

|  |  |  |  |
| --- | --- | --- | --- |
| **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**  **Domain Name : Artificial Intelligence**  **Project Title : AI-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies (RoC)** | | | |
| **1.** | **Name of the Student (s)** |  |
|  | |  |  |  |  | | --- | --- | --- | --- | | **S.No** | **Name of the Student** | **E-Mail ID** | **Phone No.** | | **1** | Mathan Raj.M | mathanrajm2003@gmail.com | 6379152566 | | **2** | Karthikeyan M | karthikn6482@gmail.com | 9363313534 | | **3** | Monish M | monishkhokho11@gmail.com | 9345340951 | | 4 | Vijay Sanjay kumar K.r | krvskevil3@gmail.com | 8015539457 | | 5 | Varun S.P | varunleela4@gmail.com | 9345560205 | | | |
| **2.** | **Name of the Guide** | : Ms.S.Suganya |
|  | **Department/ Designation** | : CSE**/**AP |
|  | **Institutional Address** | : Chettinad College of Engineering  and Technology  NH-67, Karur-Trichy Highway,  Puliyur CF, Karur |
|  | **Phone No. & Mobile No.** | : 6374527207 |

**AI-Driven Exploration and Predictive Analysis of Company Registrations**

**Architecture**

**Start**

**|**

**|--- Define Project Objectives**

**|**

**|--- Data Collection**

**| |**

**| |---** Gather RoC Data

| |

| |--- Collect External Data (Economic, Demographic, etc.)

**|**

**|--- Data Preprocessing**

**| |**

**| |---** Data Cleaning

| |

| |--- Feature Engineering

| |

| |--- Merge RoC Data with External Data

**|**

**|--- Exploratory Data Analysis (EDA)**

**| |**

**| |---** Visualize Data

| |

| |--- Identify Patterns and Anomalies

**|**

**|--- Model Selection**

**| |**

**| |---** Choose Ensemble Algorithms (e.g., Random Forest, Gradient Boosting)

| |

| |--- Choose Time Series Forecasting Algorithms (e.g., ARIMA, LSTM)

|

|--- Data Splitting

| |

| |--- Split Data into Training and Validation Sets

**|**

**|--- Model Training and Tuning**

**| |**

**| |---** Train Ensemble Models

| |

| |--- Tune Hyperparameters

**|**

**|--- Time Series Forecasting**

**| |**

**| |---** Train Time Series Forecasting Models

| |

| |--- Tune Time Series Model Parameters

**|**

**|--- Model Evaluation**

**| |**

**| |---** Evaluate Ensemble Model(s)

| |

| |--- Evaluate Time Series Forecasting Model(s)

| |

| |--- Select the Best Models

|

**|--- Interpretability and Visualization**

**| |**

**|** |--- Explain Model Predictions

| |

| |--- Create Interactive Dashboards

**|**

**|--- Real-Time Updates and Monitoring**

**| |**

**| |---** Set Up Real-Time Data Feeds

| |

| |--- Implement Model Monitoring

|

**|--- Deployment**

**| |**

**| |---** Deploy Models to Production

| |

| |--- Create User Interfaces

**|**

**|--- Ethical Considerations and Compliance**

**| |**

**| |---** Address Data Privacy and Bias Concerns

| |

| |--- Ensure Compliance with Regulations

**|**

**|--- Feedback and Improvement**

**| |**

**| |---** Gather User Feedback

| |

| |--- Continuously Improve Models and System

|

**|--- Project Conclusion**

**| |**

**| |---** Summarize Key Findings and Insights

**|**

**End**

**Model Overview :**

The purpose of this model is to analyze historical data of company registrations with the Registrar of Companies (RoC) to uncover patterns, insights, and predict future registration trends.

**Here's how it works:**

* It collects and preprocesses data, including external factors like economic indicators.
* Utilizes ensemble techniques to explore and identify hidden patterns.
* Employs time series forecasting models to predict future registration trends.
* Presents results through interpretable visualizations and dashboards.
* Continuously monitors data for real-time updates and adapts its predictions accordingly.

In essence, this model aims to provide valuable insights and predictions to assist businesses, investors, and policymakers in understanding and navigating the business landscape.

**Dataset Observation:**

The Data provided has some missed values and null attributes which affects the prediction efficiency of the model to override it we want to do some data cleaning process and feature engineering to make the pattern of the data to be efficient

To remove null values and other missing data in a dataset using Python, you can use the Pandas library. Below is a description and Python code snippet to achieve this for your dataset with the specified columns.

**Description:**

In this code snippet, we'll use Pandas to load your dataset and then remove rows (or columns) with missing values. We'll drop any row where at least one of the columns has a missing value.

import pandas as pd

# Load your dataset, assuming it's in a CSV file

# Replace 'your\_dataset.csv' with the actual file path

df = pd.read\_csv(‘Data\_Gov\_Tamil\_Nadu.csv')

# Check for missing values in the dataset

missing\_data = df.isnull().sum()

# Print the count of missing values for each column

print("Missing Data Summary:")

print(missing\_data)

# Remove rows with any missing values

df\_cleaned = df.dropna()

# You can also remove columns with missing values using df.dropna(axis=1)

# For example, to remove columns with any missing values:

# df\_cleaned = df.dropna(axis=1)

# Save the cleaned dataset to a new file

# Replace 'cleaned\_dataset.csv' with the desired file name

df\_cleaned.to\_csv('cleaned\_dataset.csv', index=False)

# Print the first few rows of the cleaned dataset

print("\nCleaned Dataset:")

print(df\_cleaned.head())

# Optional: You can also fill missing values with a specific value if needed

# For example, to fill missing values with 0:

# df\_filled = df.fillna(0)

**Ensemble Learning :**

To use ensemble algorithms for your dataset, you can follow these steps. I'll provide a sample code snippet using the scikit-learn library in Python. In this example, I'll use a Random Forest Classifier for a classification task, but you can adapt it for other ensemble methods and regression tasks as needed**.**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.ensemble import RandomForestClassifier**

**Accuracy Evaluation Method:**

**Classification Task:**

If the project involves classifying trends (e.g., "upward trend" or "downward trend" in company registrations):

Accuracy: Calculate the accuracy of the model, which is the ratio of correctly predicted trends to the total predictions.

Regression Task:

If the project involves predicting numerical trends (e.g., the number of new company registrations):

Mean Absolute Error (MAE): Measure the absolute differences between predicted and actual values and calculate their mean.

Mean Squared Error (MSE): Measure the squared differences between predicted and actual values and calculate their mean.

Root Mean Squared Error (RMSE): Take the square root of the MSE, which provides a measure in the same units as the target variable.

These evaluation metrics will help assess the accuracy and performance of the predictive model in the project. The specific metric to use depends on whether the project involves classification or regression tasks.

In practice, it's also beneficial to consider other relevant metrics or domain-specific evaluation criteria depending on the project's specific goals and requirements.