ETL & LucidChart (Twilla)

Preprocessing (Twilla)

Categorical variables changed using pd.factorize

Changing outcomes for classification to binary

Identify variables

Because of obvious multicollinearity, perform PCA (rather than RFE)

Try it with 2 Principal Components first and graph that, diff colors for depressed & not

Examine 95% variance cutoff to determine appropriate number of components (47)

Train\_test\_split

StandardScaler for X\_train & X-test – (Fit\_transform X\_train, only Fit X\_test

Perform PCA with optimal number of components

To balance the dataset and avoid over/underfitting the data, perform SMOTE (multiple times? ONLY ONCE!)

Train\_test\_split

RandomForestClassifier

Parameter tuning

Performance metrics