ANALYSIS OF GEOLOCATIONAL DATA FOR ACCOMODATION USING K - MEANS CLUSTERING ALGORITHM

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

NAME	REG. NO.
HARIPRIYA R	201008025
LEENA N	201008042
MINI M	201008045
RITHIKA S	201008062

Project Guide

Dr. L. R. SUDHA

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ANNAMALAI UNIVERSITY

ANNAMALAI NAGAR – 608 002

MAY 2024

ANNAMALAI UNIVERSITY

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINERRING

This is to certify that the Project report titled **"ANALYIS OF GEOLOCATIONAL DATA FOR ACCOMODATION USING K-MEANS CLUSTERING ALGORITHM"** is the bonafide record of the work done by

NAME	REG. NO.
HARIPRIYA R	**201008025**
LEENA N	**201008042**
MINI M	**201008045**
RITHIKA S	**201008062**

in partial fulfillment of the requirements for the Degree of Bachelor of Engineering in Computer Science and Engineering during the period 2023 - 2024.

Dr. L. R. SUDHA, M.E., Ph.D.,
Associate Professor
Department of Computer Science and Engineering
Project Guide

Dr. R. BHAVANI, M.E., Ph.D.,
Professor & Head
Department of Computer Science and Engineering

Internal Examiner

External Examiner

Place: Annamalai Nagar
Date:

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to **Dr. R. BHAVANI, M.E., Ph.D., Professor and Head**, **Department of Computer Science and Engineering** for giving us the opportunity to undertake this project.

We express our deep sense of gratitude to our project guide **Dr. L. R. SUDHA, M.E., Ph.D.**, **Associate professor**, **Department of Computer science and Engineering** for her persistent interest in this project and her guidance and constant encouragement during this project without which the project would not have been materialized.

We would like to convey our heartiest thanks to our project coordinator **Dr. R. MADHANMOHAN, M.E., Ph.D., Associate Professor**, **Department of Computer science and Engineering** for all his help and support. He with his extreme patience has guided us in situation of need for which we are extremely grateful.

We would like to thank our **Review Committee Member, M. KALAISELVI GEETHA, M.E., Ph.D., Professor, Department of Computer science and Engineering** for her valuable suggestion and ideas throughout this project work.

We would like to thank all our friends for supporting and encouraging us. We wish to thank all the technical staff members, in-charge of our department laboratories for fulfilling all our project needs and offering us a timely help.

Above all we are indebted to our beloved parents, whose blessings and best wishes have gone a long way in making this final year project work a grand success.

HARIPRIYA R
LEENA N
MINI M
RITHIKA S

TABLE OF CONTENT

ABSTRACT

A significant source of knowledge about locations and local human behavior may be obtained by analyzing geolocational data. In the fast-paced and busy environment an average person migrates from one place to another place and it is difficult for that person to search and identify the best accommodation with his preferences. A person tries to locate himself in a place where his preferences and interests are more and closer to his previous livelihood. So, using data visualization and clustering processes, it is possible to identify amenity-rich locations within a particular radius by taking into account a variety of factors, including the overall availability of restaurants, department shops, gyms, and other facilities. K-Means Clustering is used to group geo-locational data that are applied on the geo-locational data obtained from Here Geocoding and Search API v7 (Application programming Interface) in order to find the best places to stay for a person in a specific city by categorizing accommodation for the people based on their preferences on amenities, location, and budget. This project will produce intelligent user suggestions by analyzing geo-locational data and user preferences to locate spot with rich, average, and low amenities and map those amenities.

Keywords: Machine Learning, Data Visualization, Data Cleaning, K-Means Clustering, FourSquare API, Geolocational Data, Exploratory Data Analysis.

ABSTRACT (TAMIL)

புவிஇருப்பிட தரவுகளை பகுப்பாய்வு செய்வதன் மூலம் இருப்பிடங்கள் மற்றும் உள்ளூர் மனித நடத்தை பற்றிய அறிவின் குறிப்பிடத்தக்க ஆதாரம் பெறப்படலாம். வேகமான மற்றும் பரபரப்பான சூழலில், ஒரு சராசரி நபர் ஒரு இடத்திலிருந்து மற்றொரு இடத்திற்கு இடம்பெயர்கிறார், அந்த நபர் தனது விருப்பங்களுடன் சிறந்த தங்குமிடத்தைத் தேடுவது மற்றும் அடையாளம் காண்பது கடினம். ஒரு நபர் தனது விருப்பங்களும் ஆர்வங்களும் தனது முந்தைய வாழ்வாதாரத்திற்கு நெருக்கமாக இருக்கும் இடத்தில் தன்னைக் கண்டுபிடிக்க முயற்சிக்கிறார். எனவே, தரவு காட்சிப்படுத்தல் மற்றும் கிளஸ்டரிங் செயல்முறைகளைப் பயன்படுத்தி, உணவகங்கள், டிபார்ட்மென்ட் கடைகள், ஜிம்கள் மற்றும் பிற வசதிகளின் ஒட்டுமொத்த கிடைக்கும் தன்மை உட்பட பல்வேறு காரணிகளை கணக்கில் எடுத்துக்கொள்வதன் மூலம் ஒரு குறிப்பிட்ட சுற்றளவில் வசதிகள் நிறைந்த இடங்களை அடையாளம் காண முடியும். K-Means Clustering என்பது, ஒரு குறிப்பிட்ட நகரத்தில் ஒருவர் தங்குவதற்கு சிறந்த இடங்களைக் கண்டறியும் பொருட்டு, Here Geocoding மற்றும் Search API v7 (Application programming Interface) இலிருந்து பெறப்பட்ட புவி-இருப்பிட தரவுகளில் பயன்படுத்தப்படும் புவி-இருப்பிடத் தரவைக் குழுவாக்கப் பயன்படுகிறது. வசதிகள், இருப்பிடம் மற்றும் வரவு செலவுத் திட்டம் ஆகியவற்றின் அடிப்படையில் மக்களுக்கான தங்குமிடங்களை வகைப்படுத்துவதன் மூலம். புவிசார்-இருப்பிட தரவு மற்றும் பயனர் விருப்பத்தேர்வுகளை பகுப்பாய்வு செய்வதன் மூலம், பணக்கார, சராசரி மற்றும் குறைந்த வசதிகள் உள்ள இடத்தைக் கண்டறிந்து, அந்த வசதிகளை வரைபடமாக்குவதற்கு இந்தத் திட்டம் புத்திசாலித்தனமான பயனர் பரிந்துரைகளை உருவாக்கும்.

LIST OF FIGURES

LIST OF ABBREVIATIONS

SI. NO.	ABBREVIATION	EXPANDED FORM
1	CSV	Comma Separated Value
2	GPS	Global Positioning System
3	ML	Machine Learning
4	API	Application Programming Interface
5	URL	Uniform Resource Locator
6	REST API	Representational State Transfer Application Programming Interface
7	HTTP	Hypertext Transfer Protocol
8	JSON	Javascript Object Notation
9	EDA	Exploratory Data Analysis
10	GPU	Graphic Processing Unit
11	WCSS	Within-Cluster-Sum-of-Square

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Analyzing geo-locational data enables research into places and regional human behavior. Those who travel frequently may find it difficult to locate the suitable area to reside. In 2021, India accounted for 1.57% of total international tourist visits. India welcomed 17.9 million more international visitors in 2020 compared to 2019, an increase of 3.5%. India is the eighth most visited country in the Asia-Pacific region and now holds the 22nd place worldwide. It would be challenging for them to find a location to stay and enjoy their vacation because India is attracting a lot of attention from tourists. Also, the people migrating to different place may find difficulties to locate an ideal place with their priorities and preferences.

We thus recommend which would be ideal for them based on the place they choose as well as their preferences for the area. Individuals who move to a new place will likely have particular preferences and considerations, therefore analysis of geo-location is used to pinpoint the optimal locations. The situation would be hassle-free and time-saving if the consumers lived close to their preferred locations. The methodology can be applied to any location of one's choosing, and can vary according to user preference. To make the data points in each group more comparable to one another than those in the other groups, the data points are only separated into a number of groups. In other words, the goal is to group data items based on the characteristics they have in common. K-Means clustering is the best clustering technique for grouping things depending on how similar they are. K-Means clustering is an algorithm for Unsupervised Learning, which clusters an unlabeled dataset into distinct groups. The variable K represents the number of clusters that the algorithm will create. For instance, if K is set to 2, then the algorithm will create two distinct groups. This process is used to group unlabeled data into different categories without any prior training. The algorithm takes in the unlabeled dataset as input, divides it into K clusters, and repeats this process until it finds the optimal cluster.

The burgeoning popularity of online platforms for booking accommodations has underscored the importance of effectively analyzing geospatial information to cater to the diverse preferences and requirements of travelers. This project delves into the intricate domain of geolocation analysis for accommodation, leveraging state-of-the-art clustering algorithms to unearth meaningful insights and patterns within the spatial distribution of accommodation options.

