ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

ΕΡΓΑΣΙΑ 3η – TSK REGRESSION

ΘΩΜΑΣ ΚΥΡΙΑΚΟΣ ΠΡΑΒΙΝΟΣ, 9937

1.ΕΙΣΑΓΩΓΗ

Η εργασία στοχεύει στη διερεύνηση της ικανότητας των μοντέλων TSK στην μοντελοποίηση πολυμεταβλητών, μη γραμμικών συναρτήσεων. Είναι διαθέσιμα δύο datasets, το πρώτο εκ των οποίων θα χρησιμοποιηθεί για μια απλή διερεύνηση της διαδικασίας εκπαίδευσης και αξιολόγησης μοντέλων αυτού του είδους, καθώς και για μια επίδειξη τρόπων ανάλυσης και ερμηνείας των αποτελεσμάτων. Το δεύτερο είναι ένα σύνολο δεδομένων υψηλότερης πολυπλοκότητας που θα χρησιμοποιηθεί για μια πληρέστερη διαδικασία μοντελοποίησης, η οποία θα περιλαμβάνει μεταξύ άλλων προ-επεξεργαστικά βήματα όπως επιλογή χαρακτηριστικών (feature selection), καθώς και μεθόδους βελτιστοποίησης των μοντέλων μέσω της διασταυρωμένης επικύρωσης (cross validation).

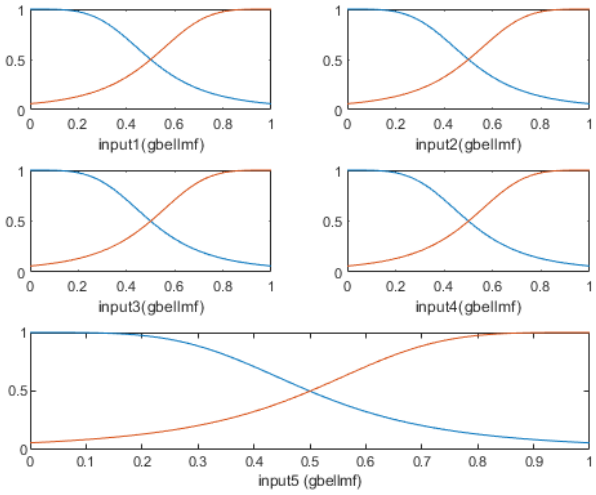
2.ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ Α

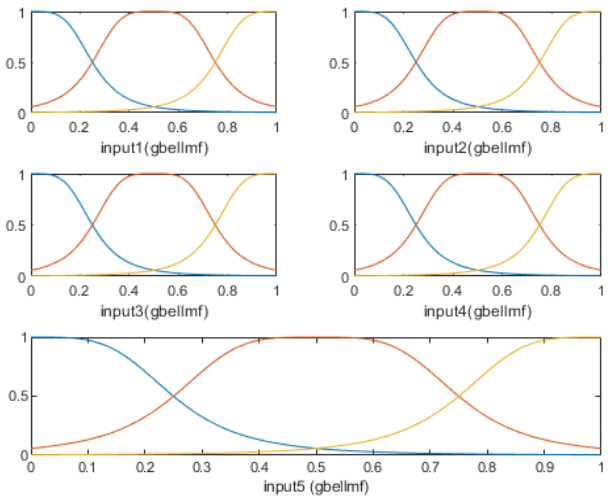
Το πρώτο μέρος της εργασίας απαιτεί την εκπαίδευση τεσσάρων μοντέλων με διαφορετικά χαρακτηριστικά, όπως φαίνεται στον παρακάτω πίνακα. Το σύνολο δεδομένων Airfoil Self-Noise από το αποθετήριο UCI, το οποίο περιέχει 1503 δείγματα και 6 χαρακτηριστικά, το τελευταίο εκ των οποίων είναι το εξαρτημένο χαρακτηριστικό ή χαρακτηριστικό-στόχος. Αρχικά, τα δεδομένα πρέπει να χωριστούν σε τρεις μη επικαλυπτόμενες ομάδες: εκπαίδευση (D_{trn}), επικύρωση (D_{val}) και δοκιμή (D_{test}). Το πρώτο σύνολο, όπως υποδηλώνουν τα ονόματα, χρησιμοποιείται για την εκπαίδευση, το δεύτερο για την αποφυγή της υπερεκπαίδευσης και το τρίτο για τη δοκιμή της απόδοσης του τελικού μοντέλου. Σύμφωνα με τη θεωρία ο διαχωρισμός γίνεται σε 60%, 20%, και 20%, του αρχικού συνόλου δεδομένων, στο κάθε υποσύνολο αντίστοιχα.

	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

Τα τέσσερα μοντέλα εκπαιδεύονται χρησιμοποιώντας ένα υβριδικό μοντέλο, σύμφωνα με την οποία οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται μέσω της μεθόδου της οπισθοδιάδοσης (backpropagation algorithm), ενώ οι παράμετροι της πολυωνυμικής συνάρτησης εξόδου βελτιστοποιούνται μέσω της μεθόδου των ελαχίστων τετραγώνων (Least Squares). Ο τύπος των συναρτήσεων συμμετοχής ορίζεται να είναι bell-shaped και η αρχικοποίησή τους γίνεται με τρόπο ώστε τα διαδοχικά ασαφή σύνολα να παρουσιάζουν, σε κάθε είσοδο, βαθμό επικάλυψης περίπου 0.5

Αρχικοποίηση Συναρτήσεων Συμμετοχής:





Για την ακρίβεια της εκτίμησης της πραγματικής συνάρτησης από καθένα από τα παραπάνω μοντέλα, θα χρησιμοποιηθούν οι εξής δείκτες απόδοσης:

- MSE: Το μέσο τετραγωνικό σφάλμα μεταξύ της εξόδου του μοντέλου και της πραγματικής εξόδου.
- Συντελεστής προσδιορισμού R2: ο συντελεστής αυτός, παρέχει πληροφορία για το ποσοστό της διακύμανσης της πραγματικής εξόδου που ¨εξηγείται¨ από το μοντέλο μας.
- Δείκτης Normalized Mean Square Error
- Δείκτης Non-Dimensional Error Index

(Οι μαθηματικοί τύποι δίνονται στην εκφώνηση της εργασίας)

