Experiment 4

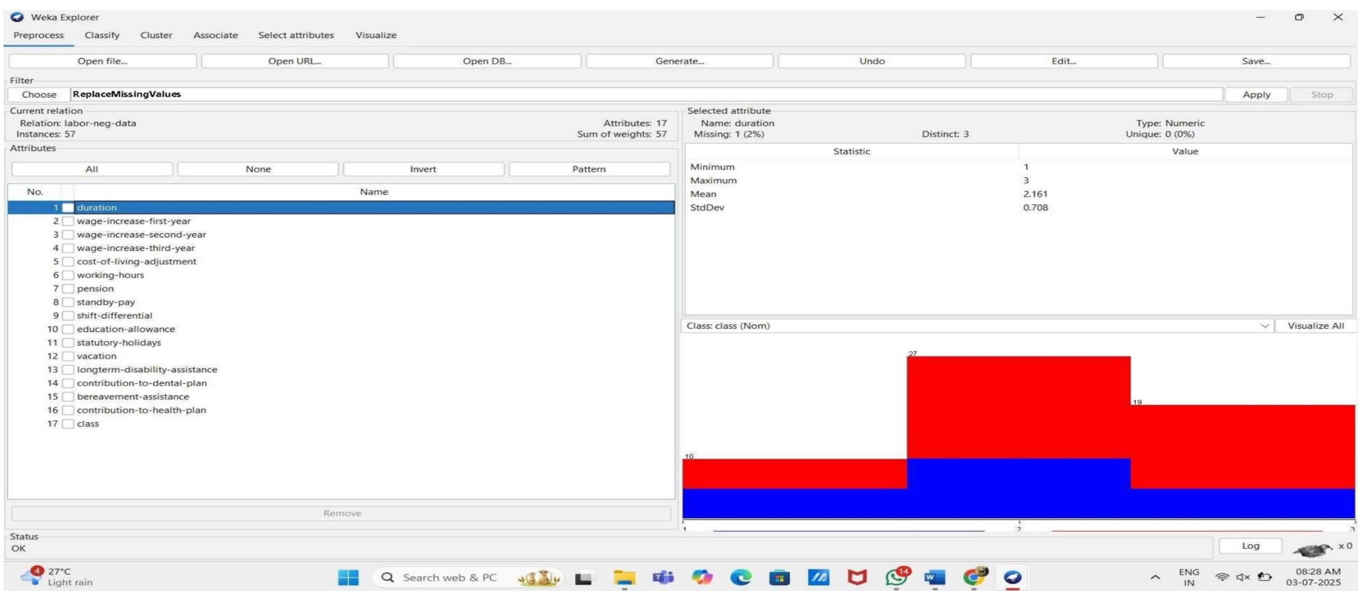
Title: Apply Preprocessing techniques on dataset using filters: Remove, ReplaceMissingValues, ReplaceMissingWithUserConstant, ReplaceWithMissingValue, Descritize. Also do the result analysis before and after preprocessing.

Filter 1: Remove

The Remove filter in Weka is an unsupervised attribute filter used to delete specific attributes (columns) from a dataset. It is commonly applied during data preprocessing to eliminate irrelevant, redundant, or non-informative features such as ID numbers or metadata that do not contribute to the learning process. Located under filters → unsupervised → attribute → Remove, this filter allows users to specify which attributes to remove using the -R option, where attributes are indexed starting from 1. For example, using
-R 1,3 will remove the first and third attributes from the dataset. The Remove filter is essential for simplifying the dataset and improving model performance by focusing only on the most relevant attributes.

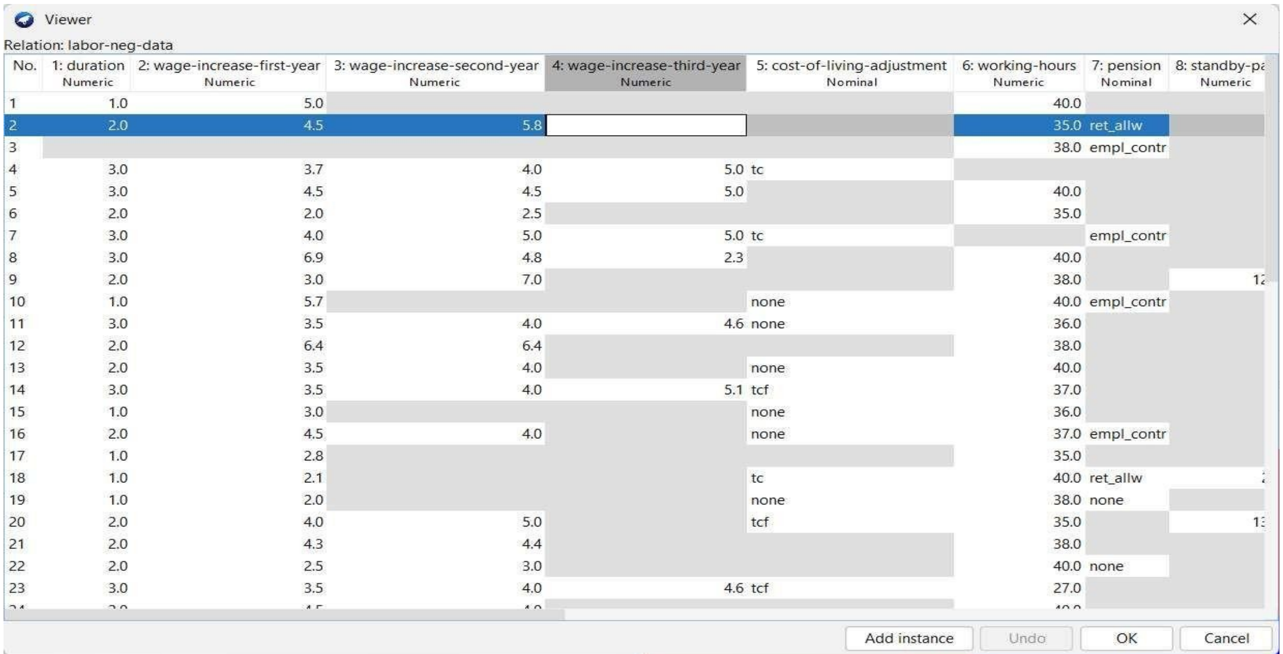
Dataset: labor.arff

Step-1: Upload dataset in Weka.



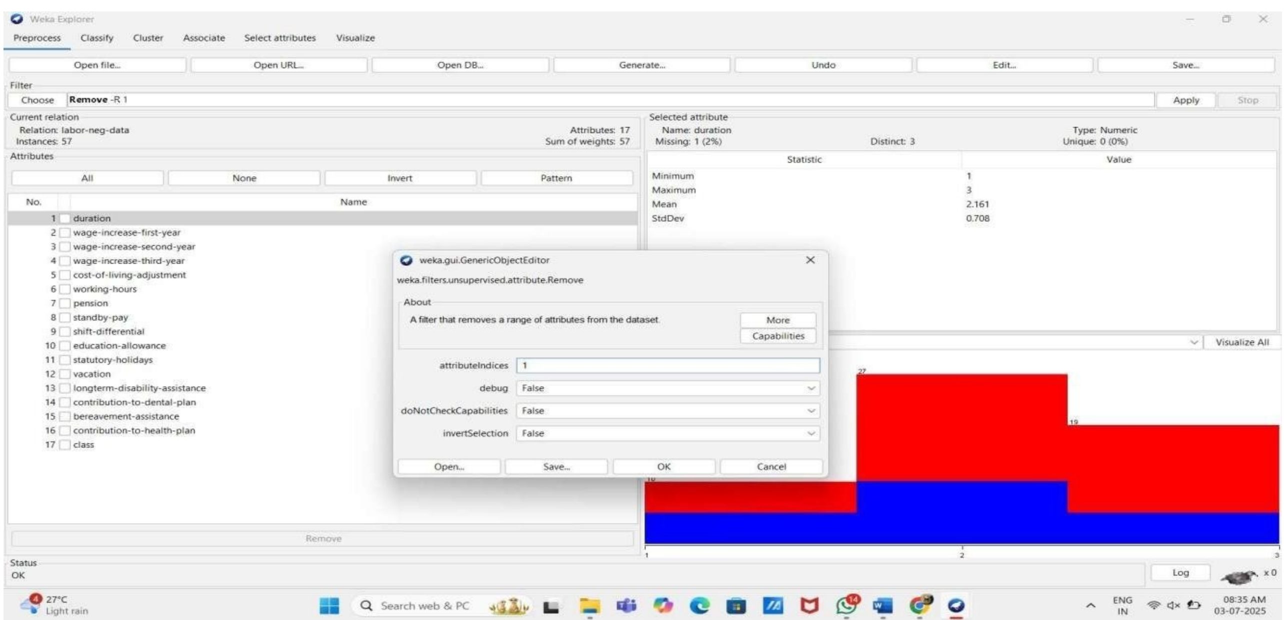
The labor.arff dataset is related to labor negotiations, typically involving attributes about employee benefits and working conditions. It consists of 57 instances (rows) and 17 attributes, including variables like duration, wage-increase, pension, vacation, contribution- to-health-plan, and more. The class attribute is usually a nominal variable indicating the outcome or status of the labor negotiation (e.g., "good" or "bad"). The dataset contains both numeric and nominal attributes, and may include missing values, as seen in the duration attribute.

Step-2: dataset in table format.



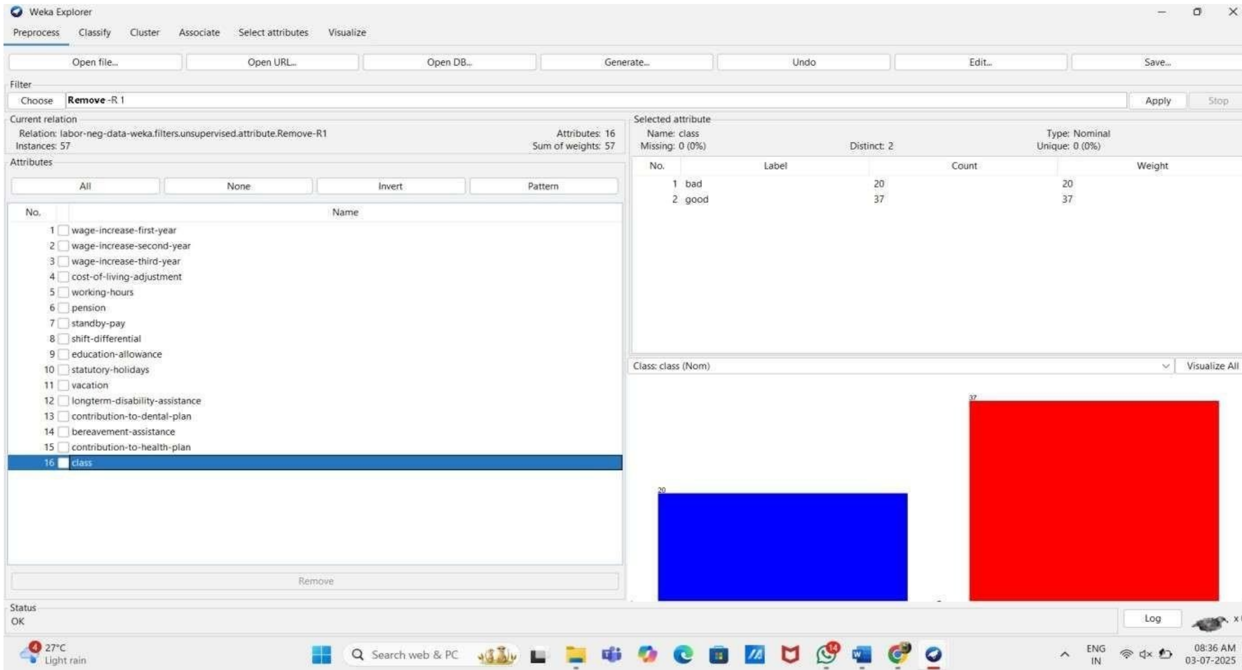
This screenshot shows the tabular view of the labor.arff dataset in Weka's Instance Viewer. Each row represents an instance (or record) from labor negotiations data, and each column is an attribute related to employment terms.

