ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Τεχνητή Νοημοσύνη

Χειμερινό Εξάμηνο 2021-2022

Εργασία: 2

Απόστολος Γρηγόρης 3190335

Ευάγγελος Γεωργακόπουλος 3150019

ID3 με πρόωρο τερματισμό

Δομή του project:

myID3.py

Αρχικά γίνεται η προσκόμιση του Dataset,η δημιουργία του λεξιλογίου και η μετατροπή των κειμένων σε διανύσματα ιδιοτήτων με τιμές 0 ή 1, οι οποίες δείχνουν ποιες λέξεις του λεξιλογίου περιέχει το κείμενο. Χρησιμοποιήθηκε για αυτό ο κώδικας του Εργαστηρίου υπ' αριθμό 8. Παραμετροποιήθηκε κατάλληλα, ώστε ο αριθμός των training data, το μέγεθος του λεξιλογίου και ο αριθμός n που δηλώνει τον αριθμό των συχνότερων λέξεων που θέλουμε να παραλείπεται απ' το λεξιλόγιο, να εισάγονται ως υπερπαράμετροι.

Έπειτα δηλώνονται συναρτήσεις που υπολογίζουν την εντροπία και το κέρδος πληροφορίας.

Στη συνέχεια ακολουθεί ο αλγόριθμος του ID3 Δέντρου Αποφάσεων που έχει υλοποιηθεί με πρόωρο τερματισμό της επέκτασης. Αποτελείται από τις:

- partition_classes: χωρίζει τα training data με κριτήριο την τιμή ενός χαρακτηριστικού (στη δικιά μας περίπτωση τα χαρακτηριστικά αυτά είναι οι λέξεις)
- find_best_feature: προσπελαύνει όλες τις λέξεις και επιστρέφει εκείνη που αν επιλεχθεί προσφέρει το μέγιστο κέρδος πληροφορίας
- MyDecisionTree: είναι η κλάση του δέντρου μας. Αποτελείται από:
    - Τον constructor της

    - Την fit: αναδρομική συνάρτηση που βρίσκει σε κάθε βήμα τη λέξη που προσφέρει το μέγιστο κέρδος πληροφορίας, και χωρίζει τα training data με βάση αυτή, δημιουργώντας έτσι το δέντρο. Σταματά είτε όταν το 95% των παραδειγμάτων του κόμβου ανήκει στην ίδια κατηγορία, είτε όταν το δέντρο φτάσει το max βάθος που ορίζεται ως υπερπαράμετρος.
    - predict: ανατρέχοντας το δημιουργηθέν δέντρο, «προβλέπει» το αποτέλεσμα (0 ή 1) των δεδομένων που λαμβάνει σαν είσοδο.
- DecisionTreeEvalution: συγκρίνει τα αποτελέσματα της predict με τα γνωστά αποτελέσματα και υπολογίζει τις τιμές των μεθόδων αξιολόγησης του δέντρου μας.

Τέλος, καλούνται οι αντίστοιχες συναρτήσεις για τη δημιουργία του ID3 δέντρου μας και την αξιολόγησή του και αυτές τις sklearn για την DecisionTreeClassifier, ώστε να συγκρίνουμε τα αποτελέσματά μας.

Ακολουθούν πίνακες που εκφράζουν τα αποτελέσματα της ID3 μας, σε συνάρτηση με τις παραμέτρους που χρησιμοποιήθηκαν και τα αντίστοιχα αποτελέσματα της DecisionTreeClassifier.

Το αρχείο myID3.py εκτελέστηκε με τις εξής παραμέτρους:

Α/Α	Αριθμός συχνότερων λέξεων	Μέγεθος δεδομένων εκπαίδευσης	Αριθμός συχνότερων λέξεων προς παράλειψη	Μέγιστο βάθος δέντρου
1	500	2500	50	11
2	500	5000	50	12
3	500	7500	50	13
4	500	10000	50	13
5	500	12500	50	14
6	500	15000	50	14
7	500	17500	50	14
8	500	20000	50	14
9	500	22500	50	14
10	500	25000	50	15

Ως αριθμός συχνότερων λέξεων και αριθμός συχνότερων λέξεων προς παράλειψη επιλέχθηκαν μετά από αρκετές δοκιμές εκείνοι που έφερναν ικανοποιητικά αποτελέσματα σε συνάρτηση με το χρόνο εκτέλεσης.

