Trip Advisor Scores

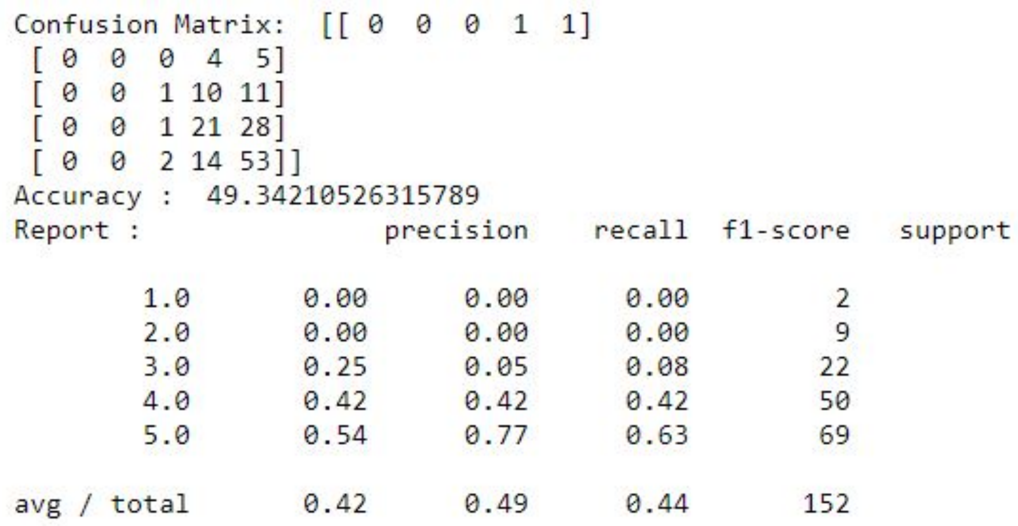
Preprocessing of dataset

1. Replaced ***Period of stay*** column with labels of seasons.
2. Applied ***Label encoder*** to the multi label features mostly those are non numerical to transform into numerical form as our algorithm takes numeric arrays as input.
3. For categorical variables where no such ordinal relationship exists, the integer encoding is not enough. In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results In this case, a **One-hot encoding** can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value. This improved the performance of the model.

ALGORITHMS

1. Random Forest Classifier

scores that are predicted are the leaf nodes of the various multi label features and the probability of each score in each scenario, random forest gave the better accuracy than other machine learning models.



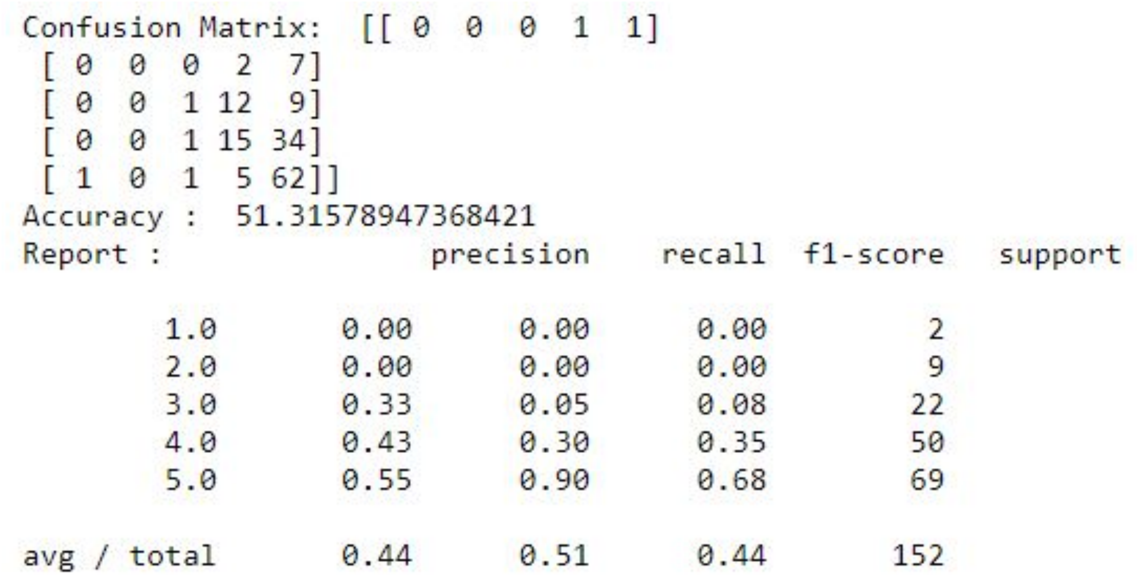
```
Confusion Matrix:  [[ 0  0  0  1  1]
 [ 0  0  0  4  5]
 [ 0  0  1 10 11]
 [ 0  0  1 21 28]
 [ 0  0  2 14 53]]
Accuracy :  49.34210526315789
Report :              precision    recall  f1-score   support

        1.0       0.00      0.00      0.00         2
        2.0       0.00      0.00      0.00         9
        3.0       0.25      0.05      0.08        22
        4.0       0.42      0.42      0.42        50
        5.0       0.54      0.77      0.63        69

avg / total       0.42      0.49      0.44       152
```

