Preterm Birth Prediction Using Microbiome Data

Team: Data Explorers

Authors: Kalpkumar Shah: kshah223@uic.edu
Naivik Patel: npate431@uic.edu
Ricardo Diaz: rgonza82@uic.edu
Tina Khajeh: tkhaje2@uic.edu
Victor Marqueti: vpasto2@uic.edu

Institution: University of Illinois, Chicago

Date: 12/06/2023

Abstract:

This study focuses on analyzing the vaginal microbiome in pregnant women to identify potential indicators of preterm birth. We use advanced data analysis and machine learning techniques to explore the correlation between the microbiome composition and the incidence of births occurring before 37 weeks of gestation. Our findings seek to contribute to the broader understanding of maternal and infant health, potentially aiding in predicting and preventing preterm births.

Introduction:

Preterm birth, defined as childbirth occurring before the completion of 37 weeks of gestation, poses significant health risks for infants and is a matter of global health concern. The vaginal microbiome, which plays a crucial role in a woman's reproductive health, has been hypothesized to influence the likelihood of preterm birth. This study embarks on an in-depth analysis of the vaginal microbiome in pregnant women to unearth possible biological markers and correlations that could predict preterm birth.

Our project narrows its focus on the age-related aspects of the microbiome and its diversity, considering these factors as potential indicators of preterm births. While the study initially considered a broader range of demographic factors, it has since honed in on more biologically pertinent features. The primary goal is to develop a predictive model using machine learning or deep neural network techniques to assess the risk of preterm birth and identify the crucial biomarkers associated with preterm birth. This model aims to differentiate between births occurring before and after 37 weeks, considering maternal age as a significant variable.

The hypothesis guiding our research is twofold:

- Null Hypothesis: Age does not have a significant impact on preterm birth.
- Alternate Hypothesis: Age has a significant impact on preterm birth.

Our approach involves a comprehensive analysis of a detailed dataset encompassing various aspects of the vaginal microbiome. Through this analytical lens, we aim to provide valuable insights that could further scientific understanding in this domain and offer practical implications for prenatal care and intervention strategies.

Dataset link: https://www.synapse.org/#!Synapse:syn26133770/wiki/612541

Literature Review:

The link between the vaginal microbiome and preterm birth has been a subject of increasing research interest. Studies have shown that the composition of the vaginal microbiome can significantly influence maternal and neonatal health outcomes. Our review focuses on key studies that have explored this relationship, highlighting the methodologies, findings, and implications of this research.

Several studies have identified specific microbial profiles associated with an increased risk of preterm birth. For instance, a shift away from Lactobacillus-dominated microbiota has been correlated with adverse pregnancy outcomes, including preterm birth. The mechanisms through which these microbial shifts impact pregnancy outcomes are complex and multifaceted, involving factors like inflammation, immune responses, and hormonal changes.

Research has also delved into the role of maternal age as a factor influencing the vaginal microbiome. Age-related microbiome composition changes could contribute to the varied risks of preterm birth observed across different age groups. Understanding these age-related dynamics is crucial for developing targeted interventions.

The literature review establishes a foundation for our study, situating it within the broader context of existing research. It also highlights gaps and limitations in current knowledge, underscoring the need for further investigation, particularly regarding the predictive capabilities of microbiome analysis for preterm birth.

Project Goals:

The primary goal of our project is to harness the power of data analysis and machine learning to predict the risk of preterm birth based on the analysis of the vaginal microbiome and to identify the crucial biomarkers. By focusing on maternal age and microbial diversity, we aim to uncover significant patterns and correlations that could be used in clinical settings to identify women at higher risk of preterm delivery.

Hypotheses:

Our research is driven by two central hypotheses:

1. **Null Hypothesis:** There is no significant correlation between maternal age and the incidence of preterm birth. This hypothesis posits that age, as an isolated factor, does not significantly influence the likelihood of preterm delivery.

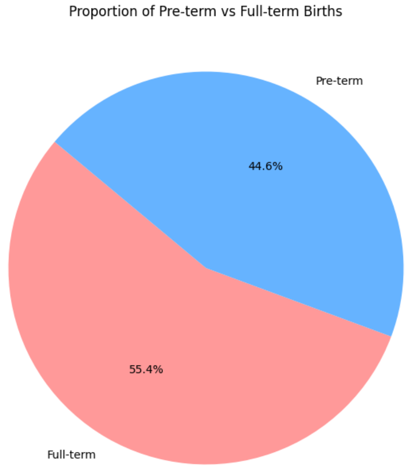
2. **Alternate Hypothesis:** Maternal age significantly impacts the incidence of preterm birth. This hypothesis suggests that variations in maternal age could correlate with distinct changes in the vaginal microbiome, consequently affecting the risk of preterm delivery.

