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

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Εργασία 3η : Regression με χρήση TSK μοντέλων

Εαρινό εξάμηνο 2023/24

Δεϊρμεντζόγλου Ιωάννης
Τομέας Ηλεκτρονικής και Υπολογιστών
Α.Ε.Μ.: 10015
Email: deirmentz@ece.auth.gr

Περιεχόμενα

ΕΙΣΑΓΩΓΗ

Η παρούσα εργασία έχει ως στόχο την επίλυση προβλήματος παλινδρόμησης μέσω της εφαρμογής και αξιολόγησης ασαφών νευρωνικών μοντέλων TSK. Η μελέτη επικεντρώνεται στη διερεύνηση της ικανότητας των μοντέλων αυτών να μοντελοποιούν μη γραμμικές συναρτήσεις πολλών μεταβλητών, χρησιμοποιώντας δύο διαφορετικά σύνολα δεδομένων από το UCI repository. Η εργασία αποτελείται από 2 μέρη:

Στο 1^{ο} μέρος, εφαρμόζονται TSK μοντέλα σε ένα απλούστερο σύνολο δεδομένων (Airfoil Self-Noise dataset), με στόχο την εξοικείωση με τη διαδικασία εκπαίδευσης και αξιολόγησης των μοντέλων. Η εκπαίδευση των μοντέλων πραγματοποιείται με διάφορες παραλλαγές των παραμέτρων, ενώ αξιολογούνται με τη χρήση δεικτών απόδοσης όπως το MSE, το RMSE και ο συντελεστής προσδιορισμού R^2.

Στο 2^{ο} μέρος, τοποθετείται ένα πιο πολύπλοκο σύνολο δεδομένων (Superconductivity dataset) με υψηλή διαστασιμότητα, όπου γίνεται χρήση τεχνικών όπως η επιλογή χαρακτηριστικών (feature selection) και η διασταυρωμένη επικύρωση (cross validation) για την περαιτέρω βελτιστοποίηση των μοντέλων. Η μεθοδολογία περιλαμβάνει τη μείωση της πολυπλοκότητας μέσω της χρήσης της μεθόδου Subtractive Clustering (SC). Σε πρώτο βαθμό, γίνεται εκπαίδευση σε μοντέλα με διαφορετικές παραμέτρους (ακτίνα clusters και αριθμός features). Με βάση το μέσο σφάλμα, εξάγεται το βέλτιστο μοντέλο, όπου και μελετάται με ορισμένες παραμέτρους αξιολόγησης.

Η εργασία ολοκληρώνεται με την ανάλυση των αποτελεσμάτων και τη σύγκριση των αποδόσεων των μοντέλων, παρέχοντας σημαντικά συμπεράσματα για τη βελτίωση της ακρίβειας και της αποδοτικότητας των TSK μοντέλων στην παλινδρόμηση.

1. Εφαρμογή σε απλό dataset

Εισαγωγή Δεδομένων

Αρχικά, εισάγονται τα δεδομένα από το αρχείο “airfoil_self_noise.dat”. Το αρχείο αυτό περιέχει τα δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση και την αξιολόγηση των μοντέλων. Τα δεδομένα αυτά αφορούν θόρυβο αεροτομής και το dataset περιέχει συνολικά 1503 στοιχεία, με συνολικά 6 χαρακτηριστικά (5 είσοδοι και μία έξοδος). Εφ’ όσον ο αριθμός των features του συστήματος είναι μικρός σε αριθμό, για τον συνδυασμό των εισόδων χρησιμοποιείται η τεχνική του grid partitioning. Οπότε το σύνολο των ασαφών κανόνων που θα δημιουργηθούν για το σύστημα προκύπτει από τον συνδυασμό κάθε διαφορετικού χαρακτηριστικού (2^6 ο συνολικός αριθμός των κανόνων).

Διαχωρισμός Δεδομένων

Η γραμμή διαχωρίζει το σύνολο δεδομένων σε τρία μέρη: δεδομένα εκπαίδευσης (60%), δεδομένα επικύρωσης (20%) και δεδομένα δοκιμών (20%). Το σύνολο εκπαίδευσης (training set) χρησιμοποιείται για την εκπαίδευση των μοντέλων, ενώ το σύνολο επικύρωσης (validation set) για τη ρύθμιση των υπερπαραμέτρων και για αποφυγή του δεδομένου της υπερεκπαίδευσης και το σύνολο δοκιμών (test set) για την τελική αξιολόγηση και τον υπολογισμό των μετρικών έπειτα από το training. Ο διαχωρισμός των δεδομένων είναι 60%, 20% και 20% αντίστοιχα. Για το shuffling και τον διαχωρισμό των δεδομένων χρησιμοποιήθηκε η συνάρτηση *split_scale,* που βρίσκεται στο αρχείο split_scale.m από τα μοντέλα TSK που είναι αναρτημένα στην ιστοσελίδα του μαθήματος.

Η συναρτηση **split_scale** έχει ως εισόδους τις μεταβλητές : `data`(τα δεδομένα που θα διαχωριστούν και θα προεπεξεργαστούν) και `preproc`(παράμετρος που καθορίζει το είδος προεπεξεργασίας) . Έχει ως εξόδους τα σύνολα `trnData` (δεδομένα εκπαίδευσης) , `valData` (δεδομένα επικύρωσης) και `tstData` (δεδομένα δοκιμών). Αρχικά, τα δεδομένα ανακατεύονται τυχαία μέσω της συνάρτησης `randperm`, η οποία επιστρέφει έναν τυχαίο πίνακα με δείκτες. Στη συνέχεια, τα δεδομένα χωρίζονται και ανάλογα με την τιμή της παραμέτρου `preproc`, εφαρμόζονται δύο είδη προεπεξεργασίας:

- Περίπτωση 1: Κανονικοποίηση (Normalization) στο unit hypercube.
- Περίπτωση 2: Τυποποίηση (Standardization) σε μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση.

TSK Μοντέλα Εκπαίδευσης

Από την εκφώνηση ζητείται να εκπαιδευθούν 4 TSK μοντέλα για το παραπάνω dataset. Αυτά τα μοντέλα έχουν διαφορετικές παραμέτρους, όπως ο αριθμός και ο τύπος των συναρτήσεων συμμετοχής (Membership Functions) και η μορφή της εξόδου (constant ή linear). Το κάθε μοντέλο έχει 5 εισόδους και πλήθος MFs για κάθε είσοδο και τύπο εξόδου του sugeno μοντέλου, που καθορίζεται ως εξής:

	Πλήθος συναρτήσεων συμμετοχής	Μορφή εξόδου
TSK_model_1	2	Singleton
TSK_model_2	3	Singleton
TSK_model_3	2	Polynomial
TSK_model_4	3	Polynomial

Πίνακας 1: Ταξινόμηση μοντέλων προς εκπαίδευση.

Τα δύο μοντέλα χρησιμοποιούν 2 ασαφή σύνολα για κάθε μεταβλητή εισόδου, ενώ ο χώρος των μεταβλητών εισόδου στα άλλα δύο μοντέλα χωρίζεται στα τρία. Η άλλη διαφορά μεταξύ των μοντέλων είναι η μορφή της εισόδου. Σε δύο μοντέλα η έξοδος είναι μία σταθερή τιμή, ενώ στα άλλα δύο TSK μοντέλα η έξοδος έχει πολυωνυμική μορφή.

Οι παράμετροι αξιολόγησης των μοντέλων στην εργασία παλινδρόμησης με χρήση μοντέλων TSK είναι οι εξής:

1. **MSE (Mean Squared Error - Μέσο Τετραγωνικό Σφάλμα)**:
    - Το MSE μετράει το μέσο όρο των τετραγώνων των διαφορών μεταξύ των πραγματικών τιμών και των τιμών που προβλέπει το μοντέλο. Είναι δείκτης που εκφράζει το πόσο κοντά είναι οι προβλέψεις του μοντέλου στις πραγματικές τιμές.
    - **Συμβολισμός**: Όσο μικρότερο είναι το MSE, τόσο πιο ακριβείς είναι οι προβλέψεις του μοντέλου. Ο τύπος του MSE είναι:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - \hat{f}(x_i))^2$$

2. **RMSE (Root Mean Squared Error - Ρίζα Μέσου Τετραγωνικού Σφάλματος)**:
    - Το RMSE είναι απλώς η τετραγωνική ρίζα του MSE και χρησιμοποιείται για να φέρει το σφάλμα πίσω στην αρχική μονάδα μέτρησης των δεδομένων.
    - **Συμβολισμός**: Το RMSE είναι εύκολα κατανοητό διότι εκφράζεται στις ίδιες μονάδες με την αρχική τιμή που προβλέπεται. Υπολογίζεται ως:

$$RMSE = \sqrt{MSE}$$

Η αντίστοιχη συναρτηση βρίσκεται στο αρχείο RMSE.m .

3. **R^2 (R-Squared - Συντελεστής Προσδιορισμού)**:
   - Ο συντελεστής προσδιορισμού R^2 μετράει το ποσοστό της διακύμανσης των δεδομένων που εξηγείται από το μοντέλο. Με άλλα λόγια, δείχνει πόσο καλά το μοντέλο ταιριάζει στα δεδομένα.
   - **Συμβολισμός**: Το R^2 κυμαίνεται από 0 έως 1, όπου 1 σημαίνει ότι το μοντέλο εξηγεί πλήρως τη διακύμανση των δεδομένων. Ο τύπος του είναι:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Οπου SS_{res} είναι το άθροισμα των τετραγώνων των σφαλμάτων και SS_{tot} η συνολική διακύμανση των δεδομένων.

Η αντίστοιχη συναρτηση βρίσκεται στο αρχείο R2.m .

4. **NMSE (Normalized Mean Squared Error - Κανονικοποιημένο Μέσο Τετραγωνικό Σφάλμα)**:
   - Το NMSE είναι μια κανονικοποιημένη εκδοχή του MSE, το οποίο προκύπτει διαιρώντας το MSE με τη διακύμανση των πραγματικών δεδομένων.
   - **Συμβολισμός**: Το NMSE βοηθά στην κατανόηση του σφάλματος σε σχέση με τη διακύμανση των δεδομένων. Ο τύπος είναι:

$$NMSE = \frac{\sigma_e^2}{\sigma_x^2}$$

όπου σ_e^2 είναι το μέσο τετραγωνικό σφάλμα και σ_x^2 η διακύμανση των πραγματικών δεδομένων.

Η αντίστοιχη συναρτηση βρίσκεται στο αρχείο NMSE.m .

5. **NDEI (Normalized Mean Deviation Index - Κανονικοποιημένος Δείκτης Μέσης Απόκλισης)**:
   - Το NDEI είναι η τετραγωνική ρίζα του NMSE και εκφράζει το κανονικοποιημένο σφάλμα σε σχέση με την τυπική απόκλιση των πραγματικών δεδομένων.
   - **Συμβολισμός**: Ένα μικρό NDEI υποδεικνύει ότι το μοντέλο προβλέπει με ακρίβεια, ενώ μια μεγαλύτερη τιμή δείχνει μεγαλύτερο σφάλμα. Υπολογίζεται ως:

$$NDEI = \frac{\sigma_e}{\sigma_x} = \sqrt{NMSE}$$

Η αντίστοιχη συναρτηση βρίσκεται στο αρχείο NDEI.m .

Στο αρχειο **TSK_Regression_1stPart.m** βρισκεται το συνολο του κωδικα για ορίζονται τα τέσσερα διαφορετικά TSK μοντέλα, και δημιουργείται το αντίστοιχο FIS. Ακολουθει η εκπαιδευση των μοντελων με την χρήση της συνάρτησης anfis, όπου ορίζονται και τα αντίστοιχα χαρακτηριστικά του training.

