Ex.No:3

Date:18.03.2022

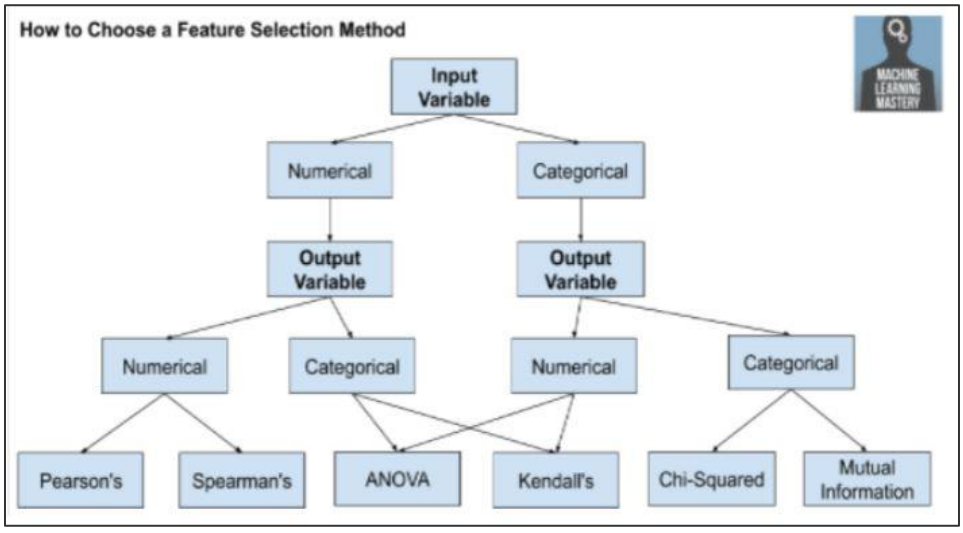
FEATURE SELECTION METHODS

AIM:

To select the features from the given dataset using any two of the appropriate feature selection methods and reduce the size of the dataset attributes.

FEATURE SELECTION METHODS:

- ✓ Choose the appropriate feature selection methods from the table given below, based on the dataset taken (input variable and output variable type).



- ✓ The data chosen was weather dataset which has numerical input data and categorical output, thus the methods chosen were "ANOVA" and "Kendall's".

1) ANOVA - (Analysis of Variance):

- It is a parametric statistical hypothesis test for determining whether the means from two or more samples of data (often three or more) come from the same distribution or not.
- It checks the impact of various factors by comparing groups (samples) on the basis of their respective mean.
- We can use this only when:
  1. The samples have a normal distribution.
  2. The samples are selected at random and should be independent of one another.
  3. All groups have equal standard deviations.
- One-way anova is a type of hypothesis test where only one factor is considered. We use F-statistic to perform a one-way analysis of variance.

2) **Kendall's:**

- The strength of the association between two variables is known as the correlation test.
- Kendall's method is rank-based correlation coefficients, are known as non-parametric correlation.
- Kendall Rank Correlation Coefficient formula:

$$\tau = \frac{\text{Number of concordant pairs-Number of discordant pairs}}{n(n-1)/2}$$

- Where,
  - Concordant Pair: A pair of observations (x1, y1) and (x2, y2) that follows the property: x1>x2 and y1>y2 (or) x1<x2 and y1<y2
  - Discordant Pair: A pair of observations (x1, y1) and (x2, y2) that follows the property: x1>x2 and y1<y2 (or) x1<x2 and y1>y2
  - n: Total number of samples
  - The pair for which x1=x2 and y1=y2 are not classified as concordant or discordant and are ignored.

Correlation Coefficient for a Direct Relationship	Correlation Coefficient for an Indirect Relationship	Relationship Strength of the Variables
0.0	0.0	None/trivial
0.1	−0.1	Weak/small
0.3	−0.3	Moderate/medium
0.5	−0.5	Strong/large
1.0	−1.0	Perfect

DATASET DESCRIPTION:

- Name : weatherHistory.csv
- Link : https://www.kaggle.com/datasets/muthuj7/weather-dataset
- Number of Rows : 96454
- Number of Columns : 8
- Columns Name : Type
  - Temperature (C) : Float
  - Apparent Temperature (C) : Float
  - Humidity : Float
  - Wind Speed (km/h) : Float
  - Wind Bearing (degrees) : integer
  - Visibility (km) : Float
  - Pressure (millibars) : Float
  - Summary : Categorical
- Class Labels:
  - Partly Cloudy, Mostly Cloudy, Overcast, Foggy, Clear, Breezy, Dry, Windy, Drizzle, Rain.

