Unsupervised Learning — Final Project

Roni Navon (ID 207858937)   Sharon Kamensky (ID 211576723)

April 2025

Abstract

The goal of this project was to explore hidden structures in the *Forest CoverType* dataset using unsupervised learning, without relying on the known forest type labels (*CoverType*). After applying dimensionality reduction techniques (PCA, UMAP, t-SNE), we ran several clustering algorithms. The combination of PCA with KMeans emerged as particularly effective, producing interpretable and stable clusters while preserving approximately 95% of the variance with just 14 components. We then compared the resulting clusters to the original labels using similarity metrics (NMI, ARI), feature importance analyses (Kruskal-Wallis, η^2), and internal evaluation measures (Silhouette). The results showed that the clusters revealed structures based on topographical and shading features, differing significantly from the original forest-type classification. This demonstrates the potential of unsupervised learning to uncover alternative and meaningful structures in complex data. The findings emphasize the power of unsupervised learning as an exploratory tool for discovering patterns in unlabelled data, particularly when human labels may reflect only partial aspects of the dataset's underlying structure.
The code is available at: GitHub Repository.

1 Introduction

Unsupervised learning focuses on identifying hidden and intrinsic structures within data, without relying on predefined labels. In this project, we aimed to explore whether meaningful partitions could be uncovered in the *Forest CoverType* dataset — which includes observations from forested areas in the United States with topographical, climatic, and ecological attributes — using dimensionality reduction and clustering techniques. Although the dataset contains classification labels (*CoverType*), we intentionally ignored them in the initial phase to enable an unguided discovery of internal patterns. A sample of 10,000 observations, preserving the original data distribution, was selected as the basis for applying dimensionality reduction methods (PCA, t-SNE, UMAP) and a variety of clustering algorithms (KMeans, DBSCAN, GMM). Beyond visual illustration and computational enhancement, the reduced-dimensional representations provided early indications of potential group structures and served as a tool for estimating the number of clusters. The comparison of approaches was based on a combination of criteria: internal metrics (such as Silhouette), similarity to true labels (such as NMI and ARI), stability across runs, graphical interpretability, and computational time — all aimed at identifying the most effective approach for revealing the dataset's underlying structure.

2 Methods

2.1 Dataset

This project was based on the publicly available *Forest CoverType* dataset, which contains over 580,000 observations describing forested areas in the United States. Each observation consists of 54 features, including continuous, binary, and categorical variables—such as topography, distances to water sources and roads, soil type, and hill shade at different times of day. Each instance is labelled with a ground truth (*Cover Type*) indicating the forest type, categorized into one of seven classes. Although the dataset includes these labels, they were excluded from the initial learning process and were used solely for evaluation purposes. The objective was to investigate whether meaningful groupings could be discovered in an unsupervised manner, relying solely on the data's intrinsic structure.

2.2 Preprocessing and Sampling

Prior to applying clustering and dimensionality reduction techniques, a systematic preprocessing pipeline was carried out. Continuous features were normalized to a uniform scale, while categorical variables (e.g., *Soil Type* and *Wilderness Area*) were already provided in one-hot encoded form, requiring no further transformation. Although some binary variables—especially those related to *Soil Type*—exhibited very low variance, they were retained intentionally, allowing PCA to determine their contribution. This contrasts with common practices in supervised learning, where such features are often removed in advance. For initial exploration, a normalized sample of 10,000 observations was selected and used for dimensionality reduction and visual embedding (t-SNE, UMAP). In a later stage, the analysis was extended to 80% of the dataset (approximately 465,000 instances), using stratified sampling by *Cover Type* to preserve the original distribution and avoid bias.

2.3 Outlier Detection Using Isolation Forest

To ensure that the cluster structure was not influenced by extreme values or outliers, we applied the Isolation Forest algorithm to the normalized data. The results revealed an exceptionally low outlier rate—only 0.1% (10 out of 10,000 observations). These anomalous observations were found to be significantly distant from the data centroid and located on the fringes, suggesting their minimal influence on the overall structure. This finding supports the reliability of the resulting clusters and reduces concerns about noise or edge-case effects distorting the analysis.

