الجمهورية الجزائرية الديمقراطية الشعبية
The People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة محمد بوضياف بالمسيلة
University Mohamed Boudiaf of M'sila

كلية الرياضيات والإعلام الآلي
Faculty of Mathematics and Informatics

قسم الإعلام الآلي
Department of Computer Science

**Domain:** Mathematics and Computer Science

Dissertation Presented to Fulfill the Partial Requirement
for Master's Degree in Computer Science
**Specialty:** **BUSINESS INTELLIGENCE AND OPTIMISSATION**

**Prepared By:** AMROU Marwa, LADJLAT Refayda

**Supervised By:**

MEHENNI TAHAR

ENTITLED

# UNSUPERVISED LEARNING FOR THE IDENTIFICATION OF HOMOGENEOUS FOREST LANDSCAPES

**Jury Members**

| | |
|---|---|
| Abdallah Tharafi | President |
| Tahar Mehenni | Supervisor |
| Abdelghani Boudaa | Examiner |

**Academic Year 2023/2024**

# Dedication

**(وآخِرُ دَعوَاهُمْ أنِ الحَمدُ لله ربِ العالمينَ)**

***<u>مروة</u>***

**الحمد للَّه وكفى والصَلاة على الحبيب المُصطفى وأهله ومن أوفى أما بعد**

**الحمد للَّه الذي وفقنا لِتثمين هذه الخطوة في مسيرتنا الدراسية بمذكرتنا هاته ثمرة الجهد والنجاح بفضله تعالى**

**مهداة:**

**إلى من لا يضاهيهما أحد في الكون، إلى من أمرنا الله ببرّهما، إلى من بذلا الكثير، وقدّما ما لا يمكن أن يردّ، إليكما تلك الكلمات أمي و أبي الغاليان قرة عيناي**

**إلى من أعتمد عليهم في كل موقف صغير أم كبير إخوتي سمير عيسى وعبد القدير**

**إلى مؤنساتي غالياتي وحبيباتي أخواتي ابتسام ياسمين صفاء**

**إلى الطفولة التي ملأت عالم عائلتي إلى أول حفيد بالعائلة بسمتي وبهجتي <u>وسيم</u> وباقي الأحفاد سراج غيث و سند**

**إلى اصدقائي، الذين جمعتني بهم هذه الكُليّة ولكل من كان له فضل في هذا النجاح شكرا ...**

**<u>رفيدة</u>**

**من قال أنا لها ... نالها**

**و أنا لها إن أبت و رغما عنها أتيت بها**

**الحمد لله شكرا حباً و امتنانا على البدء و حسن الختام**

**إلى السراج الذي لا ينطفئ نوره بقلبي أبدا (والدي العزيز) ..**

**إلى شمعتي في الليالي المظلمة والتي سهلت لي شدائد بدعائها إلى الانسانة العظيمة أمي .**

**إلى ضلعي الثابت و أمان أيامي إلى من شددت عضدي بها أختي الحبيبة .**

**إلى خير أيامي و صفوتها إلى قرة عيني اخوتي .**

**لكل من كان عوننا و سنداً في هذا الطريق ، للأصدقاء الأوفياء إلى أصحاب الشدائد شكرا.**

**أهديكم ثمرة نجاحي الذي أتممته بفضل الله عزوجل فالحمد لله على ما وهبني، اللهم اجعلني مباركةً أينما كُنت..**

# Acknowledgement

*First and foremost, heartfelt gratitude and praises go to the Almighty Allah who guided us through and through.*

*We would like to thank our supervisor teacher, Dr. Mehenni Tahar , your mentorship, feedback, and dedication to our growth have been instrumental. Thank you for steering me in the right direction.*

*We would also like to thank all the Jury Members, who have agreed to review this work.*

*To everyone who contributed, whether directly or indirectly, your assistance mattered. From late-night brainstorming sessions to technical troubleshooting, your collective efforts made this journey possible.*

*Thanks*

# List of Contents

**General Introduction** ..... 1
CHAPTER 1 ..... 3
**FOREST LANDSCAPE** ..... 3
**1. Definition** ..... 4
**2. Types of Forest Landscapes** ..... 4
**1.2 Boreal forests** ..... 4
**2.2 Temperate forests** ..... 5
**2.3 Tropical forests** ..... 5
**2.4 Montane Forests** ..... 5
**3. Parameters of Forest Landscape** ..... 6
**4. Advantages of the Forest Landscape** ..... 7
**5. Challenges facing the forest landscape** ..... 8
**6. Solutions to address challenges facing the forest landscape** ..... 8
CHAPTER 2 ..... 10
**UNSUPERVISED MACHINE LEARNING TECHNIQUES** ..... 10
**1.Definition** ..... 11
**2.Models of Unsupervised Classification (Clustering)** ..... 11
**2.1 k-means Clustering Algorithm** ..... 11
**2.1.1 Definition** ..... 11
**2.1.2 Advantages and disadvantages of k-means clustering algorithm** ..... 12
**2.1.3 Types of k-Means Clustering** ..... 13
**2.1.4 Principles of k-Means Clustering** ..... 14
**2.2 Hierarchical clustering algorithm** ..... 15
**2.2.1 Definition** ..... 15
**2.2.2 Types of Hierarchical clustering algorithm** ..... 15
**2.2.3 Advantages and disadvantages of hierarchical clustering algorithms** ..... 16
**2.3 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** ..... 17
**2.3.1 Definition** ..... 17
**2.3.2 Parameters Required for DBSCAN Algorithm** ..... 17
**2.3.3 Advantages and disadvantages of DBSCAN algorithm** ..... 18
**2.4 OPTICS clustering algorithm** ..... 18
**2.4.1 Definition** ..... 18
**2.4.2 Main steps of OPTICS algorithm** ..... 19
**2.4.3 Advantages and disadvantages of OPTICS algorithm** ..... 19
**2.5 MEAN SHIFT clustering algorithm** ..... 20

**2.5.1 Definition** ..... 20
**2.5.2 Method of Mean Shift** ..... 20
**2.5.3 Advantages and Disadvantages of Mean-Shift Clustering** ..... 22
**2.6 Affinity propagation clustering algorithm** ..... 22
**2.6.1 Definition** ..... 22
**2.6.2 Key-steps of Affinity propagation** ..... 22
**2.6.3 Advantages and Disadvantages of Affinity Propagation Clustering** ..... 23
**3. Uses of Unsupervised classification (Clustering)** ..... 24
**4.Importance of Unsupervised classification (Clustering)** ..... 24
**5.The Most Common Challenges In Clustering Algorithms:** ..... 25
CHAPTER 3 ..... 26
**1. Presentation of study area** ..... 27
**2. Description of the dataset** ..... 28
**2.1 Data collect** ..... 28
**2.2. Description of plants and trees in Djebel Messaad Forest** ..... 28
**2.3. Description of flower species in Djebel Messaad Forest** ..... 30
**2.4. Description of the dataset attributes** ..... 32
**2.5. Preprocessing** ..... 34
**3.Implementation** ..... 34
**3.1 Programming language** ..... 34
**3.2 Development environment** ..... 35
**3.2.1 Anaconda** ..... 35
**3.2.2 TensorFlow** ..... 35
**3.2.3 Sklearn** ..... 35
**3.2.4 Pandas** ..... 35
**4. Application of K-means Clustering Algorithm on the dataset of plants** ..... 36
**5. Application of K-means Clustering Algorithm on the dataset of flowers** ..... 37
**6. Application of Hierarchical Clustering Algorithm on the dataset of plants** ..... 40
**7. Application of Hierarchical Clustering Algorithm on the dataset of flowers** ..... 41
**8. Discussion** ..... 40
**General Conclusion** ..... 44
**REFERENCES** ..... 45

# List of Figures

**Fig. 1.1** Boreal Forests 4

**Fig. 1.2:** Temperate Forests 5

**Fig. 1.3:** Tropical Forests 5

**Fig. 1.4:** Montane Forests 6

**Fig. 2.1:** Understanding K-Means Clustering 12

**Fig. 2.2:** Agglomerative Hierarchical Clustering 15

**Fig. 2.3:** Divisive Hierarchical Clustering 16

**Fig. 2.4:** Components of DBSCAN 18

**Fig. 2.5:** OPTICS Clustering 19

**Fig. 2.6:** Mean Shift Clustering 21

**Fi. 2.7:** Affinity propagation clustering of mall customers. 23

**Fig. 3.1:** Djebel Messaad Forest Map 27

**Fig. 3.2:** A sample of the dataset of plants 33

**Fig. 3.3:** A sample of the dataset of Flowers 34

**Fig. 3.4**: K-means code in python on the plants dataset. 36

**Fig. 3.5:** Clusters of plants using K-means 36

**Fig. 3.6**: K-means code on the flowers dataset. 38

**Fig. 3.7:** Clusters of the flowers using K-means 38

**Fig. 3.8**: Hierarchical code on the plants dataset 40

**Fig. 3.9:** Clusters of the plants using Hierarchical 40

**Fig. 3.8**: Hierarchical code on the flowers dataset 41

**Fig. 3.9:** Clusters of the flowers using Hierarchical 41

# List of Tables

**Table 3.1:** Summary of the plants and trees in the study area 29

**Table 3.2:** Summary of the flowers located in the study area 32

**Table 3.3:** Detailed clusters of plants using K-means 37

**Table 3.4**: Detailed clusters of flowers using K-means 39

# General Introduction

In the late 1990s and early 2000s the environmental conservation community began scaling up its interventions beyond protected areas to larger areas such as hotspots, ecoregions and landscapes. [1]

Global forests are under immense pressure from a suite of human activities, such as agricultural expansion and natural resource exploitation, in addition to global environmental change. Large-scale forest restoration is essential to ensure the continued flow of vital, forest related ecosystem services, including carbon sequestration, biodiversity conservation, and livelihood contributions. [2]

The continuous advancement in artificial intelligence (AI) technologies is witnessing an expanding array of applications across various domains. This evolution enhances its role in improving daily life and propelling technological progress forward. Additionally, machine learning plays a vital role in creating programs that consume raw data, build sophisticated models from that data, and make accurate predictions based on new data. Consequently, in this dissertation, we propose an approach based on unsupervised machine learning techniques to identify homogeneous forest landscapes. Our model is applied using data collected from From the Jabal Massed forest in M'sila.

The structure of this dissertation is as follows:

The first chapter presents forest landscapes, defining various types such as boreal, temperate, tropical, and montane forests. It discusses parameters and advantages of forest landscapes, as well as ecological services they provide and challenges they face, along with potential solutions.

The second chapter focuses on unsupervised modeling, presenting definitions and various clustering models including: k-means, Hierarchical, DBSCAN, OPTICS, Mean Shift, and Affinity Propagation clustering algorithms. Each algorithm principles, advantages, and disadvantages are discussed, along with their uses and importance. Common challenges encountered in clustering algorithms are also addressed.

The third chapter implements selected models for analyzing forest landscapes using real datasets collected from various sources. It describes the datasets, implementation details including programming languages and tools and the utilization of the proposed clustering algorithms.

Overall, our research aims to highlight the potential of machine learning algorithms in identification and clustering of different forest environments in a specific region. By leveraging these techniques, we can understand environmental diversity in order to supporting decision-making and developing restoration plans.

# CHAPTER 1

# FOREST LANDSCAPE

## Introduction

A forest is a large area of wooded land, characterized by a high density of trees and often other plants such as shrubs, bushes and grasses. Forests can vary in size, floristic composition and structure depending on factors such as climate, relief, soil type and human activity. They play a vital role in the ecological balance of the planet by providing oxygen, storing carbon, harboring rich biodiversity and regulating the water cycle and climate. Forests are also places of recreation, culture and spirituality for many people around the world.

## 1. Definition:

We define an **Intact Forest Landscape (IFL)** as an unbroken expanse of natural ecosystems within the zone of current forest extent, showing no signs of significant human activity and large enough that all native biodiversity, including viable populations of wide-ranging species, could be maintained. Although all IFL are within the forest zone, some may contain extensive naturally tree-less areas, including grasslands, wetlands, lakes, alpine areas, and ice. This definition builds on the definition of Frontier Forest that was developed by World Resources Report (WRI). [3]

## 2. Types of Forest Landscapes

### 1.2 Boreal forests

Also known as taiga or snow forests, Boreal forests are a type of forest ecosystem that spans the high-latitude regions of the Northern Hemisphere. They are characterized by a cold climate, long and harsh winters, short summers, and a predominance of coniferous trees such as spruce, fir, and pine (Fig. 1.1). Boreal forests are often found in areas with low temperatures and short growing seasons, resulting in slow tree growth and the presence of permafrost in some regions. [4]

**Fig. 1.1** Boreal Forests

## 2.2 Temperate forests

Temperate forests are a type of forest ecosystem that occurs in regions with moderate climate and distinct seasons, typically between the Polar Regions and the tropics. They are characterized by a mix of deciduous and coniferous trees and experience seasonal variations in temperature, precipitation, and daylight hours (Fig. 1.2). Temperate forests are known for their diverse plant and animal species, including mammals, birds, reptiles, and amphibians. [5]

**Fig. 1.2:** Temperate Forests

## 2.3 Tropical forests

Tropical forests are a type of forest ecosystem found in the Earth's tropical regions, typically between the Tropic of Cancer and the Tropic of Capricorn. They are characterized by high levels of biodiversity, dense vegetation, and a warm and humid climate throughout the year (Fig. 1.3). Tropical forests are home to a wide variety of plant and animal species, many of which are unique to these regions. [6]

**Fig. 1.3:** Tropical Forests

## 2.4 Montane Forests

Montane forests, also known as mountain forests, are a type of forest ecosystem that occurs in mountainous regions at high elevations. They are characterized by cooler temperatures, higher levels of precipitation, and distinct vegetation zones determined by elevation (Fig. 1.4).

