CSE 634 Data Mining

CLASSIFICATION PROJECT REPORT

By:
Debjyoti Roy - 112070373
Pavan Kumar Peela - 112045988
Sai Teja - 111987438
Siddharth Harinarayanan – 112026390

INTRODUCTION

The goal of this project is to use an Internet-based classification tool - WEKA - to build two types of classifiers: a descriptive and a non-descriptive classifier, and compare these two approaches based on the results;

1. Descriptive Classifier:

Use a Decision Tree tool to generate sets of discriminant rules describing the content of the data.

2. Non-Descriptive Classifier:

Use Neural Networks tool to build your Classifier.

DATA PREPARATION:

The given dataset is a real-life classification data with TYPE DE ROCHE (Rock Type) as a CLASS attribute. There are 98 records with 48 attributes and a total of 6 classes.

For a Classifier to train properly, a paramount facet is to have a top-hole dataset, in this event, we define the approach of data cleaning below;

- Format Conversion:

  The first and foremost aspect is to load the data into WEKA, the given dataset is in .xls format, but WEKA does not support the format, so we embraced the supported format of .csv and converted the given file into a .csv file.

- Structure Cleansing:

  We observed that the .xls file has empty sheets which are of no purpose, so we discarded them.

- Segregation of Classes:

  C1 : R. Carbonatees AND R. Carbonatees impures (both classes were combined)
  C2 : Pyrate (Pyyrite and Pyrite classes are combined and changed)
  C3 : Charcopyrite
  C4 : Galene
  C5 : Spahlerite
  C6 : Sediments terrigenes

- Miswrites in Class Names:

  On a broad perspective, the classes are segregated into 6 types. A first glance into the data, we found that “R Carbonatees” and “R. Carbonatees impures” are separately tagged, so we blended them into one category. There was another observation that “Pyrate” is spelled as “Pyyrite and Pyrite”, these have been handled.

- Redundant Columns:

  Weka does not allow to load the .csv if there are any duplicated attribute columns. In this event, we removed the additional Type De Roche column which was appearing twice.

- Redundant Rows:

  Additionally, there were two undesirable rows present at the end of the file which we removed as well.

- Fortuitous Value:

  Concealed in the “Li” attribute there is a value with value “< 0.3”. This was causing a problem when implementing the decision tree classifier. So, for this sake, we replaced it to “0.2”.

- Missing Values:

  After triumphantly loading of data into WEKA, we observed amplitude of values missing in some columns specifically in Co and Mo, to fill in these missing values we used “ReplaceMissingValues” from filter section to populate the missing values of attributes in the dataset with its corresponding mean and mode.

On Completion of this process the output we get is an impeccable dataset which can now be used to build a classifier for accurate classifications.

DATA PREPROCESSING:

In consonance with the problem statement defined we need to build 2 Decision Trees (*Descriptive Classifier*) on data after implementing two different data discretization techniques and also a Neural Network (*Non-Descriptive Classifier*) on the data which we cleaned and after implementation of a normalized technique.

- Decision Tree Classifier:

  We built a Decision Tree *(Descriptive Classifier)* on two sets of data after applying two different types of data discretization techniques. A circumspect reason to use discretization is that we need to classify This reduces the number of values for continuous attribute values. This results in achieving better accuracies at a faster pace, and also easier for the algorithm to learn.

- Neural Network Classifier:

  In order to build a neural network (non-descriptive classifier), we used the Normalize preprocess filter which is in-built in Weka.

CLASSIFIER CONSTRUCTION:

Firstly, we will be discretizing the data using two techniques namely equal-width binning and equal-frequency binning.

Equal-width Binning pertains to discretize values by creating bins of equal width intervals whereas the Equal-frequency Binning pertains to discretize values by creating bins in which each bin consists of an approximately same number of values.

Cross-Validation:

In our experimentation, we have used 10-fold cross-validation to train the model on the training and validation data.

EXPERIMENT 1 (Full Classification):

1. Equal Width binning Decision Tree Classification

```
Classifier Model J48 pruned tree
K2O = '(-inf-0.82]'
|   Pb = '(-inf-4103.75]'
|   |   Sc = '(-inf-2.45]'
|   |   |   Fe2O3* = '(-inf-0.455]': R. Carbonatees AND R.
Carbonatees impures (74.0/1.0)
|   |   |   Fe2O3* = '(0.455-inf)'
|   |   |   |   S = '(-inf-2021]': R. Carbonatees AND R.
Carbonatees impures (4.0/1.0)
|   |   |   |   S = '(2021-inf)': Pyrate (4.0)
|   |   Sc = '(2.45-inf)': Charcopyrite (3.0/1.0)
|   Pb = '(4103.75-5694.5]': Spahlerite (1.0)
|   Pb = '(5694.5-inf)': Galene (3.0)
K2O = '(0.82-inf)': Sediments terrigenes (9.0)

Number of Leaves: 7, Size of the tree: 12
```

Discriminant Rules:

Rule 1. If the K2O is less than 0.82, Pb is less than 4103.75, Sc is less than 2.45, and Fe2O3 is less than 0.455 then the type of rock is R. Carbonatees and R. Carbonatees impure.

Rule 2. If the K2O is less than 0.82, Pb is less than 4103.75, Sc is less than 2.45, Fe2O3 is more than 0.455, and S is less than 2021, then the type of rock is R. Carbonatees and R. Carbonatees impure.

Rule 3. If the K2O is less than 0.82, Pb is less than 4103.75, Sc is less than 2.45, Fe2O3 is more than 0.455, and S is less than 2021, then the type of rock is Pyrate

Rule 4. If the K2O is less than 0.82, Pb is less than 4103.75, then Sc is more than 2.45, then the type of rock is Charcopyrate.

Rule 5. If the K2O is less than 0.82, and Pb is more than 4103.75 and less than 5694.5, then the type of rock is Spahlerite.

Rule 6. If the K2O is less than 0.82, and Pb is more than 5694.5, then the type of rock is Galene.

Rule 7. If K2O is more than 0.82, then the type of rock is Terrigenes.

