Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

# AI-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies (RoC)

**Ramya Venkatakrishnan[1], Mohanavarshini Sivagurunathan[2], Lalitha Siva[3], Subiksha Subramani[4], Arjun Paramarthalingam[5]**

Department of Computer Science and Engineering, University College of Engineering, Villupuram, TamilNadu, India-605103

**E-mail:** [1*]ramyavbe@gmail.com, [2]mohanasivagurunathan@gmail.com, [3]lalithasiva0301@gmail.com, [4]subi31052004@gmail.com, [5]arjun_ucev@ymail.com

## Abstract

This study presents a comprehensive analysis of company registration trends in India, focusing on data sourced from the Registrar of Companies (RoC). The study investigates the temporal patterns of company registrations, emphasizing principal business activities and classifications into public and private sectors. Leveraging advanced data pre-processing techniques, The Study explores the spatial and temporal dynamics of company registrations. Furthermore, a suite of machine learning algorithms, including Linear regression, Decision tree, Random Forest, GBM, SVM, KNN, Naive Bayes, and Weka, is employed to predict future registration trends. Visualization of insights is facilitated through the use of Tableau. The findings provide valuable insights for stakeholders in business, policy, and investment sectors, aiding informed decision-making and strategic planning.

**Keywords:** Company Registration Trends, Registrar of Companies (Roc), Exploratory Data Analysis, Predictive Analytics, Machine Learning Algorithms, Weka, Tableau Visualization.


*Extended version of article received from 4th INTERNATIONAL CONFERENCE on Artificial Intelligence,5G Communication and Network Technologies*
*DOI: https://doi.org/10.36548/jucct.2024.1.006*
*Received: 04.03.2024, received in revised form: 01.04.2024, accepted: 13.04.2024, published: 02.05.2024*

## 1. Introduction

In the dynamic landscape of India's business ecosystem, understanding the trends and patterns of company registrations is paramount for informed decision-making and strategic planning. This study delves into the exploration and prediction of company registration trends within the region, utilizing data sourced from the Registrar of Companies (RoC). The focus extends beyond mere enumeration, delving into the temporal dynamics to identify peak registration years and the geographical concentration of registered companies, with a spotlight on the RoC as the epicenter of registration activities. Moreover, the study delves deeper into the principal business activities undertaken by these registered companies, recognizing the diverse economic sectors driving India's entrepreneurial landscape. Additionally, the analysis distinguishes between public and private sector entities, shedding light on the distribution and dynamics within each sector [1-5].

Employing a comprehensive methodology encompassing data loading, pre-processing, exploration, and predictive modelling, this study harnesses a spectrum of analytical tools, including advanced machine learning algorithms such as linear regression, decision trees, random forests, gradient boosting machines (GBM), support vector machines (SVM), k-nearest neighbors (KNN), and Naive Bayes. Furthermore, we leverage the capabilities of the Tableau visualization platform and the Weka toolset to facilitate intuitive exploration and robust prediction of company registration trends. By unveiling the nuanced insights derived from this interdisciplinary approach, the study aims to empower stakeholders with actionable intelligence, enabling proactive strategies to navigate India's dynamic business landscape [6-10].

## 2. Problem Description and Motivation

In India, the dynamics of business registration play a pivotal role in understanding economic trends, regional development, and sectoral growth. The Registrar of Companies (RoC) serves as the primary repository of data regarding company registrations, offering a wealth of information for analysis and interpretation. However, navigating through this vast dataset to extract meaningful insights presents a formidable challenge. The research endeavors to address this challenge by employing advanced data exploration and predictive

analytics techniques to uncover patterns, trends, and forecasts related to company registrations in India.

### 2.1 Complexity of Company Registration Data

The sheer volume and complexity of company registration data in India pose significant hurdles to traditional analytical approaches. With millions of companies registered across various regions and sectors, manual analysis becomes impractical and inefficient. Moreover, the heterogeneous nature of company data, including diverse business activities and ownership structures, further complicates the analysis process. As such, there is a pressing need for sophisticated analytical methodologies capable of handling the intricacies inherent in company registration datasets.

### 2.2 Motivation for Exploration and Prediction

Understanding the temporal and spatial trends in company registrations is of paramount importance for policymakers, investors, and business stakeholders. Identifying regions experiencing a surge in entrepreneurial activity, discerning emerging sectors poised for growth, and predicting future trends in company registrations can inform strategic decision-making and resource allocation. By leveraging AI-driven exploration and prediction techniques, we aim to uncover actionable insights from RoC data, enabling stakeholders to make informed decisions and capitalize on emerging opportunities.

