Symbiosis Institute of Technology, Nagpur

Department of Computer Science and Engineering

Batch: 2022–2026

Course Name: Data Science

Course Code: 0705210707

Semester: VII

1. ABSTRACT

This project examines the process by which workers enroll in MGNREGA, the Mahatma Gandhi National Rural Employment Guarantee Act, due to a break-down of workers by category.

The information is broken down by state and category in the dataset on how many workers are registered, how many job cards they applied to, how many job cards they were issued, and how various social and gender groupings are engaged.

We cleaned the data and subsequently applied various machine-learning methods, including the simple linear regression, multiple linear regression, K-means, DBSCAN, XGBoost, random forest, and logistic regression.

Exploratory analysis revealed that there is a correlation between the application of job cards and the activity of the workers. Prediction was done using the regression models to determine the number of job cards to issue. Clustering was done based on the pattern of participation by the states. The participation of the workers was categorized into low, moderate, and high by the logistic regression.

The findings indicate that the number of applied job cards, the total number of registered workers as well as the participation of women are the key predictors of participation. The ensemble model (Random Forest) had an accuracy of 0.96, which means that patterns of participation are forecastable and similar across the entire country.

---

2. KEYWORDS

Data Science, Machine Learning, MGNREGA, Worker Participation, Regression, Clustering, Random Forest, XG Boost, Logistic Regression

---

3. INTRODUCTION

3.1 Background

The MGNREGA is a scheme that assists individuals in rural India to get employment and livelihood better. It provides assured work to the rural families in the form of community projects. Due to this, much information has been gathered regarding job cards, workers and the number of individuals who enrolled. This data can be used with machine learning to reveal some latent trends, demonstrate the effectiveness of the program, and make improved policy decisions.

3.2 Problem Statement

It is aimed at applying machine learning to the research and find the tendencies of the worker participation in MGNREGA and discover which social and economic factors impact the issuance of job cards and the activity of the working force

3.3 Objectives

1. Clean data for analysis.
2. To apply Exploratory Data Analysis (EDA) for initial insights.
3. To apply regression models for predicting issued jobcards
4. To apply clustering algorithms for grouping states.
5. To classify participation levels using Logistic Regression.
6. To visualize trends through graphs.

---

4. LITERATURE REVIEW

Author / Year	Title / Approach	Methods Used	Findings
Rumela Ghosh (2013)	*"A Bird's Eye View into Mahatma Gandhi National Rural Employment Guarantee Act*	Policy analysis, descriptive statistics	Reviewed the implementation and efficacy of MGNREGA across India.
B. Jayakumar & S. Prabakar (2024)	*"Leveraging Digital Innovation to Enhance MGNREGA's Impact on Rural Empowerment"*	Digital governance, GIS, secondary data analysis	Explored how digital interventions improved inclusion and scheme transparency.
Utkarsh Arora & Gaurav Arora (2023)	*"Analysis of the Impacts of MGNREGA using Spatial Data Analysis"*	Spatial analysis, machine learning, econometrics	Investigated spatial factors affecting MGNREGA implementation at state and district levels.

---

5. METHODOLOGY

5.1 Dataset Description

The dataset includes:

5.1 Dataset Description

5.1.1 Dataset Dimensions

- **Shape:** 2,676,899, 18 rows
- **Time Coverage:** Multiple years starting from 2014-2023.
- **Geographic Coverage:** Multiple states, districts, blocks, and gram panchayats across India.

	id	year	state_name	district_name	block_name	gp_name	applied_jobcards	issued_jobcards	registered_workers_scs	registered_workers_sts	registered_workers_others
0	0.0	2014-2015	Andaman And Nicobar	Nicobars	Nancowry	Chowra Tc	309.0	309.0	0.0	646.0	2.0
1	1.0	2014-2015	Andaman And Nicobar	Nicobars	Nancowry	Kamorta Tc	612.0	612.0	0.0	1255.0	191.0
2	2.0	2014-2015	Andaman And Nicobar	Nicobars	Nancowry	Katchal Tc	452.0	452.0	0.0	435.0	307.0
3	3.0	2014-2015	Andaman And Nicobar	Nicobars	Nancowry	Nancowry Tc	236.0	236.0	0.0	590.0	1.0
4	4.0	2014-2015	Andaman And Nicobar	Nicobars	Nancowry	Teressa Tc	332.0	332.0	0.0	580.0	1.0

fig.1 Overview of Dataset

5.1.2 Variable Categories

Geographical Variables:

- year: Time period of data collection.
- state_name: State-level administrative division.
- district_name: District-level administrative division.
- block_name: Block-level administrative division.
- gp_name: Gram Panchayat name.

