ΕΡΓΑΣΙΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΝΟΗΜΟΣΥΝΗΣ - ΑΣΑΦΩΝ ΣΥΣΤΗΜΑΤΩΝ

Επίλυση προβλήματος παλινδρόμησης με χρήση μοντέλων TSK

Μπούζιος Κωνσταντίνος
ΑΕΜ: 8957
Αναφορά τρίτης εργασίας Regression
email: kmpouzio@ece.auth.gr
Εαρινό Εξάμηνο 2023

Περιεχόμενα

1. Εφαρμογή σε απλό Dataset……3
   1.2 Αποτελέσματα των τεσσάρων μοντέλων…….4
      1.2.1 TSK MODEL 1……...4
      1.2.2 TSK MODEL 2……...9
      1.2.3 TSK MODEL 3……14
      1.2.4 TSK MODEL 4……19
      1.2.5 Συγκεντρωτικός πίνακας……23
   1.3 Συμπεράσματα……...24

2. Εφαρμογή σε Dataset με υψηλή διαστασιμότητα…27
   2.1 Αποτελέσματα……27
      2.1.1 Membership plots……..28
      2.1.2 Καμπύλη μάθησης……31
      2.1.3 Πραγματικές και τιμές πρόβλεψης……32
      2.1.4 Συγκεντρωτικός πίνακας……32
   2.2 Συμπεράσματα……32

1. Εφαρμογή σε απλό Dataset

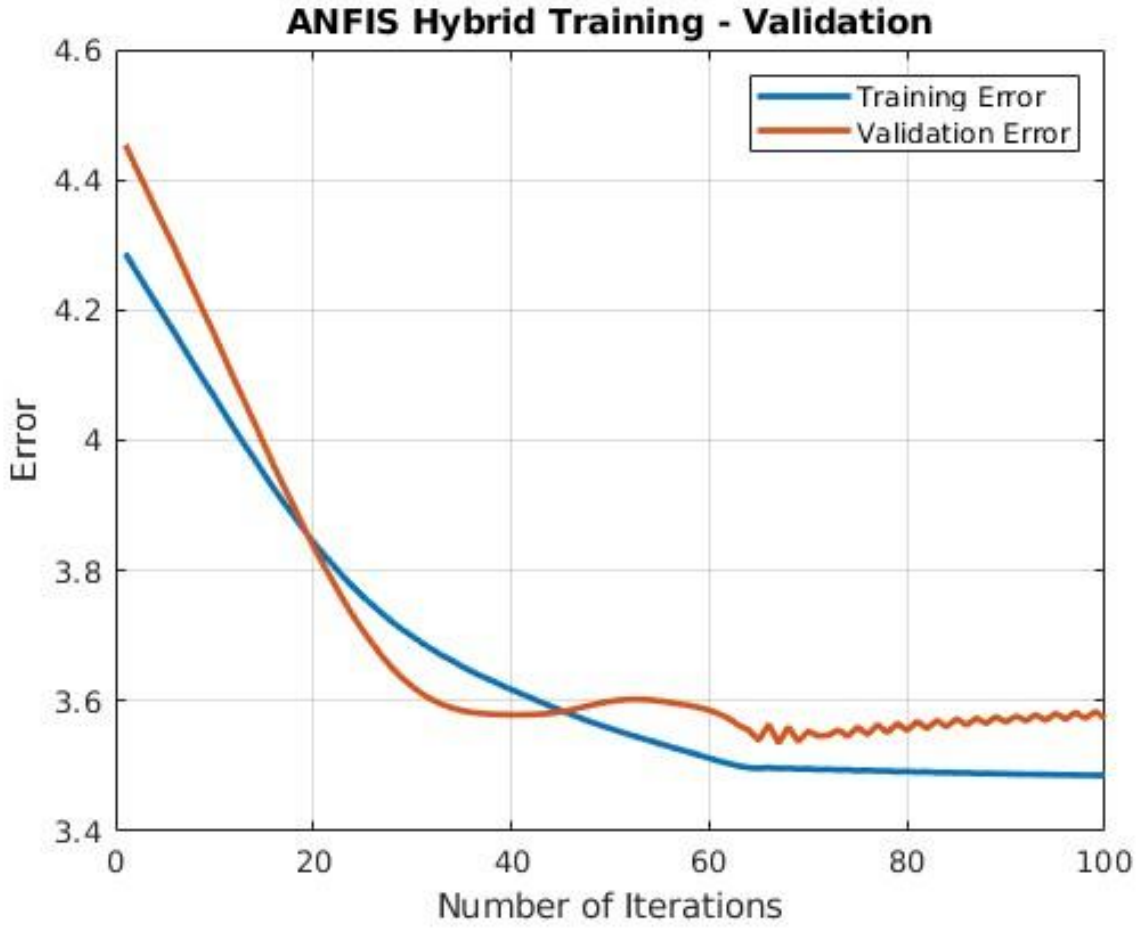
Αρχικά εφαρμόζονται τα μοντέλα TSK σε ένα απλό πολυμεταβλητό Dataset που έχει 6 χαρακτηριστικά. Το script όπου υλοποιείται η εφαρμογή των TSK ονομάζεται regression1.m. Πρώτα από όλα, φορτώνονται τα δεδομένα airfoil_self_noise μέσω της εντολής load. Έπειτα μέσω της έτοιμης συνάρτησης που υπάρχει στο e-learning χωρίζονται σε δεδομένα εκπαίδευσης (60%), δεδομένα επικύρωσης (20%) και σε δεδομένα ελέγχου απόδοσης. Σε αυτήν την έτοιμη συνάρτηση split_scale.m επιλέγεται μέθοδος προεπεξεργασίας το Normalization to unit hypercube. Σε αυτό το στάδιο ορίζονται δύο συναρτήσεις, η μία για την R^2 και η άλλη για το NMSE.

Ύστερα από αυτό, αρχικοποιείται ο πίνακας Fis με τα τέσσερα διαφορετικά μοντέλα πάνω στα οποία γίνεται η εκπαίδευση, η επικύρωση και ο έλεγχος του Dataset. Στην συνέχεια, γίνεται η εκπαίδευση και η επικύρωση μέσω της εντολής anfis του matlab. Τέλος, βγαίνουν τα χρήσιμα σχήματα όπως τα membership plots καθώς και οι καμπύλες μάθησης.

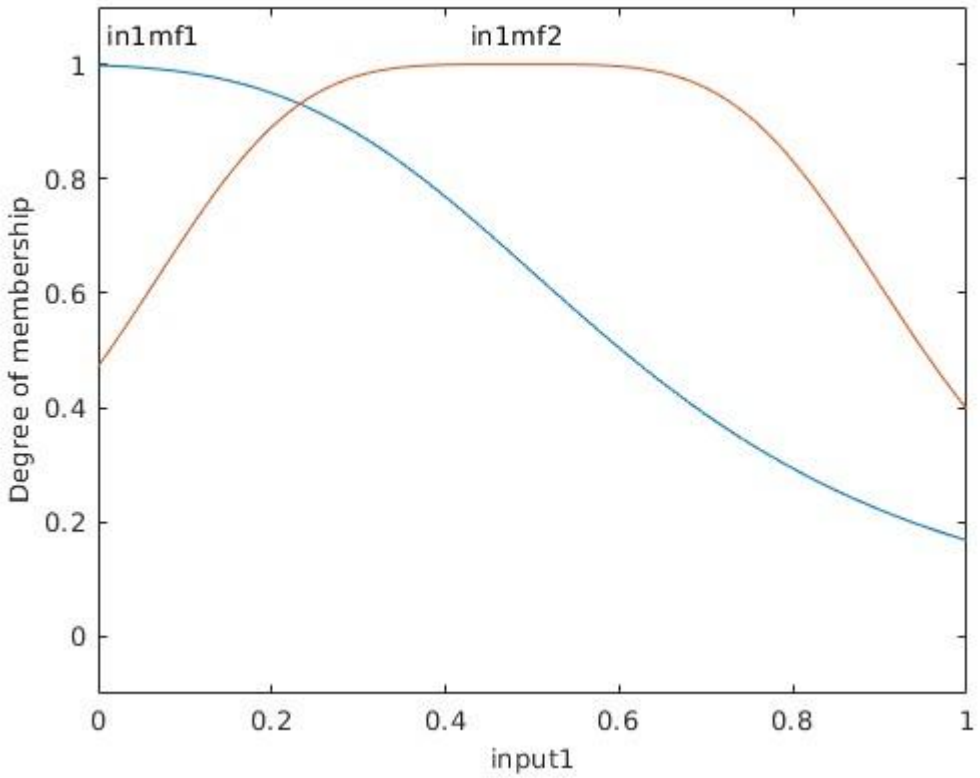
Παρακάτω απεικονίζονται τα αποτελέσματα της διαδικασίας.

1.2 Αποτελέσματα των τεσσάρων μοντέλων

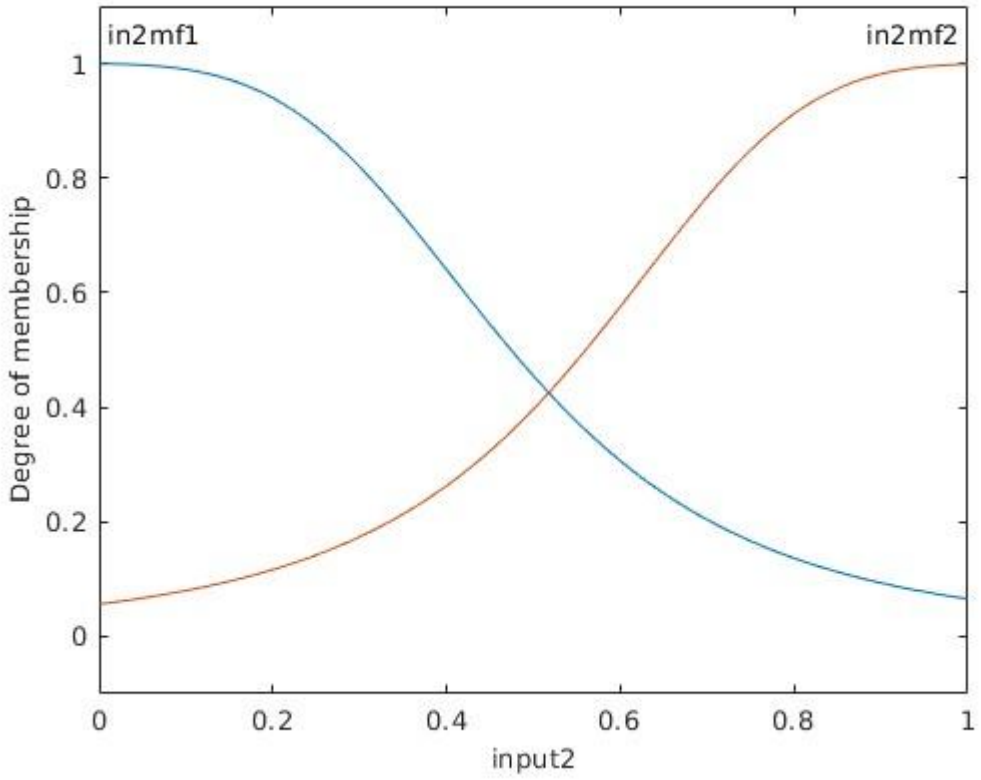
Για το μοντέλο TSK model 1



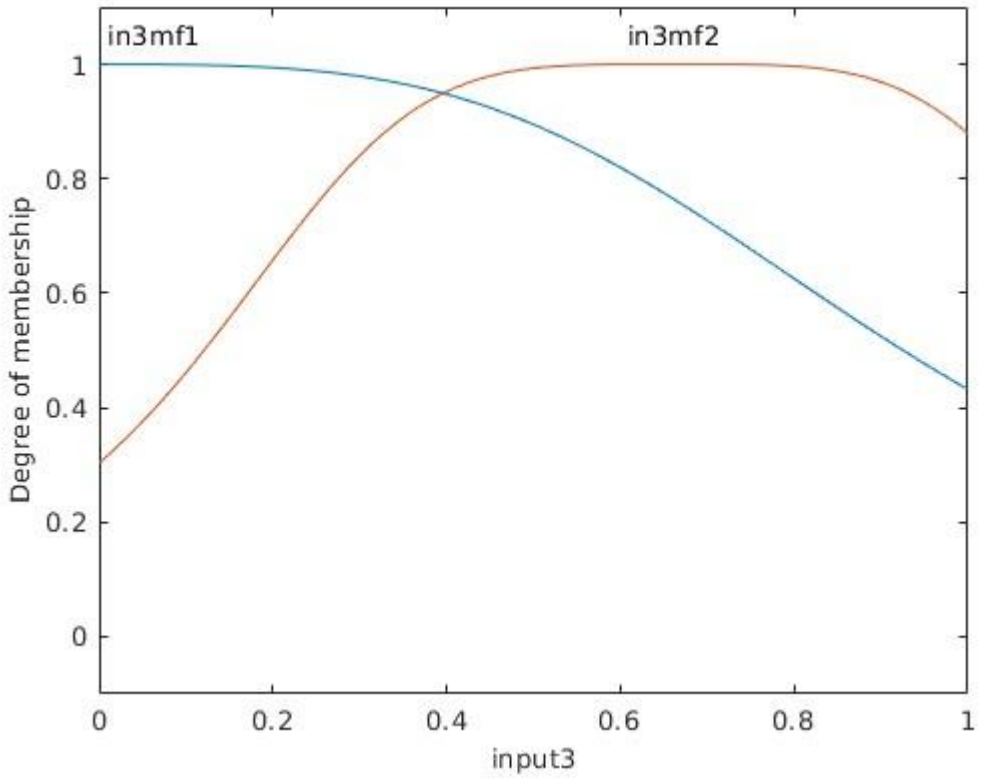
Εικόνα 1 - TSK_model_1 - Καμπύλη μάθησης



