Project Documentation: Student Evaluation Clustering Analysis

Introduction

This project involves analyzing and clustering student evaluations using various machine learning techniques, including Principal Component Analysis (PCA), K-Means clustering, and Agglomerative Clustering. The goal is to identify distinct clusters of students based on their evaluation responses, providing insights into different groups within the student population.

Data Description

- **Dataset:** `turkiye-student-evaluation_generic.csv`
- **Description:** The dataset contains evaluation scores of students across various classes and instructors. Each row represents a student's response, and the columns include demographic details, course information, and responses to multiple evaluation questions.
- **Key Columns:**
  - `instr`: Instructor identifier
  - `class`: Class identifier
  - `Q1-Q28`: Responses to evaluation questions (the primary focus of the clustering analysis)

Steps Involved and Methods

1. **Loading and Exploring Data:**
   - **Load Data:** The dataset is loaded into a pandas DataFrame.
   - **Statistical Summary:** Descriptive statistics of the dataset are obtained using `describe()` to understand the distribution of responses.
   - **Data Types and Missing Values:** Data types are checked, and missing values are identified, although the dataset does not contain any null values.
2. **Data Visualization:**
   - **Instructor and Class Distribution:** Count plots are used to visualize the distribution of students across different instructors and classes.
   - **Question Mean Scores:** The mean scores of the evaluation questions are calculated and visualized using bar plots, providing insights into the overall response trends.
   - **Correlation Matrix:** A heatmap is generated to visualize the correlation between different evaluation questions, helping to identify patterns and relationships in the data.
3. **Dimensionality Reduction:**
   - **Principal Component Analysis (PCA):** PCA is applied to reduce the dimensionality of the dataset from 28 features (questions) to 2 principal components. This helps in visualizing and simplifying the clustering process while retaining the majority of the variance in the data.
4. **Clustering Analysis:**

- **K-Means Clustering:**
  - **Elbow Method:** The elbow method is applied to determine the optimal number of clusters by plotting distortions (inertia) against the number of clusters.
  - **Cluster Formation:** Based on the elbow method, K-Means clustering is performed with the optimal number of clusters, and the results are visualized in a scatter plot with the centroids marked.
- **Agglomerative Clustering:**
  - **Cluster Formation:** Agglomerative Clustering is performed as an alternative method to identify student clusters. The results are visualized in a scatter plot, highlighting the hierarchical structure of the clusters.

5. **Cluster Interpretation:**
   - The clusters formed by both K-Means and Agglomerative Clustering are analyzed by counting the number of students in each cluster, providing insights into the distribution of student responses across the identified clusters.

Methodology

1. **Data Preparation:**
   - The dataset is first explored for any missing values and anomalies.
   - Mean scores of evaluation questions are computed to understand general trends.
2. **Principal Component Analysis (PCA):**
   - PCA is employed to reduce the dimensionality of the data, making the clustering process more efficient while retaining the essence of the original dataset.
3. **Clustering:**
   - **K-Means Clustering:** An iterative method that partitions the data into k clusters based on distance from cluster centroids.
   - **Agglomerative Clustering:** A hierarchical clustering method that builds nested clusters by progressively merging or splitting them based on distance metrics.
4. **Visualization:**
   - The results of the clustering methods are visualized to aid in the interpretation of the clusters formed and to assess the effectiveness of the clustering algorithms.

Future Work

- **Hyperparameter Tuning:** Further tuning of the clustering algorithms, such as experimenting with different numbers of clusters or linkage methods in Agglomerative Clustering, could improve the model's accuracy.
- **Advanced Clustering Techniques:** Explore other clustering methods like DBSCAN or Gaussian Mixture Models for potentially better results.
- **Evaluation Metrics:** Incorporate internal evaluation metrics like Silhouette Score or Davies-Bouldin Index to quantitatively assess the quality of the clusters.
- **Additional Features:** Include other features or external datasets that may provide more context or explanatory power to the clustering.

Conclusion

The project successfully utilized PCA, K-Means clustering, and Agglomerative Clustering to group students based on their evaluation responses. The clustering revealed distinct groups within the student population, which can be further analyzed to tailor educational strategies or identify patterns in student feedback. The use of PCA effectively reduced the dimensionality, making the clustering process more efficient and interpretable.

Results

- **K-Means Clustering:**
  - **Optimal Clusters:** 3 clusters based on the elbow method.
  - **Cluster Centroids:** Visualized and analyzed for interpretation.
- **Agglomerative Clustering:**
  - **Clusters Formed:** 2 distinct clusters were identified and visualized.
- **Explained Variance by PCA:** The first two principal components retained a significant portion of the variance, making the PCA transformation effective.

Summary

This project demonstrates the application of clustering techniques to student evaluation data. Through data visualization, PCA, and clustering methods, distinct groups within the student population were identified, offering insights into student feedback patterns. The methodology and results provide a foundation for further analysis and refinement of clustering approaches.