**RESEARCH ARTICLE**

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 08, Issue 02 February

Check for updates

http://dx.doi.org/10.47392/IRJASH.2026.008

# Entre Locate: An Intelligent Business Location Recommendation System Based on Spatial Analytics and Data-Driven Decision Support

*Mrs. S Subasree[1], Ms. Ashwini R[2], Ms. Janani V[3], Ms. Sowmiya S[4]*

*[1]Assistant Professor, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India*

*[2,3,4]UG Student, Department of Computer Science and Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India*

***Emails:*** *subashiva@gmail.com[1], ashwinirbtechcse@gmail.com[2], jananivbtechcse@gmail.com[2], sowmiyas3008@gmail.com[2]*

**Article history**

*Received: 23 December 2025*
*Accepted: 27 January 2026*
*Published: 26 February 2026*

***Keywords:***

*Business Location Recommendation, Spatial Analytics, Geospatial Data Mining, K-Means Clustering, Decision Support System.*

**Abstract**

*Choosing the right location is one of the most important decisions for the success of a new business. Many entrepreneurs still rely on manual surveys and personal judgment, which can be time-consuming and inaccurate. This paper presents EntreLocate, a smart business location recommendation system designed to support entrepreneurs with data-driven insights. The system uses real-world location data to analyze existing business patterns and identify areas with high business potential. After cleaning and organizing the data, similar locations are grouped to highlight promising zones for new ventures. The results are displayed using interactive maps, making them easy to understand and interpret. In addition, EntreLocate includes a profit estimation feature that considers rental costs and other expenses to help users evaluate financial feasibility. Experimental results show that the proposed system effectively supports business owners in selecting suitable and profitable locations, offering a practical and scalable solution for modern business planning.*

## 1. Introduction

The quick expansion of urban areas has turned the decision on where to locate your business into a very important and difficult decision. In the past, business owners have often chosen locations based on manual surveys and their own judgment—methods that are definitely very time-consuming, expensive, and also susceptible to subjective biases [4]. Still, the fate of a business depends very much on the location-related factors such as the number of potential customers and the distance to competitors. Due to the availability of massive amounts of location-based data, data-driven approaches can now provide a better solution [1], [3]. The system EntreLocate that we present here, combines geospatial analytics and business knowledge to work where traditional methods fail. OpenStreetMap API is used to gather data, and K-Means clustering, an unsupervised learning method, is tested to be efficient in detecting high-potentials business areas [6, 8]. Embedding the analytical model with the interactive map visualization, EntreLocate guides small business owners to the right location decision. Besides spatial suggestion, EntreLocate has also a profit estimation module to

evaluate how rental cost, operational expenses, and revenue calculated from cluster features may affect the entrepreneur's profit.

## 2. Objective

This project is aimed at developing and implementing an intelligent and data-driven recommendation system for business locations that will help entrepreneurs in finding the right spots for new businesses. It intends to lessen the weight of manual surveys and subjective decisions by a method that combines the use of geospatial analytics and real-world venue data [1], [3]. Locally relevant information gathered through OpenStreetMap API is handled in such a way that it is spatially accurate and dependable [6]. In order to raise the quality of the data, techniques such as noise removal, outlier detection, and category generalization are implemented to the data [7]. The system uses an unsupervised learning technique with the help of the K-Means clustering algorithm, to study the lifestyle patterns of the neighborhoods and henceforth to identify the similar business zones in the form of clusters [8]. The most suitable number of clusters is picked through the use of the validation methods such as the Elbow Method and the Silhouette Score [9], [10]. The end result is expected to be a set of location recommendations that are accurate, interpretable, and feasible, and supported by interactive visualization that would allow entrepreneurs to take data-backed, cost-effective business decisions in store placement studies [14], [15].

## 3. Literature Review

In [1], Sonia Khetarpaul et al. presented a clustering-based framework for business site selection using urban spatial data. The study compared K-Means, Hierarchical Agglomerative Clustering, and DBSCAN to identify suitable restaurant locations in New Delhi. However, this approach was limited by its dependence on the accuracy of API-derived location data. In [2], Ashok Kumar P introduced a FindKeywordMissing Algorithm that uses geographical demand to suggest places when a user enters a query in the search log (e.g., "find best gym near me"). The goal of this strategy is to use search trends to identify unmet business need and user intent. In [3], Goushuai Zhao proposed a location recommendation framework based on multi-source urban big data, integrating geospatial analysis and Haversine distance for spatial evaluation. While the multi-factor fusion improves recommendation accuracy, the framework becomes increasingly complex due to heterogeneous data integration. In [4], Han et al. proposed prescriptive analytics–driven restaurant location recommendation framework based on spatial data mining. The major strength of this work lies in its prescriptive focus and practical validation using real datasets. However, the approach is primarily domain-specific to the restaurant industry, limiting its generalizability to other business types. In [5], Iranzad and Liu reviewed Random Forest–based feature selection techniques, emphasizing their robustness and effectiveness for high-dimensional and noisy datasets, thereby improving model accuracy and interpretability.