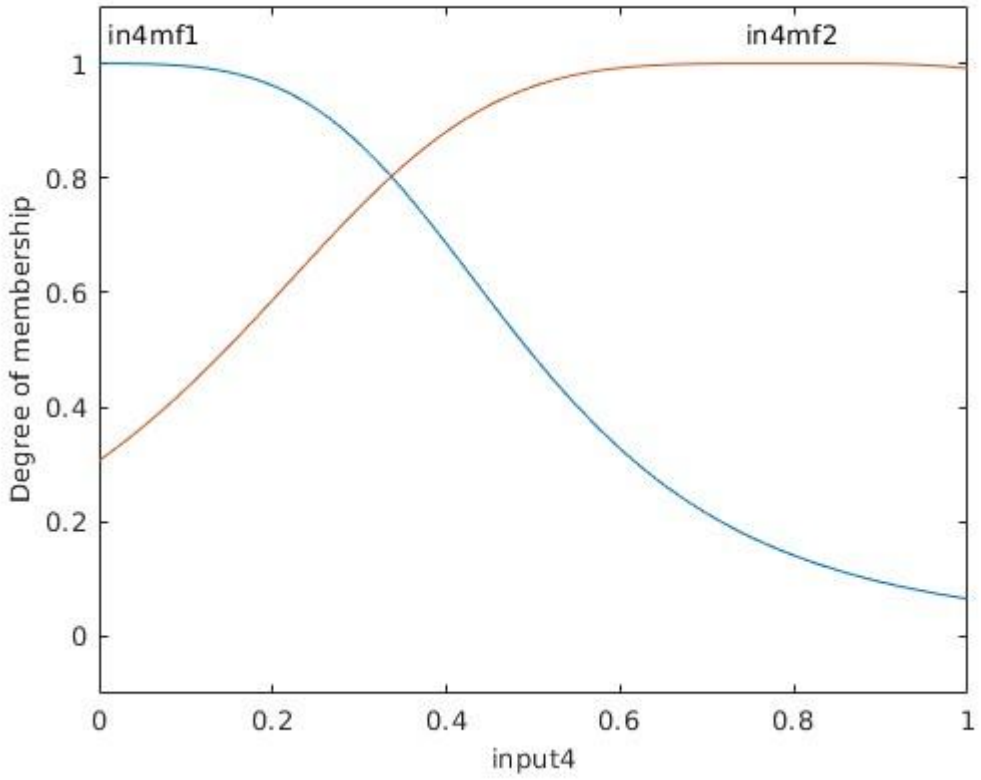
Εικόνα 2 - Membership function of Characteristic 1



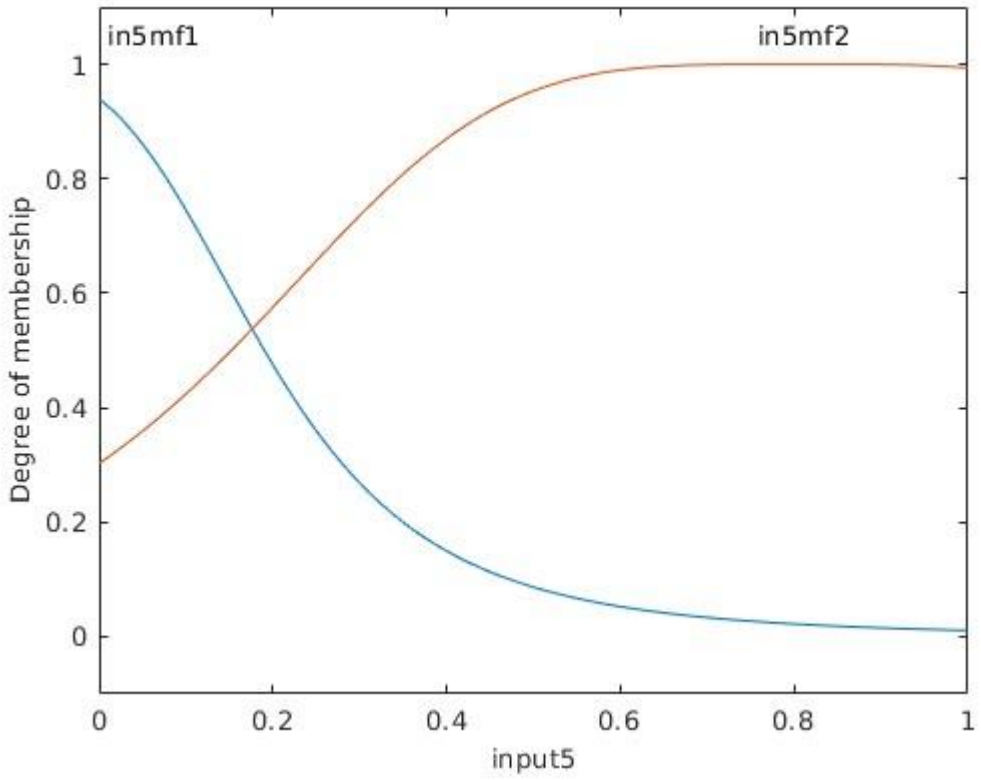
Εικόνα 3 - Membership function of Characteristic 2



Εικόνα 4 - Membership function of Characteristic 3

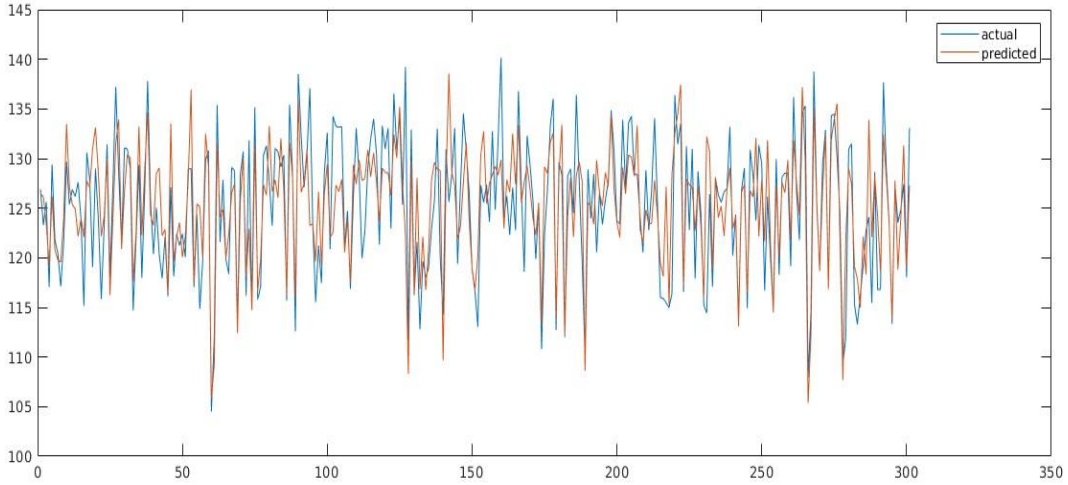


Εικόνα 5 - Membership function of Characteristic 4



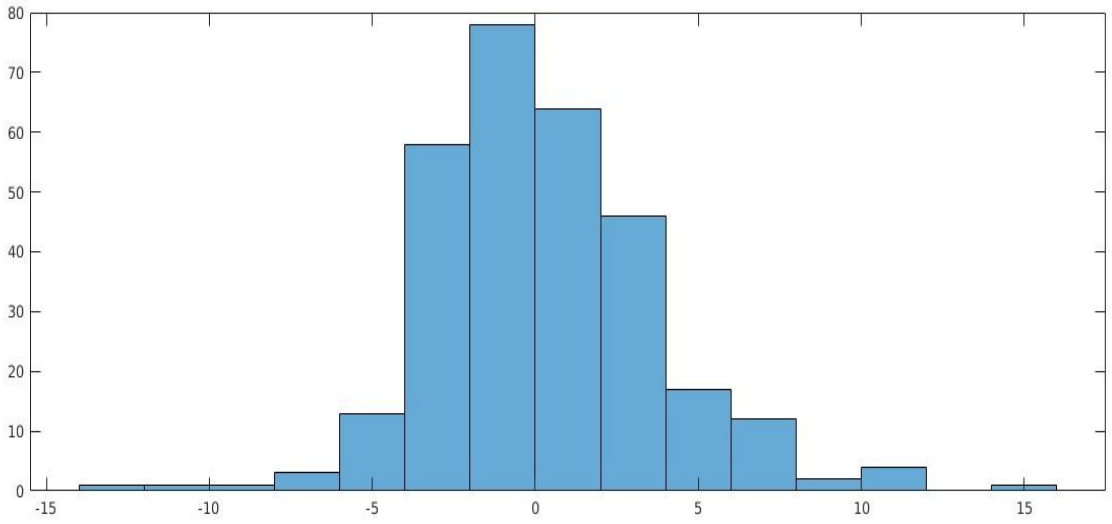
Εικόνα 6 - Membership function of Characteristic 5

Στην εικόνα 7 φαίνεται μια σύγκριση μεταξύ των πραγματικών τιμών (με μπλε) και των τιμών πρόβλεψης (με κόκκινο).



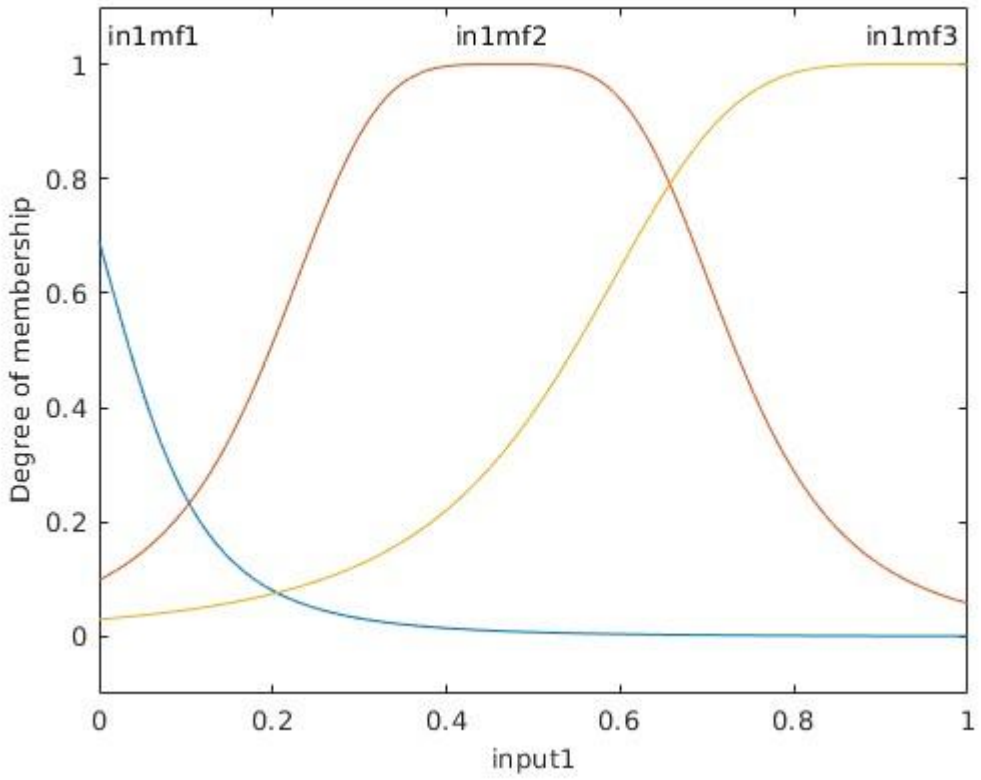
Εικόνα 7 - Πραγματικές τιμές και τιμές πρόβλεψης

Στην εικόνα 8 παρουσιάζεται το ιστόγραμμα των σφαλμάτων πρόβλεψης. Παρατηρείται ότι υπάρχει μια τάση να βγαίνουν μικρότερες από τις αναμενόμενες τιμές, εξού και οι περισσότερες τιμές είναι αρνητικές

