CP 318: Data Science for Smart City Applications
Project 2:Analysis of Land Use Classification in India

Team: Explorers
Authors: Alok Mishra (21401), Sushil Jangra (21242), Sital Kumar Sahu (21764)
Submitted to: Prof Punit Rathore, Assistant Professor (IISc)

1 Introduction

1.1 Motivation

The project aims to offer valuable insights for decision-making in agriculture, environmental conservation, and land management. The focus is on understanding how to use land more efficiently to meet the demands of a growing population while maintaining ecological balance. The analysis is designed to be a practical resource for policymakers, researchers, and stakeholders, providing evidence-based recommendations to enhance agricultural practices and land stewardship. Ultimately, the goal is to contribute to a more informed and adaptive approach to sustainable development in the face of increasing global challenges. Weak land use policies can cause environmental degradation, habitat loss, urban sprawl, strained infrastructure, climate impact, social inequalities, natural disasters, agricultural loss, and economic decline. Effective policies are crucial for sustainable development and community well-being. The analysis can be useful for providing what kind of land use policies different states should adapt.

1.2 Problem Statement

The **NDAP** dataset, with 4677 rows and 16 columns, provides detailed insights into India's land use statistics, including country, state, district, and time. It includes identifiers, indicators, and time columns, enabling a comprehensive understanding of land distribution, utilization, and environmental implications. By employing data processing, descriptive analysis, clustering, and dimension reduction techniques, the project aims to provide actionable insights for policymakers.
Novelty: Examining existing analyses and noting the distinctive features of our dataset is crucial. It is worth emphasizing that the literature pertaining to our specific dataset is notably scarce. The absence of a substantial body of prior research introduces a level of uniqueness to our study, providing an opportunity to offer fresh insights and perspectives that can enrich the current knowledge landscape.

2 Analysis

2.1 Data Processing

Data preprocessing is a crucial step in the data analysis pipeline that encompasses a series of procedures aimed at refining raw datasets to enhance their quality and suitability for analysis. Here we performed multiple tasks such as meticulous **removal of missing values**, meticulous **elimination of duplicate entries**, robust detection and **removal of outliers**, and judicious **scaling(standard scaling) of data** to ensure a more uniform and meaningful interpretation of results. The overarching goal of data preprocessing is to create a well-structured and reliable dataset, setting the stage for more accurate and insightful analyses.

Outlier Detection Techniques:

We employed various outlier detection techniques like IQR (Interquartile Range), z-score, visual inspection etc, among which the Z-score method yielded the most promising results for our specific dataset. The Z-score is a statistical measure that quantifies how far a data point is from the mean of a group of data. In our implementation, we considered Z-scores for each feature, calculated as the absolute difference between each data point and the mean of its respective feature, normalized by the standard deviation.

Interpretation and Considerations:

The Z-score threshold was set to 3, which indicates that any data point with a Z-score beyond this threshold was considered an outlier. The choice of this threshold may vary based on the characteristics of the dataset and the desired level of outlier sensitivity.

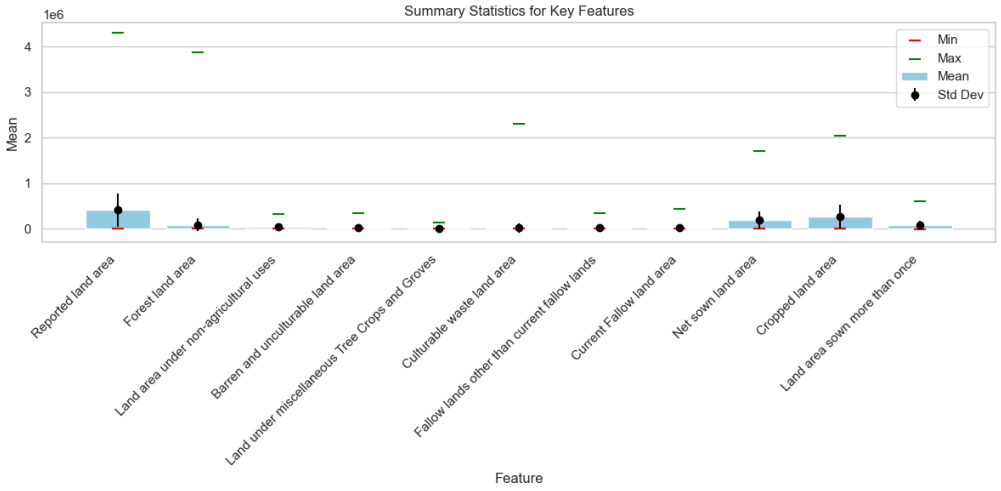
It's crucial to note that the outlier removal process resulted in a more refined dataset (4213, 16), which is expected to enhance the quality and reliability of subsequent analyses.

