Aastha Bhatt March 30, 2023

Title:

Likelihood of being Diabetic Based on Symptoms and Characteristics

Abstract:

Cases of diabetes are increasing mostly due to the lifestyle of the people today. Diabetes is also well known to put one at high risk of heart diseases and blood pressure. Thus, it is important to diagnose diabetes early so the person can get the right medications and treatments. In order to diagnose quickly, people need to be aware of their symptoms and what they indicate towards.

This dataset provides 520 records of patients in Sylhet Diabetes Hospital in Bangledesh. The dataset provides the age, gender, if the patient has any of the 15 symptoms, and if they were diabetic or not. The symptoms are the follwing: polyuria polydipsia, sudden weight loss, weakness, polygaphia, genital thursh, visual blurring, itching, irritability, dealyed healing, partial

When looking at the common symptom that diabetic patients had it was polyuria, and polydispia. Being male also increased the likelihood of being diabetic versus females. Out of the three classification models, K-Nearest Neighbors, Naïve Bayes, Decision Tree, the highest accuracy was seen using the K-Nearest Neighbor.

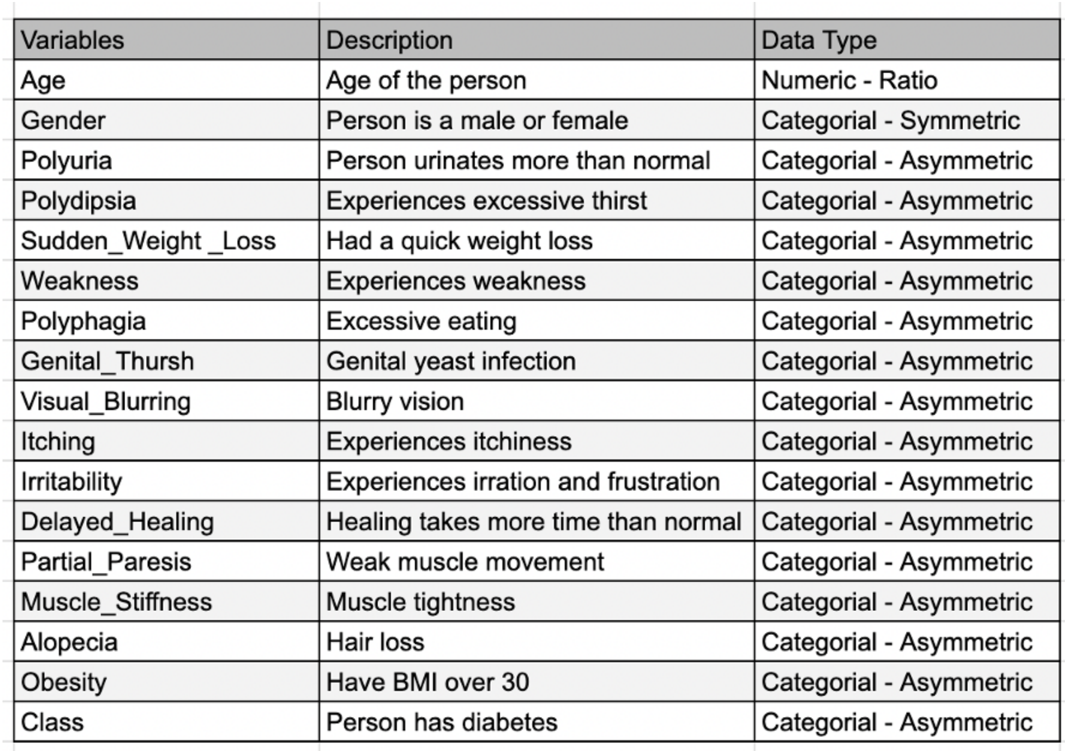
Introduction:

The dataset discussed in this report is the symptoms patients had and whether they had diabetes or not. This report will discuss the various classification models used to predict if a person has diabetes or not based on the symptoms they show. The purpose of trying to predict whether one is diabetic or not is so a person can become aware and go to the doctor early on because they are showing multiple symptoms that likely indicate diabetes. Seeking the doctor early means, early on diagnosis which leads to sooner treatment.

The report will first discuss the source of the data and describe the variables of the dataset. After that, describe the methods and the steps taken to produce the results and findings. After discussing this, the results of the tests and the code ran will be shown through tables and graphs. In the conclusion section, the findings, limitations, and future recommendations on further analysis will be discussed.

