ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

Τμημα ΗΜΜΥ. Τομεας Ηλεκτρονικης

Ασαφη Συστηματα (Υπολογιστικη Νοημοσυνη)

Εργασία 3

Επίλυση προβλήματος παλινδρόμησης με χρήση μοντέλων TSK

Συντάκτης:
Χρήστος Χουτουρίδης [8997]
cchoutou@ece.auth.gr

Διδάσκοντες:
Θεοχάρης Ιωάννης
theochar@ece.auth.gr

Χαδουλός Χρήστος
christgc@auth.gr

26 Οκτωβρίου 2025

1. Εισαγωγη

Στην παρούσα εργασία μελετούμε τη μοντελοποίηση παλινδρόμησης με ασαφή νευρωνικά μοντέλα τύπου Takagi–Sugeno–Kang (TSK). Τα TSK συνδυάζουν ασαφή διαμέριση του χώρου εισόδων (μέσω συναρτήσεων συμμετοχής και κανόνων IF–THEN) με παραμετρικά (συνήθως γραμμικά) συμπεράσματα στην έξοδο. Η πρακτική σημασία τους έγκειται στην ικανότητα (α) να προσεγγίζουν μη γραμμικές σχέσεις με μικρό αριθμό κανόνων, (β) να εκπαιδεύονται αποδοτικά με *υβριδικές* μεθόδους (οπισθοδιάδοση για τις MF και Ελάχιστα Τετράγωνα για τις παραμέτρους εξόδου), και (γ) να παρέχουν διαγνωστικές απεικονίσεις (MFs, καμπύλες μάθησης, σφάλματα) χρήσιμες για ερμηνεία και ρύθμιση.

Θεωρούμε δύο σενάρια:

1. Το *Airfoil Self-Noise* ως εισαγωγικό παράδειγμα με 5 εισόδους, και
2. Το *Superconductivity* με υψηλή διαστασιμότητα, όπου εφαρμόζουμε επιλογή χαρακτηριστικών και αναζήτηση πλέγματος (grid search) πάνω σε 5-fold cross validation.

1.1. Παραδοτέα

Τα παραδοτέα της εργασίας αποτελούνται από:

- Την παρούσα αναφορά.
- Τον κατάλογο **source**, με τον κώδικα της Matlab.
- Το σύνδεσμο με το αποθετήριο που περιέχει τον κώδικα της Matlab καθώς και αυτόν της αναφοράς.

2. Υλοποιηση

Υλοποιήθηκε κοινή υποδομή συναρτήσεων για **διαχωρισμό συνόλων**, **προ-επεξεργασία**, **αξιολόγηση** και **συστηματική απεικόνιση αποτελεσμάτων**, ώστε να εξυπηρετούνται και τα δύο σενάρια. Η αναπαραγωγή γίνεται εκτελώντας τα scripts `scenario1.m` και `scenario2.m`, τα οποία καλούν τις βοηθητικές συναρτήσεις.

2.1. Διαχωρισμός δεδομένων — `split_data()`

Η `split_data` δέχεται το πλήρες dataset και διαχωρίζει σε **train**/**validation**/**test** με ποσοστά (εδώ 60/20/20), επιστρέφοντας *ξεχωριστά* τα X και y για κάθε υποσύνολο. Χρησιμοποιούμε `seed` για *αναπαραγωγιμότητα* (`rng(seed,'twister')`), ώστε τυχόν συγκρίσεις μεταξύ μοντέλων να γίνονται επί των ίδιων χωρισμάτων.

2.2. Προεπεξεργασία δεδομένων — `preprocess_data()`

Η προεπεξεργασία είναι απαραίτητη για σταθερή εκπαίδευση. Υλοποιήθηκαν δύο τρόποι: **Min–Max** κλιμάκωση σε $[0, 1]$ και **z-score** (μέση τιμή/τυπική απόκλιση). Σε όλα τα σενάρια *υπολογίζονται τα στατιστικά **μόνο** στο train* και εφαρμόζονται σε val/test (*no leakage*). Πειραματιστήκαμε με z-score, ωστόσο καταλήξαμε σε Min–Max (mode=1), που αποδείχθηκε επαρκής και σταθερή επιλογή.

2.3. Αξιολόγηση — `evaluate()`

Η `evaluate` υπολογίζει **MSE**, **RMSE**, R^2, **NMSE** και **NDEI** βάσει των προβλέψεων και της αλήθειας (`pred, truth`). Ο δείκτης $R^2 = 1 - \frac{\mathrm{SS_{res}}}{\mathrm{SS_{tot}}}$ και τα κανονικοποιημένα μεγέθη (NMSE, NDEI= $\sqrt{\mathrm{NMSE}}$) επιτρέπουν συγκρίσεις μεταξύ ρυθμίσεων.

2.4. Εμφάνιση αποτελεσμάτων

Υλοποιήθηκαν οι `plot_results1` (Σενάριο 1) και `plot_results2` (Σενάριο 2) που:

- δημιουργούν **συνοπτικό** σχήμα με *όλες* τις MF ανά είσοδο (πριν/μετά),
- σχεδιάζουν **καμπύλες μάθησης** (train/val),
- δίνουν **Predicted vs Actual** και **σφάλμα πρόβλεψης (residuals)**,
- και αποθηκεύουν συστηματικά αρχεία εικόνων στους φακέλους `figures_scn1` / `figures_scn2`.

