Classifying Indicators of Diabetes

MIT 5742 – Spring 2023

Avery Goeters, Hadley Bunton, Steven Johnson, Vignan Reddy Putakarapu, and Vishal Reddy Musku

**Table of Contents**

[Problem 3](#_Toc1260862762)

[Description of Data 4](#_Toc1755328583)

[Data Analysis 4](#_Toc734184315)

[Results & Findings 8](#_Toc1176810837)

[Bibliography 11](#_Toc90542329)

[Appendix 12](#_Toc756421525)

# Problem

Diabetes is a chronic health condition where the body is unable to regulate its glucose levels and produce insulin. When food is digested, it is broken down into sugars that are released into the bloodstream (Teboul*,* 2021). As sugars are released, the pancreas releases insulin that is used by cells within the body to convert the sugar into energy (Teboul*,* 2021). A person who has been diagnosed with diabetes is unable to efficiently convert the sugar to energy and must check their blood sugar levels multiple times a day, take medications, and follow other healthy lifestyle choices in order to avoid serious health problems such as heart disease, vision loss, and kidney disease (Center for Disease Control, 2022).

Diabetes is very prevalent, and in many cases, unknown. In the United States alone, over 37 million people have diabetes and, of those, nearly 20% have no idea they have it (Center for Disease Control, 2022). There are three main types of diabetes: Type I, Type II, and Gestational Diabetes (Center for Disease Control, 2022). Type I diabetes accounts for 5-10% of all cases and is believed to be an autoimmune disorder in which the body attacks itself and stops producing insulin. Type II diabetes is the most common type (accounting for 90-95% of all cases) and leads to abnormal levels of blood sugar due to an inefficient use of insulin (Center for Disease Control, 2022). Gestational diabetes occurs during a woman’s pregnancy and can lead to diabetes later in life for the mother and the child (Center for Disease Control, 2022).

As the seventh leading cause of death in the Unites States (Center for Disease Control, 2022), it is imperative to take steps towards preventing the spread of this disease. While there is no cure, we can use data analysis to identify and model the major risk factors of diabetes. Equipped with the results of the analysis, we can then educate the public on the steps that can be taken to minimize the risk of developing diabetes and its impact on a person’s quality of life.

# Description of Data

The Diabetes Health Indicators Dataset was compiled by The Behavioral Risk Factor Surveillance System (BRFSS) which sends out a yearly survey via telephone to over 400,000 Americans collecting information. These individuals are asked questions regarding various health-related “risk behaviors, chronic health conditions, and the use of preventative services” (Teboul, 2021) to determine the variables that represent the most important risk factors leading to diabetes. Our dataset, published on Kaggle, is composed of 70,692 survey responses and represents a balanced population in which half of the population did not have diabetes and the other half had prediabetes or diabetes. It contains 21 independent variables including blood pressure, cholesterol, body mass index (BMI), history of smoking, history of stroke, history of heart disease, physical activity, fruit intake, vegetable intake, and having healthcare (Teboul, 2021) all of which were believed to explain, to some extent, the presence of diabetes in a person’s life.

The dependent variable in this data set is diabetes. A ‘1’ indicates the presence of prediabetes or diabetes whereas a ‘0’ indicates no diabetes. To identify the independent variables that were the greatest predictors of diabetes, analysis was performed using Python, Weka, and Excel. Linear regression, logistic regression, data trees, association rules, and cluster analysis were all utilized to accomplish this with the results and conclusions summarized below.

# Data Analysis

To begin our analysis, we removed five variables (General Health, Mental Health, Physical Health, Any Healthcare, and Income) based on the logic that these variables were too subjective in their questioning or lacked sufficient evidence to contribute to the instance of someone having diabetes.

**Regression**

After removing those variables, we ran both linear and logistic regression using Excel as well as sklearn and statsmodel in Python to find the accuracy of the 16-variable model. This gave us a baseline that we would later compare our new model to. In both regression analyses, we made diabetes the dependent variable and then measured the relationship of each variable to it. The results of the linear regression are in Figure 1. The adjusted R square was 27.45% which means that a significant number of diabetes cases could be explained by the independent variables.

Because our dependent variable was categorical, we had to conduct a logistic regression using the coefficients calculated from the linear regression. The Confusion Matrix for the logistic regression can be seen in Figure 2. Using all 16 independent variables, the model predicted diabetes with 73.38% accuracy, 72.53% precision, and 75.28% recall. The True Positive Rate was 75.28% and the False Positive Rate was 28.52%. These values are all dependent on the cut off value which is used to compare the probability of a correct match. The probability of a correct match is the probability of diabetes if that person is known to have diabetes and is 1 minus that probability if they do not. We chose to cut off at 0.5, because it provided the best balance of accuracy, precision and recall as seen in Figure 3.

