UNIVERSITI SAINS MALAYSIA

School of Computer Sciences

Universiti Sains Malaysia

CDS503

Machine Learning

1) Study the data set carefully and answer the questions below:

a. Report the class distribution. Is this a balanced or unbalanced data set?

NO	YES
155	136

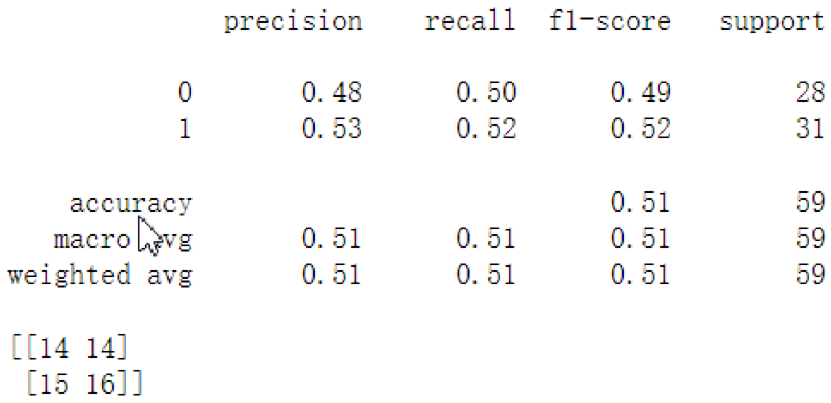
In Depression, 136 were yes and 155 were No

This is a balanced data set

b. Please select and justify a suitable metric to evaluate the performance of your classification model.

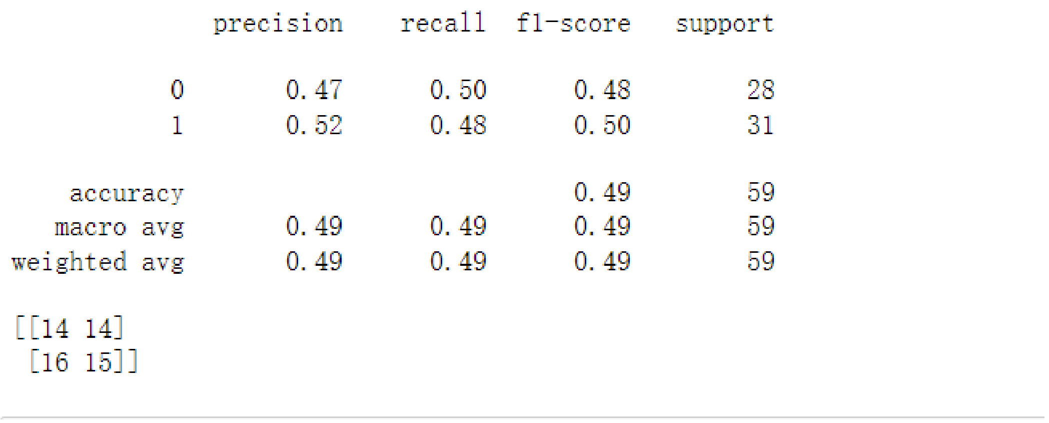
The method to evaluate the performance of classification models is confusion matrix, the general idea of which is to count the number of times category A instances are predicted (classified) to be category B.

Freq：



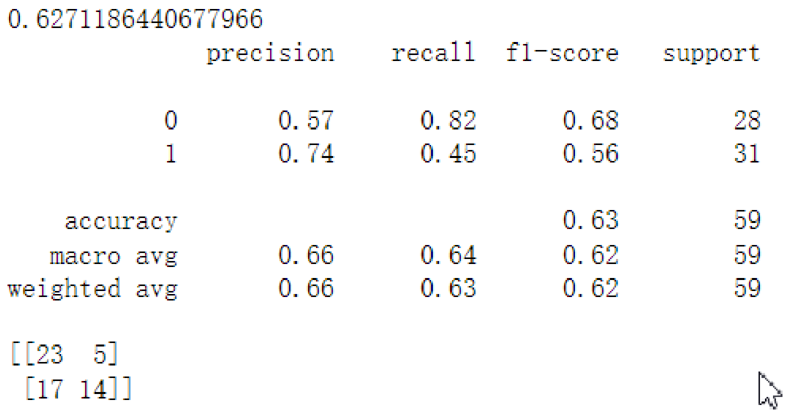
```
              precision    recall  f1-score   support

           0       0.48      0.50      0.49        28
           1       0.53      0.52      0.52        31

    accuracy                           0.51        59
   macro avg       0.51      0.51      0.51        59
weighted avg       0.51      0.51      0.51        59

[[14 14]
 [15 16]]
```