The primary objective of this project is to develop a comprehensive understanding of the geospatial dynamics inherent in the accommodation sector, thereby facilitating informed decision-making processes for both consumers and service providers. By harnessing the power of clustering algorithms, such as k-means, hierarchical clustering, and density-based spatial clustering of applications with noise (DBSCAN), this endeavor aims to partition the geographical landscape into distinct clusters, each representing a unique concentration of accommodation facilities. Through meticulous analysis and evaluation, these clusters will be scrutinized to identify spatial patterns, trends, and anomalies, thus offering invaluable insights into the underlying factors shaping the distribution of accommodation options.

Furthermore, this project endeavors to explore the efficacy of various clustering algorithms in delineating spatial clusters based on diverse criteria, including proximity to key landmarks, amenities, and transportation hubs, as well as price range and accommodation type. By employing a multifaceted approach to geolocation analysis, this study seeks to uncover nuanced insights that transcend conventional metrics, thereby enhancing the granularity and accuracy of accommodation recommendations and decision-support systems.

In addition to its practical implications for travelers seeking accommodation and service providers aiming to optimize their offerings, this project contributes to the broader landscape of geospatial analysis by showcasing the versatility and efficacy of clustering algorithms in extracting actionable insights from large-scale geolocation datasets. Through rigorous experimentation and validation, the findings of this study are poised to inform and enrich existing methodologies for geospatial analysis, paving the way for future advancements in the field of location-based services and spatial data analytics.

In essence, this project serves as a testament to the transformative potential of geolocation analysis and clustering algorithms in revolutionizing the accommodation sector, empowering stakeholders with the knowledge and tools needed to navigate and capitalize on the complex interplay between geography, consumer preferences, and market dynamics.

1.1.1 MACHINE LEARNING

Machine learning (ML) is a type of artificial intelligence (AI) focused on building computer systems that learn from data. The broad range of techniques as shown in Fig.1.1 encompasses software applications to improve their performance over time. Machine learning algorithms are trained to find relationships and patterns in data. They use historical data as input to make predictions, classify information, cluster data points, reduce dimensionality and even help generate new content, as demonstrated by new ML-fueled applications such as ChatGPT, Dall-E 2 and GitHub Copilot.

Machine learning is widely applicable across many industries. Recommendation engines, for example, are used by e-commerce, social media and news organizations to suggest content based on a customer's past behavior. Machine learning algorithms and machine vision are a critical component of self-driving cars, helping them navigate the roads safely. In healthcare, machine learning is used to diagnose and suggest treatment plans. Other common ML use cases include fraud detection, spam filtering, malware threat detection, predictive maintenance and business process automation.

While machine learning is a powerful tool for solving problems, improving business operations and automating tasks, it's also a complex and challenging technology, requiring deep expertise and significant resources. Choosing the right algorithm for a task calls for a strong grasp of mathematics and statistics. Training machine learning algorithms often involves large amounts of good quality data to produce accurate results. The results themselves can be difficult to understand -- particularly the outcomes produced by complex algorithms, such as the deep learning neural networks patterned after the human brain. And ML models can be costly to run and tune.

According to the "2023 AI and Machine Learning Research Report" from Rackspace Technology, 72% of companies surveyed said that AI and machine learning are part of their IT and business strategies, and 69% described AI/ML as the most important technology. Companies that have adopted it reported using it to improve existing processes (67%), predict business performance and industry trends (60%) and reduce risk (53%).

Supervised Learning: In supervised learning, the algorithm is trained on a labeled dataset. This means that for every input data point, there is a corresponding output label. The goal of supervised learning is to learn the mapping from input to output so that the algorithm can make predictions on new, unseen data.

Classification: When the output variable is a category, such as "spam" or "not spam" in email filtering.

Regression: When the output variable is a continuous value, such as predicting house prices based on features like square footage, number of bedrooms, etc.

Unsupervised Learning: In unsupervised learning, the algorithm is given a dataset without any labels. The algorithm tries to find patterns or intrinsic structures in the data on its own.

Clustering: Grouping similar data points together based on some similarity measure.
Dimensionality Reduction: Reducing the number of features or variables in the dataset while preserving its structure.

Anomaly Detection: Identifying data points that deviate from the norm or exhibit unusual behavior.

Fig. 1.1 Machine Learning Techniques

1.1.2 GEOLOCATIONAL ANALYSIS

Geolocational analysis stands at the forefront of understanding spatial patterns, trends, and relationships within the accommodation industry. By harnessing the power of geographic information systems (GIS), spatial databases, and advanced analytical techniques, researchers and practitioners can unlock valuable insights from geolocational data to inform decision-making, optimize resource allocation, and enhance the overall guest experience.

At its core, geolocational analysis involves the examination of spatially referenced data points, such as the location of accommodations, points of interest, transportation networks, and demographic information, to uncover meaningful patterns and relationships. By visualizing data on maps and utilizing spatial analysis techniques, researchers can identify spatial clusters, hotspots, and trends that offer valuable insights into market dynamics, customer behaviour, and competitive landscapes. One of the primary applications of geolocational analysis in the accommodation industry is site selection and location-based decision-making as shown in Fig.1.2. By overlaying demographic data, market demand projections, and

competitive landscapes onto spatial maps, businesses can identify optimal locations for new accommodations, expansion opportunities, and strategic partnerships.

Moreover, geolocational analysis serves as a powerful tool for destination marketing and tourism promotion. By leveraging spatial data visualization techniques, businesses can create compelling maps, interactive applications, and immersive experiences that showcase the unique attractions, amenities, and experiences offered by different accommodation destinations. Geospatial insights enable businesses to target specific market segments, tailor marketing campaigns, and create personalized experiences that resonate with travellers interests, preferences, and aspirations.

Geolocational data on accommodations, including location coordinates, types, prices, ratings, and amenities, were collected and cleaned in a CSV file. This involved addressing missing values, removing duplicates, and standardizing numerical features for clustering. Key visualizations like histograms and box plots were utilized to showcase trends. Geocoding & Search APIs were used to fetch accurate geographical information. The clustered data was visualized on a map using tools like Matplotlib, Seaborn, or Folium to understand spatial distribution.

Fig.1.2 Geolocation Mapping

Geolocational Mapping refers to the process of visualizing and analyzing data on maps to understand spatial relationships, patterns, and trends. It involves overlaying geospatial data onto maps to gain insights into the geographic distribution of various phenomena, such as locations of accommodations, points of interest, transportation networks, demographic information, and more.

1.1.3 K-MEANS CLUSTERING

K-Means clustering is a fundamental technique in unsupervised machine learning that plays a pivotal role in the analysis of geolocational data within the accommodation industry. As one of the most widely used clustering algorithms, K-Means offers a simple yet effective approach to partitioning data points into distinct groups based on similarity criteria. By iteratively assigning data points to the nearest cluster centroid and updating cluster centroids based on the mean of data points assigned to each cluster, K-Means enables researchers and practitioners to uncover spatial patterns, segment markets, and optimize resource allocation in the accommodation sector.

At its core, K-Means clustering aims to minimize the within-cluster sum of squares, thereby maximizing the similarity of data points within each cluster while maximizing dissimilarity between clusters. This objective makes K-Means particularly well-suited for tasks such as customer segmentation, market analysis, and spatial pattern recognition within the accommodation industry. By partitioning accommodation data into homogeneous clusters based on geographic proximity, amenities, pricing, and customer preferences, businesses can gain valuable insights into market segments, identify opportunities for targeted marketing, and tailor offerings to meet the diverse needs of different customer groups. One of the key advantages of K- Means clustering is its scalability and efficiency, making it suitable for analyzing large-scale geolocational datasets commonly encountered in the accommodation industry.

The various visualization of K Means clustering are Scatterplots (Fig.1.4), Histogram (Fig.1.5), Decision Trees (Fig.1.6), Bar Charts (Fig.1.7), Box Plot (Fig.1.8) respectively. Furthermore, K-Means clustering offers flexibility in defining the number of clusters, allowing researchers to adapt the algorithm to different

analytical objectives and business requirements. Through techniques such as the elbow method, silhouette analysis, and hierarchical clustering, researchers can determine the optimal number of clusters that best captures the underlying structure of the data and aligns with specific business goals. This flexibility enables accommodation providers to tailor clustering analyses to their unique contexts, whether it be market segmentation, customer profiling, or spatial pattern recognition. K-Means clustering offers a powerful framework for analyzing geolocational data within the accommodation industry.

