e-scaling-and-feature-selection-1

October 4, 2024

[3]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score,confusion_matrix
df=pd.read_csv("/content/income(1) (1).csv",na_values=[" ?"])
df
```

[3]:

	age	JobType	EdType	maritalstatus	\
0	45	Private	HS-grad	Divorced	
1	24	Federal-gov	HS-grad	Never-married	
2	44	Private	Some-college	Married-civ-spouse	
3	27	Private	9th	Never-married	
4	20	Private	Some-college	Never-married	
…	…	…	…	…	
31973	34	Local-gov	HS-grad	Never-married	
31974	34	Local-gov	Some-college	Never-married	
31975	23	Private	Some-college	Married-civ-spouse	
31976	42	Local-gov	Some-college	Married-civ-spouse	
31977	29	Private	Bachelors	Never-married	

	occupation	relationship	race	gender	capitalgain	\
0	Adm-clerical	Not-in-family	White	Female	0	
1	Armed-Forces	Own-child	White	Male	0	
2	Prof-specialty	Husband	White	Male	0	
3	Craft-repair	Other-relative	White	Male	0	
4	Sales	Not-in-family	White	Male	0	
…	…	…	…	…	…	
31973	Farming-fishing	Not-in-family	Black	Male	594	
31974	Protective-serv	Not-in-family	White	Female	0	
31975	Adm-clerical	Husband	White	Male	0	
31976	Adm-clerical	Wife	White	Female	0	
31977	Prof-specialty	Not-in-family	White	Male	0	

	capitalloss	hoursperweek	nativecountry	\

```
0                0            28   United-States
1                0            40   United-States
2                0            40   United-States
3                0            40          Mexico
4                0            35   United-States
…              …            …              …
31973            0            60   United-States
31974            0            40   United-States
31975            0            40   United-States
31976            0            40   United-States
31977            0            40   United-States

                              SalStat
0       less than or equal to 50,000
1       less than or equal to 50,000
2                greater than 50,000
3       less than or equal to 50,000
4       less than or equal to 50,000
…                               …
31973   less than or equal to 50,000
31974   less than or equal to 50,000
31975   less than or equal to 50,000
31976   less than or equal to 50,000
31977   less than or equal to 50,000

[31978 rows x 13 columns]
```

[4]:
```
df.isnull().sum()
```

[4]:
```
age                 0
JobType          1809
EdType              0
maritalstatus       0
occupation       1816
relationship        0
race                0
gender              0
capitalgain         0
capitalloss         0
hoursperweek        0
nativecountry       0
SalStat             0
dtype: int64
```

[5]:
```
missing=df[df.isnull().any(axis=1)]
missing
```

[5]:
```
       age JobType          EdType        maritalstatus occupation  \
8       17     NaN            11th        Never-married        NaN
17      32     NaN    Some-college   Married-civ-spouse        NaN
29      22     NaN    Some-college        Never-married        NaN
42      52     NaN            12th        Never-married        NaN
44      63     NaN         1st-4th   Married-civ-spouse        NaN
…     …      …              …                  …        …
31892   59     NaN       Bachelors   Married-civ-spouse        NaN
31934   20     NaN         HS-grad        Never-married        NaN
31945   28     NaN    Some-college   Married-civ-spouse        NaN
31967   80     NaN         HS-grad              Widowed        NaN
31968   17     NaN            11th        Never-married        NaN

          relationship    race   gender  capitalgain  capitalloss  \
8            Own-child   White   Female            0            0
17             Husband   White     Male            0            0
29           Own-child   White     Male            0            0
42     Other-relative   Black     Male          594            0
44             Husband   White     Male            0            0
…                …      …       …           …           …
31892          Husband   White     Male            0            0
31934  Other-relative   White   Female            0            0
31945            Wife   White   Female            0         1887
31967   Not-in-family   White     Male            0            0
31968        Own-child   White     Male            0            0

       hoursperweek   nativecountry                         SalStat
8                 5   United-States   less than or equal to 50,000
17               40   United-States   less than or equal to 50,000
29               40   United-States   less than or equal to 50,000
42               40   United-States   less than or equal to 50,000
44               35   United-States   less than or equal to 50,000
…              …             …                                …
31892            40   United-States            greater than 50,000
31934            35   United-States   less than or equal to 50,000
31945            40   United-States            greater than 50,000
31967            24   United-States   less than or equal to 50,000
31968            40   United-States   less than or equal to 50,000

