Project Report

Evaluation of various Classification Models for Diabetes prediction

by

Ali A. Zamin

December 6, 2022

Abstract

It has been noticed that the number of diabetic patients is increasing day by day from young to old people. According to WHO (World health organization), the number of people with diabetes has increased over the years. Diabetes can cause many problems to the body if it remains unidentified and untreated. Therefore, it is very important to have an accurate system, through which diabetes doesn't go undetected and it accurately classifies patients with diabetes or no diabetes. The objective of this project is to train various classification models on patients' data and compare the performance of these models to see which classification technique performs better and predicts accurately, that is, classifying patients with diabetes and patients with no diabetes, with maximum accuracy. To do this we will use different classification methods, such as Naïve Bise, Decision tree, Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant Analysis. And, then we will evaluate the performance of these models using various evaluation metrics.

Introduction

What is Diabetes?

First of all, let's familiarize ourselves with diabetes in a bit of detail. According to CDC, "Diabetes is a chronic (long-lasting) disease that affects how your body turns food into energy". It is considered one of the fatal diseases which cause an increase in blood sugar. With diabetes, your body doesn't make enough insulin or can't use it as well as it should. When there isn't enough insulin or cells stop responding to insulin, too much blood sugar stays in your bloodstream. Over time, that can cause serious health problems, such as heart disease, vision loss, and kidney disease.

Types of Diabetes

There are three main types of diabetes: type 1, type 2, and gestational diabetes (diabetes while pregnant).

Type 1 Diabetes

Type 1 diabetes is thought to be caused by an autoimmune reaction (the body attacks itself by mistake). This reaction stops your body from making insulin. Approximately 5-10% of the people who have diabetes have type 1. Symptoms of type 1 diabetes often develop quickly. It’s usually diagnosed in children, teens, and young adults. If you have type 1 diabetes, you’ll need to take insulin every day to survive. Currently, no one knows how to prevent type 1 diabetes.

Type 2 Diabetes

With type 2 diabetes, your body doesn’t use insulin well and can’t keep blood sugar at normal levels. About 90-95% of people with diabetes have type 2. It develops over many years and is usually diagnosed in adults (but more and more in children, teens, and young adults). You may not notice any symptoms, so it’s important to get your blood sugar tested if you’re at risk. Type 2 diabetes can be prevented or delayed with healthy lifestyle changes, such as:

- Losing weight.
- Eating healthy food.

- Being active.

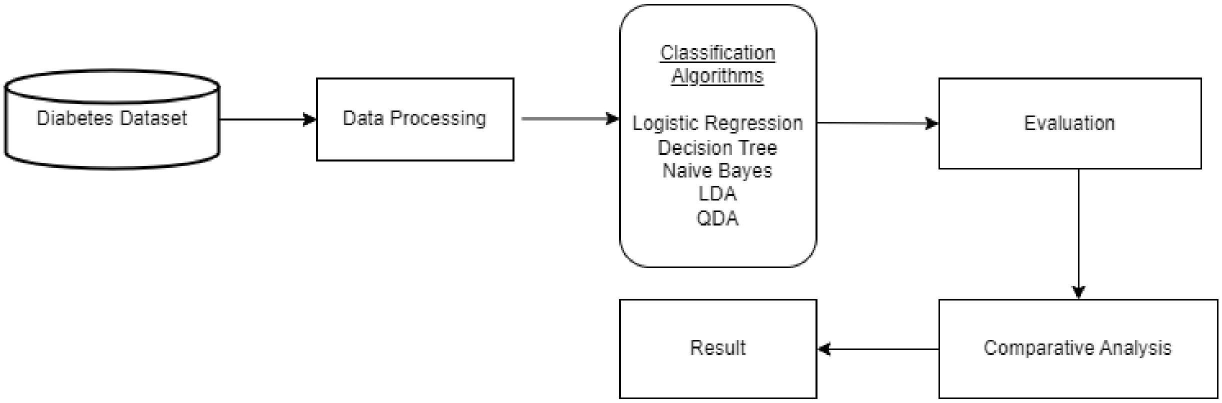
Gestational Diabetes

Gestational diabetes develops in pregnant women who have never had diabetes. If you have gestational diabetes, your baby could be at higher risk for health problems. Gestational diabetes usually goes away after your baby is born. However, it increases your risk for type 2 diabetes later in life. Your baby is more likely to have obesity as a child or teen and develop type 2 diabetes later in life.

