Statistical and Predictive Modeling II (DATA 2204)
Assignment #3 – Discriminant Analysis (15% of Final Grade)
Professor: Sam Plati

Mr. John Hughes is looking at developing an LDA model for his **heartfailure.csv** dataset and evaluate its effectiveness. If you recall the dataset has the following variables.

Independent Variables

age: age of the patient (years)
anaemia: decrease of red blood cells or hemoglobin (boolean)
high blood pressure: if the patient has hypertension (boolean)
creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
diabetes: if the patient has diabetes (boolean)
ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
platelets: platelets in the blood (kiloplatelets/mL)
sex: woman or man (binary)
serum creatinine: level of serum creatinine in the blood (mg/dL)
serum sodium: level of serum sodium in the blood (mEq/L)
smoking: if the patient smokes or not (boolean)
time: follow-up period (days)

Dependent Variable

death event: if the patient deceased during the follow-up period (0-Alive, 1-Deceased)

Below are the results of the Optimized Logistical Regression model (with SMOTE):

Note: Outcome 0: Alive, Outcome 1: Deceased

```
Optimized Model

Model Name: LogisticRegression(class_weight='balanced', random_state=100)

Best Parameters: {'clf__C': 0.01, 'clf__penalty': 'l2'}

[[36  5]
[ 5 14]]

precision        recall   f1-score   support

Outcome 0        0.88       0.88       0.88        41
Outcome 1        0.74       0.74       0.74        19

accuracy                               0.83        60
macro avg        0.81       0.81       0.81        60
weighted avg     0.83       0.83       0.83        60
```

The Ask:

1. Create a PowerPoint (PPT) presentation that includes the following:
   a. Cover Page (Title, Name (1st and last) and Student Number)
   b. Analysis Statement (i.e. LDA vs. Logistical Regression Model) ***– 2%***
   c. Identify and explain **three (3)** key insights from the Pandas Profile Report ***– 3%***
   d. Present and explain the entire Classification Report for **both the Standard and Optimized LDA model,** but first use **SMOTE** to ensure that the dataset is balanced. ***– 5%***
   e. How does the **Optimized LDA Model** results (i.e. Classification Report) compare to the Optimized Logistical Regression Results (Classification Report)? Please provide **three (3)** key insights. ***– 3%***
   f. State and explain **two (2) recommendations** for Mr. John Hughes for next steps. ***– 2%***

Attention: Please ensure that all key facts are in your slides and not in the notes section

Hint: Leverage the code from Wk5a-LDAQDA

Random State = 100 for all section

2. Provide a copy of your HTML Python Code

Please post your PowerPoint Document (.ppt) and HTML of Python Code via assignments under Assignment #3 by 11:59 p.m. on Friday, June 18th, 2021

Grading Rubric

	Needs Improvement	**Average**	**Above Average**	**Comments**
1. Create a PowerPoint (PPT) presentation that includes the following: a. Cover Page (Title, Name (1st and last) and Student Number) b. Analysis Statement (i.e. LDA vs. Logistical Regression Model) ***– 2%*** c. Identify and explain **three (3)** key insights from the Pandas Profile Report ***– 3%*** d. Present and explain the entire Classification Report for **both the Standard and Optimized LDA model** , but first use **SMOTE** to ensure that the dataset is balanced. ***- 5%*** e. How does the Optimized LDA Model results compare to the Optimized Logistical Regression Results? Please provide **three (3) key insights**. ***– 3%*** f. What **two (2)** recommendations do you have for Mr. John Hughes for next steps? Please explain your answers. ***– 2%*** 2. HTML Copy of your Python Code				
Needs Improvement –Missing the minimum requirements stated in the assignment requirements. **Average –Meets the minimum requirements stated in the assignment requirements.** **Above Average –Exceeds the requirements that are stated in the assignment requirements.**				