**Chapter 5 Unsupervised Learning**

The first step to creating a predictive model for battery energy density was the application of unsupervised learning techniques to the dataset of battery materials. The primary objective was to explore the inherent structure within the dataset by identifying natural groupings of materials without relying on predefined labels. Through dimensionality reduction and clustering algorithms, this chapter aims to uncover insights into the characteristics that differentiate battery materials and evaluate the strengths and weaknesses of each method used.

**5.1. Dimensionality Reduction**

**5.1.1 Principal Component Analysis (PCA)**

A graph with a line

AI-generated content may be incorrect.Principal Component Analysis was employed to simplify the dataset by transforming the original variables into a smaller number of uncorrelated components while retaining most of the original variability. The dataset included five numerical features: Molecular Weight, Capacity per Gram (mAh/g), Voltage (V), Efficiency (%), and Energy Density (Wh/kg). Before applying PCA, all variables were standardized to ensure that each feature contributed equally to the analysis.

The PCA results revealed that the first two principal components explained around 60% of the total variance in the data. This strong cumulative variance justified the decision to reduce the dimensionality to two components for visual analysis. The third component adds only ~19% more — which is a smaller gain compared to the first two. A third component would require 3D plotting, making it harder to visually communicate patterns or clusters. Reducing to 2 dimensions simplifies clustering or classification tasks, helping avoid overfitting and reducing noise from less informative components. The 2D PCA scatter plot showed discernible clusters, suggesting that materials with similar electrochemical behaviors were naturally grouped together. Notably, the first component was most influenced by Capacity and Energy Density, while the second component reflected variation in Efficiency and Voltage. These axes provided A graph with green and orange dots

AI-generated content may be incorrect.an interpretable projection of high-dimensional data, facilitating downstream clustering.

**5.1.2 Factor Analysis**

To investigate the latent structures potentially driving the correlations among features, Factor Analysis was conducted using two underlying factors. This statistical technique aims to identify latent constructions that explain observed correlations, which may correspond to broader conceptual groupings like energy output potential and operational stability.

A chart with numbers and symbols

AI-generated content may be incorrect.The factor loadings indicated that Factor 1 was strongly aligned with Energy Density and Capacity, suggesting an underlying trait related to energy storage capacity. Factor 2 was more strongly associated with Efficiency and Voltage, potentially representing operational performance or conversion efficiency. While Factor Analysis enhanced the interpretability of variable interrelationships, it did not yield as clear a visual separation as PCA. Nevertheless, the insights from this method reinforced the thematic distinctions in the dataset and validated the choice of variables used in clustering.

**5.2. Clustering Techniques and Evaluation**

**5.2.1 K-Means Clustering**

K-Means clustering was first applied to identify groups of materials with similar feature profiles. This method partitions the dataset into non-overlapping clusters by minimizing intra-cluster variance. To determine the optimal number of clusters, both the Elbow Method and Silhouette Score were employed. The Elbow plot showed a pronounced bend at four clusters, which corresponded to the peak silhouette score, indicating optimal clustering structure.

A graph of different types of graphs

AI-generated content may be incorrect.

The resulting four clusters, when projected onto the PCA space, were well-separated and formed distinct groupings. Each cluster could be interpreted based on dominant properties such as high energy density or exceptional efficiency. However, K-Means assumes that clusters are spherical and of similar size, which may not always reflect real-world variability. In this case, the assumption was reasonable, as evidenced by the clustering quality and coherence across multiple dimensions.

**5.2.2 Hierarchical Clustering**

A diagram of a clustering graph

AI-generated content may be incorrect.Agglomerative Hierarchical Clustering using Ward’s linkage was applied to gain insights into the nested structure of the dataset. The method iteratively merged clusters to minimize within-cluster variance. The dendrogram generated from this process suggested a natural division into four clusters, in agreement with the K-Means results.

Although the silhouette score for Hierarchical Clustering was slightly lower than that of K-Means, the cluster formation process added interpretability. For instance, some clusters merge late in the hierarchy, suggesting stronger dissimilarity and supporting their distinction in the final partition. A notable limitation of this method is its computational complexity and inability to reassign observations once merged. Still, it served as a valuable validation tool and offered a layered understanding of cluster relationships.

**5.2.3 DBSCAN**

The DBSCAN algorithm was used to explore the data's density-based structure and to identify outliers. Unlike K-Means or Hierarchical Clustering, DBSCAN does not require the number of clusters to be predefined and can discover arbitrarily shaped clusters. A k-distance graph helped identify a suitable eps value, with 2 being selected based on the observed elbow point. A minimum sample size of 5 was used, reflecting common practice.

A graph with a line

AI-generated content may be incorrect.While DBSCAN detected a few meaningful clusters, it also identified a significant number of points as noise or outliers. These included materials with extreme property values or those not conforming to dominant patterns. The silhouette score for DBSCAN was higher than the other clustering algorithms, mainly due to the exclusion of noise from the ***prime-dataset***. The noise could be the data points that are extreme in nature. The method's strength lies in its robustness to outliers and flexibility in capturing non-linear cluster boundaries. For datasets with more irregular group structures or greater density variation, DBSCAN could provide superior insights.

**5.3 Critical Evaluation**

Each technique applied offered unique benefits and challenges. PCA emerged as an indispensable tool for data visualization and provided a highly interpretable two-dimensional projection of the dataset. The clarity of group separations in PCA plots justified its use as a basis for visualizing clustering results. Factor Analysis contributed by revealing 2 factors that contribute to the underlying variance of the data, though it was more valuable for interpretation than for visualization or clustering support.

Among the clustering algorithms, DBSCAN was the most effective in terms of silhouette score and interpretability. Unlike others it did not have a requirement for a predefined number of clusters. It also effectively dealt with noise (extreme outliers) Hierarchical Clustering closely mirrored K- Means results and provided additional context through its dendrogram, although its scalability remains a concern for larger datasets. DBSCAN, while being the most effective of the three, captured the broader cluster structure and was successful in flagging unusual or unique materials and may be particularly useful in anomaly detection contexts.

Overall, the consistency of findings across PCA, DBSACN, and Hierarchical Clustering enhances confidence in the results. The choice of 2 factors appears robust and well-supported by both quantitative and qualitative assessments. Similarly, DBSCAN algorithm revealed that the 2 clusters formed are well defined and well separated. Each method complemented the others, contributing to a more comprehensive understanding of the dataset.

**5.4 Conclusion**

The application of unsupervised learning techniques to the battery materials dataset revealed a well-defined structure, with materials naturally falling into four clusters based on shared physicochemical properties. The alignment between PCA visualizations and clustering outcomes highlighted the coherence of the data. Factor Analysis deepened the interpretability of these findings by suggesting latent factors driving material behavior.

K-Means and Hierarchical Clustering emerged as the most effective clustering methods, both identifying similar groupings. DBSCAN added value by highlighting potential outliers and offering an alternative, density-based perspective. The convergence of results across methods indicates a stable and meaningful cluster structure, which provides a strong foundation for further investigations.

These findings have practical implications for the classification and selection of battery materials. By identifying distinct material categories, researchers can better target experiments and optimize battery performance. Future work may build on these results through supervised learning approaches, time-dependent modeling, or integration with domain-specific knowledge to guide the development of next-generation energy storage materials.