Montane forests often exhibit a gradient of vegetation types, ranging from coniferous forests at lower elevations to subalpine forests and alpine meadows at higher elevations. [7]

**Fig. 1.4:** Montane Forests

## 3. Parameters of Forest Landscape

Forest landscapes can be characterized by various parameters that describe their spatial and ecological features. Some key parameters include:

- **Biological Diversity :** Refers to the diversity of living organisms in the forest, including plants, animals, and microorganisms. Biological diversity is an important indicator of the ecosystem's health in the forest.

- **Land Size and Geographic Configuration :** Includes the size of the forest, its terrain, and geographic distribution. Large forests with high biological diversity tend to be more stable and provide more natural resources.

- **Tree Density and Arrangement :** Tree density and distribution affect the forest's structure and the ecosystem as a whole.

- **Environmental History :** Includes historical factors such as natural fires and past human intervention, which can significantly influence forest composition and biological diversity.

- **Soil Quality :** Soil properties such as chemical composition and moisture availability affect the health of the forest and the growth of plants within it.

- **Genetic Diversity :** Relates to the diversity of genes within different species in the forest, contributing to the forest's ability to adapt to environmental changes, diseases, and pest attacks.

These factors related to biological diversity in forests are crucial for ensuring the health and sustainability of forest ecosystems. By balancing these factors, we can enhance the ability to

adapt to environmental changes and preserve the natural environment for current and future generations. [8]

## 4. Advantages of the Forest Landscape

Forest landscapes offer numerous advantages that benefit both the environment and human well-being. Some of the key benefits include:

- **Climate regulation :** Forests help to mitigate climate change by absorbing and storing carbon, and they can also help protect communities from the impacts of climate change. [9]

- **Aesthetic beauty and relaxation :** Forests provide a visually appealing environment that promotes feelings of relaxation and well-being.

- **Water conservation and filtration :** Forests play an important role in the global water cycle, filtering out pollution and chemicals, and improving the quality of water available for human use.

- **Soil conservation and erosion control :** Trees increase soil permeability, reducing surface runoff, soil erosion, and sedimentation of streams.

- **Energy savings :** Strategically placed trees can help reduce heating and cooling costs by providing shade and windbreaks.

- **Noise reduction :** Trees absorb and block noise from the urban environment.

- **Wildlife and plant diversity :** Forests provide habitats for plants and animals, including many iconic species.

- **Economic benefits :** Forests enhance community economic stability by attracting businesses and tourists, and they can increase property values.

- **Recreational opportunities :** Forests offer nature-based recreational opportunities, such as hiking, birdwatching, and camping.

- **Health benefits :** Time spent in forests has been shown to have positive benefits on conditions including cardiovascular disease, respiratory concerns, diabetes, and mental health.

- **Biodiversity conservation :** Forests are home to over 80% of terrestrial biodiversity, and their protection is crucial for maintaining global biodiversity.

These advantages highlight the importance of forests for both environmental and human well-being, and they underscore the need for sustainable forest management and conservation efforts. [10], [11], [12].

## 5. Challenges facing the forest landscape

The challenges facing the world's forests are diverse and significant. Some of the key challenges include:

- **Deforestation and Forest Degradation :** These challenges pose significant threats to forests worldwide. These activities result in the loss of forest cover, biodiversity, and ecosystem services, leading to various environmental and social impacts. [13]
- **Climate Change :** Climate change impacts forests through altered precipitation patterns, increased temperatures, and extreme weather events such as storms and wildfires. These changes can affect forest health, species composition, and the distribution of ecosystems. [14]
- **Illegal Logging and Wildlife Trade :** Illegal logging contributes to deforestation and forest degradation, leading to habitat loss and ecosystem disruption. Wildlife trade threatens forest biodiversity by exploiting species for various purposes, including pets, food, and traditional medicine. [15]
- **Land Use Change :** Conversion of forested land for agriculture, urban development, and infrastructure projects further diminishes forest cover and fragments ecosystems. Land use change often leads to habitat loss, soil erosion, and water pollution. [16]

## 6. Solutions to address challenges facing the forest landscape

Addressing the challenges faced by forests requires a multifaceted approach that addresses root causes and promotes sustainable practices. Some solutions include:

- **Remoting sustainable management practices :** including reforestation, effective management of forest resources, and encouraging response to desertification and forest degradation, must be developed and strengthened.
- **Encouraging sustainable agriculture :** Sustainable agriculture that respects forests and biodiversity can be encouraged, reducing the need to convert more forest land for agricultural purposes. [17]
- **Use of modern technology:** Modern technology such as remote sensing and geographic information systems can be used to monitor changes in forest coverage, predict potential threats, and take appropriate action. [18]

- **Strengthening legislation and its implementation :** Legislation on forest protection must be strengthened and effectively implemented, including combating illegal desertification and enforcing laws related to the illegal wildlife trade.
- **Strengthening the protection of vital areas:** Protection of important biological areas should be increased by expanding networks of natural reserves and strengthening procedures for their effective protection and management. [19]

## Conclusion

By protecting and restoring forest landscapes, we can safeguard biodiversity, mitigate climate change, and secure vital ecosystem services for the well-being of both nature and humanity. It's imperative that we recognize the value of forests and work together to ensure their preservation and resilience in the face of ongoing challenges.

In matter or recognizing the value of forests, the next chapter presents some machine learning techniques to be applied for identifying various landscapes of forest.

# CHAPTER 2

# UNSUPERVISED MACHINE LEARNING TECHNIQUES

## Introduction

Unsupervised classification stands as a powerful tool in the realm of data analysis and machine learning, offering a means to extract meaningful insights and structure from unlabeled datasets. Through methods such as clustering and dimensionality reduction, it uncovers hidden patterns and groupings within data, providing valuable understanding without the need for explicit guidance.

## 1.Definition

Unsupervised classification is a machine learning technique used in data analysis and pattern recognition, particularly in the field of remote sensing and image processing. In unsupervised classification, the algorithm clusters input data into groups or classes based on the inherent structure and similarities within the data itself, without the need for pre-labeled training data. This means that the algorithm identifies patterns and relationships within the data autonomously, without prior knowledge or guidance from a human expert.

The goal of unsupervised classification is to divide a dataset into classes or groups without having previously directed the data. This type of classification is used to understand structural or structural relationships in data and discover hidden or unknown patterns. It is generally used in analyzing and exploring data to understand it better. [20]

## 2.Models of Unsupervised Classification (Clustering):

Unsupervised classification, particularly in the context of clustering, involves grouping unlabeled data points based on similarities. Clustering is a technique of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar to each other than those in other groups (clusters). Clustering is a widely used technique in machine learning, data analysis, data mining, image analysis, and other fields. Some common models and algorithms used in clustering are presented in the following.

## 2.1 k-means Clustering Algorithm

### 2.1.1 Definition

The k-Means Clustering algorithm is one of the most popular clustering algorithms. It is an iterative algorithm that partitions a dataset into k clusters. The algorithm starts with an initial set of k centroids (cluster centers) randomly selected from the data points. The algorithm then assigns each data point to the nearest centroid, forming k clusters. The mean of each cluster

becomes the new centroid, and the algorithm repeats until the centroids no longer move significantly (Fig. 2.1 )

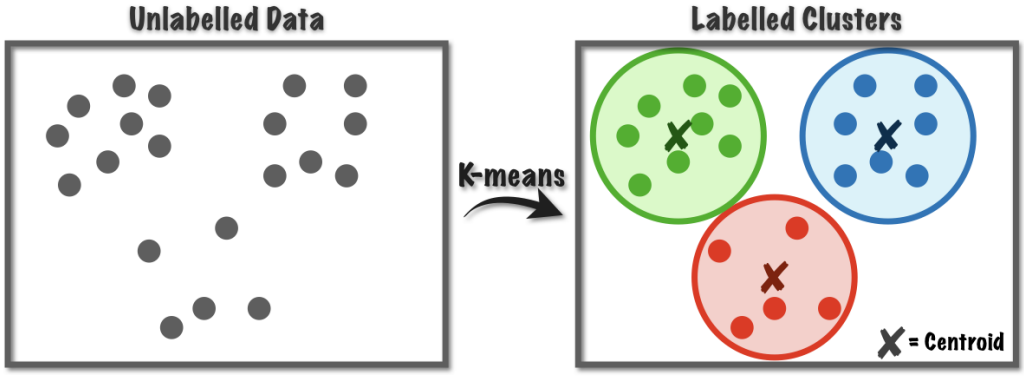



**Fig. 2.2:** Understanding K-Means Clustering

### 2.1.2 Advantages and disadvantages of k-means clustering algorithm

- **Advantages:** The k-means clustering algorithm offers several advantages:
  - **Simplicity:** K-means is straightforward and easy to implement, making it accessible even for those with limited experience in clustering algorithms.
  - **Efficiency :** It is computationally efficient and can handle large datasets efficiently, making it suitable for applications with high-dimensional data.
  - **Scalability :** K-means scales well with the number of data points, making it suitable for both small and large datasets.
  - **Versatility :** It can be applied to various types of data, including numerical and categorical, making it versatile for a wide range of applications.
  - **Interpretability:** The clusters generated by k-means are easy to interpret and understand, making it useful for exploratory data analyst.
  - **Speed :** It converges quickly, especially with random initialization, making it suitable for applications where real-time or near-real-time clustering is required.
- **Disadvantages of K-means:** K-means clustering is a popular unsupervised machine learning algorithm used for grouping similar data points together. However, it has several disadvantages that should be considered:
  - **Determining the Number of Clusters (K):** One of the major challenges in K-means clustering is determining the optimal number of clusters (K) beforehand. The algorithm requires the user to specify the number of clusters in advance, which can be difficult to determine, especially if there is no prior knowledge

about the data. Choosing an inappropriate value for K can lead to suboptimal clustering results.

- **Sensitivity to Initialization:** K-means clustering is sensitive to the initial centroid positions. Different initializations can lead to different clustering results, as the algorithm may converge to different local optima. This issue can be partially mitigated by using multiple random initializations and selecting the best result, but it adds computational complexity.
- **Inability to Handle Non-Convex Clusters:** K-means clustering assumes that the clusters are convex and spherical in nature. It struggles to identify clusters with complex shapes or non-convex geometries, such as clusters with different densities or elongated shapes.
- **Sensitivity to Outliers:** K-means clustering is sensitive to outliers or noise in the data. Outliers can significantly affect the positioning of the centroids, leading to poor clustering results. In such cases, it is necessary to preprocess the data to remove or handle outliers before applying K-means.
- **Equal Cluster Sizes:** K-means clustering tends to produce clusters with roughly equal sizes, even if the underlying data distribution has clusters of varying densities or sizes. This can lead to inaccurate representations of the true cluster structure.
- **Difficulty with High-Dimensional Data:** K-means clustering can struggle with high-dimensional data due to the curse of dimensionality. As the number of dimensions increases, the distances between data points become less meaningful, and the algorithm may fail to identify meaningful clusters.
- **Lack of Cluster Interpretation:** K-means clustering is a purely data-driven approach and does not provide any inherent interpretation or meaning for the resulting clusters. The clusters are formed based solely on the similarity of data points, and it is up to the analyst to interpret and assign meaning to the clusters based on domain knowledge or further analysis.

### 2.1.3 Types of k-Means Clustering

There are several variations of the k-Means Clustering algorithm, depending on the type of data and the objectives of clustering. Here are some of the common types of k-Means Clustering:

- **Hard Clustering :** In hard clustering, each data point belongs to only one cluster. This is the most common type of clustering used in the k-Means Clustering algorithm.

- **Soft Clustering :** In soft clustering, each data point can belong to multiple clusters with different degrees of membership. Soft clustering is also known as fuzzy clustering.
- **Online Clustering :** In online Clustering, the algorithm Can handle streaming data, where new data points arrive continuously. The algorithm updates the centroids and the clusters incrementally.
- **Spectral Clustering :** In spectral clustering, the algorithm uses the eigenvectors of the data similarity matrix to cluster the data. Spectral clustering is useful for non-linearly separable data.

