

## Experiment-1.4

**Student Name: Ashish Kumar**

**Branch: CSE AIML**

**Semester: 01**

**Subject Name: Artificial Intelligence Lab**

**UID: 23MAI10008**

**Section/Group: 23MAI-1**

**Date of Performance:**

**Subject Code: 23CSH-621**

### **Aim of the Experiment :**

Aim of the Experiment is to explore the high dimensionality issues in the machine learning and Apply the three different feature selection techniques to the high dimensional cancer dataset downloaded from the UCI repository.

### **Objective of the Experiment :**

Task to be done for this experiment is that we have to explore the high dimensionality issues in the machine learning. Use the high dimensional cancer dataset downloaded from the UCI repository and apply different feature selection techniques which are:

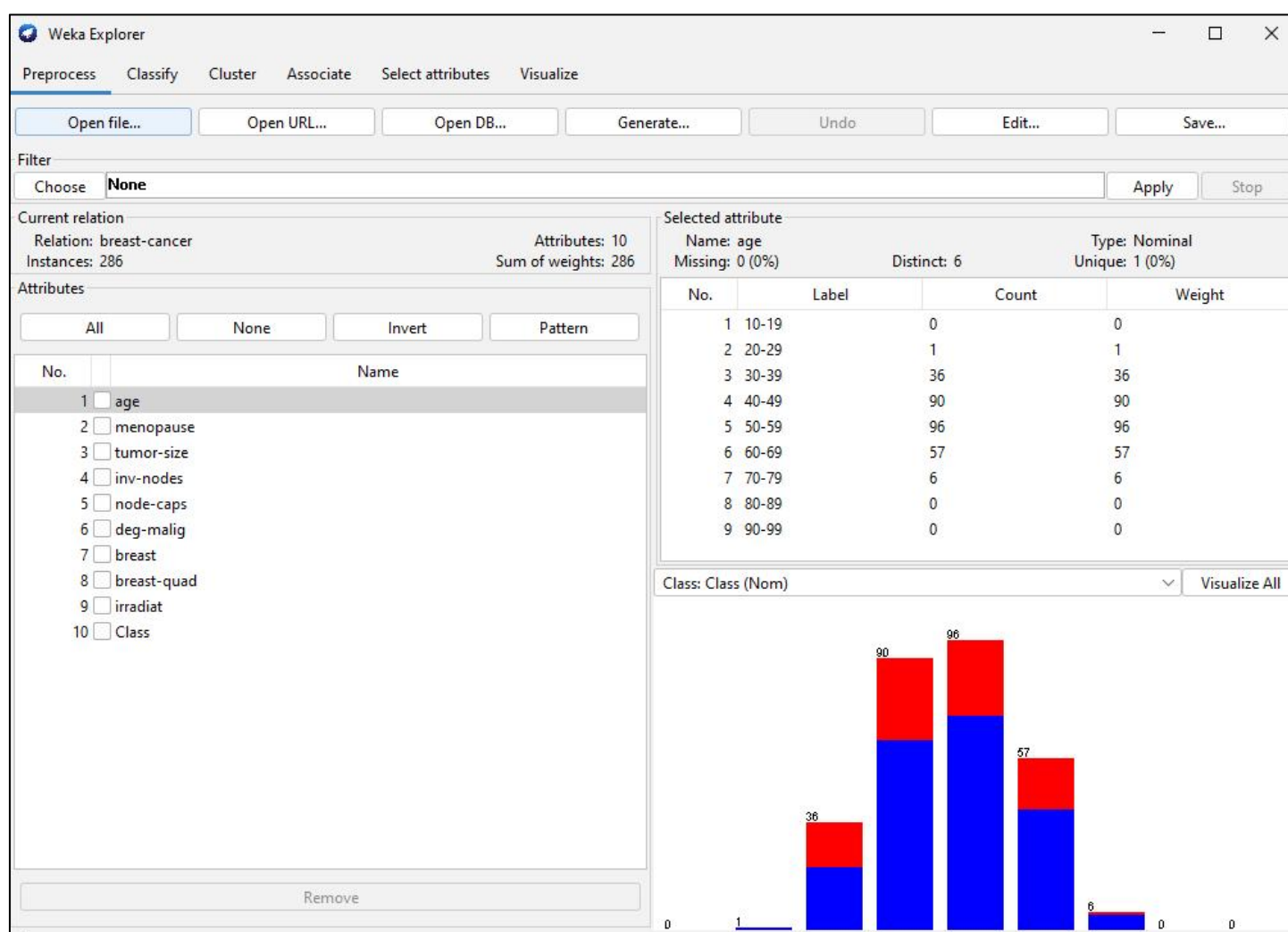
- a) filtering method
- b) Wrapper method
- c) PCA(Principal Component Analysis) method
- d) CorrelationAttributeEval method.

