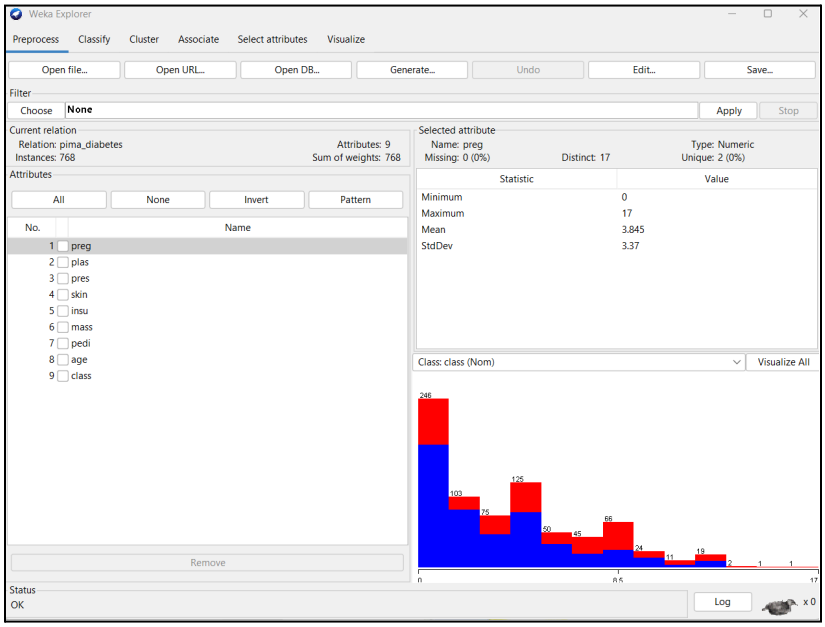
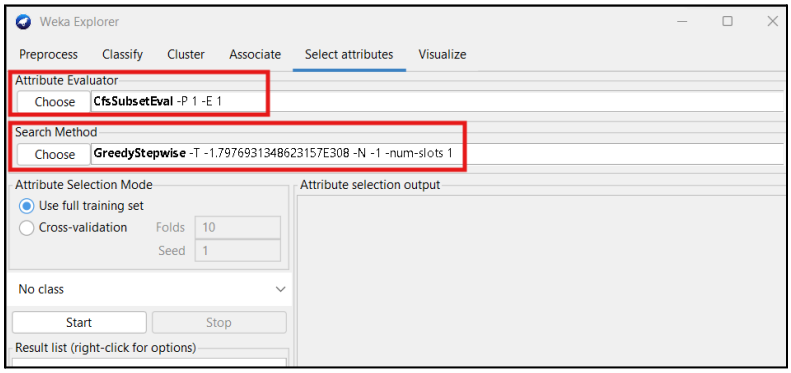
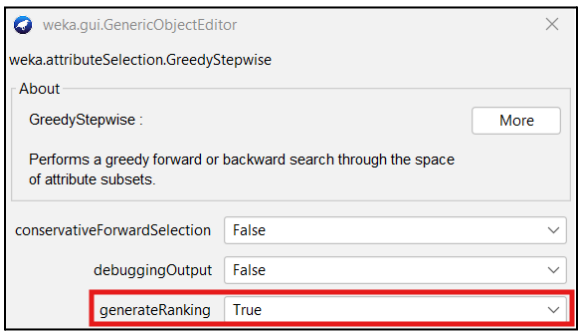
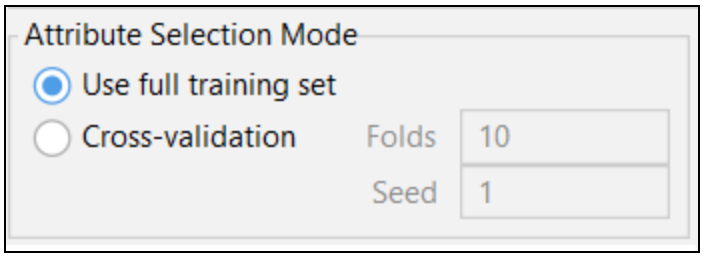
Ardientes, Mary Faith P.	Mr. Kho
IT31S1	09 - 24 - 25

