

**Optimal Accommodation Solutions for Immigrants: Leveraging Geolocational Data Analysis**

By

**SHOBALOJU ABDULLAHI ABOLA JI**

**180211052**

*A PROJECT REPORT SUBMITTED FOR THE DEGREE OF BACHELOR IN ENGINEERING*

*IN THE DEPARTMENT OF ELECTRONIC AND COMPUTER ENGINEERING, FACULTY OF ENGINEERING, LAGOS STATE UNIVERSITY.*

SUPERVISED BY

**DR. AJASA ABIODUN AFIZ**

*DECEMBER 2023*

**DECLARATION**

This project is totally original, approved by my academic supervisor, and carried out through my own efforts. When collaboration with others or the use of data produced by other researchers is utilised, due credit is given in the form of an acknowledgement or clear distinction through references, if that is thought to be appropriate.

This work is being submitted in order to complete the requirements for the Electronics and Computer Engineering Bachelor of Engineering (B. Eng.) degree.

**SHOBALOJU ABDULLAHI ABOLAJI** **180211052**

Name of Student Matriculation Number

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student’s Signature

**CERTIFICATION**

This thesis is an official declaration that the project work was carried out under my supervision at the Department of Electronic and Computer Engineering at Lagos State University, Epe Campus, by **SHOBALOJU ABDULLAHI ABOLAJI**, with the matriculation number **180211052**.

Additionally, the Department has formally acknowledged and accepted it as an important reference document.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. A. A. AJASA **Date**

**Project Supervisor**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. A. A. Adenowo **Date**

**Head of Department**

**ACKNOWLEDGEMENT**

I give thanks to God, the divine being, for showing me his constant love during every section of this research project.

My sincere thanks goes out to the people who have been instrumental in helping me complete my research thesis. Throughout my academic endeavour, their unfailing support, counsel, and encouragement have been priceless. I am indebted to my parents, Mr. Kazeem Shobaloju and Mrs Abidemi Shobaloju, who provided me with constant financial support, affection and encouragement, from the initial stage till the completion of this project. I would like to extend my appreciation to my siblings, Nurat, Semiat, Fadil, and Kharimat, who rendered financial guidance and support, my friends and peers for creating a supportive and intellectually engaging academic environment. Their viewpoints and conversations have deepened my grasp of the topic and enhanced the value of this study experience.

In addition, I would like to express my sincere gratitude to my supervisor, Dr. Afiz Ajasa Abiodun, for his outstanding leadership and mentoring. His knowledge, constructive feedback, and unwavering support have all greatly influenced the course of this study. His devotion and commitment to my academic growth are greatly appreciated.

**Abstract**

The growing patterns of global migration makes it difficult to provide immigrants with appropriate recommendations for suitable and customised housing. This study tackles this challenging problem by utilizing geolocational data analysis techniques, conquering the limitations of conventional recommendation techniques. By combining comparable neighbourhoods based on amenities, safety, and cultural diversity, this study makes use of a diverse set of unsupervised machine learning techniques, which includes k-means Clustering, Agglomerative Hierarchical Clustering, and Principal Component Analysis to improve precision. With the use of a rich dataset that includes dwelling attributes, demographic data, and geolocation information, this study contributes to enhancing existing accommodation recommendation models and offers intuitive information for better housing for immigrants.

**Table of Content**

**DECLARATION\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2**

**CERTIFICATION\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_3**

**ACKNOWLEDGEMENT\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_4**

**ABSTRACT\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_5**

**CHAPTER 1 INTRODUCTION\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_9**

* 1. **Background Study and Motivation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_9**
  2. **Problem Statement\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_11**
  3. **Objectives of the Study\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_12**
  4. **Scope of the Study\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_12**
  5. **Significance of the Study\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14**
  6. **Organization of Dissertation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_15**

**CHAPTER 2 LITERATURE REVIEW\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_16**

**2.1 Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_16**

**2.2 Conventional Unsupervised Machine Learning Algorithms and their application\_\_\_\_18**

**2.2.1 Gaussian Mixture Model\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_18**

**2.2.1.1** The Expectation-Maximization (EM) Algorithm\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20

2.2.1.2 Music Recommendation System (MRS)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_21

2.2.2 Fuzzy C-means clustering Algorithm (FCM)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_22

2.2.2.1 Smartvote\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_23

2.2.2.2. Surabaya Tourism Destination Recommendation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_24

**2.3 Recent Techniques to Tackle Similar Tasks\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_27**

2.3.1 Internet-Based Travel Portals\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_27

2.3.2 Blockchain-powered Solutions\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_27

**2.4 Other Techniques That Can Be Utilized\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_28**

2.4.1 Dimensionality reduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_28

2.4.1.1 t-Distributed Stochastic Neighbor Embedding (t-SNE)\_\_\_\_\_\_\_\_\_\_\_\_\_\_29

2.4.2 Autoencoders\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_29

**2.5 Research Gap\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_29**

**2.6 Proposed Technique\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_30**

**2.7 Summary of the Chapter\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_30**

**CHAPTER 3 METHODOLOGY\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_31**

**3.1 Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_31**

3.1.1 Clustering\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_33

3.1.1.1. Data Collection\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_33

3.1.1.2. Data Pre-processing\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_33

3.1.1.3. Feature Selection\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_33

3.1.2 Dimensionality reduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_36

3.1.2.1. Data Collection\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_36

3.1.2.3. Determining the co-variance matrix\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_36

3.1.2.4. Determining eigenvectors and eigenvalues\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_36

3.1.2.5. Selecting the top K eigenvectors\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_37

**3.2 Limitations of the Methodology\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_38**

**3.3 Results Comparison\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_39**

**3.4 Summary of the Chapter\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_39**

**CHAPTER 4 RESULTS AND DISCUSSION\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_41**

**4.1 Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_41**

**4.2 K-means Clustering\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_41**

**4.3 Agglomerative Hierarchical Clustering (AHC)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_47**

**4.3 Principal Component Analysis (PCA)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_54**

**4.4 Comparative Analysis\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_61**

**CHAPTER 5 CONCLUSION AND RECOMMENDATIONS\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_63**

**5.1 Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_63**

**5.2 Recommendation for future works\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_64**

**REFERENCES\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_66**

**Chapter 1**

**INTRODUCTION**

**1.1 BACKGROUND STUDY AND MOTIVATION**

Geolocational data analysis is the technique of removing pertinent observations and patterns from geospatial data, and it enables businesses and individuals to comprehend the geographical aspects of many occurrences [1]. Geolocational data analysis is essential in many industries, including urban planning, transportation, marketing, environmental monitoring, and emergency response, as a result of the growing use of location-based services and applications in increasingly interconnected society. Geolocational data analysis has a long history, and it has evolved significantly due to major advances in technology and processing capacity [2]. Ancient civilizations used crude maps, radar systems, and surveys to navigate and record territorial boundaries, which is where geolocational data analysis got its start [3]. The invention of instruments like compasses and astrolabes made it possible for early geographers to gather fundamental geographic information.

The 1960s saw the development of computers and digital technologies, which paved the way for Geographic Information Systems (GIS) [4]. Due to his work on the Canada Geographic Information System (CGIS) in 1963, Roger Tomlinson is frequently referred to be the "father of GIS" [4]. This was the first time that geographical data was collected, stored, and analyzed systematically. With the launching of remote sensing satellites like Landsat 1 in 1972 [5] and SPOT 1 in 1986 [6], the 1970s and 1980s saw a considerable advancement in the interpretation of geolocational data. High-resolution photos of the Earth's surface were taken by these satellites, allowing for more thorough study and mapping. Additionally, the Global Positioning System (GPS) rollout in 1993 gave users precise location data for a variety of uses [7].

Advancements in database technologies made it easier to combine geographic data with conventional relational databases in the 1990s. Geolocational data could now be stored, retrieved, and analyzed more quickly owing to this integration. Oracle Spatial and PostgreSQL with PostGIS are two examples of strong spatial databases that may be used to manage geospatial data [8]. Geolocational data analysis was revolutionized by the broad adoption of the internet and the creation of web-based mapping services. Launched in 2005 [9], Google Maps brought interactive online maps to a wide audience. This signalled the advent of the geospatial web and the subsequent incorporation of geolocation into several web-based programmes and services. Geolocational data analysis now has more options due to the exponential growth of digital data, improvements in computing power, and data analytics. Gigantic volumes of geolocational data become accessible with the growth of smart phones, IoT devices, and social media platforms.

Global positioning systems (GPS), satellites, mobile devices, sensors, social media check-ins, and internet-connected gadgets are just a few of the many ways that geolocation data is gathered [10]. To help pinpoint the precise position of items, events, or people, this data offers precise coordinates, timestamps, and other pertinent information. Analysts can gain valuable insights, make wise decisions, and enhance the effectiveness and efficiency of numerous operations by utilizing this geographical data.

Data science, geographical analysis, and machine learning approaches are all used in the study of geolocation data [11]. Data cleansing, visualization, geocoding (assigning geographic coordinates to non-geographic data), spatial clustering, spatial interpolation, and spatial prediction are frequent activities that are included. These methods help analysts recognize spatial patterns, spot anomalies, forecast future trends, and comprehend the connections between various geographic features better.

However, geolocational data analysis generates significant privacy and ethical issues. Establishing strong protections is essential to preserve people's privacy rights and ensure ethical data management practices as the gathering and usage of geolocational data grows. Geolocational data analysis has become a potent tool for learning about our physical world in this era of rapidly increasing technology and interconnection. Organizations and people may use the amount of geospatial data at their disposal to drive innovation, make better decisions, and have a beneficial impact on society. The potential for geolocation data analysis to change industries and deepen our understanding of the surroundings is enormous as the subject develops.

For immigrants and their new communities, housing is crucial in promoting social integration and cohesion. The development of surroundings where immigrants feel welcomed, can be assisted, supported, and connected by advising housing options that take into account elements like closeness to existing immigrant communities, cultural centres, language resources, and support networks. This in turn fosters a sense of belonging, promotes engagement and teamwork, and develops links within the community. There is a unique opportunity to use this data-driven approach for immigrant accommodation recommendations because geolocational data is becoming more widely available and data analysis techniques like Agglomerative Hierarchical clustering, K-means clustering, Fuzzy C-means clustering, and Principal Components Analysis [12] are improving. By utilizing these technologies, creative solutions that effectively match immigrants can be created with the best housing possibilities while taking the aforementioned variables into account. Finding appropriate housing that suits their unique needs and preferences is one of the difficulties immigrants sometimes confront when transferring to a new country or area. An immigrant's overall experience and wellbeing in their new environment can be considerably impacted by the process of obtaining suitable housing. Using geolocational data analysis can improve the experience of immigrants suggesting housing options that match their tastes, occupations, cultural proximity, and accessibility to necessary facilities.

**1.2 PROBLEM STATEMENT**

The current methods for helping immigrants find suitable homes frequently rely on manual search techniques, individual referrals, or generic housing listings. These methods fall short in addressing the unique requirements and preferences of each immigrant, which might result in poor housing decisions and subsequent integration challenges. The intricate interactions between home qualities, immigrant preferences, community factors, and access to amenities necessitate a thorough investigation in order to determine the best housing options for immigrants. Traditional approaches are frequently insufficient to address the multifaceted nature of the issue and successfully connect immigrants with housing options that meet their unique needs.

