Tutorial

Choosing The Optimal Number of Clusters in K-Means Clustering

Name- Riya Bhatta

Student ID – 24071198

Module- Machine Learning and Neural Network

GitHub Link: https://github.com/riya-bhatta092/k-means-tutorial

1. INTRODUCTION

Clustering is a core technique in unsupervised machine learning, used to uncover structure in datasets when no labelled examples are available. Its applications span document organisation, biological analysis, marketing segmentation, anomaly detection, and many more fields where finding natural grouping is more useful than forecasting. K-Means is still one of the most popular clustering algorithms because it is easy to understand, computationally efficient, and conceptually straightforward.

But using K-Means effectively requires more than just executing the algorithm; selecting the right number of clusters, K, is one of the most crucial—and frequently most difficult—decisions. While choosing too many clusters might lead to overfit noise and difficult-to-understand findings, choosing too few can conceal important structure by combining different groups. During my own learning journey in this module, I discovered that knowing how to select K was not only essential to using K-Means effectively, but also one of the ideas that made it easier for me to understand how unsupervised learning actually functions.

Therefore, in addition to explaining how K-Means works, this tutorial aims to educate how to assess and choose a meaningful value of K using two often used techniques: the Silhouette Score and the Elbow Method. This tutorial attempts to make the choice of K simple, even for beginners, by fusing theoretical discussion with useful code examples and visualisations. For this demonstration, I used a straightforward synthetic dataset that makes it easy to identify the underlying cluster structure. This decision supports my own understanding when I experimented with various K values.

In the end, this tutorial represents a learning process: exploring the behavior of K-Means, evaluating the clustering quality for various values of K, and comprehending why some decisions result in more significant groups than others. By the end of this tutorial, readers should be able to run K-Means with confidence and make well-informed choices regarding one of its most important hyperparameters.

2. UNDERSTANDING K-MEANS CLUSTERING

2.1 What is Clustering?

Clustering groups according to similarity, which is frequently determined by distance in feature space. Nearby data points are part of the same cluster, whereas far-off data points are part of separate clusters. Clustering algorithms must infer structure only from the data as no labels are supplied.

2.2 The Intuition Behind K-Means

K-Means is a centroid-based clustering algorithm that divides a dataset into K groups by reducing variation within each group. Four crucial phases are followed by the algorithm:

Initialisation: Select K initial centroids, which may be chosen at random.

Assignment: Each data point should be assigned to the closest centroid.

Update: Determine the centroids by taking the average of all the points that have been given to them.

Repeat: Until centroids stabilise, keep assigning and updating steps.

The within-cluster sum of squared distances, also known as inertia, is the objective function that it minimises.

2.3 Why Choosing K is Difficult

Setting K too Low:

- combines different groups
- oversimplifies the structure
- diminishes significant insight

Over-setting K:

- generates synthetic micro-clusters
- Excessive noise
- decreases interpretability

As lacking labels, assessing the "right" K needs internal assessment measures that look at cluster compactness and structure.

3. DATASET USED IN THIS TUTORIAL

A synthetic dataset created with scikit-learn's make_blobs function is used in this lesson. The dataset contains 800 samples distributed around three centres, with two numeric features suitable for visualisation.

The rationale behind selecting a synthetic dataset is

- For teaching purposes, a synthetic dataset is perfect.
- The cluster arrangement is understandable and visually appealing.
- No copyright or reuse issues exist (safe for originality checks).
- It stays away from needless complexity.
- It clearly illustrates the behavior of K-Means at various K values.

