Technische Hochschule Deggendorf
Fakultät Angewandte Informatik
Studiengang Master Künstliche Intelligenz und Data Science

Erklärbares Clustering in der Fernerkundung

Explainable clustering in remote sensing

Masterarbeit zur Erlangung des akademischen Grades:
Master of Science (M.Sc.)
an der Technischen Hochschule Deggendorf

Vorgelegt von:
Shivam Goyal
Matrikelnummer: 12203030

Am: 31. January 2026

Prüfungsleitung:
Prof. Dr. Andreas Fischer

Ergänzende Prüfende:
Zineddine Bettouche

Abstract

Understanding how unsupervised clustering methods form their decision boundaries remains a major challenge in remote sensing explainability. Although algorithms such as k-means are computationally efficient and widely used, their cluster assignments are difficult to interpret and provide limited insight into the underlying physical or semantic structure of the data. Existing explainable approaches for unsupervised learning offer partial transparency, but often lack rule-based semantic explanations and provide limited support for incorporating domain knowledge in an interpretable manner.

This thesis aims to contribute to this research direction by adding explainability to clustering frameworks that combines k-means clustering and Latent Dirichlet Allocation (LDA) by replacing k-means with X-kMeans variants to enhance interpretability while quantifying divergence from the original k-means assignments and confidence in the resulting explanations. Cluster labels produced by k-means are approximated using rule-based decision trees, enabling explicit and human-readable descriptions of cluster boundaries. Both greedy and non-greedy Iterative Mistake Minimization strategies are investigated to analyze trade-offs between model compactness, fidelity, and semantic richness, as well as the transferability of learned rules to unseen data. LDA is subsequently applied to derive semantic topics from cluster-aligned feature representations.

Beyond structural explainability, this work introduces two novel metrics—the Topic Alignment Factor (TAF) and the Word Explainability Confidence Score (WECS)—to quantitatively assess the reliability and semantic consistency of topic–word associations. In addition, Cluster Fidelity is used to measure how faithfully X-kMeans reproduces k-means cluster assignments. The proposed framework is evaluated on subsets of the UCMerced remote sensing dataset. Experimental results indicate that non-greedy decision trees achieve improved semantic alignment and topic coherence, while greedy trees retain compactness and comparable fidelity to k-means.

Overall, this work contributes a confidence-aware structural–semantic explainable clustering framework that improves transparency, trustworthiness, and interpretability of unsupervised learning in remote sensing applications.

Contents

1 Introduction

Unsupervised learning plays a central role in remote sensing analysis, where large volumes of high-dimensional data are generated without reliable or complete ground truth labels. Algorithms such as k-means[1] are valued for their simplicity, computational efficiency, and ability to uncover underlying structure or groupings within unlabeled data. By grouping data points based on similarity, clustering techniques enable exploratory analysis, pattern discovery, and data-driven decision-making across diverse domains such as bioinformatics, agriculture and many other fields. Among clustering algorithms, k-means remains one of the most widely adopted due to its scalability and ease of implementation. However, despite its popularity, k-means suffers from a critical limitation: its results are often difficult to interpret. Cluster assignments are based on distances to centroids in a potentially high-dimensional space, making it challenging for users to understand why a data point belongs to a particular cluster.

In modern applications,reliability is becoming as important as interpretability and transparency. Stakeholders, ranging from domain experts to policymakers, increasingly demand not only high-performing models but also explainable models that can justify their decisions. This is particularly true in sensitive contexts where trust and accountability are essential. In remote sensing, explainability is very important, as model outputs are often used to support environmental monitoring, land-use analysis, and decision-making by domain experts who require transparent and trustworthy models. While explainability has been extensively studied in supervised learning, providing explanations for unsupervised methods remains a significantly more challenging problem. The absence of labelled data complicates both the interpretation of learned structures and the validation of explanatory models.

Several approaches have been proposed to address this challenge. One promising direction for explainable clustering involves the use of interpretable surrogate models that approximate the behaviour of black-box clustering algorithms. In this context, a family of methods seeks to replace opaque cluster assignments with decision-tree based explanations, where hierarchical threshold-based rules describe the partitions created by the clustering algorithm. Such trees provide users with explicit human-readable explanations of how clusters are formed through a sequence of simple, interpretable rules to trace the assignment of any point to a cluster thereby improving transparency and interpretability.

The key challenge, however, is to design such explainable trees while maintaining a balance between human-readable structural simplicity and the preservation of the semantic meaning underlying the clustered data, without substantially compromising clustering performance. Striking the balance between semantic richness and simplicity of explanations remains an open problem. Another important aspect worth exploring is the extent to which the decision rules generated by these approaches remain transferable to unseen data.

To address these challenges, this thesis proposes a hybrid explainable clustering framework that extends a k-means and Latent Dirichlet Allocation (LDA) pipeline with interpretable

decision-tree surrogates (X-kMeans). Building on prior work such as Iterative Mistake Minimization (IMM) [2] and Non-Greedy tree induction strategies introduced in "Don't be greedy, plant bigger trees!" [3], the framework approximates k-means cluster assignments using rule-based decision trees to make implicit decision boundaries explicit. By comparing different tree construction strategies, the approach analyzes trade-offs between interpretability, semantic richness, clustering fidelity, and the transferability of learned rules to unseen data. Semantic topic modelling is subsequently applied to clusters derived from rules based decision trees. Beyond structural explainability, this work introduces two novel metrics—the Topic Alignment Factor (TAF) and the Word Explainability Confidence Score (WECS)—to quantitatively assess the semantic consistency and reliability of topic–word associations, complemented by a Cluster Fidelity measure that evaluates how faithfully the surrogate models reproduce the original k-means clustering. The semantic segmentation of derived topics is supported through domain-expert visual analysis and interpretation of word-level semantics, with TAF and WECS providing quantitative measures to assist these expert-driven interpretations. The proposed framework is evaluated on subsets of the UCMerced remote sensing dataset. Experimental results indicate that Non-Greedy tree strategies tend to capture richer semantic meaning, while Greedy tree strategies favour simpler and more interpretable explanations, demonstrating that meaningful and transferable semantic explanations can be obtained without substantially compromising clustering performance.

Based on this framework, the thesis addresses the following research questions.(1) How can the implicit decision boundaries and semantic meanings underlying clustering in remote sensing data be made explicit through the integration of domain knowledge, and how can the reliability of these explanations be quantitatively assessed? (2)To what extent do the derived semantic meanings and decision rules remain transferable to previously unseen data? The remainder of this thesis is organized as follows…

Chapter 2 – Theoretical Background This chapter introduces the theoretical foundations required for the proposed framework, It goes deeper into Clustering, Explainable Clustering, k-means Clustering,Decision Trees, Greedy and Non-Greedy Tree construction And Latent Dirichlet Allocation

Chapter 3 – Related Work This chapter discusses the work already done in this field and how this work differes or compliments on already existing work

Chapter 4 – Methodology This chapter presents the proposed explainable clustering framework in detail. It explains the data and methods used to test and evaluate the proposed framework

Chapter 5 – Experimental Evaluation This chapter describes the data setup and results of the tests conducted.It compares the propossed explainable pipeline results with previously established blackbox pipeline and tries to evaluate the accuracies of the proposed certainty scores

Chapter 6 – Explainability and Semantic Analysis This chapter focuses on deriving semantic meaning from decision rules and topic structures. It tries to complete the explainability pipeline from implicit k-means cluster semantic meaning to explicit semantic rules to LDA topic modeling to class identification

Chapter 7 – Conclusion and Future Work This chapter summarizes the main contributions and findings of the thesis. It discusses the limitations of the current framework and outlines potential directions for future work

2 Theoretical Background

This chapter provides the mathematical and conceptual foundation for the methods used in this thesis.

It introduces clustering principles, explains the need for explainable clustering, different possible approaches to clustering and presents key algorithms including k-means, decision trees[4], Latent Dirichlet allocation (LDA)[5], Greedy and Non-Greedy tree-building approaches.

2.1 Clustering

Clustering is the unsupervised process of grouping a set of objects so that those within the same group (cluster) are more similar to each other than to those in different groups.

2.1.1 Overview

For this work, given a dataset $\mathcal{X}$, the aim is to find a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ such that:

Hard: Hard Clustering[6] ensures that every data point in the dataset belongs to at least one cluster, thereby achieving complete coverage of the data space. This property guarantees that no data point remains unclassified, which is essential for comprehensive data partitioning. Algorithms such as *K-Means* and *Hierarchical Clustering*[7][8] are typically hard because they assign every data instance to a cluster, even if the assignment may be uncertain near cluster boundaries.

$\bigcup_{i=1}^{k} C_i = \mathcal{X}$.

Exclusive: Exclusivity refers to the requirement that each data point belongs to only one cluster. It enforces clear boundaries between clusters, preventing overlap or ambiguity in membership. Most traditional hard clustering techniques satisfy exclusivity as well — every point is assigned to exactly one cluster.

$C_i \cap C_j = \emptyset for i \neq j$.

However, certain clustering methods relax one or both of these constraints. For example, *DBSCAN*[9] and other density-based approaches may violate hard clustering by labeling outliers or noise points that do not belong to any cluster. Conversely, *Fuzzy C-Means*[10] and *Gaussian Mixture Models (GMM)*[11] relax exclusivity by allowing data points to belong partially to multiple clusters with probabilistic or fuzzy memberships.

Thus, the degree to which a clustering algorithm satisfies both conditions depends on its underlying formulation and the intended application context.

2.1.2 Distance and Similarity Measures

The similarity of two points is often expressed by a distance metric $d(x, y)$. Clustering uncovers hidden structure and is widely used in data mining, biology, marketing, social networks, and earth observation. In clustering, the choice of distance (or dissimilarity) metric plays a crucial role in determining how data points are grouped. Different metrics capture different notions of similarity and can significantly affect the resulting clusters:

- The **Euclidean distance**[12] (L2 norm) represents the straight-line distance between two points in multidimensional space. For points $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$, it is given by:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  It is the most common distance metric, assuming features are equally scaled and uncorrelated. It performs well for compact, spherical clusters in continuous feature space.

- The **Cosine distance**[13] measures the angular separation between two vectors, focusing on their orientation rather than magnitude. It is derived from cosine similarity:

$$CosineSimilarity = \frac{x \cdot y}{\|x\| \|y\|}$$

$$CosineDistance = 1 - CosineSimilarity$$

  Cosine distance is particularly useful for high-dimensional or sparse data, such as text or document embeddings, where the direction of vectors is more meaningful than their absolute magnitude.

2.1.3 Categories of Clustering

Clustering algorithms can be broadly divided into four major categories based on how they define and detect clusters: Partitional, Hierarchical, Density-based, and Model-based methods. Each approach follows a different principle for grouping data points and has unique strengths and limitations.For this thesis Partitional, Hierarchical and Model-based clustering is used

  - **Partitional methods** Partitional clustering [8],[6] divides a dataset into a fixed number of clusters. It optimizes either similarity within a cluster or dissimalirity between different clusters. The most common example is the **K-Means** algorithm, which minimizes the sum of squared distances between data points and their respective cluster centroids:

$$Objective : \min \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

While these clusters are computationally efficient and work effective for large datasets with well-separated, spherical clusters, they tend to be sensitive to initial conditions and have difficulty capturing clusters with irregular shapes

- **Hierarchical methods** Hierarchical clustering builds a nested hierarchy of clusters represented as a dendrogram. Two main strategies are used:
  - **Agglomerative (bottom-up):** Initially each observation is treated as its cluster, and subsequently clusters may be merged based on a linkage criterion.
  - **Divisive (top-down):** Initially all observations are in one cluster, the cluster is recursively split based on splitting criteria.

  Hierarchical clustering does not necessitate pre-specification of the number of clusters and produces interpretable results, however, it is computationally expensive ($O(n^3)$) and sensitive to noise and outliers.
- **Model-based methods** [14],[11] Model-based clustering assumes that the data is generated from a mixture of underlying probability distributions, often Gaussian. Each cluster corresponds to one component of the mixture, and parameters are estimated using the **Expectation-Maximization (EM)**[11] algorithm. A typical example is the **Gaussian Mixture Model (GMM)**. Model-based methods provide a probabilistic interpretation and can handle overlapping clusters, but they often assume a specific distributional form and are computationally intensive for high-dimensional data.

2.1.4 Evaluation

Evaluating clustering quality is essential since clustering is an unsupervised task and ground-truth labels are typically unavailable. Internal validation metrics assess the compactness (cohesion) and separation of clusters. Commonly used metrics include the **Silhouette Coefficient**, **Calinski–Harabasz Score**

Internal indices are used to choose k and assess partition quality when no ground truth exists.

- **Silhouette coefficient:** The Silhouette Coefficient quantifies how similar each data point is to its own cluster compared to other clusters. For each sample i:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

  where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance. The coefficient ranges from -1 to $+1$:
  - $s(i) \approx 1$: sample is well-matched to its own cluster.
  - $s(i) \approx 0$: sample lies between clusters.
  - $s(i) \approx -1$: sample may be misclassified.

A higher average silhouette score indicates better clustering performance. silhouette score is quite populat, but given the higher computatuional requirements of Silhouette coefficient, using silhouette score for choosing k value was not feasible for this thesis

- **Calinski–Harabasz score:** Also known as the *Variance Ratio Criterion*, the Calinski–Harabasz (CH) Index is defined as:

$$CH = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{n-k}{k-1}$$

where $Tr(B_k)$ and $Tr(W_k)$ are the traces of the between-cluster and within-cluster dispersion matrices, respectively. A higher CH score or sudden drop in CH score implies compact and well-separated clusters. It is computationally efficient and suitable for large datasets, which is why k value in this thesis is based on Calinski Harabasz score

2.2 Explainable Clustering

2.2.1 Motivation

While clustering reveals latent structure, conventional algorithms like k-means produce results that are difficult for non-experts to interpret. In high-stakes domains such as healthcare or environmental management, stakeholders need confident, transparent, human-readable explanations for cluster assignments to ensure trust and accountability.

2.2.2 Goals

Traditional clustering algorithms often operate as black boxes,[15] providing cluster assignments without revealing the reasoning behind them. **Explainable clustering** aims to improve the transparency and interpretability of these models by providing insights into both the overall structure of the clustering and the individual data assignments. Interpretability in clustering can be categorized into two main types: **global interpretability**[16] and **local interpretability**[17] following general interpretability frameworks in machine learning.

Global interpretability focuses on understanding the overall behavior and structure of the clustering model. It aims to explain how and why clusters are formed by identifying the dominant features, relationships, and patterns that characterize each cluster. This includes describing clusters through representative centroids or prototypes, determining which features have the strongest influence on the clustering process, and analyzing how clusters differ from or relate to one another within the feature space. In addition, global interpretability seeks to reveal the approximate decision boundaries that partition the data into distinct clusters. Such explanations allow clusters to be interpreted in meaningful,

domain-specific terms, for example by associating a cluster with particular demographic, behavioral, or visual characteristics.

Local interpretability concentrates on explaining individual cluster assignments. Its objective is to answer questions such as why a specific data point was assigned to one cluster rather than another. This perspective provides instance-level explanations and is particularly valuable for identifying edge cases or points that lie near cluster boundaries. Local interpretability is often achieved through approaches such as locally interpretable surrogate models that approximate clustering behavior in the neighborhood of a given instance, counterfactual reasoning that examines how minimal changes in feature values would alter the cluster assignment, or prototype-based analyses that contrast representative and atypical examples within each cluster. This form of interpretability is especially important in domains where individual decisions carry significant consequences, such as healthcare, finance, or personalized recommendation systems.

- **Comparison of Interpretability Types**

Aspect	Global Interpretability	Local Interpretability
Scope	Entire clustering model	Individual data instance
Goal	Understand global cluster structure	Explain a specific assignment
Methods	Centroids, feature importance, summary rules	Local models, counterfactuals, prototypes
Use Case	Transparency, domain insight	Debugging, individualized explanations

2.2.3 Approaches

Explainable clustering can be achieved through different methodological strategies depending on whether interpretability is built directly into the model or applied after clustering. These strategies can be broadly categorized into **intrinsic interpretability**[15],[16], **post-hoc explainability**[18],[16], and **hybrid approaches**[19] following general interpretability frameworks in machine learning.

Intrinsic interpretability: Intrinsic interpretability refers to clustering models that are inherently transparent by design. The internal mechanisms of such models are directly understandable to humans, and the resulting clusters can be easily described or visualized. The clustering process itself produces interpretable outputs such as centroids, prototypes, or rule-based cluster definitions. Cluster formation and decision boundaries can be directly understood from model parameters. Examples include hierarchical clustering, where dendrograms illustrate the hierarchical structure of merges and splits, and rule-based clustering, where clusters are described by simple logical conditions. Although such models are easy to interpret, they may not capture highly complex or non-linear data structures.

Post-hoc explainability: Post-hoc explainability refers to methods that provide explanations *after* a complex clustering model has been trained. Here, interpretability is achieved by applying separate tools or algorithms that analyze the clustering results. Post-hoc methods explain black-box clustering models such as Gaussian Mixture Models, spectral clustering, or deep

learning-based clustering. Explanations are generated through feature attribution, visualization, or surrogate models, enabling the use of high-performance but less transparent clustering techniques. Common examples include using SHAP[18] or LIME[17] which though designed primarily for supervised learning, are applied in conjunction with surrogate or auxiliary models to identify important features influencing cluster membership, employing dimensionality reduction techniques (e.g., t-SNE[20], UMAP[21]) to visualize high-dimensional clusters, and training local surrogate models to approximate the clustering decision process.

Hybrid approaches: Hybrid approaches integrate both intrinsic and post-hoc strategies to combine their respective strengths. They build models with intrinsic interpretability in certain parts that are further refined with post-hoc explanations to enhance understanding. These approaches combine interpretable model components with explainability tools, balancing accuracy, flexibility, and transparency. Examples include prototype-based neural networks that maintain interpretability in latent space, combining interpretable embeddings with post-hoc local explanations, and hybrid frameworks that generate both global summaries and instance-level insights. Hybrid approaches are increasingly used in practical settings.

2.3 k-Means Clustering

2.3.1 Objective

The *k-means* algorithm is one of the most widely used partitional clustering methods due to its simplicity, scalability, and efficiency. It aims to partition a dataset of n observations $\{x_1, x_2, \ldots, x_n\}$ into k non-overlapping clusters $\{C_1, C_2, \ldots, C_k\}$ such that each observation belongs to the cluster with the nearest mean (centroid).

Formally, k-means minimizes the **Within-Cluster Sum of Squares (WCSS)** objective function:

$$\min_{\{C_i\},\{\mu_i\}} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,$$

where:

- μ_i is the centroid (mean vector) of cluster C_i, defined as $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$,
- $\| \cdot \|$ denotes the Euclidean norm.

The objective seeks compact and well-separated clusters by minimizing the squared distance between data points and their corresponding centroids. This makes k-means particularly effective when clusters are convex and approximately spherical in shape.

2.3.2 Properties and Limitations

The k-means clustering algorithm optimizes a non-convex objective function, and as a result, it guarantees convergence only to a *local minimum*. Consequently, the quality of the final clustering solution depends strongly on the initial placement of the cluster centroids. Different

initializations may lead to different local optima, resulting in variability in the clustering outcomes.

This sensitivity to initialization is a well-known limitation of k-means. Poorly chosen initial centroids can cause the algorithm to converge to suboptimal solutions. To mitigate this issue, initialization strategies such as *k-means++* have been proposed, which select initial centroids probabilistically based on their distance from previously chosen centroids. This approach improves robustness and typically leads to more stable and higher-quality clustering results.

Another important assumption underlying k-means is that clusters are approximately spherical and of similar size. Since similarity is measured using Euclidean distance, the algorithm performs best when clusters are isotropic and well separated. In cases where clusters are elongated, overlapping, or exhibit significantly different variances, k-means may fail to capture the true underlying structure of the data.

The algorithm is also sensitive to outliers. Because cluster centroids are computed as arithmetic means of the assigned data points, extreme values can disproportionately influence centroid positions, potentially distorting the clustering results. This sensitivity necessitates careful preprocessing steps such as outlier removal or robust normalization.

Despite these limitations, k-means remains highly scalable and computationally efficient, even for large datasets and high-dimensional feature spaces. Its relatively low computational complexity makes it well suited for exploratory data analysis and as a baseline method against which more sophisticated clustering techniques can be compared.

In summary, k-means offers an efficient mechanism for producing compact and well-separated clusters, but achieving meaningful and stable results requires careful selection of the number of clusters k, appropriate initialization, and thorough data preprocessing.

2.4 Decision Trees

Decision trees are hierarchical, rule-based models that recursively partition the feature space using axis-aligned splits of the form:

$$x_j \leq \tau$$

where x_j represents a feature and τ is the threshold value defining the split. Each internal node performs such a test, dividing the data into two subsets based on whether the condition is satisfied. This recursive partitioning continues until a stopping criterion is reached, forming a tree-like structure composed of decision nodes and terminal leaves.

A path from the root node to a leaf corresponds to a sequence of conditions that can be expressed as a human-readable *if–then* rule. For example:

$$rainfall \leq 500\,mm \ \wedge \ soil_{pH} > 7 \ \Rightarrow \ cluster2$$

Such rules provide intuitive explanations for how data points are grouped or classified, making decision trees highly suitable for interpretable and explainable modeling.

Splitting Criteria During training, a decision tree selects the feature and threshold that best split the data into more homogeneous subsets with respect to the target variable. For supervised

tasks, this is typically achieved by minimizing an impurity measure.In this project, it is done using Iterative mistake minimization

In explainable clustering, the target labels are cluster assignments provided by k-means.

2.5 Greedy and Non-Greedy Tree Construction

Decision trees are a natural vehicle for explainable clustering because they yield human-readable rules. Two broad strategies exist for building such trees:

2.5.1 Greedy Algorithms

A Greedy algorithm builds the tree top-down, making locally optimal choices at each node.

Given cluster labels $y(x)$ from k-means, for node u containing data X_u, the number of misclassified points is

$$mistakes(u) = |\{x \in X_u : y(x) \neq \hat{y}(x)\}|.$$

where $\hat{y}(x)$ is a function satisfying a condition The best split is chosen to minimize mistakes (or cost).

The Iterative Mistake Minimization (IMM) algorithm follows this strategy, producing exactly k leaves, each representing a cluster.Its **advantage** is Fast and easy to implement trees which are compact and interpretable. Its **limitations** are that it may yield suboptimal global cost and is sensitive to data order and local optima

2.5.2 Non-Greedy Algorithms

Non-Greedy methods relax the best local split requirement and incorporate global cost or randomness.

Key innovations include **Cost-based splitting:** directly minimize the k-means objective when choosing splits, **Leaf budget:** allow up to $(1 + \delta)k$ leaves to prevent harmful forced splits and **Center sharing:** if a center lies near the threshold, include it in both child node

These strategies yield more faithful approximations of the k-means clustering at the expense of a slightly larger or more complex tree.

2.6 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data such as text corpora, image patches, or any "bag-of-words" representation [5]. It discovers latent topics, where each topic is a distribution over a fixed vocabulary, and each document is a mixture of these topics. Blei, Ng, and Jordan introduced LDA as a Bayesian hierarchical model that generalizes earlier mixture-of-unigrams models and probabilistic latent semantic analysis.

Generative Process Given a fixed number of topics K, a vocabulary of V unique words, and hyperparameters α (the document–topic Dirichlet prior) and β (the topic–word Dirichlet prior), LDA assumes the following generative process.

For each topic $k = 1, \ldots, K$:

- Draw a topic–word distribution $\phi_k \sim \mathrm{Dirichlet}(\beta)$.

For each document d with N_d words:

- Draw topic proportions $\theta_d \sim \mathrm{Dirichlet}(\alpha)$.
- For each word position $n = 1, \ldots, N_d$:
  - Choose a latent topic $z_{d,n} \sim \mathrm{Categorical}(\theta_d)$.
  - Choose a word $w_{d,n} \sim \mathrm{Categorical}(\phi_{z_{d,n}})$.

