Step 02

Data collection

- Essential considerations
- Tools and
- Resources

Data collection is a crucial step in building a machine learning model. It involves gathering the necessary data that will be used to train the model.
Here are some key considerations must be made when collecting data for our machine learning model.

Data quality : Ensuring that the data collected is accurate, reliable, and relevant to the problem being solved.

Data quantity : Having enough data to train the machine learning model and generate meaningful insights.

Data privacy and security : Ensuring that data collection methods comply with relevant privacy and security regulations and that sensitive information is protected.

Ethical considerations : Ensuring that data collection methods are ethical and do not cause harm or discrimination to any individuals or groups.

Now where can we find this data? Here are some potential sources:

Publicly available datasets :

There are many websites that offer free datasets for public use. Some popular examples are :

- Kaggle click here
- UCI Machine Learning Repository click here
- Data.gov click here
- Google Dataset Search click here

Scraping data :

we can also scrape data from websites using tools like

- BeautifulSoup
- Scrapy
- Selenium, Octaparse, ParseHub, WebHarvy

This methods can be useful if we need specific data that is not available in a pre-existing dataset.

APIs :

Many web services provide APIs that allow users to access their data in a structured format. Some popular APIs,

- Twitter API
- Facebook API
- Google Maps API

Data brokers :

Data brokers are companies that specialize in collecting and selling data.

Crowdsourcing :

Crowdsourcing is a technique that involves outsourcing tasks to a large group of people.

Platforms like Amazon Mechanical Turk and CrowdFlower provide access to a large pool of workers who can collect and label data.

Creating our own dataset :
If you cannot find the data we need from existing sources, we can create our own dataset through some research.

Remember, the quality of our data will impact the accuracy of our model. So it's important to carefully consider our data sources and ensure they are reliable and relevant to our problem.
