Hierarchical Clustering - Sponges

Background:

My data is on sea sponges. They are squishy, underwater animals that we use in the shower to scrub our bodies. I obtained my data from the UCI Machine Learning Repository [1]. I chose sponges because I found it fascinating that there are at least 75 different ways to say sponge in spanish (yes, for some reason, the only dataset on sponges (for clustering) on the UCI Machine Learning Repository was in spanish). I had decided I wanted to find out how a clustering model would group up the different sponges.

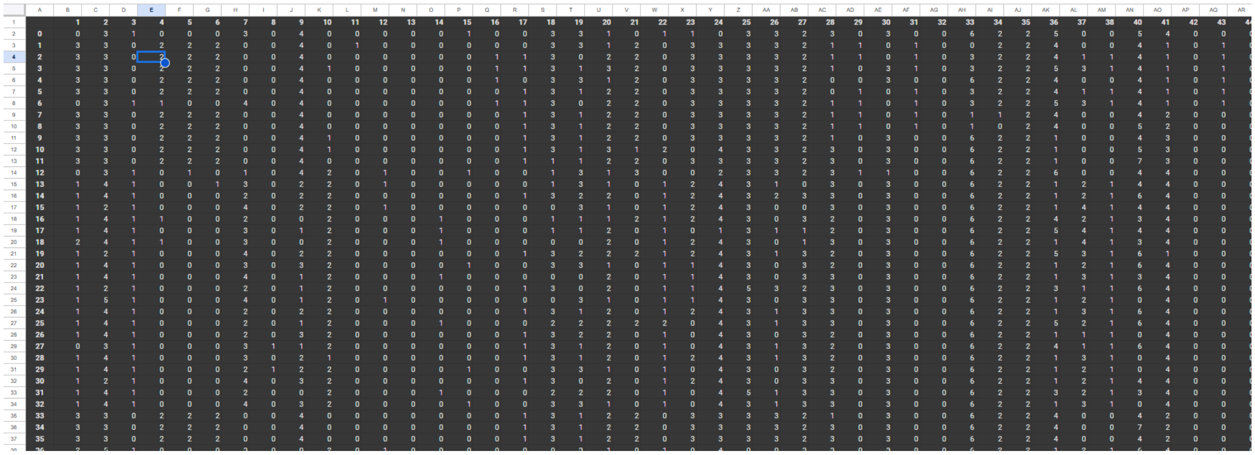
Methods:

First, I preprocessed my data. However, in order to do so, I took a look at the data using the pandas Dataframe.Style variable. The following picture shows a screenshot of the data after uploading it to Google Sheets:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
0	AAPTOS_AAPTOS	1_CAPA	SIN_CAPA_INTERNA_DEL_CORTEX	SI	NO	NO	NO	3	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	SI	NO	NO	SIN_ESPICULA_PRINCIPAL_ESTILO	SIN_ESF
1	ALECTONA_MILLARI	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	SI	NO	NO	NO	NO	NO	NO	SIN_ESPICULA_PRINCIPAL_ESTILO	SIN_ESF
2	CLIONA_CARTERI	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	SI	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
3	CLIONA_CELATA	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	2_TIPOS	NO	NO	NO	NO	NO	SI	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
4	CLIONA_LABYRINTHICA	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	SI	NO	SIN_ESPICULA_PRINCIPAL_ESTILO	SIN_ESF
5	CLIONA_SCHMIDTI	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
6	CLIONA_VIRIDIS	1_CAPA	SIN_CAPA_INTERNA_DEL_CORTEX	SI	SI	NO	NO	4	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	SI	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
7	DIPLASTRELLA_BISTELLATA	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
8	DIPLASTRELLA_ORNATA	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
9	LAXOSUBERITES_ECTYONIMUS	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	2_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
10	LAXOSUBERITES_FERRERHERNANDEZI	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	2_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
11	LAXOSUBERITES_RUGOSUS	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	NO	SI	NORMAL	
12	OXYCORDYLA_PELLITA	1_CAPA	SIN_CAPA_INTERNA_DEL_CORTEX	SI	NO	SI	NO	1	NO	SIN_TILOSTILOS_ADICIONALES	3_TIPOS	NO	SI	NO	NO	SI	NO	NO	NORMAL	SIN_ESF
13	POLYMASTIA_AGGLUTINARIS	2_CAPAS	TANGENCIAL	SI	NO	NO	SI	3	NO	INTERMEDIARIOS	3_TIPOS	NO	SI	NO	NO	NO	NO	NO	NORMAL	SIN_ESF
14	POLYMASTIA_CONIGERA	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	2	NO	INTERMEDIARIOS	3_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
15	POLYMASTIA_CORTICATA	2_CAPAS	PERPENDICULAR	SI	NO	NO	NO	4	NO	INTERMEDIARIOS	3_TIPOS	NO	SI	NO	NO	NO	NO	NO	FUSIFORME	SIN_ESF
16	POLYMASTIA_ECTOFIBROSA	2_CAPAS	TANGENCIAL	SI	SI	NO	NO	2	NO	ECTOSOMICOS_DISPERSOS	3_TIPOS	NO	NO	NO	SI	NO	NO	NO	NORMAL	
17	POLYMASTIA_FUSCA	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	3	NO	ECTOSOMICOS_EN_RAMILLETES	3_TIPOS	NO	NO	NO	SI	NO	NO	NO	NORMAL	
18	POLYMASTIA_GRIMALDI	3_CAPAS	TANGENCIAL	SI	SI	NO	NO	3	NO	ECTOSOMICOS_DISPERSOS	3_TIPOS	NO	NO	NO	SI	NO	NO	NO	FUSIFORME	
19	POLYMASTIA_HIRSUTA	2_CAPAS	PERPENDICULAR	SI	NO	NO	NO	4	NO	INTERMEDIARIOS	3_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
20	POLYMASTIA_INFLATA	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	3	NO	INTERMEDIARIOS_Y_ECTOSOMICOS	3_TIPOS	NO	NO	NO	NO	SI	NO	NO	SIN_ESPICULA_PRINCIPAL_ESTILO	SIN_ESF
21	POLYMASTIA_INFRAPILOSA	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	4	NO	ECTOSOMICOS_EN_RAMILLETES	3_TIPOS	NO	NO	NO	SI	NO	NO	NO	FUSIFORME	
22	POLYMASTIA_INVAGINATA	2_CAPAS	PERPENDICULAR	SI	NO	NO	NO	2	NO	ECTOSOMICOS_EN_RAMILLETES	3_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
23	POLYMASTIA_LITTORALIS	2_CAPAS	TANGENCIAL_Y_PERPENDICULAR	SI	NO	NO	NO	4	NO	ECTOSOMICOS_EN_RAMILLETES	3_TIPOS	NO	SI	NO	NO	NO	NO	NO	FUSIFORME	SIN_ESF
24	POLYMASTIA_MAMMILLARIS	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	2	NO	INTERMEDIARIOS	3_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
25	POLYMASTIA_MARTAE	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	2	NO	ECTOSOMICOS_EN_RAMILLETES	3_TIPOS	NO	NO	NO	SI	NO	NO	NO	POLITILOTA	
26	POLYMASTIA_POLYTYLOTA	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	2	NO	INTERMEDIARIOS_Y_ECTOSOMICOS	3_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
27	POLYMASTIA_RADIOSA	1_CAPA	SIN_CAPA_INTERNA_DEL_CORTEX	SI	NO	NO	NO	3	SI	ECTOSOMICOS_EN_RAMILLETES	3_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
28	POLYMASTIA_ROBUSTA	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	3	NO	INTERMEDIARIOS	2_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
29	POLYMASTIA_SPINULA	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	2	SI	INTERMEDIARIOS	3_TIPOS	NO	NO	NO	NO	SI	NO	NO	SIN_ESPICULA_PRINCIPAL_ESTILO	SIN_ESF
30	POLYMASTIA_TENAX	2_CAPAS	PERPENDICULAR	SI	NO	NO	NO	4	NO	INTERMEDIARIOS_Y_ECTOSOMICOS	3_TIPOS	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
31	POLYMASTIA_TISSIERI	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	2	NO	ECTOSOMICOS_DISPERSOS	3_TIPOS	NO	NO	NO	SI	NO	NO	NO	POLITILOTA	
32	POLYMASTIA_UBERRIMA	2_CAPAS	TANGENCIAL	SI	NO	NO	NO	4	NO	INTERMEDIARIOS_Y_ECTOSOMICOS	3_TIPOS	NO	NO	NO	NO	SI	NO	NO	SIN_ESPICULA_PRINCIPAL_ESTILO	SIN_ESF
33	PROSUBERITES_EPIPHYTUM	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
34	PROSUBERITES_LONGISPINA	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
35	PROSUBERITES_RUGOSUS	SIN_CORTEX	SIN_CAPA_INTERNA_DEL_CORTEX	NO	SIN_CORTEX	SIN_CORTEX	SIN_CORTEX	0	NO	SIN_TILOSTILOS_ADICIONALES	1_TIPO	NO	NO	NO	NO	NO	NO	SI	SIN_ESPICULA_PRINCIPAL_ESTILO	
36	PROTELEIA_SOLLASI	3_CAPAS	TANGENCIAL_Y_PERPENDICULAR	SI	NO	NO	NO	3	NO	ECTOSOMICOS_DISPERSOS	3_TIPOS	NO	SI	NO	NO	NO	NO	NO	FUSIFORME	SIN_ESF

When I saw the data, I noticed that column 0's data were all unique from one another, and that column 39 had '?'s as values. I looked back at the website where I got the data from, and it mentioned that the data has missing values. So I removed columns 0 and 9 from the data.

