Group 2

• Summary of the report.

The report guides us through a classic problem: Predict Survivors in Titanic Disaster. It focuses on 2 main aspects: features and Machine Learning models. Features are treated quite carefully, by filling missing value by an interesting method MICE, transforming features and examining their relationship, then finalized the list of features used to train ML models. In the next part, it introduces several methods and the metrics to evaluate them. The last part shows tuning process, which yield a better test score.

• Describe the strengths of the report.

It detailly describes workflow of the project. It shows the focus on processing and analysing features of the dataset which is very important in Data Science. Many experiments were conducted to compare models from different methods and different settings.

• Describe the weaknesses of the report.

There are several points

- Need to relate the importance of features to the real scenarios. Many articles did inform the disaster to us, such as https://www.history.com/topics/early-20th-century-us/titanic#section_8. It gives a precious information that "In compliance with the law of the sea, women and children boarded the boats first; only when there were no women or children nearby were men permitted to board." That really explained why women and children have higher chance to survive
- Not persuasive explanation of not using Precision + Recall: I think this case is different from disease diagnosis in which sensitivity for positive class is very important.

• Evaluation on quality of writing (1-5): Is the report clearly written? Is there a 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. 4

It's a bit wordy, lack of visualization. The first figure in page 4 does not have legend

The organization is okay in general, but the subsection notion in section 4 seems not – after introducing 3 models, it jumps to experiment. I think it should be split to 4.1 Model: 4.1.1 …; 4.2 Experiment.

• 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

He speaks fluently, and successfully convey all ideas and information from report to the presentation.

• 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. 4

This work does apply an unfamiliar method to fill missing values. It also applies VIF to check multicollinearity and verify its implication by experiments.

About the self-proposed question, "I then considered if it would be possible to try and simplify the logistic regression model to prevent the model from overfitting on irrelevant features and avoid collinearity, just like in the random forest case." I think you may try ensemble of Logistics Regression with random picked features (from a set of features) for each

• Confidence on your assessment (1-3) (3- I have carefully read the paper and checked the results, 2- I just browse the paper without checking the details, 1- My assessment can be wrong) 3

But not run the code again