CURNEU MEDTECH INNOVATIONS PRIVATE LIMITED

SD03Q09 - 2

Srinidhi S T (1832052)

Problem Statement - 2:

Predict which patient has diabetes from Diabetes Database.csv and try to understand the dataset attributes and try to figure out type ML model suits and build from scratch.

Problem Analysis:

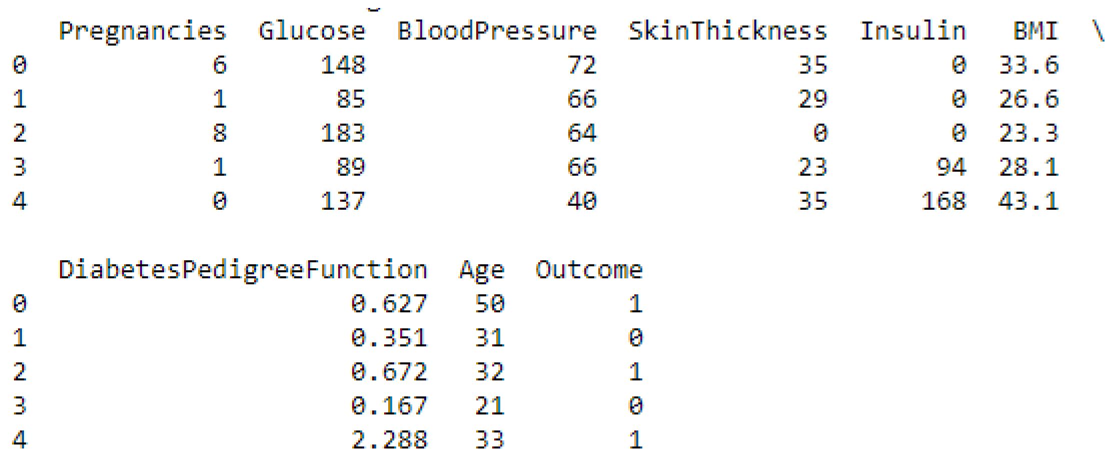
Diabetes is a disease that occurs when the blood glucose level becomes high, which ultimately leads to other health problems such as heart diseases, kidney disease, etc. Diabetes is caused mainly due to the consumption of highly processed food, bad consumption habits. The objective of this project is to build a predictive machine learning model, Random Forest Classifier to predict based on diagnostic measurements whether a patient has diabetes or not based on information about the patient such as Pregnancies (Number of times pregnant), Glucose , Blood Pressure, Skin Thickness, Insulin, BMI (Body mass index), Diabetes Pedigree Function (Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)), Age, Outcome: Class variable (0 - non-diabetic patient, 1 – diabetic patient).

Dataset:

Data before Pre-processing:

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1

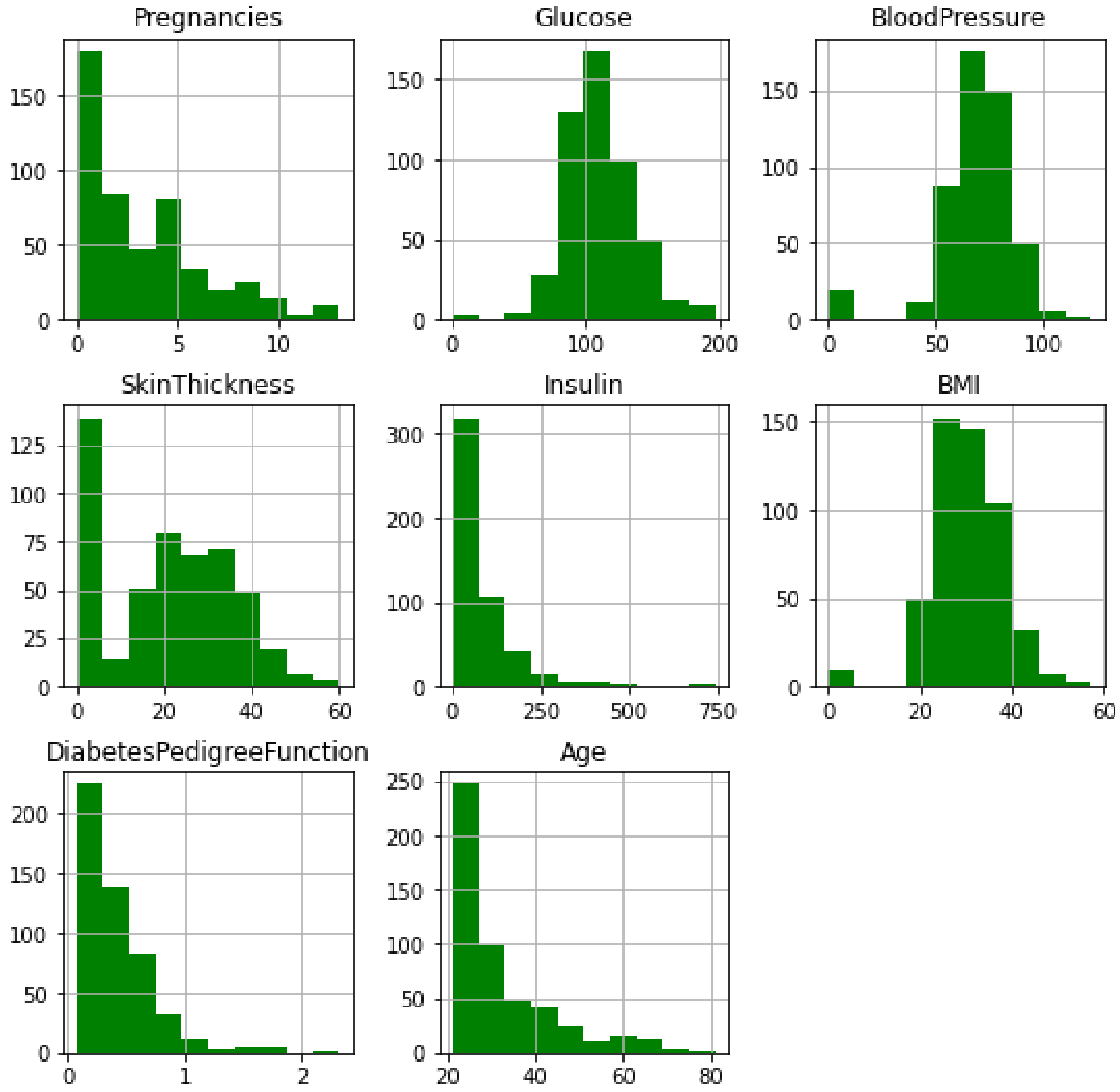
Data after Pre-processing:



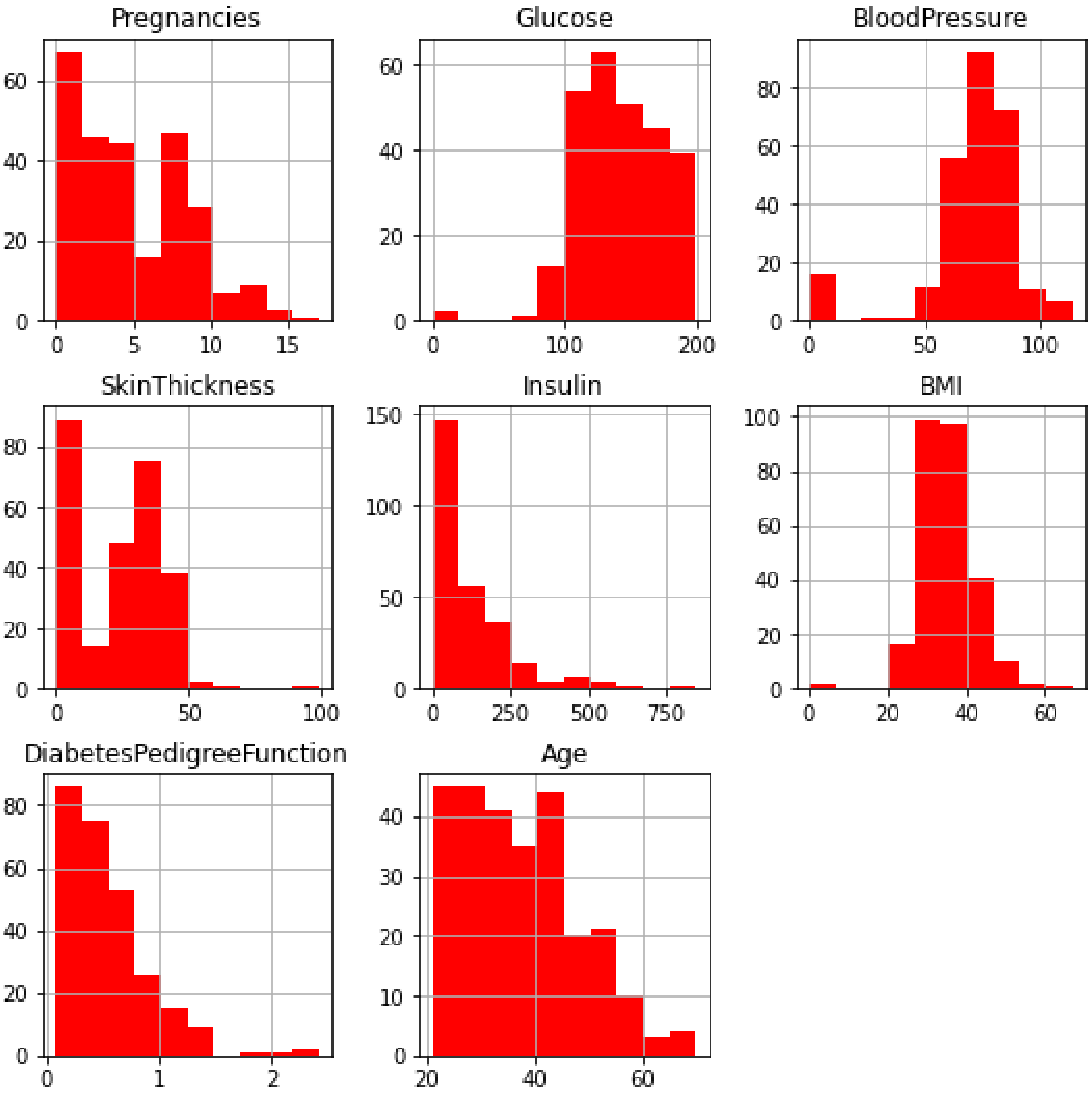
```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
```

Data Exploration and cleaning:

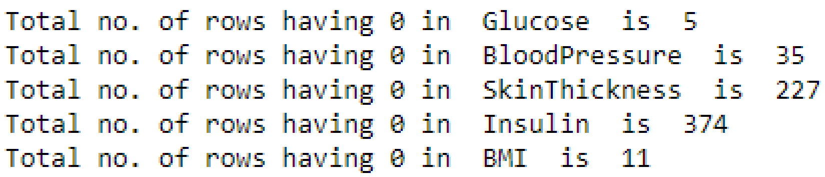
Based on Patients without diabetes



Based on Patients with diabetes



When analyzing the histogram we can identify that there are some outliers in some columns.



```
Total no. of rows having 0 in  Glucose  is  5
Total no. of rows having 0 in  BloodPressure  is  35
Total no. of rows having 0 in  SkinThickness  is  227
Total no. of rows having 0 in  Insulin  is  374
Total no. of rows having 0 in  BMI  is  11
```

Blood pressure: By observing the data we can see that there are 0 values for blood pressure. And it is evident that the readings of the data set seem wrong because a living person cannot have a diastolic blood pressure of zero. By observing the data we can see 35 counts where the value is 0.

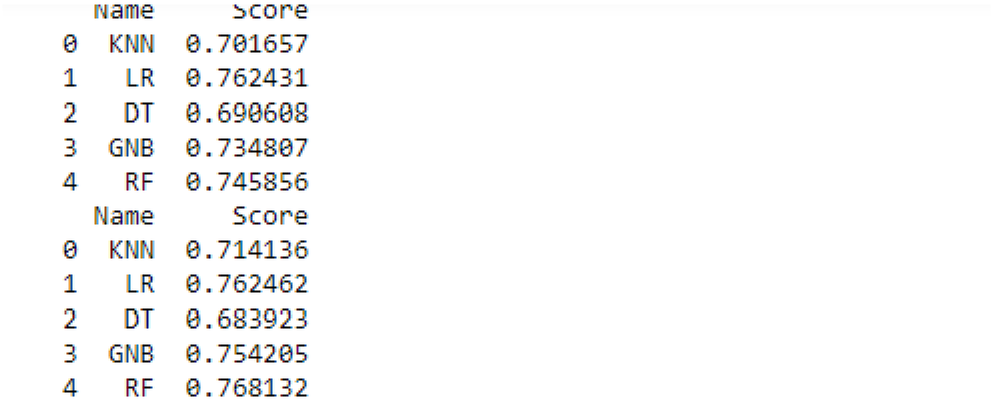
Plasma glucose levels: Even after fasting glucose levels would not be as low as zero. Therefore zero is an invalid reading. By observing the data we can see 5 counts where the value is 0.

Skin Fold Thickness: For normal people, skin fold thickness can't be less than 10 mm better yet zero. Total count where value is 0: 227.

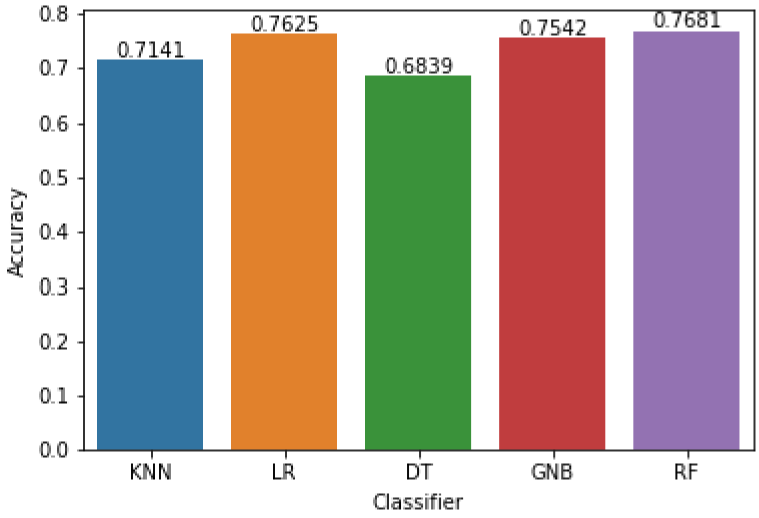
BMI: Should not be 0 or close to zero unless the person is really underweight which could be life-threatening

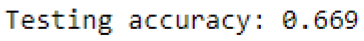
By the end of the data exploration and cleaning process, I can conclude that this given data set is incomplete. So by removing the rows with "Blood Pressure", "BMI" and "Glucose" values as zero we can get more accurate outcome.

Output:



```
  Name     Score
0  KNN  0.701657
1   LR  0.762431
2   DT  0.690608
3  GNB  0.734807
4   RF  0.745856
  Name     Score
0  KNN  0.714136
1   LR  0.762462
2   DT  0.683923
3  GNB  0.754205
4   RF  0.768132
```





```
Testing accuracy: 0.669
```

We can see that the Logistic Regression, Gaussian Naive Bayes, and Random Forest have performed better when compared with other algorithms. From the base level we can observe that the Random Forest performs better than the other algorithms.