Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρονικής και Υπολογιστών

Εργασία Regression στην Υπολογιστική Νοημοσύνη

Εργασία του
Φώτη Αλεξανδρίδη, ΑΕΜ: 9953
faalexandr@ece.auth.gr

22 Οκτωβρίου 2023

Περιεχόμενα

Κεφάλαιο 1

Πρόλογος

Η εργασία αυτή πραγματεύεται την σχεδίαση ενός TSK model για την μοντελοποίηση πολυμεταβλητών μη γραμμικών μοντέλων και την εφαρμογή του σε πρόβλημα τύπου regression σε δύο εφαρμογές: με ένα απλό dataset και με ένα dataset υψηλής διαστασιμότητας.

Κεφάλαιο 2

Παραδοτέα

Το παραδοτέο της εργασίας αποτελούνται από τα ακόλουθα αρχεία:

- function `split_dataset.m`, συνάρτηση η οποία χωρίζει το συνολικό dataset σε σύνολα εκπαίδευσης, επικύρωσης και δοκιμής
- function `extract_xy_data.m`, συνάρτηση η οποία χωρίζει το εκάστοτε μέρος του dataset σε εισόδους και έξοδο
- script `tsk_regression_plain.m`, το οποίο υλοποιεί τα μοντέλα και εμφανίζει τα αποτελέσματα για την απλή περίπτωση dataset
- script `tsk_regression_dimensional.m`, το οποίο υλοποιεί τα μοντέλα και εμφανίζει τα αποτελέσματα για την περίπτωση dataset με υψηλή διαστασιμότητα

Επιπρόσθετα, έχουμε και τα δύο παρεχόμενα datasets, τα αρχεία `airfoil_self_noise.dat` και `superconduct.csv`.

Η εργασία υλοποείται με τα δύο scripts, και τα αποτελέσματα παράγονται αυτοτελή τόσο σε διαγράμματα όσο και στην έξοδο των προγραμμάτων για τις μετρικές.

Κεφάλαιο 3

Υλοποίηση και αποτελέσματα

3.1 Απλό dataset

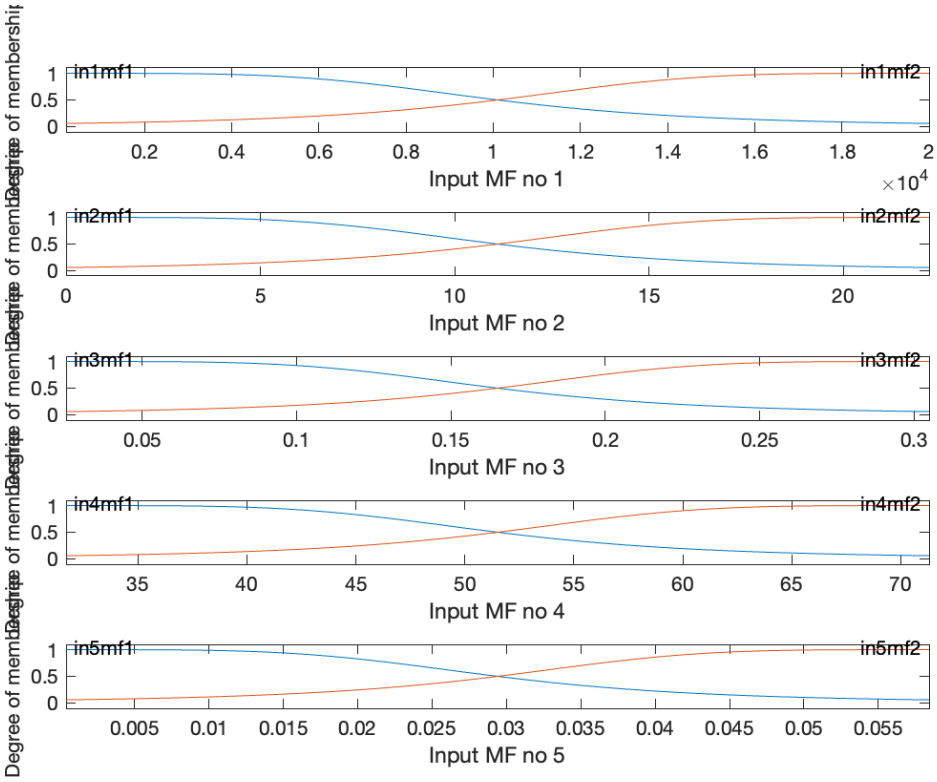
Αρχικά, χρησιμοποιούμε το dataset `airfoil_self_noise.dat`. Πρόκειται για ένα dataset με 5 features. Αφού δημιουργήσουμε τα σύνολα εκπαίδευσης, επικύρωσης και δοκιμής, δημιουργούμε τα τέσσερα μοντέλα με βάση τις προδιαγραφές που μας δίνονται από την εκφώνηση [1] και φαίνονται παρακάτω:

Model #	# of MF	Output Type
1	2	Singleton
2	3	Singleton
3	2	Polynomial
4	3	Polynomial

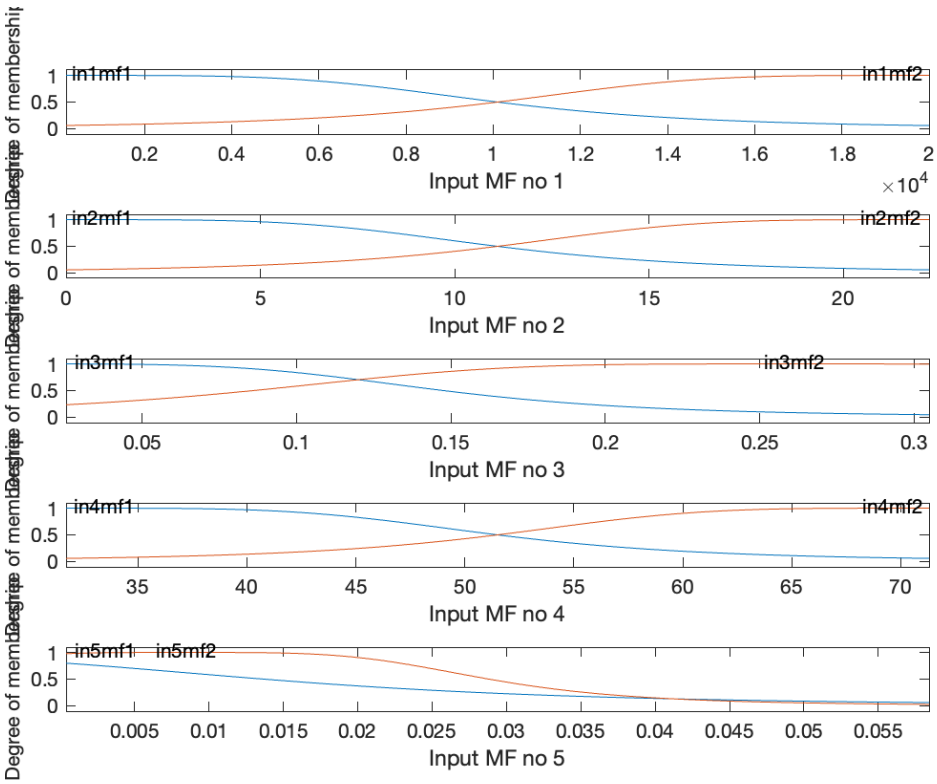
3.1.1 Model 1

Για το μοντέλο 1:

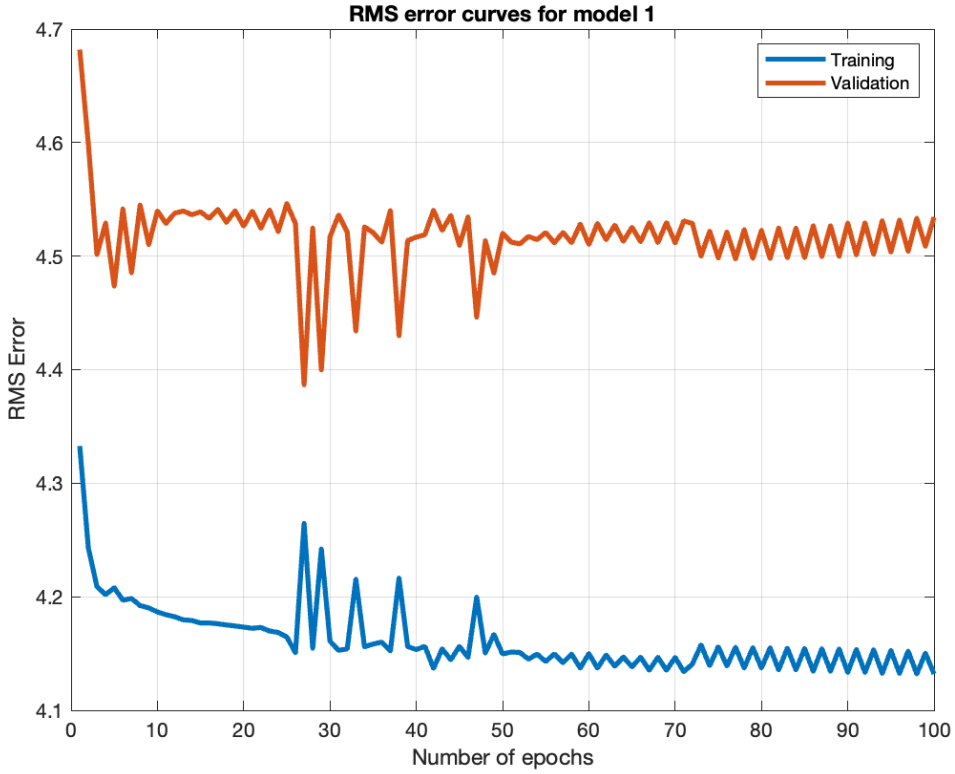
Συναρτήσεις συμμετοχής πριν την εκπαίδευση:



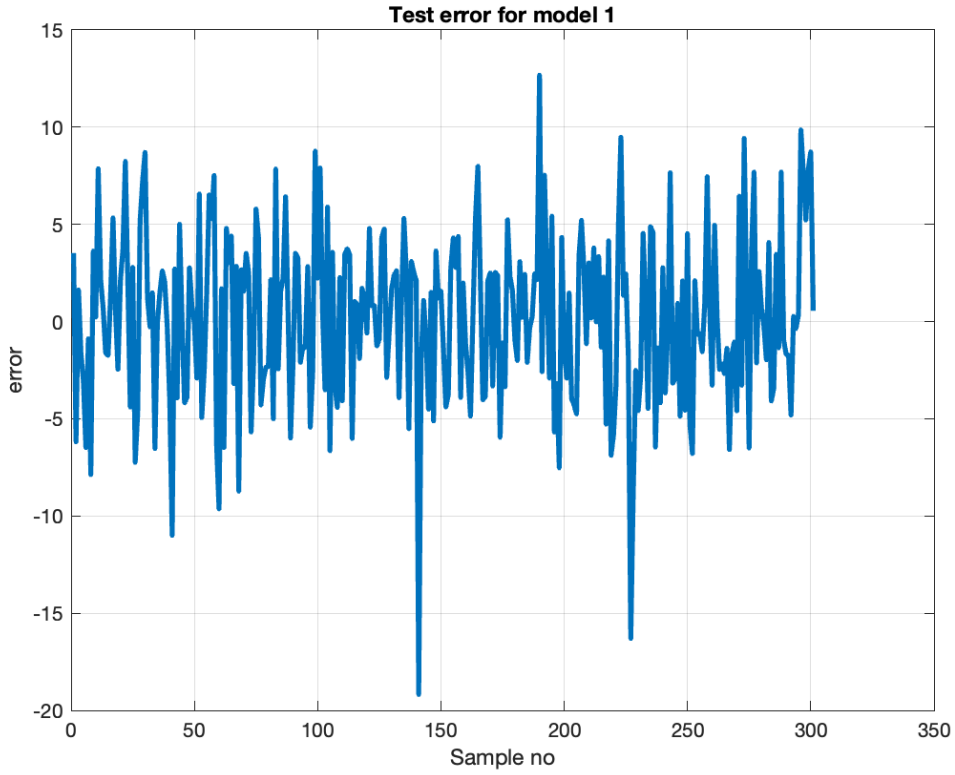
Συναρτήσεις συμμετοχής μετά την εκπαίδευση:



Καμπύλες εκμάθησης:



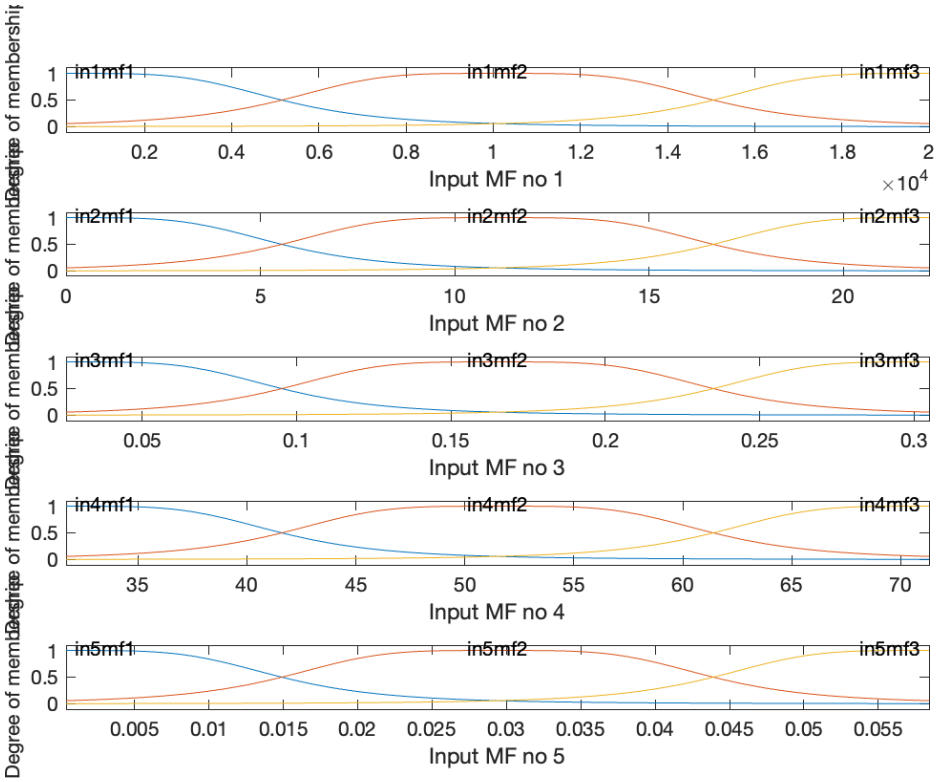
Σφάλματα στο σύνολο δοκιμής:



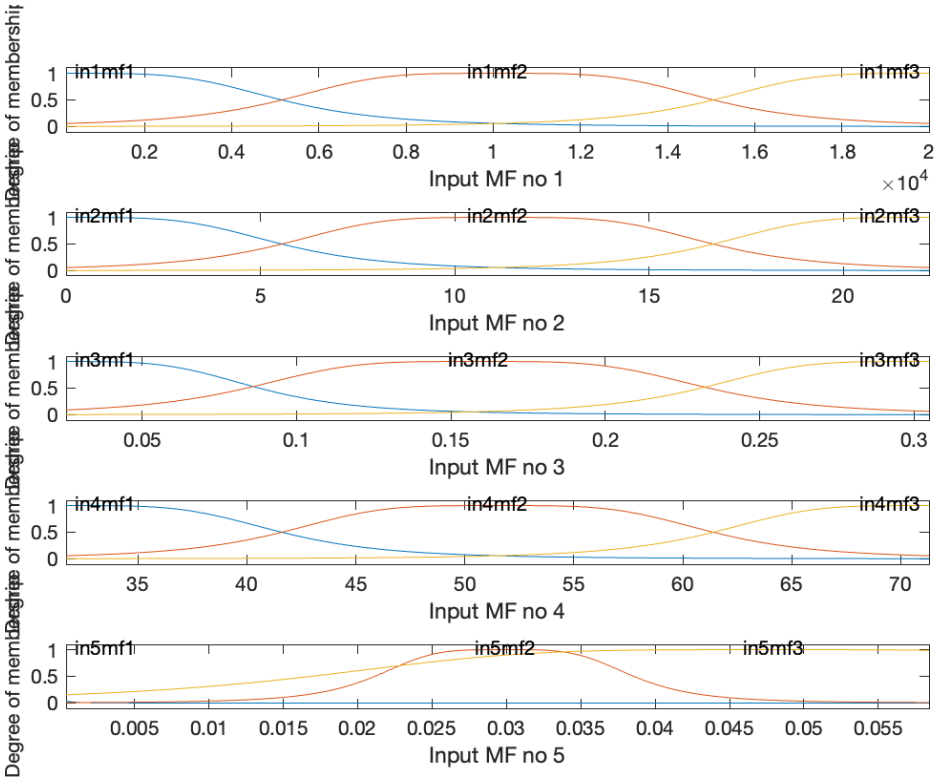
3.1.2 Model 2

Για το μοντέλο 2:

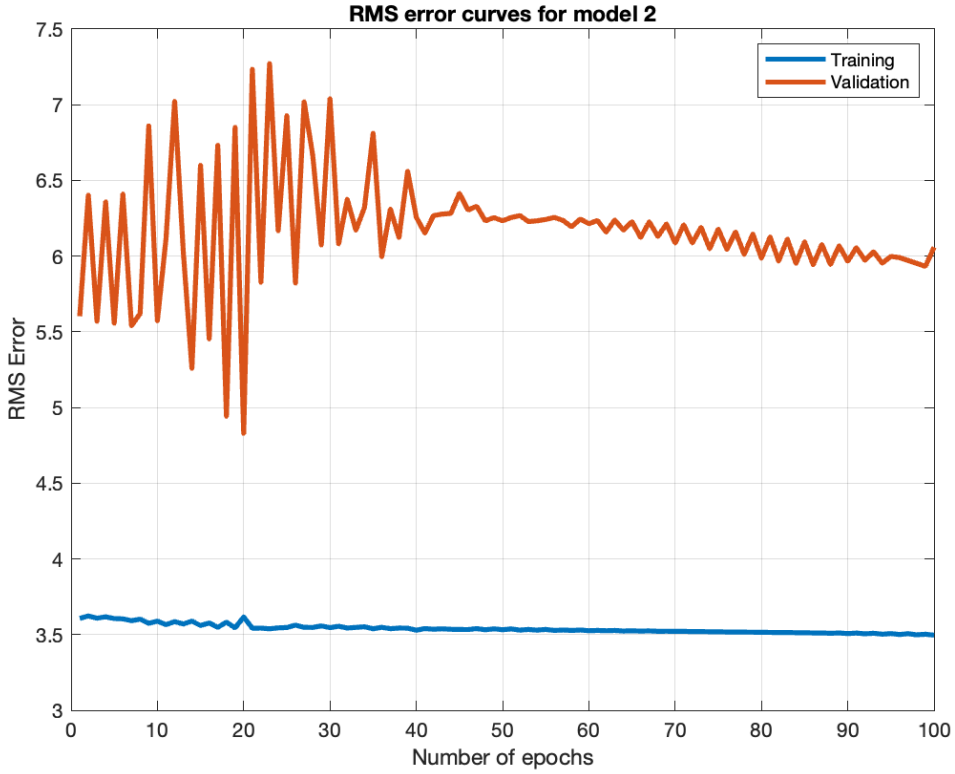
Συναρτήσεις συμμετοχής πριν την εκπαίδευση:



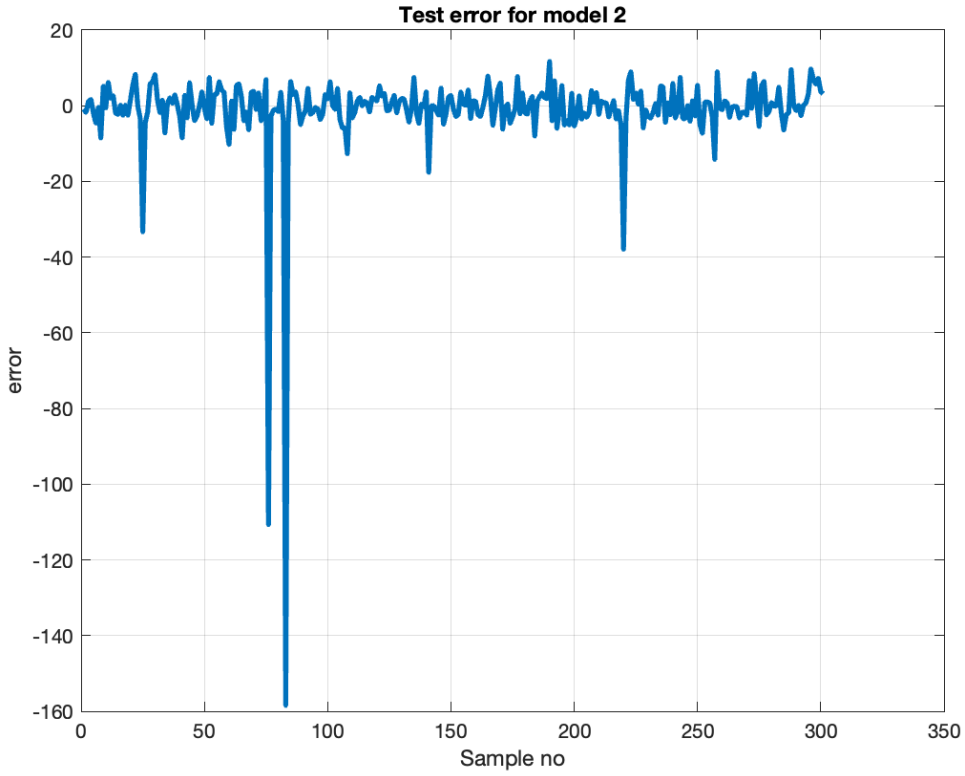
Συναρτήσεις συμμετοχής μετά την εκπαίδευση:



Καμπύλες εκμάθησης:



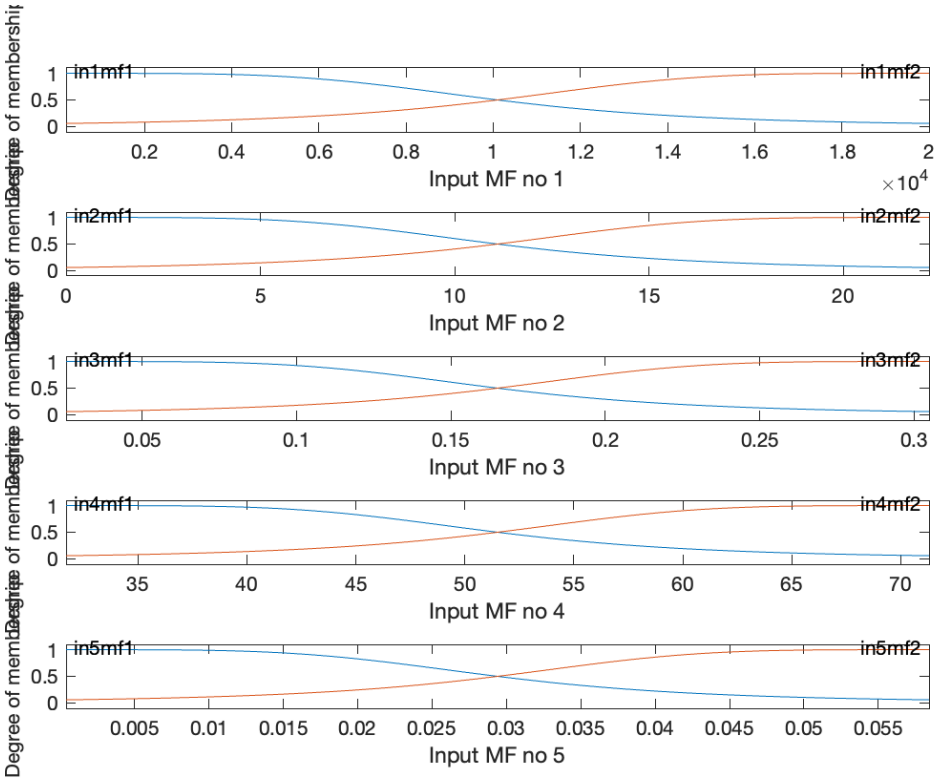
Σφάλματα στο σύνολο δοκιμής:



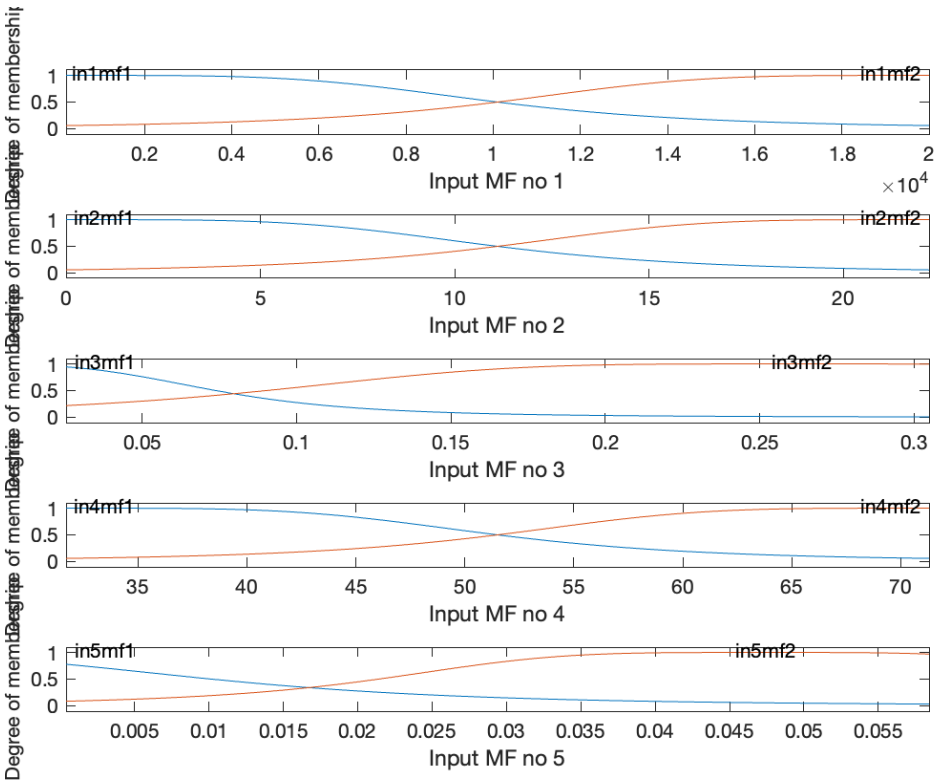
3.1.3 Model 3

Για το μοντέλο 3:

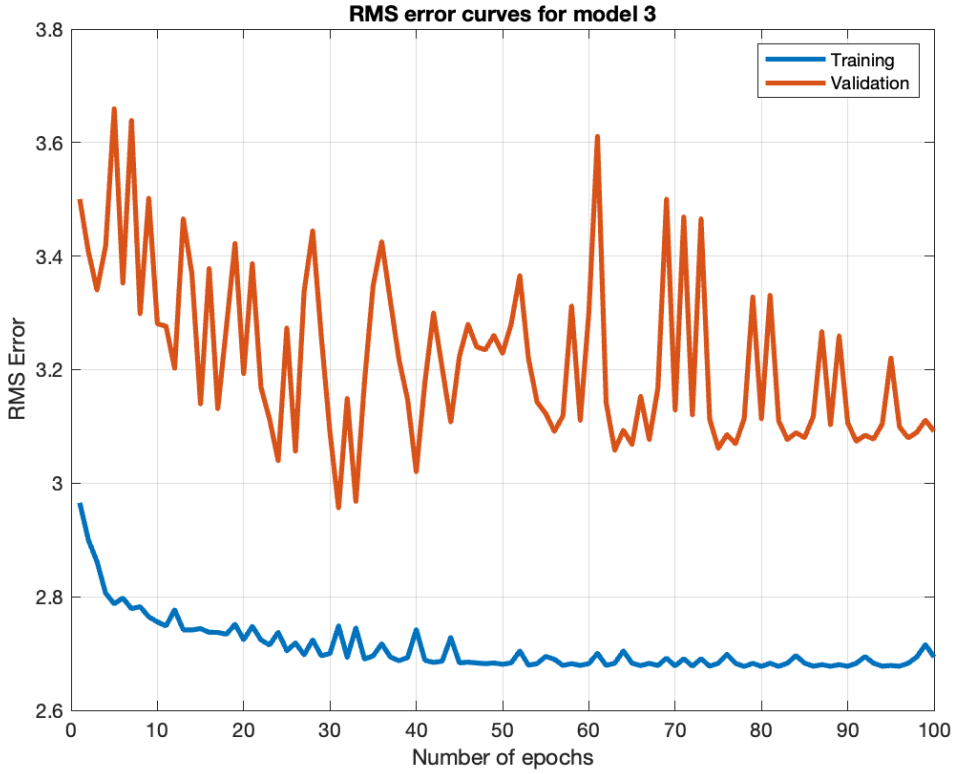
Συναρτήσεις συμμετοχής πριν την εκπαίδευση:



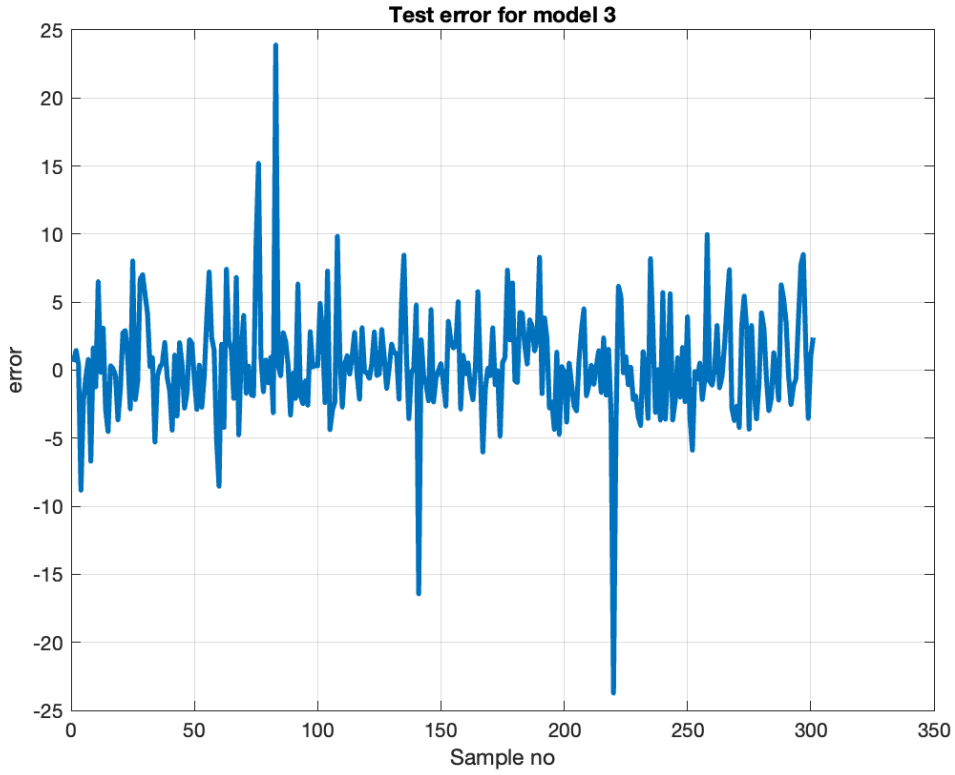
Συναρτήσεις συμμετοχής μετά την εκπαίδευση:



Καμπύλες εκμάθησης:



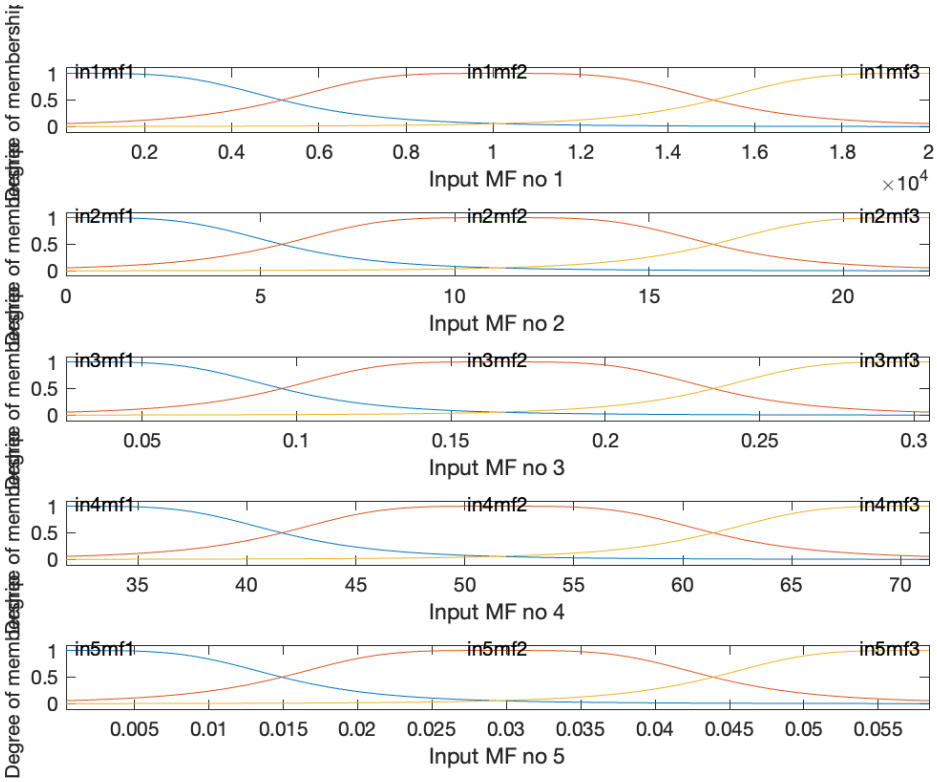
Σφάλματα στο σύνολο δοκιμής:



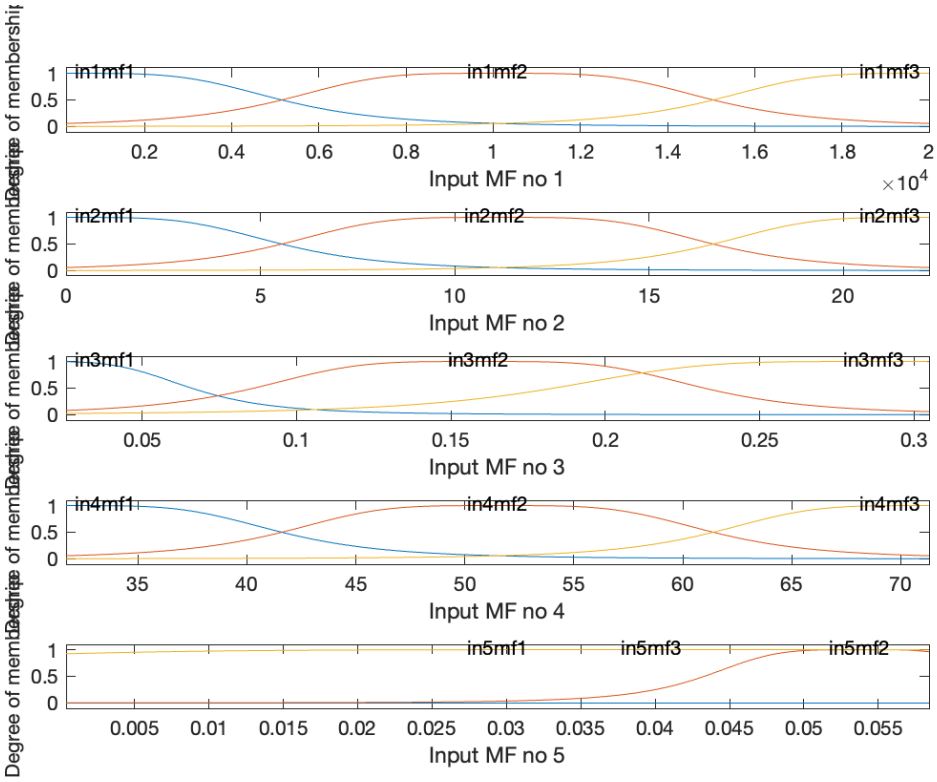
3.1.4 Model 4

Για το μοντέλο 4:

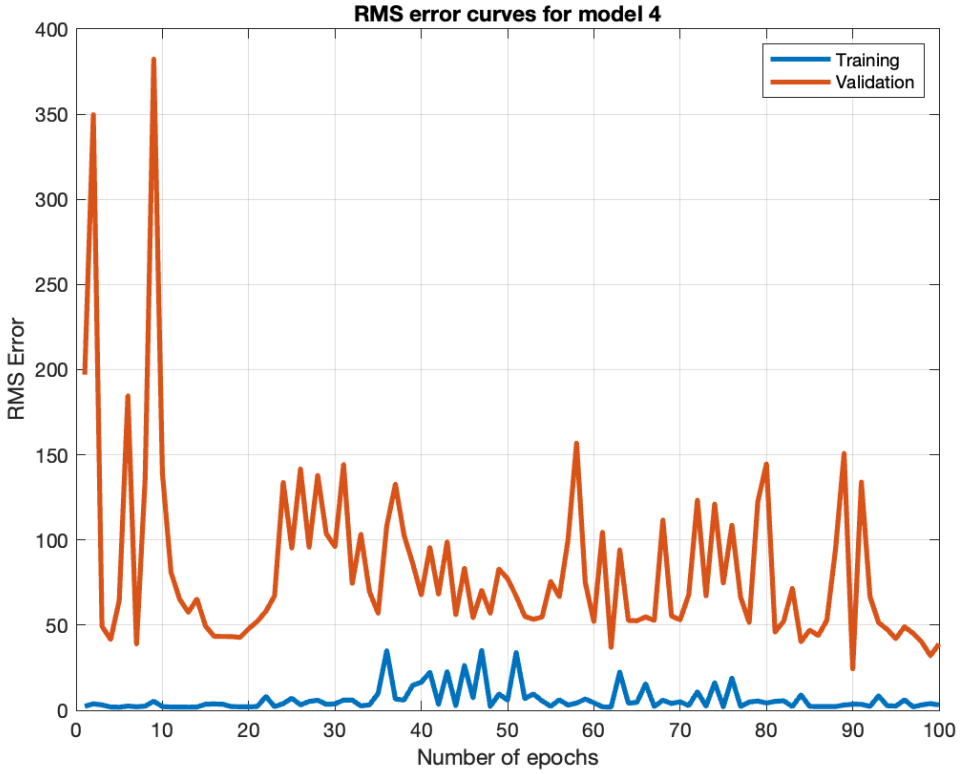
Συναρτήσεις συμμετοχής πριν την εκπαίδευση:



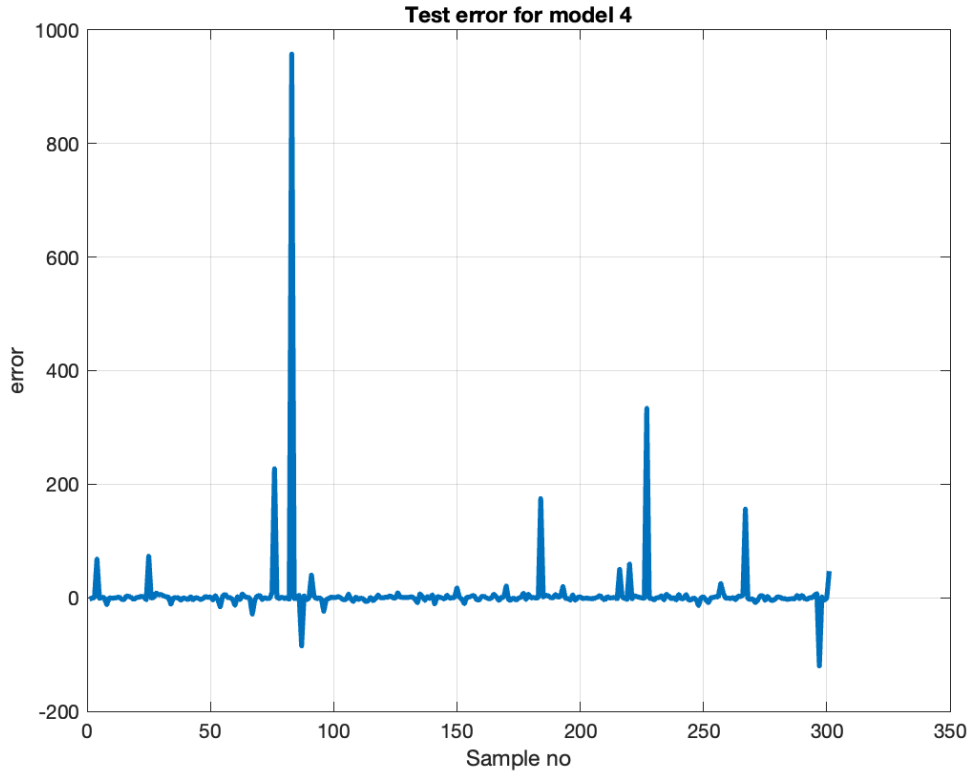
Συναρτήσεις συμμετοχής μετά την εκπαίδευση:



Καμπύλες εκμάθησης:



Σφάλματα στο σύνολο δοκιμής:



3.1.5 Metrics

Για τους ζητούμενους δείκτες απόδοσης, έχουμε τον παρακάτω πίνακα:

Model #	RMSE	NMSE	NDEI	R^2
1	4.399035	0.418001	0.646530	0.581999
2	12.217252	3.224108	1.795580	-2.224108
3	4.022012	0.349421	0.591119	0.650579
4	62.733435	85.008179	9.219988	-84.008179

3.1.6 Σχολιασμός

Από τις καμπύλες εκμάθησης του 2ου μοντέλου, αλλά και από τις μετρικές των μοντέλων 2 και 4, δηλαδή αυτών με τον μεγαλύτερο αριθμό παραμέτρων, παρατηρούμε ότι οδηγούμαστε σε υπερεκμθάθηση, δηλαδή τα μοντέλα μαθαίνουν τους τύπους τον δεδομένων εισόδου αντί να μάθουν όντως χρήσιμες πληροφορίες, και αυτό φαίνεται από την χειρότερη τους απόδοση στο σύνολο δοκιμής. Μάλιστα, ειδικά για το μοντέλο 2 παρατηρούμε την καμπύλη του σφάλματος επικύρωσης να μην μειώνεται, αλλα να παρουσιάζει ταλάντωση, που είναι σημάδι υπερεκμάθησης όταν παράλληλα η καμπύλη του σφάλματος εκπαίδευσης μειώνεται.

