CAPSTONE PROJECT – ADVANCED STATISTICAL MODELLING

**Predicting The Onset of Diabetes – An investigation into the optimal statistical model for predicting the onset of gestational diabetes**

**William Belcher**

Gestational diabetes mellitus (GDM) is a subtype of diabetes diagnosed during the second part of pregnancy, gestation, and remains until the baby is born. Currently, all women are screened for GDM during their routine 24-to-28-week check-up [https://www.pregnancybirthbaby.org.au/gestational-diabetes]. This test is quite lengthy, taking a couple of hours. With the rapid increase in computing power and use of machine learning models in a clinical environment, it now seems plausible for a time-efficient screening process to be implemented for GDM. In this paper, the methods of Logistic Regression

**Introduction**

Gestation diabetes mellitus (GDM) is a form of diabetes that occurs during pregnancy and persists until after the birth of the child. [https://www.diabetesaustralia.com.au/about-diabetes/gestational-diabetes/] Diagnosis typically happens in around the 24th to 28th week of the pregnancy. GDM can be attributed to the increased insulin resistance caused by the hormone blocking nature of the placenta. During pregnancy, the need for insulin can be as high as 2 to 3 times higher than normal. If a woman already suffered from insulin resistance, the pancreas may be unable to cope with the heightened demand, leading to higher blood glucose levels and a diagnosis of GDM. Currently, women are referred to an oral glucose tolerance test at a pathology lab to determine if they have GDM. This test requires fasting from the previous night and can take between one to two hours. With the swift increase in computation and of machine learning models in the clinical setting, it now seems plausible for a statistical model to assist in the early detection and diagnosis of gestational diabetes mellitus. The early detection of GDM enables a management plan to be developed and enacted earlier, minimising the impact of the condition on the expecting mother

**Data**

The Diabetes data set, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains several medical predictors and one target variable [https://www.kaggle.com/uciml/pima-indians-diabetes-database]. The medical predictors had many types, being either numerically discrete or continuous with the response variable being a binary 1 or 0, with a 1 representing the patient being diabetic. The dataset set is a subset of a larger database, with each observation being taken from a female patient over the age of 20 and of Pima Indian heritage. The following table details all 9 variables:

|  |  |
| --- | --- |
| Variable | Description |
| Pregnancies | Number of pregnancies (Discrete) |
| Glucose | Plasma glucose concentration |
| BloodPressure | Diastolic blood pressure (mm Hg) |
| SkinThickness | Triceps skin fold thickness (mm) |
| Insulin | 2-Hour serum insulin (mu U/ml) |
| BMI | Body mass index |
| DiabetesPedigreeFunction | Diabetes pedigree function |
| Age | Age (years) |
| Outcome | Response (binary) |

Table 1: List of variables and their associated types

In total, there were 768 observations. It is worth noting not much information about the DiabetesPedigreeFunction field was supplied. However, for this report it is assumed that this function returns some information based on the family’s history of gestational diabetes.

*Data Pre-processing*

Data pre-processing is a technique in which the raw, imported data is transformed into a more meaningful and usable format. A summary of the dataset showed that minimum value for: Glucose, BloodPressure, SkinThickness, Insulin and BMI, was zero. It is unreasonable to have zero for these values so the rows containing zeros in these columns were removed. Missing value imputation was not utilised in order to not introduce unnecessary variance into the models. The size of the cleaned data that was then utilised in the construction of the models contained 9 columns, 8 being predictors and 1 being a response, and a total of 392 observations.

**Methods**

**Results and Discussion**

**Conclusion**

**References**

* [4] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA <http://www.rstudio.com/>

Mode: Desktop. Version 1.25042