

**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

**Name : Shruti Chaube**  
**Roll No : 66**

**Experiment No. : 2**

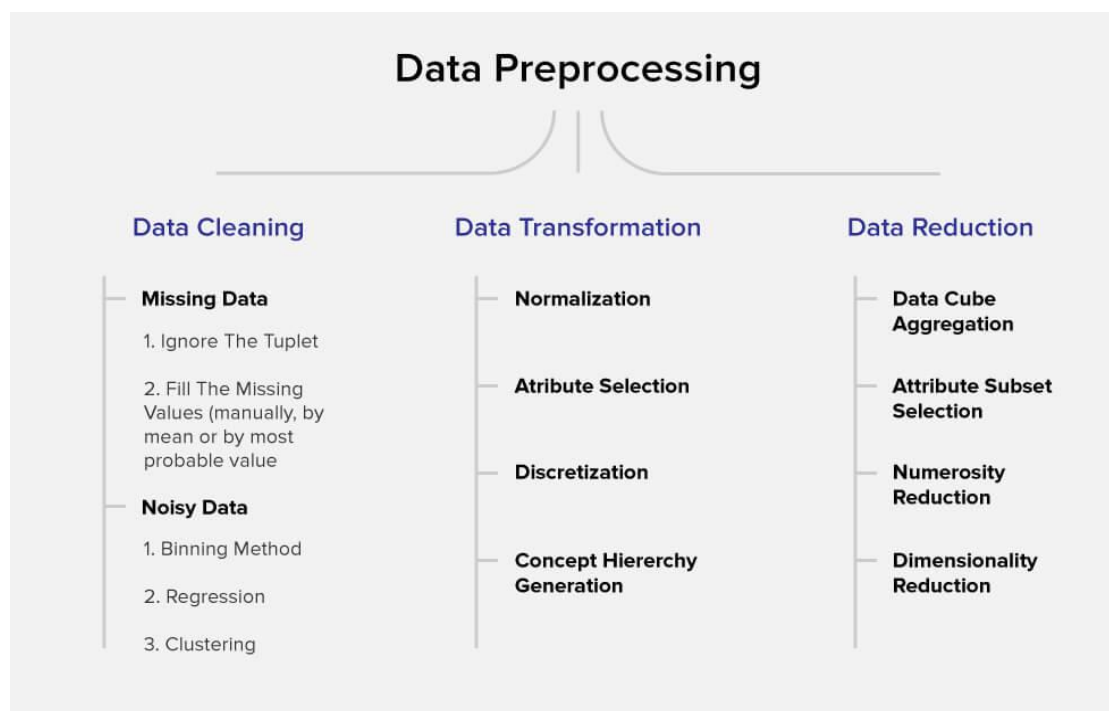
**Title :** Tutorial-2 Solving exercises in Data preprocessing

**Problem Definition :** Perform data preprocessing methods - attribute selection, discretization and filling missing values on “bank-data” file using WEKA.

**Pre-requisite :** Weka software, Data exploration

**Theory :**

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



**Steps Involved in Data Preprocessing:**

**1. Data Cleaning:**

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

**(a). Missing Data:**

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

**1.Ignore the tuples:**

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

**2.Fill the Missing values:**

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

**(b). Noisy Data:**

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

**1. Binning Method:**

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

**2. Regression:**

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

**3. Clustering:**

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

**1. Normalization:**

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

**2. Attribute Selection:**

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

**3. Discretization:**

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

**4. Concept Hierarchy Generation:**

Here attributes are converted from level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

**3. Data Reduction:**

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

**1. Data Cube Aggregation:**

Aggregation operation is applied to data for the construction of the data cube.

**2. Attribute Subset Selection:**

The highly relevant attributes should be used, rest all can be discarded. For

**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

performing attribute selection, one can use level of significance and p-value of the attribute. the attribute having p-value greater than significance level can be discarded.

**3. Numerosity Reduction:**

This enable to store the model of data instead of whole data, for example: Regression Models.

**4. Dimensionality Reduction:**

This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction.

