ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

Επίλυση προβλήματος παλινδρόμησης με χρήση μοντέλων TSK

Κωνσταντίνος Ανδρέου
9521
andreouk@ece.auth.gr

Περιεχόμενα :

Περιγραφή εργασίας και ζητούμενα

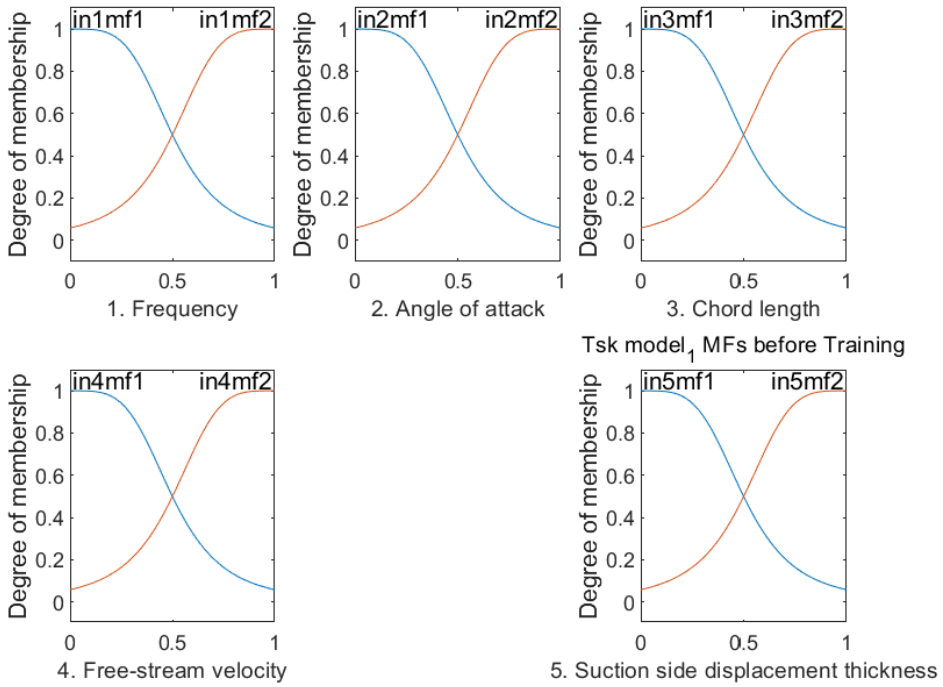
Στόχος της εργασίας αυτής είναι να διερευνηθεί η ικανότητα των μοντέλων TSK στη μοντελοποίηση πολυμεταβλητών, μη γραμμικών συναρτήσεων. Συγκεκριμένα, επιλέγονται δύο σύνολα δεδομένων από το UCI repository με σκοπό την εκτίμηση της μεταβλητής στόχου από τα διαθέσιμα δεδομένα, με χρήση ασαφών νευρωνικών μοντέλων. Το πρώτο σύνολο δεδομένων 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, πολυπλοκότερο σύνολο δεδομένων θα χρησιμοποιηθεί για μια πληρέστερη διαδικασία μοντελοποίησης, η οποία θα περιλαμβάνει μεταξύ άλλων προεπεξεργαστικά βήματα όπως επιλογή χαρακτηριστικών (feature selection), καθώς και μεθόδους βελτιστοποίησης των μοντέλων μέσω της διασταυρωμένης επικύρωσης (cross validation).

Πρώτο ζητούμενο , εφαρμογή σε απλό dataset

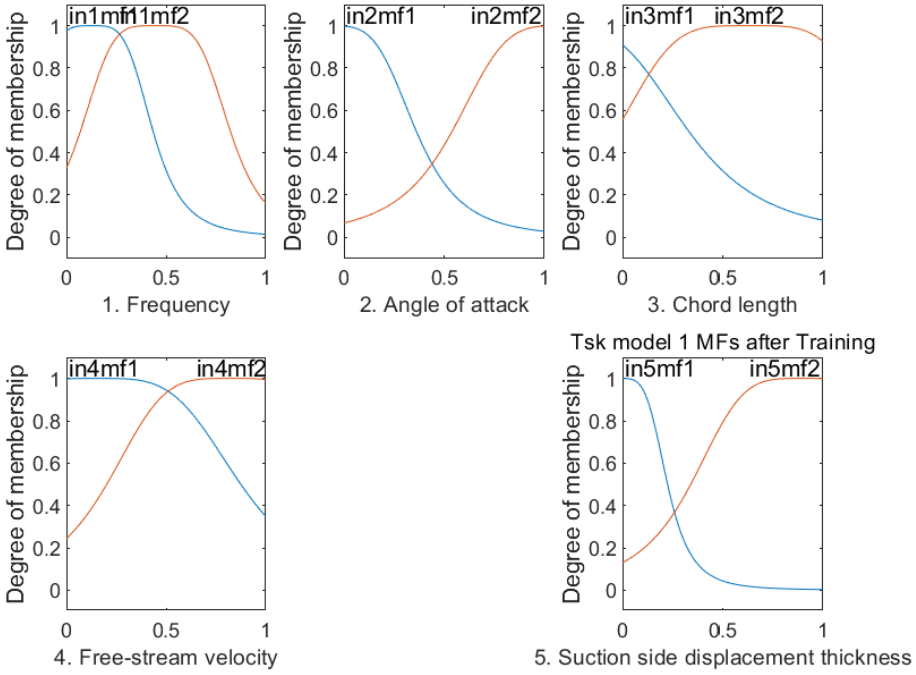
Στην πρώτη φάση της εργασίας, επιλέγεται από το UCI repository το Airfoil Self-Noise data set, το οποίο περιλαμβάνει 1503 δείγματα (instances) και 6 χαρακτηριστικά (features). Αφού πήραμε τα δεδομένα μας , σε πρώτη φάση τα χωρίσαμε σε δύο μη επικαλυπτόμενα υποσύνολα Dtrn, Dval, Dchk απο τα οποία το πρώτο το χρησιμοποιούμε για να εκπαιδεύσουμε το μοντέλο , το δεύτερο για επικύρωση και αποφυγή υπερεκπαιδευσης και το τρίτο για έλεγχο του μοντέλου. Τα δεδομένα τα χωρίσαμε σε 60% εκπαίδευσης , 20% επικύρωσης και 20% επαλήθευσης. Χρησιμοποιούμε για βελτιστοποίηση παραμέτρων backpropagation για ολό το σετ δεδομένων ενώ για την έξοδο ελάχιστα τετράγωνα (least squares). Αρχικοποίηση επισήσης της bell-shaped συναρτήσεις συμμετοχής έτσι ώστε να σύνολα μας να έχουν βαθμό επικάλυψης 0.5 για κάθε είσοδο. Εκπαιδευσαμε 4 μοντέλα TSK με σκοπό στο καθένα να μεταβάλλονται η μορφή της εξόδου και το πλήθος των συναρτήσεων συμμετοχής για κάθε μεταβλητή εισόδου. Για τα πρώτα δύο μοντέλα έχουμε έξοδο singleton ενώ για τα άλλα δύο polynomial

