DOUGLAS COLLEGE

FUNDAMENTALS OF MACHINE LEARNING IN DATA SCIENCE

CSIS 3290

PREPROCESSING

IN SCIKIT-LEARN (SKLERAN)

FATEMEH AHMADI

Installing Scikit-learn


```
(base) C:\Users\Paris>conda install -c anaconda scikit-learn
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: D:\Anaconda

  added / updated specs:
    - scikit-learn


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    joblib-1.1.1               |   py39haa95532_0         410 KB  anaconda
    scikit-learn-1.2.0         |   py39hd77b12b_0         7.7 MB  anaconda
    threadpoolctl-2.2.0        |     pyh0d69192_0          16 KB  anaconda
    ------------------------------------------------------------
                                           Total:         8.1 MB

The following NEW packages will be INSTALLED:

  joblib             anaconda/win-64::joblib-1.1.1-py39haa95532_0
  scikit-learn       anaconda/win-64::scikit-learn-1.2.0-py39hd77b12b_0
  threadpoolctl      anaconda/noarch::threadpoolctl-2.2.0-pyh0d69192_0

The following packages will be SUPERSEDED by a higher-priority channel:
```

Preprocessing : General Information

In [1]:
```
import numpy as np
import pandas as pd
from sklearn import preprocessing
```

In [2]:
```
data1=pd.read_csv('F:/00-Douglas College/1- Semester 1/3- Machine Learning in Data Science(3290)/Slides/smartphone.csv')
```

In [3]:
```
data1.head()
```

Out[3]:

	Product Name	Product URL	Brand	Sale Price	Mrp	Discount Percentage	Number Of Ratings	Number Of Reviews	Upc	Star Rating	Ram
0	XOLO T1000 (Black, 4 GB)	https://www.flipkart.com/xolo-t1000-black-4-gb...	XOLO	14153	14153	0	333	130	MOBDMKDAKQGCYZ6D	3.8	1 GB
1	GIONEE Pioneer P3 (White, 4 GB)	https://www.flipkart.com/gionee-pioneer-p3-whi...	GIONEE	6500	6500	0	437	78	MOBDRKHTA3UXHAVD	3.6	512 MB
2	KARBONN Titanium S4 (Black, 4 GB)	https://www.flipkart.com/karbonn-titanium-s4-b...	KARBONN	13298	13298	0	28	7	MOBDRYWHA3ZU9BRT	3.3	1 GB
3	KARBONN Titanium S4 (White, 4 GB)	https://www.flipkart.com/karbonn-titanium-s4-w...	KARBONN	14990	14990	0	28	7	MOBDRYWHFVVQHQVZ	3.3	1 GB
4	Micromax Bolt A71 (Black, 165 MB)	https://www.flipkart.com/micromax-bolt-a71-bla...	Micromax	6499	7499	13	61	8	MOBDSMAJ5UUJUDDE	3.1	512 MB

Preprocessing : General Information

```
In [4]: data1.shape
Out[4]: (1513, 11)

In [5]: data1.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1513 entries, 0 to 1512
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Product Name         1513 non-null   object
 1   Product URL          1513 non-null   object
 2   Brand                1513 non-null   object
 3   Sale Price           1513 non-null   int64
 4   Mrp                  1513 non-null   int64
 5   Discount Percentage  1513 non-null   int64
 6   Number Of Ratings    1513 non-null   int64
 7   Number Of Reviews    1513 non-null   int64
 8   Upc                  1513 non-null   object
 9   Star Rating          1513 non-null   float64
 10  Ram                  1513 non-null   object
dtypes: float64(1), int64(5), object(5)
memory usage: 130.1+ KB
```

Preprocessing: General Information

```
In [6]: data1.describe()
```

Out[6]:

	Sale Price	Mrp	Discount Percentage	Number Of Ratings	Number Of Reviews	Star Rating
count	1513.000000	1513.000000	1513.000000	1.513000e+03	1513.000000	1513.000000
mean	17616.582948	19350.060145	6.567746	3.616383e+04	3889.980172	3.865962
std	20373.673405	22981.244617	9.601012	1.221186e+05	15190.369690	1.050301
min	1190.000000	1190.000000	0.000000	0.000000e+00	0.000000	0.000000
25%	7199.000000	7849.000000	0.000000	2.500000e+02	25.000000	3.800000
50%	11499.000000	12990.000000	0.000000	2.858000e+03	287.000000	4.300000
75%	17999.000000	19979.000000	12.000000	1.549000e+04	1397.000000	4.400000
max	149999.000000	189999.000000	62.000000	1.340123e+06	213834.000000	4.800000

