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

**NAME: KARTHIKA A**

**REGISTER NUMBER: 61772131018**

## INTRODUCTION:

The AI-Driven Exploration and Prediction of Company Registration Trends with Registration of Companies (ROC) directs to capitalise artificial intelligence and data analytics to provide valuable insights into the registration of companies. By analysing historical data from the Registrar of Companies (ROC) and utilizing advanced machine learning algorithms, it seeks to understand, predict, and visualize trends in company registrations. The insights generated from this analysis will be valuable for government agencies, investors, entrepreneurs, and researchers.

## OBJECTIVES:

- ✓ Analyse historical ROC data to identify trends and patterns in company registrations and develop machine learning models for predicting future registration trends.

- ✓ Create a user-friendly web-based dashboard for visualizing and exploring registration data and provide actionable insights to help government agencies and businesses make informed decisions.

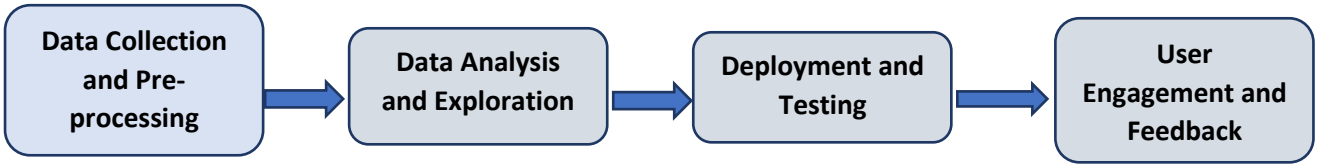
## Project Phases-

- ✓ Data Collection and Pre-processing:
  - Gather historical data from the ROC, including company registration dates, types of companies, industry classifications.
  - Clean, pre-process, and structure the data for analysis.

- ✓ Data Analysis and Exploration:
  - Utilize descriptive statistics, data visualization techniques, and exploratory data analysis to identify trends and patterns in company registrations.
  - Identify correlations between registration trends and external factors (e.g., economic indicators, policy changes).
  - Train machine learning models to predict future company registration trends based on historical data.

- ✓ Deployment and Testing:
  - Deploy the web-based dashboard on a secure and scalable platform.
  - Conduct extensive testing to ensure the system's reliability and performance.
  - Implement security measures to protect sensitive data.

- ✓ User Engagement and Feedback:
  - Promote the dashboard to government agencies, businesses, investors, and researchers.
  - Gather user feedback to continuously improve the system and add features based on user needs.




**Expected Outcomes:**

- ➢ Real-time insights into company registration trends.
- ➢ Predictive models for forecasting future trends.
- ➢ User-friendly web-based dashboard for data exploration.
- ➢ Informed decision-making for government and business stakeholders.

**Impact:**

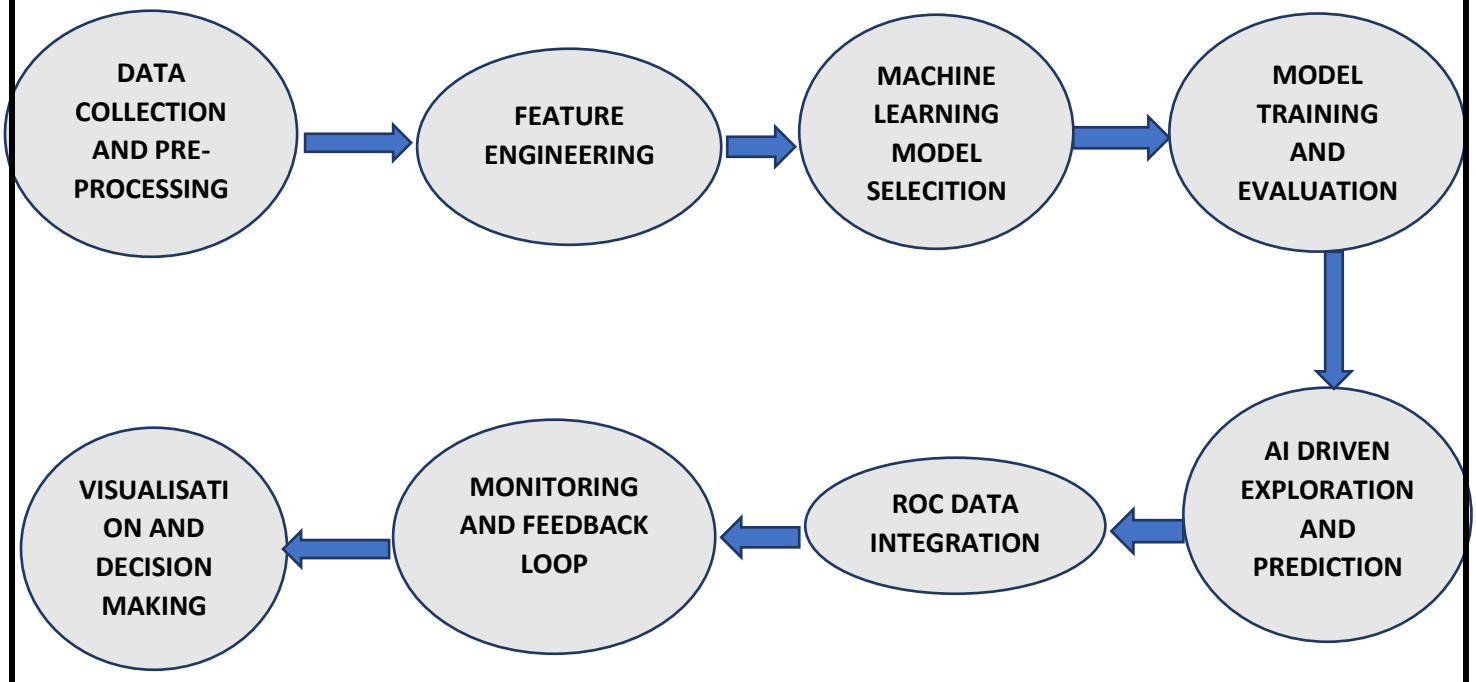
- ✓ Government agencies can use the insights to upgrade policies and regulations to support specific industries or regions.
- ✓ Investors can make data-driven decisions about where to allocate resources.
- ✓ Researchers can use the data for academic and market analysis.

