Data Science Project Report: Analysis of Maternal Healthcare Data in India (2008-2019)

Executive Summary

This report presents a comprehensive data science analysis of maternal healthcare indicators across Indian districts from 2008 to 2019. The project employs exploratory data analysis, regression modeling, classification algorithms, and clustering techniques to understand patterns in antenatal care, institutional deliveries, and infant health outcomes. The findings reveal significant regional disparities and identify key factors influencing maternal healthcare quality.

Table of Contents

1. Introduction

Background

Maternal and infant healthcare remains a critical public health priority in India. Programs like Janani Suraksha Yojana (JSY) and Antenatal Care (ANC) initiatives aim to improve healthcare access and outcomes for pregnant women and newborns across the country.

Objectives

This analysis aims to:

- Understand patterns and trends in maternal healthcare indicators
- Identify factors influencing infant mortality and institutional deliveries
- Classify districts based on healthcare performance
- Discover natural groupings of districts with similar healthcare characteristics

2. Dataset Overview

Data Source

The dataset contains maternal healthcare records from 2008-2019 covering various Indian states and districts.

Dataset Dimensions

- **Rows**: 8,285 records
- **Columns**: 200 features
- **Memory Usage**: 14.73 MB

Key Features Include:

- Women registered for Antenatal Care (ANC)
- Institutional vs. home deliveries
- Janani Suraksha Yojana (JSY) registrations
- Infant deaths reported
- C-section deliveries
- Post-partum care statistics

---

3. Data Preprocessing

Process

Data preprocessing is the foundational step that ensures data quality and prepares it for analysis.

Steps Performed:

1. **Column Name Standardization**

○ Converted all column names to lowercase
○ Removed special characters and extra spaces
○ Replaced spaces with underscores for easier coding

2. **Missing Value Treatment**

○ Identified columns with missing data
○ Filled numerical columns with median values
○ Filled categorical columns with mode (most frequent value)
○ Rationale: Median is robust to outliers; mode preserves categorical integrity

3. **Duplicate Removal**

○ Checked for and removed duplicate records
○ Result: 0 duplicates found

4. **Data Type Conversion**

○ Converted categorical variables appropriately
○ Ensured numerical columns were in proper format

Why This Process is Important

Raw data often contains inconsistencies, missing values, and formatting issues that can lead to incorrect analysis. Preprocessing ensures:

- Consistency across all data entries
- Prevention of errors during analysis
- Improved model performance
- Reliable and valid conclusions

Output

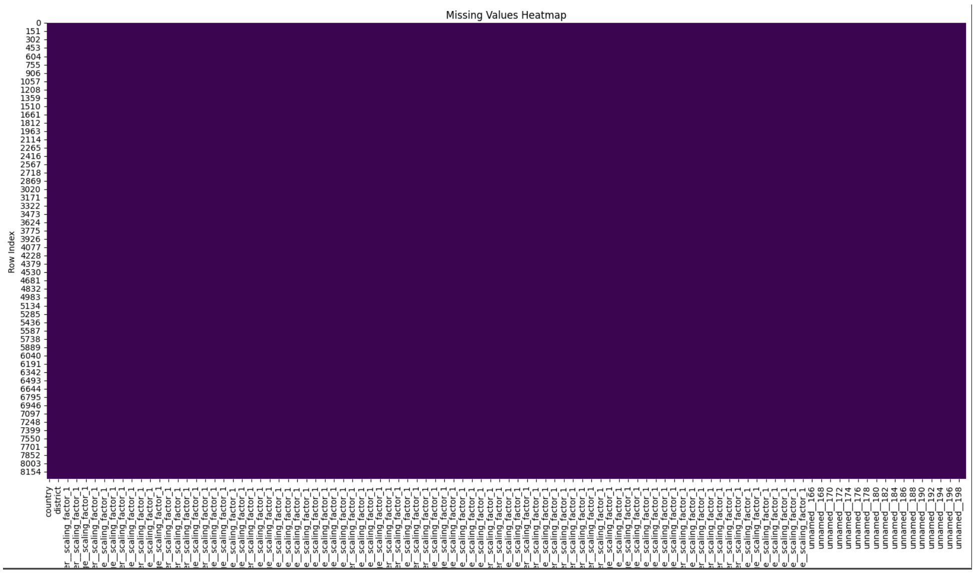


Figure 1: Missing Values Heatmap

Observations

- Several columns had 100% missing values (shown as complete vertical bands in the heatmap)
- Most healthcare indicator columns had less than 5% missing data
- Missing data was more prevalent in optional metrics rather than core indicators

Conclusion

After preprocessing, the dataset was reduced to relevant columns with complete or imputed data, ready for analysis. The cleaned dataset maintained data integrity while removing noise and inconsistencies.

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

Process

Univariate analysis examines each variable independently to understand its distribution, central tendency, and spread.

Why This is Performed

Understanding individual variables helps identify:

- Data distribution patterns (normal, skewed, etc.)
- Outliers and unusual values
- Range and variability of data
- Potential data quality issues

Output Images

Figure 2: Distribution of Records by State

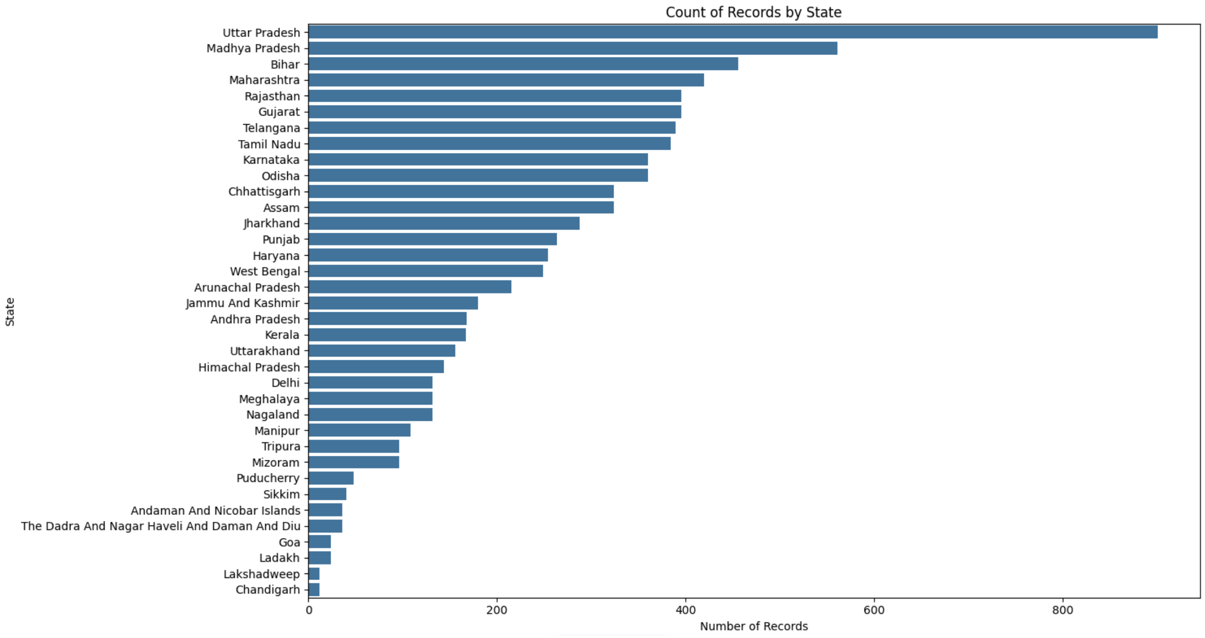
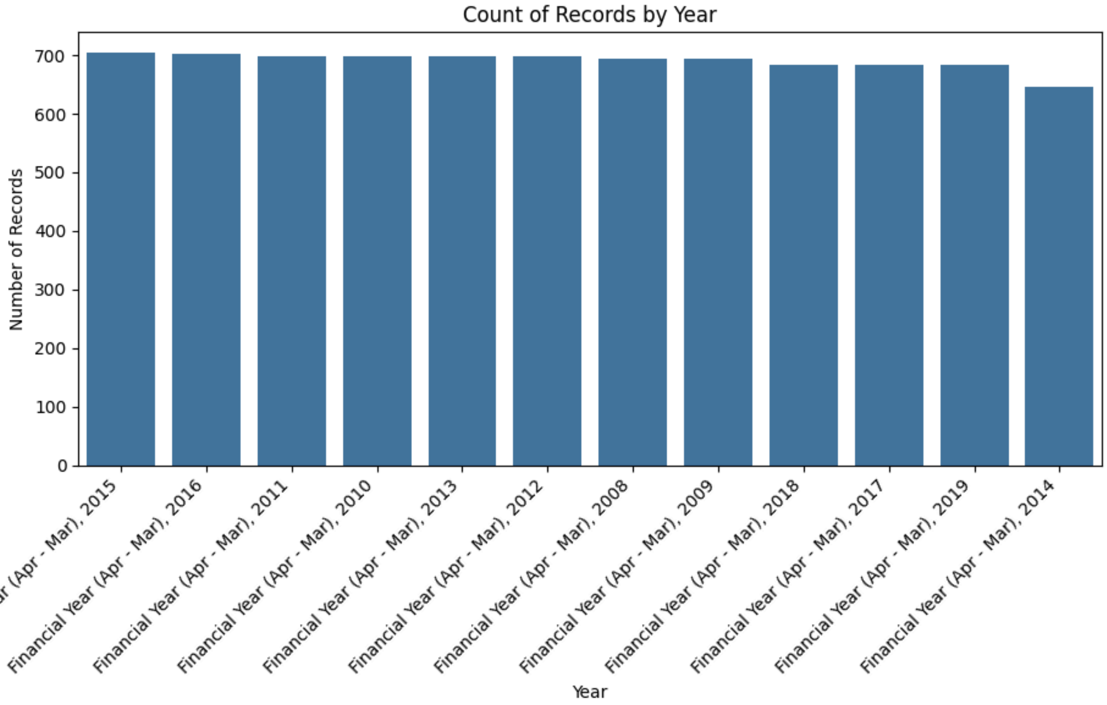