Job Card Variables:

- applied_jobcards: Number of job cards applied for.
- issued_jobcards: Number of job cards actually issued.
- active_jobcards: Number of currently active job cards.

Worker Registration by Category:

- registered_workers_scs: Scheduled Caste workers.
- registered_workers_sts: Scheduled Tribe workers.
- registered_workers_others: Other category workers.
- total_registered_workers: Total registered workers.
- registered_workers_women: Female workers registered.

Active Worker Statistics:

- active_workers_scs: Active Scheduled Caste workers.
- active_workers_sts: Active Scheduled Tribe workers.
- active_workers_others: Active workers from other categories.
- total_active_workers: Total active workers.

active_workers_women: Active female workers.

5.2 Preprocessing Steps

1. Removed null and duplicate entries
2. Dropped irrelevant columns like id, gp_name, etc.
3. Created derived column:
   female_worker_proportion = registered_workers_women / total_registered_workers
4. Encoded categorical variables (e.g., state_name) using LabelEncoder
5. Standardized numerical features for consistent scaling

```
Missing Values:
 year                          0
state_name                     0
district_name                  0
block_name                     0
gp_name                       16
applied_jobcards               0
issued_jobcards                0
registered_workers_scs         0
registered_workers_sts         0
registered_workers_others      0
total_registered_workers       0
registered_workers_women       0
active_jobcards                0
active_workers_scs             0
active_workers_sts             0
active_workers_others          0
total_active_workers           0
active_workers_women           0
dtype: int64
```

fig.2 Missing Values Analysis

5.3 Model Workflow

To analyze and predict participation by workers we applied supervised and unsupervised models.

1. Simple Linear Regression: A simple regression model, which is applied to estimate the number of active workers depending on jobcards applied. It assists in determining a straight proportionality between two variables.

2. Multiple Linear Regression: It is an overextension of linear regression with several predictors which may include applied jobcards, total registered workers, and state-level data to enhance the accuracy of prediction.
3. K-Means Clustering: This is an uncontrolled algorithm employed to classify states into three participation groups, i.e., low, moderate, and high groups, according to the worker-related features.

4. DBSCAN (Density-Based Spatial Clustering): Dense clusters and outliers are identified which will highlight areas where the participation behaviour of the workers is abnormal than that of the remainder of the dataset.

5. XGBoost Regression: A gradient boosting model that is known to be a high performance and accuracy model. It estimates the quantity of issued jobcards and prioritizes the significance of the input features.

6. Random Forest Regression This is an ensemble model where a number of decision trees are used to enhance the reliability of prediction. It had the best accuracy to verify that there are high levels of relationships among variables.

7. Logistic Regression: It is a classification algorithm, which was used to categorize areas according to the worker and jobcard characteristics into low, moderate, and high participation areas.

---

6. IMPLEMENTATION

6.1 Environment

- Language used: Python
- Platform used: VS Code
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost etc.

6.2 Steps Performed

1. Loaded and cleaned the dataset
2. Conducted Exploratory Data Analysis (EDA)
3. Applied regression, clustering, and classification models
4. Evaluated model performance using R^2, Accuracy, and RMSE
5. Created visualizations for better interpretation

---

7. RESULTS AND DISCUSSION

7.1 Exploratory Data Analysis

- Explored gender dynamics more deeply, a new column was engineered to capture the proportion of female workers relative to total registered workers. This showed that Female workers make up 40–50% of total workforce in most states.

```
df['female_worker_proportion'] = df['registered_workers_women'] / df['total_registered_workers']
df.head()
```