3. Σενάριο 1 — Εφαρμογή σε απλό Dataset (Airfoil Self-Noise)

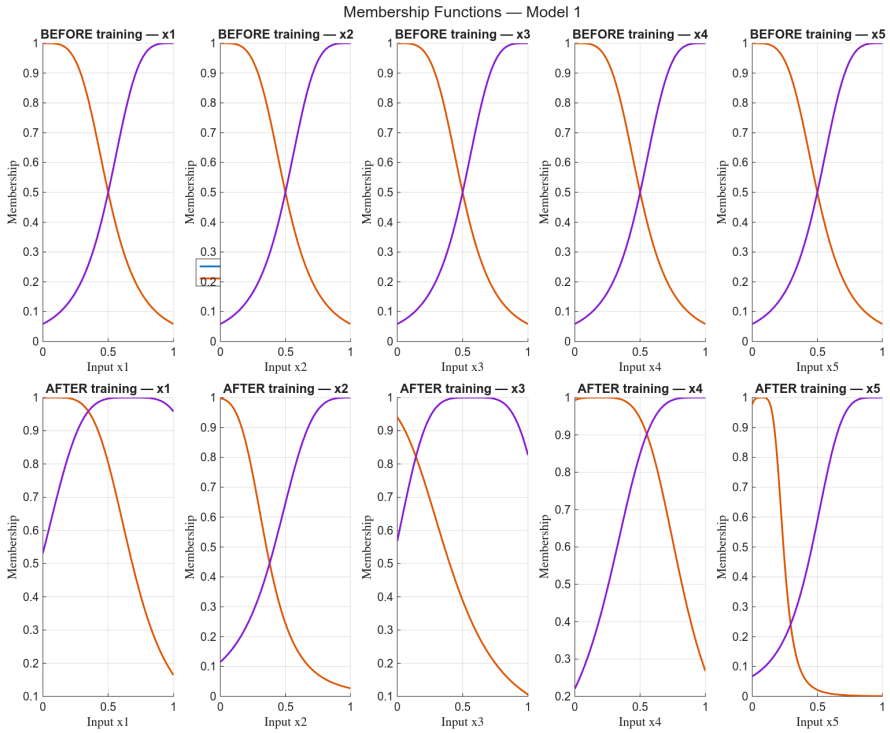
Για την παρούσα εργασία εκπαιδεύτηκαν 4 TSK μοντέλα με **bell-shaped** συναρτήσεις συμμετοχής (gbellmf) μέσω `genfis` (Grid Partition) και `anfis` (υβριδική βελτιστοποίηση, 100 epochs). Τα μοντέλα είναι:

- **Model 1:** `TSK0-2MF` — *constant* έξοδος (singleton), 2 MF ανά είσοδο.
- **Model 2:** `TSK0-3MF` — *constant*, 3 MF ανά είσοδο.
- **Model 3:** `TSK1-2MF` — *linear* έξοδος, 2 MF ανά είσοδο.
- **Model 4:** `TSK1-3MF` — *linear*, 3 MF ανά είσοδο.

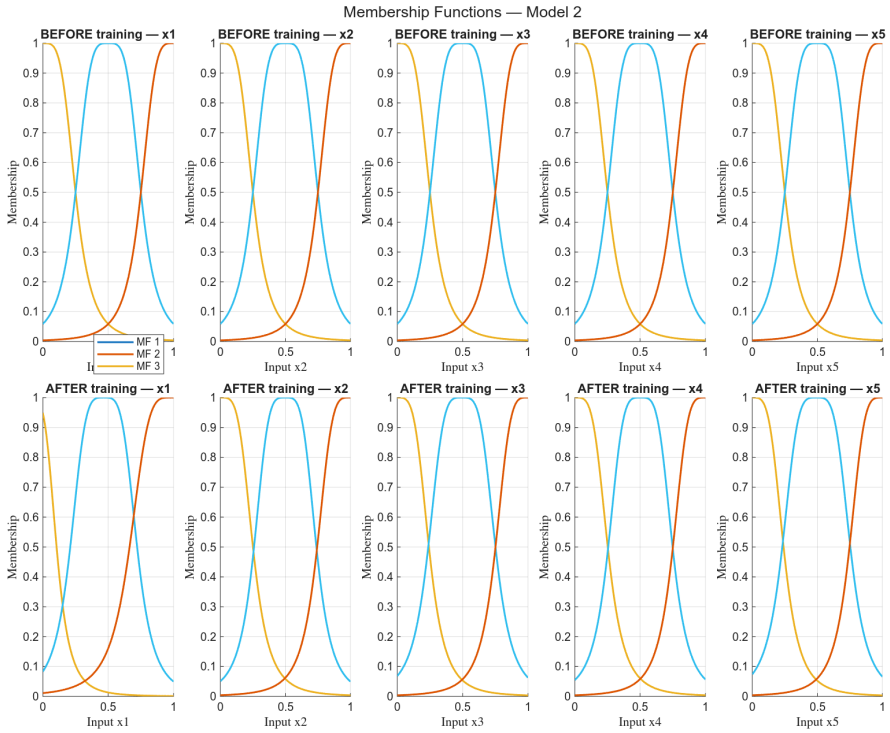
Για κάθε μοντέλο, ως *τελικό* επιλέγεται αυτό με το μικρότερο σφάλμα στο *validation*.

