EX. No : 01 STUDY OF WEKA, RAPID MINER TOOLS AND UCI REPOSITORY DATASETS
DATE :

AIM:

To explore the various features of Weka, Rapid miner Tools and UCI Repository datasets.

PROCEDURE:

FUNDAMENTAL TERMS:

Feature/Attribute: A single column of data is called a feature. It is a component of an observation and is also called an attribute of a data instance. Some features may be inputs to a model (the predictors) and others may be outputs or the features to be predicted.

Attribute values: Attribute values are numbers or symbols assigned to an attribute.

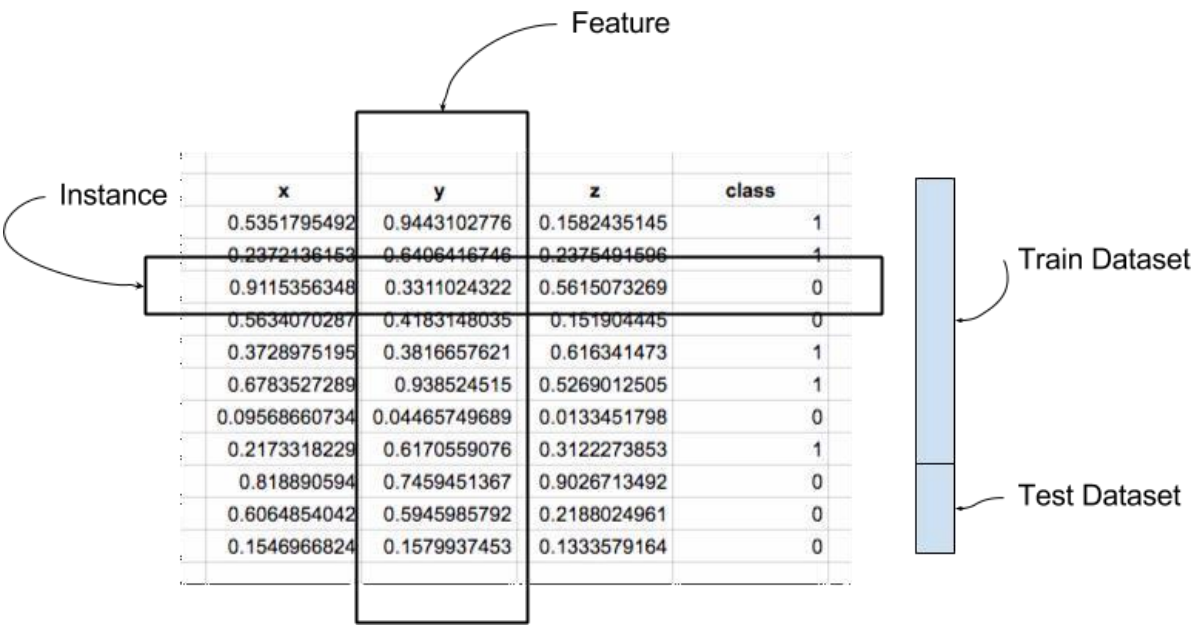
Target attribute: Target attribute is a special attribute which corresponds to the label of each instance.

Instance: Each row in the dataset is called the instance.

Datasets: A collection of instances is a dataset.

Training Dataset: A dataset that is fed into the machine learning algorithm to train the model.

Testing Dataset: A dataset that is used to validate the accuracy of the model but is not used to train the model.



WEKA:

1. Download and install weka,
2. In the window, select the explorer button from the available five buttons.
3. The weka supports two common formats for files:
   ARFF-Attribute Relation File Format
   CSV-Comma Separated Values **EXPLORER**:

The explorer window contains preprocess, classify, cluster, associate, select attribute and visualize from which select preprocess.

OPEN FILE:

To open the default dataset into the machine.

OPEN URL:

To access the dataset in the website.

OPENDB:

To open the database which the user saved in the machine.

CHOOSE:

To select the filter option.

EDIT:

To set the filled dataset before and after the filter.

FILTERS:

To filter or tune the data.

1. REMOVE(ATTRIBUTE) :

A filter that removes a range of attributes from the dataset.

1. Open file button is clicked
2. Choose weka and open the data folder
3. Choose **heart-disease.arff**
4. In filter tab, click choose button
5. Choose unsupervised and select attribute
6. Select remove filter
7. Specify the attribute index in the filter editor window
8. Apply button is clicked
9. Choose the edit button to see the output data after filtering the attribute

2. REMOVE WITH VALUES:

Filters instances according to the value of the attribute.

1. Open file button is clicked
2. Choose heart-disease.arff
3. Choose unsupervised and select instance
4. Select “remove with values” filter
5. Set attribute index to 2 and split point to 60
6. The output contains column dataset in which the second column contains only the values which is above the split point
7. Choose the **Edit button** to view the filtered dataset

3. REPLACE WITH MISSING VALUES:

Replace all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

1. Choose unsupervised and select attributes.
2. Select “remove missing values” filter.
3. Click “edit” and delete any one of the data.
4. Select **“Replace Missing Values”** filter
5. After applying the filter, the deleted values or any missing values are replaced by taking the mean values.
6. Click **Edit** button to view the updated dataset

4. REMOVE PERCENTAGE:

A filter that removes a given percentage of a database.

1. Choose unsupervised and select instance.
2. Select “remove percentage” filter.
3. Set percentage as “50.0”
4. After filter is applied , from the dataset 50% of the instance are removed

5. REMOVE FREQUENT VALUES:

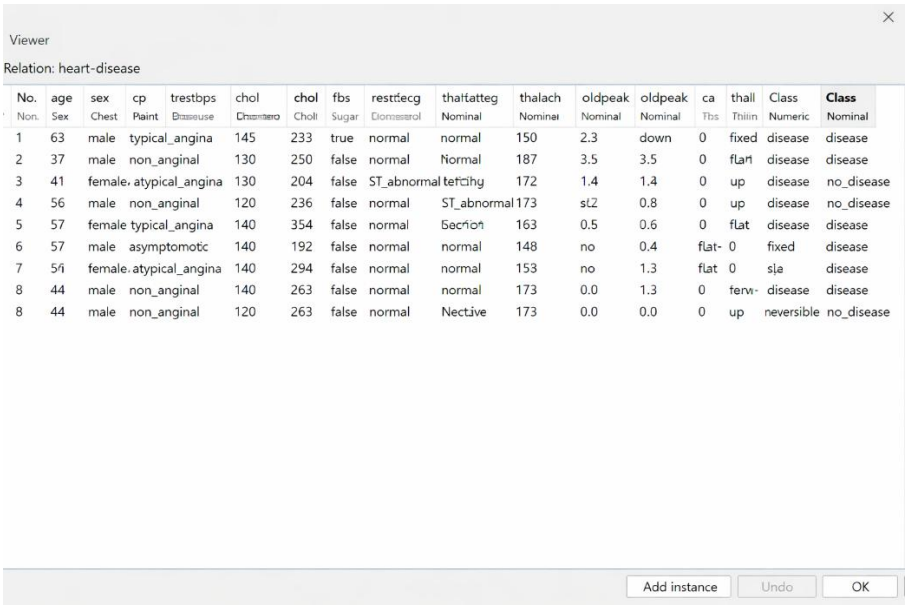
Determine which values of attribute or retained and filters the instances accordingly.

