**Discovering the Predictive Attributes of Diabetes Mellitus in the Pima Indian Population**

Maurica Nez

Colorado State University Global

MIS581: Business Intelligence and Data Analytics

Instructor: Dr. Morad

Due Date: 10/1/2023

Module 7: Portfolio Milestone 2

**Abstract**

Diabetes is a disease ailing and killing millions of people around the world every year. Medical costs worldwide are projected to reach $627.3 billion by 2035 (Zhang & Liu, 2022). Type II diabetes mellitus may be preventable. This research investigates the predictive ability of variables from the Pima Indians Diabetes data set using data mining techniques. What variables are correlated with diabetes? What variables can predict diabetes? Using statistical analysis software, this research found correlations with 4 variables. The predictive analysis confirmed the correlations. A classification tree model selected the same 4 variables between both partitions of the data. The predictive model can determine the presence of disease based on glucose levels, pregnancy, age, and insulin levels. This data was limited to a few hundred participants once the data was cleaned. A larger data set would be ideal for further analysis.

**INTRODUCTION**

The World’s Health Organization has declared diabetes to be a serious global issue (WHO, 2016). Diabetes Mellitus is one of the top four “priority noncommunicable diseases (NCDs) targeted for action by world leaders” (WHO, 2016, p. 6). Diabetes mellitus “is the leading cause of kidney failure” (Vital Signs, 2018, p. 1). In 2012, diabetes killed 1.5 million people and was associated with an additional “2.2 million deaths, by increasing the risks of cardiovascular and other diseases” (WHO, 2016, p.6). The instances of diabetes in the world population have almost quadrupled “since 1980” (WHO, 2016, p. 6).

The pathology of diabetes mellitus is dependent on the type. Cells rely on basic sugar to provide energy to operate at optimum levels. Insulin facilitates the transport of sugar into the cell. The absence of insulin within the cells prevents cells from receiving sugar and the cells die due to lack of energy (Klandorf & Stark, 2022). Sugar has limited places to go, and the kidneys become the primary mechanism for the body to remove the sugar from the blood stream. The kidneys then become overtaxed and cannot effectively perform their primary function to regulate the blood. Once the kidneys succumb to the disease, treatment options are limited to kidney dialysis or transplant (Vital Signs, 2018).

There are three types of diabetes mellitus. Type 1 diabetes mellitus is not preventable because it is hereditary and is triggered by environmental factors (Klandorf & Stark, 2022). Children have the highest instances of type 1 diabetes mellitus. The mechanism is autoimmune related where the insulin producing cells in the pancreas are destroyed by the body’s T lymphocytes (Klandorf & Start, 2022). Insulin is no longer produced, and sugar is unable to transport into the cell. There is no cure for type 1 diabetes mellitus. However, it is treatable with insulin injections (Klandorf & Stark, 2018).

Type 2 diabetes mellitus is the opposite of type 1 and is more common. Instead of an absence of insulin, too much insulin is produced causing the cells to reduce the number of receptors (Klandorf & Stark, 2022). The cells become insulin resistant due to the lack of transport receptors. The initial symptoms are less pronounced when compared to type 1 and early detection is difficult (Chang et al., 2022). Therefore, type 2 has been labeled “the silent killer” (Klandorf & Stark, 2022, p. 1). Type 2 diabetes does not have a cure, but it is treatable. The advancement of type 2 diabetes mellitus symptoms may be slowed, and kidney failure may be preventable (Chang et al., 2022). However, the difficulty lies in the timing of discovering the condition because it may take a decade to present with symptoms (Abbasi et al., 2016). Treatment for type 2 diabetes mellitus is more effective the earlier that it is discovered (WHO, 2016).

Gestational diabetes is a temporary condition that appears during pregnancy. The risk for developing type 2 diabetes mellitus later increases for those who have suffered from gestational diabetes. Moreover, the child from the affected pregnancy also has increased risk of developing type 2 diabetes during childhood and suffering from obesity (Klandorf & Stark, 2022).

