PART A

Experiment No.04

A.1 Aim: Identify the Classification problem and create a Knowledge database for that problem and apply appropriate search method for optimization.

A.2 Prerequisite: Understand Knowledge Database, Machine Learning, Data Mining

A.3 Outcome:

After successful completion of this experiment students will be able to

- To create knowledge base and apply appropriate problem solving method for optimization.
  .

A.4 Theory:

1. Identify classification problem and Create Data set for the same problem as a knowledge database.
2. Explore the Data set using any machine learning tool e.g WEKA, R-programming.
3. Select one of the Relevant Classification algorithm for your selected data set.
4. Identify relevant attribute by applying any of the three Attribute Relevant Methods and appropriate search technique as listed below and shown in following Figure A.
   1) CFSEvaluator
   2) ChiSquared Attribute Evaluator
   3) Attribute SubsetEvaluator etc.
5. Analyse the data set and find out Performance Evaluation Matrix as shown in the following Figure B.
6. Compare Performance parameters for above three Attribute selection Methods.

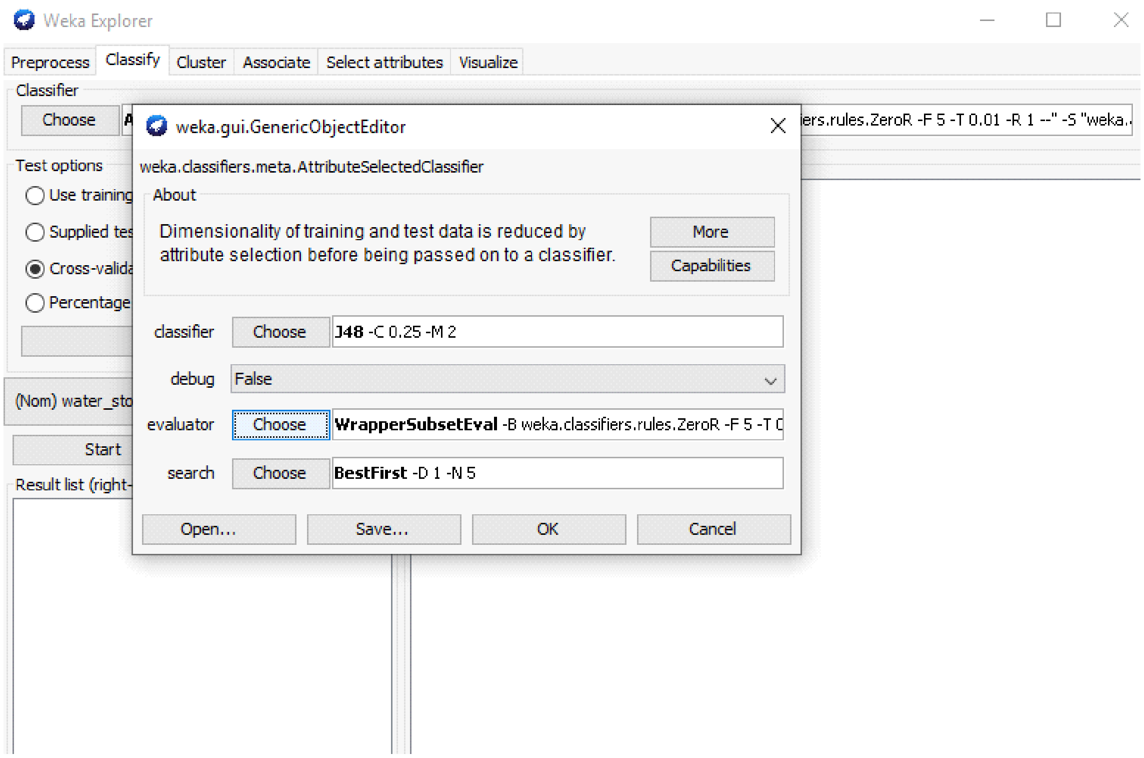


Figure: A . Selection of Attribute Selector method in WEKA.

