Name : Asha Belcilda

Roll no : 225229104

Lab6.Pandas Data Cleaning

LabelEncoder in Scikit Learn

In [1]:

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
```

In [2]:

```
le=LabelEncoder()
df=pd.DataFrame(data={'col1':['foo','bar','foo','bar'],'col2':['x','y','x','z'],'col3':[1,2,3,4]})
```

In [3]:

```
df.apply(le.fit_transform)
```

Out[3]:

	col1	col2	col3
0	1	0	0
1	0	1	1
2	1	0	2
3	0	2	3

One Hot Encoder

In [4]:

```
df=pd.DataFrame({'A':['a','b','a'],'B':['b','a','c'],'C':[1,2,3]})
df
```

Out[4]:

	A	B	C
0	a	b	1
1	b	a	2
2	a	c	3

In [5]:

```
pd.get_dummies(df,prefix=['col1','col2'])
```

Out[5]:

	C	col1_a	col1_b	col2_a	col2_b	col2_c
0	1	1	0	0	1	0
1	2	0	1	1	0	0
2	3	1	0	0	0	1

MinMaxScaler

In [6]:

```
from sklearn .preprocessing import MinMaxScaler
mm_scaler=MinMaxScaler(feature_range=(0,1))
df2=pd.DataFrame({'col1':[5,-41,-67],'col2':[23,-53,-36],'col3':[-25,10,17]})
mm_scaler.fit_transform(df2)
```

Out[6]:

```
array([[1.        , 1.        , 0.        ],
       [0.36111111, 0.        , 0.83333333],
       [0.        , 0.22368421, 1.        ]])
```

Binarizer

In [7]:

```
from sklearn.preprocessing import Binarizer
dfb=pd.DataFrame({'col1':[110,200],'col2':[120,800],'col3':[310,400]})
bin=Binarizer(threshold=300)
bin.fit_transform(dfb)
```

Out[7]:

```
array([[0, 0, 1],
       [0, 1, 1]], dtype=int64)
```

Imputer

In [8]:

```
import numpy as np
from sklearn.impute import SimpleImputer
imp_mean=SimpleImputer(missing_values=np.nan,strategy='mean')
df=pd.DataFrame({'col1':[7,2,3],'col2':[4,np.nan,6],'col3':[np.nan,np.nan,3],'col4':[10,np.nan,9]}
print(df)
imp_mean.fit_transform(df)
```

```
   col1  col2  col3  col4
0     7   4.0   NaN  10.0
1     2   NaN   NaN   NaN
2     3   6.0   3.0   9.0
```

Out[8]:

```
array([[ 7. ,  4. ,  3. , 10. ],
       [ 2. ,  5. ,  3. ,  9.5],
       [ 3. ,  6. ,  3. ,  9. ]])
```

De-duplication or Entity Resolution and String Matching

In [9]:

```
pip install dedupe
```

```
Collecting dedupe
  Downloading dedupe-2.0.23-cp39-cp39-win_amd64.whl (96 kB)
Requirement already satisfied: numpy>=1.20 in c:\users\harsmitha\anaconda3\lib\site-
packages (from dedupe) (1.21.5)
Collecting categorical-distance>=1.9
  Downloading categorical_distance-1.9-py3-none-any.whl (3.3 kB)
Collecting BTrees>=4.1.4
  Downloading BTrees-5.0-cp39-cp39-win_amd64.whl (992 kB)
Requirement already satisfied: typing-extensions in c:\users\harsmitha\anaconda3\lib
\site-packages (from dedupe) (4.1.1)
Collecting dedupe-variable-datetime
  Downloading dedupe_variable_datetime-1.0.0-py3-none-any.whl (3.9 kB)
Collecting dedupe-Levenshtein-search
  Downloading dedupe_Levenshtein_search-1.4.5-cp39-cp39-win_amd64.whl (14 kB)
Collecting doublemetaphone
  Downloading DoubleMetaphone-1.1-cp39-cp39-win_amd64.whl (28 kB)
Collecting highered>=0.2.0
  Downloading highered-0.2.1-py2.py3-none-any.whl (3.3 kB)
Collecting affinegap>=1.3
  Downloading affinegap-1.12-cp39-cp39-win_amd64.whl (16 kB)
Collecting simplecosine>=1.2
  Downloading simplecosine-1.2-py2.py3-none-any.whl (3.2 kB)
Collecting zope.index
  Downloading zope.index-5.2.1-cp39-cp39-win_amd64.whl (95 kB)
Requirement already satisfied: scikit-learn in c:\users\harsmitha\anaconda3\lib\site
-packages (from dedupe) (1.0.2)
Collecting haversine>=0.4.1
  Downloading haversine-2.8.0-py2.py3-none-any.whl (7.7 kB)
Collecting persistent>=4.1.0
  Downloading persistent-5.0-cp39-cp39-win_amd64.whl (157 kB)
Requirement already satisfied: zope.interface>=5.0.0 in c:\users\harsmitha\anaconda3
\lib\site-packages (from BTrees>=4.1.4->dedupe) (5.4.0)
Collecting pyhacrf-datamade>=0.2.0
  Downloading pyhacrf_datamade-0.2.6-cp39-cp39-win_amd64.whl (184 kB)
Requirement already satisfied: cffi in c:\users\harsmitha\anaconda3\lib\site-package
s (from persistent>=4.1.0->BTrees>=4.1.4->dedupe) (1.15.0)
Collecting PyLBFGS>=0.1.3
  Downloading PyLBFGS-0.2.0.14-cp39-cp39-win_amd64.whl (54 kB)
Requirement already satisfied: setuptools in c:\users\harsmitha\anaconda3\lib\site-p
ackages (from zope.interface>=5.0.0->BTrees>=4.1.4->dedupe) (61.2.0)
Requirement already satisfied: pycparser in c:\users\harsmitha\anaconda3\lib\site-pa
ckages (from cffi->persistent>=4.1.0->BTrees>=4.1.4->dedupe) (2.21)
Collecting dedupe-variable-datetime
  Downloading dedupe_variable_datetime-0.1.5-py3-none-any.whl (4.8 kB)
Requirement already satisfied: future in c:\users\harsmitha\anaconda3\lib\site-packa
ges (from dedupe-variable-datetime->dedupe) (0.18.2)
Collecting datetime-distance
  Downloading datetime_distance-0.1.3-py3-none-any.whl (4.1 kB)
Requirement already satisfied: python-dateutil>=2.6.0 in c:\users\harsmitha\anaconda
3\lib\site-packages (from datetime-distance->dedupe-variable-datetime->dedupe) (2.8.
2)
Requirement already satisfied: six>=1.5 in c:\users\harsmitha\anaconda3\lib\site-pac
kages (from python-dateutil>=2.6.0->datetime-distance->dedupe-variable-datetime->ded
upe) (1.16.0)
Requirement already satisfied: joblib>=0.11 in c:\users\harsmitha\anaconda3\lib\site
-packages (from scikit-learn->dedupe) (1.1.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\harsmitha\anaconda3
\lib\site-packages (from scikit-learn->dedupe) (2.2.0)
Requirement already satisfied: scipy>=1.1.0 in c:\users\harsmitha\anaconda3\lib\site
-packages (from scikit-learn->dedupe) (1.7.3)
Installing collected packages: PyLBFGS, persistent, pyhacrf-datamade, datetime-dista
nce, BTrees, zope.index, simplecosine, highered, haversine, doublemetaphone, dedupe-
variable-datetime, dedupe-Levenshtein-search, categorical-distance, affinegap, dedup
e
Successfully installed BTrees-5.0 PyLBFGS-0.2.0.14 affinegap-1.12 categorical-distan
ce-1.9 datetime-distance-0.1.3 dedupe-2.0.23 dedupe-Levenshtein-search-1.4.5 dedupe-
```

```
variable-datetime-0.1.5 doublemetaphone-1.1 haversine-2.8.0 highered-0.2.1 persisten
t-5.0 pyhacrf-datamade-0.2.6 simplecosine-1.2 zope.index-5.2.1
Note: you may need to restart the kernel to use updated packages.
```

In [10]:

```
pip install fuzzywuzzy
```

```
Collecting fuzzywuzzy
  Downloading fuzzywuzzy-0.18.0-py2.py3-none-any.whl (18 kB)
Installing collected packages: fuzzywuzzy
Successfully installed fuzzywuzzy-0.18.0
Note: you may need to restart the kernel to use updated packages.
```

In [11]:

```
import dedupe
```

In [12]:

```
import fuzzywuzzy
```

In [ ]:

```

```