**WEKA Data mining Software Report**

**Group 8 – Submitted on behalf of Akarsh Pokkunuru – 801043808**

The Weka software workflow involves importing the dataset as the first step. The Iris dataset is used for the following tasks:

1. Decision tree classification
2. Hierarchical clustering
3. Association rule mining

The initial dataset features and statistics are shown as follows:

A screenshot of a computer

Description automatically generated

The dataset contains 3 classes and each class with 4 features. The dataset is preprocessed by removing missing values, normalizing the values between 0 and 1.

The individual feature attributes are plotted as follows:

A picture containing screenshot

Description automatically generated

A screenshot of a computer

Description automatically generated

**Decision Tree classification:**

The first task at hand is to classify the data. The J48 decision tree classifier yields good classification performance compared to the Decision Stump algorithm. A 10-fold cross validation is also applied during the testing phase. The results are shown below:

=== 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: 0.04 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 144 96 %

Incorrectly Classified Instances 6 4 %

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

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.980 0.000 1.000 0.980 0.990 0.985 0.990 0.987 Iris-setosa

0.940 0.030 0.940 0.940 0.940 0.910 0.952 0.880 Iris-versicolor

0.960 0.030 0.941 0.960 0.950 0.925 0.961 0.905 Iris-virginica

Weighted Avg. 0.960 0.020 0.960 0.960 0.960 0.940 0.968 0.924

=== Confusion Matrix ===

a b c <-- classified as

49 1 0 | a = Iris-setosa

0 47 3 | b = Iris-versicolor

0 2 48 | c = Iris-virginica

As we can see the overall classification accuracy is 96% and the Precision and recall scores are 100% and 98% respectively. The overall Weka workflow is shown below:

A screenshot of a computer

Description automatically generated

The next step is to visualize the results using the tree visualizer and is shown as follows:

A close up of a map

Description automatically generated

Thus, the decision tree classification performs quite well with the Iris dataset. The next task at hand is to perform Hierarchical clustering.

**Hierarchical clustering:**

The next task at hand is to perform clustering. The workflow in Weka software is as follows:

A screenshot of a computer

Description automatically generated

A single link clustering strategy is used on the Iris dataset and its corresponding results are printed below:

=== Run information ===

Scheme: weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Relation: iris

Instances: 150

Attributes: 5

sepallength

sepalwidth

petallength

petalwidth

class

Test mode: evaluate on training data

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

Cluster 0

(((((((((((((((((((((0.0:0.03254,0.0:0.03254):0.00913,(0.0:0.03254,0.0:0.03254):0.00913):0.00332,((0.0:0.02778,0.0:0.02778):0.00476,0.0:0.03254):0.01244):0,0.0:0.04498):0.0051,0.0:0.05008):0.00364,0.0:0.05371):0.00437,(0.0:0.05085,0.0:0.05085):0.00724):0.01535,(0.0:0.06731,0.0:0.06731):0.00612):0.00188,0.0:0.07531):0.00196,0.0:0.07728):0.00536,((((((0.0:0.04383,0.0:0.04383):0.00625,0.0:0.05008):0,0.0:0.05008):0.00279,(((((0.0:0.03254,0.0:0.03254):0.01129,0.0:0.04383):0.00116,0.0:0.04498):0.0051,0.0:0.05008):0.00279,((0.0:0,0.0:0):0,0.0:0):0.05287):0):0.00522,0.0:0.05808):0.01919,((0.0:0.04498,0.0:0.04498):0.01549,0.0:0.06047):0.0168):0.00536):0.00165,0.0:0.08429):0.00356,(((0.0:0.02778,0.0:0.02778):0.04371,((0.0:0.04498,0.0:0.04498):0.01394,0.0:0.05893):0.01256):0.00809,0.0:0.07958):0.00826):0.00212,0.0:0.08996):0.00321,0.0:0.09317):0.00598,(0.0:0.0678,0.0:0.0678):0.03135):0.00292,0.0:0.10206):0.01316,0.0:0.11523):0.01375,(0.0:0.12263,(0.0:0.10346,0.0:0.10346):0.01917):0.00634):0.00241,0.0:0.13139):0.12414,0.0:0.25553)