Με το ίδιο κριτήριο επιλέξαμε να εκχωρούμε μέγιστο βάθος δέντρου το $\log_2$ του μεγέθους δεδομένων εκπαίδευσης.

Αποτελέσματα:

MyID3 (με τα test data)

A/A	1	2	3	4	5	6	7	8	9	10
Accuracy	0.7076	0.7057	0.7009	0.7129	0.7208	0.7094	0.7080	0.7149	0.7140	0.7225
Precision	0.6663	0.6637	0.6864	0.6882	0.6839	0.6972	0.6692	0.7021	0.6959	0.7080
Recall	0.8317	0.8340	0.7398	0.7786	0.8213	0.7401	0.8227	0.7464	0.7599	0.7574
F1	0.7399	0.7392	0.7121	0.7306	0.7463	0.7180	0.7381	0.7236	0.7265	0.7319

MyID3 (με τα train data)

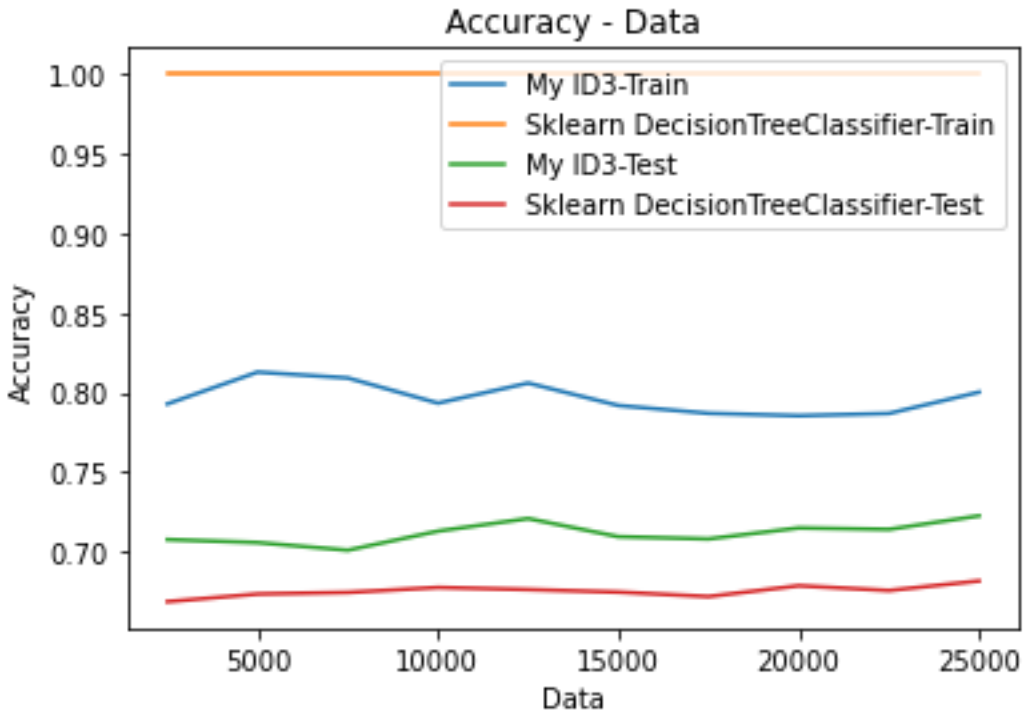
A/A	1	2	3	4	5	6	7	8	9	10
Accuracy	0.7928	0.8128	0.8091	0.7933	0.8059	0.7918	0.7869	0.7854	0.7868	0.8002
Precision	0.7455	0.7567	0.7910	0.7647	0.7555	0.7758	0.7359	0.7690	0.7619	0.7823
Recall	0.9048	0.9321	0.8465	0.8536	0.9098	0.8233	0.8986	0.8178	0.8363	0.8320
F1	0.8175	0.8353	0.8178	0.8067	0.8255	0.7989	0.8092	0.7926	0.7973	0.8064

