Textual description: Did the team include a textual description of their solution? Is it understandable? Is the described process correct?

Code quality: Is the code readable and concise? Does it use the powerful features of the libraries, or does it re-implement everything from scratch?

Results: Is the final result correct? Are all the assumptions well justified? Are there textual comments/visualizations to convince you of the final result?

Textual description: - Task 1.1 should have been explained a lot more, the functions that are used and the results they give. Task 1.2 what is you null hypnotises here? All of task 1 is generally poorly commented, I get no insight on the thought process that lead to the solutions you created, neither do I get any reasoning for why the null hypnotises can be rejected or the math/statistics behind it (can’t tell if you understand what you are doing). Again, for task 2; lack of explanation. Task 2.C why is this threshold chosen?

+

Code quality: - Start code for task 1 poorly commented, all code poorly commented. Task 2.B error codes, most likely from the calculation of precision (should check if the denominator is 0). For task 2.C I don’t think load\_iris, SelectKBest and chi2 was allowed to be used, only the two models, do you even use your greedy search algorithm???

+ Uses the libraries in a good way.

Results: - Task 1.1 what significant level did you use that made you reject you null hypnotises? I’m not sure if the spearman correlation could be used as a p-value. Task 2, training and testing data is standardized separately, but test data should be standardized with the mean and std from the test set, because……. Since you want to use the F1 score as a metric you need to be aware that this value might not always be defined as precision might divide by zero, this is not done (lucky that this edge case never happen).

+ Beautiful plots for task 1! Task 2, results are good and seems to make sens.

Review:

To start of with there is very little explanation of what your code does. This gets a little better for task 2 a and b, but it is only some comments in the code and not textual descriptions. This makes it hard to follow your process of thought through out the exercise. There are a lot of different aspects of the tasks that could have been explained, e.g. in task 1 you use the p-value returned by scipy.stats.spearmanr to reject or keep your H0 hypothesis, what is this hypothesis? It is only specified for part A, is it the same for all other tasks? What is the alpha that you have chosen as a limit for rejecting H0? All these things should have been specified or explained. Also, in the documentation for scipy it says that “The p-values are not entirely reliable but are probably reasonable for datasets larger than 500 or so” this is also a thing that easily could have been mentioned to show that you have thought about it and that you know what the results you are using are good for. Good visualizations in task 1!

Task 2 are also lacking more textual description other than the results and conclusions. For standardizing you test data it could be better to use the mean and standard deviation used to standardize the training set. This is because of it is hard to know and prove that the data in the test set is fully representative of the data in the training set, even though the data is shuffled before it is split. There is also a lot of runtime warnings in task 2. In part B this might be because of the way you calculate precision, the denominator might be zero for some thresholds which in turn will, together with recall, might affect the F1 score. This could have been mentioned in a textual description. Since the F1 score might be undefined you should have shown some understanding about this as you decided to use it as the metric you wanted to maximize. For task 2.C I don’t think load\_iris, SelectKBest and chi2 could be used, only the two models. Do you even use your greedy search algorithm?

The results you arrive to and the way you visualize them is good and deserves some acknowledgement.

Grades:

Textual description: 4

Code quality: 4.5

Results: 5.5