Τεχνητή νοημοσύνη – 2η προγραμματιστική εργασία

Σμυρνιούδης Νικόλαος (3170148) – Χουλιαράς Δημήτριος (3170180) – Τζώρτζης Σαράντος (3170157)

Οι αλγόριθμοι που υλοποιήσαμε:

- naive Bayes (πολυμετάβλητη μορφή Bernoulli)
- ID3
- Λογιστική παλινδρόμιση (με L2 κανονικοποίηση)

Αρχιτεκτονική

Η αρχιτεκτονική που τηρήσαμε είναι η συγκεκριμένη έτσι ώστε να μοιάζει με την αρχιτεκτονική του πακέτου scikit-learn

Ο χειρισμός των αρχείων ο , χωρισμός σε δεδομένα εκπαίδευσης / επικύρωσης / αξιολόγησης γίνεται απο το datasets.py. Τις μεθόδους αυτές χρησιμοποιεί το classifierTesting.py για να τρέξει τους αλγορίθμους μηχανικής μάθησης για δυο διαφορετικές περιπτώσεις χρήσης.

Για εκπαίδευση classifier και αξιολόγηση τους στο validation dataset :

```
def train_and_test(clf, X_train , Y_train , X_test , Y_test):
    """
    TRains the clf on the given training set
    and returns the accuracy on the test set
    (needs the be the validation set if we are opting to
    tune hyperparameters)
    """
```

Και για την τελική αξιολόγηση και αυτόματη δημιουργία γραφημάτων μέσω Matplotlib η συνάρτηση:

```
def testingSuite(classifier , X_test , Y_test , X_train , Y_train):
    """
    Trains and tests the classifier and outputs time elapsed
    also needs to create graphs for precision , accuracy , recall , f1 score
                                    for training set , and test set
```

Όσον αφορά τους classifiers δημιουργούμε μιά διαφορετική κλάση για κάθε διαφορετικό αλγόριθμο μηχανικής μάθησης η οποία απαιτούμε να υλοποιεί τις εξής δύο μεθόδους.

```
def fit(self,X , Y):
```

```
def predict(self,X):
```

Συγκεκριμένα έχουμε τα αρχεία naiveBayes.py , ID3.py , logisticRegression.py που συντελούν τις δικές μας υλοποιήσεις των αλγορίθμων και τα αρχεία sciNaive.py sciID3.py sciLogistic.py που είναι οι έτοιμες υλοποιήσεις του scikit-learn ενσωματωμένες σε μιά κλάση που υλοποιεί τις μεθόδους fit,predict που ορίσαμε παραπάνω.

Τέλος κάθε αλγόριθμος χρησιμοποιεί την κλάση featureExtractor.emailFeatureExtractor που βρίσκεται στο αρχείο featureExtractor.py για να εξαχθούν τα διανύσματα ιδιοτήτων σε κάθε κλήση της predict και fit. Επίσης στην αρχικοποίηση του featureExtractor σε κάθε ταξονομιτή δίνεται ώς όρισμα ένας αριθμός m που χρησιμοποιείται στην **αυτόματη επιλογή ιδιοτήτων με βάση το κέρδος πληροφορίας τους.**

Για να τρέξουμε τον κάθε αλγόριθμο αρκεί να εκτελέσουμε python (η python3) sciNaive.py (η logisticRegression.py η ID3.py) και το κάθε python πρόγραμμα θα τρέξει την συνάρτηση testingSuite και θα παράξει τα διαγράμματα μετρικών και τους αντίστοιχους πίνακες αποτελεσμάτων.

Αποτελέσματα LogisticRegression:

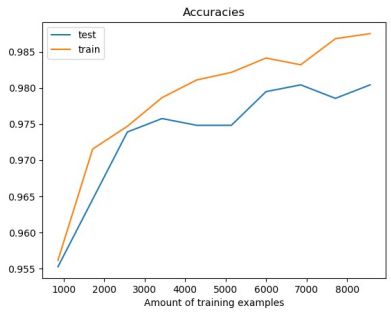
Για να βρούμε τις κατάλληλες υπεπαραμέτρους τρέξαμε τον αλγόριθμο για κάθε συνδυασμό τιμών των υπερπαραμέτρων στα δεδομένα επικύρωσης και κρατήσαμε τις υπερπαραμέτρους με το καλύτερο accuracy στα δεδομένα επικύρωσης. Συγκεκριμένα η κάθε παραμέτρους μπορούσε να πάρει τις εξής τιμές:

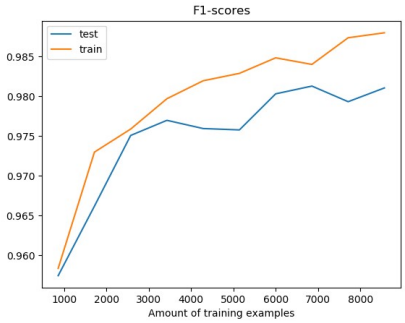
```
learning_rates = [0.1 , 0.01 , 0.001]
max_iters = [1 , 3 , 5 , 10 ]
regularization_terms = [0.1 , 0.001 , 0.0001]
attributes = [100 , 500 , 1000]
```

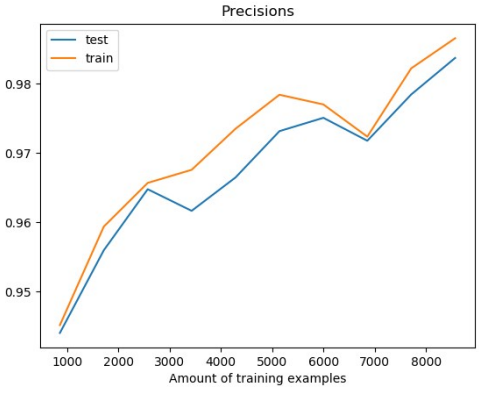
Και τελικά ο καλύτερος συνδυασμός παραμέτρων ήταν:

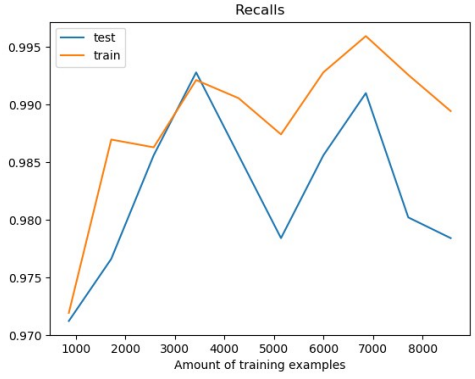
```
Best parameters :  [0.1, 3, 0.0001, 1000]
```

Τα τελικά διαγράμματα των μετρικών είναι:









Και οι πίνακες αποτελεσμάτων :

	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.95837023915	0.956156716418	0.971929036605	0.945184538109
3	1714	0.972990923179	0.971548507463	0.986975072985	0.95939751146
4	2571	0.975891567604	0.974696828358	0.986301369863	0.965699208443
5	3428	0.979709502162	0.978661380597	0.992140130249	0.967586508979
6	4285	0.98196794301	0.981110074627	0.990568156299	0.973515780181
7	5142	0.982899295853	0.982159514925	0.987424208399	0.978415665332
8	5999	0.984851860102	0.984141791045	0.992813833371	0.977016574586
9	6856	0.984024850233	0.983208955224	0.995957781271	0.972374479281
10	7713	0.987378532336	0.98682369403	0.992589265664	0.982222222222
11	8570	0.988003139365	0.987523320896	0.989445317763	0.986565158979

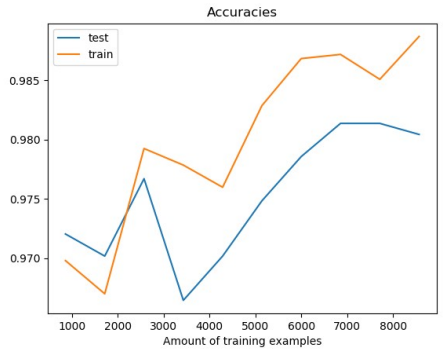
Για τις δοκιμές στα δεδομένα εκπαίδευσης:

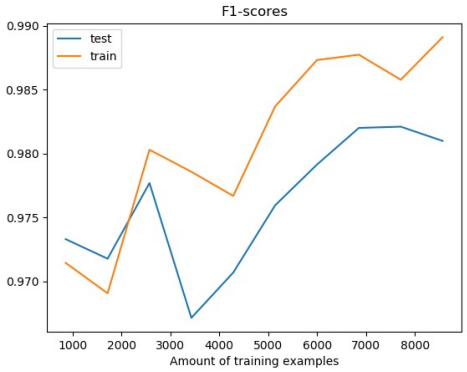
	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.957446808511	0.955265610438	0.971223021583	0.944055944056
3	1714	0.966192170819	0.96458527493	0.976618705036	0.955985915493
4	2571	0.975088967972	0.973904939422	0.985611510791	0.964788732394
5	3428	0.976991150442	0.975768872321	0.992805755396	0.961672473868
6	4285	0.975957257346	0.974836905871	0.985611510791	0.966490299824
7	5142	0.975784753363	0.974836905871	0.978417266187	0.973166368515
8	5999	0.980322003578	0.979496738117	0.985611510791	0.975088967972
9	6856	0.981300089047	0.980428704567	0.991007194245	0.971781305115
10	7713	0.979335130279	0.978564771668	0.980215827338	0.978456014363
11	8570	0.981064021641	0.980428704567	0.978417266187	0.983725135624

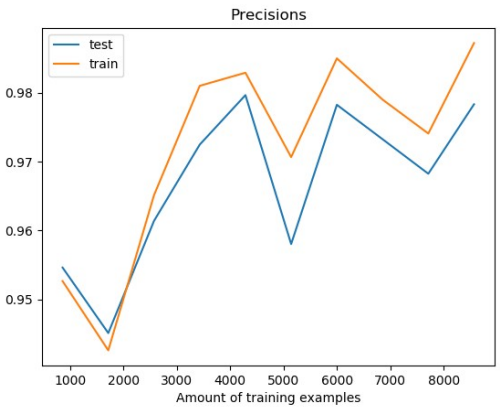
Για τις δοκιμές στα δεδομένα αξιολόγησης

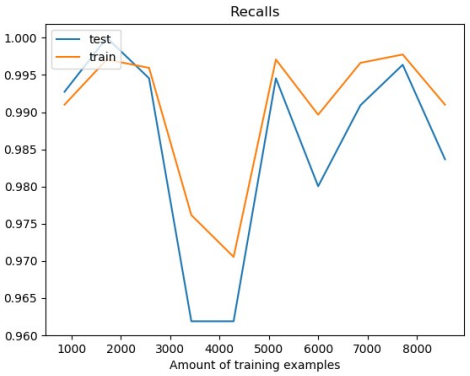
Αντίστοιχα αποτελέσματα για την υλοποίηση της λογιστικής παλινδρόμισης με την βιβλιοθήκη scikit:

Τελικά διαγράμματα μετρικών:









Και οι πίνακες αποτελεσμάτων:

	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.973309608541	0.972041006524	0.992740471869	0.95462478185
3	1714	0.971781305115	0.970177073625	1	0.945111492281
4	2571	0.977698483497	0.97670083877	0.994555353902	0.961403508772
5	3428	0.967153284672	0.966449207829	0.961887477314	0.97247706422
6	4285	0.970695970696	0.970177073625	0.961887477314	0.97966728281
7	5142	0.975957257346	0.974836905871	0.994555353902	0.958041958042
8	5999	0.979147778785	0.978564771668	0.980036297641	0.978260869565
9	6856	0.982014388489	0.981360671016	0.990925589837	0.973262032086
10	7713	0.982110912343	0.981360671016	0.996370235935	0.968253968254
11	8570	0.980995475113	0.980428704567	0.983666061706	0.978339350181

Για τα δεδομένα αξιολόγησης

	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.971453763915	0.969799440299	0.991005172026	0.952658884565
3	1714	0.969074418096	0.967000932836	0.997076680908	0.942602040816
4	2571	0.98030101815	0.979244402985	0.995952327412	0.965134016126
5	3428	0.978584310189	0.977845149254	0.976163705869	0.981016949153
6	4285	0.976691559176	0.975979477612	0.970541938385	0.98291960829
7	5142	0.983693843594	0.982859141791	0.997076680908	0.970665499124
8	5999	0.987324733595	0.98682369403	0.98965594783	0.985004476276
9	6856	0.987742366838	0.987173507463	0.99662693951	0.979014800088
10	7713	0.985780937569	0.985074626866	0.997751293007	0.974094401756
11	8570	0.989114577488	0.988689365672	0.991005172026	0.987231182796

