ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών

Εργασία 4:
«Επίλυση Προβλήματος Ταξινόμησης με χρήση μοντέλων TSK»

Υπολογιστική Νοημοσύνη

Διδάσκων: Θεοχάρης Ιωάννης

Γιάννης Τατσόγλου, 9568

Σεπτέμβριος, 2023

Περιεχόμενα

Μέρος 1 - Εφαρμογή σε Απλό Dataset

Για το πρώτο κομμάτι της εργασίας καλούμαστε να εκπαιδεύσουμε 4 μοντέλα, με διαφορετικά χαρακτηριστικά όπως αναγράφονται στον πίνακα που ακολουθεί.

Μοντέλο	Τύπος Μοντέλου	Ακτίνα Clusters
model_1	Class Dependent	0.2
model_2	Class Independent	0.2
model_3	Class Dependent	0.9
Model_4	Class Independent	0.9

Το dataset που θα χρησιμοποιηθεί είναι το Haberman's Survival, το οποίο περιέχει 306 δείγματα από 3 χαρακτηριστικά και 1 έξοδο που παίρνει τις τιμές 1 ή 2.

Αρχικά διαχωρίζεται το σύνολο των δεδομένων σε υποσύνολα μεγέθους 60%, 20% και 20% του συνολικού, τα οποία αποτελούν τα σύνολα εκπαίδευσης (training), επαλήθευσης (validation) και ελέγχου (test).

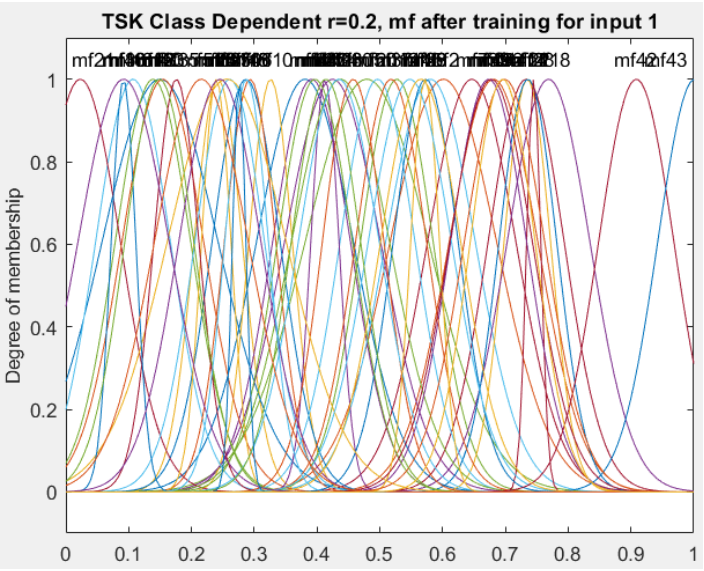
Η ακτίνα των clusters επιλέγεται αυθαίρετα σε ακραίες τιμές, έτσι ώστε ο αριθμός των κανόνων ανάμεσα στα μοντέλα να παρουσιάζει σημαντική διακύμανση.

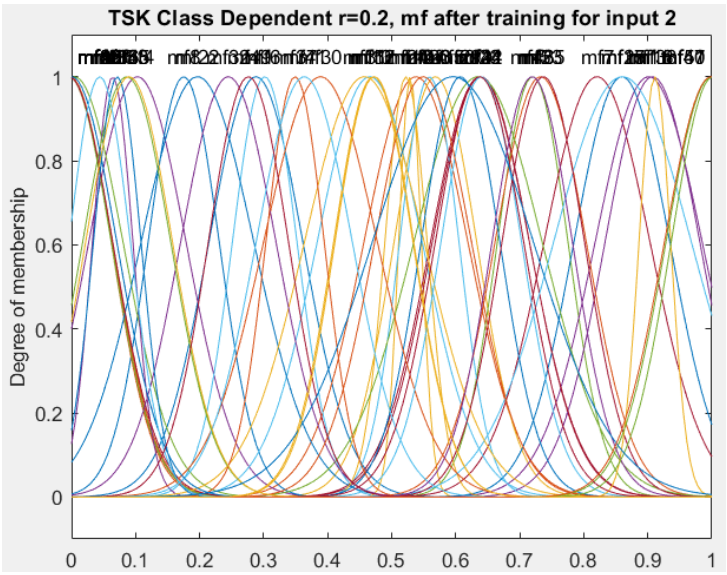
Έπειτα προχωρούμε στην υλοποίηση των μοντέλων όπως αυτά περιγράφονται στον πίνακα παραπάνω, και την εκπαίδευσή τους.