**Ethical Considerations:**

- ✓ Ensure the privacy and security of sensitive registration data.
- ✓ Avoid bias in machine learning models and provide transparency in predictions.

**Future Prospects:**

- ➢ Expand the scope to include international registration data for broader insights.
- ➢ Incorporate natural language processing to analyse company registration documents for additional context.



## IMPORTING DATASET:

The given dataset is imported into jupyter notebook. The required modules are imported to perform the cleaning operation.

- The necessary libraries are imported by the following commands:

  pip install pandas

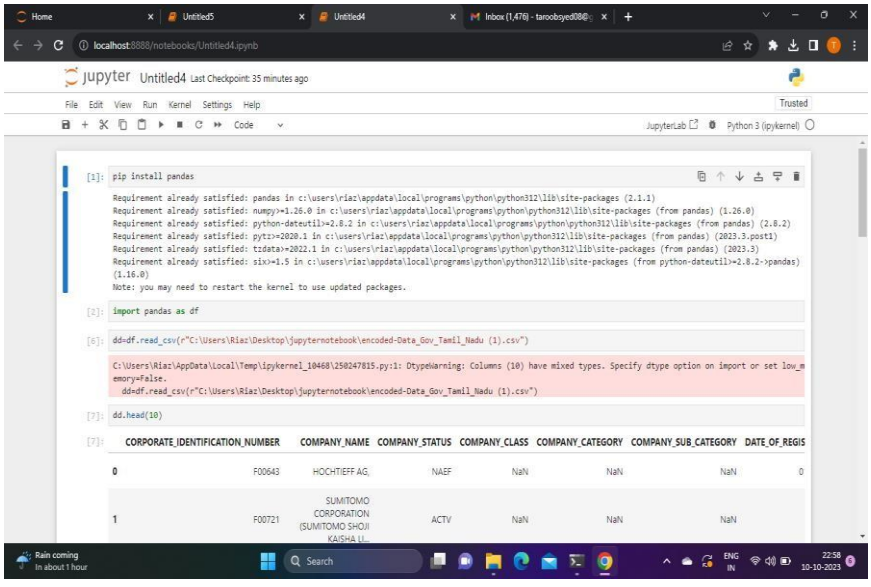
  import pandas as df

- The missing data and the duplicate data are handled by the following commands:

  dd.drop_duplicates()

  dd.dropna(inplace=True)

- Clean the pre-processed data which includes the renaming of columns.
- The cleaned dataset file is saved.



## DATA CLEANING:

Data cleaning means fixing bad data in the dataset. The bad data could be empty cells, data in wrong format and wrong data.

Data cleaning is a critical step in preparing data for the prediction of company registration using AI-driven exploration. Clean and well-structured data is essential for building accurate and reliable predictive models. Here are the steps to clean the data:

### Handling Missing Data:

- Missing data in the dataset is identified and handled. Missing data can significantly impact the quality of predictions.
- Options for handling missing data include:
- Rows with missing values are handled.
- Imputing missing values with the mean, median, or mode of the respective column.

### Handling Duplicates:

- Duplicate records are checked and removed, as duplicate entries can skew the analysis and modeling results.
- drop_duplicates method in pandas is used to remove duplicate rows.

### Data Transformation:

- Categorical variables are converted into numerical format using one-hot encoding or label encoding. This is necessary for many machine learning algorithms.
- Normalize or scale numerical features to ensure that they are on a common scale, especially if you plan to use algorithms sensitive to feature scaling.

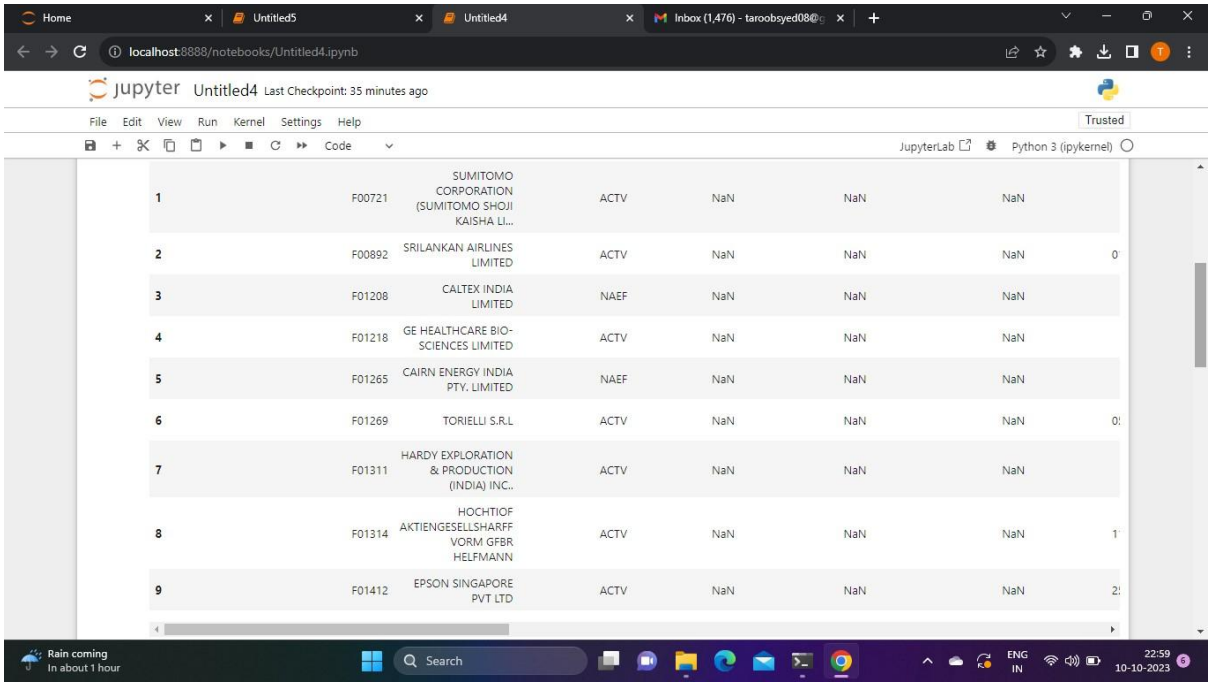
## Data Splitting:

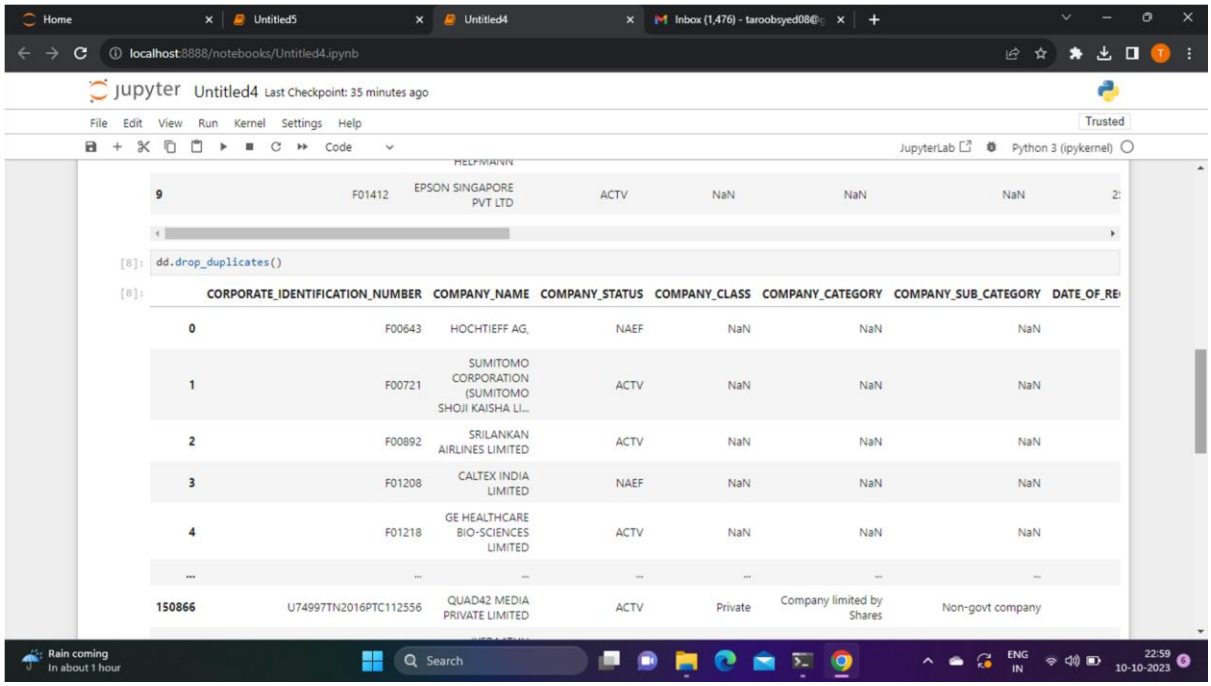
Dataset is split into training and testing sets for model evaluation. Typically, the training set is used to train the model, and the testing set is used to evaluate its performance.

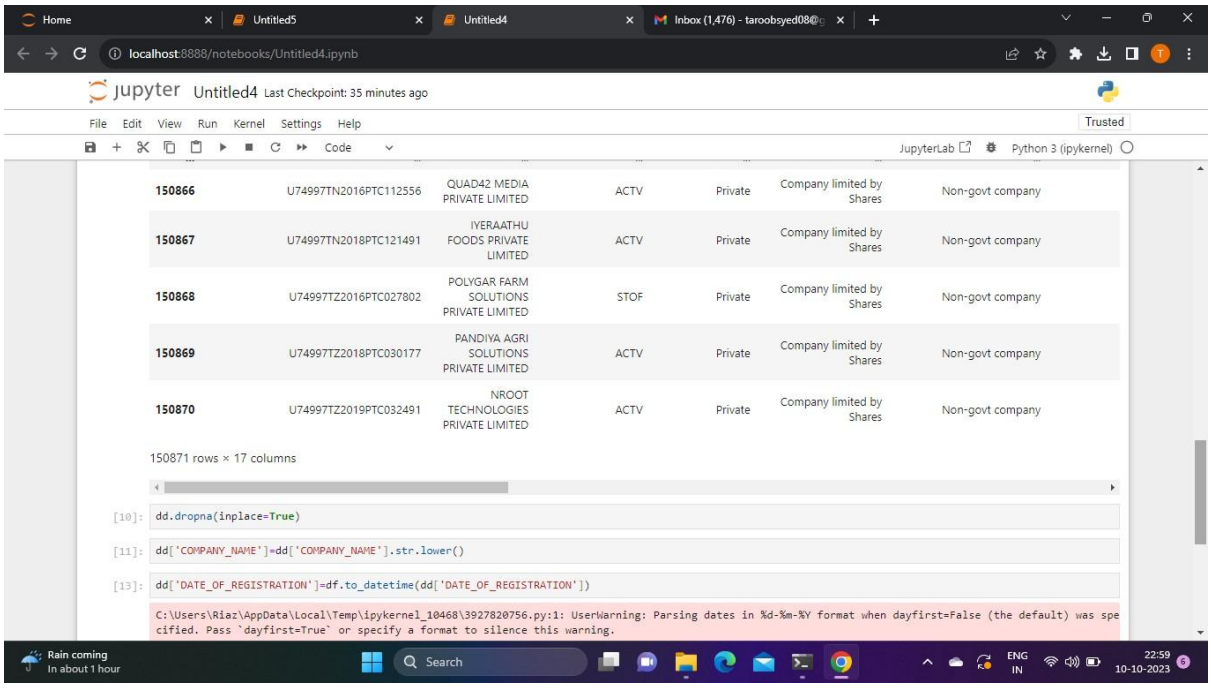
## Data Quality Checks:

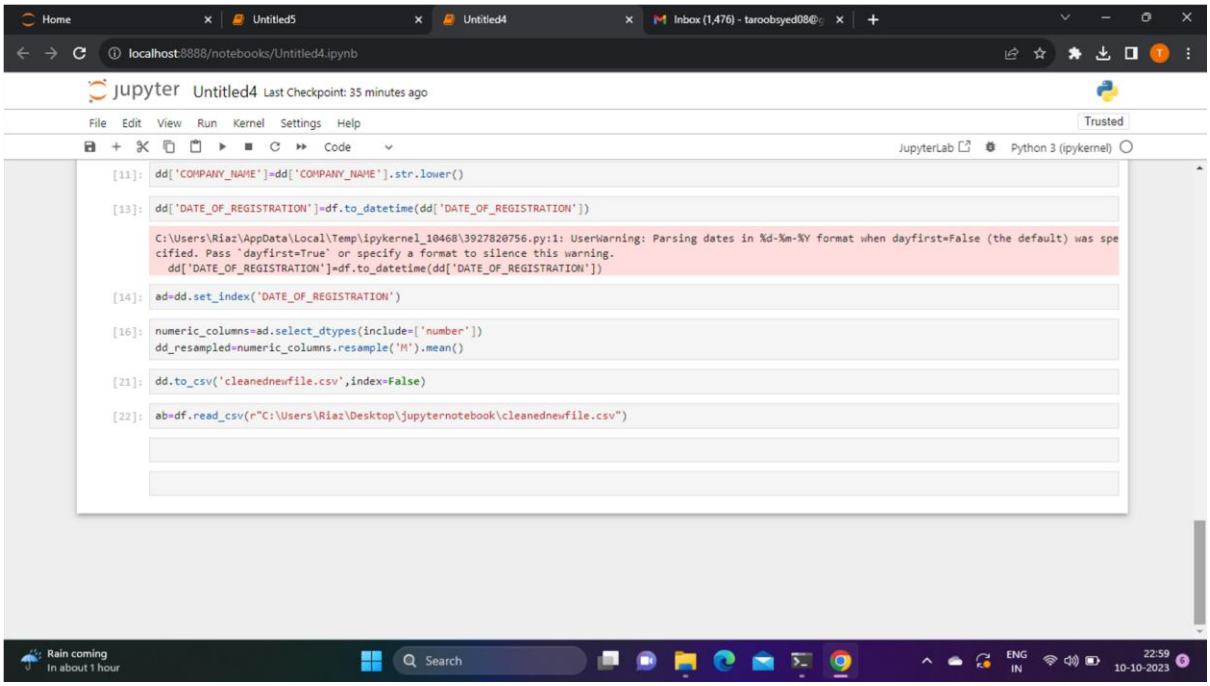
Any anomalies or data quality issues is checked. This includes verifying that data is within expected ranges and adheres to domain-specific rules.

The python code for importing the dataset and functions:



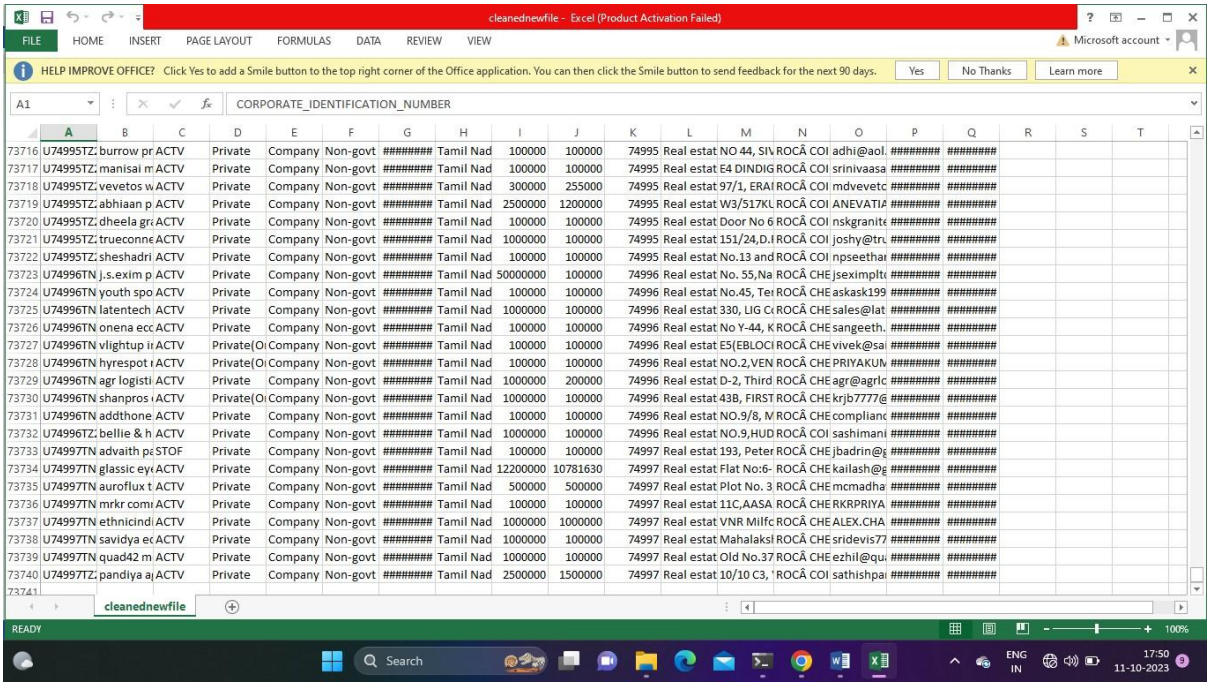

Home
Untitled5
Untitled4
Inbox (1,476) - taroobsyed08@g
localhost:8888/notebooks/Untitled4.ipynb
jupyter Untitled4 Last Checkpoint: 35 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
9 F01412 EPSON SINGAPORE PVT LTD ACTV NaN NaN NaN
[8]: dd.drop_duplicates()
[8]:
CORPORATE_IDENTIFICATION_NUMBER COMPANY_NAME COMPANY_STATUS COMPANY_CLASS COMPANY_CATEGORY COMPANY_SUB_CATEGORY DATE_OF_RE
0 F00643 HOCHTIEFF AG, NAEF NaN NaN NaN
1 F00721 SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI... ACTV NaN NaN NaN
2 F00892 SRILANKAN AIRLINES LIMITED ACTV NaN NaN NaN
3 F01208 CALTEX INDIA LIMITED NAEF NaN NaN NaN
4 F01218 GE HEALTHCARE BIO-SCIENCES LIMITED ACTV NaN NaN NaN
... ... ... ... ... ... ...
150866 U74997TN2016PTC112556 QUAD42 MEDIA PRIVATE LIMITED ACTV Private Company limited by Shares Non-govt company
Rain coming In about 1 hour
Search
ENG IN
22:59 10-10-2023


Home
Untitled5
Untitled4
Inbox (1,476) - taroobsyed08@g
localhost:8888/notebooks/Untitled4.ipynb
jupyter Untitled4 Last Checkpoint: 35 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
150866 U74997TN2016PTC112556 QUAD42 MEDIA PRIVATE LIMITED ACTV Private Company limited by Shares Non-govt company
150867 U74997TN2018PTC121491 IYERAATHU FOODS PRIVATE LIMITED ACTV Private Company limited by Shares Non-govt company
150868 U74997TZ2016PTC027802 POLYGAR FARM SOLUTIONS PRIVATE LIMITED STOF Private Company limited by Shares Non-govt company
150869 U74997TZ2018PTC030177 PANDIYA AGRI SOLUTIONS PRIVATE LIMITED ACTV Private Company limited by Shares Non-govt company
150870 U74997TZ2019PTC032491 NROOT TECHNOLOGIES PRIVATE LIMITED ACTV Private Company limited by Shares Non-govt company
150871 rows × 17 columns
[10]: dd.dropna(inplace=True)
[11]: dd['COMPANY_NAME']=dd['COMPANY_NAME'].str.lower()
[13]: dd['DATE_OF_REGISTRATION']=df.to_datetime(dd['DATE_OF_REGISTRATION'])
C:\Users\Riaz\AppData\Local\Temp\ipykernel_10468\3927820756.py:1: UserWarning: Parsing dates in %d-%m-%Y format when dayfirst=False (the default) was spe
cified. Pass `dayfirst=True` or specify a format to silence this warning.
Rain coming In about 1 hour
Search
ENG IN
22:59 10-10-2023




## OUTPUT AFTER CLEANING:

After the data cleaning is done, the number of cells in the Excel sheet in the final dataset is reduced to 73741 by removing the duplicate data, wrong data.



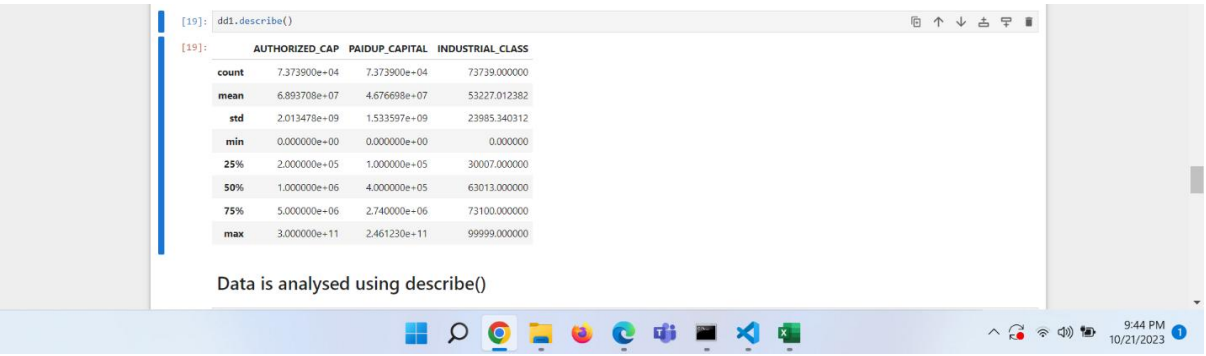
## DATA ANALYSIS:

Data analysis is a systematic approach which follows the process of inspecting, cleaning, transforming and interpreting data to extract valuable insights.

