Quantitative Analysis in the Medical Field

Using K-NN and Decision Trees to Predict Probabilities of the Diagnosis of Diabetes

Braden C. Drewery, *University of North Alabama*

Dec 7, 2024

Using Logistic Regression and Decision Trees to Predict Probabilities of the Diagnosis of Diabetes

Braden C. Drewery

1 Introduction

This paper is based on data within a dataset containing only female patients. I will use K-NN and decision trees to develop an intuitive model to predict the diagnosis of these female patients having diabetes. This diagnosis will be judged on health metrics. I have health metrics on the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI (body mass index), diabetes pedigree function, and age. These health metrics will be designated as the predictor variables that will assist me in predicting my predicted variable labeled outcome. All predictor variables are numeric variables, but the predicted variable is binary. The variable labeled outcome represents whether the patient is diagnosed with diabetes.

These predictor variables are crucial in determining the outcome of the diagnosis of diabetes in a patient. The number of female pregnancies is telling because of the risk of developing gestational diabetes, typically appearing during the second or third trimester. Gestational diabetes will also increase the patient's risk of developing Type 2 diabetes. Glucose levels will be maintained at normal levels naturally when the patient does not have diabetes. With diabetes, glucose levels will rise higher than optimal. However, medications for diabetes can cause the glucose levels to be lower than optimal. (Clinical Diabetes & American Diabetes Association, 2018). Blood pressure alone is not a factor that can be used to diagnose diabetes, but people with diabetes are twice as likely to have high blood pressure. (Diabetes and High Blood Pressure, 2023). Skin thickness is a predictor variable for diabetes because it is higher depending on the duration of diabetes. It has been found that there are connective tissue changes in diabetes mellitus (diabetes). Diabetes mellitus results in hyperglycemia when there are significantly high blood sugar levels in someone that would typically be controlled in someone who naturally produces insulin. (Collier et al., 1989).

Insulin is a necessary factor that goes into finding out if someone has diabetes or not. With type 1 diabetes, the cells of the pancreas stop producing insulin. These people would have to inject insulin for the rest of their lives. However, with type 2 diabetes, where the pancreas doesn't make enough insulin, or it isn't as efficient as it should be. This is known as insulin resistance. (Department of Health & Human Service, n.d., 2021). BMI is a risk factor for the diagnosis of diabetes as well. Studies have shown that higher BMIs are associated with a higher risk of developing type 2 diabetes complications. (Southern Medical Association, 2023). The Diabetes Pedigree Function is the following predictor variable extremely important for the diagnosis. It measures the genetic risk for developing diabetes, and the risk is mostly correlated to type 2 diabetes. (Akmeşe, 2022). The final predictor variable in the dataset is age. The likelihood of developing type 2 diabetes increases as you get older. People over 45 maintain a higher risk of developing type 2 diabetes, but it is becoming more prevalent for diagnosis at a younger age. (Cherney, 2022). All predictor variables listed prior are compiled in one dataset. The dataset that I used to try to predict the probability of a diabetes diagnosis is from Kaggle. (*PIMA Indians Diabetes Database*, 2016). Kaggle is a website that holds a multitude of public datasets. The data was originally collected from the National Institute of Diabetes and Digestive and Kidney Diseases. There is a public notebook for the dataset as well, and an acknowledgment of the ADAP learning algorithm used in the original dataset for predicting the diabetes diagnosis through a citation. The dataset contains data from the 1980s, and the original article was published on November 9, 1988. (Smith et al., 1988). I would say that my data source is very reliable and reputable. All predictor variables in the dataset are numeric. The outcome or predicted variable is a binary numeric variable. The diagnosis of whether the patient has diabetes will be 0 or 1. The patient would have a result of 0 if they weren't found to have diabetes. The patient would then have a result of 1 if they were found to have diabetes. Later, we will treat this variable as a nominal one. The dataset contains 768 total instances or observations. The eight variables that are used to predict the outcome variable seem to be extensive enough to provide a good means to achieve high accuracy.