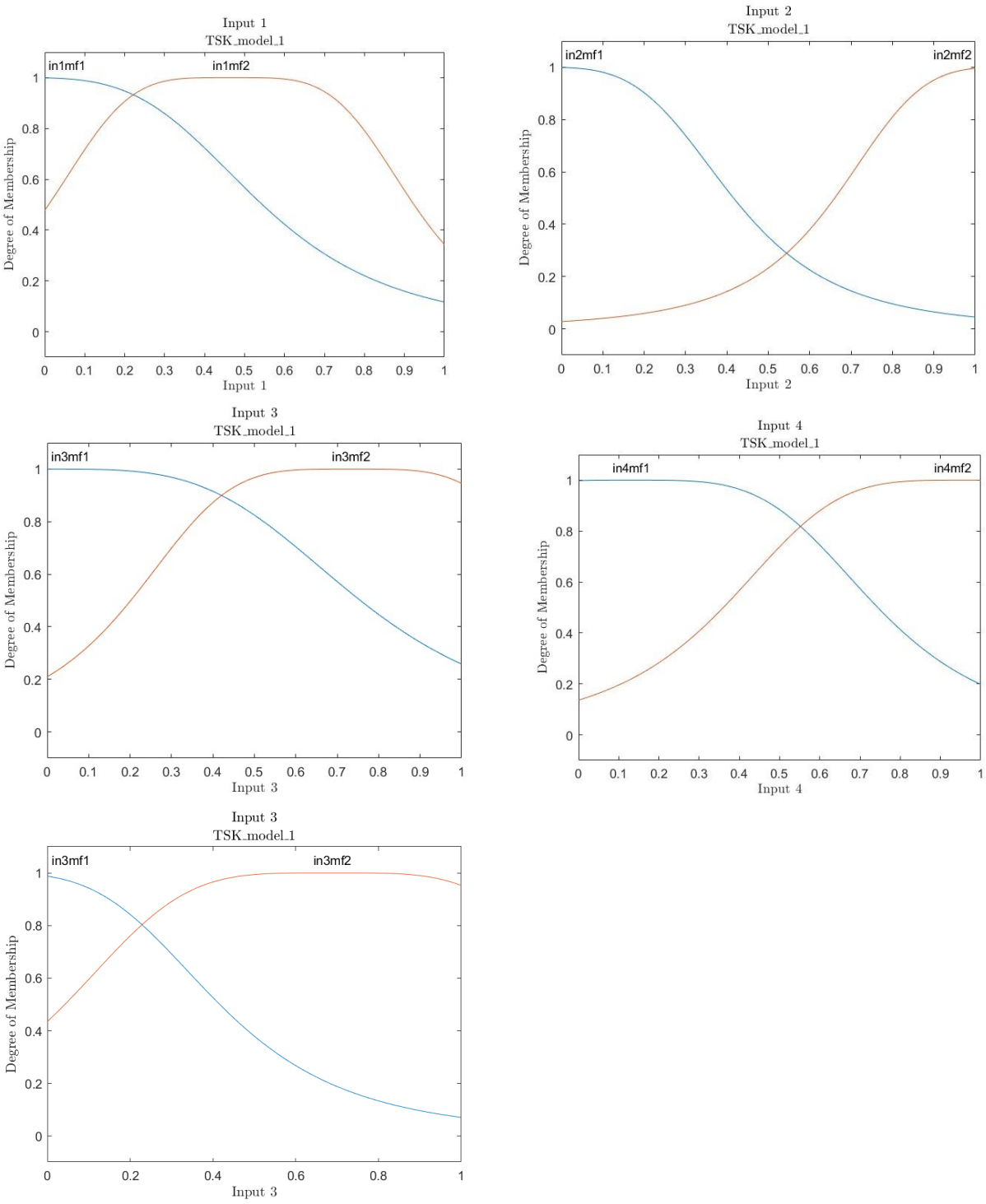
Καμπύλες Εκμάθησης

Η καμπύλη εκμάθησης κάθε μοντέλου, δηλαδή η εξέλιξη του σφάλματος εκπαίδευσης και επικύρωσης κατά τη διάρκεια των επαναλήψεων, οπτικοποιείται με τη συνάρτηση `plot`.

Σφάλμα Πρόβλεψης

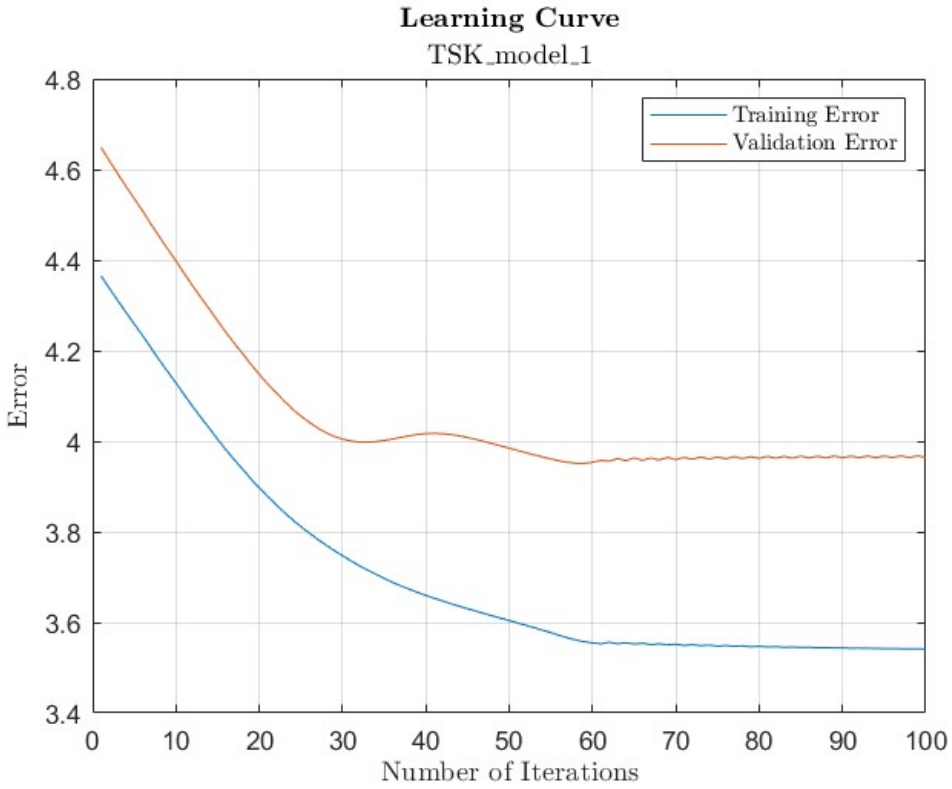
Αφού εκπαιδευτούν τα μοντέλα, υπολογίζεται το σφάλμα πρόβλεψης για τα δεδομένα δοκιμών, δηλαδή η διαφορά ανάμεσα στις προβλεπόμενες τιμές και τις πραγματικές τιμές.

Model 1: Input→ 2 Membership Functions, Output→ Singleton

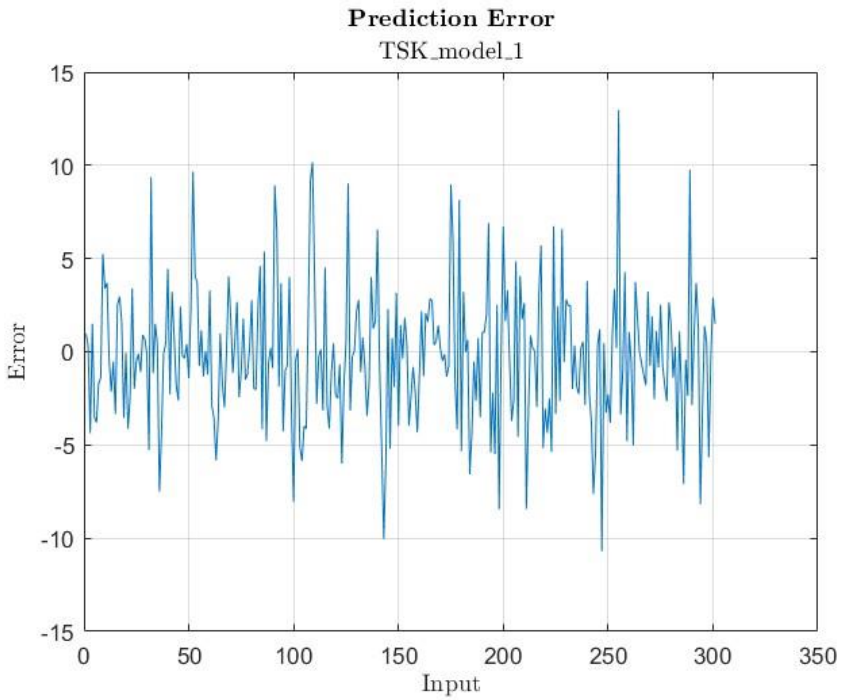


Εικόνα 1.1 : Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 1^{ου} μοντέλου

Ακολούθως, δίνονται οι καμπύλες μάθησης (training) του μοντέλου καθώς και τα σφάλματα κατά την εφαρμογή του εκπαιδευμένου μοντέλου (testing) στο test set:



Εικόνα 1.2 : Καμπύλες μάθησης 1ου μοντέλου

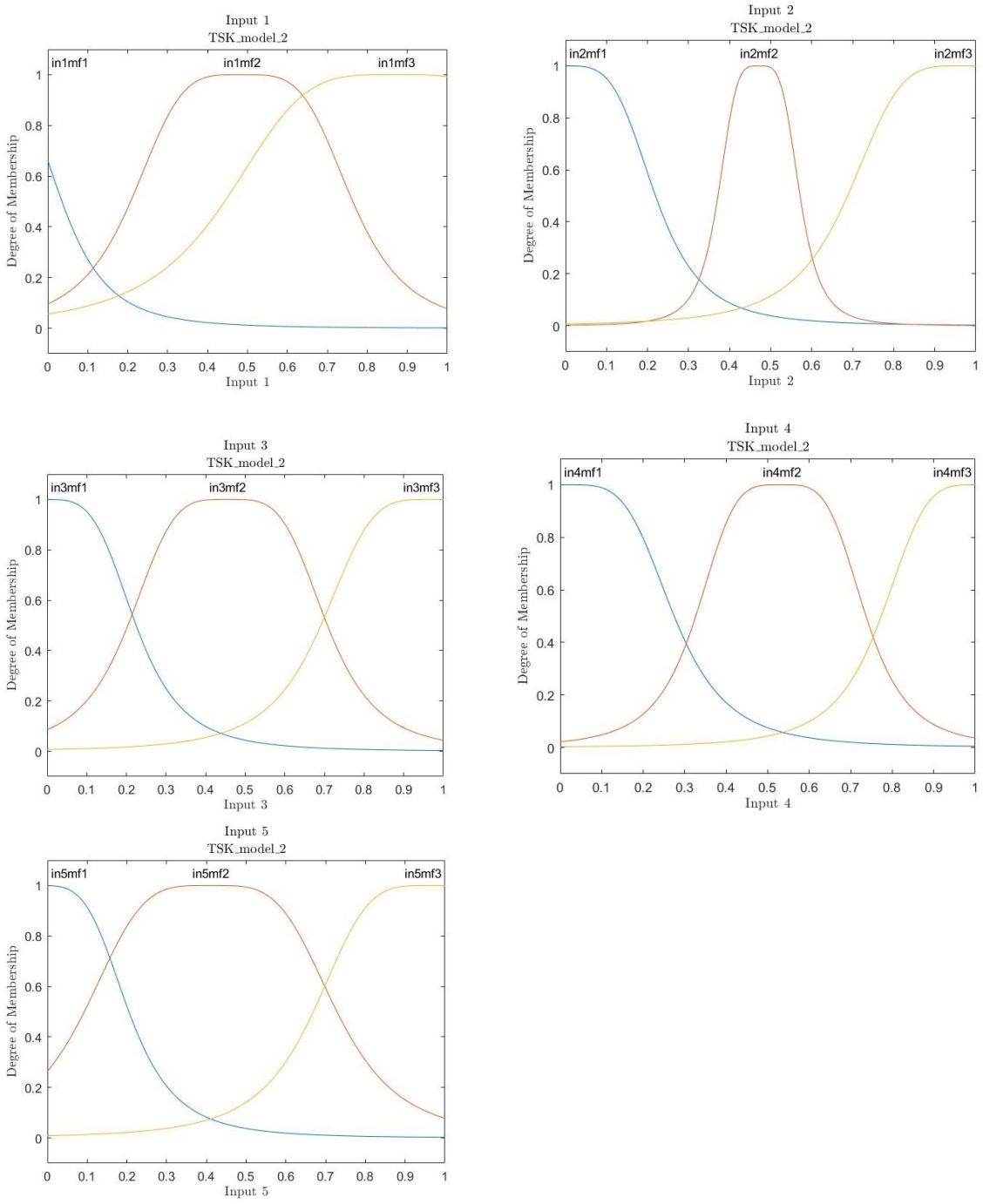


Εικόνα 1.3 : Σφάλματα πρόβλεψης κατά την εφαρμογή του 1^{ου} μοντέλου στο test set

Τέλος παρατίθενται οι ζητούμενες μετρικές αξιολόγησης του 1ου μοντέλου στον παρακάτω πίνακα :

