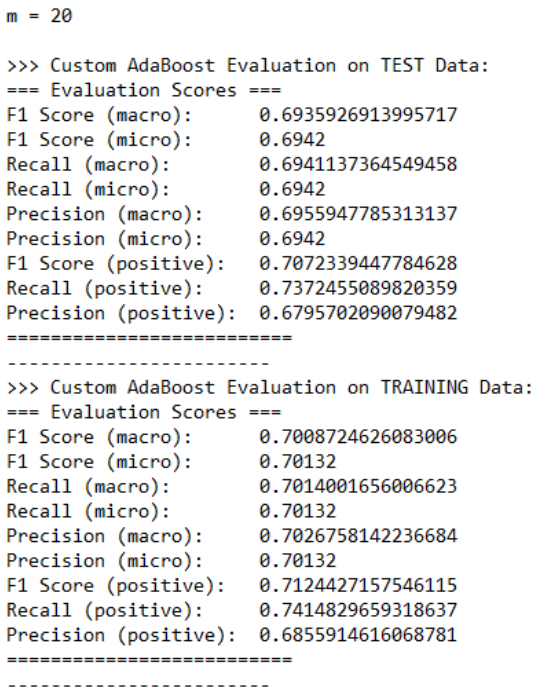
Εργασία 2

Το πρόγραμμα αρχικά κάνει μια προεπεξεργασία του csv αρχείου με τα δεδομενα, και παράγει δύο csv αρχεία κατάλληλα για χρήση από τους αλγορίθμους μηχανικής μάθησης. Τα παραγώμενα csv είναι τα test_data και train_data και εχουν την ακόλουθη μορφή: header τα ονόματα των 5000 λέξεων(features) του λεξιλογίου, 25000 γραμμές για τα δεδομένα εκπαίδευσης ή αξιολόγησης, και 5000 στήλες με τιμές 0 ή 1 ανάλογα με το αν η αντίστοιχη λέξη βρίσκεται στο review ή όχι. Για την παραγωγή αυτών των αρχείων, το αρχικό csv χωρίζεται σε 50% δεδομένα επικύρωσης και 50% δεδομένα εκπαίδευσης. Μετά, αφαιρούμε τις 50 πιο συχνές και 500 πιο σπάνιες λέξεις στο training_data, και από αυτές που απομένουν, κρατάμε στο λεξιλόγιο μας τις 5000 με το μεγαλύτερο information gain, που υπολογίζεται απο την συνάρτηση calculate_information_gain. Κρατάμε τις ίδιες λέξεις και στο test_data. Στη συνέχεια φέρνουμε τα αρχεία στην επιθυμιτή μορφή και τα εξάγουμε. Στο υπόλοιπο της εργασίας χρησιμοποιούνται αυτά τα δεδομένα, περασμένα σε pandas dataframes.

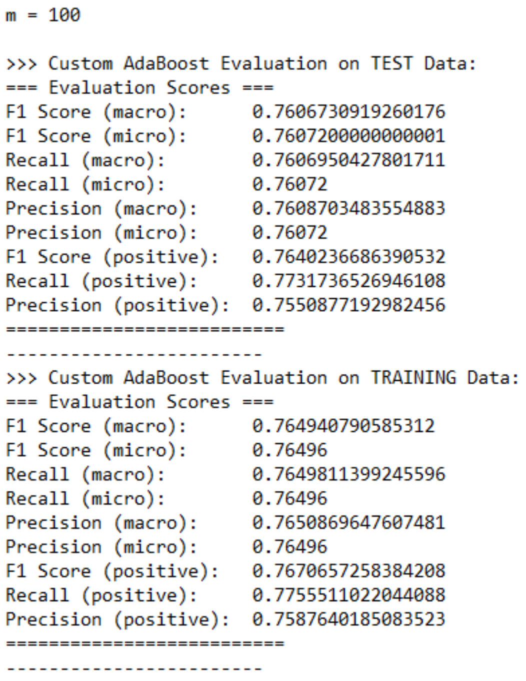
Στην εργασία κάνουμε δικές μας υλοποιήσεις για τους Naive Bayes, Random Forest και Adaboost, και συγκρίνουμε τις επιδόσεις τους με τις αντίστοιχες έτοιμες της sklearn, με κατά το δυνατόν ίδιες υπερπαραμέτρους. Στις περιπτώσεις που οι υπερπαράμετροι δεν είναι ιδιες, αυτό συμβαίνει γιατί σε κάποια υλοποίηση η εκτέλεση του προγράμματος έπαιρνε πάρα πολλή ώρα.

Αρχικά για τον adaboost, κάναμε ελέγχους με διάφορες τιμές της υπερπαραμέτρου m (αριθμός stumps), και καταλήξαμε στην τιμή

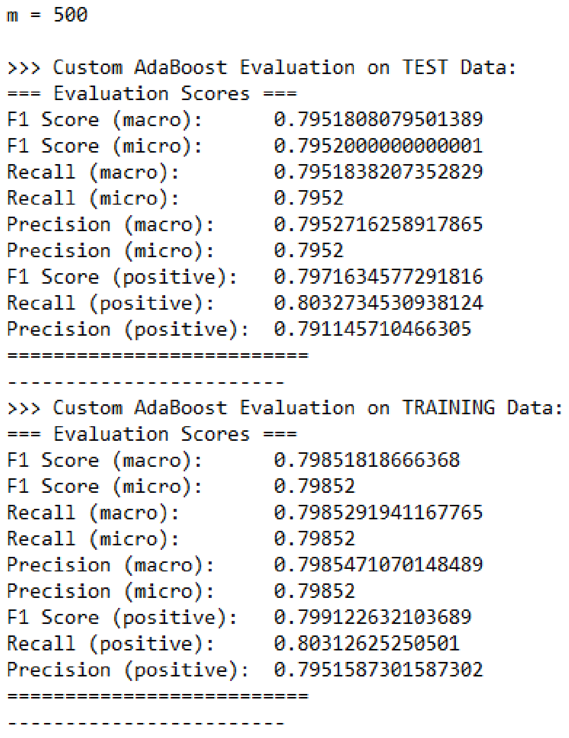
2000. Ακολουθούν τα evaluation scores για 25000 training samples, με διαφορετικές τιμές για το m:



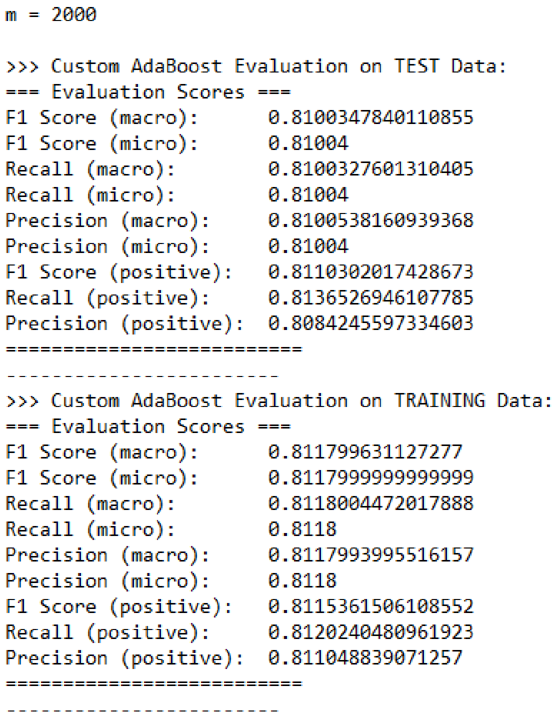
```
m = 20

>>> Custom AdaBoost Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.6935926913995717
F1 Score (micro):       0.6942
Recall (macro):         0.6941137364549458
Recall (micro):         0.6942
Precision (macro):      0.6955947785313137
Precision (micro):      0.6942
F1 Score (positive):    0.7072339447784628
Recall (positive):      0.7372455089820359
Precision (positive):   0.6795702090079482
==========================
------------------------
>>> Custom AdaBoost Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.7008724626083006
F1 Score (micro):       0.70132
Recall (macro):         0.7014001656006623
Recall (micro):         0.70132
Precision (macro):      0.7026758142236684
Precision (micro):      0.70132
F1 Score (positive):    0.7124427157546115
Recall (positive):      0.7414829659318637
Precision (positive):   0.6855914616068781
==========================
------------------------
```



```
m = 100

>>> Custom AdaBoost Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.7606730919260176
F1 Score (micro):       0.7607200000000001
Recall (macro):         0.7606950427801711
Recall (micro):         0.76072
Precision (macro):      0.7608703483554883
Precision (micro):      0.76072
F1 Score (positive):    0.7640236686390532
Recall (positive):      0.7731736526946108
Precision (positive):   0.7550877192982456
==========================
------------------------
>>> Custom AdaBoost Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.764940790585312
F1 Score (micro):       0.76496
Recall (macro):         0.7649811399245596
Recall (micro):         0.76496
Precision (macro):      0.7650869647607481
Precision (micro):      0.76496
F1 Score (positive):    0.7670657258384208
Recall (positive):      0.7755511022044088
Precision (positive):   0.7587640185083523
==========================
------------------------
```

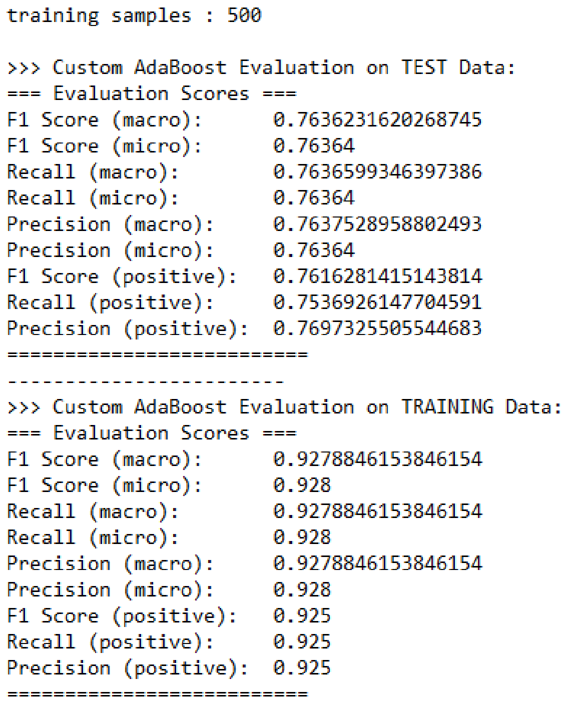


```
m = 500

>>> Custom AdaBoost Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.7951808079501389
F1 Score (micro):       0.7952000000000001
Recall (macro):         0.7951838207352829
Recall (micro):         0.7952
Precision (macro):      0.7952716258917865
Precision (micro):      0.7952
F1 Score (positive):    0.7971634577291816
Recall (positive):      0.8032734530938124
Precision (positive):   0.791145710466305
==========================
------------------------
>>> Custom AdaBoost Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.79851818666368
F1 Score (micro):       0.79852
Recall (macro):         0.7985291941167765
Recall (micro):         0.79852
Precision (macro):      0.7985471070148489
Precision (micro):      0.79852
F1 Score (positive):    0.799122632103689
Recall (positive):      0.80312625250501
Precision (positive):   0.7951587301587302
==========================
------------------------
```

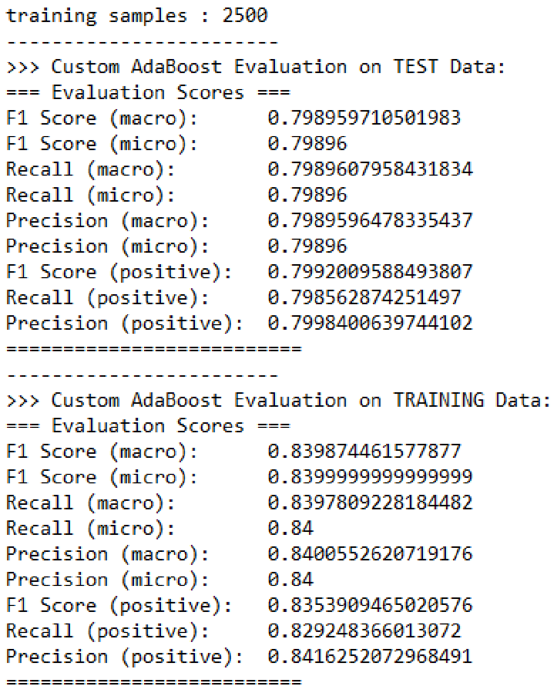


```
m = 2000

>>> Custom AdaBoost Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8100347840110855
F1 Score (micro):       0.81004
Recall (macro):         0.8100327601310405
Recall (micro):         0.81004
Precision (macro):      0.8100538160939368
Precision (micro):      0.81004
F1 Score (positive):    0.8110302017428673
Recall (positive):      0.8136526946107785
Precision (positive):   0.8084245597334603
==========================
------------------------
>>> Custom AdaBoost Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.811799631127277
F1 Score (micro):       0.8117999999999999
Recall (macro):         0.8118004472017888
Recall (micro):         0.8118
Precision (macro):      0.8117993995516157
Precision (micro):      0.8118
F1 Score (positive):    0.8115361506108552
Recall (positive):      0.8120240480961923
Precision (positive):   0.811048839071257
==========================
------------------------
```

