ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

Υπολογιστική Νοημοσύνη

ΤΙΤΛΟΣ

Επίλυση προβλήματος παλινδρόμησης με χρήση μοντέλων TSK

Ονοματεπωνυμο: Χρήστος Πεντερίδης
ΑΕΜ: 10111
Email: chrelipen@ece.auth.gr

1. Εισαγωγή

Στην παρούσα εργασία μελετάται το πρόβλημα της παλινδρόμησης μέσω της χρήσης TSK (Takagi–Sugeno–Kang) ασαφών συστημάτων. Τα συστήματα αυτά επιτρέπουν την αναπαράσταση πολύπλοκων σχέσεων μεταξύ εισόδων και εξόδου με τρόπο που συνδυάζει την ευελιξία των νευρωνικών δικτύων και την ερμηνευσιμότητα των ασαφών λογικών συστημάτων.

Συγκεκριμένα, τα TSK μοντέλα αποτελούνται από ένα σύνολο κανόνων τύπου “If–Then”, όπου το τμήμα “If” ορίζει συνθήκες συμμετοχικών συναρτήσεων (MFs) και το “Then” δίνει μια γραμμική ή σταθερή εξίσωση για την έξοδο.

Για την κατασκευή των TSK μοντέλων εφαρμόστηκαν δύο βασικές μέθοδοι:

- Grid Partitioning (Grid-Based FIS Generation):
  Μια πλήρως εποπτική μέθοδος παραγωγής του αρχικού συστήματος κανόνων, όπου ορίζονται αριθμοί συναρτήσεων συμμετοχής (Membership Functions) για κάθε είσοδο. Ο αριθμός των κανόνων προκύπτει από τον καρτεσιανό πολλαπλασιασμό των MF.

- Subtractive Clustering:
  Αυτοματοποιημένη μέθοδος ομαδοποίησης των δεδομένων εισόδου, όπου η δημιουργία κανόνων βασίζεται στη θέση των κέντρων των κλάστερ, με χρήση παραμέτρων όπως η ακτίνα επιρροής (*Cluster Influence Range*).

Η διαδικασία περιλαμβάνει:

1. Προεπεξεργασία των δεδομένων με κανονικοποίηση.
2. Κατασκευή αρχικών TSK συστημάτων με τις παραπάνω μεθόδους.
3. Εκπαίδευση μέσω Adaptive Neuro-Fuzzy Inference System (ANFIS).
4. Αξιολόγηση μέσω metrics όπως RMSE, R^2, NMSE, και NDEI.

2. Μέρος Α – Dataset χαμηλής διαστασιμότητας

2.1 Περιγραφή προβλήματος – Dataset *Airfoil Self-Noise*

Στο πρώτο μέρος της εργασίας χρησιμοποιείται το γνωστό dataset **Airfoil Self-Noise** από το UCI Machine Learning Repository. Το dataset αυτό περιέχει **1503 παρατηρήσεις** που αντιστοιχούν σε **μετρήσεις ακουστικού θορύβου** από πτερύγια αεροτομών υπό διαφορετικές συνθήκες ροής.

Το σύνολο δεδομένων αποτελείται από **5 εισόδους** και **1 έξοδο**:

#	**Μεταβλητή**	**Περιγραφή**
1	Frequency (Hz)	Συχνότητα ροής
2	Angle of attack (degrees)	Γωνία προσβολής
3	Chord length (m)	Μήκος χορδής
4	Free-stream velocity (m/s)	Ταχύτητα ροής
5	Suction side displacement thickness (m)	Πάχος οριακού στρώματος
6	Sound pressure level (dB)	Έξοδος: στάθμη θορύβου

Στόχος της παρούσας μελέτης είναι η **μοντελοποίηση της στάθμης θορύβου** συναρτήσει των πέντε παραμέτρων λειτουργίας, με χρήση TSK ασαφών μοντέλων.

Για την επεξεργασία του dataset πραγματοποιήθηκε:

- **Κανονικοποίηση** των εισόδων στο [0,1] (unit hypercube scaling),
- **Τυχαίος διαχωρισμός** σε 60% training, 20% validation (checking), και 20% testing σύνολα.

Η παλινδρόμηση αντιμετωπίζεται ως **TSK FIS Regression**, και η αρχική κατασκευή των συστημάτων γίνεται μέσω **Grid Partitioning** με διάφορους αριθμούς MF και τύπους εξόδου (σταθερή ή γραμμική).

2.2 Διαχωρισμός σε σύνολα Εκπαίδευσης, Επικύρωσης και Δοκιμών

Η κατανομή των δεδομένων πραγματοποιήθηκε με **τυχαίο διαχωρισμό (random permutation)** ώστε να διασφαλιστεί η ομοιόμορφη κατανομή παρατηρήσεων στα τρία σύνολα:

- **60% των δεδομένων** χρησιμοποιήθηκε για **εκπαίδευση** (*training set*),
- **20%** για **επικύρωση** κατά την εκπαίδευση (*validation/checking set*),
- **20%** για **τελική αξιολόγηση** του μοντέλου (*test set*).

Ο διαχωρισμός εφαρμόστηκε με χρήση custom συνάρτησης `split_scale()`, η οποία:

- **Εκτελεί κανονικοποίηση** των εισόδων στην ενότητα [0, 1] (προεπεξεργασία επιλογής `preproc=1`),
- Επιστρέφει έτοιμα datasets για χρήση στα FIS μοντέλα.

Η διατήρηση ξεχωριστού validation set είναι κρίσιμη για το ANFIS, ώστε να παρακολουθείται το φαινόμενο του overfitting και να καθορίζεται η βέλτιστη στιγμή τερματισμού της εκπαίδευσης.

2.3 Εκπαίδευση TSK μοντέλων μέσω Grid Partitioning

Για την κατασκευή των TSK μοντέλων στο χαμηλής διαστασιμότητας dataset εφαρμόστηκε η τεχνική **Grid Partitioning**, μέσω της συνάρτησης `genfisOptions` της MATLAB. Η μέθοδος αυτή απαιτεί:

- **Καθορισμό αριθμού συναρτήσεων συμμετοχής (MFs)** ανά είσοδο,
- **Επιλογή τύπου MF** (π.χ. `gbellmf` – generalized bell),
- **Τύπο εξόδου**: σταθερή (`constant`) ή γραμμική (`linear`).

Κατασκευάστηκαν **4 διαφορετικά TSK μοντέλα** με τις εξής παραμέτρους:

Μοντέλο	Αριθμός MF	MF τύπος	Τύπος εξόδου
1	2	gbellmf	constant
2	3	gbellmf	constant
3	2	gbellmf	linear
4	3	gbellmf	linear

Κάθε μοντέλο εκπαιδεύτηκε με χρήση του **ANFIS (Adaptive Neuro-Fuzzy Inference System)** αλγορίθμου της MATLAB, με ρυθμίσεις:

- **Epochs:** 100
- **Training Rate:** 0.01
- **Error Goal:** 0
- **Step size decrease rate:** 0.9
- **Step size increase rate:** 1.1

Η εκπαίδευση παρακολουθούσε **ταυτόχρονα το training και validation error**, επιτρέποντας την έγκαιρη διακοπή σε περίπτωση overfitting.

