Assignment #1: Machine Learning in Computational Biology

07/04/2025
Ioanna Elissavet Gavra

1. Introduction

The gut microbiota is increasingly linked to various health outcomes, including obesity, cardiovascular diseases and metabolic disorders. The aim of this assignment was the development of a Body Mass Index (BMI) prediction algorithm using microbiome data through machine learning models. This report outlines the dataset, preprocessing steps, feature selection methods, model implementations, and performance evaluation.

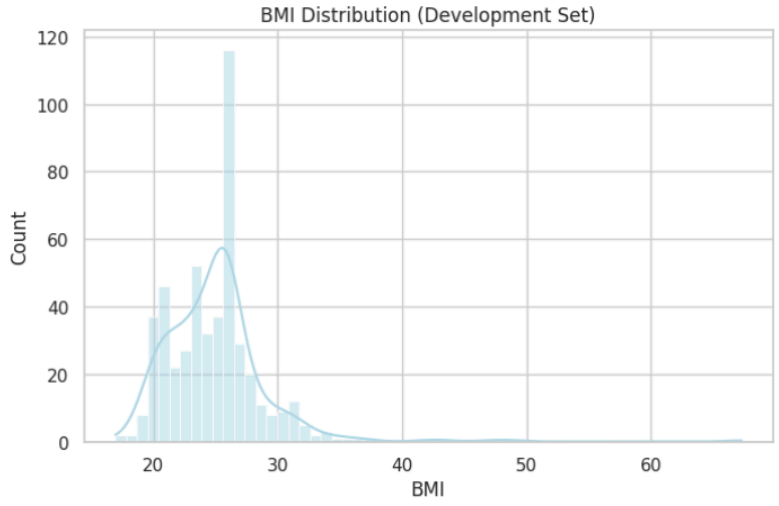
2. Dataset exploration

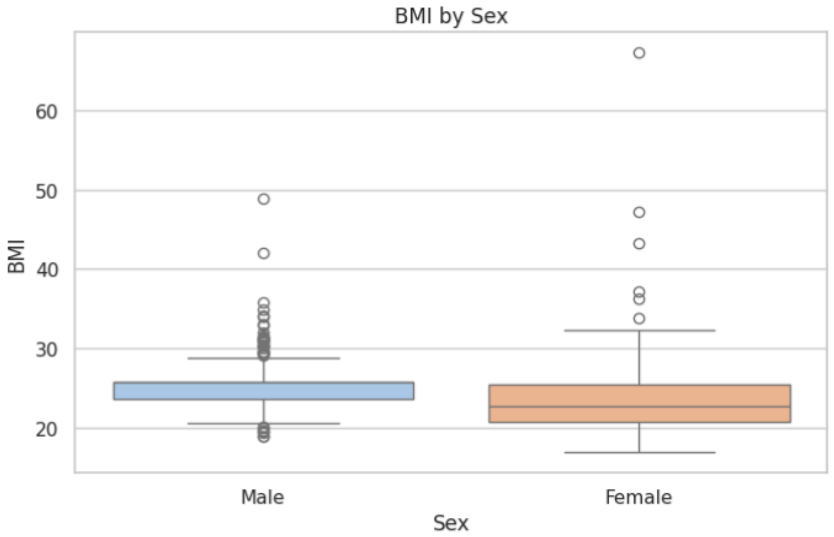
The given dataset consists of microbiome data, coming from six different projects, with various bacterial species. Each sample includes the microbiome data along with demographic information such as sex, host age, and BMI. The goal is to leverage these microbiome features to predict BMI using machine learning models.

Demographic data of the two given datasets, development & evaluation sets.

