PHASE 1: PROBLEM DEFINITION AND DESIGN THINKING

PROBLEM DEFINITION

Core Problem:

To conduct AI-driven analysis on the details of companies registered with the ***RoC***

Key Objectives:

- We will unearth hidden patterns within the data.
- We will gain a comprehensive understanding of the landscape of registered companies.
- We will also predict future trends of company registrations.

Significance:

- It would help perform informed decision-making for a range of stakeholders, including businesses, investors, and policymakers.

Approach:

- We will use cutting-edge AI techniques to build the project.
- We will rely on the historical and current data of company registrations to perform the predictions.

Expected Outcome:

We hope to create predictive models that can accurately anticipate and predict the registration of new companies in the future.

DESIGN THINKING

Data Source:

We will use the dataset with TN company details which has info like the company's name, status, class, category, registration date, authorized capital, paid-up capital and more. This dataset is available in the following link: **https://tn.data.gov.in/resource/company-master-data-tamil-nadu-upto-28th-february-2019.** We will download the dataset from the website, and create a dataframe using python's pd.read_csv() function.

Data Preprocessing:

- First , we will go through the dataset to ensure it is optimized and organized for analysis.

- **Handling Missing values:** In the next step, we will handle missing values (NA) using methods like **Imputation** ,where these missing values are filled with estimated values such as mean or average of that column and **Deletion** ,where any rows and columns with excessive missing values will be deleted to maintain data Integrity.

- **Handling Categorical values**: After this, we will handle categorical data by converting them into numbers using methods like ***One-hot encoding*** ,wherein we will transform categories into binary columns in which each category gets its own column marked with 0 or 1 and ***label encoding*** ,where we will assign a unique number to each category ,turning them into numerical labels.

Exploratory Data Analysis (EDA):

- **Understanding the Distribution:** In this phase, first we will understand the distribution by looking at how the data is spread and see if there are any patterns. We will do this by using ***Historgams***, which will help us visualize the frequency of the data and ***Box plots,*** which will help us in spotting outliers and understanding the data.

- **Understanding relationships between features**: Next, we'll see how different company attributes relate to one another using methods like ***scatter plots*** and ***Correlation matrices***.

- **Examining Unique characters:** After this, we'll try and find traits that make some companies stand out. To do this, we'll use methods like ***Value Count ,*** which will help us tally unique entries for categorical data and ***Pie Charts ,*** which will help us visualize proportions of the categories present.

Feature Engineering:

In this phase, we'll be enhancing our dataset with new features to improve our predictions by transforming and combining the available data to extract maximum value. We hope to perform this by using the following methods:

- **Aggregation:** Based on existing features, statistics like mean, sum, and variance are computed to create new insights.

- **Interaction Features**: Two or more features are combined by multiplying or dividing them in order to uncover hidden relationships.

- **Temporal Features**: In this method, the columns with date and time will be broken down into day, month, year, or even time of day to analyze trends and seasonality.

Predictive Modelling:

In this phase, we'll choose AI techniques which will predict future company registrations and deploy models using them. Our main aim here is to perform Time Series analysis. Some algorithms and methods that we are planning to use to perform the same are mentioned below.

- ARIMA
- Prophet
- Long Short Term Memory
- GARCH
- Exponential Smoothing methods: A method in which past observations are weighted with a decay factor.
- State Space models and Kalman filters: These are models that consider both the observed data and the Underlying states.

Only those that gives us the highest accuracy will be selected as the final model.

Model Evaluation:

In this phase, we'll evaluate the models using metrics to test their strength and reliability. Some metrics that we plan to use in our project are mentioned below.

- **Accuracy:** This metric would help us understand the fraction of predictions the model got right.
- **Precision:** This metric focuses on the positive predictions our model makes.
- **F1-score:** It is the harmonic mean of precision and Recall. This will be used in case our data has uneven class distribution.
- **Mean Absolute Errors:** This will give us an average of the absolute errors between the predicted values and the actual values.
- **Mean Absolute Percentage Errors**: This will help us represent the average errors as percentage.

Using these metrics, we will be able to confidently gauge the strength and reliability of our model.

By following this plan, we aim to gain valuable insights and predict registration trends of new companies in the future.