Figure 3: Distribution of Records by Year



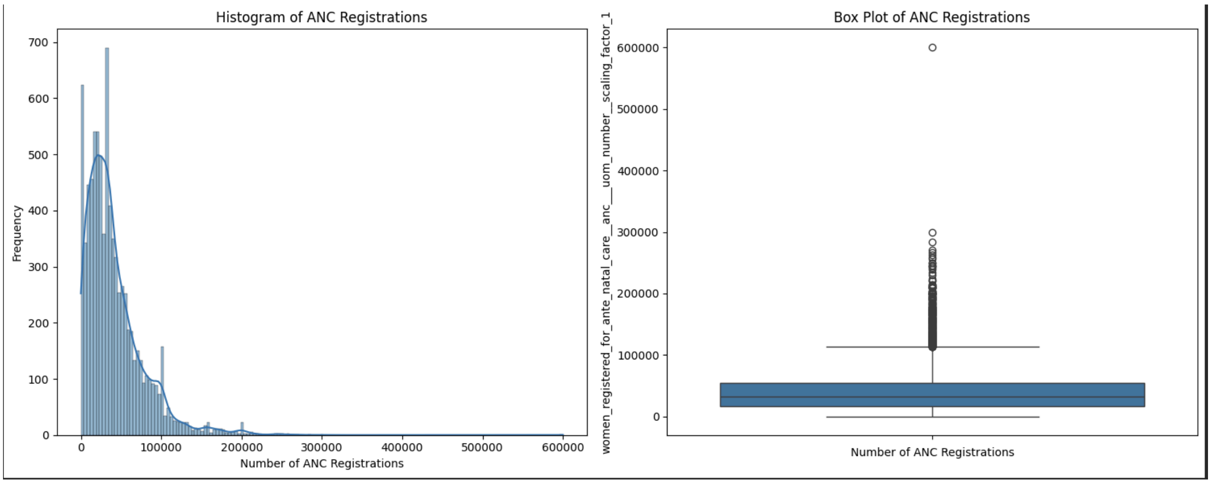
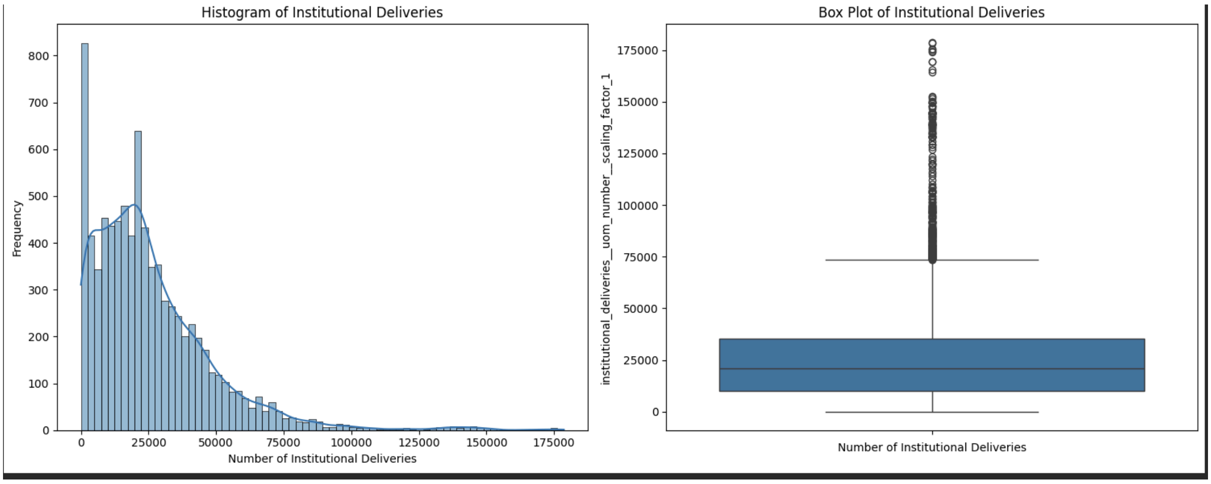


Figure 4: ANC Registrations Distribution

Figure 5: Institutional Deliveries Distribution



Observations

1. **State-wise Distribution**: Uttar Pradesh has the highest number of records (901), indicating either larger population or better data collection
2. **Temporal Distribution**: Year 2015 shows the highest number of records (704), suggesting improved data collection over time
3. **ANC Registrations**:
   ○ Mean: 41,383 registrations per district
   ○ Distribution is right-skewed with significant outliers
   ○ Most districts have 16,000-55,000 registrations
4. **Institutional Deliveries**:
   ○ Similar right-skewed pattern
   ○ Presence of many outliers indicates high-performing districts
   ○ Box plot shows median around 30,000 deliveries

Conclusion

The univariate analysis reveals substantial variation in healthcare metrics across districts. The right-skewed distributions suggest that while most districts have moderate performance, a few excel significantly. This variation warrants further investigation into factors driving these differences.

4.2 Bivariate and Multivariate Analysis

Process

This analysis explores relationships between multiple variables to identify correlations, trends, and dependencies.

Why This is Performed

Understanding relationships between variables helps:

- Identify which factors influence outcomes
- Discover patterns not visible in single-variable analysis
- Inform predictive modeling decisions
- Reveal geographical and temporal trends

