

WEKA

TABLE OF CONTENTS

Made by BIDUSA Spain teaching team

BIDUSA



Big Data
Unites
Sciences
and Arts



DAVID JESÚS F.DEZ OLMOS,
ARIADNA GARCÍA BONILLA AND
DANIEL MARTÍNEZ CAMPOS

INDEX

1

WHAT'S WEKA?

- Data mining and dataset
- Algorithms
- How will we do?

2

INSTALLATION

1

What's WEKA?

Weka is a software for data mining and machine learning. It includes multiple tools used in data research and analysis, allowing preprocessing, classification (applying different algorithms), and visualization of data.

It enables the analysis of massive datasets and the construction of predictive models by examining various relationships between parameters such as age, gender, location, preferences, consumption, etc.

WHAT IS DATA MINING?

Data mining is a science that studies data with the aim of discovering patterns, relationships, and useful or meaningful knowledge from large datasets. It utilizes techniques and advanced analysis algorithms to explore and analyze data, with the goal of extracting valuable information that can be used for decision-making, trend prediction, or identification of hidden patterns.

This data is grouped into repositories called datasets or data warehouses.

WHAT IS A DATASET?

A dataset is a collection of data that represents information about a specific topic or is gathered for a particular purpose. These datasets can take various forms and sizes, containing information of different types.

In many cases, a dataset is considered immutable if direct changes cannot be made to the original data once the dataset has been created. This means that, after creation, records or attributes within the dataset cannot be added, deleted, or modified.

WHEN ARE DATASET CREATED?

They mostly are created when we no longer use data from a transactional or live database, which are common in companies. For example, when the accounting for a year is closed, and that data can no longer be changed.

The creation of a dataset can occur at any time and depends on the specific need or purpose for which the data is collected. For example:

- **Scientific Research:**

Researchers can create datasets when conducting experiments or studies in fields such as biology, medicine, psychology, etc. The data collected during these research efforts is structured into datasets for analysis.

- **Companies and Organizations:**

Businesses gather data on customers, sales, operations, among other aspects, to make informed decisions.

- **Market Research:**

Market research companies can create datasets by gathering information on consumer preferences, market trends, and purchasing behaviors.

- **Sensors and Devices:**

Internet of Things (IoT) devices and sensors generate large amounts of data. This data can be organized into datasets for analysis and application in areas such as environmental monitoring, health, logistics, etc.

- **Machine Learning Experiments:**

When training machine learning models, datasets can be created containing examples and labels to teach the model to perform specific tasks.


- **Web Data Collection:**

Web data mining can involve creating datasets by extracting information from websites for further analysis.



HOW ARE DATASETS CREATED?

There are two ways to create a dataset:



1. Searching Web pages for already made datasets. For example:

- On the web:


<https://www.kaggle.com/datasets> (Real datasets search engine)

- On the web of each government:

<https://datos.gob.es/es/catalogo> (Government of Spain)

- On the EU website:

<https://ec.europa.eu/eurostat/data/database> (EU-wide database).



2. Creating our own dataset.

For that we need to use an ETL (Extract, Transform and Load) tool.

It can be a simple spreadsheet like **Excel** or **Calc**.

Professionals use tools like Talend, AWS Glue, Pentaho, etc.

WHAT IS AN ALGORITHM?

An algorithm is a set of rules or instructions followed to perform a specific task. In the field of data science, algorithms play a central role in tasks such as:

- **Data analysis**, which is the process of exploring, visualizing, and understanding datasets to extract useful information.
- **Classification**, to assign a label or category to an object or instance based on certain features or attributes.
- **Regression**, to predict or estimate the value of a variable based on the value of one or more predictor variables.
- **Pattern extraction**, a crucial process for uncovering hidden information and gaining a better understanding of the nature of the data.

TYPES OF ALGORITHMS IN DATA MINING

- Association Algorithms

These algorithms discover interesting patterns in datasets, especially those related to the frequency of occurrence of elements in transactions or events.