While the previously discussed predictor variables are irreplaceable, there are crucial input variables that could've been included. Someone could have a history of being a smoker. Smoking cigarettes has been proven to increase the risk of someone developing type 2 diabetes. (*Smoking and Diabetes*, 2022). Abnormal triglyceride or HDL cholesterol levels could be a telltale sign of a positive diagnosis. (Bitzur et al., 2009). Specific ethnic backgrounds have also been linked to having a higher risk of having diabetes. Plenty of public databases contain studies and a multitude of data to support the idea that these input variables are risk factors for diabetes. There are some missing points in the dataset. All missing values have been set to 0 automatically. It is easy to pick apart which ones would be missing values due to many of the variables not being able to be 0 naturally. Most of the variables would result in the patient being dead if they were at 0 (BMI, Insulin, SkinThickness, BloodPressure, Glucose). This means we should compute some measure of central tendency to replace all the missing values for my attributes. The means of each of the attributes would be the best measure of the central tendency to implement for replacing these missing values.

The question is whether we can accurately predict the diagnosis of diabetes purely based on health metrics. My family has a history of diabetes, and it would be beneficial to know the risk factors most associated with a positive diagnosis to ensure the highest chance of preventing it in the future generations of my family. The genetic passing of diabetes would be nearly impossible to avoid, but the other risk factors are manageable or trackable. The supervised learning models that I will be implementing to make my predictions are K-NN and decision trees. These models should provide me with the data needed to proceed and identify the most important risk factors for diagnosing diabetes accurately.

2 Data Analysis

There is more depth to the variables in the backend helping with the diagnosis's accuracy. There are 652 total missing values in the original dataset, but they were all initialized to zero. The missing values are scattered amongst the attributes: Glucose (5), BloodPressure (35), SkinThickness (227), Insulin (374), and BMI (11). Since there are many missing values, you can't just disregard those observations from your dataset. We will fix the missing value issue by finding a central tendency of all the attributes that contain 0 values where 0 is illogical. The central tendency that makes the most sense with my data would be the mean. All values that are 0 within these attributes will be substituted with the mean found within it. All 0s will be temporarily taken out when finding the mean of the rest of the observations to ensure the data is more accurate. The number of pregnancies ranges from 0 to 17, and the higher the number of pregnancies, the higher the risk of gestational or type 2 diabetes. There are some outliers within my variables, like the 17 max pregnancies listed in the previous sentence, and K-NN will sort out those outliers to ensure a more accurate prediction for future observations. K-NN is an instance-based algorithm that finds the nearest instances to an observation where the number of nearest looked at depends on the set value of k for that iteration. Knowing this, K-NN will focus on closer results and outliers will not be recognized as a nearby neighbor once we identify the ideal k value for our dataset. The glucose variable tells you the plasma glucose concentration in an observation after an oral glucose tolerance test.

There is a normal range for glucose levels, and someone may be at risk of diabetes if their levels do not fit within the normal range. The BloodPressure variable measures just the Diastolic reading on a blood pressure test. The SkinThickness variable refers to the skinfold that is produced with the triceps, and the skinfold would be measured in millimeters. Then, the Insulin variable takes a 2-hour Serum and tests post-glucose insulin levels. The purpose of this 2-hour test is to attempt to predict insulin resistance. Insulin resistance can show that someone may be at risk of diabetes. The BMI variable divides your weight by your height and squares the result. BMI is a very rough way of finding body fat composition, but the higher BMIs are at a higher risk of being diagnosed with diabetes. The DiabetesPedigreeFunction is a variable that provides an idea of the diabetes history in relatives and the genetic relation you may have to those relatives within the scope. My age variable is just tracking the age of the patients. People aged above 45 have a higher risk of developing diabetes. My final variable is my outcome variable which tells me whether the person has diabetes.