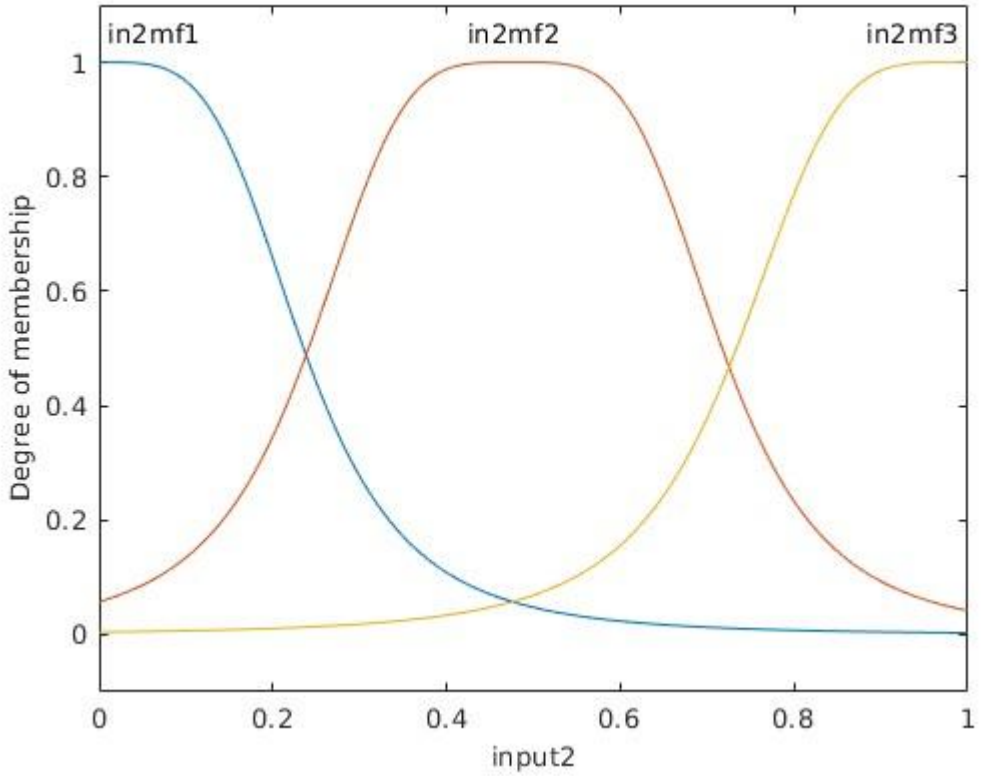


Εικόνα 8. Ιστόγραμμα σφαλμάτων πρόβλεψης

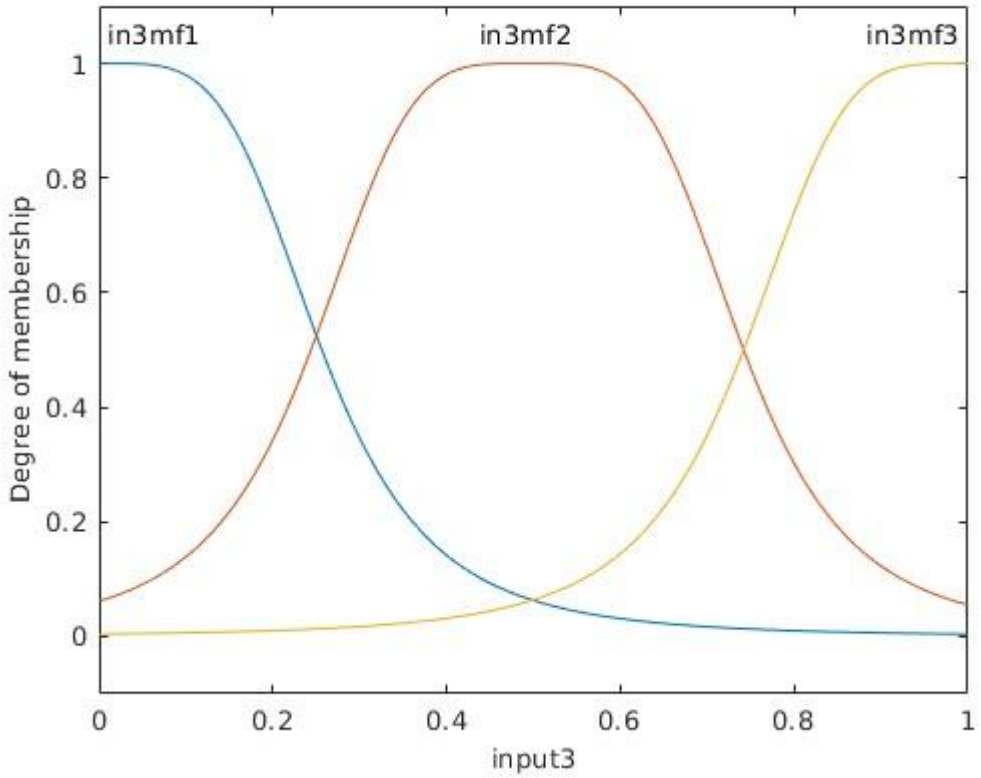
Για το TSK_model_2



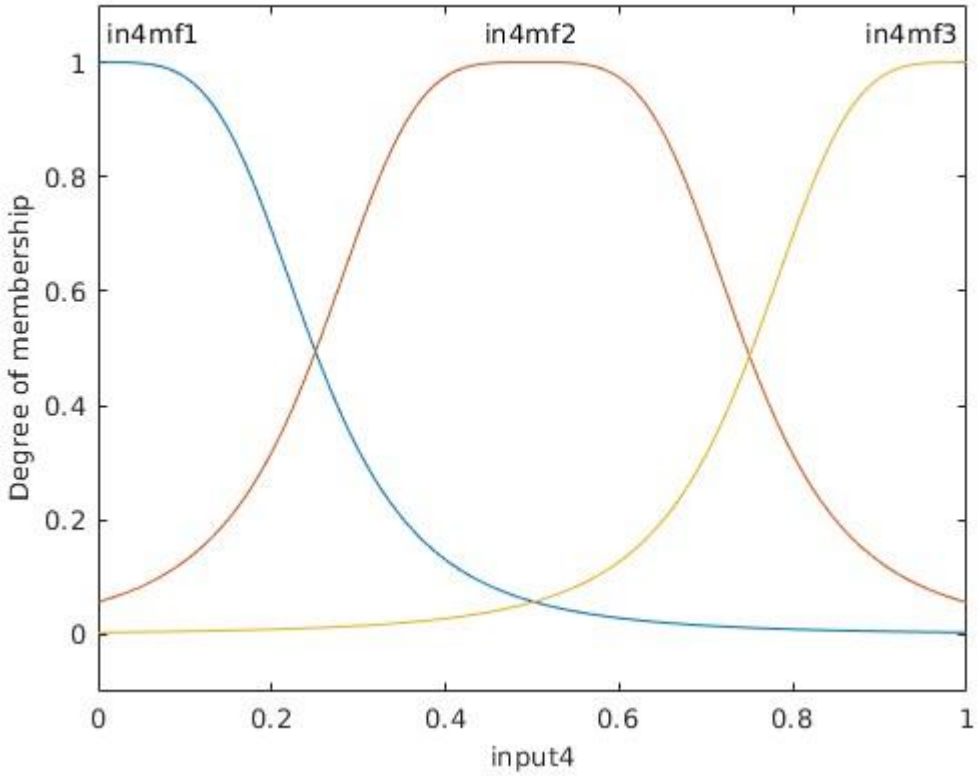
Εικόνα 9. Membership function for input 1



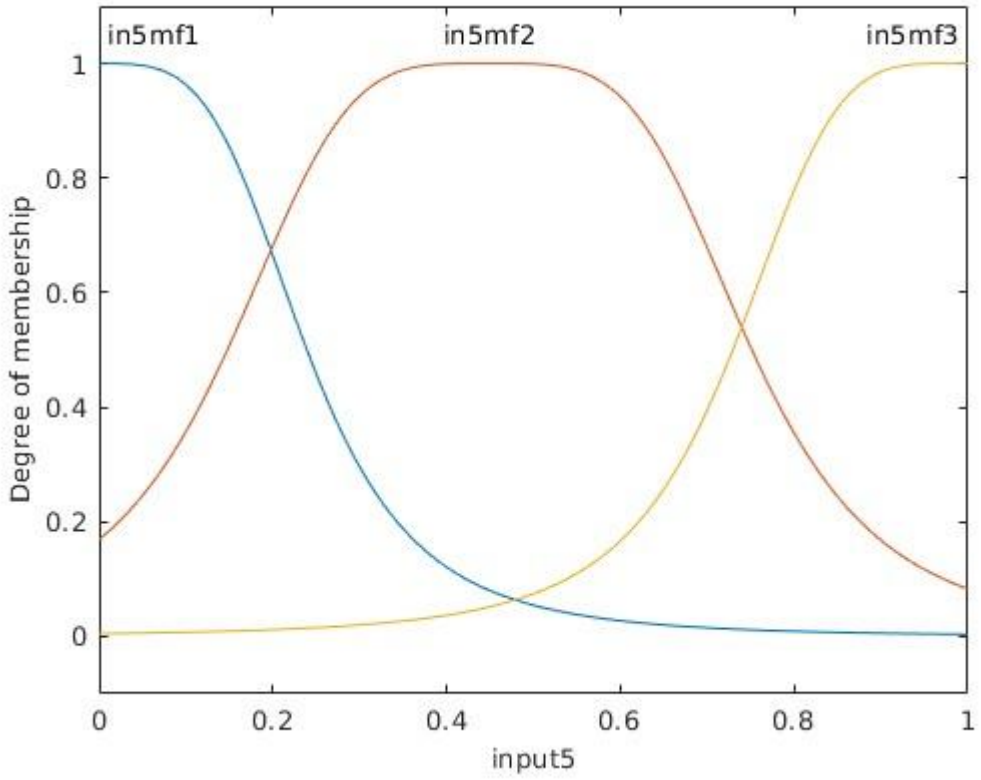
Εικόνα 10. Membership function for input 2



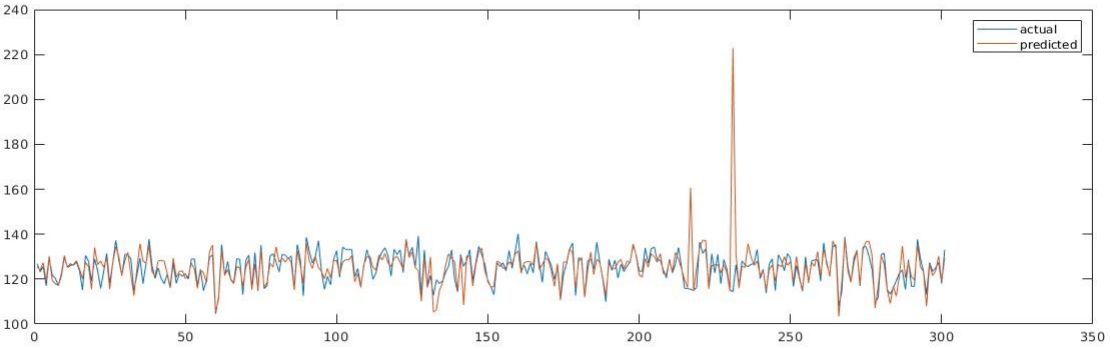
Εικόνα 11. Membership function for input 3



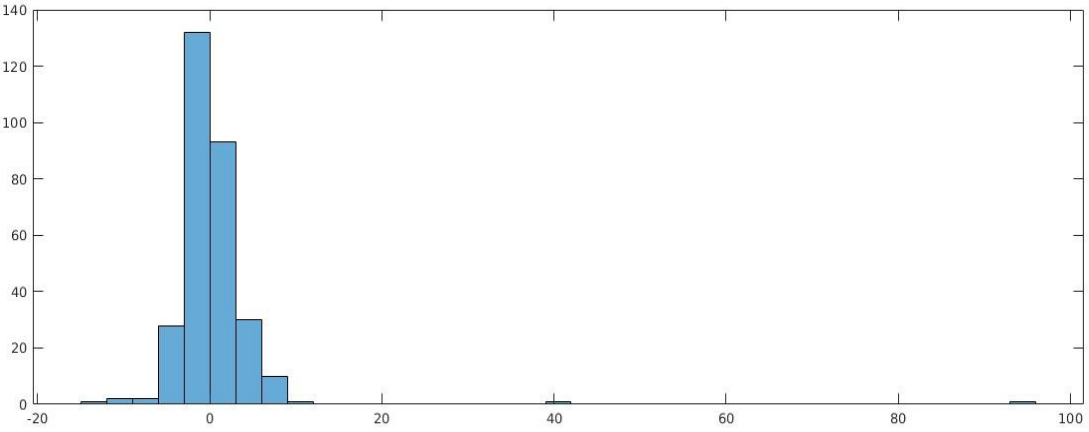
Εικόνα 12. Membership function for input 4



