

**S.THARUN**

**19MID0031**

**CSI3010 – DATA WAREHOUSING AND DATA MINING**

**FACULTY : CHELLATAMILAN T**

**22-02-2022**

CYCLE SHEET 1

# Load the Dataset file into WEKA. This Dataset consists of set of attributes and a class label. Use this dataset and apply the following data pre-processing transformations.

## Determine, how many instances and attributes contained in the dataset?

There are 14 instances and 5 attributes ( 4 inputs and 1 output)

Table

Description automatically generated

## Give the class label and number of instances in each class. Which class is dominant in the dataset?

Class label 🡺 Output label 🡺 play {Yes,No}

Yes 🡺 9 instances 🡺 highly dominant class

No 🡺 5 instances

A screenshot of a computer

Description automatically generated

## Which Attribute has a sparse set of attributes from the mean? And Why?

Among the four attributes,

* Outlook
* Temperature
* Humidity
* Windy

Two attributes among them are continuous and sparsity can be found based on Standard deviation which can be calculated only for numerical / continuous values.

The attributes are

* Humidity
* Temperature

Temperature is sparser since it has highest Standard deviation as shown below

Table

Description automatically generated

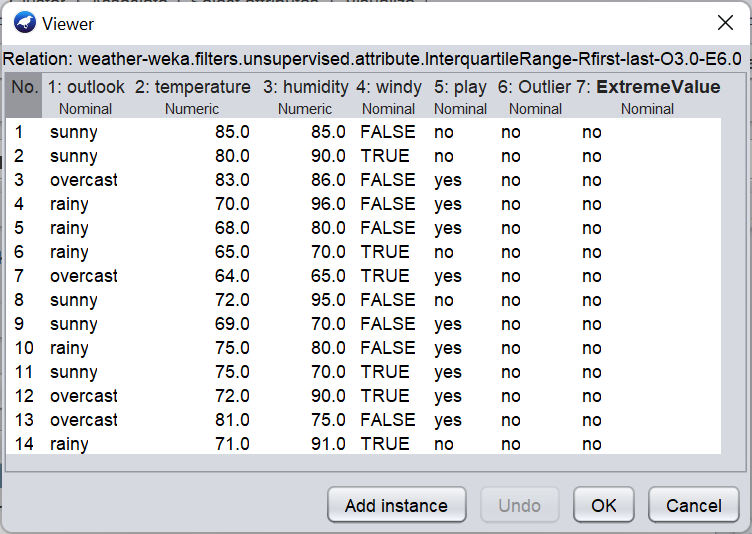
Graphical user interface

Description automatically generated

## Find how many outliers in the Dataset? What should we do on the outlier?

By applying Interquartile range filter, outliers can be found. An outlier column and Extreme value column will be created. In our example, there are no outliers. If there is any outlier, the specific instance will have Outlier attribute value as “Yes”.

Outliers can be removed (Remove with values filter) or imputed (replace with values filter) with values like mean, median and mode. Since we don’t have outliers, we are not doing that.



Graphical user interface, application, email

Description automatically generated