KNOWLEDGE ENGINEERING LAB (CSE 4.1.7)

4. Introduction to the WEKA machine learning toolkit

Create an ARFF (Attribute-Relation File Format) file and read it in WEKA. Explore the purpose of each button under the preprocess panel after loading the ARFF file. Also, try to interpret using a different ARFF file, weather.arff, provided with WEKA.

Waikato Environment for Knowledge Analysis (**Weka**) is a suite of machine earning software written in java, developed at the University of Waikato, New Zealand in 1993. It is free software licensed under the GNU(General Public License).

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces and so weka is indirectly called as Predictive modeling. Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. Weka is used to extract and analyze knowledge. **Weka** prefers to load data in the ARFF format. ARFF is an acronym that stands for Attribute-Relation File Format. It is an **extension** of the CSV file format where a header is used that provides metadata about the data types in the columns.

The applications in the WEKA software are:

- Explorer- Graphical interface to perform the data mining tasks on raw material.
- Experimental- Allows users to execute different experimental variations on data set.
- Knowledge flow- Explorer with drag & drop functionality and supports incremental learning from previous results
- Simple CLI(Command Line Interface)- Simple interface for executing commands from a terminal.
- Workbench- Combines all GUI interfaces into one.

Weka Explorer tries to support all the standard algorithms The Weka Knowledge Explorer is an easy to use graphical user interface that harnesses the power of the weka software. Each of the major weka packages Filters, Classifiers, Clusterers, Associations, and Attribute Selection is represented in the Explorer along with a Visualization tool which allows datasets and the predictions of Classifiers and Clusterers to be visualized in two dimensions.

Different Panels in Weka Explorer are:

- Data preprocessing- The preprocess panel is the start point for knowledge exploration. From this panel you can load datasets, browse the characteristics of attributes and apply any combination of Weka's unsupervised filters. to the data.
- Clustering- The classifier panel allows you to configure and execute any of the weka classifiers on the current dataset. You can choose to perform a cross validation or test on a separate dataset. Classification errors can be visualized in a pop-up data visualization tool. If the classifier produces a decision tree it can be displayed graphically in a pop-up tree visualizer.
- Classification- From the cluster panel you can configure and execute any of the weka clusters on the current dataset. Clusters can be visualized in a pop-up data visualization tool.
- Regresion(attribute)- From the associate panel you can mine the current dataset for association rules using the weka associators.
- Feature Selection (select attributes) -This panel allows you to configure and apply any combination of weka attribute evaluator and search method to select the most pertinent attributes in the dataset. If an attribute selection scheme transforms the data then the transformed data can be visualized in a pop-up data visualization tool.
- Visualization- This panel displays a scatter plot matrix for the current dataset.

Creating arff file for WEKA

- If you have a XLSX file then you need to convert it into a CSV(Comma Separated Values)File
- Then Open the CSV File with a text editor eg .Notepad++

- Append header relation eg.@relation 'name of the dataset'
- After that append the file with headers equal to the number of instances in your XLSX file eg.@attribute max numeric @attribute min numeric @attribute mean numeric @attribute median numeric. This means the file has four columns excluding the class label.
- Add the class label relation eg. @attribute CLASS {0,1} This has 2 classes mainly 0 and 1.
- After that append the header with @data and then save the file as .Arff

To read the ARFF file in WEKA we need to follow the following steps:

- Open the WEKA software and go to Tools.
- In Tools you can find ARFF viewer
- In ARFF viewer , click on file then open & search for the data set.
- Then another window will open named ARFF –viewer which shows the selected .arff file.

Now in the WEKA software there is an applications sections which shows different applications (Explorer, Experimenter, Knowledge Flow, Simple CLI) click on Explore to start preprocessing.

In the Preprocessing panel you can Analyze the data

Weather data Set

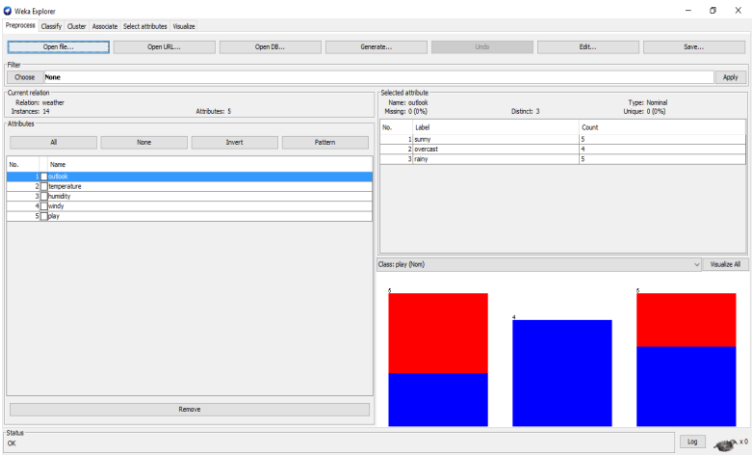
Relation: weather

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	**play** Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Click on open File & select the data set. As you open the data set you can find 4 sections

