Assignment 5: Study of open source analytical software

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 platform. Use R, WEKA and Python

OBJECTIVE :

To study:

- Concept of open source analytical software. (WEKA & R & 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:

WE

Waikato Environment for Knowledge Analysis (Weka), developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License, and the companion software to the book "Data Mining: Practical Machine Learning Tools and Techniques". Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.[1] The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains,[2][3] but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces

Features of WEKA

The features of Weka are:

- **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 used from the command line. This can be very useful for scripting large jobs.
- **Java API**: It is written in Java and provides a 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.
- 49 data preprocessing tools
- 76 classification/regression algorithms
- 8 clustering algorithms
- 15 attribute/subset evaluators + 10 search algorithms for feature selection.
- 3 algorithms for finding association rules
- 3 graphical user interfaces
- “The Explorer” (exploratory data analysis)
- “The Experimenter” (experimental environment)
- “The KnowledgeFlow” (new process model inspired interface)

Installation of WEKA

All versions of Weka can be downloaded from the Weka download webpage.

Select the version of Weka that you would like to install then visit the Weka download page to locate and download your preferred version of Weka.

Your options include:

- Install the all-in-one version of Weka for Windows or Mac OS X.
- Install Java and Weka separately for Windows or Mac OS X.
- Install the standalone version of Weka for Linux and other platforms.

1. All-in-one installation version

Weka provides an all-in-one installation version for Windows and Mac OS X.

This installation includes both the Weka platform that you can use for predictive modeling, as well as the version of Java needed to run the Weka platform.

Windows:

On Windows the all-in-one version of Weka is provided as a self-extracting executable.

You must choose whether you would like the 32-bit version of the package or the 64-bit version of the package. If you have a modern version of Windows, you should select the 64-bit version.

On the Weka download webpage, these packages are called:

- Self-extracting executable for 64-bit Windows that includes Oracle's 64-bit Java.
- Self-extracting executable for 32-bit Windows that includes Oracle's 32-bit Java.

The download is about 100 megabytes. After you have downloaded the package, double click on the icon to start the installation process.

Follow the prompts for the installation and Weka will be added to your Program Menu.

Start Weka by clicking on the bird icon.

Mac OS X:

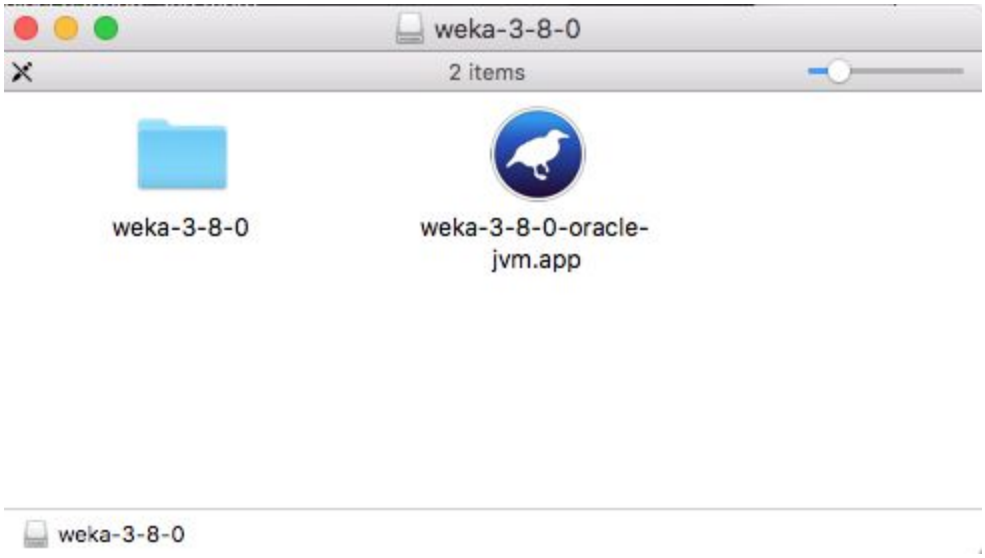
On OS X the all-in-one version of Weka is provided as a disk image.

On the Weka download webpage, this package is called:

- Disk image for OS X that contains a Mac application including Oracle's Java.

The download is about 120 megabytes. The disk image includes two versions of Weka, one with the Java version bundled and one standalone. I recommend installing both.

Drag both the folder and the icon into your Applications folder.



2.Install Java and WEKA separately

You may already have the Java Runtime Environment or Java Development Kit installed on your workstation or you may like to install Java separately from Weka so that you can use Java with other applications.

Weka provides a version that you can download that does not include the Java Runtime Environment.

I recommend this installation of Weka if you would like to access the data files and documentation provided with the Weka installation.

Weka requires at least Java 1.7 installed.

If you do not have Java installed and would like to install Java separately from Weka, you can download Java from the Java Download webpage. The webpage will automatically determine the version of Java you need for your workstation and download the latest version. The Java download is about 60 megabytes.

Windows:

Weka provides a version for Windows that does not include Java.

You must choose whether you would like the 32-bit version of the package or the 64-bit version of the package. If you have a modern version of Windows, you should select the 64-bit version.

On the Weka downloads page, this version is named as follows:

- Self-extracting executable for 64-bit Windows without a Java VM.
- Self-extracting executable for 32-bit Windows without a Java VM.

The download is about 50 megabytes. After you have downloaded the package, double click to start the installation process. Follow the prompts for the installation and Weka will be added to your Program Menu.

Start Weka by clicking on the bird icon.

