Assignment: Pre-Processing techniques using Weka

Real world databases are highly influenced to noise, missing and inconsistency due to their queue size so thedata can be pre-processed to improve the quality of data and missing results and it also improves the efficiency.

There are 3 pre-processing techniques they are:

1) Add

2) Remove

3) Normalization

4) Handling of missing values

- **Mark Missing Values**
- **Remove instances with Missing Data**
- **Impute Missing Values (replace the missing values with some other value)**

1) Open Start ->Programs -> Accessories ->Notepad

2) Type the following training data set with the help of Notepad for Weather Table.

```
@relation weather
@attribute outlook {sunny,rainy,overcast}
@attribute temparature numeric
@attribute humidity numeric
@attribute windy {true,false}
@attribute play {yes,no}
 @data
sunny,85.0,85.0,false,no
overcast,80.0,90.0,true,no
sunny,83.0,86.0,false,yes
rainy,70.0,86.0,false,yes
rainy,68.0,80.0,false,yes
rainy,65.0,70.0,true,no
overcast,64.0,65.0,false,yes
sunny,72.0,95.0,true,no
```

sunny,69.0,70.0,false,yes

rainy,75.0,80.0,false,yes

3) After that the file is saved with .arff file format.

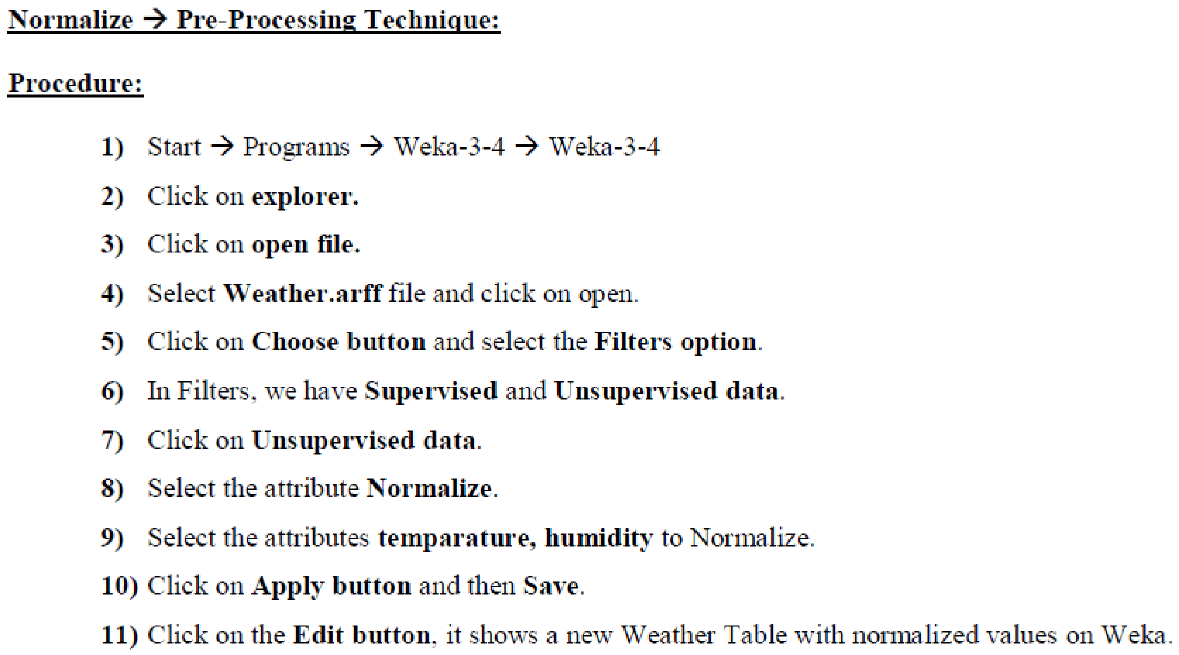
4) Minimize the arff file and then open Start □ Programs □ weka-3-4.

5) Click on weka-3-4, then Weka dialog box is displayed on the screen.

6) In that dialog box there are four modes, click on explorer.