[1816 rows x 13 columns]
```

[6]:
```
df2=df.dropna(axis=0)
df2
```

[6]:
```
       age       JobType         EdType        maritalstatus  \
0       45       Private        HS-grad             Divorced
1       24   Federal-gov        HS-grad        Never-married
```

```
2       44       Private   Some-college   Married-civ-spouse
3       27       Private            9th        Never-married
4       20       Private   Some-college        Never-married
…    …           …              …                 …
31973   34     Local-gov        HS-grad        Never-married
31974   34     Local-gov   Some-college        Never-married
31975   23       Private   Some-college   Married-civ-spouse
31976   42     Local-gov   Some-college   Married-civ-spouse
31977   29       Private      Bachelors        Never-married

             occupation     relationship    race   gender  capitalgain  \
0          Adm-clerical    Not-in-family   White   Female            0
1          Armed-Forces        Own-child   White     Male            0
2        Prof-specialty          Husband   White     Male            0
3          Craft-repair   Other-relative   White     Male            0
4                 Sales    Not-in-family   White     Male            0
…                   …              …     …       …          …
31973   Farming-fishing    Not-in-family   Black     Male          594
31974   Protective-serv    Not-in-family   White   Female            0
31975      Adm-clerical          Husband   White     Male            0
31976      Adm-clerical             Wife   White   Female            0
31977    Prof-specialty    Not-in-family   White     Male            0

       capitalloss  hoursperweek   nativecountry  \
0                0            28   United-States
1                0            40   United-States
2                0            40   United-States
3                0            40          Mexico
4                0            35   United-States
…              …            …              …
31973            0            60   United-States
31974            0            40   United-States
31975            0            40   United-States
31976            0            40   United-States
31977            0            40   United-States

                            SalStat
0       less than or equal to 50,000
1       less than or equal to 50,000
2                greater than 50,000
3       less than or equal to 50,000
4       less than or equal to 50,000
…                             …
31973   less than or equal to 50,000
31974   less than or equal to 50,000
31975   less than or equal to 50,000
31976   less than or equal to 50,000
```

```
31977   less than or equal to 50,000

[30162 rows x 13 columns]
```

[7]:
```
sal=df['SalStat']
```

[8]:
```
df2['SalStat']=df2['SalStat'].map({' less than or equal to 50,000':0,' greater
 ↪than 50,000':1})
print(df2['SalStat'])
```

```
0        0
1        0
2        1
3        0
4        0
        ..
31973    0
31974    0
31975    0
31976    0
31977    0
Name: SalStat, Length: 30162, dtype: int64
```

```
<ipython-input-8-4b0b6c30c323>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df2['SalStat']=df2['SalStat'].map({' less than or equal to 50,000':0,' greater
than 50,000':1})
```

[9]:
```
sal2=df2['SalStat']
sal2
```

[9]:
```
0        0
1        0
2        1
3        0
4        0
        ..
31973    0
31974    0
31975    0
31976    0
31977    0
Name: SalStat, Length: 30162, dtype: int64
```

[10]:
```
sal3=pd.concat([sal,sal2],axis=1)
sal3
```

[10]:

```
                            SalStat  SalStat
0       less than or equal to 50,000      0.0
1       less than or equal to 50,000      0.0
2                greater than 50,000      1.0
3       less than or equal to 50,000      0.0
4       less than or equal to 50,000      0.0
…                                 …      …
31973   less than or equal to 50,000      0.0
31974   less than or equal to 50,000      0.0
31975   less than or equal to 50,000      0.0
31976   less than or equal to 50,000      0.0
31977   less than or equal to 50,000      0.0

[31978 rows x 2 columns]
```

[11]:
```
df2
```

[11]:

```
       age       JobType         EdType        maritalstatus  \
0       45       Private        HS-grad             Divorced
1       24   Federal-gov        HS-grad        Never-married
2       44       Private   Some-college   Married-civ-spouse
3       27       Private            9th        Never-married
4       20       Private   Some-college        Never-married
…    …           …              …                 …
31973   34     Local-gov        HS-grad        Never-married
31974   34     Local-gov   Some-college        Never-married
31975   23       Private   Some-college   Married-civ-spouse
31976   42     Local-gov   Some-college   Married-civ-spouse
31977   29       Private      Bachelors        Never-married

             occupation     relationship    race   gender  capitalgain  \
0          Adm-clerical    Not-in-family   White   Female            0
1          Armed-Forces        Own-child   White     Male            0
2       Prof-specialty          Husband   White     Male            0
3          Craft-repair   Other-relative   White     Male            0
4                Sales    Not-in-family   White     Male            0
…                    …              …     …       …          …
31973   Farming-fishing    Not-in-family   Black     Male          594
31974   Protective-serv    Not-in-family   White   Female            0
31975      Adm-clerical          Husband   White     Male            0
31976      Adm-clerical             Wife   White   Female            0
31977    Prof-specialty    Not-in-family   White     Male            0

       capitalloss  hoursperweek   nativecountry  SalStat
```

```
0                0            28   United-States        0
1                0            40   United-States        0
2                0            40   United-States        1
3                0            40          Mexico        0
4                0            35   United-States        0
…              …            …            …      …
31973            0            60   United-States        0
31974            0            40   United-States        0
31975            0            40   United-States        0
31976            0            40   United-States        0
31977            0            40   United-States        0

[30162 rows x 13 columns]
```

[12]:
```
new_data=pd.get_dummies(df2,drop_first=True)
new_data
```

[12]:
```
       age  capitalgain  capitalloss  hoursperweek  SalStat  \
0       45            0            0            28        0
1       24            0            0            40        0
2       44            0            0            40        1
3       27            0            0            40        0
4       20            0            0            35        0
…    …          …          …            …       …
31973   34          594            0            60        0
31974   34            0            0            40        0
31975   23            0            0            40        0
31976   42            0            0            40        0
31977   29            0            0            40        0

       JobType_ Local-gov  JobType_ Private  JobType_ Self-emp-inc  \
0                   False             True                  False
1                   False            False                  False
2                   False             True                  False
3                   False             True                  False
4                   False             True                  False
…                   …               …                     …
31973                True            False                  False
31974                True            False                  False
31975               False             True                  False
31976                True            False                  False
31977               False             True                  False

       JobType_ Self-emp-not-inc  JobType_ State-gov  …  \
0                          False              False  …
1                          False              False  …
2                          False              False  …
```

```
3                          False                False  …
4                          False                False  …
…                          …                  …  …
31973                      False                False  …
31974                      False                False  …
31975                      False                False  …
31976                      False                False  …
31977                      False                False  …

       nativecountry_ Portugal  nativecountry_ Puerto-Rico  \
0                         False                       False
1                         False                       False
2                         False                       False
3                         False                       False
4                         False                       False
…                         …                         …
31973                     False                       False
31974                     False                       False
31975                     False                       False
31976                     False                       False
31977                     False                       False

       nativecountry_ Scotland  nativecountry_ South  nativecountry_ Taiwan  \
0                         False                False                  False
1                         False                False                  False
2                         False                False                  False
3                         False                False                  False
4                         False                False                  False
…                         …                  …                      …
31973                     False                False                  False
31974                     False                False                  False
31975                     False                False                  False
31976                     False                False                  False
31977                     False                False                  False

       nativecountry_ Thailand  nativecountry_ Trinadad&Tobago  \
0                         False                           False
1                         False                           False
2                         False                           False
3                         False                           False
4                         False                           False
…                         …                             …
31973                     False                           False
31974                     False                           False
31975                     False                           False
31976                     False                           False
31977                     False                           False
```

```
       nativecountry_ United-States  nativecountry_ Vietnam  \
0                              True                   False
1                              True                   False
2                              True                   False
3                             False                   False
4                              True                   False
…                              …                      …
31973                          True                   False
31974                          True                   False
31975                          True                   False
31976                          True                   False
31977                          True                   False

       nativecountry_ Yugoslavia
0                          False
1                          False
2                          False
3                          False
4                          False
…                          …
31973                      False
31974                      False
31975                      False
31976                      False
31977                      False

