Data Warehousing & Data Mining LAB - G2
EXPERIMENT 4

- ASHISH KUMAR
- 2K18/SE/041

Aim:- Explain missing data treatment in WEKA on a sample dataset.

Theory:- Data is rarely clean and often you can have corrupt or missing values. It is important to identify, mark and handle missing data when developing machine learning models in order to get the very best performance. We will learn to do this using WEKA.

STEPS TO FOLLOW:-

Mark Missing Data

- Open the Weka Explorer.
- Load the Pima Indians onset of diabetes dataset.
- Click the "Choose" button for the Filter and select NumericalCleaner, it us under unsupervized.attribute.NumericalCleaner.
- Click on the filter to configure it.
- Set the attributeIndicies to 6, the index of the mass attribute.
- Set minThreshold to 0.1E-8 (close to zero), which is the minimum value allowed for the attribute.
- Set minDefault to NaN, which is unknown and will replace values below the threshold.
- Click the "OK" button on the filter configuration.
- Click the "Apply" button to apply the filter.
- Click "mass" in the "attributes" pane and review the details of the "selected attribute". Notice that the 11 attribute values that were formally set to 0 are now marked as Missing.

Remove Missing data

- Click the "Choose" button for the Filter and select RemoveWithValues, it us under unsupervized.instance.RemoveWithValues.
- Click on the filter to configure it.
- Set the attributeIndicies to 6, the index of the mass attribute.
- Set matchMissingValues to "True".
- Click the "OK" button to use the configuration for the filter.
- Click the "Apply" button to apply the filter.
- Click "mass" in the "attributes" section and review the details of the "selected attribute".
- Notice that the 11 attribute values that were marked Missing have been removed from the dataset.

