Project title: Customer churn prediction

Phase 3: Development

Part 1

In this part you will begin building your project by loading and preprocessing the dataset.

Begin conducting the Customer churn prediction by collecting and preprocessing the data.

Collect and preprocess the Customer data for analysis.

Data Preprocessing:

- **Data preprocessing is a crucial step within the statistics analysis and gadget gaining knowledge of pipeline.**

- **It includes a sequence of strategies and operations finished on uncooked statistics to clean, organize, and transform it right into a layout that is suitable for analysis or device mastering version schooling.**

- **Data preprocessing goals to enhance the first-class of the records, making it greater reliable and conducive to generating accurate consequences.**

Here are some common tasks and techniques involved in data preprocessing:

Data Cleaning:

- **Handling missing values: Deciding how to deal with missing data, whether by imputing values or removing incomplete records.**

- **Outlier detection and treatment: Identifying and handling data points that significantly deviate from the norm.**

Noise reduction:

- **Smoothing noisy data through techniques like filtering.**

Data Transformation:

- **Data normalization: Scaling numerical features to a standard range (e.g., between 0 and 1) to ensure that they have similar influence in the analysis.**

- **Encoding categorical variables: Converting categorical data into numerical format, such as one-hot encoding or label encoding.**

- **Feature engineering: Creating new features or modifying existing ones to capture more meaningful information from the data.**

- **Dimensionality reduction: Reducing the number of features while retaining essential information, using methods like Principal Component Analysis (PCA).**

Data Integration:

- **Merging or joining datasets: Combining data from multiple sources into a single dataset for analysis.**

Aggregation: Summarizing data at a higher level of granularity, such as aggregating daily sales into monthly totals.

Data Reduction:

- **Sampling: Reducing the size of a large dataset by randomly selecting a representative subset.**

- **Binning: Grouping continuous data into discrete bins to simplify analysis.**

- **Filtering: Selecting a subset of data based on specific criteria.**

Data Standardization:

- **Ensuring that data follows a consistent format and structure.**

- **Date and time format conversion: Converting date and time data into a uniform format.**

- **Currency conversion: Converting monetary values into a common currency.**

Data Scaling:

- **Scaling numerical data to a common range to prevent some features from dominating the analysis.**

Data preprocessing is an iterative process that may involve several of these steps in various orders, depending on the specific dataset and the analysis goals. Proper data preprocessing is essential for improving the accuracy and effectiveness of machine learning models, as well as for making data more accessible for traditional statistical analysis.

Here is the data preprocessing codes along with the output of the given dataset:

Importing the libraries:

Import three basic libraries which are very common in machine learning and will be used every time you train a model

- **NumPy: it is a library that allows us to work with arrays and as most machine learning models work on arrays NumPy makes it easier**

- **Matplotlib: this library helps in plotting graphs and charts, which are very useful while showing the result of your model**

- **Pandas: pandas allows us to import our dataset and also creates a matrix of features containing the dependent and independent variable.**

```
#Connect the google drive for reading the dataset # Connect the google drive
from google.colab import drive
drive.mount("/content/drive")
```

```
Mounted at /content/drive
```

```
# Preparing Dataset
# Import the dataset
import pandas as pd
dataset = pd.read_csv("/content/drive/MyDrive/BIT/Customer-churn.csv")
```

```
print(dataset)
```

```
      customerID  gender  SeniorCitizen Partner Dependents  tenure PhoneService
0     7590-VHVEG  Female              0     Yes         No       1           No
1     5575-GNVDE    Male              0      No         No      34          Yes
2     3668-QPYBK    Male              0      No         No       2          Yes
3     7795-CFOCW    Male              0      No         No      45           No
4     9237-HQITU  Female              0      No         No       2          Yes
5     9305-CDSKC  Female              0      No         No       8          Yes
6     1452-KIOVK    Male              0      No        Yes      22          Yes
7     6713-OKOMC  Female              0      No         No      10           No
8     7892-POOKP  Female              0     Yes         No      28          Yes
9     6388-TABGU    Male              0      No        Yes      62          Yes
10    9763-GRSKD    Male              0     Yes        Yes      13          Yes
11    7469-LKBCI    Male              0      No         No      16          Yes
12    8091-TTVAX    Male              0     Yes         No      58          Yes
13    0280-XJGEX    Male              0      No         No      49          Yes
14    5129-JLPIS    Male              0      No         No      25          Yes
15    3655-SNQYZ  Female              0     Yes        Yes      69          Yes
16    8191-XWSZG  Female              0      No         No      52          Yes
17    9959-WOFKT    Male              0      No        Yes      71          Yes
18    4190-MFLUW  Female              0     Yes        Yes      10          Yes
19    4183-MYFRB  Female              0      No         No      21          Yes
20    8779-QRDMV    Male              1      No         No       1           No
21    1680-VDCWW    Male              0     Yes         No      12          Yes
22    1066-JKSGK    Male              0      No         No       1          Yes
23    3638-WEABW  Female              0     Yes         No      58          Yes
24    6322-HRPFA    Male              0     Yes        Yes      49          Yes
25    6865-JZNKO  Female              0      No         No      30          Yes
26    6467-CHFZW    Male              0     Yes        Yes      47          Yes
27    8665-UTDHZ    Male              0     Yes        Yes       1           No
28    5248-YGIJN    Male              0     Yes         No      72          Yes

      MultipleLines InternetService OnlineSecurity ...  \
0   No phone service             DSL             No ...
1                 No             DSL            Yes ...
2                 No             DSL            Yes ...
3   No phone service             DSL            Yes ...
4                 No     Fiber optic             No ...
```