3.ΕΚΠΑΙΔΕΥΣΗ

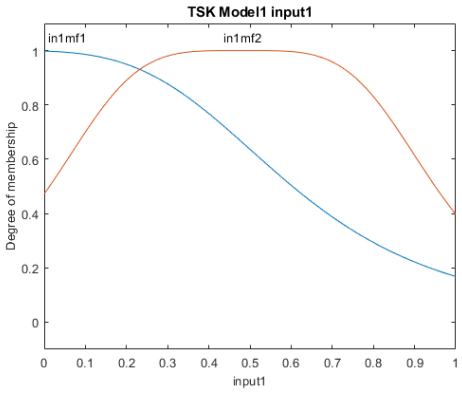
Γίνεται αρχικοποίηση των μοντέλων, σύμφωνα με τις δοθείσες προδιαγραφές και μετά από την ολοκλήρωση της εκπαίδευσης λαμβάνουμε τα εξής :

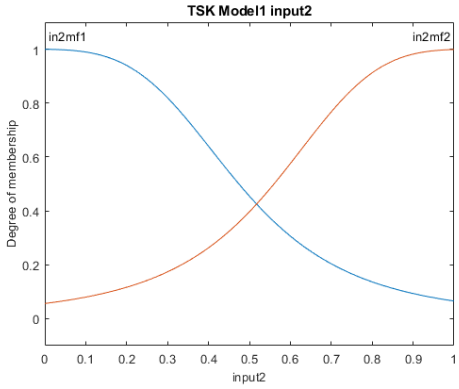
- Τελικές μορφές των ασαφών συνόλων
- Διαγράμματα μάθησης
- Διάγραμμα σφαλμάτων πρόβλεψης
- Τιμές των δεικτών απόδοσης RMSE, NSME, NDEI,R^2

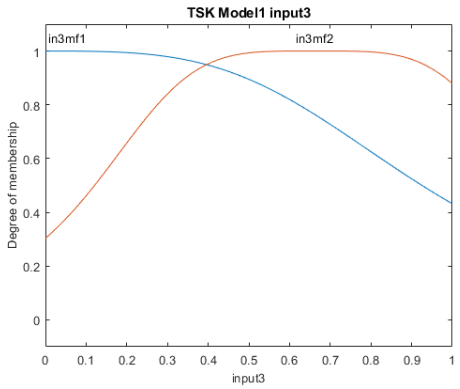
Κάθε μοντέλο εκπαιδεύεται για 100 Epochs.