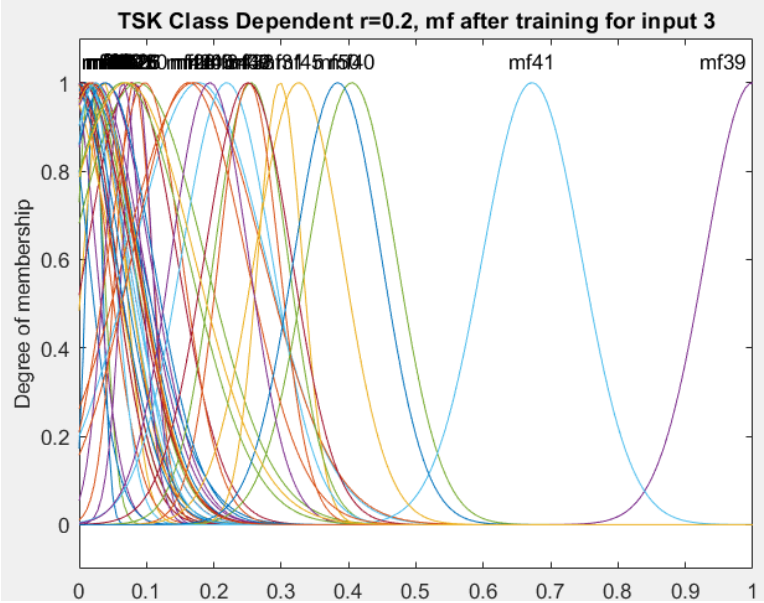
Αξιολόγηση Μοντέλων

Μοντέλο 1

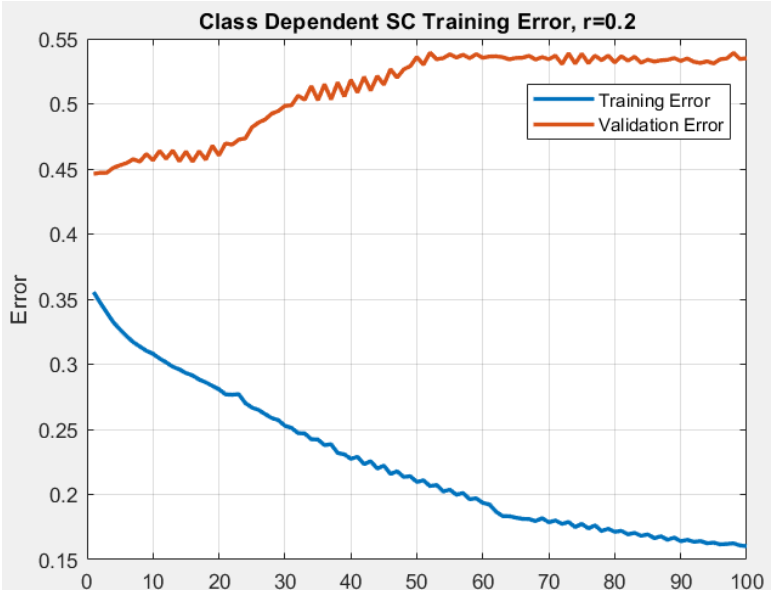
Συναρτήσεις συμμετοχής







Καμπύλη Μάθησης

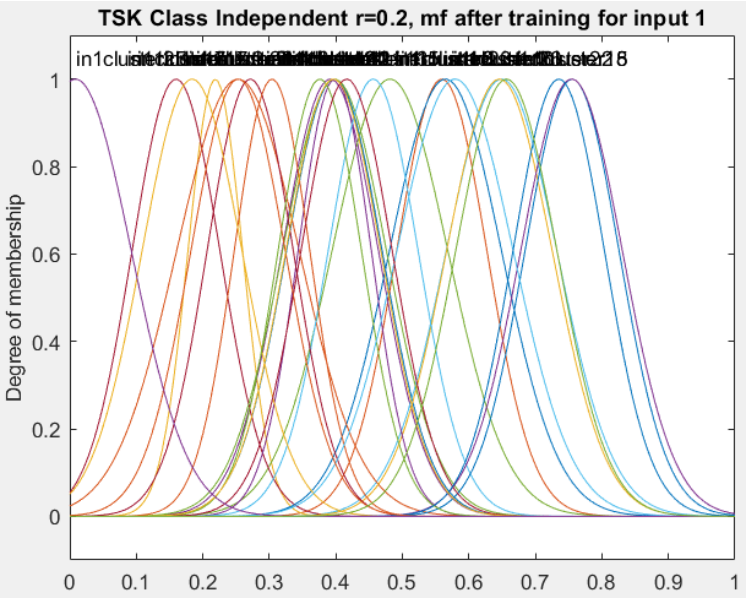


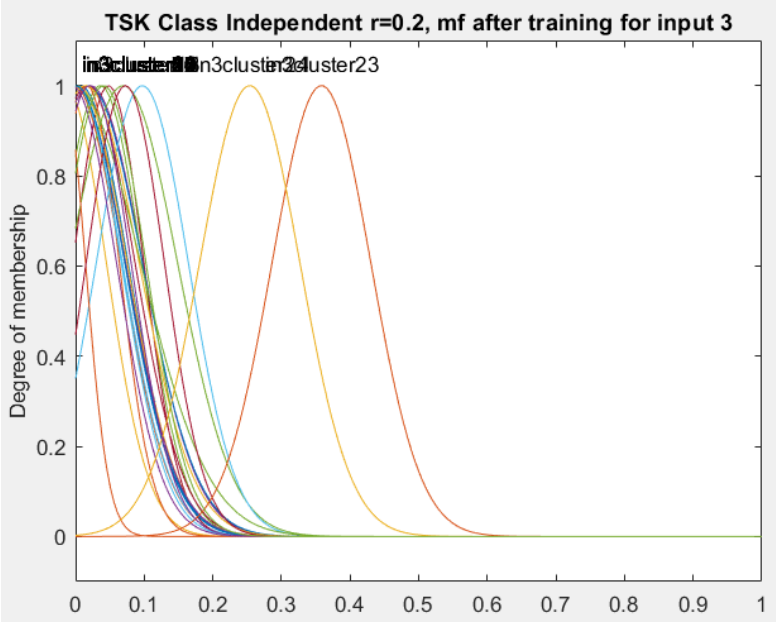
Πίνακας Σφαλμάτων Ταξινόμησης

Actual / Predicted	**1**	**2**
1	33	8
2	11	9

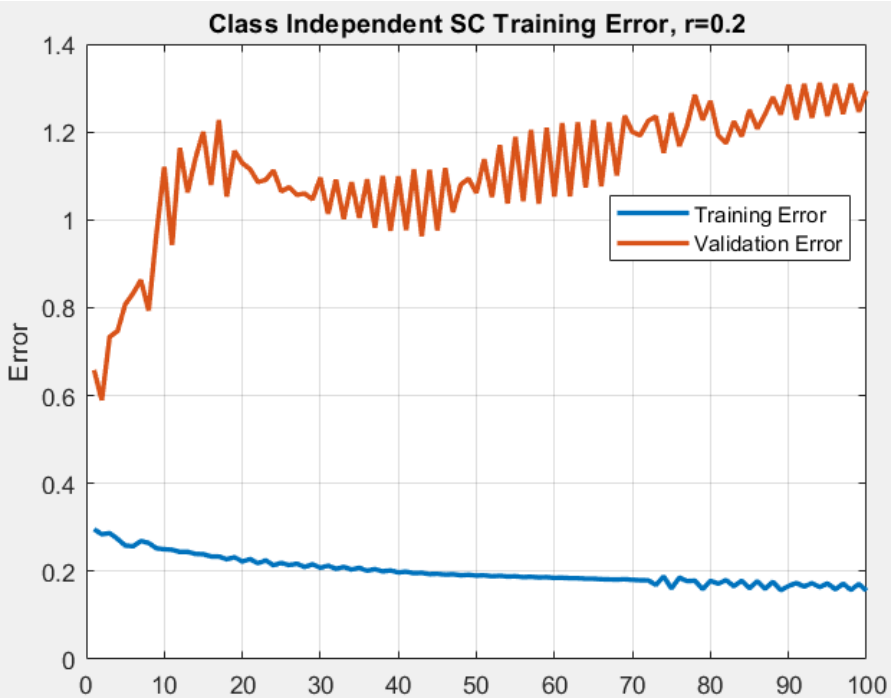
Μοντέλο 2

Συναρτήσεις Συμμετοχής





Καμπύλη Μάθησης

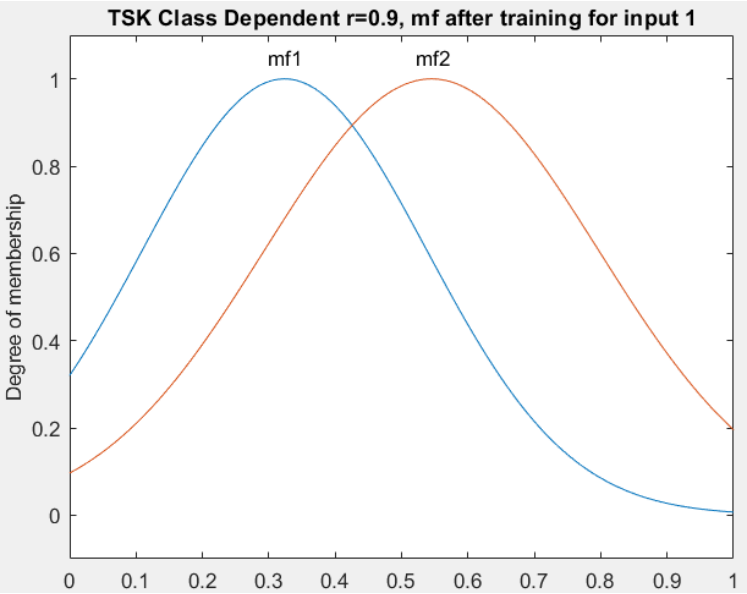


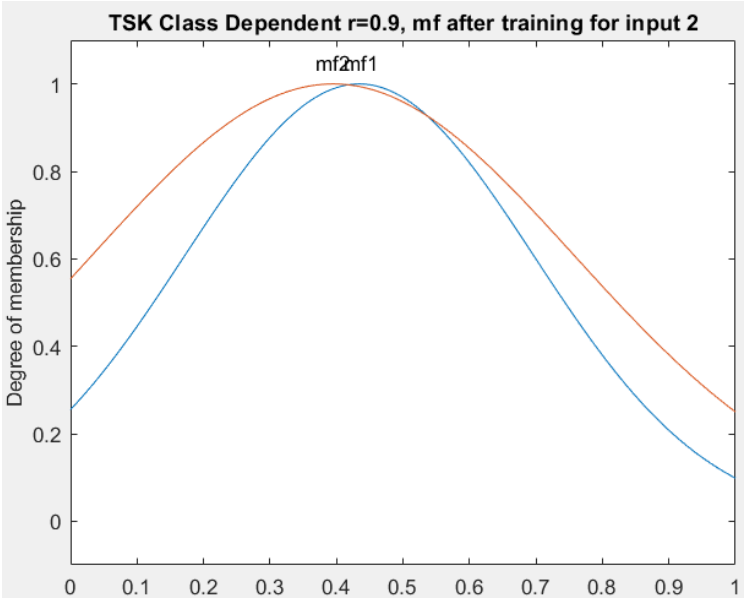
Πίνακας Σφαλμάτων Ταξινόμησης

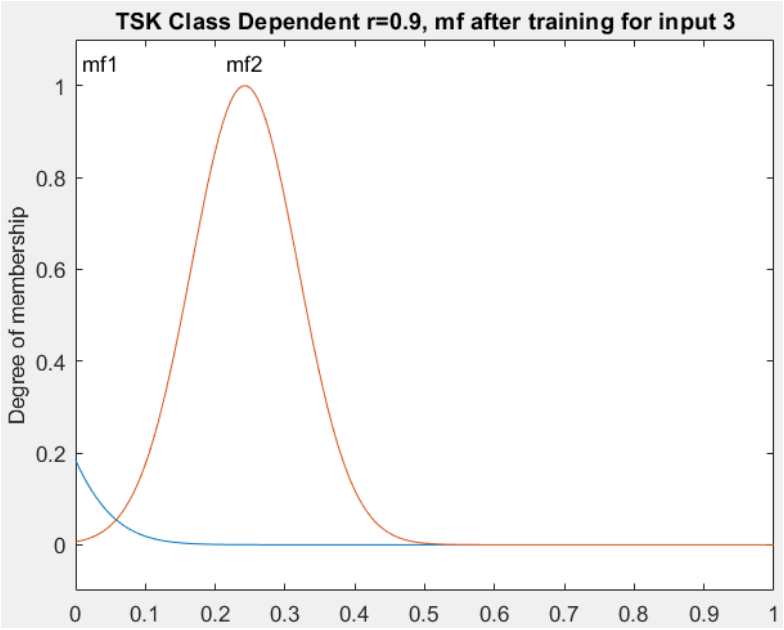
Actual / Predicted	**1**	**2**
1	27	14
2	10	10