5		Yes	Fiber optic		No	...
6		Yes	Fiber optic		No	...
7	No phon service e		DSL		Yes	...
8		Yes	Fiber optic		No	...
9		No	DSL		Yes	...
10		No	DSL		Yes	...
11		No	No	No interne service t		...
12		Yes	Fiber optic		No	...
13		Yes	Fiber optic		No	...
14		No	Fiber optic		Yes	...
15		Yes	Fiber optic		Yes	...
16		No	No	No interne service t		...
17		Yes	Fiber optic		Yes	...
18		No	DSL		No	...
19		No	Fiber optic		No	...
20	No phon service e		DSL		No	...
21		No	No	No interne service t		...
22		No	No	No interne service t		...
23		Yes	DSL		No	...
24		No	DSL		Yes	...

dataset.dropna

<bound method DataFrame.dropna of customerID gender SeniorCitizen Partner Dependents tenure PhoneService \

0	7590-VHVEG	Female	0	Yes	No	1	No
1	5575-GNVDE	Male	0	No	No	34	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes
3	7795-CFOCW	Male	0	No	No	45	No
4	9237-HQITU	Female	0	No	No	2	Yes
5	9305-CDSKC	Female	0	No	No	8	Yes
6	1452-KIOVK	Male	0	No	Yes	22	Yes
7	6713-OKOMC	Female	0	No	No	10	No
8	7892-POOKP	Female	0	Yes	No	28	Yes
9	6388-TABGU	Male	0	No	Yes	62	Yes
10	9763-GRSKD	Male	0	Yes	Yes	13	Yes
11	7469-LKBCI	Male	0	No	No	16	Yes
12	8091-TTVAX	Male	0	Yes	No	58	Yes
13	0280-XJGEX	Male	0	No	No	49	Yes
14	5129-JLPIS	Male	0	No	No	25	Yes
15	3655-SNQYZ	Female	0	Yes	Yes	69	Yes
16	8191-XWSZG	Female	0	No	No	52	Yes
17	9959-WOFKT	Male	0	No	Yes	71	Yes
18	4190-MFLUW	Female	0	Yes	Yes	10	Yes
19	4183-MYFRB	Female	0	No	No	21	Yes
20	8779-QRDMV	Male	1	No	No	1	No
21	1680-VDCWW	Male	0	Yes	No	12	Yes
22	1066-JKSGK	Male	0	No	No	1	Yes
23	3638-WEABW	Female	0	Yes	No	58	Yes
24	6322-HRPFA	Male	0	Yes	Yes	49	Yes
25	6865-JZNKO	Female	0	No	No	30	Yes
26	6467-CHFZW	Male	0	Yes	Yes	47	Yes
27	8665-UTDHZ	Male	0	Yes	Yes	1	No
28	5248-YGIJN	Male	0	Yes	No	72	Yes

		MultipleLines	InternetService	OnlineSecurity	...	\
0	No	phone service	DSL	No	...	
1		No	DSL	Yes	...	
2		No	DSL	Yes	...	

3	No	phone	service	DSL		Yes	...
4		No	Fiber optic			No	...
5		Yes	Fiber optic			No	...
6		Yes	Fiber optic			No	...
7	No	phone	service	DSL		Yes	...
8		Yes	Fiber optic			No	...
9		No	DSL			Yes	...
10		No	DSL			Yes	...
11		No	No	No	internet	service	...
12		Yes	Fiber optic			No	...
13		Yes	Fiber optic			No	...
14		No	Fiber optic			Yes	...
15		Yes	Fiber optic			Yes	...
16		No	No	No	internet	service	...
17		Yes	Fiber optic			Yes	...
18		No	DSL			No	...
19		No	Fiber optic			No	...
20	No	phone	service	DSL		No	...
21		No	No	No	internet	service	...
22		No	No	No	internet	service	...
23		Yes	DSL			No	...
24		No	DSL			Yes	

dataset.isnull()

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneServi
							PhoneServi
0	False	False	False	False	False	False	Fal
1	False	False	False	False	False	False	Fal
2	False	False	False	False	False	False	Fal
3	False	False	False	False	False	False	Fal
4	False	False	False	False	False	False	Fal
5	False	False	False	False	False	False	Fal
6	False	False	False	False	False	False	Fal
7	False	False	False	False	False	False	Fal
8	False	False	False	False	False	False	Fal
9	False	False	False	False	False	False	Fal
10	False	False	False	False	False	False	Fal
11	False	False	False	False	False	False	Fal

dataset.info

```
<bound method DataFrame.info of      customerID  gender  SeniorCitizen Partner Dependents  tenure PhoneService \
0    7590-VHVEG  Female              0     Yes         No       1           No
1    5575-GNVDE    Male              0      No         No      34          Yes
2    3668-QPYBK    Male              0      No         No       2          Yes
3    7795-CFOCW    Male              0      No         No      45           No
4    9237-HQITU  Female              0      No         No       2          Yes
5    9305-CDSKC  Female              0      No         No       8          Yes
6    1452-KIOVK    Male              0      No        Yes      22          Yes
7    6713-OKOMC  Female              0      No         No      10           No
8    7892-POOKP  Female              0     Yes         No      28          Yes
9    6388-TABGU    Male              0      No        Yes      62          Yes
10   9763-GRSKD    Male              0     Yes        Yes      13          Yes
11   7469-LKBCI    Male              0      No         No      16          Yes
12   8091-TTVAX    Male              0     Yes         No      58          Yes
13   0280-XJGEX    Male              0      No         No      49          Yes
14   5129-JLPIS    Male              0      No         No      25          Yes
15   3655-SNQYZ  Female              0     Yes        Yes      69          Yes
16   8191-XWSZG  Female              0      No         No      52          Yes
17   9959-WOFKT    Male              0      No        Yes      71          Yes
18   4190-MFLUW  Female              0     Yes        Yes      10          Yes
19   4183-MYFRB  Female              0      No         No      21          Yes
20   8779-QRDMV    Male              1      No         No       1           No
21   1680-VDCWW    Male              0     Yes         No      12          Yes
22   1066-JKSGK    Male              0      No         No       1          Yes
23   3638-WEABW  Female              0     Yes         No      58          Yes
24   6322-HRPFA    Male              0     Yes        Yes      49          Yes
25   6865-JZNKO  Female              0      No         No      30          Yes
26   6467-CHFZW    Male              0     Yes        Yes      47          Yes
27   8665-UTDHZ    Male              0     Yes        Yes       1           No
28   5248-YGIJN    Male              0     Yes         No      72          Yes

      MultipleLines InternetService       OnlineSecurity     ... \
0  No phone service             DSL                   No     ...
```

1	No	DSL	Yes	...	
2	No	DSL	Yes	...	
3	No phone service	DSL	Yes	...	
4	No	Fiber optic	No	...	
5	Yes	Fiber optic	No	...	
6	Yes	Fiber optic	No	...	
7	No phone service	DSL	Yes	...	
8	Yes	Fiber optic	No	...	
9	No	DSL	Yes	...	
10	No	DSL	Yes	...	
11	No	No	No internet service	...	
12	Yes	Fiber optic	No	...	
13	Yes	Fiber optic	No	...	
14	No	Fiber optic	Yes	...	
15	Yes	Fiber optic	Yes	...	
16	No	No	No internet service	...	
17	Yes	Fiber optic	Yes	...	
18	No	DSL	No	...	
19	No	Fiber optic	No	...	
20	No phone service	DSL	No	...	
21	No	No	No internet service	...	
22	No	No	No internet service	...	
23	Yes	DSL	No	...	
24	No	DSL	Yes	...	

dataset.describe

<bound method NDFrame.describe of customerID gender SeniorCitizen Partner Dependents tenure PhoneService \

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
0	7590-VHVEG	Female	0	Yes	No	1	No
1	5575-GNVDE	Male	0	No	No	34	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes
3	7795-CFOCW	Male	0	No	No	45	No
4	9237-HQITU	Female	0	No	No	2	Yes
5	9305-CDSKC	Female	0	No	No	8	Yes
6	1452-KIOVK	Male	0	No	Yes	22	Yes
7	6713-OKOMC	Female	0	No	No	10	No
8	7892-POOKP	Female	0	Yes	No	28	Yes
9	6388-TABGU	Male	0	No	Yes	62	Yes
10	9763-GRSKD	Male	0	Yes	Yes	13	Yes
11	7469-LKBCI	Male	0	No	No	16	Yes
12	8091-TTVAX	Male	0	Yes	No	58	Yes
13	0280-XJGEX	Male	0	No	No	49	Yes
14	5129-JLPIS	Male	0	No	No	25	Yes
15	3655-SNQYZ	Female	0	Yes	Yes	69	Yes
16	8191-XWSZG	Female	0	No	No	52	Yes
17	9959-WOFKT	Male	0	No	Yes	71	Yes
18	4190-MFLUW	Female	0	Yes	Yes	10	Yes
19	4183-MYFRB	Female	0	No	No	21	Yes
20	8779-QRDMV	Male	1	No	No	1	No
21	1680-VDCWW	Male	0	Yes	No	12	Yes
22	1066-JKSGK	Male	0	No	No	1	Yes
23	3638-WEABW	Female	0	Yes	No	58	Yes
24	6322-HRPFA	Male	0	Yes	Yes	49	Yes
25	6865-JZNKO	Female	0	No	No	30	Yes