## 4. Proposed Approach

The proposed EntreLocate system is a smart company location suggestion platform that facilitates entrepreneurs to pick the most suitable sites for their new ventures through decision-support logic, knowledge from unsupervised machine learning, and geospatial data analytics.

### 4.1. Data Collection

The information/facts utilized for writing this project were obtained through reliable open data sources. One major source is the OpenStreetMap API that delivers comprehensive and up-to-date location-based data [6]. It allows the system to associate business categories, customer check-ins, and neighborhood activity patterns, which is the foundation of its analysis and recommendation of the best business locations. Urban site selection has heavily relied on POI-driven business analysis [12], [13].

**Table 1 Data Set Sample Columns**

| Column Name | Description |
|---|---|
| Name | Name of the venue |
| Rating | Ratings of the venue |
| Address | Address of the venue |
| Latitude | latitude of the current venue |
| Longitude | longitude of the current venue |

Through the frontend interface, user inputs such as the city name and the business type preferred are collected and sent to the backend where API

requests are dynamically created to fetch the relevant spatial data.

### 4.2. Data Processing

The data fetched via the OpenStreetMap API is first subjected to a preprocessing step, which basically takes care of and gets rid of null values and missing entries. Then, the needed properties are extracted. Locational anomalies are detected with the help of latitude–longitude bounding boxes indicated to the present area in TABLE 2 and the Median absolute deviation technique. The venues outside the specified geographic limits are removed so as not to affect the clustering results adversely [7]. Table 3 shows Generalizing Venue

**Table 2 Generalizing Venue**

| Specific Category | Generalized Category |
|---|---|
| Bakery | Retail |
| Grocery | |
| Supermarket | |
| Pharmacy | Health & Medicine |
| Hospital | |
| Optician | |

Besides, a category generalization step as shown in TABLE 3 is carried out, where the layers of shop types (e.g., bakery, supermarket, grocery) are illustrated as one of the broader categories such as retail. This normalizing step lessens the redundancies, increases the clustering accuracy, and makes the business type comparison more meaningful.

### 4.3. Exploratory Data Analysis

Even before deciding to use clustering as a technique, the data on business venues goes through a round of exploratory data analysis and from the study of business venues, suggestions for new business location can be derived from clustering technique too.

**Table 3 Latitude Longitude Bounding Boxes**

| Venue | Latitude | Longitude | Bounding Box Range | Status |
|---|---|---|---|---|
| Shop A | 12.9345 | 79.1342 | 12.80–12.99, 79.10–79.20 | Within Range |
| Shop B | 13.4521 | 78.2451 | 12.80–12.99, 79.10–79.20 | Outlier |
| Restaurant C | 12.8765 | 79.1852 | 12.80–12.99, 79.10–79.20 | Within Range |
| Gym D | 14.0254 | 80.1356 | 12.80–12.99, 79.10–79.20 | Outlier |

#### 4.3.1. Venue Data Distribution

Figure 1.1 shows the geographical spread of venues in the study area. There is a clear difference in the number of venues where the downtown areas are seen to be crowded with venues while the suburban areas have lesser venues. This difference in the number of venues in the different parts of the city shows the activity of the city at different locations and hence is a good motivation to the clustering technique to find the two kinds of business zones, dense and sparse. FIGURE 1.2 illustrates the cluster-wise distribution of data points after the clustering algorithm has identified the clusters. The number of venues in each cluster lies within the range of one to six. As an example, the clusters around indices 8, 11, and 16, are relatively dense. The existence of many low-density clusters signifies the spatial heterogeneity of the dataset and at the same time, points to the algorithm's capability to detect different levels of venue concentration. To conclude, the venues data exhibit uneven geographical distribution and variation in density which supports the choice of density-aware clustering for the case of business location analysis

