STAT 542 Final Project

Microbiome Data Analysis

Submitted by

Manan Mehta (mananm2) (Lead)

Darren Wang (hsiangw2)

A report for the partial completion of

STAT 542 (Statistical Learning)

University of Illinois at Urbana-Champaign

12/11/2020

1 Project Description and Summary

Increasing amount of studies have shown the importance of the role microbiome plays in the human body, especially its relationship with health, disease and longevity. Metagenomic tools and high-throughput sequencing have allowed scientists to analyze unbiased quantification of all microbes constituting the microbiome. In this project, data analysis was performed on a microbiome data processed from the American Gut database at this Biocore GitHub repository to capture the relationship between microbiome and the host's body mass index as well as alcohol consumption habit. To deal with sparsity and compositional nature of microbiome data, specialized data transformation methods were proposed. Supervised-learning methods were applied to capture the presence of underlying clusters in microbiome. Though naturally-formed clusters were observed in the reduced space of data spanned by the microbiome, the clusters failed to exhibit differences in hosts' health condition and lifestyle indicators. Supervised-learning methods including KNN, Random Forest and AdaBoost were used to model body mass index and alcohol consumption habit using microbiome data and demographic attributes. Our study shows little predictive ability of the microbiome on the two outcome variables. Due to time limitation, several possible directions for improving the results along with justification of our obtained results were outlined in the Collaborator's Question Section.

2 Literature Review

The human gut microbiome composition has been extensively studied in the last 2 decades to analyze the relationships between gut composition and other physical attributes. Some of these physical attributes include the gut-brain axis [1], gastrointestinal diseases [2], and even deciphering genomics [3]. Since our project involved analyzing a compositional and sparse data set of gut microbiota composition, we primed our basic knowledge of gut bacteria using relevant articles from literature. We provide a summary below.

To begin with, [1] provides a simple introduction to the human gut microbiota composition, along with some relationships between bacterial species and diseases which have been modeled in the past. A particularly useful discussion is provided for the taxonomic classification of the gut bacteria, using detailed examples for the taxa names. For example, the most common Phyla of bacteria found in the human gut include Firmicutes and Bacteriodetes, which was also seen from our data set. This article provided a good context for the importance of modeling gut bacteria, along with an introduction to bacterial taxonomy.

In [4], the authors summarized several methods developed to quantify the relative abundances of microbial taxa. Details about how data was collected and quantified were also outlined. Using a compositional microbiome data, alpha- and beta-diversity can be used to analyze the data at a community level. Differential analysis methods based on compositional data were introduced, including the use of the Dirichlet class of distributions for modeling the data, the use of Kent distribution to model square root-transformed data, and using logistic normal class to model log-ratio transformed data. Modified regression analysis methods that add linear constraints on coefficients to deal with compositional covariates were also introduced. Thus, this article provided a context for how to address high dimensional and compositional data.

Further, [5] illustrated obstacles in analyzing compositional data using an example with negative correlation bias. Centred-log-ratio (clr) transformation was introduced as an data processing method before applying compositional data analysis (CoDa). Principal component analysis (PCA) was used on clr-transformed data to (1) Summarize the data, (2) Analyze taxa that drive difference between groups, and (3) Determine if there are taxa with correlated abundances. A detailed case study was performed as an example. Overall, this article introduced us to the clr transformation, which we use extensively for the pre-processing of our data set.

Lastly, [6] discussed bacteria that were proved to be associated with human health conditions. High levels of Bacteroidetes and Firmicutes, along with low levels of Proteobacteria, Actinobacteria, Fusobacteria, Verrucomicrobia were expected to be observed in healthy adults. This is also attested by our data set. The Firmicutes/Bacteroidetes (F/B) ratio was introduced as an indicator of host's health condition, the ratio is also reported to increase from birth to adulthood. Several other orders, genus, and classes of bacteria were claimed to be associated with human health. This article introduced the use of F/B ratio as another covariate to be analyzed, which we use in some unsupervised learning tasks.

