Unsupervised Learning Using K-Means Clustering on California Housing Data

Zahra Mosavi
Master in AI and Automation
University West
Trollhättan, Sweden
zamo0008@student.hv.se

Masoumeh Mosavi
Master in AI and Automation
University West
Trollhättan, Sweden
mamo0084@student.hv.se

Divine Ezeilo
Master in AI and Automation
University West
Trollhättan, Sweden
diez0001@student.hv.se

***Abstract*—This assignment focuses on exploring unsupervised learning techniques using the K-Means clustering algorithm applied to the California Housing dataset. The objective was to group housing data based on geographic and socioeconomic similarities, specifically using longitude, latitude, and median income as features. Several K-Means models were trained with different numbers of clusters, and the Silhouette Score was used to evaluate the clustering quality and determine the optimal cluster configuration. The results showed that k = 2 achieved the highest Silhouette Score of 0.55, followed by k = 3 with a score of 0.525, indicating that two clusters provide the best balance of cohesion and separation. Visualization of the clustering revealed clear geographic patterns, with one cluster representing higher-income coastal regions and the other representing lower-income inland regions. A comparative analysis with DBSCAN showed that K-Means produced more meaningful and interpretable clusters, confirming its suitability for continuous spatial data such as the California housing dataset.**

***Index Terms*—unsupervised, cluster, K-mean, Silhouette Score,**

I. Introduction

Unsupervised learning is a type of machine learning that involves the identification of concealed patterns or clusters in data without the use of predetermined labels [1]. One of the most widely used algorithms in this category is K-Means clustering [2], which partitions data into k distinct groups based on feature similarity. The algorithm iteratively assigns data points to the nearest cluster centroid and updates the centroids until convergence, minimizing the within-cluster variance. Its simplicity, efficiency, and scalability make K-Means a foundational method for exploratory data analysis and pattern recognition [1], [2].

This assignment uses the California Housing dataset [3] and K-Means clustering to find significant correlations between socioeconomic characteristics and geographic location. Using characteristics like longitude, latitude, and median income, the goal is to find natural groupings that correspond to various California home market regions. By utilizing the Silhouette Score to evaluate these clusters, one may estimate the ideal number of clusters and evaluate the quality of the clustering. K-Means is useful for comprehending real-world data distributions, as this research shows by providing insights on regional housing trends, such as the differences between coastal high-income and interior low-income areas.

II. K-Means Clustering and Evaluation

The California Housing dataset was preprocessed to guarantee consistent feature scaling and enhanced model performance prior to the application of K-Means clustering. The spatial and socioeconomic dimensions of the housing data were represented by the selected features, which included longitude, latitude, and median income. StandardScaler from scikit-learn was employed to standardize all features, resulting in a zero mean and unit variance, as K-Means is sensitive to feature magnitude [2], [4]. This step ensured that no single attribute dominated the clustering process due to scale differences [3]. This step ensured that no single attribute dominated the clustering process due to scale differences.

The K-Means algorithm was then applied with various values of k (number of clusters), ranging from 1 to 9, to explore different clustering configurations. For each configuration, the Silhouette Score [5] was calculated to evaluate cluster quality, measuring how well each data point fit within its assigned cluster compared to other clusters. The analysis revealed that k = 2 achieved the highest Silhouette Score of 0.55, followed by k = 3 with a score of 0.525. Based on these results, k = 2 was selected as the optimal configuration, effectively dividing the dataset into two distinct and interpretable clusters representing different geographic and income regions across California.

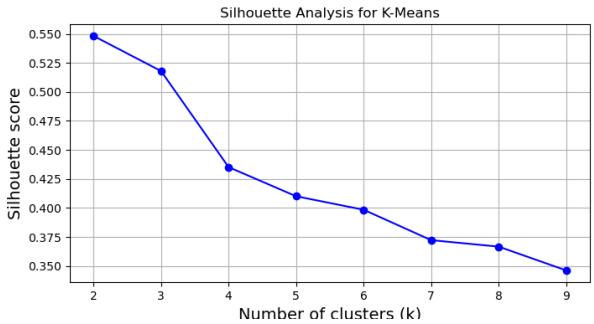

Fig. 1. Silhouette analysis for K-Means clustering with k values from 2 to 9. The highest score of 0.55 was achieved at $k = 2$, indicating the optimal number of clusters.

Fig. 2. K-Means clustering results on the California Housing dataset with $k = 2$, showing distinct geographic clusters based on longitude and latitude.

III. Results and Discussion

The K-Means clustering analysis produced clear and interpretable results on the California Housing dataset. The Silhouette Score evaluation revealed that k = 2 was the optimal number of clusters, achieving a score of 0.55, followed by k = 3 with a score of 0.525. This indicates that two clusters provide the best balance between cluster cohesion and separation as shown in figure 1.

The scatter plot of K-Means clustering figure 2 clearly shows the division between two major geographic regions. Cluster 0 primarily corresponds to coastal areas, characterized by higher median income, while Cluster 1 represents inland areas with lower median income.The K-Means decision boundary visualization figure 3 how the algorithm separates data based on distance to the cluster centroids, resulting in distinct, linearly separable regions.

When visualized against median income figure 3, the two clusters show a noticeable difference in income distribution, confirming that geographical location is strongly associated with economic variation across California.