## Algorithm/ Steps for Experiment :

**Step 1:** Download the **Cancer dataset** from UCI repository.

**Step 2:** Convert the csv file into **arff** file using the WEKA Tool.

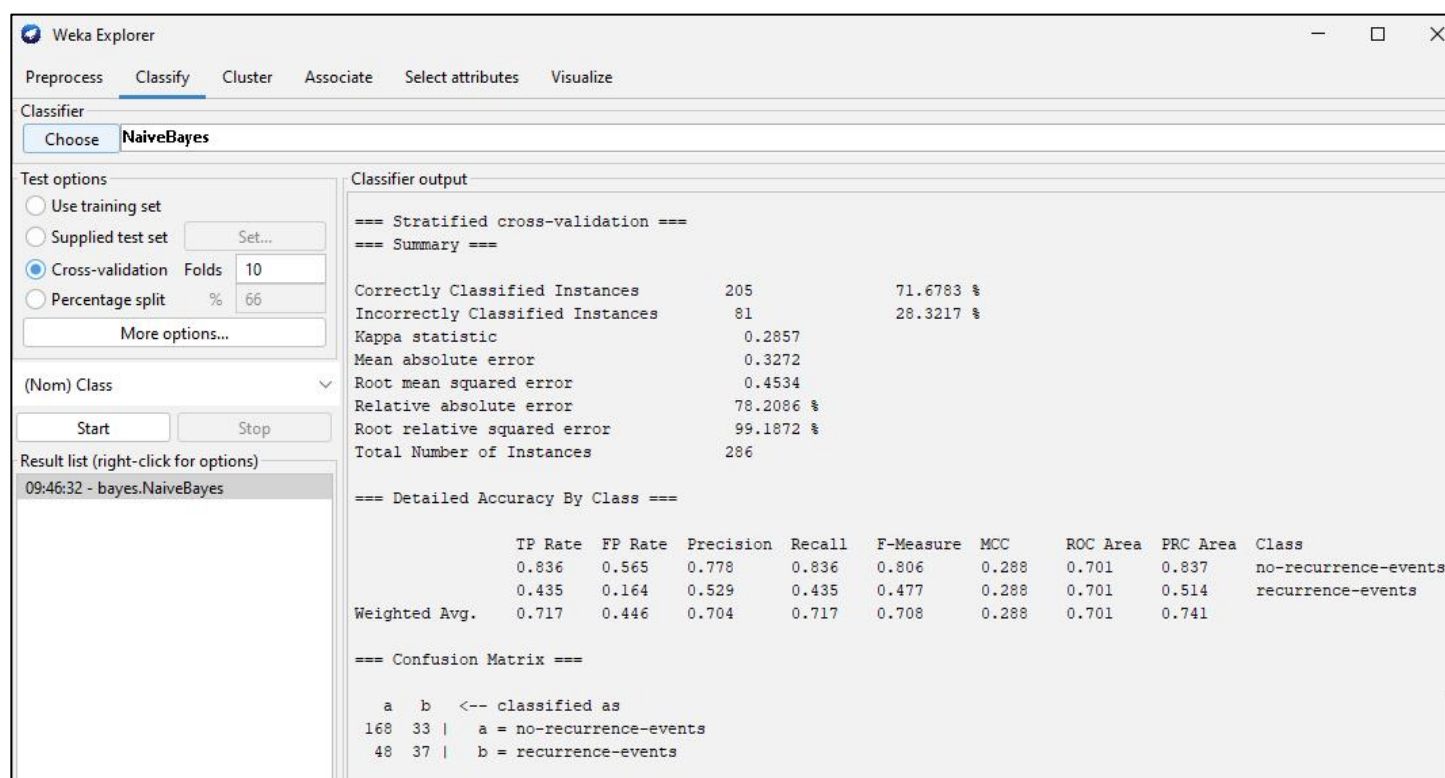
**Step 3:** Now open the Cancer dataset in the WEKA Tool using ‘**Open file**’ option.



