Report

This is a multi-class classification problem. The goal is to predict the length of stay for a patient in the hospital based on the provided attributes. The length of stay is divided into several classes (0-10, 11-20, More than 100 days, etc.). The attributes include information about the hospital (e.g. Hospital_code, Hospital_type_code, City_Code_Hospital, etc.), the patient (e.g. Age, Type of Admission, Severity of Illness, etc.), and the ward (e.g. Department, Ward_Type, Bed Grade, etc.).

Task1

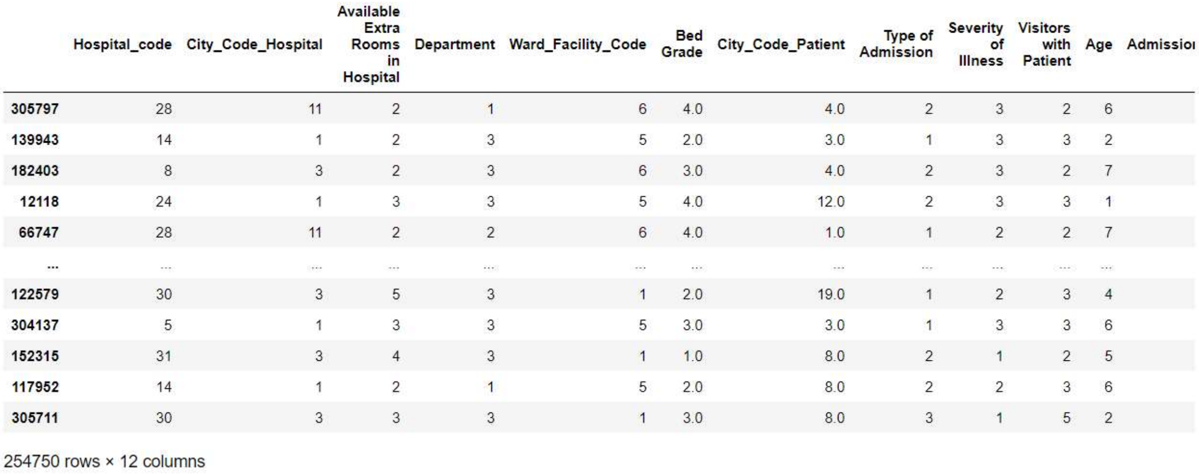
1. Import the libraries.
2. Import the data file hospital.csv. Segregate the data into X and y. Randomly divide the data into 80% for training and 20% for testing.
3. Finding the missing value with inappropriate value for each variable and assigning them to NaN like strings in int variable and vice versa. Here, if we replace with mean or median for bed Grade which is fractional, so avoided instead we use mode (most frequently occurring value).
4. Encode the categorical variables.
5. Make categorical variables.
6. Dropped the 'case_id' attribute because it's unique.
7. Naive Bayes Algorithm:
    a. Compute prior probabilities of each class.
    b. Compute the likelihood probabilities for each feature given in each class.
    c. Using the Bayes theorem, compute the posterior probabilities of each class given each feature.
    d. Predict the class with the highest posterior probability.
8. Cross-validation: 5-fold cross-validation can be performed by dividing the training data into 5 equal parts. Then use 4 parts for training the model and 1 part for validation. Repeat this process 5 times with a different part for validation each time.

```
9. Average accuracy with cross validation:  0.283269872423945
10.  Final Test Accuracy:  0.2840566511744756
11.  Final Test Accuracy with Gaussian NB:  0.3275185278231378
   Final Test Accuracy with Multinomial NB:  0.2525750533852531
```

Task2

1. Import the libraries.
2. Import the data file hospital.csv. Segregate the data into X and y. Randomly divide the data into 80% for training and 20% for testing.
3. Finding the missing value with inappropriate value for each variable and assigning them to NaN like strings in int variable and vice versa. Here, if we replace with mean or median for bed Grade which is fractional, so avoided instead we use mode (most frequently occurring value).
4. Encode the categorical variables.
5. Make categorical variables.

