Health Mangement Information System(HMIS)

Name: Arya Kashikar

PRN: 22070521036

Sec: A, Sem: VII

Subject: Data Science

Faculty In-charge: Dr Piyush Chauhan

Project Objective and Problem Statement

The project developed a complete data science pipeline to analyze and model state-wise healthcare utilization patterns in India using Health Management Information System (HMIS) data. HMIS collects complex monthly data from hospitals nationwide, but this data is difficult to interpret directly for policymakers. The key challenges addressed:

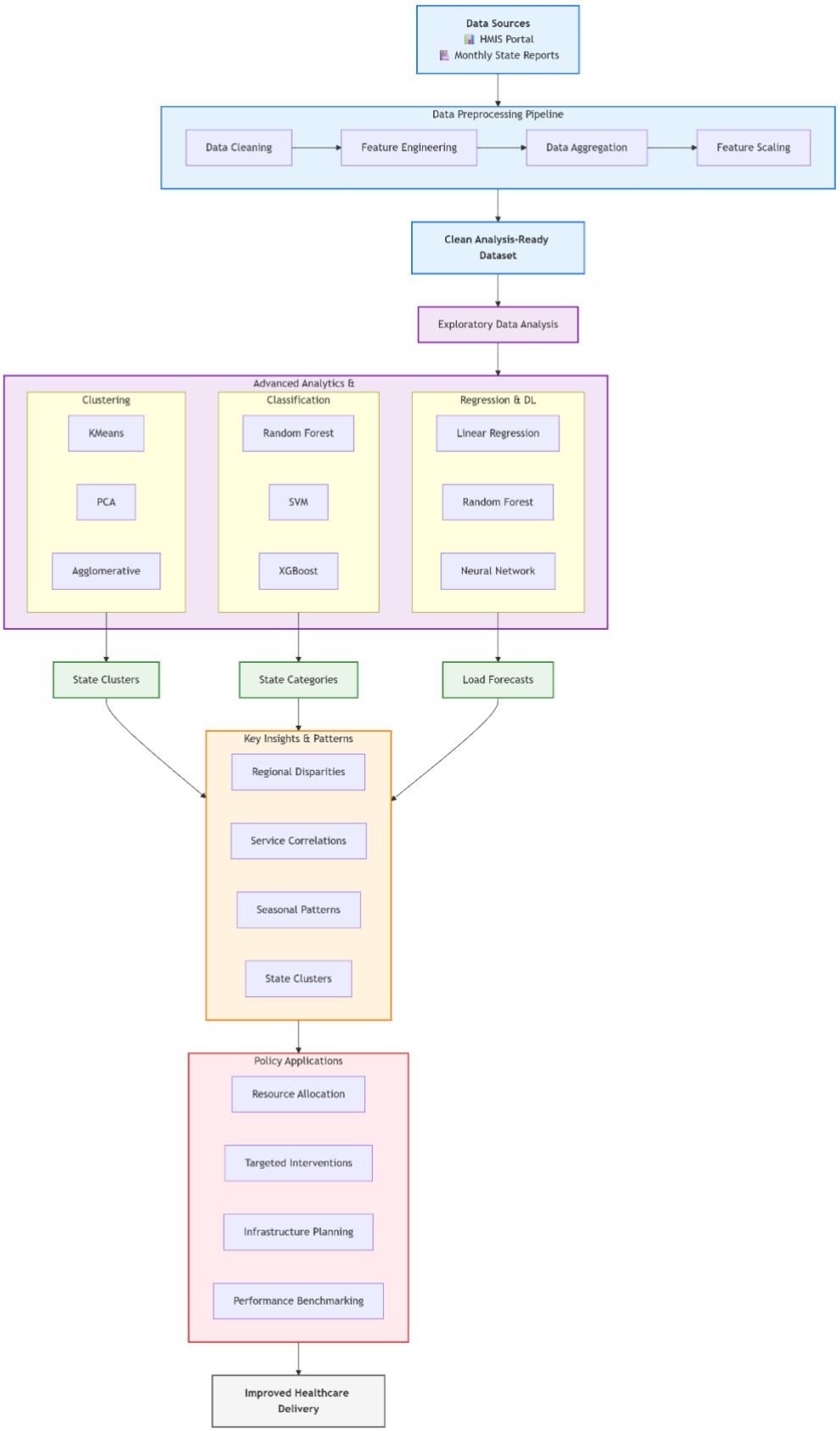
- Comparing healthcare performance between Indian states
- Grouping states with similar healthcare service patterns
- Predicting hospital inpatient load for planning resources

The aim was to provide actionable insights to improve healthcare delivery and policy decisions across different Indian states.

Dataset Description and Preprocessing

- The dataset is from HMIS, recording state-wise monthly healthcare usage and indicators such as patient attendance, lab tests, maternal health, emergency visits, and inpatient occupancy.
- Ten major states were selected for in-depth analysis:
  Kerala, Tamil Nadu, Maharashtra, Gujarat, Karnataka, Rajasthan, Uttar Pradesh, Madhya Pradesh, Bihar, and West Bengal.
- Data preprocessing steps:
    - Imported CSV dataset from the HMIS portal.
    - Standardized and trimmed column names.
    - Removed rows with missing values in critical indicators and target variables.
    - Checked and eliminated duplicate records.
    - Validated numeric data types.
    - Converted the "Month" column into datetime format and created "YearMonth" for time series analysis.
- Data was grouped state-wise and month-wise to form aggregated monthly healthcare metrics.
- Scaled features using StandardScaler to ensure equal weighting during machine learning and clustering.

Solution:


Data Sources
HMIS Portal
Monthly State Reports
Data Preprocessing Pipeline
Data Cleaning
Feature Engineering
Data Aggregation
Feature Scaling
Clean Analysis-Ready Dataset
Exploratory Data Analysis
Advanced Analytics &
Clustering
KMeans
PCA
Agglomerative
Classification
Random Forest
SVM
XGBoost
Regression & DL
Linear Regression
Random Forest
Neural Network
State Clusters
State Categories
Load Forecasts
Key Insights & Patterns
Regional Disparities
Service Correlations
Seasonal Patterns
State Clusters
Policy Applications
Resource Allocation
Targeted Interventions
Infrastructure Planning
Performance Benchmarking
Improved Healthcare Delivery

Exploratory Data Analysis (EDA)

State-Wise Healthcare Utilization Patterns

- Southern and urban states like Delhi, Kerala, and Karnataka showed consistently higher healthcare service utilization across most indicators.
- States such as Arunachal Pradesh and Manipur showed lower and more fluctuating utilization, indicating irregular access to healthcare.
- Maharashtra and Madhya Pradesh demonstrated moderate and stable usage, suggesting balanced healthcare systems.

Patient Flow and Trends

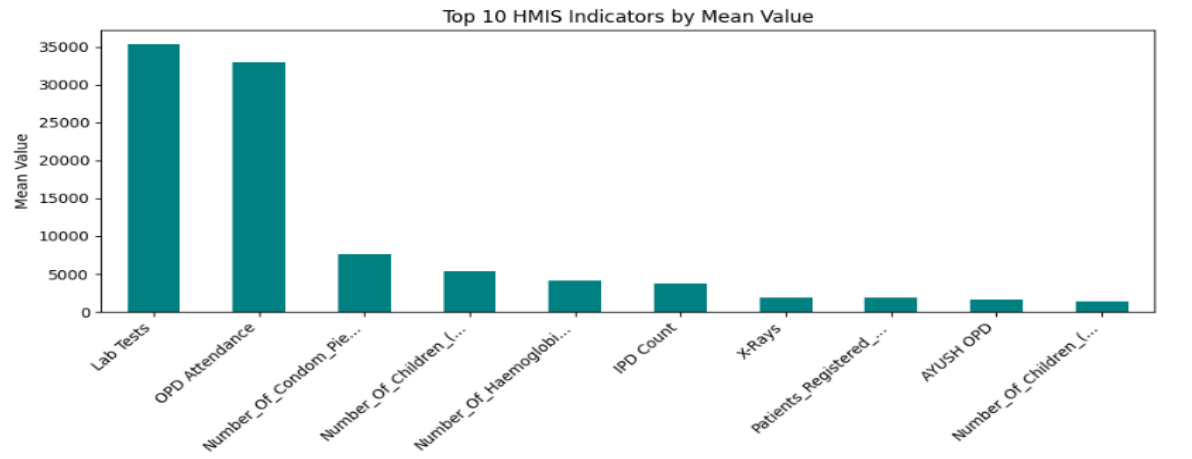
- Outpatient Department (OPD) attendance and lab tests showed strong upward trends in well-developed states.
- Emergency Department visits influenced increases in inpatient headcount, indicating hospital occupancy depends on emergency inflow.
- Seasonal variations were evident, with some months showing increased patient loads, potentially reflecting disease seasonality and rural health drives.

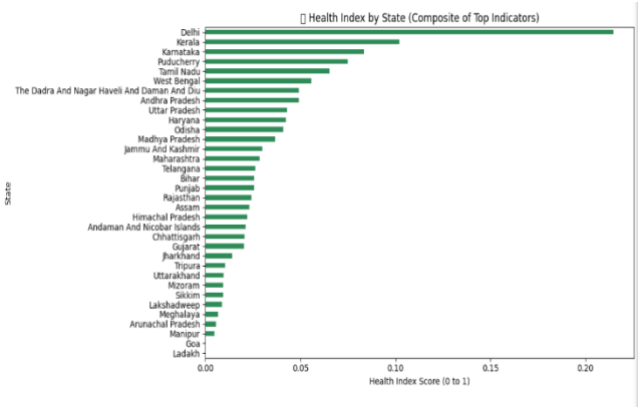
Feature Relationships and Variability

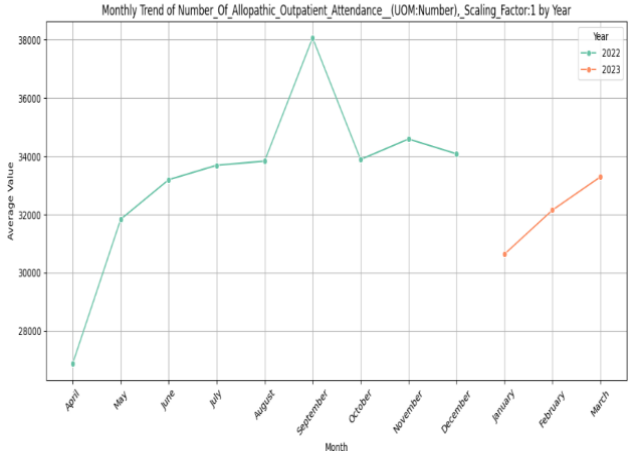
- Strong positive correlations existed between outpatient attendance, lab tests, and inpatient counts.
- Condom distribution, a public health outreach indicator, showed low correlation with inpatient occupancy but highlighted varied public health outreach effectiveness across states.
- There were significant disparities in service usage, with Kerala having better healthcare access and Bihar or Uttar Pradesh lagging behind in healthcare utilization.