For example, in problems like market basket analysis, where the goal is to identify associations or purchase patterns among different products.

- Clustering Algorithms

These algorithms group data into homogeneous sets. They are useful for identifying underlying patterns, groups, or segments in unlabeled data.

For example, a marketing analyst may want to segment a store's customers into different groups to personalize marketing strategies. In this scenario, we could use a clustering algorithm to group customers into homogeneous segments based on their buying behaviors.

- Classification Algorithms

These algorithms build predictive models from datasets.

For example, an economic analyst wants to predict whether a specific company will succeed or fail based on certain financial and performance metrics. In this context, we could use a classification algorithm to construct a model that categorizes companies into two classes: "successful" or "failed."

HOW IS IT DONE THEN?

1. The theme of the experiment is decided:

This step involves identifying the theme or objective of the experiment. It is crucial to clearly define what is being investigated or evaluated before initiating any data analysis.

2. The dataset is generated or chosen:

Here, the decision is made on how the data for the experiment will be obtained. It may involve generating data through experiments or choosing an existing dataset that is relevant to the study's theme.

3. The type of algorithm is chosen:

The type of algorithm to be used for data analysis is selected. This depends on the type of task being performed, such as classification, regression, clustering, or association.

4. The software that contains the algorithm is chosen:

The software or tool containing the implementation of the selected algorithm is chosen. In this case, it will be Weka.

5. The computer executes the algorithm:

In this step, the algorithm is executed on the computer using the dataset. The computer performs the necessary calculations and processing to apply the algorithm to the data.

6. The computer draws the conclusions:

After executing the algorithm, the computer generates results. These results may include predictions, identified patterns, statistical summaries, among others, depending on the task and algorithm used.

7. The analyst analyzes the conclusions:

An analyst reviews the results generated by the program. This involves a deeper interpretation of the findings, assessing the validity of the conclusions, and considering the context of the experiment.

8. A report is generated with the conclusions:

Finally, a report is created documenting the results of the experiment. This report may include details about the experiment's design, the data used, the choice of algorithm, the results obtained, and the analyst's interpretations. This report can be shared with other stakeholders or used for decision-making based on the results.

WHAT PROGRAM ARE WE GOING TO USE?

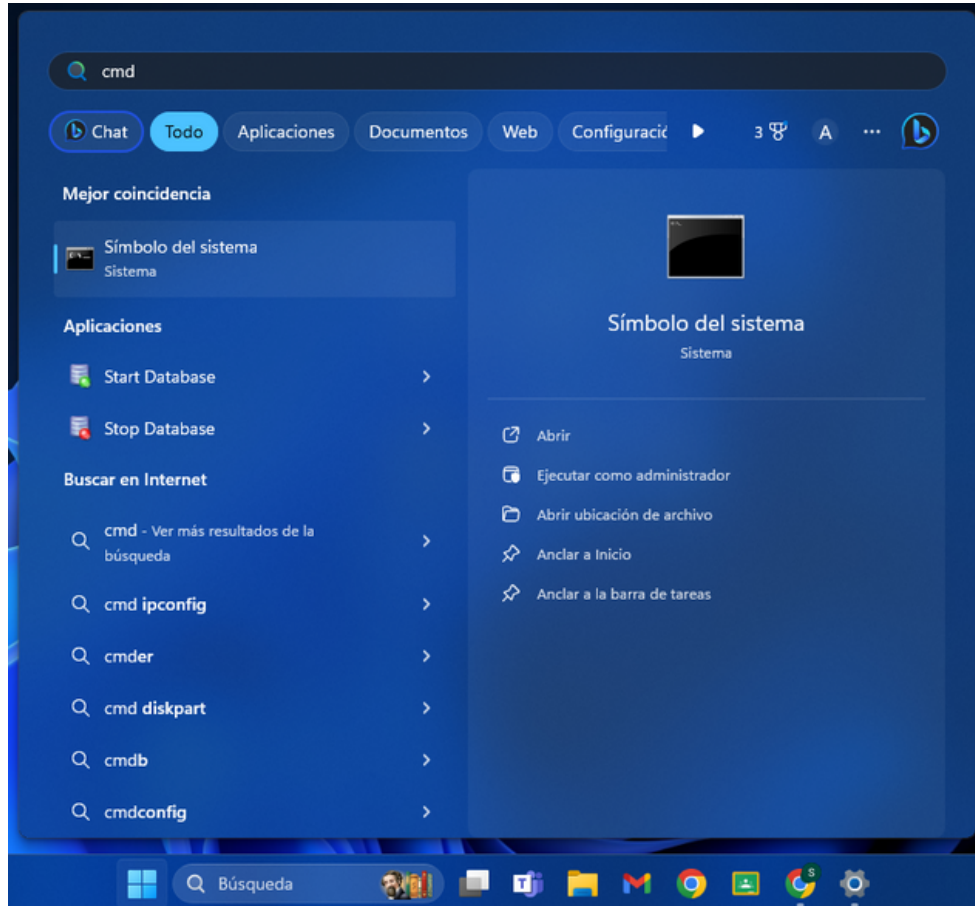
In this project, we will use Weka, a free and open-source software developed by the Department of Computer Science at the University of Waikato in New Zealand. It is programmed in Java and is known for its cross-platform compatibility, making it suitable for operating systems such as Windows, Linux, and Mac.

