The General Data Science Pipeline

It is helpful to understand the Data Science Project lifecycle as it will guide the creation of the project step by step.

There is no „official" version, but it generally looks like this:

1. Defining and Understanding the Problem
2. Data Collection
3. Data Cleaning and Preparation
4. Exploratory Data Analysis
5. Model Building and Deployment

1. Defining and Understanding the Problem:

For this part communication is very important, the problem needs to be clearly defined and then turned into a Data Science problem so clear steps can be defined on how to solve it.

2. Data Collection

In this step it is important to collect the data that is right for the job. It is a good idea to collect more data than you need as data sets can be incomplete and messy.

3. Data Cleaning and Preparation

This is the longest step, it's where you prepare the data you will be using.

Raw data can include missing entries, duplicates, extreme outliers among many other things

Spend as much time as you can on this step as bad data will produce bad models

4.Exploratory Data Analysis

This part is used to summarize the main characteristics of a data set, it will help understand what the collected data can tell us.

In this part you can generate visual representations of your data, which can help answer questions, see patterns or anomalies. This can help with building the model later on.

There are many tools you can use for this part depending on the problem

5. Model Building and Deployment

Models usually fall into two categories, Supervised or Unsupervised

Supervised models have a known outcome, it has labeled input and output data. For example, a model trained on spam emails, would include examples of both spam and non-spam emails, from which it would „learn" what a spam email looks like

Unsupervised models don't have a known outcome, they are used to compare patterns and trends in data. For example if you wanted to know the shopping habits of your customers based on their preferences.

Unsupervised learning can be problematic however as it doesn't have a set output, it has no way of confirming it's accuracy and this can lead to bias

In summary this is the part that, once the model is deployed, should help answer the questions and solve the problem we posed in the beginning

Closing thoughts:

It is important to take each step seriously as neglecting one of them could result on having to do your whole project from scratch. If you can't clearly define the problem and end up collecting incorrect data then it doesn't matter how well you were able to clean and prepare it.