Multi-label Disease Prediction Based on Gut Microbiome

Amando Jimenez
ajimenez@ucsd.edu

Emerson Chao
emchao@ucsd.edu

Renaldy Herlim
rherlim@ucsd.edu

Abstract

In this study, we will be exploring the gut microbiome of Latin American immigrants to determine what factors of their gut microbiome affect metabolic diseases. The goal of our project is to determine what metabolic diseases/disorders an individual has based on their gut microbiome and other supporting information on the individual. To achieve our goal, we will utilize a binary-relevance model for each disease type to see if we can predict if an individual has that specified disease according to their microbiome factors.

1. Introduction

Metabolic diseases afflict millions of people in the US, with diseases such as diabetes, high blood pressure, and obesity affecting Latinos and other ethnic minority groups at a significantly higher rate. One factor contributing to this multifaceted disparity is the fact that minority groups are severely underrepresented in clinical research and health studies, resulting in a continued lack of insight and solutions for these groups. Our project seeks to further metabolic disease research and expand representation in such fields by studying how the gut microbiomes of Hispanic populations are correlated with the prevalence of certain diseases. As the field of gut microbiome research has grown, scientists have discovered links between gut microbiomes and diseases like diabetes and obesity, as well as differences in gut microbiomes by race and ethnicity (Duvallet, et al.; Ross, et al.). Machine learning solutions that can accurately determine these links and scale across diverse gut microbial populations could provide useful general and specific solutions for different diseases and different groups. As such, the main goal of our project is to use the Study of Latinos (SOL) gut microbiome dataset to implement and train machine learning models that can determine an individual's metabolic disorders based on their gut microbiome.

2. Literature review

We will be primarily be building off of this study Gut microbiome composition in the Hispanic Community Health Study/Study of Laftinos is shaped by geographic relocation, environmental factors, and obesity - PubMed (nih.gov) by Kaplan et al. in which they used the SOL dataset and focused on observing the microbiome of immigrants that migrated at a young age vs later stage in life and found key differences in microbiome and obesity rates. Some of the faults in the study are that there are no visualizations on model validation and model performance scores. Although we will be focusing on metabolic disorders we hope to improve on the paper by performing cross validation techniques to more robustly test our model performances and provide clarity to how well our model performs.

For dimensional analysis of the feature table we utilized the study Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data. by Armstrong et al. that discusses the key techniques of analyzing high-dimensional microbiome data and how dimensionality reduction techniques can be applied. This study shines light on the special characteristics of microbiome data and provides literature on previous successful studies, together with the pitfalls and common mistakes to avoid. This paper provided a solid foundation in understanding microbiome data for beginners in the field, in particular it introduced us to beta-diversity analysis, PCoA, and UMAP which are methods we will apply to our dataset.

To further study how dimensionality reduction techniques are used in the microbiome field, we looked into the article Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data by Armstrong, George et al. This article compares approaches of dimensionality reduction in particular between PCoA and UMAP and the benefits and limitations of each method. The visualizations in this article helped us understand the powers of each method, and gave us useful recommendations for the selection of parameters. The article also showed us how we can utilize Qiime2 to perform analyses of microbiome data which is helpful in our learning of Qiime2.

The literature Human gut microbiome viewed across age and geography by Yatsunenko, Tanya et al. is a study performed on a dataset that is similar to ours, in which there are many geographic and demographic variables associated with the data. This study provided some useful analyses techniques, in particular, we took interest in the Unifrac distance plots and the statistical analyses techniques used in the study to pick out distinctive features in their dataset, and we plan to incorporate these techniques into our analyses of the SOL dataset.

3. Data Description

For the purposes of this project, we will be using the gut microbiome dataset collected by the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). The HCHS/SOL is a

population-based study of about 16,000 self-identifying Hispanic/Latino adults from the ages of 18-74 who were selected from randomly sampled census block areas within Chicago, IL; Miami, FL; Bronx, NY; San Diego, CA (Lavange, et al.). The study focused on gathering information on the health status of the US Latino/Hispanic population in order to address the lack of research on minorities.

For our project, we will be using 1835 self-collected stool samples that have gone through 16S rRNA gene amplicon sequencing and other bioinformatics preprocessing which is explained more thoroughly in (Kaplan, et al.). The dataset is available on Qiita (https://qiita.ucsd.edu/) ID: 11666.

After collecting the data from Qiita, we were left with a single metadata table and a single feature table. The feature table columns were OTU ID's that referenced the metadata and the rows were genome sequences; each column in the table represents the counts of each sequence in a particular sample. The metadata contained information about the participants socioeconomic conditions, their country of origin, and their medical history which includes information about whether or not a given sample had a certain metabolic disease. A subset of the diseases will be selected as our classification targets and a few metadata columns such as age and gender will be used as features in addition to the entire feature table of genome sequences.

As we will be focusing on investigating the relationship between the gut microbiome and metabolic disorders, both tables will be extremely essential for building machine learning models to predict whether or not someone has a given metabolic disease. Additionally, since our main objective is to detect the presence, not severity, of diseases, the metabolic disease data is encoded in a binary format which will lend itself to our purposes.

After dropping the samples with missing values and identifying the samples that existed in both the feature table and metadata, we found 1750 samples that we could use for our analysis, which is sufficient for building our binary classification models.

