YILDIZ TECHNICAL UNIVERSITY

FACULTY OF MECHANICAL ENGINEERING

DEPARTMENT OF INDUSTRIAL ENGINEERING

END3900 DATA MINING ASSIGNMENT-3

by

18069015- AYŞE YEŞİL

March, 2022

ISTANBUL

1. Introduction

In this assignment I have chance to investigate how to apply decision-tree and learn to apply algorithm on data set.I worked on a dataset where there are informations about on plants.By applying the J48 Classifier algorithm on this data set, I observed how can be apply Decision-tree classification on this dataset.

2. Dataset

I looked through the datasets on Weka and decided to do the homework on the iris.arff dataset, which is one of the datasets you can access that is a ready-made dataset from in weka. This dataset was created in 1988 by Michael Marshall.This dataset also contains 4 numeric, predictive attributes, and the class. The attributes are sepal length ,sepal width ,petal length ,petal width and class.Each classes contain of 50 instances and represent a type of iris plant which are consist of Iris Setosa,Iris Versicolour and Iris Virginica. In summary,there are 5 attributes and 150 instances.I chose this dataset because I thought I thought it was appropriate for the decision tree application since numeric and categorical data are available at the same time.

3. Method

In this study,I will apply decision tree claasification algorithm to classify. Decision Tree is a supervised learning technique that can be used for both classification and Regression problems.Decision tree learning uses a decision tree to go from observation about an item to conclusion about the item’s target value. Three components of decision tree are root node,link, and leaf node. Root represents the test condition for different attributes, the branch represents all possible outcomes that can be there in the test, and leaf nodes contain the label of the class to which it belongs.There are many specific decision-tree algorithms:ID3,C4.5,CART,Random Forest,CHAİD in Python. If I had chosen to do it in Python, I would have used ID3..As it is based on concept pf entropy and information gain which we learn in the lecture this subject. But I did the homework in weka, so I used the J48 Classifier algorithm which it is an algorithm to generate a decision tree that is generated by C4.5 (an extension of ID3). It is also known as a statistical classifier.

I will mention about the advantages and disadvantages of the decision tree algorithm.

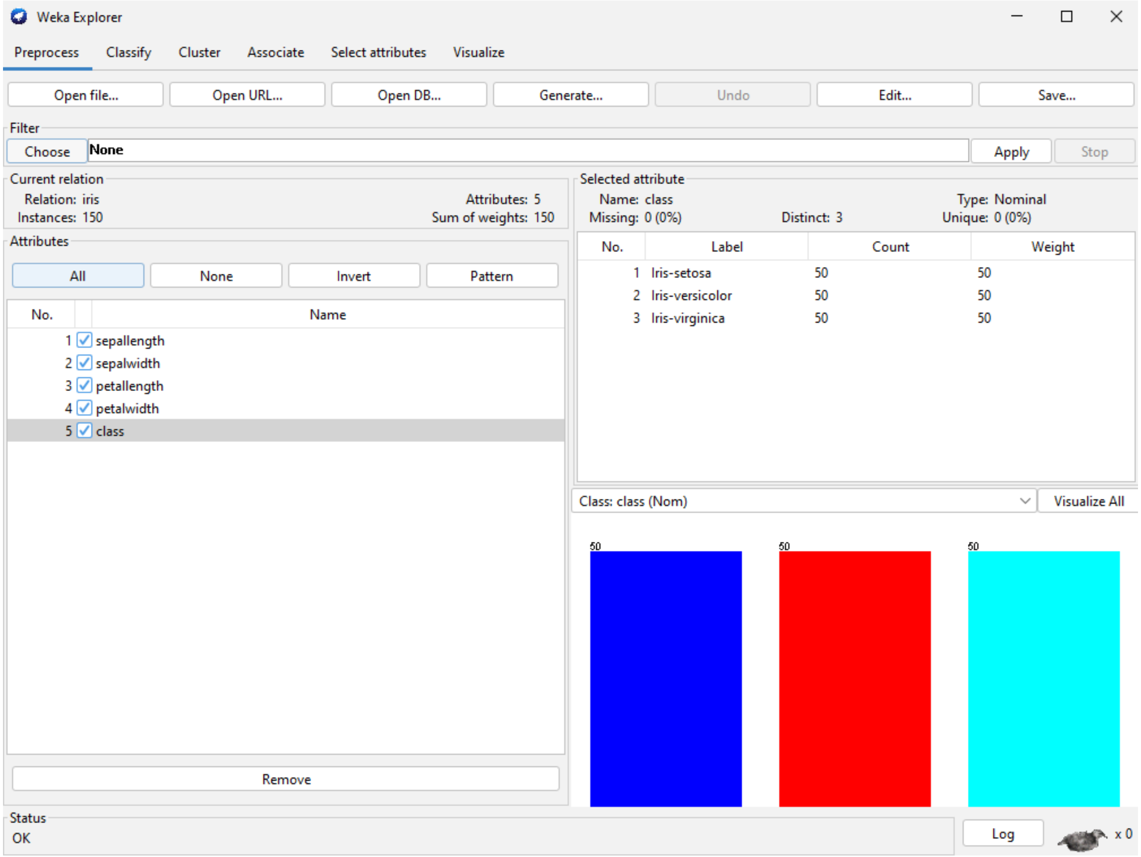
PROS	CONS
• Easy to understand and interpret	• It tends to overfit.
• The data can work with numerical and categorical features.	• Decision tree often involves higher time to train the model.
• Requires little data preprocessing	• The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.
• Fast for inference	• For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
• Feature selection happens automatically	