Step-3: Apply Remove filter to duration attribute.



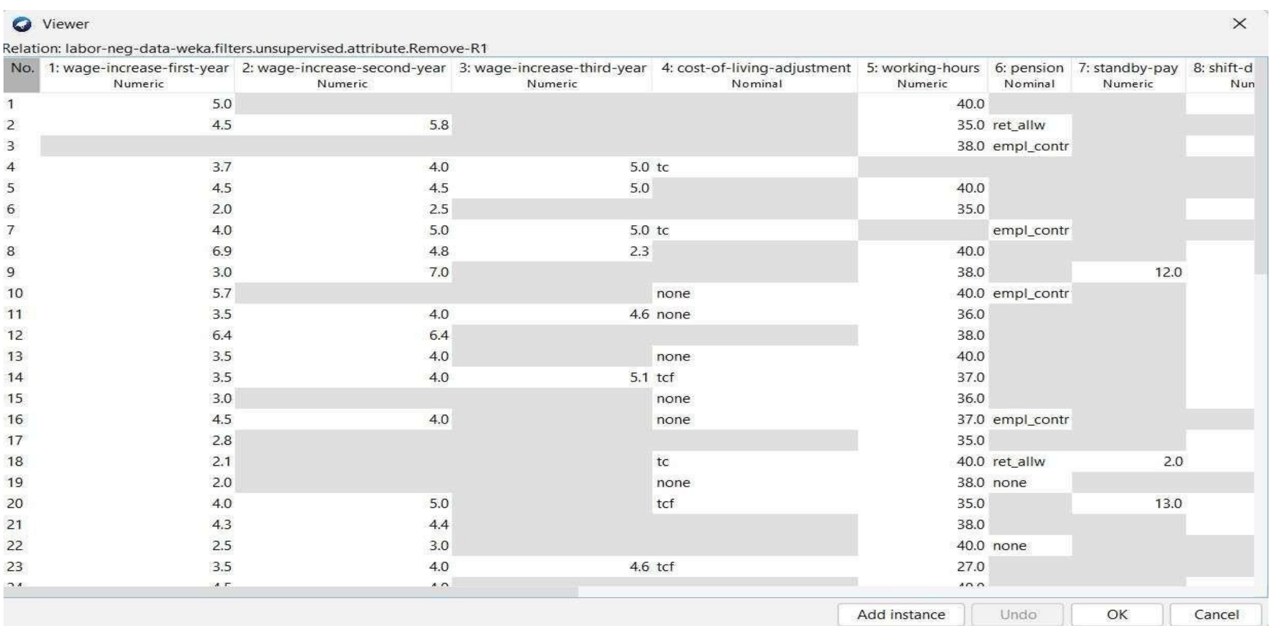
This screenshot shows the Remove filter being applied in Weka to delete the first attribute (duration) from the labor-neg-data dataset. The attributeIndices field is set to 1, and the filter configuration window is open for editing before applying the change.

Step-4: Apply Remove filter to duration attribute.



This Screenshot shows that duration attribute is removed after applying Remove filter.

Step-5: Dataset in table format after applying Remove filter.


Viewer

Relation: labor-neg-data-weka.filters.unsupervised.attribute.Remove-R1

No.	1: wage-increase-first-year Numeric	2: wage-increase-second-year Numeric	3: wage-increase-third-year Numeric	4: cost-of-living-adjustment Nominal	5: working-hours Numeric	6: pension Nominal	7: standby-pay Numeric	8: shift-d Nun
1	5.0				40.0			
2	4.5	5.8			35.0	ret_allw		
3					38.0	empl_contr		
4	3.7	4.0	5.0	tc				
5	4.5	4.5	5.0		40.0			
6	2.0	2.5			35.0			
7	4.0	5.0	5.0	tc		empl_contr		
8	6.9	4.8	2.3		40.0			
9	3.0	7.0			38.0		12.0	
10	5.7			none	40.0	empl_contr		
11	3.5	4.0	4.6	none	36.0			
12	6.4	6.4			38.0			
13	3.5	4.0		none	40.0			
14	3.5	4.0	5.1	tcf	37.0			
15	3.0			none	36.0			
16	4.5	4.0		none	37.0	empl_contr		
17	2.8				35.0			
18	2.1			tc	40.0	ret_allw	2.0	
19	2.0			none	38.0	none		
20	4.0	5.0		tcf	35.0		13.0	
21	4.3	4.4			38.0			
22	2.5	3.0			40.0	none		
23	3.5	4.0	4.6	tcf	27.0			

Add instance Undo OK Cancel

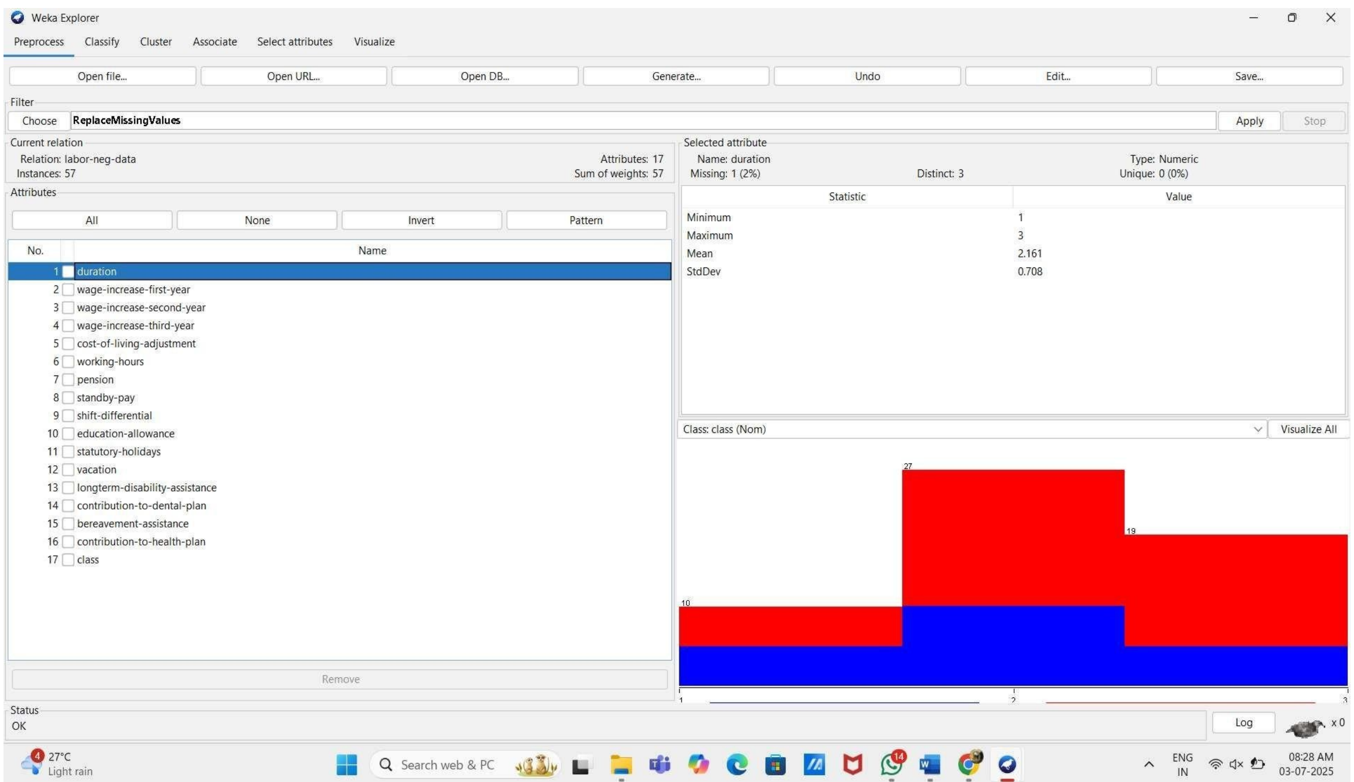
This screenshot shows the table format of dataset after removing duration attribute.

Filter 2: ReplaceMissingValues

The ReplaceMissingValues filter in Weka is an unsupervised filter used to automatically handle missing data in a dataset. When applied, it replaces any missing values in numeric attributes with the mean of the non-missing values, and for nominal (categorical) attributes, it replaces missing values with the mode (most frequent value). This filter is located under filters → unsupervised → attribute → ReplaceMissingValues. It is commonly used during the preprocessing stage to ensure that machine learning algorithms receive complete input data, as many models cannot handle missing values directly. This filter helps maintain dataset integrity while avoiding the loss of valuable data due to deletion of incomplete records.

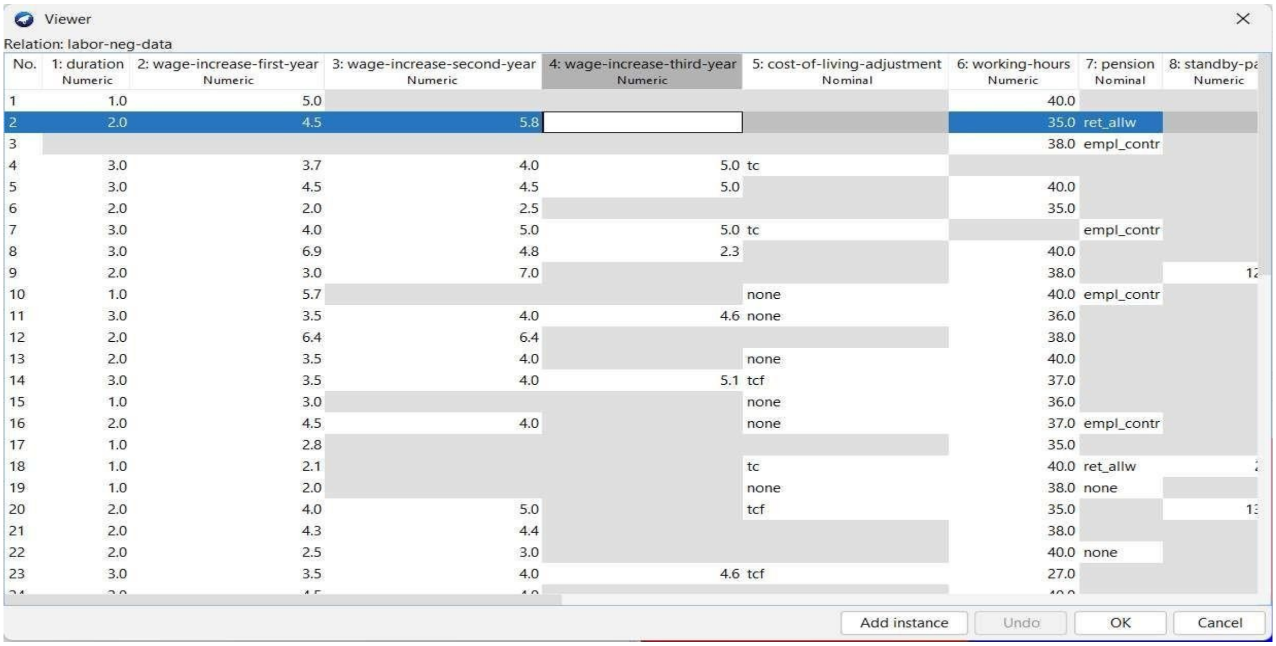
Dataset: labor.arff

Step-1: Upload dataset in Weka.