These hypotheses stem from the observation that preterm birth rates vary across different age groups and the hypothesis that these variations might be partly attributable to age-related changes in the vaginal microbiome. By testing these hypotheses, we aim to contribute to the ongoing discourse on the determinants of preterm birth and potentially pave the way for new predictive tools in prenatal care.

Methodology

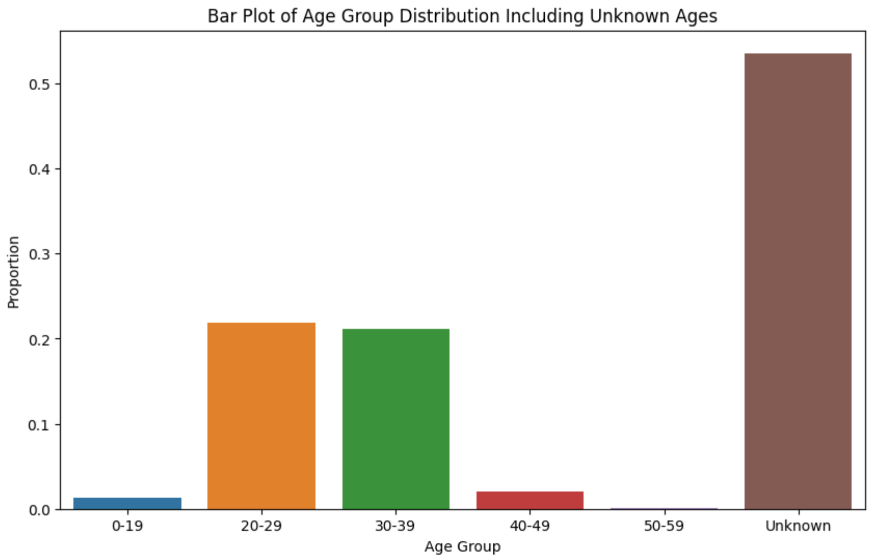
Data Description:

Our study utilizes a comprehensive dataset detailing the vaginal microbiome of pregnant women. The dataset includes a wide array of variables such as specimen details, participant ID, term status (was_term), delivery week (delivery_wk), age of the mother, and various microbial components. This rich dataset allows for a nuanced analysis of the microbiome in the context of preterm birth.

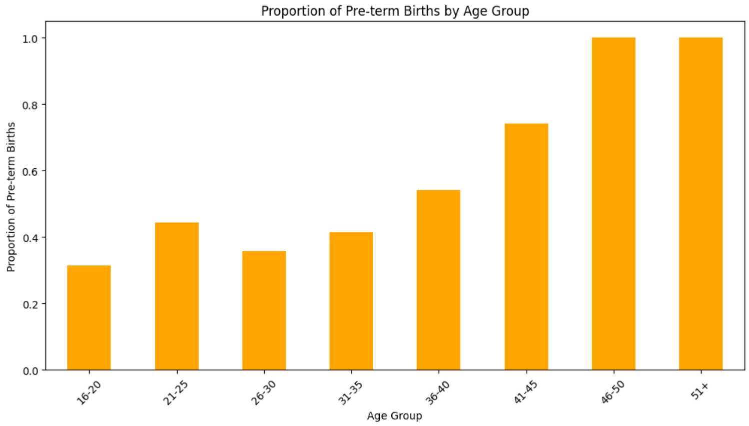


Data Cleaning and Preprocessing:

The initial step in our analysis involved streamlining the dataset to focus on variables relevant to our study. Non-essential columns were excluded from refining the dataset for a more targeted analysis, particularly those not directly related to our hypotheses (such as 'NIH Racial Category' and 'NIH Ethnicity Category').



A significant aspect of data cleaning involved handling missing values, especially in the 'age' column. Ages recorded as 'Unknown' were treated as missing data and were converted to NaN (Not a Number) values. These NaN values were then excluded from the dataset to ensure the integrity and accuracy of our analysis. This preprocessing step was crucial to preparing the dataset for the subsequent statistical and machine-learning analyses.



