Gabriela Serrano  
JuanCarlos Jimenez  
CSC 597 - Statistical Learning with Applications in R  
Professor Vanessa Aguiar-Pulido

**Symptom Significance in Diabetes Diagnosis**

1. **Introduction**

The human body collectively utilizes its many systems to maintain homeostasis. Because of this complex nature, there lies the possibility of pathological disturbances due to one or many factors, which can be devastating given the right circumstances. A terrifying example of this is the Diabetes Mellitus (DM) disease. DM reduces the body’s ability to use Insulin to regulate blood glucose levels, which play an essential part of our bodies’ balance. In the last 20 years, the number of adults diagnosed with Diabetes has doubled, challenging our health, families, and economy. DM has grown into an epidemic; unfortunately, projected to continue worsening. In the United States alone, 32.3 million people already have Diabetes, and 96 million adults have Prediabetes. Of the former, 20% are undiagnosed and of the latter, more than 80% are unaware of their heightened risk for the condition. Given all this information, our efforts are focused on using statistical learning models to predict Diabetes as early as possible most accurately. Afterward, using this same model, we gather the most relevant predictors of the disease, so that patients have insight into when it might be critical to see a physician.

**Types of Diabetes and Their Impact on the Body:** There are various ways DM can present itself; Type 1 Diabetes Mellitus (T1DM), Type 2 Diabetes Mellitus (T2DM), Gestational Diabetes, or Prediabetes. Firstly, T1DM occurs when the body has an autoimmune reaction that stops it from producing insulin, the hormone that regulates blood glucose. Hence, exogenous insulin is needed to maintain healthy levels. T2DM Diabetes is characterized by the bodies ill use of its insulin. As a result of it developing over many years, it is typically diagnosed in adults. T2DM can largely be prevented or delayed with a healthy lifestyle. Approximately 90 to 95 percent of people diagnosed have T2DM, and 5 to 10 percent have T1DM. Prediabetes is a predecessor to T2DM, but equally as serious because it also increases your chances of heart disease and stroke. Lastly, Gestational Diabetes is a special type of Diabetes which presents itself throughout pregnancy where there was no previous case of diabetes in the mother. It goes away after birth yet increases risk of T2DM later. It not only increases risk for the mother, but also for the child. The child will have an increased risk of childhood obesity and more likely to develop T2DM.

As of 2022, DM is the 7th leading cause of death in the U.S. and predicted to be within the top 5 by 2030 for high income countries (e.g., U.S., China, India). As a result of the extra care needed to treat this disease, the medical cost for individuals diagnosed with Diabetes is twice as high compared to individuals with no diabetes. Moreover, the yearly cost of diabetes on the U.S. economy due to medical expenses and lost work and wages averages $327 billion. Thankfully because of T2DM’s preventability, detection through analyses of symptom assessments like the one we examine here, we may ease a burden on our health care systems and national financial expenditures.

Currently, the literature has found accurate ways in which to classify similar datasets, but we have found no records which outline the symptoms that cause the most cause for concern and how they factor into model classifications. The dataset we’ve used is an early-stage diabetes risk prediction dataset that has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and have been approved by a doctor. It has 16 features and 520 instances that suffice in creating these predictions. It can be found at <https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset>.

1. **Methods**

**Data**

Included as part of our response variables are age and gender. Response variables that were symptoms include presence of, Polyuria, Polydipsia, sudden weight loss, weakess, Polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial Paresis, muscle stiffness, Alopecia, and Obesity. Finally, our predictor value is a positive or negative diabetes diagnosis. Of the predictor’s values, there are 320 positive and 200 negative diabetes diagnosis. Before beginning to utilize the data we ensured that there were no missing or incompatible values in the tuples of our dataset by manually parsing through our Comma Separated Value (CSV) file and using the built-in functions included in the R programming language. Furthermore, we set a random seed generator to be able to reproduce our models.

**Techniques**

We used the following learning methods to build our classifiers:

* **Logistic Regression**
* **Decision Tree**
* **Random Forest**
* **Naïve Bayes Classifier**

**Evaluation**

For evaluating the performance of classifiers we’re using K-fold Cross Validation, where we divide our dataset into 20 different parts. First, we remove the initial part and fit the model on the 19 other parts to see how good the predictions are by computing the MSE on the first part. We repeat this a total of 20 different times taking out a different part each time. Finally, averaging the 20 different MSE’s we get an estimated validation (test) error rate for new observations.

1. **Results**

Performance metrics we can use:

* **Accuracy**
* **Sensitivity**
* **Specificity**
* **Precision**
* **F1-Score (Calculate Recall + Precision)**
* **~~ROC Curve~~**
* **~~AUC~~**
* **~~Precision Recall Curve~~**

**Table 1. Classification Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | | Kappa | Accuracy | Sensitivity | Specificity |
| Logistic Regression | | 0.8199 | 0.9157 | 0.8837 | 0.9672 |
| Decision Tree | | 0.7197 | 0.8679 | 0.9535 | 0.8689 |
| Random Forest | | 0.9589 | 0.9807 | 0.8372 | 0.9508 |
| Naïve Bayes | TRUE | 0.6648 | 0.8309 | 0.9070 | 0.9016 |
| FALSE | 0.7563 | 0.8863 | 0.9016 | 0.9016 |

(Results in blue are taken from confusion matrix instead of CV – not optimal)

The confusion matrix returns a set of 4 values. The true positive, true negative, false positive, and false negative predictions. The Accuracy of our models shown above are a result of the true positive and negative values over the total amount of predictions. Sensitivity (or recall) shows the number of positive values which were accurately classified. Hence, sensitivity takes the true positive prediction over the combination of the true positive plus the false negative values. Specificity shows the opposite of Sensitivity, such that we find the number of negative values accurately classified. Therefore, specificity is calculated by taking true negative predictions over the combination of the true negative plus false positive values.

Taking this into consideration, besides finding the best accuracy, our aims in finding the right model for our classifier would be to also have a high specificity. Although high sensitivity would be ideal, it can carry less weight because going to verify a diagnosis with a physician does not carry much concern. Being positive but believing you may be fine because you received a negative prediction would cause more long-term harm. In our classification models, the Random Forest model gives us high accuracy along with high specificity. Therefore, we chose this model to find our most important response values.

Current Citations

* Kaul K, Tarr JM, Ahmad SI, Kohner EM, Chibber R. Introduction to diabetes mellitus. Adv Exp Med Biol. 2012;771:1-11. doi: 10.1007/978-1-4614-5441-0\_1. PMID: 23393665.
* Hilary King, Ronald E Aubert, William H Herman; Global Burden of Diabetes, 1995–2025: Prevalence, numerical estimates, and projections. Diabetes Care 1 September 1998; 21 (9): 1414–1431. https://doi.org/10.2337/diacare.21.9.1414
* Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012). Prediabetes: a high-risk state for diabetes development. Lancet (London, England), 379(9833), 2279–2290. https://doi.org/10.1016/S0140-6736(12)60283-9

<https://www.cdc.gov/diabetes/basics/prediabetes.html>