# cebd1160 Final Project: Predicting occurrences of Diabetes

| Name | Date |

|:-------|:---------------|

|Nevin Murad | March 28, 2020|

-----

### Resources

My repository includes the following:

- Python script for my analysis: ` TermpaperV2.ipynb`

- ReadMe Description

-----

## Research Question

Using the PIMA Diabetes data base, can we develop a machine learning model that will predict using one or more diagnostic test results which patient will develop diabetes?

### Abstract

Close to 9% of the adult population worldwide has diabetes. Using a dataset of patient of Pima Indian heritage, we used several medical predictor variables and one target variable identifying if the patient as or not diabetes to build, train and test a Logistic Regression models to predict if a given patient will develop diabetes. In general, the regressor performs quite well predicting correctly 78.125% of the sample. However, potential improvements are required especially with respect to the high number of false positives that can generate stress and anxiety for healthy patients.

### Introduction

In 2017 it was estimated that 425 million people had diabetes worldwide representing 8.8% of the adult population. Diabetes at least doubles a person's risk of early death. Since recent trends suggest that rates of diabetes in the population will continue to rise, it is important to be able to predict who may be affected by diabetes in order to proactively address their risk factors.

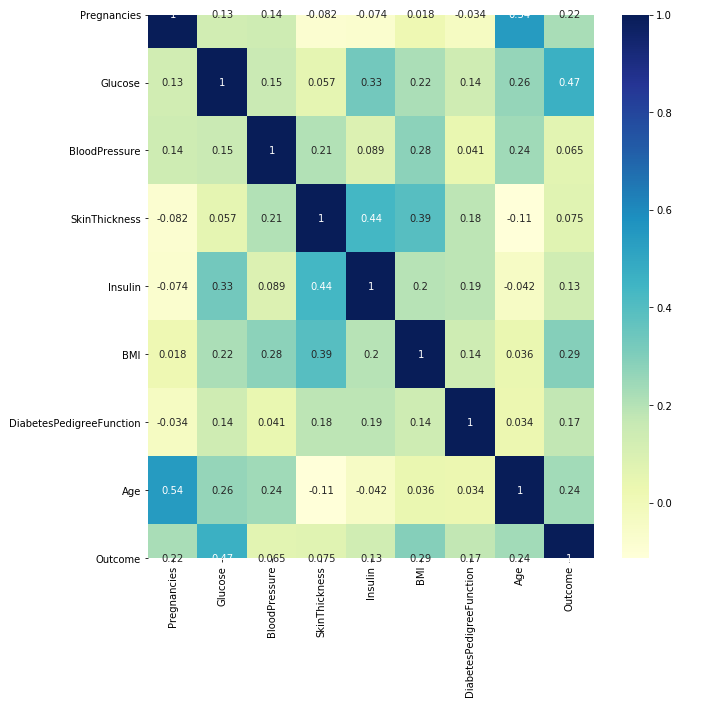
The dataset (<https://www.kaggle.com/uciml/pima-indians-diabetes-database>) originally came from the National Institute of Diabetes and Digestive and Kidney Diseases and includes only female patients at least 21 years old of Pima Indian heritage. The dataset consists of several medical predictor variables and one target variable identifying if the patient as or not diabetes. The predictor variables includes the number of pregnancies the patient has had (Pregnancies), the plasma glucose concentration a 2 hours in an oral glucose tolerance test (Glucose), their diastolic blood pressure (BloodPressure, mm Hg), their triceps skin fold thickness (SkinThickness, mm), their 2-Hour serum insulin (Insulin, mu U/ml), their body mass index (BMI), their diabetes pedigree function (DiabetesPedigreeFunction), their age (Age, years) and their Outcome (Outcome, 1 = tested positive for diabetes, 0 = tested negative for diabetes).

### Methods

The method used for modelling this data was a Logistic Regression models. Logistic regression models are good models for categorical (binary) dependant variable like the one we have as our Outcome variable. The pseudocode can be found at <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logistic%20regression#sklearn.linear_model.LogisticRegression.html>. This method was chosen because Logistic regression have been widely used to predict the risk of developing a given disease (in our case diabetes), based on observed characteristics of the patient in our case the number of pregnancies, the plasma glucose concentration, the diastolic blood pressure, the triceps skin fold thickness, the serum insulin, the body mass index, the diabetes pedigree function and the age.

We slit our dataset into two subsets of data: 1- "training data" and, 2- "testing data". The training data subset will be a bigger subset in order to train the model and the testing data subset will be smaller and used to test the newly trained model on real data. We use a 75/25 split meaning that 75% of the dataset will randomly be selected to constitute the training data subset and the remaining 25% will constitute the test data subset.

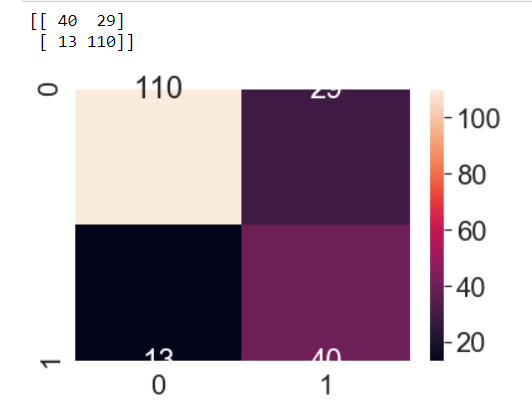
The figure below shows the Correlation Matrix.



### Results

The accuracy score of our model which corresponds to the fraction of samples predicted correctly was 0.78125. The Recall score (also known as sensitivity) which corresponds to the fraction of positives events that is predicted correctly by the model was 0.5797 and the Precision score which is the fraction of predicted positives events that are actually positive was 0.7547.

The figure below shows the confusion matrix where we can see that there were 110 true positives (57.3%), 29 false positives (14.5%), 13 false negatives (6.8%) and 40 true negatives 20.8%).



### Discussion

We can see that in general the regressor performs quite well but with potential for improvements. Predicting the unset of a disease like diabetes is important in order to proactively address their risk factors. But telling a patient that he/she is likely to have the disease when it is not the case (false positive) can create huge stress and anxiety. Our model produces 14.5% of false positive which is high. We would need to test deleting some variables like Age which seems to show some selection bias since we see a significant left-heavy (younger age) to see if we can increase the accuracy of the model and its recall and precision scores.

### References

The links referenced are imbedded in the discussion above.

-------