

The Establishment of a Tunable Data Pipeline for Investigation of the Association of Gut Microbiome Populations with Disease States

Thomas McIntyre

Frank Vasquez

Abstract

Our goal was to develop a data science pipeline for assessment of the gastrointestinal (gut) microbiota in adult patients with kidney disease. To this end, we analyzed the American Gut Project's open-source data collected from Qiita (Gonzalez et al., 2018). The data consisted of approximately 400 patients with chronic kidney disease and 800 patients with diabetes, which is the cause of nearly 50% of cases of advanced kidney disease in the US. All patients contributed stool samples allowing 16S rRNA profiling of the gut microbiota by the American Gut Project team, and answered questions by survey on their clinical characteristics. We used this 16S rRNA data set alongside clinical information obtained by survey to develop a workflow and preliminary understanding of gut microbiota changes in kidney disease. This paper details the techniques and findings from the American Gut Project data and makes recommendations for future research.

Introduction

This study is part of an ongoing research initiative to understand changes in the gut microbiota that may contribute to kidney disease that is led by Julia Scialla, a nephrologist and clinical investigator at UVA Health. The goal of this component of the project is to create an analytical pipeline that contains bioinformatics data processing, file conversion methods, exploratory data analysis examples, and machine learning examples to apply to this question utilizing Python. The modules contain simple machine learning methods, and also introduce an automatic machine learning (AutoML) package that fits complex models. The final product will present a modeling notebook where a user can interactively adjust the parameters of a random forest and examine the model metrics as well as the feature importance of the taxonomic level of choice, from as high a rank as phylum to as granular as genus. The analysis described in this paper primarily focuses on the genus level of the microbiome and explains how these tools can help guide future research by combining bioinformatics and data science methodologies to the question of the gut microbiome in chronic kidney disease (CKD).

Background

The American Gut Project was launched in November of 2012 as a collaborative effort between the Earth Microbiome Project (EMP) and the Human Food Project. The aim of this collaborative endeavor was to evaluate the composition of gut microbiota in a “self-selected citizen scientist cohort.” As of May of 2017, the American Gut Project includes more than 15,096 samples from 11,336 human participants (McDonald et al., 2018). The accumulation of this data represents nearly 500 million (with 48,599 unique) 16S rRNA V4 gene fragments.

The 16S rRNA dataset from The American Gut Project was used for this project. The use of 16S ribosomal rRNA sequences is founded on the inherent differences when compared against human rRNA. Specifically, there are small and large subunits of rRNA, and those small and large subunits can be further subdivided, as found in both eukaryotes (e.g. humans), and prokaryotes (e.g. bacteria) (Lakna, 2018). However, these ribosomal rRNA subunits are different in size and their constituent subunits respectively, when comparing prokaryotic rRNA to eukaryotic rRNA. The 16S rRNA is unique to only prokaryotes, and can therefore be used to filter for microbiota in the gut. In particular, the 16S rRNA gene has also demonstrated its utility in reconstructing phylogenetic information, as it has been noted to slowly develop over time, thereby allowing for species delineation among stool samples (Woose & Fox, 1977).

The 16S dataset contains survey information of all the people participating in the American Gut Project. The two main groups we will be investigating are people who reported by survey that they were clinically diagnosed with kidney disease and people who reported they were clinically diagnosed with diabetes. Samples from all other individuals will comprise the control group. The data are anonymized and publicly available and thus this was considered non-human subjects research by the UVA Health Sciences Research Institutional Review Board.

Qiita

The American Gut Project data that we are using for this study are available on Qiita (pronounced chee-tuh), which is an open source microbial study management platform that is hosted by the University of California San Diego (UCSD). In order to access the data on Qiita, a user must only create an account and verify their email, after which they have access to all uploaded studies, can upload their own data, or use some of the built in analysis tools that the platform has made available. The Project ID for the American Gut Project is 10317, and that ID

must be searched in order to access the samples that have been uploaded by the American Gut Project.




Public Studies

Expand for analysis (artifact count)	Title	Study ID	Samples	Preparation Data Types	Principal Investigator	Publications	Qiita EBI submission
44	Qiita Multidata Example	10768	23	Metabolomic, Metagenomic, 16S	Rob Knight		
10	Analysis of Bacterial Communities Present on Fermented Foods: From Raw Sample to Analyzed Multi-Omics Data in Less than 48 Hours 10317.000026993	10395	34	16S	Rob Knight		ERP015077
2120	American Gut Project GOLD Non-CVD 16S stool CVD 90908 10317.000001004	10317	35420	16S, 18S, Full Length Operon, Metagenomic, ITS	Rob Knight	29795809 , 10.1128/mSystems.00031-18	ERP012803

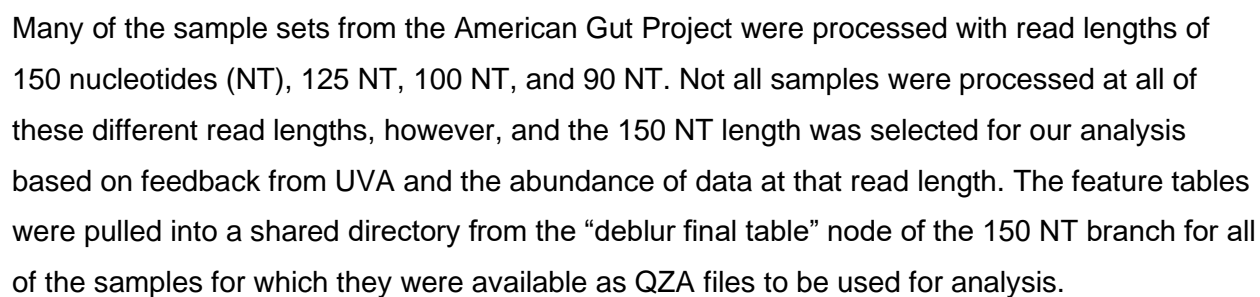
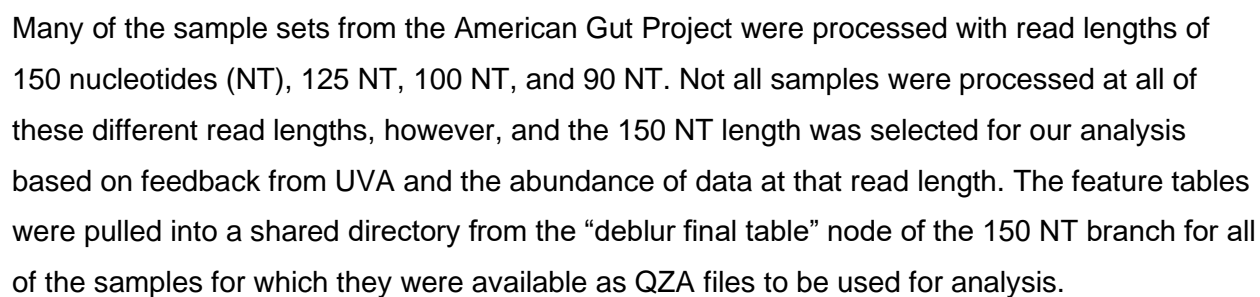
Showing 1 to 3 of 3 entries (filtered from 661 total entries) Previous 1 Next

While the raw sequencing data files from the American Gut Project are available via Qiita, we opted not to use them due to computational and time constraints. Additionally, availability of preprocessed feature tables for a significant number of the samples made starting with the raw sequencing files largely unnecessary. To remain consistent and produce an analysis/pipeline that will be transferable to the broader UVA initiative on the gut microbiome in CKD, we are using the 16S rRNA sequence data from the American Gut Project, which are available alongside a number of other data types on Qiita.

