**10th feb’21 DIABETES PREDICTION**

**AIM:**

 Predict Diabetes is there or not given the values of glucose level and blood pressure.

**DATASET:**

This dataset contains records of people with glucose level and blood pressure values

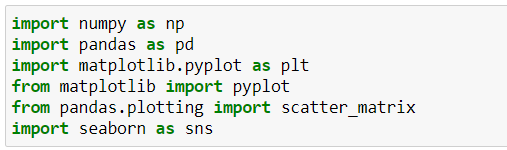
and the target class column as Diabetes or not. 995 records.

**METHOD:**

Build Naïve Bayes Classifier model using the above training data and

predict for new patient the possibility of diabetes.

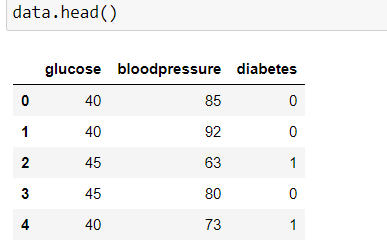
**IMPORTING THE LIBRARIES:**



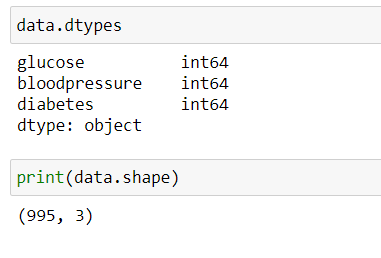
**IMPORTING THE DATA:**



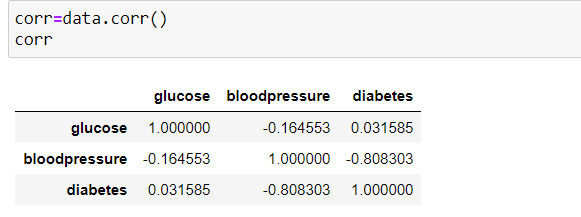
**SAMPLE DATA:**



**DATA ANALYSIS:**



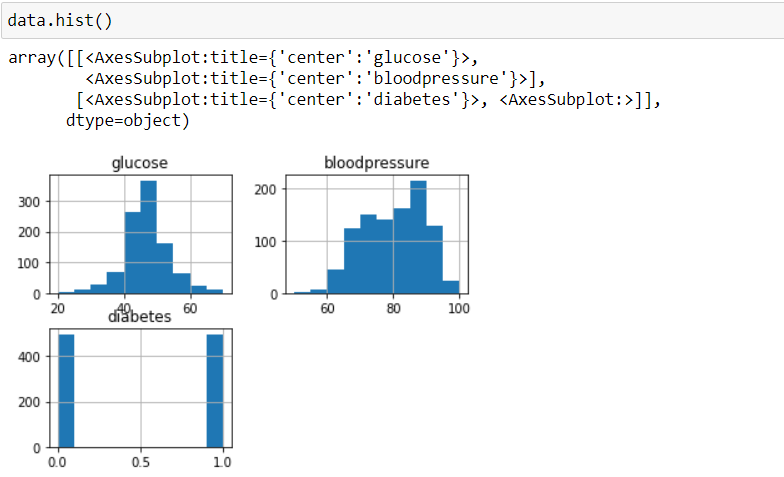
**CORRELATION BETWEEN THE DATA:**



**DATA VISUALIZATION:**

**HISTPLOT:**

A histogram is a classic visualization tool that represents the distribution of one or more variables by counting the number of observations that fall within disrete bins.



From the above plot its clear that

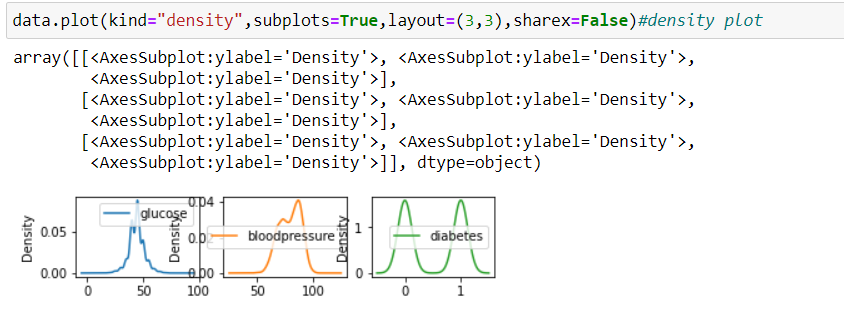
the values of glucose lies between 20 to 85

the values of blood pressure lies between 60 to 100

the values of diabetes are 0 and 1

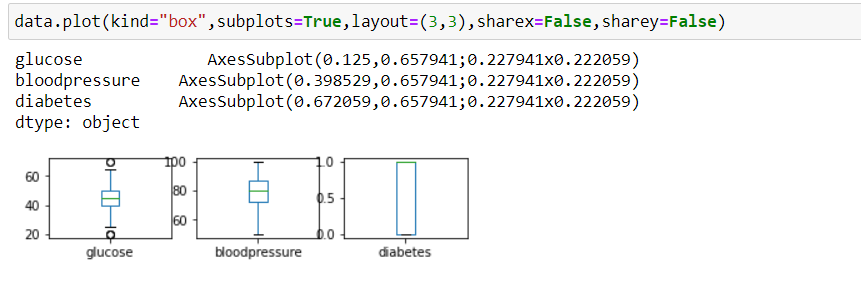
**DENSITY PLOT:**

A density plot is a representation of the distribution of a numeric variable. It uses a kernel density estimate to show the probability density function of the variable (see more). It is a smoothed version of the histogram



**BOXPLOT:**

A box plot shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range.

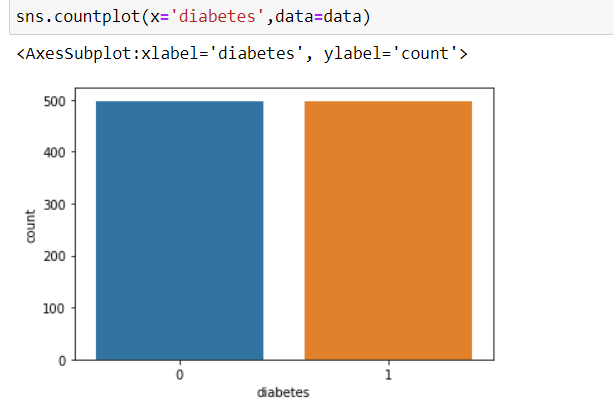


From the above plot we can interpret that the median values of dataset

Glucose – 55

Bloodpressure-80

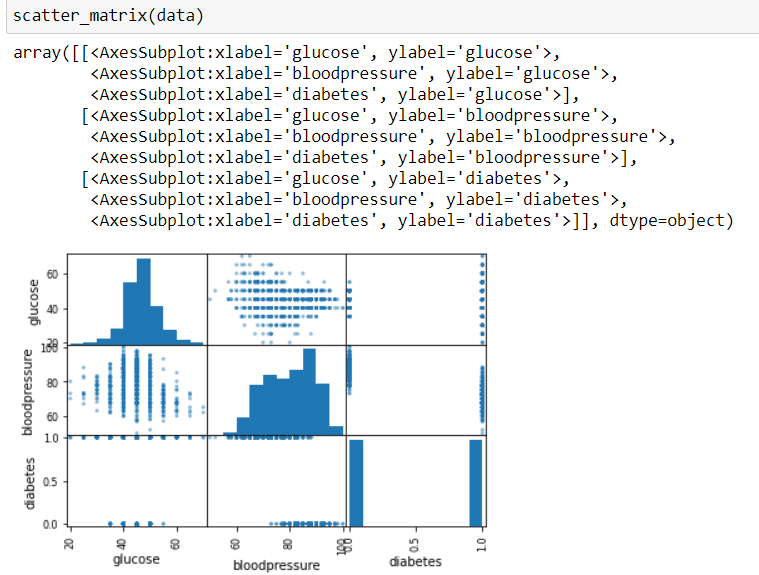
**COUNTPLOT:**



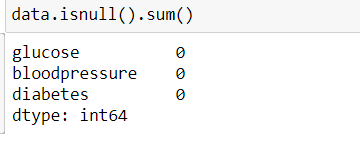
From the above plot we can visualize that the dataset contains equal number of diabetes patients and non-diabetes patients records

**SCATTER MATRIX:**

A **scatter** matrix is a estimation of covariance **matrix** when covariance cannot be calculated or costly to calculate.

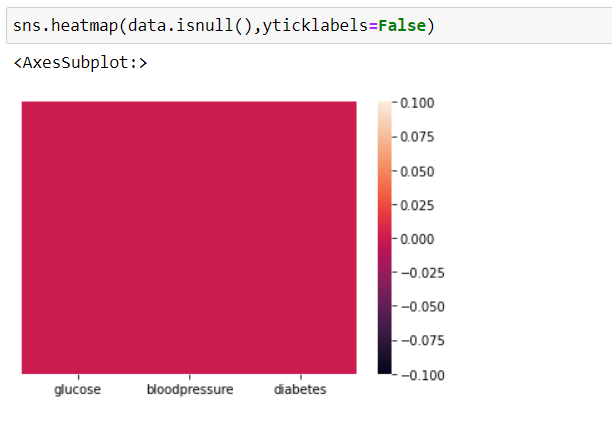


**DATA WRANGLING:**



Hence the data has no missing values

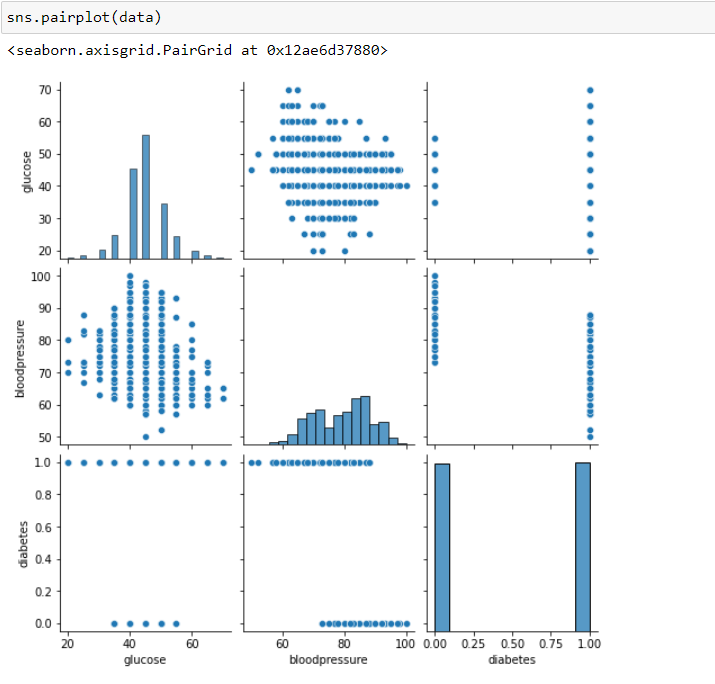
**HEAT MAP:**



To confirm there is no missing values

**PAIRPLOT:**

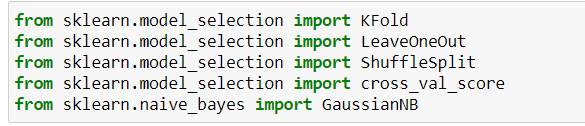
To visualize the relationship between each variable in a matrix form.



**DATA PREPARTION:**



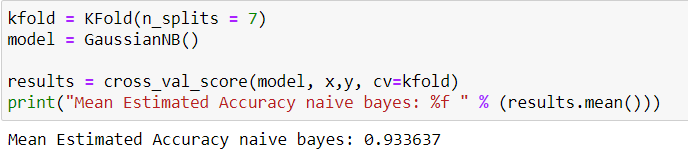
**IMPORTING THE LIBRARIES:**



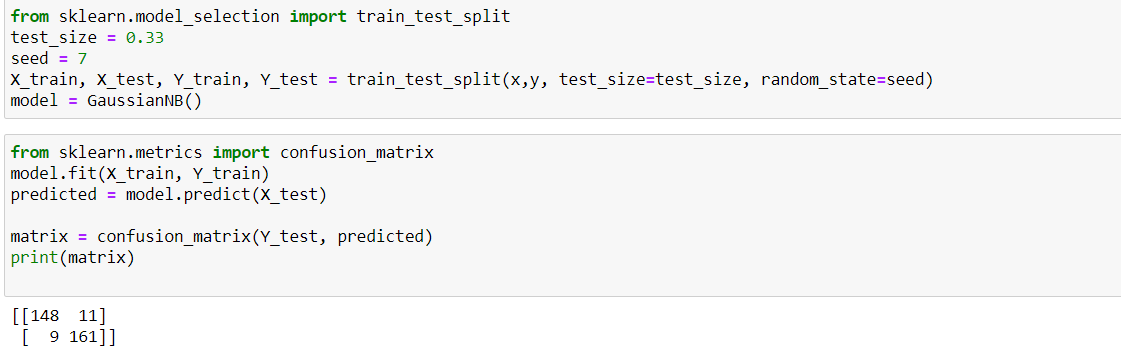
**EVALUTATION TECHINQUE:**

**K-FOLD CROSS VALIDATION:**

In**K Fold cross validation,** the data is divided into k subsets. Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the test set/ validationset and the other k-1 subsets are put together to form a training set*.* The error estimation is averaged over all k trials to get total effectiveness of our model*.* As can be seen, every data point gets to be in a validation set exactly once, and gets to be in a training set k-1 times*.*This significantly reduces bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set**.** Interchanging the training and test sets also adds to the effectiveness of this method. **As a general rule and empirical evidence.**

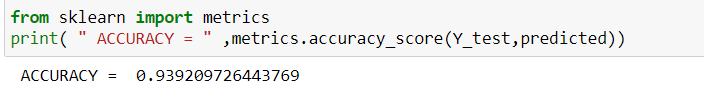


**TRAIN -TEST SPLIT AND CONFUSION MATRIX:**

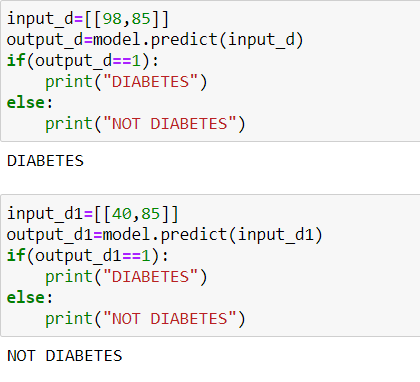


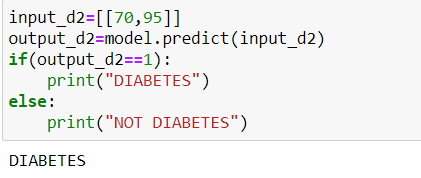
From the above confusion matrix we can say that the model has predicted 318 values correctly and incorrectly predicted only 11 values while testing

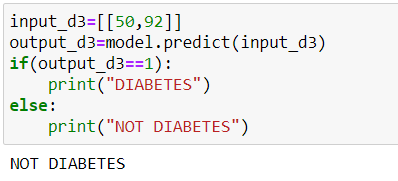
**CLASSIFICATION METRICS**



**PREDICTIONS:**







**ANALYSIS:**

* The model used is Naïve Bayes Classifier.
* The model accuracy via cross validation is 93.36%
* The model accuracy via accuracy score is 93.92%