1. Choose unsupervised and select instances.
2. Select “remove frequent values” filter.
3. Specify the attribute index as 2.

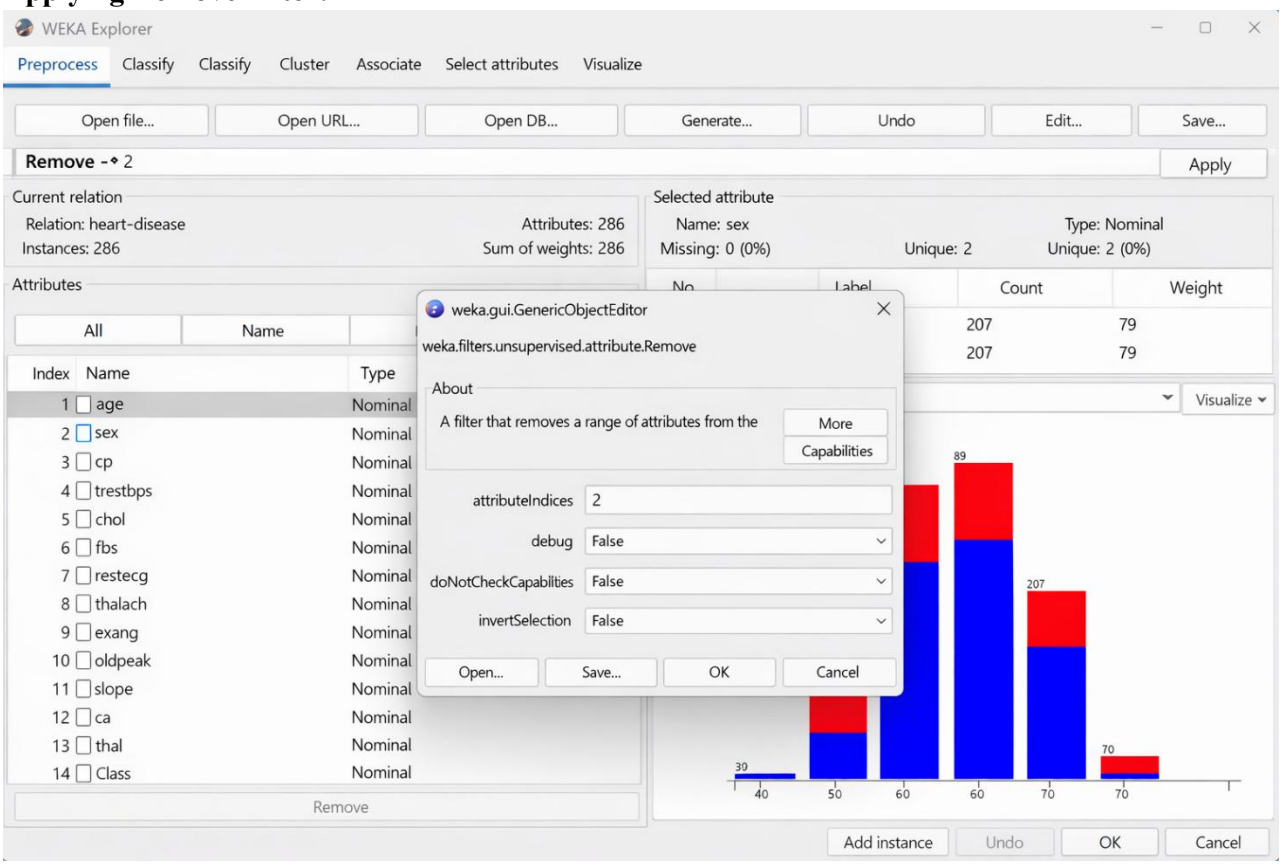
4. When apply is clicked the less frequently repeated values are removed

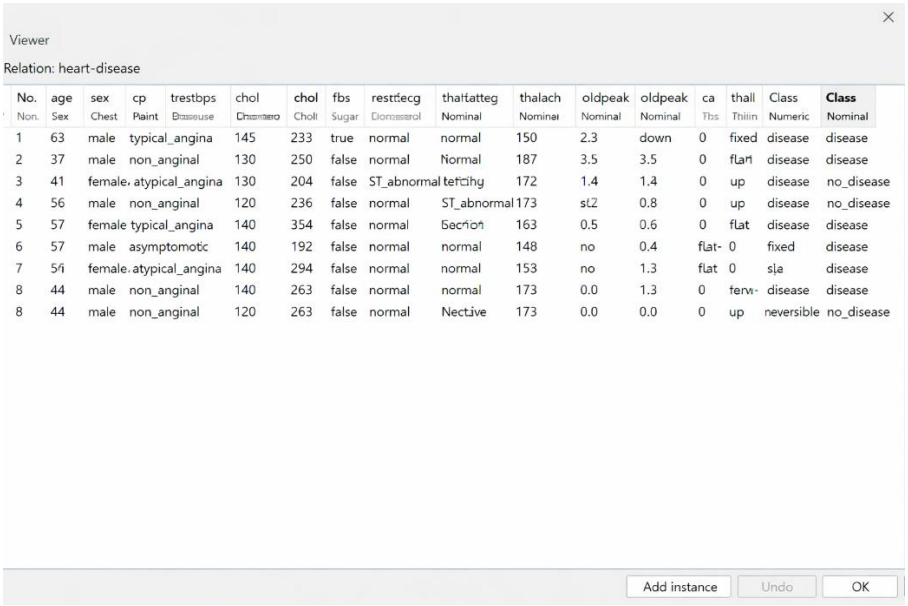
OUTPUT:

Dataset: heart-disease.arff

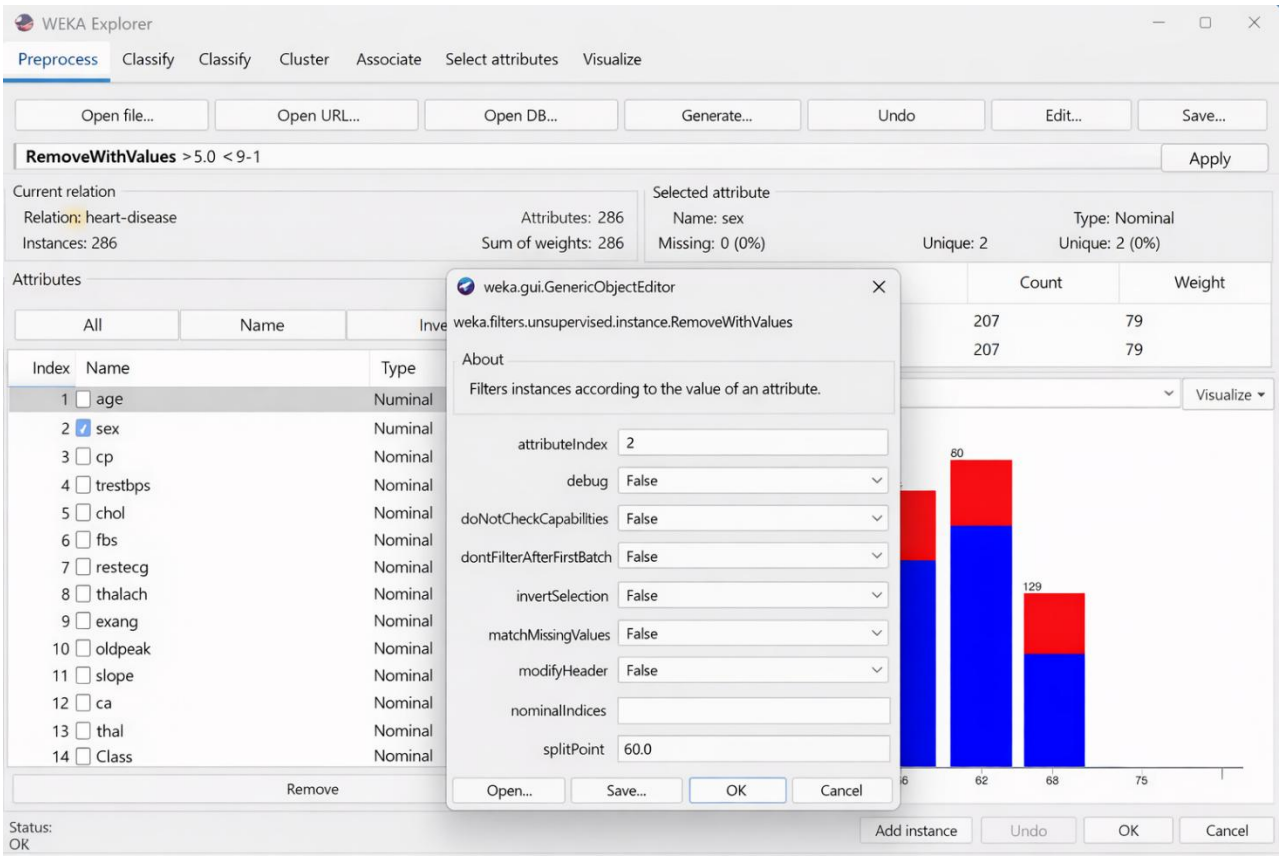


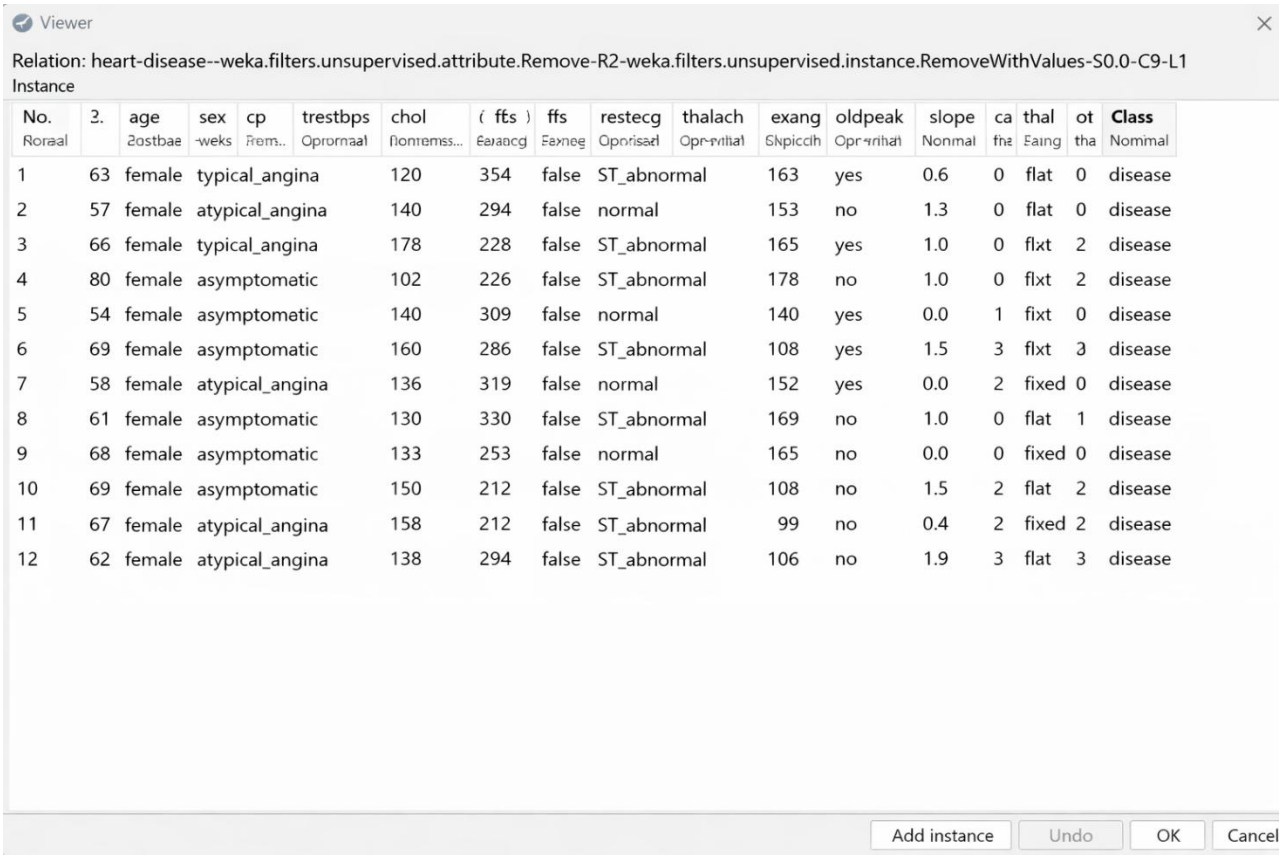
Applying Remove Filter:



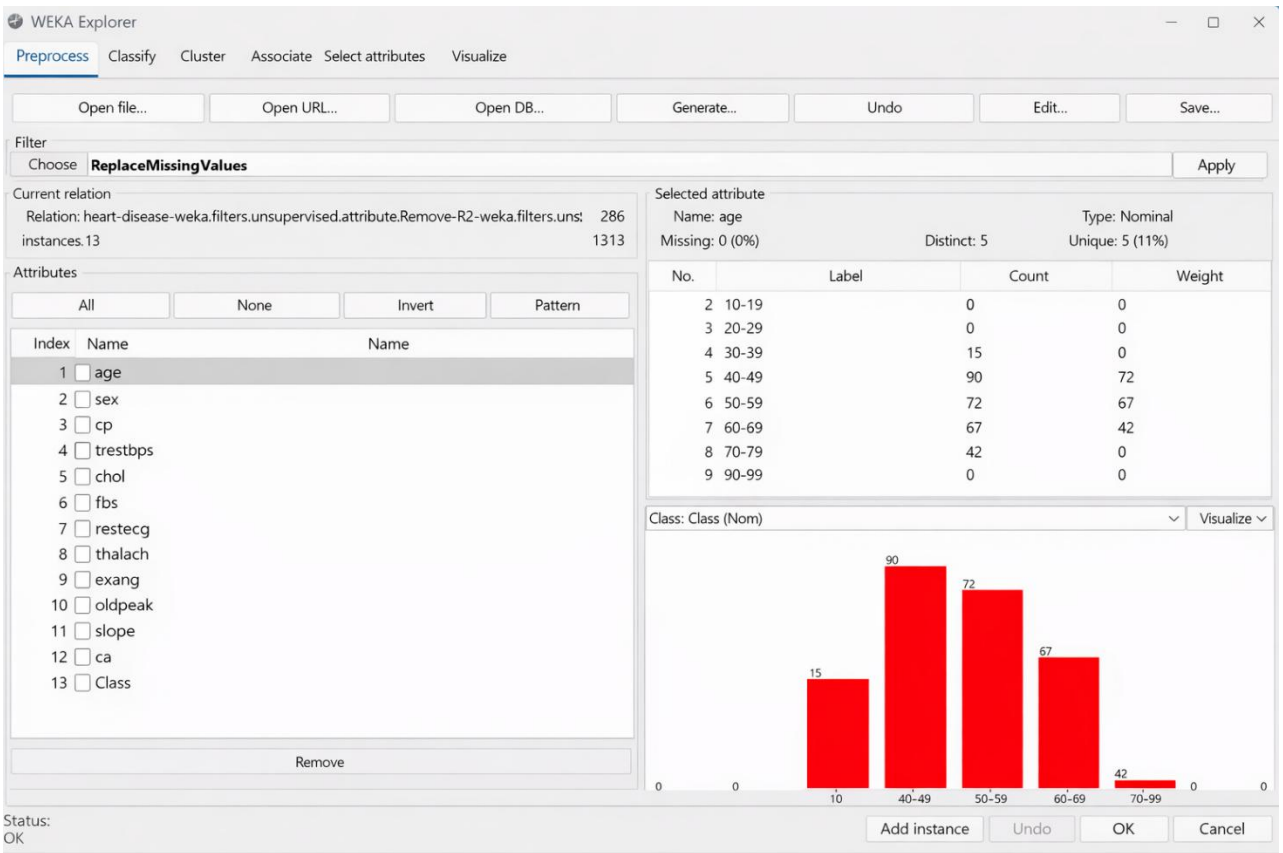


Applying Remove with values Filter





Applying Replacing Missing Value Filter



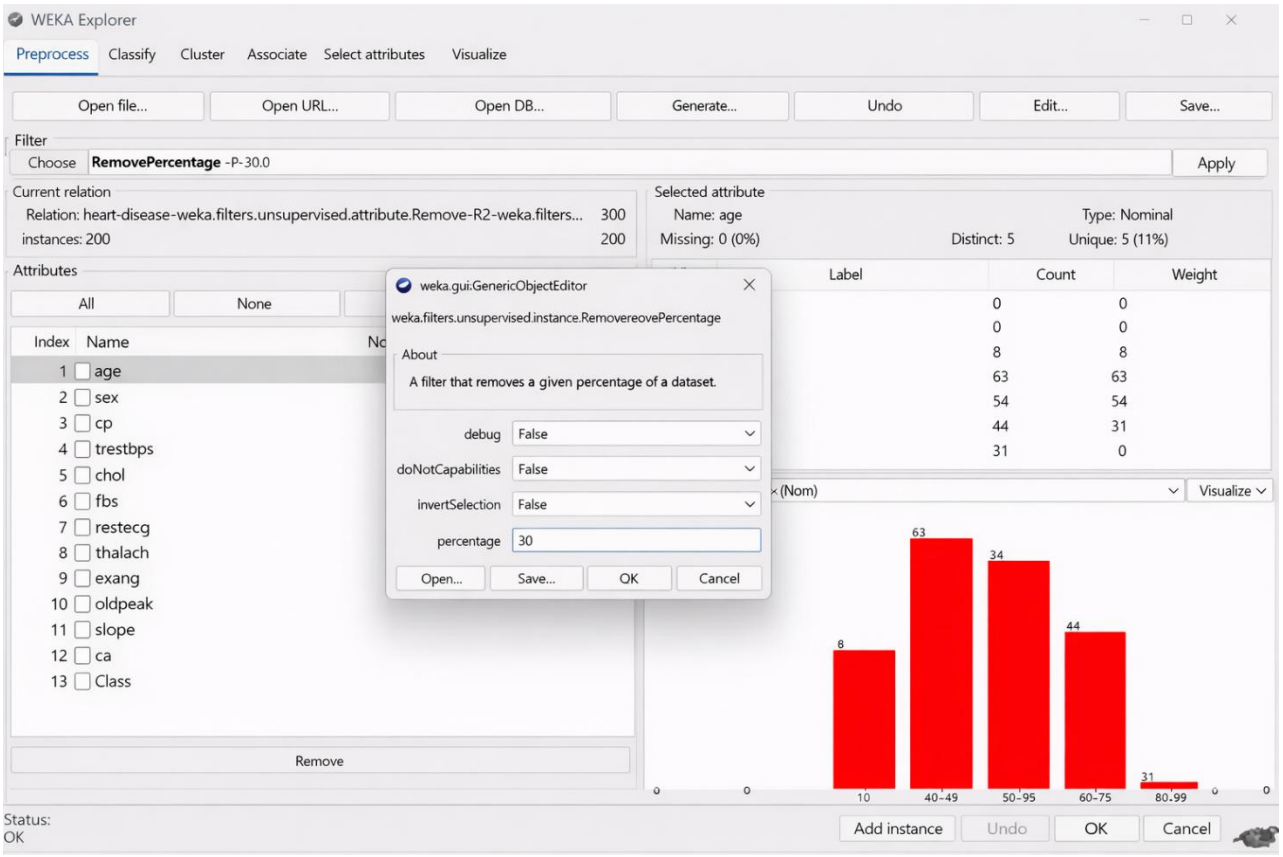
Viewer

Relation: heart-disease-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.intsaince.Remove.RemoveWithValues-S0.0-C9-L1

