edureka!

MASTERING DATA SCIENCE WITH EDUREKA

Getting Started with Data Science

Don't just Learn it, Master it!

TABLE OF CONTENTS

TABLE OF CONTENTS

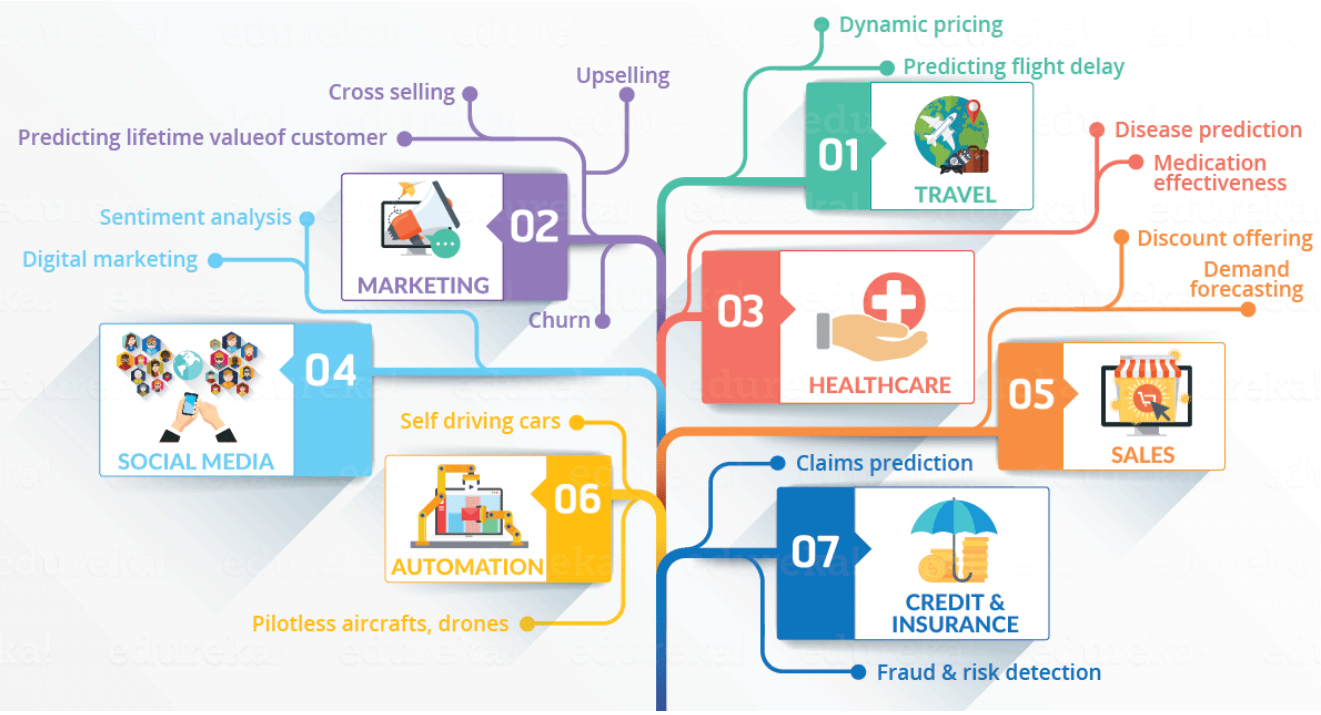
Chapter 1

INTRODUCTION TO DATA SCIENCE

The term Data Science has emerged recently with the evolution of mathematical statistics and data analysis. Data Science also known as data-driven science, makes use of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, i.e either structured or unstructured.

With the help of Data Science in the next few years, we will be able to predict the future as claimed by researchers from MIT. They already have reached a milestone in predicting the future, with their awesome research in various domains.

1.1 Data Science Domains



1.2 How to solve a problem in Data Science?

Problems in Data Science are solved using Algorithms. But, the biggest thing to judge is which algorithm to use and when to use it?
Basically, there are 5 kinds of problems that you can face in Data Science.

