Carnegie Mellon University
Africa

Name: Emmanuel Iradukunda

AndrewID: eiraduku

DATA, INFERENCE & APPLIED MACHINE LEARNING (COURSE 17-975)
Kaggle Bonus Report.

Date: 4th December 2021

Libraries used:

pandas as pd, numpy as np, math, matplotlib.pyplot as plt, ttest_1samp from scipy.stats, ttest_ind_from_stats, from scipy.stats, statistics, RandomForestClassifier, from sklearn.ensemble

Steps used

The first step was to read the data (train, test, and gender submission), secondly some features were extracted, and those features are 'Pclass', 'Sex','Age','Survived','SibSp','Parch','Fare','Embarked' and were extracted to both train and test set.

Next step was to fill the nan value for Age feature for both train and test and the nan were filled with the average of the ages. Moreover, male gender was replaced randomly with 1 and female with 0, S Embarked by 1, C Embarked by 2, and Q Embarked by 3.

I used random forest classifier, with 100 estimators, maximum depth of 5, number of job was made to -1 and criterion was entropy. After that the model was fitted with the trained selected variables and survival variable and find prediction of the test selected features.

The appropriate model was accurate so far was Random Forest classifier with 100 number of estimators

```
rf = RandomForestClassifier(n_estimators=100,max_depth=5,n_jobs=-1,criterion='entropy')
```

The accuracy resulted with Random Forest classifier, according to the extracted data and the numbers filled to transform categorical features is 0.85. I rejected the others that were not providing the accurate prediction and stay with Random Forest classifier at the end.

```
The accuracy = 0.8540965207631874
```

Finally, I combined my survival prediction with the passenger ID in its data frame and checked the result with different submission depending on the adjustment I made on the number of depths, estimators, and other parameters. The obtained result from the competition was

Thank you.