I encountered no troubles difficulties while applying the method.

4. Implementation

The details of the work done in this section should be explained step by step, supported by screenshots. The software chosen to implement the method used in the study and why it was chosen should be explained here.

Figure 1:Firstlly, I chesed explorer part in applications in Weka.After that I opened iris.arff file and choosed all attributes to evaluate.



Filter: None

Current relation
Relation: iris Attributes: 5
Instances: 150 Sum of weights: 150

Selected attribute
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	Iris-setosa	50	50
2	Iris-versicolor	50	50
3	Iris-virginica	50	50

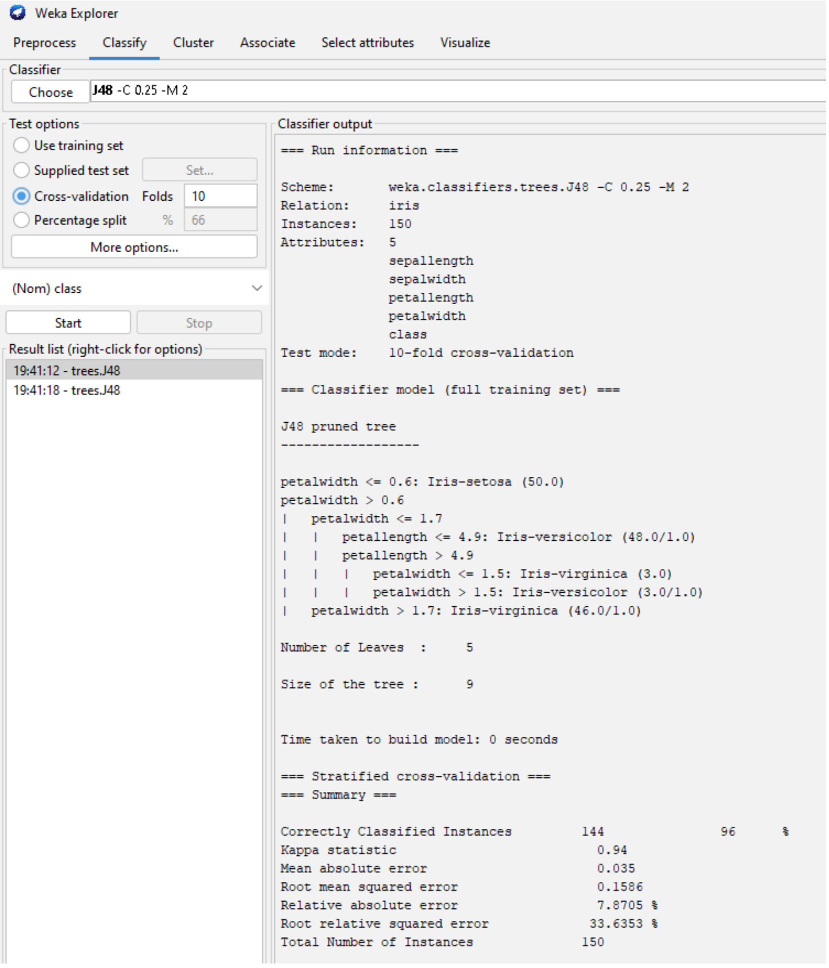
Attributes

No.	Name
1	☑ sepallength
2	☑ sepalwidth
3	☑ petallength
4	☑ petalwidth
5	☑ class

Class: class (Nom)

