# Dataset Name: **Online Shoppers Purchasing Intention Dataset Data Set**

Data Set Information: The dataset consists of feature vectors belonging to 12,330 sessions.  
The dataset was formed so that each session  
would belong to a different user in a 1-year period to avoid  
any tendency to a specific campaign, special day, user  
profile, or period.

# Attribute Information:

* The dataset consists of 10 numerical and 8 categorical attributes.
* The 'Revenue' attribute can be used as the class label.
* "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.
* The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.
* The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session.
* The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.
* The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction.
* The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date.
* For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

Instructions: Based on the above dataset the students are expected to follow the below mentioned steps.

|  |  |  |
| --- | --- | --- |
| STEPS | Instructions | MARKS |
| 1.Data Ingestion: | Read the dataset (tab, csv, xls, txt, inbuilt dataset). Do the descriptive statistics and do null value condition check, write an inference on it**.** | 5 |
| 2. Kmeans: | Apply the scaled data on the K-Means algorithm, identify how many number of clusters is optimized cluster using elbow graph and table which plots error vs number of clusters | 10 |
| 3. Hierarchical: | Apply the scaled data on the Hierarchical algorithm (Agglomerative clustering), identify how many number of clusters is optimized cluster using dendogram and table which plots error vs number of clusters, write a details inference on it | 10 |
| 4. Convert Unsupervised data into Supervised data: | Selected the output of K-means/hierarchical cluster either one and bind the data back to the original data, group the data based on the cluster and given an inference | 5 |
| 5. PCA : | Apply PCA with scaled data and identify the cluster using K-Means and Hierarchical, group the data based on the cluster and given an inference | 10 |
| 6. Data Split: | Split the data into test and train, apply Data without PCA build classification model Logistic regression, KNN, Navies Bayes, Decision Tree and write inference | 10 |
| 7. Split train and test | Split the data into test and train, apply Data after PCA build classification model Logistic regression, KNN, Navies Bayes, Decision Tree and write inference | 10 |
| 8. Final Model: | Compare all the model and write an inference which model is best/optimized and give more insight about the final model | 10 |