Weka's user-friendly interface has contributed to its adoption in educational environments, where it is used to teach fundamental concepts of data mining and machine learning.

2

Installation

1. In the Windows search bar, we will enter "cmd" to access the command prompt.



2. Inside the terminal, we will use the command "java --version" to check if the Java Runtime Environment (JRE) is installed.

```
Símbolo del sistema
Microsoft Windows [Versión 10.0.22631.3007]
(c) Microsoft Corporation. Todos los derechos reservados.

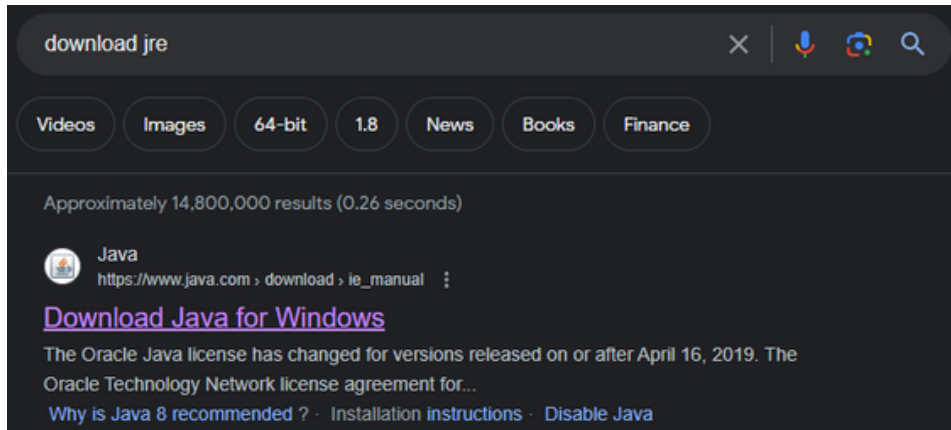
C:\Users\HP>java --version
java 21 2023-09-19 LTS
Java(TM) SE Runtime Environment (build 21+35-LTS-2513)
Java HotSpot(TM) 64-Bit Server VM (build 21+35-LTS-2513, mixed mode, sharing)
```

If the JRE is installed, it will display its version.

If this screen does not appear, it means that it is not installed, so Weka could not be executed. Therefore, we will need to download it by following these steps:

1. Search on Google for JRE.

https://www.java.com/es/download/ie_manual.jsp



2. Download it.

Download Java for Windows

Version 8 Update 401 (filesize: 64.43 MB) Why is Java 8 recommended ?

