

## **Assignment 5**

**AIM:** Study of platform for Implementation of Assignments. Download the open source software of your interest. Document the distinct features and functionality of the software platforms WEKA, R and Python.

### **OBJECTIVE:**

To study

- Concept of open source analytical software (WEKA, R and Python)
- Concept of statistical analysis.
- Distinct features and functionality of open source software.
- Open source tools and verify execution of programs on different inputs dataset.

### **THEORY:**

#### 1. Introduction:

##### 1) *Introduction of WEKA*

Weka is open source software under the GNU General Public License. System is developed at the University of Waikato in New Zealand. "Weka" stands for the Waikato Environment for Knowledge Analysis. The software is freely available at <http://www.cs.waikato.ac.nz/ml/weka>. The system is written using object-oriented language Java. There are several different levels at which Weka can be used. Weka provides implementations of state-of-the-art data mining and machine learning algorithms. Weka contains modules for data pre-processing, classification, clustering and association rule extraction.

##### 2) *Introduction of R*

R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. S was created by John Chambers while at Bell Labs. There are some important differences, but much of the code written for S runs unaltered. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the *R Development Core Team*, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S. R is a GNU project. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. R uses a command line interface; there are also several graphical front-ends for it.

### 3) Introduction to Python

Python is a widely used general-purpose, high level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions: Python 2 and Python 3. Both are quite different.

## 2. Concept/Working:

### a) Steps to download and configure the WEKA

- i) Download Weka from <http://www.cs.waikato.ac.nz/ml/weka/>
- ii) Choose a self-extracting executable (including Java VM)
- iii) After download is completed, run the self-extracting file to install Weka, and use the default set-up.



### iv) Working of WEKA

The general working steps are given below by considering the example of Hierarchical clustering.

1. Select a dataset for example iris.
2. Select option Cluster
3. Choose cluster type: Hierarchical Cluster
4. Select cluster mode: Training set. Click on Start.

### v) Features of WEKA

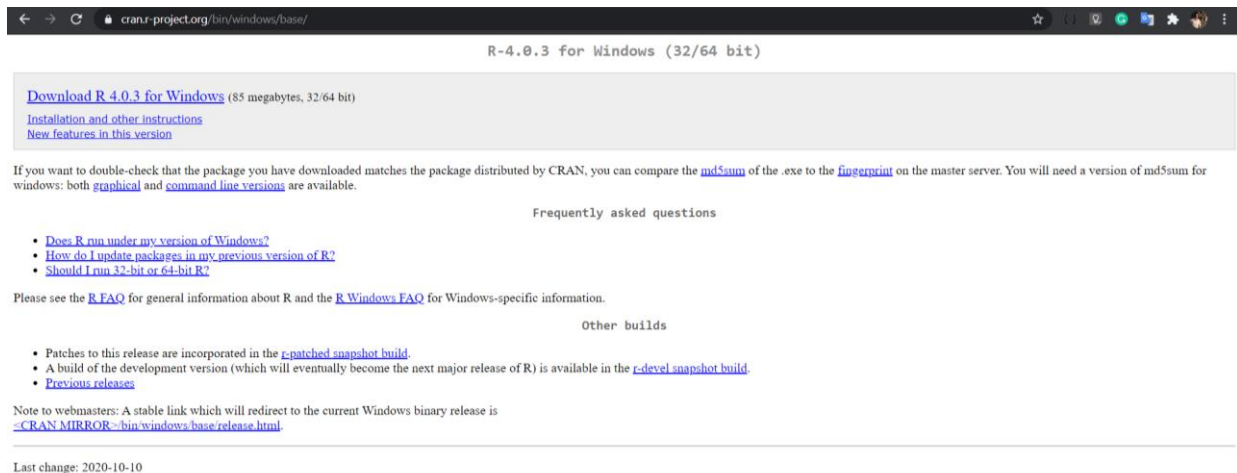
Main features of Weka include:

1. 49 data pre-processing tools
2. 76 classification/regression algorithms
3. 8 clustering algorithms
4. 15 attribute/subset evaluators + 10 search algorithms for feature selection.
5. 3 algorithms for finding association rules
6. 3 graphical user interfaces

- a. “The Explorer” (exploratory data analysis)
- b. “The Experimenter” (experimental environment)
- c. “The Knowledge Flow” (new process model inspired interface)

*b) Steps to download and configure the R*

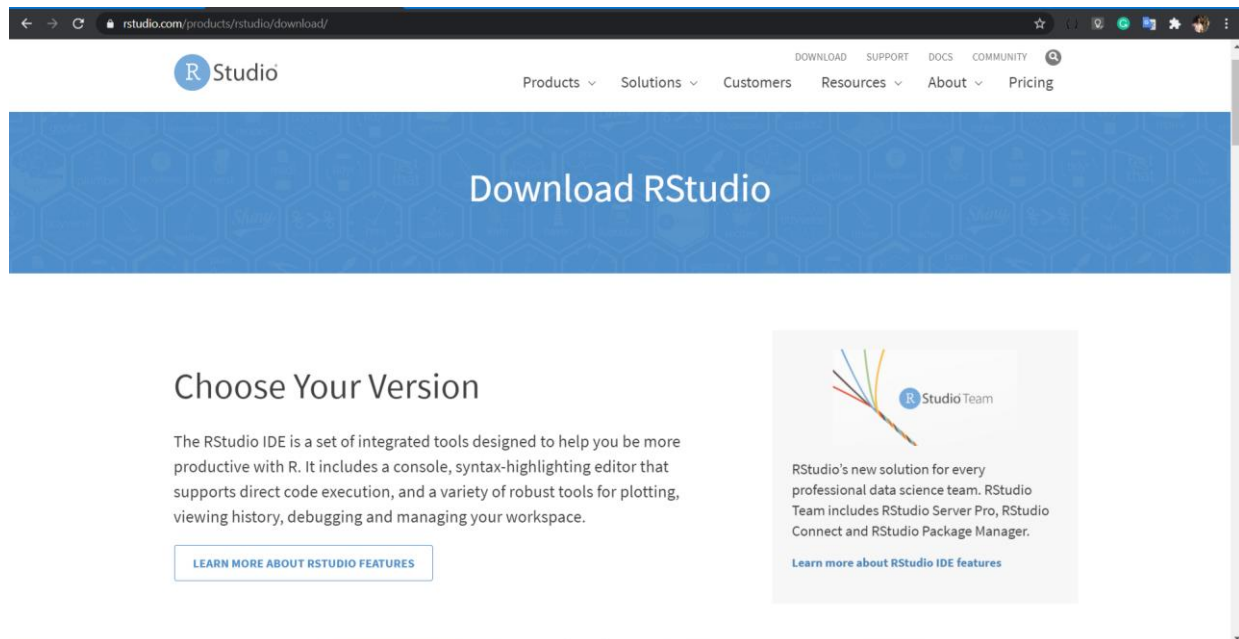
1. Install r-base : Write this command in Command Prompt :  
sudo apt-get install r-base or From CRAN for Windows.



2. Type R on terminal/Command line to get command line for R programming.
3. URL for Rstudio :
  - (a) <http://www.rstudio.com/products/rstudio/download/>
  - (b) Write this command in Command Prompt: sudo apt-get install r-base

OR

  - (2) Open Ubuntu Software Center
    - (a) Search: R-studio
    - (b) Install: R-Studio



#### a) Working of R

The default panes:

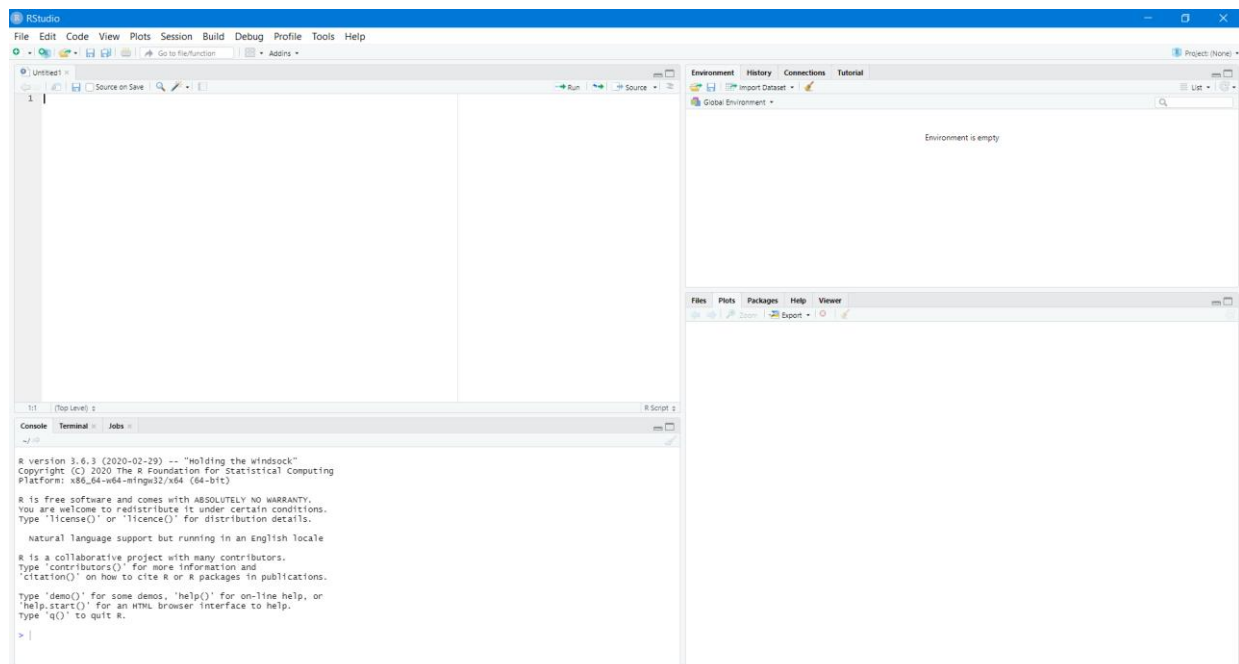
- Console (entire left)
- Workspace/History (tabbed in upper right)
- Files/Plots/Packages/Help (tabbed in lower right)

1. Download package needed for that program.
2. Open RStudio.
3. Install the Required Package. Go to Packages - Click on "Install". You will get "Install Packages" Window.

Install from: Package Archive File(tar.gz)

Package archive: (Browse the path where you have stored the Package\_name.tar.gz )

Click Install



4. In Packages - Tick the Package\_name in user library. Click on Update. Click on OK.

5. Write Following Code in R Script (File - New File - R Script) and Save it as "Program1.R"

```
library(e1071) data(iris) set.seed(123)
```

```
cm1 <- cmeans(iris[,1:4],10)
```

```
bc1<-
```

```
  bclust(iris[,1:4],3,base.centers=20,iter.base=50,base.method="cmeans"
```

```
)
```

6. Click on run Icon.

7. Write following commands on 'console' which is in 'RStudio'.

```
> data("iris")
```

```
> library(class)
```

```
> library(e1071)
```

```
> pairs(iris[1:4], main = "Iris Data(red=setosa,green=versicolor,
blue=virginica)",pch= 21, bg = c("red","green3","blue")
[unclass (iris$Species)])
```

```
> data(iris)
```

```
> summary(iris)
```

```
> classifier<-naiveBayes(iris[,1:4], iris[,5])
```

```
> table(predict(classifier, iris[, -5]), iris[,5])
```

b) Features of R Statistical Features

i) Implement a wide variety of statistical and graphical techniques.

ii) R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages.

iii) For computationally intensive tasks, C, C++, and Fortran code can be linked and called at run time.