IN-BUILT FUNCTIONS & PACKAGES:

1) ANOVA:

- The **sklearn.feature_selection.SelectKBest -** Select features according to the k highest scores.
- The **sklearn.feature_selection.f_classif** - Compute the ANOVA F-value for the provided sample.
- The **SelectKBest(score_func=<function f_classif>, k=3) –** K->Number of top features to select.
- The fit_transform(X[, y]) - Fit to data, then transform it, Fits transformer to X and y with optional parameters fit_params and returns a transformed version of X.

2) Kendall's:

- The **numpy.random.rand –** random.rand(d0, d1, ..., dn) **-**Used to generate random values in a given shape, Create an array of the given shape and populate it with random samples from a uniform distribution over [0, 1).
- The **numpy.random.seed** - random.seed(self, seed=None) - Reseed a legacy MT19937 BitGenerator.
- The **scipy.stats.kendalltau -** Calculate Kendall's tau, a correlation measure for ordinal data, Kendall's tau is a measure of the correspondence between two rankings. Values close to 1 indicate strong agreement, and values close to -1 indicate strong disagreement.
- The **pandas -** The import pandas portion of the code tells Python to bring the pandas data analysis library into your current environment.
- The **pandas.read_csv -**used to read the .csv file and store it in an variable.

PROGRAM:

1) ANOVA - CODE:

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
import pandas as pd
#input nuerical, output categorical
dataset=pd.read_csv("weatherHistory.csv")
print(dataset)
#storing and removing the output class from the dataset
TargetClass = dataset[["Summary"]]
dataset.pop("Summary")
print(dataset) #after removal of output class
featureSelection = SelectKBest(score_func=f_classif, k=3) #no. of features to be reduced
Feature_selected = featureSelection.fit_transform(dataset, TargetClass["Summary"])
print("Original dataset          : ",dataset.shape)
print("Features Selected Dataset : ",Feature_selected.shape)
Feature_selected #feature that are selected after peforming anova
```

OUTPUT:

```
       Temperature (C)  Apparent Temperature (C)  Humidity  Wind Speed (km/h)  \
0             9.472222                  7.388889      0.89            14.1197
1             9.355556                  7.227778      0.86            14.2646
2             9.377778                  9.377778      0.89             3.9284
3             8.288889                  5.944444      0.83            14.1036
4             8.755556                  6.977778      0.83            11.0446
...                ...                       ...       ...                ...
96448        26.016667                 26.016667      0.43            10.9963
96449        24.583333                 24.583333      0.48            10.0947
96450        22.038889                 22.038889      0.56             8.9838
96451        21.522222                 21.522222      0.60            10.5294
96452        20.438889                 20.438889      0.61             5.8765

       Wind Bearing (degrees)  Visibility (km)  Pressure (millibars)  \
0                         251          15.8263               1015.13
1                         259          15.8263               1015.63
2                         204          14.9569               1015.94
3                         269          15.8263               1016.41
4                         259          15.8263               1016.51
...                       ...              ...                   ...
96448                      31          16.1000               1014.36
96449                      20          15.5526               1015.16
96450                      30          16.1000               1015.66
96451                      20          16.1000               1015.95
96452                      39          15.5204               1016.16

            Summary
0      Partly Cloudy
1      Partly Cloudy
2      Mostly Cloudy
3      Partly Cloudy
4      Mostly Cloudy
...              ...
96448  Partly Cloudy
96449  Partly Cloudy
96450  Partly Cloudy
96451  Partly Cloudy
96452  Partly Cloudy
```

```
[96453 rows x 8 columns]
       Temperature (C)  Apparent Temperature (C)  Humidity  Wind Speed (km/h)  \
0             9.472222                  7.388889      0.89            14.1197
1             9.355556                  7.227778      0.86            14.2646
2             9.377778                  9.377778      0.89             3.9284
3             8.288889                  5.944444      0.83            14.1036
4             8.755556                  6.977778      0.83            11.0446
...                ...                       ...       ...                ...
96448        26.016667                 26.016667      0.43            10.9963
96449        24.583333                 24.583333      0.48            10.0947
96450        22.038889                 22.038889      0.56             8.9838
96451        21.522222                 21.522222      0.60            10.5294
96452        20.438889                 20.438889      0.61             5.8765

       Wind Bearing (degrees)  Visibility (km)  Pressure (millibars)
0                         251          15.8263               1015.13
1                         259          15.8263               1015.63
2                         204          14.9569               1015.94
3                         269          15.8263               1016.41
4                         259          15.8263               1016.51
...                       ...              ...                   ...
96448                      31          16.1000               1014.36
96449                      20          15.5526               1015.16
96450                      30          16.1000               1015.66
96451                      20          16.1000               1015.95
96452                      39          15.5204               1016.16

