Data Mining

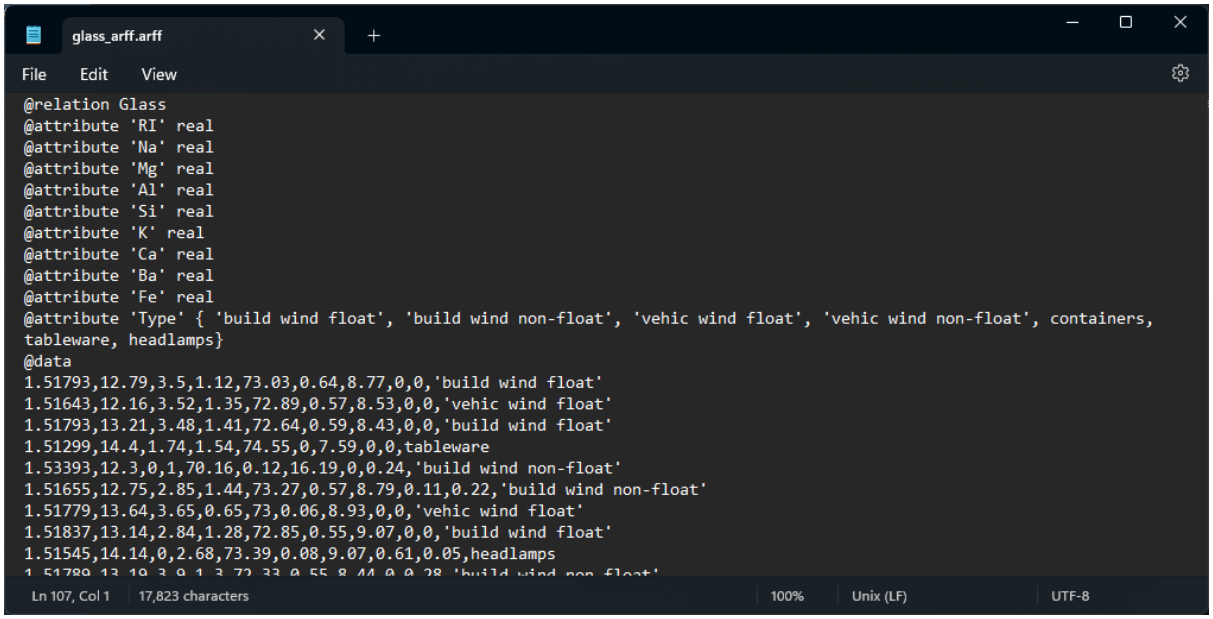
Assignment 3

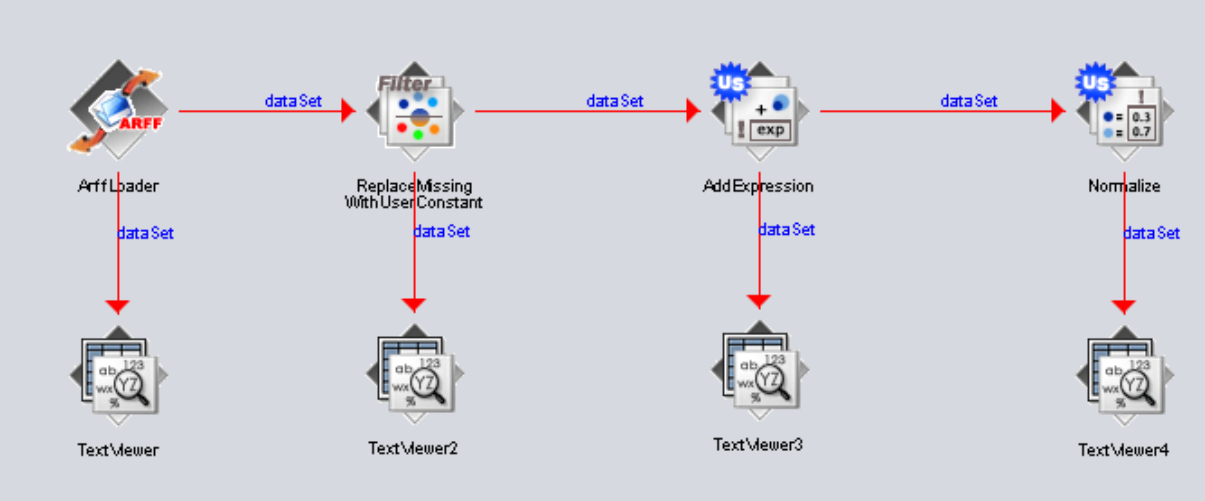
Name: Ayush Kumar

Reg. no.: 20214284

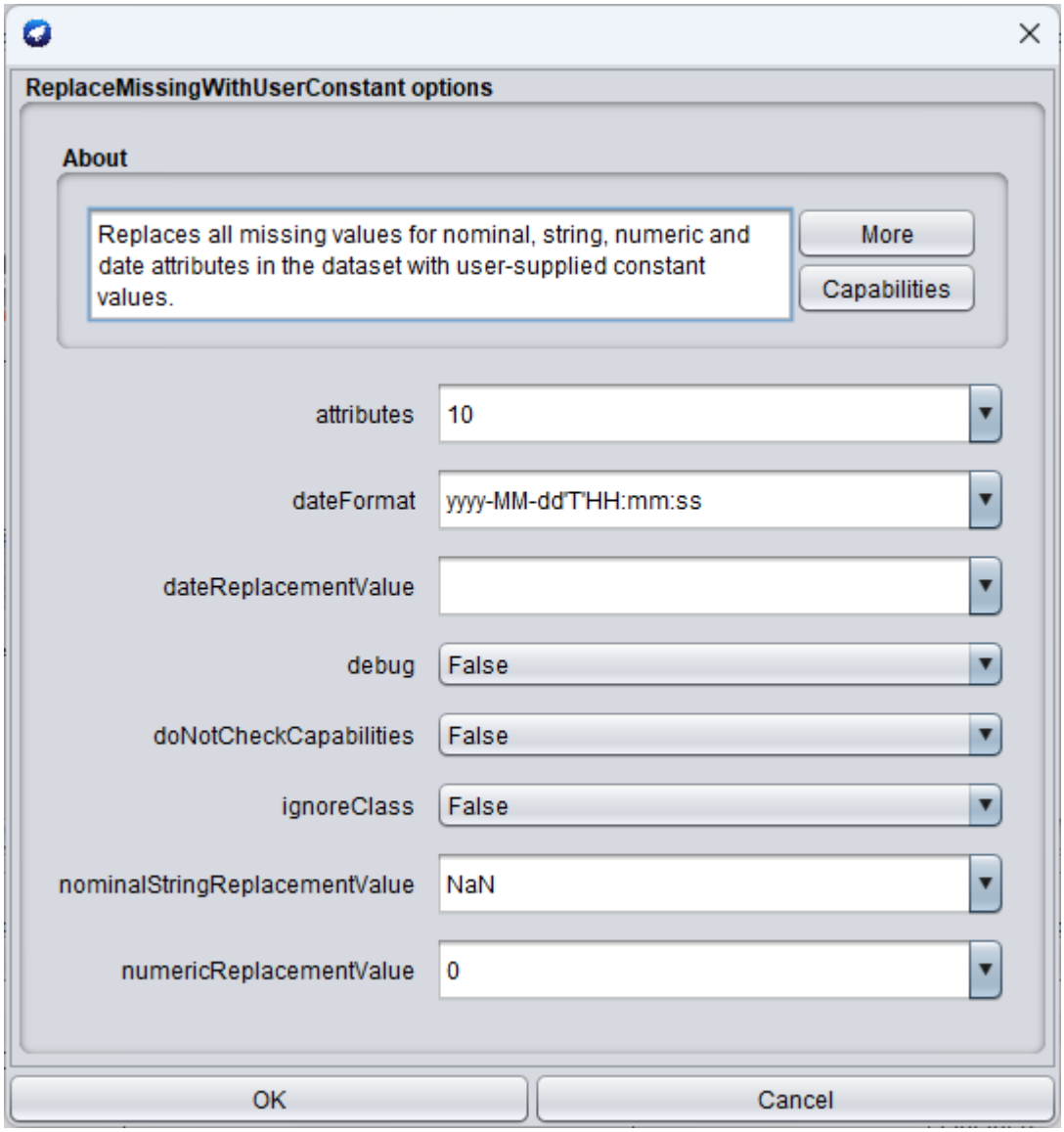
Group: CSE-6B1

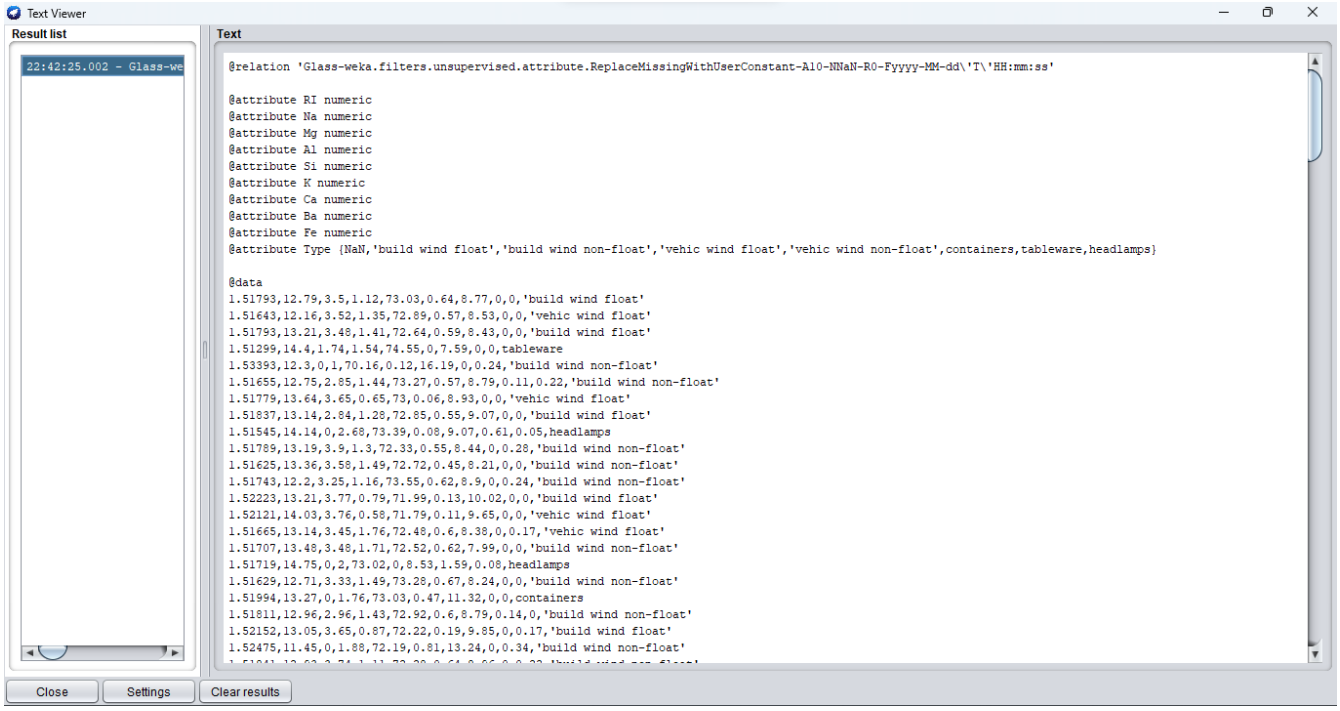
Q1.Use of filters in Weka Knowledge Flow: Use Glass dataset and do the following operations using appropriate filters.