Programming Features

R is an interpreted language.

v) R's data structures include vectors, matrices, arrays, data frames (similar to tables in a relational database) and lists.

vi) R supports procedural programming with functions and, for some functions, object-oriented programming with generic functions.

c) *Steps to download and configure Python*

i) Download the latest version of Python.

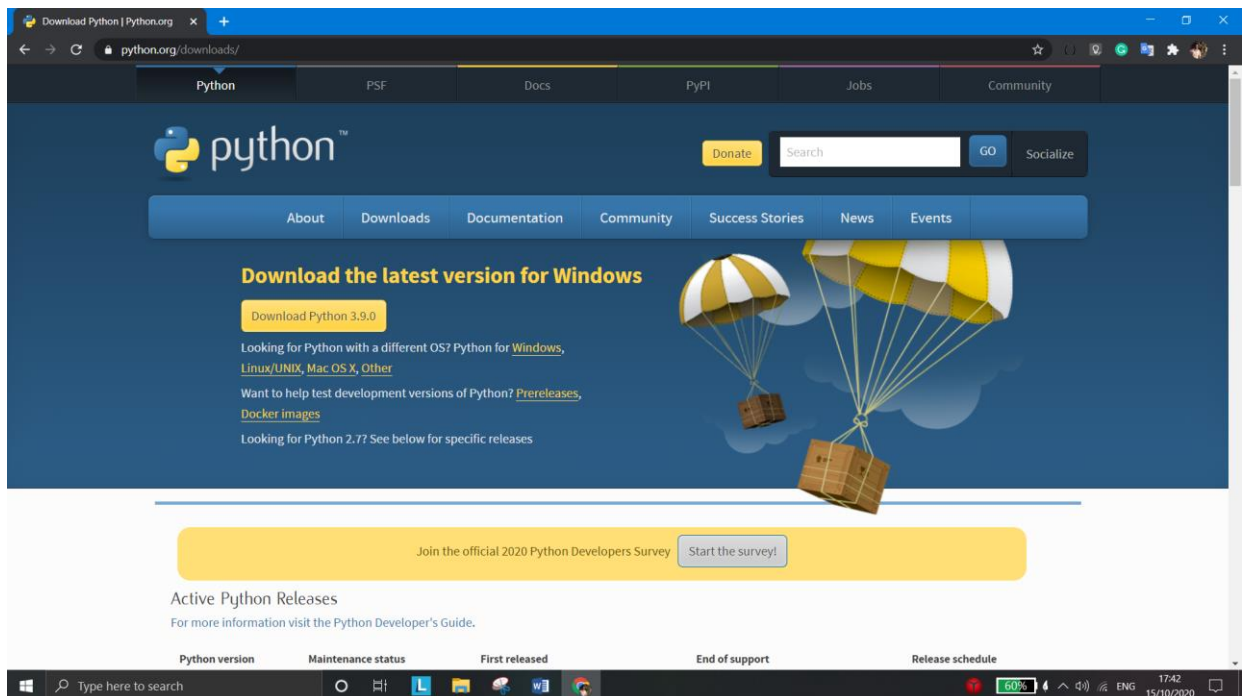
ii) Run the installer file and follow the steps to install Python

iii) During the install process, check Add Python to environment variables.