## A) Wrapper Method:

- 1) Go to **Select Attributes** tab and in **Attribute Evaluator**, click on **‘Choose’** button.
- 2) Select the **Classifier Subset Evaluator** from the given list.
- 3) Click on Classifier Subset Evaluator → A dialog box will appear.
- 4) In the dialog box, choose the **Naive Bayes** classifier.
- 5) Select the attribute **(Nom) class** and click on **‘Start’** button.

## Before applying attribute evaluator:



The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'NaiveBayes'. Under 'Test options', 'Cross-validation' is selected with 'Folds' set to 10. The '(Nom) Class' dropdown is set to 'Class'. The 'Start' button has been clicked, and the 'Classifier output' pane displays the following results:

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      205           71.6783 %
Incorrectly Classified Instances    81           28.3217 %
Kappa statistic                    0.2857
Mean absolute error                 0.3272
Root mean squared error             0.4534
Relative absolute error             78.2086 %
Root relative squared error         99.1872 %
Total Number of Instances          286

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.836   0.565    0.778     0.836    0.806     0.288    0.701    0.837    no-recurrence-events
               0.435   0.164    0.529     0.435    0.477     0.288    0.701    0.514    recurrence-events
Weighted Avg.   0.717   0.446    0.704     0.717    0.708     0.288    0.701    0.741

=== Confusion Matrix ===

  a  b  <-- classified as
168 33 |  a = no-recurrence-events
 48 37 |  b = recurrence-events
  
```

## After applying attribute evaluator:

**Weka Explorer**

Preprocess   Classify   Cluster   Associate   **Select attributes**   Visualize

Attribute Evaluator  
Choose **ClassifierSubsetEval** -B weka.classifiers.bayes.NaiveBayes -T -H "Click to set hold out or test instances" -E DEFAULT

Search Method  
Choose **BestFirst** -D 1 -N 5

Attribute Selection Mode  
☒ Use full training set  
☐ Cross-validation   Folds: 10   Seed: 1

(Nom) Class  
Start   Stop

Result list (right-click for options)  
09:57:07 - BestFirst + ClassifierSubsetEval

Attribute selection output

```

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

Search Method:
  Best first.
  Start set: no attributes
  Search direction: forward
  Stale search after 5 node expansions
  Total number of subsets evaluated: 58
  Merit of best subset found: 0.776

Attribute Subset Evaluator (supervised, Class (nominal): 10 Class):
  Classifier Subset Evaluator
  Learning scheme: weka.classifiers.bayes.NaiveBayes
  Scheme options:
    Hold out/test set: Training data
    Subset evaluation: classification error