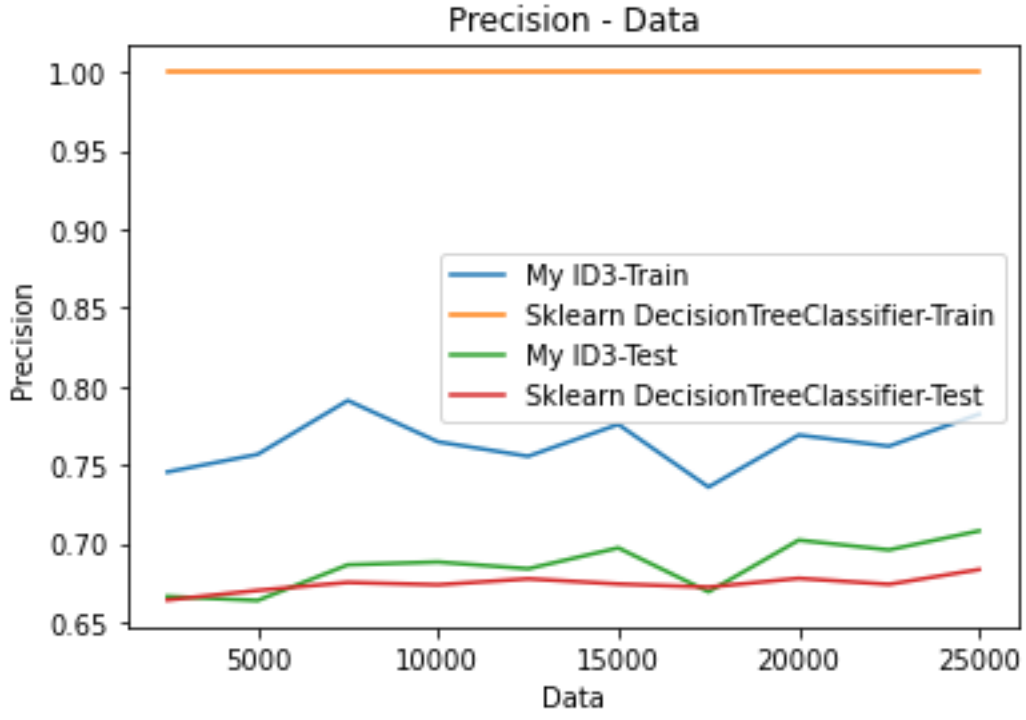
Sklearn DecisionTreeClassifier (με τα test data)

A/A	1	2	3	4	5	6	7	8	9	10
Accuracy	0.6686	0.6734	0.6744	0.6775	0.6762	0.6746	0.6718	0.6786	0.6756	0.6816
Precision	0.6640	0.6701	0.6752	0.6737	0.6775	0.6742	0.6722	0.6778	0.6740	0.6834
Recall	0.6828	0.6833	0.6724	0.6885	0.6724	0.6759	0.6704	0.6809	0.6802	0.6766
F1	0.6733	0.6766	0.6738	0.6810	0.6750	0.6751	0.6713	0.6794	0.6770	0.6800

