In this report, we document the approach we took to predict the survival of passengers aboard the Titanic using a random forest classifier. We describe the methods and techniques we used, including data preprocessing, feature engineering, model training, and model evaluation. We also report the performance of our model based on cross-validation on the training dataset.

Data Preprocessing The first step in our approach was to preprocess the data, which involved cleaning and transforming the raw data into a suitable format for our machine learning model. We used the following steps to preprocess the data:

Handling missing values:

We filled the missing 'Age' values with the median age from the training dataset.

We filled the missing 'Fare' values with the median fare from the training dataset.

We filled the missing 'Embarked' values with the mode (most frequent value) from the training dataset.

Dropping irrelevant features:

We removed the 'Name', 'Ticket', and 'Cabin' columns from the dataset. The 'Name' and 'Ticket' columns were not expected to have a significant impact on the survival prediction. The 'Cabin' column was dropped due to a large number of missing values, which would make it difficult to use as a reliable predictor without significant imputation or additional feature engineering.

Feature Engineering Next, we performed feature engineering to transform the categorical features into numerical values that our model could understand. We used the following techniques:

One-hot encoding:

We applied one-hot encoding to the 'Sex', 'Embarked', and 'Pclass' categorical features. This created binary columns for each category, with a value of 1 indicating the presence of the category and 0 otherwise.

Model Training After preprocessing the data and performing feature engineering, we trained a random forest classifier on the training dataset. We chose the random forest classifier because it is a robust and versatile algorithm that can handle non-linear relationships between features and can often achieve high predictive performance with minimal tuning.

We used the following hyperparameters for our random forest classifier:

**n\_estimators**: 100

**random\_state**: 42

Model Evaluation To evaluate the performance of our model, we used 5-fold cross-validation on the training dataset. This provided us with a more reliable estimate of the model's performance, as it assessed the model on different subsets of the data, reducing the risk of overfitting.

We obtained a mean cross-validation score of approximately 0.807, which indicates that our model correctly predicted the survival status for about 80.7% of the passengers on average during cross-validation.

Conclusion In this report, we have presented our approach to predict the survival of Titanic passengers using a random forest classifier. We preprocessed the data, engineered features, and trained a random forest classifier to achieve a mean cross-validation score of approximately 0.807 on the training dataset. While this performance is reasonably good, there is room for improvement by exploring other classifiers, tuning hyperparameters, or employing more advanced feature engineering techniques. Additionally, it is important to evaluate the model on a separate test dataset to get a more accurate understanding of how well it generalizes to unseen data.