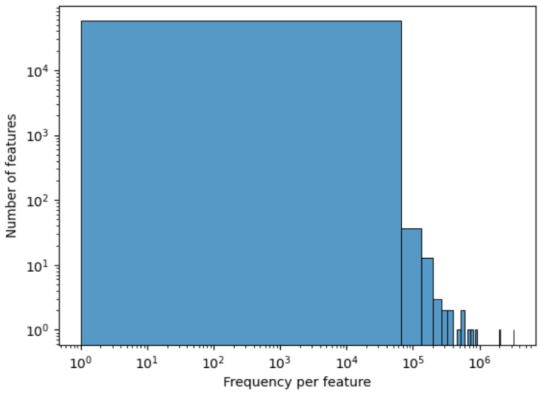
We will then do some preprocessing on the amount of features (columns) of our feature table. We plotted the count/frequency of samples that each feature appears on in our feature table dataset. By looking at the 'Frequency per Feature Analysis' plot:

Frequency per Feature Analysis

Frequency per feature

	Frequency
Minimum frequency	1.0
1st quartile	1.0
Median frequency	2.0
3rd quartile	3.0
Maximum frequency	3,303,337.0
Mean frequency	415.8038818329519

Frequency per feature detail (csv | html)



we see that alot of our features have very low sample frequencies, which means that theres a lot of irrelevant features/columns that have little importance and could be dropped in our data.

Table summary

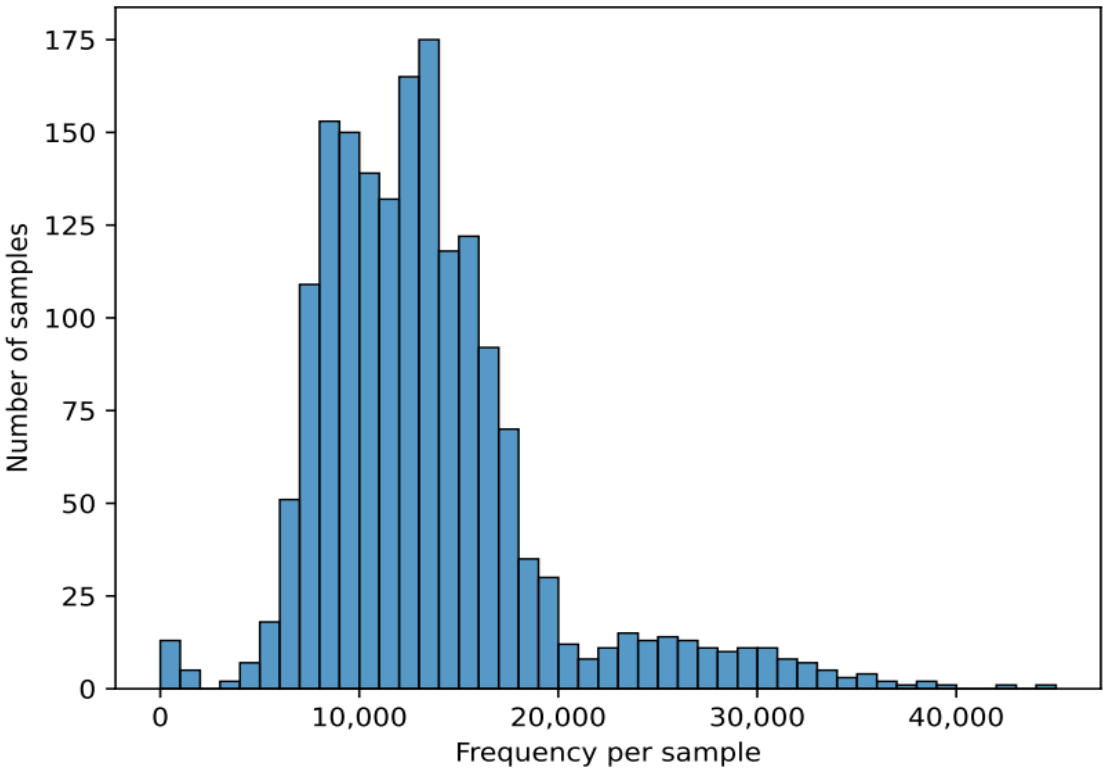
Metric	Sample
Number of samples	1,750
Number of features	57,241
Total frequency	23,801,030

Frequency per sample

	Frequency
Minimum frequency	4.0
1st quartile	9,675.75
Median frequency	12,749.0
3rd quartile	15,808.75
Maximum frequency	45,156.0
Mean frequency	13,600.588571428572

