Semester	T.E. Semester VI – Information Technology
Subject	Data Mining and business Intelligence
Subject Professor In-charge	Prof. Deepali Nayak
Assisting Teachers	Prof. Vidhya Chitre
Laboratory	L007

Student Name	Ashutosh Engavle	
Roll Number	15101B0041	
Grade and Subject Teacher's Signature		

Experiment No	1B	
Aim	An introduction to the WEKA Explorer environment Creating an ARFF file and reading it into WEKA Preprocessing data	
Resources / Apparatus Required	Hardware: Computer System	Software: WEKA
Theory of Operation	***Tasks*** The WEKA GUI Chooser window is used to launch WEKA's graphical environments. WEKA Explorer is an environment for exploring data with WEKA. In this lab, we will be focusing on creating an ARFF file and reading it into WEKA, and using the WEKA Explorer. **Creating an ARFF file** Attribute-Relation File Format (ARFF) is a file format recognized by WEKA. An ARFF file typically has a *.arff* extension and contains two sections – a Header section and a Data section. A separate file named *ARFF.doc* explaining the ARFF specifications has been uploaded herewith. You are required to go through the file and understand the specifications, before you proceed further. Now follow the steps given below to create an ARFF file.	

Copy the data given in the file *lab#1data.doc*, to an Excel sheet.
Save the data set as CSV format.
Open it with a word processor and format it according to the ARFF specifications. Save as *data1.arff*.

The WEKA Explorer

Section Tabs

At the very top of the window, just below the title bar, is a row of tabs. When the Explorer is first started only the first tab is active. The tabs are as follows: Pre-process, Classify, Cluster, Associate, Select Attributes, and Visualize.

Status Box

The status box appears at the very bottom of the window. It displays messages that keep you informed about what's going on.

Opening files

The first button at the top of the preprocess section ***Open File*** enables us to load data into WEKA. Clicking that button brings up a dialogue box allowing you to browse for the data file on the local file system. Using the Open File button, read in the ARFF file you already created in this lab.

The Current Relation

Once the data has been loaded, the Preprocess panel shows a variety of information. The **Current Relation** box displays three entities – the name of the relation, the number of attributes in the data, and the number of instances in the data.

Attributes

Below the **Current Relation** box is a box titled **Attributes**. There are three buttons and beneath them is a list of attributes in the current relation. The three buttons – **All**, **None**, and **Invert** can be used to select desired attributes from the list.

When you click on different rows in the list of attributes, the fields change in the box to the right titled **Selected Attribute**. This box displays the characteristics of the currently highlighted attribute, namely – **Name**, **Type**, **Missing**, **Distinct**, and **Unique**.

Below these is a list showing more information about the values stored in this attribute, which differ depending on its type. For instance, if the attribute is numeric, the list gives four statistics describing the distribution of value in the data – the minimum, maximum, mean, and standard deviation. And below these is a colored histogram, color-coded according to the attribute chosen as the ***Class*** using the box above the histogram. Note that only nominal ***Class*** attributes will result in a color-coding. After

pressing the Visualize All button, histograms for all the attributes are shown in a separate window

Desired attributes can be removed by using the **Remove** button below the list of attributes. This can be undone by clicking the Undo button which is located in the top-right corner of the **Preprocess** panel. The **Edit** button next to it can be used to modify your data manually in a dataset editor.

You are expected to explore, observe and understand the purpose of each button under the preprocess panel after loading the ARFF file you prepared in this lab. Also, try to interpret what you observe using a different ARFF file, *weather.arff*, provided with WEKA.

Part II – Data Preprocessing

Objective: Understanding the purpose of unsupervised attribute/instance filters for preprocessing the input data.

Tasks

Open the file lab1 – data preprocessing.arff provided to you and carry out the following preprocessing tasks.

Follow the steps mentioned below to configure and apply a filter.

