# DANA4830 – Fall 2021

Submitted by:

Mohamed Ghayaas Anjum and Umme Salma

**Summary statistics for the predictors respective to PEX and traditional treatment:**

| Variable | Group 1 : Patient with or without Pex = 0 | | Patient with or without Pex = 1 | | |
|---|---|---|---|---|---|
| | Mean | Standard Deviation | Mean | Standard Deviation | t-test results |
| Age | 41.91 | 9.57 | 40.27 | 8.98 | t = 1.238, pvalue = 0.2176 |
| Duration of Staying in hospitals | 6.48 | 2.89 | 6.96 | 3.35 | t = -0.88443, pvalue = 0.3779 |
| Subclinical Examination – wbc | 9.66 | 3.07 | 11.36 | 4.49 | t = -2.852, pvalue = 0.005025 |
| Prothrombin | 86.02 | 17.93 | 105.05 | 108.94 | t = -1.4849, pvalue = 0.1417 |
| Triglycerid | 21.54 | 13.84 | 33.56 | 28.51 | t = -3.3386, df = 112.74, p-value = 0.001142 |

Here, the entire data is divided into two groups based on treatment method. Out of 165, 83 patients are treated using pex treatment and 82 are treated with traditional method as per Vietnam's ministry of health guidelines. Summary statistics describe the mean and standard deviation of numerical data for both pex and traditional treatments as per Vietnam health Ministry guidelines.

We can observe from the table that the mean is greater for pex = 1 for variables like Duration of Staying in hospitals, Subclinical Examination – wbc, Prothrombin and Triglycerid.

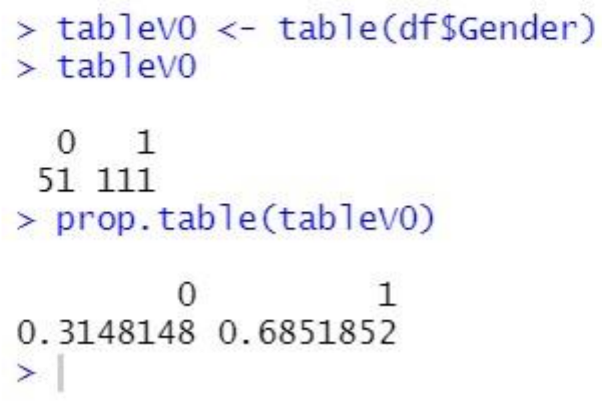
When we look at the standard deviation values for pex = 1, we find that Duration of Staying in hospitals, Subclinical Examination – wbc, Prothrombin and Triglycerid have greater standard deviation values when compared to pex = 0

We can observe a significant difference in mean for predictors like  Subclinical Examination – wbc, Prothrombin and Triglycerid for two groups of patients where one group belongs to pex and the other belongs to traditional treatment.

When we look at the standard deviation for two groups of patients, we observe that prothrombin and triglyceride show a significant increase in standard deviation value.

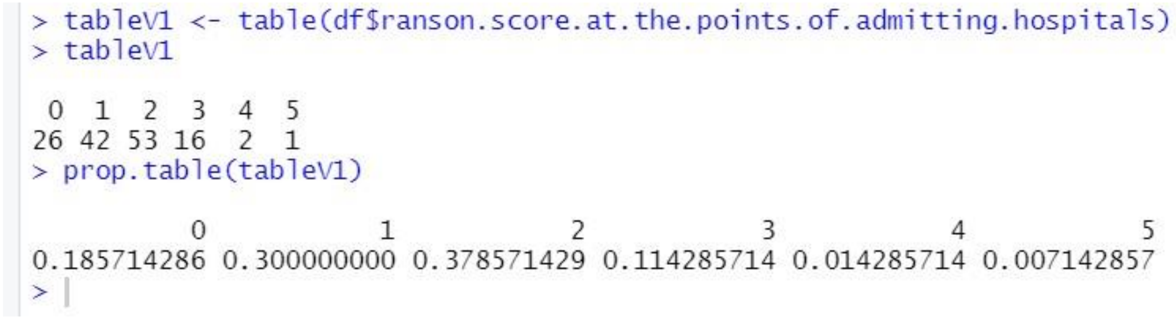
Categorical Predictors:

The frequency of variables for each category and its percentages are calculated using R



```
> tableV0 <- table(df$Gender)
> tableV0

  0   1 
 51 111 
> prop.table(tableV0)

        0         1 
0.3148148 0.6851852 
> 
```

There are 51 patients who are male and 111 who are female



```
> tableV1 <- table(df$ranson.score.at.the.points.of.admitting.hospitals)
> tableV1

 0  1  2  3  4  5 
26 42 53 16  2  1 
> prop.table(tableV1)

          0           1           2           3           4           5 
0.185714286 0.300000000 0.378571429 0.114285714 0.014285714 0.007142857 
> 
```

Out of all the patients, 26 have ranson score 0, 42 have ranson score 1, 53 have score 2, 16 have ranson score of 3, 2 have ranson score of 4 and only 1 patient has ranson score 5.

The frequency calculation for other variables like Apache 2 score at the point of admitting to hospitals, CTSI Score at the point of admitting to hospitals and Imre score at the point of admitting to hospital have been tabulated as follows:

| Variable | Frequency | Percentage |
|---|---|---|
| Gender(0) | 51 | 31.48 |
| Gender(1) | 111 | 68.52 |
| Ranson Score (0) | 26 | 18.57 |
| Ranson Score (1) | 42 | 30 |

| | | |
|---|---|---|
| Ranson Score (2) | 53 | 37.85 |
| Ranson Score (3) | 16 | 11.42 |
| Ranson Score (4) | 2 | 1.42 |
| Ranson Score (5) | 1 | 0.71 |
| Apache 2 score at the point of admitting to hospitals (0) | 33 | 23.57 |
| Apache 2 score at the point of admitting to hospitals (1) | 9 | 6.42 |
| Apache 2 score at the point of admitting to hospitals (2) | 22 | 15.71 |
| Apache 2 score at the point of admitting to hospitals (3) | 12 | 8.57 |
| Apache 2 score at the point of admitting to hospitals (4) | 21 | 15 |
| Apache 2 score at the point of admitting to hospitals (5) | 5 | 3.57 |
| Apache 2 score at the point of admitting to hospitals (6) | 13 | 9.28 |

| | | |
|---|---|---|
| Apache 2 score at the point of admitting to hospitals (7) | 11 | 7.85 |
| Apache 2 score at the point of admitting to hospitals (8) | 6 | 4.28 |
| Apache 2 score at the point of admitting to hospitals (9) | 4 | 2.85 |
| Apache 2 score at the point of admitting to hospitals (10) | 1 | 0.71 |
| Apache 2 score at the point of admitting to hospitals (11) | 1 | 0.71 |
| Apache 2 score at the point of admitting to hospitals (12) | 1 | 0.71 |
| Apache 2 score at the point of admitting to hospitals (16) | 1 | 0.71 |
| CTSI Score at the point of admitting to hospitals(0) | 3 | 2.65 |
| CTSI Score at the point of admitting to hospitals(1) | 2 | 1.76 |

| | | |
|---|---|---|
| CTSI Score at the point of admitting to hospitals(2) | 14 | 12.38 |
| CTSI Score at the point of admitting to hospitals(3) | 32 | 28.31 |
| CTSI Score at the point of admitting to hospitals(4) | 43 | 38.05 |
| CTSI Score at the point of admitting to hospitals(5) | 4 | 3.53 |
| CTSI Score at the point of admitting to hospitals(6) | 9 | 7.96 |
| CTSI Score at the point of admitting to hospitals(8) | 5 | 4.42 |
| CTSI Score at the point of admitting to hospitals(10) | 1 | 0.88 |
| Imre score at the point of admitting to hospital(0) | 30 | 21.27 |
| Imre score at the point of admitting to hospital(1) | 43 | 30.49 |
| Imre score at the point of admitting to hospital(2) | 55 | 39 |
| Imre score at the point of admitting to hospital(3) | 11 | 7.8 |

|  |  |  |
|---|---|---|
| Imre score at the point of admitting to hospital(4) | 2 | 1.41 |

**Linear Discriminant Analysis:**

i. What is the accuracy rate?

Accuracy rate is 100 percent

ii. Is one type of misclassification more likely than the other?

