Prediction of Early Onset Diabetes Using Classification Algorithms

By The Hardworking Procrastinators
Christopher Panlasigui[1], Jayashree Kapoor[2], Tia Whiteman[3], Wai Moon Chin[4]
[1,2,3,4]City University of N.Y. - Baruch College Zicklin School of Business
Data Mining for Business Analytics

Abstract

Over a third of the US population is afflicted by pre-diabetes or diabetes which leads to other health complications and oftentimes results in deaths. Early detection of the disease can avoid or delay complications. With the aid of data mining methods, we are going to construct predictive classification models to detect the early onset of diabetes from the diabetes dataset obtained from the UCI Machine Learning Repository. These models will be assessed using (k=10)-fold cross-validation and 80:20 re-sampling split. The final and best model will be used as a tool for individuals, families, and healthcare professionals to use to detect the early onset of the disease to help prevent or delay health complications.

Data Set

We obtained our dataset from UCI Machine Learning Repository. The data was originally collected using direct questionnaires from patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. It has 520 observations with 17 variables, which are all categorical except for Age. Our target variable is the response Class, which is categorical. There are no specific challenges in our dataset, nor does it contain any missing values. However, our dataset restricts our data mining procedures to classification. We intend to use all variables.

Data Mining Task

We would like to determine whether someone is at risk (positive) or not at risk (negative) for the early onset of diabetes based on a set of symptoms. More specifically, what combination of symptoms is a good indicator?

Methods and Models

Since all but one of the variables are categorical, our predictive data mining task is restricted to classification methods. Thus, we intend to implement multiple logistic regression, Naïve Bayes classifier, and random forest methods. We may perform binning data transformation on the Age variable to turn it into ordinal categorical data to improve model performance. Finally, categorical values will be re-coded.