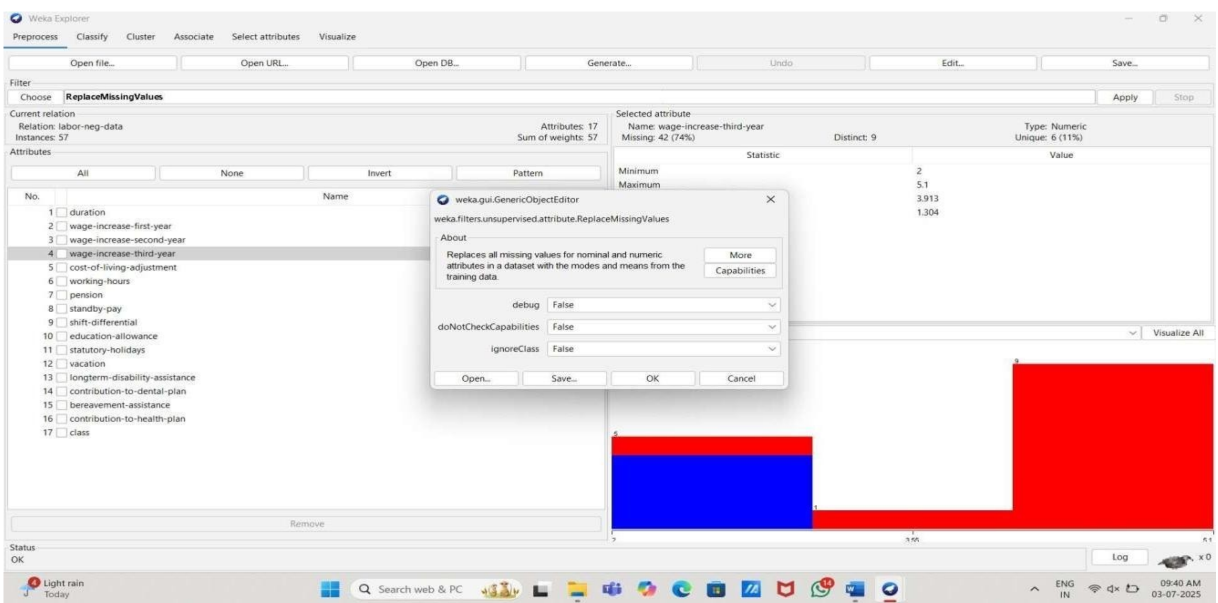
The labor.arff dataset is related to labor negotiations, typically involving attributes about employee benefits and working conditions. It consists of 57 instances (rows) and 17 attributes, including variables like duration, wage-increase, pension, vacation, contribution-to-health-plan, and more. The class attribute is usually a nominal variable indicating the outcome or status of the labor negotiation (e.g., "good" or "bad"). The dataset contains both numeric and nominal attributes, and may include missing values, as seen in the duration attribute.

Step-2: dataset in table format.



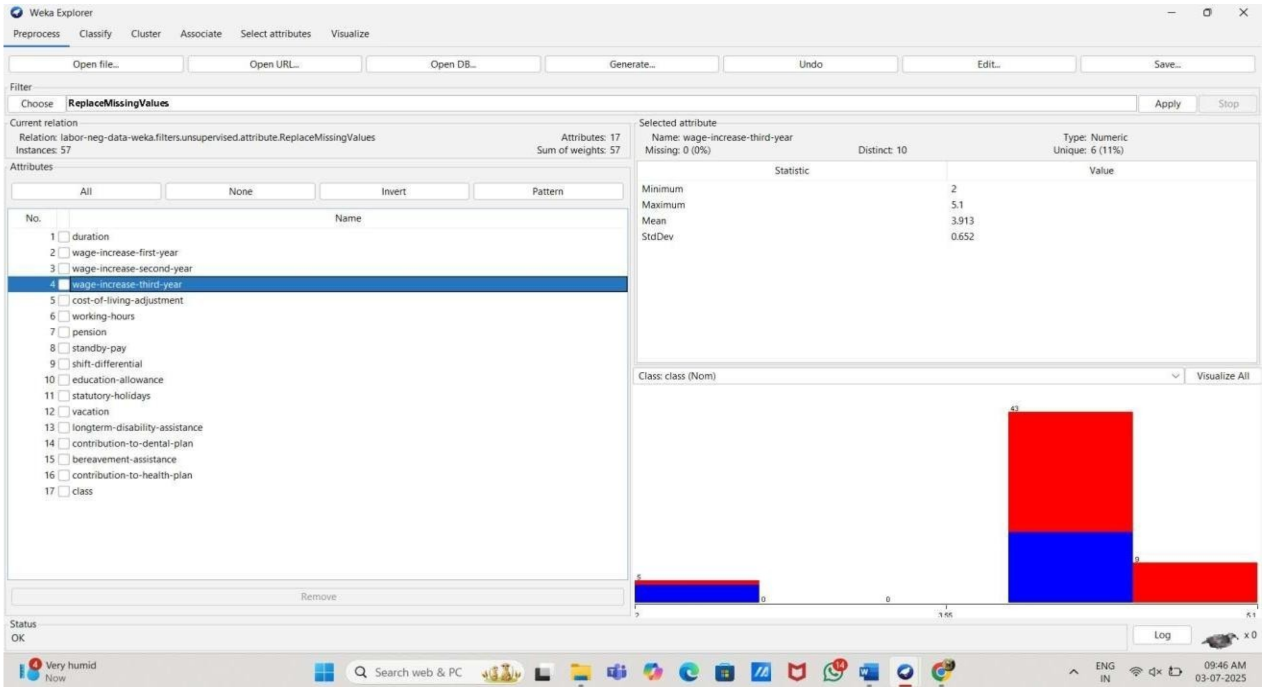
This screenshot shows the tabular view of the labor.arff dataset in Weka's Instance Viewer. Each row represents an instance (or record) from labor negotiations data, and each column is an attribute related to employment terms.

Step-3: Configure the parameters of ReplaceMissingValues filter.



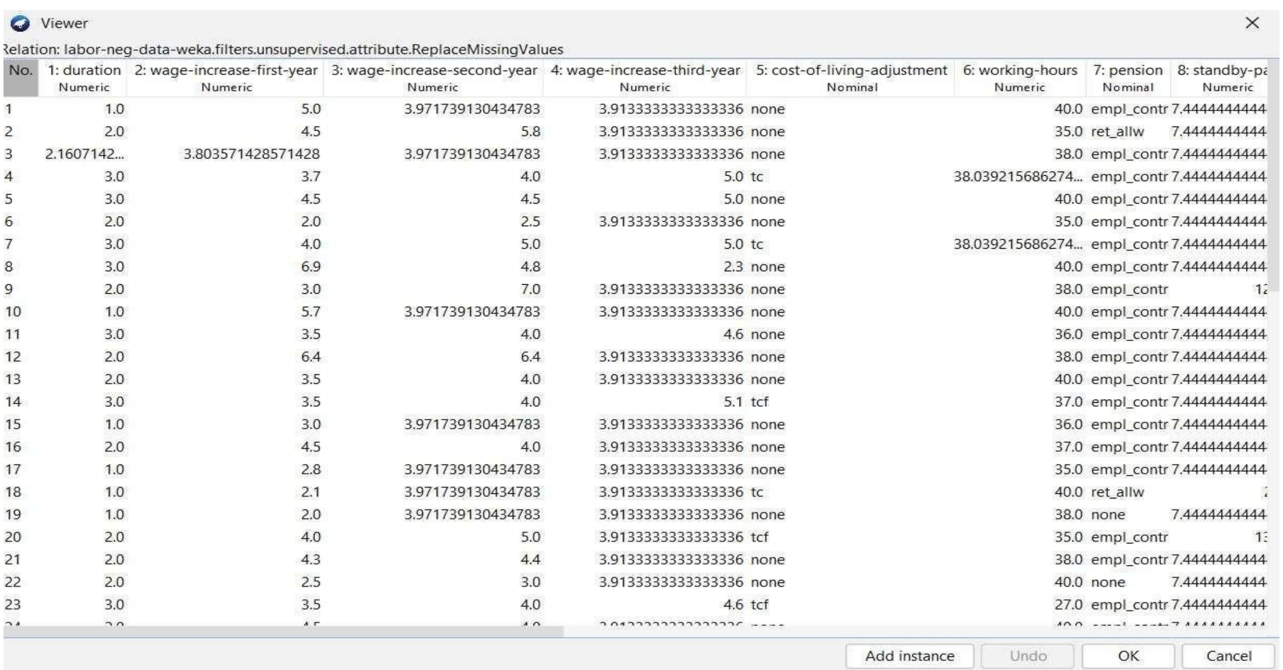
A filter named ReplaceMissingValues is selected to handle missing data in the attribute wage-increase- third-year, which has 42 missing values out of 57 instances. A configuration window for the filter is also open.

Step-4: Apply ReplaceMissingValues filter to wage-increase-third-year attribute.



This Screenshot shows that before applying filter wage-increase-third-year attribute has 42 missing values and after applying filter it becomes 0.

Step-5: Dataset in table format after applying Remove filter.