Joint Probability For a single document d with words w, topic assignments z, topic proportions θ, and topic–word distributions ϕ, the joint probability is

$$p(w_d, z_d, \theta_d, \Phi \mid \alpha, \beta) = p(\theta_d \mid \alpha) \prod_{k=1}^{K} p(\phi_k \mid \beta) \prod_{n=1}^{N_d} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid z_{d,n}, \Phi) \tag{2.1}$$

For an entire corpus D of M documents:

$$p(W, Z, \Theta, \Phi \mid \alpha, \beta) = \prod_{k=1}^{K} p(\phi_k \mid \beta) \prod_{d=1}^{M} \left[p(\theta_d \mid \alpha) \prod_{n=1}^{N_d} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid z_{d,n}, \phi) \right] \tag{2.2}$$

Marginal Likelihood [5]

The probability of the observed words W is obtained by marginalising over the latent variables Θ, Φ, and Z:

$$p(W \mid \alpha, \beta) = \sum_{Z} \int p(W, Z, \Theta, \Phi \mid \alpha, \beta)\, \mathrm{d}\Theta\, \mathrm{d}\Phi \tag{2.3}$$

which is intractable to compute exactly, motivating approximate inference methods.

Inference Methods Two major families of algorithms are commonly used to estimate the hidden variables and topic–word distributions.

Variational Bayes [22] A variational distribution $q(\theta, z)$ with free parameters (typically denoted γ and φ) is introduced to approximate the true posterior. The Evidence Lower Bound (ELBO) maximised during inference is

$$\mathcal{L} = E_q[\log p(w, z, \theta \mid \alpha, \beta)] - E_q[\log q(z, \theta)] \tag{2.4}$$

Collapsed Gibbs Sampling [23] Alternatively, the parameters θ and ϕ can be analytically integrated out, and each topic assignment $z_{d,n}$ is sampled from its conditional distribution given all other assignments, using word–topic and document–topic count statistics.

Model Parameters and Interpretability The topic–word distributions ϕ_k represent each topic k as a probability vector over the vocabulary. The document–topic distributions θ_d represent each document as a K-dimensional mixture vector over topics.

Relevance to This Work In this thesis, LDA is used as a post-clustering semantic analysis tool:

- *Input:* Pixels are interpreted as visual words and images as visual documents. Their feature vectors are first clustered using k-means and its explainable variants, thereby generating the discrete vocabulary required for subsequent topic-based analysis.
- *Label Conditioning:* Labels from regular k-means or from the X-kMeans decision tree are used to organise documents.
- *Output:* LDA provides topic–word distributions and per-sample topic proportions, which reveal semantic topics associated with each cluster or decision-tree leaf.

This combination of symbolic rules (decision trees) and probabilistic semantics (LDA topics) produces a dual explanation: threshold-based cluster boundaries and meaningful latent topics that can be inspected or visualised.

2.7 Explainable clustering with surrogate trees

Early work on explainable k-means developed axis-aligned surrogate trees that faithfully approximate centroid assignments while yielding human-readable rules. The Iterative Mistake Minimization (IMM) algorithm builds a tree greedily until exactly k leaves are formed—one per cluster—minimising misassignments at each split. Later, Non-Greedy refinements relaxed the k-leaf constraint with a leaf budget and center sharing near thresholds, improving global cost at the expense of slightly larger trees. These ideas motivate the Greedy vs. Non-Greedy X-kMeans surrogates used throughout this thesis.

2.8 End-to-end explainable cluster analysis

Beyond tree surrogates, clustering can be combined with post-hoc interpretability (e.g., topic models) to provide both statistical fidelity and semantic narratives of clusters. This approach works on this line by coupling tree rules with LDA-derived topics and by introducing word-level confidence measures (WTAC/WECS) to quantify semantic alignment.

3 Related Work

This chapter provides a detailed review of the prior work that informs and motivates this thesis, as well as other explainability techniques that offer comparable perspectives. It examines foundational contributions in explainable clustering, interpretable tree-based models, and unsupervised model attribution, highlighting how these approaches address the challenges of interpretability in the absence of labels. By situating the proposed framework within this broader research landscape, the chapter clarifies the conceptual gaps that remain in current literature—particularly in explainability for unsupervised methods and remote sensing—and establishes how this work builds upon, diverges from, and extends existing methodologies. This contextual grounding ensures that the motivations, novelty, and relevance of the thesis are clearly articulated

3.1 Explainable k-Means and k-Medians

Moshkovitz et al. [2] introduced the first formal algorithmic framework for constructing *interpretable surrogate trees* that faithfully approximate the assignments produced by k-means and k-medians. Their main contribution is the **Iterative Mistake Minimization (IMM)** algorithm, which recursively builds an axis-aligned decision tree where each leaf corresponds to a cluster.

At each node, IMM greedily selects a feature j and threshold τ that minimize the number of *mistakes*, defined as data points whose k-means label differs from the majority label of the resulting child nodes. Formally, for a node u with data X_u with centers (x), the algorithm minimizes:

$$Mistakes(i, \tau) = \sum_{x \in X_u} 1[(x_i \leq \tau) \neq (\mu(x)_i \leq \tau)]$$

The process continues until exactly k leaves are formed, ensuring that each cluster is represented by a human-readable rule such as:

> *"IF rainfall $\leq$ 500 mm AND soil_pH $>$ 7 THEN cluster 2."*

The authors prove that the resulting trees approximate the clustering cost within a constant factor of the optimal k-means or k-medians objective. This seminal work established the paradigm of **explainable clustering**, showing that unsupervised learning can yield interpretable, rule-based models.

Comparison with thesis Approach This work builds upon the IMM algorithm proposed in the referenced study and replaces the k-means in the original kMeans–LDA pipeline with an x-means–based decision tree framework. In addition, this thesis investigates a modified version of the IMM algorithm and examines the transferability of the semantic rules derived from both IMM variants. Furthermore, a cluster fidelity metric is introduced to quantitatively assess the fidelity of X-kMeans to k-means.

3.2 Explainable k-Means: Don't Be Greedy, Plant Bigger Trees!

Makarychev et al., [3] identified limitations in the purely Greedy nature of IMM. Since IMM enforces exactly k leaves, it may be forced into suboptimal splits, particularly when cluster centers lie near decision boundaries.

To address this, they introduced two key innovations:

1. **Leaf-budgeted trees:** The algorithm allows up to $(1 + \delta)k$ leaves, where $\delta > 0$ is a relaxation factor that permits more flexible partitions.
2. **Center sharing:** This approach performs data splits in a manner that allows a cluster center to be assigned to either side of a partition. By permitting centers to be shared across multiple branches of the decision tree, the algorithm avoids irrevocably separating data points from their nearest centers at early stages of the hierarchy. This flexibility enables the construction of larger yet more expressive trees.

Comparison with thesis Approach The Non-Greedy modification of the IMM algorithm in this work is inspired by the center-sharing strategy proposed in "Don't Be Greedy, Plant Bigger Trees!", enabling a detailed analysis of the center-sharing variant of the IMM algorithm.

3.3 Feature-Free Explainable Data Mining in SAR Images

Karmakar et al. [24] proposed a domain-specific explainability framework titled *Feature-Free Explainable Data Mining in SAR Images Using Latent Dirichlet Allocation*[24]. This approach applies probabilistic topic modeling to Synthetic Aperture Radar (SAR) imagery, offering both clustering and interpretation without manual feature engineering.

Key Contributions

- **Feature-free modeling:** Image patches or pixel regions are treated as visual words' and the image as a visual document, allowing the use of Latent Dirichlet Allocation (LDA) to uncover latent visual topics directly from data.
- **Topic-based explanations:** Each discovered LDA topic corresponds to distinct SAR phenomena, such as terrain textures, land-water boundaries, or structural patterns.
- **Symbolic interpretability:** The model maps the learned topics back to the spatial domain, generating interpretable visual and textual explanations that describe the occurrence and meaning of each topic.

Comparison with thesis Approach The work on Feature-Free Explainable Data Mining in SAR images is particularly relevant to this research, as it demonstrates that LDA, a generative probabilistic model, can produce explainable clustering results in the image domain. Building on this idea, the present thesis integrates the topic modeling capability of LDA with the semantic information extracted via X-kMeans, resulting in a unified framework for explainable clustering and topic modeling.

3.4 A Comprehensive Framework for Explainable Cluster Analysis

While previous works focus on algorithmic and domain-specific methods, Alvarez-García, Ibar Alonso, and Arenas-Parra[25] propose an end-to-end pipeline for practical explainable cluster analysis. Their framework integrates the following steps:

- **Data preprocessing:** handling missing values, outliers, and normalization;
- **Dimensionality reduction:** applying Sparse PCA or Multiple Correspondence Analysis;
- **Clustering:** using standard algorithms such as k-means or hierarchical methods;
- **Surrogate modeling:** training supervised classifiers to predict cluster assignments;
- **Post-hoc explainability:** leveraging SHAP values to identify both global and local feature contributions.

Comparison with thesis Approach End to end pipeline in A Comprehensive Framework for Explainable Cluster Analysis provides complete, operational framework for explainable clustering, where interpretability is integrated into the entire data mining pipeline. The present thesis replaces the post-hoc explainability of this framework with a combination of feature-based semantic rules, semantic segmentation of LDA topics, and TAF and WECS scores, enabling a human readable and quantifiable assessment of cluster explanations.

3.5 Algorithm-Agnostic Explainability for Unsupervised Clustering

Algorithm-Agnostic Explainability for Unsupervised[26] tackles the lack of explainability in unsupervised clustering by adapting model-agnostic explainability ideas from supervised learning to the clustering setting. It introduces G2PC, which provides global insight into the features that distinguish clusters, and L2PC, provides local insight into what makes individual samples belong to a particular cluster

The authors evaluate these methods across five major clustering families partition-based, density-based, model-based, hierarchical, and fuzzy methods using low-dimensional synthetic data and high-dimensional rs-fMRI FNC data from schizophrenia studies and control subjects. They demonstrate on low-dimensional, ground-truth synthetic data that they can be paired

with multiple clustering algorithms to identify the features most important to differentiating clusters and the utility of the methods for high-dimensional datasets by analyzing rs-fMRI FNC data and identifying cross-domain connectivity associated with schizophrenia

Comparison with thesis Approach Algorithm-Agnostic Explainability for Unsupervised Clustering treats explanation as a separate, post-hoc step for any clustering algorithm, providing feature importance scores without generating human-interpretable rules or guaranteeing fidelity to the original clusters. In contrast, the present thesis embeds explainability directly into the clustering pipeline, producing explicit semantic rules, topic-based segmentation, and quantitative metrics (cluster fidelity, TAF, WECS) to assess how well the explanations preserve the original cluster structure.

3.6 Label-Free Explainability for Unsupervised Models

Label-Free Explainability for Unsupervised Models[27] focuses on the difficulty of interpreting unsupervised black-box models, where outputs are latent representations whose dimensions lack direct semantic meaning. The authors propose a new label-free explainability framework that extends existing post-hoc methods to settings without labels. They introduce two complementary ideas: label-free feature importance, which identifies influential input features, and label-free example importance, which highlights influential training samples. Their approach relies on constructing an auxiliary scalar objective that enables standard attribution methods to operate on latent representations. Importantly, these extensions can be applied as lightweight wrappers around existing techniques, without increasing model or computational complexity. The framework preserves key theoretical guarantees, including completeness and invariance to transformations of the latent space. Through experiments on autoencoders and contrastive models, the authors show that the explanations are stable and informative. The paper also uses the framework to compare representations learned from different pretext tasks. Finally, it demonstrates that strength of disentanglement in VAEs does not necessarily relate to interpretability of saliency maps, challenging a common assumption in representation learning.

Comparison with thesis Approach Label-Free Explainability for Unsupervised Models introduces a paradigm for post-hoc explainability in unsupervised learning by extending feature and example importance methods to operate without requiring labels, thereby highlighting influential features and training examples for black-box representations. In contrast, the present work instead embeds interpretability directly into the clustering pipeline by combining X-kMeans clustering with LDA-based topic modeling and explicit semantic rule extraction. Unlike the label-free explainability methods, which aim to interpret latent representations (such as autoencoder outputs) through feature contributions, thesis approach produces structured, human readable semantic rules and evaluates them using dedicated metrics such as cluster fidelity, TAF, and WECS scores to quantitatively assess how well explanations preserve the original clustering structure.

3.7 Grad-CAM Family

Grad-CAM (Gradient-weighted Class Activation Mapping)[28] is a post-hoc explainability method that produces visual explanations by highlighting image regions most influential for a model's prediction. It works by using the gradients of a target class with respect to the final convolutional feature maps to compute importance weights. These weights are combined with the feature maps to generate a coarse, class-discriminative heatmap. Grad-CAM improves over earlier CAM methods by not requiring architectural changes or global average pooling. Grad-CAM++[29] extends Grad-CAM by using higher-order gradients, enabling better localization when multiple instances of the same class appear in an image. Smooth Grad-CAM++[30] reduces visual noise by averaging explanations over noisy input perturbations. Score-CAM[31] removes the need for gradients altogether and instead uses forward-pass class scores to weight activation maps. EigenCAM[32] is a class-agnostic variant that uses principal components of feature maps to generate explanations, capturing dominant activation patterns without relying on gradients or class labels. Together, these variants trade off between faithfulness, stability, and computational cost while improving localization quality.

Comparison with thesis Approach The Grad-CAM family of methods is primarily designed for providing visual explanations of deep neural network predictions, particularly in image classification tasks, by highlighting regions in input images that most influence the model's output. These approaches are inherently supervised, rely on backpropagation through trained networks, and produce heatmaps rather than explicit, human-readable rules. In contrast, the present thesis focuses on unsupervised clustering and explainability, integrating X-kMeans clustering with LDA-based topic modeling to extract explicit semantic rules for cluster interpretation. Unlike Grad-CAM, which explains predictions post-hoc without quantifying how faithfully the explanation reflects underlying structure, this work embeds interpretability into the pipeline and introduces metrics such as cluster fidelity, TAF, and WECS to evaluate how accurately the extracted rules preserve the original clustering structure.

4 Methodology

This chapter outlines and explains the methodologies employed throughout the study. It provides a detailed description of the complete analytical pipeline, including data preprocessing, feature engineering, clustering strategies, explainability mechanisms, and evaluation procedures. By systematically presenting each methodological component, this chapter establishes a clear foundation for understanding how the proposed framework operates, how its outputs are generated, and how the Greedy and Non-Greedy approaches are compared. The Cluster Fidelity experiment quantifies the fidelity of X-kMeans label assignment to k-means label assignment.The LDA experiment evaluates the transferability of the explicitly derived semantic rules and LDA topics across datasets. By jointly leveraging LDA topic representations and decision tree–based rules, the proposed approach produces interpretable outputs that are suitable for semantic segmentation by domain experts. This enables the systematic incorporation of domain knowledge into the learning process, thereby enhancing both the interpretability and practical relevance of the resulting clustering outcomes.TAF and WECS quantifies the reliability of integrating the cluster semantic meaning with topic semantic meanings. The goal is to ensure transparency, reproducibility, and a rigorous grounding for the results and interpretations discussed in later chapters.

4.1 Overview

Traditional k-means produces opaque Voronoi partitions that are difficult to interpret. The objective is to develop an explainable clustering framework that balances clustering quality (low error) with human interpretability (simple rules). X-kMeans builds an axis-aligned decision tree that explains k-means cluster assignments through simple threshold rules.

- **IMM-based approach:** Extend k-means with threshold trees where nodes split on single features to minimize mistakes or cost.
- **Greedy IMM:** Splits the data in two parts in such a way that both parts have different labels. The splits cannot have overlapping labels.
- **Non-Greedy IMM with Buffer Zone:** extends the Greedy algorithm by sharing centres across a buffer around each threshold, such that datapoints and centers in the buffer zone go into both splits (overlapping splits possible), both while calcumating mistakes and splitting the data, allowing a final leaf count of $k + c$ (where c is the number of centres which went into both splits).

Integration of recent work: Incorporate ideas from multiple researches. The core idea uses decision-tree structures to describe cluster assignments with higher interpretability and robustness.

Both variants are subjected to two independent tests and an alternative word–topic analysis is done for one experiment to show how this approach can be used for interpretability and class discovery.

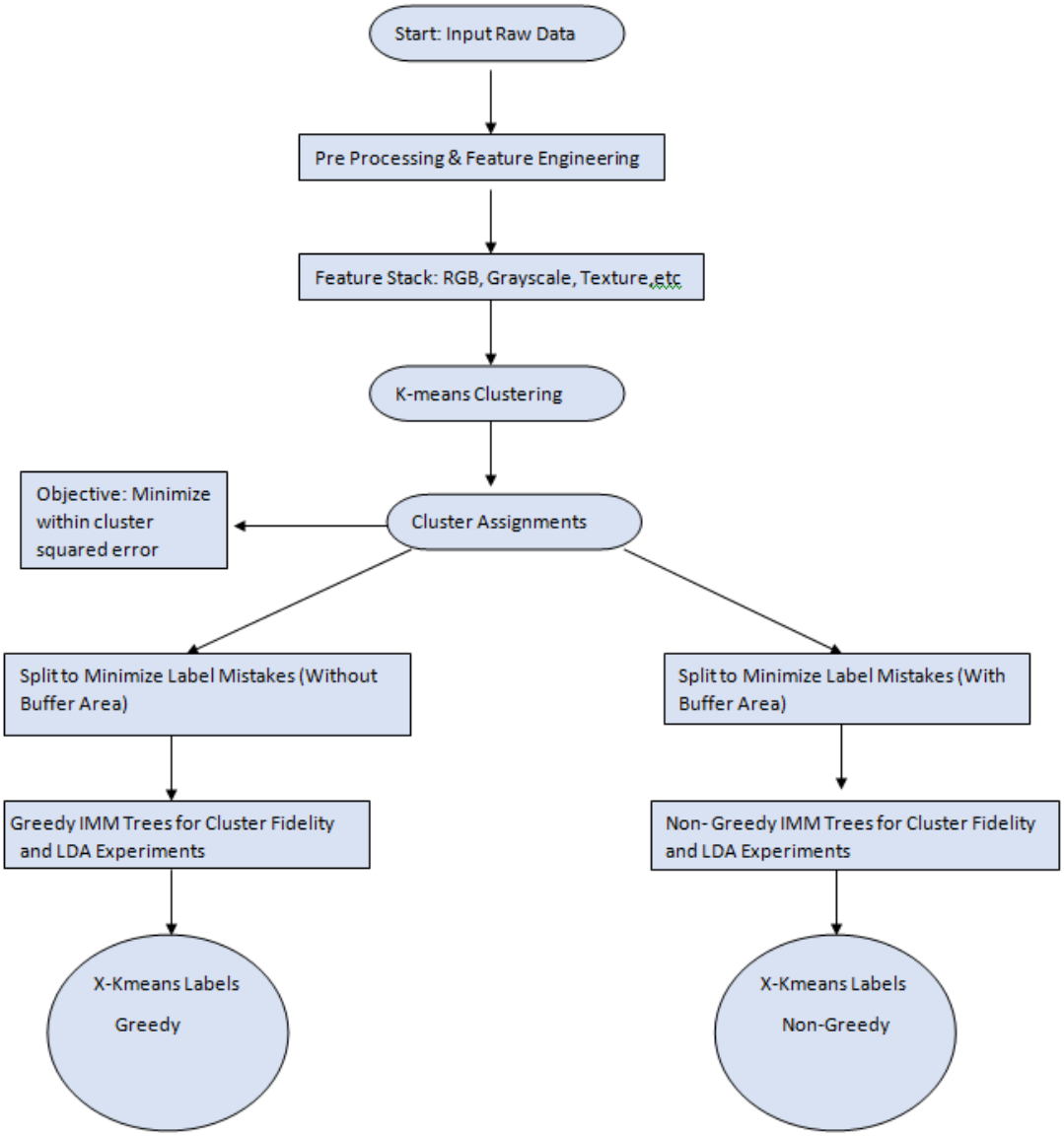

Figure 4.1: WorkFlow to generate X-kMeans Decision Trees and labels using k-means

Figure 4.1 illustrates the workflow for generating X-kMeans Decision Trees. The process begins with raw data ingestion, followed by preprocessing and feature engineering to construct a consolidated feature stack. K-means clustering is then applied, with the objective of

minimizing the within-cluster sum of squared errors. Based on the resulting clusters, Greedy and Non-Greedy IMM trees are constructed. These trees are subsequently used to generate X-kMeans labels, which serve as the input for cluster fidelity analysis and LDA experiments.

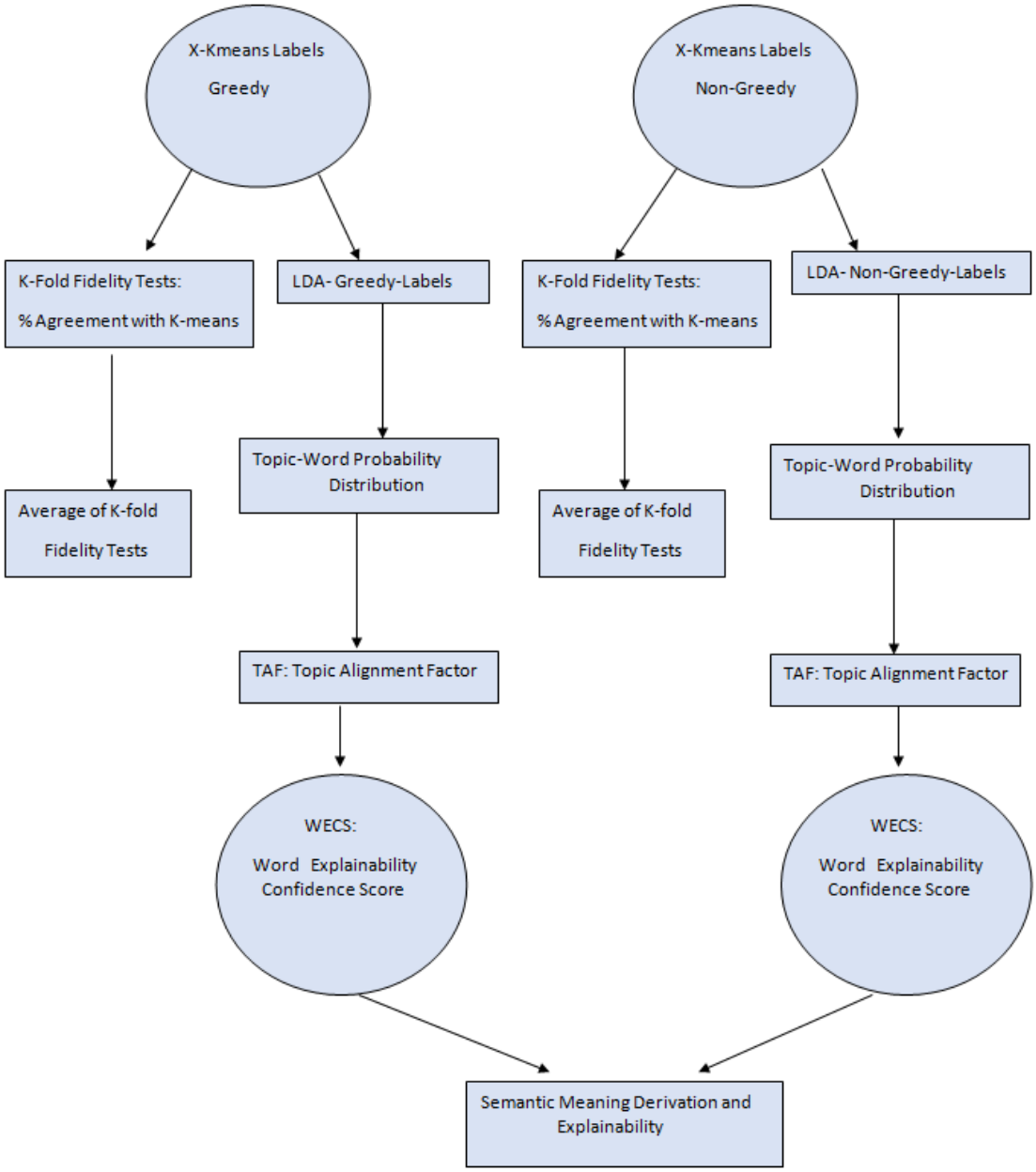

Figure 4.2: Workflow to test the X-kMeans Decision Trees and Derive Semantic Meaning

Figure 4.2 presents the workflow for utilizing labels generated by X-kMeans decision trees in cluster fidelity and LDA-based semantic analysis. For cluster fidelity evaluation, k-fold stratified fidelity tests are performed, and the mean fidelity across all folds is reported as the final score. For the LDA experiments, labels produced by the X-kMeans decision trees are used as inputs to the LDA model. The resulting topic–word probability distributions are used to compute the Topic Alignment Factor and Word Explainability Confidence Score, which together facilitate semantic interpretation of clusters and enhance model explainability.

4.2 Approach

Baseline Started with k-means clustering to obtain cluster centers.