Data and Data Preparation:

I received this dataset from Kaggle, and was uploaded two years ago by Larxel in 2022. The data in the dataset comes from patients in Sylhet Diabets Hospital in Bangledesh. The attribute “age” is a numerica- ratio data type that lists the age of the patient. The gender attribute, is a categorical - symmetric data type, with a range of two values which are Male and Female. The other 15 attributes that are going to be described are all categrorial- symmetric data type. The two values are 1 and 0, 1 is to indicate the patient has the symptom and 0 if they don’t. The variable “polyuria” indicates wheher the patient urinates more than normal. Polydipsia indicates if the patient experiences excessive thirst. The other attribute is a suddenn weight loss, and if they experience weakness. After this is the attribute polygaphia, which indicates if the patient is eating excessively. The other symptom is genital thursh, which is to indicate if the patient is expericining genital yeast infection. Follwing that are visual blurring, itching, irritability, delayed healing, and mucle stiffness. One of the other symptoms is partieal paresis which is weak muscle movement. Another attribute is alopecia which is hair loss. The last characteristic/symptom is obestiy which is when the patient has a body mass index over 30. The last column in the data set is under the name “class”, this is indication of whether the patient is diabetic or not. Below is a table, summarizing the above information of variables, description of the data in each variable, and the data type.



Variables	Description	Data Type
Age	Age of the person	Numeric - Ratio
Gender	Person is a male or female	Categorial - Symmetric
Polyuria	Person urinates more than normal	Categorial - Asymmetric
Polydipsia	Experiences excessive thirst	Categorial - Asymmetric
Sudden_Weight _Loss	Had a quick weight loss	Categorial - Asymmetric
Weakness	Experiences weakness	Categorial - Asymmetric
Polyphagia	Excessive eating	Categorial - Asymmetric
Genital_Thursh	Genital yeast infection	Categorial - Asymmetric
Visual_Blurring	Blurry vision	Categorial - Asymmetric
Itching	Experiences itchiness	Categorial - Asymmetric
Irritability	Experiences irration and frustration	Categorial - Asymmetric
Delayed_Healing	Healing takes more time than normal	Categorial - Asymmetric
Partial_Paresis	Weak muscle movement	Categorial - Asymmetric
Muscle_Stiffness	Muscle tightness	Categorial - Asymmetric
Alopecia	Hair loss	Categorial - Asymmetric
Obesity	Have BMI over 30	Categorial - Asymmetric
Class	Person has diabetes	Categorial - Asymmetric

There was no missing data so no preprocessing steps were done in that aspect.

Methods:

After importing the data, I did an exploratory analysis by running a summary of the data, part of this can be seen in Table 1 and Table 2. This helped me understand the variables in the dataset. Then I looked at the age of the patients in the dataset visually thorugh a histogram as seen in Figure 1. I decided to take out the varaible from my analysis because diabetes can be diagnosed at any age and this characterstic is not going to be a main factor in determining whether a person has diabetes or not. Then I converted the “class” column which indicates whether a person has diabetes or not to a factor. Then I partitioned the dataset into training and testing using the holdout method. I set the random seed for repeadesbility by doing “set.seed(4567)”. Then I created an index variable to perform the 70/30 split. After this, I checked the proportion of the training and testing partition to make sure the data was split evenly and properly. Then, I ran the first classification model, K-Nearest Neighbor with repeated 10-fold cross-validation. Evaluated the classifier using the confusion matrix, which in code language used the confusion matrix function. The second classifier used was decision tree, also using cross validation. I plotted a simple representation first, then used the rpart.plot to plot a visually appealing decision tree, this can be seen in Figure 2. I also evaluated this model using the confusion matrix. The last classification model performed was the Naive Bayes using cross validation. Similar to the other two models, I evaluated the model using a confusion matrix. After deriving the confusion matrix, I calculated the error rate of all the models.

Results:

Table 1 shows the average of the symptoms among the patients. In the dataset, if the patient has the sysmptom then it is indicated by 1 and if not then 0. So, the higher the average, that means the more patients in the dataset show that symptom. Looking at the table, it can be said that most of the symptoms average is around the .40-.50 mark which is half. So mostly there is a good split of patients who do exhibit the symptoms and not. This is a good thing, because in the world, some people do experience a symptom while others don't, even if they have the same disease. The symptoms that are lower than the half mark are the follwing symptoms and characterstics: genital thrush, irritability, muscle stuffness, alopecia, and obesity. Table 2, shows the number of patients who are diabetic and who are not. Figure 1 shows a histrogram of the age of the patients in the dataset. There is a bell curve shape to the historgram, with mostly the patients age being around 40-50ish.