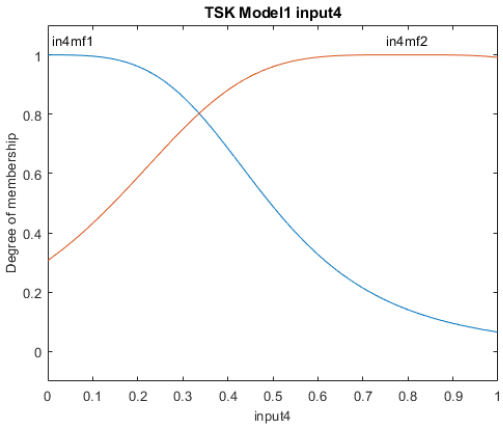
3.1.ΜΟΝΤΕΛΟ 1

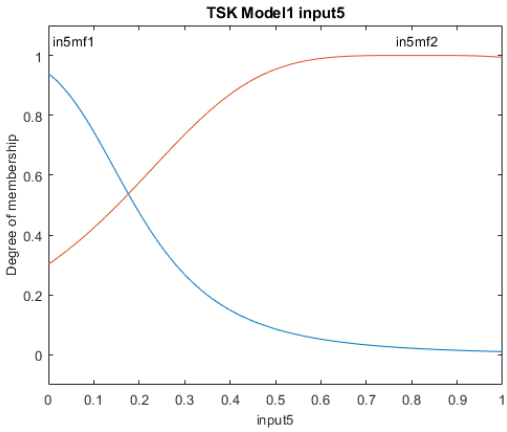
Τελικές μορφές των ασαφών συνόλων



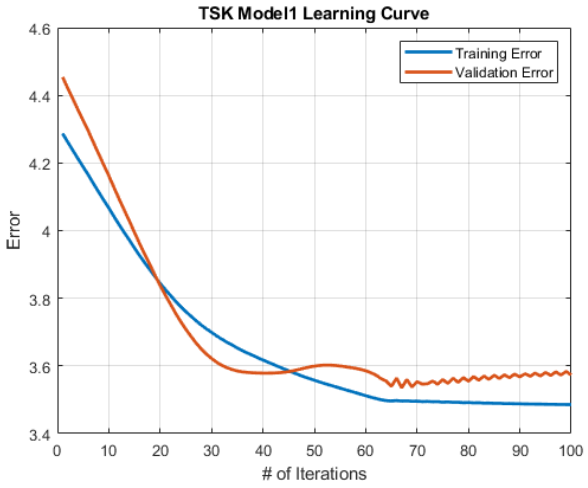




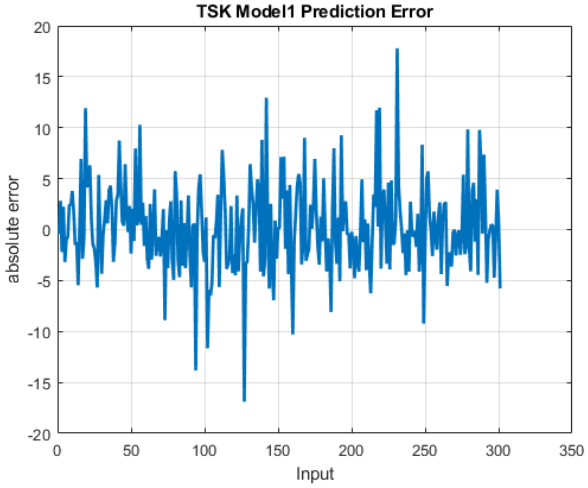




Διάγραμμα μάθησης



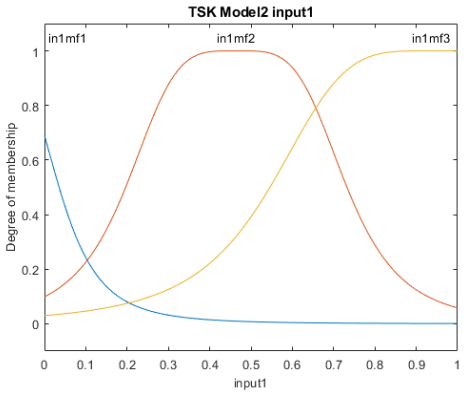
Διάγραμμα σφαλμάτων πρόβλεψης

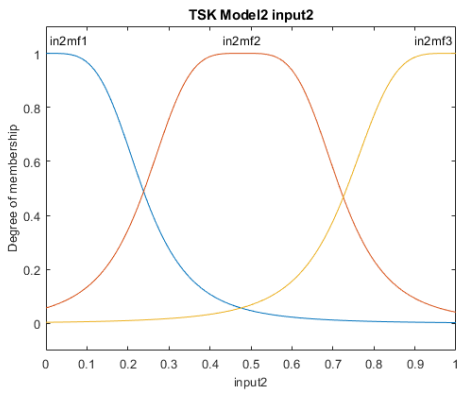


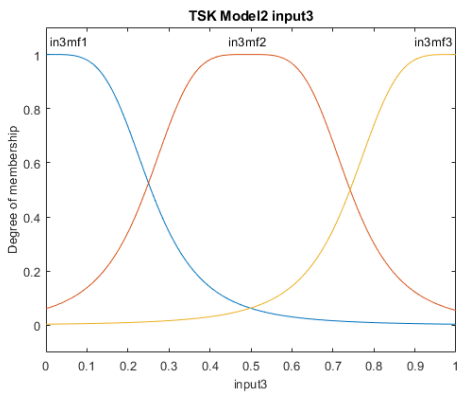
Σύμφωνα με το Διάγραμμα Μάθησης, παρατηρούμε πώς μετά από έναν αριθμό επαναλήψεων, το Validation Error σε σχέση με το Training Error αρχίζει να συγκλίνει αρκετά. Το γεγονός ότι είναι πολύ κοντά τα σφάλματα μεταξύ τους είναι σημάδι ότι ο διαχωρισμός των δεδομένων ήταν αποτελεσματικός.

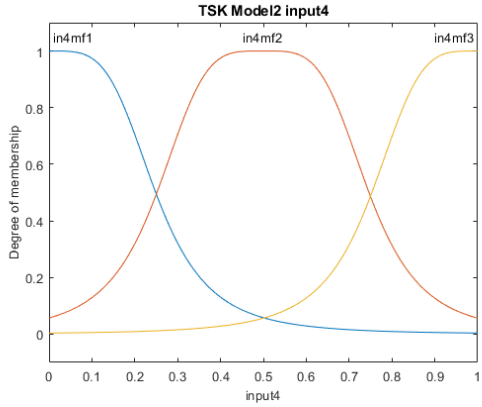
3.2.ΜΟΝΤΕΛΟ 2

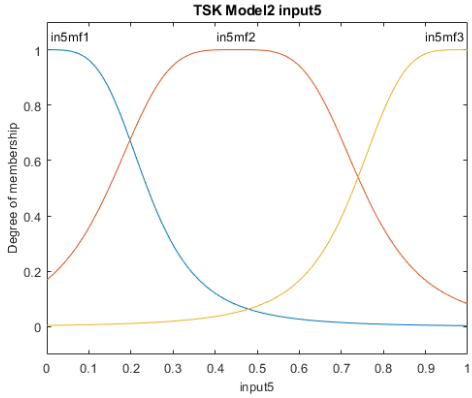
Τελικές μορφές των ασαφών συνόλων



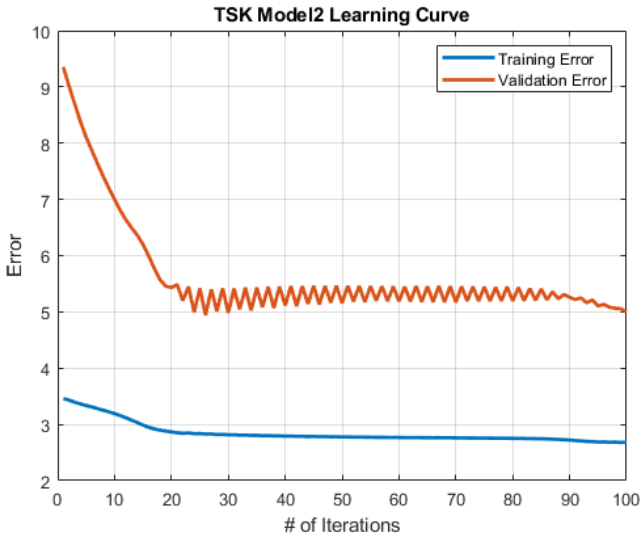




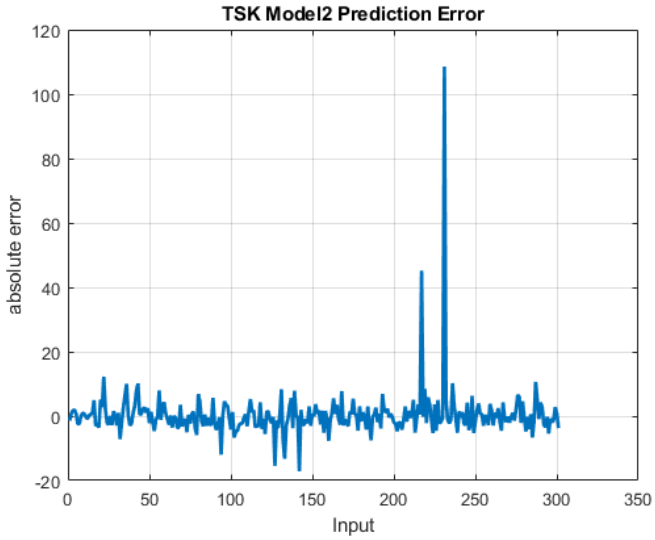




Διάγραμμα μάθησης



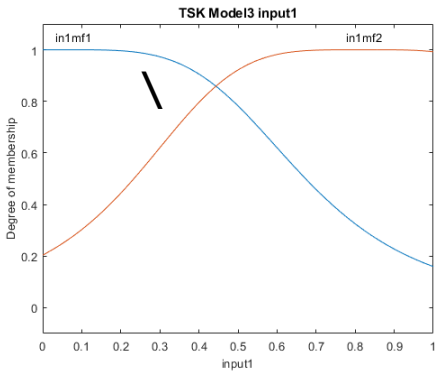
Διάγραμμα σφαλμάτων πρόβλεψης

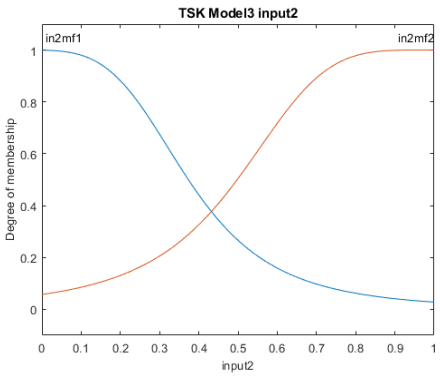


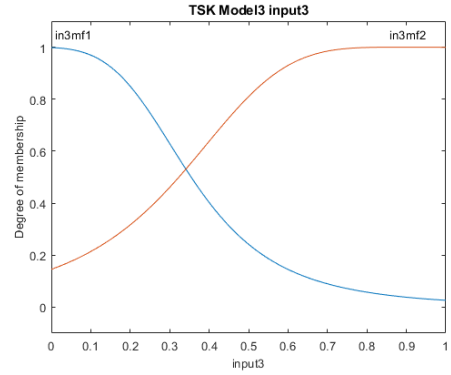
Από το διάγραμμα μάθησης, του δεύτερου μοντέλου, παρατηρούμε, παρά τις ταλαντώσεις, ότι η απόκλιση του Validation Error σε σχέση με το Training Error είναι αρκετά μεγαλύτερη από το Μοντέλο 1, παραμένωντας όμως σταθερη.

