**Diabetics Prediction Using Logistic Regression**

1. **Introduction**
   1. **Background**

Diabetes is a very common problem among human beings of any age. Diabetes mellitus (DM), commonly known as diabetes, is a group of [metabolic disorders](https://en.wikipedia.org/wiki/Metabolic_disorder) characterized by a high blood sugar level over a prolonged period of time. Symptoms often include frequent urination, increased thirst, and increased appetite. If left untreated, diabetes can cause many complications. Acute complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death. Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, damage to the nerves, damage to the eyes and cognitive impairment. As of 2019, an estimated 463 million people had diabetes worldwide (8.8% of the adult population). Rates are similar in women and men. Trends suggest that rates will continue to rise. Diabetes at least doubles a person's risk of early death. In 2019, diabetes resulted in approximately 4.2 million deaths. It is the 7th leading cause of death globally. Hence it would be helpful if it can be predicted whether a person is more prone to diabetes based on certain medical conditions.

* 1. **Problem**

The dataset providing details of certain numbers of persons under scrutiny with different medical conditions might be helpful to prepare an algorithm which may be helpful to predict whether a person under consideration may have diabetes or not.

* 1. **Interest**

It would be naturally of interest to medical researcher or practioners. At the same point of time it would be beneficial for an individual being to ascertain whether he/she is in high risk zone or not when it comes to being diagnosed with diabetes.

1. **Dataset**
   1. **Data Source**

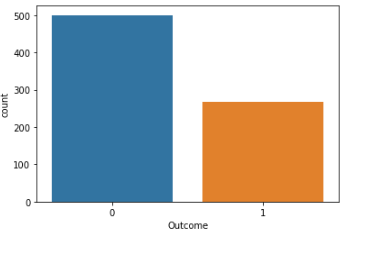
The data being used for this study was made available by “National Institute of Diabetes and Digestive and Kidney Diseases” as part of the Pima Indians Diabetes Database. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above. This dataset was available in Kaggle dataset as an open dataset for study. The dataset has a total of 8 features to determine the final outcome i.e. whether the concerned person is diabetic or not. For this purpose a total of 768 samples being considered

* 1. **Data Checking**

Fortunately none of the column of the available dataset is having any missing value. This helps us to minimize the procedure of data cleaning and preparation. A preliminary check establishes that all the columns are having complete data in terms of binary figure to determine the final outcome.

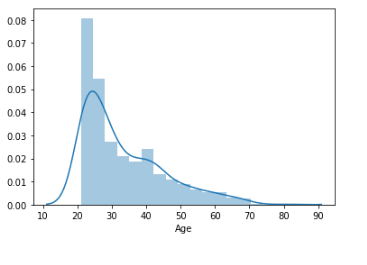
1. **Exploratory Analysis**
   1. **Calculation of Outcome**

At first the overall data is being monitored to check the ratio of sample having positive diabetic condition against negative diabetic condition. The test shows that out of 768 samples 500 samples are negative i.e. non-diabetic whereas 268 samples are positive i.e. diabetic. The following figure depicts the same

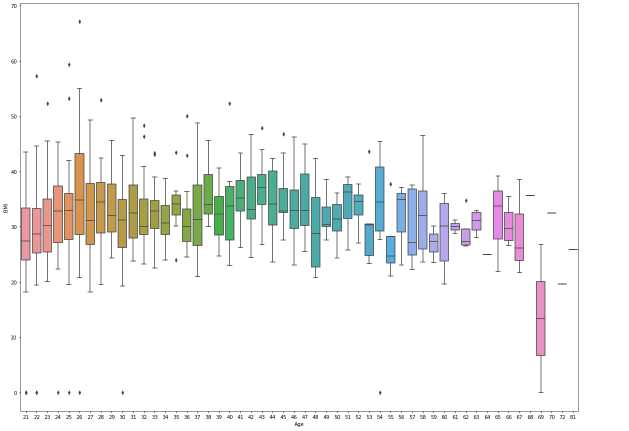
****

* 1. **Variation with age**

Among the different features available it would be interesting to know how the outcome varies with the age. The following figure depicts the understanding of the same

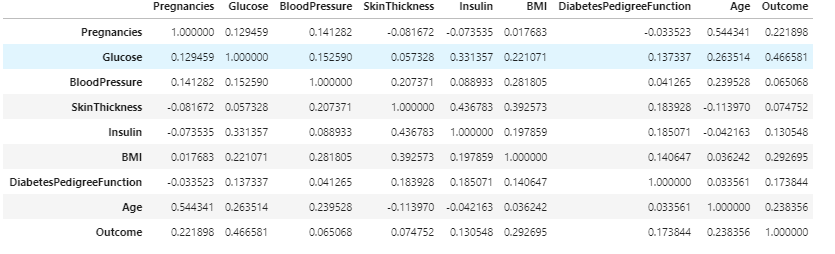
****

A box plot analysis of the outcome at different age may of interest in this case

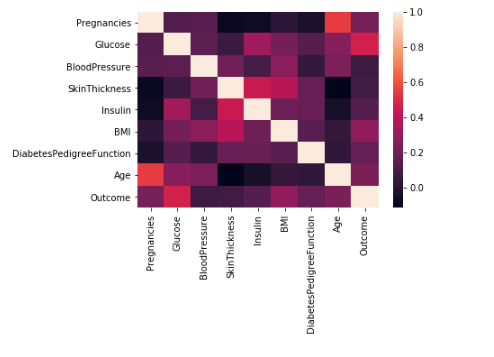


* 1. **Correlation Analysis**

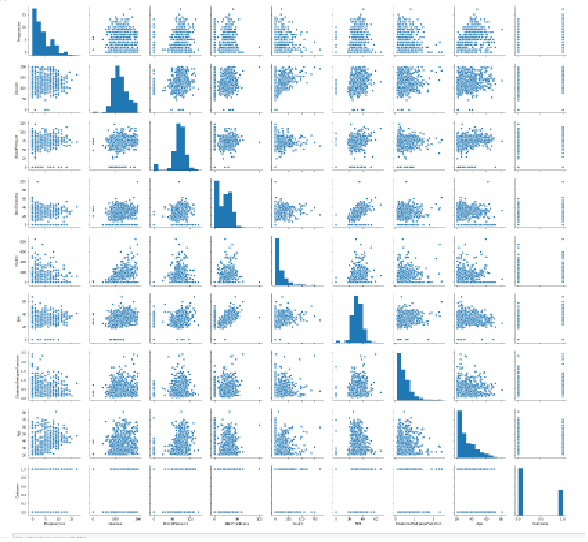
Now it would be important to analyse the correlation between different medical conditions. Though our major aim would be to find out the correlation of the final outcome with different medical conditions, a study of the correlation among themselves won’t be a bad idea. The following table depicts the same



Also a visual representation of the above data in terms of heatmap may be of helpful



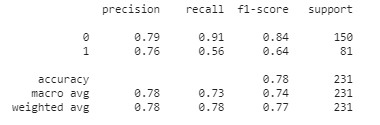
The following figure will depict the scatter plots of different features among themselves as well as the final outcome to visualize the relationship



1. **Predictive Modelling**

For the purpose of predictive modelling logistic regression is being used after splitting the dataset in training set and test set in the ration of 7:3 i.e. 70% of the dataset is being used to train the model and to obtain a satisfactory reliability of the model whereas 30% of the dataset is used to check the level of accuracy of developed model.

After the model development the accuracy of the model is being checked through classification report. The result of the same is as follows:



It signifies a fairly satisfactory accuracy in the range of 72-80%. Also the confusion matrix result as [[136,14],[36,45]] summarizes a fairly accurate model

1. **Conclusions**

The dataset being used though of a certain ethnicity the procedural development of the model may be used for a dataset which reduces this bias. Also the medical conditions being considered for this study is limited to a few nos only and there may be other different reasons which may help to determine the outcome more accurately and thus may increase the accuracy of the model.