## #Capstone1- Pima Indians-Predict the Onset of Diabetes

**Introduction**

The population for this study was the Pima Indian population near Phoenix, Arizona. The population has been under continuous study since 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases because of its high incidence rate of diabetes.

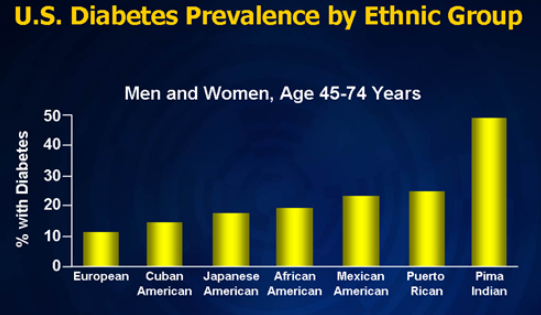
For the purposes of this dataset, diabetes was diagnosed according to World Health Organization Criteria, which stated that if the 2 hour post-load glucose was at least 200 mg/dl at any survey exam or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during the course of routine medical care.

Given the medical data we can gather about people, we should be able to make better predictions on how likely a person is to suffer the onset of diabetes, and therefore act appropriately to help. We can start analyzing data and experimenting with algorithms that will help us study the onset of diabetes in Pima Indians.

**Background on Diabetes and Description**

Diabetes is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. To study the reason that leading to diabetes, a cluster of dataset about Pima Indian Diabetes was used. It is consisted of 8 predict variables and 1 response variable.

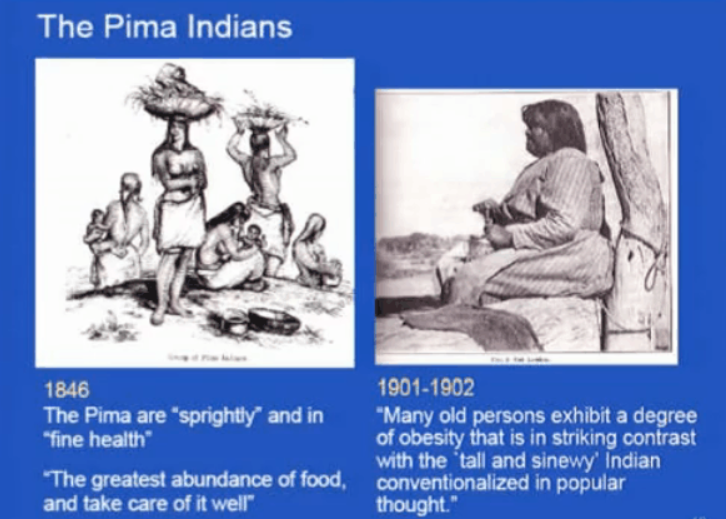
This project emphasizes on predicting the onset of diabetes in Pima Indians. Dataset has been taken from UCI ML repository. Source for this data is from National Institute of Diabetes and Digestive and Kidney Diseases. This contains data in particular, for all patients who are females and at least 21 years old of Pima Indian heritage.



***Fig1- Depicts the onset of diabetes in various ethnic group in US, Pima Indians being the highest***

**Target Audience**

The research people is Pima Indian Female diabetics, which are 768 in the dataset used. The goal of this project is to predict the probability that individual females have diabetes and detect subgroups of characteristics that are at higher risk of diabetes .The higher risk subgroups contain four variables, pregnancies, glucose, BMI and Diabetes pedigree function



***Fig2- Depicts the health history of Pima Indians***

**Objective**

## Final objective of the analysis is to make predictions on whether a person is to suffer the diabetes given the features in the dataset by determining the accuracy using various Machine learning models.

## Data Acquisition- Small dataset! Why?

Dataset has been taken from UCI ML repository .Using dataset for 768 patients, 9 variables were taken to fit in a different statistics model to predict the probability that individual females have diabetes using Train and test dataset (80:20)

Our study begins with an in-depth look of how researchers that used the same dataset approached the same problem. It helped in gaining understanding of the data and pave the way for my project,

In 1988, Smith, Everhart, Dickson, Knowler, and Johannes performed an evaluation of using an early neural network model, to forecast the onset of diabetes mellitus in a high-risk population of Pima Indians. They argued that the neural network approach would provide strong results when *“the sample size is small, the form of underlying functional relationship is not known, and the underlying functional relationships involve complex interactions and inter correlations among a number of variables“*

So, I went further and started analyzing the dataset to understand the correlations among number of variables using various machine learning models.

