**What is Data Science**

[Data Science](https://www.geeksforgeeks.org/what-is-data-science/) is a field that deals with extracting meaningful information and insights by applying various algorithms preprocessing and scientific methods on structured and unstructured data. This field is related to Artificial Intelligence and is currently one of the most demanded skills. Data science comprises mathematics, computations, statistics, programming, etc to gain meaningful insights from the large amount of data provided in various formats.

## ****What is Data Analytics****

[Data Analytics](https://www.geeksforgeeks.org/data-analytics-and-its-type/) is used to get conclusions by processing the raw data. It is helpful in various businesses as it helps the company to make decisions based on the conclusions from the data. Basically, data analytics helps to convert a Large number of figures in the form of data into Plain English i.e., conclusions which are further helpful in making in-depth decisions. Below is a table of differences between Data Science and Data Analytics:

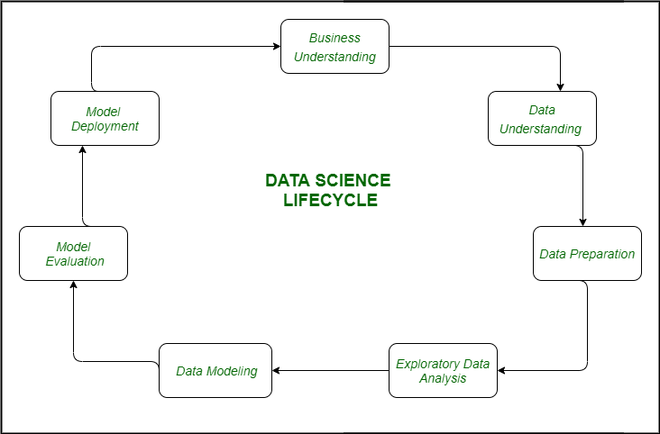
**Difference between Data Analytics and Data Science**

| **Feature** | **Data Science** | **Data Analytics** |
| --- | --- | --- |
| Coding Language | Python is the most commonly used language for data science along with the use of other languages such as C++, Java, Perl, etc. | The Knowledge of Python and R Language is essential for Data Analytics. |
| Programming Skills | In-depth knowledge of programming is required for data science. | Basic Programming skills is necessary for data analytics. |
| Use of Machine Learning | Data Science makes use of machine learning algorithms to get insights. | Data Analytics does not use machine learning to get the insight of data. |
| Other Skills | Data Science makes use of Data mining activities for getting meaningful insights. | Hadoop Based analysis is used for getting conclusions from raw data. |
| Scope | The scope of data science is large. | The Scope of data analysis is micro i.e., small. |
| Goals | Data science deals with explorations and new innovations. | Data Analysis makes use of existing resources. |
| Data Type | Data Science mostly deals with unstructured data. | Data Analytics deals with structured data. |
| Statistical Skills | Statistical skills are necessary in the field of Data Science.. | The statistical skills are of minimal or no use in data analytics. |

### Data Science Process Life Cycle

Some steps are necessary for any of the tasks that are being done in the field of data science to derive any fruitful results from the data at hand.

* [**Data Collection**](https://www.geeksforgeeks.org/sources-of-data-collection/) – After formulating any problem statement the main task is to calculate data that can help us in our analysis and manipulation. Sometimes data is collected by performing some kind of survey and there are times when it is done by performing scrapping.
* [**Data Cleaning**](https://www.geeksforgeeks.org/data-cleansing-introduction/) – Most of the real-world data is not structured and requires cleaning and conversion into structured data before it can be used for any analysis or modeling.
* [**Exploratory Data Analysis**](https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/) – This is the step in which we try to find the hidden patterns in the data at hand. Also, we try to analyze different factors which affect the target variable and the extent to which it does so. How the independent features are related to each other and what can be done to achieve the desired results all these answers can be extracted from this process as well. This also gives us a direction in which we should work to get started with the modeling process.
* [**Model Building**](https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/) – Different types of machine learning algorithms as well as techniques have been developed which can easily identify complex patterns in the data which will be a very tedious task to be done by a human.
* [**Model Deployment**](https://www.geeksforgeeks.org/cloud-deployment-models/) – After a model is developed and gives better results on the holdout or the real-world dataset then we deploy it and monitor its performance. This is the main part where we use our learning from the data to be applied in real-world applications and use cases.