The preprocess section allows filters to be defined that transform the data in various ways. The Filter box is used to set up filters that are required. At the left of the Filter box is a Choose button. By clicking this button it is possible to select one of the filters in Weka. Once a filter has been selected, its name and options are shown in the field next to the Choose button. Clicking on this box brings up a GenericObjectEditor dialog box, which lets you configure a filter. Once you are happy with the settings you have chosen, click OK to return to the main Explorer window.

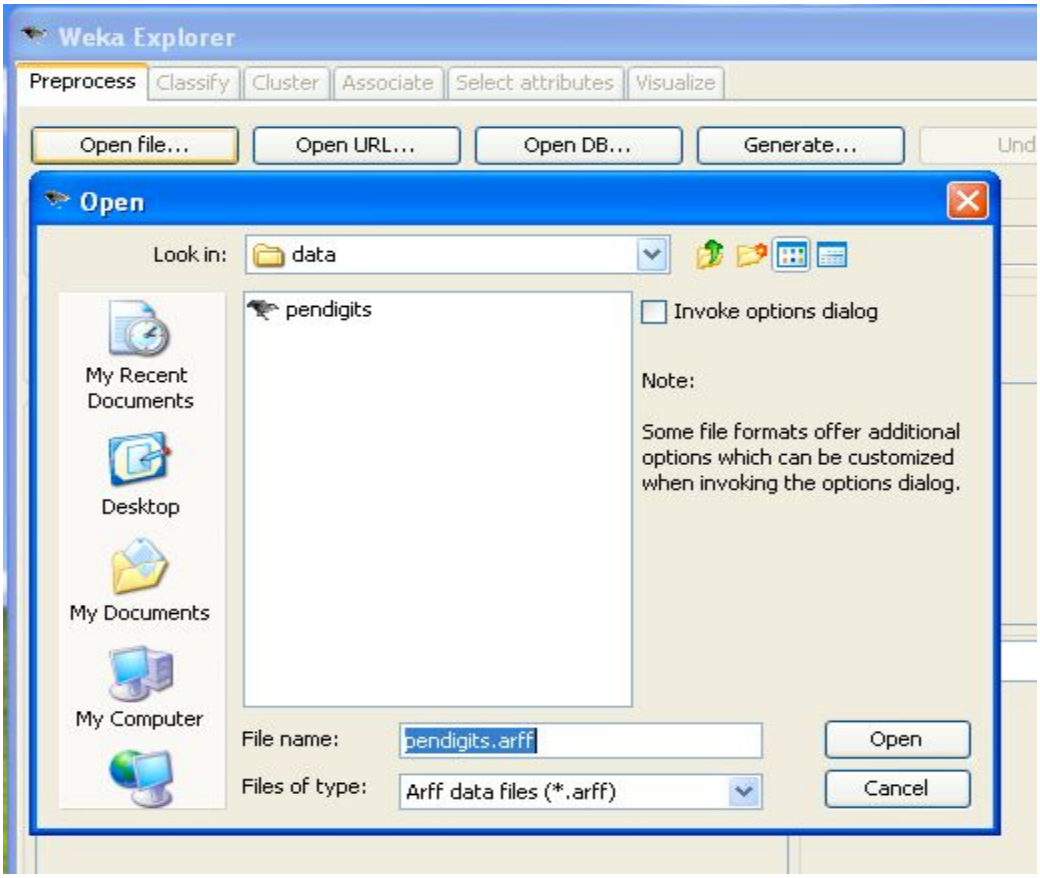
Now you can apply it to the data by pressing the Apply button at the right end of the Filter panel. The Preprocess panel will then show the transformed data. The change can be undone using the Undo button. Use the Edit button to view your transformed data in the dataset editor.

Try each of the following **Unsupervised Attribute Filters**.
(Choose -> weka -> filters -> unsupervised -> attribute)

- Use **ReplaceMissingValues** to replace missing values in the given dataset.