6. Dropped the unnecessary attributes.
7. Output of X_train data:

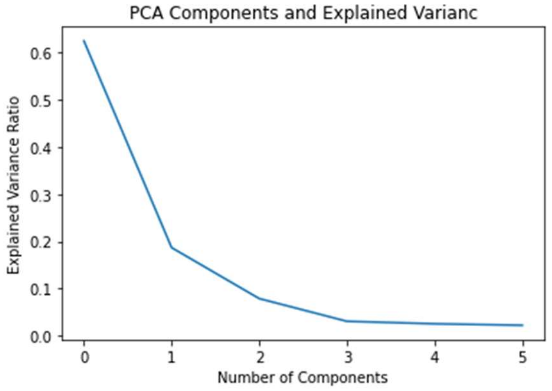


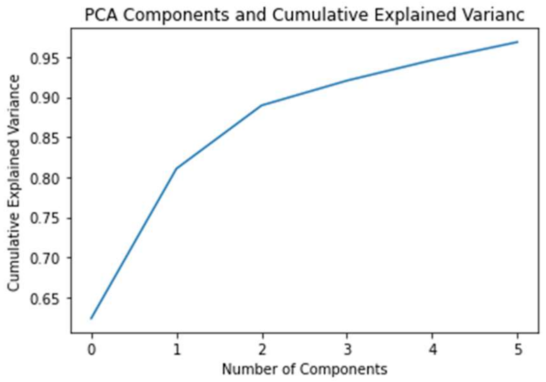
	Hospital_code	City_Code_Hospital	Available Extra Rooms in Hospital	Department	Ward_Facility_Code	Bed Grade	City_Code_Patient	Type of Admission	Severity of Illness	Visitors with Patient	Age	Admissio
305797	28	11	2	1	6	4.0	4.0	2	3	2	6	
139943	14	1	2	3	5	2.0	3.0	1	3	3	2	
182403	8	3	2	3	6	3.0	4.0	2	3	2	7	
12118	24	1	3	3	5	4.0	12.0	2	3	3	1	
66747	28	11	2	2	6	4.0	1.0	1	2	2	7	
...	...	...	...	...	...	...	...	...	...	...	...	
122579	30	3	5	3	1	2.0	19.0	1	2	3	4	
304137	5	1	3	3	5	3.0	3.0	1	3	3	6	
152315	31	3	4	3	1	1.0	8.0	2	1	2	5	
117952	14	1	2	1	5	2.0	8.0	2	2	3	6	
305711	30	3	3	3	1	3.0	8.0	3	1	5	2	

254750 rows × 12 columns

8. Naive Bayes Algorithm:
    a. Compute prior probabilities of each class.
    b. Compute the likelihood probabilities for each feature given in each class.
    c. Using the Bayes theorem, compute the posterior probabilities of each class given each feature.
    d. Predict the class with the highest posterior probability.
9. Cross-validation: 5-fold cross-validation can be performed by dividing the training data into 5 equal parts. Then use 4 parts for training the model and 1 part for validation. Repeat this process 5 times with a different part for validation each time.
10. PCA can be performed to reduce the number of features in the data. Select the number of components that preserve 95% of the total variance. We have plotted the graph of the results of PCA to visualize the explained variance by each component and the cumulative explained variance.

Explained_variance_ratio:
```
[0.62397804 0.18673151 0.0789282  0.03087592 0.025617
77 0.02249491]
```





```
Average accuracy with cross validation:  0.3633523061825319
   11.  Test set accuracy score is 0.363899007662354
```

Task3

1. Import the libraries.
2. Import the data file hospital.csv. Segregate the data into X and y. Randomly divide the data into 80% for training and 20% for testing.
3. Finding the missing value with inappropriate value for each variable and assigning them to NaN like strings in int variable and vice versa. Here, if we replace with mean or median for bed Grade which is fractional, so avoided instead we use mode (most frequently occurring value).
4. Encode the categorical variables.
5. For outlier detection, identify outliers by finding features that have values greater than mean + 3 * standard deviation.
6. Make categorical variable.
7. Dropped the ‘case_id' attribute because it's unique.
8. Naive Bayes Algorithm:
    a. Compute prior probabilities of each class.
    b. Compute the likelihood probabilities for each feature given in each class.
    c. Using the Bayes theorem, compute the posterior probabilities of each class given each feature.
    d. Predict the class with the highest posterior probability.
9. Cross-validation: 5-fold cross-validation can be performed by dividing the training data into 5 equal parts. Then use 4 parts for training the model and 1 part for validation. Repeat this process 5 times with a different part for validation each time.

```
10. Average accuracy with cross-validation:  0.280909375247389
11. Final Test Accuracy:  0.28012383698073645
```

Extra Tasks

1. Import the libraries.
2. Segregate the data into X and y.
3. Encode the categorical variables.
4. Finding the missing value with inappropriate value for each variable and assigning them to NaN like strings in int variable and vice versa. Here, if we replace with mean or median for bed Grade which is fractional, so avoided instead we use mode (most frequently occurring value).
5. Dropped the unnecessary attributes.
6. Naive Bayes Algorithm:
    a. Compute prior probabilities of each class.
    b. Compute the likelihood probabilities for each feature given in each class.
    c. Using the Bayes theorem, compute the posterior probabilities of each class given each feature.
    d. Predict the class with the highest posterior probability.
7. Cross-validation: 5-fold cross-validation can be performed by dividing the training data into 5 equal parts. Then use 4 parts for training the model and 1 part for validation. Repeat this process 5 times with a different part for validation each time.

```
   8. Average accuracy with cross-validation:  0.3637348495832172


   9. Classification Report
              precision    recall  f1-score   support

           1       0.29      0.08      0.13      4689
           2       0.38      0.47      0.42     15561
           3       0.40      0.63      0.49     17603
           4       0.25      0.11      0.15     10981
           5       0.09      0.02      0.04      2357
           6       0.35      0.40      0.37      7128
           7       0.00      0.00      0.00       554
           8       0.29      0.00      0.01      2031
           9       0.16      0.04      0.06       941
          10       0.50      0.00      0.01       552
          11       0.34      0.37      0.35      1291

    accuracy                           0.37     63688
   macro avg       0.28      0.19      0.18     63688
weighted avg       0.33      0.37      0.32     63688
```