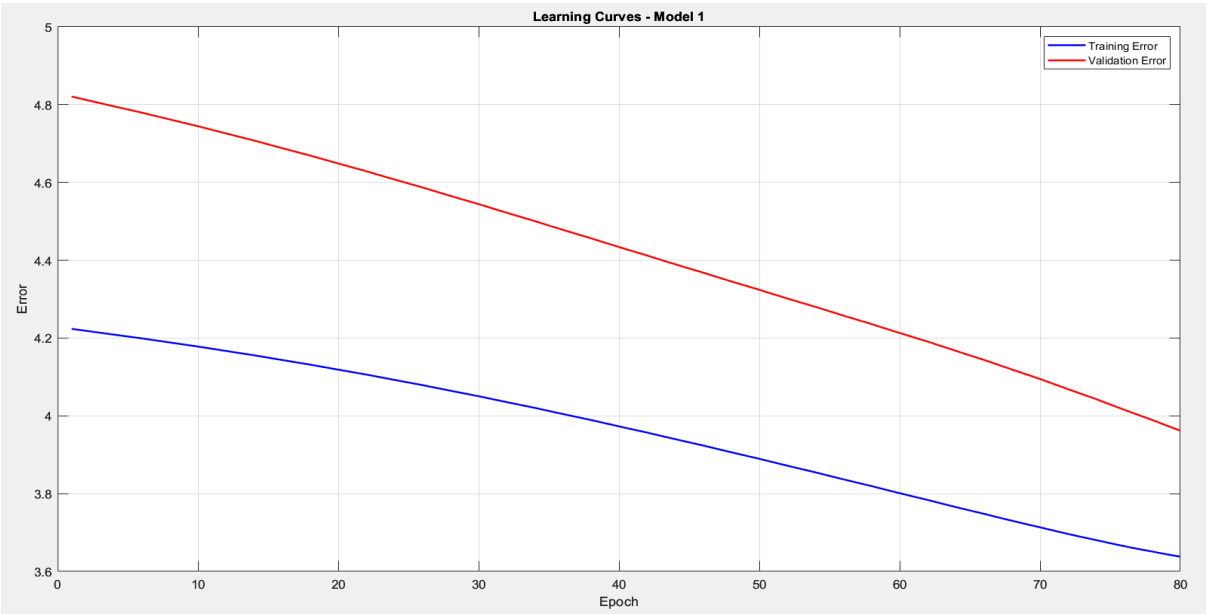
Για κάθε μοντέλο:

- Καταγράφηκαν **καμπύλες εκπαίδευσης (learning curves)**,
- Παρατηρήθηκαν οι **συναρτήσεις συμμετοχής (MFs)** πριν και μετά την εκπαίδευση,
- Υπολογίστηκαν δείκτες αξιολόγησης: **RMSE, NMSE, NDEI, R^2**,
- Καταγράφηκε το **σφάλμα πρόβλεψης (prediction error)**.

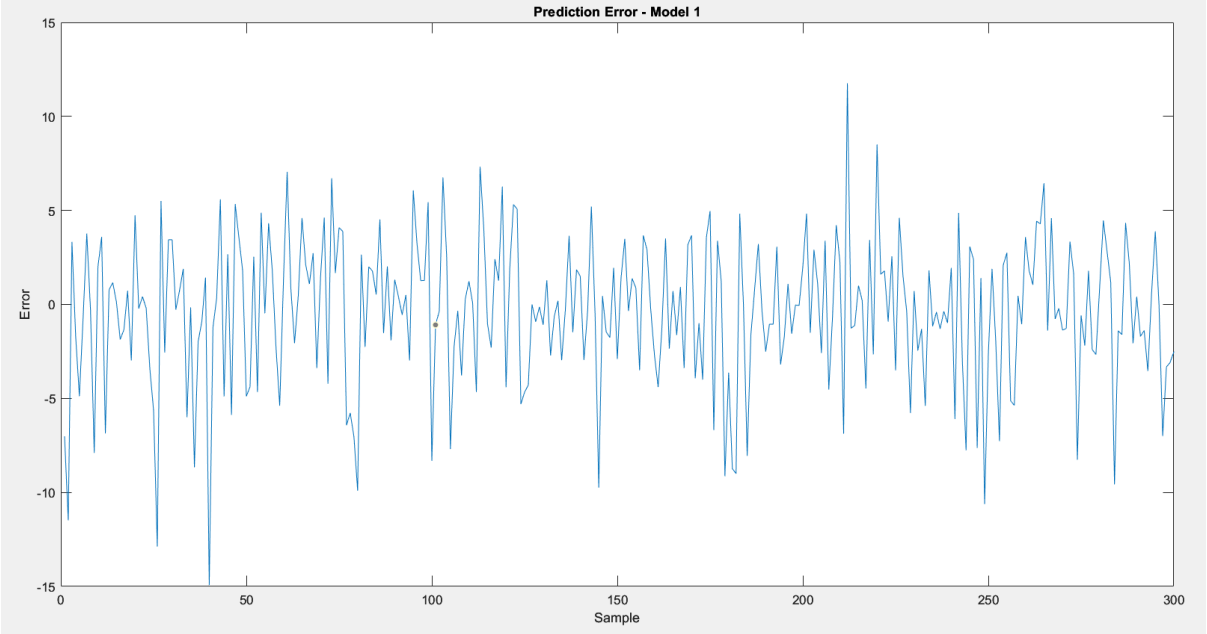
Model 1: Constant Output – 2 MF

1. Εκπαίδευση: 100 epochs
2. Παρατηρήθηκε σταθερή μείωση training error, με ήπια μεταβολή στο validation error.
3. Το prediction error ήταν περιορισμένο, χωρίς εμφανή overfitting.

