DATA MINING

Project Report

Group Members:

1) Dhivahar Perumal	111465042
2) Soham Mehta	111496015
3) Manu Mathew	111492994
4) Arjun Mathew Dan	111492985
5) Shilpa Mary George	111492833

1. INTRODUCTION

The aim of the project is to learn about the various classification tools and machine learning techniques. We build two types of classifiers i.e. a descriptive and a non-descriptive classifier. We are using the Weka(Waikato Environment for Knowledge Analysis) tool which provides a suite of machine learning algorithms for data mining tasks. We use it for data preparation, pre-processing, classification and visualization.

2. ABOUT DATA

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

The Classes for which we have train our models are:
C1: R. Carbonatees AND R. Carbonatees impures
C2: Pyrate
C3: Charcopyrite
C4: Galene
C5: Spahlerite
C6: Sediments terrigenes

Most important attributes are: S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3 as they have the maximum values provided for training our model.

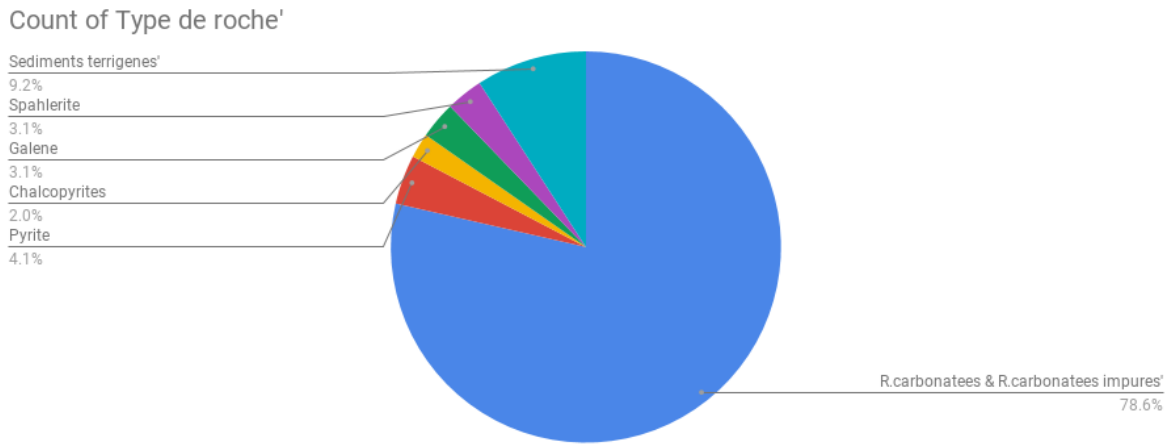
Since this is a real life experimental data, it also contains a lot of missing data and duplicate data. There are also many instances where is the data is not properly formatted or in the correct format. These all corrections are made in the data preparation step.

3. OBSERVATIONS

The data provided has following characteristics. The insights can be obtained about the same by plotting these relevant values. The main observation from the given data is that there is a high variability in the types of rocks in the given data set.

Type de roche'	Number of Rocks
Chalcopyrites	2
Galene	3
Pyrite	4
R.carbonatees & R.carbonatees impures'	77
Sediments terrigenes'	9
Spahlerite	3
Grand Total	98

The data provided has such high variance that 78% of the rocks belong only a single category. The plot for the given data is shown below:

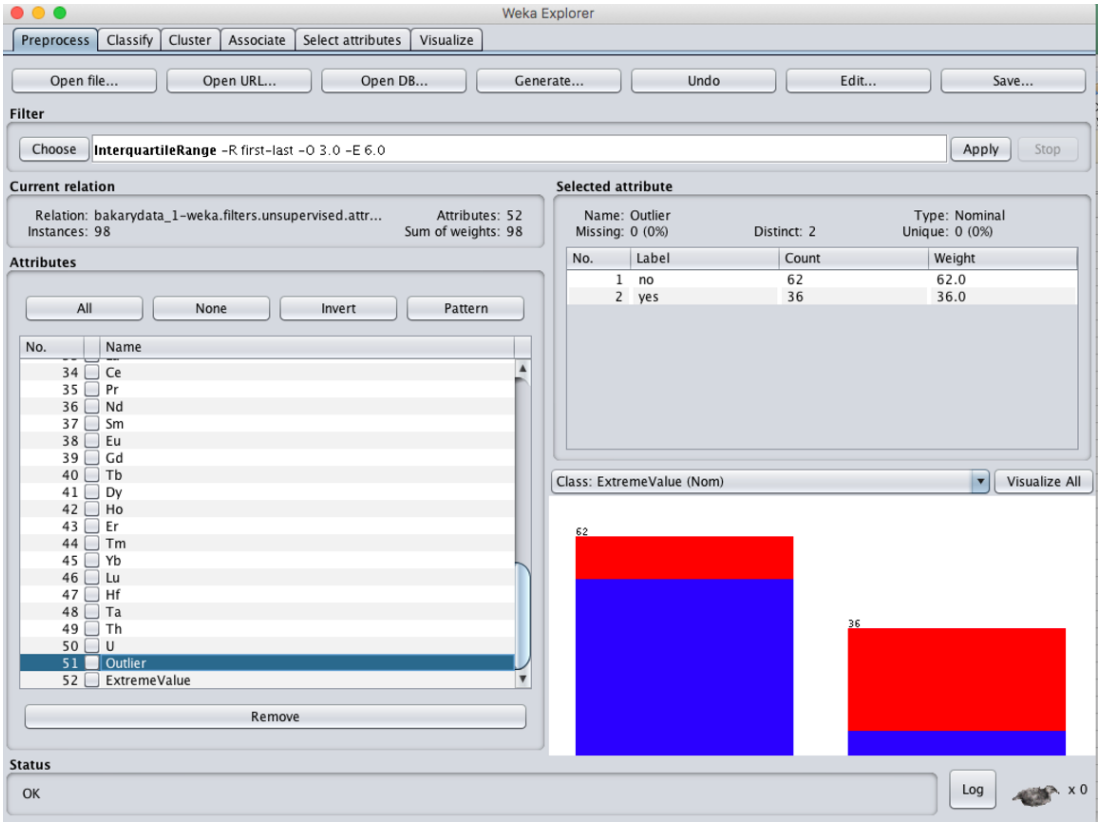


4. DATA PREPARATION

The data given initially had a lot of missing values, duplicate values and values with incorrect formatting. All these inaccuracies are removed while preparing the data.

4.1 Finding Outliers

We first identified the values which were acting as outliers in the given dataset. This can be performed using Weka method "InterQuartile Range"[1] to find the Outlier and Extreme values in a given dataset.



4.2 Cleaning Misformatted Data Values

Some values are not in a proper format as required. Weka can train the model only based on values which are either numerical or nominal. *E.g. Li- Sédiments terrigènes is given a value <0.3 which is replaced with 0.25.*

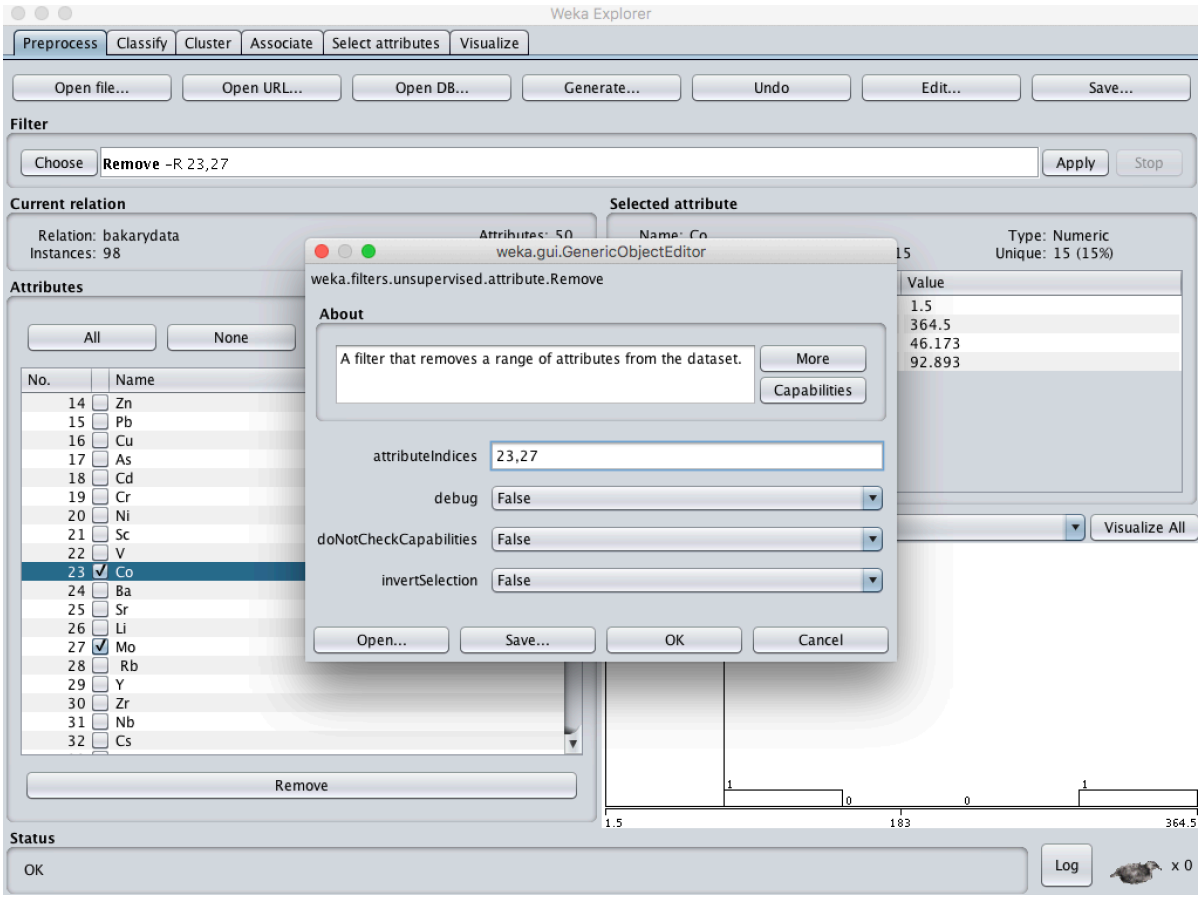
4.3 Solving Class Label Discrepancies

There were a few values in which the class labels were misspelled. This caused the model to consider this misspelled value as another class label. These values were corrected to obtain a uniform class label *E.g. pyyrite and pyrite*