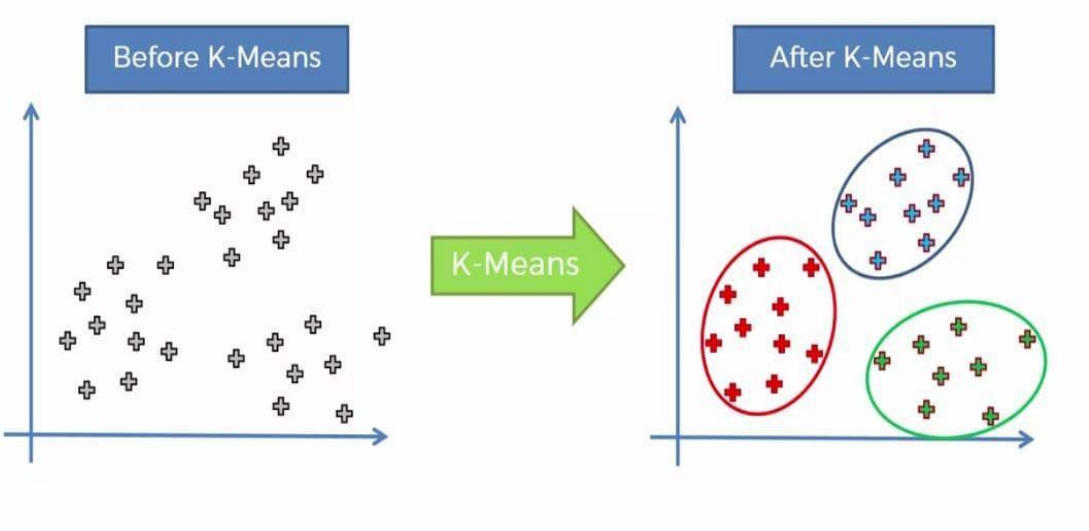

Fig. 1.3 K-means clustering

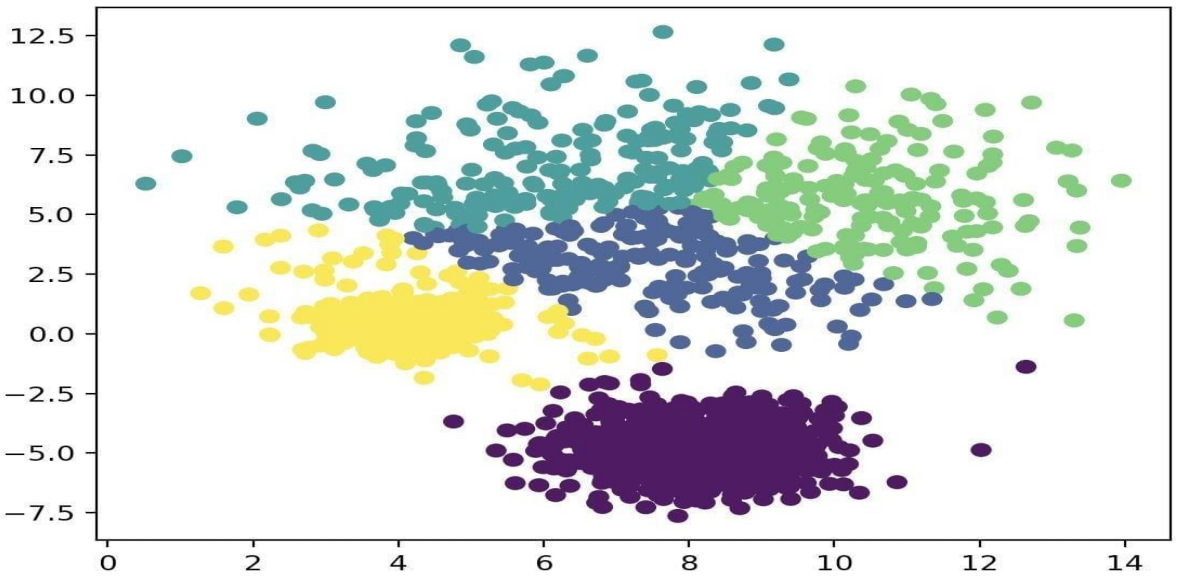

Fig. 1.4 Scatterplot Visualization of K-Means Clustering

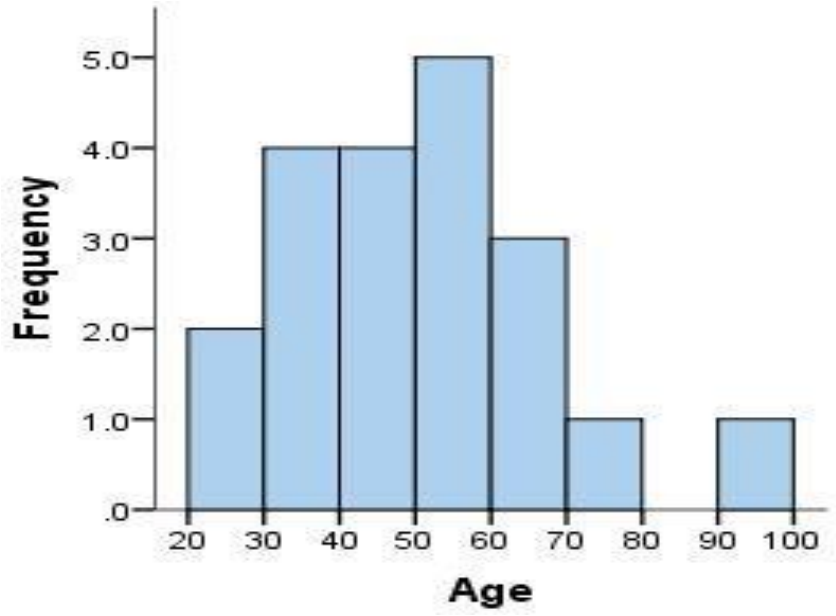

Fig. 1.5 Histogram Visualization of K-Means Clustering

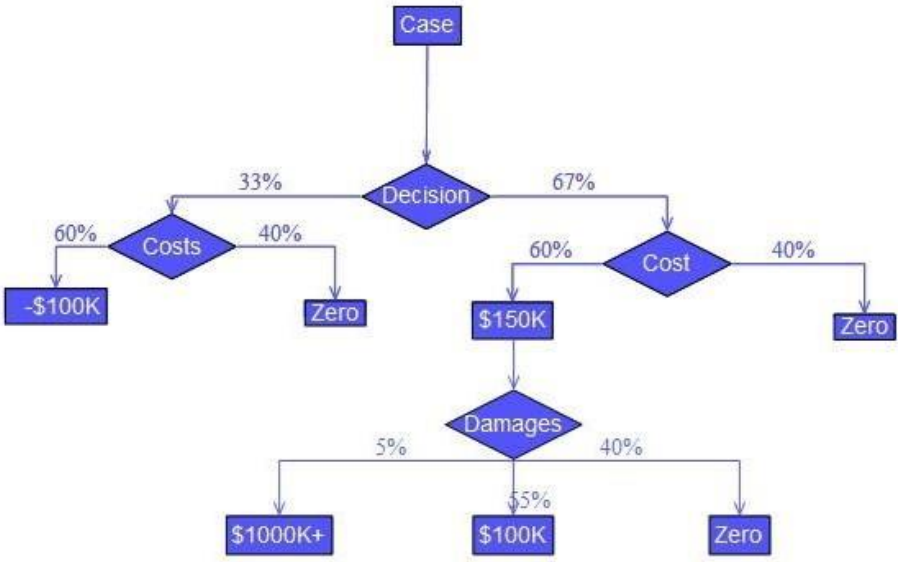

Fig. 1.6 Decision Tree Visualization of K-Means Clustering

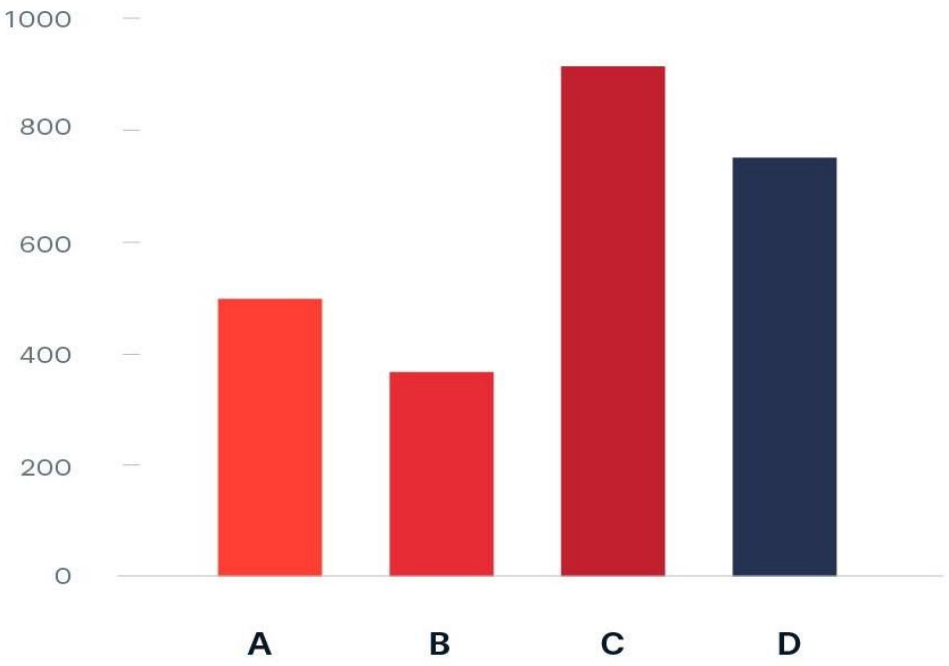

Fig. 1.7 Bar Chart Visualization in K-Means Clustering

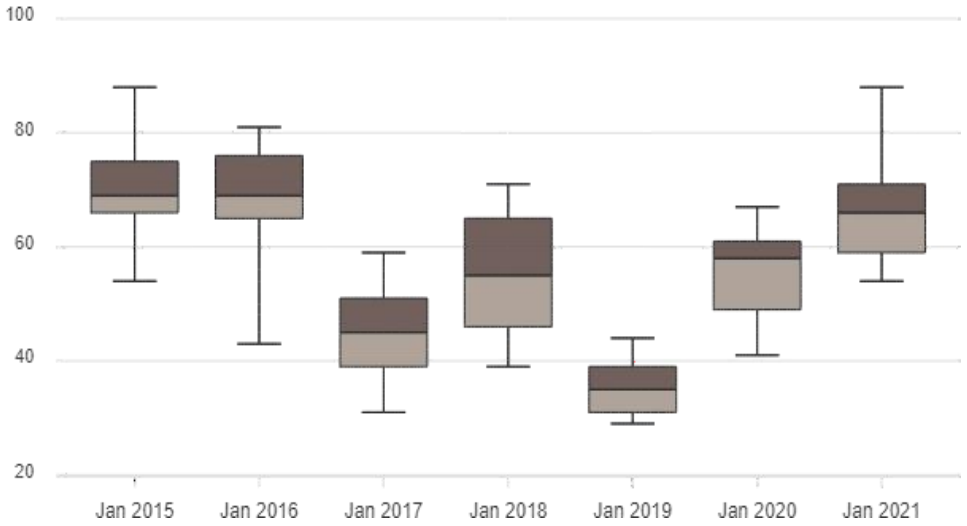

Fig. 1.8 Box Plot Visualization in K-Means clustering

K-MEANS APPLICATION

The application of K-means clustering is vast and diverse across various fields. Here are some common applications:

1. Customer Segmentation: Businesses use K-means clustering to segment customers based on their purchasing behavior, demographics, or preferences. This segmentation helps in targeted marketing strategies and personalized customer experiences.

2. Image Segmentation: In image processing, K-means clustering is used to partition an image into distinct regions or segments based on pixel similarity. This technique is widely used in medical imaging, object recognition, and computer vision applications.

3. Anomaly Detection: K-means clustering can be employed for anomaly detection by identifying data points that deviate significantly from the rest of the dataset. This is useful in fraud detection, network security, and quality control.

4. Recommendation Systems: K-means clustering can assist in building recommendation systems by grouping users or items with similar characteristics. This helps in suggesting products, movies, or content to users based on their preferences and behavior.

5. Genetic Analysis: In bioinformatics, K-means clustering is utilized to analyze gene expression data and identify patterns or clusters of genes with similar expression profiles. This aids in understanding gene functions and disease mechanisms.

6. Text Mining: K-means clustering is applied in text mining to group similar documents or articles together based on their content or semantic similarity. This is useful in document classification, topic modeling, and sentiment analysis.

7. Market Segmentation: K-means clustering helps marketers segment markets based on geographical, demographic, or behavioral characteristics of consumers. This enables businesses to tailor their products and marketing strategies to specific market segments.

8. Spatial Data Analysis: In geographical information systems (GIS), K-means clustering is used to analyze spatial data and identify clusters of geographical

locations with similar attributes. This is valuable in urban planning, environmental monitoring, and resource management.

9. Stock Market Analysis: K-means clustering is employed in financial analysis to group stocks with similar price movements or financial characteristics. This aids investors in portfolio diversification and risk management.

10. Healthcare: K-means clustering is used in healthcare for patient segmentation, disease diagnosis, and treatment planning. It helps in identifying patient subgroups with similar medical histories or symptoms for personalized healthcare delivery.

Performance Analysis

Performance analysis in the context of this research refers to the systematic evaluation and assessment of the effectiveness, accuracy, and robustness of machine learning models in their ability to cluster locations accurately. It involves quantifying various performance metrics such as accuracy, sensitivity, specificity, and precision to gauge the model's ability to locate the places on map based on the user preferences reliably. Through meticulous analysis and comparison of these metrics, performance analysis aims to provide insights into the strengths and limitations of different classification methodologies, ultimately guiding the development of more effective clustering of locations for better accommodation.

K- MEANS PERFORMANCE EVALUATION

1. Accuracy: Measures the proportion of correctly classified instances, providing an overall assessment of the model's performance.

2. Precision: Quantifies the proportion of true positive predictions among all positive predictions, emphasizing the model's ability to avoid false positives.

3. Recall (Sensitivity): Calculates the proportion of true positive predictions among all actual positive instances, indicating the model's ability to capture all relevant cases.

4. F1-Score: Harmonic mean of precision and recall, offering a balanced measure of a model's performance no imbalanced datasets.

K MEANS ALOGRITHM

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else.

Step-7: The model is ready.

Stopping Criteria for K-Means Clustering

1. There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:
2. Centroids of newly formed clusters do not change
3. Points remain in the same cluster Maximum number of iterations is reached
4. We can stop the algorithm if the centroids of newly formed clusters are not changing.
