**Initial Results and Code**

Link to Github: https://github.com/adchan11/CIND820

## Data Cleaning and Preparation

Inconsistencies with the data were addressed such as removal of missing variables and outliers. Missing values were imputed using the median. Numeric attributes that are skewed were normalized and all numeric attributes were scaled by min max scaling. There are 12 different class variables (i.e. complications) in the dataset but they are all severely imbalanced. Therefore, a new class variable titled ‘any\_complication’ will be derived from all 12 class variables such that 0/FALSE means the patient had no complications at all and 1/TRUE means the patient had one or more complications.

## Exploratory Analysis

The dataset consists of 1700 rows and 125 columns (119 integer and 6 numeric). 7.5% of the dataset contains missing values but there are no duplicate rows. Descriptive statistic revealed blood pressure variables had values of 0 as well as discrepancies between values reported by different healthcare units. Therefore, rows with values of 0 in either blood pressure variable were removed and if there were discrepant values, the mean was calculated. Otherwise, the remaining value was coalesced if one was missing. Similarly, it is redundant to have both systolic and diastolic pressure so systolic pressure was chosen as a proxy for blood pressure. Histograms of attribute distribution showed that some numeric variables were skewed and collected on different scales. The class variable suffers from some class imbalance while many categorical input variables are heavily skewed. Boxplots with t-tests showed that AST\_BLOOD and ALT\_BLOOD had no statistical significance in relation to any\_complication. Correlation analysis was performed using the point biserial correlation coefficient and t-test between numeric attributes and the binary class variable, Pearson correlation coefficient and t-test between numeric attributes, phi correlation coefficient and chi-square test between binary attributes, and Cramér's V correlation coefficient and chi-square test between categorical attributes. Some multicollinear trends exist which will be addressed in the data preparation section. Interestingly, some comorbidities such as obesity and chronic pneumonia do not show statistically significant correlation with any\_complication. While beta blockers and calcium chain blockers decrease the chance of having a complication, use of opioid drugs, lidocaine and liquid nitrates increase the chance of having a complication. Timing of the therapeutic intervention also does not seem to play a role in developing a complication.

## Dimensionality Reduction

Features that are not statistically significant were removed and some features that are highly correlated to each other were also removed to prevent multicollinearity. Features were ranked by importance using a learning vector quantization (LVQ) model, where the most important ones were selected.

## Predictive Modeling & Validation

The training and test dataset were randomly created in a 7:3 split ratio. SMOTE was performed on the imbalanced class variable. Classification algorithms such as random forest, naïve Bayes, stepwise logistic regression and neural networks were applied for predictive modeling. Confusion matrices were created to compare accuracy, precision, recall and F1 score of each method. Other metrics such as the ROC curve, AUC, and training run time of each mode were measured. A summary of the results is presented below, where the best performing metric is highlighted in yellow.

| **Model** | **Random Forest** | **Naïve Bayes** | **Stepwise Logistic Regression** | **Neural Networks** |
| --- | --- | --- | --- | --- |
| Accuracy | 0.6925 | 0.6329 | 0.6885 | 0.6349 |
| Precision | 0.7712 | 0.8268 | 0.7799 | 0.7634 |
| Recall | 0.6921 | 0.4901 | 0.6689 | 0.5662 |
| F1 Score | 0.7295 | 0.6154 | 0.7201 | 0.6502 |
| AUC | 0.693 | 0.668 | 0.693 | 0.652 |
| Training Run Time | 4.690 s | 0.009 s | 0.651 s | 0.395 s |