Studies indicate that the native American population has the highest instances of diabetes when compared to other populations (Vital Signs, 2018). A possible explanation includes genetic predisposition, a lack of healthcare, education, and the cultural food paradigm. The Native American population is in crisis attempting to manage the treatment for individuals affected because access to healthcare is expensive and limited. Understanding the condition and trusting treatment is another barrier that must be overcome to improve the situation. Within the Native American tribes, food is linked to family and celebration of togetherness. The food often prepared is what their ancestors were fed when held in captivity from commodities provided by the American government. The commodities became considered their traditional food. Eating poorly became a generational learned behavior.

Diabetes Mellitus is a global problem that has no cure (Anonymous, 2002). However, the earlier the discovery, the more positive the prognosis (WHO, 2016). Treatment is possible for type 2 diabetes mellitus. However, finding a means to prevent the condition would be ideal (Anonymous, 2002). Identifying the risk factors of Type 2 Diabetes Mellitus is a preliminary step in the research process. Any research that supports combating a worldwide problem is worth pursuing.

The Pima Indian Diabetes data set is selected to represent the Native American tribes affected by diabetes mellitus. The tribe has been greatly impacted by the effects of the disease and has data readily available for research purposes.

**OBJECTIVES**

* To identify what risk factors may be used to predict the occurrence of Type 2 Diabetes Mellitus.
* To explore the variables of the data set.
* To discover what variables are correlated with the outcome variable.
* To determine if the correlated variables can predict the outcome variable.
* To determine what additional variables would be helpful for future analysis.
* To identify what variable may accommodate preventative measures a community may take to address the determined variables.

**OVERVIEW OF STUDY**

This research will use data mining techniques to investigate the variables of the Pima Indians Diabetes data set. Descriptive and predictive analysis will be performed using three different analytical tools. The distribution of the data will be analyzed using Tableau and SAS. The R programming language will be used to clean, partition the data, and to create a classification tree. The train and test model will be compared to one another using sensitivity and specificity testing. The correlation matrix and models will identify what variables can predict diabetes by answering each research question and determine support for hypothesis statements.

**Research Questions**

This research seeks to determine what variables in the data set have a correlation and predictive relationship with the outcome variable. Both research questions will explore the nature of the relationship between the independent and dependent variables.

1. What independent variables have a correlation with the outcome variable?
2. What independent variables predict the outcome variable?

**Hypothesis**

The hypotheses questions will investigate the research questions as follows. The first set of hypotheses statements will determine which variables have a correlation with the outcome variable.

H1: The independent variables will not have a positive correlation coefficient with the dependent outcome variable.

H2: The independent variables will have a positive correlation coefficient with the dependent outcome variable.

The second set of hypotheses statements will determine if the variables that were found to have a correlation with the outcome variable may be used to predict the outcome variable.

H1: None of the independent variables will predict the dependent variable.

H2: An independent variable will be found to predict the dependent variable.

**LITERATURE REVIEW**

Due to how extensively type 2 diabetes mellitus has ailed people around the globe, researchers have prioritized finding a solution. Previous studies have explored various methods that may be used to predict and diagnose type 2 diabetes mellitus. Recent studies demonstrate a shift in the methods of diabetes research moving towards machine learning and data mining techniques.

One study, found in the *Neural computing & applications,* was conducted within the last year, and made headway in improving patient access to diagnostic tests. The study sought to establish a means for using machine learning techniques for “an e-diagnosis system” that can diagnose type 2 diabetes mellitus (Chang et al., 2022, p. 1). They take it a step further to fully describe how machine learning works to gain credibility within the medical sector. Machine learning has not been accepted within the medical community because “the internal decision-making process” is not trusted (Chang et al., 2022, p 1).

To combat the mistrust, the researchers used R coding to examine the Pima Indians diabetes dataset and build models that are easy to interpret. The diagnostic machine learning models used are Naïve Bayes classifier, random forest classifier, and J48 decision tree. The best model was selected based on the “accuracy, precision, sensitivity, and specificity” (Chang et al., 2022, p. 1). They found the Naïve Bayes model is suitable for binary classification when the features are more refined. The random forest model was better when more features were required.

The authors intend to streamline the diagnostic process where an electronic medical record may be updated by the patient at home with mobile diagnostic equipment. The patient’s doctor may use the e-diagnostic system to assess the patient’s condition remotely. An in-person appointment would no longer be necessary and would provide medical care to people living in remote areas such as the participants in the dataset.