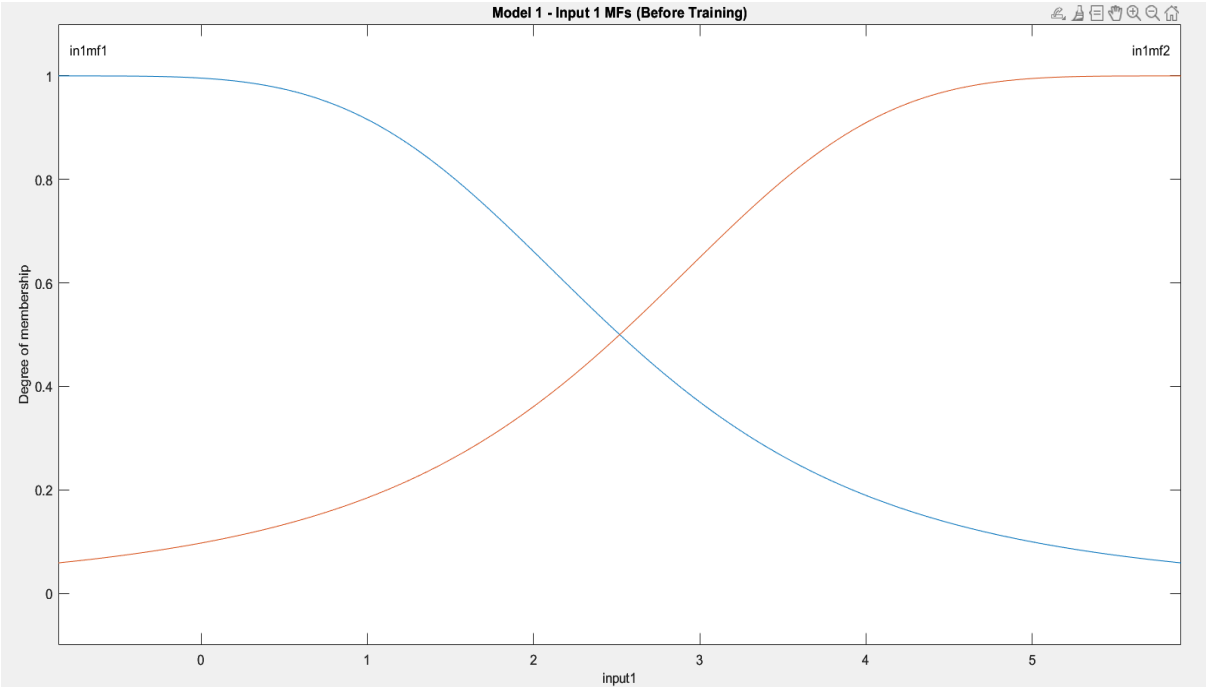
Learning Curve – Model 1



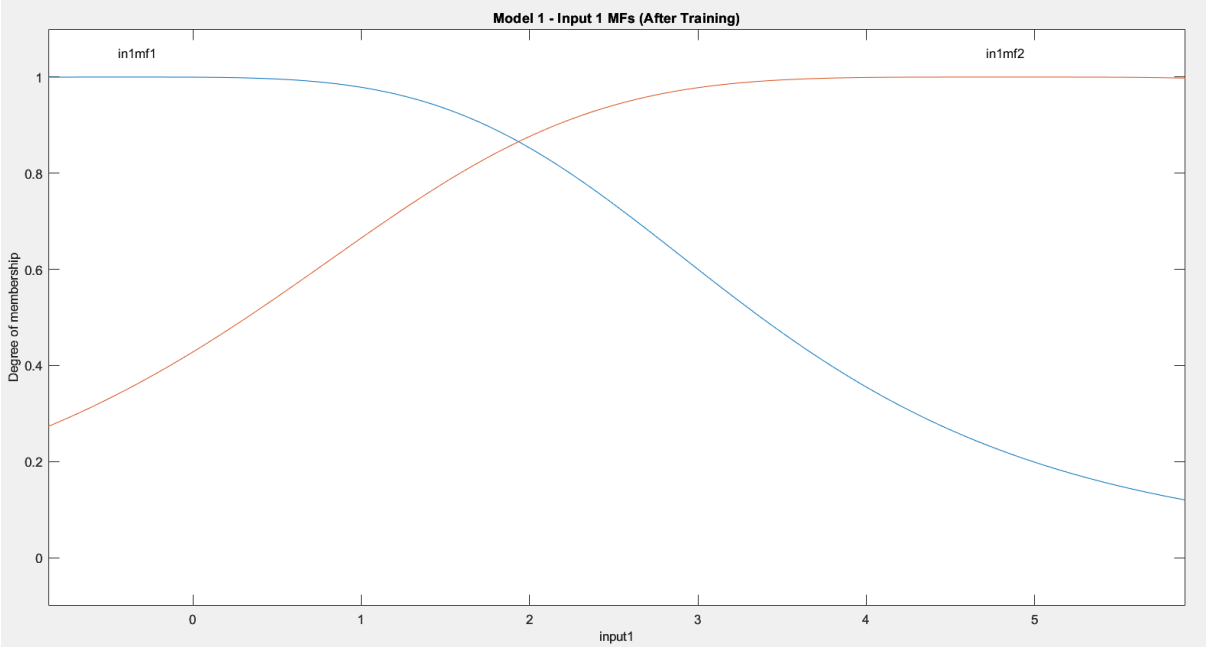
Prediction Error – Model 1



MFs Before Training – Input 1 (Model 1)



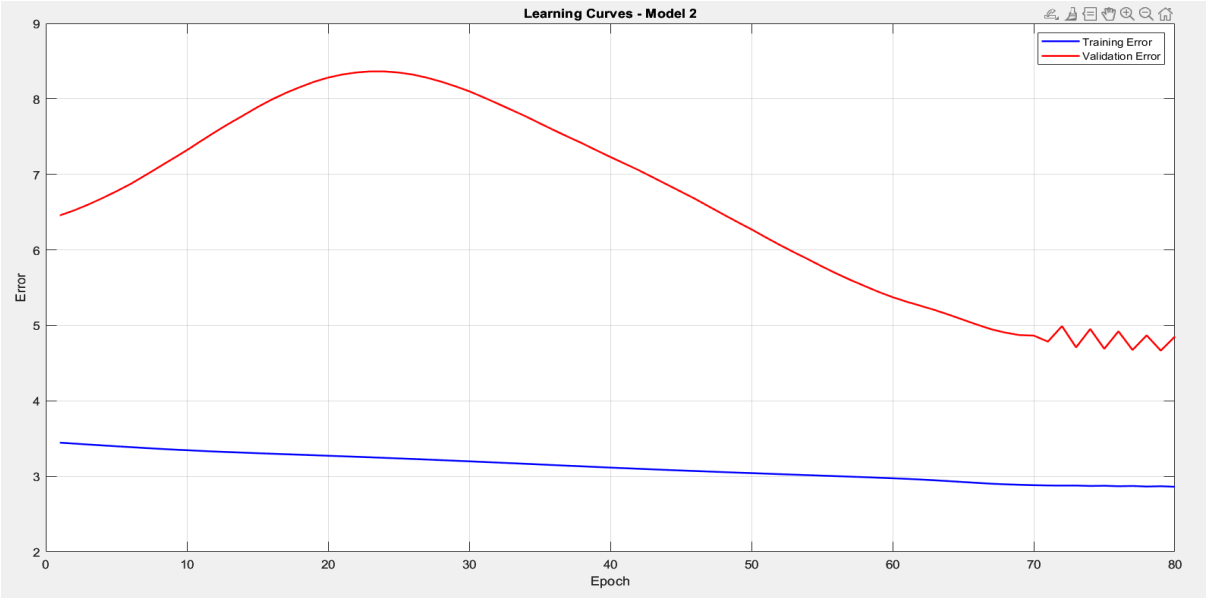
MFs After Training – Input 1 (Model 1)



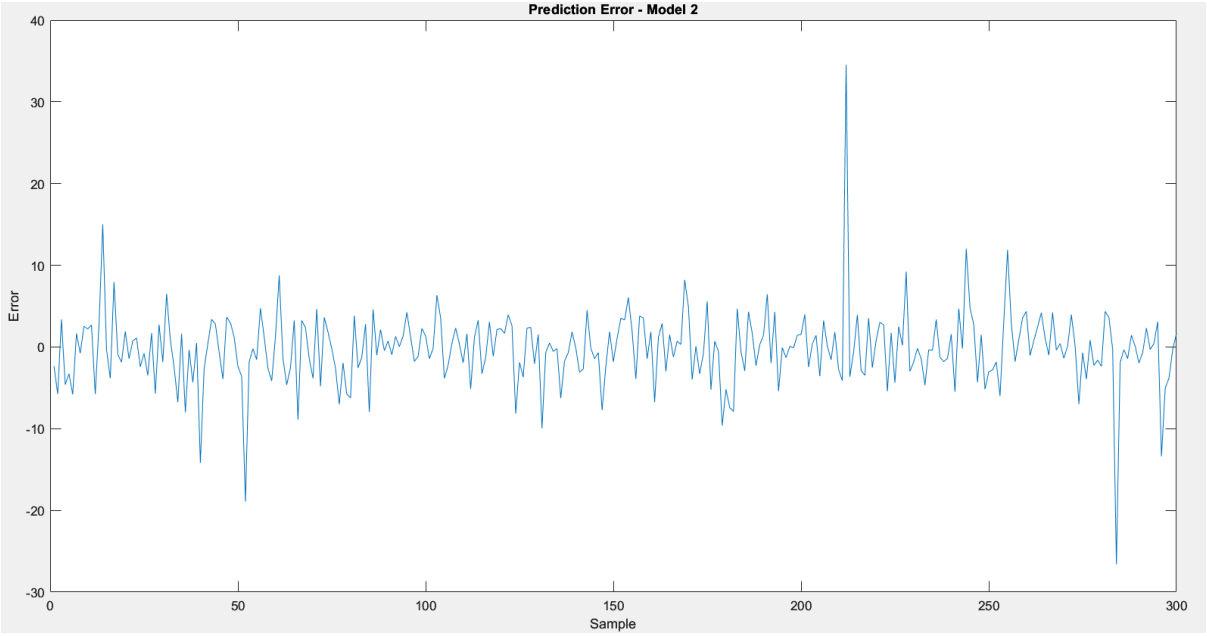
Model 2: Constant Output – 3 MF

Ο αυξημένος αριθμός MF οδήγησε σε μεγαλύτερη ευκαμψία. Παρατηρήθηκε μικρή βελτίωση στην ακρίβεια αλλά με αυξημένο αριθμό κανόνων.

