**SAS Shootout Problem – Diabetes Data Mining Contest**

Thomas Richardson

1. Introduction

The problem, as outlined by SAS, is a data mining exercise to establish the means to determine three main goals: how diabetes influences and individual’s health care costs, the age and gender distribution of people with diabetes, and how effective preventative measures directed at a specific population based on their BMI would be. The data set as provided contains information about 40 different parameters for 50,000 people in the United States. It is assumed that this data set provides an accurate “snap-shot” of the US population.

• Problem or subject to be studied

• Background (Literature Review)

2. Proposed approach

For the first problem, we need to establish some function to describe health care costs for individuals as a function of diabetes, that is, we need to find a relationship between the various parameters provided and health care. It may also be useful to find relations for total costs, Medicare costs, and Medicaid costs independently. We intend to use fuzzy clustering or fuzzy regression for this goal.

The second problem, finding diabetes distributions as a function of age and gender is pretty straightforward. For this, we just need to find the relevant percentages falling into each category. From this, information about a critical onset age may revel itself, or a gender bias for the disease.

The third problem will likely be the most difficult, as it is rather open ended beyond the statement of the problem. In determining how effective the proposed measure will be, we must analyze the onset of diabetes and / or medical costs for people in the stated BMI categories, and determine relevant correlations. This is difficult since the data provides only a snapshot, and not a time series of data points. Therefore, we have little knowledge of how these parameters changed for individuals over time, though we can extrapolate the onset of particular things (diabetes, heart conditions, BMI changes) by trends in the data.

It would be useful to first use a clustering algorithm or neural net that could accurately place data points into the correct diabetes category based on the parameters provided. A more “continuous” clustering range of placement tags, such as “very likely”, “likely”, “not very likely”, etc., would probably be more useful than a more discrete clustering. The trained weights on a neural net, if used, may also be observed to indicate how much a particular set of parameters influences diabetes onset. The parameters from these models could likely be used to assess which parameters or sets of parameters are most correlated to an individual having diabetes.