### 2.1.4 Principles of k-Means Clustering

The k-Means Clustering algorithm is based on several principles. Here are some of the key principles:

- **Distance Metric:** The algorithm uses a distance metric (such as Euclidean distance) to measure the distance between data points and centroids. The data points are assigned to the nearest centroid based on the distance metric.
- **Centroid Initialization :** The initial centroids can affect the final clustering result. Randomly selecting the initial centroids may result in suboptimal clustering. Therefore, some methods (such as k-Means++) are used to initialize the centroids more effectively.
- **Convergence Criteria:** The algorithm repeats the assignment and centroid update steps until the centroids no longer move significantly. The convergence criteria can be defined based on the maximum number of iterations, the minimum change in the centroid position, or the minimum change in the objective function (such as the sum of squared distances between data points and centroids).
- **Scalability:** The k-Means Clustering algorithm can handle large datasets efficiently. However, the computational complexity of the algorithm increases with the number of data points and the number of clusters. Therefore, some methods (such as mini-batch k-Means) are used to cluster large datasets faster.

Here are the steps of the k-Means Clustering algorithm:

- Choose the number of clusters k.
- Randomly select k data points from the dataset as the initial centroids.
- Assign each data point to the nearest centroid, forming k clusters.
- Calculate the mean of each cluster, and set the mean as the new centroid. [21]

## 2.2 Hierarchical clustering algorithm

### 2.2.1 Definition

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. A hierarchical clustering method can be thought of as a set of ordinary (flat) clustering methods organized in a tree structure. These methods construct the clusters by recursively partitioning the objects in either a top-down or bottom-up fashion. In this paper we present a new hierarchical clustering algorithm using Euclidean distance. To validate this method we have performed some experiments with low dimensional artificial datasets and high dimensional fMRI dataset. Finally the result of our method is compared to some of existing clustering methods.

The purpose of data clustering algorithm is to form clusters (groups) of data points such that there is high intra-cluster and low inter-cluster similarity. There are different types of clustering methods such as hierarchical, partitioning, grid and density based. Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. [22]

Hierarchical algorithms can be expressed in terms of either graph theory or matrix algebra. A dendrogram, a special type of tree structure, is often used to visualize a hierarchical clustering.

## 2.2.2 Types of Hierarchical clustering algorithm

There are two types of Hierarchical Clustering:

1- **Agglomerative Hierarchical Clustering:** In Agglomerative Hierarchical Clustering, each data point is considered as a separate cluster, and then clusters are merged iteratively until all data points belong to a single cluster (Fig. 2.2). The merging is done based on the similarity between the clusters. The similarity can be measured using various distance metrics, such as Euclidean distance, Manhattan distance, and Cosine similarity.

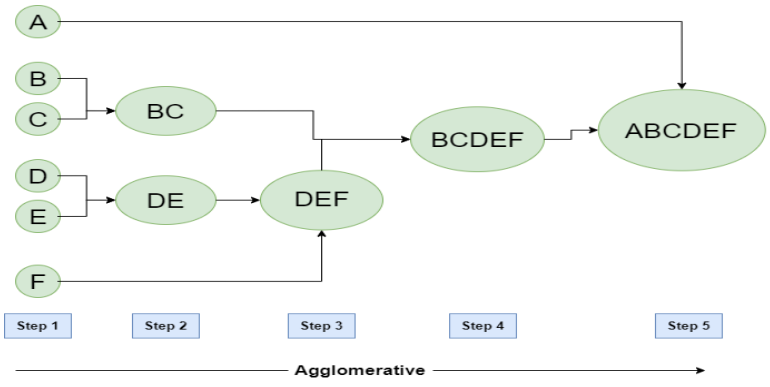



**Fig. 2.2:** Agglomerative Hierarchical Clustering

2- **Divisive Hierarchical Clustering:** In Divisive Hierarchical Clustering, all data points belong to a single cluster initially, and then clusters are divided iteratively until each data point belongs to a separate cluster. The division is done based on the dissimilarity between the clusters (Fig. 2.3). The dissimilarity can be measured using the same distance metrics used in Agglomerative Hierarchical Clustering.

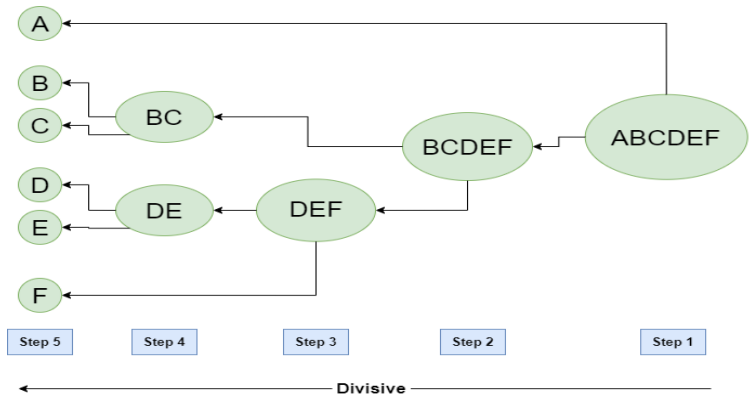



**Fig. 2.3:** Divisive Hierarchical Clustering

Overall, hierarchical clustering algorithms provide a hierarchical decomposition of the data, allowing for the identification of clusters at different levels of granularity. This hierarchical structure offers insights into the relationships and structures within the dataset, making hierarchical clustering a valuable tool for exploratory data analysis and pattern recognition.

### 2.2.3 Advantages and disadvantages of hierarchical clustering algorithms

❖ **Advantages**

- **No prespecified number of clusters:** Unlike k-means and other partitioning methods, hierarchical clustering does not require specifying the number of clusters beforehand, making it suitable for exploratory data analysis.
- **Flexibility :** Hierarchical clustering can handle various types of data and distance measures, allowing for flexibility in clustering different types of datasets.
- **Interpretability :** The dendrogram produced by hierarchical clustering provides a clear and interpretable representation of the clustering process, making it easy to understand and communicate the results.
- **Robustness to Noise :** Hierarchical clustering is robust to noise and outliers since it considers the entire dataset during the clustering process and does not rely on initializations like some other methods.

- **Agglomerative and Divisive Approaches:** Hierarchical clustering algorithms offer both agglomerative (bottom-up) and divisive (top-down) approaches, allowing users to choose the method that best suits their data and objectives.

❖ **Disadventages:**

- It requires prior knowledge of the number of clusters.
- It is sensitive to the initial choice of centroids.
- It does not handle outliers and noise in the data.
- It can converge to a suboptimal solution. [21]

## 2.3 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

### 2.3.1 Definition

Density-based clustering is a technique that identifies clusters based on the density of data points in the dataset. It works by identifying regions of high density and separating them from regions of low density. [21]

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is designed to discover the clusters and the noise in a spatial Database. Ideally, we would have to know the appropriate parameters *Eps* and *MinPts* of each cluster and at least one point from the respective cluster. Then, we could retrieve all points that are density-reachable from the given point using the correct parameters. To get this information in advance for all Clusters of the database, there is a simple and effective heuristic to determine the parameters *Eps* and *MinPts* of the "thinnest", i.e. least dense, cluster in the database. Therefore, DBSCAN uses global values for *Eps* and *MinPts*, i.e. the same values for all clusters. The density parameters of the "thinnest" cluster are good candidates for these global parameter values specifying the lowest density which is not considered to be noise. [23]

### 2.3.2 Parameters Required for DBSCAN Algorithm

- **Eps:** It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to '*eps*' then they are considered neighbors. If the *eps* value is chosen too small then a large part of the data will be considered as an outlier. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the *eps* value is based on the k-distance graph.

- **MinPts:** Minimum number of neighbors (data points) within *eps* radius. The larger the dataset, the larger value of *MinPts* must be chosen. As a general rule, the minimum *MinPts* can be derived from the number of dimensions *D* in the dataset as, *MinPts* ≥ *D+1*. The minimum value of *MinPts* must be chosen at least 3. [24]

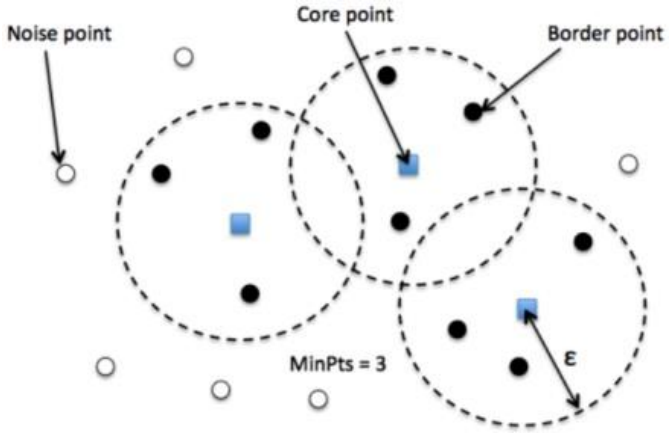



**Fig. 2.4:** Components of DBSCAN

### 2.3.3 Advantages and disadvantages of DBSCAN algorithm

❖ **Advantages**

- Efficiently identify irregular clusters and noise in large space databases.
- Resistant to noise and the ability to distinguish individual points as unclassified points
- It does not require knowing the number of clusters in advance, making it suitable for cases where you do not know the prior structure of the data. [25]

❖ **Disadvantages:**

- Sensitivity to parameters such as epsilon and MinPts.
- Performance impacted by data size and dimensions.
- Difficulty in handling clusters with highly variable densities. [26]

## 2.4 OPTICS clustering algorithm

### 2.4.1 Definition

OPTICS (Ordering Points to Identify the Clustering Structure) is a clustering algorithm that identifies clusters in space databases based on the order of points and the hierarchical clustering structure they provide. This is done by constructing a reachability curve that represents the reachable distance from each point relative to the nearest dense point. This algorithm offers advantages such as the ability to handle clusters with variable densities and

providing a hierarchical clustering structure that allows data to be examined at different levels of detail (Fig. 2.5).

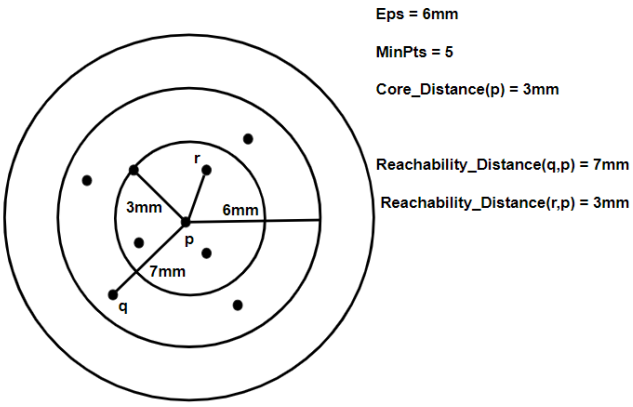



**Fig. 2.5 :** OPTICS Clustering

### 2.4.2 Main steps of OPTICS algorithm

- **Calculating Reachability Distance :** This step involves calculating the reachability distance between each point in the dataset and other points.
- **Core Distance Calculation :** Core distance is used to determine core points in the dataset.
- **Ordering Points:** This step involves ordering points based on their reachability distance to create the reachability plot.
- **Clustering:** Clustering involves extracting clusters from the reachability plot to determine cluster boundaries.

### 2.4.3 Advantages and disadvantages of OPTICS algorithm

❖ **Advantages**

- **Ability to Identify Arbitrary-Shaped Clusters** : OPTICS can identify clusters of arbitrary shapes and sizes, making it suitable for datasets with complex structures.
- **No Need to Specify the Number of Clusters**: Unlike some clustering algorithms (e.g., k-means), OPTICS does not require the number of clusters to be specified in advance, which makes it more flexible. [27]
- **Robustness to Noise**: OPTICS is robust to noise in the data due to its density-based nature, which helps in handling noisy datasets effectively.

❖ **Disadvantages :**

- **Computationally Expensive**: OPTICS can be computationally expensive, especially for large datasets, due to its density-based calculations and the need to construct a reachability plot. [28]
- **Sensitive to Parameter Settings**: OPTICS requires careful tuning of parameters such as the minimum number of points and the radius for core points, and improper parameter settings may affect its performance. [29]
- **Challenges with High-Dimensional Data**: Like many clustering algorithms, OPTICS may struggle with high-dimensional data where the curse of dimensionality comes into play. [30]

## 2.5 MEAN SHIFT clustering algorithm

### 2.5.1 Definition

Mean Shift is an unsupervised clustering algorithm used for data clustering in multidimensional space. The algorithm aims to locate the high-density centroids (or centers) in the data and gradually shift points towards these central points until they settle into potential cluster centers. Mean Shift relies on the concept of average deviation to determine changes in density, and then moves points towards the density increase.