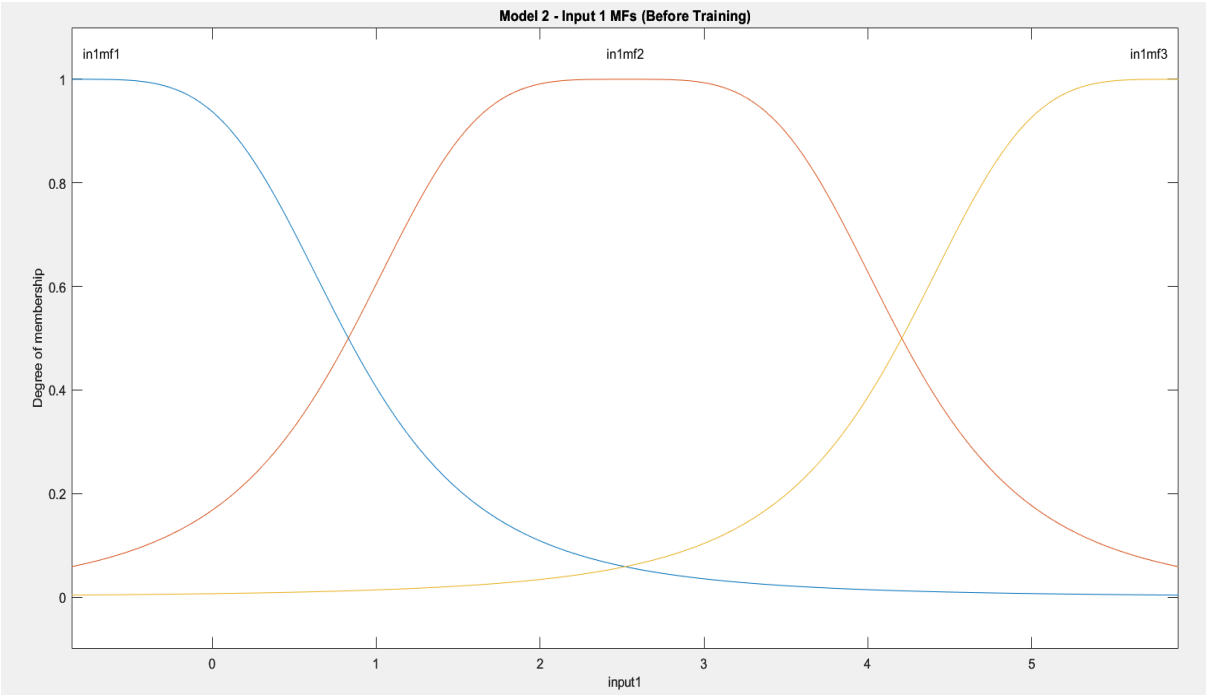
Learning Curve – Model 2

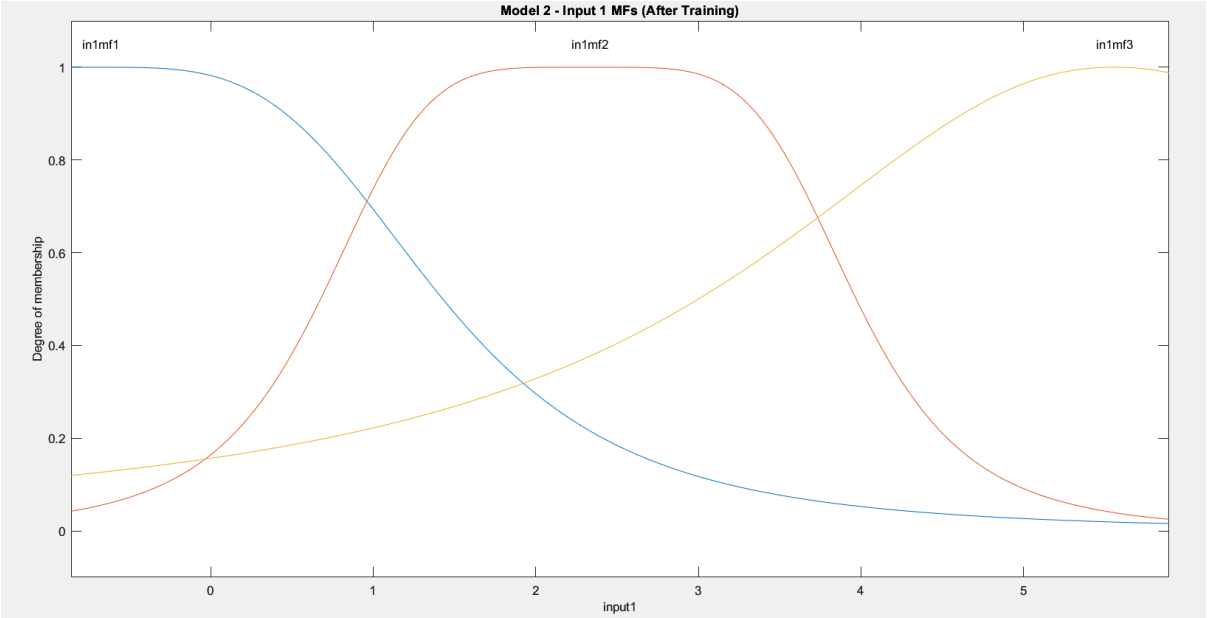


Prediction Error – Model 2



MFs Before/After Training – Model 2

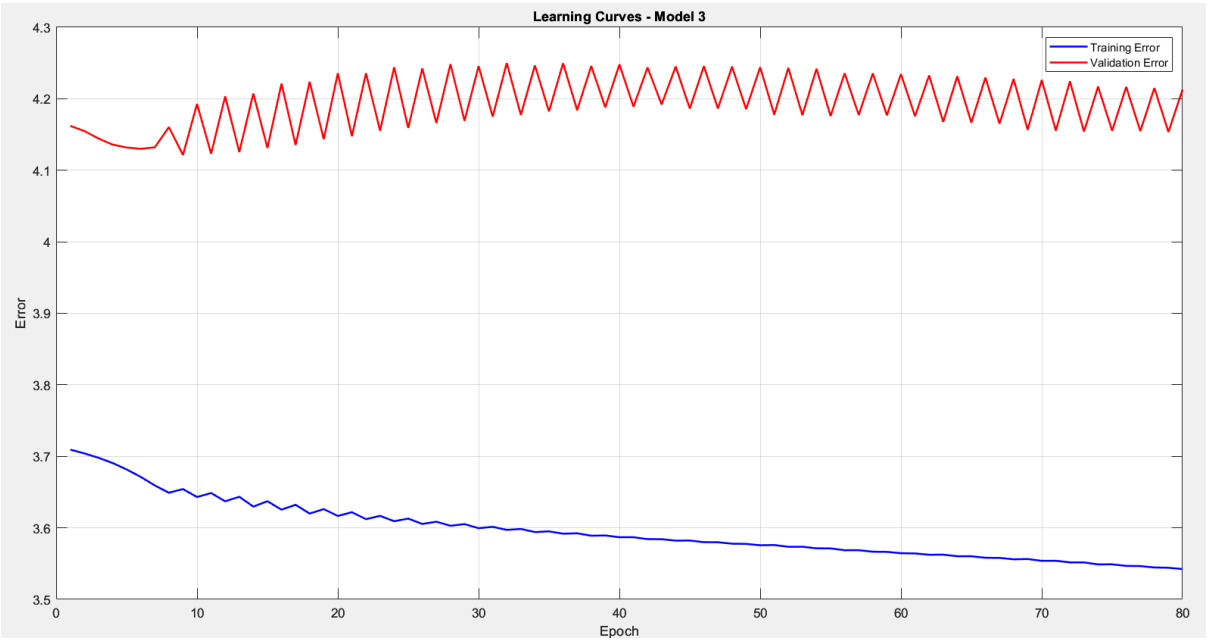




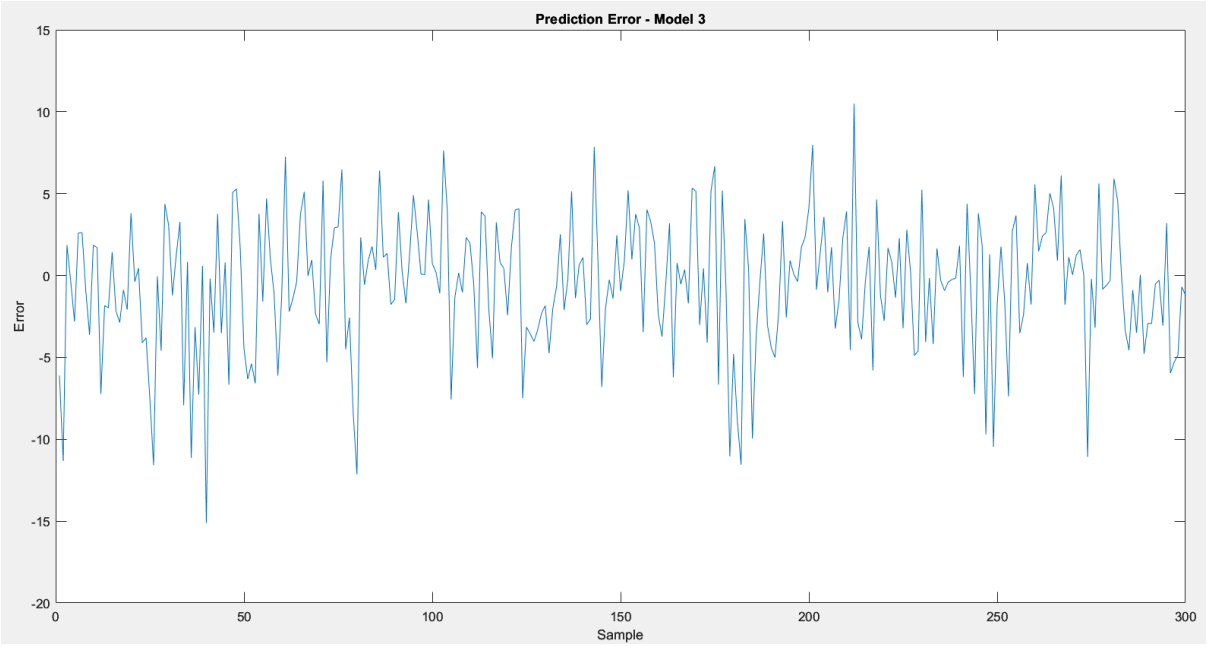
Model 3: Linear Output – 2 MF

Η γραμμική έξοδος προσφέρει καλύτερη προσαρμογή στα δεδομένα. Παρατηρήθηκε πτώση του training error, αλλά το validation error εμφάνισε κυκλικές αυξομειώσεις – πιθανή υπερπροσαρμογή.