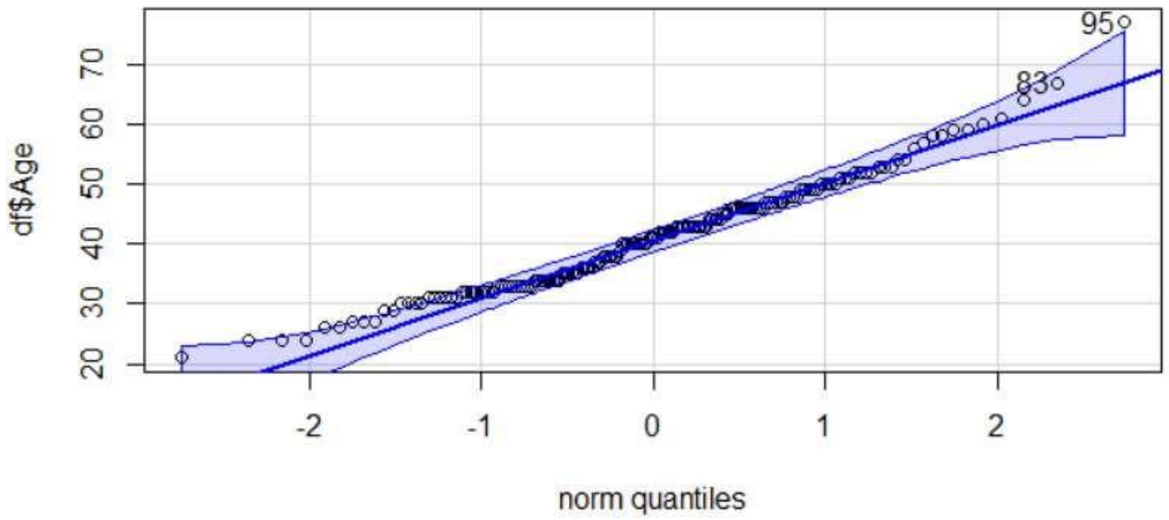
Yes, one type of misclassification is more likely than the other because out of 130 features, we selected only 19 features. Hence there are greater chances of misclassification in one group when compared to the other.

iii. Select two to three patients who were misclassified as receiving the PEX treatment and two to three who were misclassified as receiving the traditional treatment. The goal is to determine why they are misclassified.

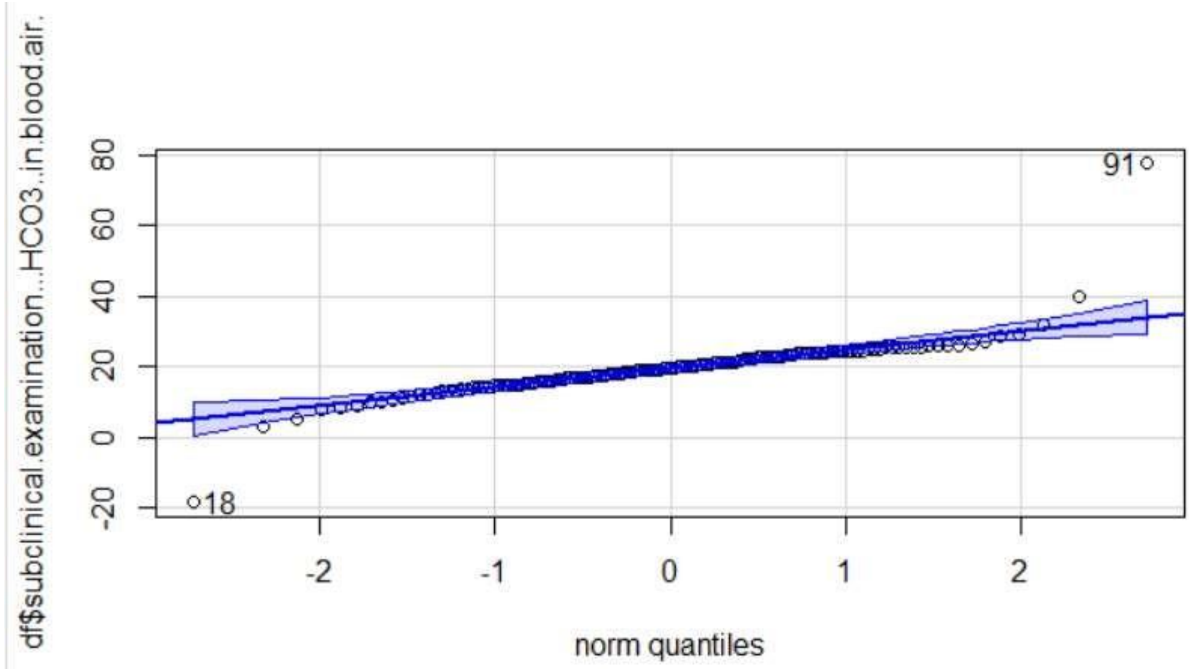
Misclassification occurs when patient is grouped under wrong category. In this present scenario, patient took pex treatment but has been grouped under traditional treatment. Similarly, patient who took the traditional treatment might have been incorrectly grouped under pex treatment. The reason behind this must be measurement or observational error.

