ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΣΑΦΗ ΣΥΣΤΗΜΑΤΑ

8ο ΕΞΑΜΗΝΟ

ΕΡΓΑΣΙΑ #3

ΕΙΣΗΓΗΤΗΣ: ΘΕΟΧΑΡΗΣ Ι.

Όνομα : ΘΑΝΑΣΗΣ ΧΑΡΙΣΟΥΔΗΣ

A.E.M. : 9026

ΘΕΣΣΑΛΟΝΙΚΗ, 28 Φεβρουαρίου 2020

Πίνακας Περιεχομένων

1. Εφαρμογή σε απλό dataset 5
1.1. Κώδικας 1ου μέρους εργασίας 5
1.2. Φόρτωση & Προ-επεξεργασία dataset 5
1.3. Διαχωρισμός του dataset 5
1.4. Ζητούμενα Εργασίας 6
1.5. Εκπαίδευση TSK μοντέλων 6
1.5.1 Model 1: (2) Input MFs, Singleton Output 6
1.5.2 Model 2: (3) Input MFs, Singleton Output 9
1.5.3 Model 3: (2) Input MFs, Polynomial Output 12
1.5.4 Model 4: (3) Input MFs, Polynomial Output 15
1.6. Συμπερασματική ανάλυση 18
2. Εφαρμογή σε high-dimensional dataset 20
2.1. Κώδικας 2ου μέρους εργασίας 20
2.2. Φόρτωση & Προ-επεξεργασία dataset 20
2.3. Μείωση Διαστασιμότητας Dataset 20
2.4. Grid Search: Εύρεση Βέλτιστου Συνδυασμού Αριθμού Features – Αριθμού Κανόνων 21
2.4.1 Ακτίνες αναζήτησης για Subtractive Clustering (SC) 21
2.4.2 Cross Validation 23
2.4.3 Αποτελέσματα – Optimum Grid Point 23
2.5. Τελικό Μοντέλο με βάση το Optimum Grid Point 26
2.5.1 Απόδοση Τελικού Μοντέλου 26
2.5.2 Συναρτήσεις Συμμετοχής (MFs) Τελικού Μοντέλου 27
2.6. Συμπερασματικές Παρατηρήσεις – Σχόλια 29

Πίνακας Εικόνων

Εικόνα 1: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 1ου μοντέλου 6
Εικόνα 2: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 1ου μοντέλου 7
Εικόνα 3: Καμπύλες μάθησης 1ου μοντέλου 8
Εικόνα 4: Σφάλματα πρόβλεψης κατά την εφαρμογή του 1ου μοντέλου στο test set 8
Εικόνα 5: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 2ου μοντέλου 9

Εικόνα 6: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 2ου μοντέλου....10
Εικόνα 7: Καμπύλες μάθησης 2ου μοντέλου....11
Εικόνα 8: Σφάλματα πρόβλεψης κατά την εφαρμογή του 2ου μοντέλου στο test set....11
Εικόνα 9: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 3ου μοντέλου....12
Εικόνα 10: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 3ου μοντέλου....13
Εικόνα 11: Καμπύλες μάθησης 3ου μοντέλου....14
Εικόνα 12: Σφάλματα πρόβλεψης κατά την εφαρμογή του 3ου μοντέλου στο test set....14
Εικόνα 13: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 4ου μοντέλου....15
Εικόνα 14: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 4ου μοντέλου....16
Εικόνα 15: Καμπύλες μάθησης 4ου μοντέλου....17
Εικόνα 16: Σφάλματα πρόβλεψης κατά την εφαρμογή του 4ου μοντέλου στο test set....17
Εικόνα 17: (Μέση) Ακτίνα SC ως προς αριθμό κανόνων και αριθμό χαρακτηριστικών (grid point)....23
Εικόνα 18: 3D plot του μέσου (τελικού) validation error για κάθε grid point ως προς τα 5 folds....25
Εικόνα 19: Καμπύλες μάθησης τελικού TSK μοντέλου με overfitting (75 epochs)....26
Εικόνα 20: Σφάλματα πρόβλεψης κατά την εφαρμογή του τελικού TSK μοντέλου στο test set....27
Εικόνα 21: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου τελικού μοντέλου....28
Εικόνα 22: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου τελικού μοντέλου....28

Πίνακας Πινάκων

Πίνακας 1: Παράμετροι των TSK μοντέλων που θα εκπαιδευτούν....6
Πίνακας 2: Μετρικές Απόδοσης του 1ου μοντέλου....9
Πίνακας 3: Μετρικές Απόδοσης του 2ου μοντέλου....12
Πίνακας 4: Μετρικές Απόδοσης του 3ου μοντέλου....15
Πίνακας 5: Μετρικές Απόδοσης του 4ου μοντέλου....18

Πίνακας 6: Συγκεντρωτικές Μετρικές Απόδοσης όλων των μοντέλων.....18
Πίνακας 7: Μέση τιμή ακτινών που αντιστοιχούν στους συνδυασμούς αριθμών κανόνων και αριθμών χαρακτηριστικών (grid points).....22
Πίνακας 8: Μέσο (τελικό) validation error για κάθε grid point ως προς τα 5 folds.....24
Πίνακας 9: Ακτίνες SC του optimal grid point σε κάθε cross-validation run.....26
Πίνακας 10: Μετρικές Απόδοσης του τελικού TSK μοντέλου.....27
Πίνακας 11: Indices των 21 πιο σημαντικών features από ReliefF (K=100) 29

1. Εφαρμογή σε απλό dataset

1.1. Κώδικας 1ου μέρους εργασίας

Ο κώδικας σε MATLAB που υλοποιεί το μέρος α' της εργασίας βρίσκεται στο αρχείο */3/matlab/main_a.m*. Εκεί υπάρχει μόνο η λογική της εκτέλεσης, ενώ βοηθητικές κλάσεις και συναρτήσεις υπάρχουν στο φάκελο */Matlab Helpers/**.

1.2. Φόρτωση & Προ-επεξεργασία dataset

Το dataset του ερωτήματος, *CCPP*, αποτελείται από 9568 δείγματα (data points) με (4) features και μία τιμή εξόδου το καθένα. Δεν υλοποιήθηκε κάποια μέθοδος προ-επεξεργασίας του dataset καθώς τα αποτελέσματα ήταν καλά. Το μόνο που γίνεται μετά την φόρτωση είναι ο έλεγχος για διπλότυπα δείγματα (~40 δείγματα αφαιρέθηκαν).

1.3. Διαχωρισμός του dataset

Για το διαχωρισμό του dataset στα **training** (χρησ. στο training), **validation** (χρησ. στο training για αποφυγή overfitting) και **testing** (χρησ. στο testing – άγνωστο a priori) subsets δημιουργήθηκε μια μέθοδος, *AnfisWrapper.partition()*, η οποία δέχεται σαν ορίσματα το dataset και τα ποσοστά του διαχωρισμού και επιστρέφει τα τρία subsets παραπάνω. Ο τρόπος που υλοποιεί το splitting είναι μία τύπου round-robin ανάθεση στοιχείων στα subsets βάσει των ποσοστών. Για παράδειγμα, για splitting 60% - 20% - 20% η μεταβλητή *split* θα είναι *[0.6, 0.2, 0.2]*, ενώ η ανάθεση θα γίνει ως εξής:

```
split10 = 10 * split = [6, 2, 2];
Για κάθε 10άδα στοιχείων του dataset, do:
      Βάλε 6 στο training
      Βάλε 2 στο validation
      Βάλε 2 στο testing
end
```

Έτσι προκύπτει μια δίκαια ανάθεση με τον τελικό αριθμό στοιχείων σε κάθε ένα από τα 3 subsets να ισούται με *(x*|dataset| ± 1)*, με x το ποσοστό του splitting ανά subset.

1.4. Ζητούμενα Εργασίας

Ζητείται να εκπαιδευθούν τέσσερα (4) TSK μοντέλα για το παραπάνω dataset με (4) εισόδους το καθένα και πλήθος MFs για κάθε είσοδο και τύπο εξόδου του sugeno μοντέλου, που καθορίζεται ως εξής:

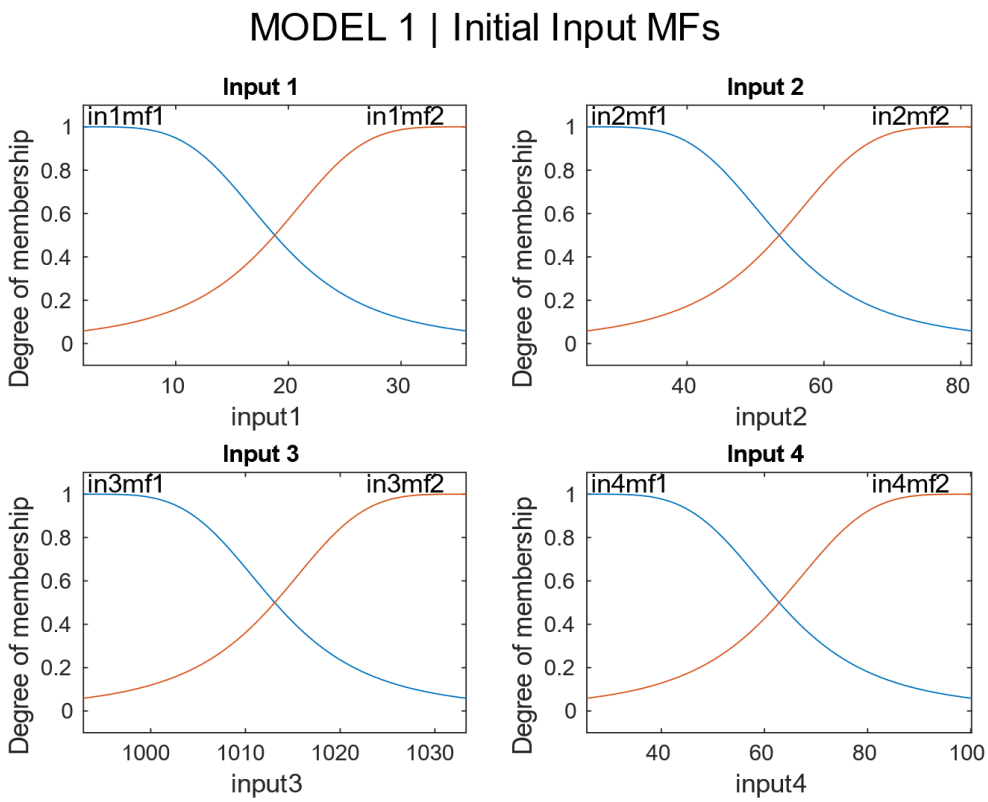
	Πλήθος MFs Μεταβλητών Εισόδου	Τύπος Εξόδου Κανόνων Sugeno
TSK Model 1	2	Singleton
TSK Model 2	3	Singleton
TSK Model 3	2	Linear
TSK Model 4	3	Linear

