Data Collection and Preprocessing Phase

Date	15 july 2024
Team ID	740040
Project Title	Predicting co2 emissions by countries using machine learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	**Description**
Data Overview	Basic statistics, dimensions, and structure of the data.
Univariate Analysis	Exploration of individual variables (mean, median, mode, etc.).
Bivariate Analysis	Relationships between two variables (correlation, scatter plots).
Multivariate Analysis	Patterns and relationships involving multiple variables.
Outliers and Anomalies	Identification and treatment of outliers.
Data Preprocessing Code Screenshots	

<table>
<tr><td>Loading Data</td><td>

READING THE DATASET

```
[ ] #Reading the dataset
    data=pd.read_csv("/content/Indicators.csv")
```

DATASET

```
[ ] data.shape
```

```
(5656458, 6)
```

```
[ ] #Representing first 5 values from the dataset
    data.head()
```

	CountryName	CountryCode	IndicatorName	IndicatorCode	Year	Value
0	Arab World	ARB	Adolescent fertility rate (births per 1,000 wo...	SP.ADO.TFRT	1960	1.335609e+02
1	Arab World	ARB	Age dependency ratio (% of working-age populat...	SP.POP.DPND	1960	8.779760e+01
2	Arab World	ARB	Age dependency ratio, old (% of working-age po...	SP.POP.DPND.OL	1960	6.634579e+00
3	Arab World	ARB	Age dependency ratio, young (% of working-age ...	SP.POP.DPND.YG	1960	8.102333e+01
4	Arab World	ARB	Arms exports (SIPRI trend indicator values)	MS.MIL.XPRT.KD	1960	3.000000e+06

</td></tr>
<tr><td>Handling Missing Data</td><td>

HANDLING MISSING DATA

```
[ ] #Returns true if any columns having null values
    data.isnull().any()
```

```
CountryName      False
CountryCode      False
IndicatorName    False
IndicatorCode    False
Year             False
Value            False
dtype: bool
```

```
[ ] #Used for finding the null values
    data.isnull().sum()
```

```
CountryName      0
CountryCode      0
IndicatorName    0
IndicatorCode    0
Year             0
Value            0
dtype: int64
```

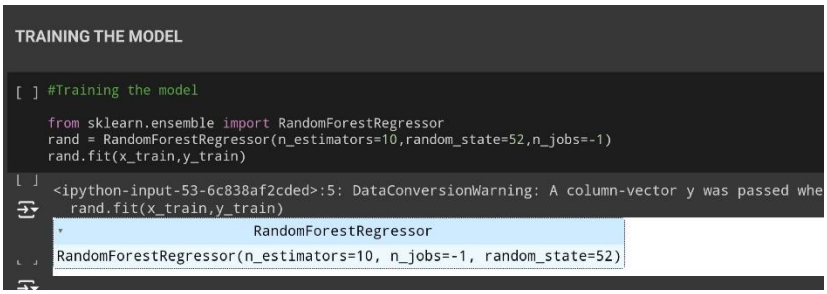
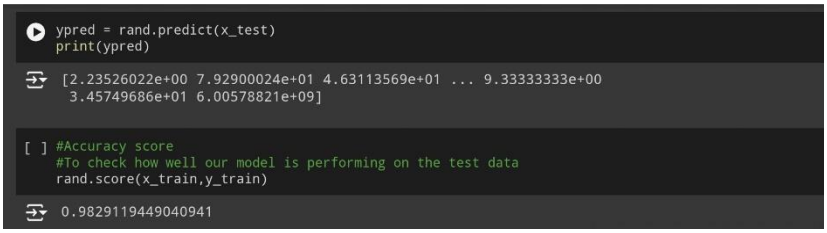
</td></tr>
<tr><td>Splitting data</td><td>

SPLITTING DATA

```
[ ] #Splitting dataset into train and test
    x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=1)
    print(x_train.shape)
    print(x_test.shape)
    print(y_train.shape)
    print(y_test.shape)
```

```
(4525166, 4)
(1131292, 4)
(4525166, 1)
(1131292, 1)
```

</td></tr>
<tr><td></td><td></td></tr>
</table>

Model training	
Model evaluation	
Saving the model