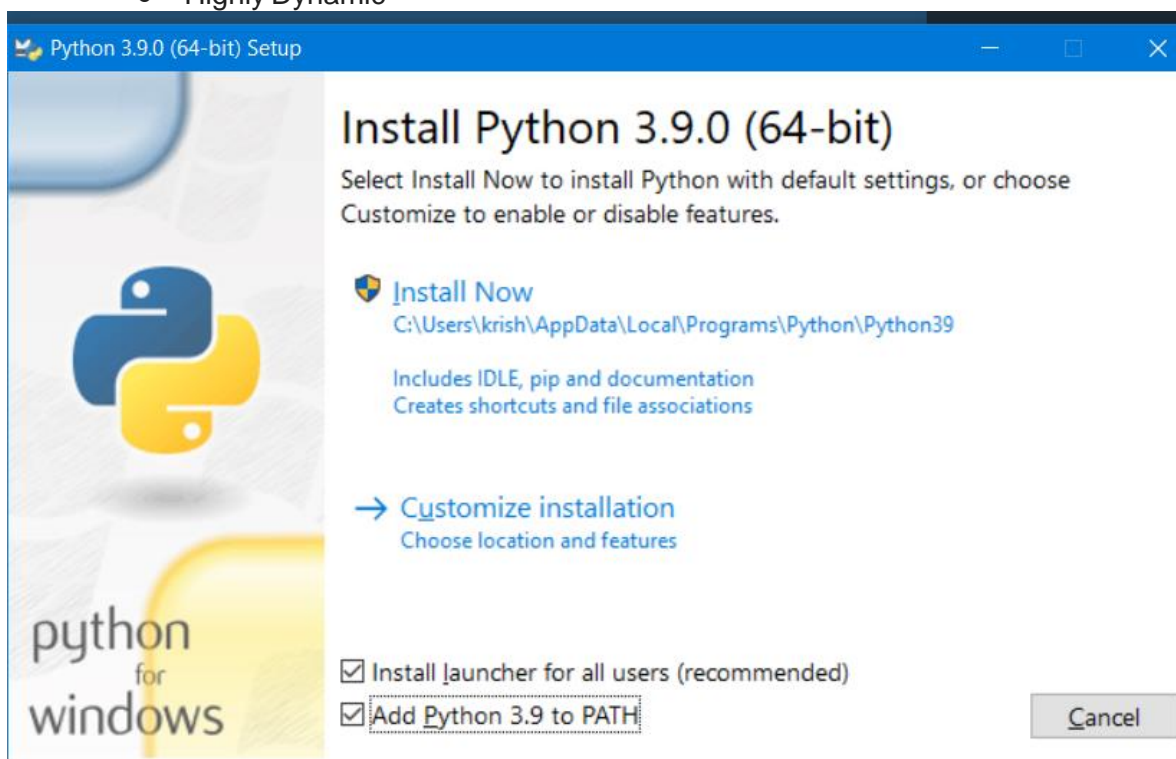
iv) This will add Python to environment variables, and you can run Python from any part of the computer.

v) Also, you can choose the path where Python is installed.

### Python website



- **Features of Python**
- Easy to Code
  - Open Source and Free
  - Support for GUI
  - Object-Oriented Approach
  - High-Level Language
  - Integrated by Nature
  - Highly Portable
  - Highly Dynamic



## Advantages/Disadvantages:

- **Advantages of WEKA.**

Free availability:

- Under the GNU General Public License
  - Portability
  - Fully implemented in the Java programming language and thus runs
- on almost any modern computing platforms
  - Windows, Mac OS X and Linux
  - Comprehensive collection of data preprocessing and modeling
- techniques
  - Supports standard data mining tasks: data preprocessing, clustering,
- classification, regression, visualization, and feature selection.
  - Easy to use GUI
  - Provides access to SQL databases
  - Using Java Database Connectivity and can process the result
- returned by a database query.
- The obvious advantage of a package like Weka is that a whole range of data preparation, feature selection and data mining algorithms are integrated. This means that only one data format is needed, and trying out and comparing different approaches becomes really easy. The package also comes with a GUI, which should make it easier to use.

- **Disadvantages of WEKA.**

- Sequence modeling is not covered by the algorithms included in the Weka distribution.
- Not capable of multi-relational data mining.
- Memory bound.
- Do not implement the newest techniques. For example the MLP implemented has a very basic training algorithm (backprop with momentum), and the SVM only uses polynomial kernels, and does not support numeric estimation. Therefore, it will be necessary to combine WEKA with some of the other tools like Netlab or SVM\_torch.
- Though the software is for free: the documentation for the GUI is quite limited.
- Limited scaling. For difficult tasks on large datasets, the running time can become quite long, and java sometimes gives an OutOfMemory error. This problem can be reduced by using the '-mxm' option when calling java, where x is memory size (eg '50m'). For large datasets it will always be necessary to reduce the size to be able to work within reasonable time limits.
- The GUI does not implement all the possible options. Things that could be very useful, like scoring of a test set, are not provided in the GUI, but can be called from the command line interface. So sometimes it will be necessary to switch between GUI and command line.

