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

**PROJECT TITLE: REGISTER OF COMPANIES**

**(ROC)**

**CONTENTS:**

**1.Introduction**

**2.Objective**

**3.Methodology**

3.1Data Collection

3.2Data Preprocessing

3.3Data Splitting

**4.Machine Learning Approach**

4.1Feature Engineering

4.2Model Selection

4.3Model Training

**5.Evaluation**

**6.Prediction**

**7.Monitoring and Updating**

**8.Interpretability**

**9.Deployment**

**10.Compilance**

**11.Dataset**

**12.Program**

**13.Conclusion**

**1.INTRODUCTION:**

In today's fast-paced business environment, staying ahead of company registration trends is crucial for both government entities and businesses. The Registrar of Companies (RoC) plays a pivotal role in this, serving as the custodian of vital data related to company registrations. Leveraging the power of artificial intelligence and machine learning, this project aims to revolutionize the way we explore and predict company registration trends.

By building a robust machine learning model, we can analyze historical registration data, identifying patterns, and uncovering valuable insights. This not only assists RoC in making informed decisions but also provides businesses with a strategic advantage in understanding market dynamics and competition.

**2.OBJECTIVE:**

To develop a machine learning model that leverages historical data from the Registrar of Companies to analyze and predict company registration trends. The primary goals are to identify patterns, insights, and factors influencing registration fluctuations, enabling proactive decision-making for regulatory and business stakeholders.

**3.METHODOLOGY:**

**3.1Data Collection:**

Gather historical company registration data from RoC, including details about new registrations, closures, and various attributes such as industry, location, and size.

**3.2Data Preprocessing:**

Clean the data by handling missing values, outliers, and ensuring consistency in formatting.

**3.3Data Splitting:**

Split the data into training and testing sets. For time series data, consider using techniques like time-based splitting to preserve the temporal structure.

**4.MACHINE LEARNING APPROACH:**

**4.1Feature Engineering:**

Create relevant features that can help your model make predictions. This could include time series features, industry-specific variables, and geographical data.

## **Importance of feature engineering:**

Feature engineering is important in traditional [machine learning](https://www.turing.com/kb/learn-machine-learning-the-self-starter-way) concepts. The following are the importance of feature engineering:

1. **Enhanced model performance with well-engineered features:**

When feature engineering techniques are carried out on features in a dataset, machine learning models are provided with reliable data that enables them to provide better accuracy and results.

1. **Improved data representation and pattern extraction:**

Properly engineered or transformed features provide reliable and detailed insights into data. This also aids data scientists or analysts in drawing out valuable conclusions from it.

**3. Dimensionality reduction and prevention of overfitting:**

 Dimensionality reduction involves removing or filtering un useful or irrelevant features which in turn will yield better model performance, especially in high dimension data. Dimensionality reduction reduces the chance of model overfitting.

**4. Handling missing data effectively:**

Feature engineering involves methods in which missing data are handled without harming model performance.

**5. Incorporating domain knowledge into the model:**

Applying feature engineering techniques allows us to include domain knowledge by selecting useful features and removing irrelevant features in the dataset before training in the machine learning model.

**4.2Model Selection:**

Choose appropriate machine learning algorithms for your problem. Time series forecasting models like ARIMA or machine learning models like decision trees, random forests, or neural networks might be suitable.

**4.3Model Training:**

Train your chosen model on the training data. Optimize hyper parameters to improve performance.

**5.Evaluation:**

Assess the model's performance using appropriate metrics (e.g., Mean Absolute Error, Root Mean Squared Error, or accuracy for classification tasks).

**6.Prediction:**

Use the trained model to make predictions for future company registration trends. You can use this to identify potential hotspots of registrations or predict when registration numbers might increase or decrease.

**7.Monitoring and Updating:**

Continuously monitor the model's performance and retrain it as new data becomes available. This ensures that it remains accurate and relevant.

**8.Interpretability:**

Ensure that the model's predictions are interpretable and can be explained to stakeholders. This might involve using techniques like feature importance analysis.

**9.Deployment:**

Implement the model into a production environment where it can be used for ongoing predictions. This could be a web application or integrated into RoC's systems.