Mac OS X:

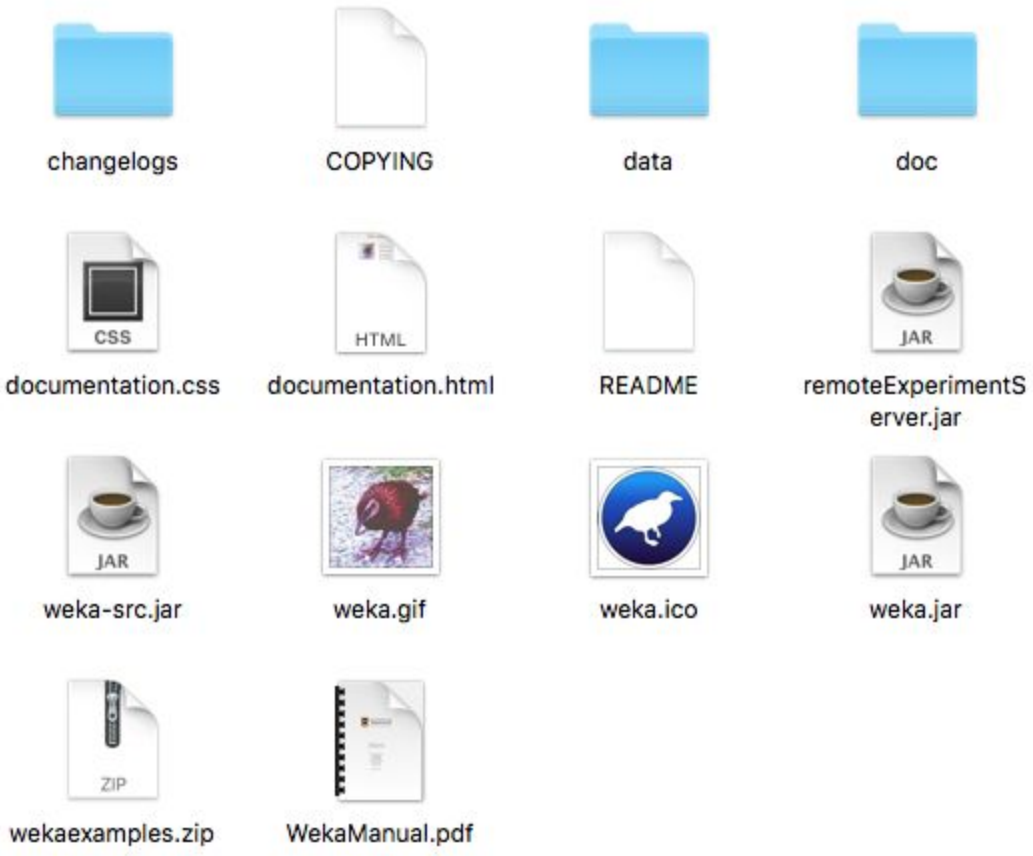
There is only a single download version of Weka for OS X.

It is a disk image that includes both the version of Weka bundled with Java as well as the standalone version.

On the Weka download webpage, this package is called:

- Disk image for OS X that contains a Mac application including Oracle's Java.

The download is about 120 megabytes. Open the disk image and drag the standalone version of Weka (the folder) into your Applications folder.



3.Install on Linux

Weka also provides a standalone version that you can install on Linux and other platforms.

Weka runs on Java and can be used on all platforms that support Java.

It is a zip file and has the following name of the Weka download webpage:

- Zip archive containing Weka.

Download the zip file and unzip it.

You can also start Weka on the command line, assuming Java is in your path.

```
    35147 14 Apr 12:58 COPYING
    16171 14 Apr 12:58 README
  6621937 14 Apr 12:58 WekaManual.pdf
     1938 14 Apr 12:58 changelogs
      918 14 Apr 12:58 data
      578 14 Apr 12:58 doc
      510 14 Apr 12:58 documentation.css
     1863 14 Apr 12:58 documentation.html
    42900 14 Apr 12:58 remoteExperimentServer.jar
 10759024 14 Apr 12:58 weka-src.jar
    30414 14 Apr 12:58 weka.gif
   359270 14 Apr 12:58 weka.ico
 10997325 14 Apr 12:58 weka.jar
 14758799 14 Apr 12:58 wekaexamples.zip
```

1. Change directory into your Weka installation directory. For example

```
cd /Applications/weka-3-8-0
```

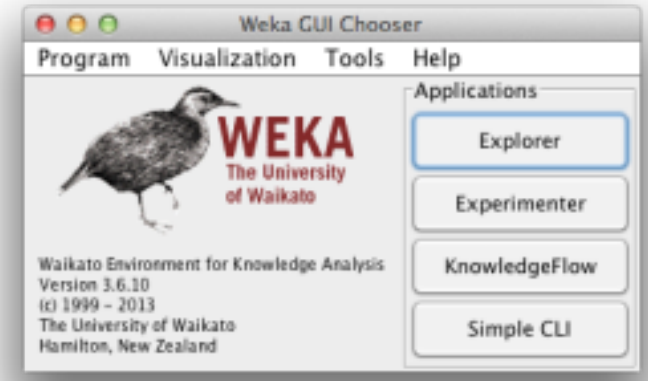
2. Start the Java virtual machine with the *weka*.jar file, For example:

```
java -jar weka.jar
```

Start WEKA

Start Weka. This may involve finding it in the program launcher or double clicking on the weka.jar file. This will start the Weka GUI Chooser.

The Weka GUI Chooser lets you choose one of the Explorer, Experimenter, KnowledgeExplorer and the Simple CLI (command line interface).



R

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, data mining surveys, and studies of scholarly literature databases show substantial increases in popularity; as of September 2020, R ranks 9th in the TIOBE index, a measure of popularity of programming languages.
A GNU package, the official R software environment is written primarily in C, Fortran, and R itself (thus, it is partially self-hosting) and is freely available under the GNU General Public License. Pre-compiled executables are provided for various operating systems. Although R has a command line interface, there are several third-party graphical user interfaces, such as RStudio, an integrated development environment, and Jupyter, a notebook interface.