Table 1: Data Description

Variable Name	Description	Range
Pregnancies	Number of pregnancies	{0,17}
Glucose	Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test (GTT)	{44,199}
BloodPressure	Diastolic Blood Pressure (mm Hg)	{24,122}
SkinThickness	Triceps Skin Fold Thickness (mm)	{7,99}
Insulin	2-Hour Serum Insulin (µU/ml)	{14,846}
BMI	(Weight in kg / Height in m) ^2	{18.2,67.1)
DiabetesPedigreeFunction	Genetic Risk Formula	{0.078,2.42)
Age	Years	{21,81}

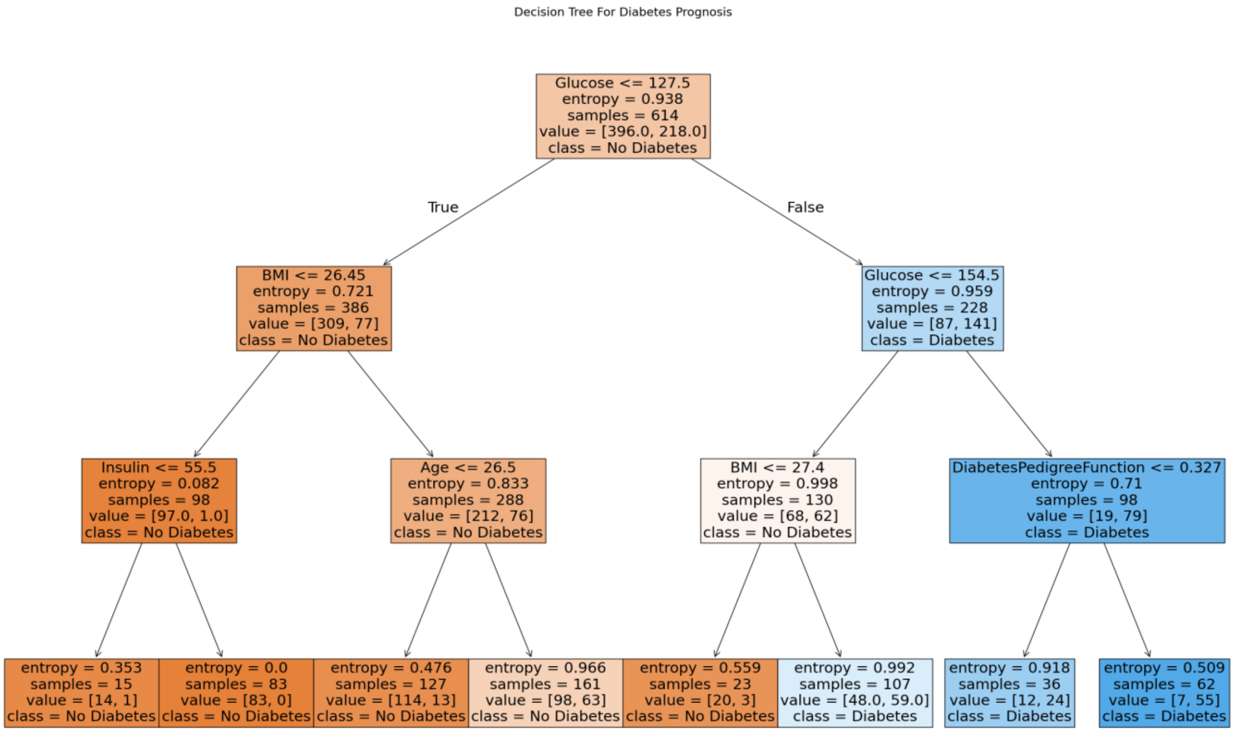
Outcome	0 if Diabetes Diagnosis is Negative / 1 if Diabetes Diagnosis is Positive	Binary {0,1}

Table 2: Summary Table

	Count	Mean	Std	Min	25%	50%	75%	Max	Median
Pregnancies	768	3.85	3.37	0	1	3	6	17	3
Glucose	768	121.69	30.44	44	99.75	117	140.25	199	117
Blood Pressure	768	72.39	12.10	24	64	72	80	122	72
Skin Thickness	768	29.11	8.79	7	25	29	32	99	29
Insulin	768	155.77	85.02	14	121.5	156	156	846	156
BMI	768	32.46	6.88	18.2	27.5	32.4	36.6	67.1	32.4
Diabetes Pedigree Function	768	0.47	0.33	0.08	0.24	0.37	0.63	2.42	0.37
Age	768	33.24	11.76	21	24	29	41	81	29
No Diabetes	500	-	-	-	-	-	-	-	-
Diabetes	268	-	-	-	-	-	-	-	-