There were a few values which belong to the same class. These were combined to obtain a uniform label for both. *E.g. carbonates and impures belong to a single group.*

4.4 Dropping Irrelevant Attributes

There were a few attributes which had a lot of missing values. Imputing these values would have caused an incorrect prediction. Its best to drop these attributes since they provide none or sometimes negative impact on training[2]. *E.g. MO and CO have less than 20 values*

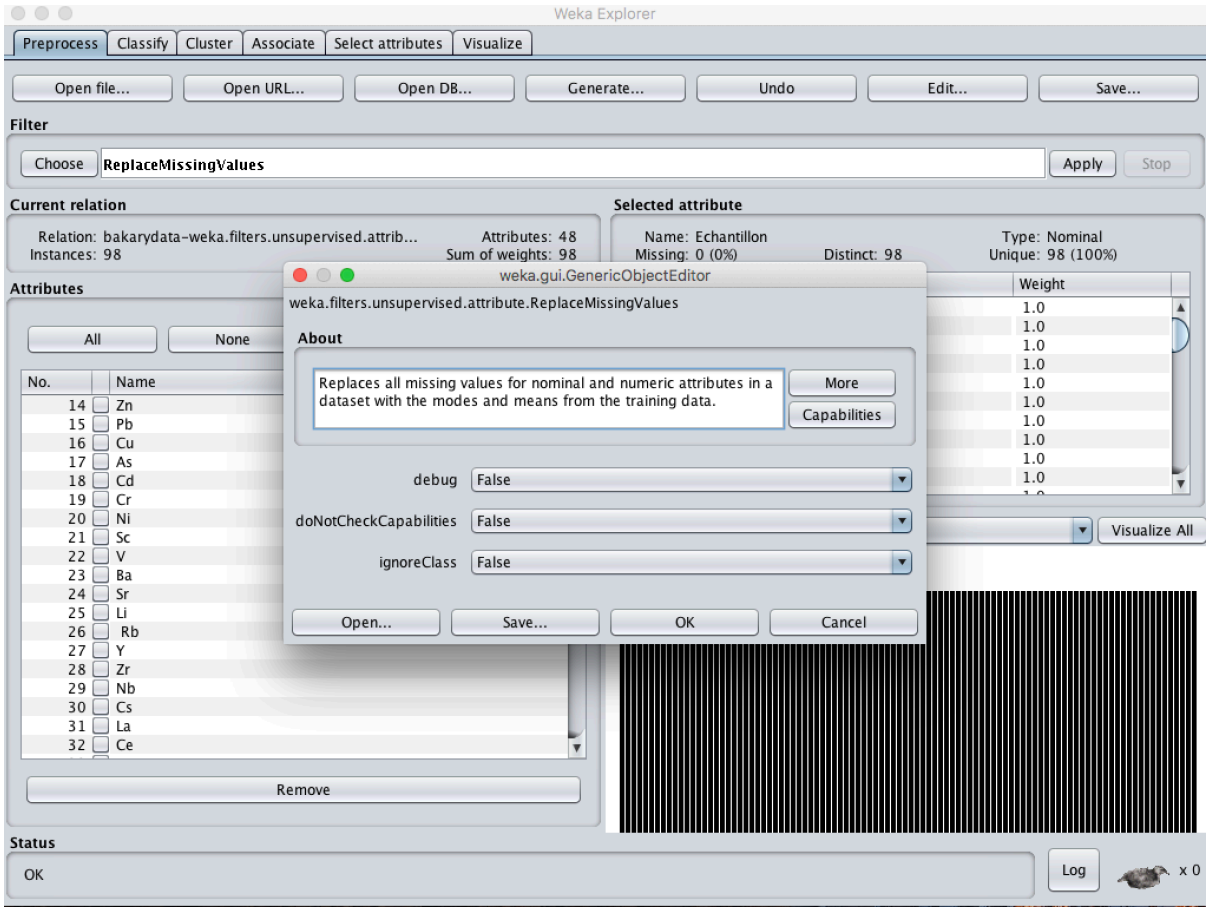


4.5 Dropping Duplicate and Irrelevant Attributes

There are multiple rows in the given dataset which are repeating or they have no values at all. These all rows are just dropped so as to create a rigid structure where all rows and attributes are filled. E.g. There is a complete empty row and a duplicate header.

4.6 Filling Missing Values

Finally, after all the cleaning is done, the missing values need to be filled so that the model can be trained. There are multiple ways in which these values can be filled. One method is to fill 0 in the missing value. However, most of the times it provides inaccurate results as 0 is itself a value used by some of the other cells. Hence it is best approach to fill it with either mean, median or mode. We have filled these missing values with the mean of that attribute.

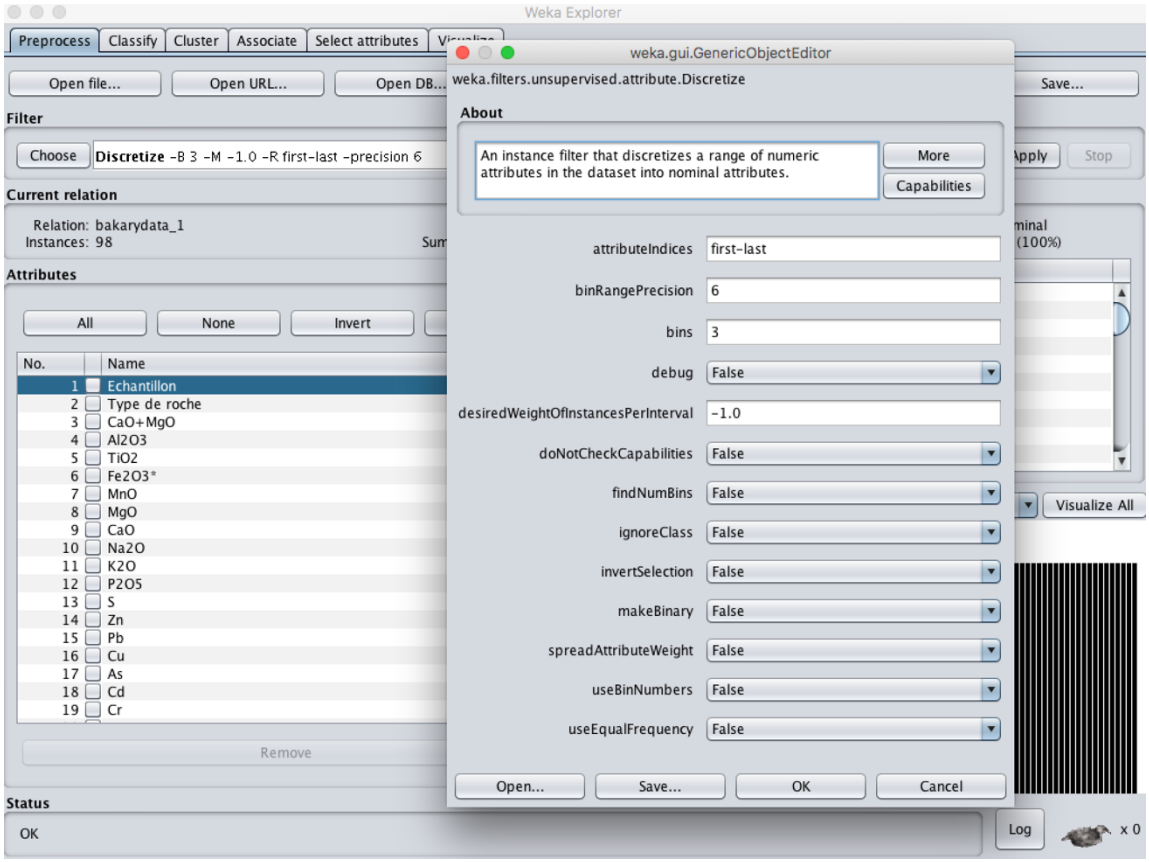


5. DATA PREPROCESSING

Here, we use our prepared data and process it using data discretization. The dataset after cleaning is divided into two data sets namely PD1 and PD2. The discretization is performed using the method of binning with equal width and equal frequency respectively. Also the cleaned project data is normalized to get the dataset PD.