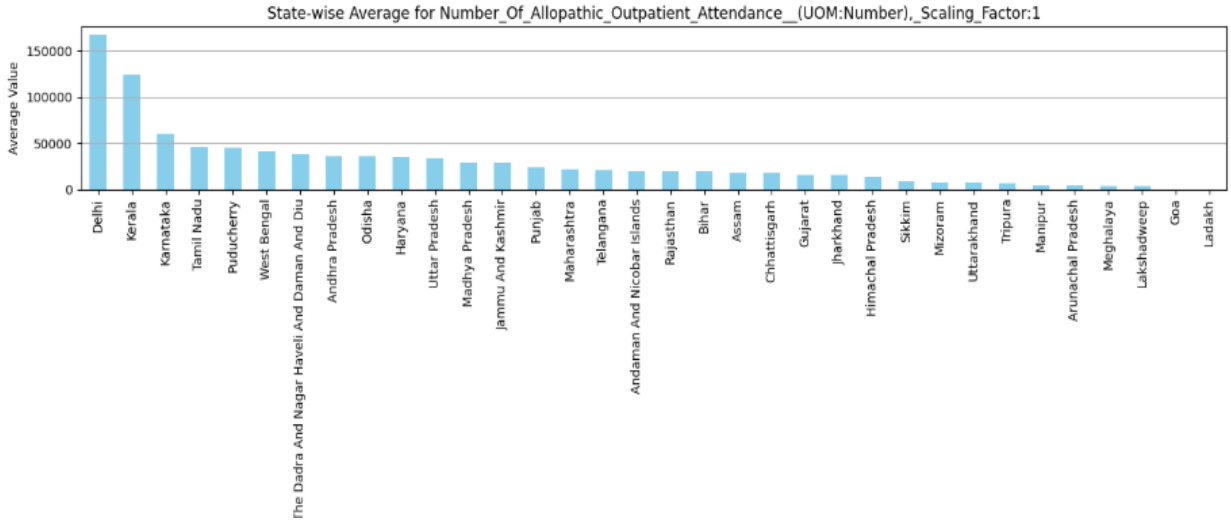
Key Attributes Used

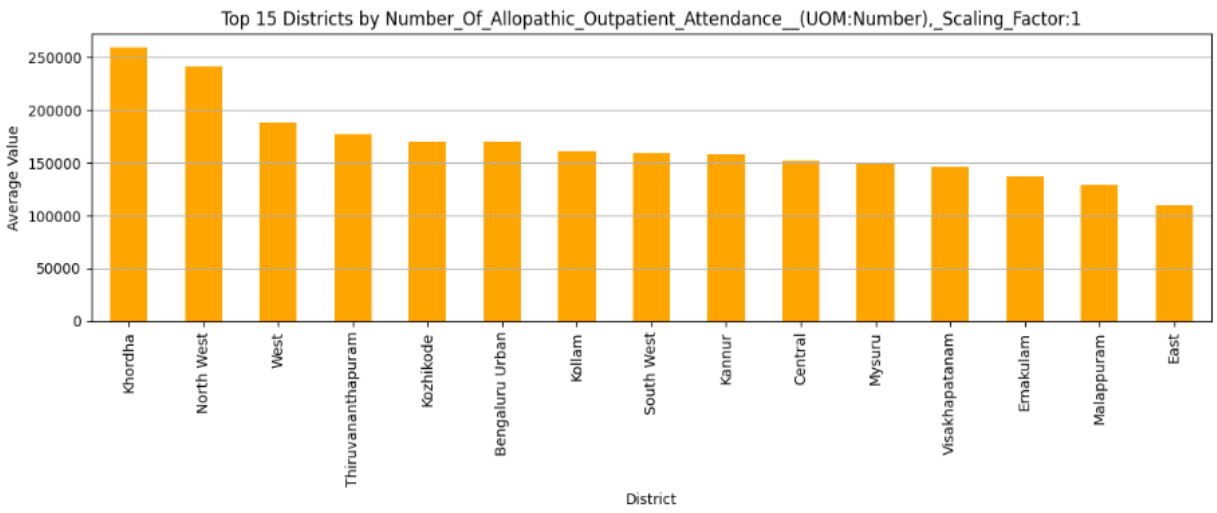
Category	Feature Name	Meaning
OPD Services	Number_Of_Allopathic_Outpatient_Attendance	Total patients who visited hospital outpatient units
Lab Diagnostics	Number_Of_Lab_Tests_Done	Total lab tests conducted
Public Health Outreach	Number_Of_Condom_Pieces_Distributed	Indicator of reproductive health awareness programs
Maternal Health	Number_Of_Haemoglobin Tests Conducted	Indicates maternal & anemia screening coverage
Emergency Care	Patients_Registered_At_Emergency_Department	Shows acute patient inflow / critical cases
Hospital Capacity / Occupancy (Target for Regression)	In-Patient_Head_Count_At_Midnight	Number of admitted patients at midnight → reflects bed occupancy


Top 10 HMIS Indicators by Mean Value
Mean Value
35000
30000
25000
20000
15000
10000
5000
0
Lab Tests
OPD Attendance
Number_Of_Condom_Pie...
Number_Of_Children_(...
Number_Of_Haemoglobi...
IPD Count
X-Rays
Patients_Registered_...
AYUSH OPD
Number_Of_Children_(...


Health Index by State (Composite of Top Indicators)
State
Delhi
Kerala
Karnataka
Puducherry
Tamil Nadu
West Bengal
The Dadra And Nagar Haveli And Daman And Diu
Andhra Pradesh
Uttar Pradesh
Haryana
Odisha
Madhya Pradesh
Jammu And Kashmir
Maharashtra
Telangana
Bihar
Punjab
Rajasthan
Assam
Himachal Pradesh
Andaman And Nicobar Islands
Chhattisgarh
Gujarat
Jharkhand
Tripura
Uttarakhand
Mizoram
Sikkim
Lakshadweep
Meghalaya
Arunachal Pradesh
Manipur
Goa
Ladakh
0.00
0.05
0.10
0.15
0.20
Health Index Score (0 to 1)

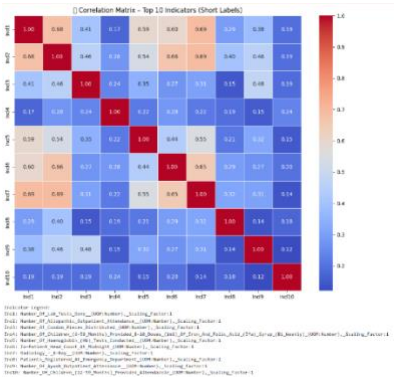

Monthly Trend of Number_Of_Allopathic_Outpatient_Attendance__(UOM:Number),_Scaling_Factor:1 by Year
Year
2022
2023
Average Value
38000
36000
34000
32000
30000
28000
April
May
June
July
August
September
October
November
December
January
February
March
Month

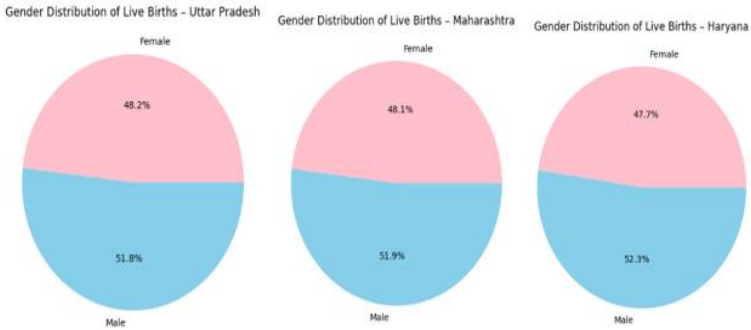

State-wise Average for Number_Of_Allopathic_Outpatient_Attendance__(UOM:Number),_Scaling_Factor:1
Average Value
150000
100000
50000
0
Delhi
Kerala
Karnataka
Tamil Nadu
Puducherry
West Bengal
The Dadra And Nagar Haveli And Daman And Diu
Andhra Pradesh
Odisha
Haryana
Uttar Pradesh
Madhya Pradesh
Jammu And Kashmir
Punjab
Maharashtra
Telangana
Andaman And Nicobar Islands
Rajasthan
Bihar
Assam
Chhattisgarh
Gujarat
Jharkhand
Himachal Pradesh
Sikkim
Mizoram
Uttarakhand
Tripura
Manipur
Arunachal Pradesh
Meghalaya
Lakshadweep
Goa
Ladakh