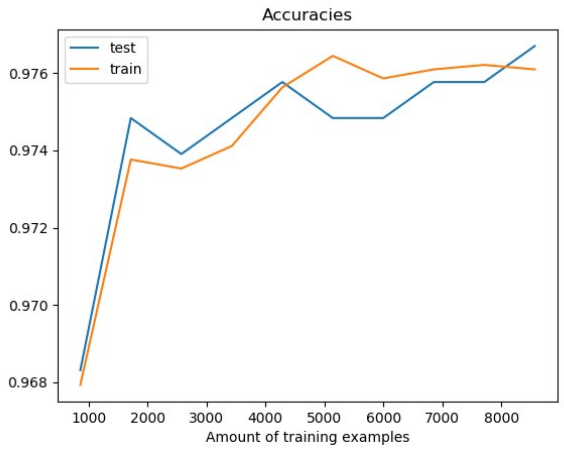
Για τα δεδομένα εκπαίδευσης

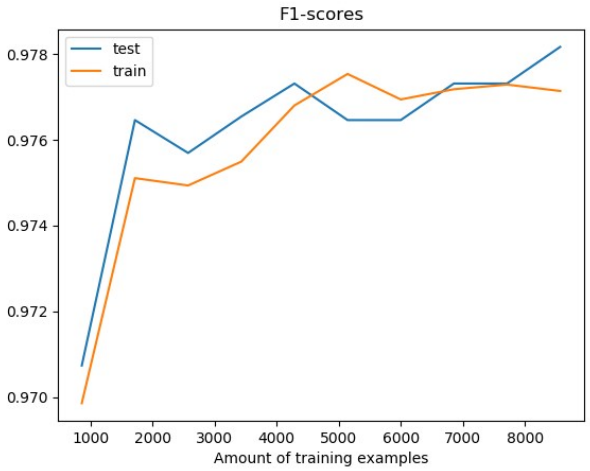
Αποτελέσματα Naive Bayes:

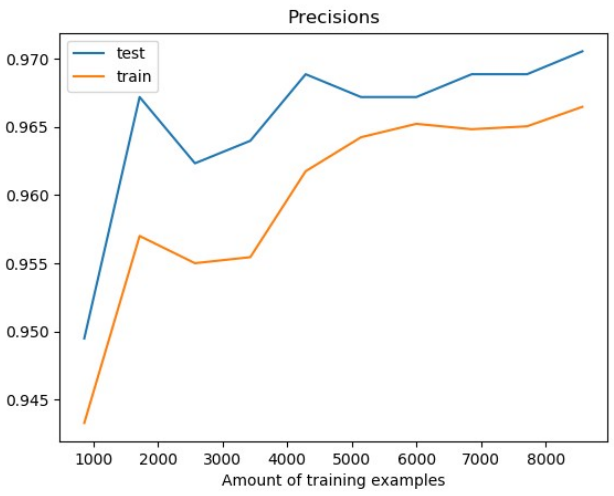
Για να βρούμε την υπερπαράμετρο m (πλήθος των ιδιοτήτων) εκπαιδεύσαμε τον αλγόριθμο στα δεδομένα εκπαίδευσης μία φορά για κάθε διαφορετική τιμή του m και κρατήσαμε το m με το οποίο ο αλγόριθμος τα πήγε καλύτερα στα δεδομένα επικύρωσης. Συγκεκριμένα αυτό το m ηταν το 1000.

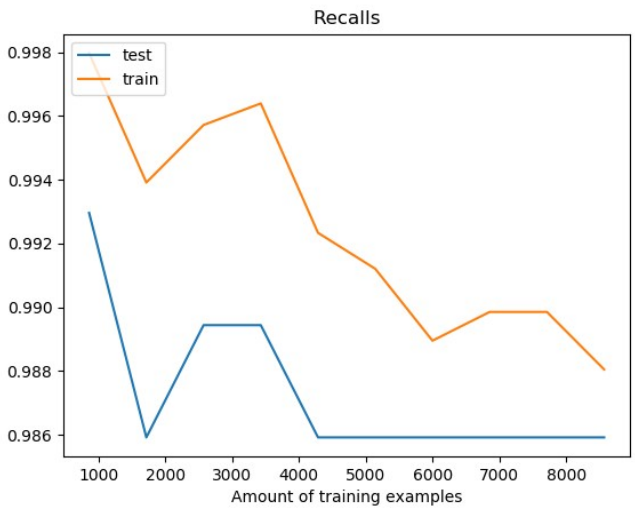
```
Best parameters :  [1000]
```

Τα αποτελέσματα των μετρικών ήταν:









Και οι πίνακες των αποτελεσμάτων:

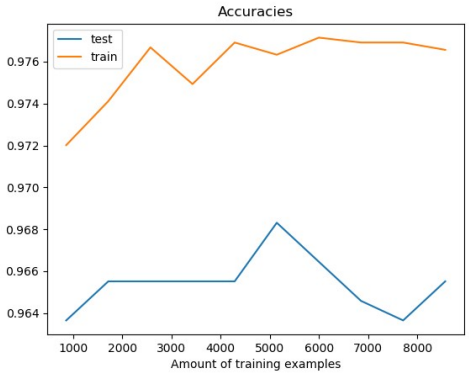
	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.97074010327	0.968313140727	0.992957746479	0.949494949495
3	1714	0.976460331299	0.974836905871	0.985915492958	0.967184801382
4	2571	0.975694444444	0.973904939422	0.989436619718	0.962328767123
5	3428	0.976542137272	0.974836905871	0.989436619718	0.96397941681
6	4285	0.977312390925	0.975768872321	0.985915492958	0.968858131488
7	5142	0.976460331299	0.974836905871	0.985915492958	0.967184801382
8	5999	0.976460331299	0.974836905871	0.985915492958	0.967184801382
9	6856	0.977312390925	0.975768872321	0.985915492958	0.968858131488
10	7713	0.977312390925	0.975768872321	0.985915492958	0.968858131488
11	8570	0.978165938865	0.97670083877	0.985915492958	0.970537261698

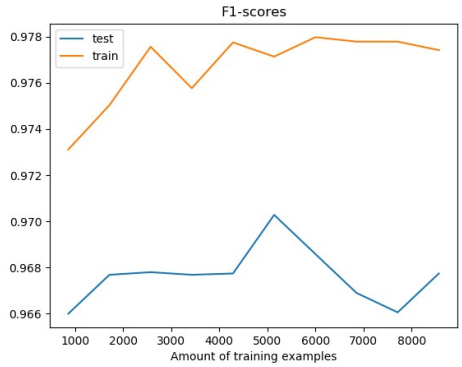
Σε δεδομένα αξιολόγησης

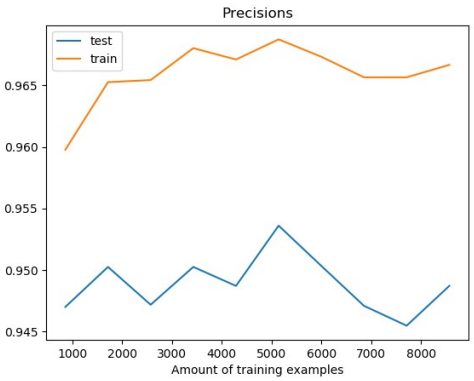
	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.969863013699	0.967933768657	0.997970230041	0.943295672564
3	1714	0.975107865914	0.973763992537	0.993910690122	0.957003257329
4	2571	0.974936513194	0.973530783582	0.995714930086	0.955007570841
5	3428	0.975491278428	0.97411380597	0.996391520072	0.95544982699
6	4285	0.976800976801	0.975629664179	0.992331980153	0.96174863388
7	5142	0.977535587189	0.976445895522	0.991204330176	0.964238701185
8	5999	0.976941071627	0.975862873134	0.988949030221	0.965221219459
9	6856	0.977179116108	0.97609608209	0.989851150203	0.964827434601
10	7713	0.977287909152	0.976212686567	0.989851150203	0.965039577836
11	8570	0.977138396342	0.97609608209	0.988046910239	0.966468122656

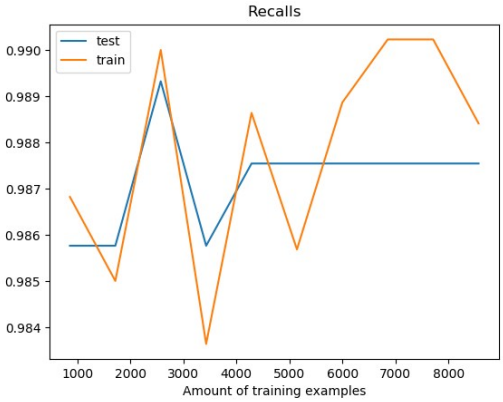
Σε δεδομένα εκπαίδευσης

Αντίστοιχα αποτελέσματα για την υλοποίηση του Naive Bayes με την βιβλιοθήκη scikit:









Και οι πίνακες αποτελεσμάτων:

	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.965998256321	0.963653308481	0.985765124555	0.947008547009
3	1714	0.96768558952	0.965517241379	0.985765124555	0.95025728988
4	2571	0.967798085292	0.965517241379	0.989323843416	0.947189097104
5	3428	0.96768558952	0.965517241379	0.985765124555	0.95025728988
6	4285	0.967741935484	0.965517241379	0.987544483986	0.948717948718
7	5142	0.97027972028	0.968313140727	0.987544483986	0.953608247423
8	5999	0.968586387435	0.966449207829	0.987544483986	0.950342465753
9	6856	0.966898954704	0.96458527493	0.987544483986	0.947098976109
10	7713	0.966057441253	0.963653308481	0.987544483986	0.945485519591
11	8570	0.967741935484	0.965517241379	0.987544483986	0.948717948718

Για δεδομένα αξιολόγησης

	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.97311225633	0.972014925373	0.986821177005	0.959779005525
3	1714	0.975033738192	0.97411380597	0.985003408316	0.965263861055
4	2571	0.977563383442	0.976679104478	0.990002272211	0.965433192998
5	3428	0.975769187423	0.974930037313	0.9836400818	0.968023255814
6	4285	0.977752808989	0.976912313433	0.988638945694	0.967103800845
7	5142	0.977137064985	0.976329291045	0.985685071575	0.968736042876
8	5999	0.97797752809	0.977145522388	0.98886616678	0.967326072461
9	6856	0.977787749607	0.976912313433	0.990229493297	0.965654775094
10	7713	0.977787749607	0.976912313433	0.990229493297	0.965654775094
11	8570	0.97741826761	0.9765625	0.988411724608	0.966666666667

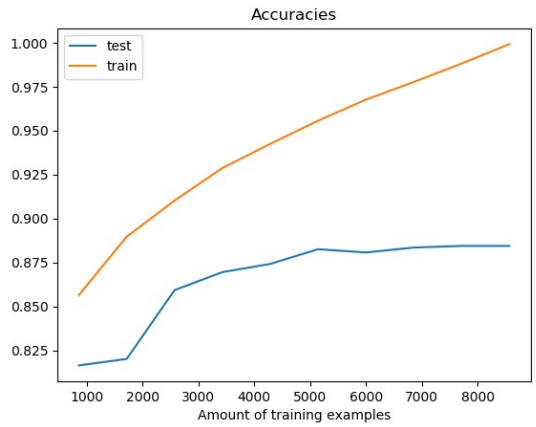
Για δεδομένα επικύρωσης

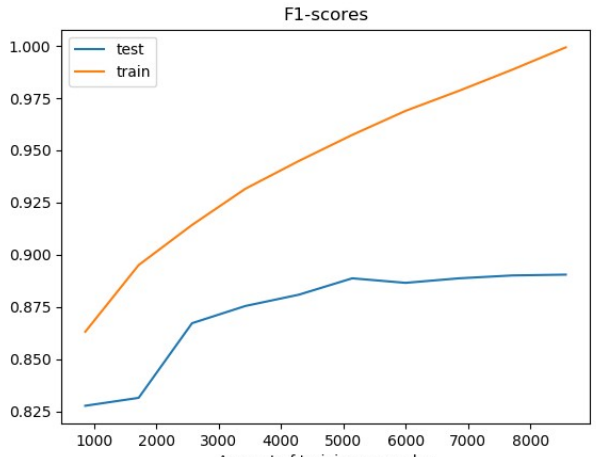
Αποτελέσματα ID3:

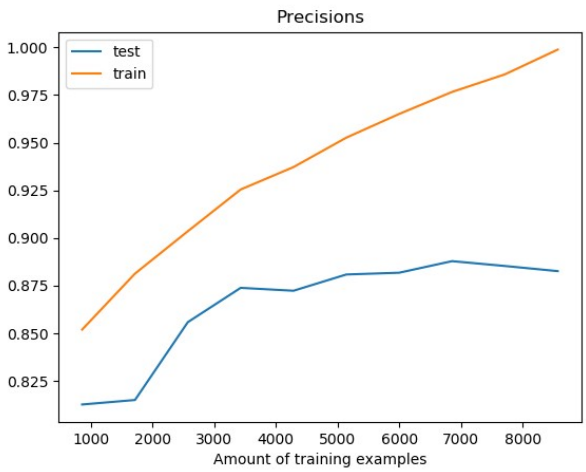
Για να βρούμε την υπερπαράμετρο m (πλήθος των ιδιοτήτων) εκπαιδεύσαμε τον αλγόριθμο στα δεδομένα εκπαίδευσης μία φορά για κάθε διαφορετική τιμή του m και κρατήσαμε το m με το οποίο ο αλγόριθμος τα πήγε καλύτερα στα δεδομένα επικύρωσης. Συγκεκριμένα αυτό το m ηταν το 1000.

