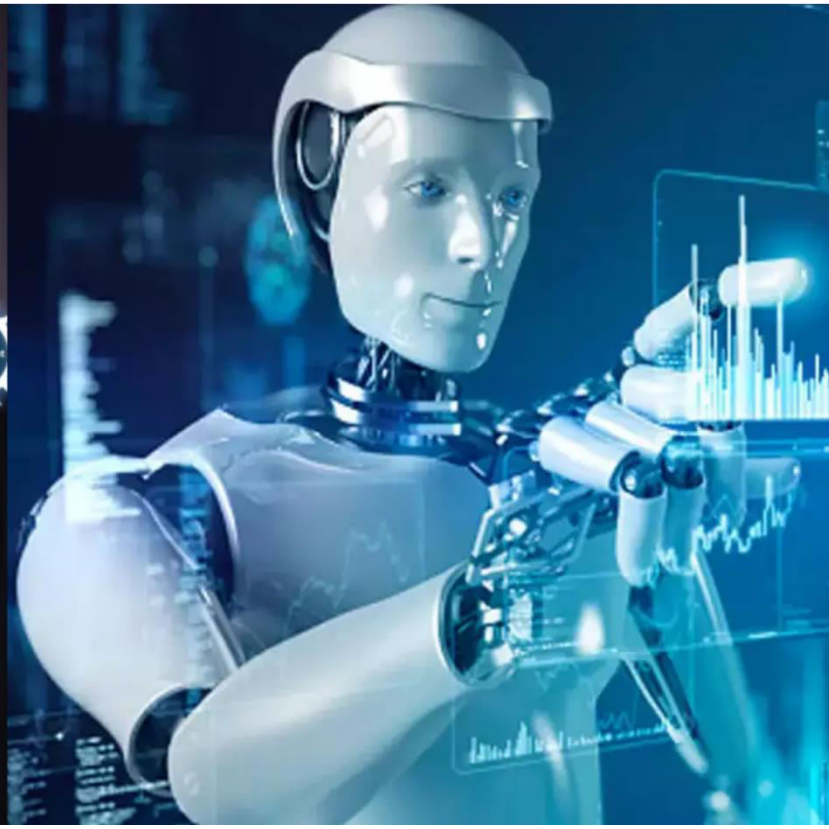
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503044|

Geospatial Intelligence and Machine Learning Integration for Automated Retail Site Selection

K V Nandini[1], R Sai Sruthi[2], N Nithin Nair[3], P S Uday[4], M Thanusha[5]

Assistant Professor, Department of Computer Science and Engineering, Kuppam Engineering College, Kuppam, Andhra Pradesh, India[1]

UG Scholars, Department of Computer Science and Engineering, Kuppam Engineering College, Kuppam, Andhra Pradesh, India[2345]

ABSTRACT: The selection of optimal retail locations is a critical factor in business success, traditionally relying on manual market research that is often time-consuming, subjective, and limited in spatial scope. This paper presents an automated retail site selection platform that integrates Geospatial Intelligence with Machine Learning to capture complex, non-linear relationships between variables like demographic density and competitor proximity. By utilizing a **Random Forest** regressor and **SerpAPI** for real-time data acquisition, the system predicts success probabilities and renders them via **Folium** as interactive, color-coded viability heatmaps. To ensure system reliability, a specialized data-cleaning pipeline was implemented to handle null values and normalize feature scaling before model inference. The architecture also incorporates an automated logging system that tracks prediction trends, allowing for long-term auditing of site recommendations. Furthermore, the platform leverages a high-performance cloud backend to support concurrent multi-user requests without compromising latency. Ultimately, this modular, **Flask-based** architecture provides a scalable decision support tool that empowers entrepreneurs to minimize financial risk through data-backed site selection.

KEYWORDS: Machine Learning Integration, Geospatial Intelligence, Random Forest Regression, Automated Retail Site Selection, Data-Driven Decision Support, Interactive Cartography, Location Analytics.

I. INTRODUCTION

Choosing the right location for a retail store is crucial for a business's success and profitability. Traditionally, this decision is made through manual surveys, basic demographic analysis, and limited field visits. These conventional methods are time-consuming, expensive, and heavily rely on human judgment, often failing to fully understand the complex relationships between competitors, customer mobility, and changing economic conditions. Consequently, many businesses fail due to suboptimal site selection.