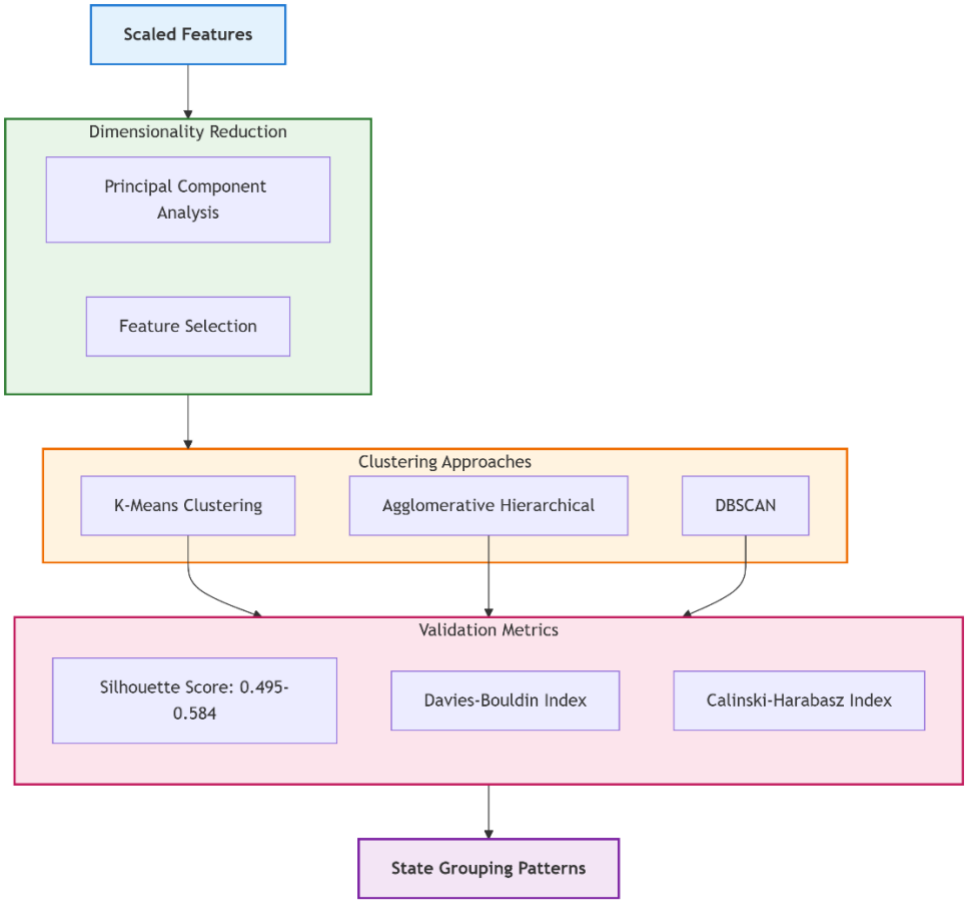

```
# Filter high correlation pairs
high_corr = correlations[correlations > 0.60]
print(" Highly Correlated Pairs:\n", high_corr)
```

```
 Highly Correlated Pairs:
 Number_Of_Lab_Tests_Done__(UOM:Number),_Scaling_Factor:1                 Radiology_-_X-Ray__(UOM:Number),_Scaling_Factor:1
0.692859
Radiology_-_X-Ray__(UOM:Number),_Scaling_Factor:1                        Number_Of_Lab_Tests_Done__(UOM:Number),_Scaling_Factor:1
0.692859
                                                                         Number_Of_Allopathic_Outpatient_Attendance__(UOM:Number),_Scaling_Factor:1
0.691825
Number_Of_Allopathic_Outpatient_Attendance__(UOM:Number),_Scaling_Factor:1  Radiology_-_X-Ray__(UOM:Number),_Scaling_Factor:1
0.691825
                                                                         Number_Of_Lab_Tests_Done__(UOM:Number),_Scaling_Factor:1
0.679363
Number_Of_Lab_Tests_Done__(UOM:Number),_Scaling_Factor:1                 Number_Of_Allopathic_Outpatient_Attendance__(UOM:Number),_Scaling_Factor:1
0.679363
Number_Of_Allopathic_Outpatient_Attendance__(UOM:Number),_Scaling_Factor:1  In-Patient_Head_Count_At_Midnight_(UOM:Number),_Scaling_Factor:1
0.664315
In-Patient_Head_Count_At_Midnight_(UOM:Number),_Scaling_Factor:1         Number_Of_Allopathic_Outpatient_Attendance__(UOM:Number),_Scaling_Factor:1
0.664315
Radiology_-_X-Ray__(UOM:Number),_Scaling_Factor:1                        In-Patient_Head_Count_At_Midnight_(UOM:Number),_Scaling_Factor:1
0.647415
In-Patient_Head_Count_At_Midnight_(UOM:Number),_Scaling_Factor:1          Radiology_-_X-Ray__(UOM:Number),_Scaling_Factor:1
0.647415
dtype: float64
```





Clustering Analysis

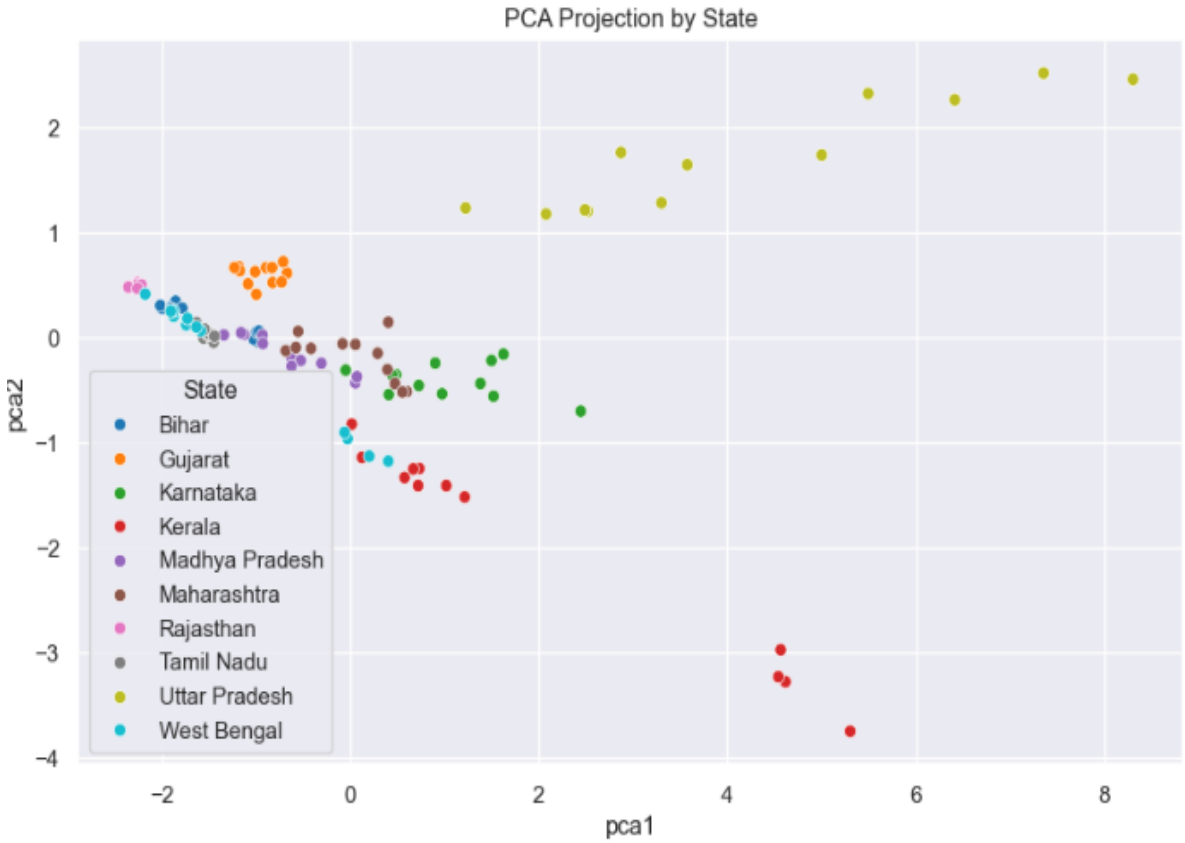


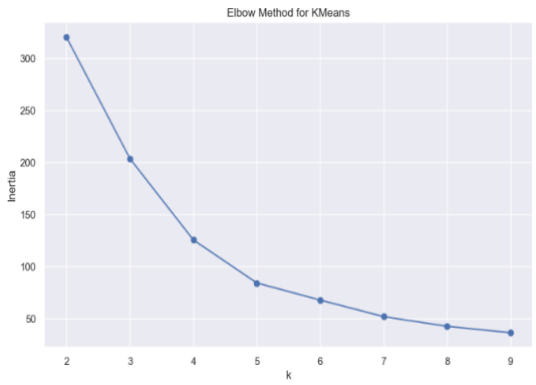
Methods Used

- Principal Component Analysis (PCA) for dimensionality reduction.
- KMeans clustering to group states based on healthcare indicators.
- Agglomerative hierarchical clustering validated KMeans findings through dendrograms showing hierarchical state groupings.
- DBSCAN clustering identified core clusters and outliers with states like Kerala and West Bengal appearing as distinct due to unique healthcare metrics.