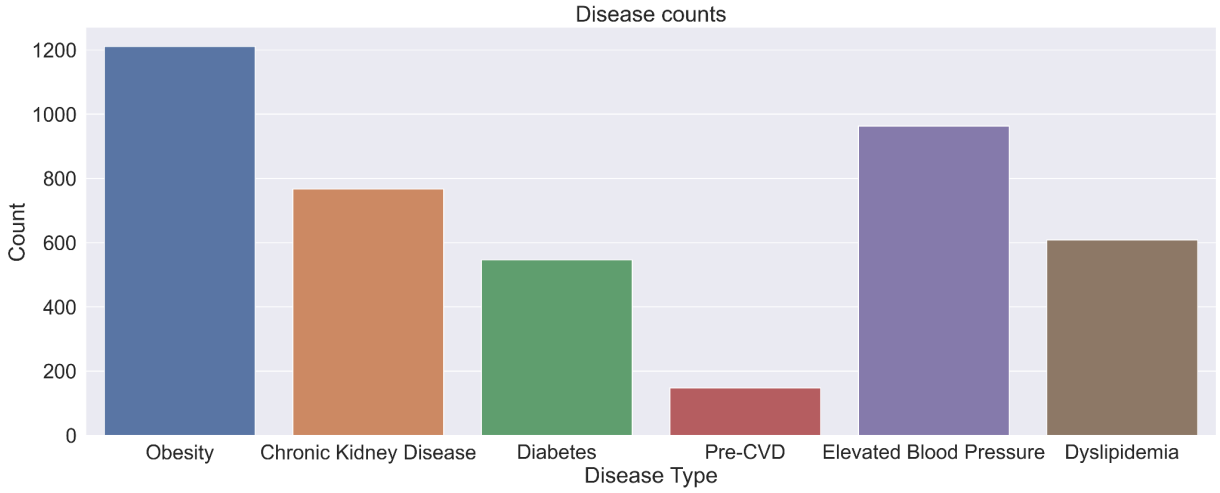
Summary of the Frequencies Table

The above summary is calculated with the 1750 samples that were found post initial preprocessing of the feature table.

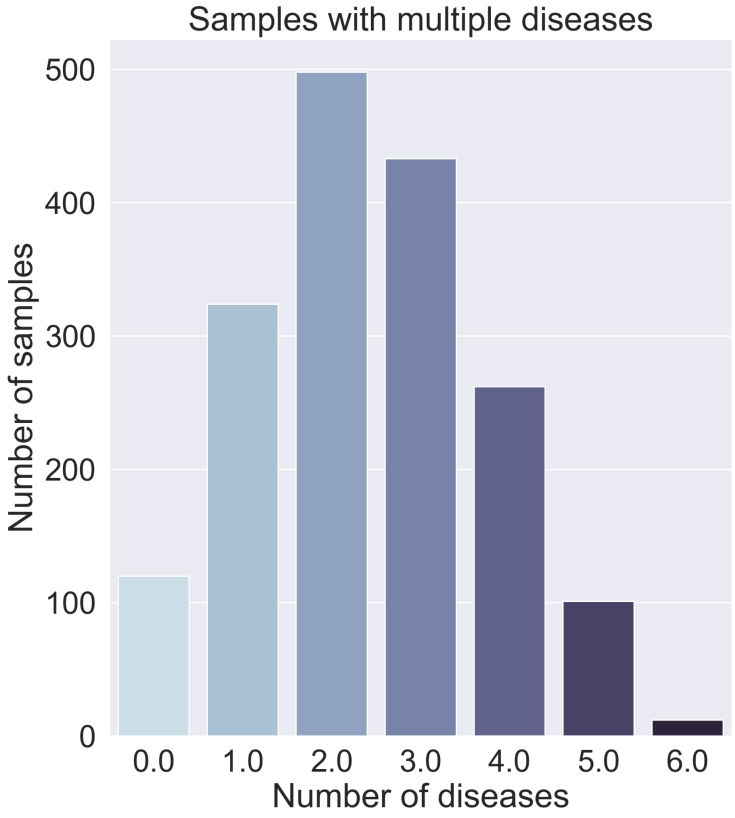


Histogram of Number of Samples & Frequency Counts Per Sample

Within the 1750 samples, we found that there were many samples that contained the metabolic diseases that we want to investigate such as obesity, elevated blood pressure, diabetes and dyslipidemia. We found the following disease counts present in the metadata. As we can see in the graph below, pre-cvd had by far the lowest presence within our dataset, indicating that there is a large class imbalance within the pre-cvd column. This class imbalance will need to be addressed in our pre-processing in order to prevent our model from overfitting.



After further data exploration, we found that many samples have varying amounts and varying combinations of diseases, as can be seen in the graph below, which means a binary classifier would not work. Though we considered using a regression model to predict the total number of diseases per sample, we ultimately decided on exploring multi-label approaches that could predict the specific diseases each sample has.



4. Methods

4.1 Data Preparation

The first step of our pipeline was to process the data and to prepare it for machine learning development. For this step we primarily used Pandas to manipulate the metadata table. The metadata was subsetted for only the diseases we wanted to study: obesity, diabetes, dyslipidemia, chronic kidney disease (ckd), pre-cardiovascular disease (pre-cvd), and elevated blood pressure. Next, the control samples (marked by "BLANK") and missing and null values were dropped. Finally, the ckd and diabetes columns were mapped to binary values and all metadata values were type casted to integers. Originally, the ckd and diabetes columns had a range of values denoting the severity of the disease; mapping these values to binary keeps these columns uniform with the other disease columns and simplifies our classification task by reducing our values to 0, not having a specific disease, and 1, having a specific disease. The feature table was also cleaned by setting a threshold value for minimum number of reads per sequence and filtering out any sequences that did not exceed that threshold. Our threshold of 100,000 reads was heuristically determined by balancing the amount of information retained and the speed at which our models could be trained and run.

