Executive Summary: Pima Indian Diabetes

The Pimas are descendants of the ancient Hohokam, who have inhabited the Sonoran desert and Sierra Madre regions for centuries and they adapted to a desert lifestyle of hard labor. But with new settlements, their lifestyle changed to the one with less physical labor. Their low-fat, high-carbohydrate diet changed to one that derived more than 40 percent of its energy from fat. Around the same time, there was a rise in the number of diabetic people in this community. In 1995 and 2010, the National Institute of Diabetes and Digestive and Kidney Diseases conducted a cross-sectional study on the sudden rise of diabetes mellitus in the Pima Indian community. This dataset is a subset of the larger dataset that was collected during the survey with medical indicators like Age, Glucose, BMI, Diabetes pedigree function, Blood pressure, Insulin levels, triceps skin thickness, and the number of pregnancies. We focus only on the data of women aged 21 years and older (768 samples) for our study. We present our main findings in the following paragraph.

Diabetes is a medical condition that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. In such conditions, the blood sugar level remains high in the body. Our study on the Pima Indian dataset, reveals that the level of glucose in the body is the prime indicator of diabetes mellitus among these women. The human body gets its energy from the chemical reaction between glucose and insulin and any situation which negatively affects this chemical reaction imbalances the glucose level in the body, potentially leading to a high glucose level. This supports our finding as glucose being the prime variable in this study. With glucose as the only predictor, we achieved an accuracy of `75` percent. Along with glucose, other medical indicators like the number of pregnancies, BMI, and insulin levels were found to be important. After including these variables in the model, we could achieve an accuracy level of `79` percent, and the model also explained `33.5` percent of the variance in the data. While developing a better understanding of the data, we studied the change in the probability of diabetes for each variable against our prime predictor variable, glucose. We studied the trend only for regions with a sufficient number of samples. Our observation says that the probability of being diabetic varies significantly with the number of pregnancies (for values `0`, `3`, `6` and `9`) in Pima Indian women; women with more pregnancies are at a higher risk. Similar is the case with higher BMI values for log BMI values of `1.3`, `1.4`, `1.5`, `1.6`, and `1.7`), the high-value BMI curve shows a rapid increase in probability with an increase in glucose level. We also observed the insulin variable with four values, `40`, `110`, `180`, and `250`. At lower levels of insulin, the probability changed quickly but that was not the same for higher levels of insulin. We also studied other medical indicators like diabetes pedigree function(DPF), Age, skin thickness, and blood pressure. After careful observation, we conclude that in the Pima Indian community, DPF (for log values `-0.8`, `-0.6`, `-0.4`, `-0.2` and `0`) didn't show a significant impact on the probability of being diabetic which follows the original theory of change in lifestyle as the most probable cause. For the age (for values `21`, `30`, `40`, `50` and `60`) of `30` to `50` years, the slight change in glucose level rapidly increases the probability of being diabetic, and females with more number of pregnancies in this age range are highly probable to be diabetic in Pima Indian community. We did not find anything interesting with the skin thickness variable, apart from the fact that it is highly correlated to body mass index variable. The different diastolic blood pressure levels didn't show much variation for different values of blood pressure (for values `60`, `70`, `80`, `90`, and `100`). For high as well as low values, the plots were not very different. We believe that the model with Glucose, number of pregnancies, log-transformed BMI along with Insulin, is the best model and these variables are the best predictor variables to predict diabetes mellitus in Pima Indian females.

Certain improvements can be done on top of this work like conducting an in-depth study with larger sample size. More samples of insulin levels would surely improve the model. For a thorough analysis, we can do a randomized controlled experiment to firmly establish a causal relation between these medical indicators and diabetes mellitus in the Pima Indian community.