Selected attributes: 1,2,4,6,7,8 : 6
    age
    menopause
    inv-nodes
    deg-malig
    breast
    breast-quad
  
```

**Weka Explorer**

Preprocess   **Classify**   Cluster   Associate   Select attributes   Visualize

Classifier  
Choose **NaiveBayes**

Test options  
☐ Use training set  
☐ Supplied test set   Set...  
☒ Cross-validation   Folds: 10  
☐ Percentage split   %: 66  
 More options...

(Nom) Class  
Start   Stop

Result list (right-click for options)  
 09:46:32 - bayes.NaiveBayes  
 10:01:50 - bayes.NaiveBayes

Classifier output

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      211      73.7762 %
Incorrectly Classified Instances    75      26.2238 %
Kappa statistic                    0.2984
Mean absolute error                 0.3505
Root mean squared error             0.4363
Relative absolute error             83.7621 %
Root relative squared error         95.4487 %
Total Number of Instances          286

=== Detailed Accuracy By Class ===

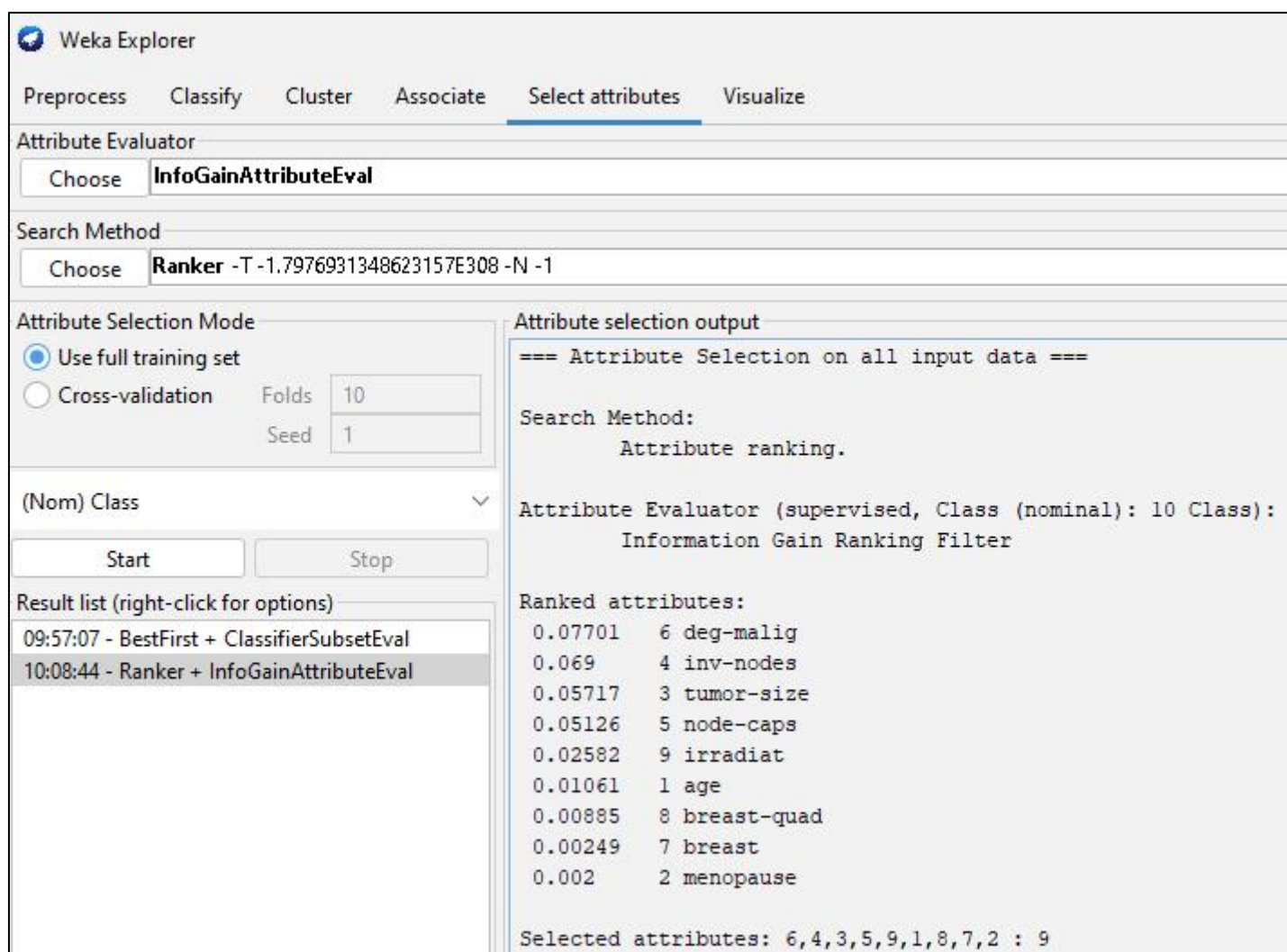
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
      0.891   0.624   0.772    0.891   0.827    0.312  0.686   0.797   no-recurrence-events
      0.376   0.109   0.593   0.376   0.460    0.312  0.686   0.494   recurrence-events
Weighted Avg.   0.738   0.471   0.718   0.738   0.718    0.312  0.686   0.707

=== Confusion Matrix ===

  a  b  <-- classified as
179 22 |  a = no-recurrence-events
 53 32 |  b = recurrence-events
  
```

## B) Filtering Method:

- 1) Go to **Select Attributes** tab and in **Attribute Evaluator**, click on 'Choose' button.
- 2) Select the **InfoGainAttributeEval** from the given list.
- 3) Click on **Search method** → choose **Ranker** → click on Ranker → a dialog box will appear.
- 4) In the dialog box, keep **num to select** as -1.
- 5) Select the attribute **(Nom) class** and click on 'Start' button.



The screenshot shows the Weka Explorer interface with the 'Select attributes' tab selected. The 'Attribute Evaluator' section has 'InfoGainAttributeEval' chosen. The 'Search Method' section has 'Ranker -T -1.7976931348623157E308 -N -1' chosen. The 'Attribute Selection Mode' section has 'Use full training set' selected. The 'Attribute selection output' section shows the results of the attribute selection process.

**Attribute Evaluator**

Choose **InfoGainAttributeEval**

**Search Method**

Choose **Ranker -T -1.7976931348623157E308 -N -1**

**Attribute Selection Mode**

☒ Use full training set  
☐ Cross-validation Folds: 10 Seed: 1

**(Nom) Class**

Start Stop

**Result list (right-click for options)**

09:57:07 - BestFirst + ClassifierSubsetEval  
10:08:44 - Ranker + InfoGainAttributeEval

**Attribute selection output**

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

Search Method:  
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 Class):  
Information Gain Ranking Filter

**Ranked attributes:**

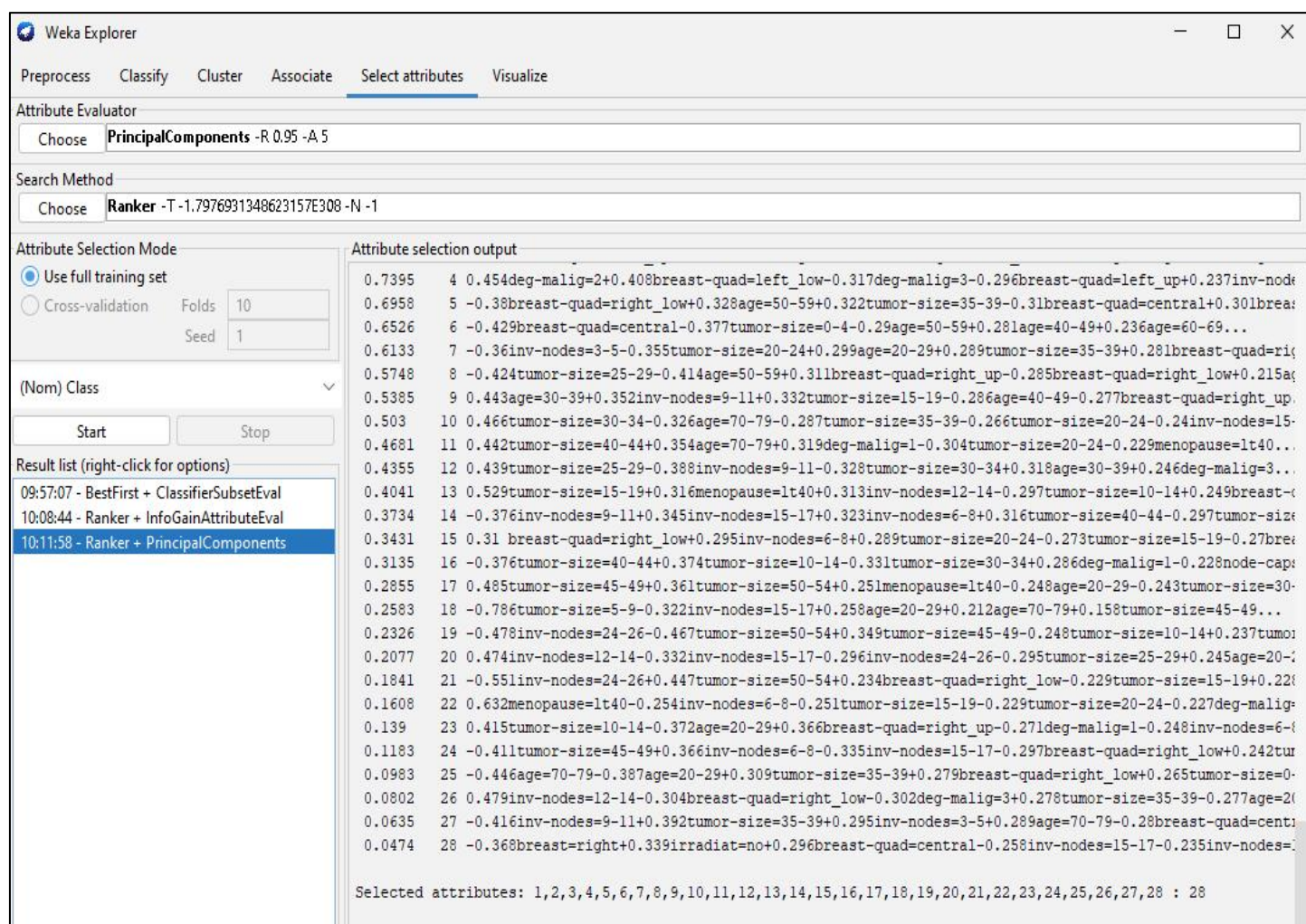
0.07701	6	deg-malig
0.069	4	inv-nodes
0.05717	3	tumor-size
0.05126	5	node-caps
0.02582	9	irradiat
0.01061	1	age
0.00885	8	breast-quad
0.00249	7	breast
0.002	2	menopause