✓ 0.1s

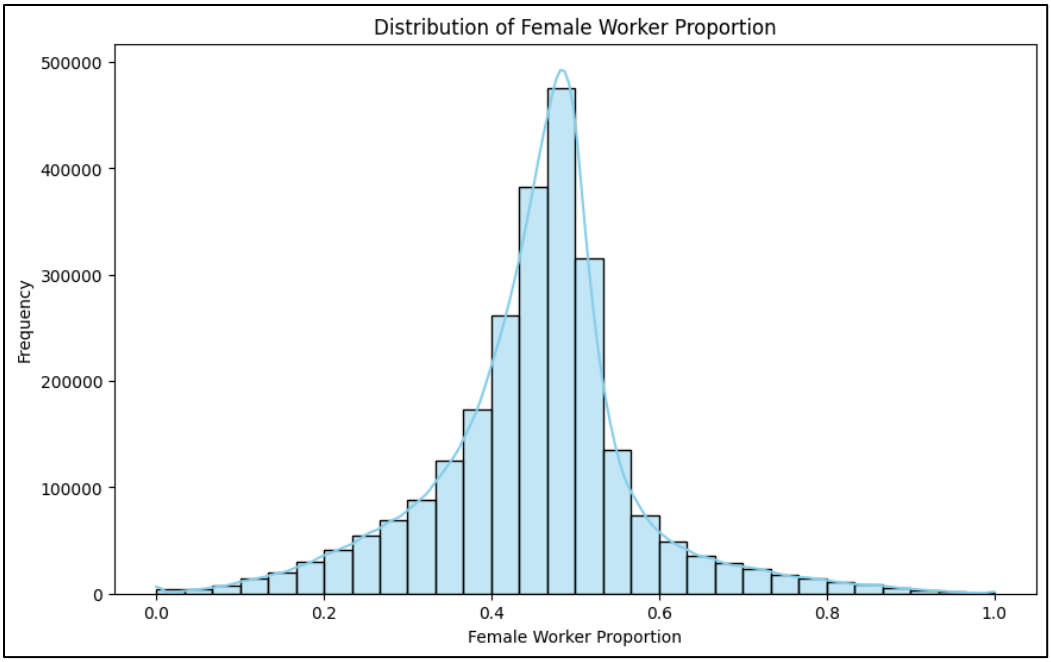

fig.3 Distribution of Female Worker Proportion

- State-wise averages of registered and active workers were visualized using grouped bar charts. States like West Bengal, Kerala, and Assam showed high participation levels. However, some states displayed a notable gap between total registration and active worker count, possibly indicating inefficiencies in job allocation, seasonal inactivity, or worker attrition.

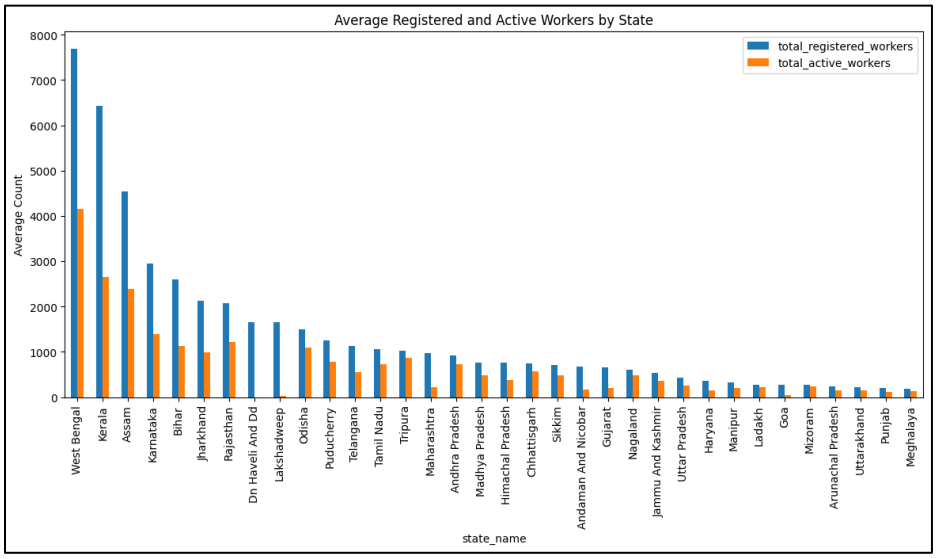

fig.4 State-wise Averages of Registered and Active Workers

7.2 Model Description

To analyze and predict worker participation we used **supervised** and **unsupervised** models.

- **Simple Linear Regression:**
  A basic regression model used to predict the number of *active workers* based on *applied jobcards*. It helps identify direct proportionality between two variables.