To address these issues, geolocational data analysis techniques must be used to create a recommendation system that helps immigrants locate the best housing. It is possible to offer individualized and accurate recommendations that take into account factors like proximity to immigrant communities, cultural amenities, accessibility to services, transportation options, and other pertinent criteria by leveraging the power of geospatial data, housing market information, and immigrant preferences.To match immigrants with housing options that suit their unique preferences and promote a more seamless transition and integration, the task at hand is to create algorithms, make use of the data sources that are already accessible, and apply the right analytical tools.

**1.3 OBJECTIVES OF THE STUDY**

The focus of this research is based on the following objectives

1. To employ data-driven insights to create numerous algorithms that use the gathered geolocational data to provide the best housing alternative for immigrants.
2. To evaluate and contrast three clustering and dimensionality reduction approaches used in the study of geolocational data.
3. To determine the most effective data analysis technique for making accommodation recommendations.
4. To enhance current methods for data analysis and accommodation recommendation.

**1.4 SCOPE OF THE STUDY**

The scope of this study is limited to:

1. The geographic area or regions where the geolocational data and housing market statistics are available are the focus of this investigation. The conclusions and suggestions are restricted to this region alone, and they might not be readily transferable to other areas with various housing markets and demographic makeups.

2. This study makes use of geolocational data from public records, mapping services, and real estate databases. The range of coverage and precision is constrained by the quality and availability of the data sources used.

3. Based on the data that is available, this study takes into account a variety of housing characteristics, including location, size, amenities, and rental costs. However, the availability of data and its applicability to immigrant preferences may affect the inclusion of particular dwelling characteristics.

4. This study considers immigrant preferences that can be gleaned through surveys or other data sources. Depending on the availability of data and the study's particular focus, the range of preferences taken into account may be restricted to elements like proximity to other cultures, access to services, and transportation options.

This work creates a recommendation system using methods for geospatial data analysis. The recommendation system's application is restricted to the approaches and algorithms used in this study, and the effectiveness of the system is assessed using the data at hand and the evaluation metrics of choice.

In order to evaluate the effectiveness of the recommendation system, this study chooses evaluation metrics. The chosen measures' shortcomings may not fully reflect all facets of suggestion efficacy or may not perfectly match the objectives of immigrant accommodation suitability.

**1.5 SIGNIFICANCE OF THE STUDY**

The significance of this study lies in its potential to address critical societal and policy challenges while offering practical benefits:

1. Personalised accommodation recommendations that take into account their tastes and needs will help this study greatly improve the integration of immigrants into their new communities. This may result in quicker adaption, less loneliness, and more social cohesiveness.

2. Efficient Resource Allocation: This study can help governments, NGOs, and housing providers optimise resource allocation by finding the best geospatial data analytic tools for accommodation recommendations. This makes sure that the housing options with the biggest positive impact are given priority when limited resources are available.

3. Enhanced Housing Access: Language and cultural obstacles, which immigrants frequently encounter when looking for acceptable accommodation, may be lessened as a result of this research. It promotes fair access to housing possibilities by advising accommodations that are both reasonably priced and culturally appropriate.

4. Data-Driven Policy Insights: The results of this study can help in the development of evidence-based immigration and housing support policies. This research can be used by policymakers to develop and put into practise more successful immigration settlement and integration plans, which will eventually benefit both immigrants and host communities.

5. Community Well-Being: Communities can become more diversified and vibrant as a result of effective immigrant assimilation. Immigrants are more likely to positively contribute to the economic and social fabric of their new home and to the general experience of the community when they are able to find suitable housing and feel welcomed.

6. Research Advancement: By illustrating the usefulness of exploratory data analysis in a social setting, this study makes a contribution to the fields of data analysis and recommendation systems. It exemplifies how cutting-edge analytics methods may handle problems in the real world outside of the scope of conventional corporate applications.

7. Ethical Considerations: The work emphasises the significance of ethical issues in data-driven decision-making, creating a precedent for responsible data use in related applications by eliminating potential biases in the recommendation system and fostering fairness and transparency.

8. Knowledge Exchange: Non-profit organisations, community groups, and housing providers can all benefit from the study's findings in addition to academics and policymakers. The sharing of expertise may inspire group initiatives to provide support and housing for immigrants.

In conclusion, the importance of this work goes beyond its technical and scholarly components. It could have a good effect on immigrant lives, support inclusive communities, and advance the proper use of data to address difficult societal problems.

**1.6 ORGANIZATION OF DISSERTATION**

This section aims to provide a clear framework by outlining the content and organisation of the following chapters, ensuring a methodical and persuasive development of concepts throughout the research. The structure and organisation of this study, as well as its framework, have been thoroughly discussed in this preliminary chapter. The study is divided into five chapters.

The first part of this chapter introduces the research topic and describes the background information: basic principles, history and applications of geolocational data analysis on immigration and housing challenges. The second part outlines the research problem statements which describe the specific issues that this research project aims to address. The third section list the research objectives. The scope and limitations of this study are discussed extensively in the fourth section. The final section of this chapter outlines the significance and contribution of this study to accommodation recommendation for immigrants. Chapter 2 discusses the review of relevant literature on immigrants housing challenges, exploration of previous research on geolocational data analysis for accommodation recommendations, discussion of relevant theories and frameworks and identification of gaps in the existing literature. The explanation of the research methodology used in this study, data collection methods, data preprocessing and cleaning techniques, details on the machine learning techniques employed and explanations of how recommendations are generated are discussed in chapter 3.

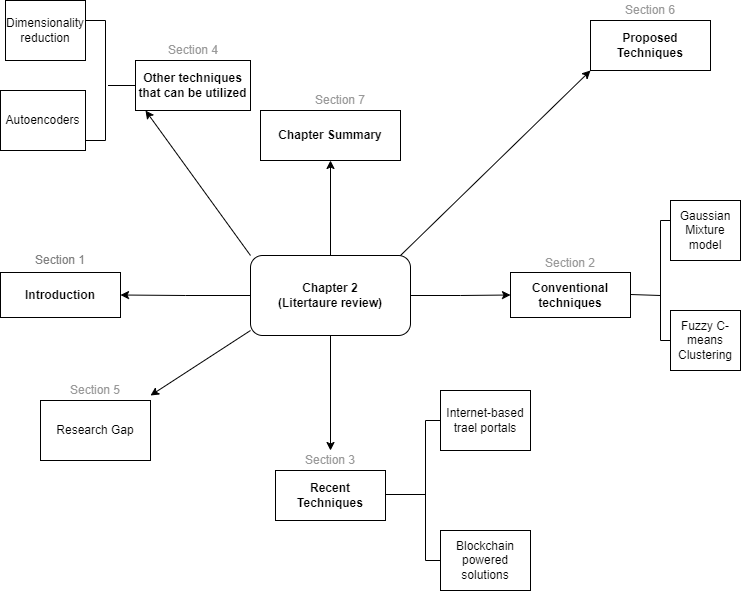
**Chapter 2**

**LITERATURE REVIEW**

**2.1 INTRODUCTION**

In this chapter, related works on various geolocational data analysis techniques utilized for accommodation recommendation shall be thoroughly reviewed. Before then, there is a need to evaluate why unsupervised machine learning models are the most suitable geolocational data analysis technique for accommodation recommendation.

Unsupervised machine learning is a subset of machine learning in which the algorithm discovers patterns and structures in data without direct supervision or samples with labels [36]. In other words, it doesn't need a set output or target variable. In order to find hidden patterns, group related data points, or reduce dimensionality, the algorithm investigates the data's intrinsic structure and relationships [37]. A class of recommendation systems are designed from unsupervised machine learning algorithms which has the capability to group data points together based on their resemblances [38], reduce the number of features or dimensions in data while upholding its important structure [39], detect infrequent or abnormal data points [40], and generate new data that is identical to the input data distribution [41]. The potential uses of unsupervised machine learning are in abundance in any area where clustering [42], image segmentation, object recognition [43], fraud detection [44], sentiment analysis [45], and collaborative filtering [46] are required. Similar technologies such as Netflix Recommendation System (NRS) [47], and Monolith [48] have been done. The works in [47] and [48] primarily concentrate on updating existing parameters while performing real-time training on less data. Hotel recommendation system (HRS) have been employed for several years to locate a hotel swiftly and efficiently based on clustering and Rankboost algorithm. [49]



*Figure 2.1 Overall Visualization of the Literature review*

**2.2 CONVENTIONAL UNSUPERVISED MACHINE LEARNING ALGORITHMS AND THEIR APPLICATIONS**

This section describes multiple Machine Learning algorithms employed in the modelling of recommendation systems and presents mathematical formulations for each model.

**2.2.1 Gaussian Mixture Model**

The Gaussian Mixture Model (GMM) is a probabilistic approach primarily employed for the estimation of density and clustering. It is presumptively formed from a combination of different Gaussian distributions, each of which represents a different cluster. Data points can be members of numerous clusters at once because to the probability assignments made by GMM. Applications for pattern recognition and machine learning make extensive use of this model [50]. In one dimension, the probability density function of a Gaussian distribution is denoted by

(2.1)

(2.2)

Where and are the mean and variance of the distribution respectively. Suppose there are K amount of clusters (where and are also approximated for each K), as a linear function of the densities of each of these K distributions, the probability density is defined below, Where  is the "mixing coefficient" for every kth distribution. The maximum log-likelihood approach is used to estimate the parameters. This can be achieved by computing

(2.3)

(2.4)

A random variable, , such that is defined.

From Bayes' theorem,

(2.5)

The derivative of with respect to and should be zero in order for the log-likelihood function to have the highest value. Consequently, by rearranging the aforementioned terms, and setting the derivative of to zero, with respect to ,

(2.6)

Similarly, the following equations can be obtained by taking the derivative with regards to and

respectively.

(2.7)

And

(2.8)

In light of this, it is evident that closed form estimation of the parameters is not possible. The Expectation-Maximization algorithm is useful in this situation.

**2.2.1.1 The Expectation-Maximization (EM) Algorithm**

The Expectation-Maximization (EM) algorithm is a repetitive technique for determining maximum-likelihood estimates for model variables when the data is insufficient, contains some missing data, or contains hidden parameters. As a fresh batch of data is estimated, EM selects some arbitrary values for the missing data points. Recursively using the new values to fill in the gaps until the values are fixed, they are then utilised to predict a better initial date. The two most crucial steps that are iteratively carried out to update the model parameters until the model convergence in the Expectation-Maximization (EM) technique are the estimation step (E-step) and maximisation step (M-step). In the estimation step, one of the first model parameters like the mean and covariance matrix is initialized. Using the most recent parameter values, it is possible to calculate the posterior probabilities of each data point belonging to each centroid. A latent variable is frequently used to express these possibilities. On the basis of the present parameter values, the hidden variable values are finally estimated. In the maximization step, the estimated latent variable is used to update the parameter values. By averaging the weighted data using the probabilities associated with the latent variables, the mean of the cluster point is updated. The squared disparities between the data points and the mean are weighted averaged using the relevant latent variable probabilities, and the covariance matrix is likewise updated in this manner. By averaging the probabilities of each latent variable for each component, the mixing coefficient is updated.

