Data Warehousing & Data Mining LAB - G2
EXPERIMENT 3

- ASHISH KUMAR
- 2K18/SE/041

Aim:- 1. Briefly describe WEKA key features.
2. Describe understanding of any 3 inbuilt dataset in WEKA.

Theory:-

Features

The Weka machine learning workbench is a modern platform for applied machine learning. Weka is an acronym which stands for Waikato Environment for Knowledge Analysis. It is also the name of a New Zealand bird the Weka.

Some of the features include,

- **Open Source:** It is released as open source software under the GNU GPL. It is dual licensed and Pentaho Corporation owns the exclusive license to use the platform for business intelligence in their own product.
- **Graphical Interface:** It has a Graphical User Interface (GUI). This allows you to complete your machine learning projects without programming.
- **Command Line Interface:** All features of the software can be use from the command line. This can be very useful for scripting large jobs.
- **Java API:** It is written in Java and provides an API that is well documented and promotes integration into your own applications. Note that the GNU GPL means that in turn your software would also have to be released as GPL.
- **Documentation:** There books, manuals, wikis and MOOC courses that can train you how to use the platform effectively.

The WEKA workbench provides three main ways to work on your problem: The Explorer for playing around and trying things out, the Experimenter for controlled experiments, and the Knowledge Flow for graphically designing a pipeline for your problem.

Datasets

The WEKA machine learning tool provides a directory of some sample datasets. These datasets can be directly loaded into WEKA for users to start developing models immediately.
The WEKA datasets can be explored from the "C:\Program Files\Weka-3-8\data" link. The datasets are in .arff format.

1) contact-lens.arff

contact-lens.arff dataset is a database for fitting contact lenses. It was donated by the donor, Benoit Julien in the year 1990.

```
@relation contact-lenses

@attribute age                  {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip   {myope, hypermetrope}
@attribute astigmatism          {no, yes}
@attribute tear-prod-rate       {reduced, normal}
@attribute contact-lenses       {soft, hard, none}

@data
%
% 24 instances
%
young,myope,no,reduced,none
young,myope,no,normal,soft
young,myope,yes,reduced,none
young,myope,yes,normal,hard
young,hypermetrope,no,reduced,none
young,hypermetrope,no,normal,soft
young,hypermetrope,yes,reduced,none
young,hypermetrope,yes,normal,hard
pre-presbyopic,myope,no,reduced,none
pre-presbyopic,myope,no,normal,soft
pre-presbyopic,myope,yes,reduced,none
pre-presbyopic,myope,yes,normal,hard
pre-presbyopic,hypermetrope,no,reduced,none
pre-presbyopic,hypermetrope,no,normal,soft
pre-presbyopic,hypermetrope,yes,reduced,none
pre-presbyopic,hypermetrope,yes,normal,none
presbyopic,myope,no,reduced,none
presbyopic,myope,no,normal,none
presbyopic,myope,yes,reduced,none
presbyopic,myope,yes,normal,hard
presbyopic,hypermetrope,no,reduced,none
presbyopic,hypermetrope,no,normal,soft
presbyopic,hypermetrope,yes,reduced,none
presbyopic,hypermetrope,yes,normal,none
```

Database: This database is complete. The examples used in this database are complete and noise-free. The database has 24 instances and 4 attributes.

Attributes: All four attributes are nominal. There are no missing attribute values.
The four attributes are as follows:

1) Age of the patient: The attribute age can take values:
- young
- pre-presbyopic
- presbyopic

2) Spectacle prescription: This attribute can take values:
- myope
- hypermetrope

3) Astigmatic: This attribute can take values

- no
- yes

4) Tear production rate: The values can be

- reduced
- normal

5) Class: Three class labels are defined here. These are:

- the patient should be fitted with hard contact lenses.
- the patient should be fitted with soft contact lenses.
- the patient should not be fitted with contact lenses.

