CS-1390/ PHY-1390-1
Introduction to Machine Learning
Monsoon 2024
Assignment 2

Clustering Task

Data: *Provided separately with the classroom post*.

Data Description

Column Name	**Description**
Variant	Indicates the specific variant or type for each sample.
Sample_ID	Unique identifiers for each sample, often combining multiple reference codes.
Sample_Code	A unique code for each sample, typically used as the primary identifier for each record.
Entry_Date	The date when the sample was recorded or added to the database (in standard date format).
Origin	Geographical origin of the sample, including country and region where applicable.
Source_Category	Describes the source of the sample, such as Natural_Source (environmental/other) or Medical_Source (clinical).
Genetic_Group	Genetic group identifier based on molecular analysis, useful for genetic comparisons.
Intra_Group_Similarity	A numerical value representing the genetic similarity within the same group.
Inter_Group_Difference	A numerical value representing the genetic difference between different groups.
Reference_ID	External reference identifier associated with the sample, used for cross-referencing with other databases.
Genomic_Data	Genomic data identifier for samples where molecular sequencing information is available.
Resistance_Profile	Describes the resistance profiles detected for each sample, indicating genetic markers linked to treatment resistance.
Predicted_Characteristics	Characteristics predicted from molecular data or other methods, indicating the predicted properties of the sample.

Any clustering algorithm can be used.

Submission Guidelines

1. Submit a python notebook (with cell outputs, we will not run your code).
2. Include an extensive report within the python notebook throughout your code. Use markdown.
3. Grading emphasis will be on coding and plotting/describing clusters.
4. Use of libraries is not allowed. Code everything from scratch.
5. Any code similarities with your peers and code from the internet, as well as any AI generated code will result in a 0, along with further severe consequences under discretion of the professor.