No.	1: age Nominal	2: sex Nominal	cp Nominal	cp Nominal	trestbps Numeric	5: chol Numeric	7: fbs Nominal	7: restecg Nominal	8: thalach Numeric	9: exang Numeric	Oldpeak Numeric	11: slope Nominal	12: 11 Nominal	**Class** Nominal
1	63	female	typical_angina	120	354	false	ST_abnormal	163	yes	0.6	flat	0	0	disease
2	57	female	atypical_angina	140	294	false	normal	153	no	1.3	flat	0	0	disease
3	66	female	typical_angina	178	228	false	ST_abnormal	165	yes	1.0	flat	2	2	disease
4	80	female	asymptomatic	102	226	false	ST_abnormal	178	no	1.0	flat	1	1	disease
5	54	female	asymptomatic	140	309	false	normal	140	yes	0.0	up	1	1	disease
6	67	female	asymptomatic	160	286	false	ST_abnormal	108	yes	1.5	flat	3	3	disease
7	58	female	atypical_angina	136	319	false	normal	152	yes	0.0	up	2	2	disease
8	61	female	asymptomatic	130	330	false	ST_abnormal	169	yes	0.0	flat	0	1	disease
9	68	female	asymptomatic	133	253	false	normal	165	no	0.0	flat	1	2	disease
10	69	female	asymptomatic	160	212	false	ST_abnormal	108	no	1.5	flat	2	2	disease
11	67	female	asymptomatic	152	212	false	ST_abnormal	99	no	0.4	flat	2	2	disease
12	62	female	atypical_angina	138	294	false	ST_abnormal	106	no	1.9	flat	3	3	disease

Status: Add instance | Undo | OK | Cancel

Applying Remove Percentage Filter



Viewer

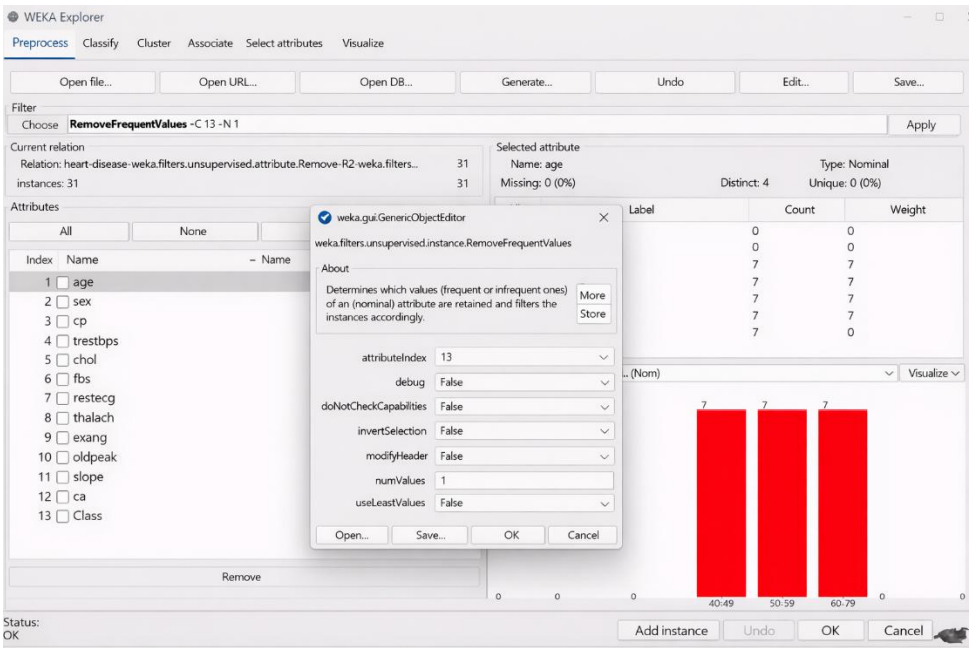
Relation: heart-disease-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.instance.RemovePercentage-P30.0

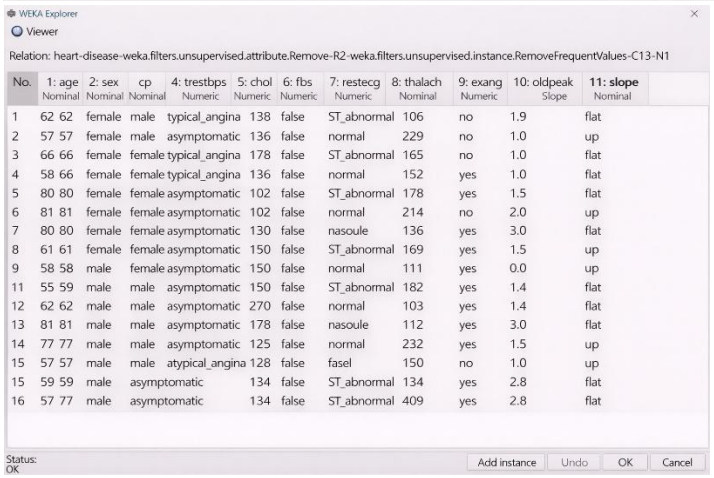
No.	1: age Nominal	2: sex Nrestbps	cp Nominal	ttrestbps Abominal	5: chol Nominal	6: fbs Nominal	7: restecg Nominal	8: rrigh Sonical	thalach Sonical	9: exang Nominal	10: oldpeak (Aponines)	12: slope Numinal	**Class** Nominal
1	62	female	typical_angina	138	false	ST_abnormal	106	106	no	1.9	3	disease	
2	57	female	asymptomatic	136	false	normal	229	152	no	1.0	2	disease	
3	66	female	typical_angina	178	false	ST_abnormal	178	165	no	1.0	2	disease	
4	58	female	typical_angina	228	false	normal	152	173	yes	1.0	2	disease	
5	80	female	asymptomatic	102	false	ST_abnormal	178	178	no	1.5	2	disease	
6	81	female	asymptomatic	130	false	normal	178	169	yes	1.0	1	disease	
7	81	female	asymptomatic	102	false	ST_abnormal	163	178	ne	2.0	2	disease	
8	80	female	asymptomatic	159	false	normal	214	228	no	3.0	1	disease	
9	81	female	asymptomatic	198	false	(fssy)	163	209	no	3.0	2	disease	
10	80	female	asymptomatic	199	false	nasoucle	318	152	no	3.0	1	disease	
11	81	female	asymptomatic	100	false	normal	168	109	no	0.0	2	disease	
12	81	female	asymptomatic	130	false	fiasel	228	213	usv	3.0	1	disease	
13	66	female	asymptomatic	120	false	normal	209	228	yes	2.0	4	disease	
14	74	female	asymptomatic	100	false	faset	121	106	no	1.9	3	disease	