Σίγουρα επομένως καταλήγουμε στο συμπέρασμα ότι ο μικρότερος αριθμός ασαφών συνόλων αποδίδει καλύτερα στο συγκεκριμένο σύνολο δεδομένων. Επίσης, μεταξύ του πρώτου και του τρίτου μοντέλου παρατηρούμε μια μικρή βελτίωση απόδοσης στο

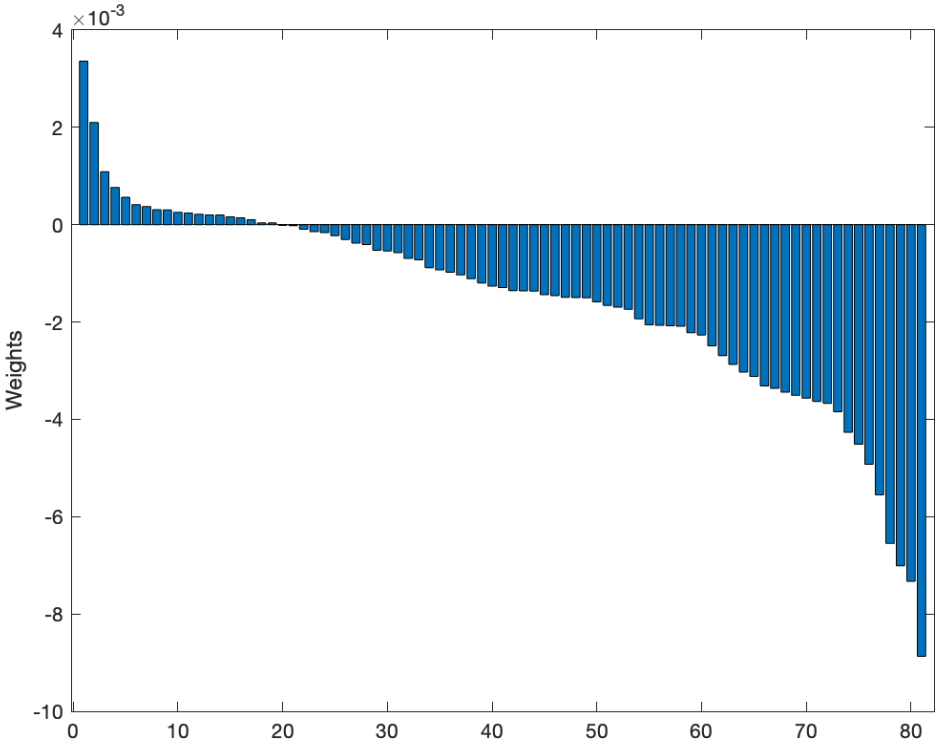
τρίτο μοντέλο, οπότε αποφαινόμαστε ότι η πολυωνυμική έξοδος εκφράζει καλύτερα το πρόβλημα μοντελοποίησής μας, κάτι που είναι και λογικό αν σκεφτούμε την πηγή των δεδομένων μας.

3.2 Dataset με υψηλή διαστασιμότητα

3.2.1 Αποτελέσματα

Εδώ, χρησιμοποιούμε το dataset `superconduct.csv`. Πρόκειται για ένα dataset με 81 features, κάτι που καθιστά απαγορευτική την χρήση απλού μοντέλου με όλες τις παραμέτρους εισόδου, λόγω του αρκετά μεγάλου αριθμού κανόνων που θα χρειαζόντουσαν για την διαμέριση του χώρου εισόδου. Αφού δημιουργήσουμε τα σύνολα εκπαίδευσης, επικύρωσης και δοκιμής, εφαρμόζουμε την μέθοδο της αναζήτησης πλέγματος για να αποφανθούμε για τις τιμές δύο παραμέτρων που θα μειώσουν την διαστασιμότητα του προβλήματος: την επολογή ενός υποσυνόλου των παραμέτρων εισόδου, και την εφαρμογή διαμέρισης διασκορπισμού, συγκεκριμένα την ακτίνα διαμέρισης.

Για την επιλογή του βέλτιστου αριθμού παραμέτρων, εφαρμόζουμε τον αλγόριθμο ReliefF. Λαμβάνουμε το παρακάτω ραβδόγραμμα:



Κατόπιν παρότρυνσης από την εκφώνηση, επιλέγουμε τρεις πιθανές τιμές για καθεμία από τις δύο παραμέτρους, δηλαδή:

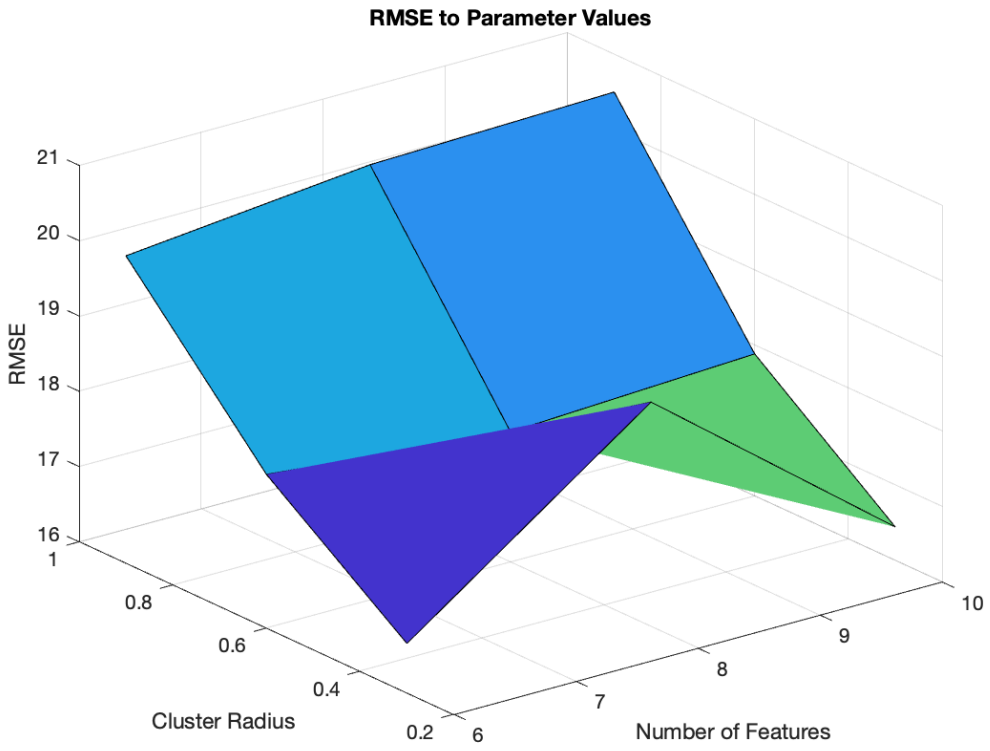
$$N_{features} = 6, 8, 10$$

$$C_{radius} = 0.3, 0.6, 0.9$$

Για καθέναν από τους 9 συνδυασμούς παραμέτρων, εφαρμόζουμε την διαδικασία εκπαίδευσης 5 φορές και λαμβάνουμε τον μέσο όρο του μέσου τετραγωνικού σφάλματος για να αξιολογήσουμε τον συνδυασμό παραμέτρων.

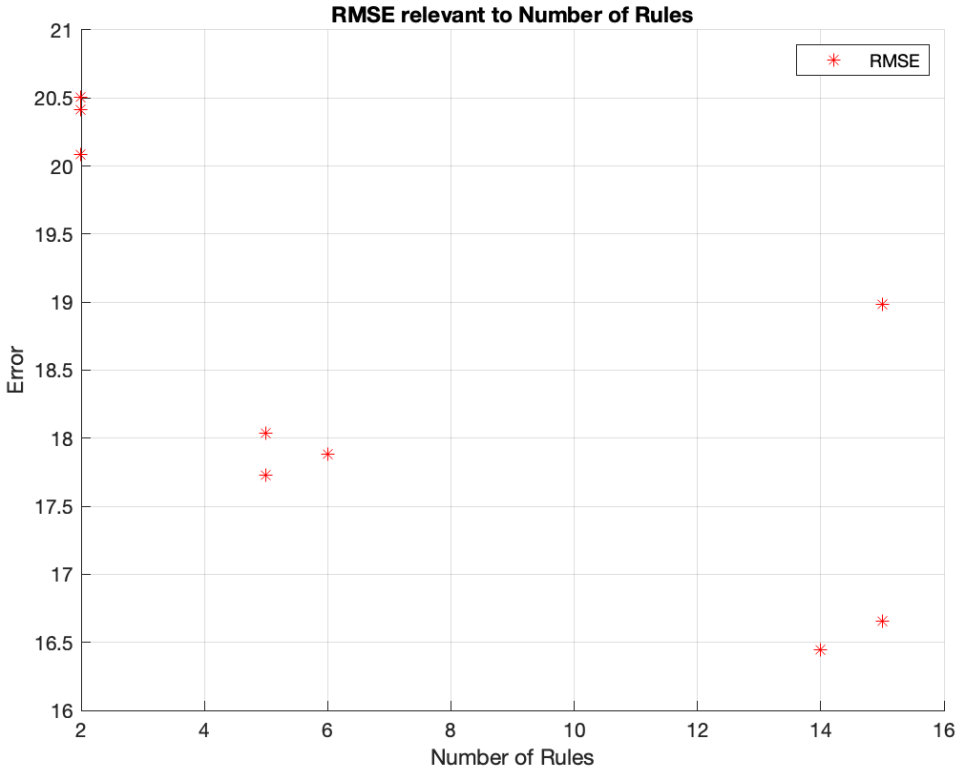
Σε κάθε συνδυασμό κρατάμε τις πρώτες $N_{features}$ πιο σημαντικές παραμέτρους που βγήκαν από τον αλγόριθμο ReliefF και εφαρμόζουμε διαμέριση διασκορπισμού με ακτίνα διαμέρισης C_{radius}.

Έτσι, λαμβάνουμε το παρακάτω διάγραμμα μέσου τετραγωνικού σφάλματος για κάθε ζεύγος παραμέτρων:

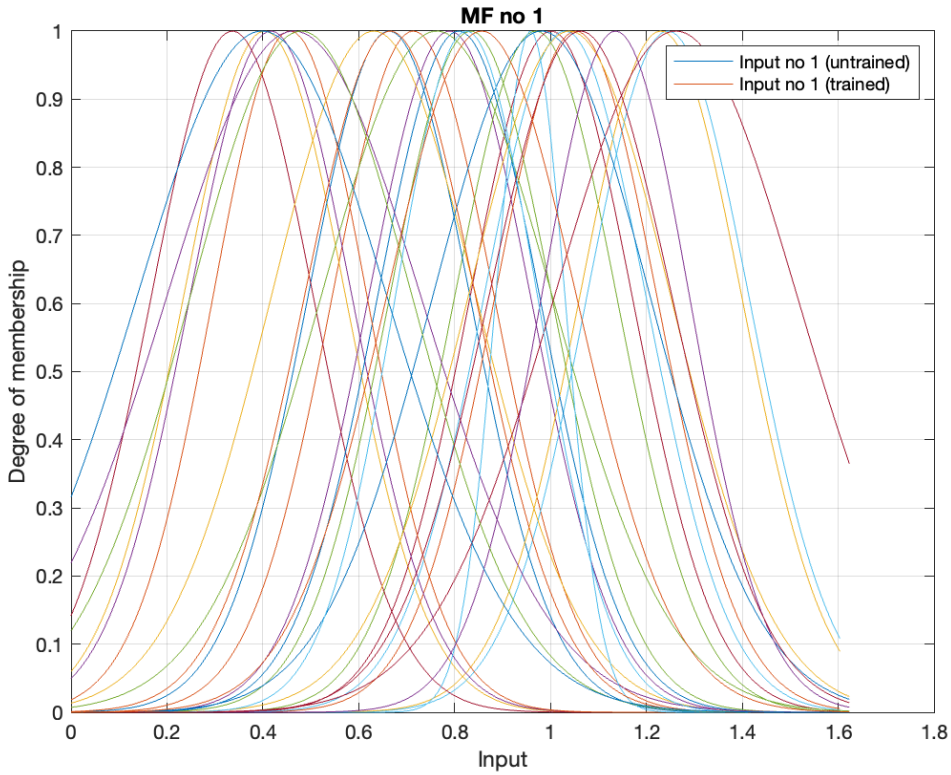


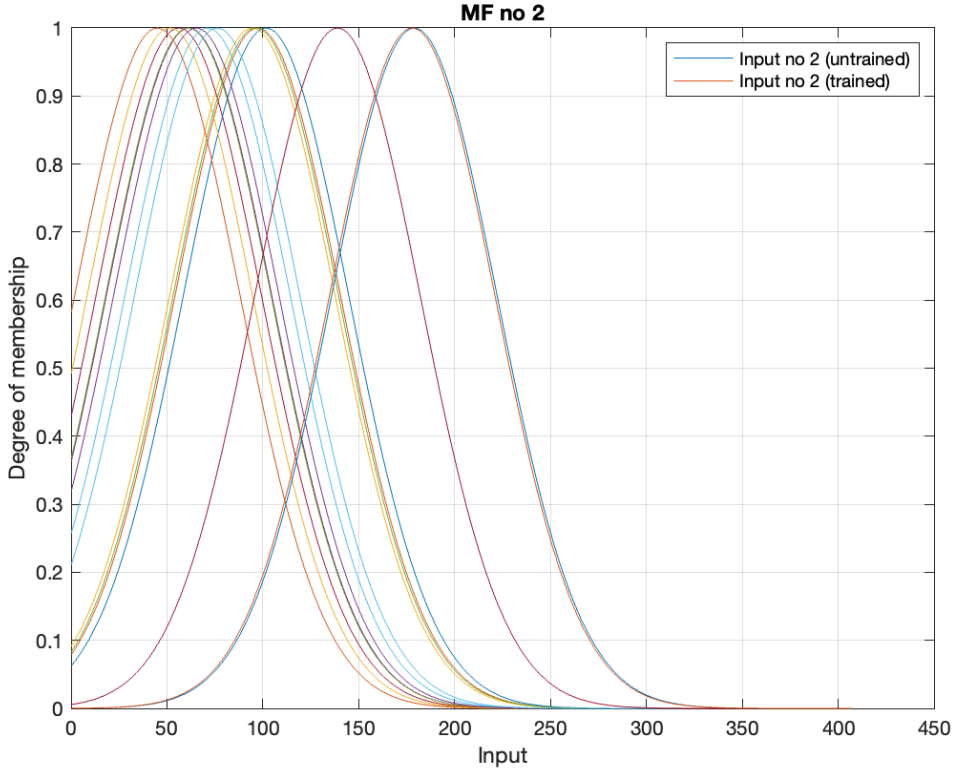
Απ' όπου βλέπουμε ότι πετυχαίνουμε τα καλύτερα αποτελέσματα για
$N_{features} = 10, C_{radius} = 0.3$.

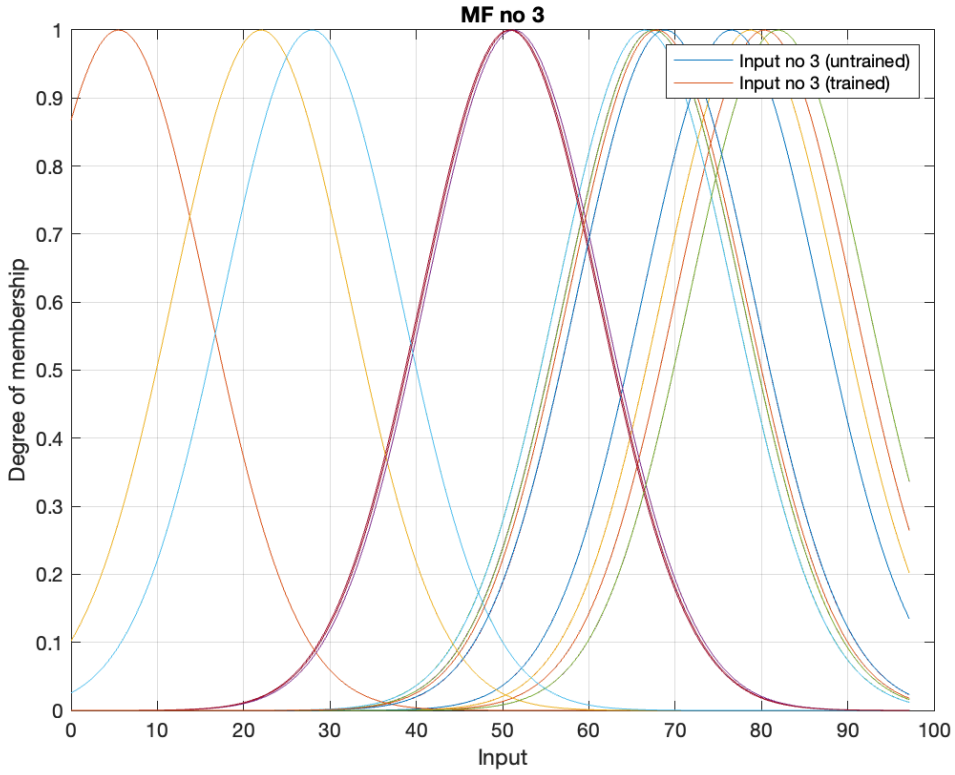
Επίσης, έχουμε και το διάγραμμα που παρουσιάζει το μέσο τετραγωνικό σφάλμα σε συνδυασμό με τον αριθμό των ασαφών κανόνων:

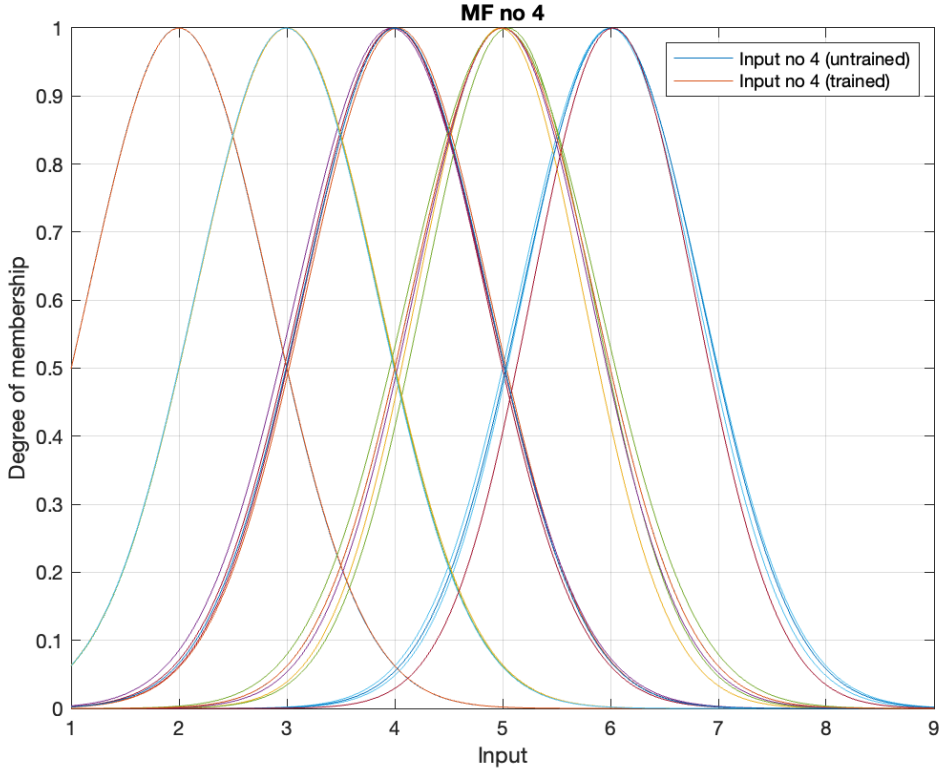


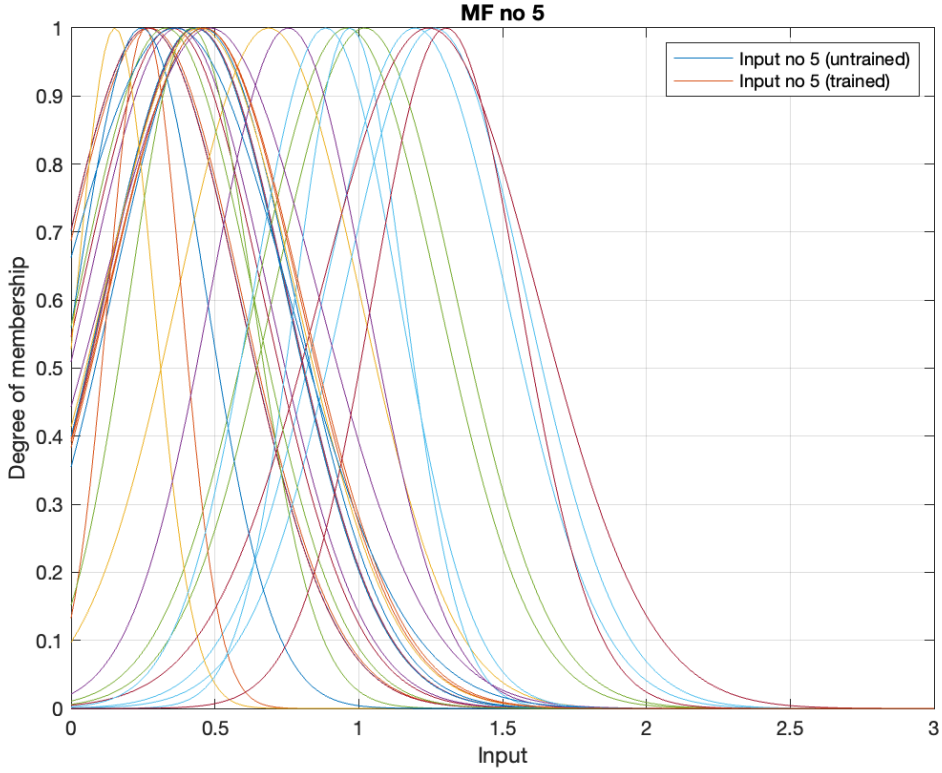
Για το βέλτιστο μοντέλο, αφότου επανεκπαιδεύτηκε, έχουμε τις συναρτήσεις συμμετοχής:

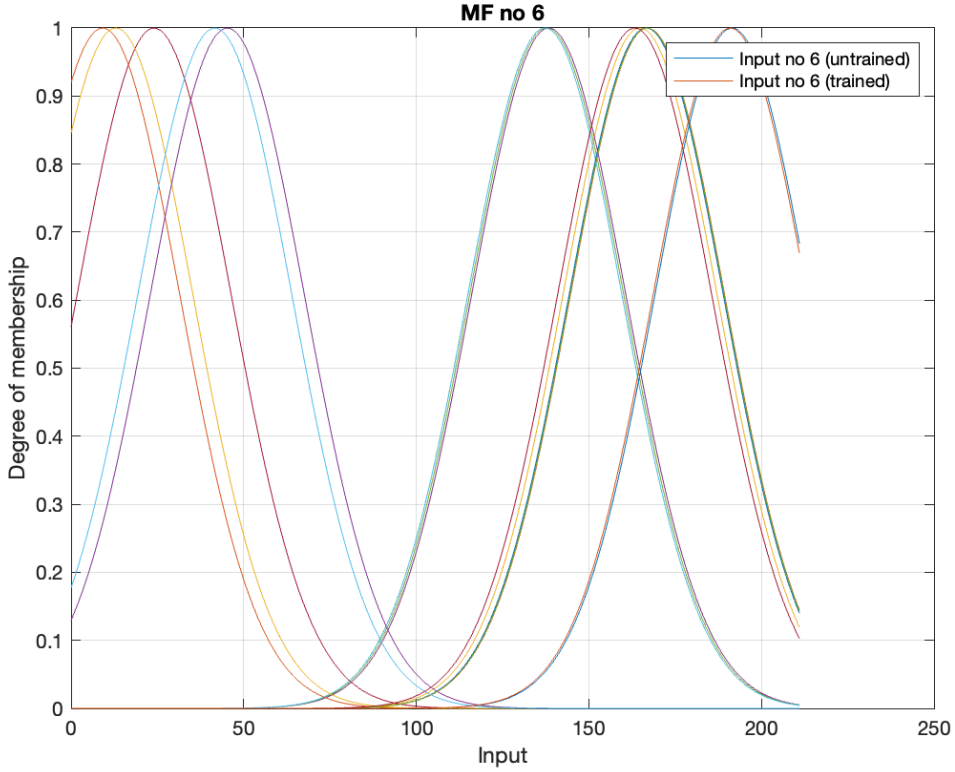


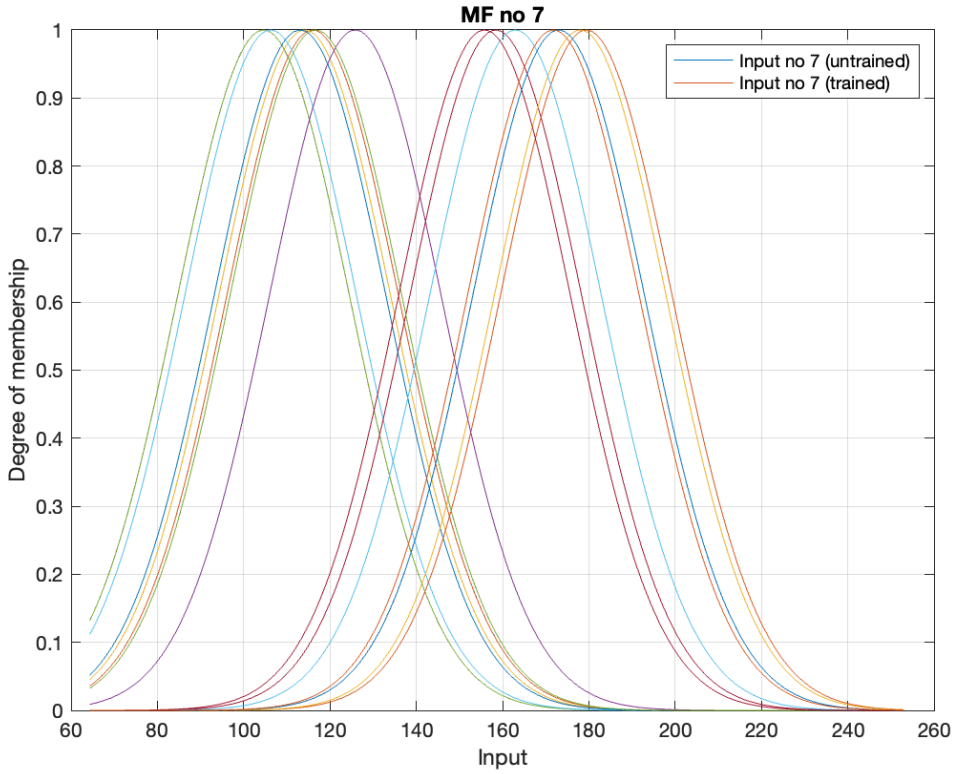

MF no 2
Input no 2 (untrained)
Input no 2 (trained)
Degree of membership
Input
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250
300
350
400
450

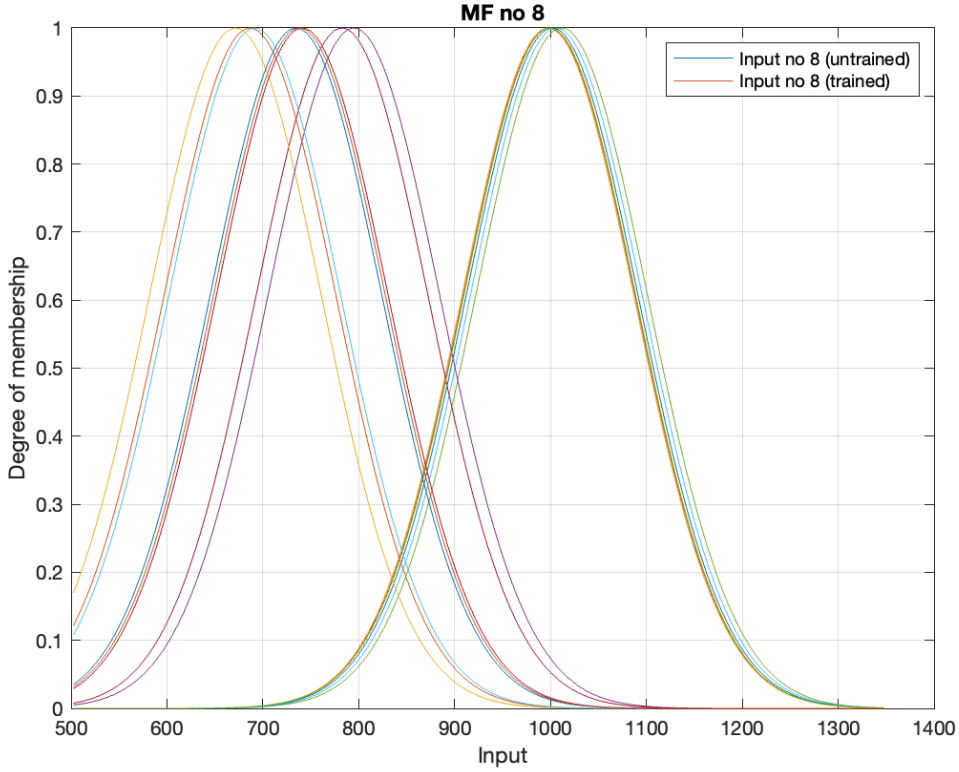

MF no 3
Input no 3 (untrained)
Input no 3 (trained)
Degree of membership
Input
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
10
20
30
40
50
60
70
80
90
100

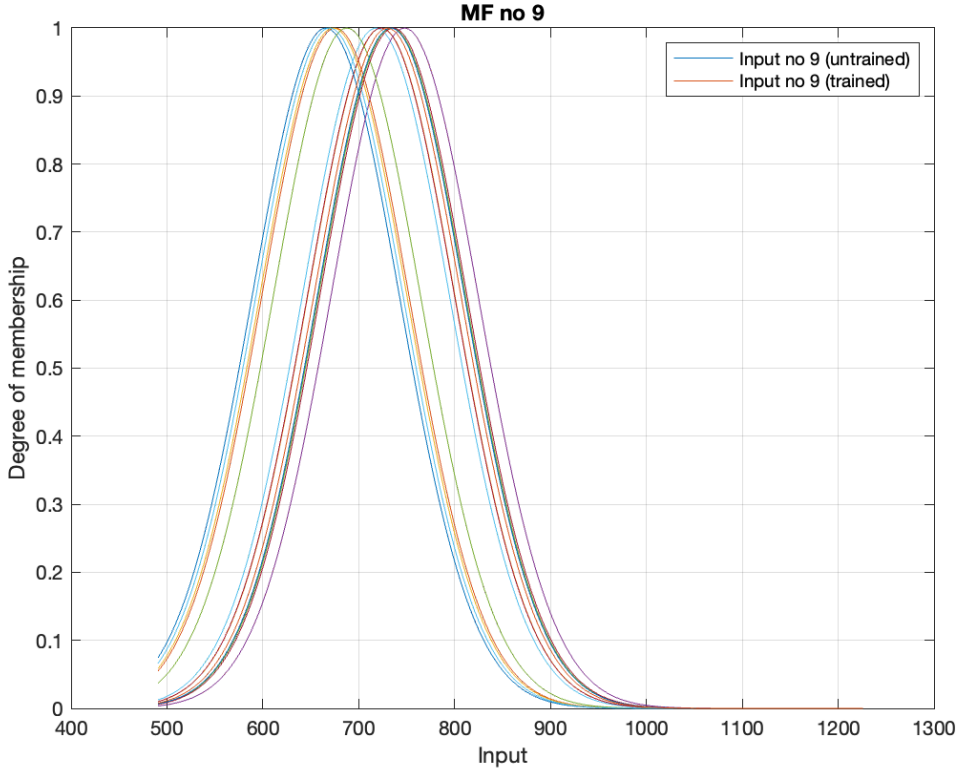

MF no 4
Input no 4 (untrained)
Input no 4 (trained)
Degree of membership
Input
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
2
3
4
5
6
7
8
9

