# **Symptom Significance in Diabetes Diagnosis**

Gabriela Serrano and JuanCarlos Jimenez

CSC 597: Statistical Learning with Applications in R

Dr. Vanessa Aguiar-Pulido

May 2, 2022

## 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 any number of 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 plays an essential part in our bodies' balance (Kaul et al., 2012). In the last 20 years, the number of adults diagnosed with Diabetes has doubled, challenging our health, families, and economy (CDC 2021). DM has grown into an epidemic; unfortunately, projected to continue worsening (King et al., 1998). In the United States alone, 37 million people already have Diabetes, and 96 million adults have Prediabetes (CDC 2021). Of the former, 20% are undiagnosed and of the latter, more than 80% are unaware of their heightened risk for the condition (CDC 2021). Due to the severity of this outlook, patients must have insight into when it becomes 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*** (CDC 2021). Firstly, T1DM occurs when the body has an autoimmune reaction that stops it from producing insulin, the hormone that regulates blood glucose. Exogenous insulin is then needed to maintain healthy levels (CDC 2021). T2DM Diabetes is characterized by the bodies ill use of its insulin (CDC 2021). 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 (CDC 2021). Approximately 90 to 95 percent of people diagnosed have T2DM while 5 to 10 percent have T1DM (CDC 2021). Prediabetes is a predecessor to T2DM, but equally as serious because it also increases your chances of heart disease and stroke (CDC 2021). Lastly, Gestational Diabetes is a special type of Diabetes which presents itself during pregnancy when there was no previous case of diabetes in the mother (CDC 2021). It goes away after birth yet increases risk of T2DM later (CDC 2021). It not only increases risk for the mother, but also for the child. The child will have an increased risk of childhood obesity and is more likely to develop T2DM (CDC 2021).

Taking this into consideration, the objective of this investigation is to find a proactive solution to this issue. We want to reproduce accurate supervised learning models to detect whether a patient has diabetes based on their given set of symptoms. We do this to ultimately, examine our findings by gathering the most indicative predictors of the disease based on a consensus of models that achieve high accuracy. The dataset we plan to work on has already been used to create accurate models for predicting DM - the most accurate being the Random Forest Algorithm, evaluated using 10-fold Cross Validation (Islam et al., 2019). The goal is to produce accurate results that will allow us to examine the importance of predictor variables. The symptoms currently known to be associated with a Diabetes diagnosis are polyuria, sudden weight loss, obesity, polydipsia, and polyphagia (Bettencourt-Silva et al., 2019 & CDC 2022). Additionally, men are more likely to develop DM due to the location in which bodyfat is typically stored -the lower abdomen (CDC 2022). We expect to see these same symptoms in our results, yet we seek to discover which symptoms are the most indicative of the presence of DM and report their order of importance.

## 2. Methods

### Data

The dataset we've used is an early-stage diabetes risk prediction dataset that has been collected using direct questionnaires from the patients of the Sylhet Diabetes Hospital in Sylhet, Bangladesh. The responses of the patients have been overseen by a doctor at the hospital. All in all, it has 16 features and 520 instances. Included as part of our p**redictor variables** are age and gender. **Predictor variables** that are symptoms include presence of *polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity*. Finally, our **target variable** is a *positive or 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. We set the whole dataset to be interpreted as a factor: for the symptoms, 0 represents its absence and 1 its presence, and for sex we had 'Female' and 'Male'. The only variable we kept as an integer was the age of the patient. This means that all the variables we will be working on are categorical except for age, which is numerical.

To visualize how the data was distributed, we generated two histograms, one for age distribution and another for sex distribution. We can observe that the ages in Figure 1 are near a normal distribution, with the most common age being around 40 years old. As for the sex distribution shown in Figure 2, we can observe that there are more male than female patients in our dataset, but they are not the overwhelming majority. Moreover, of the target variable, there are 320 positive and 200 negative diabetes diagnoses, which achieves the balance needed for dependable classification accuracy. Overall, the data is well disbursed.