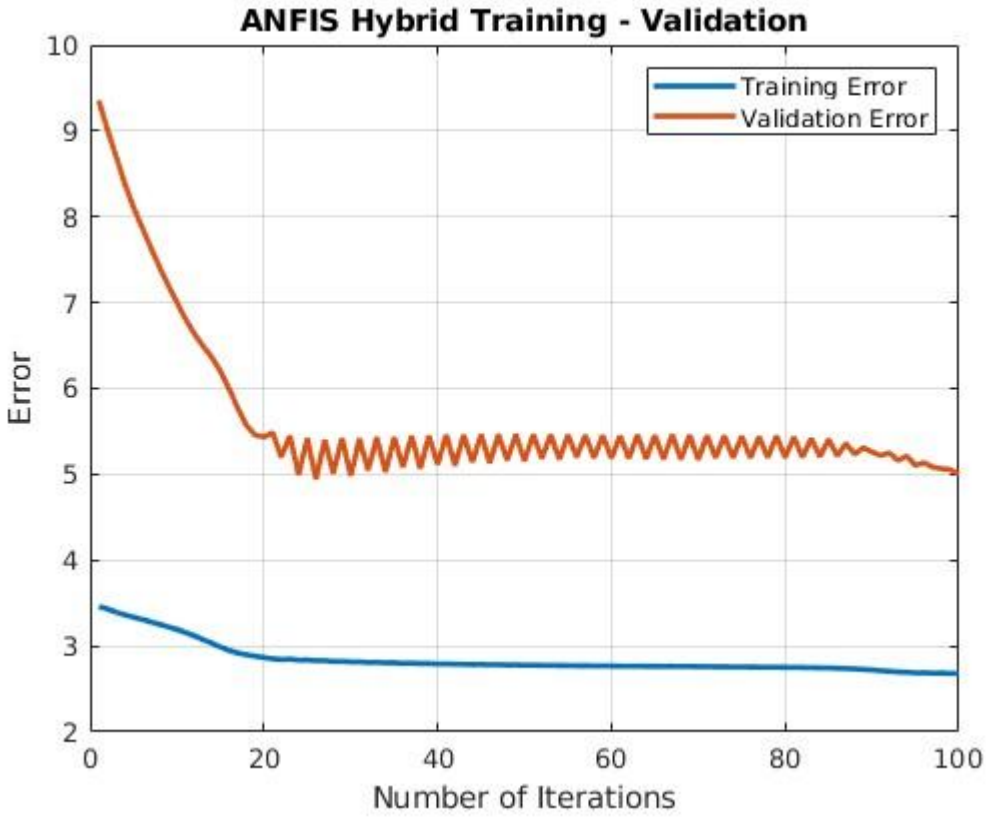
Εικόνα 13. Membership function for input 5



Εικόνα 14. Πραγματικές τιμές και τιμές πρόβλεψης

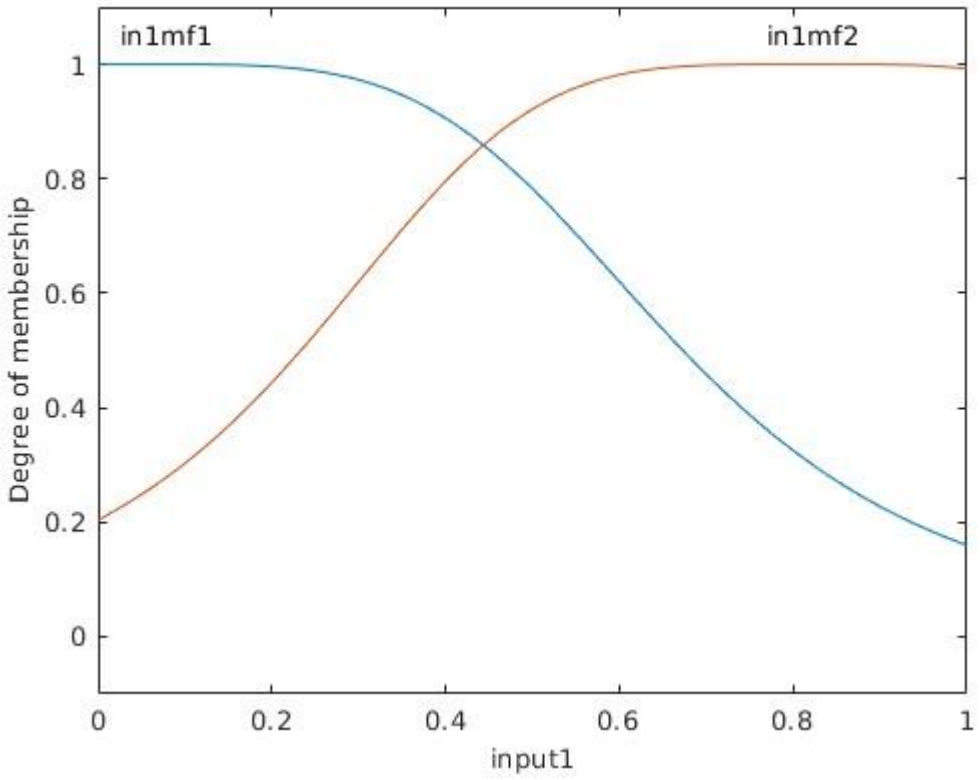


Εικόνα 15. Ιστόγραμμα σφάλματος πρόβλεψης

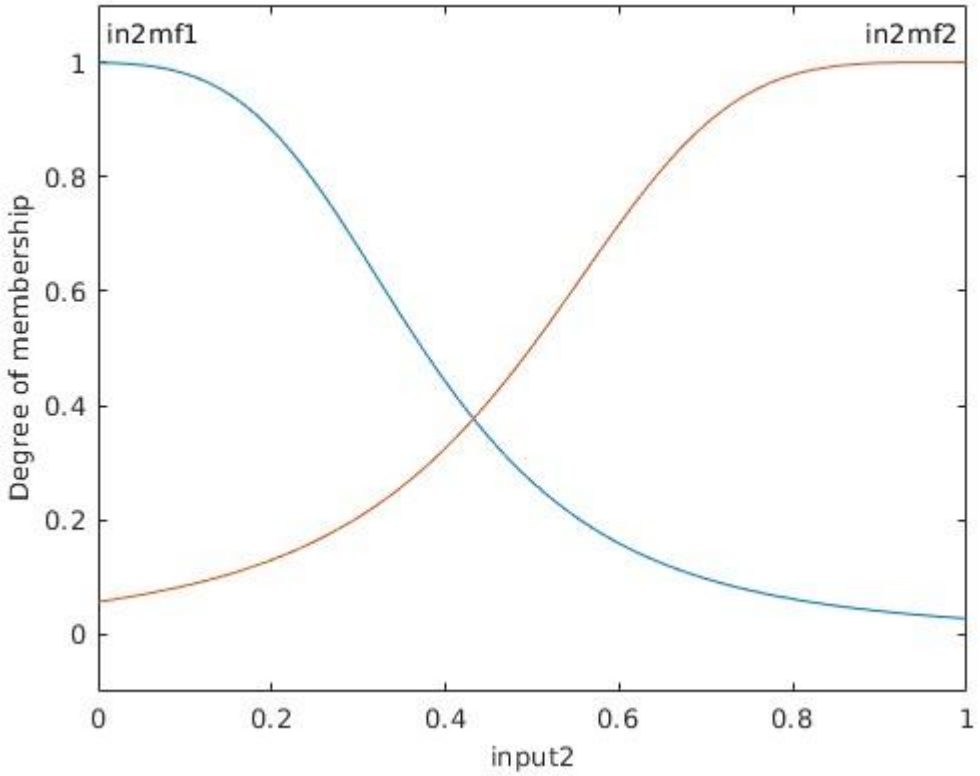


Εικόνα 15α - Καμπύλη μάθησης για το TSK_Model_2

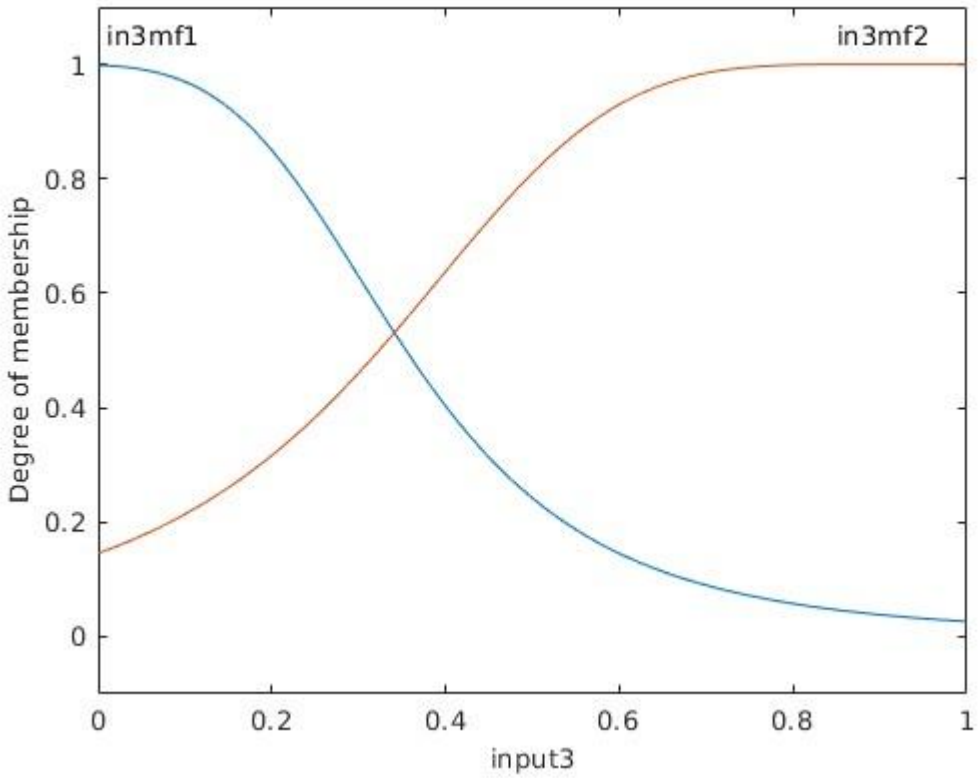
Για το TSK_model_3



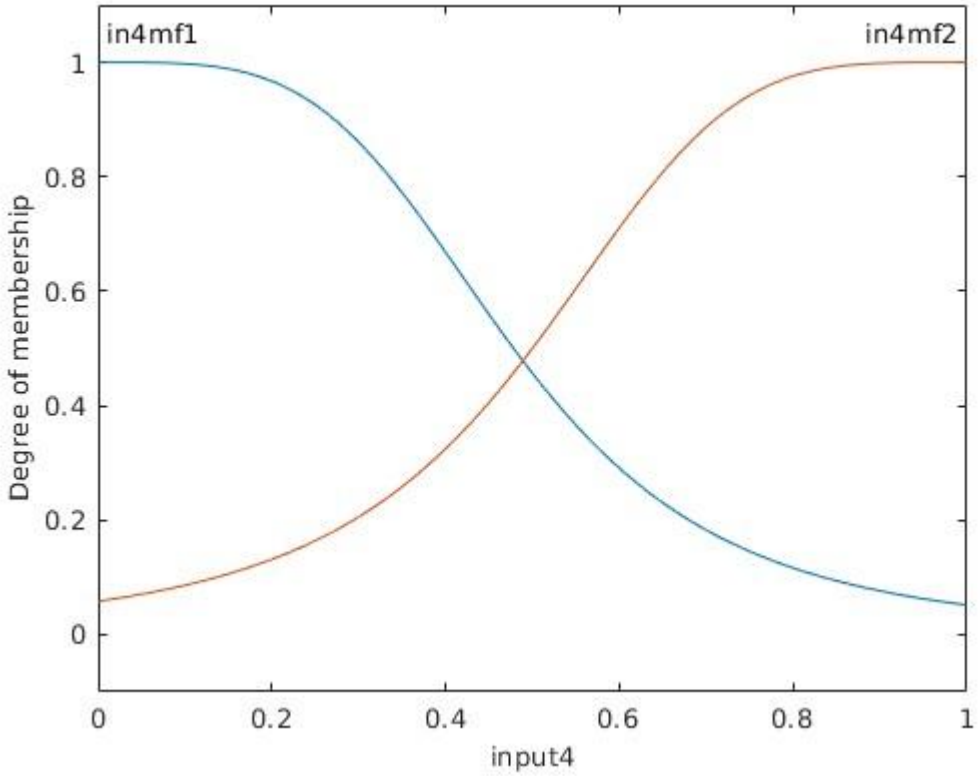
Εικόνα 16. Membership function of input 1



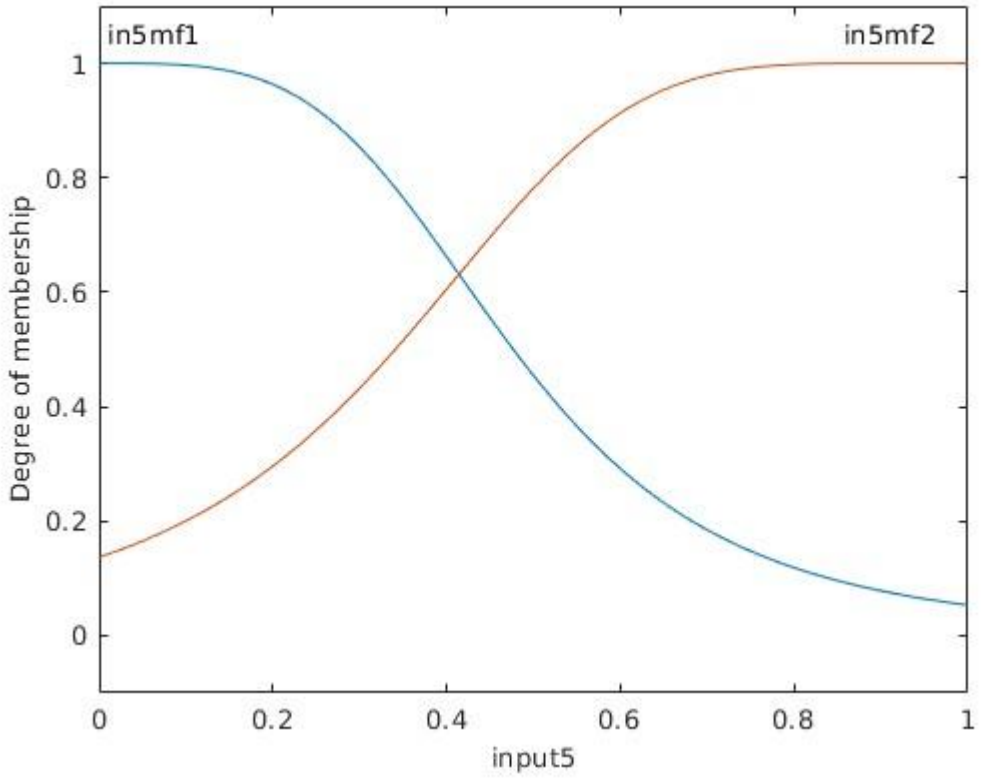
Εικόνα 17. Membership function of input 2



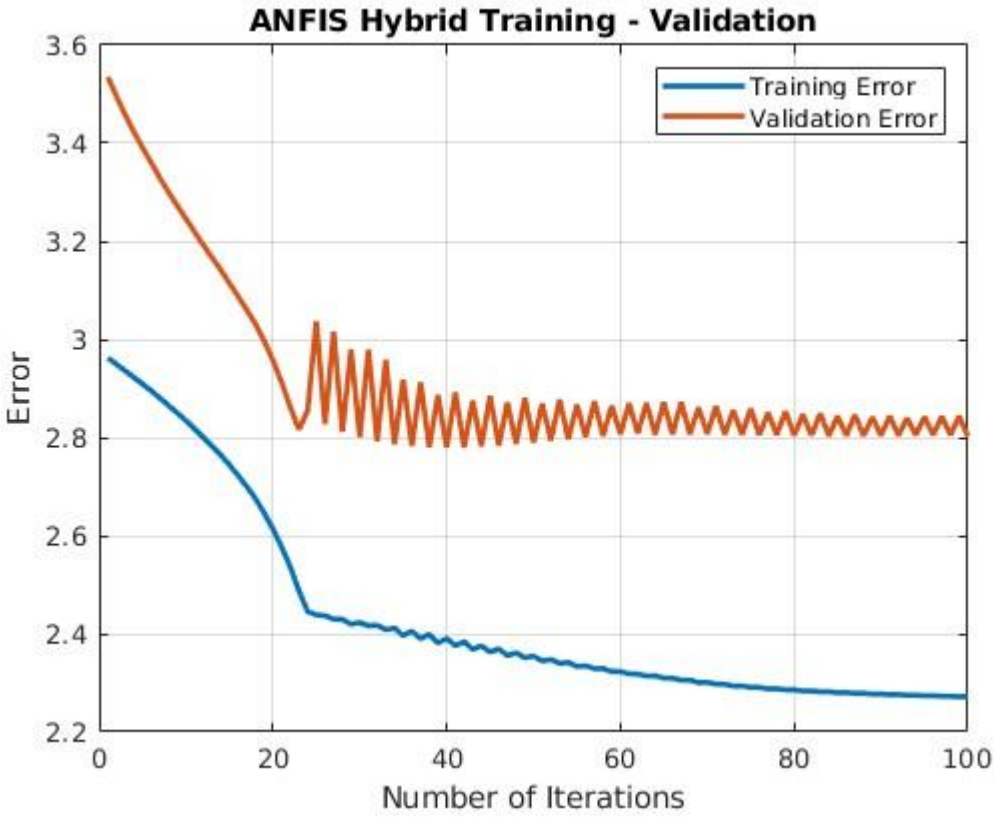
Εικόνα 18. Membership function of input 3



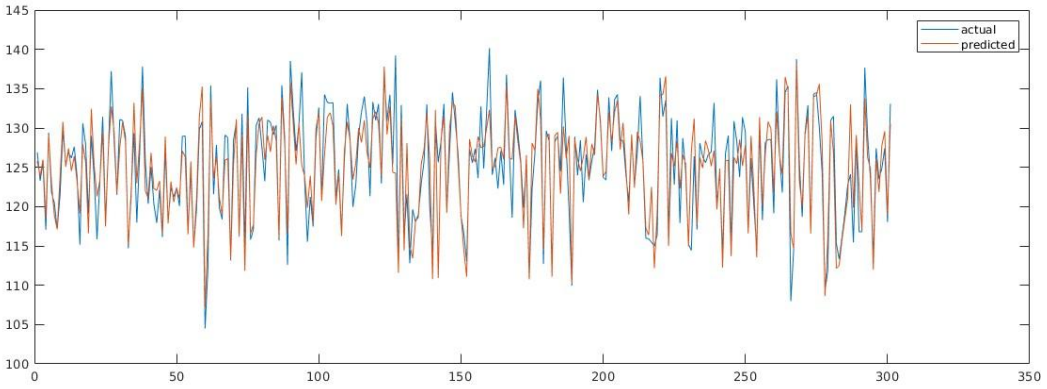
Εικόνα 19. Membership function of input 4



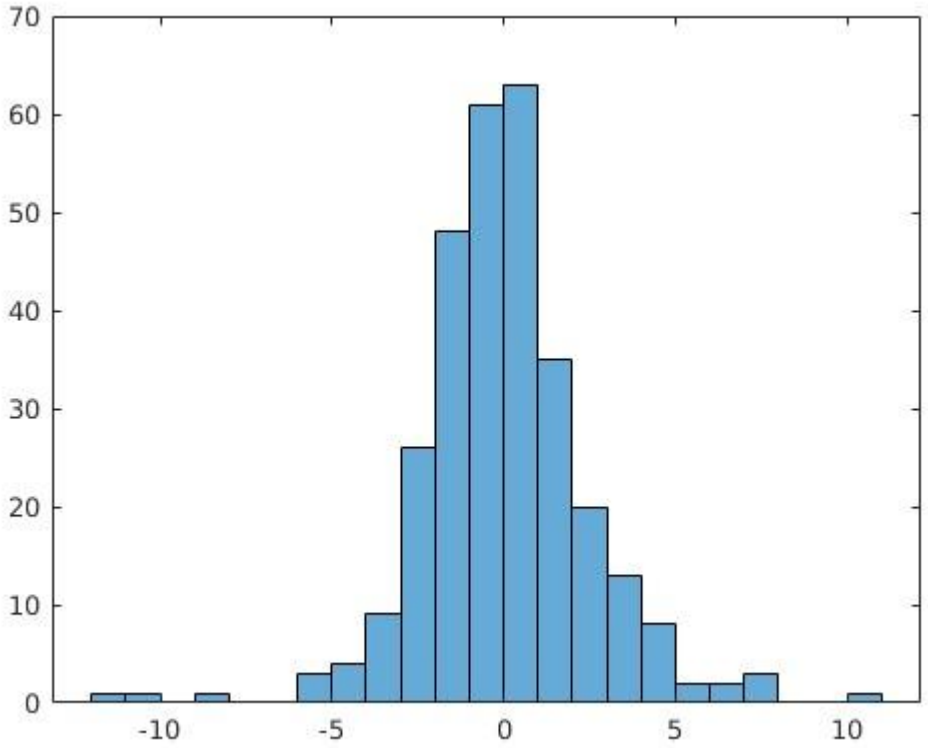
Εικόνα 20. Membership function of input 5