From the summary visualizations and statistics that was explored in the Data Description section, we see that most features only appear in less than 3 samples, therefore we are going to drop the features that appear less than 3 times in order to reduce noise. We managed to reduce 57,241 features to 3,988 features, but only lost about 400,000 reads (23.8 million to 23.4 million total frequency/reads), this will significantly speed up our downstream analysis process as well as reduce the size and noise of our dataset.

Table summary

Metric	Sample
Number of samples	1,750
Number of features	3,988
Total frequency	23,492,118

As mentioned in the data description section, the pre-cvd column was imbalanced, there were many more samples without pre-cvd than samples with pre-cvd, 1602 compared to 148 samples. As such we had to ensure that the classes within the pre-cvd column were balanced in order to prevent our model from overfitting to the majority class. We did this by undersampling and randomly selecting 148 negative samples. Although, we are losing data by undersampling we do not risk overfitting our model on the majority class.

4.2 Features Analysis

A major goal of this project was to utilize qiime2 to perform our analysis in order to demonstrate its capabilities with the hopes of increasing adoption of the package in the bioinformatics field. Qiime2 is extremely useful for microbiome analyses and provides a lot of functionality for exploring microbiome diversity, performing dimensional analysis and creating machine learning models, which is extremely useful for our project.

In order to understand our frequency table data and extract useful information to relate to our metadata, we performed a dimensional analysis on our frequency feature table. We used the Qiime2 core_metrics method to perform a multitude of dimensionality analysis techniques. The first thing that we did is rarefy our table. Rarefying is a process of subsampling from all samples so that the sum of frequencies in each sample is equal to the specified *sampling depth* and any samples with frequency sum less than the sampling depth will not be included in the resulting table. This process normalizes the data and assures that we don't have outliers in our dataset that can skew our dimensionality analysis and ML model. We still need to do further research on the most optimal way to rarefy our table so that we do not make improper assumptions of our data that could lead to invalid results of our project, for example, we do not want to lose too much feature information in our data and drop too many/little samples that could affect our findings

and credibility since different dimensionality reduction techniques yield visually and statistically very different results on the same data.

With the rarified table, we extracted the distance matrices (with Jaccard, Bray-Curtis, weighted and unweighted unifrac distance matrices) and created PCoA matrices to reduce the high-dimensional feature table to 2 and 3 dimensions. Distance matrices are useful as inputs for further statistical analysis that we will perform to identify clustering or patterns in our frequencies table and prove the significance of said findings. For example, we will be able to demonstrate a separation between groups by performing beta-diversity analysis of the distance/dissimilarity matrices and follow with a statistical validation with the PERMANOVA (Permutational multivariate analysis of variance) test.

We also created plots of the PCoA embedding matrix using Qiime2's Emperor plotting library to visualize our data. We plan on using the new feature table in the low-dimension as an input to our machine learning model. We also used another dimension reduction technique Uniform Manifold Approximation & Projection (UMAP) on the data which has been proven useful from previous successful literature (Armstrong, et al.). Specifically, we are using a supervised-UMAP approach where we give the UMAP algorithm our disease type label as targets to perform supvervised clustering in the reduced 2D space. To do this supervised method, we would need to filter our dataset to only include samples with a single disease type, and in doing so reduced our dataset down to about 300 samples.

4.3 Modeling

Because our task involves determining which specific diseases a sample has, we took a multi-label approach to building our machine learning model. A basic binary relevance model was implemented as a baseline model. An independent binary classifier was created, trained, and tuned for each class (disease), and each sample is then inputted into each classifier to see if that sample has that particular disease or not.

To create our classifiers we used qiime2's classify-samples method, which implements scikit-learn's gradient boosting classifier with log loss, as the base model for each classifier. One of the biggest advantages of using qiime2 over scikit-learn, is that it provides a lot more information about the model such as model performance visualizations, feature importance and probability tables by simply executing the classify-samples function. Additionally, it automatically splits the data into training and test sets and utilizes stratified k-fold cross validation to test the model.

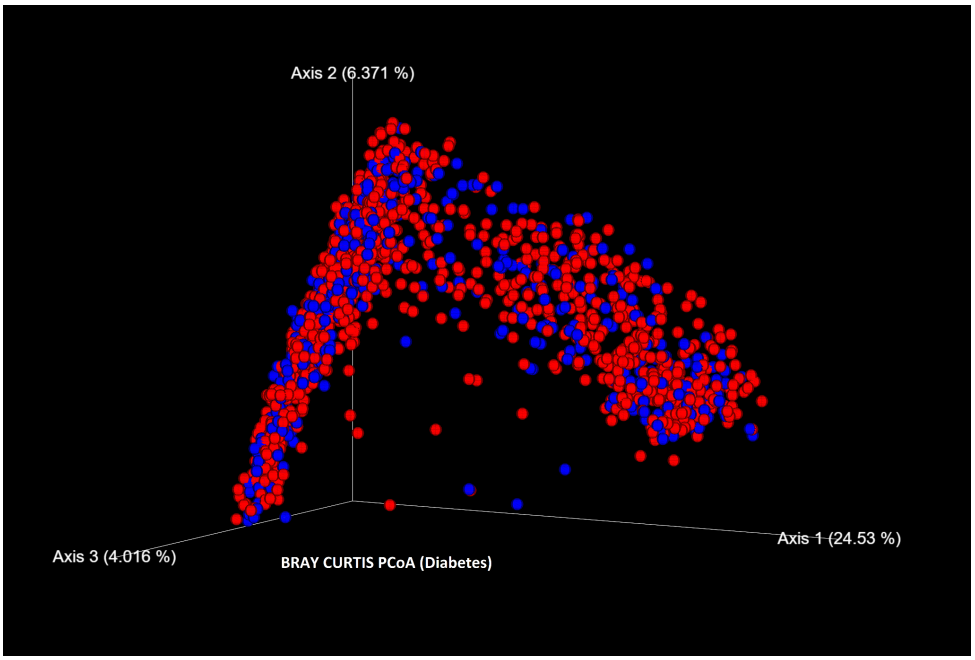
While this binary relevance model is quick to implement and train, it assumes independence between all of our classes, which is likely not true in our case. Studies have shown diseases like obesity and diabetes are correlated, so finding a model that can account for such relations would help the accuracy and scalability of our model to more and other diseases.

