Ezra Denny and Ebuka Enyinnaya

Prof. Oscar Aguilar

DATA-101

11/10/2023

[Predict Diabetes (kaggle.com)](https://www.kaggle.com/datasets/whenamancodes/predict-diabities)

**Final Project Report**

Abstract

Diabetes is a chronic disease that affects millions of people worldwide. It is characterized by high blood glucose levels that can lead to serious complications such as cardiovascular disease, kidney failure, and blindness. Early detection and prevention of diabetes are crucial for reducing the burden of this disease. The purpose of this project is to use a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) to predict whether a female patient of Pima Indian heritage has diabetes, based on certain diagnostic measurements. This dataset is available from the website: https://www.kaggle.com. Kaggle is a website that provides over 50,000 free public datasets. The dataset contains information on 768 patients, so there are 768 rows and 9 columns for the 9 variables. These variables include things such as age, number of pregnancies, blood pressure, body mass index, insulin level, and glucose level. The dataset also indicates whether the patient has diabetes or not. This dataset has no missing values. However, some of the measurements are zero, which may indicate missing data or erroneous entries, or perhaps no indication of diabetes at all. The goal of this project is to apply machine learning/data analytics techniques to this dataset and build a predictive model that can accurately classify the patients as having diabetes or not.

**Table Of Contents**

* Table of Contents
* Introduction
* Reason for project topic
* Data section
  + Where data is coming from
  + Description of row and columns
  + Exploratory Analysis
* Modeling
* Conclusion

**Introduction**

In this study we will explore the field of diabetes prediction using data analytics, using extensive datasets from electronic health records, genetic information, lifestyle factors, and more. Our goal is to find out the complex health intricacies and highlight the importance of onset diabetes. Through the application of advanced analytics techniques, our goal is to apply machine learning/data analytics techniques to this dataset and build a predictive model that can accurately classify the patients as having diabetes or not.

**Reason for project topic**

By using data analytics to anticipate the emergence of diabetes, our aim is to play a role in finding meaningful preventive strategies, finding early intervention protocols, and finding personalized healthcare methods. The advantages of precise diabetes prediction go beyond individual well-being, influencing the invention of healthcare resources, the formulation of policies, and nurturing a collaborative solution to alleviate the escalating impact of this metabolic disorder. Through this, our goal is to make a significant contribution to the ongoing discourse on diabetes by using data analytics to revolutionize healthcare practices and enhance public health outcomes at large.

List of questions we want to answer:

* What is the distribution of the patients’ age, number of pregnancies, blood pressure, body mass index, insulin level, and glucose level?
* What is the correlation between the diagnostic measurements and the diabetes outcome?
* How many patients have diabetes and how many do not?
* What are the characteristics of the patients who have diabetes and who do not?
* Can we build a logistic regression model that can predict whether a patient has diabetes or not, based on the diagnostic measurements?
* How accurate is our model and how can we evaluate its performance?
* Can we improve our model by using feature engineering, feature selection, or regularization techniques?
* Can we use other classification algorithms, such as decision trees, random forests, or support vector machines, and compare their results with our logistic regression model?
* Can we use our models to make predictions on patients within the dataset who are not diagnosed as diabetic? Such are predicting the likelihood that such individuals are diabetic.
* Can we use clustering algorithms, such as k-means or hierarchical measurement ring, to group the patients into different clusters based on their diagnostic measurements?
* What are the similarities and differences between the clusters and how do they relate to the diabetes outcome?
* Can we use principal component analysis or t-distributed stochastic neighbor embedding to reduce the dimensionality of the dataset and visualize the clusters?

**Data Section**

**Where data is coming from**

The foundation of our data-driven exploration into diabetes prediction relies on a comprehensive collection of information from diverse sources. Primary among these are the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), providing a rich repository of patient glucose levels, BMIs, and Insulin levels. Genetic data contributes a crucial dimension, offering insights into the hereditary factors influencing an individual's susceptibility to diabetes. Lifestyle factors, including dietary habits, physical activity levels, and socio-economic indicators.

**Description of rows and columns**

Each row represents an individual case or observation, providing a comprehensive snapshot of various factors relevant to diabetes prediction. These factors, summarized in columns, encompass a diverse array of information ranging from demographic details and medical history to genetic markers and lifestyle variables. Commonly included columns involve age, body mass index (BMI), blood pressure, skin thickness, insulin levels, Pregnancies, Diabetes Pedigree Function, and clinical measurements such as glucose and cholesterol levels. Each column serves as a distinct dimension, contributing to the landscape of data essential for training and validating predictive models.

**Exploratory Analysis**

In this section, we will try to answer all the questions we wanted to know about the dataset.

