ECS 171 - Mid-Quarter Progress Report

Group 6: **Arjun Ashok** (arjashok@ucdavis.edu), **Tej Sidhu** (tejsidhu@ucdavis.edu), **Taha Abdullah** (tmabdullah@ucdavis.edu), **Ayush Tripathi** (atripathi@ucdavis.edu), **Devon Streelman** (djstreelman@ucdavis.edu)

Overview

Our goal with this project was to use a neural network to classify individuals into three classes: no diabetes, pre-diabetes, and diabetes. Since we are using a neural network though, we wanted a dataset with a lot of features and data points. As a result, we needed to do extensive exploratory data analysis and preprocessing in order to make sure that our data is clean and good enough for our model.

Due to the large number of features (22), we couldn't handle all the features as a single group. So we split our features into three categories: numeric, ordinal, nominal. The numeric features are our features that have a numeric value and in order to clean up these features, we standardized the data using the standard scaler. We used standard scaler over min-max scaler because standard scaler handles outliers and larger variances better than min-max scaler. For the categorical features, a good portion of the data was stored as a boolean because it would be using a true or false statement regarding some pre-existing condition or activity. Hence, these data points are stored in our nominal category and can be left alone because there is no cleaning or adjustment needed for them. Lastly, we have our ordinal data in which each number corresponds to some underlying value, like education level. For this group, we looked into one-hot encoding. The problem with one-hot encoding is that it would increase our collinearity, and since the numbers are already ordered with respect to the other numbers, the neural network should already be able to accurately draw relationships between the data and features. Hence, we decided to also leave this part of the dataset untouched.

Distributions

In 22 features, our dataset consists of 3 numeric, 14 nominal, and 5 ordinal variables. The data seems to be collected on the average age of 53, with the income distribution skewed left, indicating bias of the data towards the older upper middle class. The data suggests that the participants in the study are relatively fit, a large percentage exhibiting good mental and physical health while having good nutritional habits. Our target (Diabetes_012) describes a large imbalance in the sample, where over 84% of the participants were not found to have diabetes compared to the 14% that do.

Outliers

We have done outlier detection and evaluation on our data, and found a lot of the data is represented in binary, so we were able to ignore those in our outlier detection. The main features we wanted to focus on in regards to outliers were bmi and age. Age was found to have no outliers, while bmi had quite a few. Upon discussion, we concluded that the outliers, mostly above the IQR, should remain included in our data. This is because it demonstrates the different ranges at which diabetes can occur, and a shift up in our data may actually result in a better model as higher bmi typically correlates with diabetes risk.

Feature Selection

Our approach to feature selection was four-fold:

(a) Check for collinearity between features: the more related features are (in a linear manner), the more likely they are to be redundant **(lower is better)**

(b) Check for target correlation: the higher the trend between the target feature and every feature, the more likely they will be able to contribute the model's eventual predictive power **(higher is better)**

(c) Regularized Regression: if we utilize some sort of penalty during regression modeling on the feature weights (especially in LASSO), we can conduct feature selection by removing any features whose eventual weights are very close to zero; if any features aren't zero-coefficient, we can still use the information about their weights to determine which features are not as useful, so long as everything has been normalized so the scale of the feature doesn't impact the weight **(higher is better)**

(d) Context-driven selection: the final step is to consider the actual context of what we are predicting to determine what features are redundant, irrelevant, or otherwise not that useful during the modeling process. We combine this analysis with the others to cross-check that just because no linear relationship is present, we aren't removing features unnecessarily (especially in process (a) and (b))

With this approach, we determined two features that seemed to under-perform in all of these tests relative to the other features. Combined with our manual pruning and considering the context of the prediction, we decided that **diff_walk** (difficulty in walking) and **smoker** (is a smoker or not, dichotomous) are both redundant and don't contribute enough to justify their inclusion. There were other features that weren't perfect throughout all tests, but given that we had a large dataset and no surefire way to determine non-linear relationships, we opted to err on the side of keeping the features rather than discard unnecessarily.

Up-Sampling

When we were looking through the instances of each output, we observed a major disparity in the number of instances for each diabetes target class: 0 - no diabetes (213703 instances), 1 - prediabetes (4631 instances), and 2 - diabetes (35346 instances). As a result, we decided to resample our dataset using up-sampling (through sk-learn resample()) to reproduce instances of the minority classes (1 and 2) so that each class had the same instance count as class 0. This is done through the bootstrap process.