Data Types (click on the tabs)

 16S
 Full Length Operon
 18S
 ITS
 Metagenomic

The 16S data is divided up into sample batches based on when the American Gut Project data came into Qiita and how it was processed. Each sample batch contains a tree that displays the various ways that the raw data was processed, and allows access to the underlying data at that step via selection of a node of interest in the graph.



Due to compatibility issues, the web scraper designed for pulling down the feature tables was abandoned for a manual download approach. From the total of over 15,000 samples collected by the American Gut Project, 1340 were lost in the form of four sample batches that did not have 150 NT data. For the purposes of the analysis, the loss of these samples is acceptable, as we are still working with thousands more samples than the UVA pilot study contains, and these samples do not constitute a large portion of the populations of interest.

Data processing

As mentioned above, the 150NT processed feature tables were extracted from Qiita, and then saved onto a virtual environment provided by UVA. These files came in QZA format, which is a QIIME2 Artifacts data file. QIIME2 is a powerful and extensible microbiome analysis package (Bolyen et al., 2019). Two main R packages were utilized to get the survey information, sequence data, and taxonomy mappings into tabular format for analysis. The first package used was QIIME2R, which is regularly maintained and used for reading QIIME2 artifacts (Bisanz, 2018). The methods in this package were used to read the 150NT feature tables into R data frames. The structure of these files had the sample IDs as the columns and DNA sequences as rows. The counts within the cells of the table correspond to the number of times that sequence was present in a sample. The sample ID was a crucial connection between the sequencing data and the survey information. After using the QIIME2R package to read in the feature tables and join the survey information, the next step was to map the genetic sequences to their respective taxonomic values.

The R package DADA2 was utilized to map the kingdom, phylum, class, order, family, and genus to the DNA sequences. The DADA2 package implements a method to make taxonomy assignments based on exact matching between amplicon sequence variants (ASVs) and sequenced reference strains. Recent research suggests that this exact matching is the only appropriate way to assign species to 16S gene fragments (Callahan et al., 2016). DADA2 can utilize the Silva reference database for mapping which is regularly updated (Yilmaz et al., 2013). Using this reference file we were able to map the different taxonomy layers onto each DNA sequence which was utilized to visualize and examine different gut microbiome populations of chronic kidney disease and diabetes patients. Following the process laid out above, an R method was constructed to loop through all of the 150NT files to produce our final dataset for analysis.

Exploratory Data Analysis (EDA)

Genus was the primary taxonomic level of interest in our analysis, therefore higher taxonomic levels were not examined in detail. After filtering down to the more common sequences across the feature tables we were left with 810 unique genera. The Shannon Diversity Index (also referred to as alpha diversity) of all the participants in the study was calculated to begin understanding the makeup of the genera within the diabetes group against the group without diabetes as well as the kidney disease group against the group without kidney disease. The Shannon Diversity Index is a common way to better understand the diversity of species within a community. This paper utilizes this measurement to analyze the within-sample diversity of the microbiomes at the genus level. The Shannon Index also helps visualize the microbiome across different demographics from the survey information. The Shannon Diversity Index can be calculated using the formula below:

$$H = - \sum_{i=1}^s p_i \ln(p_i)$$

where

H = the Shannon index value

p_i = the proportion of individuals found in the i th species

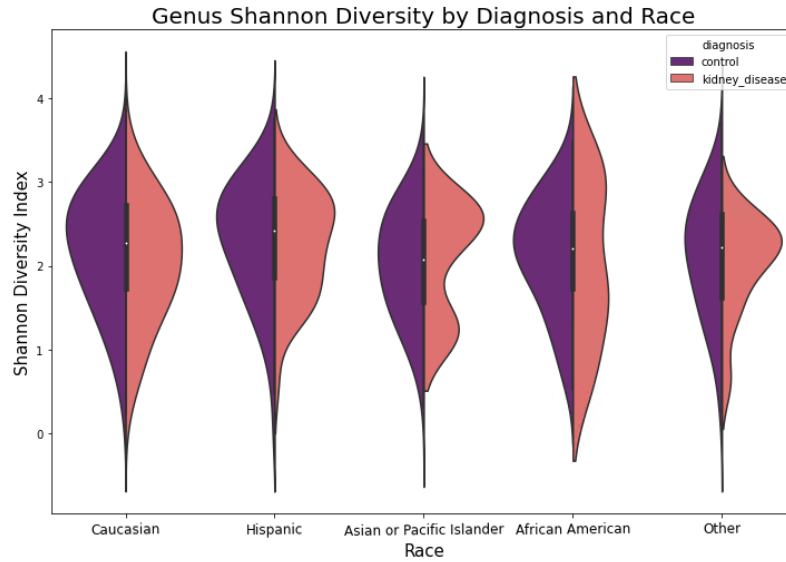
\ln = the natural logarithm

s = the number of species in the community

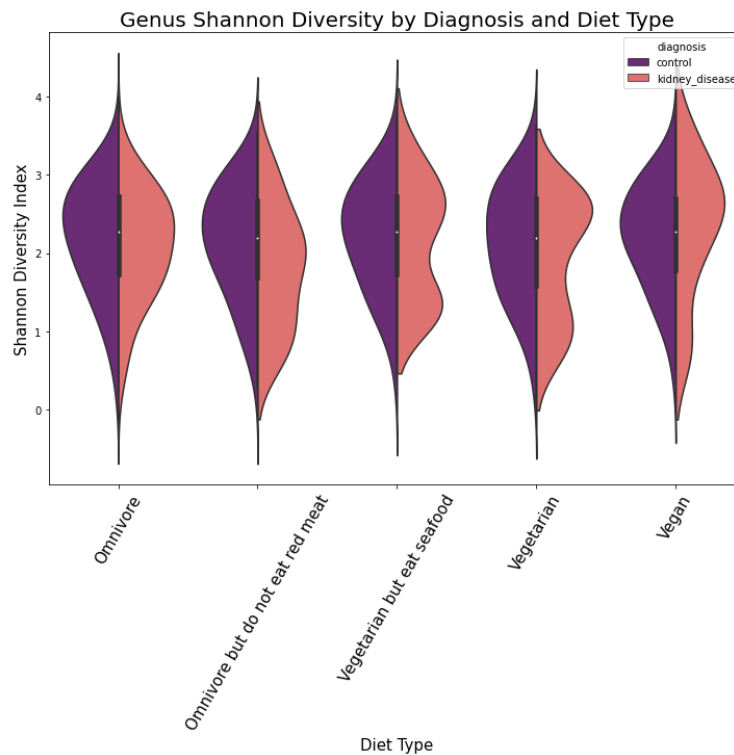
The full EDA notebook can be found [here](#), however this paper will highlight some of the interesting visualizations produced by the notebook.