**Decision Tree**

Having calculated the accuracy of the 16-variable model, we used the decision tree model to identify the importance of each variable. We used both Weka and sklearn in Python to build these decision trees. We used sklearn’s “DecisionTreeClassifier” to build our model and the “train\_test\_split()” method to train our model in Python. In Weka, we used the “J48” decision tree classifier to perform our analysis. From the decision tree model, we were able to identify the most significant predictors of diabetes by using information gain which reduces the uncertainty (entropy) at each child node of the decision tree.

The results from both Python and Weka tools were found to be consistent in their results. Model accuracy in predicting diabetes was 71.52% and 72.52% respectively (see Figures 4 and 5) and they identified the same significant predictors of diabetes. These variables were access to healthcare (NoDocbcCost), cholesterol check in the past 5 years (CholCheck), high blood pressure (HighBP), gender (Sex), history of stroke (Stroke), smoking (Smoker), consumption of vegetables (Veggies), high cholesterol (HighChol), and body mass index (BMI).

**Association Rules**

Next, we ran our dataset through association rules using aprirori and association\_rules in Python to gain further insight into the variables most significant in predicting diabetes. Before building the model, we had to convert two variables (Age and Education) into binary numbers (either 0 or 1). To convert these into binary numbers, we researched the risk factors of these two variables with respect to diabetes to determine, at what level, these variables became greater risk factors for diabetes (see chart below).

|  |  |
| --- | --- |
| **Independent Variable** | **Lower Risk of diabetes if...** |
| Age | Less than or equal to 45 years old |
| Education | High School Graduate |

Once converted, we ran the associations rules model in Python. The results from the association rules appear in Figure 6. The association rules model identified the greatest predictors of diabetes to be cholesterol check in past 5 years (CholCheck), access to health care coverage (AnyHealthcare), age, high blood pressure (HighBP), consumption of vegetables one or more times daily (Veggies), and age.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Antecedents** | **Consequents** | **Antecedent Support** | **Consequent Support** | **Support** | **Confidence** | **Lift** |
| Diabetes | Cholesterol Check | 0.5 | 0.975 | 0.497 | 0.993 | 1.018 |
| Diabetes | Any Healthcare | 0.5 | 0.955 | 0.480 | 0.960 | 1.005 |
| Diabetes | Age | 0.5 | 0.848 | 0.469 | 0.938 | 1.105 |
| Diabetes | High Blood Pressure | 0.5 | 0.563 | 0.376 | 0.753 | 1.336 |
| Diabetes | Age, Veggies | 0.5 | 0.665 | 0.354 | 0.709 | 1.065 |

**Cluster Analysis**

The last model we ran to determine the most significant variables in predicting diabetes was cluster analysis. We converted all the numeric data type variables into categorical data types (‘T’ or ‘F’ values). We used the “SimpleKMeans” cluster algorithm in the Weka tool for this analysis. We chose the number of clusters as 2 (“cluster 0” and “cluster 1”) because, ideally, one of those clusters would contain the variables that have a high influence on predicting diabetes and the other cluster would contain the variables that do not. As this was an unsupervised learning method, we excluded our dependent/target variable(Diabetes\_binaries).

From this model, we observed that the greatest number of ‘True(T)’ instances for high blood pressure (HighBP), high cholesterol (HighChol), and body mass index (BMI) were clustered in cluster0 and the greatest number of ‘False(F)’ instances for the same variables were clustered in cluster1. We tested the accuracy of our model by including our dependent variable using the ‘classes to clusters' evaluation. ‘Cluster 0’ was assigned to class ‘T’ which indicates the presence of diabetes, and ‘Cluster 1’ was assigned to class ’F’ indicating no diabetes. According to this model, high blood pressure (HighBP), high cholesterol (HighChol), and body mass index (BMI) are the top determinants in predicting the disease. The accuracy score of this model was 68.31%. The results for this model can be seen in figures 7 and 8.

# 

# Results & Findings

**Results Summary**

After running our dataset through the decision tree, association rules, and cluster analysis models, we compared the predictors of diabetes each model produced to see if any overlapped. After this comparison, we identified four variables that appeared most frequently. These four variables were high blood pressure (HighBP), cholesterol check in past 5 years (CholCheck), consumption of vegetables one or more times daily (Veggies), and high cholesterol (HighChol). These four variables became our new model.

