ΕΡΓΑΣΙΑ ΣΤΑ ΑΣΑΦΗ ΣΥΣΤΗΜΑΤΑ

ΟΝΟΜΑ: ΚΩΝΣΤΑΝΤΙΝΟΣ
ΕΠΙΘΕΤΟ: ΛΕΤΡΟΣ
ΣΧΟΛΗ: ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ: ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧ. ΚΑΙ ΜΗΧ. ΥΠΟΛΟΓΙΣΤΩΝ
ΑΕΜ: 8851
ΕΞΑΜΗΝΟ: 8ο
ΕΤΟΣ: 2019

Επίλυση προβλημάτων παλινδρόμησης με χρήση μοντέλων TSK

Ομάδα 3 – S02

Περιεχόμενα

Περιγραφή του Προβλήματος

Στόχος αυτής της εργασίας είναι η διερεύνηση της ικανότητας των TSK μοντέλων στη μοντελοποίηση πολυμεταβλητών μη γραμμικών συναρτήσεων, με χρήση ασαφών νευρωνικών μοντέλων. Η εργασία διακρίνεται σε δύο τμήματα στα οποία θα χρησιμοποιηθούν δύο διαφορετικά σετ δεδομένων. Σκοπός του πρώτου τμήματος είναι η εκπαίδευση και αξιολόγηση τεσσάρων TSK μοντέλων με διαφορετικό πλήθος συναρτήσεων συμμετοχής εισόδου και διαφορετική μορφή εξόδου. Στη συνέχεια, στο δεύτερο μέρος γίνεται χρήση εναλλακτικών μεθόδων αντιμετώπισης του παραπάνω προβλήματος, καθώς το πλήθος χαρακτηριστικών του δεύτερου σετ δεδομένων καθιστά τις μεθόδους που χρησιμοποιήθηκαν προηγουμένως απαγορευτικές, ενώ παράλληλα γίνεται διαχωρισμός του σετ δεδομένων σε τμήματα για την αναζήτηση του μοντέλου με το μικρότερο σφάλμα.

Εφαρμογή στο Σετ Δεδομένων Combined Cycle Power Plant (CCPP)

Προετοιμασία του Σετ Δεδομένων

Το Combined Cycle Power Plant Dataset της UCI αποτελείται από δεδομένα μίας μονάδας συνδυασμένου κύκλου και περιέχει 9568 δείγματα, κάθε ένα από τα οποία περιγράφεται από 4 χαρακτηριστικά (Features). Συγκεκριμένα, τα χαρακτηριστικά αυτά είναι η μέση ωριαία θερμοκρασία (Temperature - T), η μέση ωριαία ατμοσφαιρική πίεση (Ambient Pressure - AP), η μέση ωριαία σχετική υγρασία (Relative Humidity - RH) και η μέση ωριαία πίεση καυσαερίων (Exhaust Vacuum - V). Χρησιμοποιώντας τα δεδομένα αυτά, επιδιώκουμε να προβλέψουμε τη μέση ενεργειακή απόδοση της μονάδας ανά ώρα.

Πραγματοποιούμε διαχωρισμό του σετ δεδομένων σε τρία μη επικαλυπτόμενα υποσύνολα ως εξής:

1. 60% : Σετ Εκπαίδευσης – training set
2. 20% : Σετ Επικύρωσης – validation set
3. 20% : Σετ Ελέγχου – check set

Περιγραφή της Διαδικασίας Εκπαίδευσης

Η εκπαίδευση γίνεται με την υβριδική μέθοδο, δηλαδή οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται με τη μέθοδο Backpropagation και οι παράμετροι της πολυωνυμικής συνάρτησης εξόδου βελτιστοποιούνται με τη μέθοδο Least Squares. Τα τέσσερα μοντέλα TSK προς εκπαίδευση διακρίνονται με βάση τον παρακάτω πίνακα.

	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

Σχήμα 1: Πίνακας προδιαγραφών των TSK Μοντέλων

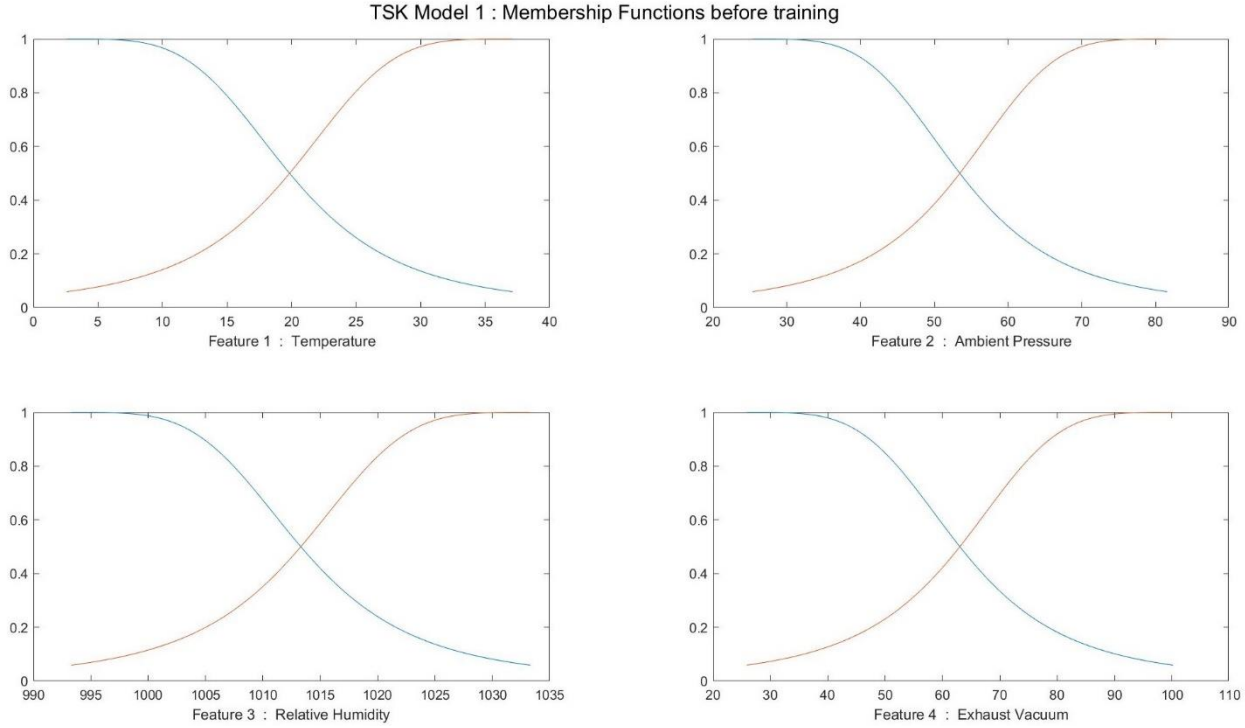
Αρχικά, δημιουργούμε, με τη συνάρτηση genfis() του MATLAB, το μοντέλο προς εκπαίδευση με βάση τα χαρακτηριστικά του πίνακα ανάλογα με τον αριθμό του μοντέλου και τη μέθοδο Grid Partition, δίνοντας ως είσοδο τα δεδομένα εκπαίδευσης.

Στη συνέχεια εκπαιδεύουμε το μοντέλο με χρήση της συνάρτησης anfis() του MATLAB για 250 εποχές, προχωρούμε στην αξιολόγηση του μοντέλου και τέλος υπολογίζουμε τις ζητούμενες μετρικές σφάλματος $MSE, RMSE, R^2, NMSE, NDEI$. Οι καμπύλες εκμάθησης σχεδιάζονται με βάση το σφάλμα $RMSE$.

Αποτελέσματα TSK Μοντέλων και Μετρικές Σφάλματος

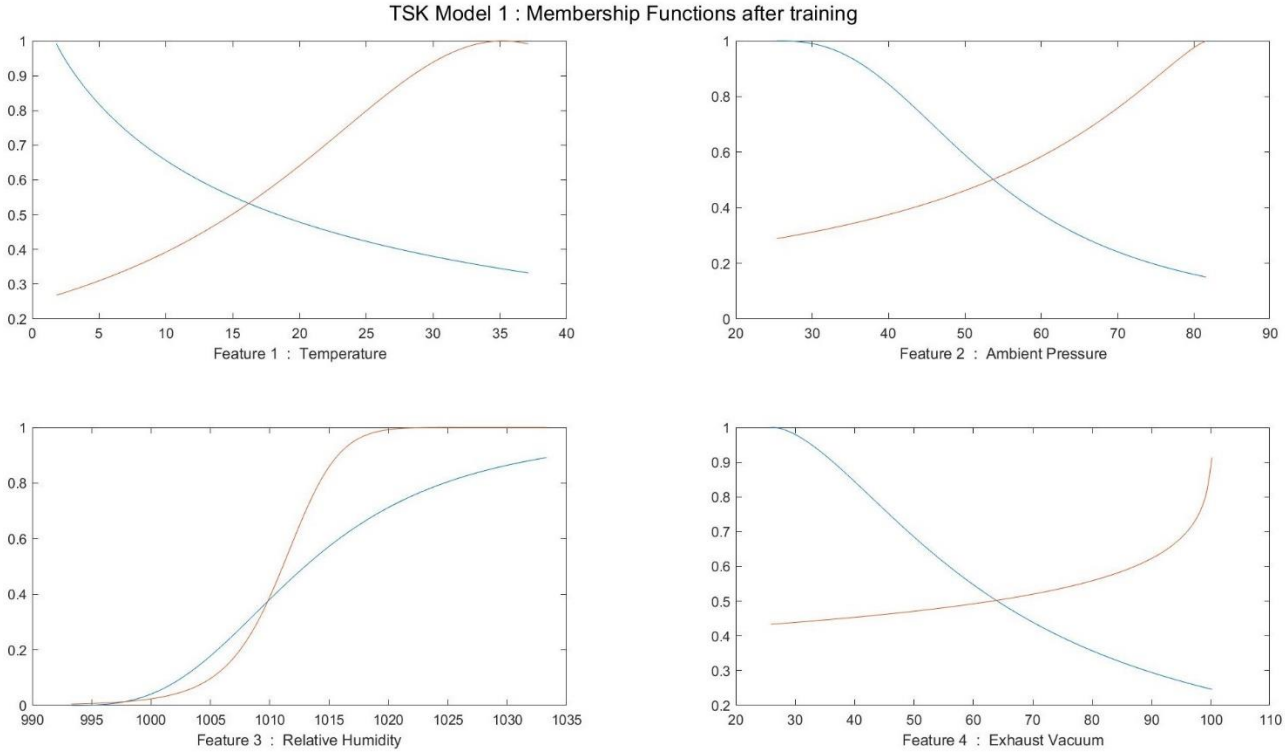
TSK Μοντέλο 1

Στο πρώτο μοντέλο TSK χρησιμοποιούμε 2 συναρτήσεις συμμετοχής τύπου Bell-Shaped με επικάλυψη 0.5 για κάθε μεταβλητή εισόδου ενώ η μορφή της εξόδου είναι Singleton (Constant). Οι συναρτήσεις αυτές πριν τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 2.



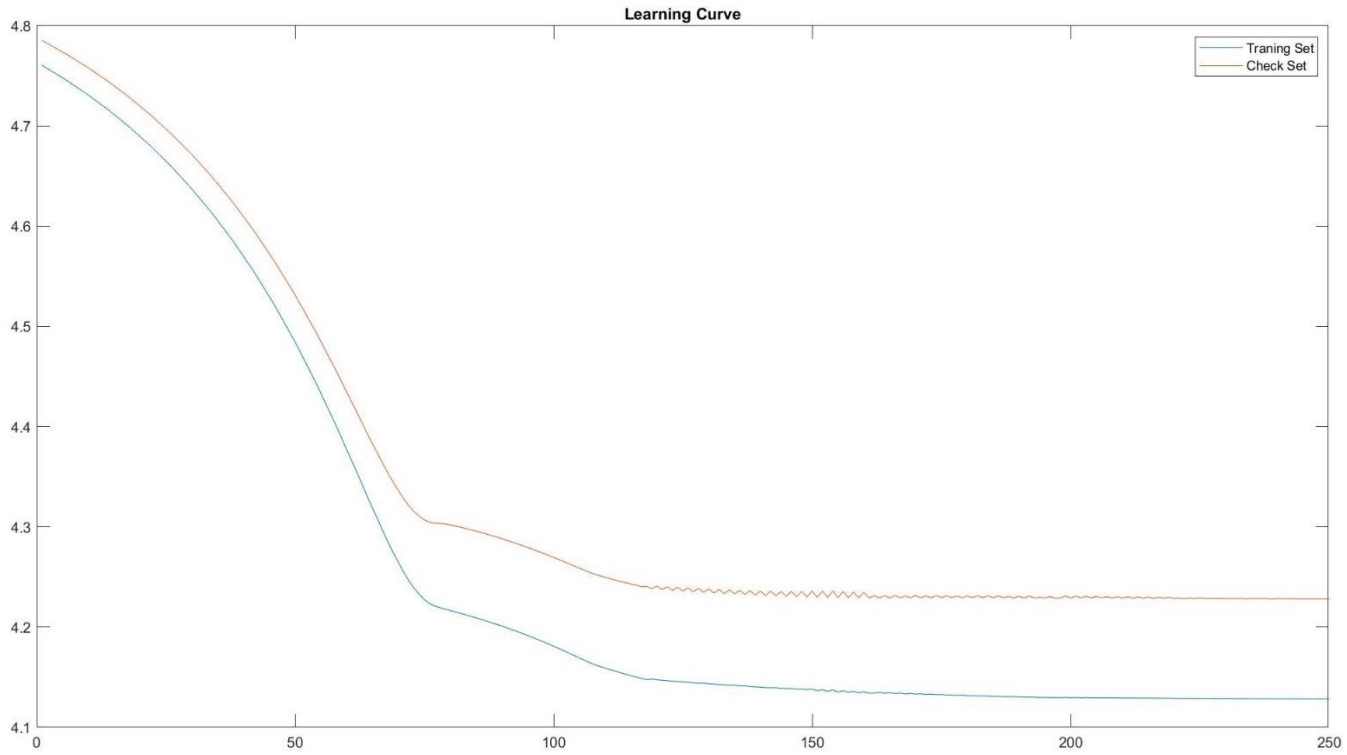
Σχήμα 2: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 1

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



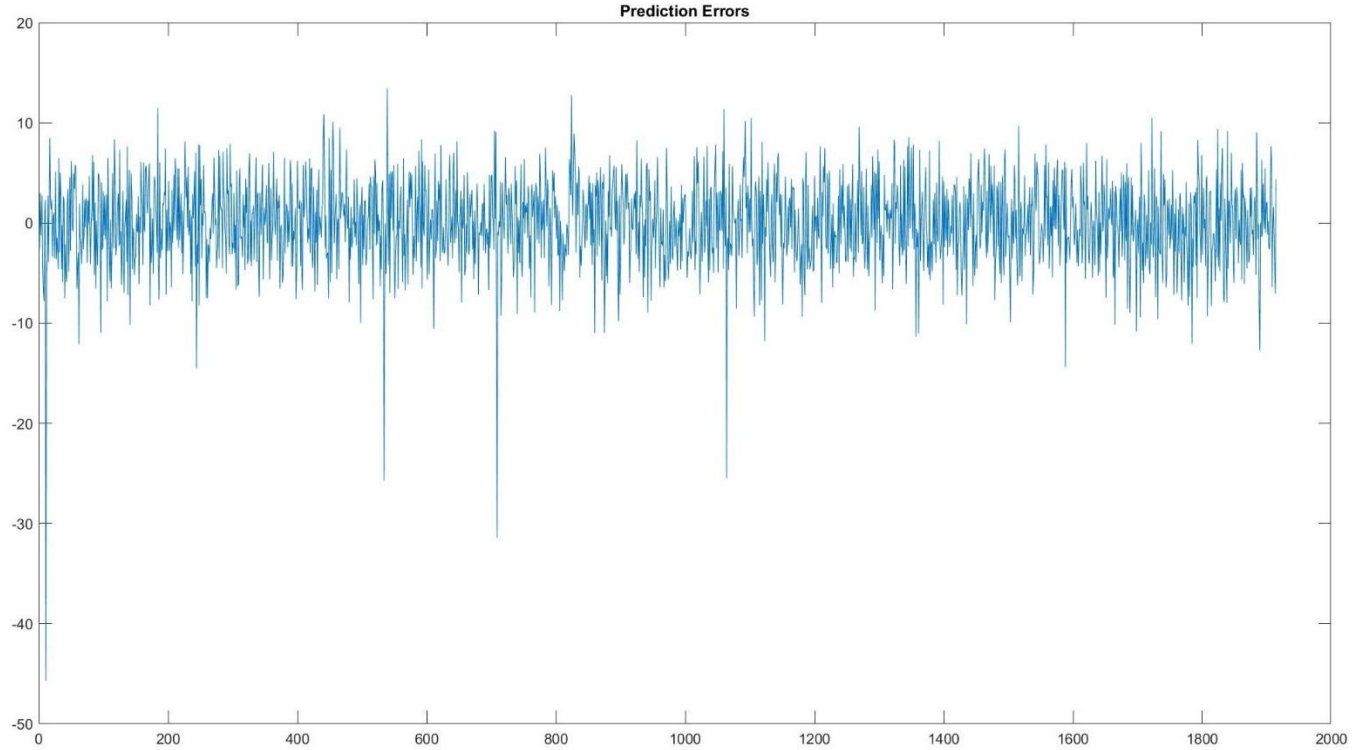
Σχήμα 3: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 1

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

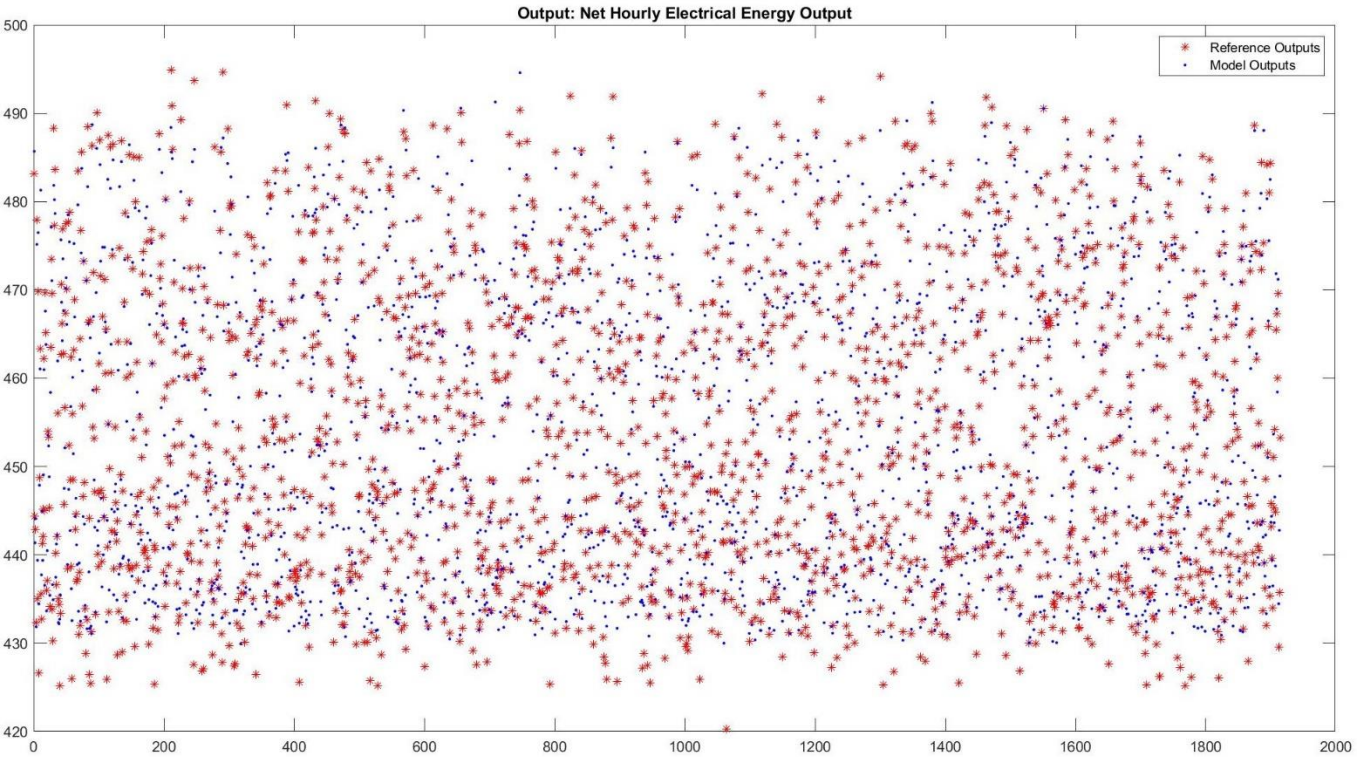


Σχήμα 4: Καμπύλες Εκμάθησης - TSK Μοντέλο 1

Τέλος, βλέπουμε τα σφάλματα πρόβλεψης και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.



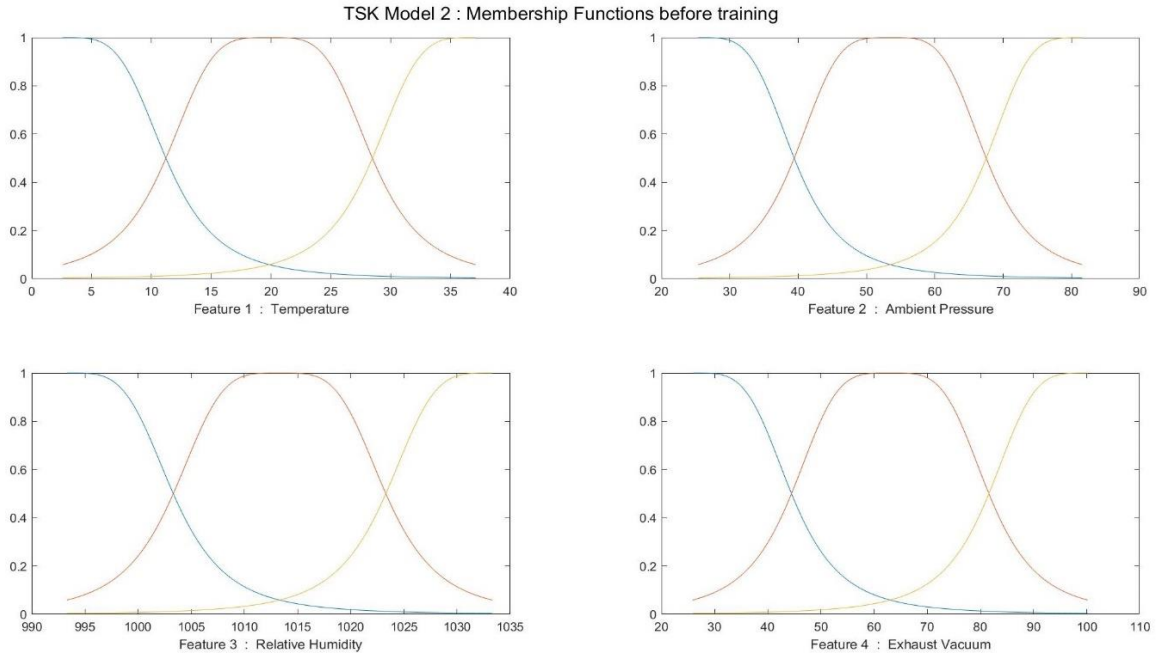
Σχήμα 5: Σφάλματα Πρόβλεψης - TSK Μοντέλο 1



Σχήμα 6: Πραγματική και Εκτιμήτρια Έξοδος - TSK Μοντέλο 1

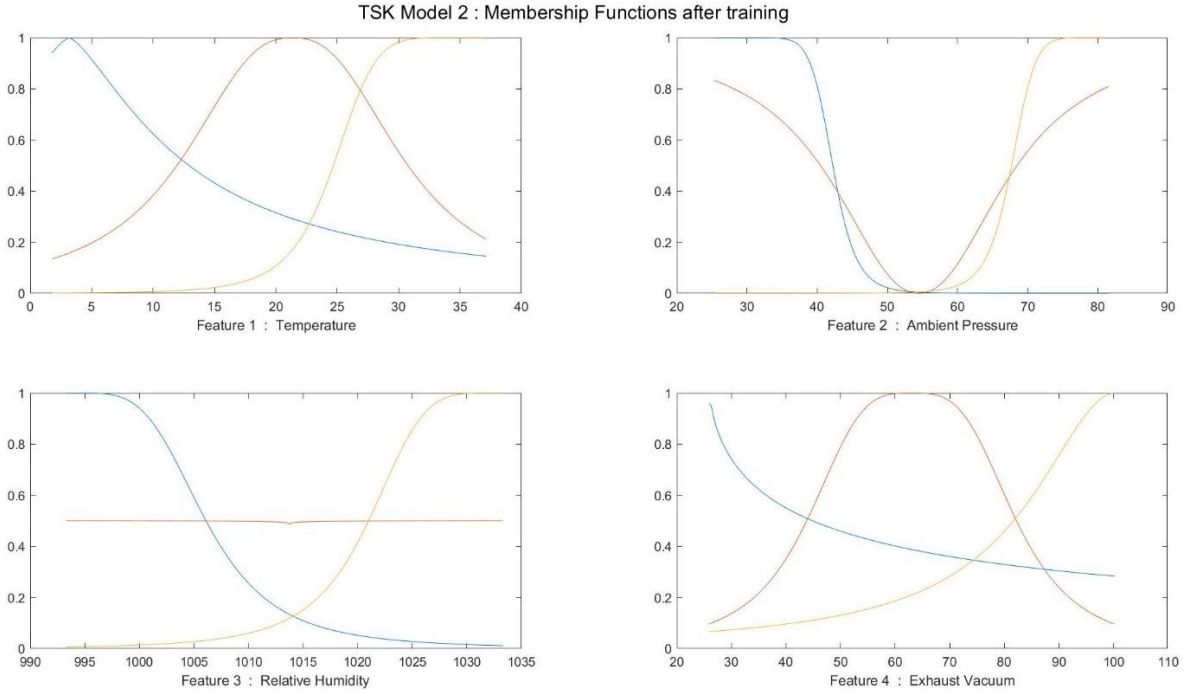
TSK Μοντέλο 2

Στο δεύτερο μοντέλο TSK χρησιμοποιούμε 3 συναρτήσεις συμμετοχής τύπου Bell-Shaped με επικάλυψη 0.5 για κάθε μεταβλητή εισόδου ενώ η μορφή της εξόδου είναι Singleton (Constant). Οι συναρτήσεις αυτές πριν τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 7.