The goal of the Mean Shift clustering algorithm is to identify the modes or high-density regions in the dataset and assign each data point to its corresponding mode or cluster. This is achieved by iteratively shifting each data point towards the nearest mode in the feature space until convergence. The final result is a set of clusters, where each cluster represents a group of data points that share similar characteristics or belong to the same high-density region in the data space. In essence, Mean Shift aims to discover the underlying structure of the data by finding dense regions and grouping data points accordingly. It is particularly effective for datasets with complex and irregular shapes, as it does not make any assumptions about the shape or size of the clusters. Instead, it adapts dynamically to the density distribution of the data, making it a versatile and powerful clustering algorithm.

### 2.5.2 Method of Mean Shift

The Mean Shift algorithm is a non-parametric clustering technique that aims to locate the modes or dense regions in the data space. It works by iteratively shifting data points towards

the mode of the underlying data distribution until convergence. Here's a general overview of the method:

- **Initialization :** Begin by selecting a set of initial data points as the starting centroids or modes.
- **Kernel Estimation :** Define a kernel function (e.g., Gaussian kernel) that determines the influence of neighboring points on each data point. Kernel Estimation represents the direction and magnitude of the shift towards the mode. This is computed as the weighted average of the differences between the data point and its neighbors, weighted by the kernel function.
- **Update Data Points :** Shift each data point according to its mean shift vector. This step moves the data points towards regions of higher density.
- **Convergence Check :** Repeat steps 3 and 4 until convergence criteria are met. Convergence can be defined based on a maximum number of iterations or when the mean shift vectors become small.
- **Cluster Assignment :** Assign each data point to the mode towards which it converged. Data points that converge to the same mode are grouped into the same cluster.

The key idea behind Mean Shift is that each data point is attracted to regions of higher density in the data space. By iteratively shifting towards these dense regions, the algorithm effectively identifies clusters in the data without requiring prior knowledge of the number of clusters. [31]

This algorithm is particularly useful for clustering datasets with irregular shapes and varying densities, as it does not require prior knowledge of the number of clusters and adapts dynamically to the data distribution.

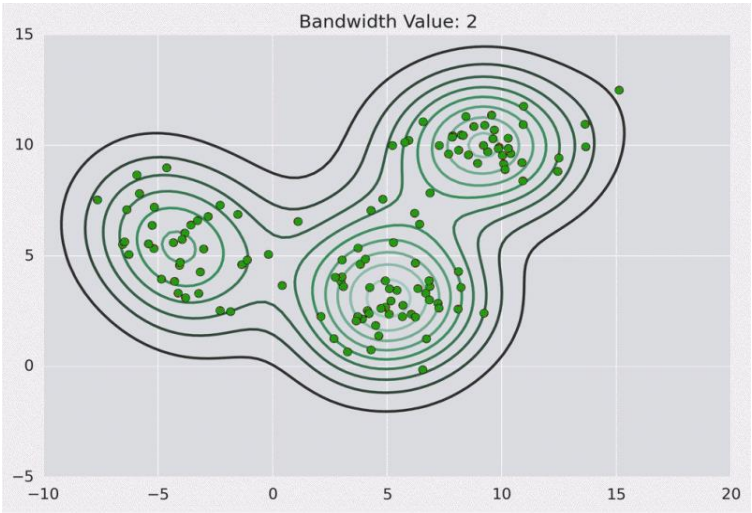



**Fig. 2.6 :** Mean Shift Clustering

### 2.5.3 Advantages and Disadvantages of Mean-Shift Clustering

❖ **Advantages:**

- It does not need to make any model assumption as like in K-means or Gaussian mixture.
- It can also model the complex clusters which have nonconvex shape.
- It only needs one parameter named bandwidth which automatically determines the number of clusters.
- It does not need to make any model assumption as like in K-means or Gaussian mixture.
- It can also model the complex clusters which have nonconvex shape.

❖ **Disadvantages:**

- We do not have any direct control on the number of clusters but in some applications, we need a specific number of clusters.
- It cannot differentiate between meaningful and meaningless modes.
- We do not have any direct control on the number of clusters but in some applications, we need a specific number of clusters. [32]

## 2.6 Affinity propagation clustering algorithm

### 2.6.1 Definition

The Affinity Propagation clustering algorithm is a modern approach to data clustering that aims to discover exemplars or representative examples from the data points. It does so by identifying a set of exemplars that best represent the data, without requiring the user to specify the number of clusters beforehand. The algorithm relies on the principle of message passing between data points to update their responsibility and availability values.

### 2.6.2 Key-steps of Affinity propagation

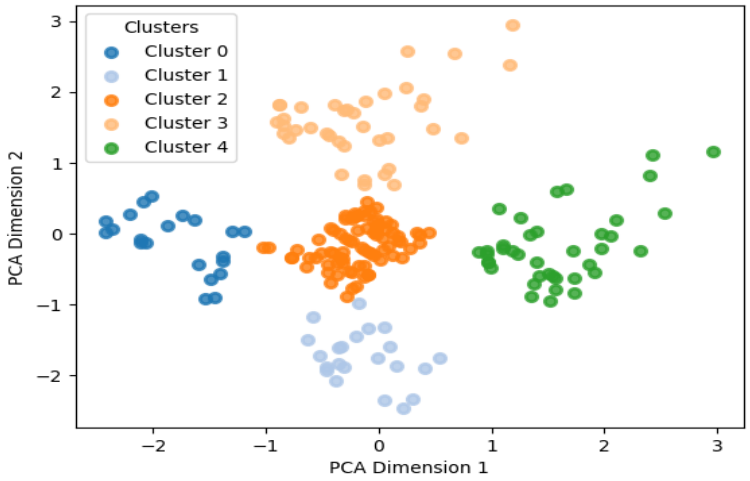
- **Similarity Matrix Calculation :** Compute a similarity matrix that measures the similarity between pairs of data points. This can be based on any suitable similarity metric, such as Euclidean distance or correlation.
- **Preference Initialization :** Initialize a preference matrix that reflects the initial preference of each data point to be chosen as an exemplar. This preference value influences the likelihood of a data point becoming an exemplar.
- **Message Passing :** Iteratively update the responsibility and availability values for each data point. The responsibility of data point to serve as the exemplar for data point is

updated based on the similarities and availability values, while the availability of data point is updated based on the responsibilities.

- **Exemplar Selection** : After convergence, select exemplars based on the responsibility and availability values. Each data point is assigned to the exemplar with the highest combined responsibility and availability.
- **Cluster Assignment :** Assign each data point to its corresponding exemplar to form clusters.

The Affinity Propagation method identifies exemplars or representative examples from the data points based on their similarities and preferences. It does not require the user to specify the number of clusters beforehand and can automatically determine the optimal number of clusters from the data.

This method is particularly suitable for applications where identifying representative examples is important, such as image segmentation, gene expression analysis, and natural language processing. [33]




**Fi. 2.7:** Affinity propagation clustering of mall customers.

### 2.6.3 Advantages and Disadvantages of Affinity Propagation Clustering

❖ **Advantages**

- **Automatic Clustering:** Affinity Propagation automatically determines the number of clusters, which can be advantageous in scenarios where the optimal number of clusters is not known beforehand.

- **No Need for Predefined Cluster Shapes:** Affinity Propagation does not assume any predefined shapes for the clusters, making it suitable for datasets with irregular or non-convex clusters. [34]
- **Automatic Cluster Number Determination:** Affinity Propagation automatically determines the number of clusters, reducing the need for manual intervention.

❖ **Disadvantages:**

- **Sensitivity to Parameters:** Affinity Propagation requires careful selection of parameters such as the preference values, which can affect the clustering results. [35]
- **Computational Complexity:** The algorithm can be computationally expensive, especially for large datasets, due to its iterative nature and the need to maintain a similarity matrix
- **Overfitting Potential:** If the preference value is not appropriately set, Affinity Propagation may lead to overfitting by selecting too many exemplars, resulting in an overly complex model. [34]

## 3. Uses of Unsupervised classification (Clustering)

- **Data Exploration :** Clustering helps in exploring the underlying structure of data by identifying natural groupings or clusters among data points.
- **Customer Segmentation :** In marketing and customer relationship management, clustering is used to segment customers into distinct groups based on their purchasing behavior, preferences, demographics, or other relevant factors. [36]
- **Genomic Data Analysis:** Clustering techniques are applied to genomic data to identify patterns or clusters of genes with similar expression profiles across different samples. [37]
- **Image Segmentation:** In image processing and computer vision, clustering is used for segmenting images into regions with similar characteristics such as color or intensity. [38]
- **Anomaly Detection:** Clustering can be used to identify outliers or anomalies in data, which deviate significantly from the normal behavior of the majority of data points. [39]
- **Document Clustering:** In natural language processing, clustering is employed to group similar documents together based on their content or topics. [40]

## 4.Importance of Unsupervised classification (Clustering)

- **Understanding Data Nature :** Unsupervised classification helps in understanding the nature and structure of data without prior knowledge of existing categories or patterns.
- **Pattern Discovery :** Clustering data can uncover hidden patterns or natural groupings in the data that may not be evident beforehand.

- **Dimensionality Reduction :** Clustering techniques can be used to reduce the dimensions of data, making it easier to analyze and interpret.
- **Performance Improvement :** Data clustering can be used to enhance the performance of predictive models by reducing noise or providing additional insights about categories.
- **Data Exploration :** Clustering can be utilized for data exploration and discovering relationships among different elements without the need for preconceived assumptions about the expected data structure.

## 5.The Most Common Challenges In Clustering Algorithms:

- **Determining the Optimal Number of Clusters**: This challenge involves finding the appropriate number of clusters in the data, which is often subjective and may require domain knowledge or validation techniques. [41]
- **Sensitivity to Initialization**: Many clustering algorithms are sensitive to the initial placement of centroids or seeds, which can lead to different clustering results for the same dataset. [42]
- **Handling Outliers**: Outliers can significantly impact the clustering process by skewing the centroids and affecting the formation of clusters. Robust techniques are required to handle outliers effectively. [43]
- **Scalability**: Some clustering algorithms may struggle to handle large datasets efficiently due to their computational complexity, resulting in increased runtime and memory requirements. [44]
- **Interpreting Results**: Interpreting and evaluating the quality of clustering results can be challenging, especially in high-dimensional or complex datasets where visual inspection is not feasible. Validating cluster quality objectively is crucial but can be difficult. [45]

## Conclusion

Unsupervised classification, or clustering, is a powerful method for analyzing data without predefined labels. It uncovers patterns and structures within datasets, aiding in tasks like exploration and anomaly detection. Despite challenges like determining optimal clusters, it remains crucial for insights in fields like machine learning and data analysis. Continued advancements ensure its relevance for uncovering hidden knowledge in large datasets.

# CHAPTER 3

# IDENTIFICATION OF HOMOGENEOUS FOREST LANDSCAPES

## Introduction

Data collection is the first and essential step in any data analysis process, as a variety of data must be collected from different sources such as databases, websites, text files, and other sources. We apply algorithms to it to extract knowledge and make decisions. In this chapter, we provide a description and sources for the data, and then we apply clustering algorithms to it.

## 1. Presentation of study area

Djebel Messaad is a mountainous region located in the M'sila province of northern Algeria, approximately 300 kilometers southeast of the capital city of Algiers. This area is part of the Saharan Atlas mountain range (Fig. 3.1).

The climate in the region is classified as Mediterranean, with hot, dry summers and mild, relatively wet winters. Annual precipitation averages around 400-600 mm, with most rainfall occurring between October and April. This precipitation pattern, coupled with the varied topography, has given rise to a diverse array of vegetation types and ecosystems.

Djebel Messaad is home to a rich diversity of forest ecosystems, including evergreen oak woodlands dominated by Algerian oak (Quercus canariensis) and Aleppo pine (Pinus halepensis) forests. At higher elevations, cedar (Cedrus atlantica) and juniper (Juniperus spp.) stands can be found, while the lower slopes support maquis shrublands and sparse vegetation adapted to the semi-arid conditions.


**Fig. 3.1:** Djebel Messaad Forest Map

The diversity of forest ecosystems, coupled with the complex terrain and varying environmental conditions, make Djebel Messaad an ideal study area for exploring techniques to identify homogeneous forest landscapes. The findings from such research could contribute

to the sustainable management and conservation of these valuable natural resources, while also supporting the livelihoods of local communities that depend on them.

## 2. Description of the dataset

### 2.1 Data collect

During the process of compiling our dataset, we consulted and utilized several references that guided us in its creation. We would like to highlight a few of them.

