MATS SCHOOL OF INFORMATION TECHNOLOGY

Where Futures are Coded....

COURSE FILE

DATA WAREHOUSING AND DATA MINING

PROGRAM: MCA | SEMESTER: II | SESSION: 2023-24

PREPARED BY: PROF. (DR.) OMPRAKASH CHANDRAKAR

MATS UNIVERSITY
Highest NAAC Ranking Pvt. University in Chhattisgarh
ACCREDITED WITH GRADE B++ NAAC

MATS School of Information Technology

COURSE FILE

COURSE: MCA 202 - DATA WAREHOUSING AND DATA MINING

PROGRAM: MCA | SEMESTER: II | SESSION: 2023-24

Course Teacher: Prof. (Dr.) Omprakash Chandrakar

S No.	Description	Remarks
1	Course Curriculum	
2	Lesson Plan	
3	Evaluation Policy	
4	Lab Manual	
5	Previous 2 Years Question Papers	
6	Question Bank	
7	Model Question/Answer	
8	Lecture Notes/Study Material	
9	Suggested MOOC Courses	
10	Quiz/Unit Test/Mid Term Question Papers	
11	Result Analysis	
12	Best 2 Answer Books	
13	Final Continuous Internal Evaluation Marks	
13	Attendance Sheet	

SYLLABUS

PROGRAM: MCA SEMESTER: II WEF: 2023-24

Course Code: MCA 202	Credit: 04	Course: Data Warehousing and Data Mining	L: 03 | T: 01 | P: 00

Prerequisites:	Basic Database Management Concepts
Objectives:	To understand the need for analysis of large dataset, data mining models and methods to discover interesting patterns from such dataset. To understand the need for analysis of large dataset, multidimensional data modelling, OLAP and various data mining techniques.
Course Outcome:	Upon successfully finishing the course, students will have the capability to:

No.	Course Outcome	BT Level
CO1	Appreciate the multidisciplinary field of data mining, its need and the importance.	Understand
CO2	Apply various pre-processing techniques on the data before mining.	Analyze, Apply
CO3	Understand, design and create a data warehouse and perform OLAP operations on it.	Apply
CO4	Appreciate and apply the concept of association rule mining.	Apply
CO5	Appreciate and apply the concept of classification and clustering.	Apply

Program Outcome:	Upon successfully finishing the program, students shall be able to:

No.	Program Outcome
PO1	Attain proficiency in computer science, enabling the identification and resolution of computational problems through the application of computational knowledge.
PO2	Acquire the capability to design, develop, test, and maintain systems, components, products, or processes in accordance with specified requirements.

	PO3	Develop an understanding of the professional and ethical roles and responsibilities associated with the field.
	PO4	Recognize the importance of lifelong learning and develop the ability to pursue ongoing education and growth.
	PO5	Gain comprehensive knowledge in programming languages, database systems, operating systems, software engineering, web and mobile technology, and remain informed about pertinent contemporary issues.
	PO6	Demonstrate the adept use of modern tools, models, and languages to address software development challenges effectively.
	PO7	Cultivate the ability to communicate and convey knowledge proficiently, ensuring effective presentation and dissemination of information.

Program Outcomes and Course Outcomes Mapping:

Course Outcomes	Program Outcomes						
	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	√	√	√	√	√		
CO2	√	√	√	√	√		
CO3	√	√	√	√	√		
CO4	√	√	√	√	√	√	
CO5	√	√	√	√	√	√	

Syllabus

No.	Module Description		BT Level	Hours
1	**Introduction to Data Mining**			8
	1.1	Introduction to Data Science: Data mining, Machine Learning, Deep Learning, Artificial Intelligence, Data Warehouse, Big Data	Understand	2
	1.2	Data Mining, Knowledge Discovery from Data (KDD) Framework	Understand	2
	1.3	Types of data for Data Mining	Understand	1
	1.4	Data Mining: Confluence of multiple disciplines	Understand	1
	1.5	Data Mining Applications	Understand	2
2	**Data Preprocessing**			**10**
	2.1	Data types: Nominal attributes, Binary attributes, Ordinal attributes	Understand	2
	2.2	Statistics of data: Central tendency, dispersion of data - Range, quartiles, Variance and standard deviation	Understand	2

	2.3	Covariance and correlation analysis	Understand	2
	2.4	Data quality, Data cleaning: Missing values, Noisy data, Data integration	Understand	2
	2.5	Data transformation: Normalization, Discretization	Understand	2
3	**Data warehousing and Online Analytical Processing**			**10**
	3.1	Introduction to Data Warehouse	Analyze	1
	3.2	Data Warehouses Architecture: The three-tier architecture, ETL, Enterprise data warehouse and data mart	Analyze	2
	3.3	Data cube: a multidimensional data model	Understand	2
	3.4	Schemas for multidimensional data models: stars, snowflakes, and fact constellations	Analyze	2
	3.5	Concept hierarchies	Analyze	1
	3.6	OLAP operations	Analyze	2
4	**Association Rule Mining**			**10**
	4.1.	Market basket analysis	Apply	2
	4.2.	Frequent itemsets	Apply	2
	4.3.	Apriori algorithm: finding frequent itemsets	Apply	2
	4.4.	Generating association rules from frequent itemsets	Apply	2
	4.5	From association analysis to correlation analysis		2
5	**Classification and Cluster Analysis**			**12**
	5.1	Introduction to Classification	Understand	1
	5.2	Decision tree induction	Understand	2
	5.3	Attribute selection measures: Information gain, Gain ratio	Understand	2
	5.4	Naïve Bayesian classification	Understand	2
	5.5	Cluster Analysis	Understand	1
	5.6	Partitioning methods	Understand	2
	5.7	k-Means: a centroid-based technique	Understand	2

Text Books/ Resources:	1. Han, J. and Kamber, M. - Data Mining: Concepts & Techniques, 3rd Edition - Morgan Kaufmann Publishers: TB#1 2. Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publications
Reference Books/ Resource	1. Mohammed J. Zaki Wagner Meira Jr - Data Mining and Machine Learning: Fundamental Concepts and Algorithms 2. Pujari, A. - Data Mining techniques - Universities Press 3. Pudi, V. and Radhakrishnan, P. - Data Mining - Oxford University Press

	4. Larose, D. - Data Mining Methods & Models - Wiley-India 5. Berry, M. and Linoff, G. - Data Mining Techniques - Wiley-India

MATS School of Information Technology

LESSON PLAN

COURSE: MCA202 DATA WAREHOUSING AND DATA MINING

PROGRAM: MCA SEMESTER: II SESSION: 2023-24

No.	Module Description		Reference	BT Level	Hours
1	**Introduction to Data Mining**		**TB#1: Ch-1**		**8**
	1.1	Introduction to Data Science: Data mining, Machine Learning, Deep Learning, Artificial Intelligence, Data Warehouse, Big Data		Understand	2
	1.2	Data Mining, Knowledge Discovery from Data (KDD) Framework	1.1 - 1.2	Understand	2
	1.3	Types of data for Data Mining	1.3	Understand	1
	1.4	Data Mining: Confluence of multiple disciplines	1.5	Understand	1
	1.5	Data Mining Applications	1.6	Understand	2
2	**Data Preprocessing**		**TB#1: Ch-2**		**10**
	2.1	Data types: Nominal attributes, Binary attributes, Ordinal attributes	2.1	Understand	2
	2.2	Statistics of data: Central tendency, dispersion of data - Range, quartiles, Variance and standard deviation		Understand	2
	2.3	Covariance and correlation analysis	2.2.3	Understand	2
	2.4	Data quality, Data cleaning: Missing values, Noisy data, Data integration	2.4	Understand	2
	2.5	Data transformation: Normalization, Discretization	2.5.1-2.5.2	Understand	2
3	**Data warehousing and Online Analytical Processing**		**TB#1: Ch-3**		**10**
	3.1	Introduction to Data Warehouse	3.1	Analyze	1
	3.2	Data Warehouses Architecture: The three-tier architecture, ETL, Enterprise data warehouse and data mart	3.1.1 -3.1.2	Analyze	2
	3.3	Data cube: a multidimensional data model	3.2.1	Understand	2
	3.4	Schemas for multidimensional data models: stars, snowflakes, and fact constellations	3.2.2	Analyze	2
	3.5	Concept hierarchies	3.2.3	Analyze	1
	3.6	OLAP operations	3.3.1	Analyze	2

4	**Association Rule Mining**		**TB#1: Ch-4**		**10**
	4.1.	Market basket analysis	4.1.1	Apply	2
	4.2.	Frequent itemsets	4.1.2	Apply	2
	4.3.	Apriori algorithm: finding frequent itemsets	4.2.1	Apply	2
	4.4.	Generating association rules from frequent itemsets	4.2.2	Apply	2
	4.5	From association analysis to correlation analysis	4.3.1-4.3.2		2
5	**Classification and Cluster Analysis**		**TB#1: Ch-6, 8**		**12**
	5.1	Introduction to Classification	6.1.1	Understand	1
	5.2	Decision tree induction	6.2	Understand	2
	5.3	Attribute selection measures: Information gain, Gain ratio	6.2.2	Understand	2
	5.4	Naïve Bayesian classification	6.3.2	Understand	2
	5.5	Cluster Analysis	8.1.1	Understand	1
	5.6	Partitioning methods	8.2	Understand	2
	5.7	k-Means: a centroid-based technique	8.2.1	Understand	2

Text Books/ Resources:	1. Han, J. and Kamber, M. - Data Mining: Concepts & Techniques, 3rd Edition - Morgan Kaufmann Publishers: TB#1 2. Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publications
Reference Books/ Resources	1. Mohammed J. Zaki Wagner Meira Jr - Data Mining and Machine Learning: Fundamental Concepts and Algorithms 2. Pujari, A. - Data Mining techniques - Universities Press 3. Pudi, V. and Radhakrishnan, P. - Data Mining - Oxford University Press 4. Larose, D. - Data Mining Methods & Models - Wiley-India 5. Berry, M. and Linoff, G. - Data Mining Techniques - Wiley-India

MATS School of Information Technology

EVALUATION POLICY

COURSE: MCA 202 - DATA WAREHOUSING AND DATA MINING

PROGRAM: MCA | SEMESTER: II | SESSION: 2023-24

Evaluation Parameter	Duration	Minimum Frequency	Marks	Weightage for CCE [30]	Remarks
Quiz	10 minutes	3 (Best of 2 will be considered)	10	3 x 2 = 6	Preferably on completion of every two modules.
Unit Test	1 hour	1	30	6	As per exam schedule.
Other Parameters (Assignment/ Presentation/ Poster/ Activity)	As per instruction	2	10	4	Preferably on completion of every two modules.
Midterm Exam (Theory)	2 hours	1	50	14	As per exam schedule.
Midterm Exam (Practical)	2 hours	1	35	15	As per exam schedule.
Term End Exam (Theory)	2.5 hours	1	70	70	As per university exam schedule.
Term End Exam (Practical)	2 hours	1	35	35	As per university exam schedule.

मैट्स विश्वविद्यालय
MATS UNIVERSITY

University Campus: Gullu, Aarang, Raipur – 493441 | Raipur Campus: MATS Tower, Pandri, Raipur – 492004
Phone : +91-771-40789 95/96/98 Fax : +91-771-40789 97
E-Mail:registrar@matsuniversity.ac.in Website : www.matsuniversity.ac.in

MATS School of Information Technology

Index

Question Bank

MCA 202 - DATA WAREHOUSING AND DATA MINING

QUESTION BANK

MCA 202 - DATA WAREHOUSING AND DATA MINING

Module 1: Introduction to Data Mining

Short Questions [2 Marks]

1. What is data mining? List any three applications of data mining.
2. List the steps involved in the process of knowledge discovery from large database.
3. What are the various data mining functionalities? List them.
4. Explain any two data mining functionalities in brief.
5. What is transactional data? Briefly describe giving example.
6. What are the measures of interestingness of a pattern?
   List the measures based on which a pattern's interestingness is determined.
7. Mining frequent patterns leads to the discovery of interesting associations and correlations within data. State weather the statement is true or false. Justify your answer.
8. What is classification? List any three applications of it.
9. What is regression? List any three applications of it.
10. What is cluster analysis? List any three applications of it.
11. What is outlier analysis? List any three applications of it.

Long Questions [8-10 Marks]

1. Data mining turns a large collection of data into knowledge. Justify the statement giving an example.
2. Describe the steps involved in data mining when viewed as a process of knowledge discovery with diagram.
3. List the kinds of data that can be mined. Describe any two of them in brief.
   OR
   Classify the different types of data on which mining can be performed. Explain any two of them in brief.

OR

Differentiate between followings:

1. Structured vs. unstructured data
2. Stored vs. streaming data

4. Write short note on confluence of Machine learning and Data mining.
5. Describe the confluence of Data mining and Data science.
6. Write a detailed note on the applications of data mining.
7. Suppose your task as a software engineer at MATS University is to design a data mining system to examine their university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and their cumulative grade point average (GPA). Describe the architecture you would choose. What is the purpose of each component of this architecture?
8. Explain the following term and describe how they are interrelated: Data mining, Machine Learning, Deep Learning, Artificial Intelligence, Data Warehouse, and Big Data
9. List and describe the five primitives for specifying a data mining task.
10. Describe three challenges to data mining regarding data mining methodology and user interaction issues.

QUESTION BANK

MCA 202 - DATA WAREHOUSING AND DATA MINING

UNIT 2: DATA PRE-PROCESSING

Short Questions [2 Marks]

1. When does a data can be considered as quality data?
2. List the factors that make a quality data.
   What are the elements that define data quality? List them.
3. List any four data pre-processing task.
4. What does it mean by noisy data? Explain giving an example.
5. Which techniques are used to overcome the noisy data problem?
6. Define following terms giving example: accuracy, completeness, and consistency.
7. Give examples of data that are inaccurate, incomplete and inconsistent.
8. Define the following terms giving example:
   Timeliness, Believability and Interpretability
9. What is data smoothing technique? List them.
10. Explain how regression is used for data smoothing.
11. Explain how outlier analysis is used for data smoothing.
12. Illustrate how binning is used for data smoothing.
13. What does it mean by data integration?
14. Why does the data reduction required during the pre-processing stage in data mining? List any four data reduction techniques.

Long Questions [8-10 Marks]

1. Describe any four Data Pre-processing task in brief.
2. Discuss any five techniques to deal with missing values.
3. What kind of data problem is overcome by smoothing techniques? Describe any three data smoothing techniques.
4. Describe data cleaning giving any three examples.
5. Explain any three data reduction techniques in brief.

6. Describe Attribute Subset Selection methods in brief.
7. Describe data cube giving the definition of base cuboid, apex cuboid, and lattice of cuboids.
8. What is data transformation? Briefly describe any 5 strategies of data transformation.
9. What is histogram?

   The following data are a list of AllElectronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

   1. Create a histogram for price using singleton buckets

   2. Create an equal-width histogram for price ($ 10)
10. Suppose that the minimum and maximum values for the attribute income are Rs. 12,000 and Rs. 98,000, respectively. We would like to map income to the range [0.0, 1.0]. How do you transform a value of Rs. 73,600 and Rs. 20,000 using min-max normalization?

    Use min-max normalization methods by setting min = 0 and max = 1 to normalize the following group of data: 200, 300, 400, 600, 1000
11. The following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

    Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps.

मैट्स विश्वविद्यालय
MATS UNIVERSITY

University Campus: Gullu, Aarang, Raipur – 493441 | Raipur Campus: MATS Tower, Pandri, Raipur – 492004
Phone : +91-771-40789 95/96/98 Fax : +91-771-40789 97
E-Mail:registrar@matsuniversity.ac.in Website : www.matsuniversity.ac.in

MATS School of Information Technology

Index

Lab Manual | MCA 202 Data Warehousing and Data Mining | 2023-24

Roll No:	Name:

No.	Practical Title	Page No	Due Date	Submission Date	Evaluation*		Teacher's Signature with Date
					Practical	Viva	
1.	Exploring ARFF file format, loading dataset and performing pre-processing operations on the given dataset.						
2.							
3.							
4.							
5.							