### 2.3 Focus Areas

Principal Business Activity and Sector Classification Central to our analysis is the examination of principal business activities as per company records. By categorizing companies based on their primary areas of operation, we seek to identify sector-specific trends and patterns in company registrations. Additionally, distinguishing between public and private sector companies allows for a nuanced understanding of the dynamics shaping different segments of the Indian economy. Through this granular analysis, we aim to provide stakeholders with valuable insights into sectoral dynamics and competitive landscapes.

### 2.4 Leveraging Advanced Analytical Techniques

To extract actionable insights from the complex RoC dataset, we employ a combination of data preprocessing, exploratory data analysis (EDA), and predictive modeling techniques. By harnessing the power of machine learning algorithms such as linear regression, decision trees, random forests, gradient boosting machines (GBM), support vector machines (SVM), k-nearest neighbors (KNN), and naive Bayes, we aim to develop robust predictive models capable of forecasting future company registration trends. Additionally, we utilize visualization tools such as Tableau and machine learning software like Weka to facilitate the exploration and interpretation of results.

## 3. Methodology

The methodology involves data collection from data.gov.in, pre-processing to handle missing values and outliers, and exploration through descriptive statistics and visualization. Predictive modeling employs machine learning algorithms like linear regression, decision trees, and ensemble methods, evaluated for accuracy and performance using techniques such as cross-validation. Visualization tools like Tableau aid in presenting insights, while Weka facilitates model building and comparison for predicting company registration trends. The Figure.1 shown below depicts the system architecture.

