Introduction to Data Analysis (DATA 1200)
Final Project (25% of Final Grade)
Professor: Ritwick Dutta

For your final project you will using the **Happy_Dataset.csv** dataset to help Mr. John Hughes with a problem. Let's look at the data first:

The dataset has 143 observations and 7 variables:

Independent Variables

X1 = the availability of information about the city services
X2 = the cost of housing
X3 = the overall quality of public schools
X4 = your trust in the local police
X5 = the maintenance of streets and sidewalks
X6 = the availability of social community events

Attributes X1 to X6 have values 1 to 5.

Dependent Variable

D = decision attribute (D) with values 0 (unhappy) and 1 (happy)

The Ask:

Mr. John Hughes is looking to identify the best model to help predict a student's change in standardized score after training.

He would like you to create forecast models using the following algorithms:

- **Naïve Bayes**
- **Logistical Regression**
- **Neural Network***

Attach a separate HTML copy of your Python Code with your submission.

***Please use the following assumptions in your Neural Network Model:**

**MLPClassifier(hidden_layer_sizes=(6,3,2),
activation='relu',solver='adam',
max_iter=10000,random_state=100)**

In addition, Mr. John Hughes would also like you to create a correlation heatmap and pair plot with **two (2)** key insights for each visualization. Finally, determine which is the best model to use for the **Happy_Dataset.csv** and identify **two (2)** recommendations to improve the model you have chosen.

What is required?

Mr. John Hughes would like a PowerPoint Presentation (see below for details) and HTML copy of the Jupyter Notebook used to solve the problem.

Please post your PowerPoint Document (.ppt or .pptx) and Jupyter Notebook in HTML (.html) format via assignments under Final Project by Friday, December 15th, 2023

Note: 50% Grade Penalty for missing Jupyter Notebook HTML file

PowerPoint Detail Requirements (Number of Slides is a Guideline)

Cover Slide

- Name (First and Last)
- Student Number
- Title: Final Project – DATA 1200

Slide 1 ***(1%)***

- Rational Statement (summary of the problem(s) to be addressed by the PPT)

Slides 2-3 ***(2%)***

- Present the correlation heatmap and explain **two (2)** key insights.

Slides 4-8 ***(18%)***

- Present the Confusion Matrix/Classification Report for each of the ***three (3) algorithms (i.e. Naïve Bayes, Logistical Regression, and Neural Network)***.
- Explain and justify **three (3) key insights** gained from the Classification Report for each algorithm (i.e., Precision, Recall, F1, Support for both summary and detailed metrics). ***Note a total of nine (9) key insights are required.***

Slide 9-10 ***(4%)***

- Recommend **one (1) model from the analysis** that should be utilized. Please justify your answer.
- State and justify **two (2)** possible improvements that can be made to increase the usability of the model you have chosen.

Code Requirements

Python Script using Jupyter Notebook (then convert to .html). **Note: 50% Grade Penalty for missing Jupyter Notebook HTML file.**

Note: 50% Grade Penalty for missing Jupyter Notebook HTML file