**Data Description**

We have 768 instances and the following 8 attributes of female patients at least 21 years old of Pima Indian heritage-

* **Pregnancies**: Number of times pregnant
* **Glucose**: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
* **Blood Pressure**: Diastolic blood pressure (mm Hg)
* **Skin Thickness**: Triceps skin fold thickness (mm)
* **Insulin**: 2-Hour serum insulin (mu U/ml)
* **BMI**: Body mass index (weight in kg/(height in m)^2)
* **Diabetes Pedigree Function**
* **Age**: Age (years)
* **Outcome**: Class variable (0 or 1)

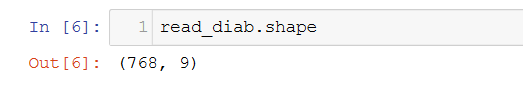
## 

***Fig3- Data description of Pima Indian dataset used from Python Notebook***

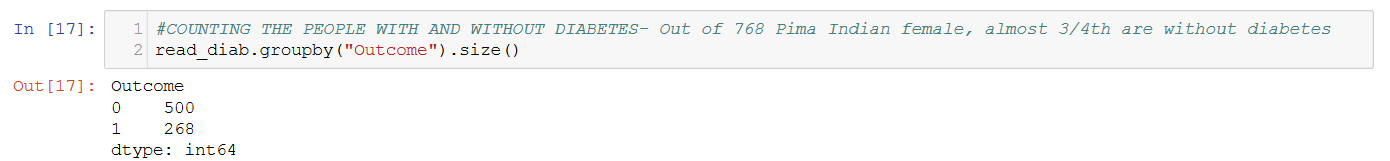
**Data Wrangling**

Steps required to cleanse and modify the data into a necessary format for analysis.

1. CSV files was loaded with 768 records
2. Validated the size of the dataset.

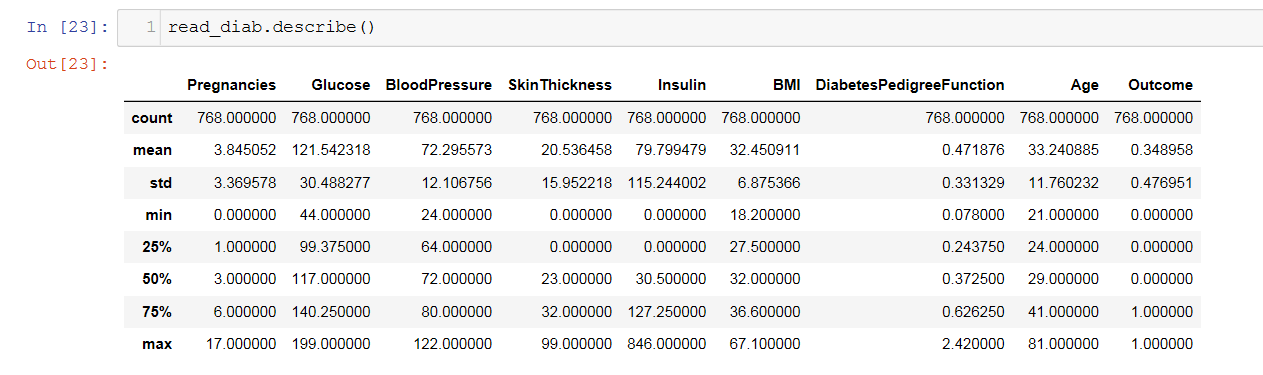


1. Counting the people with and Without diabetes (Out of 768 records). Upon examining the distribution of class values, it was noticed that there are 500 negative instances (65.1%) and 258 positive instances (34.9%).