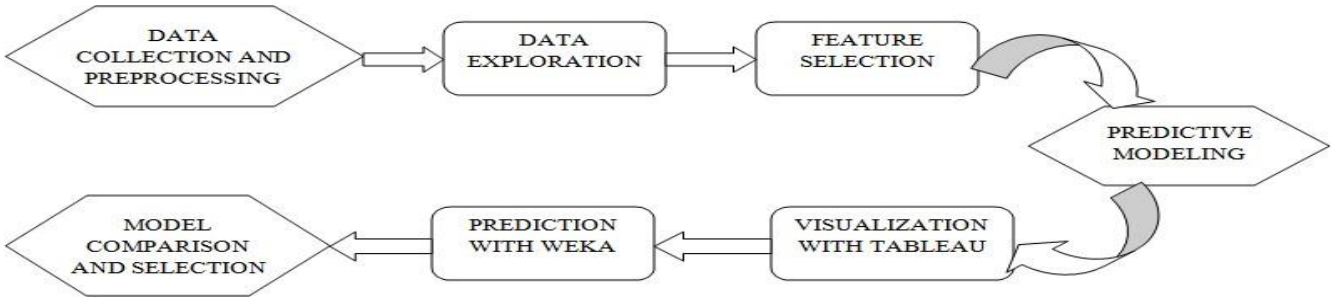


**Figure 1.** System Architecture.

### 3.1 Data Collection and Pre-Processing

**A. Data Source:** The dataset from data.gov.in comprises over 10,000 instances with 17 attributes as shown in the Figure 2. It includes information on company registrations in India, covering variables such as registration date, principal business activity, company type (public/private), and geographic location. The dataset offers insights into trends in company registrations over time and across different regions and sectors. Analysis of this dataset enables exploration and prediction of company registration patterns, crucial for understanding economic activity and business dynamics in India.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CORPORATE_IDENTIFICATION_NUMBE | COMPANY_NAME | COMPANY_STATUS | COMPANY_CLASS | COMPANY_CATEGORY | COMPANY_SUB_CATEGOR | DATE_OF_REGISTRATIO | REGISTERED_STAT | AUTHORIZED_CAI | PAIDUP_CAPITAL | INDUSTRIAL_CLASS | PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CI | REGISTERED_OFFICE | REGISTRAR_OF_COMPANIE | EMAIL_ADDR | LATEST_Y | LAT |
| 2 | F00643 | HOCHTIEFF AG, | NAEF | NA | NA | NA | 01-12-1961 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | AMBLE SIDE, NO.8(O | ROC DELHI | NA | NA | NA |
| 3 | F00721 | SUMITOMO CORPI | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | FLAT NO. 6, 1st FLOO | ROC DELHI | shuchi.chug@ | NA | NA |
| 4 | F00892 | SRILANKAN AIRLI | ACTV | NA | NA | NA | 01-03-1982 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | SRILANKAN AIRLINES | ROC DELHI | shree16us@ye | NA | NA |
| 5 | F01208 | CALTEX INDIA LIM | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | GOLD CREST 24 55 N | ROC DELHI | NA | NA | NA |
| 6 | F01218 | GE HEALTHCARE B | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | FF-3 Palani Centre1 | ROC DELHI | karthick9999@ | NA | NA |
| 7 | F01265 | CAIRN ENERGY IN | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | WELLINGTON PLAZA | ROC DELHI | neerja.sharma | NA | NA |
| 8 | F01269 | TORIELLI S.R.L | ACTV | NA | NA | NA | 05-09-1995 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 6, Mangayarkarsi Nu | ROC DELHI | chennai@torie | NA | NA |
| 9 | F01311 | HARDY EXPLORAT | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 5TH FLOOR,WESTMI | ROC DELHI | venkatesh.v@I | NA | NA |
| 10 | F01314 | HOCHTIOF AKTIEN | ACTV | NA | NA | NA | 11-04-1996 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | NEW NO.86, OLD NO | ROC DELHI | kumar@intern | NA | NA |
| 11 | F01412 | EPSON SINGAPOR | ACTV | NA | NA | NA | 25-04-1997 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 7C CEATURY PLAZA5E | ROC DELHI | NA | NA | NA |
| 12 | F01426 | CARGOLUX AIRLIN | ACTV | NA | NA | NA | 11-06-1997 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | OFFICE NO 91MEENA | ROC DELHI | NA | NA | NA |
| 13 | F01468 | CHO HEUNG ELECT | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 129, MANPUR VILLAG | ROC DELHI | chowelaccoun | NA | NA |
| 14 | F01543 | NYCOMED ASIA PA | ACTV | NA | NA | NA | 27-10-1998 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | A D 46 1ST STREET4( | ROC DELHI | NA | NA | NA |
| 15 | F01544 | CHERRINGTON AS | ACTV | NA | NA | NA | 01-05-2000 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 10HADDOWS ROAD | ROC DELHI | NA | NA | NA |