Output Images

Figure 6: Correlation Matrix of Key Health Indicators

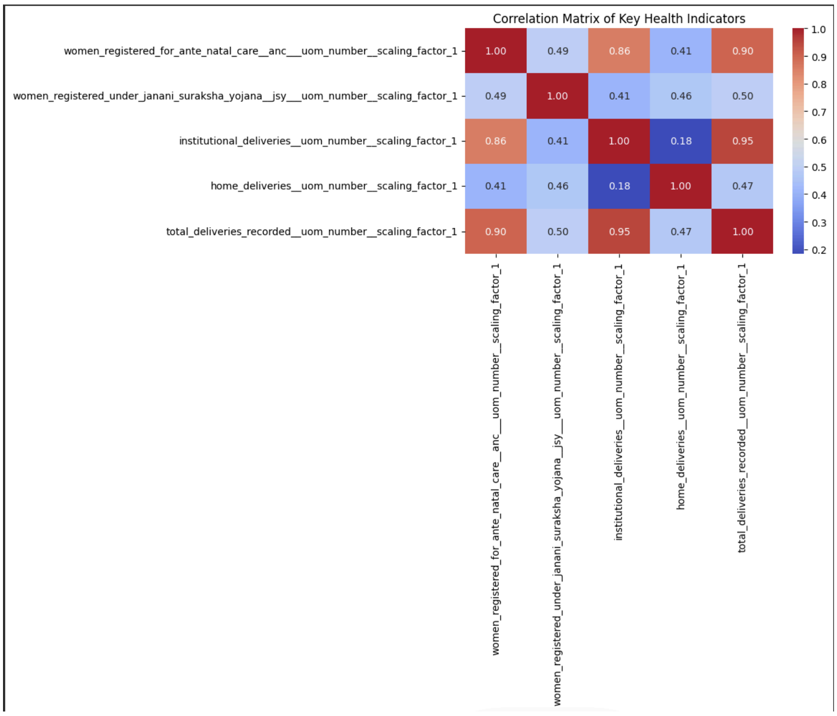


Figure 7: Total Institutional Deliveries by State

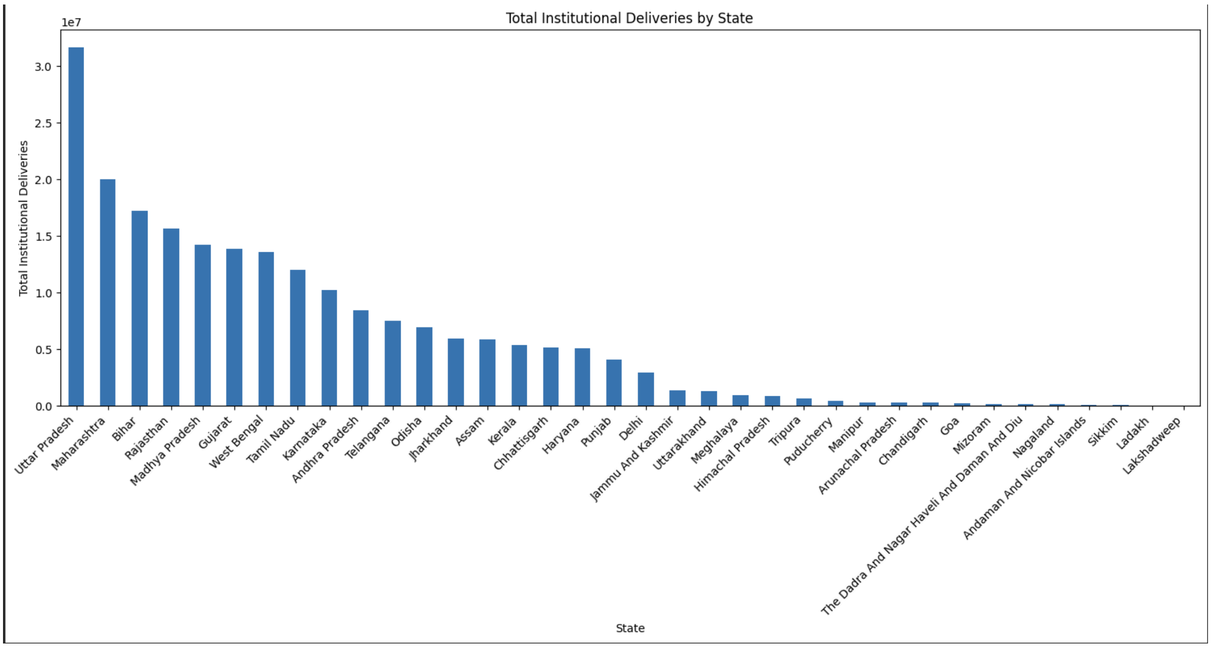
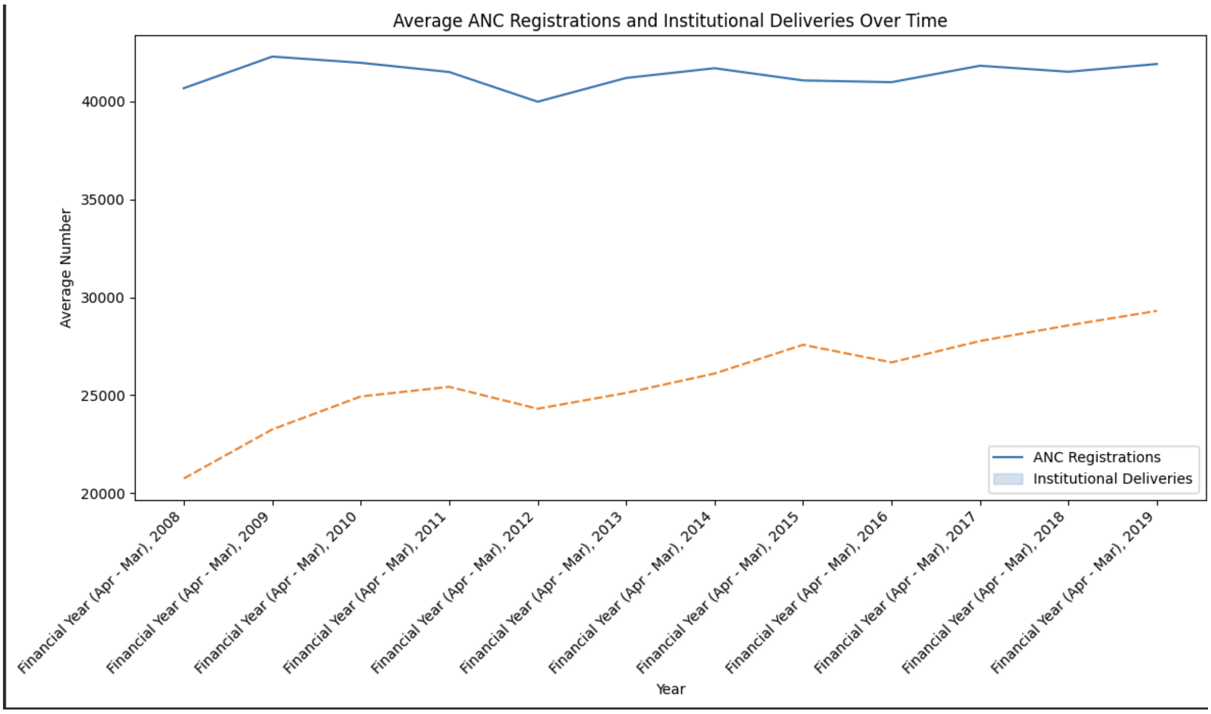


Figure 8: Time Trends in ANC and Deliveries



Observations

1. **Correlation Analysis**:

○ Strong positive correlation (0.98) between ANC registrations and JSY registrations
○ High correlation (0.94) between institutional deliveries and total deliveries
○ Home deliveries show negative correlation with institutional deliveries

2. **Geographic Trends**:

○ Uttar Pradesh leads in total institutional deliveries
○ Significant disparities exist between states

- Some smaller states show better per-capita performance

3. **Temporal Trends**:

- Clear upward trend from 2008 to 2019
- Average ANC registrations increased steadily
- Institutional deliveries grew faster than ANC registrations
- Indicates improving healthcare infrastructure and awareness

Conclusion

The bivariate analysis confirms that maternal healthcare indicators are highly interdependent. The positive trend over time demonstrates the success of government initiatives like JSY. However, regional disparities remain significant and require targeted interventions.

4.3 Additional Relationship Analysis

Process

Focused analysis on specific variable pairs to understand their direct relationships.