Πίνακας 1: Παράμετροι των TSK μοντέλων που θα εκπαιδευτούν

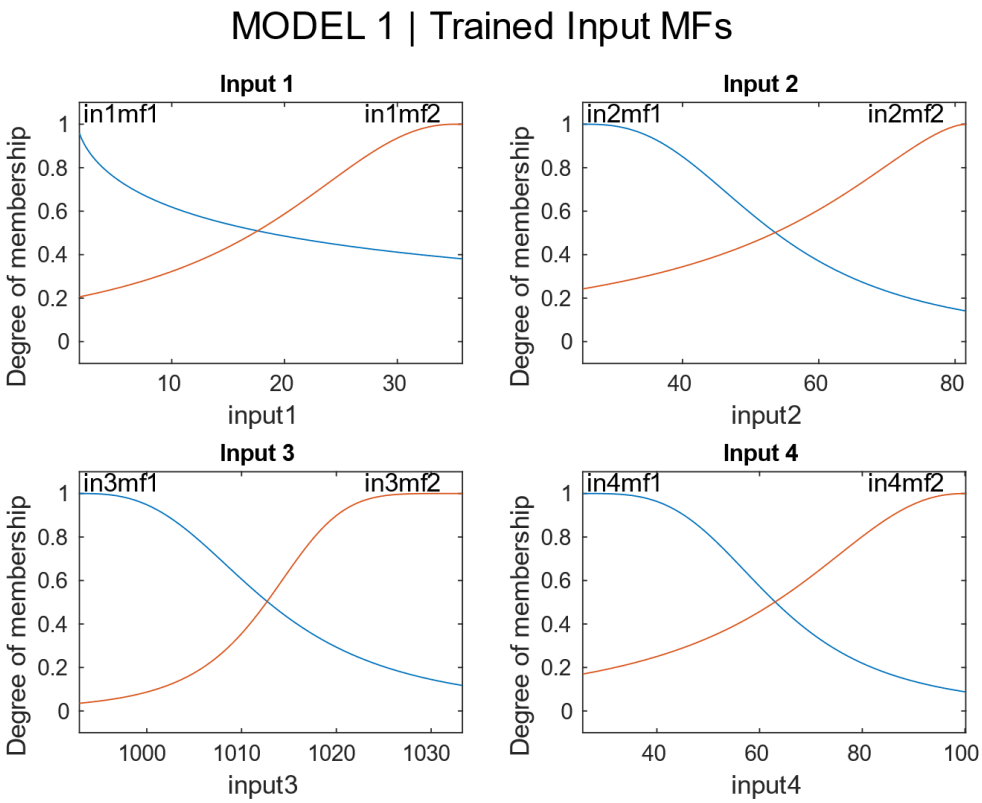
1.5. Εκπαίδευση TSK μοντέλων

1.5.1 Model 1: (2) Input MFs, Singleton Output

Αρχικά, παρατίθενται οι αρχικές και τελικές μορφές των MFs για τις ασαφείς τιμές των ασαφών μεταβλητών εισόδου των κανόνων:

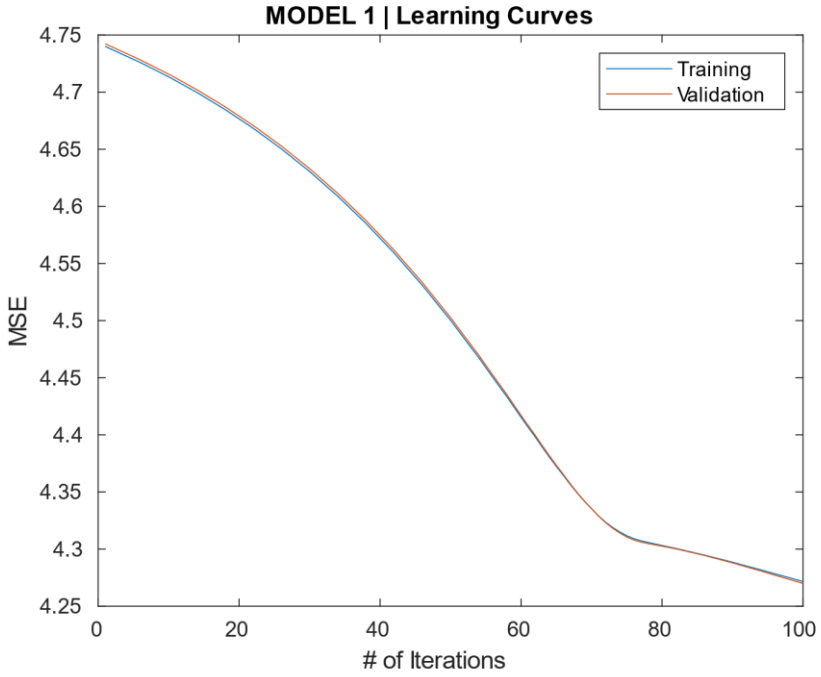


Εικόνα 1: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 1ου μοντέλου

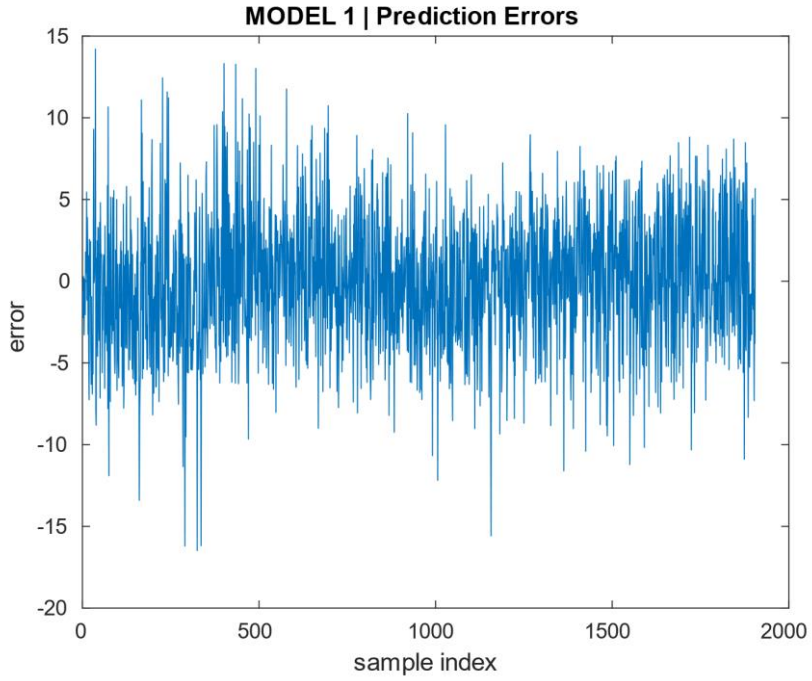


Εικόνα 2: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 1ου μοντέλου

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



Εικόνα 3: Καμπύλες μάθησης 1ου μοντέλου



Εικόνα 4: Σφάλματα πρόβλεψης κατά την εφαρμογή του 1ου μοντέλου στο test set

Τέλος, δίνονται οι ζητούμενες μετρικές απόδοσης του μοντέλου:

Metric →	MSE	RMSE	NMSE	R^2	NDEI
Value →	16.627	4.078	0.062	0.938 (**93.8%**)	0.249

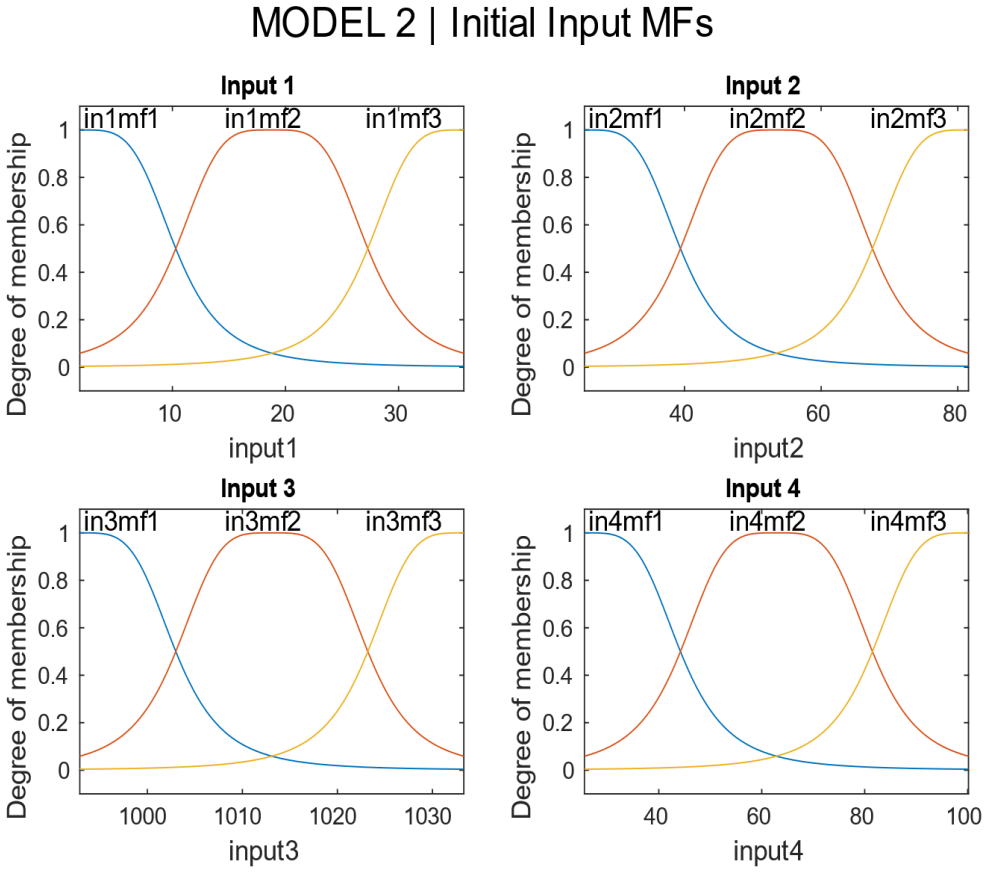
Πίνακας 2: Μετρικές Απόδοσης του 1ου μοντέλου

Μια πρώτη ανάλυση:

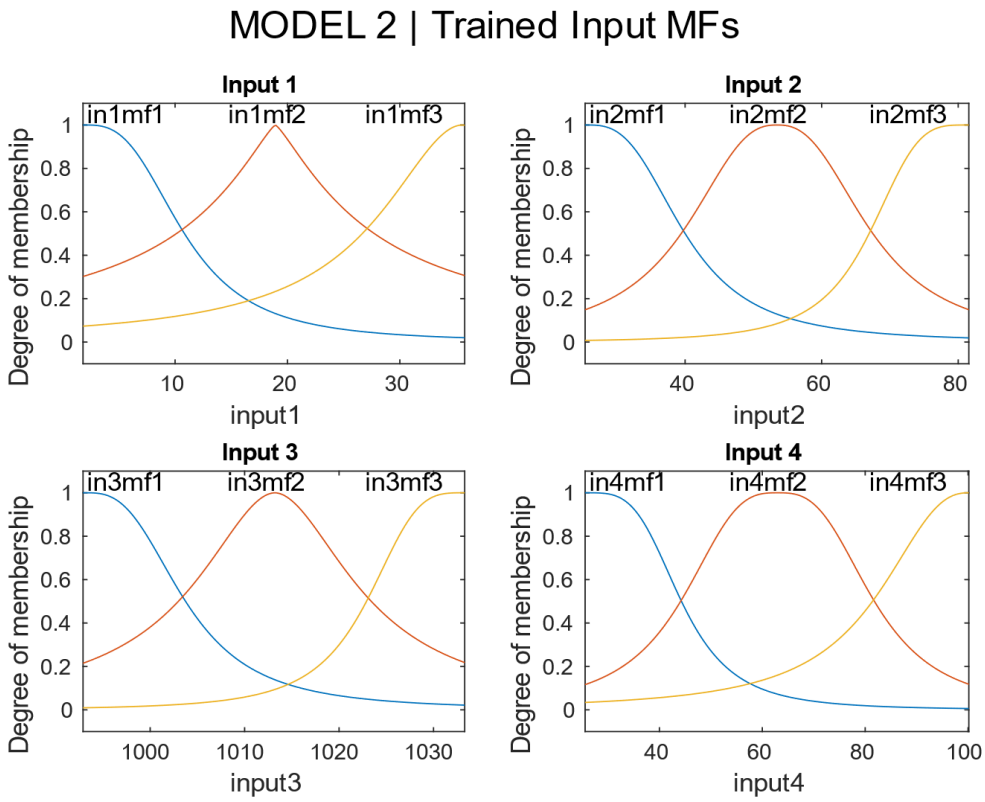
Το εκπαιδευμένο μοντέλο απέδωσε καλά στο regression task στο testing subset με τον δείκτη απόδοσης R^2 να είναι περίπου 94%. Επίσης όπως φαίνεται από τα learning curves δεν υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting – ***epochNumber=100***). Τα παραπάνω επιβεβαιώνονται και στο διάγραμμα των prediction errors που βρίσκονται γύρω από το μηδέν και με σχετικά μικρή για τα μεγέθη της εξόδου διακύμανση.

1.5.2 Model 2: (3) Input MFs, Singleton Output

Αρχικά, παρατίθενται οι αρχικές και τελικές μορφές των MFs για τις ασαφείς τιμές των ασαφών μεταβλητών εισόδου των κανόνων:

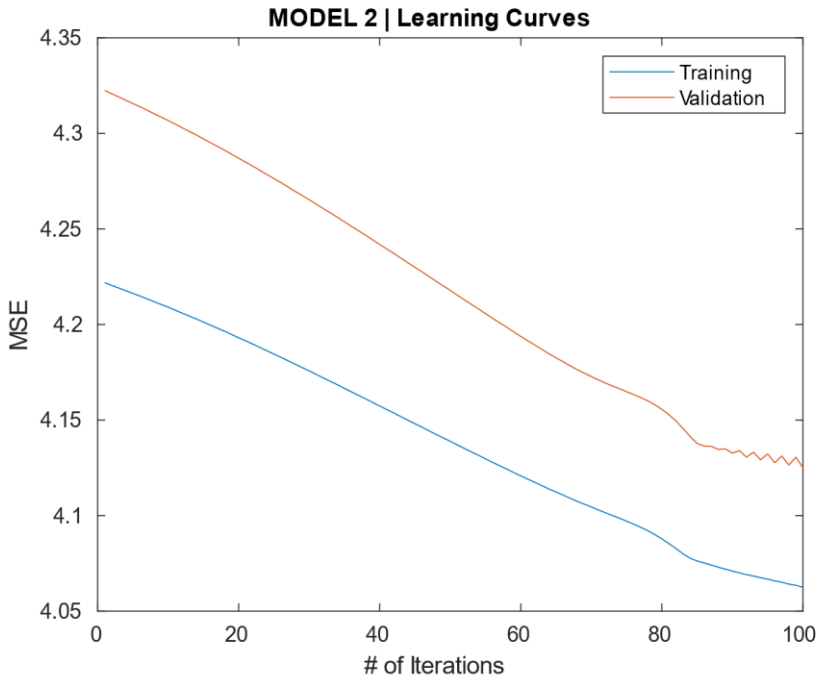


Εικόνα 5: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 2ου μοντέλου

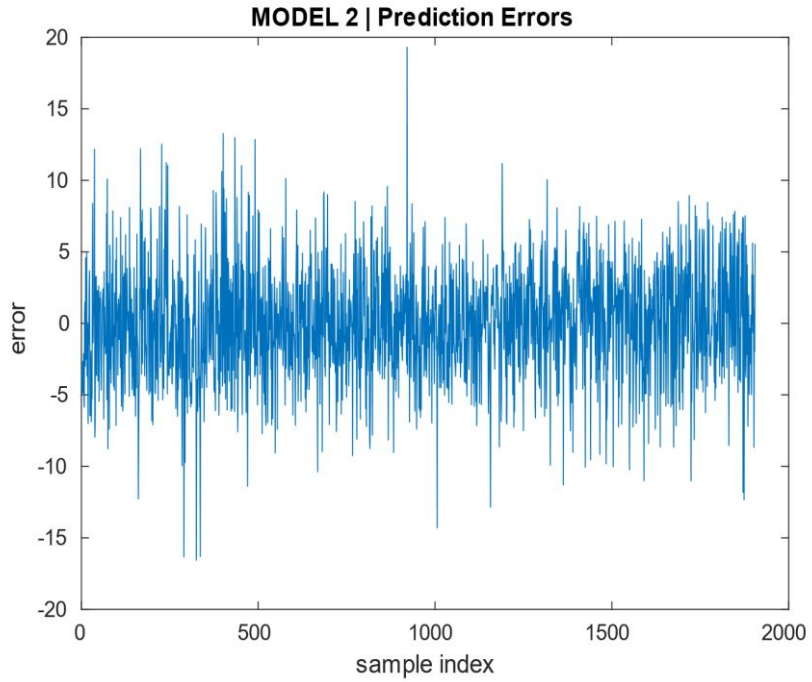


Εικόνα 6: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 2ου μοντέλου

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



Εικόνα 7: Καμπύλες μάθησης 2ου μοντέλου



Εικόνα 8: Σφάλματα πρόβλεψης κατά την εφαρμογή του 2ου μοντέλου στο test set

Τέλος, δίνονται οι ζητούμενες μετρικές απόδοσης του μοντέλου:

Metric →	MSE	RMSE	NMSE	R^2	NDEI
Value →	15.679	3.96	0.058	0.9423 (**94.23%**)	0.24

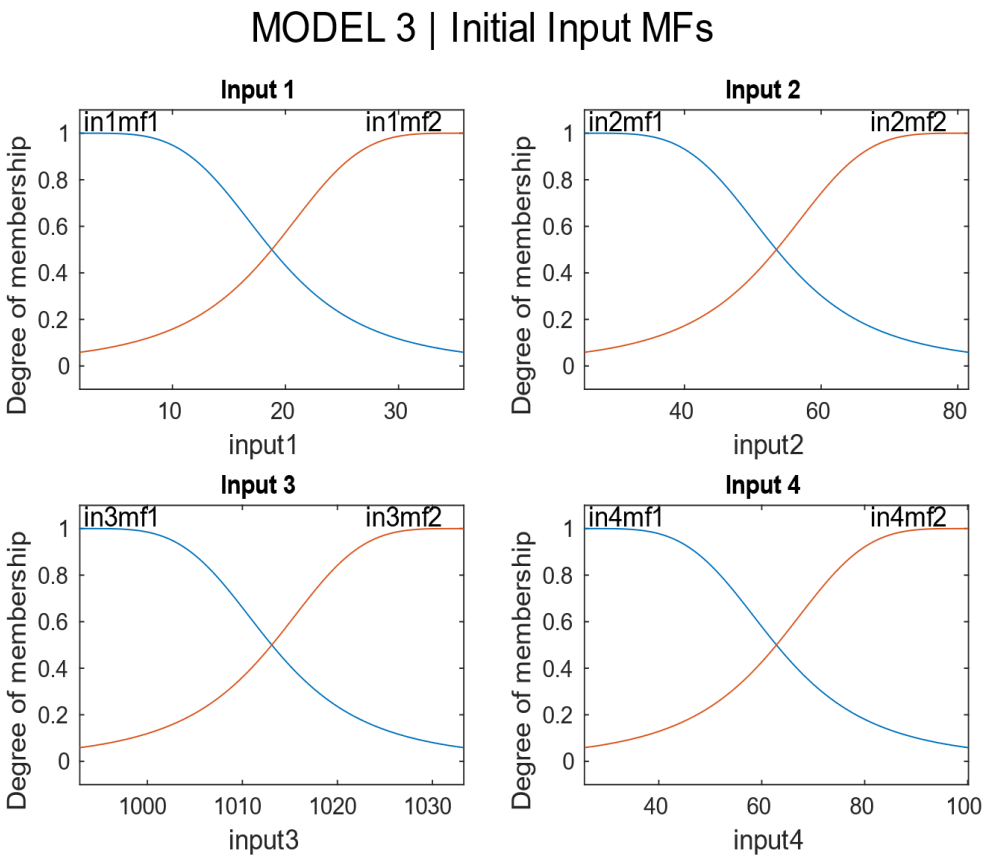
Πίνακας 3: Μετρικές Απόδοσης του 2ου μοντέλου

Μια πρώτη ανάλυση:

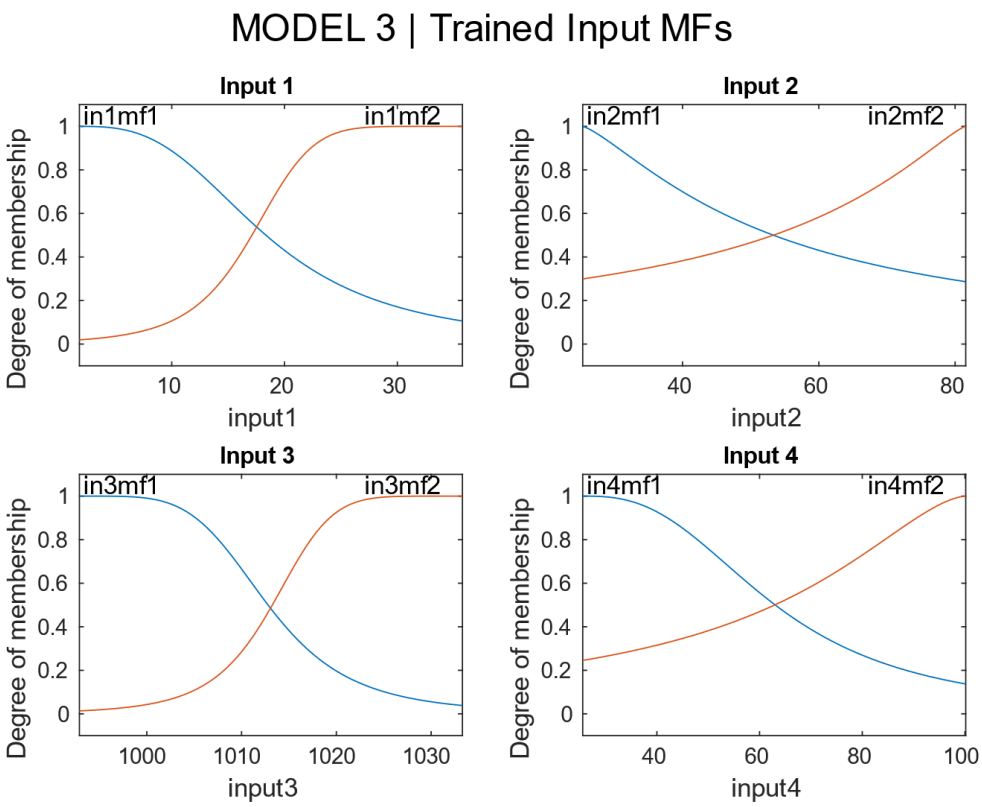
Το δεύτερο εκπαιδευμένο μοντέλο απέδωσε λίγο καλύτερα από το πρώτο στο regression task στο testing subset με τον δείκτη απόδοσης R^2 να είναι λίγο μεγαλύτερος από 94%. Επίσης όπως φαίνεται από τα learning curves δεν υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting – ***epochNumber=100***). Τα παραπάνω επιβεβαιώνονται και στο διάγραμμα των prediction errors που βρίσκονται γύρω από το και με σχετικά μικρή για τα μεγέθη της εξόδου διακύμανση.

1.5.3 Model 3: (2) Input MFs, Polynomial Output

Αρχικά, παρατίθενται οι αρχικές και τελικές μορφές των MFs για τις ασαφείς τιμές των ασαφών μεταβλητών εισόδου των κανόνων:

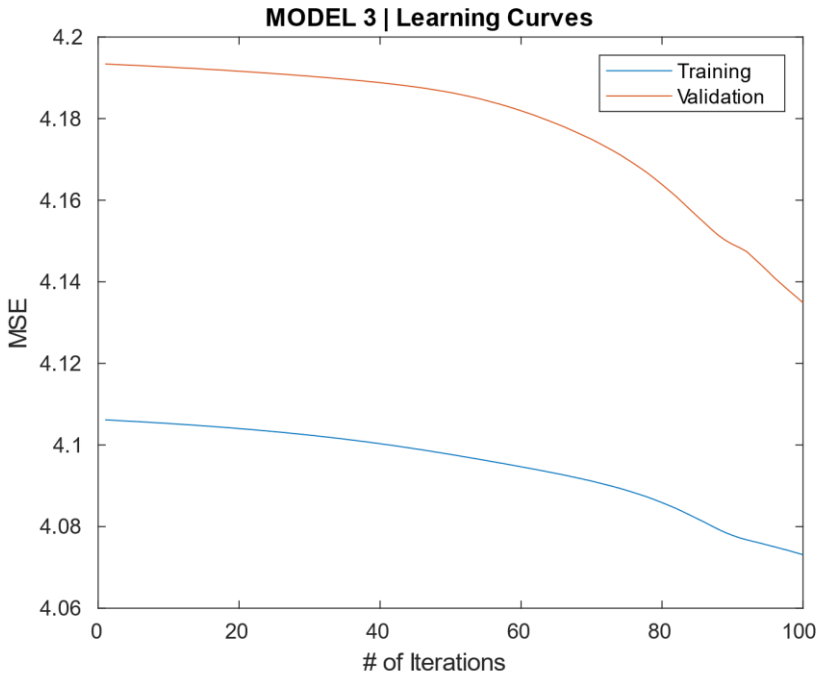


Εικόνα 9: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 3ου μοντέλου

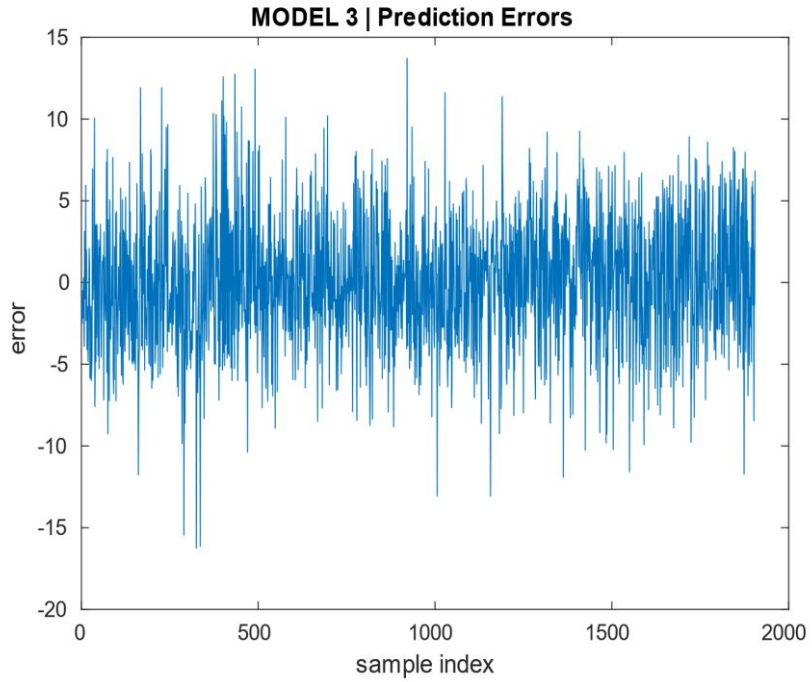


Εικόνα 10: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 3ου μοντέλου

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



Εικόνα 11: Καμπύλες μάθησης 3ου μοντέλου



Εικόνα 12: Σφάλματα πρόβλεψης κατά την εφαρμογή του 3ου μοντέλου στο test set

Τέλος, δίνονται οι ζητούμενες μετρικές απόδοσης του μοντέλου:

Metric →	MSE	RMSE	NMSE	R^2	NDEI
Value →	15.364	3.92	0.057	0.9433 (**94.33%**)	0.238

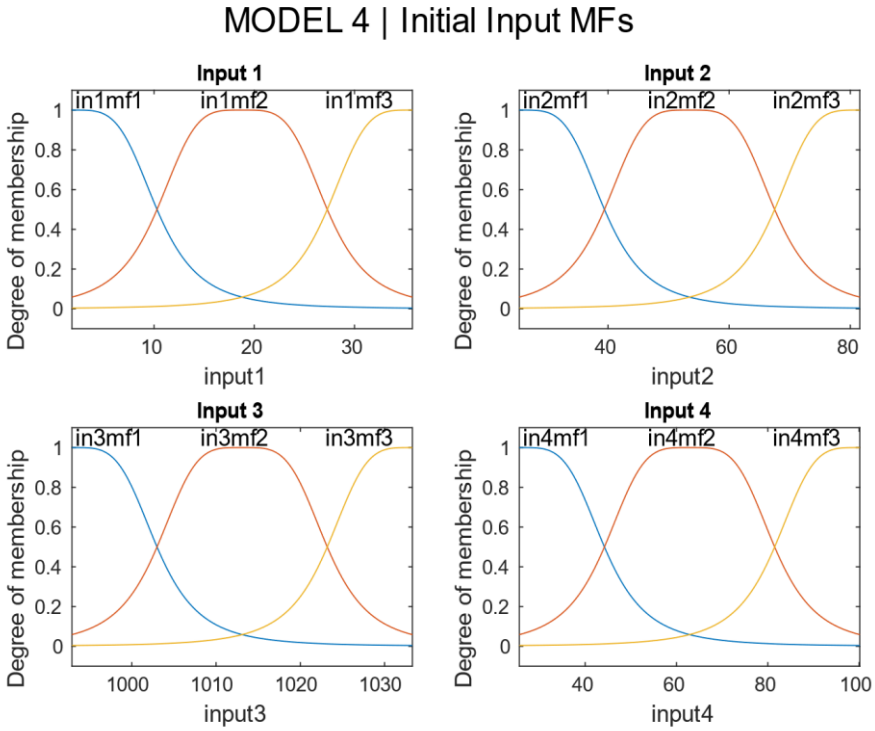
Πίνακας 4: Μετρικές Απόδοσης του 3ου μοντέλου

Μια πρώτη ανάλυση:

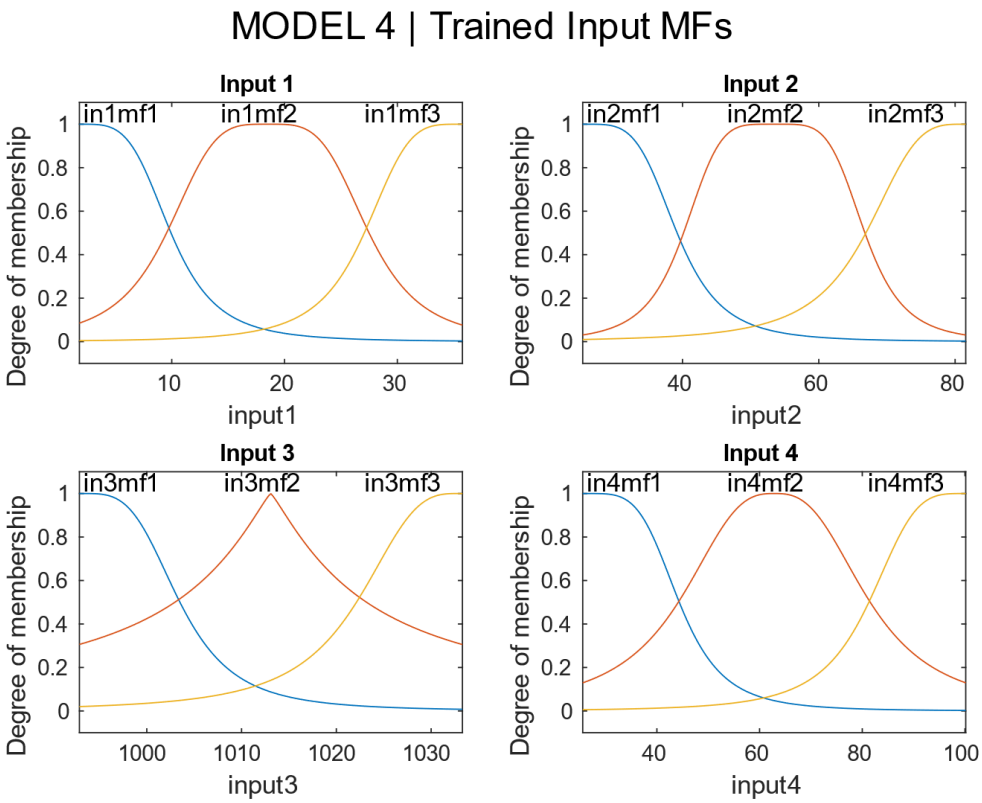
Το τρίτο εκπαιδευμένο μοντέλο απέδωσε λίγο καλύτερα από το πρώτο (έχουν ίδιο αριθμό MFs των μεταβλητών εισόδου) στο regression task στο testing subset με τον δείκτη απόδοσης R^2 να είναι 94.33% (vs. 93.8%). Επίσης όπως φαίνεται από τα learning curves δεν υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting – ***epochNumber=100***). Τα παραπάνω επιβεβαιώνονται και στο διάγραμμα των prediction errors που βρίσκονται γύρω από το και με σχετικά μικρή για τα μεγέθη της εξόδου διακύμανση.

