**Initial Report and Exploratory Data Analysis**

Marjorie Wilke

**Summary**

This software analyzes a dataset linking 36 features of health data to various gastrointestinal disease states. The data are very clean just needing conversion to numeric values. The dataset can be found at https://www.kaggle.com/datasets/amanik000/gastrointestinal-disease-dataset

This software could serve as a diagnostic tool to guide doctors in diagnosing gastrointestinal disorders.

The preliminary results indicate there is not a simple solution to diagnosing these disorders with the given data. All the features have very low correlation with the disease states. Most of the features are binary. Those that are continuous do not appear to be related to the target or each other with the exception of features describing weight being closely correlated (Body Mass Index, Weight, and Obesity Status). Linear regression is not feasible. Because the target is categorical, 6 different disease states, clustering was the best approach. Clustering indicates that 4 disease states may be more accurate than 6. The 6 disease states were only weakly correlated with each other. Furthermore, it takes 38 features to reach 90% accuracy for prediction. Because of the large number of features needed, I avoided using KNN. Logistic regression is best with 2 classes, so I used the decision tree algorithm. The best score was very low at 0.17 despite tuning the hyperparameters. Furthermore, the best score only split the data into 2 of the 6 diagnoses. At this point I am concerned that there is no signal to find in these data. Even random forest gave an out of box score of 0.17.

I am hoping neural nets will be able to pull a signal out of these data. I have questions about how to visualize the clusters with so many features. Is there a way to project them onto 2 or 3 dimensions? I did try dimensionality reduction to no avail. Is there a way to at least see what features and cutoffs point to a particular category in the target? It seems like decision trees would do this but the scores are so low I don’t trust the output. Clearly software would have to be provided to any doctor using the output because of the number of features involved in suggesting a diagnosis.

“Biology has too many variables and not enough equations.”

~Margie 1996

Link to Jupyter notebook