AI – DRIVEN EXPLORATION AND PREDICTION OF COMPANY REGISTRATION TRENDS WITH REGISTER OF COMPANIES(ROC)

TEAM MEMBER

711121106046: LOGU PRAKASH V

Phase 5 submission document

Project Title: Exploration and prediction of ROC

Phase 5: Project Documentation & Submission

Topic: In this section we will document the complete

Project and prepare it for submission.

Introduction:

* Design: Utilize AI to forecast future company registrations using historical data
* Applicability: Informs business decisions and guides policy making
* Technology:AI algorithms for trend prediction and pattern recognition
* Coding:Python,AI libraries ,data manipulation,and predictive modeling
* Architecture:An AI business astrologer predicting company registration trends
* Transformation:Transforms decision making from reactive to proactive
* Real world analogy:A financial analyst predicting market trends for investors

Data collection: Use publicly accessible data sources or gain access to RoC databases. The types, locations, and registration dates of recently registered businesses should all be included in this data.

Data Preprocessing: Make sure the data is accurate and consistent by cleaning and preprocessing it. This could entail eliminating duplication, standardising data formats, and addressing missing values.

Feature extraction is the process of locating pertinent variables or features that support trend analysis. These could include registration kinds, geographic locations, and industry sectors.

Time Series Analysis: Track changes in company registrations over time by using AI techniques such as time series analysis. This may show long-term growth patterns, yearly variations, or seasonal trends.

Use natural language processing (NLP) to glean insightful information from registration documents' textual content.

Dataset Link: https://tn.data.gov.in/resource/company-master-data-tamil-nadu-upto-28th-february-2019

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CORPORATE\_IDENTIFICATION\_NUMBER | COMPANY\_NAME | STATUS | CLASS | CATEGORY | SUB\_CATEGORY | REGISTRATION | REGISTERED STATE | AUTHORIZED CAP | PAIDUP CAPITAL | INDUSTRIAL CLASS |
| F00643 | HOCHTIEFF AG, | NAEF | NA | NA | NA | 01-12-1961 | Tamil Nadu | 0 | 0 | NA |
| F00721 | SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LIMITED) | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F00892 | SRILANKAN AIRLINES LIMITED | ACTV | NA | NA | NA | 01-03-1982 | Tamil Nadu | 0 | 0 | NA |
| F01208 | CALTEX INDIA LIMITED | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01218 | GE HEALTHCARE BIO-SCIENCES LIMITED | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01265 | CAIRN ENERGY INDIA PTY. LIMITED | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01269 | TORIELLI S.R.L | ACTV | NA | NA | NA | 05-09-1995 | Tamil Nadu | 0 | 0 | NA |
| F01311 | HARDY EXPLORATION & PRODUCTION (INDIA) INC.. | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01314 | HOCHTIOF AKTIENGESELLSHARFF VORM GFBR HELFMANN | ACTV | NA | NA | NA | 11-04-1996 | Tamil Nadu | 0 | 0 | NA |
| F01412 | EPSON SINGAPORE PVT LTD | ACTV | NA | NA | NA | 25-04-1997 | Tamil Nadu | 0 | 0 | NA |
| F01426 | CARGOLUX AIRLINES INTERNATIONAL S A | ACTV | NA | NA | NA | 11-06-1997 | Tamil Nadu | 0 | 0 | NA |
| F01468 | CHO HEUNG ELECTRIC INDUSTRIAL COMPANY LIMITED | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01543 | NYCOMED ASIA PACIFIC PTE LIMITED | ACTV | NA | NA | NA | 27-10-1998 | Tamil Nadu | 0 | 0 | NA |
| F01544 | CHERRINGTON ASIA LTD | ACTV | NA | NA | NA | 01-05-2000 | Tamil Nadu | 0 | 0 | NA |
| F01563 | SHIMADZU ASIA PACIFIC PTE LIMITED | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01565 | CORK INTERNATIONAL PTY LIMITED | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01566 | ERBIS ENGG COMPANY LIMITED | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01589 | RALF SCHNEIDER HOLDING GMBH | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01593 | MITRAJAYA TRADING PRIVATE LIMITED | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01618 | HEAT AND CONTROL PTY LIMITED | ACTV | NA | NA | NA | 13-07-1999 | Tamil Nadu | 0 | 0 | NA |
| F01628 | DIREX SYSTEMS LIMITED | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01641 | NMB-MINEBEA THAI LIMITED | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01643 | ARROW INTERNATIONAL INC | ACTV | NA | NA | NA | 02-11-1999 | Tamil Nadu | 0 | 0 | NA |
| F01694 | GAMBRO CHINA LTD | ACTV | NA | NA | NA | 14-06-2000 | Tamil Nadu | 0 | 0 | NA |
| F01703 | OBARA CORPORATION | NAEF | NA | NA | NA | 17-07-2000 | Tamil Nadu | 0 | 0 | NA |
| F01752 | CIPTA WAWASON MAJU ENGINEERING SDM BHD | ACTV | NA | NA | NA | 24-01-2001 | Tamil Nadu | 0 | 0 | NA |
| F01753 | AUCHAN INTERNATIONAL S.A. | ACTV | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |
| F01767 | TOSHIBA PLANT SYSTEMS AND SERVICES CORPORATION | NAEF | NA | NA | NA | 08-03-2001 | Tamil Nadu | 0 | 0 | NA |
| F01768 | YAMAZEN CORPORATION | NAEF | NA | NA | NA | NA | Tamil Nadu | 0 | 0 | NA |

\*procedure:

1. Language of Programming: - Python's large library and frameworks make it the most widely used language for machine learning. Libraries like pandas, scikit-learn, NumPy, and others are available for use.

2. Integrated Development Environment (IDE): Select an IDE for coding and conducting experiments related to machine learning. Jupyter Notebook, Google Colab, and conventional IDEs like PyCharm are a few well-liked choices.

3. Machine Learning Libraries: - A number of machine learning libraries are required, such as:

- scikit-learn for creating and assessing models for machine learning.

- PyTorch or TensorFlow for deep learning, as required.

- For gradient boosting models, use CatBoost, LightGBM, or XGBoost.

4. Data Visualisation Tools: - For data exploration and visualisation, programmes like Matplotlib, Seaborn, or Plotly are indispensable.

5. Tools for Preprocessing Data: - Preprocessing, cleaning, and manipulation of data are made easier with libraries like pandas.

6. Data Collection and Storage: - Depending on your data source, you may need databases (like SQLite or PostgreSQL) or web scraping tools (like BeautifulSoup or Scrapy) for data storage.

7. Version Control: - Git and other version control systems are useful for tracking changes made to your code and facilitating teamwork.

8. Notebooks and Documentation: - Resources for recording your work, like Markdown for generating README files and documentation or Jupyter Notebooks.

9. Hyperparameter Tuning: Tools from scikit-learn such as GridSearchCV or RandomizedSearchCV can be used to help with hyperparameter tuning.

10. Web Development Tools (for Deployment): - If you want to build a web application for model deployment, you need be familiar with web development tools like HTML, CSS, and JavaScript for the front end and Flask or Django for the back end.

\*DESIGN THINKING PROCESS:

1. Empathise:

Recognise the demands and difficulties faced by each party engaged in the process of predicting the price of a home, including investors, real estate agents, appraisers, sellers, and buyers.

• Use surveys and interviews to learn more about the information users find most important for making decisions and what factors they value in property assessment.

2. Defining the problem:

Using machine learning, how may we be able to anticipate housing values more openly and accurately?

Determine the primary objectives and success standards for the project, such as raising user confidence in the valuation

process, decreasing bias, or enhancing prediction accuracy.

3. Ideate:

Come up with innovative ideas and data sources to improve the precision and openness of housing price forecasts.

To develop a wide range of ideas, such as the use of alternative data, new algorithms, or enhanced visualisation approaches, foster interdisciplinary collaboration.

4.Prototype:

Using the concepts from the ideation stage, develop machine learning models in prototype form.

• Put these prototypes through testing and iterations to identify the most accurate and user-friendly methods.