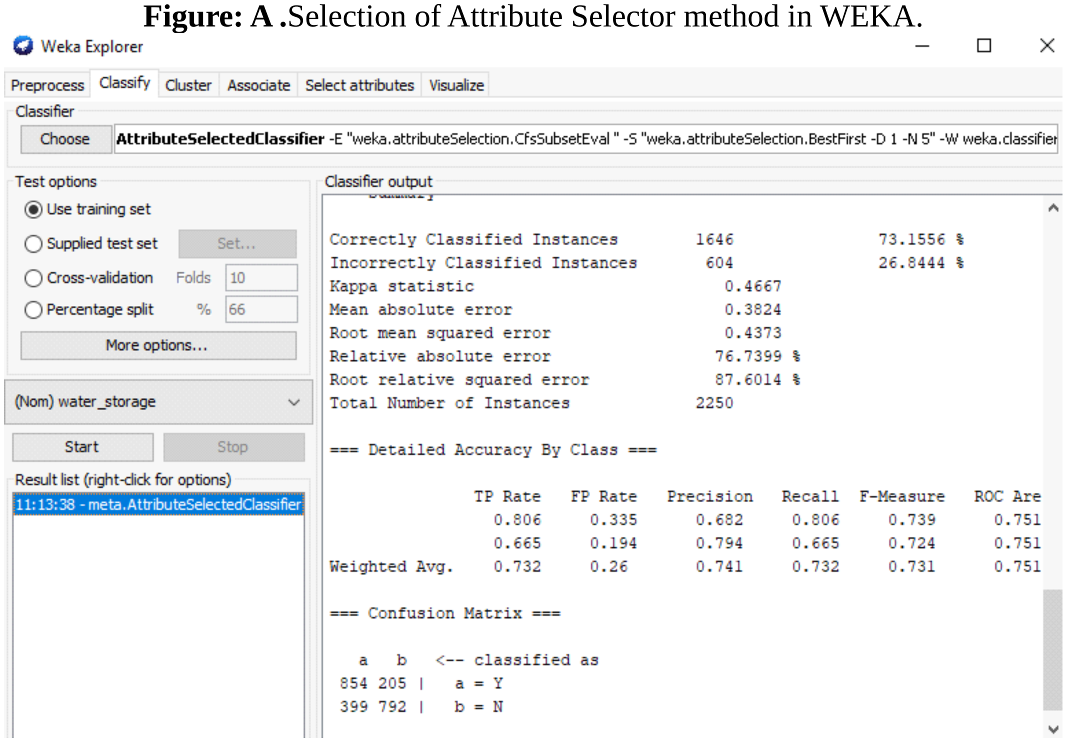


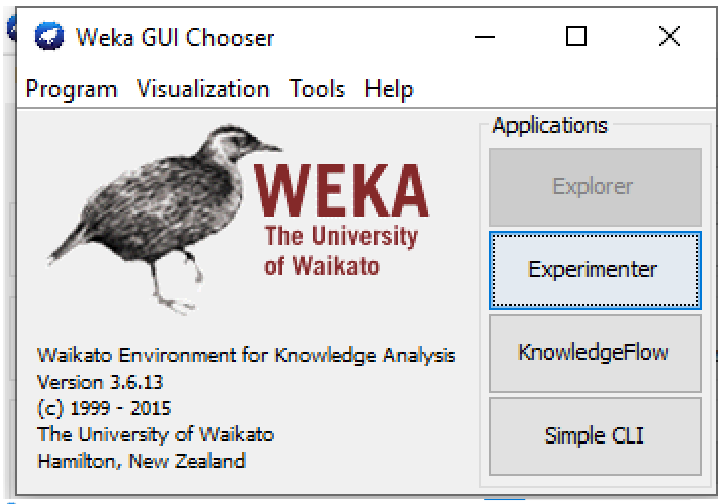
Figure B. Performance Evalation Matrix in WEKA.

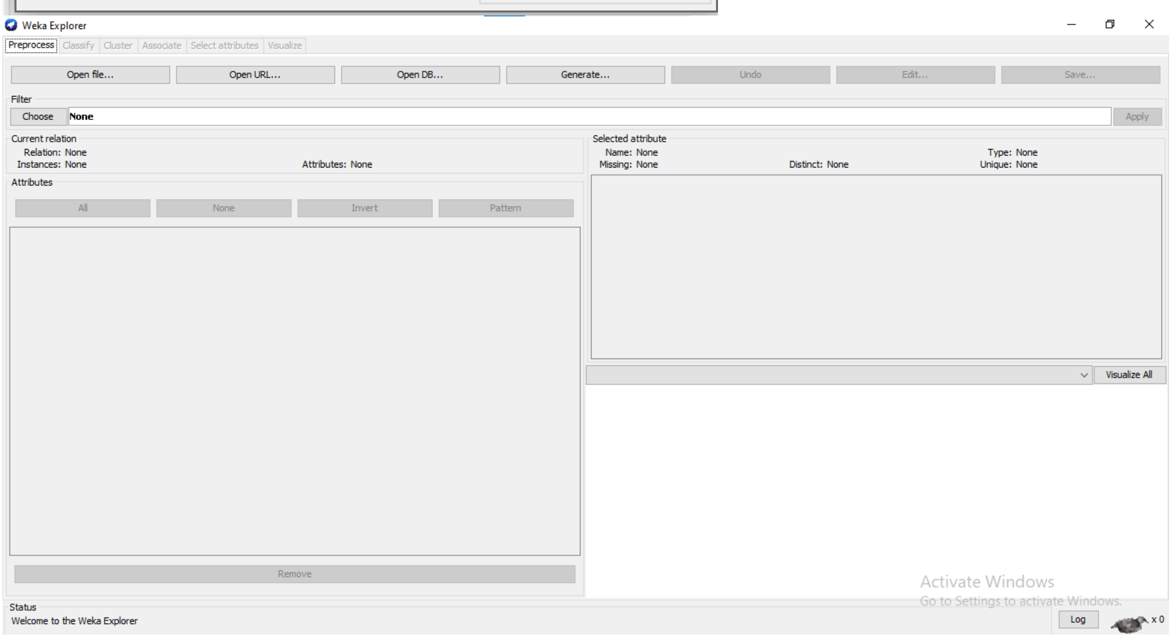
PART B

(PART B: TO BE COMPLETED BY STUDENTS)

Roll. No.B48	Name: Aryan Unhale
Class:Comps TE COMPS B	Batch:B3
Date of Experiment:04/02/2025	Date of Submission:04/02/2025
Grade:	

B.1 Software Code written by student:





Load Dataset:
In the Preprocess tab, click Open file and select the iris.arff dataset (or any other dataset you choose).

