Given our dataset, and after doing exploratory data analysis, we discovered that our dataset represents an imbalanced multi-label classification problem. This problem is hard to be contained or accurately predicted by any model without doing a very specific data preprocessing. For that reason, Dr. Fatima proposed this paper for us to tackle our problem: “MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation”. MLSMOTE, a data preprocessing and over sampling algorithm, stands for Multilabel Synthetic Minority Over-sampling Technique which is exactly what we need in our project to tackle the problem of imbalanced labels and minority cases. The way that MLSMOTE works, is that it generates and adds new synthetic instances that belongs to minority classes to the original dataset, in way that it converts an imbalanced labeled problem to a somehow balanced one (based on the number of the new samples generated as user input). Then applying any classification model on the new dataset will results in a better performance and higher accuracy.

The original code of MLSMOTE is written in Python is available online through: <https://gist.github.com/niteshsukhwani/6be61cf42d14eb78c52397f9ceacb569>

Applying MLSMOTE as it is didn’t work. In the paper, the authors stated that it also works for categorical data, however, when applying the code on our data (which is a mix of categorical and numerical data) it didn’t work because of the constraints of needing numerical input. As a result, we added the code of doing hot encoding for the categorical features and generate a new dataset which will be passed then to MLSMOTE. The resulted dataset (after applying MLSMOTE) will therefore be passed to a classification model and the results is slightly better. And as we can noticed, the higher the number of samples generated the better accuracy we get.