**Dataset Details:**

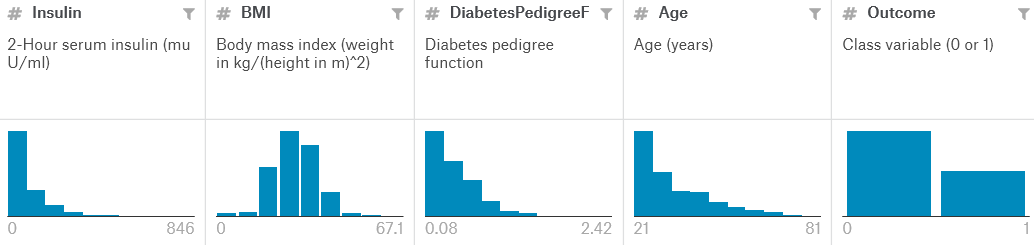
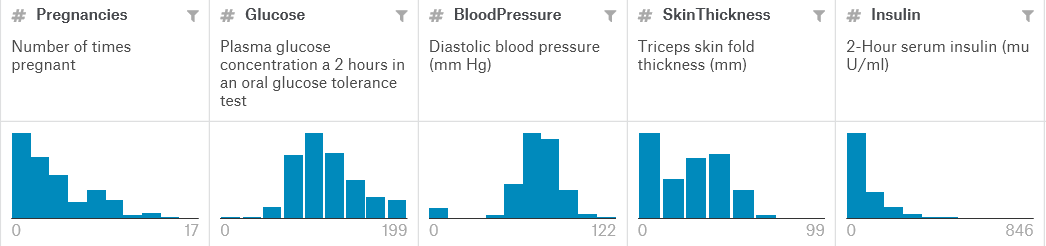
* Data Split

Used 90% of data to train the model and 10% to test. Cross validation was used to pick best 10% percent of the dataset.

Here we are given pima-Indians-diabetes-database(768\*9).   
Where feature sets i.e. Xn (8) are:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Pregnancies | Glucose | Blood Pressure | Skin Thickness | Insulin | BMI | Diabetes Pedigree Function | Age |

And, the training label (Y) which we are feeding to our K-NN classifier is “Outcome”.



**Algorithm Description:**

As we know K-NN classifies the dataset points with respect to the K neighbors.  
The main idea of *data preprocessing* is to remove outliers with acceptable values. Here in this diabetic dataset case many patient showed null values for important feature set which could totally negate our prediction classifier. So, in order to get a relevant model these null values were replaced by the mean values of respective features.

**Algorithm Results:**

* Confusion matrix

So we can see from the matrix that here false positive= 116, false negative=88, true positive=412 and true negatives=152.

[[412 88]

[116 152]]

Max Accuracy is with k = 21.000000 and Accuracy is 0.748698

**Runtime:**

The wall clock time for the program is 78.3020754643324 seconds.