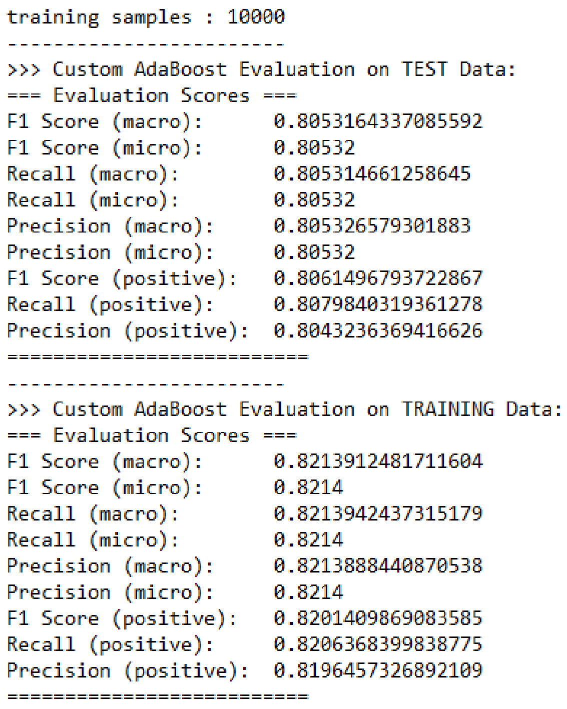
Ακολουθούν τα evaluation scores για m=2000 και διαφορετικά training sizes:



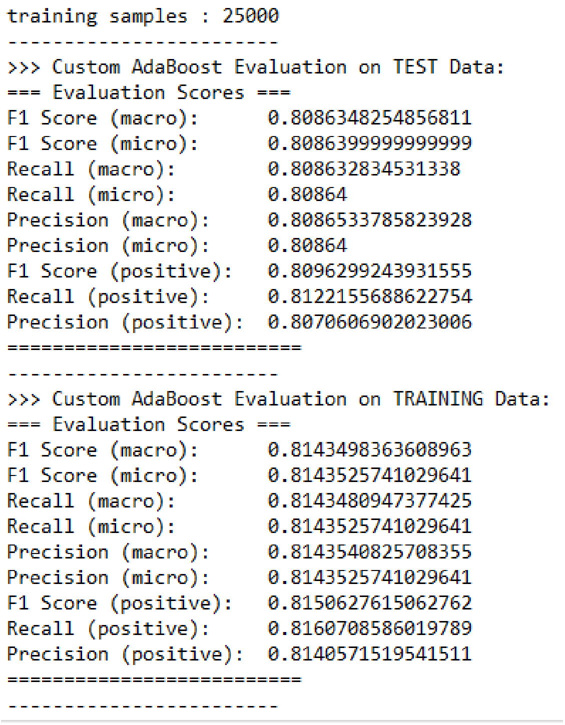
```
training samples : 500

>>> Custom AdaBoost Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.7636231620268745
F1 Score (micro):       0.76364
Recall (macro):         0.7636599346397386
Recall (micro):         0.76364
Precision (macro):      0.7637528958802493
Precision (micro):      0.76364
F1 Score (positive):    0.7616281415143814
Recall (positive):      0.7536926147704591
Precision (positive):   0.7697325505544683
==========================
------------------------
>>> Custom AdaBoost Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.9278846153846154
F1 Score (micro):       0.928
Recall (macro):         0.9278846153846154
Recall (micro):         0.928
Precision (macro):      0.9278846153846154
Precision (micro):      0.928
F1 Score (positive):    0.925
Recall (positive):      0.925
Precision (positive):   0.925
==========================
```



```
training samples : 2500
------------------------
>>> Custom AdaBoost Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.798959710501983
F1 Score (micro):       0.79896
Recall (macro):         0.7989607958431834
Recall (micro):         0.79896
Precision (macro):      0.7989596478335437
Precision (micro):      0.79896
F1 Score (positive):    0.7992009588493807
Recall (positive):      0.798562874251497
Precision (positive):   0.7998400639744102
==========================
------------------------
>>> Custom AdaBoost Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.839874461577877
F1 Score (micro):       0.8399999999999999
Recall (macro):         0.8397809228184482
Recall (micro):         0.84
Precision (macro):      0.8400552620719176
Precision (micro):      0.84
F1 Score (positive):    0.8353909465020576
Recall (positive):      0.829248366013072
Precision (positive):   0.8416252072968491
==========================
```

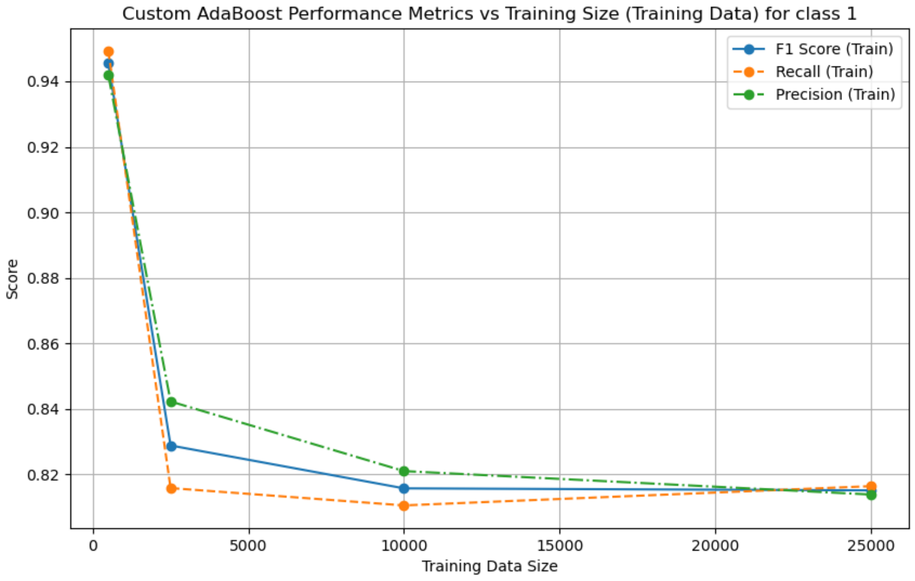


```
training samples : 10000
------------------------
>>> Custom AdaBoost Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8053164337085592
F1 Score (micro):       0.80532
Recall (macro):         0.805314661258645
Recall (micro):         0.80532
Precision (macro):      0.805326579301883
Precision (micro):      0.80532
F1 Score (positive):    0.8061496793722867
Recall (positive):      0.8079840319361278
Precision (positive):   0.8043236369416626
==========================
------------------------
>>> Custom AdaBoost Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8213912481711604
F1 Score (micro):       0.8214
Recall (macro):         0.8213942437315179
Recall (micro):         0.8214
Precision (macro):      0.8213888440870538
Precision (micro):      0.8214
F1 Score (positive):    0.8201409869083585
Recall (positive):      0.8206368399838775
Precision (positive):   0.8196457326892109
==========================
```

