ML3 mini project walkthrough

The student's approach to analyzing the National Health and Nutrition Examination Survey (NHANES) dataset can be summarized in the following steps:

1. Import Libraries : Import necessary Python libraries for data manipulation, visualization, and clustering.

2. Load Data : Load the dataset into a pandas DataFrame.

3. Data Dictionary : Load a data dictionary that describes the dataset.

4. Initial Data Exploration : Check the shape of the DataFrame and the number of missing values.

5. Data Cleaning : Fill missing values in dietary columns with zeros based on the assumption that they are question answers.

6. Demographic Data Cleaning : Fill missing values in demographic columns with zeros.

7. Laboratory Data Cleaning : Fill missing values in laboratory columns with zeros.

8. Final Missing Value Handling : Fill any remaining missing values in the dataset with zeros.

9. Feature Selection : Select numerical features for clustering and drop non-essential columns.

10. Data Transformation : Scale the numerical features using MinMaxScaler.

11. Optimal Cluster Identification : Use silhouette scores and the elbow method to find the optimal number of clusters for KMeans.

12. KMeans Clustering : Perform KMeans clustering with the identified optimal number of clusters.

13. Correlation Analysis : Analyze the correlation of features with the cluster labels to identify highly correlated features.

14. Principal Component Analysis (PCA) : Perform PCA to reduce dimensionality while retaining 95% of the variance.

15. KMeans with PCA : Apply KMeans clustering on the principal components and visualize the clusters in a scatter plot.

16. DBSCAN Clustering : Attempt DBSCAN clustering on the principal components and evaluate the results.

17. Conclusion : Conclude the analysis by summarizing the steps taken and the findings from the clustering analysis.

Throughout these steps, the student uses a combination of data preprocessing techniques, clustering algorithms, and visualization tools to segment the patient data into meaningful clusters. They also make decisions based on statistical measures like silhouette scores and the explained variance ratio from PCA.

****