Publication date: January 16, 2024

Important information about Oracle Java licensing

The Oracle Java license has changed for versions released on or after April 16, 2019.

The [Oracle Technology Network license agreement for Oracle Java SE](#) is substantially different from previous Oracle Java licenses. This license permits certain uses, such as personal and development use, free of charge (although there may be other uses authorized in earlier Oracle Java licenses that are no longer available). Please review the conditions carefully before downloading and using this product. You can check out the [FAQ here](#).

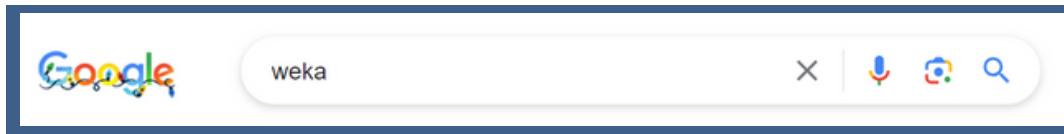
Commercial license and support is available with a low-cost [Java SE subscription](#).

[Download Java](#)

By downloading Java, you confirm that you have read and agree to the terms of the [Oracle Technology Network License Agreement for Oracle Java SE](#)

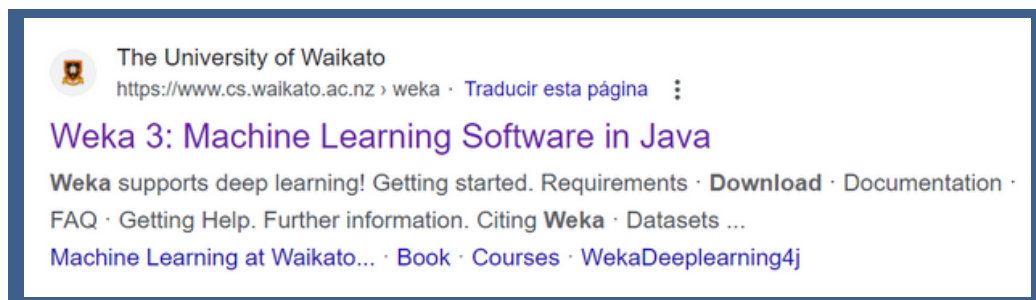
Once the JRE is installed, we can download Weka by following these steps:

1. Search on Google for "Weka".



2. Go to The University of Waikato's website.

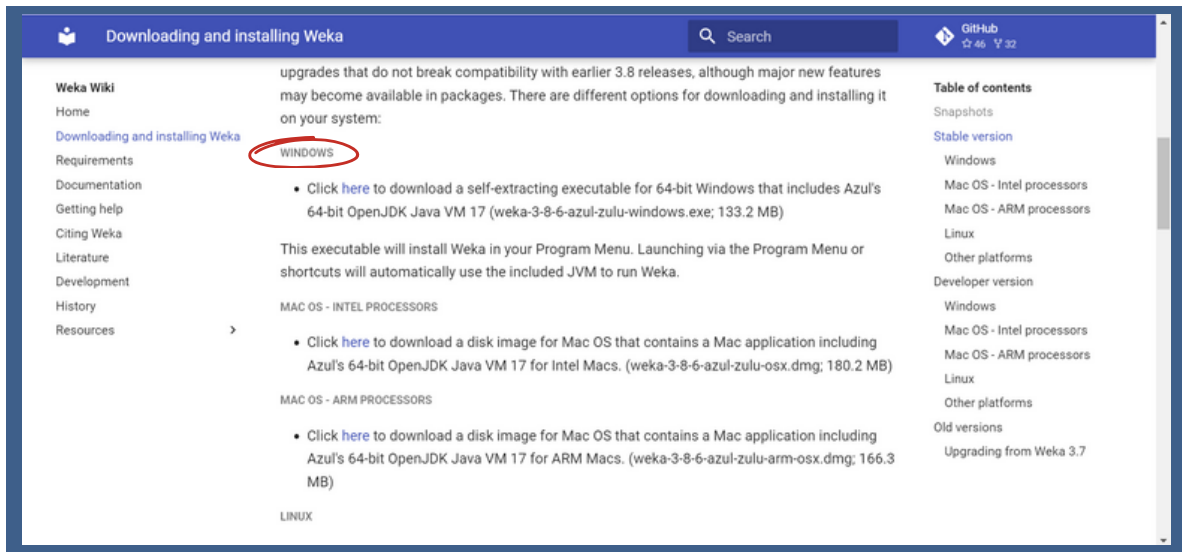
<https://www.cs.waikato.ac.nz/ml/weka>



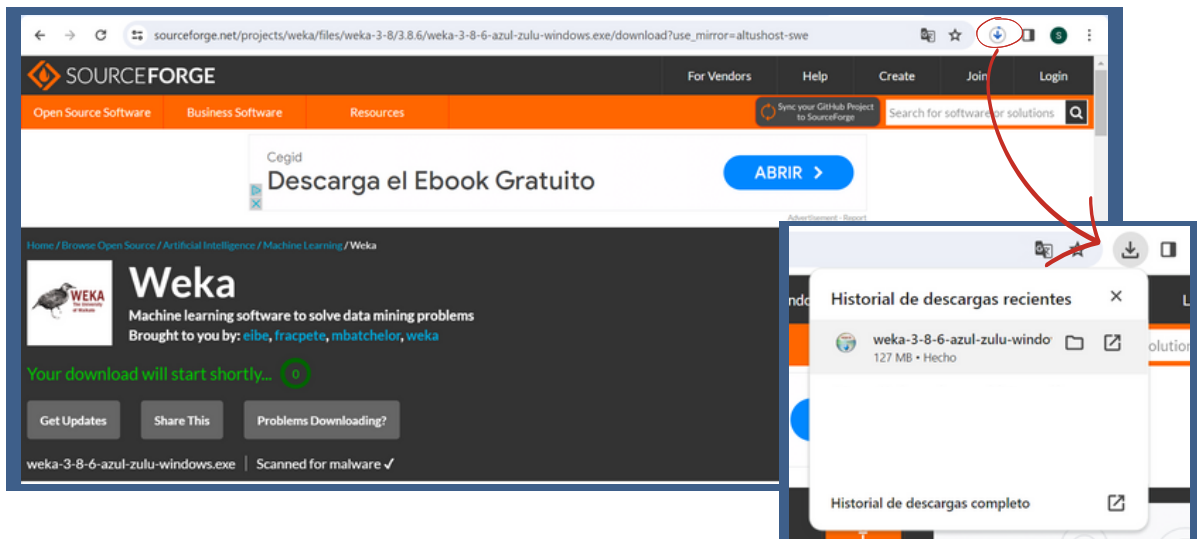
3. Once on the page, you'll see this menu, but if not, you will need to select it clicking in "Software".
Then, choose the second option ("Download") at the bottom of the page.