Preprocessing: Checking Null Values

In [7]:
```
data1.isnull()
```

Out[7]:

	Product Name	Product URL	Brand	Sale Price	Mrp	Discount Percentage	Number Of Ratings	Number Of Reviews	Upc	Star Rating	Ram
0	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
1508	False	False	False	False	False	False	False	False	False	False	False
1509	False	False	False	False	False	False	False	False	False	False	False
1510	False	False	False	False	False	False	False	False	False	False	False
1511	False	False	False	False	False	False	False	False	False	False	False
1512	False	False	False	False	False	False	False	False	False	False	False

1513 rows × 11 columns

In [8]:
```
data1.isnull().sum()
```

Out[8]:
```
Product Name           0
Product URL            0
Brand                  0
Sale Price             0
Mrp                    0
Discount Percentage    0
Number Of Ratings      0
Number Of Reviews      0
Upc                    0
Star Rating            0
Ram                    0
dtype: int64
```

Preprocessing: Dropping NAs


```
In [10]: data1.dropna(axis=0)
         data1.dropna(axis=1)
```

Out[10]:

	Product Name	Product URL	Brand	Sale Price	Mrp	Discount Percentage	Number Of Ratings	Number Of Reviews	Upc	Star Rating	Ram
0	XOLO T1000 (Black, 4 GB)	https://www.flipkart.com/xolo-t1000-black-4-gb...	XOLO	14153	14153	0	333	130	MOBDMKDAKQGCYZ6D	3.8	1 GB
1	GIONEE Pioneer P3 (White, 4 GB)	https://www.flipkart.com/gionee-pioneer-p3-whi...	GIONEE	6500	6500	0	437	78	MOBDRKHTA3UXHAVD	3.6	512 MB
2	KARBONN Titanium S4 (Black, 4 GB)	https://www.flipkart.com/karbonn-titanium-s4-b...	KARBONN	13298	13298	0	28	7	MOBDRYWHA3ZU9BRT	3.3	1 GB
3	KARBONN Titanium S4 (White, 4 GB)	https://www.flipkart.com/karbonn-titanium-s4-w...	KARBONN	14990	14990	0	28	7	MOBDRYWHFVVQHQVZ	3.3	1 GB
4	Micromax Bolt A71 (Black, 165 MB)	https://www.flipkart.com/micromax-bolt-a71-bla...	Micromax	6499	7499	13	61	8	MOBDSMAJ5UUJUDDE	3.1	512 MB
...	...	...	...	...	...	...	...	...	...	...	...
1508	Kekai Prime (Sea Blue, 32 GB)	https://www.flipkart.com/kekai-prime-sea-blue-...	Kekai	5499	5499	0	0	0	MOBGYYUWTY8DJUES	0.0	2 GB
1509	GIONEE S11 (Gold, 64 GB)	https://www.flipkart.com/gionee-s11-gold-64-gb...	GIONEE	8990	8990	0	0	0	MOBGYYUX6EBEHJCF	0.0	4 GB
1510	Kekai Prime (Sea White, 32 GB)	https://www.flipkart.com/kekai-prime-sea-white...	Kekai	5499	5499	0	0	0	MOBGYYUXQZBZ4DAY	0.0	2 GB
1511	Telefono S1 (Interstellar Black, 32 GB)	https://www.flipkart.com/telefono-s1-interstel...	Telefono	5990	5990	0	0	0	MOBGYZ8ZYCKJFEWV	0.0	3 GB
1512	Telefono S1 (Space Blue, 32 GB)	https://www.flipkart.com/telefono-s1-space-blu...	Telefono	5990	5990	0	0	0	MOBGYZ94RVSAXKSW	0.0	3 GB

1513 rows × 11 columns

Preprocessing: Renaming the Columns

```
In [11]: data1=data1.rename(columns={'Product Name':'Name', 'Product URL':'URL'})

In [12]: data1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1513 entries, 0 to 1512
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Name                 1513 non-null   object
 1   URL                  1513 non-null   object
 2   Brand                1513 non-null   object
 3   Sale Price           1513 non-null   int64
 4   Mrp                  1513 non-null   int64
 5   Discount Percentage  1513 non-null   int64
 6   Number Of Ratings    1513 non-null   int64
 7   Number Of Reviews    1513 non-null   int64
 8   Upc                  1513 non-null   object
 9   Star Rating          1513 non-null   float64
 10  Ram                  1513 non-null   object
dtypes: float64(1), int64(5), object(5)
memory usage: 130.1+ KB

In [ ]:
```