Output Images

Figure 9: ANC Registrations vs. Institutional Deliveries

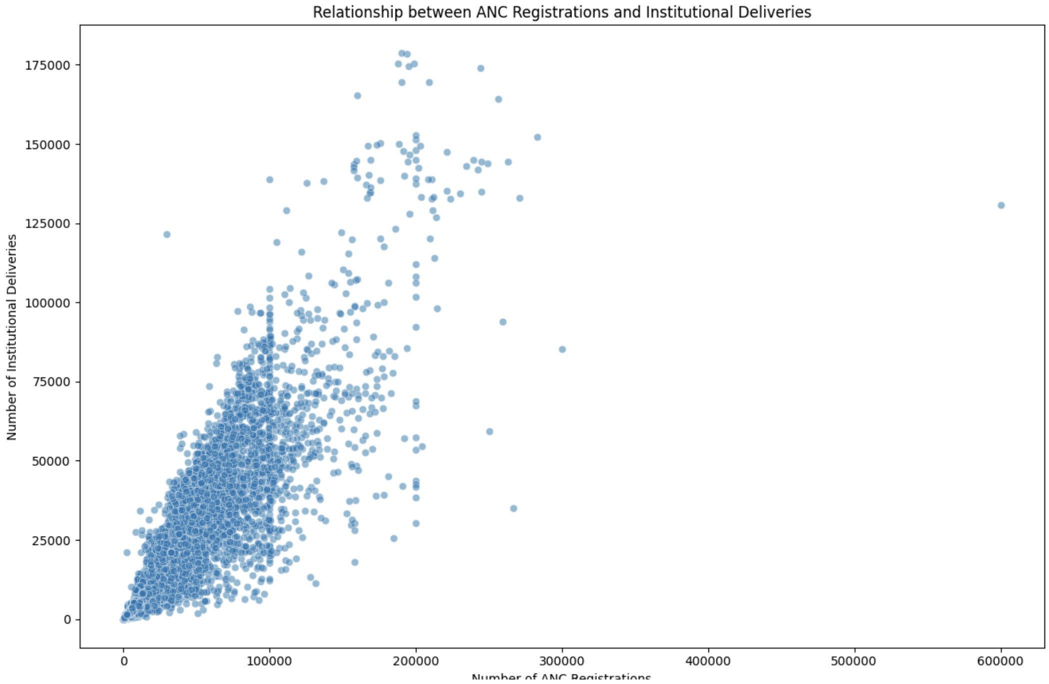


Figure 10: C-section Deliveries by Facility Type

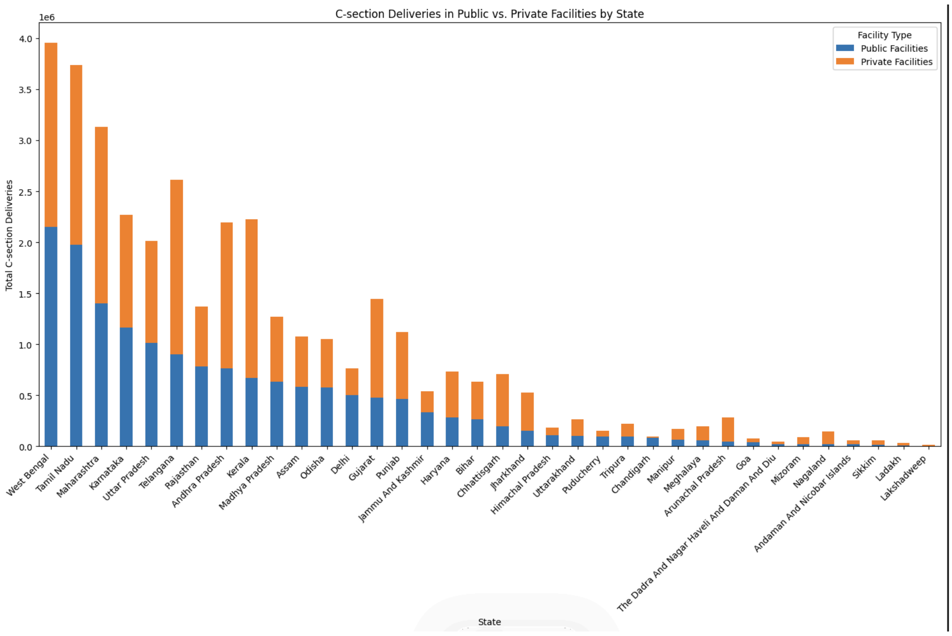


Figure 11: ANC vs. JSY Registrations

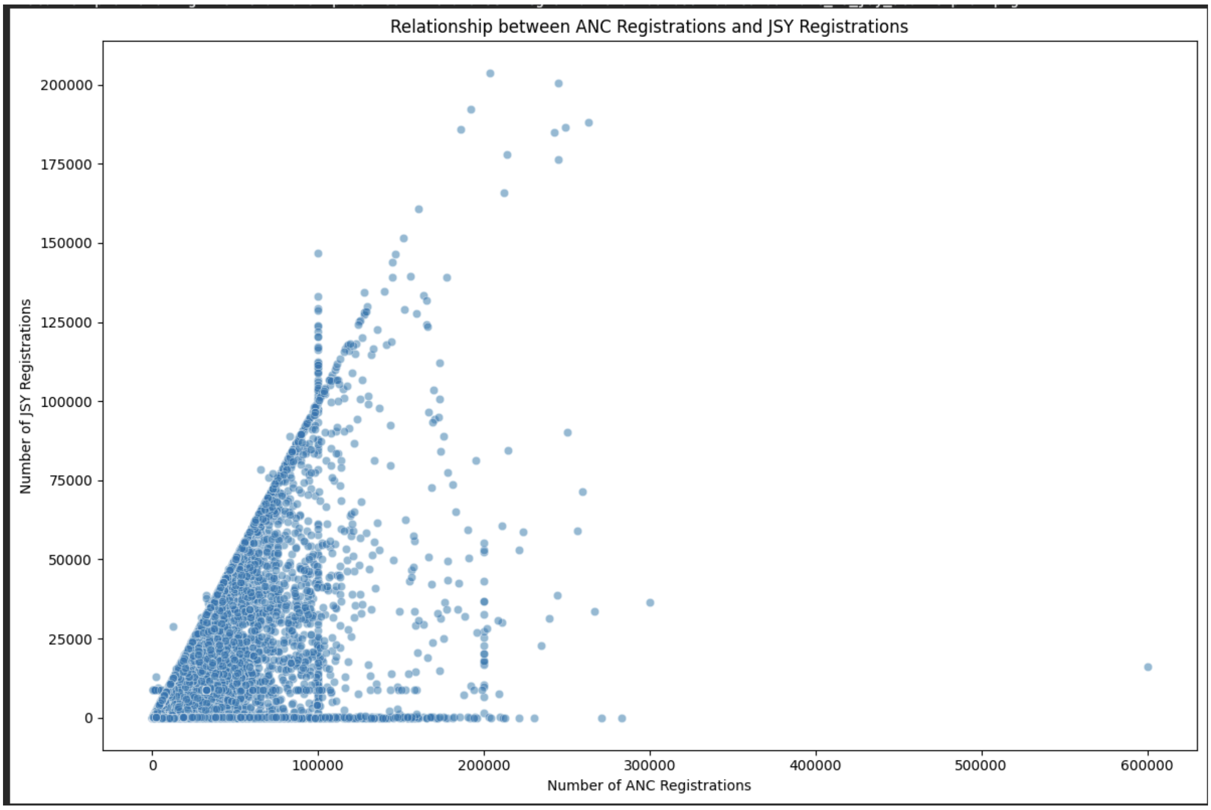
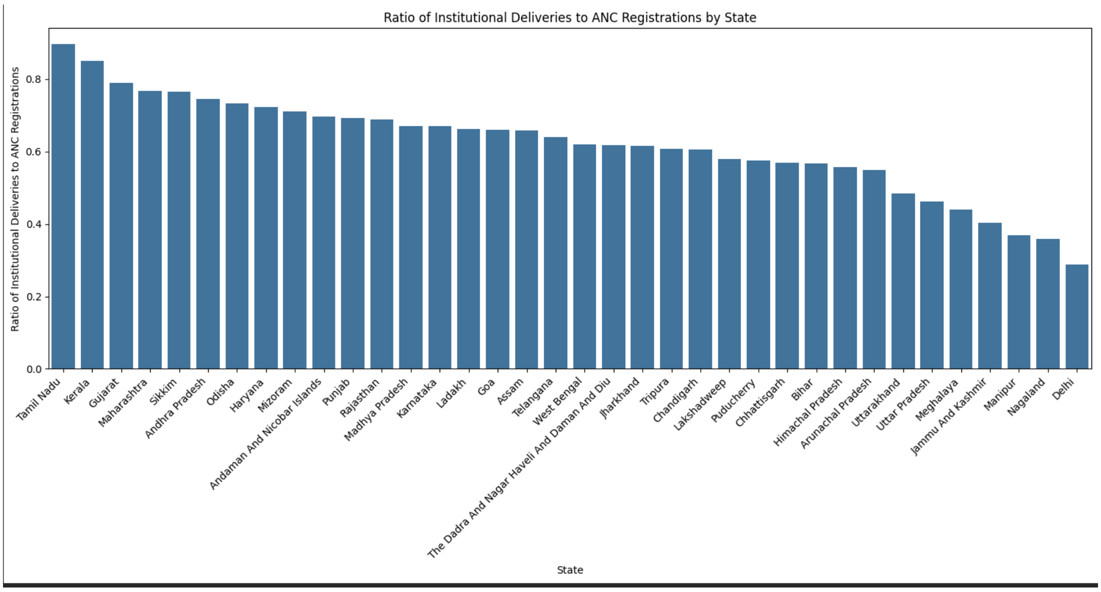


Figure 12: Delivery to ANC Ratio by State



Observations

1. **ANC and Institutional Deliveries**: Strong linear relationship confirms that better antenatal care leads to more institutional deliveries

2. **C-section Analysis**:

○ Uttar Pradesh performs most C-sections in public facilities
○ Tamil Nadu and Madhya Pradesh show higher private facility usage
○ Indicates varying healthcare infrastructure models

3. **JSY Program Effectiveness**: Districts with higher ANC registrations also show higher JSY participation, indicating successful program integration

4. **Delivery-to-ANC Ratio**: Some states achieve >90% conversion, while others lag, suggesting varying healthcare access and quality

Conclusion

These targeted analyses reveal that healthcare program success depends on integrated service delivery. States effectively linking ANC with institutional delivery services achieve better outcomes.

5. Regression Analysis

5.1 Simple Linear Regression

Process

Linear regression models the relationship between predictor variables and a continuous outcome variable using a straight-line equation.