DATASET

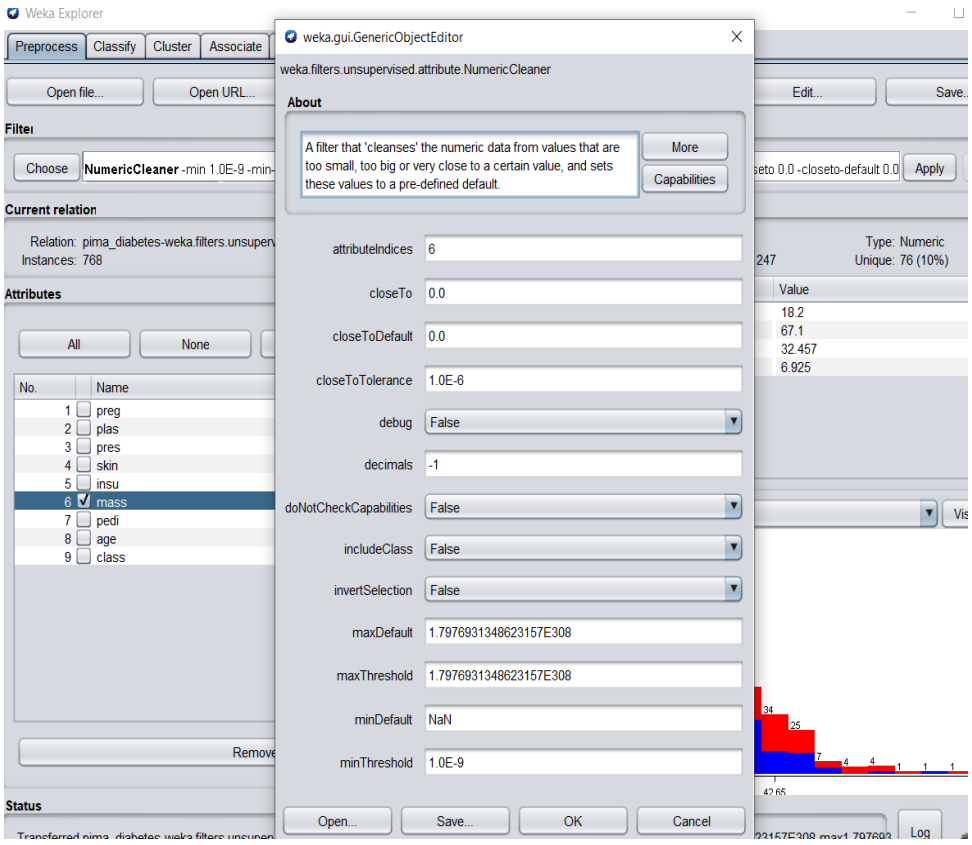
diabetes.arff

The database of this dataset is Pima Indians Diabetes. This dataset predicts whether the patient is prone to be diabetic in the next 5 years. The patients in this dataset are all females of at least 21 years of age from Pima Indian Heritage. It has 768 instances and 8 numerical attributes plus a class. This is a binary classification dataset where the output variable predicted is nominal comprising of two classes.
Some attributes such as blood pressure (pres) and Body Mass Index (mass) have values of zero, which are impossible. These are examples of corrupt or missing data that must be marked manually.

RESULTS:-

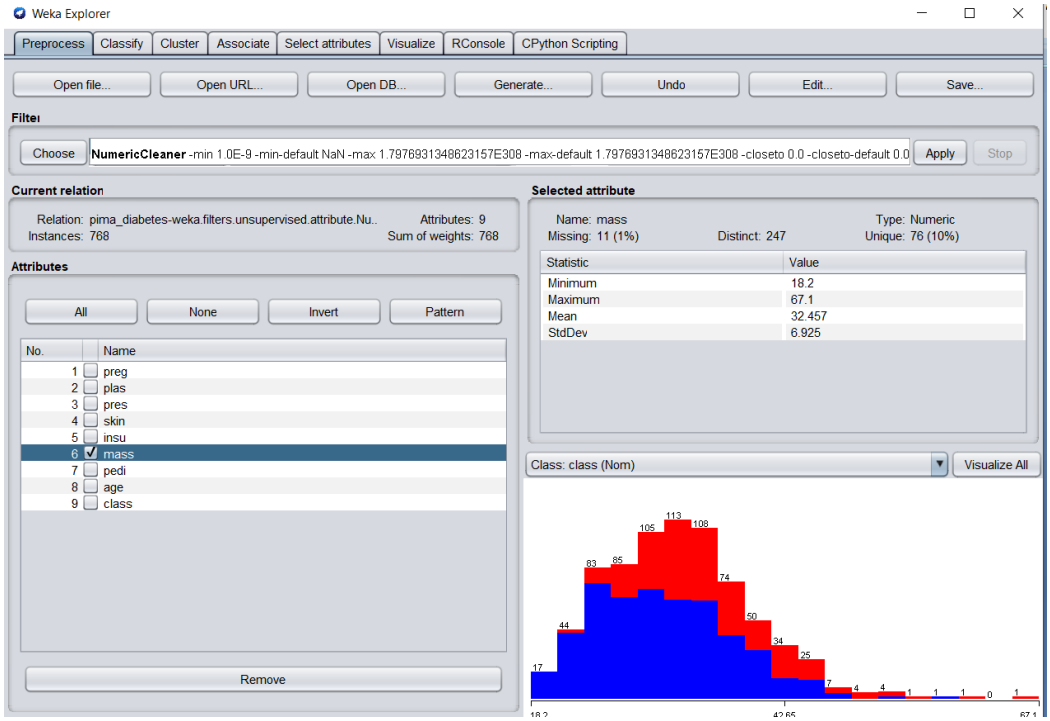
Mark missing data

Applying filter



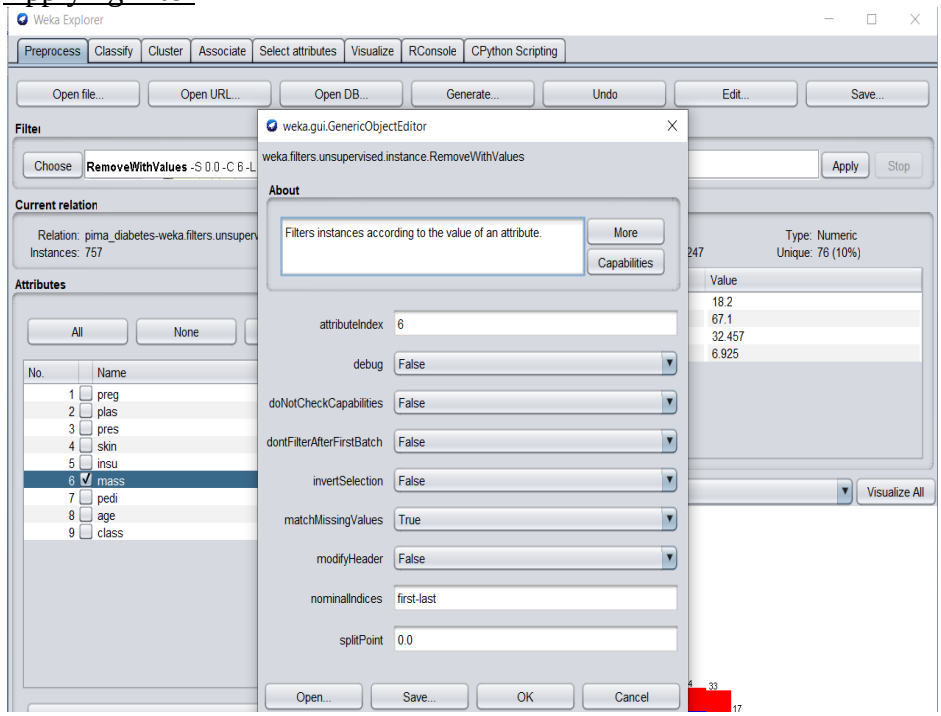
Setting mass attribute

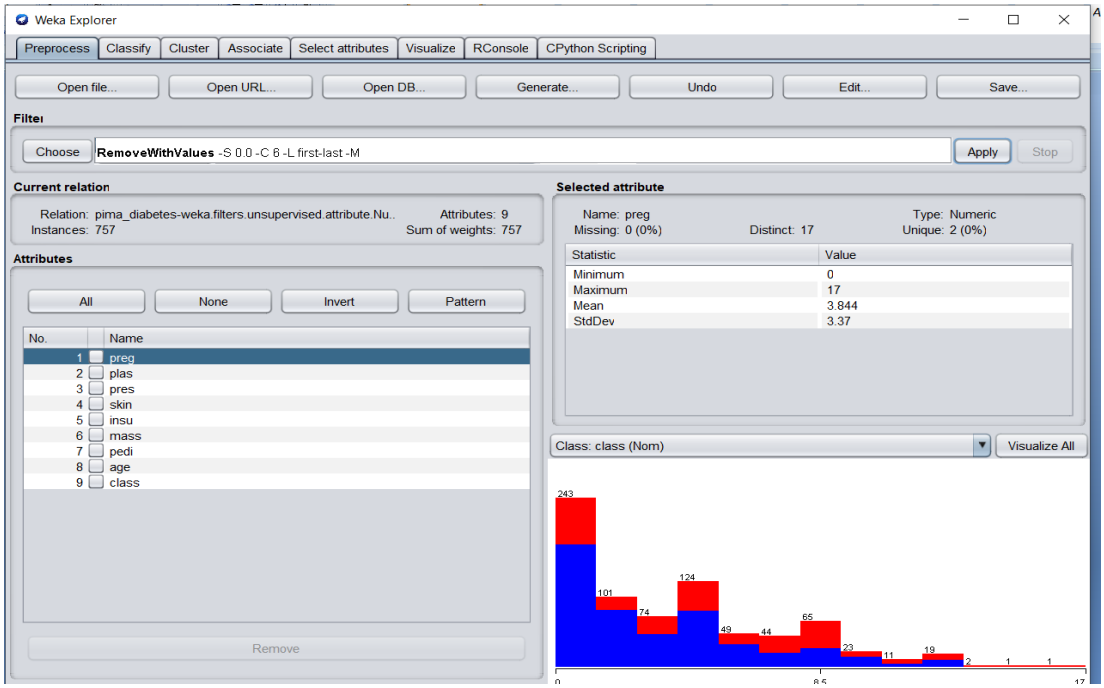
- Here you can see 11 attribute values that were formally set to 0 are now marked as Missing.



