Date:30 OCTOBER 2023

Class-1

Data Science:

In this course, we have 3 phases, first one is **Data Science Foundation Preparation**(self study , which is going to take 2 weeks), phase 2 is **Python Statistics Machine Learning Advances ML Concepts SQL Tableau Deployment Projects**(Which is going to take 4 months to complete).

Phase 3 would be Capstone projects + 1 live/client project (Active Mentoring of SMEs and Project Team) PAT(Placement Assistance)

To understand the **Python , Statistics, Machine Learning, Advances ML ,Concepts SQL** , students must have the basic knowledge about Database , SQL, statistics and Python .

In this session, we will be seeing how the real time projects is performed.

**Who can pursue this course?**

* Working Professionals, who want to switch to Data Science
* Data Professionals, Business Analyst, Software Engineers
* Senior Professionals, Managers
* Students, Beginners from any background
* Anyone, who wants to learn Data Science

Data science is not restricted to any domain.

**Introduction about Data Science:**

Data Science is nothing but it is the science of analyzing data and extracting some meaningful insights or instructions from data.

Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structured and unstructured data. (Ref: Wikipedia)

Structured data means the data in the form of some kind of order, like tables, etc.

Unstructured data includes the data in the form of images, videos, audios etc.

Data Science is the field of study that combines domain expertise, programming skills and knowledge of mathematics and statistics to extract actionable insights. Data Science an be simply stated as “Insights from Data”.

As a data scientist, you need to understand the domain by doing thorough research.

**Why we need to get instruction from the data?**

Getting the insights and instructions from the data is very important, it can be any type of data.

Example: Let us consider, if you’re running a business and you have an organization and in that organization , people are leaving every year and you have a data of attrition:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name of Employee** | **Age** | **Role** | **Attrition Details (left or not)** | **Year** |
| X  Y | 35  24 | HR  Developer | Yes  No | 2021  2022 |

**Fig: Information about the employees**

Suppose you’ve started your company from 2020 and this is the data from 3 years, you’ll see that attrition is increasing and you have to control the attrition. How would you control it?

First, we have to look through the factor that is influencing people to leave an organization. It could be due to;

No promotion, low salary, boredom, no or less job satisfaction, no challenges at work etc. These factors are nothing but insights that you’ve derived from the data from the people who have left the organization. Based on the data, you will try to control the attrition. With the help of other peoples, you can help to increase the job satisfaction of employees by providing more salary, promotions and other different kinds of incentives.

Data science helps to get some insights from the data by analyzing and studying them.

**Evolution of Data Science:**

****

**Fig: Evolution of Data Science**

When computers were not the part of our daily lives, and was not accessible everywhere, we used to physically store the data and it used to be analyzed manually. When it was not possible to handle manually as the data became large, computers were used to store the data. Slowly, all the data were digitalized .

The first thing that came to the picture was a centralized (Mainframe).

**Centralized** refers to system where all processing and data storage is handled. The single system is used. It is responsible for processing all the requests and managing the data.

After that, **Data Warehouse** was introduced where the data is collected on extensive scale to perform analytics and warehouse usually stores the data for combining different databases. It was usually used by an organization to store their details.

**Business Intelligence** basically was procedural infrastructure that select, store and analyzes some business data and it is usually presented in user friendly manner with reports, charts, graphs.

**Data Mining** is basically a large data in which we try to find patterns and relationships that tends to solve problem in data analysis. It is used to predict future trends and situations.

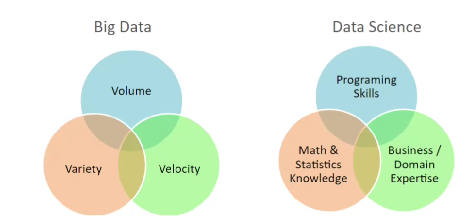
**Business Analytics** is a process of transforming data into insights to make certain decisions for your business. Data visualization, data mining is used for business insights.

**Data science**, as we all know, is a science of studying data for some purpose. It combines everything like mathematics, statistics.

From storing the data and analyzing them manually till data science, it has evolved rapidly. It took more than 50+ years to come up in this phase called Data Science.

In this course, we are learning about only Data Science.

**Big Data vs Data Science:**

****

Big Data refers to amount of data we have.

**When do you say data is big?**

Based on the volume (How big is your data), variety (Structured or unstructured data) and velocity (At what speed data is getting generated).

Talking about Data Science, It is a field which requires programming skills(Python is more preferred since it is easier to understand and implement), mathematics and statistics knowledge and business/domain expertise.