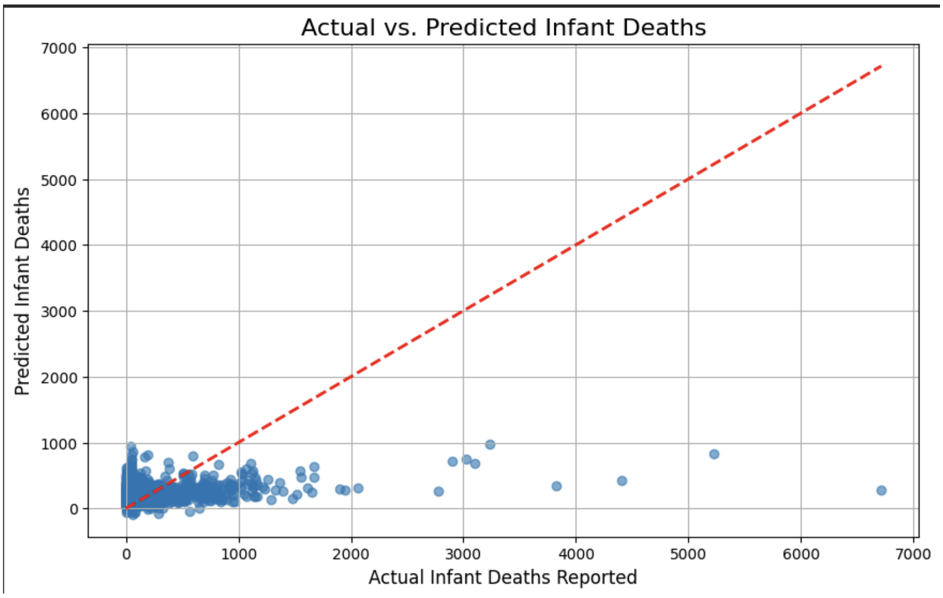
Why This is Performed

To predict infant deaths based on delivery and ANC registration data, and to quantify the relationship strength between these variables.

Target Variable: Infant deaths reported
Features: Total deliveries recorded, Women registered for ANC

Output Image

Figure 13: Linear Regression - Actual vs. Predicted



Observations

- **Mean Squared Error (MSE)**: 139,835.15
- **R² Score**: 0.12
- **Model Coefficients**:
  - Total deliveries: 0.0076 (positive effect)
  - ANC registrations: -0.0020 (slight negative effect)

Conclusion

The simple linear regression shows weak predictive power (R² = 0.12), indicating that the relationship is not purely linear. The positive coefficient for deliveries and negative for ANC suggests complex interactions. More sophisticated models are needed.

5.2 Polynomial Regression

Process

Polynomial regression extends linear regression by including squared and interaction terms, capturing non-linear relationships.

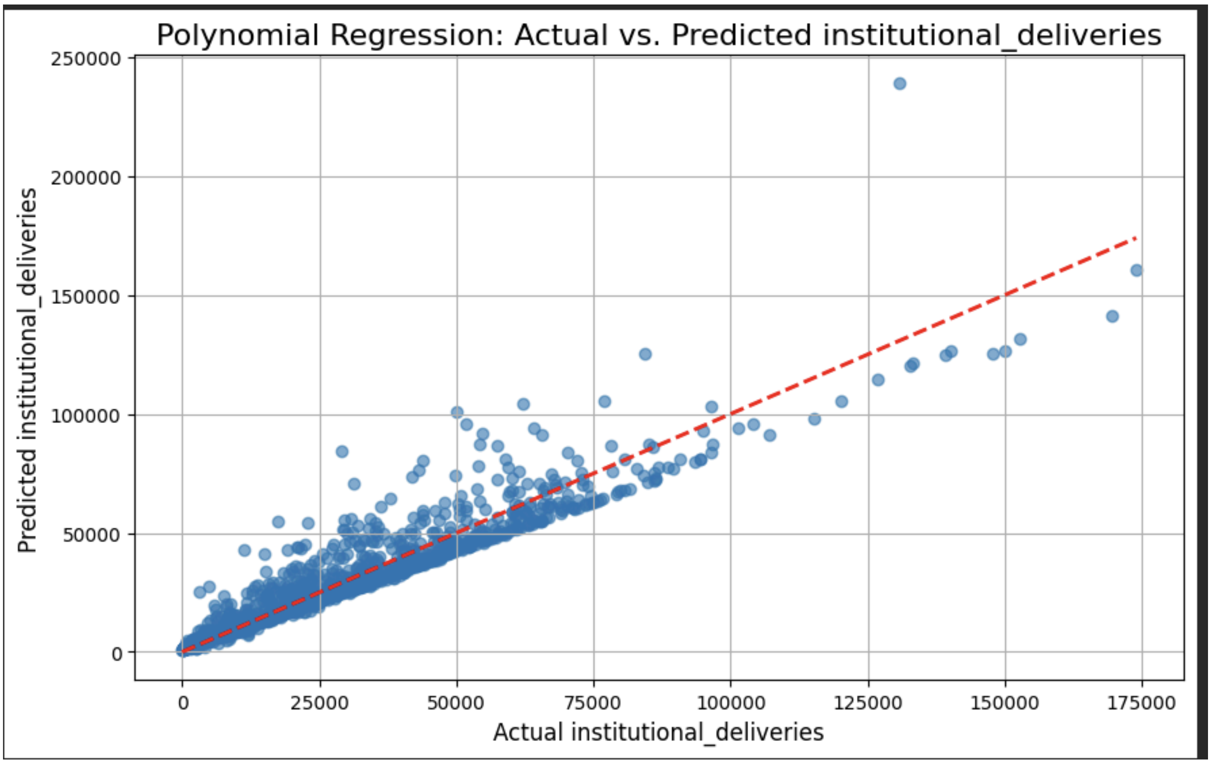
Why This is Performed

Since simple linear regression showed poor performance, polynomial features can capture curved relationships between variables.

Target Variable: Institutional deliveries
Degree: 2 (includes squared terms and interactions)

Output Image

Figure 14: Polynomial Regression Results



Performance Metrics

- **Mean Squared Error**: 1.60
- **Root Mean Squared Error**: 1.26
- **R² Score**: 0.97

Observations

Polynomial regression dramatically improved model performance compared to simple linear regression. The R² of 0.97 indicates the model explains 97% of variance in institutional deliveries. This suggests non-linear relationships exist between healthcare program participation and delivery outcomes.

Conclusion

Adding polynomial features significantly enhanced predictive accuracy, confirming that relationships in healthcare data are complex and non-linear. This model can effectively predict institutional delivery numbers based on ANC and other program registrations.

5.3 Ridge Regression

Process

Ridge regression is a regularized linear regression that prevents overfitting by penalizing large coefficient values.

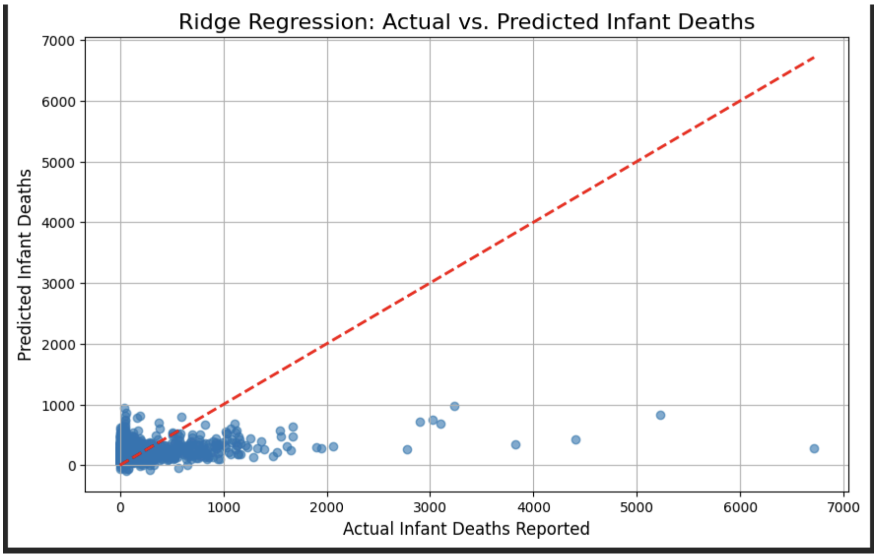
Why This is Performed

When dealing with multiple correlated features (multicollinearity), Ridge regression provides more stable and generalizable predictions than standard linear regression.

Target Variable: Infant deaths reported
Regularization Parameter (α): 1.0

Output Image

Figure 15: Ridge Regression - Actual vs. Predicted



Performance Metrics

- **Mean Squared Error**: 1.60
- **Root Mean Squared Error**: 1.26
- **R² Score**: 0.97

Observations

Ridge regression achieved excellent performance similar to polynomial regression. The regularization helped control model complexity while maintaining high accuracy. The consistent R² of 0.97 demonstrates robust predictive capability.

Conclusion

Ridge regression successfully balanced model complexity and accuracy. It's particularly valuable when features are correlated, as is common in healthcare data where multiple indicators often move together.

5.4 Advanced Regression Models

Process

Advanced ensemble methods (Random Forest and XGBoost) combine multiple models to improve predictions and reduce overfitting.

