KNOWLEDGE ENGINEERING LAB (CSE 4.1.7)

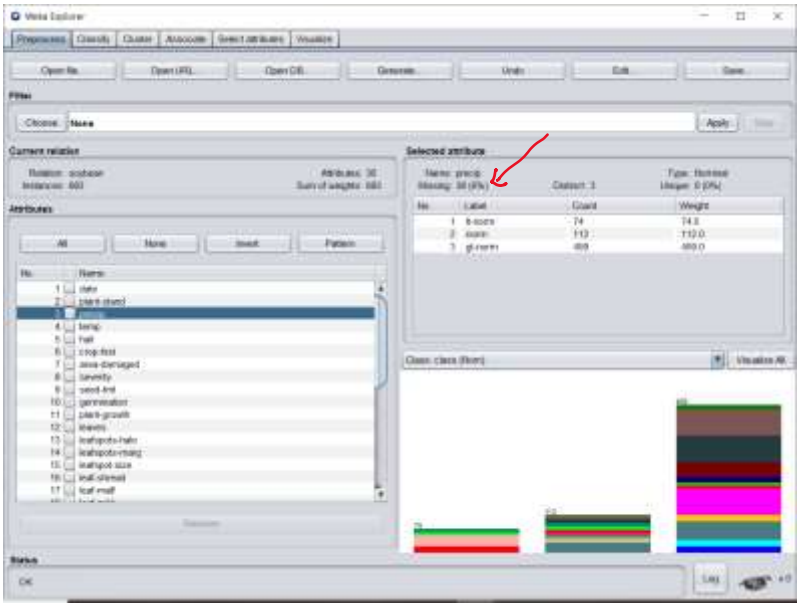
5. Performing data preprocessing in Weka – Part1

Study Unsupervised Attribute Filters such as "ReplaceMissingValues" to replace missing values in the given dataset, "Add" to add the new attribute Average, Discretize to discretize the attributes into bins. Explore Normalize and Standardize options on a dataset with numerical attributes.

Finding missing values in the dataset:

1. Launch Weka-> click on the tab Explorer
2. Load a dataset. (Click on "Open File" & locate the datafile)
3. Click on PreProcess tab & then look at your lower R.H.S. bottom window click on drop down arrow and choose "No Class"
4. Click on "Edit" tab, a new window opens up that will show you the loaded datafile. By looking at your dataset you can also find out if there are missing values in it or not. Also please note the attribute types on the column header. It would either be 'nominal' or 'numeric'.

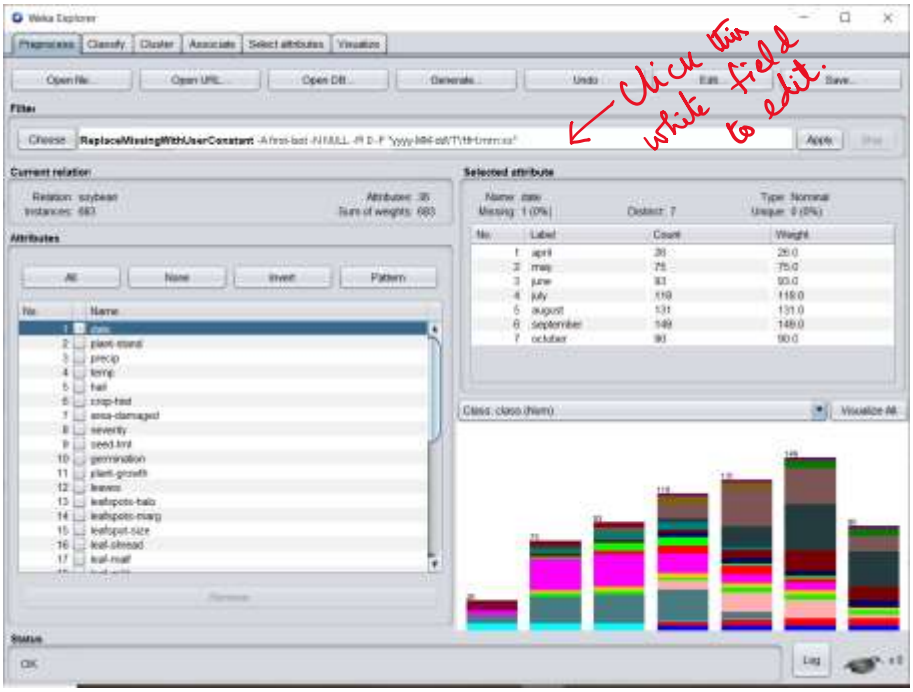
If your data has missing values then its best to clean it first before you apply any forms of algorithm to it. Please look below at Figure, you will see the highlighted fields are blank that means the data at hand is dirty and it first needs to be cleaned.



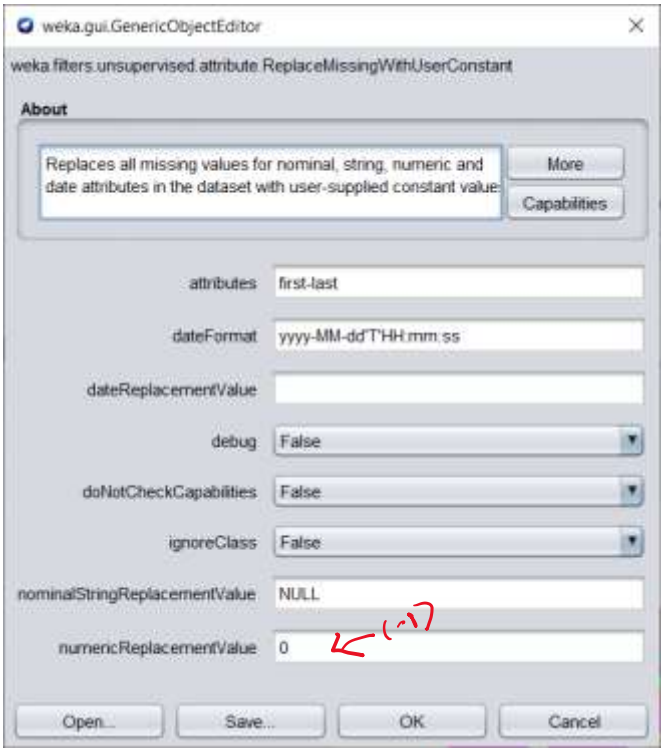
Replace Missing Values:

To clean the data, you apply “Filters” to it. Generally the data will be missing with values, so the filter to apply is “ReplaceMissingWithUserConstant” (the filter choice may vary according to our need). Click on Choose button below

- Filters
    - Unsupervised
        - Attribute
            - ReplaceMissingWithUserConstant



A good choice for replacing missing numeric values is to give it values like -1 or 0 and for string values it could be NULL.



Click on Ok and then Apply

Now, all nominal missing values are replaced by NULL and numeric values with 0.

"Add" to add the new attribute Average:

Let us assume we have dataset features X and Y in the given dataset. Requirement is to add another feature that is the average of X and Y.

- Open the CSV file
- Here X is considered as a1 and Y is considered as a2
- Click on choose button select Add Expression Filter (Weka->Filters->Unsupervised->Attribute)
- Click on the text box next to the choose button where AddExpression is appearing.
- Type the expression (a1+a2)/2 in expression text box
- Click on ok and then apply after the choose.