RQ 2.1: Summarize the results of your experiments for Problem 2. Are your results “good” or “bad”, and importantly, how do you determine that?

it depends on the baseline, we should compare metrics to some baseline methonds. And we have a imbalanced dataset so Accuracy might not be a good metric, instead F1 will be a better one because it is the harmonic mean of precision and recall

RQ 2.2: In Problem 4, you implemented undersampling. How did your results change compared to Problem 2? How would “oversampling” have looked like for this task?

For imbalanced data we have a better f1-score for the most prevailed classes such as S,M and poor value for those less amounted classes to get a better overall accuracy but a lower macro or weighted avg f1-score. While in balanced data this problem are relatively fixed, especially the macro avg is much more better.

RQ 2.3: Why is it important to do a hyperparameter search before drawing conclusions about the performance of a model? Why do you think it is often not done, anyway? Why should you never tune hyperparameters on the test set?

- different hyperparameter will yield different performance on a same (type of)model, so we need to optimize hyperparameter to get a most desiring performance.
- If its oftenly not done, it might because it would be too time consuming if we go through all hyperparameters over a wide range and fine grid. Some imperical knowledge from similar data sctructure investigated by researchers before will help good enough.
- if we use test set as valid set to tune the hyperparameters, then it will formally be the valid set and we need to find some brand new unseen data as our new testing set.