01 Is this A or B?	Classification Algorithm
02 Is this weird?	Anomaly Detection Algorithm
03 How much or how many?	Regression Algorithm
04 How is this organized?	Clustering Algorithm
05 What should I do next?	Reinforcement Learning Algorithm

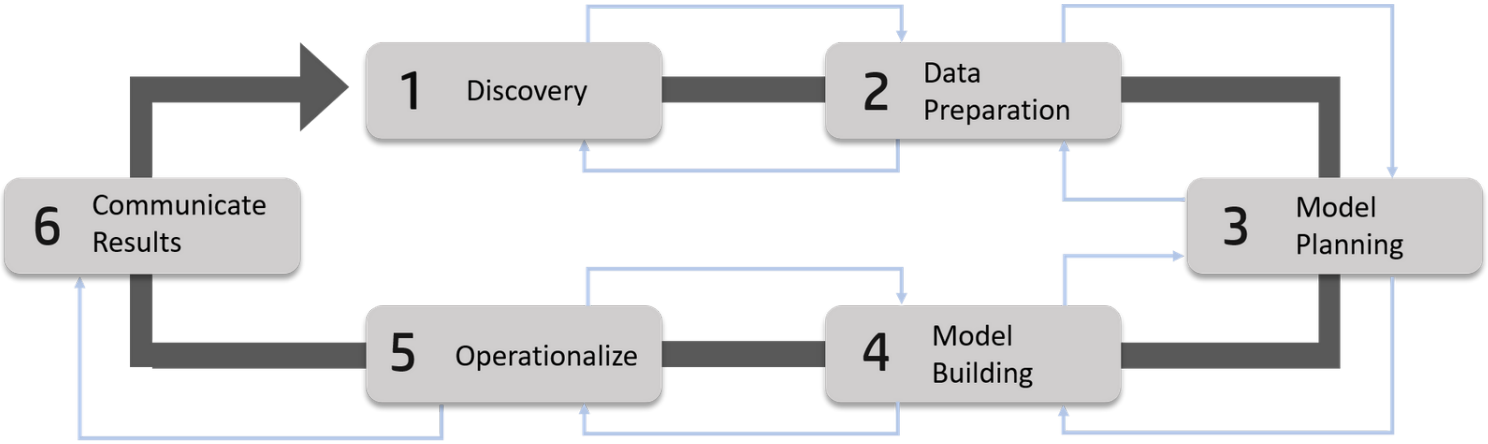
How these algorithms work?

These algorithms are based on human psychology. We like being appreciated, right? Computers implement these algorithms and expect to be appreciated when being trained. How? Let's see.
Rather than teaching the computer what to do, you let it decide what to do, and at the end of that action, you give either positive or negative feedback.
It's just like training your dog. You cannot control what your dog does, right? But you can scold him when he does wrong. Similarly, maybe patting him on the back when he does what is expected.
With each feedback, your system is learning and hence becomes more accurate in its next decision.
This type of learning is called *Reinforcement Learning*.

Chapter 2

DATA SCIENCE LIFECYCLE

A common mistake made in Data Science projects is rushing into data collection and analysis, without understanding the requirements or even framing the business problem properly. Therefore, it is very important for you to follow all the phases throughout the lifecycle of Data Science to ensure the smooth functioning of the project.



1

DISCOVERY

Before you begin the project, it is important to understand the various specifications, requirements, priorities and required budget. You must possess the ability to ask the right questions. Here, you assess if you have the required resources present in terms of people, technology, time and data to support the project. In this phase, you also need to frame the business problem and formulate initial hypotheses (IH) to test.

2

DATA PREPARATION

In this phase, you require an analytical sandbox in which you can perform analytics for the entire duration of the project. You need to explore, preprocess and condition data prior to modeling. Further, you will perform ETLT (extract, transform, load and transform) to get data into the sandbox. Let's have a look at the Statistical Analysis flow below.

3

MODEL PLANNING

Here, you will determine the methods and techniques to draw the relationships between variables. These relationships will set the base for the algorithms which you will implement in the next phase. You will apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools. Let's have a look at various model planning tools.

R Programming

SQL Analysis Services

SAS/ ACCESS

4

MODEL BUILDING

In this phase, you will develop datasets for training and testing purposes. Here you need to consider whether your existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing). You will analyze various learning techniques like classification, association and clustering to build the model.

You can achieve model building through the following tools.

SAS Enterprise Miner	WEKA	SPCS Modeler
Matlab	Alpine Miner	Statistica

5

OPERATIONALIZE

In this phase, you will deliver final reports, briefings, code and technical documents. Moreover, sometimes a pilot project is also implemented in a real-time production environment. This will provide you a clear picture of the performance and other related constraints on a small scale before full deployment.

6

COMMUNICATE RESULTS

Now it is important to evaluate if you have been able to achieve the goal that you had planned in the first phase. So, in the last phase, you identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Phase 1.