| 16 | F01563 | SHIMADZU ASIA P | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | FIRST FLOOR, NO.10( | ROC DELHI | kousik@vsnl.c | NA | NA |
| 17 | F01565 | CORK INTERNATIC | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | ARJAY APEX CENTRE I | ROC DELHI | NA | NA | NA |
| 18 | F01566 | ERBIS ENGG COMI | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 39,2nd Main Road,R | ROC DELHI | NA | NA | NA |
| 19 | F01589 | RALF SCHNEIDER I | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | FLAT C, 'SAI VASANTI | ROC DELHI | NA | NA | NA |
| 20 | F01593 | MITRAJAYA TRADI | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | OLD NO 148 NEW NO | ROC DELHI | NA | NA | NA |
| 21 | F01618 | HEAT AND CONTR | ACTV | NA | NA | NA | 13-07-1999 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | A40 OLD NO 26 6TH I | ROC DELHI | nirajagopal@ | NA | NA |
| 22 | F01628 | DIREX SYSTEMS LI | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | F-1, FIRST FLOOR, Q | ROC DELHI | direx@vsnl.co | NA | NA |
| 23 | F01641 | NMB-MINEBEA TH | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | Level - 2 Regus, Alti | ROC DELHI | stsogawa@mi | NA | NA |
| 24 | F01643 | ARROW INTERNAT | ACTV | NA | NA | NA | 02-11-1999 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | BLUE HAVEN, NO.19, | ROC DELHI | NA | NA | NA |
| 25 | F01694 | GAMBRO CHINA L | ACTV | NA | NA | NA | 14-06-2000 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 5 IST FLOOR IST STRI | ROC DELHI | NA | NA | NA |
| 26 | F01703 | OBARA CORPORAT | NAEF | NA | NA | NA | 17-07-2000 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | INDIA BRANCH OFFI | ROC DELHI | joe@obara.co. | NA | NA |
| 27 | F01752 | CIPTA WAWASON | ACTV | NA | NA | NA | 24-01-2001 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 141 AYYAI SHANMUI | ROC DELHI | NA | NA | NA |
| 28 | F01753 | AUCHAN INTERNA | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | RK Tower, No. 115, 5 | ROC DELHI | pverma@vkver | NA | NA |
| 29 | F01767 | TOSHIBA PLANT S | NAEF | NA | NA | NA | 08-03-2001 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | HOTEL AMBASSADOF | ROC DELHI | NA | NA | NA |
| 30 | F01768 | YAMAZEN CORPOI | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | PLOT 69, SIVANANDA | ROC DELHI | NA | NA | NA |
| 31 | F01770 | OWL INTERNATIO | ACTV | NA | NA | NA | 22-03-2001 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | NO 1 SAPTHAGIRI CC | ROC DELHI | NA | NA | NA |
| 32 | F01826 | LEXMARK INTERN | ACTV | NA | NA | NA | 16-08-2001 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | APEEJAY BUSINESS C | ROC DELHI | NA | NA | NA |
| 33 | F01830 | FLUID ENERGY CO | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | FLUID ENERGY CONT | ROC DELHI | jeeva@fecindi | NA | NA |
| 34 | F01861 | WATCH GUARD TE | ACTV | NA | NA | NA | 21-11-2001 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 54/2, paulwells roa | ROC DELHI | chennaiadmir | NA | NA |
| 35 | F01878 | SINAR JERUIH SDI | ACTV | NA | NA | NA | 24-12-2001 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | 57/4 SEVEN HILL APA | ROC DELHI | accounts.ho@ | NA | NA |
| 36 | F01918 | SIPLEC INTERNATI | ACTV | NA | NA | NA | 23-09-1995 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | III FLOOR, PARADE V | ROC DELHI | svrajacdm@ya | NA | NA |
| 37 | F01935 | INTELSAT GLOBAL | ACTV | NA | NA | NA | 20-05-2005 | Tamil Nadu | 0 | 0 | NA | Agriculture & allied | TPL HOUSE 2ND FLOI | ROC DELHI | NA | NA | NA |