**Components of Data Science Process**

Data Science is a very vast field and to get the best out of the data at hand one has to apply multiple methodologies and use different tools to make sure the integrity of the data remains intact throughout the process keeping data privacy in mind. Machine Learning and Data analysis is the part where we focus on the results which can be extracted from the data at hand. But Data engineering is the part in which the main task is to ensure that the data is managed properly and proper data pipelines are created for smooth data flow. If we try to point out the main components of Data Science then it would be:

* [**Data Analysis**](https://www.geeksforgeeks.org/difference-between-data-analytics-and-data-analysis/) –  There are times when there is no need to apply advanced deep learning and complex methods to the data at hand to derive some patterns from it. Due to this before moving on to the modeling part, we first perform an exploratory data analysis to get a basic idea of the data and patterns which are available in it this gives us a direction to work on if we want to apply some complex analysis methods on our data.
* [**Statistics**](https://www.geeksforgeeks.org/statistics-with-python/) – It is a natural phenomenon that many real-life datasets follow a normal distribution. And when we already know that a particular dataset follows some known distribution then most of its properties can be analyzed at once. Also, descriptive statistics and correlation and covariances between two features of the dataset help us get a better understanding of how one factor is related to the other in our dataset.
* [**Data Engineering**](https://www.geeksforgeeks.org/difference-between-data-science-and-data-engineering/)– When we deal with a large amount of data then we have to make sure that the data is kept safe from any online threats also it is easy to retrieve and make changes in the data as well. To ensure that the data is used efficiently Data Engineers play a crucial role.
* **Advanced Computing**
  + [**Machine Learning**](https://www.geeksforgeeks.org/machine-learning/) – Machine Learning has opened new horizons which had helped us to build different advanced applications and methodologies so, that the machines become more efficient and provide a personalized experience to each individual and perform tasks in a snap of the hand earlier which requires heavy human labor and time intense.
  + [**Deep Learning**](https://www.geeksforgeeks.org/introduction-deep-learning/) – This is also a part of Artificial Intelligence and Machine Learning but it is a bit more advanced than machine learning itself. High computing power and a huge corpus of data have led to the emergence of this field in data science.

## ****Steps for Data Science Processes:****

**Step 1: Defining research goals and creating a project charter**

* Spend time understanding the goals and context of your research.Continue asking questions and devising examples until you grasp the exact business expectations, identify how your project fits in the bigger picture, appreciate how your research is going to change the business, and understand how they’ll use your results.

#### **Create a project charter**

A project charter requires teamwork, and your input covers at least the following:

1. A clear research goal
2. The project mission and context
3. How you’re going to perform your analysis
4. What resources you expect to use
5. Proof that it’s an achievable project, or proof of concepts
6. Deliverables and a measure of success
7. A timeline

**Step 2: Retrieving Data**

Start with data stored within the company

* Finding data even within your own company can sometimes be a challenge.
* This data can be stored in official data repositories such as databases, data marts, data warehouses, and data lakes maintained by a team of IT professionals.
* Getting access to the data may take time and involve company policies.

**Step 3: Cleansing, integrating, and transforming data-**

**Cleaning:**

* Data cleansing is a subprocess of the data science process that focuses on removing errors in your data so your data becomes a true and consistent representation of the processes it originates from.
* The first type is the interpretation error, such as incorrect use of terminologies, like saying that a person’s age is greater than 300 years.
* The second type of error points to inconsistencies between data sources or against your company’s standardized values. An example of this class of errors is putting “Female” in one table and “F” in another when they represent the same thing: that the person is female.

**Integrating:**

* Combining Data from different Data Sources.
* Your data comes from several different places, and in this sub step we focus on integrating these different sources.
* You can perform two operations to combine information from different data sets. The first operation is joining and the second operation is appending or stacking.