**Selected attributes:** 6,4,3,5,9,1,8,7,2 : 9



## C) PCA (Principal Component Analysis):

- 1) Go to **Select Attributes** tab and in **Attribute Evaluator**, click on 'Choose' button.
- 2) Select the **Principal Components** from the given list.
- 3) Click on **Search method** → choose **Ranker** → click on Ranker → a dialog box will appear.
- 4) In the dialog box, keep **num to select** as -1.
- 5) Select the attribute (**Nom**) class and click on 'Start' button.



The screenshot shows the Weka Explorer interface with the 'Select attributes' tab selected. The 'Attribute Evaluator' is set to 'PrincipalComponents -R 0.95 -A 5'. The 'Search Method' is set to 'Ranker -T -1.7976931348623157E308 -N -1'. The 'Attribute Selection Mode' is set to 'Use full training set'. The '(Nom) Class' is set to 'Start'. The 'Result list' shows the 'Ranker + PrincipalComponents' method selected. The 'Attribute selection output' shows a list of 28 attributes selected.

**Attribute Evaluator:** Choose **PrincipalComponents -R 0.95 -A 5**

**Search Method:** Choose **Ranker -T -1.7976931348623157E308 -N -1**

**Attribute Selection Mode:** ☒ Use full training set ☐ Cross-validation Folds: 10 Seed: 1

**(Nom) Class:** Start Stop

**Result list (right-click for options):**

- 09:57:07 - BestFirst + ClassifierSubsetEval
- 10:08:44 - Ranker + InfoGainAttributeEval
- 10:11:58 - Ranker + PrincipalComponents