Preprocessing: Dropping a Column and Replacing a Value and Filling the NAs

```
In [17]: data1=data1.drop(columns='Upc')

In [18]: data1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1513 entries, 0 to 1512
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Name                  1513 non-null   object
 1   URL                   1513 non-null   object
 2   Brand                 1513 non-null   object
 3   Sale Price            1513 non-null   int64
 4   Mrp                   1513 non-null   int64
 5   Discount Percentage   1513 non-null   int64
 6   Number Of Ratings     1513 non-null   int64
 7   Number Of Reviews     1513 non-null   int64
 8   Star Rating           1513 non-null   float64
 9   Ram                   1513 non-null   object
dtypes: float64(1), int64(5), object(4)
memory usage: 118.3+ KB
```

```
In [19]: data1=data1.replace('?','np.nan')

In [20]: data1.isnull().sum()

Out[20]: Name                  0
         URL                   0
         Brand                 0
         Sale Price            0
         Mrp                   0
         Discount Percentage   0
         Number Of Ratings     0
         Number Of Reviews     0
         Star Rating           0
         Ram                   0
         dtype: int64

In [22]: data1=data1.fillna(0)

In [23]: data1=data1.fillna({'Ram':0,'Mrp':1000,'Brand':'Nothing'})
```

Preprocessing: Duplication Check and Drop

```
In [25]: data1.duplicated()

Out[25]: 0       False
         1       False
         2       False
         3       False
         4       False
                 ...
         1508    False
         1509    False
         1510    False
         1511    False
         1512    False
         Length: 1513, dtype: bool
```

```
In [28]: data1=data1.drop_duplicates()
```

```
In [29]: data1=data1.drop_duplicates(['Mrp'])
```

Crosstab

```
In [9]: table2=pd.crosstab(data1.Brand,data1.Ram)

In [10]: table2
```

Out[10]:

Brand \ Ram	1 GB	1.5 GB	12 GB	2 GB	256 MB	3 GB	4 GB	512 MB	6 GB	8 GB
ASUS	4	0	1	2	0	4	2	0	1	1
Alcatel	2	0	0	2	0	6	0	0	0	0
Apple	0	0	0	13	0	1	29	0	19	0
BlackZone	1	0	0	6	0	3	0	0	0	0
Bluboo	0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
Zoom	0	0	0	0	0	2	0	0	0	0
iball	1	0	0	0	0	0	0	0	0	0
mobiistar	0	0	0	0	0	2	0	0	0	0
realme	0	0	6	3	0	17	21	0	17	25
ringme	0	0	0	19	0	0	0	0	0	0

67 rows × 10 columns

Pivot Table

```
In [41]: pd.pivot_table(data1,index=["Ram"],columns=["Brand"],values='Star Rating',aggfunc=np.sum)
```

Out[41]:

Brand / Ram	ASUS	Alcatel	Apple	BlackZone	Bluboo	Brown	Celkon	Coolpad	GIONEE	HPL	...	Voto	Wizphone	XOLO	YU	Zen	Zoom	iball	mobiistar	rea
1 GB	NaN	NaN	NaN	NaN	NaN	NaN	NaN	7.2	7.2	NaN	...	NaN	NaN	7.6	NaN	3.4	NaN	3.3	NaN	N
1.5 GB	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
12 GB	4.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1
2 GB	8.1	NaN	27.2	6.8	3.6	3.5	3.6	NaN	NaN	NaN	...	10.1	6.6	NaN	NaN	NaN	NaN	NaN	NaN	N
256 MB	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.9	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
3 GB	16.7	3.7	NaN	3.1	NaN	NaN	NaN	7.2	12.0	NaN	...	3.7	NaN	NaN	3.8	NaN	0.0	NaN	7.5	N
4 GB	8.4	NaN	55.1	NaN	NaN	NaN	NaN	NaN	12.2	NaN	...	NaN	NaN	NaN	4.0	NaN	NaN	NaN	NaN	
512 MB	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.6	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
6 GB	4.5	NaN	18.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
8 GB	4.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1

10 rows × 57 columns