***Journal completion and timely submission are the pre-requisite to appear in any of the practical evaluation.**

Course Code: MCA 202 | Course Title: Data Warehousing and Data Mining			
Course Credits: 04	Total Hours: 50	[Lectures: 04, Tutorial: 00, Practical: 04]	
Prerequisites:	Basic Database Management Concepts		
Course Objectives:	To understand the need for analysis of large dataset, data mining models and methods to discover interesting patterns from such dataset. To understand the need for analysis of large dataset, multidimensional data modelling, OLAP and various data mining techniques.		
Programme Outcomes:	Upon successfully finishing the program, students shall be able to:		
	No.	**Program Outcome**	
	PO1	Attain proficiency in computer science, enabling the identification and resolution of computational problems through the application of computational knowledge.	
	PO2	Acquire the capability to design, develop, test, and maintain systems, components, products, or processes in accordance with specified requirements.	
	PO3	Develop an understanding of the professional and ethical roles and responsibilities associated with the field.	
	PO4	Recognize the importance of lifelong learning and develop the ability to pursue ongoing education and growth.	
	PO5	Gain comprehensive knowledge in programming languages, database systems, operating systems, software engineering, web and mobile technology, and remain informed about pertinent contemporary issues.	
	PO6	Demonstrate the adept use of modern tools, models, and languages to address software development challenges effectively.	
	PO7	Cultivate the ability to communicate and convey knowledge proficiently, ensuring effective presentation and dissemination of information.	
Course Outcomes:	Upon successfully finishing the course, students will have the capability to:		
	No.	**Course Outcome**	**Level**
	CO1	Appreciate the multidisciplinary field of data mining, its need and the importance.	Understand

	CO2	Apply various pre-processing techniques on the data before mining.	Analyze, Apply
	CO3	Understand, design and create a data warehouse and perform OLAP operations on it.	Apply
	CO4	Appreciate and apply the concept of association rule mining.	Apply
	CO5	Appreciate and apply the concept of classification and clustering.	Apply

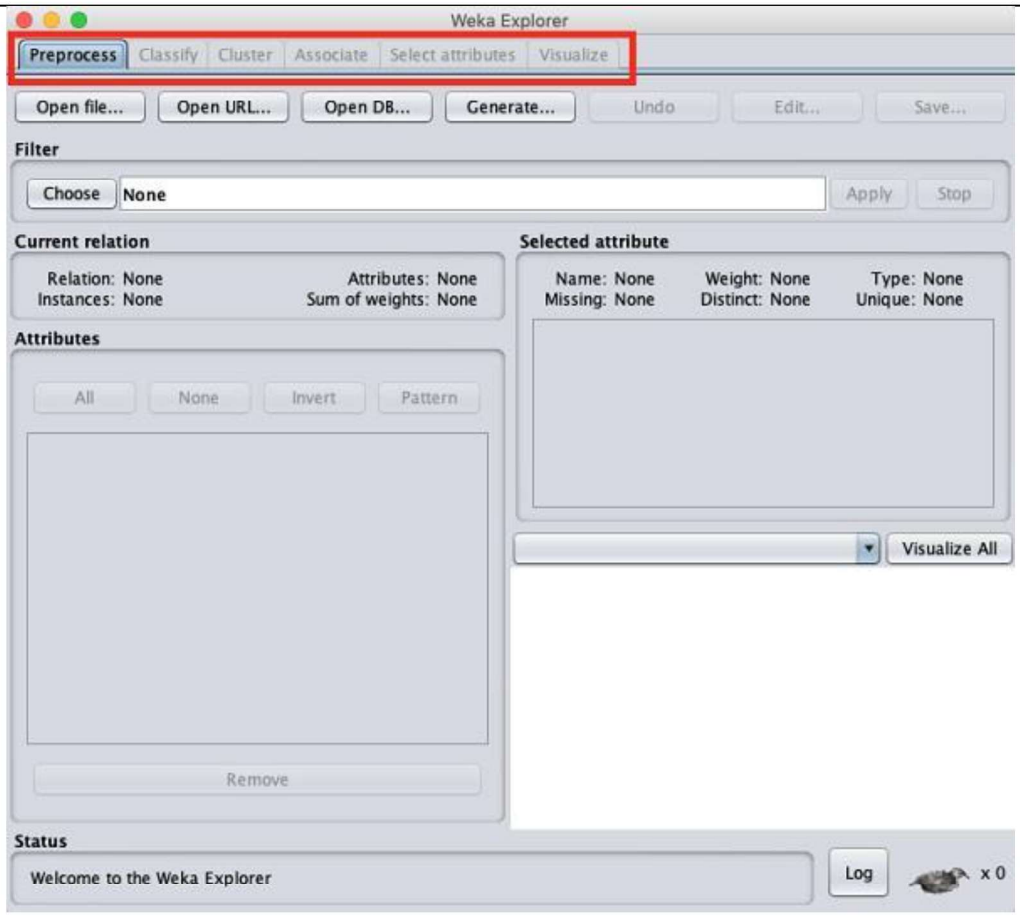
Practical List - Course Outcomes Matrix

P #	Practical Title	CO1	CO2	CO3	CO4	CO5
1.	Exploring ARFF file format, loading dataset and performing pre-processing operations on the given dataset.	√	√			
2.						
3.						
4.						
5.						

Tools and Technology	Weka 3.7.8
Computing Environment:	Processor: Intel® Core™ i5 Processors Memory: 16 GB / 12 GB Operating System: Windows 10
Learning Resources:	https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf https://www.tutorialspoint.com/weka/index.htm
Instructions for Lab Teachers:	Continuous assessment of the student lab work and journal in the prescribed format must be followed.
Instructions for Lab Students:	1. Do not disturb machine Hardware / Software Setup. 2. All the students should sit according to their roll numbers starting from their left to right. 3. Do not change the terminal on which you are working without lab teacher's permission. 4. Refer the learning resources related to the programs to be implemented in the lab. 5. Given Journal / paper work must be finished before you enter in the lab. 6. Strictly observe the instructions given by the teacher/Lab Instructor.

Practical List

Practical No: 1	**Exploring ARRF file format, loading dataset and performing pre-processing operations on the given dataset.**
Objective(s):	Study the Weka Explorer and perform pre-processing operations.
Pre-requisite:	Familiarity with database concepts.
Keywords/ Concepts:	.arrf, loading data, pre-processing
Solution:	Must contain code and output.
Nature of submission:	Handwritten
Learning Recourses:	**Weka - Launching Explorer** Let us look into various functionalities that the explorer provides for working with big data. When you click on the **Explorer** button in the **Applications** selector, it opens the following screen –



On the top, you will see several tabs as listed here –

- Preprocess
- Classify
- Cluster
- Associate
- Select Attributes
- Visualize

Under these tabs, there are several pre-implemented machine learning algorithms. Let us look into each of them in detail now.

Preprocess Tab

Initially as you open the explorer, only the **Preprocess** tab is enabled. The first step in machine learning is to preprocess the data. Thus, in the **Preprocess** option, you will select

	the data file, process it and make it fit for applying the various machine learning algorithms. Classify Tab The **Classify** tab provides you several machine learning algorithms for the classification of your data. To list a few, you may apply algorithms such as Linear Regression, Logistic Regression, Support Vector Machines, Decision Trees, RandomTree, RandomForest, NaiveBayes, and so on. The list is very exhaustive and provides both supervised and unsupervised machine learning algorithms. Cluster Tab Under the **Cluster** tab, there are several clustering algorithms provided - such as SimpleKMeans, FilteredClusterer, HierarchicalClusterer, and so on. Associate Tab Under the **Associate** tab, you would find Apriori, FilteredAssociator and FPGrowth. Select Attributes Tab **Select Attributes** allows you feature selections based on several algorithms such as ClassifierSubsetEval, PrinicipalComponents, etc. Visualize Tab Lastly, the **Visualize** option allows you to visualize your processed data for analysis. As you noticed, WEKA provides several ready-to-use algorithms for testing and building your machine learning applications. To use WEKA effectively, you must have a sound knowledge of these algorithms, how they work, which one to choose under what circumstances, what to look for in their processed output, and so on. In short, you must have a solid foundation in machine learning to use WEKA effectively in building your apps.

In this chapter, we start with the first tab that you use to preprocess the data. This is common to all algorithms that you would apply to your data for building the model and is a common step for all subsequent operations in WEKA.

For a machine learning algorithm to give acceptable accuracy, it is important that you must cleanse your data first. This is because the raw data collected from the field may contain null values, irrelevant columns and so on.

Weka - Loading Data

In this chapter, you will learn how to preprocess the raw data and create a clean, meaningful dataset for further use.

