Trip Recommendation System

Our Team Members

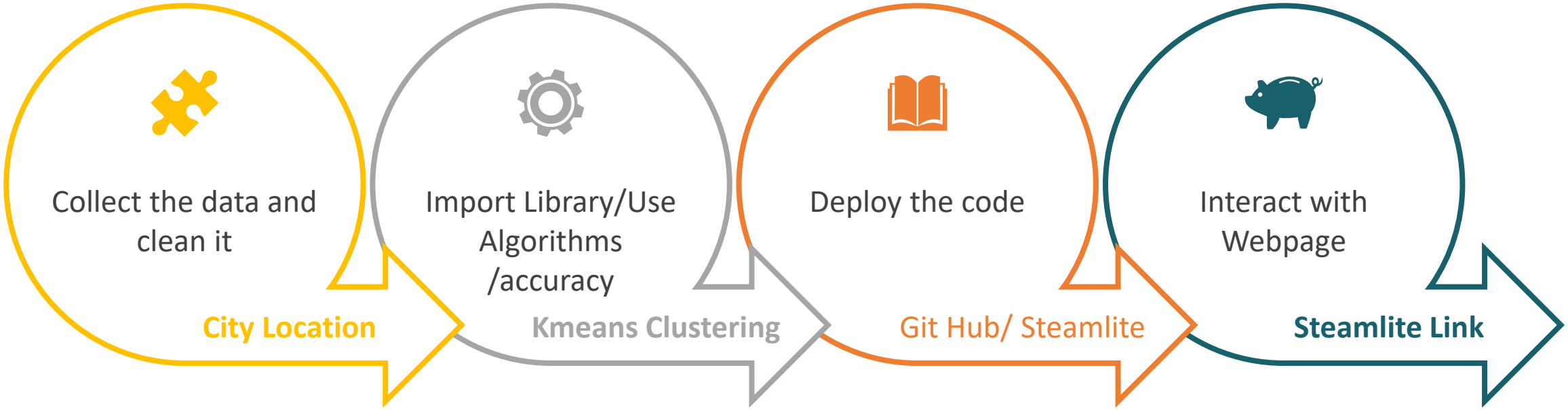
- Rushikesh Kharode (211090053)
- Vitthal Chavan (211090056)
- Shanti Trivedi (221091910)

Introduction

- The recommendation project for locations aims to provide users with suggestions of similar locations based on their input. Leveraging Geocoding and Clustering Techniques, the project enhances the user experience by offering personalized recommendations tailored to their preferences or interests.
- Our goal is to improve recommendation effectiveness through personalized strategies, accurate geocoding, advanced clustering, and dynamic updates.

Flow of Project

- Key points of Project Flowchart



Data Collection:

1.Source Identification: Searched government databases, open data portals, and specialized platforms for location datasets.
2.Quality Assessment: Evaluate metadata, documentation, and sample data for completeness, accuracy, and relevance.
3.Selection Process: Prioritize datasets containing latitude and longitude coordinates or easily geocodable information.
4.Data Access and Preprocessing: Download datasets, clean, filter, and format them to ensure consistency and compatibility.
5.Integration and Validation: Geocode locations, integrate datasets, validate for accuracy, and document sources and methods for transparency.

Import Libraries:

1.Streamlit: Streamlit is a Python library used for building interactive web applications for machine learning and data science projects. It allows developers to create user-friendly interfaces with minimal code.
2.Pandas: Pandas is a powerful data manipulation library in Python used for data analysis and manipulation. It provides data structures and functions to efficiently work with structured data, such as data frames.
3.scikit-learn (sklearn): Scikit-learn is a machine learning library in Python that provides tools for data mining and data analysis. It includes various algorithms for clustering, classification, regression, and more.
4.KMeans: KMeans is a clustering algorithm implemented in scikit-learn. It is used for partitioning data into 'k' clusters based on similarity.