A comparison with DBSCAN [6] figure 5 reveals key difference between the Kmean and DBSCAN algorithms. DBSCAN identified only one large cluster and several noise points, suggesting that the dataset does not have sharp density separations suitable for DBSCAN's density-based approach. On the other hand, the dataset was successfully divided into spatially coherent and significant groups using K-Means.

Overall, K-Means outperformed DBSCAN in terms of interpretability and clustering quality, producing distinct geographic and income-based patterns that reflect real-world housing market trends. The results highlight K-Means as a robust method for analyzing large-scale continuous data where clusters are well-distributed and relatively spherical.

IV. Conclusion

This study explored unsupervised learning techniques by applying K-Means and DBSCAN clustering algorithms to

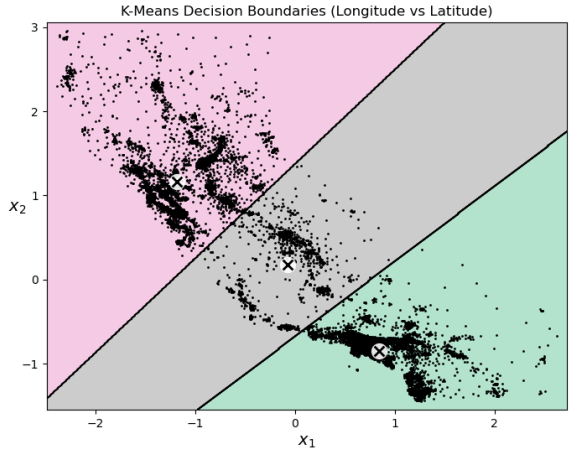

Fig. 3. K-Means decision boundaries illustrating how the algorithm separates data points into three clusters based on proximity to centroids.

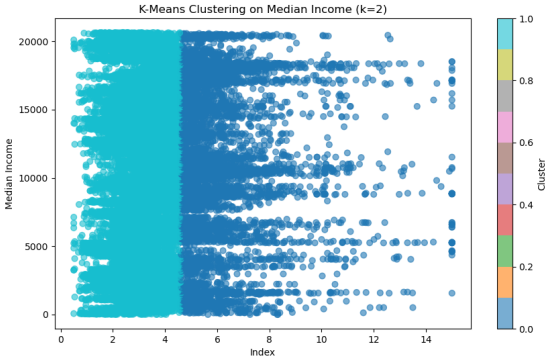

Fig. 4. Distribution of median income across the two K-Means clusters ($k = 2$), showing clear income differences between coastal and inland regions.

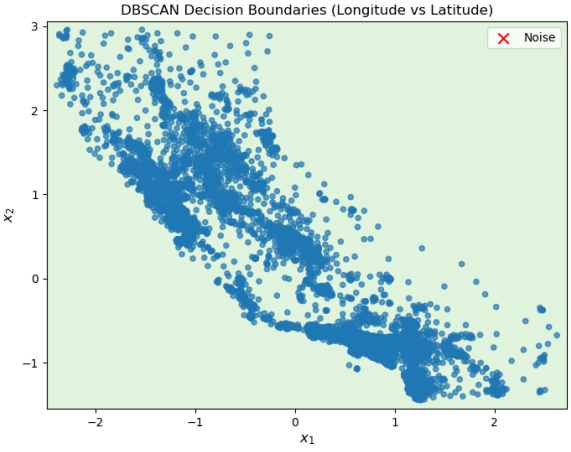

Fig. 5. DBSCAN decision boundaries showing one large cluster and scattered noise points, indicating limited cluster separation for this dataset.

the California Housing dataset to uncover geographic and socioeconomic patterns. Using longitude, latitude, and median income as input features, the goal was to identify natural groupings of housing data and assess the effectiveness of each clustering approach.

The Silhouette Score analysis indicated that K-Means performed best with two clusters (k = 2), achieving a score of 0.55, signifying strong and well-defined cluster separation. The resulting clusters revealed clear spatial and income distinctions — with one cluster representing coastal, higher-income areas, and the other representing inland, lower-income regions. Visualization of decision boundaries confirmed that K-Means created simple, interpretable partitions aligned with real-world geographic divisions.

DBSCAN, on the other hand, showed ineffectiveness for the continuous density distribution of this dataset, identifying only one significant cluster and several noise spots. In general, K-Means fared better than DBSCAN, offering more insightful and comprehensible information about the California real estate market. K-Means is a reliable and effective approach for finding structured patterns in big, continuous geographical datasets, according to the results.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[2] "K-Mean CLustering Algorithm," Scikit-learn Datasets, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.htmlk-means

[3] "California Housing Dataset," *Scikit-learn Datasets*, 2024. [Online]. Available: https://scikit-learn.org/stable/datasets/real_world.htmlcalifornia-housing-dataset

[4] C. Wongoutong, "The impact of neglecting feature scaling in k-means clustering," *PLOS ONE*, vol. 19, no. 12, p. e0310839, Dec. 2024. doi:10.1371/journal.pone.0310839.

[5] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, Portland, OR, USA, 1996, pp. 226–231.

[7] D. Nelson, M. Mosavi, Z. Mosavi, K Mean CLustering, GitHub repository, 2025. [Online]. Available: https://github.com/Divine-Nelson/kmean_clustering