1.5.4 Model 4: (3) Input MFs, Polynomial Output

Αρχικά, παρατίθενται οι αρχικές και τελικές μορφές των MFs για τις ασαφείς τιμές των ασαφών μεταβλητών εισόδου των κανόνων:

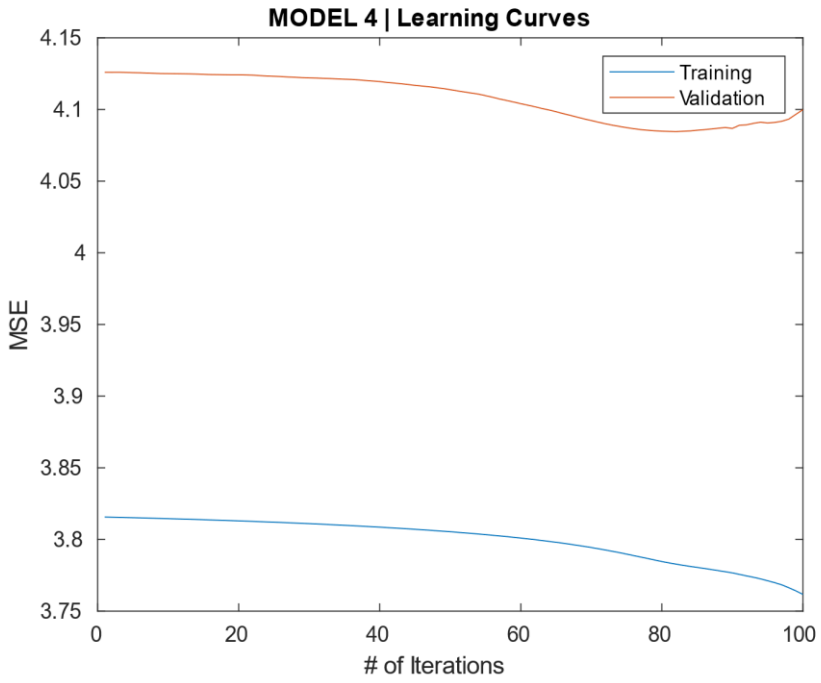


Εικόνα 13: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 4ου μοντέλου

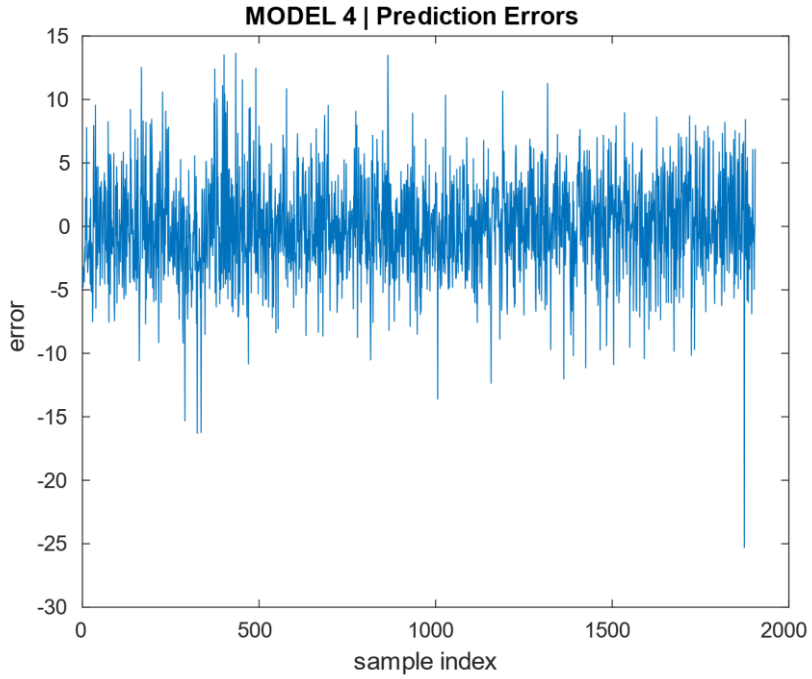


Εικόνα 14: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 4ου μοντέλου

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



Εικόνα 15: Καμπύλες μάθησης 4ου μοντέλου



Εικόνα 16: Σφάλματα πρόβλεψης κατά την εφαρμογή του 4ου μοντέλου στο test set

Τέλος, δίνονται οι ζητούμενες μετρικές απόδοσης του μοντέλου:

Metric →	MSE	RMSE	NMSE	R^2	NDEI
Value →	14.848	3.853	0.054	0.946 (**94.6%**)	0.232

Πίνακας 5: Μετρικές Απόδοσης του 4ου μοντέλου

Μια πρώτη ανάλυση:

Το τέταρτο μοντέλο από αυτά που εκπαιδεύτηκαν απέδωσε λίγο καλύτερα από το δεύτερο (έχουν ίδιο αριθμό MFs των μεταβλητών εισόδου) στο regression task στο testing subset με τον δείκτη απόδοσης R^2 να είναι 94.6% (vs. 94.23%). Όμως, όπως φαίνεται από τα learning curves υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting – ***epochNumber=100***) για το ίδιο epochNumber με τα υπόλοιπα. Να τονισθεί ότι, αν και είναι λίγο παράδοξο, για 80 epochs το παραπάνω μοντέλο ενώ δεν παρουσιάζει overfitting δεν υπάρχει καμία βελτίωση της απόδοσης (μάλιστα υπάρχει και μία μικρή χειροτέρευση – 94.59%). Γι' αυτό δεν παραθέτονται τα στοιχεία του 4ου μοντέλου χωρίς overfitting (epochNumber=80).

1.6. Συμπερασματική ανάλυση

Αρχικά παραθέτονται συγκεντρωτικά οι μετρικές απόδοσης όλων των μοντέλων:

Metric → ↓ Model	MSE	RMSE	NMSE	R^2	NDEI
Model 1	16.627	4.078	0.062	0.938 (**93.80%**)	0.249
Model 2	15.679	3.960	0.058	0.942 (**94.23%**)	0.240
Model 3	15.364	3.920	0.057	0.943 (**94.33%**)	0.238
Model 4	14.848	3.853	0.054	0.946 (**94.6%**)	0.232

Πίνακας 6: Συγκεντρωτικές Μετρικές Απόδοσης όλων των μοντέλων

Βάσει των παραπάνω μετρικών απόδοσης φαίνεται ότι:

- 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 -singleton- έξοδο σε

πολυωνυμική –polynomial-) οδηγεί σε καλύτερα αποτελέσματα (διαφορά μοντέλου 1 από 3, διαφορά μοντέλου 2 από 4)

- Όλα τα μοντέλα, όπως έχει αναφερθεί και στις πρώτες αναλύσεις, παρουσιάζουν αρκετά ικανοποιητική απόδοση παλινδρόμησης (regression) με τον δείκτη απόδοσης R^2 να κυμαίνεται περίπου μεταξύ 94% και 95%. Ωστόσο το **μοντέλο 4** επιλέγεται ως το πιο αποδοτικό με βάση τον παραπάνω πίνακα.
- Αξίζει να σημειωθεί ότι το μοντέλο (2) (3 συναρτήσεις συμμετοχής, singleton έξοδος) είναι πολύ κοντά σε απόδοση από το μοντέλο 3 (2 συναρτήσεις συμμετοχής, polynomial έξοδος) το οποίο εν μέρη είναι αναμενόμενο.
- Τέλος, στα πρώτα τρία μοντέλα δεν παρατηρείται overfitting σε αντίθεση με το τέταρτο, γεγονός που φαίνεται από την απόκλιση της καμπύλης μάθησης στο validation set, η όποια δεν εμφανίζει μονότονα φθίνουσα μορφή (όπως εμφανίζει στο training set). Ο αριθμός των epochs (επαναλήψεων) στην εκπαίδευση του κάθε μοντέλου FNN είναι 100.

2. Εφαρμογή σε high-dimensional dataset

2.1. Κώδικας 2ου μέρους εργασίας

Ο κώδικας σε MATLAB που υλοποιεί το μέρος β' της εργασίας βρίσκεται στο αρχείο */3/matlab/main_b.m*. Εκεί υπάρχει μόνο η λογική της εκτέλεσης, ενώ βοηθητικές κλάσεις και συναρτήσεις υπάρχουν στο φάκελο */Matlab Helpers/**.

2.2. Φόρτωση & Προ-επεξεργασία dataset

Το dataset του ερωτήματος, *superconduct*, αποτελείται από 21263 δείγματα (data points) με 81 features και μία τιμή εξόδου το καθένα. Σε ότι αφορά τη προ-επεξεργασία του dataset, αρχικά, γίνεται έλεγχος για διπλότυπα δείγματα (~66 δείγματα αφαιρέθηκαν). Κατόπιν κανονικοποιούμε το dataset (πλην της τελευταίας στήλης) ως προς το εύρος (range), κλιμακοποιούμε (scale) δηλαδή τα σημεία στο εύρος [0,1]. Κατόπιν, κάνουμε ένα μικρό smoothing (*SmoothingFactor*=0.05). Σκόπιμο είναι να αναφερθεί ότι οι παραπάνω μέθοδοι – παράμετροι προ-επεξεργασίας του dataset έχουν προέλθει με τη μέθοδο trial-and-error. Για παράδειγμα, αν δεν γίνει το smoothing, οι παραγόμενες ακτίνες SC (με τη μέθοδο που περιγράφεται στη παρ. 2.4.1 παρακάτω) είναι απαγορευτικά μικρές (καθώς οδηγούν σε NaNs ή runtime errors) και έτσι δεν θα ήταν δυνατό να δοκιμαστούν όλοι οι δυνατοί συνδυασμοί στο grid. Τέλος, το splitting του dataset σε training, validation & testing subsets γίνεται με βάση τον αλγόριθμο που παρουσιάστηκε στο 1ο μέρος του παρόντος (βλ. παράγραφος 1.3).

