**Predict the onset of diabetes based on diagnostic measures**

DS 740: FINAL PROJECT: Connie Sosa: 12/16/2016

**Executive Summary**

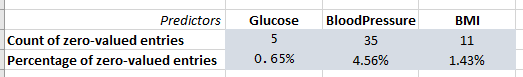
1. **INTRODUCTION**

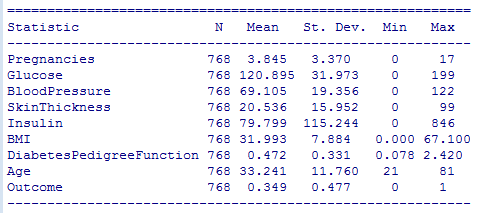
The Pima Indians Diabetes Database [dataset](https://www.kaggle.com/uciml/pima-indians-diabetes-database/downloads/diabetes.csv) obtained from kaggle site, contains 768 cases of female patients of Pima Indian heritage and each case contains nine variables. This analysis attempts to predict the onset of diabetes based on diagnostic measures such as Glucose, Blood Pressure, BMI, and Diabetes Pedigree Function.

The target audience for this analysis are health care providers and medical insurance companies. By having predictions of which patients might be at a higher risk of being diabetic, a possible early diabetes intervention program may be offered or further screening if necessary.

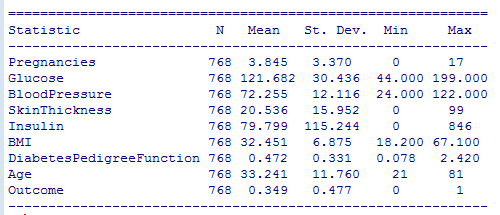
1. **DATA QUALITY**

Preprocessing of data is required prior to analyzing of this dataset. Important predictors, Blood Pressure, Glucose, and BMI, have missing or zero-valued entries. These missing data account for about 5% of the entire dataset (see Figure 1). The mean imputation method is employed to address this issue. Tables 1 and 2 show the summary statistics of the before and after mean imputation been applied to the data.

 ***Figure 1***



***Table 1.*** Data Summary Before Imputation



***Table 2.*** Data Summary After Imputation

1. **DATA MINING TECHNIQUES**

Method 1: Logistic Regression.

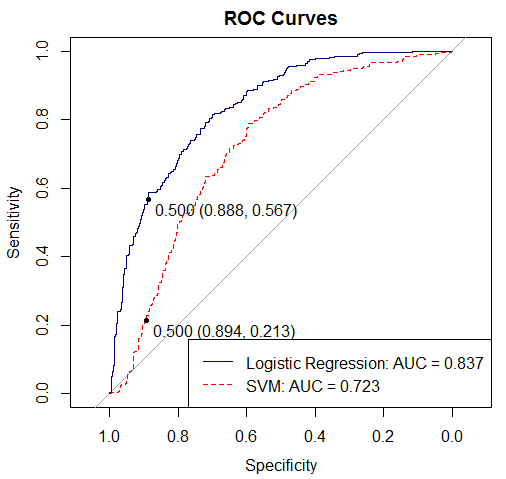
Method 2: Support Vector Machine (SVM)

This analysis is intended to predict whether a patient is diabetic, it is a two-category classification problem. Both logistic regression and Support Vector Machine (SVMs) methods are good techniques for this type of classification problem. These two methods can also provide an easy interpretation for the analysis. If the two categories are well separated, then SVMs tend to perform better, but if there exists much overlapping between the two categories, then logistic regression typically outperforms SVMs.

1. **RESULTS OF THE ANALYSIS**

This analysis used eight predictors to classify whether a patient is diabetic (*Outcome* of *1*) or not (*Outcome* of *0*). The classification is checked against the dataset’s response variable(*Outcome*). Model assessment compares the ROC curves between the two models. As shown in Figure 2, Logistic Regression is the better model to fit the data. It has a higher value of the area under the curve (AUC) of 0.837 compared with 0.723 for SVM.

Comparison of the overall accuracy at a threshold of 0.5 for the models shows the following results. Sensitivity, which measures the true positive rate, indicates that logistic regression is able to correctly classify a patient being diabetic 57% of the time, while SVM model at 21% of the time. The specificity, measures the model’s ability to correctly identify those who are not diabetic (true negative rate), is at about 89% for both logistic regression and SVMs. These two models have high specificity and low sensitivity at 0.5 threshold. The threshold may be adjusted to increase the true positive rate/decrease the true negative rate of the test.

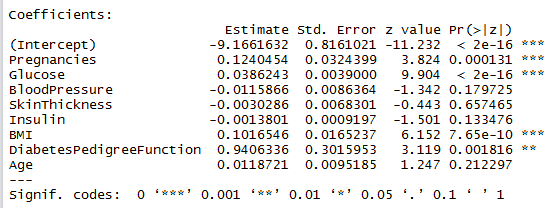


***Figure 2.*** ROC Curves for Logistic Regression and Support Vector Machine

1. **CONCLUSION**

Logistic Regression is shown to be the preferred model over SVMs. Based on the logistic regression model results shown in Table 3 below, we have a strong evidence to believe that when predicting the onset of diabetes based on these eight diagnostic measures, the number of pregnancies, glucose level, BMI, and Diabetes Pedigree Function are positively associated with the probability of diabetes onset. Out of these, glucose level and BMI may be closely monitored to lower the probability of diabetes onset.

Please note that the prediction based on this model will not be applicable to male patients nor non-Pima Indian heritage patients. Whether this model can be generalized to a broader patient base will need to be examined and studied further.



***Table 3.*** Summary of the Logistic Regression fitting.

**REFERENCES**

<https://www.kaggle.com/uciml/pima-indians-diabetes-database>

<https://www.kaggle.com/usersumit/d/uciml/pima-indians-diabetes-database/genetic-programming-generate-best-ml-algo>

<http://stackoverflow.com/questions/25835643/replacing-missing-values-in-r-with-column-mean>

https://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/