**Experimental Process Design**

**Experiment:**

**Problem Analysis:**

**Objective:**

To select features for each age group, perform clustering based on feature differences, and minimize the impact of age-related variations (considering the results from Question 1). The optimal hyperparameter, i.e., the number of labels, will be selected as an estimate of the number of categories based on a certain evaluation criterion.

**Data Preprocessing:**

1. **Classify samples by age group.**
2. **Remove the top 95% and bottom 5% of samples for each feature** to reduce the impact of outliers.

**Feature Engineering:**

1. **Analyze the feature differences for each age group.**
   * **Remove features with low variability** and keep those with high variability.
   * Since the dataset is large, some features may need to be discarded. An alternative approach is **dimensionality reduction**.

**Reasoning:**  
If a feature has very low variability but is still used, it will increase the similarity between samples. (For example, in a 2D plane with K-means clustering, selecting a feature with low variability will reduce the distance between samples, thereby affecting clustering results.)

**Note:**  
Different features may need to be selected for different age groups because the microbial species may exhibit different characteristics during different periods. That is, the variability of a specific feature may differ across age groups.

**Model Building:**

**Clustering Model Comparison:**

A comparison was conducted among the following clustering models:

* **K-means**
* **Learning Vector Quantization**
* **Hierarchical Clustering**
* **Density-Based Clustering**
* **Gaussian Mixture Clustering**

**1) K-means:**

* K-means typically measures sample differences using **Minkowski distance** and clusters data by minimizing the intra-cluster vector error.
* It is an intuitive clustering method but **not suitable for this experiment** because:
  + Since we are clustering within a specific age group, and feature magnitudes may be age-dependent (based on Question 1), the clustering results could be correlated with age instead of potential species.
  + For instance, in the age group (1-10), the final clusters might group samples from different eras rather than grouping based on microbial species.
  + Therefore, **K-means fails to mitigate the influence of age variations** and is discarded.

**2) Learning Vector Quantization (LVQ):**

* LVQ clusters data by learning **prototype vectors**.
* It **requires labeled samples**, which makes it unsuitable for this task.

**3) Hierarchical Clustering:**

* This method clusters data from bottom to top.
* It is **only suitable for small datasets**, whereas our dataset has over 2 million records.
* **Not feasible** for this experiment.

**4) Density-Based Clustering:**

* Expands clusters by linking data points based on density.
* **Similar to K-means**, it does not effectively mitigate the influence of age variations.
* **Not suitable** for this experiment.

**5) Gaussian Mixture Clustering (GMM):**

* Assumes that the samples are generated from a **mixture of Gaussian distributions**.
* Clusters are formed by identifying the prototype Gaussian distributions.

**Key Considerations:**

* We need to determine whether our data can be assumed to follow this model.
* If we assume that different microbial species within a specific age group exhibit **multidimensional Gaussian distributions**, then the sample space can be considered as generated from a **Gaussian mixture model**.
* According to **the Law of Large Numbers in natural sciences**, this assumption is likely to hold.
* Since this assumption is valid, we **can ignore the impact of age variations** on the clustering effectiveness.

Thus, **GMM is the chosen model**.

**Optimization:**

**Performance Evaluation:**

* The **Akaike Information Criterion (AIC)** is used for evaluation.
* **Silhouette Score is not used** because it evaluates clustering similarly to K-means and does not account for age-related influences.

**Visualization:**

1. **Plot the optimal number of clusters** for different age groups.
2. **Generate scatter plots** to visualize the clustering results.

**Potential Experiment Improvements:**

* Since we have **over 2 million samples**, directly feeding them into the model **may introduce excessive noise**.
* To mitigate this, we can incorporate:
  1. **Resampling techniques**
  2. **Cross-validation**  
     These methods can help **enhance model performance** and reduce noise effects.