Kidney Disease Demographics EDA

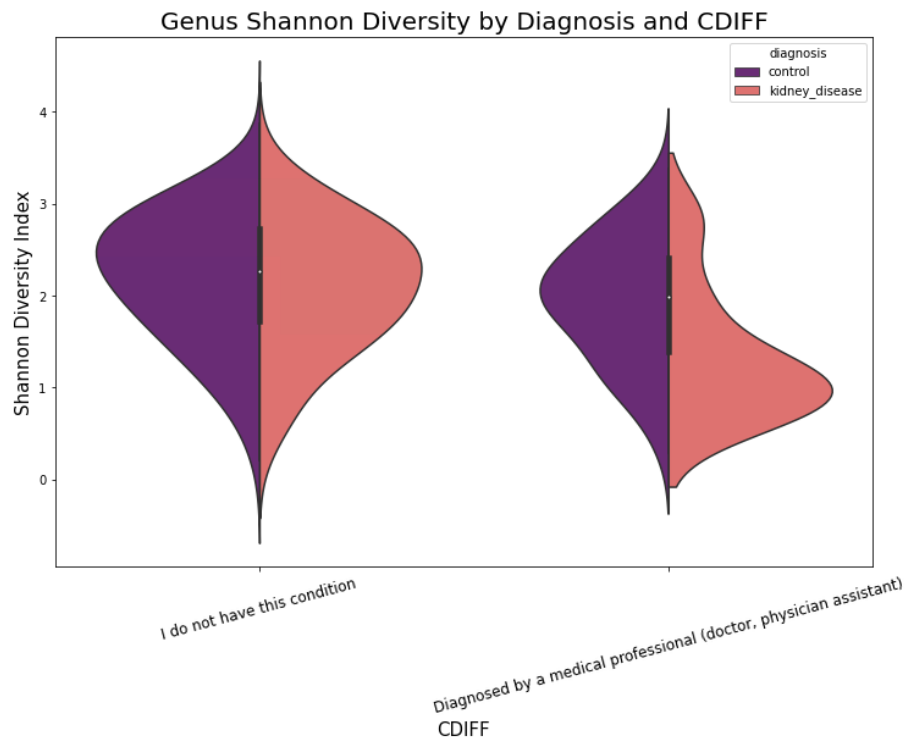
First, the microbiome diversity comparing patients with kidney disease to patients without kidney disease is examined. Three areas of the demographic survey questions of potential interest are presented as an example. These areas include the participants' self-reported race/ethnicity, diet type, and other health concerns. Viewing the plot below, the diversity of the participants' microbiome may differ by race/ethnicity. For instance, the pattern in the caucasian and hispanic subgroups are similar but differ from some of the other race groups qualitatively.



The following plot focuses on what the participants self-reported diets. Viewing the control groups (purple), the microbiome diversity appears similar across all the diet types, however when you look at the kidney disease groups (pink) there may be some different patterns across these diet types that could represent true differences or imprecision due to the smaller samples size of patients with CKD.

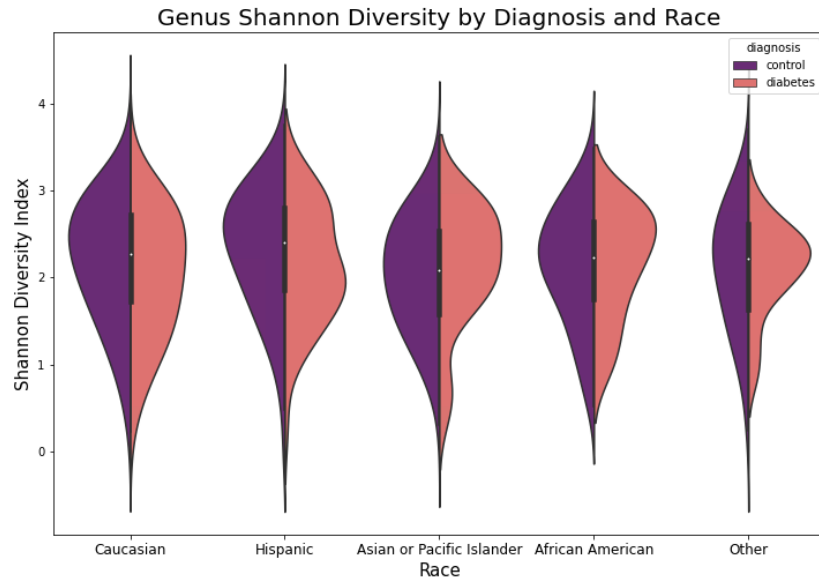


The final plot for the kidney disease visualizations shows that patients also diagnosed with *Clostridium difficile* infection (C.diff), appear to have a less diverse microbiome. For patients that do not have this infection the diversity appears similar between control and kidney disease. However, if they do have this condition, then the patient's diversity appears much lower. Again, sample size may be a factor in understanding these visualizations and these are suggestions for initial exploratory analyses and not inference generation.