The estimation and maximization steps are repeated until the change in the parameters or the log-likelihood falls below a predetermined entry or until the maximum number of iterations is reached. The latent parameters are basically updated based on the current parameter values in the estimation step. The parameter values are updated using the estimated latent variables during the maximization step. Iterations of this method are done until the model eventually converges.

Multiple recommendation systems have been built using GMM in recent years

**2.2.1.2 Music Recommendation System (MRS)**

A music recommendation system based on Gaussian Mixture Model Yang Lu et.al [57] was designed in

2015. It was developed to employ GMM in three-dimensional data to analyse user preferences for various musical moods. Initially, only music segments lasting 30 seconds were retrieved, pre emphasized, and normalised. All music was pre-processed at this point. Pre-emphasis is added to the hamming window with a coefficient of -0.9375. Ten training moods and forty songs were used to run this model. The models of GMMV+ and GMMV- for the positive and negative shaft ends, which represent the emotional intensity, are established. To train GMMV+, high intensity training samples that represent the emotions of annoyance and excitement are used, while GMMV- is trained using low intensity training samples that represent the emotions of sadness and peace. The two models depict the intensely optimistic and the bad mood. On both ends of the valence shaft, models of GMMV+ and GMMV- that represent positive and negative emotions respectively are set. A two-layer classifier is created during the experiment. A test sample first ascertains the mood's strength using the first layer classifier, then following the second classifier, it ascertains the positive and negative emotions. GMMV+ is used as an example to show how the data normalisation process works to extract the feature parameters for the user's music. The characteristic vector for the V shaft is set to Cxv, and the posteriori probability given that the value for θv+ is in the GMMV+ model. The value in θv+ for the training characteristic vector Cv+ of GMMV+ should represent the maximum posteriori probability when the model parameter is , and it is displayed as

(2.9)

The minimum value is denoted as

(2.10)

The data point C in the axis of V is indicated as

(2.11)

Where K stands for the quantity of training model samples. The normalized data points for C on the axis of V- are also denoted as

(2.12)

Similarities between two users A and B are indicated by

(2.13)

Using this data, it is possible to determine the resemblance to judge the same preference points for challenging consumers, and then popularize the new music.

**2.2.2 Fuzzy C-means clustering Algorithm (FCM)**

The concept of fuzzy set theory was first put forth by Lufti Asker Zadeh in 1965 [51], and it characterised the uncertainty of belonging using a membership function. Unreliable class membership function is provided by the usage of fuzzy sets. Richard Bellman, Lufti Zadeh, and Enrique Ruspini's work was the first to suggest using fuzzy set theory for cluster analysis [52]. Integration of fuzzy logic with data mining techniques has become one of the key constituents of recommendation models in handling challenges posed by massive collections of natural data. The central idea in fuzzy clustering is the non-unique partitioning of the data into a collection of clusters. The data points are assigned membership values for each of the clusters and fuzzy clustering algorithm allow the clusters to grow into their natural shapes [53]. In fuzzy c-means clustering, k clusters are selected, and each cluster's data point is given a random coefficient. This is carried out repeatedly until the algorithm converges. Every point x has a set of coefficients that indicate its degree of membership in the kth cluster, wk(x). When using fuzzy c-means, a cluster's centroid is the average of all points that have been weighted according to how much they belong to the cluster.

(2.14)

where m is the hyper-parameter that determines the cluster's degree of fuzziness. The cluster will ultimately be fuzzier the higher it gets. A finite collection of n elements P = {p1, p2,....,pn} is divided by the FCM algorithm according to a predetermined criterion, into a group of c fuzzy clusters. A list of c cluster centres C = {c1, c2,....,cn} and a partition matrix is produced by the algorithm given a finite collection of data.

where every element determines the extent to which element is associated to cluster .

The primary objective of FCM is to minimize the function:

(2.15)

Where

(2.16)

Multiple recommendation systems have been built using FCM algorithms in recent years.

**2.2.2.1 Smartvote**

Switzerland uses the online voting system Smartvote for local, cantonal, and national elections [54]. The Smartvote system compares the profiles of voters and candidates. Based on commonalities between voters and candidates, Smartvote recommends candidates. When smartvote was initially created for the 2003 Swiss parliamentary elections, a large portion of the population adopted it right away. Smartvote has been providing its services for communal and cantonal elections since 2004. By 2018, more than 200 elections in Switzerland alone had employed smartvote [55]. With the use of smartvote, voters are connected with politicians and political parties that support the same policies. By answering 75 questions on current policy concerns on a standardised questionnaire, a voter can construct a political profile. The profiles of the parties and candidates running are then compared to those already compiled. The candidates themselves respond to the smartvote questionnaire, in contrast to other online voting tools. The voter is given a list of candidates after completing the questionnaire, which is arranged in descending order based on how closely each candidate or party matches the voter's profile [55]. The algorithm used to match the voter profiles with those of candidates or parties is the Euclidean distance

(2.17)

where a and b are two separate points in Euclidean space, and are Euclidean vectors.

To generate the recommendations, Smartvote computes the "MatchPoints" using

(2.18)

where and are the responses of online voters *v* and political candidate *c* to questions *i* respectively. The points of agreement between voter, candidate and questions is denoted by .

A perfect match exists if there is a "yes-yes” and “no-no” combination and a bonus is assigned in this case, which is denoted by *y*. The values presented in common questions are shown in table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| V(⇓) C(⇒) | Yes | Probably Yes | Probably No | No | Total |
| Yes | 100+50 | 75 | 25 | 0 | 250 |
| Probably Yes | 75 | 100 | 50 | 25 | 250 |
| Probably No | 25 | 50 | 100 | 75 | 250 |
| No | 0 | 25 | 75 | 100+50 | 250 |
| Total | 250 | 250 | 250 | 250 |  |

*Table 2.1. Points of agreement in smartvote system*

Consideration of each question's relevance to voters is the next step in the computation of matching. For every question, there is a natural selection process which is done by weighting additions, subtractions and equality. The appropriate match points are then multiplied by the factors 0.5, 1 and 2.

(2.19)

where represents the matching score of voter *v* and candidate *c* to the questions *i.*

Finally, the match between voter, candidate and questions is computed with

(2.20)

The highest potential matchscore, which in turn depends on the voters' responses, determines how the matching point is calculated. Smartvote can be used to provide recommendations from complete lists, in which case the matching values are calculated using the list's candidates' mean averages.

**2.2.2.2. Surabaya Tourism Destination Recommendation**

Raymond et.al [56] in 2018 implemented the fuzzy c-means clustering algorithm into developing a desktop application. The purpose of this application was to use the FCM algorithm to evaluate user input data and produce recommendations for travel destinations that would be appropriate for their needs. This application requires the user to enter the following pieces of information: total budget, total day trip, total number of passengers, and travel destination. An administrator needs to enter information about the hotels, eateries, and tourist attractions in Surabaya in order to turn the input data into the output data. Thirty data on tourist locations, fifty data on hotels, and fifty data on restaurants were gathered for this study. Name, address, phone number, and average cost per night for a basic room are the details required for hotels in Surabaya. Name, address, phone number, and typical cost of a standard meal are the pieces of information required for restaurants in Surabaya. However, the information required for tourist locations is name, address, and telephone number. The Surabaya Tourism Board, Google, or an on-site survey were all used to collect the data, which was then saved in a database to be further analyzed by a clustering algorithm.

There are three steps to clustering once all the data have been collected. The clustering of tourist destinations by latitude and longitude is the initial step. Based on distance, this will divide thirty tourist locations into three groupings. Following the initial clustering, each cluster will be once more categorized according to how far the hotels and restaurants are from the tourist attractions. The average cost of hotel rooms and restaurant meals is the final clustering step. As a result, the customer will receive recommendations in three different tables, one for each hotel, restaurant, and tourist attraction. The user can also view the distance between each location and the forecasted budget allocation in this way. By selecting the context menu button in each row, they may also view more information about each proposed location, such as the full address, phone number, brief description, and photo.

The techniques that have been reviewed thus far are soft clustering techniques. Although, soft clustering algorithms require faster computational time, hard clustering algorithms have faster convergence rate [58], which is why they will be employed in this study.

**2.3 RECENT TECHNIQUES TO TACKLE SIMILAR TASKS**

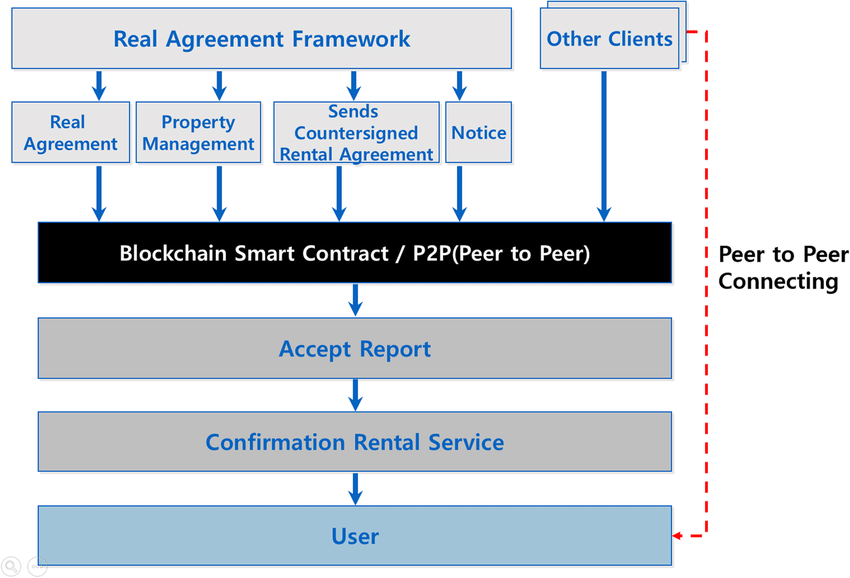
The problem of effectively recommending accommodations for immigrants has led to the development of a number of strategies that make use of a range of technologies and techniques. Present-day tactics combine conventional housing platforms with state-of-the-art technologies to guarantee customised and culturally aware solutions. A few noteworthy methods are as follows:

**2.3.1 Internet-Based Travel Portals**

Immigrants looking for short-term or long-term housing have been using Booking.com, Airbnb, and Zillow, among other traditional internet accommodation services, extensively. These platforms use property facts, reviews, and user preferences to suggest accommodations that would be appropriate.

**2.3.2 Blockchain-powered Solutions**

Blockchain technology is being investigated by a few up-and-coming platforms to improve the security, trust, and openness of the accommodation recommendation process. Because blockchain maintains the accuracy of data, transactions go more smoothly, and immigrants feel more secure.



*Fig 2.3, Blockchain-based process for house rental*

**2.4 OTHER TECHNIQUES THAT CAN BE UTILIZED**

Multiple recent techniques can be utilized for recommending suitable accommodation for immigrants. They are discussed in the following section.