Prediabetes

With prediabetes, blood sugar levels are higher than normal, but not high enough for a type 2 diabetes diagnosis. Prediabetes raises your risk for type 2 diabetes, heart disease, and stroke. But there’s good news. If you eat healthy foods, regular exercise, and maintain a healthy weight, you can reverse it.

Design/Approach



The above figure demonstrates our design and approach to achieving our objective. The following are the steps.

1. Start by data cleaning\pre-processing our data.
2. Train different classification models with pre-processed data.
3. Use evaluation matrixes to get the performance scores of each model.
4. Compare and analyze the performance of each model
5. Show what was discovered or found in the result.

Data Information

The dataset is acquired from Kaggle (https://www.kaggle.com/datasets/mathchi/diabetes-data-set) but it, in fact, originates from the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The variables involved in the datasets are:

	Attributes	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index (Kg/m^2)
7	DiabetesPedigreeFunction	Diabetes pedigree function (indicates the function which scores likelihood of diabetes based on family history)
8	Age	Age (years)
9	Outcome	Class label (0 and 1). 1 indicates positive diabetes test. 0 indicates negative test.

There are 8 input variables and 1 output variable, with a total of 768 instances and all the data points are of numeric type. The data is slightly imbalanced, that is the number of two classes, 1 and 0, are not relatively equal. The data consists of 65% of class 0 is 35% of class 1.

Processing

Tools and Implementation

Python programming language is a very easy and great tool for data analysis and machine learning. It has many great libraries and tools for data analytics and machine learning and very easy and simple to implement them. Therefore, we used Python coding to visualize and build our model. The coding was done using Jupiter notebook, which is a very famous and interactive code editor used for easy data visualization. Some of the libraries we used to for data analysis and modeling are sklearn, Pandas, and MatPlotLib.

Data Pre-processing

In data pre-processing, we check for the missing values to clean them as they lead to decrease in the accuracy of the models. For our dataset, we first checked for any null/Nan values to see if there is any missing values. It turned out that there were no Null/Nan values.

Next, we counted the numbers of zeros for each feature (except for “Insulin” and “Pregnancy” attribute). We observed that feature “SkinThickness” contained considerably a lot of zeros values, which was around 227 counts. Since It could affect the accuracy of our model, It is completely removed from the dataset.

Next, we removed all the rows containing zeros in features “Glucose", "BloodPressure", "SkinThickness", "BMI", "DiabetesPedigreeFunction", and "Age”, considering it as missing values, as zero is invalid values for these features since it doesn’t make sense to have zero values for them. For instance, it’s not possible for a person to 0 blood pressure because there must be some kind of blood pressure in the body. As a result, our dataset reduced to 724 datapoints.

Training and Testing

Selecting a proper model allows you to generate accurate results when making a prediction. To do that, you need to train your model by using a specific portion of your dataset. Then, you test the model against another another set amount of data. To do that you can use a train test split method to divide your dataset with specific ratio. In our case, we have divided dataset into 80:20 ratio, with 80% being training set (579 samples) and 20% testing/validating set (145 samples).

Cross validation

We will use cross validation in our data to better validate our model. To give an overview, “Cross validation (CV) provides the ability to estimate model performance on unseen data not used while training”. One of the critical pillars of validating a learning model before putting them in production is making accurate predictions on unseen data. The unseen data is all types of data that a model has never learned before. This is where cross validation comes handy, as it basically validates a model with multiple iteration, using the testing set that the model hasn’t seen before.

Evaluation methods

For evaluating the performance of our classification algorithms, we used some evaluation matrixes, namely, Accuracy, Receiver operating characteristic (ROC) curve and Area under the curve (AUC).