3.1. Συναρτήσεις συμμετοχής

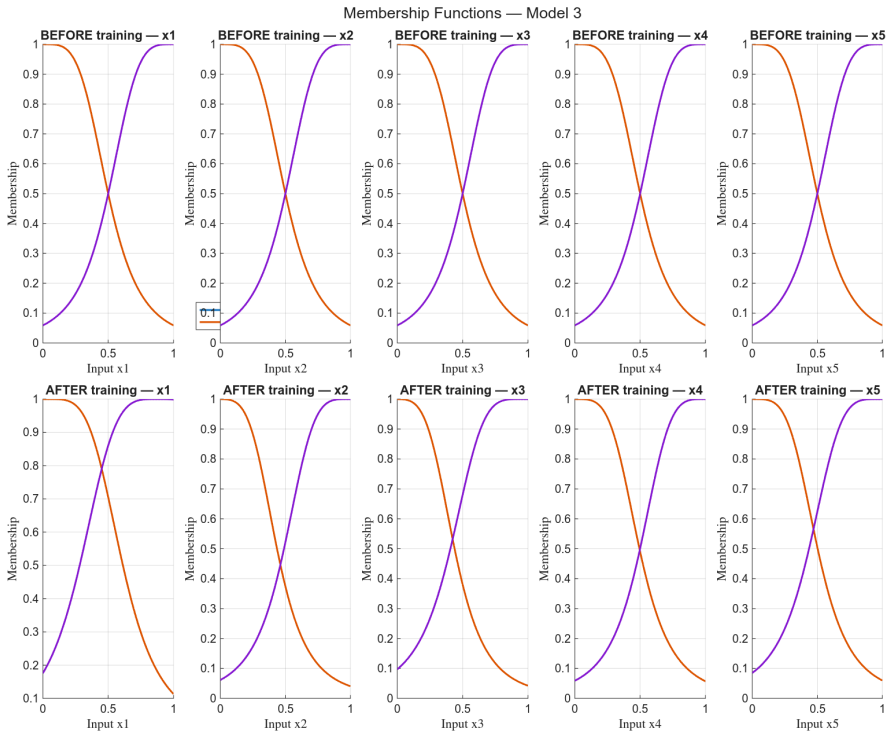
Οι συναρτήσεις συμμετοχής πριν και μετά την εκπαίδευση φαίνονται παρακάτω.



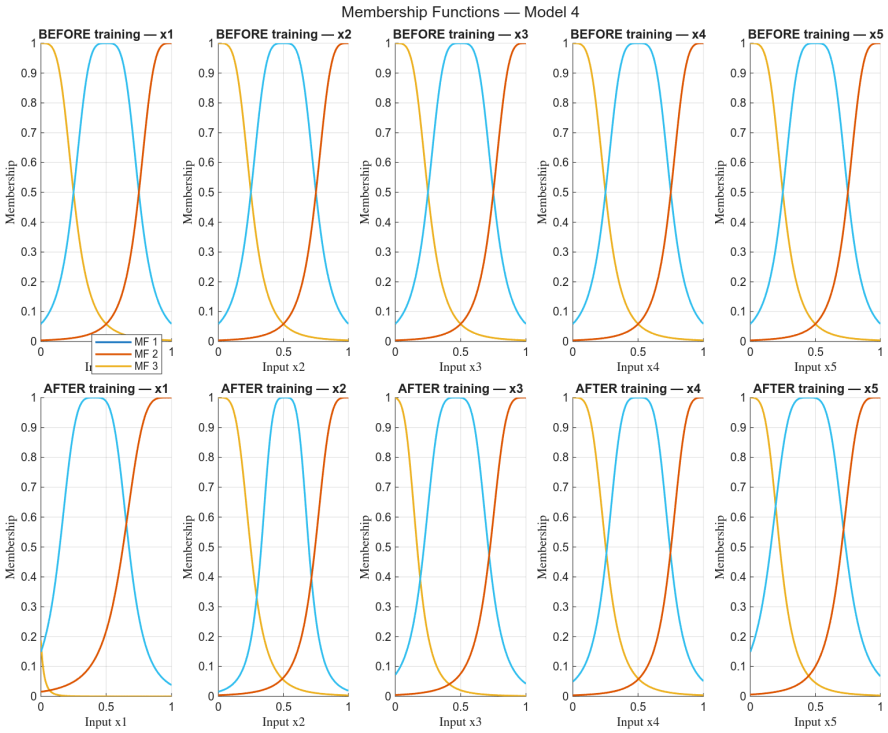
Σχήμα 1: Συναρτήσεις συμμετοχής πριν/μετά για το Model 1 (TSK0-2MF).



Σχήμα 2: Συναρτήσεις συμμετοχής πριν/μετά για το Model 2 (TSK0-3MF).



Σχήμα 3: Συναρτήσεις συμμετοχής πριν/μετά για το Model 3 (TSK1-2MF).

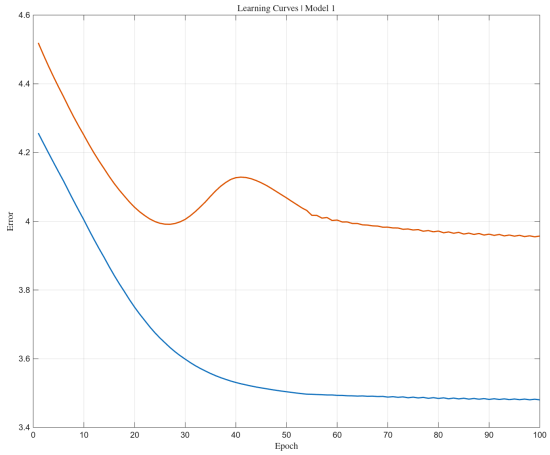


Σχήμα 4: Συναρτήσεις συμμετοχής πριν/μετά για το Model 4 (TSK1-3MF).