A study from 2023 shows that machine and deep learning methods are beginning to take the lead in research. The techniques Deepa and Kumar used were Support Vector Machine, Artificial Neural Networks, K-nearest neighbor, and logistic regression. Their study sought to find out “how to rapidly and effectively identify and assess diabetes” (Deepa & Kumar, 2023, p. 2). They also utilized the Pima Indians Diabetes Data Set along with others to find the most effective predictive model. As with the previously mentioned study, they found that each model had its merit, and the performance was dependent on the data used. They concluded that future studies use “more samples as well as further machine learning and data mining techniques” (Deepa & Kumar, 2023, p. 12).

Another research team surveyed the research at the time and sought to “link data assessment to diagnosis and appropriate decision-making” for treatment of diabetes mellitus (Kavakiotis et al., 2017, p. 4). They stated that there is “an evident gap in research on diabetes with respect to data mining and machine learning” (Kavakiotis et al., 2017, p. 10). Their survey of the literature confirmed that the main biomarker for diagnosis is blood glucose levels.

Another study, found in the *BMC endocrine disorders* publication*,* conducted an extensive survey of studies that compiled data from over 14,000 patients with type 2 diabetes mellitus. Blood glucose levels have previously been the main indicator for type 2 diabetes mellitus. This study sought to “identify all the metabolites… that may be useful for the diagnosis or treatment of diabetes (Long et al., 2020, p.2). A comparison of the metabolite biomarkers demonstrated a significant difference between patients classified as pre-diabetes and patients with type 2 diabetes. Indicating that predicting the onset of the disease may be possible.

**RESEARCH DESIGN:**

**Methodology**

The data set contains the “diagnostic measurements” that are associated with diabetes mellitus (Kaggle, 2016, p. 1). There are nine variables in the Pima Indians Diabetes dataset. The dataset contains 768 rows that document the variables for each participant in the study that collected the data. All participants in the study “are females at least 21 years old of Pima Indian heritage” (Kaggle, 2016, p. 1). There are 768 rows in the data set. One row for each patient in the study.

Three tools will be used to conduct the data analysis portion of this study. Tableau will provide the visuals for the descriptive statistical evaluation of the variables in the data set. RStudio will be used to clean and partition the data set into a training and test set. For the predictive statistical analysis, a variable correlation, logistic regression, and confusion matrix will be created in RStudio to create the model and determine the accuracy using the partitioned data. Enterprise Miner will be the final tool used. A classification tree will be created and compared to the results of the logistic regression to confirm the results of both models.

The nine variables in the Pima Indian Diabetes data set are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome. The Pregnancies variable is an integer and records the number of pregnancies a woman has had. The Glucose variable is an integer that documents the concentration of sugar in the blood. The Blood Pressure variable is an integer that records the patient’s blood pressure. The Skin Thickness variable is an integer that provides the measurement for the patient's skin thickness. The Insulin variable is an integer that documents the concentration of insulin in the patient’s blood. The BMI variable is a floating (has decimal points) variable that documents the patient’s body mass index score. The Diabetes Pedigree Function variable is a floating variable that calculates the possibility of the patient developing diabetes when considering the patient’s family history. The Age variable is an integer that accounts for the patient's age at the time of the study. The Outcome variable is binary and records whether the patient suffers from diabetes (1) or not (0). Figure 1 displays additional details for each variable included in the data dictionary (R-Project, n.d.).

Figure 1

*Data Dictionary*

*A table of medical data

Description automatically generated*

**Methods**

The descriptive analysis will be conducted in Tableau and SAS to determine the data distribution visualizing histograms. To test the first set of hypotheses statements, a correlation analysis will be conducted using RStudio to determine the correlation coefficient value of each independent variable. For the second set of hypotheses statements, the train and test partitions will be used to create a classification tree using RStudio. A confusion matrix will test the sensitivity and specificity of the models. The variables the classification tree selects will be compared to the variables found in the correlation matrix to determine which variables predict diabetes.

**Limitations**

The data set contains variables that have been examined using various methods. However, the literature does encourage the continued study of the variables using multiple models. Using data mining techniques appears to be relatively new for analyzing the data set. Very few articles referenced data mining and machine learning techniques and such research was relatively new. More models, combinations of models, more data sets, and different variables need to be explored for future research.