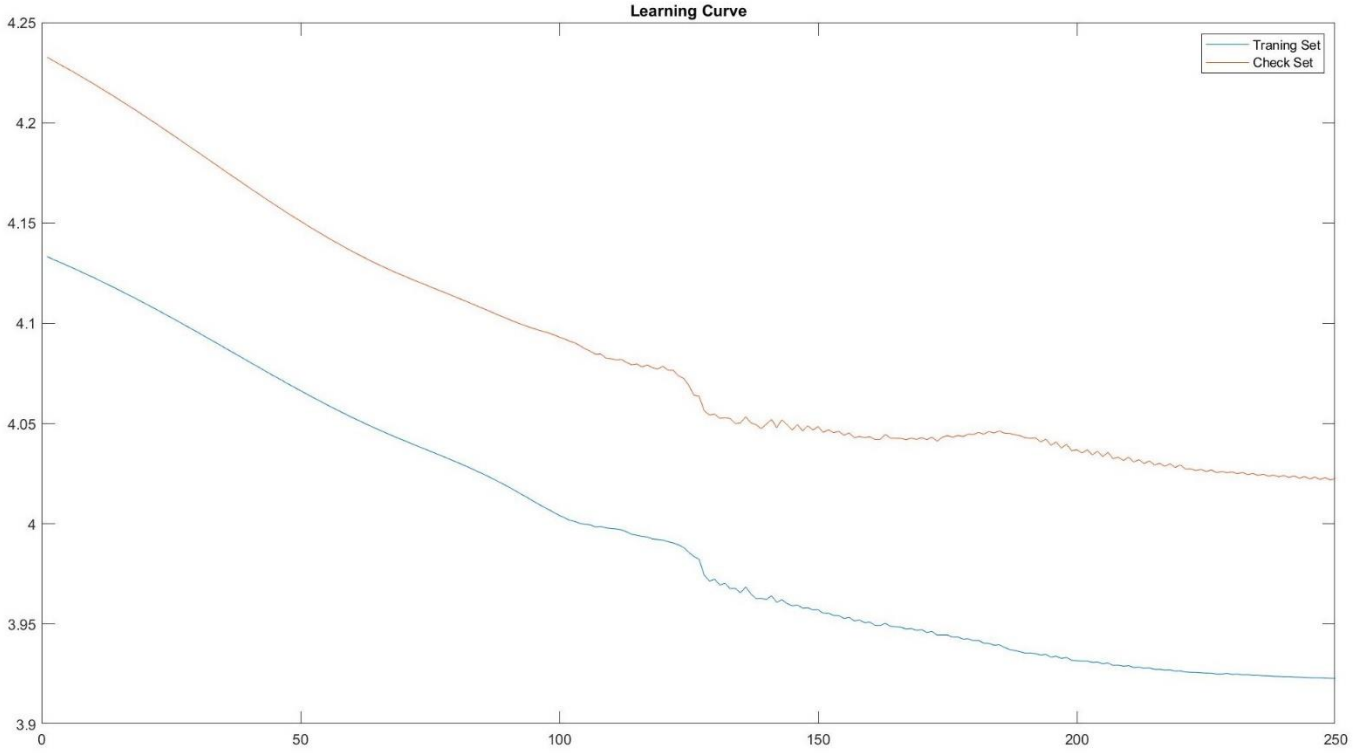
Σχήμα 7: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 2

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



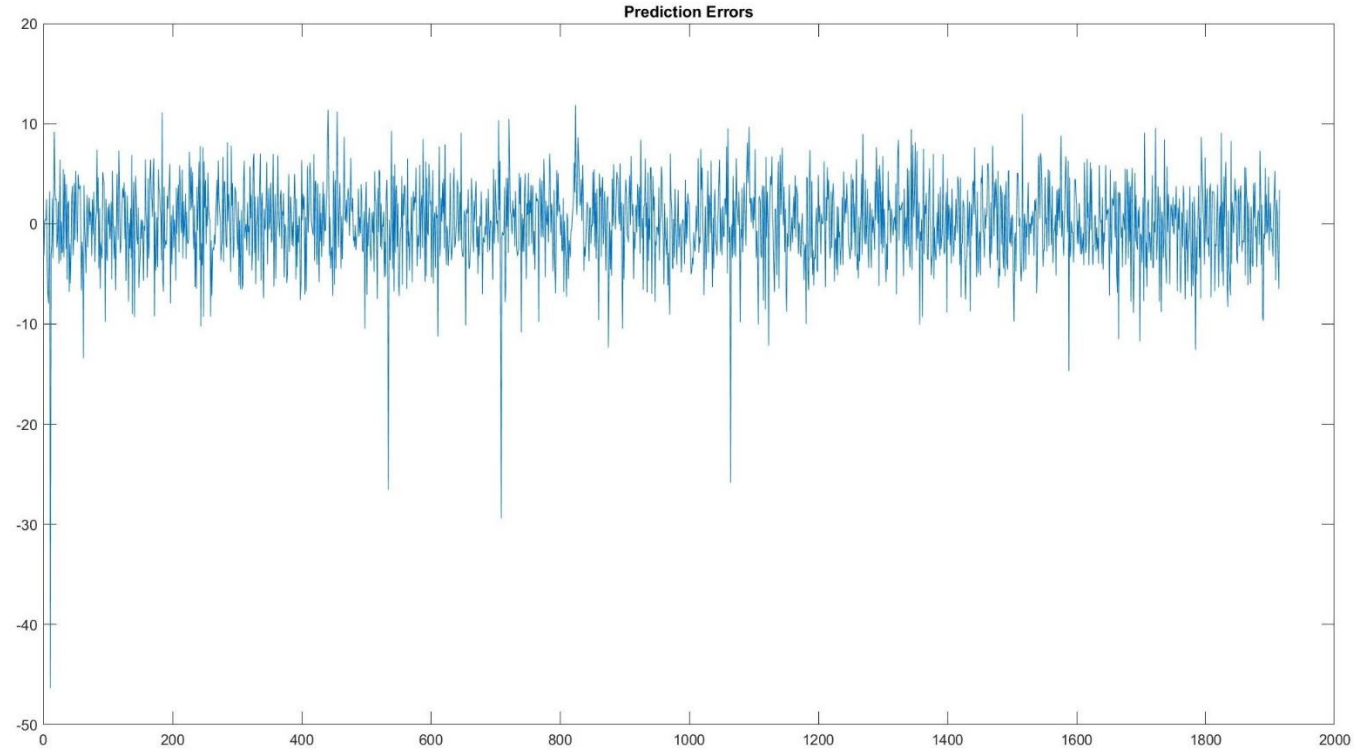
Σχήμα 8: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 2

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

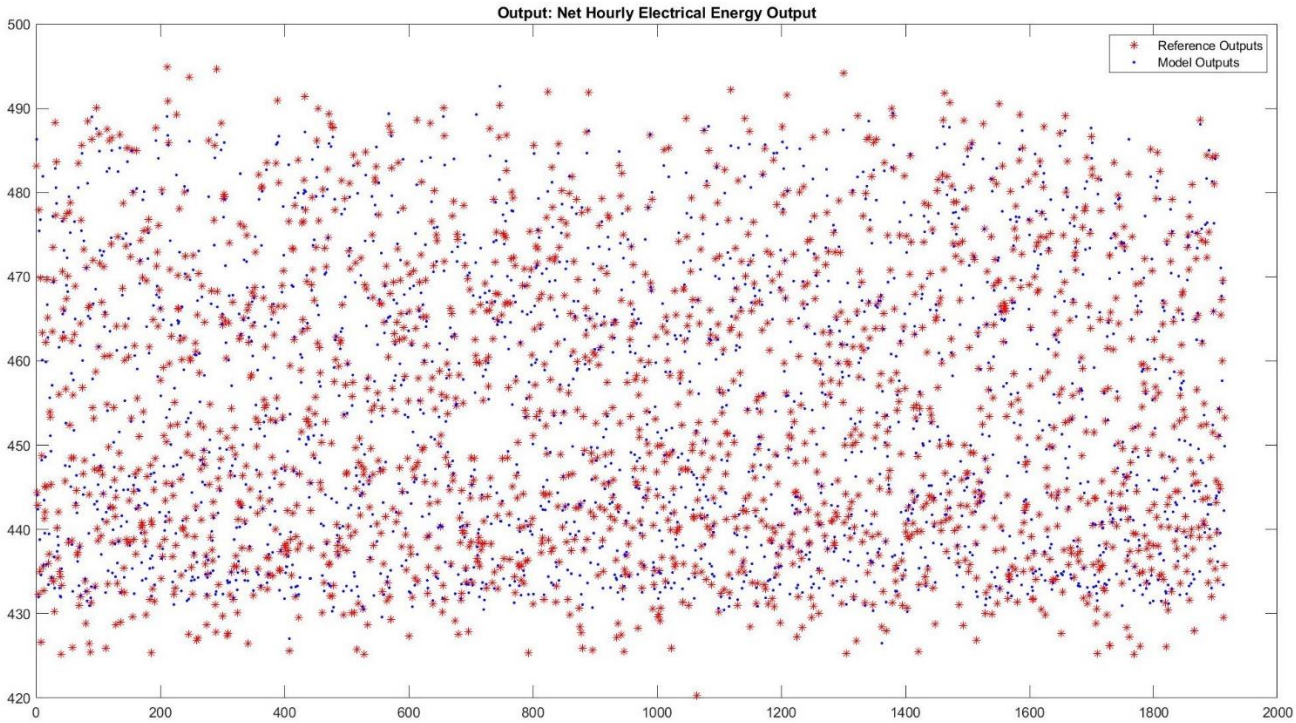


Σχήμα 9: Καμπύλες Εκμάθησης - TSK Μοντέλο 2

Τέλος, βλέπουμε τα σφάλματα πρόβλεψης και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.



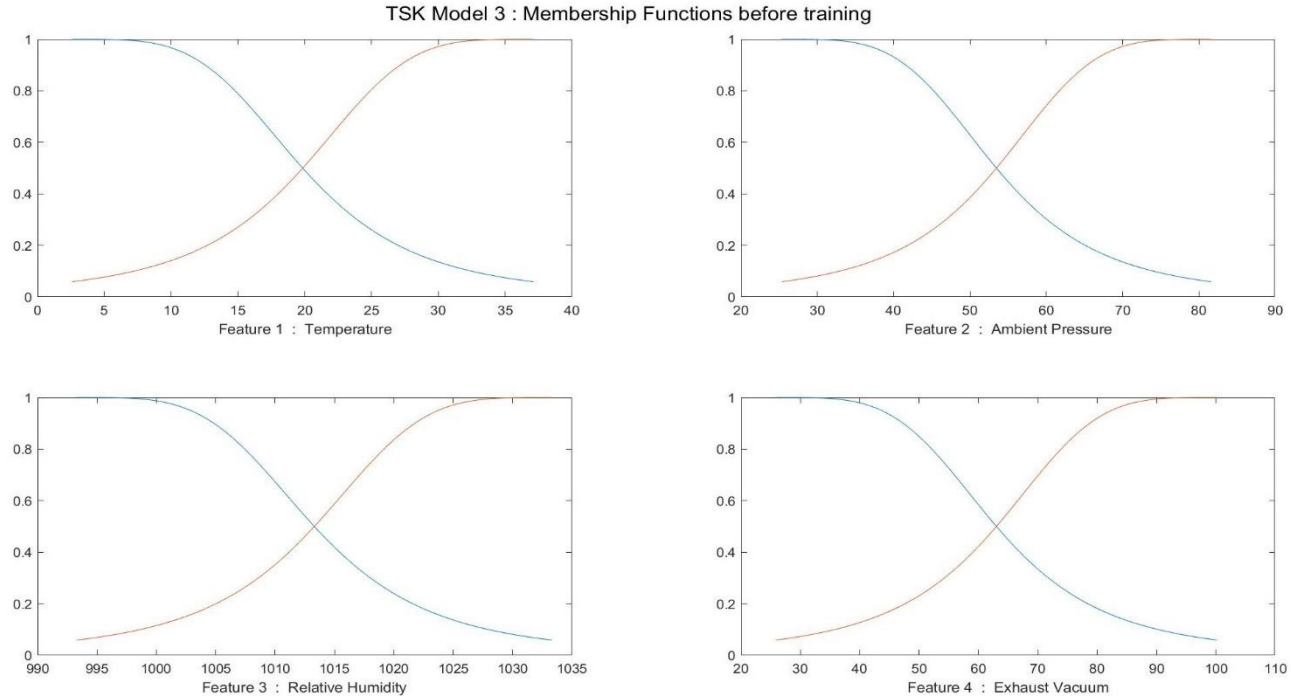
Σχήμα 10: Σφάλματα Πρόβλεψης - TSK Μοντέλο 2