Data analysis is done using Python- jupyter notebook.

It includes various various methods:

- Data Collection and Pre-processing.
- Conduct EDA to gain insights into the dataset.
- Feature and Model Selection.
- Data Splitting
- Model training and evaluation.
- Deployment and Monitoring.

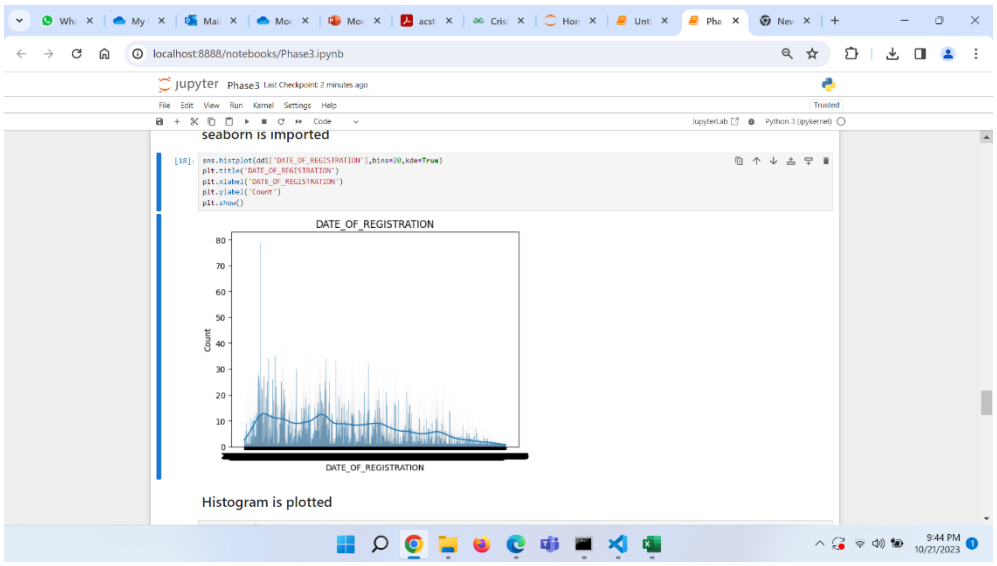



## DATA VISUALIZATION:

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations.

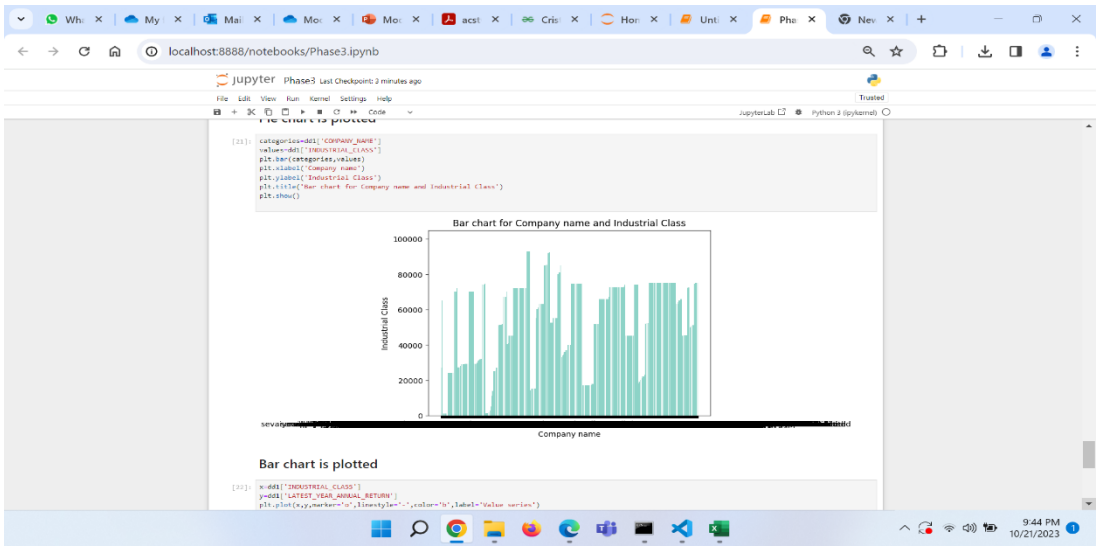
In python, it is done using matplotlib library that is imported to visualize different charts.