**Ethical Considerations**

The primary ethical concern of this project is that the participants provided informed consent that allowed the use of their data (O’Leary, 2017). There is no indication in the data set documentation that the participants were uninformed (R-project, 2023). The data set does not provide any participant identification. For the purposes of this study, the participants are anonymous. Therefore, the ethical principle of anonymity is upheld (O’Leary, 2017).

**FINDINGS**

The summary statistics are displayed in Figure 2. The mean for the number of pregnancies of the participants is 3.3. The number of pregnancies for participants ranged from 0 to 17. The mean glucose value is 122.6. The glucose value ranged from 56 to 198. The mean blood pressure is 70.66 with a variable range of 24 to 110. The mean for the triceps skinfold is 29.15 with a variable range of 7 to 63. The mean for insulin is 156.06 with a variable range of 14 to 846. The mean for body mass index is 33.09 with a variable range of 18.20 to 67.10. The mean for pedigree is 0.52 with a variable range of 0.08 to 2.42. The mean age is 30.86 with a variable range of 21 to 81. 262 participants did not have diabetes and 130 did have diabetes.

Figure 2

*Summary Statistics of Pima Indians Diabetes Data Set*

A screenshot of a computer

Description automatically generated

Figure 3 displays the correlation matrix for the data set. Values closest to positive 1 are selected. Glucose and insulin have a correlation value of 0.6. Age and pregnancy had a correlation value of 0.7. Therefore, the first null hypothesis is not supported. Figure 4 displays the distribution of the Age variable. The Age data is skewed to the right. Figure 5 displays the distribution of the Glucose variable. The Glucose variable is skewed to the right. Figure 6 displays the distribution of the Insulin variable. The data for Insulin is skewed to the right. Figure 7 displays the distribution of Pregnancies. The Pregnancy variable is skewed to the right.

Figure 3

*Correlation Matrix*

A screenshot of a computer

Description automatically generated

Figure 4

*Distribution for Age*

A graph with many blue bars

Description automatically generated with medium confidence

Figure 5

*Distribution of Glucose*

A screenshot of a graph

Description automatically generated

Figure 6

*Distribution of the Insulin Variable*

A screenshot of a graph

Description automatically generated

Figure 7

*Distribution of Pregnancies*

*A screen shot of a graph

Description automatically generated*

There are two versions of the Classification Tree model. Figure 8 displays the variables the model selected using the train partition of the data set. The model displays the glucose, insulin, age, and pregnancy status that will determine the diagnosis of diabetes mellitus. Figure 9 displays the model constructed with the test partition of the data set. The glucose, age, and pregnancy status were the final variables selected to predict the diabetes status. The sensitivity value for the Train model was 0.73 (Figure 10). The specificity value for the train model was .91 (Figure 10). For the test model, the sensitivity improved with a value of 0.75 (Figure 11). The specificity was reduced to 0.89 (Figure 11). The variables selected by the Classification Tree model align with the variables found to have high correlation values. Therefore, the second null hypothesis is not supported.

Figure 8

*Classification Tree from Train Data*

A screenshot of a computer

Description automatically generated

Figure 9

*Classification Tree from Test Data*

A screenshot of a computer

Description automatically generated

Figure 10

*Confusion Matrix, Sensitivity, and Specificity Testing for the Train Partition Tree*

A screenshot of a computer code

Description automatically generated

Figure 11

*Confusion Matrix, Sensitivity, and Specificity Testing for the Test Partition Tree*

A screenshot of a computer code

Description automatically generated

**CONCLUSION**

Medical professionals may find the outcome of this study useful because data mining can be used to predict diabetes. The symptoms of diabetes mellitus are known. Symptoms appear slowly after years of developing the disease. Treating the patient prior to the onset of symptoms may help improve the patient’s prognosis. The variables in the data set may be used to determine which patients have diabetes mellitus. The data analysis may provide a foundation or reference for future studies.