2.4 Dimensionality Reduction

In order to reduce dimensionality, improve performance, and enable visual accessibility of the data, three dimensionality reduction methods were employed: PCA, t-SNE, and UMAP. PCA was used to reduce the dataset to 14 components that preserved about 95% of the variance, while reducing dependencies among features. Nonlinear projection methods such as t-SNE were designed for visualization, allowing an intuitive identification of hidden structures such as cluster boundaries and outlier observations. The combination of these methods helped to assess the impact of the reduction on the structure of the data—both from a computational and a visual standpoint—thus supporting an informed choice for the primary method for further analysis.

2.5 Clustering

KMeans was used as the principal algorithm for unsupervised clustering, both after dimensionality reduction (via PCA) and on the full normalized dataset, to compare the group structures between the different spaces. The choice of KMeans was based on its computational efficiency, ease of application on large datasets, and its ability to provide clear and interpretable output. Additionally, KMeans was used at an earlier stage to estimate the optimal number of clusters as part of an independent exploratory process (see Section 2.6). The cluster label assigned to each observation was stored in the dataset, which then served as the basis for subsequent statistical analyses, comparisons with ground truth labels, and visual representations.

2.6 Determining the Optimal Number of Clusters

To understand the internal structure of the data and identify natural groupings without relying on labels, we conducted an analysis to determine the optimal number of clusters—a key step in unsupervised learning. For this purpose, we applied two common approaches: the Elbow Method, which examines the decrease in inertia, and the Silhouette measure, which assesses the quality of separation between clusters. In both approaches—applied to the original data and to the PCA-reduced data—the optimal number of clusters was found to be $k = 6$ (see Figure 1). Although the numerical result was identical, the performance was superior after dimensionality reduction: the Silhouette score was higher (0.154 compared to 0.144) and the runtime was significantly shorter (2.2 seconds compared to 3.4 seconds). While the difference in execution time is not substantial on the smaller sample, it may become significant when running on the full dataset. Additionally, the t-SNE visualizations generated for both versions demonstrated that dimensionality reduction not only preserved the overall data structure but even sharpened the cluster boundaries. Based on these findings, we decided to use the results from the PCA-transformed data for further data analysis and clustering. Since PCA retains about 95% of the variance, there is no concern about a significant loss of important information; on the contrary, the dimensionality reduction contributed to enhancing the clarity, interpretability, and efficiency of the clustering process.