## DATA VISUALIZATION USING HISTOGRAM:



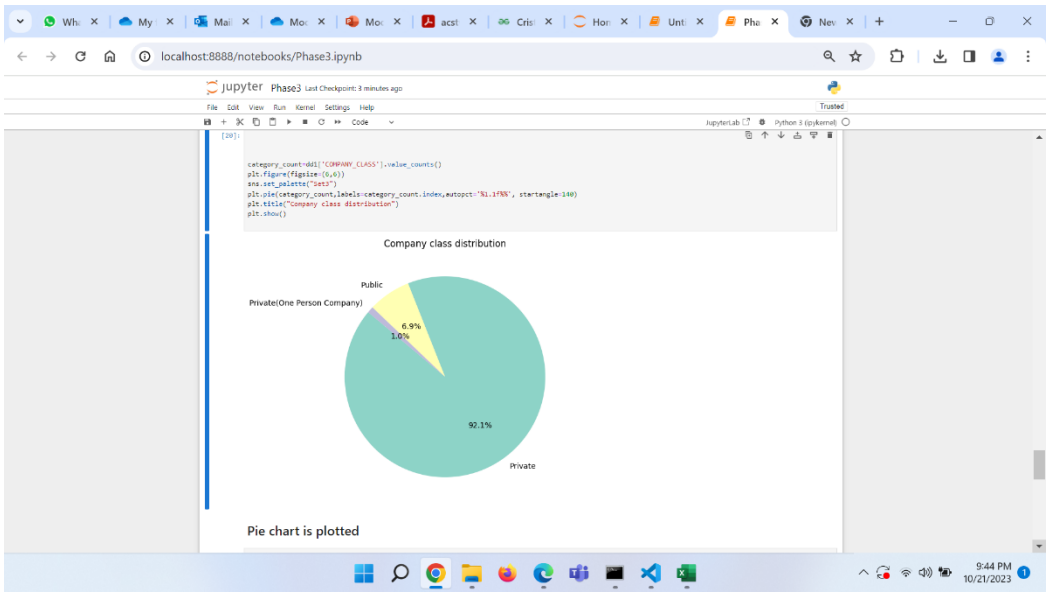

Histogram is plotted for the “DATE_OF_REGISTRATION” and “COUNT” in x and y axes respectively.

## DATA VISUALIZATION USING BARCHART:



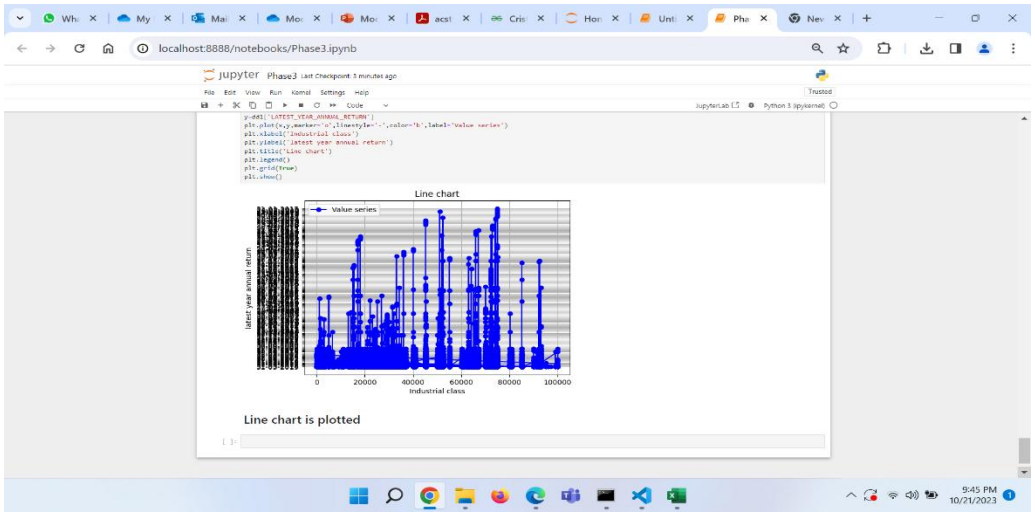

Bar chart is plotted for “COMPANY_NAME” and “INDUSTRIAL_CLASS” in x and y axes respectively.

## DATA VISUALIZATION USING PIE CHART:




Data visualization is done using pie chart for “COMPANY_CLASS_DISTRIBUTION”.

## DATA VISUALIZATION USING LINECHART:




Data visualization is done using line chart for “INDUSTRIAL_CLASS” and “LATEST_YEAR_ANNUAL_RETURN”.

**MODEL:**

For the model development and evaluation, the **RANDOM FOREST CLASSIFIER** is used which is a Machine Learning model.

This model combines multiple decision trees to make predictions.

It contains a number of decision trees on various subset of the given dataset and takes the average to improve the predictive accuracy of the dataset.

This model is known for its robustness , ability to handle high dimensional data and resistance to over-fitting.

This model takes less training time when compared to other algorithms.

It predicts the output with high accuracy for larger datasets too.

It can perform both classification and regression.

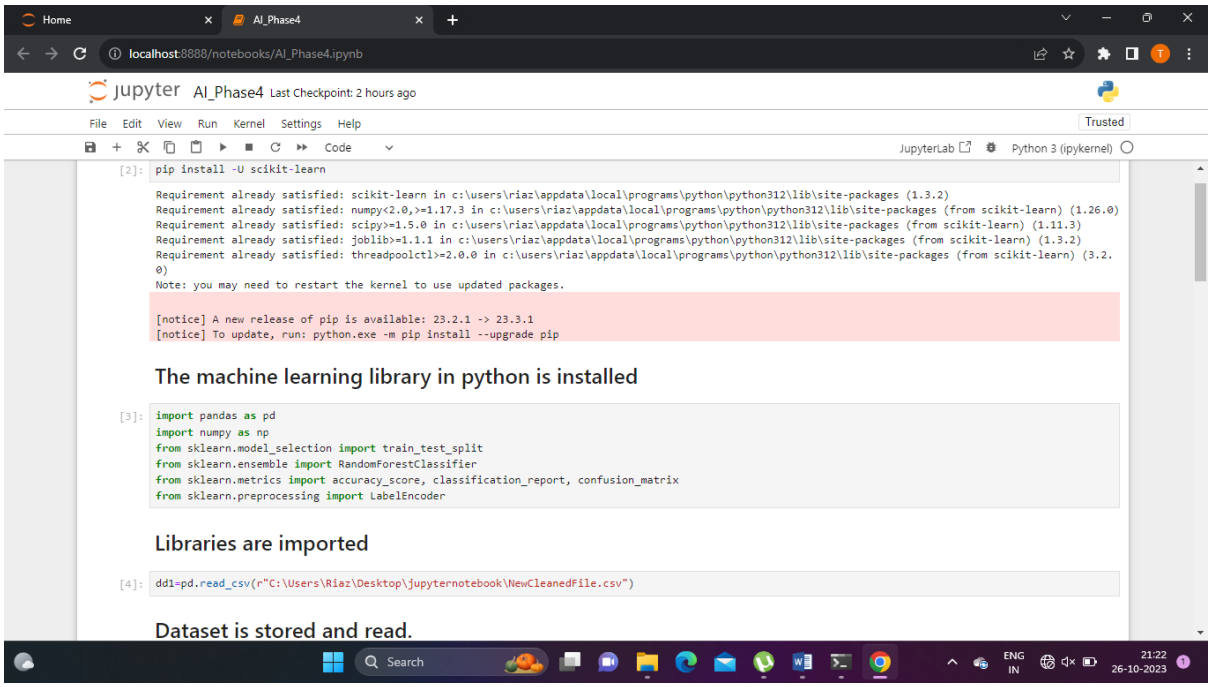
It can maintain accuracy though larger proportion of data is missing.