Accuracy

Accuracy is the quintessential classification metric. It is pretty easy to understand and easily suited for binary classification problem. It simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.

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

ROC-AUC

We specifically chose ROC and AUC evaluation methods as they work great when the data is imbalanced, which is true in our case. The Receiver Operator Characteristic (ROC) is a performance measurement curve and is a “probability curve that plots the TPR(True Positive Rate) against the FPR(False Positive Rate) at various threshold values”. Whereas Area Under the Curve (AUC) is simply the value of the area under the ROC curve. It is “the measure of the ability of a classifier to distinguish between classes”.

In our project we will plot the ROC curves of classification models and calculate/display the AUC score of each of those model’s roc curve. Using ROC curve and AUC score, we can tell how much our models are capable of distinguishing between two classes (0 and 1). By analogy, the Higher the AUC, the better the model is at distinguishing patients whether they have diabetes or not.

Modeling

For our project we used 5 different classification models, namely, Naïve Bayes, Decision tree, Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant Analysis. Below we discuss those model in a bit detail.

Naïve Bayes

The Naive Bayes is classification technique and a supervised learning algorithm based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. Naive Bayes is a purely statistical model and can be used for binary classification. It is one of the simplest and most effective classification algorithms which helps in building the fast predicting models that can make quick predictions.

Decision Tree

A Decision Tree is a non-parametric supervised learning method used for classification and regression. It is another great technique for binary classification as it creates a model that predicts the value of a target variable by learning simple decision rules (if-else) inferred from the data features. The simple mechanism of how decision tree work is that for predicting a class label for a record, we start from the root of the tree. We compare the values of root attribute with the record attribute. On the basis of comparison, we follow the branch corresponding to the value and jump to next node.

Linear Discriminant Analysis

Linear Discriminant Analysis, also known as LDA, is another important binary classification technique for distinguishing feature variables between healthy individuals and patient's data. It is utilized to deduce feature variables and help the supervised learning method with finding an arrangement of base vectors.

Although, it's alternative, Principal Component Analysis (PCA), is another famous method used among most data analysts, but generally LDA is preferred over PCA because LDA directly deals with discrimination between classes. While PCA does not pay any attention to the underlying class structure since it has lower error rates.