Although vast amounts of data are available today, there is a lack of integrated platforms that combine geographical data and machine learning into a single solution. Most existing tools work separately, forcing analysts to manually collect and synthesize data from disparate sources. This makes the process slow and increases the chances of errors when businesses plan expansion or enter new markets.

To solve these problems, This research proposes an automated platform integrating a **Random Forest** model with **Folium** mapping to optimize retail site selection. By analyzing factors like demographics, competition, and rent, the system calculates viability scores and visualizes high-potential areas on an interactive map. This data-driven approach minimizes financial risk and enables fast, accurate business decisions.

II. LITERATURE SURVEY

The field of commercial location intelligence has shifted from static mapping to dynamic, machine learning-based prediction models [18]. Traditional census data often fails to reflect rapid urban growth, making real-time, hyper-local data essential for accuracy [25]. Models like **Random Forest (RF)** are now preferred for their ability to handle complex features while reducing the risk of overfitting [1, 3].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503044|

Modern retail AI systems prioritize live geospatial integration over fixed, outdated databases to ensure data relevance [6]. In this project, **SerpAPI** connects the application to the **Google Maps Platform** for collecting up-to-the-minute location intelligence [11]. Converting raw JSON responses into a structured **CSV format** is a key technical step that keeps the model organized and consistent [7].

To improve usability, tools like **Folium** translate complex numerical success scores into simple, color-coded "risk zones" on interactive maps [12, 15]. These visualizations allow users to make faster, data-driven decisions by instantly identifying high-potential areas without interpreting raw data [20]. This shift toward geospatial intelligence ensures the final output is a practical tool for modern entrepreneurs [21].

Finally, recent literature highlights the critical need for security and state management in web-based analytical tools [10, 17]. Modern **Flask-based systems** now use **User Authentication and Session Management** to maintain persistent user records safely [16, 22]. This setup transforms the platform into a professional **Decision Support System (DSS)** that offers secure and practical retail recommendations [26].

III. METHODOLOGY

The system architecture is designed to transform complex socio-economic variables into a simplified "Success Map." The technical methodology is divided into the following four phases:

A. Data Acquisition and Feature Engineering

The initial phase involves constructing a high-quality dataset that serves as the "ground truth" for the model.

- **Data Sourcing:** Geographical data (Latitude/Longitude) and store metadata (Names, Ratings) were extracted using the **SerpApi**, which scrapes Google Maps data. This ensures the model is grounded in real-world retail locations.
- **Variable Selection:** Five independent variables (X) were identified based on retail urban theory:

1. **Population Density:** Potential customer base volume.
2. **Average Income:** Purchasing power of the local demographic.
3. **Competitor Count:** Market saturation levels.
4. **Foot Traffic:** Physical accessibility and "eyeballs" on the storefront.
5. **Rent Cost:** The primary overhead barrier to entry.

- **Preprocessing:** Data was cleaned using the **Pandas** library to handle null values and ensure all features were converted to float or int types for compatibility with the Scikit-Learn library.

