Data Mining

Report on WEKA Dataset

Contents:

Introduction:

Data Mining:

Data mining is the process of identifying patterns in massive data sets using techniques that combine machine learning, statistics, and database systems.

WEKA:

The Waikato Environment for Knowledge Analysis (WEKA), created at the University of Waikato in New Zealand, is free software distributed under the terms of the GNU General Public License. Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

Retrieving the Data:

- When we open the WEKA application the home page should look like the figure 1 below. We can find many operations in the Weka GUI Chooser we are going to use “explorer” option to process our data.

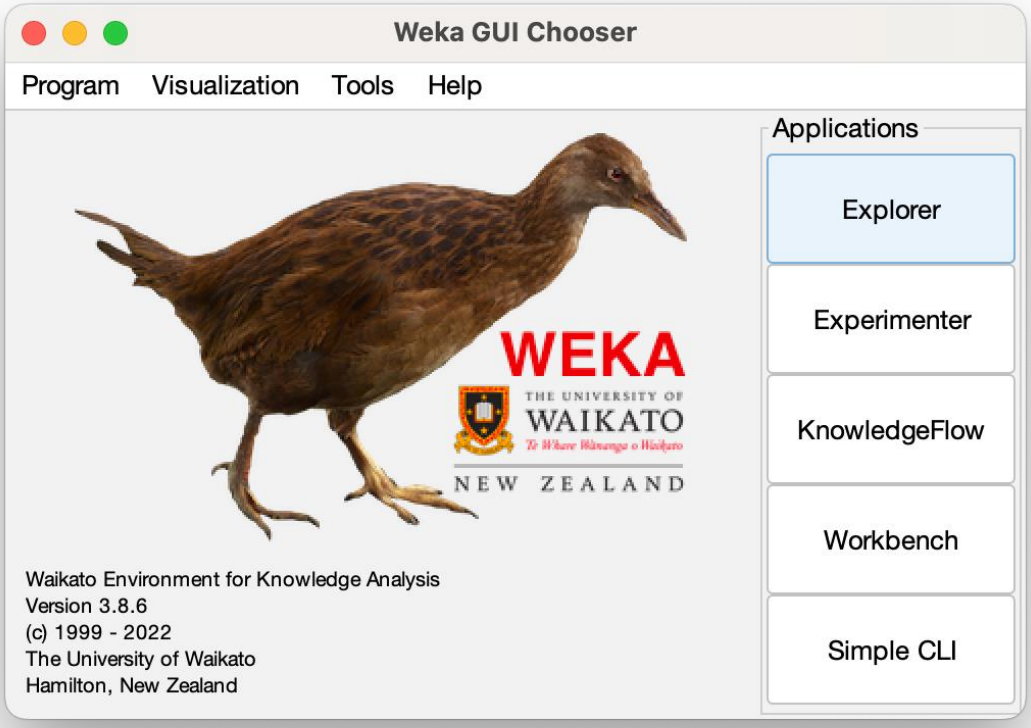


Figure 1 WEKA Homepage

- After we enter the explorer, we will find a normal java interfaced application then to load data we have many options like open file, open URL, open DB and generate. We already have our data saved in computer, so we choose open file and choose the csv file. It should look the figure 2 after data was loaded.

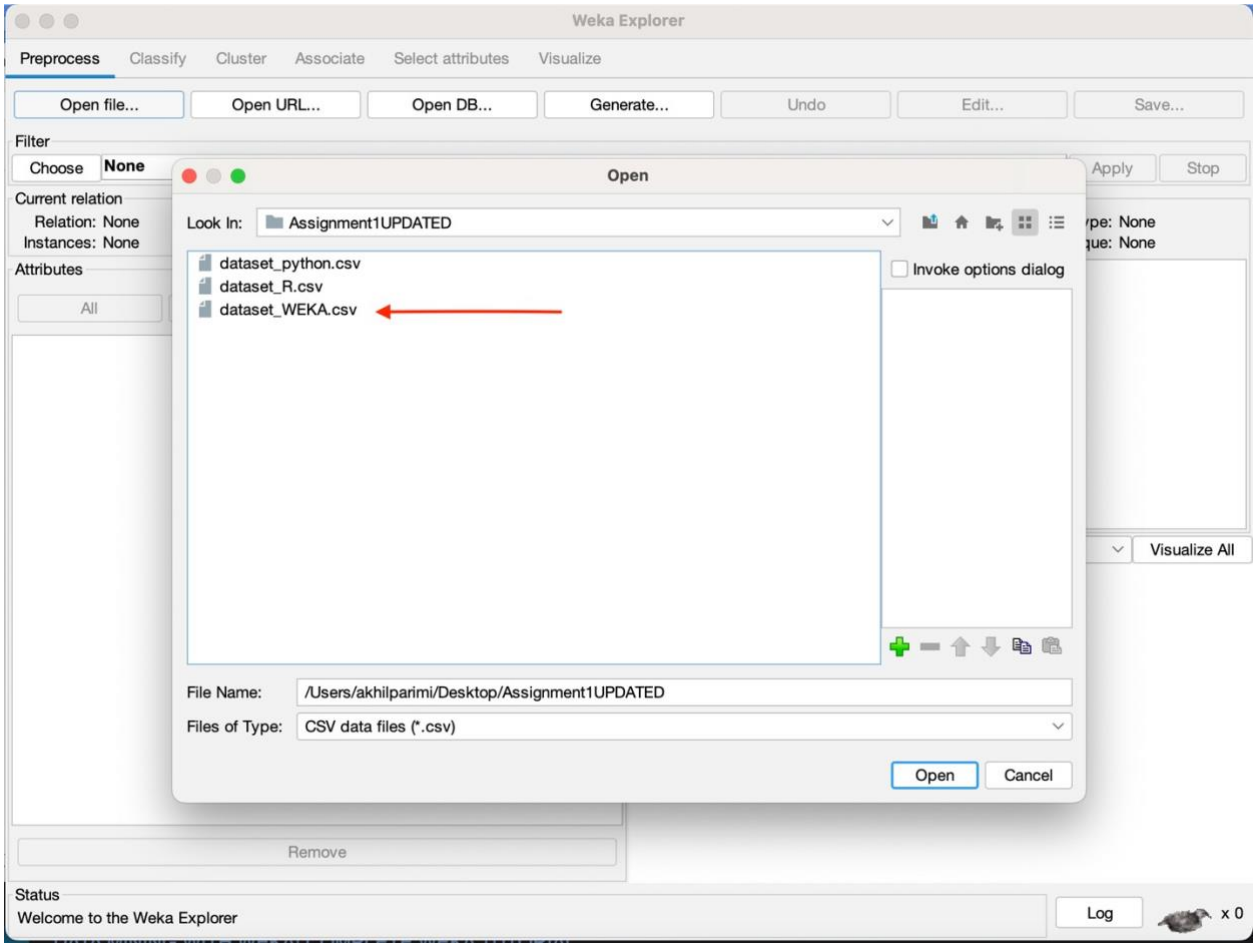


Figure 2 Loading the data

Glimpse of Data:

- The Figure 3 represents the processed file in Weka in which all the attributes are chosen. Here we can observe the name, type, number of missing values, unique values,

minimum value, maximum value, mean, standard deviation of the values and visualized bar graph of the attribute chosen which is bmi in this figure.

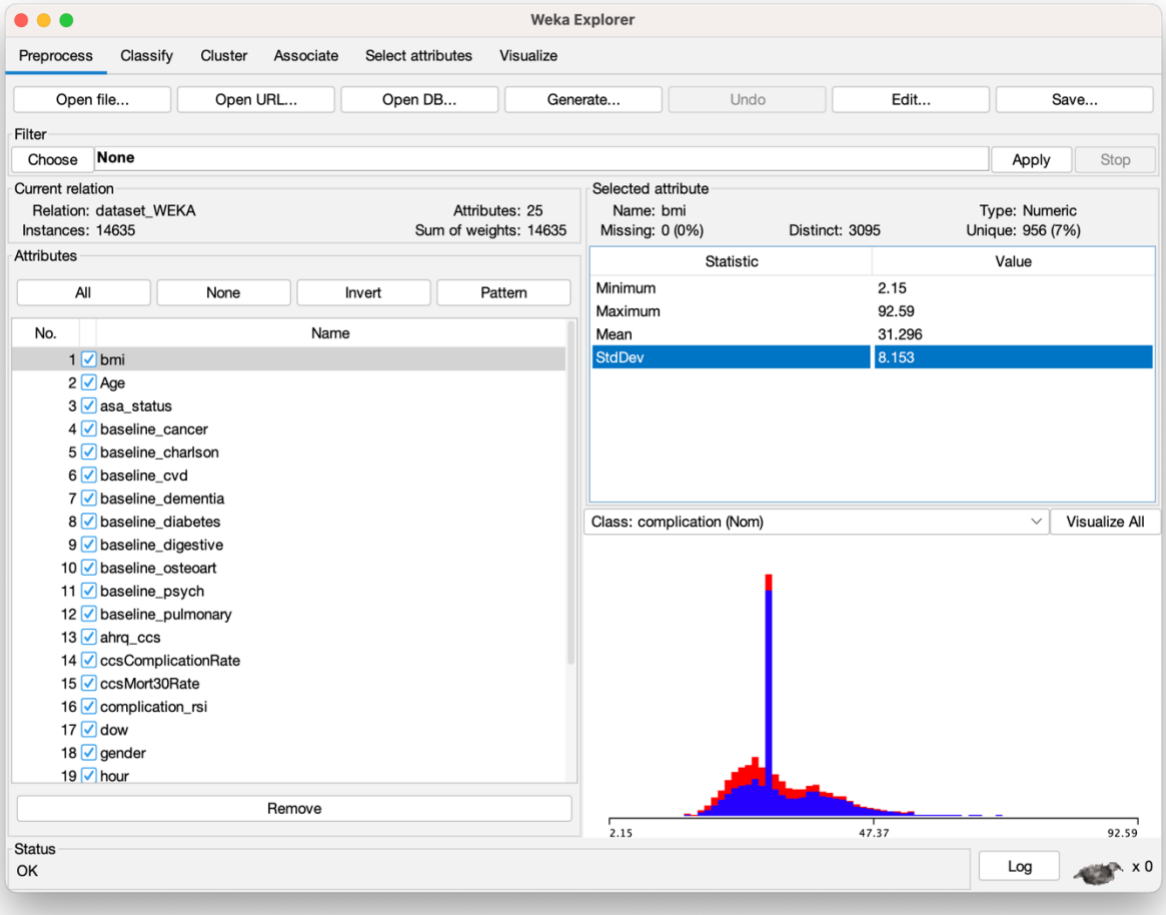


Figure 3 Processed file in WEKA.

- When we click on visualize all we will get a glimpse of all the attributes within one class visualized graph like from the figure 4 in which can choose any attribute we want.

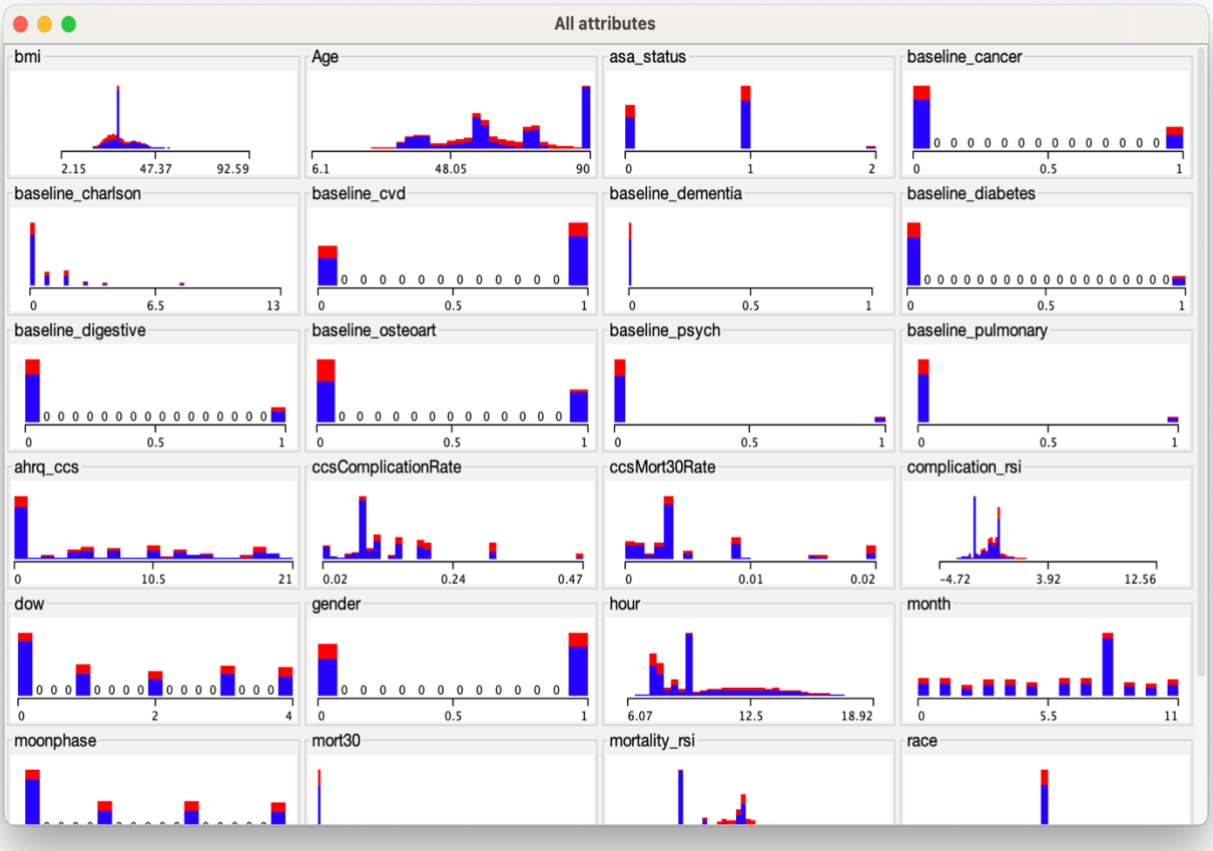


Figure 4 Glimpse of all attributes of one class

- There is a visualize option which is indicated with red arrow in the tool bar, when we click on that option a plot matrix of all the attributes will show as in the figure 5. In this we can

adjust plot size, pointer size and jitter as we like them to be, and we can find them below the plot matrix.

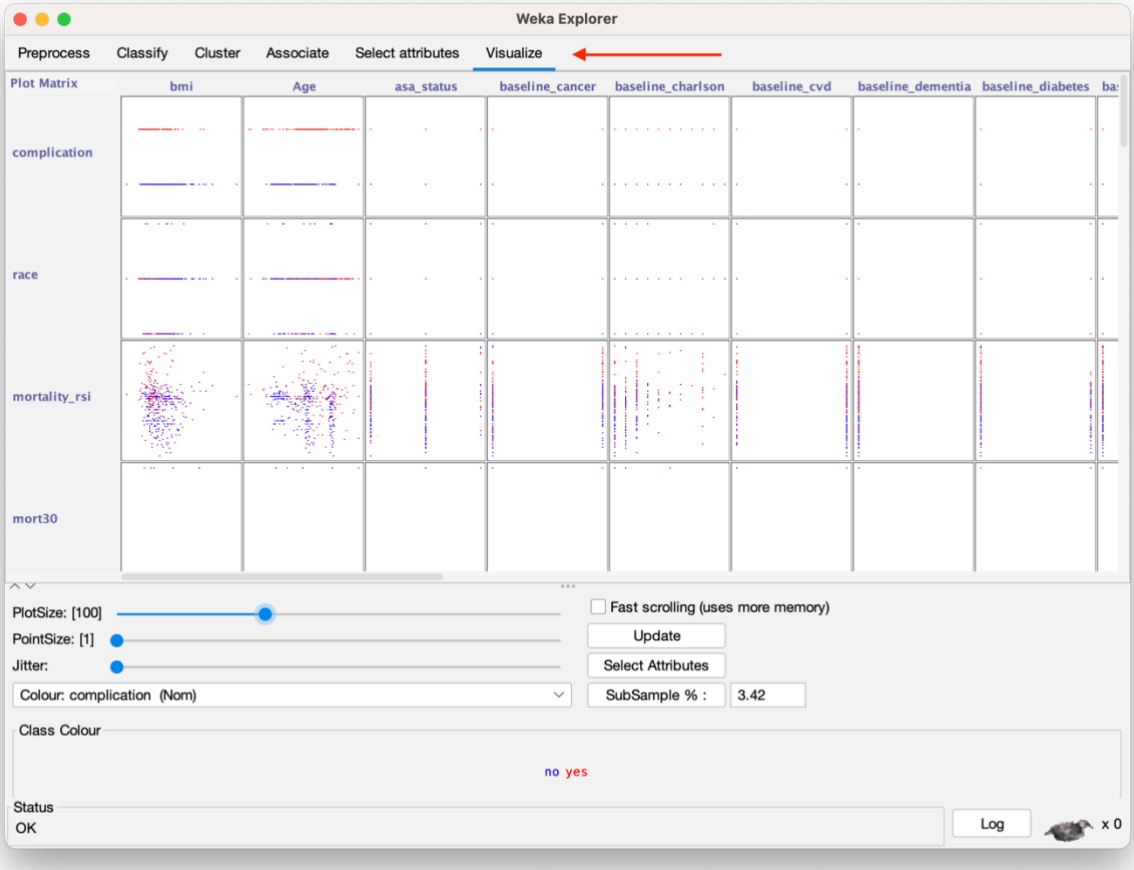


Figure 5 Plot matrix of all attributes

Checking for Missing Data:

- In the dataset for the Weka (Dataset_WEKA.csv) there is no missing data it can also be seen in the Weka as from the figure 6. We can individually see all the attributes unique values and missing data there. In this case we choose the attribute “morality_rsi”.

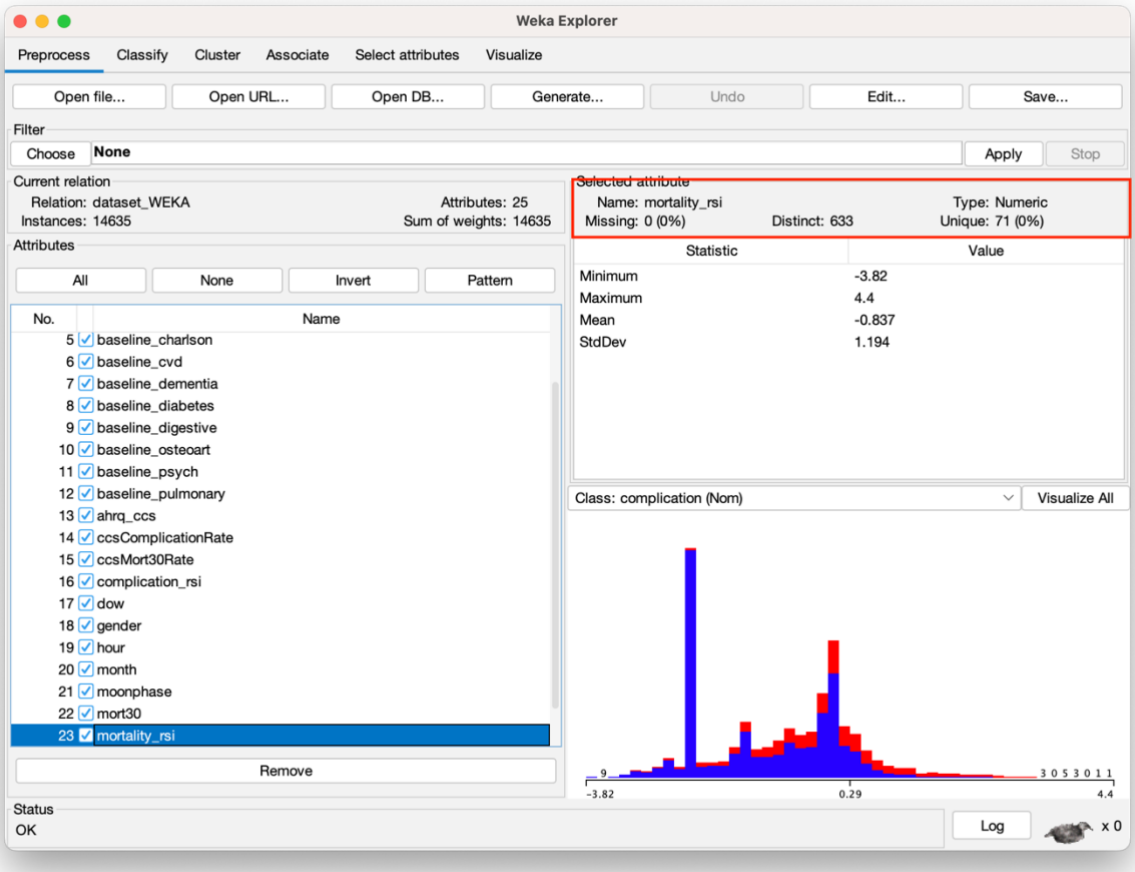


Figure 6 Missing values

Data Exploration:

- In the given dataset there is so much data i.e., we have 26 columns and 14636 rows. Here are some of the random plot graphs of the given dataset.
- In the figure 7, the x-axis represents bmi and the y-axis represents complication. Here we can adjust the jitter and change the colors of the pointers and select different instances as well.

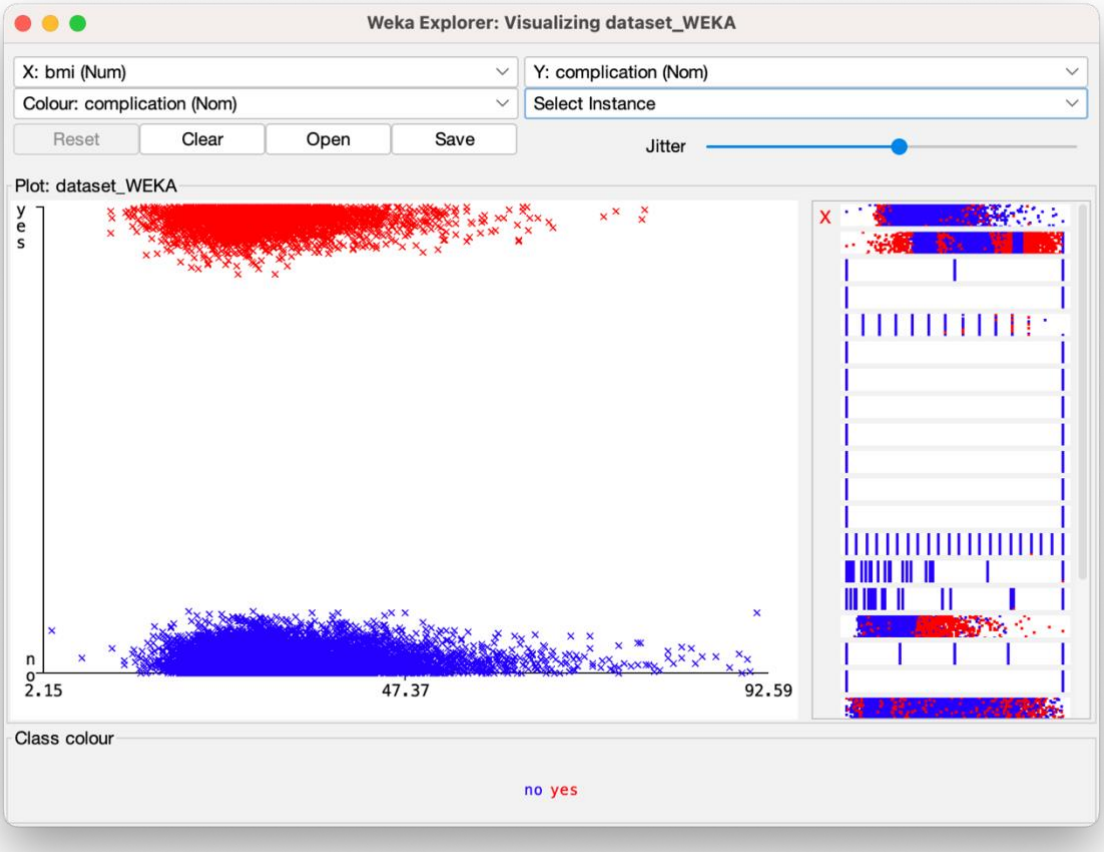


Figure 7

- In figure 8, the x-axis represents Age, the y-axis represents mortality_rsi and we can also see the type of attribute it is in the bracket beside the attribute names.

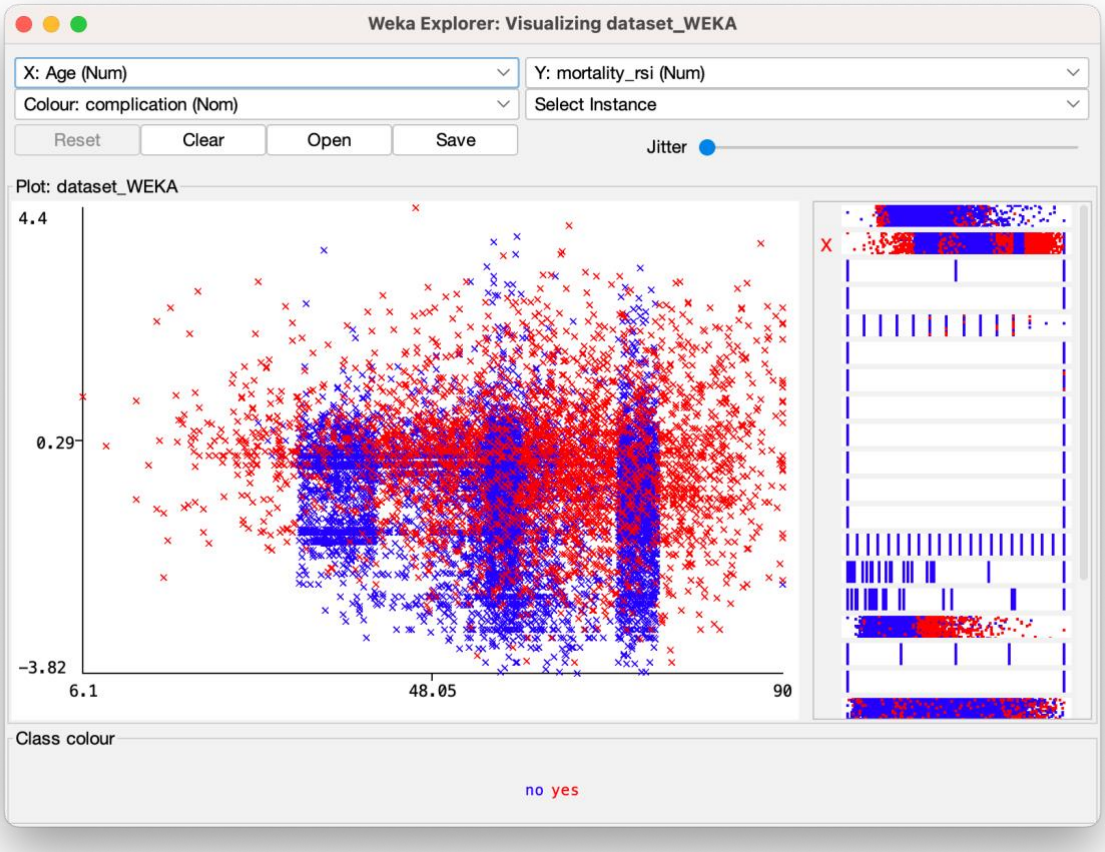
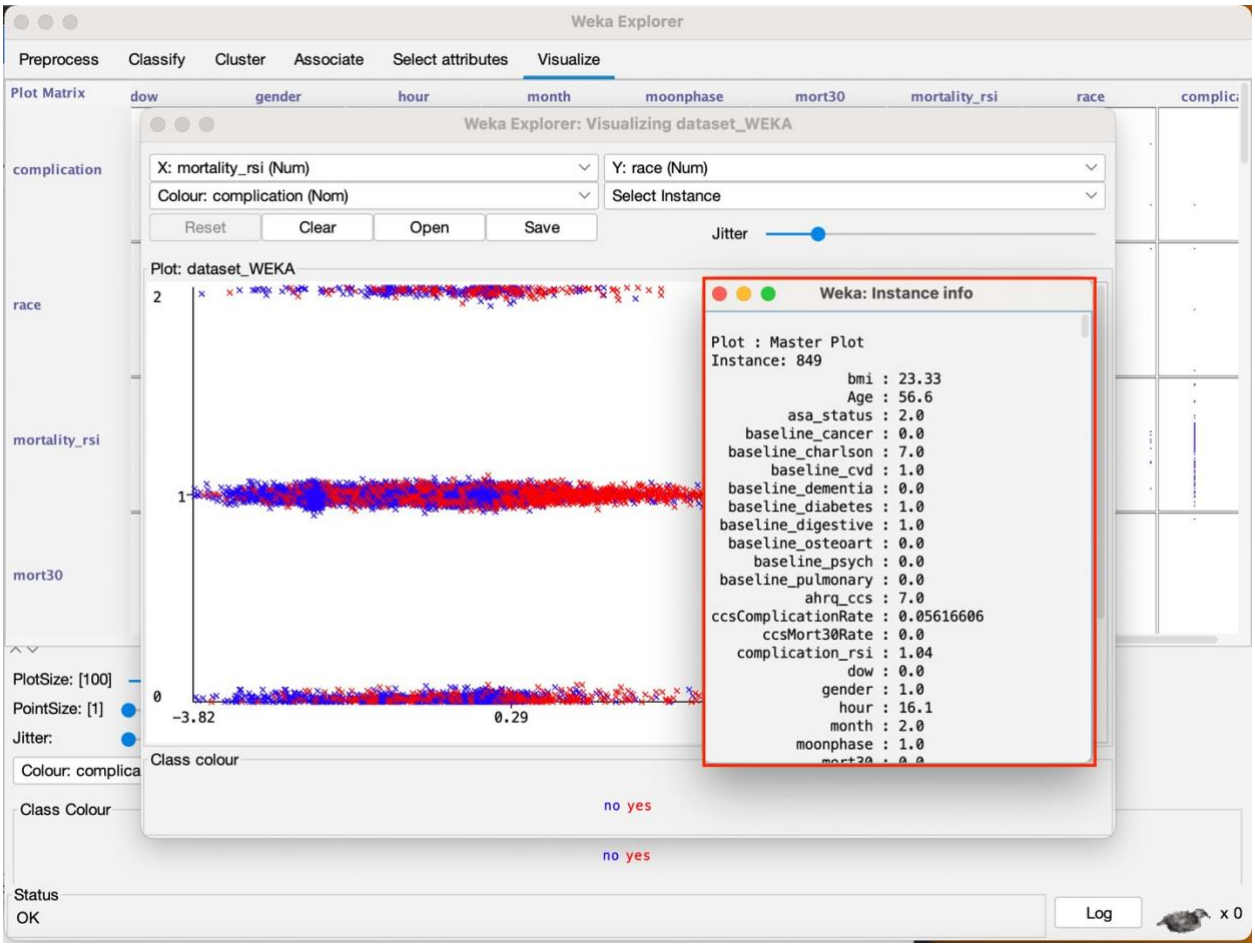


Figure 8

•

In the figure 9, x-axis represents mortality_rsi and y-axis represents race. If we click on any pointer (pointed as x in graph), we will get the information about that instance as shown in the figure.

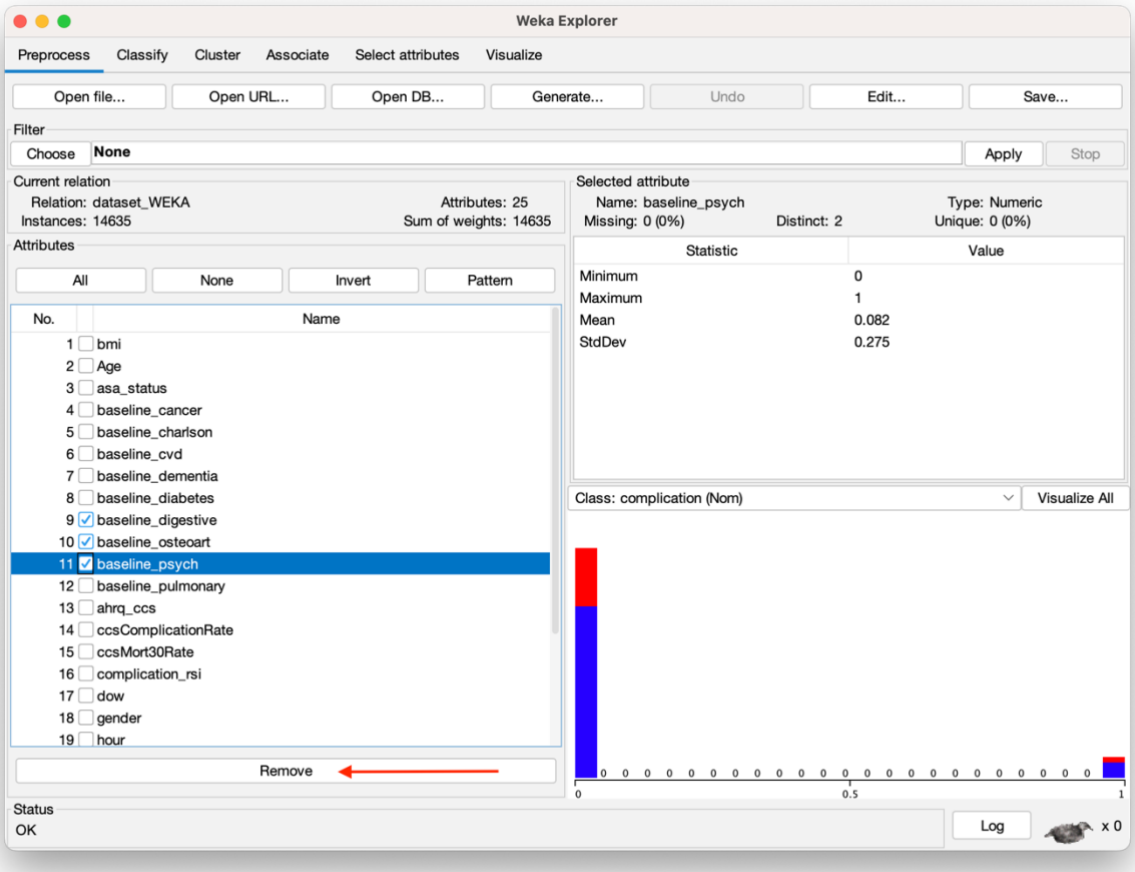


-

Figure 9

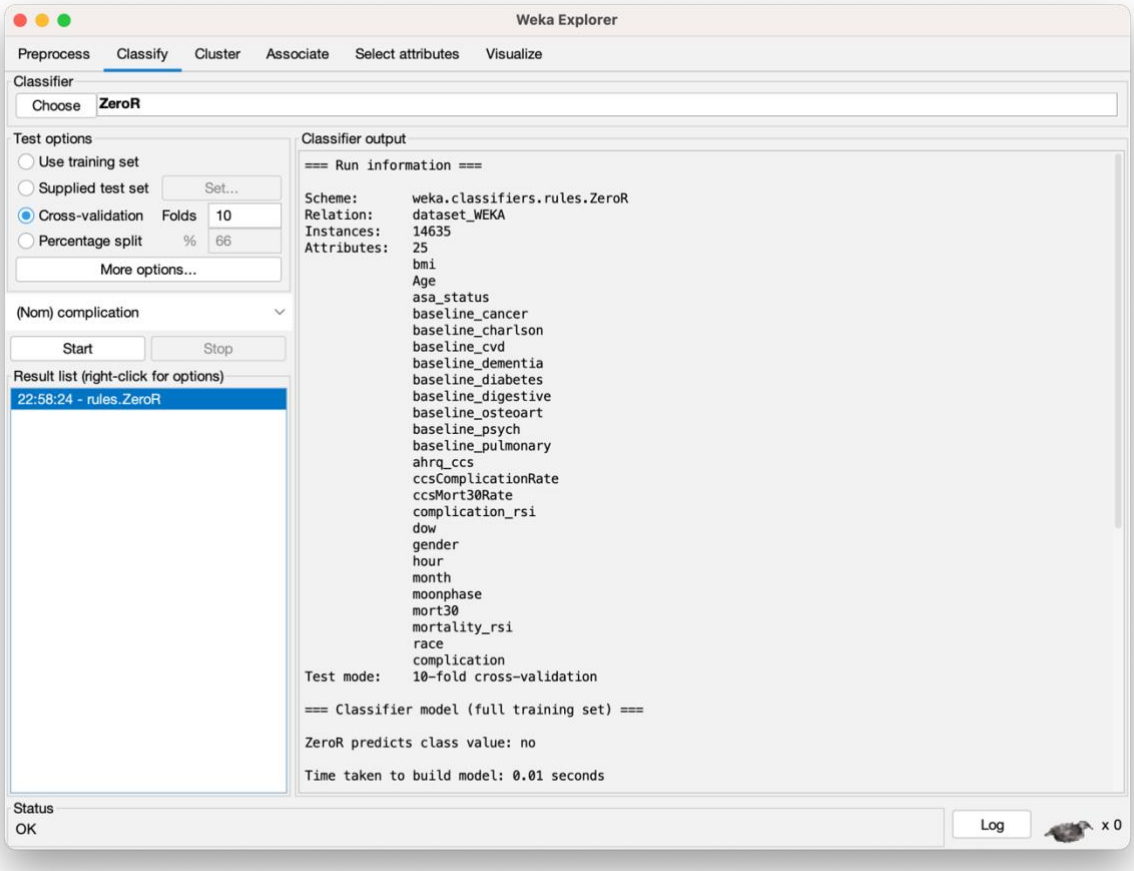
We can remove attributes we don’t need in the preprocess page by selecting the attributes and clicking the remove option shown in the figure below pointed by a red arrow.

•



We can also find the classifier output by going to classify tool and click start option marked the figure below. We will have run information, classifier model, stratified crossvalidation and summary.

-


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) complication
Start
Stop
Result list (right-click for options)
22:58:24 - rules.ZeroR
Classifier output
=== Run information ===
Scheme: weka.classifiers.rules.ZeroR
Relation: dataset_WEKA
Instances: 14635
Attributes: 25
bmi
Age
asa_status
baseline_cancer
baseline_charlson
baseline_cvd
baseline_dementia
baseline_diabetes
baseline_digestive
baseline_osteoart
baseline_psych
baseline_pulmonary
ahrq_ccs
ccsComplicationRate
ccsMort30Rate
complication_rsi
dow
gender
hour
month
moonphase
mort30
mortality_rsi
race
complication
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
ZeroR predicts class value: no
Time taken to build model: 0.01 seconds
Status
OK
Log
x 0