Sudhin Domala, Shariq Syed, Allan Lee
CSE 352 Project: Data Prep Explanation
110475495

Removing attributes:

My criteria for removing the attributes was if over 25% of the data for that column was missing. I thought this would be a good amount because if we have to fill in more than 25% of the missing data with a value then it creates a bias towards that value during the learning process. On WEKA, I went to the attributes and examined each of them to find the percentage of data missing from it. If it exceeded 25% then I checked off the box for it. After finding all attributes to remove from the data set, I clicked on the remove button on the program. As a result, I removed 8 attributes with them being:

1) Pb as it was missing 55% of its data
2) As as it was missing 72% of its data
3) Cd as it was missing 70% of its data
4) Ni as it was missing 40% of its data
5) Sc as it was missing 50% of its data
6) Co as it was missing 84% of its data
7) Mo as it was missing 88% of its data
8) Li as it was missing 39% of its data

Filling in missing data:

To fill in missing data in each attribute column, I went to WEKA tool and clicked on the filter. I then went under supervised folder, then under the attribute folder, and selected the filter called ReplaceMissingValues. I applied that filter to the data which replaced missing data from each attribute with the mean of the numeric values.

Discretizing the data:

To discretize the data and bin them, I went to find another filter on the WEKA tool. I went under unsupervised folder and then under the attribute folder. There, I clicked on Discretize which is another filter. I double clicked on the filter to edit the settings and changed the number of bins from 10 to 3 or 4 on a tool. To end the Discretizing phase, I saw if I need to replace whatever values in the bins by the appropriate numbers given the nature of the Bakary Data.

- Converted file to CSV and removed extra “type de Roche” column
- Attached CSV is in submissions.