Linear Discriminant Analysis

Statistical Prediction Modeling ***DATA 2204***

Assignment No3

Due Date : ,10th March 2022

Jumana Obeid

100832624

DURHAM COLLEGE
SUCCESS MATTERS

b. Rational Statement (summary of the problem or problems to be addressed by the PPT) – 2%

1) Limitations of logistic regression, why we tried LDA :

• Unstable with Well Separated Classes. Logistic regression can become unstable when the classes are well separated.

• Unstable with Few Examples. Logistic regression can become unstable when there are few examples from which to estimate the parameters.

The LDA algorithm is going to compute the probability of being classified as Benign (No Cancer) or Malignant (Cancer) , Discriminant score will separate the two classes as mentioned before to Benign and Malignant. Having the below features in the dataset :

The independent variables:

1. Clump Thickness – 1-10
2. UofCSize - Uniformity of Cell Size 1-10
3. UofShape - Uniformity of Cell Shape 1-10
4. Marginal Adhesion - 1-10
5. SECSize - Single Epithelial Cell Size 1-10
6. Bare Nuclei - 1-10
7. Bland Chromatin - 1-10
8. Normal Nucleoli - 1-10
9. Mitoses - 1-10

Dependent Variable Class –

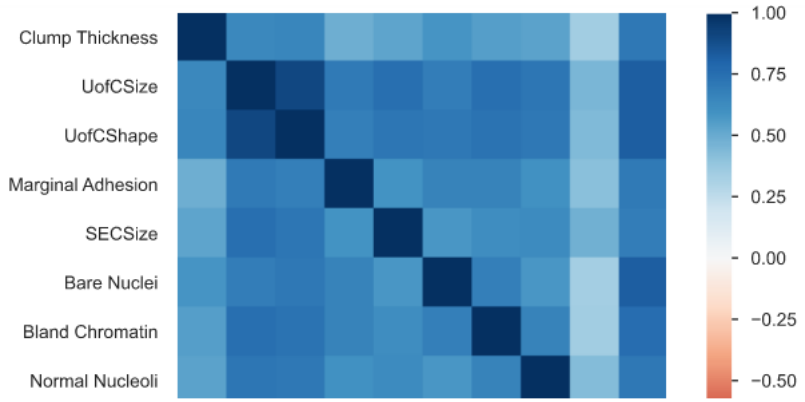
1. Benign (i.e. No Cancer) - 2,
2. Malignant (i.e. Cancer) - 4

DURHAM COLLEGE
SUCCESS MATTERS

c.Identify and explain two (2) key insights from the Pandas Profile Report – 2%

There is high correlation between UofSize and UofCShape and we can exclude UofCShape from our test , ensuring that the dataset features are independent, In other words ensuring the accuracy of our test results.

The mean of Clump Thickness is 4.442 in the data
Set and the maximum is 10, minimum is 1.
The data could be skewed a little and having more data
Could be beneficial .



DURHAM COLLEGE
SUCCESS MATTERS

d. Present and explain three (3) key insights from the Optimized LDA classification report, but first use SMOTE to ensure that the dataset is balanced. – 6%

1)Using SMOTE

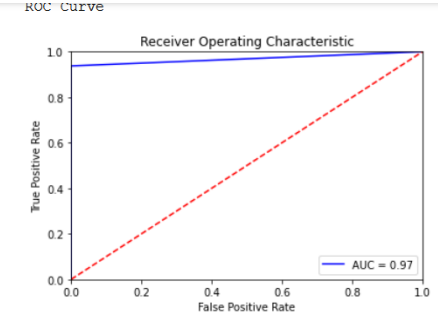
In [12]:

```
#Create x and y variables
x = dataset.drop('Class', axis=1).to_numpy()
Y = dataset['Class'].to_numpy()

#Create Train and Test Dataset
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,Y,test_size = 0.2,stratify=Y,random_state = 100)

#Fix the imbalanced Classes
from imblearn.over_sampling import SMOTE
smt=SMOTE(random_state=100)
x_train_smt,y_train_smt = smt.fit_resample(x_train,y_train)
```

2) The F1 score is 98% , the model is optimal and the Area under the graph is approximately the largest , the model could be overfitting.



3)Precision(out of all predictions we got 98% right) and Recall (out of all Correct Predictions the truths, we got 98% right).

4) ROC , The ROC curve shows **the trade-off between sensitivity (or TPR) and specificity (1 – FPR)**. **tells us** Classifiers that give curves closer to the top-left corner indicate a better performance.

DURHAM COLLEGE
SUCCESS MATTERS

e. Compare the Optimized LDA to the Optimized Logistical Regression model (from page 1) identifying three (3) key insights. – 3%

1)The performance of the two models approximately the same , I believe that this is because we have only two classes , where logistic regression performs well.Both score 98% in F1 score.

2)AUC-(Area under the curve , the bigger the better) and ORC -shows **the trade-off between sensitivity (or TPR) and specificity (1 – FPR).for both models indicates that these models performs very well. There could be an overfitting .**

3)Precision and Recall tells us that both models perform very well.
Precision(out of all predictions we got 98% right) and Recall (out of all Correct Predictions the truths, we got 98% right).

f. State and explain two (2) recommendations for Mr. John Hughes for next steps. – 2%

1-Gather more data for better accuracy .
2-Use important features(from independent variables) only.

DURHAM COLLEGE
SUCCESS MATTERS