DS CAT 1 QB UNIT 1

Q . 2) Explain Data Science Process in detail?

Ans . 2)

The data science process is a systematic approach to solving a data problem. It provides a structured framework for articulating your problem as a question, deciding how to solve it, and then presenting the solution to stakeholders.

Data Science is all about a systematic process used by Data Scientists to analyze, visualize and model large amounts of data. A data science process helps data scientists use the tools to find unseen patterns, extract data, and convert information to actionable insights that can be meaningful to the company.

This aids companies and businesses in making decisions that can help in customer retention and profits. Further, a data science process helps in discovering hidden patterns of structured and unstructured raw data. The process helps in turning a problem into a solution by treating the business problem as a project.

Components of Data Science Process

1. **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.**

2. **Statistics – 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.**

3. **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.**

4. **Advanced Computing**

    a. **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.**

    b. **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.**

Q . 3) Explain first Phases Of Data Science Process?

Ans . 3)

Phases Of Data Science Process

1. **Discovery**

   The first phase is discovery. This is where we define and understand the problem. This involves asking the right questions and determining all the required factors such as technology required, number of people, data, budget and also decide on estimated deadlines. This is the most essential phase as the whole data science life cycle revolves around solving the business problem. So it is necessary that the business problem is defined on the first hypothesis level before proceeding further.

2. **Understanding data**
   After defining the business problem, the next step is understanding the data. This includes a wide array of data that can be accessed. Enterprises usually store their data in data warehouses. We need to closely work with our peers to understand what information is stored and what is the required information to be used. This step involves describing what data is needed, how relevant are they and finally extracting the required data.

3. **Data preparation**
   Next step is data preparation. Here we filter out the data applicable for the problem, merge different datasets, clean the data - eliminating inaccurate data, treating missing values and outliers. This phase is also known as data munging. Here, we also convert the data into desired format, eliminate columns that are not needed and derive new elements from the data acquired. This is arguably the most time consuming step but is also essential as our model will only be as accurate as our data. After this step, we can easily use data for the further phases.

4. **Data analysis**
   This is the part where exploratory data analysis is done. Here, we analyze our data, look at possible relations between various features and get an understanding of how much effect each variable has over our final prediction or target. We make use of graphs like bar graphs to visualize the distribution of data, pie charts to describe the parts of a whole and scatter plots to visualize relationship between two or more variables. In this step, we get an idea of what features to consider for building our machine learning model.

5. **Model planning**
   In this phase, we decide on our machine learning model based on the business problem. We select the suitable model - classification, clustering or regression. Once the model family is decided, we carefully choose one algorithm to implement amongst the wide range of algorithms present in each model family. Often this step is done while performing data analysis.

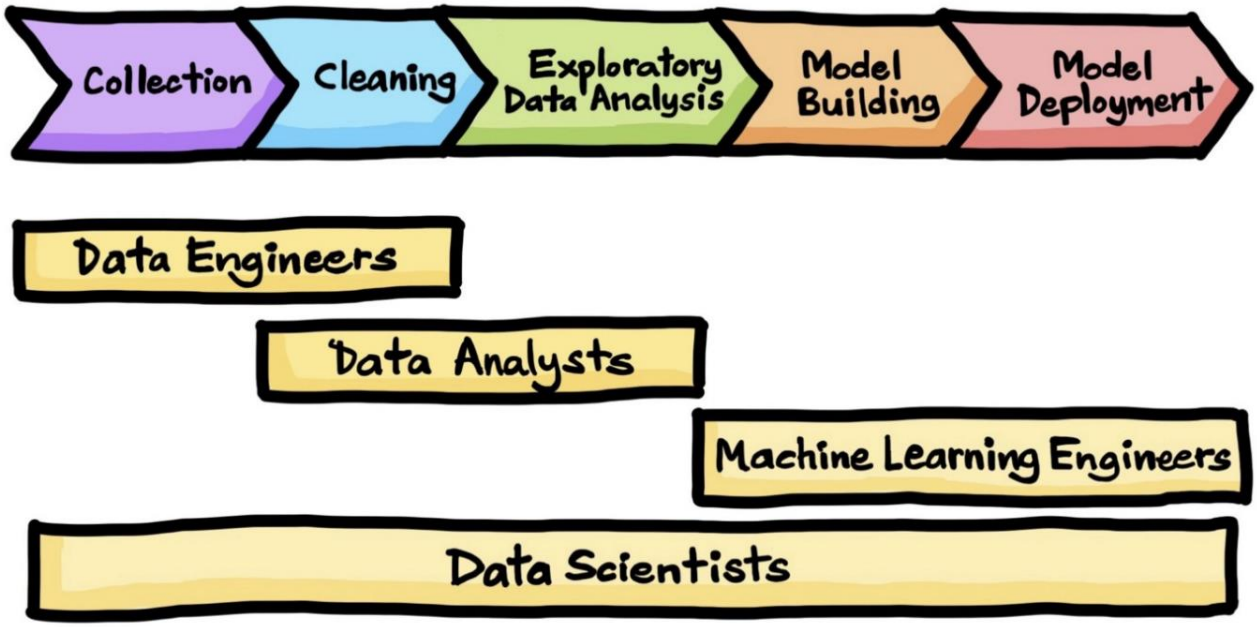
6. **Model building and deployment**
   In this phase, we create our machine learning model. We split the dataset into train and test data. We fit the train data values in the chosen algorithm and allow the machine to learn. We tune the hyperparameters of the models, adjust weights to improve the results. We additionally need to make sure that there is no generalization error and that the model performs well with other similar datasets too. The model is evaluated by feeding unseen data into it. If we are not happy with the results, we need to go back and make changes to our model until preferred metrics are achieved. After going through rigorous tests, the model is finally deployed in the preferred environment.

7. **Communication of results**
   In this phase, we reflect back to our original goal that we set in the first phase. We check if we reached our goal and the extent to which the results have been obtained. We communicate our findings to the stakeholders. This is where presentation plays a key role. Data visualization is used to convey the information in an easy way so that others could also understand the story the data told us and understand the performance of the proposed solution. This helps them in making informed decisions. This is the final step in the data science life cycle.

Q . 5) Write Short note on Data Science Process Lifecycle?

Ans . 5)

Data Science Life Cycle



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

1. **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.**

2. **Data Cleaning – 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.**

3. **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.**

4. **Model Building – 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.**

5. **Model Deployment – 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.**

Q . 6) Explain any Two Phases of Data Science Process?

Ans . 6)

Two Phases Of Data Science Process

1. **Discovery**

   The first phase is discovery. This is where we define and understand the problem. This involves asking the right questions and determining all the required factors such as technology required, number of people, data, budget and also decide on estimated deadlines. This is the most essential phase as the whole data science life cycle revolves around solving the business problem. So it is necessary that the business problem is defined on the first hypothesis level before proceeding further.

2. **Understanding data**

   After defining the business problem, the next step is understanding the data. This includes a wide array of data that can be accessed. Enterprises usually store their data in data warehouses. We need to closely work with our peers to understand what information is stored and what is the required information to be used. This step involves describing what data is needed, how relevant are they and finally extracting the required data.