Assumptions:

- Multivariate normality: Scores on predictors are independently and randomly sampled from a population, and that the sampling distribution of any linear combination of predictors is normally distributed.



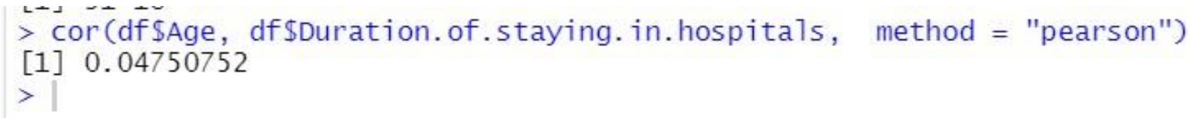
In AP Dataset, when we check if the variables follow normal distribution using qqplot, we can observe that not every point in dataset falls on the reference diagonal line, there are some points which are above the line. Hence we can say that this assumption of multivariate normality is not met.



We can observe from qqplot that the subclinical examination for HCO3 in blood has few data points that do not follow the diagonal reference line. Hence we can conclude that assumption of multivariate normality is not met.

- A lack of multicollinearity.

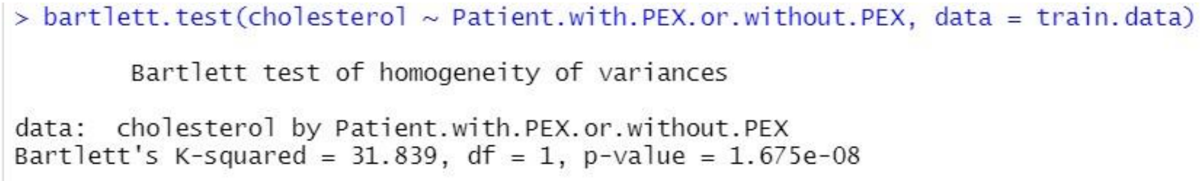
There exists correlation between predictor variables in AP dataset for age and duration of stay in hospitals.



```
> cor(df$Age, df$Duration.of.staying.in.hospitals,  method = "pearson")
[1] 0.04750752
> |
```

Hence we cannot see the lack of multicollinearity. This assumption is not met.

- Equal dispersion matrices: Box's M test is used to examine equal variance dispersion to assess the equality of variances within groups formed by nonmetric variables or bartlette's test can be used to check if samples have equal variances.