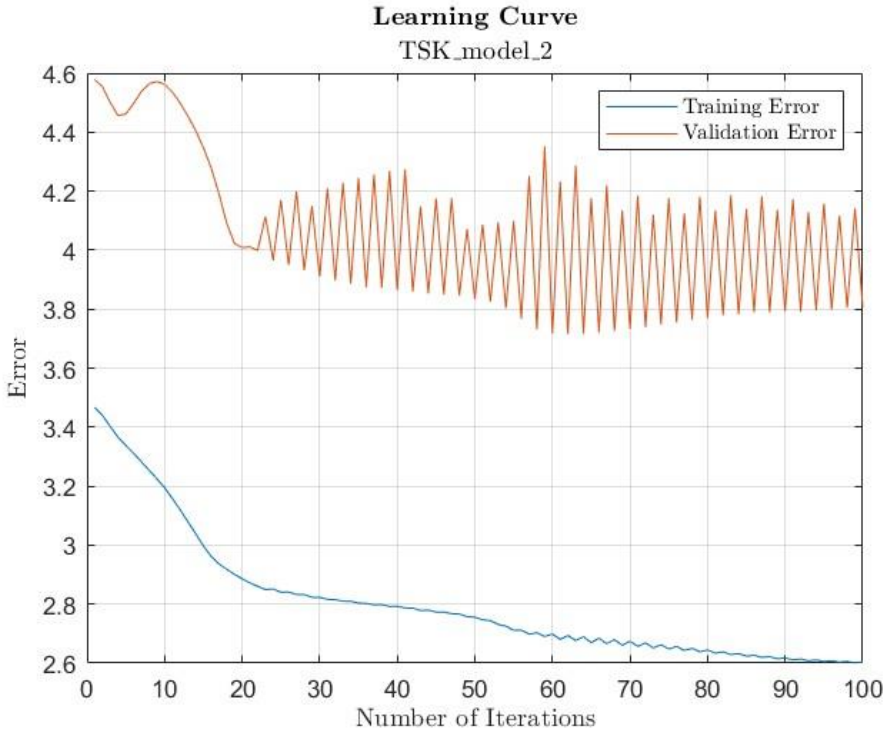
Metric	MSE	RMSE	NMSE	NDEI	R^2
Value	13.813	3.6308	0.29938	0.54716	0.70062

Model 2: Input→ 3 Membership Functions, Output→ Singleton

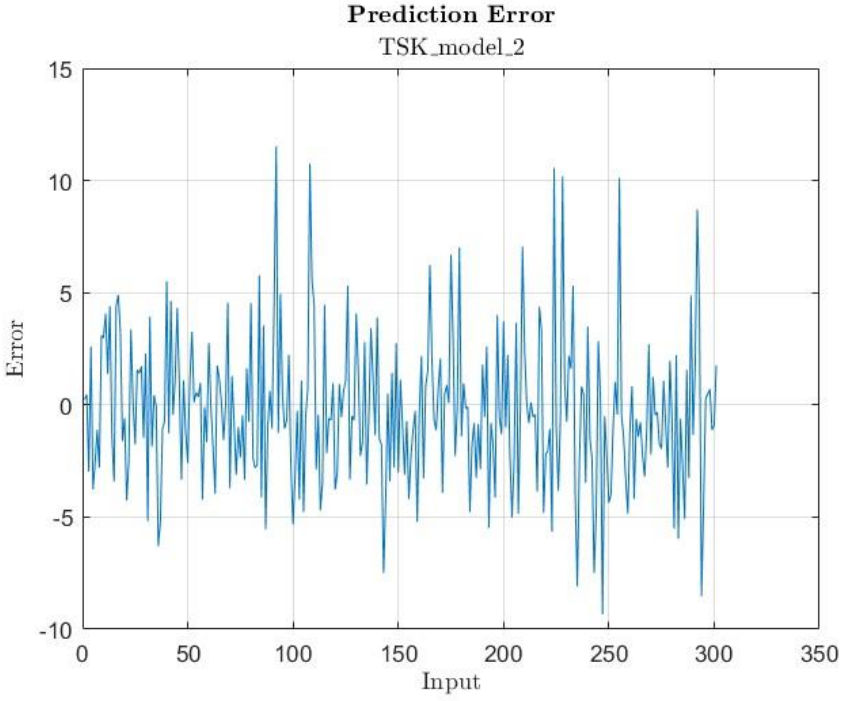


Εικόνα 1.4 : Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 2ου μοντέλου

Ακολούθως, δίνονται οι καμπύλες μάθησης (training) του μοντέλου καθώς και τα σφάλματα κατά την εφαρμογή του εκπαιδευμένου μοντέλου (testing) στο test set:



Εικόνα 1.5 : Καμπύλες μάθησης 2ου μοντέλου

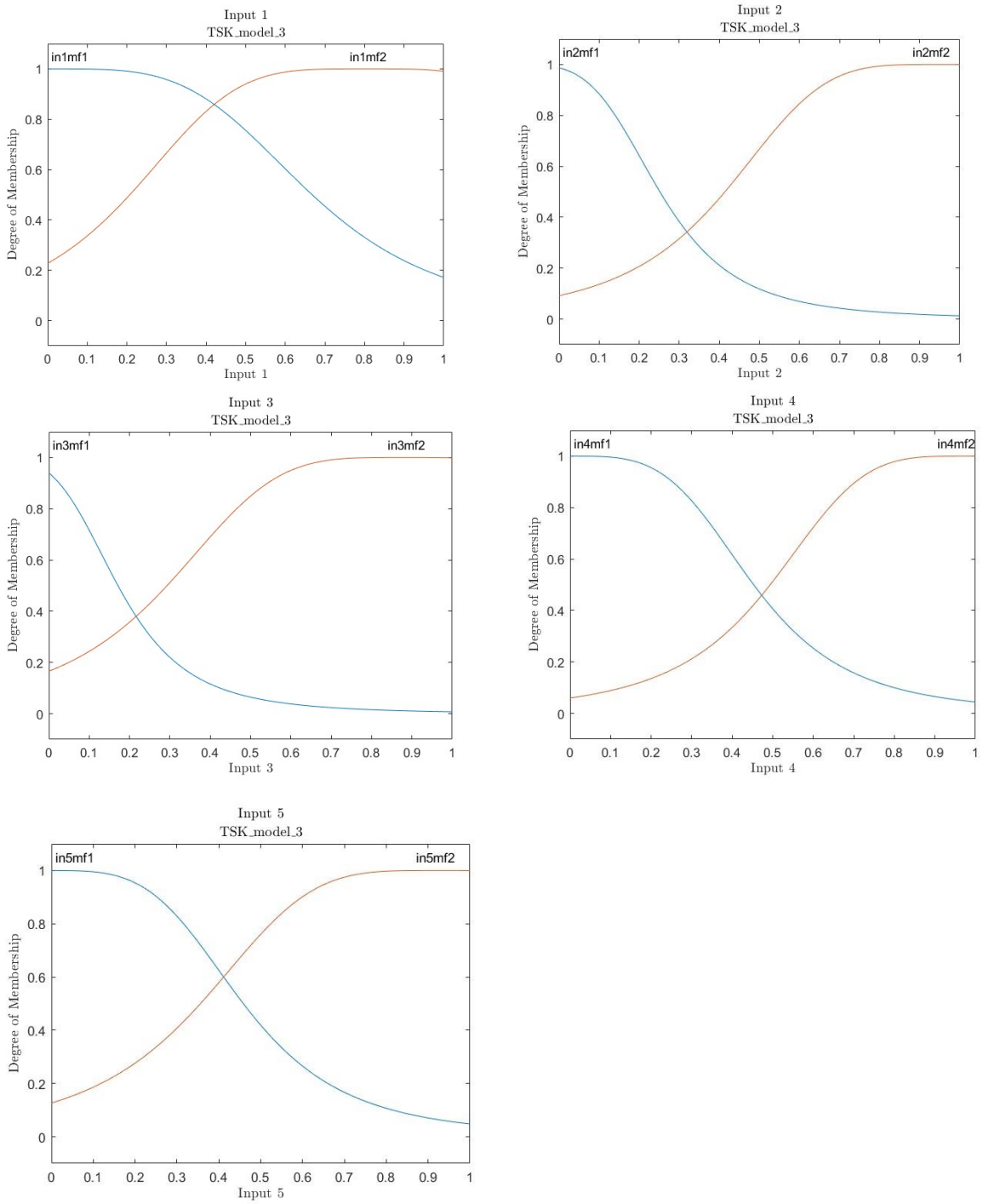


Εικόνα 1.6 : Σφάλματα πρόβλεψης κατά την εφαρμογή του 2ου μοντέλου στο test set

Τέλος παρατίθενται οι ζητούμενες μετρικές αξιολόγησης του 2ου μοντέλου στον παρακάτω πίνακα :

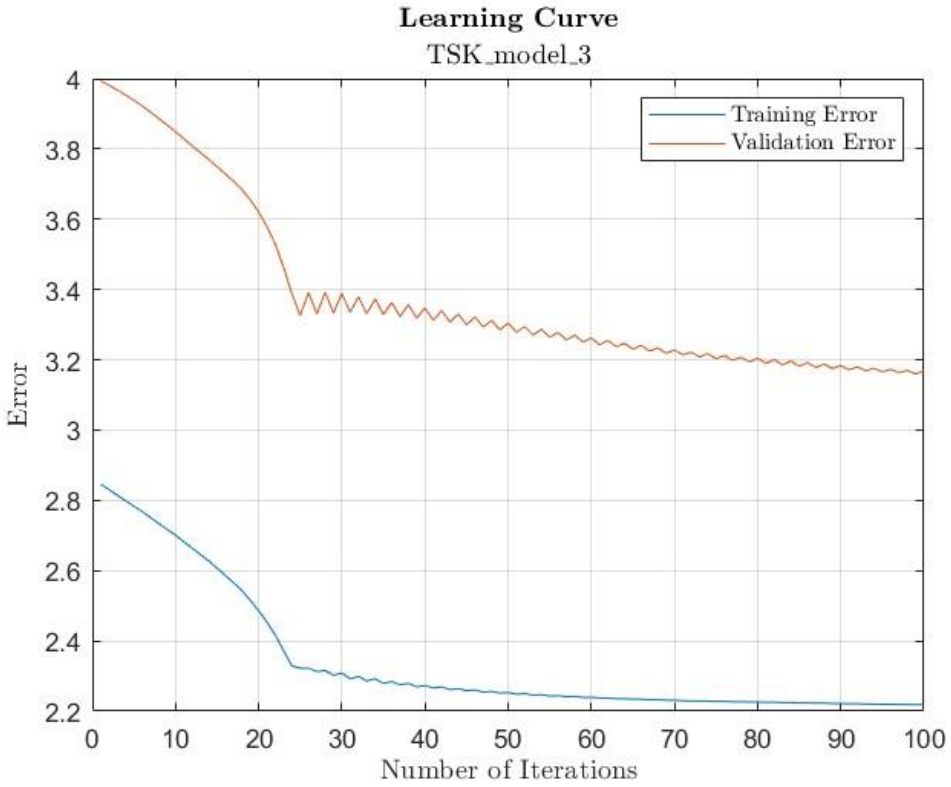
Metric	MSE	RMSE	NMSE	NDEI	R^2
Value	10.895	3.3008	0.24743	0.49743	0.75257

Model 3: Input→ 2 Membership Functions, Output→ Polynomial

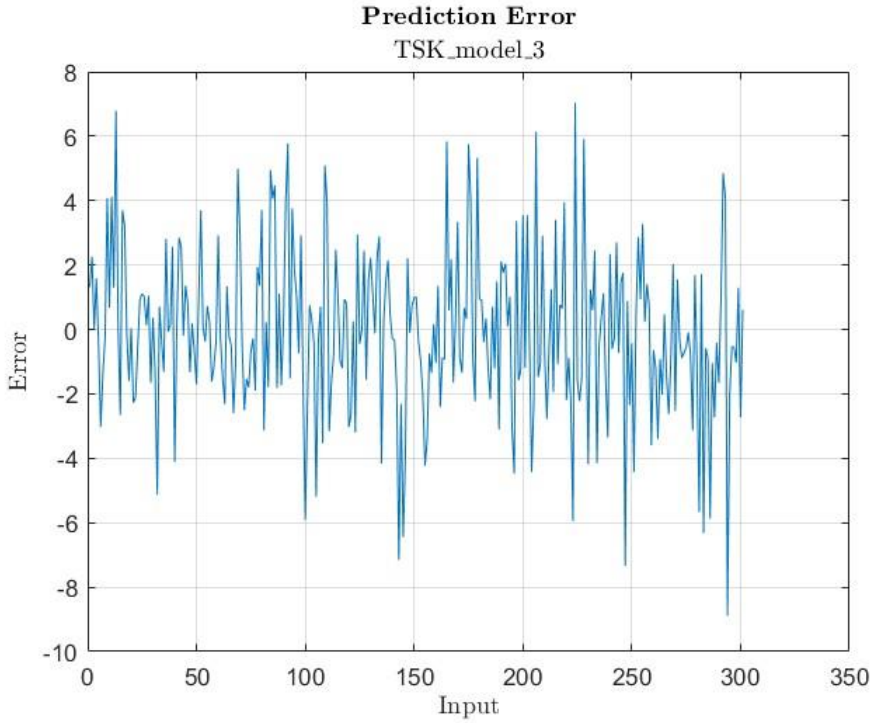


Εικόνα 1.7 : Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 3ου μοντέλου

Ακολούθως, δίνονται οι καμπύλες μάθησης (training) του μοντέλου καθώς και τα σφάλματα κατά την εφαρμογή του εκπαιδευμένου μοντέλου (testing) στο test set:



Εικόνα 1.8 : Καμπύλες μάθησης 3ου μοντέλου

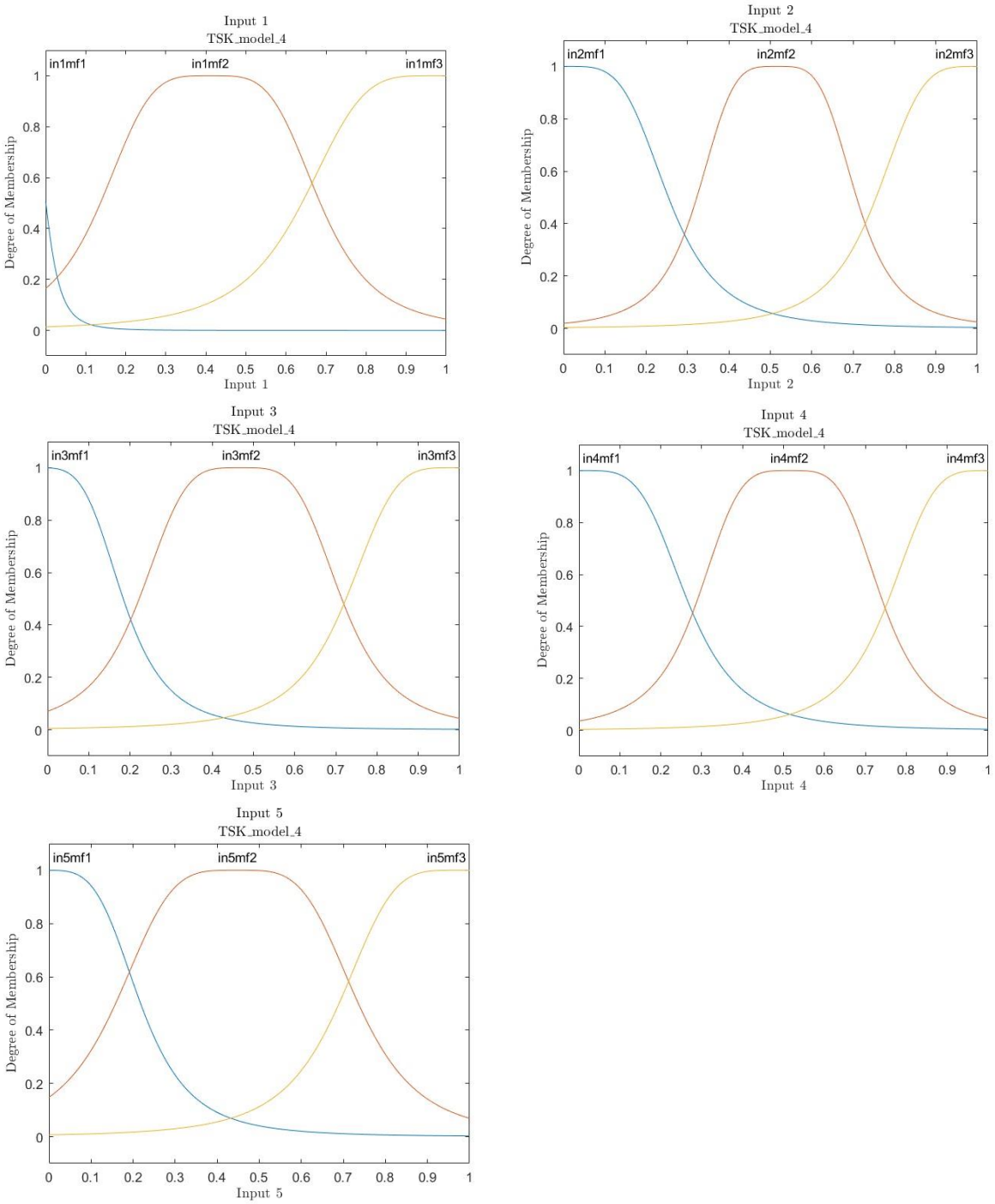


Εικόνα 1.9 : Σφάλματα πρόβλεψης κατά την εφαρμογή του 3^{ου} μοντέλου στο test set

Τέλος παρατίθενται οι ζητούμενες μετρικές αξιολόγησης του 3ου μοντέλου στον παρακάτω πίνακα :

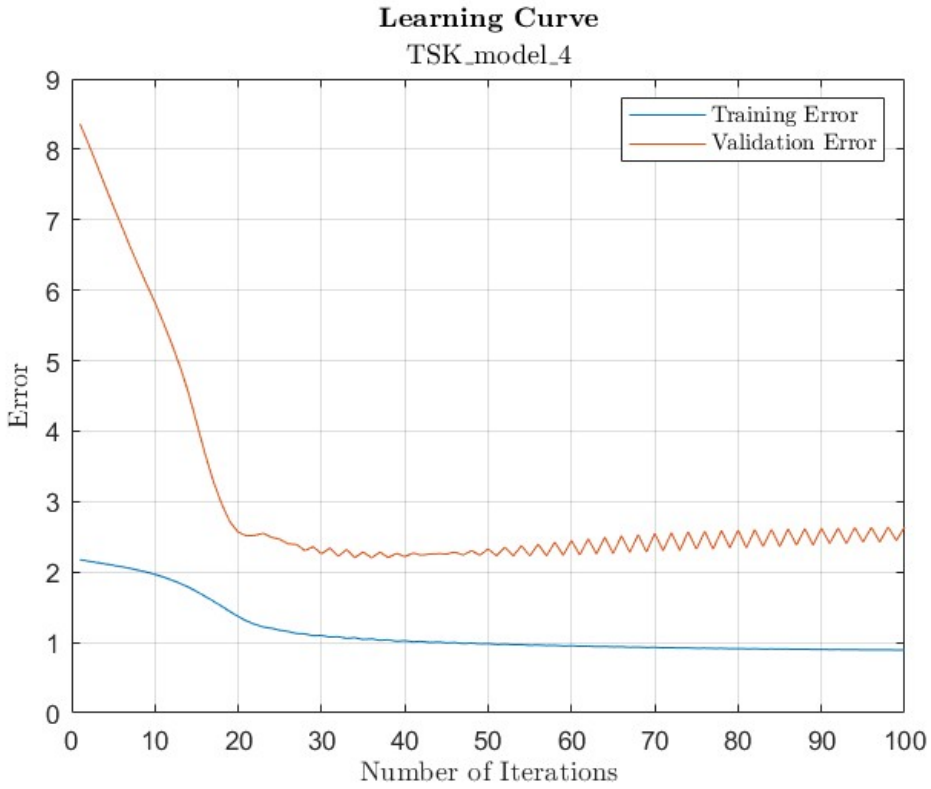
Metric	MSE	RMSE	NMSE	NDEI	R^2
Value	6.4347	2.5367	0.14613	0.38227	0.85387

Model 4: Input→ 3 Membership Functions, Output→ Polynomial

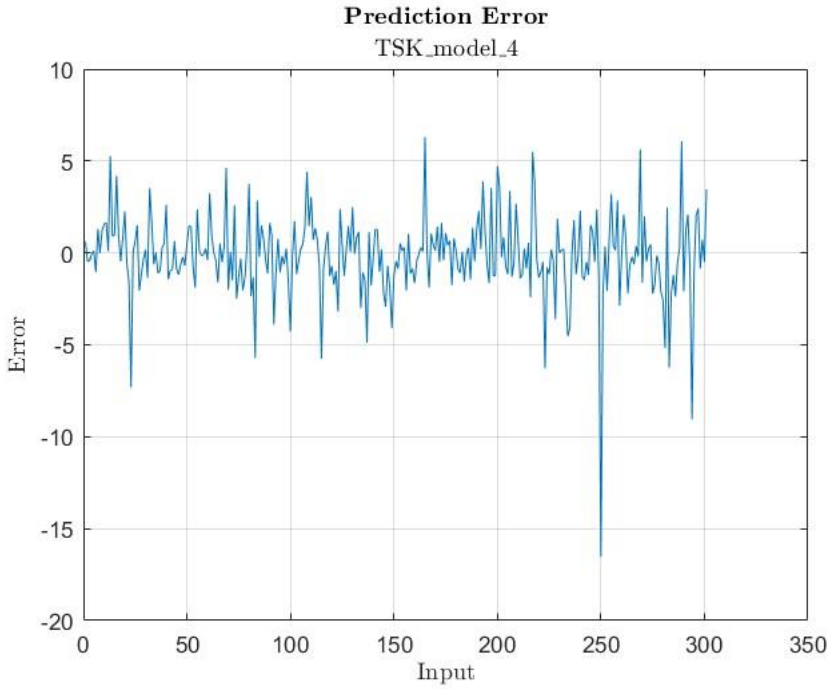


Εικόνα 1.10 : Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 4ου μοντέλου

Ακολούθως, δίνονται οι καμπύλες μάθησης (training) του μοντέλου καθώς και τα σφάλματα κατά την εφαρμογή του εκπαιδευμένου μοντέλου (testing) στο test set:



Εικόνα 1.11 : Καμπύλες μάθησης 4ου μοντέλου



Εικόνα 1.12 : Σφάλματα πρόβλεψης κατά την εφαρμογή του 4^{ου} μοντέλου στο test set

Τέλος παρατίθενται οι ζητούμενες μετρικές αξιολόγησης του 4ου μοντέλου στον παρακάτω πίνακα :

Metric	**MSE**	**RMSE**	**NMSE**	**NDEI**	**R^2**
Value	5.0512	2.2475	0.11471	0.33869	0.88529

Σύγκριση των Αποτελεσμάτων Εκπαίδευσης

I. Διαμέριση του Χώρου Εισόδου

Στην εκπαίδευση των τεσσάρων μοντέλων τύπου TSK, παρατηρούμε σημαντικές διαφορές στη διαμέριση του χώρου των εισόδων, δηλαδή στη μορφή και την κατανομή των συναρτήσεων συμμετοχής που χρησιμοποιούνται για τον ορισμό των ασαφών συνόλων σε κάθε μεταβλητή εισόδου.

1ο Μοντέλο: Στο πρώτο μοντέλο, το οποίο χρησιμοποιεί δύο ασαφή σύνολα ανά μεταβλητή εισόδου, παρατηρείται μετατόπιση των συναρτήσεων συμμετοχής. Τα δεύτερα ασαφή σύνολα (που αντιστοιχούν στο "υψηλό" επίπεδο κάθε εισόδου) καλύπτουν μεγαλύτερο εύρος στον χώρο των τιμών εισόδου, συγκριτικά με τα πρώτα ασαφή σύνολα. Αυτό σημαίνει ότι το μοντέλο τείνει να "ευνοεί" την περιγραφή των δεδομένων με τις υψηλότερες τιμές εισόδου. Ωστόσο, υπάρχει μία εξαίρεση: στη δεύτερη μεταβλητή εισόδου, η κατανομή των συναρτήσεων συμμετοχής παραμένει περισσότερο συμμετρική, γεγονός που υποδεικνύει μια πιο ισορροπημένη κατανομή των τιμών της μεταβλητής.

2ο Μοντέλο: Στο δεύτερο μοντέλο, το οποίο χρησιμοποιεί τρία ασαφή σύνολα ανά μεταβλητή, η κατανομή των συναρτήσεων συμμετοχής αλλάζει ακόμη περισσότερο. Το δεύτερο ασαφές σύνολο (το "μεσαίο") καλύπτει τον περισσότερο χώρο στον χώρο των εισόδων, ενώ τα πρώτα και τρίτα σύνολα ("χαμηλό" και "υψηλό") συμπιέζονται και περιορίζονται στα άκρα. Αυτό ισχύει για όλες τις μεταβλητές εισόδου, με εξαίρεση τη δεύτερη, όπου η κατανομή των ασαφών συνόλων είναι πιο ισομερώς κατανεμημένη και συμμετρική.