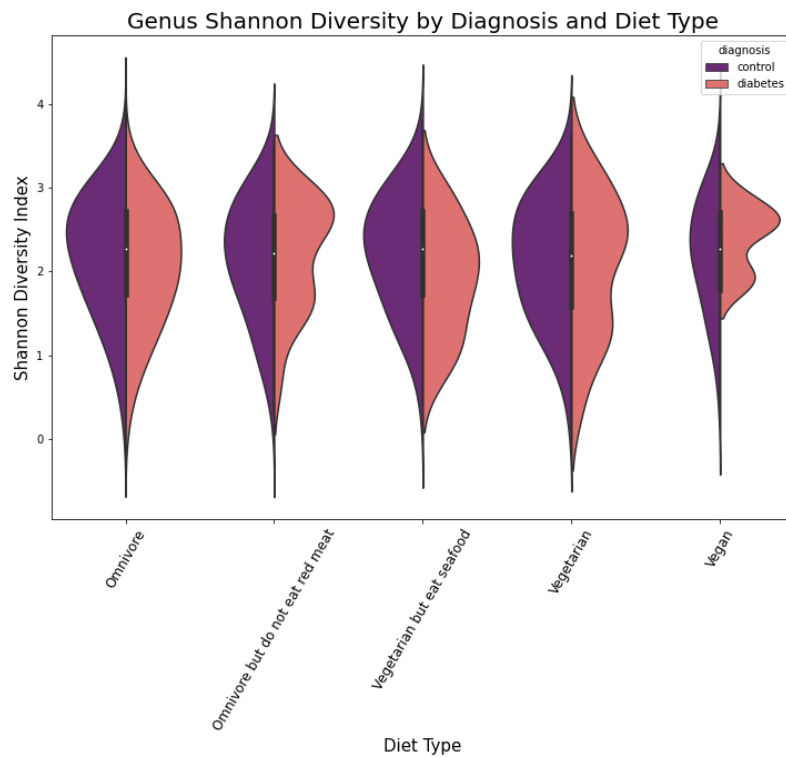


Diabetes Demographics EDA

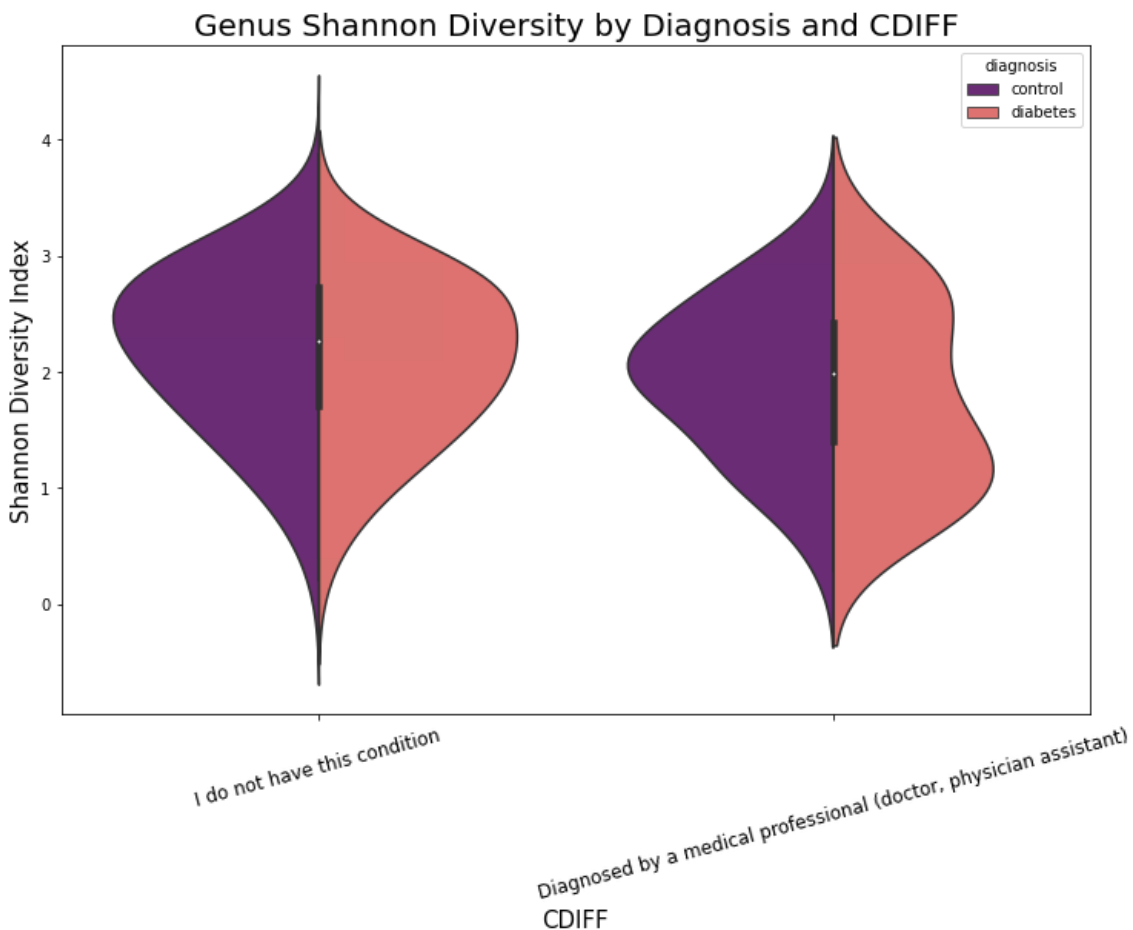
Due to the importance of diabetes as a cause of CKD we evaluated EDA stratified by self-reported diabetes status as well. We did not note a visually apparent difference between diabetic and non-diabetic patients by race/ethnicity.



The next plot below shows the Shannon diversity of the patients with and without diabetes separated out by their self-reported diet type. This violin plot has similar findings as the plot in the kidney disease section. As previously mentioned, the control group is distributed similarly across all of the diet types. However, when we look at the diabetes groups, there is a visible difference in the populations' microbiome diversity, which could be due in part to smaller sample size.



The last plot displayed is looking at the diabetes groups' diversity and whether or not having clostridium difficile infection has an impact on the microbiome. We did not note clear differences in the alpha diversity on this plot although the pattern in the diseased group (here diabetes) may be shifted lower than non-diseased, which was a possible trend also seen for kidney disease.



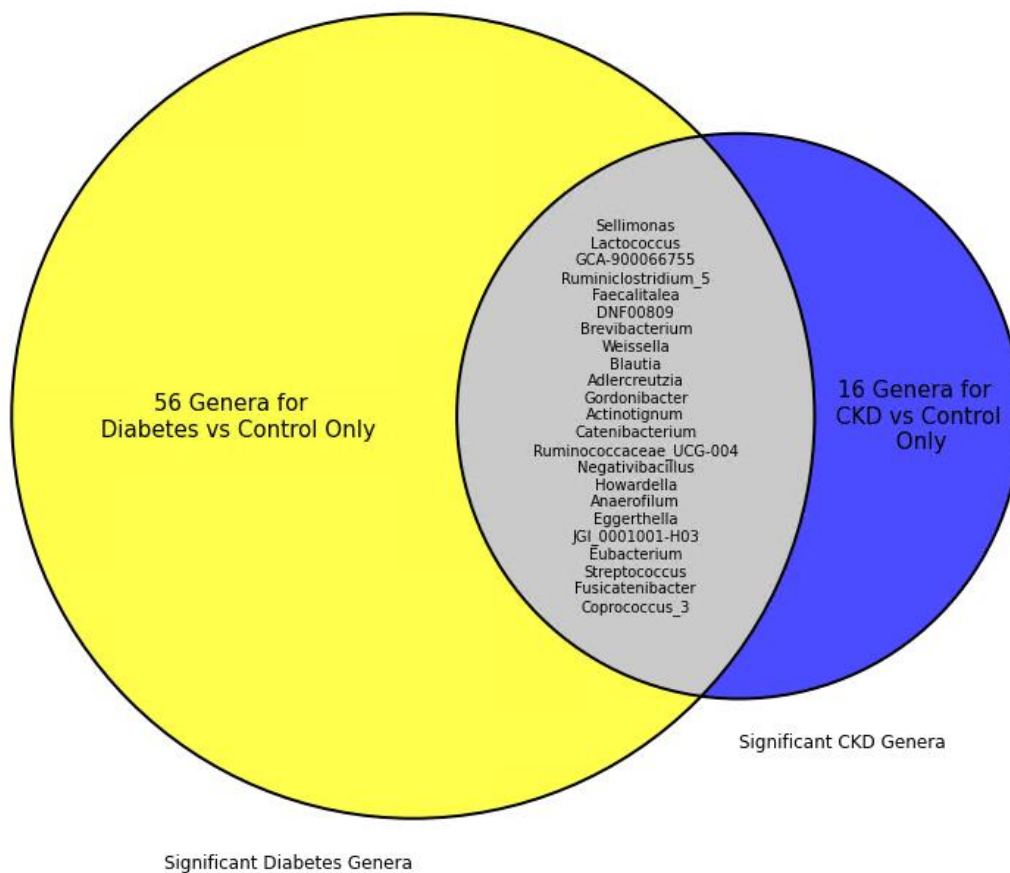
The previous six plots give a brief look into the exploratory data analysis performed for this project. The full analysis can be found on the project's Github page ([GitHub - fav3ba/bioinformatics-pipeline: A pipeline for visualizing and analyzing bioinformatics data.](#)). From this type of analysis our workflow can help select categorical variables of interest that should be considered in models developed to gain more insight into the microbiomes of people currently living with these diseases.

