Training report

How to deal with class imbalance?

In some situation there exist class imbalance in our dataset. In this situation, our classifier will tend to ignore the minority classes

Solution #1

- We train on what we have and then correct the metric
- **Problem:**
  - Biased classification, where the model favors the majority classes or achieves better performance on majority classes

C0	6940
C1	25552
C2	15650

When we train our classifier without class balancing and we can see how performance on C0 class is completely ignored.

|	C0	C1	C2
C0	0	0	0
C1	1548	5777	1892
C2	0	0	1566

Weighted scores:

- After our model has made its decision, we calculate the performance of the model where we weight the class specific metrics with the number of examples in each class:

$$P = 0.920$$
$$R = 0.681$$
$$F = 0.783$$

Solution #2

- We train on the stratified samples, where we organize groups based on the class distribution which we control.

Stratification

- We split the original population in the strata and then we are sampling our sample from the strata. We want to avoid minority issues and we want to create a strata where units have equal distribution.

- For each strata I trained a model and added an averaging of their results → **ANSAMBLE**

  -
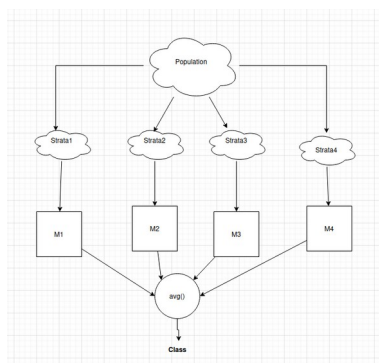
```
stratas = stratified_sampling(X=X_train, y=y_train, class_distrib=[0.333, 0.3333, 0.333], strata_size_proba=0.3, strata_num=10)
models = []
for xs, ys in stratas:
    model = GeoModel()
    m = train(model, xs, ys, epochs=50)
    models.append(m)
```

With ansamble:

- F score = 0.99