Viewer

Relation: labor-neg-data-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: duration Numeric	2: wage-increase-first-year Numeric	3: wage-increase-second-year Numeric	4: wage-increase-third-year Numeric	5: cost-of-living-adjustment Nominal	6: working-hours Numeric	7: pension Nominal	8: standby-pa Numeric
1	1.0	5.0	3.971739130434783	3.9133333333333336	none	40.0	empl_contr	7.4444444444
2	2.0	4.5	5.8	3.9133333333333336	none	35.0	ret_allw	7.4444444444
3	2.1607142...	3.803571428571428	3.971739130434783	3.9133333333333336	none	38.0	empl_contr	7.4444444444
4	3.0	3.7	4.0	5.0	tc	38.039215686274...	empl_contr	7.4444444444
5	3.0	4.5	4.5	5.0	none	40.0	empl_contr	7.4444444444
6	2.0	2.0	2.5	3.9133333333333336	none	35.0	empl_contr	7.4444444444
7	3.0	4.0	5.0	5.0	tc	38.039215686274...	empl_contr	7.4444444444
8	3.0	6.9	4.8	2.3	none	40.0	empl_contr	7.4444444444
9	2.0	3.0	7.0	3.9133333333333336	none	38.0	empl_contr	12
10	1.0	5.7	3.971739130434783	3.9133333333333336	none	40.0	empl_contr	7.4444444444
11	3.0	3.5	4.0	4.6	none	36.0	empl_contr	7.4444444444
12	2.0	6.4	6.4	3.9133333333333336	none	38.0	empl_contr	7.4444444444
13	2.0	3.5	4.0	3.9133333333333336	none	40.0	empl_contr	7.4444444444
14	3.0	3.5	4.0	5.1	tcf	37.0	empl_contr	7.4444444444
15	1.0	3.0	3.971739130434783	3.9133333333333336	none	36.0	empl_contr	7.4444444444
16	2.0	4.5	4.0	3.9133333333333336	none	37.0	empl_contr	7.4444444444
17	1.0	2.8	3.971739130434783	3.9133333333333336	none	35.0	empl_contr	7.4444444444
18	1.0	2.1	3.971739130434783	3.9133333333333336	tc	40.0	ret_allw	2
19	1.0	2.0	3.971739130434783	3.9133333333333336	none	38.0	none	7.4444444444
20	2.0	4.0	5.0	3.9133333333333336	tcf	35.0	empl_contr	13
21	2.0	4.3	4.4	3.9133333333333336	none	38.0	empl_contr	7.4444444444
22	2.0	2.5	3.0	3.9133333333333336	none	40.0	none	7.4444444444
23	3.0	3.5	4.0	4.6	tcf	27.0	empl_contr	7.4444444444

Add instance | Undo | OK | Cancel

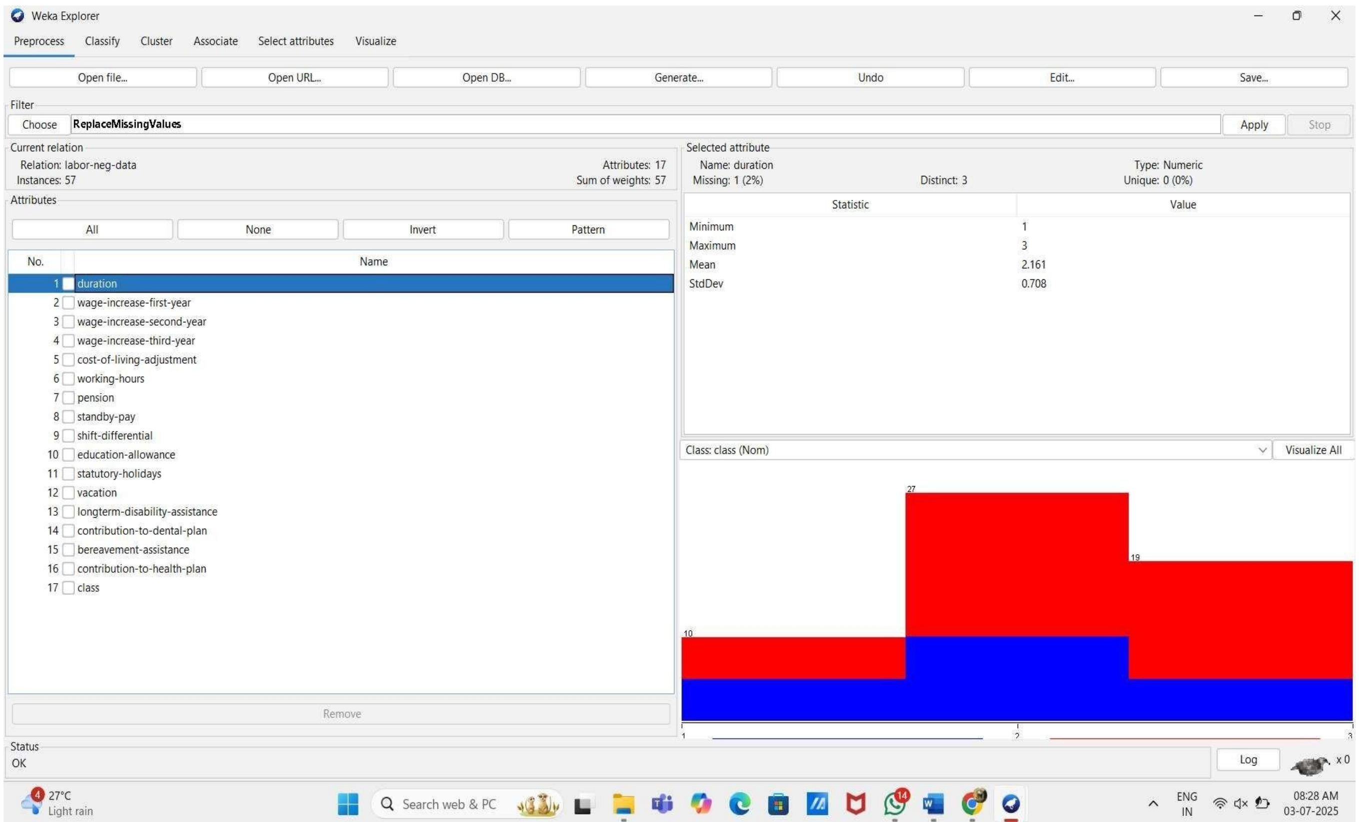
This screenshot shows the table format of dataset after Replacing missing values of wage-increase-third-year attribute.

Filter 3: ReplaceMissingWithUserConstant

The ReplaceMissingWithUserConstant filter in WEKA replaces all missing values in a dataset with a user- specified constant. It allows setting different constants for nominal and numeric attributes, providing control over how missing data is handled.

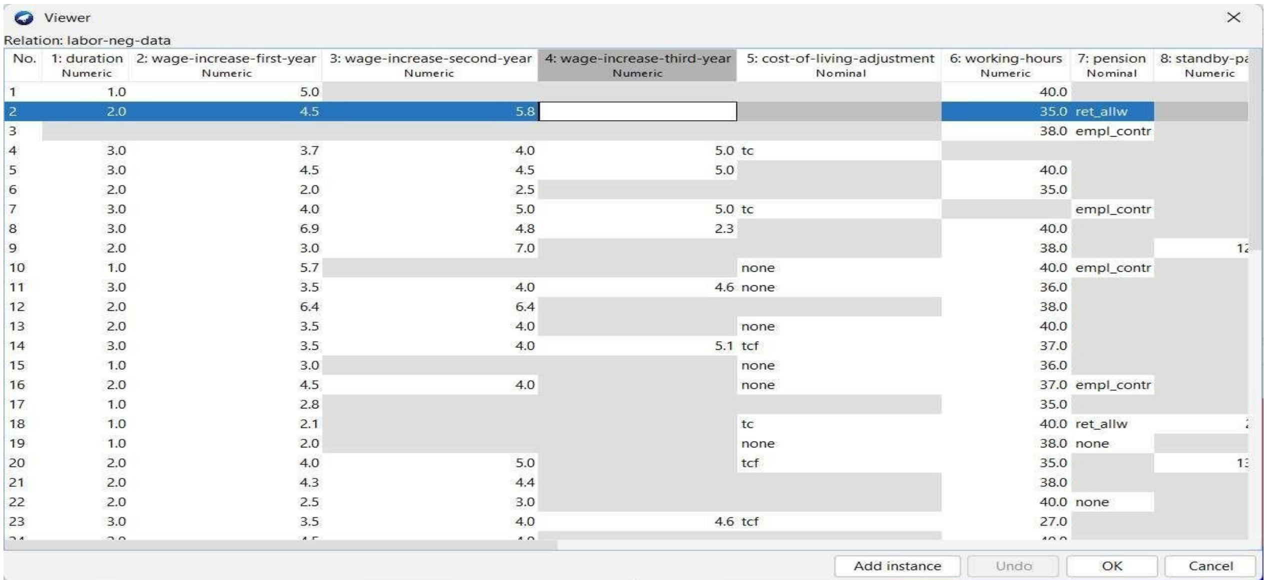
Dataset: labor.arff

Step-1: Upload dataset in Weka.



The labor.arff dataset is related to labor negotiations, typically involving attributes about employee benefits and working conditions. It consists of 57 instances (rows) and 17 attributes, including variables like duration, wage-increase, pension, vacation, contribution- to-health-plan, and more. The class attribute is usually a nominal variable indicating the outcome or status of the labor negotiation (e.g., "good" or "bad"). The dataset contains both numeric and nominal attributes, and may include missing values, as seen in the duration attribute.

Step-2: dataset in table format.

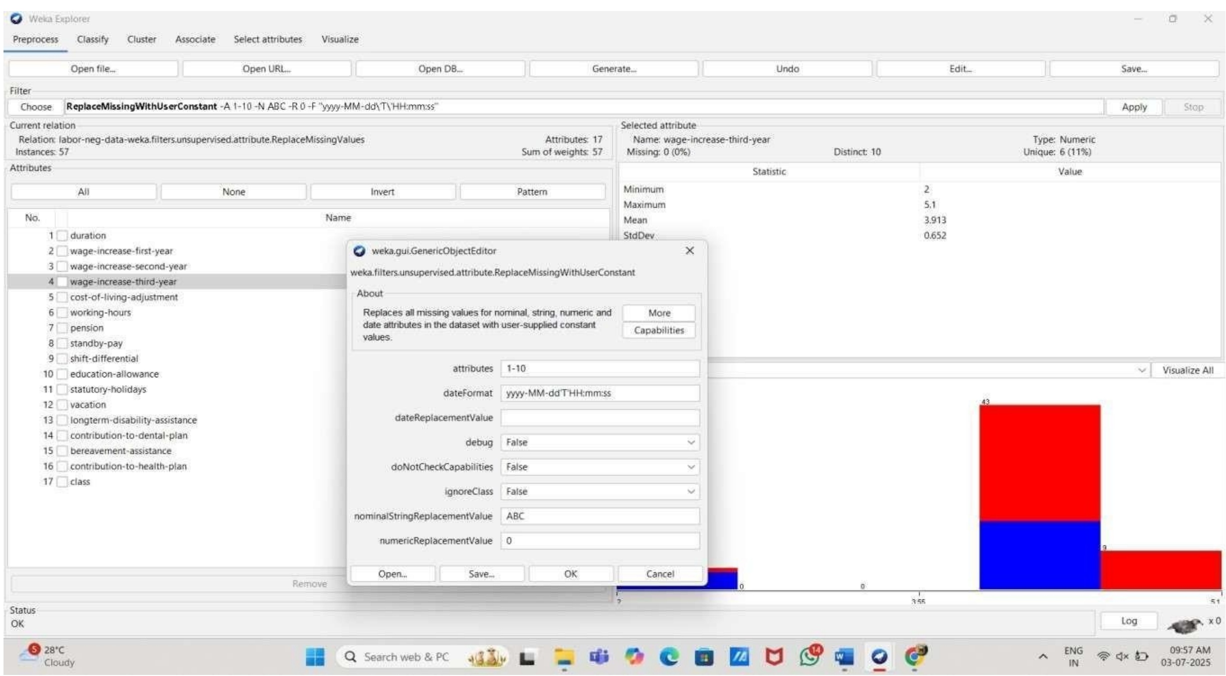

Viewer
Relation: labor-neg-data

No.	1: duration Numeric	2: wage-increase-first-year Numeric	3: wage-increase-second-year Numeric	4: wage-increase-third-year Numeric	5: cost-of-living-adjustment Nominal	6: working-hours Numeric	7: pension Nominal	8: standby-pa Numeric
1	1.0	5.0				40.0		
2	2.0	4.5	5.8			35.0	ret_allw	
3						38.0	empl_contr	
4	3.0	3.7	4.0	5.0	tc			
5	3.0	4.5	4.5	5.0		40.0		
6	2.0	2.0	2.5			35.0		
7	3.0	4.0	5.0	5.0	tc		empl_contr	
8	3.0	6.9	4.8	2.3		40.0		
9	2.0	3.0	7.0			38.0		1
10	1.0	5.7			none	40.0	empl_contr	
11	3.0	3.5	4.0	4.6	none	36.0		
12	2.0	6.4	6.4			38.0		
13	2.0	3.5	4.0		none	40.0		
14	3.0	3.5	4.0	5.1	tcf	37.0		
15	1.0	3.0			none	36.0		
16	2.0	4.5	4.0		none	37.0	empl_contr	
17	1.0	2.8				35.0		
18	1.0	2.1			tc	40.0	ret_allw	
19	1.0	2.0			none	38.0	none	
20	2.0	4.0	5.0		tcf	35.0		1
21	2.0	4.3	4.4			38.0		
22	2.0	2.5	3.0			40.0	none	
23	3.0	3.5	4.0	4.6	tcf	27.0		