3 Unsupervised Learning

3.1 Data Pre-processing

Our data set consists of 9511 samples, each with 32,961 Operational Taxonomic Unit (OTU) variables measured in relative abundance. Each column highlights the taxonomic hierarchy of the microbiome in question, in the Phylum-Class-Order-Family-Genus-Species format. Since the data provides relative abundance, the dataframe is compositional in nature i.e. $\sum_j \{\text{feature}\}_{ij} = 1 \ \forall i$. In such cases, traditional methods of analysis and interpretation are rendered useless [5, 7]. Furthermore, since not all microbes are present in all samples, the data is very sparse. For example, 92.83% of the variables have less than 1% nonzero entries. Lastly, there are several outliers in the data which may severely affect learning results. We deal with these challenges in the following ways:

3.1.1 Dealing with Compositional Data

As explained in [7], compositional data should not be directly analyzed using standard statistical techniques that assume independence of underlying variables. In case of compositional data, the data points do not map to a Eucledian space but to a hyperplane referred to as the Aitchison Simplex, as first explained by J. Aitchison in a classic article [8]. A widely used and convenient data-transformation technique is the centered-log-ratio (clr) transformation. We employ this technique to transform our data before employing any machine learning technique. In general, if $\mathbf{X}$ is a vector of numbers that contains d parts i.e. $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_d]$, then $\text{clr}(\mathbf{X})$ can be computed as:

$$\mathbf{X}_{\text{clr}} = \left[\log\left(\frac{\mathbf{x}_1}{g_x}\right), \log\left(\frac{\mathbf{x}_2}{g_x}\right), ... \log\left(\frac{\mathbf{x}_d}{g_x}\right)\right]$$

where g_x is the geometric mean of all values in $\mathbf{X}$ [8]. After using this transformation, most multivariate analysis techniques can be applied. Further, since the shape of the data is reconstituted, some univariate analyses are also likely to be valid.
Since our data has several zeros, division by g_x is not numerically stable. If our data represents the counts per taxon through the process of random sampling, some zero values could arise simply by random chance, while others arise because of true absence of the taxon in the environment. Unfortunately, using Bayesian methods to estimate the likelihood of zero values with the compositional analysis was not possible in our project time frame. Thus, for numerical stability, we substitute the zero values in the data by a small relative value of 1e-10 prior to applying the clr transformation.

3.1.2 Dealing with Sparsity

Sparsity of data poses a challenge for supervised learning, as classic clustering algorithms that measure distance between samples are likely to produce compromised results. Distance measurements like Euclidean or Mahalanobis distance are computationally expensive in the high dimensional space and are likely to be distorted. We deal with this sparsity using some intuitive strategies:

1. Merge variables by taxonomic rank:
   OTU variables follow hierarchical ordering of taxonomic rank and in our analysis, several

strategies including grouping OTUs by kingdom, phylum, order and genus were examined. For example, to reduce dimensionality, we merged all the species belonging to one genus as a single variable. This automatically reduces sparsity too, as different columns contribute values from different samples, while preserving the nature of the bacterial classification.

2. Greedy merging:
   A greedy approach was examined to further reduce the sparsity of OTU variables. Variables containing (1) less than a threshold number of non-zero values or (2) having maximum value less than the mean of the data set were merged into a separate column named 'Other Species'. The value of the threshold was tuned differently for different analysis techniques by conducting several experiments and observing the results.

3. PCA:
   PCA was performed to (1) analyze clustering results using the first few PCs and (2) visualize clustering results from the overall high-dimensional data.

3.1.3 Dealing with outliers and missing values

In order to deal with missing values from the data ('Not Provided' strings), we consider three different methods: (1) Discard any sample containing missing values in one or more variables, (2) Treat 'Not provided' as a separate level in the variables, (3) Impute missing values. Since the data was collected from multiple sources using different methods, popular imputation methods like MICE, mean/median value imputation are likely to fail here. Treating 'Not provided' as a separate level was also not justified since we don't have strong evidence to believe that these missing values occur in a systematic way. Moreover, there's only a small subset of samples containing missing values in the data(around 6.5%). We thus choose to discard samples with missing values. This left us with 8,311 samples for all analyses.