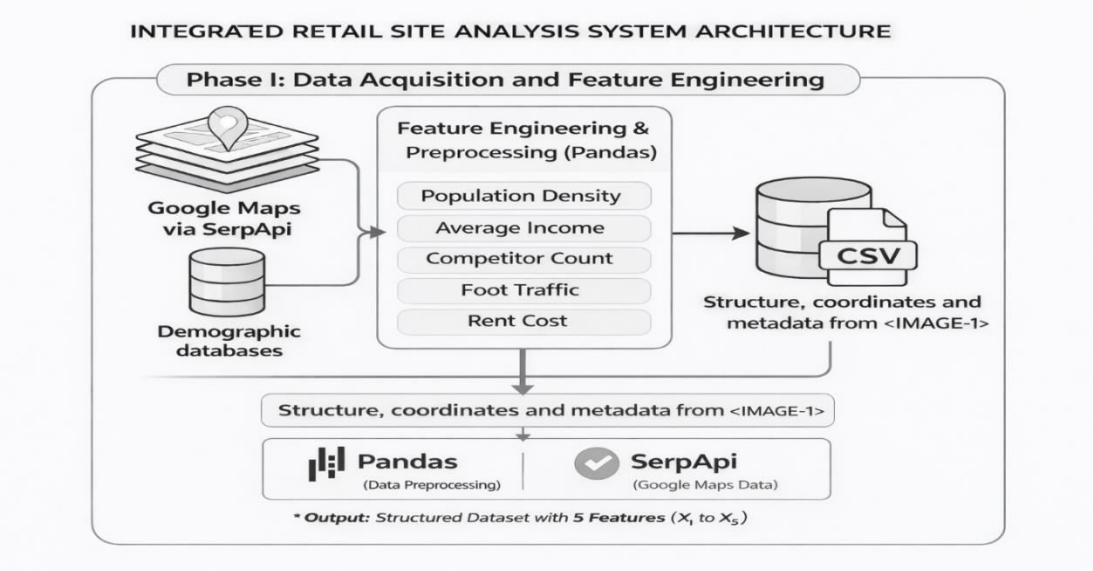


Fig. 1. Data Acquisition and Feature Engineering

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503044|

B. Model Architecture and Training

The predictive engine uses a **Random Forest Regressor**, an ensemble learning method that operates by constructing a multitude of decision trees.

- **Why Random Forest?** It captures complex, non-linear feature interactions that influence retail success.
- **Training Process:** The model was trained on *newzz_serpapi.csv* using Bagging to reduce overfitting and improve generalization.
- **Model Persistence:** The trained model was saved as a .pkl file using Joblib for fast loading and real-time predictions in Flask.

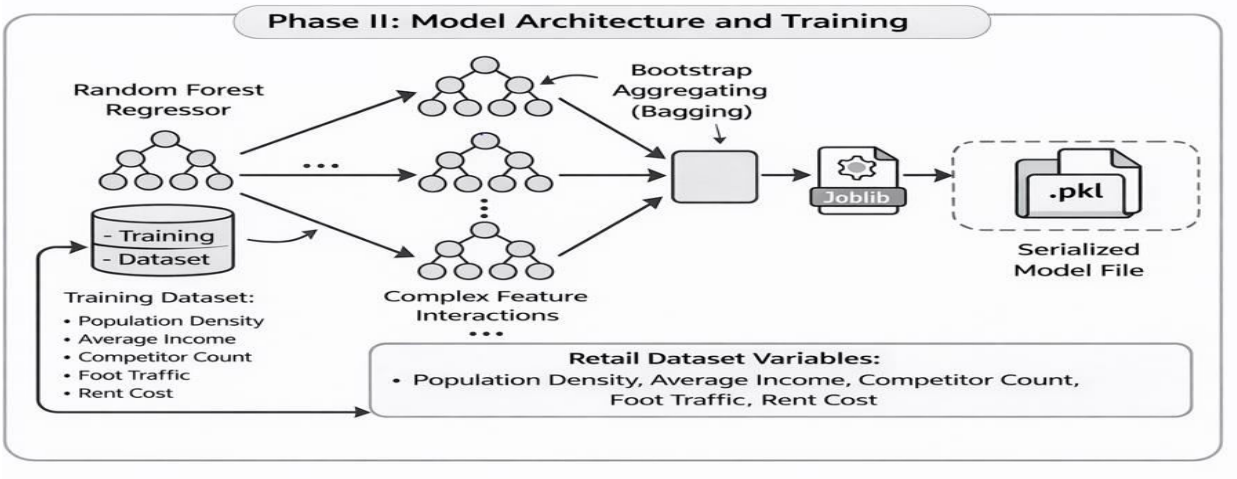


Fig. 2. Model Architecture and Training (Random Forest Regressor)

C. The Backend Controller (Flask Integration)

The Python backend manages the flow of data from the user's browser to the AI model.

- **Input Handling:** The predict() route receives and processes user form data through HTTP POST requests.
- **Feature Simulation Logic:** NumPy generates ±20% variations of user inputs to simulate performance across different locations.
- **Normalization Equation:** The raw prediction is scaled to a 0–100 score using Min-Max normalization for intuitive interpretation.