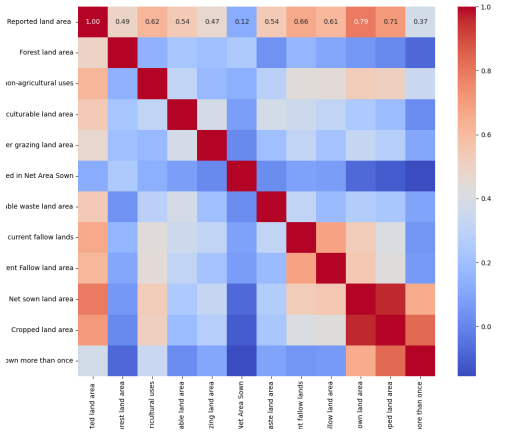
Process	Total Instances	Data Shape (Rows, Columns)
Initial Shape	**4677**	**(4677, 16)**
Missing Values Removal	**203**	**(4474, 16)**
Duplicate Values Removal	**09**	**(4465, 16)**
Outlier Detection and Removal	**252**	**(4213, 16)**
Final Shape	**4213**	**(4213, 16)**

Table 1: Data Preprocessing Summary

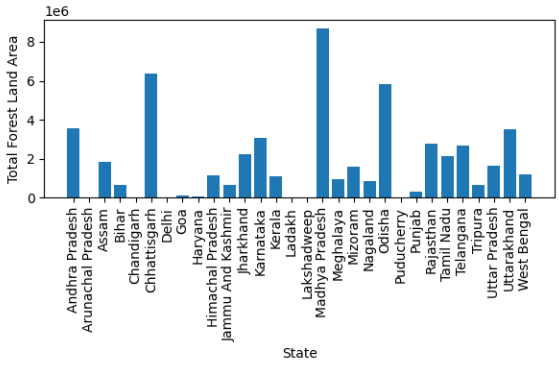
2.2 Descriptive Analysis and Exploratory Data Analysis (EDA)

Through Exploratory Data Analysis (EDA), we gained insights into the range and extremes of various features within our dataset(Summary Statistics). We visualized the variability in values across different states for these features, exemplified by the illustration below depicting how the forest area changes with each state. Additionally, we generated a correlation heatmap among features. During subsequent analysis, special attention was given to highly correlated features to ensure the selection of only one representative feature for multiple highly correlated features.**(Note: area in hectares)**





(a) Heatmap of Correlation among features



(b) statewise forest area

Figure 1: Summary statistics, Heatmap for correlation and variability of forest land with states

2.3 Clustering Analysis

In our analysis, we strategically applied **KMeans** and **Gaussian Mixture Model (GMM)** clustering algorithms, tailoring feature combinations to our specific objectives. The ensuing presentation offers a comprehensive exploration of results, enabling insightful comparisons between the outcomes of both algorithms. Various combination of features were used for clustering, e.g. "net sown area" and "land area sown more than once". To determine optimal clusters, the **Elbow curve and Akaike Information Criterion (AIC) score** were employed for KMeans and GMM, respectively, streamlining our analysis. The meticulous selection of algorithms, features, and cluster determination methodologies collectively fortified the reliability and depth of our analytical findings, contributing to a more nuanced understanding of the underlying patterns.

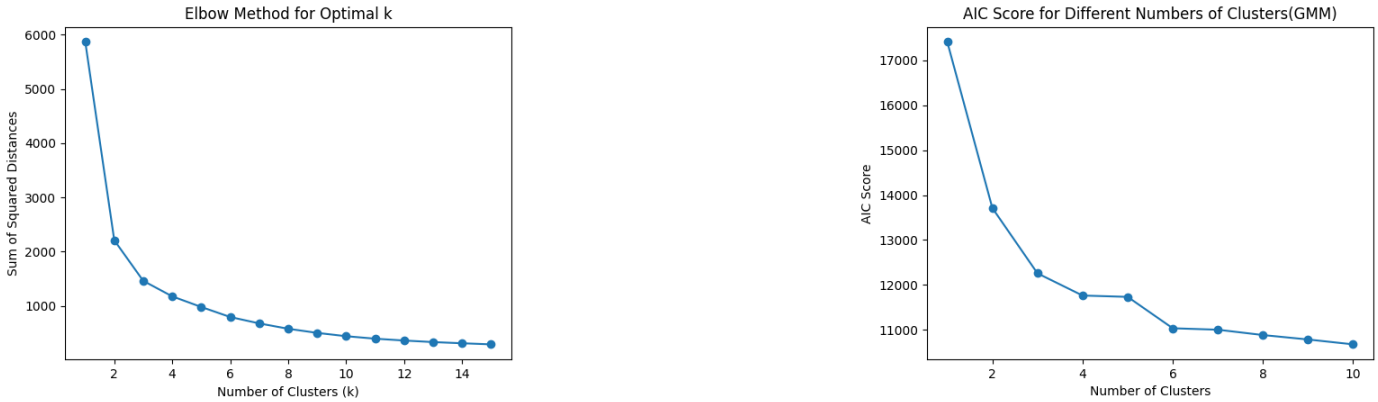


Figure 2: Comparison of Clustering Methods: Elbow Method for K-means vs. AIC Score for Gaussian Mixture Model (GMM)

2.3.1 Clustering results:

Based upon Elbow method and AIC Score four clusters were chosen (as shown below). Comparisons of both the algorithms was done based upon time and accuracy **(Silhouette Score)**. It was found that performance wise both the algorithms were almost similar. Further using the obtained clusters, we identified the districts associated with each cluster. This information was then utilized to determine the cluster affiliation of each state. Implementing similar land use policies for states within the same cluster can be a strategic approach.

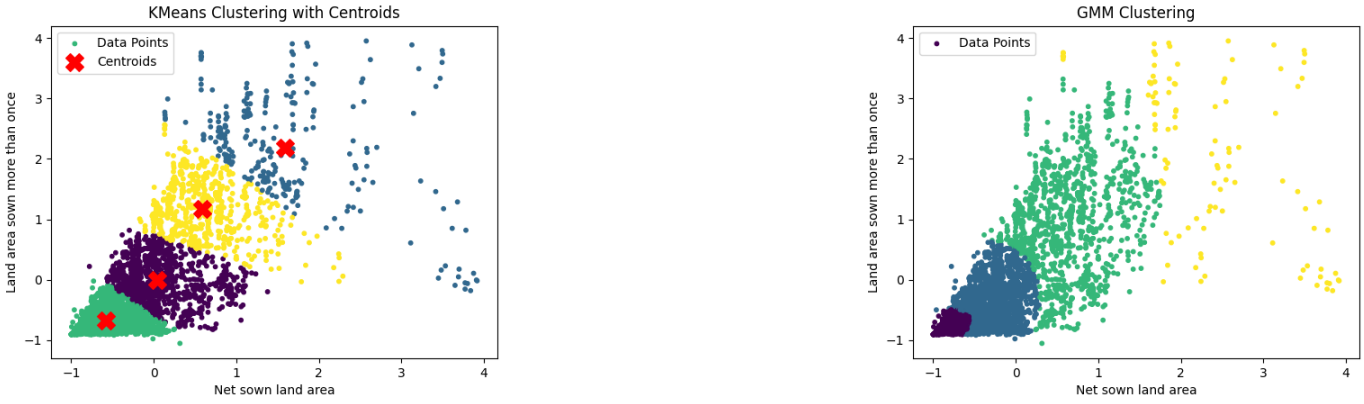
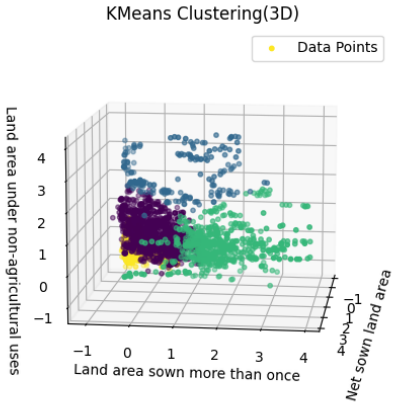


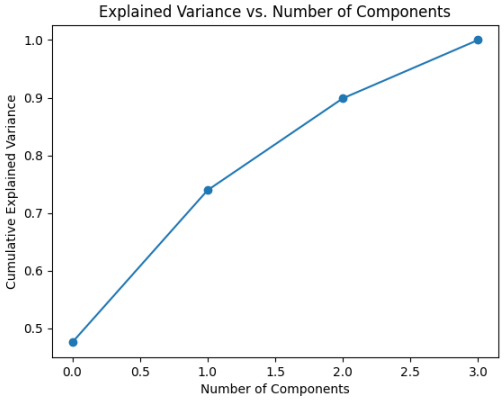
Figure 3: Comparison of Clustering Algorithms: K-means vs. Gaussian Mixture Model (GMM)

3D: The similar kind of analysis was performed using various combinations of features, including more than two at a time. One such is shown here:

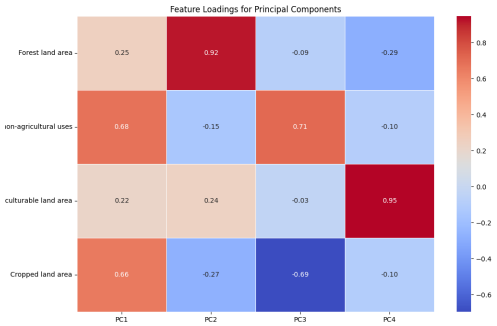


2.4 Dimension Reduction

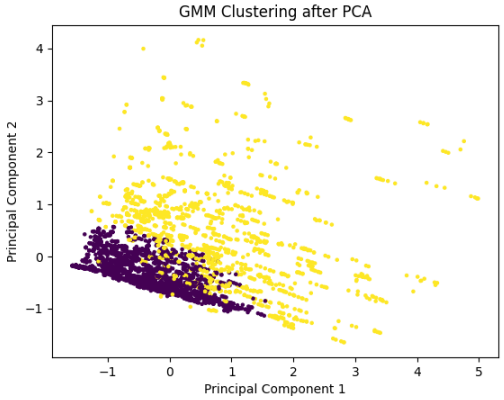
4 features of interest ['Forest land area','Land area under non-agricultural uses','Barren and unculturable land area','Cropped land area'] were taken at once.
The utility of **Linear Dimensional Reduction Technique(PCA)** in our analysis (including visualizing information for more than 3 features at once) is evident. However **non-linear dimensional reduction technique(t-SNE)** was not much helpful(shown below) as preserving overall variance and capturing only linear relationships was needed here and there were no non-linear patterns. Initially, we focused on four features of interest for analysis. However, due to challenges in visualization and analysis, we applied PCA. Consequently, we discovered that a significant portion of the information could be condensed into 2 principal components **(> 90 % of data variance)**, facilitating a more accessible and insightful analysis, as illustrated below(GMM applied after PCA).

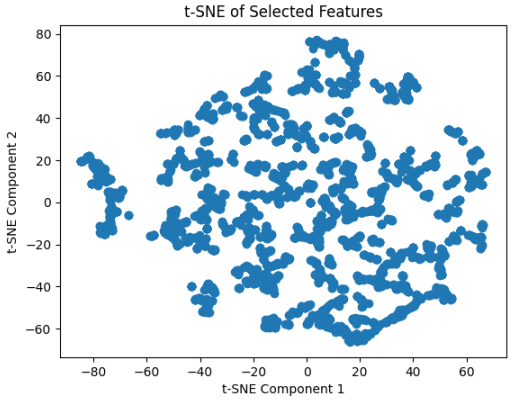


(a) Explained variance with number of components for PCA



(b) Feature loading for principal components





3 Discussion on Results

Using the obtained clusters, we identified the districts associated with each cluster. This information was then utilized to determine the cluster affiliation of each state. Implementing similar land use policies for states within the same cluster can be a strategic approach.

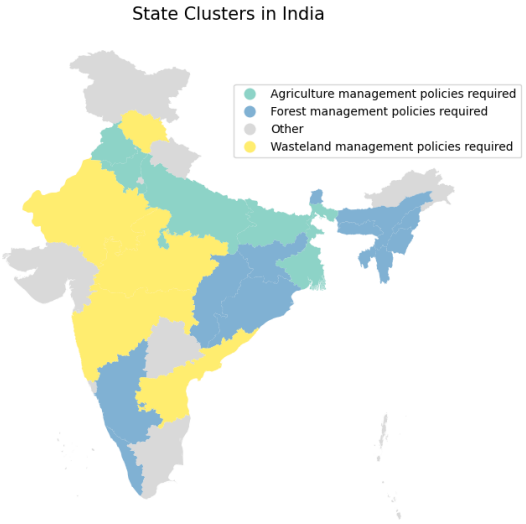


Figure 6: India's map depicting the types of policies that various states should adopt.

4 Conclusion and Self-Reflection

This project has been an instructive journey, imparting valuable lessons about the intricacies inherent in real-world datasets. It has underscored the importance of meticulous preprocessing and has significantly enhanced our ability to tackle challenges in model selection. The collaborative nature of our efforts has emphasized the pivotal role of effective communication and teamwork, particularly in the realms of exploratory data analysis and unsupervised learning.

Beyond the educational aspects, our analysis yields substantial insights into land use patterns, presenting valuable information for the Indian government and society. This knowledge contributes to sustainable agricultural development, optimal land management, and well-informed decision-making. The overarching impact extends to fostering food security and promoting environmental sustainability.

Throughout the analysis, we grappled with challenges such as missing values, inconsistencies, and outliers. These complexities necessitated the application of meticulous data preprocessing and outlier detection techniques. Looking forward, future iterations could benefit from an even more comprehensive dataset, allowing for a deeper and more nuanced exploration of the underlying patterns and trends.

5 Sources and References

The dataset used for this analysis was obtained from The National Data and Analytics Platform (NDAP) available at `https://ndap.niti.gov.in/dataset/6795?tab=data&filter_id=1924`.NDAP serves as a valuable resource for accessing diverse datasets relevant to India's socio-economic landscape. Additionally, course lecture slides and workbooks provided valuable guidance in understanding and applying the analytical techniques employed in this project.