**Project Summary**

In summary, our project was quite interesting, yet pretty challenging at the same time. Even though we expected at first to achieve better results at the end, we are happy with what we tried and how we explored different models and ways to approach this difficult problem. The main achievements we got were the models that showed quite interesting results that show how good models in some metrics can only look good because we don’t see the other perspectives; what I mean by that is for example, we had models that had accuracy of 99% but the f-measure was very low. This opened our eyes on the challenge of this problem, and we started applying advanced techniques to try and tackle it such as meta learning and MLSMOTE. We did some improvements but still the project needed more work. We believe more oversampling could have had even better results, maybe more careful pre-processing or merging classes also since we had lots of classes (200+), so it was quite hard to achieve good results without effort, even with effort we found it challenging and we believe this project should be given its time to study carefully because it’s quite unique and hard to come to a convergent solution. The main limitations for us were time and, to be honest, the way the course made us to try different models even though they could not give us what we need in some cases, but we did try irrelevant things; however, it was quite useful too because it opened our eyes on the challenge of this problem and what difficulties we can face so it wasn’t for nothing.

We think this project needs more preprocessing before trying advanced techniques such as oversampling and maybe merging classes after careful analysis with someone in the medical field because we need to make sense if it’s a real world project. Maybe some of the limitations, such as the small amount of data given that we have a 200+ classes multilabel problem are indeed beyond a data scientist’s control, but one must try their best to do what can be done with what they’re given.

Another difficulty we faced is the rather long training time for most models, caused mainly by the large number of output classes (over 200) and features (over 800 before feature selection, over 300 after) for this multi-label problem. Many models took days to finish training, especially with grid search and cross-validation, and this made us struggle to finish some tasks by the deadline.

Concerning the research part, we implemented what we picked as topics, and we had slight improvement in some cases as we mentioned in the notebooks, but not by a lot. For example, meta learning did not give us better results, not a considerable difference between it and the models before meta learning. We spent actually more time on the research part and applying to other datasets to submit our research phases, and we believe we should have spent more time exploring how we can combine the advanced techniques to get better results. But again, we think the data is not enough for 200+ classes and more work needed to be done for sure, and it’s definitely an advanced project.

The way we approached the research part was trying out each technique by itself, and none of them gave us results that we are satisfied of. What we were looking for was good f-scores, but we only had improvements in accuracy, and it’s a problem because we are interested in the true positives for this type of project, especially because it’s medical data. One technique for example, MLSMOTE, gave us slightly better results in SVM models compared to before and it was a big deal because we were struggling with f-measures. When we realized we needed more processing using such technique, it was too late because we implemented it at the end of the semester, and we believe better results can be achieved using this technique combined with others we tried earlier, or we did not try yet such as advanced oversampling and multi-label research areas that can be explored for this type of project.