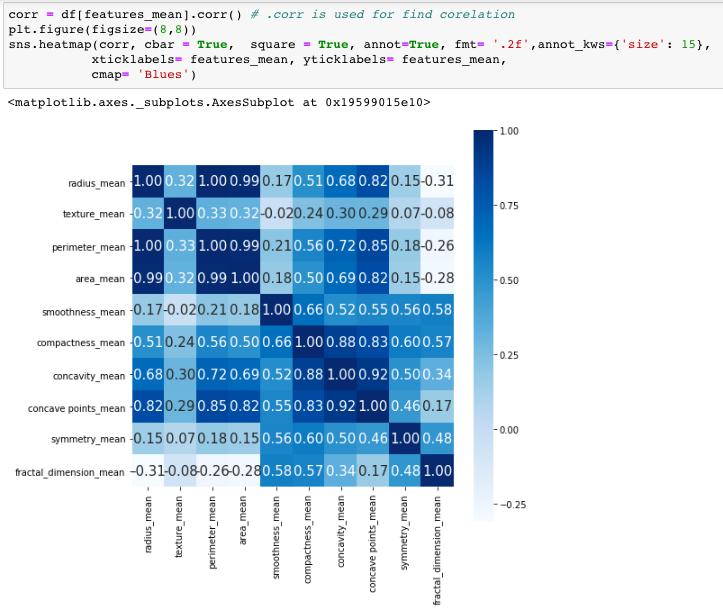
Chapter 8 RESULTS

* In the below figure 8.3 it is showing the dependencies of the different column attributes (chosen as the feature column) on each other.

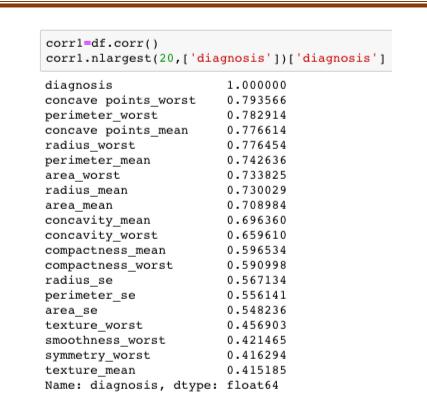


**Figure 8.3 Correlation Square showing relation between each columns**

* Finding the correlation between useful columns and diagnosis after dropping the less useful columns as shown in figure 8.4.

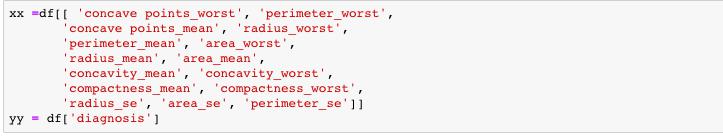
|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 39 |

Chapter 8 RESULTS



**Figure 8.4 Correlation between useful columns and diagnosis**

* While training the data, X is chosen based on the correlation with respect to diagnosis column (>0.5) and the Y is the Diagnosis column itself as shown in figure 8.5 below.



**Figure 8.5 Choosing X and Y for Training the model**

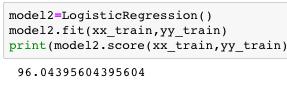
* Scikit-Learn has a function *train-test-split()* which can be used to randomly split the data into training and testing set as shown in figure 8.6 below.



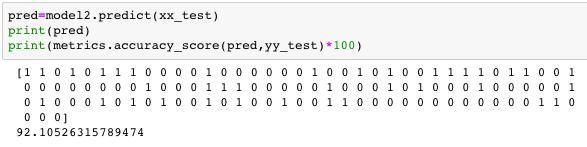
**Figure 8.6 Splitting data into training and testing set in ratio 4:1**

|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 40 |

Chapter 8 RESULTS

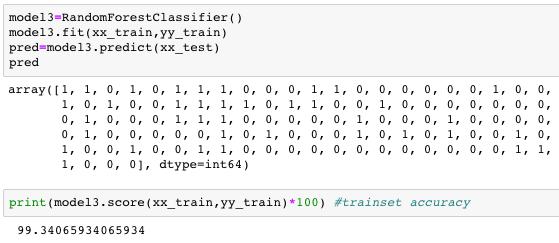


**Figure 8.7 Training Accuracy using Logistic Regression**



**Figure 8.8 Testing Accuracy using Logistic Regression**

* In order to improve the accuracy of the model considering the same feature columns another learning method i.e. Random Forest Classifier as seen in figure 8.9 and 8.10.



**Figure 8.9 Training Accuracy using Random Forest**



**Figure 8.10 Testing Accuracy using Random Forest**

Hence the training as well as the testing accuracy obtained in the Random Forest algorithm is relatively higher than the Logistic regression algorithm considering the same feature columns for training in either case.

|  |  |  |
| --- | --- | --- |
| Department of CSE, RNSIT | 2018-2019 | 41 |