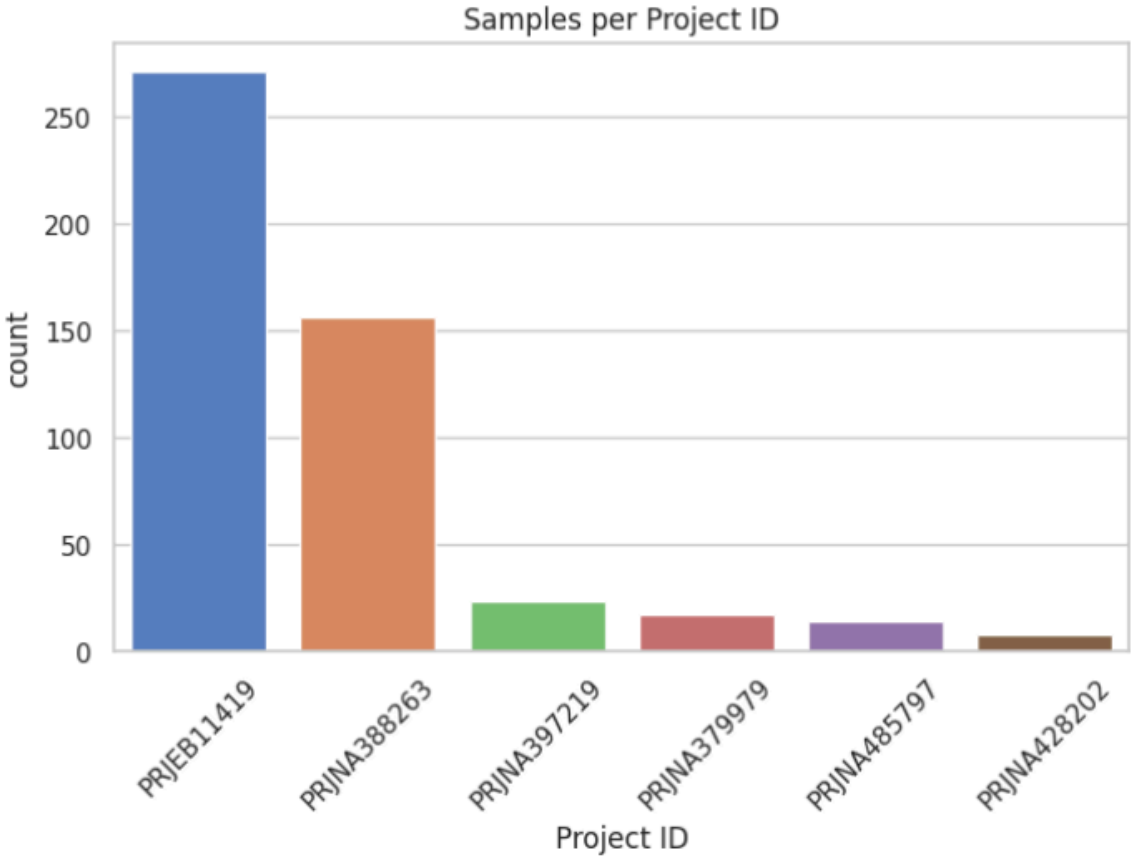
Dataset	N of samples	Male (%)	Female (%)	Age	BMI
Development	489	302 (61.8%)	187 (38.2%)	46.7 ± 15.7	24.9 ± 4.2
Evaluation	211	142 (67.3%)	69 (32.7%)	46.3 ± 16.1	24.8 ± 4.0

Sex and age previously established correlation with BMI, and will be excluded from feature exploration in order to be used later as confounding factors. Males are 61.8% of our dataset and females the 38.2%. Below one can see the BMI distribution across all samples and then the BMI per sex.