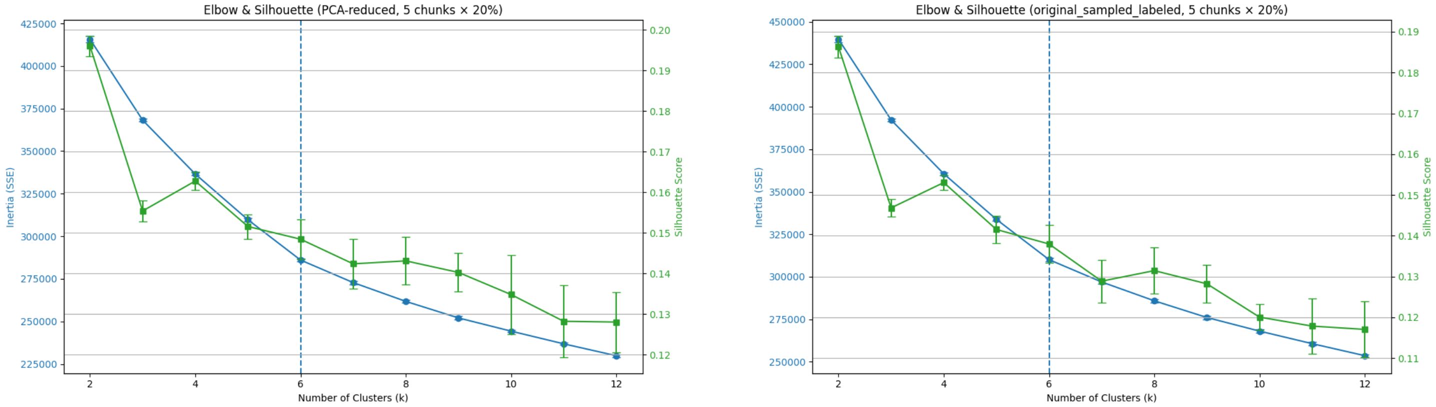

Figure 1: Comparison of Elbow and Silhouette results before and after dimensionality reduction. **Left:** Results on the PCA-reduced dataset. **Right:** Results on the original normalized dataset. Both approaches indicate $k = 6$ as the optimal number of clusters.

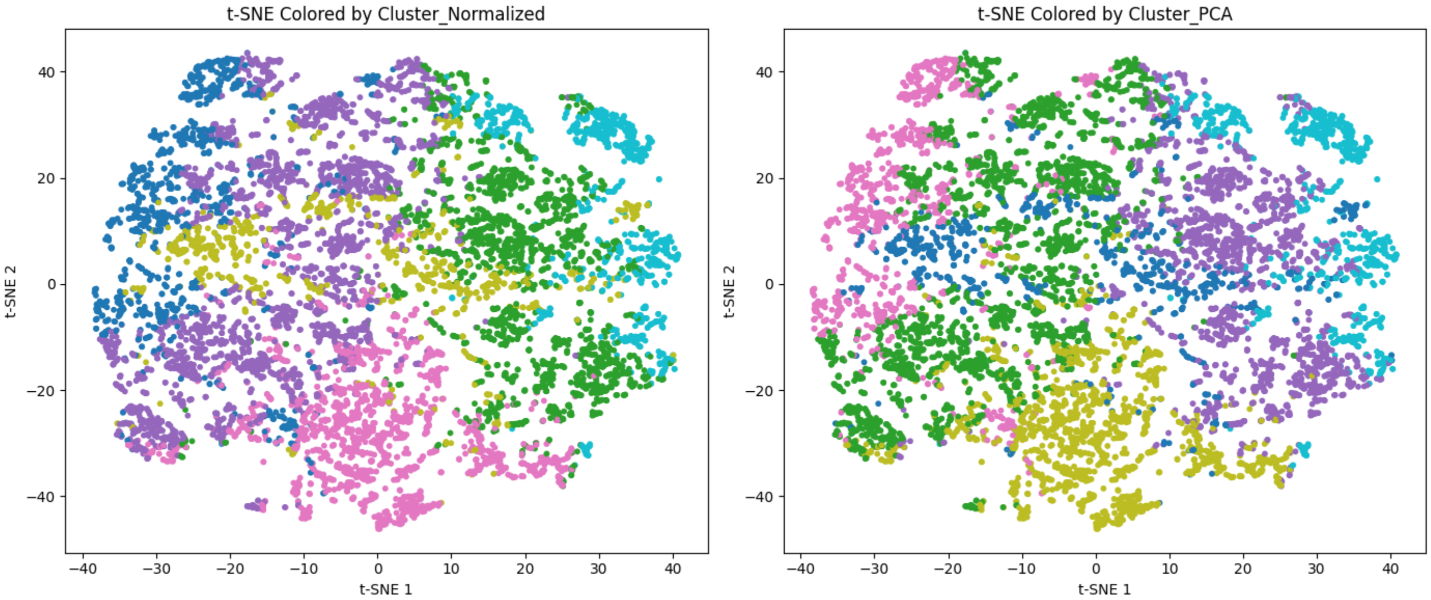

Figure 2: t-SNE visualization of cluster assignments for $k = 6$. **Left:** Clusters based on the full normalized dataset. **Right:** Clusters based on the PCA-reduced dataset.