Μοντέλο 3

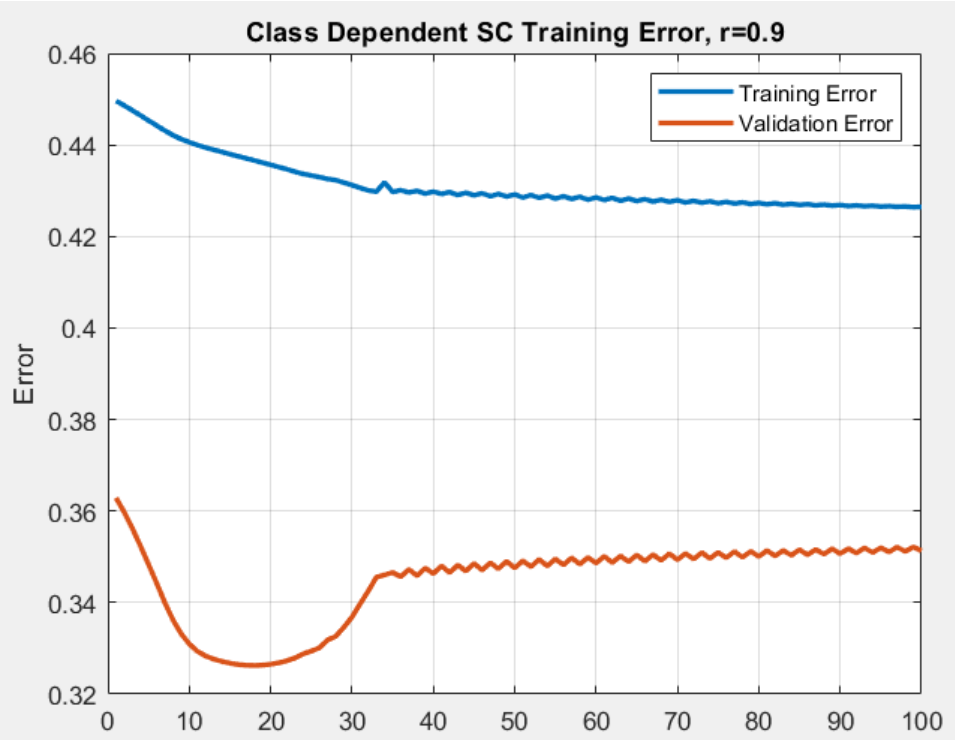
Συναρτήσεις Συμμετοχής







Καμπύλη Μάθησης

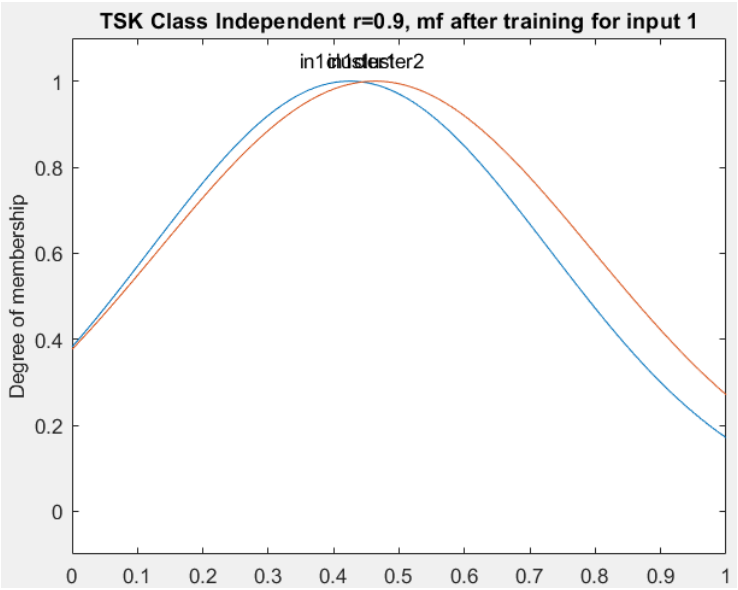


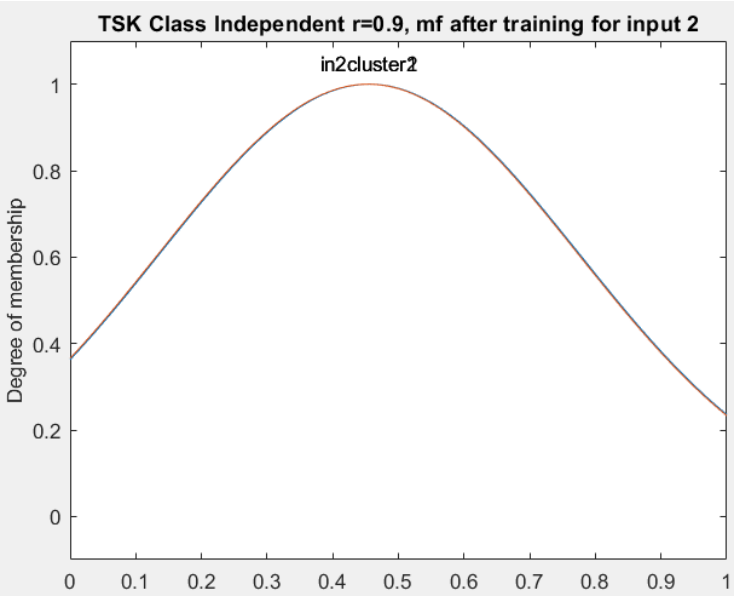
Πίνακας Σφαλμάτων Ταξινόμησης

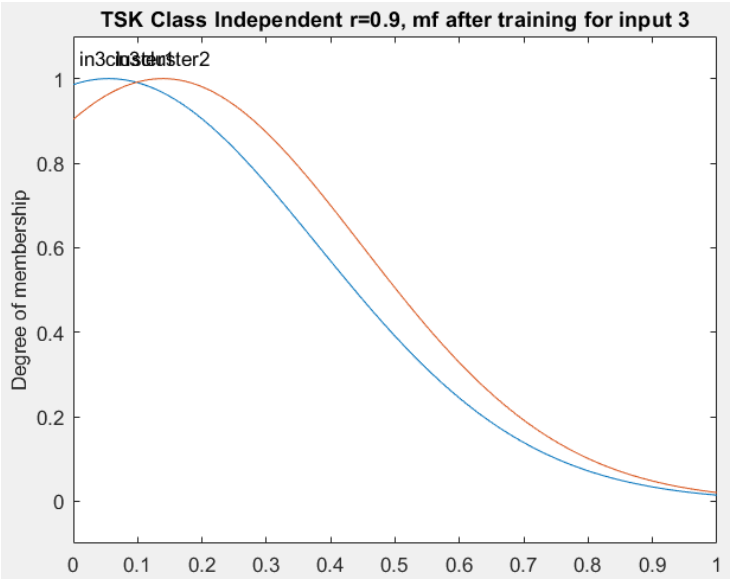
Actual / Predicted	**1**	**2**
1	39	2
2	16	4