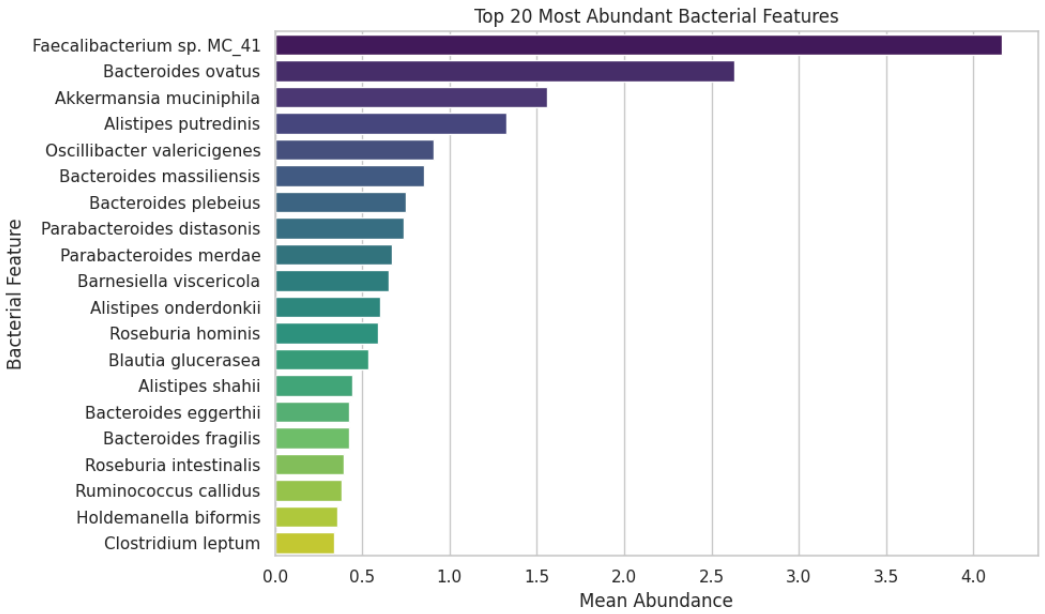


The project from where each sample was derived is also a confounding factor, are the experiment circumstances may affect the counted microbiota abundances.