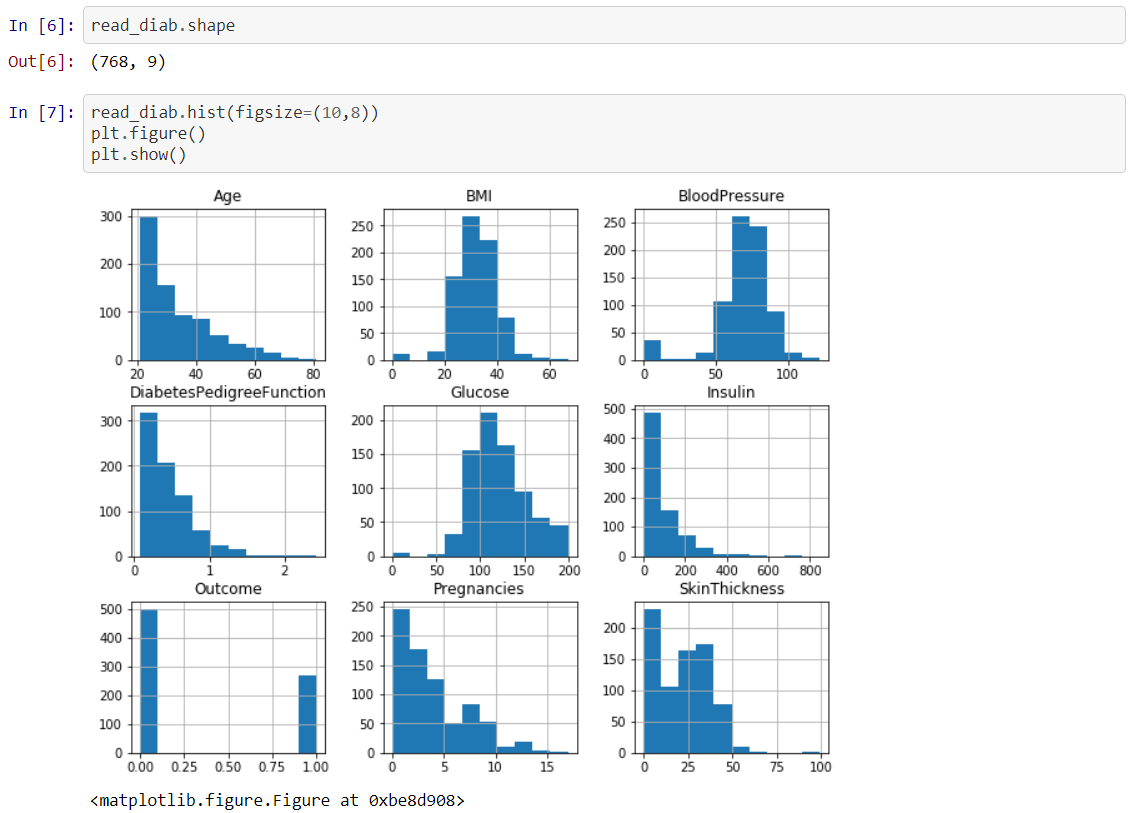
1. The pregnancies and age attributes are integers.
2. The population is generally young, less than 50 years old.
3. Some attributes where a zero value exist seem to be errors in the data (e.g. Glucose, BMI and blood pressure)
4. Filled in missing values, mostly ‘0’ observed for Glucose, BMI and blood pressure. To fill in missing values, mode was used for each attributes for better analysis purpose.

Post data wrangling missing values were filled and same was used in prediction of accuracy



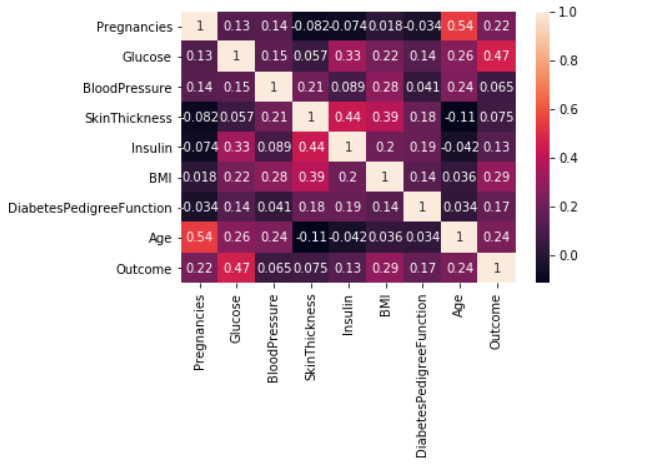
**Exploratory Data Analysis**

Most of the initial data analysis was regarding the shape of the dataset and exploration of each attributes in the data.