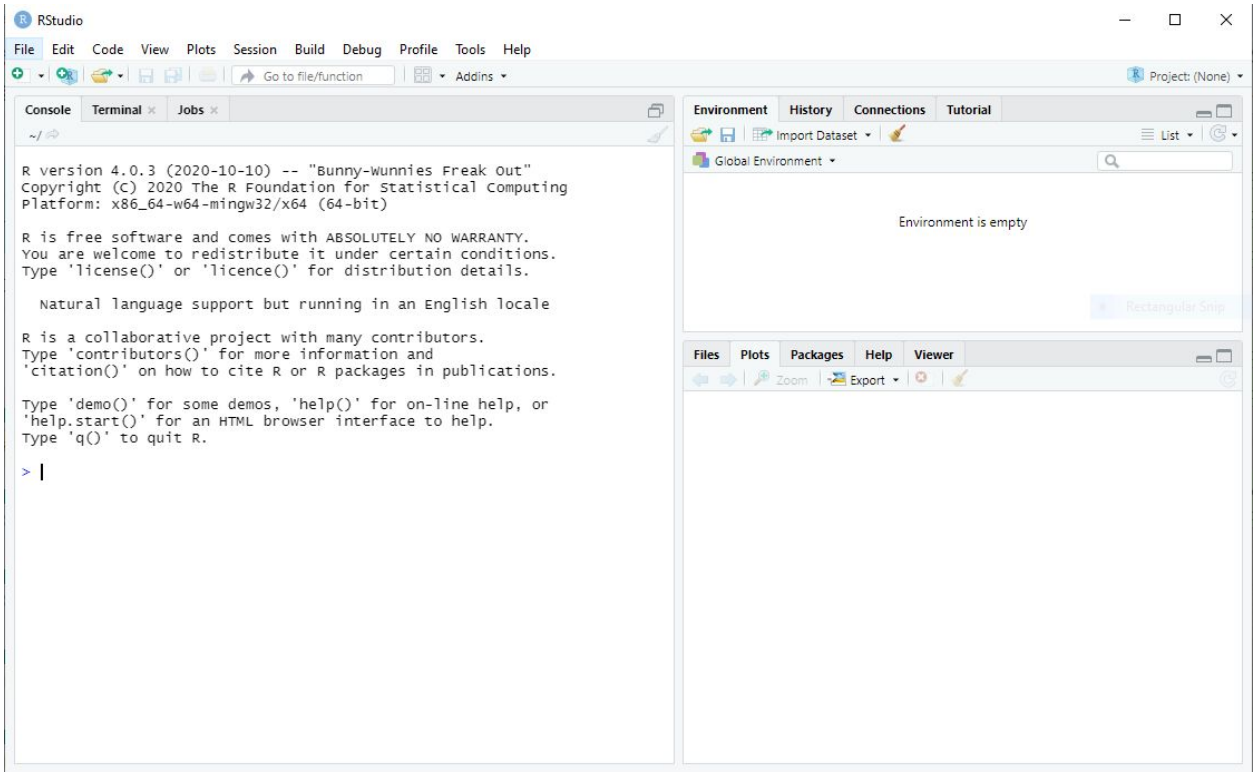
Features of R

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R −

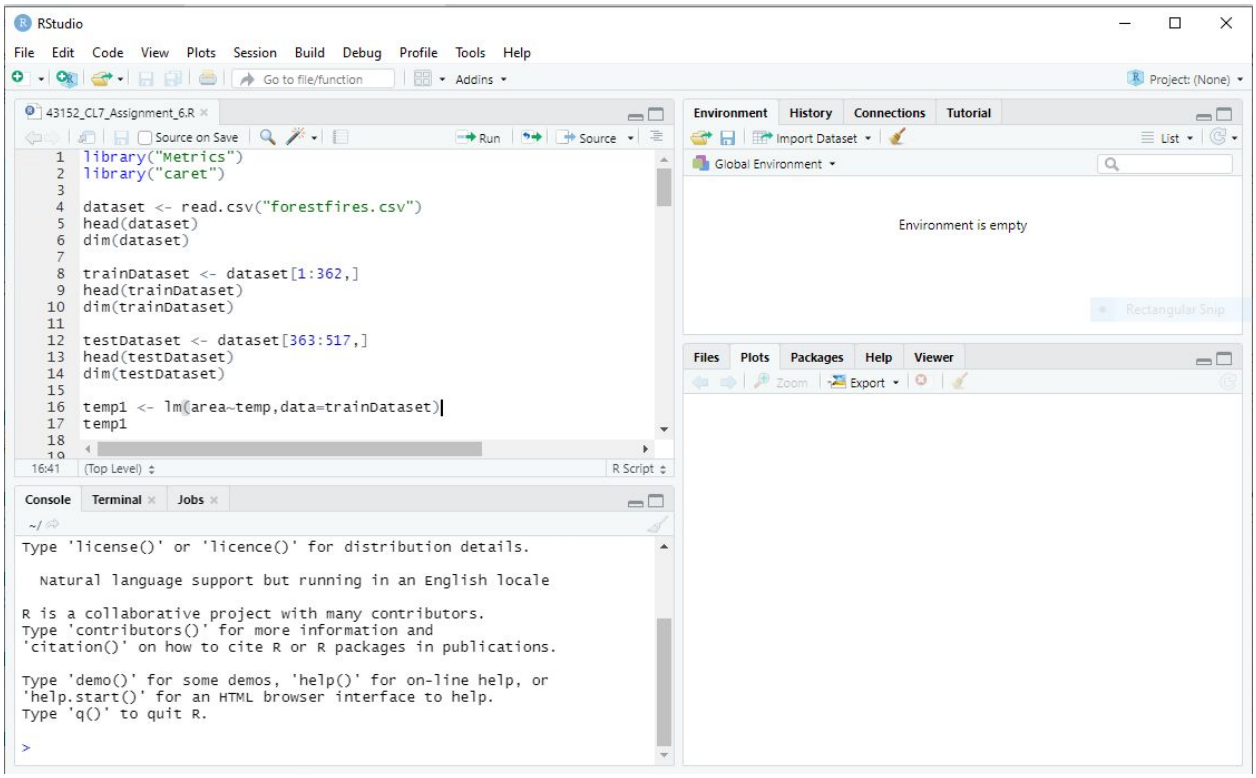
- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.

- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

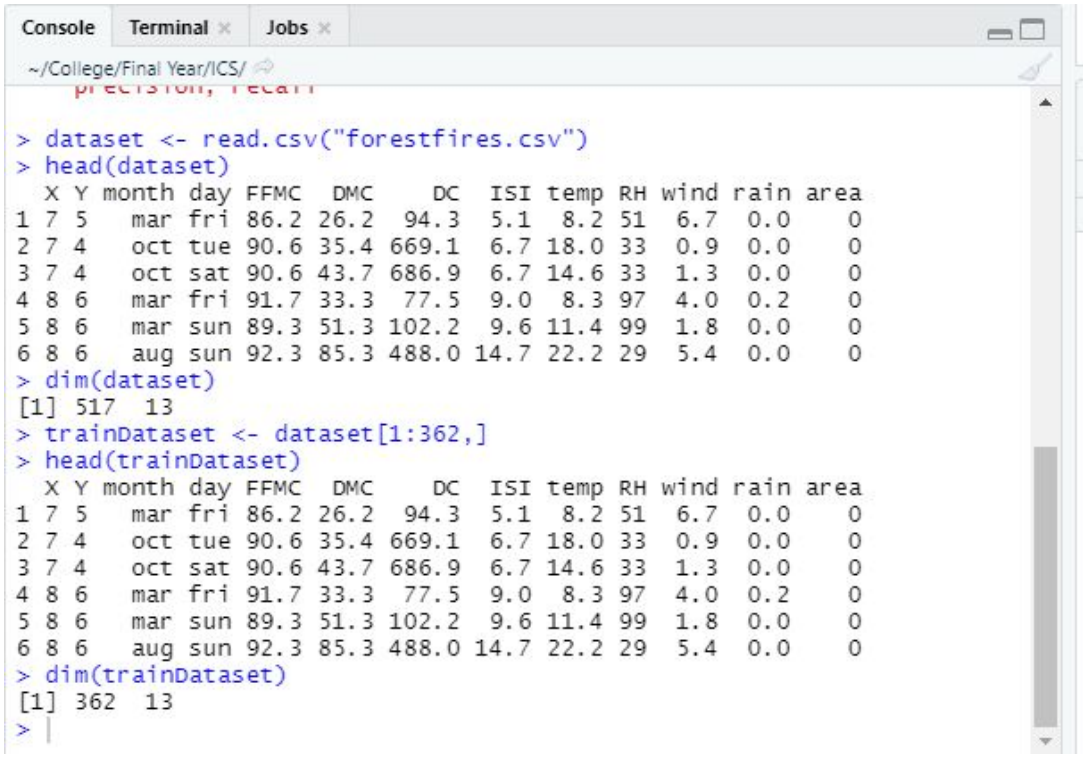
ON Startup, Rstudio looks like the below picture:



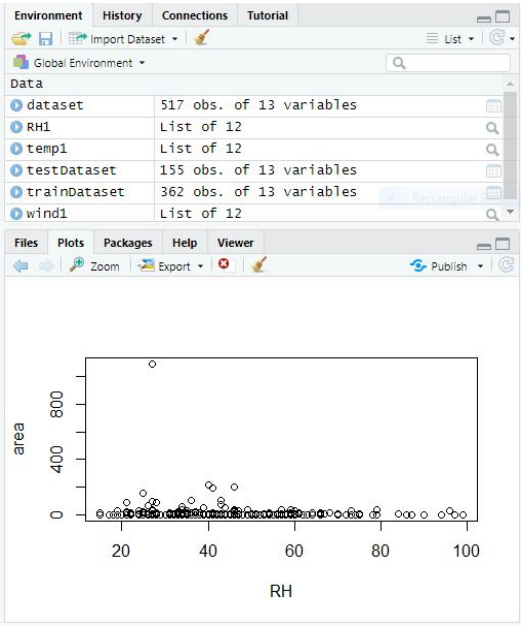
After loading R file in Rstudio the following output can be seen, which shows the RScript in the left hand corner and bottom left displays the console output:



Console output can be seen in the following ways:



Graphs and datasets that have been loaded are seen in the right section of RStudio:



Installing R

1. MAC Users

To Install R

1. Open an internet browser and go to www.r-project.org.
2. Click the "download R" link in the middle of the page under "Getting Started."
3. Select a CRAN location (a mirror site) and click the corresponding link.
4. Click on the "Download R for (Mac) OS X" link at the top of the page.
5. Click on the file containing the latest version of R under "Files."
6. Save the .pkg file, double-click it to open, and follow the installation instructions.
7. Now that R is installed, you need to download and install RStudio.

To Install RStudio

1. Go to www.rstudio.com and click on the "Download RStudio" button.
2. Click on "Download RStudio Desktop."

3. Click on the version recommended for your system, or the latest Mac version, save the .dmg file on your computer, double-click it to open, and then drag and drop it to your applications folder.

To Install the SDSFoundations Package

1. Download SDSFoundations to your desktop (make sure it has the ".tgz" extension).
2. Open RStudio.
3. Click on the Packages tab in the bottom right window.
4. Click "Install."
5. Select install from "Package Archive File."
6. Select the SDSFoundations package file from your desktop.
7. Click install. You are done! You can now delete the SDSpackage file from your desktop.

2. Windows

To Install R:

1. Open an internet browser and go to www.r-project.org.
2. Click the "download R" link in the middle of the page under "Getting Started."
3. Select a CRAN location (a mirror site) and click the corresponding link.
4. Click on the "Download R for Windows" link at the top of the page.
5. Click on the "install R for the first time" link at the top of the page.
6. Click "Download R for Windows" and save the executable file somewhere on your computer. Run the .exe file and follow the installation instructions.
7. Now that R is installed, you need to download and install RStudio.