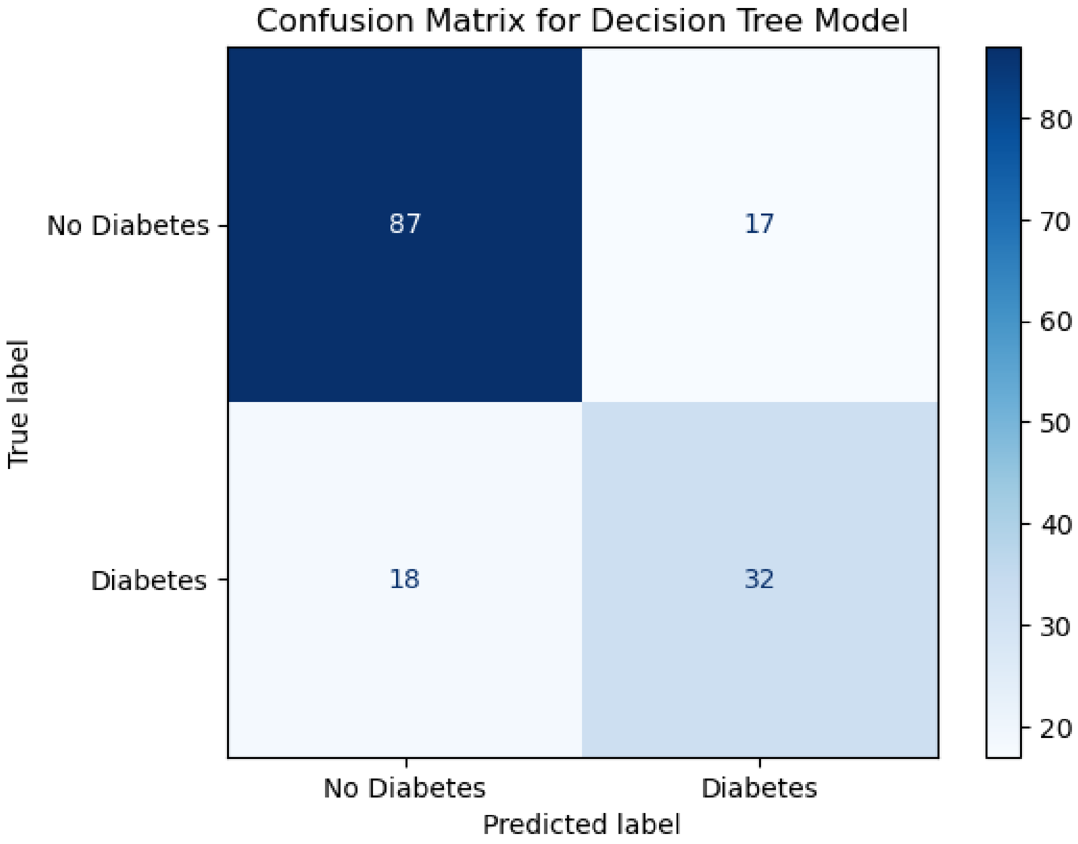
The summary table above highlights the key differences between my attributes. The end of the summary table contains the count of patients with diabetes, but there is also a count for the patients that do not have diabetes. The summary table highlights the need for normalization since the scales across the input variables vary. For example, the Diabetes Pedigree Function is on a much lower scale than Glucose levels. To ensure there is no bias when calculating the Euclidean distance for K-NN, putting the attributes on the same or similar scale is essential unless you want one attribute to be weighted more. For predicting diabetes, there are specific signals for individuals who may be more prone to receiving a positive diagnosis. High glucose levels, higher number of pregnancies, high blood pressure, high insulin levels, high BMI, higher pedigree score, and older age are all heavy risk factors. The table, paired with the information stated in the previous sentence, shows that individuals with values above 140.25 for glucose levels, diastolic blood pressure greater than 84, insulin levels above 156, BMI above 36.6, a pedigree score above 0.63, and age above 41 are at a high risk of having diabetes.