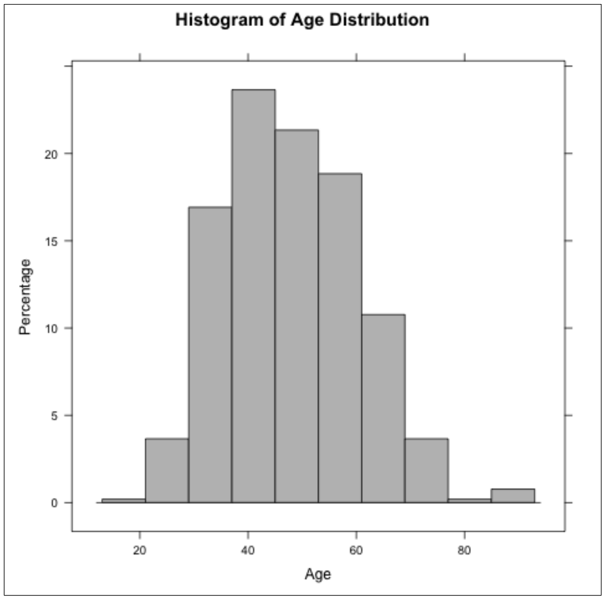

**Figure 1. Histogram of Age Distribution**

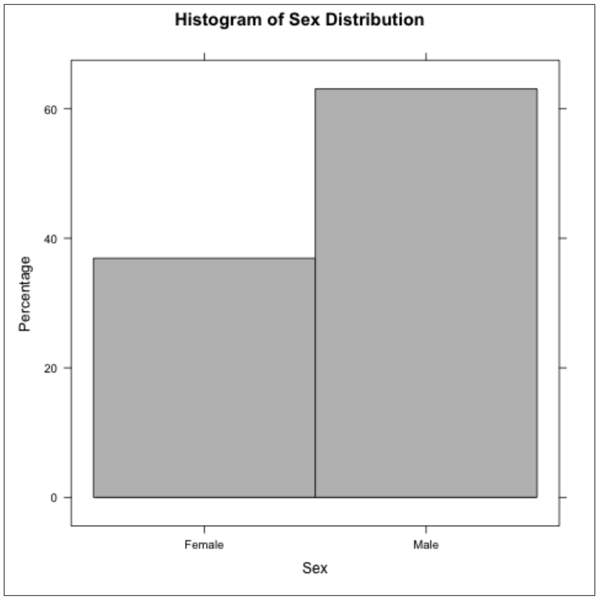

**Figure 2. Histogram of Sex Distribution**

### Techniques

Before applying the methods, we set a random seed generator to be able to reproduce our models. We used the following supervised learning algorithms to build our classifiers:

### Logistic Regression

$$Y_i = B_0 + B_1X_1 + B_2X_2 + \cdots + B_pX_p + \varepsilon$$

Logistic regression is a process of modeling the probability of a discrete outcome given a set of input variables. In our case, the discrete outcome is a binary classification- a positive or negative diagnosis of DM.

### Decision Tree

Decision trees take a set of predictors (the symptoms) as rules to sort observations down the tree from the root to some terminal node or in other words, whether the patient has diabetes or not. They are easy to interpret and identify important variables, as well as make predictions fast. One of the drawbacks is they can have high variability in performance.

### Random Forest

Random Forest uses a collection of decision trees to predict a response. We selected this approach to be part of our project because it does not require the normalization of features and because it handles non-linear parameters efficiently.

### Naïve Bayes Classifier

The algorithm uses Bayes' Theorem with strong (naive) independence assumptions between the features to predict a response.

**Naïve Assumption:** ***P(datapoint | class) = P(feature$_1$ | class) * … * P(feature$_n$ | class)***

It is simple and fast to use but its conditional independence assumption does not always hold. In most situations, and especially real-life (such as a diagnosis), the features usually show some form of dependency.

## Evaluation

For evaluating the performance of classifiers, we used K-fold Cross Validation (k = 20) contained in the caret library in R. This procedure divides a limited dataset into 20 non-overlapping folds. Each of the 20 folds is given an opportunity to be used as a held-back test set, whilst all other folds collectively are used as a training dataset. A total of 20 models are fit and evaluated on the 20 hold-out test sets and the mean performance is reported.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

Doing cross-validation with this specific package has many advantages, such as it is easy to manipulate and indicates a group of settings destined to refine each model. We divided the dataset using 80 percent (416 observations) for training purposes and 20 percent (104 observations) for testing the predictions. Taking the resulting settings from cross-validation into account, we trained each classification model. We generated confusion matrices using the holdout set, to evaluate their performance.