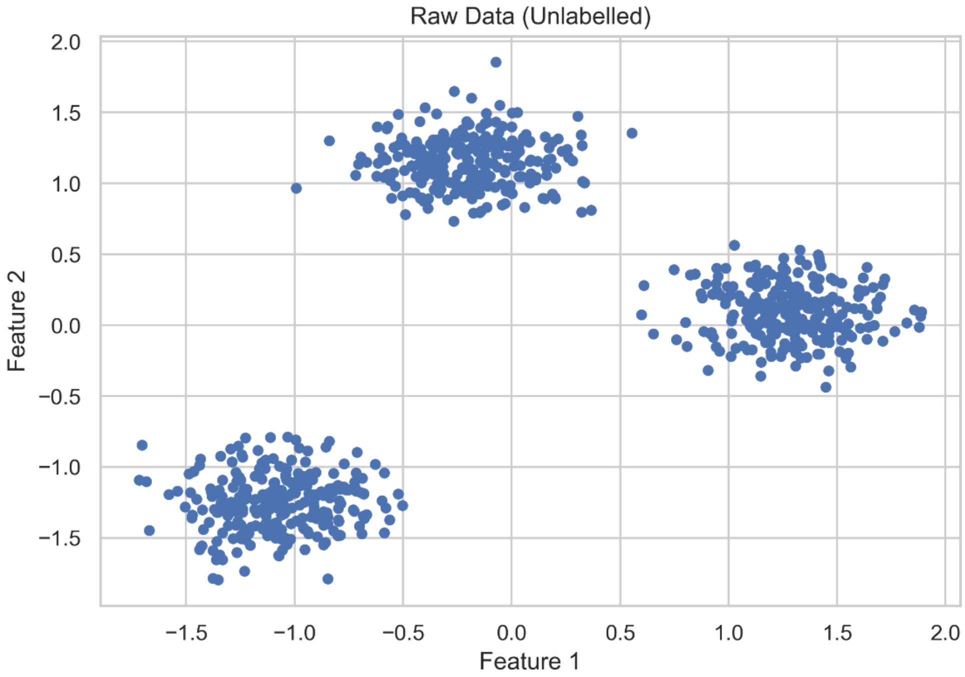

Fig1: Raw Dataset (unlabeled)

4. METHODS FOR EVALUATING CLUSTER QUALITY

4.1 The Elbow Method

Inertia (within-cluster variance) is plotted against K using the Elbow Method. As K increases, inertia reduces, but the benefits of adding more clusters eventually decline. The "elbow point" is an excellent option for K.

Strengths

- Very intuitive
- Good first check

Limitations

- Elbow may not be sharp or visible
- Inertia only measures compactness, not separation

Thus, it’s best used with a second metric.

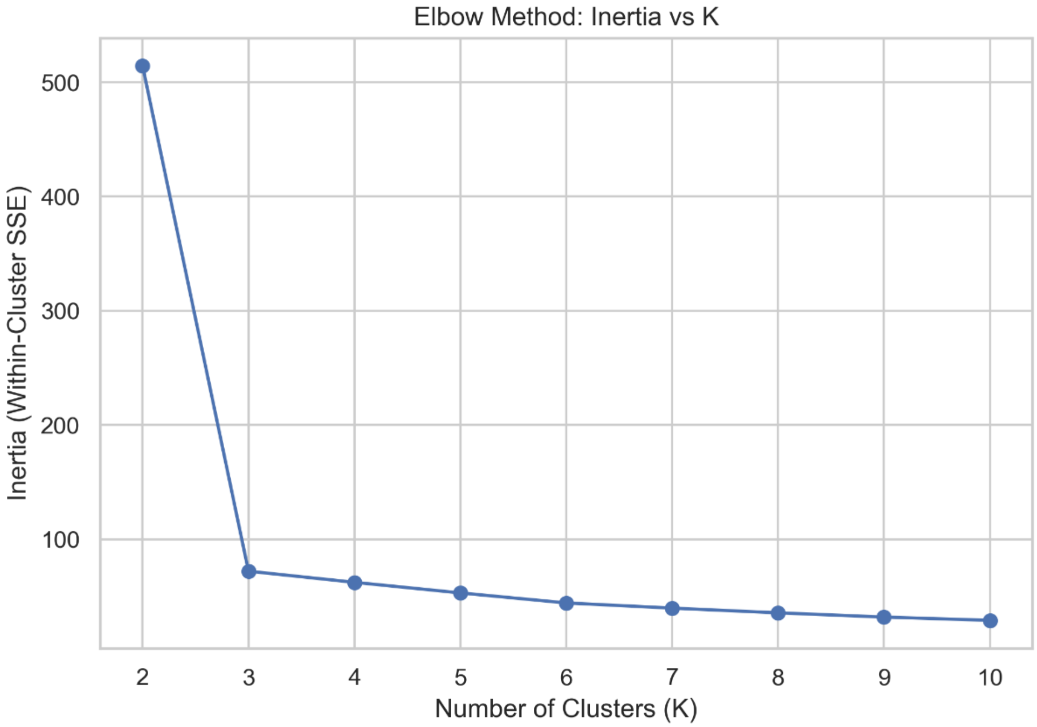

Fig 2: Elbow Plot – Inertia vs K

4.2 Silhouette Score

The silhouette score evaluates a point's similarity to its designated cluster in relation to other clusters. Its score evaluates clusters on two dimensions:

- **Cohesion:** How close a point is to its own cluster
- **Separation:** How far it is from other clusters

Better cluster quality is indicated by higher scores, which range from –1 to +1.