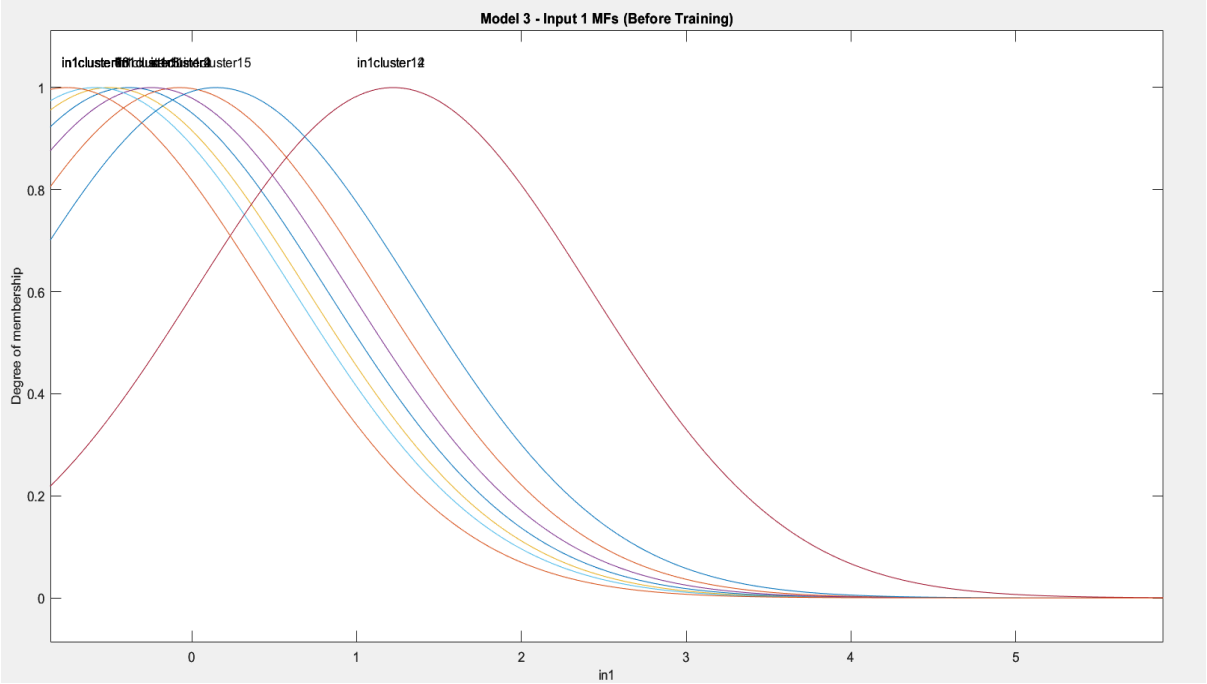
Learning Curve – Model 3

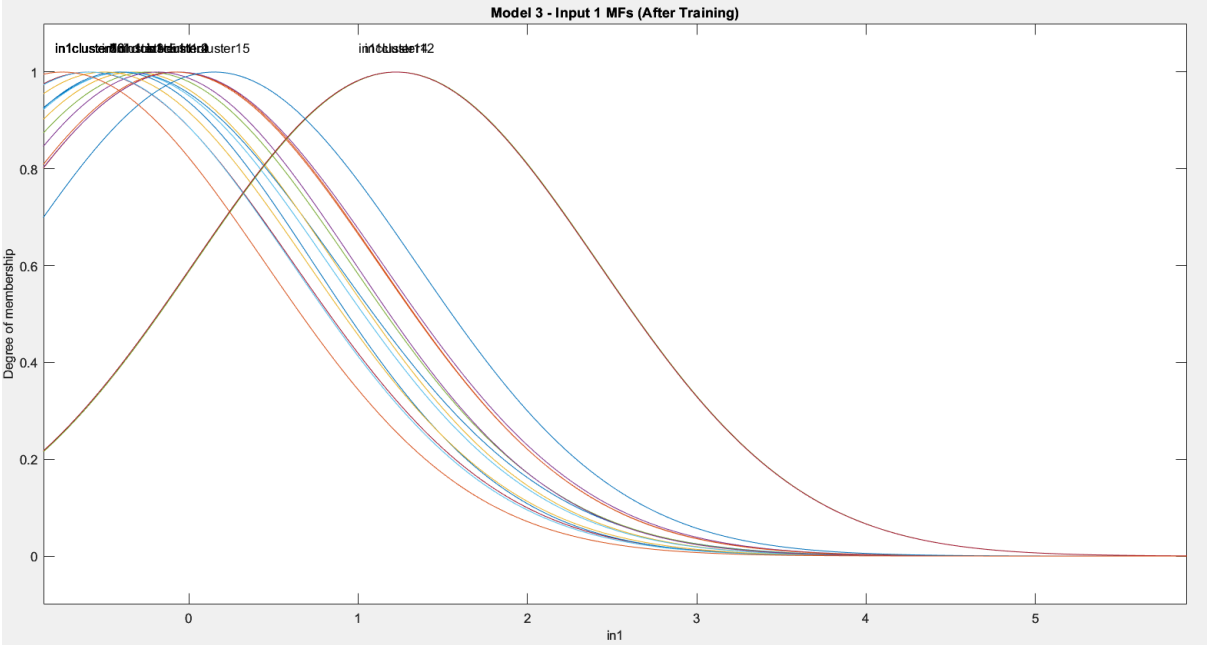


Prediction Error – Model 3



MFs Before/After Training – Model 3

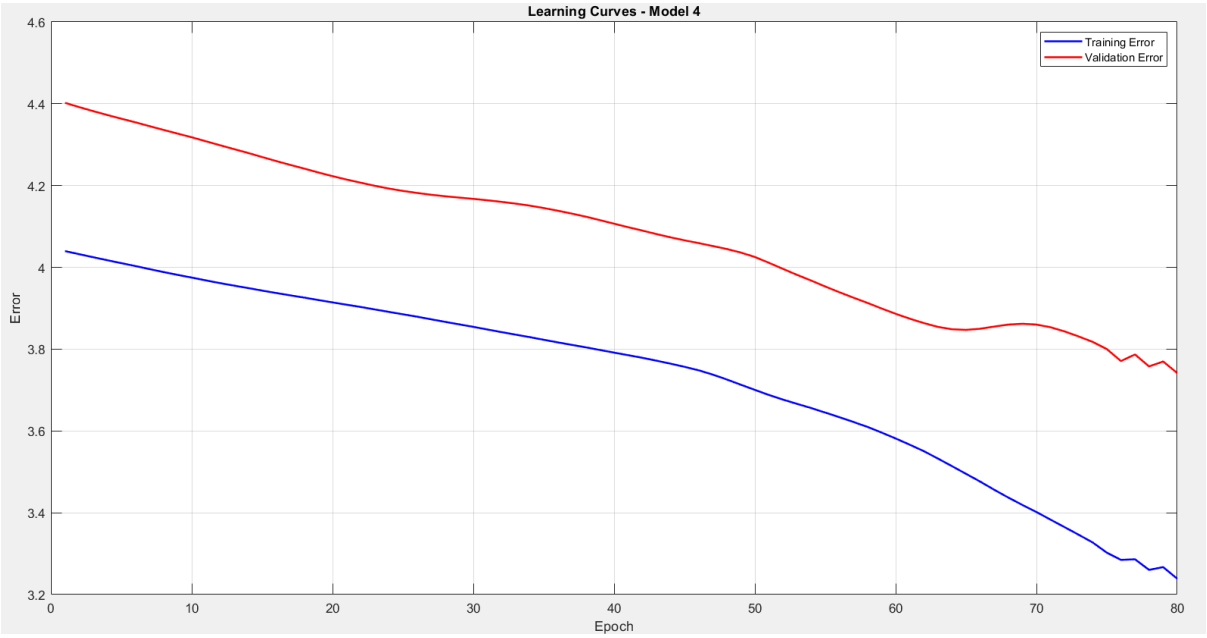




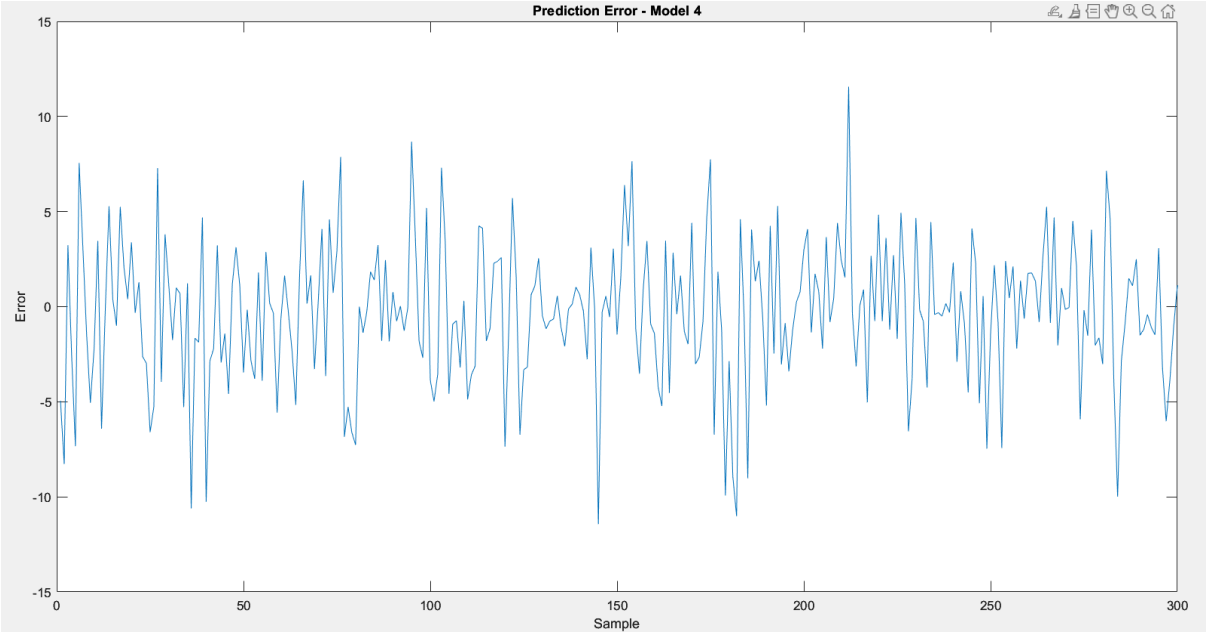
Model 4: Linear Output – 3 MF

Το πιο ευέλικτο μοντέλο: 3 MF και γραμμική έξοδος. Παρουσίασε την **καλύτερη συνολική απόδοση**, με συνεχώς φθίνουσα καμπύλη σφάλματος και χαμηλό prediction error.

