ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 6, June 2022

Geolocation Analysis Using Machine Learning

[1] Sakshi Rajesh Sinha, [2] Prof. Sumedh Pundkar

[1] [2] Dept. Computer Science and Technology, Usha Mittal Institute of Technology, S.N.D.T, Santacruz(W), Mumbai, India.

Corresponding Author Email: [1] sakshisinha1501@gmail.com, [2] sumedhpundkar@gmail.com

***Abstract*— A new journey commences every time a student leaves his/her home for education or work thereby leading themselves to self-discovery and self-reliance. But along with new adventures comes various challenges such as unfamiliar health care systems, personal safety issues, financial problems, etc, but the major problem of them all is accommodation issues. Students and young adults often face difficulties when it comes to immigrating to new cities or states for pursuing higher studies from colleges or for work purposes. As different people have different priorities and interests, finding a suitable place that fits their budget and have easy access to their daily requirements for sustainable living becomes a challenge.**

The objective of this project is to create a system to find the best accommodation for the user in a particular city by classifying the user based on the preferences given by them such as budget, proximity to a certain location, daily necessities, etc. This system can be expanded and further be used for various purposes such as finding a suitable location for any business (for eg. restaurants/cafes or stationery shops are best suited near an educational institution) or for the best area of land for crop cultivation for maximum yield etc.

Index Terms – Machine learning, Residential Location, GIS Spatial Analysis, Recommendation Model, K-means Clustering

I. INTRODUCTION

The internet has become a huge source of information. One can search for almost anything and find valuable insights. The data uploaded on the internet is increasing tremendously with every passing second. Earlier we often used the internet for educational purposes, mainly for finding information. But due to increasing interaction with the internet on a daily basis, it has become a major influencing factor in our decision-making process.

From what to wear to what to eat and which movie to watch to where to meet, everything is decided using what the internet suggests us to do. But one must wonder how it works. The answer to this question is a recommendation system. This system takes into account one's likes and dislikes, preferences, and opinions on different things, daily activities on the internet, websites visited and time spent on each of them. Basically, it gets a general idea of who we are to give us suggestions by grouping similar people and analyzing what they prefer to do.

In India, the major reason for migration is either work / employment or education. These migrations mainly take place from developing states/cities to developed states/cities. As they are economically and educationally developed, they provide better education, better facilities, a higher standard of living, etc. Also, the migration from one developing state to another developing state can be seen because of renowned educational institutions such as IITs, NITs, etc. Hence the development of these states in one way or another depends on the migrants as they feed upon the unemployed and students. According to the Economic Survey of India 2017, the internal migration in India is nearly 9 million every year. Census states that 3% of total migration is for the purpose of education and covers the age group of 14-20 years of age and 14.7% of total migration is for the purpose of work/employment and covers the age group of 21-29 years of age. The migration does offer a large set of opportunities and a high scale of development that one might not find in rural areas, helping one to become self-reliant and leading them to self-discoveries, but as one embarks on this new journey of development and growth, along with it also come various challenges and impediments that one might face because of changing environment, culture, language, etc.

Some of the adversities are improper health care services, personal safety issues that one might have to deal with depend- ing on the neighborhood, especially finding a safer residence for women, and various factors resulting in financial problems depending on the cost of living as it differs from place to place depending on the locality, infrastructure, etc, income plays a main factor in decision making, daily expenditure, house rent, loans, etc. but most severe of them all is the accommodation issues faced by them. Finding a suitable place becomes difficult as one has to consider factors that suit their budget, have easy access to their daily requirements, proximity to a particular college/company, etc.

Hence there is a need for a personalized recommendation system that takes into consideration all the necessary parame- ters and searches intensively before recommending a suitable location for accommodation.

