**Synapse Task 3**

**Q. 3.1) Let’s say you are given a large amount of textual data- messages, emails, books, etc. Before performing any operations on this data, it is necessary to clean and preprocess the data (removing unnecessary words or symbols, etc.). Explain how you would go about preprocessing. What different steps would be followed? Why are they necessary?**

Ans) Data Cleaning is an essential step that we cannot miss before creating good Machine Learning models. This is because machine learning uses data and program to create an algorithm which in turn can be used on other datasets to yield desired output. If the data is not up to the mark, we cannot expect good algorithms. Let’s take an example. Consider the data collected from an online form for selections in an interview and our goal is to predict the chances he/she will get placed. The primary data needed from a student would be: name, resume, mobile number, email-id, address, overall CGPA, and in case of too many applicants, father’s name, mother’s name, github id, linkedIn id and D.O.B. But what if the applicants don’t fill D.O.B section, or the applicants apply twice using two different accounts, or instead of writing name in the format “first-name last-name” he/she ends up writing “first-name” or “first-name father-name last-name”. What if the student applied for the form and never sat for the interview? These can create change in the prediction of the model and reduce the accuracy of the same. It’s like finding the factorial of 10 but we miss out 5, or add 1 twice or thrice, or add 9 instead of 6, leading to output with less accuracy. That’s why data cleaning comes into picture. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Dataset can contain useless or missing data not only because of user’s mistake, but also due to different sources of data. Collecting a Data from an interviewee can be through forms, company registries any many more.

Now, how can we go about cleaning the data? There is no specific way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset, but considering the above example, we can do it the following way:

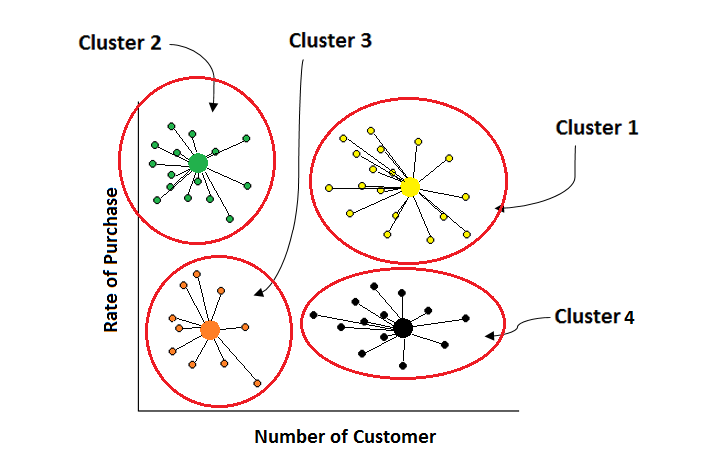
1. Remove duplicate or irrelevant observations: Irrelevant data are those which has no relation with our aimed prediction. For Example: We don’t need parent’s name to predict the chances of getting placed. Sometimes we may also find duplicate data due to different resources.
2. Fix structural errors: D.O.B written as mm/dd/yyyy fashion instead of dd/mm/yyyy, or name not in the specified format comes under fix structural errors
3. Filter unwanted outliers: Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. In terms of Regression, we are unable to plot these on the graph. Thus, it is suggested to remove the data. But sometimes these data can be proved vital for prediction, so we need to change our model rather than cleaning. Of course, if the outliers don’t contribute much, it is better to remove it.
4. Handle missing data: Just as we read above, there might be cases where data was never received whether the interviewee was placed if he never came for the interview, in that case missing values arise which needs to be sorted by either removing it or by replacing it with a value

**Q. 3.2) Have you ever wondered how streaming platforms like Netflix work and how they recommend movies or shows based on your current watch? How does a bank decide which customers get loans and which do not? This all is done using Unsupervised learning. Machine Learning is internally subdivided into different parts- one of them is Unsupervised learning. The technique used for these kinds of problems is known as Clustering. So, for this task, explain what clustering is and describe any two types of clustering.**

Ans) Clustering is basically grouping of data based on who similar they are. Taking Netflix as an example, Netflix has different varieties of movies and series. If we were to group them based on their types such as romantic, comedy, action, fiction etc we can just deploy these suggestions to the user as soon as the user starts watching any one kind of the data. Unlike supervised learning, where the model learns from labeled data, clustering doesn't require labeled data. Instead, it identifies patterns and structures within the data on its own.

There are many types of clustering, but the two main types include:

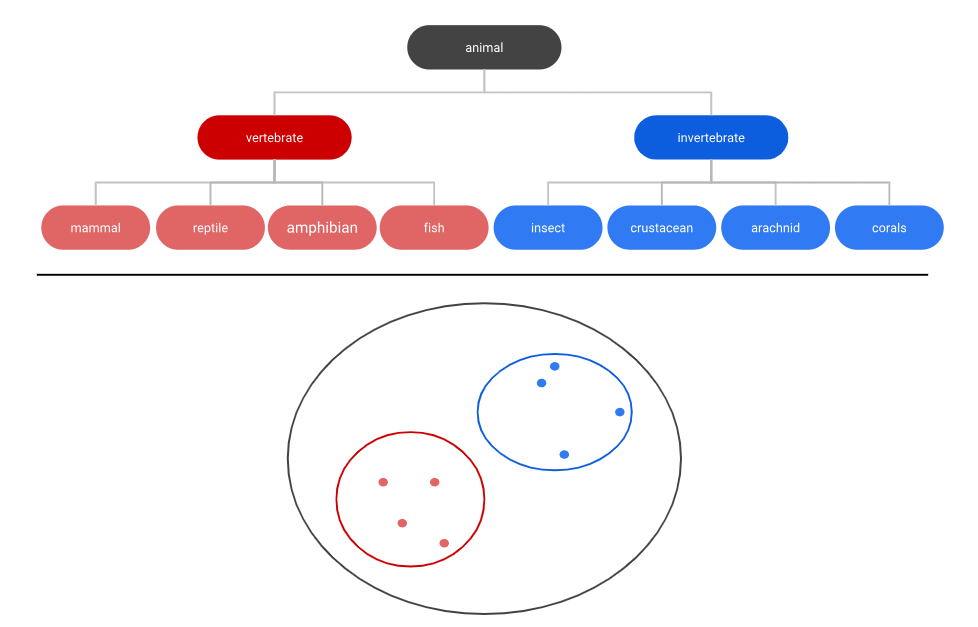
1. K-mean Clustering- In this type of clustering, every data is divided into k number of groups based on their features. The clustering begins with randomly deciding any k number of points from which we’ll start grouping. Then every data is plotted or grouped according to how close or similar the data is to any of the centroids. The process doesn’t stop here. Based on the group we again assign new centroid points and again repeat the process until there’s no major change observed further.



In the following diagram, the big points are the centroids and the data is divided into four groups. Thus, k is 4. Each and every data are grouped based on the similar centroid.

K-Means is used in market segmentation, image compression, and organizing large sets of documents, among other areas.

1. Hierarchical Clustering: Hierarchical clustering creates a tree of clusters. Suppose we want to create a model to find out what classification of animal Kingdom it belongs to, So we’ll first group data into a big cluster which will contain five major kingdoms, then in every kingdom, we’ll group the data based on how they are further classified (Example in Kingdom animalia, there are two classifications, vertebrates and invertebrates) and so on.



Hierarchical clustering is used in bioinformatics (e.g., to create phylogenetic trees), social network analysis, and customer segmentation.