5.Test:

Using actual data and scenarios, test the machine learning models to get input from users and stakeholders.

• Evaluate the models' compliance with the established objectives and success criteria and make necessary modifications in response to user input.

6.Implement:

• Create a machine learning solution that is ready for production to estimate house values by combining the most effective algorithms and data sources.

• Use model interpretability tools and other transparency techniques to make sure people know how predictions are made.

7.Evaluate:

After the machine learning model is put into practise, keep an eye on its performance to make sure it continues to be accurate and applicable in a real estate market that is constantly changing.

• Get user input and insights to determine what needs to be improved.

\*DESIGN INTO INNOVATION

1. Data Collection:

Gather a comprehensive dataset that includes features such as

location, size, age, amenities, nearby schools, crime rates, and other

relevant variables.

2.Data Preprocessing:

Clean the data by handling missing values, outliers, and

encoding categorical variables. Standardize or normalize numerical

features as necessary.

**IMPORTING LIBRARIES:**

Importing necessary libraries to do basic things on the Dataset

**Program:**

import pandas as pd

import numpy as np

import seaborne as sns

import matplotlib.pyplot as plt

print(‘Successfully imported’)

**Output :**

Successfully imported

**Program :**

frame=pd.concat([data.isnull().sum() , data.nunique() , data.dtypes] , axis=1, sort=False)

frame

**VISUALIZATION AND PRE-PROCESSING OF DATA:**

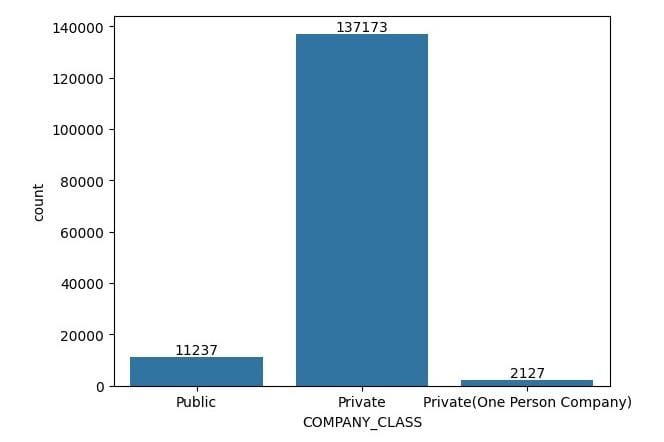
In[1]

ax = sns.countplot(data = data , X = ‘COMPANY\_CLASS’);

#Add the value of each parameters on the plot

ax.bar\_label(ax.containers[0]);

Out[1]



**Program :**

**In[1]**

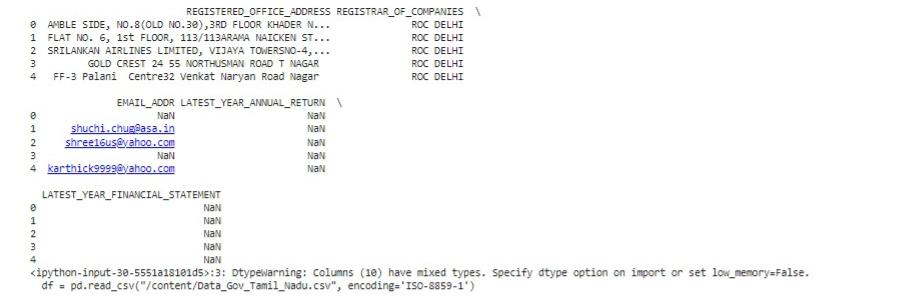
frame=pd.concat([data.isnull().sum() , data.nunique() , data.dtypes] , axis=1, sort=False)

frame

file\_path = 'Data\_Gov\_Tamil\_Nadu.csv'

df = pd.read\_csv("/content/Data\_Gov\_Tamil\_Nadu.csv", encoding='ISO-8859-1')

print(df.head())