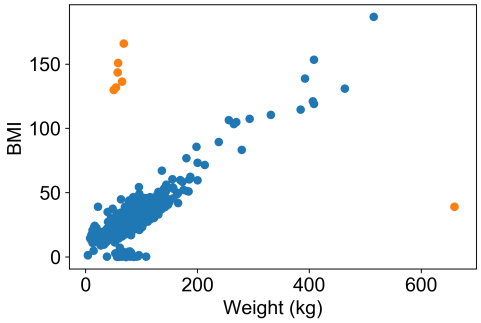


Figure 1: Outliers in weight and BMI

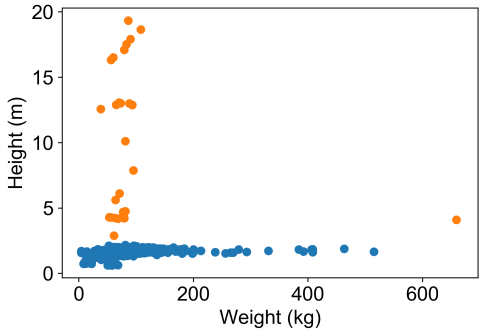


Figure 2: Outliers in height

The heaviest man on record weighed 594.8 kg and had a BMI of 180. Using this information, we identified outliers as as shown in Figure 1. There were also a group of samples with very low weights and BMI above 100. Intuitively, these must be severely overweight teens or children, however they all aged over 30, and hence should be excluded.
Further, with BMI and weights available, heights of all samples in meters can be calculated using the ratio of weight and BMI. Through this, we further identified a group of outliers as samples with height over 2.5 meters as shown in Figure 2. These were discarded from the data set too, leaving the final analyzable data set with 7,711 samples after all outliers were removed.

3.2 Results

Given the high-dimensional and sparse data, unsupervised learning algorithms gave very little signal. No firm clusters were found on applying clustering algorithms to several different versions of modified data. The data here was modified using either greedy merging of variables, or by taxonomic rank. We examined the results when the OTUs are merged together while maintaining the taxonomic classification. For example, at the genus level, all the species belonging to the same genus were merged together into their individual genus. Doing this across the data set reduced the number of OTU variables from 32000 (individual species) to 832 (individual genuses). This can logically be extended further up the hierarchy tree, to the family, order, class, or phylum levels. This method progressively reduces sparsity and dimensions. On applying K-means clustering to the Phylum level data (44 variables only), we see the clusters as shown in Figure 3 which are plotted on the first 3 PCs for visualization.

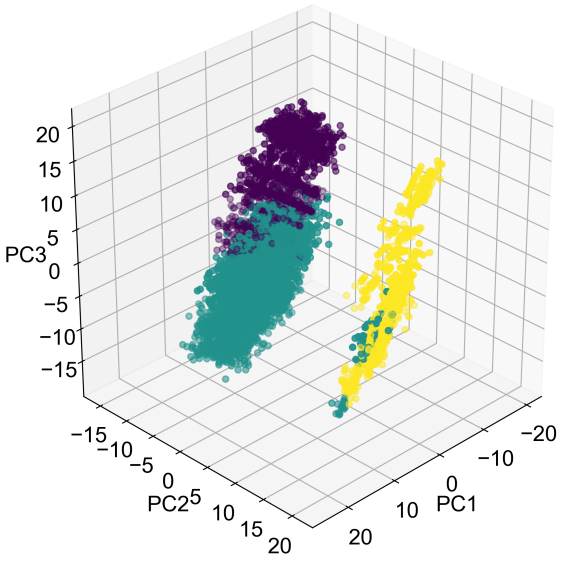


Figure 3: Example of k-means clusters plotted on 3 PCs

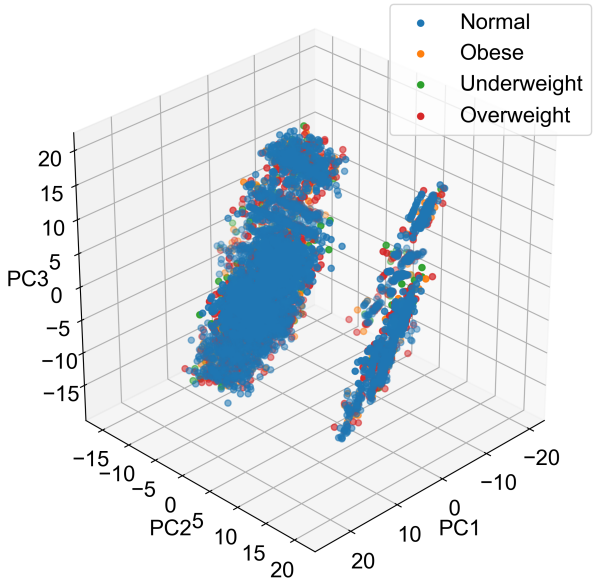


Figure 4: Categorical BMI plotted on first 3 PCs

In this case, the best clusters were obtained for k = 3, which indirectly translates to the three main phylums i.e. Bacteriodetes, Fermicutes, and all other phylums merged together into a third cluster. This trend can be seen for all levels of taxonomy. Unfortunately, it does not give us any new information w.r.t. the categorical variables we model. For example, Figure 4 plots the categorical BMI on to the first 3 PCs. No trend is seen in this case, and this is the same with all other categorical variables. This leads us to believe that there is very little signal in the data, which is further explained in Section 6.