IMM Trees

- Initially, mistake-based splitting where each label can go in only one side of the binary split (Greedy Approach).
- Modified version: Taking buffer radius around thresholds for calculating mistakes where multiple labels can be there on both sides of the binary split (Non-Greedy Approach).

Non-Greedy vs Greedy Approach

- **Leaf budget:** Allow up to $k + c$ leaves (Non-Greedy) compared to k leaves(Greedy), improving flexibility and reducing bad splits in Non-Greedy compared to Greedy.
- **Center sharing:** Duplicate centers close to a threshold into both branches (Non-Greedy), preventing cost blow-ups.
- **Explainability Complexity:** More complex rules(Non-Greedy) to extract semantic meaning of a cluster compared to Greedy.

Decision-tree integration Use tree induction methods with pruning to generate compact rule sets. Trade-off between cluster fidelity and simplicity via a user-controlled parameter Buffer Ratio(Non-Greedy) which decides buffer area around threshhold where all points belong to both splits.

Additional explainability Create a custom framework for global and local feature importance to validate IMM tree rules. Compare IMM-based rules vs word topic assignment in LDA to provide explainability.

4.3 Data Processing and Feature Engineering

Here a comprehensive description of the preprocessing steps applied prior to experimentation is provided.

4.3.1 Datasets

Primary dataset used was UCmerced dataset. This thesis evaluates the performance of k-Means and X-kMeans clustering in a non-specialized, general-purpose setting, reflecting realistic conditions with minimal prior assumptions. Because LDA represents images as a bag of words, it may lose spatial information. To address this, features derived from RGB data were used to capture informative visual characteristics while preserving limited spatial context.

Additionally, since X-kMeans is sensitive to low-cardinality features, dominant orientation was processed to provide complementary spatial cues while maintaing low dimensionality. Following is the final feature stack

F0-2 Normalized RGB channels All pixel intensities are converted to floating point and scaled to the range $[0, 1]$ by division with 255.0. This normalization ensures consistency across images, allowing learned rules to remain transferable while preserving the semantic meaning of visual words, independent of varying image acquisition settings

F3 Grayscale Intensity (1 feature) A luminance channel is computed:

$$I_{gray} = 0.299R + 0.587G + 0.114B$$

then scaled to $[0, 1]$.

F4 Local Binary Pattern (LBP) (1 feature) Texture at each pixel is encoded using an 8-neighbour, radius-1 uniform LBP operator, normalised to $[0, 1]$. LBP captures local texture micro-patterns.

F5 Gradient Magnitude (1 feature) The Sobel operator estimates horizontal and vertical derivatives G_x, G_y; gradient magnitude is $\sqrt{G_x^2 + G_y^2}$, then quantised into 256 bins and normalised.

F6-8 Global Mean RGB (3 features) The image-wide mean of each RGB channel is replicated to all pixels, supplying global colour context.

F9 LBP Histogram Entropy (1 feature) The entropy $H = -\sum_i p_i \log p_i$ of the LBP histogram summarises overall texture complexity and is broadcast to every pixel.

F10 Edge Density (1 feature) Using the Canny edge detector, the ratio of edge pixels to total pixels provides a global edge-density measure, again broadcast.

F11 Dominant Edge Orientation (1 feature) From gradient angles $\theta = \operatorname{atan2}(G_y, G_x)$ a 4-bin orientation histogram is computed; the bin of maximum count is encoded (normalised to $[0, 1)$) and broadcast.

Final Feature Stack All features are concatenated along the channel dimension to produce a tensor of shape $(x \cdot y) \times 12$, where the 12 channels correspond to: R, G, B; grayscale; LBP; gradient magnitude; mean R, mean G, mean B; LBP entropy; edge density; dominant orientation.

4.4 Experiments

This subsection discusses the type of experiments conducted

4.4.1 Experiment 1 – Cluster Fidelity Evaluation

Objective: Quantify how faithfully the X-kMeans decision tree reproduces the original k-means labels. The purpose of computing cluster fidelity is to assess how well the rule-based X-kMeans decision tree aligns with the original k-means clustering assignments. While a high fidelity score indicates strong alignment, and is highly desirable, a non-perfect score does not inherently suggest that the X-kMeans model is inferior in capturing semantic structure. Unlike standard k-means, which assigns equal importance to all features after normalization, the X-kMeans framework produces a hierarchically structured rule set that assigns differential significance to features based on their contribution to the separation of clusters.

This aspect becomes particularly valuable in high-dimensional or multimodal datasets, such as UCmerced baseball diamond, where the semantic relevance of certain features (e.g., grayscale, texture, green intensity) is stronger than others. In such scenarios, even when cluster fidelity shows some misalignment between X-kMeans and k-means, the X-kMeans decision tree may offer superior semantic grounding by focusing on the most discriminative features—thereby improving interpretability and contextual alignment. Thus, cluster fidelity should be viewed not as a sole indicator of performance but as one component in a broader interpretability–performance tradeoff in tandem with ground truth/domain knowledge and topic-based explainability metrics like TAF and WECS.

Cross-Validation Setup To evaluate the robustness and stability of the proposed explainable clustering framework, a stratified k-fold cross-validation strategy was employed. In this context, a *fold* refers to one of k approximately equal-sized, non-overlapping partitions of the dataset, constructed such that the distribution of k-means cluster labels is preserved across partitions.

In this study, $k = 3$ folds were used. This choice reflects a trade-off between statistical robustness and dataset size. Given the limited number of samples in the UCMerced Airplane and baseball subsets and hardware limitations using three folds ensures that each training split contains a sufficiently large number of samples to construct stable and interpretable decision trees, while still allowing for meaningful evaluation on held-out data.

Procedure:

1. **k-Means Clustering** Standard k-means clustering was first performed on the complete dataset to obtain reference cluster assignments and centroids. These assignments serve as the *ground-truth* cluster labels for subsequent explainability evaluation.

2. **Decision Tree Construction** For each fold $i \in \{1, 2, 3\}$:

   a) The training subset (2 folds,67%) was used to construct two interpretable models:

      i. the **Greedy IMM Decision Tree**, and

      ii. the **Non-Greedy IMM Decision Tree**.

   b) Both trees were trained to replicate the cluster structure derived from k-means, using feature thresholds that maximize alignment with the cluster assignments while maintaining interpretability.

3. **Fidelity Measurement (per fold)** Each model was evaluated on the held-out test subset (1 fold, 33%) of the same fold. The **Cluster Fidelity** for fold i was computed as:

$$\text{Fidelity}_i = 100 \times \frac{N_{correct}}{N_{test}}$$

where $N_{correct}$ denotes the number of datapoints correctly assigned to their k-means cluster, and N_{test} denotes the total number of datapoints in the test subset.

This metric quantifies how accurately the interpretable decision tree reproduces the original k-means partitioning.

4. **Cross-Fold Aggregation** The final cluster fidelity was obtained by averaging the fold-wise fidelity scores across all three folds:

$$Fidelity_{final} = \frac{1}{3}\sum_{i=1}^{3} Fidelity_i$$

This aggregation ensures that the reported fidelity score is stable, unbiased, and representative of the model's performance across the full dataset.

4.4.2 Experiment 2 – Kmeans-LDA experiments

Objective: These experiments investigates whether the semantic structure learned by the explainable clustering framework (**X-kMeans**) is consistent with that of traditional **k-means** clustering when analyzed through a probabilistic topic model. Specifically, Latent Dirichlet Allocation (LDA) is employed to evaluate how well the clusters generated by k-means and X-kMeans reflect coherent, interpretable topics in the feature space.Topic modeling provided ny LDA can be analysed by domain expert to integrate domain knowledge into to learning process

Beyond alignment, this experiment also evaluates the *transferability* of learned cluster representations. It investigates whether the explainable decision tree model can meaningfully assign semantic labels to previously unseen data, indicating generalizable learning rather than mere overfitting to the training set.

Background and Motivation:

1. **Why k-Means?** K-means is a foundational clustering algorithm that partitions data into k groups based on feature similarity, minimizing intra-cluster variance. It is efficient and widely used in remote sensing, image segmentation, and unsupervised representation learning. However, k-means lacks transparency: while it provides cluster centroids, it offers no explanation for why specific data points belong to certain clusters or which features drive the separation.

   In this framework, k-means serves as a baseline to establish the latent structure of the dataset, providing reference clusters that the X-kMeans model later seeks to explain in a human-interpretable way.

2. **Why LDA?** Latent Dirichlet Allocation (LDA) is a probabilistic generative model that represents complex datasets as mixtures of latent topics, each corresponding to recurring

semantic patterns. Traditionally used in natural language processing, LDA has recently been adapted for visual and remote-sensing data, where words correspond to local image features or pixel-level clusters.

Here, LDA serves a novel role: rather than discovering new clusters, it is used to interpret existing clusters by analyzing how their internal data distributions align with coherent topics. This enables the evaluation of how well k-means and X-kMeans capture meaningful, interpretable structures..

Experimental Design:

1. **Tree Construction (Training Phase)** The X-kMeans decision trees (both *Greedy* and *Non-Greedy* versions) are built using 80% of the images in the datasets. These trees translate the numeric cluster boundaries of k-means into human-interpretable decision rules based on the most discriminative features.

2. **Label Generation (Testing Phase)** The remaining 20% of unseen images in the dataset are used to evaluate the generalizability of both models:
   a) Regular k-means is applied to assign labels directly.
   b) The trained X-kMeans trees assign labels by traversing the learned feature-based rules.

LDA Topic Modelling Each image or data patch is treated as a visual document represented by a bag of visual words. Three separate LDA models are trained:

1. One using the k-means labels, and

2. Two using the X-kMeans labels(Greedy and Non-Greedy).

All 3 X-kMeans models produce topic–word distributions probabilities $P(T \mid W)$ along with visual result images. These distributions provide a means to examine whether clusters derived from X-kMeans correspond to semantically meaningful topics similar to, or more coherent than, those derived from regular k-means.

Label–Topic Explainability Analysis The label–topic explainability analysis aims to interpret how the semantic structure derived from the X-kMeans clusters (labels) aligns with the latent semantic themes identified by LDA topics. The process involves four key stages, as outlined below.

1. **Derivation of Label Semantics (from X-kMeans):** Each cluster label obtained from X-kMeans is interpreted by analyzing its defining feature rules within the decision tree. These rules, combined with domain knowledge (e.g., change in RGB values in overcast conditions, vegetation, or urban surfaces), are used to infer the probable meaning and physical characteristics of each cluster.

2. **Derivation of Topic Semantics (from LDA):** From the LDA output, the $P(T \mid W)$ distributions—representing the probability of each topic given a word—are examined to derive the dominant properties and semantic meaning of each topic. This provides insight into which topics correspond to distinct surface types, material compositions, or land cover patterns.

3. **Semantic Grouping and Cross-Interpretation:** Using the interpretations from the previous steps, both labels (from X-kMeans) and topics (from LDA) are cross-examined by a domain expert. This approach helps in the integraion of domain knowledge in the learning process Topics are grouped or associated with semantic classes (e.g., vegetation areas, aircraft parts, or urban surfaces) based on their feature characteristics and visual or physical correspondence with the clusters.This approach allows

4. **Explainability Confidence Evaluation:** The **Topic Alignment Factor (TAF)** and **Word Explainability Confidence Score (WECS)** are computed to quantitatively assess the strength of alignment between X-kMeans words and LDA topics.

4.5 Alternative Explainability via Word–Topic Distributions

As an alternative to methods such as SHAP and GradCAM, this work proposes an explanation framework that integrates Topic Alignment Factors (TAF), Word Explainability Confidence Scores (WECS), interpretable decision-tree rules, topic–word probability distributions from LDA, domain knowledge, and visual analysis. By incorporating topics as intermediate semantic structures, the framework links low-level decision-tree features with higher-level patterns discovered in the data. Together, these components form a coherent mechanism for deriving, validating, and communicating the semantic meaning of clusters, while the TAF and WECS scores quantify the reliability and consistency of each topic–word relationship, thereby providing confidence-aware explainability in an unsupervised setting:

Decision-Tree → Word Mapping

Each leaf (cluster) of the X-kMeans tree defines a word label. All test pixels assigned to a leaf form a document set for those words.

Topic Distributions per Leaf

For each word(leaf or cluster) w and topic t, $P(t \mid w)$ is computed by aggregating document-level topic proportions of images routed to word w.

Analysis

Identify dominant topics for each word, compare $P(t \mid w)$ between Greedy and Non-Greedy trees, and visualize as stacked heatmaps.

Topic Alignment Factor

The **Topic Alignment Factor (TAF)** quantifies the degree of agreement between the expected and observed topic proportions for a given word. It forms the core component of the *Word Explainability Confidence Score (WECS)* and ensures that each topic's contribution reflects not only its probability but also its empirical consistency with the data.

Formally, for a word w and topic t, the Word Topic Alignment confidence $x_t(w)$ is defined as:

$$x_t(w) = 1 - \left| \left(\frac{[P(t \mid w) \times N_w] - N_t(w)}{N_w} \right)^{\alpha} \right| \tag{4.1}$$

where:

- $P(t \mid w)$ is the probability of topic t given word w, obtained from the LDA topic–word distribution,
- N_w represents the total number of datapoints (e.g., pixels or image patches) associated with word w, and
- $N_t(w)$ denotes the number of datapoints within word w that were actually assigned to topic t by LDA.
- α serves as a tunable parameter that controls the strictness of the TAF. α must be greater than 0. value greater than 0 and lesser than 1 means stricter penalty, 1 means linear penalty and value above 1 means softer penalty

The alignment confidence penalizes discrepancies between the expected topic count (as predicted by the probabilistic distribution $P(t \mid w)$) and the observed topic count (as determined by LDA assignments).

If these two quantities match perfectly, $x_t(w) = 1$, indicating ideal consistency between probabilistic and empirical distributions. As the deviation between them increases, $x_t(w)$ decreases, signaling weaker alignment between the topic model's predictions and actual assignments.

In essence, $x_t(w)$ serves as a *local consistency measure* for each topic–word pair. It captures how faithfully a topic's probabilistic representation reflects the true semantic composition of the word's data points. By incorporating this term into the overall WECS formulation, the framework ensures that topics with poor alignment have a diminished influence on the explainability score, while those with strong correspondence are rewarded.

This design makes the Topic Alignment Factor a crucial intermediary between the statistical inference of LDA and the interpretability evaluation of the explainable clustering pipeline.

4.5.1 Word Explainability Confidence Score

The overall **Word Explainability Confidence Score (WECS)** for a given word w is obtained by combining the topic alignment factors across all topics, weighted by their respective topic probabilities. Formally, it is defined as:

$$\mathrm{WECS}(w) = \sum_{t=1}^{T} \left(x_t(w) \times P(t \mid w) \right)$$

where T denotes the total number of topics in the model.

This formulation ensures that:

- That WECS is normalized to range [0,1], but the topics with higher probability for a given word (i.e., more influential topics) contribute more to the overall score.
- Topics with poor alignment (low $x_t(w)$ values) reduce the overall confidence proportionally.

Thus, *WECS* integrates both probabilistic significance and empirical consistency, providing a balanced measure of how well a word's cluster-level representation aligns with the topic model's inferred structure.

4.5.2 Advantages of TAF and WECS over Standard Alternatives

Problem-Specific: Quantifying confidence Explanations, Not Clusters or Topics

Most existing evaluation scores—such as topic coherence, silhouette coefficient, or Kullback–Leibler (KL) divergence—focus on assessing *cluster quality*, *model fit*, or *distributional divergence*. However, they do not measure how *reliable* the underlying units of explanation are.

The proposed **Topic Alignment Factor (TAF)** is designed specifically to evaluate *explanation reliability* rather than cluster compactness or separation.

TAF penalizes discrepancies between the expected and actual word–topic assignments by considering:

- **Theoretical expectation:** $P(t \mid w) \times N_w$
- **Empirical observation:** $N_t(w)$

For instance, if the topic model predicts that 30% of this word belongs to Topic 5, but Latent Dirichlet Allocation (LDA) assigns 70% of its occurrences to Topic 5, this mismatch indicates unbalanced priors,data drift and model–topic misalignment. None of the standard evaluation measures quantify such over- or under-allocation errors at the word level.

Customizable Strictness and Normalization

The proposed framework introduces a tunable parameter α, which allows users to control how severely misalignments are penalized. A lower α imposes stricter penalization, making the confidence estimation more sensitive in low-tolerance domains.

Furthermore, the **Word Explainability Confidence Score (WECS)** re-normalizes TAF values to the $[0, 1]$ range, ensuring that the resulting scores are both interpretable and comparable across words.

This combination of user-controllable strictness and normalized interpretability enables the method to balance *flexibility* (domain-specific tuning) and *rigor* (consistent mathematical abstraction).

4.6 Results Presentation

Results are reported in separate subsections for each dataset and for each approach (Greedy vs. Non-Greedy). Included are tables (Probability of topic given word ($P(T \mid W)$), Topic Alignment Factor (TAF), Word Explainability Confidence Score (WECS)), figures (decision-tree diagrams, distributions, example images of results), and comparative discussion.

4.7 Summary

This methodology provides a complete experimental pipeline to quantify fidelity of explainable decision-tree surrogates to baseline k-means, compare semantic topic structures, and offer an unique interpretability route through word–topic probability analysis.

5 Experimental Evaluation

This chapter presents and critically analyses the results of the various experiments conducted as part of this study. It systematically evaluates the performance of both the Greedy and Non-Greedy explainable clustering approaches across multiple datasets, highlighting the behavior of the models under different visual and semantic conditions. Quantitative outcomes—such as cluster fidelity, Topic Alignment Factor (TAF), and Word Explainability Confidence Scores (WECS)—are examined alongside qualitative findings derived from decision-tree structures and visual inspection of cluster assignments.

The chapter further interprets these results in relation to the research questions, identifying patterns, strengths, and limitations of the proposed methods and provides a final result.It does so by discussing the analytical results of experments on the two datasets in section 1 and section 2 and finally concluding all thwe results in section 3 By integrating numerical metrics with semantic and visual analysis, this chapter provides a comprehensive understanding of how well the framework captures meaningful structure in remote-sensing imagery and how the two algorithmic variants differ in their interpretability performance.

5.1 Airplane(UCMerced)

Dataset Description: Dataset contains $n = 100$ image patches of size 256×256 pixels. Each pixel is a datapoint.

Size: 6,537,216 samples

Dimensionality: 3

Derived features: 9 (grayscale, LBP, gradient magnitude, mean RGB, LBP entropy, edge density, dominant orientation)

Reason: The UCMerced Land Use dataset (Airplane subset) was selected as it visually differentiates between distinct classes, making the interpretation of semantic meaning more straightforward. This characteristic is particularly valuable for evaluating explainability in scenarios where ground truth labels are unavailable. The airplane scenes exhibit clear geometric and textural patterns such as runways, aircraft bodies, and varying ground surfaces, which challenge the model to discriminate between subtle spatial and structural variations. Such diversity provides an effective basis for assessing how well the explainable X-kMeans decision tree captures feature importance, hierarchical structure, and spatial boundaries in a visually interpretable manner. Furthermore, the dataset's moderate size of 100 images makes it suitable for controlled experimentation and cross-validation, while maintaining sufficient complexity to test the generalizability and transferability of the learned semantics

5.1.1 Cluster Fidelity for airplane experiment

Cluster Fidelity experiment involves performing 3 fold cross validation and finding cluster fidelity

Non-Greedy Cluster fidelity for Airplane Dataset

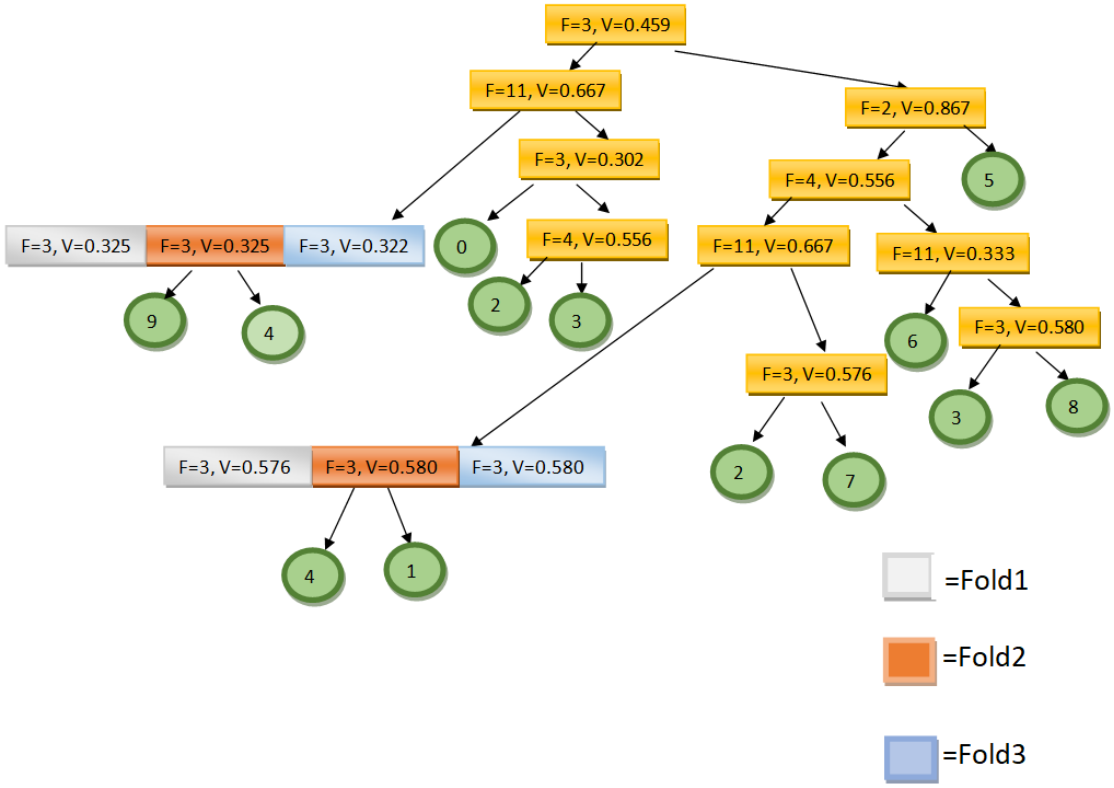

Figure 5.1: Non-Greedy Cross Validation Decision Trees

Figure(5.1) shows **Non-Greedy IMM decision trees** for the Airplane dataset across three cross-validation folds. **F3** is grayscale, **F2** is blue, **F4** is LBP and **F11** is dominant orientation. Internal nodes (shown as **yellow rectangles**) denote feature–threshold splits, where value of F indicates the feature index and V denotes the threshold value, while leaf nodes (shown as **green circles**) correspond to k-means cluster labels. Due to center sharing, some clusters appear along multiple decision paths, reflecting smoother and less restrictive decision boundaries compared to Greedy IMM.

Node colors indicate the three cross-validation folds: light grey for **Fold 1**, light brown for **Fold 2**, and light blue for **Fold 3**. Yellow colour nodes are shared across all Folds. Each fold represents a different train–test partition of the dataset. Minor variations in split thresholds

and tree topology across folds highlight sensitivity to data partitioning, whereas the overall structure and dominant decision rules remain consistent.

OutCome

The three-fold stratified cross-validation yields cluster fidelity scores of 91.33%, 91.32%, and 91.30% for the respective Non-Greedy folds with an average fidelity of 91.31%. The decision tree shows slight variations in threshold values of Feature 3 at height of 1 in the left half of the decision tree

Greedy Cluster fidelity for Airplane Dataset

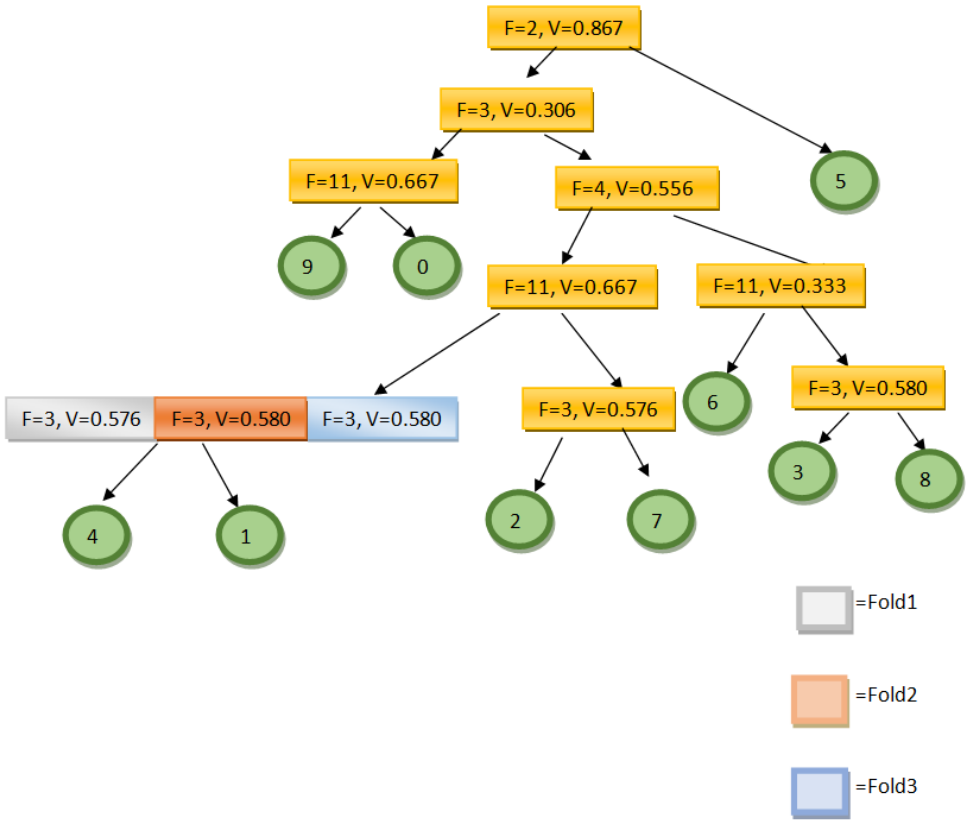

Figure 5.2: Greedy Cross Validation Decision Tree

Figure(5.2) shows **Greedy IMM decision trees** for the Airplane dataset across three cross-validation folds. **F3** is grayscale, **F2** is blue, **F4** is LBP and **F11** is dominant orientation. Internal nodes (shown as **yellow rectangles**) denote feature–threshold splits, where value of F indicates the feature index and V donating the threshold value, while leaf nodes(shown as **green circles**) correspond to k-means cluster labels. Unlike the Non-Greedy approach, each

cluster is associated with a single unique decision path, reflecting the strict one-to-one mapping enforced by Greedy optimization.

Node colors indicate the three cross-validation folds: light grey for **Fold 1**, light brown for **Fold 2**, and light blue for **Fold 3**. Yellow colour nodes are shared across all Folds. Each fold represents a different train–test partition of the dataset. Minor variations in split thresholds and tree topology across folds highlight sensitivity to data partitioning, whereas the overall structure and dominant decision rules remain consistent.

OutCome

The three-fold stratified cross-validation yields cluster fidelity scores of 92.80%, 92.77%, and 92.76% for the respective Greedy folds with the average fidelity of 92.78%. The decision tree shows slight variations in threshold values of Feature 3 at height of 1

Conclusion

The high and consistent fidelity across folds indicates that the X-kMeans decision tree generalizes well to data, with minimal dependence on specific training samples. The marginal variation (less than 0.05%) confirms the stability and robustness of the model's learned partitioning logic.

Furthermore, as shown in Figures 5.1 and 5.2, the resulting decision trees are structurally identical across folds, differing only by minor variations in numerical thresholds for a few nodes. This structural consistency suggests that feature selection and hierarchical splits are not random artifacts of a particular training subset but reflect genuinely discriminative relationships within the data.

Overall, these results demonstrate that the X-kMeans framework preserves the underlying k-Means clustering structure with high fidelity while providing an interpretable, rule-based representation. The stable fidelity scores also reinforce the reliability of using X-kMeans as a surrogate explainable model for large-scale or unseen data without significant loss in cluster assignment accuracy.

5.1.2 Experiment-2 Explainability via LDA (Airplane Experiment)

In this experiment, the objective is to evaluate the **semantic interpretability** of clusters generated by three different clustering methods:

1. Regular **k-Means**,
2. **Greedy X-kMeans**, and
3. **Non-Greedy X-kMeans**,

through the use of **Latent Dirichlet Allocation (LDA)**.

The experiment is conducted on the **Airplane** class from the **UC Merced** dataset, which contains a total of 100 images. The dataset is divided into an 80:20 ratio, where the first 80 images are used to construct the decision trees for both the Greedy and Non-Greedy variants of X-kMeans. The remaining 20 images serve as a test set to assess the *transferability* of the learned cluster representations.

Clustering Pipelines

Each test image is passed through three clustering pipelines:

1. Regular k-Means,
2. Greedy X-kMeans Decision Tree, and
3. Non-Greedy X-kMeans Decision Tree.

This process produces three distinct sets of cluster labels corresponding to the same 20 test images. These label sets are subsequently used as inputs to the LDA model, where each image is treated as a visual document and its cluster assignments as visual words.

Topic Modelling via LDA

From the trained LDA model, the topic–word probability distribution $(P(T \mid W))$ is obtained for both Greedy and Non-Greedy decision tree outputs. This distribution represents how likely each word (i.e., decision-tree leaf or cluster) is associated with each latent topic. To evaluate how reliably a word's inferred meaning transfers to its assigned topic, a Topic Alignment Factor (TAF) is computed, measuring the agreement between expected and observed topic assignments. Finally, the Word Explainability Confidence Score (WECS) is calculated to summarize the overall contribution of each word to the LDA results, providing an interpretable measure of how strongly and consistently a word supports the semantic structure of the topic model.

Decision Trees

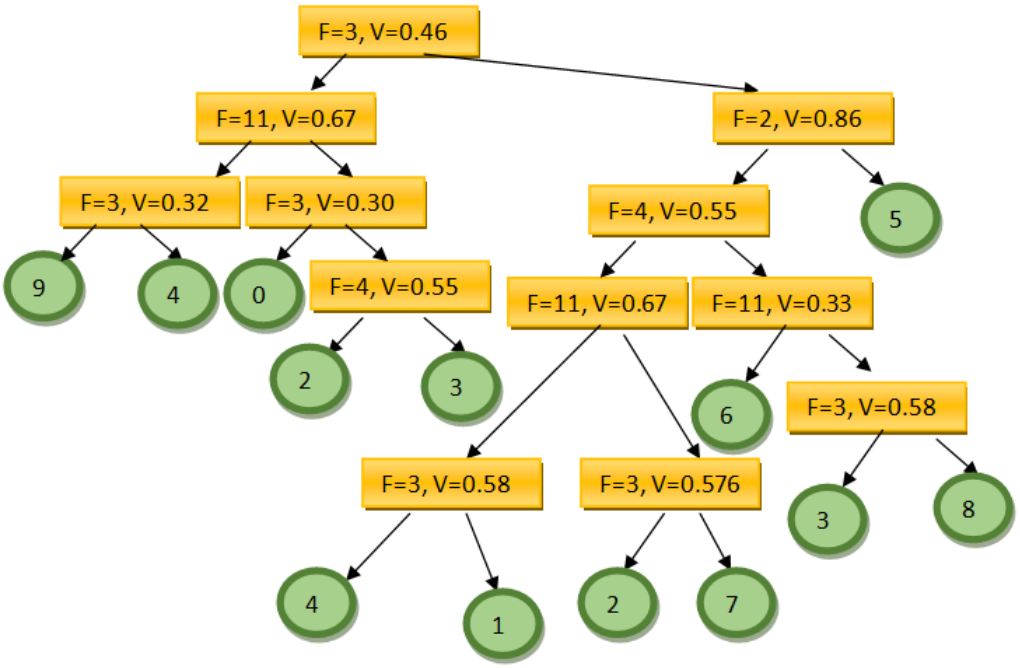

Figure 5.3: Airplane Non-Greedy Decision Tree

Figure 5.3 presents the decision tree generated for the Airplane dataset using the Non-Greedy X-kMeans approach. **F3** is grayscale, **F2** is blue, **F4** is LBP and **F11** is dominant orientation. Unlike the Greedy variant, the Non-Greedy method does not permanently commit to the earliest locally optimal split. Instead, it evaluates multiple branching alternatives at each stage, allowing several feature-based separation paths to remain valid for a single cluster label. As a result, some labels—such as label 2—emerge from more than one root-to-leaf path in the tree

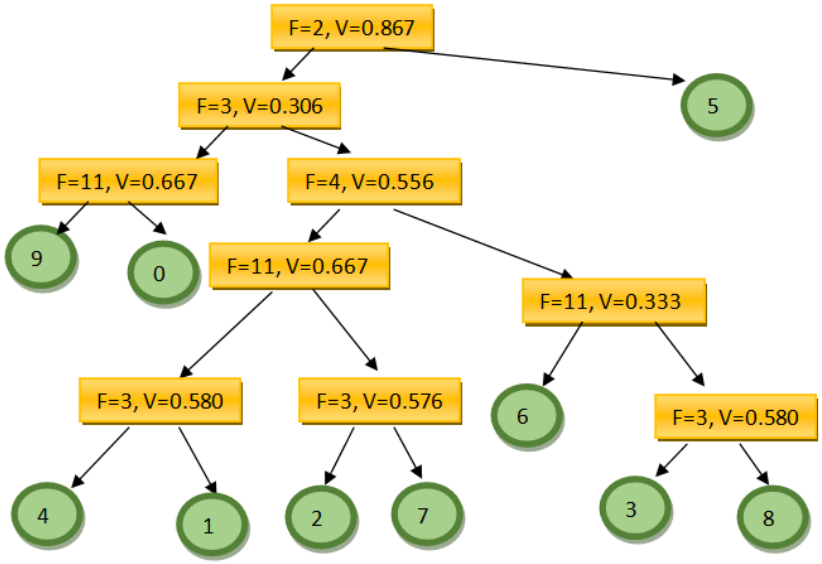

Figure 5.4: Airplane Greedy Decision Tree

Figure 5.4 presents the decision tree generated for the Airplane dataset using the Greedy X-kMeans approach. **F3** is grayscale, **F2** is blue, **F4** is LBP and **F11** is dominant orientation. In contrast to the Non-Greedy formulation, the Greedy version selects the locally optimal split at each node without maintaining a buffer region around decision boundaries. As a result, each cluster label is associated with a single deterministic feature pathway, even if multiple reasonable separation patterns exist in the data. This produces a more compact and computationally efficient tree, but can also result in reduced semantic richness, as fine-grained variations within a class may be absorbed into neighbouring branches. Consequently, the Greedy decision tree tends to offer simpler but less expressive explanations compared to the multi-path structures observed in the Non-Greedy approach.

Explainability Metrics

To quantitatively assess the degree of alignment between topics and words, two explainability metrics are computed:

- **Topic Alignment Factor (TAF):** Measures the consistency between expected and

observed topic assignments.

- **Word Explainability Confidence Score (WECS):** Integrates TAF values across topics, weighted by their respective probabilities.

Evaluation Outputs

For each dataset configuration, three heatmaps are presented:$P(T \mid W)$ – Topic–Word Probability Distribution,TAF – Topic Alignment Factor Scores, and WECS – Word Explainability Confidence Scores.

These results collectively enable the evaluation of how effectively the X-kMeans framework captures semantically coherent topics and aligns them with meaningful visual patterns when compared to standard k-Means.

LDA results — analysis & comparison

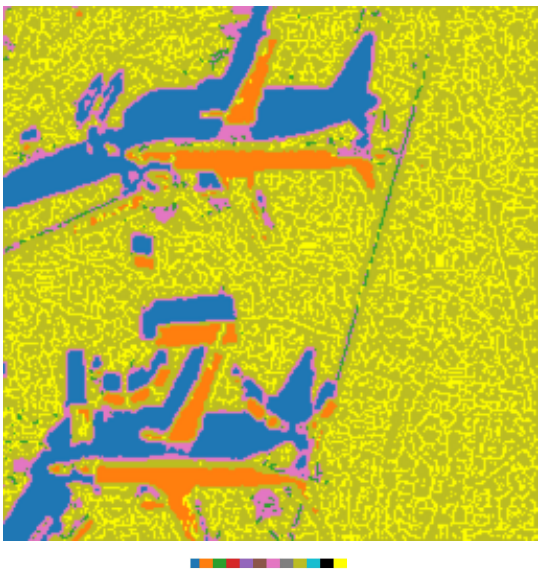

Figure 5.5: Non-Greedy X-kMeans Image 1

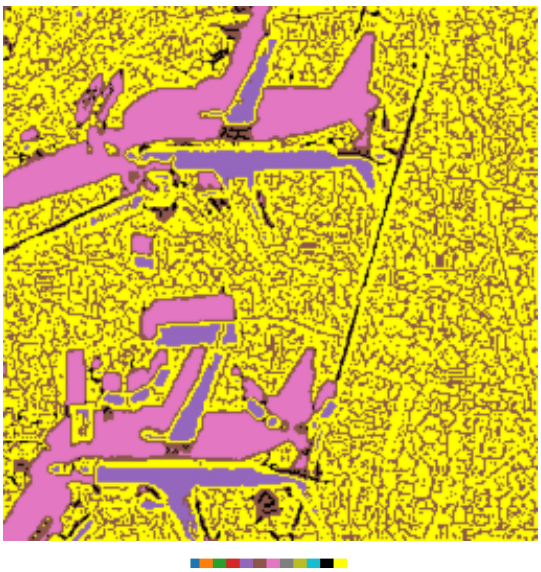

Figure 5.6: Greedy X-kMeans image 1

Figure 5.7: Original image 1

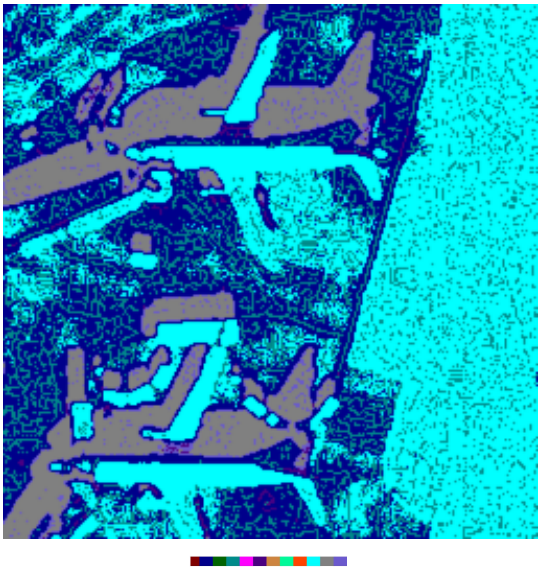

Figure 5.8: Regular k-means image 1

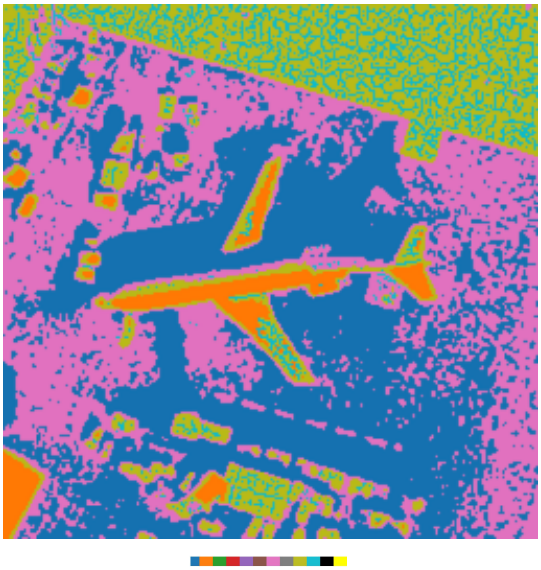

Figure 5.9: Non-Greedy X-kMeans Image 2

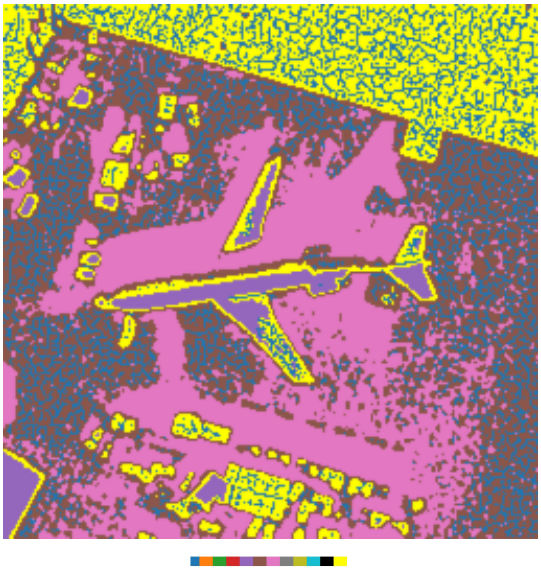

Figure 5.10: Greedy X-kMeans Image 2

Figure 5.11: Original Image 2

Figure 5.12: Regular k-means image 2

Figures (5.5-5.12) compare the **regular k-means** and both X-kMeans results with the help of the original. Below different kmneans results are the topic colors legends indicating which topic is assigned which color in the ascending order. **Topic 0** is the first color in the legend and **Topic 11** is the last

Analysis: A comparison of the Greedy and Non-Greedy decision trees for the Airplane experiment shows that a large proportion of leaf nodes follow identical or nearly identical feature pathways. Only a small number of leaves in the Non-Greedy tree diverge into multiple paths, reflecting areas where the algorithm retains ambiguity to capture alternative interpretations of similar regions.Non-Greedy results tend to better differentiate metallic edges from road edges in some cases(Figure 5.5 and 5.6). Otherwise, the differences observed in the downstream LDA topic assignments are relatively small, as the semantic structure induced by both trees remains largely consistent

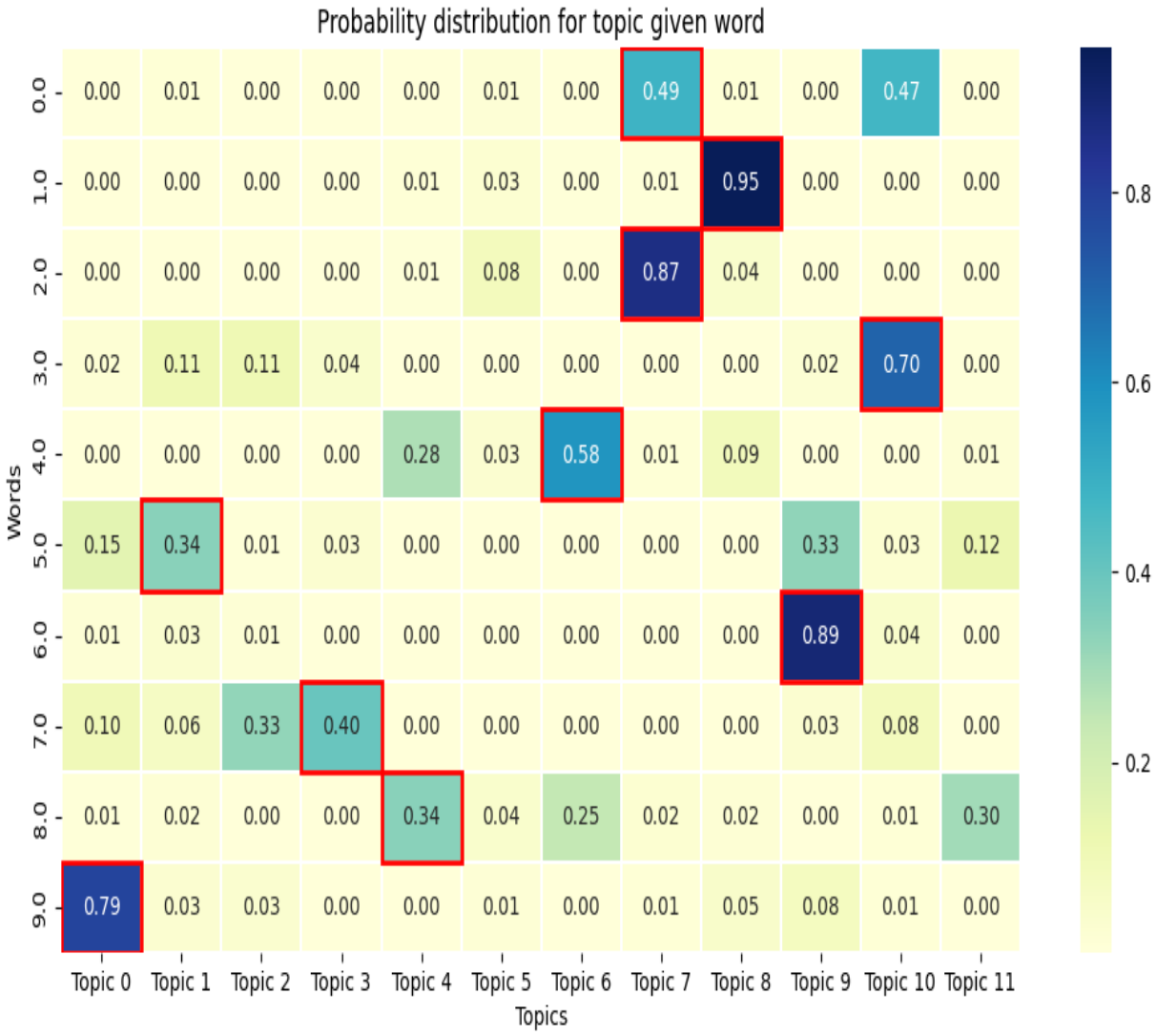

Figure 5.13: Airplane(Non-Greedy) $P(T \mid W)$

The heatmap in Figure 5.13 illustrate the conditional probability distribution of topics given each word for the Non-Greedy airplane experiment. For six words out of ten, the distribution is sharply peaked, with at least one topic receiving a dominant probability (more than 50%). This concentration simplifies the semantic interpretation process, since a clear correspondence can often be established between a word and its most strongly associated topic. In contrast,four words have more diffused distributions where a word exhibits small(less than 50%), scattered probabilities across multiple topics. These words maybe carrying semantic meaning that aligns with multiple classes, thus making their interpretability more difficult, something expected from Non-Greedy approach as multiple paths to a word can cause its semantic meaning to be

more diverse but less interperatible.

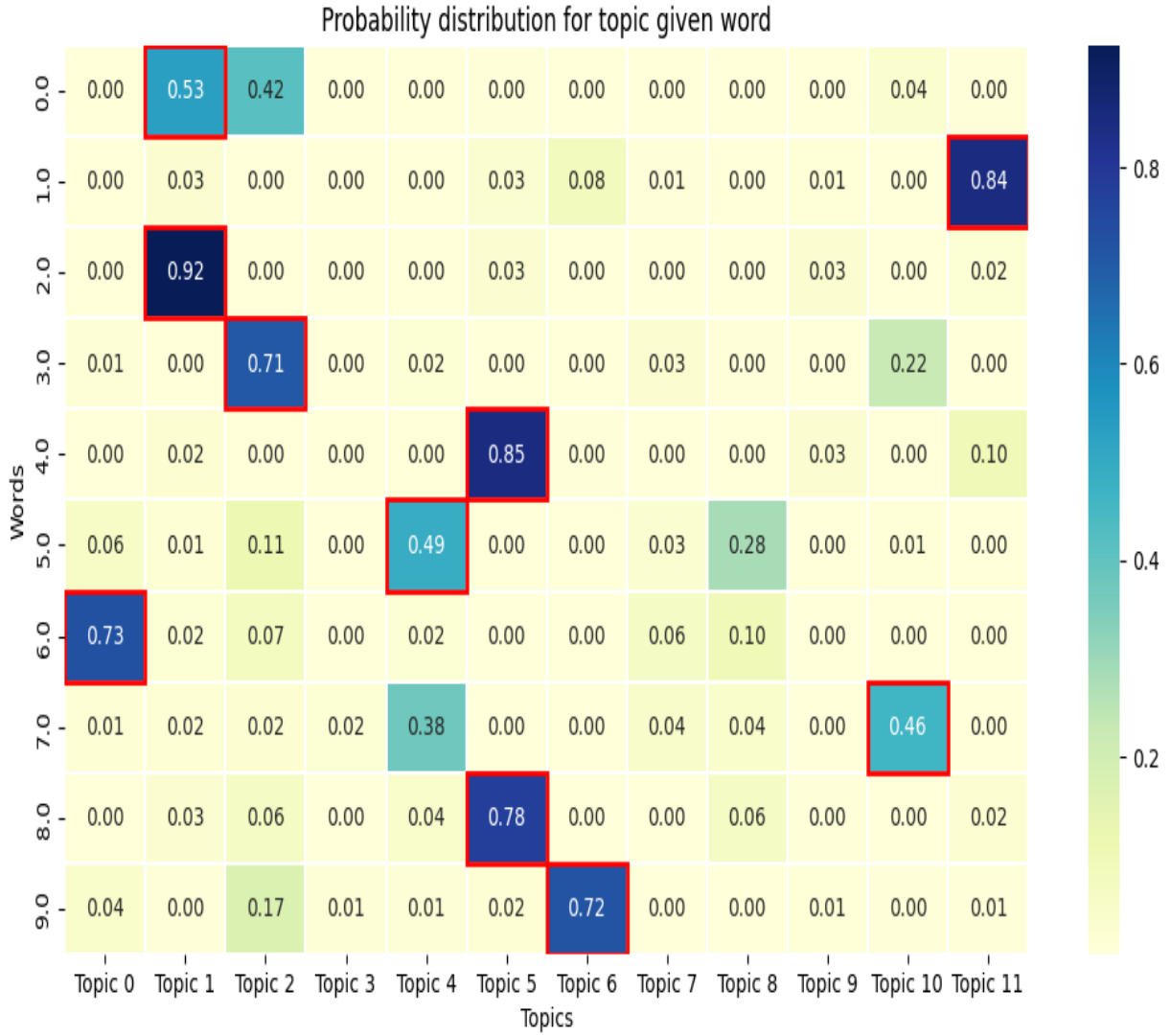

Figure 5.14: Airplane(Greedy) $P(T \mid W)$

The heatmap in Figure 5.14 illustrate the conditional probability distribution of topics given each word for the Greedy airplane experiment. For the majority of words(eight), the distribution is sharply peaked, with at least one topic receiving a dominant probability (more than 50%). This concentration simplifies the semantic interpretation process, since a clear correspondence can often be established between a word and its most strongly associated topic. In contrast, two words have more diffused more diffuse distributions—where a word exhibits small(less than 50%), scattered probabilities across multiple topics. Notably, this proportion is approximately half

of that observed in the Non-Greedy approach, which is consistent with the Greedy strategy's tendency to suppress boundary-case semantic variations. While this results in reduced semantic diversity, it yields easier to interperate decisive topic–word associations.