We used the following params for the classify-samples method: {test_size = 0.3, cv = 10, random_state = 100}. The gradient boosting classifiers used the default sklearn parameters with a random_state=100.

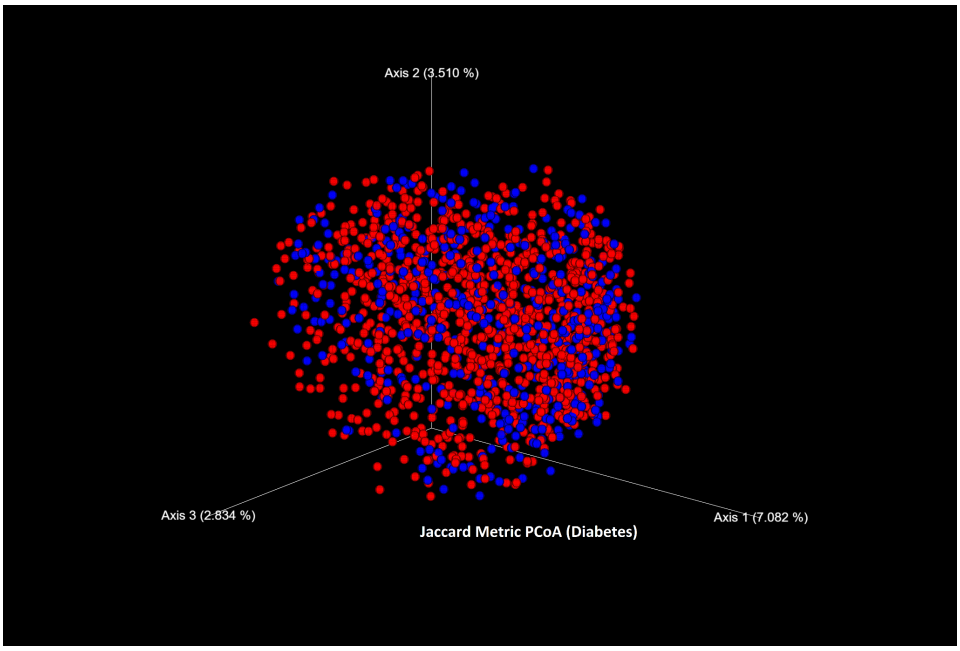
5. Results

5.1 Dimensionality Reduction Results

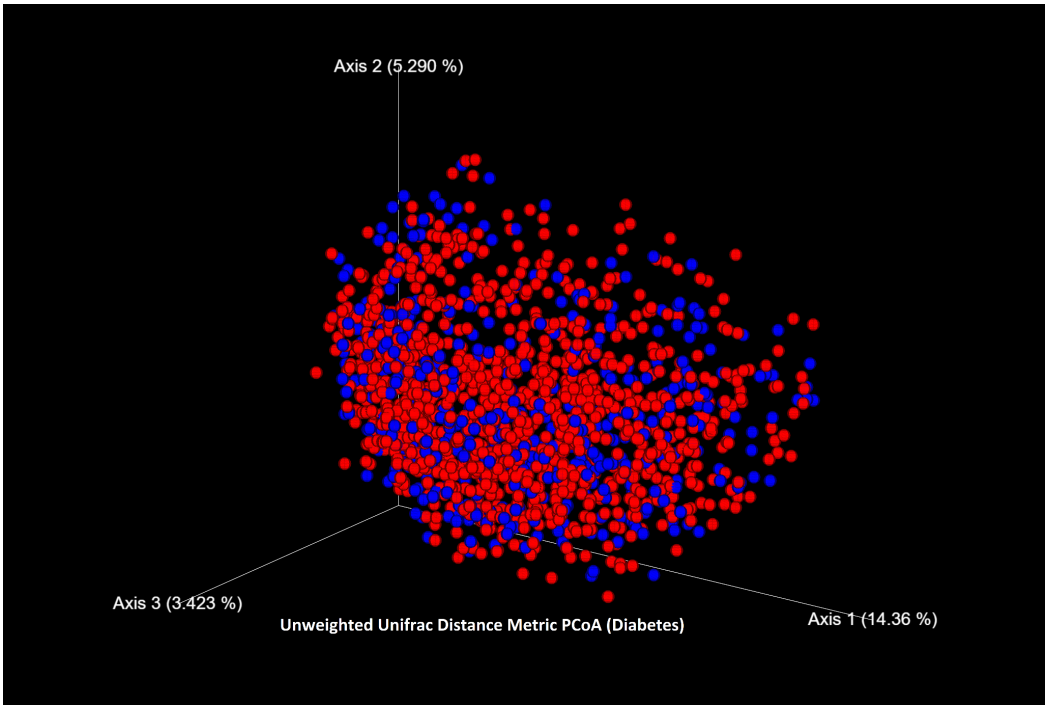
For our first attempt at visualizations, we plotted PCoA plots in the 3D space using a variety of distance metrics for comparison, but found that clustering is difficult to see. We believe this is caused because of varying amounts and varying combinations of diseases that individuals/samples have, as previously mentioned in our metadata analysis.



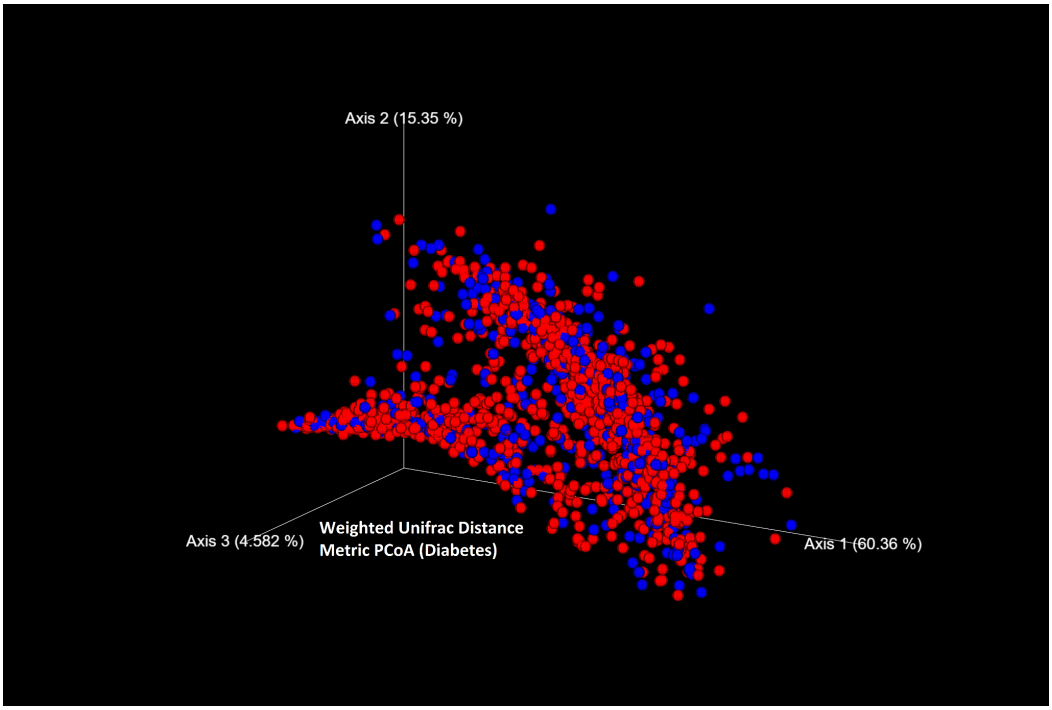
Bray Curtis PCoA (Diabetes)



Jaccard PCoA (Diabetes)

