Name : Yang, Po-Yen
Student ID : 20561878

Group 2

1. Summary of the report
   Jasper is working on the Titanic dataset, and he has transformed the data into numerical values as he would perform machine learning algorithms on these data, and machine learning methods tend to accept numerical values as inputs. For filling in the missing values of the Pclass/Embarked ordered pair, he used the median fare that was closest to the fare of the missing values. While filling in the missing values for age with the mean or median would result in the distribution heavily centered around the middle, which is inconsistent with the original data, hence, he proposed the approach of applying regression on each missing age value. Additionally, he generated some new features from the dataset, and drew a correlation heatmap to show which variables were more important. He also plotted some graphs for visualization. The methods that he used are the k-Nearest Neighbors (KNN), Logistic Regression, and Random Forest. For the KNN model, he used n = 29, and the training accuracy and validation accuracy that he received are 75% and 70.9% respectively. As for logistic regression, he utilized 10-fold cross validation to estimate the test error and received the average validation training accuracy of 0.8249. For the random forest, he received validation accuracy of 0.835 with 10-fold cross validation, 150 estimators, and maximum depth of 7. Lastly, he proposed some future analysis for the Titanic dataset.

2. Strengths of the report
   (1) He performed different methods for filling in the missing values for different variables as different methods are more suitable for different variables.
   (2) He drew lots of graphs which are great for visualization.
   (3) For the random forest method, I can see that he did lots of tests to see which parameters would result in a better accuracy for predicting this dataset.
   (4) Lastly, he provided many future analyses that he or other people may be able to improve in the future.

3. Weaknesses of the report
   (1) He did not provide a conclusion, the last part of his report is about future analysis.
   (2) The parts in which he mentioned different methods are very skewed, as he mentioned a lot of features and testing process for random forest, but only two lines for logistic regression.

4. Evaluation on quality of writing (4)
   I think the report is clearly written as it did express his thoughts clearly, and there were figures for visualization and there were no typos. I think the parts that this report can improve are as follows :

(1) He could first put the visualization as it is more related to the data, then show the feature selection as it is more related to the models that he wrote afterwards.
(2) He could add a conclusion at the very last to make the report more complete.

5. Evaluation on presentation (4)
   The presentation was clear and well organized, and the language flow was fluent. The slides were also clear and well prepared as he used different colors for highlight points. However, there are some points that I think he can improve, which are as follows :
   (1) He didn't say much about the reasons why he chose these models, and the reasons for setting the parameters for the models
   (2) He didn't mention how he chose the subset for AdaBoost and he should talk more about the strengths and weaknesses of each model

6. Evaluation on creativity (3)
   I don't think the work proposes any new ideas, as I believe these methods have been proposed for the Titanic dataset before. However, I do think that KNN can be counted as one of the state-of-the-art results, additionally, the accuracy for the random forest is relatively good and I may want to utilize the parameters of his random forest model, which I think is an extension of existing ideas.

7. Confidence on my assessment (3)
   I have carefully read the paper and checked the results.