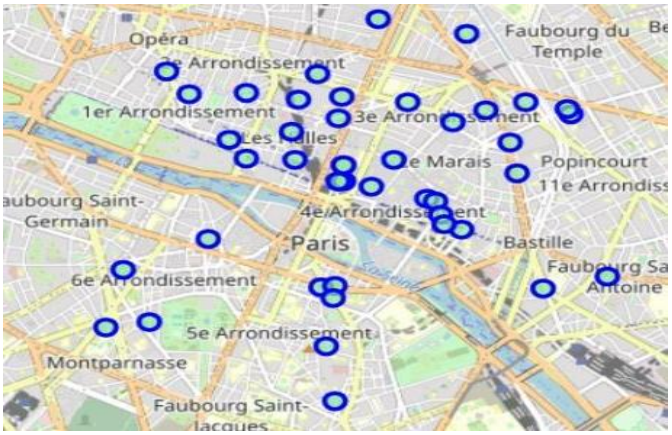


**Figure 1 Geospatial Distribution**

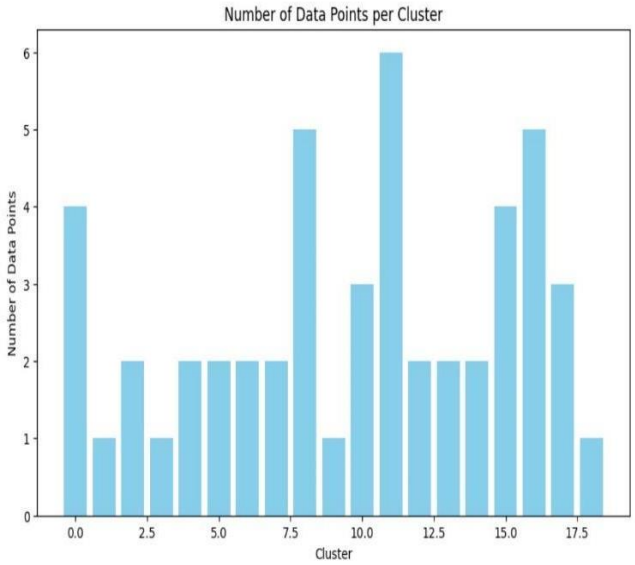


**Figure 2 Data Distribution of Venues**

### 4.3.2. Venue Rating Analysis

Venue ratings obtained from OpenStreetMap-based location data reveal differences in service quality and customer satisfaction among various regions. We can note from FIGURE 1.2 that highly rated commercial venues are mostly located in city centers, where there is a strong demand, easy accessibility, and efficient service delivery, whereas lower ratings are more frequently found in suburban areas. The combination of rating evaluation and spatial clustering allows the platform to focus on locations with high customer traffic that also receive positive feedback for giving business recommendations that are spot on.

This diversity means that some clusters correspond to comparatively smaller and more focused segments of the data, whereas other clusters point at more distinct patterns.

### 4.3.3. Outlier Inspection

Outlier detection was carried out to ensure the data points were valid, refer to TABLE 2. The venues which fell outside the study's geographical boundaries were regarded as outliers based on the bounding box limits of latitude and longitude. Shop B, for example, (Latitude 13.4521, Longitude 78.2451) was excluded from the study because it lies outside the predetermined boundary area. This helps prevent the data from being distorted by inaccurate or irrelevant site information. Moreover, outlier elimination improves the precision of clustering by allowing only those business venues which are relevant and specific to the context to be retained for subsequent investigation.

### 4.3.4. Project Implementation

After the cleaning and exploration of the data, the resultant dataset is sent to the machine learning component based on Flask for further processing. The geographic attributes, especially latitude and longitude, are normalized to the same scale before modeling. Subsequently, the K-Means clustering algorithm is utilized to classify business venues into different zones that reflect the patterns of their spatial distribution [8]. The number of clusters that best fit the data is identified through the use of internal validation methods such as the Elbow Method and Silhouette Score [9], [10]. Each cluster is then interpreted and labeled accordingly, e.g. a high potential zone, moderate zone, or low opportunity zone. The results produced are sent to the React frontend, where the clusters are represented on an interactive map using leaflet. The users can narrow their choices through the business category filter and also identify those locations which have a combination of high ratings and low competition. This complete chain takes raw geospatial data and turns it into useful and understandable recommendations, which is the main achievement of the EntreLocate system.

## 5. System Architecture

EntreLocate is built on a modular client–server architectural framework that facilitates data gathering, analysis, and decision-making when planning a business location. The system comprises a React-based frontend for user interaction, a Flask backend for data processing and coordination, a machine learning module for analytical modeling, and integrated external APIs for geospatial data acquisition. This multi-layered design not only promotes scalability, ease of maintenance and efficient data exchange among the components but also supports real-time engagements between users and the analytics services.

### 5.1. Frontend Design (User Interface & Map Visualization)

The frontend enables user interaction with the system through an interface where users first specify the city and the business area of interest (e.g. retail, food, health). Processed outputs such as:

- identified business clusters,
- density of venues,
- rating-based insights, and
- recommended areas

are shown on an interactive map. High, medium, and low opportunity zones are marked with markers and color-coded areas, thus entrepreneurs do not necessarily have to be skilled in data interpretation to use these results. Also, the interface enables filtering by rating and business type to facilitate custom-made decisions

### 5.2. Backend Services and API Workflow

The backend layer is the main director of EntreLocate. User inputs coming from the frontend are delivered through RESTful APIs to the Flask server where: The server issues the OpenStreetMap API location data request [6], The obtained output is pre-processed (null elimination, bounding-box verification, generalization) [7], The pre-processed data are fed into the ML pipeline for clustering. Planner returns the final results structured - including labels, centroids, and suggested zones - to the frontend for display.

### 5.3. Machine Learning Pipeline

This machine learning pipeline implements the K-Means algorithm for the spatial analysis of venue features [8]. Latitudes and longitudes are normalized before clustering to prevent the dominance of one scale over another. Each of the clusters obtained are supposed to represent zones of commerce with similar types of neighbors. In order to choose a suitable parameter number, different internal validation techniques, e.g. Elbow Method and Silhouette Score, are used [9], [10].

### 5.4. Integration of External API'S

External APIs are a crucial part of this project. The OpenStreetMap API provides real-world venue attributes such as name, category, rating, and coordinates [6]. The Geocoder service ensures accurate latitude–longitude extraction for user-entered locations. Thanks to this integration, EntreLocate can be used without the manual collection of datasets and its recommendations can be constantly updated with live spatial information.

### 5.5. Profit Estimation and Dashboard Module

The EntreLocate platform goes beyond the provision of location clustering to include a profit estimation tool. This tool is based on the retail profit forecasting models [15], [16].

$$Profit = Estimated\ Revenue - (Fixed\ Cost + Variable\ Cost)$$

It allows users to specify fixed costs such as rent and inventory. Then, the system combines these inputs with demand indicators derived from the cluster to estimate potential revenue and monthly profit.

### 5.6. Proposed Algorithm

Clustering is a well-known method of unsupervised learning which is applied here [8].

It segments the data points into different groups (clusters) based on their closeness to each other, typically using the Euclidean distance measure.

#### 5.6.1. K-Means Clustering

K-Means algorithm was chosen mainly due to its computational efficiency, ability to handle large datasets, and its compatibility with the analysis of spatial data. This algorithm creates K clusters by trying to keep the points in the same cluster as close as possible (low intra-cluster variance). The distance used for this is the Euclidean distance.

- Input: Preprocessed latitude, longitude, and rating
- Output: Cluster label assigned to each venue
- Purpose: Identification of dense and commercially active business zone

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

The algorithm operates iteratively by:

- It starts off with guessing where the K centroids are,
- Then finds the closest centroid for each data point.
- Next, the centroid is recalculated as an average of all the points allocated to it,
- Steps 2 - 4 are repeated until no more changes occur in the centroid position or the assignment of data points to the clusters.

This type of clustering allows the system to identify areas of high spatial concentration which are usually good locations for business activities.

#### 5.6.2. Optimal Cluster Selection Strategy

The number of clusters (K) can drastically change the way the results are interpreted and the effectiveness of the clustering itself. To make sure the data is divided in a meaningful way, the right number of K is figured out by repeating the experiment with different values of K. Moreover, internal validation measures such as the Elbow

Method, Silhouette Score, and Calinski–Harabasz Index are used to determine the optimal value of K. The in-depth research and explanation of the final K value selection will be given in Section V. Figure 2 shows Entre Locate Architecture

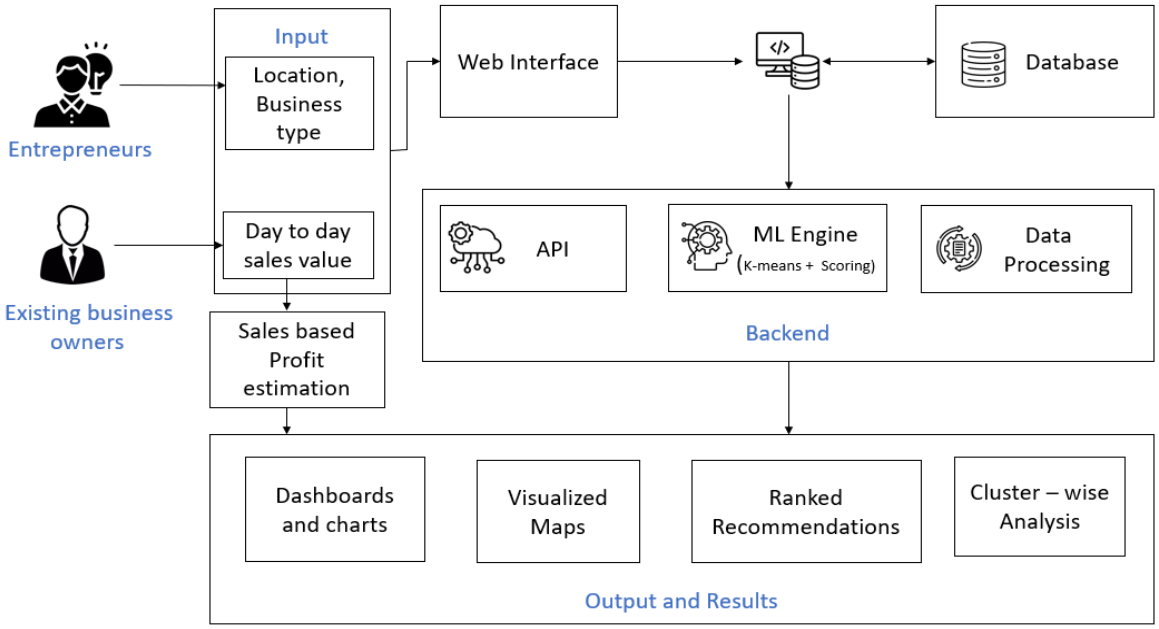


**Figure 2 Entre Locate Architecture**

## 6. Experiment and Result

Experiments were conducted using real-world venue data obtained from OpenStreetMap API. The clustering-based recommendation model was evaluated and compared against supervised machine learning classifiers—Random Forest and Logistic Regression—to assess effectiveness under unlabeled data conditions.

### 6.1. Optimal Cluster Selection Analysis

To perform the experiments, real-world venue data was used, which was obtained through the OpenStreetMap API. The authors also tested their clustering-based recommendation system model against the supervised machine learning classifiers—Random Forest and Logistic Regression—in order to demonstrate performance in unlabeled data scenarios. Choosing the right number of clusters is very important for obtaining stable and meaningful business zones. Three popular internal validation techniques were used to find the correct number of clusters (K).

#### 6.1.1. Elbow Method

The Elbow Method is based on measuring the Sum of Squared Errors (SSE) as K gets larger. Since partitioning becomes more detailed as the number of clusters goes up, SSE goes down. However, after some point, the rate of change becomes smaller, and the curve starts to resemble an "elbow".

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

In FIGURE 3.1, it can be seen that an elbow was practically detected at K=3, which meant that the clustering accuracy and the complexity of the model were balanced.

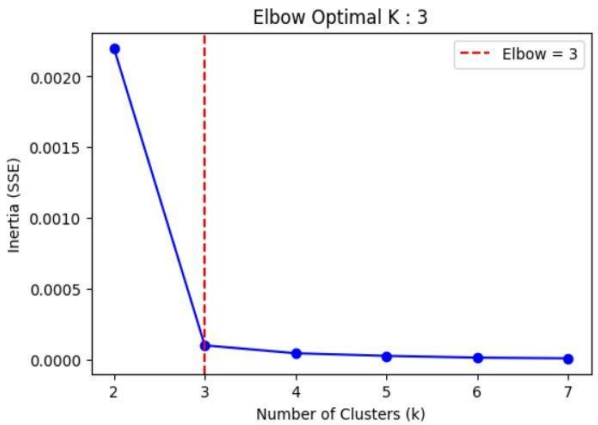


**Figure 3 Elbow Method for Optimal K Value**

### 6.1.2. Silhouette Score Analysis

The Silhouette Score is a metric that evaluates how good the clusters are by measuring the degree of similarity within the cluster and how different the other clusters are, with values that can go as low as −1 and as high as 1.

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

Besides average values of the silhouette, the silhouette difference (the maximum minus the minimum score of clusters) was also taken into consideration to evaluate the degree of the cluster homogeneity. For K = 2, two very close values of silhouette scores, 0.9996 and 0.9922, were obtained, thus the silhouette difference was very small, only 0.0074, denoting the presence of two very compact and well-balanced clusters.


**Figure 4 Silhouette Score to Determine Cluster Quality**

For K = 3, some clusters had high silhouette values, however, one cluster has its silhouette score near zero, hence the silhouette difference is very big (≈0.9996), which is an indication of uneven clusters.

### 6.1.3. Calinski-Harabasz Index

The Calinski–Harabasz Index measures how much of the total data dispersion is due to differences between clusters relative to the differences within each cluster, with higher values denoting clusters that are more well-separated. This measure proposed a most suitable number of clusters, K, as 7.

$$CH = \frac{\mathrm{Tr}(B_k)/(K-1)}{\mathrm{Tr}(W_k)/(N-K)}$$

The problem with such a choice is that it would have split up the commercial areas too much and hence made the resulting map difficult to understand, besides going against the consistency of the findings from the Elbow and Silhouette methods. So, this outcome was disregarded in the final decision-making process.

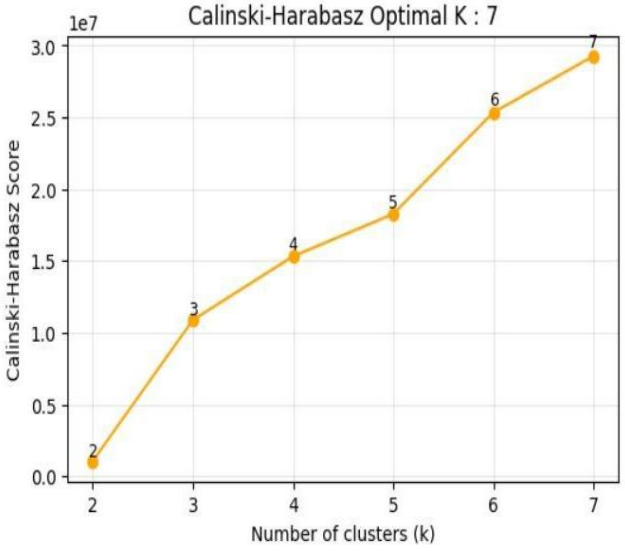


**Figure 4 Calinski–Harabasz Index**

So, this outcome was disregarded in the final decision-making process.

### 6.1.4. Final Decision On Optimal K

The three validation methods were used not as independent decision-makers but as a complementary way of working. Silhouette analysis was considered the main indicator of cluster stability, the Elbow Method was employed to avoid over-segmentation, and the Calinski–Harabasz Index was used as a statistical sanity check. A decision rule was established to solve conflicts. If the difference in silhouette among the clusters was very small, pointing to stable partitions, the result obtained using the Elbow method was considered the simplest adequate K. However, if the difference in silhouette was large, i.e., the clusters were very imbalanced, the silhouettes' stability was given more importance.

```
silhouette_difference value

--- Silhouette Analysis for k=2 ---
Cluster 1: 0.9996
Cluster 2: 0.9922
Silhouette difference (max-min): 0.0074
0.007408734275151763

--- Silhouette Analysis for k=3 ---
Cluster 1: 0.9996
Cluster 2: 0.8913
Cluster 3: 0.0000
Silhouette difference (max-min): 0.9996
0.9996114886775997
```

As a result of the adopted approach, the optimum number of clusters decision was K=2, which ensured very compact and balanced clusters, avoided the fragmentation phenomenon taking place at higher values of K, and gave the most easily interpretable business zones. Silhouette score was the main factor determining the optimum number of clusters. However, the Elbow Method and Calinski–Harabasz Index were also taken into consideration to confirm reliability and prevent bias from a single measure. The comparison of all three methods was made, and the final value of K was chosen based on the agreement between the metrics, on the one hand, and the avoidance of over-fragmentation and the business interpretability of the clusters, on the other hand.

### 6.2. Comparison of Algorithms

In order to determine whether EntreLocate was successful or not, the clustering-based recommendation model was benchmarked against Random Forest and Logistic Regression classifiers. The measures used for comparison were accuracy, precision, recall, and F1-score. Contrary to the supervised models that depend on labeled training data, the suggested clustering method functions label-free. The results seen in Figure 4 indicate that the clustering model can reach an accuracy of around 0.8, and its precision could be very high (close to 1), both of these being comparable to the performance of supervised methods. This serves to show that valuable business locations can be figured out even when there are no labels of past successes, which is especially beneficial for cities that are new or have not been explored much.

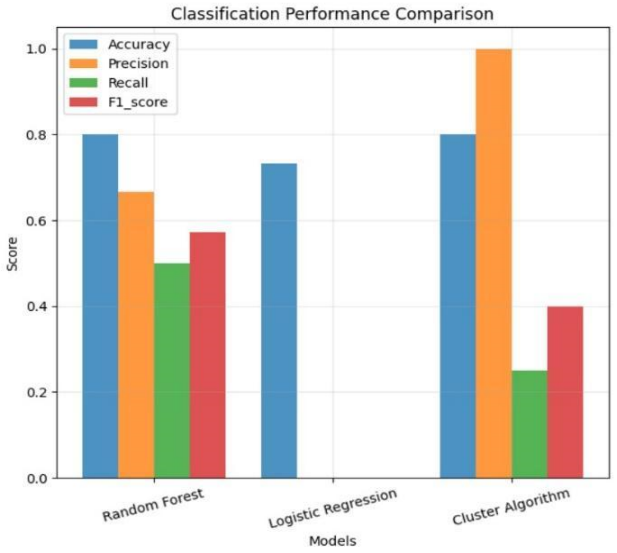


**Figure 4 Comparison of Algorithms**

### 6.3. Confusion Matrix Analysis

To better understand the recommendation model based on the clustering technique, a confusion matrix was constructed (Figure 7). It can be seen that the model was able to correctly recognize the one true high-potential zone along with eleven true low-opportunity zones. In contrast, three potential zones were mistakenly classified as low-opportunity.

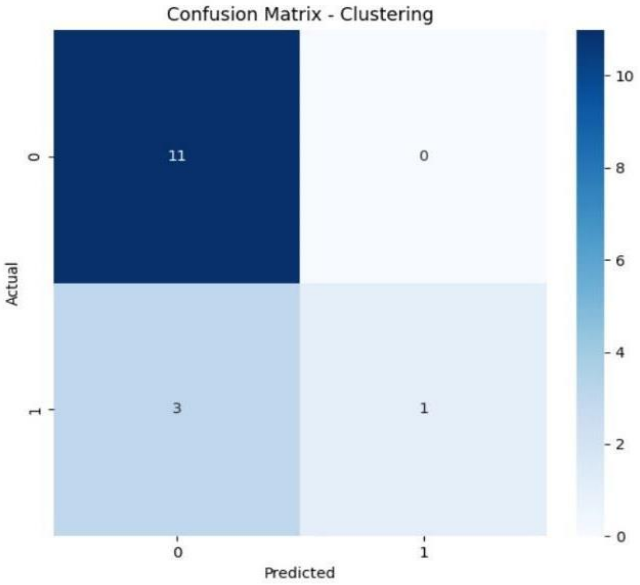


**Figure 5 Confusion Matrix**

Such a cautious approach is indeed very welcomed in business decision support tools, as it effectively reduces the number of false-positive recommendations that might lead to financial loss.