Status
OK

Figure 2: In classify part I chosed J48 Classifier explained it why I chosed in method.After I applied it to see results and tree.In this,I chosed test mode cross-validation and then percentage split since I wanted to compared their accuracy and chose best one.



Classifier: J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

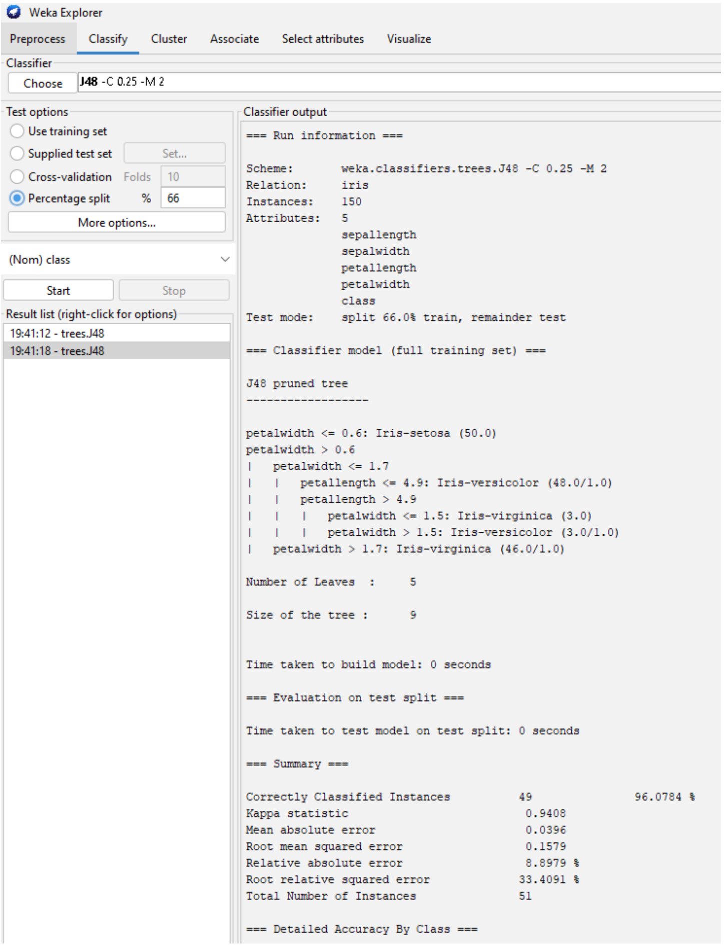
(Nom) class

Result list (right-click for options)
19:41:12 - trees.J48
19:41:18 - trees.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0 seconds

