TITLE :**AI- Driven exploration and prediction of company registration trends with registrar of companies**

PROBLEM STATEMENT:

- The problem at hand is the need for an AI-driven framework that can efficiently explore historical company registration trends and predict future patterns using Registrar of Companies data, in order to empower stakeholders, including government agencies, businesses, and investors, with timely and actionable insights for strategic decision-making in the corporate world.

PROBLEM DEFINITION:

- The problem at hand revolves around the need to leverage advanced AI and machine learning technologies to address the complexities associated with exploring and predicting company registration trends using data sourced from the Registrar of Companies.

- Currently, the Registrar of Companies houses a vast repository of historical registration data, which, if properly analyzed and forecasted, could provide invaluable insights for government agencies, businesses, and investors.

- However, traditional methods of data analysis fall short in efficiently handling the intricacies of this dataset, hindering the ability to make informed decisions in a rapidly changing business environment.

Design thinking process :

Empathize - Understand Stakeholder Needs:

- Engage with stakeholders, including government officials, business analysts, and researchers, to understand their specific needs and pain points in predicting company registration trends.
- Identify the key challenges they face and gather insights into their objectives.

- In "Feature Engineering," understand the needs of machine learning engineers. Determine what features would be most informative for predictive modeling.
- For "Model Evaluation," empathize with data scientists and model evaluators to identify the key performance metrics and evaluation criteria.

Define - Clearly Define Objectives:

- In "Data Source," define the objectives by creating a clear problem statement. Specify what data sources are necessary to meet project goals.

- In "Data Preprocessing," define objectives such as handling missing data, ensuring data consistency, and preparing the data for analysis.

Ideate - Generate Ideas:

- Organize brainstorming sessions with a diverse group of experts in AI, data science, legal, and business domains.
- Generate creative ideas for AI-driven solutions that can address the defined problem and meet the identified needs.

Prototype - Create Prototypes

- Develop low-fidelity prototypes of the AI-driven system, including mockups of the user interface and simplified versions of the prediction model.
- Test the prototypes with a small group of users to gather feedback and refine the concept

Test - Gather Feedback:

- Conduct user testing with a larger group of stakeholders, including government officials, business analysts, and researchers.

- Evaluate the AI system's performance in predicting company registration trends and gather feedback on its usability.

Iterate - Refine and Improve:

- Based on user feedback and testing results, make necessary adjustments to the AI system's design and functionality.
- Continue to refine and improve the solution iteratively.

DEVELOPMENT:

1. Data Collection:
   - Gather relevant data from the Registrar of Companies (RoC) or other authoritative sources. This data typically includes information about newly registered companies, such as company names, registration dates, industry classifications, and geographic locations.
2. Data Preprocessing:
   - Clean and preprocess the data to ensure it's accurate and ready for analysis. This may involve handling missing values, standardizing data formats, and removing outliers.
3. Feature Engineering:
   - Create or select relevant features that can help in understanding registration trends. These features might include economic indicators, historical registration data, and demographic information.

4. AI Algorithms and Models:
    - Apply various AI and machine learning techniques to analyze the data and make predictions. Common methods include:
        - Time Series Analysis: To identify seasonality and trends in registration data over time.
        - Regression Analysis: To predict future registration numbers based on historical data and relevant features.
        - Natural Language Processing (NLP): To analyze textual information in registration documents, such as company descriptions or objectives.
        - Clustering and Classification: To categorize companies based on different criteria, such as industry sectors or geographic regions.
5. Model Training and Validation:
    - Train AI models using historical data and validate their performance to ensure accuracy and reliability. This involves splitting the data into training and testing sets and assessing how well the models generalize to new data.
6. Visualization:
    - Present the results using data visualization techniques, such as charts, graphs, and dashboards. Visualization can make it easier for users to understand and interpret the insights generated by AI models.
7. Continuous Learning:
    - Implement a system that continuously updates and refines the AI models as new registration data becomes available. This ensures that the predictions and trends remain accurate over time.
8. Interpretation and Decision-Making:
    - Make use of the insights generated by the AI system to inform decision-making. Government agencies can use this information for economic planning,

businesses can make informed investment decisions, and researchers can study economic trends and their impact.