Task 1: Feature Selection









Attribute selection output

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.GreedyStepwise -R -T -1.7976931348623157E308 -N -1 -num-slots 1
Relation:     pima_diabetes
Instances:    768
Attributes:   9
              preg
              plas
              pres
              skin
              insu
              mass
              pedi
              age
              class
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes

        Ranking is the order that attributes were added, starting
        with no attributes. The merit scores in the left column
        are the goodness of the subset after the adding the
        corresponding attribute in the right column to the subset.
```

```
Attribute Subset Evaluator (supervised, Class (nominal): 9 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Ranked attributes:
 0.133    2 plas
 0.151    6 mass
 0.164    8 age
 0.161    5 insu
 0.157    1 preg
 0.153    7 pedi
 0.147    4 skin
 0.141    3 pres

Selected attributes: 2,6,8,5,1,7,4,3 : 8
```

Reflection:

i. Which features were selected?

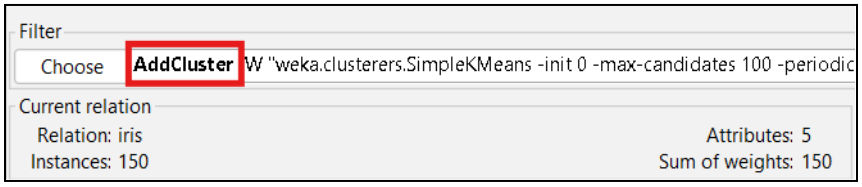
- Following the given procedure, the selected features are CfsSubsetEval with GreedyStepwise; these methods help with predicting the outcome while preventing each other from overlapping and redundancy among them.

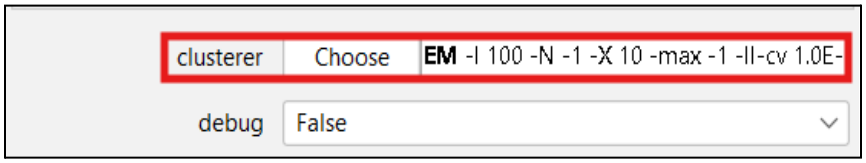
ii. What are the top 3 attributes detected by the CfsSubsetEval algorithm based on their ranked values?

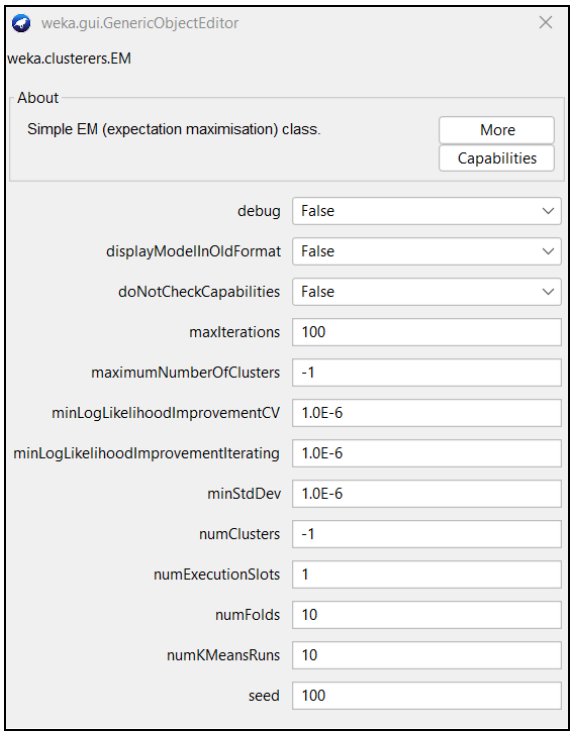
- Reviewing the output given, the rankings are listed in descending order of importance. The top 3 attributes, based on their merit scores, are
  - Age - 0.164
  - Insulin - 0.161
  - Mass - 0.151

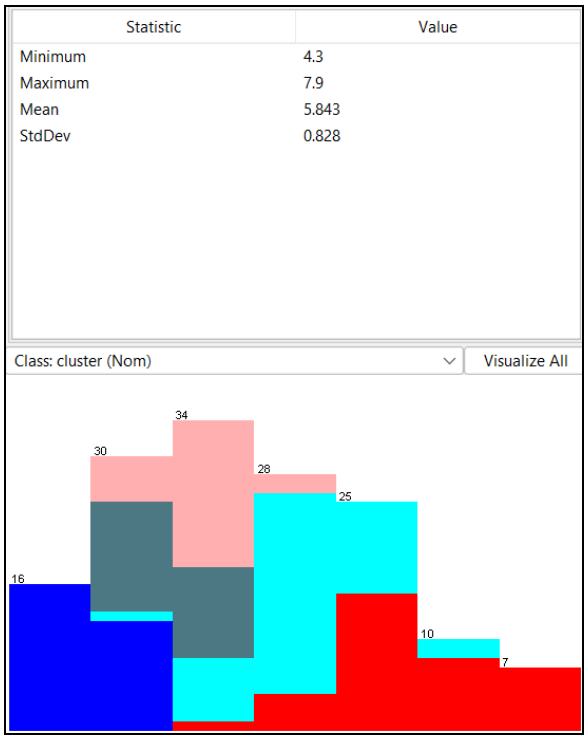
These mean the top 3 attributes that have the highest scores are the most useful for prediction.

Task 2: Instance Selection









Clusterer output

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0
Relation:     iris-weka.filters.unsupervised.attribute.AddCluster-Wweka
Instances:    150
Attributes:   6
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              cluster
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


EM
==

Number of clusters selected by cross validation: 6
Number of iterations performed: 9


                    Cluster
Attribute                0       1       2       3       4       5
                    (0.15)  (0.19)  (0.23)  (0.12)  (0.16)  (0.15)
=============================================================================
sepallength
  mean                5.5393  4.7784  6.8577  5.9993  6.2842  5.2949
  std. dev.           0.3063  0.2367  0.5168  0.3721  0.3995  0.2365

sepalwidth
  mean                2.5868   3.182  3.0914  2.7401  2.9247  3.7176
  std. dev.           0.2524  0.2566  0.2835  0.2338  0.2713  0.2809
```