**Procedure :**

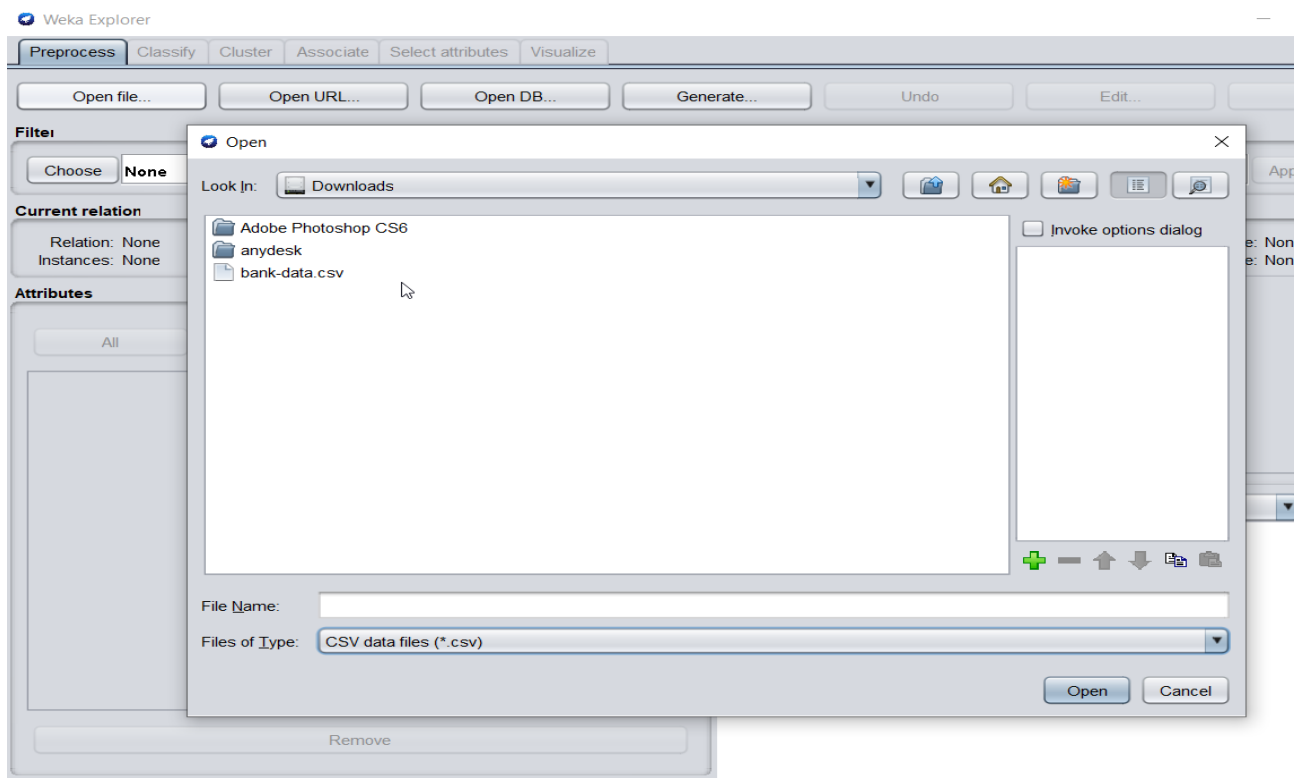
This exercise illustrates some of the basic data preprocessing operations that can be performed using WEKA. The sample data set used for this example is the "bank data" available in comma-separated format (bank-data.csv). Follow the instructions from the pdf file available at [https://www.researchgate.net/profile/.../datasets.../hw2\\_dataPreproc.pdf](https://www.researchgate.net/profile/.../datasets.../hw2_dataPreproc.pdf).