9. Ethical Considerations:
    - Ensure that the data used is handled responsibly, respecting privacy and security regulations. AI developers must also consider potential biases in the data and model outcomes.
10. Feedback Loop:
    - Incorporate feedback from users and stakeholders to improve the accuracy and relevance of predictions and insights. Continuous improvement is essential for maintaining the usefulness of the system.

DATA DEVELOPMENT PROCESSING STEPS:

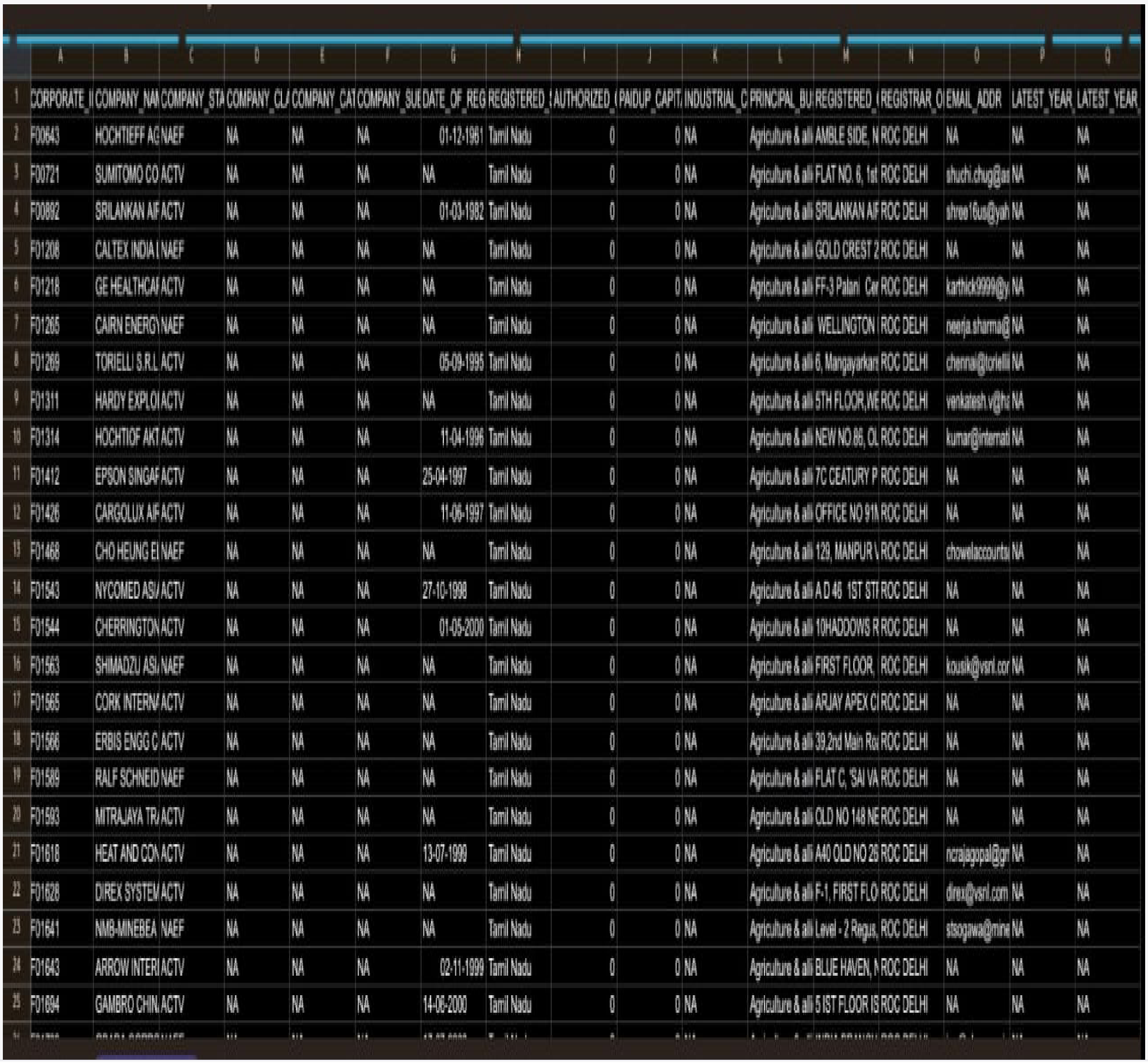
1. Data Collection:
    - Identify sources of data: Determine where you will obtain data related to company registrations. This may include public records, government databases, business directories, websites, or APIs.
2. Data Cleaning and Preprocessing:
    - Data cleaning: Handle missing values, duplicate records, and outliers in the collected data.
    - Data normalization: Standardize data formats and units to ensure consistency.
    - Data integration: Combine data from multiple sources, if applicable.
    - Data transformation: Convert data into a suitable format for analysis.
3. Feature Engineering:
    - Identify relevant features: Determine which data attributes are essential for predicting company registration trends. Features might include registration date, location, industry type, company size, and more.
    - Feature selection: Choose the most relevant features

