DOUGLAS COLLEGE – Summer 2023

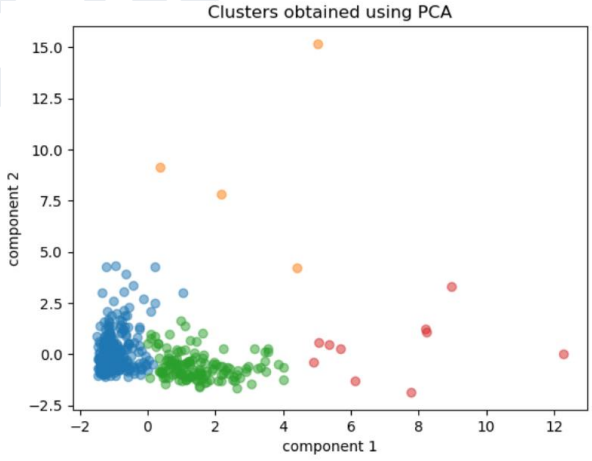
CSIS 3290 – 001 – Lab 5

- **Create a folder and name it according to naming convention stated below**
- **All the files you are required to submit for the assignment should be placed inside this folder.**
- **If cheating is determined (i.e., you shared your work with another student in the class), your work will a ZERO mark and you will face further consequences.**
- **Make sure to include all the necessary files to make sure that the code can run properly**

In this lab, we practice implementing clustering and DNN for classification models. You need study the textbook, demo code and do your own research to make sure that you can perform all the tasks describe below.

Part1 (clustering, pipeline and PCA)

1. Create a python notebook named as **Lab5_part1_ABcXXXXX** with A signifies the first letter of your **first name**, Bc signifies the first two letters of your **last name** and XXXXX denotes the last five digits of your **student ID**
2. Create a markdown cell to state the lab, **your name and student ID** with the correct heading.
3. **For each of the following section**, you need to create a **markdown heading cell** followed by a few code cells to complete the tasks. Please also put some comments in each code cell:
   a. **Load the python library**. Please load all the required python libraries in this section
   b. **Read the data**. Use the *Lab5_wholesome_customer_data.csv* file.
   c. **Drop the unneeded data.** Check if there is any null and drop them.
   d. Use the elbow method to **determine the optimal number of clusters**
   e. **Use a pipeline** to implement a standard scaler, PCA (with 2 components) and different clustering algorithms for the provided dataset.
      The clustering algorithm to be used are: AgglomerativeClustering, KMeans, Birch (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.Birch.html), and SpectralClustering (see https://scikit-learn.org/stable/modules/clustering.html). Please specify the hyperparameter n_clusters to be equal to the optimal number of clusters obtained from the previous step
   f. **Evaluate the clusters.**
      - Calculate the silhouette score, carlinszky harabaz score and davies bouldin score to find the best cluster
      - Create a dataframe to combine the result and pick the best pipe
      - Create another dataframe consisting of the PCA components and the cluster label and plot it. Below is an example of such a plot.
      - Hint: Create another pipeline consisting of the standard scaler and PCA, fit transform to get a new dataframe, add the cluster label (obtained from the best clustering algorithm) to this df and plot it.



g. (Optional step) Add the cluster label to the original dataframe. The dataset is about annual spending of wholesome customers (https://archive.ics.uci.edu/dataset/292/wholesale+customers). How do you think the clusters were formed? Which features (values) are more prominent in each cluster? What kind of customers do you think each cluster represent?

Part2 (classification)

1. Create a python notebook named as **Lab5_part2_ABcXXXXX** with A signifies the first letter of your **first name**, Bc signifies the first two letters of your **last name** and XXXXX denotes the last five digits of your **student ID**.
2. Create a markdown cell to state the lab, **your name and student ID** with the correct heading.
3. **For each of the following section**, you need to create a **markdown heading cell** followed by a few code cells to complete the tasks. Please also put some comments in each code cell.
   a. **Load the python library**. Please load all the required python libraries in this section
   b. **Read the data**. Use the *Lab5_user_behavior.csv* file.
   c. **Drop the unneeded data.** Check if there is any null and drop them.
   d. **Create dummy values for the dataframe**, do not forget to use the drop_first parameter.
   e. **Prepare the features and target variable**, use the train_test_split to split the dataframe into training and test such that the test size is 25% and specify the random state value.
   f. **Prepare the normalization layer** using the training dataset
   g. **Create the DNN sequential model for classification** that includes:
      - Normalization layer
      - Dense layer of 64 neurons
      - Dense layer of 32 neurons
      - Output layer with the number of neurons equal to the number of classes
   h. **Compile the model and display the model summary**. You should use *sparse categorical crossentropy* for the loss and use accuracy as the metric
   i. **Fit, evaluate and plot the epoch accuracy**. When fitting the model, use the batch size of 100 and determine the appropriate value for the epoch
   j. **Make prediction, plot the confusion matrix and provide the classification report**

Note on submission:

- Create a folder named as Lab5_ABcXXXXX following the naming convention.
- Put your Jupyter notebook and the original and cleaned dataset in this folder.
- Zip the file and submit it through the blackboard

LAB/ASSIGNMENT PRE-SUBMISSION CHECKLIST

- Did you follow the naming convention for your files and folder?!
- Does your submission work on another computer?!
- Double check **before** submitting