2.7 Statistical Analysis of Feature Contribution

To understand which features differentiate between the resulting groups, a statistical feature-level analysis was conducted for two grouping schemes: unsupervised clusters (*Cluster*) and supervised labels (*CoverType*).
Three main tools were used for this purpose:

- **Kruskal-Wallis test** – to assess the statistical significance of differences across groups.
- **Eta Squared** (η^2) – to estimate the effect size of each feature in explaining the group separation.
- **Random Forest** – to evaluate feature importance within each grouping. The algorithm was trained twice: once to predict *CoverType*, and once to predict the clustering outcome (*Cluster*). The differences in feature importance served as an additional indication of variables with differential impact across the two grouping approaches.

The combination of statistical tests and tree-based models enabled a deeper analysis of the influence patterns in each method, enhancing our understanding of which variables drive the separation in supervised versus unsupervised settings.

3 Results

3.1 Supervised vs. Unsupervised Learning Results

In the initial phase of analyzing the unsupervised learning outcomes, we visually examined the differences between approaches with and without dimensionality reduction. In both cases, $k = 6$ was identified as the optimal number

of clusters, according to the Elbow Method and the Silhouette score—a consistent finding that reinforced the revealed structure. Subsequently, we chose to analyze the normalized dataset without PCA, to enable direct identification of the features contributing to the clustering. While PCA highlights important variables, it obscures the direct contribution of each feature. Therefore, the original feature space was preferred for understanding and comparing clusters to the ground truth labels. To visualize the findings, we used t-SNE with dual coloring — once by *CoverType* and once by the cluster labels obtained from KMeans (see Figure 3). The comparison highlighted a substantial difference between the two approaches: differing numbers of groups, distinct boundaries, and varied compositions. This divergence served as a starting point for a deeper examination of the differences between the approaches and the role of features in shaping the data structure.
On the quantitative level, consistent results were observed:

- The **Normalized Mutual Information (NMI)** score was 0.053—indicating nearly no overlap between the algorithm's clustering and the human-labelled classification.
- The **Adjusted Rand Index (ARI)** was 0.023—also extremely low.
- **Accuracy**, computed using the Hungarian method, was approximately 29%—a surprisingly low value.

Additional metrics such as V-Measure, Homogeneity, and Completeness further supported this finding, with values ranging from 0.04 to 0.06. These results suggest that the clustering output does not replicate the human classification—though this does not necessarily indicate an error. It is possible that the algorithm uncovered an alternative latent structure, shaped by different patterns in the data—potentially reflecting topographical, environmental, or geometric characteristics that diverge from the original labelling assumptions.