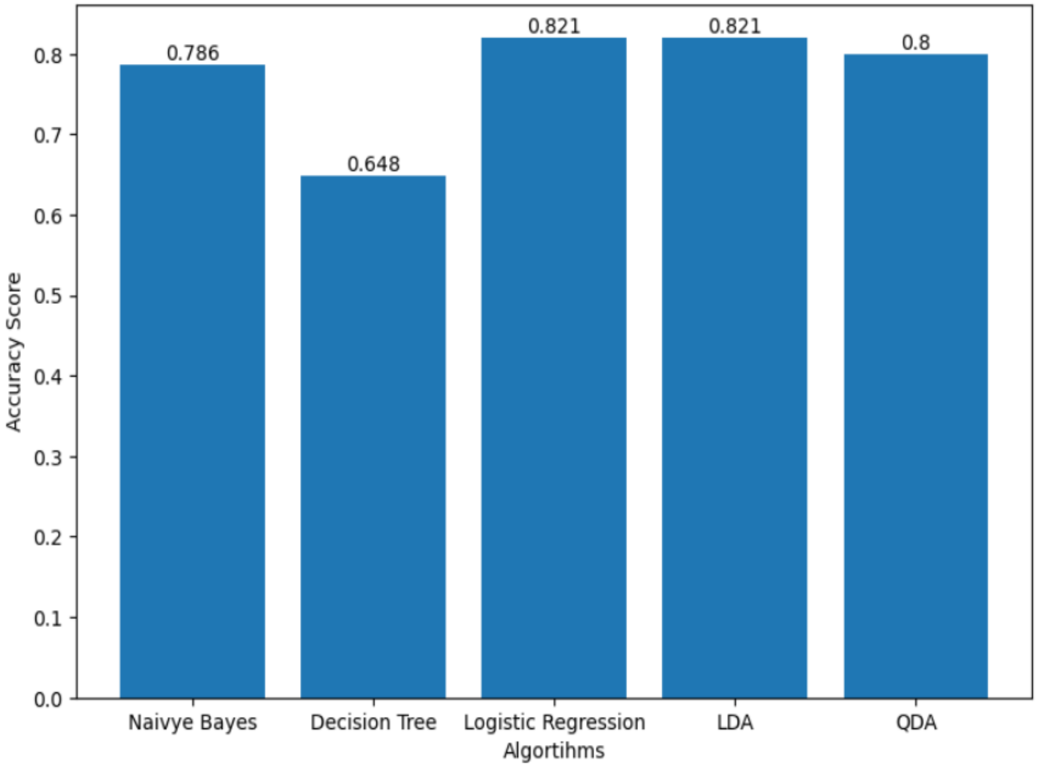
Quadratic Discriminant Analysis

Quadratic discriminant analysis is quite like Linear discriminant analysis except that it allows for non-linear separation of data. Because, here, the covariance matrix is not identical, so you cannot throw away the quadratic terms. It allows for more flexibility for the covariance matrix, and tends to fit the data better than LDA, but then it has more parameters to estimate.

Result

Accuracy

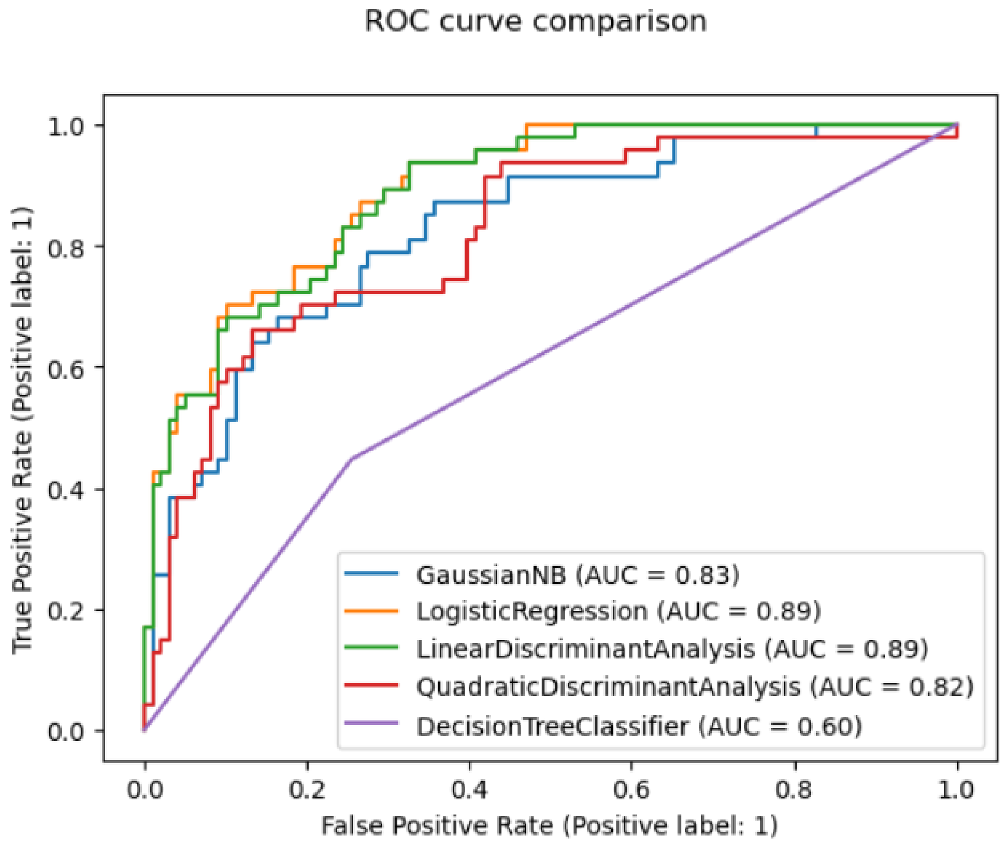
Each model's accuracy score was saved in a list and later plotted in a bar chart. The bar chart below shows the accuracy score of each classification model, showing the algorithms on the x-axis and Accuracy scores on the y-axis.