Chart 1: Decision Tree



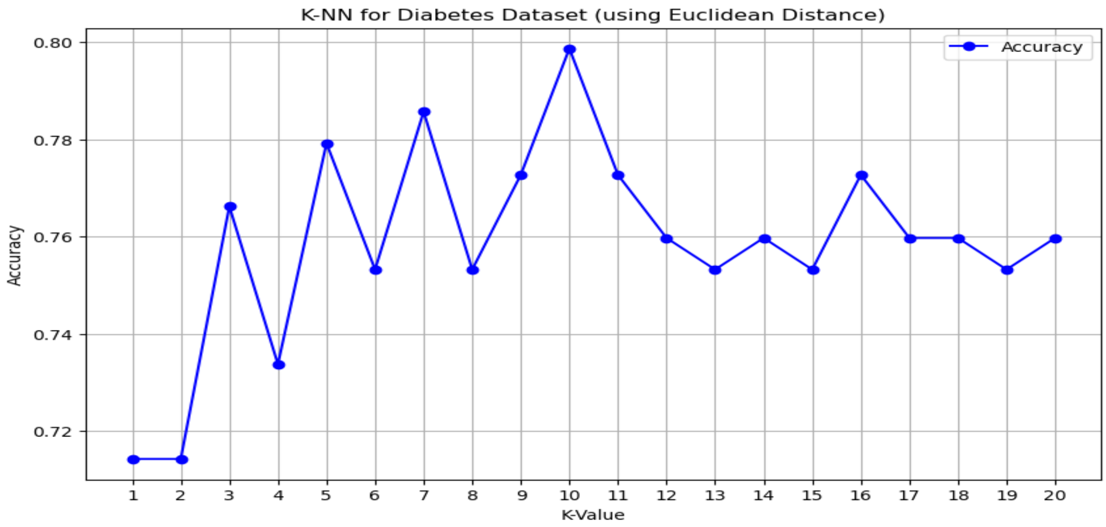
The dataset used has a binary, nominal outcome, making decision trees one of the better models to visually show the logic behind accurately predicting someone having diabetes. The decision tree above is the result of pruning because the decision tree is not close to any absolute stop criteria with its final nodes. This decision tree was created after my dataset was split into training sets and testing sets. I used 80% of the data for training and 20% for testing. A maximum depth of three was set on my decision tree to ensure that it wouldn't go past three levels of decision nodes to prevent overfitting. Entropy is within my nodes to measure the information gained when splitting data. The accuracy of the decision tree is 77%, which explains that 77% of the predictions for the test set observations aligned with their actual outcomes.

Chart 2: Confusion Matrix



This confusion matrix reiterates some information within the decision tree, but it also provides insight into the strengths and weaknesses found within our decision tree model. The confusion matrix shows that I correctly predicted 119 out of the 154 test set observations. This prediction rate is a direct reflection of the accuracy resulting from the rules set within the decision tree model. It is crucial to remember the confusion matrix is solely based on the 20% test set observations after using 80% of my data to train the model. First, I attempted to do the 2/3 training and 1/3 testing split for my data, but I found that the prediction accuracy was lower than the 4/5 training and 1/5 testing split for my test set observations. My model was better at identifying individuals without diabetes than with. A weakness that would need to be improved would be concerning false negatives found. A false negative is where we predict someone not having diabetes while they do. This is significant because this could be a life-or-death situation for someone in their near future. An adjustment to the decision tree's hyperparameters like max depth and minimum samples split could assist with bringing the number of false negatives down. Another way to decrease the likelihood of false negatives would be to try different models or techniques like Random Forest or Gradient Boosting. The final way would be to collect more observations or look for additional attributes to enhance the model's performance.