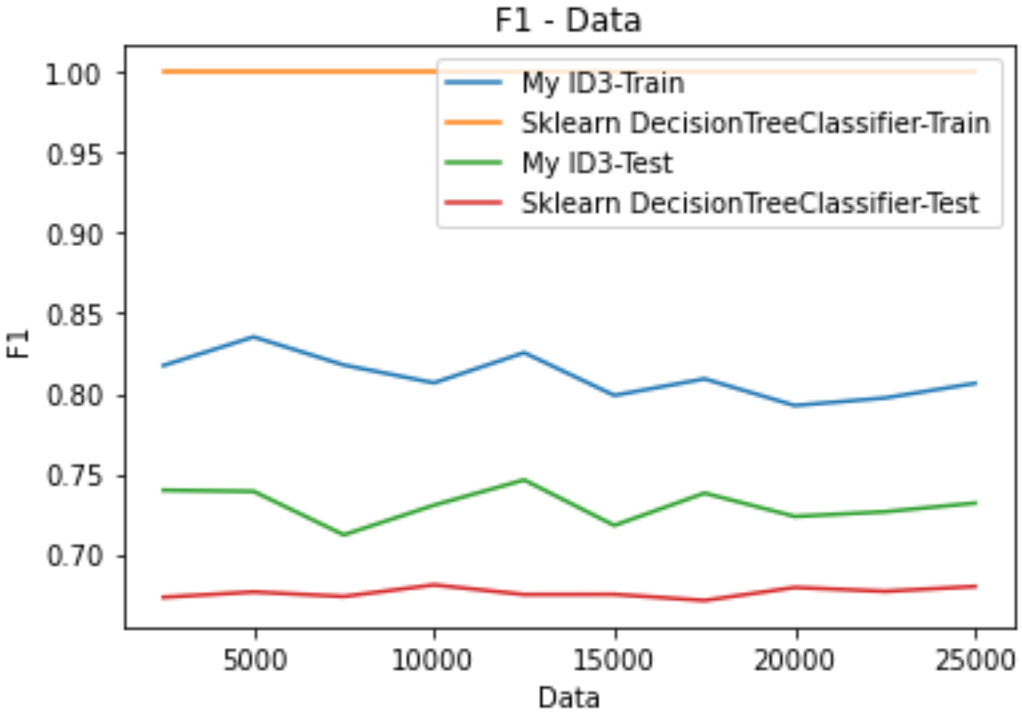
Sklearn DecisionTreeClassifier (με τα test data)

A/A	1	2	3	4	5	6	7	8	9	10
Accuracy	1	1	1	1	1	1	1	1	1	1
Precision	1	1	1	1	1	1	1	1	1	1
Recall	1	1	1	1	1	1	1	1	1	1
F1	1	1	1	1	1	1	1	1	1	1


Accuracy - Data
My ID3-Train
Sklearn DecisionTreeClassifier-Train
My ID3-Test
Sklearn DecisionTreeClassifier-Test
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
5000
10000
15000
20000
25000
Data


Precision - Data
My ID3-Train
Sklearn DecisionTreeClassifier-Train
My ID3-Test
Sklearn DecisionTreeClassifier-Test
Precision
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
5000
10000
15000
20000
25000
Data


Recall - Data
Recall
Data
1.00
0.95
0.90
0.85
0.80
0.75
0.70
5000
10000
15000
20000
25000
My ID3-Train
Sklearn DecisionTreeClassifier-Train
My ID3-Test
Sklearn DecisionTreeClassifier-Test


F1 - Data
F1
Data
1.00
0.95
0.90
0.85
0.80
0.75
0.70
5000
10000
15000
20000
25000
My ID3-Train
Sklearn DecisionTreeClassifier-Train
My ID3-Test
Sklearn DecisionTreeClassifier-Test

RandomForest χρησιμοποιώντας την υλοποίηση μας για τον ID3

Δομή του project:

myRandomForest.py

Μεταξύ άλλων βιβλιοθηκών καλεί και το myID3.py που περιεγράφηκε ανωτέρω, ώστε να δημιουργηθεί το δάσος με τα αντικείμενα της κλάσης MyDecisionTree. Επίσης χρησιμοποιεί το αρχείο για την ανάγνωση των δεδομένων και τη μετατροπή τους στην επιθυμητή και ζητούμενη μορφή. Το μέγιστο βάθος δέντρου και ο αριθμός των δέντρων ορίζονται ως υπερπαράμετροι. Η υλοποίηση του RandomForest γίνεται από την κλάση MyRandomForest, η οποία αποτελείται από:

- Τον constructor της
- Τη _bootstrapping: επιλέγει με κάποια τυχαιότητα ένα υποσύνολο δεικτών γραμμών και στηλών. Οι γραμμές μπορεί να επαναλαμβάνονται. Οι στήλες όχι.
- Τη bootstrapping: καλεί για κάθε δέντρο του δάσους μας την _bootstrapping, με σκοπό κάθε δέντρο να «εκπαιδευτεί» με διαφορετικά δεδομένα. Κρατάει τους δείκτες στηλών και γραμμών που επιλέχθηκαν για κάθε δέντρο.
- Την fit: για κάθε δέντρο του δάσους μας δημιουργεί training data από τους δείκτες που έχουν επιλεγεί και τα αρχικά training data. Έπειτα καλεί την fit της MyDecisionTree.
- RFpredict: ανατρέχοντας κάθε δέντρο του δάσους μας, «προβλέπει» το αποτέλεσμα (0 ή 1) των δεδομένων που λαμβάνει σαν είσοδο κάθε δέντρο. Έπειτα, συγκρίνει τα διαφορετικά αποτελέσματα (για τα ίδια δεδομένα) μεταξύ τους και κρατάει αυτό που πλεονεκτεί αριθμητικά. Στη συνέχεια, συγκρίνει τα τελικά αποτελέσματα με τα γνωστά αποτελέσματα και υπολογίζει τις τιμές των μεθόδων αξιολόγησης του δάσους μας.

Τέλος, καλούνται οι αντίστοιχες συναρτήσεις για τη δημιουργία του Random Forest μας και την αξιολόγησή του και αυτές τις sklearn για την RandomForestClassifier, ώστε να συγκρίνουμε τα αποτελέσματά μας.

Ακολουθούν πίνακες που εκφράζουν τα αποτελέσματα της RandomForest μας, σε συνάρτηση με τις παραμέτρους που χρησιμοποιήθηκαν και τα αντίστοιχα αποτελέσματα της RandomForestClassifier.

Το αρχείο myRandomForest.py εκτελέστηκε με τις εξής παραμέτρους:

Α/Α	Αριθμός συχνότερων λέξεων	Μέγεθος δεδομένων εκπαίδευσης	Αριθμός συχνότερων λέξεων προς παράλειψη	Μέγιστο βάθος δέντρου	Πλήθος δέντρων
1	500	2500	50	11	10
2	500	5000	50	12	10
3	500	7500	50	13	10
4	500	10000	50	13	10
5	500	12500	50	14	10
6	500	15000	50	14	10
7	500	17500	50	14	10
8	500	20000	50	14	10
9	500	22500	50	14	10
10	500	25000	50	15	10

Ως αριθμός συχνότερων λέξεων και αριθμός συχνότερων λέξεων προς παράλειψη επιλέχθηκαν μετά από αρκετές δοκιμές εκείνοι που έφερναν ικανοποιητικά αποτελέσματα σε συνάρτηση με το χρόνο εκτέλεσης.

Με το ίδιο κριτήριο επιλέξαμε να εκχωρούμε μέγιστο βάθος δέντρου το $\log_2$ του μεγέθους δεδομένων εκπαίδευσης.

Όσο ανέβαινε το πλήθος δέντρων, ανέβαιναν και οι τιμές των ποσοστών μετρήσεων. Παρ' όλα αυτά μετά την τιμή 10, το αρχείο έκανε υπερβολικό χρόνο να εκτελεστεί, οπότε επιλέχθηκε η μέγιστη τιμή με την οποία ήταν εφικτό να εκτελεστεί.

Αποτελέσματα:

myRandomForest (με τα test data)

Α/Α	1	2	3	4	5	6	7	8	9	10
Accuracy	0.7283	0.7388	0.7489	0.7365	0.7501	0.7600	0.7561	0.7503	0.7526	0.7585
Precision	0.7063	0.7117	0.7240	0.7158	0.7252	0.7455	0.7453	0.7296	0.7382	0.7402
Recall	0.7814	0.8027	0.8046	0.7846	0.8052	0.7894	0.7781	0.7954	0.7829	0.7965
F1	0.7420	0.7545	0.7622	0.7486	0.7631	0.7669	0.7613	0.7611	0.7599	0.7673

myRandomForest (με τα train data)