Add instance Undo OK Cancel

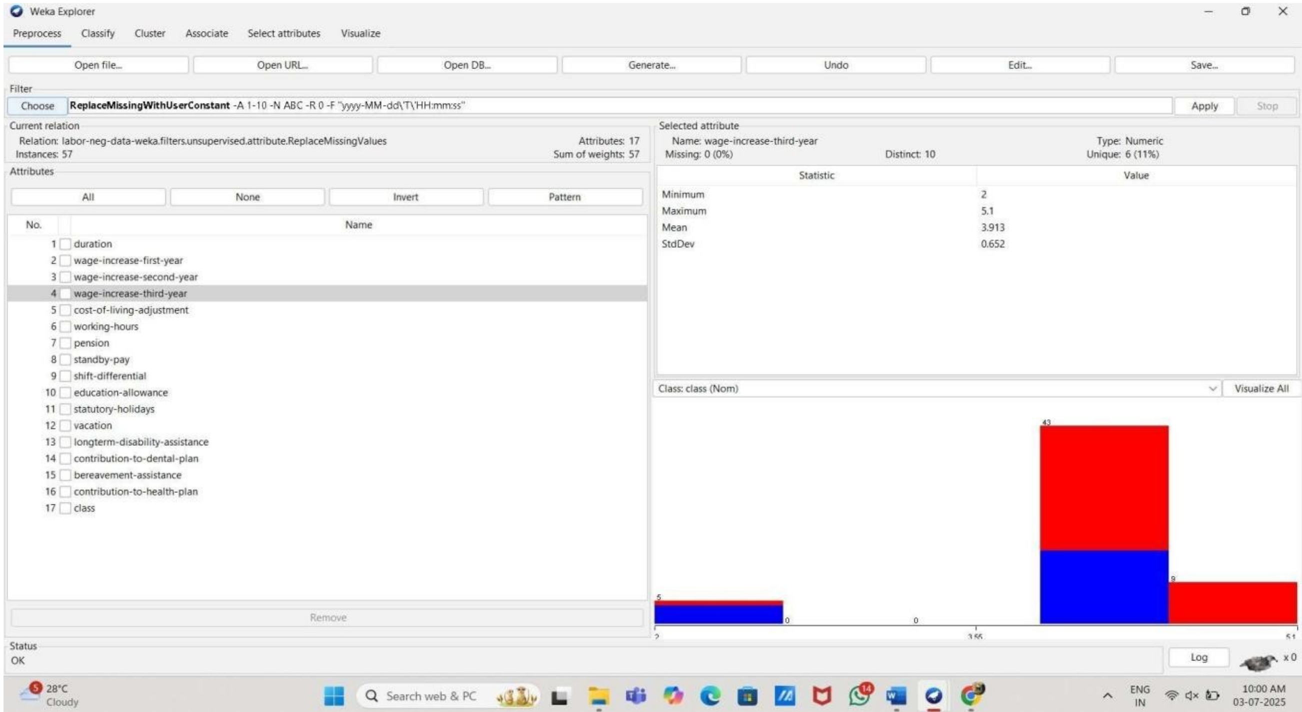
This screenshot shows the tabular view of the labor.arff dataset in Weka's Instance Viewer. Each row represents an instance (or record) from labor negotiations data, and each column is an attribute related to employment terms.

Step-3: Configure the parameters of ReplaceMissingWithUserConstant filter.



This image shows WEKA Explorer using the ReplaceMissingWithUserConstant filter, configured to replace missing numeric values with 0 and nominal/string values with "ABC". The attribute wage- increase-third-year has no missing values.

Step-4: Apply ReplaceMissingWithUserConstant filter from 1 to 10 attribute.



This Screenshot shows that ReplaceMissingWithUserConstant to 1-10 attributes and when I click on Apply then missing values are replaced by ABC.

Step-5: Dataset in table format after applying Remove filter.

Viewer

Relation: labor-neg-data-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.ReplaceMissingWithUserConstant-A1-10-NABC-R0-Fyyyy-MM-dd'T'H

No.	1: duration Numeric	2: wage-increase-first-year Numeric	3: wage-increase-second-year Numeric	4: wage-increase-third-year Numeric	5: cost-of-living-adjustment Nominal	6: working-hours Numeric	7: pension Nominal	8: standby-pa Numeric
1	1.0	5.0	3.971739130434783	3.9133333333333336	none	40.0	empl_contr	7.4444444444
2	2.0	4.5	5.8	3.9133333333333336	none	35.0	ret_allw	7.4444444444
3	2.1607142...	3.803571428571428	3.971739130434783	3.9133333333333336	none	38.0	empl_contr	7.4444444444
4	3.0	3.7	4.0	5.0	tc	38.039215686274...	empl_contr	7.4444444444
5	3.0	4.5	4.5	5.0	none	40.0	empl_contr	7.4444444444
6	2.0	2.0	2.5	3.9133333333333336	none	35.0	empl_contr	7.4444444444
7	3.0	4.0	5.0	5.0	tc	38.039215686274...	empl_contr	7.4444444444
8	3.0	6.9	4.8	2.3	none	40.0	empl_contr	7.4444444444
9	2.0	3.0	7.0	3.9133333333333336	none	38.0	empl_contr	12
10	1.0	5.7	3.971739130434783	3.9133333333333336	none	40.0	empl_contr	7.4444444444
11	3.0	3.5	4.0	4.6	none	36.0	empl_contr	7.4444444444
12	2.0	6.4	6.4	3.9133333333333336	none	38.0	empl_contr	7.4444444444
13	2.0	3.5	4.0	3.9133333333333336	none	40.0	empl_contr	7.4444444444
14	3.0	3.5	4.0	5.1	tcf	37.0	empl_contr	7.4444444444
15	1.0	3.0	3.971739130434783	3.9133333333333336	none	36.0	empl_contr	7.4444444444
16	2.0	4.5	4.0	3.9133333333333336	none	37.0	empl_contr	7.4444444444
17	1.0	2.8	3.971739130434783	3.9133333333333336	none	35.0	empl_contr	7.4444444444
18	1.0	2.1	3.971739130434783	3.9133333333333336	tc	40.0	ret_allw	2
19	1.0	2.0	3.971739130434783	3.9133333333333336	none	38.0	none	7.4444444444
20	2.0	4.0	5.0	3.9133333333333336	tcf	35.0	empl_contr	13
21	2.0	4.3	4.4	3.9133333333333336	none	38.0	empl_contr	7.4444444444
22	2.0	2.5	3.0	3.9133333333333336	none	40.0	none	7.4444444444
23	3.0	3.5	4.0	4.6	tcf	27.0	empl_contr	7.4444444444
24	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Add instance | Undo | OK | Cancel

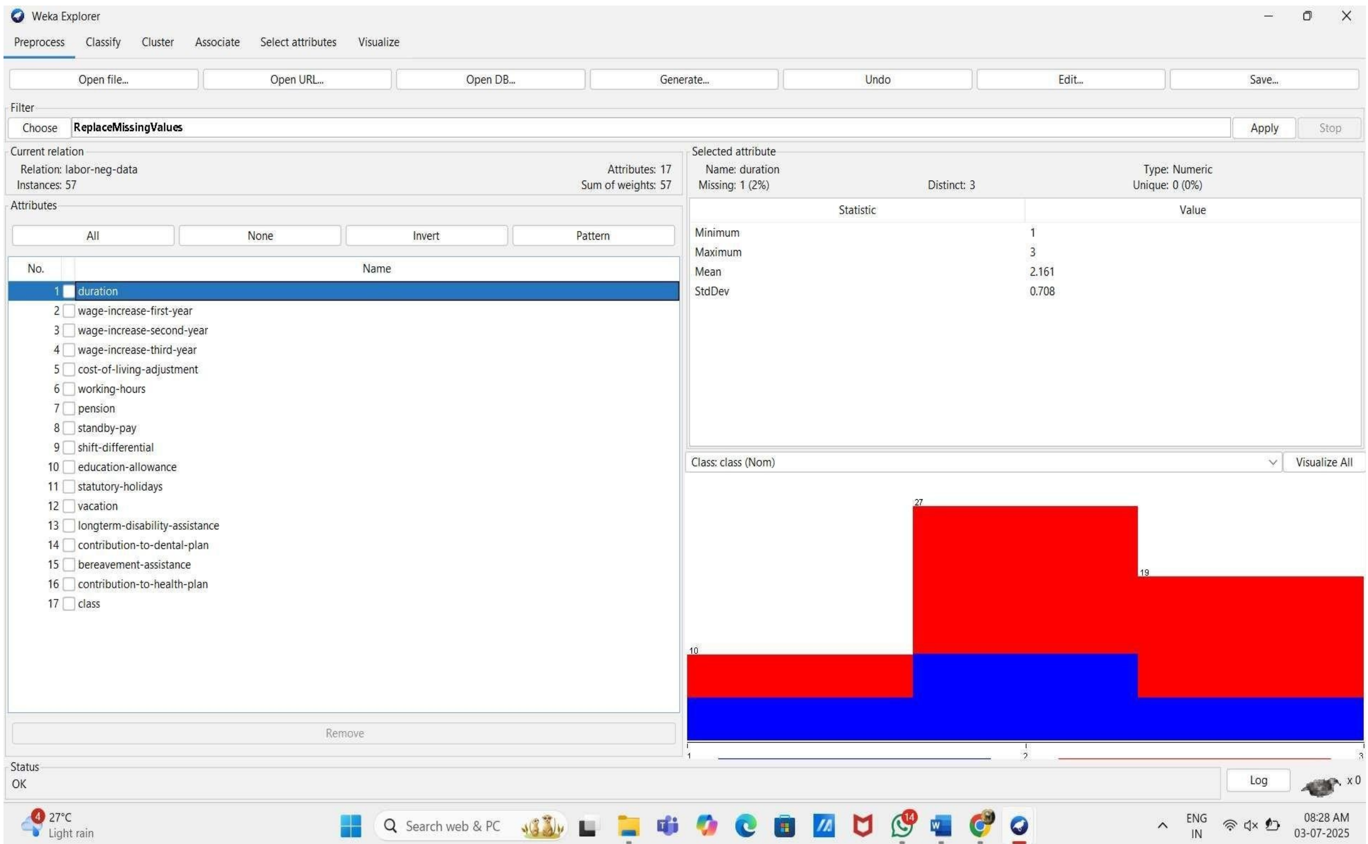
This screenshot shows the table format of dataset after Replacing missing values of wage- increase-third-year attribute.