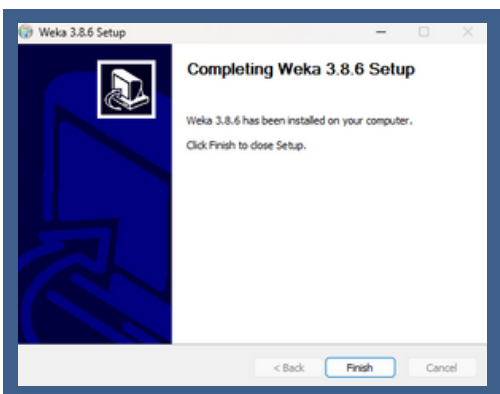
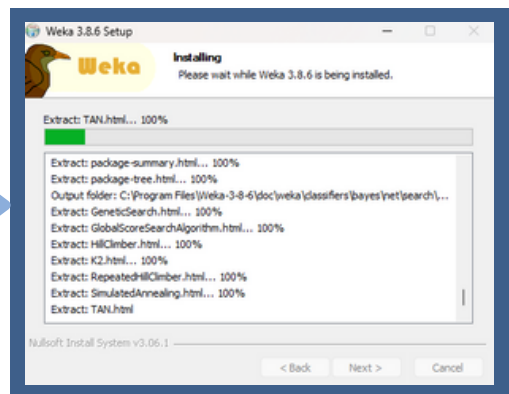
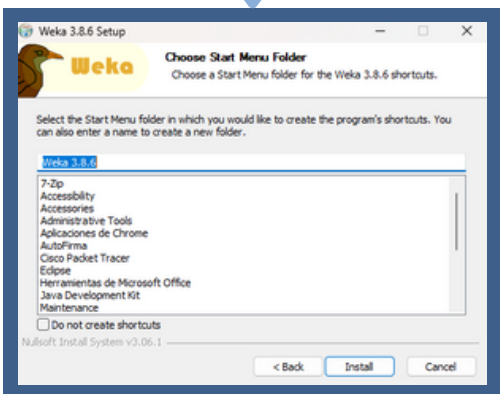
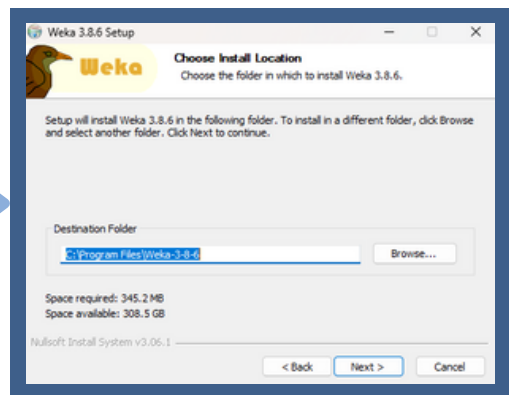
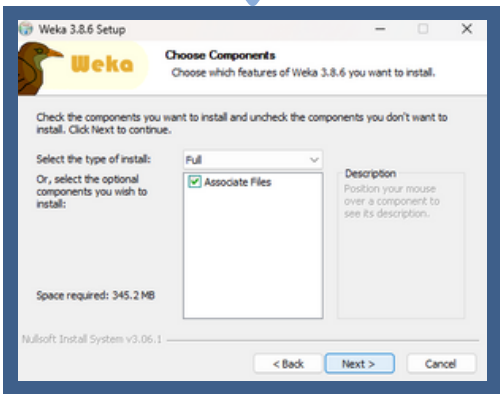
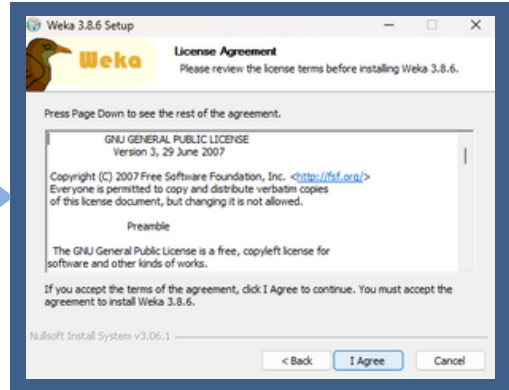
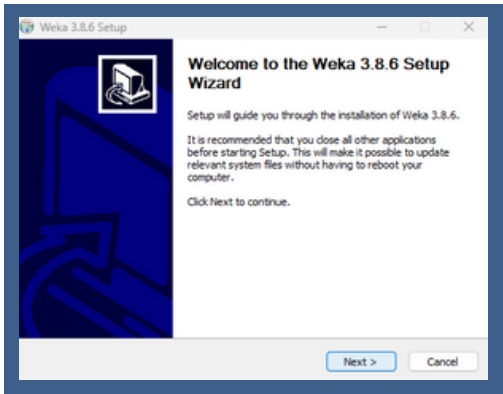


4. Select your operating system and click on “here” to start download.



5. The file will begin downloading in 5 seconds. Go to the downloads section and open the file.





The installation process has been completed.