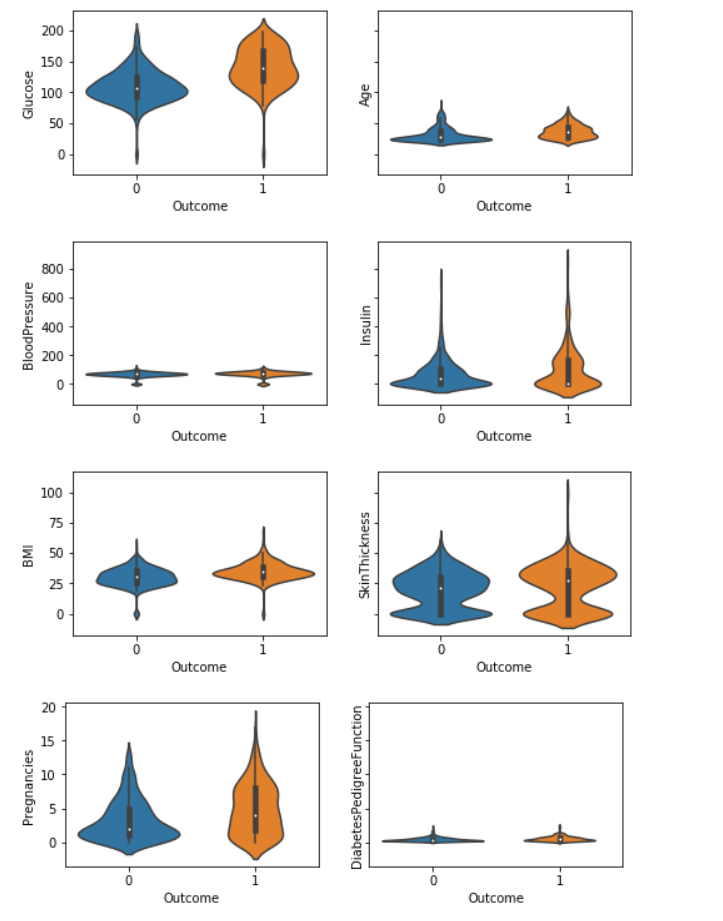
***Fig- Initial visualization of data for each attributes***