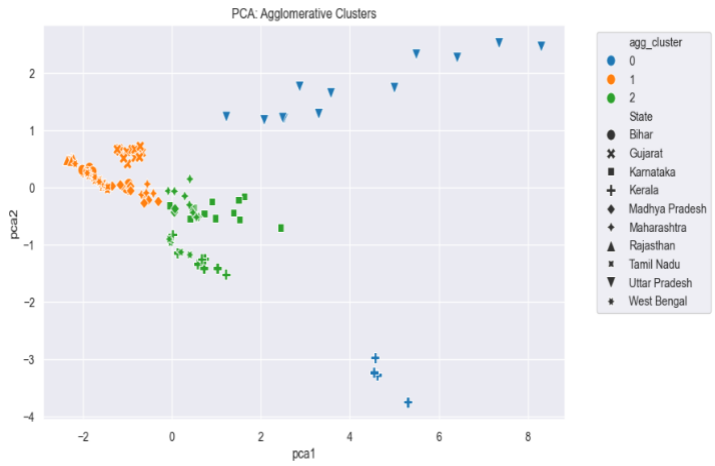
Cluster Evaluation

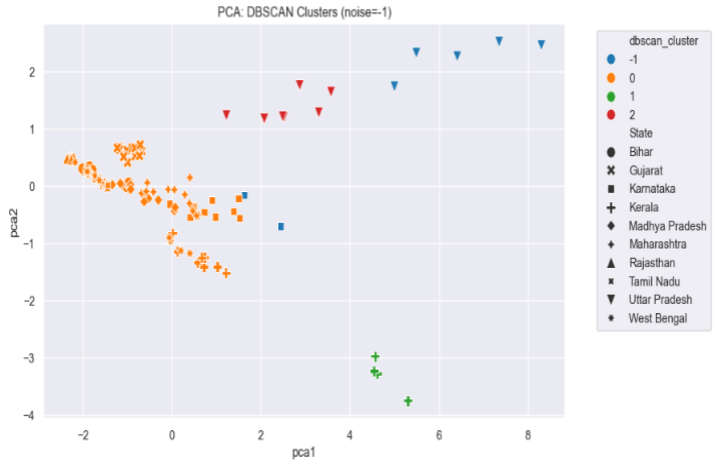
- Silhouette scores indicated satisfactory separation: KMeans and Agglomerative scored around 0.495, while DBSCAN showed better separation at 0.584.
- Davies-Bouldin Index (DBI) and Caliński-Harabasz Index (CHI) also supported cluster quality.

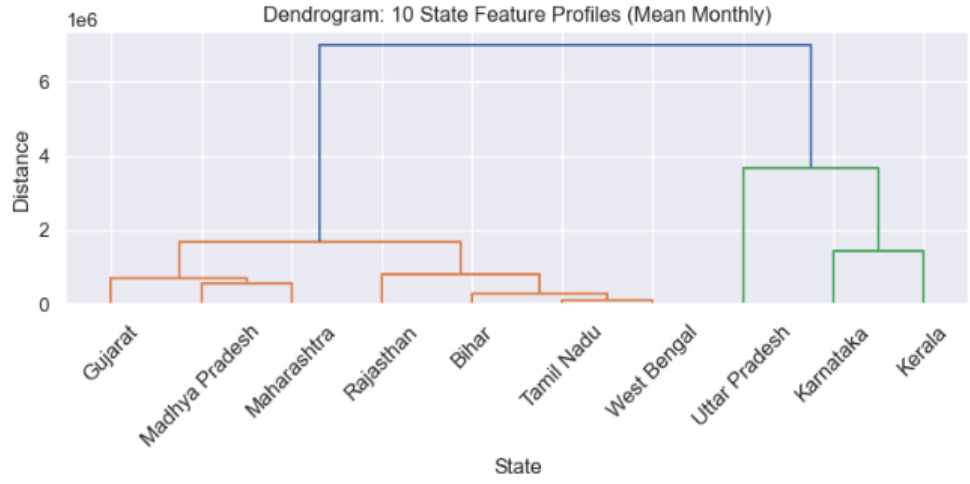

PCA Projection by State
State
Bihar
Gujarat
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Rajasthan
Tamil Nadu
Uttar Pradesh
West Bengal
pca1
pca2


Elbow Method for KMeans
Inertia
k


PCA: KMeans Clusters
kmeans_cluster
0
1
2
State
Bihar
Gujarat
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Rajasthan
Tamil Nadu
Uttar Pradesh
West Bengal
pca1
pca2



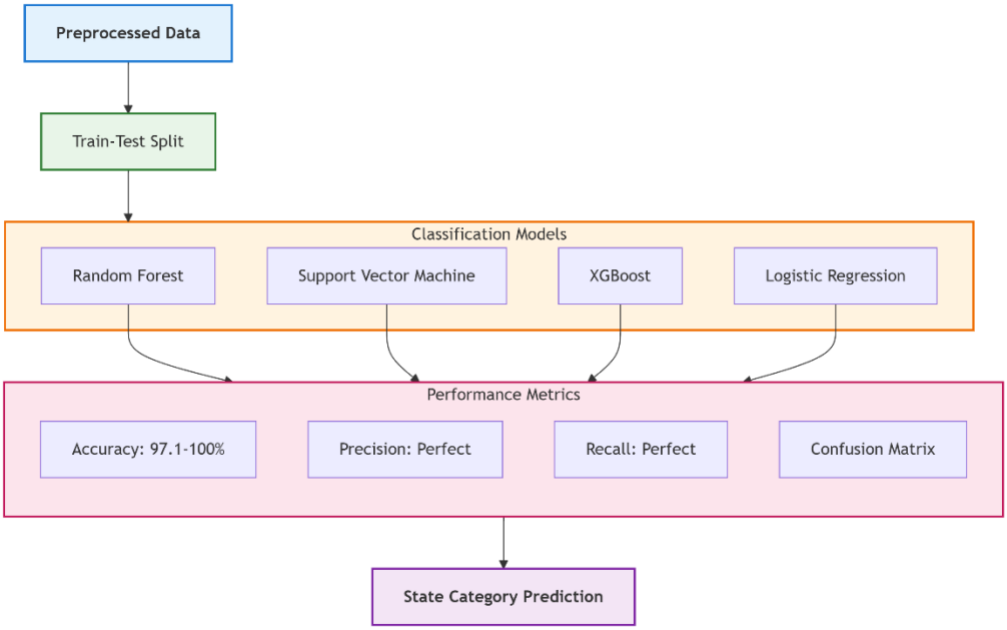




Cluster Evaluation Metrics:

- **KMeans**: Silhouette=0.495 | DBI=0.843 | CHI=133.832
- **Agglomerative**: Silhouette=0.495 | DBI=0.865 | CHI=125.829
- **DBSCAN**: Silhouette=0.584 | DBI=0.440 | CHI=74.935
- The silhouette score (~0.61) indicates a reasonably good cluster structure for the DBSCAN result, with clusters being well-separated and data points within each cluster being simi

Classification Models



Models Employed

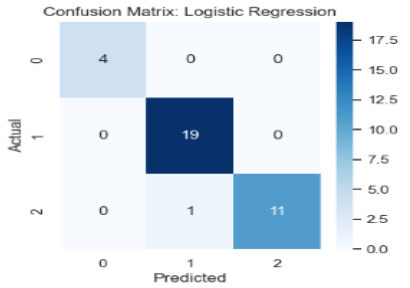
- Random Forest
- Support Vector Machine (SVM)
- Logistic Regression
- XGBoost

	Model	Accuracy
0	Logistic Regression	0.971429
1	Random Forest	1.000000
2	SVM	1.000000
3	XGBoost	1.000000

Results and Interpretation

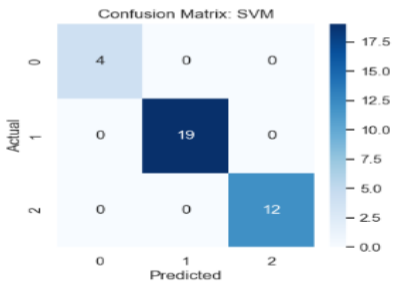
- SVM, Random Forest, and XGBoost achieved perfect accuracy, recall, and precision across all classes.
- Logistic Regression showed a slightly lower accuracy (97.1%), with only one misclassification.
- There was no class imbalance issue; model performances indicate strong feature discriminative power.
- Confusion matrices confirmed robustness and stability for potential deployment in healthcare decision systems.

```
Logistic Regression Accuracy: 0.971
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       0.95      1.00      0.97        19
           2       1.00      0.92      0.96        12

    accuracy                           0.97        35
   macro avg       0.98      0.97      0.98        35
weighted avg       0.97      0.97      0.97        35
```



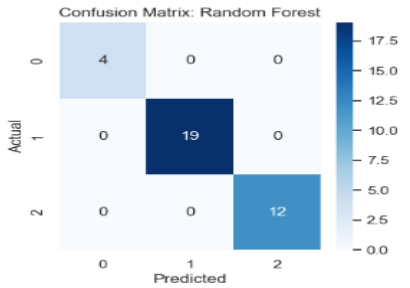

```
Logistic Regression: 5-Fold CV Accuracy = 0.922
```

```
SVM Accuracy: 1.000
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00        19
           2       1.00      1.00      1.00        12

    accuracy                           1.00        35
   macro avg       1.00      1.00      1.00        35
weighted avg       1.00      1.00      1.00        35
```



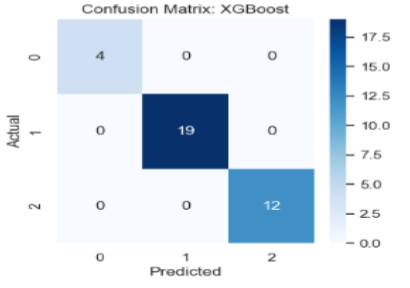

```
SVM: 5-Fold CV Accuracy = 0.930
```

```
Random Forest Accuracy: 1.000
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00        19
           2       1.00      1.00      1.00        12

    accuracy                           1.00        35
   macro avg       1.00      1.00      1.00        35
weighted avg       1.00      1.00      1.00        35
```



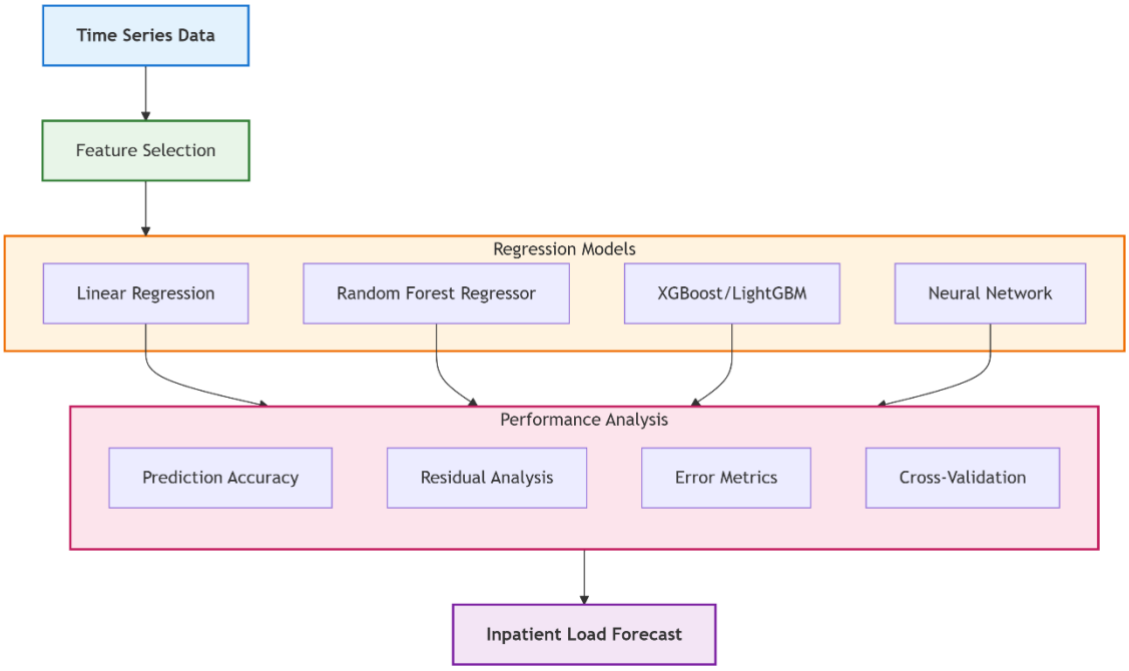

```
Random Forest: 5-Fold CV Accuracy = 0.939
```

```
XGBoost Accuracy: 1.000
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00        19
           2       1.00      1.00      1.00        12

    accuracy                           1.00        35
   macro avg       1.00      1.00      1.00        35
weighted avg       1.00      1.00      1.00        35
```




```
XGBoost: 5-Fold CV Accuracy = 0.939
```

Regression and Deep Learning Models



Regression

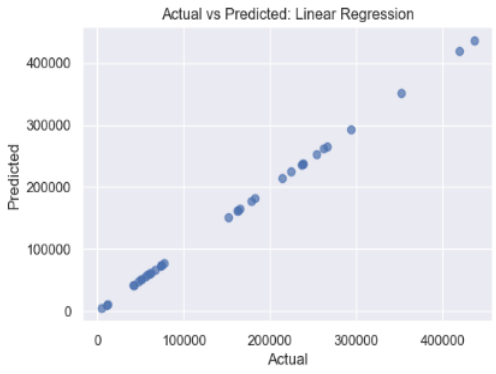
- Models used included Linear Regression, Random Forest Regressor, XGBoost, and LightGBM.
- Time series analysis revealed states like Maharashtra and Tamil Nadu consistently had the highest inpatient head counts.
- Kerala, Rajasthan, and West Bengal reported substantially lower inpatient numbers, reflecting regional policy and infrastructure differences.

