Mini-Lab Analysis

## PipeLine and Cross Validation

The below cell analyzes performance of 3 classifiers to determine whether a patient may have cardiovascular disease. Each of the 3 classifiers uses default parameters. A grid search here will be conducted and analyzed in later cells. The 3 classifiers analyzed include a Logistic Regression model, a Stochastic Gradient Descent (SGD) model, and a Full Support Vector Classifier (SVC) model. Each of the models consists of all original and new features having standardized values. The RobustScaler() function from the scikit-learn python package is used to standardize values between features. The RobustScaler() function scales features using statistical analysis that is robust to outliers in a dataset. (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html).

A one-hot encoding (OHE) has been implemented to standardize attributes such as gender and cholesterol. However, since these categorical attributes were recorded as numerical variables when recorded into the dataset, OHE is not expected to impact the performance of the classifiers. For example, gender was recorded as between 1 or 2 instead of 0 and 1, which is still indicative of an ordinal relationship. Therefore, OHE may aid in standardizing the data, but in this instance, not impact performance of the models because of the encoding that occurred when the data were recorded. Stated sufficiently, OHE may be redundant.

Each of the models are analyzed using an 80/20 train/test split. A 10-fold Cross-Validation (CV) is included with each of the classifiers.

Using the ROC AUC metric to measure performance between classifiers, each model performs equally. Indeed, the only apparent difference is that the SVC model value is within +/- 0.01, whereas the others are within +/- 0.00.

A variety of different parameters may be adjusted to improve performance of each of the models. For example, instead of using the RobustScaler() function, the StandardScalar() function may be implemented from the scikit-learn python package. The StandardScalar() function removes the mean and scales to a unit variance to standardize the data.

(https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html).

Otherwise, individual parameters for each of the models may be tuned to improve performance. For example, the below cell executes the GridSearchCV() function from the scikit-learn python package to tune model parameters to improve performance of each of the models.

The GridSearchCV() function compares the possible combinations of the parameters to achieve the best performing model. Specifically, the GridSearchCV() function uses a “fit” and a “score” analysis to find estimators for optimized, specified parameter values.

(https://scikit-learn.org/stable/modules/grid\_search.html#grid-search.)

As shown, the logistic regression model parameters being compared in the GridSearchCV() function are logistic C values and logistic penalty values. Likewise, the parameters being compared in the SGD model include α values, loss values, class weight values, and penalty values. The linear SVC model compared the C, penalty, and loss parameters to optimize performance of the model. Lastly, an SVC model comparing 2 kernel parameters was included in the analysis. The optimized values for each of the varied parameters are shown. As seen in the output, the GridSearchCV() function provided some separation between the classifiers. Specifically, the SVC model performed best having a slight increase in performance over the performance of the classifiers without the GridSearchCV() function.

# Model Advantages (10)

Each of the classifiers output similar scores with slight variation between them. The logistic regression model had a ROC AUC score of 0.7695, whereas the SGD and linear SVC models had scores of 0.7707 and 0.7694, respectively. As stated above, the best performing model was the SVC model that had a ROC AUC score of 0.7781. Therefore, while the SVC model scored best, the improved performance does not indicate that the performance is superior to the others.

While all these scores are similar and close together, the time needed to run each of the models varies. The mean fit and scores times, and the cross-validation times differ significantly between models. For example, the mean fit time for the logistic regression model is 0.3632 seconds, whereas the mean fit time for the linear SVC and SVC models have fit times of 7.7961 second and 400.3548 seconds, respectively. The fit time for the SGD model is 0.6603. The score times for each of the models are relatively similar except for SVC model. Each of the score times for the logistic regression, SGD, linear SVC and SVC are 0.0104, 0.0199, 0.01568, and 13.3632, respectively. Lastly, the cross-validation time for each of the models varied significantly with the SVC model being the longest at 1723.5489 seconds compared to 4.9318 seconds of the logistic regression model. The SGD model performs faster than the linear SVC model because the linear SVC includes an additional 5-fold cross-validation, which increases the processing time. (https://scikit-learn.org/stable/modules/svm.html).

As discussed, the performance of each of the models does not significantly differ while the time needed to run does vary significantly. This impacts efficiency of each of the models. For example, the best performing SVC model took the longest to run, and the improved difference in performance is not significant over the much faster logistic regression model. Therefore, since the logistic regression model has less than a point difference in performance, but a much faster processing time, the logistic regression model is the most efficient model.

# Interpret Feature Importance (30)

Use the weights from logistic regression to interpret the importance of different features for the classification task. Explain your interpretation in detail. Why do you think some variables are more important?

Using the parameters found from the GridSearch() function, a logistic regression model is executed in the below cell.

The above graph depicts feature weights for each of the features in the logistic regression model. Based on the above graph, the features that contribute significantly to this model include “Age,” “Cholesterol,” “Blood Pressure” and the” BMI Group”. Other attributes, like “Gender,” “Smoking,” “Alcohol” and “Activity” don't appear to as strong of indicators for the presence of cardiovascular disease. For example, each of the features “Age,” “Cholesterol, “Blood Pressure” and the” BMI Group” have strong weights that positively extend to around 0.5. Likewise, some of the remaining features have some positive extending weights, such as “Smoking,” “Alcohol” and “Activity.” These weights, however, only appear to extend to around 0.1.

Using common domain knowledge, the features “Age,” “Cholesterol,” “Blood Pressure” and the” BMI Group” appear to be consistent with attributes that would contribute significantly to predicting the presence of cardiovascular disease. For example, as a person ages, so does the heart tissue, which would likely result in more cardiovascular injuries. Similarly, a high cholesterol value is indicative of a high presence of a substance commonly known to cause artery blockage. Lastly, a similar analysis applies to the “Blood Pressure” and the” BMI Group” features. As the value in each of those features increases, the likelihood of cardiovascular injury also increases, for example through increased arterial pressure or a high body mass. The other positively extending features also depict that some behaviors, such as smoking and alcohol intake, and inactivity may contribute slightly to the presence of cardiovascular disease. This also matches the common domain knowledge of cardiovascular disease causes. The below cell executes a further logistic regression model using the 4, strongest, weighted features to analyze changes in model performance.

The above accuracy value, confusion matrix, and ROC curve depict that performance of the logistic regression model with the features “Age,” “Cholesterol, “Blood Pressure” and the” BMI Group” is nearly identical to the full model. For example, the ROC AUC value is 0.76, or identical to the initial logistic regression model. Likewise, the accuracy value is similar to the previous logistic regression model at 0.71. Therefore, based on the above analysis, the reduced complexity logistic regression model provides the most interpretable, and efficient classifier to predict the presence of cardiovascular disease.