2.3. Μείωση Διαστασιμότητας Dataset

Επειδή το δοσμένο dataset έχει high-dimensionality, ο αριθμός των απαιτούμενων κανόνων με βάση το grid partitioning στο χώρο των εισόδων – features του μοντέλου θα είναι τεράστιος. Για το λόγο αυτό, όπως αναφέρεται και στην εκφώνηση του 2ου μέρους της παρούσας εργασίας, θα χρησιμοποιήσουμε τεχνικές μείωσης της διαστασιμότητας και τεχνικές ομαδοποίησης για το διαχωρισμό του χώρου των εισόδων με σκοπό τη 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 (scatter partitioning) πριν την δημιουργία των κανόνων

2.4. Grid Search: Εύρεση Βέλτιστου Συνδυασμού Αριθμού Features – Αριθμού Κανόνων

Το σετ των πιθανών χαρακτηριστικών που θα κρατήσουμε κατά το feature subset selection με τον ReliefF (K=100) είναι NF={3,9,15,21} ενώ το σετ των πιθανών αριθμών κανόνων (clusters εισόδων) του μοντέλου θα είναι NR={4,8,12,16,20}. Για κάθε ένα συνδυασμό αριθμού χαρακτηριστικών – αριθμού κανόνων στο grid, θα αναζητήσουμε το σφάλμα εκπαίδευσης στο validation set (validation error) και μέσω αυτού θα καταλήξουμε στο βέλτιστο συνδυασμό ή βέλτιστο σημείο στο grid (optimum grid point).

Για την ακριβέστερη εύρεση του optimum grid point θα εφαρμόσουμε 5-fold cross-validation κατά την εκπαίδευση. Έτσι, για κάθε grid point, το validation error θα προκύψει ως ο μέσος όρος των (τελικών) validation errors για κάθε ένα από τα 5 folds.

2.4.1 Ακτίνες αναζήτησης για Subtractive Clustering (SC)

Ο αλγόριθμος Subtractive Clustering (SC), σε αντίθεση με των Fuzzy C-Means, 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.

Για το σκοπό αυτό, έχει αναπτυχθεί η συνάρτηση
$SubtractiveClusteringWrapper::bisectionForNfNr(nf, nr, rad_{LEFT}, rad_{RIGHT}, nr_{LEFT}, nr_{RIGHT})$
σε MatLab. Η λογική υλοποίησης της εύρεσης της ακτίνας είναι μία bisection-like λογική, όπου δίνουμε αρχικά ένα εύρος (κανονικοποιημένων) ακτινών, rad_{LEFT}=0.1 και rad_{RIGHT}=1. Στη παραπάνω μέθοδο υπολογίζονται τα clusters που αντιστοιχούν στις εκάστοτε ακτίνες-ορίσματα καθώς και στο μέσο αυτών και αποφασίζεται σε ποια μεριά θα συνεχίσει η αναδρομή. Η αναδρομή σταματάει είτε όταν βρεθεί αριθμός clusters ίσος με το ζητούμενο nr είτε μετά από συγκεκριμένο αριθμό επαναλήψεων με σχετικό μήνυμα σφάλματος. Η παραπάνω διαδικασία είναι μια εξαιρετικά χρονοβόρα διαδικασία καθώς για κάθε έναν από τους 20 συνδυασμούς (grid points) και για κάθε ένα από τα cross-validation runs θα γίνουν τουλάχιστο δύο δοκιμές για την εύρεση της ακτίνας του SC με μέσο αριθμό δοκιμών, 5.4 δοκιμές/grid point/fold.

Στο σημείο αυτό αξίζει να σημειώσουμε τις επιταχύνσεις που έχουν υλοποιηθεί για να κατεβάσουμε το χρόνο εκτέλεσης σε ρεαλιστικά επίπεδα:

- Στη παραπάνω μέθοδο χρησιμοποιείται ένα map, *"NFMap"*, το οποίο αρχικοποιείται σε κάθε fold και υπάρχει ένα για κάθε έναν από τους ζητούμενους αριθμούς χαρακτηριστικών. Στο map απομνημονεύονται οι υπολογισμένες ακτίνες για το συγκεκριμένο αριθμό χαρακτηριστικών και για το fold στο οποίο είχε αρχικοποιηθεί το map, ώστε να μην ξανα-υπολογιστούν.
- Έχουν τροποποιηθεί οι συναρτήσεις *genfis()* και *genfis2()* έτσι ώστε η διαδικασία να σταματάει όταν η *subclust()* επιστρέψει τα κέντρα των clusters και άρα έχουμε τον αριθμό αυτών. Η χρήση των τροποποιημένων συναρτήσεων (βρίσκονται στο φάκελο *Matlab Helpers/Builtins*) οδηγεί αφενός σε πολύ σημαντική επιτάχυνση και αφετέρου επιτρέπει τον υπολογισμό για πολύ μικρότερες ακτίνες, όπως για ακτίνες radii<<0.1, κάτι το οποίο δεν ήταν δυνατό προηγούμενα (σε αυτό όμως βοηθάει και η αντικατάσταση της συνάρτησης για την επίλυση του γραμμικού συστήματος του FIS matrix, από *A\b* σε *pinv(A)*b* – βλ. */Matlab Helpers/Builtins/BGenfis2.m*).

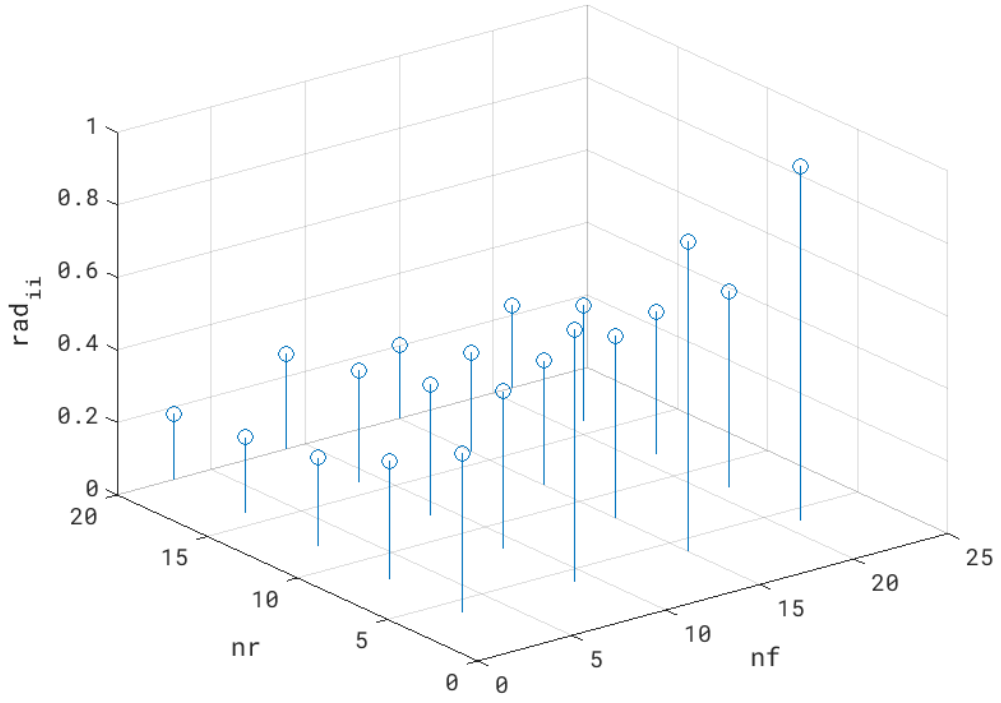
Η διαδικασία αυτή έχει ολοκληρωθεί, όσες φορές χρειάστηκε, ενώ πάντα έβρισκε την ακτίνα που υλοποιεί το εκάστοτε grid point (για το εκάστοτε training set του cross validation) που δίνονταν ως παράμετρος.

Ως παράδειγμα υπολογισμού των ακτινών του SC, παρακάτω δίνονται οι μέσες τιμές των ακτινών για κάθε ένα από τα 20 grid points (μέσοι όροι ως προς τα 5 folds του cross-validation):

NR → ↓ NF	**4**	**8**	**12**	**16**	**20**
3	0.4375000	0.3250000	0.2434375	0.2081137	0.1820898
9	0.6962500	0.4336768	0.3601563	0.3076031	0.2610156
15	0.8537500	0.5014844	0.3418750	0.2731445	0.2020410
21	0.9775000	0.5415625	0.3917969	0.3203418	0.2270898

Πίνακας 7: Μέση τιμή ακτινών που αντιστοιχούν στους συνδυασμούς αριθμών κανόνων και αριθμών χαρακτηριστικών (grid points)

Ενώ, ακολούθως, φαίνονται τα παραπάνω σε μορφή 3D stem διαγράμματος:



Εικόνα 17: (Μέση) Ακτίνα SC ως προς αριθμό κανόνων και αριθμό χαρακτηριστικών (grid point)

Από το παραπάνω διάγραμμα συμπεραίνουμε ότι αύξηση του αριθμού των κανόνων ή / και μείωση του αριθμού των features οδηγεί σε μείωση της ακτίνας του SC, κάτι το οποίο είναι απολύτως λογικό.

2.4.2 Cross Validation

Ακολούθως, θα ξεκινήσουμε την αναζήτηση πλέγματος (grid search) υλοποιώντας 5-fold cross validation στο dataset. Έτσι, για κάθε ένα από τα πέντε folds του dataset και για κάθε ένα από τους 20 συνδυασμούς, grid points, θα εκπαιδεύσουμε ένα TSK μοντέλο χρησιμοποιώντας την υβριδική μέθοδο και συγκεκριμένα τις συναρτήσεις *genfis()* και *anfis()* του MatLab, όπου ως training και validation set σε κάθε fold θα δίνονται **subsets του αρχικού training set** που προκύπτουν με τη βοήθεια της συνάρτησης *cvpartition()* του MatLab.