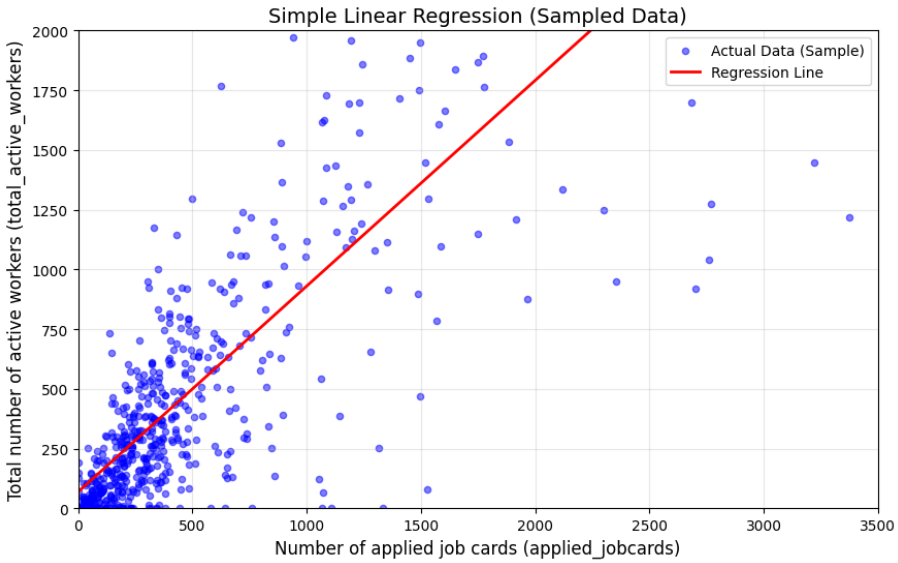

fig.5 Simple linear regression graph

An R^2 value of 0.64 indicates that around 64% of the variation in active worker count can be explained solely by the number of applied job cards, which suggests a moderately strong positive relationship.

However, the RMSE of 467 shows that the model's predictions can deviate by ±467 workers on average, which is relatively high compared to the range of values. This, combined with the wide dispersion of data points in the scatter plot, implies that while applied jobcards is an important factor, it cannot reliably predict active workers on its own.

Therefore, the simple linear regression indicates a weak but positive relationship between the number of applied job cards and the number of active workers. While the regression line shows an upward trend — implying that higher job card applications are generally associated with higher worker participation — the wide dispersion of data points suggests that the prediction power of this single variable is limited. This means that applied_jobcards alone is not a strong predictor of active workers, and other factors (such as state policies, seasonal demand, or worker category) likely play a significant role.

- **Multiple Linear Regression:**
  An extension of linear regression using multiple predictors such as *applied jobcards*, *total registered workers*, and *state-level data* to improve prediction accuracy.

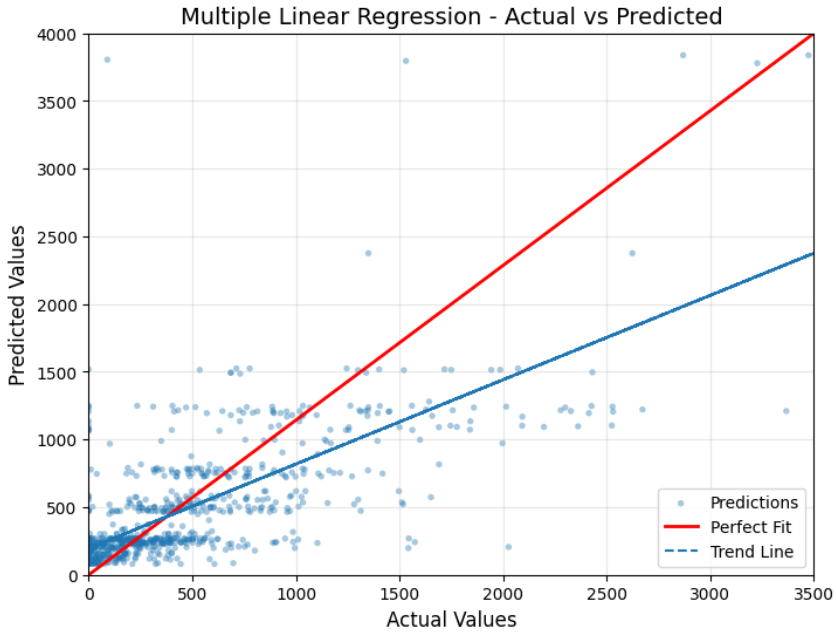

fig.6 Multiple Linear Regression graph

The model performs reasonably well for lower values of total_active_workers — predictions are clustered near the diagonal line when actual values are small. As actual values increase, the model consistently underpredicts, shown by most data points falling below the red perfect-fit line. The trend line is significantly flatter than the perfect-fit line, confirming that the model is not able to capture high-variance or high-value observations effectively.

