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

## Table of Contents

1. [Prerequisites](#prerequisites)

2. [Code Overview](#code-overview)

3. [Running the Code](#running-the-code)

4. [Interpreting the Results](#interpreting-the-results)

---

## 1. Prerequisites <a name="prerequisites"></a>

Before running the code, ensure you have the following prerequisites in place:

- Python 3.x installed on your system.

- Required libraries and packages installed. You can install the necessary packages using the following command:

```

pip install pandas numpy sklearn statsmodels matplotlib seaborn

```

---

## 2. Code Overview <a name="code-overview"></a>

The code is divided into several sections, each with a specific purpose. Here is a brief overview of each section:

1. \*\*Data Loading and Cleaning:\*\*

- Loads data from the 'Data\_Gov\_Tamil\_Nadu.csv' file.

- Checks for missing values and interpolates them.

- Converts categorical features to numerical using label encoding.

2. \*\*Data Preprocessing:\*\*

- Scales numerical columns using StandardScaler.

- Performs label encoding and one-hot encoding for categorical variables.

- Engineers date-based features, calculates interaction features, and handles missing values.

3. \*\*Exploratory Data Analysis (EDA):\*\*

- Provides insights into the dataset using visualizations and statistics.

- Displays histograms, correlation matrices, and count plots for different attributes.

4. \*\*Classification Model:\*\*

- Prepares the data for a classification task.

- Splits the data into training and testing sets.

- Trains a Random Forest classifier and evaluates its performance.

5. \*\*Time Series Forecasting:\*\*

- Assumes a time-based attribute 'DATE\_OF\_REGISTRATION'.

- Converts date data to time series data.

- Fits an ARIMA model for time series forecasting.

- Forecasts future registration trends and plots the results.

6. \*\*Prediction:\*\*

- Uses the trained classification model to make predictions on new data. Note that this step assumes you have new data for prediction.

---

## 3. Running the Code <a name="running-the-code"></a>

To run the code, follow these steps:

1. Make sure you have the prerequisites installed as mentioned in the "Prerequisites" section.

2. Download the 'Data\_Gov\_Tamil\_Nadu.csv' dataset and place it in the same directory as the code.

3. Execute the code in your Python environment. You can run the entire script, or you can run specific sections individually as needed.

- \*\*Data Loading and Cleaning\*\*: Ensure that the data file 'Data\_Gov\_Tamil\_Nadu.csv' is in the same directory, then run this section to load and clean the data.

- \*\*Data Preprocessing\*\*: Run this section to perform data preprocessing. It includes scaling, encoding, feature engineering, and imputation.

- \*\*Exploratory Data Analysis (EDA)\*\*: Execute this section to explore the dataset through various visualizations and statistical summaries.

- \*\*Classification Model\*\*: Run this section after data preprocessing to perform classification on the dataset. The code will train a Random Forest classifier and provide classification results.

- \*\*Time Series Forecasting\*\*: Run this section if your dataset contains a time-based attribute. It will perform time series forecasting and display the results.

- \*\*Prediction\*\*: Use this section if you have new data for prediction. The trained classification model will be used to make predictions.

4. Review the results and visualizations produced by each section of the code.

---

## 4. Interpreting the Results <a name="interpreting-the-results"></a>

- \*\*Data Loading and Cleaning\*\*: This section ensures that the dataset is loaded and cleaned, handling missing values and encoding categorical features.

- \*\*Data Preprocessing\*\*: The data preprocessing section prepares the data for modeling. It scales numerical columns, encodes categorical variables, engineers date features, calculates interaction features, and imputes missing values. The result is an updated dataset ready for classification, prediction, or further analysis.

- \*\*Exploratory Data Analysis (EDA)\*\*: EDA provides insights into the dataset through visualizations such as histograms, correlation matrices, and count plots. These visualizations help you understand the data distribution and relationships between variables.

- \*\*Classification Results\*\*: After running the classification model, you'll get accuracy, F1-score, and a classification report. These metrics assess the performance of the classification model.

- \*\*Time Series Forecasting\*\*: The time series forecasting section will provide a plot showing actual and forecasted registration trends. You can adjust the forecasting period as needed.

- \*\*Prediction\*\*: Use the prediction section to make predictions on new data. Ensure that you have new data in a suitable format for the model. The code will provide predictions based on the trained model.