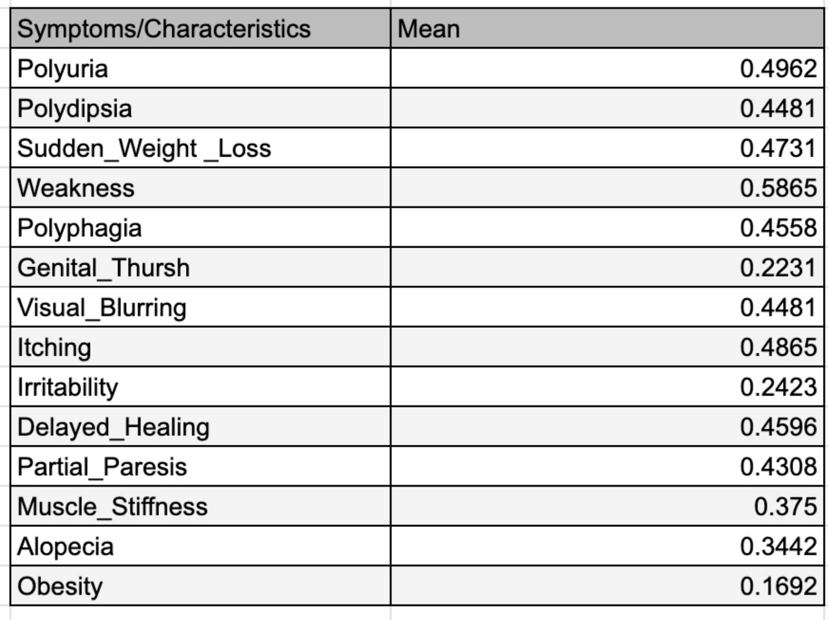
Table 3 shows the evaluation of the K-Nearst Neighbor using cross validation through a confusion matrix. The model predicted accurately that 59 patients did not have diabetes and 87 did. However, it predicted 9 patients did not have diabetes when they actually did. While it also predicted that 4 people had diabetes when they did not. This is a type 1 error which is more harmful than the 9 patients that were not diagnosed with diabetes.

Table 4 shows the evaluation of the Decision Tree using cross validation through a confusion matrix. The model predicted accurately that 45 patients did not have diabetes and 85 patients did. However, it predicted that 11 patients did not have diabetes when they did and 15 patients were predicted to have diabetes when they actually did not. So the type 1 error would consists of the 15 patients. Figure 2 shows the decision tree, which is first split with polyuria and if you do have it then there is a .95 chance that the person has diabetes. If they do not have polyuria then the decision tree splits based on gender. If they are a female then there is .71 chance they are diabetic. If you are a male, then there is only a .15 chance that they are not diabetic. In other words, if you do not have polyuria, and are male then you have a .85 chance of being diabetic.

Table 5 shows the evaluation of the Naive Bayes using cross validation through a confusion matrix. The model predicted accurately that 50 patients did not have diabetes and 87 patients did. However, it predicted that 9 patients did not have diabetes when they did and 10 patients who did not have diabetes were predicted to have diabetes. The type 1 error consists of 10 patients who were diagnosed with diabetes but did not have it.

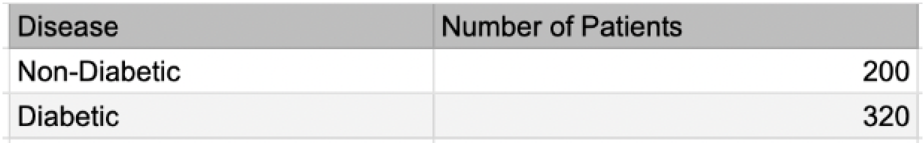
Table 6 compares the three classifaction models which are K-Nearest Neighbor, Decision Tree, and Naive Bayes. The model with the highest accuray of 0.9167 and the lowest error rate of 0.0833 is K-Nearest Neighbor. In second comes the Naive Bayes with accuracy of 0.8782 and error rate of 01218. In last comes the Decision Tree with the accuracy of 0.8333 and error rate of 0.1667. Overall, all of the models accuracy rate is above .80 and error rate is below 0.2. The precision, sensitivity, recall, and F1 fall in the same order as accuracy which is K-Nearest Neighbor, Naive Bayes, and then Decsion Tree. The only row that has a different order than the one above is specificity. Specificity is the actual number of negative cases that were correctly identified versus the total number of negative cases identified. Both K-Nearest Neighbor tie in the sense they have the same value of 0.9062 and after that is the Decision Tree at the value 0.8854.