26	6467-CHFZW	Male	0	Yes	Yes	47	Yes
27	8665-UTDHZ	Male	0	Yes	Yes	1	No
28	5248-YGIJN	Male	0	Yes	No	72	Yes

```
    MultipleLines InternetService       OnlineSecurity     ...    \
0   No phone service             DSL                   No   ...
1                 No             DSL                  Yes   ...
2                 No             DSL                  Yes   ...
3   No phone service             DSL                  Yes   ...
4                 No     Fiber optic                   No   ...
5                Yes     Fiber optic                   No   ...
6                Yes     Fiber optic                   No   ...
7   No phone service             DSL                  Yes   ...
8                Yes     Fiber optic                   No   ...
9                 No             DSL                  Yes   ...
10                No             DSL                  Yes   ...
11                No              No  No internet service   ...
12               Yes     Fiber optic                   No   ...
13               Yes     Fiber optic                   No   ...
14                No     Fiber optic                  Yes   ...
15               Yes     Fiber optic                  Yes   ...
16                No              No  No internet service   ...
17               Yes     Fiber optic                  Yes   ...
18                No             DSL                   No   ...
19                No     Fiber optic                   No   ...
20  No phone service             DSL                   No   ...
21                No              No  No internet service   ...
22                No              No  No internet service   ...
23               Yes             DSL                   No   ...
```

```
import matplotlib.pyplot as plt

X=dataset.MonthlyCharges

Y=dataset.TotalCharges

Xtrain = dataset[['gender','PaymentMethod','OnlineBackup','PaperlessBilling']]

Ytrain = dataset[['Churn']]

print(Xtrain)
```

```
    gender              PaymentMethod OnlineBackup PaperlessBilling
0   Female           Electronic check          Yes              Yes
1     Male               Mailed check           No               No
2     Male               Mailed check          Yes              Yes
3     Male  Bank transfer (automatic)           No               No
4   Female           Electronic check           No              Yes
5   Female           Electronic check           No              Yes
6     Male  Credit card (automatic)          Yes              Yes
7   Female               Mailed check           No               No
8   Female           Electronic check           No              Yes
9     Male  Bank transfer (automatic)          Yes               No
10    Male               Mailed check           No              Yes
```

11	Male	Credit card (automatic)	No internet service		No
12	Male	Credit card (automatic)	No		No
13	Male	Bank transfer (automatic)	Yes		Yes
14	Male	Electronic check	No		Yes

15	Female	Credit card (automatic)		Yes	No
16	Female	Mailed check	No internet service		No
17	Male	Bank transfer (automatic)		No	No
18	Female	Credit card (automatic)		No	No
19	Female	Electronic check		Yes	Yes
20	Male	Electronic check		No	Yes
21	Male	Bank transfer (automatic)	No internet service		No
22	Male	Mailed check	No internet service		No
23	Female	Credit card (automatic)		Yes	Yes
24	Male	Credit card (automatic)		Yes	No
25	Female	Bank transfer (automatic)		Yes	Yes
26	Male	Electronic check		Yes	Yes
27	Male	Electronic check		Yes	No
28	Male	Credit card (automatic)		Yes	Yes

```
print(Ytrain)
```

```
    Churn
0      No
1      No
2     Yes
3      No
4     Yes
5     Yes
6      No
7      No
8     Yes
9      No
10     No
11     No
12     No
13    Yes
14     No
15     No
16     No
17     No
18    Yes
19     No
20    Yes
21     No
22    Yes
23     No
24     No
25     No
26    Yes
27    Yes
28     No
```

```
from sklearn.preprocessing import OrdinalEncoder
```

```
enc = OrdinalEncoder()
```

```
enc.fit(Xtrain)
```

▾ OrdinalEncoder

```
Xtrain_encoded=enc.transform(Xtrain)

print(Xtrain_encoded)
```

```
[[0.  2.  2.  1.]
 [1.  3.  0.  0.]
 [1.  3.  2.  1.]
 [1.  0.  0.  0.]
 [0.  2.  0.  1.]
 [0.  2.  0.  1.]
 [1.  1.  2.  1.]
 [0.  3.  0.  0.]
 [0.  2.  0.  1.]
 [1.  0.  2.  0.]
 [1.  3.  0.  1.]
 [1.  1.  1.  0.]
 [1.  1.  0.  0.]
 [1.  0.  2.  1.]
 [1.  2.  0.  1.]
 [0.  1.  2.  0.]
 [0.  3.  1.  0.]
 [1.  0.  0.  0.]
 [0.  1.  0.  0.]
 [0.  2.  2.  1.]
 [1.  2.  0.  1.]
 [1.  0.  1.  0.]
 [1.  3.  1.  0.]
 [0.  1.  2.  1.]
 [1.  1.  2.  0.]
 [0.  0.  2.  1.]
 [1.  2.  2.  1.]
 [1.  2.  2.  0.]
 [1.  1.  2.  1.]]
```

