the **most frequent value (mode)**.
* **Feature Engineering:**
  + Created a new feature FamSize = SibSp + Parch + 1 to represent total family members aboard.
  + Dropped non-predictive columns such as Name, Ticket, and Cabin.
* **Feature Scaling:**
  + Applied **StandardScaler** (z-score normalization) to numerical features (Age, Fare, FamSize) within a preprocessing pipeline.
* **Categorical Encoding:**
  + Applied **One-Hot Encoding** to categorical variables (Sex, Embarked, Pclass) to convert them into numeric form.
* **Pipeline Integration:**
  + Combined all preprocessing steps into a unified **ColumnTransformer pipeline** to ensure consistent transformation during training and prediction.

1. **What machine learning model(s) did you apply, such as SVM or random forest?**

I used a **Random Forest Classifier**, an ensemble learning algorithm that builds multiple decision trees and aggregates their results for improved accuracy and robustness.

* **Key Parameters:**
  + n\_estimators = 150 (number of trees)
  + max\_depth = 6 (controls tree depth to prevent overfitting)
  + random\_state = 42 (for reproducibility)
  + n\_jobs = -1 (uses all CPU cores for faster training)
* The model was embedded within a **scikit-learn Pipeline**, ensuring preprocessing and classification were executed in one streamlined process.

1. **What are performance evaluation results on training dataset, including accuracy, confusion matrix, and F1 score, on the training dataset?**

Evaluation was done on the training set after fitting the pipeline.

The **accuracy** is 0.8732 and the **F1 score** is 0.8169

**Confusion Matrix:**

[[526 23]

[ 90 252]]

**4. What the best score and ranking at Leaderboard at Kaggle did you receive?**

The highest score received is by using the Random Forest Classifier. It scored 0.792. The rank in the leaderboard is 917.