**2.4.1 Dimensionality reduction**

In data analysis and machine learning, dimensionality reduction is an essential technique, especially for datasets with many features or dimensions [19]. Dimensionality reduction can be used to simplify the data, increase computational effectiveness, and boost recommendation algorithm performance when it comes to recommending accommodations for immigrants. The following are some application of dimensionality reduction methods in this domain:

**2.4.1.1 t-Distributed Stochastic Neighbor Embedding (t-SNE):**

One popular method for visualising high-dimensional data in two or three dimensions is t-SNE. t-SNE is a dimensionality reduction technique, albeit its main application is in visualization [59]. When handling accommodation features that display intricate interdependencies, it can be useful to capture local relationships within the data.

**2.4.2 Autoencoders**

One kind of neural network architecture called autoencoders is intended for unsupervised learning. They are composed of a decoder and an encoder, with a compressed representation of the input data represented by the middle layer [60]. The most important features are efficiently captured by the intermediate layer, which trains the autoencoder to reconstruct the input data. As a reduced-dimensional form for accommodation advice, this compressed representation can be applied.

**2.5 RESEARCH GAP**

The literature review provides a thorough summary of methods for recommending accommodations, emphasising the transition from physical models to data-driven approaches. Although previous studies have discussed the application of geolocation data to suggest accommodations, there may be a lack of attention to the complex cultural preferences of immigrant groups. Cultural factors may not have been sufficiently addressed in the literature up to this point, such as accessibility to cultural centres, houses of worship, or community centres catering to certain ethnic communities. It's possible that a lot of recommendation systems are unable to adjust dynamically to shifting neighbourhood dynamics. Over time, immigrant communities frequently see changes in amenities and demographics. To ensure ongoing relevance for immigrant residents, this research will concentrate on creating recommendation models that can adjust in real-time to changing neighbourhood factors.

**2.6 PROPOSED TECHNIQUE**

Unsupervised machine learning techniques will be employed in this study, particularly clustering algorithms such as k-means and agglomerative hierarchical clustering, and Principal component analysis to group neighborhoods based on shared characteristics. This step aims to identify distinct clusters representing areas with similar cultural amenities, safety profiles, and other relevant features. These techniques functions similarly to the supervised machine learning language that have been reviewed, but it requires less computing power and gives more accuracy It is also simple to include with the current infrastructure.

**2.7 SUMMARY OF THE CHAPTER**

The use of unsupervised machine learning methods to suggest immigrant housing holds great promise for enhancing immigrant integration and housing experiences. These systems can help immigrants discover accommodation that suits their requirements and preferences by utilising demographic data, location-based characteristics, and affordability considerations. However, addressing ethical issues and improving personalisation continue to be significant obstacles in this subject, opening the door for additional study and innovation.

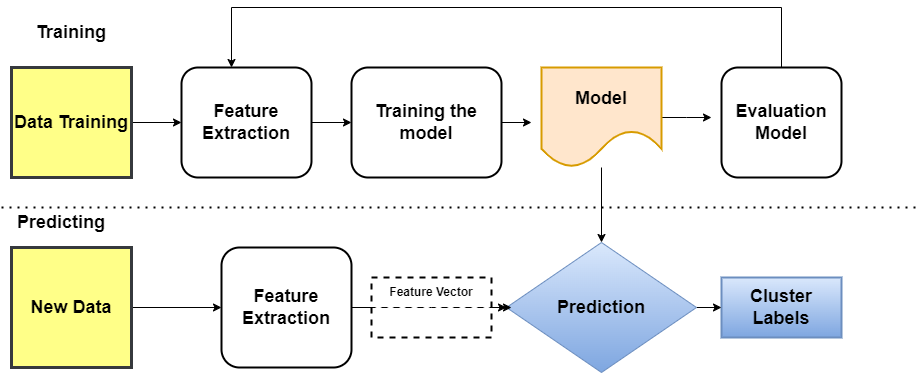
**Chapter 3**

**METHODOLOGY**

**3.1 INTRODUCTION**

In this work, computer simulations, dimensionality reduction, data analysis, and clustering are all included. Clustering and dimensionality reduction are two methods applied in this study. This study encompasses data collection and processing, data analysis, dimensionality reduction, cluster evaluation and computer simulations. The techniques utilized in this research include clustering and dimensionality reduction. The procedures for creating and assessing machine learning models to recommend suitable accommodation to immigrants are covered in this section. Figure 3.1's flowchart offers a summary of the process.

Gathering data is the first step in our technique. Extensive datasets are essential to our methodology's success. The recommendation engine's basic materials include geolocation data, demographic data, home attributes, and cultural facilities. To guarantee a comprehensive portrayal of the lodging landscape, information is gathered from a variety of trustworthy sources. A number of preprocessing procedures are applied to the raw data to guarantee its quality, applicability, and suitability for machine learning analysis. This covers encoding categorical variables, handling missing data, and normalisation. In order to prepare the dataset for the ensuing unsupervised machine learning algorithms, data preparation is essential. The choice of an effective unsupervised machine learning algorithm is a crucial step in our process. Algorithms for clustering, like k-means or hierarchical clustering, show up as the best options for organising comparable accommodations according to shared features, which makes the recommendation process easier. The preprocessed dataset is then used to train the selected unsupervised learning algorithm. In this stage, the model uses the data to find patterns, clusters, and relationships without the requirement for pre-established labels. This ability to learn on its own is a crucial asset for revealing hidden patterns in the housing market.



*Fig 3.1 Research Methodology Flowchart*

**3.1.1 Clustering**

The clustering technique is executed in the following steps:

**3.1.1.1. Data Collection:**

A comprehensive dataset of housing attributes such as location, size, amenities, rental prices, and neighbourhood demographics are gathered from kaggle. Additional data on immigrant preferences, including factors like cultural proximity, accessibility to services, occupation, and proximity to immigrant communities are also acquired.

**3.1.1.2. Data Pre-processing:**

The collected data is cleansed using python pandas and pre-processed by removing duplicates, handling missing values, and standardizing variables. Feature engineering is then performed to extract relevant features or create new variables that capture meaningful information for the clustering analysis.

**3.1.1.3. Feature Selection:**

Exploratory data analysis is conducted to gain insights into the dataset and identify the most influential features for the clustering process. Statistical techniques, such as correlation analysis or feature importance methods are utilized to select the most relevant features for clustering.

Clustering:

This study shall apply two different clustering techniques which are;

(i) K-means clustering

(ii) Agglomerative hierarchical clustering

K-means Clustering:

The K-means clustering algorithm creates clusters by utilizing the average value of the objects within each cluster [13]. The main tasks of the k-means algorithm are

1. To find the optimal number of centre points or centroids, referred to as K, through a process of evaluation.

2. To assign data point to the nearest k-centre, forming a cluster based on the data points that are in close proximity to that specific k-centre [14].

Thus, every cluster has data points with some similarities, and is distant from other clusters. Firstly, we determine the optimal number of clusters (K) using the elbow method. This method employs the concept of WCSS (Within Cluster Sum of Squares) value, which quantifies the total variances within a cluster.

Mathematically,

(3.1)

Where Ci is the cluster, Cn is total clusters, is the Cluster centroid, and d is the data point in each cluster.

To calculate the distance between data points and centroids, we use the Euclidian distance method [15] which mathematically, is;

(3.2)

Where;

(a1, b1) are the coordinates of one point.

(a2, b2) are the coordinates of the other point.

3 d is the known distance between (a1, b1) and (a2, b2).

Then the K-means clustering algorithm is applied to segment the housing dataset into K distinct clusters based on the selected features. Lastly, we appropriate scaling or normalization techniques, if required, to ensure fair comparison between different feature scales.

Agglomerative hierarchical clustering (AHC):

AHC is a bottom-up technique that eliminates the need for predefining the number of clusters. [11] .It is a method where each instance is initially treated as a separate cluster, and subsequently, these clusters are combined to form larger clusters. This process is repeated until all the clusters are merged into a single, comprehensive cluster that encompasses all the instances [12]. Firstly, the appropriate linkage criterion (e.g., Ward, complete, average, or single linkage) is determined based on the characteristics of the dataset and research objectives. Then we apply the agglomerative hierarchical clustering algorithm to the housing dataset using the chosen linkage criterion. Finally, the stopping criteria for clustering, which is the number of desired clusters is specified.

Cluster Evaluation:

The quality of the clustering results are evaluated using internal validation method, the "within-cluster sum of squares"(WCSS), which is the parameters that are ought to be minimized through the clustering process. The characteristics of each cluster is analyzed and interpreted, identifying common traits and patterns that differentiate them.

Results:

The results are plotted on a map using folium

**3.1.2 Dimensionality reduction:**

Principal component analysis technique shall be used for dimensionality reduction.

**Principal Component Analysis (PCA):**

PCA is a statistical technique that employs an orthogonal transformation [17]. Its purpose is to transform a set of variables that are correlated into a set of variables that are not correlated with each other [18]. PCA can be utilized to analyze the interrelationships among a set of variables, making it suitable for examining their connections. This makes it a highly effective technique for dimensionality reduction [19].

Assuming we have a finite dataset p1, p2, ...., pm, which has n-dimension inputs, using PCA, n-dimension data can be reduced to k-dimension, (k n) [20] .

PCA is executed on the data in the following steps:

**3.1.2.1. Data Collection**

A comprehensive dataset of housing attributes such as location, size, amenities, rental prices, and neighbourhood demographics is gathered from Kaggle. Additional data on immigrant preferences, including factors like cultural proximity, accessibility to services, occupation, and proximity to immigrant communities are also acquired.

Data Pre-processing:

The raw data is arranged such that the variance of the data is equal to 1, and the mean is equal to 0. [21]

(3.3)

**3.1.2.3. Determining the co-variance matrix**

The co-variance matrix of the raw data is determined using

(3.4)

**3.1.2.4. Determining eigenvectors and eigenvalues**

Both the eigenvector and eigenvalue of the co-variance matrix are determined.

(3.5)

**3.1.2.5. Selecting the top K eigenvectors**

The top k eigenvectors of the covariance matrix are selected to project the raw data into a subspace with a dimensionality of k. These chosen eigenvectors will serve as the new basis for the data, replacing the original basis [22].

(3.6)

By following this approach, the original raw data, which is n-dimensional, can be transformed into a reduced representation of the data with a dimensionality of k. [23]

The developed models will be simulated, evaluated and presented using python programming language.

**3.2 LIMITATIONS OF THE METHODOLOGY**

Although the unsupervised machine learning approach for recommending accommodations for immigrants shows promise in tackling the complexity of the accommodation landscape, it is crucial to recognise a number of inherent limitations that could affect the reliability and generalizability of the findings. Firstly, the quality and accessibility of the input data are critical components that determine how well the process works. The geolocational, demographic, or cultural statistics may have biases, inaccuracies, or incompleteness that could contribute noise and affect the validity of the accommodations recommendations. Within recognised clusters, homogeneity is assumed by the clustering technique. Suboptimal suggestions may arise from reliance on broad cluster characteristics that obscure specific traits crucial to particular immigrant groups, even when neighbourhoods may actually show large internal heterogeneity. The approach can find it difficult to adjust to changes in neighbourhood dynamics over time. If not taken into account in a dynamic and ongoing updating process, gentrification, urban development, or demographic shifts over time may pose a threat to the stability of clusters and the applicability of suggestions.