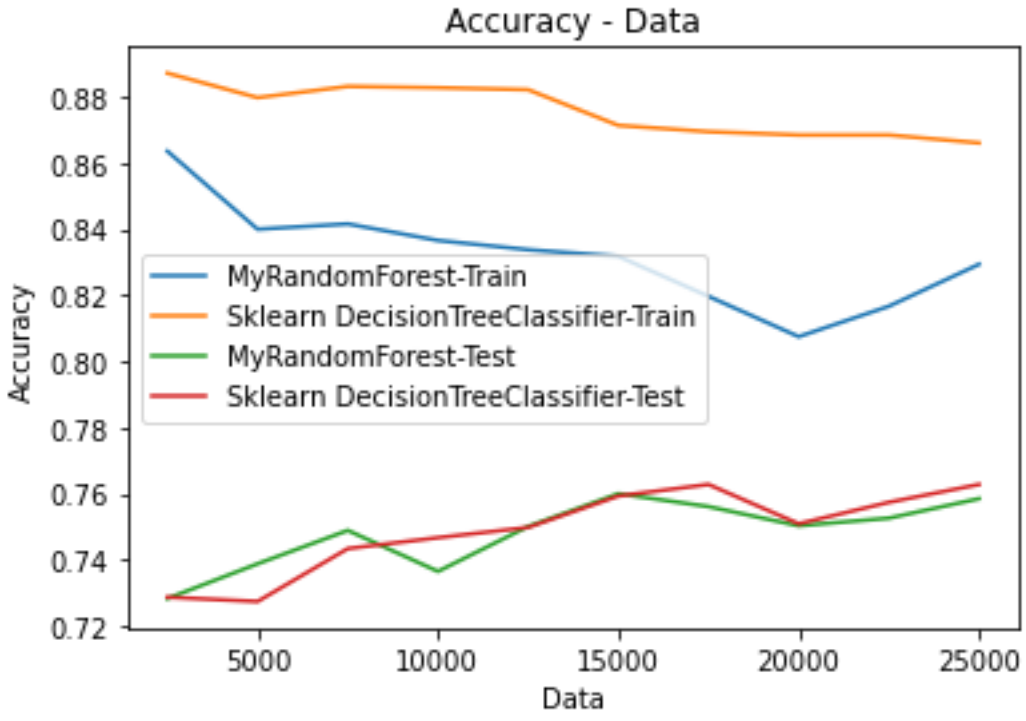
Α/Α	1	2	3	4	5	6	7	8	9	10
Accuracy	0.8636	0.8400	0.8416	0.8366	0.8338	0.8318	0.8197	0.8075	0.8167	0.8294
Precision	0.8270	0.7967	0.8042	0.8027	0.7999	0.8097	0.8027	0.7806	0.7926	0.8013
Recall	0.9282	0.9207	0.9083	0.8971	0.8944	0.8694	0.8505	0.8570	0.8592	0.8760
F1	0.8747	0.8542	0.8531	0.8473	0.8445	0.8385	0.8259	0.8170	0.8246	0.8370

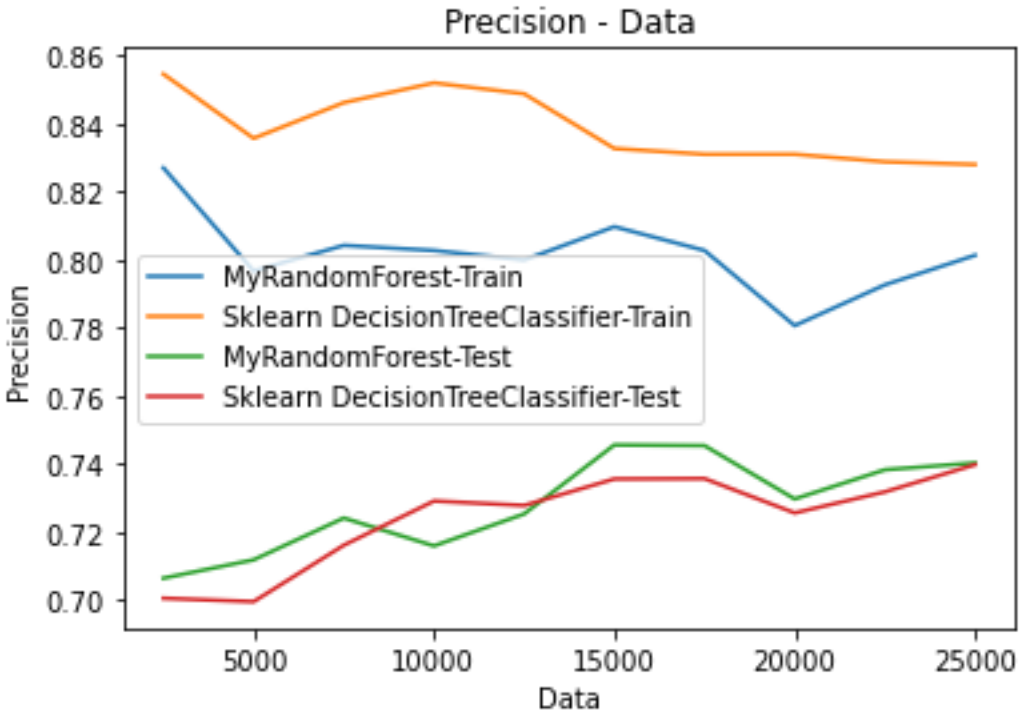
Sklearn RandomForestClassifier (με τα test data)

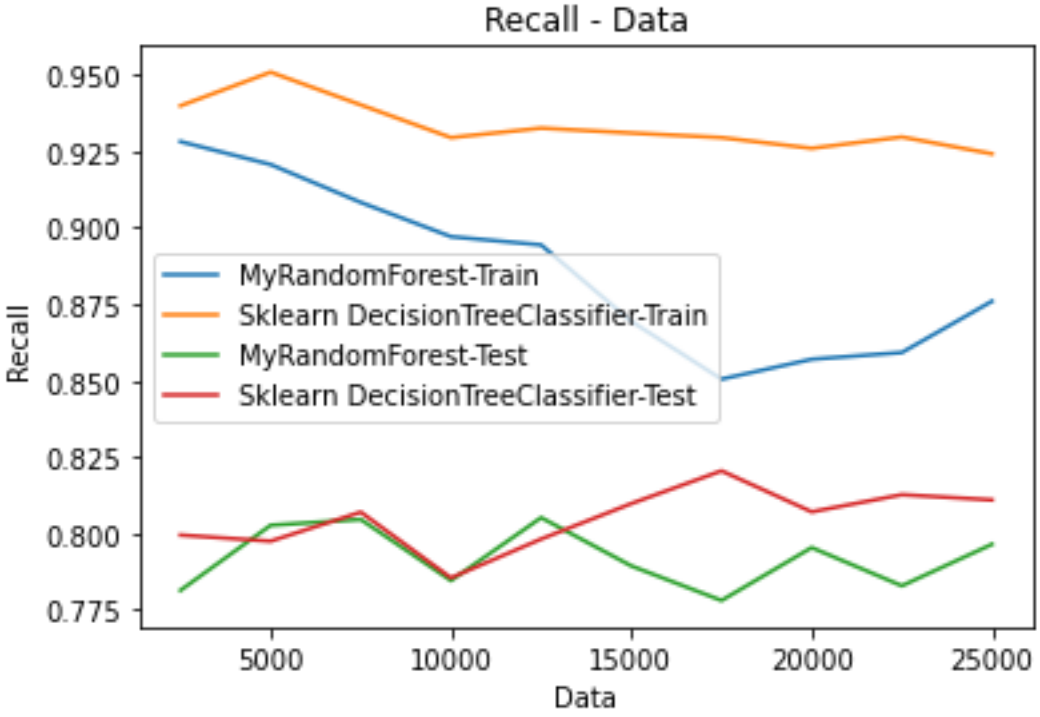
A/A	1	2	3	4	5	6	7	8	9	10
Accuracy	0.7287	0.7274	0.7434	0.7467	0.7498	0.7593	0.7628	0.7509	0.7574	0.7628
Precision	0.7004	0.6994	0.7160	0.7290	0.7277	0.7355	0.7356	0.7255	0.7317	0.7397
Recall	0.7995	0.7975	0.8070	0.7855	0.7982	0.8098	0.8205	0.8071	0.8127	0.8110
F1	0.7467	0.7452	0.7588	0.7562	0.7614	0.7709	0.7757	0.7641	0.7701	0.7737