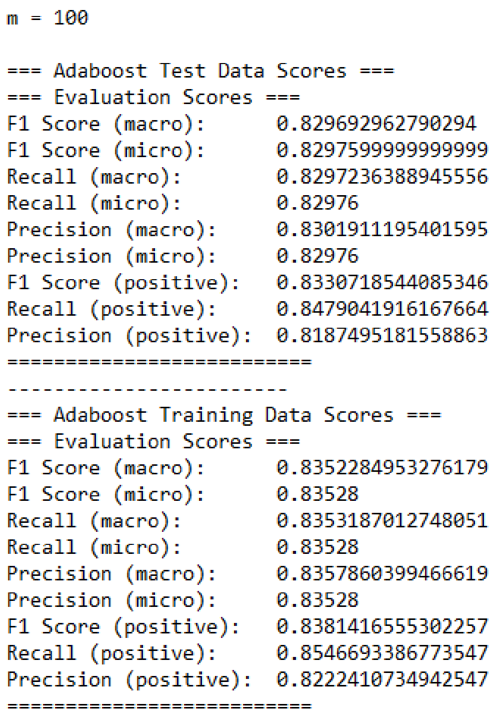


```
training samples : 25000
------------------------
>>> Custom AdaBoost Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8086348254856811
F1 Score (micro):       0.8086399999999999
Recall (macro):         0.808632834531338
Recall (micro):         0.80864
Precision (macro):      0.8086533785823928
Precision (micro):      0.80864
F1 Score (positive):    0.8096299243931555
Recall (positive):      0.8122155688622754
Precision (positive):   0.8070606902023006
==========================
------------------------
>>> Custom AdaBoost Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8143498363608963
F1 Score (micro):       0.814352574102964
Recall (macro):         0.8143480947377425
Recall (micro):         0.814352574102964
Precision (macro):      0.8143540825708355
Precision (micro):      0.814352574102964
F1 Score (positive):    0.8150627615062762
Recall (positive):      0.8160708586019789
Precision (positive):   0.8140571519541511
==========================
------------------------
```

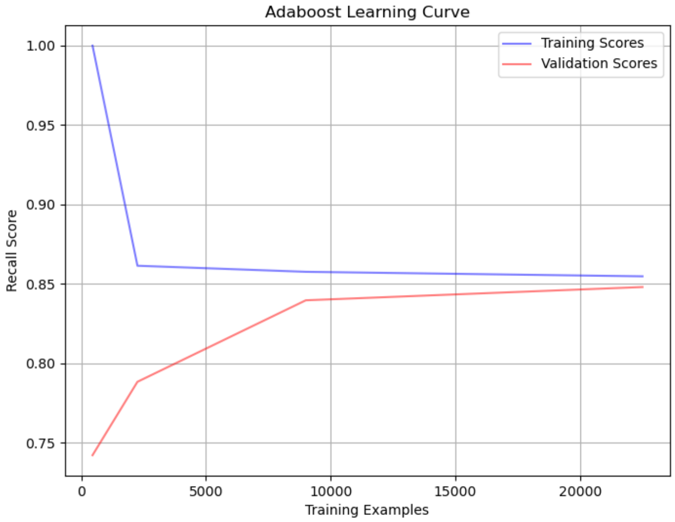
Και το plot:



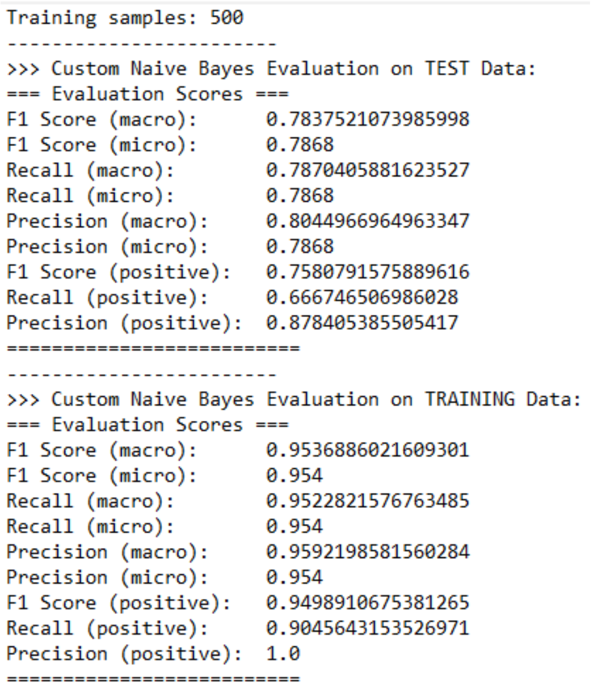
Για τον adaboost της sklearn:



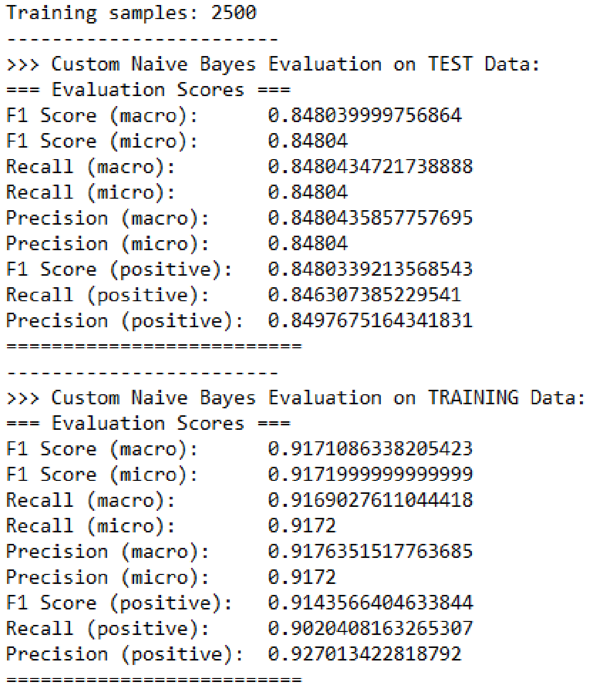
```
m = 100

=== Adaboost Test Data Scores ===
=== Evaluation Scores ===
F1 Score (macro):      0.829692962790294
F1 Score (micro):      0.8297599999999999
Recall (macro):        0.8297236388945556
Recall (micro):        0.82976
Precision (macro):     0.8301911195401595
Precision (micro):     0.82976
F1 Score (positive):   0.8330718544085346
Recall (positive):     0.8479041916167664
Precision (positive):  0.8187495181558863
==========================
------------------------
=== Adaboost Training Data Scores ===
=== Evaluation Scores ===
F1 Score (macro):      0.8352284953276179
F1 Score (micro):      0.83528
Recall (macro):        0.8353187012748051
Recall (micro):        0.83528
Precision (macro):     0.8357860399466619
Precision (micro):     0.83528
F1 Score (positive):   0.8381416555302257
Recall (positive):     0.854669338677347
Precision (positive):  0.8222410734942547
==========================
```



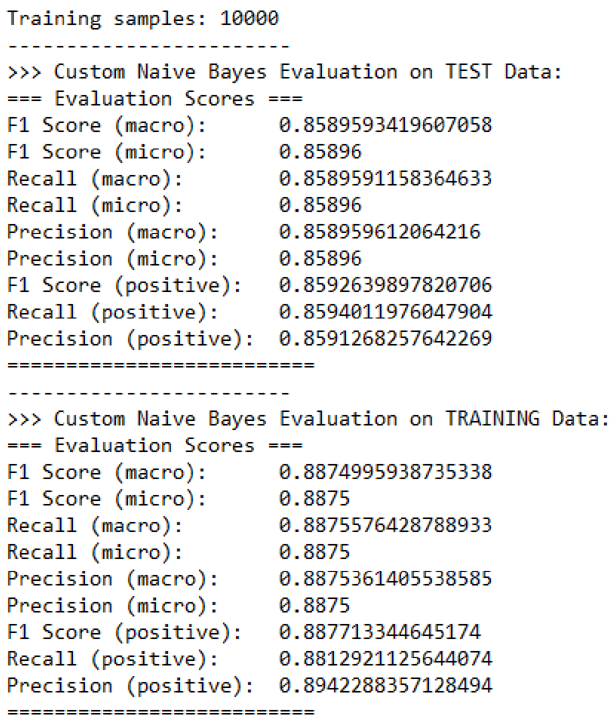
Για τον Naive Bayes τα evaluation scores είναι τα παρακάτω:



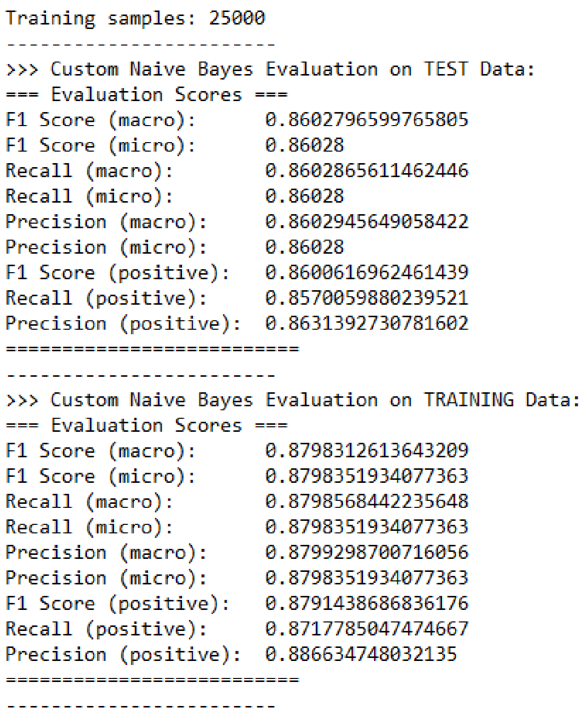
```
Training samples: 500
------------------------
>>> Custom Naive Bayes Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):      0.7837521073985998
F1 Score (micro):      0.7868
Recall (macro):        0.7870405881623527
Recall (micro):        0.7868
Precision (macro):     0.8044966964963347
Precision (micro):     0.7868
F1 Score (positive):   0.7580791575889616
Recall (positive):     0.666746506986028
Precision (positive):  0.878405385505417
==========================
------------------------
>>> Custom Naive Bayes Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):      0.9536886021609301
F1 Score (micro):      0.954
Recall (macro):        0.9522821576763485
Recall (micro):        0.954
Precision (macro):     0.9592198581560284
Precision (micro):     0.954
F1 Score (positive):   0.9498910675381265
Recall (positive):     0.9045643153526971
Precision (positive):  1.0
==========================
```



```
Training samples: 2500
------------------------
>>> Custom Naive Bayes Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):      0.848039999756864
F1 Score (micro):      0.84804
Recall (macro):        0.8480434721738888
Recall (micro):        0.84804
Precision (macro):     0.8480435857757695
Precision (micro):     0.84804
F1 Score (positive):   0.8480339213568543
Recall (positive):     0.846307385229541
Precision (positive):  0.8497675164341831
==========================
------------------------
>>> Custom Naive Bayes Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):      0.9171086338205423
F1 Score (micro):      0.9171999999999999
Recall (macro):        0.9169027611044418
Recall (micro):        0.9172
Precision (macro):     0.9176351517763685
Precision (micro):     0.9172
F1 Score (positive):   0.9143566404633844
Recall (positive):     0.9020408163265307
Precision (positive):  0.927013422818792
==========================
```