First, you will learn to load the data file into the WEKA explorer. The data can be loaded from the following sources −

- Local file system
- Web
- Database

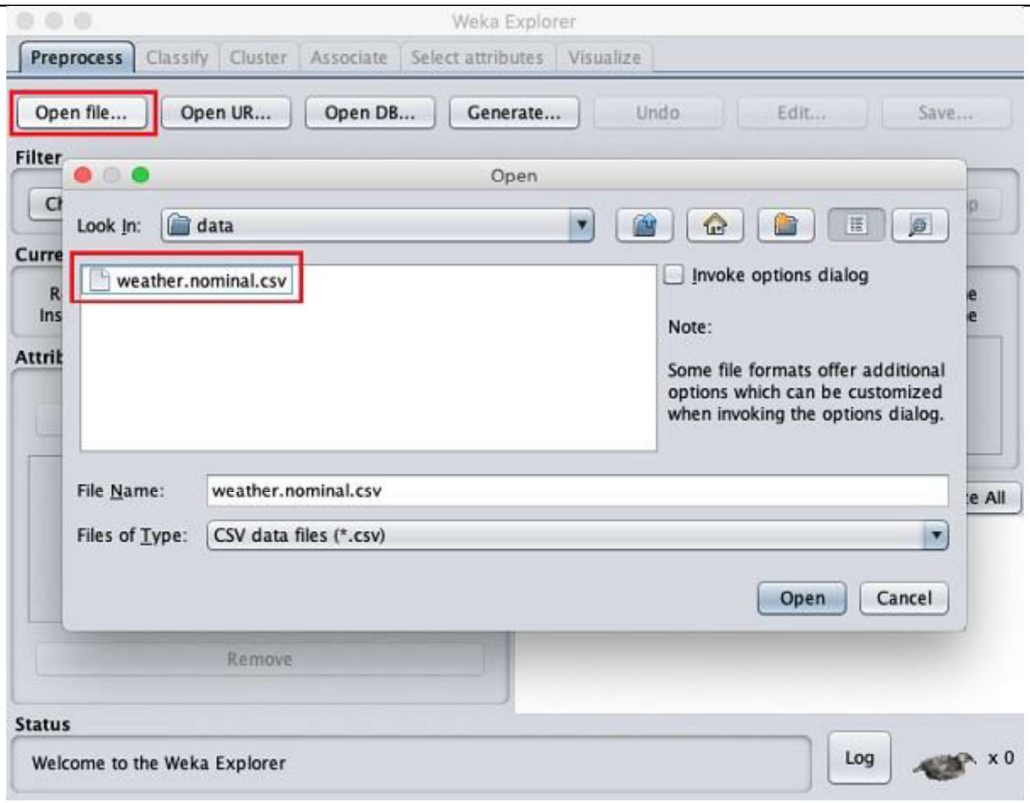
In this chapter, we will see all the three options of loading data in detail.

Loading Data from Local File System

Just under the Machine Learning tabs that you studied in the previous lesson, you would find the following three buttons −

- Open file ...
- Open URL ...
- Open DB ...

Click on the **Open file** ... button. A directory navigator window opens as shown in the following screen −

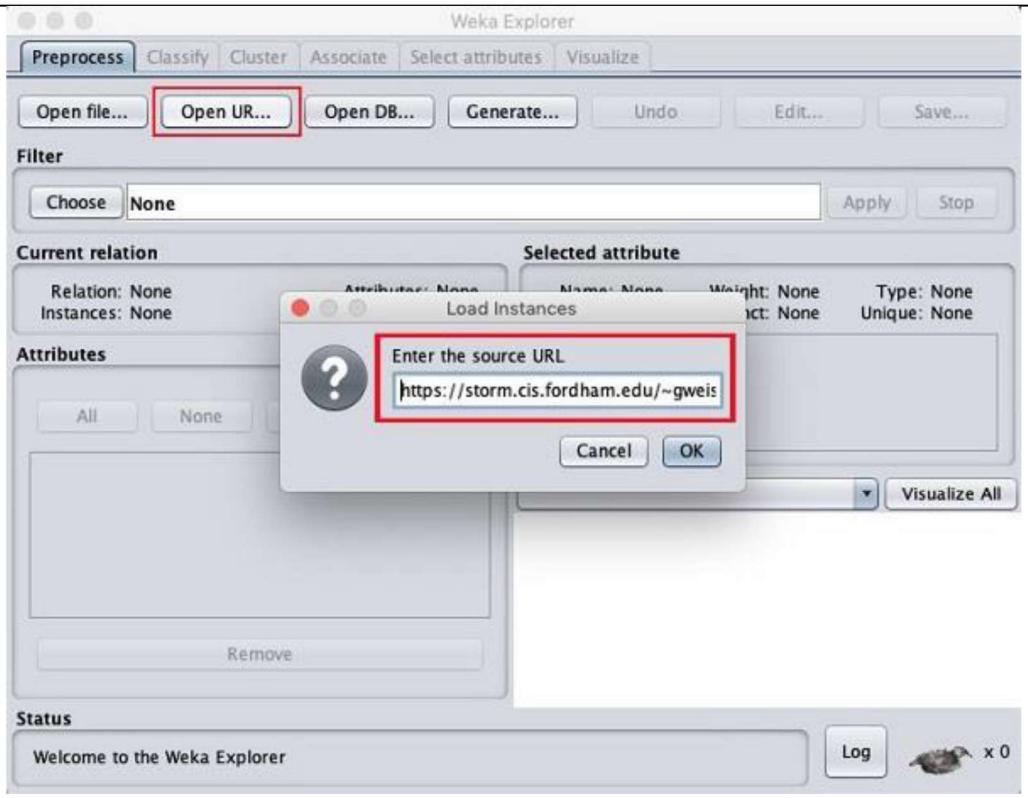


Now, navigate to the folder where your data files are stored. WEKA installation comes up with many sample databases for you to experiment. These are available in the **data** folder of the WEKA installation.

For learning purpose, select any data file from this folder. The contents of the file would be loaded in the WEKA environment. We will very soon learn how to inspect and process this loaded data. Before that, let us look at how to load the data file from the Web.

Loading Data from Web

Once you click on the **Open URL ...** button, you can see a window as follows –



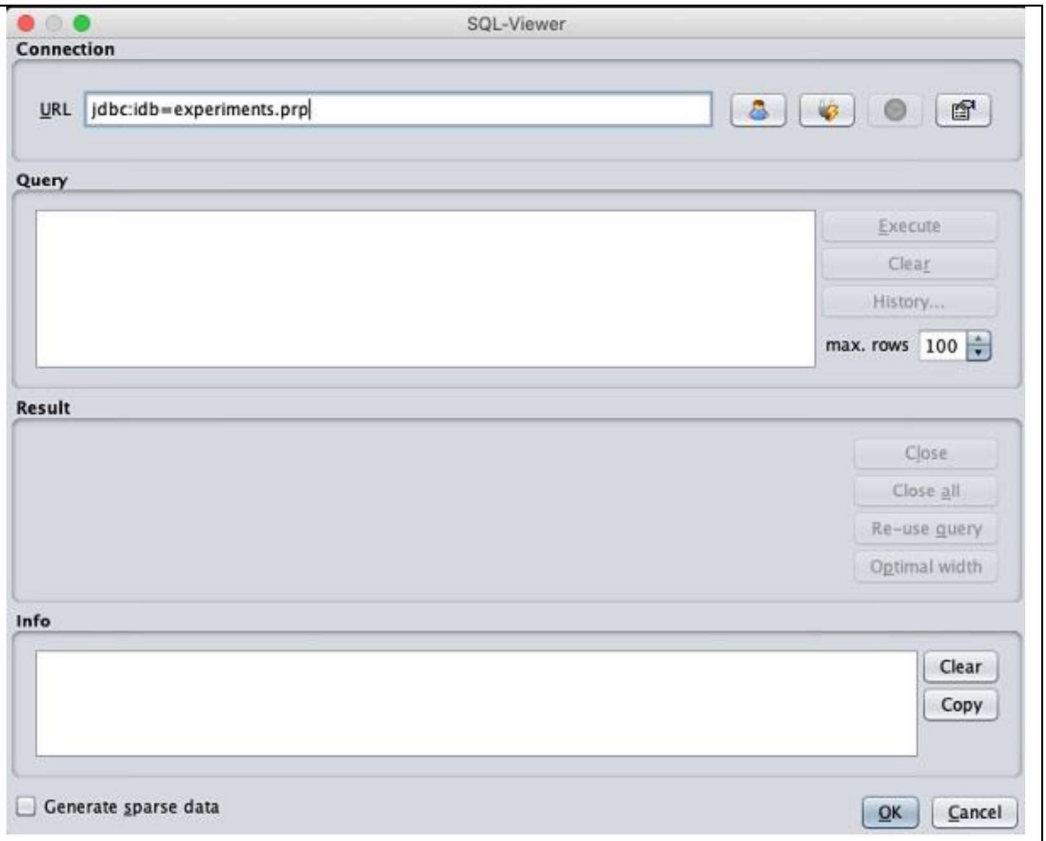
We will open the file from a public URL Type the following URL in the popup box –

https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.nominal.arff

You may specify any other URL where your data is stored. The **Explorer** will load the data from the remote site into its environment.