Chapter 3

INTRODUCTION TO MACHINE LEARNING

The term Machine Learning was first coined by Arthur Samuel in the year 1959. Looking back, that year was probably the most significant in terms of technological advancements. If you browse through the internet about 'What is Machine Learning', you'll get at least 100 different definitions. However, the very first formal definition was:

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."
>
> *Tom M. Mitchell*

What is Machine Leaning?

In simple terms, Machine learning is a subset of Artificial Intelligence (AI) which provides machines the ability to learn automatically & improve from experience without being explicitly programmed to do so. In this sense, it is the practice of getting machines to solve problems by gaining the ability to think.

How Machine Learning works?

It enables the computers or the machines to make data-driven decisions rather than being explicitly programmed for carrying out a certain task. These programs or algorithms are designed in such a way that they learn and improve over time when exposed to new data.

But wait, can a machine think or make decisions? Well, if you feed the machine a good amount of data, it will learn how to interpret, process and analyze this data by using Machine Learning Algorithms. A machine can learn to solve a problem by following any one of the following three approaches.

TYPES OF MACHINE LEARNING

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

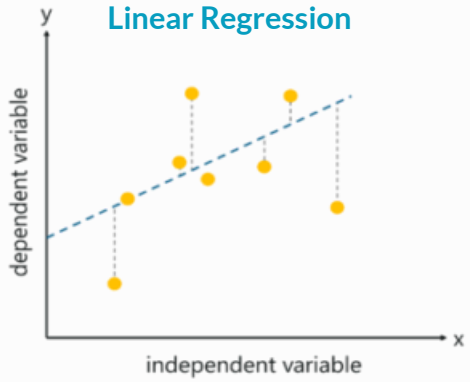
1 Supervised Learning

Supervised learning is a technique in which we teach or train the machine using data that is well labeled. To understand Supervised Learning, let's consider an analogy. As kids we all needed guidance to solve math problems. Our teachers helped us understand what addition is, and how it is done. Similarly, you can think of Supervised Learning as a type of Machine Learning that involves a guide. The labeled data set is the teacher that will train you to understand patterns in the data. Here, the labeled data set is the training data set.

Types of Supervised Learning

01. REGRESSION

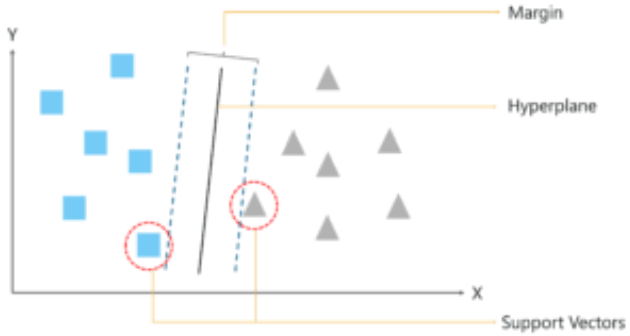
Linear Regression



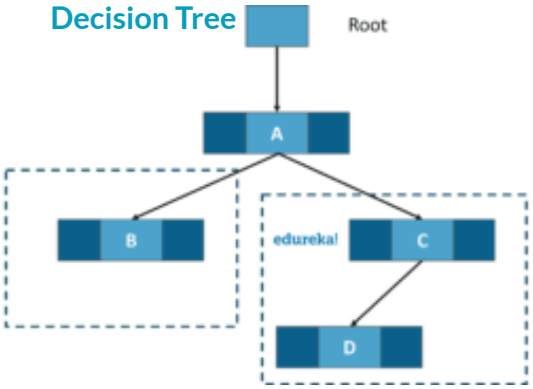
Logistic Regression

02. CLASSIFICATION

Support Vector Machines (SVM)



Decision Tree



2 Unsupervised Learning

Unsupervised learning involves training by using unlabeled data and allowing the model to act on that information without guidance. The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically detects patterns and relationships in the dataset by creating clusters in it. What it cannot do is add labels to the cluster, like it cannot say this is a group of 'apples' or 'mangoes', but it will separate all the 'apples' from 'mangoes'. Suppose we presented images of 'apples', 'bananas' and 'mangoes' to the model, so what it does, based on some patterns and relationships it creates clusters and divides the dataset into those clusters. Now if a new data is fed to the model, it adds it to one of the created clusters.