Σχήμα 11: Πραγματική και Εκτιμήτρια Έξοδος - TSK Μοντέλο 2

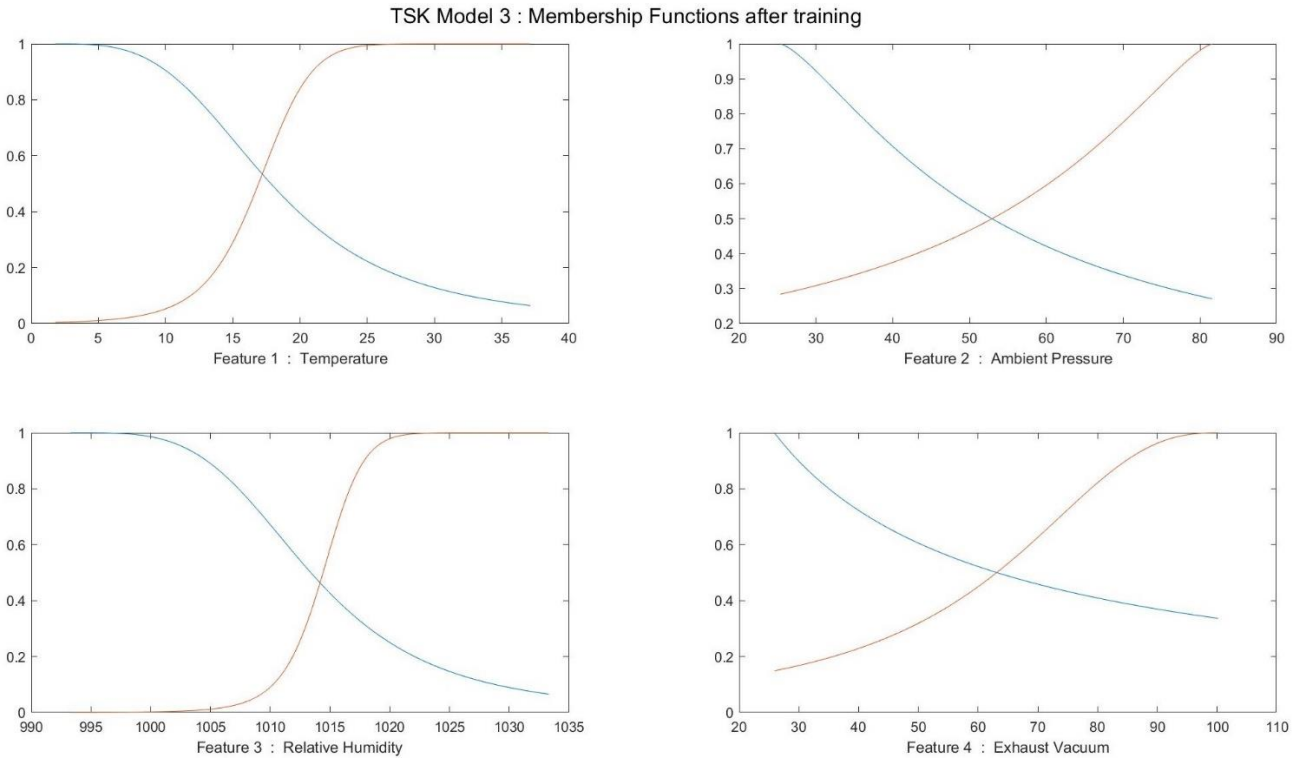
TSK Μοντέλο 3

Στο τρίτο μοντέλο TSK χρησιμοποιούμε 2 συναρτήσεις συμμετοχής τύπου Bell-Shaped με επικάλυψη 0.5 για κάθε μεταβλητή εισόδου ενώ η μορφή της εξόδου είναι Polynomial (Linear). Οι συναρτήσεις αυτές πριν τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 12.



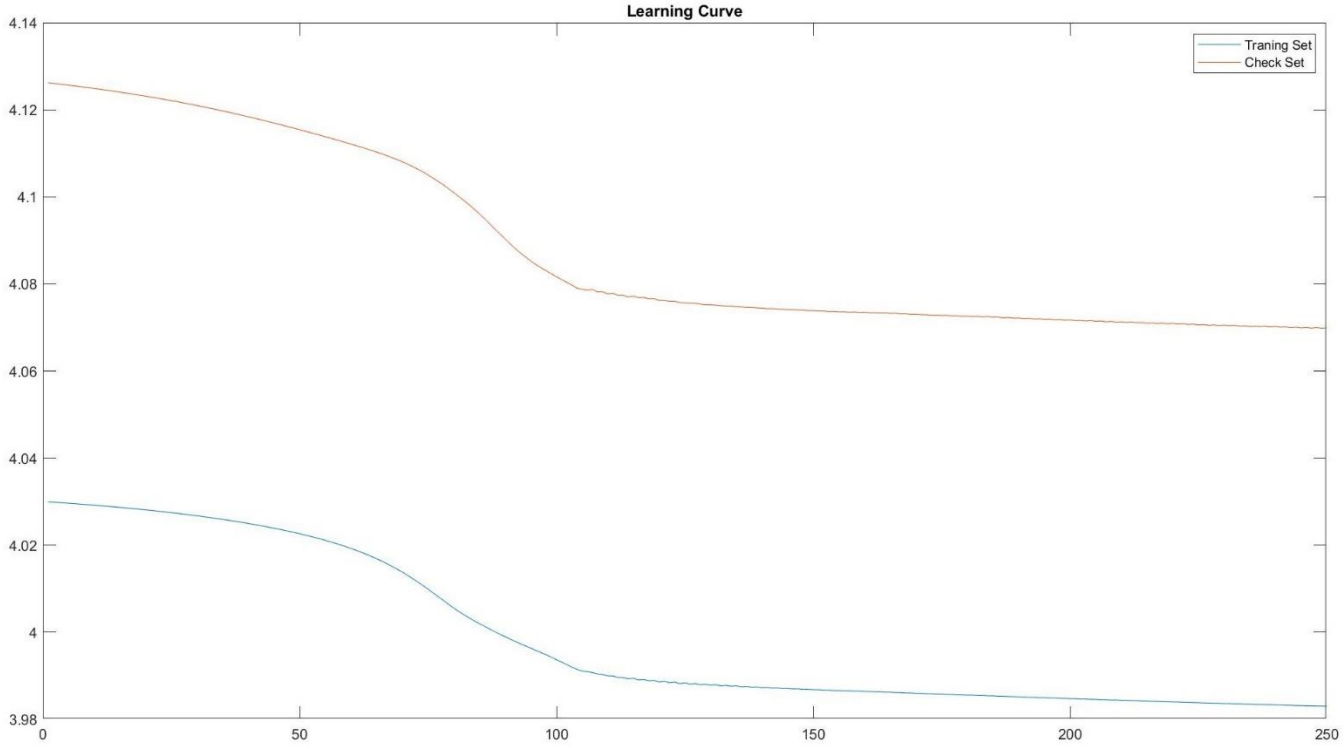
Σχήμα 12: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 3

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



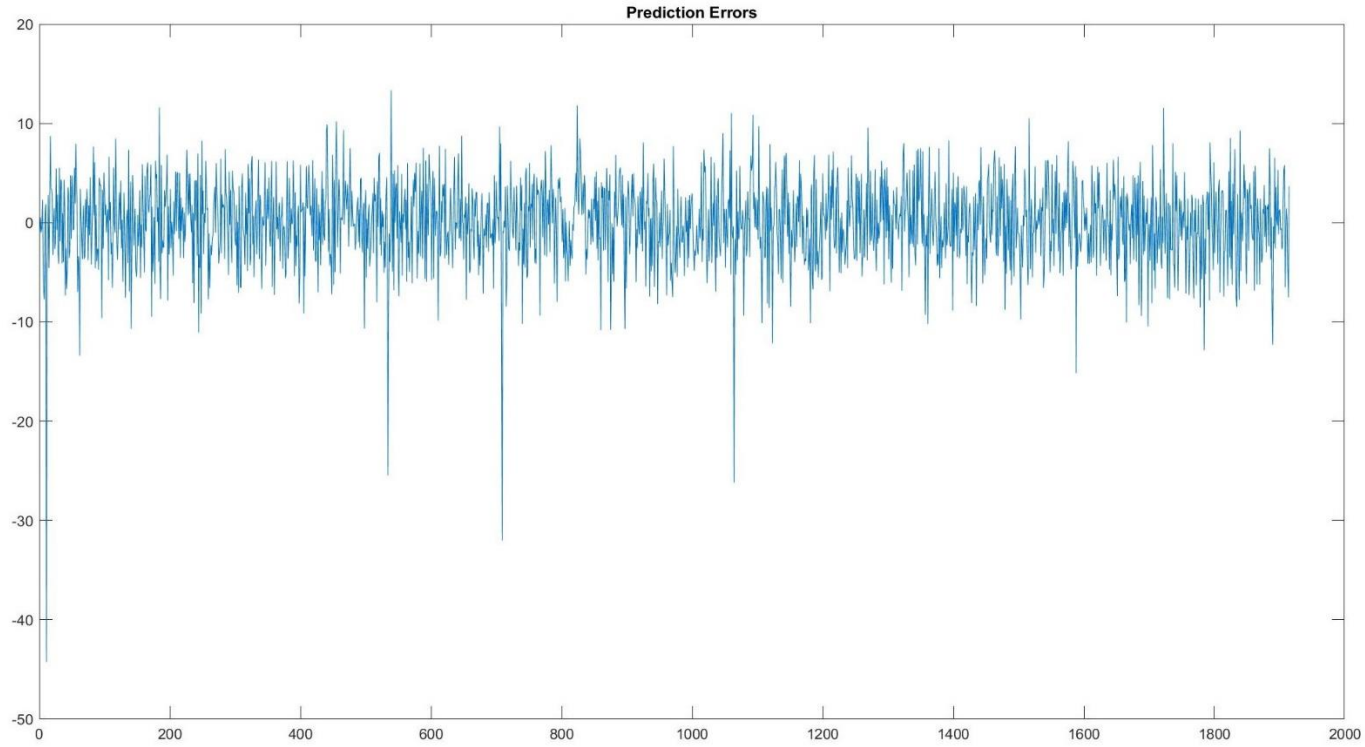
Σχήμα 13: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 3

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

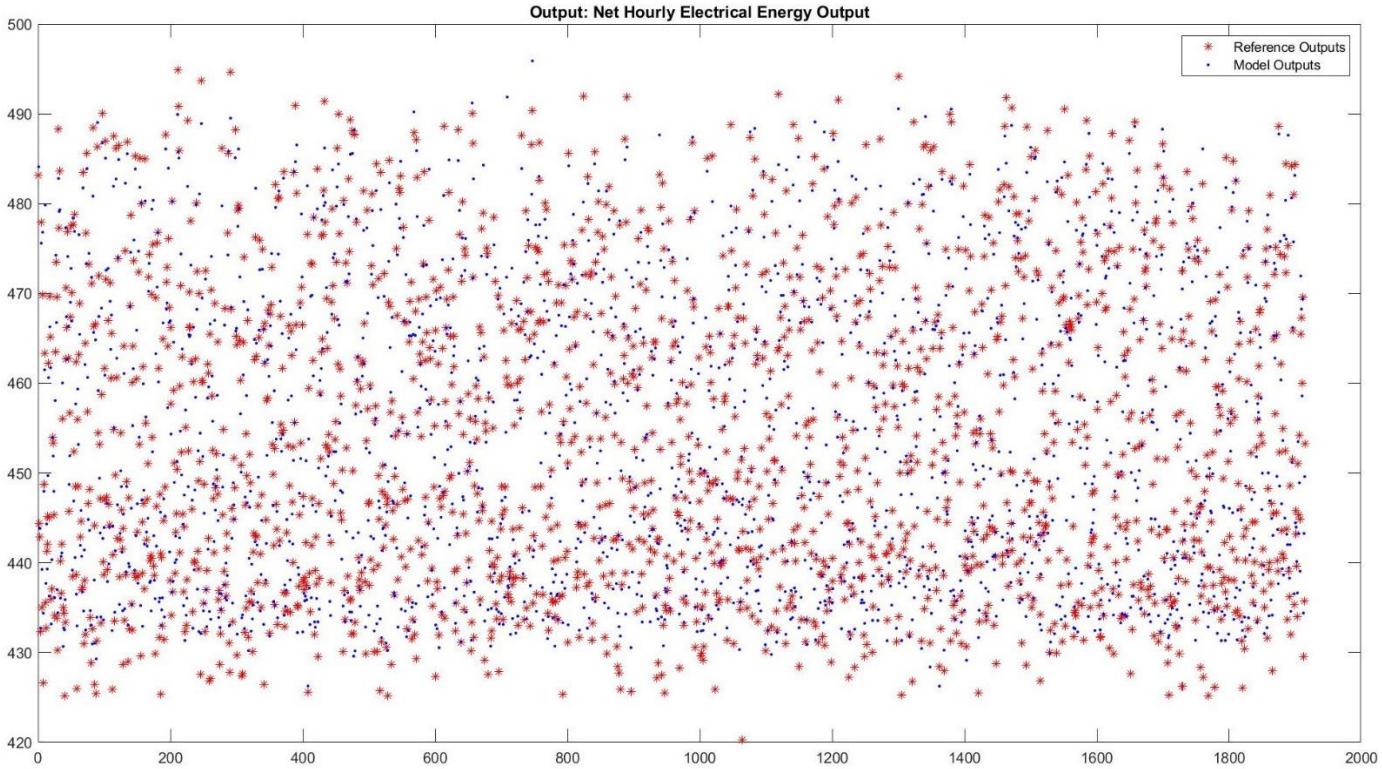


Σχήμα 14: Καμπύλες Εκμάθησης - TSK Μοντέλο 3

Τέλος, βλέπουμε τα σφάλματα πρόβλεψης και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.