Next, we ran a logistic regression and decision tree model on our 4-variable model to see how accurate this new, simplified model would be. The logistic regression analysis (see Figure 9) showed our 4-variable model predicted diabetes with 69.12% accuracy, 67.14% precision, and 74.86% recall. Our True Positive rate was 74.86% and the False Positive rate was 36.63%. The 4-variable model accuracy using the decision tree model was 68.65% (see Figure 10). In both models, our accuracy declined from the 16-variable model by 4.26% and 2.83% respectively. However, while our accuracy was slightly lower, our model adhered to the Parsimony principle and compensated for that decrease in accuracy with simplicity. Our model reduces required inputs by 75%, saving both time and money while achieving nearly the same level of accuracy.

To do one final check on our 4-variable model, we decided to create another model of four randomly chosen variables and predict its accuracy to see how our model compared. Using physical activity (PhysAct), consumption of fruit (Fruits), general health (GenHealth), and heavy alcohol consumption (HvyAlcConsump), this comparison 4-variable model only produced 55.19% accuracy, 46.51% precision and 10.38% recall and had a 10.38% True Positive Rate and a 11.94% False Positive Rate as seen in Figure 11. Our 4-variable model outperformed this randomly selected model giving further credibility to our new model.

**Recommendations & Conclusion**

With the help of data analysis, we were able to identify four variables that are strong predictors of diabetes. While helpful, our goal is to provide actionable steps that individuals can take to prevent and diagnose this deadly disease. Below we have outlined actionable steps that, if taken seriously and applied, can lead to positive results in the fight against diabetes.

Children need to have their cholesterol screened at least once around the age of 10 and once more in their late teens/early 20s. Adults should be checked every 4 to 6 years (Center for Disease Control, 2022). Those who have diabetes or struggle with obesity are recommended to be checked more often. Since high cholesterol does not show symptoms, these tests are very important. The damage from high or low cholesterol happens in your arteries and can result in heart problems or strokes (Center for Disease Control, 2022). To prevent high cholesterol, you can maintain a healthy lifestyle by eating properly, maintaining a healthy weight, and/or getting daily exercise.

Blood pressure is the measure of pressure when blood pushes through your arteries. Your blood pressure can fluctuate throughout the day and a normal rate is around 120/80 mmHg (Center for Disease Control, 2021). Preventing high blood pressure levels can also prevent heart issues and strokes. Like high cholesterol, high blood pressure does not have symptoms, so you need to be checking your blood pressure level often. To prevent this, you can practice a healthy lifestyle, manage stress, keep a healthy weight, and get daily exercise. High blood pressure and cholesterol are very similar and can be prevented in the same ways.

Consumption of vegetables once or more a day is another factor that overlaps with having diabetes or prediabetes. This can be achieved by maintaining a healthy lifestyle and relationship with food. All these factors can be controlled by being more self-aware of what you are doing to take care of your body. In conclusion, diabetes is common and often goes undiagnosed in people who have it. Taking the steps outlined above can help prevent or reduce the negative impacts of this life-altering health condition.

# Bibliography

Center for Disease Control. (2022, October 25). *By the Numbers: Diabetes in America*. Centers for

Disease Control and Prevention. Retrieved February 25, 2023, from

<https://www.cdc.gov/diabetes/health-equity/diabetes-by-the-numbers.html>

Center for Disease Control. (2022, September 30). *Diabetes Fast Facts*. Centers for Disease Control and

Prevention. Retrieved February 27, 2023, from <https://www.cdc.gov/diabetes/basics/quick-facts.html#:~:text=More%20than%2037%20million%20people,(and%20may%20be%20underreported)>.

Centers for Disease Control and Prevention. (2022, October 24). *Get a cholesterol test*. Centers for Disease Control and Prevention. Retrieved March 4, 2023, from <https://www.cdc.gov/cholesterol/cholesterol_screening.htm>

Centers for Disease Control and Prevention. (2021, May 18). *High blood pressure symptoms and causes*. Centers for Disease Control and Prevention. Retrieved March 4, 2023, from <https://www.cdc.gov/bloodpressure/about.htm>

Center for Disease Control. (2022, December 30). *Type 2 Diabetes*. Centers for Disease Control and

Prevention. Retrieved February 25, 2023, from <https://www.cdc.gov/diabetes/basics/type2.html>

Center for Disease Control. (2022, July 7). *What is Diabetes?* Centers for Disease Control and

Prevention. Retrieved February 27, 2023, from <https://www.cdc.gov/diabetes/basics/diabetes.html>

Narayan, K. M. V., Boyle, J. P., Thompson, T. J., Gregg, E. W., & Williamson, D. F. (2007, June 1). *Effect of BMI on Lifetime Risk for Diabetes in the U.S.* Diabetesjournals.org. Retrieved February 24, 2023, from <https://diabetesjournals.org/care/article/30/6/1562/30745/Effect-of-BMI-on-Lifetime-Risk-for-Diabetes-in-the>

Teboul, A. (2021, November 8). *Diabetes health indicators dataset*. Kaggle. Retrieved February 27, 2023,

from [https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset resource=download&select=diabetes\_012\_health\_indicators\_BRFSS2015.csv](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset%20resource=download&select=diabetes_012_health_indicators_BRFSS2015.csv)

YouTube. (2020, October 20). *Logistic regression for excel*. YouTube. Retrieved February 25, 2023, from

<https://www.youtube.com/watch?v=xKA-fOCENaY>

Appendix

**Figure 1: Linear Regression Analysis - Output**



**Figure 2: Logistic Regression Analysis – Confusion Matrix for 16 Variables**



**Figure 3: Logistic Regression – 0.5 Cutoff**

**Figure 4: Python Decision Tree – Accuracy**

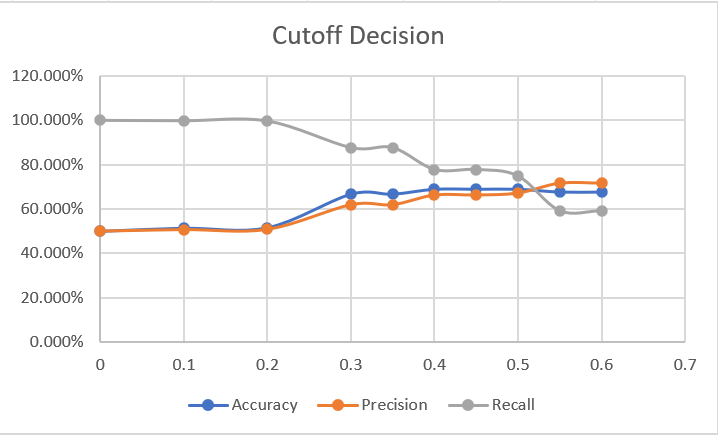


Figure 3

Graphical user interface, text, application, email

Description automatically generated

**Figure 5: Weka Decision Tree – Accuracy**

Text

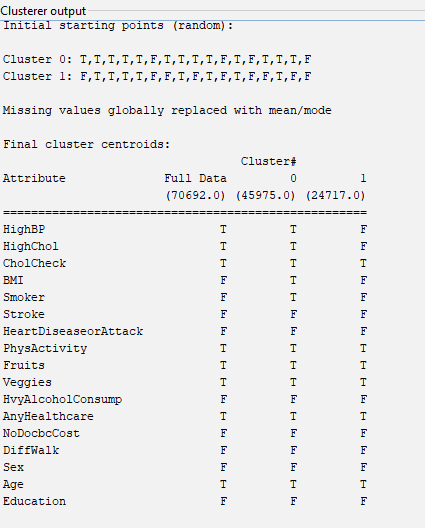
Description automatically generated

**Figure 6: Association Rules – Results**Table, Excel

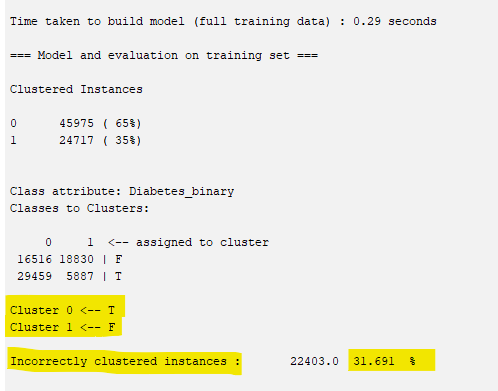
Description automatically generated**Table

Description automatically generated**

**Figure 7: Weka Cluster Analysis – Final Cluster Centroids**



**Figure 8: Weka Cluster Analysis – Accuracy of the Model**



**Figure 9: Logistic Regression – Accuracy of New Model (4 Variables)**



**Figure 10: Python Decision Tree – Accuracy of New Model (4 Variables)**

**Graphical user interface, text, application

Description automatically generated**

**Figure 11: Logistic Regression – Accuracy of Random Variable Model (4 Variables)**