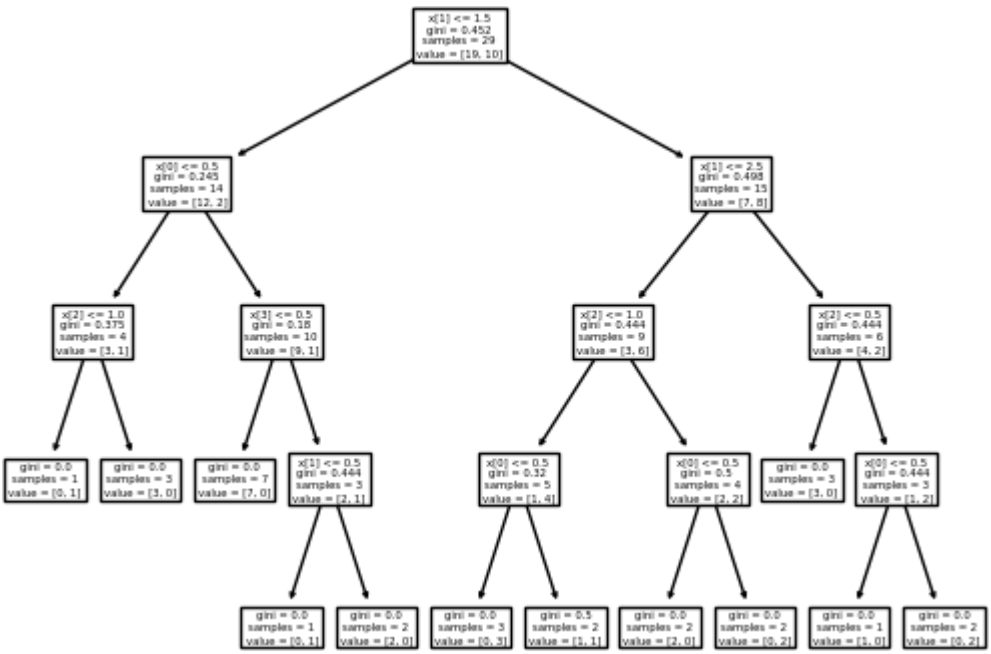
```
from sklearn import tree

clf = tree.DecisionTreeClassifier()

clf.fit(Xtrain_encoded,Ytrain)
```

▾ DecisionTreeClassifier

DecisionTreeClassifier()

```
tree.plot_tree(clf)
```

[Text(0.4642857142857143, 0.9, 'x[1] <= 1.5\ngini = 0.452\nsamples = 29\nvalue = [19, 10]'),
 Text(0.19047619047619047, 0.7, 'x[0] <= 0.5\ngini = 0.245\nsamples = 14\nvalue = [12, 2]'),
 Text(0.09523809523809523, 0.5, 'x[2] <= 1.0\ngini = 0.375\nsamples = 4\nvalue = [3, 1]'),
 Text(0.047619047619047616, 0.3, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(0.14285714285714285, 0.3, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
 Text(0.2857142857142857, 0.5, 'x[3] <= 0.5\ngini = 0.18\nsamples = 10\nvalue = [9, 1]'),
 Text(0.23809523809523808, 0.3, 'gini = 0.0\nsamples = 7\nvalue = [7, 0]'),
 Text(0.3333333333333333, 0.3, 'x[1] <= 0.5\ngini = 0.444\nsamples = 3\nvalue = [2, 1]'),
 Text(0.2857142857142857, 0.1, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(0.38095238095238093, 0.1, 'gini = 0.0\nsamples = 2\nvalue = [2, 0]'),
 Text(0.7380952380952381, 0.7, 'x[1] <= 2.5\ngini = 0.498\nsamples = 15\nvalue = [7, 8]'),
 Text(0.6190476190476191, 0.5, 'x[2] <= 1.0\ngini = 0.444\nsamples = 9\nvalue = [3, 6]'),
 Text(0.5238095238095238, 0.3, 'x[0] <= 0.5\ngini = 0.32\nsamples = 5\nvalue = [1, 4]'),
 Text(0.47619047619047616, 0.1, 'gini = 0.0\nsamples = 3\nvalue = [0, 3]'),
 Text(0.5714285714285714, 0.1, 'gini = 0.5\nsamples = 2\nvalue = [1, 1]'),
 Text(0.7142857142857143, 0.3, 'x[0] <= 0.5\ngini = 0.5\nsamples = 4\nvalue = [2, 2]'),
 Text(0.6666666666666666, 0.1, 'gini = 0.0\nsamples = 2\nvalue = [2, 0]'),
 Text(0.7619047619047619, 0.1, 'gini = 0.0\nsamples = 2\nvalue = [0, 2]'),
 Text(0.8571428571428571, 0.5, 'x[2] <= 0.5\ngini = 0.444\nsamples = 6\nvalue = [4, 2]'),
 Text(0.8095238095238095, 0.3, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
 Text(0.9047619047619048, 0.3, 'x[0] <= 0.5\ngini = 0.444\nsamples = 3\nvalue = [1, 2]'),
 Text(0.8571428571428571, 0.1, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
 Text(0.9523809523809523, 0.1, 'gini = 0.0\nsamples = 2\nvalue = [0, 2]')]



```
from sklearn.ensemble import
RandomForestClassifier clf =
RandomForestClassifier(n_estimators = 100)
clf.fit(Xtrain_encoded,Ytrain)
```

```
<ipython-input-20-b6cd1249641e>:3: DataConversionWarning: A column-vector y
  w clf.fit(Xtrain_encoded,Ytrain)
```