- **Forest Conservation Department of M'sila:** As trainees, we had the opportunity to visit the forest conservation department (Directorate) in M'sila. During our time there, they kindly provided us with valuable data about the Djebel Messaad forest in M'sila region.
- **Faculty of Biology:** In our research, we also sought help from the Faculty of Biology at the University of M'sila, where they have a specialty called Earth Biology. They graciously provided us with a link to their website, which hosts a repository of graduation research projects. This resource proved invaluable, as it helped us gain insights into the various plant species that thrive in the different soil types found in M'sila, with a particular focus on the Djebel Messaad forest ecosystem.
- **Visit to Djebel Messaad Forest:** We have the opportunity to conduct multiple visits to the study area, during which the forestry personnel graciously guided us through various regions of the forest. They provided us with valuable information about the diverse plant life and floral species found within the forest ecosystem.
- **Search on the internet:** Additionally, we supplemented our research by conducting online searches, particularly focused on identifying the specific floral species found within each plant family.

### 2.2. Description of plants and trees in Djebel Messaad Forest

After searching from various existing sources, we have summarized 18 plants and trees collected from Djebel Messaad forest in the Table 3.1, which contains the name of each plant or tree, a picture to facilitate its identification, and a brief description of it.

| Name & Picture | Description | Name & Picture | Description |
|---|---|---|---|
| Aleppo Pine (Pin d'Alep)  | a small to medium-sized tree, well-adapted to the Mediterranean climate | Phoenician juniper (Genévrier de Phénicie)  | is a small to medium-sized evergreen tree or shrub, found in the Mediterranean region |
| Evergreen oak (Le chêne vert)  | is a medium to large-sized tree , found in the Mediterranean region | Juniperus oxycedrus (Genévrier oxycedre)  | is a small to medium-sized evergreen tree or shrub, found in the Mediterranean region |
| Alpha  | small seeds inside round capsules that turn from green to brown when ripe, found in the Mediterranean region | Shrubs (les arbustes)  | Refers to natural areas typically composed of diverse vegetation such as small trees, shrubs, and grasses, often found in wild or natural areas, in English thickets. |
| Harmal  | Is an herbaceous plant, trailing leaves and small, yellow flowers. Their fruits typically grow on the ground and are usually round or oval in shape | Aloe Vera  | is characterized by its thick, fleshy leaves containing a natural gel, is widespread in arid and dry areas of Africa |
| Thuy  | Is a genus of coniferous trees and shrubs, are prized for their dense foliage, which consists of scale-like leaves that are arranged in flattened sprays. | Agave blue  | Are succulents native to the hot and arid regions, their thick, fleshy leaves that often have sharp spines at the tips? |
| Alfalfa  | Is a flowering plant in the legume family Fabaceae. It is one of the most cultivated forage crops worldwide, valued for its high nutritional content and versatility. | Sea orache (Atriplex halimus)  | Is a species of flowering plant in the family Amaranthaceous It is native to the Mediterranean region, is a shrub with silvery-gray foliage that helps it tolerate high salinity levels in its habitat. The leaves are small, alternate, and somewhat succulent. |
| prickly pear | A species of cactus native to Mexico but | Cranberry (canneberge) | Cranberries are a group of evergreen dwarf |

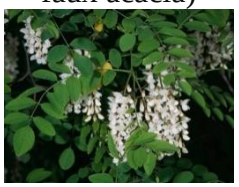
| | | | |
|---|---|---|---|
|  | also widely cultivated in the Mediterranean region, is a large, succulent cactus with flat, oval or pear-shaped pads |  | shrubs or trailing vines in the genus Vaccinium, fruits are small, round, red berries with a tart taste |
| Eucalyptus  | is a genus of flowering trees and shrubs in the myrtle family, are known for their large, aromatic leaves and showy flowers | Thyme  | is a perennial herb with culinary, medicinal, and ornamental uses, is a small, woody-stemmed plant with tiny, aromatic leaves |
| Rosemary (Romarin)  | is an evergreen herb with fragrant, needle-like leaves and small blue, purple, or white flowers and aromatic , is native to the Mediterranean region . | Menthe  | It refers to a genus of aromatic herbaceous plants within the family Lamiaceae, primarily the genus Menthe. Mint plants are known for their strong, refreshing aroma |

**Table 3.2:** Summary of the plants and trees in the study area

### 2.3. Description of flower species in Djebel Messaad Forest

After conducting research from various sources, we've compiled a summary of 25 flower species gathered from Dejbel Messaad forest in the table provided below (Table 3.2). The table includes the names of the flowers with its picture and a brief description of it.

| Name & Picture | Description | Name & Picture | Description |
|---|---|---|---|
| Aster  | the aster is a perennial plant belonging to the Asteraceae family, known for its beautiful star-shaped flowers | Dandelions (Pissenlits)  | Dandelions are common flowering plants belonging to the genus Taraxacum in the family asteraceae. Here's a detailed description of dandelions |
| Daisy (Pâquerette)  | Is characterized by its flower composed of a bright yellow center, called the capitulum, surrounded by delicate white petals arranged in multiple rows. | Sunflowers (Tournesols)  | These vibrant flowers are known for their large, bright yellow blooms that resemble the sun, hence their name, have tall, sturdy stalks and large, broad leaves |

| | | | |
|---|---|---|---|
| Red clover (Trèfle rouge)  | Is an herbaceous perennial plant, It has a slender, erect stem with trifoliate leaves, which consist of three leaflets that are typically oval to egg-shaped, with serrated margins. | Lupine (Lupinus)  | is a genus of flowering plants With hundreds of species distributed across various regions of the world, lupines exhibit a wide range of sizes, colors, and growth habits |
| Lavender (Lavande )  | Is an aromatic herbaceous plant belonging to the genus Lavandula, is renowned for its light floral fragrance and distinctive purple color. | Bee balm(Monarda)  | Also known as bee balm or bergamot, is a genus of flowering plants in the mint family, Lamiaceae. Herbaceous perennial plants, are prized for their colorful, showy flowers and aromatic foliage. |
| Verbena  | Species of annuals, perennials, and subshrubs, are known for their clusters of small, colorful flowers that bloom throughout the growing season. | Glycine max  | Widely grown for its edible bean which has numerous uses. |
| Judas tree (arbre de judee)  | is a small deciduous tree native to the Mediterranean is spectacular bright pink or purple flowers | black locust (robinier faux acacia)  | Is Flowering trees , is known for its distinctive compound leaves and fragrant white flowers, which bloom in late spring to early summer |
| sweet pea (pois de senteur)  | It refers to a flowering plant species belonging to the genus Lathyrus, particularly Lathyrus odoratus. are known for their delicate, fragrant flowers that bloom in a variety of colors | Fetuque (fescue )  | It refers to a genus of grasses within the family Poaceae, primarily the genus Festuca. |
| Coleus  | Is a genus of plants that are popular for their vibrant foliage, colorful leaves, which come in a | (Basilic) basil  | It refers to a culinary herb belonging to the genus Ocimum within the family Lamiaceae. Basil is |

| | | | |
|---|---|---|---|
| | wide range of patterns and hues. | | known for its distinctive aroma and flavor |
| sage (Sauge)  | It is characterized by its woody stems, aromatic leaves, and spikes of blue to purple flowers. | Melissa (Mélisse)  | It is lemon balm, It is known for its lemon-scented leaves and small white flowers. |
| Hyssop (Hysope)  | Is a herbaceous plant, is known for its aromatic leaves and spikes of blue, purple, or pink flowers. | rock rose (Helianthemum)  | Flowering plants commonly known as rock roses or sun roses. They belong to the family Cistaceae |
| Lily (Lilium)  | Are flowering plants that are renowned for their large, showy flowers and pleasant fragrance | Allium  | Are characterized by their long, green leaves and clustered, spherical flowers that grow atop their stems. |
| Arabidopsis  | Is a genus of flowering plants in the mustard, is model organism extensively used in plant biology research due to its small size, short life cycle, and ease of genetic manipulation. | Bergenia  | is a group of flowering plants , are known for their large, leathery leaves and their clusters of pink, red, or white flowers |
| Lunaria annua  | a biennial plant, its translucent dried fruits resembling coins | | |

**Table 3.2:** Summary of the flowers located in the study area

### 2.4. Description of the dataset attributes

The database contains 47 lines where we have included plants and flowers mentioned previously in Tables 3.1 and 3.2 with 13 columns, where each column contains properties of each plant and flower helping in the clustering process. These columns (attributes) are described as follows.

- Name of the plant or flower.

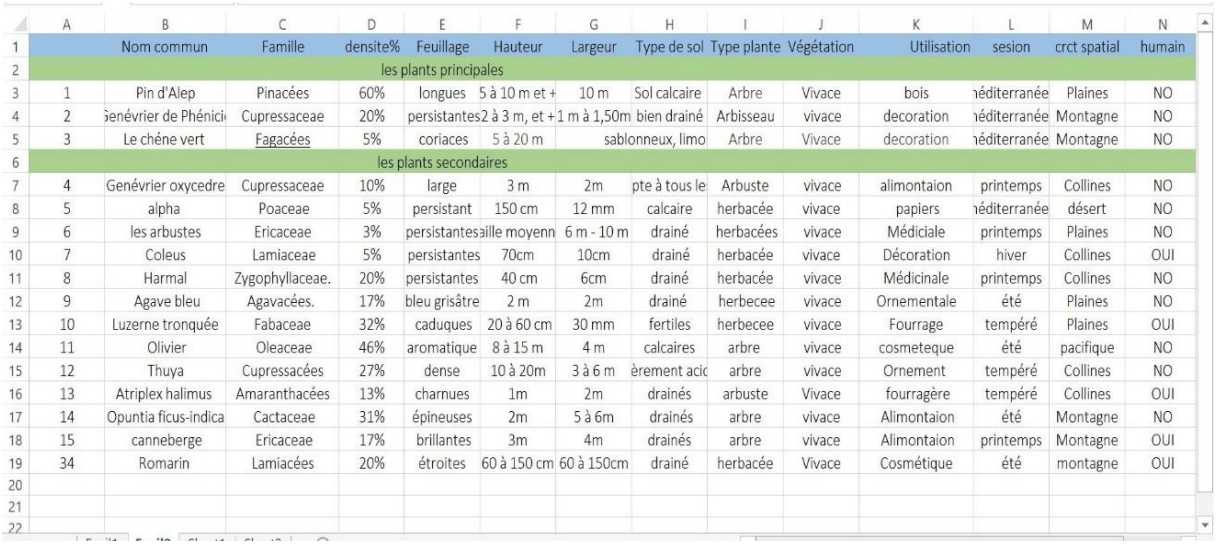
- Family of the plant or flower. We found these families: Asteraceae, fabaeae, Poaceae, Lamiaceae, Oleaceae, Poaceae, etc…

- Density of site in the Djebel Messaad forest, expressed in percentage.
- Shape of the leaves for example: large, evergreen, oval, alternate, rosette shape etc. …….
- Length of plant or flower.
- Width of the plant or flower.
- Type of soil, there are five common types: drained, fertile, organic, dry and loamy.
- Type of plant or flower (tree, shrub, Herbal).
- Vegetation, all the plants and flowers are perennial.
- Utilization for example: Medical, decoration, food Perfume, Cosmetics
- Season in which it is found (spring winter summer fall)
- Land where it is found (Hills, desert, Plains, Mountain).
- Human interventions in the growth of this plant, whether watering or grazing, with the answer being Yes or No.

After collecting all the plants and flowers found in Djebel Messaad forest, they were included in a datasets with their characteristics that help in clustering. Fig 3.2 and 3.3 shows samples of these datasets.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | | Nom commun | Famille | densite% | Feuillage | Hauteur | Largeur | Type de sol | ype plant | /égétatior | Utilisati | sesion | :rct spatia | humain |
| 3 | 16 | | | | les fleurs | | | | | | | | | |
| 4 | 17 | Aster | asteraceae | 60% | d'étoile | 10cm | 10mm | draine | herbacée | vivace | Decoration | printemps | Plaines | NON |
| 5 | 18 | pissenlits | asteraceae | 50% | rosette basale | 6 à 30cm | 15 mm | organiques | herbacée | vivace | Médiciale | printemps | tout | NON |
| 6 | 19 | pâquerette | asteraceae | 60% | ovales | 20m | 4 à 20 mm | argileux | herbacée | Vivace | Médiciale | printemps | montagne | NON |
| 7 | 20 | marguerite | asteraceae | 40% | centre jaune | 5cm | 70 cm | ordinaire | ıromatiqu | Vivace | decoration | printemps | desert | OUI |
| 8 | 21 | tournesols | asteraceae | 45% | ovales | 30 cm 50 cm | 50 cm | secs | herbacée | Vivace | nourriture | l'automne | tout | NON |
| 9 | 22 | Trèfle rouge | fabaeae | 30% | trois folioles | 40 cm | 30 cm | fertiles | herbacees | Vivace | Médiciale | printemps | tout | OUI |
| 10 | 23 | lupinus | fabaeae | 35% | forme de rosett | 2 à 5 cm | 30 à 90 cm | drainé | arbuste | Vivace | médicinales | hivers | montagne | NON |
| 11 | 24 | Glycine max | fabaeae | 35% | trifoliées | 5 à 15 cm | 5 à 10cm | drainés | arbre | Vivace | alimenatation | été | Plaines | OUI |
| 12 | 25 | arbre de judee | fabaeae | 30% | rondes | 10 m | 50 cm | calcaires | arbre | Vivace | Decoration | printemps | Plaines | OUI |
| 13 | 26 | obinier faux acac | fabaeae | 25% | folioles ovales | 12 à 30 m | 0,61 à 1,22m | humide | arbre | Vivace | Médiciale | été | Plaines | OUI |
| 14 | 27 | pois de senteur | fabaeae | 20% | alterne | 2 m | 0,7 m | drainé | ıromatiqu | Vivace | Parfum | l'automne | desert | OUI |
| 15 | 28 | fetuque | Poaceae | 20% | persistantes | 25cm | 15cm | secs | herbacée | Vivace | Decoration | printemps | Plaines | NON |
| 16 | 29 | Menthe | Lamiacées | 19% | ovales | 30 à 100 cm | 15cn | riche | herbacée | Vivace | Cosmétique | été | desert | NON |
| 17 | 30 | Basilic | Lamiacées | 15% | aromatique | 20 à 60 cm | 30 cm | acide | herbacée | Vivace | Cosmétique | printemps | Plaines | NON |
| 18 | 31 | Thym | Lamiacées | 13% | linéaires | 30 cm | 10 cm | drainé | herbacée | Vivace | Cosmétique | printemps | tout | OUI |
| 19 | 32 | Sauge | Lamiacées | 17% | ovales | 60 à 90 | 30 cm | pauvre | herbacée | Vivace | Cosmétique | printemps | tout | OUI |
| 20 | 33 | Lavande | Lamiacées | 16% | roites et lancéolé | 30 à 60 cm | 30 cm | sec | herbacée | Vivace | Parfum | printemps été | montagne | NON |
| 21 | 34 | Coleus | Lamiaceae | 5% | persistantes | 70cm | 10cm | drainé | herbacée | vivace | Décoration | hiver | Collines | OUI |
| 22 | 35 | Mélisse | Lamiacées | 15% | ovales, dentelées | 30 à 90 cm | 30 cm | drainé | ıromatiqu | Vivace | Parfum | été | Plaines | NON |
| 23 | 36 | Hysope | Lamiacées | 15% | ovales , cœur | 60 à 90 cm | 30 cm | drainés | ıromatiqu | Vivace | Alimontaion | été | montagne | OUI |
| 24 | 37 | Monarde | Lamiacées | 10% | pposées, dentée | 30 à 90 cm | 30 cm | drainé | ıromatiqu | Vivace | Médiciale | été | desert | NON |
| 25 | 38 | Helianthemum | cistacées | 9% | petites et simple | 10 à 30 cm | 20 cm | relativement | ıromatiqu | Vivace | Parfum | été | montagne | OUI |
| 26 | 39 | Lilium | Liliaceae | 9% | ncéolées à étroit | 60 cm | 30cm | ièrement acid | herbacée | Vivace | Decoration | été | tout | NON |
| 27 | 40 | Allium | Liliaceae | 8% | linéaires | 20 à 30 cm | 15à 20 cm | drainés | herbacée | Vivace | Alimontaion | printemps | montagne | NON |
| 28 | 41 | Arabette | Brassicaceae | 7% | simples | 30 à 60 cm | 10 cm | limoneux | herbacée | Vivace | Decoration | printemps | desert | NON |
| 29 | 42 | Bergénie | Brassicaceae | 5% | Persistant | 20 à 60 cm | 30 à 60 cm | humifère | herbacée | Vivace | Decoration | Printemps | montagne | NON |
| 30 | 43 | Lunaria annua | Brassicaceae | 7% | ovale | 60 à 90 cm | 30 à 45 cm | fertile | herbacée | lle ou bisa | Decoration | été | tout | NON |
| 31 | 44 | verbena | Brassicaceae | 8% | lancéolées | 60 à 90 cm | 30 à 45 cm | fertile | herbacée | nnuelle ou | Decoration | Printemps | tout | OUI |
| 32 | | | | | | | | | | | | | | |

**Fig. 3.2:**: A sample of the dataset of plants



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Nom commun | Famille | densite% | Feuillage | Hauteur | Largeur | Type de sol | Type plante | Végétation | Utilisation | sesion | crct spatial | humain |
| 2 | les plants principales | | | | | | | | | | | | | |
| 3 | 1 | Pin d'Alep | Pinacées | 60% | longues | 5 à 10 m et + | 10 m | Sol calcaire | Arbre | Vivace | bois | néditerranée | Plaines | NO |
| 4 | 2 | Genévrier de Phénici | Cupressaceae | 20% | persistantes | 2 à 3 m, et + | 1 m à 1,50m | bien drainé | Arbisseau | vivace | decoration | néditerranée | Montagne | NO |
| 5 | 3 | Le chéne vert | Fagacées | 5% | coriaces | 5 à 20 m | | sablonneux, limo | Arbre | Vivace | decoration | néditerranée | Montagne | NO |
| 6 | les plants secondaires | | | | | | | | | | | | | |
| 7 | 4 | Genévrier oxycedre | Cupressaceae | 10% | large | 3 m | 2m | pte à tous le | Arbuste | vivace | alimontaion | printemps | Collines | NO |
| 8 | 5 | alpha | Poaceae | 5% | persistant | 150 cm | 12 mm | calcaire | herbacée | vivace | papiers | néditerranée | désert | NO |
| 9 | 6 | les arbustes | Ericaceae | 3% | persistantes | aille moyenn | 6 m - 10 m | drainé | herbacées | vivace | Médiciale | printemps | Plaines | NO |
| 10 | 7 | Coleus | Lamiaceae | 5% | persistantes | 70cm | 10cm | drainé | herbacée | vivace | Décoration | hiver | Collines | OUI |
| 11 | 8 | Harmal | Zygophyllaceae. | 20% | persistantes | 40 cm | 6cm | drainé | herbacée | vivace | Médicinale | printemps | Collines | NO |
| 12 | 9 | Agave bleu | Agavacées. | 17% | bleu grisâtre | 2 m | 2m | drainé | herbecee | vivace | Ornementale | été | Plaines | NO |
| 13 | 10 | Luzerne tronquée | Fabaceae | 32% | caduques | 20 à 60 cm | 30 mm | fertiles | herbecee | vivace | Fourrage | tempéré | Plaines | OUI |
| 14 | 11 | Olivier | Oleaceae | 46% | aromatique | 8 à 15 m | 4 m | calcaires | arbre | vivace | cosmeteque | été | pacifique | NO |
| 15 | 12 | Thuya | Cupressacées | 27% | dense | 10 à 20m | 3 à 6 m | èrement acid | arbre | vivace | Ornement | tempéré | Collines | NO |
| 16 | 13 | Atriplex halimus | Amaranthacées | 13% | charnues | 1m | 2m | drainés | arbuste | Vivace | fourragère | tempéré | Collines | OUI |
| 17 | 14 | Opuntia ficus-indica | Cactaceae | 31% | épineuses | 2m | 5 à 6m | drainés | arbre | vivace | Alimontaion | été | Montagne | NO |
| 18 | 15 | canneberge | Ericaceae | 17% | brillantes | 3m | 4m | drainés | arbre | vivace | Alimontaion | printemps | Montagne | OUI |
| 19 | 34 | Romarin | Lamiacées | 20% | étroites | 60 à 150 cm | 60 à 150cm | drainé | herbacée | Vivace | Cosmétique | été | montagne | OUI |
| 20 | | | | | | | | | | | | | | |
| 21 | | | | | | | | | | | | | | |
| 22 | | | | | | | | | | | | | | |

**Fig. 3.3:** A sample of the dataset of Flowers

### 2.5. Preprocessing

Converting textual data about plants and flowers and their characteristics into binary data is an important process in data analysis and machine learning. This type of transformation can help improve the performance of models and algorithms when dealing with large data sets, Here is a description of how we convert text data to binary data: Defining text variables and properties: We have defined the properties that we would like to convert from text data to binary data.

Encoding text data into binary data: Use text data encoding techniques such as class coding. We create a binary variable for each possible value in the text variable. Assign one value to the binary variable corresponding to the value in the original data, and zero for the rest of the binary variables.

Finally, we verified the generated binary data to ensure that the conversion was successful and that the text data was correctly represented in binary.

## 3.Implementation

To assess and validate the effectiveness of the proposed method, our initial step involves transitioning into the implementation phase. Within this segment, we outline the array of tools and programming languages employed.

### 3.1 Programming language

Nowadays, there are several programming languages and each language has its own characteristics. Among its languages, our choice focused on Python.

Python is an interpreter, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. [46]

### 3.2 Development environment:

#### 3.2.1 Anaconda:

The Hub for Data Science and AI Collaboration Expert support and streamlined collaboration at every stage of the data science lifecycle. Source, build, and deploy with ease. Take your work from idea to integration alongside data scientists, open-source contributors, and partners, leveraging leading-edge tools along the way [47].

#### 3.2.2 TensorFlow:

TensorFlow is an interface for expressing machine learning algorithms and an implementation for executing such algorithms. A computation expressed using TensorFlow can be executed with little or no change on a wide variety of heterogeneous systems, ranging from mobile devices such as phones and tablets up to large-scale distributed systems of hundreds of machines and thousands of computational devices such as GPU cards. The system is flexible and can be used to express a wide variety of algorithms, including training and inference algorithms for deep neural network models, and it has been used for conducting research and for deploying machine learning systems into production across more than a dozen areas of computer science and other fields, including speech recognition, computer vision, robotics, information retrieval, natural language processing, geographic information extraction, and computational drug discovery. [48]

#### 3.2.3 Sklearn:

Scikit-learn is a key library for the Python programming language that is typically used in machine learning projects. Scikit-learn is focused on machine learning tools including mathematical, statistical and general purpose algorithms that form the basis for many machine learning technologies. As a free tool, Scikit-learn is tremendously important in many different types of algorithm development for machine learning and related technologies. [49]

#### 3.2.4 Pandas:

Pandas is a Python library used for working with data sets, It has functions for analyzing, cleaning, exploring, and manipulating data, The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008. [50].

## 4. Application of K-means Clustering Algorithm on the dataset of plants

We implemented K-Means algorithm to group the data into a certain number of clusters, and then performed verification in order to validate the results (Fig. 3.4 shows the K-means code on dataset of plants).

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import pandas as pd
file_path = 'C:/Users/HI-TECH/Downloads/Nouveau dossier (2)/nouveau 3/BDD final (1) 2.xlsx'
data = pd.read_excel(file_path)
data_cleaned = data.drop(columns=['Nom'])
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_cleaned)
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(data_scaled)
centers = kmeans.cluster_centers_
labels = kmeans.labels_
data['Cluster'] = labels

print(data)
```

**Fig. 3.4**: K-means code in python on the plants dataset.

After implementing the K-Means algorithm on the dataset of plants, we obtained three clusters, where each cluster groups plants with common characteristics. Fig. 3.5 shows the results.

| | Nom | Cluster |
|---|---|---|
| 0 | Aleppo Pine | 2 |
| 1 | Phoenician juniper | 0 |
| 2 | Evergreen oak | 2 |
| 3 | Juniperus oxycedrus | 0 |
| 4 | allies | 0 |
| 5 | Shrubs | 0 |
| 6 | Rosemary | 2 |
| 7 | cloves | 0 |
| 8 | Agave bleu | 0 |
| 9 | truncated alfalfa | 1 |
| 10 | Aloe vera | 1 |
| 11 | Thuya | 2 |
| 12 | Sea orache | 0 |
| 13 | prickly pear | 1 |
| 14 | Cranberry | 0 |
| 15 | Eucalyptus | 1 |
| 16 | Menthe | 2 |
| 17 | Thyme | 1 |

**Fig. 3.5:** Clusters of plants using K-means

In the table 3.3, we display a summary of each cluster of plants, indicating the name of each plant alongside its image with its classification into the specific cluster determined by the K-Means algorithm. It can be seen that the plants are well clustered based on their characteristics.

| **Clusters** | **Name of plant & Picture** | | | |
|---|---|---|---|---|
| Cluster 1 | Phoenician juniper  | Juniperus oxycedrus  | Cloves  | Cranberry  |
| | Sea orache  | Allies  | Agave ble  | Shrubs  |
| Cluster 2 | Truncated alfalfa  | Thyme  | Eucalyptus  | Aloe Vera  |
| | Prickly pear  | | | |
| Cluster 3 | Aleppo Pine  | Thuya  | Rosemary  | Menthe  |

**Table 3.3** : Detailed clusters of plants using K-means

## 5. Application of K-means Clustering Algorithm on the dataset of flowers

We implemented K-Means algorithm to group the flower species into a certain number of clusters, and then performed verification in order to validate the results (Fig. 3.6 shows the K-means code on dataset of flowers).