II. LITERATURE SURVEY

Recommendation systems have come into play to present the user with exactly what he had in mind while searching for an item or service [1]. The system takes into consideration the parameters input by the user so as to determine what he actually needs. Today we are surrounded by different types of applications and websites for fulfilling our daily needs and all these technologies are using the recommendation system in some way or the other, for example recommending similar types of music based on your previously created playlist or songs that you frequently listen, similar genres of movies that

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 6, June 2022

have been previously watched or searched for, and same goes for the books. The amount of information available on the internet is increasing with every tick of the clock hence providing the user a plethora of options to choose from, that's where the recommendation system comes in to help decide what to choose by taking account of the user's history and relevant suggestions, a bulk of information gets eliminated to present a few similar options.

Various studies and researches on decision-making strategies were done to analyze the methods that can be relied upon for efficient decision-making that helped many planners, students, and practitioners interested in the field of spatial location decision-making. Huff model is a type of spatial interaction model that measures the probabilities of customers based on gravity at every location [2]. Multiple regression analysis is a technique that implies factors affecting the sales of current stores would have a similar impact on stores located at another location, it is used mostly by restaurant chains, books, music, and home furnishing stores. The analog approach takes into consideration the aspects of trade and site area features to recommend a similar site [2]. The gravity model is used at 2 ends of the spectrum, where the first spectrum considers the number of sales and probability of shoppers and the second spectrum focuses on distance as a major factor for site location [2]. Neural network is a computational model that simulates functional aspects and processes information with help of a connectionist approach and is considered as a robust classifier. Machine learning is mainly used for traditional techniques like geostatistics and is widely used for spatial application and GIS [2]. Multi-criteria decision-making model is considered as an important tool to solve complex business solutions as they tend to deal with uncertainty, complexity, and conflicting objectives [2].

There are different technologies used to construct the recom- mendation system. Collaborative filtering aggregates the user's ratings and object's previous recommendations to acknowl- edge similarities between them and create a new recommen- dation based on those comparisons. It is also known as the word of mouth" approach [3]. It utilizes the rating information obtained from the evaluation made by various users on various items [4]. The algorithm in CF gives suggestions based on the user's previous point of interest and the opinions of other like-minded users. For that, they check the rating score on a number scale and navigate through the past purchase records or mine websites that were visited recently [1].

It is further divided into 2 methods: the memory-based method that exploits the user database completely to provide suggestions and the model-based method that fits a model on the user's database to make predictions [3].

Content-based filtering does not use people's opinions, instead recommends items or services based on its description and profile of the user [4]. It learns the profile of the user to understand his interest based on the features of the items or services that the user has liked. It is keyword specific and so the algorithms used are liked in the past or currently searched by the user. For example, consider that the attributes of movies are genre, ratings, actors, directors, box office performance, etc. now based on the profile created by the users where they input the parameters that they like or prefer, the system recommends movies to them based on that.

Knowledge-based filtering provides suggestions based on user's needs and preferences [6]. It does not focus on making generalizations about the users. It is also known as "case-based systems" [3]. According to studies, this approach is especially useful where the items or services are not purchased as often. Those instances include items such as real estate, automobiles, financial services, expensive luxury goods, etc as adequate reviews or ratings might not be available as they are not frequently purchased [1]. They rely on specific domain knowledge and try comparing items' specifications with user's requirements to predict if the item is useful or not.

And lastly combining any of these approaches in a particular fashion that suits the desired results is known as a hybrid system. It follows a blended approach in an attempt to use the advantages of one system to fix the disadvantages of the other [3].

Research has shown various issues and challenges that are faced by the recommendation system. Usually, maximum number of users do not rate most of the items thereby leading to a very sparse rating matrix. This arises the data sparsity problem that reduced the chances of grouping users with similar ratings. This is a major drawback of CF filtering. Cold- start problem arises when a new item and a new user are considered as a case. To elaborate further a new item that is recently added is bound to have no ratings and hence can't be recommended to users initially and similarly, a new user can't be categorized initially without having any user's past preference. One of the major issues faced by recommendation systems is the scalability problem of the algorithm with large datasets. When these algorithms are applied to relatively small datasets, they provide the best results but it becomes rather challenging to handle the large, dynamic data sets created on the basis of users' preferences, items specifications, and inter- action between both these entities. For the system to provide quality personalized recommendations, it aims to gather as much information as it can through users' data and exploit it to the fullest [3]. But at the same time, it raises concerns about users' privacy and creates a negative impression as the system knows too much about them. Hence there is a need to design these systems meticulously so as to carefully use the data extracted and ensure that the user's privacy is also maintained.

