CSE 634 – Data Mining

TYPE DE ROCHE (Types of Rock)

DATA CLASSIFICATION

-Weka 3.8.2

Team Members:

Akhil Bhutani	*110898687*
Chaitanya Kalantri	*111446728*
Sourav Mishra	*110946489*
Selina Kaur	*110936206*
Shreyas Bhatia	*111432576*

Brief Introduction of Weka and Requirement

Weka is a collection of machine learning algorithms for data mining tasks. We are provided with a dataset for Analysis. We can use this dataset in Weka to apply machine learning algorithms and gather heuristics on the dataset. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

The project aims to use Weka to build two types of Classifiers – Descriptive and Non-Descriptive.

Descriptive Classifier: We use Weka to build a decision tree in the following way. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes. A decision has two or more branches and leaf nodes represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. For descriptive classifier, we use two methods of Data Discretization. In some cases, the columns may contain so many values that that the algorithm cannot easily identify interesting patterns in the data from which to create a model. Hence, we need to perform Discretization. Discretization is the process of putting values into buckets so that there are a limited number of possible states. The buckets themselves are treated as ordered and discrete values. One can discretize both numeric and string columns depending on requirement of the project.

Non-Descriptive Classifier: We use Weka to create a Neural Networks to classify the data set provided to us. Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record. The errors from the initial classification of the first record is fed back into the network and used to modify the networks algorithm for further iterations. We use MultiLayer Perceptrons algorithm for neural networks. A multilayer perceptron is a class of feedforward artificial neural network. Multilayer Perceptron consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. Multilayer Perceptron utilizes a supervised learning technique called backpropagation for training. The multiple layers and non-linear activation distinguish Multilayer Perceptron from a linear perceptron. It can distinguish data that is not linearly separable.

Data Cleaning

Data cleansing is the process of correcting and detecting inaccurate records from the data and performing step to correct, replace or modify the data. Several algorithms are present to detect the inaccurate data. Data Cleansing is also called Data Scrubbing.

Data cleansing can be a challenging task. There are several instances where one has to make important decision about the data. There are several techniques that can be applied. If there is a missing value, we can replace it by mean or median. If we have an attribute which is sparse, then we can decide to drop the record if we cannot interpolate its values. There could be outlier in the data and we might have to replace such values to build a better classifier.

There could be several challenges to data cleaning. The most challenging problem is indeed to handle duplicates and invalid entry in data. There could not be sufficient information for the necessary transformation or corrections. The primary solution could be to delete such data which apparently leads to loss of information. This can sometimes be costly. Also, data cleansing can be time consuming and a costly step.

Using filters present in Weka, we have applied the filters as below -

1. Since we have only 6 classes. Hence, we merged R. Carbonatees impures and R. Carbonatees as one class.
2. Changed the spelling of 'Pyyrites' to 'pyrate'; 'Chalcopyrites' to 'charcopyte'; 'GalSne' to 'Galene'; 'S,diments terrogsness' to Sediments terrigenes as a part of data cleaning.
3. Removed the second row in the dataset which is empty.
4. Remove the columns which has less than 30% of the data
    a. As it won't be relevant to predict values of a column which has 70%+ data missing.
    b. Using the 'edit' option, we were able to remove the column and the
5. For other columns, replace the missing values with the mean or median.
    a. To do the perform the below tasks, go to 'Chose'-> 'Filter' -> 'ReplaceMissingValues'
    b. Thereafter replace with the respective value as per the requirement.
    c. For 'Pb' column, fill the missing values with median values.
    d. For 'Ni', 'Sc', 'Li' column, fill the missing values with mean values.
    e. For all the other missing values, replace the missing values with the mean value of that respective column.

Experiment 1:

In Experiment 1 we used all records to perform the full classification (learning), i.e. build a classifier for all classes C1- C6 simultaneously. We will perform this experiment by using a J48 Decision Tree for a Descriptive Classifier and a Multi-Layer Perceptron for a Non-Descriptive Classifier.

Decision Tree with Discretization:

We used all records to perform the full classification (learning), i.e. build a classifier for all classes C1- C6 simultaneously.

Equal Width Binning

We used a technique of equal binning with the tool. Here's the run information for the same. We optimized the bin size by setting findBinNums to True to get the optimum bin value.

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Updated_Bakarydata
Instances: 98
Attributes: 46

Echantillon	Type de roche	S	Zn	Pb	CaO+MgO	Al2O3	TiO2	Fe2O3*	Cu
MnO	MgO	Cr	CaO	Na2O	K2O	P2O5	Ni	Sc	V
Ba	Sr	Li	Rb	Y	Zr	Nb	Cs	La	Ce
Pr	Nd	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm
Yb	Lu	Hf	Ta	Th	U				

Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree

```
K2O <= 0.8
|   Zn <= 700.073684
|   |   Tb <= 0.8
|   |   |   Fe2O3* <= 0.76: R. Carbonatees AND R. Carbonatees impures (74.0)
|   |   |   Fe2O3* > 0.76
|   |   |   |   CaO <= 26.86: Pyrite (4.0)
|   |   |   |   CaO > 26.86: R. Carbonatees AND R. Carbonatees impures (3.0)
|   |   Tb > 0.8: Charcopyrite (2.0)
|   Zn > 700.073684
|   |   Pb <= 5277: Spahlerite (3.0)
|   |   Pb > 5277: Galene (3.0)
K2O > 0.8: Sediments terrigenes (9.0)
```