Input Features Used: ['year_2015-2016', 'year_2016-2017', 'year_2017-2018', 'year_2018-2019', 'year_2019-2020',

	Feature	Coefficient	Absolute_Impact
41	state_name_West Bengal	3684.444873	3684.444873
11	state_name_Assam	2299.606506	2299.606506
22	state_name_Kerala	2259.615479	2259.615479
21	state_name_Karnataka	1401.895546	1401.895546
34	state_name_Rajasthan	1126.274344	1126.274344
12	state_name_Bihar	1084.939422	1084.939422
31	state_name_Odisha	980.724926	980.724926
20	state_name_Jharkhand	876.927736	876.927736
38	state_name_Tripura	781.517683	781.517683
9	state_name_Andhra Pradesh	662.129747	662.129747
36	state_name_Tamil Nadu	636.023817	636.023817
35	state_name_Sikkim	466.158250	466.158250
13	state_name_Chhattisgarh	464.220573	464.220573
37	state_name_Telangana	409.591915	409.591915
25	state_name_Madhya Pradesh	381.365348	381.365348

fig.7 Impact of features on multiple linear regression

The multiple linear regression model indicates that geographical factors, particularly state, are the strongest determinants of active worker participation. West Bengal contributes the highest positive impact, increasing predicted worker count by approximately 3,684 workers relative to the baseline state. Assam, Kerala, and Karnataka also show strong positive influence. This suggests that policy structure, labor demand, or demographic factors in specific states are key drivers of employment activity, more so than temporal (year-based) variations.

- **K-Means Clustering:**
  An unsupervised algorithm used to group states into three participation clusters — *low*, *moderate*, and *high* — based on worker-related attributes.

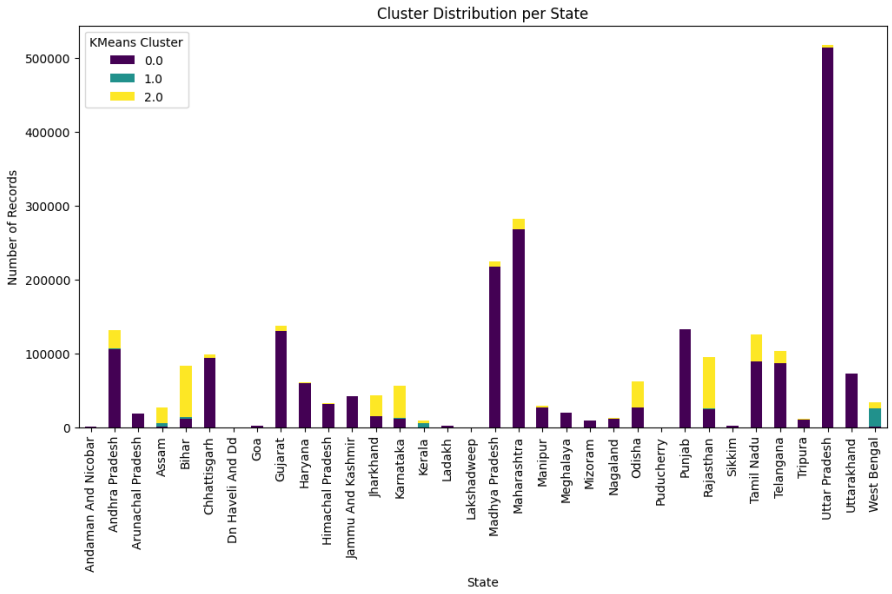

fig.8 K-means clustering graph

- **DBSCAN (Density-Based Spatial Clustering):**
  Identifies dense regions and outliers, revealing areas with unusual worker participation patterns compared to the rest of the dataset.

