MACHINE LEARNING - CS60050
ASSIGNMENT 2 REPORT
NAÏVE BAYES CLASSIFIER

- **GROUP 61:**
  G Sai Chaitanya, 18CS30018.
  G Maneesh Kumar, 18CS10020.

- **Dataset given:** Train_A.csv

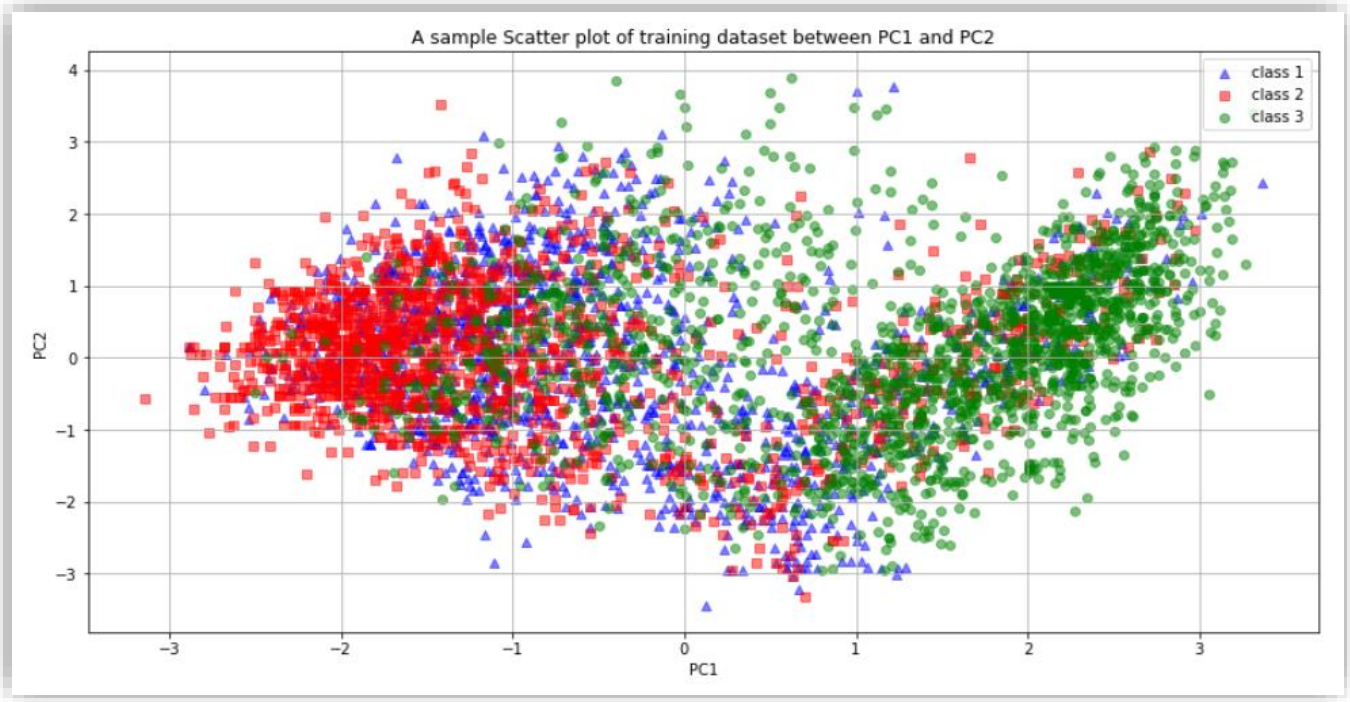
- **Naïve Bayes Classifier Implementation:**
  - The dataset given to us has the following input attributes:
    - **ID**, **Gender, Ever_Married, Age, Graduated, Profession, Work_Experience, Spending_Score, Family_Size, Var_1**
  - The target attribute is 'Segmentation'.
  - We have dropped the column 'ID', as it may not provide any information about classification and it is not considered for training the model.
  - In **Part1**, the dataset is divided into 80% for training, 20% as testing data.
  - The missing values are handled using and inbuilt 'SimpleImputer' from the sklearn library, where string attributes are filled with the most frequent Value and the numerical attributes are filled with mean.
  - The attributes whose values are strings are encoded using 'LabelEncoding' of sklearn library.
  - The Naïve bayes classifier is implemented by splitting dataset according to their class labels and necessary information is evaluated/ summarized from the train data. Log – likelihoods are computed which are used along with class priors to evaluate class probabilities. Then predictions are done on the test fold using the model summarized. The final test accuracy is evaluated on the test data.
  - In **Part 2,** the data is first normalized using 'StandardScaler' from sklearn library. Then the principle component analysis(PCA) is done on train data to get the reduced dimensions/features. These features are further used by the naïve bayes classifier for modelling and 5 fold CV is done and the final test accuracy is evaluated.
  - A sample scatter plot of training dataset is drawn between first two major principle components PC1 and PC2. A heatmap is as well drawn between principle components and original features. A scree plot is as well drawn.
  - In **Part3,** the samples containing outlier feature values are removed from the data. And then the sequential backward selection method is done to obtain the subset of features that are optimal enough to classify.
  - The data from above two steps is again used by naïve bayes classifier, 5 fold CV is done and final test accuracy is determined. The cross-validation accuracy is better after SBS than earlier(part 1).