Number of Leaves : 7

Size of the tree : 13

Time taken to build model: 0.04 seconds

Stratified cross-validation

Summary

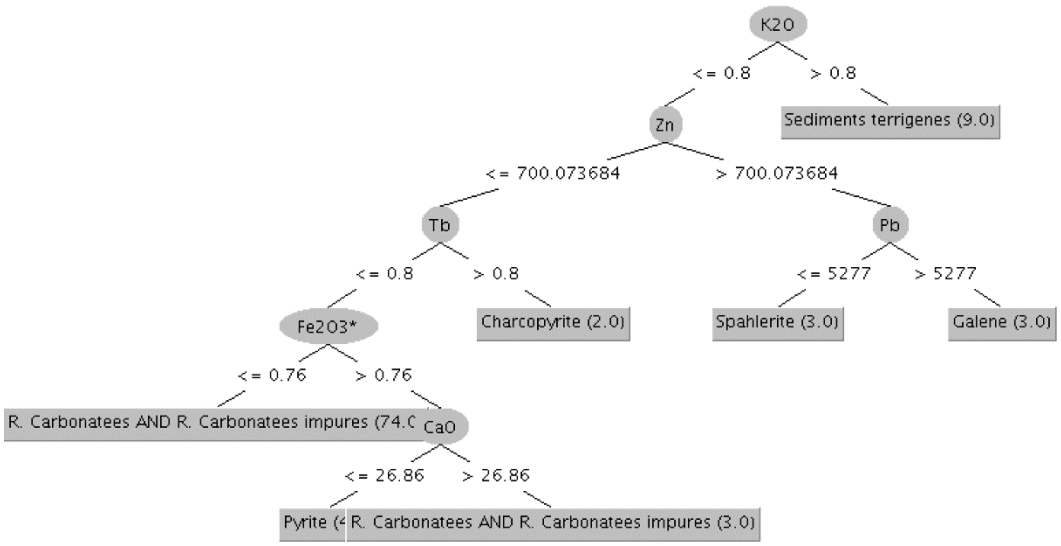
Correctly Classified Instances	90 (91.8367 %)
Incorrectly Classified Instances	8 (8.1633 %)
Kappa statistic	0.7834
Mean absolute error	0.0258
Root mean squared error	0.1525
Relative absolute error	19.2718 %
Root relative squared error	61.0517 %
Total Number of Instances	98

Detailed Accuracy By Class: -

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.974	0.048	0.987	0.974	0.980	0.911	0.961	0.980	R. Carbonatees and R. Carbonatees impures
	0.750	0.032	0.500	0.750	0.600	0.593	0.855	0.573	Pyrite
	0.000	0.000	?	0.000	?	?	1.000	1.000	Charcopyrite
	0.667	0.011	0.667	0.667	0.667	0.656	0.828	0.455	Galene
	0.667	0.011	0.667	0.667	0.667	0.656	0.828	0.455	Spahlerite
Weighted Average	0.889	0.022	0.800	0.889	0.842	0.827	0.931	0.688	Sediments terrigenes
	0.918	0.041	?	0.918	?	?	0.947	0.905	

Confusion Matrix

```
 a  b  c  d  e  f   <-- classified as
75  1  0  0  0  1 |  a = R. Carbonatees AND R. Carbonatees impures
 0  3  0  0  0  1 |  b = Pyrite
 0  2  0  0  0  0 |  c = Charcopyrite
 0  0  0  2  1  0 |  d = Galene
 0  0  0  1  2  0 |  e = Spahlerite
 1  0  0  0  0  8 |  f = Sediments terrigenes
```



Equal Frequency Binning

We used a technique of equal frequency binning with the tool. Here's the run information for the same. Equal Frequency binning does not support findBinNums setting to find optimum number of bins.

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Updated_Bakarydata-weka.filters.supervised.attribute.Discretize-Rfirst-last-precision6
Instances: 98
Attributes: 46

Echantillon	Type de roche	S	Zn	Pb	CaO+MgO	s	TiO2	Fe2O3*	Cu
MnO	MgO	Cr	CaO	Na2O	K2O	P2O5	Ni	Sc	V
Ba	Sr	Li	Rb	Y	Zr	Nb	Cs	La	Ce
Pr	Nd	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm
Yb	Lu	Hf	Ta	Th	U				

Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree

```
K2O = '(-inf-0.82]'
|   Pb = '(-inf-2695]'
|   |   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)': Pyrite (4.0)
|   |   Sc = '(2.45-inf)': Charcopyrite (3.0/1.0)
|   Pb = '(2695-5694.5]': 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

Stratified cross-validation

Summary: -

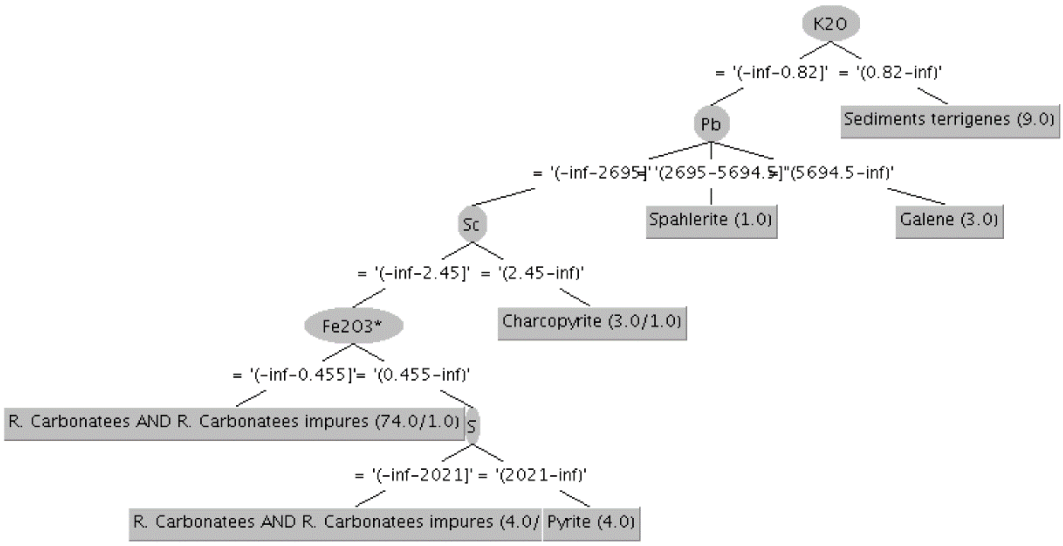
Correctly Classified Instances	91(92.8571 %)
Incorrectly Classified Instances	7 (7.1429 %)
Kappa statistic	0.7937
Mean absolute error	0.0305
Root mean squared error	0.1471
Relative absolute error	22.7938 %
Root relative squared error	58.8737 %
Total Number of Instances	98

Detailed Accuracy By Class: -

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.987	0.190	0.950	0.987	0.968	0.844	0.902	0.935	R. Carbonatees and R. Carbonatees impures
	0.750	0.021	0.600	0.750	0.667	0.655	0.977	0.721	Pyrite
	0.000	0.000	?	0.000	?	?	0.990	0.667	Charcopyrite
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Galene
	0.000	0.000	?	0.000	?	?	0.333	0.028	Spahlerite
Weighted Average	1.000	0.011	0.900	1.000	0.947	0.943	0.994	0.900	Sediments terrigenes
	0.929	0.152	?	0.929	?	?	0.901	0.891	

Confusion Matrix

