Project Proposal | Machine Learning Fall 2019

Diabetes or Nah

Colin Avidano | Quinten Caliendo | John Thomas Parrish | Kunal Patel | Raj Patel

Introduction:

Type 2 diabetes is a serious medical condition affecting over 30 million Americans and approximately 9 in 110 people worldwide (Zheng). Type 2 diabetes is not an isolated condition, and it is often preceded by dietary and overall health concerns. The Pima Indians are a Native American tribe living in Arizona, and their genetic traits have predisposed them to survival on a low-carb diet. Within the past 100 years, the Pima have developed the highest prevalence of Type 2 diabetes in the world, making them of much interest to researchers.

Goals / Results:

We would like to evaluate the performance of many classical machine learning models to accurately classify whether a given person has diabetes. Additionally, we would like to isolate certain features that correlate highly with a diabetes diagnosis.

Methods:

For the purposes of classification, we would like to compare the performance of many common supervised models including a decision tree, a neural network, an SVM, a linear model leveraging PCA, and some form of hybrid or tiered model. In addition to the above, we would also like to test at least one model for clustering analysis, more than likely K-means. We plan to cluster based on a subset of dimensions such as blood pressure and BMI. We will observe any clusters that form and trends in other variables not used in the clustering in order to infer causality.

Discussion:

We would like to achieve an accuracy of 75% with our best predictive model. With our unsupervised models, we would like to observe various patterns and correlations of the diagnosis of diabetes. We expect that high glucose, BMI and low insulin levels will have a strong correlation with the diagnosis of diabetes given the generally known nature of the diseases (Balkau). However, we expect that age will be uncorrelated with diabetes while high blood pressure and diabetes will have a positive correlation. Additionally, we believe it would be interesting to look at the correlation between genetic factors described by the Pedigree Factor of subjects vs environmental factors such as glucose and BMI.

References:

Balkau, Beverley, et al. "Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR)." *Diabetes care* 31.10 (2008): 2056-2061.

Kandhasamy, J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." Procedia Computer Science 47 (2015): 45-51.

Mani, Subramani, et al. "Type 2 diabetes risk forecasting from EMR data using machine learning." AMIA annual symposium proceedings. Vol. 2012. American Medical Informatics Association, 2012.

Zheng, Yan, Sylvia H. Ley, and Frank B. Hu. "Global aetiology and epidemiology of type 2 diabetes mellitus and its complications." *Nature Reviews Endocrinology* 14.2 (2018): 88.

Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." *Frontiers in genetics* 9 (2018).