```python
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import pandas as pd
file_path = 'C:/Users/HI-TECH/Downloads/Nouveau dossier (2)/nouveau 3/datasetsf.xlsx'
data = pd.read_excel(file_path)
data_cleaned = data.drop(columns=['Nom'])
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_cleaned)
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(data_scaled)
centers = kmeans.cluster_centers_
labels = kmeans.labels_
data['Cluster'] = labels

print(data)
```

**Fig. 3.6**: K-means code on the flowers dataset.

After implementing the K-Means algorithm, on flowers we obtained 3 clusters, where each cluster represents flowers with common characteristics (Fig. 3.7)

| | Nom | Cluster |
|---|---|---|
| 0 | Aster | 2 |
| 1 | Dandelion | 2 |
| 2 | Daisy | 1 |
| 3 | marguerite | 0 |
| 4 | Sunflowers | 2 |
| 5 | Red clover | 1 |
| 6 | Lupine | 0 |
| 7 | Glycine max | 0 |
| 8 | Judas tree | 0 |
| 9 | black locust | 1 |
| 10 | sweet pea | 1 |
| 11 | fetuque | 2 |
| 12 | Lunaria annua | 1 |
| 13 | basil | 2 |
| 14 | Bergenia | 2 |
| 15 | sage | 2 |
| 16 | Lavender | 2 |
| 17 | Coleus | 2 |
| 18 | Melissa | 1 |
| 19 | Hyssop | 1 |
| 20 | Bee balm | 1 |
| 21 | rock rose | 2 |
| 22 | Lily | 2 |
| 23 | Allium | 2 |
| 24 | Arabidopsis | 2 |

**Fig. 3.7:** Clusters of the flowers using K-means

In Table 3.4, we display the name of each flower alongside its image, along with its classification into the specific cluster determined by the K-Means algorithm. We can observe the diversity in the plants and how they have been clustered into different groups based on their characteristics.

| Clusters | Name of flower & picture | | | |
|---|---|---|---|---|
| Cluster 1 | Daisy  | Lupine  | Glycine max  | Verbena  |
| | Judas tree  | | | |
| Cluster 2 | Red clover  | Black locust  | Sweet pea  | Lunaria annua  |
| Cluster 3 | Sunflowers  | Lily  | Bergenia  | Basil  |
| | Sage  | Coleus  | Fetuque  | Rock rose  |
| | Lavender  | Allium  | Arabidopsis  | Dandelion  |
| | Aster  | | | |

**Table 3.4** : Detailed clusters of flowers using K-means

## 6. Application of Hierarchical Clustering Algorithm on the dataset of plants

we implemented Hierarchical algorithm to group the data into a certain number of clusters, and then performed verification in order to validate the results (Fig. 3.8 shows the Hierarchical code on dataset of plants).

```
In [1]: from sklearn.cluster import KMeans
        import matplotlib.pyplot as plt
        import seaborn as sns
        import pandas as pd

        # Load data
        data = pd.read_excel('C:/Users/HI-TECH/Downloads/Nouveau dossier (2)/nouveau 3/BDD final (1) 2.xlsx')

        # Select feature columns (excluding the 'Nom' column)
        X = data.drop(columns=['Nom'])

        # Apply hierarchy Clustering
        from scipy.cluster.hierarchy import dendrogram, linkage
        from scipy.cluster.hierarchy import fcluster

        # Perform hierarchical clustering
        linked = linkage(X, method='ward')
        data['Cluster_Hierarchical'] = fcluster(linked, t=3, criterion='maxclust')

        # Plot the dendrogram
        plt.figure(figsize=(10, 7))
        dendrogram(linked,
                   orientation='top',
                   labels=data['Nom'].values,
                   distance_sort='descending',
                   show_leaf_counts=True)
        plt.title('Hierarchical Clustering Dendrogram')
        plt.show()
```

**Fig. 3.8**: Hierarchical code on the plants dataset.

After implementing the Hierarchical algorithm, on flowers we obtained dendrogram of the clusters , where each cluster represents plants with common characteristics (Fig. 3.9)

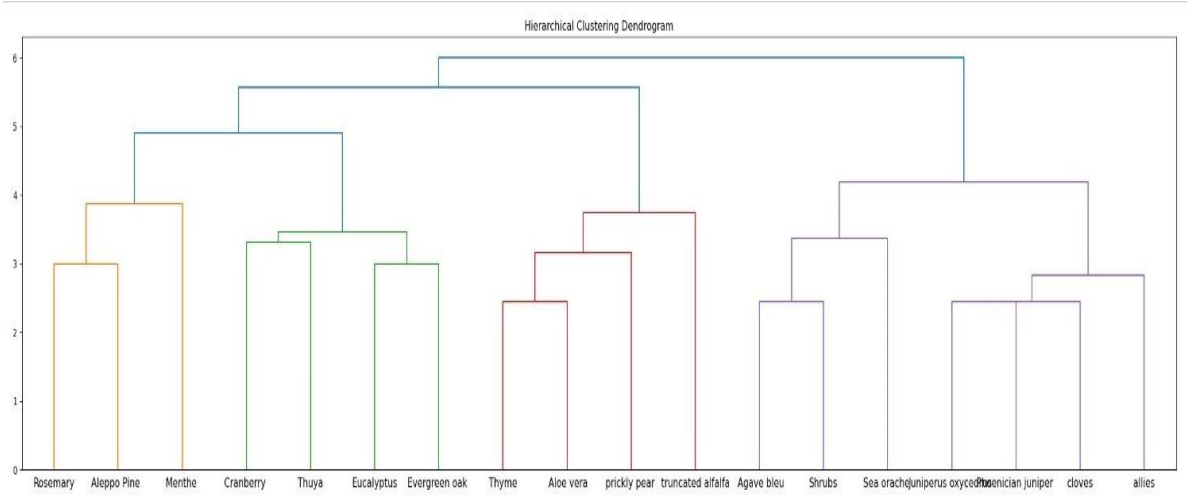



**Fig. 3.9:** Clusters of the plants using Hierarchical

## 7. Application of Hierarchical Clustering Algorithm on the dataset of flowers

We implemented Hierarchical algorithm to group the flower species into a certain number of clusters, and then performed verification in order to validate the results (Fig. 3.10 shows the Hierarchical code on dataset of flowers).

```
In [1]: from sklearn.cluster import KMeans
        import matplotlib.pyplot as plt
        import seaborn as sns
        import pandas as pd

        # Load data
        data = pd.read_excel('C:/Users/HI-TECH/Downloads/Nouveau dossier (2)/nouveau 3/datasetsf.xlsx')

        # Select feature columns (excluding the 'Nom' column)
        X = data.drop(columns=['Nom'])

        # Apply hierarchy Clustering
        from scipy.cluster.hierarchy import dendrogram, linkage
        from scipy.cluster.hierarchy import fcluster

        # Perform hierarchical clustering
        linked = linkage(X, method='ward')
        data['Cluster_Hierarchical'] = fcluster(linked, t=3, criterion='maxclust')