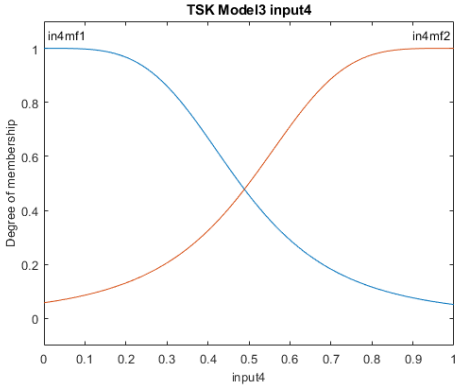
3.3.ΜΟΝΤΕΛΟ 3

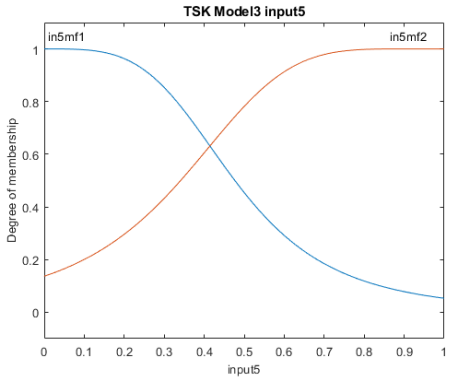
Τελικές μορφές των ασαφών συνόλων



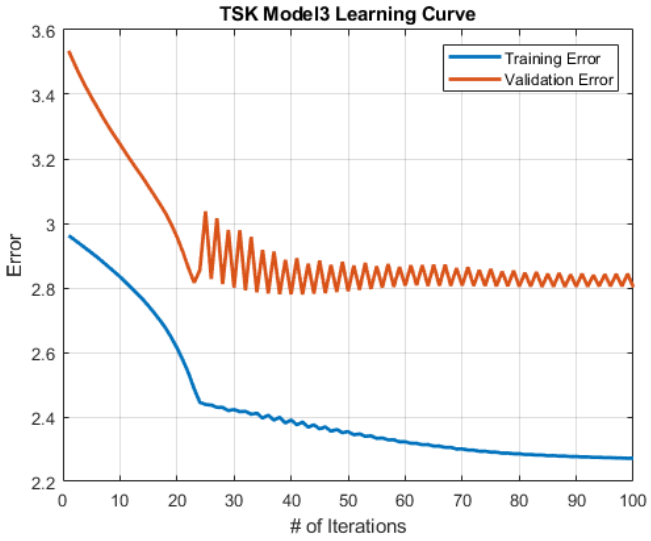




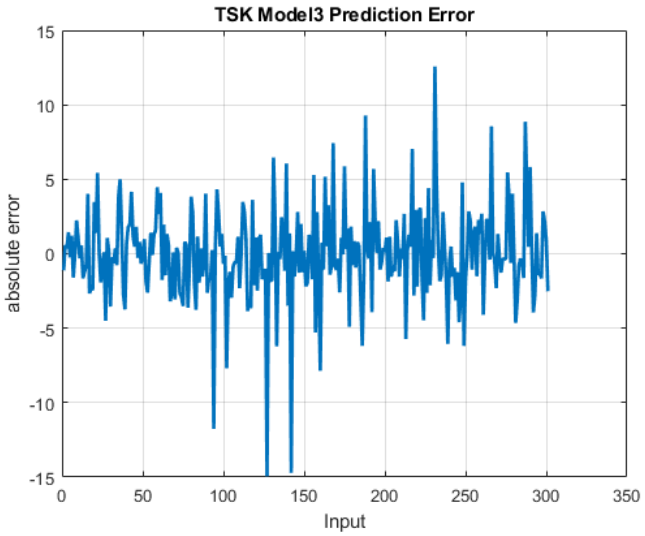




Διάγραμμα μάθησης



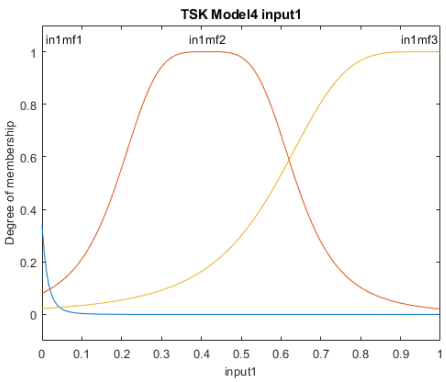
Διάγραμμα σφαλμάτων πρόβλεψης

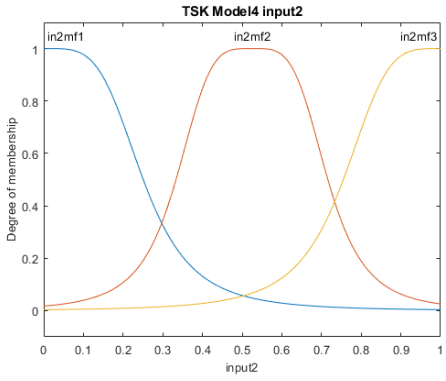


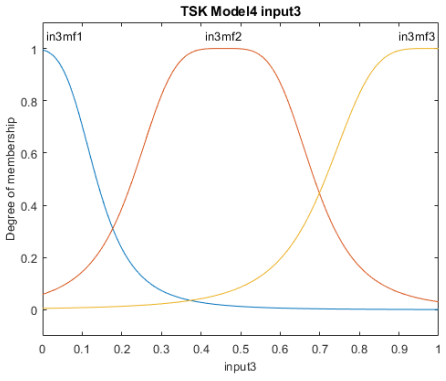
Συγκριτικά με το Μοντέλο 1, που έχει τα ίδια χαρακτηριστικά με μόνη διαφορά τον τύπο της μεταβλητής εξόδου, παρατηρούμε πως το Validation Error συγκλίνει πιο γρήγορα προς μια σταθερή τιμή. Παρόλα αυτά η απόκλιση μεταξύ των σφαλμάτων είναι αρκετή και παρατηρούμε και στο διάγραμμα με τα σφάλματα πρόβλεψης πως η διακύμανση αυτών είναι μεγαλύτερη.