[96453 rows x 7 columns]
Original dataset          :  (96453, 7)
Features Selected Dataset :  (96453, 3)
[3]: array([[ 0.89  , 14.1197, 15.8263],
            [ 0.86  , 14.2646, 15.8263],
            [ 0.89  ,  3.9284, 14.9569],
            ...,
            [ 0.56  ,  8.9838, 16.1   ],
            [ 0.6   , 10.5294, 16.1   ],
            [ 0.61  ,  5.8765, 15.5204]])
```

2) KENDALL'S - CODE:

```
from numpy.random import rand
from numpy.random import seed
from scipy.stats import kendalltau
import pandas as pd
seed(1)
#input numerical, output categorical
dataset=pd.read_csv("weatherHistory.csv")
print(dataset)
#storing and removing the output class from the dataset
TargetClass = dataset[["Summary"]]
dataset.pop("Summary")
print("after removal:\n",dataset) #after removal of output class
arr=["Temperature (C)","Apparent Temperature (C)",
"Humidity","Wind Speed (km/h)","Wind Bearing (degrees)",
"Visibility (km)","Pressure (millibars)"]
attr=[]
temp=dataset[[arr[0]]]
attr.append(temp)
Atemp=dataset[[arr[1]]]
attr.append(Atemp)
Humid=dataset[[arr[2]]]
attr.append(Humid)
WindSpeed=dataset[[arr[3]]]
attr.append(WindSpeed)
WindBearing=dataset[[arr[4]]]
attr.append(WindBearing)
visibility=dataset[[arr[5]]]
attr.append(visibility)
pressure=dataset[[arr[6]]]
attr.append(pressure)
for i in range(len(attr)-1):
   for j in range(i+1,len(attr)):
      print(arr[i]," --> ",arr[j])
      corr, p = kendalltau(attr[i],attr[j])
      print("Kendall Rank Correlation          : ",corr)
      print("Probability against null hypothesis(p) : ",p) # probability that measures the
evidence against the null hypothesis
      alpha = 0.05
      if p > alpha:          print('Samples are uncorrelated (fail to reject H0) p=%.3f' % p)
      else:          print('Samples are correlated (reject H0) p=%.3f' % p)
      print("\n")
```

OUTPUT:

```
       Temperature (C)  Apparent Temperature (C)  Humidity  Wind Speed (km/h)  \
0             9.472222                  7.388889      0.89            14.1197
1             9.355556                  7.227778      0.86            14.2646
2             9.377778                  9.377778      0.89             3.9284
3             8.288889                  5.944444      0.83            14.1036
4             8.755556                  6.977778      0.83            11.0446
...                ...                       ...       ...                ...
96448        26.016667                 26.016667      0.43            10.9963
96449        24.583333                 24.583333      0.48            10.0947
96450        22.038889                 22.038889      0.56             8.9838
96451        21.522222                 21.522222      0.60            10.5294
96452        20.438889                 20.438889      0.61             5.8765

       Wind Bearing (degrees)  Visibility (km)  Pressure (millibars)  \
0                         251          15.8263               1015.13
1                         259          15.8263               1015.63
2                         204          14.9569               1015.94
3                         269          15.8263               1016.41
4                         259          15.8263               1016.51
...                       ...              ...                   ...
96448                      31          16.1000               1014.36
96449                      20          15.5526               1015.16
96450                      30          16.1000               1015.66
96451                      20          16.1000               1015.95
96452                      39          15.5204               1016.16

            Summary
0      Partly Cloudy
1      Partly Cloudy
2      Mostly Cloudy
3      Partly Cloudy
4      Mostly Cloudy
...              ...
96448  Partly Cloudy
96449  Partly Cloudy
96450  Partly Cloudy
96451  Partly Cloudy
96452  Partly Cloudy

[96453 rows x 8 columns]
```

```
after removal:
        Temperature (C)  Apparent Temperature (C)  Humidity  Wind Speed (km/h)  \
0             9.472222                  7.388889      0.89            14.1197
1             9.355556                  7.227778      0.86            14.2646
2             9.377778                  9.377778      0.89             3.9284
3             8.288889                  5.944444      0.83            14.1036
4             8.755556                  6.977778      0.83            11.0446
...                ...                       ...       ...                ...
96448        26.016667                 26.016667      0.43            10.9963
96449        24.583333                 24.583333      0.48            10.0947
96450        22.038889                 22.038889      0.56             8.9838
96451        21.522222                 21.522222      0.60            10.5294
96452        20.438889                 20.438889      0.61             5.8765