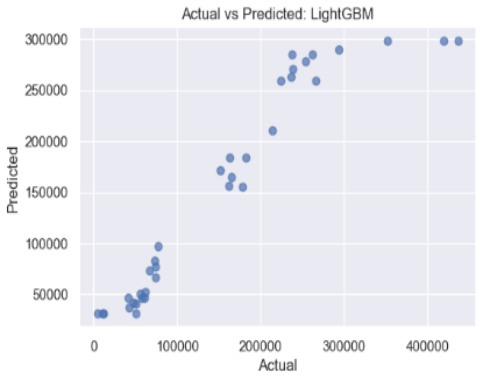
Deep Learning

- A feed-forward neural network was used for regression predictions.
- Scatter plots of actual versus predicted values showed most predictions closely aligned with true values, indicating good model fit.
- Residual plots showed residuals symmetrically scattered around zero, suggesting no systemic bias in predictions and overall model accuracy.

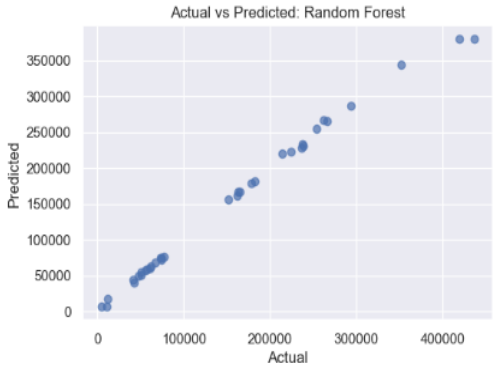
Regression:

```
Linear Regression: MAE=0.00, RMSE=0.00, R²=1.000
```

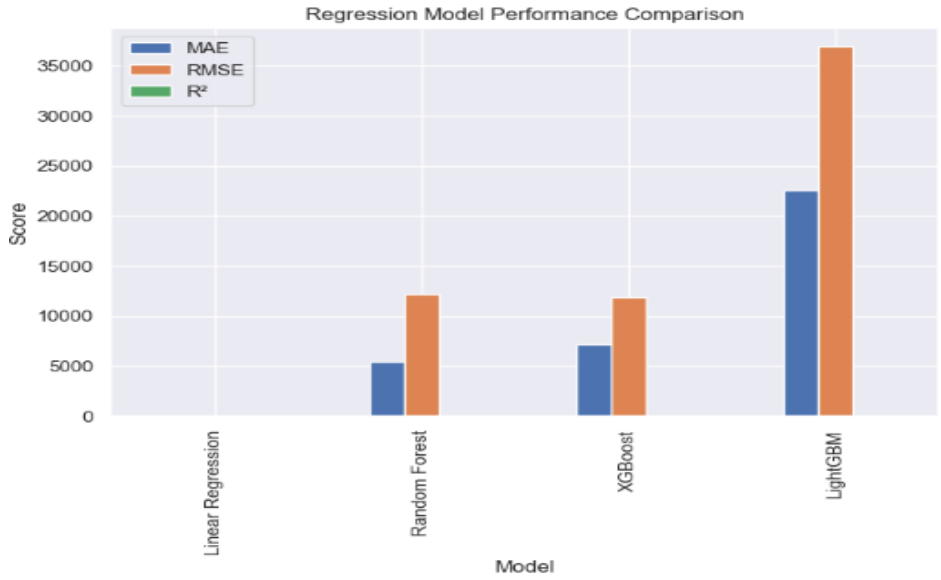


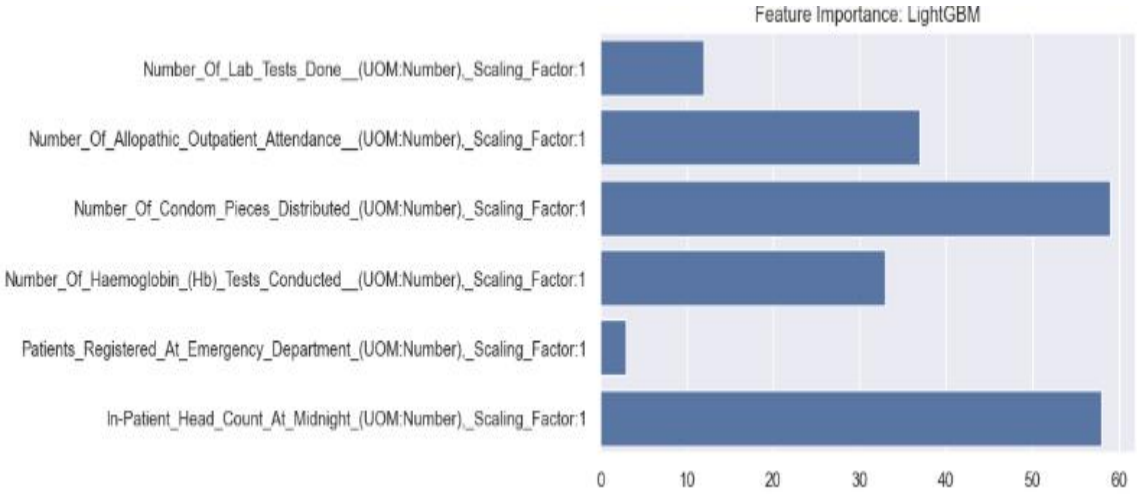


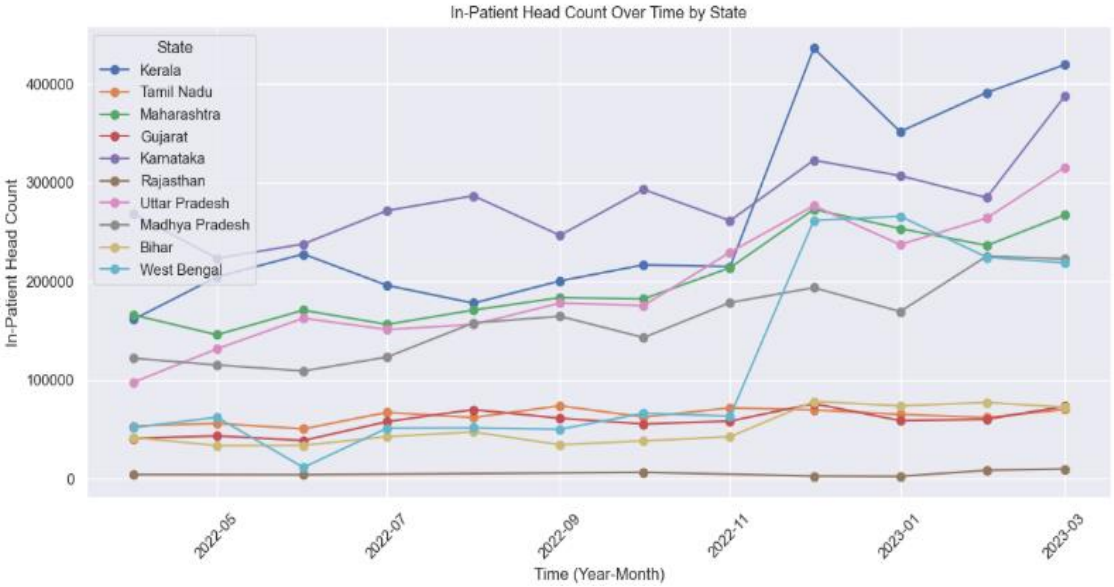

```
Random Forest: MAE=5420.46, RMSE=12178.51, R²=0.989
```