Αποτελέσματα πρώτου ζητούμενου

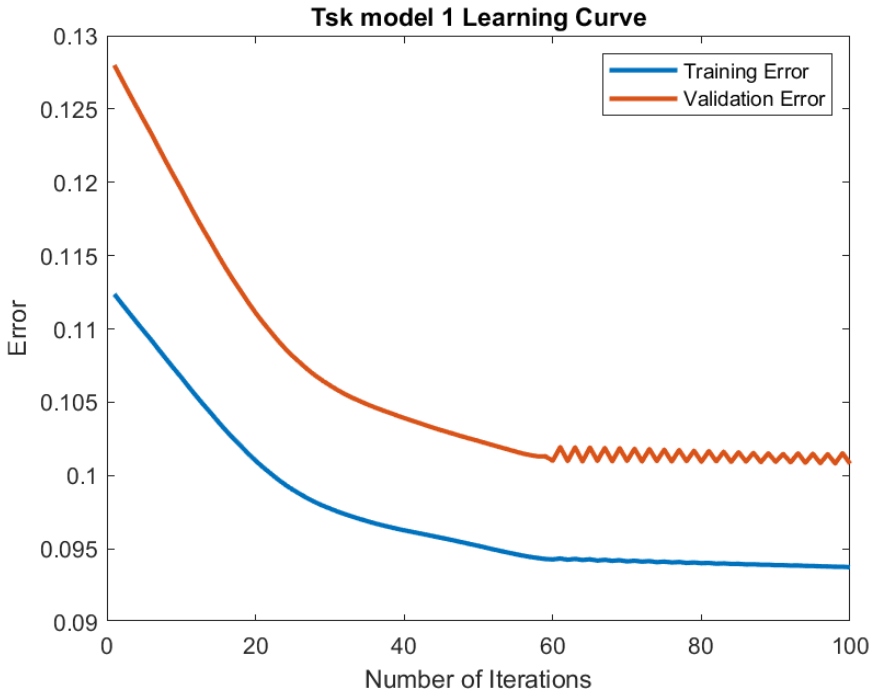
TSK_model_1



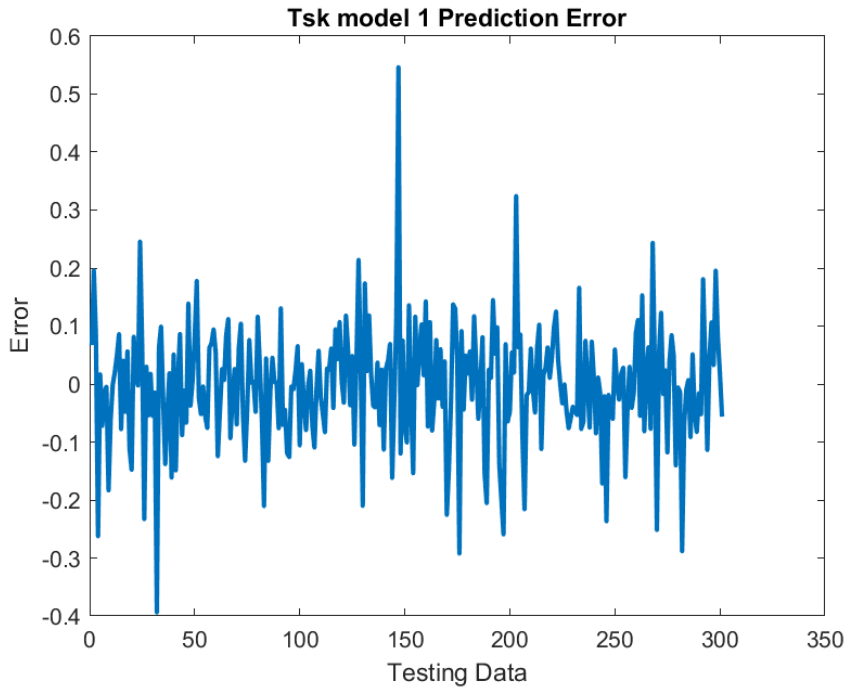
1. TSK_model1 MFs before Training(Αρχικές συναρτησεις συμμετοχης πριν την εκπαίδευση)