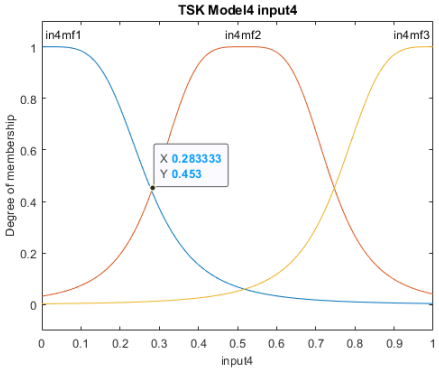
3.4.ΜΟΝΤΕΛΟ 4

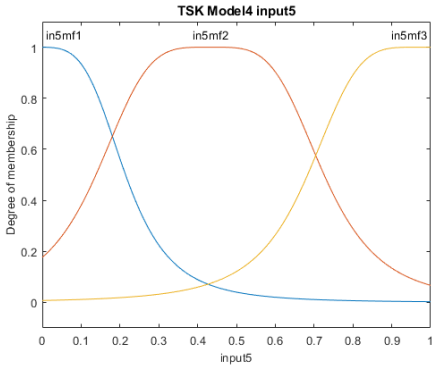
Τελικές μορφές των ασαφών συνόλων



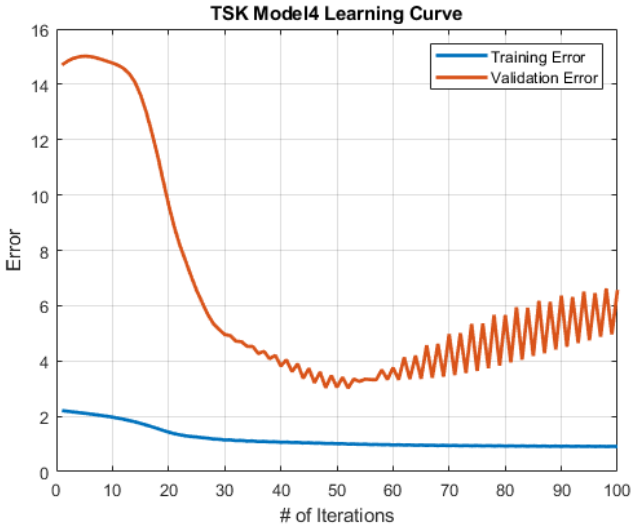




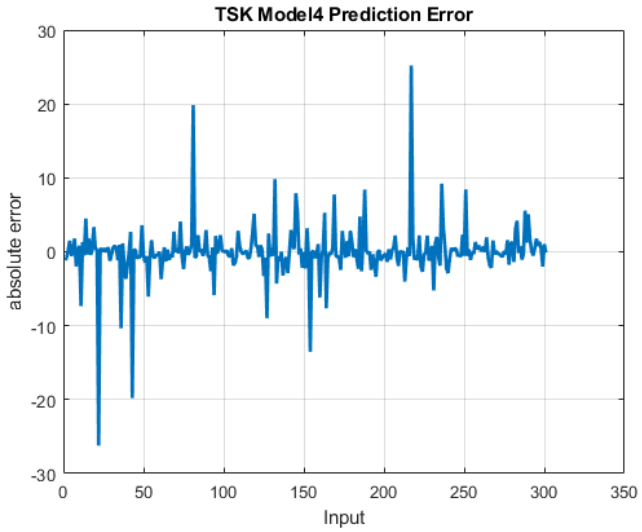




Διάγραμμα μάθησης



Διάγραμμα σφαλμάτων πρόβλεψης



Από το διάγραμμα μάθησης, του δεύτερου μοντέλου, παρατηρούμε, παρά τις ταλαντώσεις, ότι η απόκλιση του Validation Error σε σχέση με το Training Error αυξάνεται. Αυτό το γεγονός οφείλεται στο ότι στο μοντέλο εμφανίζεται φαινόμενο υπερεκπαίδευσης (overfitting).

3.5.ΑΠΟΤΕΛΕΣΜΑΤΑ

	TSK_model_1	TSK_model_2	TSK_model_3	TSK_model_4
MSE	18.472	60.375	9.6101	12.939
RMSE	4.2979	7.7701	3.1	3.5971
NMSE	0.38802	1.2682	0.20186	0.27179
NDEI	0.62291	1.1261	0.44929	0.52134
R2	0.61198	-0.26819	0.79814	0.72821

Σύμφωνα με τα παραπάνω, παρατηρείται πως η συμπεριφορά των μοντέλων με έξοδο Polynomial είναι σχετικά καλύτερη από τα μοντέλα με Singleton έξοδο.

4.ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ Β

Για την εργασία, επιλέγεται το Superconductivity dataset από το UCI Repository, το οποίο περιλαμβάνει 21.263 δείγματα, με κάθε ένα να περιγράφεται από 81 χαρακτηριστικά. Για την αντιμετώπιση του φαινομένου rule explosion, εφαρμόζονται δύο σημαντικές μέθοδοι:

Επιλογή Χαρακτηριστικών: Εδώ, αποφασίζεται πόσα από τα 81 αρχικά χαρακτηριστικά θα χρησιμοποιηθούν για την εκπαίδευση των μοντέλων.

Διαμέριση Διασκορπισμού (Clustering): Αυτή η μέθοδος επηρεάζει την ακτίνα των clusters και, κατά συνέπεια, το πλήθος των κανόνων που θα προκύψουν.