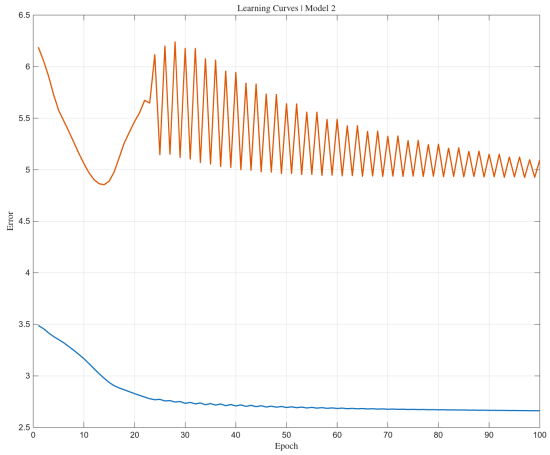
Από τα παραπάνω παρατηρούμε ότι οι MF αναπροσαρμόζονται ομαλά μετά την εκπαίδευση. Με 3 MF/είσοδο αυξάνεται η εκφραστικότητα, αλλά ενδέχεται να αυξηθεί το ρίσκο υπερεκπαίδευσης αν δεν υπάρξει επαρκές regularization/early stopping μέσω validation.

3.2. Διαγράμματα μάθησης

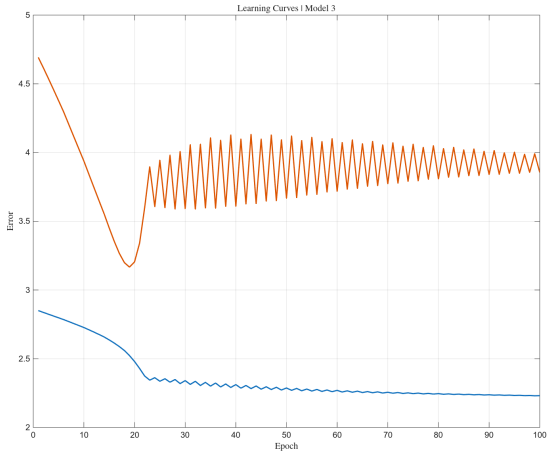
Παρακάτω παραθέτουμε τα διαγράμματα μάθησης.



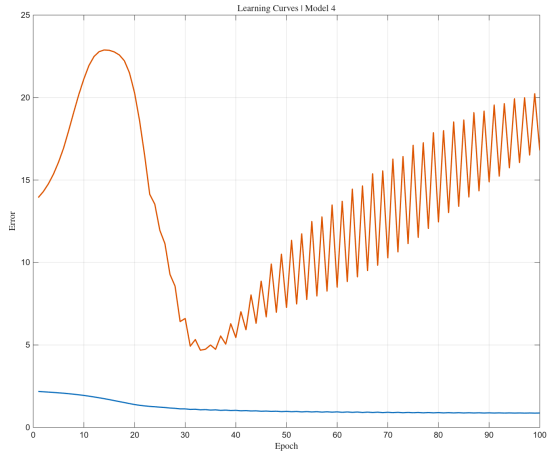
(α′) *Learning curves — Model 1*



(β′) *Learning curves — Model 2*



(γ′) *Learning curves — Model 3*



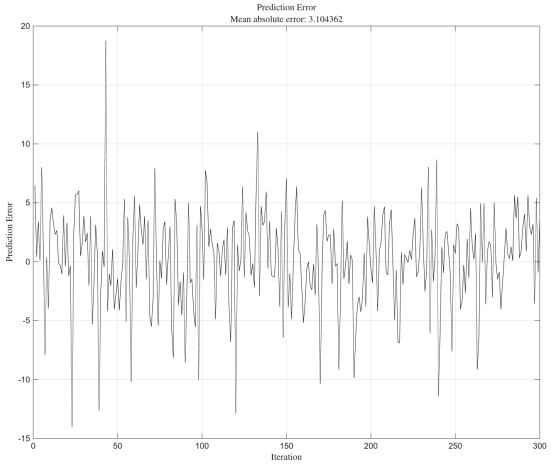
(δ′) *Learning curves — Model 4*

Σχήμα 5: *Καμπύλες μάθησης (train/validation) για τα 4 μοντέλα.*

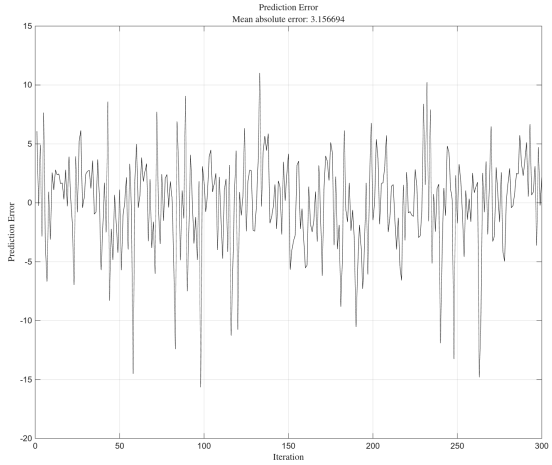
Παρατηρούμε μια σταθερή αποκλιμάκωση του train error και επιλογή εποχών μέσω του ελαχίστου validation. Τα μοντέλα με *linear* έξοδο (ιδίως με 2 MF) συγκλίνουν ταχύτερα.

3.3. Σφάλματα πρόβλεψης (residuals)

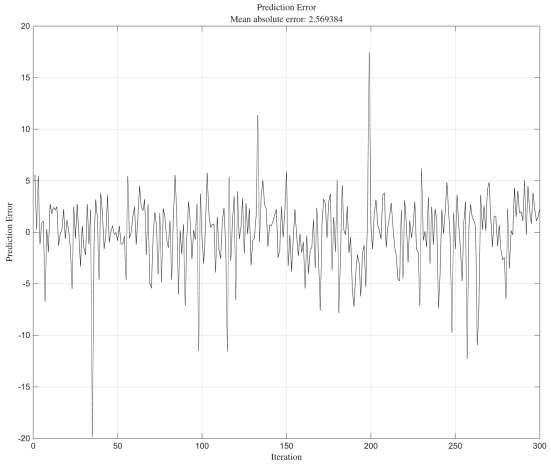
Στο σχήμα 6 βλέπουμε τα σφάλματα πρόβλεψης.



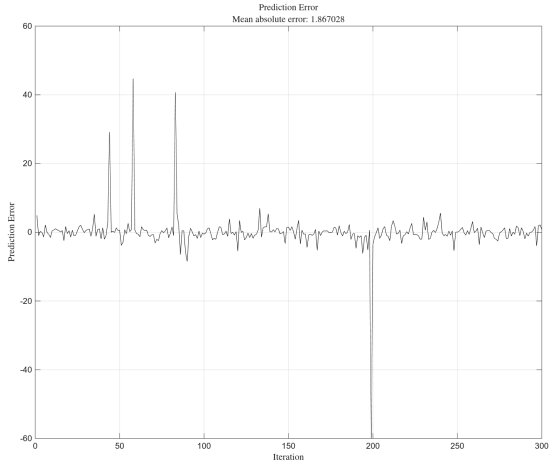
(α′) *Residuals — Model 1*



(β′) *Residuals — Model 2*



(γ′) *Residuals — Model 3*



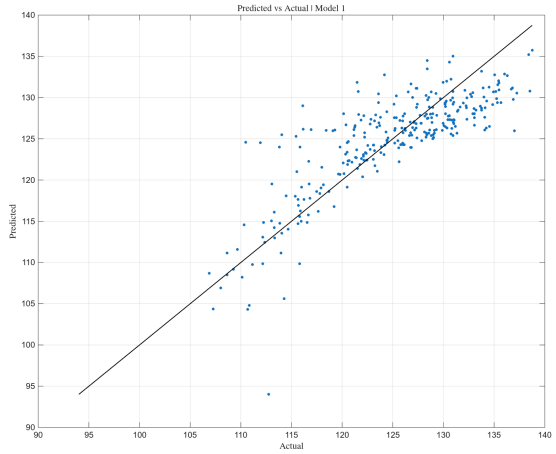
(δ′) *Residuals — Model 4*

Σχήμα 6: *Χρονοσειρές σφάλματος πρόβλεψης στα δεδομένα ελέγχου.*

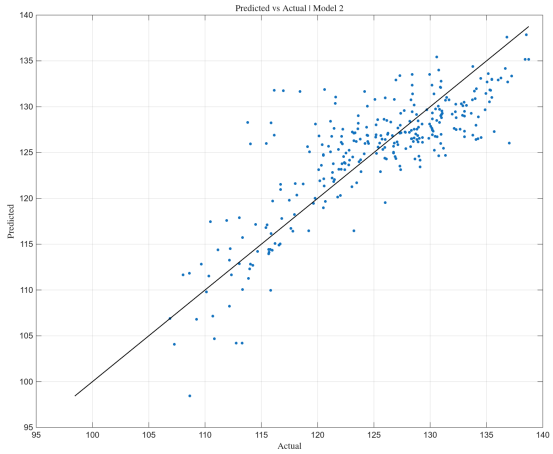
Από τα διαγράμματα παρατηρούμε ότι, για τα καλύτερα μοντέλα, τα σφάλματα πρόβλεψης (residuals) κατανέμονται τυχαία γύρω από το μηδέν, χωρίς να εμφανίζουν συγκεκριμένο μοτίβο. Αντίθετα, αν τα σφάλματα παρουσίαζαν κάποια εμφανή δομή ή συστηματική τάση, αυτό θα σήμαινε ότι το μοντέλο δεν περιγράφει επαρκώς τη σχέση εισόδου–εξόδου και πιθανόν θα χρειαζόταν μεγαλύτερη πολυπλοκότητα (π.χ. περισσότερες συναρτήσεις συμμετοχής).

