Machine Learning in Computational Biology

Georgios Leventis

7115172100024

Assignment #1

https://github.com/geoleven/Assignment-1

The exercise was done according to its description as much as possible.
At the relevant github you can find the final models, as well as the proper notebooks, the python code and metrics json file.

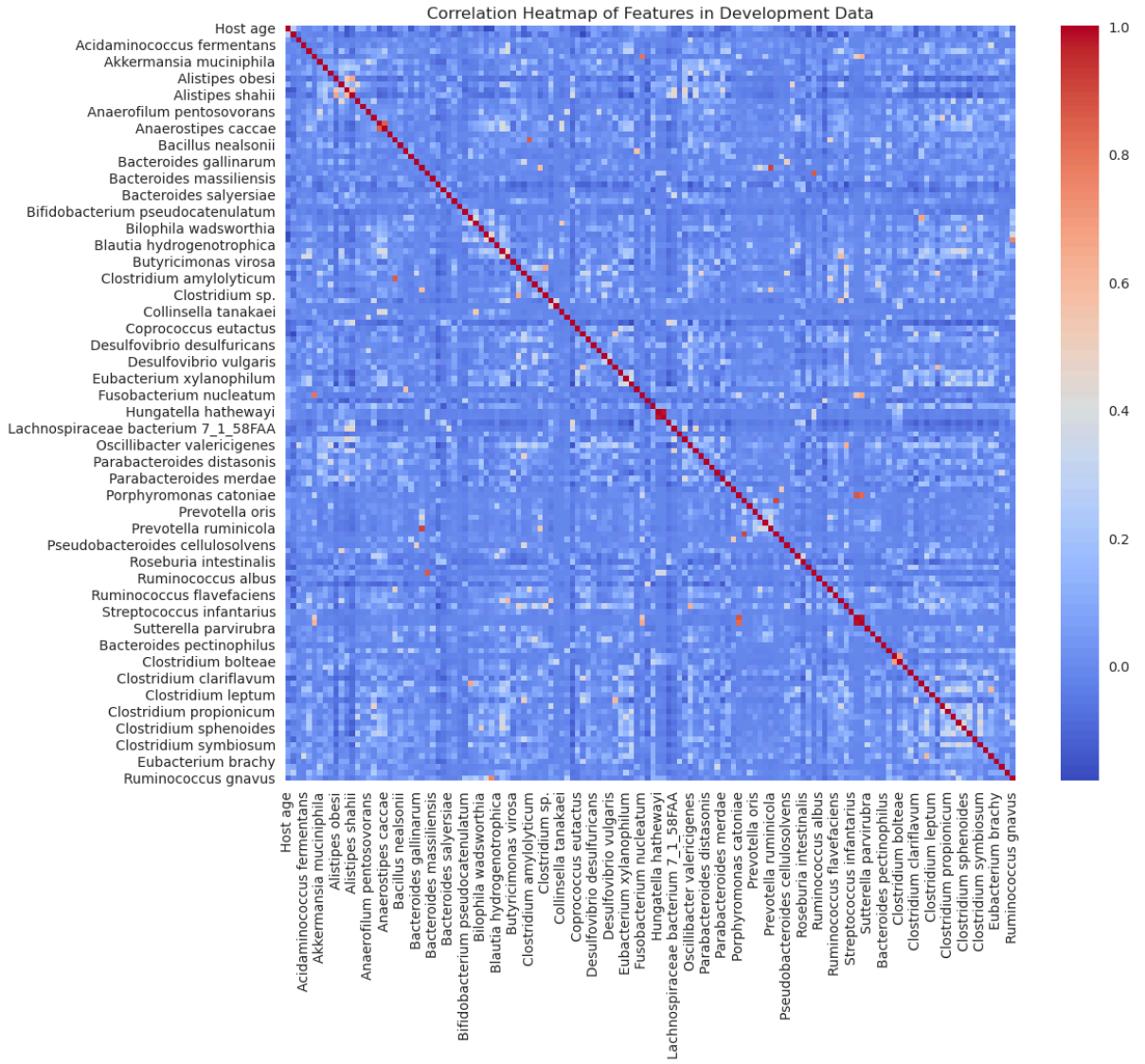
Data exploration:

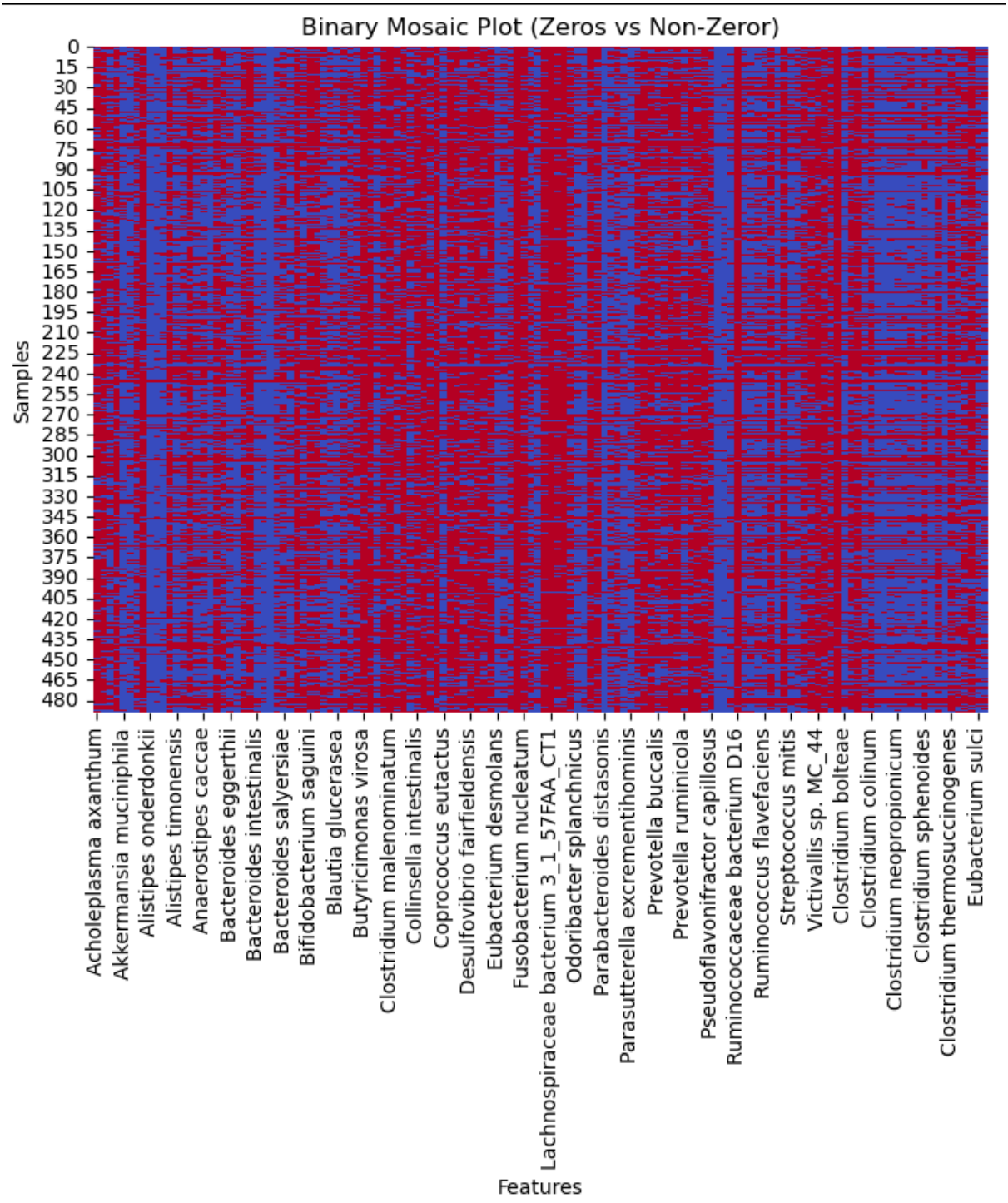
For the data exploration task I did all that is considered normal, from showing info and description of the datasets, up to calculating the zeros per feature percentage etc. I also cleaned the data from some of the irrelevant columns.