**Multiple visualization plot were used for data analysis, below is the snippet for violin and Heat map visualization-**



***Fig- Visualization for Co-relation of each attributes using Heat map***

**Violin Plot Visulaization**



***Fig :-Violin plot visualization for 8 different attributes wrt to Outcomes of diabetes***

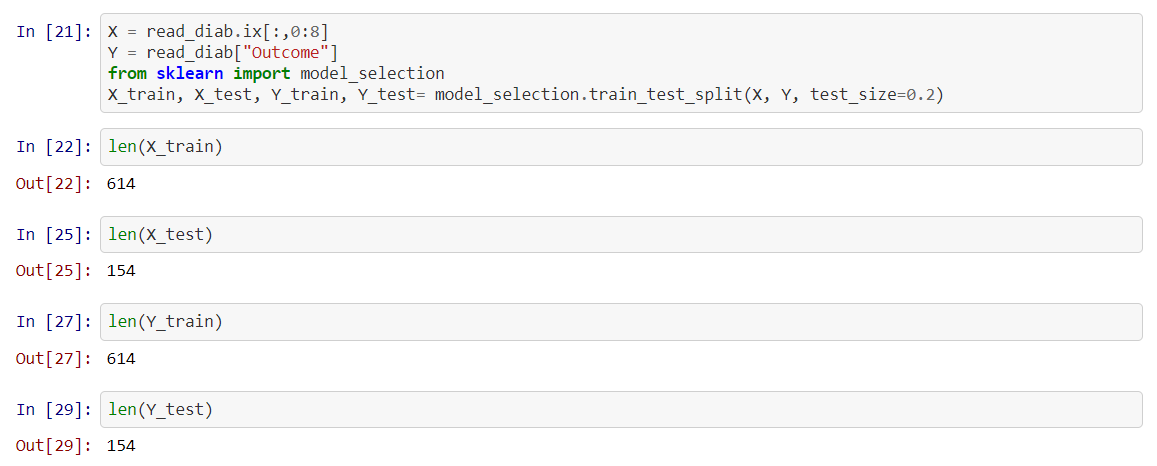
Inference from above visualizations-

1. Graph shows that the glucose ranges from 44 to 200 and there isn't much difference on outcome comparison.
2. But age and outcome comparison shows that the aged person are there who has '0' outcomes which helps to draw the thought that there is more chances of diabetes among young group of pima indians females.
3. Insulin level shows that person having onset of diabetes has high insulin level.
4. Also there seems to be an outlier in skin thickness for people having positive outcome.
5. Looking onto pregnancies plot it seems like there are more pregnant Pima indian females which has onset of diabetes*.*
6. Seems to be no obvious relationship between age and onset of diabetes.
7. Seems no obvious relationship between pedigree function and onset of diabetes.
8. It suggest that diabetes is not hereditary, or that the Diabetes Pedigree Function needs work.

**MACHINE LEARNING**

The primary goal of the machine learning (ML) section is to identify the best algorithm to predict the accuracy of onset of diabetes with various ML models. This is an unsupervised ML exercise and will focus on testing and understanding key parameters as relates to the exploratory data analysis findings above.

For validating the accuracy through each ML model, data has been split into Train and Test (80:20)



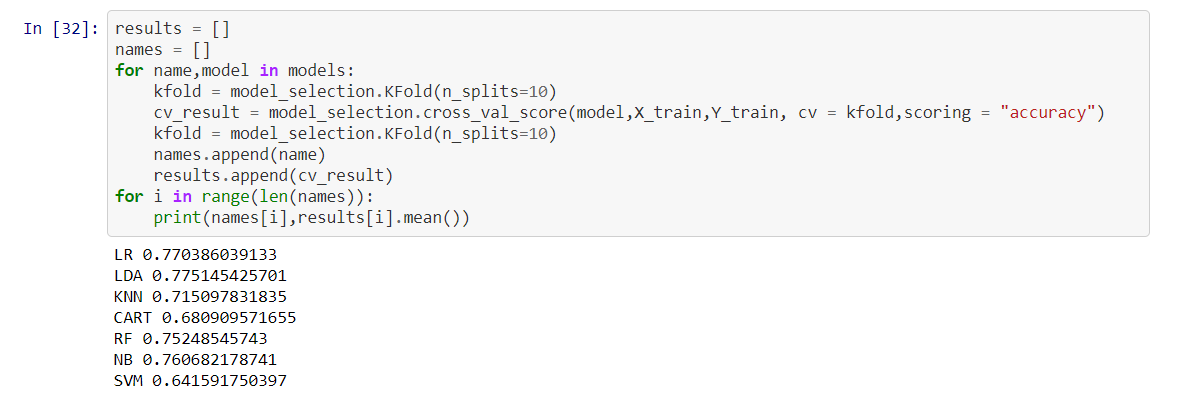
**Machine Learning Algorithms**

For this study, we’ll take a look at the performance of below algorithms:

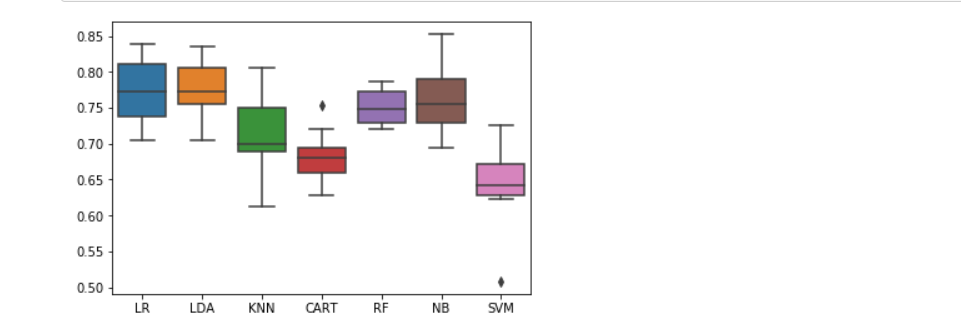
1. Logistics Regression
2. Linear Discriminant Analysis
3. K-Nearest Neighbor
4. Decisions Tree Classifier
5. Random Forest
6. Gaussian Naive Bayes
7. Support Vector Machines

Above ML algorithm will be applied first on train data and the depending upon the performance of the model, we will fit test data to validate the accuracy.

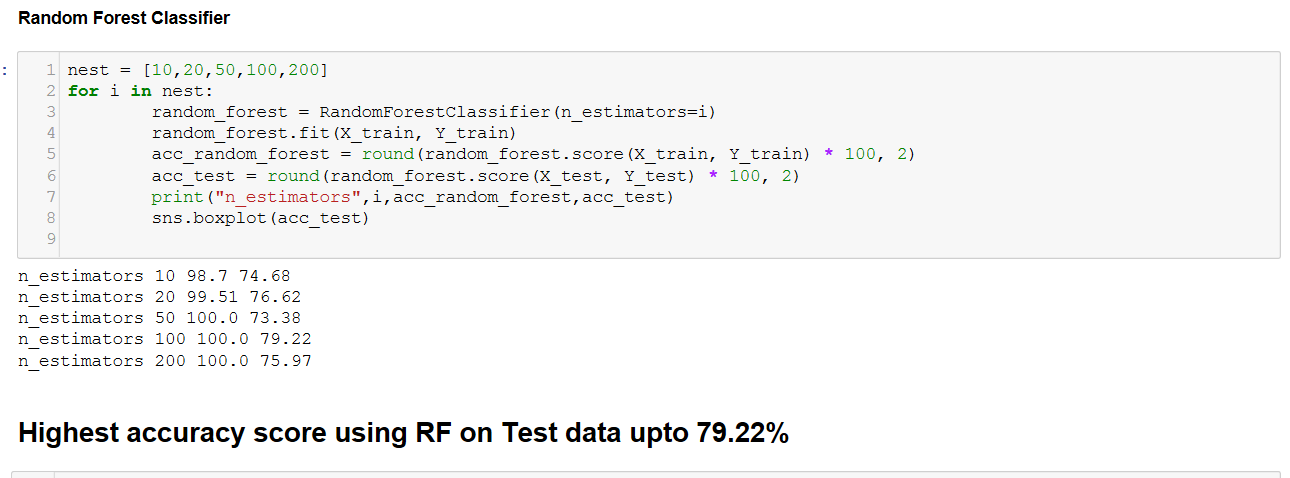
**On applying all of the above ML algorithms using Train data-**



**Below is the box plot depicting the performance of all algorithms-**



Based on above visualization, as these are above 75% accuracy, so I planned to fit test data into LR, LDA and RF model. Out of all best performance comes out from Random forest classifier model.



**Result Summary and Conclusion-**

After performing a cross-validation on the dataset, I focused on analyzing the algorithms which has more accuracy.

Based on testing, accuracy will determine the percentage of instances that were correctly classified by the algorithm. This was an important start of my analysis since gave me a baseline of how each algorithm performs.

I believe it was very interesting to see how our algorithms predict on this scale.

The data here suggests that Logistic Regression, LDA and RF performs the best on the standard, while other performed comparatively low. However, there is no clear winner between any of the algorithms.

I strongly believe that all algorithms will perform rather similarly because we are dealing with a small dataset for classification. However, the 3 algorithms should all perform better than the class baseline prediction that gave an accuracy above 75%.