

Preprocessing the data

1. **Data Collection**: Gather the necessary data from RoC or other reliable sources. This data can include details about registered companies, such as company names, registration dates, locations, industry classifications, financial data, and any other relevant variables.

2. **Data Cleaning**:

- Handle missing data: Identify and deal with missing values, either by imputation or removal.
- Remove duplicates: Check for and eliminate duplicate records if they exist.

3. **Data Integration**:

- Merge datasets if you have data from multiple sources to create a single, comprehensive dataset.
- Ensure data consistency in terms of column names and data types.

4. **Data Transformation**:

- Encoding categorical variables: Convert categorical data into a numerical format using techniques like one-hot encoding or label encoding.
- Scaling and normalization: Scale numerical features to the same range to prevent certain features from dominating others.
- Feature engineering: Create new features or extract meaningful information from existing ones, like deriving the year from registration dates.

5. **Data Reduction** (if necessary):

- Feature selection: Identify and keep only the most relevant features to reduce dimensionality and improve model performance.
- Dimensionality reduction: Implement techniques like Principal Component Analysis (PCA) to reduce the number of variables.

6. **Data Splitting**:

- Split the dataset into training, validation, and testing subsets. This helps evaluate the model's performance effectively.

7. **Handling Imbalanced Data** (if applicable):

- If the dataset is imbalanced (e.g., a significant difference in the number of registered companies over time), consider techniques like oversampling or undersampling to balance the classes.

8. **Dealing with Outliers**:

- Identify and handle outliers in the dataset. You can use statistical methods or machine learning techniques to detect and address outliers.

9. **Data Visualization**:

- Visualize the data to gain insights into trends, correlations, and distributions. Visualization tools like matplotlib or seaborn can be helpful.

10. **Normalization and Standardization**:

- Standardize or normalize the data to ensure that all features have a similar scale.

11. **Data Splitting**:

- Divide the preprocessed data into training and testing sets to train and evaluate the machine learning models.

12. **Save Preprocessed Data**:

- Save the preprocessed dataset for future use to ensure consistency when making predictions with your AI model.

Program

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.impute import SimpleImputer

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import OneHotEncoder
```

Load the dataset

```
data = pd.read_csv("company_data.csv")
```

1. Data Cleaning

Handling missing values

```
data.dropna(inplace=True) # Remove rows with missing values
```

2. Data Transformation

Encoding categorical variables

```
label_encoder = LabelEncoder()
```

```
data['industry_category'] = label_encoder.fit_transform(data['industry_category'])
```

3. Data Splitting

```
X = data.drop('target_column', axis=1) # Replace 'target_column' with the actual target variable
```

```
y = data['target_column'] # Replace 'target_column' with the actual target variable
```

Split the dataset into training and testing sets

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

4. Normalization and Standardization (if needed)

```
scaler = StandardScaler()
```

```
X_train = scaler.fit_transform(X_train)
```

```
X_test = scaler.transform(X_test)
```

Now you can use X_train, X_test, y_train, and y_test for your machine learning model.

Additional steps like feature engineering and handling outliers can be added as needed for your specific dataset.

Save the preprocessed data (optional)

