Project Writeup: Personalized Activity/Workout Recommendation System

Team: ML Mavericks

Team Members:

- Soumya Bharathi Vetukuri
- Shubham Kothiya
- Yugm Patel

Homework Assignment: Clustering K-means + 1 Additional clustering method to your dataset(s)

Business Objective

The primary goal of this project is to maximize user engagement and health outcomes by providing personalized activity recommendations and optimizing activity schedules. Through clustering analysis of Fitbit activity data, the project aims to enhance user interaction with the device, which could lead to improved health results and increased customer satisfaction.

Task Overview

The task involves a collaborative effort where each team member applies different unsupervised clustering methods to the dataset to:

1. Discover inherent groupings within the user data.
2. Interpret what these groupings indicate about user behavior and preferences.
3. Utilize insights from the clustering analysis to propose solutions to business problems such as user retention and personalized marketing.

Datasets

Utilized the "**dailyActivity_merged.csv**" from the Fitbit Fitness Tracker Dataset.

Link: https://www.kaggle.com/datasets/arashnic/fitbit

This dataset includes a range of daily activity metrics for individual users, such as total steps taken, total distance covered, calories burned, and activity intensity levels categorized into very active, fairly active, lightly active, and sedentary minutes. Each entry likely corresponds to a single user's activity recorded over one day.

Methodology

Data Wrangling and Preparation

Source: Data was sourced from the "dailyActivity_merged.csv" file, containing extensive activity metrics such as total steps, distance, active minutes, and calories burned.

Preparation Steps:

1. **Load Data:** Data was initially loaded into a pandas DataFrame to evaluate its structure and integrity.
2. **Clean Data:** Procedures were implemented to handle missing values, correct data anomalies, and eliminate redundancies.
3. **Transform Data:** The 'ActivityDate' was converted from string format to datetime to facilitate time-series analysis.
4. **Feature Engineering:** Potential new features were considered to better represent underlying patterns for clustering.
5. **Data Reduction:** Scaling techniques were applied to ensure uniform contribution of each attribute to the clustering process.

Clustering Analysis

Four clustering algorithms were applied to discover intrinsic groupings within the data:

1. **K-Means Clustering:** Implemented using TensorFlow, Scikit-learn, and PyTorch to identify user groups based on activity patterns. The optimal number of clusters was determined using the Elbow Method.
2. **DBSCAN:** Used to identify clusters with varying densities and shapes, including handling outliers that K-Means may overlook.
3. **Spectral Clustering:** Employed to capture complex cluster structures that are not necessarily globular, which might be missed by K-Means.
4. **Hierarchical Clustering:** Provided insights into the hierarchical organization of the data, helping to understand the multilevel structure of user behaviors.

Results and Interpretation

- **Cluster Characteristics:** Clusters varied significantly across user activity levels from sedentary to highly active, distinguished by metrics like 'VeryActiveMinutes' and 'Calories'.
- **Business Implications:** The clustering results are crucial for:
    - **Personalizing Activity Recommendations:** Customizing suggestions to enhance user engagement and promote healthier lifestyles.
    - **Optimizing Activity Schedules:** Determining optimal activity times to increase the effectiveness of fitness routines.

Visualizations and Further Analysis

- **Feature Distribution Across Clusters:** Visualizations such as box plots and scatter plots clarified the differences among user groups based on their activity metrics.
- **Cluster Validation:** Consistency and stability of clusters were evaluated by comparing outcomes across different clustering methods.

Conclusion

The project's comprehensive analysis of Fitbit data through multiple clustering techniques has yielded valuable insights into user engagement and activity patterns. These insights facilitate the development of targeted fitness programs and enhance user interaction with health-tracking technologies. Future directions include integrating more sophisticated models and expanding the data sources to enhance the predictive capabilities and user segmentation.