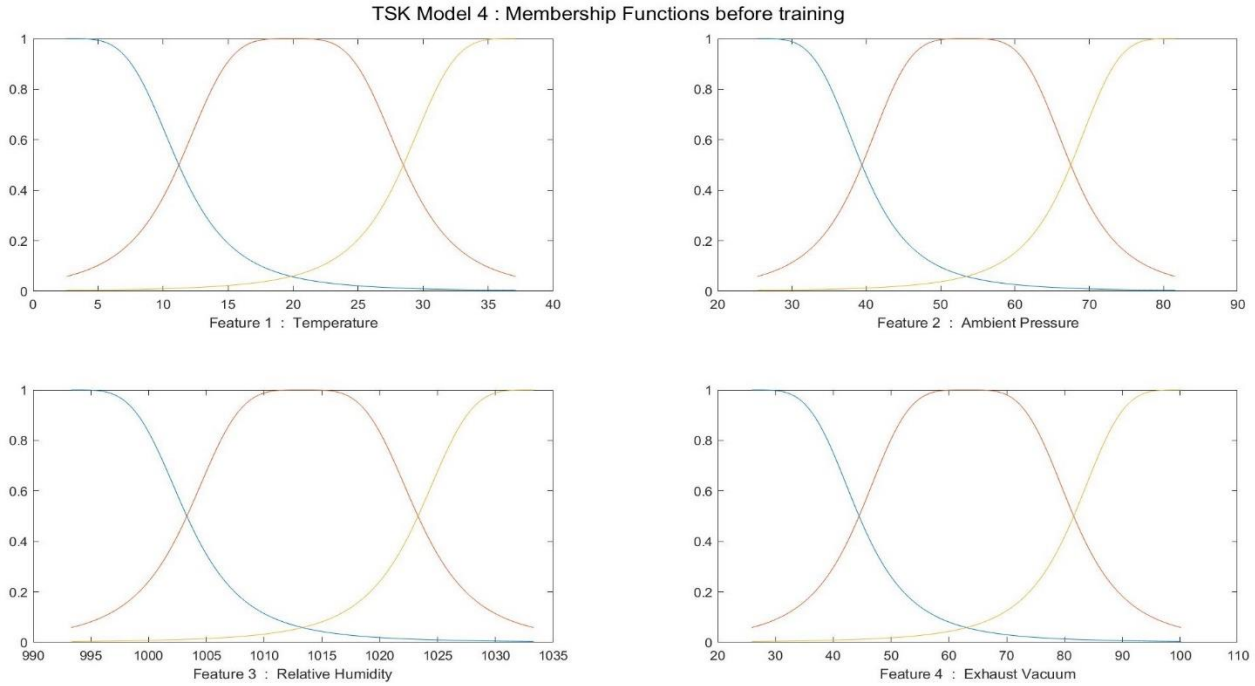
Σχήμα 15: Σφάλματα Πρόβλεψης - TSK Μοντέλο 3



Σχήμα 16: Πραγματική και Εκτιμήτρια Έξοδος - TSK Μοντέλο 3

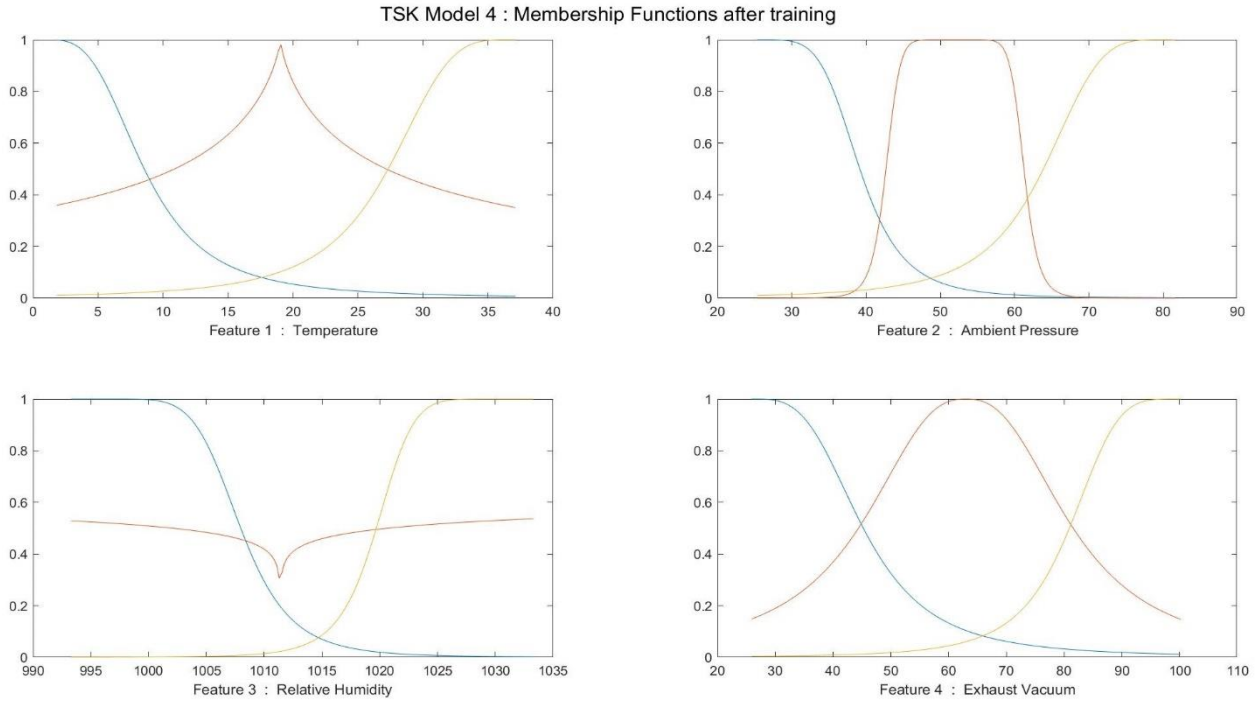
TSK Μοντέλο 4

Στο τέταρτο μοντέλο TSK χρησιμοποιούμε 3 συναρτήσεις συμμετοχής τύπου Bell-Shaped με επικάλυψη 0.5 για κάθε μεταβλητή εισόδου ενώ η μορφή της εξόδου είναι Polynomial (Linear). Οι συναρτήσεις αυτές πριν τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 17.



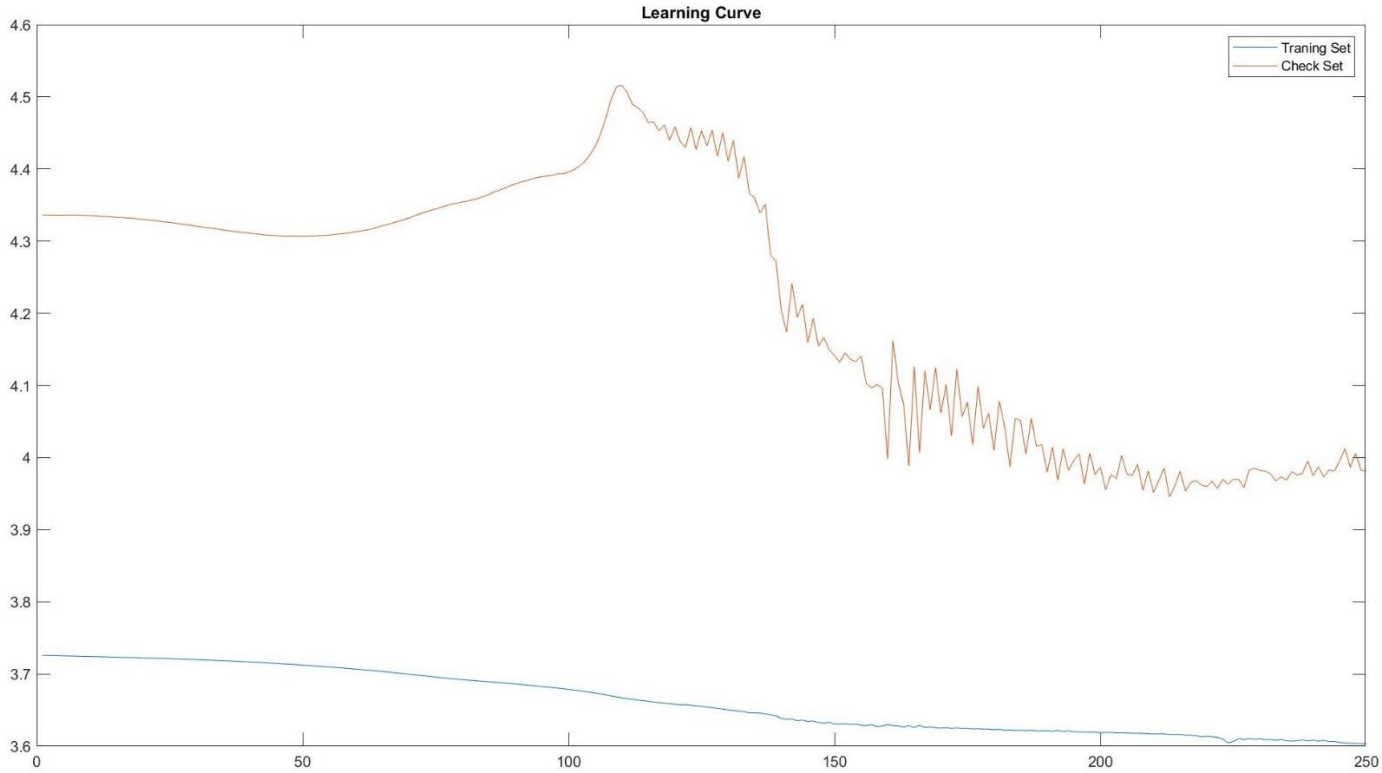
Σχήμα 17: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 4

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



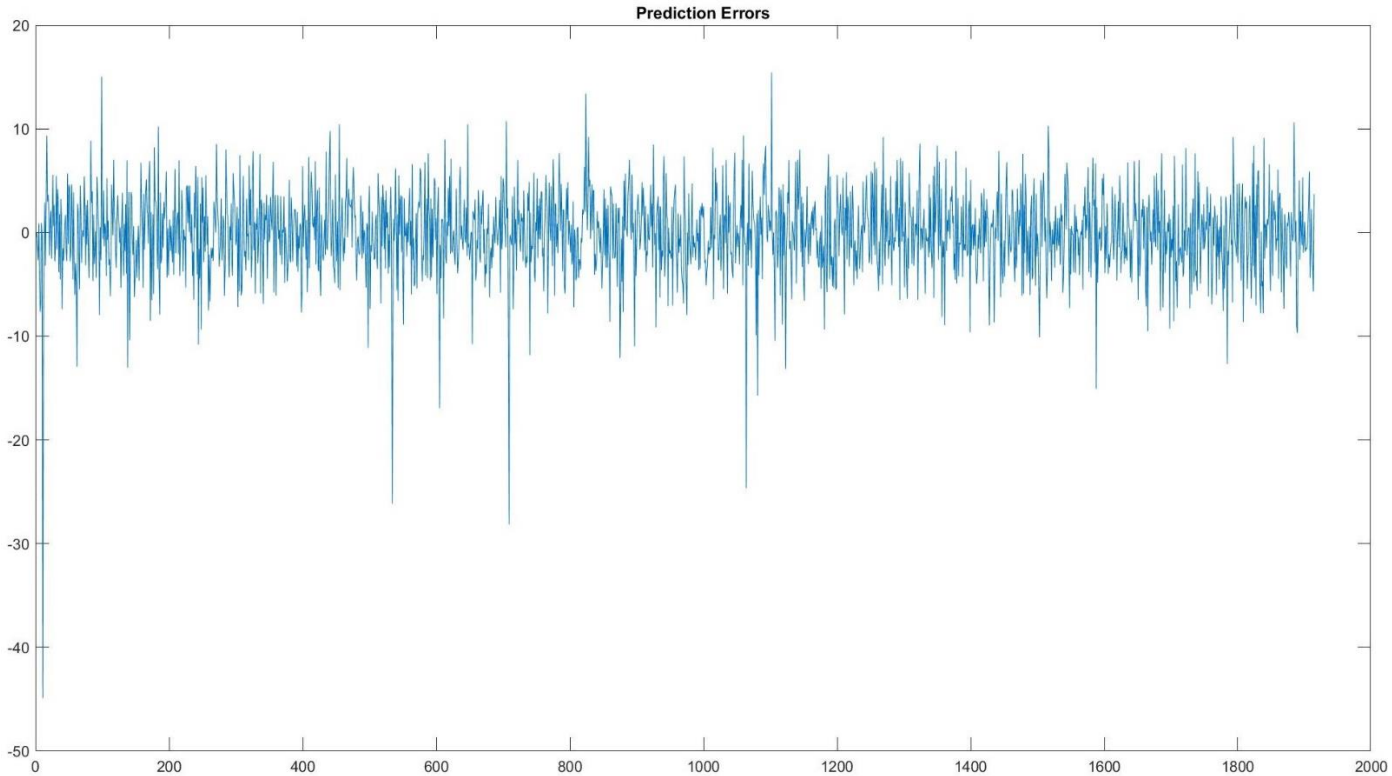
Σχήμα 18: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 4