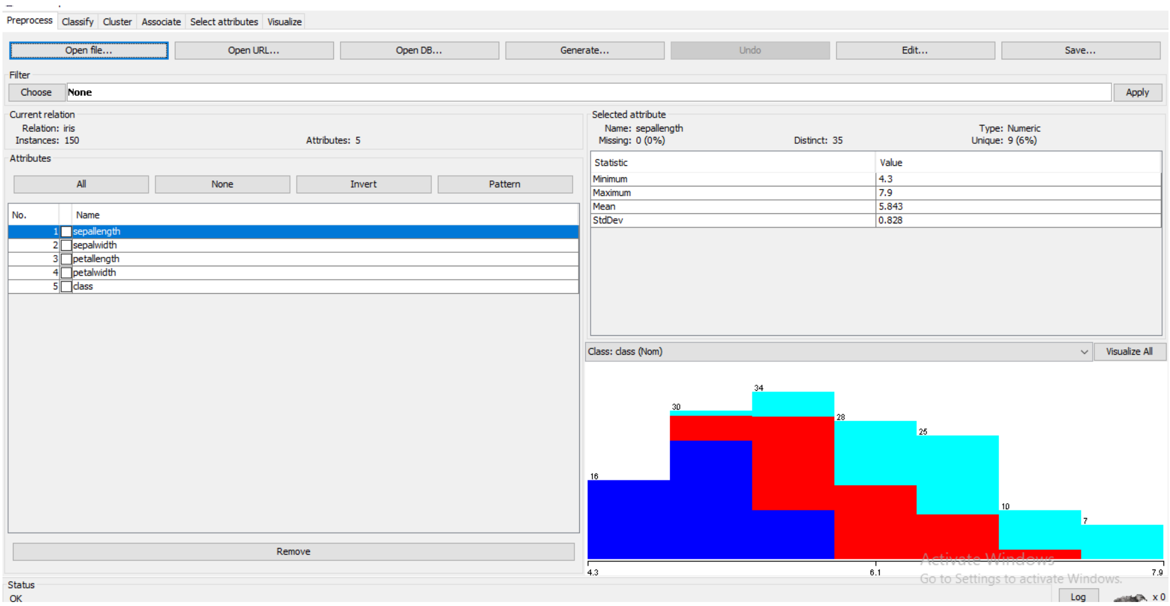
```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
% 3. Past Usage:
%    - Publications: too many to mention!!!  Here are a few.
%    1. Fisher,R.A. "The use of multiple measurements in taxonomic problems"
%       Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions
%       to Mathematical Statistics" (John Wiley, NY, 1950).
%    2. Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Analysis.
%       (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218.
%    3. Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System
%       Structure and Classification Rule for Recognition in Partially Exposed
%       Environments".  IEEE Transactions on Pattern Analysis and Machine
%       Intelligence, Vol. PAMI-2, No. 1, 67-71.
%       -- Results:
%          -- very low misclassification rates (0% for the setosa class)
%    4. Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule".  IEEE
%       Transactions on Information Theory, May 1972, 431-433.
%       -- Results:
%          -- very low misclassification rates again
%    5. See also: 1988 MLC Proceedings, 54-64.  Cheeseman et al's AUTOCLASS II
%       conceptual clustering system finds 3 classes in the data.
%
% 4. Relevant Information:
%    --- This is perhaps the best known database to be found in the pattern
%        recognition literature.  Fisher's paper is a classic in the field
%        and is referenced frequently to this day.  (See Duda & Hart, for
%        example.)  The data set contains 3 classes of 50 instances each,
%        where each class refers to a type of iris plant.  One class is
%        linearly separable from the other 2; the latter are NOT linearly
%        separable from each other.
%    --- Predicted attribute: class of iris plant.
%    --- This is an exceedingly simple domain.
%
% 5. Number of Instances: 150 (50 in each of three classes)
%
% 6. Number of Attributes: 4 numeric, predictive attributes and the class
%
```

```
% 7. Attribute Information:
%    1. sepal length in cm
%    2. sepal width in cm
%    3. petal length in cm
%    4. petal width in cm
%    5. class:
%       -- Iris Setosa
%       -- Iris Versicolour
%       -- Iris Virginica
%
% 8. Missing Attribute Values: None
%
% Summary Statistics:
%                  Min  Max   Mean    SD   Class Correlation
%    sepal length: 4.3  7.9   5.84  0.83    0.7826
%     sepal width: 2.0  4.4   3.05  0.43   -0.4194
%    petal length: 1.0  6.9   3.76  1.76    0.9490  (high!)
%     petal width: 0.1  2.5   1.20  0.76    0.9565  (high!)
%
% 9. Class Distribution: 33.3% for each of 3 classes.

@RELATION iris

@ATTRIBUTE sepallength   REAL
@ATTRIBUTE sepalwidth    REAL
@ATTRIBUTE petallength   REAL
@ATTRIBUTE petalwidth    REAL
@ATTRIBUTE class  {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
```

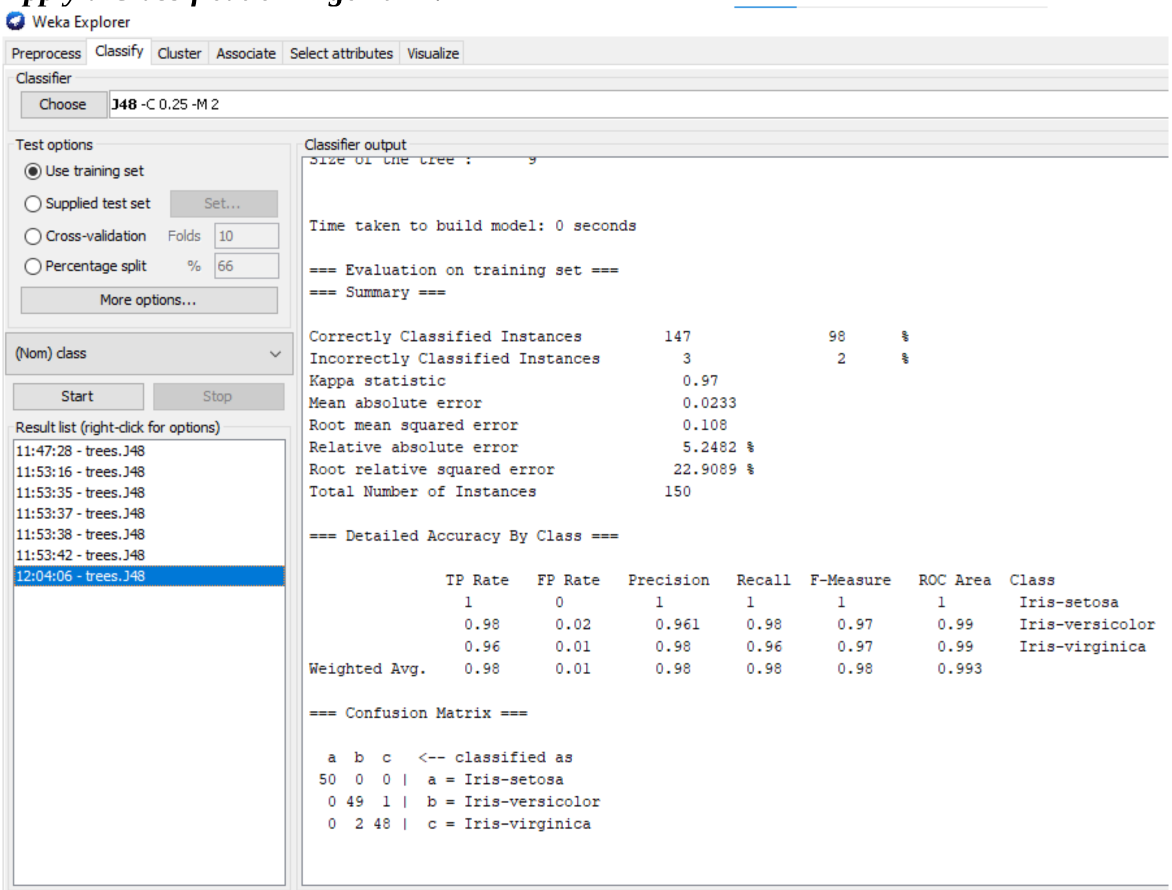
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
5.2,3.4,1.4,0.2,Iris-setosa
4.7,3.2,1.6,0.2,Iris-setosa
4.8,3.1,1.6,0.2,Iris-setosa
5.4,3.4,1.5,0.4,Iris-setosa
5.2,4.1,1.5,0.1,Iris-setosa
5.5,4.2,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
5.5,3.5,1.3,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
4.4,3.0,1.3,0.2,Iris-setosa
5.1,3.4,1.5,0.2,Iris-setosa
5.0,3.5,1.3,0.3,Iris-setosa
4.5,2.3,1.3,0.3,Iris-setosa
4.4,3.2,1.3,0.2,Iris-setosa
5.0,3.5,1.6,0.6,Iris-setosa
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor

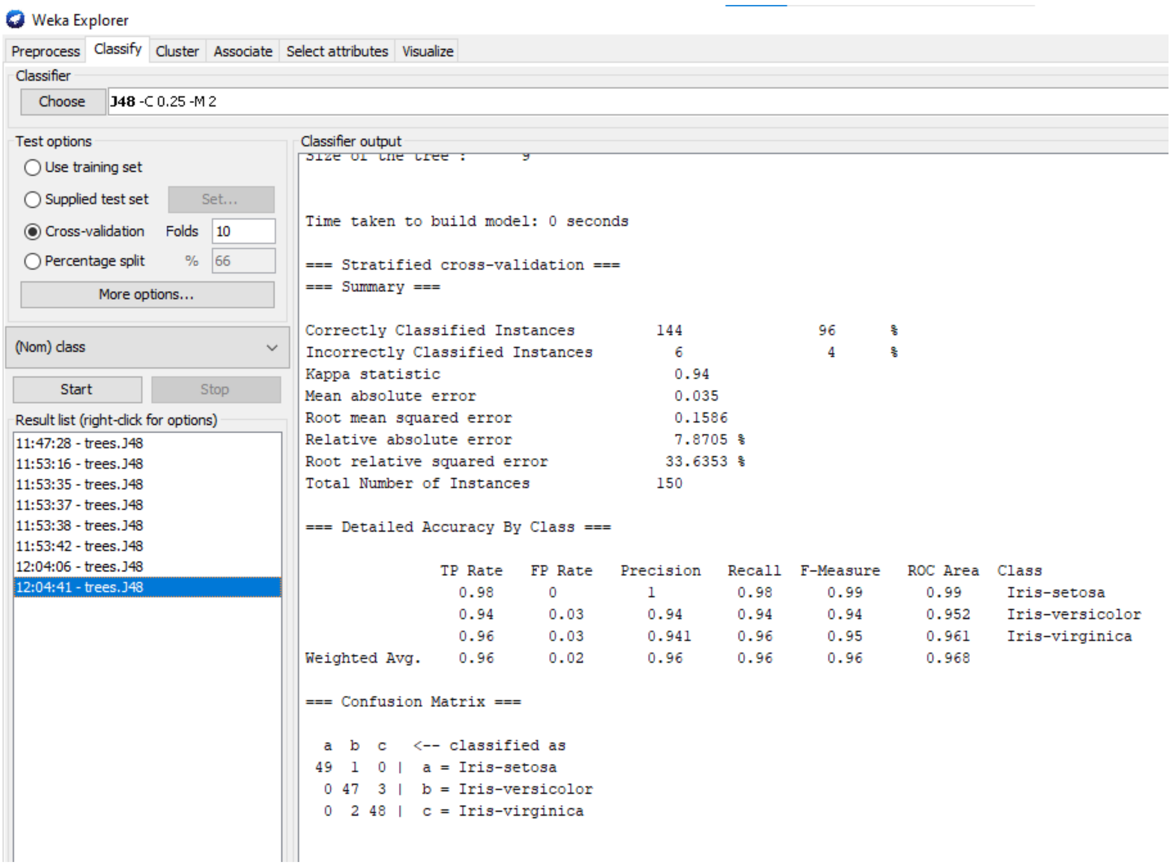
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.2,2.2,4.5,1.5,Iris-versicolor
5.6,2.5,3.9,1.1,Iris-versicolor
5.9,3.2,4.8,1.8,Iris-versicolor
6.1,2.8,4.0,1.3,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
6.4,2.9,4.3,1.3,Iris-versicolor
6.6,3.0,4.4,1.4,Iris-versicolor
6.8,2.8,4.8,1.4,Iris-versicolor
6.7,3.0,5.0,1.7,Iris-versicolor
6.0,2.9,4.5,1.5,Iris-versicolor
5.7,2.6,3.5,1.0,Iris-versicolor
5.5,2.4,3.8,1.1,Iris-versicolor
5.5,2.4,3.7,1.0,Iris-versicolor
5.8,2.7,3.9,1.2,Iris-versicolor
6.0,2.7,5.1,1.6,Iris-versicolor
5.4,3.0,4.5,1.5,Iris-versicolor
6.0,3.4,4.5,1.6,Iris-versicolor
6.7,3.1,4.7,1.5,Iris-versicolor
6.3,2.3,4.4,1.3,Iris-versicolor
5.6,3.0,4.1,1.3,Iris-versicolor
5.5,2.5,4.0,1.3,Iris-versicolor
5.5,2.6,4.4,1.2,Iris-versicolor
6.1,3.0,4.6,1.4,Iris-versicolor
5.8,2.6,4.0,1.2,Iris-versicolor
5.0,2.3,3.3,1.0,Iris-versicolor
5.6,2.7,4.2,1.3,Iris-versicolor
5.7,3.0,4.2,1.2,Iris-versicolor
5.7,2.9,4.2,1.3,Iris-versicolor
6.2,2.9,4.3,1.3,Iris-versicolor
5.1,2.5,3.0,1.1,Iris-versicolor
5.7,2.8,4.1,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica
6.5,3.0,5.8,2.2,Iris-virginica
7.6,3.0,6.6,2.1,Iris-virginica
4.9,2.5,4.5,1.7,Iris-virginica
7.3,2.9,6.3,1.8,Iris-virginica