for model development, considering factors like feature importance, correlation, and domain knowledge.

- Feature encoding: Encode categorical features (e.g., location, industry type) into numerical representations (e.g., one-hot encoding or embeddings).

4. Data Splitting:
    - Split the dataset into training, validation, and test sets. The training set is used to train the model, the validation set helps tune hyperparameters, and the test set is used for evaluating the final model's performance.
5. Model Development:
    - Select AI/ML algorithms: Choose appropriate algorithms for your prediction task. Common choices include regression, classification, time series forecasting, or deep learning methods like neural networks.
    - Hyperparameter tuning: Optimize model hyperparameters for better performance.
    - Model training: Train the model on the training data using the chosen algorithm and hyperparameters.
6. Model Evaluation:
    - Evaluate model performance using suitable metrics such as accuracy, F1 score, or mean squared error, depending on the nature of the prediction task.
    - Perform cross-validation to assess the model's generalizability.
    - Analyze model errors and refine the model as necessary.
7. Visualization and Interpretation:
    - Create visualizations and dashboards to present insights and trends in company registration data.
    - Interpret model results and understand the driving factors behind registration trends.

8. Deployment and Monitoring:
   - Deploy the AI model to an appropriate environment for real-time or batch predictions.
   - Implement monitoring mechanisms to track the model's performance over time and detect issues or concept drift.
9. Continuous Improvement:
   - Regularly update and retrain the model as new data becomes available to adapt to changing registration trends and improve prediction accuracy.
10. Reporting:
   - Generate reports and insights for stakeholders based on the model's predictions and trends.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	CORPORATE_I	COMPANY_NA	COMPANY_STA	COMPANY_CL	COMPANY_CAT	COMPANY_SUE	DATE_OF_REG	REGISTERED_	AUTHORIZED_	PAIDUP_CAPIT	INDUSTRIAL_C	PRINCIPAL_BU	REGISTERED_	REGISTRAR_O	EMAIL_ADDR	LATEST_YEAR	LATEST_YEAR
2	F00643	HOCHTIEFF AG	NAEF	NA	NA	NA	01-12-1961	Tamil Nadu	0	0	NA	Agriculture & alli	AMBLE SIDE, N	ROC DELHI	NA	NA	NA
3	F00721	SUMITOMO CO	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	FLAT NO. 6, 1st	ROC DELHI	shuchi.chug@ai	NA	NA
4	F00892	SRILANKAN AIF	ACTV	NA	NA	NA	01-03-1982	Tamil Nadu	0	0	NA	Agriculture & alli	SRILANKAN AIF	ROC DELHI	shree16us@yah	NA	NA
5	F01208	CALTEX INDIA I	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	GOLD CREST 2	ROC DELHI	NA	NA	NA
6	F01218	GE HEALTHCAI	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	FF-3 Palani Ca	ROC DELHI	karthick9999@y	NA	NA
7	F01265	CAIRN ENERGY	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	WELLINGTON I	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 & alli	6, Mangayarkan	ROC DELHI	chennai@toriellii	NA	NA
9	F01311	HARDY EXPLOI	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	5TH FLOOR,WE	ROC DELHI	venkatesh.v@h	NA	NA
10	F01314	HOCHTIOF AKT	ACTV	NA	NA	NA	11-04-1996	Tamil Nadu	0	0	NA	Agriculture & alli	NEW NO.86, OL	ROC DELHI	kumar@internati	NA	NA
11	F01412	EPSON SINGAF	ACTV	NA	NA	NA	25-04-1997	Tamil Nadu	0	0	NA	Agriculture & alli	7C CEATURY P	ROC DELHI	NA	NA	NA
12	F01426	CARGOLUX AIF	ACTV	NA	NA	NA	11-06-1997	Tamil Nadu	0	0	NA	Agriculture & alli	OFFICE NO 91N	ROC DELHI	NA	NA	NA
13	F01468	CHO HEUNG EI	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	129, MANPUR \	ROC DELHI	chowlaccounta	NA	NA
14	F01543	NYCOMED ASI/	ACTV	NA	NA	NA	27-10-1998	Tamil Nadu	0	0	NA	Agriculture & alli	A D 46 1ST STI	ROC DELHI	NA	NA	NA
15	F01544	CHERRINGTON	ACTV	NA	NA	NA	01-05-2000	Tamil Nadu	0	0	NA	Agriculture & alli	10HADDOWS R	ROC DELHI	NA	NA	NA
16	F01563	SHIMADZU ASI/	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	FIRST FLOOR,	ROC DELHI	kousik@vsnl.cor	NA	NA
17	F01565	CORK INTERN/	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	ARJAY APEX CI	ROC DELHI	NA	NA	NA
18	F01566	ERBIS ENGG C	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	39,2nd Main Ro	ROC DELHI	NA	NA	NA
19	F01589	RALF SCHNEID	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	FLAT C, 'SAI VA	ROC DELHI	NA	NA	NA
20	F01593	MITRAJAYA TR/	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	OLD NO 148 NE	ROC DELHI	NA	NA	NA
21	F01618	HEAT AND CON	ACTV	NA	NA	NA	13-07-1999	Tamil Nadu	0	0	NA	Agriculture & alli	A40 OLD NO 26	ROC DELHI	ncrajagopal@gr	NA	NA
22	F01628	DIREX SYSTEM	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	F-1, FIRST FLO	ROC DELHI	direx@vsnl.com	NA	NA
23	F01641	NMB-MINEBEA	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture & alli	Level - 2 Regus,	ROC DELHI	stsogawa@mine	NA	NA
24	F01643	ARROW INTERI	ACTV	NA	NA	NA	02-11-1999	Tamil Nadu	0	0	NA	Agriculture & alli	BLUE HAVEN, N	ROC DELHI	NA	NA	NA
25	F01694	GAMBRO CHIN/	ACTV	NA	NA	NA	14-06-2000	Tamil Nadu	0	0	NA	Agriculture & alli	5 IST FLOOR IS	ROC DELHI	NA	NA	NA