[30162 rows x 95 columns]
```

[13]:
```
columns_list=list(new_data.columns)
columns_list
```

[13]:
```
['age',
 'capitalgain',
 'capitalloss',
 'hoursperweek',
 'SalStat',
 'JobType_ Local-gov',
 'JobType_ Private',
 'JobType_ Self-emp-inc',
 'JobType_ Self-emp-not-inc',
 'JobType_ State-gov',
 'JobType_ Without-pay',
 'EdType_ 11th',
 'EdType_ 12th',
 'EdType_ 1st-4th',
 'EdType_ 5th-6th',
```

```
'EdType_ 7th-8th',
'EdType_ 9th',
'EdType_ Assoc-acdm',
'EdType_ Assoc-voc',
'EdType_ Bachelors',
'EdType_ Doctorate',
'EdType_ HS-grad',
'EdType_ Masters',
'EdType_ Preschool',
'EdType_ Prof-school',
'EdType_ Some-college',
'maritalstatus_ Married-AF-spouse',
'maritalstatus_ Married-civ-spouse',
'maritalstatus_ Married-spouse-absent',
'maritalstatus_ Never-married',
'maritalstatus_ Separated',
'maritalstatus_ Widowed',
'occupation_ Armed-Forces',
'occupation_ Craft-repair',
'occupation_ Exec-managerial',
'occupation_ Farming-fishing',
'occupation_ Handlers-cleaners',
'occupation_ Machine-op-inspct',
'occupation_ Other-service',
'occupation_ Priv-house-serv',
'occupation_ Prof-specialty',
'occupation_ Protective-serv',
'occupation_ Sales',
'occupation_ Tech-support',
'occupation_ Transport-moving',
'relationship_ Not-in-family',
'relationship_ Other-relative',
'relationship_ Own-child',
'relationship_ Unmarried',
'relationship_ Wife',
'race_ Asian-Pac-Islander',
'race_ Black',
'race_ Other',
'race_ White',
'gender_ Male',
'nativecountry_ Canada',
'nativecountry_ China',
'nativecountry_ Columbia',
'nativecountry_ Cuba',
'nativecountry_ Dominican-Republic',
'nativecountry_ Ecuador',
'nativecountry_ El-Salvador',
```

```
 'nativecountry_ England',
 'nativecountry_ France',
 'nativecountry_ Germany',
 'nativecountry_ Greece',
 'nativecountry_ Guatemala',
 'nativecountry_ Haiti',
 'nativecountry_ Holand-Netherlands',
 'nativecountry_ Honduras',
 'nativecountry_ Hong',
 'nativecountry_ Hungary',
 'nativecountry_ India',
 'nativecountry_ Iran',
 'nativecountry_ Ireland',
 'nativecountry_ Italy',
 'nativecountry_ Jamaica',
 'nativecountry_ Japan',
 'nativecountry_ Laos',
 'nativecountry_ Mexico',
 'nativecountry_ Nicaragua',
 'nativecountry_ Outlying-US(Guam-USVI-etc)',
 'nativecountry_ Peru',
 'nativecountry_ Philippines',
 'nativecountry_ Poland',
 'nativecountry_ Portugal',
 'nativecountry_ Puerto-Rico',
 'nativecountry_ Scotland',
 'nativecountry_ South',
 'nativecountry_ Taiwan',
 'nativecountry_ Thailand',
 'nativecountry_ Trinadad&Tobago',
 'nativecountry_ United-States',
 'nativecountry_ Vietnam',
 'nativecountry_ Yugoslavia']
```

[14]:
```
features=list(set(columns_list)-set(['SalStat']))
features
```

[14]:
```
['nativecountry_ Hungary',
 'JobType_ Local-gov',
 'nativecountry_ Japan',
 'EdType_ 1st-4th',
 'maritalstatus_ Widowed',
 'nativecountry_ Ecuador',
 'EdType_ 11th',
 'EdType_ Assoc-voc',
 'EdType_ Bachelors',
 'nativecountry_ China',
```