Cluster 1

((((((((((((((((((1.0:0.07344,(((1.0:0.06508,1.0:0.06508):0.00066,(1.0:0.05008,1.0:0.05008):0.01566):0.00224,1.0:0.06798):0.00546):0.00188,(1.0:0.07137,(1.0:0.05556,1.0:0.05556):0.01581):0.00395):0.00733,(1.0:0.07137,((1.0:0.04498,1.0:0.04498):0.01549,1.0:0.06047):0.01089):0.01127):0.00515,1.0:0.08779):0.00538,1.0:0.09317):0.00405,1.0:0.09722):0.0004,(1.0:0.05556,1.0:0.05556):0.04207):0.00152,(1.0:0.07344,1.0:0.07344):0.02571):0,1.0:0.09914):0.00432,(((((1.0:0.08333,1.0:0.08333):0.00613,((((1.0:0.06574,((1.0:0.05287,1.0:0.05287):0,(1.0:0.05287,(1.0:0.04498,1.0:0.04498):0.00789):0):0.01287):0.0077,(1.0:0.04498,1.0:0.04498):0.02845):0,1.0:0.07344):0.0093,(1.0:0.05287,(1.0:0.04498,1.0:0.04498):0.00789):0.02987):0.00672):0.0005,1.0:0.08996):0.00406,1.0:0.09402):0.00041,1.0:0.09443):0.00902):0.00268,1.0:0.10614):0.02,1.0:0.12614):0.00518,1.0:0.13132):0.0066,(1.0:0.12951,1.0:0.12951):0.00841):0.00697,(1.0:0.09869,1.0:0.09869):0.0462):0.01518,(((1.0:0.05008,1.0:0.05008):0.04555,1.0:0.09562):0.03389,1.0:0.12951):0.03056):0.00969,1.0:0.16976):0.83409,((((((((((((((2.0:0.08983,(2.0:0.06047,2.0:0.06047):0.02935):0.01175,2.0:0.10158):0.01245,(2.0:0.10743,((((2.0:0.07148,(2.0:0.05008,2.0:0.05008):0.02141):0.02614,(2.0:0.08504,2.0:0.08504):0.01258):0.00852,(((2.0:0.05287,2.0:0.05287):0.04475,((((2.0:0.04383,2.0:0.04383):0.03881,2.0:0.08264):0.00719,(2.0:0.07148,2.0:0.07148):0.01834):0.00487,2.0:0.0947):0.00292):0.00534,2.0:0.10296):0.00318):0,((2.0:0.09415,(2.0:0.04167,2.0:0.04167):0.05249):0.01199,(((2.0:0.08429,(2.0:0.05287,2.0:0.05287):0.03142):0.00518,((2.0:0.03254,2.0:0.03254):0.0254,2.0:0.05794):0.03152):0.00524,2.0:0.0947):0.01144):0):0.00129):0.0066):0.02063,(((2.0:0,2.0:0):0.08779,2.0:0.08779):0.01089,2.0:0.09869):0.03597):0.00139,2.0:0.13605):0.0016,2.0:0.13765):0.01061,(((2.0:0.09869,2.0:0.09869):0.02337,2.0:0.12206):0.01586,((2.0:0.07344,2.0:0.07344):0.05554,(2.0:0.12263,2.0:0.12263):0.00634):0.00895):0.01034):0.00275,2.0:0.15102):0.00299,2.0:0.15401):0.02491,2.0:0.17892):0.01985,2.0:0.19878):0.00279,2.0:0.20156):0.02691,(2.0:0.11232,2.0:0.11232):0.11615):0.0402,2.0:0.26868):0.73517)

Time taken to build model (full training data) : 0.15 seconds

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

Clustered Instances

0 50 ( 33%)

1 100 ( 67%)

As we can see the single link strategy created two clusters with 100 instances in cluster 1 and 50 other data points in the second cluster. The corresponding dendrogram and visualization is shown below:

A screenshot of a cell phone

Description automatically generated

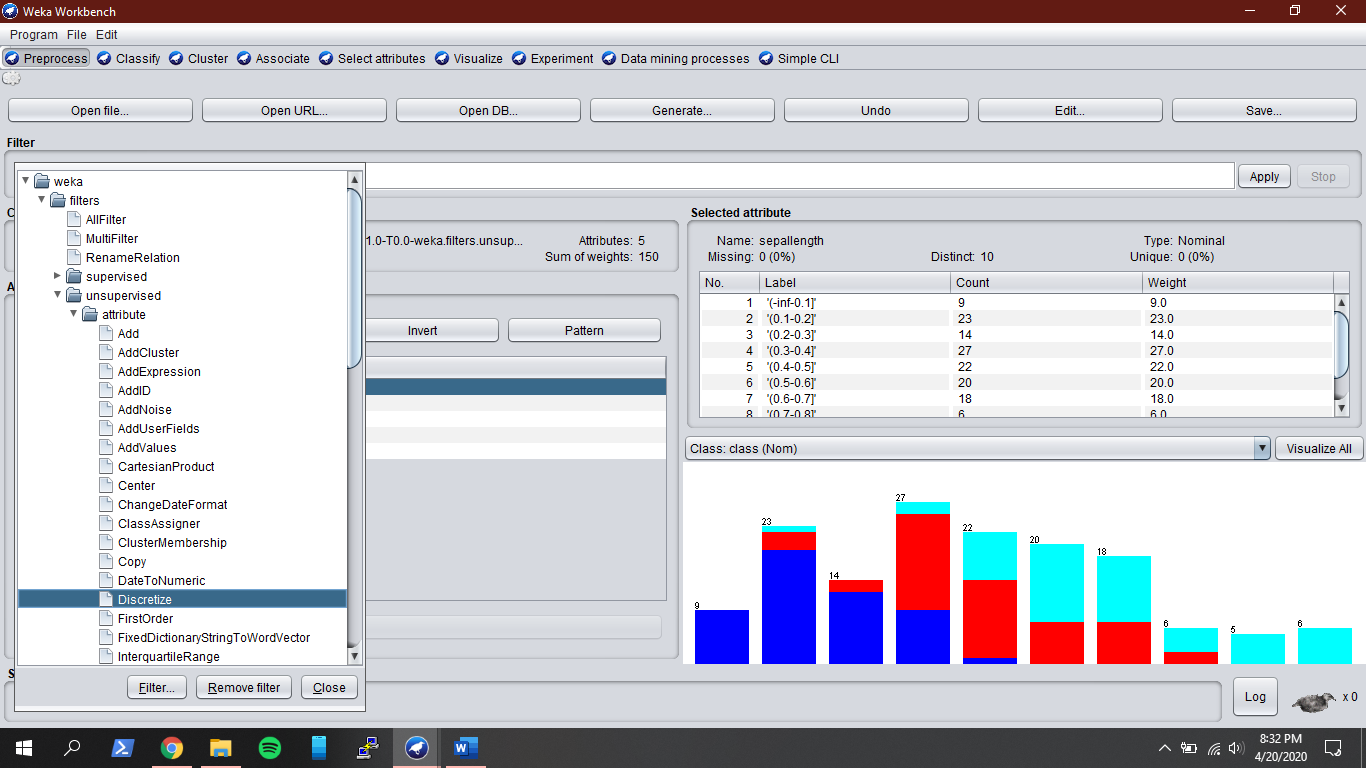
A screenshot of a social media post

Description automatically generated

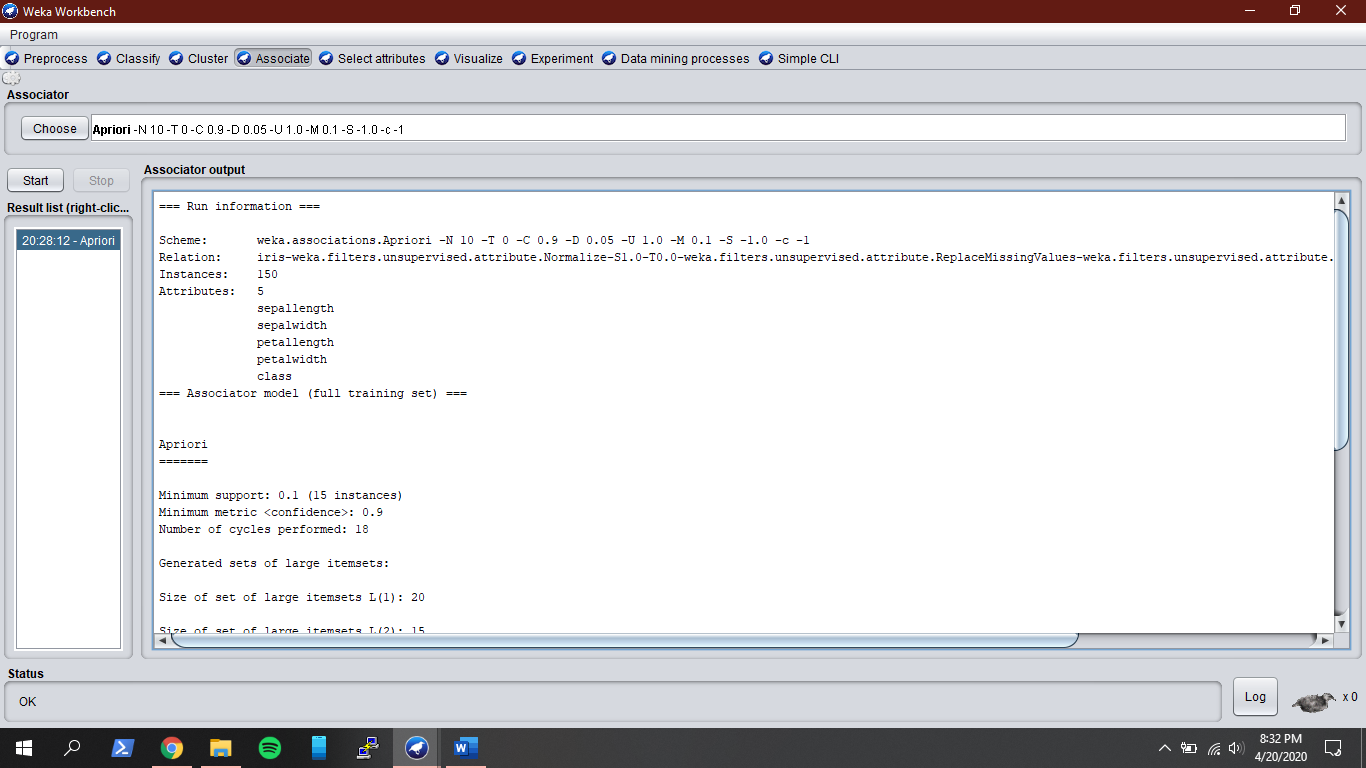
**Association Rule Mining:**

The third task is association rule mining using Apriori algorithm.

The dataset is discretized so as to apply the Apriori algorithm. The work flow is as follows:



Then the Apriori algorithm is applied to the processed dataset as follows:



The associated rules created are as following:

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: iris-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6

Instances: 150

Attributes: 5

sepallength

sepalwidth

petallength

petalwidth

class

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (15 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 15

Size of set of large itemsets L(3): 3

Best rules found:

1. petalwidth='(-inf-0.1]' 41 ==> class=Iris-setosa 41 <conf:(1)> lift:(3) lev:(0.18) [27] conv:(27.33)

2. petallength='(-inf-0.1]' 37 ==> class=Iris-setosa 37 <conf:(1)> lift:(3) lev:(0.16) [24] conv:(24.67)