- **Advantages of R.**

- R is free and open source software, allowing anyone to use and, importantly, to modify it. R is licensed under the GNU General Public License, with copyright held by The R Foundation for Statistical computing.
- The graphical capabilities of R are outstanding, providing a

fully programmable graphics language that surpasses most other statistical and graphical packages.

- R is a programming language and environment developed for statistical analysis by practicing statisticians and researchers.

- **Disadvantages of R.**

- R is not so easy to use for the novice.
- Many R commands give little thought to memory management, and so R can very quickly consume all available memory.
- The quality of some packages is less than perfect

- **Advantages of Python**

- **Less Coding** : Almost all of the tasks done in Python requires less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don't have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python to beginners.
- **Affordable** : Python is free therefore individuals, small companies or big organizations can leverage the free available resources to build applications. Python is popular and widely used so it gives you better community support.
- **Python is for Everyone** : Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and *machine learning*, automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

- **Disadvantages of Python**

- **Speed Limitations**

We have seen that Python code is executed line by line. But since Python is interpreted, it often results in slow execution. This, however, isn't a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

- **Weak in Mobile Computing and Browsers**

While it serves as an excellent server-side language, Python is much rarely seen on the client-side. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is called Carbonnelle.

The reason it is not so famous despite the existence of Brython is that it isn't that secure.

- **Design Restrictions**

As you know, Python is dynamically-typed. This means that you don't need to declare the type of variable while writing the code. It uses duck-typing. But wait,



what's that? Well, it just means that if it looks like a duck, it must be a duck. While this is easy on the programmers during coding, it can raise run-time errors.

- Underdeveloped Database Access Layers

#### **Any Software or Hardware Used:**

- **Hardware or software required for WEKA.**

Hardware:

- 4GB RAM

Software:

- Java
- 64-bit / 32-bits versions of Windows.
- 64-bit / 32-bits Linux

- **Hardware or Software required for R.**

Hardware:

- ★ The amount of RAM that you need is highly dependent on the work/analysis you will be doing. (More than 1 GB of RAM.)

Software:

- 64-bit / 32-bits versions of Windows.
- 64-bit / 32-bits Linux

- **Hardware or Software required for Python.**

Hardware:

- x86 64-bit CPU (Intel / AMD architecture) (x64 preferred)
- 4 GB RAM
- 5 GB free disk space

Software:

- 64-bit / 32-bits versions of Windows.
- 64-bit / 32-bits Linux
- Modern Operating System:
  - Windows 7 or 10
  - Mac OS X 10.11 or higher, 64-bit
  - Linux: RHEL 6/7, 64-bit (almost all libraries also work in Ubuntu)

#### **INPUT/OUTPUT:**

##### **Input:**

Dataset with the attributes or features may or may not be trained. ( .ARFF, .CSV, C4.5 and binary)

##### **Output:**

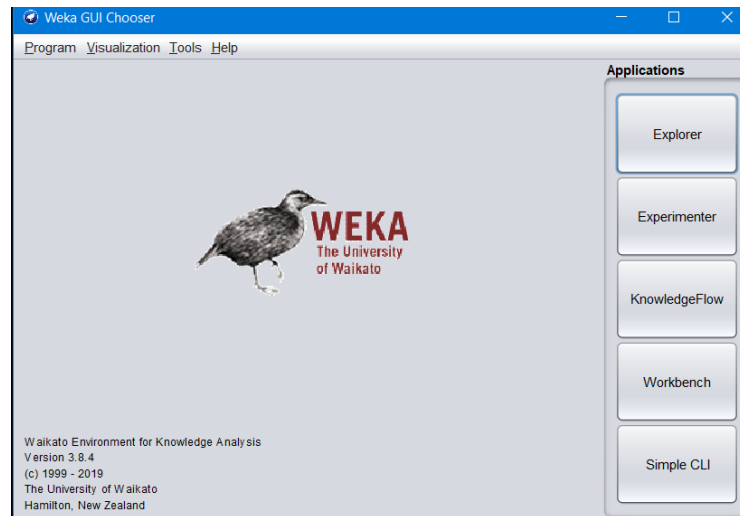
Graphical Representation of analyzed data from Dataset in case of R.

