1 Preliminary data analysis

TO DO LIST:

- Basic summary stats? Size and shape of the dataset
- Class balance
- Sparsity or something like that?
- For plots, we could try biplot, check if I understood Milan's explanation?
- The nicest we can do is reduce the dimensionality to 2 factors and plot by hue class. Or just pick 2 first factors from pca.
- More importantly we need some kind of scree plot. To justify working on the PCA.

2 Training baseline models and hypertuning

- Most tuning is done, VSM give very good results.
-
- RESULTS FROM GRIDSEARCH ARE NOT GOING TO BE REPLICABLE, RESULTS FROM FINAL ESTIMATOR BASED ON IT SHOULD BE.

3 Our 3 models

- AdaBoost, just because we have seen it in class and it could be interesting.
- I wanted some classically strong classifier like random forest or GBDT?
- Play around with the PCA optimality. Or some other modern dimensionality reduction tool like t-SNE or UMAP.

4