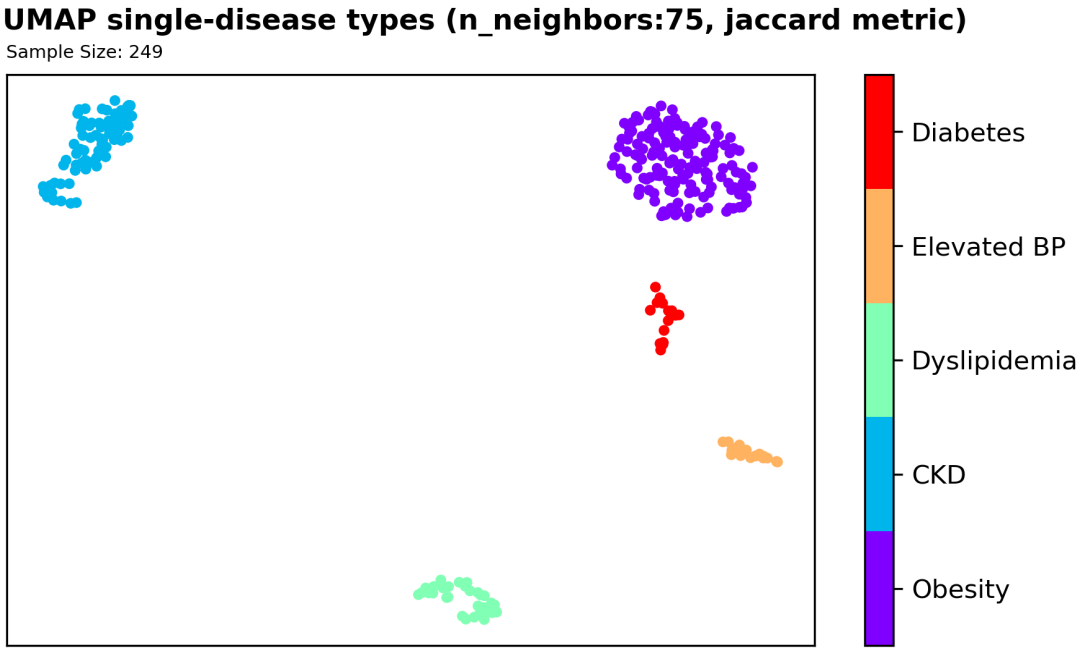


Unweighted Unifrac Distance PCoA (Diabetes)



Weighted Unifrac Distance PCoA (Diabetes)

We found significant clustering results in the 2D space using a Jaccard distance metric, which signifies that further exploration of using this method could be useful for our classification problem. However, since we have very limited sample size, and an imbalanced disease types among samples, it is hard to utilize our finding as information/features for our classification model.



Supervised UMAP Single-Disease Types

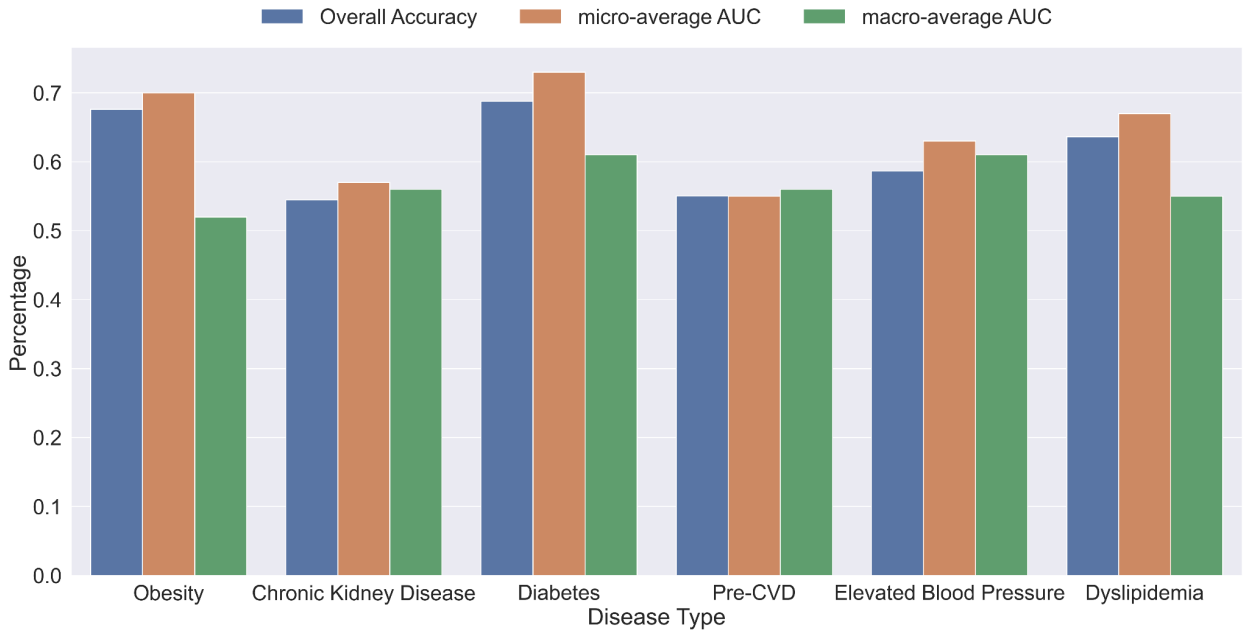
5.2 Permanova Test

PERMANOVA Test Results using Unweighted and Weighted UniFrac Distance Matrices

Disease	Unweighted UniFrac p-value	Weighted UniFrac p-value
Obesity	0.023	0.124
Chronic Kidney Disease	0.001	0.001
Diabetes	0.001	0.021
Pre-Cvd (n=296; Balanced Classes)	0.07	0.549
Elevated BP	0.005	0.035
Dyslipidemia	0.001	0.2

The table above demonstrates the p-values of the PERMANOVA tests computed for each disease type using unweighted and weighted UniFrac distance matrices. As we can see in the table above, the majority of PERMANOVA tests computed had a p-value less than 0.05 which means that the null hypothesis is rejected in those cases. This means that the observed differences between those who had a given disease and those who did not, is likely not due to chance alone and indicates that another factor/variable is leading to the differences between both groups (those with and those without a given disease). The only PERMANOVA tests that had a p-value higher than .05 were both pre-cvd tests and the obesity test using the weighted UniFrac distance matrix. We suspect that pre-cvd tests were likely affected by the class balancing, which resulted in a loss of information.

5.3 Binary Relevance Model Results



The plot above shows the performance of each individual gradient boosting binary classification model within our binary relevance model. The plot contains the overall accuracy, the macro-average, and micro-average area under the curve (AUC's) for the model's receiver operating characteristic (ROC) curves. The overall accuracy measured the proportion of samples the model correctly predicted diseases for. The AUCs essentially measure our classifier's ability to discriminate between positive and negative samples by measuring the true positive rate when given a decision threshold. The macro-average AUC is calculated by calculating the AUC for each class and then averaging the AUC overall the classes. The micro-average AUC is calculated by averaging across each sample, which essentially means that it uses every sample prediction to calculate the overall AUC.