Table 1: Average of the Symptoms Present in the Patients



Symptoms/Characteristics	Mean
Polyuria	0.4962
Polydipsia	0.4481
Sudden_Weight _Loss	0.4731
Weakness	0.5865
Polyphagia	0.4558
Genital_Thursh	0.2231
Visual_Blurring	0.4481
Itching	0.4865
Irritability	0.2423
Delayed_Healing	0.4596
Partial_Paresis	0.4308
Muscle_Stiffness	0.375
Alopecia	0.3442
Obesity	0.1692

Table 2: Number of Diabetic Patients in the Dataset



Disease	Number of Patients
Non-Diabetic	200
Diabetic	320

Figure 1 : Histogram of Age of Patients

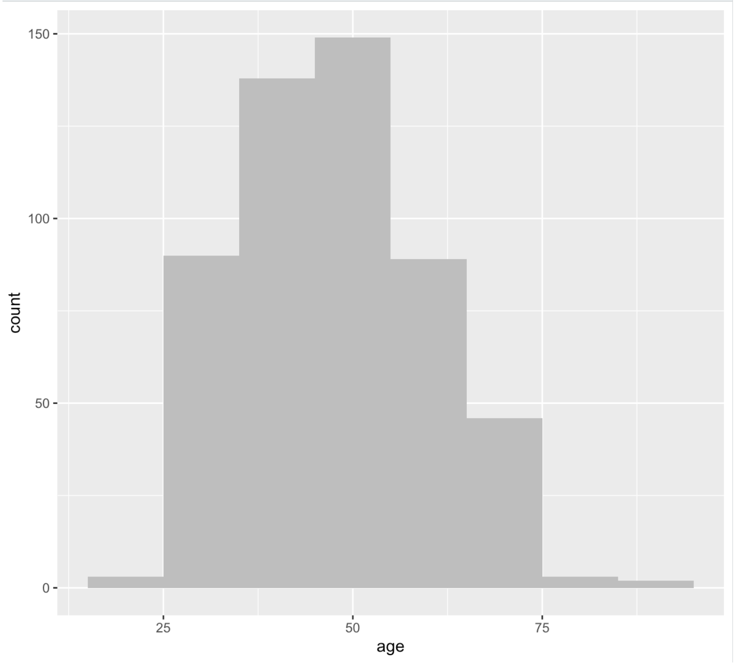
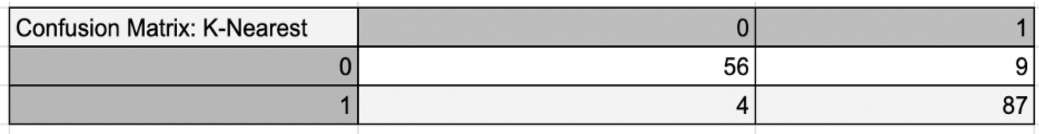


Table 3: K-Nearest Neighbor Confusion Matrix

Actual

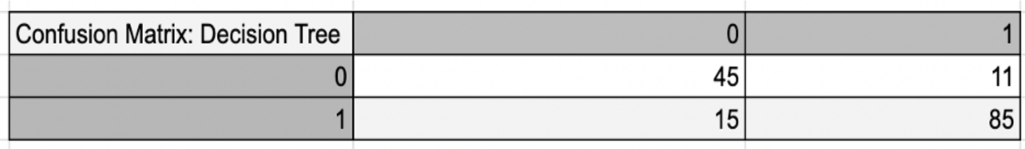


Confusion Matrix: K-Nearest	0	1
0	56	9
1	4	87

Predicted

Table 4: Decision Tree Confusion Matrix

Actual



Confusion Matrix: Decision Tree	0	1
0	45	11
1	15	85

Predicted

Figure 2: Decision Tree

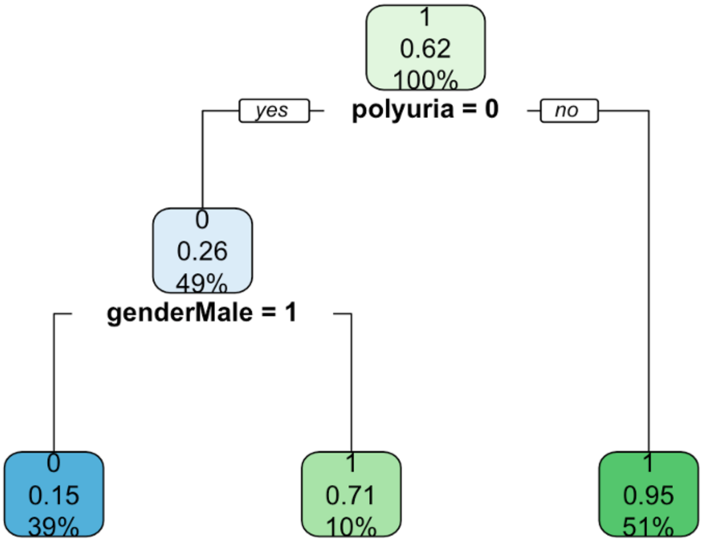
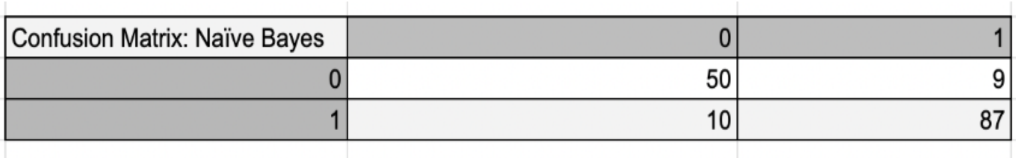


