Diabetes

Lara Mechling, Corrina Hanson, Isaac Liem

Bellevue University

DSC 450 Applied Data Science

Professor Alsaleem

October 23, 2022

Table of Contents

Introduction

“Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy” (CDC, n.d.). The disease effects over 37 million Americans (CDC, n.d.) and can cause significant health concerns. " Early detection is key in diabetes because early treatment can prevent serious complications. When a problem with blood sugar is found, doctors and patients can take steps to prevent permanent damage to the heart, kidneys, eyes, nerves, blood vessels, and other vital organs” (Falcone, 2020). Using machine learning predictive algorithms can aid in detecting diabetes in a healthy population to further aid doctors in early detection and prevention of Diabetes related complications. The datapoints used in the dataset include, among others, skin thickness, age, and BMI. These attributes of the dataset are all factors that play a role in patients with Diabetes. According to Collier et al. “skin thickness was increased and significantly related to duration of diabetes” (1989), and according to Helmer “age is a big risk factor and an estimated 14% of Americans ages 45 to 64are diagnosed with diabetes which is almost five times the rate for those 18 to 44” (2022). Along with these factors “individuals affected by excess weight, particularly obesity and morbid obesity, are more likely to develop diabetes as a related condition of their excess weight” (Understanding excess weight and its role in type 2 diabetes, n.d.), and "high blood pressure is twice as likely to strike a person with diabetes than a person without diabetes" (Johns Hopkins, n.d.). Using the dataset provided by Kaggle we touch on some of the factors most affecting a person’s predisposition too diabetes and use machine learning to predict the onset of the disease.

Business Problem

Doctors are tasked with diagnosing Diabetes every day and early detection of the disease is paramount in preventing life threatening complications. Using a machine learning model we will input patient data, including some of the largest risk factors of Diabetes, and predict if the patient has the disease. This model will be able to aid Doctors in their quest for early detection.

Method

The diabetes data was downloaded from Kaggle and imported into Python using Pandas. The .info() function was used to check for null values and check the data type for every variable. Exploratory data analysis was performed by visualizing counts, means, stand deviations, min/max, etc., of each variable. A Pandas profile report was also generated for further understanding of the data. MinMaxScaler was used to normalize – or scale – all numerical variables to increase model accuracy. For reference, all variables in the data were integers or floats.

The target variable for this project is ‘outcome.’ More precisely, does an individual have diabetes. The model chosen was a random forest classifier. Random forest classifier (RF) is an ensemble method that uses many decision trees in order to make the final prediction. "RF is a multifunctional machine learning method. It can perform the tasks of prediction and regression. In addition, RF is based on bagging, and it plays an important role in ensemble machine learning" (Zou, et al., 2018).

In order to create the model, the x and y variables were assigned. The x variable consisted of every data point, excluding the ‘outcome’ column, which was assigned as the y variable. The train, test, split function was then used to assign 70% of the data to training and 30% retained for testing. The RF classifier was instantiated using n-estimators = 50 and the model was fit. Once the model was fit, the accuracy, precision, and recall were calculated. The accuracy was 77%, the recall was 68%, and the precision was 60%. These values will change if the model is rerun but will remain relatively stable, this is because no random seed was assigned.

Feature importance was another metric to be found for this project. The function ‘feature_importances_’ from the sklearn.ensemble library was used to find how much each variable affected the outcome. Of the variables, blood pressure was the one that seemed to hold the most importance when determining whether someone has diabetes.

Results

The best evaluation metric, when it comes to classification of this type, is a confusion matrix. In this model, the goal is to correctly identify true cases of diabetes very accurately. A false positive is much less impactful than a false negative. In our test evaluation there were 229 patients. The model incorrectly predicted true 31 patients and false 22 patients. However, the model predicted 130 false cases and 46 true cases correctly. The false cases show good signs of model performance, but more tuning for true cases could be used.

Recommendations and Ethical Considerations

Further evaluation and tuning of the model plus exploring other models could prove useful in providing more impactful results. There were no ethical considerations or biases that needed to be noted during this study.

