Assignment 9 report - Unsupervised Learning Analysis Report

Introduction

This report presents the findings from an unsupervised learning analysis conducted on the California Housing dataset. The objective was to apply clustering algorithms, perform dimensionality reduction, evaluate clustering performance, and interpret emerging patterns. The dataset comprises 20,640 records and 9 numerical features describing housing and demographic characteristics in California.

Dataset Exploration

The dataset includes features such as Median Income (MedInc), House Age, Average Rooms, Average Bedrooms, Population, Average Occupancy, Latitude, Longitude, and the target variable Median House Value (MedHouseVal). All variables are numerical, and no missing values were detected, as shown by the exploratory output (0 missing values across all columns).

Data Preprocessing

Since all features were numerical, feature scaling was performed using StandardScaler to standardize the distributions before applying clustering algorithms. The scaled dataset contained 20,640 rows and 8 predictor features.

Clustering Analysis

Two clustering methods were applied: K-Means and Agglomerative Clustering. To determine the optimal number of clusters, the Elbow Method and Silhouette Score were computed for k values between 2 and 10. K-Means achieved its highest Silhouette Score of 0.3355 at k = 4, which was selected as the optimal cluster count. The inertia values decreased steadily with increasing k, consistent with typical clustering behavior.

Cluster profiling revealed distinct patterns across the four K-Means clusters. For example:

- Cluster 0: Moderate median income, average house age, and medium house values.

- Cluster 1: Slightly higher median house value and lower latitude (southern California).

- Cluster 2: Highest population density and unusually high occupancy rates.

- Cluster 3: Largest average rooms and bedrooms, indicating areas with larger homes.

Agglomerative Clustering Results

Agglomerative Clustering was also performed using the same number of clusters (k = 4). The silhouette score obtained was 0.3232, slightly lower than the K-Means score (0.3355), indicating that K-Means produced better-defined clusters. The cluster profiles from Agglomerative Clustering followed similar patterns to K-Means but with slight variations.

Dimensionality Reduction

Principal Component Analysis (PCA) was applied, reducing the dataset from 8 dimensions to 2 components for visualization. The first two principal components explained 48.85% of the total variance. Visualizations of K-Means and Agglomerative cluster assignments in PCA space revealed overlapping but distinguishable cluster regions. t-SNE was also used for non-linear dimensionality reduction, providing clearer separation for some clusters.

Model Evaluation

Silhouette scores were used to evaluate cluster quality. K-Means outperformed Agglomerative Clustering, reinforcing its selection as the primary clustering method. Cluster profiles were consistent and showed meaningful segmentation, such as identifying areas with significantly different median income and population densities.

Deployment and Monitoring Considerations

If deployed in a real-world scenario, such as a real estate analytics platform, the clustering model could help segment geographical areas based on demographic and housing characteristics. Deployment challenges include scalability, data drift, and ongoing maintenance. To ensure reliability, regular monitoring of feature distributions, silhouette scores, and periodic model retraining is recommended.

Conclusion

The unsupervised learning analysis successfully identified meaningful clusters within the California Housing dataset. K-Means with 4 clusters demonstrated the best performance. Dimensionality reduction enhanced the understanding of cluster structure, and deployment considerations were outlined to ensure long-term model effectiveness.