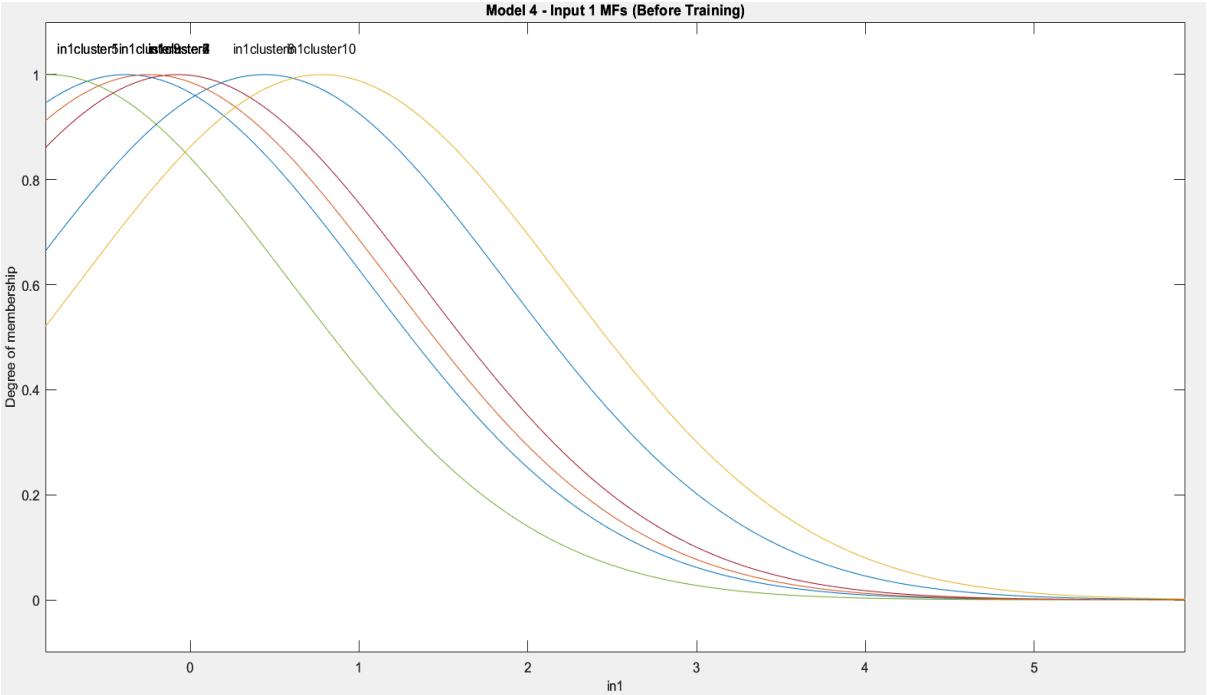
Learning Curve – Model 4

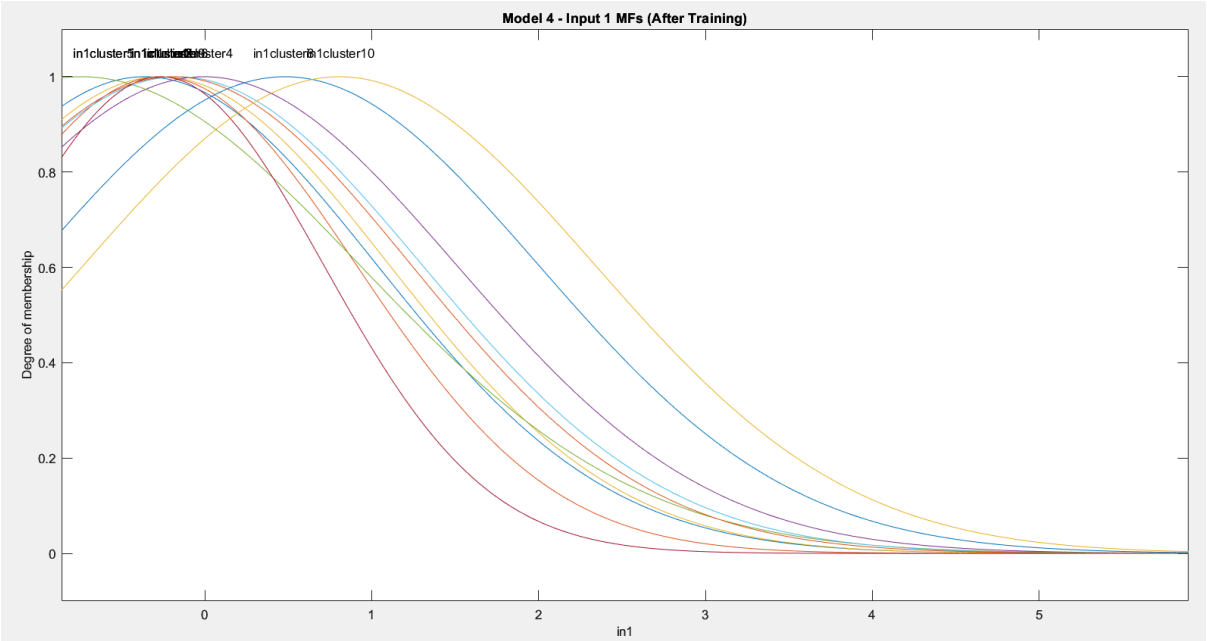


Prediction Error – Model 4



MFs Before/After Training – Model 4





2.4 Μέθοδοι αξιολόγησης και Συμπεράσματα

Για την αξιολόγηση της απόδοσης των μοντέλων TSK, χρησιμοποιήθηκαν τα παρακάτω metrics:

- RMSE (Root Mean Squared Error): Το σφάλμα ρίζας μέσου τετραγώνου, εκφράζει το μέσο σφάλμα πρόβλεψης.

- R^2 (Coefficient of Determination): Μετρά την ικανότητα του μοντέλου να εξηγεί τη μεταβλητότητα των δεδομένων.

- NMSE (Normalized Mean Squared Error): Κανονικοποιημένη μορφή του MSE, σε σχέση με τη διασπορά του στόχου.

- NDEI (Normalized Root Mean Square Error): Τετραγωνική ρίζα του NMSE, μετρά την προβλεπτική ικανότητα.

Μοντέλο	RMSE	R^2	NMSE	NDEI
1	3.9618	0.6774	0.3226	0.5680
2	4.6654	0.5527	0.4473	0.6688
3	4.1214	0.6509	0.3491	0.5909
4	3.7415	0.7123	0.2877	0.5364

Συμπεράσματα:

- Το **μοντέλο 4 (Linear, 3 MF)** επιτυγχάνει την καλύτερη γενίκευση, με **χαμηλότερο RMSE και υψηλότερο R^2**, αποδεικνύοντας ότι η χρήση περισσότερων MF και γραμμικής εξόδου ενισχύει τη μαθησιακή ικανότητα του συστήματος.

- Τα μοντέλα με **γραμμική έξοδο** (3 & 4) αποδίδουν γενικά καλύτερα από τα αντίστοιχα με **σταθερή έξοδο** (1 & 2).

- Το μοντέλο 2 με **3 σταθερές MF** έχει τη **χειρότερη απόδοση**, παρότι έχει περισσότερες συναρτήσεις συμμετοχής. Αυτό πιθανώς οφείλεται στη σταθερή μορφή εξόδου, η οποία περιορίζει την εκφραστική ικανότητα του συστήματος.

3. Β’ Μέρος: Dataset Υψηλής Διαστασιμότητας (Superconduct)

3.1 Περιγραφή Προβλήματος

Στο δεύτερο μέρος της εργασίας χρησιμοποιείται το **dataset “Superconductors”**, το οποίο προέρχεται από το UCI Machine Learning Repository. Το συγκεκριμένο dataset χρησιμοποιείται για **παλινδρόμηση**, με σκοπό την πρόβλεψη της **κρίσιμης θερμοκρασίας υπεραγώγιμων υλικών** βάσει των χημικών χαρακτηριστικών τους.

Χαρακτηριστικά:

- Περιλαμβάνει **21.263 δείγματα**.
- Κάθε δείγμα περιέχει **81 χαρακτηριστικά** (features) που περιγράφουν φυσικοχημικές ιδιότητες.
- Το τελευταίο πεδίο αφορά τη **μετρούμενη κρίσιμη θερμοκρασία** (target output).