       Wind Bearing (degrees)  Visibility (km)  Pressure (millibars)
0                         251          15.8263               1015.13
1                         259          15.8263               1015.63
2                         204          14.9569               1015.94
3                         269          15.8263               1016.41
4                         259          15.8263               1016.51
...                       ...              ...                   ...
96448                      31          16.1000               1014.36
96449                      20          15.5526               1015.16
96450                      30          16.1000               1015.66
96451                      20          16.1000               1015.95
96452                      39          15.5204               1016.16

[96453 rows x 7 columns]
Temperature (C)  -->  Apparent Temperature (C)
Kendall Rank Correlation               :  0.9615086082924876
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000
```

```
Temperature (C)  -->  Apparent Temperature (C)
Kendall Rank Correlation              :  0.9615086082924876
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Temperature (C)  -->  Humidity
Kendall Rank Correlation              :  -0.41811899378228584
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Temperature (C)  -->  Wind Speed (km/h)
Kendall Rank Correlation              :  0.01164389719692316
Probability against null hypothesis(p) :  6.10090124513629e-08
Samples are correlated (reject H0) p=0.000


Temperature (C)  -->  Wind Bearing (degrees)
Kendall Rank Correlation              :  0.020578247729092634
Probability against null hypothesis(p) :  1.221060792555378e-21
Samples are correlated (reject H0) p=0.000


Temperature (C)  -->  Visibility (km)
Kendall Rank Correlation              :  0.2689559009881763
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Temperature (C)  -->  Pressure (millibars)
Kendall Rank Correlation              :  -0.20435395110130772
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Apparent Temperature (C)  -->  Humidity
Kendall Rank Correlation              :  -0.4037436260948938
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Apparent Temperature (C)  -->  Wind Speed (km/h)
Kendall Rank Correlation              :  -0.025783968030249207
Probability against null hypothesis(p) :  3.814679994748463e-33
Samples are correlated (reject H0) p=0.000


Apparent Temperature (C)  -->  Wind Bearing (degrees)
Kendall Rank Correlation              :  0.01868223772079493
Probability against null hypothesis(p) :  4.069898284248284e-18
Samples are correlated (reject H0) p=0.000


Apparent Temperature (C)  -->  Visibility (km)
Kendall Rank Correlation              :  0.25716214507112145
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Apparent Temperature (C)  -->  Pressure (millibars)
Kendall Rank Correlation              :  -0.19106794475086228
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Humidity  -->  Wind Speed (km/h)
Kendall Rank Correlation              :  -0.1772238337622487
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Humidity  -->  Wind Bearing (degrees)
Kendall Rank Correlation              :  -0.0012840292924364873
Probability against null hypothesis(p) :  0.5547673180784156
Samples are uncorrelated (fail to reject H0) p=0.555


Humidity  -->  Visibility (km)
Kendall Rank Correlation              :  -0.30050956070173696
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000
```

```
Humidity  -->  Pressure (millibars)
Kendall Rank Correlation               :  0.02917827311771651
Probability against null hypothesis(p) :  2.854477595700446e-41
Samples are correlated (reject H0) p=0.000


Wind Speed (km/h)  -->  Wind Bearing (degrees)
Kendall Rank Correlation               :  0.05814026441222385
Probability against null hypothesis(p) :  2.9786305417032454e-160
Samples are correlated (reject H0) p=0.000


Wind Speed (km/h)  -->  Visibility (km)
Kendall Rank Correlation               :  0.07047572731162938
Probability against null hypothesis(p) :  2.4912649556770944e-225
Samples are correlated (reject H0) p=0.000


Wind Speed (km/h)  -->  Pressure (millibars)
Kendall Rank Correlation               :  -0.15356530455714892
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000


Wind Bearing (degrees)  -->  Visibility (km)
Kendall Rank Correlation               :  0.03574531646368761
Probability against null hypothesis(p) :  3.192451215300687e-59
Samples are correlated (reject H0) p=0.000


Wind Bearing (degrees)  -->  Pressure (millibars)
Kendall Rank Correlation               :  -0.04752154184903051
Probability against null hypothesis(p) :  6.333516114598305e-108
Samples are correlated (reject H0) p=0.000


Visibility (km)  -->  Pressure (millibars)
Kendall Rank Correlation               :  -0.08875308362645083
Probability against null hypothesis(p) :  0.0
Samples are correlated (reject H0) p=0.000
```

RESULT:

Thus, we have successfully reduced the size of the dataset by selecting appropriate features using ANOVA and Kendall's method (as we have used numerical input data and categorical output data).