# **Step Up - 0019M**

## Mandatory Assignment -1

**Q1. Briefly describe the steps involved in a typical data science process in your own words**

Ans. The steps involved in data science process are

1. Data collection and storage
2. Cleaning data
3. Exploration and analysis
4. Model development and Prediction
5. Model validation and deployment

* **Data collection and storage :** In this process the raw data is collected through many ways for the given problem. Before collecting raw data we need to understand the problem statement which helps us to collect relevant data.
* The data is collected through many ways such as surveys, polls, interviews,etc. Which is known as the primary data collection method.
* The collected data is used for different purposes like internet, govt records, sales reports etc. known as secondary data collection method.
* **Cleaning data :**In this process the unwanted data is cleaned
* Data is cleaned to fix erroneous data and fill up the missing data or delete the duplicate records and handle the outliers.
* **Exploration and analysis :**This step helps in data discretization as well as in “engineering” new variables that better explains the data and hence result in better predictions known as feature engineering
* This feature engineering helps to create new feature which could correlate product final outcome
* **Model development and Prediction :** In this process we will use machine learning models in which we make the machine learn from past data for future prediction.
* Which is divided into 3 parts

1. Supervised : in this the prediction is done from the label and features which is further classified into 2 types classification (categorical) and regression(numerical)
2. Un supervised : In this model is not guided by labeled data
3. Reinforcement : in this machine has its own rules

* **Model validation and deployment :** In this the data is splitted into training data and testing data
* Training data set: The first part of the data could be used to train and tune a prediction model. This is known as training data set
* Testing data set: The second part of the data that was split would act as a proxy to new data and will assess the model's prediction accuracy.
* Model deployment is the process of placing a finished machine learning model into a live environment where it can be used for its intended purpose.

**Q2. Citing relevant examples, classify the data as qualitative and quantitative types.**

**Ans .** data is classified into two types

1. Qualitative - This data is usually descriptive in nature

It in the form of text and refers to things that can be observed but cannot be measured such as colors (red, blue, yellow) or emotions (happy, sad).

It describes certain attributes and helps us to understand the ‘why’ and ‘how’ of certain observations.

types

* Nominal - it has no measure and no order

Ex. The categories under the variable “Transportation” could be car, bus, train, tram or bicycle.

* Ordinal - it has certain order and expressed in numbers

Ex. age variables are distributed as categories, child, teenager, youth, middle age and old.

1. Quantitative - The variable or data in which each datapoint has a certain unique numeric value

The value of data can be in the form of counts or numbers. This data can be identified as having a certain order and magnitude.

Typically, the data is collected by answering questions such as “How many?”, “How often?”, “How much?” falls in this category. (For example, how many students are present in the class? The answer would be a count).

Types

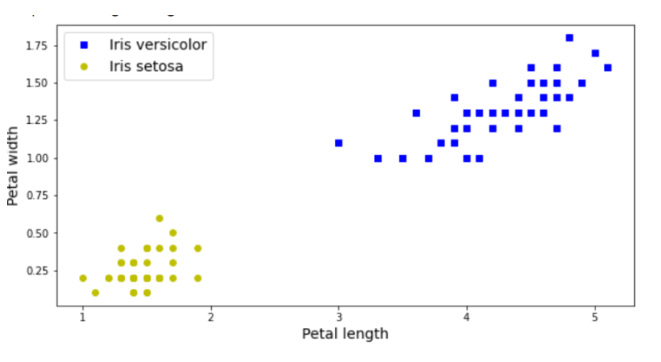
* Discrete data: When values in a data set are countable and can only take certain values, it is called discrete data.

Ex. number of books on a bookshelf, number of players required in a team, etc.

* Continuous data: The type of data that takes specific values which can be measured and falls under a given range is known as continuous data.

Ex. weight, length, temperature, speed etc

**Q3. Given below is a visual of a classical ML model. Identify the type of model and describe what it is intended to achieve?**



The given model is an unsupervised model and it is seperating iris setosa and iris versicolor with different color patterns

* Unlike supervised learning, the model here is not trained on any specific outcome. Instead the model operates upon only the input data without outputs or target variables.
* Hence the name unsupervised as the machine has to learn to make some of the data without any guidance on the outcome predicted. Here the main task is to identify what natural patterns exist in the data.
* Such a model could be used to learn which items are bought together when a customer visits the store. Common examples of unsupervised models are clustering, visualization, dimensionality reduction, and association rule learning.