Loading Data from DB

Once you click on the **Open DB** ... button, you can see a window as follows –



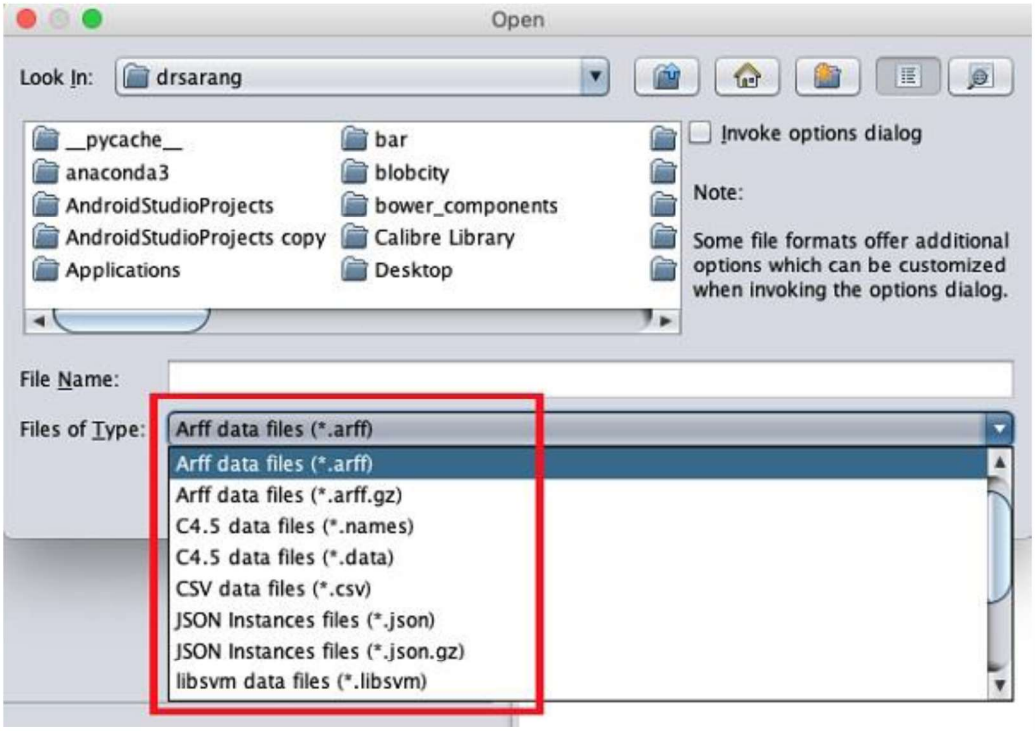
Set the connection string to your database, set up the query for data selection, process the query and load the selected records in WEKA.

WEKA supports a large number of file formats for the data. Here is the complete list −

- arff
- arff.gz
- bsi
- csv
- dat
- data
- json
- json.gz
- libsvm
- m
- names
- xrff

- xrff.gz

The types of files that it supports are listed in the drop-down list box at the bottom of the screen. This is shown in the screenshot given below.



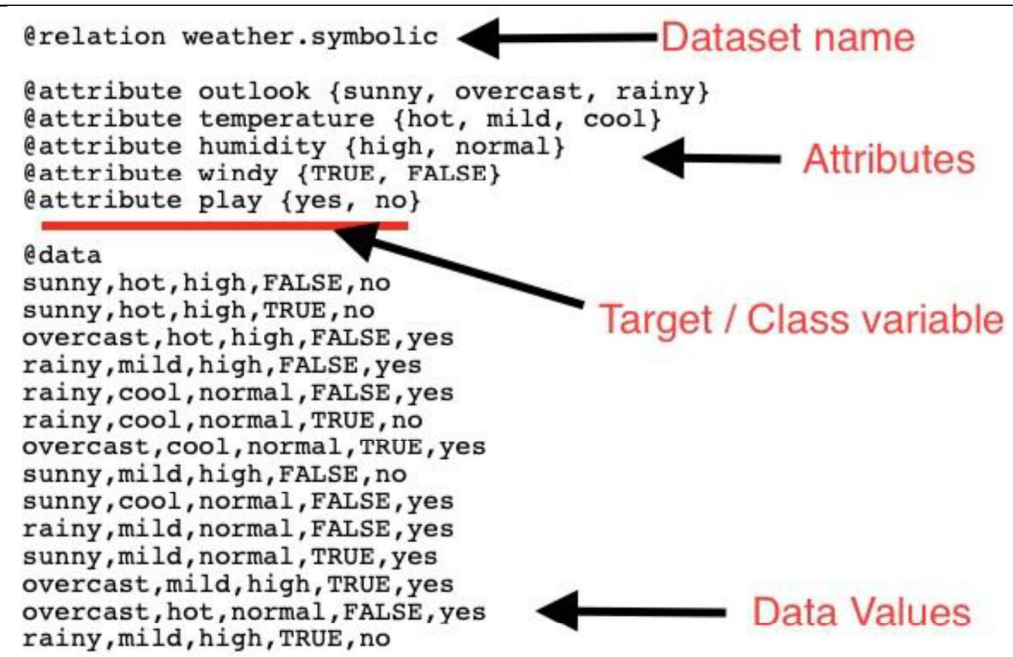
As you would notice it supports several formats including CSV and JSON. The default file type is Arff.

Arff Format

An **Arff** file contains two sections - header and data.

- The header describes the attribute types.
- The data section contains a comma separated list of data.

As an example for Arff format, the **Weather** data file loaded from the WEKA sample databases is shown below –



From the screenshot, you can infer the following points −

- The @relation tag defines the name of the database.
- The @attribute tag defines the attributes.
- The @data tag starts the list of data rows each containing the comma separated fields.
- The attributes can take nominal values as in the case of outlook shown here −

@attribute outlook (sunny, overcast, rainy)

- The attributes can take real values as in this case −

@attribute temperature real

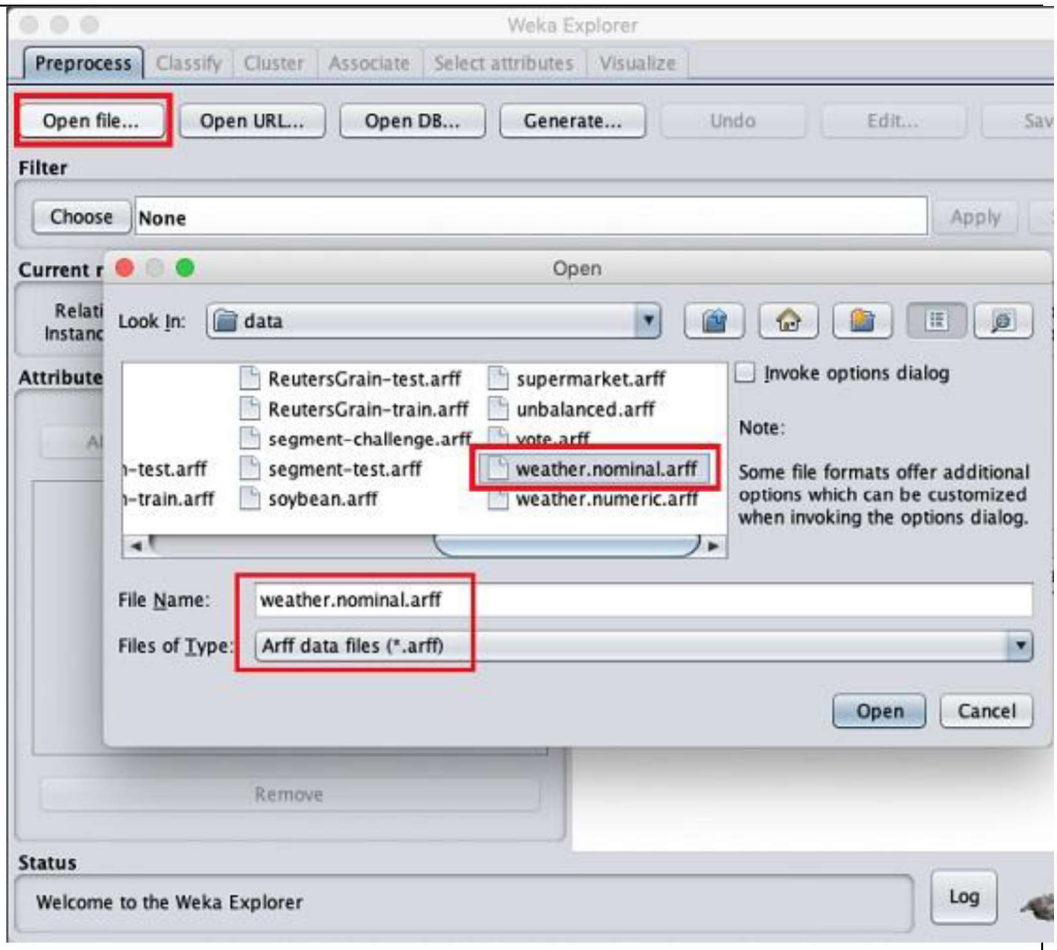
- You can also set a Target or a Class variable called play as shown here −

