1 Auto evaluation

This project started as a much bigger project than the resulting one, with exciting ideas that we later found out were not achievable in the given time frame.

At the start of the project, we had trouble regarding the group itself, but after talking with our supervisors, the problems were resolved and we finally learned how to work efficiently together. However, this discrepancy has cost us some time, since we were missing a group member.

Furthermore, we spent a lot of time writing the report itself in the first 13 weeks. This was about the parts Introduction, State of the Art, Motivation, Data description and Research questions. When thinking back on this, we had spent too much trying to write a ”perfect” report from the beginning, instead of working on the data and sorting everything out. It was the first time any of us had worked with real-world data. So we were not prepared for how much time it would take to sort the data. Furthermore, we have also reformulated the research questions, since we needed to adjust our goals to what time we had left.

Next time, we should consider working with the data way earlier (and not starting with it until the 3-week period). Moreover, we should have used our supervisors more, especially to get help with general ideas and also the code itself (both data cleaning and how the models could be implemented). Similarly, we should have used the doctors' supervision much more, for instance, to talk about missing values, and which values were most important so that we could optimize the work further in regards to the time we had and our knowledge of certain medical terms.

As mentioned we were also forced to change our research questions during the 3-week period. This was due to the fact that the EUROSCORE II model could not be fitted, without significant sacrifice, to our dataset because of missing variables. This meant that we had to drop the heavy connection to EUROSCORE II and instead focus on just building our own logistic regression model. This caused us to take a bit of a different and more exploratory approach to the dataset with a lack of a valid baseline to outperform. We also changed our second research question to something more feasible after a talk with our counselor. The first iteration was focused on finding a correlation between the features which is almost a project in and of itself.

Because of the dataset belonging to Rigshospitalet, we were also forced to use one work laptop, making it difficult to all work on the code at once. We ended up working on the code in pairs but this was not entirely optimal and we should try to find a better approach for solving this, in the future.

2 AI tool evaluation

Large Language models like Chat GPT and co-pilot have been utilized throughout this project, each used with different purposes. Since this project contained a lot of coding, numerous AI models were to be built. For instance, a Co-pilot would help implement repetitive code lines and or repetitive definitions of similar variables but with different names. Moreover, Chat GPT would be used in cases where errors couldn't be fixed by ourselves. Therefore, sometimes, Chat GPT was used as a troubleshooting tool, but despite asking for chat, a solution could not always be found.

Also, for desperate cases, we would sometimes use Chat GPT to troubleshoot problems and errors regarding data cleaning. However, when doing this, we had to be careful in how the prompts were generated. This was due to the signed contract of not sharing the data in any shape or form to other places that are not locally on the computer we were given. Therefore, to ensure not to break any rules, we had to carefully generate the prompts, which was done by explaining the problems at hand and making up ”fake” data scenarios (instead of just copy-pasting), and then chat would provide ideas for solutions, which we then would try to adapt to the ”real” data that was worked with.

But in general, Chat GPT used for coding purposes would be used as a last resort, as we have experienced that it took a lot of prompts and messages back and forth before finding a solution, and sometimes, we would not even find a solution, and we would start over with what we were doing from a new perspective. We saw that Chat GPT worked best for coding if it was a small snippet that we needed. For instance, we needed a simple loop that could open all CSV files in a folder while assigning each opened file to a certain name.

Furthermore, Chat GPT would sometimes also be used as Google, meaning that instead of searching something up or searching in books, we would ask Chat for a simple and quick explanation. We would also use Chat GPT to generate ideas/ variations on how to reformulate sentences or certain words to achieve a certain professional language in the report.

Chat has also been used sometimes to check and evaluate the ideas we had, which meant that if we had some ideas in regards to what statistical test to run or what relevant theory to include, we would explain those ideas to chat and ask it to rate and evaluate them. Hence, Chat GPT has also been used to push us in the, presumably, correct direction or just to confirm, that what we have done for now is valid.

There have also been instances where we asked Chat to write code for models and such, but often, we would find that it did not work, hence, we would quickly move away from Chat GPT and find other ways such as searching on the internet or finding code in material from previous courses.

To summarize, AI tools, mainly Chat GPT, have been used as a quick way of finding information and explanations for terms or concepts, or sometimes for code troubleshooting, although this was often a tedious process with back-and-forth messages while not nearing a solution and evaluation of some ideas we have for the project.