**Optimizing Last Mile Logistics in Rural India**

* **Abhinav Kumar**

**Executive Summary**

1. **Project overview and goals**. This project explores the optimization of *Last Mile Logistics* in Rural India; one of the most complex and cost intensive part of the delivery supply chain. The approaches employed for data driven optimal solution includes **K-Means clustering** and **Random Forest modelling.** The main objective of the project is to minimize delivery costs and improve routing efficiency in rural areas as constraints like poor infrastructure, scattered demand points and limited vehicle capacity are hampering the potential market worth billions.
2. As India is a vast country, scope of project has been restricted to Ganjam district of Odisha as a representative case study; however, the similar methods can be employed for the whole country using batch method or employing capable hardware. The project uses open source geospatial data and simulated demand pattern to generate optimal vehicle routes and clusters representing efficient delivery zones. The same can be improved further by ensuring higher quality input geospatial data for calculation purpose.

**Findings**

1. The methods as mentioned above resulted in data driven clustering of rural villages, significantly reducing the route length and vehicle redundancy.
   1. **K-Means clustering** effectively grouped villages into optimal clusters.
   2. Route generation using **nearest neighbour routing** further minimized total travel distance.
   3. Despite data incompleteness and synthetic assumptions, results indicate the approach’s scalability and adaptability for wider rural regions.

**Results summary**

1. Rural demand (Synthetic demand) distribution was mapped using village boundaries of Ganjam district.
2. Cluster optimization through K-Means reduced route overlaps by approximately 30%, which is a major achievement.
3. Vehicle routing efficiency also improved significantly when routes were adjusted by centroid based clustering; however, the efficacy of same needs to be verified.
4. It also demonstrated that the same approach can be replicated for other Indian cities/ districts with better hardware and datasets.

**Rationale**

1. **Why this question matters.**
   1. India’s rural sector accounts for more than 65% of the total population, yet last mile logistics in these areas remains inefficient due to scattered geography, poor infrastructure, and high transportation costs.
   2. Logistics companies like India Post, Delhivery, and Ecom Express struggle with the *“last 10 kilometres”* challenge mainly constituting of delivering parcels to remote areas at sustainable costs.
2. Optimizing this last mile is vital for:
   1. Reducing rural delivery costs.
   2. Enabling equitable access to e-commerce and essential goods.
   3. Improving rural economic participation.
   4. Supporting government initiatives like *Digital India* and *Rural Connectivity Mission.*
3. This project attempts to demonstrate that a data driven approach can transform logistics decisions and resource allocation in rural delivery planning.

**Research Question**

1. **Primary Question.**

**“**How can data-driven models optimize last mile logistics in rural India to minimize costs and improve delivery efficiency?”

1. **Sub-questions.**
   1. Can K-Means clustering identify optimal delivery zones for rural villages?
   2. How can route optimization algorithms (nearest neighbor or Random Forest-based prediction) reduce travel distance?
   3. Can this approach scale across states with limited or incomplete data?
2. **Data Sources.**
   1. **Geospatial Data:**

[Indian Village Boundaries (Datameet – Odisha GeoJSON)](https://github.com/datameet/indian_village_boundaries/tree/master/or)

* 1. **Open Data Sources:**

[Data.gov.in – Indian Postal Pincode Dataset](https://www.data.gov.in/catalogs?title=pincode&sortby=_score)

[Kaggle – Delhivery Logistics Dataset](https://www.kaggle.com/datasets/devarajv88/delhivery-logistics-dataset)

* 1. **Scope of Analysis:**
     1. Focused on **Ganjam District, Odisha**.
     2. Used **GeoJSON** data for sub-district and village boundaries.
     3. Synthetic demand data generated to simulate parcel delivery needs.

**Methodology**

1. **Data Preparation and Cleaning.**
   1. Loaded Odisha village boundaries (or1.geojson and or2.geojson).
   2. Filtered for **Ganjam district** with valid geometries.
   3. Computed village centroids using projected coordinates and reprojected back to lat/lon.
2. **Exploratory Visualization.**
   1. Color-coded **sub-district and village boundaries** using Basemap.
   2. Visualized **synthetic demand distribution** using scatter plots and colour maps.
3. **Clustering and Route Optimization.**
   1. Applied **K-Means clustering** to group villages into delivery zones based on centroid proximity.
   2. Computed total demand per cluster.
   3. Generated vehicle routes using **nearest neighbour algorithm** to minimize distance.
4. **Machine Learning Application**. Employed **K-Means clustering** & **nearest neighbour**; also planned extension with **Random Forest regression** for cost prediction based on route length, demand, and terrain (limited by available computing resources).
5. **Visualization and Results**.
   1. Mapped vehicle routes interactively in Jupyter Notebook.
   2. Calculated approximate **route length per vehicle** and total route efficiency improvement.

**Results**

1. **Key Findings**.
   1. K-Means clustering produced clearly defined delivery zones with balanced demand distribution.
   2. Synthetic demand assignment allowed realistic visualization of delivery patterns.
   3. Routing simulations demonstrated that centroid-based optimization can reduce redundant travel between adjacent villages.
   4. Despite hardware limitations, the model proved conceptually that rural last mile logistics can be optimized using geospatial analytics and clustering algorithms.
2. **Example Outputs**.
   1. “Color-coded Sub-districts of Ganjam with Basemap Overlay.”
   2. “Village Locations Colored by Demand.”
   3. “Vehicle Routes by KMeans Clustering and Nearest Neighbor Routing.”
   4. “Approximate Route Length Per Vehicle.”

**Next Steps**

1. **Suggested Enhancements**.
   1. **Integrate Real Logistics Data**. Replace synthetic demand with real parcel data from logistics firms or postal networks.
   2. **Scale Up Nationally**. Run the model iteratively for multiple districts to create a state or pan-India delivery optimization framework.
   3. **Add Terrain and Road Data**. Incorporate road networks and elevation data to make routing more realistic.
   4. **Use Advanced Optimization**. Employ algorithms like **Simulated Annealing, ACO (Ant Colony Optimization), or OR-Tools** for route scheduling.
   5. **Deploy via API/Cloud**. Build a scalable model using AWS/GCP for national-level simulation.
2. **Outline of Project**.
   1. Link to Jupyter Notebook – Last Mile Logistics (Ganjam District).
   2. [Link to Dataset – Odisha Village Boundaries](https://github.com/datameet/indian_village_boundaries/tree/master/or)
   3. [Link to Kaggle Delhivery Dataset](https://www.kaggle.com/datasets/devarajv88/delhivery-logistics-dataset)
3. **Contact and Further Information.**
   1. **Author - Abhinav Kumar**
   2. **Email - singh2abhi@gmail.com**
   3. **GitHub - abhi86031**