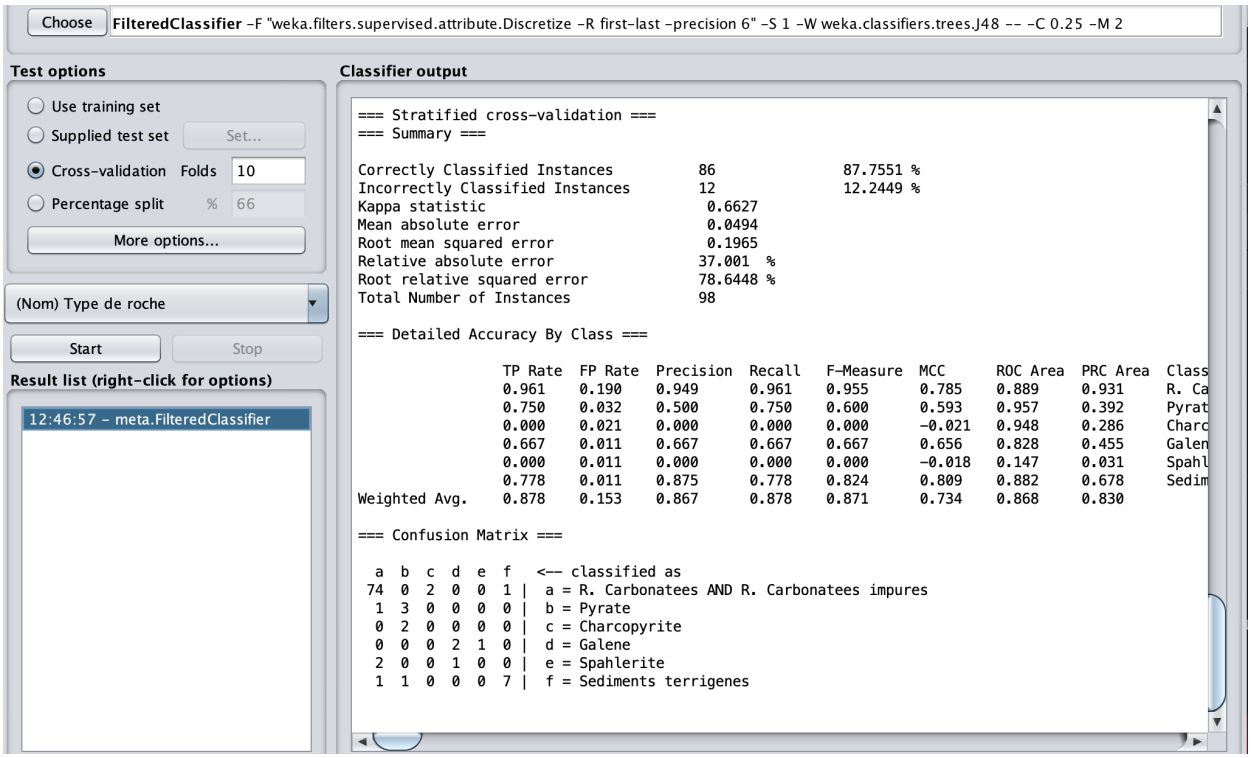


Fig. Screengrab of the output of J48 pruned tree

2. Equal Frequency Discretization Decision Tree Classification

```
Classifier Model - J48 pruned tree
K2O_1 = '(-inf-0.82]'
|   Zn_2 = '(-inf-1309.536842]'
|   |   Tb_2 = '(-inf-0.9]'
|   |   |   S_1 = '(-inf-2021]': R. Carbonatees and R.
Carbonatees impures (74.0)
|   |   |   S_1 = '(2021-inf)'
|   |   |   |   Fe2O3*_1 = '(-inf-0.455]': R. Carbonatees
and R. Carbonatees impures (3.0)
|   |   |   |   Fe2O3*_1 = '(0.455-inf)': Pyrate (4.0)
|   |   Tb_2 = '(0.9-inf)': Chalcopyrite (2.0)
|   Zn_2 = '(1309.536842-inf)'
|   |   Pb_2 = '(-inf-5694.5]': Spahlerite (3.0)
|   |   Pb_2 = '(5694.5-inf)': Galene (3.0)
K2O_1 = '(0.82-inf)': Sediments terrigenes (9.0)
Number of Leaves: 7, Size of the tree: 13
```

Discriminant Rules:

Rule 1. If the K2O is less than 0.82, Zn is less than 1309.54, Tb is less than 0.9, and S is less than 2021 then the type of rock is R. Carbonatees and R. Carbonatees impure.

Rule 2. If the K2O is less than 0.82, Zn is less than 1309.54, Tb is less than 0.9, S is more than 2021, and Fe2O3 is less than 0.455 then the type of rock is R. Carbonatees and R.Cabonatees impure.

Rule 3. If the K2O is less than 0.82, Zn is less than 1309.54, Tb is less than 0.9, S is more than 2021, and Fe2O3 is more than 0.455 then the type of rock is Pyrate.

Rule 4. If the K2O is less than 0.82, Zn is less than 1309.54, Tb is more than 0.9, then the type of rock is Chalcopyrite.

Rule 5. If the K2O is less than 0.82, Zn is less than 1309.54, and Pb is less than 5694.5 then the type of rock is Saphlerite.

Rule 6. If the K2O is less than 0.82, Zn is less than 1309.54, and Pb is more than 5694.5 then the type of rock is Galene.

Rule 7. If K2O is more than 0.82, then the type of rock is Terrigenes.

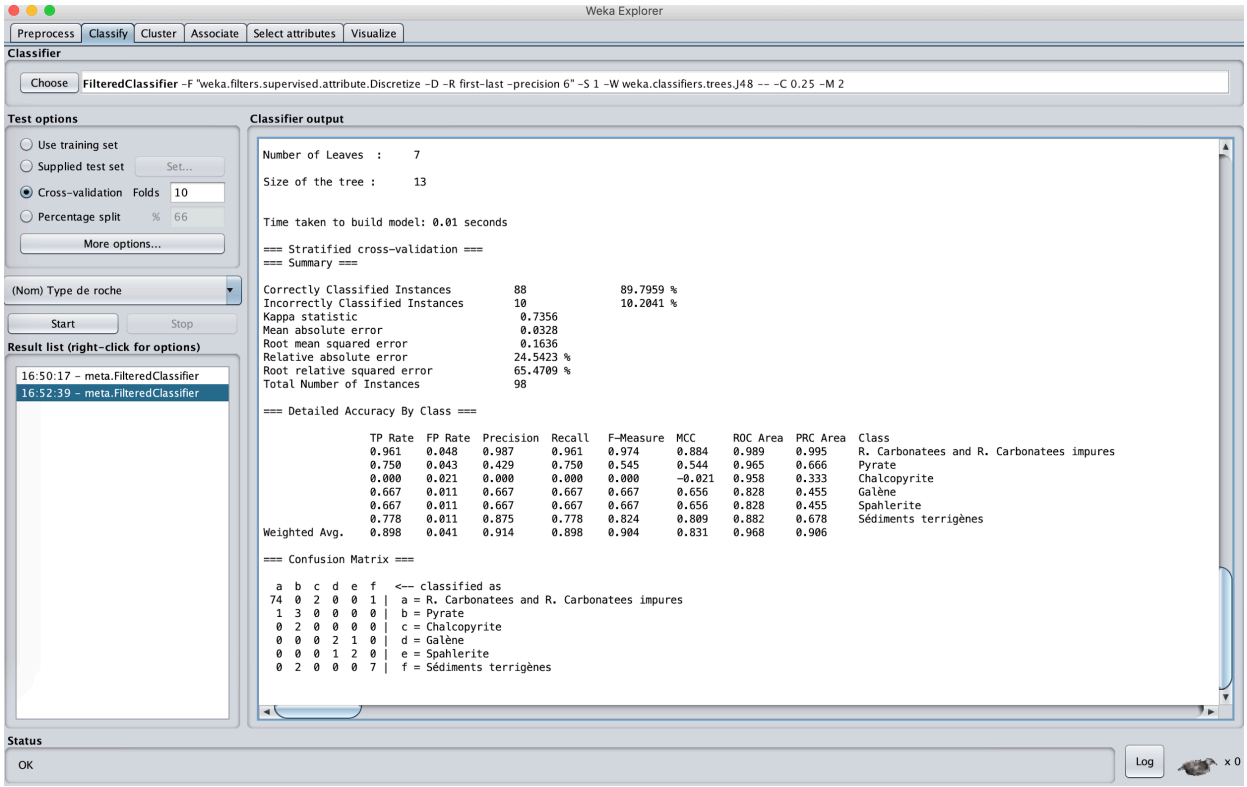


Fig. Screengrab of the output of J48 pruned tree

Non-Descriptive Classifier

For non-descriptive classifier, we used multi perceptron neural network and following are the parameters used during the classification.

1. Basic Neural Network

Parameter	Value
Learning Rate	0.3
Epochs	500
Hidden Layers	Input and output average
Percentage Split	80
Momentum	0.2
Accuracy (%)	95

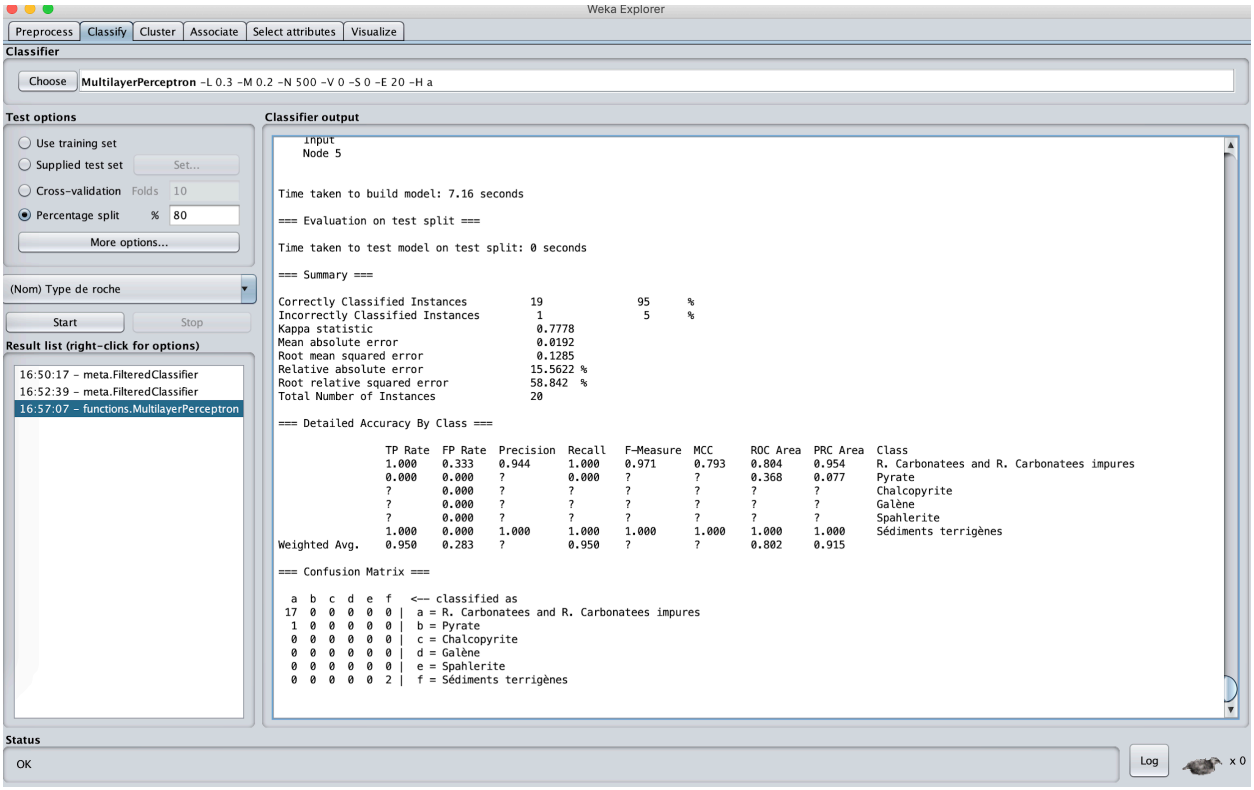


Fig. Screengrab of the output of Multilayer Perceptron Neural Network

2. Neural Network with 10-Cross Validation:

Parameter	Value
Learning Rate	0.3
Epochs	500
Hidden Layers	Input and output average
Percentage Split	80
Momentum	0.2
Accuracy (%)	88.77