Filter 4: ReplaceWithMissingValue

The ReplaceWithMissingValue filter in WEKA replaces specified attribute values with missing values. It is useful for simulating missing data or reverting imputed values back to missing for testing or preprocessing purposes.

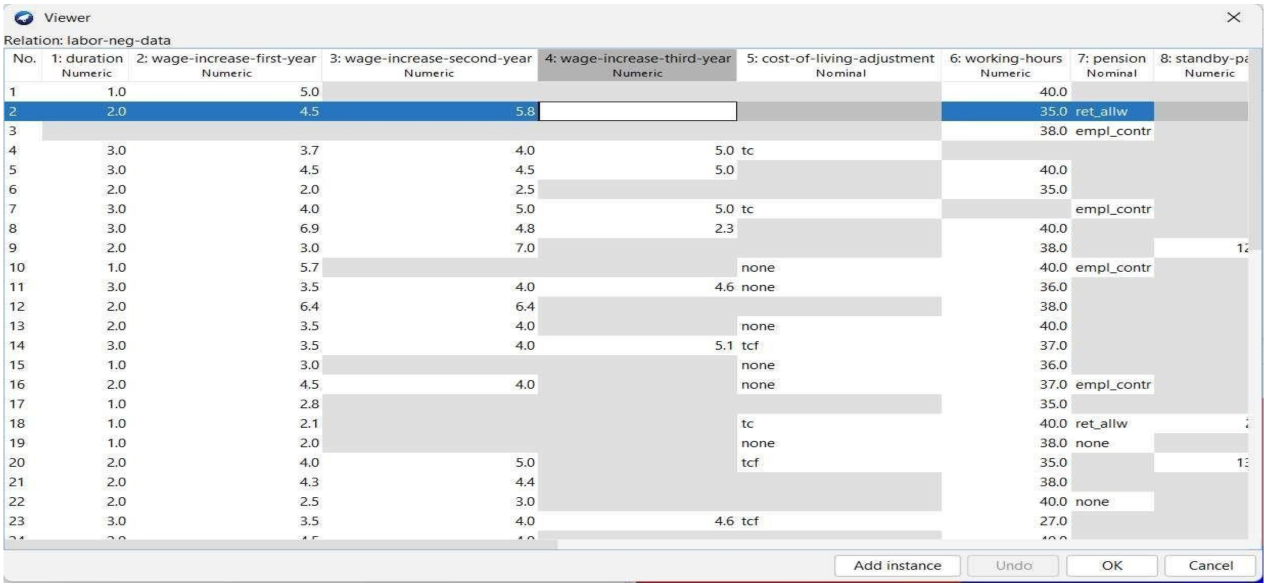
Dataset: labor.arff

Step-1: Upload dataset in Weka.



The labor.arff dataset is related to labor negotiations, typically involving attributes about employee benefits and working conditions. It consists of 57 instances (rows) and 17 attributes, including variables like duration, wage-increase, pension, vacation, contribution- to-health-plan, and more. The class attribute is usually a nominal variable indicating the outcome or status of the labor negotiation (e.g., "good" or "bad"). The dataset contains both numeric and nominal attributes, and may include missing values, as seen in the duration attribute.

Step-2: dataset in table format.

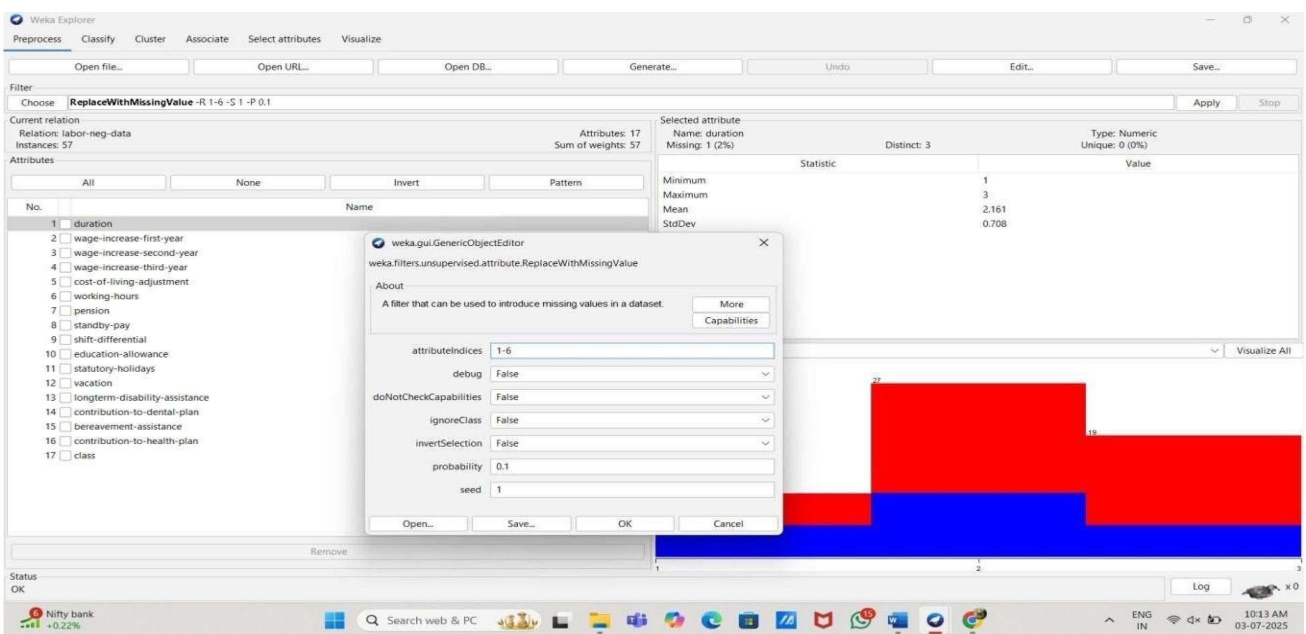

Viewer
Relation: labor-neg-data

No.	1: duration Numeric	2: wage-increase-first-year Numeric	3: wage-increase-second-year Numeric	4: wage-increase-third-year Numeric	5: cost-of-living-adjustment Nominal	6: working-hours Numeric	7: pension Nominal	8: standby-pa Numeric
1	1.0	5.0				40.0		
2	2.0	4.5	5.8			35.0	ret_allw	
3						38.0	empl_contr	
4	3.0	3.7	4.0	5.0	tc			
5	3.0	4.5	4.5	5.0		40.0		
6	2.0	2.0	2.5			35.0		
7	3.0	4.0	5.0	5.0	tc		empl_contr	
8	3.0	6.9	4.8	2.3		40.0		
9	2.0	3.0	7.0			38.0		[illegible]
10	1.0	5.7			none	40.0	empl_contr	
11	3.0	3.5	4.0	4.6	none	36.0		
12	2.0	6.4	6.4			38.0		
13	2.0	3.5	4.0		none	40.0		
14	3.0	3.5	4.0	5.1	tcf	37.0		
15	1.0	3.0			none	36.0		
16	2.0	4.5	4.0		none	37.0	empl_contr	
17	1.0	2.8				35.0		
18	1.0	2.1			tc	40.0	ret_allw	[illegible]
19	1.0	2.0			none	38.0	none	
20	2.0	4.0	5.0		tcf	35.0		[illegible]
21	2.0	4.3	4.4			38.0		
22	2.0	2.5	3.0			40.0	none	
23	3.0	3.5	4.0	4.6	tcf	27.0		

Add instance | Undo | OK | Cancel

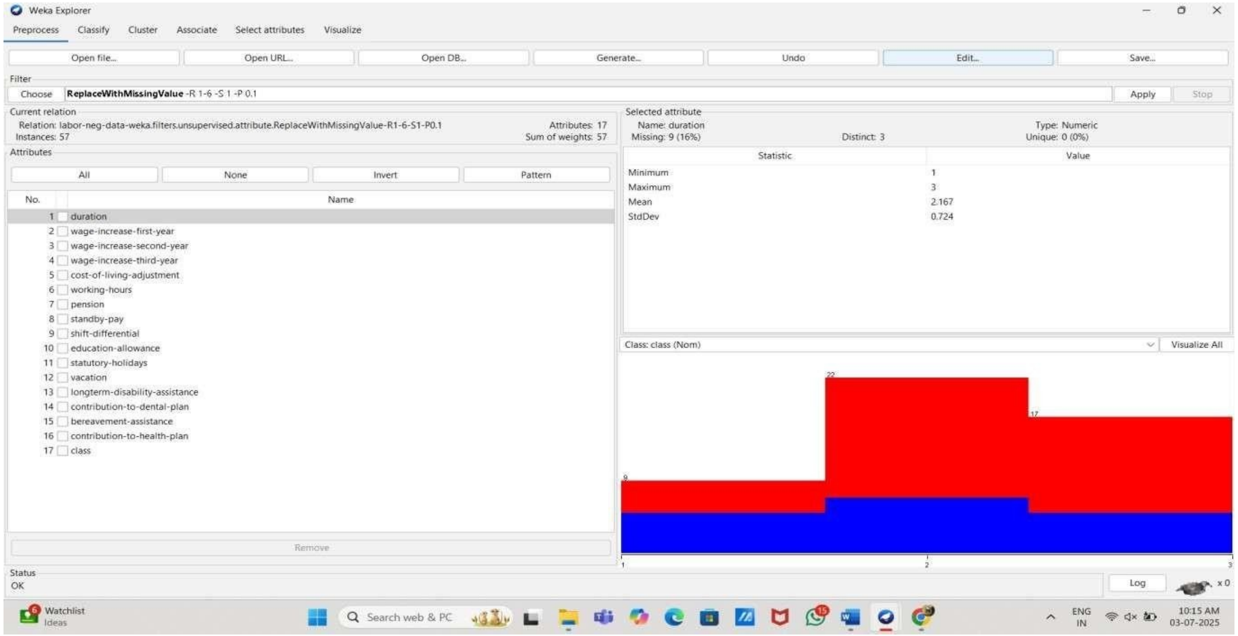
This screenshot shows the tabular view of the labor.arff dataset in Weka's Instance Viewer. Each row represents an instance (or record) from labor negotiations data, and each column is an attribute related to employment terms.

Step-3: Configure the parameters of ReplaceWithMissingValue filter.