Μοντέλο 4

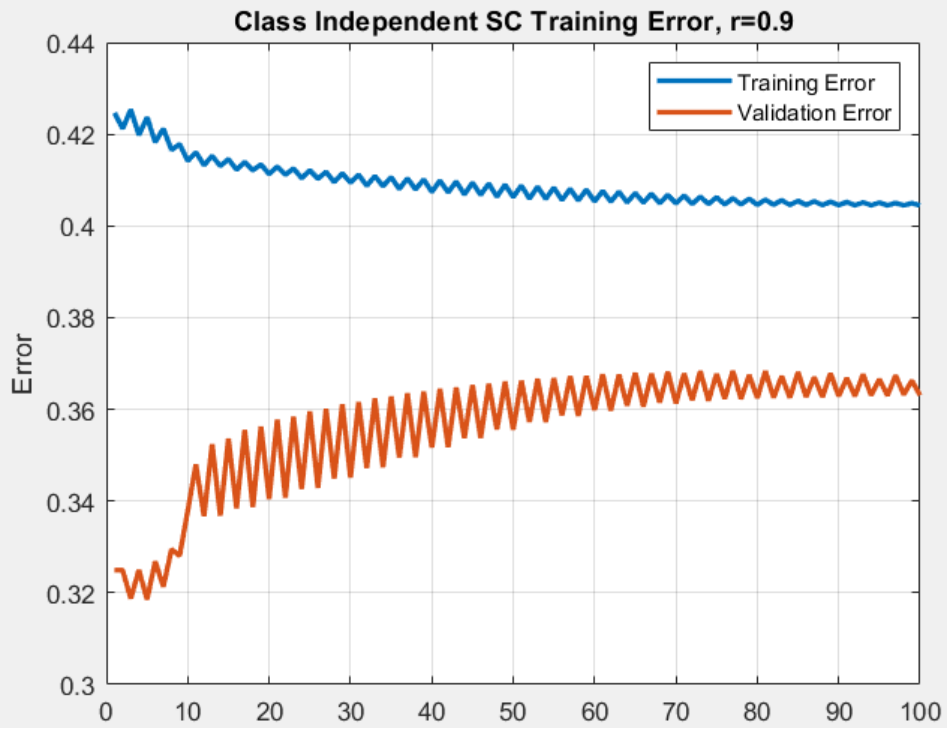
Συναρτήσεις Συμμετοχής







Καμπύλη Μάθησης



Πίνακας Σφαλμάτων Ταξινόμησης

Actual / Predicted	**1**	**2**
1	38	3
2	16	4

Μετρικές

Στον πίνακα που ακολουθεί αποτυπώνονται οι δείκτες αξιολόγησης για κάθε μοντέλο. Πιο συγκεκριμένα ο δείκτης **Overall Accuracy (OA)**, που απεικονίζει την συνολική ακρίβεια του μοντέλου, οι δείκτες **Producers και Users Accuracy (PA, UA)**, που απεικονίζουν την ακρίβεια του μοντέλου όσον αφορά τις κλάσεις παραγωγού και χρήστη, ο δείκτης $\widehat{K}$ όπου είναι ένας εκτιμητής της στατιστικής παραμέτρου.

Επιπλέον στον πίνακα αναγράφονται και οι κανόνες που παρήχθησαν από κάθε μοντέλο.

Model	OA	PA		UA		$\widehat{K}$	Rules
model_1	0.6885	0.8049	0.45	**0.75**	0.5294	**0.2651**	52
model_2	0.6065	0.6585	**0.5**	0.7297	0.4167	0.1508	26
model_3	**0.7049**	**0.9512**	0.2	0.7091	**0.6667**	0.1842	2
model_4	0.6985	0.9268	0.2	0.7037	0.5714	0.1521	2

Σχολιασμός Αποτελεσμάτων

Παρατηρώντας τον πίνακα παραπάνω, φαίνεται ότι το μοντέλο 3 έχει την καλύτερη απόδοση με βάση την τιμή των μετρικών, αφού οι τιμές τους υπερτερούν από αυτές των άλλων μοντέλων.

Δεν μπορούμε να βγάλουμε κάποιο συμπέρασμα για το πως επηρεάζεται η απόδοση του μοντέλου σύμφωνα με το πλήθος των κανόνων, αφού δεν είναι εμφανής κάποια τέτοιου είδους συσχέτιση με βάση τις τιμές του παραπάνω πίνακα.

Συγκρίνοντας τα μοντέλα βάσει του τύπου τους, Class Dependent (1, 3) και Class Independent (2, 4), βλέπουμε πως η μέθοδος Dependent Clustering παράγει περισσότερους κανόνες, και η απόδοσή της είναι εξίσου καλή η καλύτερη της απόδοσης των Class Independent μοντέλων.

Για την επίδραση της ακτίνας clustering, παρατηρούμε ότι για μικρές τιμές, προκύπτουν πολύ περισσότεροι κανόνες και συνεπώς αυξάνεται ο αριθμός των συναρτήσεων συμμετοχής. Παρατηρούμε ότι σε αυτά τα μοντέλα (1, 2), ο μεγάλος αριθμός συναρτήσεων συμμετοχής οδηγεί και σε υψηλότερο βαθμό επικάλυψης μεταξύ τους, ταυτόχρονα έχουν τον μικρότερο OA από τα υπόλοιπα μοντέλα. Συμπεραίνουμε πως η επικάλυψη των ασαφών συνόλων επιδρά αρνητικά στο μοντέλο.

Μία μέθοδος που θα μπορούσε να εφαρμοσθεί ώστε να βελτιωθεί το τμήμα υπόθεσης, θα ήταν μετά την υλοποίηση και το training του μοντέλου, να αφαιρούνται οι συναρτήσεις συμμετοχής που έχουν μεγάλη επικάλυψη, ώστε το μοντέλο να γίνεται πιο «απλό» και ταυτόχρονα να έχει λιγότερους και σαφέστερους κανόνες.

Μέρος 2 – Εφαρμογή σε Δεδομένα Υψηλής Διαστασιμότητας

Για το δεύτερο μέρος της εργασίας χρησιμοποιείται το dataset «epileptic_seizure_data.csv» το οποίο έχει μεγάλο βαθμό διαστασιμότητας, αποτελείται από 11500 δείγματα το καθένα από τα οποία περιγράφεται από 178 features και μια έξοδο που χαρακτηρίζει την κλάση του κάθε δείγματος.

Ο μεγάλος αριθμός μεταβλητών καθιστά αδύνατη την προσέγγιση του πρώτου μέρους καθώς ο αριθμός των κανόνων που θα αυξανόταν σε πάρα πολύ μεγάλο βαθμό.

Δημιουργία Δοκιμαστικών Μοντέλων

Αρχικά διαχωρίζεται το σύνολο των δεδομένων σε υποσύνολα μεγέθους 60%, 20% και 20% του συνολικού, τα οποία αποτελούν τα σύνολα εκπαίδευσης (training), επαλήθευσης (validation) και ελέγχου (test), όπως και στο πρώτο μέρος.

Θα χρησιμοποιηθεί η μέθοδος της επιλογής χαρακτηριστικών και διαμέρισης του χώρου, επομένως θα εισάγουμε δυο ελεύθερες μεταβλητές στο πρόβλημα (αριθμός χαρακτηριστικών προς επιλογή και αριθμός ομάδων που δημιουργούνται). Οι μεταβλητές θα προσεγγιστούν με την μέθοδο αναζήτησης πλέγματος ώστε να βρεθεί η βέλτιστη τιμή για αυτές και να εκπαιδευτεί το τελικό μοντέλο.

Οι τιμές που εξετάστηκαν παρατίθενται στους πίνακες παρακάτω.

- features_number = [5, 7, 9, 11]
- cluster_radius = [0.2, 0.4, 0.6, 0.8, 1]