**Figure 2.** Registrar of Company (RoC) Dataset.

**B. Data Variables:** Collect relevant variables such as company name, registration date, principal business activity, company type (public/private), etc.

**C. Data Cleaning:** Remove any duplicate or irrelevant entries. Handle missing values and outliers appropriately.

**D. Feature Engineering:** Extract additional features if necessary, such as year of registration, location of registration, etc.

**E. Normalization/Standardization:** Standardize numerical features if required for certain machine learning algorithms.

### 3.2 Data Exploration

To begin, descriptive statistics are computed for pivotal variables like registration counts per year and the distribution of principal business activities. Following this, visual exploration, is employed. Techniques such as histograms, bar charts, and heatmaps are utilized to delve into the distribution and interconnections among the variables, facilitating a comprehensive understanding. Finally, spatial analysis takes center stage, where spatial patterns of company registrations are scrutinized by geographic location utilizing maps or spatial analysis techniques, providing insights into geographical trends and concentrations [9].

### 3.3 Feature Selection

**A. Correlation Analysis:** It is used to identify correlated features and eliminate redundant ones to improve model performance.

**B. Principal Component Analysis (PCA):** It is used to perform dimensionality reduction for the dataset that contains a large number of features.

## 3.4 Predictive Modeling

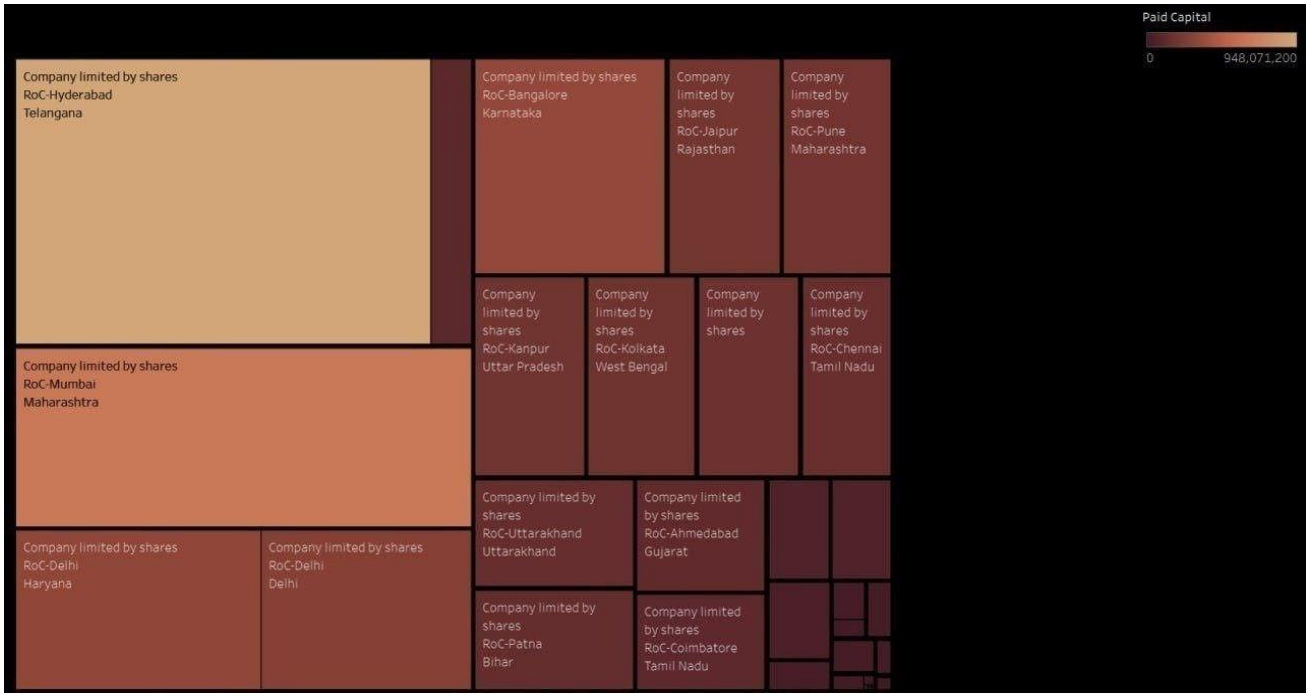


**Figure 3.** Evaluating the Performance of Data

Initially, a selection of machine learning algorithms suitable for the task is made, encompassing techniques such as Linear Regression, Decision Trees, Random Forest, Gradient Boosting Machines (GBM), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naive Bayes, among others. Following this, the dataset is divided into training (80%) and test sets (20%) to facilitate model evaluation. Hyper parameters of each model are then fine-tuned using techniques like grid search to optimize their performance. Subsequently, the performance of each model is rigorously assessed utilizing appropriate metrics such as Mean Squared Error for regression models and accuracy, precision, and recall for classification models, as depicted in Figure 3. Furthermore, ensemble methods like model averaging or stacking are explored to combine predictions from multiple models, aiming for enhanced accuracy and robustness in the prediction [10].

## 3.5 Visualization with Tableau

Step 1: Import pre-processed data into Tableau [7] for visualization.

Step 2: Design interactive dashboards to visualize company registration trends, distribution by location, principal business activities, etc as shown in Figure 4.