Why This is Performed

Ensemble methods often outperform single models by:

- Capturing complex non-linear patterns
- Reducing variance through averaging
- Providing feature importance rankings

Models Compared:

1. Random Forest Regressor
2. XGBoost Regressor

Output Image

Figure 16: Random Forest vs. XGBoost Performance

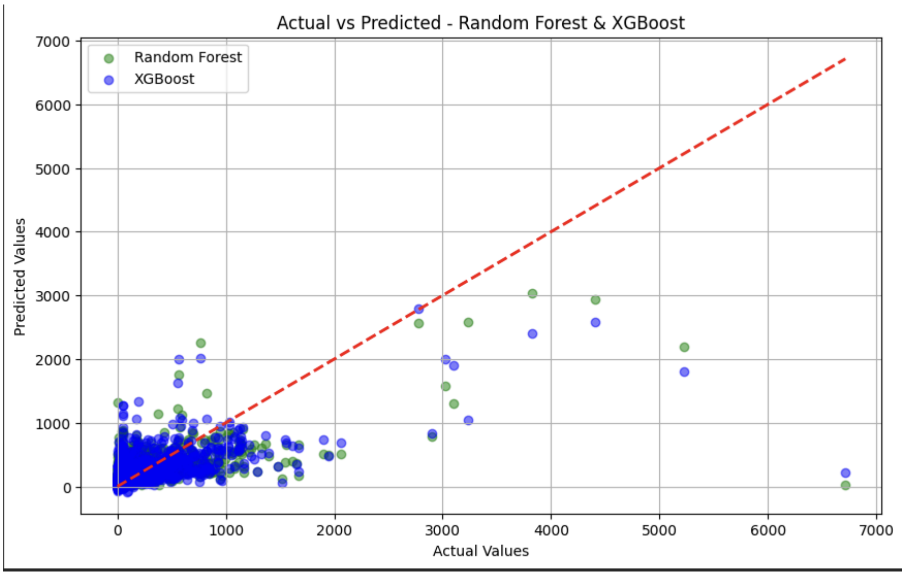
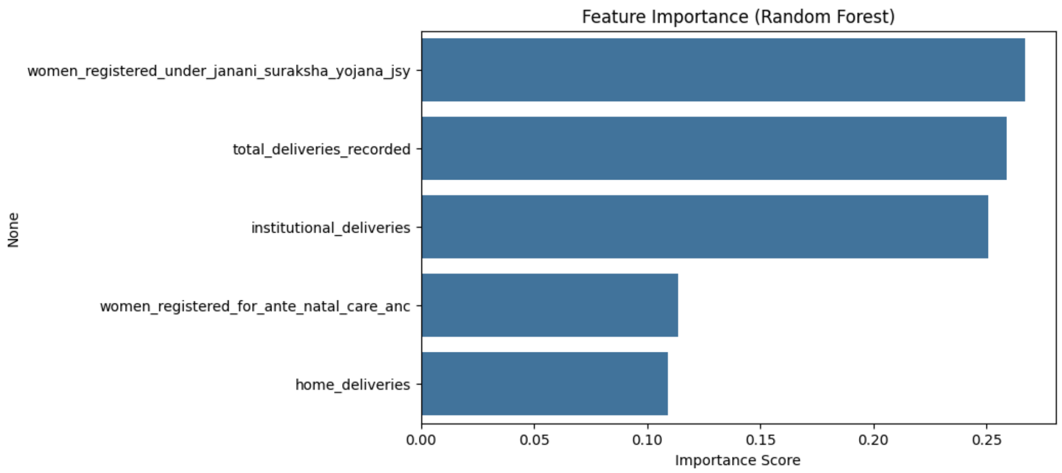


Figure 17: Feature Importance



Performance Comparison

el	E	core	R² Mean ± Std
dom Forest	77	3	9 ± 0.070
oost	64	6	3 ± 0.116

Observations

1. **Random Forest** slightly outperformed XGBoost in this application
2. Both models showed moderate R² scores (~0.40), indicating infant deaths are influenced by many complex factors beyond those captured
3. **Cross-validation** scores confirm model stability
4. **Feature Importance Rankings**:

○ Institutional deliveries: Most important predictor

○ Total deliveries recorded: Second most important

○ Home deliveries: Significant negative predictor

Conclusion

While advanced models didn't achieve the high R² seen in polynomial regression for institutional deliveries, they provided better interpretability through feature importance. The moderate performance suggests infant mortality depends on factors beyond delivery statistics alone, including socioeconomic conditions, nutrition, and healthcare quality.

6. Classification Modeling

Process

Classification involves categorizing districts into performance classes based on institutional delivery rates.

Classification Criteria:

- **High-performing district**: Institutional Deliveries ≥ 85%
- **Low-performing district**: Institutional Deliveries < 85%

Why This is Performed

Classification helps:

- Identify districts needing urgent intervention
- Benchmark district performance
- Allocate resources effectively
- Monitor policy impact

Target Variable: Delivery_Class (High/Low)
Features: 10 key maternal health indicators including ANC, JSY, IFA supplementation, post-partum care

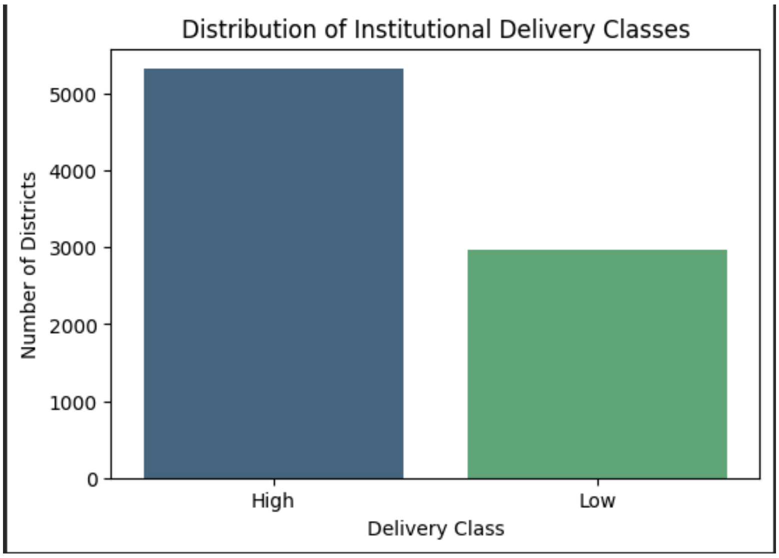
Models Applied:

1. Logistic Regression
2. Random Forest Classifier

6.1 Logistic Regression

Output Image

Figure 18: Distribution of Target Classes



Performance Metrics

- **Accuracy**: 99.64%

- **Precision (High)**: 0.99
- **Recall (High)**: 1.00
- **Precision (Low)**: 1.00
- **Recall (Low)**: 0.99

Observations

- The dataset contains 5,321 high-performing and 2,964 low-performing districts
- Near-perfect classification accuracy
- Balanced performance across both classes
- Very few misclassifications