Another significant limitation of the methodology is cluster interpretability and geographic variability. There may be issues with the interpretability of clusters produced by unsupervised machine learning techniques. Post-analysis interpretation may need to be strengthened in order to comprehend the significant distinctions across clusters and convert them into insights that policymakers or users can utilise to make decisions. Unsupervised machine learning algorithms may result in clusters that are difficult to interpret. To fully understand the notable differences within clusters and translate them into insights that users or policymakers may utilise to inform decisions, post-analysis interpretation might need to be reinforced. In order to improve the overall effectiveness and applicability of accommodation recommendations for immigrants using unsupervised machine learning, it is imperative to comprehend these limitations in order to accurately interpret the methodology's results and to direct future research endeavours to address these challenges.

**3.3 RESULTS COMPARISON**

The suggested unsupervised machine learning models are compared. This makes it easier to determine which strategy is best for the particular job of suggesting the most suitable accommodation for immigrants.

In conclusion, the goal of the results comparison section is to offer a thorough evaluation of the accommodation recommendation system from a variety of angles. Through a combination of quantitative algorithmic performance measures and qualitative user-centric feedback, the evaluation method aims to provide a comprehensive picture of how effectively the system meets the various demands of immigrant populations.

**3.4 SUMMARY OF THE CHAPTER**

The techniques employed in this research include clustering and dimensionality reduction. The clustering technique is executed by gathering comprehensive data on housing attributes and immigrant preferences. The collected data is cleansed, transformed, and standardized, with relevant features extracted. Statistical techniques are then used to identify the most relevant features for clustering. The K-means algorithm is applied to segment housing data into distinct clusters. The optimal number of clusters (K) is determined using the elbow method, and Euclidean distance is used to calculate data point centroids. The Agglomerative Hierarchical Clustering (AHC) is also applied to housing data, and linkage criteria are selected based on the dataset's characteristics. Stopping criteria specify the number of desired clusters. The quality of clustering results is then assessed using the "within-cluster sum of squares" technique. Common traits and patterns distinguishing clusters are identified. The clustering results are visualized on a map using boxplots.

Principal Component Analysis (PCA) is employed for dimensionality reduction. Using PCA, relevant housing and immigrant preference data are gathered and arranged so that the variance equals 1 and the mean equals 0. Then the covariance matrix of the raw data is determined. Both eigenvectors and eigenvalues of the covariance matrix are calculated afterwards. The top k eigenvectors are then selected to project the raw data into a lower-dimensional subspace. Locations are plotted on a map based on the reduced representation. Python programming language will be used to simulate, assess, and present the developed models. Based on data analysis and dimensionality reduction techniques, this complete methodology seeks to offer insightful conclusions and suggestions for immigrant accommodation. The results observations gotten from this study will be discussed in the next chapter.

**Chapter 4**

**RESULTS AND DISCUSSION**

**4.1 INTRODUCTION**

In this chapter, the results and learnings obtained from applying the suggested methodology will be explored. The integration of sophisticated machine learning algorithms with geolocational data analysis has ushered in a new age in the search for individualized, culturally appropriate accommodation choices. The findings are thoroughly examined in this section, which is followed by a detailed discussion aimed at interpreting the findings, their limits, and possible directions for further improvement and advancement.

**4.2 K-MEANS CLUSTERING**

Table 4.1 shows the raw, un-cleaned dataset of house rents for 20,132 vacant apartments in lagos gotten from Kaggle,



*Table 4.1 Uncleaned house rent dataset*

Table 4.2 shows the dataset, cleaned using python pandas. The duplicate and irrelevant data were removed. 52 districts were assigned codes and the data column was named district\_code. The district codes together with their respective district equivalent are represented below:

1 = abule egba

2 = agbara ijaiye

3 = agbara

4 = agege

5 = ajah

6 = alagbado

7 = alimosho

8 = allen avenue

9 = apapa

10 = badagry

11 = ejigbo

12 = festac

13 = gbagada

14 = idimu

15 = ifako

16 = igando

17 = ijora

18 = iju

19 = iju ishaga

20 = ikeja

21 = ikeja adeniyi jones

22 = ikeja gra

23 = ikorodu

24 = ikotun

25 = ikoyi

26 = irepo

27 = isolo

28 = isolo ago palace

29 = isolo jakande

30 = iyana ipaja

31 = ketu ikosi

32 = ketu ogudu

33 = ketu shangisha

34 = lagos island

35 = lekki

36 = maryland

37 = mushin

38 = ogba

39 = ojodu

40 = okeafa

41 = okota

42 = onikan

43 = oniru

44 = opebi

45 = oregun

46 = oshodi

47 = oshodi ajao

48 = oshodi mafoluku

49 = satellite town

50 = surulere

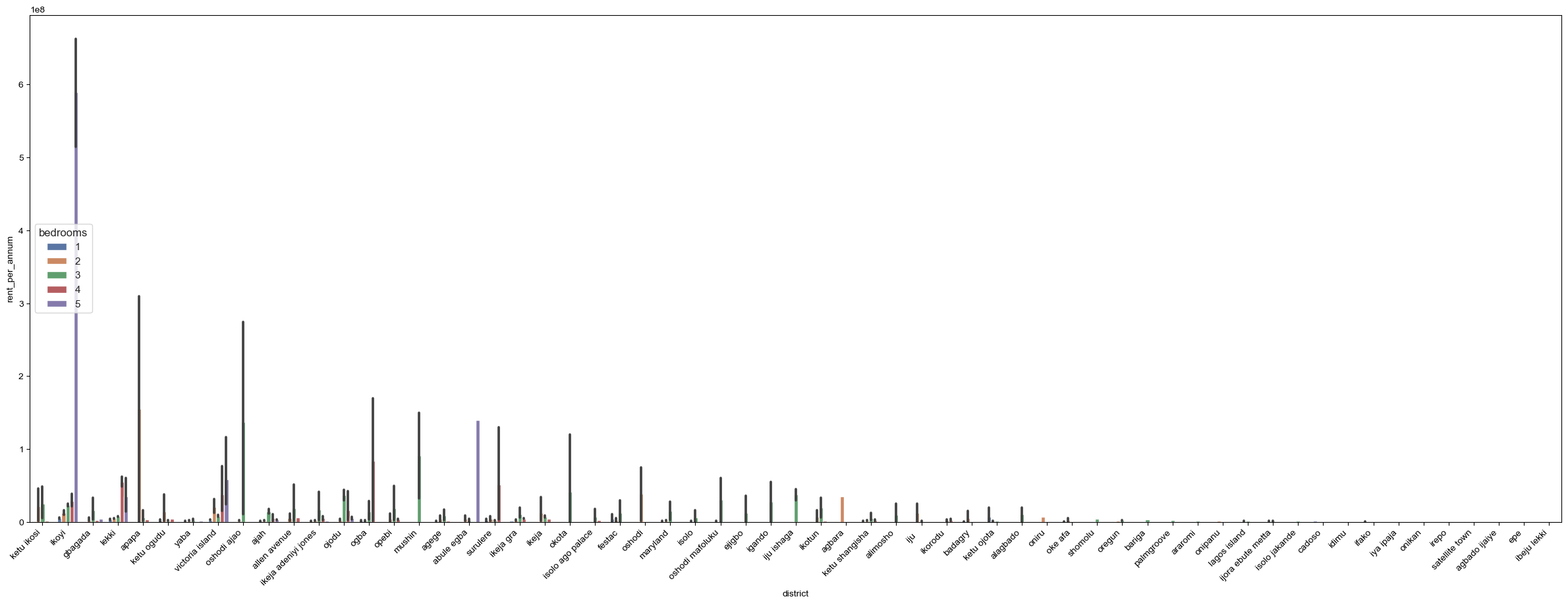
51 = victoria island

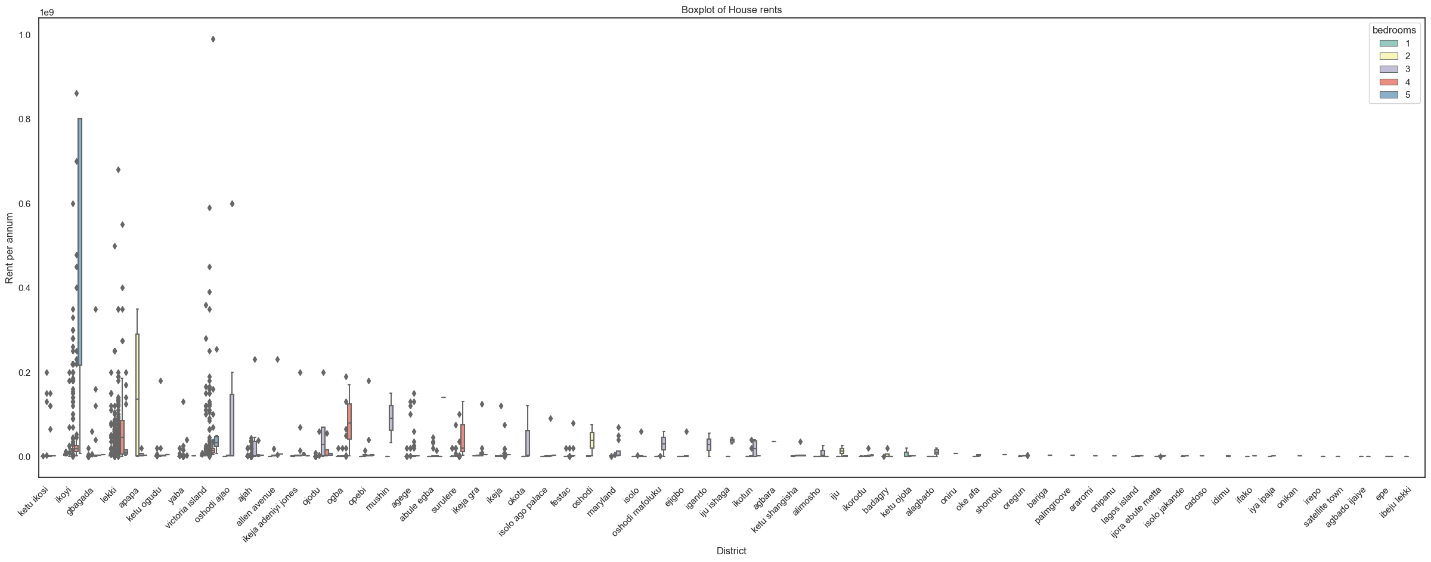
52 = yaba



*Table 4.2, Cleaned House rent dataset.*

The following figures, Figure 4.1 and Figure 4.2 show the data being visualised on a barplot and a boxplot respectively