2.4.3 Αποτελέσματα – Optimum Grid Point

Αρχικά δίνουμε ένα index σε κάθε grid point από το 1 έως το 20, σκανιάροντας τα grid points κατά τα NF, δηλ. 1 ↦ (NF(1),NR(1)), 2 ↦ (NF(1),NR(2)), κ.ο.κ . Για κάθε ένα από τα cross-validation runs κάνουμε τα εξής:

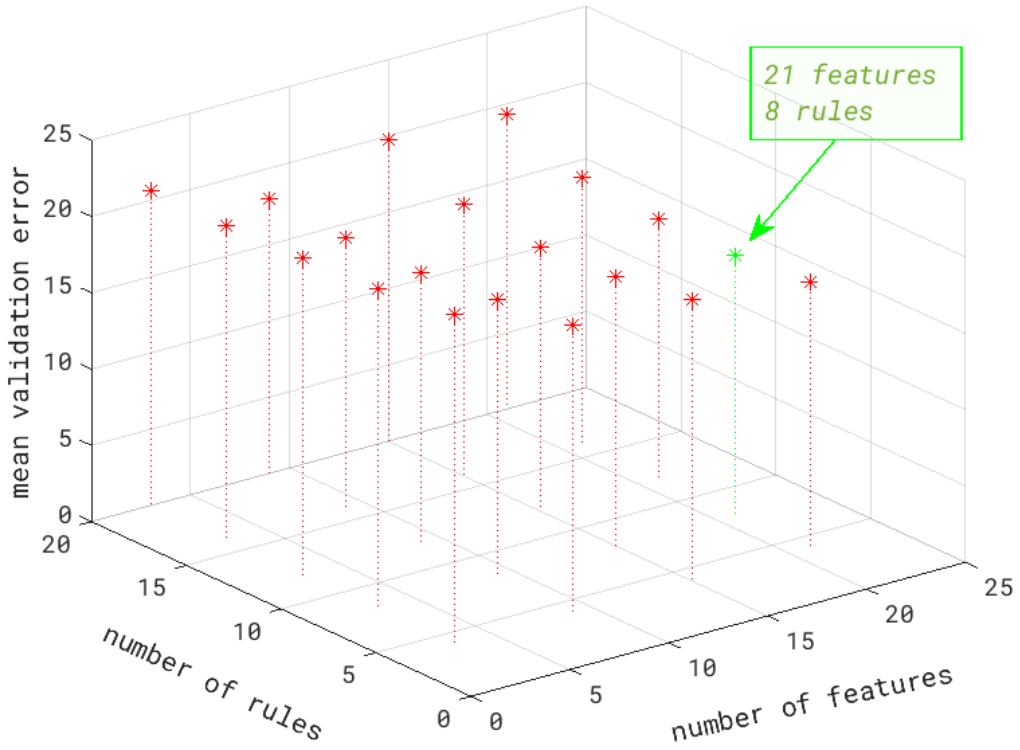
1. Εξάγουμε από το αρχικό training set, το training set και το test set που θα χρησιμοποιηθούν σε αυτό το fold, με βάση τα indices της *cvpartition()*
2. Για κάθε grid point, κάνουμε τα εξής:
    - Βρίσκουμε την ακτίνα ομαδοποίησης για το SC, για το συγκεκριμένο αριθμό κανόνων-clusters, για το συγκεκριμένο αριθμό χαρακτηριστικών και για το συγκεκριμένο training set του fold
    - Εκπαιδεύουμε με βάση το training set του fold ένα TSK μοντέλο με SC στο χώρο των εισόδων με παράμετρο την παραπάνω ακτίνα αναζήτησης
    - Λαμβάνουμε το τελικό (ως προς τα epochs) validation error του μοντέλου στο test set του fold

Έχοντας τα τελικά validation errors των grid points για κάθε cross-validation run, υπολογίζουμε το μέσο όρο αυτών και κατόπιν επιλέγουμε σαν βέλτιστο το grid point του οποίου ο μέσος όρος των (τελικών) validation errors των folds είναι ελάχιστος. Παρακάτω, παρατίθεται ο μέσος όρος των τελικών validation errors για κάθε grid point:

NR → ↓ NF	**4**	**8**	**12**	**16**	**20**
3	21.7173	21.1085	20.8277	20.6719	20.5923
9	18.8437	18.2756	17.7700	17.7407	17.9857
15	18.4378	17.6606	17.2821	17.7819	19.7807
21	17.5132	16.9533	17.0038	17.4524	19.2980

Πίνακας 8: Μέσο (τελικό) validation error για κάθε grid point ως προς τα 5 folds

ενώ, παρακάτω δίνουμε ένα 3D plot των μέσων όρων των τελικών validation errors ως προς τα runs για κάθε grid point, σημειώνοντας με ανοιχτό πράσινο χρώμα το σημείο (grid point) που οδηγεί στο ελάχιστο mean validation error:



Εικόνα 18: 3D plot του μέσου (τελικού) validation error για κάθε grid point ως προς τα 5 folds

Όπως φαίνεται από τον Πίνακα 8 καθώς και την Εικόνα 18 παραπάνω, ο βέλτιστος συνδυασμός αριθμού χαρακτηριστικών – αριθμού κανόνων (grid point) είναι το (21,8) και άρα:

$$\mathbf{nf_{opt} = 21\ features}$$
$$\mathbf{nr_{opt} = 8\ rules}\ \text{(clusters)}$$

Επίσης, από τα παραπάνω, φαίνεται ότι μείωση του αριθμού των χρησιμοποιούμενων χαρακτηριστικών (features) οδηγεί σε μεγαλύτερα μέσα validation errors, κάτι απολύτως αναμενόμενο εάν λάβουμε υπόψη την υψηλή διαστασιμότητα του dataset. Είναι επομένως λογικό ότι ο βέλτιστος συνδυασμός θα περιέχει το μεγαλύτερο από τους πιθανούς αριθμούς χαρακτηριστικών, δηλ. 21, κάτι που συμβαίνει. Ο αριθμός των clusters για δεδομένο αριθμό χαρακτηριστικών δεν φαίνεται να παίζει καίριο ρόλο στη διαμόρφωση του μέσου validation error. Έτσι, για 21 features ο αριθμός κανόνων ή clusters που δίνει το ελάχιστο μέσο validation error είναι 8, ενώ πολύ κοντά στο σφάλμα αυτού του συνδυασμού είναι και ο συνδυασμός (21,12).

2.5. Τελικό Μοντέλο με βάση το Optimum Grid Point

Έχοντας καταλήξει στο βέλτιστο συνδυασμό αριθμού χαρακτηριστικών – αριθμού κανόνων (grid point), προχωράμε στην επιλογή των 21 χαρακτηριστικών με βάση τον ReliefF και κατόπιν στην εκπαίδευση του τελικού μοντέλου στο οποίο θα θέλαμε να υπάρχουν 8 clusters στον Subtractive Clustering (SC). Από τα cross-validation runs έχουμε 5 ακτίνες για τις οποίες, στο εκάστοτε fold, ο SC φτιάχνει 12 clusters. Οι ακτίνες αυτές δίνονται ακολούθως:

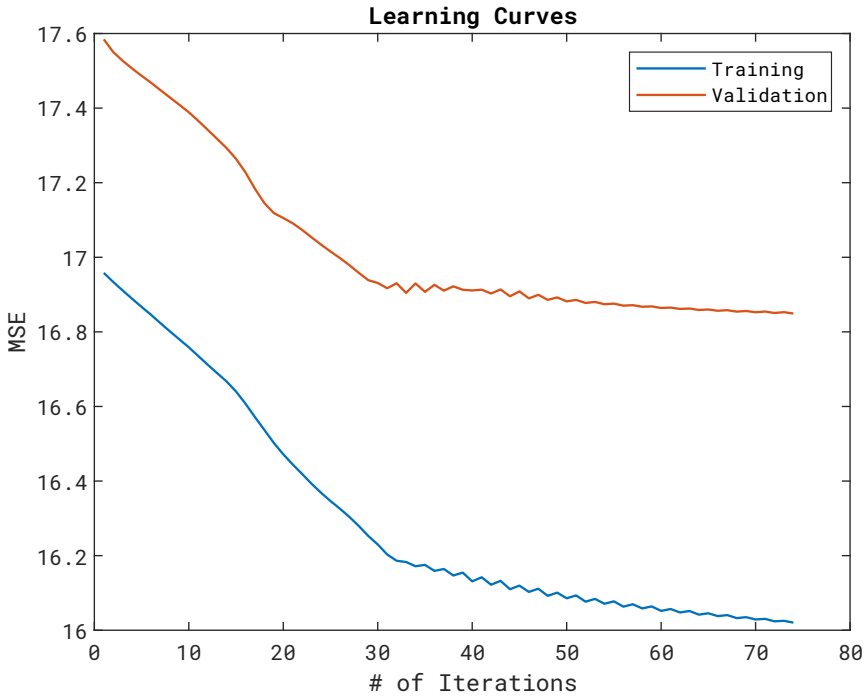
Fold →	1	2	3	4	5
Ακτίνα →	0.49375	0.55000	0.55000	0.53594	0.57812

Πίνακας 9: Ακτίνες SC του optimal grid point σε κάθε cross-validation run

Παίρνοντας τον μέσο όρο με trimming (συνάρτηση *trimmean()* στο MatLab) των ακτινών αυτών οδηγούμαστε στην (κανονικοποιημένη) ακτίνα **0.5453**. Πράγματι με αυτή την ακτίνα προκύπτει TSK με 8 κανόνες και άρα δεν χρειάζεται η εκ νέου αναζήτηση με βάση τη μέθοδο που περιεγράφηκε στην παράγραφο 2.4.1 παραπάνω (σε ολόκληρο το training set πλέον).

2.5.1 Απόδοση Τελικού Μοντέλου

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



Εικόνα 19: Καμπύλες μάθησης τελικού TSK μοντέλου με overfitting (75 epochs)

όπου όπως φαίνεται, για τον χρησιμοποιούμενο αριθμό epochs (75) δεν υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting).

