DATA MINING

Project Analysis

Group Members:

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

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.

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: