MGT 665 Final Paper

 It has long been an objective for healthcare professionals to be able to predict whether a patient will have a health condition. It is never a one-size-fits-all when a physician needs to diagnose a patient when not able to directly test. However, with a machine learning model, we may be able to get a better guess at identifying people who are at a higher risk. Diabetes is a massive health problem all around the world, but it is impossible to know exactly when someone will qualify as a diabetic. This is where I decided to implement a machine learning model to take different attributes of diabetes patients and see if there were factors, we could use to predict if someone was diabetic. The dataset I used was from the *National Institute of Diabetes and Digestive and* *Kidney Diseases* it includes only patients that are at least 21 years old, female, and of Pima Indian heritage. While this dataset is a bit niche and may not apply one-to-one with different populations, it should still give us valuable insight into the likelihood of being able to predict whether we can know if a certain person will have diabetes.

Introduction

Diabetes is a chronic condition that affects how your body regulates its blood sugar levels and how your body turns food into energy. We can detect diabetes when the blood glucose levels are higher than normal for extended periods and quite frequently. The two main types of diabetes are Type 1 and Type 2. Type 1 diabetes is an autoimmune disorder where the body attacks itself by mistake which then prevents the body from producing insulin. It is required for these people to take insulin every single day to be able to survive. It is a relatively uncommon type of diabetes, only accounting for about 10 percent of the total diabetes victims. Type 2 diabetes is the far more common type of diabetes and has developed over many years. It is usually developed from unhealthy weight management and diet, though this is not always the case. It is far more common in adults than it is in children due to these factors. For our dataset, we will not be looking at any children so this will not be relevant for us. Healthcare professionals need to be able to identify whether or not a person is at risk of developing or currently has diabetes. If the physician can do this then they will be able to let the patient know and possibly implement lifestyle changes or medication to help prevent diabetes from fully developing. Healthcare professionals also find it important to be able to predict this due to the ability for them to provide personalized healthcare to the patient and know how to better allocate their resources. The biggest problem we have with being able to accurately predict diabetes with a model is the lack of available data. As we have stated our dataset is a relatively specific population which prevents us from having more encompassing data. Another issue will be consideration of people with Type 1 diabetes. These individuals will stand out as outliers in our data because they were born with the ailment.

Data Analysis and Model Implementation

Our dataset comes from the *National Institute of Diabetes and Digestive and* *Kidney Diseases.*The dataset has 768 observations and 9 different variables. Our variables include:

* Pregnancies: Number of times pregnant
* Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
* BloodPressure: Diastolic blood pressure (mm Hg)
* SkinThickness: Triceps skin fold thickness (mm)
* Insulin: 2-Hour serum insulin (mu U/ml)
* BMI: Body mass index (weight in kg/(height in m)^2)
* DiabetesPedigreeFunction: shows the probability of diabetes based on family history
* Age: Age (years)
* Outcome: Class variable (0 or 1)

*\*Description of variables from Kaggle website, reference listed below\**

The first step of building the model is preprocessing our data, which will involve checking for any missing values and checking for outliers. There were no missing values within our data set so there was nothing we had to do for that. However, there were plenty of outliers within the dataset or 80 to be exact. This is about 10 percent of our dataset which would make plenty of sense when we take into consideration that about 10 percent of the diabetic population is Type 1 diabetic. For this, we will simply remove these outliers from our dataset. It will make our model more accurate and this data is irrelevant since their condition is not developed but rather it is genetic. It is possible some of these outliers are not Type 1 cases but we can safely assume that the vast majority are and for the simplicity of our model we will exclude this. So our final data set that we will use for the model has 688 observations and 9 variables.

            I chose to create a supervised logistic regression model for our dataset. This model will give us a binary response of yes or no for whether we can say our patient has diabetes. I felt as though this was the best option for our dataset given the simplicity of it and its interpretation. It is also a highly robust model that can handle large datasets that are high dimensional. The next step for building the model was to split the dataset into training and testing datasets. I decided to go with a 70/30 split for our data which left me with a training dataset of 486 observations and 9 variables, and with a testing dataset with 206 observations and 9 variables. This is done to train the model, with the training dataset, and then use the testing dataset to test our model's performance and see its accuracy. Once this is done we can now run our logistic regression model and interpret the results.

Model Evaluation

I used 4 different metrics to interpret the performance of my model. These are; Accuracy, Precision, Recall, and F1 score. The accuracy of the model was 78.16% and this number shows us the percentage of true positives and true negatives, and it means we can accurately predict whether a patient has diabetes with 78.16% accuracy. This is a relatively high accuracy, however, I would be cautious as a healthcare professional to take any results from this model as a “law of the land”. The Precision of the model is 75.56% which means that our model is about 75.56% accurate when predicting a true positive. The Recall of the model is 50% which indicates that the model is predicting about 50% of the positives in our dataset as a true positive. The F1 score of the model is 60.18% and the number represents the balance between Recall and Precision. It checks for any imbalances in our model, however, our model seems to indicate reasonable balance. Overall, our model seems to be able to provide reasonable results when looking to provide valuable insights to healthcare professionals. Our model is not accurate enough to be able to say with 100% confidence someone does or does not have diabetes but it provides valuable insight into who may be more at risk of developing the disease. Knowing this is the outcome of our model we can then also interpret which variables have the biggest correlation with diabetes. The significant variables, which have P-values under .05, are number of pregnancies, BMI, glucose levels, and Diabetes Pedigree Function. A physician can look at all 4 of these variables and then be able to help patients be able to proactively prevent diabetes.

Ethical Implications and Conclusion

There are many positives that can come out of a model that tries to predict the outcome of an underlying condition using just common variables. If accurate, it can save patients and physicians time and money not having to spend money on test kits to tell whether or not they have diabetes. More realistically, it provides valuable information on how a patient can prevent the development of a disease like diabetes. Ultimately, since our model is not accurate enough for us to definitively say whether or not someone has diabetes, it is much more likely to be used as a way for physicians to identify patients at risk. This is not to say that a model like ours cannot be used to create accurate predictions but I believe this would require multiple things. For one I believe that we would need a dataset that encompasses a much bigger population of people. Given the fact our dataset was very specific, it may create some biases that prevent our results from being used across different populations. Another important item to consider would be getting rid of patients who have Type 1 diabetes. These patients are not relevant to our research given the fact that their condition cannot be developed and is given to them at birth. Even though we believe we threw out the outliers in our dataset we cannot be 100% sure all of them were Type 1 patients. We also cannot be sure that all the patients in our dataset were Type 2 patients. With better data, there would be a real promise for a machine-learning model like this one.

References

Akturk, Mehmet. “Diabetes Dataset.” *Kaggle*, 5 Aug. 2020, www.kaggle.com/datasets/mathchi/diabetes-data-set?resource=download.

Balkau, Beverley, et al. “Predicting Diabetes: Clinical, Biological, and Genetic Approaches: Data from the Epidemiological Study on the Insulin Resistance Syndrome (Desir).” *Diabetes Care*, U.S. National Library of Medicine, Oct. 2008, www.ncbi.nlm.nih.gov/pmc/articles/PMC2551654/.

“What Is Diabetes?” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 5 Sept. 2023, www.cdc.gov/diabetes/basics/diabetes.html.