```
 a  b  c  d  e  f   <-- classified as
76  0  0  0  0  1 |  a = R. Carbonatees AND R. Carbonatees impures
 1  3  0  0  0  0 |  b = Pyrite
 0  2  0  0  0  0 |  c = Charcopyrite
 0  0  0  3  0  0 |  d = Galene
 3  0  0  0  0  0 |  e = Spahlerite
 0  0  0  0  9 |  f = Sediments terrigenes
```



Neural Network with Normalization:

A Multi-Layer perceptron was used as a Non-Descriptive Classifier. Before using Neural Network the data was normalized from scale 0 to 1.0. The dataset was split into two parts, the training data and test data. We used a split factor of 0.8 to split between train and test data. The training was performed to classify into 6 classes.

Summary

Correctly Classified Instances	19(95%)
Incorrectly Classified Instances	1(5%)
Kappa statistic	0.7778
Mean absolute error	0.0495
Root mean squared error	0.1343
Relative absolute error	40.0636%
Root relative squared error	61.502%
Total Number of Instances	20

Detailed Accuracy By Class: -

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.333	0.944	1.000	0.971	0.793	0.922	0.986	R. Carbonatees and R. Carbonatees impures
	0.000	0.000	?	0.000	?	?	0.737	0.167	Pyrite
	?	0.000	?	?	?	?	?	?	Charcopyrite
	?	0.000	?	?	?	?	?	?	Galene
	?	0.000	?	?	?	?	?	?	Spahlerite
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Sediments terrigenes
Weighted Average	0.950	0.283	?	0.950	?	?	0.920	0.946	

Confusion Matrix:

```
a  b  c  d  e  f   <-- classified as
17  0  0  0  0  0 |  a = R. Carbonatees AND R. Carbonatees impures
1  0  0  0  0  0 |  b = Pyrite
0  0  0  0  0  0 |  c = Charcopyrite
0  0  0  0  0  0 |  d = Galene
0  0  0  0  0  0 |  e = Spahlerite
```

Experiment 2:

In Experiment 2 we used all records to perform the contrast classification (contrast learning), i.e. contrasting class C1 with a class notC1 that contains other classes. This is a binary classification where C1 can be resembled as 1 and notC1 as 0. The class labels were modified for Transfer Learning. The class labels were changed to C1 and notC1.

Decision Tree with Discretization:

Equal width binning

We used a technique of equal binning with the tool. Here's the run information for the same. We optimized the bin size by setting findBinNums to True to get the optimum bin value.

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Updated_Bakarydata 2-weka.filters.supervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-O-B10-M1.0-Rfirst-last-precision6
Instances: 98
Attributes: 45

Echantillon	Type de roche	S	Zn	Pb	CaO+MgO	Al2O3	TiO2	Fe2O3*	Cu
MnO	MgO	Cr	CaO	Na2O	K2O	P2O5	Ni	Sc	V
Ba	Sr	Li	Rb	Y	Zr	Nb	Cs	La	Ce
Pr	Nd	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm
Yb	Lu	Hf	Ta	Th	U				

Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree

```
Ni = '(-inf-23.87]'
|   Zr = '(-inf-16.55]': R. Carbonatees AND R. Carbonatees impures (84.0/8.0)
|   Zr = '(16.55-inf)': NOT R. Carbonatees AND R. Carbonatees impures (7.0/1.0)
Ni = '(23.87-inf)': NOT R. Carbonatees AND R. Carbonatees impures (7.0)
```

Number of Leaves: 3

Size of the tree: 5

Time taken to build model: 0 seconds

Stratified cross-validation

Summary

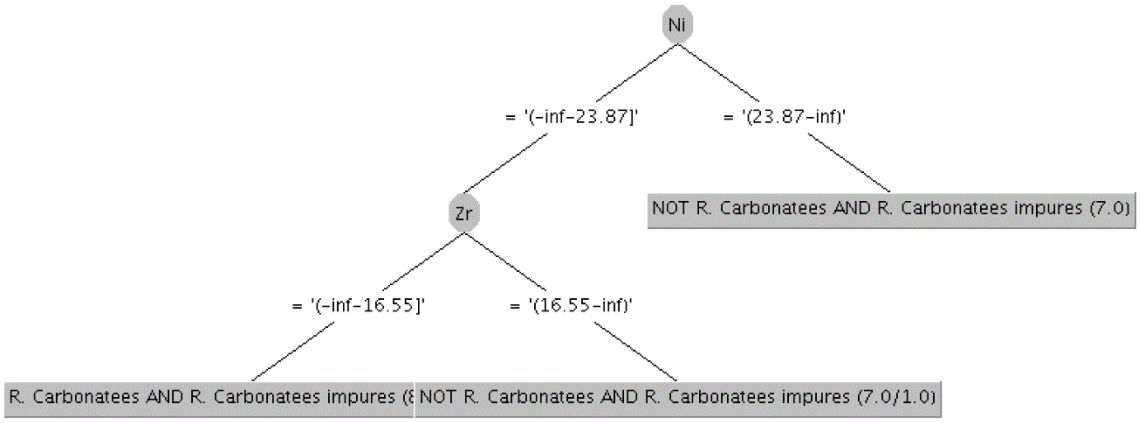
Correctly Classified Instances	83 (84.6939 %)
Incorrectly Classified Instances	15 (15.3061 %)
Kappa statistic	0.4134
Mean absolute error	0.216
Root mean squared error	0.3666
Relative absolute error	63.4166 %
Root relative squared error	89.2688 %
Total Number of Instances	98

Detailed Accuracy By Class: -

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.987	0.667	0.844	0.987	0.910	0.480	0.645	0.828	R. Carbonatees and R. Carbonatees impures
	0.333	0.013	0.875	0.333	0.483	0.480	0.645	0.482	Not R. Carbonatees and R. Carbonatees impures
Weighted Average	0.847	0.527	0.851	0.847	0.819	0.480	0.645	0.754	

Confusion Matrix

