

PRIOR KNOWLEDGE

Team ID	PNT2022TMID11573
Project Name	Efficient water quality analysis and prediction using machine learning

Prior knowledge:

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Data scientists use Python for retrieving, cleaning, visualizing and building models so if we want to use Python to learn data science.

Step 1: Setting up your machine:

The easiest way to proceed is to just download anaconda from website. It comes packaged with most of the things you will need ever. The major downside of taking this route is that you will need to wait for Continuum to update their packages, even when there might be an update available to the underlying libraries.

Step 2: Learn Scientific libraries in Python:

NumPy, SciPy, Matplotlib and Pandas Practice the NUMPY tutorial thoroughly, especially NumPy arrays. This will form a good foundation for things to come. Next, look at the Scipy tutorial. Go through the introduction and the basics and do the remaining one basis your needs.

Step 3: Effective Data Visualization:

It is important to understand what data visualization is and the purpose behind it if we want to build effective visualizations, so without further ado. Data

visualization is the representation of data in a visual format. The purpose of visualizing data is to summarize and present data with easily understandable visualizations that highlight a key message (or messages) from the data, to the readers.

Step 4: Learn Scikit-learn and Machine Learning:

Scikit-learn is the most useful library on python for machine learning. Here is a brief overview of the library. An overview of machine learning, Supervised learning algorithms like regressions, decision trees, ensemble modeling and non-supervised learning algorithms like clustering.

Step 5: Programming:

Some level of programming is required to execute a successful data science project. The most common programming languages are Python, and R. Python is especially popular because it's easy to learn, and it supports multiple libraries for data science and ML.

Step 6: Deep Learning:

Deep learning is an important element of data science, which includes statistics and predictive modeling. It is extremely beneficial to data scientists who are tasked with collecting, analyzing and interpreting large amounts of data; deep learning makes this process faster and easier.