4 Supervised Learning: Classification

4.1 Modeling BMI categorical

To model BMI categorical, we discard the variables directly associated with it i.e. BMI and weight. Two extra variables introduced in [6] were added to the data set viz. α diversity and F/B ratio. α diversity is defined to be the number of non-zero microbiome species in a sample, and F/B ratio is the sum of Firmicutes phylum divided by sum of Bacteroidetes phylum of a sample. Samples with missing values and identified outliers were also discarded, as explained in Section 3.1.3.

While performing EDA for BMI categorical, a severe class imbalance was observed, in that the data was dominated by samples with normal BMI. This is shown in Figure 5. This imbalance adversely affects modeling results. Several techniques were used to deal with this issue, which are further discussed with the results.

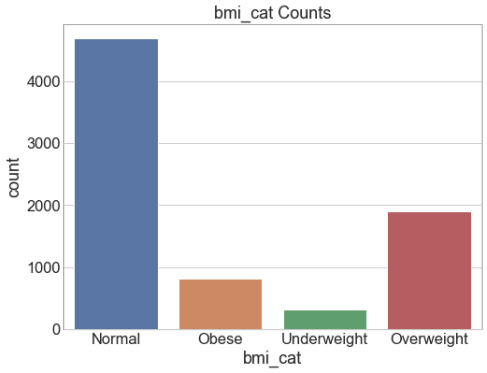


Figure 5: Class Imbalance in categorical BMI

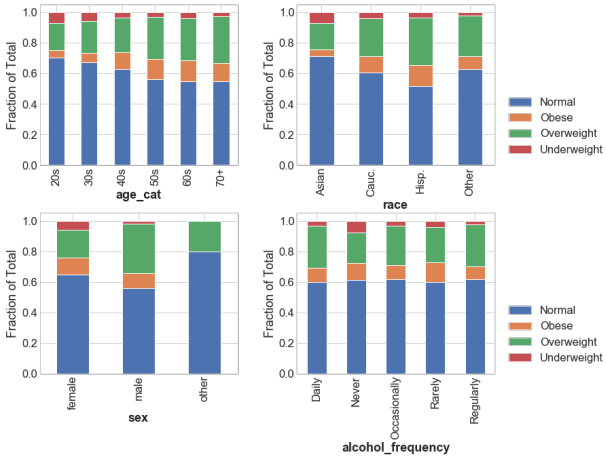


Figure 6: Univariate analysis for BMI

Further, in Figure 6, categorical BMI frequency was plotted against each categorical variable. Age demonstrated the strongest relationship with categorical BMI, the fraction of overweight samples increases with age, and vice versa for underweight samples. Race also exhibits a stronger relationship with categorical BMI, Caucasian and Hispanic people are slightly more likely to be overweight/obese. However, sample sizes from each race are extremely uneven.

Age demonstrated the strongest relationship with categorical BMI, the fraction of overweight samples increases with age, and vice versa for underweight samples. Race also exhibits a stronger relationship with categorical BMI, Caucasian and Hispanic people are slightly more likely to be overweight/obese. However, sample sizes from each race are extremely uneven.

Before applying classification models, the data was split into training and testing sets in a 9:1 ratio with stratified random sampling. **Random Forest** was used to model categorical BMI and several techniques were employed to further improve the model. 14 variables were used in the model: age_cat(ordinal), alcohol_frequency(ordinal), alpha_diversity, log_fb_ratio, K_Archaea, k_Bacteria|p_Actinobacteria, k_Bacteria|p_Bacteroidetes, k_Bacteria|p_Firmicutes, k_ Bacteria|p_Proteobacteria, k_Bacteria|p_Tenericutes, k_Bacteria|Others, Other OTUs, sex, and race. Race and sex were further encoded using one-hot encoding. Training set and testing set were both 19 dimensional. To deal with the problem of imbalance in the target variable, the model was trained using weighted F1-score as the loss function. Also, sample weights were