@attribute play (yes, no)

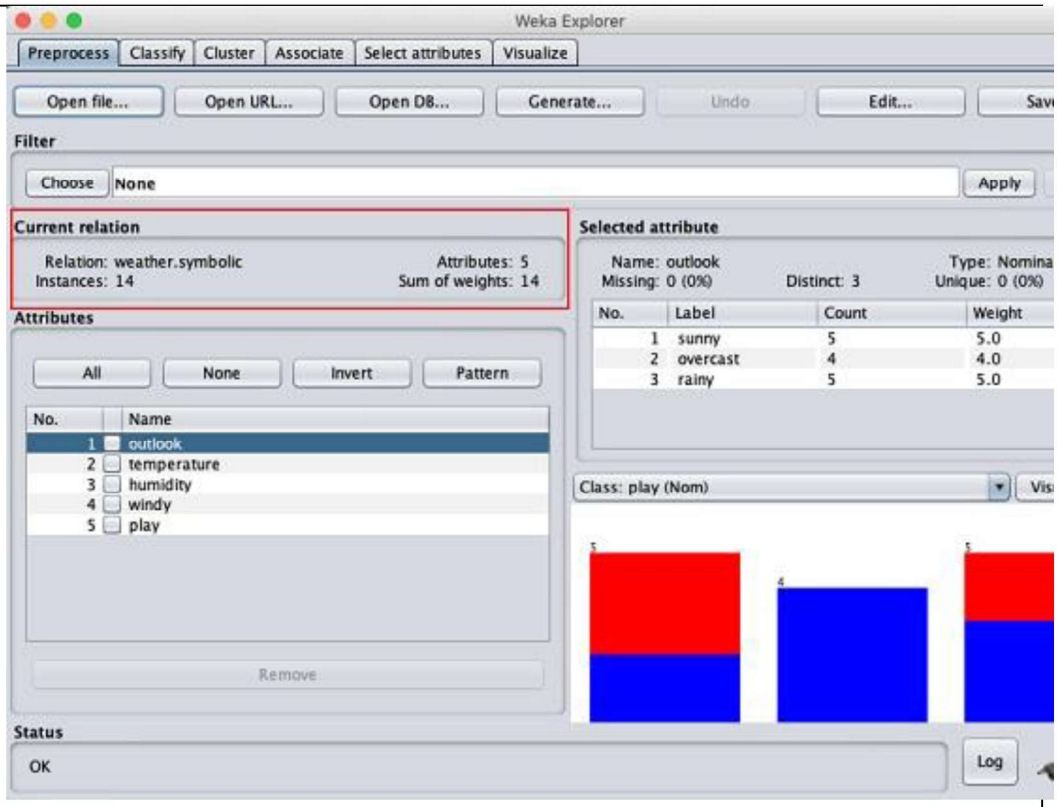
- The Target assumes two nominal values yes or no.

Other Formats

	The Explorer can load the data in any of the earlier mentioned formats. As arff is the preferred format in WEKA, you may load the data from any format and save it to arff format for later use. After preprocessing the data, just save it to arff format for further analysis. **Weka - Preprocessing the Data** The data that is collected from the field contains many unwanted things that leads to wrong analysis. For example, the data may contain null fields, it may contain columns that are irrelevant to the current analysis, and so on. Thus, the data must be preprocessed to meet the requirements of the type of analysis you are seeking. This is the done in the preprocessing module. To demonstrate the available features in preprocessing, we will use the **Weather** database that is provided in the installation. Using the **Open file** ... option under the **Preprocess** tag select the **weather-nominal.arff** file.



When you open the file, your screen looks like as shown here –



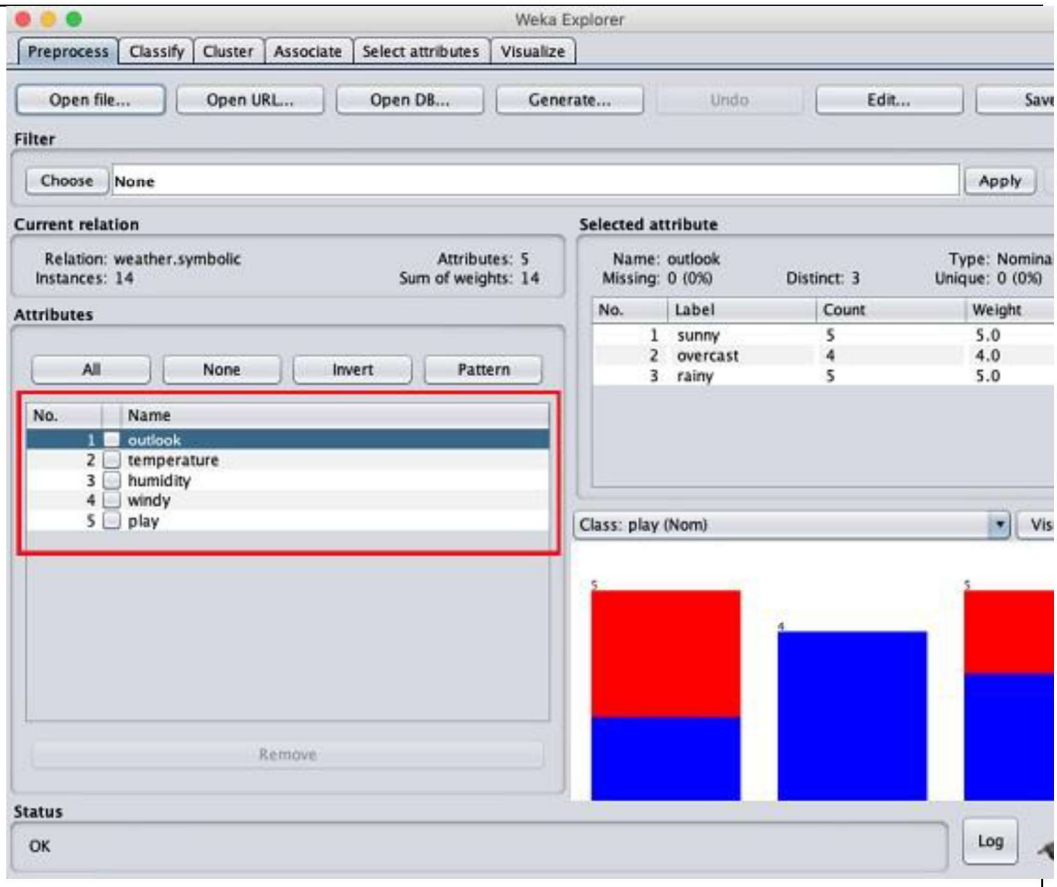
This screen tells us several things about the loaded data, which are discussed further in this chapter.

Understanding Data

Let us first look at the highlighted **Current relation** sub window. It shows the name of the database that is currently loaded. You can infer two points from this sub window –

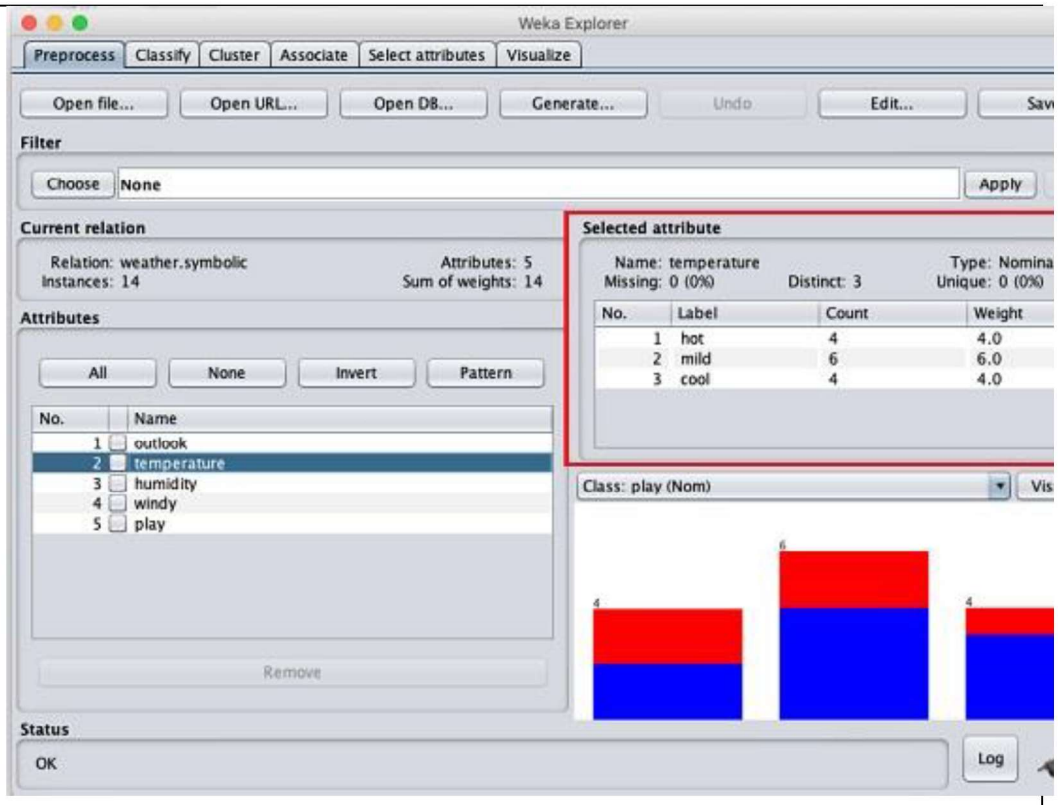
- There are 14 instances - the number of rows in the table.
- The table contains 5 attributes - the fields, which are discussed in the upcoming sections.

