Exploration of Human Microbiome Diversity through K-Means Clustering

DSC 680: Applied Data Science

Deborah L. Young

Business Problem

This project seeks to deepen our understanding of microbial distributions within human hosts, a key factor for advancing a range of microbiome restoration strategies, such as Fecal Microbiota Transplant (FMT). By developing predictive models of microbial ecosystems, the project supports the customization of microbial therapies, including FMT, aimed at rebalancing the microbiome and mitigating a spectrum of related health conditions related to microbial imbalance.

Background/History

Data is sourced from the NIH Human Microbiome Project (*NIH Human Microbiome Project*), an extensive collaborative effort involving over 300 scientists from more than 80 organizations. The HMP provides comprehensive information on microbial samples from various human body sites to explore the role of microbial communities in human health and disease. The project has systematically catalogued microbial samples from 300 adults at various body sites, serving to create reference genomes for these microbes and investigate the microbiomes of individuals with certain diseases. Gut health restoration treatments, including FMT, have emerged as innovative approaches to treating various conditions and show promise in improving whole-body health. The human microbiome's complexity and its role in health underscore the need for detailed study and predictive modeling to enhance therapeutic outcomes.

Data Explanation

This rich dataset includes microbial community profiles and whole-genome sequences, which will be instrumental in analyzing microbial diversity and abundance. It contains samples from various body sites, such as the mouth, lungs, genital tract, skin, and gut. This comprehensive approach provides a significant resource for understanding human-associated microbes.

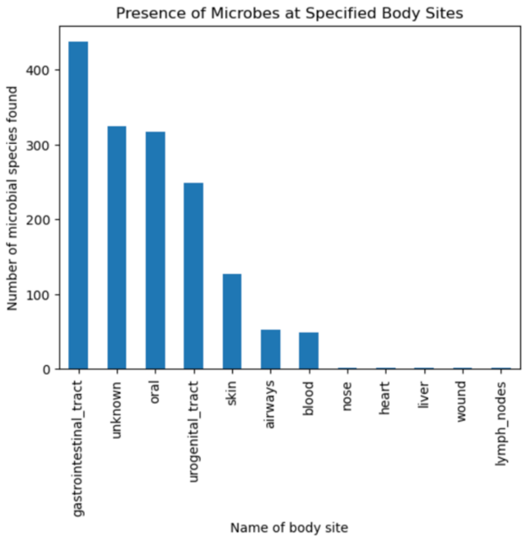
Columns included in the dataset are listed below alongside their descriptions. The relevance of these features to our project will be determined during data cleaning.

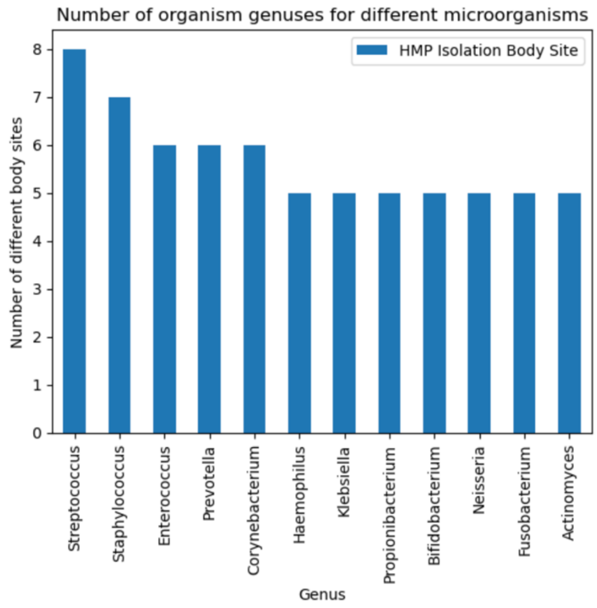
Field Name	**Description**
HMP ID	Unique identifier for each sample in the HMP.
GOLD ID	Identifier linking to the project info in the GOLD database.
Organism Name	Scientific name of the sample organism.
Domain	Highest taxonomic rank of the organisms, typically Bacteria, Archaea, or Eukaryota.
NCBI Superkingdom	Broad classification of life as recognized by NCBI.
HMP Isolation Body Site	Body site from which the sample was isolated.
Project Status	Indicates whether the project is complete or in progress.
Current Finishing Level	Completeness level of genome sequencing.
NCBI Submission Status	Status of the data submission to the NCBI database.
NCBI Project ID	Identifier for the project within NCBI.
Genbank ID	Accession number for the sequence records in GenBank.
Gene Count	Number of genes identified in the genome.
IMG/HMP ID	Identifier for the sample in the IMG & Microbiomes database specific to HMP.
HOMD ID	Identifier for the organism in the Human Oral Microbiome Database, if applicable.
Sequencing Center	Institution where the sequencing was performed.
Funding Source	Organization that provided funding for the project.
Strain Repository ID	Identifier for the organism's strain in the storage repository.

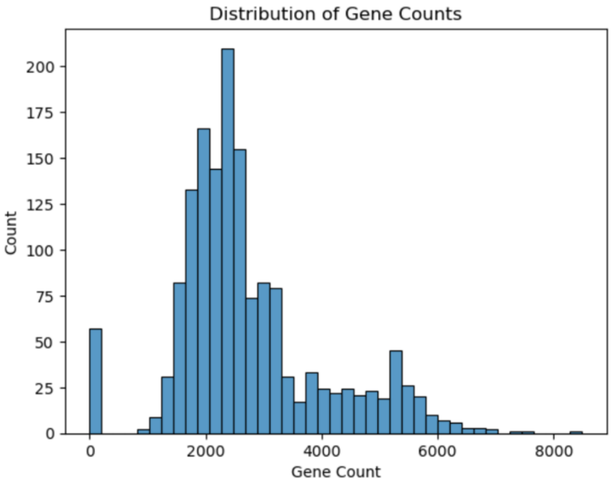
Methods

The project began with an **EDA** (Exploratory Data Analysis) to examine species distribution and gene prevalence and visualize data.

Null values were carefully imputed, duplicate entries were managed, and measures were taken to ensure the consistency of taxonomic classifications, which is critical in biological datasets.







The preprocessing and modeling evaluation of the project involved several key steps to prepare the data for clustering and to assess the effectiveness of the model. Here's a detailed breakdown of each step:

1. **Feature Selection and Encoding**:
    - The dataset was initially comprised of various features, including the body site of isolation, genus, and gene count.
    - Categorical features like the body site and genus were **one-hot encoded**, which means they were converted into a binary matrix representation for the modeling process. This step is crucial when working with categorical data that a model cannot interpret in its raw form.
    - Numerical features like gene count were standardized using **StandardScaler**. Standardization involves rescaling the features so that they have a mean of 0 and a standard deviation of 1. This process ensures that each feature contributes equally to the result and improves the convergence of the clustering algorithm.

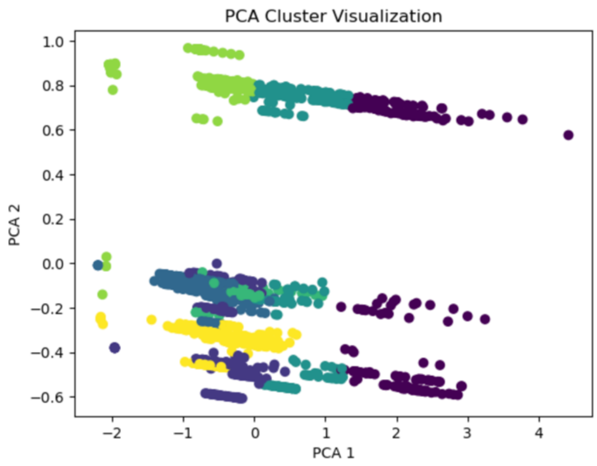
2. **Column Transformation**:
   - A **ColumnTransformer** was created to apply the appropriate transformations to the different types of features within the dataset, such as standardization for numerical features and one-hot encoding for categorical features.
3. **Data Transformation**:
   - The **fit_transform** method was called on the data to apply the transformations and prepare it for clustering.
4. **Elbow Method for Optimal Clusters**:
   - The elbow method was employed to find the optimal number of clusters (k) for the k-means algorithm. This method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and looking for the "elbow" point where the rate of decrease sharply changes. This represents a point of diminishing returns where increasing the number of clusters does not significantly improve the within-cluster variance.
5. **K-Means Clustering**:
   - K-means clustering was applied to the processed data with the chosen number of clusters (in this case, 7). K-means is an iterative algorithm that partitions the dataset into k pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.

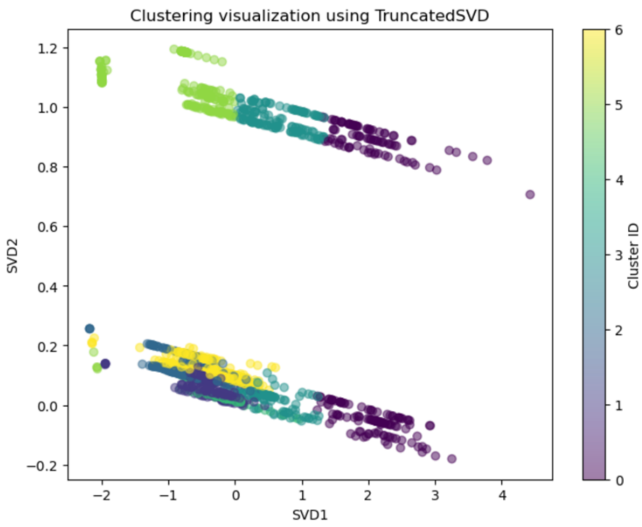
6. **Cluster Analysis**:

    - After clustering, the cluster labels were added to the dataset as a new feature to analyze the distribution of data points across the clusters. This helps in understanding the composition of each cluster.

7. **Visualization**:

    - Dimensionality reduction techniques such as PCA (Principal Component Analysis) and TruncatedSVD (Singular Value Decomposition) were used to visualize the clusters in two dimensions. These techniques reduce the number of variables of the data while preserving as much information as possible. The visualizations can provide insights into how well-separated the clusters are.





8. **Statistical Evaluation**:

    - ANOVA (Analysis of Variance) was conducted to evaluate the statistical significance of the clusters with respect to gene count. A significant F-statistic and a p-value close to zero indicated that there are statistically significant differences in the mean gene counts across different clusters.

9. **Silhouette Score**:

    - The silhouette score was calculated to measure the quality of the clusters. A score closer to +1 indicates that the sample is far away from the neighboring clusters. A score of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

Analysis

The analysis of the k-means clustering on the microbial dataset reveals a structured separation into distinct clusters. These clusters are likely representative of different ecological niches within the human body, indicated by their unique gene counts, species composition, and isolation sites. The identified clusters hold potential insights into the roles these microbial communities play in health and disease contexts.

For visualization, dimensionality reduction techniques like PCA and TruncatedSVD were used. These visualizations reinforced the separation of clusters, albeit indicating room for further improvement. For instance, the silhouette score of approximately 0.25 suggests moderate cluster separation; this is above the threshold where clusters overlap but below the level indicating strong separation. Hence, there might be potential to refine the model by adjusting the number of clusters, reconsidering the features used for clustering, or applying a different clustering technique.

Statistical analysis, including ANOVA, was performed to validate the significance of these clusters. The results, showing a high F-statistic and a p-value of zero, suggest strong statistical significance in gene count variations across different clusters. This confirms that the clusters have distinct microbial characteristics based on gene counts.

Overall, the k-means clustering approach effectively grouped organisms and revealed patterns and associations between microbial genera, gene counts, and their prevalence across different body sites. The statistical validation underscores the clusters' significance in terms of gene count variation. However, the silhouette score indicates that while there is some structure to the clusters, there is potential for further refinement of the model to enhance its interpretability and application in microbiome research and targeted therapeutic interventions.

Conclusion

The application of advanced data analytics, specifically k-means clustering, to the data from the Human Microbiome Project has provided valuable insights into the complex landscape of our microbiome. Our analysis successfully identified distinct microbial patterns that correspond to various human body sites, offering a window into the intricate relationships between microbes and their hosts. These findings not only enrich our understanding of microbial ecosystems in the human body but also lay a foundational framework for future health interventions. By demonstrating the structured nature of microbial diversity, this study paves the way for the next generation of precision medicine, where such insights could be leveraged to tailor medical treatments to individual microbial profiles.

Assumptions

The analysis assumes microbial community stability across individuals and the generalizability of findings across different populations. This project operates with the presumption that the microbial ecosystems across different human body sites are consistent among individuals, and the patterns discovered can be generalized across various populations. It assumes that the microbial samples from the NIH Human Microbiome Project accurately reflect the diversity and prevalence of microbial species across the body sites studied.

Furthermore, the project's predictive modeling presumes the stability of these microbial communities and their predictable response to restorative therapies, despite the complex and dynamic nature of the human microbiome. These assumptions are crucial as they underpin the project's analytical framework and influence the interpretation of the clustering results, guiding the development of personalized medical interventions.

Limitations

One major limitation is the reliance on clustering algorithms that may impose artificial boundaries within a continuous ecological spectrum. The diversity and complexity of the human microbiome, along with inter-individual variability, present challenges in defining clear-cut clusters. Furthermore, the clusters identified may not fully capture the functional capabilities of the microbial communities due to the focus on genetic markers rather than metabolic or phenotypic data. Our study aimed to create universally applicable models, while acknowledging the potential limitations in generalizability for applying these findings to all demographic groups due to the unique and dynamic nature of individual microbiomes.

Challenges

Challenges arose in dealing with high-dimensional data, the need for significant data preprocessing, and ensuring accurate clustering. A meticulous approach was adopted as we attempted to retain as much meaningful information as possible while eliminating extraneous noise. Despite our best efforts, the moderate silhouette score indicates that the clusters' separation is less than ideal. A score close to 0.25 suggests some level of structure within the clusters; however, it also points to a potential overlap and a lack of clear delineation between different microbial communities. This outcome signals a need for further refinement of the clustering approach.

Human microbiome data is intricate and varies significantly across different populations and locations. These challenges highlight the nuanced nature of microbial diversity and its representation in a clustered model.

Future Uses/Additional Applications

The insights from the Human Microbiome Project pave the way for innovative health solutions. They have the potential to revolutionize medical care by allowing for more personalized treatments. This could include designing specific probiotics tailored to individual needs or even more nuanced dietary recommendations that support a person's unique microbiome, enhancing not just gut health but overall wellness.

Recommendations

The diversity of the microbiome is a treasure trove of information waiting to be fully explored. Future iterations of the model could be refined, possibly by re-evaluating the number of clusters, reassessing the features used for clustering, or exploring alternative clustering techniques to enhance the precision of microbial community classification and, by extension, the insights into their roles in human health and disease.

Continuing research across different populations and environments will help us understand the vast potential of microbiome therapy. Long-term studies on treatments like FMT are crucial to ensure they are not just effective but sustainable for long-term health benefits.

Implementation Plan

Bringing the research into practice is a team effort. It involves creating a collaborative network of experts from various fields, from scientists and doctors to policymakers and ethical experts. Together, they will guide the safe and effective translation of microbiome research into treatments, ensuring that advancements in this field reach those who need them most, safely, and ethically.