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

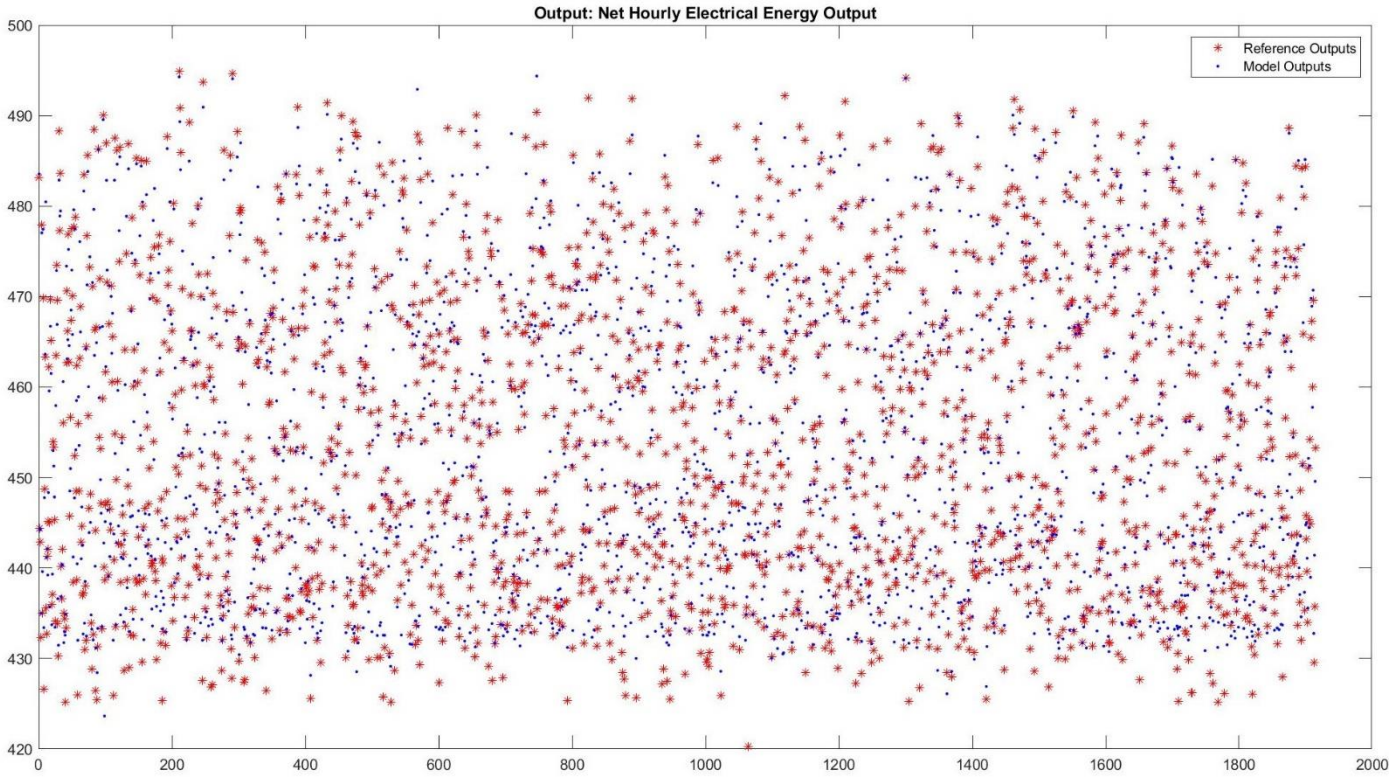


Σχήμα 19: Καμπύλες Εκμάθησης - TSK Μοντέλο 4

Τέλος, βλέπουμε τα σφάλματα πρόβλεψης και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.



Σχήμα 20: Σφάλματα Πρόβλεψης - TSK Μοντέλο 4



Σχήμα 21: Πραγματική και Εκτιμήτρια Έξοδος - TSK Μοντέλο 4

Μετρικές Σφάλματος και Χρόνοι Εκτέλεσης

Στον παρακάτω πίνακα βλέπουμε τις μετρικές σφαλμάτων και το χρόνο εκτέλεσης για τα τέσσερα μοντέλα.

TSK Model	*Number of Input MF*	*Output Format*	*MSE*	*RMSE*	R^2	*NMSE*	*NDEI*	*Elapsed Time*
Model 1	2	Singleton	18.663	4.32	0.935	0.0646	0.2542	26.017 sec
Model 2	3	Singleton	17.033	4.127	0.941	0.059	0.2429	160.3 sec
Model 3	2	Polynomial	17.641	4.2	0.939	0.0611	0.2472	70.029 sec
Model 4	3	Polynomial	16.491	4.061	0.943	0.0571	0.239	1748.639 sec

Σχήμα 22: Πίνακας Μετρικών Σφάλματος – Χρόνου Εκτέλεσης

Με βάση τις παραπάνω μετρικές σφάλματος παρατηρούμε ότι και τα τέσσερα μοντέλα παρουσιάζουν παρόμοιο σφάλμα μεταξύ τους αναφορικά με την εκτίμηση που κάνουν. Για το μοντέλο με τις τρεις συναρτήσεις συμμετοχής και πολυωνυμική μορφή εξόδου (Μοντέλο 4), το μέσο τετραγωνικό σφάλμα (MSE) είναι μικρότερο και ο συντελεστής προσδιορισμού (R^2) είναι πιο κοντά στη μονάδα σε σχέση με τα υπόλοιπα μοντέλα. Αυτό έχει ως αποτέλεσμα να είναι το βέλτιστο εκ των τεσσάρων με την έννοια ότι η

εκτιμήτρια έξοδος που παράγει βρίσκεται πιο κοντά στην πραγματική τιμή της εξόδου. Παρόλα αυτά είναι αρκετά πιο πολύπλοκο, καθώς ο χρόνος εκτέλεσης του είναι κατά πολύ μεγαλύτερος από το υπόλοιπα μοντέλα. Η μεγάλη αυτή διάρκεια οφείλεται στην επιλογή της μεθόδου, Grid Partitioning και στο πλήθος των εισόδων - συναρτήσεων συμμετοχής αφού με τη μέθοδο αυτή ο χρόνος εκτέλεσης αυξάνει εκθετικά με την αύξηση του πλήθους των εισόδων. Επίσης, παρατηρούμε ότι η καμπύλη Εκμάθησης που αφορά τα δεδομένα ελέγχου παρουσιάζει σχετικά έντονες διακυμάνσεις, καθώς επίσης και μια μικρή ανοδική πορεία της καμπύλης, προς το τέλος της εκπαίδευσης, κάτι που μπορεί να υποδεικνύει ότι συνεχίζοντας είναι πιθανό να οδηγηθούμε σε υπερεκπαίδευση. Βέβαια η χρήση του validation set συνιστά στη συνεχή εκπαίδευση χωρίς το μοντέλο να φτάνει σε υπερεκπέδευση.

Γενικότερα, η χρήση μεγαλύτερου αριθμού συναρτήσεων συμμετοχής βελτιώνει τα αποτελέσματα ιδιαίτερα όταν συνδυάζεται με γραμμική πολυωνυμική έξοδο, όπως είναι αναμενόμενο, καθώς δίνει τη δυνατότητα να χρησιμοποιούνται πιο ακριβή αποτελέσματα στην έξοδο του μοντέλου. Ωστόσο, η χρήση εξόδου Singleton μειώνει σημαντικά το χρόνο εκπαίδευσης, αλλά επιφέρει το κόστος της λιγότερο ακριβούς εκτίμησης.

Τέλος, πρέπει να σημειώσουμε ότι τα κανονικοποιημένα σφάλματα NMSE και NDEI δίνουν μια καλύτερη εικόνα για την επίδοση των μοντέλων μας σε σχέση με το MSE, καθώς μας δίνουν τη δυνατότητα να απαλείψουμε την επίδραση της μέσης τιμής και της μεταβλητότητας της υπό μοντελοποίηση διαδικασίας.

Εφαρμογή στο Σετ Δεδομένων Superconductivity

Αντιμετώπιση σετ δεδομένων υψηλής διαστασιμότητας

Το Superconductivity Dataset είναι ένα πολύ μεγαλύτερο σετ δεδομένων σε σχέση με το CCPP, καθώς περιέχει 81 διαφορετικά χαρακτηριστικά σχετικά με υπεραγώγιμα υλικά. Στόχος του τμήματος της εργασίας αυτού είναι η πρόβλεψη της κρίσιμης θερμοκρασίας με βάση τα χαρακτηριστικά αυτά. Ο μεγάλος όγκος των δεδομένων καθιστά τη χρήση της μεθόδου Grid Partition για εκπαίδευση του ζητούμενου μοντέλου πρακτικά ανέφικτη, καθώς ο χρόνος που απαιτείται είναι υπερβολικά μεγάλος. Για το λόγο αυτό, θα χρειαστεί να επιλέξουμε ένα αρκετά πιο περιορισμένο πλήθος χαρακτηριστικών, και συγκεκριμένα τα πιο αντιπροσωπευτικά του δείγματος, η επιλογή των οποίων γίνεται με χρήση του αλγορίθμου Relief.

Εύρεση βέλτιστου πλήθους Χαρακτηριστικών και Κανόνων

Αρχικά, αναδιατάσσουμε τη σειρά των δεδομένων του Dataset ώστε να υπάρχει τυχαιότητα στην σειρά με την οποία εμφανίζονται τα δεδομένα. Έπειτα διαχωρίζουμε το σετ δεδομένων ως εξής:

1. 60% : Σετ εκπαίδευσης – training set
2. 20% : Σετ επικύρωσης – validation set
3. 20% : Σετ ελέγχου – check set

Στη συνέχεια εφαρμόζουμε τον αλγόριθμο Relief επιλέγοντας ως αριθμό γειτόνων το 100 ώστε να γίνει εκτίμηση των σημαντικότερων χαρακτηριστικών με τη σειρά που εμφανίζονται στον πίνακα ranks.

Θα χρησιμοποιήσουμε το συνδυασμό των μεθόδων Grid Search και 5-Fold Cross Validation ώστε να βρούμε το μοντέλο που εκτιμάει καλύτερα την επιθυμητή έξοδο. Συγκεκριμένα η μέθοδος k-Fold Cross Validation, με τιμή k=5, αποτελείται από τα εξής βήματα:

1. Αρχικά, διαχωρίζουμε το set δεδομένων εκπαίδευσης σε δύο νέα τμήματα, ένα νέο set δεδομένων εκπαίδευσης (80% του αρχικού set εκπαίδευσης) και ένα νέο set δεδομένων επικύρωσης (20% του αρχικού set εκπαίδευσης). Για το κάνουμε αυτό αναδιατάσσουμε τα 5 folds δεδομένων κάθε φορά ως 4/5 folds για set εκπαίδευσης και 1/5 folds για set ελέγχου με όλους τους δυνατούς τρόπους δημιουργώντας τελικά πέντε νέα δευτερεύοντα μοντέλα.

2. Εκπαιδεύουμε καθένα από αυτά τα δευτερεύοντα μοντέλα και στη συνέχεια υπολογίζουμε το σφάλμα του καθενός ως το μέσο τετραγωνικό σφάλμα MSE.

3. Τέλος, υπολογίζουμε τη μέση τιμή των προηγουμένως υπολογισμένων σφαλμάτων, η οποία αποτελεί αντιπροσωπευτικό δείγμα του πραγματικού σφάλματος για το συνολικό κύριο μοντέλο.

Η παραπάνω διαδικασία συνδυάζεται με τη μέθοδο Grid Search, δηλαδή εκτελείται μια επαναληπτική διαδικασία στην οποία εφαρμόζεται συνεχώς η μέθοδος 5-Fold Cross Validation για διάφορα κύρια μοντέλα μεταβάλλοντας κάθε φορά τόσο το πλήθος των IF THEN κανόνων όσο και το πλήθος χαρακτηριστικών που λαμβάνονται υπόψιν. Έπειτα συγκεντρώνονται όλα τα μέσα σφάλματα, που υπολογίζονται όπως αναφέρθηκε προηγουμένως για κάθε κύριο μοντέλο, και επιλέγεται το βέλτιστο μοντέλο ως αυτό που παρουσιάζει το ελάχιστο μέσο σφάλμα.

Για την ομαδοποίηση και τη δημιουργία των IF THEN κανόνων χρησιμοποιείται η μέθοδος Fuzzy C-Means (FCM) ενώ οι διάφορες περιπτώσεις των μοντέλων που διερευνώνται αποτελούνται από τους συνδυασμούς πλήθους χαρακτηριστικών και IF THEN κανόνων όπως προκύπτουν από το καρτεσιανό γινόμενο των συνόλων αντίστοιχα,

$$NF \times NR = \{5, 10, 15, 20\} \times \{4, 8, 12, 16, 20\}$$

Συνεπώς εξετάζεται η απόδοση 20 διαφορετικών κύριων μοντέλων του αρχικού σετ εκπαίδευσης με βάση τα 5 δευτερεύοντα μοντέλα στα οποία διακρίνεται το καθένα από αυτά (Μέθοδος 5-Fold Validation). Για κάθε κύριο μοντέλο πραγματοποιείται εκπαίδευση, καθενός από τα πέντε δευτερεύοντα μοντέλα του (συνολικά 100 μοντέλα) για 150 εποχές το καθένα, και υπολογίζεται το σφάλμα καθενός από αυτά. Τέλος, υπολογίζεται ο μέσος όρος των 5 σφαλμάτων ο οποίος αποτελεί το κριτήριο για την εύρεση του βέλτιστου από τα κύρια μοντέλα.

