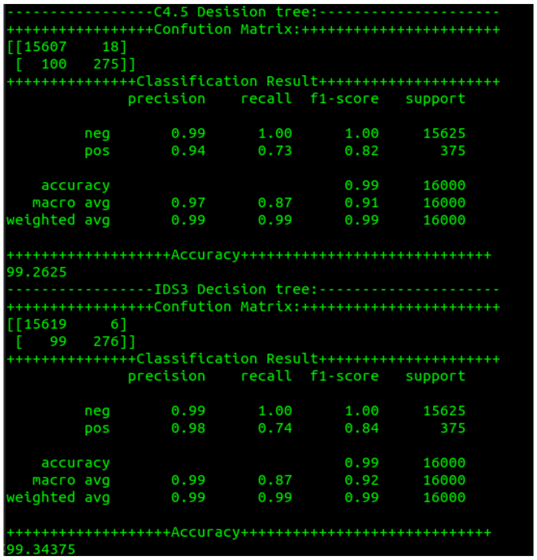
ID3 and C4.5 algorithm

Armin soltan 610396115

scania

In this program I use scikit learn library but something that is my issue is that this library optimize dicision tree algorithm for ID3 and C4.5 so I use two different criterion first one is entropy and second one is gini and i use missing value to improve accuracy and replace them with average of data. scikit learn has lot of ready packages so i don't implement each of them by myself so i use them like confusion and classification result

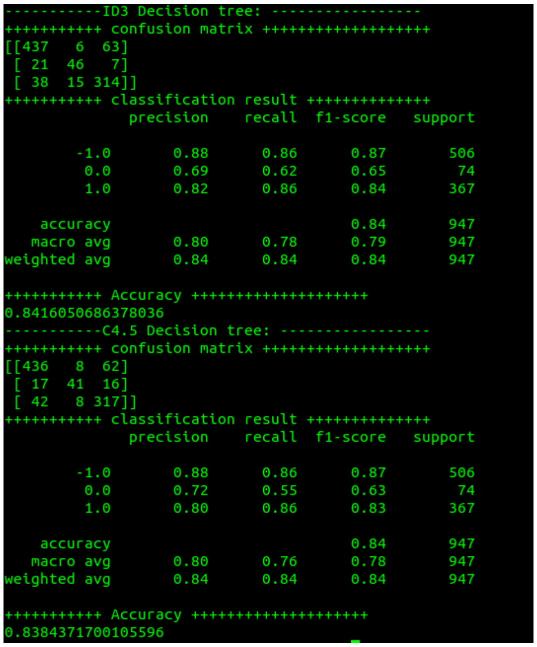


```
-----------------C4.5 Desision tree:---------------------
+++++++++++++++++Confution Matrix:+++++++++++++++++++++++
[[15607    18]
 [  100   275]]
+++++++++++++++Classification Result+++++++++++++++++++++
              precision    recall  f1-score   support

         neg       0.99      1.00      1.00     15625
         pos       0.94      0.73      0.82       375

    accuracy                           0.99     16000
   macro avg       0.97      0.87      0.91     16000
weighted avg       0.99      0.99      0.99     16000

+++++++++++++++++++Accuracy+++++++++++++++++++++++++++++
99.2625
-----------------IDS3 Decision tree:---------------------
+++++++++++++++++Confution Matrix:+++++++++++++++++++++++
[[15619     6]
 [   99   276]]
+++++++++++++++Classification Result+++++++++++++++++++++
              precision    recall  f1-score   support

         neg       0.99      1.00      1.00     15625
         pos       0.98      0.74      0.84       375

    accuracy                           0.99     16000
   macro avg       0.99      0.87      0.92     16000
weighted avg       0.99      0.99      0.99     16000

+++++++++++++++++++Accuracy+++++++++++++++++++++++++++++
99.34375
```

Figure 1: scania output

trucks

In this program since we don't have test and train data separatly i split phising data set in two parts test data and training data. 30% of data is testing data and rest of it belongs to training data. You can see scania.py and trucks.py in this folder. If you have any advice or my program has a problem I would be glad if you tell me my program's defect.



```
-----------ID3 Decision tree: -----------------
+++++++++++ confusion matrix +++++++++++++++++++
[[437   6  63]
 [ 21  46   7]
 [ 38  15 314]]
+++++++++++ classification result ++++++++++++++
              precision    recall  f1-score   support

        -1.0       0.88      0.86      0.87       506
         0.0       0.69      0.62      0.65        74
         1.0       0.82      0.86      0.84       367

    accuracy                           0.84       947
   macro avg       0.80      0.78      0.79       947
weighted avg       0.84      0.84      0.84       947

+++++++++++ Accuracy ++++++++++++++++++++
0.8416050686378036
-----------C4.5 Decision tree: -----------------
+++++++++++ confusion matrix +++++++++++++++++++
[[436   8  62]
 [ 17  41  16]
 [ 42   8 317]]
+++++++++++ classification result ++++++++++++++
              precision    recall  f1-score   support

        -1.0       0.88      0.86      0.87       506
         0.0       0.72      0.55      0.63        74
         1.0       0.80      0.86      0.83       367

    accuracy                           0.84       947
   macro avg       0.80      0.76      0.78       947
weighted avg       0.84      0.84      0.84       947

+++++++++++ Accuracy ++++++++++++++++++++
0.8384371700105596
```

Figure 2: trucks output