The data did not need a lot of cleaning per the usual meaning as there were no null or duplicate entries.

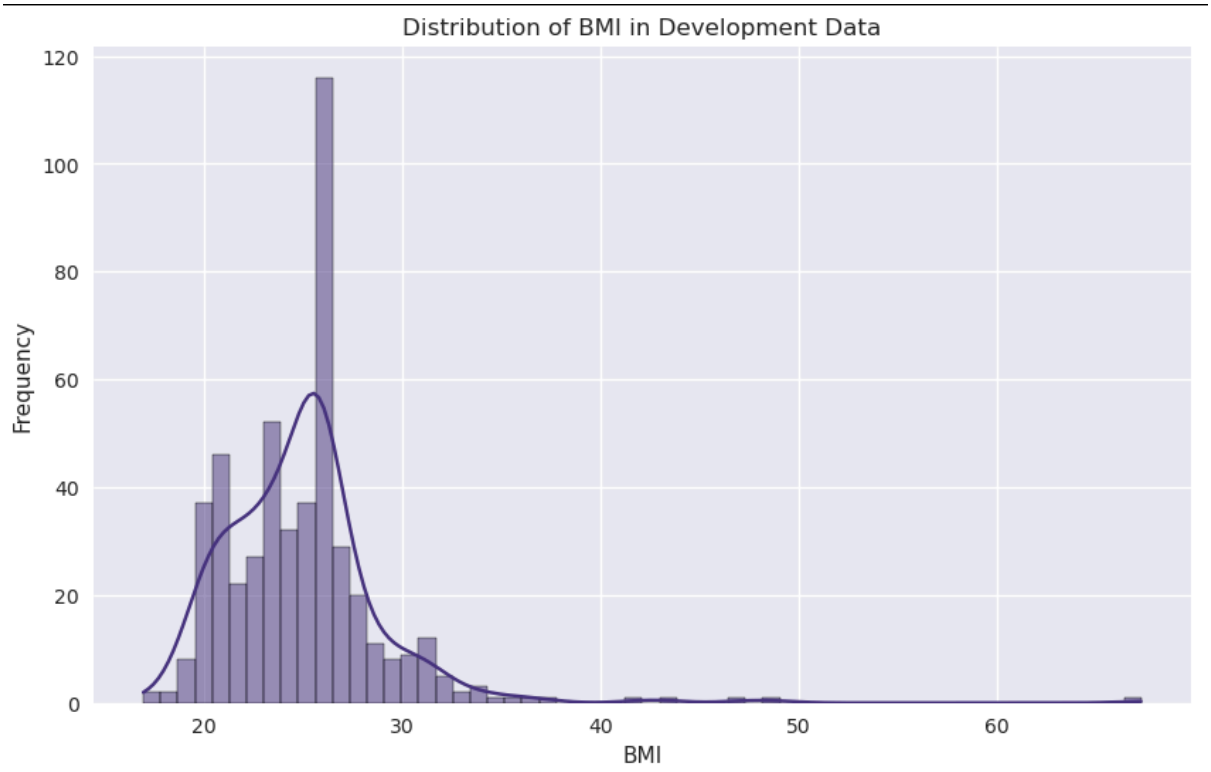
To get a better feeling of the data also the IQR was calculated and studied.

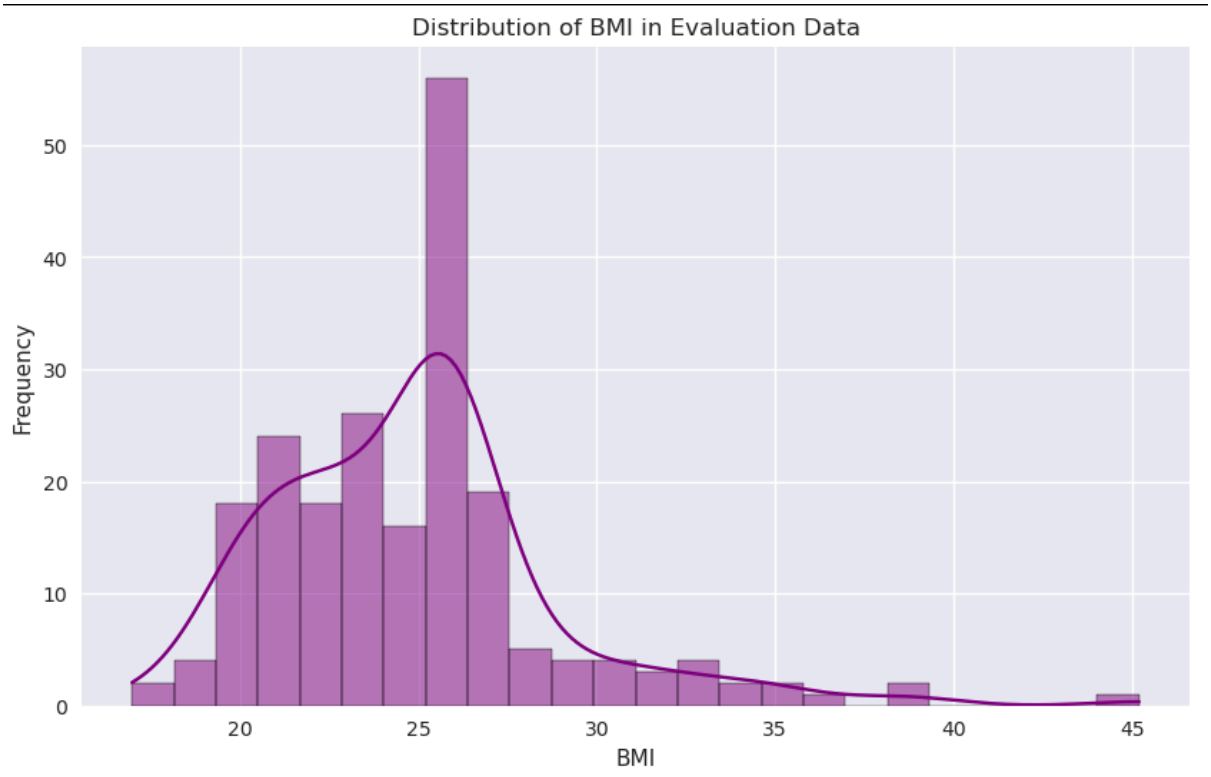
I plotted a heatmap of the features and also a mosaic of the zero vs non-zero values of the features:


Correlation Heatmap of Features in Development Data
Host age
Acidaminococcus fermentans
Akkermansia muciniphila
Alistipes obesi
Alistipes shahii
Anaerofilum pentosovorans
Anaerostipes caccae
Bacillus nealsonii
Bacteroides gallinarum
Bacteroides massiliensis
Bacteroides salyersiae
Bifidobacterium pseudocatenulatum
Bilophila wadsworthia
Blautia hydrogenotrophica
Butyricimonas virosa
Clostridium amylolyticum
Clostridium sp.
Collinsella tanakaei
Coprococcus eutactus
Desulfovibrio desulfuricans
Desulfovibrio vulgaris
Eubacterium xylanophilum
Fusobacterium nucleatum
Hungatella hathewayi
Lachnospiraceae bacterium 7_1_58FAA
Oscillibacter valericigenes
Parabacteroides distasonis
Parabacteroides merdae
Porphyromonas catoniae
Prevotella oris
Prevotella ruminicola
Pseudobacteroides cellulosolvens
Roseburia intestinalis
Ruminococcus albus
Ruminococcus flavefaciens
Streptococcus infantarius
Sutterella parvirubra
Bacteroides pectinophilus
Clostridium bolteae
Clostridium clariflavum
Clostridium leptum
Clostridium propionicum
Clostridium sphenoides
Clostridium symbiosum
Eubacterium brachy
Ruminococcus gnavus
1.0
0.8
0.6
0.4
0.2
0.0