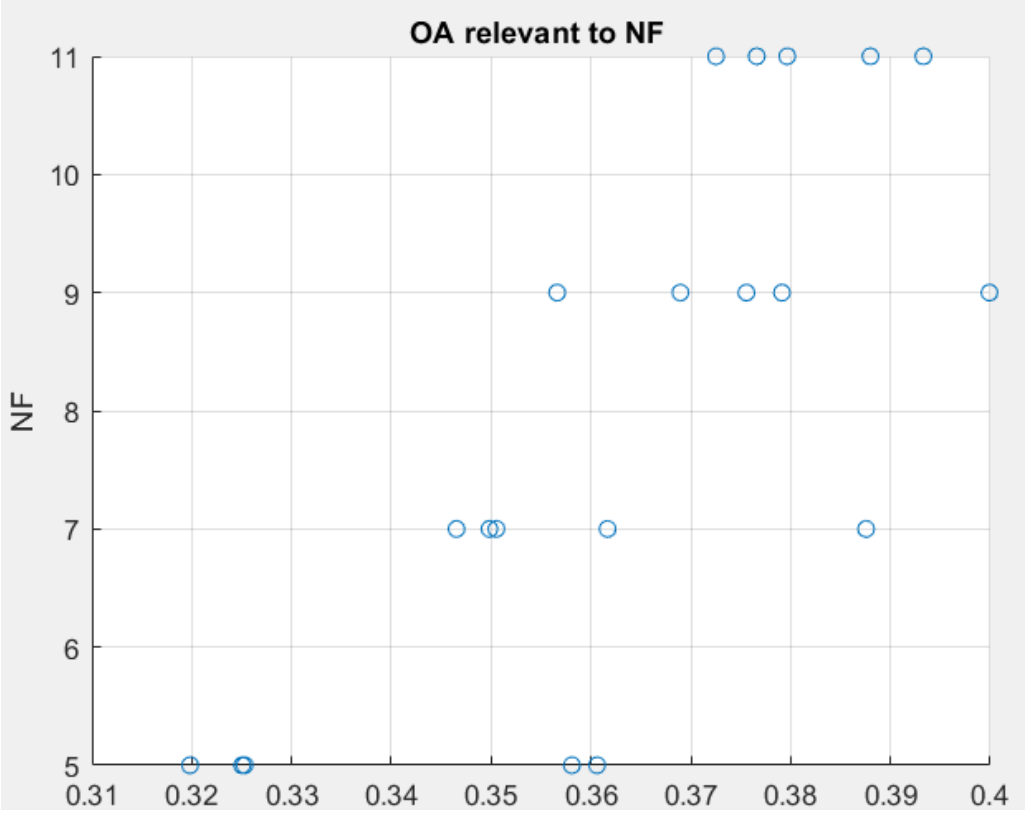
Χρησιμοποιήθηκε η μετρική Overall Accuracy (OA) για την αξιολόγηση των δοκιμαστικών μοντέλων, οι τιμές της μετρικής για τους διάφορους συνδυασμούς φαίνεται στον πίνακα παρακάτω.

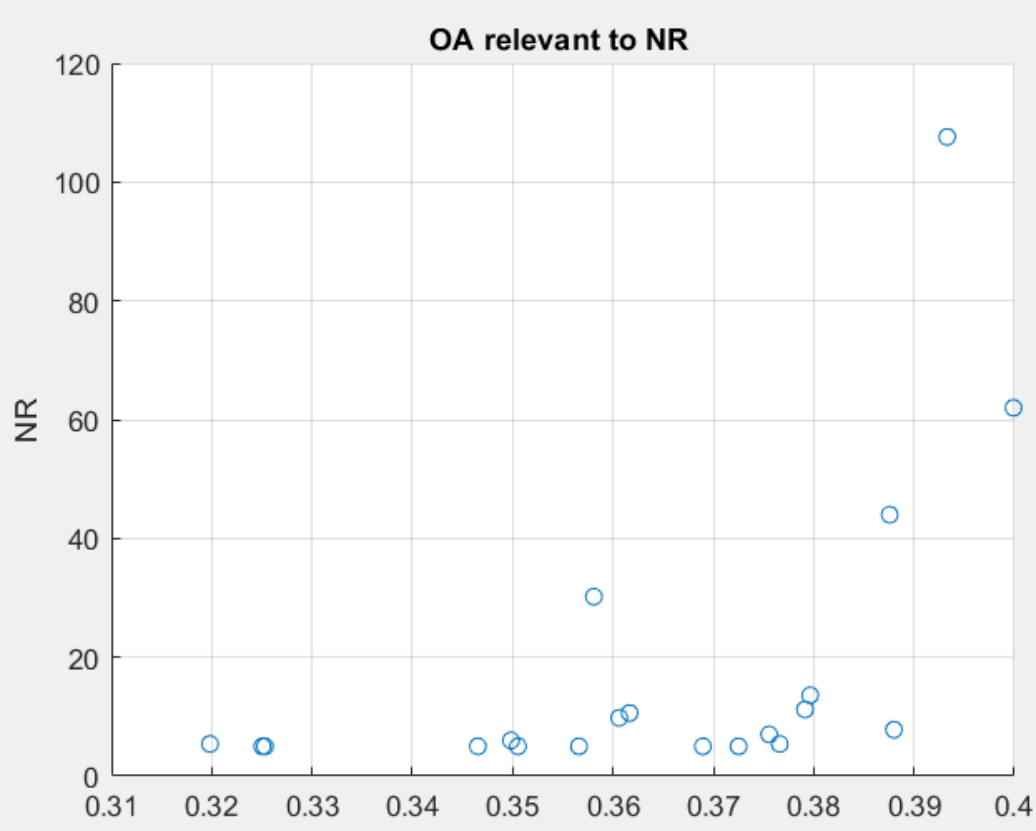
NF \ r	0.2	0.4	0.6	0.8	1
5	0.3597	0.3502	0.3239	0.3268	0.3159
7	0.3861	0.3550	0.3567	0.3540	0.3385
9	0.3966	0.3689	0.3870	0.3620	0.3515
11	0.3986	0.3717	0.3792	0.3807	0.3677

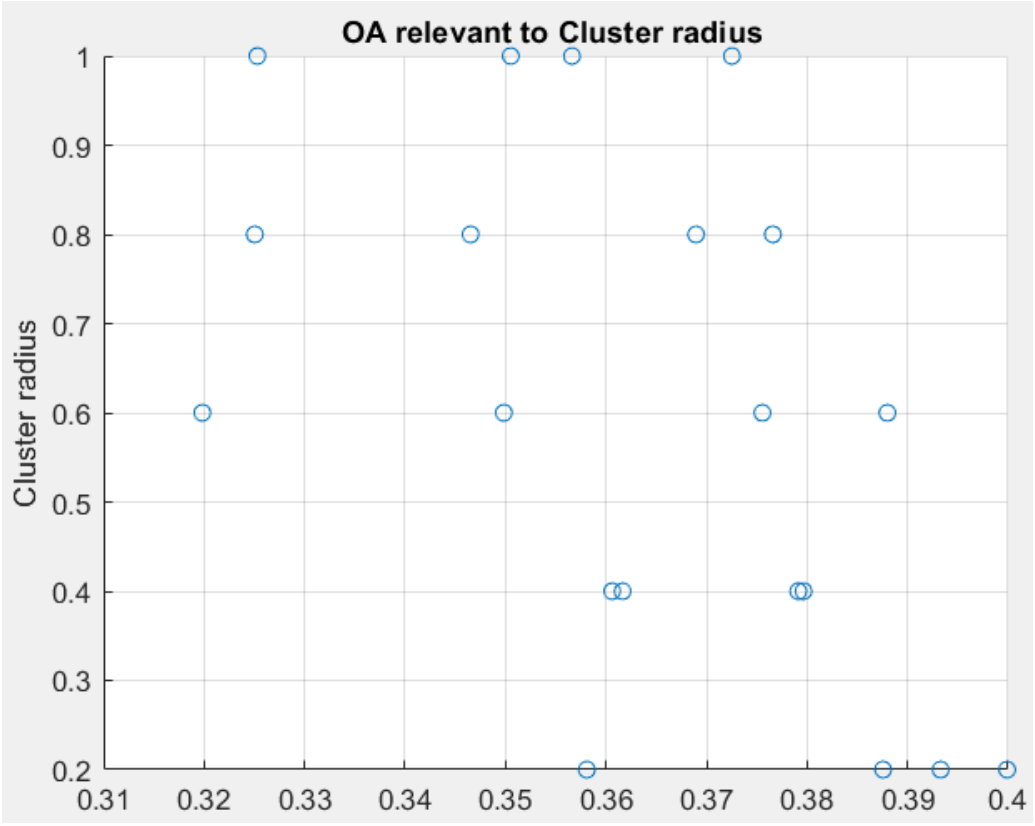
Ο πίνακας παρακάτω παρουσιάζει τον αριθμό των κανόνων που προκύπτουν μετά το cross-validation για κάθε συνδυασμό.

