DAV 5400 Project 4 (Week 13) (100 Points)

****You may work in small groups of no more than three (3) people for this project. ****

Predictive Analysis using scikit-learn

As data analysts, we're often tasked with taking data in one form and transforming it for easier downstream analysis. In this Project, you'll use what you've learned in the course to prepare data for predictive analysis and then construct a predictive model using tools available within the **`scikit-learn`** library.

The data set you will be using for this Project is comprised of medical information of individuals who complained of chest pain problems. These individuals were evaluated by medical practitioners to determine whether or not the chest pain problems were the result of a previously undiagnosed case of heart disease. The data contain 1 response/dependent variable (which indicates whether or not the individual was diagnosed with heart disease) and 10 explanatory/independent variables. A data dictionary for the dataset is provided below.

Attribute	Description
Diagnosis	Indicates whether individual has heart disease (0 = no, >0 = yes)
Age	Age if individual in years
Gender	Gender (1 = Male; 0 = Female)
PainType	Chest pain type (1 = 'Typical Angina', 2 = 'atypical angina', 3 = 'non-anginal pain', 4 = 'asymptomatic')
BloodPres	Resting systolic blood pressure in mm Hg
Chol	Serum cholesterol in milligrams/decilitre
FBSugar	Indicator: Is fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
RestECG	Resting echocardiogram results: 0 = normal; > 0 = abnormal
MaxHeart	Maximum heart rate achieved
ExAngina	Indicator: Is exercise induced angina present? (1 = true; 0 = false)
STPeakSlope	Stress test results: What is slope of peak exercise segment? (1 = 'upsloping', 2= 'flat', 3 = 'down sloping)

In this Project, we'll use **`scikit-learn`** to answer the question:

"***Which attribute or attributes are the best predictors of whether an individual has heart disease?"***

The work you will need to do for this Project can be separated into two distinct phases: all of the work required for **Phase I** can be completed using the Python and Pandas skills you developed through Week 11 of this class (i.e., without any prior **`scikit-learn`** knowledge), while Phase II will require the use of **`scikit-learn`** to assess the predictive qualities of the attributes contained within your DataFrame.

Phase I: Data Acquisition, Data Preparation & Exploratory Data Analysis (60 Points)

- Construct a professional-quality introductory narrative that explains the purpose and objectives of your work.
- Study the dataset and the associated description of the data (i.e. "data dictionary").
- Load the provided data set from your Github repository into your Jupyter Notebook.

- Create a pandas DataFrame **with a subset of the columns in the dataset**. You should include the **Diagnosis** column, which indicates **whether or not an individual was diagnosed with heart disease**, the **Age, Gender, PainType**, and **STPeakSlope** columns and **at least two other columns** of your choosing.

- Perform exploratory data analysis: show the distribution of data for each of the columns you selected (including the **Age, Gender, PainType** and **STPeakSlope** columns**)** , and show plots for Diagnosis vs. Age, Diagnosis vs. Gender, Diagnosis vs. PainType, Diagnosis vs. STPeakSlope, as well as the other columns that you selected. It is up to you to decide which types of plots to use for these tasks. Include explanatory commentary describing your EDA findings for all of the above.

- Include some text describing your preliminary conclusions about whether any of the other columns you've included in your subset *(i.e., aside from the **Diagnosis** indicator)* could be helpful in predicting whether a specific individual has heart disease.

- For both the **PainType** and **STPeakSlope** columns in your DataFrame create a set of dummy variables. This is necessary because your downstream processing in Project 4 using scikit-learn requires that categorical data values be converted to binary indicator variables.. See the pandas **get_dummies()** method we've discussed previously for one possible approach to doing this.

Phase II: Build Predictive Models (40 Points)

- Start with the data (including the dummy variables) in the pandas DataFrame that you constructed in Phase I.

- Use **scikit-learn** to determine which of the predictor columns that you selected (including the **Age, Gender, PainType**, **STPeakSlope,** and the other columns of your choosing) most accurately predicts whether or not an individual is likely to have heart disease. How you go about doing this with **scikit-learn** is up to you as a practitioner of data analytics.

- Clearly state your conclusions along with any recommendations for further analysis.

****HINT**** : If you understand the process used in the M12 DataSchool videos on Machine Learning with scikit-learn to predict iris species from four predictor variables, you should be able to apply what you've learned to complete this Project.

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Save all of your work for this project within **a single Jupyter Notebook** and submit it via the Project 4 page within Canvas. Be sure to save your Notebook using the following nomenclature : **first initial_last name_Project4**" (e.g., J_Smith_Project4). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***