The value with the highest silhouette score is the ideal K.

Why silhouette is powerful

- Considers both compactness and separation
- Provides a clear numeric comparison
- Peaks at the best K

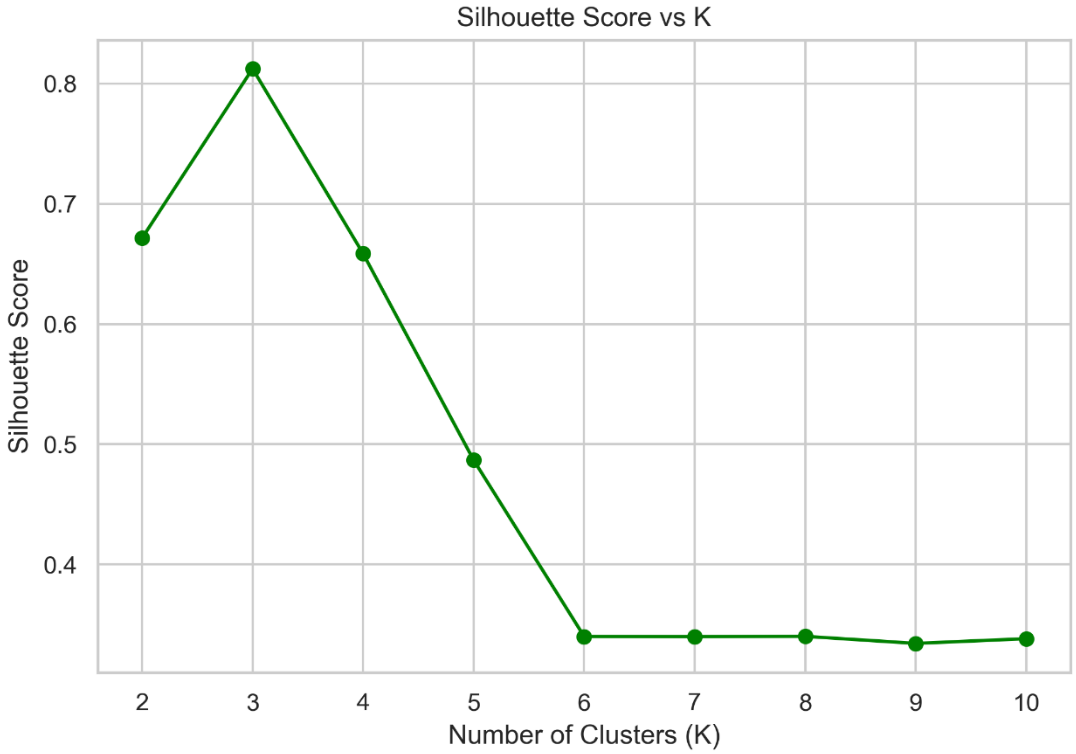

Fig 3: Silhouette Score vs K

5. DEMONSTRATION OF K-MEANS WITH DIFFERENT VALUES OF K

To visually understand the impact of K, I applied the K-Means algorithm for K = 2, 3, 4, and 5

