PCA for our data

During week 7 we used all the parameters that we have calculated to get the two principal components - ‘Number of branches', 'PD - sum of length of all branches', 'avPD - average sum of length of all branches', 'Height 1', 'Height 2', 'Height 3', 'Depth 1', 'Depth 2', 'Depth 3'.

The results didn’t allow us to separate the clonal trees in the proposed four categories. This week we tried to get the categories by using the length of the branches of clonar trees as we are expecting to be able to use their distribution to separate them.

First, we go to the histogram for the length of branches of each tree.
Then, we used PCA. For this section we got the PCA results shown in Figure 1.

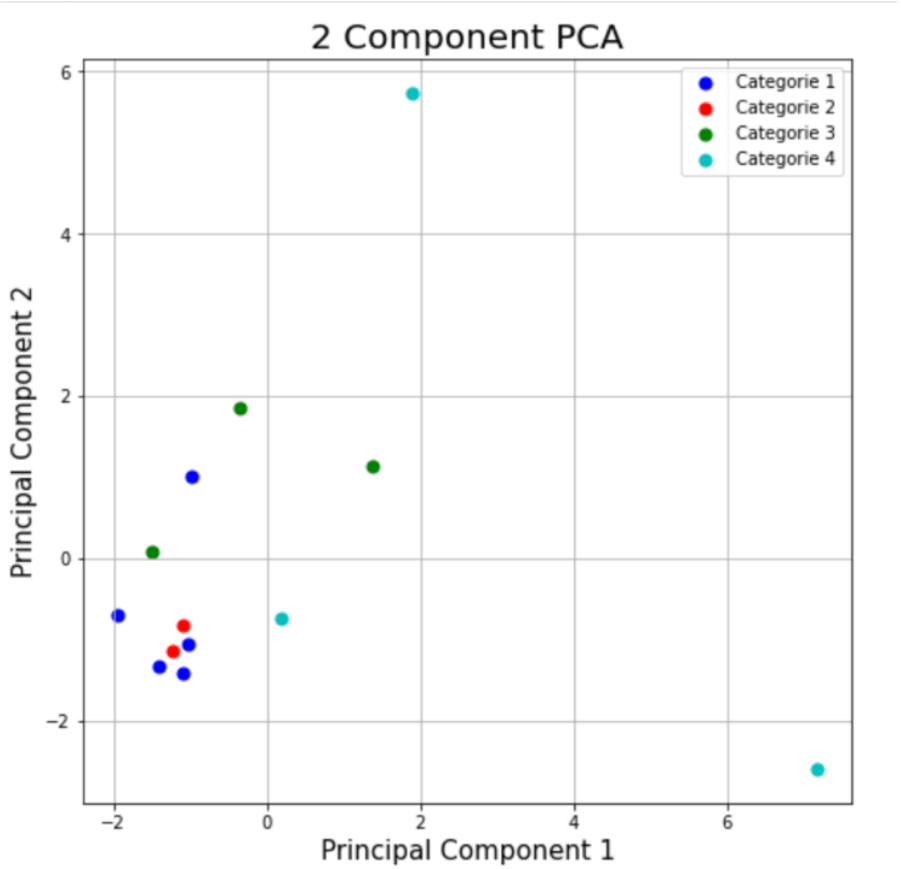

Figure 1. PCA

Then we used the results to train a KNN but as we can see in Figure 1, classes are not grouped so the performance of the KNN is not good.

Then, we tried to improve the results by normalizing data of branches by dividing the number of branches of each length into the total of branches for each clonal tree. However, by using PCA we still can't separate the categories. Figure 2.