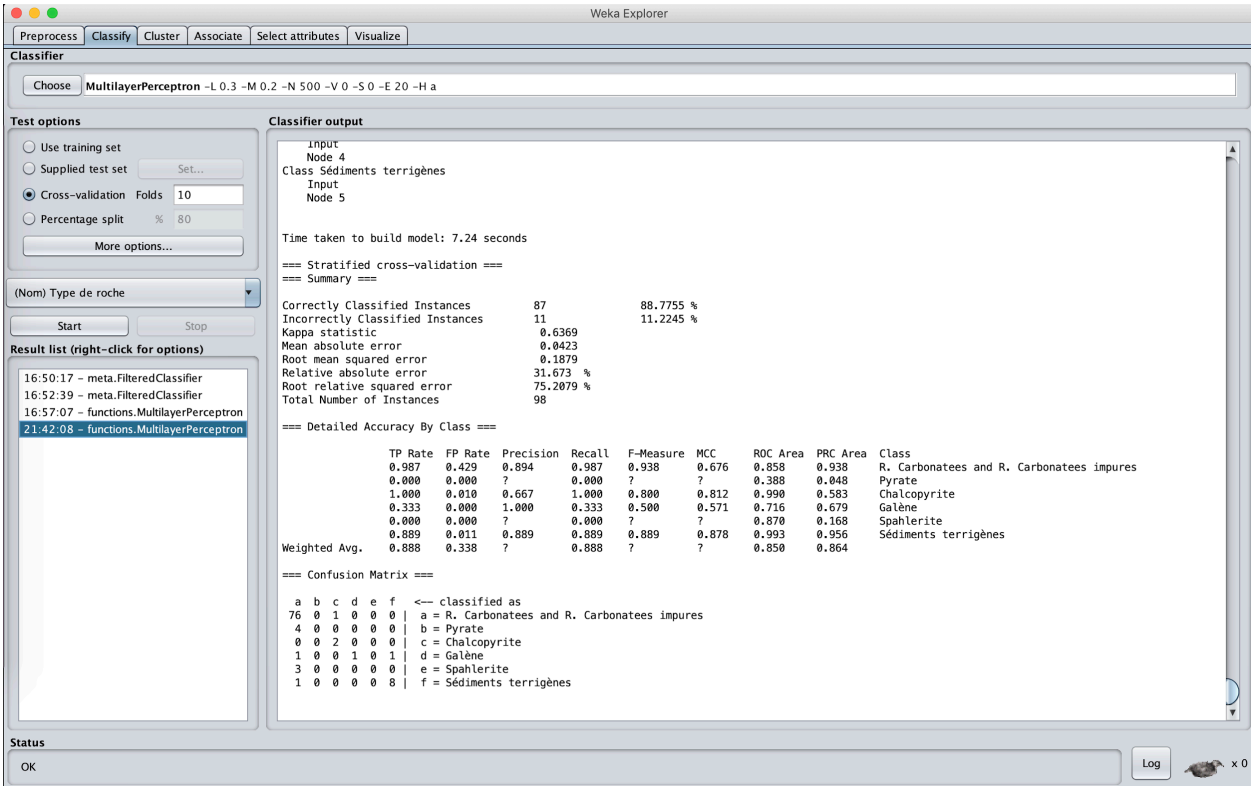


Fig. Screengrab of the output of Multilayer Perceptron Neural Network

3. Neural Network with 300 Epochs

Parameter	Value
Learning Rate	0.3
Epochs	300
Hidden Layers	Input and output average
Percentage Split	80
Momentum	0.2
Accuracy (%)	95

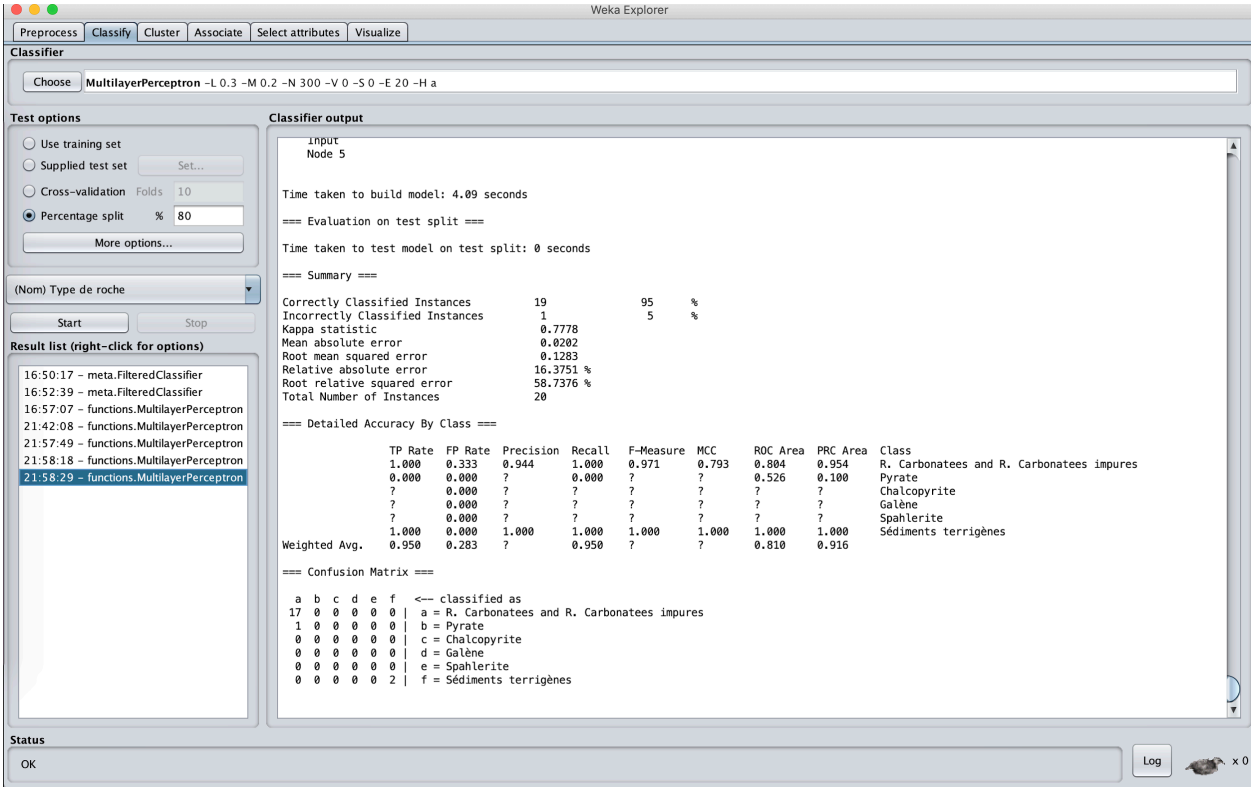


Fig. Screengrab of the output of Multilayer Perceptron Neural Network

4. Discretization under Supervised Learning

Motivation:

The main motivation to try out this method is to experiment the variation in the accurately classified instances. This is view of the property of the Supervised Discretization technique that it also considers the class values into consideration contrasting the un-supervised discretization which don't consider the class values.