Class Distribution: The instances that are classified into class labels are enlisted below:

S.No.	Class Label	No. of Instances
1.	Hard contact lenses	4
2.	Soft contact lenses	5
3.	No contact lenses	15

2) iris.arff

iris.arff dataset was created in 1988 by Michael Marshall. It is the Iris Plants database.

```
@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
5.2,3.4,1.4,0.2,Iris-setosa
4.7,3.2,1.6,0.2,Iris-setosa
4.8,3.1,1.6,0.2,Iris-setosa
5.4,3.4,1.5,0.4,Iris-setosa
5.2,4.1,1.5,0.1,Iris-setosa
5.5,4.2,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
```

Database: This database is used for pattern recognition. The data set contains 3 classes of 50 instances. Each class represents a type of iris plant. One class is linearly separable from the other 2 but the latter are not linearly separable from each other. It predicts to which species of the 3 iris flower the observation belongs. This is called a multi-class classification dataset.

Attributes: It has 4 numeric, predictive attributes, and the class. There are no missing attributes.

The attributes are:

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm
- class:
    -- Iris Setosa
    -- Iris Versicolour
    -- Iris Virginica

Summary Statistics:

	Min	Max	Mean	SD	Class Correlation
sepal length	4.3	7.9	5.84	0.83	0.7826
sepal width	2.0	4.4	3.05	0.43	-0.4194
petal length	1.0	6.9	3.76	1.76	0.9490 (high)
petal length	0.1	2.5	1.20	0.76	0.9565 (high)

Class Distribution: 33.3% for each of 3 classes.

3) diabetes.arff

The database of this dataset is Pima Indians Diabetes. This dataset predicts whether the patient is prone to be diabetic in the next 5 years. The patients in this dataset are all females of at least 21 years of age from Pima Indian Heritage. It has 768 instances and 8 numerical attributes plus a class. This is a binary classification dataset where the output variable predicted is nominal comprising of two classes.

Number of Instances: 768

Number of Attributes: 8 plus class

For Each Attribute: (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

```
Relabeled values in attribute 'class'
    From: 0                       To: tested_negative
    From: 1                       To: tested_positive

@relation pima_diabetes
@attribute 'preg' real
@attribute 'plas' real
@attribute 'pres' real
@attribute 'skin' real
@attribute 'insu' real
@attribute 'mass' real
@attribute 'pedi' real
@attribute 'age' real
@attribute 'class' { tested_negative, tested_positive}
@data
6,148,72,35,0,33.6,0.627,50,tested_positive
1,85,66,29,0,26.6,0.351,31,tested_negative
8,183,64,0,0,23.3,0.672,32,tested_positive
1,89,66,23,94,28.1,0.167,21,tested_negative
0,137,40,35,168,43.1,2.288,33,tested_positive
5,116,74,0,0,25.6,0.201,30,tested_negative
3,78,50,32,88,31,0.248,26,tested_positive
10,115,0,0,0,35.3,0.134,29,tested_negative
2,197,70,45,543,30.5,0.158,53,tested_positive
8,125,96,0,0,0,0.232,54,tested_positive
4,110,92,0,0,37.6,0.191,30,tested_negative
10,168,74,0,0,38,0.537,34,tested_positive
10,139,80,0,0,27.1,1.441,57,tested_negative
1,189,60,23,846,30.1,0.398,59,tested_positive
5,166,72,19,175,25.8,0.587,51,tested_positive
7,100,0,0,0,30,0.484,32,tested_positive
0,118,84,47,230,45.8,0.551,31,tested_positive
7,107,74,0,0,29.6,0.254,31,tested_positive
1,103,30,38,83,43.3,0.183,33,tested_negative
1,115,70,30,96,34.6,0.529,32,tested_positive
3,126,88,41,235,39.3,0.704,27,tested_negative
8,99,84,0,0,35.4,0.388,50,tested_negative
7,196,90,0,0,39.8,0.451,41,tested_positive
9,119,80,35,0,29,0.263,29,tested_positive
11,143,94,33,146,36.6,0.254,51,tested_positive
10,125,70,26,115,31.1,0.205,41,tested_positive
7,147,76,0,0,39.4,0.257,43,tested_positive
```

Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

Class Value	Number of instances
0	500
1	268

Brief statistical analysis:

Attribute number	Mean	Standard Deviation
1.	3.8	3.4
2.	120.9	32.0
3.	69.1	19.4
4.	20.5	16.0
5.	79.8	115.2
6.	32.0	7.9
7.	0.5	0.3
8.	33.2	11.8

Findings and Learning: WEKA provides many features and all of them are very useful in its own. It also provides various inbuilt datasets which can be used to analyse it and for predicative modeling as well. We have successfully understood 3 inbuilt datasets.