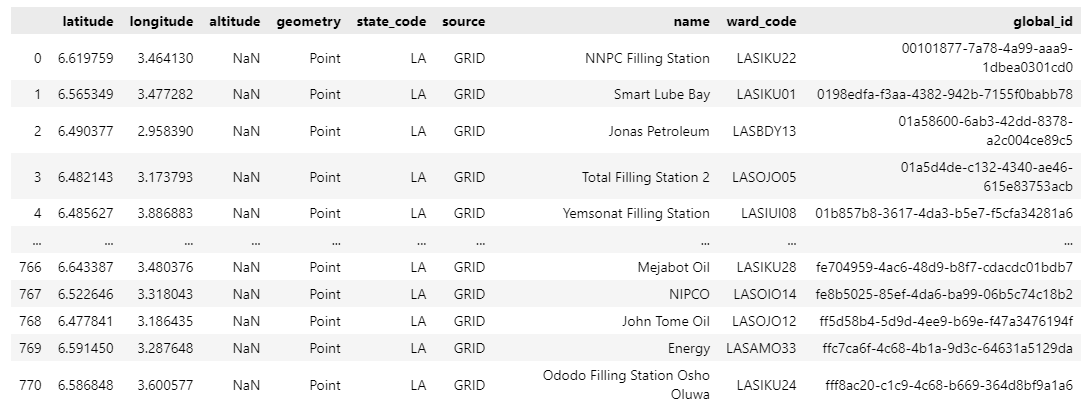


*Figure 4.1 & Figure 4.2, data plotted on barplots and boxplots respectively*

From the plots shown, it is evident that Ikoyi has the most expensive accommodation and Ibeju

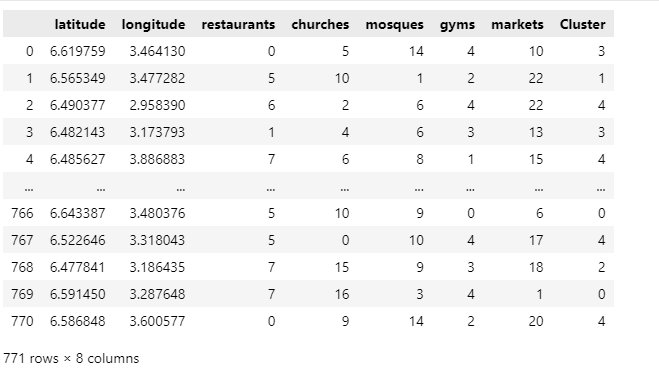
lekki has the least expensive accommodation.

Table 4.3 shows the cleaned lagos filling stations dataset obtained from kaggle.



*Table 4.3, cleaned filling stations dataset*

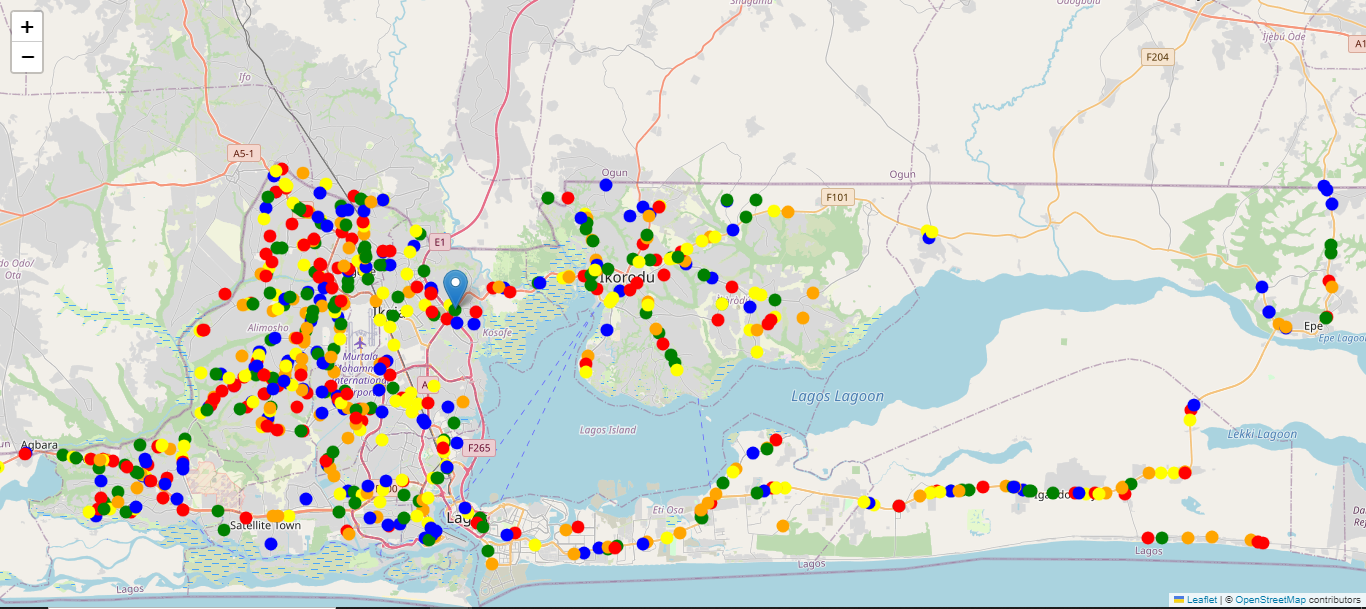
After the filling stations dataset was cleaned and using Ojota interchange as the reference point, (because it is technically at the center of Lagos map), the entire restaurants, churches, mosques, markets, and gyms within the 50 mile radius of each filling station were collected using the longitude and latitude of each filling station. The geolocational data of the aforementioned location was collected using foursquare API and each location was counted and clustered into five different clusters using k-means clustering. Table 4.4 shows the locational data after clustering.



*Table 4.4 locational data after clustering*

After the data has been clustered, the result of each clustering was plotted on the map of Lagos

Folium. This can be seen in Figure 4.3



*Figure 4.3 Results of clustering plotted on the map of lagos*

Fetching the first and last 5 data of each clusters, the following were observed:

**In the green cluster (cluster 0)**,

* churches > mosques 100%
* restaurants > gyms 90%
* mosques > markets 80%
* average accuracy: 90%

**In the Orange cluster (cluster 1)**,

* markets > churches > mosques 100%,
* churches > restaurants 100%
* restaurants > gyms 70%
* average accuracy: 90%

**In the Blue cluster (cluster 2)**,

* markets > mosques > restaurants 90%,
* markets > mosques > gyms 80%
* churches > mosques 100%
* average accuracy  90%

**In the Yellow cluster (cluster 3)**,

* markets > restaurants 80%
* churches > restaurants 80%
* churches > gyms 100%
* mosques > restaurants 90%
* average accuracy: 87.5%

**In the Red cluster (cluster 4)**,

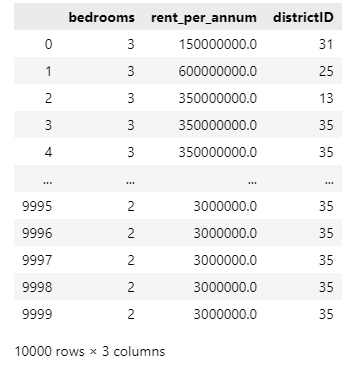
* markets > mosques > churches 100%
* restaurants > gyms 90%
* Average accuracy: 95%

Conclusion:

* overall average accuracy 90.5%
* Total Run time: 10 Seconds
* clusters with the most markets 1 and 4
* clusters with the most churches 0
* clusters with the most mosques 2 and 4
* clusters with the most restaurants 4
* clusters with the most gyms 3

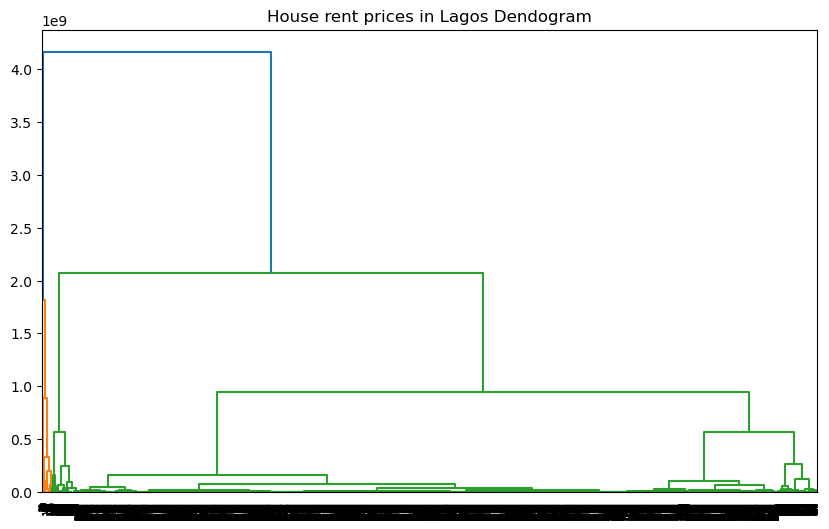
**4.3 AGGLOMERATIVE HIERARCHICAL CLUSTERING (AHC)**

Data were collected and cleaned as done in the previous section and the first 10,000 data were extracted as shown below. This was done as a result of multiple shut down of the clustering algorithm as AHC is not suitable for clustering extremely large datasets. The extracted data is shown below



