Unsupervised Learning Using K-Means Clustering on California Housing Data

Mulubrhan Abebe Nerea
MSc in Artificial Intelligence and Automation
University west, sweden
mulubrhan-abebe.nerea@student.hv.se

***Abstract*—This report evaluates unsupervised learning methods, particularly K-Means Clustering and DBSCAN, on the California Housing dataset. Our goal is to segment housing data according to geographic features, like longitude and latitude, as well as economic indicators, like median_income. We applied K-Means clustering with different numbers of clusters, figuring out the best configuration based on silhouette scores, which determined that k=2 was the best choice for clear data separation; DBSCAN, on the other hand, found 4 clusters, indicating its capacity to handle noise and identify non-spherical clusters. The analysis highlights the strengths and weaknesses of both approaches in housing market segmentation, and the knowledge gathered can be used to guide real estate marketing, housing development strategies, and urban planning.**

***Index Terms*—Unsupervised Learning, Clustering, MNIST, silhouette, DBSCAN, machine learning**

The code is available at https://github.com/Mulubrhan21/KMeans_MNIST.

I. Introduction

Clustering is an essential tool in data science, aimed at identifying group structures within datasets based on similarity. It maximizes similarity within clusters and dissimilarity between them. Hierarchical clustering was one of the first methods used by biologists and social scientists, leading to the development of cluster analysis as a branch of statistical multivariate analysis [1] [2]. Unsupervised learning techniques like K-Means Clustering play a crucial role in data exploration where labels are not available [3]. By partitioning the dataset into clusters, we can uncover hidden patterns that provide deeper insights into the data. This report aims to apply K-Means and DBSCAN to the California Housing dataset to explore segmentation based on geographic location (longitude, latitude) and economic factors (median_income). We will optimize the number of clusters using the silhouette score and compare the performance of K-Means with DBSCAN, a density-based clustering method, to better understand the distribution and characteristics of different housing clusters.

II. K-Means Clustering and Evaluation

We applied K-Means clustering to the California Housing dataset, utilizing the features of longitude, latitude, and median income. To optimize the number of clusters, we computed silhouette scores for various values of k, ranging from 2 to 9. The silhouette scores peaked at k=2 with a value of 0.548, indicating that this configuration provided the most distinct separation between clusters. This result suggests a clustering solution that effectively captures the inherent geographical and income distribution characteristics of the dataset. In addition to silhouette scores, we employed the Elbow Method to further determine the optimal number of clusters. This method involves calculating the Within-Cluster Sum of Squares (WCSS), which measures the total variance within each cluster by summing the squared distances between data points and their respective cluster centroids. By plotting WCSS against different values of k, we identified the ”elbow point,” where the rate of improvement diminishes. As we see in the Figure 1 the optimal number of clusters is k=2, as the plot showed but also a significant decrease in WCSS when k=3 followed by a plateau. This indicates that while increasing the number of clusters improves clustering quality, the marginal gains become minimal beyond this point. Preprocessing: The data was standardized before applying K-Means, which helped balance the influence of features.

Table I
Silhouette Scores for Different Values of k

k (Number of Clusters)	Silhouette Score
2	0.5482
3	0.5177
4	0.4351
5	0.3870
6	0.3985
7	0.3705
8	0.3525
9	0.3562

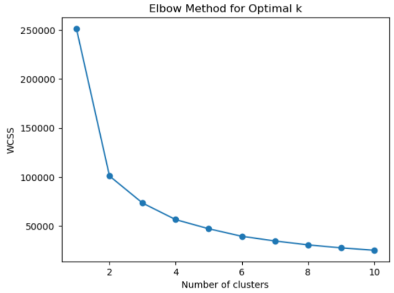

Figure 1. Elbow method of Clustering

III. Results and Analysis

The results from applying K-Means clustering and DBSCAN to the California Housing dataset reveal distinct clustering patterns, which offer insights into geographic income distribution.

A. K-Means Clustering Results

TableII summarizes the statistics for the clusters derived from K-Means with k=2, showing the differences in median income between the two clusters.