Weka supports Supervised Discretization, so we ran the data into this filter and then built the different classifier which were required to build.

```
Classifier Model - J48 pruned tree
K2O = '(-inf-0.82]'
|   Pb = '(-inf-4103.75]'
|   |   Sc = '(-inf-2.45]'
|   |   |   Fe2O3* = '(-inf-0.455]': R. Carbonatees AND R.
Carbonatees impures (74.0/1.0)
|   |   |   Fe2O3* = '(0.455-inf)'
|   |   |   |   S = '(-inf-2021]': R. Carbonatees AND R.
Carbonatees impures (4.0/1.0)
|   |   |   |   S = '(2021-inf)': Pyrate (4.0)
|   |   Sc = '(2.45-inf)': Charcopyrite (3.0/1.0)
|   Pb = '(4103.75-5694.5]': Spahlerite (1.0)
|   Pb = '(5694.5-inf)': Galene (3.0)
K2O = '(0.82-inf)': Sediments terrigenes (9.0)

Number of Leaves : 7, Size of the tree: 12
```

Discriminant Rules:

Rule 1. If the K2O is less than 0.82, Pb is less than 4103.75, Sc is less than 2.45, and Fe2O3* is less than 0.455 then the type of rock is R. Carbonatees and R. Carbonatees impure.

Rule 2. If the K2O is less than 0.82, Pb is less than 4103.75, Sc is less than 2.45, and Fe2O3* is more than 0.455 and S is less than 2021 then the type of rock is R. Carbonatees and R. Carbonatees impure.

Rule 3. If the K2O is less than 0.82, Pb is less than 4103.75, Sc is less than 2.45, and Fe2O3* is more than 0.455 and S is more than 2021 then the type of rock is Pyrate.

Rule 4. If the K2O is less than 0.82, Pb is less than 4103.75, Sc is more than 2.45 then the type of rock is Chalcopyrite.

Rule 5. If the K2O is less than 0.82, Pb is between 4103.75 and 5694.5 then the type of rock is Saphlerite.

Rule 6. If the K2O is less than 0.82, Pb is more than 5694.5 then the type of rock is Galene.

Rule 7. If K2O is more than 0.82, then the type of rock is Terrigenes.

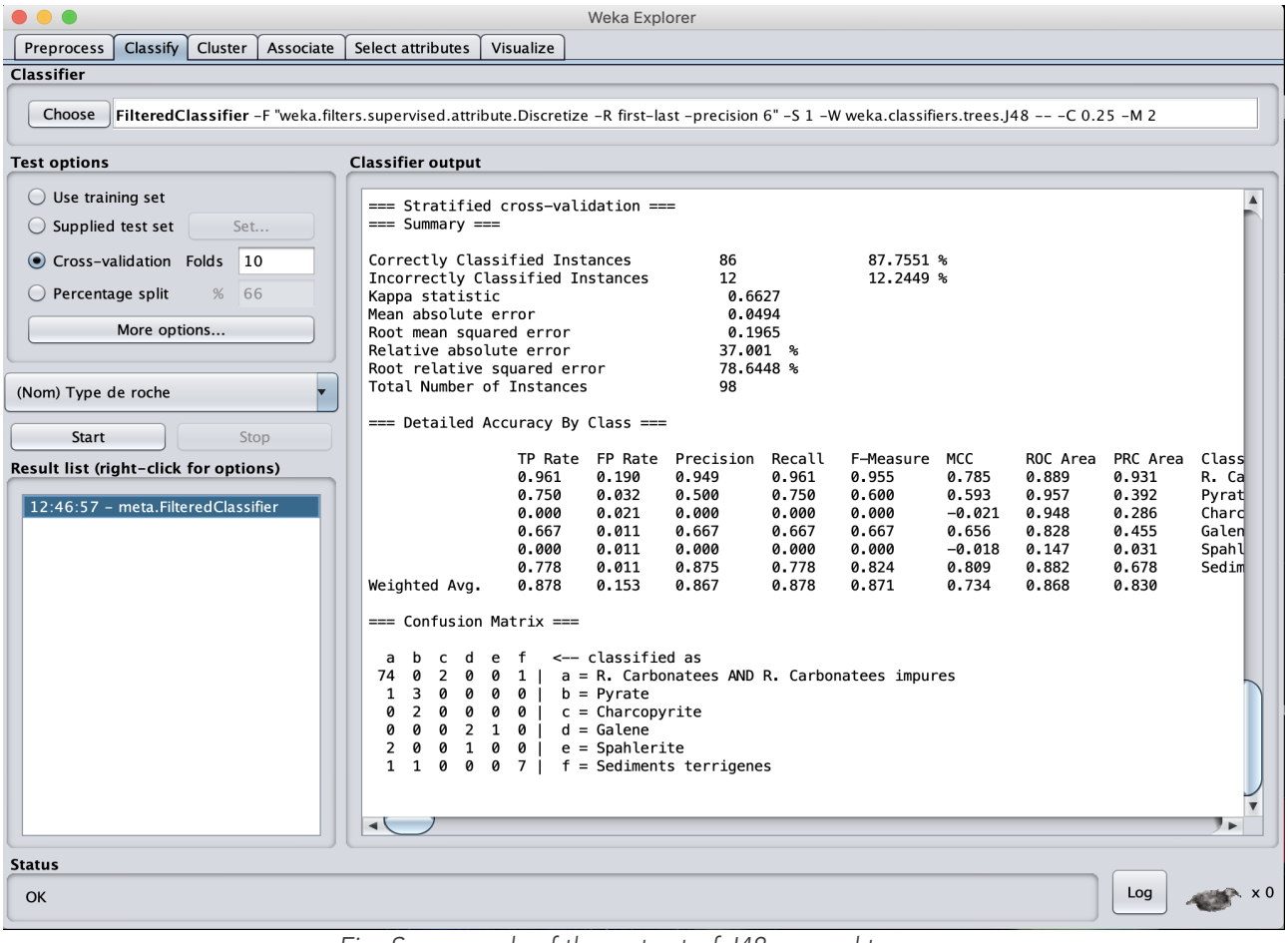


Fig. Screengrab of the output of J48 pruned tree

EXPERIMENT 2 (Contrast Classification):

In contrast classification we have just two classes, C1 which is R. Carbonatees AND R. Carbonatees impures and not C1. Following are the experiments done on this kind of classification.

1. Equal Width binning Decision Tree Classification

C1: R. Carbonatees and R. Carbonatees impure

```
Classifier Model - J48 pruned tree
CaO+MgO_5 = '(-inf-0.5]': Not C1 (10.0) CaO+MgO_5 = '(0.5-inf)'
|   As_2 = '(-inf-0.2]'
|   |   Rb _1 = '(-inf-0.1]': C1 (82.0/6.0)
|   |   Rb _1 = '(0.1-inf)': Not C1 (3.0/1.0)
|  As_2 = '(0.2-inf)': Not C1 (3.0)
Number of Leaves : 4 , Size of the tree : 7
```

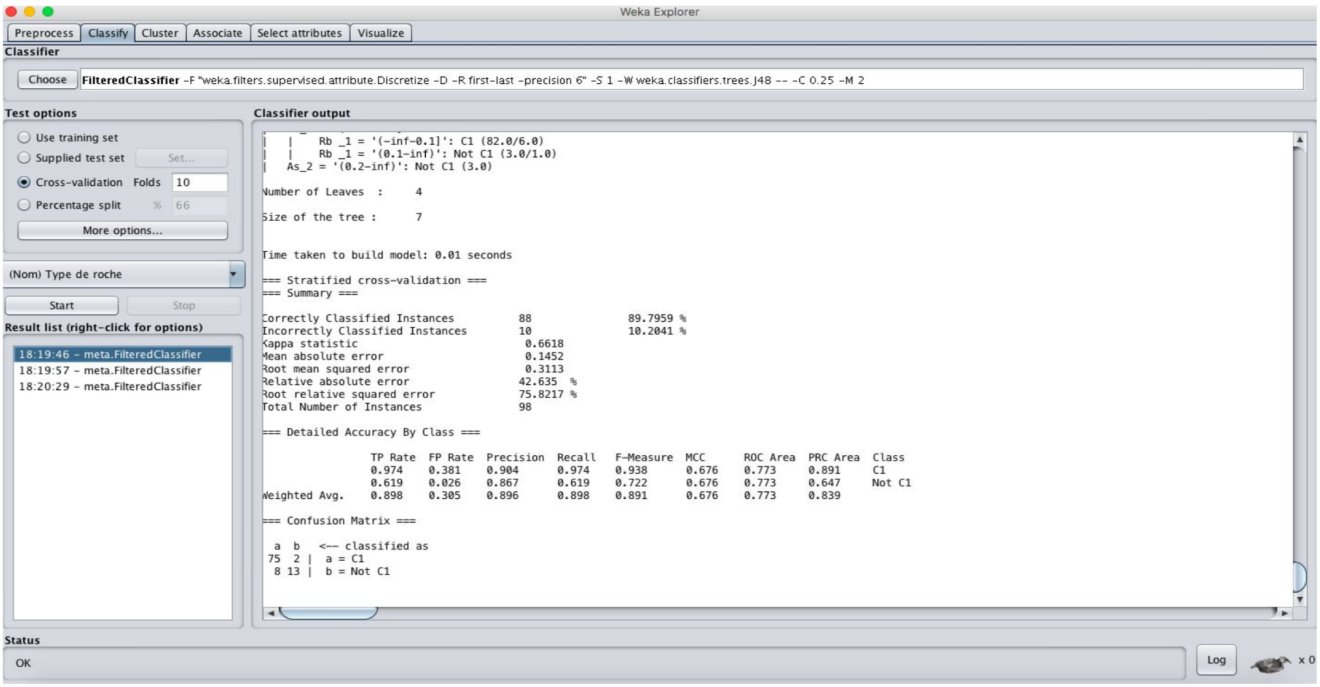
Discriminant Rules

Rule 1. If the CaO_MgO is less than 0.5 then the type of rock is Not C1.