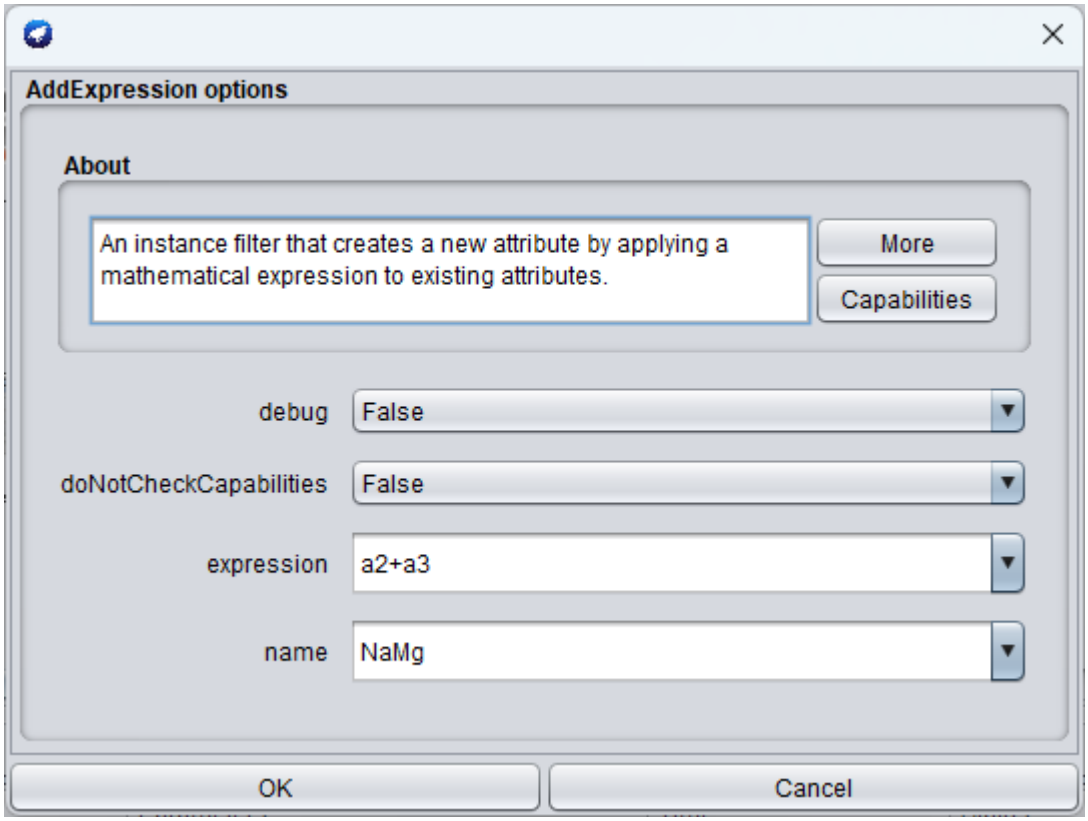


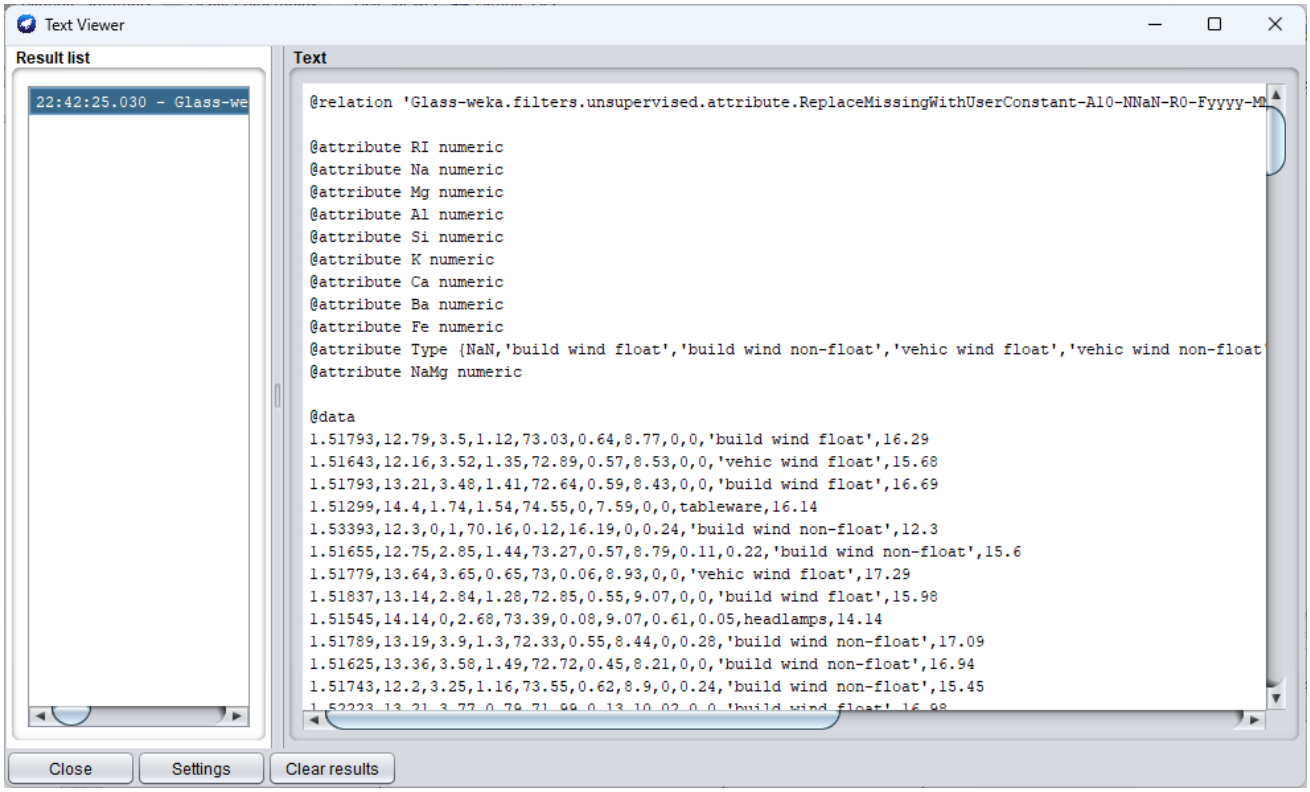
a. Mark missing values with name NaN.


ReplaceMissingWithUserConstant options
About
Replaces all missing values for nominal, string, numeric and date attributes in the dataset with user-supplied constant values.
More
Capabilities
attributes
10
dateFormat
yyyy-MM-dd'T'HH:mm:ss
dateReplacementValue
debug
False
doNotCheckCapabilities
False
ignoreClass
False
nominalStringReplacementValue
NaN
numericReplacementValue
0
OK
Cancel

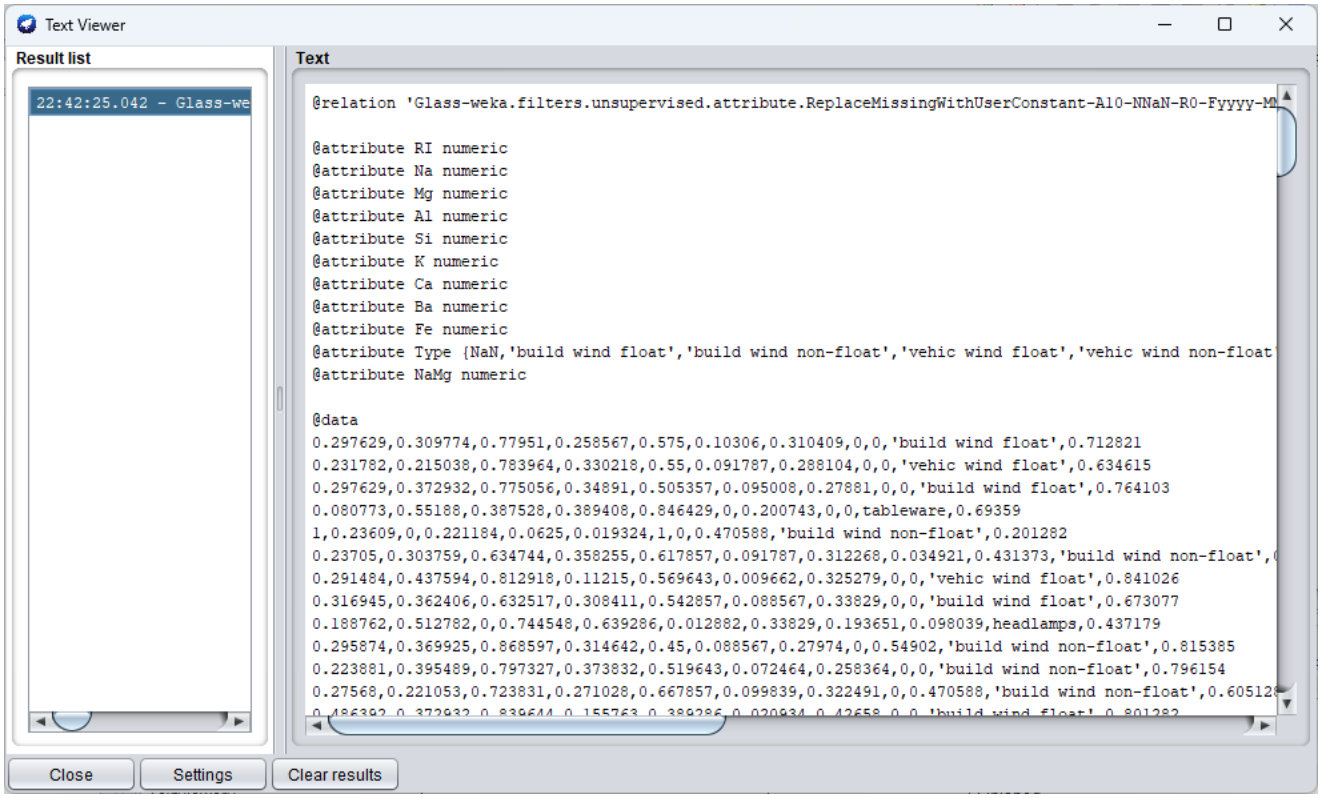


b. Marge any two similar numeric types attribute into new attributes.