To Install RStudio

1. Go to www.rstudio.com and click on the "Download RStudio" button.
2. Click on "Download RStudio Desktop."

3. Click on the version recommended for your system, or the latest Windows version, and save the executable file. Run the .exe file and follow the installation instructions.

To Install the SDSFoundations Package

1. Download SDSFoundations to your desktop (make sure it has the ".zip" extension).
2. Open RStudio.
3. Click on the Packages tab in the bottom right window.
4. Click "Install."
5. Select install from "Package Archive File."
6. Select the SDSFoundations package file from your desktop.
7. Click install. You are done! You can now delete the SDSpackage file from your desktop.

Python

Python is an interpreted, high-level and general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library

Python was created in the late 1980s as a successor to the ABC language. Python 2.0, released in 2000, introduced features like list comprehensions and a garbage collection system with reference counting.

Python 3.0, released in 2008, was a major revision of the language that is not completely backward-compatible, and much Python 2 code does not run unmodified on Python 3.

Features of Python

1) Easy to Learn and Use

Python is easy to learn as compared to other programming languages. Its syntax is straightforward and much the same as the English language. There is no use of the semicolon or curly-bracket, the indentation defines the code block. It is the recommended programming language for beginners

2) Expressive Language

Python can perform complex tasks using a few lines of code. A simple example, the hello world program you simply type print("Hello World"). It will take only one line to execute, while Java or C takes multiple lines.

3) Interpreted Language

Python is an interpreted language; it means the Python program is executed one line at a time. The advantage of being interpreted language, it makes debugging easy and portable.

4) Cross-platform Language

Python can run equally on different platforms such as Windows, Linux, UNIX, and Macintosh, etc. So, we can say that Python is a portable language. It enables programmers to develop the software for several competing platforms by writing a program only once.

5) Free and Open Source

Python is freely available for everyone. It is freely available on its official website www.python.org. It has a large community across the world that is dedicatedly working towards make new python modules and functions. Anyone can contribute to the Python community. The open-source means, "Anyone can download its source code without paying any penny."

6) Object-Oriented Language

Python supports object-oriented language and concepts of classes and objects come into existence. It supports inheritance, polymorphism, and encapsulation, etc. The object-oriented procedure helps to programmer to write reusable code and develop applications in less code.

7) Extensible

It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our Python code. It converts the program into byte code, and any platform can use that byte code.

8) Large Standard Library

It provides a vast range of libraries for the various fields such as machine learning, web developer, and also for the scripting. There are various machine learning libraries, such as Tensor flow, Pandas, Numpy, Keras, and Pytorch, etc. Django, flask, pyramids are the popular framework for Python web development.

9) GUI Programming Support

Graphical User Interface is used for the developing Desktop application. PyQT5, Tkinter, Kivy are the libraries which are used for developing the web application.

10) Integrated

It can be easily integrated with languages like C, C++, and JAVA, etc. Python runs code line by line like C,C++ Java. It makes easy to debug the code.

11) Embeddable

The code of the other programming language can use in the Python source code. We can use Python source code in another programming language as well. It can embed other language into our code.

12) Dynamic Memory Allocation

In Python, we don't need to specify the data-type of the variable. When we assign some value to the variable, it automatically allocates the memory to the variable at run time. Suppose we are assigned integer value 15 to x, then we don't need to write int x = 15. Just write x = 15.

Installation of Python

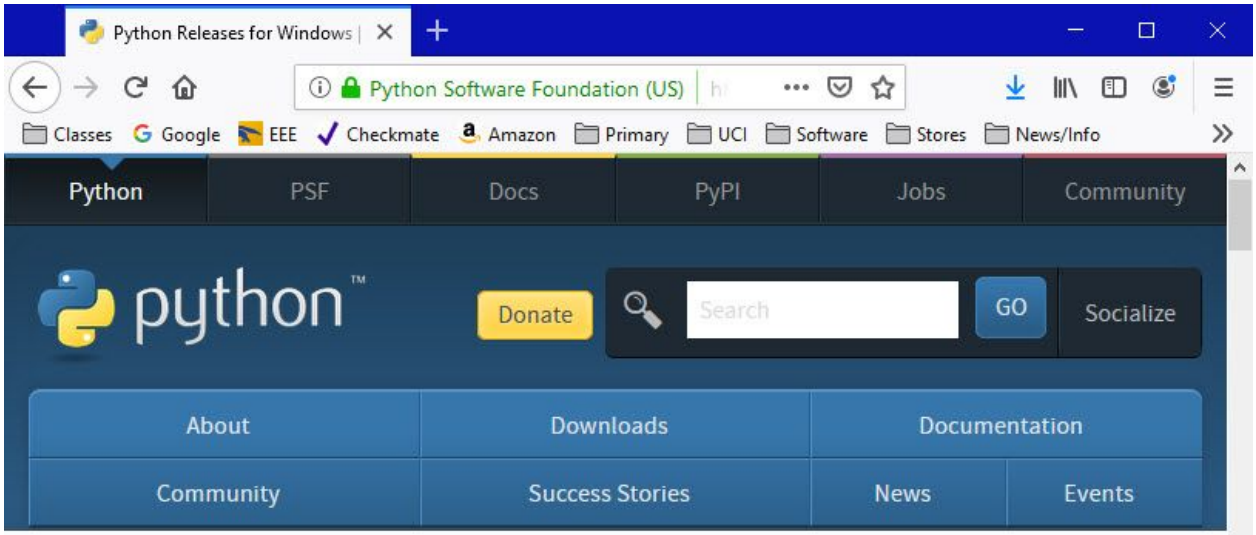
1.Windows

The Python download requires about 25 Mb of disk space; keep it on your machine, in case you need to re-install Python. When installed, Python requires about an additional 90 Mb of disk space.

Downloading

1. Click Python Download.
   The following page will appear in your browser.

Click the Windows link (two lines below the Download Python 3.7.4 button). The following page will appear in your browser.