Rule 2. If the CaO_MgO is more than 0.5, As is less than 0.2, and Rb is less than 0.1, then the type of rock is C1.

Rule 3. If the CaO_MgO is more than 0.5, As is less than 0.2, and Rb is more than 0.1, then the type of rock is Not C1.

Rule 4. If the CaO_MgO is more than 0.5, As is more than 0.2, then the type of rock is Not C1.



2. Contrast Classification with equal Frequency

```
Classifier Model, J48 pruned tree
CaO+MgO_1 = '(-inf-0.478646]': Not C1 (10.0)
CaO+MgO_1 = '(0.478646-inf)'
|   S_9 = '(-inf-0.041379]'
|   | Pb_5 = '(-inf-0.038576]'
|   |   |   Cs_5 = '(-inf-0.121621]'
|   |   |   |   Zn_9 = '(-inf-0.013859]': C1 (73.0)
|   |   |   |   Zn_9 = '(0.013859-inf)'
|   |   |   |   |   MnO_5 = '(-inf-0.07377]': C1 (3.0)
|   |   |   |   |   MnO_5 = '(0.07377-inf)': Not C1 (2.0)
|   |   |   Cs_5 = '(0.121621-inf)': Not C1 (3.0/1.0)
|   |   Pb_5 = '(0.038576-inf)': Not C1 (2.0)
|   S_9 = '(0.041379-inf)': Not C1 (5.0)
Number of Leaves : 7
Size of the tree : 13
```

Discriminant Rules

Rule 1. If the CaO_MgO is less than 0.4786 then the type of rock is Not C1.

Rule 2. If the CaO_MgO is more than 0.4786, S is less than 0.04138, Pb is less than 0.03858, Cs is less than 0.1216, and Zn is less than 0.01386, then the type of rock is C1.

Rule 3. If the CaO_MgO is more than 0.4786, S is less than 0.04138, Pb is less than 0.03858, Cs is less than 0.1216, Zn is more than 0.01386, and MnO is less than 0.0738 then the type of rock is C1.

Rule 4. If the CaO_MgO is more than 0.4786, S is less than 0.04138, Pb is less than 0.03858, Cs is less than 0.1216, Zn is more than 0.01386, and MnO is more than 0.0738 then the type of rock is Not C1.

Rule 5. If the CaO_MgO is more than 0.4786, S is less than 0.04138, Pb is less than 0.03858, and Cs is more than 0.1216, then the type of rock is Not C1.

Rule 6. If the CaO_MgO is more than 0.4786, S is less than 0.04138, and Pb is less than 0.03858, then the type of rock is Not C1.

Rule 7. If the CaO_MgO is more than 0.4786, and S is less than 0.04138, then the type of rock is Not C1.

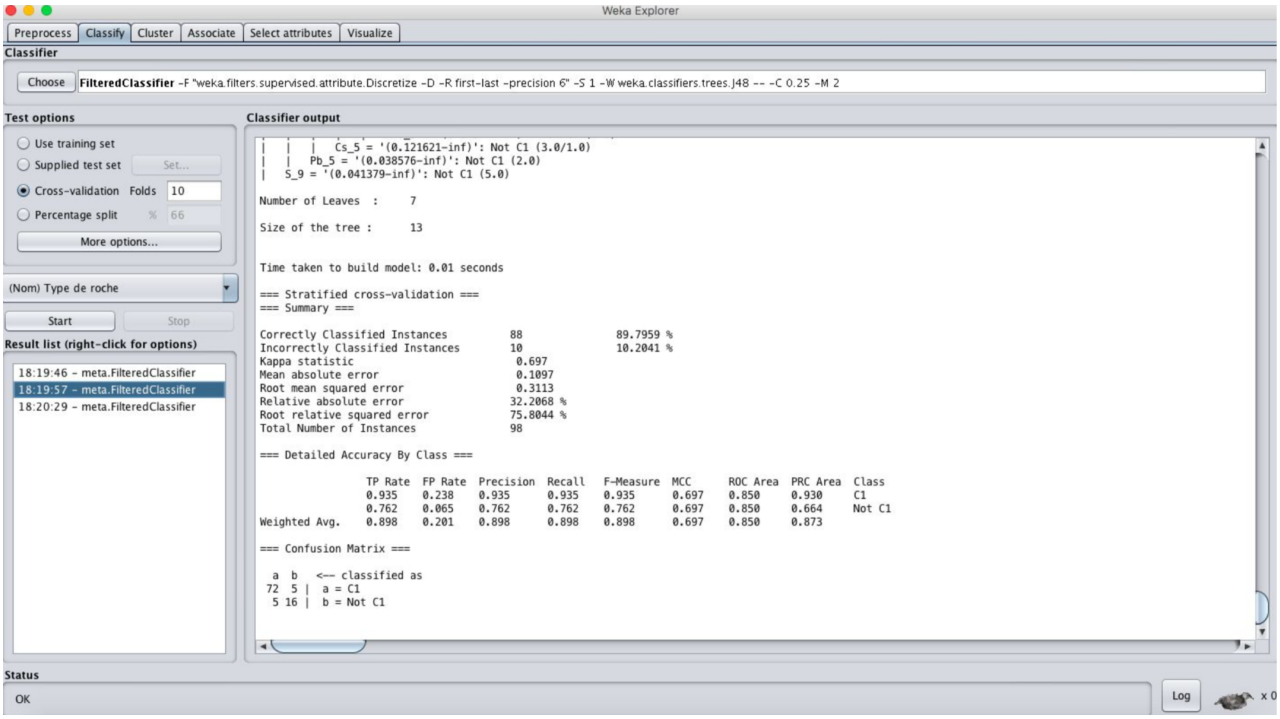


Fig. Screengrab of the output of J48 pruned tree

3. Neural Network with 10-Cross Validation:

Parameter	Value
Learning Rate	0.3
Epochs	500
Hidden Layers	Input and output average
Percentage Split	80
Momentum	0.2
Accuracy (%)	88.77