As we can see by the results, the average overall accuracy of the models was about 60.9% and all the AUC's are above 50% which means that every model performed better than random chance. Obesity and diabetes were the best performing models with an accuracy of about 67% and 69%, micro-average of about 70% and 73%, and macro-average of about 51% and 61%, respectively. The big difference between micro and macro-average AUC's of the obesity model indicates that the model performs better on the majority class than on the minority class, which indicates that we may need to do more pre-processing to improve the results.

In addition to more pre-processing, it's possible that different choices of classifiers or models could be implemented to improve our task performance. As seen in the PERMANOVA test, there are significant and obvious differences between the metabolic diseases when looking at samples with only one disease. However, because the majority of our samples have multiple diseases, it is reasonable to assume the microbiome features would look very differently, requiring models or special inputs that recognize the how having multiple diseases can impact the features and vice versa.

6. Discussion

Overall, we were unable to yield the desired results for our classification model. Due to the difficulty of multi-label classification tasks, the complexity of microbiome and disease research, and the lack of substantial data, the results yielded by our model has a lot of room for improvement. The low accuracy of our model suggests a different multi-label machine learning model, different choice of classifier model, or further data pre-processing may be required.

We attempted multiple methods to improve our model performance such as using unweighted and weighted unifrac PCoA results in our model training, but unfortunately it did not improve upon our baseline performance. We suspect the biggest limitation to our project is the lack of data and class imbalances within certain disease columns. Since we have 6 diseases we are studying, we would need a minimum of 64 samples to have one of each possible combination of diseases present. If we were to treat each combination as its own class and wanted 1,000 samples per class to train on, we would need a minimum of 64,000 unique samples.

Other models were also implemented to attempt to improve our classification performance. One model was a modified multi-label K-nearest neighbors model. This model looks at the distance of a sample's features to the features of other samples, and predicts labels by looking at the weighted average of the closest 10 samples, or neighbors. Though this model takes longer than the binary relevance model to train, there is only one classifier involved and can easily be scaled with new data points and new diseases. However, this model performed even worse than the binary relevance model, likely due to the lack of data and large variance for each class. Other distance metrics may need to be tested to see improvement in this model.

Another tested model was a Multilayer Perceptron neural net with three fully connected layers and a ReLU (rectified linear unit) activation layer using binary cross-entropy loss and optimized with Adam optimizer (modified stochastic gradient descent). This is a significantly more complex model than the knn model or binary relevance model, seeking to recognize any patterns from scratch. Using a neural net enables actual multilabel prediction; the many nodes in each fully connected layer allow for more nuanced weighting of different combinations of features, and the final layer can be customized to output however many binary labels required. Because of the complex nature of neural networks, they require enormous amounts of data, long training times, and lots of tuning. Ultimately, the neural net did not perform very well, likely due to our not having enough data.

While there remains other more complex models to test, a binary relevance model follows the current state of medical testing; there is a separate, specific test for a specific disease and sometimes multiple samples are required to conduct multiple tests. In this case, it makes sense to continue using a binary relevance model and optimize the results. However, models like the neural net that can weigh features and relationships between features against each other to recognize varying patterns could very well be the next step in diagnostic fields.

Even though our project did not yield our desired results, our project still contributes to the field of gut microbiome research by performing analysis with Python and commonly used Python packages, lowering the barrier of entry for other data scientists and equipping current researchers with new tools.

References

1. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications*, *8*(1), 1784. https://doi.org/10.1038/s41467-017-01973-8
2. Ross, M. C., Muzny, D. M., McCormick, J. B., Gibbs, R. A., Fisher-Hoch, S. P., & Petrosino, J. F. (2015). 16S gut community of the Cameron County Hispanic Cohort. *Microbiome*, *3*, 7. https://doi.org/10.1186/s40168-015-0072-y
3. Kaplan, R.C., Wang, Z., Usyk, M. *et al.* Gut microbiome composition in the Hispanic Community Health Study/Study of Latinos is shaped by geographic relocation, environmental factors, and obesity. *Genome Biol* **20**, 219 (2019). https://doi.org/10.1186/s13059-019-1831-z
4. Armstrong, G., Rahman, G., Martino, C., McDonald, D., Gonzalez, A., Mishne, G., & Knight, R. (2022). Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data. *Frontiers in bioinformatics*, *2*, 821861. https://doi.org/10.3389/fbinf.2022.821861
5. Armstrong, G., Martino, C., Rahman, G., Gonzalez, A., Vázquez-Baeza, Y., Mishne, G., & Knight, R. (2021). Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data. *mSystems*, *6*(5), e0069121. https://doi.org/10.1128/mSystems.00691-21
6. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J.,

Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., … Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222–227. https://doi.org/10.1038/nature11053

7. Lavange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J., Criqui, M. H., & Elder, J. P. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Annals of epidemiology*, *20*(8), 642–649. https://doi.org/10.1016/j.annepidem.2010.05.006