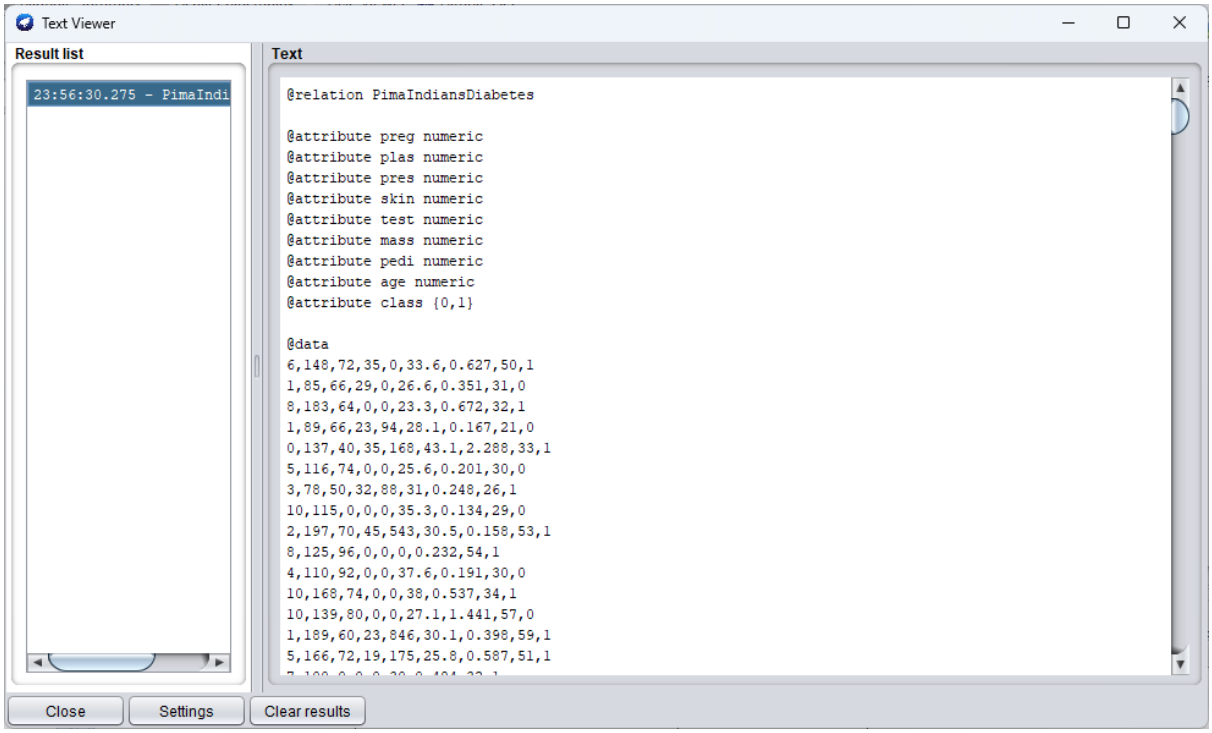


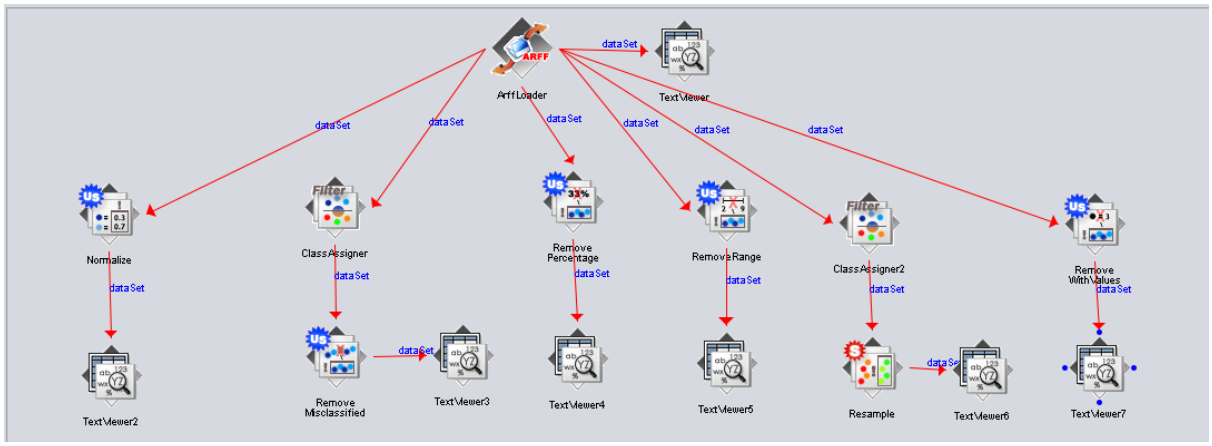


c. Select single /multi numeric attributes and convert it into numerical range between (Min-Max) ranges by formula given below: Note: normalization on other such as z-score, linear etc.

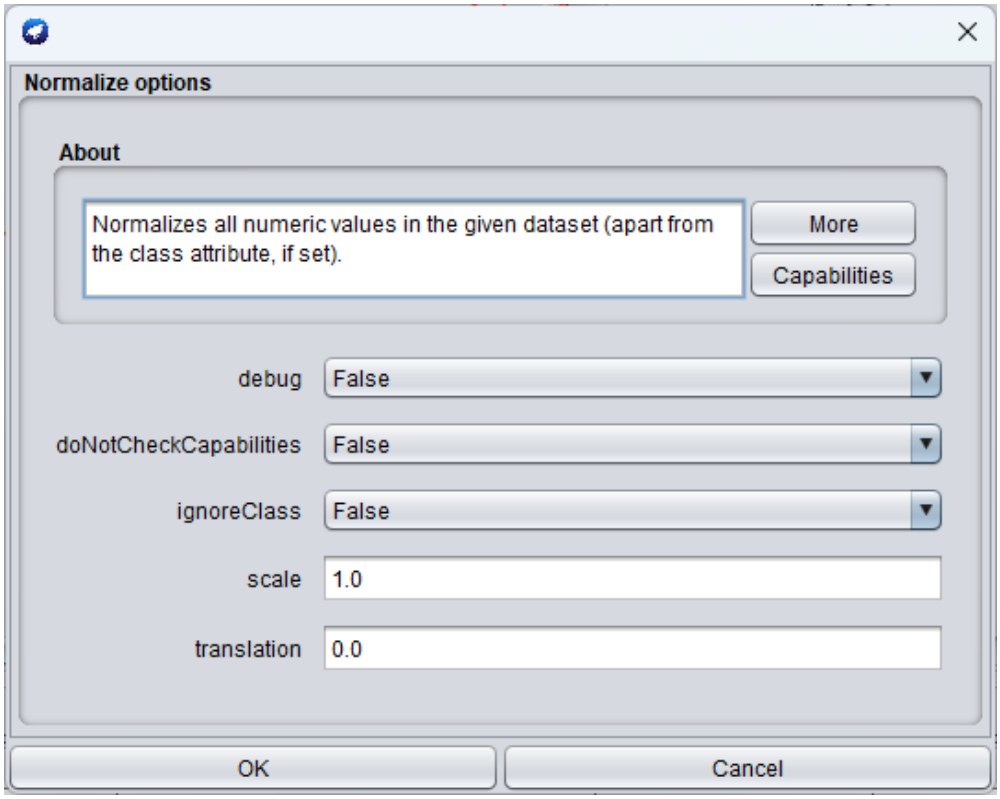


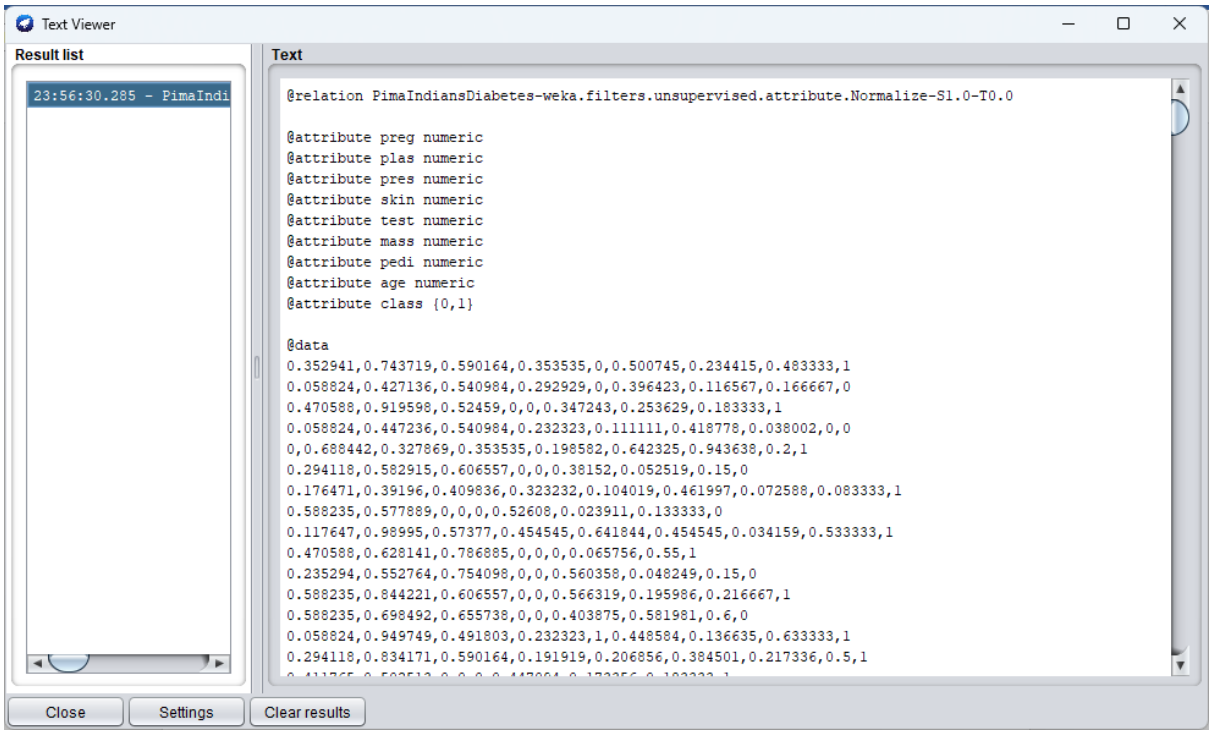
Q2. Download (https://www.kaggle.com/datasets/kumargh/pimaindiansdiabetescsv Pima Indian Diabetes Dataset (PIDD) and do following operations and save all reports in separate file using data sink .csv format.



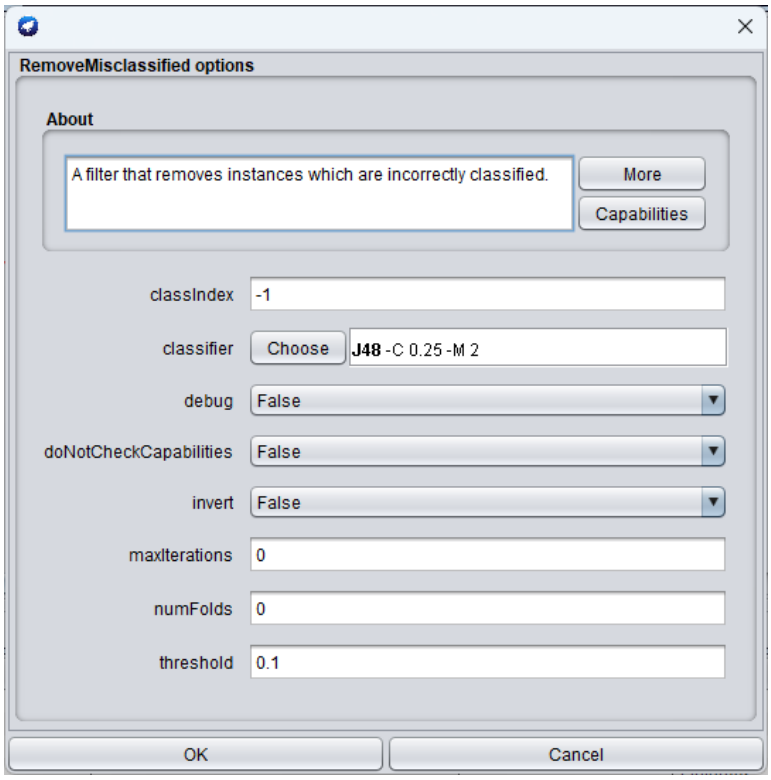


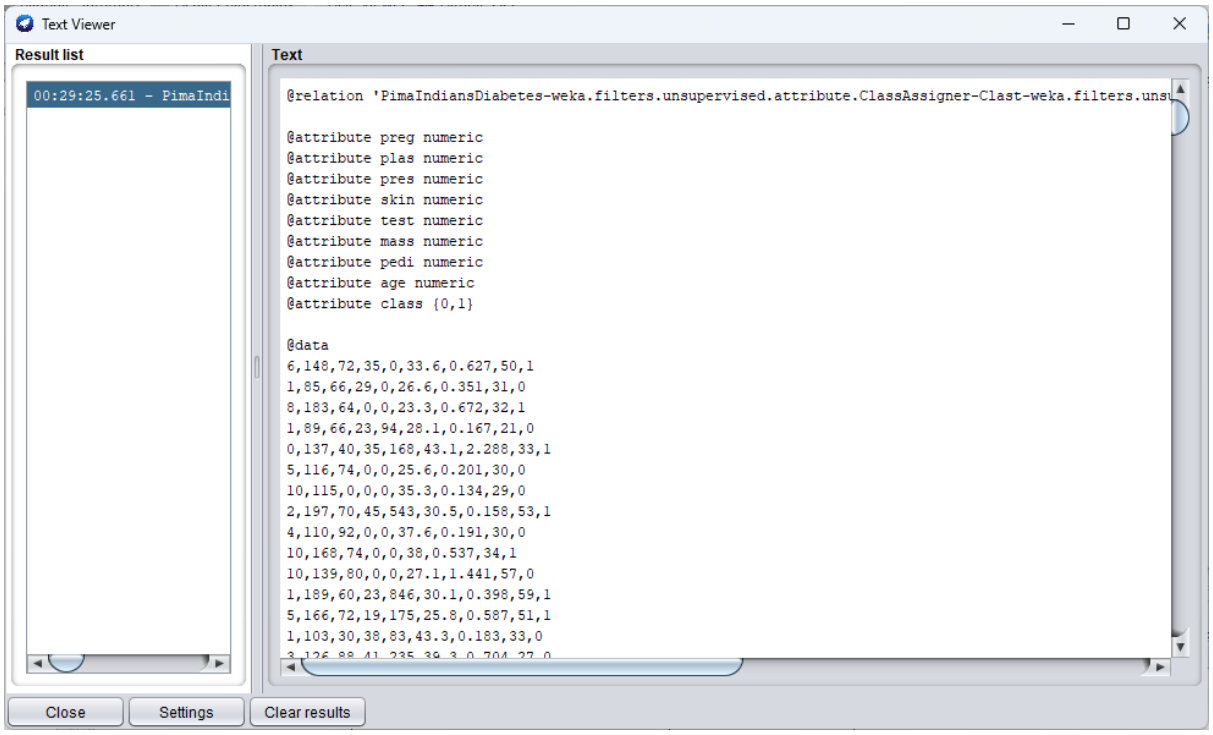
a. Normalize all numeric attributes


Normalize options
About
Normalizes all numeric values in the given dataset (apart from the class attribute, if set).
More
Capabilities
debug
False
doNotCheckCapabilities
False
ignoreClass
False
scale
1.0
translation
0.0
OK
Cancel

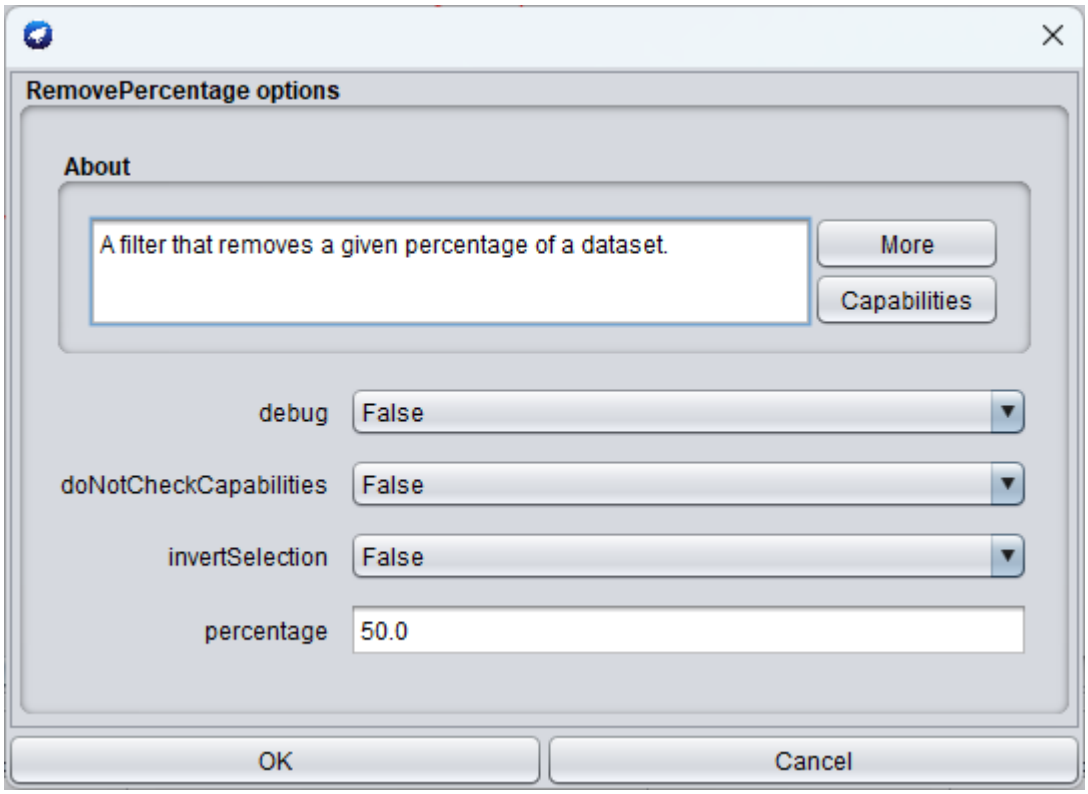

Text Viewer
Result list
23:56:30.285 - PimaIndi
Text
@relation PimaIndiansDiabetes-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
@attribute preg numeric
@attribute plas numeric
@attribute pres numeric
@attribute skin numeric
@attribute test numeric
@attribute mass numeric
@attribute pedi numeric
@attribute age numeric
@attribute class {0,1}
@data
0.352941,0.743719,0.590164,0.353535,0,0.500745,0.234415,0.483333,1
0.058824,0.427136,0.540984,0.292929,0,0.396423,0.116567,0.166667,0
0.470588,0.919598,0.52459,0,0,0.347243,0.253629,0.183333,1
0.058824,0.447236,0.540984,0.232323,0.111111,0.418778,0.038002,0,0
0,0.688442,0.327869,0.353535,0.198582,0.642325,0.943638,0.2,1
0.294118,0.582915,0.606557,0,0,0.38152,0.052519,0.15,0
0.176471,0.39196,0.409836,0.323232,0.104019,0.461997,0.072588,0.083333,1
0.588235,0.577889,0,0,0,0.52608,0.023911,0.133333,0
0.117647,0.98995,0.57377,0.454545,0.641844,0.454545,0.034159,0.533333,1
0.470588,0.628141,0.786885,0,0,0.065756,0.55,1
0.235294,0.552764,0.754098,0,0,0.560358,0.048249,0.15,0
0.588235,0.844221,0.606557,0,0,0.566319,0.195986,0.216667,1
0.588235,0.698492,0.655738,0,0,0.403875,0.581981,0.6,0
0.058824,0.949749,0.491803,0.232323,1,0.448584,0.136635,0.633333,1
0.294118,0.834171,0.590164,0.191919,0.206856,0.384501,0.217336,0.5,1
Close
Settings
Clear results

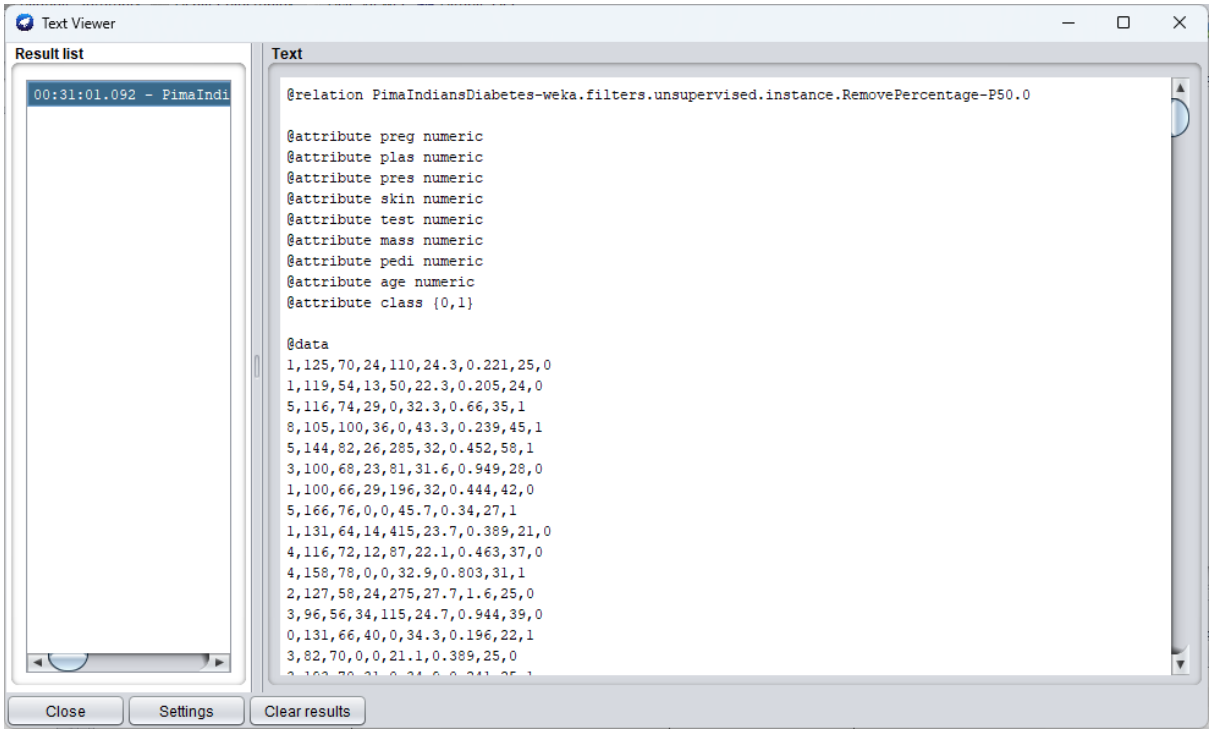
b. Remove all misclassified value by J48 class filter



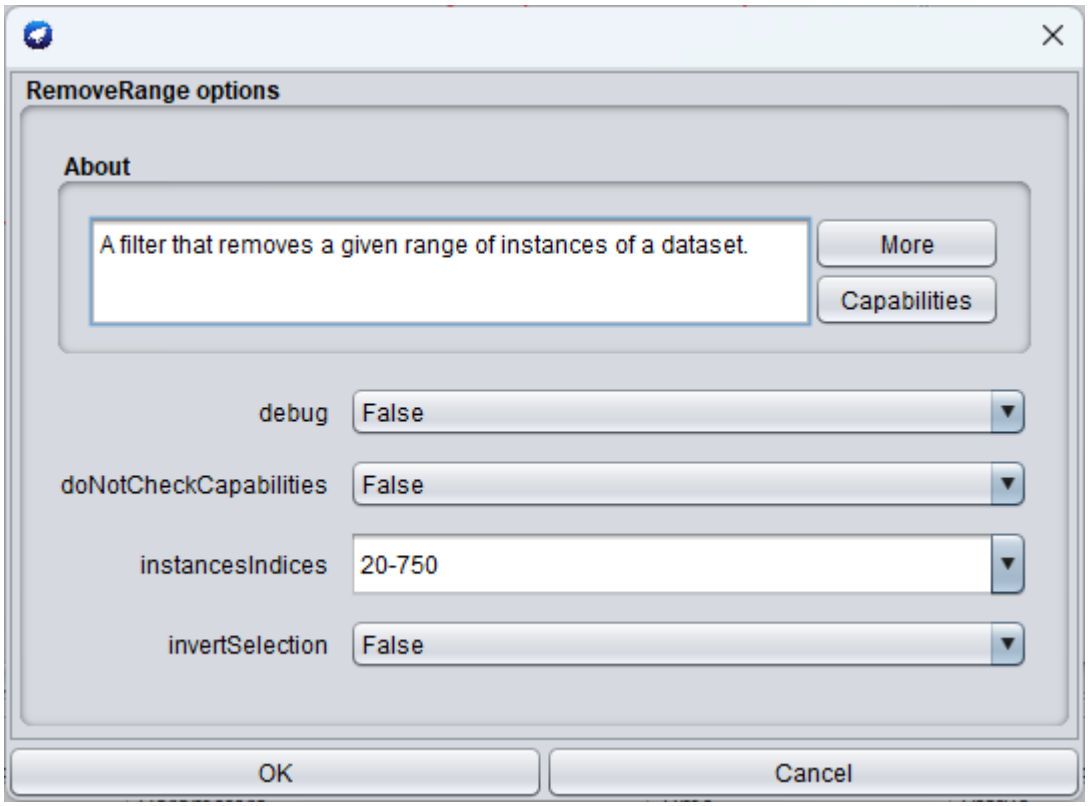


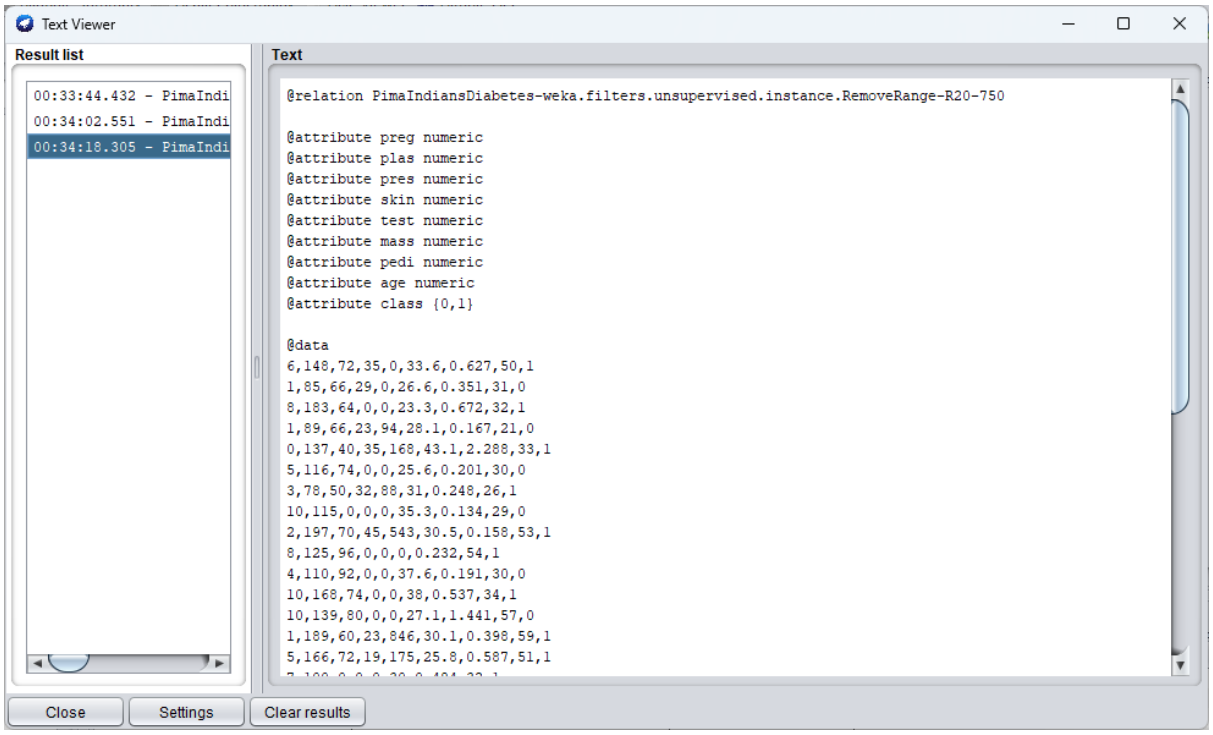
c. Remove Percent



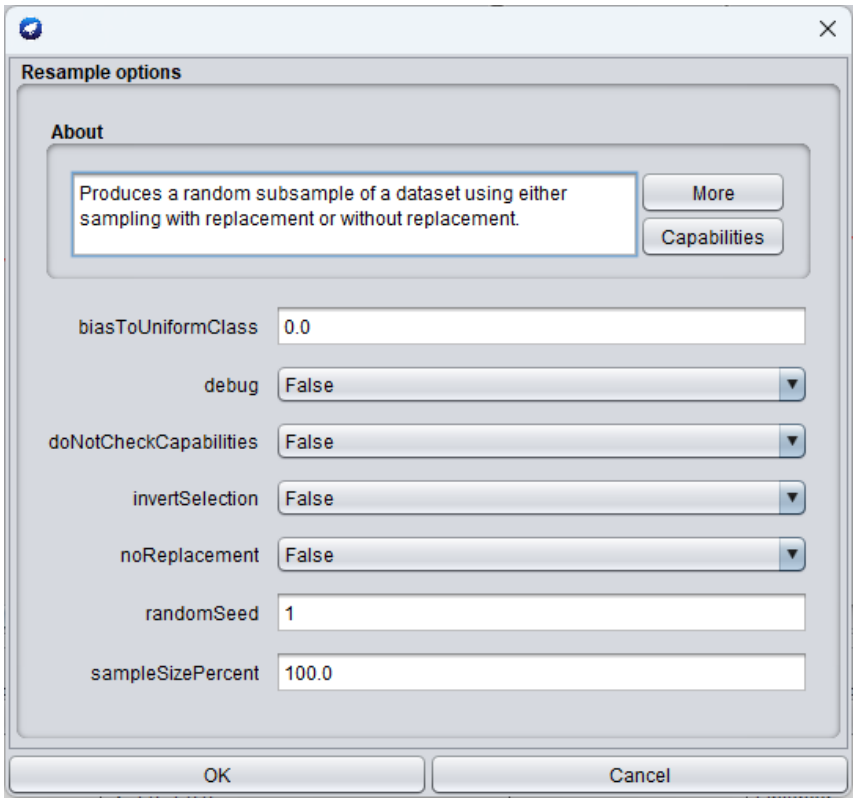


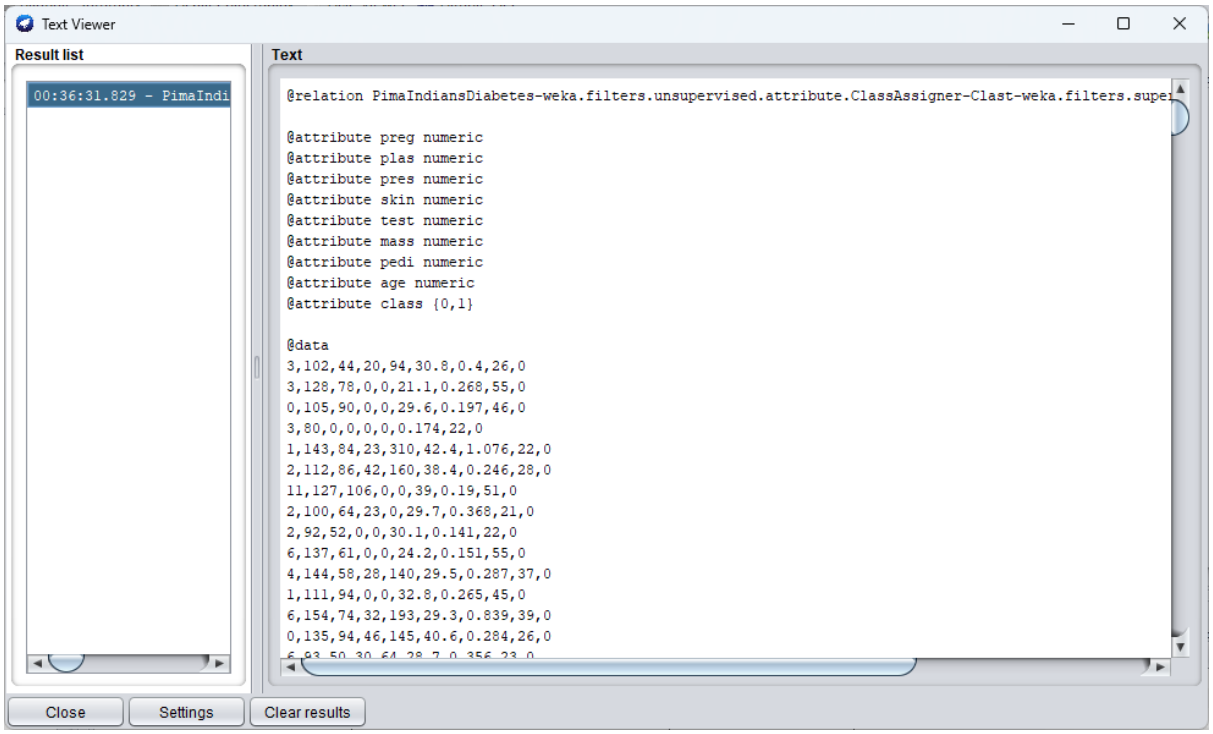
d. Remove Range





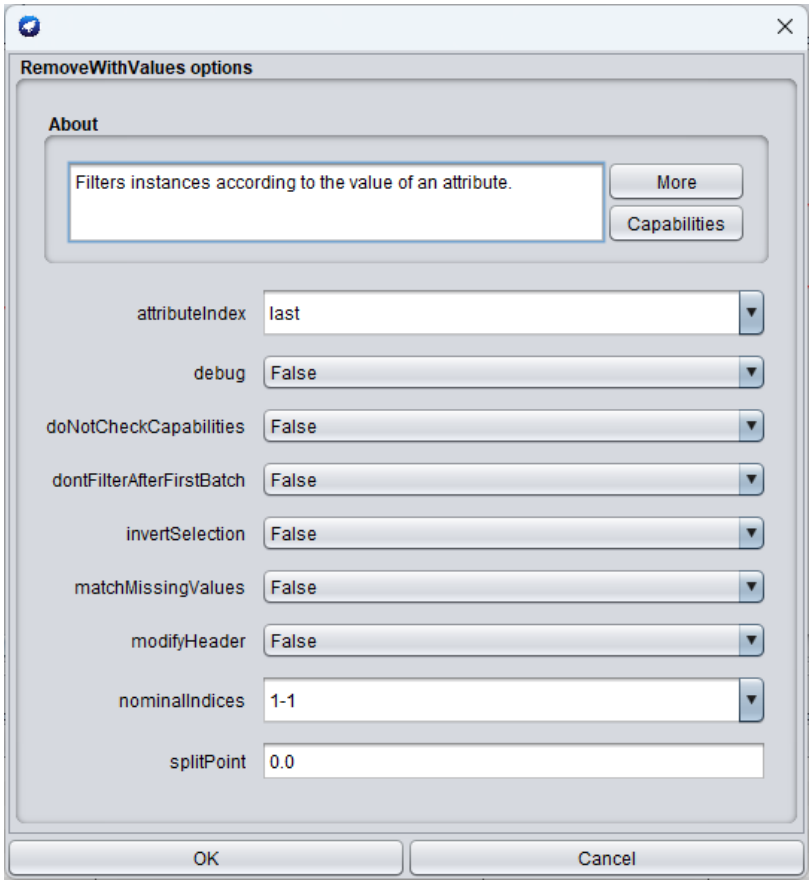
e. Resample

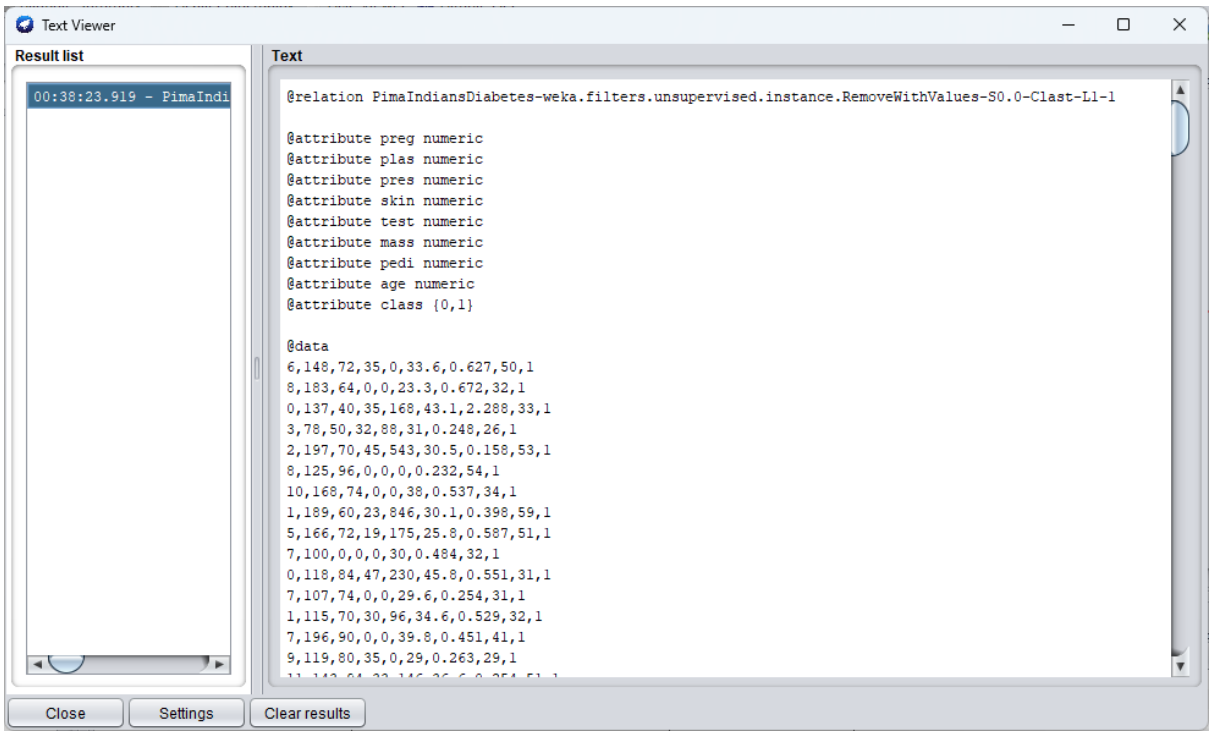

Resample options
About
Produces a random subsample of a dataset using either sampling with replacement or without replacement.
More
Capabilities
biasToUniformClass
0.0
debug
False
doNotCheckCapabilities
False
invertSelection
False
noReplacement
False
randomSeed
1
sampleSizePercent
100.0
OK
Cancel


Text Viewer
Result list
00:36:31.829 - PimaIndi
Text
@relation PimaIndiansDiabetes-weka.filters.unsupervised.attribute.ClassAssigner-Clast-weka.filters.super
@attribute preg numeric
@attribute plas numeric
@attribute pres numeric
@attribute skin numeric
@attribute test numeric