Table II
Cluster Statistics for Median Income

Cluster	Number of Data Points	Mean Median Income	Min Income	25th Percentile	Median Income	75th Percentile	Max Income
0	8,704	3.80	0.50	2.53	3.46	4.67	15.00
1	11,936	3.92	0.50	2.59	3.59	4.81	15.00

As shown in Table II Cluster 0 consists of 8,704 data points, with a mean median income of 3.80, representing regions with slightly lower-income households. Cluster 1, comprising 11,936 data points, has a mean median income of 3.92, indicating wealthier regions. As shown in Figure 2, the histogram analysis of median income in both clusters revealed a small but meaningful difference in income levels. This segmentation reflects geographical income distribution patterns, which could be useful for identifying regions requiring economic intervention or housing policy changes.

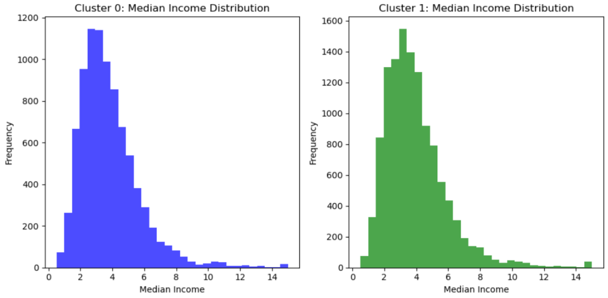

Figure 2. Histogram Income Distrubtion for the two clusters

B. Comparison with DBSCAN

We also applied the DBSCAN algorithm to the dataset, which identified 4 clusters based on data density. Unlike K-Means, DBSCAN does not require the number of clusters to be predefined and can handle noise effectively. The key comparison points between the two methods are:

As shown in Figure 3 DBSCAN identified 4 clusters, while K-Means with k=2 yielded 2 clusters. This difference reflects the distinct characteristics of the two algorithms. First, DBSCAN has the ability to handle noise by marking some data points as outliers, labeled as -1, whereas K-Means assigns all data points to clusters. This allows DBSCAN to provide more detailed segmentation, particularly in datasets with varying densities. Additionally, DBSCAN can find clusters of arbitrary shape, whereas K-Means assumes spherical clusters, which limits its flexibility. As a result, DBSCAN was able to detect smaller, dense clusters that K-Means missed. Finally, the granularity of DBSCAN's 4 clusters, compared to K-Means' 2, demonstrates that DBSCAN offers a finer level of detail, potentially uncovering more specific housing market segments. This makes DBSCAN more suitable for detecting localized patterns, especially in areas with high data density or when noise is present.

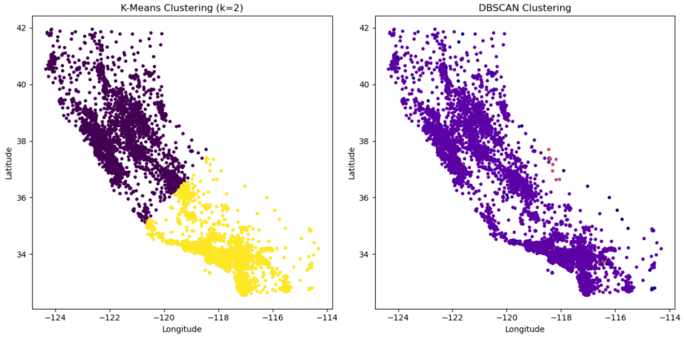

Figure 3. comparison silhouette and DBSCAN of Clustering

IV. Conclusion

The K-Means Clustering algorithm successfully divided the California Housing dataset into two broad clusters based on geographical location and median income. In addition to producing an easy-to-understand interpretation, selecting k=2 maximized the silhouette score, showing distinct cluster separation. Although K-Means offers simplicity and efficiency and is well-suited for large-scale segmentation, the comparison with DBSCAN showed that it may miss complicated data structures. DBSCAN demonstrated its capacity to deal with noise and odd forms by identifying four clusters. K-Means is still a good option for simple applications, though, as it produces findings that are easy to understand while yet delivering insightful information about how to segment the housing market.

References

[1] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *Department of Applied Mathematics, Chung Yuan Christian University*, 2023.

[2] Scikit-learn, "Support vector machines," https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html, 2024, accessed: 2024-10-21.

[3] A. Geron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2019.