Sevics:

Add instance | Undo | OK | Cancel

Applying Remove Frequent Values Filter

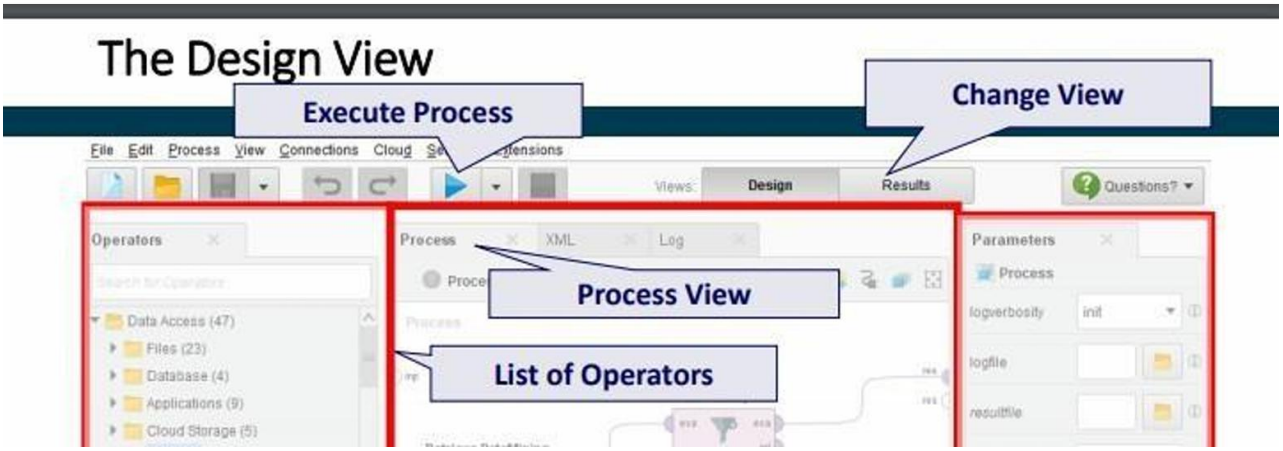




No.	1: age Nominal	2: sex Nominal	cp Nominal	4: trestbps Numeric	5: chol Numeric	6: fbs Numeric	7: restecg Numeric	8: thalach Nominal	9: exang Numeric	10: oldpeak Slope	**11: slope** Nominal
1	62 62	female	male	typical_angina	138	false	ST_abnormal	106	no	1.9	flat
2	57 57	female	male	asymptomatic	136	false	normal	229	no	1.0	up
3	66 66	female	female	typical_angina	178	false	ST_abnormal	165	no	1.0	flat
4	58 66	female	female	typical_angina	136	false	normal	152	yes	1.0	flat
5	80 80	female	female	asymptomatic	102	false	ST_abnormal	178	yes	1.5	flat
6	81 81	female	female	asymptomatic	102	false	normal	214	no	2.0	up
7	80 80	female	female	asymptomatic	130	false	nasoule	136	yes	3.0	flat
8	61 61	female	female	asymptomatic	150	false	ST_abnormal	169	yes	1.5	up
9	58 58	male	female	asymptomatic	150	false	normal	111	yes	0.0	up
11	55 59	male	male	asymptomatic	150	false	ST_abnormal	182	yes	1.4	flat
12	62 62	male	male	asymptomatic	270	false	normal	103	yes	1.4	flat
13	81 81	male	male	asymptomatic	178	false	nasoule	112	yes	3.0	flat
14	77 77	male	male	asymptomatic	125	false	normal	232	yes	1.5	up
15	57 57	male	male	atypical_angina	128	false	fasel	150	no	1.0	up
15	59 59	male	male	asymptomatic	134	false	ST_abnormal	134	yes	2.8	flat
16	57 77	male	male	asymptomatic	134	false	ST_abnormal	409	yes	2.8	flat

RAPIDMINER:

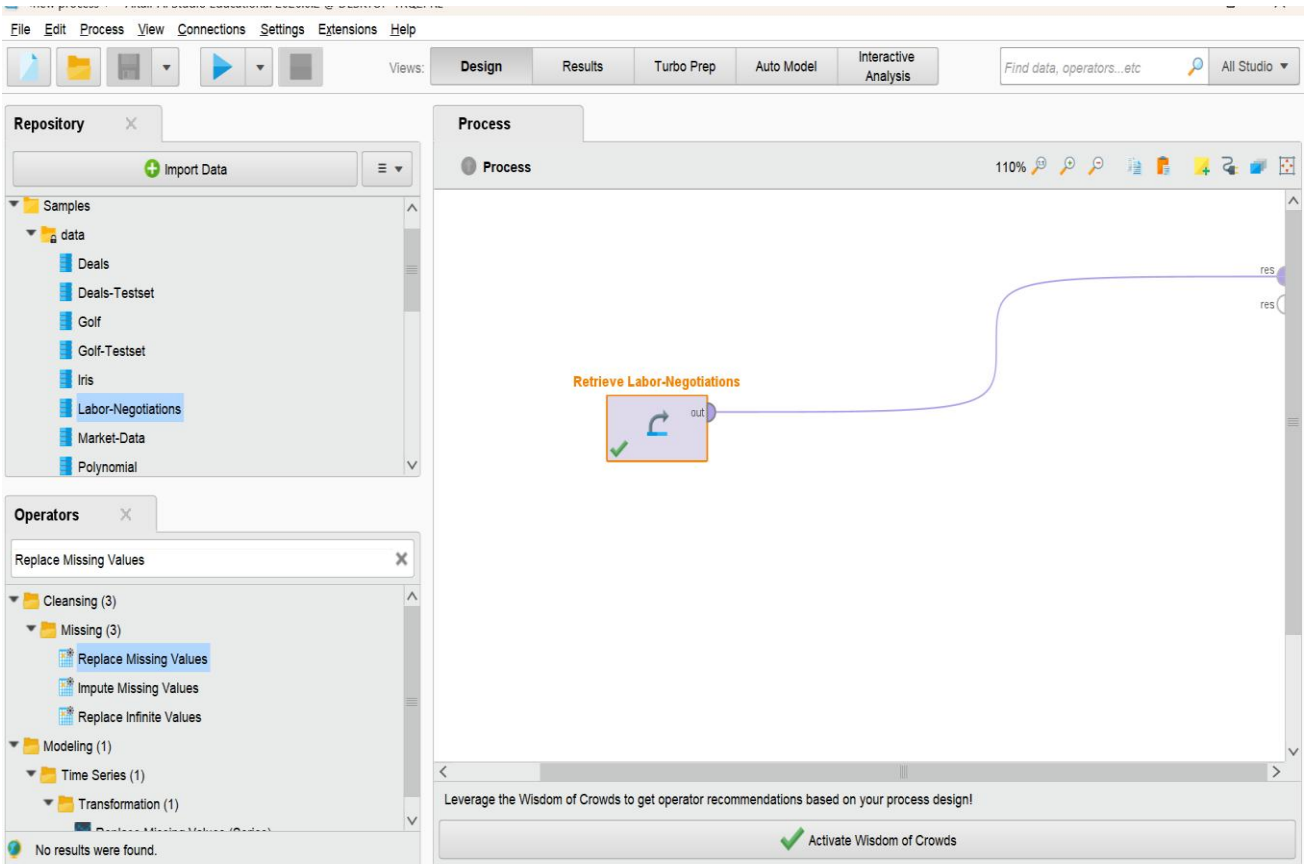
Design View

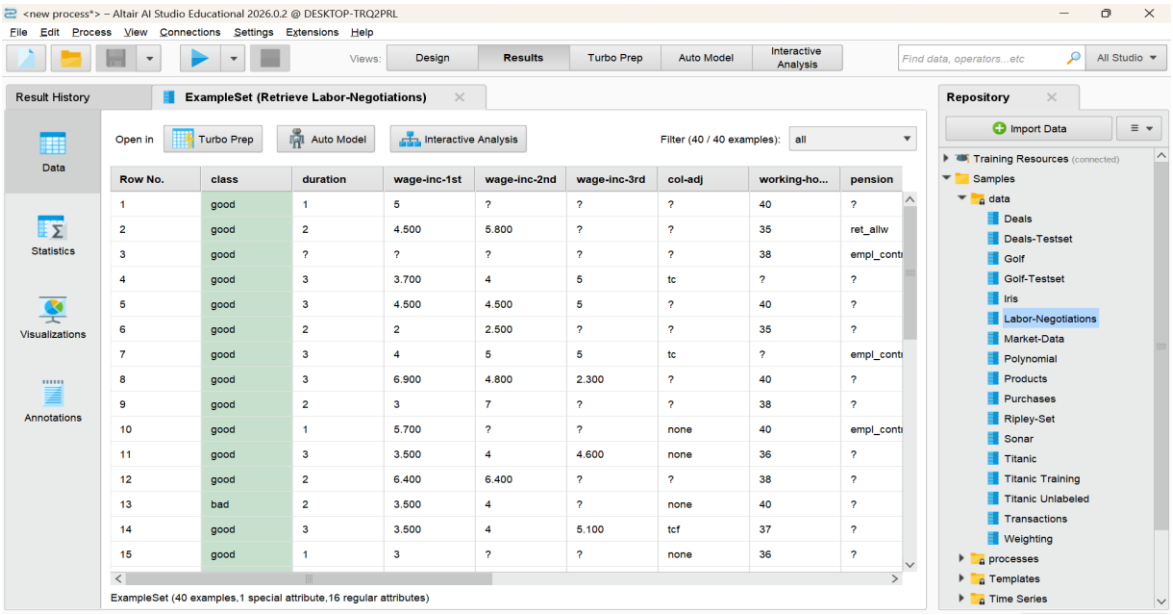


Preprocessing: Replace missing values

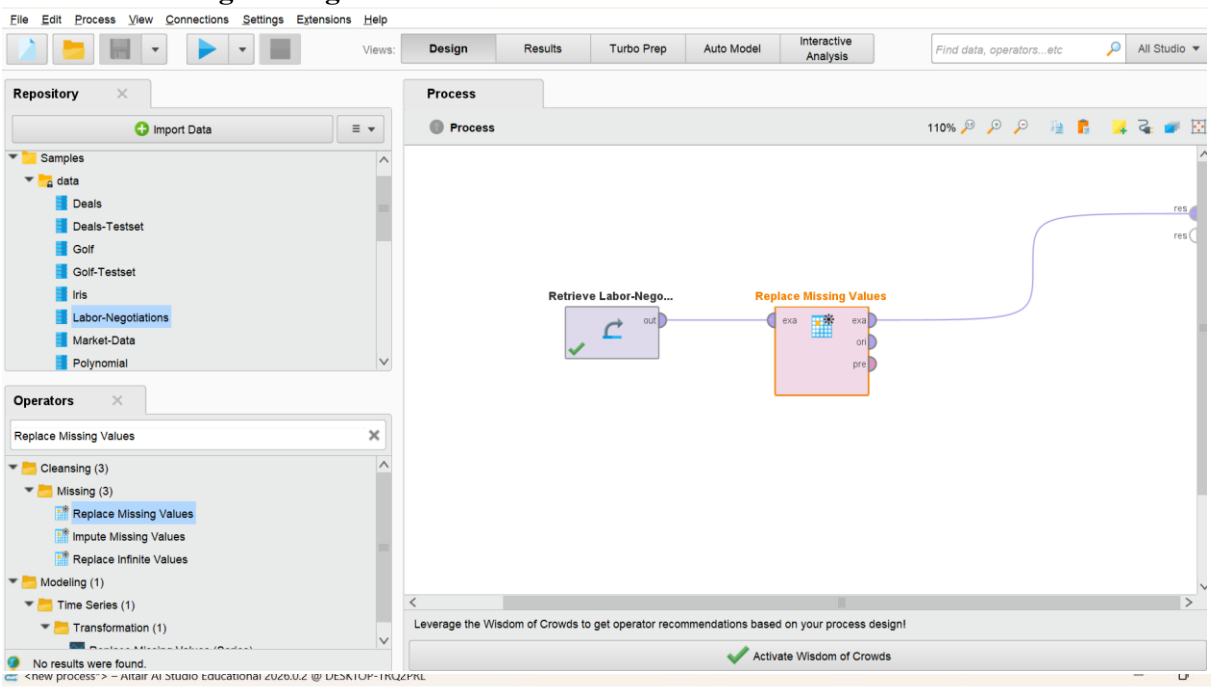
1. Load the Labor-Negotiations data set from the Samples folder.
2. Drag and drop the Replace Missing Values Operator. It applies the replacement on all attributes in the dataset which have at least one missing value.
3. Click the play button and view the output.