To start, here are histogram charts that relay the frequencies of each variable in the dataset. This helps us get a better understanding of each variable and what the more common values of them are. For example, below are the charts, and a large portion of the people in this dataset are in their early 20’s, have 0 – 3 pregnancies, blood pressure between 60-80, and BMI between 30 – 40. Another noteworthy frequency is insulin and glucose, as insulin frequencies are around 0-100 in most people, and glucose around 100-125. Knowing the frequencies of these variables can help us in making predictions with different models.

A graph of age and frequency

Description automatically generated A graph of a graph

Description automatically generated with medium confidence

A graph of blood pressure

Description automatically generated A graph of a number of objects

Description automatically generated with medium confidence A graph of insulin

Description automatically generated A graph of a number of blood glucose

Description automatically generated with medium confidence

Here we have the correlations of each variable to the target variable, which is of course the outcome of whether a person is diabetic. Notice that we didn’t include the skin thickness variable. This is because there are other factors that would be much more likely to determine diabetes, like glucose level, insulin level, and BMI. A lot of times skin thickness is more related to a person’s genetics, not necessarily correlated with diabetes. As we can see from the correlations below, skin thickness has the lowest correlation with the outcome, while variables like age, pregnancies, and BMI have a much higher correlation with the outcome variable. This is important because we now know what variables are necessary to use when making predictions using our models based on their correlations with the outcome variable.















Based on the results of each correlation, our best predictor in terms of what may determine whether someone is diabetic, would-be glucose levels, BMI, number of pregnancies, and age. An important factor to remember however, is that if a person’s insulin level is determined to be 0 or very low, this can also be a very telling indication that a person is diabetic, even though it doesn’t have a high correlation.

Next, we determined how many people in the set of data had diabetes, and how many did not.



As you can see, there are many more people without diabetes than with diabetes in the dataset. Because of this, we notice we have a lot of data that we can work with when it comes to predicting likelihoods of becoming diabetic on the people in the set who have not been diagnosed with diabetes.

Below are the characteristics we found of people in the dataset who had diabetes, and below are characteristics of the people in the dataset who do not have diabetes.

A screenshot of a computer

Description automatically generated

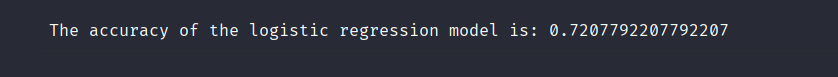
A screenshot of a computer

Description automatically generated

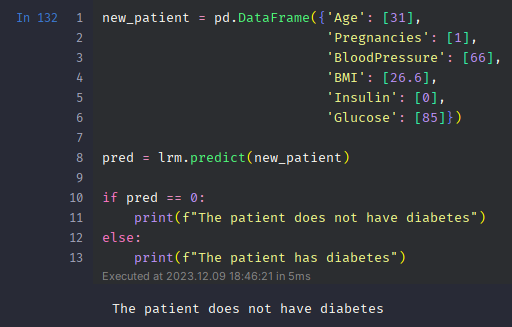
These two tables of characteristics of each patient type can help us when we build a regression model to predict diabetes in people, because it gives us a better understanding of what may determine whether somebody is diabetic. It also gives us easy test data if we wanted to test out of models manually.

In this next part, we built a logistic regression model so we could start making predictions.

The first model we will show is a logistic regression model, in which the variables are not standardized. After the model’s creation we determined the accuracy to be around 72%.



Then, we used some test data to make a prediction on that model:



The model we created can accurately predict diabetes in patients using test data, but we needed to try and make the model more accurate, so we rebuilt the same model but with standardized variables, and the accuracy of the model improved:



This improvement was around 5%

We then used some test data again to make a prediction on the new standardized logistic regression model:

A computer screen shot of a program

Description automatically generated

However before predicting on the model with our data, we must standardize it first, since our model is trained off standardized data, not just the raw data. After doing that, the model proved to be accurate once again, however of course, 5% more accurate now that the data is standardized.

To try and find a more accurate model, we made a random forest classification model that was used much like the logistic regression model to make predictions on the dataset.



The accuracy of the Random Forest model was identical in accuracy to logistic regression model. It is important to note that we standardized the data on this model and any after it as well. We did this because we already knew that the models are more accurate that way.

A screenshot of a computer program

Description automatically generated

As we assumed, the model was just as accurate as the logistc regression model, and our test data gave the same output, which we already knew to be correct.

We also used a Gradient Boosting Classifier model:



This model was not as accurate as the others, however, could still accurately predict diabetes in patients.

