Group 2 Summary:

This group use the Titanic data set to work on a classification problem.

They use several technique in data cleaning and data processing stage. For instance, dropping variable, fill in missing data with MICE and PMM, and feature transformation. They also check for multicollinearity with heatmap and discovered that PClass and Deck_X is highly correlated, so Deck_X is deleted.

They apply 4 different models with different variation in this project, include 1. Logistics regression, 2. KNN, 3. Random Forest, and 4. Decision Tree (AdaBoost). Other technique they use in the model selection stage are cross validation, backward subset selection, gird search and Ridge regularization.

There are several key finding in this project, include 1. Using MICE and PMM in data imputation may achieve a closer to original distribution result than mean and median replacement method. 2. In AdaBoost, a good mean CV score may not lead to good test performance because of the large amount of estimators. 3. KNN with K=1 is susceptible to noise and outliers.

The final result is that random forest with n_estimators =150 and max_depth=2 achieve best result in term of test accuracy.

Strength:

The report described everything in great detail, for example, it share some tips on importing data to Google Colab. The visualization in the report is also clear. For example, the graph that compare the distribution of variable “Age”, clearly demonstrate how MICE perform.

Weakness:

Only minor mistake is found. For example, a typo in 4.4 of the report “mean error” should be “mean accuracy”.

Evaluation:

Criteria	**Points**
Quality of writing	5
Presentation	4
Creativity	5
Confidence on my assessment	3

Comment:

1. Everything is written concisely, with excellent use of figures.
2. The presentation is clear, and the slide also comes with great figures and examples.
3. There is a lot of adaptations and creative idea. For example:

1. When trying to fill in the “Embarked” empty data, they use Pclass/Embarked ordered pair and filled in with the Pclass/ Embarked pair with closest median fare to the missing values. This approach is useful.
2. Adaptation of MICE and PMM to predict missing values.
3. Using VIF in refining logistic regression model