Clusterer output

```
petallength
  mean                3.8568  1.4214  5.7863  4.9992  4.5598  1.5181
  std. dev.            0.339  0.1678  0.4691  0.2252  0.1811  0.1612

petalwidth
  mean                1.1696   0.193  2.1316  1.7646  1.4315  0.3088
  std. dev.           0.1333  0.0531  0.2353  0.1698  0.1058  0.1203

class
  Iris-setosa              1 28.9688       1       1       1 23.0312
  Iris-versicolor    23.9994       1  1.0009  3.9892 25.0104       1
  Iris-virginica           1       1 35.9546 16.0281  1.0173       1
  [total]            25.9995 30.9688 37.9555 21.0173 27.0277 25.0312
cluster
  cluster1                 1 28.9014       1       1       1  1.0986
  cluster2                 1       1  35.938   1.062       1       1
  cluster3             1.023       1  1.0175 18.9553 25.0042       1
  cluster4                 1  1.0674       1       1       1 22.9326
  cluster5           23.9765       1       1       1  1.0235       1
  [total]            27.9995 32.9688 39.9555 23.0173 29.0277 27.0312


Time taken to build model (full training data) : 0.3 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       23 ( 15%)
1       28 ( 19%)
2       35 ( 23%)
3       18 ( 12%)
4       24 ( 16%)
5       22 ( 15%)
```