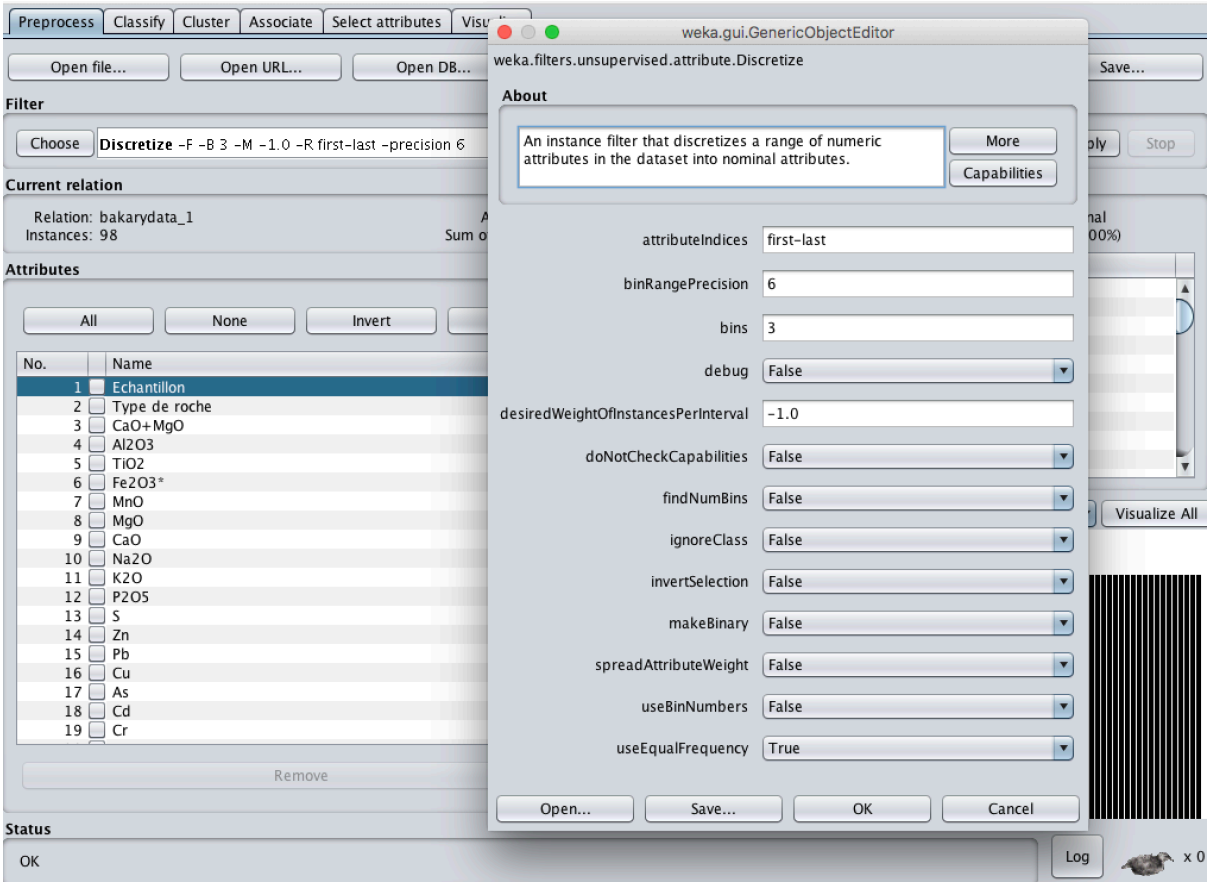
5.1 Discretization by Equal Width Binning (PD1)

Equal width binning divides the range into N intervals/ bins of the same size. We use this dataset for the Decision Trees Descriptive Classifier. The following setting were used in Weka for Discretization by Equal Width Binning.



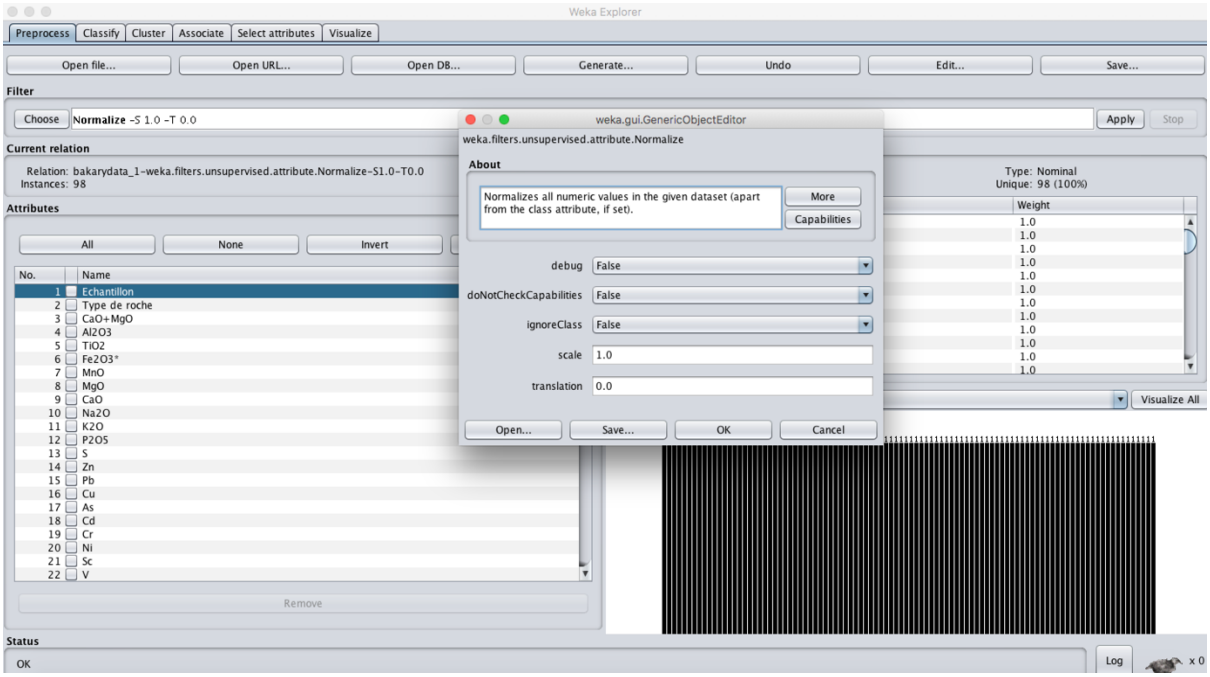
5.2 Discretization by Equal Frequency Binning (PD2)

This is also known as Equal Depth binning. It divides the range into N intervals/ bins, each containing nearly the same number of samples**.** We use this dataset for the Decision Trees Descriptive Classifier. The following setting were used in Weka for Discretization by Equal Frequency Binning.



5.3 Normalization (PD)

Here we adjust the values measured on different scales to a notionally common scale. We use this dataset for the Neural Network Non-Descriptive Classifier. The following setting were used in Weka for Normalizing.



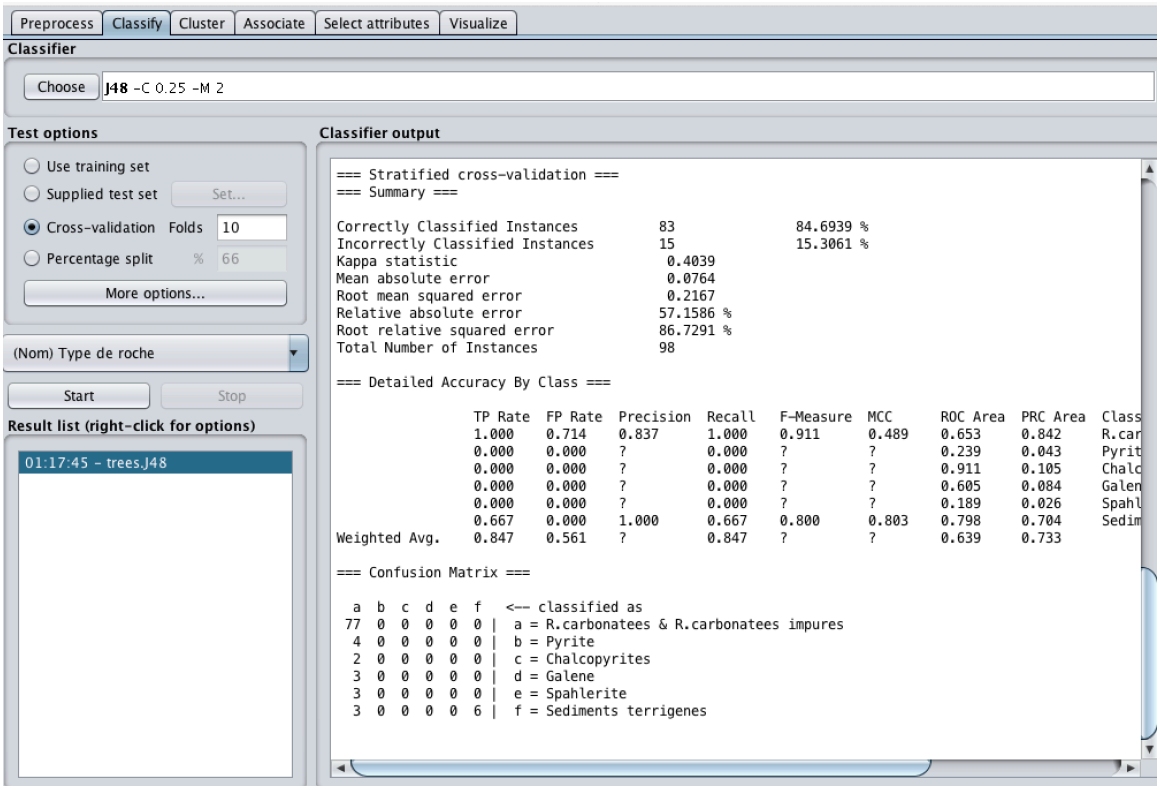
6. EXPERIMENTS

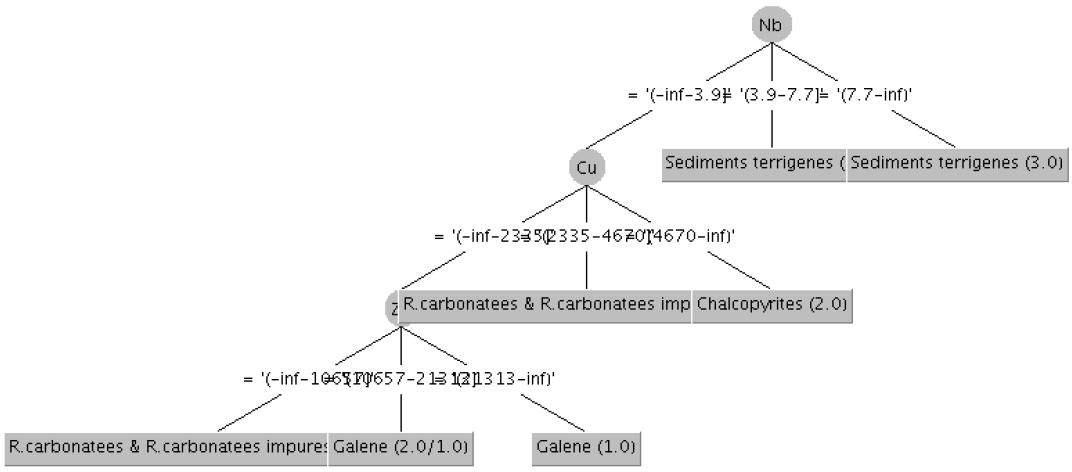
The following 3 experiments were conducted as part of this project.