Remove Missing data

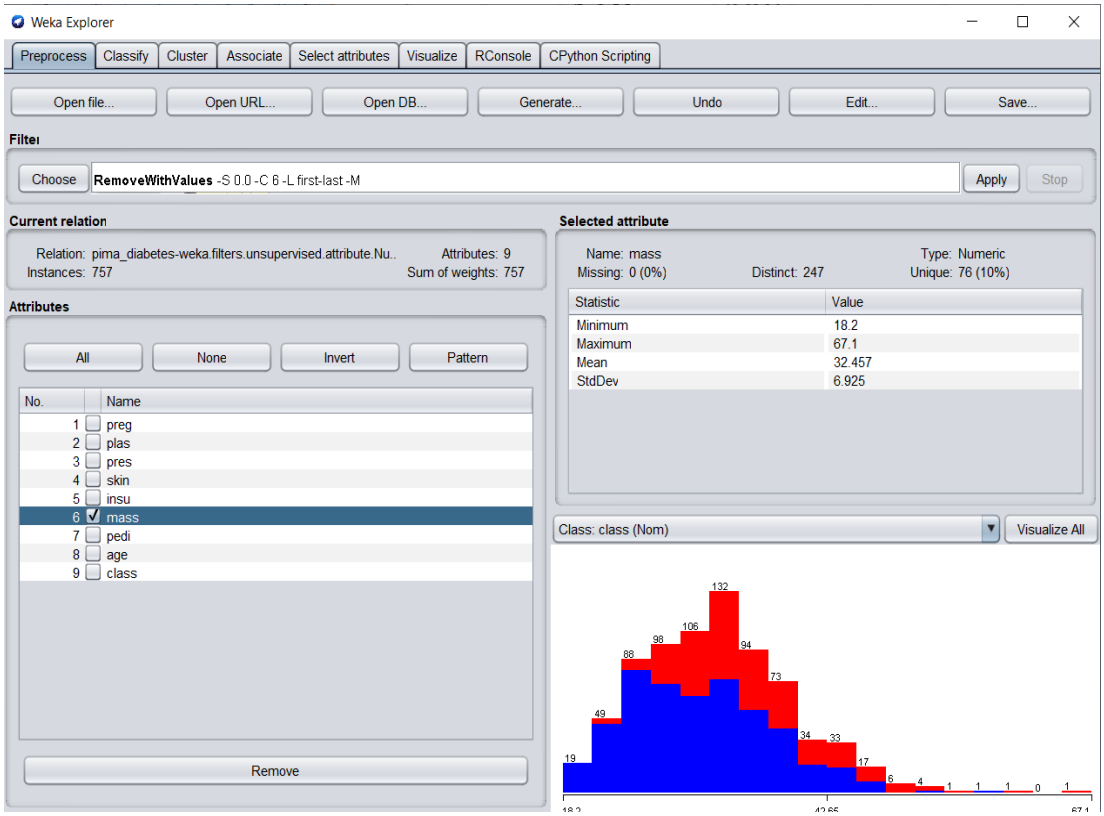
Applying filter





Setting mass attribute

- Here you can see 11 attribute values that were marked Missing have been removed from the dataset.



Findings and Learning: In this experiment we have discovered how to handle missing data in machine learning dataset using WEKA.
We learnt:

- How to mark corrupt values as missing in your dataset.
- How to remove instances with missing values from your dataset.
- How to impute mean values for missing values in your dataset.