**Joining Tables:**

* Joining tables allows you to combine the information of one observation found in one table with the information that you find in another table.

**Appending Tables:**

* Appending or stacking tables is effectively adding observations from one table to another table.

**Transforming Data**

* Certain models require their data to be in a certain shape.

**Reducing the Number of Variables**

* Sometimes you have too many variables and need to reduce the number because they don’t add new information to the model.
* Having too many variables in your model makes the model difficult to handle, and certain techniques don’t perform well when you overload them with too many input variables.
* Dummy variables can only take two values: true(1) or false(0). They’re used to indicate the absence of a categorical effect that may explain the observation.

**Step 4: Exploratory Data Analysis**

* During exploratory data analysis you take a deep dive into the data.
* Information becomes much easier to grasp when shown in a picture, therefore you mainly use graphical techniques to gain an understanding of your data and the interactions between variables.
* Bar Plot, Line Plot, Scatter Plot , Multiple Plots , Pareto Diagram , Link and Brush Diagram ,Histogram , Box and Whisker Plot .

**Step 5: Build the Models**

* Build the models are the next step, with the goal of making better predictions, classifying objects, or gaining an understanding of the system that are required for modeling.

**Step 6: Presenting findings and building applications on top of them –**

* The last stage of the data science process is where your soft skills will be most useful, and yes, they’re extremely important.
* Presenting your results to the stakeholders and industrializing your analysis process for repetitive reuse and integration with other tools

What is Machine Learning?

A subset of artificial intelligence known as machine learning focuses primarily on the creation of algorithms that enable a computer to independently learn from data and previous experiences. Arthur Samuel first used the term "machine learning" in 1959. It could be summarized as follows:

Without being explicitly programmed, machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things.

\*\***A machine can learn if it can gain more data to improve its performance.\*\***

Features of Machine Learning:

* Machine learning uses data to detect various patterns in a given dataset.
* It can learn from past data and improve automatically.
* It is a data-driven technology.
* Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Importance of Machine Learning

* Rapid increment in the production of data
* Solving complex problems, which are difficult for a human
* Decision making in various sector including finance
* Finding hidden patterns and extracting useful information from data.

Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

1. **Supervised learning**
2. **Unsupervised learning**
3. **Reinforcement learning**

### **1) Supervised Learning**

In supervised learning, sample labeled data are provided to the machine learning system for training, and the system then predicts the output based on the training data.

The system uses labeled data to build a model that understands the datasets and learns about each one. After the training and processing are done, we test the model with sample data to see if it can accurately predict the output.

1. Classification
2. Regression

### **2) Unsupervised Learning**

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

* **Clustering**
* **Association**

### **3) Reinforcement Learning**

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

Difference between supervised and unsupervised learning

|  |  |
| --- | --- |
| **Supervised Learning** | **Unsupervised Learning** |
| Supervised learning algorithms are trained using labeled data. | Unsupervised learning algorithms are trained using unlabeled data. |
| Supervised learning model takes direct feedback to check if it is predicting correct output or not. | Unsupervised learning model does not take any feedback. |
| Supervised learning model predicts the output. | Unsupervised learning model finds the hidden patterns in data. |
| In supervised learning, input data is provided to the model along with the output. | In unsupervised learning, only input data is provided to the model. |
| The goal of supervised learning is to train the model so that it can predict the output when it is given new data. | The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset. |
| Supervised learning needs supervision to train the model. | Unsupervised learning does not need any supervision to train the model. |
| Supervised learning can be categorized in **Classification** and **Regression** problems. | Unsupervised Learning can be classified in **Clustering** and **Associations** problems. |
| Supervised learning can be used for those cases where we know the input as well as corresponding outputs. | Unsupervised learning can be used for those cases where we have only input data and no corresponding output data. |
| Supervised learning model produces an accurate result. | Unsupervised learning model may give less accurate result as compared to supervised learning. |
| Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output. | Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences. |
| It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc. | It includes various algorithms such as Clustering, KNN, and Apriori algorithm. |