▾ RandomForestClassifier
RandomForestClassifier()

```
import numpy as np
arr = np.array([[1, 1, 2, 1]])
print(clf.predict(arr))
```

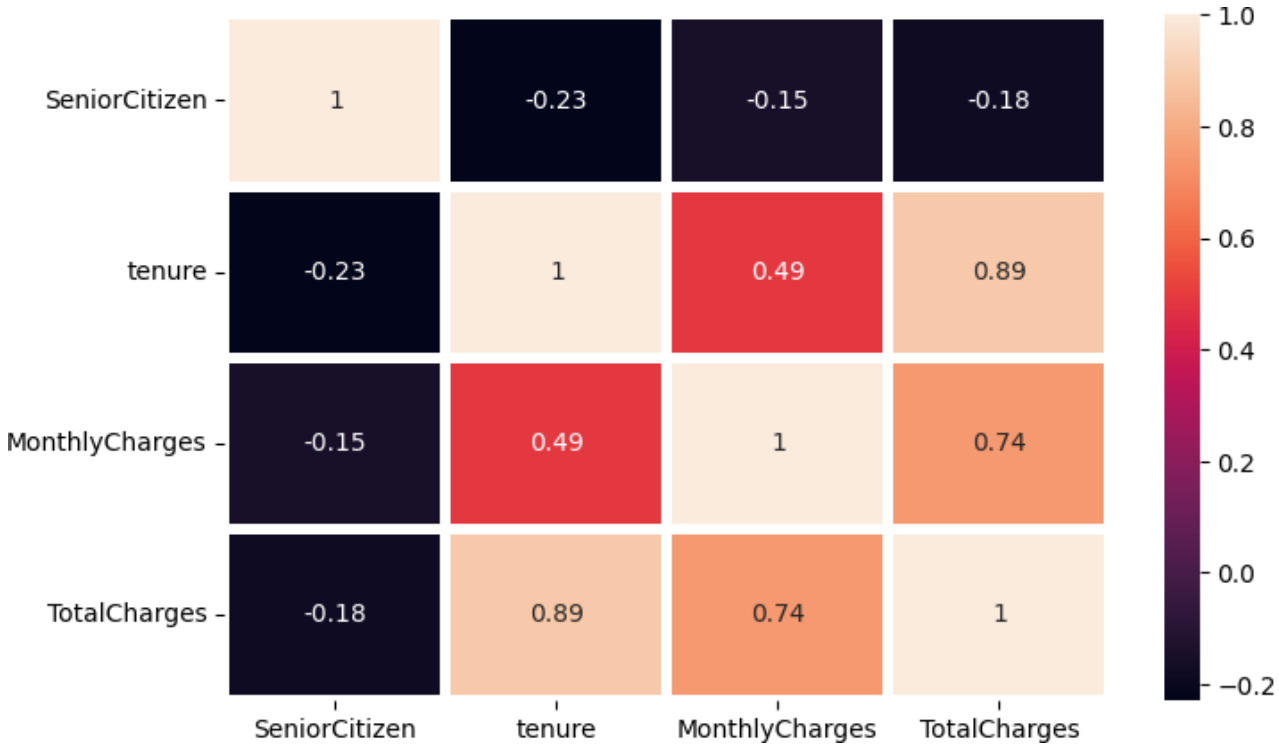
```
['No']
```

```
import numpy as np
arr1 = np.array([[1, 3, 2, 1]])
print(clf.predict(arr1))
```

```
['Yes']
```