7) Explorer shows many options. In that click on ‘open file’ and select the arff file

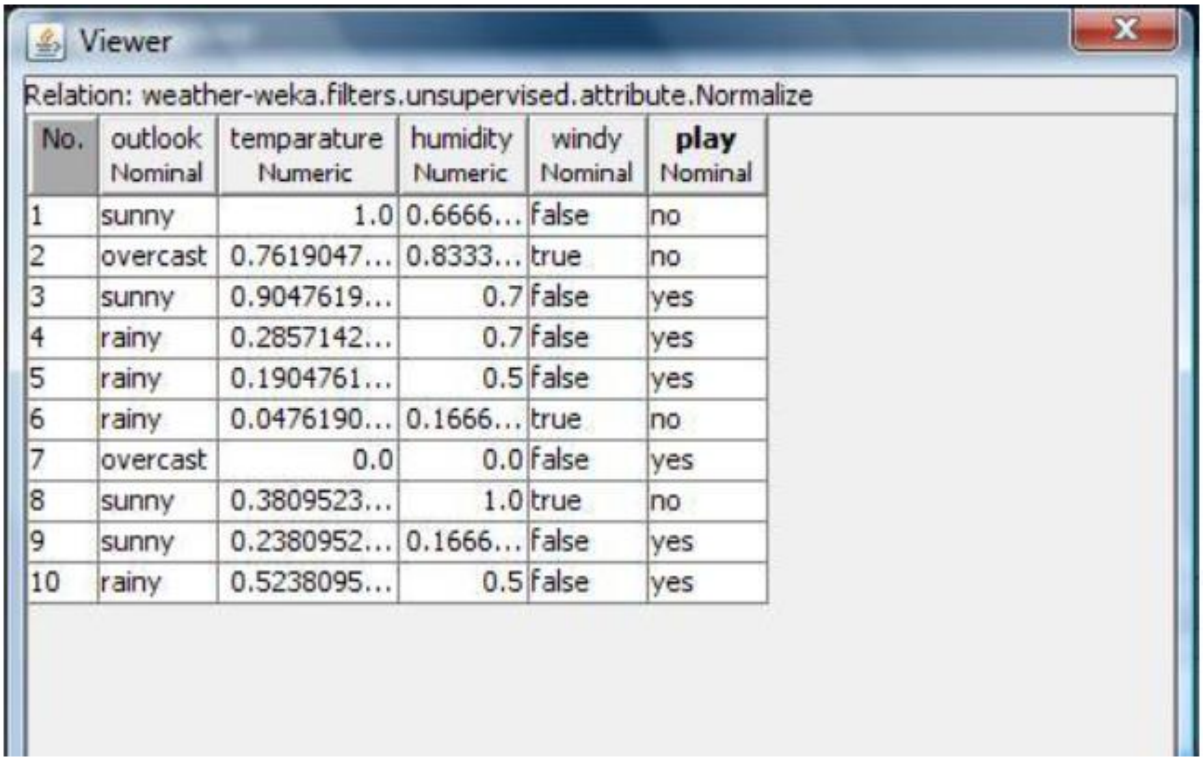
8) Click on edit button which shows weather table on weka.



Normalize → Pre-Processing Technique:

Procedure:

1) Start → Programs → Weka-3-4 → Weka-3-4
2) Click on **explorer.**
3) Click on **open file.**
4) Select **Weather.arff** file and click on open.
5) Click on **Choose button** and select the **Filters option**.
6) In Filters, we have **Supervised** and **Unsupervised data**.
7) Click on **Unsupervised data**.
8) Select the attribute **Normalize**.
9) Select the attributes **temparature, humidity** to Normalize.
10) Click on **Apply button** and then **Save**.
11) Click on the **Edit button**, it shows a new Weather Table with normalized values on Weka.



Relation: weather-weka.filters.unsupervised.attribute.Normalize

No.	outlook Nominal	temparature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	1.0	0.6666...	false	no
2	overcast	0.7619047...	0.8333...	true	no
3	sunny	0.9047619...	0.7	false	yes
4	rainy	0.2857142...	0.7	false	yes
5	rainy	0.1904761...	0.5	false	yes
6	rainy	0.0476190...	0.1666...	true	no
7	overcast	0.0	0.0	false	yes
8	sunny	0.3809523...	1.0	true	no
9	sunny	0.2380952...	0.1666...	false	yes
10	rainy	0.5238095...	0.5	false	yes

4) Handling of missing values

- **Mark Missing Values**
- **Remove instances with Missing Data**
- **Impute Missing Values (replace the missing values with some other value)**

STEP 1: Download the Pima Indians onset of diabetes dataset

https://www.kaggle.com/uciml/pima-indians-diabetes-database

Step 2: Convert the CSV File into a ARFF formatted File and then use the ARFF file for handling of missing values.