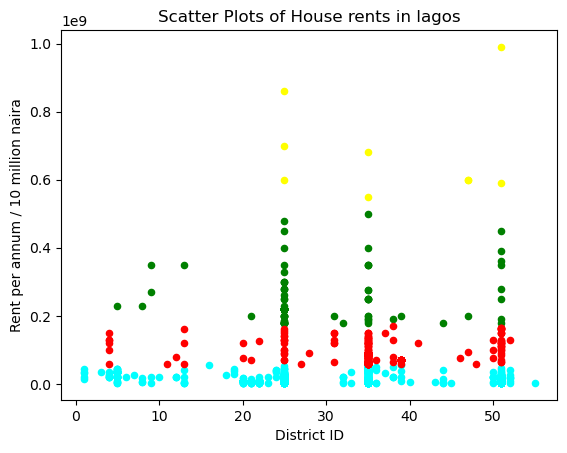
*Table 4.5 extracted data*

A dendogram was obtained after data was extracted which is shown below



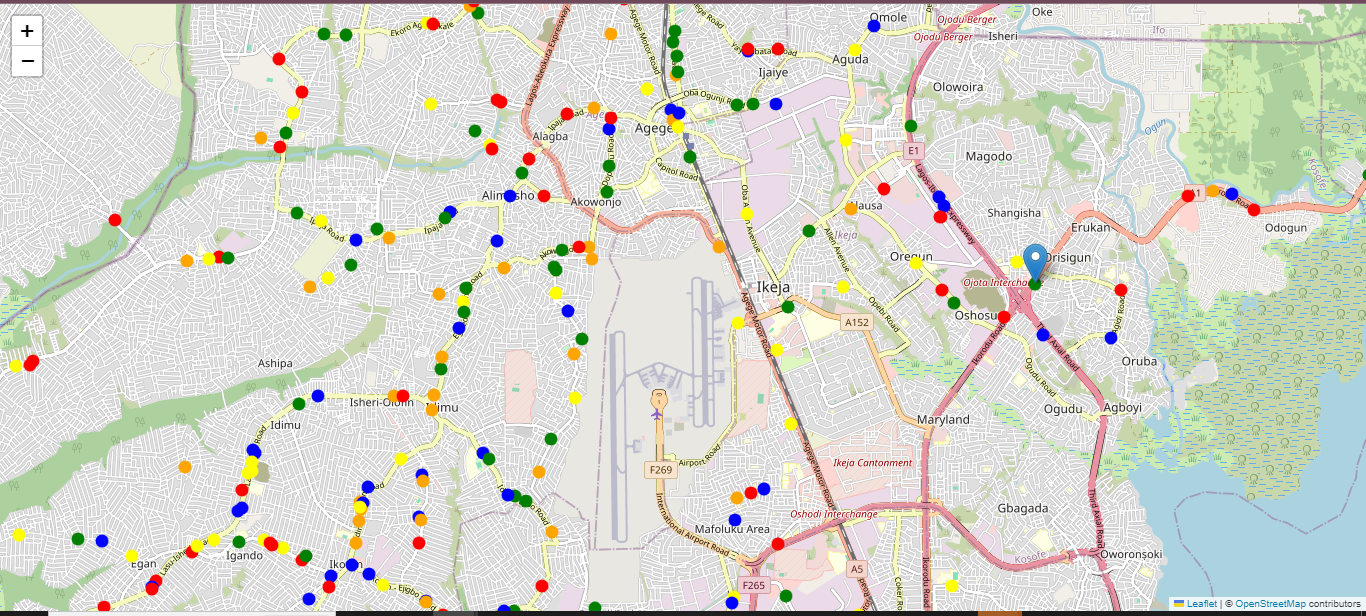
*Figure 4.4 Dendogram for extracted data*

Agglomerative Hierarchical clustering was then applied to the extracted data and the output is shown in the scatter plot below



*Figure 4.5 Scatter plots of agglomerative clustering*

Filling station data sets was cleaned as done in the previous section, and clustered using AHC. A similar result was obtained which were plotted on the map of lagos as shown below:

*Figure 4.6, agglomerative clustering Result plotted on a map*

Fetching the first and last 5 data of each clusters, the following were observed:

**In the green cluster (cluster 0)**,

* churches > mosques 90%
* restaurants > gyms 70%
* mosques > markets 70%
* average accuracy: 76.667%

**In the Orange cluster (cluster 1)**,

* markets > mosques > churches 70%,
* churches > restaurants 60%
* restaurants > gyms 90%
* average accuracy: 73.33%

**In the Blue cluster (cluster 2),**

* mosques > markets 90%,
* restaurants > gyms 80%
* mosques > churches 80%
* average accuracy  83.33%

**In the Yellow cluster (cluster 3)**,

* markets > churches > restaurants 100%
* churches > mosques 100%
* restaurants > gyms 80%
* average accuracy: 93.33%

**In the Red cluster (cluster 4),**

* markets > churches > mosques 60%
* restaurants > gyms 80%
* average accuracy 70%

conclusion:

* overall average accuracy 79.33%
* Total Runtime: 4 minutes, 45 Seconds
* cluster with the most markets 1 and 4
* clusters with the most churches 0
* clusters with the most mosques 3 and 4
* clusters with the most restaurants 4
* clusters with the most gyms 2 and 3

**4.3 PRINCIPAL COMPONENT ANALYSIS (PCA)**

Lagos house rent dataset was obtained and cleaned as in preceding sections. 100 mock dataset was created based on the following questions;

**1 .How much is your current housing budget?**

Numerical response

**2. How many bedroom accommodation do you prefer?**

Numerical response

**3. Do you work?**

responses

1-Full time

2- Part time

3- I don’t

**4. How often do you cook?**

Responses

1-Everyday

2- A couple of times a week

3- Whenever I can but not very often

4- I only help a little during holidays

5- Never, I don’t know my way around a kitchen

**5. How often do you leave home?**

Responses

1-Everyday

2-Few times a week

3-Once in a week

4- I barely leave my apartment in a month

**6. What is your religion?**

1- Christianity

2- Islam

3-Others

**7. How much do you dislike high pitched sound?**

Responses

1-I cant stand high pitched sound at all

2-i can tolerate a bit of sound if it’s good music

3-I don’t mind

4- I love it

**8. How often do you work out or exercise?**

Responses

1-Every day

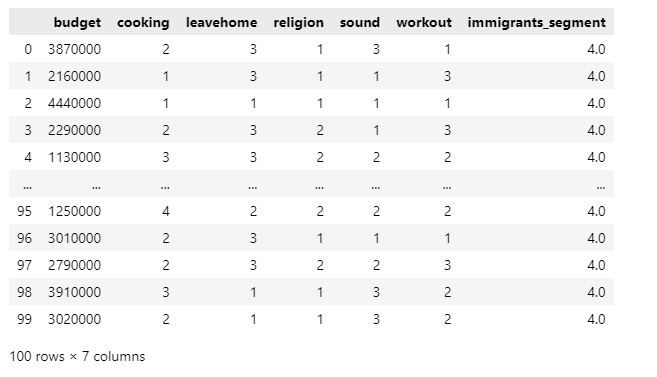
2-Few times a week

3- Once in a month

4-I don’t work out at all

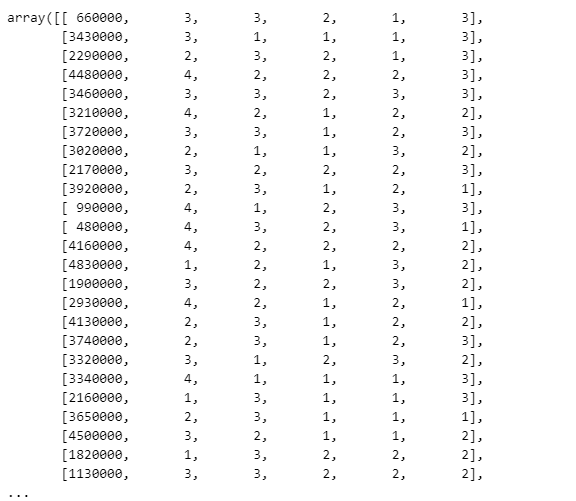
Data is obtained, cleaned and grouped into four segments; 1.0, 2.0, 3.0, 4.0, as

shown below;

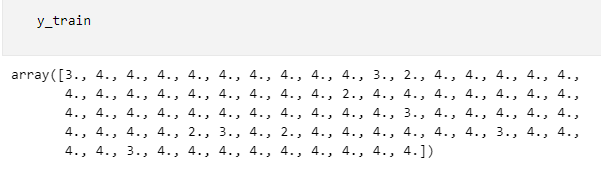


*Table 4.6 cleaned and segmented data*

Data is trained using PCA and grouped into two components namely x\_train and y\_train. The result of the trained data are shown below

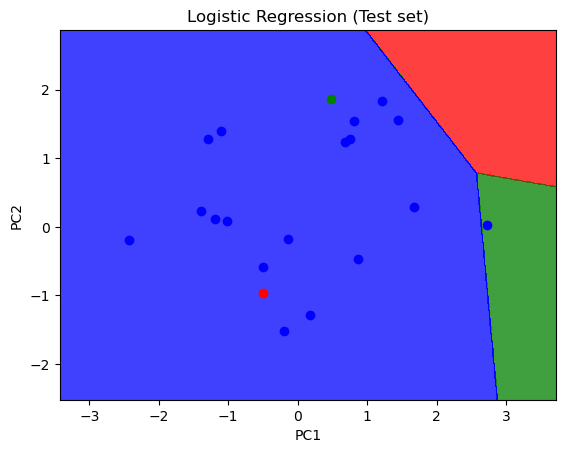


*Figure 4.7 x\_train*



*Figure 4.8 y\_train*

PCA is then applied again on the trained data and the final result is plotted as shown below



*Figure 4.9 PCA Plot*

Observation

There are only three incorrect predictions; the red dot showing in the blue region, the blue dot showing in the green region and the green dot showing in the blue region

Conclusion

* Accuracy of this analysis: 17/20 = 85%
* Total runtime: 6 seconds

**4.4 COMPARATIVE ANALYSIS**

From the examination of the algorithms used, accuracy, capacity, and speed of each algorithm are tabulated and displayed below

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Overall Accuracy | RunTime | Dataset |
| K-means clustering | 90.5% | 10 Seconds | 20,123 |
| Agglomerative Hierarchical Clustering | 79.33% | 4 minutes, 45 Seconds | 10,000 |
| Principal Component Analysis | 85% | 6 Seconds | 100 |

*Table 4.7 Examination table*

*Fig 4.10 Comparative analysis*

From the above analysis, K-means clustering has the highest accuracy, and highest data capacity.

Therefore, K-means clustering algorithm is the best technique for suitable accommodation

Recommendation.

**CHAPTER 5**

**CONCLUSION AND RECOMMENDATIONS**

**5.1 CONCLUSION**

Through the perspective of accommodation recommendation, this research has negotiated the complex confluence of unsupervised machine learning and geolocational data analysis in an effort to improve the settlement experience for immigrants. The major goal of this research was to use supervised machine learning techniques to increase the accuracy of accommodation recommendations for immigrants. Several significant findings and observations came from the inquiry. The process of defining the problem and putting a recommendation system in place is a big step in the right direction towards giving immigrant communities individualised, culturally appropriate housing recommendations. The use of unsupervised machine learning methods in the accommodation recommendation space has opened up new avenues for development. The system is positioned as a viable tool for addressing the specific needs of immigrants seeking acceptable accommodation because of its capacity to autonomously recognise patterns, establish clusters, and change dynamically.

In conclusion, the combination of unsupervised machine learning and geolocational data analysis has produced a recommendation system that might potentially herald in a new era of immigrant accommodation recommendations. This study aims to further the conversation on utilising technology to improve immigrant settlement experiences by thorough analysis, iterative improvement, and a steadfast dedication to user-centricity. Moving forward, the knowledge acquired paves the door for new developments in the field of accommodations recommendations, encouraging cultural congruence, inclusivity, and a more fulfilling sense of home for immigrant groups.

**5.2 RECOMMENDATION FOR FUTURE WORKS**

Although our methodology has advanced, it is important to recognise its limitations. The need for further improvement is highlighted by variations in data quality, cultural complexity, and algorithmic sensitivity. As the area develops, tackling these drawbacks and looking for ways to get better will add to the continuing conversation about creating more sophisticated recommendation systems. Subsequent research concerning the recommendation of housing for immigrants ought to delve into the incorporation of varied data sources. Incorporating data on language, cultural preferences, and local amenities may improve the recommendations' relevance and accuracy. A more thorough knowledge of the unique requirements and preferences of immigrant communities may result from this strategy. Also, Future studies should concentrate on integrating user feedback into the recommendation system in order to improve the user experience and satisfaction. Incorporating systems that enable immigrants to offer input on suggested accommodations might yield better insights for enhancing and optimising the algorithm. A recommendation system that is more user-centric and efficient may result from this repeated feedback loop.