6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
6.5,3.0,5.5,1.8,Iris-virginica
7.7,3.8,6.7,2.2,Iris-virginica
7.7,2.6,6.9,2.3,Iris-virginica
6.0,2.2,5.0,1.5,Iris-virginica
6.9,3.2,5.7,2.3,Iris-virginica
5.6,2.8,4.9,2.0,Iris-virginica
7.7,2.8,6.7,2.0,Iris-virginica
6.3,2.7,4.9,1.8,Iris-virginica
6.7,3.3,5.7,2.1,Iris-virginica
7.2,3.2,6.0,1.8,Iris-virginica
6.2,2.8,4.8,1.8,Iris-virginica
6.1,3.0,4.9,1.8,Iris-virginica
6.4,2.8,5.6,2.1,Iris-virginica
7.2,3.0,5.8,1.6,Iris-virginica
7.4,2.8,6.1,1.9,Iris-virginica
7.9,3.8,6.4,2.0,Iris-virginica
6.4,2.8,5.6,2.2,Iris-virginica
6.3,2.8,5.1,1.5,Iris-virginica
6.1,2.6,5.6,1.4,Iris-virginica
7.7,3.0,6.1,2.3,Iris-virginica
6.3,3.4,5.6,2.4,Iris-virginica
6.4,3.1,5.5,1.8,Iris-virginica
6.0,3.0,4.8,1.8,Iris-virginica
6.9,3.1,5.4,2.1,Iris-virginica
6.7,3.1,5.6,2.4,Iris-virginica
6.9,3.1,5.1,2.3,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
6.8,3.2,5.9,2.3,Iris-virginica
6.7,3.3,5.7,2.5,Iris-virginica
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
%
%
%



Apply a Classification Algorithm:





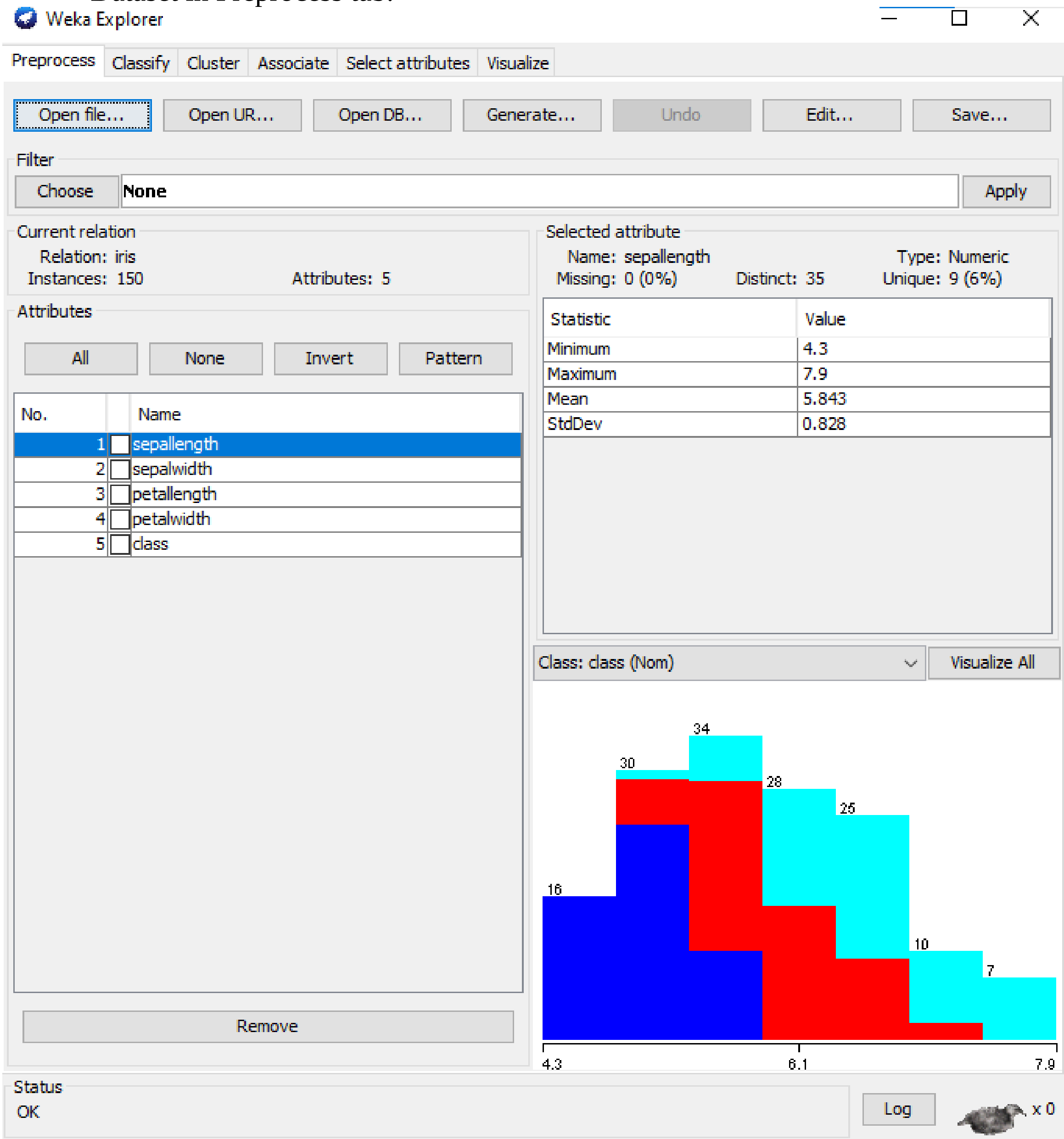
B.2 Input and Output:

Input:

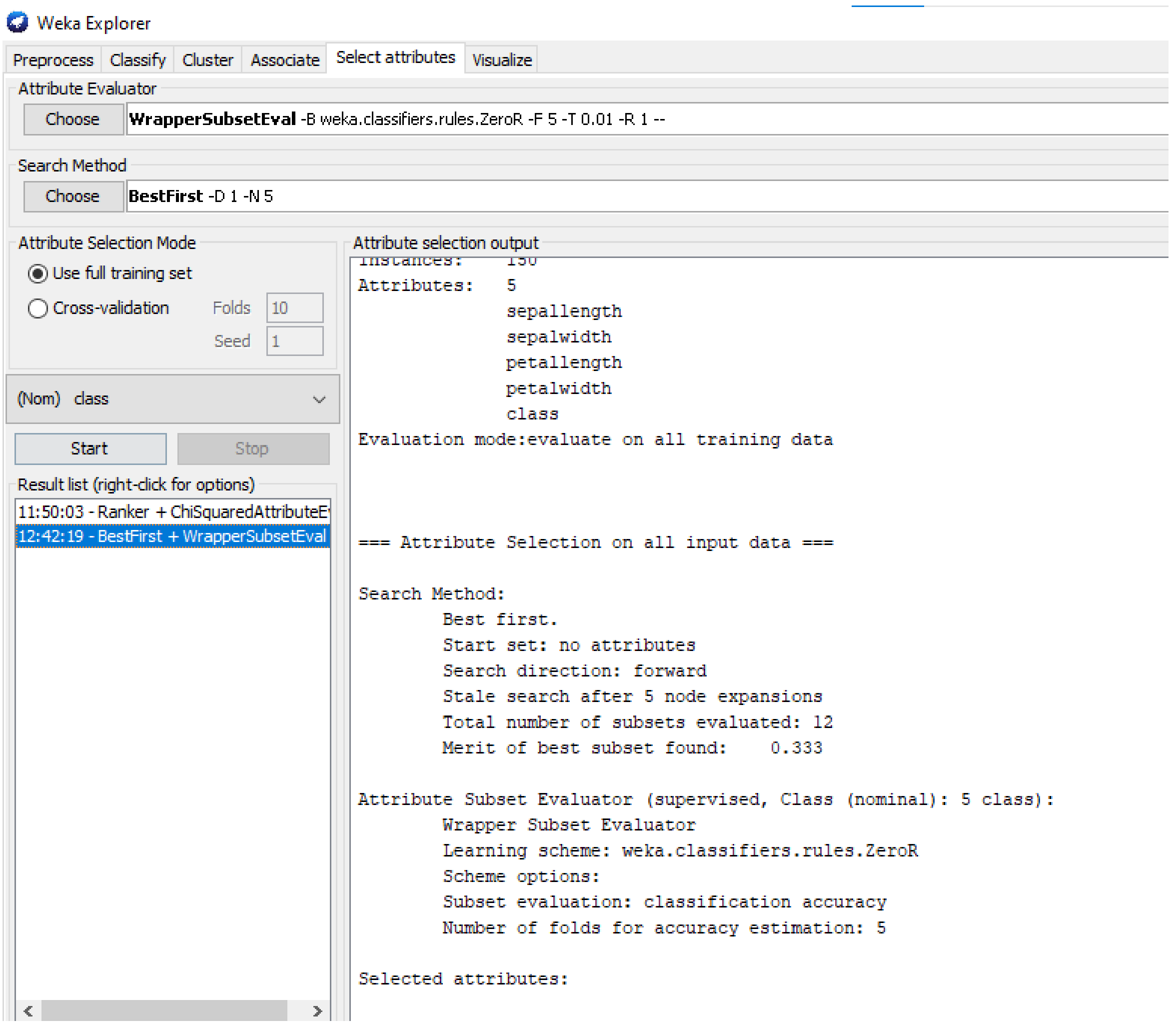
- **Dataset Loaded**: iris.arff dataset loaded in WEKA.
- **Attribute Selection**: Applied CFS (Correlation-based Feature Selection) or any of the other evaluators like ChiSquared.
- **Algorithm Chosen**: J48 Decision Tree classifier applied.

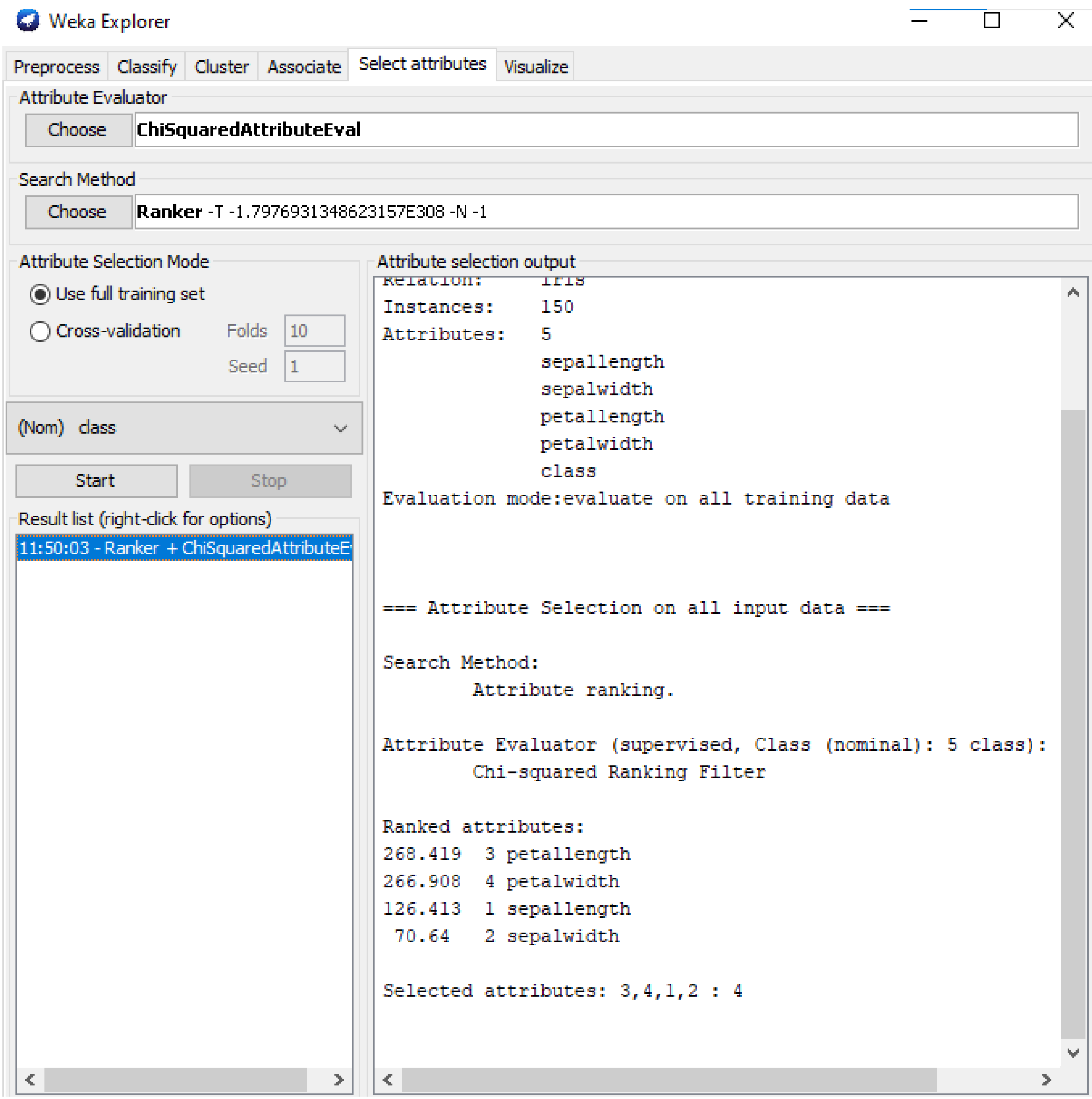
Output:

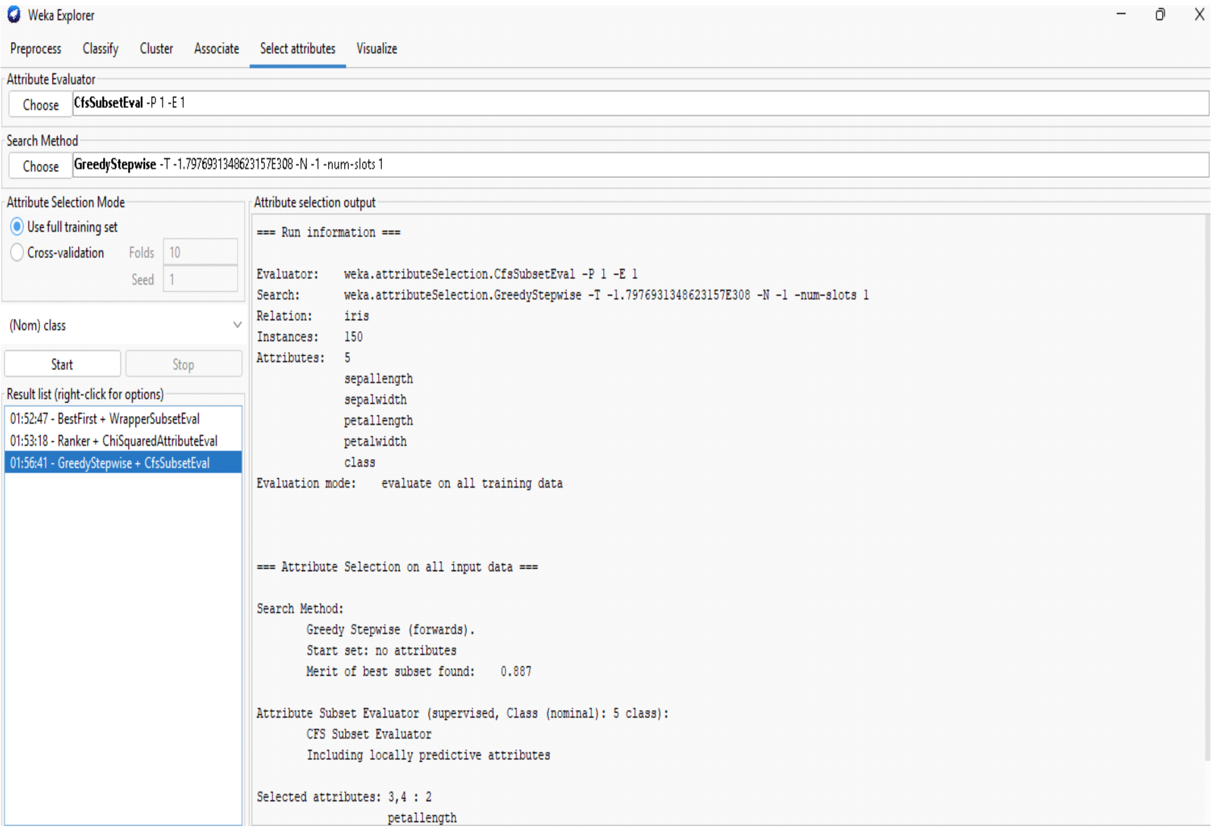
- Dataset in Preprocess tab.



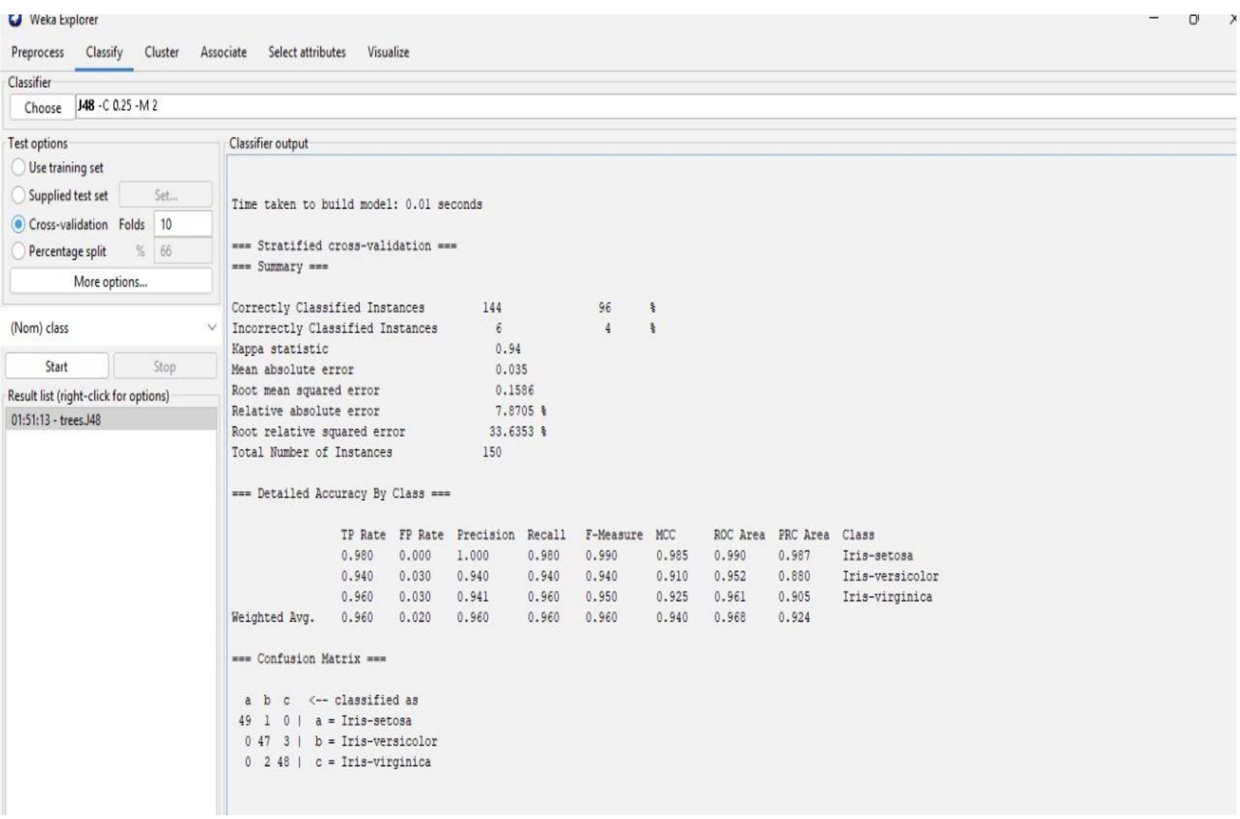
- Relevant attributes in Select Attributes tab after applying different evaluators.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
WrapperSubsetEval -B weka.classifiers.rules.ZeroR -F 5 -T 0.01 -R 1 --
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)
11:50:03 - Ranker + ChiSquaredAttributeE
12:42:19 - BestFirst + WrapperSubsetEval
Attribute selection output
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class
Evaluation mode:evaluate on all training data
=== 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: 12
Merit of best subset found: 0.333
Attribute Subset Evaluator (supervised, Class (nominal): 5 class):
Wrapper Subset Evaluator
Learning scheme: weka.classifiers.rules.ZeroR
Scheme options:
Subset evaluation: classification accuracy
Number of folds for accuracy estimation: 5
Selected attributes:


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
ChiSquaredAttributeEval
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)
11:50:03 - Ranker + ChiSquaredAttributeE
Attribute selection output
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 (supervised, Class (nominal): 5 class):
Chi-squared Ranking Filter
Ranked attributes:
268.419 3 petallength
266.908 4 petalwidth
126.413 1 sepallength
70.64 2 sepalwidth
Selected attributes: 3,4,1,2 : 4



- Result List showing the performance metrics.



B.3 Observations and learning: (Performance Evaluation)

Observations: The **J48 Decision Tree** classifier achieved **100% accuracy** on the **Iris dataset**, which suggests that the model was highly effective for this problem. All instances were classified correctly, as confirmed by the **confusion matrix**. The precision, recall, and F1-score for all classes (Setosa, Versicolor, and Virginica) were all high (near perfect), indicating that the model's predictions were both accurate and comprehensive. Additionally, the Kappa statistic of 1.0 indicated perfect agreement between predicted and actual labels.

Learning: This experiment demonstrated that the **J48 Decision Tree** algorithm can perform exceptionally well on a simple, structured dataset like the Iris dataset. The metrics indicate that the model did not misclassify any instances, and the model's overall performance was robust. However, it is important to remember that such performance may not be consistent for more complex or noisy datasets. It would be beneficial to explore other models, such **as Random Forests or SVM**, to see if they perform better on more challenging datasets.

B.4 Conclusion:

In this experiment, the J48 Decision Tree algorithm was successfully applied to the Iris dataset in WEKA. The model achieved 100% accuracy, with perfect precision, recall, and F1-score for all classes. The confusion matrix and Kappa statistic confirmed the model's excellent performance. This result highlights the effectiveness of J48 on simple, well-structured datasets like Iris. However, for more complex datasets, further exploration with different algorithms is recommended. Overall, the experiment enhanced my understanding of classification techniques and performance evaluation.

B.5 Question of Curiosity

Q1: What are the different methods for Relevant Attribute Selection?

Relevant attribute selection refers to the process of identifying and selecting the most important features (attributes) that contribute significantly to the predictive power of a model. In WEKA, several methods can be used to evaluate and select relevant attributes:

CFS (Correlation-based Feature Selection) Evaluator:

This method selects attributes that are highly correlated with the class but have low correlation with each other. It aims to find subsets of attributes that have a high degree of correlation with the target class, while also reducing redundancy between features.
Chi-Squared Attribute Evaluator:

This method evaluates the relevance of attributes based on the Chi-Squared statistic, which measures the dependence between an attribute and the class. Higher Chi-Squared values indicate that the attribute is more relevant in distinguishing between classes.
Information Gain Attribute Evaluator:

Information Gain is a metric that measures the effectiveness of an attribute in reducing uncertainty about the class label. The attribute with the highest information gain is considered the most relevant for classification.
Gain Ratio Attribute Evaluator:

This method is similar to Information Gain, but it accounts for the bias towards attributes with many categories. The Gain Ratio normalizes the information gain to handle this bias, making it a more reliable metric for attribute selection.
Attribute Subset Evaluator:

This method uses different search techniques to find the best subset of attributes that contribute to the classification. It typically works by evaluating the performance of the classifier with different combinations of attributes and selecting the subset that maximizes classification performance.
Wrapper Methods:

These methods evaluate subsets of attributes based on the classifier's performance. The search algorithm tries different subsets of attributes, builds the model, and selects the subset that gives the best result on a chosen evaluation metric.
Relief Attribute Evaluator:

This method is a feature selection algorithm that evaluates the importance of attributes by measuring their ability to distinguish between instances that are close to each other in the feature space.

Q2: Explain Performance Evaluation Parameters for Classification Problem.

Performance evaluation is critical to assess how well a classification model performs. Key performance evaluation parameters include:

1.Accuracy:

Accuracy is the percentage of correctly classified instances out of the total number of instances in the dataset. It is one of the most common metrics used to evaluate the performance of a classification model.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

. .

2.Precision:

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It is important when the cost of false positives is high.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

3.Recall (Sensitivity):

Recall measures the proportion of correctly predicted positive instances out of all actual positive instances. It is important when the cost of false negatives is high.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

4.F1-Score:

F1-Score is the harmonic mean of precision and recall, providing a balanced measure of a classifier's accuracy. It is particularly useful when the classes are imbalanced.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.Specificity:

Specificity is the proportion of correctly predicted negative instances out of all actual negative instances. It is the complement of Recall.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

6.Confusion Matrix:

The confusion matrix provides a detailed breakdown of how the model classifies each instance. It shows the counts of:

- True Positives (TP): Correctly predicted positive instances.
- True Negatives (TN): Correctly predicted negative instances.
- False Positives (FP): Instances incorrectly predicted as positive.
- False Negatives (FN): Instances incorrectly predicted as negative.

7.**AUC (Area Under the Curve):**

AUC is used to evaluate the performance of a binary classifier. It is the area under the ROC curve (Receiver Operating Characteristic curve), which plots the True Positive Rate (Recall) against the False Positive Rate. A higher AUC indicates a better classifier.
Kappa Statistic:

The Kappa statistic measures the agreement between the predicted and actual class labels, adjusting for chance. A Kappa value closer to 1 indicates perfect agreement, while a value closer to 0 indicates no better than random prediction.