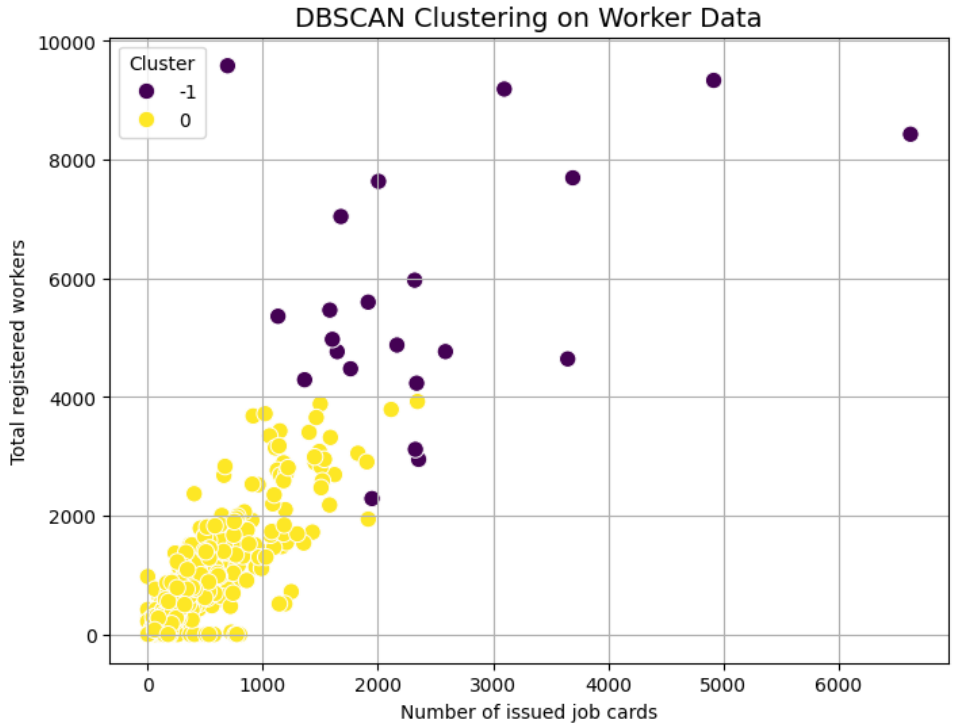

fig.9 DBSCAN Clustering on worker data

```
Number of clusters found: 1
cluster
 0    478
-1     22
Name: count, dtype: int64
```

DBSCAN identified a single dense group (Cluster 0). The majority of records (478) belong to one large, homogeneous group. This implies that most districts/blocks in dataset have similar patterns in worker-related metrics such as:

- number of issued job cards,
- number of registered and active workers, and
- proportion of female workers.

This group represents the typical or average behavior across most areas.

Here, 69 points marked as outliers (Noise = -1). DBSCAN considered 69 records as outliers because they lie far from the dense cluster. These points represent districts or gram panchayats with very different characteristics

- **XGBoost Regression:**
  A gradient boosting model known for high performance and accuracy. It predicts the number of *issued jobcards* while ranking the importance of input features.