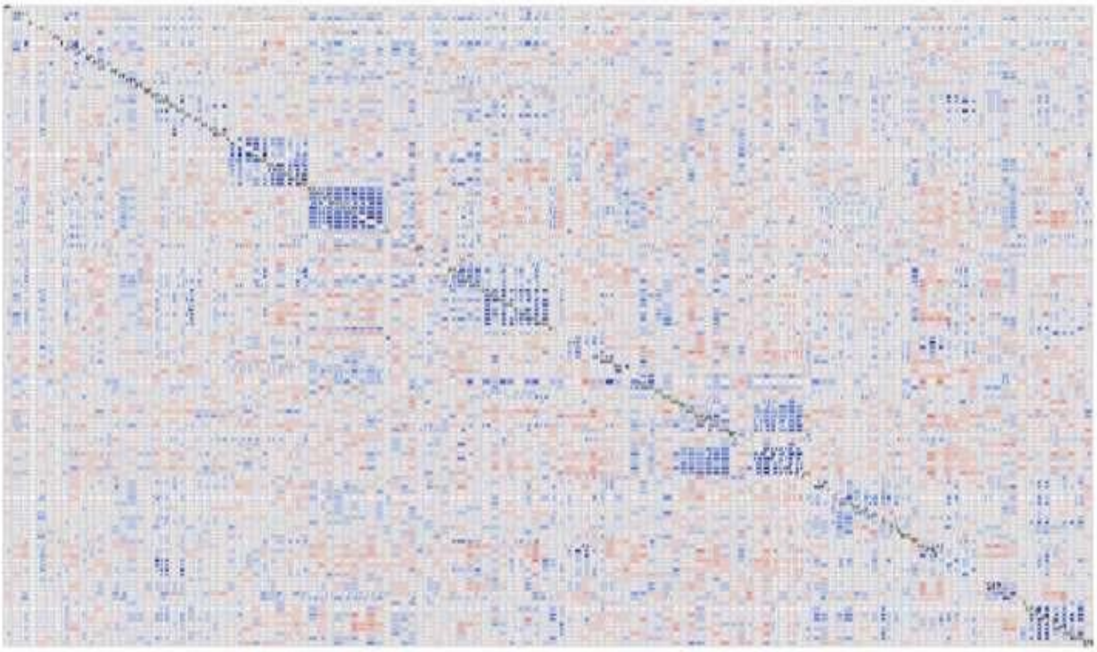
```
> bartlett.test(cholesterol ~ Patient.with.PEX.or.without.PEX, data = train.data)

        Bartlett test of homogeneity of variances

data:  cholesterol by Patient.with.PEX.or.without.PEX
Bartlett's K-squared = 31.839, df = 1, p-value = 1.675e-08
```

Null Hypothesis $H_0$: Variances are equal across all samples in dataset

Alternate Hypothesis $H_a$: Variances are not equal across all samples in dataset

Here p value less than 0.05, hence we can reject null hypothesis. Hence variances are not equal across all samples in dataset.

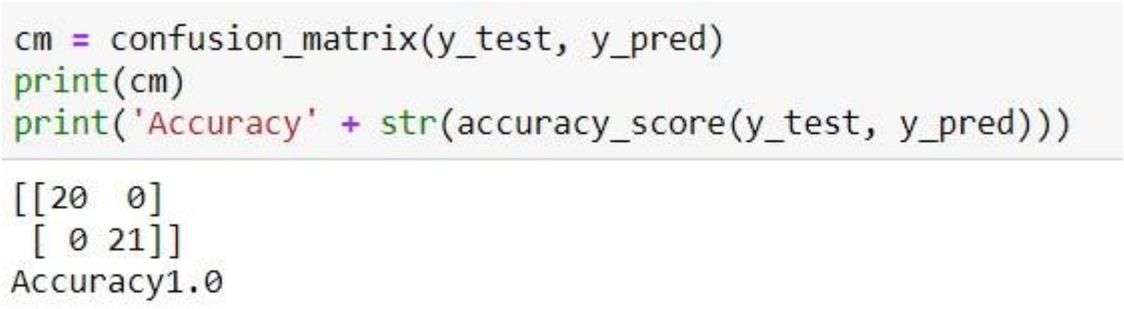
- Linear relationships: discriminant analysis assumes linear relations among the independent variables.



Corgram plot is used to plot the correlation between independent variables. Here, positive correlations are represented by blue and negative by red. From corgram plot, we can observe that there exists many positive and negative linear relationships among the predictor variables. Hence this assumption of linear relationships is met.

**Perform K-nearest Neighbour (K-NN) analysis and comparison with LDA:**

**Confusion matrix obtained from Linear Discriminant analysis:**



```
cm = confusion_matrix(y_test, y_pred)
print(cm)
print('Accuracy' + str(accuracy_score(y_test, y_pred)))
```

```
[[20  0]
 [ 0 21]]
Accuracy1.0
```

y_test : represents the actual value of y y_pred: Predicted

value of y as returned by knn algorithm.

