Industrial Human Resource Geo-Visualization

- Technologies Used: EDA, Visualization, NLP
- Domain: Resource Management
- Your name, date, etc.

Introduction

- **Workforce classification in India is vital for policy making.**

- **Traditionally classified into main & marginal workers (excluding agriculture).**

- **Existing datasets are outdated.**

- **This study updates the classification using EDA, Clustering, and Machine Learning.**

Problem Statement

- Current industrial classification of workforce is not updated.
- Difficult to capture the real employment trends.
- Need for accurate, up-to-date classification by sector, division, and class.
- Aim: Build NLP + ML pipeline to improve workforce classification.

Dataset & Preprocessing

- Source: Census/industrial workforce CSV files.
- Preprocessing:
- Removed backticks and converted codes to integers.
- Encoded categorical variables (States, NIC Name).
- Handled missing values.
- Feature engineering (ratios: rural/urban, marginal/total).

Exploratory Data Analysis (EDA)

- **Distribution of industries across India.**
- **Correlation Heatmap of features.**
- **Outlier detection using boxplots.**
- **Word clouds / frequency counts for industry names.**

NLP & Clustering

- **TF-IDF Vectorization on industry descriptions (NIC Name).**
- **KMeans Clustering (k=6) to group industries.**
- **Top terms per cluster revealed clear sector patterns.**
- **Exported cluster-labeled data for visualization.**

Classification Models

- **Models Trained:**
- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **KNN**
- **SVM**
- **Naive Bayes**
- **AdaBoost, ExtraTrees, XGBoost**
- **Best Accuracy: Random Forest / XGBoost performed strongest.**

Regression Models

- **Target: Main Workers – Total Persons**
- **Models Tested:**
- **Linear, Ridge, Lasso, ElasticNet**
- **Decision Tree, Random Forest, Gradient Boosting, XGBoost**
- **KNN Regressor**
- **Best Fit: ElasticNet (saved as final model).**

Results & Insights

- **Clustering: Segmented industries into 6 logical groups.**
- **Classification: Achieved strong accuracy in predicting sector labels.**
- **Regression: ElasticNet provided balanced performance on workforce prediction.**
- **Visualization: Correlation heatmaps and cluster terms provided insights into sectoral distribution.**

Conclusion

- **Successfully built an Industrial HR Geo-Visualization pipeline.**
- **Combined EDA, NLP clustering, Classification, and Regression.**
- **Provided updated workforce classification useful for:**
- **Policy-making**
- **Employment planning**
- **Industrial analysis.**
- **Future Scope:**
- **Deploy as Streamlit Dashboard for interactive geo-visualization.**
- **Extend to state-wise workforce mapping.**

Thank You