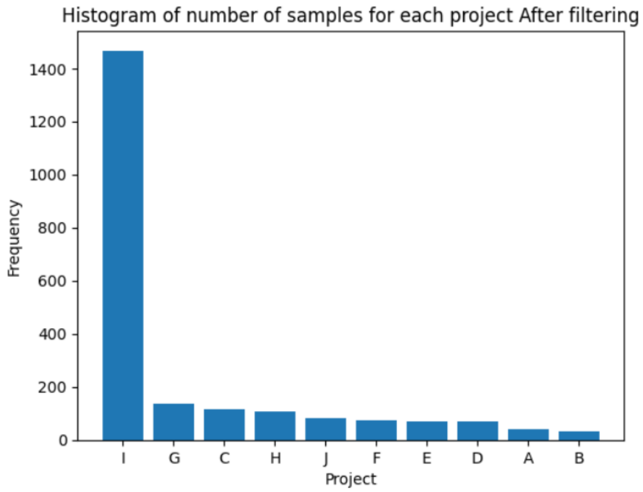
In addition, we filtered out all samples collected after week 37, and for the microbiome data, we removed microbes with a prevalence of less than 90% in the samples.

Statistical Analysis:

Our approach to data analysis included both descriptive and inferential statistical methods. Descriptive statistics provided a baseline understanding of the dataset, including the distribution of ages, term status, and various microbiome components. Inferential statistics were employed to test our hypotheses, particularly focusing on the relationship between maternal age and preterm birth.

Machine Learning Model:

To prepare the data for the machine learning model, we scaled the features to have a mean of 0 and a standard deviation of 1. Subsequently, we divided the data into training and testing sets. The samples were collected from different hospitals, with samples from projects C, E, G, I, J, and H selected for the training set and samples from projects A, B, D, and F for the testing set. Consequently, the training set size is (1980, 50), and the testing set size is (216, 50). Also, Hyperparameters were set using 5-fold cross-validation for SVM with an RBF kernel, and the number of trees in the Random Forest was set to 200.

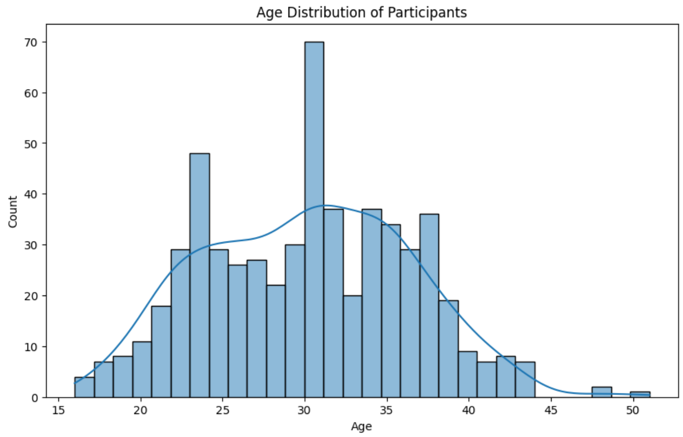


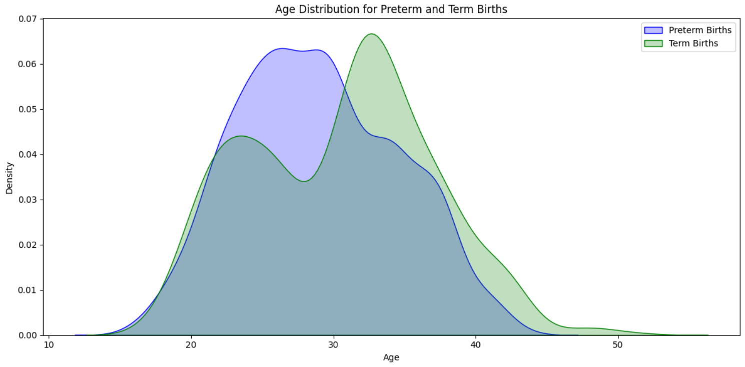
Data Analysis:

The data analysis phase of our project was multifaceted, involving both statistical analysis and machine learning techniques. Our approach included:

- **Exploratory Data Analysis (EDA):** EDA was conducted to gain an initial understanding of the data, identify patterns, and check for inconsistencies or outliers. This involved visualizing the distribution of key variables like age, term status, and various microbiome components.