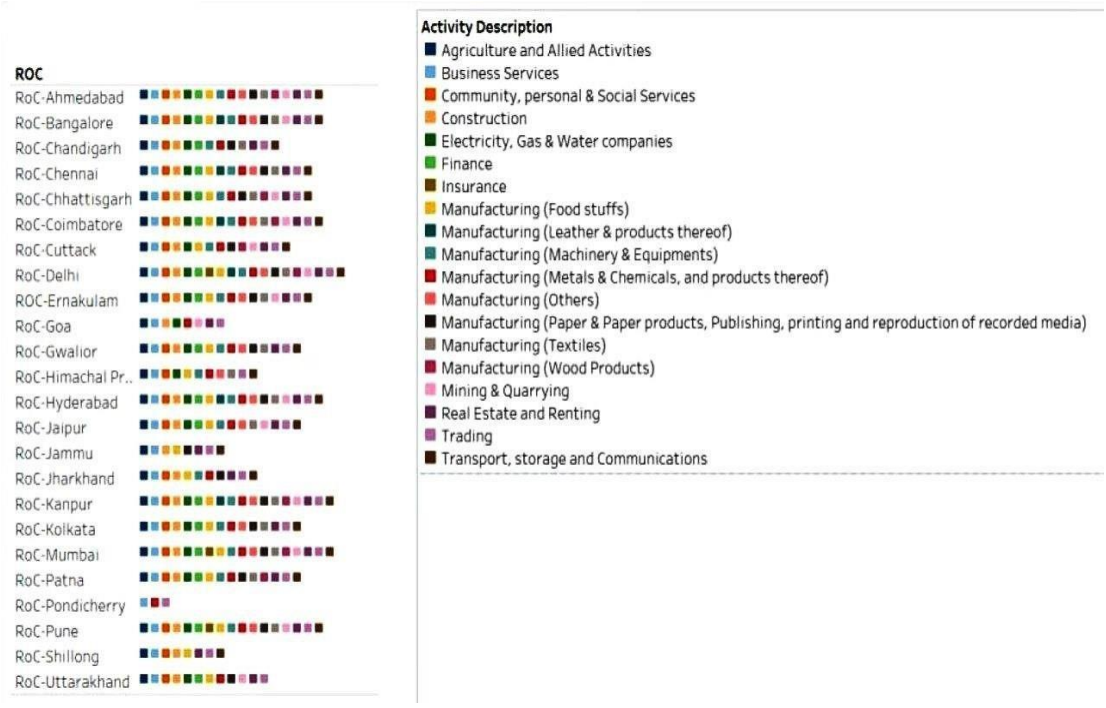


**Figure 4.** Activity description of RoC.

Step 3 : Enable dynamic filtering to allow users to explore the data interactively based on different criteria.

Step 4 : Utilize maps in Tableau to visualize spatial patterns of company registrations.

### 3.6 Prediction with Weka

Step 1: Format the dataset Attribute-Relation File Format (ARFF).

Step 2: Model Building: Implement machine learning models using Weka [5] for prediction.

Step 3 : Perform cross-validation to assess the generalization performance of the models.

Step 4 : Visualize the prediction results generated by Weka [8].

### 3.7 Model Comparison and Selection

Compare the performance of different machine learning algorithms based on evaluation metrics.

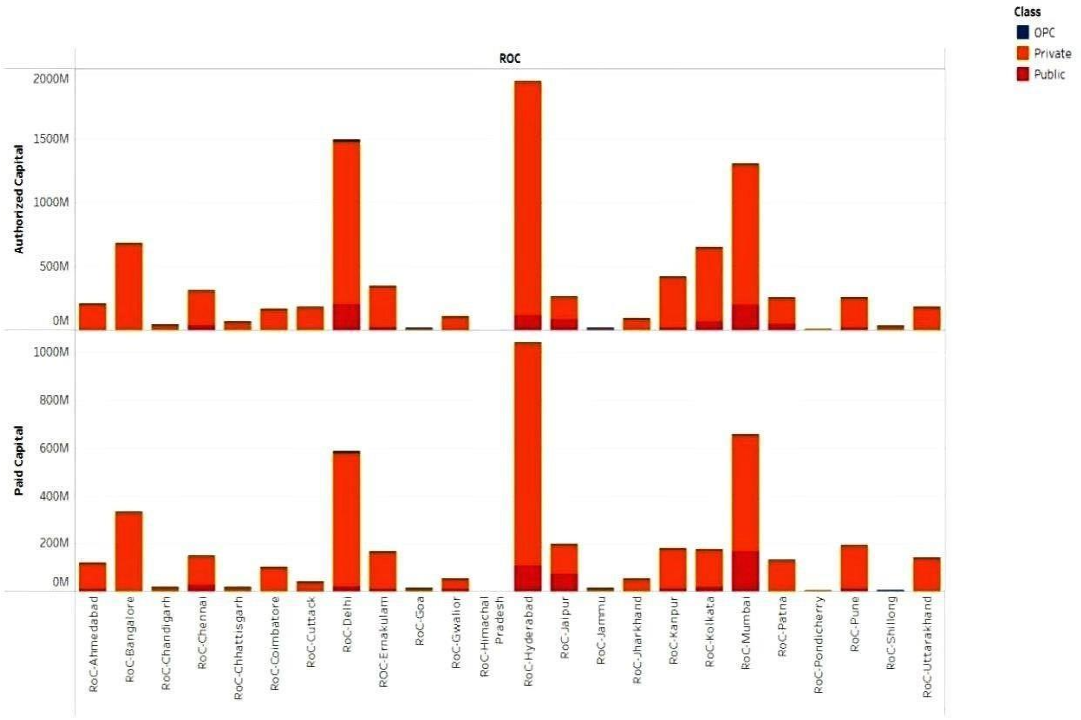


**Figure 5.** Comparison of RoC using BarChart.

Select the most appropriate models for predicting company registration trends [3] as shown in Figure 5 based on overall performance.