**Attribute selection output:**

```

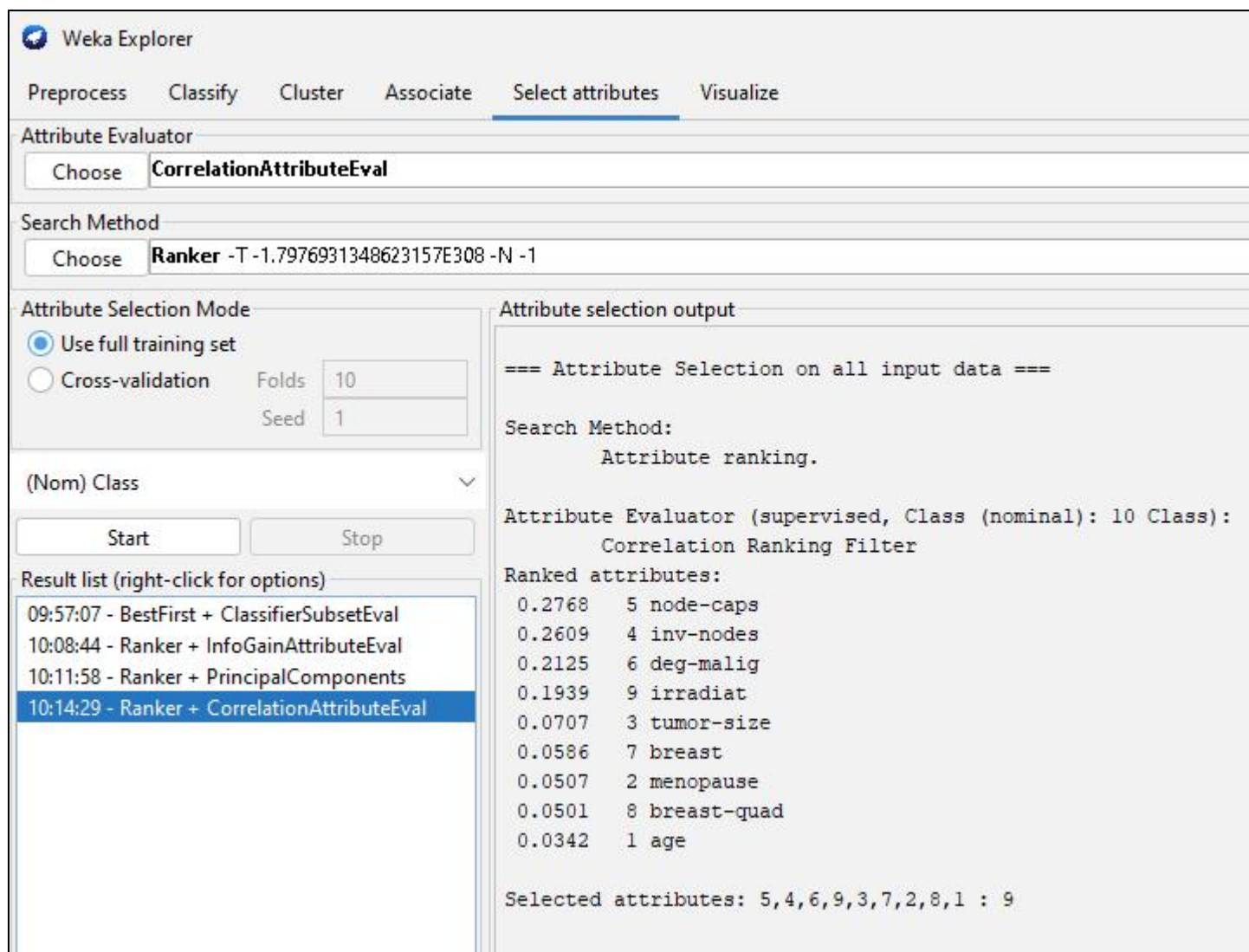
0.7395 4 0.454deg-malig=2+0.408breast-quad=left_low-0.317deg-malig=3-0.296breast-quad=left_up+0.237inv-nod
0.6958 5 -0.38breast-quad=right_low+0.328age=50-59+0.322tumor-size=35-39-0.31breast-quad=central+0.301brea
0.6526 6 -0.429breast-quad=central-0.377tumor-size=0-4-0.29age=50-59+0.281age=40-49+0.236age=60-69...
0.6133 7 -0.36inv-nodes=3-5-0.355tumor-size=20-24+0.299age=20-29+0.289tumor-size=35-39+0.281breast-quad=ri
0.5748 8 -0.424tumor-size=25-29-0.414age=50-59+0.311breast-quad=right_up-0.285breast-quad=right_low+0.215a
0.5385 9 0.443age=30-39+0.352inv-nodes=9-11+0.332tumor-size=15-19-0.286age=40-49-0.277breast-quad=right_up
0.503 10 0.466tumor-size=30-34-0.326age=70-79-0.287tumor-size=35-39-0.266tumor-size=20-24-0.24inv-nodes=15
0.4681 11 0.442tumor-size=40-44+0.354age=70-79+0.319deg-malig=1-0.304tumor-size=20-24-0.229menopause=lt40...
0.4355 12 0.439tumor-size=25-29-0.388inv-nodes=9-11-0.328tumor-size=30-34+0.318age=30-39+0.246deg-malig=3...
0.4041 13 0.529tumor-size=15-19+0.316menopause=lt40+0.313inv-nodes=12-14-0.297tumor-size=10-14+0.249breast-
0.3734 14 -0.376inv-nodes=9-11+0.345inv-nodes=15-17+0.323inv-nodes=6-8+0.316tumor-size=40-44-0.297tumor-siz
0.3431 15 0.31 breast-quad=right_low+0.295inv-nodes=6-8+0.289tumor-size=20-24-0.273tumor-size=15-19-0.27bre
0.3135 16 -0.376tumor-size=40-44+0.374tumor-size=10-14-0.331tumor-size=30-34+0.286deg-malig=1-0.228node-cap
0.2855 17 0.485tumor-size=45-49+0.361tumor-size=50-54+0.251menopause=lt40-0.248age=20-29-0.243tumor-size=30
0.2583 18 -0.786tumor-size=5-9-0.322inv-nodes=15-17+0.258age=20-29+0.212age=70-79+0.158tumor-size=45-49...
0.2326 19 -0.478inv-nodes=24-26-0.467tumor-size=50-54+0.349tumor-size=45-49-0.248tumor-size=10-14+0.237tumo
0.2077 20 0.474inv-nodes=12-14-0.332inv-nodes=15-17-0.296inv-nodes=24-26-0.295tumor-size=25-29+0.245age=20-
0.1841 21 -0.551inv-nodes=24-26+0.447tumor-size=50-54+0.234breast-quad=right_low-0.229tumor-size=15-19+0.22
0.1608 22 0.632menopause=lt40-0.254inv-nodes=6-8-0.251tumor-size=15-19-0.229tumor-size=20-24-0.227deg-malig
0.139 23 0.415tumor-size=10-14-0.372age=20-29+0.366breast-quad=right_up-0.271deg-malig=1-0.248inv-nodes=6-
0.1183 24 -0.411tumor-size=45-49+0.366inv-nodes=6-8-0.335inv-nodes=15-17-0.297breast-quad=right_low+0.242tur
0.0983 25 -0.446age=70-79-0.387age=20-29+0.309tumor-size=35-39+0.279breast-quad=right_low+0.265tumor-size=0
0.0802 26 0.479inv-nodes=12-14-0.304breast-quad=right_low-0.302deg-malig=3+0.278tumor-size=35-39-0.277age=2
0.0635 27 -0.416inv-nodes=9-11+0.392tumor-size=35-39+0.295inv-nodes=3-5+0.289age=70-79-0.28breast-quad=cent
0.0474 28 -0.368breast=right+0.339irradiat=no+0.296breast-quad=central-0.258inv-nodes=15-17-0.235inv-nodes=