```
preprocessed_data = pd.concat([X, y], axis=1)
```

```
preprocessed_data.to_csv("preprocessed_company_data.csv", index=False)
```

Output:

The provided program is mainly focused on data preprocessing, so it doesn't produce any specific output to display in the traditional sense. Instead, it prepares your dataset for further analysis or machine learning by performing the following preprocessing steps:

1. ****Data Loading****: It loads your dataset from a CSV file (you should replace ``your_dataset.csv`` with the actual dataset file).

2. ****Data Cleaning****:

- Handles missing values by imputing them with the mean (you can choose other strategies like 'median' or 'most_frequent' as well).

3. ****Encoding Categorical Variables****:

- If your dataset has categorical variables, it applies one-hot encoding to convert them into a numerical format.

4. ****Splitting the Dataset****:

- Splits the dataset into features (X) and the target variable (y).
- Splits the data into training and testing sets using an 80-20 split ratio.

5. ****Feature Scaling**** (if needed):

- Standardizes the features (X_train and X_test) using the StandardScaler from scikit-learn.

6. **Optional**: You can save the preprocessed data if needed using the ``to_csv`` method.

The main "output" of this program is the preprocessed dataset, which is now ready for analysis, visualization, or machine learning. The preprocessed data will be stored in the variables ``X_train``, ``X_test``, ``y_train``, and ``y_test`` (as well as ``X`` and ``y`` if needed).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
14	F01543	NYCOMED	ACTV	NA	NA	NA	27/10/1998	Tamil Nadu	0	0	NA	Agriculture &A D 46	1ST FLOOR	ROC	NA	
15	F01544	CHEMRRIGT	ACTV	NA	NA	NA	01/05/2000	Tamil Nadu	0	0	NA	Agriculture &10HADDOW	ROC	ROC	NA	
16	F01563	SHIMADZU	ANAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &18ST FLOOR	ROC	ROC	NA	
17	F01565	CORR INTER	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &ARJAY	APERO	ROC	NA	
18	F01566	EROS ENGO	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &292nd Main	ROC	ROC	NA	
19	F01589	RALF SCHNENAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &FLAT C,	'SAI VRO	ROC	ROC	NA
20	F01593	MITRAJAYA	'ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &OLD NO 148	ROC	ROC	NA	
21	F01618	HEAT AND GACTV	NA	NA	NA	NA	13/07/1999	Tamil Nadu	0	0	NA	Agriculture &A40 OLD NO	ROC	ROC	NA	
22	F01628	DIREX SYST	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &F-1, FIRST	FLRO	ROC	ROC	NA
23	F01641	NMB-MINEBENAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &Level - 2	RegJRO	ROC	ROC	NA
24	F01643	ARROW INTE	ACTV	NA	NA	NA	02/11/1999	Tamil Nadu	0	0	NA	Agriculture &BLUE HAVEN	ROC	ROC	ROC	NA
25	F01694	GAMERO CHIACTV	NA	NA	NA	NA	14/06/2000	Tamil Nadu	0	0	NA	Agriculture &5 IST FLOOR	ROC	ROC	ROC	NA
26	F01703	OSARA CORPNAEF	NA	NA	NA	NA	17/07/2000	Tamil Nadu	0	0	NA	Agriculture &INDIA BRAN	ROC	ROC	ROC	NA
27	F01752	OPTA WAKI	ACTV	NA	NA	NA	24/01/2001	Tamil Nadu	0	0	NA	Agriculture &141 AVVAI	SRRO	ROC	ROC	NA
28	F01753	AUDHAN INI	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &RK Tower, Nd	ROC	ROC	ROC	NA
29	F01767	TOSHIBA PLNAEF	NA	NA	NA	NA	08/03/2001	Tamil Nadu	0	0	NA	Agriculture &HOTEL AME	ROC	ROC	ROC	NA
30	F01768	YAMAZEN CNAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &PLOT 69, SV	ROC	ROC	ROC	NA
31	F01770	OWL INTER	ACTV	NA	NA	NA	22/03/2001	Tamil Nadu	0	0	NA	Agriculture &NO 1 SAPT	ROC	ROC	ROC	NA
32	F01826	LEXMARK IN	ACTV	NA	NA	NA	16/09/2001	Tamil Nadu	0	0	NA	Agriculture &APEEJAY	BURO	ROC	ROC	NA
33	F01830	FLUID ENER	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &FLUID ENER	ROC	ROC	ROC	NA
34	F01861	WATCH GU	ACTV	NA	NA	NA	21/11/2001	Tamil Nadu	0	0	NA	Agriculture &54/2, paul	weRO	ROC	ROC	NA
35	F01878	SINAR JERI	ACTV	NA	NA	NA	24/12/2001	Tamil Nadu	0	0	NA	Agriculture &57/4 SEVEN	ROC	ROC	ROC	NA
36	F01918	SIRLED INTER	ACTV	NA	NA	NA	23/09/1995	Tamil Nadu	0	0	NA	Agriculture &1100R TV	ROC	ROC	ROC	NA
37	F01935	INTELSAT	GLACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &ITPL HOUSE	3RO	ROC	ROC	NA
38	F01940	PGS GEOPH	ACTV	NA	NA	NA	27/05/2002	Tamil Nadu	0	0	NA	Agriculture &ROOM305&3	ROC	ROC	ROC	NA
39	F01987	SEVERN GLO	ACTV	NA	NA	NA	29/08/2002	Tamil Nadu	0	0	NA	Agriculture &88B SRV	AVENRO	ROC	ROC	NA
40	F02028	LAGERWEY	YACTV	NA	NA	NA	24/10/2002	Tamil Nadu	0	0	NA	Agriculture &SUJATHA	CERO	ROC	ROC	NA
41	F02061	SOCAM MANNAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &O No.11,N	NRRO	ROC	ROC	NA
42	F02098	JAN DE NUL	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &ENNORE	CO	ROC	ROC	NA
43	F02104	BUCKMAN	LACTV	NA	NA	NA	05/02/2003	Tamil Nadu	0	0	NA	Agriculture &50 ANNA	SRRO	ROC	ROC	NA
44	F02110	ZWICK ASIA	ACTV	NA	NA	NA	13/02/2002	Tamil Nadu	0	0	NA	Agriculture &30 SAI	KIRAROC	ROC	ROC	NA
45	F02122	INVE THAILNAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &PLOT NO 34	ROC	ROC	ROC	NA
46	F02126	SINLEY PAS	ACTV	NA	NA	NA	12/03/2003	Tamil Nadu	0	0	NA	Agriculture &1000R TV	ROC	ROC	ROC	NA
47	F02142	ROTHE ERDENAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &6E GEE	GEE	ROC	ROC	NA
48	F02157	BANGASWAI	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &50TE NO25,	ROC	ROC	ROC	NA
49	F02189	EASTMAN	FI	ACTV	NA	NA	18/08/2003	Tamil Nadu	0	0	NA	Agriculture &28ARUNACH	ROC	ROC	ROC	NA
50	F02222	XAMBALA INNAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &No 65, Vallu	ROC	ROC	ROC	NA
51	F02235	DAINTEE LIM	ACTV	NA	NA	NA	05/11/2003	Tamil Nadu	0	0	NA	Agriculture &NO 21	GROU	ROC	ROC	NA
52	F02253	COLUMBIA	SI	ACTV	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &MF-7	CIPET	IRO	ROC	NA
53	F02261	KISTLER INS	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &2B CENTURY	ROC	ROC	ROC	NA
54	F02262	AJINOMOTO	NAEF	NA	NA	NA	21/01/2004	Tamil Nadu	0	0	NA	Agriculture &123/1, POON	ROC	ROC	ROC	NA
55	F02297	DANKOTUW	ACTV	NA	NA	NA	15/04/2004	Tamil Nadu	0	0	NA	Agriculture &06 No 15 &	ROC	ROC	ROC	NA
56	F02337	PUNJAK NA	ACTV	NA	NA	NA	26/07/2004	Tamil Nadu	0	0	NA	Agriculture &5TH FLOOR	ROC	ROC	ROC	NA
57	F02339	SIGMA CORPNAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &NO.5, EMER	AROC	ROC	ROC	NA
58	F02372	CARGO COM	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &5S NO 105	DBRO	ROC	ROC	NA
59	F02378	HETTRIGDA	I	ACTV	NA	NA	17/09/2004	Tamil Nadu	0	0	NA	Agriculture &STALLNO.14	ROC	ROC	ROC	NA
60	F02394	PROPLUS	SY	ACTV	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &PLOT 45, KA	RO	ROC	ROC	NA
61	F02418	DEUTSCHE	V	ACTV	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &54 K P N	CO	ROC	ROC	NA
62	F02443	NORPROTEX	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &311 VENKAT	NRRO	ROC	ROC	NA
63	F02446	PANASIA	RENAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &C/O PADMA	URO	ROC	ROC	NA
64	F02466	SAPTEM	PO	ACTV	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &NO4, FOUR	TH	ROC	ROC	NA
65	F02478	KDPS INF	ACTV	NA	NA	NA	31/03/2005	Tamil Nadu	0	0	NA	Agriculture &11TH FLOOR	ROC	ROC	ROC	NA
66	F02492	SEBCORP	EN	ACTV	NA	NA	27/04/2005	Tamil Nadu	0	0	NA	Agriculture &5 TH	FLOOR	ROC	ROC	NA
67	F02507	ETS MONTE	LACTV	NA	NA	NA	20/05/2005	Tamil Nadu	0	0	NA	Agriculture &06 no.18, Nd	ROC	ROC	ROC	NA
68	F02522	ITAC UK	LTD	NAEF	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &SANMOL	PAL	ROC	ROC	NA
69	F02537	EDS GMBH	E	ACTV	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &Ground Floor	ROC	ROC	ROC	NA
70	F02594	CHANDRA	I	ACTV	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &NO 46	KK	MIRO	ROC	NA
71	F02608	SDG TRADIN	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &FLAT NO 4-B	ROC	ROC	ROC	NA
72	F02630	GLOBAL	BR	ACTV	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &SUITE NO 40	ROC	ROC	ROC	NA
73	F02673	EXPONENTI	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &NEW NO - 28	ROC	ROC	ROC	NA
74	F02694	ATHEROS	INDNAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &141, TVH	AGRO	ROC	ROC	NA
75	F02695	DECH-HARNAEF	NA	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture &PRINCE	TOVRO	ROC	ROC	NA
76	F02734	NSR	LIMIT	NAEF	NA	NA	29/05/2006	Tamil Nadu	0	0	NA	Agriculture &06 No.5 /	NRRO	ROC	ROC	NA
77	F02738	ONYX	ASIA	ACTV	NA	NA	29/05/2006	Tamil Nadu	0	0	NA	Agriculture &NO 6, RAJA	MRO	ROC	ROC	NA
78	F02749	TETRIBERICA	ACTV	NA	NA	NA	09/06/2006	Tamil Nadu	0	0	NA	Agriculture &NO. 16, ANN	RO	ROC	ROC	NA
79	F02753	POCLAIN	HYNAEF	NA	NA	NA	10/06/2006	Tamil Nadu	0	0	NA	Agriculture &KAPEEJAY	BURO	ROC	ROC	NA
80	F02758	WELCH	ALI	ACTV	NA	NA	27/06/2006	Tamil Nadu	0	0	NA	Agriculture &FLAT NO 3C,	ROC	ROC	ROC	NA
81	F02770	ONESTEEL	I	ACTV	NA	NA	11/07/2006	Tamil Nadu	0	0	NA	Agriculture &JAMALS	FAZRO	ROC	ROC	NA
82	F02772	CHARABOT	S	ACTV	NA	NA	22/07/2006	Tamil Nadu	0	0	NA	Agriculture &CHARABOT	I	ROC	ROC	NA
83	F02785	CONVANTA	I	NAEF	NA	NA	20/08/2006	Tamil Nadu	0	0	NA	Agriculture &62/63, OX	FORO	ROC	ROC	NA
84	F02793	BIDIM	SING	ACTV	NA	NA	24/08/2006	Tamil Nadu	0	0	NA	Agriculture &2nd Floor, Nd	ROC	ROC	ROC	NA
85	F02795	BEX	SHERMACTV	NA	NA	NA	24/08/2006	Tamil Nadu	0	0	NA	Agriculture &06 No. 6, Nd	ROC	ROC	ROC	NA
86	F02796	TREXTA	INDI	ACTV	NA	NA	04/09/2006	Tamil Nadu	0	0	NA	Agriculture &22, PADMAN	ROC	ROC	ROC	NA
87	F02812	TITANUS	S	AACTV	NA	NA	18/09/2006	Tamil Nadu	0	0	NA	Agriculture &9 VELLA	YD	IRO	ROC	NA
88	F02813	SSM	SCHARNAEF	NA	NA	NA	18/09/2006	Tamil Nadu	0	0	NA	Agriculture &II FLOOR	GE	RO	ROC	NA
89	F02817	PIEL	COLOR	I	NAEF	NA	21/09/2006	Tamil Nadu	0	0	NA	Agriculture &1-A, SARGUN	RO	ROC	ROC	NA
90	F02828	SCHMIDT	EN	ACTV	NA	NA	29/09/2006	Tamil Nadu	0	0	NA	Agriculture &C/o BIZPRO	SRO	ROC	ROC	NA
91	F02830	GEOPETROL	ACTV	NA	NA	NA	29/09/2006	Tamil Nadu	0	0	NA	Agriculture &Lakshmi	Ch	ARO	ROC	NA
92	F02837	ROHM	SEMIONAEF	NA	NA	NA	13/10/2006	Tamil Nadu	0	0	NA	Agriculture &SUNIT 103, IS	RO	ROC	ROC	NA
93	F02842	STAHL	CRANNAEF	NA	NA	NA	19/10/2006	Tamil Nadu	0	0	NA	Agriculture &Door No. 19	RO	ROC	ROC	NA
94	F02846	CHARABOT	S	ACTV	NA	NA	23/10/2006	Tamil Nadu	0	0	NA	Agriculture &Unit No. 104	RO	ROC	ROC	NA
95	F02871	REVO	TRADINAEF	NA	NA	NA	30/11/2006	Tamil Nadu	0	0	NA	Agriculture &A, 2A/2C, IS	RO	ROC	ROC	NA
96	F02884	TERIMU	CO	ACTV	NA	NA	15/12/2006	Tamil Nadu	0	0	NA	Agriculture &2nd Floor, C	RO	ROC	ROC	NA
97	F02898	BCE	CANCO	ACTV	NA	NA	17/01/2007	Tamil Nadu	0	0	NA	Agriculture &Citi Centre,	LRO	ROC	ROC	NA
98	F02908	S & V	INDUSIACTV	NA	NA	NA	24/01/2007	Tamil Nadu	0	0	NA	Agriculture &No. 114,Sou	TH	ROC	ROC	NA
99	F02910	ANSALDO	C	ACTV	NA	NA	30/01/2007	Tamil Nadu	0	0	NA	Agriculture &NO 53	HABERRO	ROC	ROC	NA
100	F02915	TDK	SINGAP	ACTV	NA	NA	30/01/2007	Tamil Nadu	0	0	NA	Agriculture &C/O M. ANA	NRRO	ROC	ROC	NA
101	F02924	MURATA	ELENAEF	NA	NA	NA	12/02/2007	Tamil Nadu	0	0	NA	Agriculture &714A, 7TH	FIRO	ROC	ROC	NA
102	F02937	CALYX	HEAL	ACTV	NA	NA	24/02/2007	Tamil Nadu	0	0	NA	Agriculture &S-2, Kurin	CHRO	ROC	ROC	NA
103	F02938	ALKYON	HYD	NAEF	NA	NA	26/02/2007	Tamil Nadu	0	0	NA	Agriculture &FlatC, Second	RO	ROC	ROC	NA
104	F02941	ULTRAMATE	NAEF	NA	NA	NA	01/03/2007	Tamil Nadu	0	0	NA	Agriculture &NEW NO 62,	ROC	ROC	ROC	NA
105	F02942	MERO	ASIA	ACTV	NA	NA	02/03/2007	Tamil Nadu	0	0	NA	Agriculture &SYMTEC	BURO	ROC	ROC	NA
106	F02956	UMDAN	JAY	ACTV	NA	NA	20/03/2007	Tamil Nadu	0	0	NA	Agriculture &BHA	GVAM	CI	ROC	NA
107	F02958	SIFANI	JEW	ACTV	NA	NA	20/03/2007	Tamil Nadu								