True positive = 20

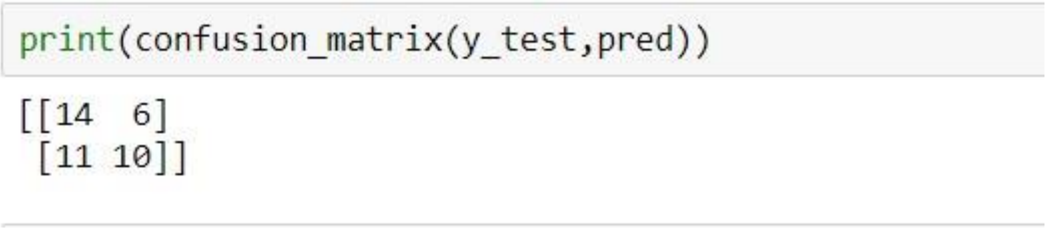
False positive = 0

False Negative = 0

True Negative = 21

We can observe from the matrix that there are zero misclassifications and hence the model is 100 % accurate.

**Confusion matrix obtained from knn for missforest imputed data:**



```
print(confusion_matrix(y_test,pred))
```

```
[[14  6]
 [11 10]]
```

y_test : represents the actual value of y

pred: Predicted value of y as returned by knn algorithm.

True positive = 14

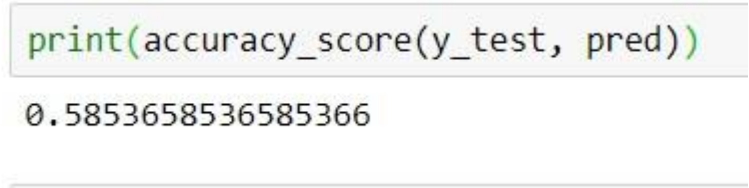
False positive = 6

True Negative = 10

False Negative = 11

11 patients belong to pex = 0 but were misclassified as pex = 1 by knn.

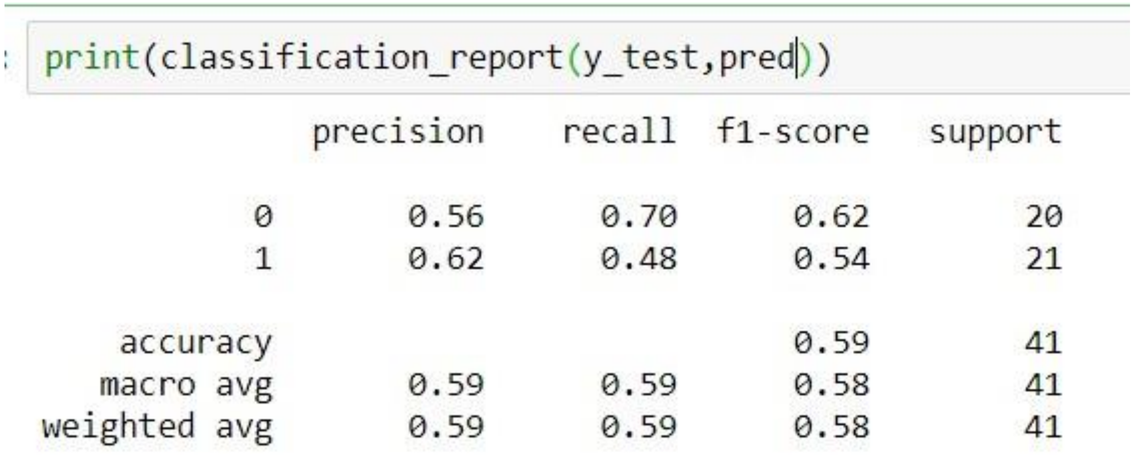
6 patients that belong to pex = 1 group were misclassified as pex = 0 by knn.



```
print(accuracy_score(y_test, pred))
```

```
0.5853658536585366
```

We can observe that the accuracy of knn model for missforest imputed dataset is 58.53 percent