```
 a  b   <-- classified as
76  1 |  a = R. Carbonatees AND R. Carbonatees impures
 7  |  b = NOT R. Carbonatees AND R. Carbonatees impures
```



Equal Frequency Binning

We used a technique of equal frequency binning with the tool. Here's the run information for the same. Equal Frequency binning does not support findBinNums setting.

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Updated_Bakarydata 2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-F-B10-M-1.0-Rfirst-last-precision6
Instances: 98
Attributes: 45

Echantillon	Type de roche	S	Zn	Pb	CaO+MgO	Al2O3	TiO2	Fe2O3*	Cu
MnO	MgO	Cr	CaO	Na2O	K2O	P2O5	Ni	Sc	V
Ba	Sr	Li	Rb	Y	Zr	Nb	Cs	La	Ce
Pr	Nd	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm
Yb	Lu	Hf	Ta	Th	U				

Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree

Hf = '(-inf-0.05]': R. Carbonatees AND R. Carbonatees impures (67.0/7.0)
Hf = '(0.05-0.15]': R. Carbonatees AND R. Carbonatees impures (16.0/2.0)
Hf = '(0.15-0.25]': NOT R. Carbonatees AND R. Carbonatees impures (1.0)
Hf = '(0.25-0.35]': R. Carbonatees AND R. Carbonatees impures (2.0/1.0)
Hf = '(0.35-0.55]': R. Carbonatees AND R. Carbonatees impures (2.0/1.0)
Hf = '(0.55-1.05]': NOT R. Carbonatees AND R. Carbonatees impures (2.0)
Hf = '(1.05-1.25]': R. Carbonatees AND R. Carbonatees impures (2.0/1.0)
Hf = '(1.25-1.65]': NOT R. Carbonatees AND R. Carbonatees impures (2.0)
Hf = '(1.65-3.6]': NOT R. Carbonatees AND R. Carbonatees impures (2.0)
Hf = '(3.6-inf)': NOT R. Carbonatees AND R. Carbonatees impures (2.0)

Number of Leaves: 10

Size of the tree: 11

Time taken to build model: 0.01 seconds

Stratified cross-validation

Summary: -

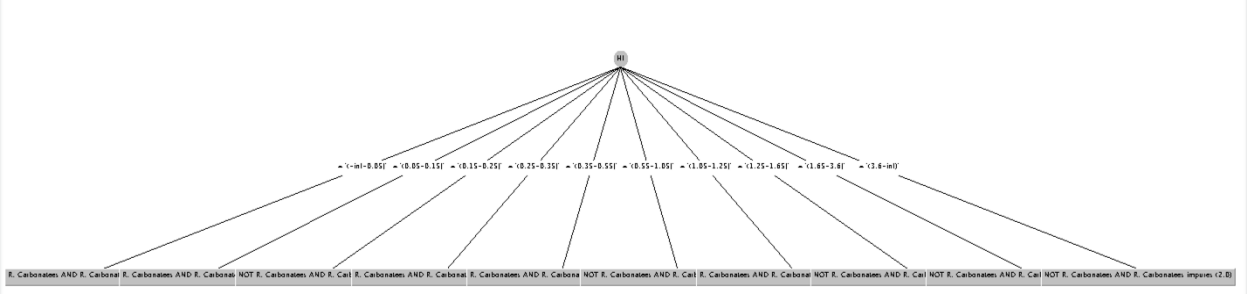
Correctly Classified Instances	80 (81.6327%)
Incorrectly Classified Instances	18 (18.3673%)
Kappa statistic	0.3403
Mean absolute error	0.2411
Root mean squared error	0.402
Relative absolute error	70.7912%
Root relative squared error	97.9039%
Total Number of Instances	98

Detailed Accuracy By Class: -

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.948	0.667	0.839	0.948	0.890	0.366	0.551	0.796	R. Carbonatees and R. Carbonatees impures
	0.333	0.052	0.636	0.333	0.437	0.366	0.551	0.388	Carbonatees AND R. Carbonatees impures
Weighted Average	0.816	0.535	0.796	0.816	0.793	0.366	0.551	0.708	

Confusion Matrix

```
 a  b   <-- classified as
73  4 |  a = R. Carbonatees AND R. Carbonatees impures
14  7 |  b = NOT R. Carbonatees AND R. Carbonatees impures
```



Neural Network with Normalization:

A Multi-Layer perceptron was used as a Non-Descriptive Classifier. Before using Neural Network the data was normalized from scale 0 to 1.0. The dataset was split into two parts, the training data and test data. We used a split factor of 0.78 to split between train and test data.

Summary

Correctly Classified Instances	20 (90.9091%)
Incorrectly Classified Instances	2 (9.0909%)
Kappa statistic	0.6333
Mean absolute error	0.0591
Root mean squared error	0.1731
Relative absolute error	45.6632 %
Root relative squared error	73.8593 %
Total Number of Instances	22

Detailed Accuracy By Class:-

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.500	0.900	1.000	0.947	0.671	0.958	0.991	R. Carbonatees and R. Carbonatees impures
	0.000	0.000	?	0.000	?	?	0.810	0.200	Pyrite
	?	0.000	?	?	?	?	?	?	Charcopyrite
	?	0.000	?	?	?	?	?	?	Galene
	0.000	0.000	?	0.000	?	?	0.857	0.250	Spahlerite
Weighted Average	1.000	0.000	0.409	?	0.909	?	0.951	0.922	Sediments terrigenes

Confusion Matrix:

```
 a  b  c  d  e  f   <-- classified as
18  0  0  0  0  0 |  a = R. Carbonatees AND R. Carbonatees impures
 1  0  0  0  0  0 |  b = Pyrite
 0  0  0  0  0  0 |  c = Charcopyrite
 0  0  0  0  0  0 |  d = Galene
 1  0  0  0  0  0 |  e = Spahlerite
 0  0  0  0  0  2 |  f = Sediments terrigenes
```

Neural Network For Binary:

A Multi-Layer perceptron was used as a Non-Descriptive Classifier. Before using Neural Network the data was normalized to a scale of 0 to 1.0. Also, the class labels were changed to support only two classes ie C1 and notC1. The dataset was split into two parts, the training data and test data. We used a split factor of 0.78 to split between train and test data.

Summary

Correctly Classified Instances	18 (90 %)
Incorrectly Classified Instances	2 (10 %)
Kappa statistic	0.4595
Mean absolute error	0.1088
Root mean squared error	0.2615
Relative absolute error	34.3918 %
Root relative squared error	71.132 %
Total Number of Instances	20

Detailed Accuracy By Class: -

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.667	0.895	1.000	0.944	0.546	0.824	0.961	R. Carbonatees and R. Carbonatees impures
	0.333	0.000	1.000	0.333	0.500	0.546	0.824	0.750	Not R. Carbonatees and R. Carbonatees impures
Weighted Average	0.900	0.567	0.911	0.900	0.878	0.546	0.824	0.929	

Confusion Matrix:

```
a  b   <-- classified as
17  0 |  a = R. Carbonatees AND R. Carbonatees impures
2  1 |  b = NOT R. Carbonatees AND R. Carbonatees impures
```

Experiment 3

In Experiment 3 we repeated Experiments 1, 2 for all records with the most important attributes as defined by the expert only. According to Experts CaO+MgO, Fe2O3, MgO, CaO, S, Zn, Pb and Cu are the most important fields. We will repeat the above experiments with the above steps.

Decision Tree with Discretization:

Equal width binning (Classification on 6 classes)

We used a technique of equal binning with a J48 Decision Tree using the tool. Here's the run information for the same. We optimized the bin size by setting findBinNums to True to get the optimum bin value.

Run information

Scheme: **weka.classifiers.trees.J48 -C 0.25 -M 2**
Relation: **Updated_Bakarydata-weka.filters.unsupervised.attribute.Remove-R1,4-5,7,10-12,17-46-weka.filters.unsupervised.attribute.Discretize-O-B10-M1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribute.Discretize-O-B10-M1.0-Rfirst-last-precision6**
Instances: **98**
Attributes: **9**

Type de roche	CaO+MgO	Fe2O3*	MgO	CaO	S	Zn	Pb	Cu

Test mode: **10-fold cross-validation**

Classifier model (full training set)

J48 pruned tree

CaO+MgO = '(-inf-6.287]': Sediments terrigenes (7.0)
CaO+MgO = '(6.287-12.094]': R. Carbonatees AND R. Carbonatees impures (0.0)
CaO+MgO = '(12.094-17.901]': Galene (1.0)
CaO+MgO = '(17.901-inf)': R. Carbonatees AND R. Carbonatees impures (90.0/13.0)

Number of Leaves: **4**

Size of the tree: **5**

Time taken to build model: 0 seconds

Stratified cross-validation

Summary: -

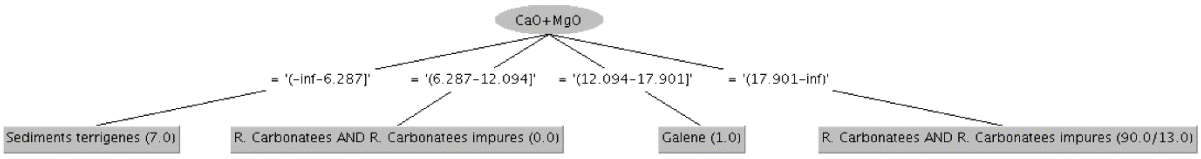
Correctly Classified Instances	84 (85.7143 %)
Incorrectly Classified Instances	14 (14.2857 %)
Kappa statistic	0.4586
Mean absolute error	0.0828
Root mean squared error	0.2092
Relative absolute error	62.0021 %
Root relative squared error	83.7536 %
Total Number of Instances	98

Detailed Accuracy By Class: -

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.667	0.846	1.000	0.917	0.531	0.629	0.825	R. Carbonatees and R. Carbonatees impures
	0.000	0.000	?	0.000	?	0.245	0.038	?	Pyrite
	0.000	0.000	?	0.000	?	0.253	0.019	?	Charcopyrite
	0.000	0.000	?	0.000	?	0.223	0.024	?	Galene
	0.000	0.000	?	0.000	?	0.202	0.027	?	Spahlerite
	0.778	0.000	1.000	0.778	0.875	0.872	0.807	0.799	Sediments terrigenes
Weighted Average	0.857	0.524	?	0.857	?	?	0.596	0.725	

Confusion Matrix

```
a  b  c  d  e  f  <-- classified as
77  0  0  0  0  0 |  a = R. Carbonatees AND R. Carbonatees impures
 4  0  0  0  0  0 |  b = Pyrite
 2  0  0  0  0  0 |  c = Charcopyrite
 3  0  0  0  0  0 |  d = Galene
 3  0  0  0  0  0 |  e = Spahlerite
 2  0  0  0  0  7 |  f = Sediments terrigenes
```



Equal width binning (Contrast learning)

We used a technique of equal binning with a J48 Decision Tree. Here's the run information for the same. We optimized the bin size by setting findBinNums to True to get the optimum bin value. As done in previous experiment for contrast learning we will rename our classes to C1 and notC1.

Run information
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Updated_Bakarydata-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R3-4,6,9-11,16-45
Instances: 98
Attributes: 9

Type de roche	CaO+MgO	Fe2O3*	MgO	CaO	S	Zn	Pb	Cu

Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree

```
Fe2O3* <= 0.45: R. Carbonatees AND R. Carbonatees impures (74.0/1.0)
Fe2O3* > 0.45
|   Zn <= 700.073684
|   |   Cu <= 331
|   |   |   CaO <= 22.84: Sediments terrigenes (10.0/1.0)
|   |   |   CaO > 22.84
|   |   |   |   S <= 2158: R. Carbonatees AND R. Carbonatees impures (4.0)
|   |   |   |   S > 2158: Pyrite (3.0)
|   |   Cu > 331: Charcopyrite (2.0)
|   Zn > 700.073684
|   |   Pb <= 5277: Spahlerite (2.0)
|   |   Pb > 5277: Galene (3.0)
```

Number of Leaves: 7

Size of the tree: 13

Time taken to build model: 0.01 seconds

Stratified cross-validation

Summary: -

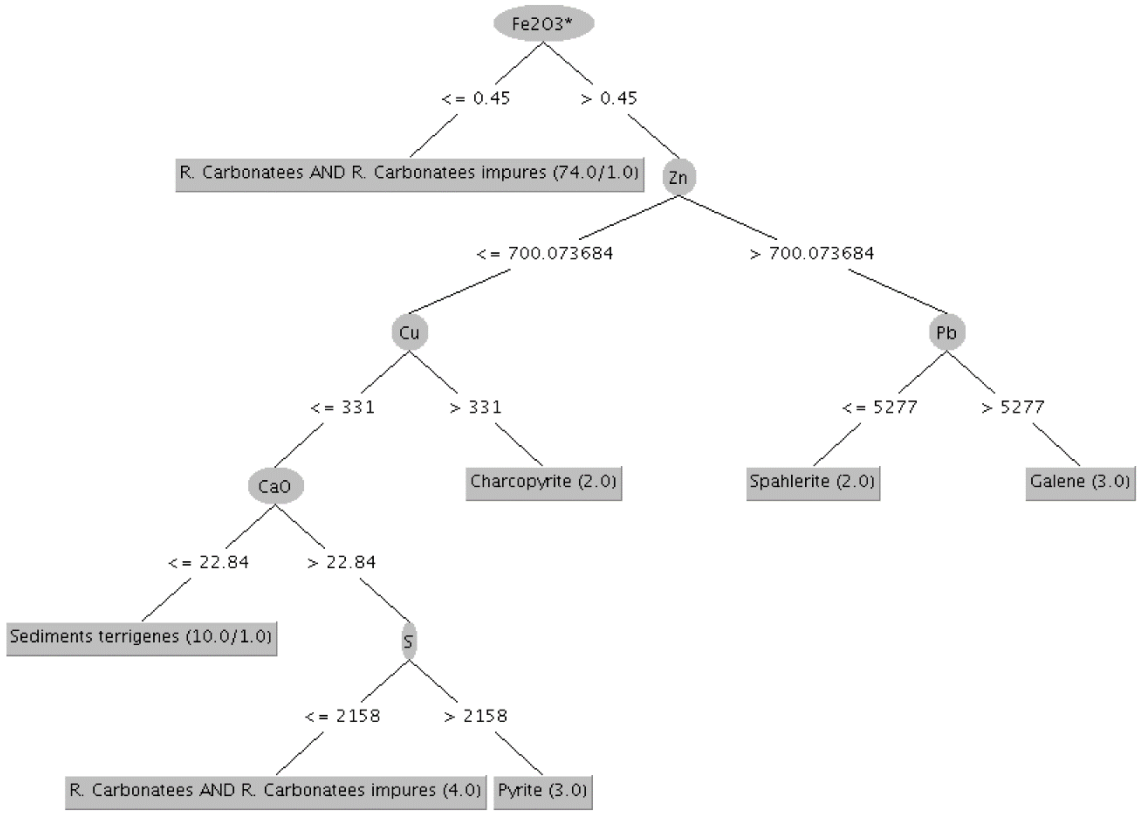
Correctly Classified Instances	83 (84.6939 %)
Incorrectly Classified Instances	15 (15.3061 %)
Kappa statistic	0.568
Mean absolute error	0.0524
Root mean squared error	0.2132
Relative absolute error	39.2087 %
Root relative squared error	85.3328 %
Total Number of Instances	98

Detailed Accuracy By Class: -

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.961	0.238	0.937	0.961	0.949	0.750	0.831	0.914	R. Carbonatees and R. Carbonatees impures
	0.250	0.021	0.333	0.250	0.286	0.263	0.582	0.114	Pyrite
	0.000	0.021	0.000	0.000	0.000	-0.021	0.729	0.135	Charcopyrite
	0.333	0.011	0.500	0.333	0.400	0.393	0.828	0.566	Galene
	0.000	0.021	0.000	0.000	0.000	-0.026	0.575	0.132	Spahlerite
	0.778	0.034	0.700	0.778	0.737	0.710	0.871	0.570	Sediments terrigenes
Weighted Average	0.847	0.192	0.829	0.847	0.837	0.676	0.814	0.799	

Confusion Matrix

```
 a  b  c  d  e  f   <-- classified as
74  1  1  0  0  1 |  a = R. Carbonatees AND R. Carbonatees impures
 2  1  0  0  0  1 |  b = Pyrite
 0  1  0  0  0  1 |  c = Charcopyrite
 0  0  0  1  2  0 |  d = Galene
 2  0  0  1  0  0 |  e = Spahlerite
 1  0  1  0  0  7 |  f = Sediments terrigenes
```



Equal Frequency Binning (classification on 6 classes)

We used a technique of equal frequency binning with a J48 classifier. Here's the run information for the same. Equal Frequency binning does not support findBinNums setting.

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Updated_Bakarydata-weka.filters.unsupervised.attribute.Discretize-F-B10-M1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribute.Remove-R1,4-5,7,10-12,17-46-weka.filters.unsupervised.attribute.Discretize-F-B10-M1.0-Rfirst-last-precision6
Instances: 98
Attributes: 9[Type de roche,CaO+MgO,Fe2O3*,MgO,CaO,S,Zn,Pb,Cu]
Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree:-

R. Carbonatees AND R. Carbonatees impures (98.0/21.0)

Number of Leaves: 1

Size of the tree: 1

Time taken to build model: 0 seconds

Stratified cross-validation

Summary

Correctly Classified Instances	77 (78.5714%)
Incorrectly Classified Instances	21 (21.4286%)
Kappa statistic	0
Mean absolute error	0.1238
Root mean squared error	0.2492
Relative absolute error	92.6592%
Root relative squared error	99.7704%
Total Number of Instances	98

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.786	1.000	0.880	?	0.465	0.771	R. Carbonatees and R. Carbonatees impures
	0.000	0.000	?	0.000	?	?	0.182	0.035	Pyrite
	0.000	0.000	?	0.000	?	?	0.094	0.020	Charcopyrite
	0.000	0.000	?	0.000	?	?	0.133	0.025	Galene
	0.000	0.000	?	0.000	?	?	0.142	0.031	Spahlerite
	0.000	0.000	?	0.000	?	?	0.435	0.088	Sediments terrigenes
Weighted Average	0.786	0.786	?	0.786	?	?	0.423	0.618	

Confusion Matrix