Εικόνα 21. Καμπύλη μάθησης για το TSK_Model_3

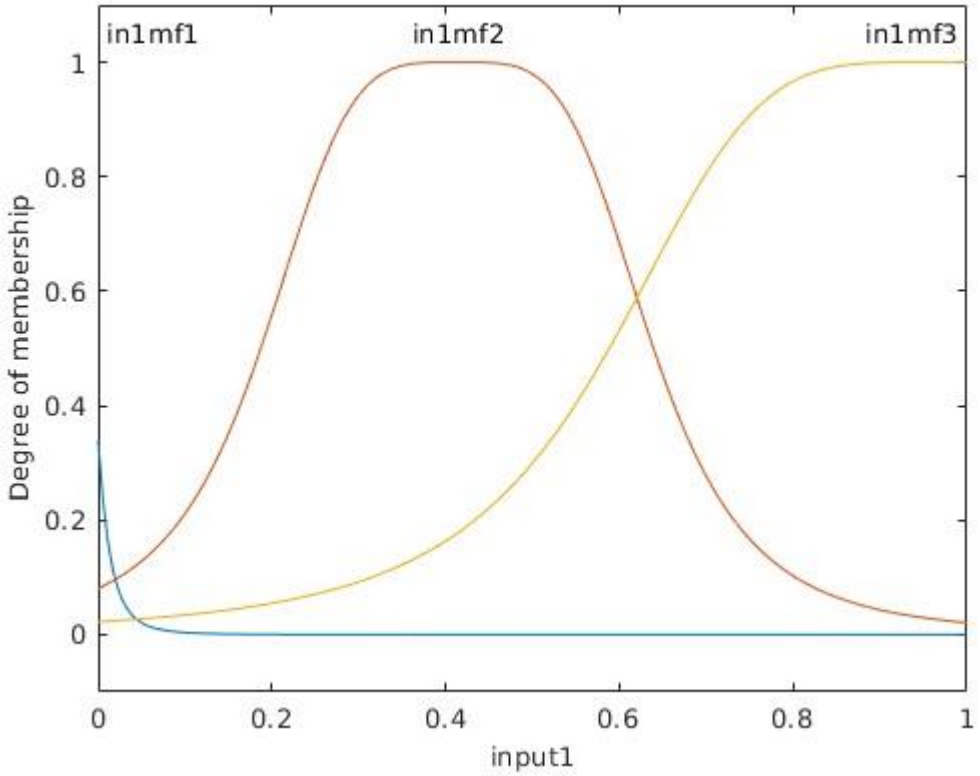


Εικόνα 22. Πραγματικές τιμές με τις τιμές πρόβλεψης

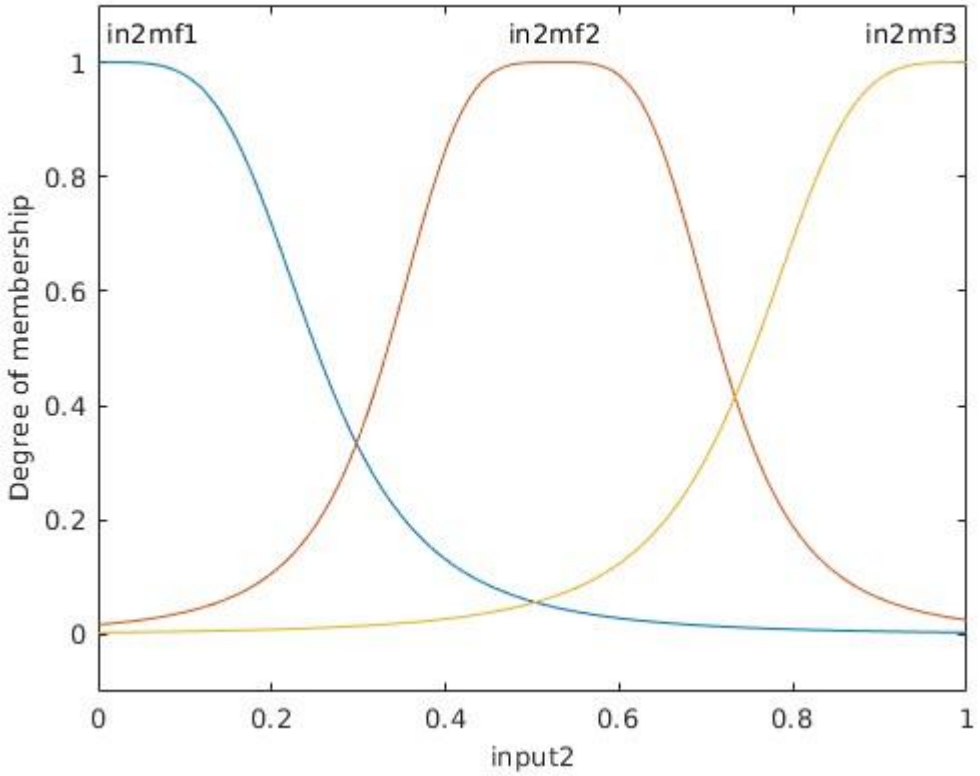


Εικόνα 23. Ιστόγραμμα σφάλματος πρόβλεψης για το TSK_Model_3

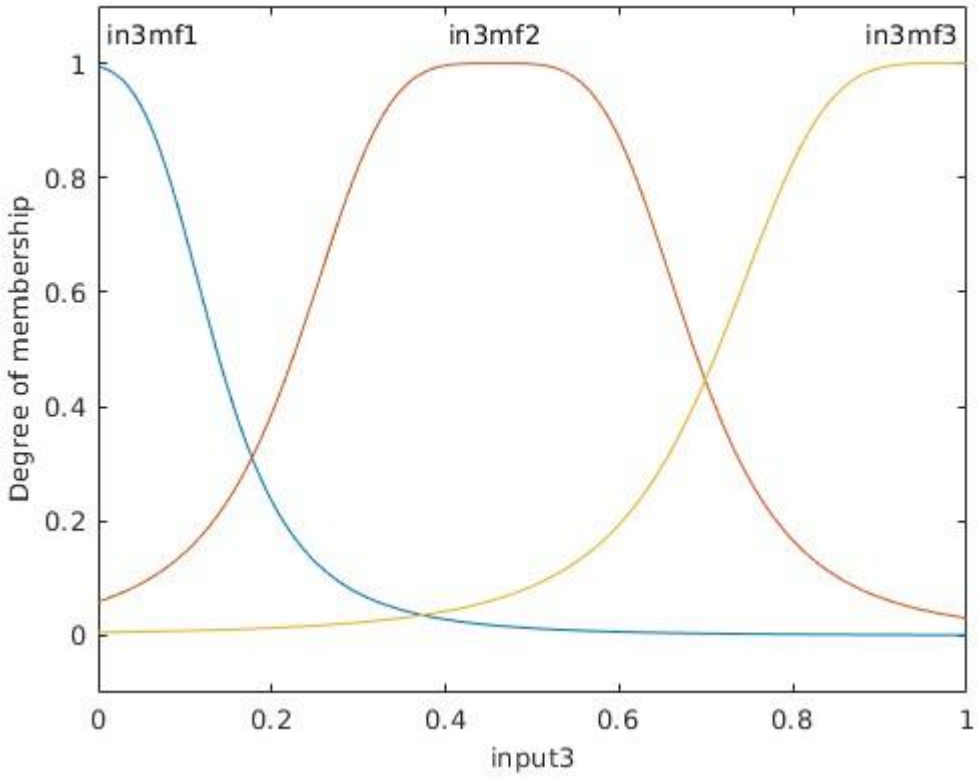
Για το TSK_model_4



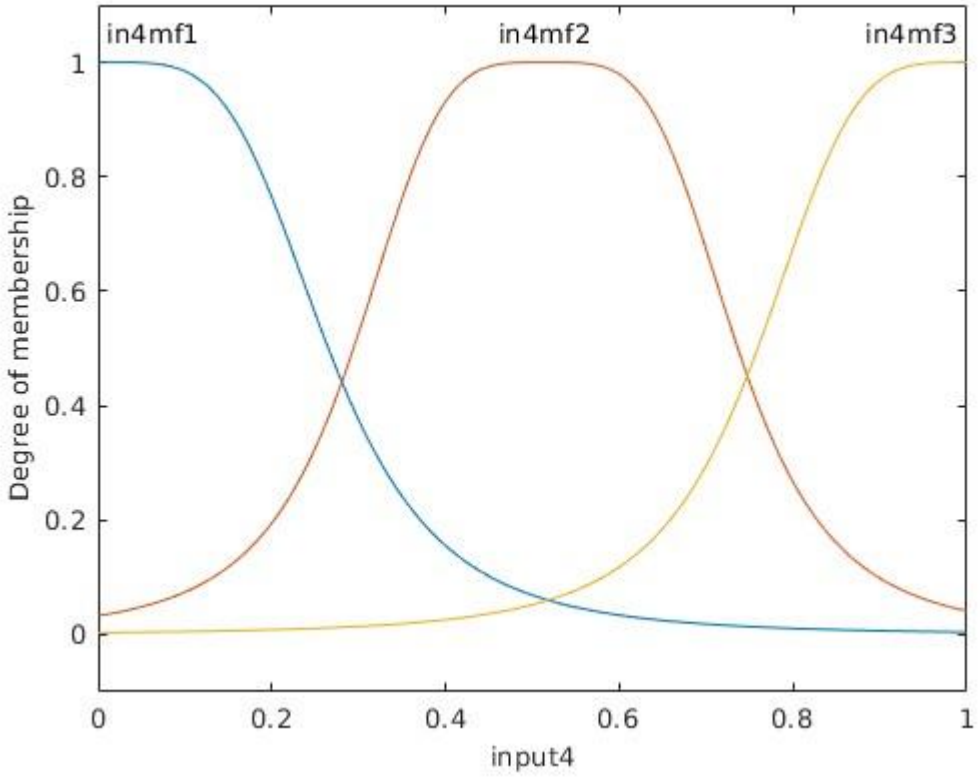
Εικόνα 24. Membership function for input 1



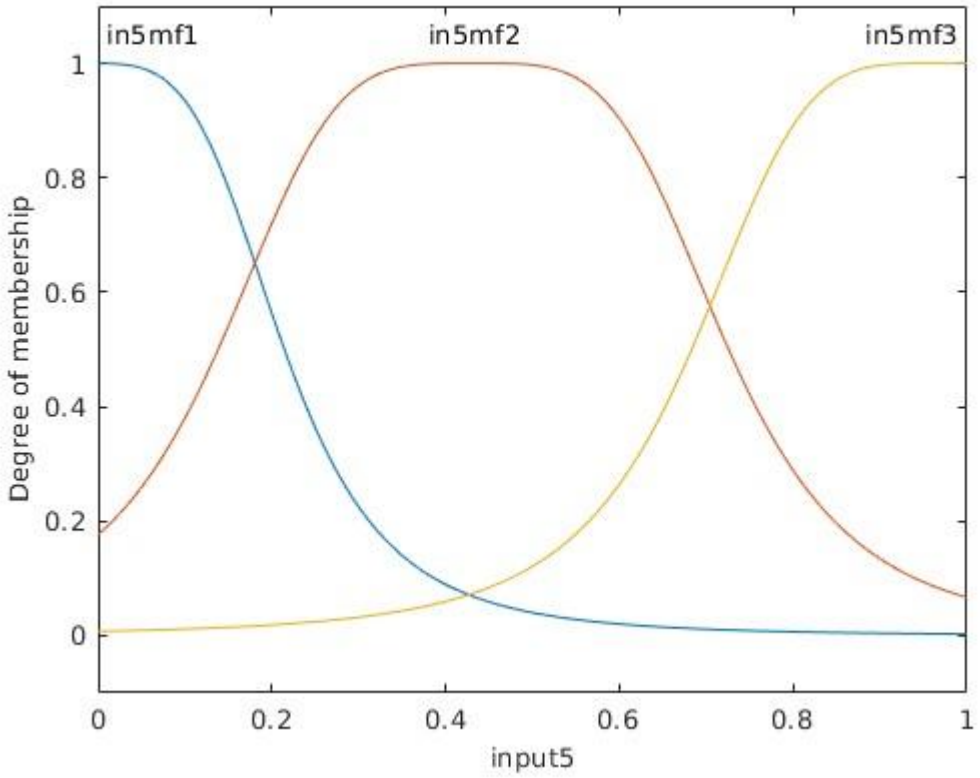
Εικόνα 25. Membership function for input 2



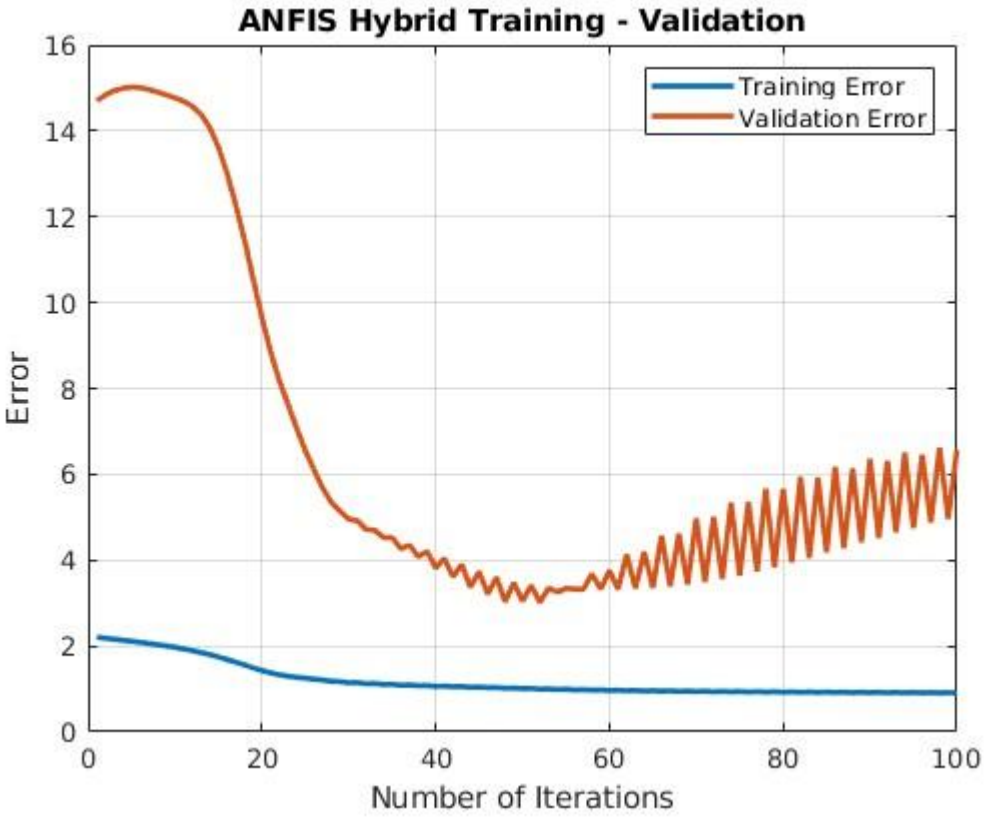
Εικόνα 26, Membership function for input 3



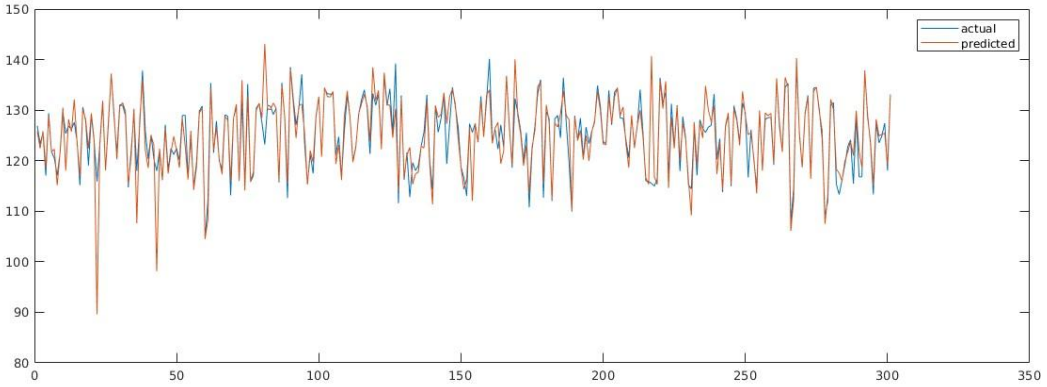
Εικόνα 27. Membership function for input 4