assigned inversely proportional to class frequencies in the input data.
3-fold stratified cross-validation was used to tune parameters. Five parameters were tuned, the best set of parameters, as well as their search space is as summarized as:
max_depth: 30 (5 - 50), max_features: 6 (1 - 10), min_samples_leaf: 10 (10 - 30), min_samples_split: 10 (10 - 30), n_estimators: 300 (50 - 500).
The best model achieved 0.484 F1-Score in testing data. Though reducing min_samples_leaf and min_samples_split could increase testing F1 score slightly to around 0.51, that compromised the accuracy of less frequent classes (obese, overweight). Therefore, this model is considered the best because of its ability to correctly classify more than half of the overweight and obese samples in the testing data. The result is shown in Figure 7.

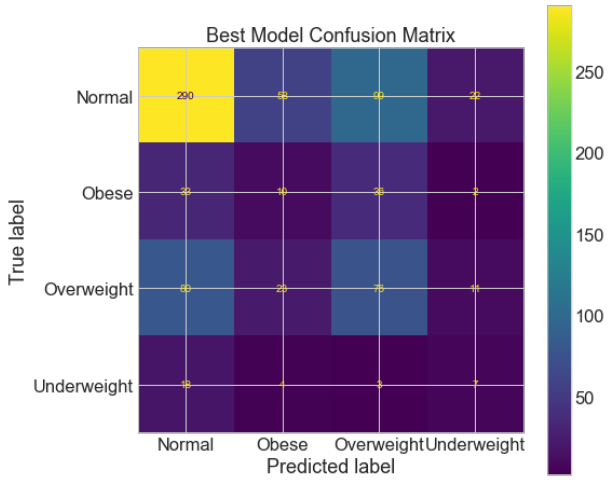


Figure 7: Best Model Confusion Matrix for categorical BMI classification

4.2 Modeling Alcohol Frequency

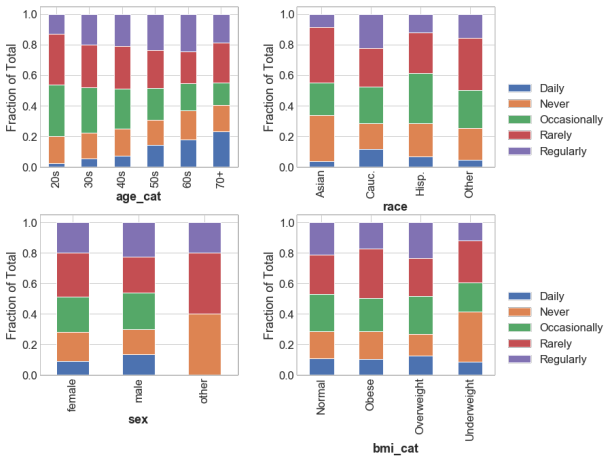


Figure 8: Univariate analysis for alcohol frequency

To model alcohol frequency, first samples missing values for alcohol frequency were excluded. Next, samples collected from babies, children and teens were also excluded because people in these age-groups would not drink and their OTU variables were not useful. The same preprocessing steps taken for modeling BMI categorical were also performed. Unlike the previous case, class imbalance was not severe enough to raise concern, thus classification algorithms were directly applied.
As a part of univariate analysis, each level in alcohol frequency was plotted as fraction of total against categorical variables, as seen in Figure 8. An almost exponential increase in proportion of daily drinkers was observed in the upper-left plot; the proportion of other