```
a  b  c  d  e  f   <-- classified as
77  0  0  0  0  0 |  a = R. Carbonatees AND R. Carbonatees impures
 4  0  0  0  0  0 |  b = Pyrite
 2  0  0  0  0  0 |  c = Charcopyrite
 3  0  0  0  0  0 |  d = Galene
 3  0  0  0  0  0 |  e = Spahlerite
 9  0  0  0  0  0 |  f = Sediments terrigenes
```

Equal Frequency Binning (contrast learning)

We used a technique of equal frequency binning with a J48 Decision Tree. Here's the run information for the same. Equal Frequency binning does not support findBinNums setting.

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Updated_Bakarydata 2-weka.filters.unsupervised.attribute.Remove-R1,4-5,7,10-12,17-46-weka.filters.unsupervised.attribute.Discretize-F-B10-M-1.0-Rfirst-last-precision6
Instances: 98
Attributes: 9[Type de roche,CaO+MgO,Fe2O3*,MgO,CaO,S,Zn,Pb,Cu]
Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree

Fe2O3* = '(-inf-0.065]': R. Carbonatees AND R. Carbonatees impures (10.0)
Fe2O3* = '(0.065-0.085]': R. Carbonatees AND R. Carbonatees impures (10.0)
Fe2O3* = '(0.085-0.135]': R. Carbonatees AND R. Carbonatees impures (9.0)
Fe2O3* = '(0.135-0.165]': R. Carbonatees AND R. Carbonatees impures (12.0)
Fe2O3* = '(0.165-0.235]': R. Carbonatees AND R. Carbonatees impures (10.0)
Fe2O3* = '(0.235-0.285]': R. Carbonatees AND R. Carbonatees impures (9.0)
Fe2O3* = '(0.285-0.405]': R. Carbonatees AND R. Carbonatees impures (10.0/1.0)
Fe2O3* = '(0.405-0.715]': NOT R. Carbonatees AND R. Carbonatees impures (9.0/4.0)
Fe2O3* = '(0.715-2.78]'
| S = '(-inf-84.5]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(84.5-129]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(129-189]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(189-266.5]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(266.5-382]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(382-659.5]': NOT R. Carbonatees AND R. Carbonatees impures (1.0)
| S = '(659.5-866.5]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(866.5-1758]': R. Carbonatees AND R. Carbonatees impures (4.0)
| S = '(1758-13467]': NOT R. Carbonatees AND R. Carbonatees impures (2.0)
| S = '(13467-inf)': NOT R. Carbonatees AND R. Carbonatees impures (3.0)
Fe2O3* = '(2.78-inf)': NOT R. Carbonatees AND R. Carbonatees impures (9.0)

Number of Leaves: 19

Size of the tree: 21

Time taken to build model: 0 seconds

Stratified cross-validation

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Updated_Bakarydata 2-weka.filters.unsupervised.attribute.Remove-R1,4-5,7,10-12,17-46-weka.filters.unsupervised.attribute.Discretize-F-B10-M-1.0-Rfirst-last-precision6
Instances: 98
Attributes: 9 [Type de roche,CaO+MgO,Fe2O3*,MgO,CaO,S,Zn,Pb,Cu]

Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree

Fe2O3* = '(-inf-0.065]': R. Carbonatees AND R. Carbonatees impures (10.0)
Fe2O3* = '(0.065-0.085]': R. Carbonatees AND R. Carbonatees impures (10.0)
Fe2O3* = '(0.085-0.135]': R. Carbonatees AND R. Carbonatees impures (9.0)
Fe2O3* = '(0.135-0.165]': R. Carbonatees AND R. Carbonatees impures (12.0)
Fe2O3* = '(0.165-0.235]': R. Carbonatees AND R. Carbonatees impures (10.0)
Fe2O3* = '(0.235-0.285]': R. Carbonatees AND R. Carbonatees impures (9.0)
Fe2O3* = '(0.285-0.405]': R. Carbonatees AND R. Carbonatees impures (10.0/1.0)
Fe2O3* = '(0.405-0.715]': NOT R. Carbonatees AND R. Carbonatees impures (9.0/4.0)
Fe2O3* = '(0.715-2.78]'
| S = '(-inf-84.5]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(84.5-129]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)

| S = '(129-189]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(189-266.5]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(266.5-382]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(382-659.5]': NOT R. Carbonatees AND R. Carbonatees impures (1.0)
| S = '(659.5-866.5]': NOT R. Carbonatees AND R. Carbonatees impures (0.0)
| S = '(866.5-1758]': R. Carbonatees AND R. Carbonatees impures (4.0)
| S = '(1758-13467]': NOT R. Carbonatees AND R. Carbonatees impures (2.0)
| S = '(13467-inf)': NOT R. Carbonatees AND R. Carbonatees impures (3.0)
Fe2O3* = '(2.78-inf)': NOT R. Carbonatees AND R. Carbonatees impures (9.0)

Number of Leaves : 19

Size of the tree: 21

Time taken to build model: 0 seconds

Stratified cross-validation

Summary

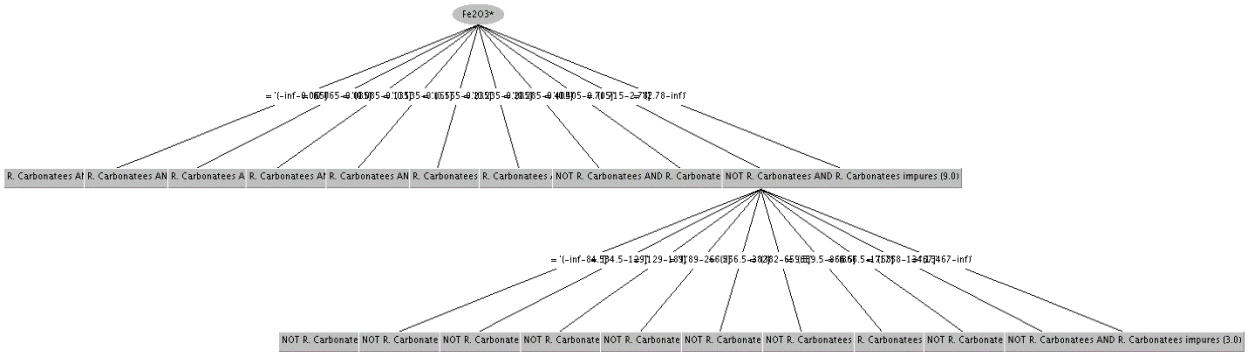
Correctly Classified Instances	89 (90.8163%)
Incorrectly Classified Instances	9 (9.1837%)
Kappa statistic	0.7225
Mean absolute error	0.0969
Root mean squared error	0.2669
Relative absolute error	28.458%
Root relative squared error	65.0035%
Total Number of Instances	98

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.948	0.238	0.936	0.948	0.942	0.723	0.877	0.939	R. Carbonatees and R. Carbonatees impures
	0.762	0.052	0.800	0.762	0.780	0.723	0.877	0.777	Not R. Carbonatees and R. Carbonatees impures
Weighted Average	0.908	0.198	0.907	0.908	0.907	0.723	0.877	0.904	

Confusion Matrix

a b <-- classified as
73 4 | a = R. Carbonatees AND R. Carbonatees impures
5 16 | b = NOT R. Carbonatees AND R. Carbonatees impures



Neural Network with Normalization (Classification):

A Multi-Layer perceptron was used as a Non-Descriptive Classifier. Before using Neural Network the data was normalized from scale 0 to 1.0. The dataset was split into two parts, the training data and test data. We used a split factor of 0.83 to split between train and test data.

Summary

Correctly Classified Instances	17(94.4444%)
Incorrectly Classified Instances	1 (5.5556%)
Kappa statistic	0.775
Mean absolute error	0.0639
Root mean squared error	0.1483
Relative absolute error	50.9164 %
Root relative squared error	65.6565 %
Total Number of Instances	18

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.333	0.938	1.000	0.968	0.791	0.911	0.981	R. Carbonatees and R. Carbonatees impures
	0.000	0.000	?	0.000	?	?	0.706	0.167	Pyrite
	?	0.000	?	?	?	?	?	?	Charcopyrite
	?	0.000	?	?	?	?	?	?	Galene
	?	0.000	?	?	?	?	?	?	Spahlerite
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Sediments terrigenes
Weighted Average	0.944	0.278	?	0.944	?	?	0.910	0.938	

Confusion Matrix