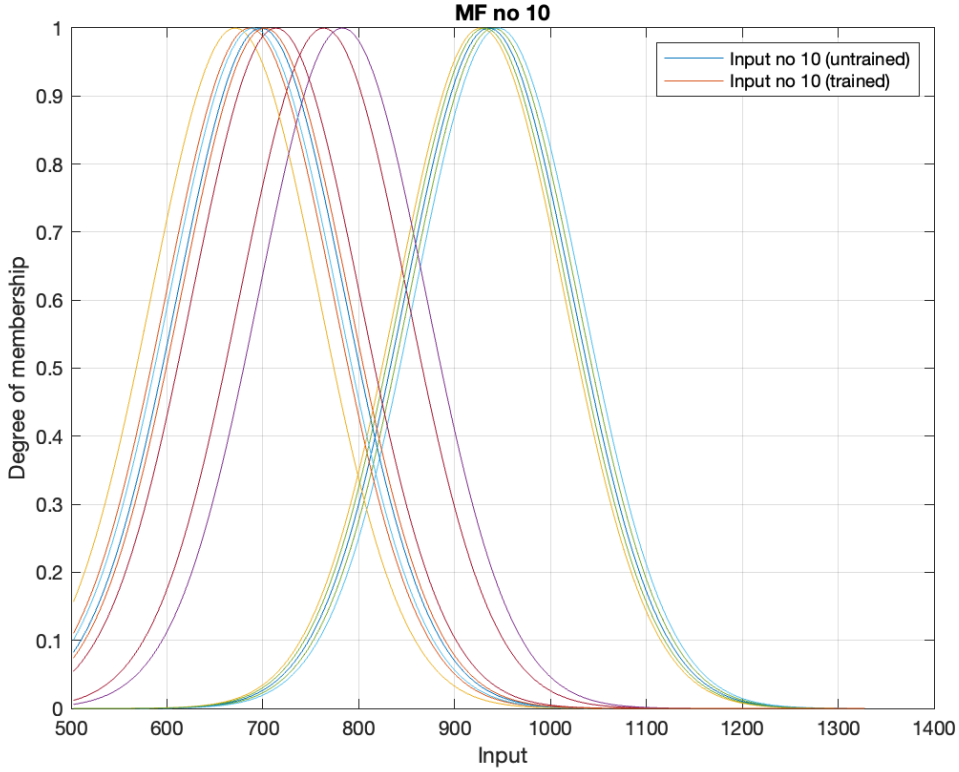

MF no 5
Input no 5 (untrained)
Input no 5 (trained)
Degree of membership
Input
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3


MF no 6
Input no 6 (untrained)
Input no 6 (trained)
Degree of membership
Input
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250

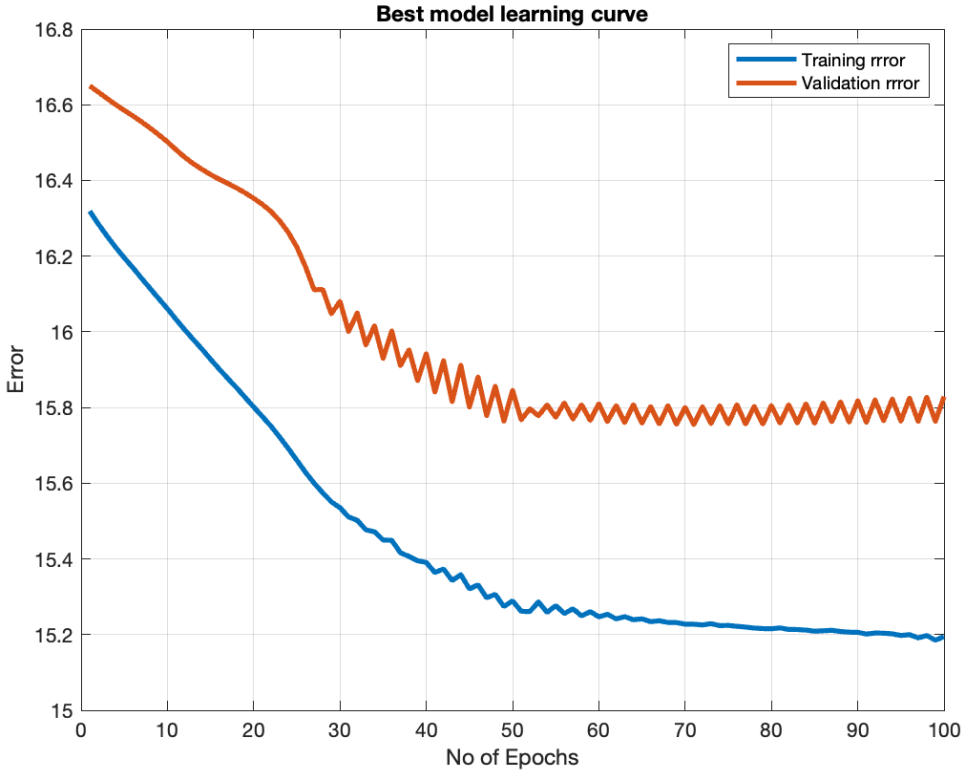

MF no 7
Input no 7 (untrained)
Input no 7 (trained)
Degree of membership
Input
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
60
80
100
120
140
160
180
200
220
240
260


MF no 8
Input no 8 (untrained)
Input no 8 (trained)
Degree of membership
Input
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
500
600
700
800
900
1000
1100
1200
1300
1400

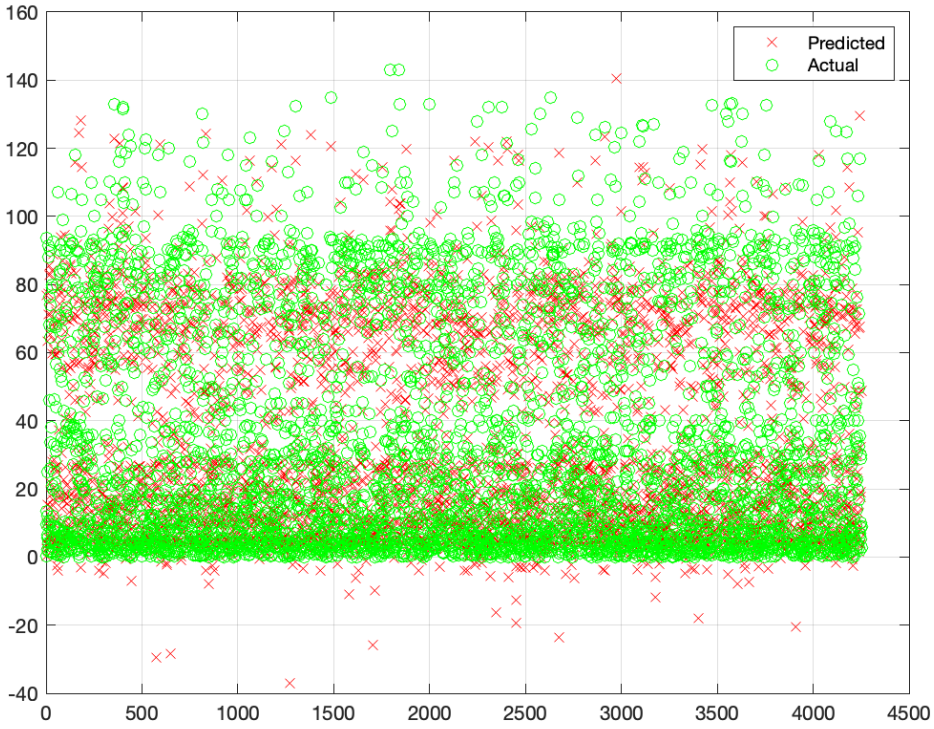

MF no 9
Input no 9 (untrained)
Input no 9 (trained)
Degree of membership
Input
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
400
500
600
700
800
900
1000
1100
1200
1300



Καθώς και τις καμπύλες εκμάθησης:



Και τα σφάλματα πρόβλεψης στο σύνολο δοκιμών:



Για το μοντέλο αυτό έχουμε τις παρακάτω μετρικές:

Name	Value
RMSE	15.550008
NMSE	0.209409
NDEI	0.457612
R^2	0.790591

3.2.2 Σχολιασμός

Παρατηρούμε πως παρόλο που χρησιμοποιούμε έναν σχετικά μικρό αριθμό παραμέτρων (10) σε σχέση με τις συνολικές παραμέτρους εισόδου (81), επιτυγχάνουμε αρκετά καλό αποτέλεσμα εκπαίδευσης. Οι μετρικές μας (συγκεκριμένα το R^2 και NMSE) μας δείχνουν ότι έχουμε μια αρκετά καλή εκμάθηση του συνόλου δεδομένων μας, ενώ βλέποντας τις καμπύλες μάθησης, μπορούμε να αποφανθούμε ότι δεν έχουμε υπερεκμάθηση.

Από το σχήμα με τον αριθμό κανόνων συναρτήσει του σφάλματος, το ελάχιστο μέσο τετραγωνικό σφάλμα το πετυχαίνουμε για 14 κανόνες. Αν εφαρμόζαμε διαμέριση διασκορπισμού με 2 ή 3 ασαφή σύνολα ανά είσοδο, τότε το σύνολο ων κανόνων θα αυξανόταν σε 2^14 ή 3^14 κανόνες, αριθμός που θα εισήγαγε πολύ μεγάλη χρονική καθυστέρηση και αύξηση στις απαιτήσεις υπολογιστικής ισχύος για την εκπαίδευση. Πιθανότατα όμως θα πετυχαίναμε καλύτερα αποτελέσματα, καθώς παρατηρούμε πως

το μέσο τετραγωνικό σφάλμα είναι σχετικά υψηλό σε σχέση με την περίπτωση του απλού συνόλου δεδομένων, οπότε μπορούμε να σκεφτούμε πως δεν αξιοποιούνται στο έπακρο οι δυνατότητες του TSK μοντέλου. Όπως συμβαίνει συνήθως σε τομείς μηχανικής μάθησης, έχουμε και εδώ ένα tradeoff ανάμεσα στην απλότητα του μοντέλου (κατ' επέκταση και στην υπολογιστική ισχύ και τον χρόνο που απαιτείται για την εκπαίδευσή του) και στην ακρίβεια των αποτελεσμάτων που λαμβάνουμε.

Bibliography

[1] *Regression*.