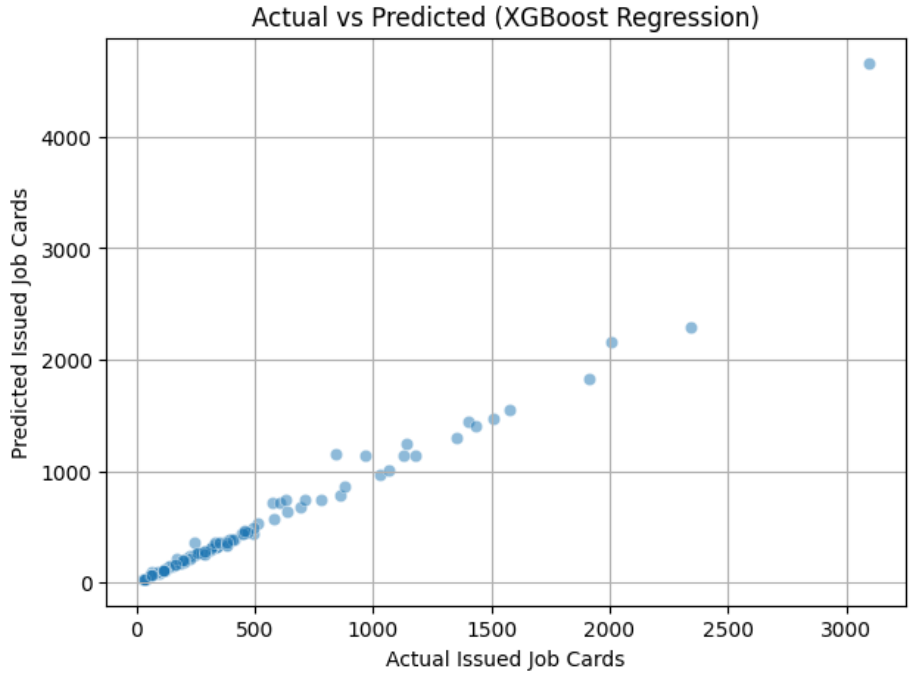

fig.10 XGBoost Regression graph

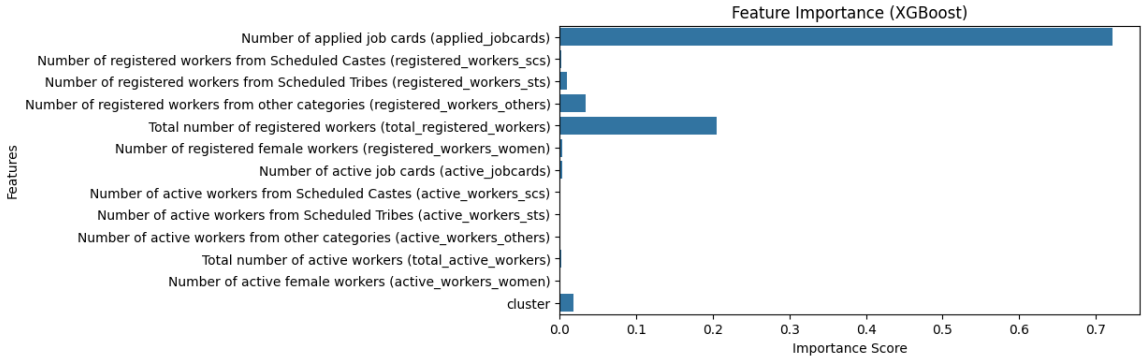

fig.11 XGBoost feature importance graph

There's a strong and predictable relationship between the number of applied job cards and issued job cards. The dataset is clean and consistent enough for XGBoost to learn effectively. It shows that demographic and category-based worker counts (like SC/ST/Others) influence the issuance slightly but are secondary factors.

DBSCAN earlier found only one cluster (most data is dense and similar), and this XGBoost result supports that — the dataset follows a clear trend rather than having diverse group behaviors.

- **Random Forest Regression:**
  An ensemble model combining multiple decision trees to improve prediction reliability. It achieved the highest accuracy, confirming strong relationships among variables.

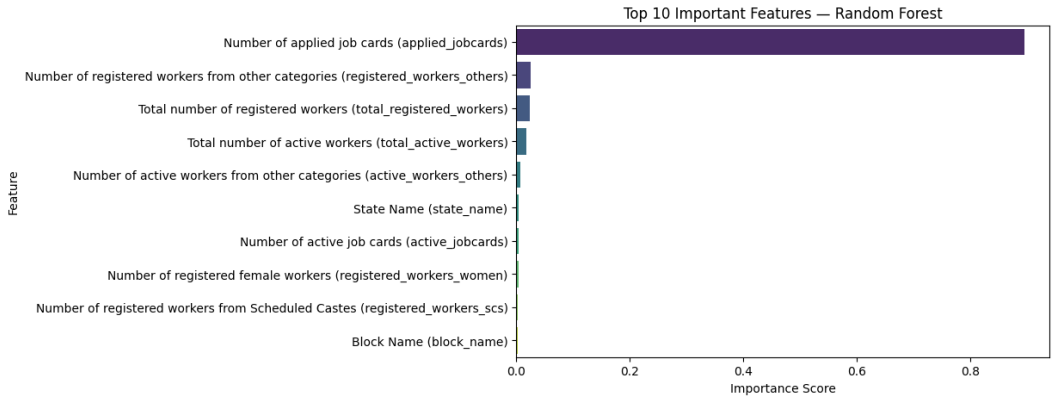

fig.12 Random Forest feature importance graph

The Random Forest model achieved 96.4% accuracy, confirming that job card issuance can be reliably predicted using applied job cards and worker registration metrics.

The dataset shows high consistency across states and low noise, aligning with DBSCAN result (dense, homogeneous data) and indicating a stable employment pattern under MGNREGA, i.e the MGNREGA process seems consistent across states.

- **Logistic Regression:**
  A classification algorithm applied to categorize regions into *low*, *moderate,* and *high* participation levels based on worker and jobcard features.

