

What is Data Science? | Introduction to Data Science | Data Science for Beginners | Simplilearn

Data Science is an increasingly important field, with an ever increasing demand for data scientists. It is used for a variety of tasks, from predictive analysis like predicting delays in airlines or predicting demand for certain products, to creating promotional offers and choosing the most efficient routes for certain journeys. Mohan Mohan discussed the need for data science and definitions, as well as the differences between business intelligence and data science. He also discussed the prerequisites for learning data science. Lastly, he mentioned how data science can be used in politics to create personalized messages tailored to the voters.

The first step in data science is asking the right questions and exploring the data. This helps to identify the problem that needs to be solved and serves as the basis for the modelling process. After modelling, results need to be visualized and communicated to those who need to know them. Business intelligence relies heavily on structured data, while data science involves much more complexity, such as machine learning and the extrapolation of future trends like sales. Data science goes beyond just presenting what has happened in the past and seeks to understand why certain behavior has occurred.

Python is becoming increasingly popular in data science for its ease of use and the variety of libraries it supports for data science, machine learning, and powerful visualization through matplotlib. SAS is a well-established tool, and R provides excellent visualization during development. Spark is an excellent computing engine for distributed data analysis or machine learning. Additionally, there are standard tools such as Informatica Data Stage, Talend, and AWS Redshift that can be used for on-the-cloud operations. Raw data is collected, processed and analyzed before being fed into the analytic system to create output which is then formatted in a way that is useful for stakeholders.

Decision tree is primarily used for classification and can also be used for regression. It is a clustering mechanism which determines which objects belong to which cluster based on their scores. One advantage of decision tree is that it's very easy to understand why a certain object has been classified in a certain way. Data scientists explore the data, looking at its structure and removing any columns that don't add value from an analytical perspective. Data must be cleaned and prepared in order for the system to work properly, although the way of doing this can vary from project to project. If there are too many missing values in few records of large data sets, it's ok to get rid of those entire rows.

Data preparation is an essential step before analyzing or applying data. Model planning follows, and which model to use depends on the problem you're trying to solve. For example, if it is a regression problem, 80% of the training data can be used to train a machine learning model. The training process may have to be iterative, and MATLAB is a popular tool for educational purposes. As an example, data scientists might build a model based on diamond carats in order to predict the price of a 1.35 carat diamond. This would involve passing the information through a linear regression model or creating an appropriate model for the task.

The demand for data scientists is currently huge and the supply is very low, creating a large gap. Gaming and healthcare are two industries that are particularly reliant on data science,