Python >>> Downloads >>> Windows

Python Releases for Windows

- Latest Python 3 Release - Python 3.7.4
- Latest Python 2 Release - Python 2.7.16

Stable Releases

- Python 3.7.4 - July 8, 2019

  Note that Python 3.7.4 *cannot* be used on Windows XP or earlier.

  - Download Windows help file
  - Download Windows x86-64 embeddable zip file
  - Download Windows x86-64 executable installer
  - Download Windows x86-64 web-based installer
  - Download Windows x86 embeddable zip file
  - Download Windows x86 executable installer
  - Download Windows x86 web-based installer
- Python 3.6.9 - July 2, 2019

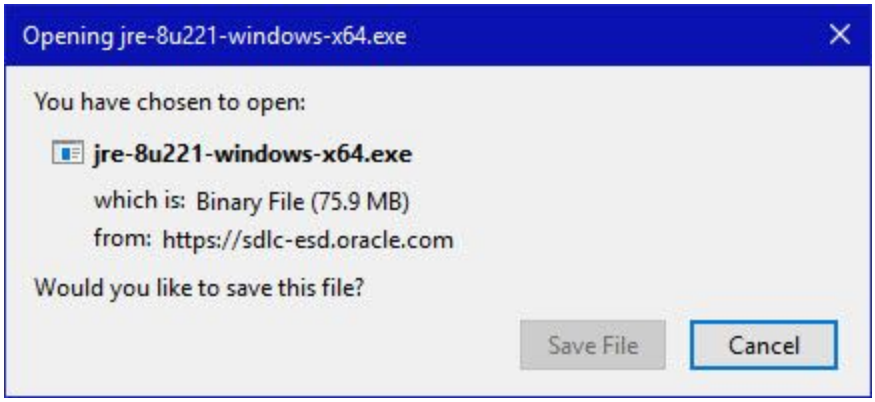
  Note that Python 3.6.9 *cannot* be used on Windows XP or earlier.

Pre-releases

- Python 3.8.0b3 - July 29, 2019
  - Download Windows help file
  - Download Windows x86-64 embeddable zip file
  - Download Windows x86-64 executable installer
  - Download Windows x86-64 web-based installer
  - Download Windows x86 embeddable zip file
  - Download Windows x86 executable installer
  - Download Windows x86 web-based installer
- Python 3.8.0b2 - July 4, 2019
  - Download Windows help file
  - Download Windows x86-64 embeddable zip file
  - Download Windows x86-64 executable installer
  - Download Windows x86-64 web-based installer
  - Download Windows x86 embeddable zip file

Click on the Download Windows x86-64 executable installer link under the top-left Stable Releases.

The following pop-up window titled Opening python-3.74-amd64.exe will appear.

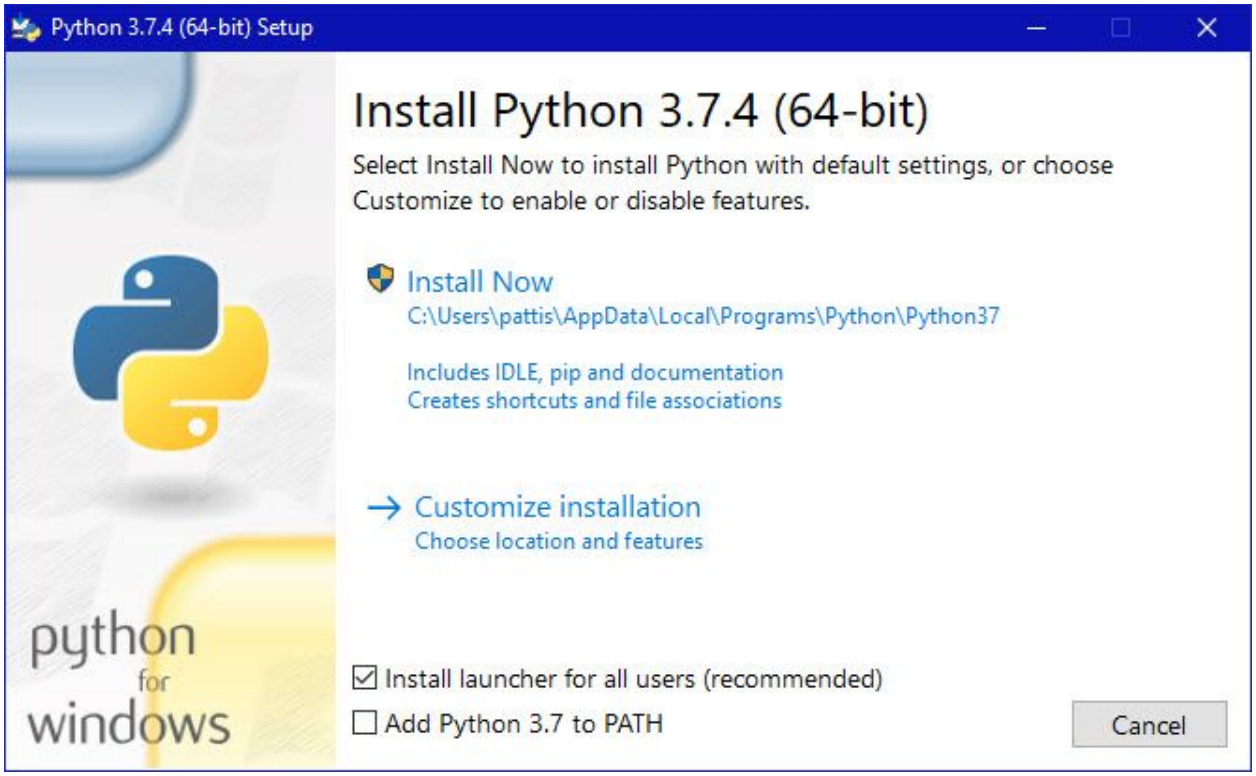


1. Click the Save File button.
   The file named python-3.7.4-amd64.exe should start downloading into your standard download folder. This file is about 30 Mb so it might take a while to download fully if you are on a slow internet connection (it took me about 10 seconds over a cable modem).
   The file should appear as
   