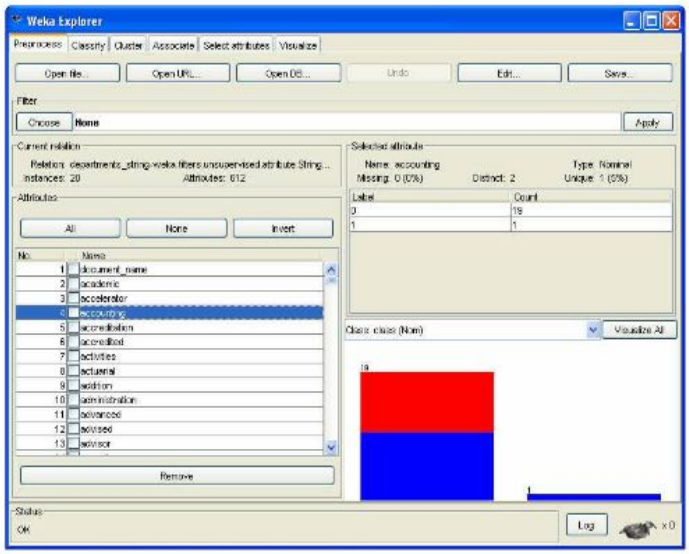
	● Use the filter **Add** to add the attribute Average. ● Use the filter **AddExpression** and add an attribute which is the average of attributes M1 and M2. Name this attribute as AVG. ● Understand the purpose of the attribute filter **Copy.** ● Use the attribute filters **Discretize** and **PKIDiscretize** to discretize the M1 and M2 attributes into five bins. (NOTE: Open the file afresh to apply the second filter since there would be no numeric attribute to dicretize after you have applied the first filter.) ● Perform **Normalize** and **Standardize** on the dataset and identify the difference between these operations. ● Use the attribute filter **FirstOrder** to convert the M1 and M2 attributes into a single attribute representing the first differences between them. ● Add a nominal attribute Grade and use the filter **MakeIndicator** to convert the attribute into a Boolean attribute. ● Try if you can accomplish the task in the previous step using the filter **MergeTwoValues**. ● Try the following transformation functions and identify the purpose of each ● ***NumericTransform*** ● ***NominalToBinary*** ● ***NumericToBinary*** ● ***Remove*** ● ***RemoveType*** ● ***RemoveUseless*** ● ***ReplaceMissingValues*** ● ***SwapValues*** Try the following **Unsupervised Instance Filters.** (Choose -> weka -> filters -> unsupervised -> instance) ● Perform ***Randomize*** on the given dataset and try to correlate the resultant sequence with the given one. ● Use ***RemoveRange*** filter to remove the last two instances. ● Use ***RemovePercent*** to remove 10 percent of the dataset. ● Apply the filter ***RemoveWithValues*** to a nominal and a numeric attribute **Loading and Preprocessing and visualization of Data file** ● Load data file in formats: ARFF, CSV, C4.5,binary

- Import from URL or SQL database (using JDBC)
- Preprocessing filters
  - Adding/removing attributes
  - Attribute value substitution
  - Discretization
  - Time series filters (delta, shift)
  - Sampling, randomization
  - Missing value management
  - Normalization and other numeric transformations



Finding a minimal set of attributes that preserve the class distribution

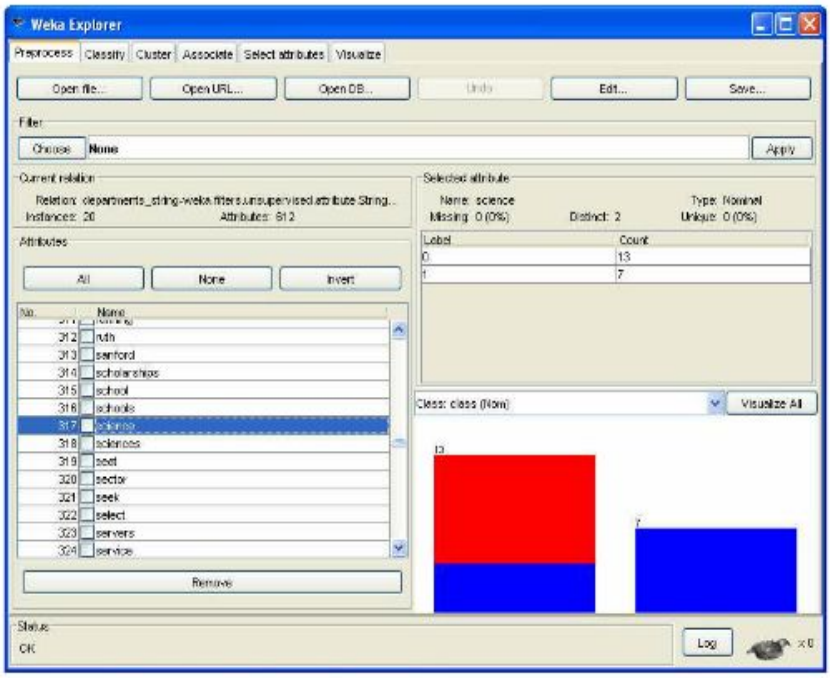
Attribute relevance with respect to the class – not relevant attribute (*accounting*)



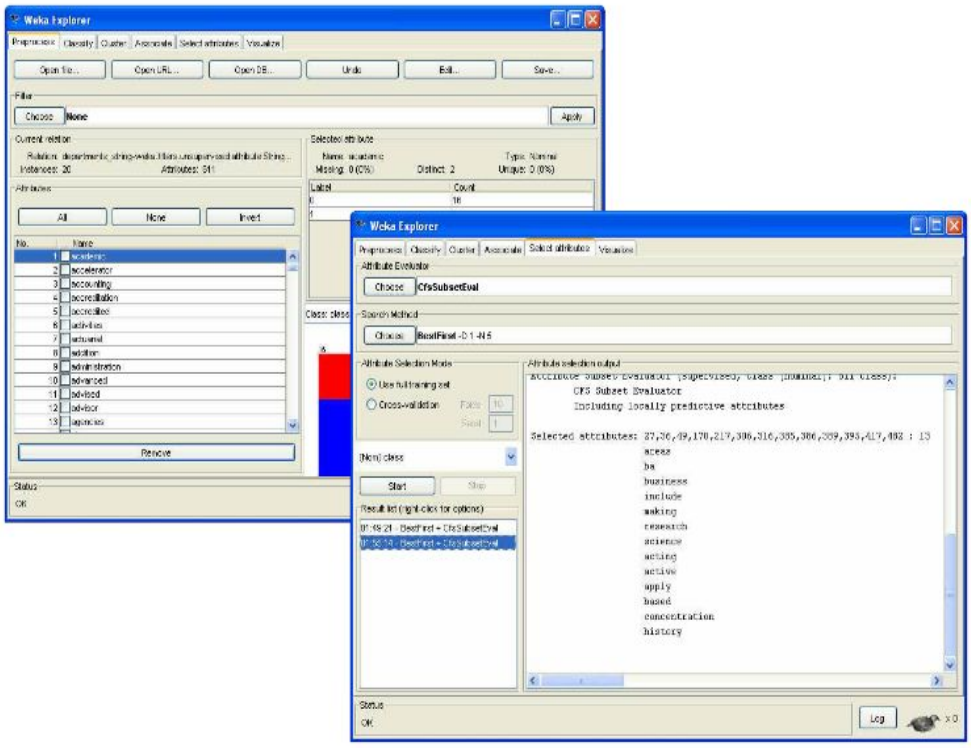
IF accounting=1 THEN class=A (Error=0, Coverage = 1 instance → **overfitting**)

IF accounting=0 THEN class=B (Error=10/19, Coverage = 19 instances → **low accuracy**)

Attribute relevance with respect to the class – relevant attribute (*science*)



IF accounting=1 THEN class=A (Error=0, Coverage = 7 instance)
IF accounting=0 THEN class=B (Error=4/13, Coverage = 13 instances)

	Attribute Selection (without document_name) 
Conclusion	Thus we have studied the preprocessing steps in WEKA