02: Pre-Analysis Plan

The purpose of the pre-analysis plan is to get you to commit to a framework for doing your analysis that makes sense, given the tools you've seen so far and the data you have. Pre-analysis plans are common in experimental fields, in order to bring some discipline to the data analysis.

You should address the following questions explicitly:

- What is an observation in your study?

An observation in our study is a single survey response from an individual participant. Each observation includes the person's demographic details, health behaviors, and diabetes status.

- Are you doing supervised or unsupervised learning? Classification or regression?

We will be doing supervising learning because we are using labeled data (diabetes status) to train our model for predicting outcomes. This is classification because we are attempted to predict a categorical outcome, whether a person has diabetes or not depending on their health indicators.

- What models or algorithms do you plan to use in your analysis? How?

We want to analyze someone's probability of having diabetes depending on their health factors and identify the most predictive factors. To do our analysis we plan on creating a baseline logistic regression to look at the linear relationships between variables and their influence on diabetes risk. Next, we plan on applying decision trees to capture the non-linear patterns amongst these health indicators. Lastly, we will implement random forests to mitigate overfitting/sensitivity of decision tress. By combing multiple decision trees, we hope it will provide the most robust predictions on diabetes likelihood. Of note, we might use PCA to reduce dimensionality of our data as certain indicators, like BMI and physical activity, might be correlated leading to redundancy.

- How will you know if your approach "works"? What does success mean?

Success will be if the model reliability identifies individuals at risk for diabetes. To know if our approach works, we will look the confusion matrix and resulting accuracy, sensitivity/recall, specificity, F1 score, and MCC scores. F1 and MCC will most likely be the most indicative of success as our dataset is imbalanced with more healthy individuals than diabetic ones. Success will than hopefully be high levels for these measures.

- What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?

One major weakness we anticipate being an issue is the class imbalance inherently in the data as there are far fewer diabetes cases than healthy cases. We will need to use some technique to adjust the class weights so that it doesn't cause discrepancies in our model. Otherwise, we will most likely just look at F1 MCC oppose to accuracy to mitigate this issue.

- Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?

To prepare our data for analysis, we will implement several feature engineering techniques. We'll apply one-hot encoding to categorical variables such as 'Sex', 'Education', and 'Income', transforming them into binary features that machine learning algorithms can interpret more effectively. For numerical features like BMI, Age, and health-related variables, we'll use normalization techniques such as Min-Max scaling or Standard scaling to ensure no single feature dominates due to its scale. Given the presence of correlated variables, we plan to use Principal Component Analysis (PCA) on groups of related numerical variables to reduce dimensionality and address multicollinearity. This is particularly useful for health indicators that show correlation, such as General Health, Mental Health, Physical Health, Difficulty Walking. We'll also employ feature selection techniques to identify the most important predictors of diabetes risk. Additionally, we'll create meaningful interaction terms between variables, such as BMI * Age, to capture non-linear relationships, and to observe the combined effects of obesity and aging. For features that are continuous/categorical, mean/mode imputation will be used, and features with an abundance of missing data will be dropped.

- Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.

To effectively communicate and present our results, we'll use a variety of visual and numerical methods. We'll create a confusion matrix to visualize our model's performance in classifying diabetes risk, showing true positives, true negatives, false positives, and false negatives. The ROC curve, which plots the true positive rate (recall) against the false positive rate, and PR curves will both be generated to illustrate our model's discrimination ability across different thresholds, which is particularly useful for imbalanced datasets. Furthermore, the AUC (Area Under the Curve) will quantify overall how well the model is able to distinguish between classes. We'll produce a feature importance of different predictors in our model, helping to identify the most influential factors in diabetes risk. A comprehensive performance metrics table will be presented, comparing different models using metrics such as accuracy (overall correctness), precision (percentage of correctly predicted positives), recall (percentage of actual positives that were detected), F1 score (metric that measures the balance between precision and recall), and Matthews Correlation Coefficient (metric that considers all four confusion matrix elements and provides a single value that measures a balanced measure of the model's performance) .

Research Question: ***“What individuals are most at risk of developing diabetes based on their health indicators?”***