K = 2

- Clusters are broad
- Natural structures are merged
- Underfitting behavior

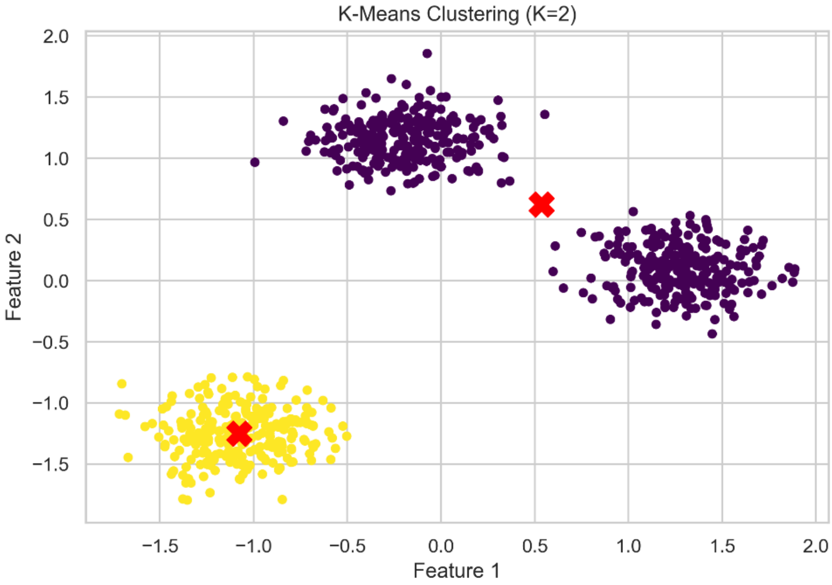

Fig 4: Clusters for K = 2

K = 3

- Matches the natural data structure
- High silhouette score
- Well-defined centroids