Ωστόσο, η εκτέλεση του αλγορίθμου για τις συγκεκριμένες τιμές του συνόλου κανόνων και χαρακτηριστικών και ιδιαίτερα για τις υψηλότερες τιμές των 20 χαρακτηριστικών και κανόνων εκτός από ιδιαίτερα χρονοβόρα ήταν και αποτυχημένη. Το σφάλμα Check Error που επέστρεψε η ANFIS, μετά το πέρας των 19 ωρών και 26 λεπτών προσομοίωσης, αποτελούνταν από μεγάλο πλήθος τιμών NaN. Για την επίλυση του παραπάνω προβλήματος μειώνουμε τις τιμές των κανόνων και χαρακτηριστικών και επαναλαμβάνουμε την εκτέλεση του αλγορίθμου. Οι νέες τιμές είναι οι εξής:

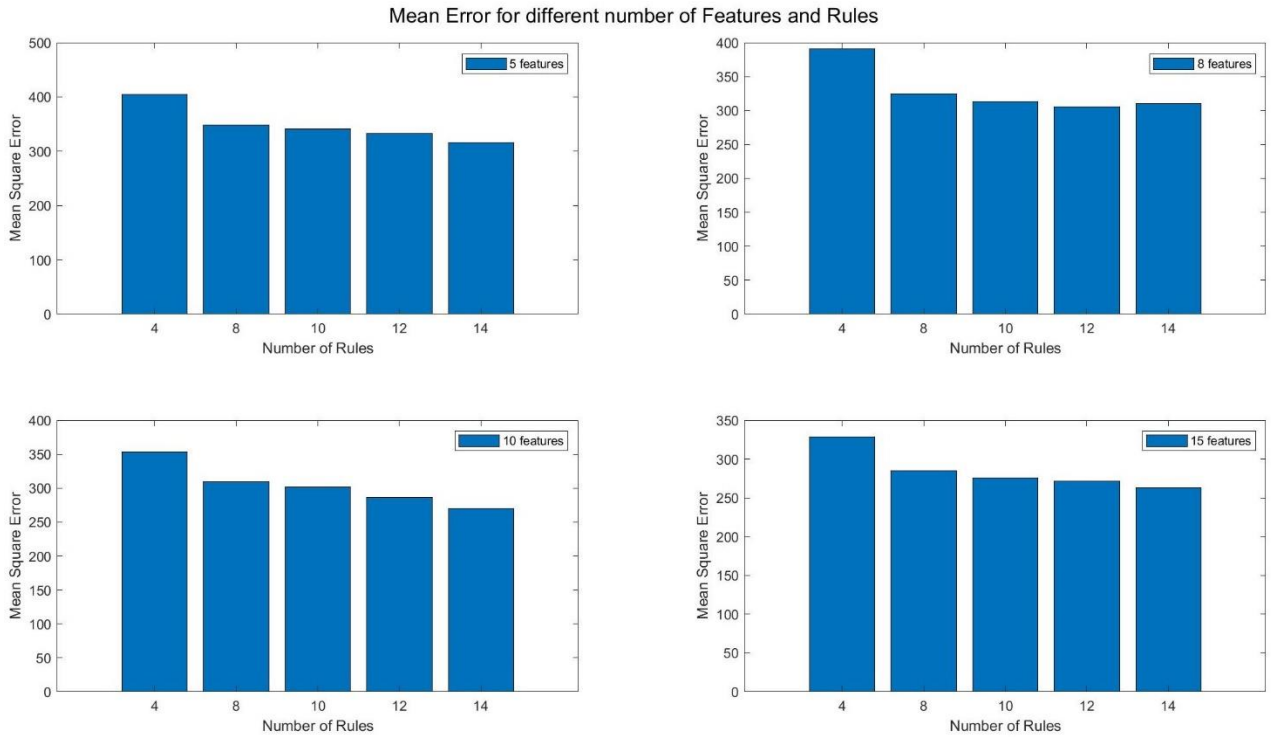
$$NF \times NR = \{5, 8, 10, 15\} \times \{4, 8, 10, 12, 14\}$$

Στον παρακάτω πίνακα παρουσιάζεται το μέσο σφάλμα των MSE των 5 δευτερευόντων μοντέλων για καθένα από τα 20 διαφορετικά κύρια μοντέλα.

Number of Rules / Number of Features	4 Rules	8 Rules	10 Rules	12 Rules	14 Rules
5 Features	404.3285	348.0685	340.5908	332.6154	316.4561
8 Features	391.2954	324.2558	313.2237	305.7468	310.6164
10 Features	353.7305	309.3503	301.8013	286.2236	269.9682
15 Features	328.4132	285.1113	275.6476	271.5057	263.4929

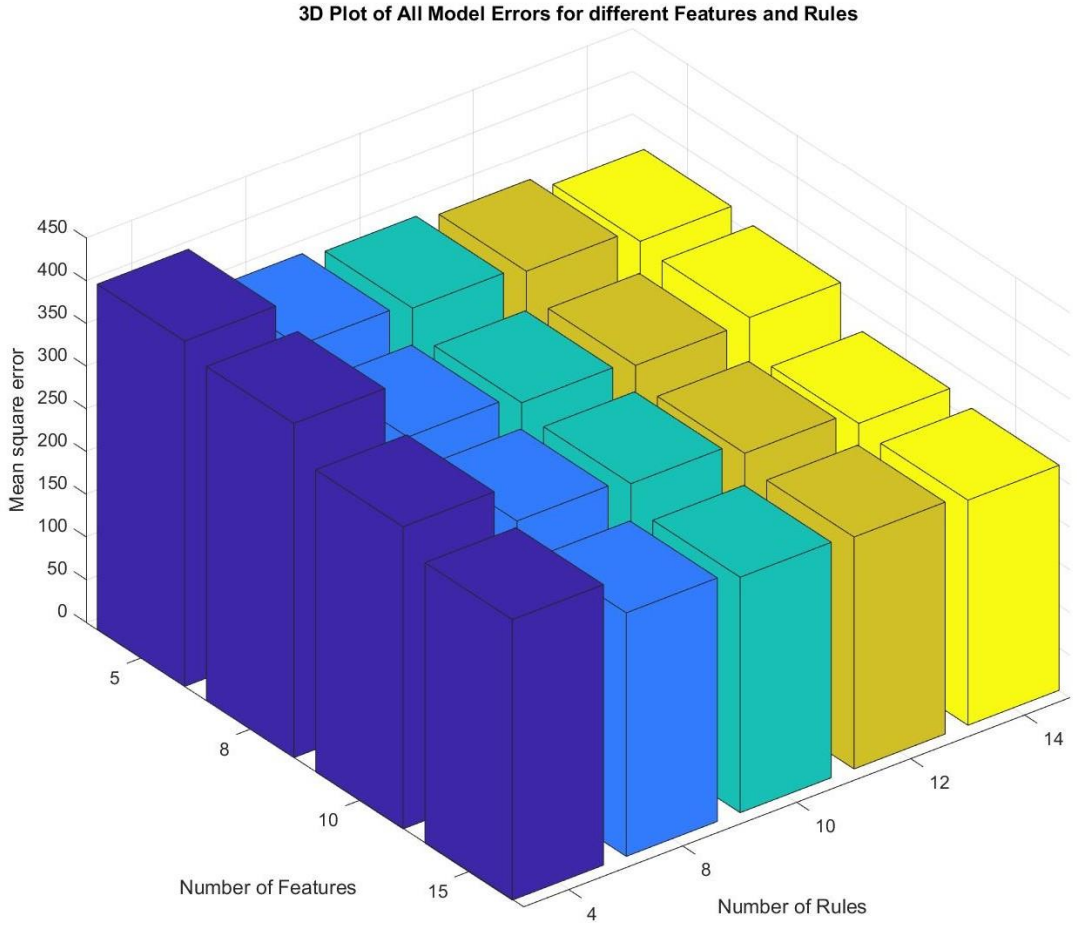
Σχήμα 23: Πίνακας Μέσου Σφάλματος για τα διάφορα μοντέλα

Στα παρακάτω διαγράμματα φαίνονται γραφικά οι τιμές του μέσου σφάλματος για τις διάφορες τιμές χαρακτηριστικών και κανόνων.



Σχήμα 24: Μέσο σφάλμα μοντέλων για τις διάφορες τιμές πλήθους χαρακτηριστικών και κανόνων

Τέλος, τα παραπάνω σφάλματα παρουσιάζονται και σε ένα κοινό διάγραμμα τριών διαστάσεων.

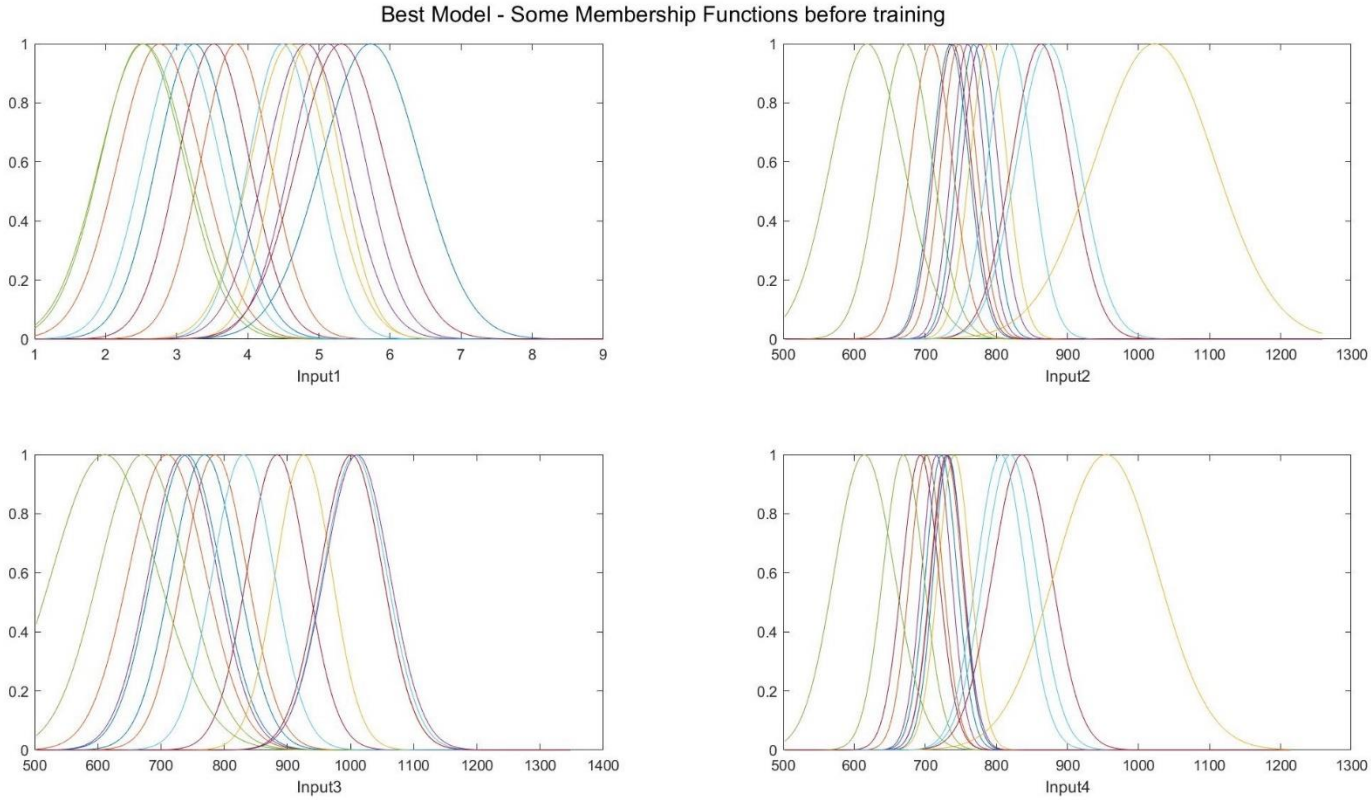


Σχήμα 25: Κοινό 3D Διάγραμμα Μέσου Σφάλματος των διάφορων μοντέλων

Από τα παραπάνω είναι εμφανές ότι το βέλτιστο, από τα εξεταστέα μοντέλα, είναι αυτό με τα **15 χαρακτηριστικά** και τους **14 κανόνες**. Παρατηρούμε ότι όσο αυξάνεται η πολυπλοκότητα του μοντέλου (πλήθος χαρακτηριστικών και κανόνων) τόσο αυξάνεται και ο χρόνος εκτέλεσης του αλγορίθμου, ωστόσο δεν βελτιώνεται έντονα η ικανότητα εκτίμησης του μοντέλου. Επίσης, παρατηρήθηκε, όπως ήταν αναμενόμενο, ότι η εκπαίδευση των τελευταίων σε σειρά μοντέλων (δηλαδή αυτών με το μεγαλύτερο πλήθος χαρακτηριστικών και κανόνων) διαρκούσε κατά πολύ περισσότερο σε σχέση με αυτήν των πρώτων μοντέλων. Ο συνολικός χρόνος εκτέλεσης του αλγορίθμου με τις νέες τιμές κανόνων και χαρακτηριστικών ήταν περίπου 7 ώρες και 26 λεπτά (26753.563 δευτερόλεπτα).

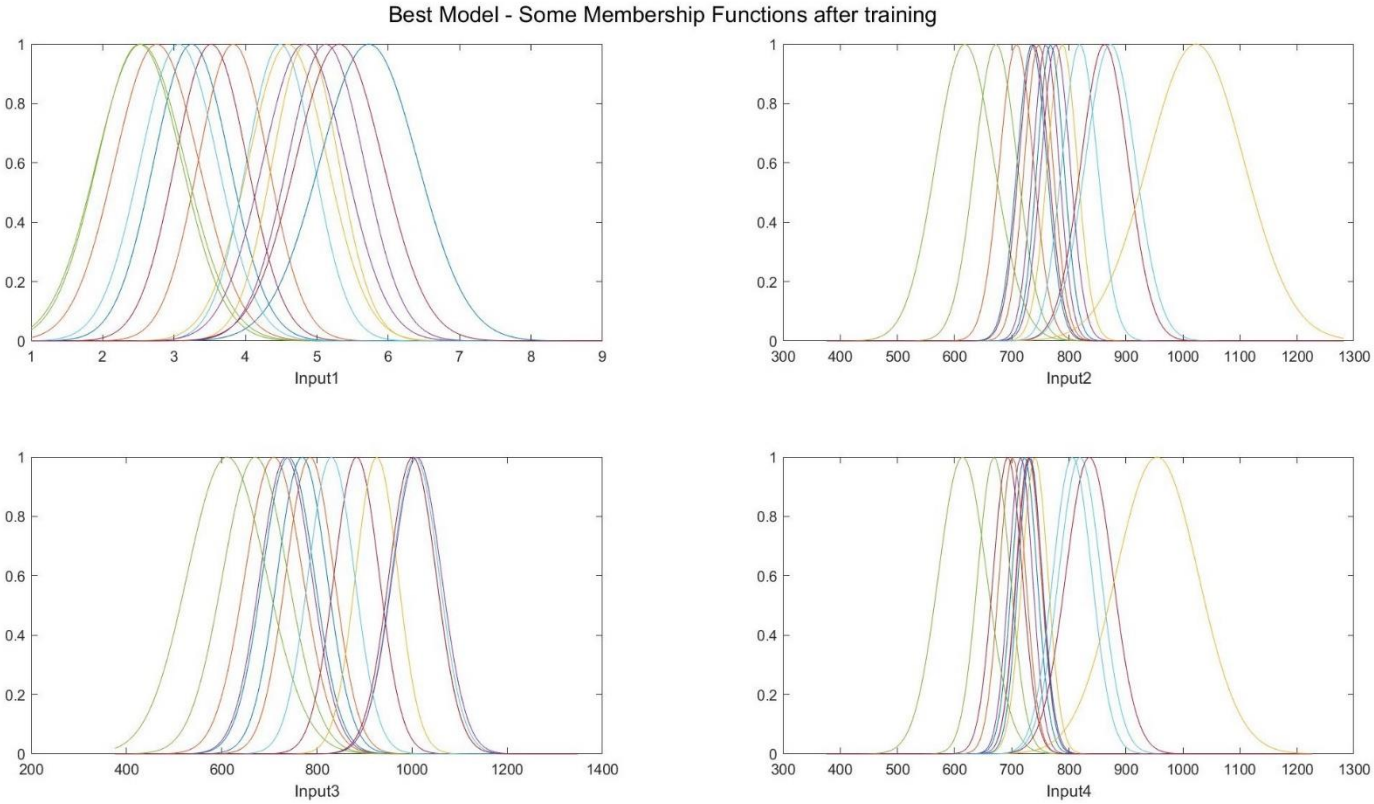
Εκπαίδευση βέλτιστου TSK μοντέλου

Αρχικά παρουσιάζουμε ορισμένες από τις συναρτήσεις συμμετοχής του βέλτιστου μοντέλου πριν την εκπαίδευσή του.