Next, I noticed that there were values in Spanish that I could not make sense of. I also needed numerical values, so I applied label encoding to all of the categorical data.



My next thoughts were to reduce the number of attributes, since there were 45 attributes, which I thought was a lot of attributes. So I used the correlation coefficients between the attributes to determine irrelevant and redundant attributes. Irrelevant attributes were determined when all correlations with the other attributes were between 0.2 and -0.2. This would mean that the attribute did not have any correlation with the other attributes. Redundant attributes were determined when any correlation with the attribute was greater than / equal to 0.5, or less than / equal to -0.5. This would mean that the attribute was very similar to another attribute, and thus could be represented by that similar attribute. Here is my formula, it is a nested loop that compares every attribute to each other:

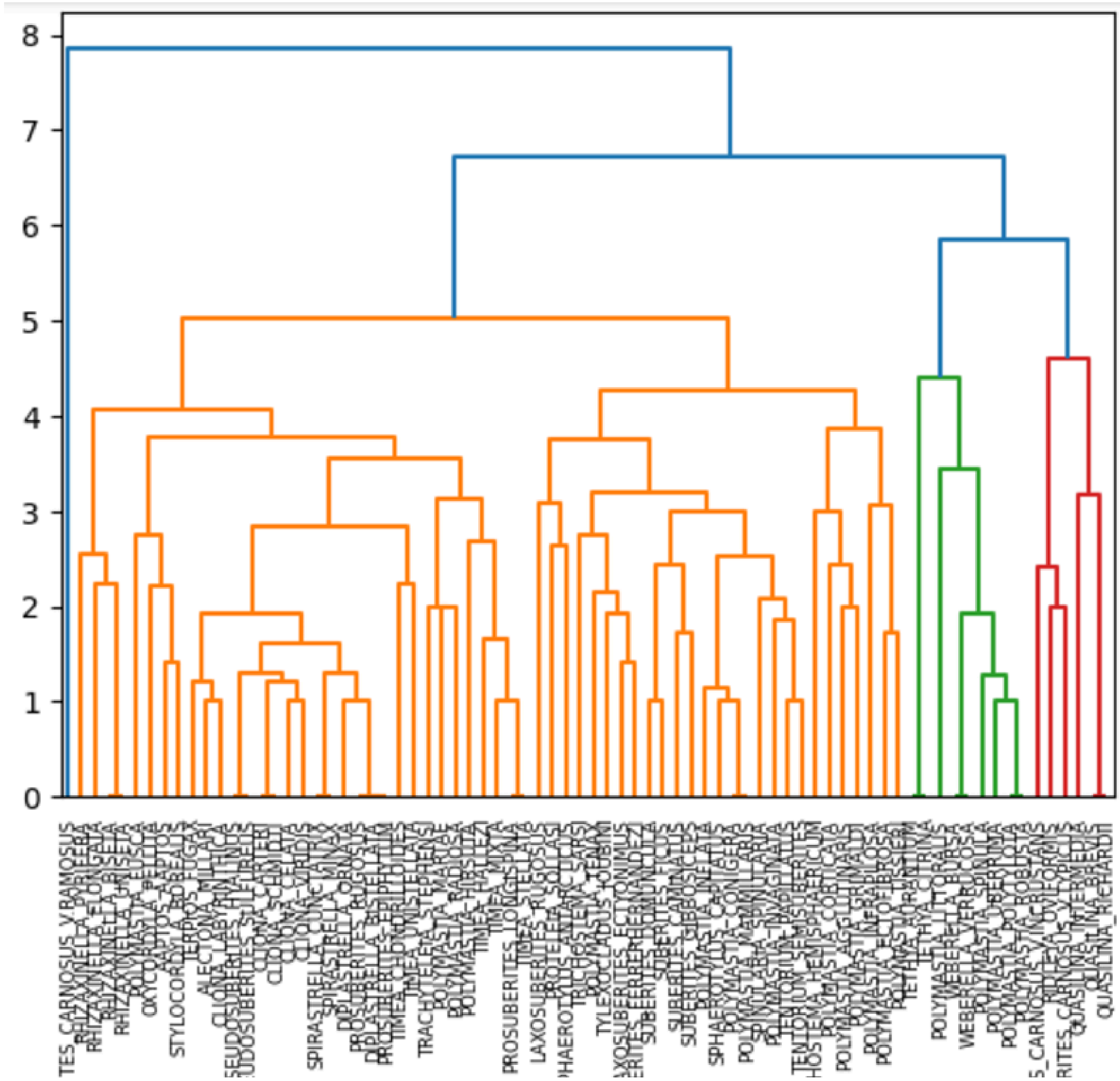
```
for i in range(len(encoded_data.columns)):
    removed = False
    # assume the data is irrelevant
    irrelevant = True
    column1 = encoded_data.iloc[:,i]
    for j in range(i + 1, len(encoded_data.columns)):
        column2 = encoded_data.iloc[:,j]
        correlation = np.corrcoef(column1, column2)
        # if there's a tiny bit of correlation with one of the attributes
        if (correlation[0, 1] >= 0.2 or correlation[0, 1] <= -0.2):
            irrelevant = False
        # reduce redundant features
        if correlation[0, 1] >= 0.5 or correlation[0, 1] <= -0.5:
            reduced_data.drop(encoded_data.columns[i], axis=1, inplace=True)
            removed = True
            break
    # reduce irrelevant features
    if irrelevant:
        reduced_data.drop(encoded_data.columns[i], axis=1, inplace=True)
```