Η κύρια πρόκληση εδώ είναι η **πολύ υψηλή διαστασιμότητα του συνόλου δεδομένων** (81 είσοδοι). Σε τέτοιες περιπτώσεις, η απευθείας χρήση Grid Partitioning ή η μη επιλεγμένη εκπαίδευση μέσω Subtractive Clustering δεν είναι εφικτή ή απαιτεί υπερβολικά μεγάλο υπολογιστικό κόστος.

Για την αντιμετώπιση αυτής της πρόκλησης, αξιοποιήθηκε τεχνική **προεπεξεργασίας μέσω επιλογής χαρακτηριστικών** (feature selection), ώστε να βελτιωθεί η αποδοτικότητα και η γενίκευση του μοντέλου.

3.2 Επιλογή Χαρακτηριστικών

Η χρήση του αλγορίθμου **ReliefF** κρίθηκε απαραίτητη λόγω της **υψηλής διαστασιμότητας** του dataset. Ο ReliefF είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος επιλογής χαρακτηριστικών, ο οποίος εκτιμά τη **σημαντικότητα κάθε χαρακτηριστικού** βάσει της ικανότητάς του να διακρίνει μεταξύ κοντινών παραδειγμάτων διαφορετικών εξόδων (targets).

Στην παρούσα εργασία, ο ReliefF εφαρμόστηκε πάνω στο εκπαιδευτικό σύνολο (training set), με σκοπό την επιλογή ενός υποσυνόλου από τα 81 διαθέσιμα χαρακτηριστικά. Συγκεκριμένα:

Ο αλγόριθμος καλείται με τη συνάρτηση:

[Index, ~] = relieff(trnData(:,1:end-1), trnData(:,end), 10);

επιστρέφει τα **index** των χαρακτηριστικών ταξινομημένα κατά σημαντικότητα.

Η επιλογή του πλήθους των χαρακτηριστικών (π.χ. 5, 10, 15, 20, 25) γίνεται σε επόμενο βήμα, κατά τη **διαδικασία Grid Search**, ώστε να διερευνηθεί η επίδοση του μοντέλου συναρτήσει του αριθμού των εισόδων.

Η επιλογή χαρακτηριστικών όχι μόνο μειώνει το υπολογιστικό κόστος, αλλά συμβάλλει και στην **αποφυγή υπερπροσαρμογής (overfitting)**, καθώς εξαλείφει τις μη σχετικές ή θορυβώδεις εισόδους.

3.3 Grid Search για Βελτιστοποίηση Μοντέλου

Για τον προσδιορισμό του **βέλτιστου συνδυασμού παραμέτρων** ενός TSK συστήματος πάνω σε δεδομένα υψηλής διαστασιμότητας, εφαρμόστηκε μέθοδος **Grid Search με 5-Fold Cross Validation**.

Παράμετροι που εξετάστηκαν:

Cluster Influence Radius (`r_a`): επηρεάζει τη δημιουργία κανόνων στο Subtractive Clustering. Τιμές: {0.2,0.3,0.4,0.5,0.6}

Αριθμός επιλεγμένων χαρακτηριστικών: {5,10,15,20,25}

Αυτά προκύπτουν από τα κορυφαία χαρακτηριστικά του πίνακα `Index` του ReliefF.