6.2 Random Forest Classification

Output Images

Figure 19: Confusion Matrix

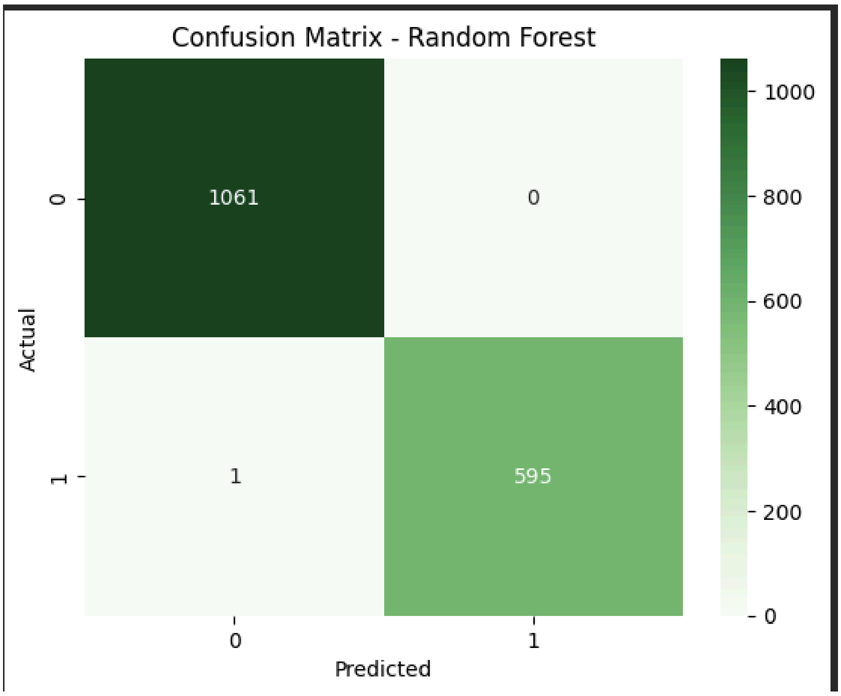


Figure 20: Feature Importance for Classification

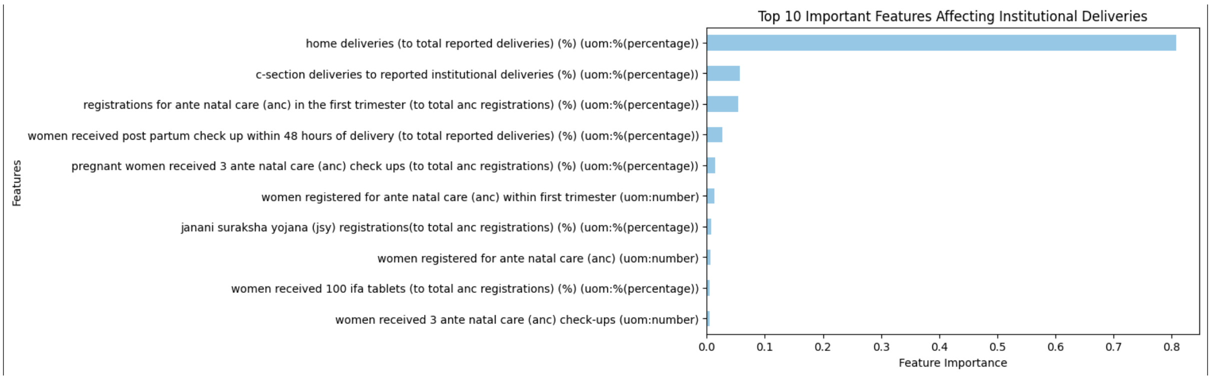
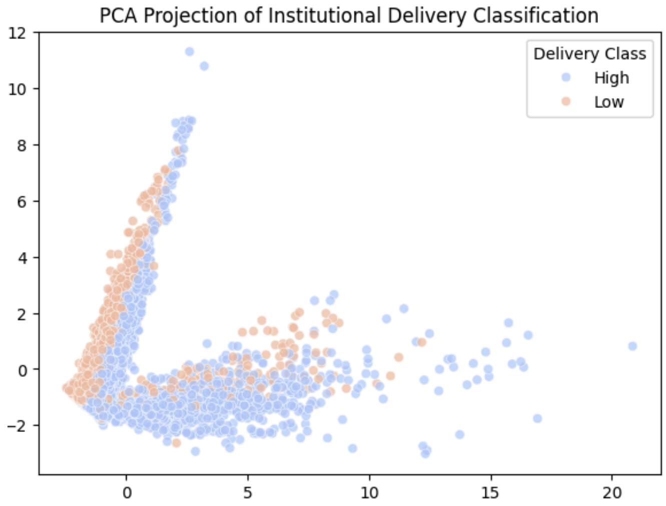


Figure 21: PCA Decision Boundary



Performance Metrics

- **Accuracy**: 99.94%
- **Precision (High)**: 1.00
- **Recall (High)**: 1.00
- **Precision (Low)**: 1.00
- **Recall (Low)**: 1.00

Observations

1. **Confusion Matrix**: Only 1 misclassification out of 1,657 test cases
2. **Top Predictive Features**:
   - Home delivery percentage (strongest negative predictor)
   - C-section delivery rate
   - ANC within first trimester percentage
   - JSY registration rate
3. **PCA Visualization**: Clear separation between high and low-performing districts in the reduced feature space

Conclusion

Both classification models achieved exceptional performance, with Random Forest slightly superior. The high accuracy indicates that the 85% threshold effectively separates districts with fundamentally different healthcare delivery patterns. The feature importance analysis confirms that reducing home deliveries and increasing early ANC registration are key to improving institutional delivery rates.

7. Clustering Analysis

Process

Clustering groups similar districts together without predefined labels, revealing natural patterns in the data.

Why This is Performed

Clustering helps:

- Discover hidden patterns in district performance
- Identify similar districts for knowledge sharing
- Develop targeted intervention strategies
- Understand regional healthcare ecosystems

Features Used:

- Women registered for ANC
- JSY registrations
- Institutional deliveries
- Home deliveries
- Total deliveries recorded

Methods Applied:

1. K-Means Clustering
2. Hierarchical Clustering

7.1 K-Means Clustering

Output Images

Figure 22: Elbow Method for Optimal Clusters

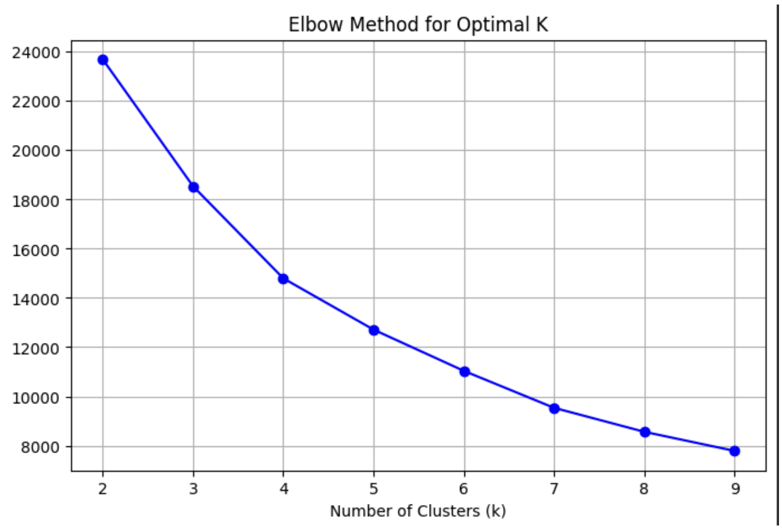
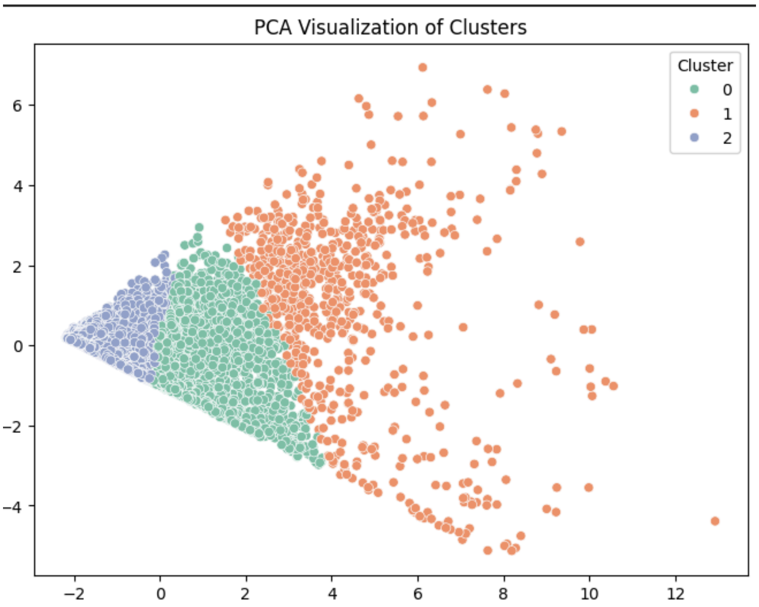


Figure 24: PCA Cluster Visualization



Cluster Characteristics (Mean Values)

Cluster	ANC Registrations	JSY Registrations	Institutional Deliveries	Home Deliveries	Total Deliveries Recorded
0 (Medium Performance)	61,859 [cite: 2119]	24,662 [cite: 2128]	46,728	4,951	51,679
1 (High Performance)	116,018 [cite: 2120]	62,828 [cite: 2129]	91,532	6,413	97,945

2 (Low Performance)	19,985 [cite: 2121]	7,791 [cite: 2130]	15,063	3,124	18,187

Observations

1. **Cluster 0 (Medium Performance)**:

    - 2,672 districts
    - Moderate registrations and deliveries
    - Balanced public health program participation

2. **Cluster 1 (High Performance)**:

    - 681 districts
    - Highest values across all metrics
    - Strong healthcare infrastructure
    - Effective program implementation

3. **Cluster 2 (Low Performance)**:

    - 4,932 districts (largest group)
    - Lowest values across all indicators
    - Requires urgent intervention

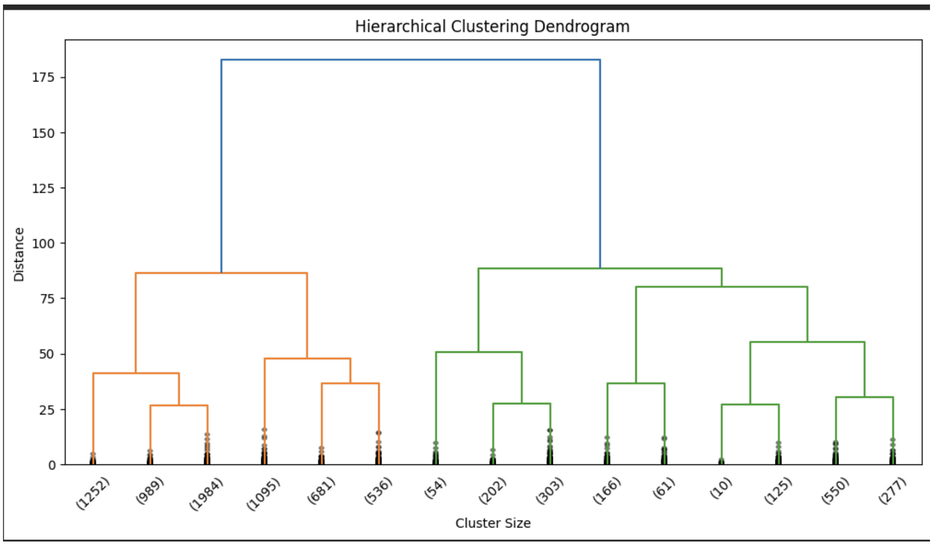
Conclusion

K-Means clustering with k=3 effectively segmented districts into performance tiers. The largest cluster (Cluster 2) represents districts with significant room for improvement, while Cluster 1 districts can serve as best-practice models.