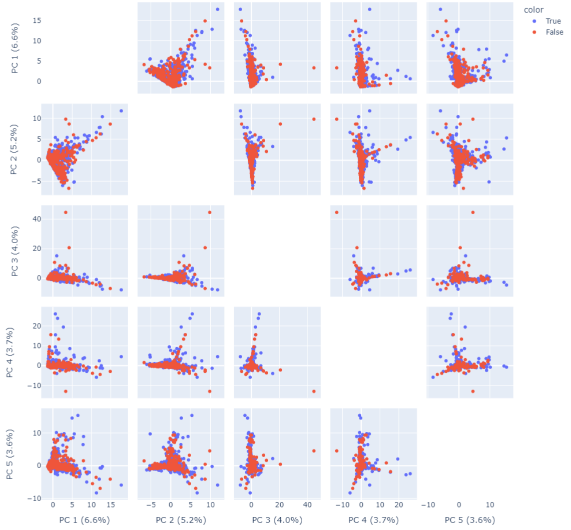
- **Statistical Testing:** To address our hypotheses, we performed statistical tests to ascertain the relationship between maternal age and the incidence of preterm births. This involved using chi-square tests for categorical data and t-tests or ANOVA for continuous data, depending on the nature of the variables under examination.





- **Machine Learning Model:** We developed predictive models based on machine learning, utilizing Random Forest and SVM for predicting preterm cases. The dataset included microbiome data at three levels: species, genus, and family. Initially, we plotted PCA plots using PC1 to PC5 to analyze the data separation for two classes at different levels. The results showed no specific separation of the data.

  Subsequently, we ran SVM models on the microbiome data at each level and chose the taxonomy level with higher performance, which was family. Finally, we used microbiome taxa at the family level and clinical data as input for the model. Hyperparameters for Random Forest and SVM were tuned using 5-fold cross-validation. The model was trained on the training set and tested 20 times on the testing set. We ranked features using Random Forest and SVM for biomarker discovery, which involves identifying highly important features. We selected the top 10 features with high importance.

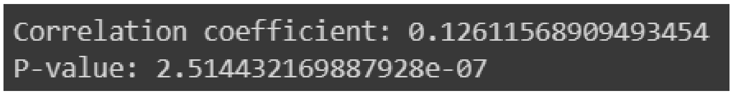


PCA plot for PC1 to PC5 at family level

Results:

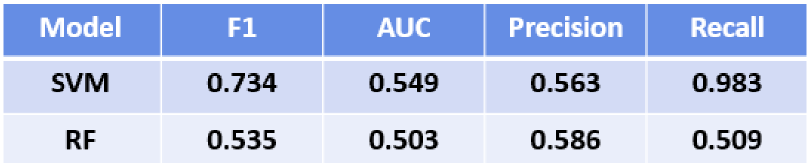
The results of our study were revealing in several aspects:

- **Statistical Analysis Results:** The statistical tests provided insights into the relationship between maternal age and preterm birth.



```
Correlation coefficient: 0.12611568909493454
P-value: 2.514432169887928e-07
```

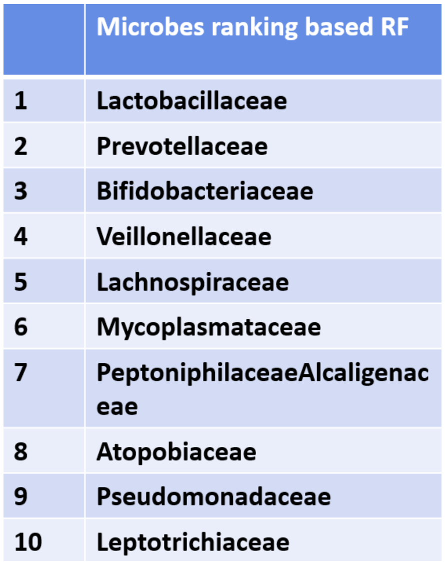
- **Machine Learning Model Performance:** The predictive performance of our developed models is presented in the following table. Given the medical nature of our problem, we employed metrics such as F1 score, AUC, precision, and recall to assess the model's performance for both case and control samples. The results represent the mean value from 20 runs on the test set. As indicated in the table, SVM outperformed Random Forest according to the F1 score, achieving a score of 0.734.



