SImran Bhalla

Data Mining - Dr. Guha

October 8th 2020

Lit Review and Research Question

The data set that I have chosen for my project is predicting whether a patient is diabetic or not. This dataset is really important to me because this health issue is very prevalent in my family. The girls on my Dad's side of the family have to constantly make sure they are taking precautions to avoid becoming diabetic. Being able to create a logistic regression analysis on health issues like this is very helpful for patients to look at and figure out what they can do to avoid a health issue or figuring out if they are at risk or not. Being able to look through data like this is fascinating and I am excited to see what I can come up with and share with my family.

The binary condition and value in this data set looks at whether or not a patient is likely to become diabetic or not. This outcome can be determined by looking at and basing it off of the categorical values. If the outcome was a 1, that would mean that the patient is likely to get diabetes. On the other hand if our outcome was 0, that would mean that it is unlikely for the patient to get diabetes. Some of the categories that are looked at are how many pregnancies the patient had, glucose level, blood pressure level, the thickness of skin, Insulin level, and age. It is also important to note that the people who were asked to fill out the dataset are of age 21+. I know that with a logistic regression model, the curve we look for is made by looking at different categorical values and determining the odds of getting the targeted value. This targeted value is the binary classification or in our case, the likelihood of a patient being diabetic. I really like this dataset as well because there are no major outliers that will throw off my analysis or mess with my model after interpreting what the logistic regression would look like.

Sources and Citations

[1] Frost, J., Adusei, C., Peeyush, Siyabonga, Sreeja, Narayanan, J., . . . Sachin. (2019, June 13). Identifying the Most Important Independent Variables in Regression Models. Retrieved October 08, 2020, from https://statisticsbyjim.com/regression/identifying-important-independent-variables/

[2] T anyildizderya. (2019, September 10). Diabetes Prediction with Logistic Regression. Retrieved October 08, 2020, from https://www.kaggle.com/tanyildizderya/diabetes-prediction-with-logistic-regression

[3] What is Logistic Regression? (n.d.). Retrieved October 08, 2020, from https://www.statisticssolutions.com/what-is-logistic-regression/