5. Even after multiple iterations, if we are getting the same centroids for all the clusters, we can say that the algorithm is not learning any new pattern, and it is a sign to stop the training.
6. Another clear sign that we should stop the training process is if the points remain in the same cluster even after training the algorithm for multiple iterations.
7. Finally, we can stop the training if the maximum number of iterations is reached.

1.2 PROBLEM STATEMENT

Analyzing geolocational data for accommodation purposes poses challenges in extracting actionable insights due to its volume and complexity. This project aims to utilize K-Means clustering algorithm to identify spatial patterns, customer preferences, and market segments within accommodation data. The goal is to develop a methodology to optimize pricing strategies, resource allocation, and personalized recommendations in accommodation.

1.3 OBJECTIVE

1. To develop an Exploratory Data Analysis model to examine and visualize geospatial data and to identify potential accommodation sites.
2. To recommend suitable places for accommodation by analyzing geo locational data after considering customer preferences on amenities, locations and budget.

1.4 ORGANISATION OF THE THESIS

The report is organized as follows.

- Chapter 1 introduces the project and about the Machine learning. It also describes the problem statement along with objective of the work.
- Chapter 2 contains the literature survey.
- Chapter 3 discusses the block diagram, tools and technologies.
- Chapter 4 contains the dataset description followed by the implementation details of each module and their experimental results.
- Chapter 5 contains the conclusion followed by references of the geolocational analysis.

CHAPTER 2

LITERATURE SURVEY

In [1], Atharva Mohite et al., explored the application of clustering algorithms, specifically K- Means and DBSCAN, in assisting students to find suitable accommodations based on preferences such as amenities, budget, and location proximity. The study employed exploratory data analysis techniques, including descriptive statistics, univariate and multivariate visualization, to analyze accommodation details in different city neighborhoods. The results demonstrated the effectiveness of K-Means for structured clustering and DBSCAN for flexible outlier detection.

In [2], Joelson Antonio Dos Santos et al., delved into the challenges and opportunities of applying hierarchical density-based clustering to large datasets using MapReduce. It referenced works by Naldi and Campello on the computational complexity of hierarchical clustering methods, highlighting the need for efficient parallelization schemes.

In [3], Weikai Yang et al., reviewed various techniques and approaches used for organizing and analyzing big data. This included traditional hierarchical clustering algorithms such as agglomerative and divisive methods, as well as more recent advancements like density-based clustering and graph-based clustering. Additionally, the survey covered topics such as constrained clustering, interactive clustering, and knowledge-driven clustering, highlighting the challenges and limitations faced by existing methods and the need for more adaptable and user-centric approaches.

In [4], Youguo Li et al., applied K-Means clustering algorithm combining the largest minimum distance algorithm and traditional K-Means would involve reviewing various techniques and advancements in K-Means clustering and related clustering algorithms. This included studies on the challenges and limitations of traditional K-Means clustering, such as its sensitivity to initial centroid selection and susceptibility to local minima. Research in this area covered methods proposed to address these issues, such as initialization strategies like K-Means++ and alternative distance metrics. Additionally, the survey explored related work on clustering algorithms that aim to improve clustering performance, enhance convergence speed,

or handle data with specific characteristics such as high dimensionality or uneven cluster sizes.

In [5], Jie Yang et al., discussed the challenges in identifying spatially contiguous areas characterized by similar social interaction features in cities. It highlighted the multidimensional nature of social interactions, encompassing aspects such as activity locations, venue functions, activity types, time periods, and demographic features. The study emphasized the use of dynamic geo-social records from sources like social media, mobile phones, and sensors as proxies for human mobility and social connectivity. By leveraging the high-dimensional and fine-grained nature of social media data, researchers aimed to extract spatial, temporal, topical, demographic, and contextual features to gain insights into the distribution of social interactions in space and time, ultimately contributing to the understanding of urban dynamics and social behavior.

In [6], Yash. M. Nemani et al., provided the use of mobile tourism applications utilizing Google API to provide information about places and their descriptions, as well as searching for nearby locations based on the user's current location. The project aimed to address the challenge of users having to visit multiple websites to plan their city visits, which may not always meet their specific requirements. By leveraging GPS and the Four-Square API instead of Google API, the City Tour Traveler system offered a more tailored and efficient approach to city exploration, allowing users to input their travel time and receive a customized travel plan for the city they are visiting.

In [7], Haang Viet Long et al., covered a range of topics related to recommendation systems. It included studies on the effects of overload on social media users, the integration of visual and textual information for improved recommendations, personalized recommender systems for product-line configuration processes, recommendation systems for fashion retail e-commerce, and a novel approach for product prediction using artificial neural.

In [8], Likhitha D et al., assisted students to find suitable accommodations based on preferences such as amenities, budget, and location proximity. The study employed exploratory data analysis techniques, including descriptive statistics, univariate and multivariate visualization, to analyze accommodation details in

different city neighborhoods. The results demonstrated the effectiveness of K-Means for structured clustering and DBSCAN for flexible outlier detection. The project's potential for machine learning algorithm integration, user feedback systems, and expansion to cover more locations and aspects like transportation and local culture highlighted its significance in aiding international students and workers in settling into new environments.

In [9], Hari Dheeraj et al., explored the use of advanced techniques such as K Nearest Neighbor Clustering, Artificial Neural Networks, and Decision Making to recommend accommodations for individuals, particularly students, migrating to new locations. The document highlighted the development of an enhanced clustering method, HC-PE, for dynamical systems, showcasing its superior performance in reducing complexity and preserving system models. Additionally, the study emphasized the application of K-Means clustering and Here Geocoding & Search API for scheduling journeys and swiftly exploring new cities, demonstrating the effectiveness of these methods in analyzing customer proximity to various amenities. The research also delved into the significance of exploratory analysis of geo-location data, customer segmentation studies, and the utilization of location-based social networks for data retrieval, providing valuable insights for marketers and travelers alike.

In [10], Gong, M et al., in the field of object re-identification addressed the challenges of unsupervised domain adaptation. Recent research has focused on leveraging clustering techniques, contrastive learning, and memory-based methods to improve the robustness and accuracy of re- identification models. Methods such as Uncertainty-aware Clustering for Unsupervised Domain Adaptive Object Re-identification (UCF) have shown promising results by effectively handling label noise and reducing the performance gap between unsupervised and supervised tasks. Additionally, the use of hierarchical clustering and uncertainty-aware collaborative instance selection has demonstrated significant improvements in model performance. Overall, this highlighted the importance of innovative techniques in enhancing object re-identification systems for real-world applications.

In [11], Daraio, E et al., explored the integration of Location-Based Social Network (LBSN) data with mobility data to analyze citizens' activities in urban environments. It discussed the challenges of collecting LBSN data, such as privacy concerns and connectivity issues, and highlights the implicit feedback on user habits provided by mobility data. The document introduced a multidimensional model that describes recurrent citizens' activities using a new pattern type called the generalized activity pattern. Real-world data from Foursquare check-ins, taxi services, and free-floating car sharing are used to evaluate the proposed approach, demonstrating the complementarity and interchangeability of data sources in various spatio- temporal conditions.

In [12], Sharma, S. et al., compared the Single Linkage and Complete Linkage in Agglomerative Hierarchical Cluster Analysis for identifying tourist segments revealed a growing interest in utilizing different clustering methods to segment tourist populations effectively. Few previous Journal highlighted the importance of selecting the appropriate linkage method, with findings indicating that Complete Linkage may offer advantages in certain scenarios. This research contributed to the field by providing insights into the performance of clustering techniques in tourism segmentation, offering valuable implications for businesses and policymakers seeking to enhance marketing strategies and tailor services to diverse tourist segments.

In conclusion, the array of research papers and projects showcased the evolving landscape of data-driven methodologies, spanning clustering algorithms, recommendation systems, and geolocational data analysis. From enhancing accommodation search for students to refining tourism applications, these endeavors exemplify the practical applications of data science. Moreover, the focus on leveraging dynamic geo-social records underscore the potential for technology to deepen our understanding of urban dynamics and human behavior. Together, these contributions signify the enduring relevance and impact of data-driven approaches in tackling modern challenges.

CHAPTER 3

METHODOLOGY

3.1 BLOCK DIAGRAM OF THE PROPOSED SYSTEM

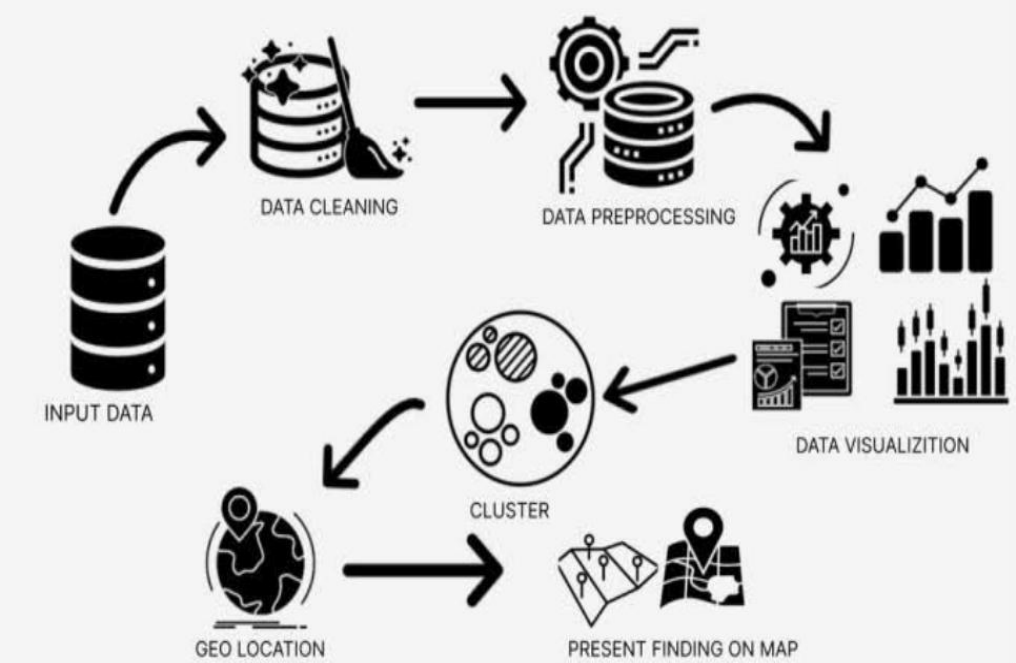

Fig. 3.1 Block diagram of the proposed system

Figure 3.1 shows the block diagram of the proposed system. Geolocational data on accommodations, including location coordinates, types, prices, ratings, and amenities, were collected and cleaned in a CSV file. This involved addressing missing values, removing duplicates, and standardizing numerical features for clustering. Key visualizations like histograms and box plots were utilized to showcase trends. Geocoding & Search APIs were used to fetch accurate geographical information. The K-means clustering algorithm was applied to analyze the data, identifying groups or clusters within accommodations. The clustered data was visualized on a map using tools like Matplotlib, Seaborn, or Folium to understand spatial distribution.

3.2 SYSTEM REQUIREMENTS

System requirements are the minimum specifications needed for hardware, software, and other resources to run a computer system or software application effectively.

3.2.1 HARDWARE REQUIREMENTS

Processor: A modern multi-core CPU (e.g., Intel Core i5/i7 or AMD Ryzen series) for faster data processing and model training.

RAM: At least 8GB of RAM is recommended for handling large datasets and running machine learning models efficiently. More RAM may be necessary for extremely large datasets or complex models.

Ethernet Connection: Stable internet connection with a minimum speed of 10 Mbps.

Optimal performance- wired ethernet connection is preferred over wifi.

3.2.2 SOFTWARE REQUIREMENTS

Operating System: Most popular operating system, Windows is used for data analysis.

Data Analysis Tools: Need software for data manipulation, visualization, and modelling. Commonly used tools include:

- Python: Utilize libraries like Pandas, NumPy, Matplotlib/Seaborn for data manipulation and visualization, and scikit-learn for machine learning models.
- Google Colab: A cloud-based platform provided by Google that enables users to write and execute Python code in a collaborative environment, offering free access to GPU resources for machine learing tasks.

3.2.3 INTERNET CONNECTION

A stable internet connection may be necessary for accessing online datasets, APIs, or cloud-based services.

3.3 TOOLS AND TECHNOLOGIES

The following tools are used in this project.

- **Python**

Python is an essential component of sentiment analysis because it is a universal language that can be used for various tasks, especially sentiment analysis, not just for data analysis and machine learning. Python has a rich set of libraries and frameworks that makes it easy to work with data and build models.

- **Google Colab Notebook**

Google Colab is a free Jupyter notebook that allows to run Python in the browser without the need for complex configuration. It comes with Python installed and has all the main Python libraries installed. It also comes integrated with free GPUs.

