ETL & LucidChart (Twilla)

Tables created – combination of LabelEncoder and pd.factorize to create unique id’s for each table.

Data separated into 4 csv files to ease in loading the tables. Used psycopg2 to create the database (dass), create the tables and then load the tables

LucidChart used to create the data model

Preprocessing (Twilla)

What data we went with and maybe a little bit about it – not the detail on ratings but more a general idea of what was in the data. Reading it in was different than what we had done with csv files in the past – had to use \t separator instead of commas.

Special note on the 4648 majors that were combined down to 80, with 38 rows being dropped due to having fewer than 10 with those majors – not sure if all that is needed.

More with extreme/severe depression and anxiety while stress levels were normal.

Boxplots used to find outliers on age and family size, then used mean to replace those values.

Removal of invalid 0’s for most fields since valid values did not include 0’s

Nulls in country were replaced by the country used most often

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

Performance metrics/Parameter tuning