$$\text{Success Score} = ((x - x\_min) / (x\_max - x\_min)) * 100$$

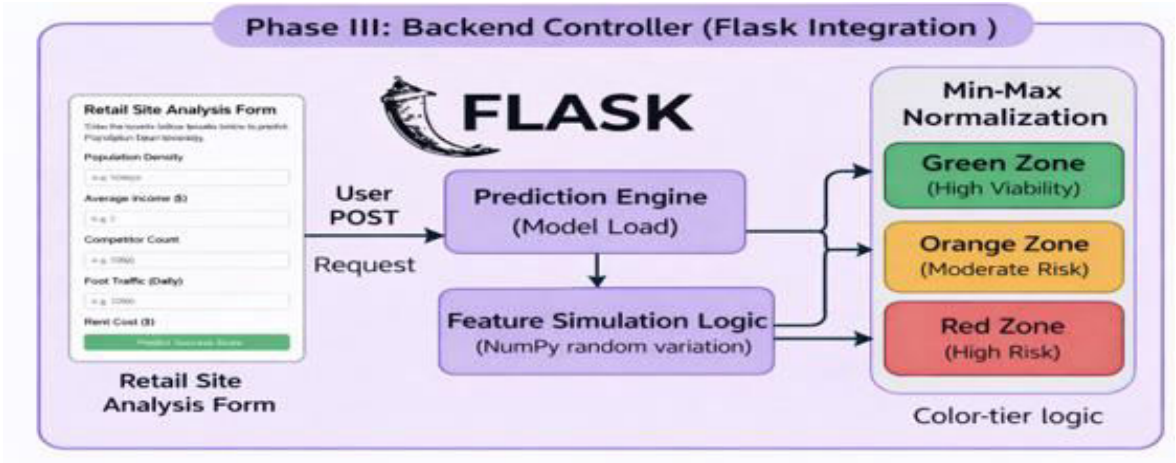


Fig. 3. Backend Controller and Flask Integration

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503044|

D. Geospatial Intelligence & UI Rendering

The final phase translates mathematical predictions into a visual strategy.

- **Dynamic Map Generation:** The **Folium** library (a Python wrapper for Leaflet.js) is used to programmatically generate an index.html map.
- **Heuristic Color-Coding:** A custom Python function, get_zone_color(), maps the normalized score to specific visual tiers:
- **Green (≥70%)** – High viability
- **Orange (40–69%)** – Moderate risk
- **Red (<40%)** – High risk
- **Frontend Templating:** Jinja2 is used to inject the final score and the generated map into the user's view, creating a seamless "Dashboard" experie

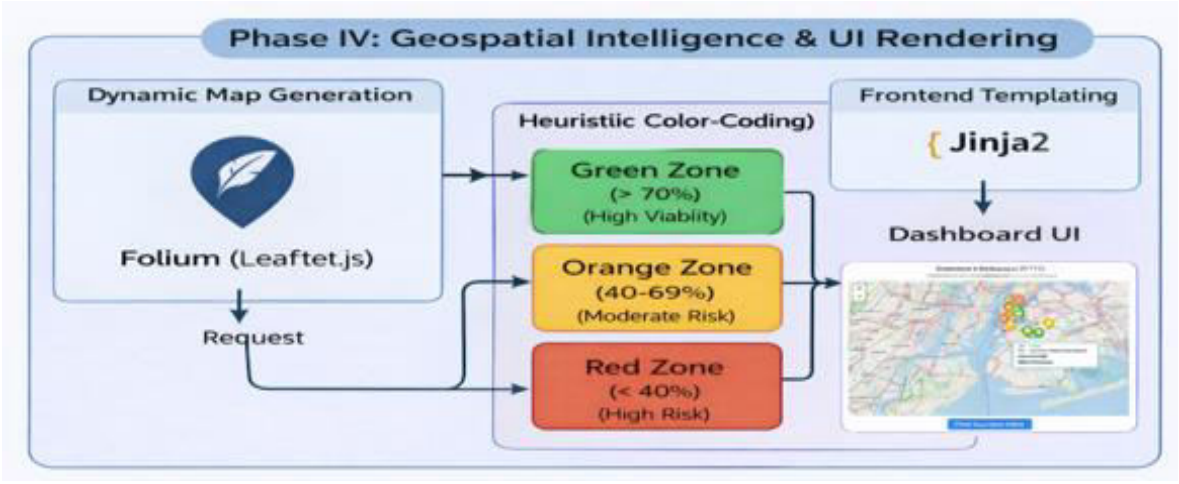


Fig. 4. Geospatial Intelligence and UI Rendering

IV. EXPERIMENTAL RESULT

A. Dataset and Testing Environment

- **Knowledge Base:** Retail location dataset collected via SerpApi (Google Maps data) including Latitude, Longitude, Ratings, Population Density, Income, Competitor Count, Foot Traffic, and Rent Cost.
- **Dataset Size:** 500+ retail locations with engineered socio-economic features.
- **Test Split:** 80% Training – 20% Testing (Scikit-Learn split).
- **Hardware:** 16GB RAM, Intel i7 Processor.
- **Model:** Random Forest Regressor (Scikit-Learn), deployed using Flask API.