Step 3:

Mark Missing Values

Mark Missing Values

The Pima Indians dataset is a good basis for exploring missing data.Some attributes such as blood pressure (pres) and Body Mass Index (mass) have values of zero, which are impossible. These are examples of corrupt or missing data that must be marked manually.You can mark missing values in Weka using the NumericalCleaner filter. The recipe below shows you how to use this filter to mark the 11 missing values on the Body Mass Index (mass) attribute.

1. Open the Weka Explorer.

2. Load the Pima Indians onset of diabetes dataset.

3. Click the “Choose” button for the Filter and select NumericalCleaner, it us under unsupervized.attribute.NumericalCleaner.

4. Click on the filter to configure it.

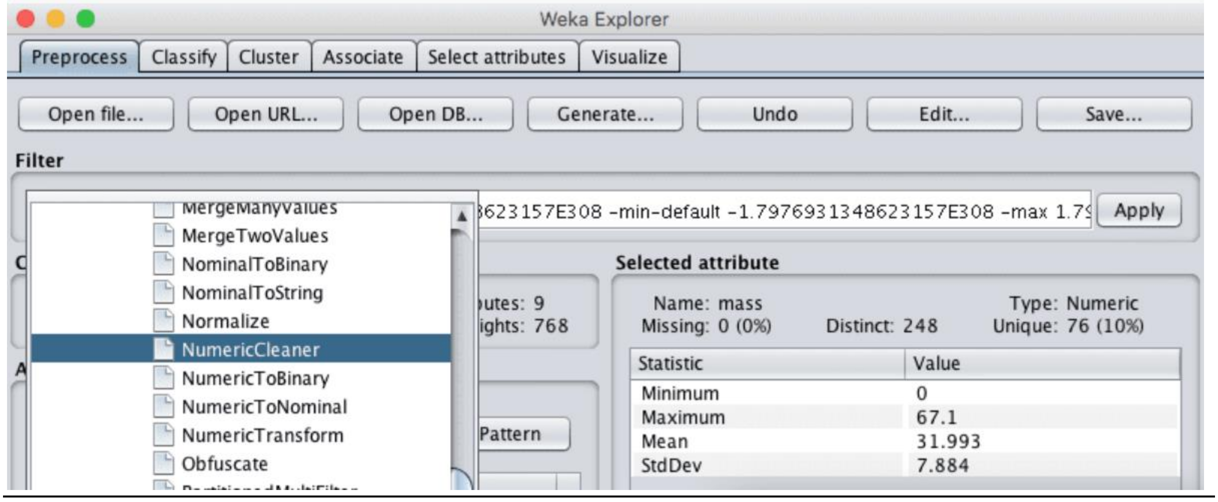
5. Set the attributeIndicies to 6, the index of the mass attribute.

6. Set minThreshold to 0.1E-8 (close to zero), which is the minimum value allowed for the attribute.

7. Set minDefault to NaN, which is unknown and will replace values below the threshold.

8. Click the “OK” button on the filter configuration.

9. Click the “Apply” button to apply the filter.



3623157E308 -min-default -1.7976931348623157E308 -max 1.79

butes: 9
ights: 768

Name: mass
Type: Numeric
Missing: 0 (0%)
Distinct: 248
Unique: 76 (10%)

Statistic	Value
Minimum	0
Maximum	67.1
Mean	31.993
StdDev	7.884

Remove Missing Data:

Now that you know how to mark missing values in your data, you need to learn how to handle them.A simple way to handle missing data is to remove those instances that have one or more missing values.You can do this in Weka using the RemoveWithValues filter.

Continuing on from the above recipe to mark missing values, you can remove missing values as follows:

1. Click the “Choose” button for the Filter and select RemoveWithValues, it us under unsupervized.instance.RemoveWithValues.