Population Differences EDA

For the 810 unique genera that were present in our samples, we wanted to examine which (if any) had significantly different population levels in the gut microbiome of each of our groups. A series of two-way t-tests were conducted to determine which genera exhibited significantly

different population levels for patients with vs. without kidney disease and patients with vs. without diabetes. There were 39 genera with different mean population sizes by kidney disease status. For the comparison by diabetes status, 79 genera showed significantly different populations. Of the genera that displayed different population values between the groups, 23 were different across diseased and non-diseased patients in both comparisons. By removing that intersection we are left with 16 genera whose populations differentiate patients with and without kidney disease but not diabetes, and 56 genera that differentiate those with and without diabetes but not kidney disease.

Relationship between Significant Genera vs. Control by Disease



At a more granular level, the diabetes cohort was divided into subpopulations of Type I diabetes and Type II diabetes. These groups were then subject to the same series of two way t-tests as the kidney and full diabetes group. The Type I diabetes group had 23 genera that differed significantly from its control group, and the Type II diabetes group had 66 genera that differed significantly from its control group. There were 6 genera that appeared in the significant t-tests for both Type I and Type II diabetes. As we move onto the modeling part of our analysis, the t-

tests described above have given us a greatly reduced set of parameters that will allow us to train a classifier that can predict whether an individual is likely to have diabetes or kidney disease.

Using Gut Microbiota to Classify Disease Status - LDA

Two example models were run for each of the kidney disease and diabetes comparisons. The first example model was a linear discriminant analysis (LDA) model. An LDA model creates a linear separation in an n -dimensional space where n corresponds to the number of predictor features in the model. On either side of the separation boundary lies one of the target classes. This model is characterized by high bias and low variance, so there is a low probability of overtraining, but the model often does not fit well on complex data. There are a number of modifications that can be performed to an LDA model such as L1/L2 normalization, but a simple LDA model was constructed for the purposes of this pipeline and demonstration. The LDA model for kidney disease reported an accuracy of 56.9%, with sensitivity of 50.0% and specificity of 61.1%. The LDA model for diabetes reported an accuracy of 57.1% with sensitivity of 79.3% and specificity of 46.7%. *Accuracy* is defined as the total number of correct predictions over the total number of patients. *Sensitivity* is defined as the proportion of people who are predicted to have the disease compared to the number of all people with disease irrespective of their test result. *Specificity* is the proportion of healthy people who tested negative compared to the total number of healthy individuals irrespective of their test result. Based on the accuracy metrics, both LDA models are slightly better than guessing at predicting disease state based solely on bacterial populations in the gut microbiome. The LDA model for kidney disease exhibited a lower sensitivity and a higher specificity than the LDA model for diabetes, which means from a practical standpoint that the model could be better at ruling in kidney disease than ruling it out. In contrast the diabetes model could be better for screening.

Metric	LDA – Kidney	LDA - Diabetes
Accuracy	56.9%	57.1%
Sensitivity	50.0%	79.3%
Specificity	61.1%	46.7%

Using Gut Microbiota to Classify Disease Status - Tree-based Pipeline Optimization Tool (TPOT)

The next modeling approaches gave an example of AutoML modules in Python. The goal of using AutoML packages was to see if a complex model with greater variance could overcome the shortcomings of the simple LDA model. The specific package used for this analysis was TPOT, which uses a tree-based structure to represent a model pipeline for predictive modeling problems. It utilizes different data preparation, modeling algorithms, and model hyperparameters to attempt to optimize certain predictive metrics (Le et al., 2019). TPOT utilizes many of the machine learning tools and techniques supplied within the popular Python ML package scikit-learn. After running the TPOT model fitting optimizer on the training dataset, the package exports the top cross-validated machine learning pipeline for the respective problem to a “.py” formatted file. TPOT was used to produce a pipeline to fit a classifier for both of the kidney disease and diabetes comparisons. Beginning with the kidney disease classifier, the algorithm ran for 100+ iterations and the top cross-validation accuracy was 60.4%. The model produced was a scikit-learn pipeline consisting of a min max scalar transformation, followed by an L1 normalization of the features, and ended with a stack ensemble model of two gradient boosted classifiers. This model was much more complex than the simple LDA model, however it only achieved a hold-out accuracy of 55.1%. This model misclassified one extra observation in comparison to the LDA model above. This builds on the previous evidence of the difficulty of separating patients with and without kidney disease based on the dataset.

The diabetes vs. control TPOT classification model also produced a complex modeling approach to this problem. The pipeline consisted of stacking 3 separate classifiers (Multinomial Naive Bayes, XGBoost, and Stochastic Gradient Descent) before passing the parameters into a final XGBoost classifier. This model involved ensembling multiple complex algorithms into a single classification method. This model led to a cross-validation accuracy of 62.9%, and a hold-out accuracy of 59.3%. This training and hold-out accuracy outperformed the LDA model for diabetes, however the improvement was not substantial.

Metric	TPOT - Kidney	TPOT - Diabetes
Accuracy	55.1%	59.3%
Sensitivity	47.6%	68.6%
Specificity	59.5%	47.5%

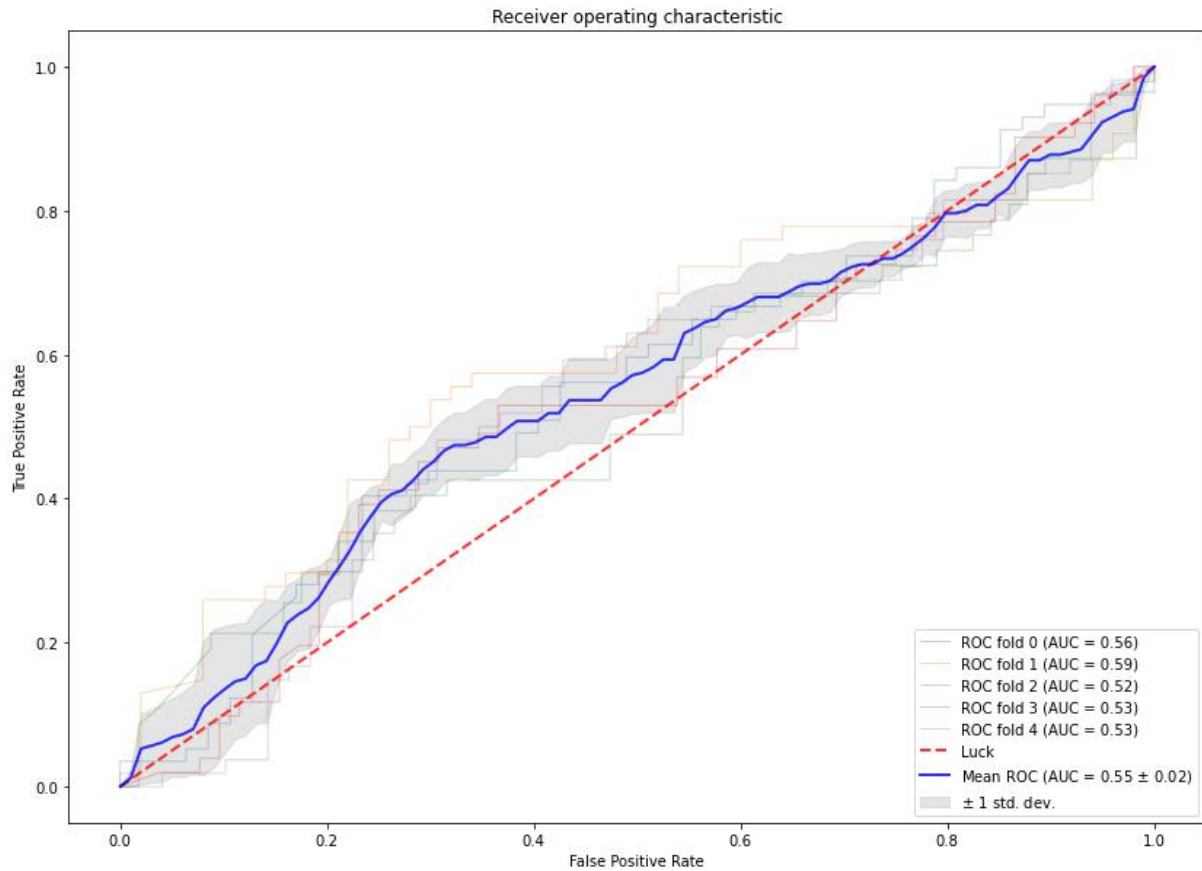
The performance of the complex models in this section performed essentially on par with the simple LDA approach. This result helps illustrate that this data presents a difficult problem to accurately classify just based on the genus counts. The goal of this project, however, is not to build the best possible classifier, but rather to gain insights on the microbiome makeup of patients with kidney disease and diabetes. Therefore, the following section will introduce random forest models, which allow for easy insight into the feature importance.

Identifying Top Features of the Gut Microbiome by Disease Status - Random Forest

At a high-level, a Random Forest is a model which is composed of many decision trees. A decision tree has a hierarchical tree-like structure that consists of different nodes where decisions are made. A Random Forest is a collection of many decision trees, which uses the classifications of each tree within the forest to come up with the final decision for that instance. The structure of this model is very useful when trying to determine the most important features that are leading to the specific classifications. The modeling notebook that corresponds to this report has a section for the user to interactively change the parameters of random forest models and run the fitting algorithm utilizing K-fold cross-validation. The main goal of this section is to investigate the feature importance of the models, thus the max_depth in these examples is set to 1 and the n_estimators is set to 1,000. This forces each decision tree to make the classification based on a single genus, while also having the model produce a large number of trees to come to the final decision. This will help increase the interpretation of the feature importances since all the trees are based on a single genus.

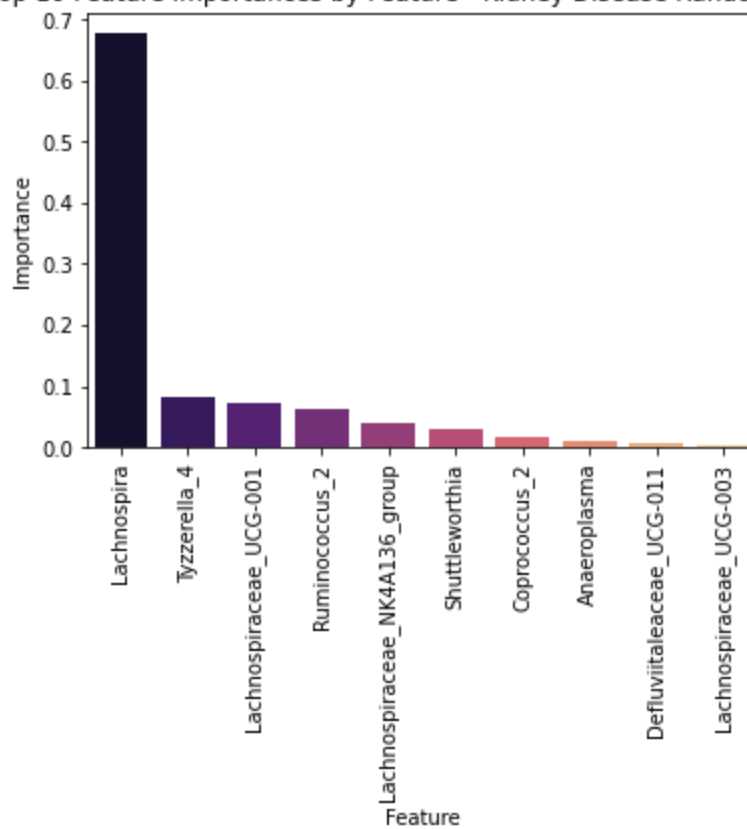
Kidney Disease Random Forest

The chart below displays the different receiver operating characteristic (ROC) plots for the 5 folds; testing sets for kidney disease vs. control.



The random forest with tree depth equal to one approach led to results slightly better than guessing across all folds. The training AUC for each fold was approximately 0.63. Thus, there was not significant overfitting in the cross-validated model. This model generated the following plot showing the 10 most significant genera for classifying kidney disease in this instance.

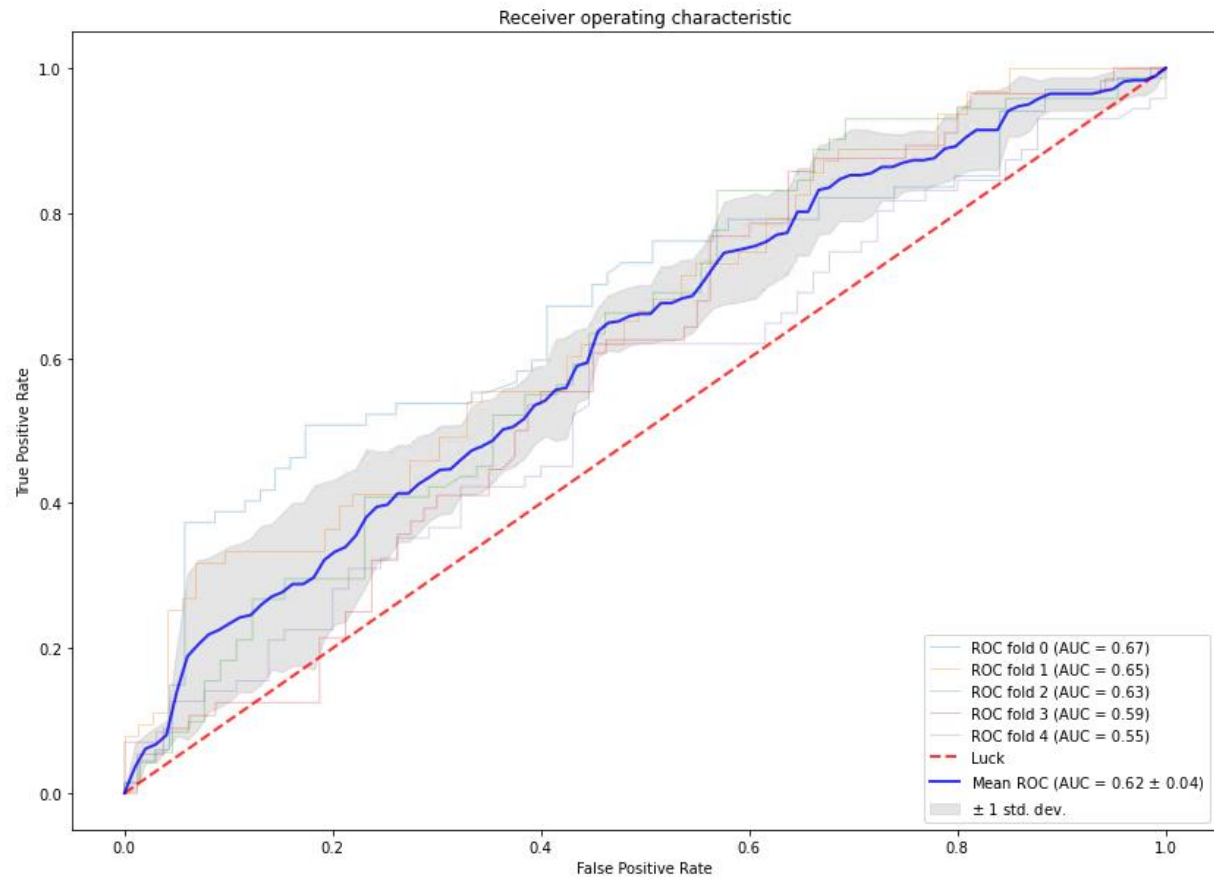
Top 10 Feature Importances by Feature - Kidney Disease Random Forest



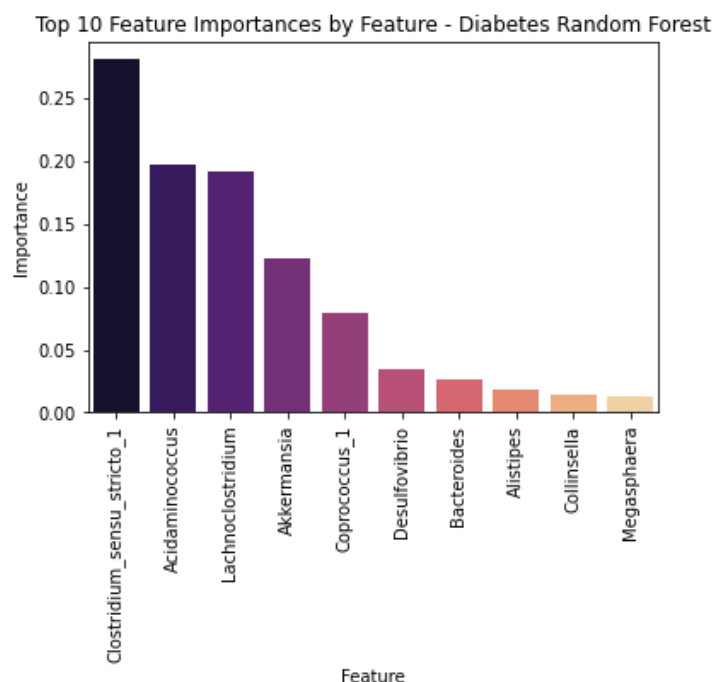
The plot quickly indicates that the genus, *Lachnospira*, is a top candidate as a genera that differs in gut microbiome between patients with and without kidney disease. There are several other genera that demonstrate some importance. However with the tree depth of one approach, *Lachnospira* dominates the decision making of the random forest.

Diabetes Random Forest

The same approach was utilized when investigating the diabetes vs. control population. The chart below shows the ROC plots for the different folds.



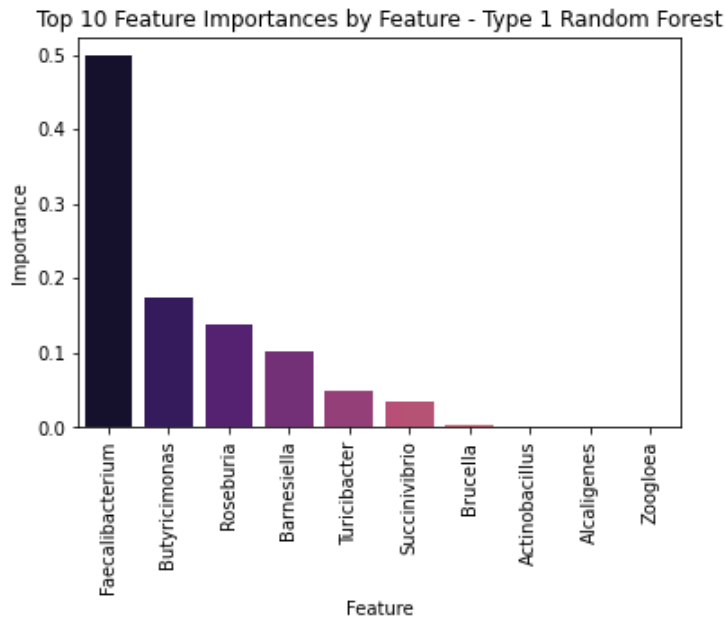
The random forest for diabetes showed more performance lift away from randomly guessing, which was in line with the prior models results. The training AUC for each fold was approximately 0.68. Thus, there was not significant overfitting in the cross-validated model. This model generated the following plot showing the 10 most significant genera for classifying diabetes. This feature importance plot shows that more than one genus was often important in making the final classification if the patient had diabetes.



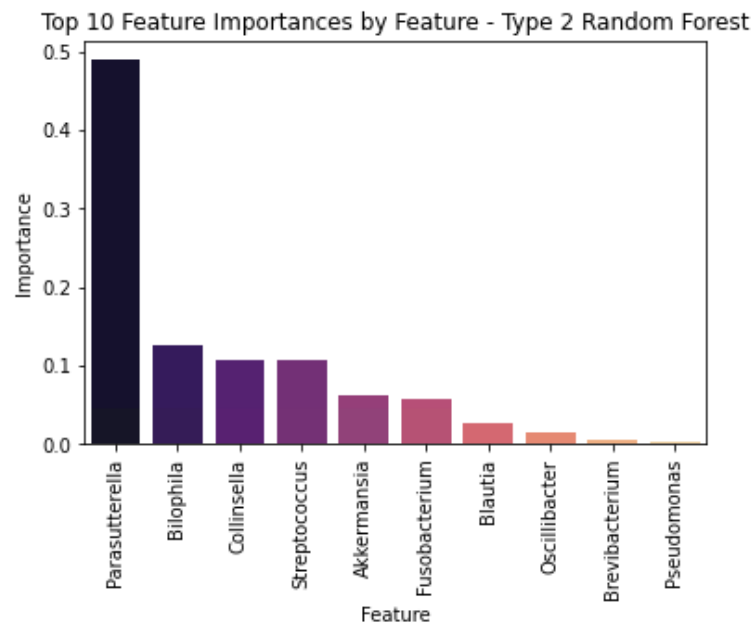
To gain a more holistic view of the diabetes landscape, while also demonstrating other use cases of this modeling approach, the following section separates out Type 1 and Type 2 diabetes and investigates the leading genera in classifying the respective diseases.

Type 1 vs Type 2 Diabetes Case Study

The same approach for the two previous models was used in calculating the feature importances within Type 1 and Type 2 diabetes patients. The models utilized the features discussed in the EDA portion of this paper, not including the intersection between Type 1 and Type 2 diabetes. There were only 52 observations specifically labeled Type 1 diabetes in the dataset, thus the results were relatively volatile. However, the feature importances for Type 1 can be seen in the plot below.



The leading feature in predicting Type 1 diabetes in this random forest was the genus Faecalibacterium. Interestingly, this genus is not present in the top 10 features of the combined diabetes random forest. There were 242 observations specifically labeled Type 2, which led to less volatile results for the following feature importance plot.



When separating specifically Type 2 diabetes, it is apparent that there are different leading indicators of having the disease. This section shows a potential use case of narrowing down to specific populations based on categorical demographic survey answers. It is intuitive that Type 1 and Type 2 diabetes patients likely have different microbiome compositions, however by filtering down the data sets utilizing different categorical variables a biologist could use this

modeling approach to gain initial insight into the microbiome makeup. Example use cases of this tool can be edited and used within the modeling notebook to help guide future research.

Future Work and Use of Pipeline

The notebooks created for this project serve as an example bioinformatics pipeline using R and Python packages. The notebooks detail a process that begins with data mining from the open source website Qiita, and then follow through with data aggregation, preprocessing, exploratory analysis, and finally construction of multiple machine learning models and extraction of feature importance. The primary purpose of this pipeline is not so much the predictive power of its output models as is, but rather to serve as both a tutorial and an easily modifiable set of steps that can be altered by a user for a variety of bioinformatics purposes. The entire pipeline can be used at once, or each of the individual notebooks may be taken separately based on the user's needs.

While the ingestion notebook is specific to data that comes from Qiita in QZA files, this notebook serves as an example of how to apply bioinformatic specific R libraries to extract data into easily readable CSV files. The preprocessing notebook details how one can combine extracted bioinformatics data (from the first notebook) with metadata and other human-readable features that can be used for further analysis. It also details how one can combine all data into a single data frame that can be used for visualization and modeling. A user can utilize the second notebook as an example of how to clean data for future analysis, and would additionally be able to append new data as it comes in to fortify and validate results of previous analyses. The third notebook provides a tutorial on a variety of visualizations to examine differences between samples and among varying demographics, as well as code that calculates the Shannon diversity index (also known as alpha diversity or within sample diversity) of each individual, thereby demonstrating how specific metrics can be added to a dataset if not already present. The final notebook provides examples and code for feature selection and classifications based machine learning models. It provides one example of a high bias, low variance model (LDA) and one example of a low bias, high variance model. The notebook is set up such that a user can easily modify the hyperparameters for the existing algorithms, but also so that any additional models from the scikit-learn library may be easily deployed. This notebook additionally provides metrics such as the confusion matrix and ROC curves that can be used to evaluate predictive algorithms. Lastly, and perhaps most importantly, this notebook provides a method of extracting feature importance from our predictive models for hypothesis generation.

From a biological perspective, the feature importance represents not just necessarily which features have a high correlation with a disease state, but which features are potential causes or consequences of disease that deserve further study. In terms of gut microbiome specifically, identifying the important genera in particular disease states could lead to discovery of previously unknown pathologies or diagnostic methods. For other medical applications, feature importance may have different implications, but it can always be used as a tool to identify future research areas.

Acknowledgements

The authors would like to thank Julia Scialla and Peter Gedeck for their mentorship through this project. The authors would also like to acknowledge their classmates and the support of Jennie Ma, Sue Haas, Binu Sharma, and Pankaj Kumar.

References

- Bisanz, J. E. (2018). QIIME2R. *qiime2R: Importing QIIME2 artifacts and associated data into R sessions*. computer software, Github. Retrieved from <https://github.com/jbisanz/qiime2R>.
- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., ... Caporaso, J. G. (2019). Reproducible, interactive, Scalable and Extensible Microbiome Data Science using QIIME 2. *Nature Biotechnology*, 37(8), 852–857. <https://doi.org/10.1038/s41587-019-0209-9>
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2016). Dada2: High-resolution sample inference from Illumina Amplicon Data. *Nature Methods*, 13(7), 581–583. <https://doi.org/10.1038/nmeth.3869>
- Gonzalez, A., Navas-Molina, J. A., Kosciulek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A. D., Orchanian, S. B., Sanders, J. G., Shorenstein, J., Holste, H., Petrus, S., Robbins-Pianka, A., Brislawn, C. J., Wang, M., Rideout, J. R., Bolyen, E., ... Knight, R. (2018). Qiita: Rapid, web-enabled microbiome

meta-analysis. *Nature Methods*, 15(10), 796–798. <https://doi.org/10.1038/s41592-018-0141-9>

Lakna. (2018, June 28). *Difference between prokaryotic and eukaryotic ribosomes - in tabular form*. Pediaa.Com. Retrieved August 10, 2022, from <https://pediaa.com/difference-between-prokaryotic-and-eukaryotic-ribosomes/>

Le, T. T., Fu, W., & Moore, J. H. (2019). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1), 250–256. <https://doi.org/10.1093/bioinformatics/btz470>

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., DeRight Goldasich, L., Dorrestein, P. C., Dunn, R. R., Fahimipour, A. K., Gaffney, J., Gilbert, J. A., Gogul, G., Green, J. L., Hugenholtz, P., ... Gunderson, B. (2018). American Gut: An open platform for citizen science microbiome research. *MSystems*, 3(3). <https://doi.org/10.1128/msystems.00031-18>

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11), 5088–5090. <https://doi.org/10.1073/pnas.74.11.5088>

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., & Glöckner, F. O. (2013). The Silva and “all-species living tree project (LTP)” taxonomic frameworks. *Nucleic Acids Research*, 42(D1). <https://doi.org/10.1093/nar/gkt1209>