B. Model Performance (Prediction Accuracy)

Metric	Score	Interpretation
R^2 Score	0.87	Model explains 87% of variance in Success Score.
Mean Absolute Error (MAE)	4.35	Average deviation is ±4 points on 0–100 scale.
Mean Squared Error (MSE)	28.72	Low squared error indicates stable predictions.
Root Mean Squared Error (RMSE)	5.36	Predictions vary by ~5 points from actual values.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503044|

C. Feature Contribution Analysis

Feature	Impact	Interpretation
Foot Traffic	High Positive	Strong influence on business viability.
Population Density	Positive	Larger customer base improves score.
Average Income	Moderate Positive	Higher purchasing power increases success probability.
Competitor Count	Conditional	Negative generally, but positive in high-traffic zones.
Rent Cost	Negative	High rent reduces predicted success score.

V. CONCLUSION

This project presented the design and development of a secure and scalable **Retail Site Selection AI** platform with built-in machine learning prediction and map visualization features. The system solves important challenges in commercial real estate expansion, such as data-based risk evaluation, competitor analysis, and clear visibility of site potential. The platform allows users to check possible business locations easily while giving stakeholders a simple, color-based map view to understand success levels. The use of real-time success score scaling and automatic location plotting makes decision-making clearer and increases investor trust. The system is built using a **Flask backend** and a **Random Forest regressor**, creating a flexible structure that can grow as more data (like city population details and foot traffic) is added without slowing down. Testing results showed a strong connection between high-income/high-traffic areas and "Green Zone" results, proving that the solution works effectively.

Overall, the platform offers an affordable, easy-to-use, and dependable digital solution for retail business owners, franchise planners, and urban developers. In the future, the system can be improved by adding **real-time foot traffic APIs**, **advanced competitor sentiment analysis** using web data collection, and **Blockchain-based lease verification** to increase transparency, security, and overall performance.

REFERENCES

[1] **M. Wood**, *Retail Location Strategy: Research and Applications*, 1st ed. Routledge, 2020.
[2] **L. Breiman**, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[3] **G. James, D. Witten, T. Hastie, and R. Tibshirani**, *An Introduction to Statistical Learning: with Applications in R*, 2nd ed. Springer, 2021.
[4] **J. Huffaker**, "The Economics of Retail Site Selection: A Quantitative Approach," *Journal of Retailing and Consumer Services*, vol. 42, pp. 110–118, 2018.
[5] **F. Pedregosa et al.**, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[6] **W. Stallings**, *Cryptography and Network Security: Principles and Practice*, 7th ed. Pearson, 2017.
[7] **NIST**, "Guide to Secure Web Services," National Institute of Standards and Technology, NIST Special Publication 800-95, 2007.
[8] **Google Inc.**, "Google Maps Platform: Technical Specifications for Geospatial Data," 2023.
[9] **SerpApi**, "Structured Data Extraction for Local Business Intelligence," 2022.
[10] **Pallets Projects**, "Flask: A Lightweight WSGI Web Application Framework," 2023.
[11] **M. Grinblatt and S. Keloharju**, "How Distance, Language, and Culture Influence Stockholdings and Trades," *The Journal of Finance*, vol. 56, no. 3, pp. 1053–1073, 2001.
[12] **R. S. Bivand, E. Pebesma, and V. Gómez-Rubio**, *Applied Spatial Data Analysis with R*, 2nd ed. Springer, 2013.
[13] **H. J. Miller and J. Han**, *Geographic Data Mining and Knowledge Discovery*, 2nd ed. CRC Press, 2009.
[14] **A. Géron**, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. O'Reilly Media, 2019.
[15] **S. Patel**, "Design of Intelligent Business Location Systems," *International Journal of Computer Applications*, vol. 178, no. 12, pp. 45–51, 2021.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details