- **Pandas**

Pandas is a Python package providing fast, flexible and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical and real-world data analysis in Python.

- **Scikit-learn**

Scikit-learn, often abbreviated as sklearn, is a comprehensive machine learning library in Python, offering tools for data preprocessing, model building, evaluation, and more. With a user-friendly interface and extensive documentation, sklearn simplifies the implementation of various machine learning algorithms. From classification and regression to clustering and dimensionality reduction, sklearn provides a wide range of algorithms and utilities for both supervised and unsupervised learning tasks.

- **NumPy**

NumPy is a powerful Python library for numerical computing. It provides support for arrays, matrices, and mathematical functions, making it efficient for performing mathematical operations on large datasets. It is widely used in data science, machine learning, and scientific computing due to its speed and versatility.

- **Matplotlib**

Matplotlib is a powerful plotting library for Python widely used for creating static, interactive, and publication-quality visualizations. It provides a flexible interface for constructing various types of plots, including line plots, scatter plots, bar plots, histograms, pie charts, and more.

- **Seaborn**

Seaborn is a Python data visualization library based on matplotlib. It provides a high- level interface for drawing attractive and informative statistical graphics. While Seaborn is not specifically designed for machine learning, it is commonly used in conjunction with machine learning tasks to visualize data distributions, relationships, and patterns, aiding in data exploration, model evaluation, and result interpretation.

- **Folium**

Folium is like Google Maps for Python code. It creates interactive maps with markers, shapes, and layers, all using Python. It is a great tool for visualizing geographic data and exploring locations within the code.

- **GeoPy**

GeoPy is a Python library that work with geographical data and perform geocoding (converting addresses into coordinates) and reverse geocoding (finding addresses from coordinates). It provides a simple interface to access

various geocoding services, such as Google Maps, Bing Maps, and OpenStreetMap. GeoPy is helpful for tasks like finding distances between locations, getting location information, and adding spatial context to the projects.

- **SciPy**

  SciPy is an open-source Python library used for scientific and technical computing. It builds on NumPy and provides additional functionality for optimization, interpolation, integration, linear algebra, signal processing, and much more. SciPy is widely used in fields such as physics, engineering, biology, and finance for solving complex mathematical problems and performing advanced data analysis tasks.

- **MiniSom**

  MiniSom is a Python library for creating self-organizing maps (SOMs), which are a type of artificial neural network used for dimensionality reduction and clustering of data. MiniSom provides a lightweight and efficient implementation of SOMs, making it suitable for handling large datasets. It's often used in tasks such as data visualization, feature extraction, and exploratory data analysis, particularly in fields like data mining, pattern recognition, and machine learning.

- **Four Square API**

  The FourSquare API provides developers with access to extensive location-based data and services. It enables venue search, details retrieval, recommendations, check-ins geocoding, and more. With features like venue categories, photos, and tips, it enhances location-aware applications and social networks. Developers can integrate it to create personalized experiences and innovative location-based services.

CHAPTER 4

IMPLEMENTATION

INTRODUCTION

The implementation of our project revolves around harnessing the power of K-means clustering to analyze geolocational data for accommodation purposes. Leveraging this technique, we aim to efficiently segment geographic regions into clusters based on similarities in accommodation features, facilitating targeted analysis and decision-making processes. By utilizing K-means clustering, we can effectively identify patterns and trends within the data, enabling us to offer tailored insights and recommendations for accommodation providers and stakeholder

4.1 DATA COLLECTION

- We collected data from various platforms such as Google, Kaggle amd Google docs.
- Thedata contains detailed description of Geo Locational data including latitude, longitude, amenities, price and user reviews.
- The dataset has been combined in a .csv file which encompasses 126×61 rows and columns as shown in Fig. 4.1.

	GPA	Gender	breakfast	calories_chicken	calories_day	calories_scone	coffee	comfort_food	comfort_food_reasons	comfort_food_reasons_coded	cook	comfort_
0	2.4	2	1	430	NaN	315.0	1	none	we dont have comfort	9.0	2.0	
1	3.654	1	1	610	3.0	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	1.0	3.0	
2	3.3	1	1	720	4.0	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	1.0	1.0	
3	3.2	1	1	430	3.0	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	2.0	2.0	
4	3.5	1	1	720	2.0	420.0	2	Ice cream, chocolate, chips	Stress, boredom, cravings	1.0	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	
120	3.5	1	1	610	4.0	420.0	2	wine. mac and cheese, pizza, ice cream	boredom and sadness	NaN	3.0	
121	3	1	1	265	2.0	315.0	2	Pizza / Wings / Cheesecake	Loneliness / Homesick / Sadness	NaN	3.0	

Fig.4.1. Sample Data form the Dataset

4.2 DATA CLEANING

- Cleaning of data involved removing any duplicates or irrelevant entries and addressed missing values by either imputing them based on neighboring data or removing records with significant missing information.
- Next, standardized the data by scaling numerical features to ensure equal importance during clustering.
- Also we considered outlier detection and removal to prevent them from skewing cluster boundaries.
- Additionally, performed feature engineering to extract relevant information such as distance from key landmarks or amenities.

- Finally, validated the cleanliness of the data through exploratory data analysis to ensure it aligns with the project objectives and is ready for K-means clustering analysis as shown in Fig.4.2.

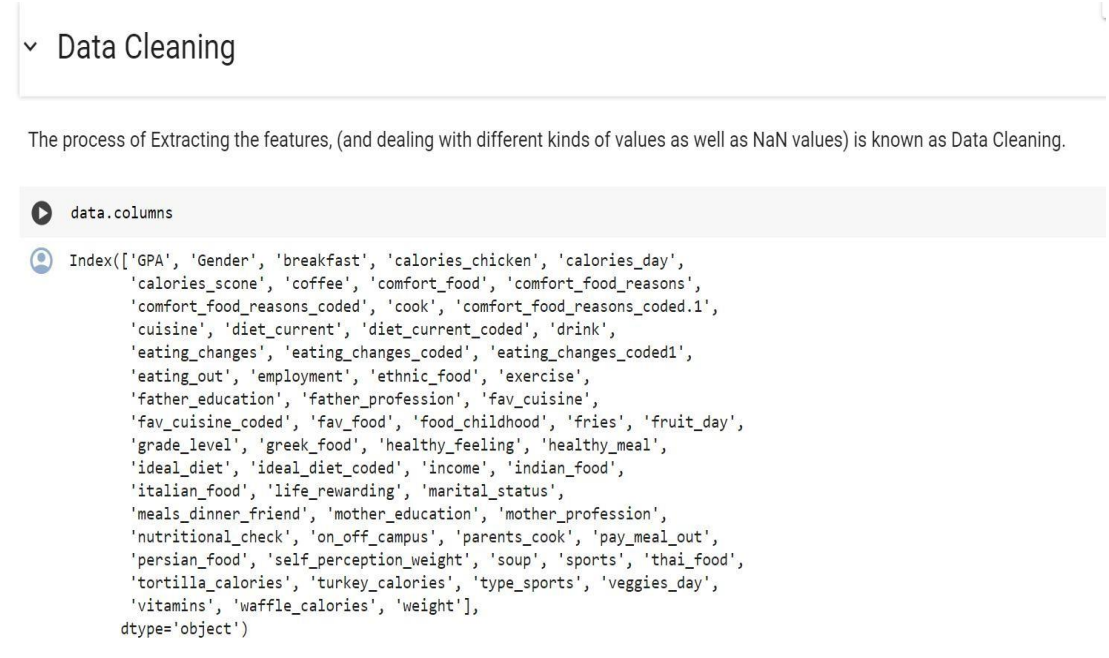

Fig. 4.2 Data Cleaning

4.3 DATA VISUALISATION

- To visualize the results of a geolocational data analysis project for accommodation using K-means clustering, we started by plotting the raw data points on a map to understand the distribution of accommodation locations.
- Then, visualized the data points using various visualisation methods such as pair plot(Fig.4.3.1) and Box Plot(Fig.4.3.2) where scatter plots show how two variables are related to each other helping us understand any patterns or connection between them.
- Finally, this provided interactive features allowing users to explore the individual data points for deeper insights into accommodation distribution.