Here, the accuracy of each model are following; Naive Naïve Bayes 78%, Decision Tree 64%, Logistic Regression 82%, LDA 82%, and QDA 80. From the above result we can see all models are good enough for our binary classification, having at least 64% accuracy which is good. However, Logistic Regression and LDA turns out to be the two best performing models, giving highest accuracy of 82%. Whereas Decision tree gives the lowest accuracy, that is, 64%.

ROC-AUC Curve

For our ROC-AUC curve, we plotted all our ROC curves of the models in one graph, along with their AUC scores, making it easy for us to visualize and analyze the data. The x-axes represent the False Positive Rate while y-axes represent True Positive Rate.



Analyzing the ROC curves above, we observe the following AUC scores for each model; Naive Naïve Bayes 0.83, Decision Tree 0.60, Logistic Regression 0.89, LDA 0.89, and QDA 0.82. The basic understanding of the ROC-AUC curve is that the more ROC curve is tilted towards top left with the high area under the curve (AUC), the better the model's performance is. In our case, the Higher the AUC, the better the model is at distinguishing patients whether they have diabetes or not. The result shows that Logistic Regression and LDA gives the highest AUC score, i.e., 0.89. On the other hand, Decision Tree gives the lowest AUC score of 0.60. This result indicates that logistic regression and LDA are the two best performing models for our data, while Decision Tree is lowest performing model. As a result, both Logistic Regression and LDA have the highest tendency of correctly predicting or classifying patients with diabetes or no diabetes given the patients' data.

Discussion and Conclusion

In this project, we presented our approach in the design and development of a diabetes prediction model and their performance evaluation. Our main objective was train various classification models on patients' data and compare the performance of these models to see which classification technique performs better and predicts accurately, that is, classifying patients with diabetes and patients with no diabetes, with maximum accuracy. As a result, we found out that Logistic Regression and LDA are the two models that perform best on our data. We confirm that by using two evaluation matrixes, that is, accuracy and ROC-AUC, as they both showed the same result. It might not be surprising why logistic regression performs best here as it is popular and known for performing well for binary classification. However, it is worth noting that LDA performs well on our data compared to other popular classification model like Decision tree.

Though this project gives us some results and idea on how to use diabetes dataset for diabetes prediction in patients, there are still some factors that has to addressed for future researchers in this domain. For instance, further research can be conducted to improve the accuracy of the model furthermore. In addition, our model was trained on a very low amount of data, i.e., 768 instances. Therefore, if given more data or trained in large dataset, the accuracy of the model can be further improved. For future, the model should be trained on scaled data to see if it improves the performance of the model. Also, further research can be done to see if the model, given very few features, could predict with the same accuracy, whether patients has diabetes or not. Most importantly, the model has to be further explored to see if there is a technique that could be applied on the model which would make the system not only predict the diabetes but also its type. Since our dataset is about female diabetic patients, it's is reasonable to think that the model should be capable to distinguish diabetes type and specify the type of diabetes the patient has.

Overall, the factors that cause diabetes in people is Genetics, lifestyle or environment. There isn't a cure yet for diabetes, but losing weight, eating healthy food, and being active can really help.

Reference

- **https://www.kaggle.com/code/baturalpsert/diabetes-classification-roc-curve/notebook#ROC-Curve**
- **https://www.kaggle.com/code/baturalpsert/diabetes-classification-roc-curve/notebook**
- **https://www.youtube.com/watch?v=O_5kf_Kb684**
- **https://scikit-learn.org**
- **https://prezi.com/p/ezef9wri0ymy/predictive-analytics-in-healthcare-for-diabetes-prediction/?frame=386d1a205382fd219761ce7d9cbf26afa1129a37**
- **https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75**
- **https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=A366DD819AEB5535F9DAC39FD42FAB31?sequence=1**
- **https://www.geeksforgeeks.org/quadratic-discriminant-analysis/**