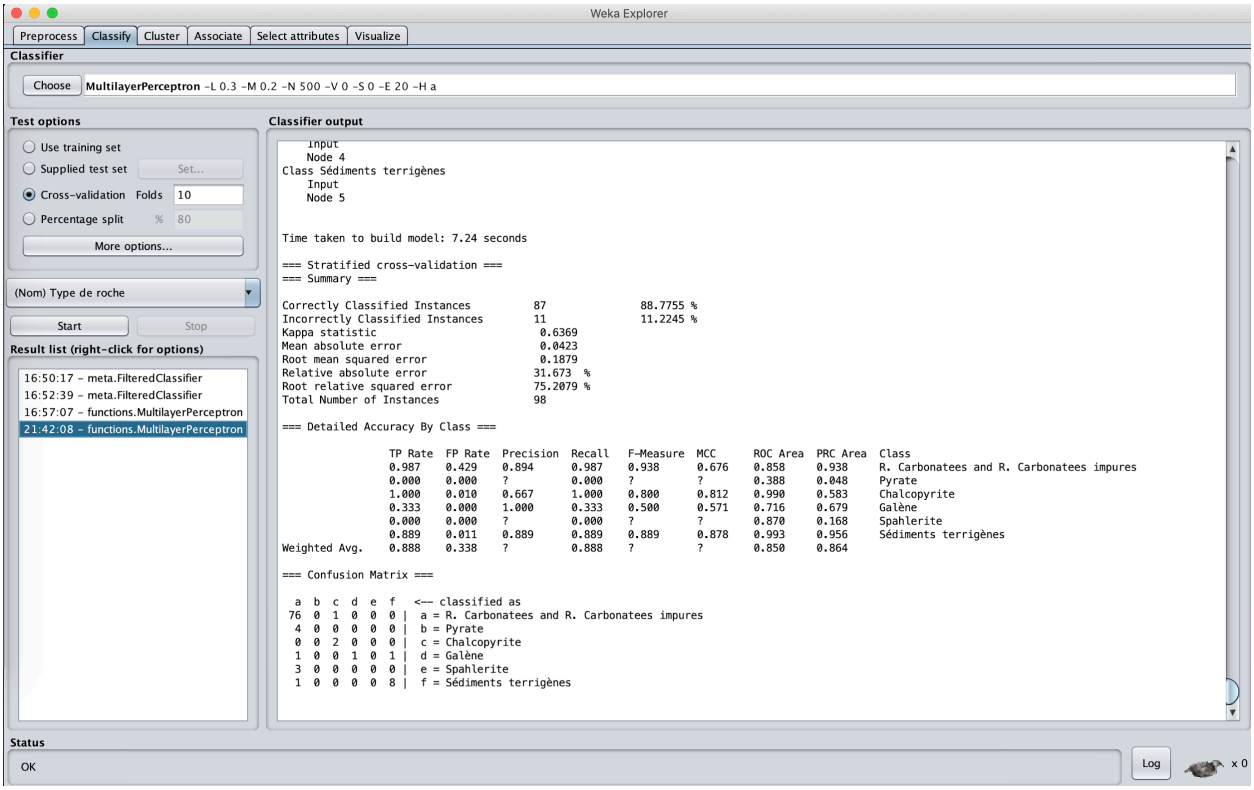


Fig. Screengrab of the output of Multilayer Perceptron Neural Network

4. Discretization under Supervised

```
Classifier Model, J48 pruned tree
Fe2O3* = '(-inf-0.455]': C1 (74.0/1.0)
Fe2O3* = '(0.455-inf)'
|   S = '(-inf-1547.5]'
|   |   K2O = '(-inf-0.795]': C1 (4.0)
|   |   K2O = '(0.795-inf)': Not C1 (4.0)
|   S = '(1547.5-inf)': Not C1 (16.0)

Number of Leaves  : 4, Size of the tree :     7
```

Discriminant Rules

Rule 1. If the Fe2O3 is less than 0.455, then the type of rock is C1.

Rule 2. If the Fe2O3 is more than 0.455, S is less than 1547.5, and K2O is less than 0.795 then the type of rock is C1.

Rule 3. If the Fe2O3 is more than 0.455, S is less than 1547.5, and K2O is more than 0.795 then the type of rock is Not C1.

Rule 4. If the Fe2O3 is more than 0.455, S is more than 1547.5 then the type of rock is Not C1.

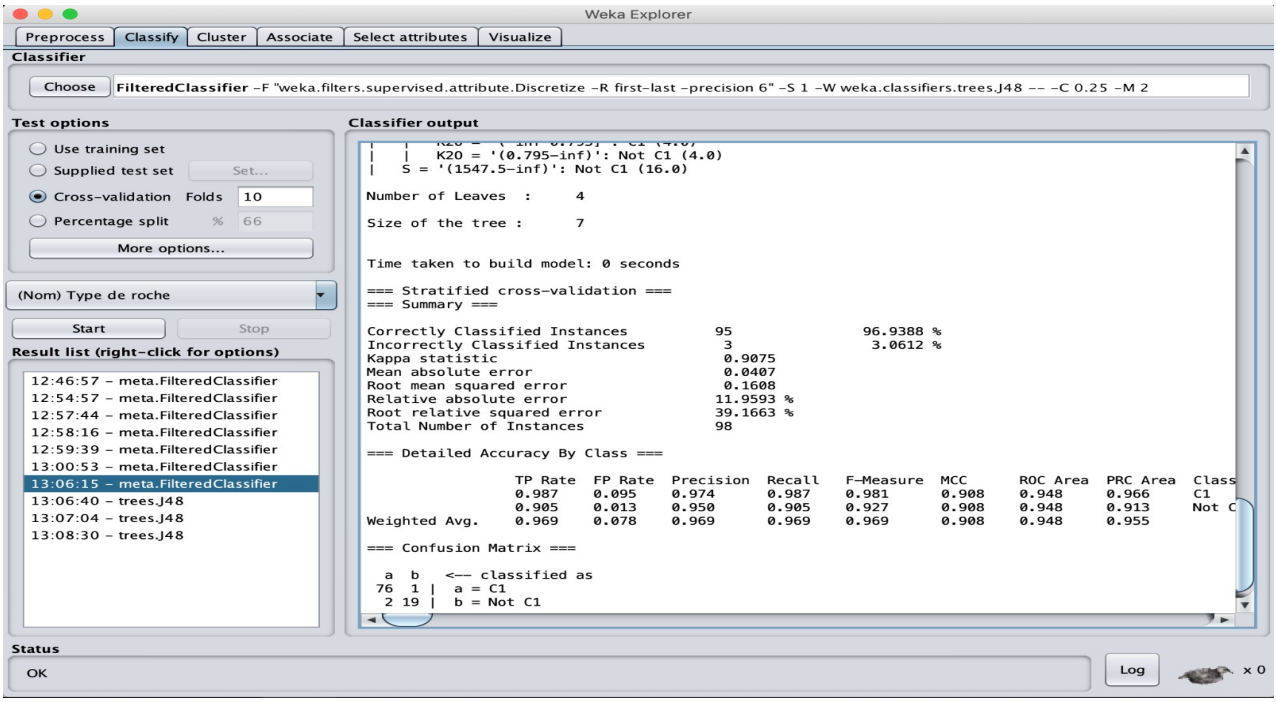


Fig. Screengrab of the output of J48 pruned tree

EXPERIMENT 3 (MOST IMPORTANT FEATURES):

As suggested by an expert, the following are the most important features of the dataset: S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3.

1. Most Important Features using Equal Width Binning

```
Classifier Model, J48 pruned tree
Fe2O3* = '(-inf-0.455]': R. Carbonatees and R. Carbonatees
impures (74.0/1.0)
Fe2O3* = '(0.455-inf)'
|   Pb = '(-inf-4103.75]'
|   |   CaO = '(-inf-23.73]': Sédiments terrigènes
(12.0/3.0)
|   |   CaO = '(23.73-inf)'
|   |   |   S = '(-inf-2021]': R. Carbonatees and R.
Carbonatees impures (5.0/1.0)
|   |   |   S = '(2021-inf)': Pyrate (3.0)
|   Pb = '(4103.75-5694.5]': Spahlerite (1.0)
|   Pb = '(5694.5-inf)': Galène (3.0)
Number of Leaves  :     6
Size of the tree :      10
```

Discriminant Rules

Rule 1. If the Fe2O3 is less than 0.455, then the type of rock is R. Carbonatees and R. Carbonatees impure.

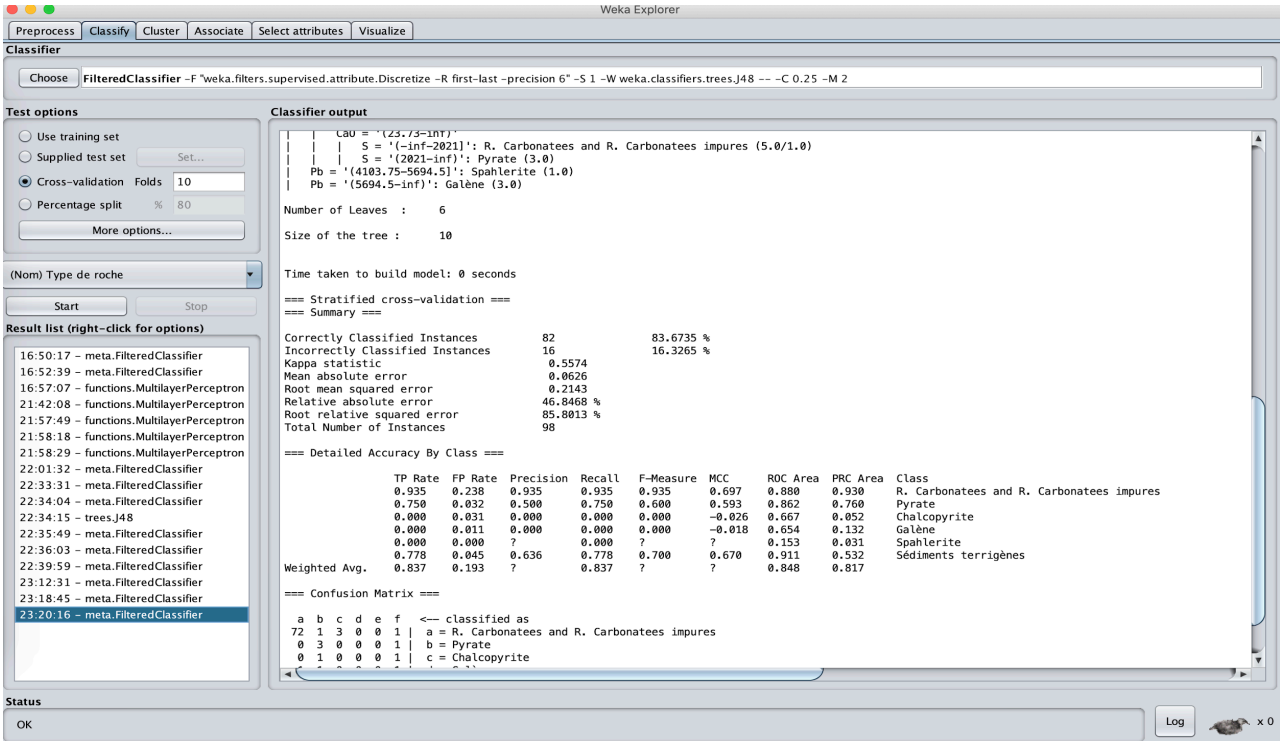
Rule 2. If the Fe2O3 is more than 0.455, Pb is less than 4103.75, and CaO is less than 23.73 then the type of rock is Sediments terrigenes.