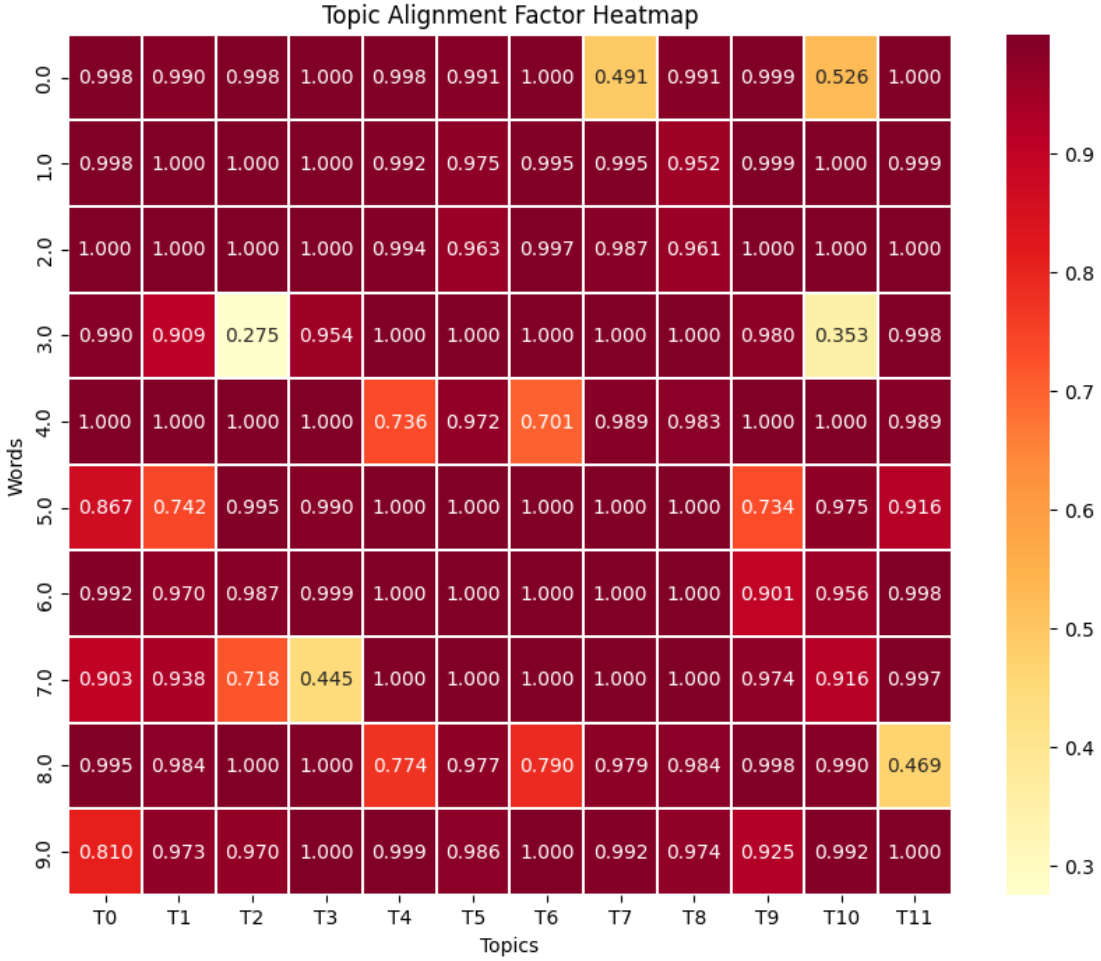

Figure 5.15: Airplane(Non-Greedy) TAF

The heatmap in Figure 5.15 displays the **Topic Alignment Factor (TAF)** values for the Non-Greedy experiment. This visualization allows evaluation of how effectively the probabilistic assignment $P(T \mid W)$ corresponds to the actual topic allocations produced by LDA. For instance, **Word 0** exhibits a relatively low TAF for **Topic 7**, whereas **Word 2** shows a noticeably higher alignment score. This indicates that, although both words may have their highest respective probabilities belonging to Topic 7, Word 2 aligns more consistently with the empirical topic assignments and therefore provides a more reliable semantic explanation of Topic 7, as confirmed by visual analysis. Such word topic level analysis can be very useful and TAF gives us a metric how much faith LDA itself has in that analysis.

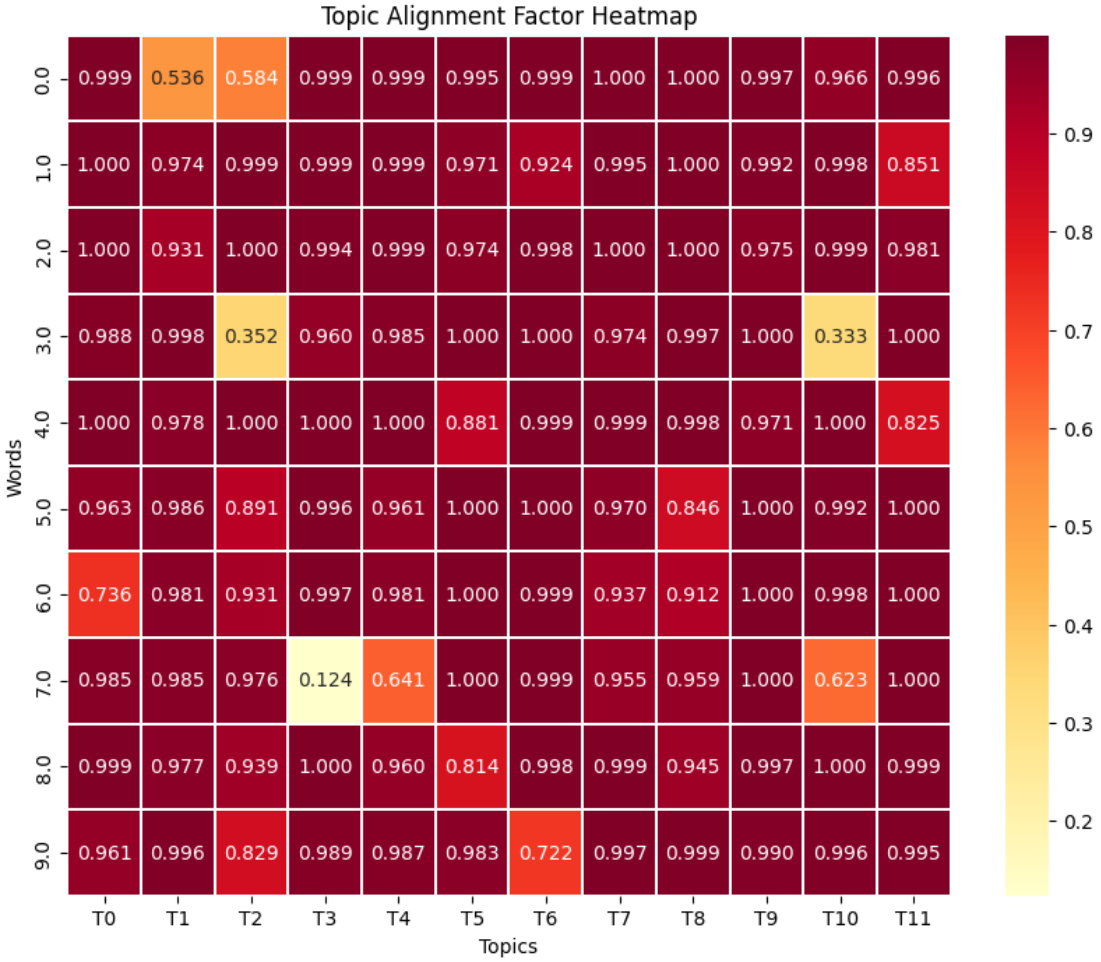

Figure 5.16: Airplane(Greedy) TAF

The heatmap in Figure 5.16 displays the **Topic Alignment Factor (TAF)** values for the Greedy experiment.This visualization allows evaluation of how effectively the probabilistic assignment $P(T \mid W)$ corresponds to the actual topic allocations produced by LDA.For instance,For instance, **Word 0** exhibits a relatively low TAF for **Topic 1**, whereas **Word 2** shows a noticeably higher alignment score. These patterns are similar to Non-Greedy approach topic 7. Consequently, the TAF heatmap provides a direct mechanism to assess and compare how well individual words contribute to the interpretation of topic semantics within the explainable clustering framework.

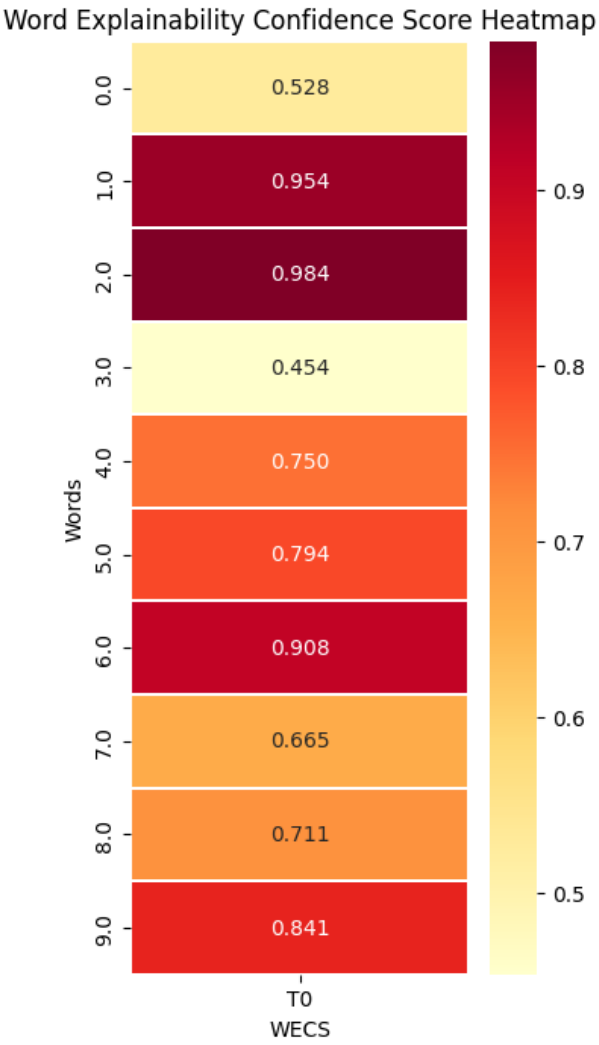

Figure 5.17: Airplane(Non-Greedy) WECS

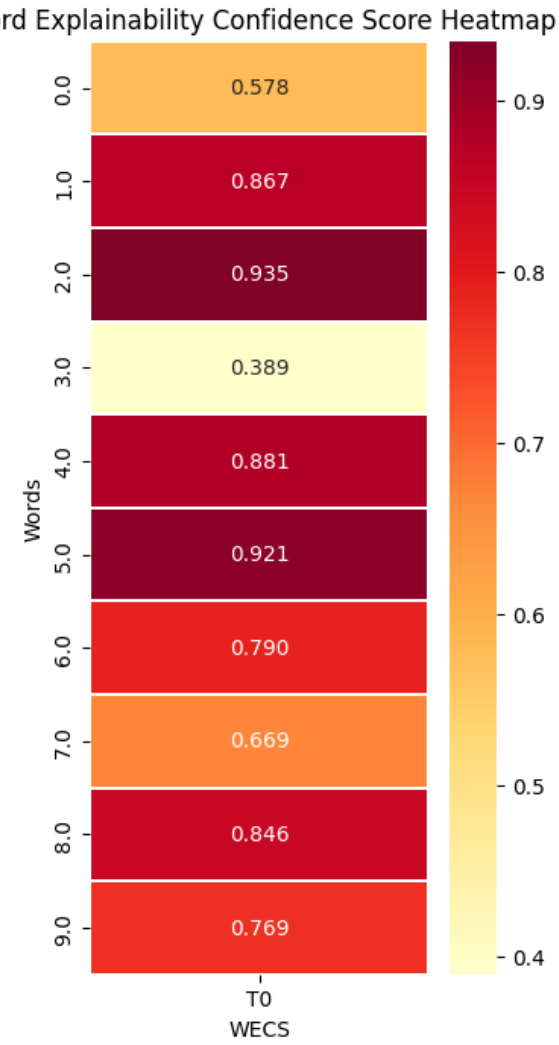

Figure 5.18: Airplane(Greedy) WECS

Figures 5.17 and 5.18 present the **Word Explainability Confidence Scores (WECS)** for the airplane experiment. A comparison of the two heatmaps shows that the WECS values for corresponding words in the **Greedy** and **Non-Greedy** variants generally differ within a margin of approximately 10–20%, which can be mostly attributed to different initial conditions of LDA.

Conclusion:

While there is some noticeable difference between the LDA outputs generated by Non-Greedy and Greedy approach it is not significant.This can be attributed to relative ease of differentiating between different objects or classes in the image which leads to easier capture of the semantic meaning. Despite these small differences, the overall topic–word structures remain consistent, indicating that both decision-tree construction strategies capture similar semantic behaviour in the data, with the Non-Greedy approach providing slightly stronger alignment in a few cases. This is further supported by the WECS scores, which remain close in value for corresponding words across both approaches, indicating that both trees produce similar levels of topic–word alignment. Overall, the limited deviation between the two models suggests that, for this dataset, the Greedy strategy does not significantly weaken semantic interpretability.

5.2 Baseball Diamond(UCMerced)

Dataset Description: Dataset contains $n = 100$ image patches of size 256×256 pixels. Each pixel is a datapoint.

Size: 6,537,216 samples

Dimensionality: 3

Derived features: 9 (grayscale, LBP, gradient magnitude, mean RGB, LBP entropy, edge density, dominant orientation)

Reason: The UCMerced Land Use dataset (Baseball Diamond subset) was selected because it contains visually similar classes that differ only through subtle semantic variations—for example, different types of grass, clay, and sand observed under varying illumination conditions. This makes the dataset more challenging than categories with strong visual separability and is therefore well suited for evaluating whether the Non-Greedy approach offers advantages over the Greedy variant. In such scenarios, even small differences in feature boundaries, textures, or lighting should be reflected in the structure of the learned decision rules, allowing a meaningful comparison of the two methods in terms of their ability to extract fine-grained and semantically relevant distinctions.

5.2.1 Cluster Fidelity for baseball diamond experiment

Cluster Fidelity experiment involves performing 3 fold cross validation and finding cluster fidelity

Non-Greedy Cluster fidelity for baseball diamond Dataset

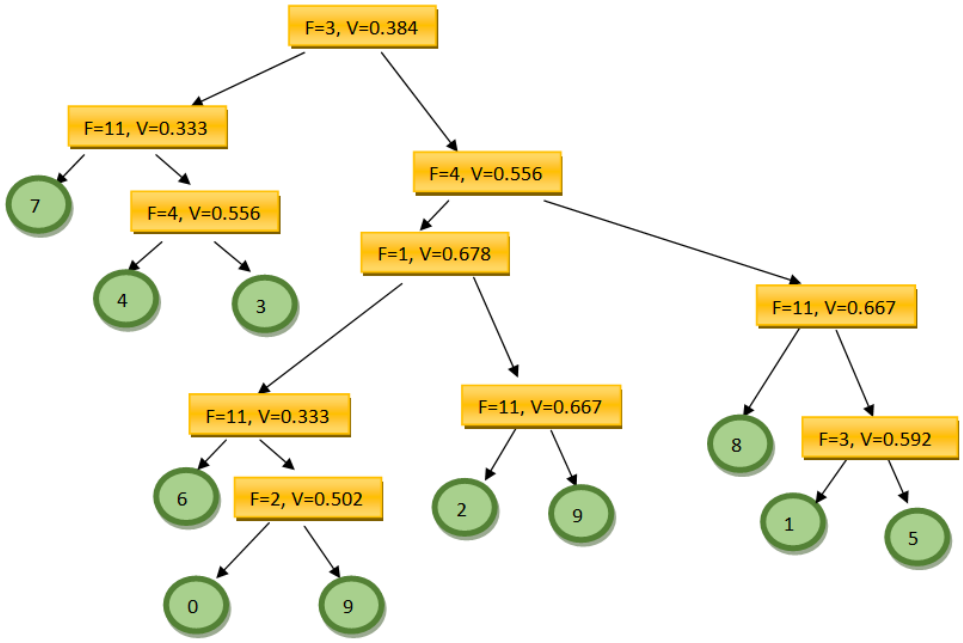

Figure 5.19: Non-Greedy Cross Validation Decision Trees

Figure(5.19) shows **Non-Greedy IMM decision trees** for the Baseball diamond dataset across three cross-validation folds.**F3** is grayscale, **F2** is blue,**F1** is Green, **F4** is LBP and **F11** is dominant orientation. Internal nodes (shown as **yellow rectangles**) denote feature–threshold splits, where value of F indicates the feature index and V donating the threshold value, while leaf nodes(shown as **green circles**) correspond to k-means cluster labels. Due to center sharing, some clusters appear along multiple paths, reflecting smoother and less restrictive decision boundaries compared to Greedy IMM.

All 3 folds in this experiment had identical decision trees, indicating dominant decision rules remain consistent across all folds.

OutCome

The three-fold stratified cross-validation yields cluster fidelity scores of 86.22%, 86.17%, and 86.19% for the respective Non-Greedy folds with average fidelity of 86.19%. The decision trees for all 3 folds are completely identical

Greedy Cluster fidelity for baseball diamond Dataset

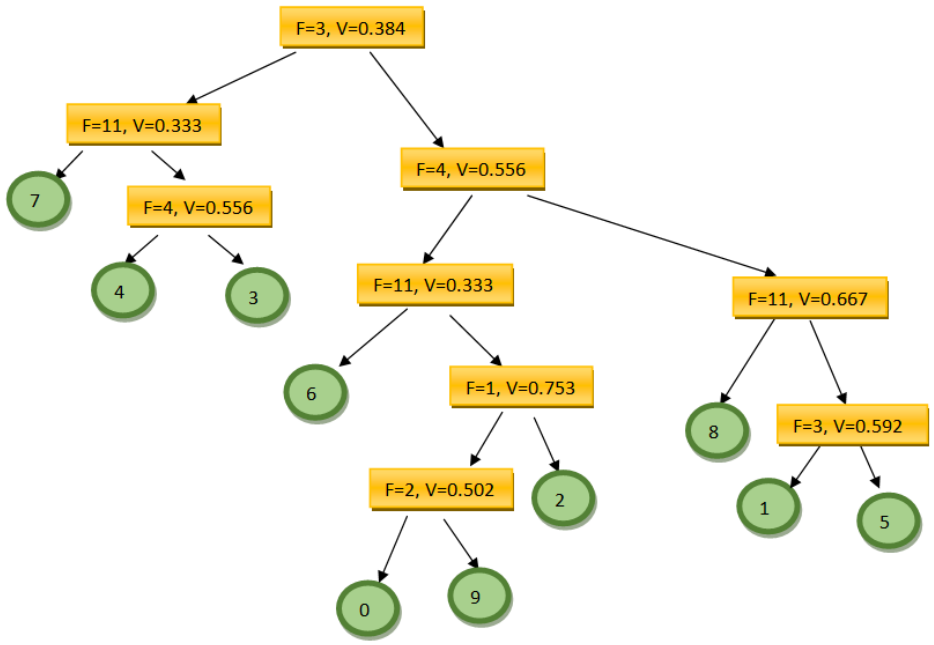

Figure 5.20: Greedy Cross Validation Decision Trees

Figure(5.20) shows **Greedy IMM decision trees** for the Baseball diamond dataset across three cross-validation folds. **F3** is grayscale, **F2** is blue,**F1** is Green, **F4** is LBP and **F11** is dominant orientation. Internal nodes (shown as **yellow rectangles**) denote feature–threshold splits, where value of F indicates the feature index and V donating the threshold value, while leaf nodes(shown as **green circles**) correspond to k-means cluster labels. Unlike the Non-Greedy approach, each cluster is associated with a **single unique decision path**, reflecting the strict one-to-one mapping enforced by Greedy optimization.

All 3 folds in this experiment had identical decision trees.

OutCome

The three-fold stratified cross-validation yields cluster fidelity scores of 83.23%, 83.20%, and 83.23% for the respective Greedy folds with average fidelity of 83.22%. The decision trees for all 3 folds are completely identical indicating dominant decision rules remain consistent across all folds

Conclusion

The high and consistent fidelity across folds indicates that the X-kMeans decision tree generalizes well to data, with minimal dependence on specific training samples. The marginal variation (less than 0.05%) confirms the stability and robustness of the model's learned partitioning logic.

Furthermore, as shown in Figures 5.19 and 5.20, the resulting decision trees are identical across folds. This structural consistency suggests that feature selection and hierarchical splits are not random artifacts of a particular training subset but reflect genuinely discriminative relationships within the data.

Overall, these results demonstrate that the X-kMeans framework preserves the underlying k-Means clustering structure with high fidelity while providing an interpretable, rule-based representation. The stable fidelity scores also reinforce the reliability of using X-kMeans as a surrogate explainable model for large-scale or unseen data without significant loss in cluster assignment accuracy.

5.2.2 Experiment-2 Explainability via LDA (baseball diamond Experiment)

In this experiment, the objective is to evaluate the **semantic interpretability** of clusters generated by three different clustering methods:

1. Regular **k-Means**,
2. **Greedy X-kMeans**, and
3. **Non-Greedy X-kMeans**,

through the use of **Latent Dirichlet Allocation (LDA)**.

The experiment is conducted on the **Baseball diamond** class from the **UC Merced** dataset, which contains a total of 100 images. The dataset is divided into an 80:20 ratio, where the first 80 images are used to construct the decision trees for both the Greedy and Non-Greedy variants of X-kMeans. The remaining 20 images serve as a test set to assess the *transferability* of the learned cluster representations.

Clustering Pipelines

Each test image is passed through three clustering pipelines:

1. Regular k-Means,
2. Greedy X-kMeans Decision Tree, and
3. Non-Greedy X-kMeans Decision Tree.

This process produces three distinct sets of cluster labels corresponding to the same 20 test images. These label sets are subsequently used as inputs to the LDA model, where each image is treated as a document and its cluster assignments as visual words.

Topic Modelling via LDA

From the trained LDA model, the topic–word probability distribution $(P(T \mid W))$ is obtained for both Greedy and Non-Greedy decision tree outputs. This distribution represents how likely each word (i.e., decision-tree leaf or cluster) is associated with each latent topic. To evaluate how reliably a word's inferred meaning transfers to its assigned topic, a Topic Alignment Factor (TAF) is computed, measuring the agreement between expected and observed topic assignments. Finally, the Word Explainability Confidence Score (WECS) is calculated to summarize the

overall contribution of each word to the LDA results, providing an interpretable measure of how strongly and consistently a word supports the semantic structure of the topic model.

Decision Trees

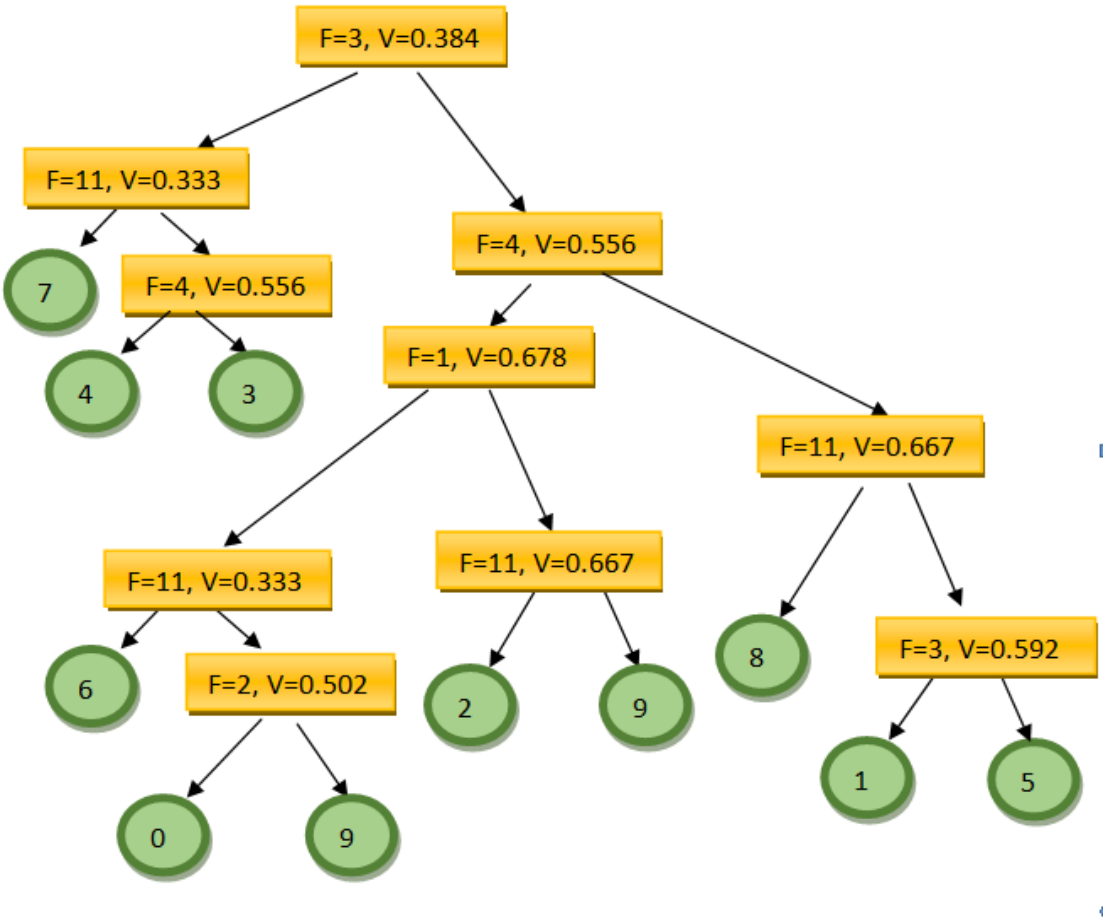

Figure 5.21: Baseball Diamond Non-Greedy Decision Tree

Figure 5.21 presents the decision tree generated for the Baseball diamond dataset using the **Non-Greedy X-kMeans** approach. **F3** is grayscale, **F2** is blue,**F1** is red, **F4** is LBP and **F11** is dominant orientation.

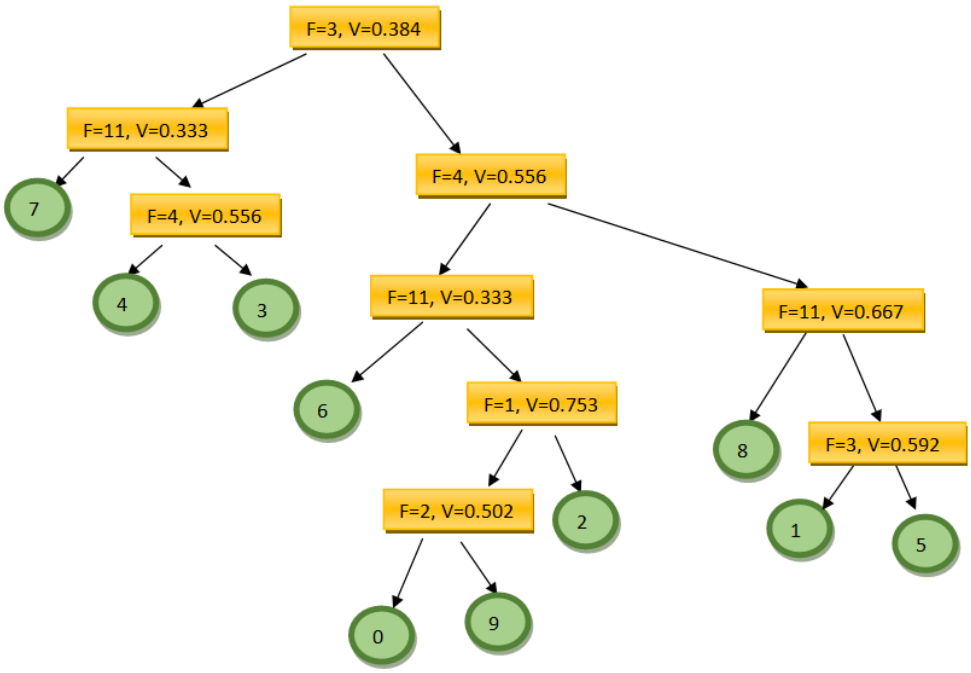

Figure 5.22: Baseball Diamond Greedy Decision Tree

Figure 5.22 presents the decision tree generated for the Baseball diamond dataset using the **Greedy X-kMeans** approach. **F3** is grayscale, **F2** is blue,**F1** is red, **F4** is LBP and **F11** is dominant orientation.

Explainability Metrics

To quantitatively assess the degree of alignment between topics and words, two explainability metrics are computed:

- **Topic Alignment Factor (TAF):** Measures the consistency between expected and observed topic assignments.
- **Word Explainability Confidence Score (WECS):** Integrates TAF values across topics, weighted by their respective probabilities.

Evaluation Outputs

For each dataset configuration, three heatmaps are presented:$P(T \mid W)$ – Topic–Word Probability Distribution,TAF – Topic Alignment Factor Scores, and WECS – Word Explainability Confidence Scores.

These results collectively enable the evaluation of how effectively the X-kMeans framework—particularly the Non-Greedy variant—captures semantically coherent topics and aligns them with meaningful visual patterns when compared to both Greedy X-kMeans and standard k-Means.

LDA results — analysis & comparison

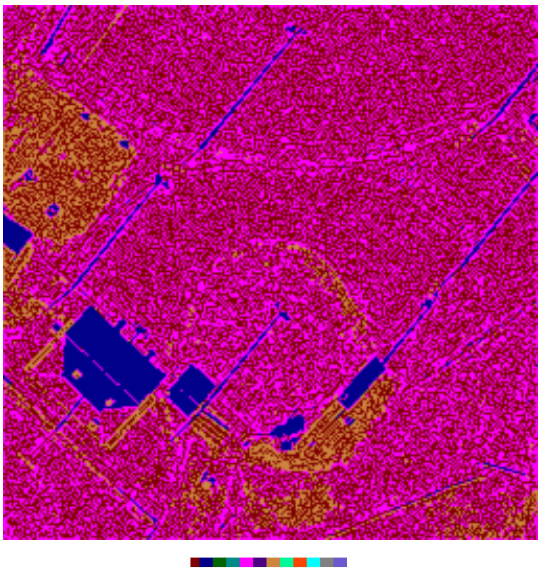

Figure 5.23: Non-Greedy image 1

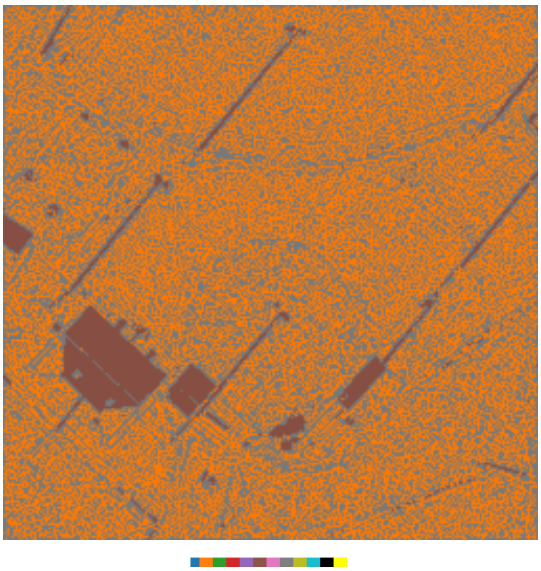

Figure 5.24: Greedy image 1

Figure 5.25: Original image 1

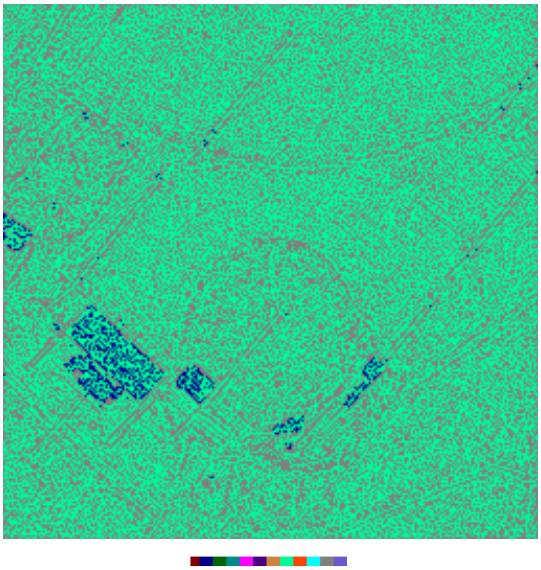

Figure 5.26: Regular k-means image 1

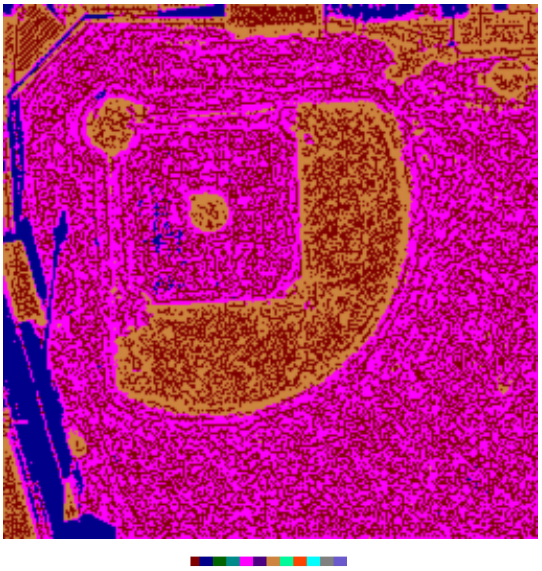

Figure 5.27: Non-Greedy image 2

Figure 5.28: Greedy image 2

Figure 5.29: Original image 2

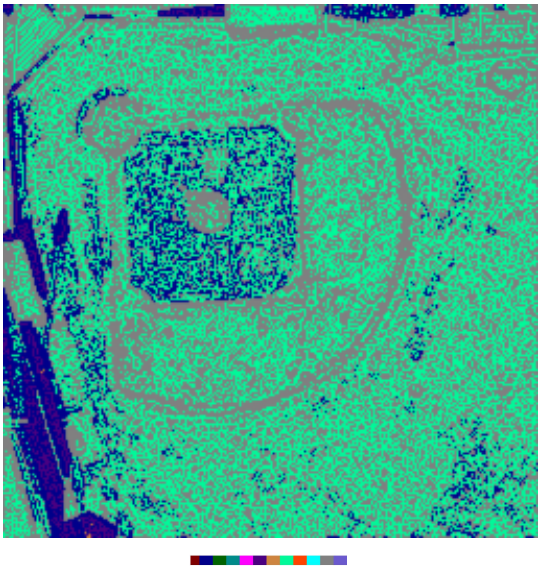

Figure 5.30: Regular k-means image 2

Figures (5.23-5.30) compare the **regular k-means** and both X-kMeans results with the help of the original. Below different kmneans results are the topic colors legends indicating which topic is assigned which color in the ascending order. **Topic 0** is the first color in the legend and **Topic 11** is the last

Analysis: A comparison of the Greedy and Non-Greedy decision trees for the Baseball experiment reveals noticeable differences in their ability to separate semantically meaningful regions. In the first example(Figure 5.23), the Non-Greedy approach distinguishes the concrete areas from the rest of the baseball field more accurately than both standard k-means and the Greedy X-kMeans variant. In the second example, the Non-Greedy method again provides clearer separation between clay and grass regions, whereas the other approaches produce more mixed label assignments. However, all three methods exhibit limitations in certain cases—for instance, misclassifying clay regions and building rooftops as same—indicating that some semantic ambiguities remain challenging to resolve without more specialized feature engineering or higher-resolution domain cues

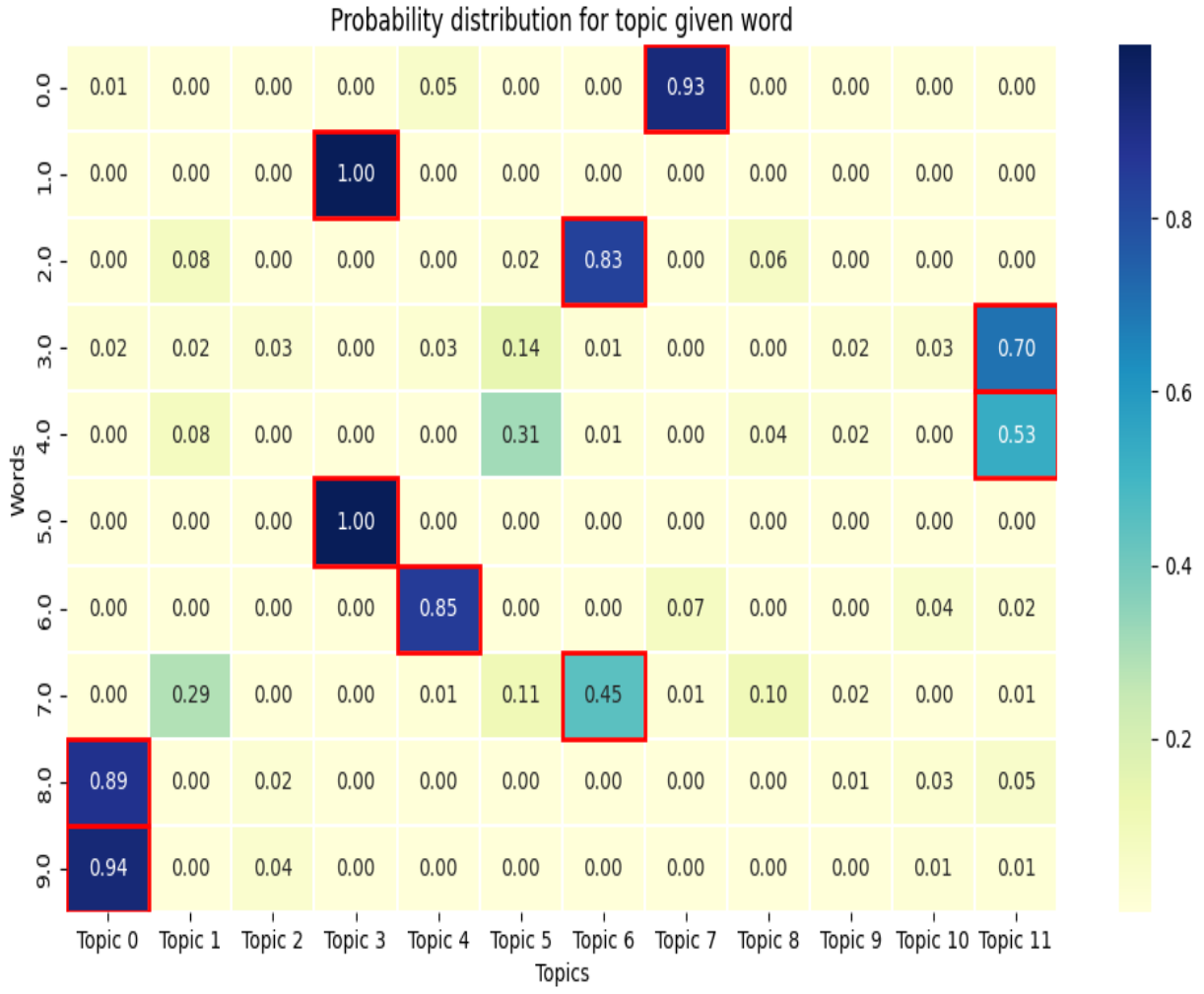

Figure 5.31: Baseball(Non-Greedy) $P(T \mid W)$

The heatmap in Figure 5.31 illustrate the conditional probability distribution of topics given each word for the Non-Greedy baseball diamond experiment. For the majority of words(eight), the distribution is sharply peaked, with at least one topic receiving a dominant probability (more than 50%). This concentration simplifies the semantic interpretation process, since a clear correspondence can often be established between a word and its most strongly associated topic. In contrast,two words have more diffused distributions where a word exhibits small(less than 50%), scattered probabilities across multiple topics. These words maybe carrying semantic meaning that aligns with multiple classes, thus making their interpretability more difficult, something expected from Non-Greedy approach as multiple paths to a word can cause its semantic meaning to be more diverse but less interperatible. Compared to the airplane experiment, a greater number of words in the baseball experiment exhibit a dominant topic assignment. This difference may be attributed to the reduced number of words with multiple contributing paths in the baseball case (one) relative to the airplane case (three).

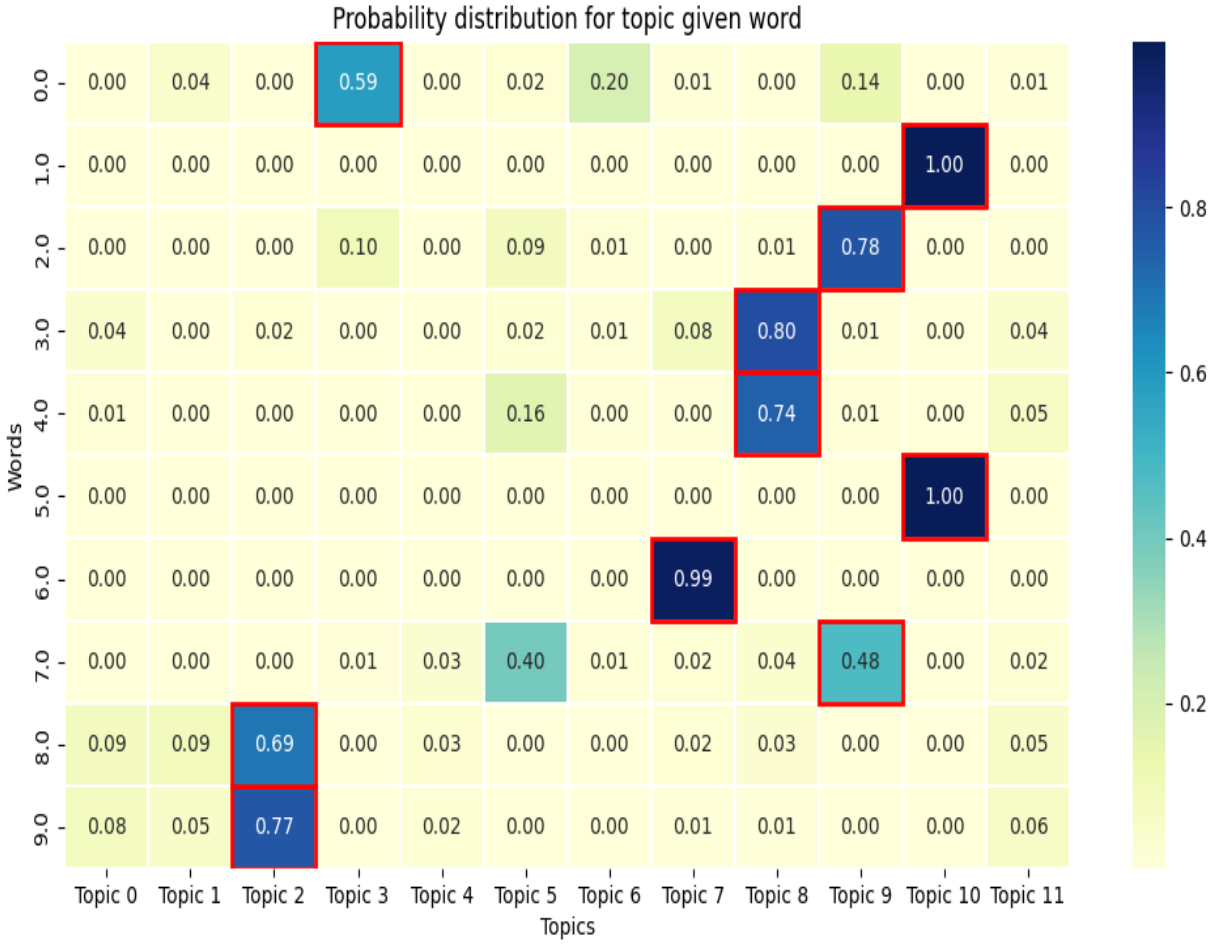

Figure 5.32: Baseball(Greedy) $P(T \mid W)$

The heatmap in Figure 5.32 illustrate the conditional probability distribution of topics given each word for the Greedy baseball diamond experiment. For the majority of words(nine), the distribution is sharply peaked, with at least one topic receiving a dominant probability (more than 50%). This concentration simplifies the semantic interpretation process, since a clear correspondence can often be established between a word and its most strongly associated topic. In contrast, just one word have more diffused more diffuse distributions—where a word exhibits small(less than 50%), scattered probabilities across multiple topics. Notably, this proportion is approximately half of that observed in the Non-Greedy approach, which is consistent with the Greedy strategy's tendency to suppress boundary-case semantic variations. While this results in reduced semantic diversity, it yields easier to interperate decisive topic–word associations.

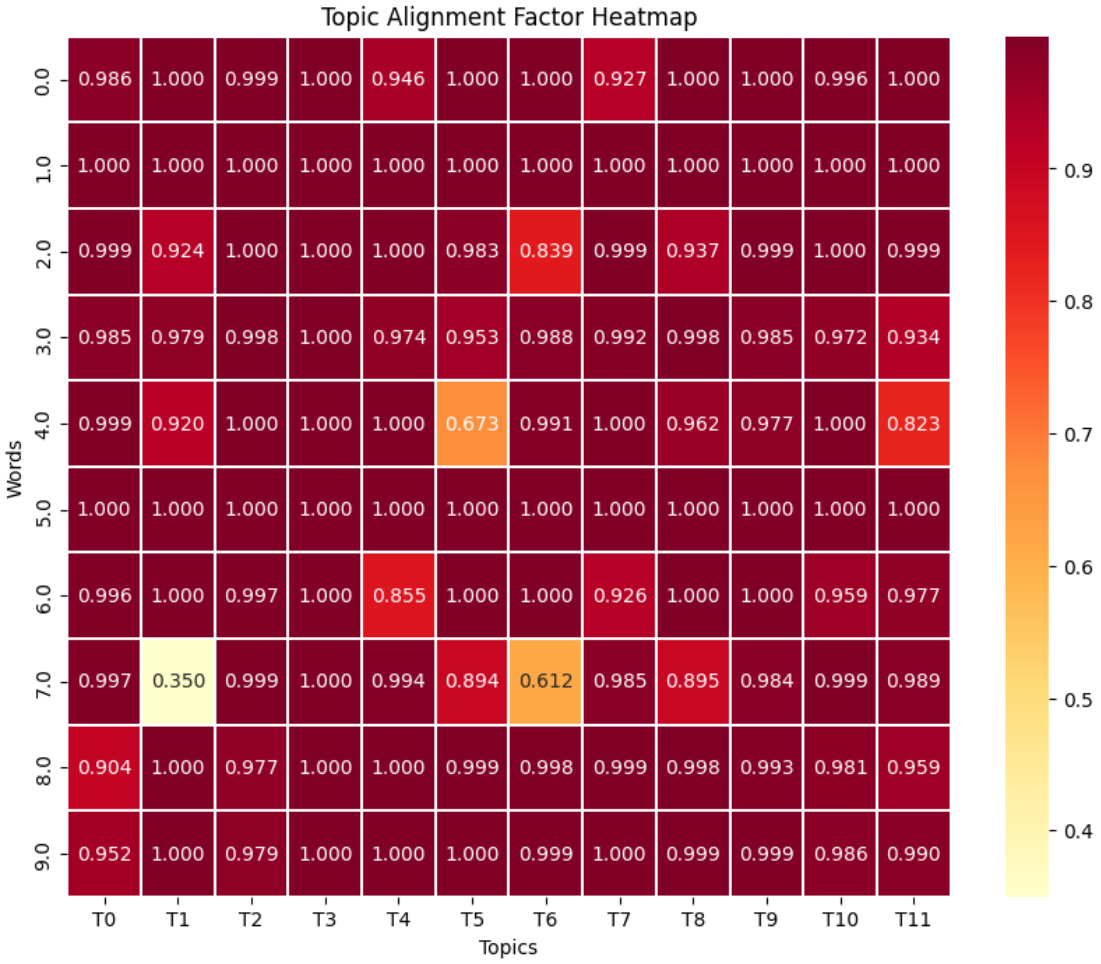

Figure 5.33: Baseball(Non-Greedy) TAF

The heatmap in Figure 5.33 displays the **Topic Alignment Factor (TAF)** values for the Non-Greedy experiment. This visualization allows evaluation of how effectively the probabilistic

assignment $P(T \mid W)$ corresponds to the actual topic allocations produced by LDA. For most words, their corresponding TAFs seems to be higher than 0.8 with just three exceptions. This suggests even though Non-Greedy experiments generate more difficult to interpret rules than the Greedy approach, it seems to extract better semantic meanings.