We get the accuracy of successful classification in weka.

#### **OPERATIONAL STEPS REQUIRED:**

- **Steps to operate WEKA**

Choose "WEKA 3.8" from Programs. The first interface that appears looks like the one below.

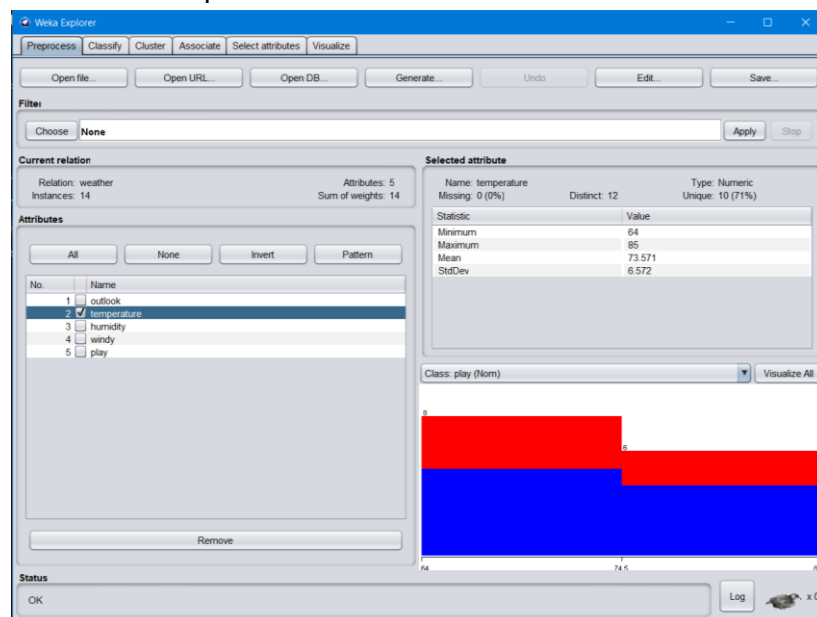


Click on “Explorer”

To load a data set from library eg. ‘weather.arff.’ So file click on “Open File” and browse the path for ‘weather.arff’.

Select the attributes.

Under the Classify tab, click ‘Choose’ and select a classifier from the drop-down menu. E.g.: ‘Decision Stump’



Once, a classifier is chosen, select percentage split and leave it with its default values. The default ratio is 66% for training and 34% for testing.

Click ‘Start’ to train and test the classifier.

- **Steps to operate R**

Open RStudio.

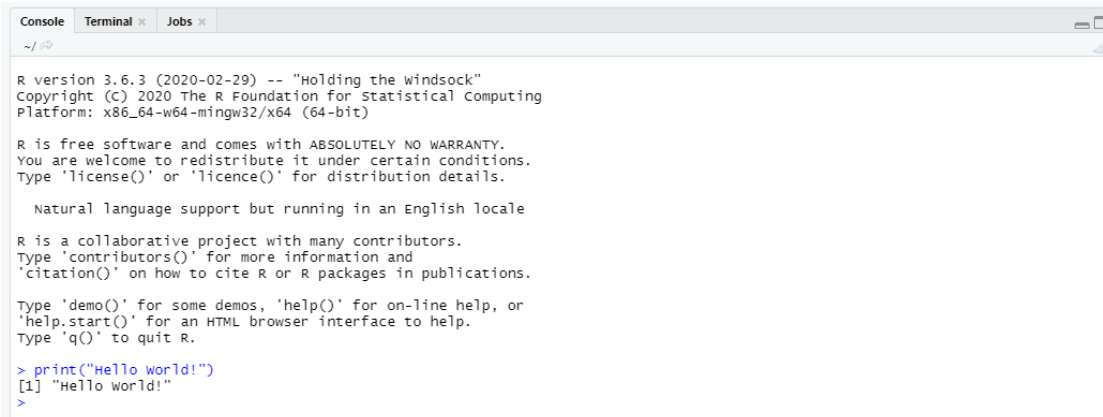
Go to Packages - Click on "Install". You will get "Install Packages" Window.