## 4. Result and Discussion

The findings of our analysis have several implications for various stakeholders, including policymakers, investors, and business owners. By accurately predicting company registration trends, decision-makers can anticipate market demands, identify emerging opportunities, and formulate effective strategies for growth and investment.

Furthermore, the insights gained from our analysis can inform regulatory policies aimed at fostering entrepreneurship, promoting economic development, and ensuring equitable distribution of business opportunities across different regions and sectors.

Overall, our study demonstrates the value of leveraging AI-driven approaches and predictive analytics to gain actionable insights from Registrar of Companies data. The Table.1 shows the model performance comparison in fundamental analysis.

**Table 1.** Models Performance Comparison in Fundamental Analysis

| **Metrices** | **LR** | **SLR** | **RF** | **RT** | **MLP** | **ZeroR** | **REP Tree** | **DT** |
|---|---|---|---|---|---|---|---|---|
| Correlation coefficient | -0.4118 | -0.4156 | 0.2502 | 0.2088 | 0.0019 | -0.6056 | -0.3798 | 0.3798 |
| Mean absolute error | 11.3834 | 11.2191 | 9.737 | 11.5787 | 11.0693 | 9.6857 | 9.6893 | 9.4422 |
| Root mean squared error | 21.6899 | 21.1145 | 15.848 | 20.6991 | 18.4742 | 15.4202 | 15.5499 | 15.5683 |
| Relative absolute error | 117.527 | 115.831 | 100.529 | 119.543 | 114.284 | 100.000 | 100.037 | 97.4855 |
| Root relative squared error | 140.659 | 136.928 | 102.774 | 134.233 | 119.805 | 100.000 | 100.841 | 100.960 |

Where,

LR – Linear Regression

SLR – Simple Linear Regression

RF – Random Forest

RT – Random Tree

MLP – Multi-Layer Perceptron

REP Tree – Reduced Error Pruning Tree

DT – Decision Tree

By combining advanced analytical techniques with intuitive visualization tools, we empower stakeholders to make informed decisions and navigate the dynamic landscape of business registration trends in India.

## 5. Conclusion

In conclusion, our study provides a comprehensive analysis of company registration trends in India, leveraging AI-driven approaches and machine learning algorithms. By exploring RoC data, we uncovered significant insights into registration patterns, geographical concentrations, and sector classifications. Our predictive models, validated through various algorithms and tools, offer actionable insights for stakeholders. This research contributes to the understanding of business dynamics and can inform strategic decision-making processes. As India continues to experience economic growth and business expansion, our findings serve as a valuable resource for anticipating future trends and fostering sustainable development in the corporate sector.

## 6. Future Works

Proposal of potential avenues for future research and development: Incorporation of additional data sources for more comprehensive analysis (e.g., economic indicators, industry reports). Exploration of advanced machine learning techniques or ensemble methods for improved prediction accuracy. Investigation of the impact of external factors (e.g., regulatory changes, economic events) on company registration trends. Development of predictive models tailored to specific industry sectors or regions within India. Integration of real-time data streams for dynamic updates and forecasting.

## References

[1] Akash Patel, Devang Patel, Seema Yadav.(2021),"Prediction of stock market using Artificial Intelligence", Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021). 2021.1-6

[2] Sohrab Mokhtari, Kang K.Yen,Jin Liu.(2021),” Effectiveness of Artificial Intelligence in Stock Market Prediction based on Machine Learning”, 183(7),1-8,

[3] Popescu, Cristina Raluca Gh, and Poshan Yu, eds. Intersecting Environmental Social Governance and AI for Business Sustainability. IGI Global, 2024.

[4] Semenkevich, Ekaterina. "Use of AI-Driven Targeted Marketing on Instagram by Online Retail SMEs." (2024).

[5] Cockburn, Iain M., Rebecca Henderson, and Scott Stern. The impact of artificial intelligence on innovation. Vol. 24449. Cambridge, MA, USA: National bureau of economic research, 2018.

[6] Milica Mitrović , Slađana Janković, Snežana Mladenović .(2022), “Prediction of Daily Demand for Goods using Weka Software too”l,5th Logistics International Conference,283-292.