A lot of technical and scientific literature has been reviewed to understand the pros and cons of recommendation systems, as each of them has their own set of strengths and weaknesses. It is necessary to compare them in order to find a suitable method for the said system. CF suggests items based on similar likes and preferences but works poorly for users whose preferences do not agree or disagree consistently with other user groups. Challenges faced by collaborative filtering are data sparsity, scalability, cold-start, etc. Even

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 6, June 2022

content-based filtering faces a cold-start problem as they need to collect ad- equate ratings to provide quality suggestions [5] [3]. Whereas knowledge-based doesn't face any cold start, ramp up, or data sparsity issues. They are best for casual exploration as they can provide a wide range of recommendations depending on the knowledge base. Hence after comparing all the methods, we observe that knowledge-based filtering has more set of strengths such as being sensitive to user's preferences and mapping user's needs to products, items and a very few weaknesses such as static suggestion ability and need for knowledge engineering thereby fulfilling requirements needed to build the said system [3].

III. METHODOLOGY

The flow chart describes the ideology behind the building of the recommendation system where necessary parameters are selected to train the model which precisely describes the user's daily requirements and preferences. The data thus fetched undergoes the cleaning and visualization stage for better analysis. After that, the algorithm is applied to the clean data. The model is then connected with the Foursquare API so as to extract relevant geolocation information according to the user's preferences. The results thus obtained are displayed on a map.

To further elaborate:

1. Fetch the dataset from the relevant location. Setting up an environment for data analysis using jupyter notebook. From an ocean of parameters selection of the relevant and essential parameters (income, proximity to college or office, etc.).
2. Cleaning and preparing dataset for analysis (Using Pan- das).
   The data collected via survey is raw and not fit for analysis as it contains various anomalies that may further lead to improper and unexpected results. Hence data cleaning is important. The process of cleaning includes
   - Removal of irrelevant and duplicate data,
   - Handle missing data,
   - Removing improper data to improve accuracy and
   - Fixing structural errors such as typos, naming conventions, incorrect capitalization.

   Here all the irrelevant data is removed or corrected to create a well-defined dataset. The most relevant features are then extracted into pandas data frame.
   ' xi vj ' is the Euclidean distance between xi and vj. 'ci' is the number of data points in ith cluster.
   'c' is the number of cluster centers.
   To calculate K-Means:
   Random value of cluster center is selected.
   Distance is calculated between each data point and cluster centers.
3. The data point is assigned to the cluster whose distance from the cluster center is minimum.
   New cluster is recalculated using:Visualizing the data using boxplots.
   After collecting the data, graphs are used to understand and visualize them. General study of trends followed by the useful in comparing distributions between several groups and datasets.

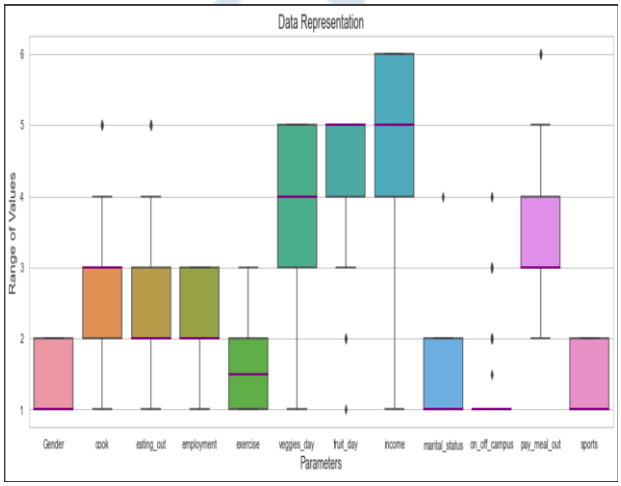

Fig. 1. Data Visualization Using Boxplots.