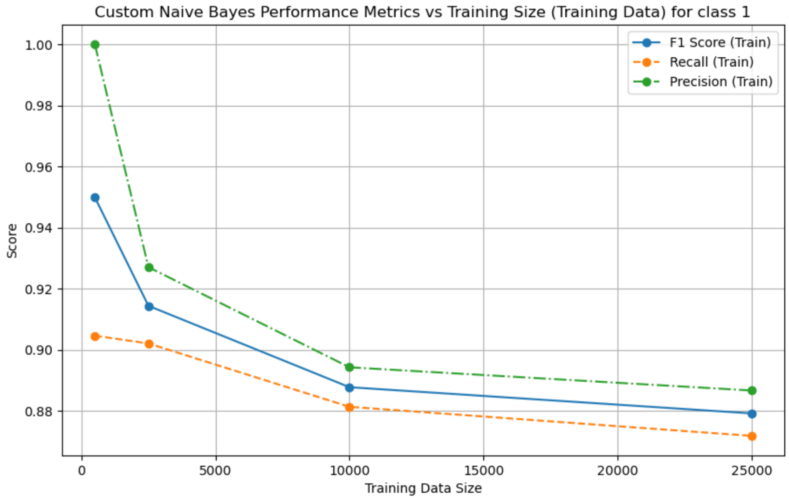


```
Training samples: 10000
------------------------
>>> Custom Naive Bayes Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8589593419607058
F1 Score (micro):       0.85896
Recall (macro):         0.8589591158364633
Recall (micro):         0.85896
Precision (macro):      0.858959612064216
Precision (micro):      0.85896
F1 Score (positive):    0.8592639897820706
Recall (positive):      0.8594011976047904
Precision (positive):   0.8591268257642269
==========================
------------------------
>>> Custom Naive Bayes Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8874995938735338
F1 Score (micro):       0.8875
Recall (macro):         0.8875576428788933
Recall (micro):         0.8875
Precision (macro):      0.8875361405538585
Precision (micro):      0.8875
F1 Score (positive):    0.887713344645174
Recall (positive):      0.8812921125644074
Precision (positive):   0.8942288357128494
==========================
```

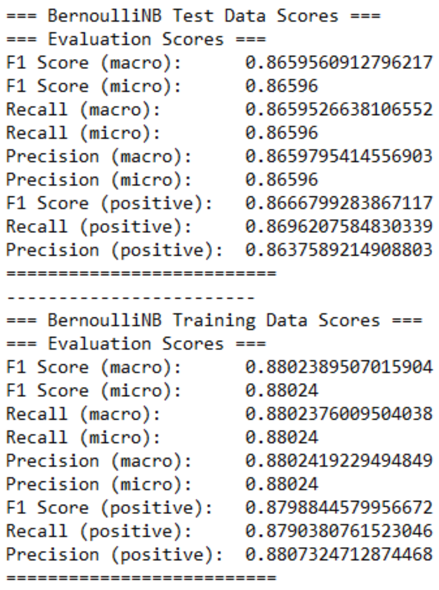


```
Training samples: 25000
------------------------
>>> Custom Naive Bayes Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8602796599765805
F1 Score (micro):       0.86028
Recall (macro):         0.8602865611462446
Recall (micro):         0.86028
Precision (macro):      0.8602945649058422
Precision (micro):      0.86028
F1 Score (positive):    0.8600616962461439
Recall (positive):      0.8570059880239521
Precision (positive):   0.8631392730781602
==========================
------------------------
>>> Custom Naive Bayes Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.8798312613643209
F1 Score (micro):       0.8798351934077363
Recall (macro):         0.8798568442235648
Recall (micro):         0.8798351934077363
Precision (macro):      0.8799298700716056
Precision (micro):      0.8798351934077363
F1 Score (positive):    0.8791438686836176
Recall (positive):      0.8717785047474667
Precision (positive):   0.886634748032135
==========================
------------------------
```

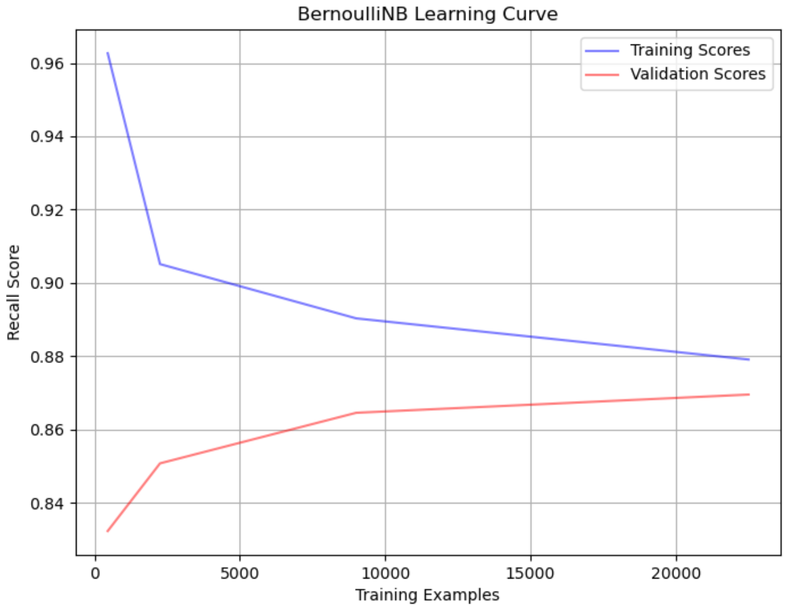
Και το plot:



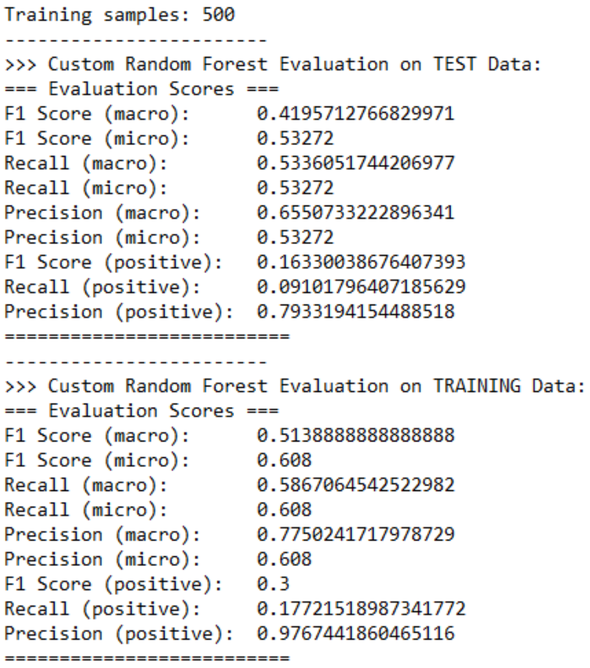
Για τον Bernoulli Naive Bayes της sklearn:



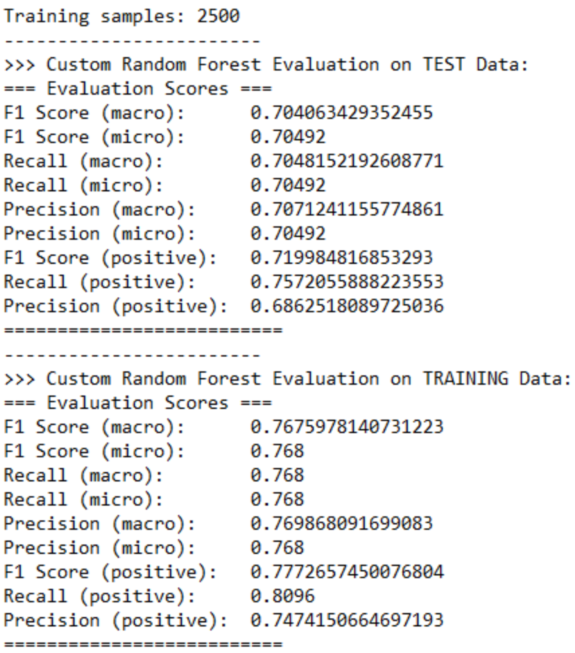
```
=== BernoulliNB Test Data Scores ===
=== Evaluation Scores ===
F1 Score (macro):      0.8659560912796217
F1 Score (micro):      0.86596
Recall (macro):        0.8659526638106552
Recall (micro):        0.86596
Precision (macro):     0.8659795414556903
Precision (micro):     0.86596
F1 Score (positive):   0.8666799283867117
Recall (positive):     0.8696207584830339
Precision (positive):  0.8637589214908803
==========================
------------------------
=== BernoulliNB Training Data Scores ===
=== Evaluation Scores ===
F1 Score (macro):      0.8802389507015904
F1 Score (micro):      0.88024
Recall (macro):        0.8802376009504038
Recall (micro):        0.88024
Precision (macro):     0.8802419229494849
Precision (micro):     0.88024
F1 Score (positive):   0.8798844579956672
Recall (positive):     0.8790380761523046
Precision (positive):  0.8807324712874468
==========================
```



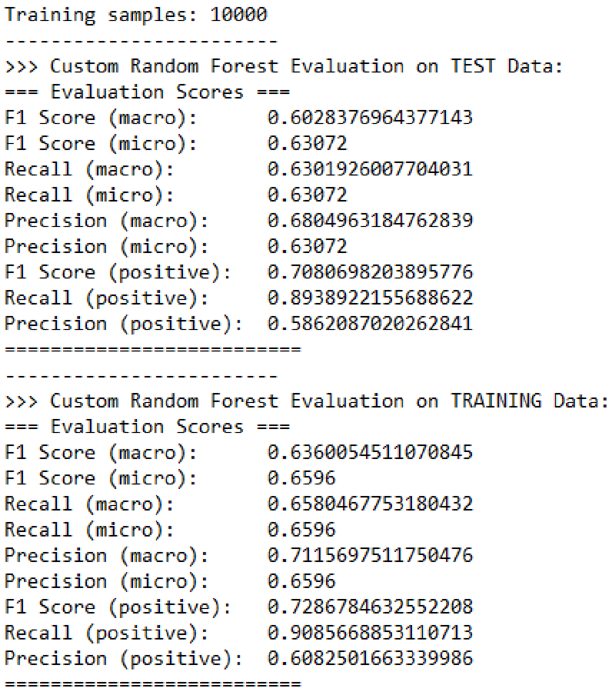
Για τον random forest χρησιμοποιήσαμε τον ID3 του φροντιστηρίου, δημιουργώντας 23 δέντρα με 70 τυχαία features. Ακολουθούν τα evaluation scores:



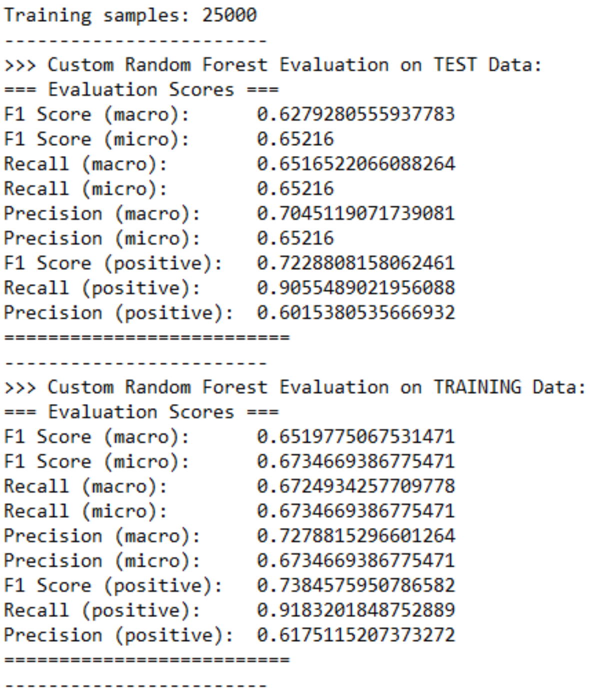
```
Training samples: 500
------------------------
>>> Custom Random Forest Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.4195712766829971
F1 Score (micro):       0.53272
Recall (macro):         0.5336051744206977
Recall (micro):         0.53272
Precision (macro):      0.6550733222896341
Precision (micro):      0.53272
F1 Score (positive):    0.16330038676407393
Recall (positive):      0.09101796407185629
Precision (positive):   0.7933194154488518
=========================
------------------------
>>> Custom Random Forest Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.5138888888888888
F1 Score (micro):       0.608
Recall (macro):         0.5867064542522982
Recall (micro):         0.608
Precision (macro):      0.7750241717978729
Precision (micro):      0.608
F1 Score (positive):    0.3
Recall (positive):      0.17721518987341772
Precision (positive):   0.9767441860465116
=========================
```



```
Training samples: 2500
------------------------
>>> Custom Random Forest Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.704063429352455
F1 Score (micro):       0.70492
Recall (macro):         0.7048152192608771
Recall (micro):         0.70492
Precision (macro):      0.7071241155774861
Precision (micro):      0.70492
F1 Score (positive):    0.719984816853293
Recall (positive):      0.7572055888223553
Precision (positive):   0.6862518089725036
=========================
------------------------
>>> Custom Random Forest Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.7675978140731223
F1 Score (micro):       0.768
Recall (macro):         0.768
Recall (micro):         0.768
Precision (macro):      0.769868091699083
Precision (micro):      0.768
F1 Score (positive):    0.7772657450076804
Recall (positive):      0.8096
Precision (positive):   0.7474150664697193
=========================
```