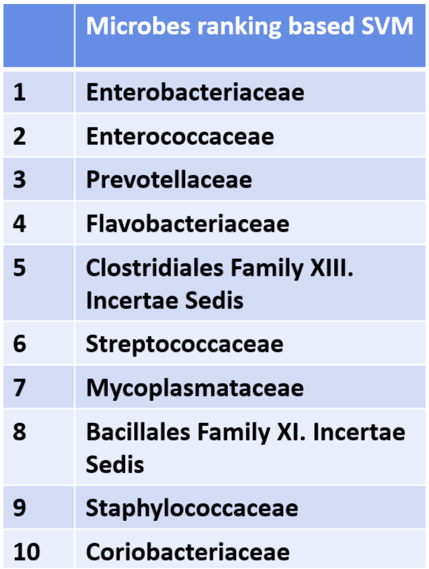
Model	F1	AUC	Precision	Recall
SVM	0.734	0.549	0.563	0.983
RF	0.535	0.503	0.586	0.509

Furthermore, we ranked the important features for preterm prediction, and the results are illustrated in the following two tables. The top 10 microbes are reported for both the Random Forest and SVM models. The significance of most of these microbes has been validated in the literature.

We ranked the important features for preterm prediction, and the results are illustrated in the following two tables. The top 10 microbes are reported for both the Random Forest and SVM models. The significance of most of these microbes has been validated in the literature

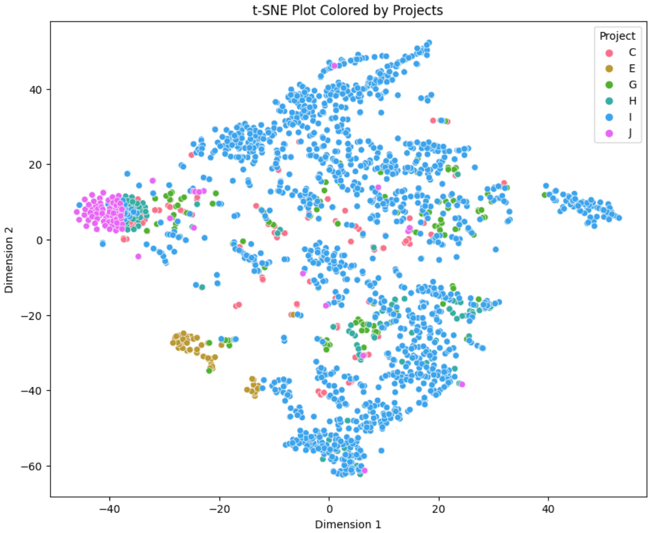


	Microbes ranking based RF
1	Lactobacillaceae
2	Prevotellaceae
3	Bifidobacteriaceae
4	Veillonellaceae
5	Lachnospiraceae
6	Mycoplasmataceae
7	PeptoniphilaceaeAlcaligenaceae
8	Atopobiaceae
9	Pseudomonadaceae
10	Leptotrichiaceae



	Microbes ranking based SVM
1	Enterobacteriaceae
2	Enterococcaceae
3	Prevotellaceae
4	Flavobacteriaceae
5	Clostridiales Family XIII. Incertae Sedis
6	Streptococcaceae
7	Mycoplasmataceae
8	Bacillales Family XI. Incertae Sedis
9	Staphylococcaceae
10	Coriobacteriaceae

The model performances were observed to be close to random, prompting an investigation into the reasons behind this outcome. To gain insights, we generated a 2D scatter plot using t-SNE and color-coded the samples based on the collection hospital, as depicted in the following plot. Remarkably, distinct clusters were formed by samples from different projects, suggesting that each project contributed to its cluster. Importantly, the data reveals the presence of non-biological biases, which may contribute to the observed lower performance.



Discussion:

In this project, we highlight the potential of the vaginal microbiome as a predictive biomarker for preterm birth, underscoring its significance in prenatal care. The identification of such biomarkers opens avenues for targeted interventions, particularly for populations at a higher risk of preterm birth.

While acknowledging the strengths of our methodology, which includes robust statistical methods and machine learning techniques for biomarker discovery, aligning with existing literature, we remain mindful of the study's limitations. Notably, the presence of potential biases has been identified as a factor contributing to lower model performance. A comprehensive exploration of these biases revealed non-biological factors. Addressing these factors in future analyses has the potential to enhance the accuracy of our predictive models.

Future Work:

The study opens several avenues for future research. This includes exploring additional variables that may influence preterm birth, applying our model to different populations, and integrating our findings into clinical practice.

Future endeavors could prioritize the refinement and validation of the predictive model by incorporating larger and more diverse datasets, including aspects like alpha diversity. Information visualizations such as PCoA plots can glean insights into diversity metrics. This has the potential to augment the model's accuracy and provide valuable insights into the dataset.

To address potential biases, utilizing approaches for bias mitigation before applying machine learning models should be considered. Additionally, given the high dimensionality of the data, employing dimension reduction techniques and neural networks, such as autoencoders, may further enhance model performance.