4. Clustering the data using K-means clustering.

K-Means clustering algorithm is implemented on the dataset of the population to categorize them into separate groups on the basis of some similarities. It is one of the simplest unsupervised learning algorithm used to train the ML model when dealing with unlabelled data. It assigns data points to clusters such that there is minimum sum of squared distance between data points and the cluster center. The less the variation the more the similarity among data points in a cluster. This algorithm aims at minimizing the objective function called the sum of squared errors that is given by:

where,

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (||xi\text{-}vj||)$$

population, based on surveys. Boxplot is a graph that indicates how the data is spread out as it takes up less space, it is

where,

$$vi = \left(\frac{1}{ci}\right) \sum_{j=1}^{ci} (x)$$

'c_i' represents the number of data points in i^{th} cluster.

1 Then the distance is recalculated between each data point and new cluster center.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 6, June 2022

2 This process is continued until the cluster centers don't change positions anymore.

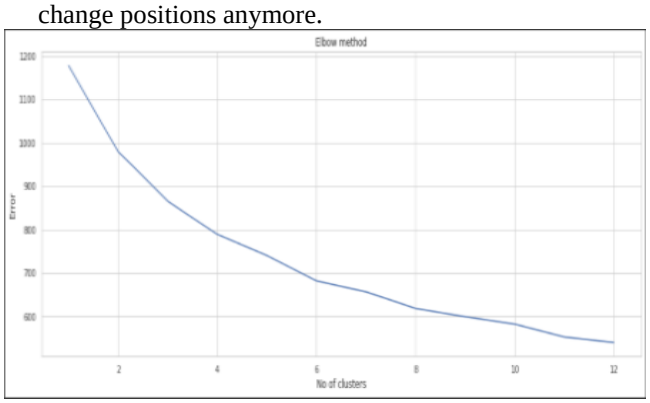

Fig. 2. Sum of Squared Errors.

Principal component analysis is an unsupervised machine learning algorithm used for dimensionality reduction. With the help of orthogonal transformation, it covers correlated features into linearly uncorrelated features. These new features are called principal components, they remove noise by reducing large no. of components down to a few important ones.

To Calculate Principal Component Analysis:

1. Each feature is standardized to have a mean of 0 and a variance of 1.
2. A covariance matrix is obtained by computation so as to show pairwise correlation between each feature.
3. Eigen vectors and their eigen values of the covariance matrix are calculated.
4. Top N eigen vectors are selected to become N principal components.

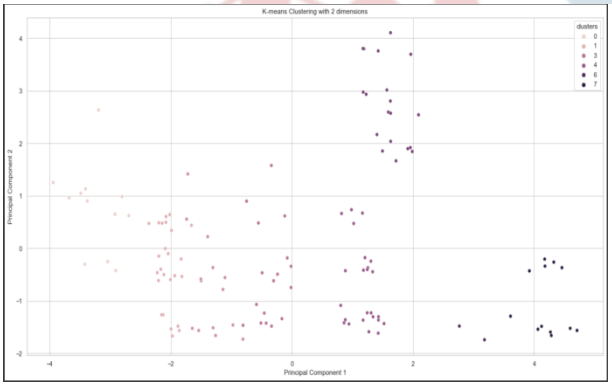

Fig. 3. K-Means Clustering Algorithm.