**Data Science vs Business Analytics**

|  |  |
| --- | --- |
| Data Science | Business Analytics |
| Involves programming/coding. | Minimal/No coding required |
| Data Science uses both structured and unstructured data. | Business Analytics use mostly structured data. |
| Data Science extensively involves Machine Learning and Artificial Intelligence as tools. | Business Analytics use traditional statistical and forecasting techniques. |

**Business Analytics Classification:**

**Level 1: Descriptive Analytics**

This is the basic level of analytics every organization has. It is used to describe the data. Here’s how;

Presenting data in informative way or understandable manner. In this analytics, we are basically trying to understand what happened in past. For Example: Average sales made by the company in the last year. You’re going to summarize your data from the representative value called “Average Sales”. so, Average sales is going to describe your data with some single value. Say, you have sales of 12 months. You will be describing your 12 months data in a single informative way ; which is called Descriptive Analytics.

**Level 2: Diagnostic Analytics**

Without performing Descriptive analytics, you cannot perform Diagnostic Analytics. This analytics talks mutually about why it happened. Diagnostic analytics talks about the root cause.

Let’s link the example of Descriptive analytics here, suppose, Average sales is low, it will diagnose why the sales are low. It tries to understand the sectors which has resulted the sales to go down.

It helps to provide the decision making capacity based on the root cause and helps to solve the problem.

**Level 3: Predictive Analytics**

It is one of the powerful analytics which helps the business. Here, we use machine learning or AI to do predictive analytics; which predicts the future.

Example: Companies can use this analytics to predict their sales in upcoming year. Netflix uses this predictive model, they predict what the customers would like to see next based on the past data.

This analytics tells you what kind of decision you should make further. It can be applied to any domain, like, ecommerce, health sectors, entertainment, stocks exchange. Data is the most important factor for any analytics, concise data should be used.

**Level 4: Prescriptive Analytics**

This is the final level of analytics. By the name, you can tell that it is used to recommend or prescribe your certain decisions. This analytics tells you what should be your next move. When you have prescriptive analytics in your organization, it means that you have completely automated processes.

Example: When the Predictive Analytics tells you that the sales in the upcoming year is going to be low, Prescriptive analytics gives you the suggestions to improve your tactics.

As no company has implemented this level, yet if any organizations has reached to this prescriptive analytics stage, then it is called evolved organization.

**Data Science:**

A key part of data science is data analysis and preparation. Let’s dig a little deeper….

There are two important thing in data science, one is data analysis and anothr is data preparation.

Before we do certain predictions and before we prepare model for those predictions, we have to follow certain steps, then only the model can learn data and then it can do certain predictions. The steps are:

**1.Domain Analysis:**

It is necessary to understand the domain prior to proceeding. This helps to gain a business perspective on the analysis which follows.

Suppose you’re collecting the data about diabetic patients, which contains the information about Blood Pressure and Insulin. As a domain expert, you must have knowledge about all the small details. In this case, you have to at least know about BP and Insulin. If you are new to any domain, first you should do research about all of its components.

**2.Data Analytics/Analysis:**

After getting well information of the domain, the next thing is data analytics. Analytics then allow us to arrive at patterns and relationships within the data. This includes tools, graphical analysis.

Example: We have the data of the patient and their age is given.

|  |
| --- |
| Age\_of\_Patients |
| 20 |
| 40 |
| 25 |
| 20 |

**Fig: An example for data analysis**

Suppose, you want to find out what is the maximum age they all(Patients) belong to and you have around 500 data . You can do analysis with the help of graph. In this list of data, the answer would be 20.

We can use bar-graph, histogram, pie-chart, etc.

**3.Data Preparation:**

Data preparation is also known as data preprocessing/Data cleaning/data wrangling. This is an essential stage of data science. Once the data is prepared, we may use it for modelling to perform predictive analytics. The cleaned data only could give you better prediction.

You cannot just use the raw data. There might be some features that might be irrelevant. There might be missing values, extreme values(which might be extra compared to your other data) or corrupted data which can actually affect your analysis. You have to understand what kind of cleaning is required for your data.

There are certain methods to handle the missing values, which are called NAN values(Non-applicable).

Outliers are the extreme values which are not near to the normal values. There are certain

methods to handle this value. We will see a lot about data preparation technique when we reach to machine learning model.

Suppose you have a data about patient only. We have 50 patients and we are going to predict whether they have diabetes or not.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Age** | **BP** | **Insulin** |
| Harry | 45 | NaN | NaN |
| Aayu | 30 | 110/30 | NaN |
| Henrry | 50 | 120/80 | 231 mIU/L |

Lets say that some of the fields of BP and Insulin level are empty. When data containing fields are empty, It would not give you predictions. These are the important factors that you have to handle to ensure the data is authentic and useable to perform any kind of operations.

Hence, these three steps are crucial part to follow before building the Data Model

After that, the data goes into the Machine Learning Model (ML Model) and gives you the right predictions and answers based on the list of collected data.