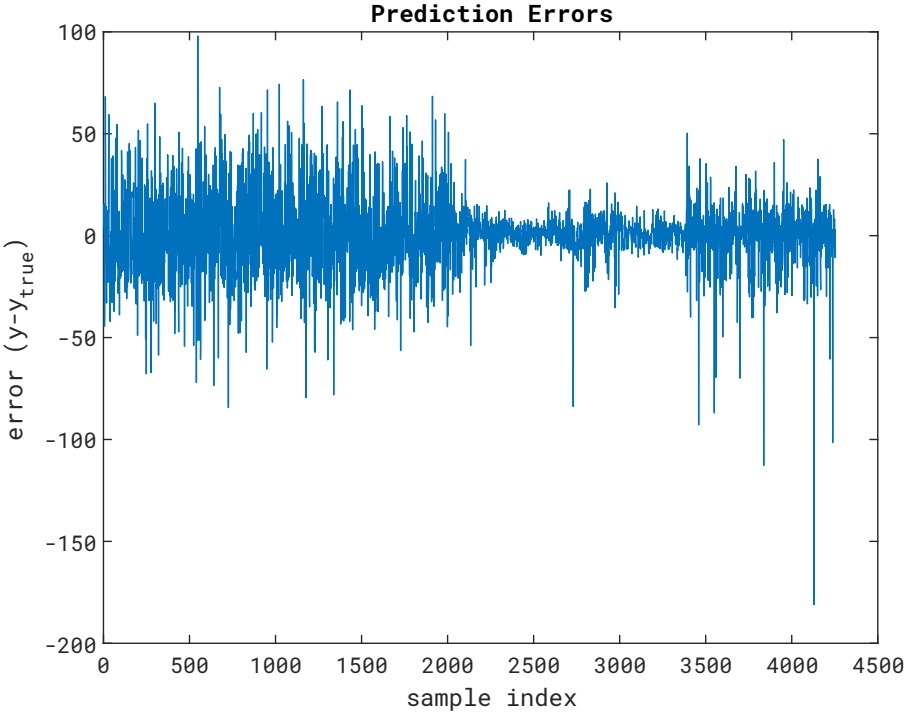
Ακολουθούν, οι ζητούμενες μετρικές απόδοσης του τελικού μοντέλου:

Metric →	MSE	RMSE	NMSE	R^2	NDEI
Value →	290.4961	17.0439	0.3117	0.6883 (**68.83%**)	0.5583

Πίνακας 10: Μετρικές Απόδοσης του τελικού TSK μοντέλου

Όπως φαίνεται το μοντέλο με τον βέλτιστο συνδυασμό αποδίδει με δείκτη R^2 λίγο κάτω από 69%, αποτέλεσμα σχετικά ικανοποιητικό. Κάτι που ίσως οδηγούσε σε μικρή αύξηση του παραπάνω ίσως να ήταν κάποιο tweaking στο SmoothingFactor στη φάση της προ-επεξεργασίας του dataset.

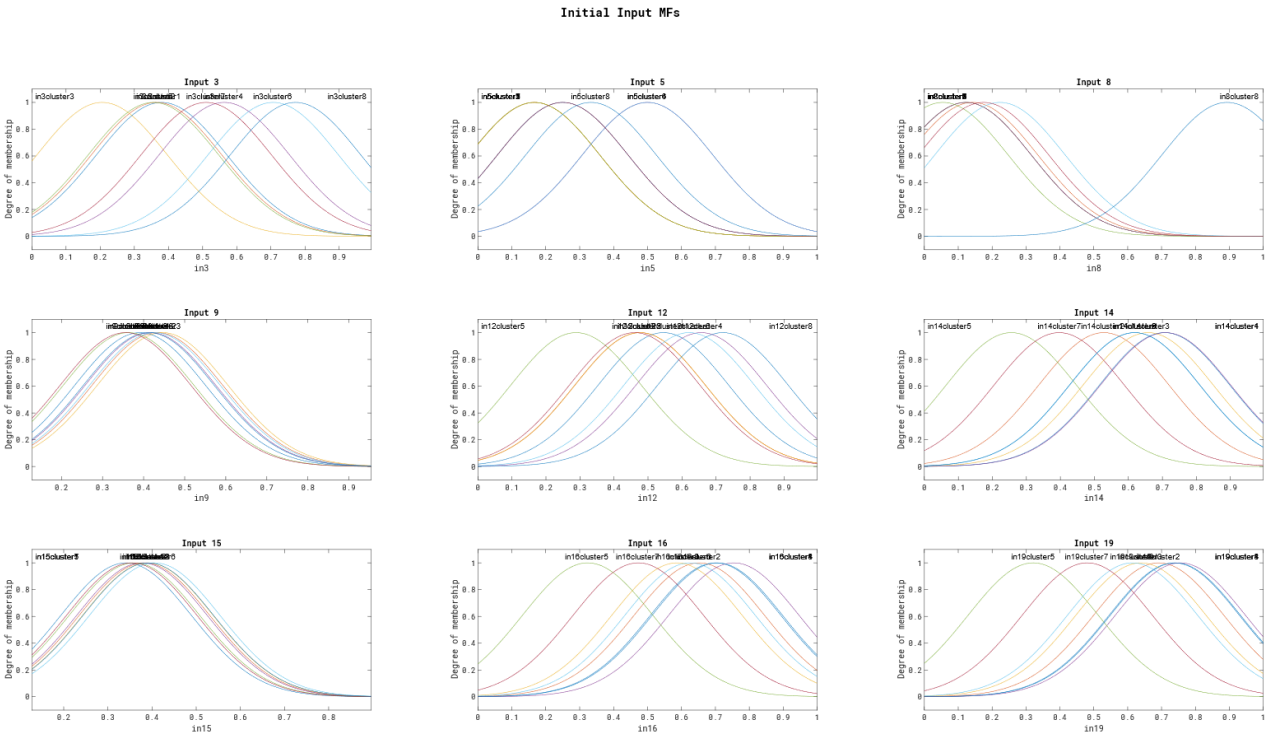
Τέλος, σε ότι αφορά την απόδοση του τελικού μοντέλου, δίνεται το διάγραμμα με τα σφάλματα πρόβλεψης του μοντέλου ως προς τις πραγματικές τιμές του test set:



Εικόνα 20: Σφάλματα πρόβλεψης κατά την εφαρμογή του τελικού TSK μοντέλου στο test set

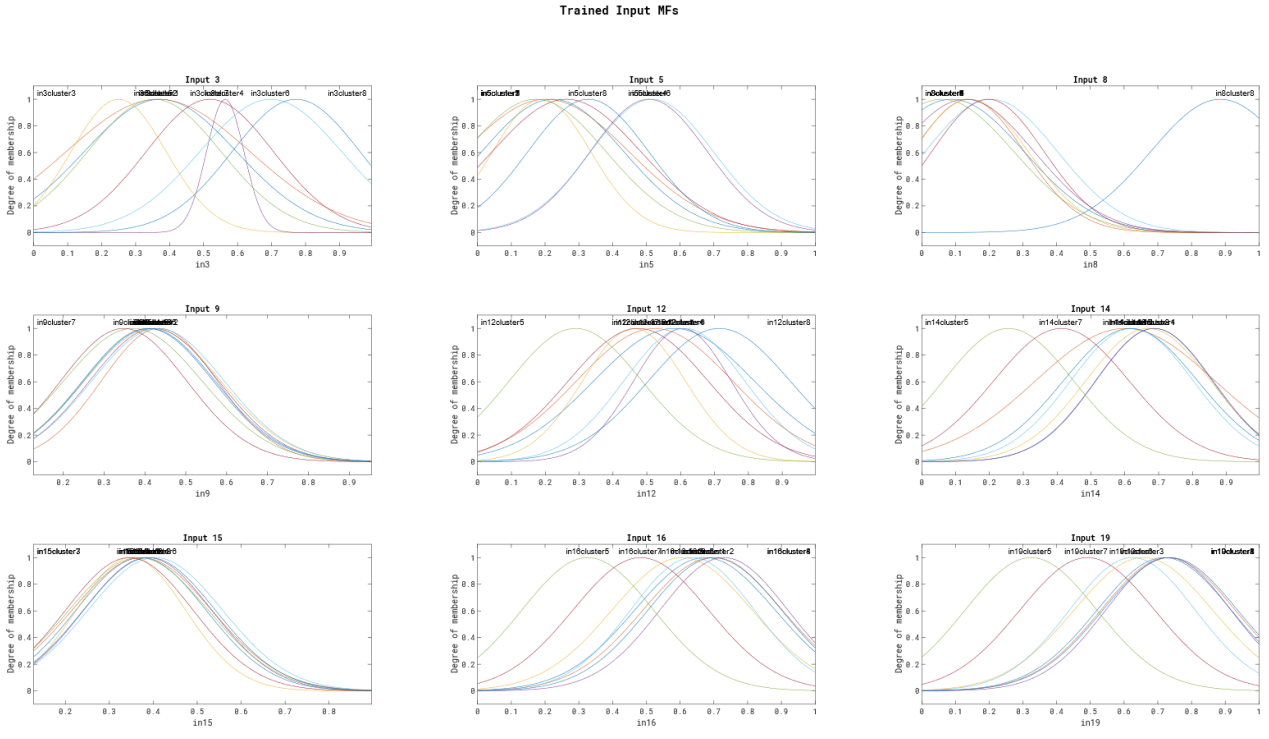
2.5.2 Συναρτήσεις Συμμετοχής (MFs) Τελικού Μοντέλου

Παρακάτω, δίνονται οι αρχικές συναρτήσεις συμμετοχής (MFs) των 8 λεκτικών τιμών - clusters για τις 21 εισόδους - features του τελικού TSK μοντέλου (έχουν επιλεγεί τυχαία 9 είσοδοι για απεικόνιση):



Εικόνα 21: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου τελικού μοντέλου

ενώ, ακολούθως δίνονται οι τελικές μορφές των παραπάνω MFs:



Εικόνα 22: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου τελικού μοντέλου

2.6. Συμπερασματικές Παρατηρήσεις – Σχόλια

Παρατηρούμε ότι για το δοσμένο dataset η ανάπτυξη ενός FNN κρατώντας μόνο τα 20 (σημαντικότερα) από τα συνολικά 81 χαρακτηριστικά οδηγεί σε ικανοποιητικά αποτελέσματα πρόβλεψης (regression). Αυτό φαίνεται στους δείκτες απόδοσης που δίνονται στον Πίνακα 10 παραπάνω και ειδικά στους δείκτες RMSE και R^2. Προφανώς, οι δείκτες απόδοσης αυτοί είναι αρκετά χειρότεροι σε σύγκριση με τους αντίστοιχους του μη-υψηλής διαστασιμότητας dataset του $1^{ου}$ μέρους, κάτι απολύτως αναμενόμενο αν λάβουμε υπόψη το feature subset selection που έλαβε μέρος στο dataset αυτού του μέρους της εργασίας.

Επίσης η χρήση μόνο 8 ασαφών κανόνων σε σύγκριση με τους 2^{21} (για 2 λεκτικές τιμές) ή 3^{21} (για 3) που θα απαιτούνταν για grid partitioning, φανερώνει το πλεονέκτημα της χρήσης ομαδοποίησης στο χώρο των εισόδων (scatter partitioning), με το μειονέκτημα της χειροτέρευσης της απόδοσης του μοντέλου.

Τέλος, για λόγους πληρότητας, παραθέτουμε τα 21 πιο σημαντικά features που επιλέχθηκαν με τον ReliefF για την εκπαίδευση του τελικού μοντέλου:

Indices των 21 πιο σημαντικών features																				
1	63	66	71	78	69	31	58	12	81	28	37	16	7	14	26	67	36	76	56	27

Πίνακας 11: Indices των 21 πιο σημαντικών features από ReliefF (K=100)

Θεσσαλονίκη, 28/2/2020