

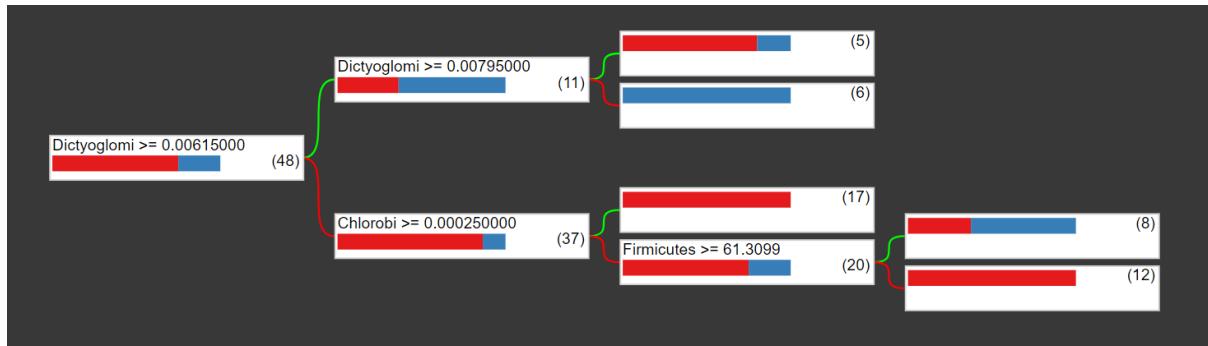
La reproducció no és només cosa de dos: hackejant els secrets del microbioma seminal

Para el análisis del microbioma hemos decidido darle tres enfoques distintos, creando árboles de decisión, heatmaps y pairplots.

Árboles de decisión

El código utilizado para entrenar a los Árboles de decisión están en el fichero “Decission_Tree_Maraton”. Se ha utilizado un split de 0.15 para separar el dataset de training del de validación y se ha hecho uso de la librería tensorflow para el tratamiento de datos y el entrenamiento.

Empezaremos por el árbol que analiza el dataset de pylum:



```
[11] model_1.compile(metrics=["accuracy"])
    evaluation = model_1.evaluate(test_ds, return_dict=True)

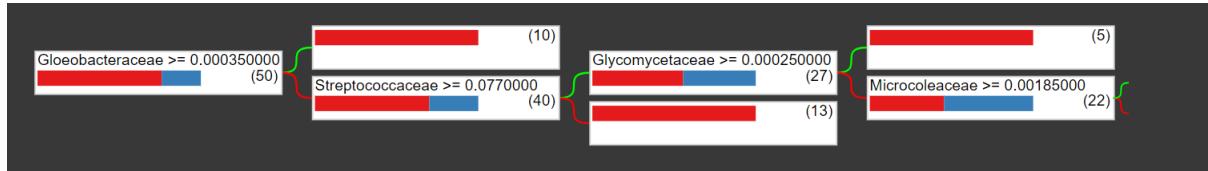
    for name, value in evaluation.items():
        print(f"{name}: {value:.4f}")

1/1 [=====] - 1s 1s/step - loss: 0.0000e+00 - accuracy: 1.0000
loss: 0.0000
accuracy: 1.0000
```

Observamos principalmente dos datos curiosos. Por un lado, cuando la bacteria Dictyoglomi tiene una concentración mayor o igual a 0.006, el 75% de los pacientes son estériles.

Y por otro lado, cuando además la concentración de la bacteria Chiorobi des mayor o igual a 0.00025, el porcentaje de pacientes estériles sube hasta el 85%.

Con el dataset de family conseguimos el siguiente árbol:



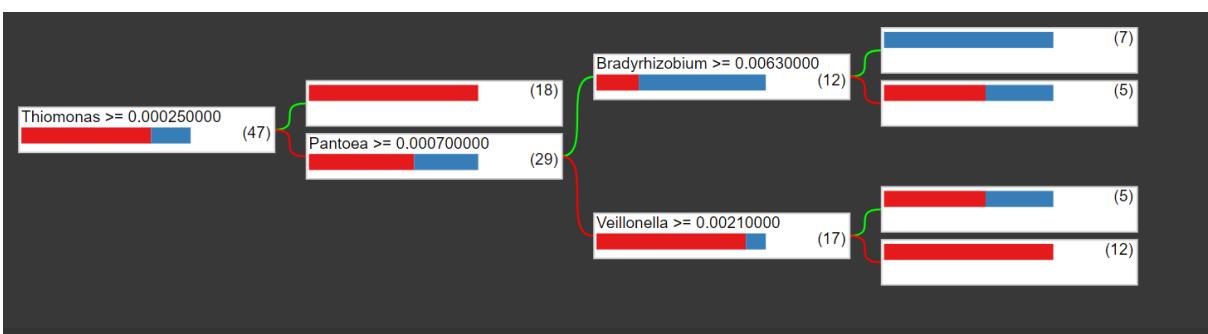
```
[13] model_1.compile(metrics=["accuracy"])
      evaluation = model_1.evaluate(test_ds, return_dict=True)

      for name, value in evaluation.items():
          print(f"{name}: {value:.4f}")

1/1 [=====] - 1s 1s/step - loss: 0.0000e+00 - accuracy: 0.8333
loss: 0.0000
accuracy: 0.8333
```

En el que si la bacteria *Gloeobacteraceae* es igual o superior a 0.00035, el 74% de los pacientes son estériles.

Por último tenemos el modelo de género:



```
[11] model_1.compile(metrics=["accuracy"])
      evaluation = model_1.evaluate(test_ds, return_dict=True)

      for name, value in evaluation.items():
          print(f"{name}: {value:.4f}")

1/1 [=====] - 3s 3s/step - loss: 0.0000e+00 - accuracy: 0.8889
loss: 0.0000
accuracy: 0.8889
```

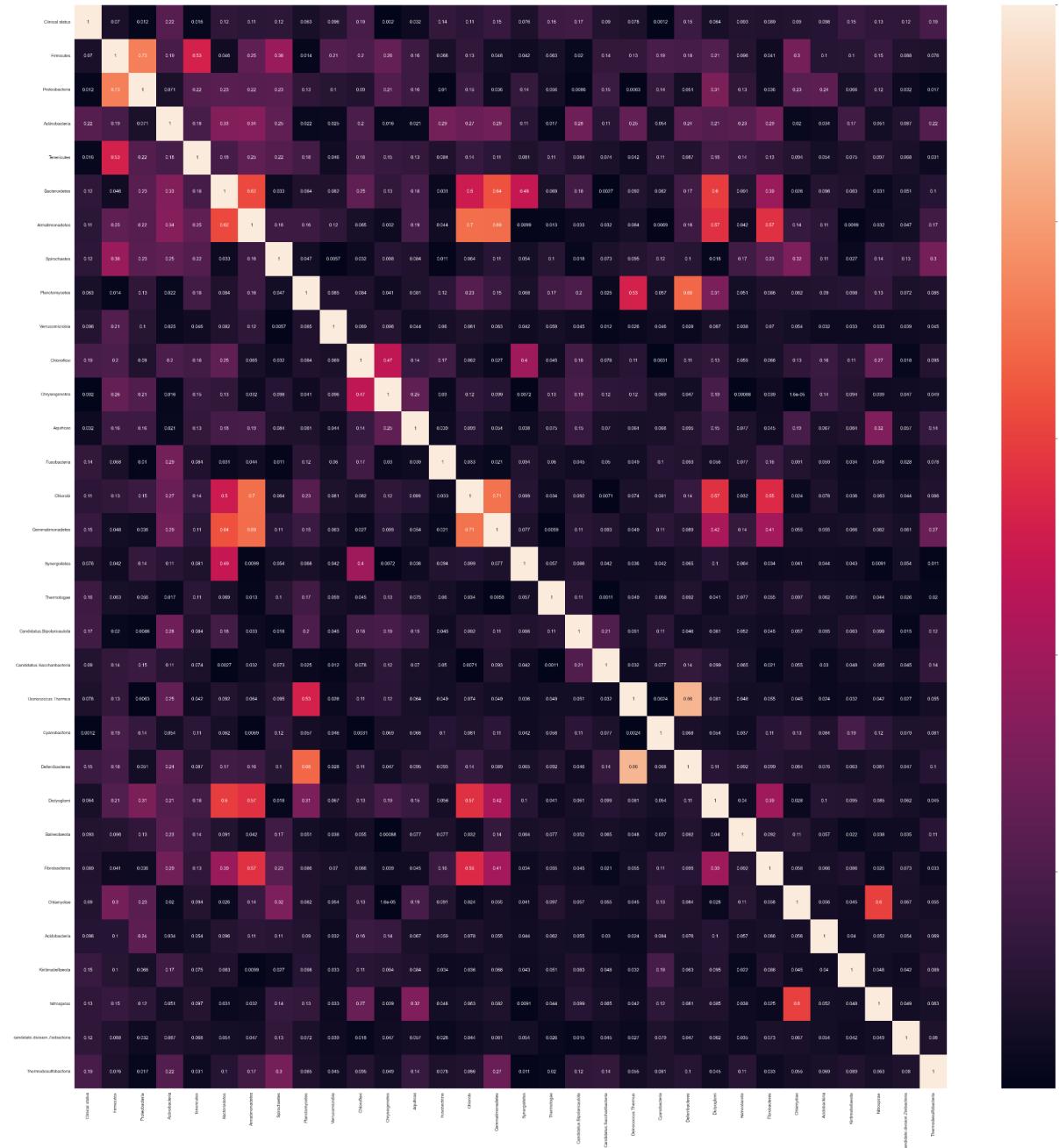
En este caso, cuando la bacteria *Thiomonas* es igual o superior a 0.00025, el 76% de los hombres son estériles y en el dataset de training si se le suma que la *Pantoea* sea menor a 0.0007, todos los casos son de esterilidad.

Bajo cada árbol de decisión se ha añadido una captura enseñando el accuracy con el dataset de test, para ver la fiabilidad del árbol. También cabe destacar que la cantidad de datos es muy baja y que la fiabilidad se podría mejorar aumentándolos.

Heatmaps

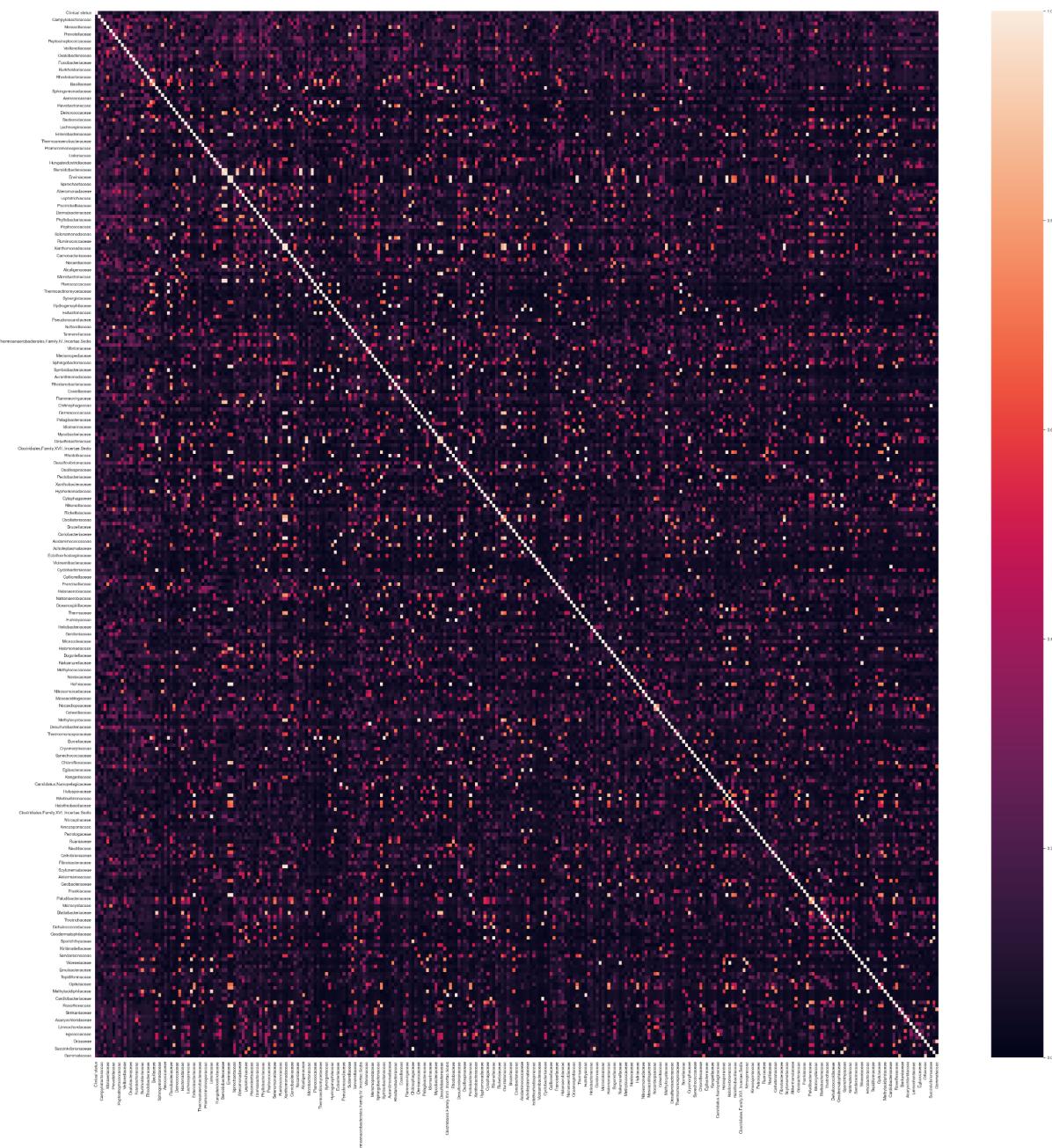
A continuación tenemos algunos heatmaps que hemos hecho con las correlaciones entre microbiotas y el estado clínico de las personas. Tanto en Genus como en Family tenemos tantos datos que la gráfica deja de ser útil para encontrar correlaciones concretas, pero podemos observar bastante correlación entre algunos microbiotas. Pese a que muchas de las correlaciones en los apartados de Genus y Family se deben a que los microbiotas en cuestión son de la misma familia, también podemos observar algunas correlaciones que no son tan directas, sobretodo en el apartado Pylum, aunque habría que hacer un estudio más extenso para poder confirmar que estas correlaciones son relevantes y no son casualidad por la poca cantidad de datos utilizados en el estudio. Debajo de cada heatmap hemos anotado los pares de microbiotas más correlacionados entre ellos.

Pylum



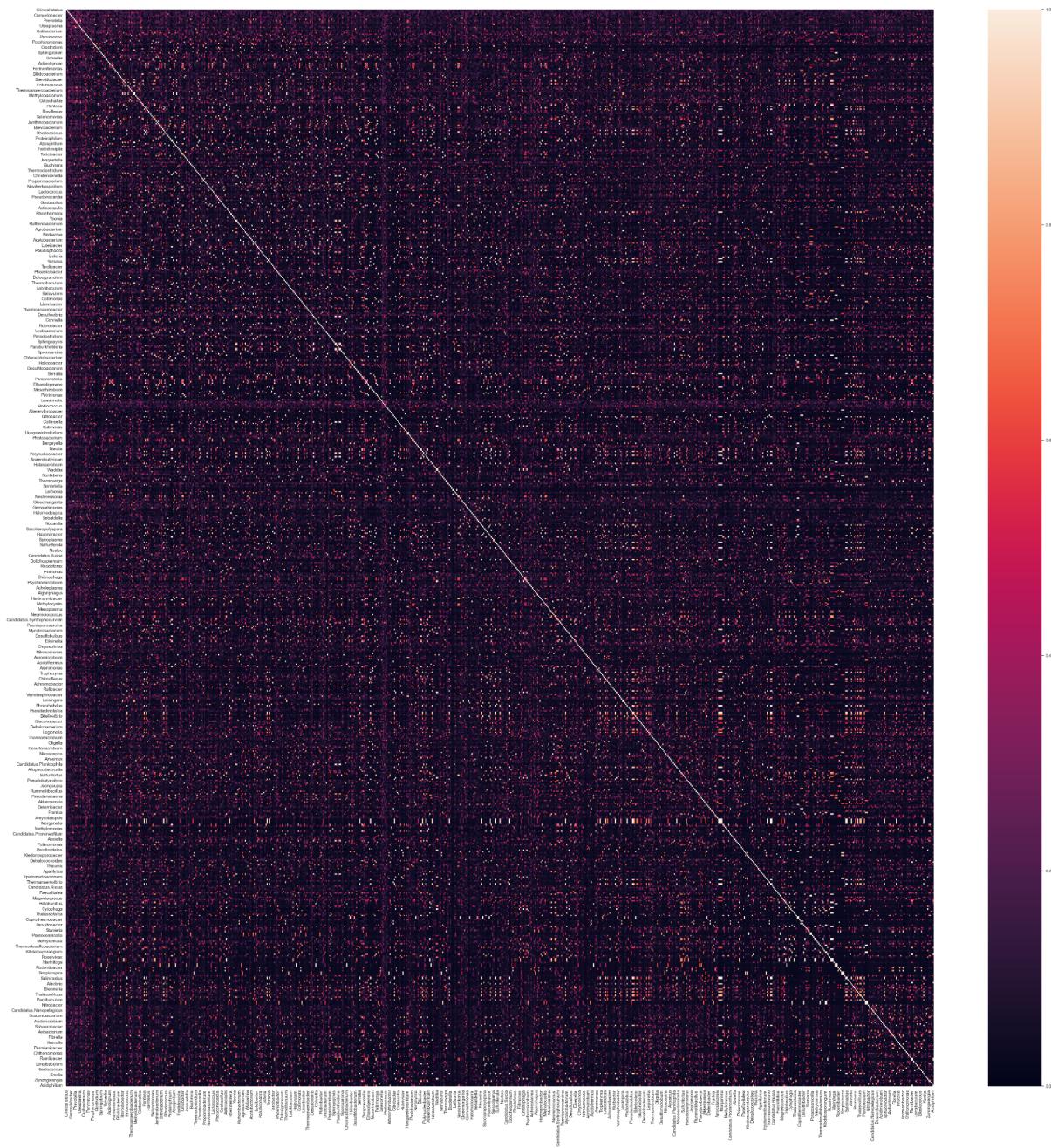
[('Firmicutes', 'Proteobacteria'), ('Bacteroidetes', 'Armatimonadetes'), ('Bacteroidetes', 'Gemmamatimonadetes'), ('Armatimonadetes', 'Chlorobi'), ('Armatimonadetes', 'Gemmamatimonadetes'), ('Planctomycetes', 'Deferrribacteres'), ('Chlorobi', 'Gemmamatimonadetes'), ('Deinococcus.Thermus', 'Deferrribacteres')]

Family



[('Dysgonamonadaceae', 'Leuconostocaceae'), ('Dysgonamonadaceae', 'Streptosporangiaceae'), ('Clostridiales.Family.XVII..Incertae.Sedis', 'Acidobacteriaceae'), ('Clostridiales.Family.XVII..Incertae.Sedis', 'Rhodocyclaceae'), ('Acidobacteriaceae', 'Rhodocyclaceae'), ('Psychromonadaceae', 'Cardiobacteriaceae'), ('Moritellaceae', 'Budviciaceae')]

Genus



[('Rahnella', 'Plesiomonas'), ('Rahnella', 'Saliniradius'), ('Rahnella', 'Aliivibrio'), ('Candidatus. Annandia', 'Plesiomonas'), ('Candidatus. Annandia', 'Saliniradius'), ('Candidatus. Annandia', 'Aliivibrio'), ('Candidatus. Ishikawaella', 'Plesiomonas'), ('Candidatus. Ishikawaella', 'Saliniradius'), ('Candidatus. Ishikawaella', 'Aliivibrio'), ('Morganella', 'Plesiomonas'), ('Morganella', 'Saliniradius'), ('Morganella', 'Aliivibrio')]

Durante la elaboración de los heatmaps, para poder tratar los metadatos de “City of residence 1”, “Clinical status”, “Alcohol intake” y “Antibiotics (last month)”, los cuales son considerados interesantes para tratar durante el proyecto, sus valores se han convertido de

categóricos a numéricos para poder ser tratados para la correlación, asignando un identificador único para cada posible valor.

En el caso de “Clinical status”, se ha asignado un valor de 0 cuando el paciente es infértil, de lo contrario un 1. Con “Alcohol intake”, se ha asignado un 0 cuando el paciente “Never” nunca bebe, 1 en el caso de “Sporadic” y 2 en el caso de “Frecuent”. Finalmente, con “Antibiotics (last month)” se ha asignado un 0 en los casos de “None” y 1 en los de “Yes”.

Durante la búsqueda de correlaciones, el dataset indica que hay una correlación de un 75% entre “City of residence 1” y “Clinical status”, lo cual implicaría que el lugar de residencia impacta la fertilidad. No obstante, después de revisar el dataset, muchos de los casos de infertilidad coinciden en una misma ciudad, lo cual crea un bias respecto encontrar esta correlación. Por lo tanto, se puede considerar que no es posible llegar a ninguna conclusión fiable respecto a esta relación, y que más datos serían necesarios.

Usando el script “Cluster_finding” para encontrar clusters de microbiotas que pudiesen estar relacionados con los metadatos, hemos conseguido en el caso de genus hallar clusters de microbiotas que estaban relacionados con “Total sperm number ($\times 10^6$)” y “Radiation exposure”.

Correlaciones encontradas:

- Correlation between Rhodocytophaga, ['Starkeya'] with Total sperm number ($\times 10^6$): 0.6411
- Correlation between Microvirga, ['Rhodocytophaga', 'Starkeya'] with Total sperm number ($\times 10^6$): 0.6086
- Correlation between Starkeya, ['Rhodocytophaga'] with Total sperm number ($\times 10^6$): 0.6666
- Correlation between Luteipulveratus, ['Stella', 'Maritalea'] with Radiation exposure: 0.6696
- Correlation between Stella, ['Luteipulveratus', 'Maritalea'] with Radiation exposure: 0.6887
- Correlation between Maritalea, ['Luteipulveratus', 'Stella'] with Radiation exposure: 0.7006

Pairplot

A continuación hemos hecho un pairplot para encontrar parejas que indiquen si alguna correlación entre microbiotas están también relacionadas con el estatus clínico de la persona. Por desgracia no hemos encontrado ninguna relación relevante. Únicamente hemos podido construir el pairplot con los datos de Pylum ya que los otros dos datasets tenían demasiados microbiotas para poder construir el pairplot entero con la capacidad de computación que teníamos accesible.