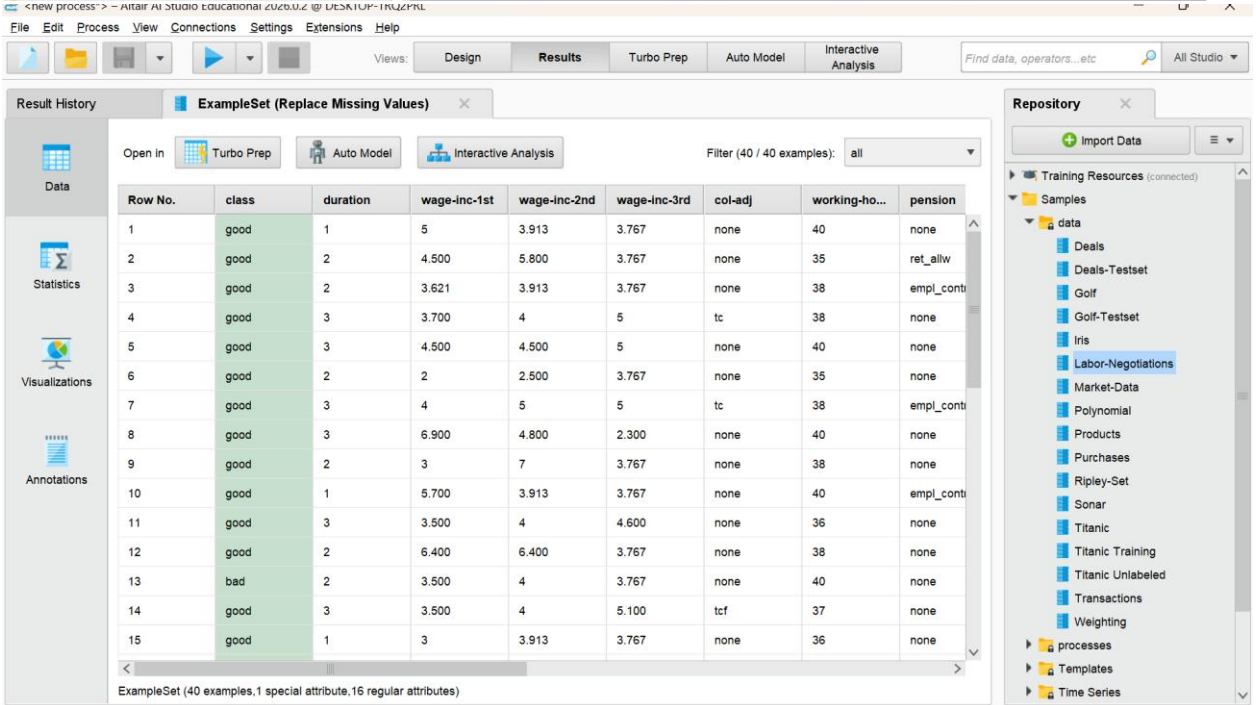
Dataset with missing values:



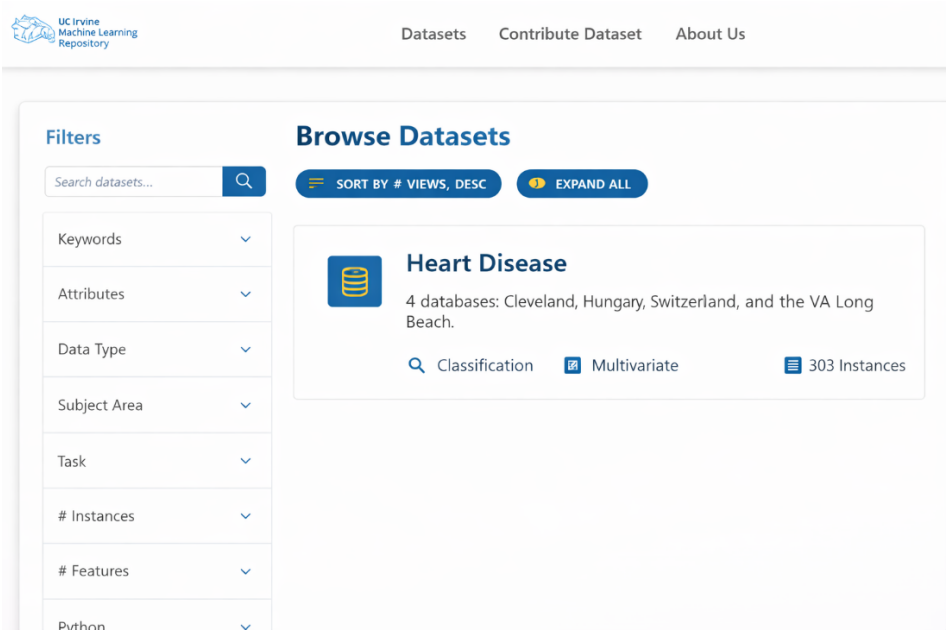


Dataset after filling missing values:

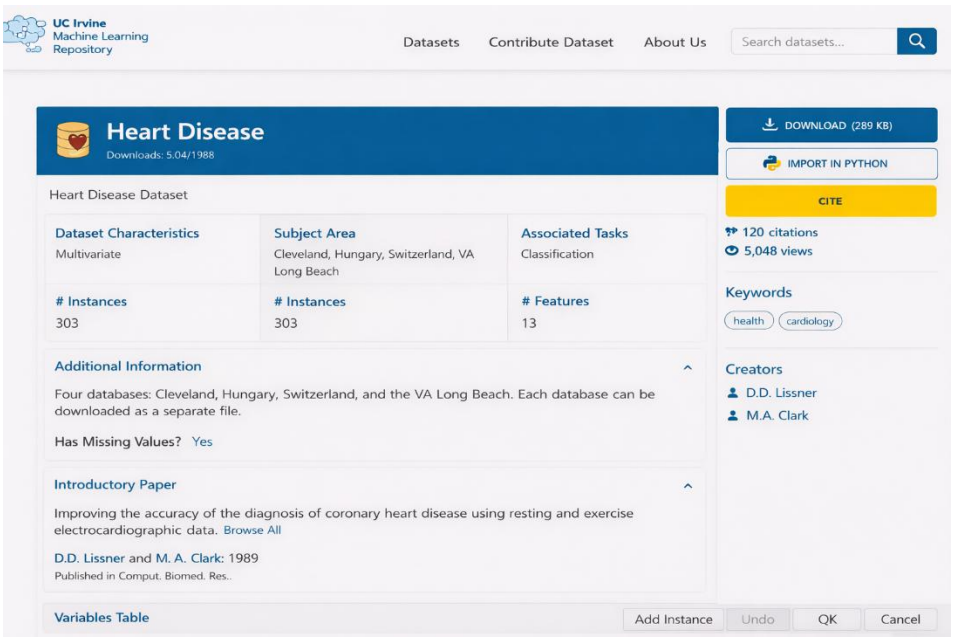




UCI Repository:



Sample dataset:



Conclusion: The various features of Weka Tool, Rapidminer Tool and UCI Repository datasets have been explored.

MARK ALLOCATION	
Conduct of Experiment(30)	
Record Observation (20)	
Viva (10)	
Total (60)	
Signature of the Faculty with Date	