feature-selection

October 4, 2024

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
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
```

[1]:

	age	JobType	EdType	maritalstatus	occupation	relationship	race	gender	capitalgain	capitalloss	hoursperweek	nativecountry
0	45	Private	HS-grad	Divorced	Adm-clerical	Not-in-family	White	Female	0	0	28	United-States
1	24	Federal-gov	HS-grad	Never-married	Armed-Forces	Own-child	White	Male	0	0	40	United-States
2	44	Private	Some-college	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States
3	27	Private	9th	Never-married	Craft-repair	Other-relative	White	Male	0	0	40	United-States
4	20	Private	Some-college	Never-married	Sales	Not-in-family	White	Male	0	0	40	United-States
…	…	…	…	…	…	…	…	…	…	…	…	…
31973	34	Local-gov	HS-grad	Never-married	Farming-fishing	Not-in-family	Black	Male	594	0	40	United-States
31974	34	Local-gov	Some-college	Never-married	Protective-serv	Not-in-family	White	Female	0	0	40	United-States
31975	23	Private	Some-college	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	40	United-States
31976	42	Local-gov	Some-college	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	40	United-States
31977	29	Private	Bachelors	Never-married	Prof-specialty	Not-in-family	White	Male	0	0	40	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]
```

[2]:
```
df.isnull().sum()
```

[2]:
```
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
```

[3]:
```
missing=df[df.isnull().any(axis=1)]
missing
```

[3]:
```
       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]
```

[4]:
```
df1=df.dropna(axis=0)
df1
```

[4]:
```
       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]
```

[5]:
```
sal=df['SalStat']
```

[6]:
```
df1['SalStat']=df1['SalStat'].map({' less than or equal to 50,000':0,' greater␣
 ↪than 50,000':1})
print(df1['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-6-7d934405fb92>: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
  df1['SalStat']=df1['SalStat'].map({' less than or equal to 50,000':0,' greater
than 50,000':1})
```

[7]:
```
sal2=df1['SalStat']
dfs=pd.concat([sal,sal2],axis=1)
dfs
```

[7]:

	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]
```

[8]:
```
df1
```

[8]:
```
       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]
```

[9]:
```
new_data=pd.get_dummies(df1, drop_first=True)
new_data
```

[9]:

```
       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]
```

[10]:
```
columns_list=list(new_data.columns)
print(columns_list)
```

```
['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']
```

[11]:
```
features=list(set(columns_list)-set(['SalStat']))
print(features)
```

```
['maritalstatus_ Married-civ-spouse', 'nativecountry_ Italy', 'EdType_ Masters',
'nativecountry_ Hungary', 'EdType_ 11th', 'nativecountry_ Japan', 'occupation_
Craft-repair', 'relationship_ Unmarried', 'race_ Black', 'EdType_ Assoc-voc',
'EdType_ HS-grad', 'JobType_ Self-emp-not-inc', 'EdType_ Some-college',
'nativecountry_ Ireland', 'maritalstatus_ Married-AF-spouse', 'nativecountry_
England', 'EdType_ 1st-4th', 'nativecountry_ Outlying-US(Guam-USVI-etc)',
'occupation_ Protective-serv', 'nativecountry_ Philippines', 'nativecountry_
Hong', 'nativecountry_ Guatemala', 'nativecountry_ Haiti', 'JobType_ Local-gov',
'relationship_ Other-relative', 'maritalstatus_ Widowed', 'occupation_ Priv-
house-serv', 'nativecountry_ Honduras', 'race_ White', 'nativecountry_
Scotland', 'nativecountry_ Iran', 'occupation_ Transport-moving', 'race_ Asian-
Pac-Islander', 'nativecountry_ Jamaica', 'nativecountry_ Yugoslavia',
'relationship_ Not-in-family', 'nativecountry_ Greece', 'nativecountry_ Taiwan',
'nativecountry_ Nicaragua', 'occupation_ Armed-Forces', 'JobType_ Self-emp-inc',
'nativecountry_ India', 'nativecountry_ South', 'nativecountry_ Mexico',
'occupation_ Machine-op-inspct', 'nativecountry_ China', 'relationship_ Wife',
'gender_ Male', 'occupation_ Farming-fishing', 'nativecountry_ Columbia',
'nativecountry_ Canada', 'nativecountry_ France', 'EdType_ 9th', 'EdType_
5th-6th', 'occupation_ Prof-specialty', 'capitalloss', 'maritalstatus_ Married-
spouse-absent', 'nativecountry_ Germany', 'nativecountry_ Portugal', 'JobType_
Private', 'race_ Other', 'age', 'nativecountry_ Trinadad&Tobago',
'nativecountry_ Holand-Netherlands', 'EdType_ Bachelors', 'EdType_ Prof-school',
'nativecountry_ Dominican-Republic', 'relationship_ Own-child', 'nativecountry_
Poland', 'nativecountry_ Thailand', 'EdType_ 7th-8th', 'nativecountry_ United-
States', 'maritalstatus_ Separated', 'nativecountry_ Laos', 'nativecountry_
Ecuador', 'hoursperweek', 'occupation_ Exec-managerial', 'occupation_ Sales',
'occupation_ Other-service', 'nativecountry_ Vietnam', 'JobType_ Without-pay',
'capitalgain', 'nativecountry_ El-Salvador', 'occupation_ Handlers-cleaners',
'occupation_ Tech-support', 'nativecountry_ Peru', 'maritalstatus_ Never-
married', 'nativecountry_ Puerto-Rico', 'EdType_ Doctorate', 'EdType_ 12th',
'nativecountry_ Cuba', 'JobType_ State-gov', 'EdType_ Preschool', 'EdType_
Assoc-acdm']
```

[12]:
```
y=new_data['SalStat'].values
```

[13]:
```
print(y)
```

```
[0 0 1 … 0 0 0]
```

[14]:
```
x= new_data[features].values
print(x)
```

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

[15]:
```
train_x,test_x,train_y,test_y=train_test_split(x,y,test_size=0.3,random_state=0)
```

[16]:
```
KNN_classifier=KNeighborsClassifier(n_neighbors=5)
```

[17]:
```
KNN_classifier.fit(train_x,train_y)
```

[17]:
```
KNeighborsClassifier()
```

[19]:
```
prediction=KNN_classifier.predict(test_x)
```

[20]:
```
confusionmatrix=confusion_matrix(test_y,prediction)
print(confusionmatrix)
```

```
[[6176  647]
 [ 808 1418]]
```

[21]:
```
accuracy_score=accuracy_score(test_y,prediction)
print(accuracy_score)
```

```
0.8392087523483258
```

[22]:
```
print("Misclassied smples: %d" % (test_y !=prediction).sum())
```

```
Misclassied smples: 1455
```

[24]:
```
df.shape
```

[24]:
```
(31978, 13)
```