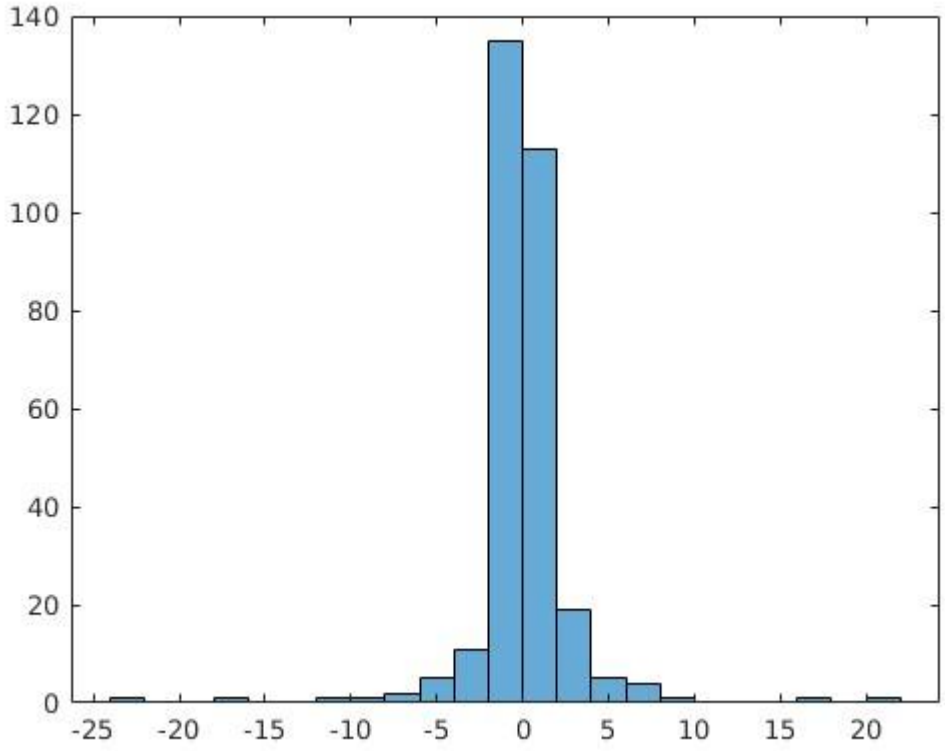
Εικόνα 28. Membership function for input 5



Εικόνα 29. Καμπύλη μάθησης για το TSK_Model_4



Εικόνα 30. Πραγματικές τιμές και τιμές προβλέψεις



Εικόνα 31. Ιστόγραμμα σφάλματος πρόβλεψης για το TSK_Model_4

1.2.5 Συγκεντρωτικός πίνακας

	TSKmodel1	TSKmodel2	TSKmodel3	TSKmodel4
R^2	0.6120	0.7428	0.7981	0.7282
RMSE	4.2979	3.3407	3.1000	3.5971
NMSE	0.3867	0.2691	0.2012	0.2709
NDIE	0.6219	0.5719	0.4485	0.5205

Πίνακας 1. Σύγκριση αποτελεσμάτων για τα μοντέλα με βάση τις μετρικές

1.3 Συμπεράσματα

Μπορούν να βγουν αρκετά χρήσιμα συμπεράσματα από όλη την ανάλυση. Καταρχήν και οι τέσσερις προβλέψεις ήταν αρκετά καλές γιατί όπως παρατηρείται το R^2 ήταν πάντα μεγάλο και το NMSE μικρό. Το μοντέλο TSK_3 το οποίο είναι πολυωνυμικό και χρησιμοποιεί 2 συναρτήσεις συμμετοχής ήταν το καλύτερο μοντέλο, με R^2 που αγγίζει το 0.8 και μόλις 0.2 NMSE. Γενικά παρατηρείται ότι τα μοντέλα που έχουν πολυωνυμική έξοδο φαίνεται πως μοντελοποιούν καλύτερα το πρόβλημα του θορύβου αεροτομής (airfoil noise) σε σχέση με τα μοντέλα οπισθοδιάδοσης. Επίσης παρατηρείται ότι τα μοντέλα που χρησιμοποιούν τρεις συναρτήσεις συμμετοχής είναι πιο επιρρεπή στο φαινόμενο της υπερεκπαίδευσης (overfitting). Αυτό ειδικά φαίνεται ξεκάθαρα στην εικόνα 29 όπου παρουσιάζεται η καμπύλη μάθησης για το TSK_Model_4. Επίσης, στο καλύτερο μοντέλο που είναι το πολυωνυμικής εξόδου με 2 συναρτήσεις συμμετοχής, το TSK_Model_3 δηλαδή, παρατηρείται από τα plots των membership functions ότι έχει ισορροπημένη προς μικρή επικάλυψη που φτάνει μέχρι το 0.5 περίπου ενώ στα άλλα μοντέλα, σε μερικές εισόδους η επικάλυψη δεν είναι και η καλύτερη. Παραδείγματος χάρην, στην εικόνα 24 παρατηρούμε το membership plot για το TSK_4 για την πρώτη είσοδο, στο οποίο μία συνάρτηση συμμετοχής δεν συμμετέχει σχεδόν καθόλου. Ενώ στα άλλα σχεδιαγράμματα η επικάλυψη μπορεί να

φτάσει σε τεράστιο ποσοστό, πράγμα που σημαίνει ότι το μοντέλο δεν μπορεί να διακρίνει καλά την απόφαση. Οπότε με δύο συναρτήσεις συμμετοχής το συγκεκριμένο πρόβλημα μπορεί να μοντελοποιηθεί και να δωσει καλύτερα αποτελέσματα καθώς υπάρχει και καλύτερη ισορροπία στα memberships functions.

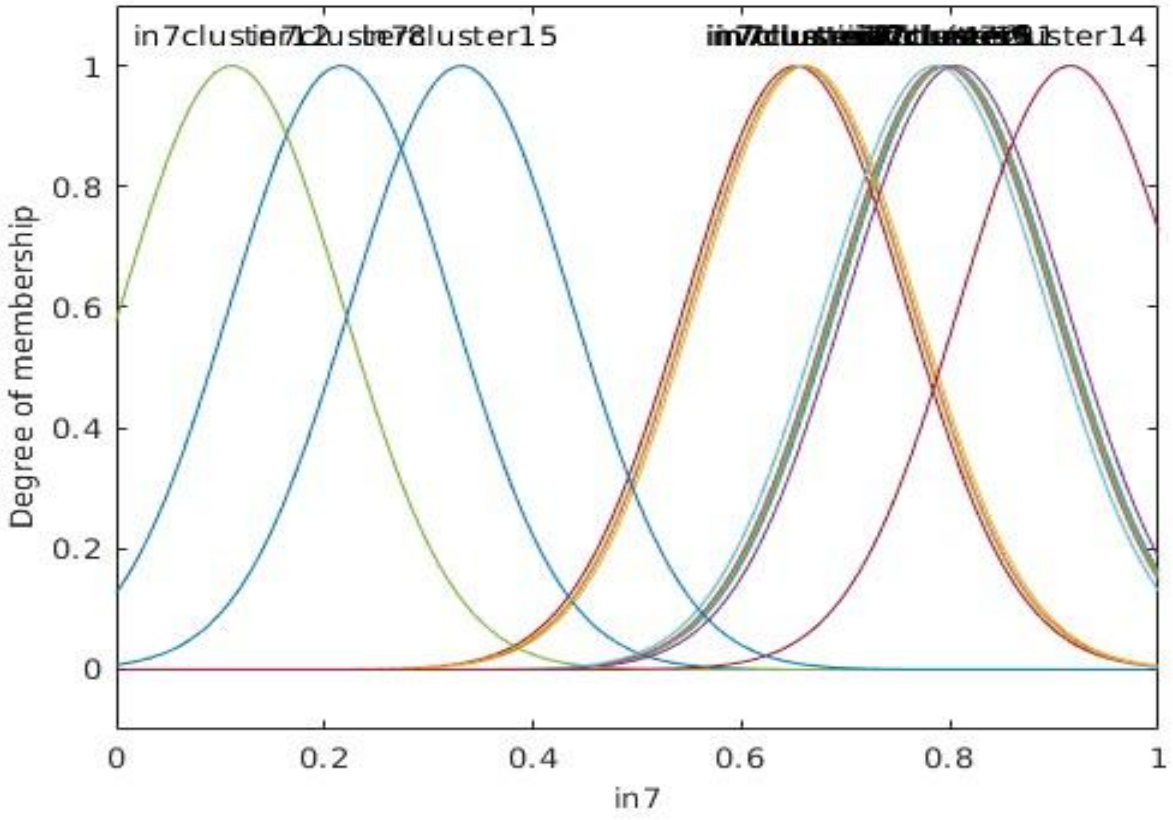
2. Εφαρμογή σε Dataset με υψηλή διαστασιμότητα

Στο δεύτερο κομμάτι της εργασίας το dataset ήταν υψηλότερης διαστασιμότητας και γι αυτόν τον λόγο απαιτήθηκε να γίνει διαφορετική προσέγγιση του προβλήματος. Αρχικά επιλέχθηκε ένας πίνακας με τους αριθμούς των χαρακτηριστικών του dataset που θα χρησιμοποιηθεί στην εκπαίδευση και στην επικύρωση, καθώς και μερικές υποψήφιες τιμές για την ακτίνα ra που καθορίζει ουσιαστικά το πλήθος των κανόνων που θα προκύψουν. Για τον αριθμό των χαρακτηριστικών επιλέχθηκαν οι παρακάτω τιμές:

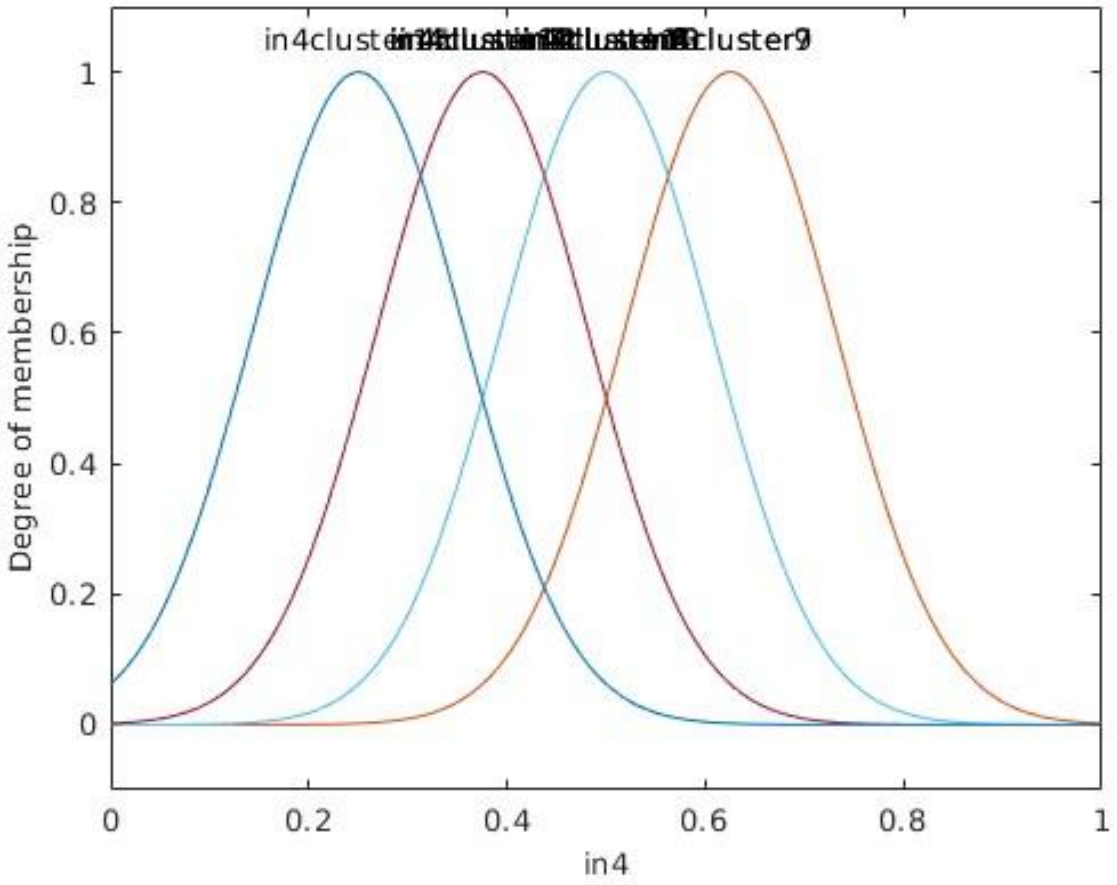
αριθμός χαρακτηριστικών = [5 10 15 20 25]

τιμές ακτίνας cluster ra = [0.1 0.2 0.3 0.4 0.5 0.6 0.8]