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

Correctly Classified Instances         144               96      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150
```

Figure 3:When split data %66 train ,the accuracy rate is 96.1%.And it is better cross-validation mode.



J48 -C 0.25 -M 2

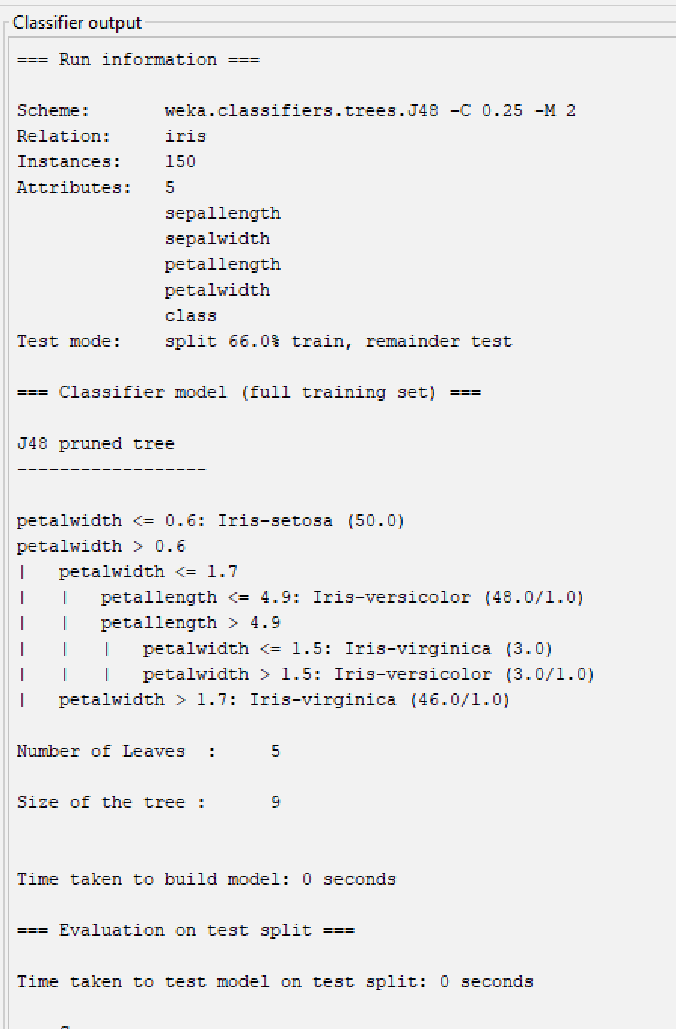
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) class

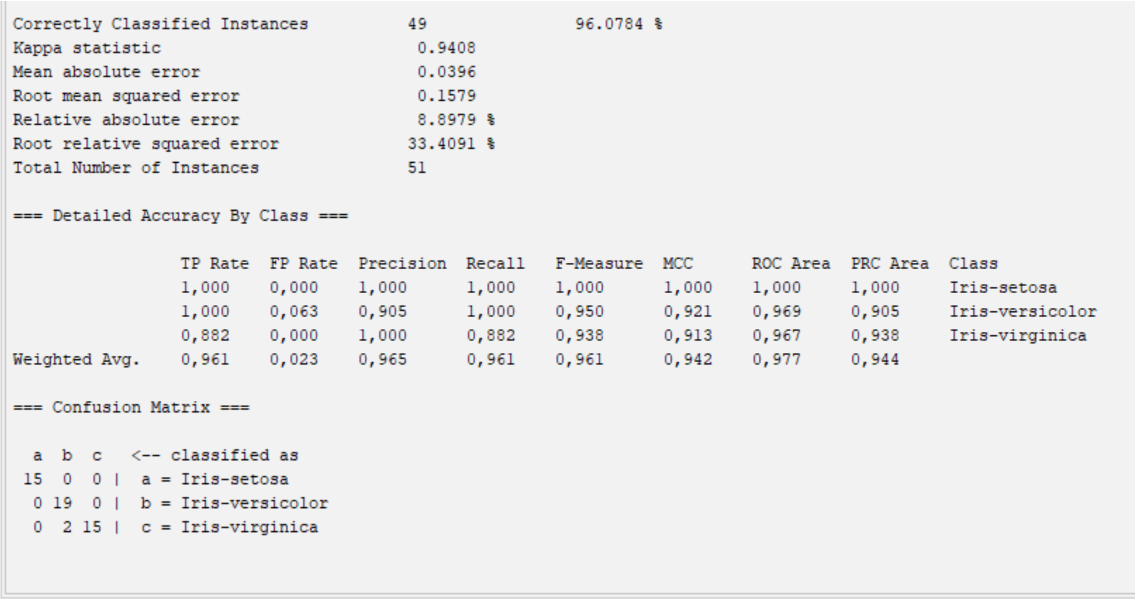
19:41:12 - trees.J48
19:41:18 - trees.J48

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          49               96.0784 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===
```

Figure4: In below pictures are classifer output result.

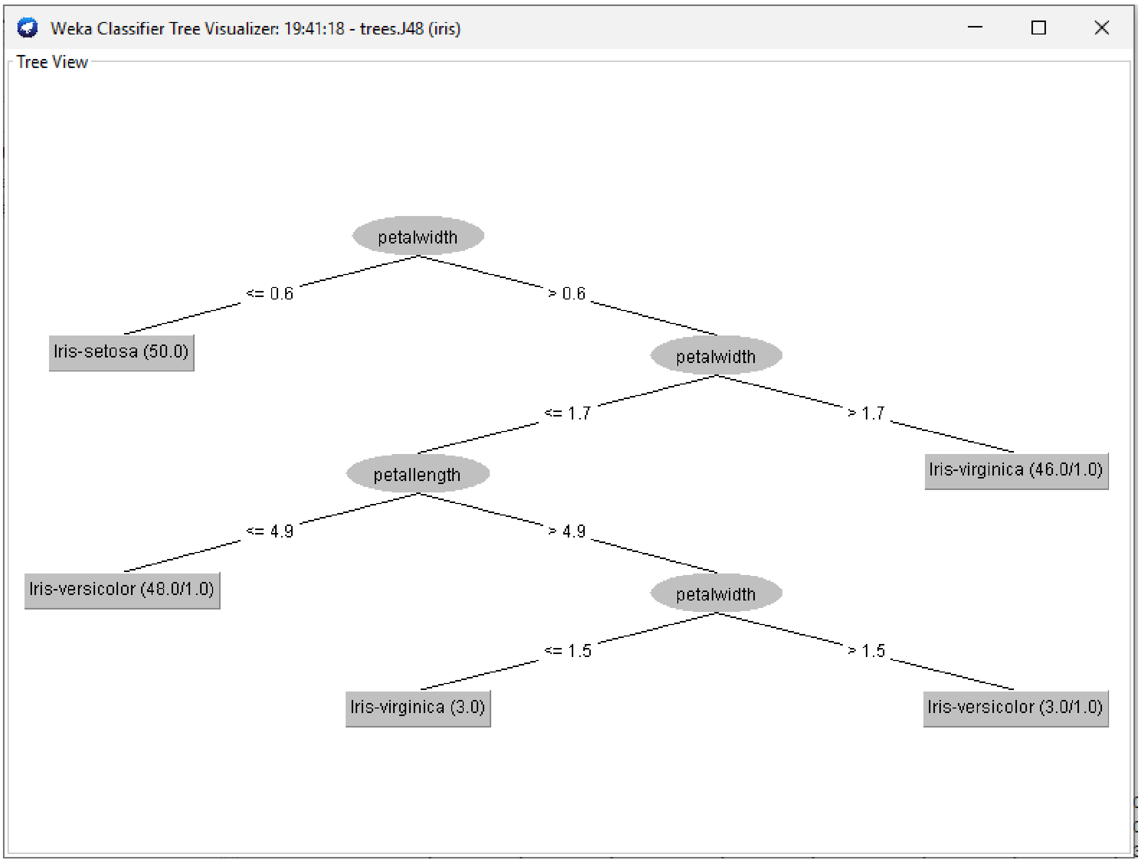


```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds
```



```
Correctly Classified Instances          49               96.0784 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     Iris-setosa
                 1,000    0,063    0,905      1,000    0,950      0,921    0,969     0,905     Iris-versicolor
                 0,882    0,000    1,000      0,882    0,938      0,913    0,967     0,938     Iris-virginica
Weighted Avg.    0,961    0,023    0,965      0,961    0,961      0,942    0,977     0,944

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  0  0 |  a = Iris-setosa
  0 19  0 |  b = Iris-versicolor
  0  2 15 |  c = Iris-virginica
```

Figure 5:I applied decision-tree classification algorithm on dataset and visualized it as a tree.


Weka Classifier Tree Visualizer: 19:41:18 - trees.J48 (iris)
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
> 1.7
Iris-virginica (46.0/1.0)
petallength
<= 4.9
> 4.9
Iris-versicolor (48.0/1.0)
petalwidth
<= 1.5
> 1.5
Iris-virginica (3.0)
Iris-versicolor (3.0/1.0)

5. Results and Evaluation

I can say that this decision-tree classification which I use J48 Classifier algorithm gives a very clear insight about the classification of plants.

For example when I havea input like that:

Petal with:1.5, petal length:5.8 the plants type ırıs-virginica (3.0).We can easy make conclusion and decide plants type with this tree after gave the inputs.

6. Resources

May 5, 2022,WEKA Dataset, Classifier And J48 Algorithm For Decision Tree

Afroz Chakure, Jul 6, 2019, Decision Tree Classification An introduction to Decision Tree Classifier