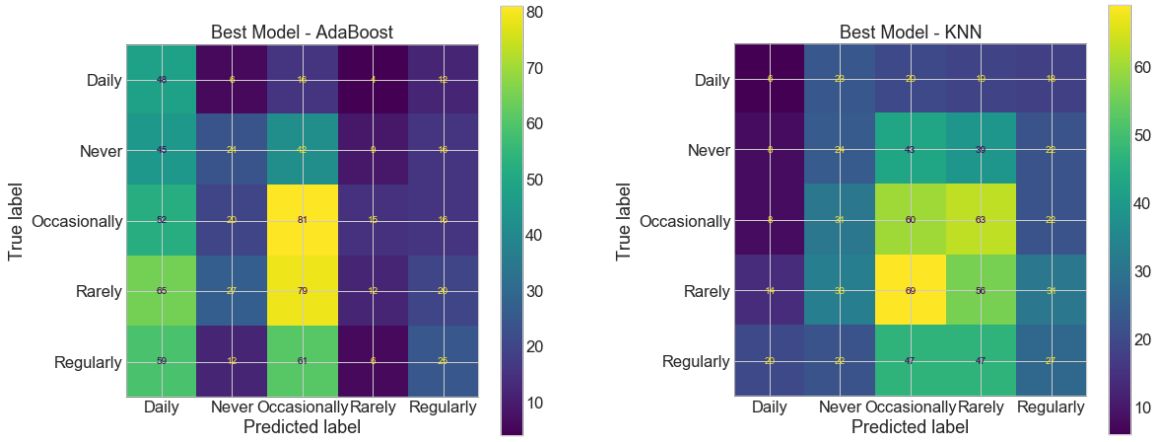
levels also demonstrated linear patterns with the growth of age. Relationships between alcohol frequency and race, categorical BMI, and sex were weak considering the uneven sample size between levels in these variables.
Before applying classification models, the data was split into training and testing sets in a 9:1 ratio with stratified random sampling. For alcohol frequency, we used **AdaBoost** and **KNN** were used to model the data.
18 variables were used to model alcohol frequency: age_cat(ordinal), bmi_cat, weight_kg, height_m, alpha_diversity, log_fb_ratio, K_Archaea, k_Bacteria|p_Actinobacteria, k_ Bacteria|p_Bacteroidetes, k_Bacteria|p_Firmicutes, k_Bacteria|p_Proteobacteria, k_Bacteria|p_ Tenericutes, k_Bacteria|Others, Other OTUs, sex, and race. Race, bmi_cat, and sex were further encoded using one-hot encoding. Training set and testing set were both 25 dimensional.

Using AdaBoost

When using AdaBoost, 3-fold stratified cross-validation was used to tune parameters with weighted F1-score, similar to the approach used with random forest while modeling categorical BMI. Sample weights were assigned inversely proportional to class frequencies in the input data. Decision tree model was used as the base learner for AdaBoost. Five parameters were tuned, the best set of parameters, as well as their search space is as summarized as:
Decision Tree: max_depth: 1 (1-5), min_samples_leaf: 10 (10-50), min_samples_split: 30 (5-50).
AdaBoost: n_estimators: 300 (50-500).
The best model achieved 0.274 F1-Score on testing data. See Figure 9a for the confusion matrix.

Using KNN

Similarly, 3-fold stratified cross-validation was used to tune the number of neighbors k with a weighted F1-score. However, sample weights were assigned equally to fit the KNN model. The optimal number of neighbors to use was found to be 10. Best weighted F1 score was 0.233 on the testing set. See Figure 9b for the confusion matrix.



(a) (b)

Figure 9: Alcohol Frequency classification using (a) AdaBoost and (b) KNN

5 Supervised Learning: Regression

After using supervised learning for classification of categorical variables, we now analyze the BMIs of samples (float) using regression. The standard methods of regression do not apply to compositional data as we observe in our original data set. Thus, before performing regression analysis, we use the centered-log-ratio transformation on the data to deal with its compositional nature. Next, we follow the same procedure as used for unsupervised learning and classification to remove outliers from the data. On doing this, we are left with 7,711 (n) samples and ∼32000 (p) covariates to analyze.

Linear Regression

Since $p >> n$, linear regression does not yield a good (or stable) result if applied directly. Hence, to control the bias-variance trade-off in linear regression, we examined different methods of merging OTU variables together. First, we examined the results when the OTUs are merged together while maintaining the taxonomic classification. For example, at the genus level, all the species belonging to the same genus were merged together into their individual genus. Doing this across the data set reduced the number of OTU variables from ∼32000 (individual species) to 832 (individual genuses). This can logically be extended further up the hierarchy tree, to the family, order, class, or phylum levels. However, this type of merging did not yield any useful results. The average RMSE using 832 covariates, for example, was 9.63 which gives us no useful insights into the BMI.
To deal with this, a more controlled merging strategy was implemented. Columns were merged together which belonged to the following set:

$$\{\textbf{col} \;:\; \mathcal{N}(\text{col}) < i \;\; \textbf{OR} \;\; \max(\text{col}) < \text{mean}\}$$

where the function $\mathcal{N}(\cdot)$ gives the number of non-zero values in a column and mean is the mean all non-zero values in the data set.