We evaluate using the following metrics: F-1 scores, Accuracy, Sensitivity and Specificity. Using the confusion matrix, which shows the amount of true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) predictions, we calculate the following performance metrics with accuracy being the main metric used to determine the quality of our models.

$$Accuracy = \frac{TP + TN}{TP+FP+FN+TN} \qquad Precision = \frac{TP}{TP + FP}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP}$$

Because our dataset is relatively balanced (320 positive and 200 negative values), we can consider this metric as reliable, because the model does not classify the data for any majority predictor value. F-1 scores were calculated and represent a weighted average of the values of Precision (the extent of error caused by false positives) and Sensitivity. An F1 score reaches its best value at 1 and worst value at 0.

Besides finding the best F-1 score and Accuracy, our aims in finding the best classifying model would require having a high Sensitivity/Recall value. Although having high Specificity would be ideal, it does not need to be prioritized with the same seriousness as we attribute to Sensitivity. This is because verifying a false-positive diagnosis with a physician does not cause long-term harm when compared to receiving the illegitimate reassurance of a false-negative.

Lastly, we will evaluate variable importance through a consensus on the resulting models that are deemed accurate.

## 3. Results

In our classification models, the Random Forest model gives us the best results. We utilized nine variables randomly sampled as candidates at each split (mtry=9) as the configuration used in R. Additionally, we set the number of trees to five hundred (ntree=500) and the number of nodes to be seventy-seven (nrnodes=77). The combination of these settings improved the results on our Random Forest algorithm. Specifically, achieving a perfect value for Sensitivity on our test set.

**Table 1. Classification Results**

| Classification Model | F-1 Score | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| **Logistic Regression** | 0.9157 | 0.9327 | 0.8837 | 0.9672 |
| **Decision Tree** | 0.8913 | 0.9038 | 0.9535 | 0.8689 |
| **Random Forest (default)** | 0.9655 | 0.9712 | 0.9767 | 0.9672 |
| **Random Forest (mtry = 9, ntree = 500, nrnodes = 77)** | 0.9663 | 0.9712 | 1.0000 | 0.9508 |
| **Naïve Bayes (laplace = 0, usekernel = F, adjust = 1)** | 0.8864 | 0.9038 | 0.9070 | 0.9016 |

Furthermore, we utilized varImp() on each model. This function tracks the changes in model statistics for each predictor and accumulates the reduction in the statistic when each predictor's feature is added to the model. We summarized the top six variables on the table presented below.

**Table 2. Variable Importance Results**

| **Logistic Regression** | | **Decision Tree** | | **Random Forest** | | **Naïve Bayes** | |
|---|---|---|---|---|---|---|---|
| **Sex (male)** | **100.00** | **Polydipsia** | **100.00** | **Polyuria** | **100.00** | **Polyuria** | **100.00** |
| **Polydipsia** | **80.76** | **Polyuria** | **86.01** | **Polydipsia** | **71.83** | **Polydipsia** | **96.18** |
| **Polyuria** | **76.84** | **Sex (male)** | **70.85** | **Sex (male)** | **44.30** | **Sudden weight loss** | **67.86** |
| **Itching** | **53.33** | **Sudden weight loss** | **51.60** | **Age** | **23.43** | **Sex** | **67.38** |
| **Genital thrush** | **39.50** | **Partial paresis** | **44.24** | **Alopecia** | **12.10** | **Partial paresis** | **62.93** |
| **Irritability** | **39.33** | **Irritability** | **7.28** | **Sudden weight loss** | **11.12** | **Polyphagia** | **45.75** |

We will continue to examine and analyze the results obtained by all the techniques applied in the following section.