Although recall is quite low, it is the high precision that among other things indicates that the zones endorsed by EntreLocate are trustworthy.

### 6.4. Performance Radar Analysis

In FIGURE 8 radar chart compares evaluation metrics from various perspectives. The proposed model showcases string accuracy and very high precision. However, recall is a bit lower (approximately 0.25).This is in line with the system's objective of offering more reliable recommendations thus fewer aggressive predictions.

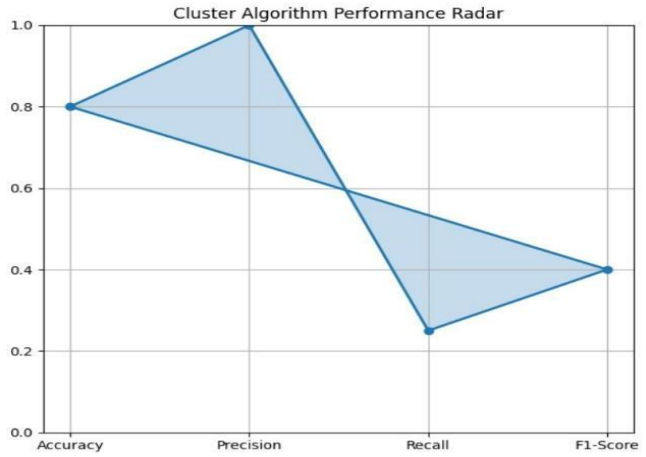


**Figure 6 Performance Radar Analysis**

Additional features to include footfall estimation, proximity to same-type of businesses, and temporal demand patterns, may help to increase recall without losing precision.

### 6.5. Output Generation

The given clusters are changed to user-level recommendations:

- Flask sends cluster labels + centroids to React
- The map shows clustered areas and recommendations.

The system recommends business locations by identifying clusters that have high commercial activity but with lower average ratings. This means the system locates areas where there is a high demand but low supply of quality service. Suggestions also look at the lack of a particular business category in a cluster, thus, the entrepreneur is given the choice of targeting the area where the provision of the same kind of service is limited or the existing one is below the level. The new business entering strategy, therefore, encourages the moving along of customer flows while avoiding the places of high competition.

## 7. Limitations and Future Enhancements

### 7.1. Data Dependency and Quality Constraints

EntreLocate system depends on the OpenStreetMap platform for worldwide venue information, the quality of its suggestions is tied to how complete and accurate the OSM data is for various countries and regions. Certain areas may have very few or even outdated entries, which can have an impact on the quality of clustering. The present study mainly depends on static snapshots of spatial data, and it may therefore not be able to fully account for short-term changes like seasonal demand, temporary events, or recent market changes. The profit estimation component relies on the user's inputs for fixed and variable costs; therefore, the lack of standardized regional cost datasets may limit the accuracy of the financial forecasts.

### 7.2. Scalability Considerations

EntreLocate is conceived as a global platform and could work for any city or country that is mapped by OpenStreetMap. But global usage on the scale of the whole world creates problems of very large data volumes and different characteristics of regions.

One may need to process venue data from different locations in the world that require:

- the use of different clustering parameters for different urban densities,
- the use of efficient caching to reduce repeated API calls,
- the use of distributed computation to be able to handle millions of venue records.

In addition to Google Maps, Yelp, or Justdial that are consumer platforms and only show highly rated existing shops for customers, EntreLocate, on the other hand, carries out entrepreneur-focused analytics so as to suggest locations for setting up new business. This, in turn, requires more extensive processing and continuous updates to be able to stay effective at a global scale.

### 7.3. Future Research Directions

Future work will include continuing to develop EntreLocate as a globally adaptive decision-support platform. One way to enhance the system includes integrating live map OpenStreetMap updates, implementing user feedback to improve refining recommendations, and developing an enhanced profit estimation model that takes into account

seasonal trends, and local economic factors. Changing clustering methods depending on the density of urban and rural areas, and including more features like competitor's distance and demographic indicators, may also help the system to produce more accurate recommendations. These upgrades will facilitate EntreLocate in offering more tailored and financially well-informed advice to entrepreneurs all over the world based on spatial decision-support studies [15] and profit modeling approaches [16].