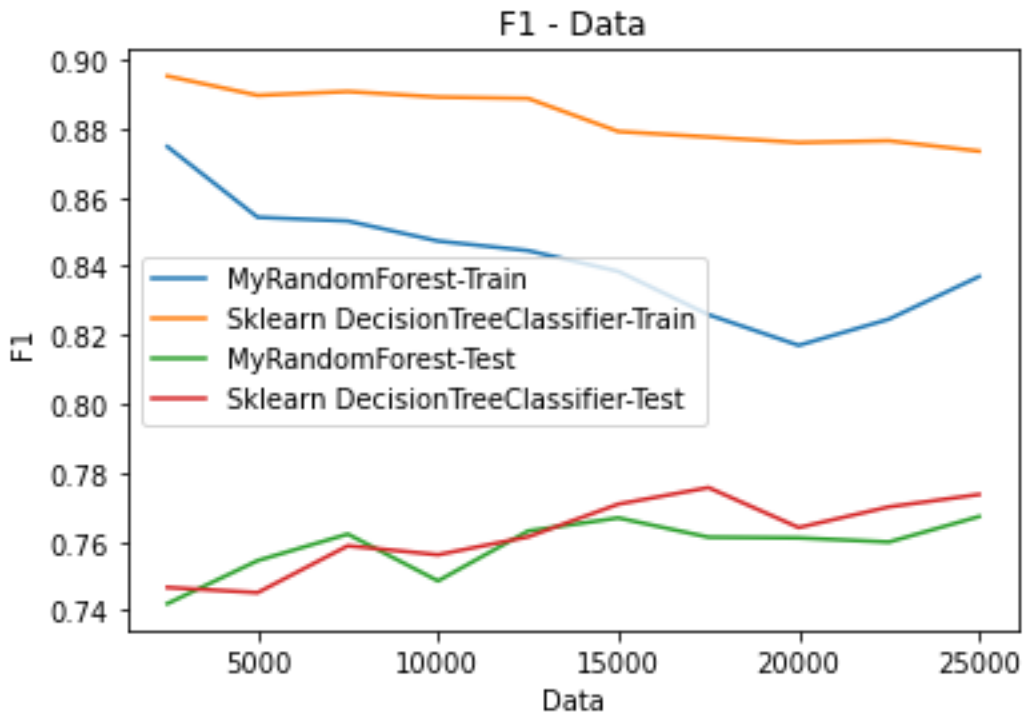
Sklearn RandomForestClassifier (με τα train data)

A/A	1	2	3	4	5	6	7	8	9	10
Accuracy	0.8872	0.8798	0.8832	0.8828	0.8822	0.8714	0.8695	0.8685	0.8685	0.8661
Precision	0.8546	0.8357	0.8462	0.8520	0.8488	0.8327	0.8310	0.8310	0.8288	0.8280
Recall	0.9399	0.9509	0.9402	0.9295	0.9326	0.9310	0.9295	0.9260	0.9297	0.9242
F1	0.8952	0.8896	0.8907	0.8891	0.8887	0.8791	0.8775	0.8759	0.8764	0.8734




Precision - Data
Precision
Data
MyRandomForest-Train
Sklearn DecisionTreeClassifier-Train
MyRandomForest-Test
Sklearn DecisionTreeClassifier-Test
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
5000
10000
15000
20000
25000


Recall - Data
Recall
Data
MyRandomForest-Train
Sklearn DecisionTreeClassifier-Train
MyRandomForest-Test
Sklearn DecisionTreeClassifier-Test
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.775
5000
10000
15000
20000
25000


F1 - Data
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
F1
MyRandomForest-Train
Sklearn DecisionTreeClassifier-Train
MyRandomForest-Test
Sklearn DecisionTreeClassifier-Test
5000
10000
15000
20000
25000
Data