**Results :**

Attribute Selection

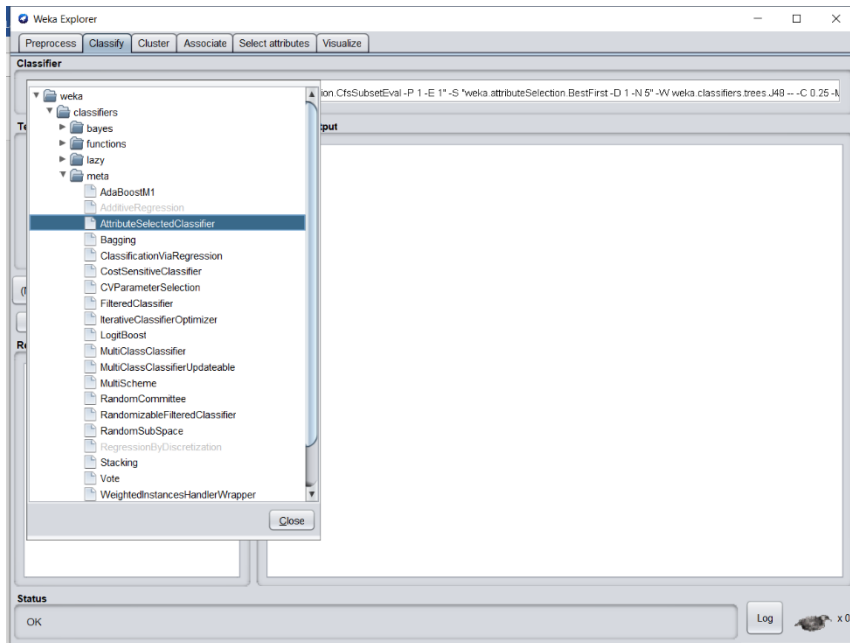
Steps:

1. Select the file

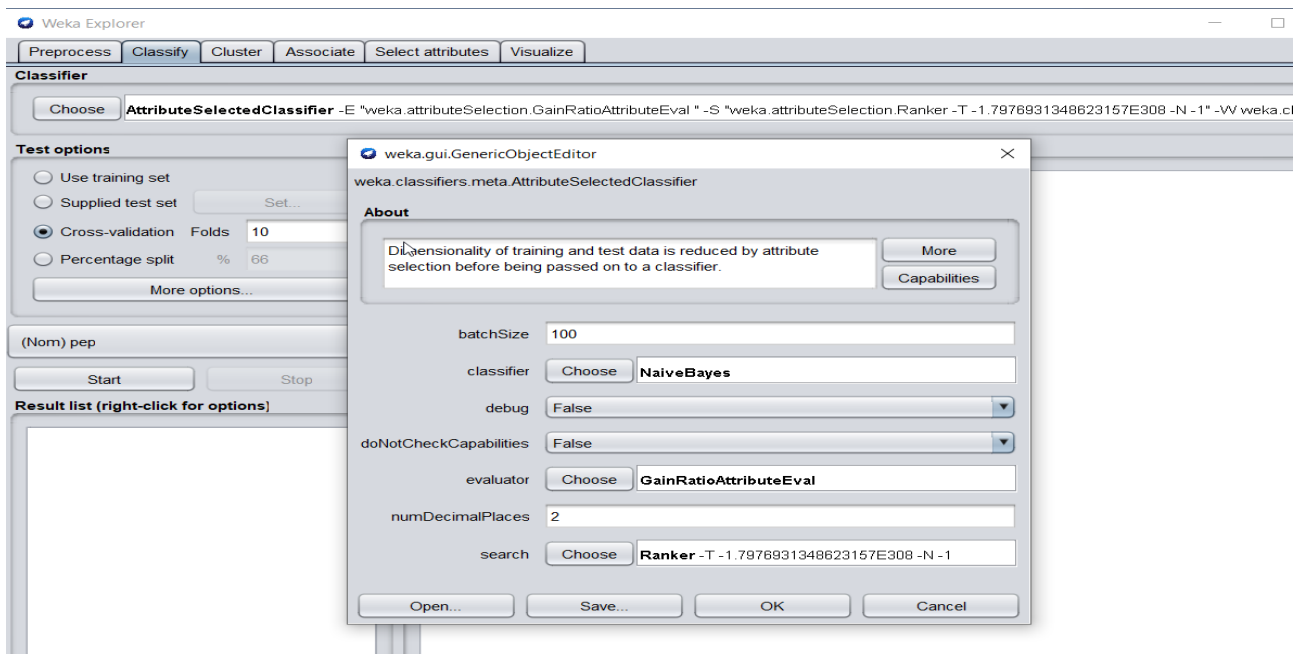


**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

2. Go to classify



3.



**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

OUTPUT:

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'AttributeSelectedClassifier'. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' pane displays the following results:

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===

=== Summary ===

Metric	Value
Correctly Classified Instances	385
Incorrectly Classified Instances	215
Kappa statistic	0.2722
Mean absolute error	0.427
Root mean squared error	0.4678
Relative absolute error	86.0487 %
Root relative squared error	93.9157 %
Total Number of Instances	600

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
0.558	0.288	0.619	0.558	0.587	0.273	0.707	0.652	
0.712	0.442	0.657	0.712	0.683	0.273	0.707	0.741	
Weighted Avg.	0.642	0.372	0.640	0.642	0.640	0.273	0.707	

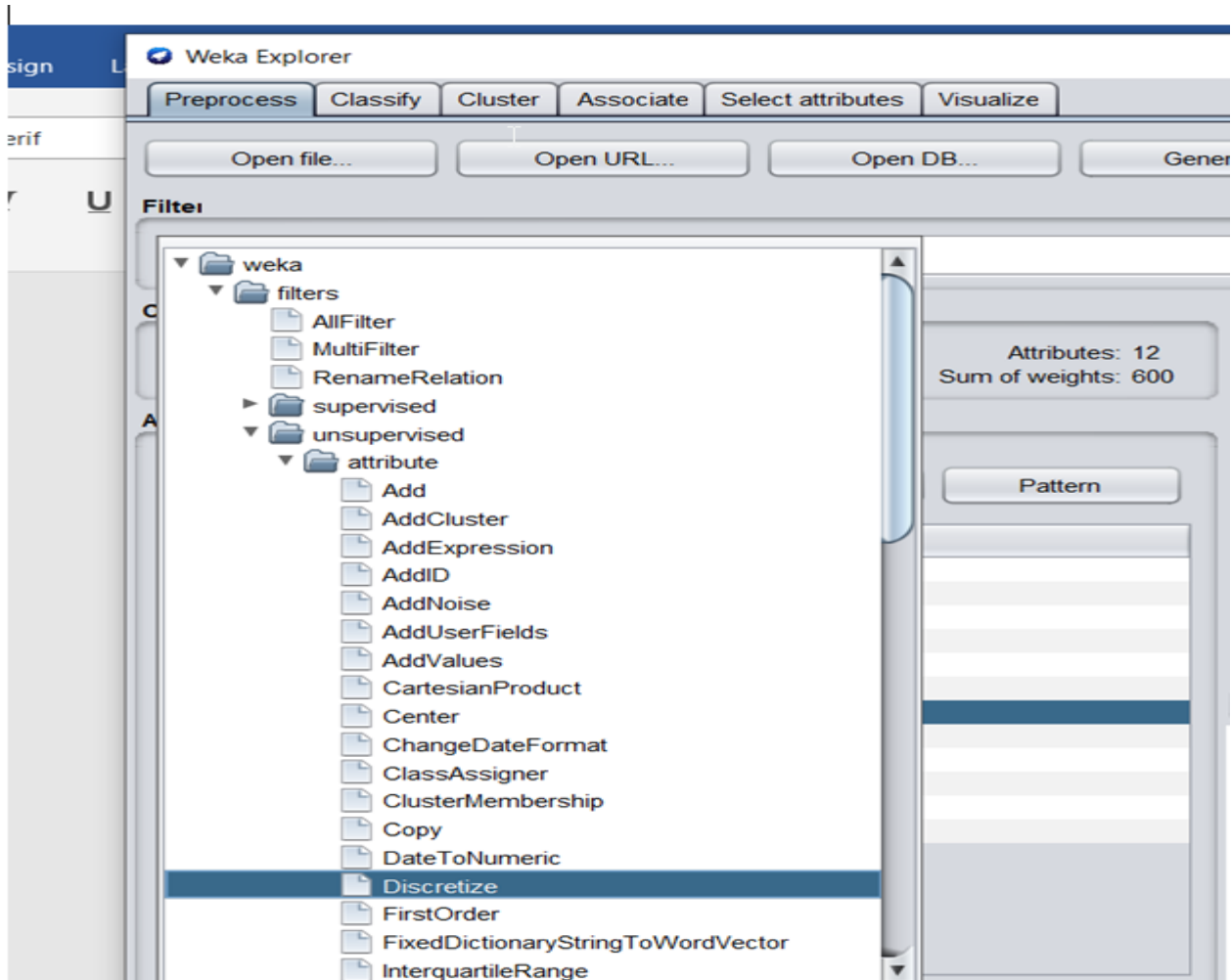
=== Confusion Matrix ===

```
a b <-- classified as
153 121 | a = YES
94 232 | b = NO
```

The 'Status' bar at the bottom shows 'OK'.

**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

2. Discretization:



**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

OUTPUT:

The screenshot shows the Weka Explorer Classifier window. The 'Classifier' dropdown is set to 'J48 -C 0.25 -M 2'. Under 'Test options', 'Cross-validation' is selected with 'Folds' set to 10. The 'Classifier output' pane displays the following results:

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===  
=== Summary ===

Correctly Classified Instances	512	85.3333 %
Incorrectly Classified Instances	88	14.6667 %
Kappa statistic	0.7034	
Mean absolute error	0.2142	
Root mean squared error	0.3439	
Relative absolute error	43.166 %	
Root relative squared error	69.0328 %	
Total Number of Instances	600	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.818	0.117	0.855	0.818	0.836	0.704	0.881	0.839	YES
	0.883	0.182	0.852	0.883	0.867	0.704	0.881	0.867	NO
Weighted Avg.	0.853	0.152	0.853	0.853	0.853	0.704	0.881	0.854	

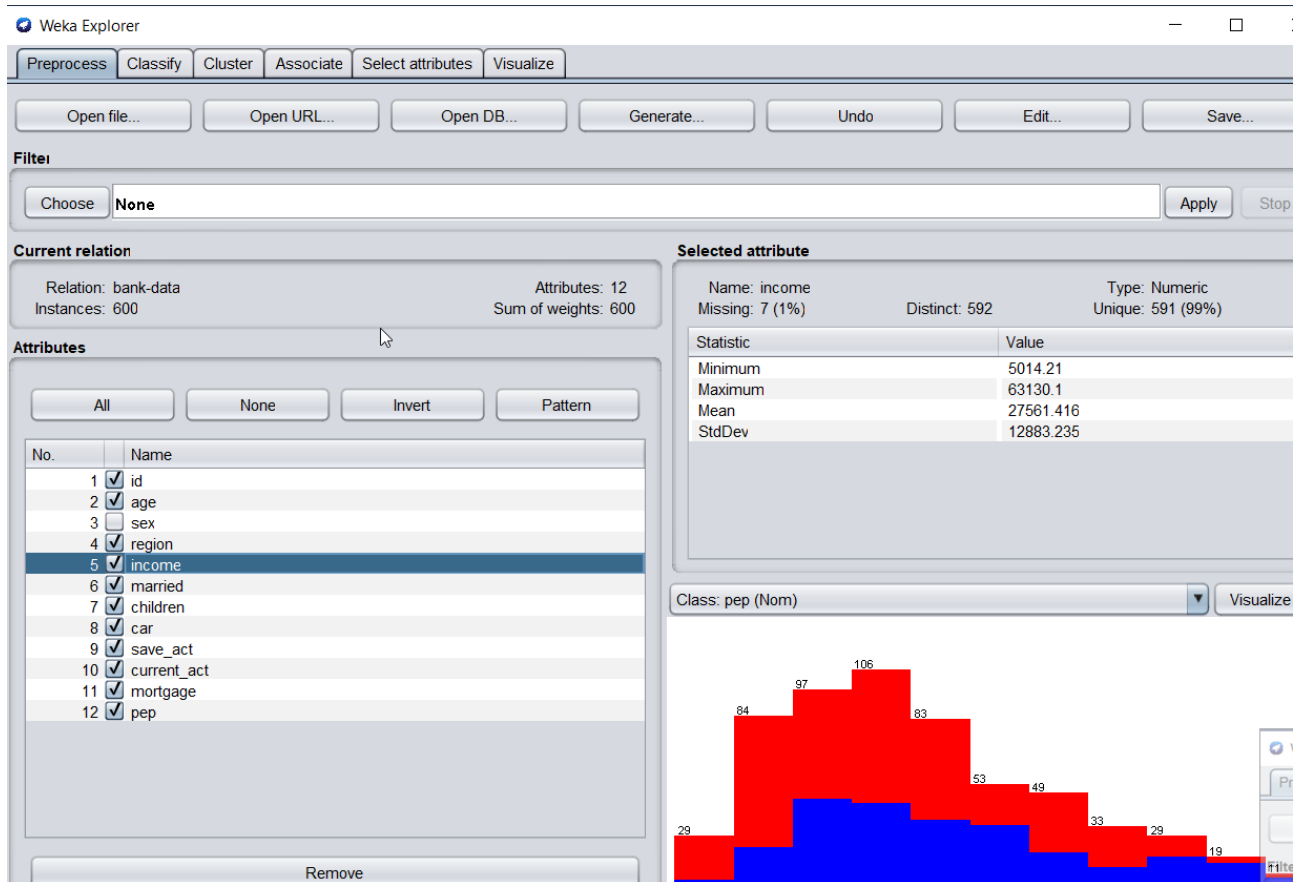
=== Confusion Matrix ===

a	b	<-- classified as	
224	50		a = YES
38	288		b = NO

The 'Result list' on the left shows two entries: '08:43:04 - trees.J48' and '08:45:34 - trees.J48'. The 'Status' bar at the bottom shows 'OK' and a 'Log' button.

**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

### 3. Filling Missing Value





# Don Bosco Institute of Technology, Mumbai 400070

## Department of Information Technology

### 2) REPLACE WITH MISSING VALUE

The screenshot shows the Weka Explorer interface. The 'Preprocess' tab is active, and the 'ReplaceWithMissingValue' filter is applied to the 'income' attribute. The 'Selected attribute' list on the right shows 'income' with 525 distinct values and 68 missing values (11%). The 'Class: pep (Nom)' is selected. A histogram of the 'income' attribute is visible on the right side of the filter dialog.