**REFERENCES:**

[1] van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Berlin.ISBN 978-3- 642-19344-6. 2011 1

[2] Kim Cotzias, "The history of Geolocation: modern technology born from Ancient human interest". <https://www.geopointe.com/2017/07/11/geolocation-history/> 2017 1

[3] Siem Peters, "What is geolocation: History and use-cases"<https://theapicompany.com/blog/history-and-explanation-of-geolocation/> 2020 1

[4] Jack Dangermond: History of GIS , timeline of early history & the future of gis, <https://www.esri.com/en-us/what-is-gis/history-of-gis> 2023 1

[5] Laura Roccio, "Landsat science" ,<https://landsat.gsfc.nasa.gov/about/history/> 2

[6] "SPOT - Looking down on earth", [https://web.archive.org/web/20131006213713/http://www.cnes.fr/web/CNES-en/1415-spot.php 2013](https://web.archive.org/web/20131006213713/http://www.cnes.fr/web/CNES-en/1415-spot.php%202013) 2

[7] Thuy Mai, "Global Positioning System history", NASA TV, 2012 2

[8]Ph Rigaux, Michel Scholl and Agnes Voisard,

"Spatial databases: with application to GIS", Morgan Kaufmann, 2002

[9] Elizabeth Reid"A look back at 15 years of mapping the world", 2020 2

[10] Yuval Shavitt and NoaZilberman, "A Geolocation Databases Study",

[IEEE Journal on Selected Areas in Communications](journal/IEEE-Journal-on-Selected-Areas-in-Communications-0733-8716) 29(10):2044-2056

DOI:[10.1109/JSAC.2011.111214](http://dx.doi.org/10.1109/JSAC.2011.111214" \t "_blank), 2011

[12] Daniel Johnson, "Unsupervised machine learning: algorithm, types with examples", <https://www.guru99.com/unsupervised-machine-learning.html>, 2023 4

[11] P.B. Nagpal and P.A. Mann,

"Comparative study of density based clustering algorithms" 2011[11]

[12] Teng Li, AminRezaeipanah and Elsayed M. Tag El Din, "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement" Journal of king Saud University - Computer and Information Sciences, volume 34, Issue 6, part B, 2022

[13] Abiodun M. Ikotun, Absalom E. Ezugwu , Laith Abualigah, Belal Abuhaija , andJia Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data", Information Sciences, 2022.

[14] Kristina P Sinaga, Miin-Shen Yang, IEEE access 8, 80716-80727, 2020

[15] R Suwanda, Z Syahputraand E M Zamzami, "Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K", journal of physics conference series, volume 1556, 2020

[16] [OuafaAmira](/author/Amira%2C+Ouafa), [Jiang-She Zhang](/author/Zhang%2C+Jiang-She) &

[Junmin Liu](/author/Liu%2C+Junmin)"Fuzzy c-means clustering with conditional probability based K–L information regularization", Journal of statistical computation and simulation, Vol 91, 2021.

[17] Hai Lan, Chi Zhang, Ying-Yi Hong, Yin He and Shuli Wen

"Day-ahead spatiotemporal solar irradiation forecasting using frequency-based hybrid principal component analysis and neural network", Applied Energy 247, 389-402, 2019

[18] H. Abdi and L. J. Williams, "Principal component analysis", Wiley Interdiscipl. Rev. Comput. Statist., vol. 2, no. 4, pp. 433-459, 2010.

[19] G Thippa Reddy, M Praveen Kumar Reddy, KuruvaLakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, GautamSrivastava, Thar Baker, "Analysis of dimensionality reduction techniques on big data", Ieee Access 8, 54776-54788, 2020

[20]   
Huimin Zhao, JianjieZheng, JunjieXu and Wu Deng,

"Fault diagnosis method based on principal component analysis and broad learning system", IEEE Access 7, 99263-99272, 2019.

[21] Ian T. Jolliffe and Jorge Cadima, "Principal component analysis: a review and recent developments", Philos Trans A Math Phys Eng Sci. Apr 13; 374(2065): 20150202.doi: 10.1098/rsta.2015.0202, 2016

[22]NemaSalema and Sahar Hussein, "Data dimensional reduction and principal components

analysis", 16th International Learning & Technology Conference, 2019

[23]Salih Hasan, B. M. ., &Abdulazeez, A. M, "A Review of Principal Component Algorithm

for Dimensionality Reduction". A Review of Principal Component Analysis Algorithm for

Dimensionality Reduction. Journal of Soft Computing and Data Mining, 2(1), 20–30, 2021

[24]Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: self-supervised learning

via redundancy reduction. In Proc. 38th International Conference on Machine Learning (eds.

Meila, M. & Zhang, T.) 12310–12320 (PMLR, 2021).

[25] Ning Li, Martin Shepperd and Yuchen Guo, "A systematic review of unsupervised learning

techniques for software defect prediction", Information and Software Technology

Volume 122, , 106287, June 2020

[26] S. Ahmad, I. Ullah, F. Mehmood, M. Fayaz and D. Kim, "A Stochastic Approach Towards Travel Route Optimization and Recommendation Based on Users Constraints Using Markov Chain," in IEEE Access, vol. 7,

pp. 90760-90776, 2019, DOI: 10.1109/ACCESS.2019.2926675.

[27] Q. Xu, J. Wang, and B. Xiao, "Personalized location recommendation for location-based social networks," 2017

IEEE/CIC International Conference on Communications in China (ICCC), 2017, pp. 1-6, DOI:

10.1109/ICCChina.2017.8330459.

[28] R. Yue-Qiang, W. Ze, S. Xiao-Na and S. Shi-Min, "A Multi-Element Hybrid Location Recommendation

Algorithm for Location Based Social Networks," in IEEE Access, vol. 7, pp. 100416-100427, 2019, DOI:

10.1109/ACCESS.2019.2929313.

[29] M. Ludewig and D. Jannach, "Learning to Rank Hotels for Search and Recommendation from Session-based

Interaction Logs and Meta Data," in RecSys Challenge '19: Proceedings of the Workshop on ACM Recommender Systems Challenge, September 2019 Article No.: 5Pages 1–5https://doi.org/10.1145/3359555.3359561.

[30] K. Gao et al., "Exploiting Location-Based Context for POI Recommendation When Traveling to a New Region,"

in IEEE Access, vol. 8, pp. 52404-52412, 2020, DOI: 10.1109/ACCESS.2020.2980982.

[31] Karen Xie, Linchi Kwok, Chih-Chien Chen, and Jiang Wu, “To share or to access? Travelers’ choice on the types of accommodation-sharing services” Journal of Hospitality and Tourism Management, 2020

[32] Lee S, Kim D. "Distributed-based hierarchical clustering system for large-scale semiconductor wafers". 2018 IEEE international conference on industrial engineering and engineering management (IEEM); Piscataway. pp. 1528–1532.

[33] Catherine Withol de Wenden, "International Migration as a World Issue", [IMISCOE Research Series](/bookseries/13502) book series (IMIS),2023

[34] Basavesh D, Laharishree S, Sthuthi S, Tejaswini N , and  Vidya R, "Hostel Finder: Location-Based Recommendation System for Hostels and PGS with Transit Information" , International Journal of Research Publication and Reviews, Vol 4, no 5, pp 4482-4489 , 2023

[35] Aang Munawar, Gen Gen Gendalasari , I Made Gede Ariestova Kurniawan, D Purnomo, Nur Haris Ependi, Rulinawaty, Muhammad Isa Indrawan , Muhammad Sadri, "Cluster Application with K-Means Algorithm on the Population of Trade and Accommodation Facilities in Indonesia", Virtual Conference on Engineering, Science and Technology (ViCEST) 2020.

[36] R Ravi Kumar, M Babu Reddy and P Praveen, "A review of feature subset selection on unsupervised learning", Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 163-167, 2017

[37] Florian Hahne, Wolfgang Huber, Robert Gentleman, Seth Falcon, R Gentleman, VJ Carey, "Unsupervised machine learning", Bioconductor case studies, 137-157, 2008

Kumar V. and Mahapatra D, "A study of clustering techniques in customer segmentation", 2013

L.J.P. van der Maaten and G.E. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, 2008.

Dal Pozzolo, A., Boracchi, G., Caelen, O., & Bontempi, G, "Credit card fraud detection: A realistic modeling and a novel learning strategy", 2015

Radford, A., Metz, L., and Chintala, S, "Unsupervised representation learning with deep convolutional generative adversarial networks", 2015

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, 2003.

A. Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems (NIPS), 2012.

R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," Proceedings of the IEEE Symposium on Security and Privacy, 2010

B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," Proceedings of the Association for Computational Linguistics (ACL), 2004.

G. Linden, B. Smith, and J. York, "AmazonRecommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, 2003.

Xavier Amatriain, Justin Basilico, "Recommender systems in industry: A netflix case study", Recommender systems handbook, 385-419, 2015

Haneul Kim "Monolith: TikTok’s Real-time Recommender System", haneulkim.medium.com, 2022

Gao Huming, Li Weili, "A hotel recommendation system based on collaborative filtering and rankboost algorithm", Second international conference on multimedia and information technology 1, 317-320, 2010

Douglas Reynolds, "Gaussian Mixture Models", Encyclopedia of biometrics 741 (659-663), 2009

Sezai Tokat , Kenan Karagul, Yusuf Sahin, Erdal Aydemir, "Fuzzy c-means clustering-based key performance indicator design for warehouse loading operations", Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 8, Part B, Pages 6377-6384, 2022

R.Suganya, R.Shanthi, "Fuzzy C- Means Algorithm- A Review", International Journal of Scientific and Research Publications, Volume 2, Issue 11, 2012

Pal N.R, Pal K, Keller J.M. and Bezdek J.C, “A Possibilistic Fuzzy c-Means Clustering Algorithm”, IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530, 2005.

James Thurman and Urs Gasser"Three Case Studies from Switzerland: Smartvote", Berkman Center Research Publication, 2009.

Andreas Ladner, Jan Fivaz and Joëlle Pianzola, "Voting Advice Applications and Party Choice: Evidence From Smartvote Users in Switzerland", International Journal of Electronic governance, 2012

Raymond Sutjiadi, Edwin Meinardi Trianto and Adriel Giovani Budihardjo, "Surabaya Tourism Destination Recommendation Using Fuzzy C-Means Algorithm", Journal of Telecommunication, Electronic and Computer Engineering, 2018

Yang Lu1, Xuemei Bai and Feng Wang, "Music Recommendation System Design Based on Gaussian Mixture Model", 4th International Conference on Mechatronics, Materials, Chemistry and Computer Engineering (ICMMCCE), 2015

Dehariya VK, Shrivastava SK, Jain RC. Clustering of image data set using k-means and fuzzy k-means algorithms. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, pp. 386–391, 2010.

[59] Laurens van der Maaten, Geoffrey Hinton, “Visualizing Data using t-SNE”, Journal of Machine Learning Research 9 (2008) 2579-2605, 2008

[60] Shuangshuang Chen and Wei Guo, “Auto-Encoders in Deep Learning—A Review with New Perspectives”, Jiangsu Provincial Key Constructive Laboratory for Big Data of Psychology and Cognitive Science, Yancheng Teachers University, Yancheng 224002, China. 2023