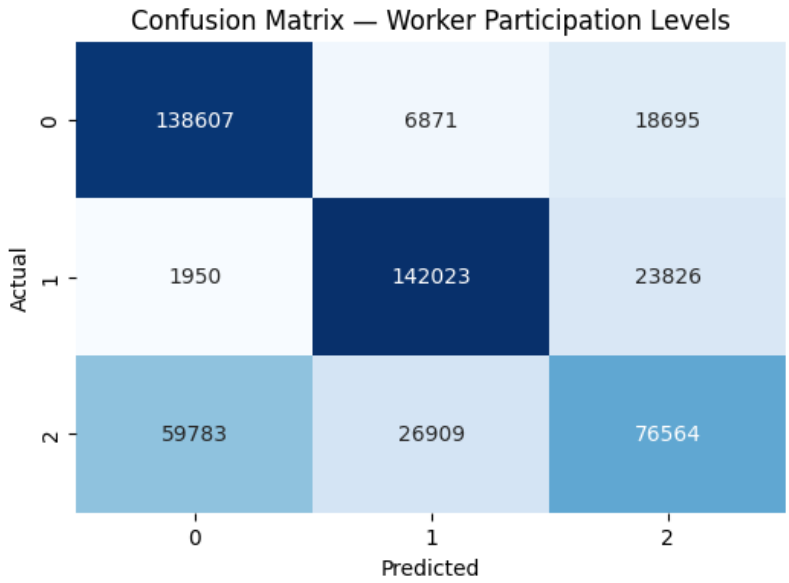

fig.13 Confusion Matrix for worker participation levels

The Logistic Regression model achieved 72.13% accuracy in classifying regions into Low, Moderate, and High participation categories.

It performed best in identifying high participation states and moderately well in detecting low participation regions.

Misclassifications mostly occurred for moderate participation zones, indicating overlapping job card and worker characteristics between categories.

Overall, the model confirms that participation trends are predictable, and jobcard and registration variables play a dominant role in determining regional activity levels.

7.3 Model-Wise Results

Model	Goal	Performance	Insights
Simple Linear Regression	Predict total active workers from applied jobcards	$R^2 = 0.64$	Moderate linear relation, but other factors affect outcomes
Multiple Linear	Predict active workers using	$R^2 = 0.87$	State-level variation has strong influence

Regression	multiple factors		
KMeans Clustering	Cluster states by worker participation	3 Clusters	Cluster 0: Low, Cluster 1: High, Cluster 2: Moderate participation
DBSCAN	Identify dense clusters and outliers	1 Dense Cluster + 69 Outliers	Dataset mostly uniform, few abnormal regions
XGBoost Regression	Predict issued jobcards	$R^2 = 0.90$	Strong predictive accuracy, important features: applied jobcards, registered workers
Random Forest Regression	Predict issued jobcards	$R^2 = 0.96$	Highest accuracy, stable and consistent results
Logistic Regression	Classify participation levels	Accuracy = 72%	Accurately detects high and low participation; moderate overlaps

---

7.3 Feature Importance

1. Applied Jobcards — strongest predictor (~90%)
2. Total Registered Workers — strong influence
3. SC/ST Worker Ratio — moderate impact
4. Female Worker Proportion — mild influence

---

8. CONCLUSION

We successfully analysed MGNREGA worker participation using multiple machine learning models.

The regression model achieved strong prediction accuracy whereas clustering revealed distinct participation of groups across the different states.

Key Takeaways:

- Jobcard applications and total registered workers are the strongest predictors of active participation.

- Random Forest and XGBoost provided excellent predictive performance ($R^2 \geq 0.90$).
- Logistic Regression effectively classified participation into meaningful categories.
- Data patterns across states are largely consistent, validating MGNREGA's implementation efficiency.

---

9. FUTURE WORK

1. Incorporate spatial visualization using Folium and GeoPandas for interactive maps.
2. Apply time-series forecasting models to predict future participation trends.
3. Integrate demographic and economic indicators for deeper analytical insights.

---

10. REFERENCES

1. Ghosh, R. (2013). *A Bird's Eye View into Mahatma Gandhi National Rural Employment Guarantee Act.* World Bank Discussion Paper.
2. Jayakumar, B., & Prabakar, S. (2024). *Leveraging Digital Innovation to Enhance MGNREGA's Impact on Rural Empowerment. International Journal of Computational Engineering Science and Emerging Networks (IJCESEN).*
3. Arora, U., & Arora, G. (2023). *Analysis of the Impacts of MGNREGA using Spatial Data Analysis.* IIIT Delhi Repository. https://repository.iiitd.edu.in
4. Jiawei Han & Micheline Kamber, *Data Mining: Concepts and Techniques*