Relation: bank-data-weka.filters.unsupervised.attribute.ReplaceWithMissingValue-Rfirst-last-S1-P0.1-weka.filters.unsupervised.attribute.ReplaceWithMissingValue-Rfirst-last-S1-P0.1

Attributes: 12  
Sum of weights: 600

Selected attribute

Name: income  
Missing: 68 (11%)  
Distinct: 525  
Type: Numeric  
Unique: 523 (87%)

Statistic Value  
Minimum 0  
Maximum 63130.1  
Mean 27090.558  
StdDev 13229.936

Class: pep (Nom)

OUTPUT:-  
FILLING MISSING VALUE WITH USER CONSTANT

### 3) FILLING MISSING VALUE WITH USER CONSTANT

The screenshot shows the Weka Explorer interface. The 'Preprocess' tab is active, and the 'ReplaceMissingWithUserConstant' filter is applied to the 'income' attribute. The 'Selected attribute' list on the right shows 'income' with 525 distinct values and 0 missing values (0%). The 'Class: pep (Nom)' is selected. A histogram of the 'income' attribute is visible on the right side of the filter dialog.

Relation: bank-data-weka.filters.unsupervised.attribute.ReplaceMissingWithUserConstant-Afirst-last-NTRUE-R0-F"yyy-MM-dd'T'HH:mm:ss"

Attributes: 12  
Sum of weights: 600

Selected attribute

Name: income  
Missing: 0 (0%)  
Distinct: 525  
Type: Numeric  
Unique: 523 (87%)

Statistic Value  
Minimum 0  
Maximum 63130.1  
Mean 24020.295  
StdDev 15133.835

Class: pep (Nom)

Noisy data is a meaningless data generated due to faulty data collection following ways:-

1. Binning Method:

**Don Bosco Institute of Technology, Mumbai 400070**  
**Department of Information Technology**

**Conclusion :**

In this experiment we have study about various aspect of Data Processing with theory as well as using weka tool.

**References :**

<https://www.youtube.com/watch?v=aDMzPC5IO4c>  
[https://www.researchgate.net/profile/...datasets.../hw2\\_dataPreproc.pdf](https://www.researchgate.net/profile/...datasets.../hw2_dataPreproc.pdf)