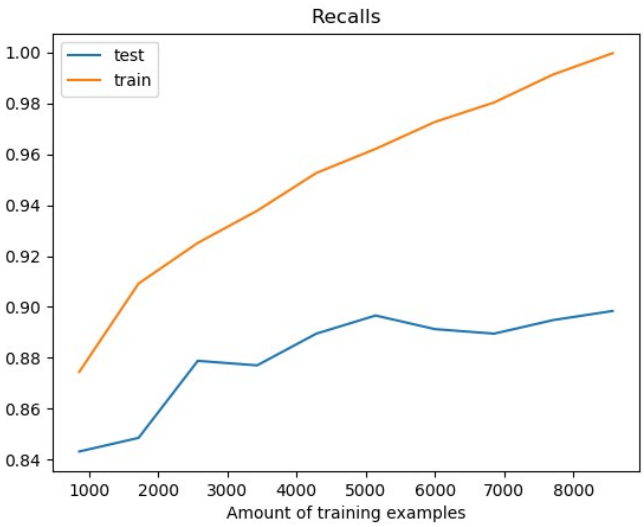
```
Best parameters :  [500]
```

Τα αποτελέσματα των μετρικών ήταν:









Και οι πίνακες με τα αποτελέσματα:

	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.827646544182	0.816402609506	0.843137254902	0.812714776632
3	1714	0.831441048035	0.820130475303	0.848484848485	0.815068493151
4	2571	0.867194371152	0.85927306617	0.878787878788	0.855902777778
5	3428	0.875444839858	0.869524697111	0.877005347594	0.873889875666
6	4285	0.880847308032	0.874184529357	0.889483065954	0.872377622378
7	5142	0.888692579505	0.8825722274	0.896613190731	0.880910683012
8	5999	0.886524822695	0.880708294501	0.891265597148	0.881834215168
9	6856	0.888691006233	0.883504193849	0.889483065954	0.887900355872
10	7713	0.890070921986	0.884436160298	0.894830659537	0.885361552028
11	8570	0.890459363958	0.884436160298	0.898395721925	0.882661996497

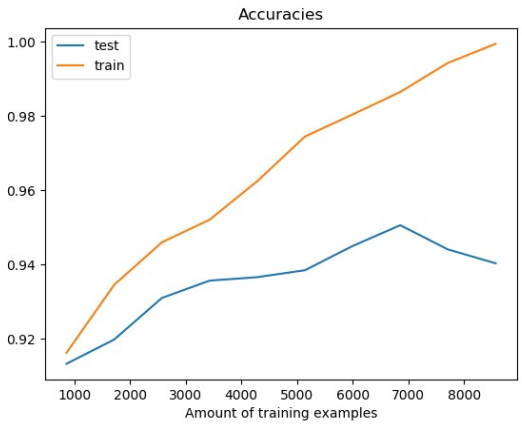
Δεδομένα αξιολόγησης

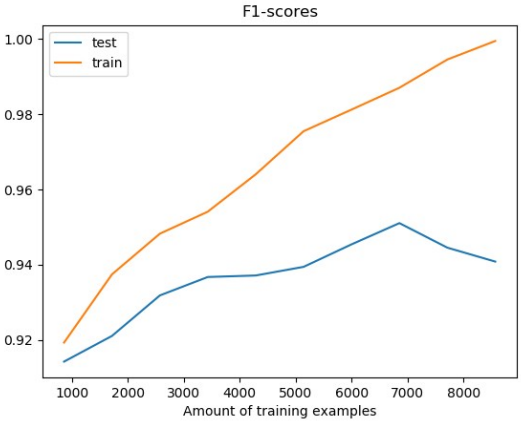
	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.86308530753	0.85645988806	0.87446472842	0.851998243303
3	1714	0.895052141114	0.889692164179	0.909172864548	0.881363338431
4	2571	0.91425389755	0.910214552239	0.925174667568	0.903587937486
5	3428	0.931601925445	0.928754664179	0.937795807978	0.925489323843
6	4285	0.944897731083	0.942513992537	0.952670723462	0.937250554324
7	5142	0.957389549226	0.955690298507	0.962136578769	0.95268913189
8	5999	0.968907845998	0.967700559701	0.972729321614	0.96511627907
9	6856	0.97851760207	0.977728544776	0.980392156863	0.976650202066
10	7713	0.988650410158	0.988222947761	0.991435654722	0.985880770955
11	8570	0.999324172111	0.999300373134	0.999774622493	0.998874127449