```
 a  b  c  d  e  f   <-- classified as
15  0  0  0  0  0 |  a = R. Carbonatees AND R. Carbonatees impures
 1  0  0  0  0  0 |  b = Pyrite
 0  0  0  0  0  0 |  c = Charcopyrite
 0  0  0  0  0  0 |  d = Galene
 0  0  0  0  0  0 |  e = Spahlerite
 0  0  0  0  0  2 |  f = Sediments terrigenes
```

Neural Network with Normalization (Contrast Learning):

A Multi-Layer perceptron was used as a Non-Descriptive Classifier. Before using Neural Network the data was normalized from scale 0 to 1.0. The dataset was split into two parts, the training data and test data. We used a split factor of 0.84 to split between train and test data. The class labels were changed to C1 and notC1 to perform contrast learning.

Summary

Correctly Classified Instances	15 (93.75%)
Incorrectly Classified Instances	1 (6.25%)
Kappa statistic	0.7647
Mean absolute error	0.1166
Root mean squared error	0.2547
Relative absolute error	35.4521 %
Root relative squared error	64.9436 %
Total Number of Instances	16

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.333	0.929	1.000	0.963	0.787	0.923	0.982	R. Carbonatees and R. Carbonatees impures
	0.667	0.000	1.000	0.667	0.800	0.787	0.923	0.833	Not R. Carbonatees and R. Carbonatees impures
Weighted Average	0.938	0.271	0.942	0.938	0.932	0.787	0.923	0.954	

Confusion Matrix

```
 a  b   <-- classified as
13  0 |  a = R. Carbonatees AND R. Carbonatees impures
 1  2 |  b = NOT R. Carbonatees AND R. Carbonatees impures
```

RESULTS

This project required performing various experiments using data mining concepts. In the first experiment, we used all the records to perform a full classification and used a J48 Decision Tree for a Descriptive Classifier and a Multi-Layer Perceptron for a Non-Descriptive Classifier.

When we use a decision tree with discretization and used equal binning we used 98 instances and 46 attributes, we correctly classified 90 instances and incorrectly 8 instances. The mean absolute error was 0.0258. When we use equal frequency binning instead of equal width binning with the same number of instances and attributes, we correctly classify 91 instances and incorrectly classify 7 instances. The mean absolute error in this case is 0.0305.

For the first experiment we also use neural network with normalization with the scale lying between 0 to 1.0. The split factor between train and test was 0.8. We correctly classified 19 instances and 1 wrong instance. The mean absolute error was determined to be 0.0495.

For the second experiment, we use all records to perform the contrast classification i.e contrasting class C1 with a class notC1 that contains the other classes. When use decision tree with discretization and equal width binning. There were 96 instances and 45 attributes and we classified 83 instances and incorrectly 15 instances. The mean absolute error 0.216. The use of equal frequency binning saw 80 correct instances and 18 incorrectly. The mean square error 0.2411. We used neural network with normalization we correctly classified 20 instances and 2 wrong instances. The mean absolute error was 0.0591.

The third experiment required repeating experiments 1 and 2 for all records with the most important attributes. Decision tree with discretization with equal binning led to 84 correctly classified instances and 14 incorrectly classified instances. For equal width binning, we have 83 correctly classified instances and 15 incorrectly classified instances with a mean absolute error of 0.0524. Equal frequency binning saw 77 instances being correctly classified and 21 incorrectly with a mean absolute error of 0.1238. When applied with contrast learning we got 89 correct instances and 9 incorrect instances with a mean absolute error of 0.0969. Neural network with normalization(83% split) led to 17 correctly classified instances 1 incorrectly classified instance. Upon using contrast learning we got 15 correct and 1 incorrect instances.

CONCLUSION

We performed various experiments on the data set provided to use and used certain data mining techniques to obtain the results. The analyses were done on Weka and the results were as expected based on the various techniques used. The project was successfully executed based on the directives provided and the various results have been documented above.

BIBLIOGRAPHY

- https://en.wikipedia.org/wiki/Artificial_neural_network
- http://www.saedsayad.com/decision_tree.htm
- https://www.tutorialspoint.com/data_mining/dm_dti.htm
- https://en.wikipedia.org/wiki/Weka_(machine_learning)
- http://www.cs.waikato.ac.nz/~ ml/weka/index.html
- https://www.mathworks.com/products/neural-network.html?requestedDomain=www.mathworks.com
- http://www.simbrain.net/
- https://en.wikipedia.org/wiki/Multilayer_perceptron/
- https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/discretization-methods-data-mining
- https://en.wikipedia.org/wiki/Data_cleansing