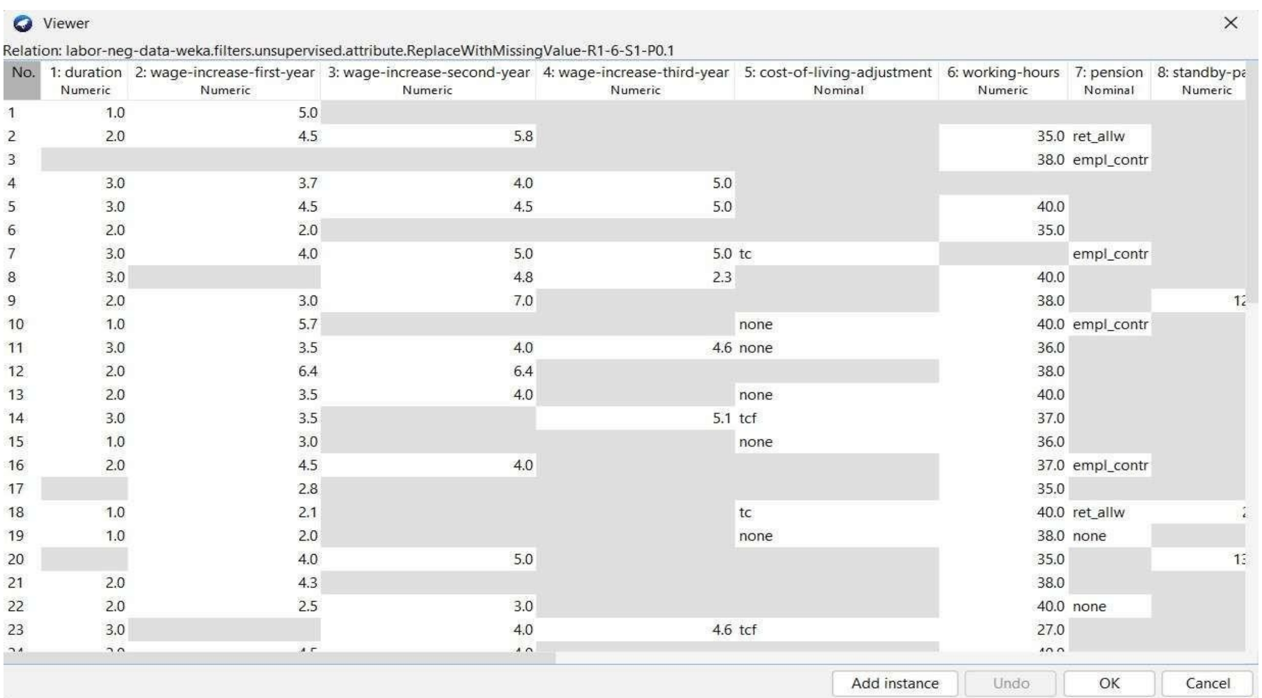
This image shows WEKA Explorer using the ReplaceWithMissingValue filter to introduce missing values in attributes 1 to 6 with a 10% probability. The attribute duration now has 1 missing value (2%).

Step-4: Apply ReplaceWithMissingValue filter from 1 to 6 attribute.



This Screenshot shows that ReplaceWithMissingValue to 1-6 attributes and when I click on Apply then values are replaced by missing values.

Step-5: Dataset in table format after applying Remove filter .



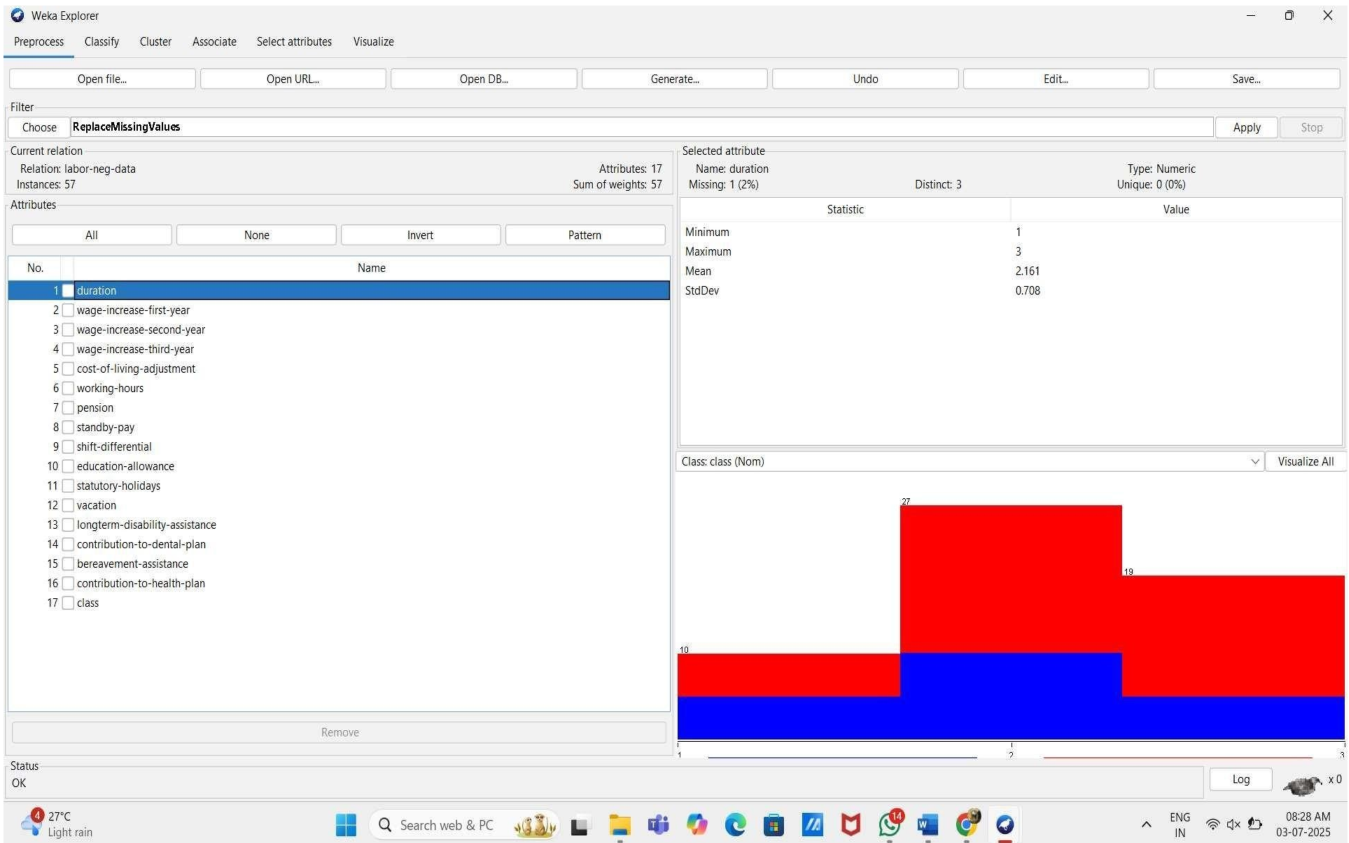
This screenshot shows the table format of dataset in which values are replaced by missing values.

Filter 5: Descritize

The Discretize filter in WEKA converts numeric attributes into nominal ones by dividing their range into intervals or bins. This is useful for algorithms that require categorical input or for simplifying data analysis. Binning can be done using equal-width or equal-frequency methods.

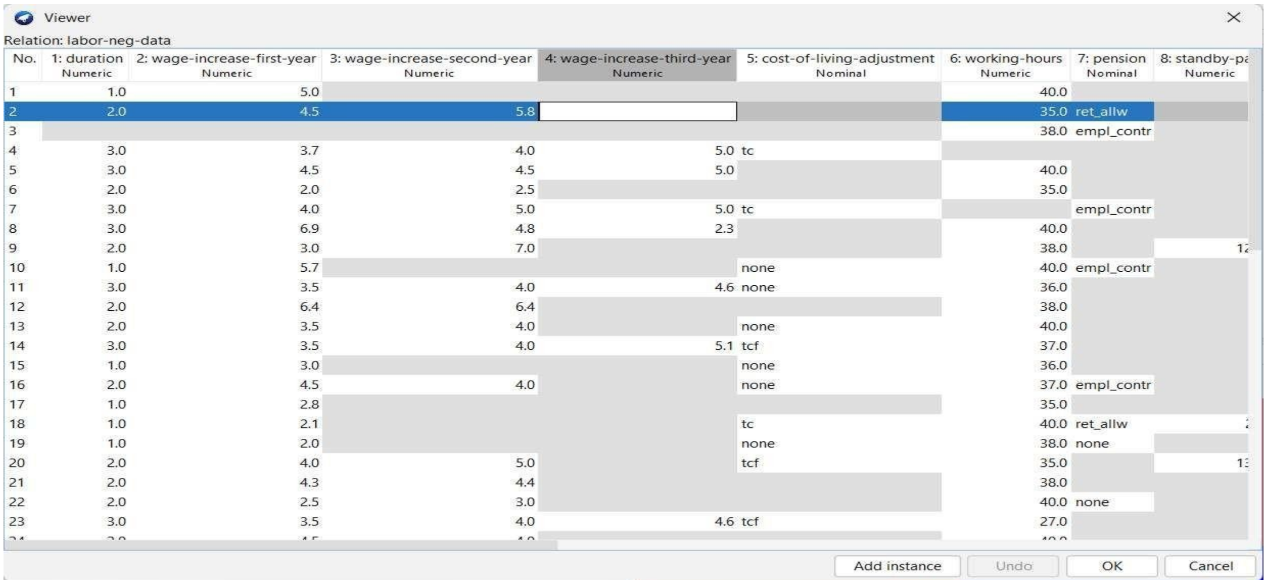
Dataset: labor.arff

Step-1: Upload dataset in Weka.



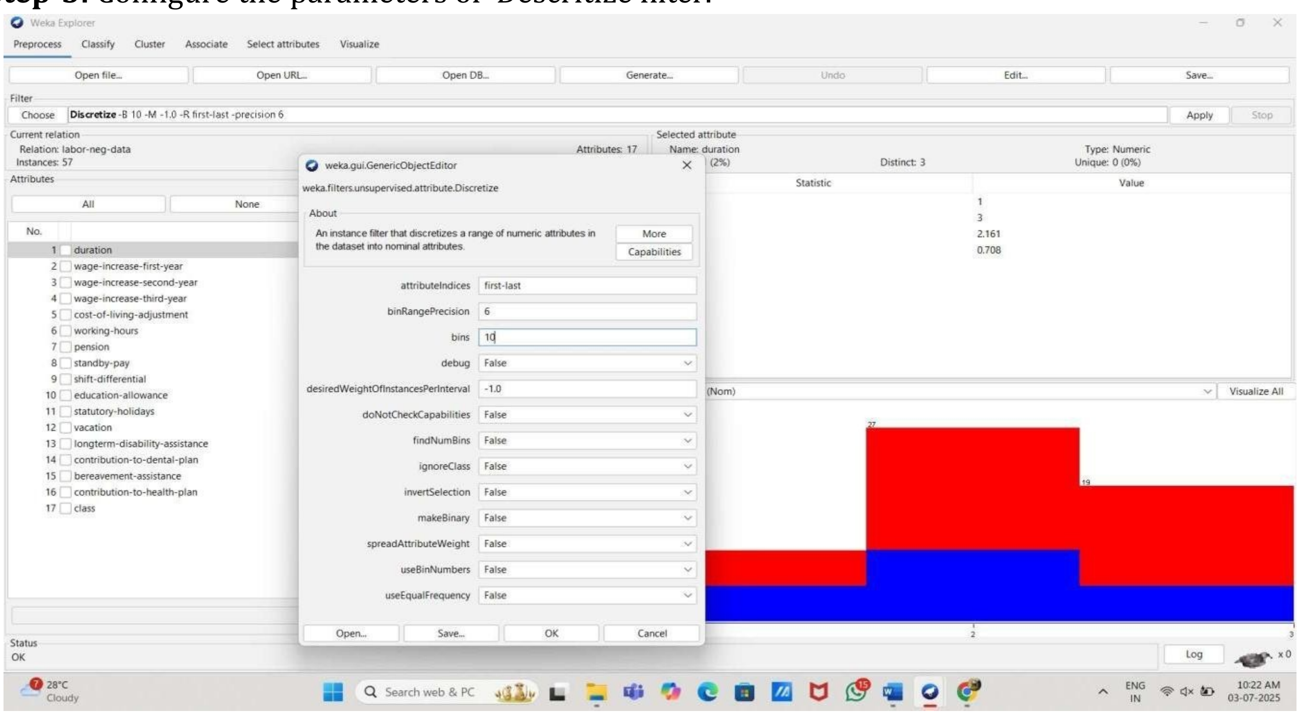
The labor.arff dataset is related to labor negotiations, typically involving attributes about employee benefits and working conditions. It consists of 57 instances (rows) and 17 attributes, including variables like duration, wage-increase, pension, vacation, contribution- to-health-plan, and more. The class attribute is usually a nominal variable indicating the outcome or status of the labor negotiation (e.g., "good" or "bad"). The dataset contains both numeric and nominal attributes, and may include missing values, as seen in the duration attribute.

Step-2: dataset in table format.



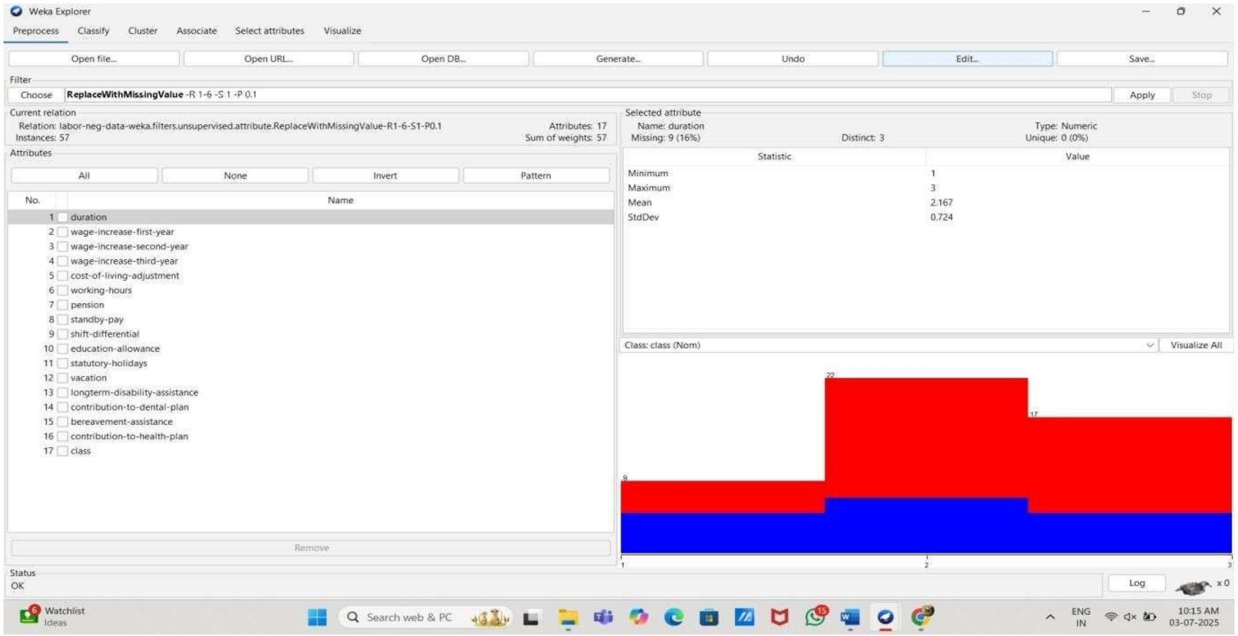
This screenshot shows the tabular view of the labor.arff dataset in Weka's Instance Viewer. Each row represents an instance (or record) from labor negotiations data, and each column is an attribute related to employment terms.

Step-3: Configure the parameters of Descritize filter.



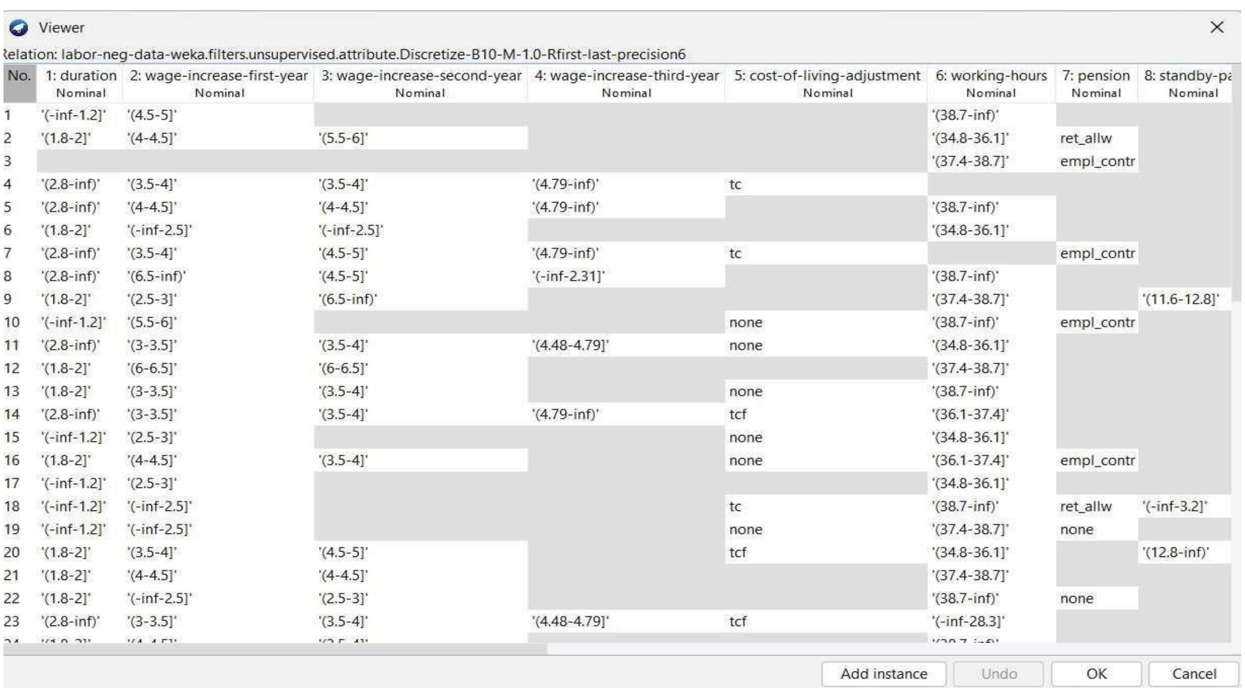
This image shows the use of the Discretize filter in WEKA, configured to convert numeric attributes into nominal ones using 10 bins with a bin range precision of 6, applied to all attributes in the dataset.

Step-4: Apply descritize filter from to all attribute.



This Screenshot shows that descritize to all attributes and attributes are divided in 10 bins.

Step-5: Dataset in table format after applying Remove filter .



Viewer

Relation: labor-neg-data-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6

No.	1: duration Nominal	2: wage-increase-first-year Nominal	3: wage-increase-second-year Nominal	4: wage-increase-third-year Nominal	5: cost-of-living-adjustment Nominal	6: working-hours Nominal	7: pension Nominal	8: standby-pa Nominal
1	'(-inf-1.2]'	'(4.5-5]'				'(38.7-inf)'		
2	'(1.8-2]'	'(4-4.5]'	'(5.5-6]'			'(34.8-36.1]'	ret_allw	
3						'(37.4-38.7]'	empl_contr	
4	'(2.8-inf)'	'(3.5-4]'	'(3.5-4]'	'(4.79-inf)'	tc			
5	'(2.8-inf)'	'(4-4.5]'	'(4-4.5]'	'(4.79-inf)'		'(38.7-inf)'		
6	'(1.8-2]'	'(-inf-2.5]'	'(-inf-2.5]'			'(34.8-36.1]'		
7	'(2.8-inf)'	'(3.5-4]'	'(4.5-5]'	'(4.79-inf)'	tc		empl_contr	
8	'(2.8-inf)'	'(6.5-inf)'	'(4.5-5]'	'(-inf-2.31]'		'(38.7-inf)'		
9	'(1.8-2]'	'(2.5-3]'	'(6.5-inf)'			'(37.4-38.7]'		'(11.6-12.8]'
10	'(-inf-1.2]'	'(5.5-6]'			none	'(38.7-inf)'	empl_contr	
11	'(2.8-inf)'	'(3-3.5]'	'(3.5-4]'	'(4.48-4.79]'	none	'(34.8-36.1]'		
12	'(1.8-2]'	'(6-6.5]'	'(6-6.5]'			'(37.4-38.7]'		
13	'(1.8-2]'	'(3-3.5]'	'(3.5-4]'		none	'(38.7-inf)'		
14	'(2.8-inf)'	'(3-3.5]'	'(3.5-4]'	'(4.79-inf)'	tcf	'(36.1-37.4]'		
15	'(-inf-1.2]'	'(2.5-3]'			none	'(34.8-36.1]'		
16	'(1.8-2]'	'(4-4.5]'	'(3.5-4]'		none	'(36.1-37.4]'	empl_contr	
17	'(-inf-1.2]'	'(2.5-3]'				'(34.8-36.1]'		
18	'(-inf-1.2]'	'(-inf-2.5]'			tc	'(38.7-inf)'	ret_allw	'(-inf-3.2]'
19	'(-inf-1.2]'	'(-inf-2.5]'			none	'(37.4-38.7]'	none	
20	'(1.8-2]'	'(3.5-4]'	'(4.5-5]'		tcf	'(34.8-36.1]'		'(12.8-inf)'
21	'(1.8-2]'	'(4-4.5]'	'(4-4.5]'			'(37.4-38.7]'		
22	'(1.8-2]'	'(-inf-2.5]'	'(2.5-3]'			'(38.7-inf)'	none	
23	'(2.8-inf)'	'(3-3.5]'	'(3.5-4]'	'(4.48-4.79]'	tcf	'(-inf-28.3]'		

Add instance | Undo | OK | Cancel

This screenshot shows the table format of dataset in which Range is provided to all attributes.

Experiment Outcome:

The aim of this experiment was to apply and analyze various preprocessing filters in WEKA using the labor.arff dataset. Filters like Remove, ReplaceMissingValues, ReplaceMissingWithUserConstant, ReplaceWithMissingValue, and Discretize were used to clean, transform, and prepare the data. The changes observed before and after applying each filter showed how preprocessing improves data quality and makes it more suitable for analysis and machine learning tasks.