Norm：



```
              precision    recall  f1-score   support

           0       0.47      0.50      0.48        28
           1       0.52      0.48      0.50        31

    accuracy                           0.49        59
   macro avg       0.49      0.49      0.49        59
weighted avg       0.49      0.49      0.49        59

[[14 14]
 [16 15]]
```

c. Given the size of the data set, which validation option (e.g., percentage split, k-fold cross validation) do you think is suitable to be used in your machine learning experiments. Specify the validation option you are selecting for your machine learning experiments. Briefly explain the reason for using the validation option.

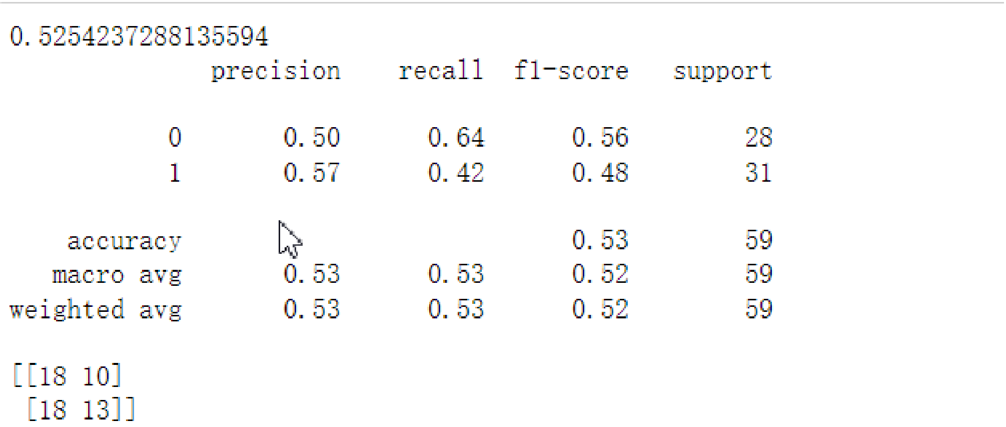


```
0.6271186440677966
              precision    recall  f1-score   support

           0       0.57      0.82      0.68        28
           1       0.74      0.45      0.56        31

    accuracy                           0.63        59
   macro avg       0.66      0.64      0.62        59
weighted avg       0.66      0.63      0.62        59

[[23  5]
 [17 14]]
```

（2）Which feature representation produces a better model? Explain how you determine the best performing model based on the performance metric you have selected. Can you explain why one feature representation is better than the other?

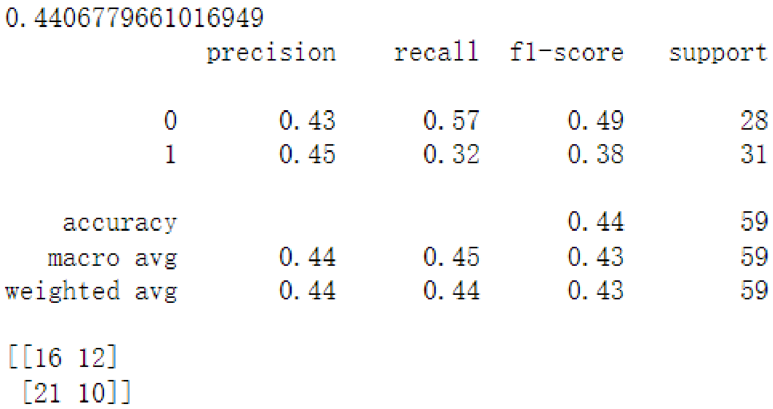
Classification problem evaluation indicators:Accuracy -- Accuracy Accuracy (error) - Precision， Recall rate - Recall， F1 score， Recall, Precision and F1-score were used to evaluate the quality of a model. Accuracy is The percentage of the total sample that predicted correct results,Precision:Of all the samples that were predicted to be positive that were actually positive,Recall :The probability of being predicted to be a positive sample in a sample that is actually positive,F1 is The balance between Precision and recall.When choosing between two different groups of accuracy and recall, consider the F-1 Score combining both.so,SVM is the best model.