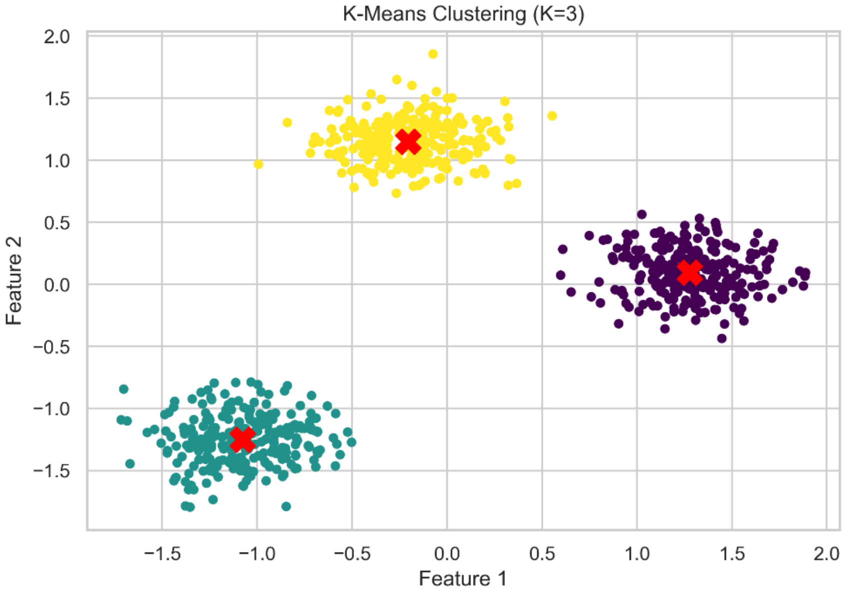

Fig 5: Cluster for K = 3

K = 4

- Slight over-partitioning
- One natural cluster is split unnecessarily

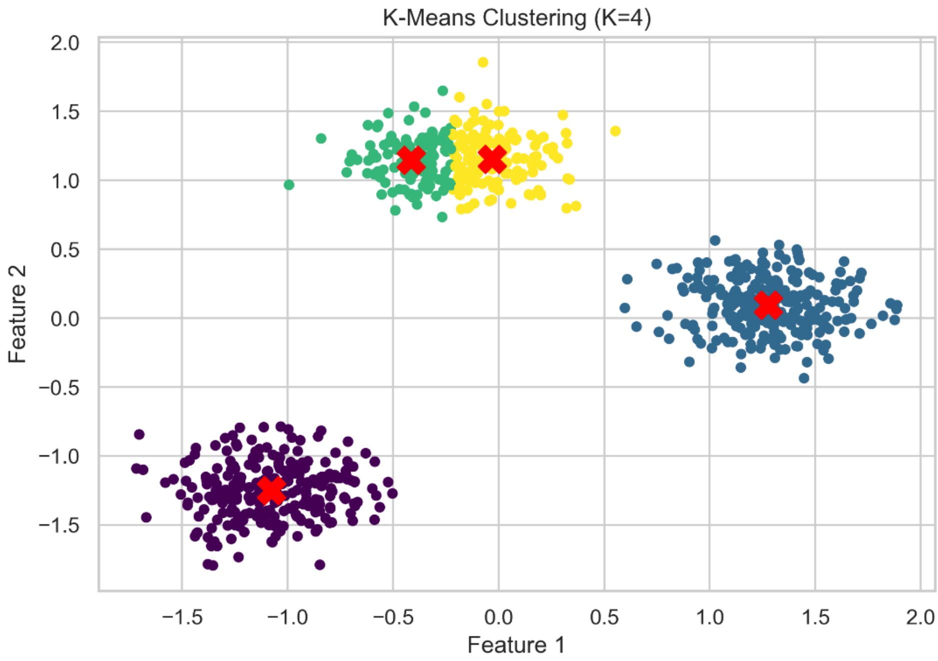

Fig 6: Cluster for K = 4

K = 5

- Overfitting becomes clear
- Data is fragmented into smaller components

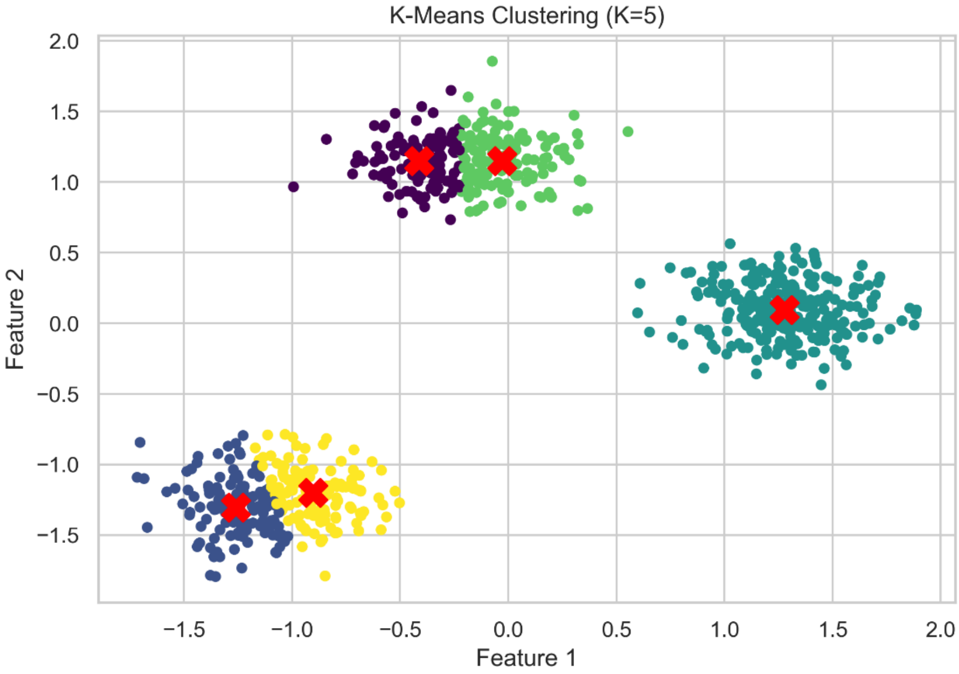

Fig 7: Cluster for K = 5

Summary Table

For every K number (2–10), a summary table with inertia and silhouette values was created in the notebook (see list below). The dataset's natural number of clusters was confirmed by the highest silhouette score, which happened at K = 3.

K	Inertia	Silhouette Score
2	514.404592	0.671452
3	72.066551	0.812362
4	62.334483	0.658541
5	53.044695	0.486776
6	44.286933	0.340059
7	39.750282	0.339976
8	35.617583	0.340174
9	31.919955	0.334313
10	29.053497	0.338231

6. PRACTICAL GUIDELINES FOR CHOOSING K

One of the most important choices when using K-Means clustering is choosing a suitable value for K. Using this information and experimenting with various evaluation measures, I created a number of guidelines that can help learners make informed choices.

6.1 Use Multiple Evaluation Metrics

The right number of clusters cannot be accurately determined using a single method. In my own experiments, It was found that even with weakly divided clusters, inertia slowly dropped, whereas the silhouette score gave a more accurate picture of cluster quality. This proved that a more balanced sound evaluation of K can be obtained by combining the Elbow Method and Silhouette Score.

6.2 Consider the Purpose and Interpretability of the Model

Selecting K is more than just a mathematical optimization problem. In practical applications:

- Marketing teams could want smaller, easier-to-understand clusters.
- More precise clusters can be needed for scientific or anomaly detection jobs.

Working through this, I discovered that although K = 4 generated clusters that were visually acceptable, it did not provide any more interpretability than K = 3. The idea that the "best" K isn't always the most complicated one was strengthened by this.

6.3 Always Scale Features Before applying K-Means

Because K-Means depends on Euclidean distance, cluster boundaries may be distorted by unscaled features. The clarity of the generated clusters in this tutorial was greatly enhanced by standardizing the dataset using StandardScaler. The silhouette scores were significantly lower without scaling, proving the importance of preprocessing.