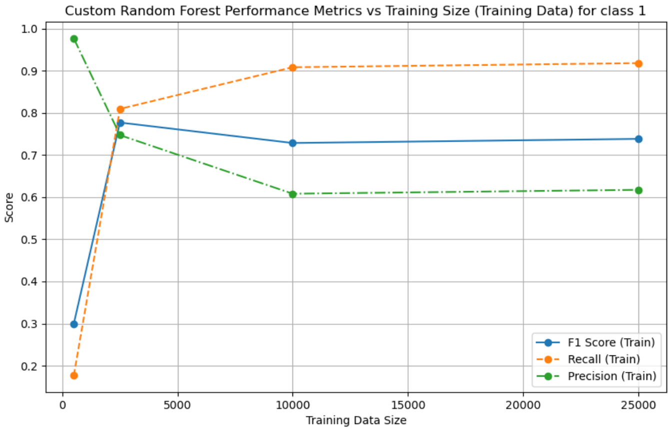


```
Training samples: 10000
------------------------
>>> Custom Random Forest Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.6028376964377143
F1 Score (micro):       0.63072
Recall (macro):         0.6301926007704031
Recall (micro):         0.63072
Precision (macro):      0.6804963184762839
Precision (micro):      0.63072
F1 Score (positive):    0.7080698203895776
Recall (positive):      0.8938922155688622
Precision (positive):   0.5862087020262841
=========================
------------------------
>>> Custom Random Forest Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.6360054511070845
F1 Score (micro):       0.6596
Recall (macro):         0.6580467753180432
Recall (micro):         0.6596
Precision (macro):      0.7115697511750476
Precision (micro):      0.6596
F1 Score (positive):    0.7286784632552208
Recall (positive):      0.9085668853110713
Precision (positive):   0.6082501663339986
=========================
```

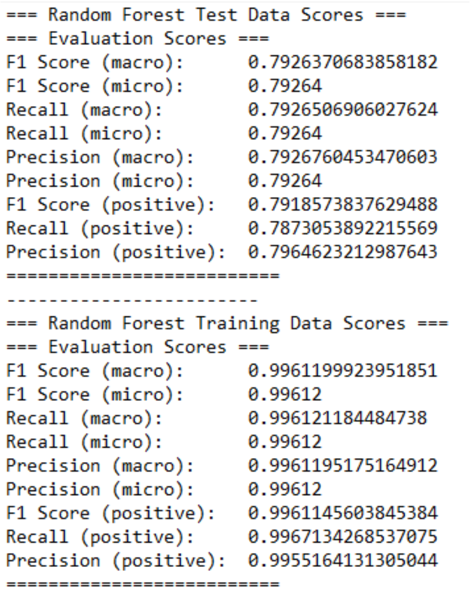


```
Training samples: 25000
------------------------
>>> Custom Random Forest Evaluation on TEST Data:
=== Evaluation Scores ===
F1 Score (macro):       0.6279280555937783
F1 Score (micro):       0.65216
Recall (macro):         0.6516522066088264
Recall (micro):         0.65216
Precision (macro):      0.7045119071739081
Precision (micro):      0.65216
F1 Score (positive):    0.7228808158062461
Recall (positive):      0.9055489021956088
Precision (positive):   0.6015380535666932
=========================
------------------------
>>> Custom Random Forest Evaluation on TRAINING Data:
=== Evaluation Scores ===
F1 Score (macro):       0.6519775067531471
F1 Score (micro):       0.6734669386775471
Recall (macro):         0.6724934257709778
Recall (micro):         0.6734669386775471
Precision (macro):      0.7278815296601264
Precision (micro):      0.6734669386775471
F1 Score (positive):    0.7384575950786582
Recall (positive):      0.9183201848752889
Precision (positive):   0.6175115207373272
=========================
------------------------
```

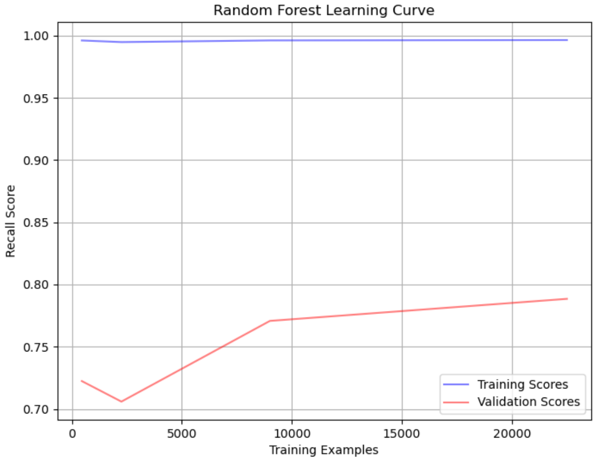
Και το plot:



Για την Random Forest της sklearn:



```
=== Random Forest Test Data Scores ===
=== Evaluation Scores ===
F1 Score (macro):      0.7926370683858182
F1 Score (micro):      0.79264
Recall (macro):        0.7926506906027624
Recall (micro):        0.79264
Precision (macro):     0.7926760453470603
Precision (micro):     0.79264
F1 Score (positive):   0.7918573837629488
Recall (positive):     0.7873053892215569
Precision (positive):  0.7964623212987643
=========================
-----------------------
=== Random Forest Training Data Scores ===
=== Evaluation Scores ===
F1 Score (macro):      0.9961199923951851
F1 Score (micro):      0.99612
Recall (macro):        0.996121184484738
Recall (micro):        0.99612
Precision (macro):     0.9961195175164912
Precision (micro):     0.99612
F1 Score (positive):   0.9961145603845384
Recall (positive):     0.9967134268537075
Precision (positive):  0.9955164131305044
=========================
```


Random Forest Learning Curve
Recall Score
Training Examples
Training Scores
Validation Scores