3. petallength='(-inf-0.1]' petalwidth='(-inf-0.1]' 33 ==> class=Iris-setosa 33 <conf:(1)> lift:(3) lev:(0.15) [22] conv:(22)

4. petalwidth='(0.4-0.5]' 21 ==> class=Iris-versicolor 21 <conf:(1)> lift:(3) lev:(0.09) [14] conv:(14)

5. petallength='(0.7-0.8]' 18 ==> class=Iris-virginica 18 <conf:(1)> lift:(3) lev:(0.08) [12] conv:(12)

6. sepallength='(0.1-0.2]' petalwidth='(-inf-0.1]' 17 ==> class=Iris-setosa 17 <conf:(1)> lift:(3) lev:(0.08) [11] conv:(11.33)

7. sepalwidth='(0.4-0.5]' class=Iris-setosa 16 ==> petalwidth='(-inf-0.1]' 16 <conf:(1)> lift:(3.66) lev:(0.08) [11] conv:(11.63)

8. sepalwidth='(0.4-0.5]' petalwidth='(-inf-0.1]' 16 ==> class=Iris-setosa 16 <conf:(1)> lift:(3) lev:(0.07) [10] conv:(10.67)

9. petallength='(0.5-0.6]' 26 ==> class=Iris-versicolor 25 <conf:(0.96)> lift:(2.88) lev:(0.11) [16] conv:(8.67)

10. petalwidth='(0.7-0.8]' 23 ==> class=Iris-virginica 22 <conf:(0.96)> lift:(2.87) lev:(0.1) [14] conv:(7.67)

Thus, the tasks of classification, rule mining and clustering is complete. The complete project file along with the results can be found in the their individual project files.