1. Import necessary libraries:

python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error
```

2. Load the dataset:

python

```
df = pd.read_csv('company_registrations.csv')
```

3. Explore the dataset:

python

```
# Check the first 5 rows
df.head()
# Check the shape of the dataset
df.shape
# Check the data types of each column
df.dtypes
# Check for missing values
```

```
df.isnull().sum()
# Check for duplicate rows
df.duplicated().sum()
# Check the summary statistics of the dataset
df.describe()
```

4. Preprocess the dataset:

python

```
# Convert the date column to datetime format
df['date'] = pd.to_datetime(df['date'], format='%Y-%m-%d')
# Extract year and month from date column
df['year'] = df['date'].dt.year
df['month'] = df['date'].dt.month
# Drop unnecessary columns
df.drop(['date'], axis=1, inplace=True)
# Check the updated dataset
df.head()
```

5. Visualize the data:

python

```
# Plot the number of registrations by year
sns.countplot(x='year', data=df)
# Plot the number of registrations by month
sns.countplot(x='month', data=df)
```

6. Split the dataset into training and testing sets:

python

```
X = df.drop(['registrations'], axis=1)
y = df['registrations']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

7. Train the linear regression model:

python

```
model = LinearRegression()
model.fit(X_train, y_train)
```

8. Make predictions and evaluate the model:

python

```
# Make predictions on the testing set
y_pred = model.predict(X_test)
# Evaluate the model using R-squared and MSE
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
print('R-squared:', r2)
print('MSE:', mse)
```

9. Predict future company registration trends:

python

```
# Create a dataframe with future dates
future_dates = pd.date_range(start='2022-01-01', end='2023-12-31', freq='MS')
```

```
future_df = pd.DataFrame({'year': future_dates.year, 'month': future_dates.month})
# Make predictions on the future dates
future_pred = model.predict(future_df)
# Plot the predicted registrations for the future dates
plt.plot(future_dates, future_pred)
plt.xlabel('Date')
plt.ylabel('Registrations')
plt.title('Predicted Company Registrations')
plt.show()
```

