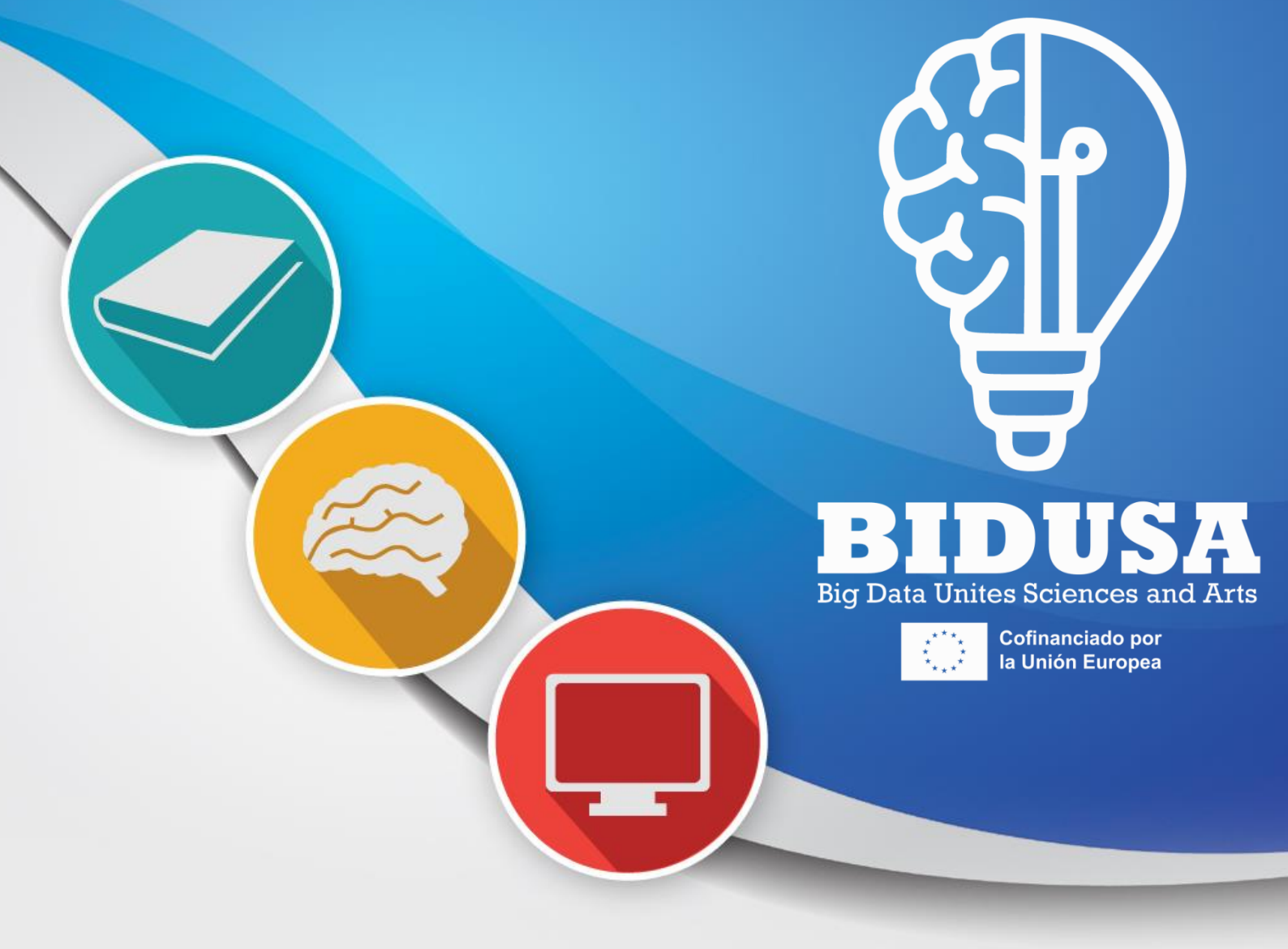


BIDUSA

Big Data Unites Sciences and Arts

Cofinanciado por
la Unión Europea

Data mining concepts

Session I

In this session, introductory concepts to data mining and its use will be explained through examples.

Erasmus+
Enriqueciendo vidas, abriendo mentes.

Part A – Data mining

A.- What is data mining?

It is the science that studies data with the aim of finding trends, patterns or simply statistics. The data is grouped into repositories called **datasets or datawarehouse**.

B.- What is dataset?

A dataset is a fixed or immutable database. We normally call them dead databases or information systems, since their usefulness is to inform about their data.

C.- When are dataset created?

They 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. Then they become datasets or informational or dead.

D.- How are dataset created?

Two ways:

- ✓ 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) or on the EU website: https://ec.europa.eu/eurostat/data/database (EU-wide database).

- ✓ 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.

CURIOSITY: What color cars do pigeons shit on the most?*

Well, it turns out that in red cars, 18%, and in less green cars, 1%.

** According to a study by the Halfords Group company. https://ornithology.com/red-cars-and-bird-poop/*

Part B – Algorithms

A.- What is an algorithms?

An algorithm is a step-by-step procedure that must always be done the same. For example, medical procedures with patients in the emergency room. In our case, they will be mathematical procedures to follow.

B.- What types of algorithms exist in data mining?

Basically, the following:

Association algorithms: allow you to extract statistical information from the datasets. For example, the pigeon experiment.

Clustering algorithms: allow you to create groups of data in a dataset, based on certain criteria. For example, types of customers of a company by purchasing volume.

Classification algorithms: allow predicting future trends or patterns based on datasets with past data. For example, predictions about the world economy, corporate stock markets, etc.

C.- How is it done then?

Basically, the following:

1.- The theme of the experiment is decided.
2.- The dataset is generated or chosen.
3.- The type of algorithm is chosen.
4.- The software that contains the algorithm is chosen.
5.- The computer executes the algorithm.
6.- The computer draws the conclusions.
7.- The analyst analyzes the conclusions.
8.- A report is generated with the conclusions.

D.- What program are we going to use?

Weka software from the University of Waikato in New Zealand. It is free, open source, compatible with Windows, Linux and Mac, and programmed in Java. It contains all the previous algorithms and its use is very simple.

WEKA download → https://waikato.github.io/weka-wiki/downloading_weka/

Part C – Tasks until June 2024

A.- What dataset will we use?

We will create our own dataset with up to 20 questions that we will ask all of our students and teachers, about 7,000 people. Of them, 12 questions will be from the EU knowledge test at the initial level available on the Internet. Plus up to 8 more that each center will be able to choose from the aspects that interest you most.

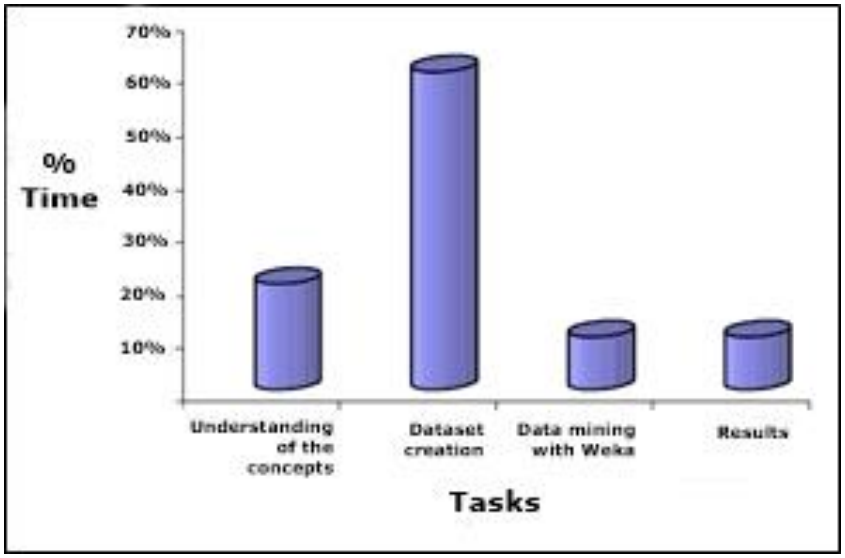
B.- What algorithm will we use?

We will use the A Priori associative algorithm that will allow us to extract statistics by country, by age, by gender, etc.

C.- How will we do it?

1.- We will use an online Google Form to create the questions and record the answers.
2.- We will convert it to a spreadsheet format.
3.- We will process it to make it compatible with Weka.
4.- We will pass the A Priori algorithm to obtain statistics.
5.- We will represent the information with online infographics.

D.- How much time is dedicated to each part?



OWN QUESTIONS: choose your topics

You can start thinking about your own questions: what social networks do you use? How many hours do you spend a day with your cell phone?, etc..