**10.Compliance:**

Ensure that your project complies with data privacy regulations and that the data used is obtained and handled ethically.

**11.DATASET:**

* The data set is the important element that contributes to the accuracy of performing data analysis.
* The data set considered was an ROC dataset that contained

1. CORPARATE\_ID

2.COMPANY\_NAME

3.COMPANY\_\_STATUS

4.COMPANY\_CLASS

5.COMPANY\_CATE GROYS

6.COMPANY\_SUB\_CATEGORY

7.DATE\_OF\_REGISTRATION

8.REGISTERED\_STATE

9.AUTHORIZED\_CAP

10.PAIDUP\_CAPITAL

11.INDUSTRIAL\_CLASS

12.PRINCIPAL\_BUSINESS\_ACTIVITY\_AS\_PER\_CIN

13.REGISTERED\_OFFICE\_ADDRESS

14.REGISTERAR\_OF\_COMPANIES

15.EMAIL\_ADDRESS

16.LATEST\_YEAR\_ANNUAL\_RETURN

17.LATEST\_YEAR\_FINANCIAL\_STATEMENT

**12.PROGRAM:**

**A program for analyzing a Register of Companies(ROC) using machine learning algorithm in python, following steps that provides analyzing dataset using pandas and numpy libraries.**

**1.Import the libraries:**

import numpy as np # linear algebra

import pandas as pd # data processing

#Input data files are available in the read-only “../input/”directory

IMPORT OS

## for dirname, \_, filenames in os.walk('/DATASET/input'):

## for filename in filenames

## print(os.path.join(dirname, filename)) **2.Loading the data:**

**[In]:**

dataset=pd.read\_csv("/dataset/input/Data\_Gov\_Tamil\_nadu/registered\_companies.csv")

print(dataset.columns)

 **[Out]:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CORPORATE  \_IDENTIFICATION  \_NUMBER | COMPANY  \_NAME | COMPANY  \_CLASS | DATE\_OF  \_REGISTRATION | REGISTERED  \_STATE | PRINCIPAL  \_BUSINESS  \_ACTIVITY  \_AS\_PER\_CIN | REGISTERED  \_OFFICE  \_ADDRESS | REGISTRAR  \_OF  \_COMPANIES | EMAIL\_ADDR | LATEST\_YEAR\_ANNUAL\_RETURN |
| |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  |   F00643 | HOCHTIEFF AG, | NA | 01-12-1961 | Tamil Nadu | Agriculture & allied | AMBLE SIDE, NO.8(OLD NO.30),3RD FLOOR KHADER NAWAZ KHAN ROAD,NUGAMBA | ROC DELHI | NA | NA |
| F00721 | SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LIMITED) | NA | NA | TamilNadu | Agriculture & allied | FLAT NO. 6, 1st FLOOR, 113/113ARAMA NAICKEN STREET, NUNGAMBAKKAM | ROC DELHI | shuchi.chug@asa.in | NA |
| F00892 | SRILANKAN AIRLINES LIMITED | NA | 01-03-1982 | Tamil  Nadu | Agriculture & allied | SRILANKAN AIRLINES LIMITED, VIJAYA TOWERSNO-4, KODAMBAKKAM HIGH ROAD, NUNGAMBAKKA | ROC DELHI | shree16us@yahoo.com | NA |
| **…..** | **…** | **…** | **….** | **…** | **…** | **…** | **…** | **…** | **…** |
| U74997TZ2016PTC027802 | POLYGAR FARM SOLUTIONS PRIVATE LIMITED | Private | NA | Tamil Nadu | Real estate renting and business activities | Real estate renting and business activities | ROC COIMBATORE  ROC COIMBATORE | [prashanthramana@gmail.com](mailto:prashanthramana@gmail.com)  sathishpandiya@gmail.com | NA |
| U74997TZ2018PTC030177 | PANDIYA AGRI SOLUTIONS PRIVATE LIMITED | Private | NA | Tamil Nadu | Real estate renting and business activities | 10/10 C3, Venkatasamy StreetRailyway Mens Colony, Kavundampalayam | ROC COIMBATORE | nroottechnologies@gmail.com | NA |
| U74997TZ2019PTC032491 | NROOT TECHNOLOGIES PRIVATE LIMITED | Private | NA | Tamil Nadu | Real estate renting and business activities | 139/1BPUDHUKOTTAI ROAD, MAPILLAI NAYAKKANPATTI |  |  | NA |