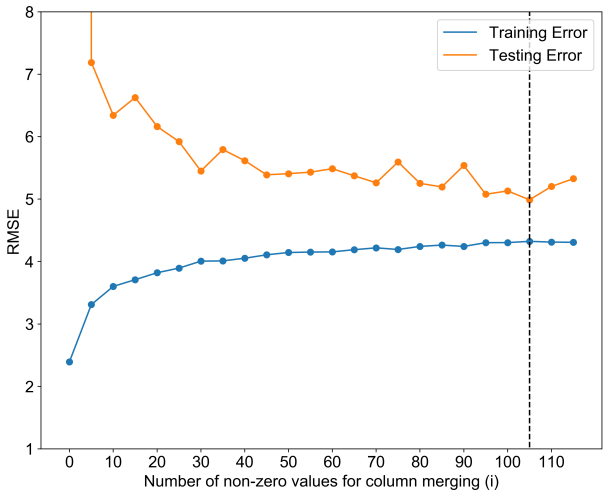
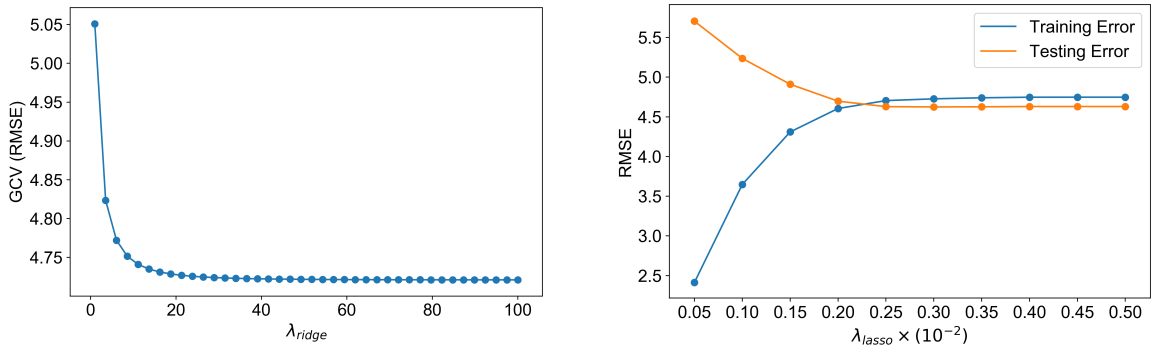


Figure 10: Linear Regression Results

The logic behind this formulation is to deal with sparsity and dimensionality together, in that if variables have only a few sample values, their contribution to the overall data composition is small. We varied the value of i in the above plot to observe the results depicted in Figure 10.
We see the plot resembles a typical bias-variance trade-off plot, with testing error decreasing up to a certain i and then increasing again. For these values of i from the above equation, we see that the number of covariates analyzed ranges from 5792 ($i = 0$) to 1212 ($i = 120$). The optimum is always found at about 1272 columns ($i = 105$), which we found to be the best case with linear regression, The best case testing RMSE for linear regression was found to be ~ 4.95.

Ridge and Lasso Regression

Since the data is high-dimensional, this is a classic case for the usefulness of penalized regression models. We examined both Ridge (L2 penalty) and Lasso (L1 penalty) regression models to uncover some interesting insights. To obtain the best value for parameter λ_{ridge}, we use generalized cross validation RMSE. To obtain the best parameter estimate for Lasso λ_{lasso}, we try different values on a grid and plot the corresponding training and testing errors. Both plots are shown in Figure 11.



(a) (b)

Figure 11: Parameter tuning for (a) Ridge and (b) Lasso Regression

From the plots in Figure 11, we see that the curves do not resemble a typical bias-variance trade-off curve. The cross-validation error decreases up to a certain value of the penalty ($\sim$ 20 for ridge and $\sim$3e-03 for Lasso) after which there is no change with increasing penalty. On further examination, we found that these arise as the penalty discards all variables, and only fits the intercept to the model i.e. predicting the mean of the training data gives the best error. A further discussion is provided in Section 6.
Lastly, we compare the three regression approaches with a 10-fold cross validation using a 90-10 split. The errors are shown in the boxplots in Figure 12.