Filter

Choose **RemoveWithValues** S 0.0 -C last -L first-last

Current relation

Relation: iris-weka.filters.unsupervised.attribute.AddClu... Attributes: 6

Instances: 0 Sum of weights: 0

weka.gui.GenericObjectEditor

weka.filters.unsupervised.instance.RemoveWithValues

About

Filters instances according to the value of an attribute. More Capabilities

attributeIndex	last
debug	False
doNotCheckCapabilities	False
dontFilterAfterFirstBatch	False
invertSelection	False
matchMissingValues	False
modifyHeader	False
nominalIndices	first-last
splitPoint	0.0

Open... Save... OK Cancel

Selected attribute

Name: sepallength Type: Numeric

Missing: 0 (0%) Distinct: 0 Unique: 0 (0%)

Statistic	Value
Minimum	NaN
Maximum	NaN
Mean	NaN
StdDev	NaN

Class: cluster (Nom) Visualize All

0

NaN NaN NaN

Reflection:

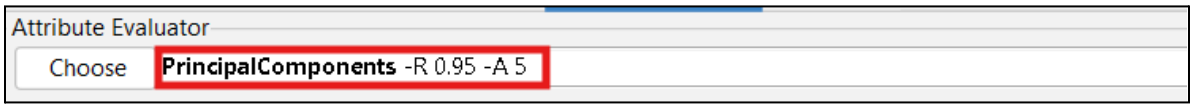
i. What is the purpose of the newly created attribute “Cluster”?

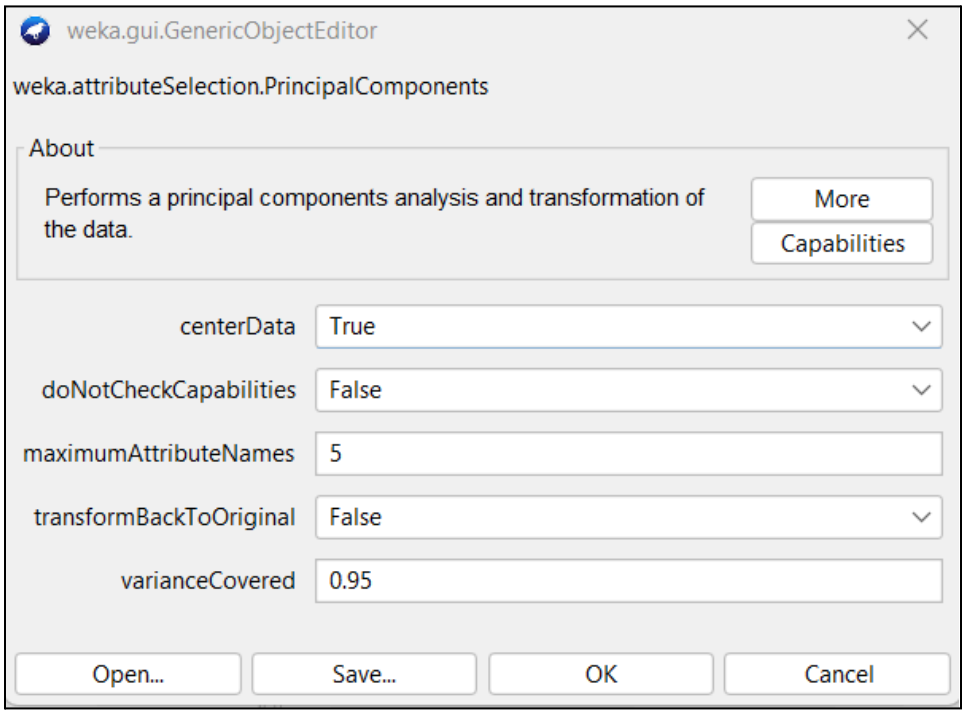
- Navigating the filter to choose the cluster is a feature added to the datasat to assign each instance to a cluster based on the patterns shown in the data using the select algorithm (EM). It helps to label instances with a nominal value, to organize them into which cluster they belong to. It overall filters noisy data and understands how the groups of data behave.