An example of Clustering model:

Clustering falls under the category of unsupervised machine learning algorithms. Here, no predefined notion of a label is allocated to the groups/clusters formed,

For e.g.,in the real estate field, we may want to understand and divide the various property locations based on the value, importance, etc. Clustering algorithms can process through the data on various properties in terms of locations and price and the cost of living in those areas and identify various groups of property on the basis of probable price and find clusters (For eg group them into highly affluent, affordable, developing etc)

**Q4. Big data comes with its own set of challenges. Highlight and discuss two biggest challenges of modern big data.**

Ans. Modern ‘Big Data’

* With the rapid acceleration in data generation due to the growth in the field of technology, specifically in smart devices the need to store and capture data in various forms such as text, video, audio etc became imperative
* Modern databases store data in raw form and use a dynamic schema that can handle different formats of data.The term Big Data is used in the data definition to describe the data that is in the petabyte range or higher values than that.
* The characteristics of Big Data are described as the 5Vs: variety, volume, value, veracity, and velocity.
* Nowadays, web-based eCommerce has spread vastly, business models based on Big Data have evolved, and they treat data as an asset itself. And there are many benefits of Big Data as well, such as reduced costs, enhanced efficiency, enhanced sales, etc to name a few.

**Challenges with Modern/Big Data:**

### **1. Data growth issues**

One of the most pressing challenges of Big Data is storing all these huge sets of data properly. The amount of data being stored in data centers and databases of companies is increasing rapidly. As these data sets grow exponentially with time, it gets extremely difficult to handle.

Most of the data is unstructured and comes from documents, videos, audios, text files and other sources. This means that you cannot find them in databases.

Solution

In order to handle these large data sets, companies are opting for modern techniques, such as compression, tiering, and deduplication. Compression is used for reducing the number of bits in the data, thus reducing its overall size. Deduplication is the process of removing duplicate and unwanted data from a data set.

Data tiering allows companies to store data in different storage tiers. It ensures that the data is residing in the most appropriate storage space. Data tiers can be public cloud, private cloud, and flash storage, depending on the data size and importance.

### **2.Lack of data professionals**

To run these modern technologies and Big Data tools, companies need skilled data professionals. These professionals will include data scientists, data analysts and data engineers who are experienced in working with the tools and making sense out of huge data sets.

Companies face a problem of lack of Big Data professionals. This is because data handling tools have evolved rapidly, but in most cases, the professionals have not. Actionable steps need to be taken in order to bridge this gap.

Solution

Companies are investing more money in the recruitment of skilled professionals. They also have to offer training programs to the existing staff to get the most out of them.

Another important step taken by organizations is the purchase of data analytics solutions that are powered by artificial intelligence/machine learning. These tools can be run by professionals who are not data science experts but have basic knowledge. This step helps companies to save a lot of money for recruitment.

**Q5. “80 percent of a data scientist’s valuable time is spent simply finding, cleansing, and organizing data, leaving only 20 percent to actually perform analysis.”. Justify the above statement in your own words.**

Ans. 80 percent of a data scientist’s valuable time is spent simply finding, cleansing, and organizing data as it takes more time

Data scientists may need to contact different departments to beg for the data they need and wait weeks for it to be delivered, only to find that it doesn’t provide the information they need or has serious quality issues. At the same time, responsibility for data governance often falls to them, since corporate-level governance policies are confusing, inconsistent, or difficult to enforce.

Even when they can get their hands on the right data, data scientists need to spend time exploring and understanding it. For example, they might not know what a set of fields in a table is referring to at first glance, or data may be in a format that can’t be easily understood or analyzed. There is usually little to no metadata to help, and they may need to seek advice from the data’s owners to make sense of it.

Once they wrangle the data, there’s yet another laborious task to perform: preparing it for analysis. This step involves formatting, cleaning, and sometimes sampling the data. In some cases, they may also have to perform scaling, decomposition, and/or aggregation transformations on the data before they are ready to start training their models.

In testing data we use machine which is already trained with the information so it is simple to make analysis in small amount of time