@attribute mass numeric
@attribute pedi numeric
@attribute age numeric
@attribute class {0,1}
@data
3,102,44,20,94,30.8,0.4,26,0
3,128,78,0,0,21.1,0.268,55,0
0,105,90,0,0,29.6,0.197,46,0
3,80,0,0,0,0,0.174,22,0
1,143,84,23,310,42.4,1.076,22,0
2,112,86,42,160,38.4,0.246,28,0
11,127,106,0,0,39,0.19,51,0
2,100,64,23,0,29.7,0.368,21,0
2,92,52,0,0,30.1,0.141,22,0
6,137,61,0,0,24.2,0.151,55,0
4,144,58,28,140,29.5,0.287,37,0
1,111,94,0,0,32.8,0.265,45,0
6,154,74,32,193,29.3,0.839,39,0
0,135,94,46,145,40.6,0.284,26,0
Close
Settings
Clear results

f. Remove With Values

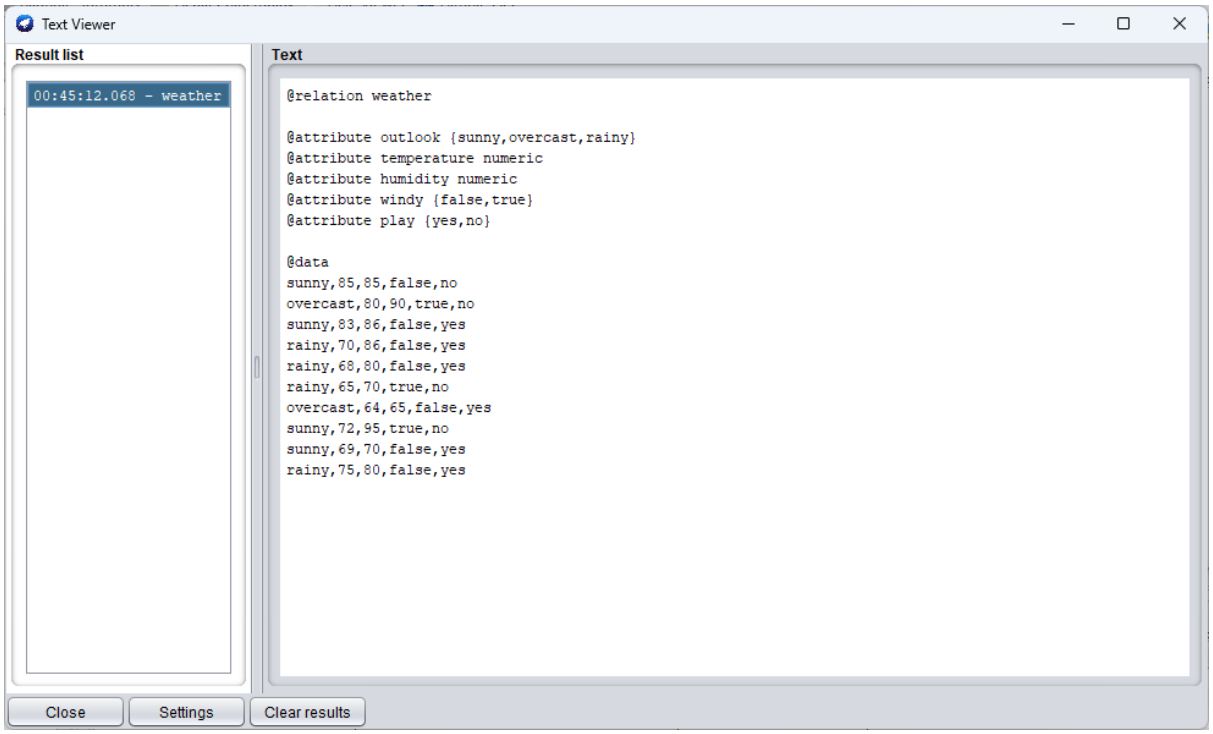
instances with class=0 removed



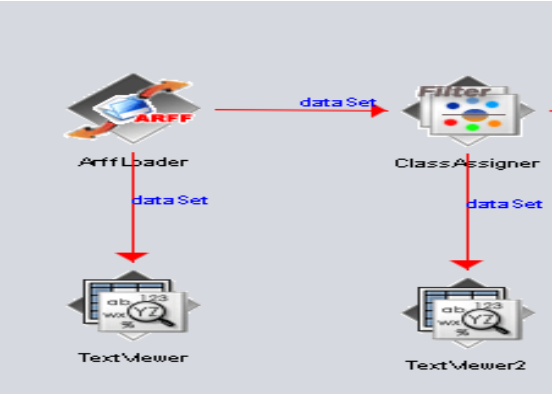


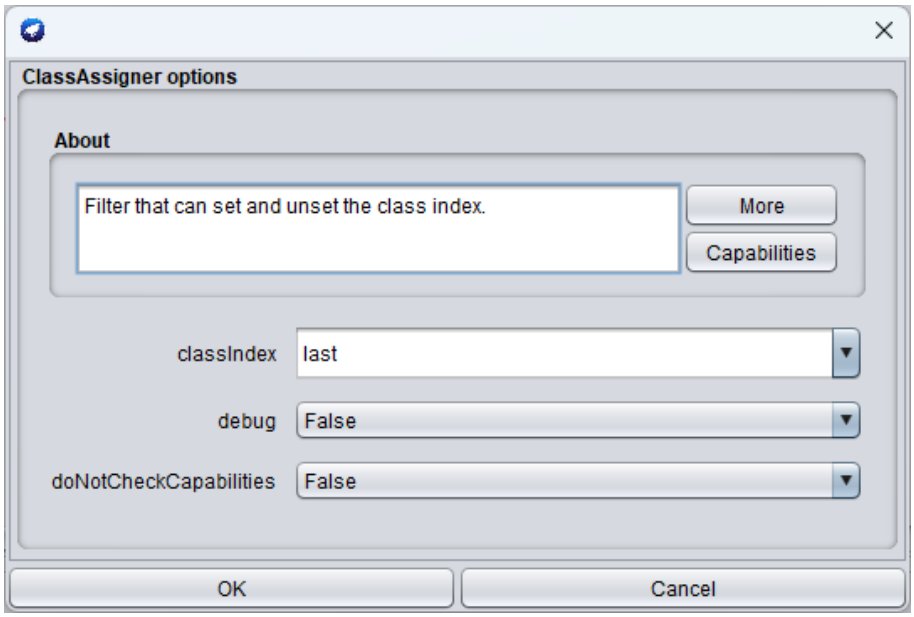
Q3.Training -test layout maker using Weka Knowledge Flow:

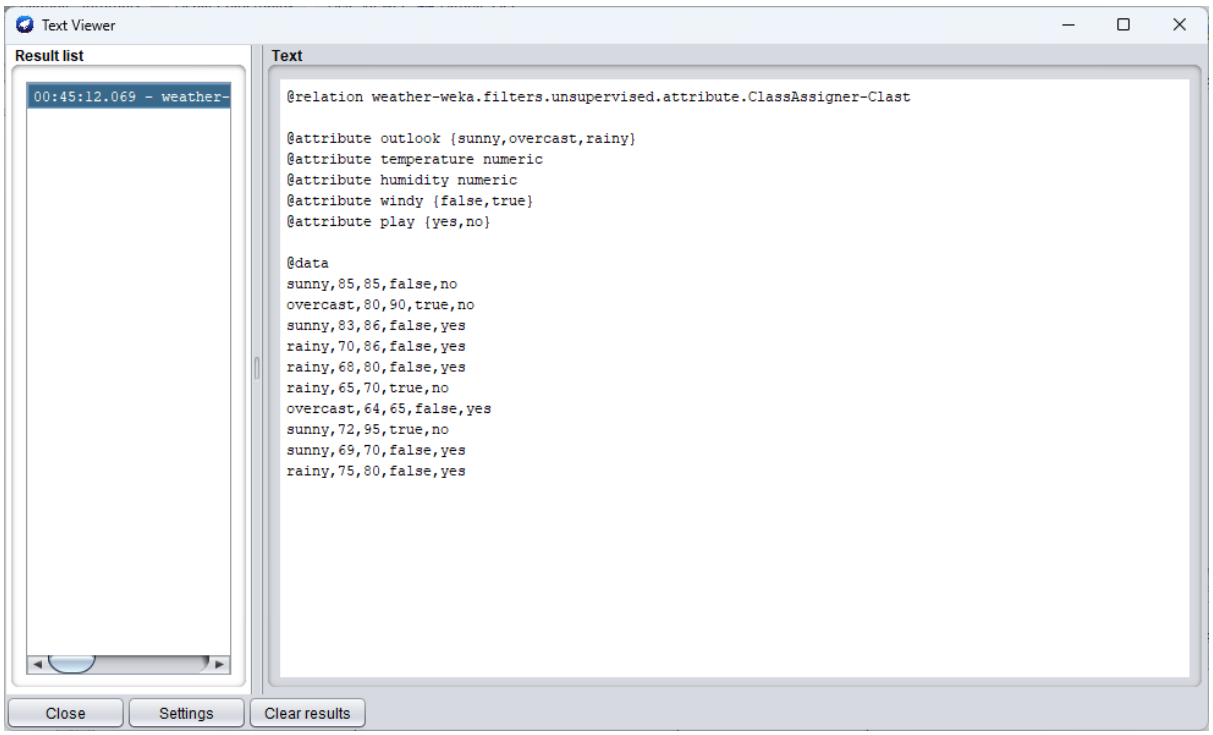
Load weather.arrf dataset and do the following operation using:



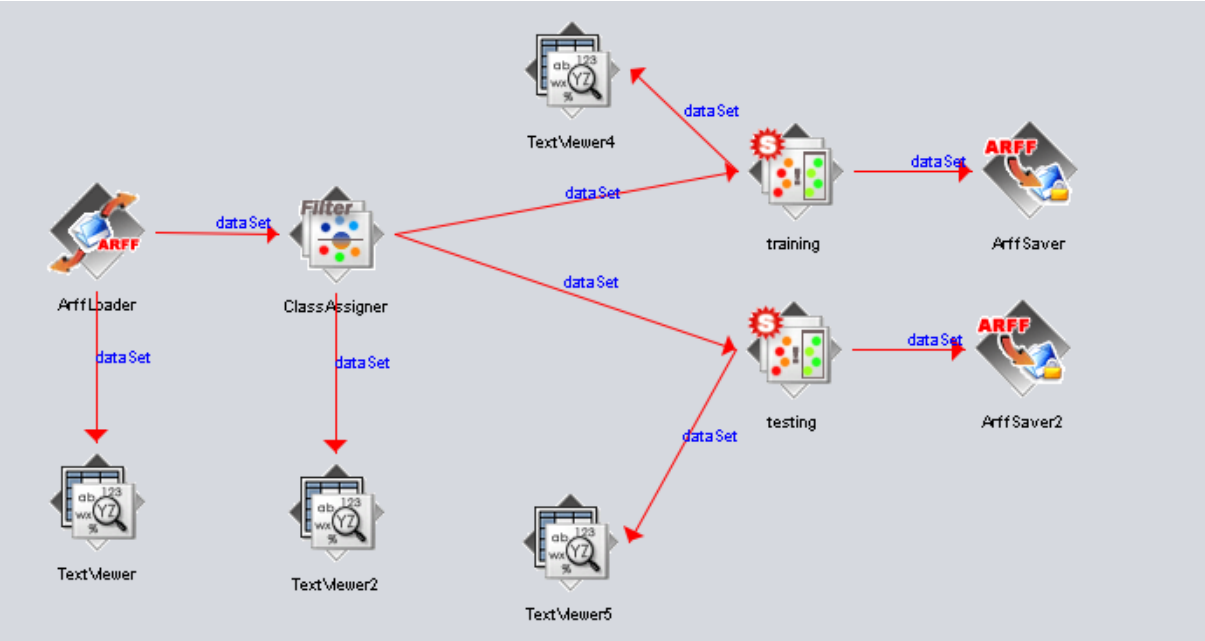
a. Make play attribute as Class label.

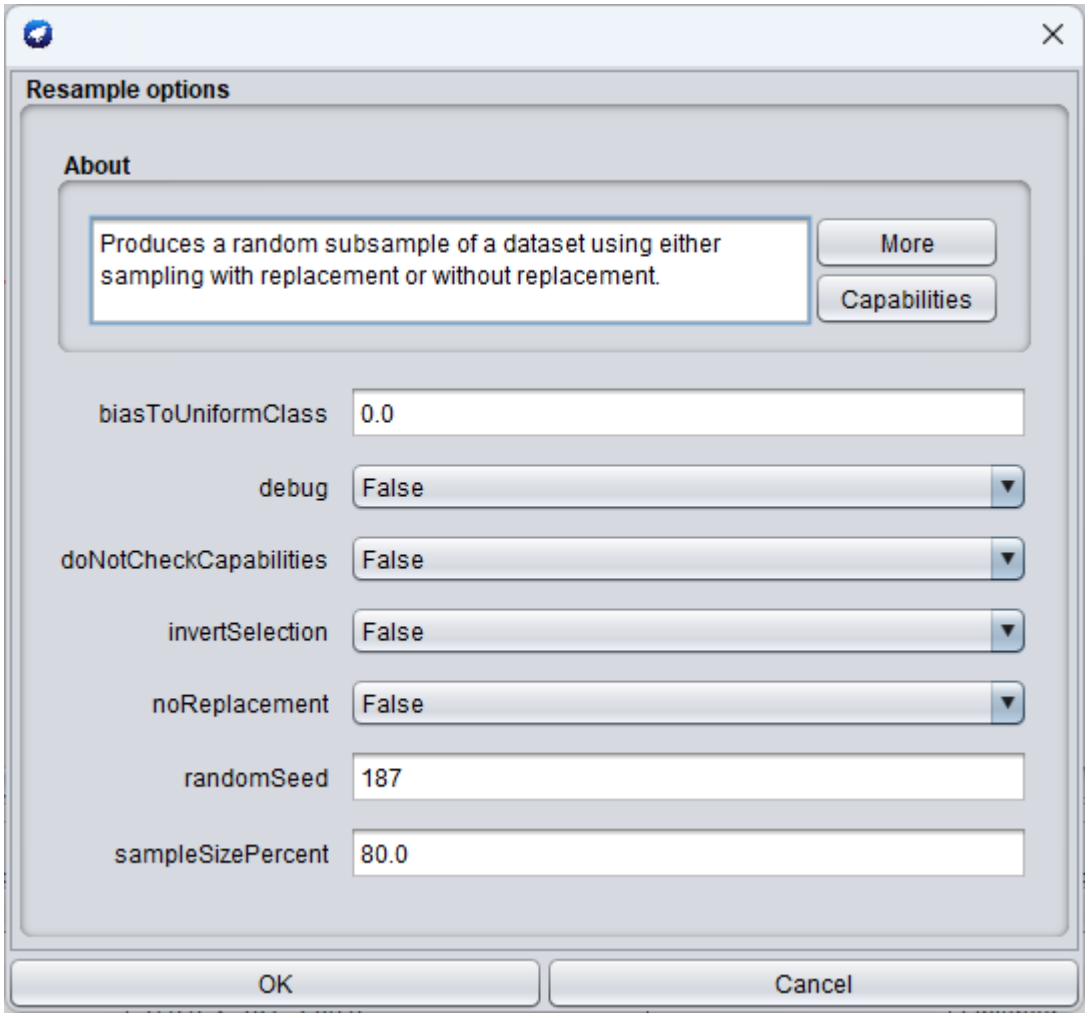


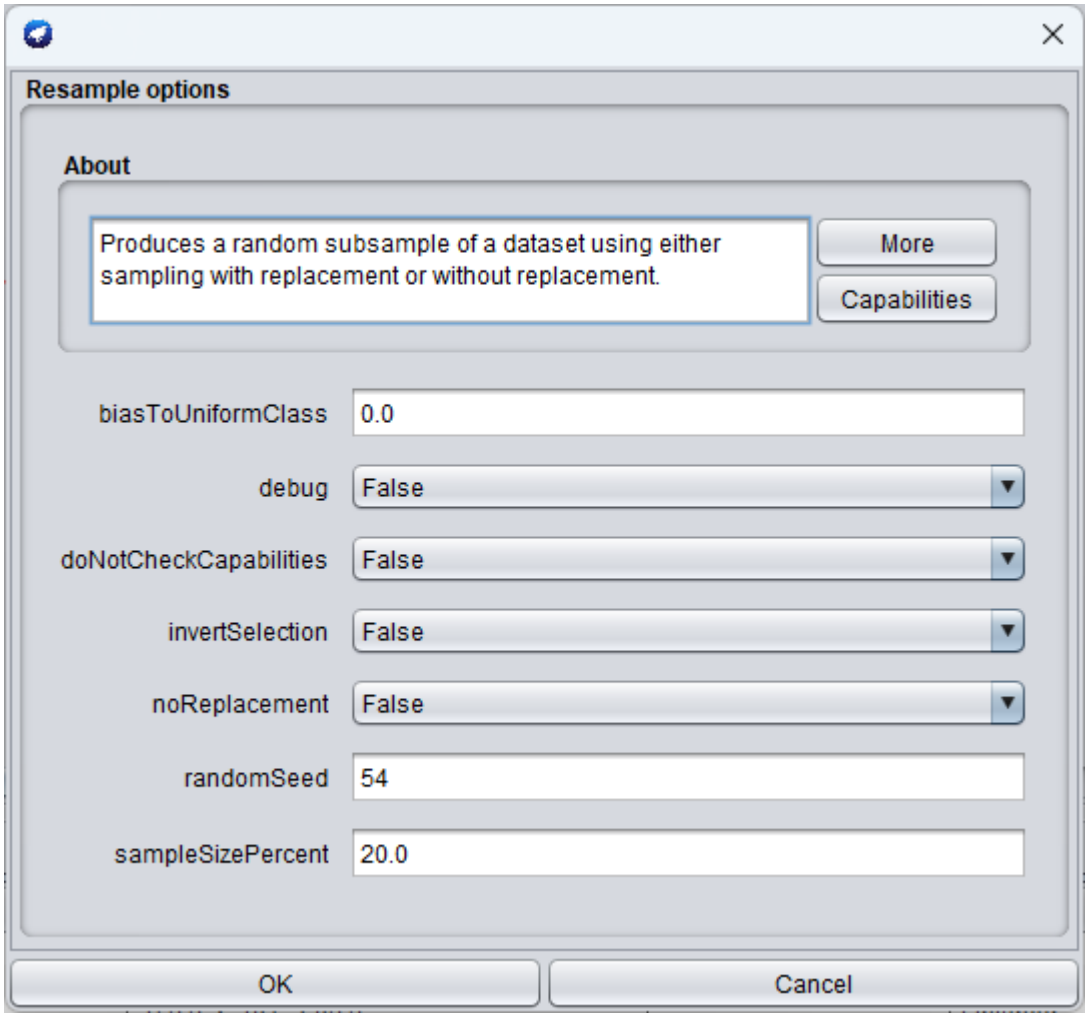




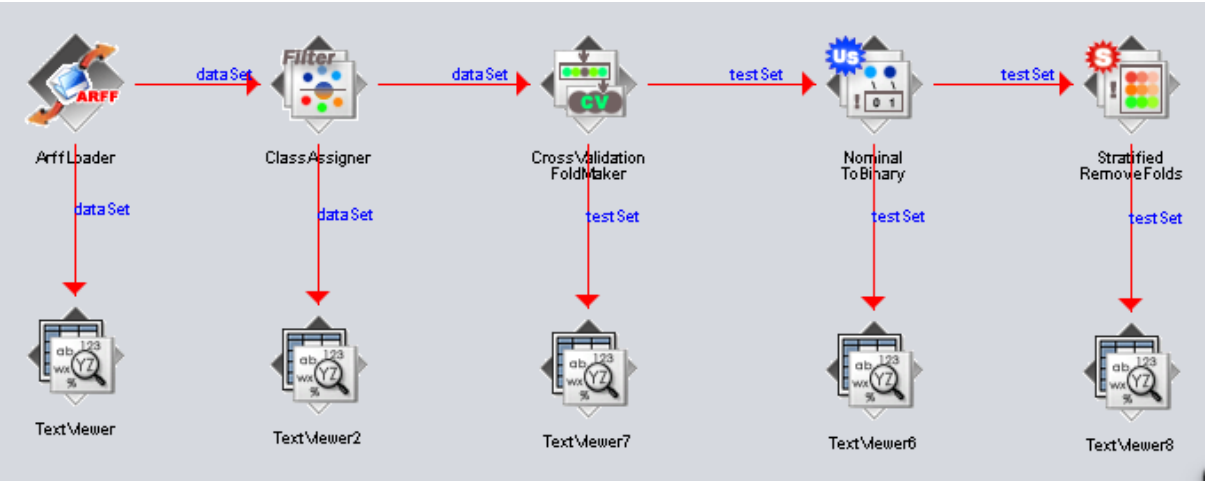
b. Split dataset into training (80%)and testing(20%) part and save both splits in different folds in directory location

