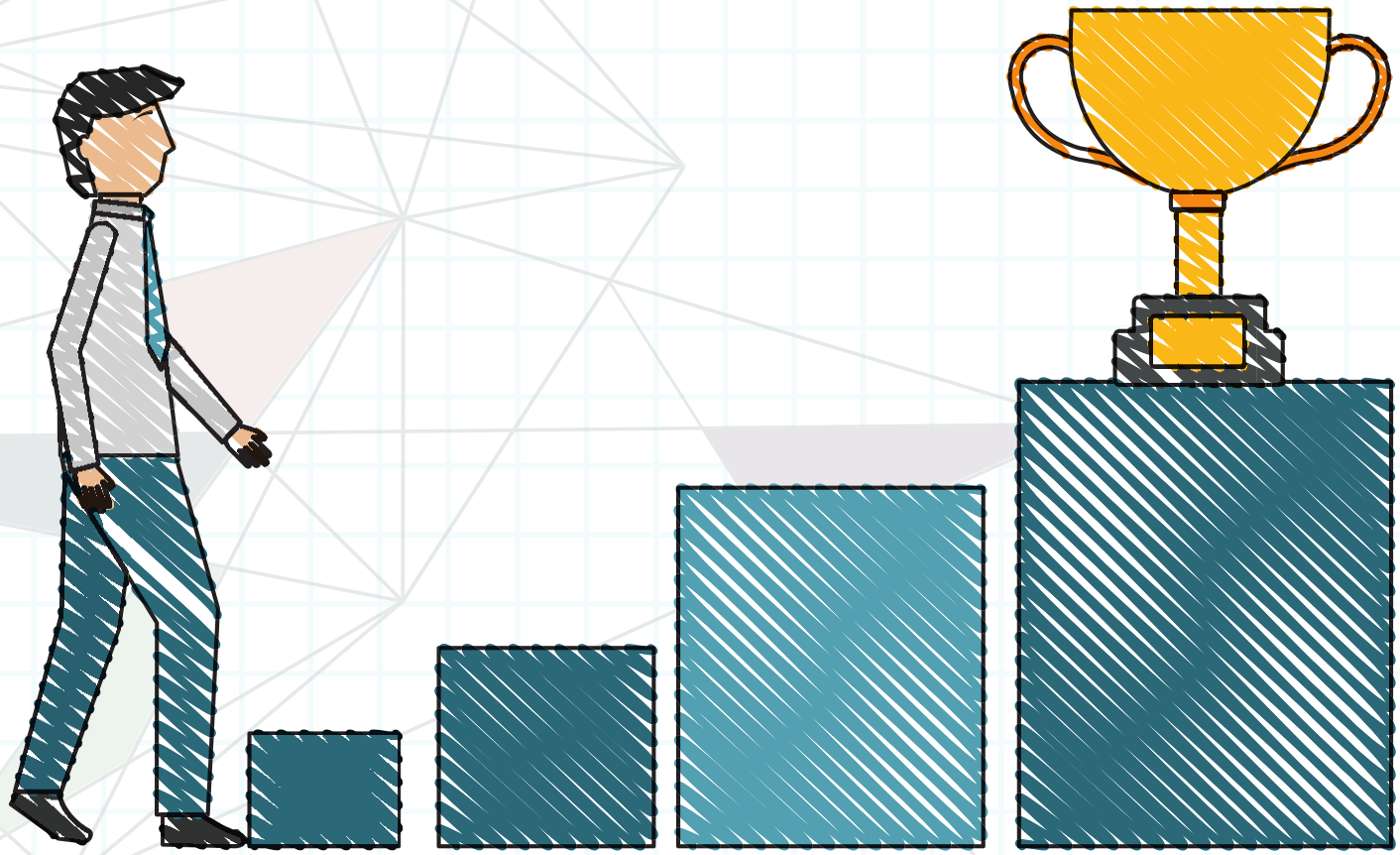
7 STEPS INVOLVED IN A DATA SCIENCE PROJECT



Understanding the business problem

@learn.machinelearning

- You should ask relevant questions which makes you understand the problem which you are going to solve
- You should ask multiple WHY? questions and get answers from the client or the stakeholder or the person who told you to do the project.

Data acquisition

@learn.machinelearning

- After deciding what features or metrics to use to solve the business problem.
- Next step is to gather the data.
- You may use sources like Databases, API's, Web scraper, online repositories etc...

Data preparation

- This step involves 2 important things Data cleaning, Data transformation.
- Data cleaning is like check missing values, inconsistency datatypes, duplicate values etc.. (Check our post on data pre-processing to see what are the most used techniques)
- Data transformation is a process of modifying the data based on predefined rules.

Exploratory data analysis

@learn.machinelearning

- EDA helps you to understand what exactly you can do with the data.
- This is the most important step.
- Through EDA you can find what features are the most important in the model building.
- You can also find useful insights through EDA.

Data modeling

@learn.machinelearning

- This is the most important part where you will be finding the model the best fits the business requirement.
- You will be doing multiple iterations on the test and train data to find the best performing model.

Visualization and communication

@learn.machinelearning

- This is where you will show all the things which you did and fond during the previous steps to your client, stakeholders or the person who gave you the project.
- You will be creating reports or dashboards to show your business finding in a powerful way (visualizations) to make them understand easily.

Deploy & maintenance

- Test your best performing model multiple times before deploying it into production.
- You will be using reports and dashboards for real-time analytics.
- It is also important to monitor the model performance in the real world and retraining it if the performance degrades.