Bella Cho

Professor Cai

CIS 3920

15 December 2020

Prediction of Diabetes Mellitus at the Early Stage Using Data Mining Techniques

1.Introduction

1.1 Diabetes Mellitus

Diabetes Mellitus, also known as Diabetes, is a chronic metabolic disorder that is characterized by high blood sugar level over a prolonged period of time. There are two types of diabetes: type one diabetes is due to the pancreas not producing enough insulin, on the other hand, type two diabetes is due to the cells of the body not responding properly to the produced insulin. Over time diabetes can affect everybody part and cause damages to numerous organs and systems, it also increases risks for strokes, heart and blood vessel diseases, which can be life threatening. Diabetes is dangerous not only because it can be lethal, but also because it is asymptomatic and can proceed unnoticed till it’s very last stage, where it is too late for medical attentions.

1.2 Data Mining Techniques

The purpose of this project is to predict the likelihood of having diabetes based on certain health conditions to detect diabetes during an early stage. Hopefully with early detection, patients can receive early treatment and prevent diabetes from developing further. The data of this project is *Early-Stage Diabetes Risk Prediction Dataset*[[1]](#footnote-1) from *UCI Machine Learning Repository*. The source of this data is: M M Faniqul Islam[[2]](#footnote-2), Rahatara Ferdousi[[3]](#footnote-3), Sadikur Rahmanand Humayra[[4]](#footnote-4), and Yasmin Bushra[[5]](#footnote-5). This data is collected using direct questionnaires from the patients of Sylhet Diabetes hospital in Sylhet, Bangladesh. It has 16 attributes: Age, Sex, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesity, and class. Attribute Age’s datatype is integer, while every other attribute’s datatype is object (character). (Note: The character datatypes were converted to factor datatypes.)

1.3 Data

This is a prediction, supervised learning, and classification project. With these natures of the project, I implemented and compared the prediction accuracies between logistic regression algorithm and classification tree with random forest algorithm to develop the optimal model for predicting the likelihood of diabetes during an early stage. The training dataset was 80% of the data, and the testing dataset was 20% of the data. The threshold to determine if a patient is likely to be diagnosed with diabetes is 0.5.

2 Methodology and Analysis

For both methods, I computed models based on my testing dataset and made predictions based on the testing dataset. Then, I compared the predicted values to my testing dataset value by generating confusion matrixes. Based on the corresponding confusion matrix, the prediction accuracy was calculated with formula (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative). The precision was calculated with formula True Positive / (True Positive + False Positive).

2.1 Logistic Regression Model Prediction Result

Based on logistic regression model confusion matrix (figure 1), the accuracy and precision for logistic regression is 0.9423 and 0.9683 respectfully.

2.2 Classification Tree with Random Forest Model Prediction Result

Based on classification tree with random forest model confusion matrix (figure 2), the accuracy and precision for logistic regression is 0.9808 and 0.9846 respectfully.

2.3 Result Analysis

Although both logistic regression and classification tree with random forest are two of the most popular algorithms for supervised data prediction, in case of this project, classification tree with random forest method outperformed logistic regression. In addition, figure 3 sheds lights on which diabetes related symptoms are more prominent compared to other symptoms. Based on figure 3, polyuria, polydipsia, gender (male), and age are the most critical symptoms.

3. Conclusion

According to the World Health Organization, diabetes was “the seventh leading cause of death in 2016.”[[6]](#footnote-6) In addition, the number of people with diabetes almost tripled from 1980 to 2014.[[7]](#footnote-7) Diabetes is one of the fastest growing, life threatening diseases. In order to detect diabetes in its relatively long asymptomatic phase, people not only should maintain healthy diets and lifestyles, but also should be able to properly assess certain diabetes related symptoms and receive proper medical attention which will delay the development of diabetes.

Figure 1

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Test Values | | |
| Predicted Values |  | Negative | Positive |
| Negative | 37 (True Negative) | 4 (False Negative) |
| Positive | 2 (False Positive) | 61 (True Positive) |

Figure 2

|  |  |  |  |
| --- | --- | --- | --- |
| C.T. Random Forest | Test Values | | |
| Predicted Values |  | Negative | Positive |
| Negative | 38 (True Negative) | 1 (False Negative) |
| Positive | 1 (False Positive) | 63 (True Positive) |

Figure 3

Chart

Description automatically generated

1. Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.  
   Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.

   https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. [↑](#footnote-ref-1)
2. Queen Mary University of London, United Kingdom m.islam **'@'** smd17.qmul.ac.uk [↑](#footnote-ref-2)
3. Metropolitan University Sylhet, Bangladesh rahatara **'@'** metrouni.edu.bd [↑](#footnote-ref-3)
4. Metropolitan University Sylhet, Bangladesh rahmansadik004 **'@'** gmail.com [↑](#footnote-ref-4)
5. Metropolitan University Sylhet, Bangladesh humayrabushra234 **'@'** gmail.com [↑](#footnote-ref-5)
6. https://www.who.int/news-room/fact-sheets/detail/diabetes [↑](#footnote-ref-6)
7. https://www.who.int/news-room/fact-sheets/detail/diabetes [↑](#footnote-ref-7)