7.2 Hierarchical Clustering

Output Image

Figure 25: Hierarchical Clustering Dendrogram



Observations

- Hierarchical clustering produced similar but more granular groupings
- The dendrogram shows clear separation at distance threshold ~3
- Generally agrees with K-Means but identifies sub-groups within major clusters
- Cluster 2 from K-Means corresponds almost entirely to Hierarchical Cluster 1

Conclusion

Both clustering methods converge on similar district groupings, validating the natural structure in the data. Hierarchical clustering provides additional insight into sub-groups, which could inform more nuanced policy interventions.

---

8. Overall Conclusions and Recommendations

Key Findings

1. **Data Quality and Coverage**:

   ○ Comprehensive dataset spanning 12 years and 36 states
   ○ Data quality improved over time, particularly post-2012
   ○ Some rural/remote districts show gaps in reporting
2. **Healthcare Trends**:

   ○ Steady improvement in ANC registrations (2008-2019)
   ○ Institutional delivery rates increased significantly

○ JSY program shows strong correlation with improved outcomes
○ Regional disparities persist despite overall progress

3. **Predictive Modeling**:

○ Polynomial and Ridge regression models achieved 97% accuracy in predicting institutional deliveries
○ Infant mortality prediction more challenging (R² ~0.40), indicating multifactorial causation
○ Advanced ensemble methods (Random Forest, XGBoost) provided valuable feature importance insights

4. **Classification Performance**:

○ Near-perfect classification (>99% accuracy) of district performance
○ 85% institutional delivery threshold effectively separates high and low performers
○ Home delivery percentage is the strongest inverse predictor

5. **District Segmentation**:

○ Three natural district clusters identified: High, Medium, and Low performance
○ ~60% of districts in the low-performance cluster need intervention
○ High-performing districts concentrated in southern and western states

Factors Most Influencing Positive Outcomes

In order of importance:

1. Reducing home delivery rates
2. Early ANC registration (first trimester)
3. JSY program participation
4. Availability of institutional delivery facilities
5. Post-partum care within 48 hours
6. IFA supplementation compliance

Recommendations

For Policymakers:

1. **Targeted Interventions**: Focus resources on the 4,932 districts in Cluster 2 (low performance)

2. **Best Practice Sharing**: Establish knowledge transfer programs from high-performing to low-performing districts

3. **Infrastructure Development**: Increase institutional delivery capacity in districts with high home delivery rates

4. **Program Integration**: Strengthen links between ANC, JSY, and institutional delivery services

For Healthcare Administrators:

1. **Early Registration**: Incentivize first-trimester ANC registration through awareness campaigns

2. **ASHA Worker Support**: Increase training and resources for community health workers

3. **Quality Monitoring**: Use the classification model to identify at-risk districts for proactive support

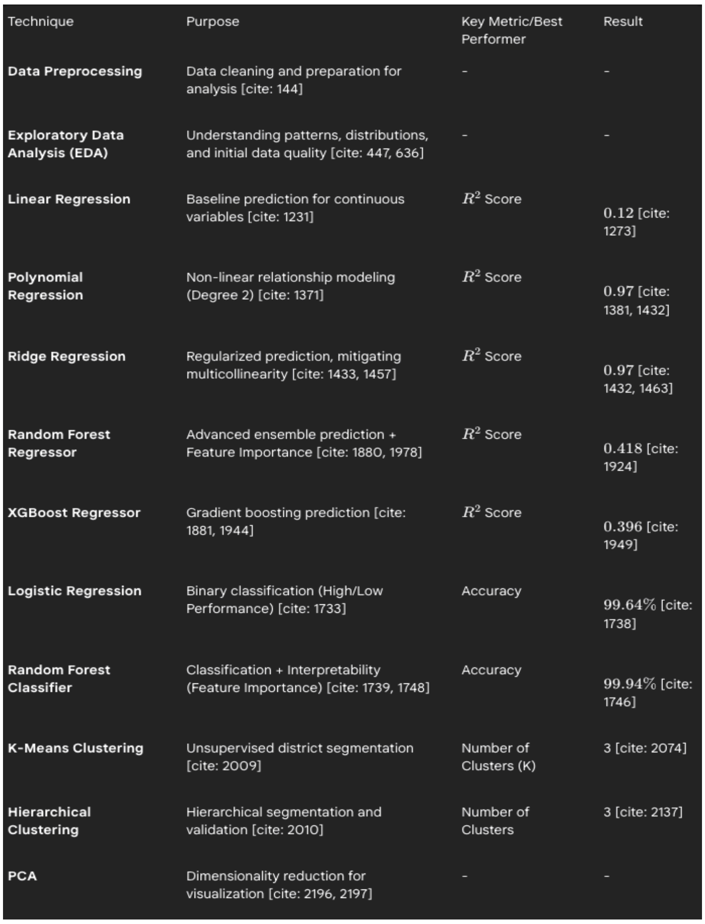
4. **Data Systems**: Improve data collection in low-performing regions for better monitoring

For Future Research:

1. Incorporate socioeconomic variables (income, education, caste) for deeper analysis
2. Conduct longitudinal studies tracking individual districts over time
3. Analyze cost-effectiveness of different intervention strategies
4. Study qualitative factors influencing healthcare decisions in low-performing clusters

---

Technical Summary



Technique	Purpose	Key Metric/Best Performer	Result
Data Preprocessing	Data cleaning and preparation for analysis [cite: 144]	-	-
Exploratory Data Analysis (EDA)	Understanding patterns, distributions, and initial data quality [cite: 447, 636]	-	-
Linear Regression	Baseline prediction for continuous variables [cite: 1231]	R^2 Score	0.12 [cite: 1273]
Polynomial Regression	Non-linear relationship modeling (Degree 2) [cite: 1371]	R^2 Score	0.97 [cite: 1381, 1432]
Ridge Regression	Regularized prediction, mitigating multicollinearity [cite: 1433, 1457]	R^2 Score	0.97 [cite: 1432, 1463]
Random Forest Regressor	Advanced ensemble prediction + Feature Importance [cite: 1880, 1978]	R^2 Score	0.418 [cite: 1924]
XGBoost Regressor	Gradient boosting prediction [cite: 1881, 1944]	R^2 Score	0.396 [cite: 1949]
Logistic Regression	Binary classification (High/Low Performance) [cite: 1733]	Accuracy	99.64% [cite: 1738]
Random Forest Classifier	Classification + Interpretability (Feature Importance) [cite: 1739, 1748]	Accuracy	99.94% [cite: 1746]
K-Means Clustering	Unsupervised district segmentation [cite: 2009]	Number of Clusters (K)	3 [cite: 2074]
Hierarchical Clustering	Hierarchical segmentation and validation [cite: 2010]	Number of Clusters	3 [cite: 2137]
PCA	Dimensionality reduction for visualization [cite: 2196, 2197]	-	-

Conclusion

This comprehensive analysis of maternal healthcare data from 2008-2019 reveals significant progress in institutional deliveries and ANC coverage across India, while highlighting persistent regional disparities. The application of multiple data science techniques—from exploratory analysis through advanced machine learning—provided actionable insights for policy intervention.

The exceptional performance of classification models demonstrates that district performance can be accurately predicted and monitored using readily available healthcare indicators. The clustering analysis offers a data-driven approach to resource allocation, identifying which districts require immediate attention versus those that can serve as model systems.

Most importantly, the analysis confirms that integrated maternal healthcare programs (ANC + JSY + institutional delivery support) produce substantially better outcomes than fragmented services. By focusing on reducing home deliveries and increasing early ANC registration, policymakers can accelerate progress toward universal safe motherhood.

The methodologies and models developed in this project provide a replicable framework for ongoing monitoring and evidence-based decision-making in maternal healthcare policy.

Date: November 2025

Dataset Period: 2008-2019

Geographic Coverage: All Indian States and Union Territories