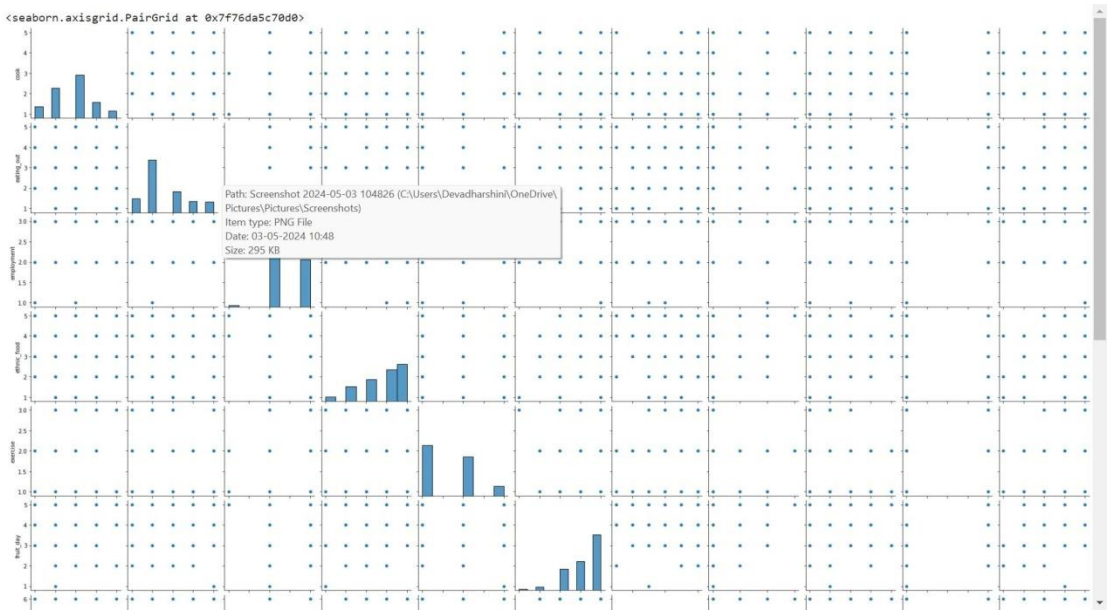

Fig. 4.3 Pair Plot

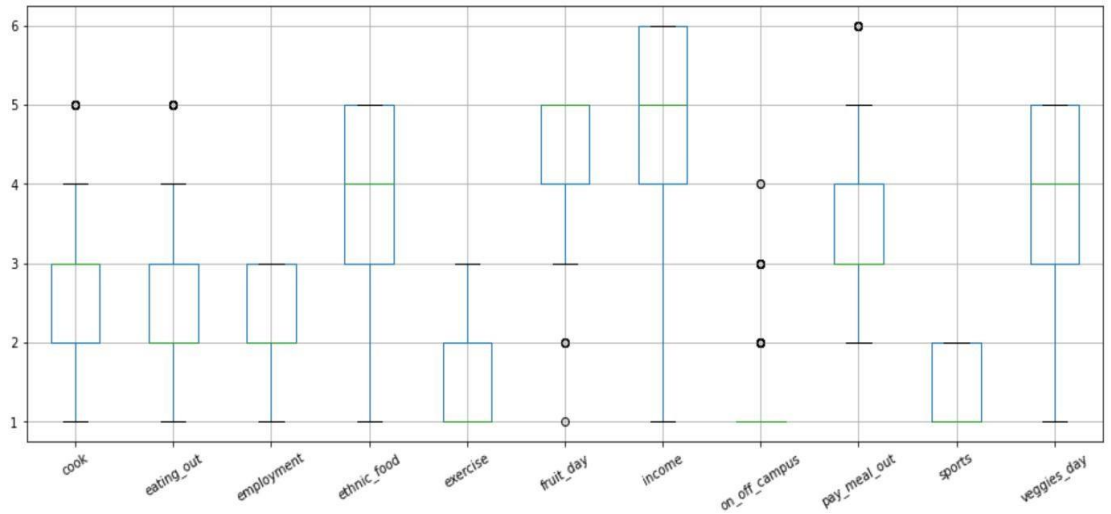

Fig. 4.4 Box Plot

4.4 GEOLOCATION ANALYSIS

- Analysis of geolocational data using FourSquare API
- Foursquare City Guide, commonly known as Foursquare, is a local search-and- discovery mobile app developed by Foursquare Labs Inc.
- The API provides personalized recommendations of places to go near a user's current location based on users' preferences on location, amenities, budget.

4.5 CLUSTERING ALGORITHM

- The K-means clustering algorithm was employed in this project to partition geolocation data pertaining to accommodation facilities into distinct clusters based on their spatial proximity as referred in Fig 4.5.
- Initially, the algorithm randomly initializes cluster centroids and assigns each accommodation facility to the nearest centroid.
- Subsequently, the centroids are recalculated as the mean coordinates of the facilities assigned to each cluster.
- This process iterates until convergence, where the cluster assignments stabilize, or a predefined stopping criterion is met.
- By iteratively optimizing cluster centroids, K-means effectively delineates spatial clusters of accommodation options, enabling the identification of geographical patterns and trends within the dataset.
- This algorithmic approach provides a structured framework for analysing and organizing geolocation data, facilitating informed decision-making processes for both travellers and service providers in the accommodation sector.

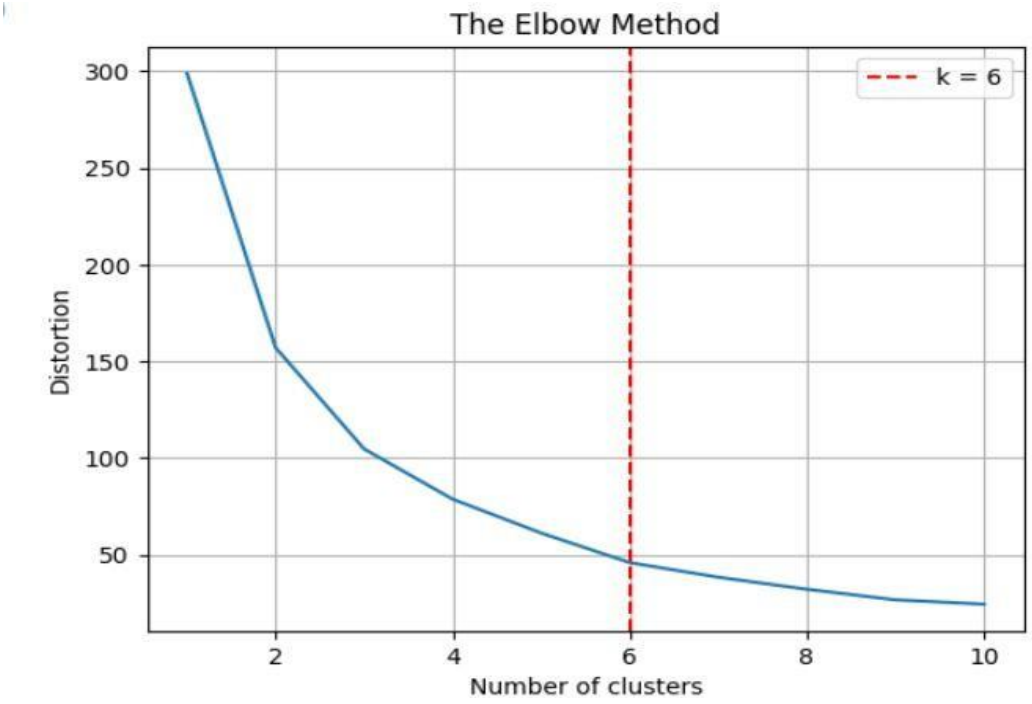

Fig. 4.5 Elbow Curve

The elbow method is a technique used to determine the optimal number of clusters in a k-means clustering algorithm. It involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow" point, where the rate of decrease in WCSS slows down significantly. This point indicates the optimal number of clusters where adding more clusters doesn't significantly reduce the WCSS.

4.6 CLUSTERED OUTPUT

- The clustered output of the analysis of geolocation for accommodation reveals insightful spatial patterns and groupings within the accommodation sector.
- Through the application of K-means clustering algorithm, the geographical landscape is partitioned into distinct clusters, each representing a unique concentration of accommodation facilities.
- These clusters showcase a nuanced distribution of accommodations

based on various factors such as proximity to landmarks, amenities, transportation hubs, and price range.

- Additionally, the clustered output facilitates targeted decision-making processes for travellers seeking accommodations and empowers service providers with valuable insights to optimize their offerings and marketing efforts.
- Overall, the clustered output underscores the significance of geolocation analysis and clustering algorithms in unravelling the spatial dynamics of the accommodation sector, thereby informing strategic initiatives and enhancing the overall user experience.