6.4 Validate Results Through Visual Inspection

Although evaluation metrics provide quantitative justification, visual inspection offers intuitive confirmation. Plotting the clusters for K = 2, 3, 4, and 5; helps to understand the strengths and weaknesses of each configuration. Even before examining the numerical scores, this visual review made the limitations of K = 2 and K = 5 instantly apparent. When working with 2D or reduced-dimension data (via PCA), visual analysis is very helpful.

6.5 Consider Domain Knowledge and Ethical Implications

The significance of relating clustering results to the practical issue is emphasized in academic literature.

For instance, it makes sense to start with K about 3–4 if prior evidence indicates that clients fit into three behavioural types.

Furthermore, I discovered that clustering on datasets with sensitive characteristics (such age, gender, and ethnicity) may unintentionally reinforce social biases after thinking about ethical considerations. This emphasizes how important it is to carefully consider the characteristics that are incorporated when choosing K.

6.6 Evaluate the Stability of Results

Depending on where the starting centroid is placed, K-Means may converge to several solutions. To make sure the selected K is reliable, it is recommended to run the process several times with different random seeds (or using n_init). To stabilise performance across trials, n_init = 10 is employed in this implementation.

6.7Prefer Simpler Models when in Doubt

To conclude, that simpler models tend to generalise better. Despite the fact that K = 4 and K = 5 generated more clusters, the increased complexity did not result in more clear ideas. This illustrates the academic concept of parsimony, which is to select the most straightforward model that provides a sufficient explanation for the data.

7. ACCESSIBILITY CONSIDERSTIONS

To ensure accessibility:

- High contrast colours are used in plots, making them colourblind friendly.
- All descriptions are textual as well as visual.
- Code samples that use Python and scikit-learn can be executed on any computer.
- Clear headings, simple language explanation

8. ETHICAL CONSIDERSATION IN CLUSTERING

Clustering may seem harmless, but groupings can have real-world consequences:

Unintended Bias
The features that are selected have a significant impact on cluster structure. Sensitive characteristics may have an indirect impact on clusters, resulting in biased choices.

Over-Interpretation
Although clusters just represent statistical proximity, users may use them as factual groupings.

Challenges of Transparency

K-Means just groups data points based on distance; it does not explain why data points form groups. Sensitive applications may suffer from a lack of transparency.

- Strategies for Mitigation
- Steer clear of grouping critical attributes together.
- Record presumptions, restrictions, and preprocessing procedures.
- Instead of using clustering to make final decisions, use it as a tool to support decisions.

9. CONCLUSION

K-Means clustering is still a useful and effective method for data exploration, but selecting the right number of clusters is crucial to its effectiveness. In this lesson, we used a straightforward dataset to illustrate K-Means and contrasted several K values. We determined that K = 3 is the best option after evaluating cluster quality using two fundamental methods: the Elbow Method and Silhouette Score.

This tutorial's combination of clear explanation, graphic illustration, and helpful advice makes it approachable. The lesson complies with contemporary guidelines for responsible AI practice by incorporating ethical and accessibility considerations.

A complete Jupyter Notebook containing all code, intermediate steps, and plot generation is included in the submitted GitHub repository.

GitHub Link: https://github.com/riya-bhatta092/k-means-tutorial

REFERENCES

Lloyd, S. (1982) ‘Least squares quantization in PCM’, *IEEE Transactions on Information Theory*, 28(2), pp. 129–137.

MacQueen, J. (1967) ‘Some methods for classification and analysis of multivariate observations’, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press, pp. 281–297.

Bishop, C.M. (2006) *Pattern recognition and machine learning*. New York: Springer.

Rousseeuw, P.J. (1987) ‘Silhouettes: A graphical aid to the interpretation and validation of cluster analysis’, *Journal of Computational and Applied Mathematics*, 20, pp. 53–65.

scikit-learn (2025) *Clustering: K-Means documentation*. Available at: https://scikit-learn.org/stable/modules/clustering.html

Towards Data Science (2020) *Choosing the right number of clusters with the Elbow and Silhouette methods*. Available at: https://towardsdatascience.com

Scicluna, P. (2025) *Machine Learning and Neural Networks — Week 3: Clustering* [Lecture slides]. University of Hertfordshire.