# Using Human Microbial Data to Predict Geographical Location

Leshan Viswanath, Alex Rojas, and Dr. Harinder Singh, PhD

**J. Craig Venter Institute**

## Abstract

**Background:** With improvements in technologies, we are now able to better study and understand the complex human microbiome; these technologies have enabled us to predict the diverse physiological states based as a function of the role and makeup of microbial communities around the body. We were able to make predictions on composition and site of a microbiome based upon the geographical location, city, state, and country.

**Method:** We were able to make these predictions by building models based on concepts of machine learning within the python programming language. Using the sci-kit learn library, we were able to successfully build machine learning models that can make predictions based upon the data obtained from Forensic Microbiome Database. We focused primarily on Random Forest and applied it to our datasets.

**Result:** The Random Forest model that we built performed very well achieving an accuracy score of >80. Meaning that our model was very successful in predicting our data. However, we did run into various errors throughout the process of building the model. Through various troubleshooting mechanisms we were able to improve performance slightly.

**Conclusion:** After performing this test, we were able to identify the best specifications that can be added to our model to produce accurate predictions. Due to current restrictions on our computers we weren't able to test the models as extensively as we would have liked, but in the future, we plan on testing the models more extensively and improve predictions.

## Introduction

A microbiota is a community of microorganisms that live inside a host all over the body, and the composition of such a community can include viruses, bacteria, fungi, and other microbes. The "microbiome" refers to the collective genome of these species, and it can be read using sequencing technology such as illumina. In this project, we sought to find an efficient and accurate model that can make predictions of geographical location based upon data from the Forensic Microbiome Database on microbial data of composition and site. By the end of the project we hoped to have a clear understanding of not only which model was the best at making predictions, but as well as why the model was so successful. Building models to make predictions is important because it could allow us to detect the prevalence of certain microorganisms in regions. For instance, if a model predicts that a region, based upon a multitude of factors, is at high risk for a type of viral infection; we could take early action to prevent the virus from spreading and causing damage. Evidently, predictive modeling has a large application within the scientific space, as a result it is incredibly useful to research this topic.

## Methodology

Initially, we started off by researching the different methods such as Classification/Regression/Tree-based. We first looked into classification method, which works by assigning categories to a collection of data to make predictions. Next, we researched Regression models which work by examining the influence of one or more independent variables on a dependent variables. Tree based models which are used when dependent variable is continuous.

**Random Forest**

- The random forest is a model made up of many decision trees
- Random Forest uses a random sampling of training data points when building trees and random subsets of features considered when splitting nodes
- While training each tree learns from a random sample of the data
- These samples are *bootstrapped*, meaning the samples are drawn with replacement.

  Resulting in sometimes samples being used multiple times in a single tree

Below is an explanation of the process that we went through in writing our model


Below is an explanation of ways we quantified the performance of our models, and an explanation of how it functions

**F1 Score**

- F1 Score is the weighted average of Precision and Recall
- This score takes both false positives and false negatives into account

**Accuracy Score**

- It is the number of correct predictions made divided by the total number of predictions made
- This function computes subset accuracy

**Recall**

- Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes

**MCC Score**

- Known as Matthews correlation coefficient
- It takes into account true and false positives and negatives
- *n* is the total number of observations

**Precision Score**

- Precision is the ratio of correctly predicted positive observations to the total predicted positive observations

***Figure 1 shows the geographical locations of the stool data***




## Results

***Figure 2 shows the scores from classification***

| Location | F1 Score | Percision Score | Accuracy Score | Recall Score | MCC Score |
|---|---|---|---|---|---|
| City | 0.9545760754 | 0.9579678069 | 95.45760754 | 0.9545760754 | 0.9485401686 |
| State | 0.9603503031 | 0.9638902688 | 96.03503031 | 0.9603503031 | 0.955066402 |
| Country | 0.9727648927 | 0.9819479051 | 97.27648927 | 0.9727648927 | 0.9669218396 |

***Figure 3. Confusion Matrix***

| Country | Argentina | Australia | Bangladesh | Barbados | Belgium | Burkina Faso | Canada | Chile | China | Colombia | Gambia | Germany | Ghana | India | Indonesia | Ireland | Italy | Japan | Kenya | Malawi | Mali | Philippines | Russia | South Africa | Spain | Sweden | Taiwan | Tanzania | Thailand | USA | Uganda | United Kingdom |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Argentina | 36 | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | 2 | | |
| Australia | | 32 | | | | | | | 2 | | | | | | | 1 | | | | | | | | | | | | | | 4 | | |
| Bangladesh | | | 54 | | | | | | | | 5 | | | | | | | | | 3 | 6 | 1 | | | | | | | | 4 | | |
| Barbados | | | | 31 | | | | | | | | | | | | 2 | | | | | | | | | | | | | | 14 | | |
| Belgium | | | | | 45 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Burkina Faso | | | | | | 2 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Canada | | | | | | | 17 | | | | | | | | | 1 | | | | | | | | | | | | | | | | |
| Chile | | | | | | | | 79 | | | | | | | | | | | | | | | | | | | | | | 30 | | |
| China | | | | | | | | | 278 | | | | | | | 1 | | 1 | | 2 | | | | | | | | | | 2 | | |
| Colombia | | | | | | | | | | 19 | | | | | | 1 | | | | | | | | | | | | | | 7 | | |
| Gambia | | | | | | | | | | | 112 | | | | | 1 | | | 6 | | 1 | | | | | | | | | 5 | | |
| Germany | | | | | | | | | | | | 132 | | | | | | | | | | | | | | | | | | | | |
| Ghana | | | | | | | | | | | | 1 | 4 | | | | | | | | | | | | | | | | | | | |
| India | | | | | | | | | | | | | 92 | | 2 | | | | | 5 | | | | | | | | | | 5 | | |
| Indonesia | | | | | | | | | 2 | | | 1 | | 45 | | | | | | | | | | | | | | | 1 | 4 | | |
| Ireland | | | | | | | | | | | | | | | 966 | | 2 | | | | | | | | | | | | | 13 | | |
| Italy | | | | | | | | | | | | | | | | 99 | | | 2 | | | | | | | | | | | 9 | | |
| Japan | | | | | | | | | 1 | | | | | | | | 889 | | | | | | | | | | | | | 2 | | |
| Kenya | | | | | | | | | | | 12 | | | | | | | 102 | 5 | 2 | | | | | | | | | | 5 | | |
| Malawi | | | | | | | | | | | | | | | | | 1 | | 1035 | | | | | | | | | | | 2 | | |
| Mali | | | | | | | | | | | 5 | | | | | | | | 2 | 6 | 64 | | | | | | | | | 2 | | |
| Philippines | | | | | | | | | | | | | | | | 1 | | 1 | | | | 23 | | | | | | | | | | |
| Russia | | | | | | | | | | | | | | | | | | | | | | | 2 | | | | | | | | | |
| South Africa | | | | | | | | | | | | | | | | | | | | | | | | 29 | | | | | | | 2 | |
| Spain | | | | | | | | | | | | | | | | 5 | | | | | | | | | 103 | | | | | 15 | | |
| Sweden | | | | | | | | | | | | | | | | | | | | | | | | | | 3 | | | | 1 | | |
| Taiwan | | | | | | | | | 6 | | | | | | | | | | 7 | | | | | | | | 31 | | | 2 | | |
| Tanzania | | | | | | | | | | | | | | | | | | | | | | | | | | | | 20 | | | | |
| Thailand | | | | | | | | 1 | 6 | | | | | | 2 | 1 | | 1 | | | | | | | | | | | 82 | 25 | | |
| USA | | | | | | | | | | | | | | | | | | 2 | | 3 | | | | | | | | | | 2975 | | |
| Uganda | | | | | | | | | | | | | | | | | | 6 | | | | | | | | | | | | 1 | 43 | |
| United Kingdom | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 2664 |

***Figure 4. Country***

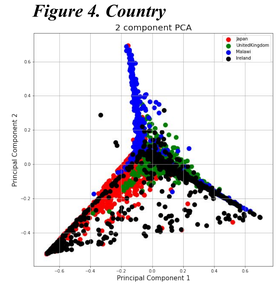



***Figure 5. State***

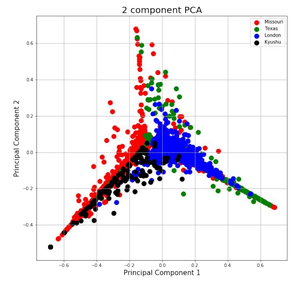
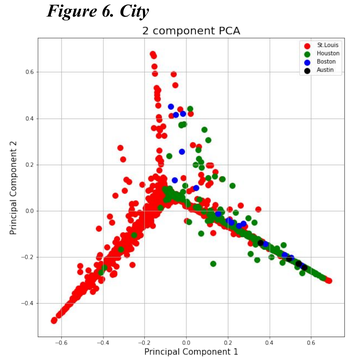



***Figure 6. City***



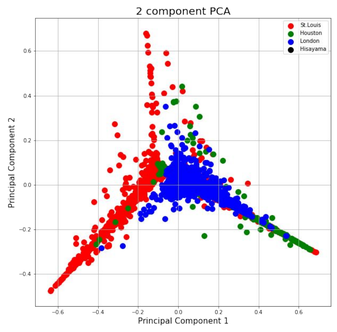




***Figure 7. US State PCA***

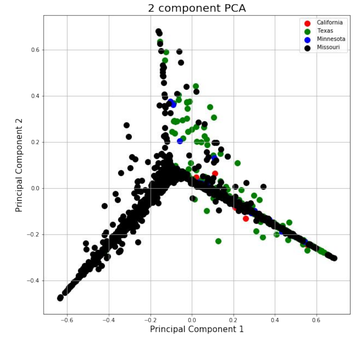



***Figure 8. US Cities PCA***

## Conclusion

**Conclusion in Analysis**

In our analysis we learnt that using a classification based Random Forest, produces the most accurate predictions of stool data. Our research, showed us that Random Forest is the best model to use when trying to make predictions on microbiome data.

**Future Testing**

In future testing, we plan on testing our model with more larger microbial data. In doing this we will be able to test the capabilities of our model and improve it for better predictions.

## References

- Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
- /@tonester524. (2019, August 4). Understanding Random Forest. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2
- J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.