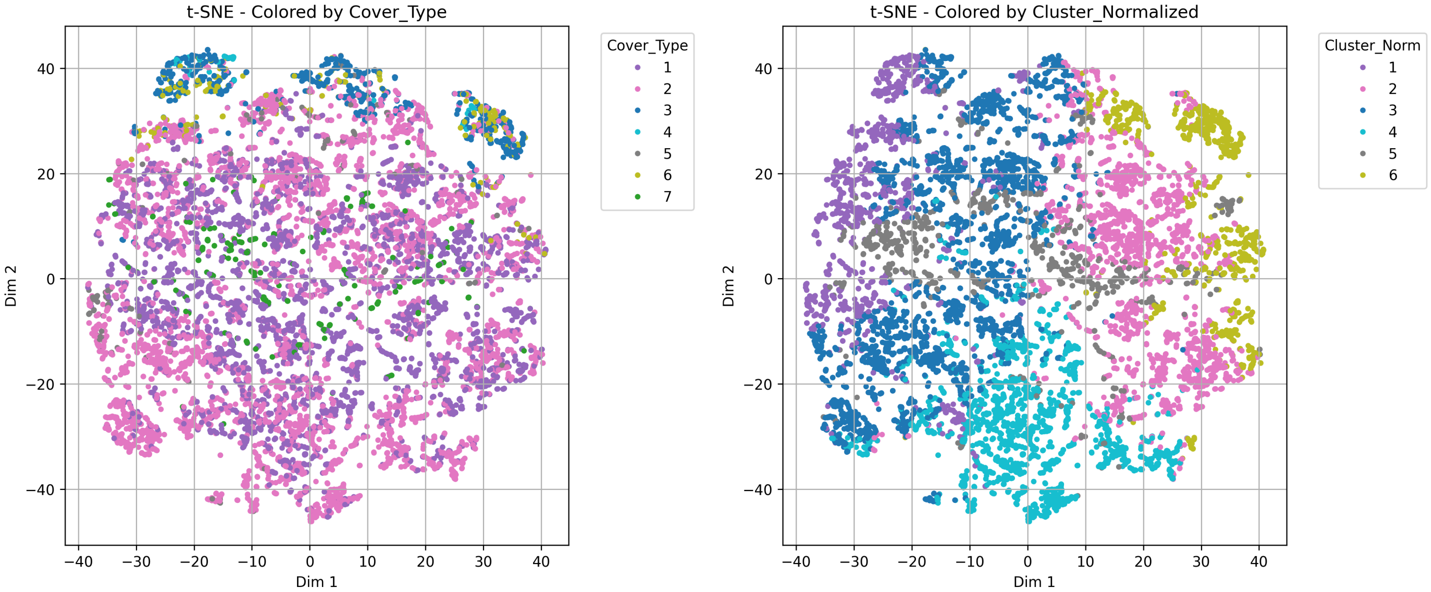

Figure 3: t-SNE projection of the dataset colored by two different groupings: on the left – by *CoverType* (supervised learning), and on the right – by the clustering result of KMeans (unsupervised learning). The figure illustrates the substantial differences between the approaches, both in the geometric structure and the number of groups.

3.2 Feature Contribution Analysis: Supervised vs. Unsupervised Learning

To understand how each feature contributes to group separation under the two partitioning strategies (*CoverType* and *Cluster*), an in-depth feature importance analysis was conducted. The findings revealed significant differences between supervised and unsupervised learning, both in terms of feature ranking and the statistical strength of their impact.

3.2.1 Differences in Feature Rankings

As a first step, we compared the feature importance rankings as measured by two distinct models: one trained on the *CoverType* labels (supervised), and the other on the *Cluster Normalized* labels (unsupervised). Feature importance was computed using Random Forest for each model (see Figure 4).
The comparison highlighted two notable features:

- The **Elevation** feature ranked highly in both models, suggesting it is a strong differentiator in both the supervised labels and the unsupervised clusters.
- In contrast, features such as **Hillshade_3pm**, **Aspect**, and **Hillshade_9am** ranked highly in the unsupervised model but were much less significant in the supervised model.

This discrepancy suggests that the unsupervised model may have captured a different structure within the data—possibly one driven by environmental or lighting conditions, rather than formal categories like soil or vegetation types.

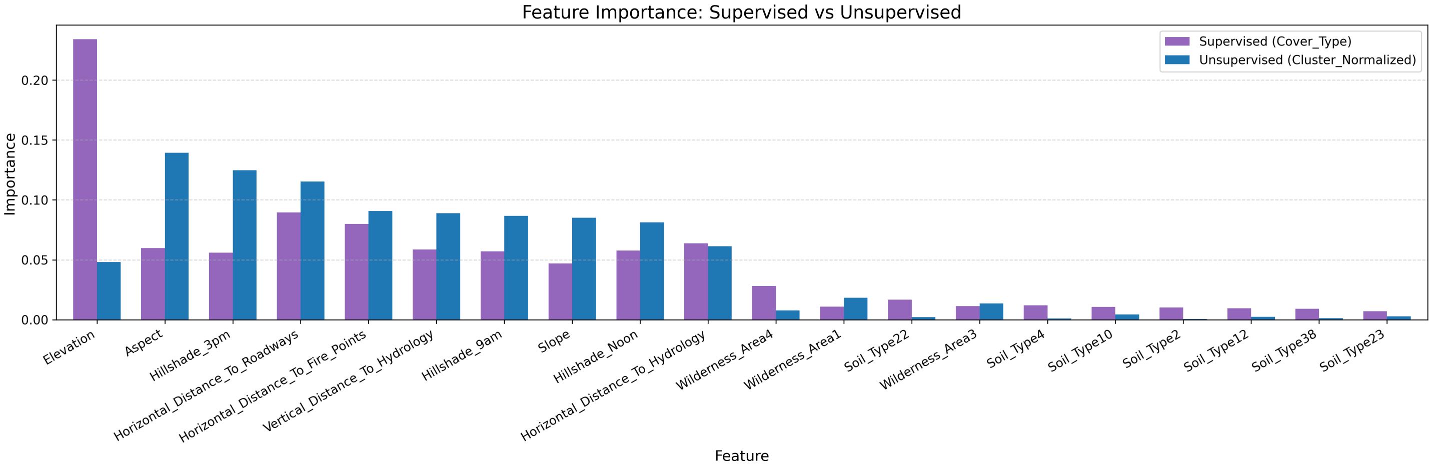

Figure 4: Bar chart showing the importance of the top features as measured by Random Forest, comparing supervised grouping (*CoverType*) with unsupervised grouping (*Cluster Normalized*).

3.2.2 Statistical Analysis of Feature Contribution

To understand which features distinguish between the groups in supervised learning (*CoverType*) versus those formed through unsupervised learning (*Cluster*), we conducted a comparative analysis using the Eta Squared (η^2) metric—a statistical measure that quantifies the proportion of variance explained by each feature. The calculation was performed for each continuous variable using the Kruskal-Wallis test, separately for the ground truth labels and the clustering labels. This approach provided a precise estimate of each feature's contribution to explaining group structure in both settings.

Key findings:

- The feature **Elevation**, which was the most influential in the supervised classification ($\eta^2 \approx 0.52$), contributed very little to the explanation of the clusters.

- In contrast, features such as **Hillshade_3pm**, **Aspect**, and **Hillshade_9am**—which were considered marginal in the supervised model—became dominant in the unsupervised clusters ($\eta^2 > 0.6$).

This finding reflects not just a shift in feature ranking, but a fundamental difference in the structure of feature influence—evidence that each learning approach emphasizes different aspects of the data. It offers statistically significant support for the idea that unsupervised learning uncovered alternative internal structures, which led to a distinct partitioning of the dataset.

Although the results focus on the individual contribution of each feature, it is likely that the prominent differences between supervised and unsupervised groupings do not stem from any single feature in isolation, but rather from interactions among them—patterns that are not captured by standard feature importance metrics. This insight opens the door to a deeper understanding, which is further explored in the Discussion section.