## Conclusion

The paper details a smart, data-driven business location recommender system that helps startup founders pick the best sites for their new businesses. Specifically, the framework draws on OpenStreetMap's worldwide venue data and goes through a preprocessing stage before utilizing K-Means clustering to identify not only the features of the neighborhoods but also the commercial density of the areas. The findings are then accessible via a user-friendly web interface. Traditional mapping services, which largely recommend customers to visit already established popular shops, differ from EntreLocate which, on the one hand, employs an entrepreneur-centric strategy, and on the other, it points to opportunity zones where demand exists but service quality or availability is comparatively low. The addition of a profit estimation and dashboard module that accounts for fixed and variable costs makes it possible for users to assess financial viability prior to investment. A demonstration of the system's working showed that it offers understandable and reliable recommendations, achieving a level of performance. Furthermore, EntreLocate is a tool that geospatial analytics and effective business planning had long been wanting for. It is a scalable global solution that is in line with the latest spatial business analytics literature [12]–[16] and enables new business owners to make informed, low-risk decisions.

## References

[1]. Khetarpaul, Sonia, et al. “Location-Based Ideal Site Selection Using Clustering.” 2024 IEEE International Conference on Contemporary Computing and Communications (InC4), vol. 1, 2024, pp. 1–8. IEEE Xplore, https://doi.org/10.1109/InC460750.2024.10649092.

[2]. P, Ashok Kumar, et al. “Location Based Business Recommendation Using Spatial Demand.” Sustainability, vol. 12, no. 10, May 2020, p. 4124. DOI.org (Crossref), https://doi.org/10.3390/su12104124.

[3]. Zhao, Guoshuai, et al. “Location Recommendation for Enterprises by Multi-Source Urban Big Data Analysis.” IEEE Transactions on Services Computing, vol. 13, no. 6, Nov. 2020, pp. 1115–27. IEEE Xplore,https://doi.org/10.1109/TSC.2017.2747538.

[4]. Han, Shuihua, et al. “Identifying a Good Business Location Using Prescriptive Analytics: Restaurant Location Recommendation Based on Spatial Data Mining.” Journal of Business Research, vol. 179, Jun. 2024, p. 114691. ScienceDirect, https://doi.org/10.1016/j.jbusres.2024.114691.

[5]. Iranzad, Reza, and Xiao Liu. “A Review of Random Forest-Based Feature Selection Methods for Data Science Education and Applications.” International Journal of Data Science and Analytics, vol. 20, no. 2, Aug. 2025, pp. 197–211. Springer Link, https://doi.org/10.1007/s41060-024-00509-w.

[6]. Lin, J., and Oentaryo, R. J., “A Business Zone Recommender System Based on Facebook and Urban Planning Data,” Proc. International Conf. on …, 2016.

[7]. Smiti, Abir. “A Critical Overview of Outlier Detection Methods.” Computer Science Review, vol. 38, Nov. 2020, p. 100306. ScienceDirect,https://doi.org/10.1016/j.cosrev.2020.100306.

[8]. Bindra, Kamalpreet, and Anuranjan Mishra. “A Detailed Study of Clustering Algorithms.” 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2017, pp. 371–76. IEEE Xplore, https://doi.org/10.1109/ICRITO.2017.8342454

[9]. Shahapure, Ketan Rajshekhar, and Charles Nicholas. “Cluster Quality Analysis Using

Silhouette Score." 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, pp. 747–48. IEEE Xplore, https://doi.org/10.1109/DSAA49011.2020.00096.

[10]. [10] Marutho, Dhendra, et al. "The Determination of Cluster Number at K-Mean Using Elbow Method and Purity Evaluation on Headline News." 2018 International Seminar on Application for Technology of Information and Communication, 2018, pp. 533–38. IEEE Xplore,https://doi.org/10.1109/ISEMANTIC.2018.8549751.

[11]. Sitarz, Mikolaj. "Extending F1 Metric, Probabilistic Approach." arXiv:2210.11997, arXiv, 26 Oct. 2022. arXiv.org, https://doi.org/10.48550/arXiv.2210.11997.

[12]. Li, X., Liu, Y., and Chen, J., "Urban Business Site Selection Using POI Big Data and Spatial Clustering," IEEE Access, vol. 9, 2021, pp. 84532–84545. https://doi.org/10.1109/ACCESS.2021.3089124.

[13]. [Yuan, J., Zheng, Y., and Xie, X., "Discovering Regions of Different Functions in a City Using Human Mobility and POIs," Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 186–194. https://doi.org/10.1145/2339530.2339561.

[14]. Zhang, D., He, T., and Lin, S., "Data-Driven Location Recommendation for New Store Placement," Expert Systems with Applications, vol. 168, 2021, p. 114236. https://doi.org/10.1016/j.eswa.2020.11423

[15]. Huang, Q., Wong, D., and Li, Y., "Spatial Analytics for Small Business Decision Support Using Open Geospatial Data," ISPRS International Journal of Geo-Information, vol. 10, no. 5, 2021, pp. 1–18. https://doi.org/10.3390/ijgi10050321.

[16]. Wang, S., and Zhao, P., "Profit Prediction Model for Retail Stores Based on Location and Demographic Features," Journal of Retail Analytics, vol. 7, no. 3, 2022, pp. 55–67. https://doi.org/10.1109/JRA.2022.3178894.