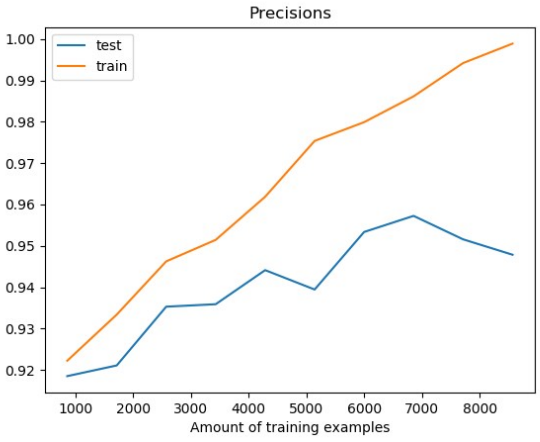
Δεδομένα εκπαίδευσης

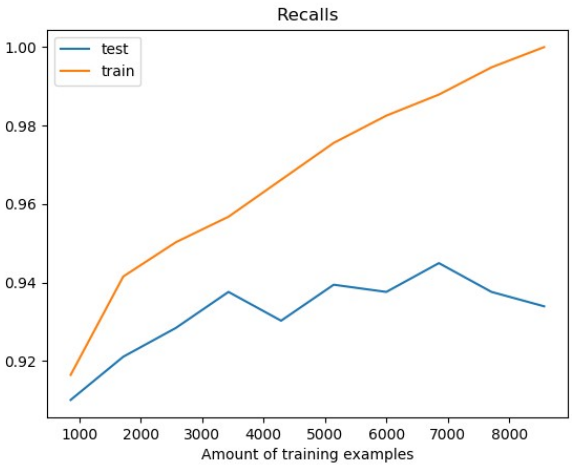
Τα αντίστοιχα αποτελέσματα απο το scikit:

Τα διαγράμματα μετρικών:









Οι πίνακες αποτελεσμάτων:

	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.914285714286	0.913327120224	0.910091743119	0.918518518519
3	1714	0.921100917431	0.919850885368	0.921100917431	0.921100917431
4	2571	0.931860036832	0.931034482759	0.928440366972	0.935304990758
5	3428	0.936755270394	0.935694315005	0.937614678899	0.935897435897
6	4285	0.937153419593	0.936626281454	0.930275229358	0.944134078212
7	5142	0.939449541284	0.938490214352	0.939449541284	0.939449541284
8	5999	0.945420906568	0.945013979497	0.937614678899	0.953358208955
9	6856	0.951061865189	0.950605778192	0.94495412844	0.957249070632
10	7713	0.944547134935	0.944082013048	0.937614678899	0.951582867784
11	8570	0.940850277264	0.940354147251	0.933944954128	0.947858472998

Δεδομένα αξιολόγησης

	A	B	C	D	E
1	n	f1-score	accuracy	recall	precision
2	857	0.919343967648	0.916277985075	0.916461366181	0.922244759973
3	1714	0.937451220872	0.93458488806	0.941545352744	0.933392539964
4	2571	0.948262375684	0.946012126866	0.950279955207	0.946253345227
5	3428	0.954103852596	0.952075559701	0.956774916013	0.95144766147
6	4285	0.964022346369	0.962453358209	0.966181410974	0.961872909699
7	5142	0.975478669802	0.974463619403	0.975587905935	0.975369458128
8	5999	0.981212256766	0.980410447761	0.982530795073	0.979897252625
9	6856	0.987021705079	0.986473880597	0.98790593505	0.986139056562
10	7713	0.994514720699	0.994286380597	0.994848824188	0.99418084154
11	8570	0.99944040291	0.999416977612	1	0.998881431767

Δεδομένα εκπαίδευσης