Deliverable 1

1. Choose of dataset

   There are several medical predictor variables for diabetes diagnosis, such as age, insulin level, blood pressure, etc. This project will attempt to do predict whether or not the patients have diabetes based on several medical predictor variables. I will be using the Pima Indians Diabetes Database from the National Institute of Diabetes and Digestive and Kidney Disease.

2. Methodology
   1. Data Pre-processing
      Since a patient can either have diabetes or do not have diabetes, my model will study the probability of get diabetes affected by the following predictor variables of patients: the number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, and diabetes Pedigree Function, and do classification. I will split the set into three, one contains 70% of data for training the model, one contains 15% for model validation, and a 15% data set for testing my model.
   2. Machine learning model
      I will use the classification model, but I will do more research over witch model is best to be implemented.
   3. Final conceptualization
      To present the model, I will be demoing a simple web/mobile application that will implement this model by taking the number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, and diabetes Pedigree Function as inputs, and display a diagnostic prediction of whether that patient have diabetes and saving the predictions to a database.

References

https://www.kaggle.com/uciml/pima-indians-diabetes-database Data Owner: UCI Machine Learning

Version 1, Last update: Oct 06 2016

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press. http://rexa.info/paper/04587c10a7c92baa01948f71f2513d5928fe8e81