```
'race_ Asian-Pac-Islander',
'EdType_ Prof-school',
'nativecountry_ Hong',
'nativecountry_ Peru',
'relationship_ Wife',
'occupation_ Exec-managerial',
'nativecountry_ Nicaragua',
'nativecountry_ Outlying-US(Guam-USVI-etc)',
'nativecountry_ Iran',
'maritalstatus_ Separated',
'nativecountry_ Taiwan',
'gender_ Male',
'nativecountry_ England',
'nativecountry_ Thailand',
'occupation_ Sales',
'age',
'nativecountry_ India',
'occupation_ Armed-Forces',
'EdType_ Some-college',
'occupation_ Other-service',
'nativecountry_ Poland',
'nativecountry_ Holand-Netherlands',
'occupation_ Craft-repair',
'nativecountry_ Canada',
'nativecountry_ Germany',
'race_ White',
'occupation_ Machine-op-inspct',
'occupation_ Farming-fishing',
'EdType_ 5th-6th',
'maritalstatus_ Married-AF-spouse',
'relationship_ Not-in-family',
'nativecountry_ Dominican-Republic',
'JobType_ Private',
'nativecountry_ Trinadad&Tobago',
'occupation_ Protective-serv',
'EdType_ Preschool',
'capitalloss',
'nativecountry_ Haiti',
'nativecountry_ Columbia',
'nativecountry_ Italy',
'nativecountry_ Vietnam',
'maritalstatus_ Married-spouse-absent',
'nativecountry_ Yugoslavia',
'nativecountry_ Ireland',
'maritalstatus_ Married-civ-spouse',
'nativecountry_ Greece',
'occupation_ Transport-moving',
```

```
 'nativecountry_ Cuba',
 'nativecountry_ Guatemala',
 'nativecountry_ Jamaica',
 'EdType_ Doctorate',
 'JobType_ State-gov',
 'EdType_ HS-grad',
 'occupation_ Handlers-cleaners',
 'nativecountry_ El-Salvador',
 'race_ Other',
 'EdType_ 7th-8th',
 'nativecountry_ Mexico',
 'EdType_ 9th',
 'nativecountry_ South',
 'relationship_ Unmarried',
 'occupation_ Tech-support',
 'JobType_ Self-emp-not-inc',
 'hoursperweek',
 'occupation_ Prof-specialty',
 'JobType_ Without-pay',
 'JobType_ Self-emp-inc',
 'relationship_ Other-relative',
 'capitalgain',
 'maritalstatus_ Never-married',
 'nativecountry_ United-States',
 'nativecountry_ Scotland',
 'relationship_ Own-child',
 'nativecountry_ Philippines',
 'EdType_ 12th',
 'nativecountry_ France',
 'race_ Black',
 'EdType_ Masters',
 'nativecountry_ Puerto-Rico',
 'nativecountry_ Portugal',
 'occupation_ Priv-house-serv',
 'EdType_ Assoc-acdm',
 'nativecountry_ Laos',
 'nativecountry_ Honduras']
```

[15]:
```
y=new_data['SalStat'].values
print(y)
x=new_data[features].values
print(x)
```

```
[0 0 1 … 0 0 0]
[[False False False … False False False]
 [False False False … False False False]
 [False False False … False False False]
```

```
…
[False False False … False False False]
[False True False … False False False]
[False False False … False False False]]
```

[16]:
```
train_x,test_x,train_y,test_y=train_test_split(x,y,test_size=0.3,random_state=0)
```

[18]:
```
KNN_classifier=KNeighborsClassifier(n_neighbors=5)
```

[19]:
```
KNN_classifier.fit(train_x,train_y)
```

[19]:
```
KNeighborsClassifier()
```

[20]:
```
prediction=KNN_classifier.predict(test_x)
```

[21]:
```
confusionmatrix=confusion_matrix(test_y,prediction)
confusionmatrix
```

[21]:
```
array([[6176,  647],
       [ 808, 1418]])
```

[22]:
```
accuracy=accuracy_score(test_y,prediction)
accuracy
```

[22]:
```
0.8392087523483258
```

[24]:
```
print("missclassified data:",(test_y!=prediction).sum())
```

```
missclassified data: 1455
```

[26]:
```
new_data.shape
```

[26]:
```
(30162, 95)
```