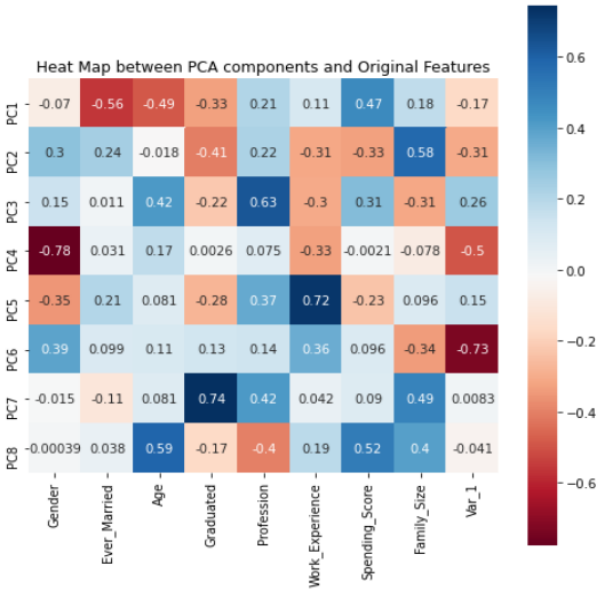
- **RESULTS(RUN ON GOOGLE COLAB):**
  - For Part 1 (Naïve Bayes Classification and 5 fold CV),

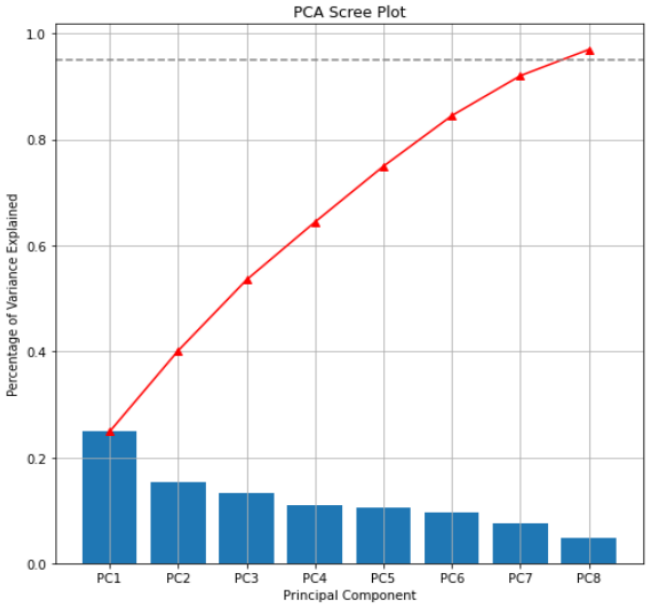
```
#################################################################
           PART 1 (NAIVE_BAYES_CLASSIFICATION)
#################################################################

 => Missing Data Handled......

 => Label Encoding Done......
------------------------------
 Naive Bayes Classification :
------------------------------
 => Accuracy in each fold: [46.9, 46.36, 48.53, 50.23, 47.91]

 => Average Accuracy after 5 fold CV: 47.99

 => Final test accuracy is 46.78
```