NF \ r	0.2	0.4	0.6	0.8	1
5	31	10	5	5	5
7	43	11	6	5	5
9	57	12	7	5	5
11	108	13	7	5	5

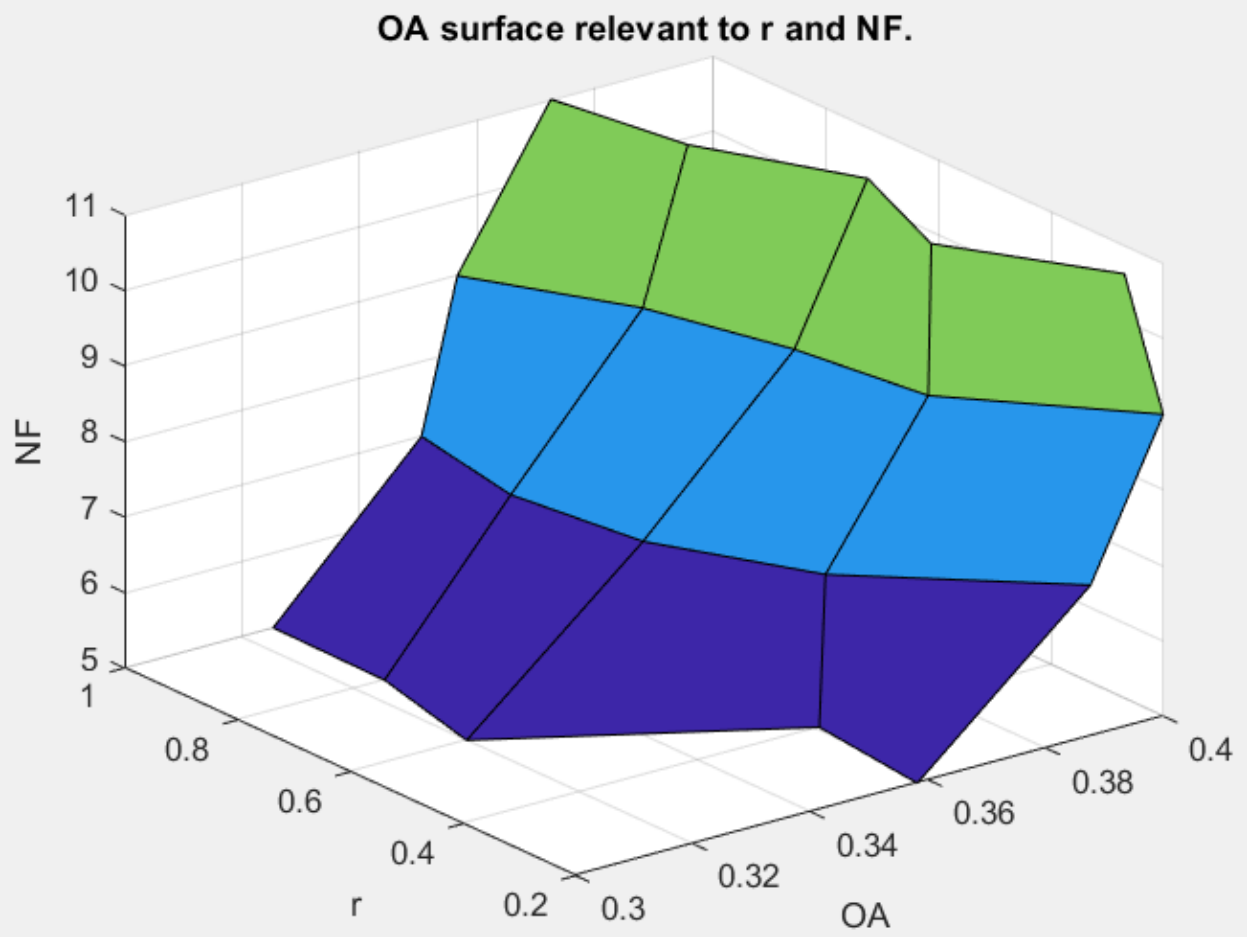
Επιπλέον παρουσιάζονται τα διαγράμματα OA – NF, OA – Cluster radius kai OA – NR.




OA relevant to NR
NR
120
100
80
60
40
20
0
0.31
0.32
0.33
0.34
0.35
0.36
0.37
0.38
0.39
0.4


OA relevant to Cluster radius
Cluster radius
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.31
0.32
0.33
0.34
0.35
0.36
0.37
0.38
0.39
0.4

Τέλος παρουσιάζεται η επιφάνεια OA – NF – Cluster radius



Σχολιασμός

Παρατηρούμε, από τα παραπάνω διαγράμματα και πίνακες, ότι η αύξηση του NF και μείωση της cluster radius οδηγεί σε αύξηση του OA και συνεπώς καλύτερο μοντέλο.

Δεν μπορούμε να γενικεύσουμε αυτό το συμπέρασμα όμως καθώς σε πολλές περιπτώσεις μεγάλος αριθμός NF και μικρή cluster radius, μπορεί να προκαλέσει overfitting. Να μάθει δηλαδή το μοντέλο καλύτερα το dataset, αλλά να χάσει την δυνατότητα της γενίκευσης.

Από τα διαγράμματα δεν μπορούμε να εξάγουμε κάποια συσχέτιση ανάμεσα στον NR και το OA. Παρόλο που το καλύτερο OA παρουσιάζεται για το μοντέλο με τους περισσότερους κανόνες, βλέπουμε ότι υπάρχουν μοντέλα με λιγότερους κανόνες που έχουν μεγαλύτερο OA από μοντέλα με περισσότερους.

Τέλος, για την πειραματική διάταξη που σχεδιάσαμε βρίσκουμε ότι ο βέλτιστος συνδυασμός τιμών είναι NF = 11 και Cluster radius = 0.2, ο οποίος δημιουργεί μοντέλο με 108 κανόνες.

Βέλτιστο Μοντέλο

Το βέλτιστο μοντέλο προέκυψε όπως αναφέρθηκε παραπάνω για NF = 11 και cluster radius = 0.2

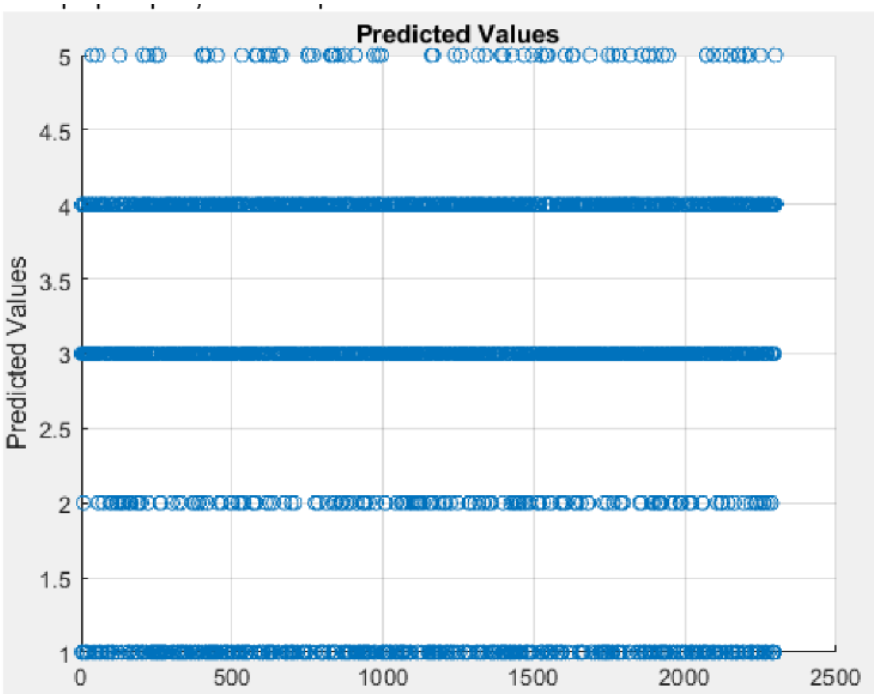
Μετά την εκπαίδευσή και αξιολόγησή του προκύπτει ο παρακάτω πίνακας σφαλμάτων ταξινόμησης.