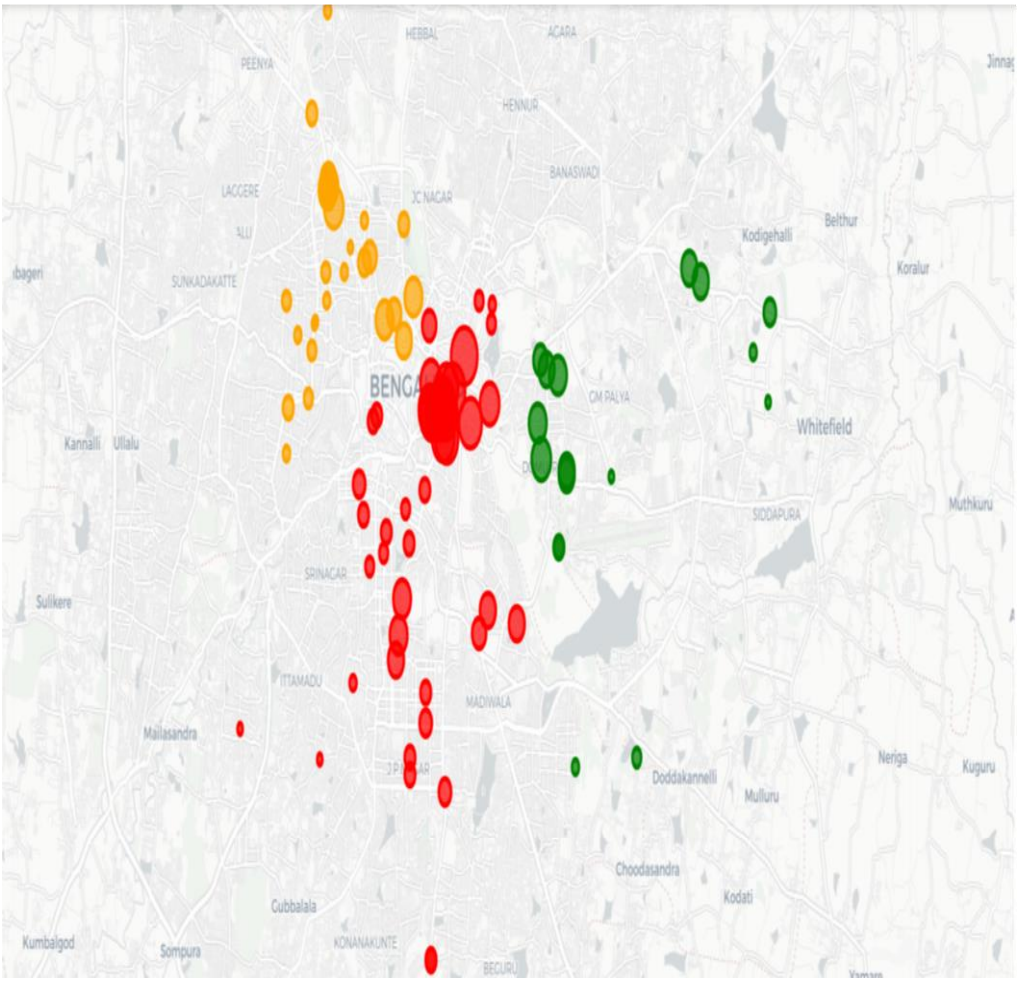

Fig. 4.6 Clustered Output

PERFORMANCE ANALYSIS

Performance of the proposed system is evaluated by four metrics namely Accuracy, Precision, Recall, and F1 Score whose equations are given below.

- Accuracy = $\frac{TN+TP}{TN+FP+TP+FN}$
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1 Score = 2 * $\frac{Precision*Recall}{Precision+Recall}$

Where,

TN is TRUE NEGATIVE

TP is TRUE POSITIVE

FN is FALSE NEGATIVE

FP is FALSE POSITIVE

In our proposed approach, we have achieved TN = 7, FP = 2, FN = 3 and TP = 18 as shown in the confusion matrix given below.

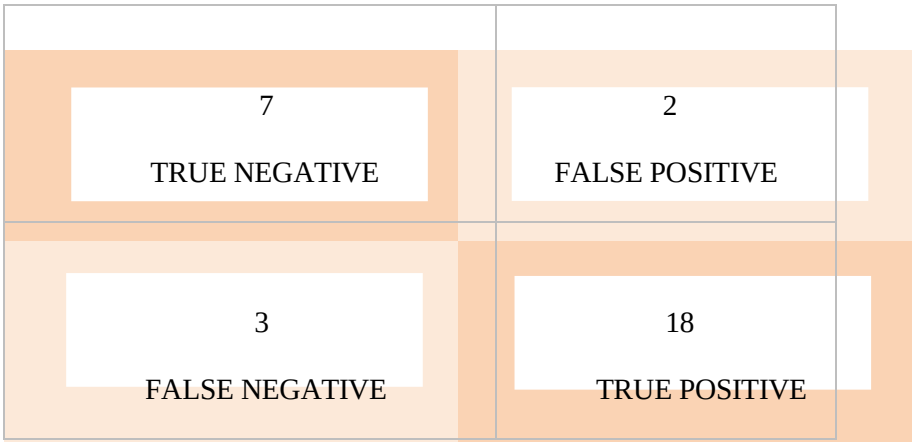

Fig. 4.7. Confusion matrix of the proposed system

Accuracy represents the number of correctly classified data instances over the total number of data instances.

$$\text{Accuracy} = \frac{7+18}{7+2+18+3}$$

In this example, Accuracy = (7 + 18)/(7 + 2 + 18 + 3) = 0.833 and in percentage the accuracy will be 83%.

Precision should ideally be 1 (high) for a good classifier. Precision becomes 1 only when the numerator and denominator are equal i.e., $TP = TP + FP$, this also means FP is zero. As FP increases the value of denominator becomes greater than the numerator and precision value decreases.

$$\text{Precision} = \frac{18}{18+2} = \frac{18}{20} = 0.9$$

Recall should ideally be 1 (high) for a good classifier. Recall becomes 1 only when the numerator and denominator are equal i.e., $TP = TP + FN$, this also means FN is zero.

As FN increases the value of denominator becomes greater than the numerator and recall value decreases.

Recall is also known as sensitivity or true positive rate and is defined as follows:

$$\text{Recall} = \frac{18}{18+3} = \frac{18}{21} = 0.85$$

So ideally in a good classifier, we want both precision and recall to be one which also means *FP* and *FN* are zero. Therefore, we need a metric that takes into account both precision and recall. *F1-score* is a metric which takes into account both *precision* and *recall* and is defined as follows:

$$\text{F1 Score} = 2 * \frac{0.9*0.85}{0.9+0.85}$$

$$= 2 * 0.437$$

$$= 0.874$$

$$= 87\ \%$$

CHAPTER 5

CONCLUSION

The utilization of geo-locational data, coupled with data visualization and clustering techniques, offers a promising approach to identify ideal accommodation options based on individual preferences and interests. By employing methodologies such as K-Means Clustering and leveraging geo- locational data from sources like Here Geocoding and Search API, it becomes feasible to categorize accommodation based on amenities, location, and budget preferences. This approach not only streamlines the process of locating suitable accommodation but also enhances the overall user experience by providing intelligent suggestions tailored to individual needs.

Furthermore, the results of the exploratory analysis on geo-location data demonstrate the effectiveness of K-Means clustering in grouping locations based on amenities, as evidenced by the clustered maps showcasing regions with varying levels of amenities such as restaurants, departmental stores, and gyms. The user-friendly nature of the developed website, which presents clustered geographic data, contributes to resolve common challenges faced by migrants in finding housing options that align with their preferences.

In conclusion, the integration of geo-locational data analysis, clustering techniques, and user-friendly interfaces holds significant potential in facilitating the search for suitable accommodation options, particularly in scenarios involving frequent travel, migration, or tourism.

5.1 FUTURE SCOPE

Future work in this area may involve enhancing the website functionality by incorporating features like directions between locations and implementing user login through platforms like Google, thereby further improving the user experience and accessibility of accommodation recommendation

REFERENCES

1. Atharva Mohite, Rushikesh Kulkarni, Moham med Salmanuddin ,"Exploratory Data Analysis", International Journel of Advances in Engineering and Management (IJAEM) 2023.

2. Joelson Antonio das santos, Joerg Sander, Talat labal Syed Youguo Li, "Hierarchical Density- Based Clustering Using Map Reduce", IEEE (2020).

3. Shixla llu, Jie Lu, Welkai Yang, "Iterative steering of hierarchical clustering", IEEE (2020)

4. Youguo Li, Haiyan Wu, "A Clustering method based on k-means Algorithm", SciVerse Science Direct 2012

5. Jie Yang, Alessandro Bozzon," Regionalization of social interactions points of interest location predictions with Geo social data", Research Gate (2018).

6. Yash.M.Nemani, Ra mesh Yadev, Mr.Manish Bhelande, "City tour a traveller based on Four Square API", International Reasearch Journel of Engineering and Technology (IRJET) 2018.

7. Hoang viet Long, "A Hybrid Action- Related K-Nearest Neighbour (HAR-KNN) Approach for Recommendation", IEEE Access 2020

8. Pooja k, Lithitha D, Ashwini, Gauri Sameer, "Exploratory analysis of geolocational Data", International. Journal of Advances in Engineering and Management (IJAEM) 2023.

9. Press V Raju, M Hari Dheera, n Dinesh, "Exploratory Analysis on Geo-Location Data for Accomodation", JETIR 2023.

10. Wang, P., Ding, C., Tan, W., Gong, M., Jia, K. and Tao, D, "Uncertainty-aware clustering for unsupervised domain adaptive objectre-identification". IEEE 2022.

11. Daraio, E., Cagliero, L., Chiusano, S. and Garza, P, "Complementing Location-Based Social Network Data with Mobility Data: A Pattern-Based Approach", IEEE 2022.

12. Sharma, S. and Batra, N, "Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering, IEEE (2019).