  - For Part 2 (PCA and Naïve Bayes on PCA features),

```
#################################################################
           PART 2 (PRINCIPLE COMPONENT ANALYSIS)
#################################################################
---------------
 EIGEN VALUES :
---------------
[2.23819469 1.37291813 1.20416194 0.98153216 0.94712449 0.85951128
 0.67758535 0.44486871]

----------------
 EIGEN VECTORS :
----------------
[[-7.04395952e-02 -5.62239013e-01 -4.87998658e-01 -3.26661951e-01
   2.13460777e-01  1.09542783e-01  4.66342752e-01  1.75143932e-01
  -1.68400568e-01]
 [ 2.96186201e-01  2.41466163e-01 -1.82702782e-02 -4.07578641e-01
   2.22135814e-01 -3.11539787e-01 -3.27112793e-01  5.82827793e-01
  -3.07280490e-01]
 [ 1.48925655e-01  1.06759466e-02  4.18735751e-01 -2.18910211e-01
   6.33665044e-01 -3.02644794e-01  3.11733119e-01 -3.12347800e-01
   2.58034725e-01]
 [-7.76651949e-01  3.12633832e-02  1.74828377e-01  2.59039903e-03
   7.52901598e-02 -3.28165147e-01 -2.11034213e-03 -7.76905349e-02
  -4.95844164e-01]
 [-3.51086979e-01  2.07523204e-01  8.07160823e-02 -2.76070973e-01
   3.74917171e-01  7.23082996e-01 -2.33478218e-01  9.58701213e-02
   1.54359073e-01]
 [ 3.91042556e-01  9.85137092e-02  1.06910061e-01  1.30865509e-01
   1.38020758e-01  3.60141536e-01  9.56988142e-02 -3.37428261e-01
  -7.32842054e-01]
 [-1.47718069e-02 -1.07078505e-01  8.09219297e-02  7.43566805e-01
   4.22620220e-01  4.15242337e-02  8.98478765e-02  4.90308693e-01
   8.28664364e-03]
 [ 3.87167635e-04  3.75936339e-02  5.89176315e-01 -1.70250416e-01
  -3.98061527e-01  1.86615407e-01  5.15088460e-01  4.02773639e-01
  -4.06401249e-02]]
```








```
-------------------------------------------------
 Naive Bayes Classification for PCA features :
-------------------------------------------------
 => Accuracy in each fold: [48.91, 47.98, 49.3, 51.4, 48.45]

 => Average Accuracy after 5 fold CV: 49.21

 => Final test accuracy is 49.5
```

- For part 3(Removing Outliers, Sequential Backward Selection and Subsequent naïve bayes classification)

```
############################################################
            PART 3 (SEQUENTIAL BACKWARD SELECTION)
############################################################

-------------------------------------------
 Number of outliers in the data removed = 146
-------------------------------------------

--------------------------------
 Sequential Backward Selection :
--------------------------------

Feature removed:Work_Experience
['Gender', 'Ever_Married', 'Age', 'Graduated', 'Profession', 'Spending_Score', 'Family_Size', 'Var_1']

Feature removed:Var_1
['Gender', 'Ever_Married', 'Age', 'Graduated', 'Profession', 'Spending_Score', 'Family_Size']

Feature removed:Ever_Married
['Gender', 'Age', 'Graduated', 'Profession', 'Spending_Score', 'Family_Size']

Feature removed:Gender
['Age', 'Graduated', 'Profession', 'Spending_Score', 'Family_Size']

-----------------------------------------------------
 Features obtained after sequential backward selection :
-----------------------------------------------------
['Age', 'Graduated', 'Profession', 'Spending_Score', 'Family_Size']

-------------------------------------------------------------------
 Naive Bayes Classification after Sequential Backward Selection :
-------------------------------------------------------------------
 => Accuracy in each fold: [46.95, 48.37, 48.45, 50.59, 48.3]

 => Average Accuracy after 5 fold CV : 48.53

 => Final test accuracy is 47.77
```