Rule 3. If the Fe2O3 is more than 0.455, Pb is less than 4103.75, CaO is more than 23.73, and S is less than 2021, then the type of rock is R. Carbonatees and R. Carbonatees impures.

Rule 4. If the Fe2O3 is more than 0.455, Pb is less than 4103.75, CaO is more than 23.73, and S is more than 2021, then the type of rock is Pyrate.

Rule 5. If the Fe2O3 is more than 0.455, Pb is more than 4103.75 and less than 5694.5, then the type of rock is Spahlerite.

Rule 6. If the Fe2O3 is more than 0.455, Pb is more than 5694.5, then the type of rock is Galene.



2. Most Important Features using Equal Frequency Discretization

```
Classifier Model, J48 pruned tree
Fe2O3*_1 = '(-inf-0.455]': R. Carbonatees and R.
Carbonatees impures (74.0/1.0)
Fe2O3*_1 = '(0.455-inf)'
|   Zn_2 = '(-inf-1309.536842]'
|   |   CaO_1 = '(-inf-23.73]': Sédiments terrigènes
(12.0/3.0)
|   |   CaO_1 = '(23.73-inf)'
|   |   |   S_1 = '(-inf-2021]': R. Carbonatees and R.
Carbonatees impures (4.0)
|   |   |   S_1 = '(2021-inf)': Pyrate (3.0)
|   Zn_2 = '(1309.536842-inf)'
|   |   Pb_2 = '(-inf-5694.5]': Spahlerite (2.0)
|   |   Pb_2 = '(5694.5-inf)': Galène (3.0)
Number of Leaves  :    6
Size of the tree : 11
```

Discriminant Rules

Rule 1. If the Fe2O3 is less than 0.455, then the type of rock is R. Carbonatees and R. Carbonatees impure.

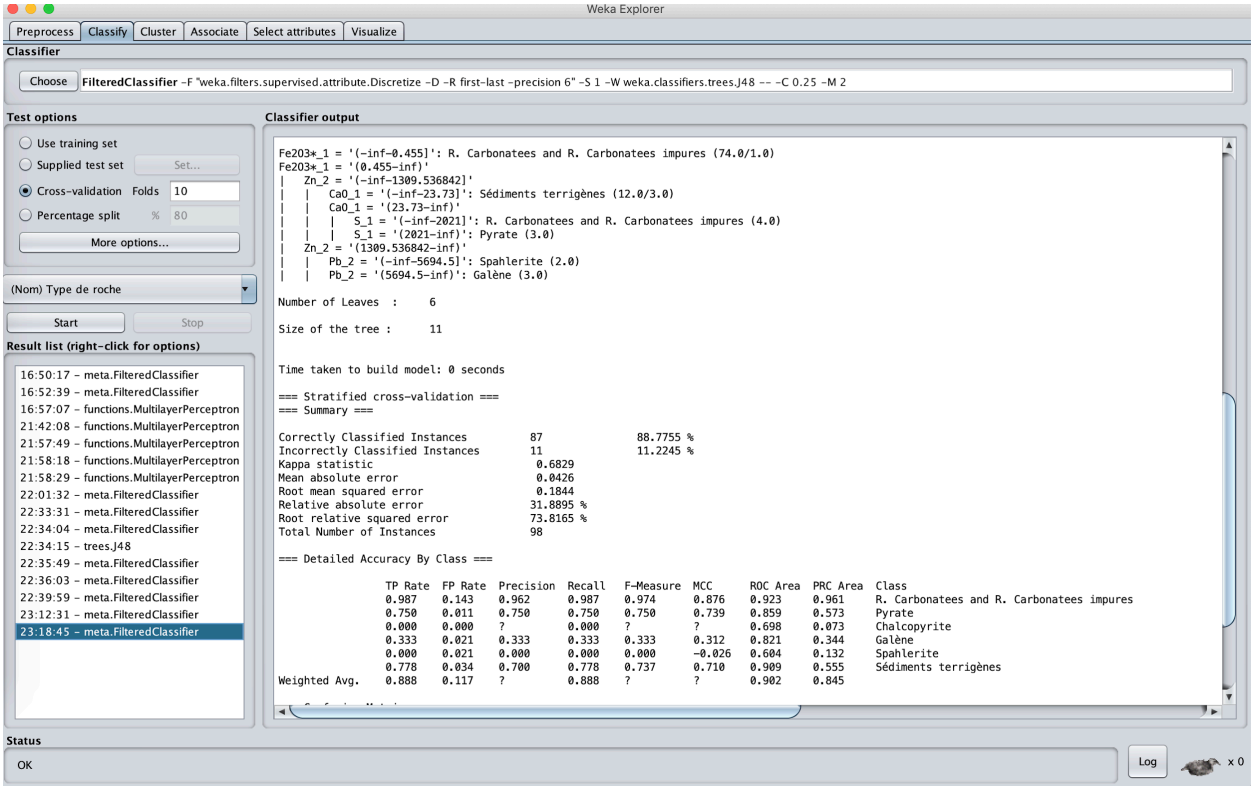
Rule 2. If the Fe2O3 is more than 0.455, Zn is less than 1309.537, and CaO is less than 23.73 then the type of rock is Sediments terrigenes.

Rule 3. If the Fe2O3 is more than 0.455, Zn is less than 1309.537, and CaO is more than 23.73, and S is less than 2021, then the type of rock is R. Carbonatees and R. Carbonatees impures.

Rule 4. If the Fe2O3 is more than 0.455, Zn is less than 1309.537, and CaO is more than 23.73, and S is less than 2021, then the type of rock is R. Carbonatees and R. Carbonatees impures.

Rule 5. If the Fe2O3 is more than 0.455, Zn is more than 1309.537, and Pb is less than 5694.5, then the type of rock is Sphalerite.

Rule 6. If the Fe2O3 is more than 0.455, Zn is more than 1309.537, and Pb is more than 5694.5, then the type of rock is Galene.



3. Most Important Features Neural Network

Parameter	Value
Learning Rate	0.3
Epochs	500
Hidden Layers	Input and output average
Percentage Split	80
Momentum	0.2
Accuracy (%)	95