My data was now reduced to 12 attributes:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		2	8	11	13	18	20	27	35	36	38	40	41
2	0	3	0	0	0	3	1	2	2	5	0	5	4
3	1	3	0	1	0	3	1	2	2	4	0	4	1
4	2	3	0	0	0	3	2	2	2	4	1	4	1
5	3	3	0	0	0	3	3	2	2	5	1	4	1
6	4	3	0	0	0	3	1	2	2	4	0	4	1
7	5	3	0	0	0	3	2	2	2	4	1	4	1
8	6	3	0	0	0	3	2	2	2	5	1	4	1
9	7	3	0	0	0	3	2	2	2	4	0	4	2
10	8	3	0	0	0	3	2	2	2	4	0	5	2
11	9	3	0	0	0	3	2	2	2	1	0	4	3
12	10	3	0	0	0	3	3	2	2	1	0	5	3
13	11	3	0	0	0	1	2	2	2	1	0	7	3
14	12	3	0	0	0	1	1	2	2	6	0	4	4
15	13	4	0	0	0	1	1	0	2	1	1	4	4
16	14	4	0	0	0	3	2	3	2	1	1	6	4
17	15	2	0	0	0	0	1	0	2	1	1	4	4
18	16	4	0	0	0	1	1	3	2	4	1	3	4
19	17	4	0	0	0	1	2	1	2	5	1	4	4
20	18	4	0	0	0	0	2	1	2	1	1	3	4
21	19	2	0	0	0	3	2	3	2	5	1	6	1
22	20	4	0	0	0	3	1	3	2	1	1	6	4
23	21	4	0	0	0	0	2	3	2	1	1	3	4
24	22	2	0	0	0	3	2	2	2	3	1	6	4
25	23	5	0	0	0	0	1	3	2	1	1	0	4
26	24	4	0	0	0	3	2	3	2	1	1	6	4
27	25	4	0	0	0	2	2	3	2	5	1	6	4
28	26	4	0	0	0	3	2	3	2	1	1	0	4
29	27	3	1	0	0	3	2	3	2	4	1	6	4
30	28	4	0	0	0	3	2	3	2	1	1	0	4
31	29	4	1	0	0	3	1	3	2	1	1	0	4
32	30	2	0	0	0	3	2	3	2	1	1	4	4
33	31	4	0	0	0	2	2	3	2	3	1	3	4
34	32	4	0	0	0	3	1	3	2	1	1	0	4
35	33	3	0	0	0	3	2	2	2	4	0	4	2
36	34	3	0	0	0	3	2	2	2	4	0	7	2
37	35	3	0	0	0	3	2	2	2	4	0	4	2
38	36	5	0	0	0	0	1	3	2	1	1	6	4
39	37	3	0	0	0	3	2	2	2	5	0	4	1

Results:

Finally, I used hierarchical clustering to get a clear picture of which sponges were grouped together early on, and which groups became the direct “subclasses” of the sponge “class.” I used an agglomerative hierarchical clustering algorithm and I used group averages between clusters for computing the inter-cluster similarity. This is my resulting dendrogram:



Conclusions:

It only took a few seconds to train the model. Perhaps I could make the threshold for irrelevant attributes a bit smaller. Other than that, it seems that the model thinks that there are 3 main subcategories under what is defined as a sponge, perhaps 4 if the threshold cutoff in the dendrogram was slightly under 5.

References:

[1] https://archive.ics.uci.edu/dataset/97/sponge