Introduction to Data Science

Notes by Mannan Ul Haq (BDS-3C)

What is Data Science?

Data Science provides special skills to learn from information. It helps us find hidden patterns in big sets of data, so we can make smart choices for future.

Why Do We Need Data Science?

We use Data Science to understand things better. It helps us make sense of the massive amount of information we gather every day. Think of all the numbers, words, and facts we collect from things like websites, apps, and sensors. It tells us what might happen next and helps us solve difficult problems.

Data vs. Information:

Data is like building blocks – numbers, words, raw facts and figures. When we process data and put it together in a smart way, we get information.

Big Data:

Big Data is like a huge amount of information. It's so big that regular tools (hardware devices) can't handle it. We need special tools and tricks to make sense of this massive amount of data.

Hence, Data Science is a field that uses scientific methods, algorithms, processes, and systems to extract valuable knowledge and insights from massive and complex datasets, commonly known as Big Data.

Five Characteristics of Big Data:

- **Volume:** How much data we have collected.
- **Velocity:** How fast data is coming at high speed.
- **Variety:** Different types of data (like numbers, words, pictures, voice-records).
- **Veracity:** How much we can trust the data (accuracy, precision, integrity, reliability).

- **Value:** How useful the data is to make useful decisions.

Data Science Life Cycle:

Here is how a Data Scientist works:

1. **Problem Understanding:** Figuring out what the problem is. It starts with understanding the problem at hand, the questions, and the answers we are trying to find.
2. **Data Acquisition:** Collecting the data required to answer the question or to solve the problem at hand.
3. **Data Wrangling:** Cleaning and getting the data ready. It involves looking for missing values. It uses knowledge to give shape to the dataset appropriate for visualizations.
4. **Data Exploration:** Data Exploration is about visualization and other statistics' measures to see whether the questions we asked, in the beginning, are being answered or not?
5. **Feature Engineering and Selection:** Picking the best parts of the data for your task.
6. **Modeling:** Making a plan to solve the problem. It is about understanding the data's behavior to make the model, which can be used for predictive analytics as described in the previous section.
7. **Deployment:** Sharing your solution with others. It can be deployed on mobile applications and web applications.
8. **Monitoring:** Keeping an eye on things to make sure they're going well. It also involves making changes to the analysis and starting over if required.