Conclusion

Proper identification of true cases of patients with diabetes is the paradigm of this study. Accurate predictions based upon common testing would accelerate the pathways to treatment for individuals beginning life with diabetes. As noted, diabetes is a chronic condition that if caught early enough can prevent severe damage to organs in the body. It is crucial to catch as many cases as possible before the disease can wreak havoc in patients' bodies. If the cost is false positives, this is a better outcome than predicting falsely when the patient does not have the disease. It is paramount that more research be done to improve patients' outcomes. The accuracy score of 77% "can indicate machine learning can be used for predicting diabetes, but finding suitable attributes, classifier and data mining methods are very important" (Zou, et al., 2018).

References

CDC. (n.d.). *The Facts, Stats, and Impacts of Diabetes*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/diabetes/library/spotlights/diabetes-facts-stats.html#:~:text=37.3%20million%20Americans%E2%80%94about%201,t%20know%20they%20have%20it.

Collier, A., Patrick, A. W., Bell, D., Matthews, D. M., Macintyre, C. C., Eing, D. J., & Clarke, B. F. (1989). Relationship of skin thickness to duration of diabetes, glycemic control, and diabetic complications in male IDDM patients. *National Library of Medicine*, 309 - 312.

Falcone, S. (2020, November 30). *Why Early Detection is Key in Diabetes*. Retrieved from My Virtual Physician: https://myvirtualphysician.com/2020/11/30/why-early-detection-is-key-in-diabetes/#:~:text=Early%20detection%20is%20key%20in%20diabetes%20because%20early%20treatment%20can,vessels%2C%20and%20other%20vital%20organs

Helmer, J. (2022, April 9). *How Age Relates to Type 2 Diabetes*. Retrieved from Web MD: https://www.webmd.com/diabetes/diabetes-link-age#091e9c5e81edf172-2-6

Johns Hopkins. (n.d.). *Diabetes and High Blood Pressure*. Retrieved from Johns Hopkins Medicine: https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure

Understanding excess weight and its role in type 2 diabetes. (n.d.). Retrieved from Honor Health: https://www.honorhealth.com/medical-services/bariatric-weight-loss-surgery/patient-education-and-support/comorbidities-type-2-

diabetes#:~:text=Being%20overweight%20(BMI%20of%2025,to%20your%20own%20insulin%20hormone

Zou, Q., Qu, K., Lou, Y., Yin, D., Ju, Y., & Tang, H. (2018, November 6). Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontier Genetics*. Retrieved from https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full

Appendix

Kaggle Dataset Variables:

Number of pregnancies

Glucose

Blood pressure

Skin thickness

Insulin

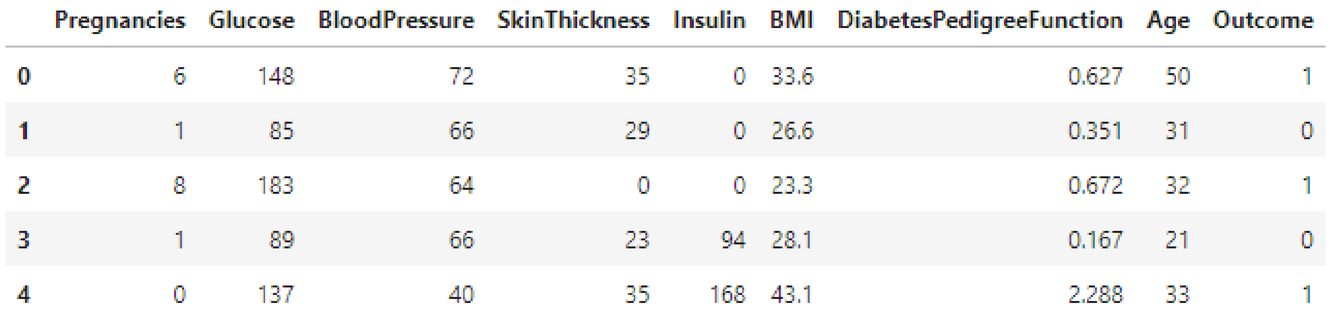
BMI

Diabetes pedigree function

Age

Outcome

Dataset Preview:



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1