Binary Mosaic Plot (Zeros vs Non-Zeror)
Samples
0
15
30
45
60
75
90
105
120
135
150
165
180
195
210
225
240
255
270
285
300
315
330
345
360
375
390
405
420
435
450
465
480
Acholeplasma axanthum
Akkermansia muciniphila
Alistipes onderdonkii
Alistipes timonensis
Anaerostipes caccae
Bacteroides eggerthii
Bacteroides intestinalis
Bacteroides salyersiae
Bifidobacterium saguini
Blautia glucerasea
Butyricimonas virosa
Clostridium malenominatum
Collinsella intestinalis
Coprococcus eutactus
Desulfovibrio fairfieldensis
Eubacterium desmolans
Fusobacterium nucleatum
Lachnospiraceae bacterium 3_1_57FAA_CT1
Odoribacter splanchnicus
Parabacteroides distasonis
Parasutterella excrementihominis
Prevotella buccalis
Prevotella ruminicola
Pseudoflavonifractor capillosus
Ruminococcaceae bacterium D16
Ruminococcus flavefaciens
Streptococcus mitis
Victivallis sp. MC_44
Clostridium bolteae
Clostridium colinum
Clostridium neopropionicum
Clostridium sphenoides
Clostridium thermosuccinogenes
Eubacterium sulci
Features

I also plotted the BMI distribution to get a feeling of the data:





After that a correlation matrix was calculated and the most relevant features plotted.

BMI prediction (regression models and metagenomic data):

An extensive baseline was made following the exact process described in the exercise. The initial models, with no modifications are stored in the model folder, as is the scaler.

For the feature selection I tried multiple ways and methods with multiple parameters each. I have noted on my code the result of several of them (best of category). Although, the very best result was from using RFE with 21 features on SVR (R^2 0.2168 for the baseline model), for the rest of the models and especially BayesianRidge the results were so bad that I decided to choose the most moderate LassoCV with a hand picked alpha of 0.2.

I even made a thorough algorithm to try XGB following the paper at:
Machine learning prediction of obesity-associated gut microbiota: identifying Bifidobacterium pseudocatenulatum as a potential therapeutic target
But the results were not astonishing given our data.

I also checked my results against known microbes that have an effect on human mass but did not see a real world correlation to our data. As I am not a biologist, I could not judge this result and didn't proceed further with it.

For all my tries, I have "pretty print" of the deltas of absolute and percentage numbers as well as thorough visualizations which can be found in the notebooks.

The final reevaluation of the trained models is quantified and stored in the models folder as a json file for further investigation.

I found out about the last steps (after evaluation) a bit before the time end for the exercise (mixing it with the bonus part unfortunately for me). As such, a small piece of code may be missing. One such small thing is the function to retrain on the whole merged dataset. I have, however, a function to make the relevant pipeline to run the whole training.

Citations:

Wu H, Li Y, Jiang Y, Li X, Wang S, Zhao C, Yang X, Chang B, Yang J, Qiao J. Machine learning prediction of obesity-associated gut microbiota: identifying *Bifidobacterium pseudocatenulatum* as a potential therapeutic target. Front Microbiol. 2025 Feb 5;15:1488656. doi: 10.3389/fmicb.2024.1488656. PMID: 39974372; PMCID: PMC11839209.

Numerical Comparison: Different Methods of Handling Zeros in Microbiome Data Analysis

https://scikit-learn.org

AI disclosure:

I use perplexity.ai (with a mixture of models -> Chatgpt, Claude, etc).
The main usage (as you ask about it) is helper functions (especially data edits for plots) and cutting on the writing in general. Also, I sometimes use it to quickly learn about some bits and bytes that I may be missing to understand something.