This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The aim here is to use the Principal Component Analysis (PCA) and/or Partial Least Squares Regression (PLS) to indicate the variables that affect the appearance of diabetes less. In other words, which variables do not have a relationship or have a partial relationship with diabetes.

Data Content:

The datasets consists of several medical predictor variables and one target variable, Outcome. In our case we will call the outcome ‘Class’. Predictor variables includes:

1. The number of pregnancies the patient has had.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test (Blood Sugar).
3. Diastolic blood pressure (Blood Pressure).
4. Body mass index (Mass).
5. Diabetes pedigree function (pedigree).
6. Age.
7. The presence of diabetes (Diabetes).