TYPES OF UNSUPERVISED LEARNING

01 Clustering	02 Association
a. Hierarchical Clustering	a. Apriori Algorithm
b. K-Means Clustering	b. FP-Growth Algorithm
c. K-NN Clustering	

3 Reinforcement Learning

Reinforcement Learning is a part of Machine learning where an agent is put in an environment. It learns to behave in this environment by performing certain actions and observing the rewards which it gets from those actions. In other terms, it is the ability of an agent to interact with the environment and find out what is the best outcome. It follows the concept of the hit and trial method. The agent is rewarded or penalized with a point for a correct or a wrong answer, and on the basis of the positive reward points gained, the model trains itself. And again once trained, it gets ready to predict the new data presented to it. This type of Machine Learning is comparatively different.

TYPES OF REINFORCEMENT LEARNING

01 Positive	02 Negative

Chapter 4

LANGUAGES TO WORK WITH DATA SCIENCE

A programming language is an arrangement of a set of instructions written in a way to generate various kinds of output. These are used in computer programs that help in implementing various algorithms and have a wide range of applications. Talking about Data Science, we have several programming languages that provide a comprehensive set of libraries for performing various Data Science functions. A Data Scientist must have command over at least one language in order to optimize the process. Some of the major languages used for Data Science are:

BEST LANGUAGES

PYTHON

JAVASCRIPT

R PROGRAMMING

SQL

SCALA

JULIA

Among all these, R is considered to be the best programming language for any statistician as it possesses an extensive catalog of statistical and graphical methods. Python, on the other hand, can do pretty much the same work as R but it is preferred by data scientists or data analysts because of its simplicity and high performance. R is a powerful scripting language and highly flexible with a vibrant community and resource bank whereas Python is a widely used object-oriented language that is easy to learn and debug.

5.1 Python for Data Science

Python is no-doubt the best-suited language for a Data Scientist. Below are a few points that will help you understand why people go with Python for Data Science:

1. **DRIVEN BY VAST AND ACTIVE COMMUNITY**
Python has one of the most known and active community which helps them in the continuous improvement of Python as a programming language. It is distributed under an open source license which makes its development easy via open source contributions.

2. **LEARNING CURVE**
With most of the programming languages, their learning curves tend to grow parabolic with time which means it is hard to grasp early but as you become familiar with this language the learning becomes easy. Whereas, in the case of Python, the learning is easy because of easy syntax and short handwriting.

3. **THIRD PARTY LIBRARIES**
Standard Python Package Installer (PIP) can help you install numerous modules that make Python interactive. These libraries and modules can interact with internet protocols, operating system calls, and many more.

4. **INTEGRATION WITH OTHER LANGUAGES**
Integration libraries like Cython and Jython make Python integrate with C/C++ and Java for cross-platform development. This makes Python even more powerful since we all know, every language has its own forte, so using these libraries you can enjoy the powerful features of each language.

5.2 R Programming for Data Science

R is basically an open-source programming and statistical language. It is a multi-purpose programming language popularly used in the field of Data Science.

1. **DATA MANIPULATION**
With R, you can easily shape the dataset into a format that could be easily accessed and analyzed by slicing large multivariate datasets.

2. **R HAS BUILT-IN FUNCTIONS FOR DATA ANALYSIS**
In R, we can use the summary built-in function to analyze summary statistics. Whereas in Python, you need to import packages such as statsmodels to do this.

3. **8000+ PACKAGES**
R has over 8000 packages that can be used to implement various statistical analysis tools related to Hypothesis Testing, Model Fitting, Clustering techniques, and Machine Learning

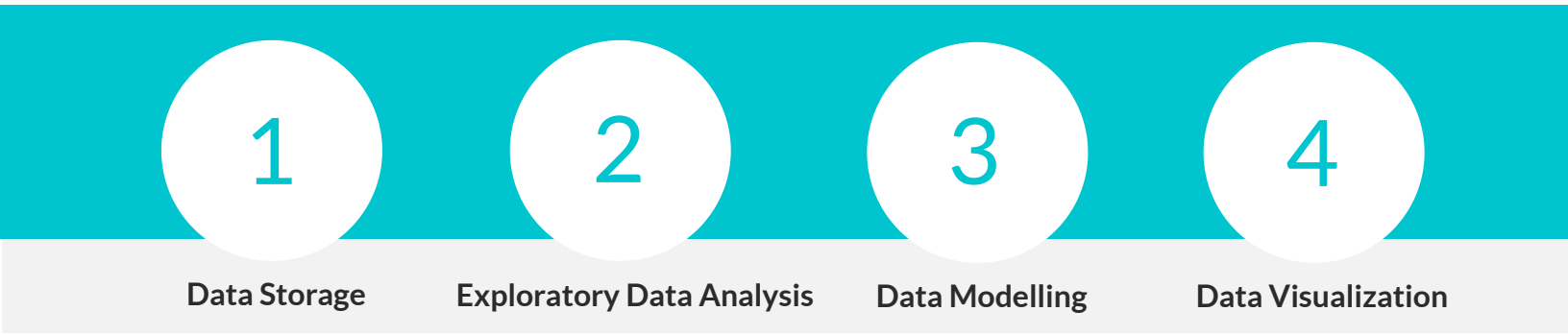
4. **DATA VISUALIZATION**
Provides a comprehensive set of functionalities for data visualization that helps in understanding a dataset and the relationship between various variables.