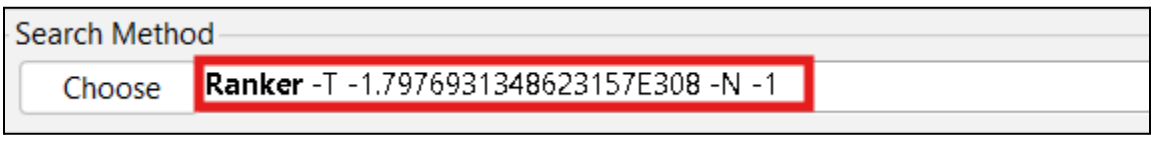
ii. After completing all the procedures, did you encounter the label “Number of clusters selected by cross-validation:”? What does this value represent?

- Yes, after completing the following procedure, it displayed the number of clusters selected by cross-validation: 6. This represents the number of clusters that EM determined using cross-validation, which evaluates different cluster counts and selects the one that maximizes model likelihood while avoiding overfitting. The EM set follows starts with a default setting og “-1” for the number of clusters, which tells it to let the data decide how many clusters best fit the distribution.

Task 3: Feature Transformation







Attribute selection output

```
=== Run information ===

Evaluator:    weka.attributeSelection.PrincipalComponents -C -R 0.95 -A 5
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (unsupervised):
        Principal Components Attribute Transformer

Covariance matrix
  0.69  -0.04   1.27   0.52  -0.28   0.03   0.25
 -0.04   0.19  -0.32  -0.12   0.12  -0.1   -0.03
  1.27  -0.32   3.11   1.3   -0.77   0.17   0.6
  0.52  -0.12   1.3    0.58  -0.32   0.04   0.28
 -0.28   0.12  -0.77  -0.32   0.22  -0.11  -0.11
  0.03  -0.1    0.17   0.04  -0.11   0.22  -0.11
  0.25  -0.03   0.6    0.28  -0.11  -0.11   0.22
```

```
eigenvalue      proportion      cumulative
  4.53848         0.86605         0.86605       0.827petallength+0.347petalwidth+0.346sepallength-0.203class=Iris-setosa+0.162class=Iris-virginica...
  0.39999         0.07633         0.94238       0.687class=Iris-versicolor-0.445sepalwidth-0.416class=Iris-virginica-0.274sepallength-0.271class=Iris-setosa...
  0.20244         0.03863         0.98101       0.677sepallength+0.469sepalwidth-0.397class=Iris-virginica+0.348class=Iris-versicolor-0.183petalwidth...

Eigenvectors
 V1      V2      V3
 0.3464 -0.274   0.6767 sepallength
-0.0813 -0.445   0.4693 sepalwidth
 0.8268  0.0763 -0.0879 petallength
 0.3469 -0.0582 -0.1829 petalwidth
-0.2034 -0.2712  0.0489 class=Iris-setosa
 0.041   0.6867  0.3477 class=Iris-versicolor
 0.1624 -0.4155 -0.3966 class=Iris-virginica

Ranked attributes:
 0.1339  1 0.827petallength+0.347petalwidth+0.346sepallength-0.203class=Iris-setosa+0.162class=Iris-virginica...
 0.0576  2 0.687class=Iris-versicolor-0.445sepalwidth-0.416class=Iris-virginica-0.274sepallength-0.271class=Iris-setosa...
 0.019   3 0.677sepallength+0.469sepalwidth-0.397class=Iris-virginica+0.348class=Iris-versicolor-0.183petalwidth...

Selected attributes: 1,2,3 : 3
```

Reflection:

i. How many principal components were created, and what percentage of total variance do they explain?

- Upon observing the output, three principal outcomes are evident in the proportions, which indicate the total variance in the dataset. These values are 0.86605 (86.605%), 0.07633 (7.633%), 0.03863 (3.863%), which explains about 98% (0.98101) of the data's variation, capturing all important patterns, making the dimensionality reduction effective.