        # Plot the dendrogram
        plt.figure(figsize=(10, 7))
        dendrogram(linked,
                   orientation='top',
                   labels=data['Nom'].values,
                   distance_sort='descending',
                   show_leaf_counts=True)
        plt.title('Hierarchical Clustering Dendrogram')
        plt.show()
```

**Fig. 3.10**: Hierarchical code on the flowers dataset.

After implementing the Hierarchical algorithm, on flowers we obtained dendrogram , where each cluster represents flowers with common characteristics (Fig. 3.11)

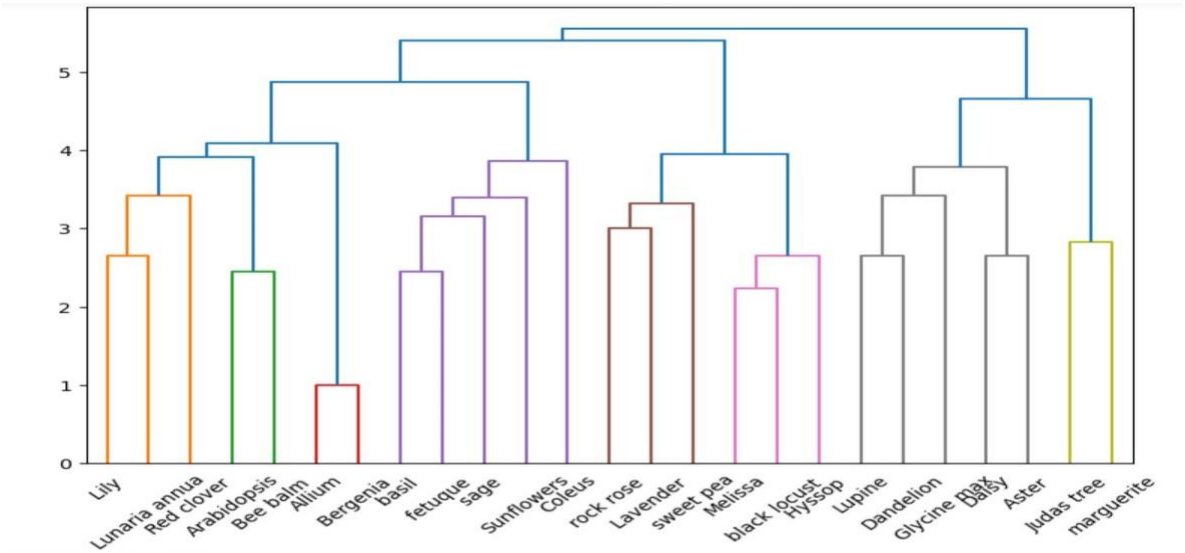



**Fig. 3.11:** Clusters of the flowers using Hierarchical

# 8. Discussion

Forests are complex and heterogeneous ecosystems, exhibiting a diverse range of characteristics across different spatial scales. Understanding this heterogeneity is crucial for effective forest management, conservation efforts, and ecosystem monitoring. One approach to characterizing forest landscapes is through the identification of homogeneous regions or clusters based on various environmental and biophysical attributes.

Clustering techniques, a branch of unsupervised machine learning, provide a powerful set of tools for identifying and delineating areas with similar forest characteristics. By grouping together spatially contiguous areas that share comparable environmental conditions, vegetation patterns, or structural properties, clustering algorithms can reveal the underlying structure and composition of forest landscapes.

In this chapter, we presented the application of clustering methods to the identification of homogeneous forest landscapes. To determine why plants or flowers might be grouped together in a K-means clustering and Hierarchical algorithm , we can consider several common characteristics that could contribute to their similarity. Here are some potential common characteristics:

1- **Mediterranean Origin and Climate Preference:** plants and flowers native to the Mediterranean region and thrive in similar climates with hot, dry summers and mild, wet winters. However, some flowers often thrive in similar climates when grown in Mediterranean regions or adapted landscapes, while more widely adaptable, often flourish in Mediterranean climates, particularly those varieties that prefer well-drained soils and partial shade.

2- **Drought Tolerance:** some plants belong to the same cluster because they are well-known for their drought tolerance, making them suitable for arid and semi-arid environments. Certain varieties also show resilience to dry conditions once established, especially in the Mediterranean or similar regions where they are cultivated.

3- **Aromatic Properties:** aromatic herbs widely used in cooking, perfumery, and traditional medicine due to their fragrant oils or for their pleasant scent in landscaping.

4- **Evergreen Nature:** Certain plants and trees retain their foliage throughout the year, while others are evergreen shrub, maintaining its leaves year-round. Moreover, some

plants and flowers are typically herbaceous and often deciduous in colder climates, they can be evergreen in milder climates.

5- **Growth Habit and Maintenance:** It can be seen that some plants in the same cluster are often used in landscaping for their aesthetic appeal and relatively low maintenance once established, while certain varieties are similarly valued in gardens for their beauty, fragrance, and ease of growth.

When considering these characteristics, it's plausible that in a clustering algorithm like K-means, these plants might be grouped together due to similarities in climate adaptability, drought tolerance, evergreen nature, aromatic qualities, and their growth habit and maintenance.

## Conclusion

After collecting the datasets and applying k means algorithm and Hierarchical algorithm , the effort and dedication are evident in the amazing results. With these technologies, we succeed in analyzing data efficiently and accurately, allowing us to deeply understand the information and derive smart decisions. Knowing that the datasets is the primary storehouse of this data, and that the algorithms of clustering are the accounting mind that extracts value.

# General Conclusion

In this thesis, we propose an approach based on the use of an unsupervised classification technique to identify homogeneous forest landscapes. Our model is evaluated using data collected from the Djebel Messaad forest in M'sila Province Algeria.

In the first chapter, we defined forest landscapes and mentioned their types, parameters, advantages, some of the services they provide, the challenges they face, and some solutions to reduce these challenges.

The second chapter focused on clustering, including its definition and algorithms (k-means, Hierarchical, DBSCAN, OPTICS, Mean shift, Affinity Propagation). For each algorithm, we mentioned its definition, its method, its algorithm, its types, and an example, in addition to its advantages and disadvantages.

In the third chapter, we talked about how to collect information from various sources to collect a datasets and then apply the k-means algorithm to it. After applying K-means and Hierarchical clustering to identify landscapes in forests, the algorithms have effectively identified distinct landscape types characterized by clusters of plants and trees with similar attributes, such as species composition, canopy density, and environmental conditions.

The results highlighted the significant influence of environmental factors on plant species distribution, with soil type, moisture levels, elevation, and sunlight exposure playing key roles in defining each landscape cluster. These clusters provide a basis for targeted forest management and conservation strategies. By understanding the specific needs and vulnerabilities of each landscape type, forest managers can implement tailored practices to protect and enhance the ecological balance within each cluster.

In conclusion, K-means clustering and Hierarchical algorithm have proven to be a powerful tool in identifying and understanding the diverse landscapes within forests. The results provide a comprehensive framework for enhancing forest management, conservation efforts, and ecological research, ultimately contributing to the sustainable stewardship of forest ecosystems.

Overall, our research highlighted the importance of data collection and analysis and the potential of using machine learning algorithms to understand environmental changes and their impacts, and to provide scientific, data-based methodologies to better protect and manage forest spaces. Therefore, it has become evident that leveraging these technologies is not merely a complementary measure but a necessity to enhance performance and attain remarkable outcomes.

# REFERENCES


[1] Myers, Norman, Mittermeier, R. A, Mittermeier, C. G, G. A, B. d. Fonseca et J. Kent, «biodiversity hotspots for conservation priorities,» *nature*, vol. 403, pp. 853-858, 2000.

[2] C. RL, «Beyond deforestation: restoring forests and ecosystem services on degraded lands,» *Science* , vol. 320, pp. 1458-1460, 2008.

[3] K. Singh, «Last.Frontier.Forests,» 1997.

[4] «(n.d.). Boreal Forest.,» [En ligne]. Available: https://www.thecanadianencyclopedia.ca/en/article/boreal-forest/ [8] WW. [Accès le 29 05 2024].

[5] N. N. G. S. (. T. F. In, «Encyclopedia.,» *Geographic Society*, 2003.

[6] (2015), UNECE et FAO, «Forest Types and Classification. In Manual on Definitions and Classifications of Forests (p. 9-20).,» pp. 9-20.

[7] Wetlands, UNESCO, (n.d.), Associated et M. Forests, «UNESCO. (n.d.). Montane Forests and Associated Wetlands,» *Sources of Knowledge*, p. p. 78.

[8] W. M. E, A et S. a. Bruce, Principles of Conservation Biology.

[9] «GPFLR: Bridging gaps, growing solutions,» cite de AFR100, 2023.

[10] «The potential of forest-based carbon offset projects in South Korea: A landowner survey,» *Forest Policy and Economics.*, n° %1106, pp. 77-84, 2019.

[11] Assessment, Ecosystem et Millennium, «Ecosystems and Human Well-being: Synthesis,» Island Press, 2005.

[12] Costanza, d. R, d. G. R, F. S. R, Grasso, H. B. M, K. Limburg, S. Naeem, R. V. O'Neill, J. Paruelo, R. G. Raskin et S. P. &. v. d. B. M., «The value of the world's ecosystem services and natural capital,» *Nature*, vol. 387, n° %16630, p. 253–260, 1997.

[13] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanov et A. .. &. T. J. R. Tyukavin, «High-resolution global maps of 21st-century forest cover change,» *Science*, pp. 850-853, 2013.

[14] G. B. Bonan, «Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests,» *Science*, pp. 1444-1449, 2008.

[15] V. Nijman, «An overview of international wildlife trade from Southeast Asia,» *Biodiversity and Conservation*, pp. 1101-1114, 2010.

[16] J. A. Foley, R. DeFries, G. P. Asner, C. B. G. Barford, S. R. Carpenter et P. K. ... & Snyder, «Global consequences of land use,» *Science*, pp. 570-574, 2005.

[17] Organization, Agriculture et F. (Food, «"Sustainable agricultural development for food security and nutrition: what roles for livestock,» 2013.

[18] M. A. Wulder, J. C. White, T. R. Loveland, C. E. Woodcock, A. S. Belward, W. B. Cohen et J. G. ... & Masek, «The global Landsat archive: Status, consolidation, and direction.",» *Remote Sensing of Environment*, pp. 271-283, 2016.

[19] A. S. Rodrigues, S. J. Andelman, M. I. Bakarr, L. Boitani, T. M. Brooks, R. M. Cowling et T. H. ... & Ricketts, «Effectiveness of the global protected area network in representing species diversity.",» *Nature*, pp. 640-643, 2004.

[20] Bishop et C. M., «Models and Expectation-Maximization mixture,» chez *Pattern Recognition and Machine Learning*, 2006.

[21] Prathmesh, «introduction to clusyering,» chez *Book of Machine learning & IA*.

[22] Nazari, Zahra, Kang, Dongshik, Asharif, M. Reza, Sung, Yulwan, Ogawa et Seiji, «A new hierarchical clustering algorithm,» *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pp. 148-152, 2015.

[23] Ester, Martin, J. Hans-Peter Kriegel, g. Sander et X. Xu, «Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 Munich, Germany,» *Research Paper.*

[24] «GeeksforGeeks,» [En ligne]. Available: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering. [Accès le 23 05 2023].

[25] M. Ester, H. P. Kriegel, J. Sander et X. Xu, «A density-based algorithm for discovering clusters in large

spatial databases with noise,» *KDD (Knowledge Discovery in Databases)*, Vols. %1 sur %296,, n° %134, pp. 226-231, 1996.

[26] J. Han et J. &. K. M. Pei, «Disadvantages of DBSCAN Algorithm,» chez *Data Mining: Concepts and Techniques*, Elsevier, Éd., 2011.

[27] Ankerst, Mihael, Breunig, M. M, Kriegel, Hans-Peter, Sander et Jörg, «OPTICS: Ordering points to identify the clustering structure,» *ACM Sigmod Record*, vol. 28, n° %12, pp. 49-60, 1999.

[28] R. J. Campello, D. Moulavi et J. Sander, «Density-Based Clustering Based on Hierarchical Density Estimates,» 2013.

[29] D. Birant et A. Kut, «ST-DBSCAN: An algorithm for clustering spatial–temporal data,» *Data & Knowledge Engineering*, vol. 60, n° %11, pp. 208-221.

[30] Aggarwal, C. C et C. K. Reddy, «Data clustering: Algorithms and applications,» CRC Press, 2013.

[31] Y. Cheng et A. Mean Shift, «A robust approach toward feature space analysis,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, n° %18, pp. 790-799, 1995.

[32] prutor.ai, «CLUSTERING ALGORITHMS – MEAN SHIFT ALGORITHM,» [En ligne]. Available: https://prutor.ai/clustering-algorithms-mean-shift-algorithm/. [Accès le 1 6 2024].

[33] B. J. Frey et D. Dueck, «Clustering by passing messages between data points,» *Science*, vol. 315, n° %15814, pp. 972-976, 2007.

[34] A. Gottlieb et R. I. Brafman, «A Generalized Affinity Propagation Algorithm for Clustering Data,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, n° %13, pp. 649-662, 2019.

[35] A. Rodriguez et A. Laio, «Clustering by fast search and find of density peaks,» *Science*, vol. 344, n° %16191, pp. 1492-1496, 2014.

[36] L. Kaufman et P. J. Rousseeuw, chez *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, 1990.

[37] M. B. Eisen, P. T. Spellman, P. O. Brown et D. Botstein, «Cluster analysis and display of genome-wide expression patterns,» *Proceedings of the National Academy of Sciences*, vol. 95, n° %125, pp. 14863-14868, 1998.

[38] A. K. Jain, M. N. Murty et P. J. Flynn, «Data clustering: a review,» *ACM computing surveys (CSUR)*, vol. 31, n° %13, pp. 264-323, 1999.

[39] V. Chandola, A. Banerjee et V. Kumar, «Anomaly detection: A survey,» *ACM computing surveys (CSUR)*, vol. 41, n° %13, p. 15, 2009.

[40] C. D. Manning, P. Raghavan et H. Schütze, chez *Introduction to information retrieval*, Cambridge University Press, 2008.

[41] A. Jain, «Data clustering: 50 years beyond K-means,» *Pattern Recognition Letters*, 2010.

[42] Vassilvitskii et D. A. S., *k-means++: The Advantages of Careful Seeding*, 2007.

[43] C. A. Reddy et C.K., «Outlier Detection: A Survey,» *ACM Computing Surveys (CSUR)*, 2013.

[44] D. Sculley, «Web-scale k-means clustering,» *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[45] M. Halkidi, Y. Batistakis et M. Vazirgiannis, «On Clustering Validation Techniques,» *Journal of Intelligent Information Systems*, 2001.

[46] «What is Python?,» [En ligne]. Available: https://www.python.org/doc/essays/blurb/. [Accès le 02 06 2024].

[47] «The Hub for Data Science and AI Collaboration,» [En ligne]. Available: https://www.anaconda.com/products. [Accès le 02 06 2024].

[48] «TensorFlow,» [En ligne]. Available: https://www.tensorflow.org/about/bib#large-scale_machine_learning_on_heterogeneous_distributed_systems. [Accès le 02 06 2024].

[49] «Scikit-Learn,» 02 08 2019. [En ligne]. Available: https://www.techopedia.com/definition/33860/scikit-learn. [Accès le 02 06 2024].

[50] McKinney, «Pandas Introduction,» 2008. [En ligne]. Available: https://www.w3schools.com/python/pandas/pandas_intro.asp. [Accès le 03 06 2024].

**ملخص**

هذه الدراسة تركز على جمع المعلومات عن غابة جبل مساعد من مختلف المصادر وإدراجها في قاعدة بيانات. يُلاحظ أهمية التحليل الفعّال لتلك البيانات لتحقيق أهداف محددة. تُعتبر قاعدة البيانات مصدرًا هامًا للاستفادة من البيانات المتاحة، K-ويتطلب ذلك استخدام تقنيات التحليل المناسبة لاستخراج رؤى قيمة. تم استكشاف في هذه المذكرة استخدام خوارزمية و خوارزمية التجميع الهرمي كأدوات رئيسية لتحليل البيانات. يتمثل هدف تطبيق خوارزميات التجميع في Means تجميع البيانات إلى مجموعات تتميز بأقصى تشابه بين البيانات النباتية والأزهار داخل كل مجموعة، وأقصى اختلاف بين المجموعات المختلفة. من خلال تحليل البيانات باستخدام هذه الخوارزميات، تمكنا من الوصول إلى نتائج مرضية تساهم في فهم أفضل للبيانات وتحقيق الأهداف المحددة.
**الكلمات المفتاحية** : منظر غابي، نظام بيئي غابي، تجميع، استخراج البيانات، تحليل البيانات

**Abstract**

This study focuses on collecting information about Djebel messaad Forest from various sources and integrating it into a database. The importance of effectively analyzing this data to achieve specific objectives is noted. The database is considered a crucial source for leveraging available data, requiring the use of appropriate analytical techniques to extract valuable insights. The memorandum explores the use of K-Means clustering and hierarchical algorithms as primary tools for data analysis. The goal of applying clustering algorithms is to group data into clusters characterized by maximum similarity within plant and flower data in each cluster, and maximum dissimilarity between different clusters. Through the analysis of data using these algorithms, we were able to achieve satisfactory results that contribute to a better understanding of the data and the attainment of specific objectives.
**Key words**: Forest landscape, forest ecosystem, clustering, data mining, data analysis.

**Résumé**

Cette étude se concentre sur la collecte d'informations sur la forêt du Djebel messad à partir de diverses sources et leur intégration dans une base de données. L'importance de l'analyse efficace de ces données pour atteindre des objectifs spécifiques est soulignée. La base de données est considérée comme une source cruciale pour exploiter les données disponibles, nécessitant l'utilisation de techniques d'analyse appropriées pour extraire des informations précieuses. Le mémorandum explore l'utilisation des algorithmes de regroupement K-Means et hiérarchique comme outils principaux pour l'analyse des données. L'objectif de l'application

de ces algorithmes de regroupement est de regrouper les données en clusters caractérisés par une similarité maximale au sein des données végétales et florales de chaque cluster, et une dissimilarité maximale entre les différents clusters. Grâce à l'analyse de données à l'aide de ces algorithmes, nous avons pu obtenir des résultats satisfaisants qui contribuent à une meilleure compréhension des données et à la réalisation des objectifs spécifiques.

**Les mots-clés :** Paysage forestier, écosystème forestier, regroupement, exploration de données, analyse de données.