5. **MASSIVE COMMUNITY SUPPORT**
R is the most sought-after technology because of its ingenuity and community support. Over 2.5 million users are using R. Companies such as TechCrunch, Google, Facebook, Mozilla, etc make use of R.

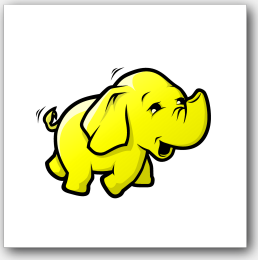
Chapter 6

DATA SCIENCE TOOLS

The main feature of Data Science tools is that you don't have to use programming languages in order to implement Data Science functions. They come with pre-defined functions, algorithms, and a very user-friendly GUI. Hence, they can be used to build convoluted Machine Learning models without the use of a programming language. Several start-ups and tech giants have been working on developing such user-friendly Data Science tools. However, since Data Science is a very vast process, it is often never enough to use one tool for the entire workflow. Let's take a look at Data Science tools used for the different stages in a Data Science process:



6.1 Data Science Tools For Data Storage



APACHE HADOOP

Apache Hadoop is a free, open-source framework that can manage and store tons and tons of data. It provides distributed computing of massive data sets over a cluster of 1000s of computers. It is used for high-level computations and data processing.

AZURE HDINSIGHT

Azure HDInsight is a cloud platform provided by Microsoft for the purpose of data storage, processing, and analytics. Enterprises such as Adobe, Jet, and Milliman use Azure HDInsights to process and manage massive amounts of data.

6.2 Data Science Tools For Exploratory Data Analysis

INFORMATICA POWERCENTER

The buzz around Informatica is explained by the fact that their revenue has rounded off to around $1.05 billion. Informatica has many products focused on Data Integration. However, Informatica PowerCenter stands out due to its Data Integration capabilities.

RAPIDMINER

There is no surprise that RapidMiner is one of the most popular tools for implementing Data Science. RapidMiner is ranked at number 1 in the Gartner Magic Quadrant for Data Science Platforms 2017, and in the Forrester Wave for predictive analytics and ML.

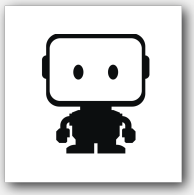
6.3 Data Science Tools For Data Modelling

H2O.AI

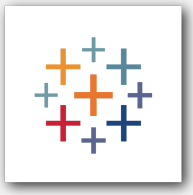
H2O.ai is the company behind open-source Machine Learning (ML) products like H2O, aimed to make ML easier for all. With around 130K Data Scientists and approx 14K organizations, the H20.ai community is growing at a strong pace. H20.ai is an open-source Data Science tool that is aimed at making Data Modeling easier.

DATAROBOT

DataRobot is an AI-driven automation platform, that aids in developing accurate predictive models. DataRobot makes it easy to implement a wide range of Machine Learning algorithms, including Clustering, Classification and Regression Models.



6.4 Data Science Tools For Data Visualization



TABLEAU

Tableau is the most popular Data Visualization tool used in the market. It allows you to break down raw, unformatted data into a processable and understandable format. Visualizations created by Tableau can easily help you understand the dependencies between the predictor variables.

QLIKVIEW

QlikView is another Data Visualization tool that is used by more than 24,000 organizations worldwide. It is one of the most effective visualization platforms for visually analyzing data to derive useful business insights.