6.1. EXPERIMENT 1 (Non-Contrast Learning):

All the records were used to perform the full classification (learning), i.e. built a classifier for all classes C1- C6 simultaneously. This experiment was repeated for the 3 datasets PD1, PD2 and PD.

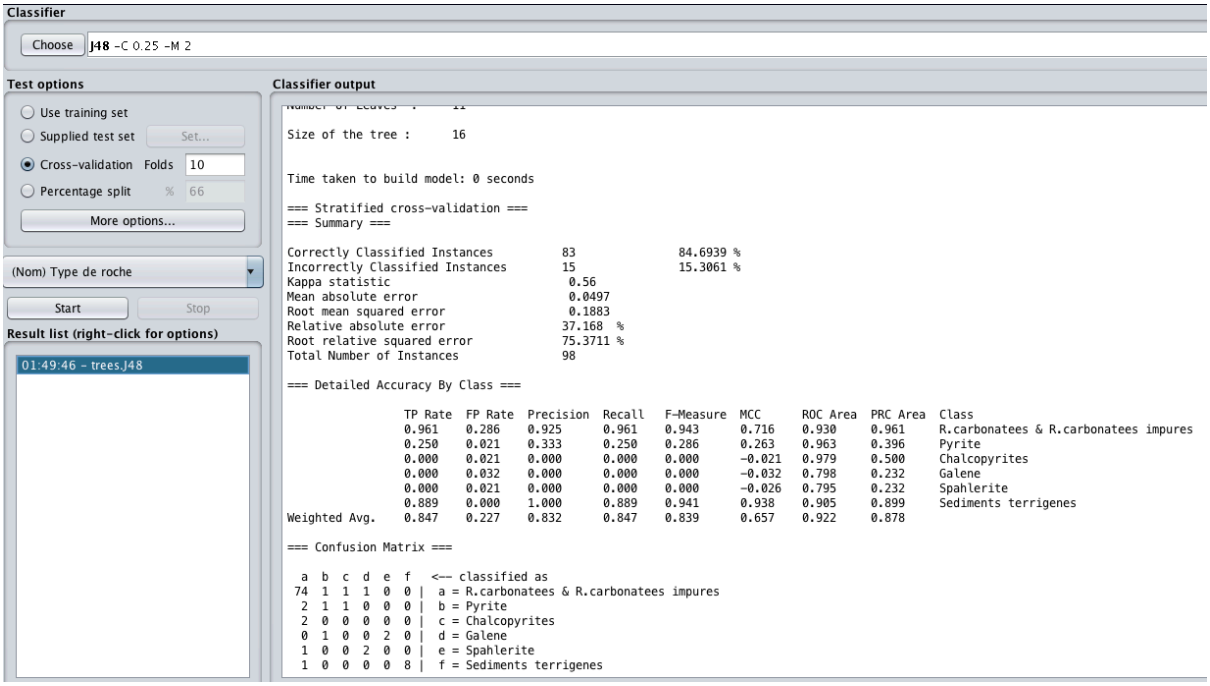
a) With PD1- Equal Width Binning:

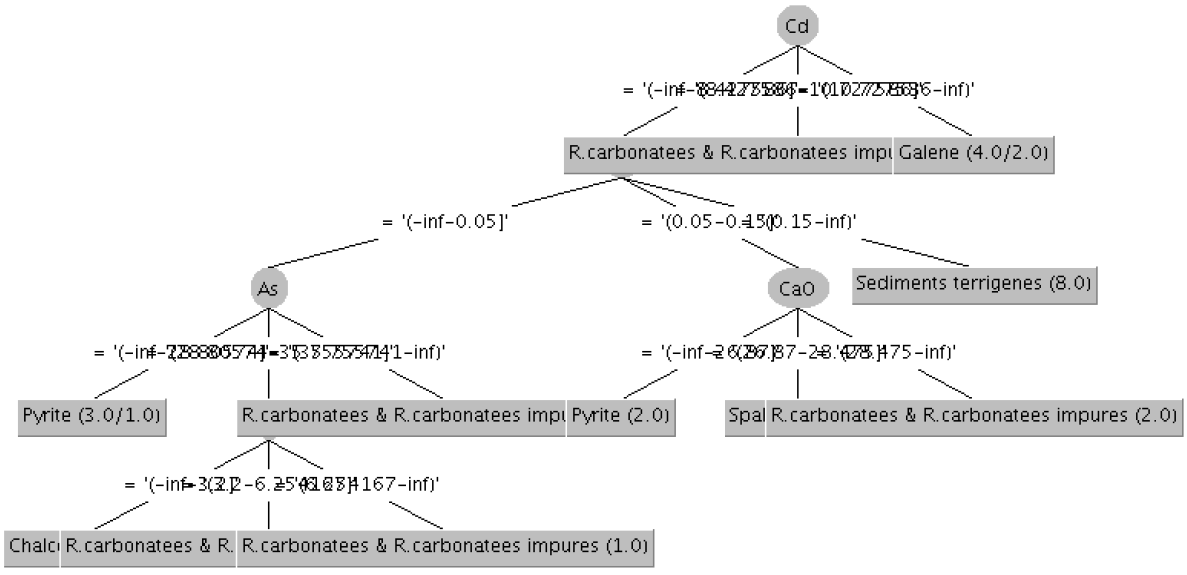




ACCURACY: 84.69%

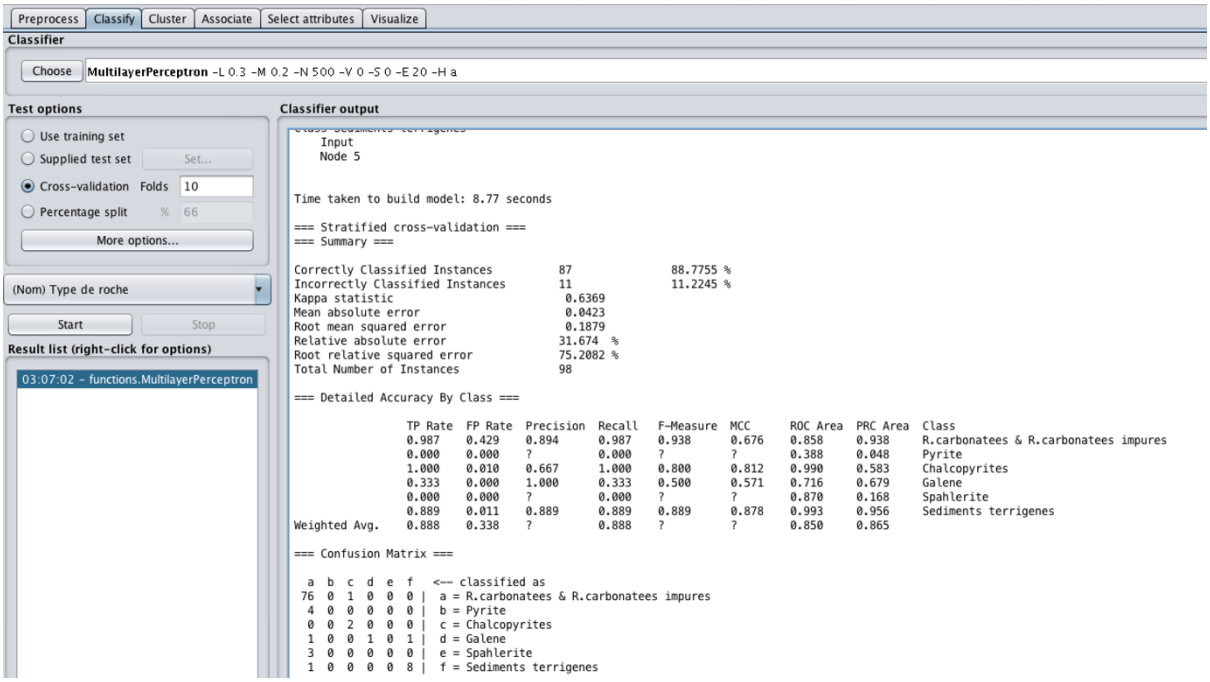
b) With PD2 - Equal Frequency Binning:





ACCURACY: 84.69%

c) With PD - Normalized:

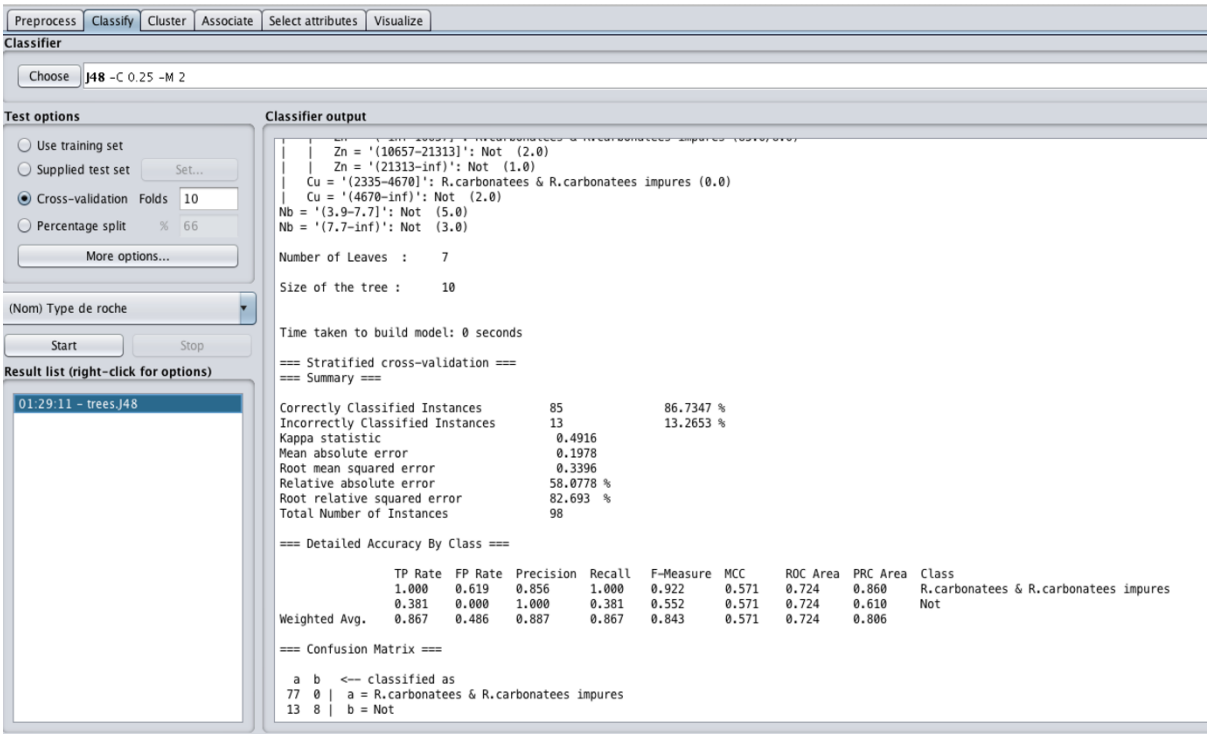


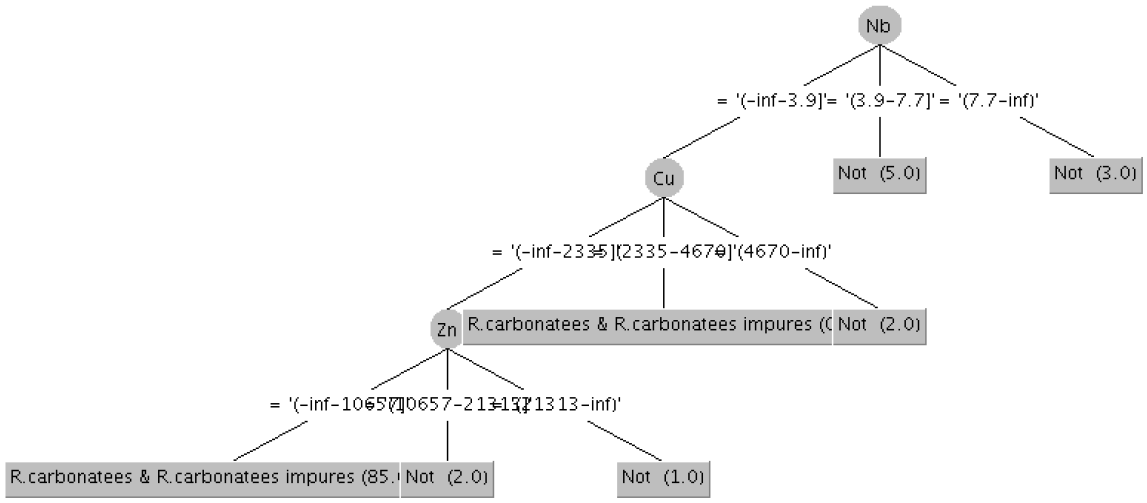
ACCURACY: 88.77%

6.2. EXPERIMENT 2 (Contrast Learning)

All the records were used to perform the contrast classification i.e. contrasting class **C1**(R. carbonatees and R. carbonatees impures) with a class **not C1** that contains other classes. This experiment was repeated for the 3 datasets PD1, PD2 and PD.

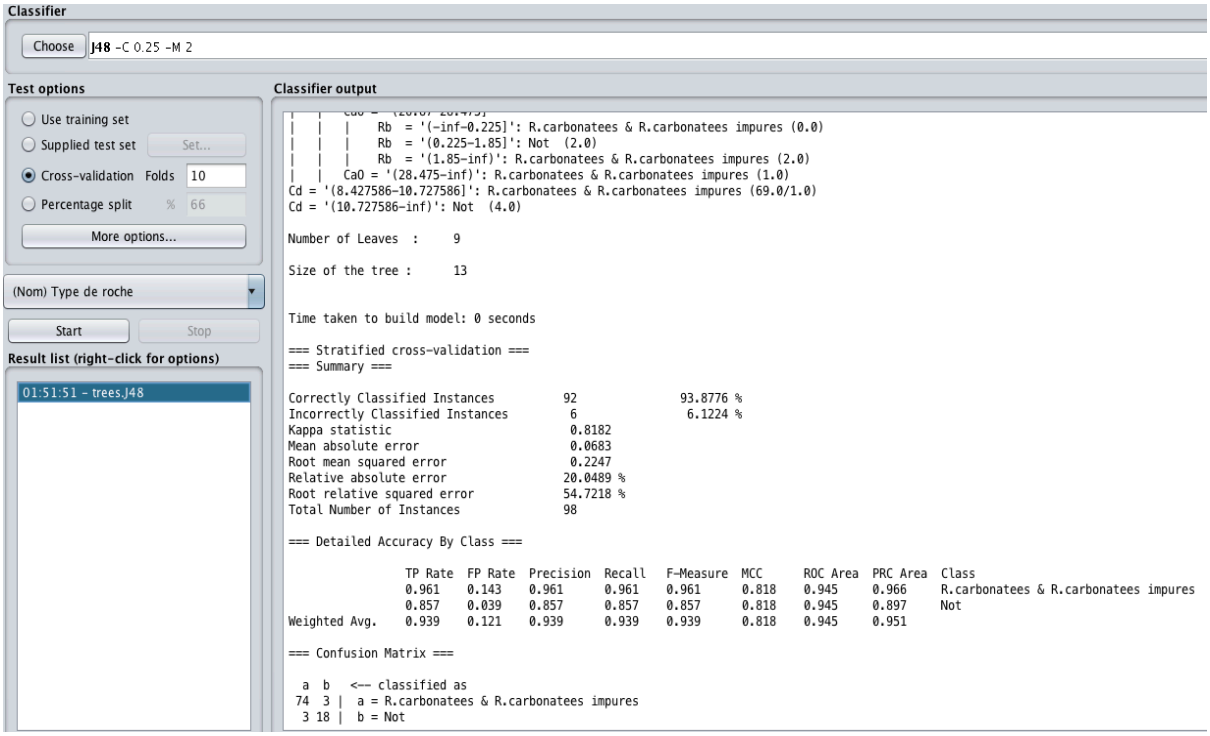
a) With PD1- Equal Width Binning:

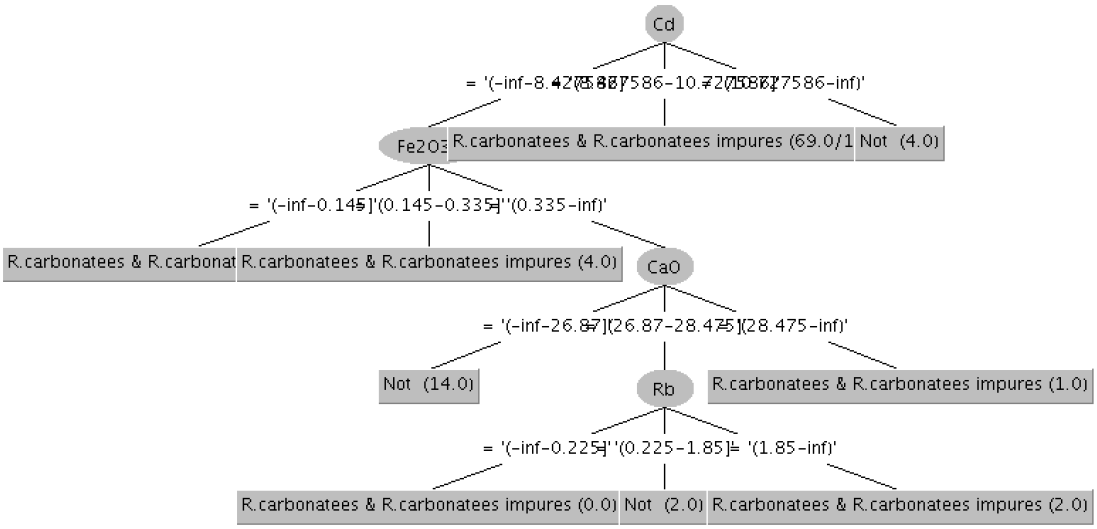




ACCURACY: 86.73%

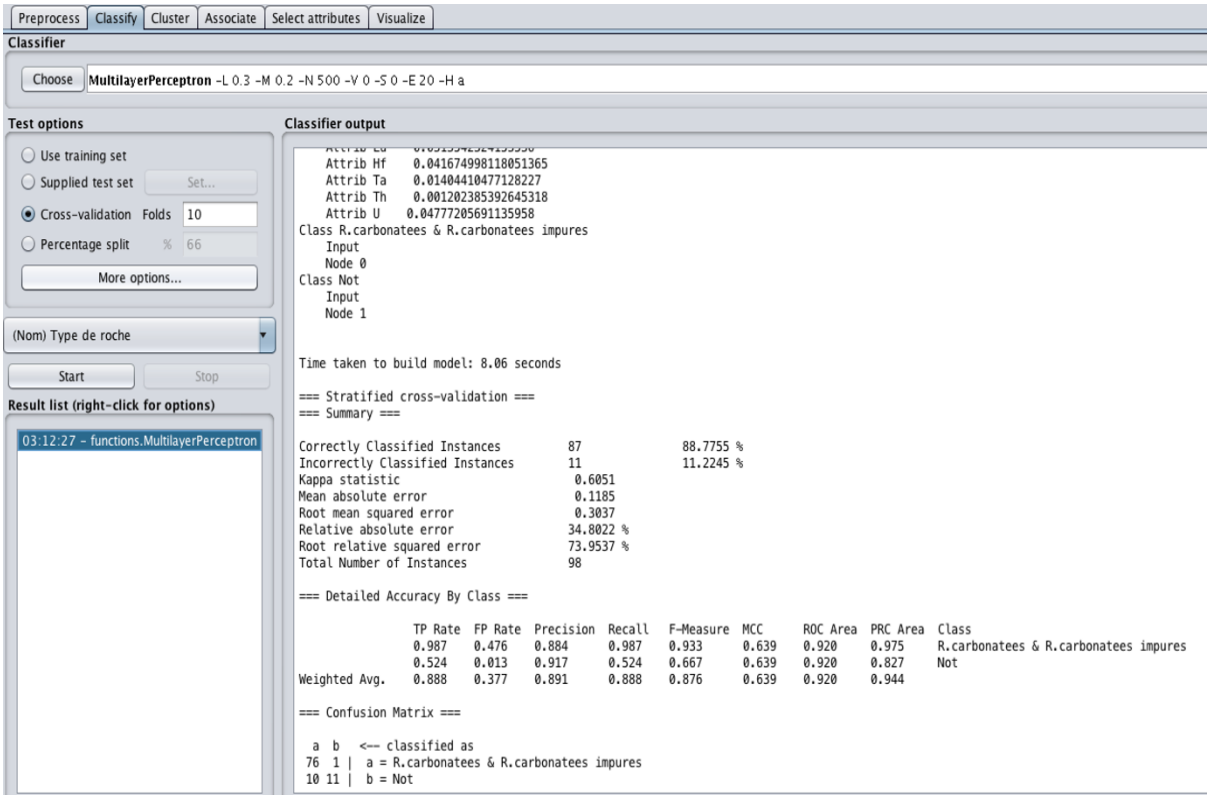
b) With PD2 - Equal Frequency Binning:





ACCURACY: 93.87%

c) With PD - Normalized:



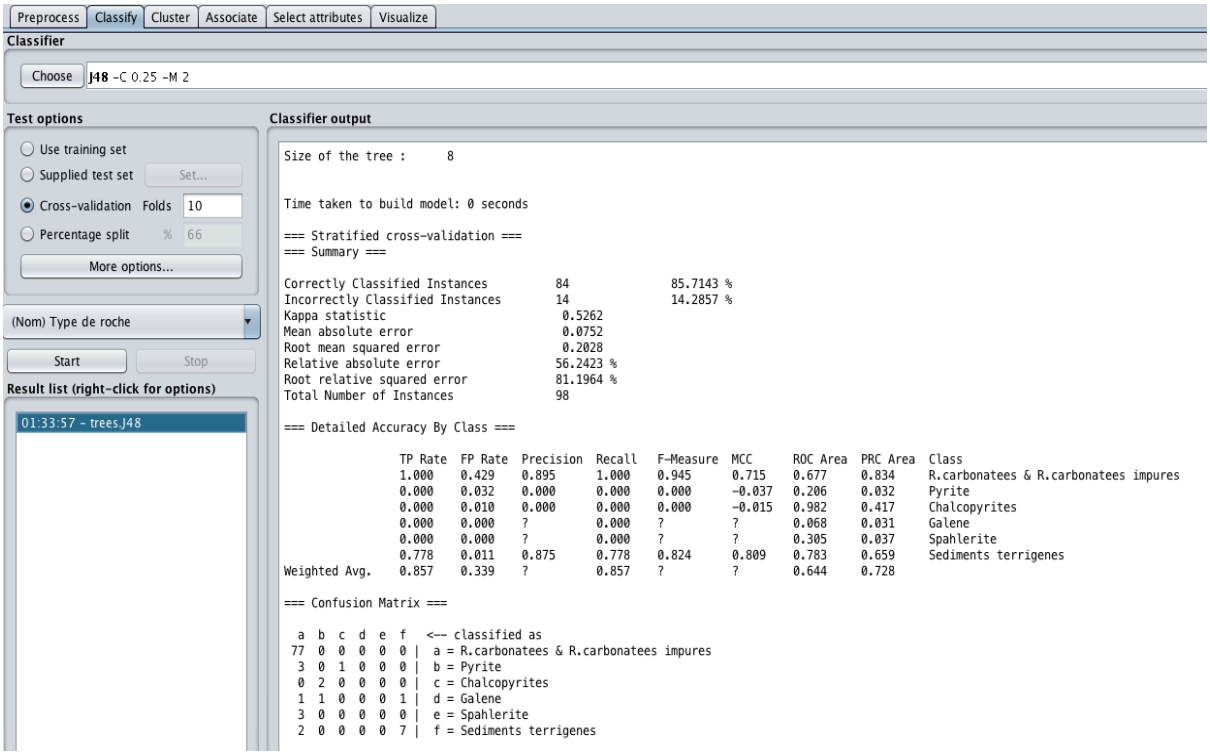
ACCURACY: 88.77%

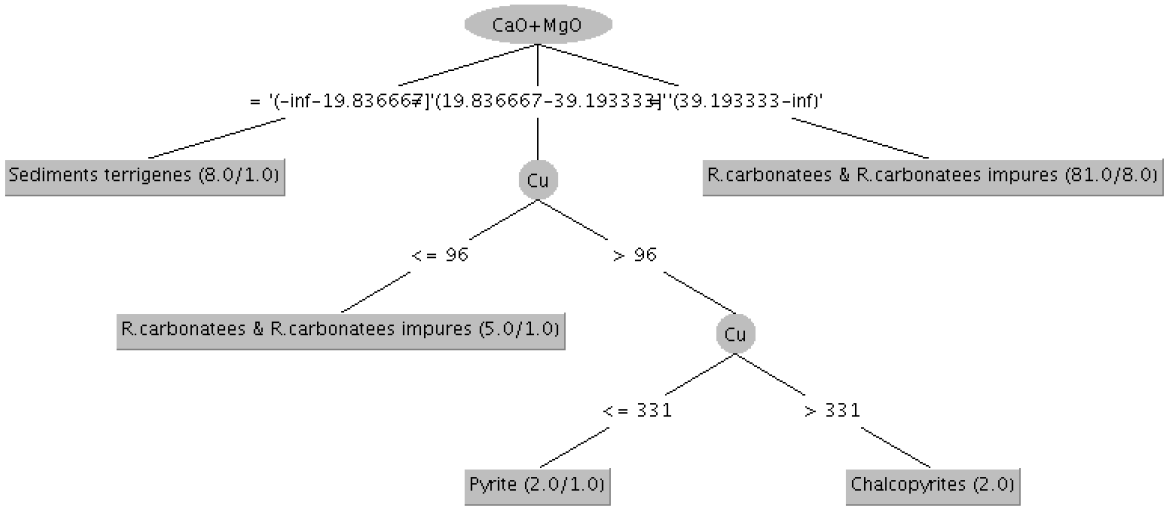
6.3. EXPERIMENT 3

Experiments 1, 2 (both contrast and non-contrast learning) were repeated for all records with the most important attributes as defined by the expert (S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3). This experiment was also repeated for the 3 datasets PD1, PD2 and PD.

a) With PD1- Equal Width Binning:

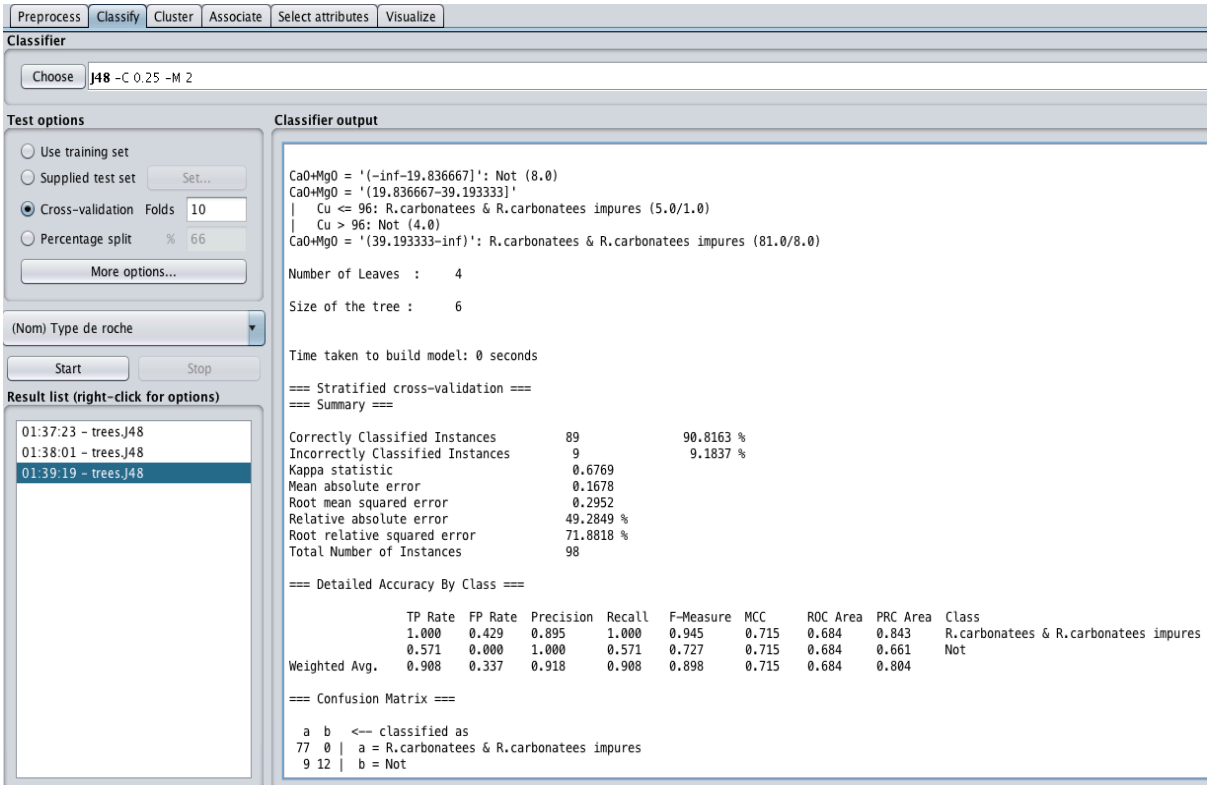
- **Repeating experiment 1:**

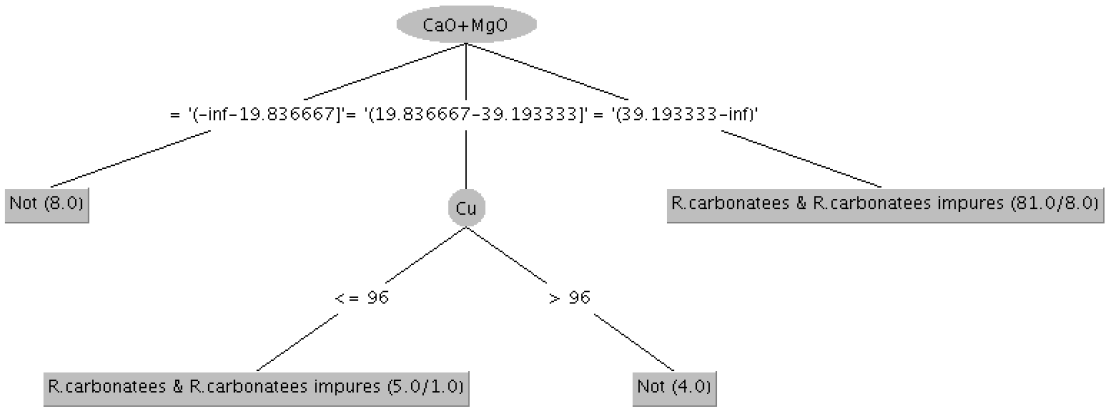




ACCURACY: 85.71%

- **Repeating experiment 2:**

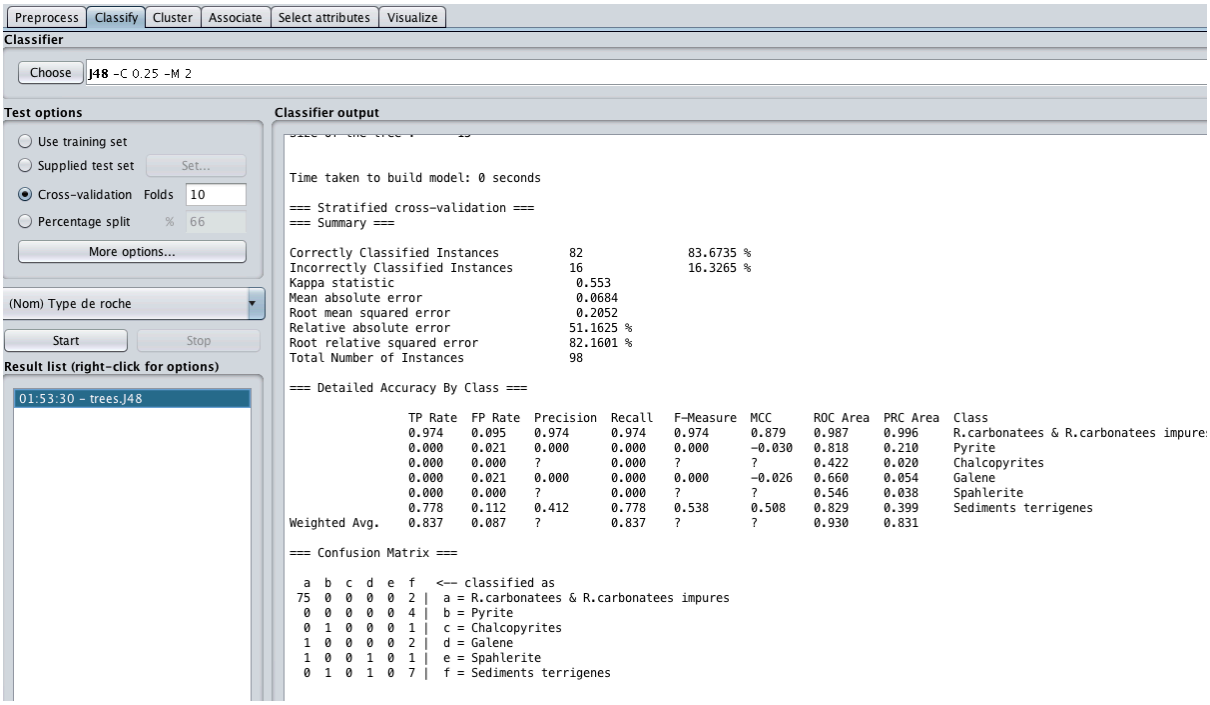


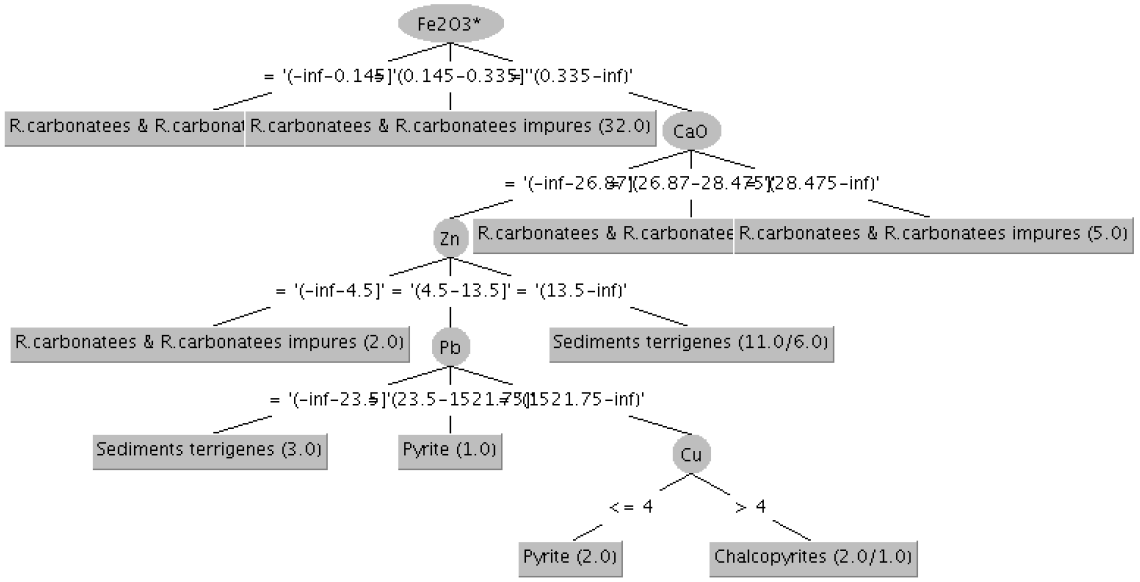


ACCURACY: 90.81%

b) With PD2 - Equal Frequency Binning:

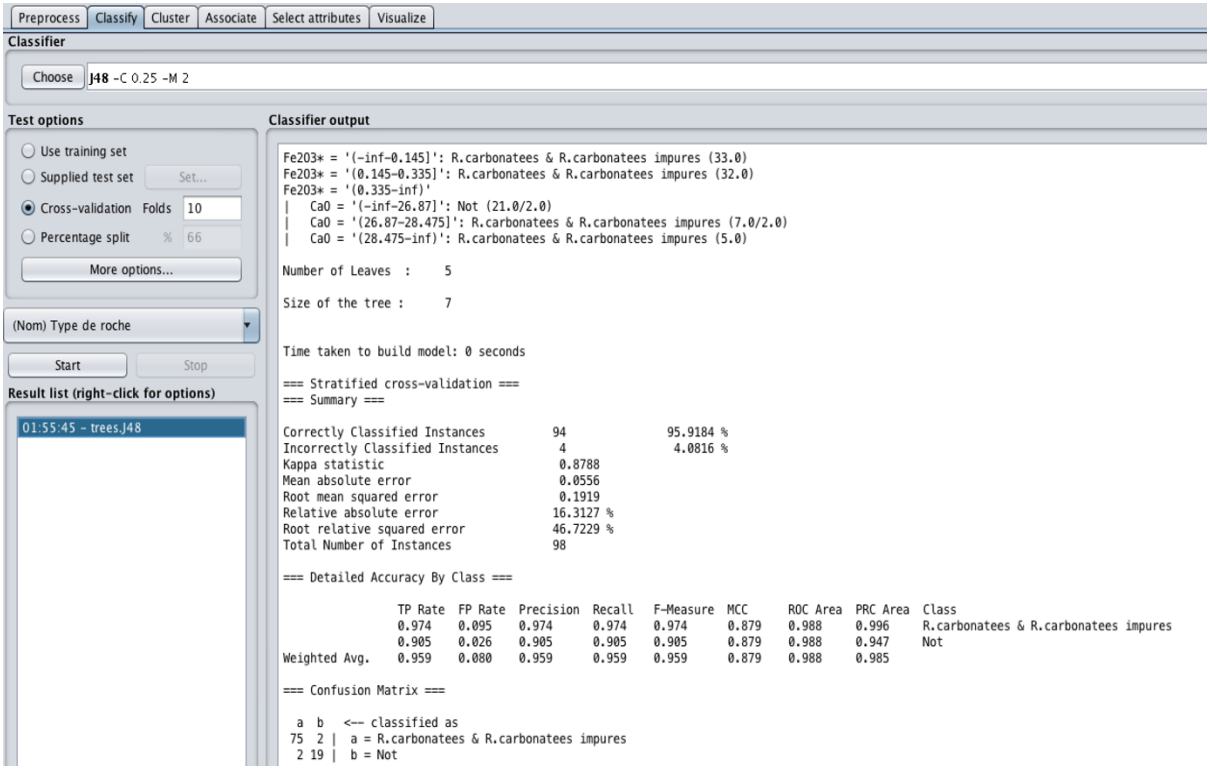
- **Repeating experiment 1:**

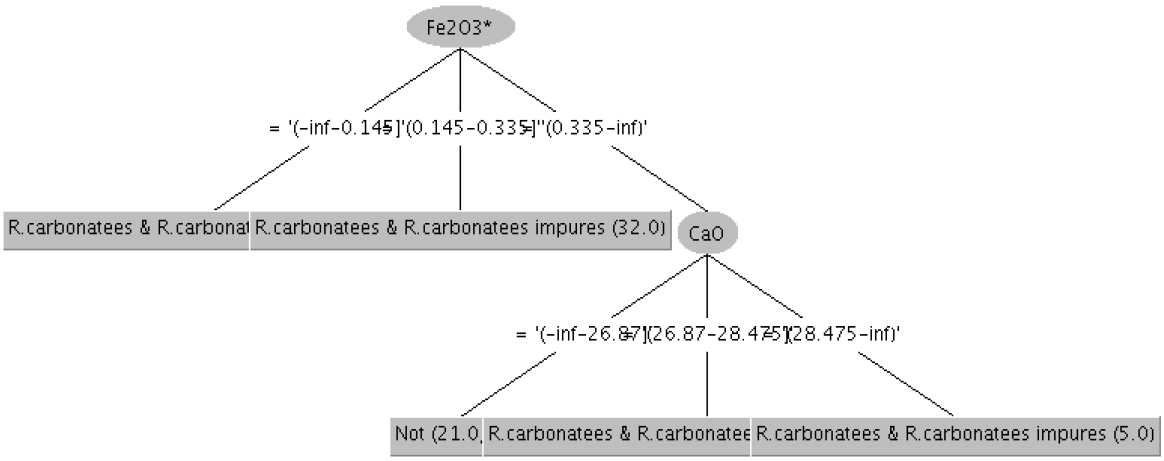




ACCURACY: 83.67%

- **Repeating experiment 2:**

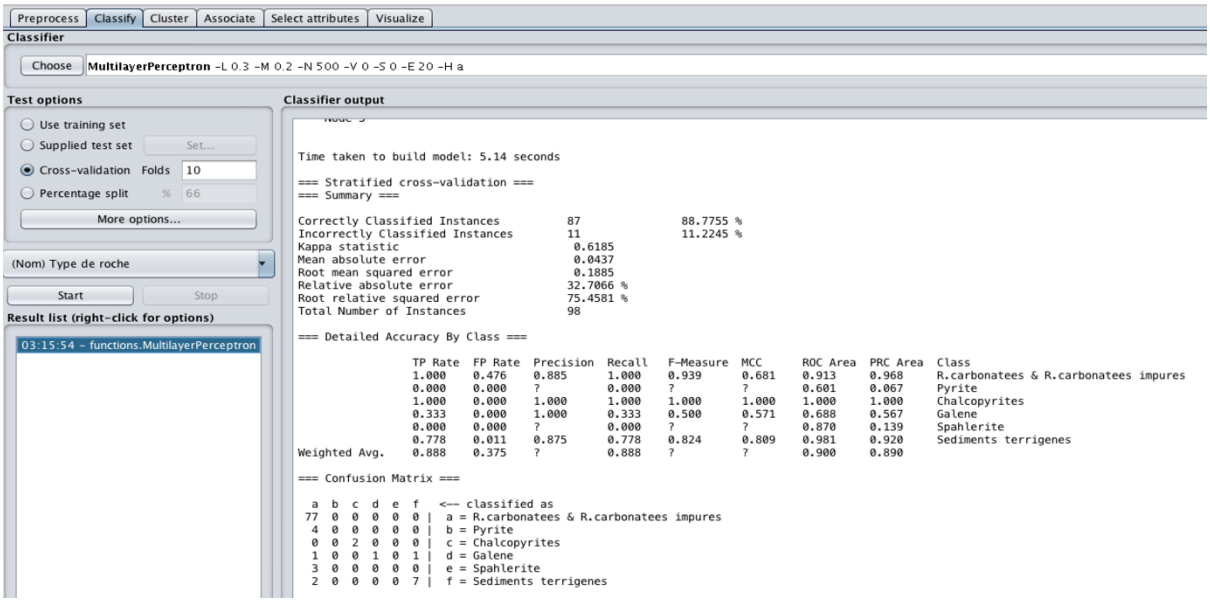




ACCURACY: 95.91%

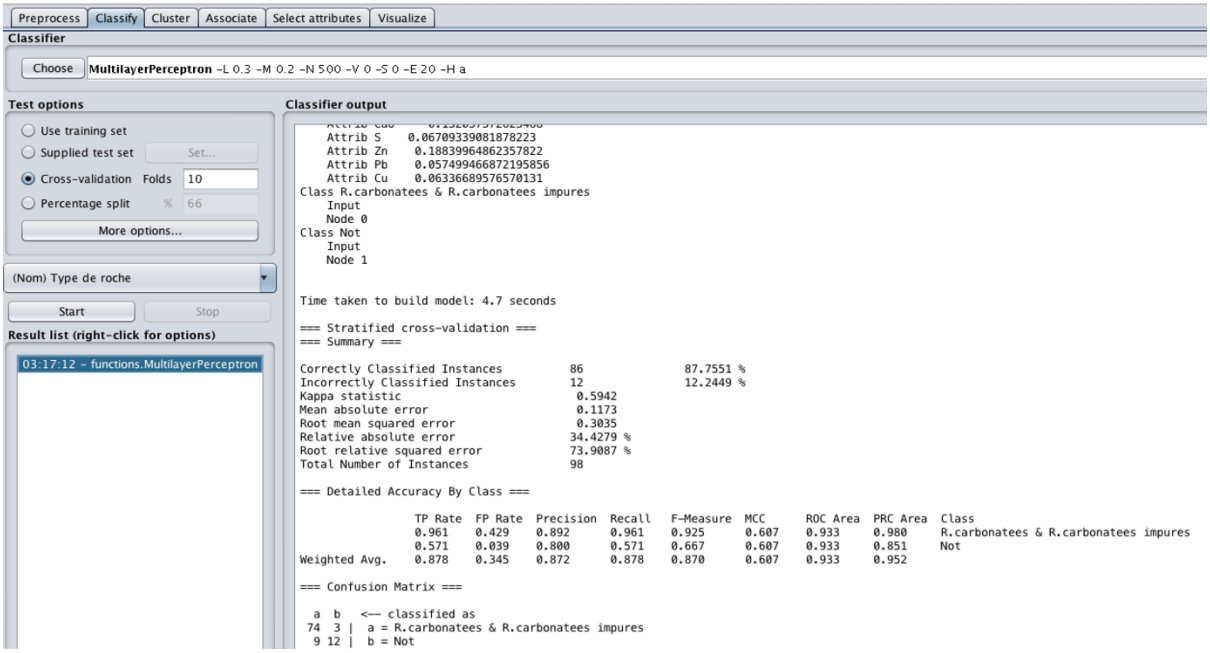
c) With PD - Normalized:

- **Repeating experiment 1:**



ACCURACY: 88.77%

- **Repeating experiment 2:**



ACCURACY: 87.75%

7. EVALUATING THE MODELS

A number of machine learning techniques were applied to the given dataset after cleaning. Each of these methods were also applied on different versions of the same dataset:

PD data represents the original normalized data.

PD1 dataset represents Equal Width Binning Discretized data.

PD2 represents the Equal Frequency Binning Discretized data.

All the models that are implemented are evaluated on the basis of the accuracy as given in the table below.

Experiment →	1	2	3	
Models ↓	Non Contrast Learning (All Attributes)	Contrast Learning (All Attributes)	Non Contrast Learning (Expert Selected Attributes)	Contrast Learning (Expert Selected Attributes)
Decision Tree (PD1)	84.69%	86.73%	85.71%	90.81%
Decision Tree (PD2)	84.69%	93.87%	83.67%	95.91%
Neural Network (PD)	88.77%	88.77%	88.77%	87.75%

8. SUMMARY

The aim of the project was to classify the given data on the basis of TYPE DE ROCHE (Rock Type). We used WEKA (Internet based Classification Tool) to build two types of classifiers: Descriptive and Non-Descriptive. For the descriptive classifier, we used Decision tree to generate sets of discriminant rules describing the content of the data. We used Neural Networks to build the Non-Descriptive Classifier.

The Data Preparation step included attributes selection, cleaning the data, filling the missing values, error correction etc. to build the Project Data. The Data Preprocessing step included 2 methods of Data Discretization Namely Equal Width Binning and Equal Frequency Binning to obtain the 2 datasets PD1 and PD2 respectively. Also, the project dataset was normalized to obtain the dataset PD. For each of these datasets 3 experiments were carried out in WEKA. The predictive accuracy was computed for the different classifiers built.

9. REFERENCES

1] Outlier Detection in Weka : https://www.youtube.com/watch?v=WrjpO7CmUoQ

2] Deleting an Attribute: https://sourceforge.net/projects/weka/files/documentation/

3] Data Cleaning using Weka : https://machinelearningmastery.com/how-to-handle-missing-values-in-machine-learning-data-with-weka/

4] Data Discretization : Data Discretization: https://www.youtube.com/watch?v=P--UFzlNGeA&t=109s

5] Equal Frequency Binning: https://www.youtube.com/watch?v=aDMzPC5IO4c