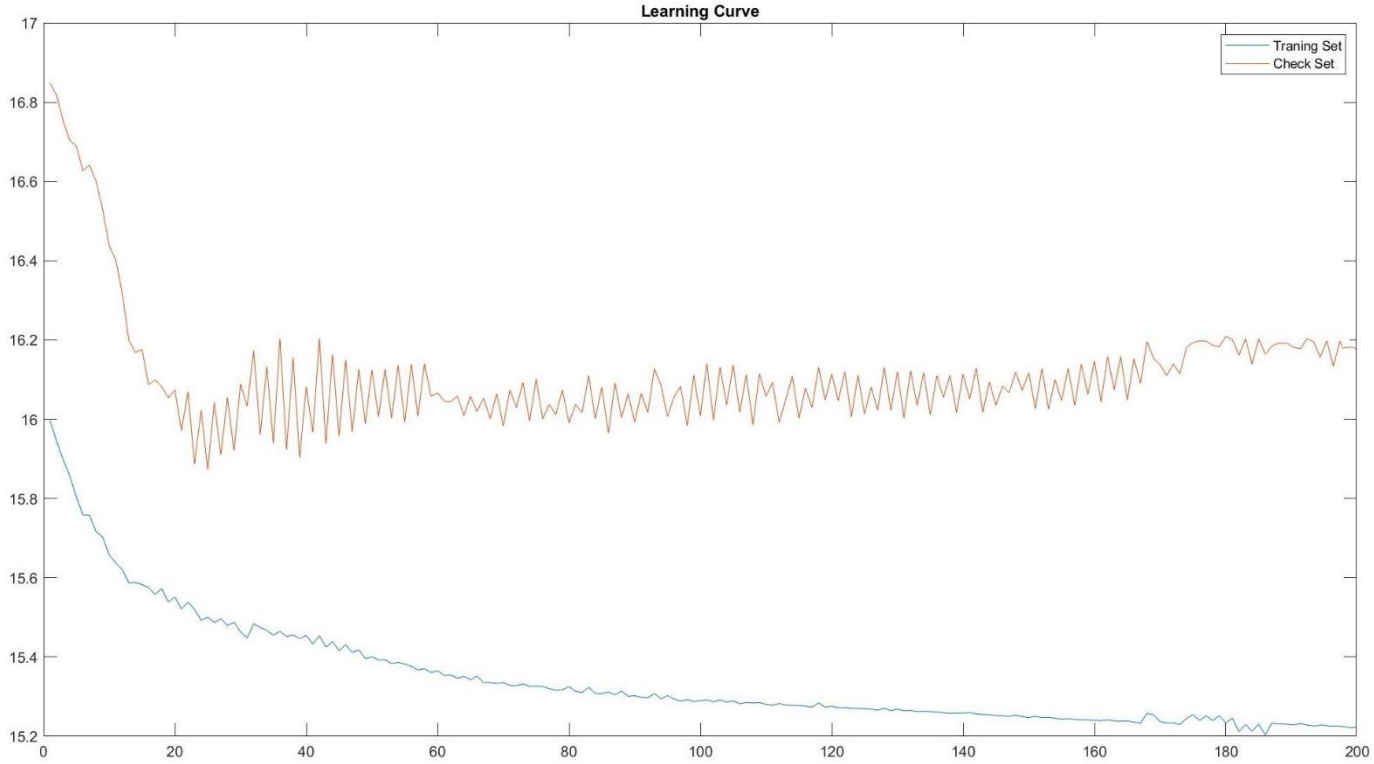
Σχήμα 26: Συναρτήσεις Συμμετοχής πριν την εκπαίδευση

Μετά από εκπαίδευση σε 200 εποχές οι παραπάνω συναρτήσεις συμμετοχής λαμβάνουν την παρακάτω μορφή.



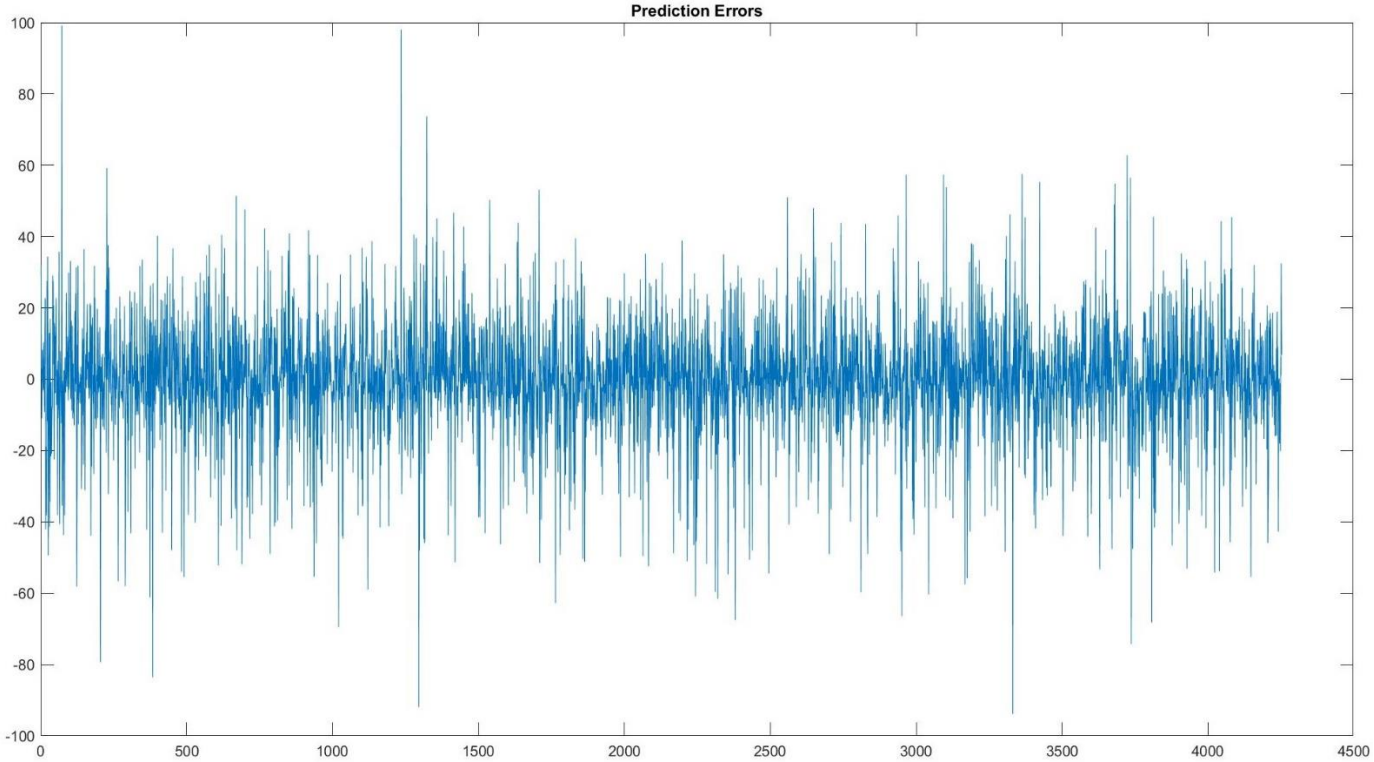
Σχήμα 27: Συναρτήσεις Συμμετοχής μετά την εκπαίδευση

Ακολουθούν οι καμπύλες εκμάθησης με βάση το RMSE στο πέρας των εποχών.



Σχήμα 28: Καμπύλες Εκμάθησης - TSK Μοντέλο

Τέλος, βλέπουμε τα σφάλματα πρόβλεψης μετά το πέρας της εκπαίδευσης.



Σχήμα 29: Σφάλματα Πρόβλεψης - TSK Μοντέλο

Μετρικές Σφάλματος και Χρόνος Εκτέλεσης

Στον παρακάτω πίνακα βλέπουμε τις μετρικές σφαλμάτων και το χρόνο εκτέλεσης για την εκπαίδευση και αξιολόγηση του βέλτιστου μοντέλου.

TSK Model	*Number of Features*	*Number of Rules*	*MSE*	*RMSE*	R^2	*NMSE*	*NDEI*	*Elapsed Time*
Optimum Model	15	14	259.758	16.117	0.7778	0.222	0.471	997.299 sec

Σχήμα 30: Πίνακας Μετρικών Σφάλματος – Χρόνου Εκτέλεσης

Το αποτέλεσμα που προέκυψε είναι ικανοποιητικό δεδομένης της μεγάλης έκτασης του Dataset όπως γίνεται εμφανές από το δείκτη R^2 ο οποίος είναι σχετικά κοντά στη μονάδα, πράγμα που σημαίνει ότι οι εκτιμώμενες έξοδοι του μοντέλου είναι ικανοποιητικά κοντά στις αντίστοιχες πραγματικές τιμές. Κάτι τέτοιο γίνεται επίσης εμφανές και από τα μικρά κανονικοποιημένα σφάλματα NMSE, NDEI.

Στην περίπτωση που είχαμε χρησιμοποιήσει τη μέθοδο Grid Partitioning, ο αντίστοιχος συνολικός χρόνος εκτέλεσης θα ήταν απαγορευτικά μεγαλύτερος του παρόντος. Είναι γνωστό ότι ο χρόνος εκτέλεσης μεγαλώνει εκθετικά με την αύξηση του πλήθους των χαρακτηριστικών αλλά και με την αύξηση των ασαφών συνόλων εισόδου όπως γίνεται εμφανές και από το πρώτο μέρος της παρούσας εργασίας. Ιδιαίτερα στην περίπτωση των τριών ασαφών συνόλων εισόδου και των 14 χαρακτηριστικών ο συνολικός χρόνος σε ώρες για τη διεκπεραίωση της διαδικασίας εκπαίδευσης και αξιολόγησης θα μπορούσε να χαρακτηριστεί αστρονομικός δεδομένων των $3^{14} \approx 4.783$ εκατομμύρια κανόνων.

Τέλος, παρατηρούμε από το Σχήμα 28 με την καμπύλη εκμάθησης του βέλτιστου μοντέλου μια μικρή αύξηση καθώς και διαταραχές στο σφάλμα ιδιαίτερα προς τις τελευταίες εποχές εκπαίδευσης, πράγμα που σημαίνει ότι είναι καλύτερα να σταματήσει η εκπαίδευση προτού οδηγηθούμε σε υπερεκπαίδευση.

Αρχεία MATLAB

1. ccppTSKModels.m : MATLAB Script – Υλοποίηση πρώτου τμήματος της εργασίας (CCPP Dataset). Ο χρήστης αρχικά επιλέγει τον αριθμό του TSK Μοντέλου (1 έως 4) που θέλει να εκπαιδεύσει, αξιολογήσει και υπολογίσει τις μετρικές σφάλματος.

2. gridSearch.m : MATLAB Script – Υλοποίηση δεύτερου τμήματος της εργασίας (Superconductivity Dataset). Ο χρήστης ενημερώνεται σε ζωντανό χρόνο για την πρόοδο της διαδικασίας εκπαίδευσης των 100 μοντέλων και τις παραμέτρους (πλήθος χαρακτηριστικών, πλήθος κανόνων, αριθμός πτυχής) του μοντέλου που εκπαιδεύεται κάθε φορά. Τέλος, δημιουργείται και ένα αρχείο με όνομα optimum_model.mat, το οποίο περιλαμβάνει τον αριθμό των χαρακτηριστικών και κανόνων του βέλτιστου μοντέλου καθώς και το απαραίτητο τμήμα του πίνακα ranks, που καθορίζει με φθίνουσα σειρά σημασίας ποιες από τις στήλες των χαρακτηριστικών χρησιμοποιήθηκαν.

3. optimumModel.m : MATLAB Script – Εκπαίδευση του βέλτιστου TSK μοντέλου που επιλέγεται με βάση το αρχείο gridSearch.m και υπολογισμός των απαραίτητων μετρικών σφάλματος.