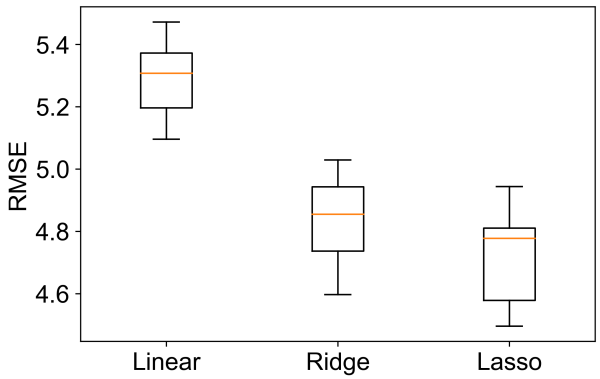


Figure 12: Comparison between 3 different regression models using 10-fold cross validation

6 Collaborator's Questions (Discussion)

Our study shows little to no power for predicting BMI, BMI category and alcohol consumption habit was possessed by microbiome. One could argue the methods taken to deal with sparsity lost most of the information in the data. However, different data-processing strategies were evaluated but not documented due to project length limitation. It was shown that relationship between the three target variables and microbiome were weak regardless of the data processing method taken. In fact, many strategies aimed to improve the model were examined and implemented, but claims were made against the predictive ability of the data set on these target variables. For example, when modeling BMI category, SMOTE oversampling was used to produce more synthesized samples in minority classes. As suggested in the original paper, it was used in conjunction with random undersampling the dominating class. Various ratios of SMOTE and random undersampling were examined, however, synthesized data from the minority classes actually impaired the model's overall predictive ability. Less power to distinguish between 'Normal' and 'Underweight' is observed when oversampling 'Underweight' samples. The problem was further simplified in search of improvement. Merging 'Obese' class and 'Overweight' class, merging 'Obese', 'Overweight' and 'Underweight' class, discarding 'Underweight' class were all examined with and without oversampling and undersampling, but none of them produced better results. Therefore, claims were made against the predictive ability of the data set on categorical BMI. Similar scenarios were encountered when modeling BMI and alcohol_frequency, thus, lack of predictive ability in the data set were claimed.

For the regression case, penalized regression models are seen to outperform linear regression, which is expected in a high-dimensional setting where the number of covariates ($\sim$32000) is higher than the number of valid samples ($\sim$8000). However, within penalized regression models, an interesting result was observed. As seen in Figure 11, the testing error decreases up to a certain penalty but does not increase again. This is because the penalty is the highest possible, after which only the intercept (training data mean) is used to predict the BMI. The class imbalance in BMIs puts most samples in the Normal BMI (18-24) range, which leads to poor prediction of underweight and overweight samples. One way to tackle this issue is to use inverse weighted sampling. Further improvement can be obtained by trying out different methods for pre-processing the compositional and sparse data, such that a better signal can be observed.

Additional concern was raised against data integrity, as mentioned in the documentation, the data consists of samples collected from over 10,000 citizen-scientists, together with an open research network, across many countries. Though testimony backing data quality for citizen science, and self-selected cohort shipping samples was made, it's likely that the quantification of microbiome was not unified enough for the scope of this project. In the unsupervised learning section, two naturally formed clusters was often observed in principle components and clr-transformed microbiome variables. The existence of such clusters, and their cross-referenced similarity in demographic and health-related variables drew our attention onto data integrity. A closer inspection into the clusters and why such clusters were formed is one possible future research direction.

References

[1] Emanuele Rinninella, Pauline Raoul, Marco Cintoni, Francesco Franceschi, Giacinto Abele Donato Miggiano, Antonio Gasbarrini, and Maria Cristina Mele. What is the healthy gut microbiota composition? a changing ecosystem across age, environment, diet, and diseases. *Microorganisms*, 7(1):14, 2019.

[2] Jasmohan S Bajaj. Alcohol, liver disease and the gut microbiota. *Nature Reviews Gastroenterology & Hepatology*, 16(4):235–246, 2019.

[3] Alexandre Almeida, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, 2019.

[4] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2(1):73–94, 2015.

[5] Gregory B Gloor and Gregor Reid. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian journal of microbiology*, 62(8):692–703, 2016.

[6] Minhoo Kim and Bérénice A Benayoun. The microbiome: an emerging key player in aging and longevity. *Translational medicine of aging*, 2020.

[7] Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing rna-seq, 16s rrna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1):15, 2014.

[8] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.