```

**Selected attributes:** 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28 : 28

## D) Correlation Attribute Evaluation:

- 1) Go to **Select Attributes** tab and in **Attribute Evaluator**, click on 'Choose' button.
- 2) Select the **CorrelationAttributeEval** from the given list.
- 3) Click on **Search method** → choose **Ranker** → click on Ranker → a dialog box will appear.
- 4) In the dialog box, keep **num to select** as -1.
- 5) Select the attribute **(Nom) class** and click on 'Start' button.



**Weka Explorer**

Preprocess   Classify   Cluster   Associate   **Select attributes**   Visualize

**Attribute Evaluator**

Choose **CorrelationAttributeEval**

**Search Method**

Choose **Ranker** -T -1.7976931348623157E308 -N -1

**Attribute Selection Mode**

☒ Use full training set  
☐ Cross-validation   Folds: 10   Seed: 1

**(Nom) Class**   v

Start   Stop

**Result list (right-click for options)**

- 09:57:07 - BestFirst + ClassifierSubsetEval
- 10:08:44 - Ranker + InfoGainAttributeEval
- 10:11:58 - Ranker + PrincipalComponents
- 10:14:29 - Ranker + CorrelationAttributeEval**

**Attribute selection output**

```

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

Search Method:
  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 Class):
  Correlation Ranking Filter

Ranked attributes:
0.2768   5 node-caps
0.2609   4 inv-nodes
0.2125   6 deg-malig
0.1939   9 irradiat
0.0707   3 tumor-size
0.0586   7 breast
0.0507   2 menopause
0.0501   8 breast-quad
0.0342   1 age

Selected attributes: 5,4,6,9,3,7,2,8,1 : 9
  
```

### **Learning outcomes (What I have learnt):**

1. I learnt about the WEKA Tool and its applications.
2. I learnt about how to create dataset in .arff format.
3. I learnt about different feature selection techniques in WEKA Tool.
4. I learnt about filtering method and wrapper method in WEKA.
5. I learnt about principal component analysis and correlation attribute in WEKA.