Group 7
Sagnik Roy (18CS10063)
Suryansh Kumar (18CS30043)

Machine Learning Assignment 2

1st November 2020

OVERVIEW

Given a data about patients admitted in a hospital, perform first a Naive Bayes Classifier on the unedited data, then perform PCA (using scikit-learn) and perform Gaussian Naive Bayes Classifier on the formatted data and finally remove outliers and Perform Sequential Backward Selection on the data formatted in 1st part. Perform 5-fold Cross Validation in all 3 parts.

GOALS

1. Split the data into 80:20 Train_Test splits
2. With the data-split obtained in the previous step, perform Naive Bayes Classification
3. Perform Principal Component Analysis on the data obtained in part 1
4. Remove outliers from the data in part 1 and perform sequential backward selection on the data.
5. Perform 5-fold Cross Validation on all three steps

PREREQUISITES

Python Inbuilt Libraries:

The python inbuilt libraries used obtaining, modifying and visualizing data are :

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn
- Math

Procedure

Pre-Processing the data:

The data is first prepared for performing the Naive Bayes Classification :

- **Handle Missing Values:** Missing values are handled by replacing them with the most frequently occurring value using **handle_missing_values()** function.
- **Handle Continuous Attribute:** The continuous attribute 'Admission_Deposit' is handled by the function **createBin()**
- **Encoding Categorical Variables:** The categorical variables are encoded into discrete numbers using the **encode_data()** function.

Naive Bayes Classification:

After the data has been processed, we perform Naive Bayes Classification on the processed data. The encoding part is done after performing Naive Bayes as encoding is not necessary for performing Naive Bayes. The Naive Bayes Classifier utilises the following helper functions:

- train()
- predict()
- getPrediction()
- getAccuracy()

5-Fold Cross Validation:

The dataset is divided into 5 equal parts by the name test_df_1 to test_df_5. The train/test is performed 5 times. Each time, 1 of the 5 parts is used as a test dataset and the remaining dataset is used to train the classifier.

Result of Part 1:

After the 5-fold cross Validation has been performed, the test accuracies were obtained as:

1. Accuracy 1 = 0.34588933551061424
2. Accuracy 2 = 0.3460777540509986
3. Accuracy 3 = 0.3476923076923077
4. Accuracy 4 = 0.34763928381714215
5. Accuracy 5 = 0.3489985313876824

Principal Component Analysis

Principal Component Analysis is done using PCA of sklearn. The number of components is so chosen that 95% of the total variance is retained.

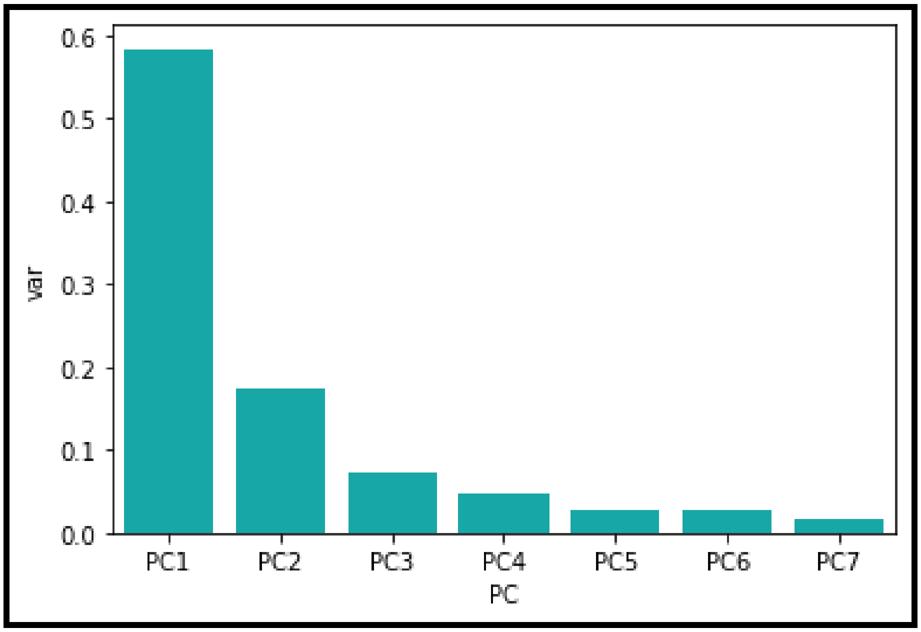
```
if pca.explained_variance_ratio_.sum()>=0.95:
        break
```

A decision taken before PCA:

Before PCA was performed on the dataset, the columns ‘case_id’ and ‘patient_id’ were dropped. This decision was taken keeping in mind the practicality of the dataset as our target attribute is the duration of stay in the hospital, which should not be affected by features like patient_id and case_id. Also, since both are unique for each patient, they are virtually key values and hence will capture a large part of the variance and hence dominate over the more effective features like Severity of Illness, Department, Ward_Type etc.

This was handled during the data preprocessing stage itself.

After performing PCA, it was found that 7 attributes, in all, contribute 95% of the variance. The scree graph was plotted as:



So, we proceed with only these 7 features and perform a Gaussian Naive Bayes Classification (since it is now continuous data):

The class **gaussClf** handled the Gaussian Naive Bayes Classification.

Result of Part 2:

After performing the Gaussian Naive Bayes Classification, 5 fold Cross Validation was performed as explained before on the dataset. The Accuracies obtained were:

1. Accuracy 1 : 0.6178872001004899
2. Accuracy 2 : 0.6183896495415149
3. Accuracy 3 : 0.6177478191911182
4. Accuracy 4 : 0.6165828203605912
5. Accuracy 5 : 0.6203168452957535

We observe that the current accuracy is a drastic improvement on the one performed on raw data using simple Naive Bayes Classifier, implying that the data was grossly overfitted by the attributes that were discarded as a result of Principal Component Analysis.

Removing outliers:

The attributes, like Admission_deposit, having continuous values are used for removing outliers. The entries for which values of maximum of these attributes exceed [mean + 3*standard deviation] are marked as outliers and removed from the dataset

Sequential Backward Selection

The training set is split into Validation set, taking an 80:20 split. Attributes are then removed if their removal increases the accuracy on the validation set. The final set of attributes obtained from sequential backward selection are:

Final Set of Attributes: ['Hospital_code', 'Available Extra Rooms in Hospital', 'Department', 'Ward_Type', 'Bed Grade', 'City_Code_Patient', 'Type of Admission', 'Visitors with Patient', 'Age']

Final Accuracy Calculated after removal of attributes:

Accuracy= 0.38102939329229996

which is an improvement on the approximate accuracy of 0.34 obtained earlier.

The complete assignment has been submitted in 2 parts: one containing qn 1 and 2 and the other containing qn 3. Due to the huge size of the dataset(>300000 entries) the computations take a lot of time, especially due to 5-fold cross validation. So to ensure that the 2 part can be computed parallely on the local machines.