3.4. Σύγκριση Actual–Predicted

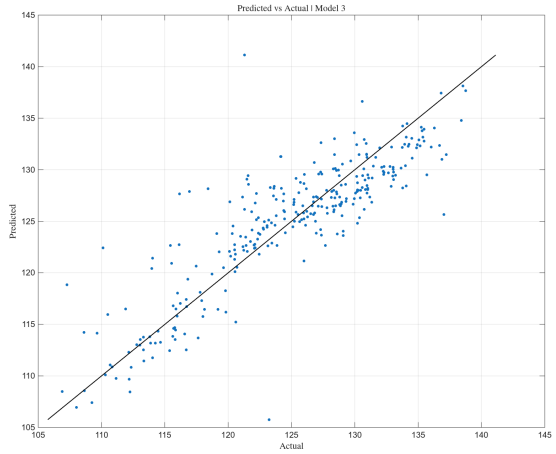
Όσων αφορά τη σύγκριση μεταξύ του μοντέλου πρόβλεψης από την εκτέλεση του script έχουμε τα αποτελέσματα του σχήματος 7:



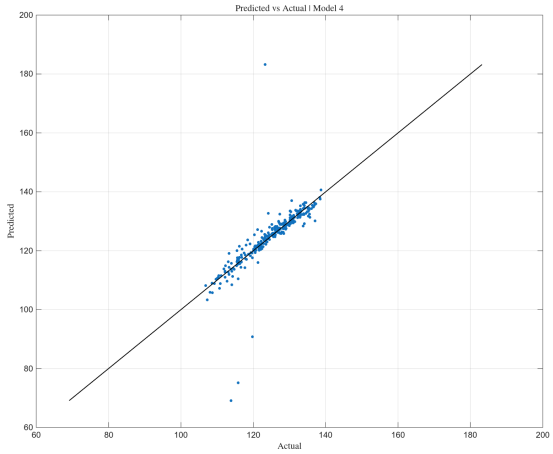
(α′) *Predicted vs Actual — Model 1*



(β′) *Predicted vs Actual — Model 2*



(γ′) *Predicted vs Actual — Model 3*



(δ′) *Predicted vs Actual — Model 4*

Σχήμα 7: *Συσχέτιση προβλέψεων με τις πραγματικές τιμές (ideal: γραμμή $y = x$).*

3.5. Συνολικά αποτελέσματα

Συνολικά έχουμε:

Metrics	**Model-1**	**Model-2**	**Model-3**	**Model-4**
MSE	16.926	17.141	12.895	30.475
RMSE	4.1142	4.1401	3.5910	5.5204
R^2	0.66468	0.66043	0.74454	0.39627
NMSE	0.33532	0.33957	0.25546	0.60373
NDEI	0.57907	0.58272	0.50543	0.77700

Πίνακας 1: *Δείκτες απόδοσης στο test set για τα 4 μοντέλα του Σεναρίου 1.*

Παρατηρούμε πως το **Model 3 (TSK1-2MF)** πετυχαίνει την καλύτερη ισορροπία (χαμηλότερα MSE/RMSE, υψηλότερο R^2), επιβεβαιώνοντας ότι *γραμμική έξοδος με μετριοπαθή διαμέριση (2 MF)* είναι επαρκής για το Airfoil. Η αύξηση σε 3 MF (Model 4) αυξάνει την πολυπλοκότητα και επιδεινώνει τη γενίκευση.

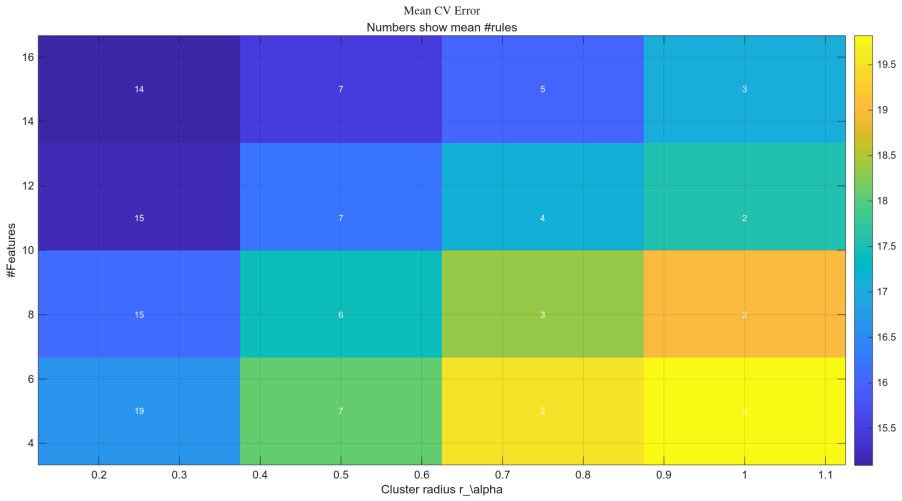
4. Σενάριο 2 — Dataset με υψηλή διαστασιμοτητα (Superconductivity)

4.1. Στόχος και μεθοδολογία

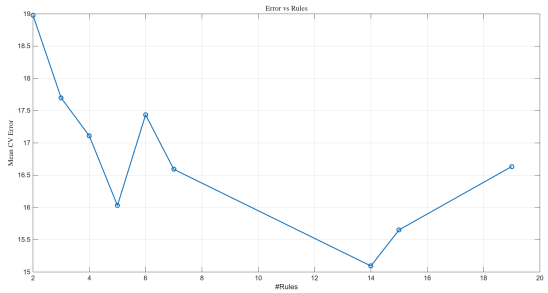
Σε αυτό το σενάριο ο στόχος μας είναι η αποδοτική μοντελοποίηση με περιορισμό της εκθετικής αύξησης κανόνων. Ακολουθούμε:

1. **Split 60/20/20** σε Train/Val/Test και **Min–Max** κλιμάκωση με *στατιστικά train*.
2. **5-fold CV grid search** πάνω σε δύο *ελεύθερες* παραμέτρους:
   - πλήθος χαρακτηριστικών $k\_\text{feat} \in \{5, 8, 11, 15\}$ (επιλογή με ReliefF),
   - ακτίνα επιρροής SC $r_\alpha \in \{0.25, 0.5, 0.75, 1.0\}$ (καθορίζει περίπου #κανόνων).
3. **Τελική εκπαίδευση** με τα βέλτιστα $(k^*_{\text{feat}}, r^*_\alpha)$ και αξιολόγηση στο test.

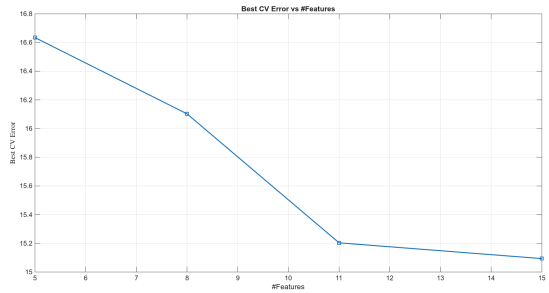
4.2. Διαγνωστικά Grid Search



Σχήμα 8: *Heatmap μέσου σφάλματος CV ανά* (k_feat, r_α) *— με επικάλυψη του #κανόνων.*



(α′) *Μέσο CV σφάλμα ως προς τον αριθμό κανόνων.*



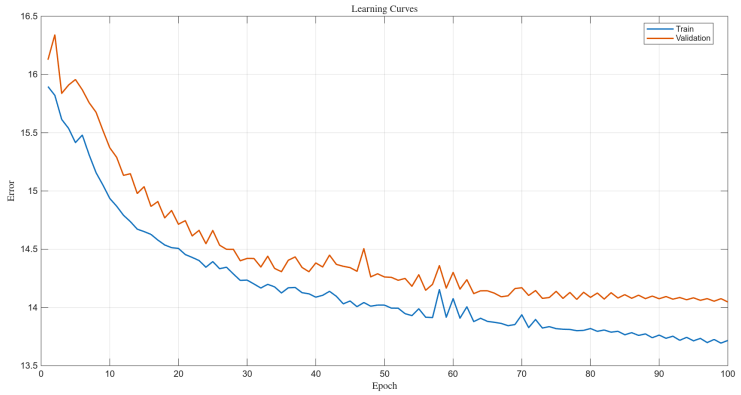
(β′) *Καλύτερο CV σφάλμα ανά πλήθος χαρακτηριστικών.*

Σχήμα 9: *Σχέση σφάλματος με πολυπλοκότητα (κανόνες) και μείωση διαστασιμότητας (χαρακτηριστικά).*

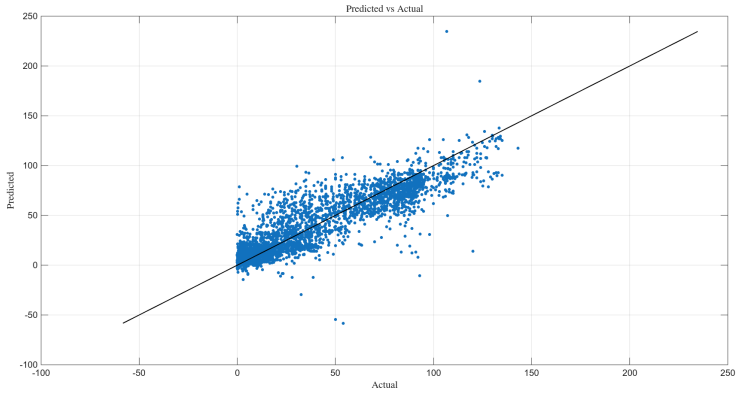
Παρατηρούμε πως υπάρχει το τυπικό *U-σχήμα* ως προς τους κανόνες: λίγοι κανόνες υποπροσαρμόζουν, υπερβολικά πολλοί οδηγούν σε υπερεκπαίδευση. Ως προς τα χαρακτηριστικά, μια *μέτρια* τιμή $k_$feat δίνει την καλύτερη ισορροπία θορύβου/πληροφορίας.

4.3. Τελικό μοντέλο: καμπύλες, προβλέψεις, residuals

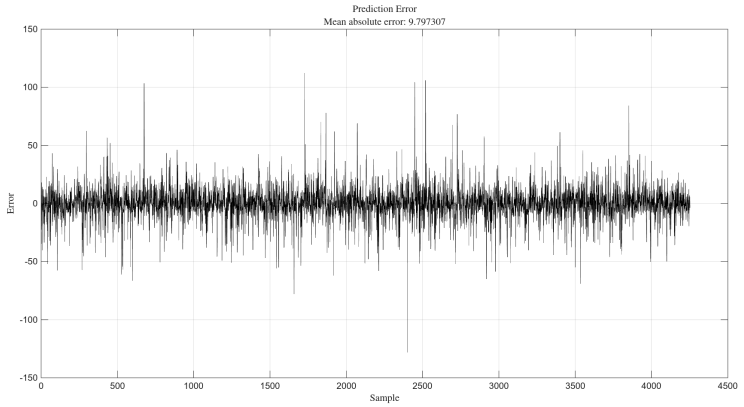
Για το τελικό μοντέλο μετά την εκπαίδευση και αξιολόγηση έχουμε:



Σχήμα 10: *Καμπύλες μάθησης του τελικού μοντέλου (train/validation).*

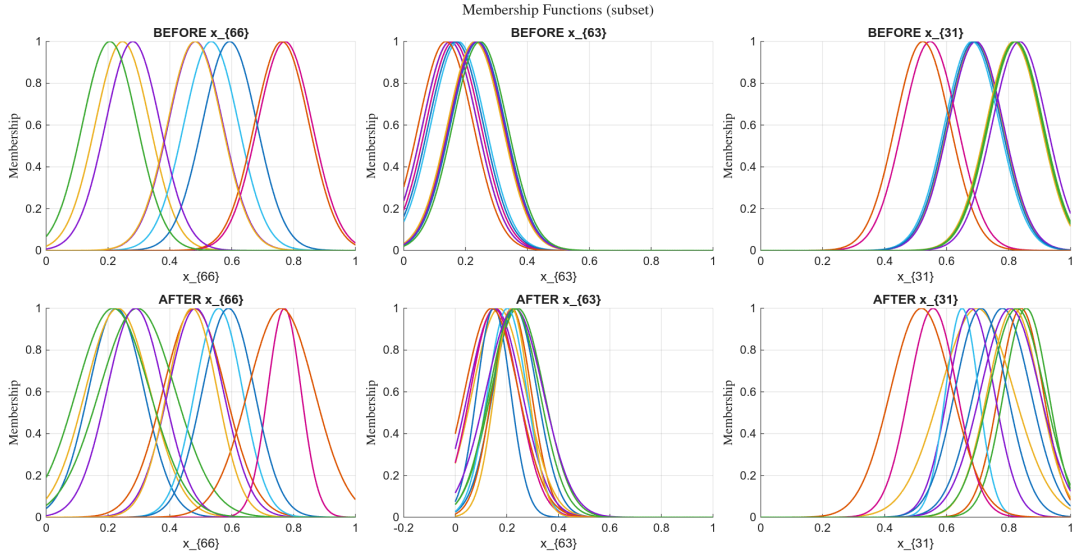


Σχήμα 11: *Σύγκριση* Predicted vs Actual *στο test set.*



Σχήμα 12: *Residuals (σειρά σφάλματος) στο test set.*

4.4. Ενδεικτικές συναρτήσεις συμμετοχής (subset)



Σχήμα 13: *Ενδεικτικές MF για επιλεγμένες εισόδους πριν/μετά την εκπαίδευση.*

4.5. Δείκτες απόδοσης τελικού μοντέλου (Test set)

Στην εκτέλεση που παραθέτουμε, το τελικό μοντέλο προέκυψε με *12 κανόνες* και η επίδοση στο test είναι:

Metric	**Τιμή**
MSE	208.506
RMSE	14.4397
R^2	0.820108
NMSE	0.179892
NDEI	0.424137

Πίνακας 2: *Απόδοση τελικού μοντέλου στο test (Superconductivity).*

Παρατηρούμε πως η επίδοση είναι ισχυρή ($R^2 \approx 0.82$) με σχετικά μικρό αριθμό κανόνων, επιβεβαιώνοντας ότι ο συνδυασμός *ReliefF + SC* ελέγχει αποτελεσματικά την πολυπλοκότητα.

4.6. Συμπεράσματα

Η επιλογή χαρακτηριστικών και η ρύθμιση της ακτίνας r_α είναι κρίσιμες. Το grid search με 5-fold CV παρείχε αξιόπιστη εκτίμηση σφάλματος και οδήγησε σε μοντέλο με *καλή γενίκευση* και *περιορισμένο* πλήθος κανόνων. Αν υιοθετούσαμε *Grid Partition* με 2–3 MF/είσοδο για ίδιο πλήθος χαρακτηριστικών, ο αριθμός κανόνων θα κλιμακωνόταν εκθετικά ($\sim m^d$), καθιστώντας το μοντέλο δυσκολότερο στην εκπαίδευση και λιγότερο ελέγξιμο. Το SC αποφεύγει αυτή την «έκρηξη κανόνων» δημιουργώντας data-driven κανόνες με συμπαγή κάλυψη του χώρου εισόδου.

4.7. Επίλογος

Αναπτύξαμε μια ολοκληρωμένη διαδικασία παλινδρόμησης με μοντέλα TSK, η οποία καλύπτει όλα τα στάδια — από τον σωστό διαχωρισμό και την προεπεξεργασία των δεδομένων, μέχρι την εκπαίδευση, την αξιολόγηση και την ανάλυση των αποτελεσμάτων.

Στο **Σενάριο 1** φάνηκε ότι ένα μοντέλο με γραμμική έξοδο και μικρό αριθμό συναρτήσεων συμμετοχής μπορεί να αποδώσει ικανοποιητικά. Στο **Σενάριο 2** διαπιστώσαμε ότι ο συνδυασμός *επιλογής χαρακτηριστικών, Subtractive Clustering και διασταυρούμενης επικύρωσης (cross validation)* επιτρέπει τη δημιουργία αποδοτικών και εύκολα ερμηνεύσιμων μοντέλων, χωρίς την εκθετική αύξηση της πολυπλοκότητας που εμφανίζεται στις παραδοσιακές μεθόδους grid partition.