Group 1 Cai and Song

Summary
The group project is about finding a suitable machine learning model on the Kaggle dataset - Survival prediction of the Titanic. Three major machine learning models are adopted, which are logistic regression, decision tree, and random forest. They divided the project into several components, which are statistical assumption, data overview and processing, model diagnostics, model selection, and result.

Various data exploratory and data pre-processing techniques are adopted, they can showcase their understanding of some special cases in the datasets, including missing values or transforming certain attributes.

The correlation coefficient is used as a model checker for the multicollinearity of the model. All data passed the multicollinearity check.

They adopted several models including logistic regression, KNN, SVM, Decision Tree, and Random Forest. Several modeling techniques are adopted, including cross-validation, and forward/backward feature selection. Random forest is the best model with 0.822 accuracies.

Strengths of the report.
Detailed information and motivation are explained in the report, with multiple data visualization graphs presented. Multiple models are used, showcased their effort in implementing the model and strong analytic skills are shown in the feature importance part.

Weakness of the report.
Overall, the report is well designed. However, the majority of the report focused on the data exploration and pre-processing part, which may be better to shift the focus to the model selection part. Readers may want to know more about the motivation, advantages, and disadvantages of selecting the models, instead of just based on accuracy.

• Evaluation on quality of writing (1-5): Is the report written? Is there good use of examples and figures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the paper, and provide details of typos.
5
The report is written in a clear way with subsections, objectives and methodology stated clearly. Sufficient graphs and visualization to support the argument. No issues with typos, formatting, or references are found.

• Evaluation on presentation (1-5): Is the presentation clear and well organized? Are the language flow fluent and persuasive? Are the slides clear and well elaborated? Please make suggestions to improve the presentation.
5
The presentation is clear and well-organized. The presentation is easy to follow with fluent language. The presentation explains clearly the motivation and results of data exploration, data processing, and model selection. I would like to suggest that the time allocation of the presentation can be more balanced with more focus on the model and conclusion part.

• Evaluation on creativity (1-5): Does the work propose any genuinely new ideas? Is this a work that you are eager to read and cite? Does it contain some state-of-the-art results? As a reviewer, you should try to assess whether the ideas are truly new and creative. Novel combinations, adaptations, or extensions of existing ideas are also valuable.
5

The data exploratory part is highly creative. They carefully examined the relationship between several important features and response variables, i.e. the survival rate. Lots of information is provided and the idea behind model selection can inspire readers. It combines several state-of-art machine learning models and ranks the performance of several models. It provides a great introduction to readers who are new to machine learning, and this is a work I would eagerly read and cite.

• Confidence in your assessment
3, I have carefully read the paper and checked the results