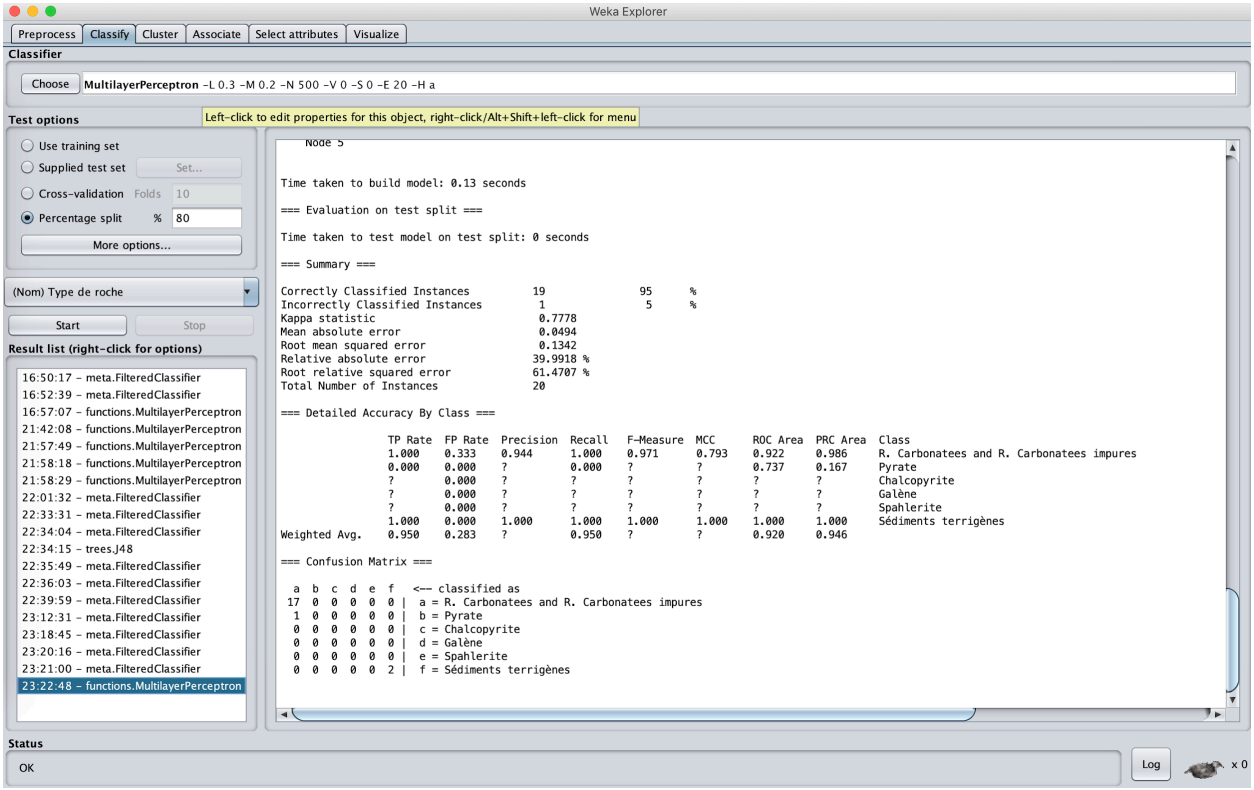


Fig. Screengrab of the output of Multilayer Perceptron Neural Network

4. Discretization Under Supervised Learning

```
Classifier Model, J48 pruned tree
Fe2O3* = '(-inf-0.455]': R. Carbonatees AND R. Carbonatees
impures (74.0/1.0)
Fe2O3* = '(0.455-inf)'
|   Pb = '(-inf-4103.75]'
|   |   CaO = '(-inf-23.73]': Sediments terrigenes
(12.0/3.0)
|   |   CaO = '(23.73-inf)'
|   |   |   S = '(-inf-2021]': R. Carbonatees AND R.
Carbonatees impures (5.0/1.0)
|   |   |   S = '(2021-inf)': Pyrate (3.0)
|   Pb = '(4103.75-5694.5]': Spahlerite (1.0)
|   Pb = '(5694.5-inf)': Galene (3.0)

Number of Leaves  : 6, Size of the tree : 10
```

Discriminant Rules

Rule 1. If the Fe2O3 is less than 0.455, then the type of rock is R. Carbonatees and R. Carbonatees impure.

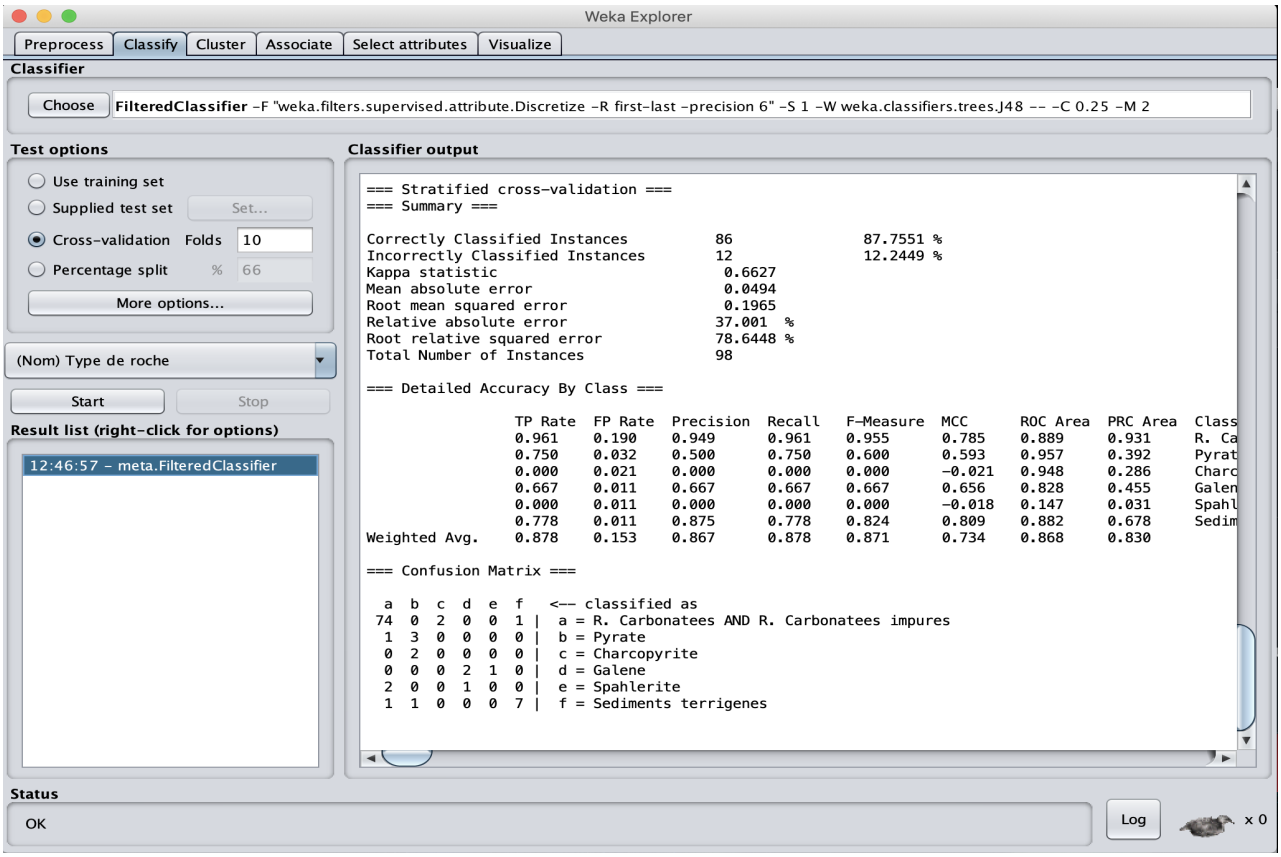
Rule 2. If the Fe2O3 is more than 0.455, Pb is less than 4103.75, and CaO is less than 23.73 then the type of rock is Sediments terrigenes.

Rule 3. If the Fe2O3 is more than 0.455, Pb is less than 4103.75, CaO is more than 23.73, and S is less than 2021, then the type of rock is R. Carbonatees and R. Carbonatees impure.

Rule 4. If the Fe2O3 is more than 0.455, Pb is less than 4103.75, CaO is more than 23.73, and S is more than 2021, then the type of rock is Pyrate.

Rule 5. If the Fe2O3 is more than 0.455, Pb is more than 4103.75 and less than 5694.5, then the type of rock is Spahlerite.

Rule 6. If the Fe2O3 is more than 0.455, Pb is more than 5694.5, then the type of rock is Galene.



Consolidated Accuracy Measures:

Title	Experiment 1	Experiment 2	Experiment 3
Unsupervised Equal Width Binning	95%	89.80%	83.67%
Unsupervised Equal Frequency Binning	88.78%	89.80%	88.78%
Neural Network	95%	88.78%	95%
Discretization under Supervised Learning	87.76%	96.93%	87.75%

CONCLUSION

On completion of all the experiments it has been found that for Full Classification of the given Dataset we achieved maximum accuracy using Unsupervised Equal Width Binning Decision Tree and also with the Multi Perceptron Neural Network. On contrary, when performing Experiment 2 (Contrast Classification), the Supervised Discretization has given the best accuracy which can be clearly understood as this type of Discretization actually considers the values of the class too while discretizing the data while un-supervised don't. For example, the unsupervised *discretize* filter only considers the attribute being discretized. While it can 'optimize' the number of bins, it does so only with respect to self-encoding. However, the supervised *discretize* filter will break the attribute into bins that provide the most information about the *class*. When most important attributes are used for performing Experiment 3, on contrasting with the obtained accuracies we found that Multi-perceptron Neural Network gives the best accuracy since the model actually takes into consideration the important features to build the classifier and by giving more epochs depending on the size of the dataset, the Neural Network can actually be trained on new or unseen data points too, that is the reason we get a high accuracy.