**Out[1]:**

**In[2]**

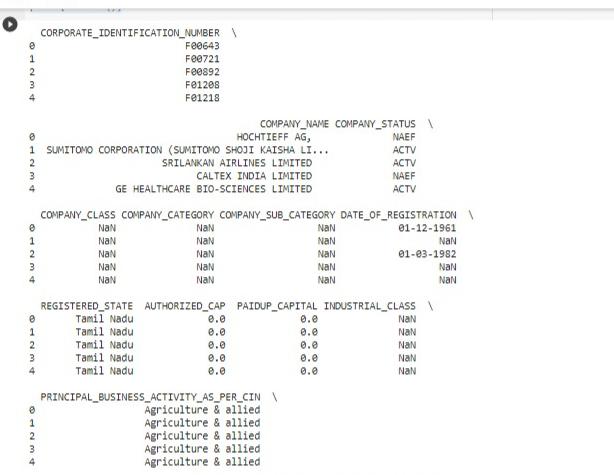
# Load the CSV file into a DataFrame

df = pd.read\_csv('Data\_Gov\_Tamil\_Nadu.csv',encoding='ISO-8859-1')

# Print the number of null values in first 10000 column

print(df.isnull().sum())

**Out[2]**

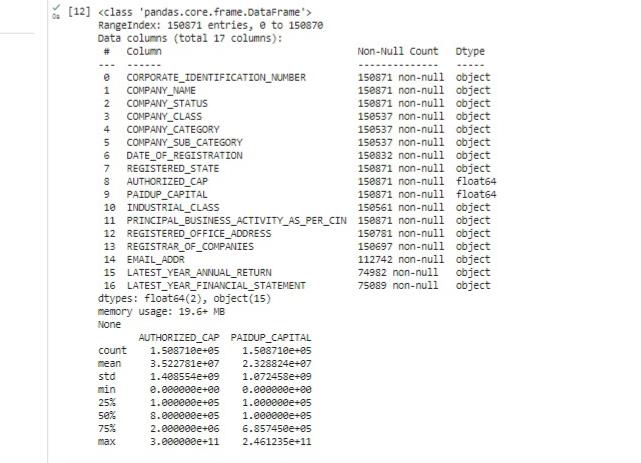
****

**In[3]**

print(df.info())

print(df.describe())

**Out[3]**

****

**VISUALIZATION AND PRE-PROCESSING OF DATA:**

# Print the columns of your dataframe to verify the presence of the 'INDEX ' column

print(df.columns)

# If the 'COMPANY\_STATUS ' column is present in your dataframe

if 'COMPANY\_STATUS ' in df.columns:

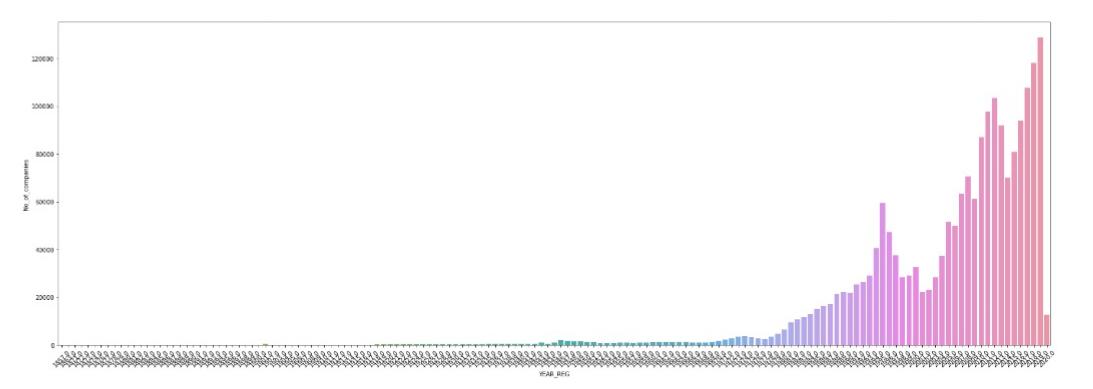
X = df.drop('COMPANY\_STATUS ', axis=1)

y = df['COUNTING']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

print(df.columns)

**Out**

 <matplotlib.axes.\_subplots.AxesSubplot at 0x7f2c8c5c6190>