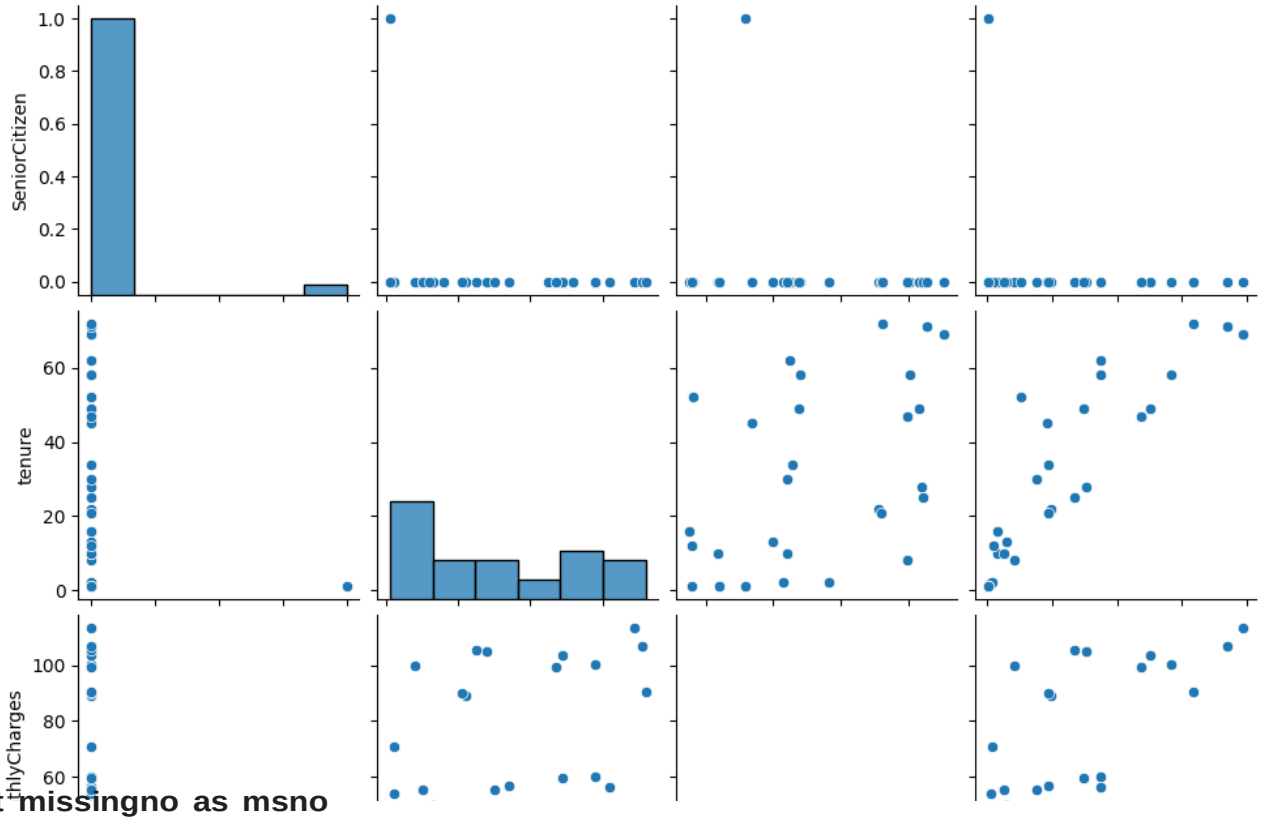
```
import seaborn as sns

plt.figure(figsize=(8,5))
sns.heatmap(dataset.corr(),annot=True,linewidth=
3) plt.show
```

```
<ipython-input-25-095c9657f905>:2: FutureWarning: The default value of numeri
  sns.heatmap(dataset.corr(),annot=True,linewidth=3)
<function matplotlib.pyplot.show(close=None, block=None)>
```

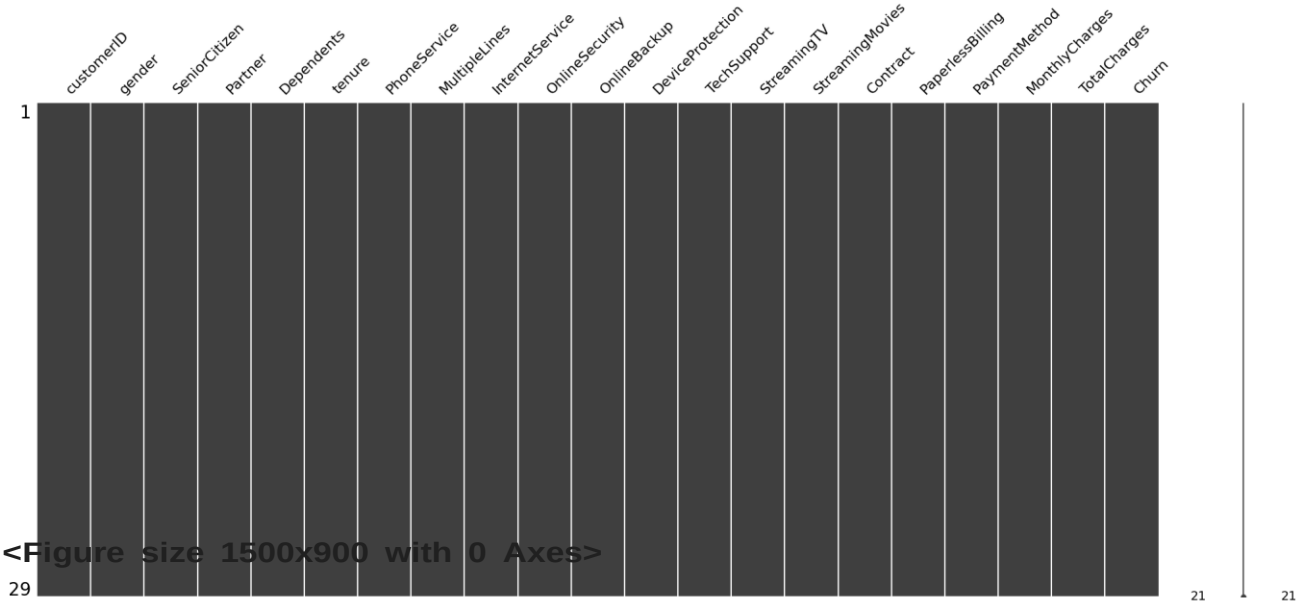



```
sns.pairplot(dataset)
```

<seaborn.axisgrid.PairGrid at 0x7813753a3e80>



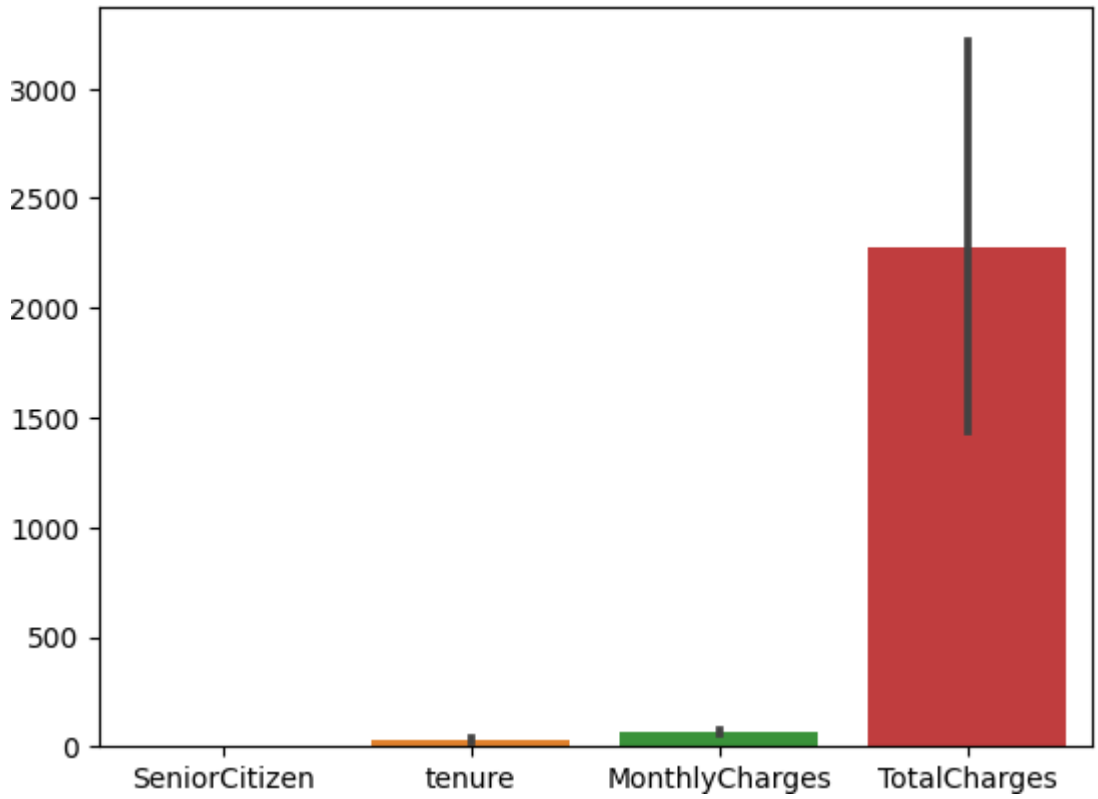

```
import missingno as msno
msno.matrix(dataset)
plt.figure(figsize=(15,9))
plt.show()
```



<Figure size 1500x900 with 0 Axes>

```
sns.barplot(dataset)
```

<Axes: >



dataset.hist()

array([[<Axes: title={'center': 'SeniorCitizen'}>,
<Axes: title={'center': 'tenure'}>],
[<Axes: title={'center': 'MonthlyCharges'}>,
<Axes: title={'center': 'TotalCharges'}>]], dtype=object)