This model run in two phases:

- Create random forest by combining N decision trees.
- Make predictions for each tree created in the previous step.

First, the machine learning library in Python is installed and the required libraries are imported in the jupyter notebook.

The data set is read as a CSV file and stored in the jupyter notebook.



```
[11]: rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
```

Initializing RandomForestClassifier

The Random Forest Classifier is initialized.

**SPLITTING AND TRAINING:**

The input feature and target variable is chosen from the dataset and stored in X and y respectively.

The data is split into two parts:

- Training set
- Testing set

```
[9]: X = dd1_encoded.drop(['COMPANY_CLASS'], axis=1)
     y = dd1_encoded['COMPANY_CLASS']
```

X contains the input feature and y contains the target variable

```
[10]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Splitting of data into training and testing sets

Training is done for the Random Forest Classifier using the training data.

```
[12]: rf_classifier.fit(X_train, y_train)
```

```
[12]: RandomForestClassifier
      RandomForestClassifier(random_state=42)
```

Trains the RandomForestClassifier

The training set is used to train the Random Forest model, while the testing set is used to evaluate the performance.

Once the model is trained, the testing set is used to evaluate the performance using various metrics like accuracy, precision, recall.

**EVALUATION:**

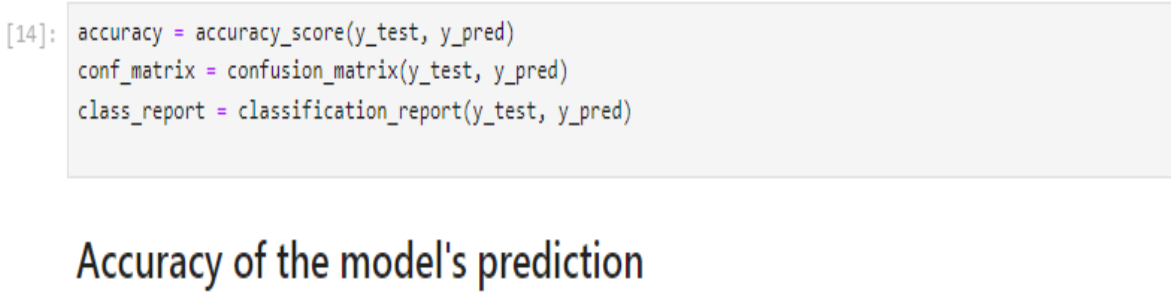
Evaluating the Random Forest model involves assessing its performance to understand how well it generalizes to unseen data.

**PREDICTION:**

Since the model is fitted into the training set, now we can predict the test result. For prediction, we create a new prediction vector ‘y_pred’.


**ACCURACY:**

It measures the proportion of correctly classified instances in the test set.




The accuracy can be assessed by using the above code.

In this code, the accuracy is calculated by using ‘accuracy_score’ by comparing the true target values ‘y_test’ with predicted values ‘y_pred’.

The result would be a decimal value between 0 and 1.

The confusion matrix and the classification report can also be assessed by using Random Forest Classifier Model.

Confusion matrix is used to determine the correct and incorrect predictions.

```
[15]: print(f'Accuracy: {accuracy}')
      print('Confusion Matrix:')
      print(conf_matrix)
      print('Classification Report:')
      print(class_report)
```

```
Accuracy: 0.9373474369406021
Confusion Matrix:
[[13436     0   121]
 [  110    45     0]
 [  693     0   343]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.99      0.97     13557
           1       1.00      0.29      0.45       155
           2       0.74      0.33      0.46      1036

    accuracy                           0.94     14748
   macro avg       0.89      0.54      0.62     14748
weighted avg       0.93      0.94      0.93     14748
```

Prints the accuracy,confusion matrix,classification report

```
[16]: accuracy_percentage = accuracy * 100

      print(f'Accuracy: {accuracy_percentage:.2f}%')
```

```
Accuracy: 93.73%
```

Calculates accuracy in percentage and prints it

The accuracy is printed in percentage.

Accuracy achieved is 93.73%.

**Conclusion:**

The AI-Driven Exploration and Prediction of Company Registration Trends with ROC project combines data analysis, machine learning, and user-friendly visualization to empower various stakeholders with valuable insights into company registration trends, ultimately contributing to informed decision-making and economic growth.