**Assignment 4**

**Random Forest**

**Submitted By: Neel Vijay Patil (11436004)**

**Submitted To: Dr. Junhua Ding**

**INFO 5505 Applied Machine Learning for Data Science**

**UNT**

**Dataset:** I have collected the data-breast Cancer dataset from Kaggle.

The main purpose of assignment is to apply RF-Model on data that consist of various descriptive features that were figured from pictures of (FNA) of breast mass. They portray qualities of the cell cores in picture.

Dataset comprises of 29 attributes processed for every cell core having unique Id no. and diagnosis represented in form of binary where (Benign by 0’s and Malignant by 1’s)

**Step 1: Importing the required libraries**

Graphical user interface, text, application

Description automatically generated

**Step 2: Importing the data-breast Cancer data and generating the data frames**

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated

Table

Description automatically generated

**Step 3: Checking for datatypes, non-null counts, data types**

There are total 33 columns and 569 entries and data types include int 64, float 64.

Table

Description automatically generated

**Step 4: Checking the statistics of the dataset using the describe function.**

Graphical user interface, application

Description automatically generated

**Checking the columns in the dataset**

Text, letter

Description automatically generated

**Step 5: Replacing the columns names void spaces with ‘\_’ and deleting the Unnamed:\_32 column which is useless as well as setting index to id.**

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, application

Description automatically generated

**Step 6: Checking the Null values using isnull().sum() and were found to be 0 for every feature**

A picture containing text

Description automatically generated

**Step 7: Mapping the diagnosis column to the binary variables (Benign by 0’s and Malignant by 1’s)**

A picture containing graphical user interface

Description automatically generated

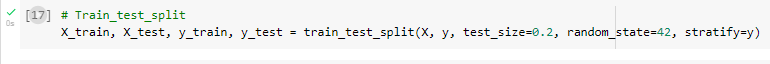
**Step 8: Importing the libraries and plotting the count plot for the benign and malignant**

Chart, bar chart

Description automatically generated

The count of Malignant was found to be 212 while Benign was found to be 357.

**Step 9: The Dataset is splitted into train and test in ratio of 80% to 20%**



**Step 10:** Designing the Random Forest model and representing the no of decision trees by using the n\_estimators. Here we have used the value of n\_estimators as 10 means number of trees are 10. Predicting results by model fitting and evaluating Y\_Predict.

Graphical user interface, text, application, chat or text message

Description automatically generated

**Step 11: Calculating the Accuracy Score of the Random Forest Model**

Graphical user interface, text

Description automatically generated

**The Accuracy was estimated as 95.61%**

**Step 12:** **Presently, I have created 6 folds by utilizing KFOLD technique, what partitions dataset into 6 sections, out of which the 20 percentage of the 6 parts will be used for testing and remaining 80% will be used for Training the dataset. Here we have calculated the cross-validation score for every Kfold split and also calculated the Accuracy score by evaluating the mean of the AUC and ROC Score**

Text

Description automatically generated with medium confidence

**The Accuracy was found to be 96.03%**

**Step 13: Generating the confusion matrix and Accuracy score for every Classification here we have set the value of n\_splits as 6.**

Text

Description automatically generated

Text

Description automatically generated

Here I have generated the confusion matrix for every classification and accuracy score for every classification

**1)**

Confusion Matrix for 1 classification

Accuracy is 0.9052631578947369

Graphical user interface

Description automatically generated with medium confidence

[[31 1]

[ 8 55]]

**2)**

Confusion Matrix for 2 classification

Accuracy is 0.9052631578947369

Graphical user interface, application

Description automatically generated

[[57 4]

[ 5 29]]

**3)**

Confusion Matrix for 3 classification

Accuracy is 0.9894736842105263

Graphical user interface, application

Description automatically generated

[[47 0]

[ 1 47]]

**4)**

Confusion Matrix for 4 classification

Accuracy is 0.9789473684210527

Graphical user interface, application

Description automatically generated

[[70 1]

[ 1 23]]

**5)**

Confusion Matrix for 5 classification

Accuracy is 0.9578947368421052

Graphical user interface, application

Description automatically generated

[[73 2]

[ 2 18]]

**6)**

Confusion Matrix for 6 classification

Accuracy is 0.9680851063829787

A screenshot of a cell phone

Description automatically generated with low confidence

[[68 3]

[ 0 23]]

**Step 14: Generating the Confusion matrix and classification report and accuracy score**

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

**Step 15: Probability predictions of every kept Tested data instance occurring in the one or the other classes.**

Text

Description automatically generated

Chart

Description automatically generated with medium confidence

**Step 16: Plotting the ROC curve I used the AUC values and plotted a line graph between the false positive and true positive value.**

Graphical user interface, text

Description automatically generated

A picture containing graphical user interface

Description automatically generated

**Conclusion:**

**The Accuracy score for the entire Random Forest model was estimated to be 96.80%. While the AUC was found to be 99.84 percentage Which indicates that our model has a good accuracy score to predict whether disease is Benign or Malignant in view of the data given.**