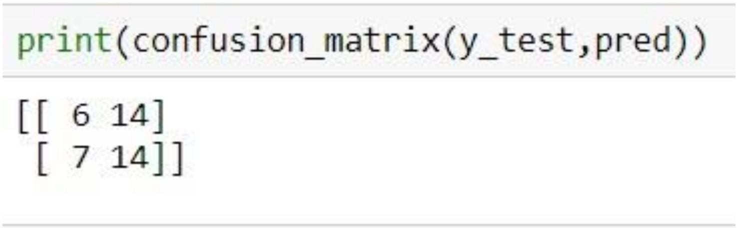
```
print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

           0       0.56      0.70      0.62        20
           1       0.62      0.48      0.54        21

    accuracy                           0.59        41
   macro avg       0.59      0.59      0.58        41
weighted avg       0.59      0.59      0.58        41
```

From the classification report, we can observe that the precision for pex = 0 is 0.56 and for pex = 1 is 0.62

Similarly, the recall for pex = 0 is 0.7 and for pex = 1 is 0.48

**Confusion matrix obtained from knn for amelia imputed data:**



```
print(confusion_matrix(y_test,pred))
```

```
[[ 6 14]
 [ 7 14]]
```

True positive = 6

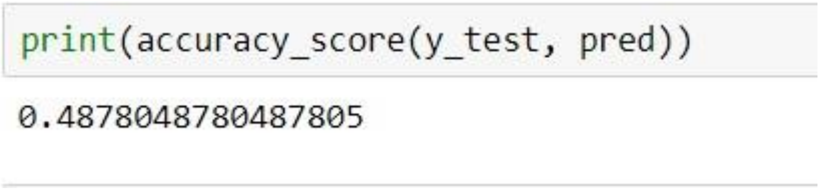
False positive = 14

True Negative = 14

False Negative = 7

7 patients belong to pex = 0 but were misclassified as pex = 1 by knn.

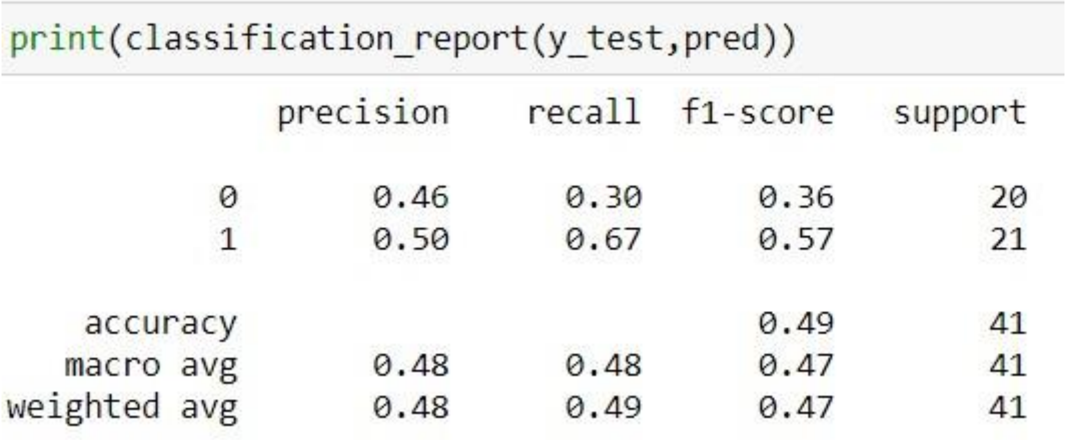
14 patients that belong to pex = 1 group were misclassified as pex = 0 by knn.



```
print(accuracy_score(y_test, pred))
```

```
0.4878048780487805
```

We can observe that the accuracy of knn model for amelia imputed dataset is 48.78 percent



```
print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

           0       0.46      0.30      0.36        20
           1       0.50      0.67      0.57        21

    accuracy                           0.49        41
   macro avg       0.48      0.48      0.47        41
weighted avg       0.48      0.49      0.47        41
```

From the classification report, we can observe that the precision for pex = 0 is 0.46 and for pex = 1 is 0.50

When we compare the two datasets, missforest and amelia, the precision is greater for missforest for both pex = 0 and pex = 1 groups. However, recall score shows a different trend. For pex = 0, precision is reduced in case of amelia when compared to missforest.

Similarly, the recall for pex = 0 is 0.3 and for pex = 1 is 0.67

But, for pex = 1, recall score has increased in amelia when compared to missforest.

Hence the accuracy of knn model using missforest imputed dataset is greater than accuracy of amelia imputed dataset.