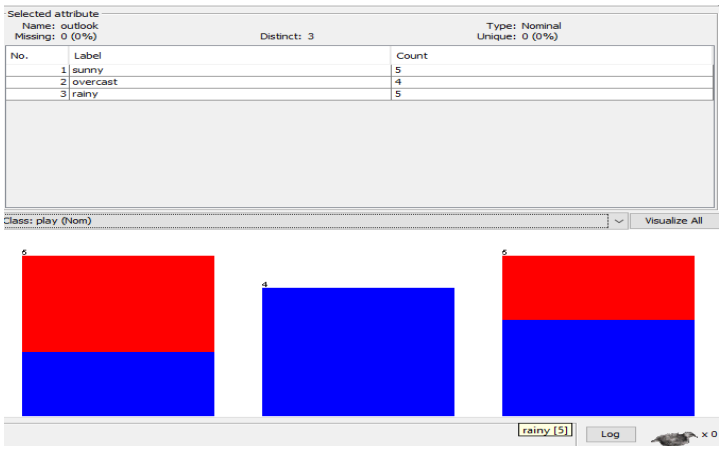
1. Current relation -shows which data set we are using
2. Attributes –shows all the attributes present in the data set

3. Selected attributes – The selected attribute to be analyze.
4. Class – shows the visual representation of a particular attribute.



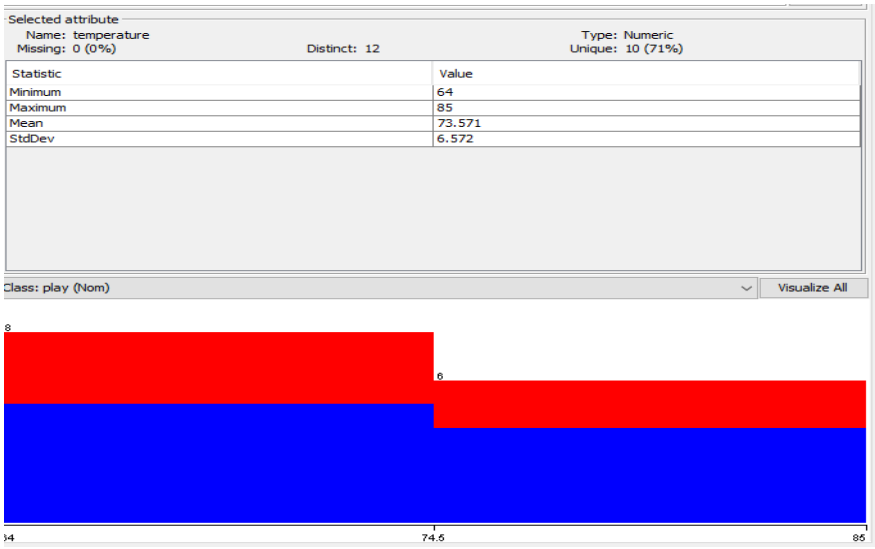
The current Relation, we have considered is Weather dataset which has 14 Instances and 5 attributes namely Outlook, temperature, humidity, sunny, play.

1.Selected attribute – Outlook



Here, we have selected the 'outlook' attribute which is of type 'nominal' where there is no missing and no unique value but the number of distinct values is 3, where each distinct value is defined with a label and the count of it. Visualizing the values we get a graph where label="sunny" has 5 counts in which 3 are NO and 2 are YES in the play(nominal) , similarly for label="overcast" which has 4 counts & all are YES and for label="rainy" it has 5 counts where 2 are NO and 3 are YES.

2.Selected attribute – Temperature

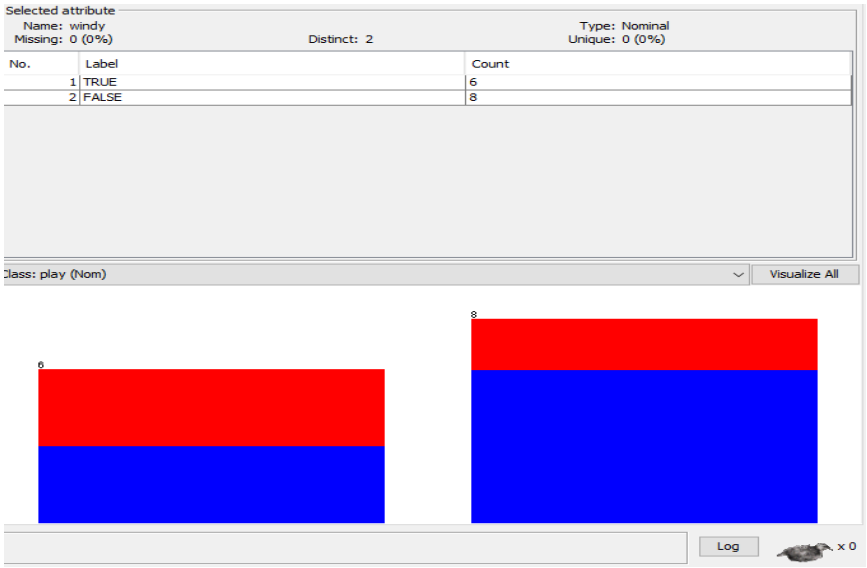


Here, we have selected the 'temperature' attribute which is of type 'numeric' where there is no missing values but 10 unique value and the number of distinct values is 12. The statistics of these distinct values is calculated by finding the Minimum, maximum, mean and standard deviation of all the values. Visualizing the values we get a graph where min to max is calculated from left to right. From min to mean (64 to 73.571) the no. of values are 8 where 5 are Yes and 3 are NO and From mean to max (73.571 to 85) the no. of values are 6 where 4 are Yes and 2 are NO in play nominal.

3.Selected attribute – Humidity

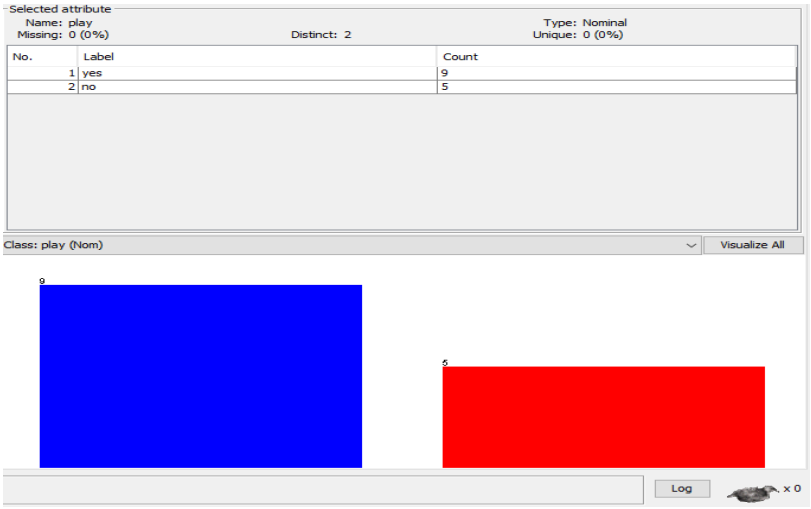
Here, we have selected the ‘humidity’ attribute which is of type ‘numeric’ where there is no missing values but 7 unique values and the number of distinct values is 10. The statistics of these distinct values is calculated by finding the Minimum, maximum, mean and standard deviation of all the values. Visualizing the values we get a graph where min to max is calculated from left to right. From min to mean (65 to 81.643) the no. of values are 7 where 6 are Yes and 1 is NO and From mean to max (73.571 to 85) the no. of values are 7 where 3 are Yes and 4 are NO in the play nominal.

4.Selected attribute – windy



Here, we have selected the ‘windy’ attribute which is of type ‘nominal’ where there is no missing and no unique value but the number of distinct values is 2, where each distinct value is defined with a label and the count of it. Visualizing the values we get a graph where label=”true” has 6 counts in which 3 are NO and 3 are YES in the play(nominal) for label=”false” it has 8 counts where 2 are NO and 6 ar YES.

5.Selected attribute – play



Here, we have selected the 'play' attribute which is of type 'nominal' where there is no missing and no unique value but the number of distinct values is 2, where each distinct value is defined with a label and the count of it. Visualizing the values we get a graph where label="Yes" has 9 counts with all YES and label="no" has 5 counts with all NO in the play(nominal).

The graph is represented in Class: Play(nominal) where No is indicated with Red colour & YES is indicated in Blue colour.