## 4. Discussion

Random Forest was determined to be the best classifier for the diabetes dataset in predicting (based on symptoms mentioned previously) if the patient did or did not have the condition. The top three predictors that determined the presence of DM in the Random Forest model were: polyuria, polydipsia, and the sex of the patient. Considering our findings in Table 2, one can clearly realize the role that Polyuria, Polydipsia, sex, and sudden weight loss (greatest-to-least affect) play in determining whether an individual may or may not have diabetes, regardless of the models used. For Naïve Bayes, Random Forest, Decision Tree, and Logistic Regression classifiers, these symptoms were always present in the 6 most significant predictors. Given the current literature on Diabetes Mellitus, these finding come as no surprise (Bettencourt-Silva et al., 2019 & CDC 2022). What is important to note from this information is the order in which they contribute to the models. Lastly, sex was a top variable of importance for the Naïve bayes, Random Forest, and Decision Tree algorithms which may indicate the need for individualized statistical learning models.

In conclusion, we have achieved the goal of producing an accurate DM diagnosis by training four proven supervised learning algorithms using a specific set of symptoms. Furthermore, we have analyzed dominance of predictors in each of our models and found their respective order of importance in a DM diagnosis. Polyuria, Polydipsia, sex, sudden weight loss, partial Paresis, and Alopecia all carry great importance, as previously found in the literature. Using our findings, we hope to educate patients and physicians on the exact weight that each symptom may or may not play in predicting DM and thus positively contribute to this area of medicine by reducing a prevalent burden on the healthcare system. Future work may include the development of a model that utilizes more detailed patient questionnaires so that the dataset may have additional features.

This dataset with more features will allow the creation of a more robust and informative classifier, by finding the significance of other symptoms that may not be currently closely associated with diabetes. Moreover, an increased instance dataset will also allow for developing statistical learning models that are trained and tested with only one sex’s data and can point to differences in significance of symptoms for each gender. In general, gathering more data and having more observations would help create more reliable statistical models.

## References

Bettencourt-Silva, R., Aguiar, B., Sá-Araújo, V., Barreira, R., Guedes, V., Marques Ribeiro, M. J., Carvalho, D., Östlundh, L., & Paulo, M. S. (2019). Diabetes-related symptoms, acute complications and management of diabetes mellitus of patients who are receiving palliative care: A Protocol for a systematic review. *BMJ Open*, *9*(6). https://doi.org/10.1136/bmjopen-2018-028604

Centers for Disease Control and Prevention. (2021, December 16). *Type 2 diabetes*. Centers for Disease Control and Prevention. Retrieved April 30, 2022, from https://www.cdc.gov/diabetes/basics/type2.html

Centers for Disease Control and Prevention. (2021, December 17). What is diabetes? Centers for Disease Control and Prevention. Retrieved April 30, 2022, from https://www.cdc.gov/diabetes/basics/quick-facts.html

Centers for Disease Control and Prevention. (2021, December 21). *Prediabetes - your chance to prevent type 2 diabetes*. Centers for Disease Control and Prevention. Retrieved April 30, 2022, from https://www.cdc.gov/diabetes/basics/prediabetes.html

Centers for Disease Control and Prevention. (2022, April 5). *Diabetes risk factors*. Centers for Disease Control and Prevention. Retrieved April 30, 2022, from https://www.cdc.gov/diabetes/basics/risk-factors.html

Centers for Disease Control and Prevention. (2022, March 15). *Diabetes and men*. Centers for Disease Control and Prevention. Retrieved April 30, 2022, from https://www.cdc.gov/diabetes/library/features/diabetes-and-men.html

Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood prediction of diabetes at early stage using data mining techniques. *Computer Vision and Machine Intelligence in Medical Image Analysis*, 113–125. https://doi.org/10.1007/978-981-13-8798-2_12

Kaul, K., Tarr, J. M., Ahmad, S. I., Kohner, E. M., & Chibber, R. (2012). Introduction to diabetes mellitus. *Advances in Experimental Medicine and Biology*, 1–11. https://doi.org/10.1007/978-1-4614-5441-0_1

King, H., Aubert, R. E., & Herman, W. H. (1998). Global burden of diabetes, 1995–2025: Prevalence, numerical estimates, and projections. *Diabetes Care*, *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. *The Lancet*, *379*(9833), 2279–2290. https://doi.org/10.1016/s0140-6736(12)60283-9