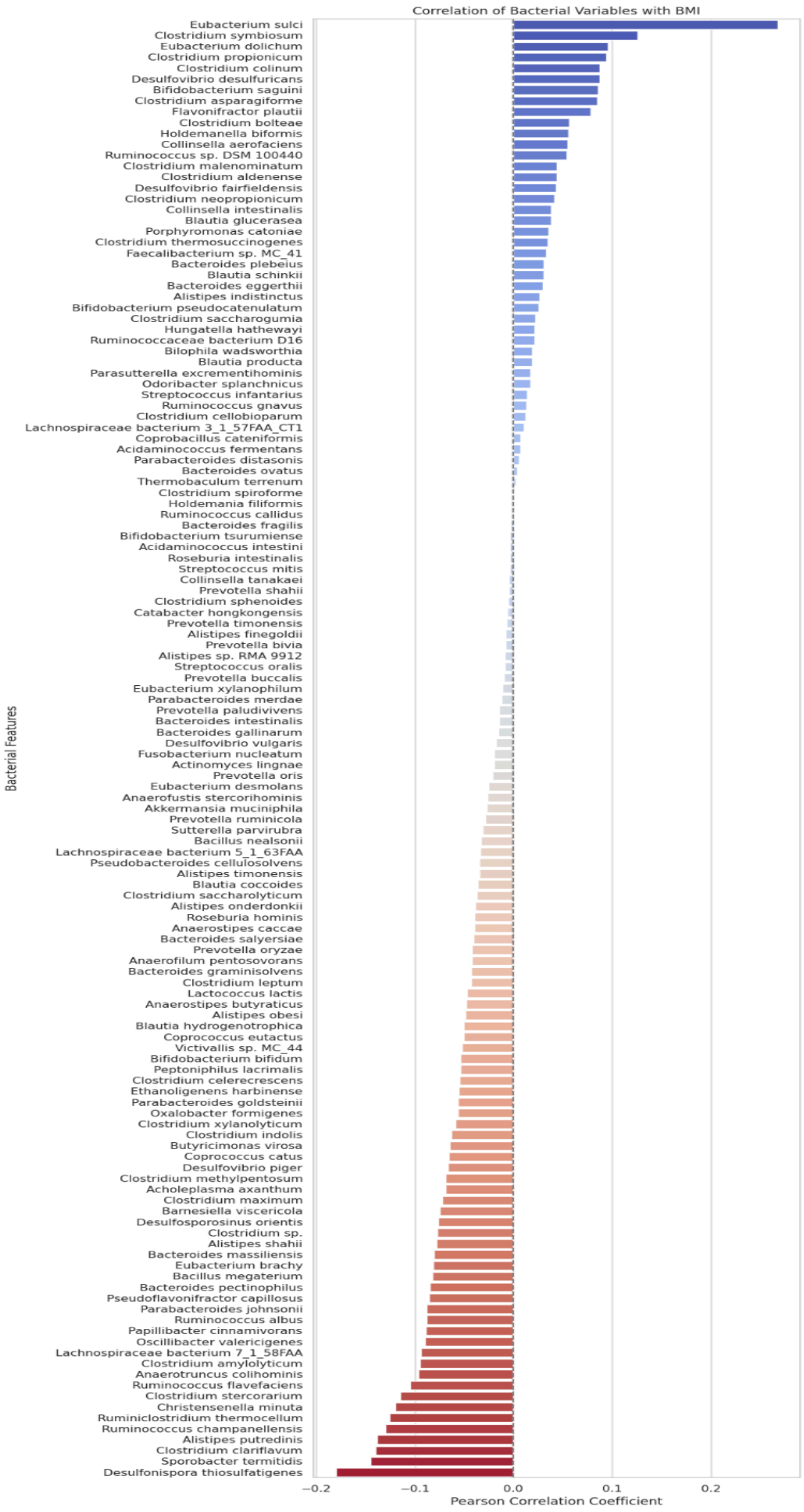


As it is reported, most of the samples come from two projects, and only a small proportion came from the four others.

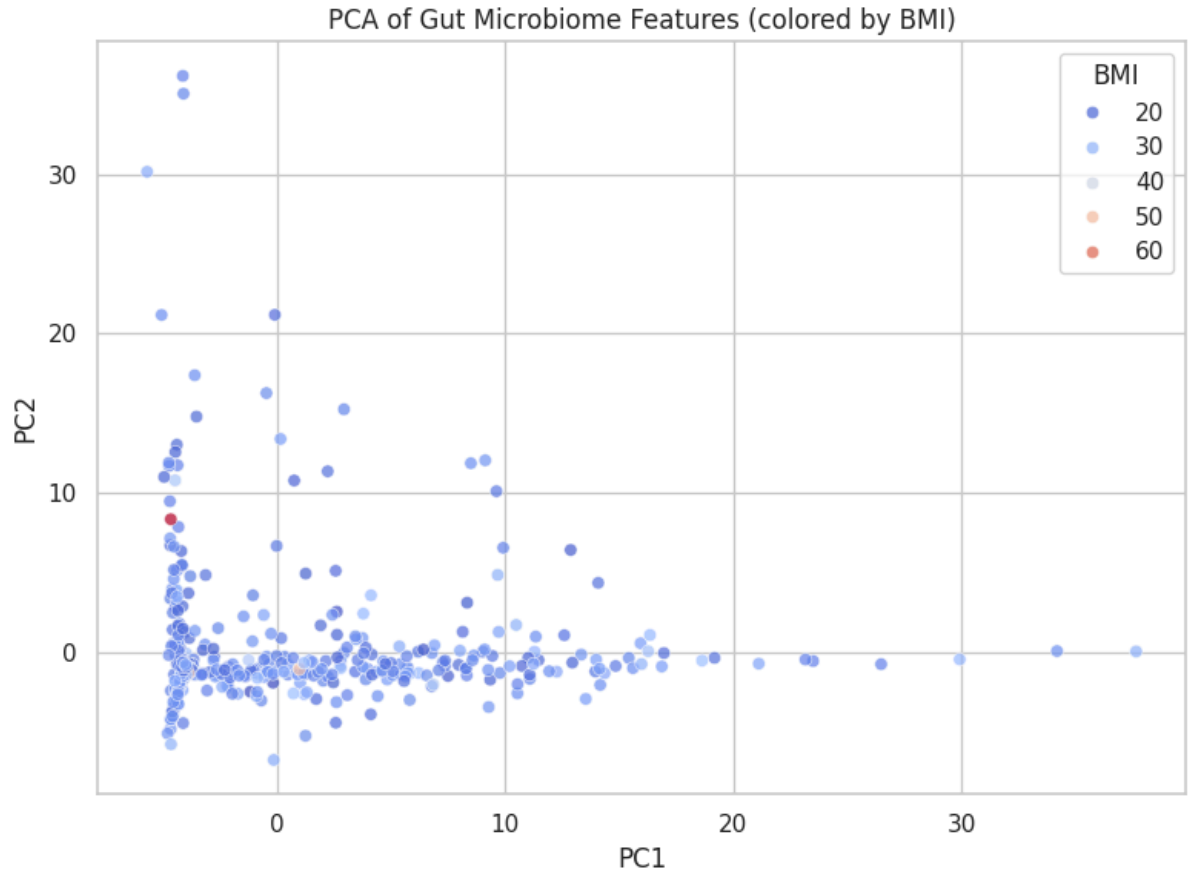
It was also useful to present the most abundant gut microorganisms to gain insight regarding the subject.



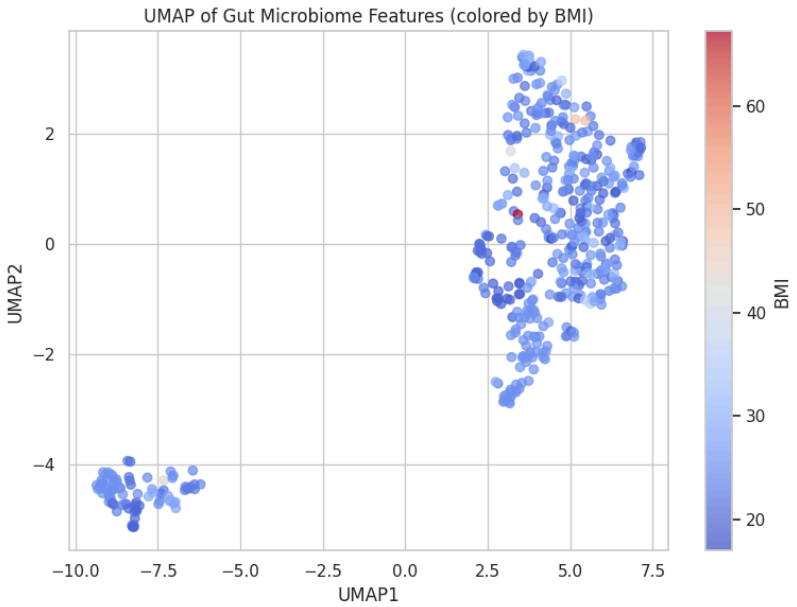
The Pearson correlation coefficient was computed in order to examine some correlation of the species with the variable of interest, BMI.


Correlation of Bacterial Variables with BMI
Eubacterium sulci
Clostridium symbiosum
Eubacterium dolichum
Clostridium propionicum
Clostridium colinum
Desulfovibrio desulfuricans
Bifidobacterium saguini
Clostridium asparagiforme
Flavonifractor plautii
Clostridium bolteae
Holdemanella biformis
Collinsella aerofaciens
Ruminococcus sp. DSM 100440
Clostridium malenominatum
Clostridium aldenense
Desulfovibrio fairfieldensis
Clostridium neopropionicum
Collinsella intestinalis
Blautia glucerasea
Porphyromonas catoniae
Clostridium thermosuccinogenes
Faecalibacterium sp. MC_41
Bacteroides plebeius
Blautia schinkii
Bacteroides eggerthii
Alistipes indistinctus
Bifidobacterium pseudocatenulatum
Clostridium saccharogumia
Hungatella hathewayi
Ruminococcaceae bacterium D16
Bilophila wadsworthia
Blautia producta
Parasutterella excrementihominis
Odoribacter splanchnicus
Streptococcus infantarius
Ruminococcus gnavus
Clostridium cellobioparum
Lachnospiraceae bacterium 3_1_57FAA_CT1
Coprobacillus cateniformis
Acidaminococcus fermentans
Parabacteroides distasonis
Bacteroides ovatus
Thermobaculum terrenum
Clostridium spiroforme
Holdemania filiformis
Ruminococcus callidus
Bacteroides fragilis
Bifidobacterium tsurumiense
Acidaminococcus intestini
Roseburia intestinalis
Streptococcus mitis
Collinsella tanakaei
Prevotella shahii
Clostridium sphenoides
Catabacter hongkongensis
Prevotella timonensis
Alistipes finegoldii
Prevotella bivia
Alistipes sp. RMA 9912
Streptococcus oralis
Prevotella buccalis
Eubacterium xylanophilum
Parabacteroides merdae
Prevotella paludivivens
Bacteroides intestinalis
Bacteroides gallinarum
Desulfovibrio vulgaris
Fusobacterium nucleatum
Actinomyces lingnae
Prevotella oris
Eubacterium desmolans
Anaerofustis stercorihominis
Akkermansia muciniphila
Prevotella ruminicola
Sutterella parvirubra
Bacillus nealsonii
Lachnospiraceae bacterium 5_1_63FAA
Pseudobacteroides cellulosolvens
Alistipes timonensis
Blautia coccoides
Clostridium saccharolyticum
Alistipes onderdonkii
Roseburia hominis
Anaerostipes caccae
Bacteroides salyersiae
Prevotella oryzae
Anaerofilum pentosovorans
Bacteroides graminisolvens
Clostridium leptum
Lactococcus lactis
Anaerostipes butyraticus
Alistipes obesi
Blautia hydrogenotrophica
Coprococcus eutactus
Victivallis sp. MC_44
Bifidobacterium bifidum
Peptoniphilus lacrimalis
Clostridium celerecrescens
Ethanoligenens harbinense
Parabacteroides goldsteinii
Oxalobacter formigenes
Clostridium xylanolyticum
Clostridium indolis
Butyricimonas virosa
Coprococcus catus
Desulfovibrio piger
Clostridium methylpentosum
Acholeplasma axanthum
Clostridium maximum
Barnesiella viscericola
Desulfosporosinus orientis
Clostridium sp.
Alistipes shahii
Bacteroides massiliensis
Eubacterium brachy
Bacillus megaterium
Bacteroides pectinophilus
Pseudoflavonifractor capillosus
Parabacteroides johnsonii
Ruminococcus albus
Papillibacter cinnamivorans
Oscillibacter valericigenes
Lachnospiraceae bacterium 7_1_58FAA
Clostridium amylolyticum
Anaerotruncus colihominis
Ruminococcus flavefaciens
Clostridium stercorarium
Christensenella minuta
Ruminiclostridium thermocellum
Ruminococcus champanellensis
Alistipes putredinis
Clostridium clariflavum
Sporobacter termitidis
Desulfonispora thiosulfatigenes
Bacterial Features
−0.2
−0.1
0.0
0.1
0.2
Pearson Correlation Coefficient

Principal Component Analysis (PCA) was applied to visualize the dataset in a reduced dimensional space. There was no clear separation observed based on BMI. This could suggest that BMI-related variation in microbiome features is subtle and needs further investigation.



(UMAP) technique was also employed to analyze the microbiome in relation to BMI. This visualization suggests potential microbiome differences associated with BMI, as reflected by the two clusters. Further investigation is needed to determine whether these groups represent distinct microbiome profiles linked to BMI categories.



3. Methodology

The preprocessing steps involved handling numerical features, including bacterial species, while excluding confounding factors. To ensure that only the most relevant features were included, a feature selection approach was implemented using the SelectKBest method with mutual information regression. This method helped identify the top ten bacterial species most informative for BMI prediction. The same feature selection process was separately applied to the validation set to ensure unbiased evaluation. Also to ensure dataset balance, the python caret package was used for data split.

Following feature selection, three regression models were implemented: Elastic Net Regression, Ridge Regression, and Bayesian Ridge Regression. These models were trained using the selected features and evaluated based on three key metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared Score (R^2).

Please note that LLM was used for coding function creation.

4. Results

Baseline model:

	RMSE	RMSE	R-squared
Elastic Net	3.28960	2.62299	-0.02311
Ridge Regression	8.85295	5.3254	-6.40991
Bayesian Ridge	3.38893	2.50788	-0.08583

Model with Feature selection:

	RMSE	RMSE	R-squared
Elastic Net	3.31522	2.63922	-0.03911
Ridge Regression	3.3385	2.60719	0.05379
Bayesian Ridge	3.31692	2.63489	-0.04017

	RMSE	RMSE	R-squared
Elastic Net	3.96502	2.86911	0.00640
Ridge Regression	3.76948	2.57328	0.10199
Bayesian Ridge	3.85427	2.73313	0.06114

5. Conclusion

The Ridge regression model seems to output the best metrics overall and on the evaluation set.