**150872 rowsx10columns**

**3.Clustering for dataset exploration:**

**[In1]:**

from sklearn.datasets import load\_iris

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

**[In2]:**

iris=load\_iris()

samples=iris.data

model=KMeans(n\_clusters=3)

model.fit(samples)

labels=model.predict(samples)

print(labels)

**[Out]:**

**[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0**

**0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1**

**1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2**

**2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2**

**2 1]**

**3.Plots:**

**[In1]:**

# plotting sepal length and petal length

xs=samples[:,0]

ys=samples[:,2]

plt.scatter(xs,ys,c=labels)

centroids=model.cluster\_centers\_

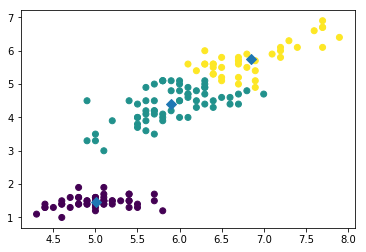
centroids\_x=centroids[:,0]

centroids\_y=centroids[:,2]

plt.scatter(centroids\_x,centroids\_y,marker='D',s=50)

plt.show()

**[Out1]:**

****

**[In2]:**

ks=range(1,6)

inertias=[]

for k in ks:

model=KMeans(n\_clusters=k)

model.fit(seeds)

nertias.append(model.inertia\_)

plt.plot(ks,inertias,'-o')

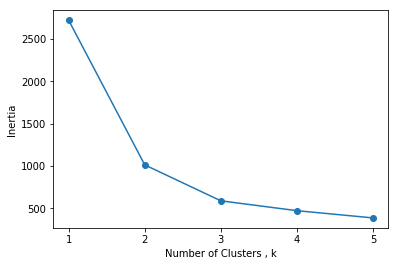
plt.xlabel('Number of Clusters , k')

plt.ylabel('Inertia')

plt.xticks(ks)

plt.show()

**[Out2]:**

****

**[In3]:**

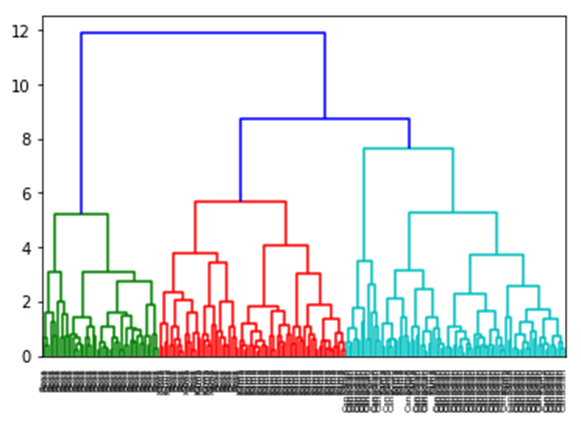
margings=linkage(company\_ class, method='complete')

#plt.figure(figsize=(50,50))

dendrogram(margings, labels= company\_ class \_list, company\_ rotation=90,company\_font\_size=6)

plt.show()

**[Out3]**

****

**13.Conclusion:**

In conclusion, AI-driven exploration and prediction of company registration trends with the Registrar of Companies (RoC) is a powerful and forward-thinking approach to understanding and forecasting business registration activities. Building a model using machine learning algorithms has the potential to provide valuable insights for various stakeholders, including government agencies, businesses, and investors. By analyzing historical data and employing predictive models, it becomes possible to anticipate future registration trends, identify economic shifts, and make informed decisions. However, it is essential to continually update and refine the model to adapt to changing market conditions and regulatory environments. Overall, the integration of AI and machine learning in the realm of company registration trends holds great promise for enhancing economic analysis and strategic planning.