# Busara Depression Prediction

The following data was gathered by the Busara Center located in Siaya County in Kenya in 2015. Over 1100 rural locals were surveyed on their economic situations, expenditures, and familial makeup. The original dataset contains 1143 objects and 75 features, with the first two being a survey and village id. Three variables; survey id and village id may be useful when examining clusters further, and the final variable is the classifier: depressed (0=not depressed, 1=depressed).

There is also a training dataset with 286 observations however, this is excluded as almost all the depression classifier values are missing/empty.

Note: Whether the participant is depressed or not is not clinically diagnosed but instead screened for using an epidemiological measure.

**Link to dataset:** https://zindi.africa/competitions/busara-mental-health-prediction-challenge/data

**Research Question:** Can we classify and predict the occurrence of depression based on a select number of socio-economic variables and determine which ones have a greater influence, if any, on the outcome?

**Proposed methods:** Feature selection techniques to reduce the dimensionality of the data to choose the most influential predictor variables. Clustering using KMeans, Hierarchical Clustering etc; Classification and Prediction using KNN, Logistic Regression and other GLMs like Lasso, Ridge, Elastic Net if multicollinearity exists.