On the left side, notice the **Attributes** sub window that displays the various fields in the database.



The **weather** database contains five fields - outlook, temperature, humidity, windy and play. When you select an attribute from this list by clicking on it, further details on the attribute itself are displayed on the right hand side.

Let us select the temperature attribute first. When you click on it, you would see the following screen –

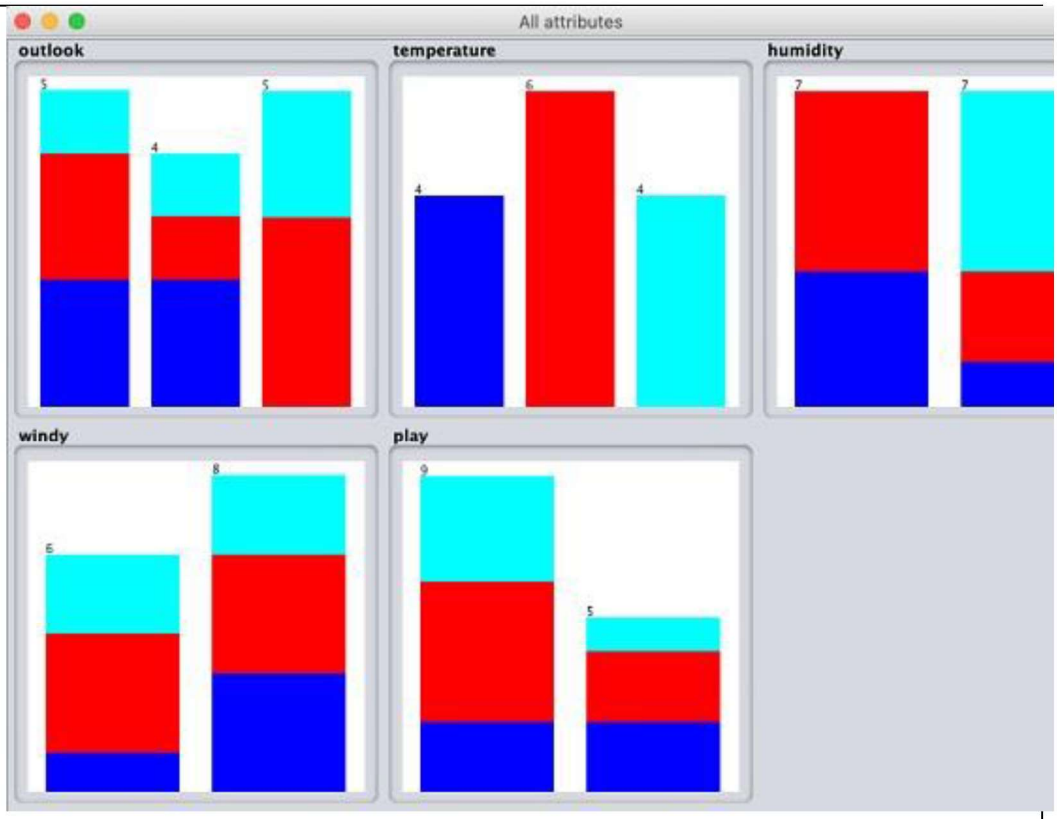


In the **Selected Attribute** subwindow, you can observe the following –

- The name and the type of the attribute are displayed.
- The type for the **temperature** attribute is **Nominal**.
- The number of **Missing** values is zero.
- There are three distinct values with no unique value.
- The table underneath this information shows the nominal values for this field as hot, mild and cold.
- It also shows the count and weight in terms of a percentage for each nominal value.

At the bottom of the window, you see the visual representation of the **class** values.

If you click on the **Visualize All** button, you will be able to see all features in one single window as shown here –



Removing Attributes

Many a time, the data that you want to use for model building comes with many irrelevant fields. For example, the customer database may contain his mobile number which is relevant in analysing his credit rating.

To remove Attribute/s select them and click on the **Remove** button at the bottom.

The selected attributes would be removed from the database. After you fully preprocess the data, you can save it for model building.

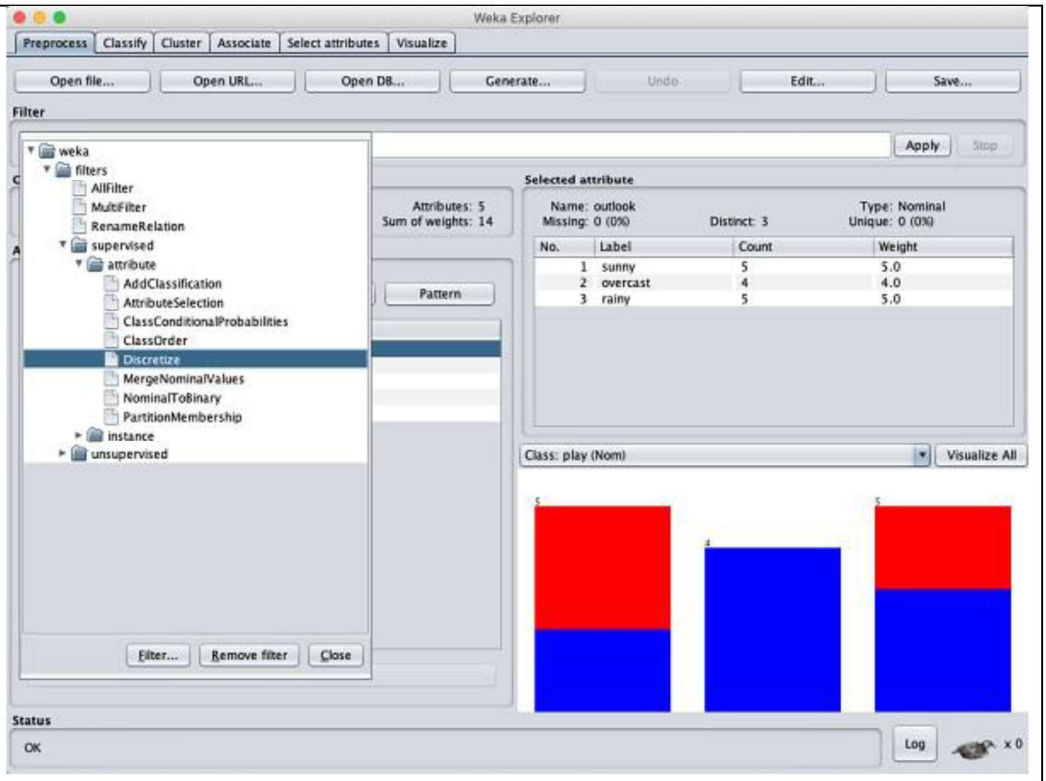
Next, you will learn to preprocess the data by applying filters on this data.

Applying Filters

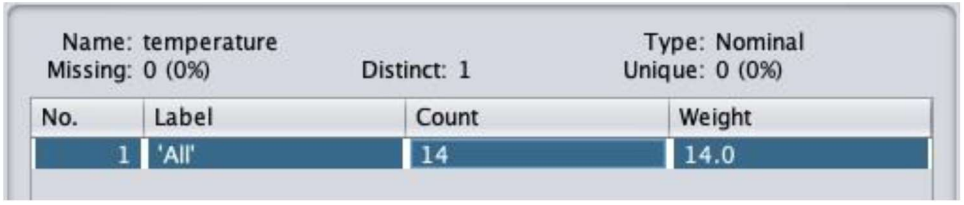
Some of the machine learning techniques such as association rule mining requires categorical data. To illustrate the use of filters, we will use **weather-numeric.arff** database that contains two **numeric** attributes - **temperature** and **humidity**.