Chart 3: K-NN



K-NN is an instance-based algorithm that finds the K nearest observations to the test set observations. Since my dataset values vary across attributes, I standardized all attributes to be on the same scale to ensure no bias with my distance calculations. I used Euclidean distance as my distance metric and looped over multiple k-values to attempt to find the best one. The best k is the k-value corresponding to the highest accuracy. Finding the best k-value can be done with the max function in Python. Two trends were found through the data analysis of the information in this chart. With small k values, the algorithm performed poorly on unseen data and seemed to memorize the training set data. With large k values, the model dilutes the influence of closer neighbors and includes too many irrelevant distant neighbors. K = 10 provides a peak accuracy of 80% while avoiding overfitting or underfitting the observations. Our best k-value being even may pose an issue of there being a tie between modal outcomes, but this can be resolved by prioritizing based on specific criteria or considering the weighted distance to the test observations. Prioritizing can be done by having the highest priority feature that focuses on information that aligns more closely with the medical importance of certain risk factors. In our dataset, cases like higher glucose levels or higher BMI would have a higher priority than lower glucose levels or lower BMI. Weighted distance is a method that involves giving more influence on the nearer neighbors in the training data based on their distance from the test observation. Both approaches aim to refine the decision-making process if a tie occurs, and this would allow the K-NN algorithm to make more reliable predictions. For predicting diabetes, it would be better to predict the individual having diabetes than not in case of a medical emergency.

3 Conclusion

Throughout my study, I was surprised by my model's performance due to the complexity of predicting diabetes. Although the model was designed using reliable input variables, I encountered unexpected challenges such as false negatives. The number of false negatives found in my model was concerning given the life-or-death implications of an accurate diagnosis. Furthermore, the accuracy differences between the (2/3 training to 1/3 testing split) and the (4/5 training to 1/5 testing split) was notable, with the latter providing the better results. This could be due to the relatively small size of the dataset, and a larger proportion of data set towards training may have provided the model with a better understanding. My K-NN model slightly outperforms the Decision Tree in terms of accuracy, but the Decision Tree's strength lies in its visual explanation and its decision-making to be interpreted easily. To further improve my model's performance, future studies could explore adding alternative models such as Random Forests or Gradient Boosting. Additionally, acquiring a larger dataset or more diverse risk factors like smoking history and triglyceride levels could refine the model's predictions. These adjustments in future studies could reduce the false negative rate and provide a more accurate tool for predicting diabetes. Refining these models could have great effects for critical medical contexts.

Works Cited

Smith, J. W., Everhart, J., Dickson, W., Knowler, W., & Johannes, R. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Annual Symposium on Computer Application in Medical Care*, 261–265. Retrieved September 26, 2024, from https://europepmc.org/articles/PMC2245318

Collier, A., Patrick, A. W., Bell, D., Matthews, D. M., Maclntyre, C. C. A., Ewing, D. J., & Clarke, B. F. (1989). Relationship of skin thickness to duration of diabetes, glycemic control, and diabetic complications in male IDDM patients. *Diabetes Care*, *12*(5), 309–312. Retrieved September 24, 2024, from https://doi.org/10.2337/diacare.12.5.309

Bitzur, R., Cohen, H., Kamari, Y., Shaish, A., & Harats, D. (2009). Triglycerides and HDL cholesterol. *Diabetes Care*, *32*(suppl_2), S373–S377. https://doi.org/10.2337/dc09-s343

PIMA Indians Diabetes Database. (2016, October 6). Kaggle. Retrieved September 24, 2024, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Clinical Diabetes & American Diabetes Association. (2018, April). *Good to Know: Factors Affecting Blood Glucose*. National Library of Medicine. Retrieved September 24, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5898168/

Department of Health & Human Services. (n.d.). (2021, October 17). *Diabetes and insulin*. Better Health Channel. Retrieved September 25, 2024, from https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-and-insulin#gestational-diabetes

Akmeşe, Ö. F. (2022). Diagnosing Diabetes with Machine Learning Techniques. *Hittite Journal of Science & Engineering*, *9*(1), 9–18. https://doi.org/10.17350/hjse19030000250

Smoking and diabetes. (2022, May 19). Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/campaign/tips/diseases/diabetes.html

Cherney, K. (2022, July 6). *Age of onset for type 2 diabetes: Know your risk*. Healthline. https://www.healthline.com/health/type-2-diabetes-age-of-onset#delaying-the-onset

Southern Medical Association. (2023, January 19). *The Southern Medical Journal (Do not Delete) - Southern Medical Association*. Retrieved September 25, 2024, from https://sma.org/southern-medical-journal/article/relation-between-bmi-and-diabetes-mellitus-and-its-complications-among-us-older-adults/

Diabetes and high blood pressure. (2023, November 3). Johns Hopkins Medicine. Retrieved September 24, 2024, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure