**Diabetes Classification problem**

**Problem Statement :**

The disease burden of diabetes is increasing by the day, diagnosis of diabetes at an early stage of the disease is essential for effective treatment and control of diabetes. So to identify accurately whether the subject under analysis has diabetes or not based on a set of variables we compare the accuracy obtained from implementing different models.

**Data:**

We have 768 observations and 9 features in this dataset,out of which 500 are non diabetic while 268 are diabetic people.Independent Features: - Pregnancies, Glucose,BP,Skin thickness,Insulin,BMI, Diabetespedigreefunction (likelihood of diabetes based on family history),Age

Target: Outcome whether subject is diabetic or not (1 ,0 ) binary outcome problem.

**Hypothesis: yet to be done**

**EDA:**

We did not find any null values however there were zero values in features(Glucose,BP,skinthickness,insulin,BMI) which did not make any sense. As our data sample had very few values imputing zero values with any other aggregate value for patients without actually knowing the patients background would be much more harm than boon so we decided to drop those observations.Now We have 392 observations and 9 features in this dataset

We performed univariate and bivariate analysis also (pearson) correlation coeff matrix to answer our **hypothesis**.

**Data Preparation:**

We split the data in 80-20 ratio with stratified sampling to keep the target variable proportion constant

**Modelling:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Recall(TPR)** | **Precision** |
| Logistic Regression | 80.45 | 58.06 | 81.81 |
| Decision tree | 81.60 | 70.96 | 75.86 |
| DT(Pruned) | 78.16 | 80.64 | 65.78 |
| Random Forest | 83.90 | 74.19 | 79.31 |

**Conclusion:**

From above table, as per the accuracy Random forest model turned out to be the best with 83.90%. But since this is a healthcare problem where predicting correctly true diabetic patient is more important than falsely claiming non diabetic as a diabetic patient. Hence as per metric Recall(TPR) which is more important for this problem DT(pruned) model would give much more better results.