Algorithm Used

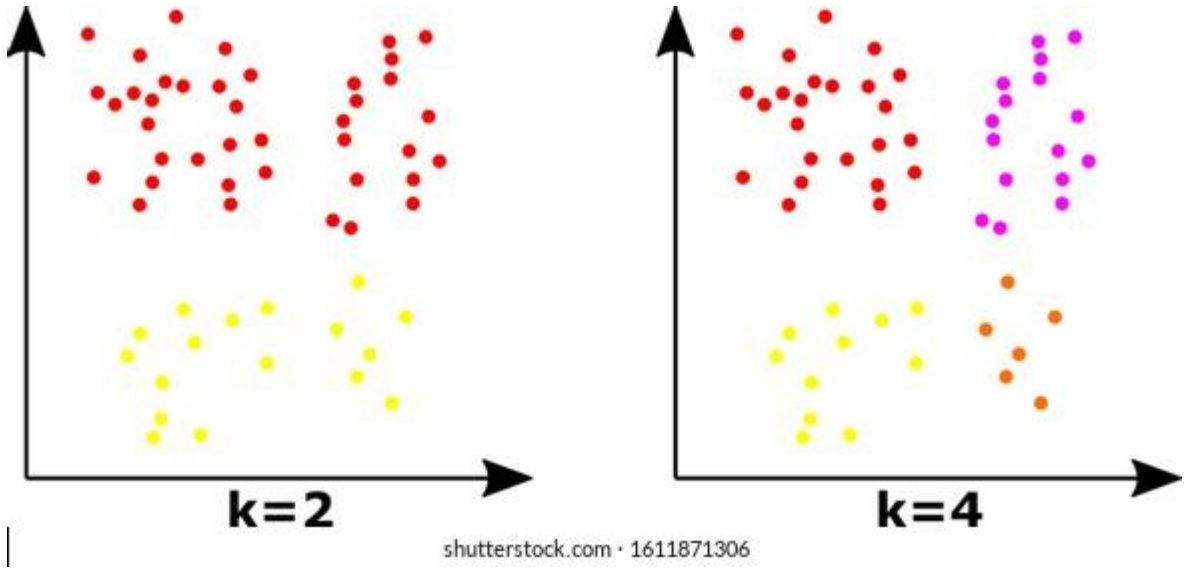
Clustering is a machine learning technique used to:
1.Group similar data points together based on their features or attributes.
2.Discover inherent structures or patterns in unlabeled datasets without prior knowledge of class labels.

K Means Algorithm : Unsupervised Learning

An iterative algorithm that divides the unlabeled dataset into k different clusters in a way that each element belongs to a single group that has similar properties.

N observations, k clusters

1: Select random K points i.e. centroids/means.
2: Assign each element/data point to their closest centroid, which will form the initial K clusters.
3: Calculate the variance and place a new centroid of each cluster.
4: Repeat step 2, which means reassign each element to the new closest centroid of each cluster.
Step-6: If any reassignment occurs, repeat from 3, else END.



Silhouette Coefficient procedure:

Assign Data Points to Clusters:
Apply a clustering algorithm to assign each data point to a cluster.
Calculate Distances: Compute the average distance:

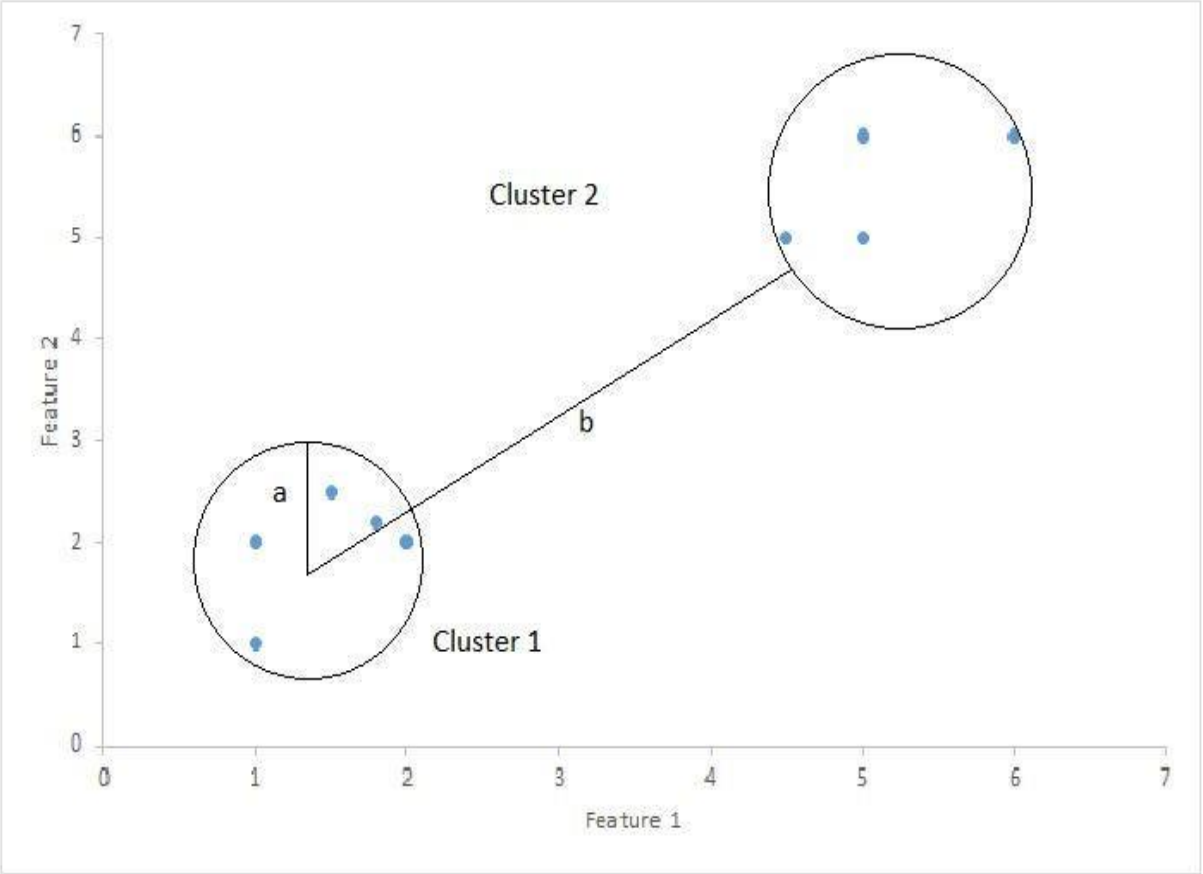
- aa: Average distance from a data point to other points within its cluster.
- bb: Average distance from the data point to points in the nearest neighboring cluster.

Compute Silhouette Coefficient: Calculate the Silhouette Coefficient for each data point using:

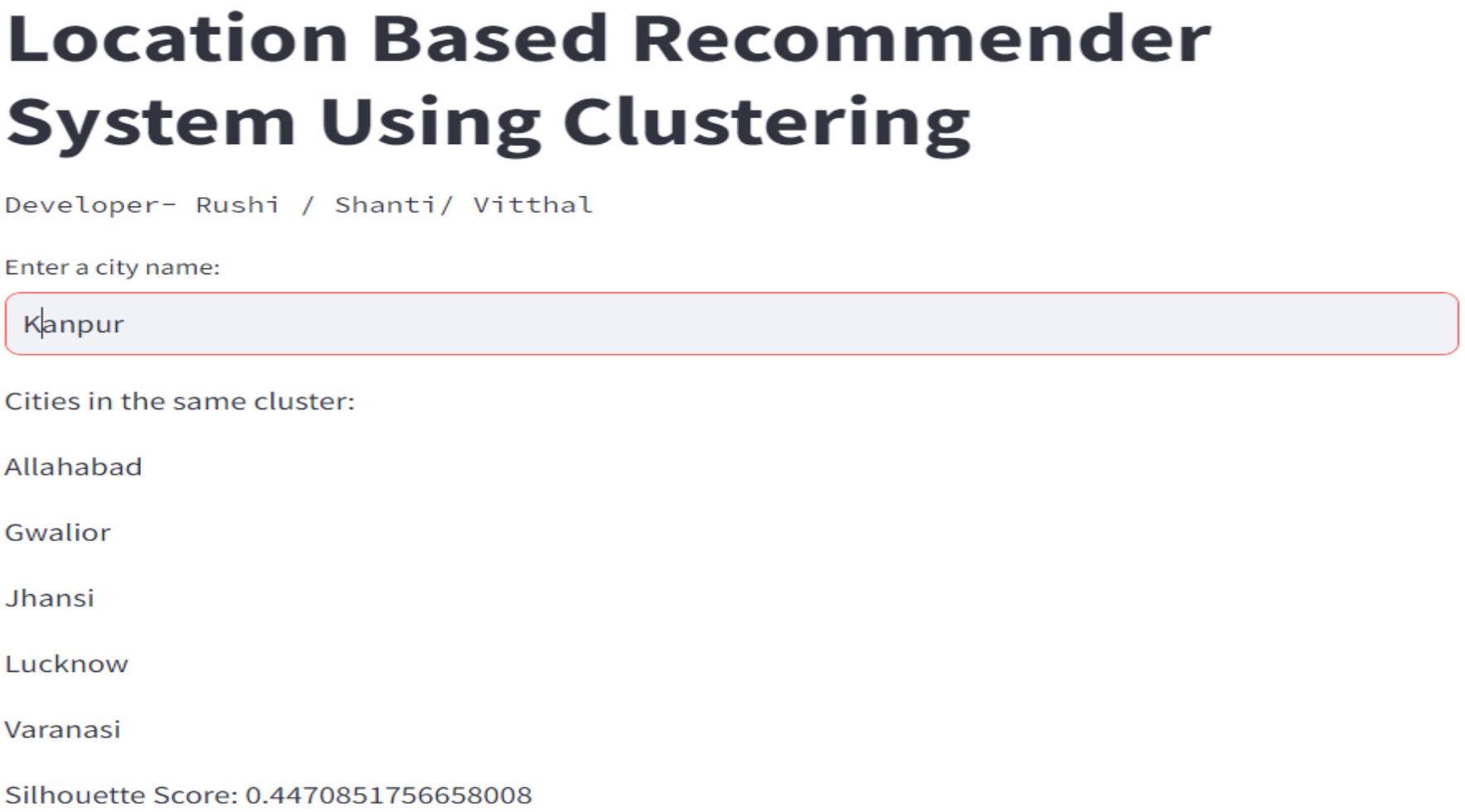
$$S = \frac{b-a}{\max(a,b)}$$

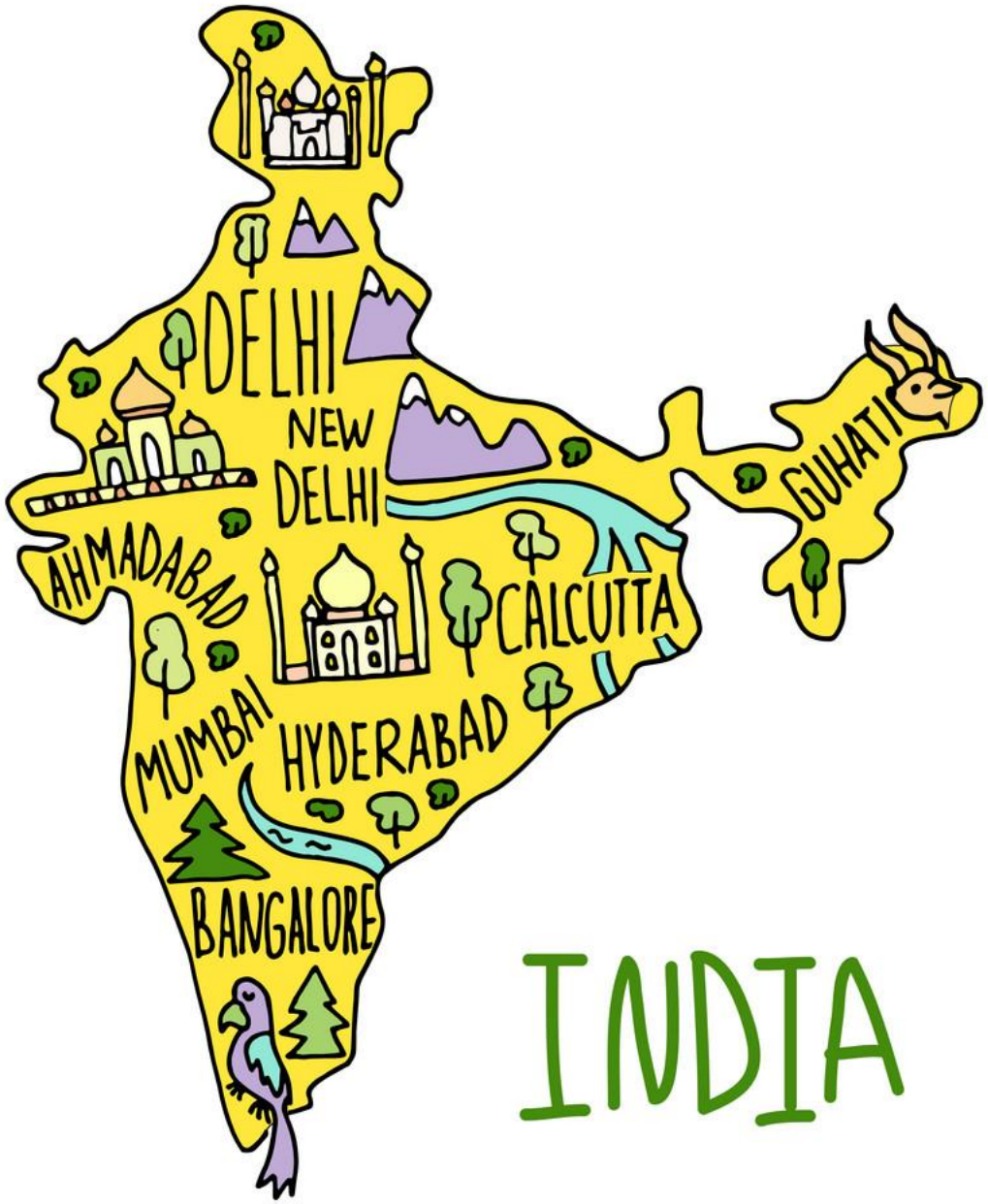
- Higher values (close to 1) indicate well-clustered data points.
- Lower values or negatives suggest misclassification.

1.Calculate Average Score: Compute the average Silhouette Coefficient across all data points for the overall clustering solution.



Web Page:


Location Based Recommender System Using Clustering
Developer- Rushi / Shanti/ Vitthal
Enter a city name:
Kanpur
Cities in the same cluster:
Allahabad
Gwalior
Jhansi
Lucknow
Varanasi
Silhouette Score: 0.4470851756658008



What we did?

Data Collection:
Gathered information for 100 cities in India, including their latitude and longitude coordinates.

Clustering with KMeans:
Utilized KMeans clustering algorithm to group cities based on their geographic coordinates.
Recommendation Generation: Developed a recommendation system where users input a city name, and the system suggests other cities within the same cluster as potential destinations.

Accuracy Measurement:
Evaluated the accuracy of the clustering algorithm using metrics like the Silhouette Score to assess the cohesion and separation of clusters.

Web Page display and final recommendation

Conclusion:

Effective Clustering Implementation:
The application successfully implements KMeans clustering to group cities based on their geographical coordinates, providing a foundation for a location-based recommender system.

User-Friendly Interface with Streamlit:
Streamlit's interactive interface allows users to easily input a city name and receive recommendations of similar cities, enhancing user experience and accessibility.

Accurate Recommendations:
By identifying the cluster of the input city, the application provides accurate and relevant city recommendations, helping users explore locations with similar geographical characteristics.

Quality Evaluation:
The inclusion of the Silhouette Score offers a measure of clustering quality, ensuring users receive reliable recommendations based on well-defined clusters.

Scope for Enhancements:
The application lays a strong groundwork with potential for future improvements, such as dynamic clustering adjustments, better error handling, and enhanced data visualizations to further enrich the user experience.

Thank You