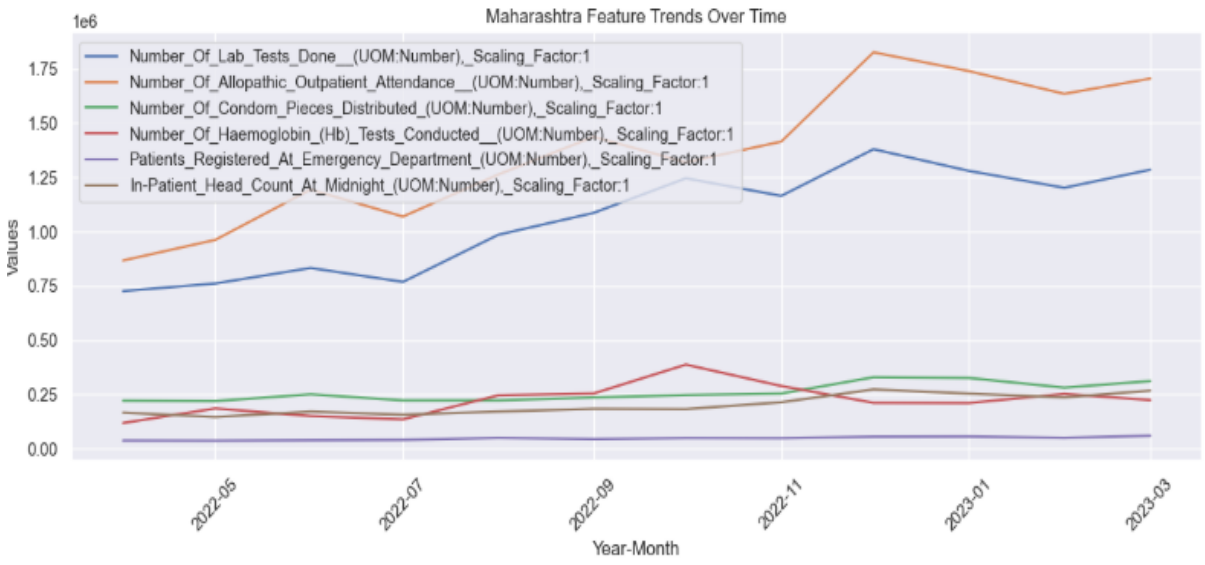
	Model	MAE	RMSE	R^2
0	Linear Regression	1.141286e-10	1.738634e-10	1.000000
1	Random Forest	5.420462e+03	1.217851e+04	0.988802
2	XGBoost	7.115779e+03	1.184785e+04	0.989401
3	LightGBM	2.252098e+04	3.691164e+04	0.897128

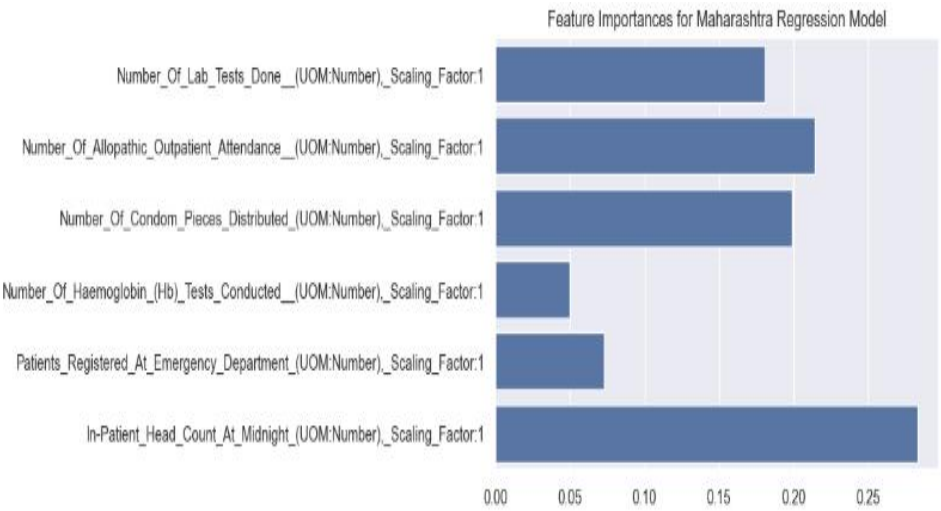




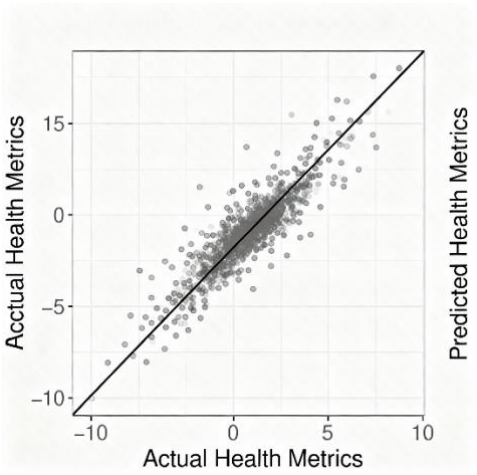


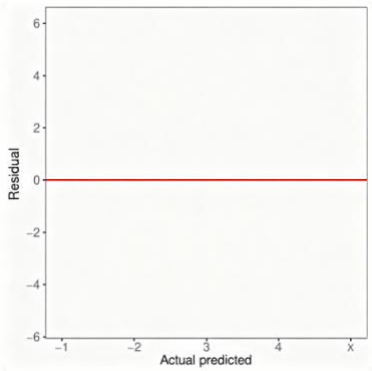
For Maharashtra:





Deep Learning





Key Insights and Implications

- The data pipeline is effective for pattern recognition, classification, and hospital load prediction based on HMIS data.
- Strong state-wise variation in healthcare usage highlights infrastructure inequality, with southern states better equipped than northern states.
- Robust model performance supports their use in operational healthcare planning and policymaking.
- Clustering helps identify similar healthcare states, enabling targeted policy interventions.
- High classification accuracy ensures reliable state-level healthcare status identification.
- Deep learning regression provides accurate, low-bias inpatient load forecasts crucial for resource allocation.