3ο Μοντέλο: Στο τρίτο μοντέλο, όπου και πάλι υπάρχουν δύο ασαφή σύνολα ανά μεταβλητή, αλλά η έξοδος του μοντέλου δεν είναι πλέον σταθερή αλλά πολυωνυμική (γραμμική), βλέπουμε μία διαφορετική συμπεριφορά στη διαμόρφωση των συναρτήσεων συμμετοχής. Εδώ, τα δεύτερα ασαφή σύνολα για κάθε μεταβλητή (τα "υψηλά") τείνουν να επεκτείνονται σε μεγαλύτερο εύρος, ενώ τα πρώτα (τα "χαμηλά") περιορίζονται ή συμπιέζονται. Μια εξαίρεση είναι η πρώτη μεταβλητή εισόδου, όπου το πρώτο ασαφές σύνολο δεν συμπιέζεται αλλά, αντίθετα, επεκτείνεται.

4ο Μοντέλο: Το τέταρτο μοντέλο χρησιμοποιεί τρία ασαφή σύνολα ανά μεταβλητή εισόδου, όπως το δεύτερο, αλλά και εδώ η έξοδος είναι πολυωνυμική. Παρατηρούνται σημαντικές αλλαγές στην κατανομή των συναρτήσεων συμμετοχής σε σχέση με τα αρχικά σύνολα, ειδικά στην πρώτη και την τέταρτη μεταβλητή εισόδου. Στην πρώτη μεταβλητή, το αριστερό ασαφές σύνολο ("χαμηλό") σχεδόν εξαφανίζεται, ενώ στην τέταρτη μεταβλητή, το μεσαίο ασαφές σύνολο επεκτείνεται σημαντικά, καλύπτοντας μεγαλύτερο μέρος του χώρου τιμών της εισόδου.

Γενικές Παρατηρήσεις:

Σε όλα τα μοντέλα, η εκπαίδευση επέφερε σημαντικές τροποποιήσεις στις συναρτήσεις συμμετοχής. Η αρχική, συμμετρική διαμόρφωση των συναρτήσεων συμμετοχής, όπως είχε οριστεί πριν την εκπαίδευση, δεν περιέγραφε επαρκώς το πρόβλημα και τα δεδομένα. Μετά την εκπαίδευση, τα σύνολα προσαρμόστηκαν έτσι ώστε να αντικατοπτρίζουν τις πραγματικές ιδιαιτερότητες των δεδομένων, κάτι που οδήγησε σε μείωση του σφάλματος πρόβλεψης. Αυτή η μείωση είναι φυσική συνέπεια του γεγονότος ότι το μοντέλο προσαρμόζεται ώστε να περιγράφει πιο ακριβώς τα δεδομένα εκπαίδευσης.

II. Αύξηση του Συνόλου των Μεταβλητών Εξόδου

Συγκρίνοντας τα μοντέλα 1 και 2, καθώς και τα μοντέλα 3 και 4, που διαφέρουν στο πλήθος των συνόλων των μεταβλητών εισόδου, παρατηρείται το εξής: τα μοντέλα τα μοντέλα 1 και 2 παρουσιάζουν μεγαλύτερο σφάλμα (RMSE) συγκριτικά με τα μοντέλα 3 και 4. Αυτό δείχνει ότι η αύξηση του πλήθους των συνόλων στις μεταβλητές εισόδου και η χρήση πιο σύνθετης πολυωνυμικής εξόδου βελτιώνει την απόδοση των μοντέλων.

Παράλληλα, ο συντελεστής R^2, ο οποίος μετρά την προβλεπτική ακρίβεια του μοντέλου, είναι Διαμέριση επίσης βελτιωμένος, υποδεικνύοντας ότι αυτά τα πιο σύνθετα μοντέλα με πολυωνυμική έξοδο είναι σε θέση να προβλέπουν με μεγαλύτερη ακρίβεια τις εξόδους με βάση τις εισόδους. Επομένως, τα μοντέλα 3 και 4 φαίνεται να είναι προτιμότερα για εφαρμογές όπου απαιτείται μεγαλύτερη ακρίβεια, ακόμα κι αν η πολυπλοκότητα της προσομοίωσης αυξάνεται..

Συνεπώς, το $4^ο$ μοντέλο είναι πιθανότατα προτιμότερο όσον αφορά την ακρίβεια και την απόδοση, καθώς συνδυάζει καλύτερη προβλεπτική ικανότητα με χαμηλότερα σφάλματα.

III. Διαφορές στην Απόδοση των Μοντέλων

Αναφορικά με την απόδοση των μοντέλων, τα μοντέλα με πολυωνυμική έξοδο (TSK μοντέλα 3 και 4) εμφανίζουν καλύτερα αποτελέσματα σε σύγκριση με εκείνα που έχουν σταθερή έξοδο (TSK μοντέλα 1 και 2). Συγκεκριμένα, το σφάλμα RMSE είναι χαμηλότερο στα μοντέλα με περισσότερες συναρτήσεις συμμετοχής και πολυωνυμική έξοδο, γεγονός που υποδεικνύει ότι η χρήση πιο σύνθετων μοντέλων βελτιώνει την απόδοση.

Αυτό μπορεί να εξηγηθεί από την ικανότητα των πολυωνυμικών μοντέλων να προσαρμόζονται καλύτερα σε μη γραμμικές σχέσεις στα δεδομένα. Παρά την πολυπλοκότητα της διαδικασίας εκπαίδευσης, που απαιτεί περισσότερο χρόνο προσομοίωσης, η τελική απόδοση των μοντέλων αυτών δικαιολογεί την αύξηση της πολυπλοκότητας.

Γενικά, το μοντέλο 4, που συνδυάζει 3 συναρτήσεις συμμετοχής για κάθε μεταβλητή εισόδου και πολυωνυμική έξοδο, είναι το πιο αποδοτικό, καθώς εμφανίζει το χαμηλότερο μέσο τετραγωνικό σφάλμα.

Υπερεκπαίδευση (Overfitting)

Επίσης, μια γενική παρατήρηση αφορά την υπερεκπαίδευση (overfitting) στα μοντέλα. Μετά από έναν συγκεκριμένο αριθμό επαναλήψεων (iterations), παρατηρείται ότι το σφάλμα επικύρωσης (validation error) παρουσιάζει διακυμάνσεις, ενώ το σφάλμα εκπαίδευσης (training error) μειώνεται με πιο αργό ρυθμό. Αυτή η διαφορά στην εξέλιξη των σφαλμάτων είναι ένδειξη υπερεκπαίδευσης, όπου το μοντέλο αρχίζει να "μαθαίνει" τα δεδομένα εκπαίδευσης πολύ καλά, αλλά δεν γενικεύει σωστά σε νέα δεδομένα.

Αυτό σημαίνει ότι τα μοντέλα, μετά από ένα συγκεκριμένο σημείο, δεν κερδίζουν ουσιαστική βελτίωση με περαιτέρω εκπαίδευση. Μάλιστα, σε πολλές περιπτώσεις, θα μπορούσαμε να διακόψουμε την εκπαίδευση νωρίτερα, χωρίς να επιτρέψουμε να εμφανιστεί υπερεκπαίδευση, και να επιτύχουμε παρόμοια επίπεδα σφάλματος.

Σύνοψη 1ου μέρους

Παραθέτονται συγκεντρωτικά οι μετρικές απόδοσης όλων των μοντέλων:

```
Minimal training RMSE = 0.893576
Minimal checking RMSE = 2.20071
            TSK_Model_1    TSK_Model_2    TSK_Model_3    TSK_Model_4
            ___________    ___________    ___________    ___________

    MSE        13.183         10.895         6.4347         5.0512
    RMSE       3.6308         3.3008         2.5367         2.2475
    NMSE      0.29938        0.24743        0.14613        0.11471
    NDEI      0.54716        0.49743        0.38227        0.33869
    R2        0.70062        0.75257        0.85387        0.88529
```

Εικόνα 1.13 : Αποτελέσματα για τις μετρικές από Matlab

- Αύξηση του πλήθους των συναρτήσεων συμμετοχής (MFs) και άρα των πιθανών τιμών ανά ασαφή μεταβλητή εισόδου οδηγεί σε καλύτερα αποτελέσματα για ίδια μορφή εξόδου του μοντέλου (διαφορά μοντέλου 1 από 2, διαφορά μοντέλου 3 από 4)
- Για ίδιο πλήθος συναρτήσεων συμμετοχής (MFs) ανά ασαφή μεταβλητή εισόδου, διατήρηση περισσότερων όρων στην έξοδο του κάθε κανόνα του μοντέλου sugeno (μετάβαση από σταθερή έξοδο σε πολυωνυμική οδηγεί σε καλύτερα αποτελέσματα (διαφορά μοντέλου 1 από 3, διαφορά μοντέλου 2 από 4) .

2. Εφαρμογή σε dataset υψηλής διαστασιμότητας

Το dataset του ερωτήματος, *superconduct*, αποτελείται από 21263 δείγματα (data points) με 81 features και μία τιμή εξόδου το καθένα. Αν χρησιμοποιούνταν η μέθοδος grid partitioning, για αυτό τον αριθμό χαρακτηριστικών, τότε θα δημιουργούνταν 2^{81} κανόνες. Για την αποφυγή αυτής της πολυπλοκότητας του μοντέλου, θα χρησιμοποιηθεί η μέθοδος του subtractive clustering. Η λογική βασίζεται στην ομαδοποίηση των χαρακτηριστικών (features) με στόχο εν τέλει να προκύψει ένα ακριβές TSK μοντέλο με πολύ λιγότερα features. Πιο συγκεκριμένα, σκοπός είναι η περαιτέρω μείωση του απαιτούμενου αριθμού ασαφών κανόνων (καθώς αυτοί θα λάβουν σαν είσοδο όχι τα features αλλά τα ομαδοποιημένα features). Οι δύο τεχνικές που θα χρησιμοποιηθούν είναι:

- Ο αλγόριθμός ReliefF για feature subset selection
- Ο αλγόριθμος Subtractive Clustering για ομαδοποίηση των features πριν την δημιουργία των κανόνων

Εισαγωγή και Διαχωρισμός Δεδομένων

Αρχικά, φορτώνεται το σύνολο δεδομένων από το αρχείο superconduct.csv . Στη συνέχεια, η τελευταία στήλη του πίνακα δεδομένων, που αντιπροσωπεύει την έξοδο (στόχο), αποθηκεύεται στη μεταβλητή dataTarget. Για την προεπεξεργασία των δεδομένων, τα δεδομένα διαχωρίζονται σε δεδομένα εκπαίδευσης (60%), δεδομένα επικύρωσης (20%) και δεδομένα δοκιμών (20%) χρησιμοποιώντας τη συνάρτηση split_scale. Το διαχωρισμό ακολουθεί η αποθήκευση των δεδομένων στόχου σε ξεχωριστές μεταβλητές.

Αρχικοποίηση Μεταβλητών για Grid Search

Σε αυτό το βήμα, αρχικοποιούνται οι βασικές μεταβλητές που θα χρησιμοποιηθούν κατά τη διάρκεια της διαδικασίας αναζήτησης grid search. Οι βασικές παράμετροι που ορίζονται περιλαμβάνουν :

- Αριθμός χαρακτηριστικών (numFeatures) που θα εξεταστούν: [5, 8, 10, 12, 15]
- Ακτίνα clustering (clusterRadius) που θα χρησιμοποιηθεί: [0.2, 0.4, 0.6, 0.8 ,1]
- Αριθμός folds (numFolds) για τη διασταυρούμενη επικύρωση: 5
- Η μέθοδος clustering που επιλέγεται είναι το Subtractive Clustering

Μέθοδος Επιλογής Χαρακτηριστικών ReliefF

Η μέθοδος ReliefF χρησιμοποιείται για την επιλογή των πιο σημαντικών χαρακτηριστικών του dataset. Η διαδικασία αυτή λαμβάνει υπόψη τον αριθμό των πλησιέστερων γειτόνων (numNearestNeighbors = 10) και εκτελείται για την εκτίμηση της βαρύτητας των χαρακτηριστικών για το πρόβλημα παλινδρόμησης που εξετάζουμε. Οπότε με βάση τον αριθμό των features που επιλέγεται σε κάθε επανάληψη, η χρήση του relieff αλγορίθμου εξάγει τα indices των χαρακτηριστικών που θα επιλεχθούν.