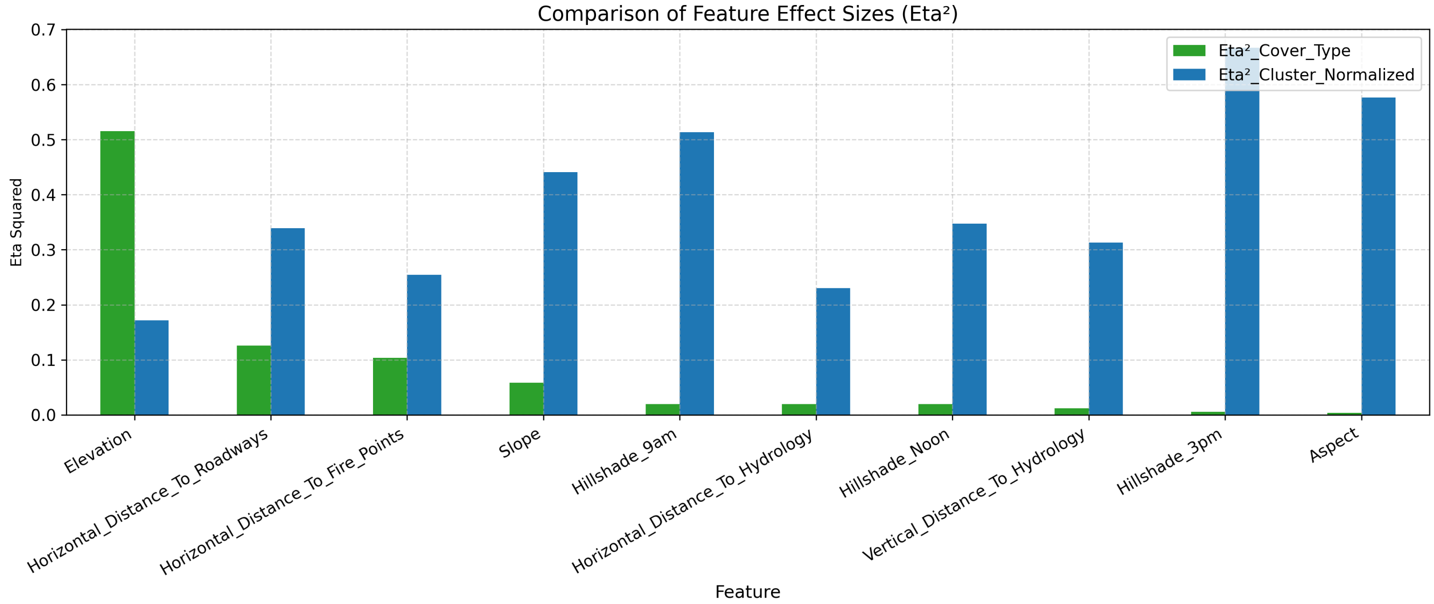

Figure 5: Comparison of Eta Squared (η^2) values for each continuous variable, based on grouping by *CoverType* (supervised learning) and *Cluster* (unsupervised learning). Higher values indicate a greater contribution of the feature to group separation under each respective approach.

4 Discussion

Beyond Labels: What Unsupervised Learning Reveals (and What It Misses)

This project began with a central question: Can the internal structure of the data, when analyzed using unsupervised learning, reproduce the known groupings? To explore this, we intentionally ignored the CoverType labels and allowed the algorithms to detect patterns independently. Labels were used only at the end, for reflective comparison. The results revealed substantial differences: while the original dataset included seven predefined categories, KMeans consistently identified only six clusters with different internal structures. For example, CoverType 4 did not appear as a standalone cluster—its observations were dispersed across multiple groups. This likely indicates that its identity depends on a subtle combination of features, not captured by geometric clustering.

This is not an error, but rather a manifestation of the methodological gap. Unsupervised learning does not aim to validate existing labels, but to uncover latent structures based on distance, density, and similarity. It challenges us to rethink what defines a group: a dominant feature, or a nuanced interaction between several? In the case of Group 4, perhaps a specific mix of elevation, soil type, and distance to water distinguishes it—yet that may not emerge from standard clustering. Moreover, the discovered clusters may reflect entirely different principles—such as lighting, drainage, or unseen ecological patterns—not the forest types originally labeled. In this sense, unsupervised learning may reveal structures that are unnamed but nonetheless real.

The strength of unsupervised learning lies in discovering novel patterns, even when they diverge from expectations. While it may miss complex feature interactions, it can detect deep order where others perceive only noise.

The comparison to the original labels was not to evaluate accuracy, but to explore the algorithm's perception: What does it see? What does it define as meaningful? And what might it capture that we, as humans, overlook?

> *What defines a group? What structures remain hidden when we rely on existing classifications? What insights can be discovered—only when we let go of them?*

References

[1] Kaggle – Forest CoverType Dataset, UCI Machine Learning Repository.
https://www.kaggle.com/...

[2] Scikit-learn Documentation – Unsupervised Learning & Clustering.
PCA Documentation, Clustering Documentation

[3] Understanding Dimensionality Reduction: PCA vs t-SNE – Carnot Research.
https://carnotresearch.medium.com/...

[4] Ghahramani, Z. (2003). *Unsupervised Learning.* Gatsby Computational Neuroscience Unit, University College London.
https://mlg.eng.cam.ac.uk/pub/pdf/Gha03a.pdf

[5] Unsupervised Learning Method Series – Exploring K-Means Clustering.
https://towardsdatascience.com/...