**PCA & LDA Problem Statement:**

**Objective:**

Visualizing high dimensional data using PCA and doing dimensionality reduction to check the explained variance using the PCA model. Training a Linear Discriminant Analysis(LDA) model to check if the product has been shipped or cancelled.

**Problem Statement:**

XYZ.com is an e-commerce company based in Argentina. Due to the covid crisis and lockdown XYZ.com is facing lots of issues from the dealer and the shipment team. XYZ.com has lots of product data where various shipping and sales details of each product have been mentioned. XYZ.com wants to find out which of the products has been shipped and which of the products has been cancelled to reduce customer escalation. As a data-scientist, we have to train a PCA model to visualize its higher-dimensional data and we have to train an LDA(Linear Discriminant Analysis) model to predict which of the products has been shipped and which of the products has been cancelled.

**Dataset Description:**

The dataset link:

https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

Feature Details:

ORDER NUMBER: Order number of the product.

QUANTITY ORDERED: Ordered quantity.

PRICEEACH: Price of each product.

ORDERLINENUMBER: Order line number of the product.

SALES: Sales of the product.

ORDERDATE:Order date of the product.

STATUS:Shipping status(i.e. Shipped or canceled or Resolved) (TARGET) STATE: state where the product needs to be shipped

COUNTRY: Country where the product to be shipped. And so on...

The complete feature details can be found in the above mentioned link.

**Questions**

**● Data Loading and Exploration.**

1. Import necessary libraries.

2. Display a sample of five rows of the data frame.

3. Check the shape of the data (number of rows and columns). Check the general information about the dataframe using the .info() method.

4. Check the percentage of missing values in each column of the data frame.

5. Check if there are any duplicate rows.

6. Write a function that will impute missing values of the columns “STATE”, “POSTALCODE”,“TERRITORY” with its most occurring label.

7. Drop “ADDRESSLINE2”,”ORDERDATE”,”PHONE” column.

8. Convert the labels of the STATUS column to 0 and 1. For Shipped assign value 1 and for all other labels (i.e. ‘Canceled’,’ Resolved’,’ On Hold’,’ In Process’, 'Disputed') assign 0.

Note we will consider everything apart from Shipped as cancel (i.e. 0).

9. Assign ‘STATUS’ column into a label variable and drop it from the original Dataframe. Convert the original dataframe to the dummy coded data.(Hint:-use pd.get\_dummies()).

10. Use StandardScaler to scale the data.

**● PCA FOR VISUALIZATION**

1. Take the help of PCA to reduce the data to 2 dimensions. Use n\_components=2.

2. Take the first and second principal components and plot a scatter plot with the labels.

3. Write the intuitions about the scatter plot.

**● PCA FOR DIMENSION REDUCTION.**

1. Fit the PCA model on the data and plot a graph between n\_components and cumulative explained variance.

2. In how many components we are getting approximately 90% of explained Variance.

● LDA. (5 points)

1. Split the dataset into two parts (i.e. 80% train and 20% test) using random\_state=42.

2. Train a Linear Discriminant Analysis(LDA) model on the train data. Do fit\_transform on the train data and only transform on the test data. Use n\_components=1.

3. Train a RandomForest classifier model on the transformed train and test data. Print the accuracy score.

Please note: You are expected to complete this part by writing queries in MongoDB Query Language (MQL) using VS Code.