**Statistics for Data Science(UE19CS203)**

**Pima Indian Diabetes Dataset Analysis**

Pradish Kapur

**Abstract:**

The purpose of this statistical study was to investigate the occurrence of diabetes in a group of women in India typically around the age of 20-30 years with various medical features which include but were not limited to the number of pregnancies a woman has had, her age, her glucose concentration etc. The study was designed in such a manner so as to assess the degree of correlation between these features and diabetes and to infer and draw a conclusion from the data by visualising it using a variety of python tools and libraries. Within the study there were no major findings or trends unearthed that was previously unheard of in the study of diabetes. The study remained quite in tune with the medical knowledge in patients who have diabetes, for example, women with high values of BMI, skin thickness and insulin were more likely to have diabetes than those who did not have such high values. In short, the results can be summarised by inferring that older, unhealthy women make up most of the population of the group diagnosed with diabetes.

**Introduction:**

Diabetes mellitus, commonly known as diabetes, is a metabolic disease that causes high blood sugar. Some of the potential complications of diabetes include: Heart and blood vessel disease. Diabetes dramatically increases the risk of heart disease, kidney failure, stroke, high blood pressure and narrowing of blood vessels (atherosclerosis), and can also lead to nerve damage. Deaths per year due to diabetes and related issues in India are in the range of 5-8 lakhs, and this number is only increasing.

What if such a life-threatening disease could be detected and even possibly curbed before it’s onset? With the data present today, we are sure that it is possible. If the factors that correlate with the onset of diabetes mellitus can be watched out for, corrective measures can be made to ones lifestyle and proper medication can be prescribed in a timely manner too, which would greatly reduce the dangers posed by this disease.

**Dataset:**

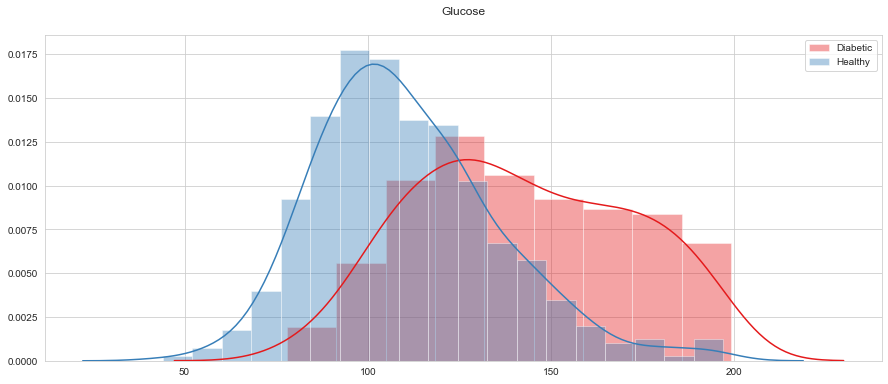
This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes or not, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old of Pima Indian heritage. The dataset consists of the following columns: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and Outcome.

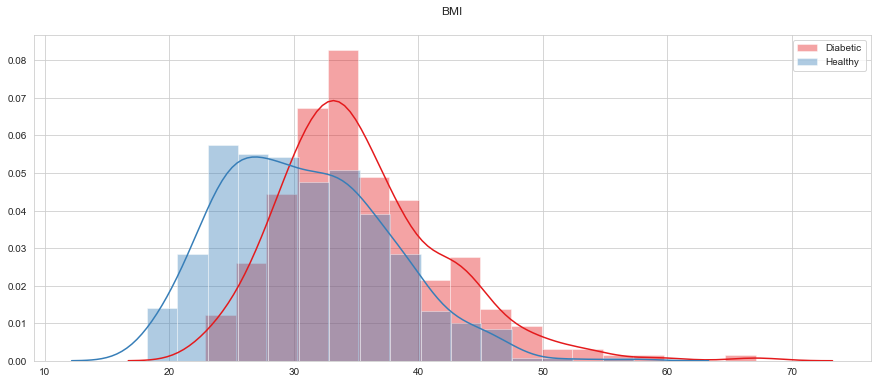
**Data Cleaning:**

The dataset contained multiple nonsensical values such as ‘0’ for skin thickness (which is not logical). Hence first these nonsensical nulls were converted to NaNs. After this the data distribution was viewed using boxplots, and it was found that there were many extreme values or ‘outliers’, although it’s to be noted that these were not removed as they seemed to possible occurrences. The NaN’s were then imputed and replaced by the median for each columns. The median was used here and not the mean as the mean is affected greatly by extreme values.

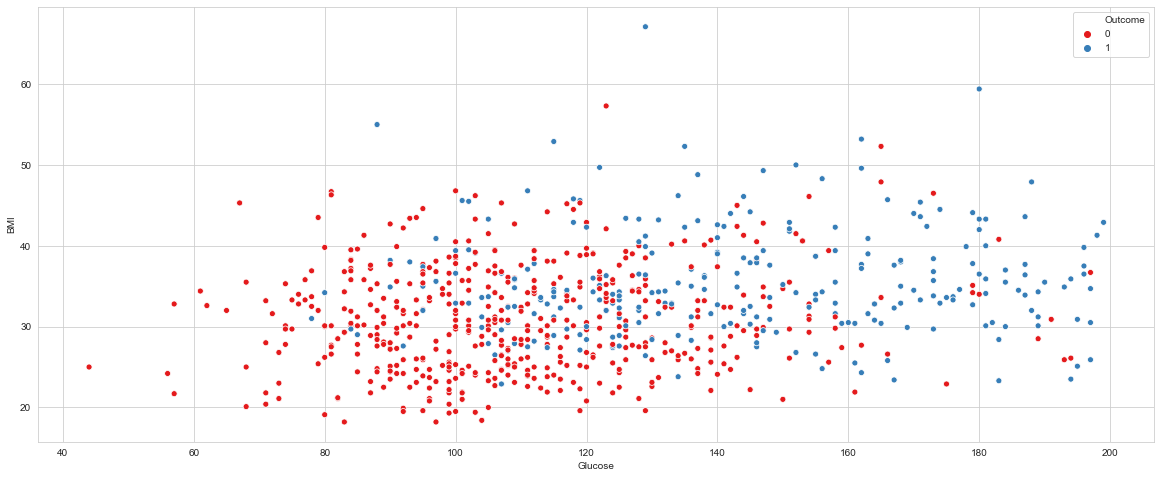
**Exploratory Data Analysis:**

The Exploratory Data Analysis was carried out by plotting various graphs and the correlation heatmap. The insights obtained were:

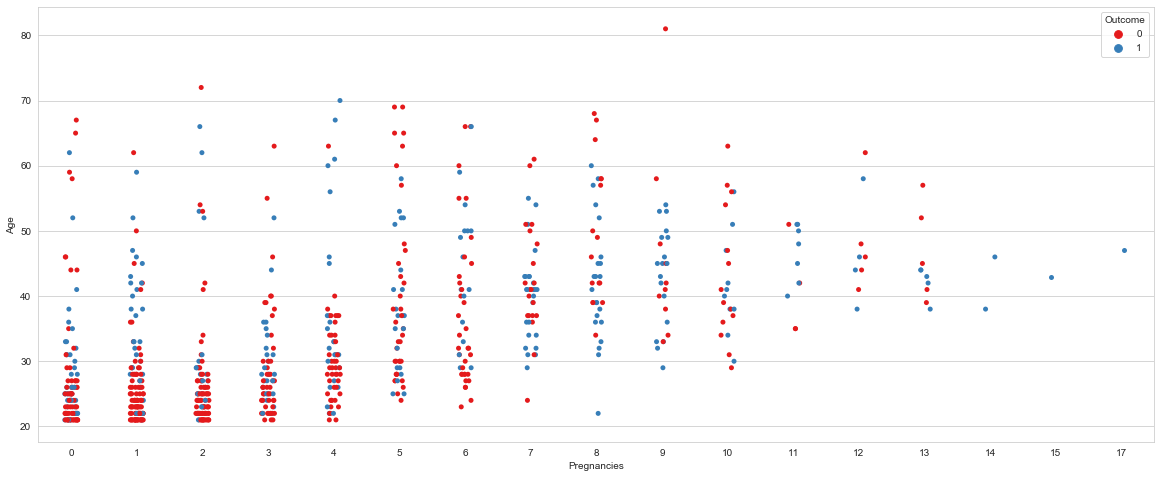
1. Higher the plasma glucose levels, higher the chance of the patient being diabetic. We obtained a median of 107 Plasma glucose concentration level for non-diabetic patients and 140 for diabetic patients. This could be due to your excessive glucose intake or decreased production of insulin. (Values are measures calculated in an oral glucose tolerance test) This can be observed in the below distribution
2. Higher body mass indexes are associated with higher rates of diabetes. Your BMI is a measure of body fat based on your height and weight. It is indicative of how much body fat you have and hence relates with levels of fitness and healthy lifestyles. Therefore, unhealthy lifestyles with unhealthy eating habits and low exercise indirectly relate with higher levels of diabetes.



1. Similar trends can be observed from the scatter plot between Glucose and BMI. Healthy patients are concentrated with BMI <= 30 and Glucose <= 120.



1. Diabetes is more prevalent at higher ages, and the risk of diabetes increases with the total number of pregnancies undergone by the woman. Healthy patients are concentrated with Age <= 30 and Pregnancies <= 6. This is because during pregnancy, the placenta makes hormones that cause glucose to build up in the blood plasma and causes gestation diabetes.



**Hypothesis Testing:**

Null hypothesis H0: 2 independent samples of Glucose (grouped by whether patient is Diabetic or not) have identical average

Alternative hypothesis Ha: 2 independent samples of Glucose (grouped by whether patient is Diabetic or not) have DO NOT identical average

This hypothesis was chosen as it would give us some clarity as to whether the Glucose reading of a patient has some relation with the patient being Diabetic or not. A t-test was performed on the samples (as they were taken from an almost-normally distributed population of Glucose values). The outcome obtained was such that the p-value was lower than the alpha value of 0.01, and hence this two-tailed test led us to reject the null hypothesis H0 and accept the alternative hypothesis Ha.

**Results and Discussion:**

It is observed that the insulin reading is the biggest determining factor of diabetes followed by the glucose reading. There are other factors that could possibly indicate the onset of diabetes such as high BMI’s and increased blood pressure among others. Hence, we can conclude that certain diagnostic measurements can show us signs of the possible onset of diabetes. If detected early on, changes can be made to lifestyle habits and timely medication can be prescribed, hence reducing the risk of major complications caused by diabetes.