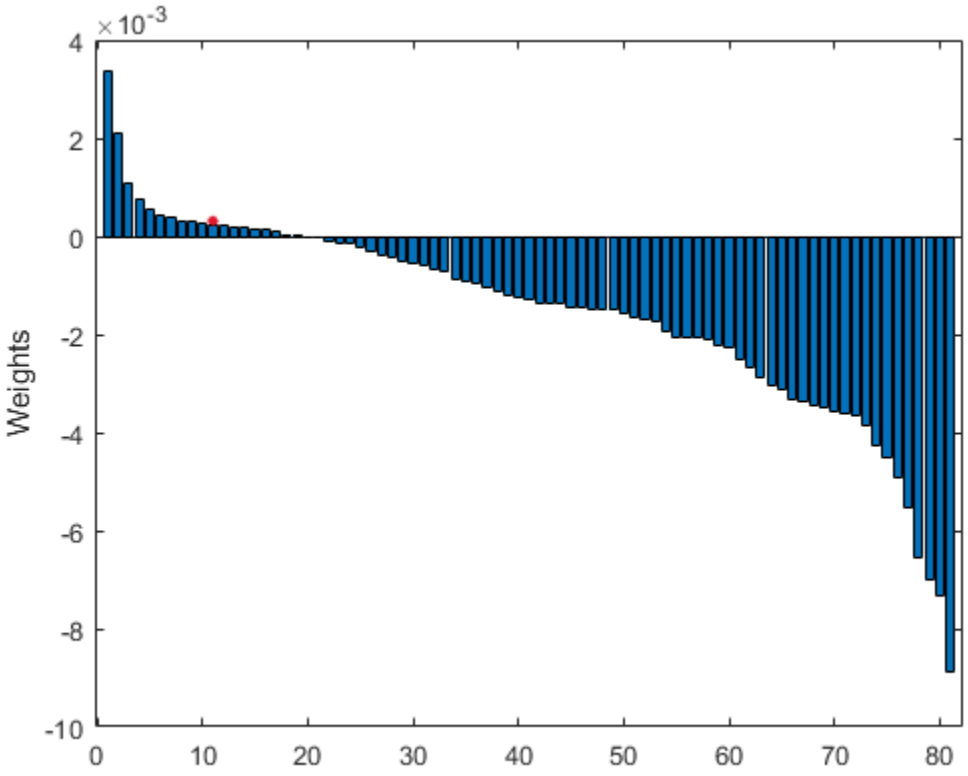
Για την επιλογή των καλύτερων παραμέτρων, χρησιμοποιείται η μέθοδος grid search, κατά την οποία δημιουργείται ένα δισδιάστατο πλέγμα, όπου κάθε σημείο του αντιστοιχεί σε μια συνδυασμένη τιμή για τις παραμέτρους. Σε αυτά τα σημεία, εφαρμόζεται μια μέθοδο αξιολόγησης για να αξιολογηθεί η απόδοση του μοντέλου. Πιο συγκεκριμένα, η μέθοδος αξιολόγησης που χρησιμοποιείται είναι η διασταυρωμένη επικύρωση (Cross Validation). Αυτό σημαίνει ότι για κάθε συνδυασμό παραμέτρων, διαχωρίζεται το σύνολο εκπαίδευσης σε δύο υποσύνολα, ένα για την εκπαίδευση του μοντέλου και ένα για την αξιολόγησή του. Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές (συνήθως 5 ή 10), και τελικά υπολογίζεται ο μέσος όρος του σφάλματος του μοντέλου.

Για κάθε διαδικασία διασταυρωμένης επικύρωσης, τα μοντέλα εκπαιδεύονται για 75 Epochs.

5.GRID SEARCH

Για την επιλογή χαρακτηριστικών, εφαρμόζεται στα δεδομένα η συνάρτηση του MATLAB, relieff για k-nearest neighbors. Η συνάρτηση αυτή επιστρέφει τους δείκτες των χαρακτηριστικών των δειγμάτων με σειρά σημαντικότητας βάσει ενός βάρους, όπως φαίνεται στο παρακάτω σχήμα. Παρατηρείται πως ήδη μέχρι το 11ο σημαντικότερο χαρακτηριστικό, το βάρος είναι πολύ μικρό. Επομένως δεν θα χρειαστεί να χρησιμοποιηθούν περισσότερα χαρακτηριστικά. Επίσης για λόγους πρακτικότητας δεν ξεπερνάμε αυτήν την τιμή διότι ο χρόνος εκπαίδευσης αυξάνεται εκθετικά στο περιβάλλον MATLAB, σε σχέση με τον αριθμό χαρακτηριστικών.

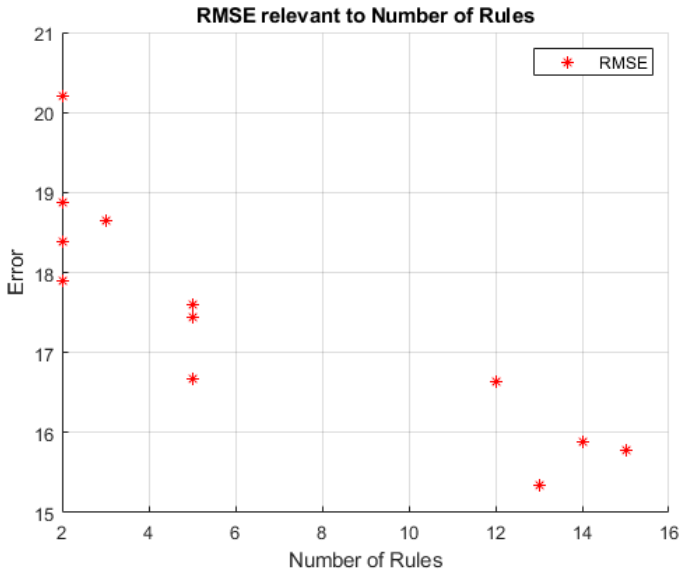
Βάρη σημαντικότητας χαρακτηριστικών:



Επομένως για τις παραμέτρους του πλέγματος αυθαίρετα διαλέγουμε:

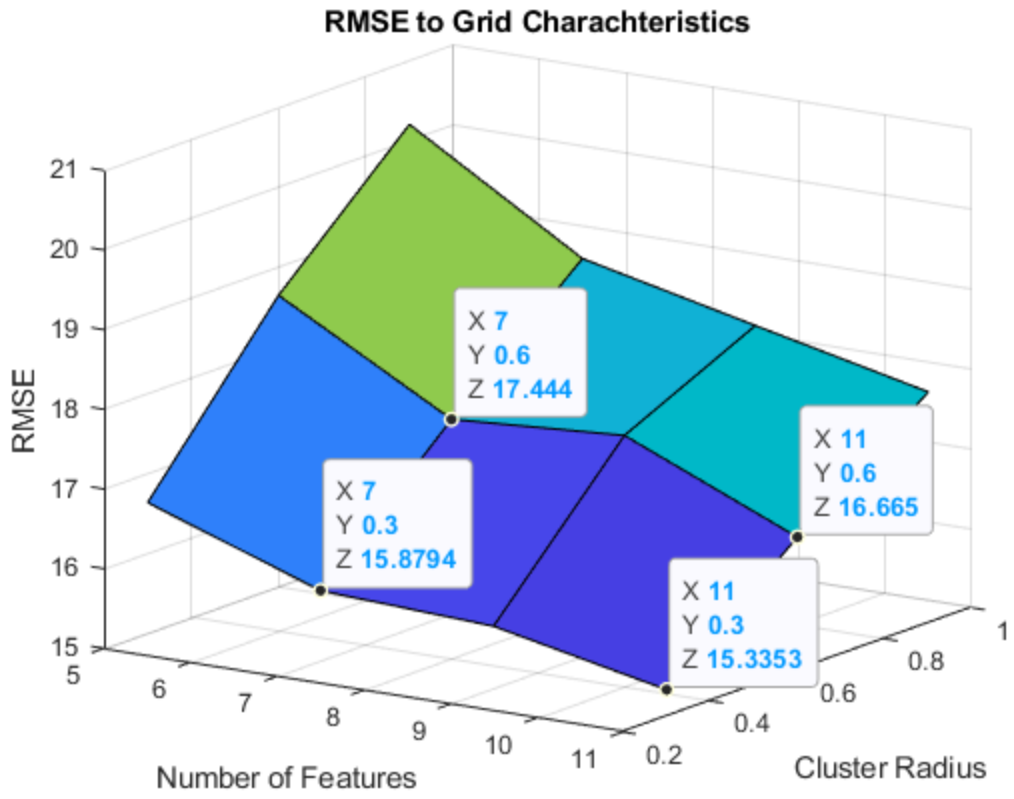
- Αριθμός χαρακτηριστικών = {5, 7, 9, 11}
- Ακτίνα cluster $r_{cluster}$ = {0.3, 0.6, 0.9}