Diabetes is a pandemic with many complex contributing factors that require analysis. The minority populations are hit hard and do not have the resources available to effectively combat the disease. A fresh perspective and an alternative form of analysis may provide new insights or lend credibility to previous analytical studies of the Pima Indian Diabetes data set. The classification tree model was useful in determining which variables in the data set predict diabetes. Insulin levels, Age, Pregnancy, and Glucose levels may be used to determine the risk of developing type 2 diabetes mellitus.

**RECOMMENDATIONS**

Further research using data mining techniques may be valuable in providing more evidence that diabetes is predictable. Comparison models, different partition parameters, and accuracy tests may be used to explore the Pima Indians Diabetes data set. In the future, an alternative approach and a different data set may be used to add to the overall research. For example, a data set containing metabolite biomarkers could explore the topic at a more intricate level using data analysis (Long et al., 2020). There are many variables associated with diabetes that merit the continued investigation using machine learning, artificial intelligence, and any other form of analysis (Luo et al., 2023).

**REFERENCES**

Abbasi, A., Sahlqvist, A.-S., Lotta, L., Brosnan, J. M., Vollenweider, P., Giabbanelli, P., Nunez, D. J., Waterworth, D., Scott, R. A., Langenberg, C., & Wareham, N. J. (2016). A Systematic Review of Biomarkers and Risk of Incident Type 2 Diabetes: An Overview of Epidemiological, Prediction and Aetiological Research Literature. *PLoS ONE*, *11*(10), 120. https://doi.org/10.1371/journal.pone.0163721

Anonymous. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or met form in. *The New England Journal of Medicine, 346*(6), 393-403. [https://csuglobal.idm.oclc.org/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fsc arlyjournals%2Freduction-incidence-type-](https://csuglobal.idm.oclc.org/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fsc%09arlyjournals%2Freduction-incidence-type-) diabeteswith%2Fdocview%2F223937844%2Fse2%3Faccountid%3D38569

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural computing & applications*, 1–17. Advance online publication. <https://doi.org/10.1007/s00521-022-07049-z>

Deepa, K., & Ranjeeth Kumar, C. (2023). Early diagnosis of diabetes mellitus using data mining and machine learning techniques. *Journal of Intelligent & Fuzzy Systems, 44*(3), 3999–4011. https://doi.org/10.3233/JIFS-222574

Kaggle. (2016). Pima Indians Diabetes Data Set. Retrieved from <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download>

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Klandorf, H., Stark, S. W., (2022). Diabetes mellitus. Magill’s Medical Guide (Online Edition). <https://eds>s-ebscohost.com.csuglobal.idm.oclc.org/eds/detail/detail?vid=1&sid=

3d6f0d06-8b3d-402aa0a7980652984094%40redis&bdata=JnNpdGU9ZWRzLWxpdmU%3d#AN=

86194054&db=ers

Long, J., Yang, Z., Wang, L., Han, Y., Peng, C., Yan, C., & Yan, D. (2020). Metabolite biomarkers of type 2 diabetes mellitus and pre-diabetes: a systematic review and meta-analysis. *BMC endocrine disorders*, *20*(1), 174. <https://doi.org/10.1186/s12902-020-00653-x>

Luo, X., Sun, J., Pan, H., Zhou, D., Huang, P., Tang, J., Shi, R., Ye, H., Zhao, Y., & Zhang, A. (2023). Establishment and health management application of a prediction model for high-risk complication combination of type 2 diabetes mellitus based on data mining. *PLoS ONE, 17*(8), 1 18. https://doi.org/10.1371/journal.pone.0289749

R-Project. (n.d.). Pima Indians Diabetes Database. R Documentation. Retrieved from

<https://search.r-project.org/CRAN/refmans/mlbench/html/PimaIndiansDiabetes.html>

Vital Signs. (2018). Native Americans with Diabetes: Better diabetes care cand decrease kidney disease. *Centers for Disease Control and Prevention.* Retrieved from <https://www.cdc.gov/vitalsigns/aian-diabetes/index.html>

Worlds Health Organization. (2016). *Global Report on Diabetes.* <https://www.who.int/publications/i/item/9789241565257>

Zhang, L., & Liu, M. (2022). Analysis of Diabetes Disease Risk Prediction and Diabetes Medication Pattern Based on Data Mining. *Computational & Mathematical Methods in Medicine*, 1–9. https://doi.org/10.1155/2022/2665339