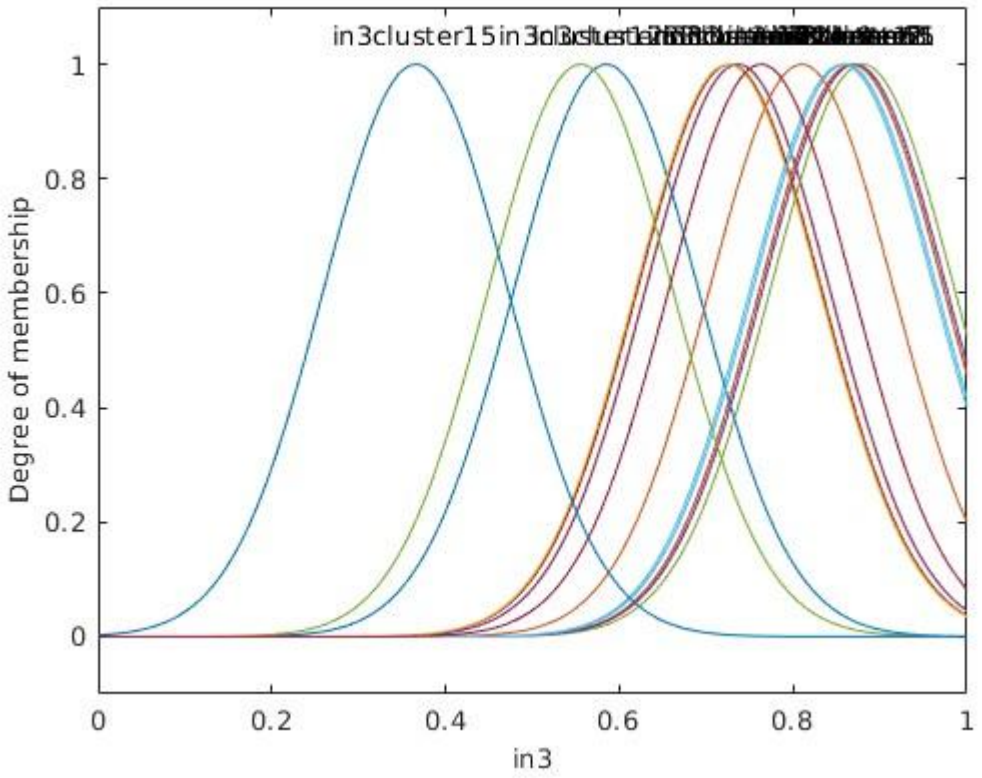
Έπειτα από την αρχικοποίηση μερικών μεταβλητών και πινάκων, τα δεδομένα επεξεργάζονται μέσω της συνάρτησης relieff. Αυτή η συνάρτηση ουσιαστικά επιστρέφει το πόσο μεγάλη η μικρή είναι η βαρύτητα του κάθε χαρακτηριστικού στο dataset. Μετά από αυτό, εφαρμόζεται ο αλγόριθμος της διασταυρωμένης επικύρωσης. Τα αποτελέσματα αποθηκεύονται σε πίνακες και το καλύτερο μοντέλο επιλέγεται για εκπαίδευση, επικύρωση και έλεγχο.Παρακάτω παρατίθενται μερικά από τα membership functions.



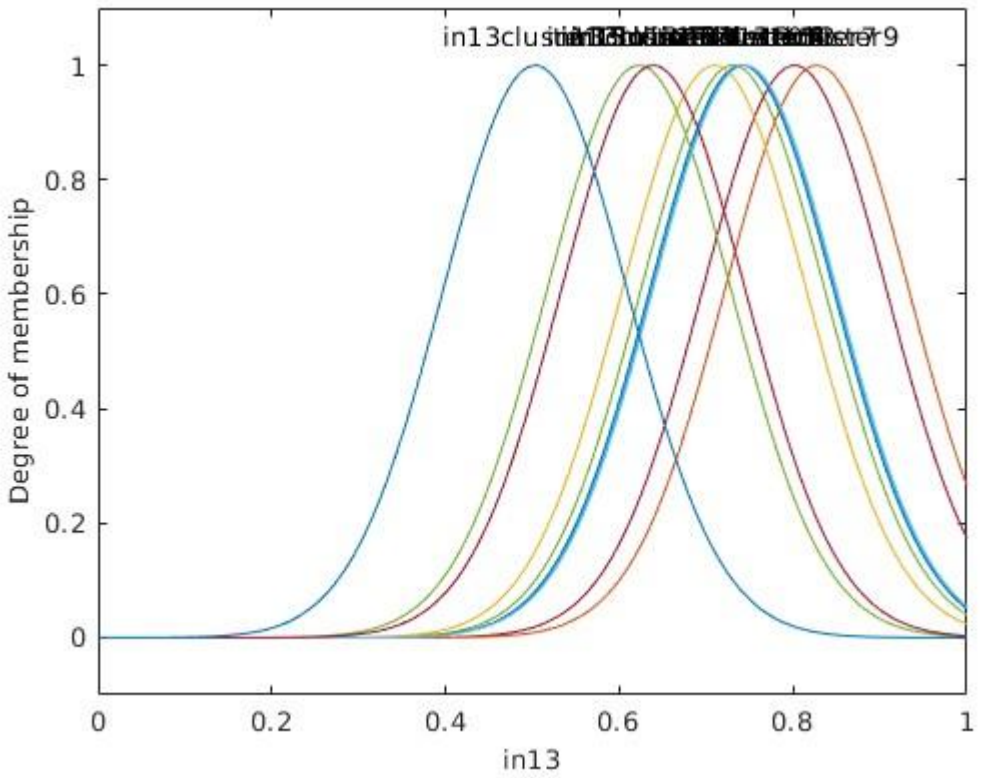
Εικόνα 32. Membership plot



Εικόνα 33. Membership plots

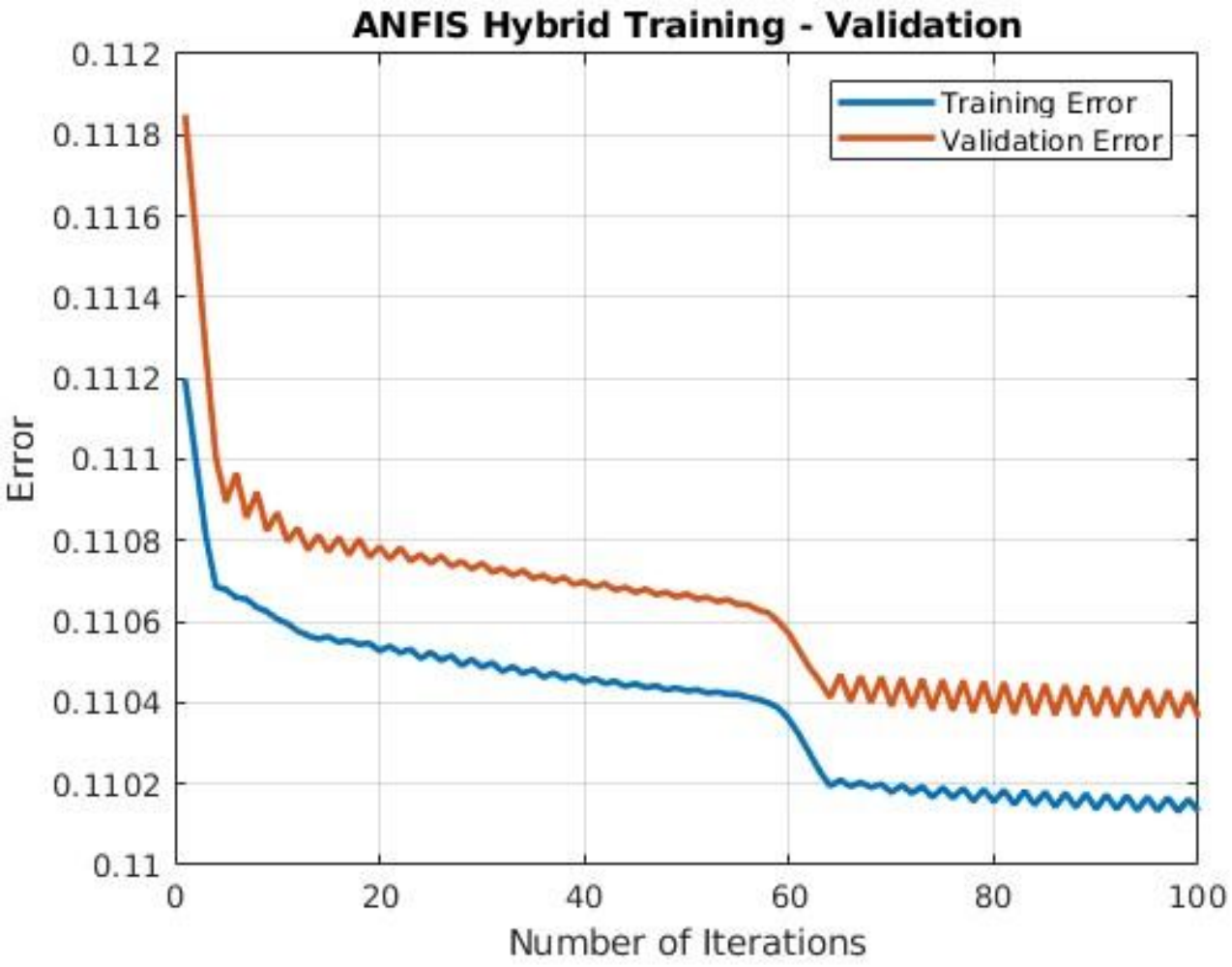


Εικόνα 34. Membership plots



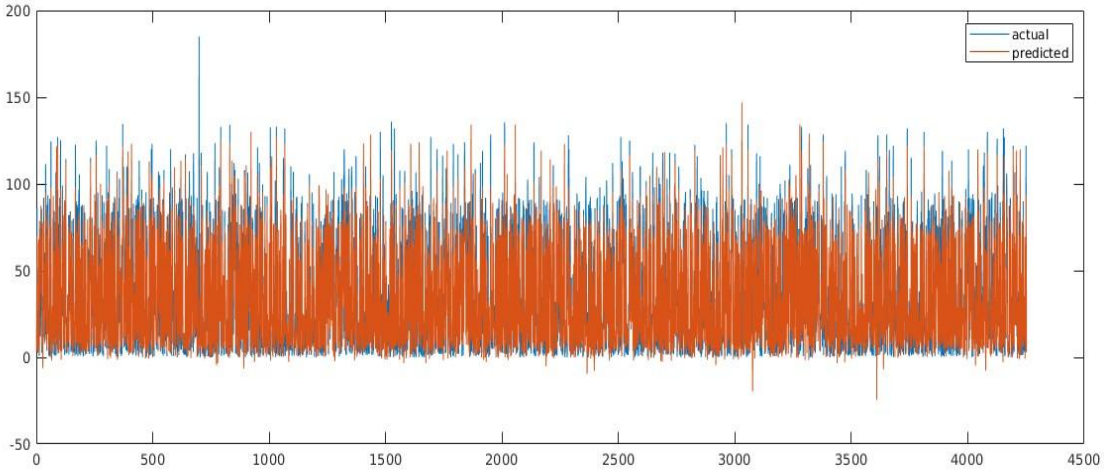
Εικόνα 35. Membership plots

Στην εικόνα 36 φαίνεται η καμπύλη μάθησης.



Εικόνα 36. Καμπύλη μάθησης

Στην εικόνα 37 απεικονίζονται οι πραγματικές τιμές με μπλε και οι προβλέψεις με κόκκινο.



Εικόνα 37. Πραγματικές τιμές με τις τιμές πρόβλεψης

Τα συγκεντρωτικά αποτελέσματα του καλύτερου μοντέλου παρουσιάζονται στον πίνακα 2.

R^2	RMSE	NMSE	NDEI
0.8056	15.1484	0.1943	0.4408

Πίνακας 2. Αποτελέσματα βέλτιστου μοντέλου

2.2 Συμπεράσματα

Φαίνεται από τα membership plots ότι η επικάλυψη είναι μεγάλη μερικές φορές, οπότε το μοντέλο χάνει την διακριτική του ικανότητα, δεν μπορεί να διακρίνει ποια είναι η σωστότερη έξοδος. Αυτό είναι φυσιολογικό γιατί η μέθοδος της

διασταυρωμένης επικύρωσης, δεν εξαντλεί πλήρως όλες τις διαθέσιμες τιμές έτσι ώστε να βρεθεί το βέλτιστο μοντέλο.

Από την άλλη μεριά, η καμπύλη μάθησης δεν δίνει άσχημα αποτελέσματα. Το μέσο τετραγωνικό σφάλμα μάθησης στα 0.11 είναι αρκετά καλό αποτέλεσμα και σίγουρα θα μπορούσε να γίνει πολύ καλύτερο αν εξαντλήσουμε περισσότερες τιμές για τα χαρακτηριστικά και την ακτίνα ra. Σε ό,τι αφορά την εικόνα 37 και την σύγκριση των πραγματικών με των τιμών πρόβλεψης παρατηρείται με μια πρώτη ματιά ότι υπάρχουν μπλε γραμμές οπότε υπάρχει ένα σφάλμα. Αν γίνουν περισσότερα plots σε συγκεκριμένες περιοχές παρατηρείται ότι το σφάλμα ναι μεν υπάρχει αλλά είναι σε φυσιολογικά πλαίσια.

Ο αριθμός κανόνων του ασαφούς συστήματος είναι 15 και εύκολα βγαίνει με την παρακάτω γραμμή κώδικα του matlab:

```
size(showrule(fis), 1)
```

Προφανώς αν επιλεχθεί grid partitioning με 2 ή 3 σύνολα ο χρόνος θα γινόταν τεράστιος καθώς με 2^15 ή 3^15 ο αριθμός των κανόνων IF-THEN θα εκτοξεύονταν.