2. Gradient Boosting

We used fully grown decision trees in Random Forests which may also lead to overfitting, so to reduce the loss and to reduce bias and variance we apply gradient boosting and it improved the accuracy by 2 %.



```
Confusion Matrix:  [[ 0  0  0  1  1]
 [ 0  0  0  2  7]
 [ 0  0  1 12  9]
 [ 0  0  1 15 34]
 [ 1  0  1  5 62]]
Accuracy :  51.31578947368421
Report :              precision    recall  f1-score   support

        1.0       0.00      0.00      0.00         2
        2.0       0.00      0.00      0.00         9
        3.0       0.33      0.05      0.08        22
        4.0       0.43      0.30      0.35        50
        5.0       0.55      0.90      0.68        69

avg / total       0.44      0.51      0.44       152
```

Breast Cancer Classification

Preprocessing Steps

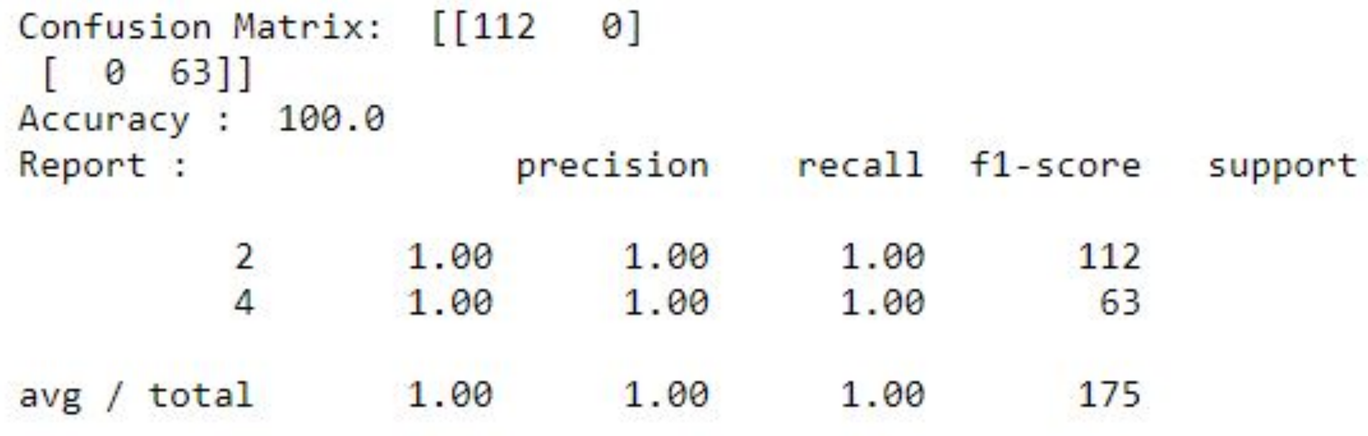
1. Handling missing values : Dataset contained 16 missing values in Bare Nuclei column, they are replaced with the mean values.
2. Standard Scaler : To make the data normally distributed with zero mean and unit variance.

Algorithm

Logistic Regression

Since this is the case of binary classification, logistic regression is one of the most powerful algorithm for binary classification, logistic regression not only gives a measure of how *relevant* a predictor is (coefficient size) but also its *direction* of association (positive or negative)

Results



```
Confusion Matrix:  [[112   0]
 [  0  63]]
Accuracy :  100.0
Report :               precision    recall  f1-score   support

          2       1.00      1.00      1.00       112
          4       1.00      1.00      1.00        63

avg / total       1.00      1.00      1.00       175
```