5. Fetch Geolocational data from Foursquare API.

An account is created on Foursquare API, where important credentials are established, necessary for creating a connection between the model and the API. Geolocation data is fetched from Foursquare API to find accommodation. The given address gets converted into their respective geo-coordinates (latitude and longitude) and is sent to the REST API. The query is set in such a way that residential locations can be found in a fixed radius around a selected point.

	name	categories	distance	latitude	longitude	country	formatted_address	postcode	region	neighborhood	fsq_id
0	Regal Arcade	Residential Building	286	19.273565	72.860287	IN	Ponam Sagar Complex (mira road), Mumbai, Mahär...		Mahārāshtra	NaN	4e47d43b7d8b91a065a12457
1	Poonam Sagar Complex	Residential Building	364	19.273686	72.861291	IN	House No.L/43 To L/46, J47/48, Mira-Bhayandar ...	401107	Maharashtra	[Mira Road East]	4f8948ede4b0ab5c87c0fe0e
2	Shantinath Empress	Residential Building	456	19.271847	72.861761	IN	Poonam Sagar Complex Mira Road East, Mira-Bhay...	401107	Mahārāshtra	[Mira Road East]	4deb56844c00071b83a8fd0
3	Happy Home	Residential Building	633	19.281481	72.867723	IN	Shanti Park Mira Road East, Thāne 401107, Mahā...	401107	Mahārāshtra	[Mira Road East]	4d8ef95eca75b60c821b0ba9
4	Asmita Ascon Acres IV	Residential Building	774	19.283585	72.857546	IN	Station Road, Srikanath Dhadway Road, Mira Roa...	401107	Maharashtra	[Mira Road East]	4e89bd7f82310042c1c2a4d0
5	Sector-8	Residential Building	828	19.273321	72.866782	IN	Shanti Nagar Sector 9, Mira Road East, Mira-Bh...		Mahārāshtra	[Mira Road East]	4ed529a88231c75eddeeb194

Fig. 4. Geolocation Data from Foursquare API.

Now for each location total no. of amenities such as no. of hospitals, restaurants, gyms, grocery stores, fruits and vegetable stores, drug stores, etc. located in the vicinity is extracted from Foursquare API. The query for each amenity is set differently considering their proximity to the location

	Latitude	Longitude	Restaurants	Hospitals	Fruits&Vegetables	Groceries	Gyms	Drugstores
0	19.283585	72.857546	21	47	0	1	11	2
1	19.285228	72.861562	33	45	1	3	16	3
2	19.287638	72.858093	29	43	0	1	10	1
3	19.283670	72.863993	41	47	1	3	17	5
4	19.281481	72.867723	48	47	1	3	12	6
5	19.289491	72.858225	26	43	0	1	9	0
6	19.273565	72.860287	18	50	1	3	8	1
7	19.273686	72.861291	21	50	1	3	9	1
8	19.271847	72.861761	16	50	1	3	8	1
9	19.286562	72.865957	39	42	1	3	20	3

Fig. 5. Amenities for each location.

6. Clustering the location using K-means clustering.

The resultant data of locations goes through K-means cluster- ing to find the best location for accommodation and results are plotted on the map using Folium.

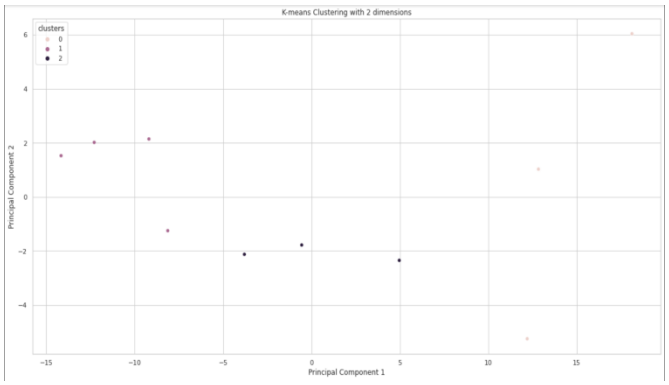

Fig. 6. K-Means Algorithm on Geolocation data.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 6, June 2022