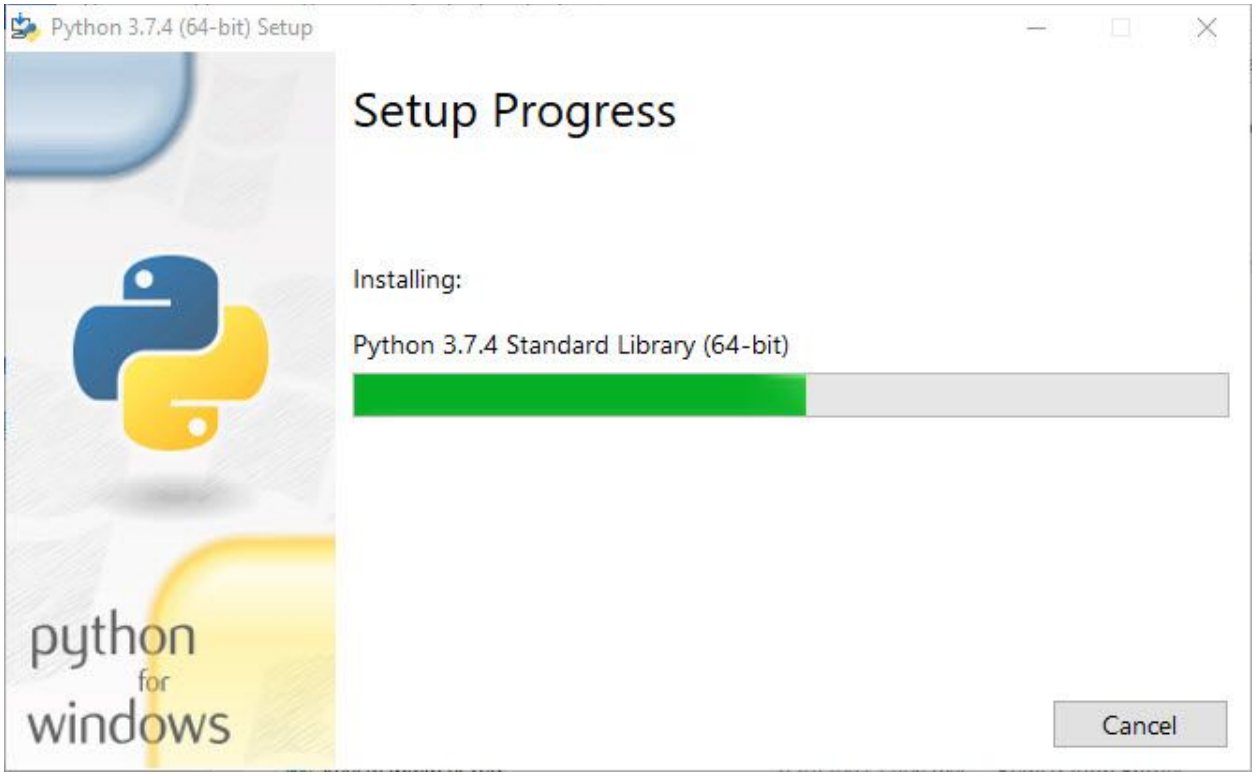
2. Move this file to a more permanent location, so that you can install Python (and reinstall it easily later, if necessary).
3. Feel free to explore this webpage further; if you want to just continue the installation, you can terminate the tab browsing this webpage.
4. Start the Installing instructions directly below.

Double-click the icon labeling the file python-3.7.4-amd64.exe.

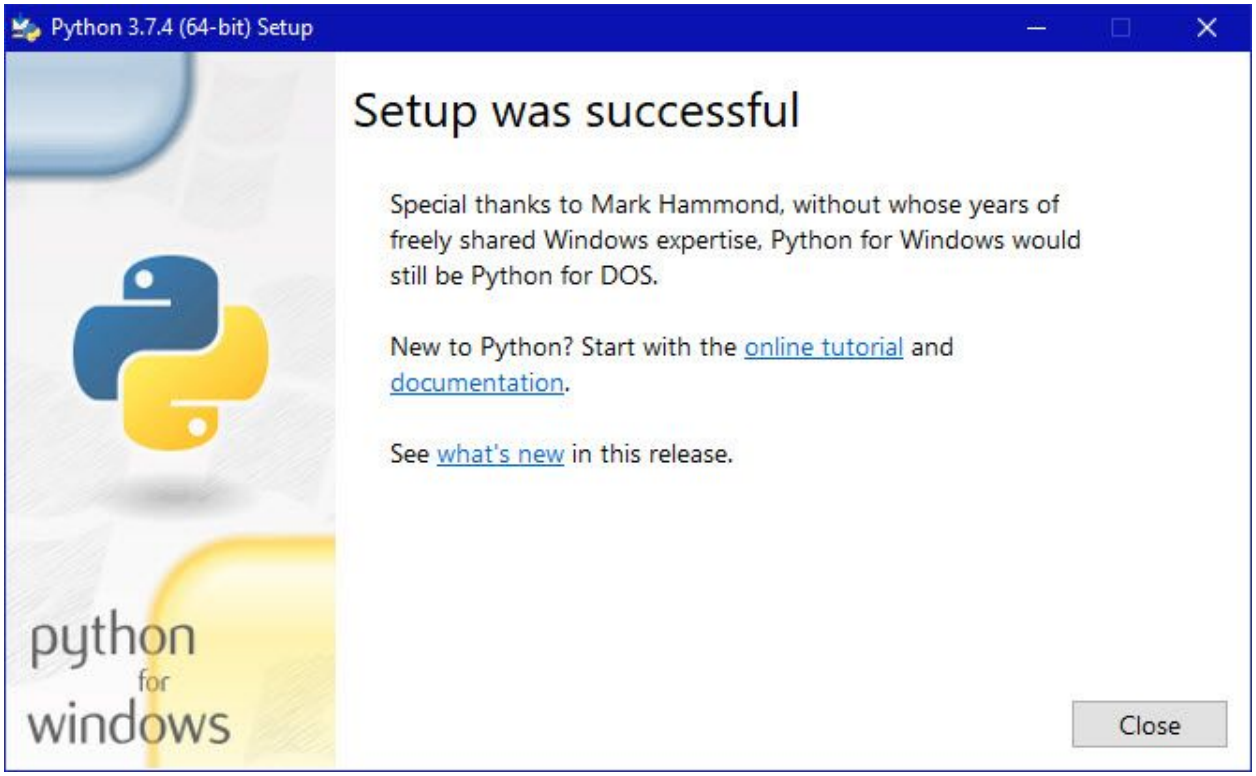
A Python 3.7.4 (64-bit) Setup pop-up window will appear.



