Health Facility Clustering Project

1. Project Overview

This project analyzes Indian government health facility data. It includes data extraction, cleaning, creation of 10 structured tables, and the development of a machine learning clustering model to group healthcare facilities based on geographic and operational characteristics.

2. Dataset Source

The dataset was downloaded from a public Government of India portal. It includes information such as State, District, Facility Name, Facility Type, Latitude, Longitude, Ownership Type, and Notional/Physical status.

3. Project Requirements

- Create 10 dimension tables.
- Build a machine learning model over the final dataset.
- Visualize cluster insights.
- Deliver KPIs summarizing analytical results.
- Share GitHub repository + deployed output.

4. Tables Created

1. State Table
2. District Table
3. Subdistrict Table
4. Facility Type Table
5. Facility Master Table
6. Location Table
7. Ownership Table
8. Physical/Notional Table
9. Administrative Mapping Table
10. Full Master Table

5. Machine Learning Pipeline

- Cleaned dataset, removed missing values, and standardized columns.
- OneHotEncoded categorical fields.
- Scaled latitude/longitude values.
- Trained K-Means clustering model with 6 clusters.
- Visualized clusters using PCA (Principal Component Analysis).

6. KPIs (Key Performance Indicators)

Data Quality KPIs
- Missing Latitude/Longitude handled.
- Missing Facility Type filled with fallback labels.
- Duplicate facility entries removed.

Model KPIs
- Number of clusters: 6
- Silhouette Score: Indicates cluster separation quality.
- Inertia Score: Shows compactness of clusters.

Operational KPIs
- Active vs inactive facility ratio.
- Facility distribution across clusters.
- Dominant facility type per cluster.

Business KPIs
- Identified high-density vs low-density health facility regions.
- Cluster insights helpful for health resource allocation and planning.

7. Final Deliverables

- Trained Machine Learning model (kmeans_health_model.pkl).
- Scaler and encoder objects.
- Clustered dataset (health_facility_clusters.csv).
- Complete Google Colab notebook.
- Project documentation (PDF).

8. Conclusion

This project demonstrates a full end-to-end data analytics and machine learning workflow—from government dataset extraction to model training, visualization, and KPI-driven insights. It is suitable for real-world analytics, operational planning, and decision-making applications in the healthcare sector.