decision tree norm：



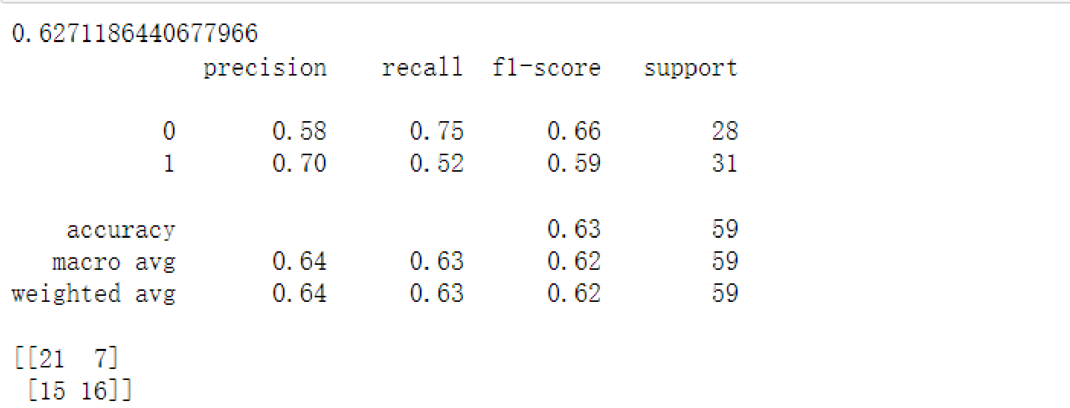
```
0.5254237288135594
              precision    recall  f1-score   support

           0       0.50      0.64      0.56        28
           1       0.57      0.42      0.48        31

    accuracy                           0.53        59
   macro avg       0.53      0.53      0.52        59
weighted avg       0.53      0.53      0.52        59

[[18 10]
 [18 13]]
```

decision tree freq：



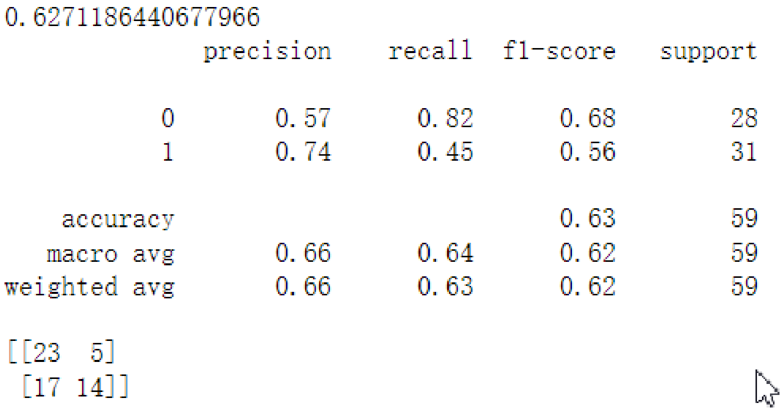
```
0.4406779661016949
              precision    recall  f1-score   support

           0       0.43      0.57      0.49        28
           1       0.45      0.32      0.38        31

    accuracy                           0.44        59
   macro avg       0.44      0.45      0.43        59
weighted avg       0.44      0.44      0.43        59

[[16 12]
 [21 10]]
```

Svm norm：



```
0.6271186440677966
              precision    recall  f1-score   support

           0       0.58      0.75      0.66        28
           1       0.70      0.52      0.59        31

    accuracy                           0.63        59
   macro avg       0.64      0.63      0.62        59
weighted avg       0.64      0.63      0.62        59

[[21  7]
 [15 16]]
```

SVM freq：



```
0.6271186440677966
              precision    recall  f1-score   support

           0       0.57      0.82      0.68        28
           1       0.74      0.45      0.56        31

    accuracy                           0.63        59
   macro avg       0.66      0.64      0.62        59
weighted avg       0.66      0.63      0.62        59

[[23  5]
 [17 14]]
```

4) Is your overall best performing model able to achieve very promising results (reach at least 0.8 of your selected performance metric)? Provide two suggestions on future strategies that can be used to improve the performance of your best performing model.

To achieve best performing model able to achieve very promising results.

Suggestions:

1.In terms of data, increase the number of data, collect more real data, generate more data, transform the data and select features.

2.Improve performance algorithmically, filter algorithms, improve performance from algorithm tuning, compare different algorithms.