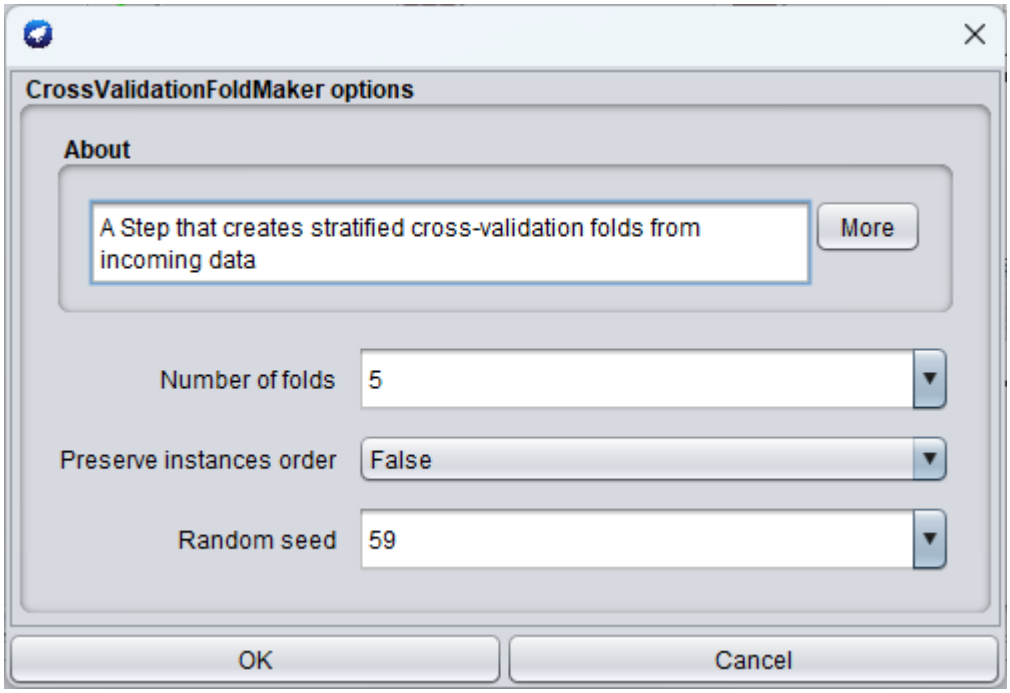


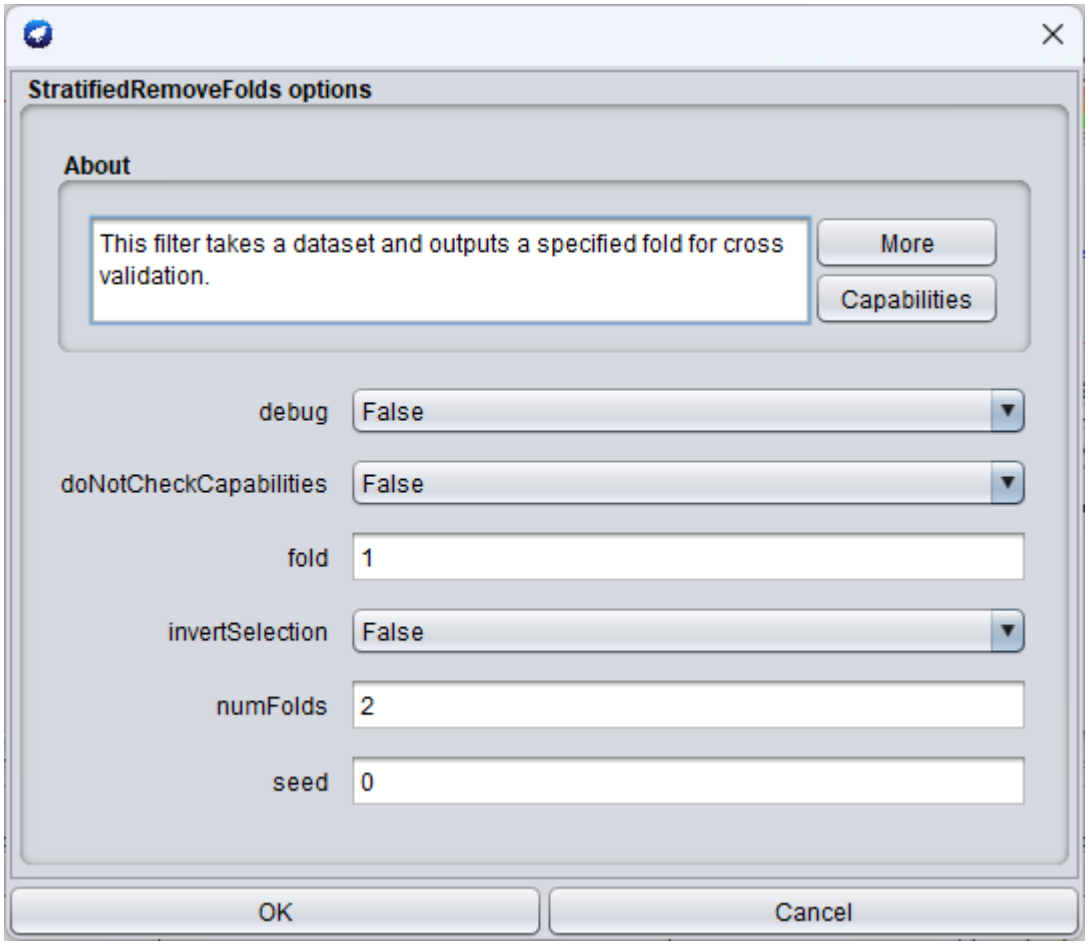

Resample options
About
Produces a random subsample of a dataset using either sampling with replacement or without replacement.
More
Capabilities
biasToUniformClass
0.0
debug
False
doNotCheckCapabilities
False
invertSelection
False
noReplacement
False
randomSeed
187
sampleSizePercent
80.0
OK
Cancel

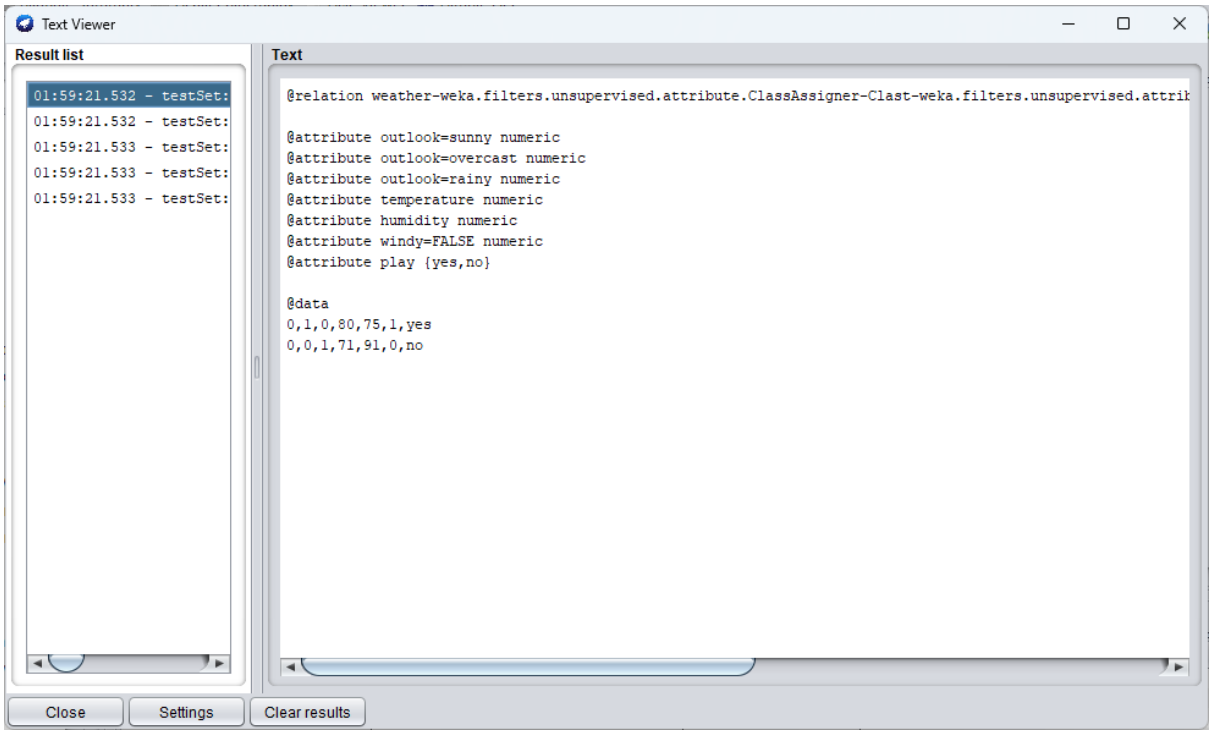


c. Prepare 5 fold cross validation testing data and filter all such folds which having all given class label nominal values.


ArffLoader
ClassAssigner
CrossValidation FoldMaker
Nominal ToBinary
Stratified RemoveFolds
dataSet
dataSet
testSet
testSet
dataSet
dataSet
testSet
testSet
testSet
TextViewer
TextViewer2
TextViewer7
TextViewer6
TextViewer8


CrossValidationFoldMaker options
About
A Step that creates stratified cross-validation folds from incoming data
More
Number of folds
5
Preserve instances order
False
Random seed
59
OK
Cancel


StratifiedRemoveFolds options
About
This filter takes a dataset and outputs a specified fold for cross validation.
More
Capabilities
debug
False
doNotCheckCapabilities
False
fold
1
invertSelection
False
numFolds
2
seed
0
OK
Cancel


Text Viewer
Result list
01:59:21.532 - testSet:
01:59:21.532 - testSet:
01:59:21.533 - testSet:
01:59:21.533 - testSet:
01:59:21.533 - testSet:
Text
@relation weather-weka.filters.unsupervised.attribute.ClassAssigner-Clast-weka.filters.unsupervised.attrib
@attribute outlook=sunny numeric
@attribute outlook=overcast numeric
@attribute outlook=rainy numeric
@attribute temperature numeric
@attribute humidity numeric
@attribute windy=FALSE numeric
@attribute play {yes,no}
@data
0,1,0,80,75,1,yes
0,0,1,71,91,0,no
Close
Settings
Clear results