Ethical Considerations

In the realm of machine learning projects, ethical considerations are paramount. They encompass ensuring fairness and transparency, safeguarding against biases in data and algorithms, maintaining privacy and security of data, and adhering to regulatory and ethical standards. Ensuring that models do not perpetuate or amplify existing social biases and are transparent in their functioning and decision-making processes is essential. An ethical ML project should also provide clear documentation and accountability measures. The aim is to complement human expertise, enhance decision-making, and improve efficiency without infringing on individual rights or reinforcing harmful stereotypes.

Our approach involves maintaining the integrity of the dataset and ensuring privacy and confidentiality regarding human subject data. The research must also avoid overpromising the benefits of microbiome mapping and consider the implications of misinterpreting microbial functions and interactions. It is important to note the limitations in generalizability for applying these findings to all demographic groups due to the unique and dynamic nature of individual microbiomes. The goal of this project is to not only to contribute to the scientific community, but also potentially impact the medical field by informing therapeutic practices.

References

NIH Human Microbiome Project - View Dataset. (n.d.). Retrieved February 14, 2024, from https://www.hmpdacc.org/hmp/catalog/grid.php?dataset=genomic

Ooijevaar, R. E., Terveer, E. M., Verspaget, H. W., Kuijper, E. J., & Keller, J. J. (2019). Clinical Application and Potential of Fecal Microbiota Transplantation. *Annual Review of Medicine*, *70*, 335–351. https://doi.org/10.1146/annurev-med-111717-122956

OpenAI. (2024). ChatGPT [Large language model]. https://chat.openai.com

Orr, M. R., Kocurek, K. M., & Young, D. L. (2018). Gut Microbiota and Human Health: Insights From Ecological Restoration. *The Quarterly Review of Biology*, *93*(2), 73–90. https://doi.org/10.1086/698021

The Human Microbiome Project - Registry of Open Data on AWS. (n.d.). Retrieved February 14, 2024, from https://registry.opendata.aws/human-microbiome-project/

Understanding K-means Clustering with Examples. (2014, July 25). *Edureka*. https://www.edureka.co/blog/k-means-clustering/

Yang, D., & Xu, W. (2020). Clustering on Human Microbiome Sequencing Data: A Distance-Based Unsupervised Learning Model. *Microorganisms*, *8*(10), 1612. https://doi.org/10.3390/microorganisms8101612

10 Questions an Audience May Have

1. What was the primary motivation behind this research?
Our main goal was to understand the distribution of microbial species across human body sites to aid in microbiome restoration treatments.

2. How did you ensure the data used in the project was reliable?
The data came from the NIH Human Microbiome Project, which is a reputable source that involves stringent data collection and validation protocols.

3. How did you address potential biases in your study?
We used clustering algorithms to identify patterns without relying on preconceived labels, reducing the risk of biases influencing our results.

4. What measures did you take to ensure the reproducibility of your research?
We've provided detailed documentation of our methods and analysis, allowing other researchers to replicate our study.

5. How does the variation in gene count across different body sites influence the microbial ecosystem's stability?
The gene count can indicate microbial diversity and functional potential. Variations suggest different ecological roles and resilience to perturbations across body sites, affecting the stability and health implications of the microbiome.

6. Could your predictive model be applied to other aspects of human health?
Yes, the predictive modeling approach could be adapted to study other complex systems within human health beyond the microbiome.

7. How do you envision the practical implementation of your findings?
Collaborations with healthcare professionals to integrate our predictive model into clinical decision-making processes for microbiome restoration treatments.

8.What would be the societal impact of advancing microbiome restoration therapies?
It could lead to breakthroughs in treating chronic diseases and improving overall human health by maintaining a healthy microbiome balance.

9. Can your model's findings be generalized to populations outside the study sample?
While the model provides valuable insights, caution should be exercised when generalizing to other populations due to potential variations in genetic backgrounds and environmental exposures.

10. How can this model be integrated into current clinical practices for diagnosing or treating microbiome-related conditions?
The model can inform the design of diagnostic tools and interventions, but careful clinical validation and integration into existing medical workflows are needed for practical applications.