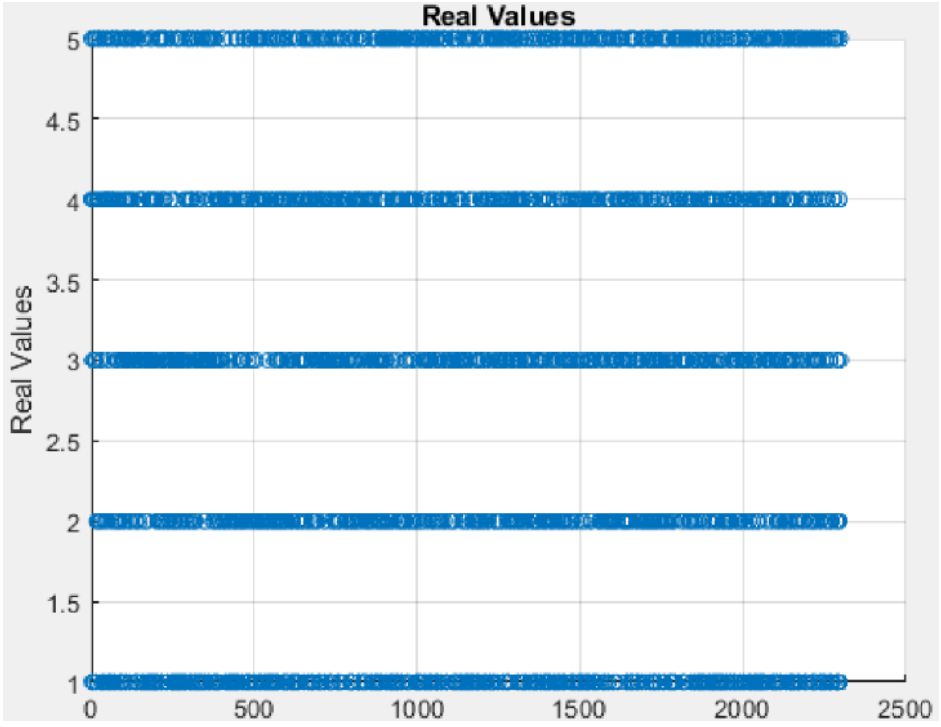
	Predicted Class				
True Class	343	66	38	12	0
	24	47	255	130	4
	4	34	269	148	5
	2	31	184	227	16
	1	10	165	256	29

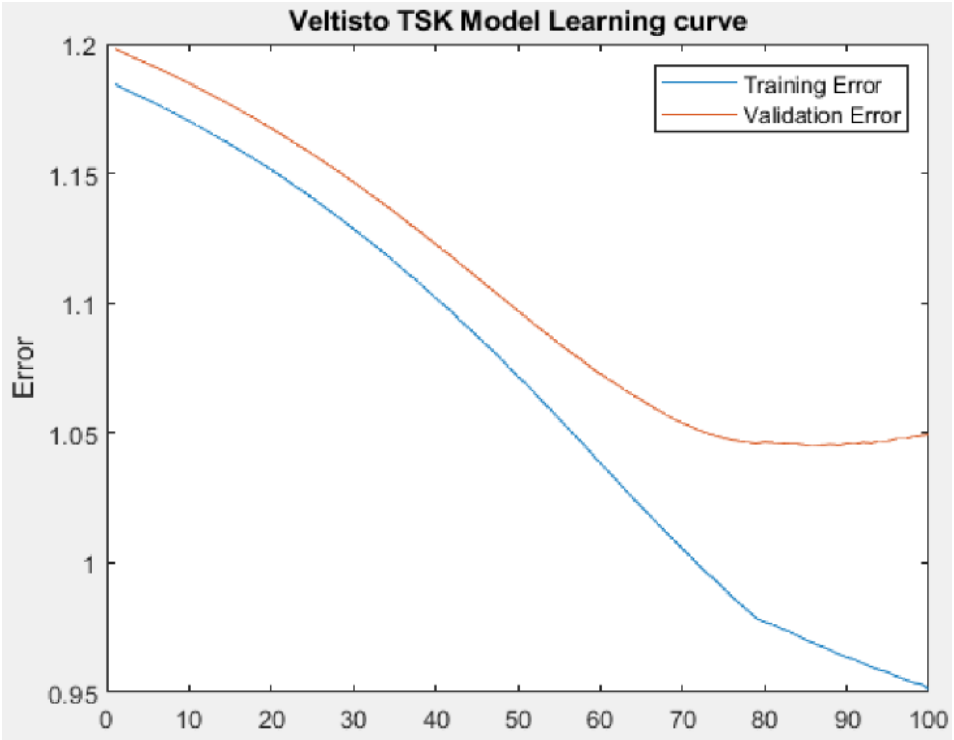
Από τον πίνακα, εξάγουμε τις μετρικές για το μοντέλο.

OA	**PA_1**	**PA_2**	**PA_3**	**PA_4**	**PA_5**	**UA_1**	**UA_2**	**UA_3**	**UA_4**	**UA_5**	**K**	**Rules**
0.3979	0.7457	0.1026	0.5848	0.4939	0.0626	0.9201	0.2481	0.2956	0.2937	0.5545	0.2474	108

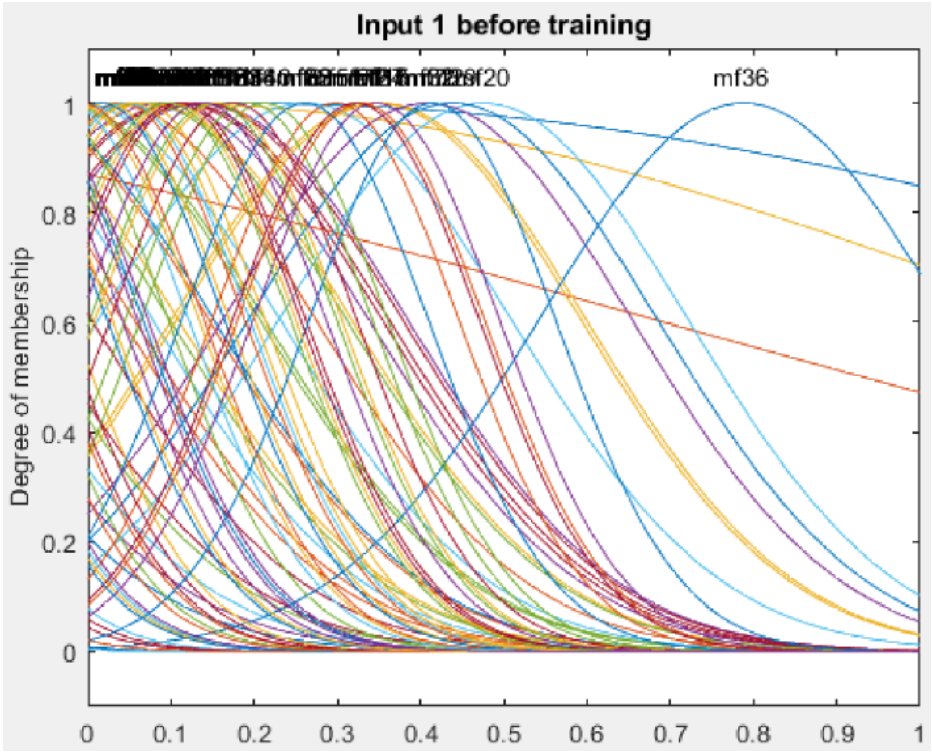
Παρατίθενται στην συνέχεια τα διαγράμματα εκπαίδευσης και προβλέψεων του μοντέλου.