Table 5: Naive Bayes Confusion Matrix

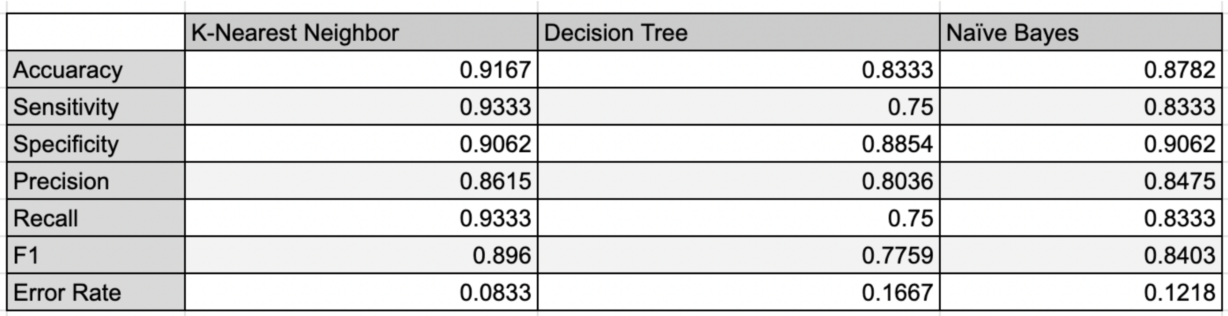
Actual



Confusion Matrix: Naïve Bayes	0	1
0	50	9
1	10	87

Predicted

Table 6: Evaluation and Comparison of the Classification Models



	K-Nearest Neighbor	Decision Tree	Naïve Bayes
Accuaracy	0.9167	0.8333	0.8782
Sensitivity	0.9333	0.75	0.8333
Specificity	0.9062	0.8854	0.9062
Precision	0.8615	0.8036	0.8475
Recall	0.9333	0.75	0.8333
F1	0.896	0.7759	0.8403
Error Rate	0.0833	0.1667	0.1218

Conclusion and Discussion:

Based on the results, the best classification model was K-Nearest neighbor, then naive bayes, third is the decision tree. All of the models had a accuracy rate higher than 0.80 and an error rate below 0.2. When it comes to type 1 error, diagnosing people who do not have diabetes with diabetes, K-Nearest Neighbor wrongly predicted 4 patients, Naive Bayes did 10, and the Decision Tree wrongly predicted 15 patients. Again, K-Nearest Neighbor had the lowest but is an error that can scare a person. On the other hand, type II error would be misdiagnosing that a person who is diabetic to not be diagnosed with diabetes. That would mean a late start on treatment leading to worse conditions and possibly death. One of the limitation on this anaylsis is we don't the type of diabetes the patient has. This dataset provided just the information whether a patient was diabetic or not but not the type. Another limitation in this analysis I did not include age because its not necereally a symptom and a person can be diagnosed at any point in their lifetime. Yes, the older you get the chances increases. This being said, for further analysis on symptoms and liklihood of diabetes, the type of diabetes can be looked at. As well as characteristics such as age, race, ethnicity and the region the person lives. In conclusion, the number of diabetic pateints are growing and people becoming aware of their symptoms and what they indicate to could encourage them to go to the doctor and start their treatment.

References:

Basic evaluation measures from the confusion matrix. (2017, September 13). Retrieved March 30, 2023, from https://classeval.wordpress.com/introduction/basic-evaluation-measures/#:~:text=Error%20rate%20(ERR)%20is%20calculated,dataset%20(P%20%2B%20N).

Latrex (2022). *Early Classification of Diabetes.* Kaggle. https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification

Sharma, P. (2022, March 09). Decoding the confusion matrix. Retrieved March 30, 2023, from https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb

Type 2 diabetes. (2023, March 14). Retrieved March 30, 2023, from https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193