Chapter 7

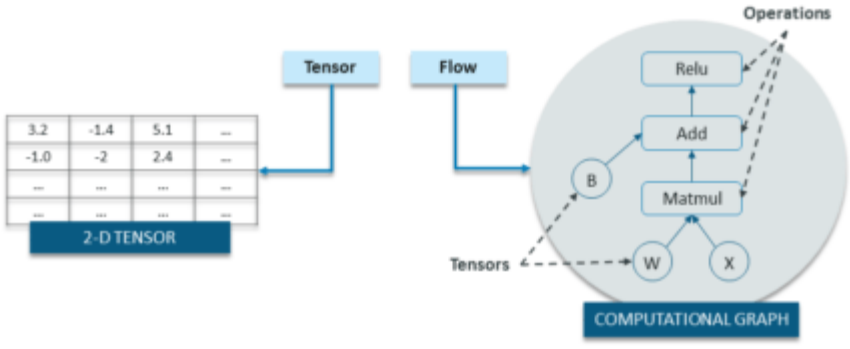
DATA SCIENCE FRAMEWORKS

This chapter will introduce you to the best frameworks present in the Data Science world and help in solving various Data Science day-to-day problems.

BEST FRAMEWORKS

7.1 TensorFlow

TensorFlow is a library based on Python that provides different types of functionality for implementing Deep Learning Models. The term TensorFlow is made up of two terms – Tensor & Flow.



In TensorFlow, the term 'tensor' refers to the representation of data as a multi-dimensional array whereas the term 'flow' refers to the series of operations that one performs on tensors as shown in the above image.

Basically, the overall process of writing a TensorFlow program involves two steps:

1. Building a Computational Graph
2. Running a Computational Graph

7.2 Scikit Learn

Scikit learn is a library used to perform Machine earning in Python. Scikit learn is an open-source library that is licensed under BSD and is reusable in various contexts, encouraging academic and commercial use. It provides a range of supervised and unsupervised learning algorithms in Python. It consists of popular algorithms and libraries. Apart from that, it also contains the following packages:

01 NumPy	02 Matplotlib	03 SciPy
NumPy is a Python package that stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object.	matplotlib.pyplot is a plotting library used for 2D graphics in Python programming language. It can be used in Python scripts, shell and web application servers.	SciPy is an open-source Python library used to solve scientific & mathematical problems. It is built on the NumPy extension and helps in data manipulation and visualization.

7.3 PyTorch

PyTorch is a Python-based scientific computing package targeted at two sets of audiences:

1. A replacement for NumPy to make use of the power of GPUs.
2. Deep Learning research platform that provides maximum flexibility and speed.

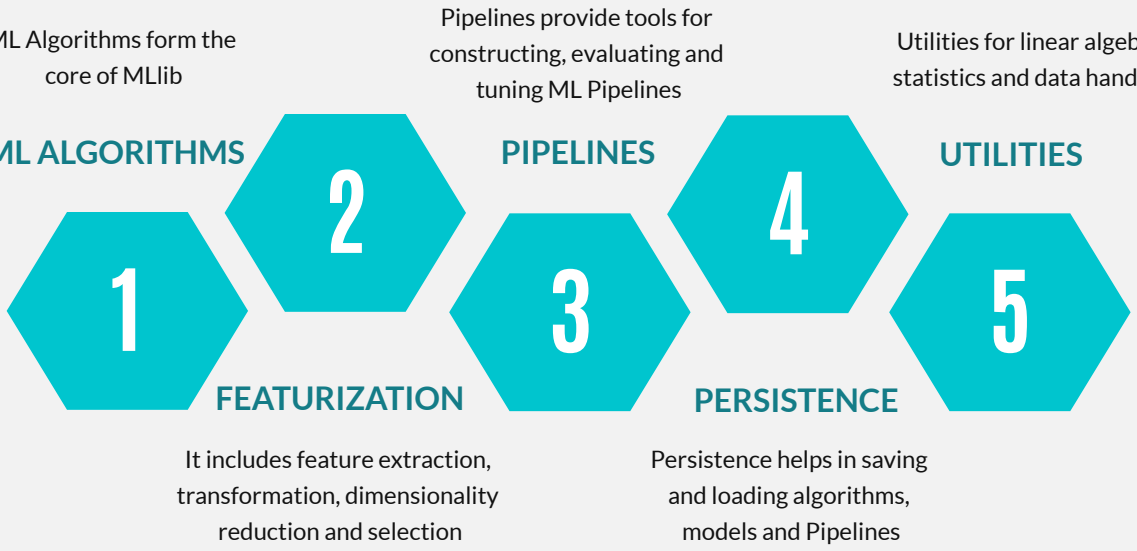
FEATURES

1. **NATIVE SUPPORT**
   Native support for Python and use of its libraries
2. **USED IN FACEBOOK**
   Actively used in the development of Facebook for all of it's Deep Learning requirements in the platform
3. **EASY TO USE API**
   PyTorch ensures an easy to use API which helps with easier usability and better understanding when making use of the API
4. **FAST PROCESSING**
   PyTorch is fast and feels native, hence ensuring easy coding and fast processing
5. **SUPPORT FOR CUDA**
   The support for CUDA ensures that the code can run on the GPU, thereby decreasing the time needed to run the code and increasing the overall performance of the system
6. **DYNAMIC COMPUTATION GRAPHS**
   Dynamic Computation Graphs are a major highlight here as they ensure the graph build-up dynamically – at every point of code execution, the graph is built along and can be manipulated at run-time

7.4 Spark MLlib

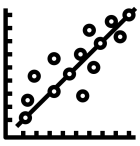
Spark MLlib is Apache Spark's Machine Learning component. One of the major attractions of Spark is the ability to scale computation massively, and that is exactly what you need for Machine Learning algorithms. But the limitation is that all Machine Learning algorithms cannot be effectively parallelized. Each algorithm has its own challenges for parallelization, whether it is task parallelism or data parallelism. **spark.mllib** contains the original API built on top of RDDs. It also provides higher-level API built on top of DataFrames for constructing ML pipelines. spark.ml is the primary Machine Learning API for Spark at the moment.

SPARK MLLIB TOOLS



MLLIB ALGORITHMS

BASIC STATISTICS



REGRESSION

CLASSIFICATION

CLUSTERING

RECOMMENDATION SYSTEM

DIMENSIONALITY REDUCTION

FEATURE EXTRACTION

OPTIMIZATION

Chapter 8

FREQUENTLY ASKED INTERVIEW QUESTIONS

As per Glassdoor and Linkedin, **Data Scientists** have consistently been ranked at number 1 because of their high demand in the business and tech industries. This chapter covers the questions which will help you in your Data Science Interviews and open up various career opportunities for a Data Science aspirant.

1. What is Data Science? How would you say it is similar or different to Business Analytics and Business Intelligence?
2. Which package is used to do data import in R and Python? How do you do data import in SAS?
3. How do you build a custom function in Python or R?
4. What is an RDBMS? Name some examples for RDBMS? What is CRUD?
5. Define a SQL query? What is the difference between SELECT and UPDATE Query? How do you use SQL in SAS, Python, R languages?
6. What is an API ? What are APIs used for?
7. What is NoSQL? Name some examples of NoSQL databases. What is a key value store? What is column storage? What is a document database?
8. What is a data warehouse?
9. What is JSON and What is XML?
10. Name some kinds of graphs and explain how you would build them in Python or R.
11. How do you check for data quality?
12. What is an outlier? How do you treat outlier data?
13. What is missing value imputation? How do you handle missing values in Python or R?
14. Why do you need a for loop? What do you do for loops in Python and R?
15. What is the advantage of using the apply family of functions in R? How do you use lambda in Python?
16. What packages are used for data mining in Python and R?
17. What is Machine Learning? What is the difference between Supervised and Unsupervised methods?
18. What are random forests and how is it different from decision trees?
19. What is logistic and linear regression? Name some packages in R and Python for building regression models.
20. What is linear optimization? Where is it used? What is the travelling salesman problem?
21. What is CART and CHAID? How is bagging different from boosting?
22. What is a Z test, Chi Square test, F test and T test?
23. What are Entropy and Information gain in the decision tree algorithm?

CAREER GUIDANCE

WHO IS A DATA SCIENCE PROFESSIONAL?
A **Data Science Professional** dons many hats in his/her workplace. They are not only responsible for business analytics but also are involved in building data products and software platforms, developing visualizations and ML algorithms.

Data Scientist

Data scientists are those who crack complex data problems with their strong expertise in certain scientific disciplines. They work with several elements related to mathematics, statistics, computer science, etc. They make a lot of use of the latest technologies in finding solutions and reaching conclusions that are crucial for an organization's growth and development.

Data Visualizers

Data Visualizers are technologists who translate data analytics into vital information for businesses to use. They manage to harness the data analytics in layman's language, to be able to communicate the outcome to all the parts of the company.

Data Architect

People with traditional programming and Business Intelligence background also good at dealing data ambiguity, have all the prerequisites to become **Data Architects**. They are often familiar with undefined and unstructured type of data and statistics. Data Architects are also creative enough to harness the data in new ways for new insights.

Data Engineers

Data Engineers are the heart and soul of 'Data Science'. They are responsible for designing, building and managing the Big Data infrastructure. They do play a major role in developing the architecture to analyze and process the data according to the business need.

Database Administrators

Database Administrators are responsible for administrating the collected data is an important task for organizations' decisions. They make use of multiple software tools to store and organize data for further analysis.

Data Analyst

Data Analysts deliver value to their companies by taking information about specific topics and then interprets, analyzes, and presents findings in comprehensive reports. Many different types of businesses use data analysts to help. As experts, data analysts are often called on to use their skills and tools to provide competitive analysis and identify trends within industries.

EDUREKA DATA SCIENCE TRAINING

DATA SCIENCE PYTHON CERTIFICATION

Weekend/Weekday | Live Class | 24 x 7 Technical Assistance

www.edureka.co/data-science-python-certification-course

DATA SCIENCE MASTERS PROGRAM

Weekend/Weekday | Live Class/SP | 24 x 7 Technical Assistance

www.edureka.co/masters-program/data-scientist-certification

MACHINE LEARNING CERTIFICATION

Weekend | Live Class | 24 x 7 Technical Assistance

www.edureka.co/machine-learning-certification-training

DEEP LEARNING WITH TENSORFLOW 2.0

Weekend | Live Class | 24 x 7 Technical Assistance

www.edureka.co/ai-deep-learning-with-tensorflow

LEARNER'S REVIEWS

S — Shehna

I did the training on **Python for Data Science from Edureka**. Instructor was very knowledgeable. The support team is very responsive. I would definitely recommend **Edureka**

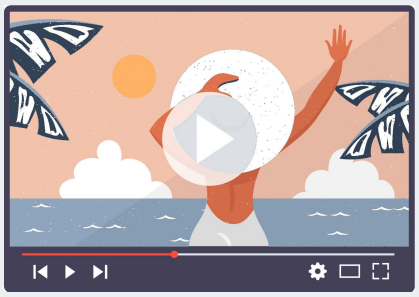
PK — Pawan Kumar

The course content is very good with easily understandable text with examples. The **support team** at back end is excellent as they give full support for any kind of doubts **24*7**.

SN — Swapnil Naresh

The course curriculum was **well structured** and the instructor taught me very well. The query were solved conveniently. Overall great learning and experience.

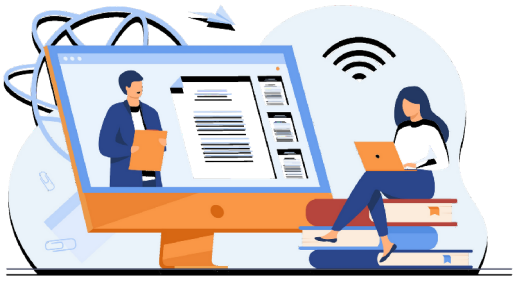
Free Resources



3000+
Video Tutorials on
YouTube

2500+ Technical
Blogs

Active
Community



30+
Free Monthly
Webinars

About Us

There are countless online education marketplaces on the internet. And there's us. We are not the biggest. We are not the cheapest. But we are the fastest growing. We have the highest course completion rate in the industry. We aim to become the largest online learning ecosystem for continuing education, in partnership with corporates and academia. To achieve that we remain ridiculously committed to our students. Be it constant reminders, relentless masters or 24 x 7 online technical support - we will absolutely make sure that you run out of excuses to not complete the course.

Contact Us

IndiQube ETA, 3rd Floor,
No.38/4,
Adjacent to Dell EMC2,
Dodanekundi,
Outer Ring Road, Bengaluru,
Karnataka - 560048

IN: 08035068112 | US: +1415 6970520
www.instagram.com/edureka.co/
www.facebook.com/edurekaIN
www.linkedin.com/company/edureka/
www.youtube.com/user/edurekaIN
t.me/s/edurekaupdates
twitter.com/edurekaIN
in.pinterest.com/edurekaco/

News & Media

INDIA TODAY

Edureka partners with **NIT Warangal** to upskill IT professionals in **AI and Machine Learning**

Deloitte.

Edureka (Brain4ce Education Solutions) tops Deloitte Tech Fast 50 2014 rankings

edureka!