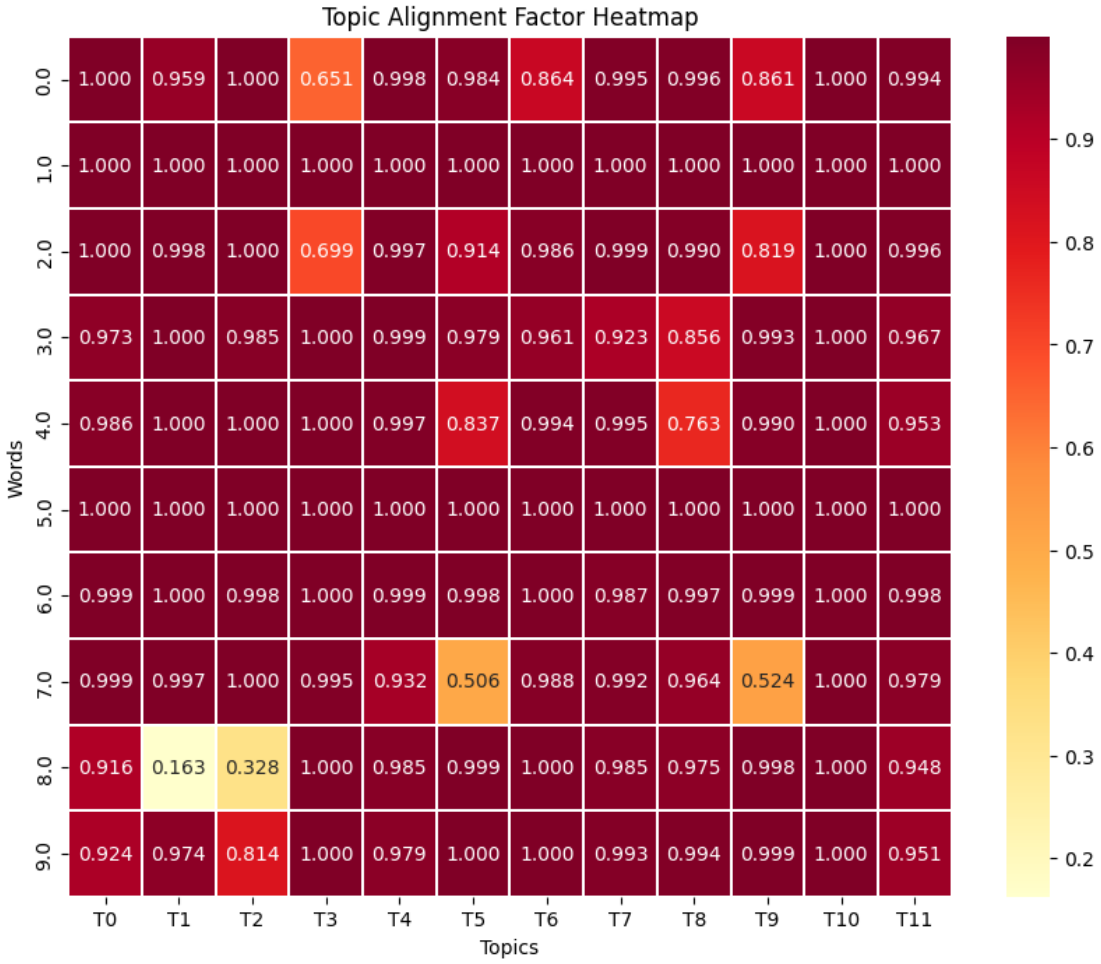

Figure 5.34: Baseball(Greedy) TAF

The heatmap in Figure 5.34 displays the **Topic Alignment Factor (TAF)** values for the Greedy experiment. This visualization allows evaluation of how effectively the probabilistic assignment $P(T \mid W)$ corresponds to the actual topic allocations produced by LDA. Overall, the TAF values are lower than those observed in the Non-Greedy experiment, indicating reduced confidence in the semantic associations inferred from this setting. While the Greedy approach yields decision rules that are more straightforward to interpret, this result suggests a loss of semantic richness, leading to weaker alignment between words and latent topics.

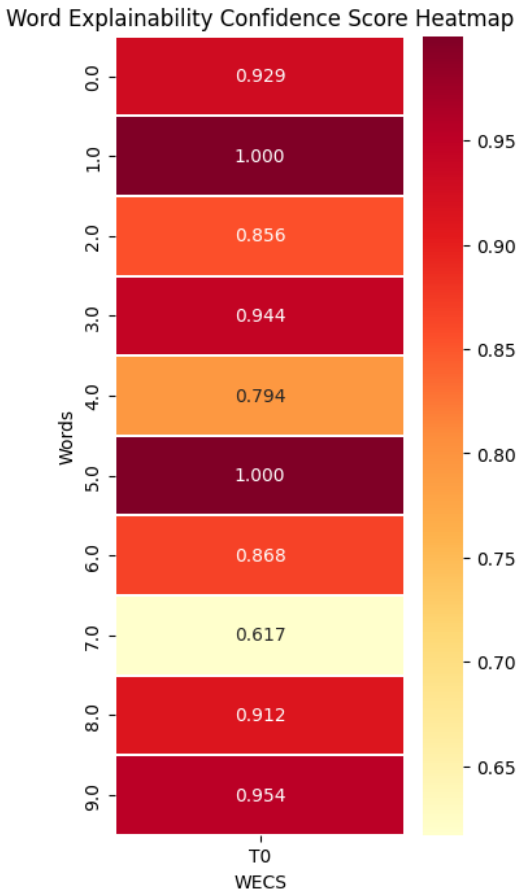

Figure 5.35: Baseball(Non-Greedy) WECS

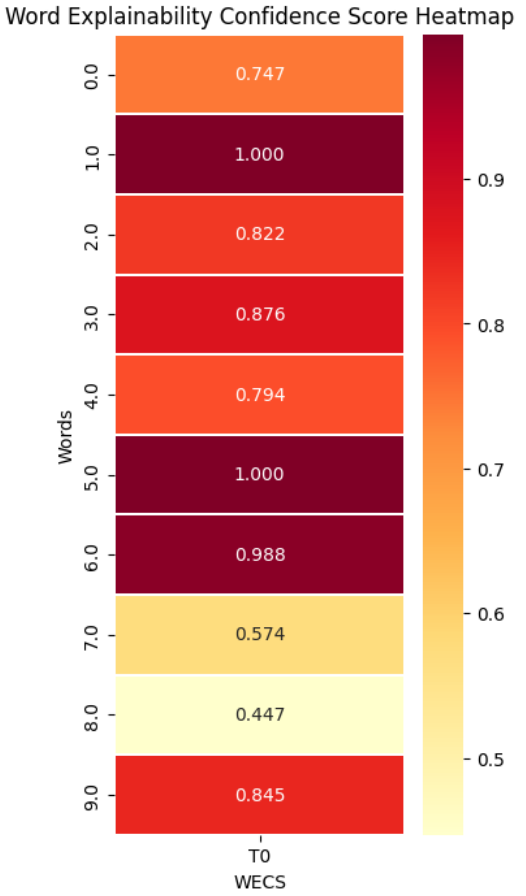

Figure 5.36: Baseball(Greedy) WECS

Figures 5.35 and 5.36 present the **Word Explainability Confidence Scores (WECS)** obtained for the Baseball experiment. A comparison of the two heatmaps indicates that, for most corresponding words, the Non-Greedy approach achieves higher WECS values than the Greedy variant. This suggests that the Non-Greedy version offers more reliable word–topic alignment and preserves semantic structure more effectively. These findings are consistent with the visual LDA topic interpretations, where the Non-Greedy model produces clearer and more coherent topic segmentation, while the Greedy approach shows a noticeable loss of semantic distinction. Together, the WECS patterns and LDA visual analysis demonstrate that the proposed scoring metric is able to capture and reflect the degradation of semantic meaning introduced by the Greedy splitting strategy.

Conclusion:

Overall, the results indicates that Non-Greedy and Greedy approach both are able to capture semantic rules of clusters. THe Non-Greedy approach shows significantly better capturing os

dematic meaning than both Non-Greedy and Greedy k-means.This could be a result of high overlap of sematic meaning of different classes like different colour grass and clays. WECS along with scores suggests that the LDA has more confidence in its word topic semantic meaning alignment for Non-Greedy than Greedy

5.3 Results and Conclusion

Overall, the results indicate that while both the Greedy and Non-Greedy approaches are capable of extracting meaningful, transferable emantic rules, their relative effectiveness depends strongly on the characteristics of the dataset and the requirements of the user. The cluster fidelity scores and tree structures suggest that the semantic meaning serived are similar across various folds of same datasets. LDA experiment shows that in datasets where the semantic separation between classes is visually distinct and structurally well defined—such as the Airplane subset of the UC Merced dataset—the Greedy strategy may offer practical advantages. Its single, strictly optimal path from root to leaf produces simpler and more compact decision trees, making subsequent semantic interpretation more direct and efficient.

However, in datasets where class boundaries are more subtle and semantic differences are harder to distinguish—such as in the Baseball Diamond subset, where multiple grass, clay, and surface types appear under varying lighting conditions—the Non-Greedy approach demonstrates clear benefits. By considering a buffer region and evaluating more globally optimal splits, it avoids premature decisions based solely on local thresholds and is better able to preserve subtle visual distinctions. This leads to higher semantic fidelity, stronger word–topic alignment, and more reliable explainability scores.

Thus, the Greedy approach is computationally simpler and may be preferred when class separability is high, while the Non-Greedy method provides superior semantic modelling in datasets characterized by fine-grained or overlapping visual classes.

The results also demonstrate the good performance of both X-kMeans approaches on unseen data, thus establishing the transferability of the semamtic meaning. LDA topics along with decision tree semantic rules provides domain expert with necessary tools to integrate domain knowledge into the learning process, while TAF and WECS scores quantify the confidence in the integration of decision tree semanmtic rules with LDA topic modeling.

6 Explainability and Semantic Analysis

This chapter demonstrates how semantic interpretations can be assigned to both the decision-tree "words" and the LDA-generated topics. The approach begins by analyzing the feature-based rules learned by the X-kMeans decision tree, where each leaf node corresponds to a "word" defined by a set of interpretable threshold-based feature splits. These rules provide a deterministic mapping between pixel properties and their final cluster assignment, allowing the characteristics of each word to be inferred directly from the structure of the tree.

For every word, the associated feature constraints (e.g., brightness ranges, texture descriptors, dominant edge orientation, etc) are interpreted using domain knowledge and visual inspection of representative image samples. This produces a semantic explanation describing what type of region or surface the word likely represents (e.g., dark shadowed areas, vegetation with strong red-edge response, high-contrast aircraft bodies, etc.).

These word-level interpretations are then connected to topics discovered by Latent Dirichlet Allocation (LDA). Since LDA provides a probability distribution $P(T \mid W)$ over topics for each word, the dominant topic for each word can be identified. In this analysis, semantic meaning for a topic is primarily inferred from the word with the highest $P(T \mid W)$, under the assumption that the most probable word contributes the strongest semantic signal to that topic. While this chapter focuses on the dominant topic per word for clarity of presentation, a more detailed analysis could consider the contribution of multiple topics for each word, weighted by their respective probabilities. Such an extension would reflect that a word may meaningfully participate in several topics, and that topics themselves may represent overlapping semantic concepts.

This structured pipeline—(1) decision-tree rule extraction, (2) word interpretation through quantitative constraints and ectracted rules, and (3) aggregation of word-level semantics to topic-level meaning through $P(T \mid W)$ and visual evidence provides a transparent mechanism for explaining the latent structure discovered in unsupervised clustering. It demonstrates how interpretable, rule-based explainability can bridge the gap between low-level feature thresholds and high-level semantic understanding in remote sensing imagery.

6.1 Airplane

This explaination experiment is done on the Non-Greedy experiment for UCmerced aircraft images

6.1.1 Word Topic Distribution

Figure 6.1: $P(T \mid W)$ for Non-Greedy Airplane Experiment

Topic 7 — Dark, Structured Surfaces with medium brightness Variants Feature–Word Rules

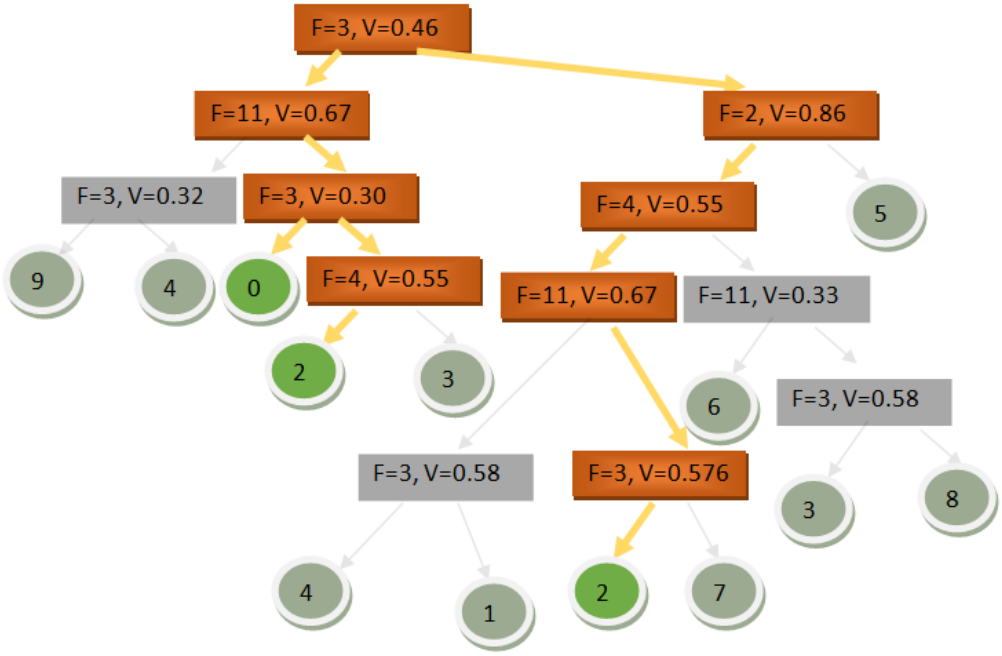

Figure 6.2: Paths to word 0 and word 2

Topic Assignment Word 0 receives highest probability from Topic 7 (48.5) Word 2 receives highest probability from Topic 7 (86.6).

Interpretation

Topic 7 primarily represents dark, structured surfaces with medium-brightness variants. Both Word 0 and Word 2 receive their highest probabilities from this topic, with Word 0 scoring 48.5 and Word 2 scoring 86.6. Word 0 follows the feature-rule sequence $F3 : 0{-}{-}0.46 \rightarrow F11 : 0.67{-}{-}1.0 \rightarrow F3 : 0{-}{-}0.3$. This combination highlights very low grayscale intensity, strong vertical or diagonal edge orientations, and a final restriction to extremely dark values. These characteristics indicate that Word 0 is associated with regions that are darker than average, such as shadows on aircraft wings or on tarmac surfaces. The strong orientation response from $F11$ suggests that these shadows are sharply angled, often corresponding to fuselage edges, angled wing tips, or pronounced Quadrant 4 shadow boundaries. Consequently, Word 0 identifies very dark, diagonally or vertically oriented shadowed regions adjacent to aircraft or other structures.

Word 2 also maps strongly to Topic 7 but includes two possible feature-rule paths. Path A corresponds to a progression from low brightness through strong diagonal or vertical orientation into a slightly brighter subrange of grayscale, concluding with lower texture values. This pattern reflects surfaces that remain dark overall but contain locally brighter shadow gradients, such as dark regions on concrete or shaded areas receiving partial reflected light. In contrast, Path B

moves through medium brightness, medium to high blue-channel responses, smooth surfaces, and vertically aligned edges before returning to lower brightness values. This path captures darker aircraft or runway regions influenced by variable lighting conditions,dark concrete or asphalt, particularly under diffuse illumination or overcast environments. In both paths, the unifying characteristic is the depiction of pronounced shadow regions or dark structural surfaces, which explains the strong association with Topic 7. Whereas Word 0 isolates the darkest and most sharply defined shadow edges, Word 2 represents a broader set of dark or shadow-affected areas with greater variability in texture and illumination.

Topic 8 — Mid to high bright Bright, Smooth Aircraft parts and runway areas with some possible texture Feature–Word Rules

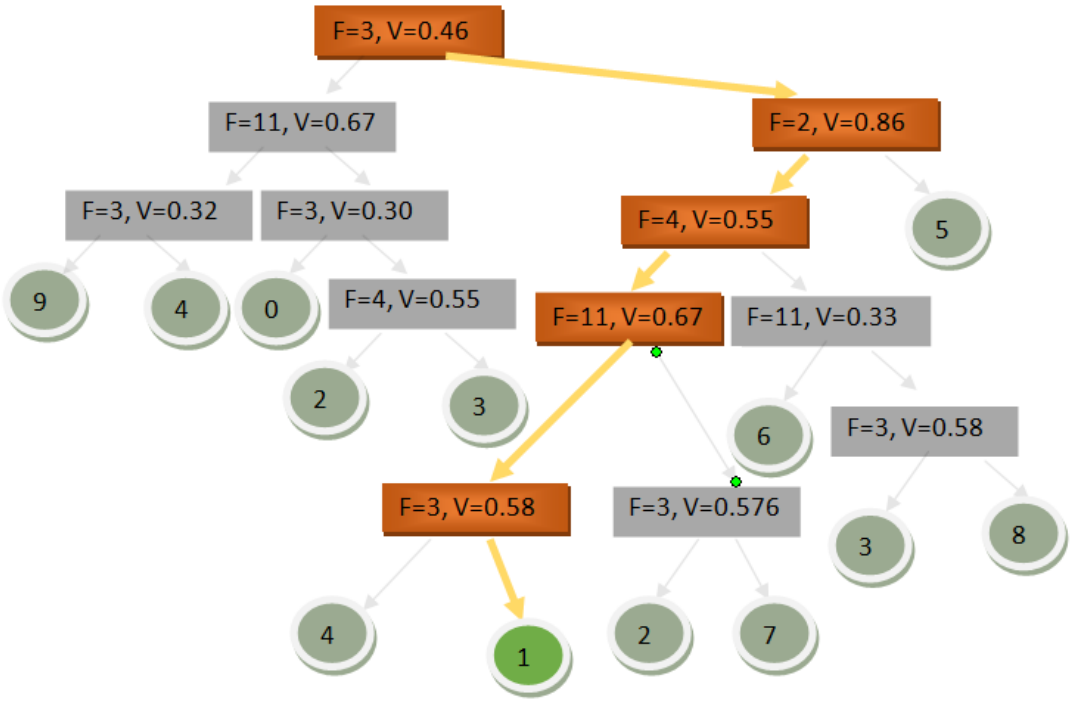

Figure 6.3: Path to Word 1

Topic Assignment

Word 1 receives highest probability from Topic 8 (95.2).

Interpretation

Topic 8 describes mid- to high-brightness, smooth aircraft surfaces and runway areas. Word 1, which shows the strongest association with this topic (95.2), follows the sequence $F3 : 0.46{-}{-}1.0 \rightarrow F2 : 0{-}{-}0.86 \rightarrow F4 : 0{-}{-}0.55 \rightarrow F11 : 0{-}{-}0.67 \rightarrow F3 : 0.58{-}{-}1.0$. These feature ranges indicate bright grayscale intensities, moderate blue-channel reflectance,

smooth textures, and horizontal or slightly slanted edge orientations. Such properties correspond to sunlit aircraft fuselage surfaces, bright runway markings, or smooth runway pavements. The low texture values further confirm their association with clean, painted, or metallic aircraft components.

Topic 10 — Mid Bright, Textured with Oblique Edges Feature–Word Rules

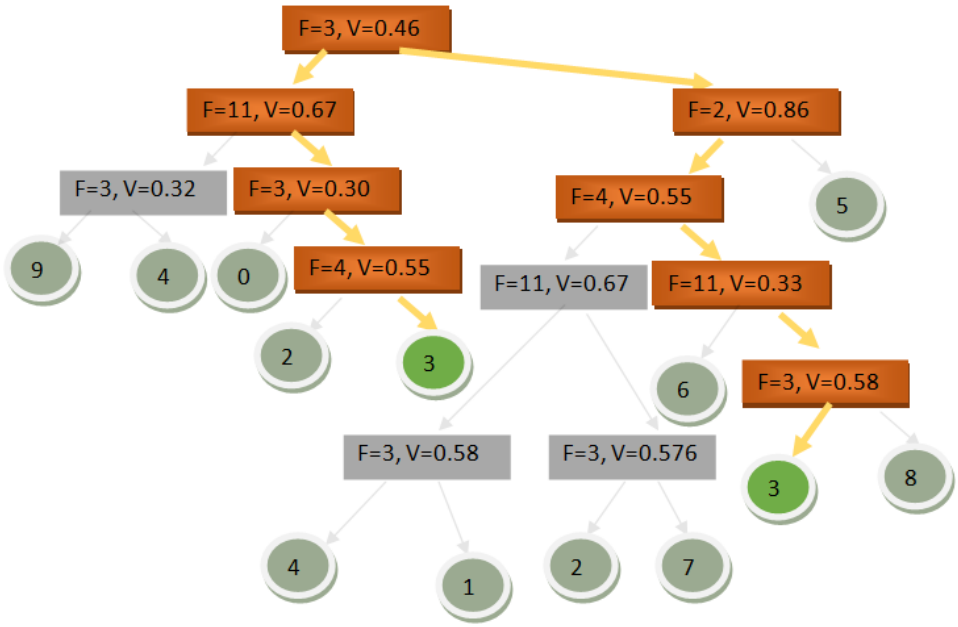

Figure 6.4: Paths to Word 3

Topic Assignment

Word 3 receives highest probability from Topic 10 (70.1).

Interpretation

Topic 10 captures mid brightness regions with noticeable texture and oblique edge orientations. Word 3, which most strongly represents this topic (70.1), exhibits two feature-rule paths reflecting both darker and brighter contexts. Path A transitions from low to mid grayscale values through strong diagonal orientation and then into higher brightness levels combined with high texture, whereas Path B stays in mid brightness range, strong blue-channel reflectance, and high texture, then moves through diagonal or vertical edges before settling into medium brightness. Despite these differences, both paths emphasize textured structures with oblique orientation typical of certain aircraft components, such as patterned wing surfaces or textured tarmac regions with strong directional edges. Its sparse appearance in the LDA results likely arises from the relatively low document probability associated with Topic 10.