Διαδικασία:

1. **5-Fold CV** στο εκπαιδευτικό σύνολο.
2. Για κάθε συνδυασμό (r_a, \text{#features}):
    - Επιλέγονται τα αντίστοιχα χαρακτηριστικά.
    - Δημιουργείται FIS μέσω `genfis` με Subtractive Clustering.
    - Εκπαιδεύεται μέσω `anfis`.
    - Καταγράφεται το **σφάλμα επικύρωσης (validation error)** για κάθε fold.
    - Υπολογίζεται ο **μέσος όρος** των 5 σφαλμάτων.
3. Ο συνδυασμός με **το ελάχιστο μέσο σφάλμα επικύρωσης** επιλέγεται ως **βέλτιστος**.

Βέλτιστες Τιμές Παραμέτρων

Η ελάχιστη μέση τιμή σφάλματος παρατηρήθηκε για:

- **Ακτίνα Clustering (r_a)**: `0.4`
- **Αριθμός χαρακτηριστικών**: `25`

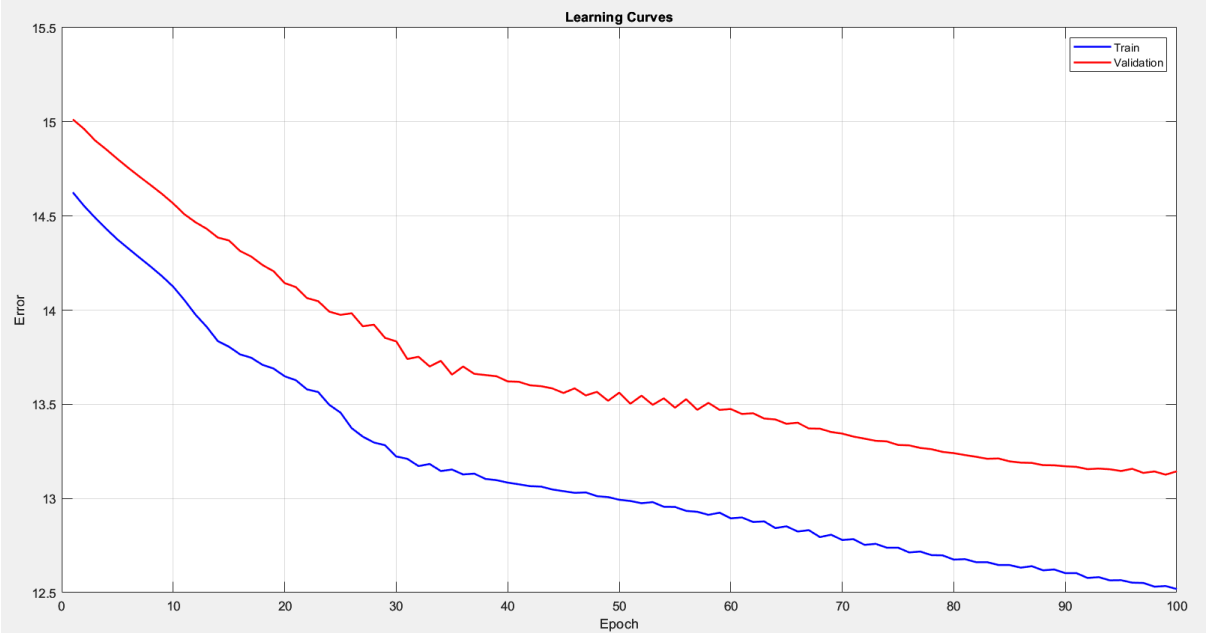
Αυτές οι τιμές χρησιμοποιήθηκαν για την **τελική εκπαίδευση** του μοντέλου.

3.4 Τελική Εκπαίδευση του Βέλτιστου Μοντέλου

Μετά την ολοκλήρωση του grid search και την επιλογή των βέλτιστων παραμέτρων , πραγματοποιήθηκε η τελική εκπαίδευση ενός TSK μοντέλου.

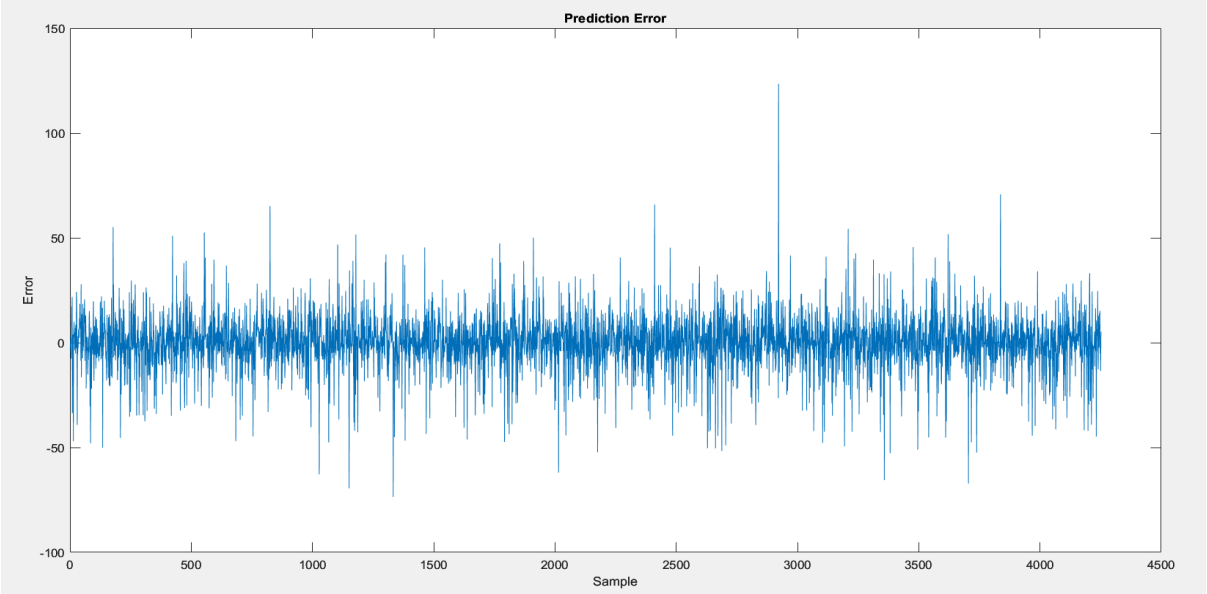
Καμπύλες Εκπαίδευσης & Επικύρωσης

Η εξέλιξη του σφάλματος κατά τη διάρκεια των 100 epochs φαίνεται παρακάτω. Παρατηρείται σαφής μονοτονική μείωση τόσο στο training όσο και στο validation error, με ελαφρώς πιο αργή αποκλιμάκωση στο validation (κόκκινη γραμμή), υποδεικνύοντας ικανοποιητική εκπαίδευση χωρίς σημαντική υπερπροσαρμογή.



Σφάλμα Πρόβλεψης

Το διάγραμμα prediction error παρουσιάζει τις αποκλίσεις μεταξύ προβλεπόμενων και πραγματικών τιμών στο validation set (~4,250 δείγματα). Η κατανομή φαίνεται ισορροπημένη γύρω από το μηδέν με μερικές εξάρσεις, που είναι αναμενόμενες σε πραγματικά δεδομένα.



Δείκτες Απόδοσης

Μετά το τέλος του training, οι τιμές των δεικτών είναι:

Δείκτης	Τιμή
RMSE	13.1264
R^2	0.8567
NMSE	0.1433
NDEI	0.3785

Το μοντέλο έχει **υψηλή ακρίβεια** ($R^2 \approx 0.86$), **χαμηλό NMSE**, και **NDEI < 0.5**, δείχνοντας σταθερή απόδοση.

Συμπερασματικά

Η συνδυαστική χρήση **ReliefF + Subtractive Clustering + ANFIS** απέδωσε ένα TSK μοντέλο ικανό να εξάγει πληροφορία από υψηλής διάστασης δεδομένα.
Οι καμπύλες εκπαίδευσης δείχνουν σταθερή μείωση σφάλματος χωρίς απότομες αναστροφές, ενώ το validation παραμένει κοντά στο training.
Η τελική απόδοση ($R^2 \approx 0.86$, NDEI ≈ 0.38) δείχνει ότι το μοντέλο γενικεύει καλά τα δεδομένα, ακόμα και με μεγάλο πλήθος χαρακτηριστικών.