Install from: Package Archive File(tar.gz)

Package archive: (Browse the path where you have stored the Package\_name.tar.gz)

Click Install.

In Packages - Tick the Package\_name in user library. Click on Update. Click on OK.  
Write Code in R Script (File - New File - R Script) and Save it as "Program1.R"  
Click on run Icon.  
Write R commands on 'console' which is in 'RStudio'.



```
Console Terminal Jobs x
~/
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (c) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> print("Hello world!")
[1] "Hello world!"
>
```

- **Steps to operate Python**

- Python comes with its own shell, called IDLE.
- Go to: File > New. Then save the file with .py extension. For example, hello.py, example.py, etc.  
You can give any name to the file. However, the file name should end with .py
- Write Python code in the file and save it



```
File Edit Shell Debug Options Window Help
Python 3.9.0 (tags/v3.9.0:9cf6752, Oct 5 2020, 15:34:40) [MSC v.1927 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> print("Hello world")
Hello world
>>>
Python 3.9.0
Hello world!
>>> |
```

## LATEST TRENDS:

- **For WEKA**  
Weka 3.8
- **For R**  
R 4.0.2
- **For Python**  
Python 3.9.0

## **APPLICATIONS:**

- **Application of WEKA:**

The WEKA system has been applied successfully in a variety of areas including the areas of agriculture, machine learning research and education.

- **Application of R:**

R applications span the universe from theoretical computational statistics and the hard sciences such as astronomy, chemistry and genomics to practical applications in business, drug development, finance, health care, marketing, medicine and much more. Because R has nearly 5,000 packages (libraries of functions) many of which are dedicated to specific applications you don't have to be an R genius to begin developing your own applications.

- **Application of Python:**

Applications of python can be seen in the fields of Web Development, Game Development, Machine Learning and Artificial Intelligence, Data Science and Data, Visualization, Desktop GUI, Web Scraping Applications, Business Applications, Audio and Video Applications, CAD Applications, Embedded Applications, etc.

## **LIMITATIONS:**

- **Limitations of WEKA**

GUI is not as well documented.

2 different Modules cannot be combined (ex. modules for both PCA and clustering without writing a Java Code).

The Weka GUI provides several built-in 'visualization' panels but these are very limited. Manipulation of data sets is not easy in Weka

- **Limitations of R**

The biggest limitation in R is the data processing model which is to load everything up in memory and process it. This not only limits the amount of data you can process but it also scales very badly for complex processes.

- **Limitations of Python**

Python is slower compared to the other programming languages. If we have a big code, it might sometimes take up-to 40 seconds for python to completely solve it whereas, java or C++ might take up-to 5 seconds. It is always not easy to convert a .py to a .exe files, there would be many restrictions. One main disadvantage of python is it cannot run two versions of python on one desktop. Many modules, libraries are third-party which means, the libraries are not done by the official python team, instead some other unknown organization has done and posted it. There might be many risks.

**CONCLUSION:**

Downloaded the open source software R-base, RStudio and WEKA. Studied the distinct features and functionality of both the software platforms. Found WEKA easier to learn but there are some limitations in case of Graphical Representations, Modifying the dataset etc. R is difficult to learn for novice but its Graphical Representation is better than WEKA. Python is an easy to use scripting language which is the best-fit for implementing machine learning algorithms