2. Click on the filter to configure it.

3. Set the attributeIndicies to 6, the index of the mass attribute.

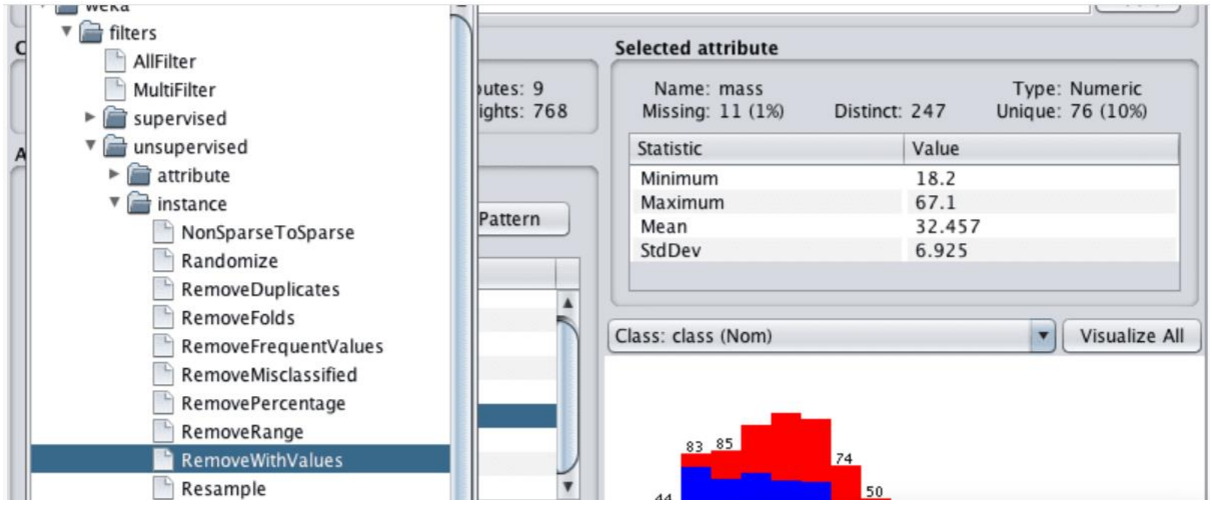
4. Set matchMissingValues to “True”.

5. Click the “OK” button to use the configuration for the filter.

6. Click the “Apply” button to apply the filter.

Click “mass” in the “attributes” section and review the details of the “selected attribute”.

Notice that the 11 attribute values that were marked Missing have been removed from the dataset



Selected attribute

Name: mass
Missing: 11 (1%)
Distinct: 247
Type: Numeric
Unique: 76 (10%)

Statistic	Value
Minimum	18.2
Maximum	67.1
Mean	32.457
StdDev	6.925

Impute Missing Values

Instances with missing values do not have to be removed, you can replace the missing values with some other value.

This is called imputing missing values.

It is common to impute missing values with the mean of the numerical distribution. You can do this easily in Weka using the ReplaceMissingValues filter.

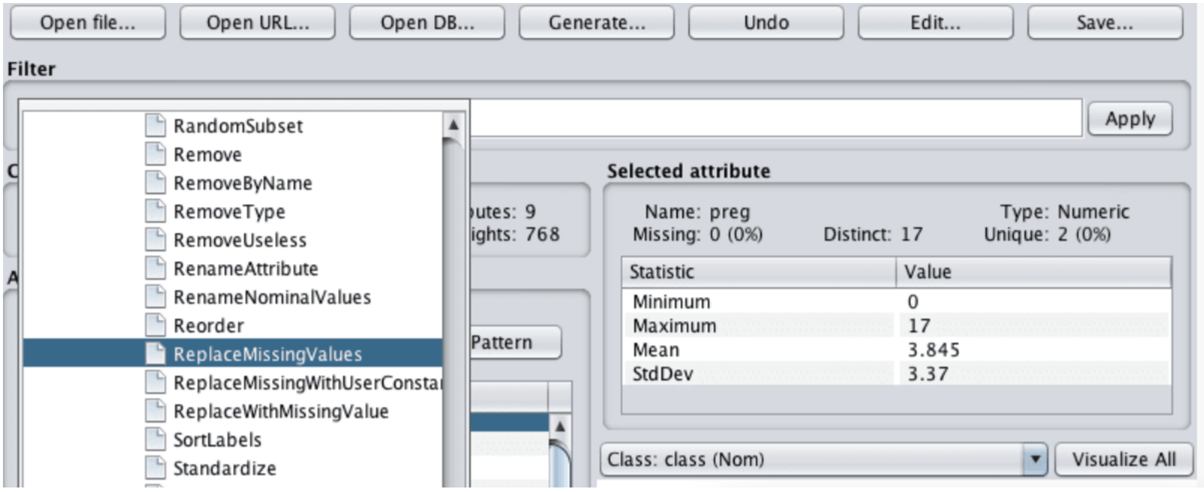
Continuing on from the first recipe above to mark missing values, you can impute the missing values as follows:

1. Click the “Choose” button for the Filter and select ReplaceMissingValues, it us under unsupervized.attribute.ReplaceMissingValues

2. Click the “Apply” button to apply the filter to your dataset.

Click “mass” in the “attributes” section and review the details of the “selected attribute”.

Notice that the 11 attribute values that were marked Missing have been set to the mean value of the distribution.



utes: 9
ights: 768

Name: preg
Missing: 0 (0%)
Distinct: 17
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

Class: class (Nom)