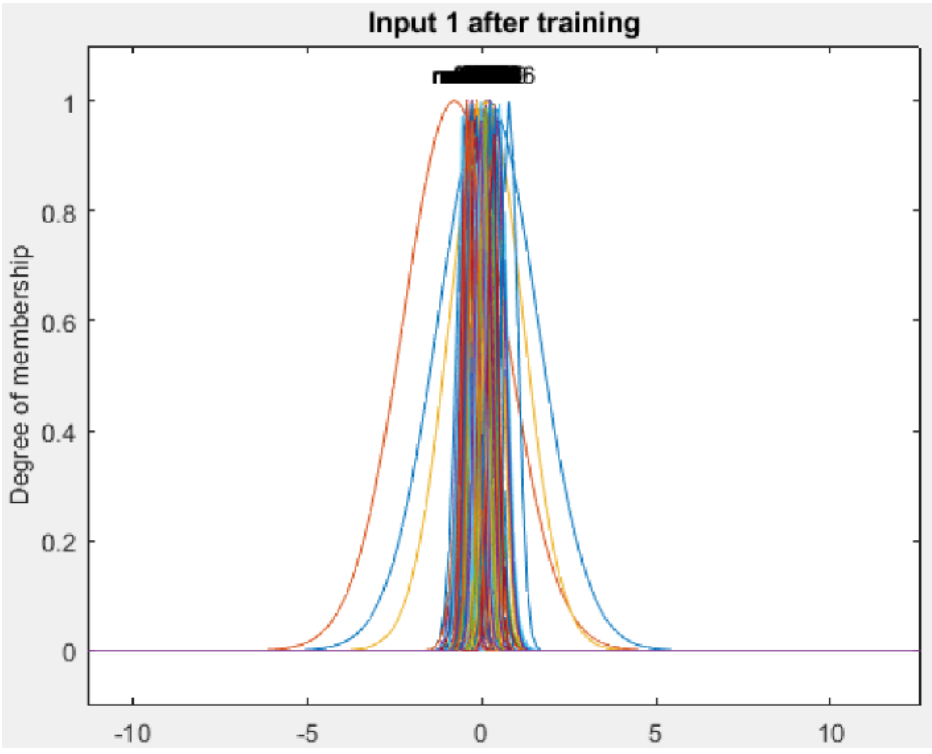


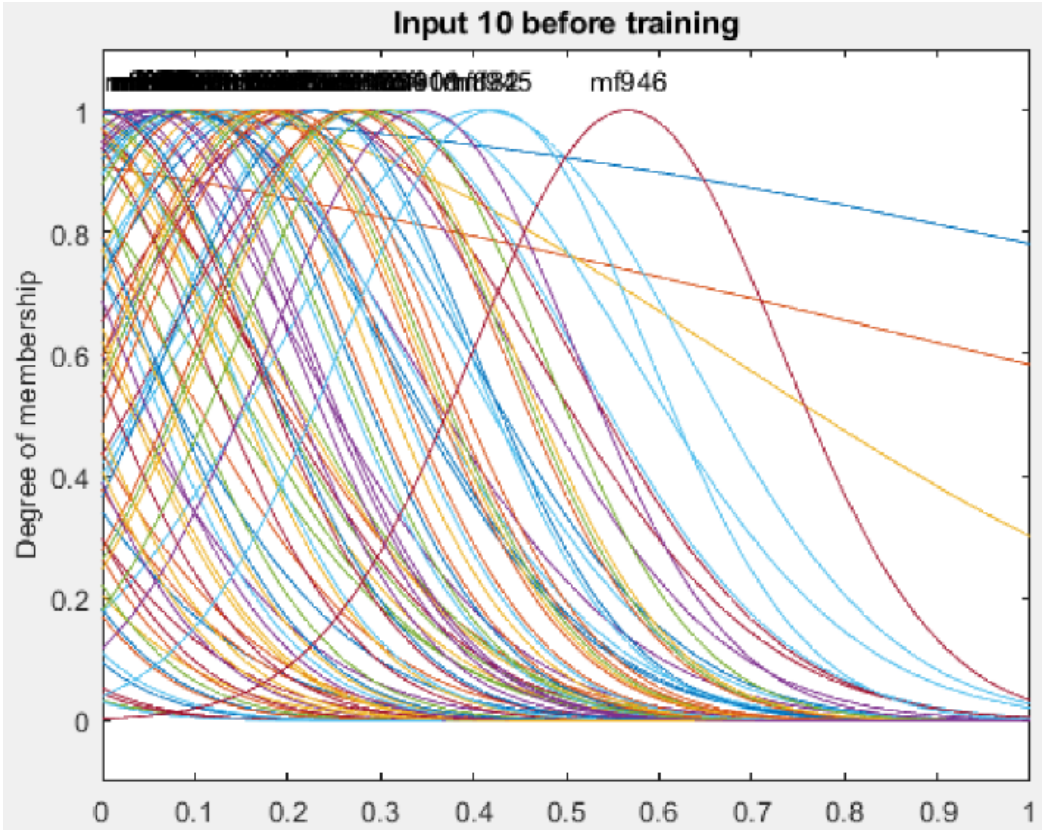

Real Values
Real Values
5
4.5
4
3.5
3
2.5
2
1.5
1
0
500
1000
1500
2000
2500

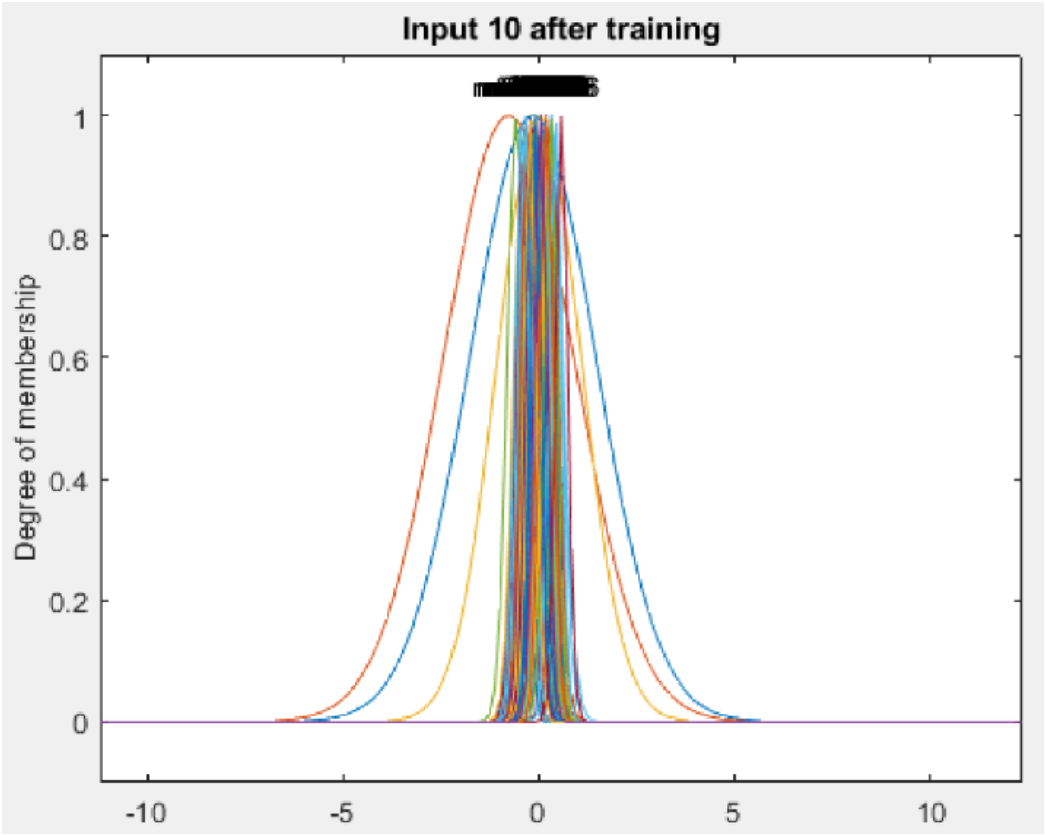

Veltisto TSK Model Learning curve
Training Error
Validation Error
Error
1.2
1.15
1.1
1.05
1
0.95
0
10
20
30
40
50
60
70
80
90
100

Και ενδεικτικά παρουσιάζονται δύο ασαφή σύνολα πριν και μετά την εκπαίδευση.






Input 10 before training
mf946
Degree of membership
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Input 10 after training
Degree of membership
1
0.8
0.6
0.4
0.2
0
-10
-5
0
5
10

Σχολιασμός

Από τις τιμές των μετρικών που προέκυψαν, μπορούμε να εξάγουμε τα εξής συμπεράσματα για το μοντέλο.

Μπορεί να προβλέπει αρκετά καλά τις κλάσεις 1 και 3, λιγότερο καλά την κλάση 4, και όχι καλά τις κλάσεις 2 και 5.

Σύμφωνα με τις τιμές των UA, καλή ακρίβεια χρήστη έχει η κλάση 1 με πολύ υψηλό ποσοστό, το οποίο πέφτει ραγδαία για τις υπόλοιπες κλάσεις.

Τα δυο παραπάνω σημεία μας οδηγούν να σκεφτούμε ότι ίσως το dataset να είναι Imbalanced σχετικά με το πλήθος των δειγμάτων που περιέχει σε κάθε κλάση.

Επίσης παρατηρούμε ότι οι συναρτήσεις συμμετοχής του εκπαιδευμένου μοντέλου παρουσιάζουν πολύ μεγάλο βαθμό επικάλυψης, που σημαίνει ότι κάποιες λεκτικές μεταβλητές ίσως να είναι περιττές.

Τέλος, ο αριθμός κανόνων είναι σε διαχειρίσιμα επίπεδα, ενώ αν είχαμε επιλέξει την μέθοδο του Grid partitioning, με δυο η τρία ασαφή σύνολα ανά είσοδο, θα είχαμε 2^{11} ή 3^{11} κανόνες αντίστοιχα. Γεγονός που θα έκανε τους υπολογισμούς πολύπλοκους και ασύμφορους.