We will convert these to **nominal** by applying a filter on our raw data. Click on the **Choose** button in the **Filter** subwindow and select the following filter –

weka→filters→supervised→attribute→Discretize



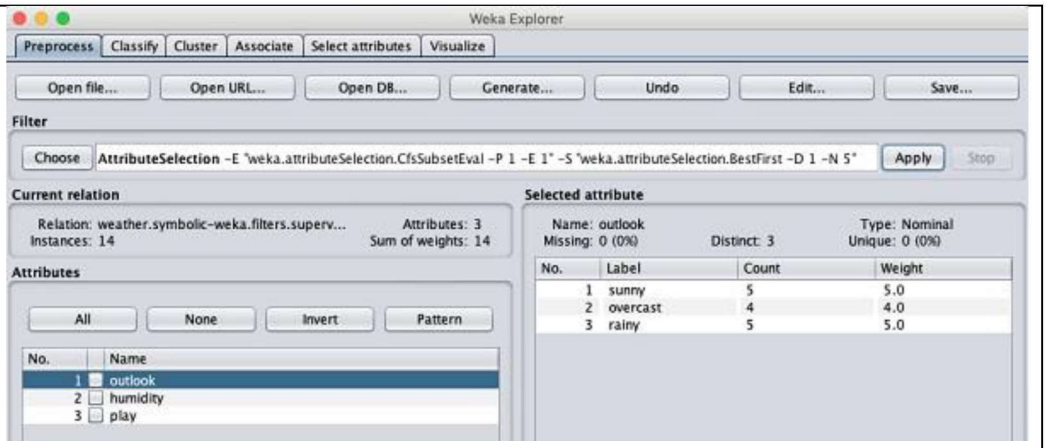
Click on the **Apply** button and examine the **temperature** and/or **humidity** attribute. You will notice that these have changed from numeric to nominal types.



Let us look into another filter now. Suppose you want to select the best attributes for deciding the **play**. Select and apply the following filter –

weka→filters→supervised→attribute→AttributeSelection

You will notice that it removes the temperature and humidity attributes from the database.



After you are satisfied with the preprocessing of your data, save the data by clicking the **Save** ... button. You will use this saved file for model building.

PROBLEM STATEMENTS

1.1 Create an ARFF File in Weka as the specification given below.

Dataset student .arff

@relation student

@attribute age {<30,30-40,>40}

@attribute income {low, medium, high} @attribute student {yes, no}

@attribute credit-rating {fair, excellent} @attribute buyspc {yes, no}

@data

30, high, no, fair, no

30, high, no, excellent, no

30-40, high, no, fair, yes

40, medium, no, fair, yes

40, low, yes, fair, yes

40, low, yes, excellent, no

30-40, low, yes, excellent, yes

30, medium, no, fair, no

	30, low, yes, fair, no 40, medium, yes, fair, yes 30, medium, yes, excellent, yes 30-40, medium, no, excellent, yes 30-40, high, yes, fair, yes 40, medium, no, excellent, no
1.2	Create weather dataset as given below.

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

1.3	Load the weather dataset. And find the answer of following questions. A. What are the values that the attribute temperature can have? B. How many instances does this dataset have? C. How many attributes does this dataset have?
1.4	A. Remove all instances in which the humidity attribute has the value high. B. Undo the change to the dataset that you just performed, and verify that the data has reverted to its original state.
1.5	Load iris dataset and perform the following statistical analysis of data. A. Discretization B. Missing Values C. Numeric Transform

Reference questions for Viva:

1. What is the ARFF file format in the context of Weka?

ARFF stands for Attribute-Relation File Format, a plain text file format in Weka used to describe datasets.

2. How do you load a dataset in Weka using an ARFF file?

To load a dataset, use the "Explorer" interface and click "Open file," selecting the ARFF file. Alternatively, use the command line with **java -cp**.

3. What are the typical pre-processing operations in Weka?

Pre-processing in Weka includes handling missing values, normalization, attribute selection, and filtering.

4. How do you handle missing values in a dataset using Weka?

Weka offers options like removing instances, replacing with mean/median, or using advanced techniques. The "ReplaceMissingValues" filter is commonly used.

5. Explain the significance of attribute selection in Weka pre-processing.

Attribute selection in Weka is crucial to improve model performance, reduce dimensionality, and mitigate the impact of irrelevant or redundant features.

6. How can you apply a filter to preprocess data in Weka?

Apply filters through the "Explorer" interface or use the command-line interface with the **weka.filters** package to apply filters programmatically.

7. What is the purpose of the "Normalize" filter in Weka during dataset pre-processing?

The "Normalize" filter in Weka is used to scale numeric attributes, ensuring they fall within a specified range. This is crucial for algorithms sensitive to attribute scales, such as distance-based methods.

8. Can you explain the concept of imputation and its role in handling missing values in Weka?

Imputation in Weka involves estimating and replacing missing values with predicted or calculated values based on the available data. It helps maintain dataset integrity by filling in gaps caused by missing information.

9. How does Weka handle categorical data, and what preprocessing steps are commonly applied to such data?

Weka often requires converting categorical data into numeric forms for modeling. Preprocessing steps include one-hot encoding, label encoding, or using specific filters like the "Nominal to Binary" filter.

10. In Weka, what is the purpose of the "Remove" filter, and when would you use it during pre-processing?

The "Remove" filter in Weka is used to eliminate specified attributes from the dataset. It is employed during pre-processing when certain attributes are deemed irrelevant or when feature selection is necessary for model building.

11. How can you check for and handle duplicate instances in a dataset using Weka?

Weka provides the "RemoveDuplicates" filter to identify and eliminate duplicate instances in a dataset during pre-processing, ensuring that the data used for modeling is unique.

12. What is the significance of the "AttributeSelection" filter in Weka, and how does it contribute to model development?

The "AttributeSelection" filter in Weka is crucial for choosing the most relevant attributes, aiding in model performance improvement and reducing computational complexity by selecting only the essential features for analysis.

मैट्स विश्वविद्यालय
MATS UNIVERSITY

MATS SCHOOL OF INFORMATION TECHNOLOGY

University Campus: Gullu, Aarang, Raipur – 493441 | Raipur Campus: MATS Tower, Pandri, Raipur – 492004

Exam: Unit Test **Program: MCA** **Semester: II**

Subject Code: 202 **Subject: Data Warehousing and Data Mining** **Syllabus Year: 2023-24**

Max. Marks: 30 Student Id: Max. Time: 1 Hour

Q. No.	Question Description	CO	PO	PSO	BTL
	Section – A (Attempt **ALL** questions. Each question carries 2 marks.)				
Q. 1		1	2	-	2
Q. 2					
Q. 3					
	Section – B (Attempt any **THREE** questions. Each question carries 8 marks.)				
Q. 4					
Q. 5					
Q. 6					
Q. 7					

मैट्स विश्वविद्यालय
MATS UNIVERSITY

MATS SCHOOL OF INFORMATION TECHNOLOGY

University Campus: Gullu, Aarang, Raipur – 493441 | Raipur Campus: MATS Tower, Pandri, Raipur – 492004

Exam: Mid Term **Program: MCA** **Semester: II**

Subject Code: 2102 **Subject: Data Warehousing and Data Mining** **Syllabus Year: 2023-24**

Max. Marks: 50 Student Id: Max. Time: 2 Hours

Q. No.	Question Description	CO	PO	PSO	BTL
	Section – A *(Attempt **ALL** questions. Each question carries 2 marks.)*				
Q. 1		1	1,5	-	2
Q. 2		3	1,2,5, 6	-	2
Q. 3		3	1,2,5, 6	-	2
Q. 4		1	1,5	-	2
	Section – B (Attempt any **FOUR** questions. Each question carries 8 marks. Answer in 300-400 words)				
Q. 5		1	1,5	-	2
Q. 6		1	1,5	-	2
Q. 7		2	1,2,5	-	2
Q. 8		2	1,2,5	-	3,4
Q. 9		1	1,5	-	2
Q. 10	A.	2	1,2,5	-	2
	Section – C (Attempt any **ONE** question. Each question carries 10 marks. Answer in 600 words)				
Q. 11		3	1,2,5, 6	-	3
Q. 12		3	1,2,5, 6	-	3