Διαδικασία Grid Search

Το grid search είναι η διαδικασία κατά την οποία δοκιμάζονται διαφορετικές παραμετροποιήσεις για να βρεθεί το καλύτερο μοντέλο. Στην συγκεκριμένη περιπτωση, εξετάζονται όλοι οι συνδυασμοί από τα numFeatures και clusterRadius. Για κάθε συνδυασμό εκπαιδεύεται ένα μοντέλο και χρησιμοποιείται η 5-fold cross validation για την εκτίμηση του σφάλματος (MSE). Για κάθε ένα συνδυασμό αριθμού χαρακτηριστικών –αριθμού κανόνων στο grid, αναζητείται το σφάλμα εκπαίδευσης στο validation set (validation error) και μέσω αυτού θα καταλήξουμε στο βέλτιστο συνδυασμό ή βέλτιστο σημείο στο grid (optimum grid point).

Cross Validation

Για αύξηση της εγκυρότητας των προσομοιώσεων και των υπολογισμών του μέσου σφάλματος, χρησιμοποιείται και η μέθοδος του 5-fold cross validation. Πιο συγκεκριμένα, η ίδια διαδικασία για κάθε ζεύγος τιμών των παραμέτρων, επαναλαμβάνεται 5 φορές με διαφορετικό διαμερισμό του training set κάθε φορά. Μετά το partitioning των training data, δημιουργούνται δύο υποσύνολα, το training και το testing, βάση των οποίων θα ακολουθήσει εκπαίδευση για να εξαχθεί το mean error.

Από το σύνολο των 5 επαναλήψεων, εν τέλει υπολογίζεται το mean error του training μοντέλου, για κάθε ζεύγος παραμέτρων και αποθηκεύεται σε έναν πίνακα τιμών. Εφ΄ όσον ολοκληρωθεί η διαδικασία για όλα τα ζεύγη των παραμέτρων εισόδου, υπολογίζεται το ελάχιστο από τα μέσα σφάλματα. Με αυτό τον τρόπο βρίσκονται οι καλύτερες παράμετροι για το TSK model.

Δημιουργία αρχικού Fuzzy Inference System (FIS) με Subtractive Clustering

Για το κάθε υποσύνολο που προκύπτει (training και validation set) χρησιμοποιείται το Subtractive Clustering για να δημιουργηθεί ένα αρχικό Fuzzy Inference System (FIS).

Substractive Clustering

Ο αλγόριθμος Subtractive Clustering (SC) δεν δέχεται απευθείας τον αριθμό των clusters που θα σχηματίσει, αλλά μια (normalized) ακτίνα αναζήτησης στο χώρο των εισόδων. Πρέπει επομένως για κάθε grid point (συνδυασμό αριθμού χαρακτηριστικών –αριθμού κανόνων) και για κάθε ένα από τα cross-validation runs (5 όσα και τα folds) να βρεθεί η ακτίνα του SC που δίνει τον ζητούμενο αριθμό clusters

για το dataset με τον εκάστοτε αριθμό χαρακτηριστικών. Μετά τη δημιουργία του FIS, γίνεται έλεγχος του αριθμού των παραγόμενων κανόνων. Αν το μοντέλο παράγει λιγότερους από 2 κανόνες, η διαδικασία παραλείπεται, καθώς ένα τέτοιο μοντέλο δεν θεωρείται επαρκές για να μοντελοποιήσει τη σχέση στα δεδομένα.

Εκπαίδευση του ANFIS μοντέλου

Αν ο αριθμός κανόνων είναι επαρκής, το **ANFIS (Adaptive Neuro-Fuzzy Inference System)** μοντέλο εκπαιδεύεται χρησιμοποιώντας τα δεδομένα εκπαίδευσης. Το αρχικό FIS που δημιουργήθηκε από το Subtractive Clustering χρησιμοποιείται ως βάση για την εκπαίδευση.

Αποτελέσματα –Optimal Grid Point

Για κάθε ένα από τα cross-validation runs κάνουμε τα εξής:

1. Εξάγουμε από το αρχικό training set, το training set και το test set που θα χρησιμοποιηθούν σε αυτό το fold, με βάση τα indices της.

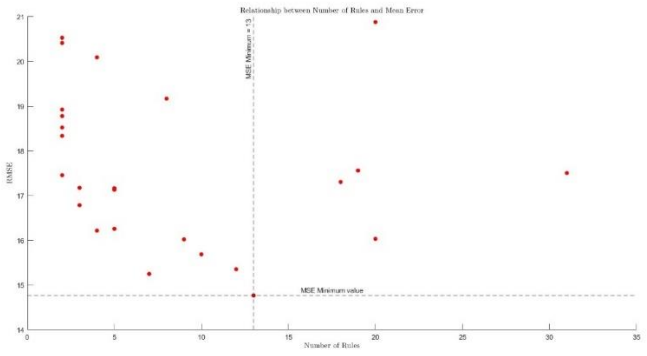
2. Για κάθε grid point, ακολουθούνται τα εξής βήματα:

- Βρίσκεται η ακτίνα ομαδοποίησης για το SC, για το συγκεκριμένο αριθμό κανόνων-clusters, για το συγκεκριμένο αριθμό χαρακτηριστικών και για το συγκεκριμένο training set του fold
- Εκπαιδεύεται με βάση το training set του fold ένα TSK μοντέλο με SC στο χώρο των εισόδων με παράμετρο την παραπάνω ακτίνα αναζήτησης
- Υπολογισμός του τελικού (ως προς τα epochs) validation error του μοντέλου στο test set του fold.

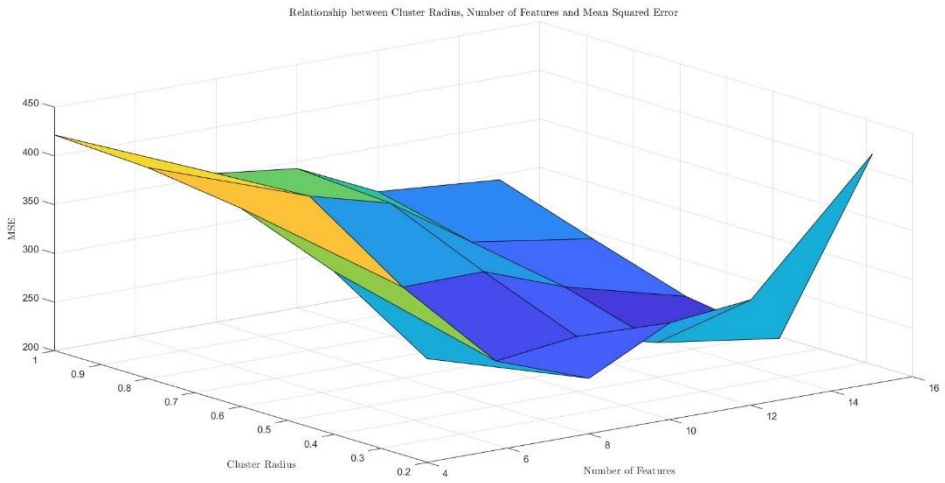
Από το grid search προέκυψε ότι :

```
Optimal Cluster Radius: 0.40
Optimal Number of Features: 15
Minimum Error (MSE): 218.0242
```

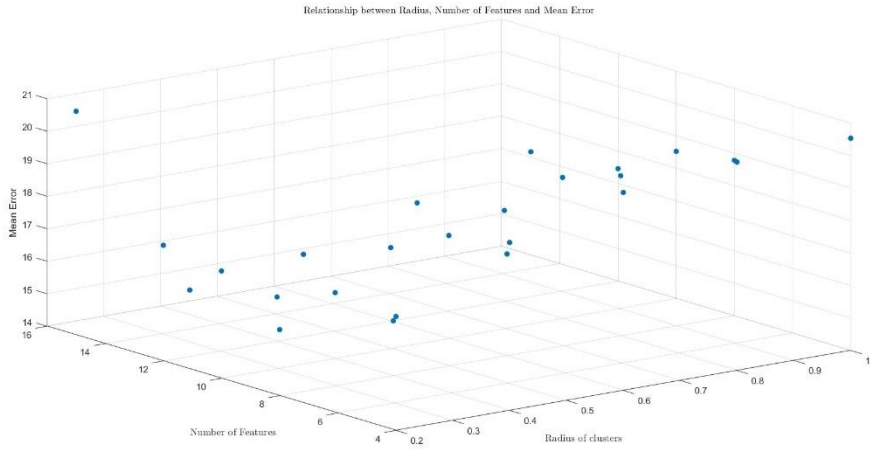
Παρακατω, παρατίθενται κάποια ενδεικτικά διαγράμματα οπου απεικονίζουν το MSE, RMSE συναρτήσει των αριθμό χαρακτηριστικών numFeatures και της ακτίνας clusterRadius με βάση τα αποτελέσματα που προέκυψαν :



Εικόνα 2.1:Διάγραμμα RMSE συναρτήσει του αριθμού των κανόνων



Εικόνα 2.2:Διάγραμμα μέσου τετραγωνικού σφάλματος συναρτήσει του αριθμού χαρακτηριστικών και της ακτίνας των clusters



Εικόνα 2.3:Διάγραμμα μέσου σφάλματος συναρτήσει του αριθμού χαρακτηριστικών και της ακτίνας των clusters

Επίσης παρατίθενται τα παρακάτω αποτελέσματα από τις μετρικές NDEI, RMSE, R^2, NMSE.

RMSE

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	17.8941	19.1064	19.8727	20.6689	20.6856
8	18.5853	16.6248	17.8722	19.4362	19.1678
10	33.0596	16.4357	18.0069	18.9272	19.2367
12	26.7277	15.5645	16.4753	17.3839	17.7883
15	19.6902	15.3022	15.3886	16.4345	17.3030

NMSE

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.2740	0.3124	033.79	0.3655	0.3661
8	0.3003	0.2366	0.2734	0.3233	0.3144
10	1.0969	0.2312	0.2776	0.3067	0.3167
12	0.7802	0.2073	0.2323	0.2588	0.2708
15	0.3478	0.2006	0.2026	0.2311	0.2562

NDEI

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.5234	0.5589	0.5813	0.6046	0.6051
8	0.5437	0.4863	0.5228	0.5685	0.5607
10	0.9671	0.4808	0.5267	0.5537	0.5627
12	0.7818	0.4553	0.4819	0.5085	0.5203
15	0.5760	0.4476	0.4501	0.4807	0.5061

R^2

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.7260	0.6876	0.6621	0.6345	0.6339
8	0.6997	0.7634	0.7266	0.6767	0.6856
10	-0.0969	0.7688	0.7224	0.6933	0.6833
12	0.2198	0.7927	0.7677	0.7412	0.7292
15	0.6522	0.7994	0.7974	0.7689	0.7438

Από τα αποτελέσματα του grid search, φαίνεται ότι η μείωση του αριθμού των χαρακτηριστικών (features) οδηγεί σε αυξημένα validation errors, κάτι αναμενόμενο λόγω της υψηλής διαστασιμότητας του dataset. Συνεπώς, είναι λογικό ο βέλτιστος συνδυασμός να περιλαμβάνει τον μέγιστο αριθμό διαθέσιμων χαρακτηριστικών, δηλαδή 15, όπως και παρατηρείται.

Αναφορικά με την ακτίνα cluster, παρατηρούμε ότι για έναν δεδομένο αριθμό χαρακτηριστικών, μικρότερες τιμές ακτίνας οδηγούν σε χαμηλότερα validation errors. Στην περίπτωση των 15 χαρακτηριστικών, η βέλτιστη ακτίνα είναι 0.4, ενώ κοντά στο ελάχιστο σφάλμα βρίσκονται και άλλες χαμηλές τιμές ακτίνας, όπως η 0.2. Αυτή η τάση υποδεικνύει ότι μικρότερη ακτίνα βελτιώνει την ακρίβεια του μοντέλου.

Επισης, για σταθερό αριθμό κρατημένων features, παρατηρειται πως το RMSE αυξάνεται όσο αυξάνεται η ακτίνα r των clusters. Από την άλλη, κρατώντας σταθερή την ακτίνα, παρατηρειται ότι αυξάνοντας τον αριθμό των features, τόσο μείωνεται το RMSE. Αυτό συμβαίνει, διότι, όσο περισσότερα features επιλεχθούν, τόση λιγότερη πληροφορία χάνεται από το dataset, ενώ ο χρόνος εκπαίδευσης αυξάνεται εκθετικά. Επιπλέον, μικρότερη ακτίνα r των clusters σημαίνει πως το μοντέλο μαθαίνει

καλύτερα το dataset, με αποτέλεσμα να πραγματοποιεί καλύτερες προβλέψεις. Η συνεχής όμως μείωση της ακτίνας οδηγεί το μοντέλο σε overfitting, με αποτέλεσμα την εξάλειψη της γενικότητας, κάτι που δεν είναι επιθυμητό. Γι' αυτό και το RMSE είναι καλύτερο για r = 0.4 και όχι για r = 0.2 με αριθμό κρατημένων features = 15.