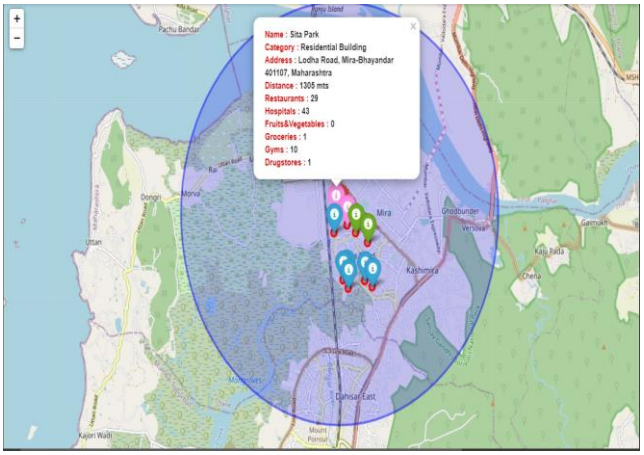

Fig. 7. Plotting Results on a Map.

IV. CONCLUSION

The literature survey helped us understand how machine learning is majorly used for geostatistics and spatial application. Also by comparing the pros and cons of different techniques used by recommendation systems, it was concluded that knowledge-based filtering is best suited for housing rec- ommendation systems. To build this system an environment was set up using Jupyter Lab. Pandas library was used for data cleaning. Boxplots helped in providing a visual repre- sentation of data for better analysis. K-Means Clustering and Principal Component Analysis algorithms were used for model training. This ML model was connected to Foursquare API from which residential locations in a fixed radius along with basic amenities were extracted using an HTTP library called requests. Lastly using Folium the results were presented on a map.

V. FUTURE SCOPE

This model can be expanded into a website or mobile application. It can also be further used for various purposes such as finding suitable locations for businesses. For example restaurants, cafes, and stationery shops are best suited near educational institutions as they are likely to make more profit there, or it can also be used to find the best area of land for the cultivation of different crops for maximum yield by taking into consideration type of soil, weather conditions, etc. of that particular place.

VI. ACKNOWLEDGMENT

I feel immense pleasure to express my gratitude to Prof. Sumedh Pundkar, without his guidance and constant motivation this project wouldn't have been a success. His creative suggestions and enthusiasm to build something new proved to be an inspiration throughout the process. Also, I would like to thank my project supervisor Prof. Narendra Gawai as his attention to detail helped me improve a lot. Through this project, I learned a lot which I hope will be useful for me in my future endeavors. Finally, I would like to thank my family and friends for their constant support and motivation.

REFERENCES

[1]. O. Aboulola, "A literature review of spatial location analysis for retail site selection," Journal of the Association for Information Systems, 08 2017.

[2]. Ojokoh, Bolanle, O. Catherine, Olayemi, Babalola, Asegunloluwa, Eyo, and Eyo, "A user-centric housing recommender system," Information Management and Business Review, vol. 10, pp. 17–24, 09 2018.

[3]. B. Kumar and N. Sharma, "Approaches, issues and challenges in recom- mender systems: A systematic review," Indian Journal of Science and Technology, vol. 9, 12 2016.

[4]. Jun, H. Jong, Kim, J. Hee, Rhee, D. Young, Chang, and S. Woo, ""seoulhouse2vec": An embedding-based collaborative filtering housing recommender system for analyzing housing preference," Sustainability, vol. 12, no. 17, 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/17/6964

[5]. Sakthivel, M., J. Udaykumar, and V. Saravana Kumar. "Progressive AODV: A Routing Algorithm Intended for Mobile Ad-Hoc Networks." *International Journal of Engineering and Advanced Technology (IJEAT), ISSN*: 2249-8958, Vol.9 no.2, PP: 70-74, 2019

[6]. Shishehchi, Saman, Banihashem, Seyed, M. Zin, N. Azan, M. Noah, and S. Azman, "Ontological approach in knowledge based recommender system to develop the quality of e-learning system," Australian Journal of Basic and Applied Sciences, vol. 6, pp. 115–123, 02 2012.