A computer screen shot of a program

Description automatically generated

As we had hypothesized initially, the model does indeed accurately predict diabetes in patients regardless of the difference in accuracy. However, when it comes to choosing which model we thought fit best, we chose the logistic regression model.

An important goal that we wanted to hit in this project was to predict the likelihood of patients who were not diagnosed as diabetic being diabetic. This is an important task to accomplish as it would prove the real-world application of such a prediction model, and its importance to the significance of the entire project.

First, we had to single out the dataset so that we were only dealing with patients who had a 0 in the ‘Outcome’ variable. This way we were only working with non-diabetic patients in the set. Then, we standardized the data for the filtered dataset, and used the Logistic Regression model that was trained on standardized data in the original dataset on each patient who had a 0 in the ‘Outcome’ column. Then, to finalize the process we created a new observation in the filtered dataset called ‘predicted\_diabetes’ which indicates a value between 0 and 1 to represent the likelihood a patient was diabetic:

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generatedNote – in the 4th column of the above table, notice that the patient has a outcome of 0, indicating that the patient is not diabetic, but in ‘predicted\_diabetes’ the patient has 77%. This means that the patient is 77% likely to be diabetic. It is also worth noting that the patient had 10 pregnancies, which was likely the largest contributor to this percentage.

This above snippet of code is the describe() function in pandas in which gives us some different characteristics of our filtered dataset, and specially our ‘predicted\_diabetes’ column. For example, the ‘max’ value for ‘predicted\_diabetes’ was 97%, which means there is someone in the dataset who is not diabetic but has a 97% chance of becoming diabetic. This person, as stated in the markdown under the table, should be evaluated immediately by medical personnel, because they are so likely to become diabetic.

We then did the exact same thing using our Random Forest model, since it was almost identical in accuracy:

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In hindsight, the differences between the two model’s outputs are relatively miniscule. The differences in numbers are so small, it isn’t noteworthy. However, it does show that both models are just as capable of performing predictions.

Finally, we move on to KMeans Clustering. The goal of this final part was to see if we could group the patients into different clusters based on their diagnostic measurements, what the cluster’s differences are and how they are related to diabetes outcome, and if we could visualize these clusters using KMeans.

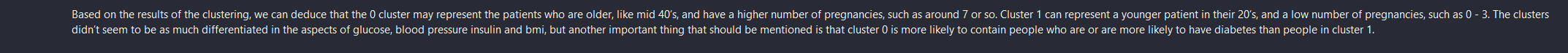
First, we started with a fresh version of the diabetes dataset. We needed to standardize the data and perform KMeans clustering on it. We used 2 clusters, an n\_init value of 20, and random\_state = 42. The cluster centers output is below:  
A computer screen shot of a program

Description automatically generated

We wanted to have an easier way to see that our clusters worked, so we made a new column to represent which cluster each patient was in:  
A screenshot of a computer

Description automatically generated

However, we still had a problem. What do these clusters represent? How are they related to diabetes outcome? Well, using the groupby() function in pandas, we were able to group the means of each variable with correlation to ‘Outcome’ by whether they were part of cluster 0, or cluster 1.

A screenshot of a computer program

Description automatically generated

Based on the results, we deduced that cluster 0 represented patients who were a bit older, such as mid 40’s, and had more pregnancies, in the 6-7 range. Cluster 1 seemed to represent a younger patient, in their 20’s, and with a lower number of pregnancies, in the 2-3 range.

So, how is this related to diabetes outcome? Well, some basic knowledge on diabetes would tell someone that an older person is more likely to get diabetes than a younger person, and a person with more pregnancies is also much more likely to become diabetic than someone with a lower amount. And it would seem our clusters would reflect that, since patients in cluster 0 are more likely to be diabetic or become diabetic than patients in cluster 1.

A screenshot of a computer

Description automatically generatedFinally, we wanted to visualize our clusters distributions, to see if we could find any patterns using t-distributed stochastic neighbor embedding:

As mentioned before, the differences in age and number of pregnancies on average in each cluster were heavily differentiated, however, it would seem a lot of the observations in the set are in different clusters but distributed similarly. This is because a lot of the patients shared many similar factors: such as similar BMI, glucose level, insulin level, etc.

**Conclusion:**

In conclusion, our exploration into diabetes prediction through data analytics has shown data that factors influencing the onset of this widespread metabolic disorder. Through exploratory analysis, we have patterns, correlations, and potential predictors across diverse datasets, establishing a foundational understanding crucial for predictive model development. Beyond individual health implications, this research holds significance in informing preventive strategies, and personalized healthcare approaches. Using this potential data, our aim is to contribute meaningfully to the ongoing discourse on diabetes, fostering a collective effort to alleviate its growing impact on global public health.