Στη συνέχεια, παίρνουμε το γράφημα RMSE συναρτήσει του πλήθους κανόνων:



Παρατηρώντας το γράφημα, φαίνονται 12 μετρήσεις καθώς το σφάλμα που υπολογίζεται είναι το μέσο σφάλμα από κάθε k-fold. Βγαίνει το συμπέρασμα πως καθώς αυξάνεται το πλήθος των κανόνων, μειώνεται το μέσο σφάλμα του μοντέλου.

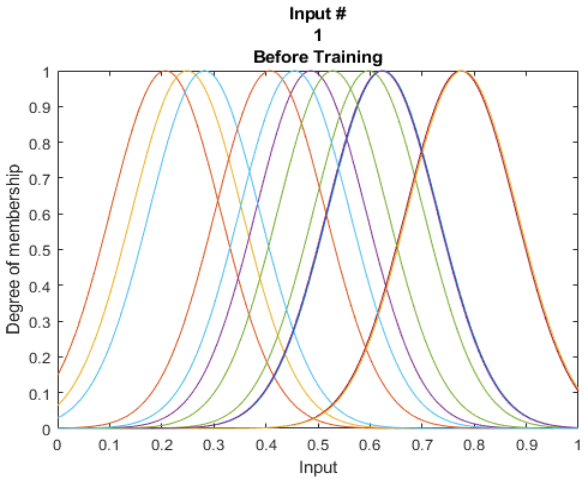
Τα υποψηφία σημεία εύρεσης του βέλτιστου μοντέλου μπορεί να διαπιστωθούν από την επιφάνεια του παρακάτω σχήματος:



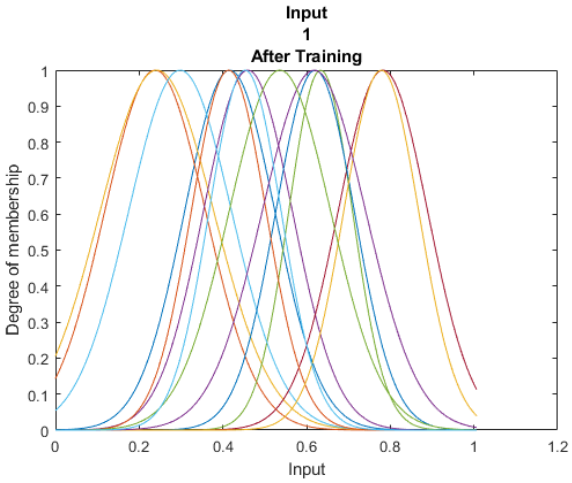
Επιλέγεται μοντέλο με αριθμό χαρακτηριστικών $feat_{num}$ = 11 και μέγεθος ακτίνας cluster, $r_{cluster}$ = 0.3, το εκπαιδεύουμε για τα αρχικά δείγματα εκπαίδευσης αλλά κρατάμε μόνο τα 11 πιο σημαντικά χαρακτηριστικά. Από την εκπαίδευση του μοντέλου έχουμε:

Αρχικές και τελικές μορφές Ασαφών Συνόλων:

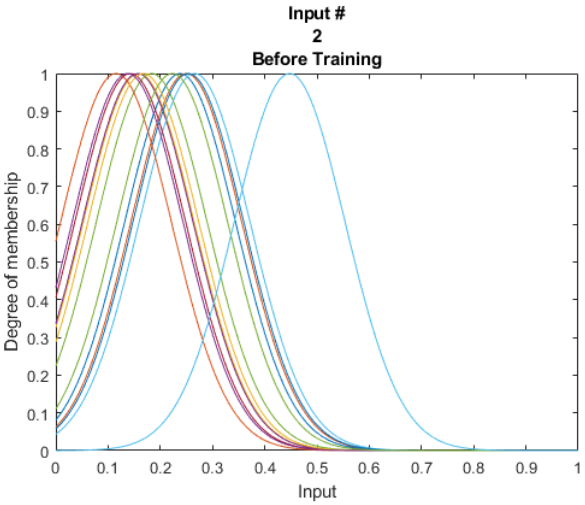
Είσοδος 1 Αρχικό



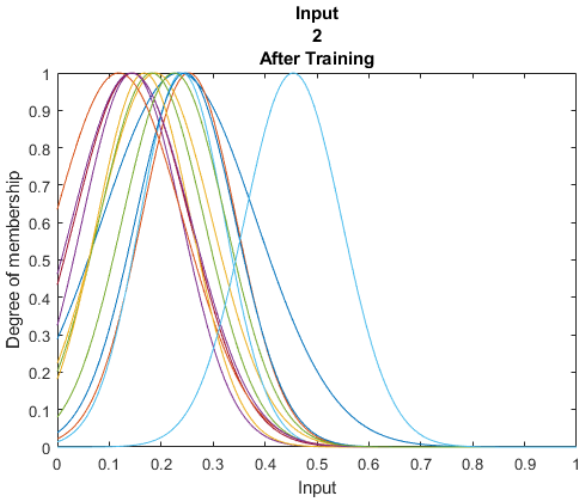
Είσοδος 1 Τελικό



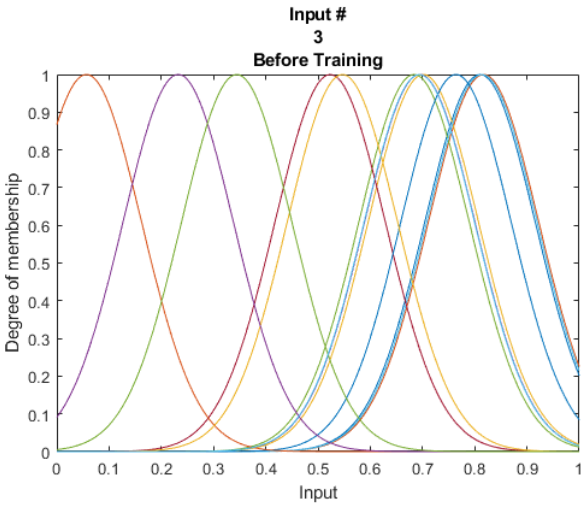
Είσοδος 2 Αρχικό



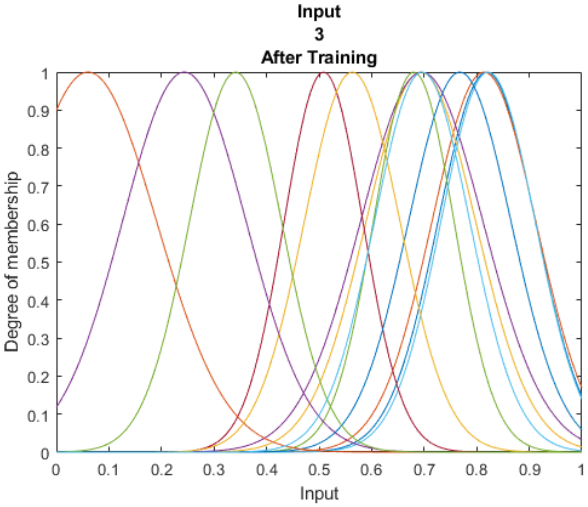
Είσοδος 2 Τελικό



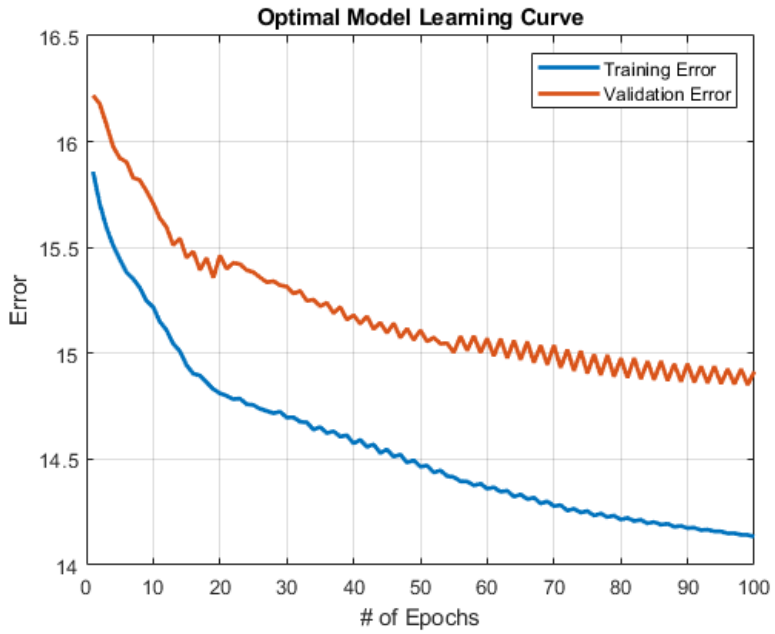
Είσοδος 3 Αρχικό



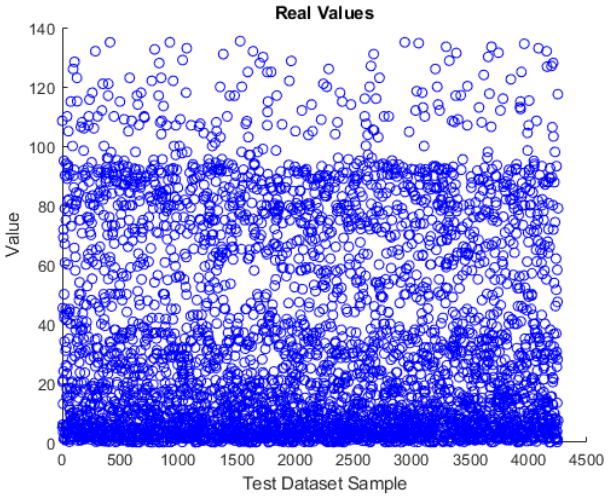
Είσοδος 3 Τελικό

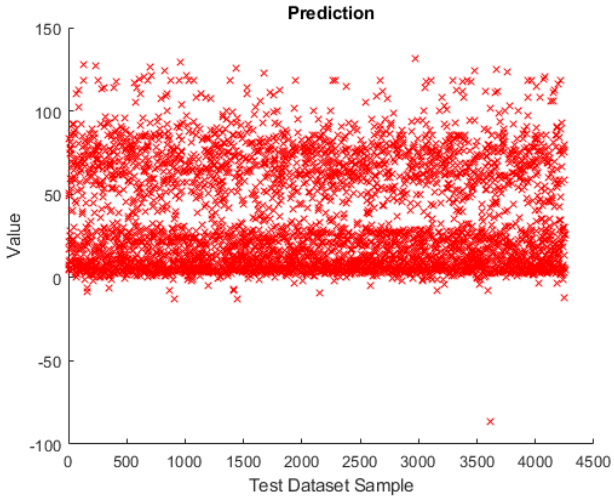


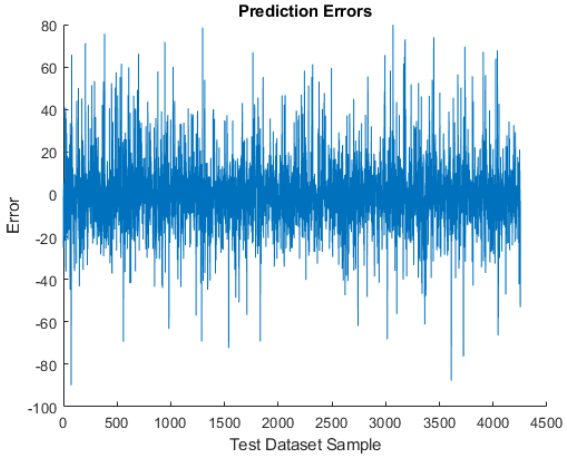
Διάγραμμα Μάθησης:



Διαγράμματα Πραγματικών – Προβλεπόμενων τιμών:







Τιμές δεικτών απόδοσης βέλτιστου μοντέλου:

Optimal Model	
MSE	235.44
RMSE	15.344
NMSE	0.20146
NDEI	0.44885
R2	0.79854

Επειδή δεν υπάρχει σταθερός διαμερισμός του αρχικού dataset, τα αποτελέσματα μπορεί να αποκλίνουν για κάθε ξεχωριστή εκτέλεση των αλγορίθμων. Στην συγκεκριμένη περίπτωση, προκύπτει από τα διαγράμματα σφάλματος πρόβλεψης, πως το μοντέλο προβλέπει σε ικανοποιητικό βαθμό το σύνολο ελέγχου, με μεμονωμένες περιπτώσεις μεγάλων σφαλμάτων να είναι όμως υπαρκτές.