1. Ensure that the Install launcher for all users (recommended) and the Add Python 3.7 to PATH checkboxes at the bottom are checked.
   If the Python Installer finds an earlier version of Python installed on your computer, the Install Now message may instead appear as Upgrade Now (and the checkboxes will not appear).
2. Highlight the Install Now (or Upgrade Now) message, and then click it.
   When run, a User Account Control pop-up window may appear on your screen. I could not capture its image, but it asks, Do you want to allow this app to make changes to your device.
3. Click the Yes button.
   A new Python 3.7.4 (64-bit) Setup pop-up window will appear with a Setup Progress message and a progress bar.



During installation, it will show the various components it is installing and move the progress bar towards completion. Soon, a new Python 3.7.4 (64-bit) Setup pop-up window will appear with a Setup was successfully message.



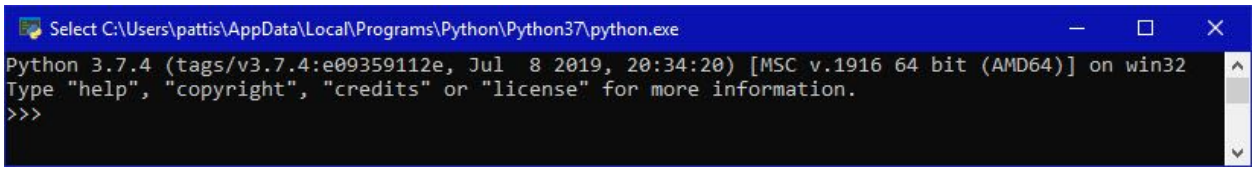
1. Click the Close button.

Python should now be installed.

Verifying

To try to verify installation,

1. Navigate to the directory C:\Users\Pattis\AppData\Local\Programs\Python\Python37 (or to whatever directory Python was installed: see the pop-up window for Installing step 3).
2. Double-click the icon/file python.exe.
   The following pop-up window will appear.



A pop-up window with the title C:\Users\Pattis\AppData\Local\Programs\Python\Python37\python.exe appears, and

inside the window; on the first line is the text Python 3.7.4 ... (notice that it should also say 64 bit). Inside the window, at the bottom left, is the prompt >>>: type exit() to this prompt and press enter to terminate Python.

You should keep the file python-3.7.4.exe somewhere on your computer in case you need to reinstall Python (not likely necessary).

You may now follow the instructions to download and install Java (you should have already installed Java, but if you haven't, it is OK to do so now, so long as you install both Python and Java before you install Eclipse), and then follows the instruction to download and install the Eclipse IDE. Note: you need to download/install Java even if you are using Eclipse only for Python)

2.Linux

This document describes how to install Python 3.6 or 3.8 on Ubuntu Linux machines.

To see which version of Python 3 you have installed, open a command prompt and run

```
$ python3 --version
```

If you are using Ubuntu 16.10 or newer, then you can easily install Python 3.6 with the following commands:

```
$ sudo apt-get update
$ sudo apt-get install python3.6
```

If you're using another version of Ubuntu (e.g. the latest LTS release) or you want to use a more current Python, we recommend using the deadsnakes PPA to install Python 3.8:

```
$ sudo apt-get install software-properties-common
$ sudo add-apt-repository ppa:deadsnakes/ppa
$ sudo apt-get update
$ sudo apt-get install python3.8
```

If you are using other Linux distribution, chances are you already have Python 3 pre-installed as well. If not, use your distribution's package manager. For example on Fedora, you would use dnf:

```
$ sudo dnf install python3
```

Note that if the version of the `python3` package is not recent enough for you, there may be ways of installing more recent versions as well, depending on you distribution. For example installing the `python3.9` package on Fedora 32 to get Python 3.9. If you are a Fedora user, you might want to read about multiple Python versions available in Fedora.

Working with Python 3

At this point, you may have system Python 2.7 available as well.

```
$ python
```

This might launch the Python 2 interpreter.

```
$ python3
```

This will always launch the Python 3 interpreter.

Setuptools & Pip

The two most crucial third-party Python packages are setuptools and pip.

Once installed, you can download, install and uninstall any compliant Python software product with a single command. It also enables you to add this network installation capability to your own Python software with very little work.

Python 2.7.9 and later (on the python2 series), and Python 3.4 and later include pip by default.

To see if pip is installed, open a command prompt and run

```
$ command -v pip
```

To install pip, follow the official pip installation guide - this will automatically install the latest version of setuptools.

Note that on some Linux distributions including Ubuntu and Fedora the `pip` command is meant for Python 2, while the `pip3` command is meant for Python 3.

```
$ command -v pip3
```

However, when using virtual environments (described below), you don't need to care about that.

Pipenv & Virtual Environments

The next step is to install Pipenv, so you can install dependencies and manage virtual environments.

A Virtual Environment is a tool to keep the dependencies required by different projects in separate places, by creating virtual Python environments for them. It solves the "Project X depends on version 1.x but, Project Y needs 4.x" dilemma, and keeps your global site-packages directory clean and manageable.

For example, you can work on a project which requires Django 1.10 while also maintaining a project which requires Django 1.8.

Conclusion:

Downloaded the open source softwares 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 a 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