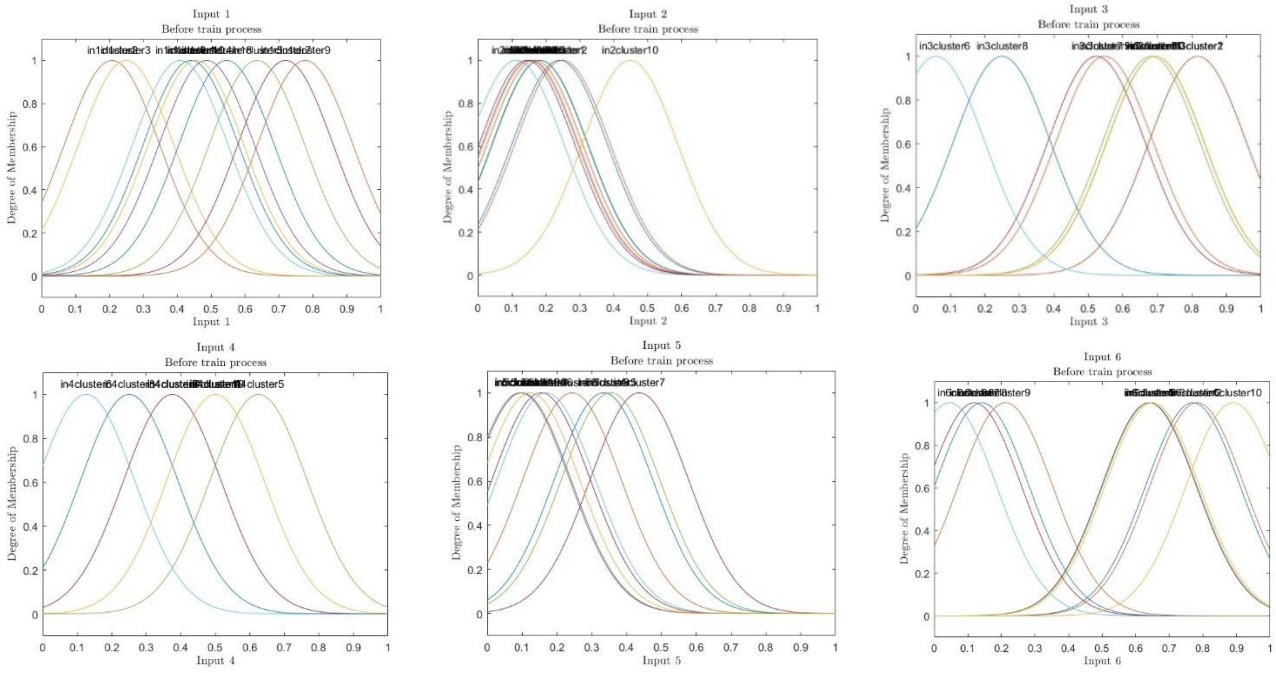
Εκπαίδευση του Βέλτιστου TSK Μοντέλου

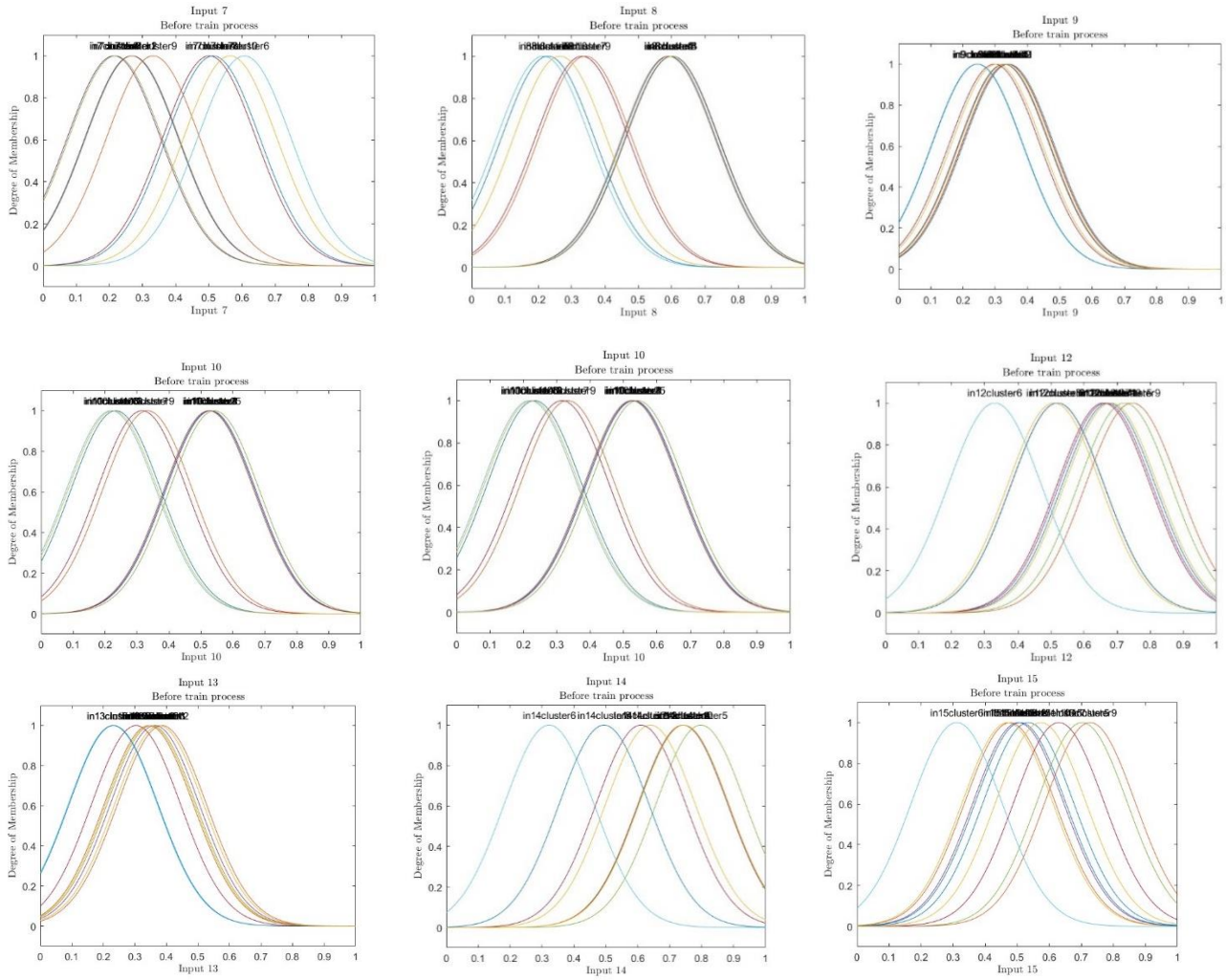
Ακολουθώντας τις παραμέτρους που προέκυψαν από τη διαδικασία του grid search, δημιουργείται το βέλτιστο TSK μοντέλο με τη χρήση της μεθόδου subtractive clustering. Το μοντέλο (fis) βασίζεται σε 25 επιλεγμένα χαρακτηριστικά και κάθε cluster έχει ακτίνα 0.4. Οι παράμετροι αυτές επιλέγονται με τη μέθοδο relieff, η οποία χρησιμοποιείται και σε αυτό το στάδιο για την κατάλληλη επιλογή των χαρακτηριστικών. Στη συνέχεια, το μοντέλο εκπαιδεύεται και αξιολογείται με τις ίδιες μετρικές που χρησιμοποιήθηκαν στο προηγούμενο μέρος της εργασίας, ενώ τα αποτελέσματα παρουσιάζονται αναλυτικά στη συνέχεια.

Αξιολόγηση της εκπαίδευσης του βέλτιστου TSK μοντέλου

Συναρτήσεις Συμμετοχής (MFs) Τελικού Μοντέλου

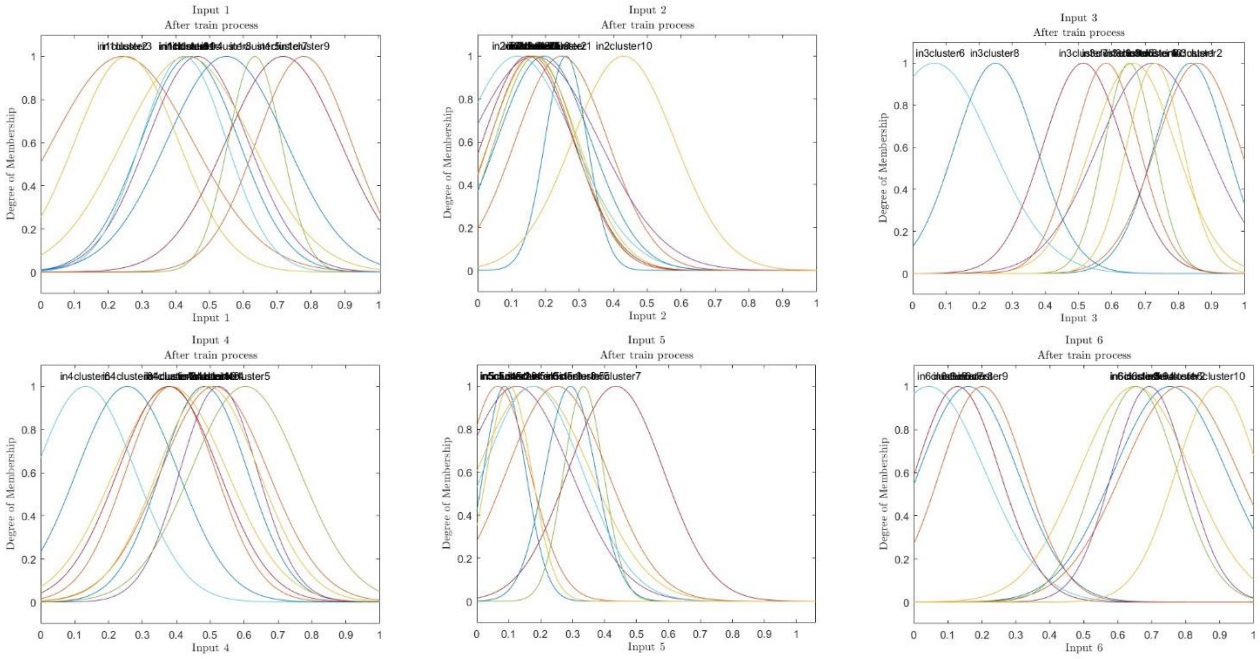
Παρακάτω, δίνονται οι αρχικές συναρτήσεις συμμετοχής (MFs)των λεκτικών τιμών - clusters για τις εισόδους-features του τελικού TSK μοντέλου:

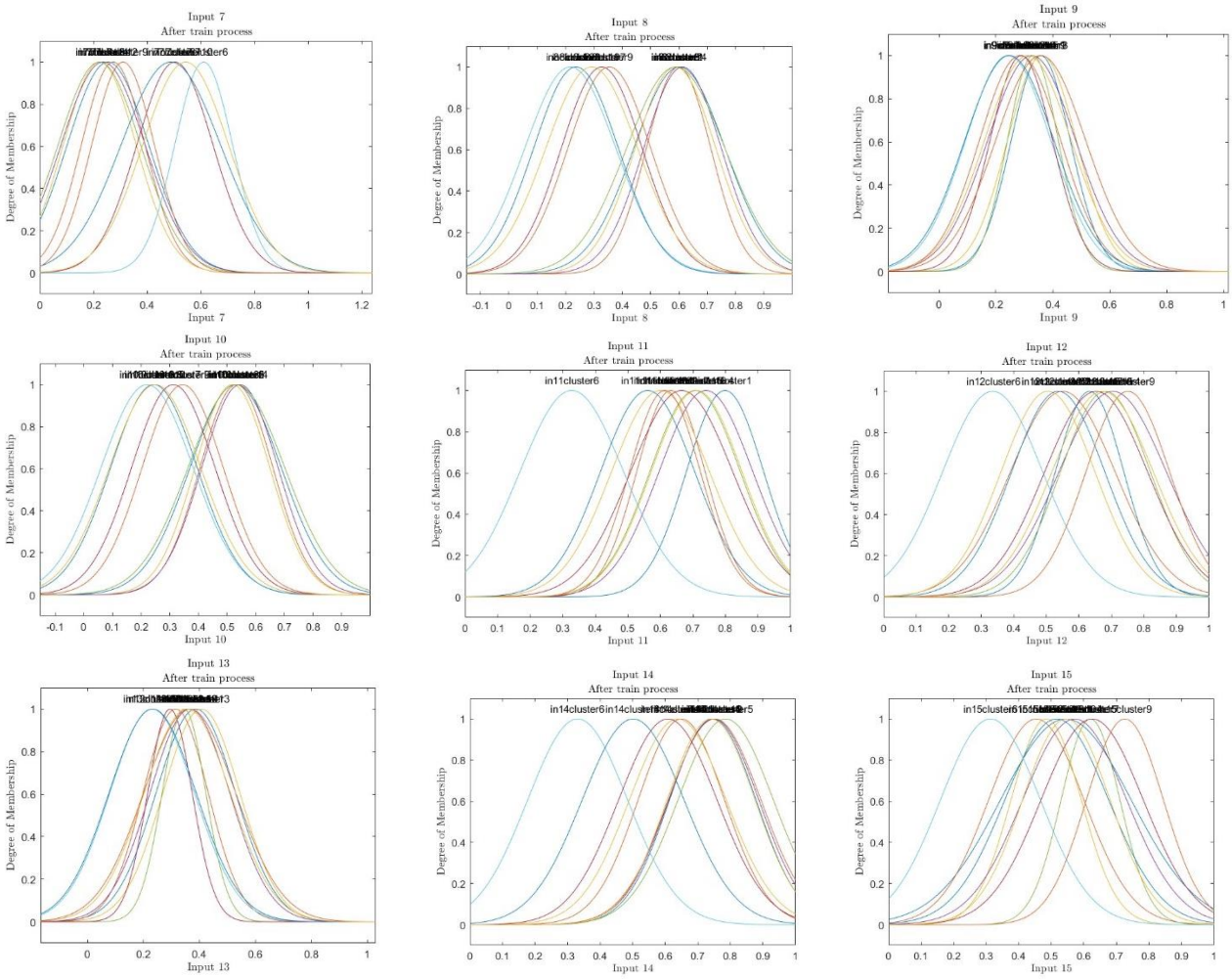




Εικόνα 2.4: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου τελικού μοντέλου

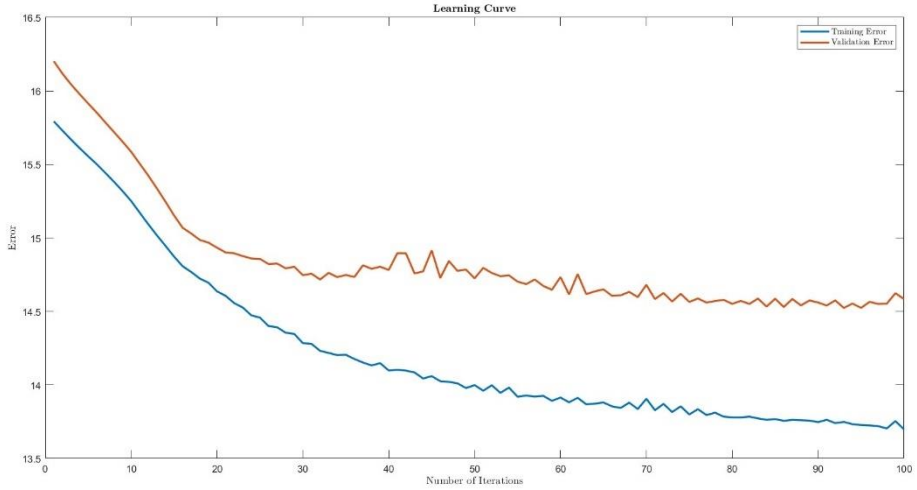
ενώ ακολούθως δίνονται οι τελικές μορφές των παραπάνω MFs:





Εικόνα 2.5 : Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου τελικού μοντέλου

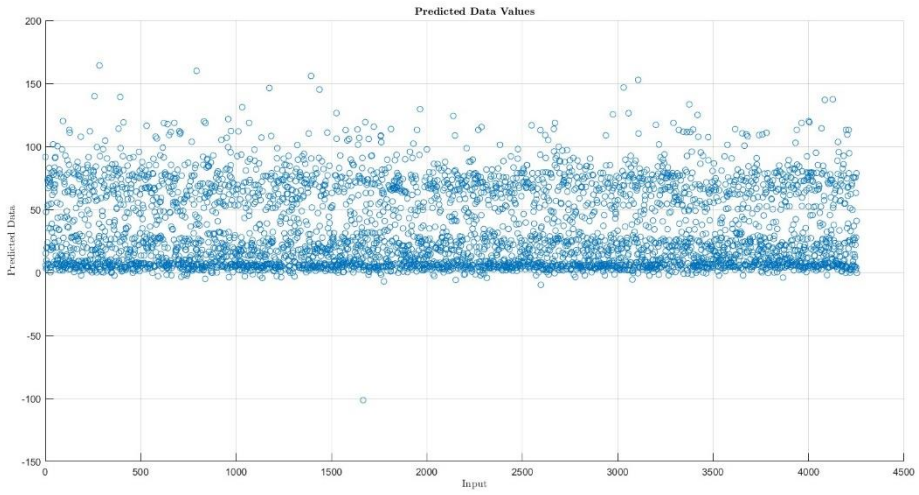
Παρακάτω, δίνονται οι καμπύλες μάθησης (learning curves) του τελικού TSK μοντέλου με τον βέλτιστο συνδυασμό χαρακτηριστικών –κανόνων:



Εικόνα 2.6: Διάγραμμα καμπυλών εκμάθησης του βέλτιστου TSK μοντέλου

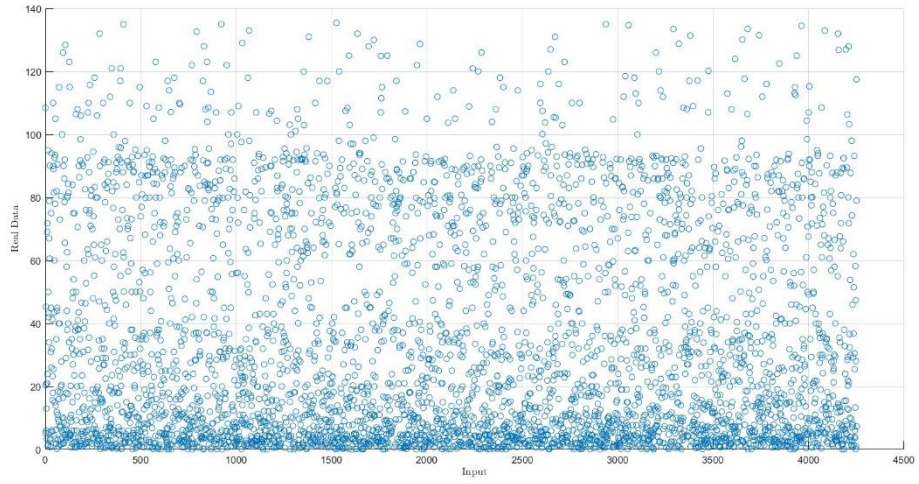
όπου όπως φαίνεται, για τον χρησιμοποιούμενο αριθμό epochs (100) δεν υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting).

Προβλέψεις τελικού μοντέλου



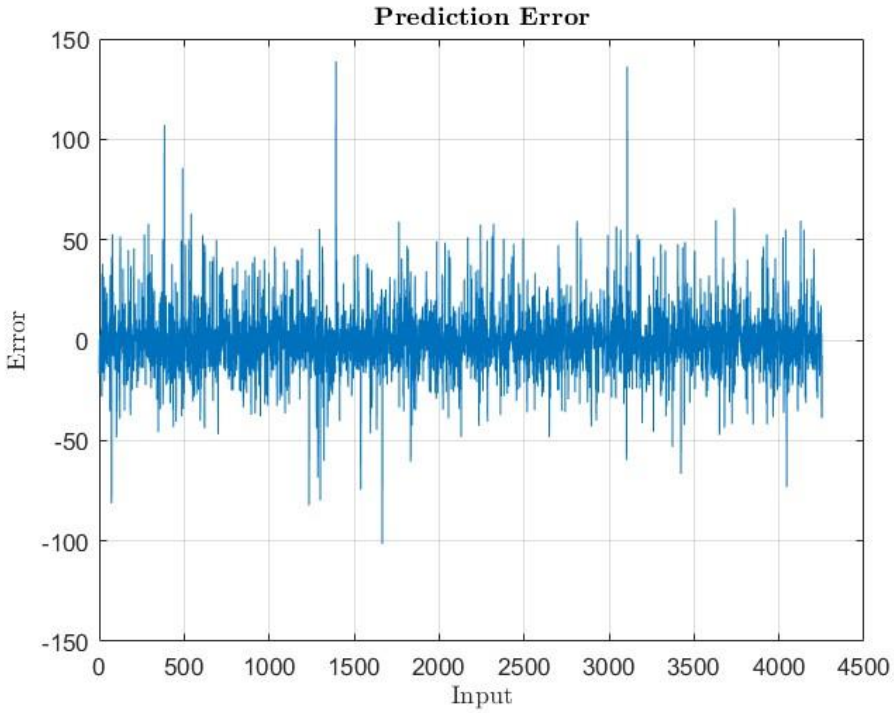
Εικόνα 2.7 : **Προβλέψεις τελικού μοντέλου**

Πραγματικές τιμές target του dataset



Εικόνα 2.8 : Πραγματικές τιμές target του dataset

Ακολούθως, σε ότι αφορά την απόδοση του τελικού μοντέλου, δίνεται το διάγραμμα με τα σφάλματα πρόβλεψης του μοντέλου ως προς τις πραγματικές τιμές του test set:



Εικόνα 2.9: Σφάλματα πρόβλεψης κατά την εφαρμογή του τελικού TSK μοντέλου στο test set

Τέλος, παρατίθενται οι μετρικές αξιολόγησης για το βέλτιστο TSK μοντέλο :

Optimal Model

MSE	225.7
RMSE	15.023
NMSE	0.19312
NDEI	0.43946
R2	0.80688

Όπως φαίνεται το μοντέλο με τον βέλτιστο συνδυασμό αποδίδει με δείκτη R2 λίγο πάνω από 80%, αποτέλεσμα ικανοποιητικό.

Σχολιασμός Αποτελεσμάτων/ Συμπεράσματα

Από το σύνολο των 83 αρχικών χαρακτηριστικών, η βέλτιστη αναπαράσταση του μοντέλου επιτυγχάνεται με την επιλογή 15 χαρακτηριστικών, χρησιμοποιώντας τον αλγόριθμο ReliefF. Αυτό σημαίνει ότι η αύξηση του αριθμού των χαρακτηριστικών συνήθως βελτιώνει την περιγραφή του μοντέλου. Ωστόσο, για μείωση της πολυπλοκότητας (ώστε να υπάρχουν λιγότεροι κανόνες IF-THEN), θα μπορούσε να επιλεχθεί μικρότερος αριθμός χαρακτηριστικών, όπως 10, διατηρώντας το μέσο σφάλμα σε παρόμοια επίπεδα.

Παρατηρείται διαφοροποίηση στις μετρικές απόδοσης των μοντέλων, ειδικά σε σχέση με τη μέθοδο grid search. Η τιμή της μετρικής RMSE είναι μεγαλύτερη σε σχέση με τα μοντέλα Takagi-Sugeno-Kang (TSK) του Α' μέρους, καθώς εκεί υπήρχε πιο ακριβής περιγραφή του μοντέλου. Αντίθετα, στο Β' μέρος, η προσέγγιση είναι περισσότερο αφαιρετική, με ελαφρώς μειωμένη ακρίβεια.

Η μέθοδος grid search δεν είναι κατάλληλη για σύνολα δεδομένων με μεγάλη διασπορά. Σε αυτές τις περιπτώσεις, η χρήση του subtractive clustering είναι προτιμότερη, αν και συνοδεύεται από μικρή θυσία στην απόδοση του training.

Τέλος, στα ασαφή σύνολα παρατηρούνται μεταβολές στα fuzzy sets, όπως ήταν αναμενόμενο, λόγω της διαδικασίας εκπαίδευσης. Αυτή η προσαρμογή των fuzzy sets επιτρέπει στο μοντέλο να εκπαιδευτεί καλύτερα και να αποδώσει με μεγαλύτερη ακρίβεια στις προβλέψεις του.