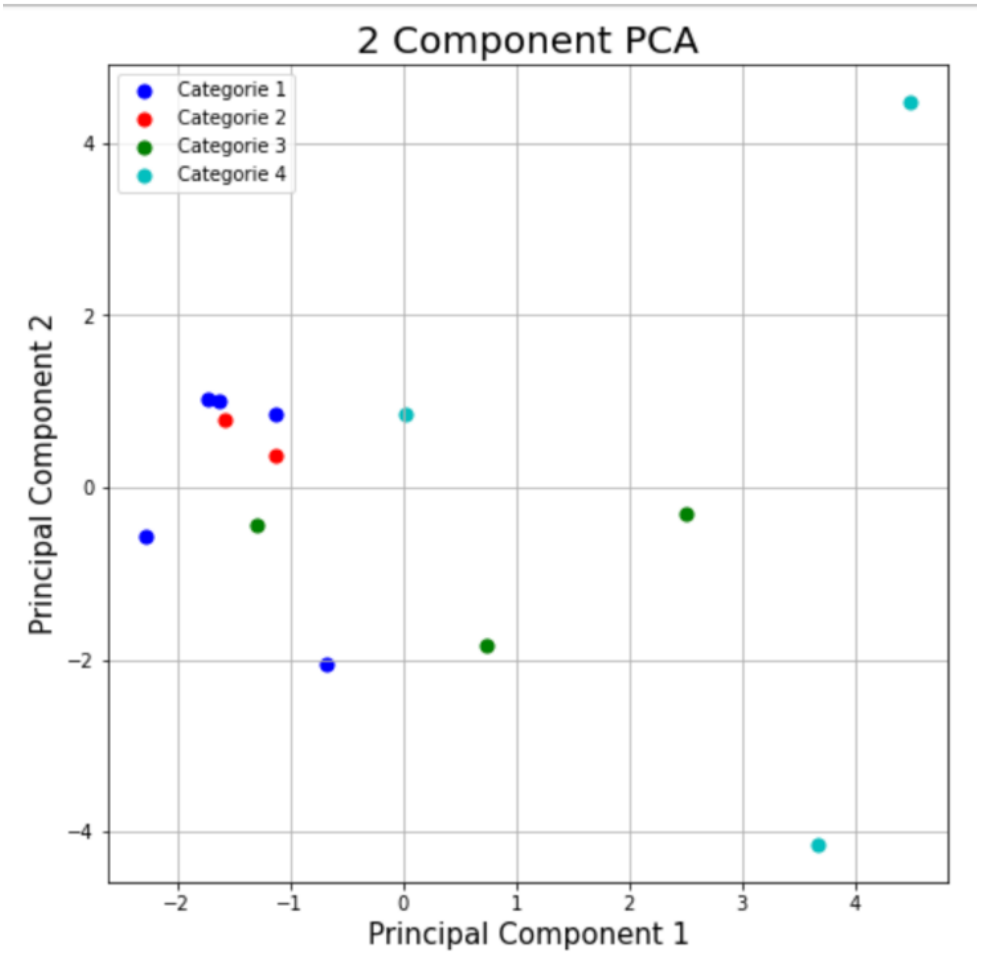

Figure 2.

Using KNN, we didn't get good results. Table 1 and Table 2.

	PC1	PC2	Categorie	KNN_zero	KNN_moy
0	-1.629169	1.000913	1	1.0	2.0
1	-1.131741	0.855318	1	1.0	2.0
2	-2.273970	-0.560297	1	1.0	1.0
3	-0.673647	-2.061912	1	4.0	3.0
4	-1.731651	1.019572	1	1.0	2.0
5	-1.127214	0.368569	2	1.0	2.0
6	-1.578294	0.775690	2	1.0	2.0
7	0.734671	-1.837121	3	4.0	3.0
8	2.513233	-0.308316	3	2.0	4.0
9	-1.291773	-0.431429	3	1.0	1.0
10	4.490959	4.481618	4	2.0	4.0
11	0.025253	0.846221	4	3.0	2.0
12	3.673343	-4.148826	4	4.0	3.0

Table 1

	True P	True N	False P	False N	Sensitivity	Specificity
1	1.0	7.0	1.0	4.0	0.200000	0.875000
2	2.0	7.0	4.0	0.0	1.000000	0.636364
3	1.0	8.0	2.0	2.0	0.333333	0.800000
4	1.0	9.0	1.0	2.0	0.333333	0.900000

Table 2

We can see that we got better results by using the metrics that we calculated during week 7 than just using the length of branches.