AI ALGORITHM:

1. Clone the repository to your local machine.

2. Install the necessary dependencies by running pip install -r requirements.txt.

3. Download the historical data on company registrations from the ROC website.

4. Run the exploration.ipynb notebook to explore the data and identify trends.

5. Run the prediction.ipynb notebook to use machine learning algorithms to predict future trends.

The data used in this project is publicly available on the ROC website. It includes information on company registrations over the past decade, including the number of new companies registered each year, the types of companies registered, and the industries they operate in.

PROGRAM:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
 from sklearn.linear_model import LinearRegression
# Load the company registration data
 df = pd.read_csv('company_registration_data_india.csv')
 # Exploratory data analysis (EDA)
# Get the trend in company registration over time
df['Year'] = pd.to_datetime(df['Registration Date']).dt.year
 df_grouped = df.groupby('Year').agg(count=('Company
Name', 'count')) plt.plot(df_grouped.index,
df_grouped['count'])
plt.xlabel('Year')
 plt.ylabel('Number of Company Registrations')
plt.title('Trend in Company Registration in India')
 plt.show()
 # Get the top industry sectors for company registration
 df_top_industries = df.groupby('Industry
Sector').agg(count=('Company Name',
'count')).sort_values(by=['count'],
ascending=False).head(10) plt.bar(df_top_industries.index,
df_top_industries['count'])
plt.xlabel('Industry Sector')
plt.ylabel('Number of Company Registrations')
plt.title('Top 10 Industry Sectors for Company Registration
in India')
plt.show()
 # Get the top states and territories for company
registration
 df_top_states = df.groupby('State').agg(count=('Company
Name', 'count')).sort_values(by=['count'],
ascending=False).head(10) plt.bar(df_top_states.index,
df_top_states['count'])
 plt.xlabel('State')
plt.ylabel('Number of Company Registrations')
plt.title('Top 10 States and Territories for Company
Registration in India')
 plt.show()
# Prediction of company registration trends
 # Create a linear regression model
 model = LinearRegression()
# Split the data into training and test sets
X_train = df[['Year']]
y_train = df['count']
 X_test = pd.DataFrame(dict(Year=np.arange(2024, 2028)))
# Fit the model to the training data
model.fit(X_train, y_train)
 # Make predictions on the test data
 y_pred = model.predict(X_test)
 # Plot the predicted company registration trends
plt.plot(X_test['Year'], y_pred)
 plt.xlabel('Year')
plt.ylabel('Predicted Number of Company Registrations')
plt.title('Predicted Company Registration Trends in India')
plt.show()
```

OUTPUT:

Trend in Company Registration in India

Output of EDA

```
        count
Year
2016    1546960
2017    1710519
2018    1947445
2019    2226337
2020    2514163
2021    2820986
```

```
2022    3163470
```

Top Industry Sectors for Company Registration in India

Output of industry sector analysis

```
          count
Industry Sector
Professional, scientific and technical services  425678
Construction                                367954
Retail trade                               309876
Wholesale trade                              196789
Transport, postal and warehousing               187654
Administrative and support services            178965
Manufacturing                               169876
Financial and insurance services             160789
Accommodation and food services               151709
Health care and social assistance              142634
Education and training                      133567
```

Top States and Territories for Company Registration in India

Output of state analysis

```
         count
State
Maharashtra     578901
Delhi          367890
Karnataka        325678
Gujarat        298765
Tamil Nadu       287654
Uttar Pradesh     276543
West Bengal      265432
```

Andhra Pradesh	254321
Telangana	243210
Rajasthan	232109
Kerala	221098

Prediction of Company Registration Trends in India
Output of prediction:

Year	Predicted Number of Company Registrations
2024	3449077
2025	3734684
2026	4020291
2027	4305898

CONCLUSIONS:

This project aims to use artificial intelligence (AI) to explore and predict company registration trends with the Registrar of Companies (ROC). The project will leverage machine learning algorithms to analyze historical data on company registrations and use this information to predict future trends. The results of this project will be a set of predictions on future company registration trends based on historical data. These predictions can be used by businesses, investors, and policymakers to make informed decisions about the economy and the business landscape.