Topic 6 — Mixed Shadow Medium-Texture Regions Feature–Word Rules

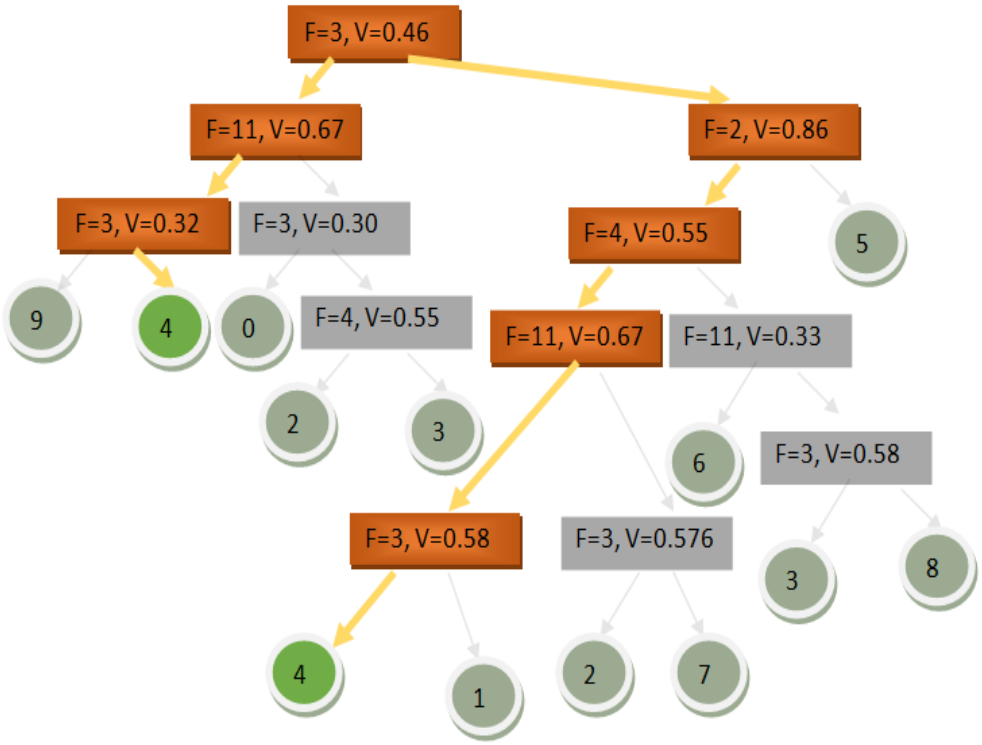

Figure 6.5: Path to Word 4

Topic Assignment

Word 4 receives highest probability from Topic 6 (58.6).

Interpretation

Topic 6 represents mixed-shadow, medium-texture regions commonly found on runways under variable lighting conditions. Word 4, which shows the strongest alignment with this topic (58.6), also follows two feature-rule paths. Path A involves low brightness values, horizontally to moderately slanted edges, and transitions into mid-range brightness, suggesting pavement surfaces with faint or diffuse shadows. Path B involves medium brightness, strong blue-channel reflectance, smooth surfaces, and variable edge directions, characteristics consistent with runway pavement under overcast or lightly shadowed conditions. Overall, Topic 6 captures runway regions exhibiting moderate brightness and subtle shadowing influenced by inconsistent illumination.

Topic 1 — Bright High-Blue, Smooth Highlights Feature–Word Rules

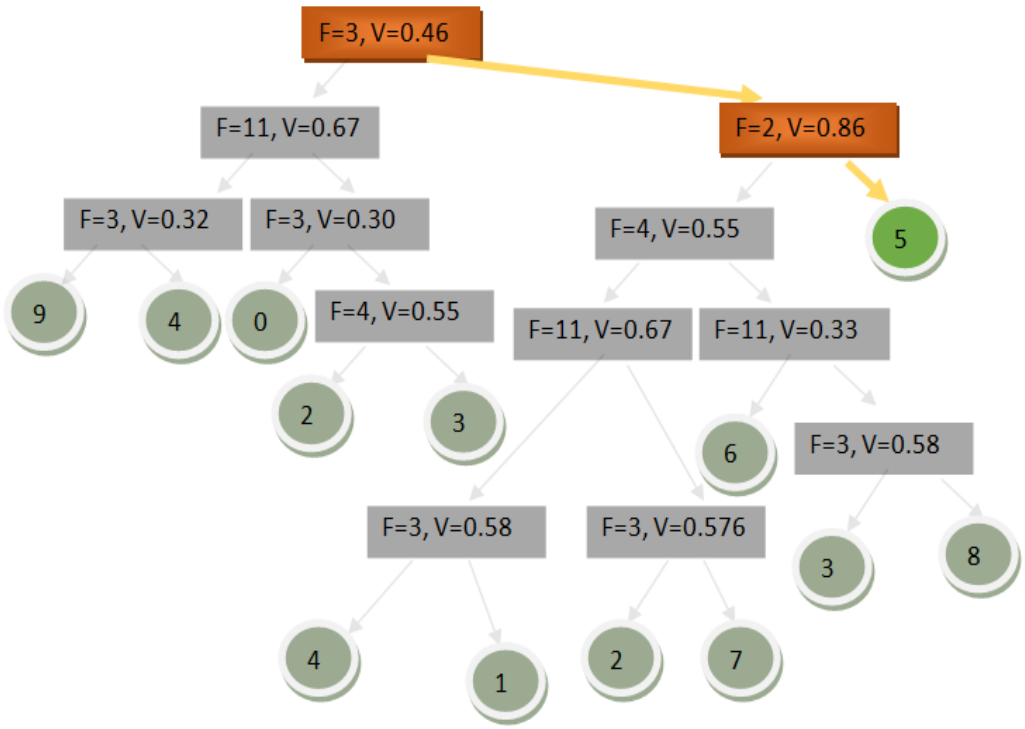

Figure 6.6: Path to Word 5

Topic Assignment

Word 5 receives highest probability from Topic 1 (34.6).

Interpretation

Topic 1 corresponds to bright, high-blue,areas although it appears at lower probability. Word 5, the defining word for this topic (34.6), is characterized by high grayscale brightness combined with very strong blue-channel intensity. These properties often occur in highly reflective or metallic surfaces on aircraft bodies, such as fuselage highlights, cockpit reflections, or tail sections reflecting the sky. Topic 1 therefore may isolates bright, sky-reflective. Since this word just takes 2 features, its semantic meaning can overlap many classes and is indifferent to many other features which may suggest why it has such a scattered probability

Topic 9 — Mid to high Bright Textured Horizontal Regions Feature–Word Rules

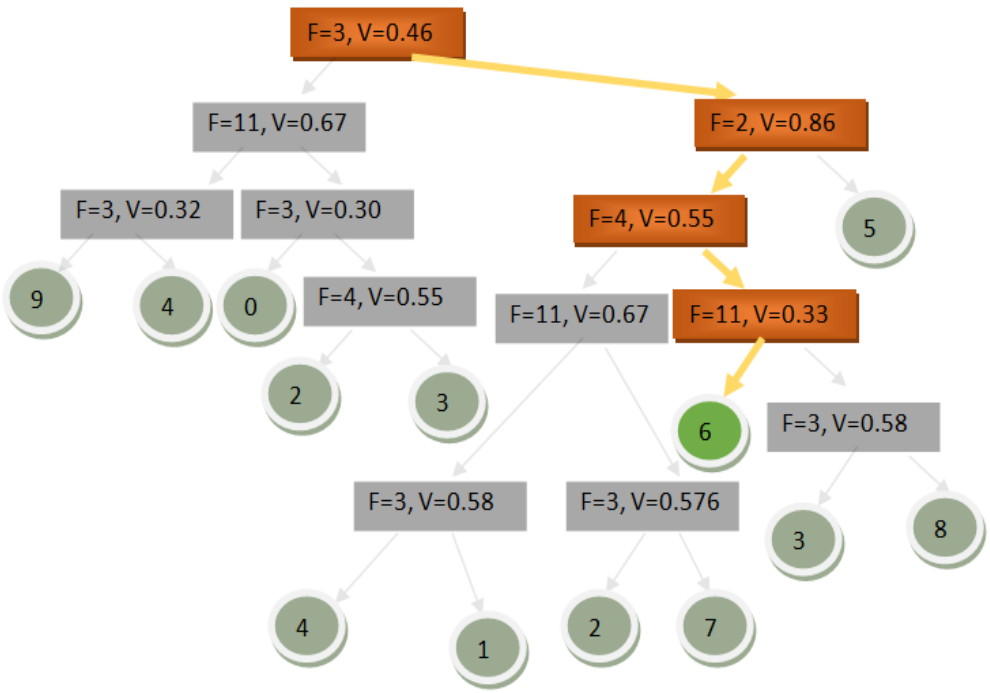

Figure 6.7: Path to Word 6

Topic Assignment

Word 6 receives highest probability from Topic 9 (94.99).

Interpretation

Topic 9 describes medium- to high-brightness textured regions with horizontal structural alignment. Word 6, which strongly defines this topic with a probability of 94.99, features mid to bright intensities, high vareity(covering 82% of the total range) blue-channel reflectance, high texture values, and horizontal edges. These characteristics indicate textured runway surfaces or other horizontally aligned ground structures exposed to varied lightings.

Topic 3 — Mid to high Bright Diagonal Smooth Surfaces Feature–Word Rules

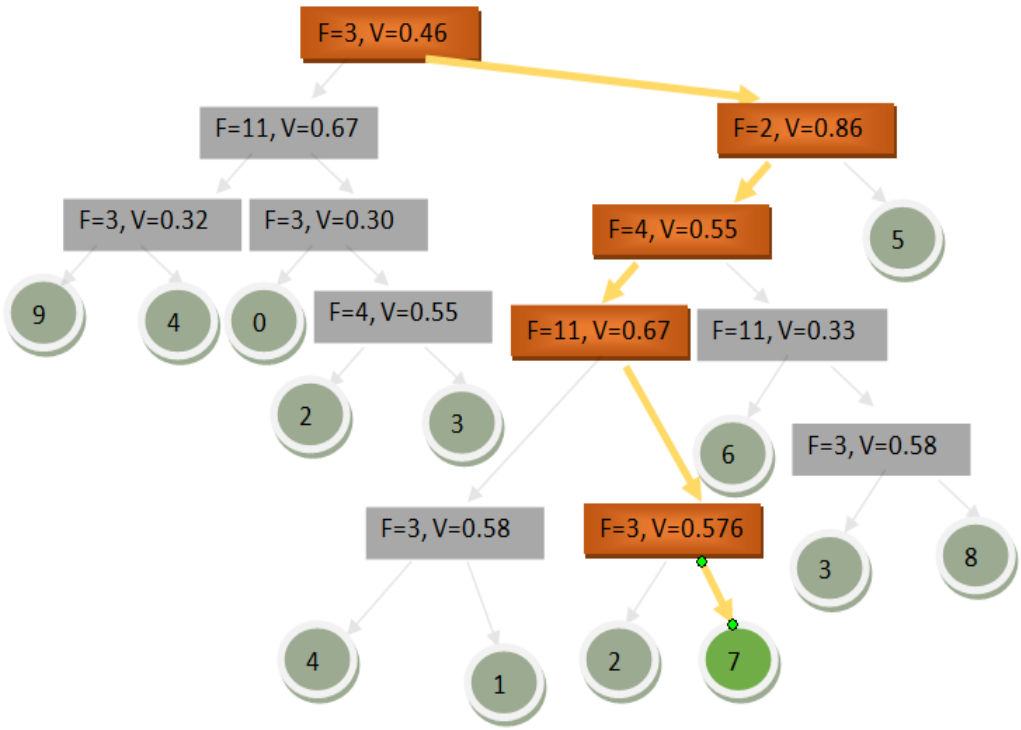

Figure 6.8: Path to word 7

Topic Assignment

Word 7 receives highest probability from Topic 3 (39.8).

Interpretation

Topic 3 captures bright, diagonally aligned smooth surfaces. Word 7 aligns most strongly with this topic (39.8), but also aligns close enough with topic 2(33).It maybe defined by mid to bright intensities, low texture, strong diagonal orientation, and smooth surfaces. These features make it representative of smooth fuselage or wing regions that appear diagonally oriented in the imagery.

Topic 4 — Moderate-High Brightness, Textured Angled Surfaces Feature–Word Rules

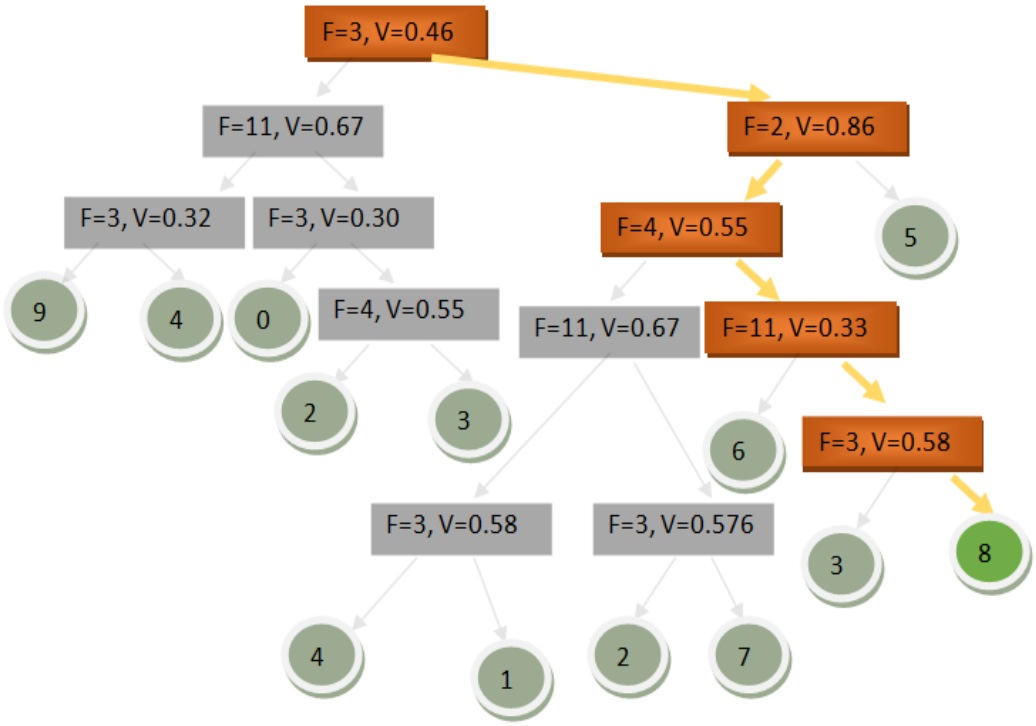

Figure 6.9: Path to word 8

Topic Assignment

Word 8 receives highest probability from Topic 4 (34.04).

Interpretation

Topic 4 represents moderately bright, textured surfaces with angled edge orientations. Word 8, although associated with this topic at a lower probability (34.04), corresponds to regions with moderate brightness, high texture, and oblique orientations. This LDA topic has sparse presence in the LDA output.

Topic 0 — Dark Horizontal Shadows Feature–Word Rules

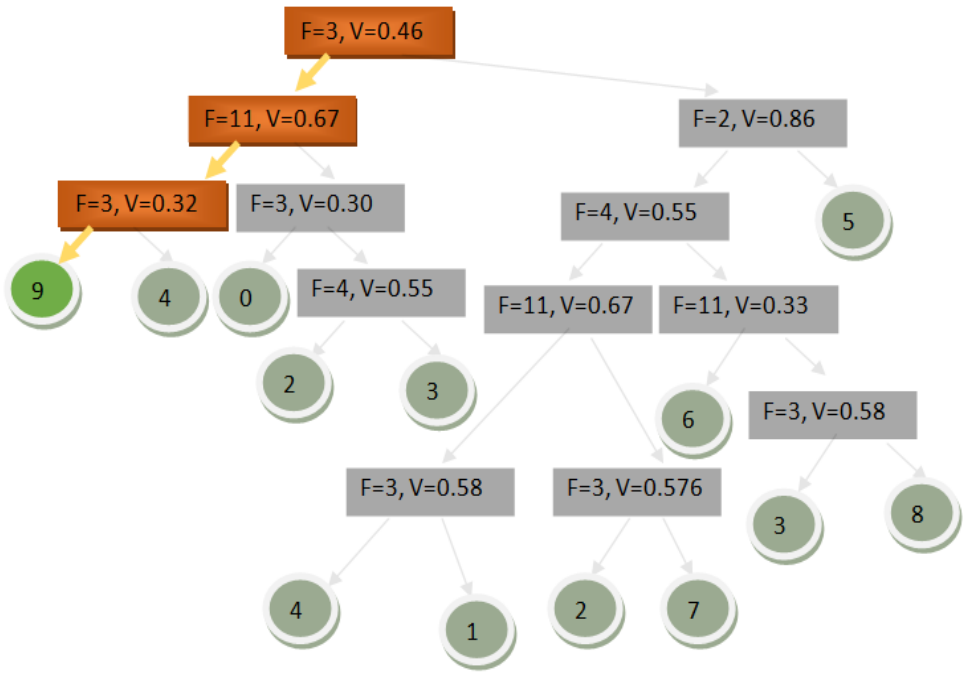

Figure 6.10: Path to Word 0

Topic Assignment

Word 9 receives highest probability from Topic 0 (78.6).

Interpretation

Topic 0 captures dark horizontal shadows. Word 9, which receives a high probability of 78.6 for this topic, is characterized by very low grayscale intensity combined with horizontal edges and a final narrowing into extremely dark values. These properties correspond to strong, elongated shadows on runways or beneath aircraft fuselages. Unlike Topic 7, which emphasizes diagonal or vertical shadow structures, Topic 0 isolates horizontally aligned shadow regions, highlighting the importance of edge orientation despite the overall similarity in depicting shadow-dominated areas.

7 Conclusion and Future Work

7.1 Conclusion

This thesis investigated how the implicit decision boundaries and latent semantic structure of k-means clustering in remote sensing data can be made explicit, interpretable, and quantitatively reliable.It also assesses the tranferability of these rules. Addressing a key gap in unsupervised explainability, the work proposed a hybrid framework that combines explainable decision-tree surrogates with probabilistic topic modeling to bridge the divide between numerical clustering behavior and semantic understanding.

By approximating k-means cluster assignments through X-kMeans decision trees, the framework translated opaque distance-based partitions into human-readable feature-threshold rules. These rules enable explicit inspection of cluster boundaries while preserving a high degree of fidelity to the original clustering. To connect structural explanations with higher-level semantics, Latent Dirichlet Allocation was applied by treating clusters as visual words and images as visual documents, allowing latent topics to emerge that reflect recurring semantic patterns in the data.

A key contribution of this work is the introduction of the Cluster Fidelity, Topic Alignment Factor (TAF) and the Word Explainability Confidence Score (WECS). Unlike traditional clustering or topic-evaluation metrics, these measures are specifically designed to assess the reliability and consistency of semantic explanations at the word–topic level, along with the faithfulness of the explainable k-Means variants to k-Means. By quantifying the alignment between probabilistic topic assignments and observed cluster-derived word distributions, TAF and WECS enable a confidence-aware evaluation of explainability and provide quantitative support for domain experts during semantic interpretation and topic validation.

Experimental evaluation on subsets of the UCMerced remote sensing dataset demonstrated that the proposed framework can generate explanations that are both interpretable and semantically meaningful without substantially compromising clustering fidelity. While different tree-construction strategies exhibit trade-offs between structural simplicity and semantic richness, the results consistently show that meaningful semantic structure can be recovered from unsupervised clustering and assessed quantitatively. The LDA-based analysis further indicates that when semantic overlap between classes is limited, both Greedy and Non-Greedy strategies perform comparably, whereas in scenarios with higher semantic ambiguity, Non-Greedy strategy better capture subtle distinctions. Additionally, evaluation on unseen data suggests that the learned decision rules and semantic interpretations exhibit a degree of transferability beyond the training set, indicating that the explanations reflect underlying data structure rather than dataset-specific artifacts.

Overall, this thesis demonstrates that unsupervised clustering in remote sensing can be made not only explainable, but also trustworthy, by jointly considering structural transparency,

semantic coherence, and confidence in explanation quality. The proposed framework establishes a foundation for integrating domain knowledge into unsupervised explainability pipelines and contributes toward more transparent, interpretable, and reliable use of clustering methods in Earth observation and related domains.

7.2 Future

Several directions exist for extending and strengthening the proposed explainable clustering framework.

First, the framework can be evaluated on larger and more diverse remote sensing datasets, including multispectral and hyperspectral imagery such as Sentinel-2 data. This would allow the analysis of how the proposed explainability measures behave under higher spectral complexity and varying spatial resolutions.

Second, although this work focuses on image-based data, the methodology is not inherently image-specific. Future studies may investigate its applicability to non-image datasets to assess the generality of the decision-tree–based explainability and the robustness of the proposed confidence measures.

Third, the framework will be integrated with ULearn [33], enabling a semi-supervised pipeline. In such a setting, domain experts could inspect decision rules, topic assignments, and confidence scores through a user interface and iteratively provide feedback to refine topic semantics and improve explainability.

Fourth, while this thesis primarily considers k-means and its explainable variants, future work may explore the use of alternative clustering techniques to study how different clustering paradigms influence rule extraction and semantic alignment.

Finally, Possibility can be explored to extend the framework to analyze the latent spaces of supervised learning models. Through this approach, it may be possible to gain insights into the internal feature organization of supervised models and assess the semantic consistency of learned representations.

Bibliography

[1] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979, doi: 10.2307/2346830. [Online]. Available: http://www.jstor.org/stable/2346830

[2] S. Dasgupta, N. Frost, M. Moshkovitz, and C. Rashtchian, "Explainable k-means and k-medians clustering," 2020, doi: 10.48550/arXiv.2002.12538. [Online]. Available: https://arxiv.org/abs/2002.12538

[3] K. Makarychev and L. Shan, "Explainable k-means: don't be greedy, plant bigger trees!" in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 1629–1642. [Online]. Available: https://doi.org/10.1145/3519935.3520056

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*, 1984, doi: 10.1201/9781315139470.

[5] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," vol. 3, 2003, pp. 993–1022.

[6] J. B. McQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.

[7] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, p. 241–254, 1967, doi: 10.1007/BF02289588.

[8] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, p. 264–323, Sep. 1999. [Online]. Available: https://doi.org/10.1145/331499.331504

[9] D. Deng, "Dbscan clustering algorithm based on density," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 2020, doi: 10.1109/IFEEA51475.2020.00199.

[10] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*, 1981, doi: 10.1007/978-1-4757-0450-1.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[12] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[13] G. Salton and M. McGill, "Introduction to modern information retrieval." New York, NY, USA: McGraw-Hill, 1983.

[14] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002, doi: 10.1198/016214502760047131.

[15] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 05 2019, doi: 10.1038/s42256-019-0048-x.

[16] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, doi: 10.48550/arXiv.1702.08608. [Online]. Available: https://arxiv.org/abs/1702.08608

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[18] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, doi: 10.48550/arXiv.1705.07874.

[19] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, p. 22071–22080, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1073/pnas.1900654116

[20] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[21] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020, doi: 10.48550/arXiv.1802.03426. [Online]. Available: https://arxiv.org/abs/1802.03426

[22] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 1999, doi: 10.1023/A:1007665907178.

[23] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl_1, pp. 5228–5235, 2004, doi: 10.1073/pnas.0307752101.

[24] C. Karmakar, C. O. Dumitru, G. Schwarz, and M. Datcu, "Feature-free explainable data mining in sar images using latent dirichlet allocation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 676–689, 2021, doi: 10.1109/JSTARS.2020.3039012.

[25] M. Alvarez-Garcia, M. Arenas-Parra, and R. Ibar-Alonso, "Uncovering student profiles. an explainable cluster analysis approach to pisa 2022," *Computers Education*, vol. 223, p. 105166, 2024, doi: 10.1016/j.compedu.2024.105166. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131524001805

[26] C. A. Ellis, M. S. E. Sendi, E. P. T. Geenjaar, S. M. Plis, R. L. Miller, and V. D. Calhoun, "Algorithm-agnostic explainability for unsupervised clustering," 2021, doi: 10.48550/arXiv.2105.08053. [Online]. Available: https://arxiv.org/abs/2105.08053

[27] J. Crabbé and M. van der Schaar, "Label-free explainability for unsupervised models," 2022, doi: 10.48550/ARXIV.2203.01928. [Online]. Available: https://arxiv.org/abs/2203.01928

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[29] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[30] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models," 2019, doi: 10.48550/arXiv.1908.01224. [Online]. Available: https://arxiv.org/abs/1908.01224

[31] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 111–119, doi: 10.1109/CVPRW50498.2020.00020.

[32] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7, doi: 10.1109/IJCNN48605.2020.9206626.

[33] S. Goyal, C. Karmakar, A. Camero, C. O. Dumitru, and M. Datcu, "Ulearn: An explainable uncertainty-aware machine learning tool for unsupervised classification," in *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, 2025, pp. 2748–2751, doi: 10.1109/IGARSS55030.2025.11242525.