2. TSK_model1 MFs after Training(Τελικές συναρτήσεις συμμετοχής μετά την εκπαίδευση)

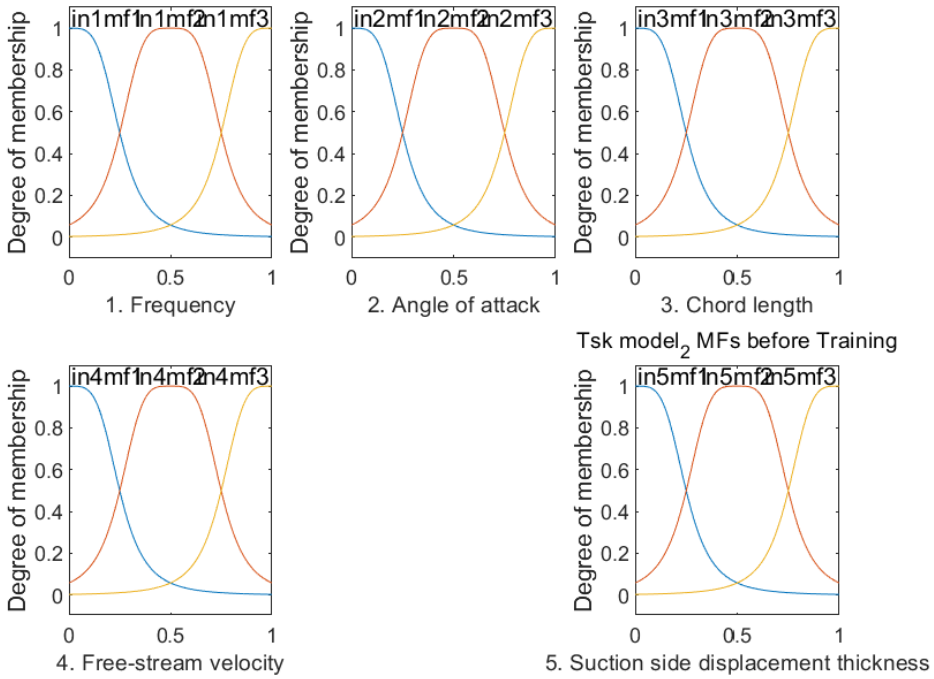


3. TSK_model_1 Learning Curve(Καμπύλες εκπαίδευσης)

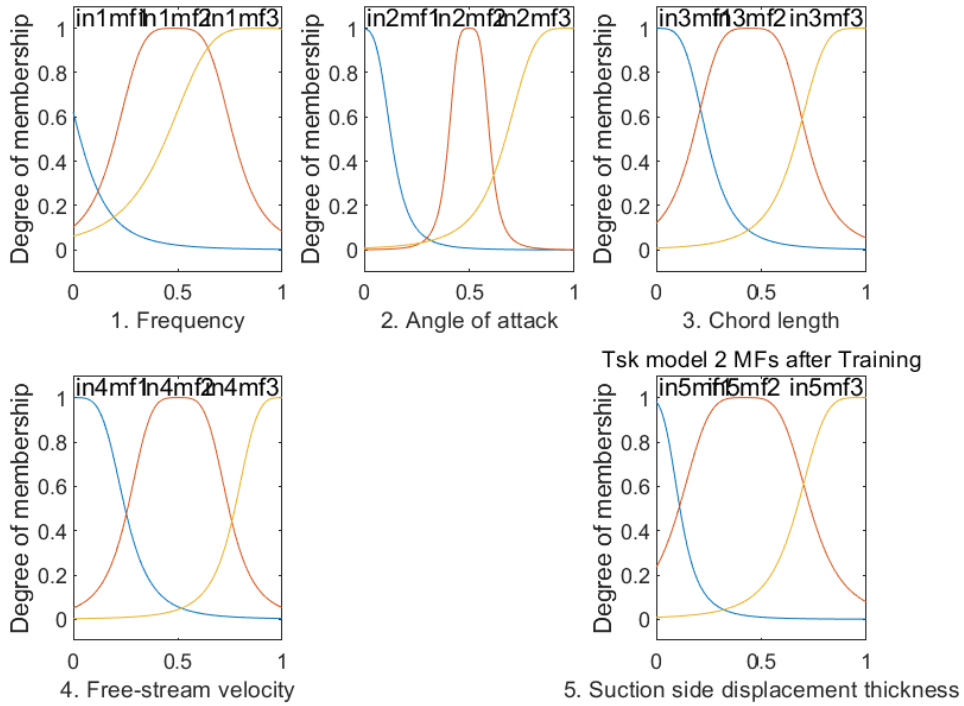


4. TSK_model_1 Prediction Error(Σφάλμα πρόβλεψης)

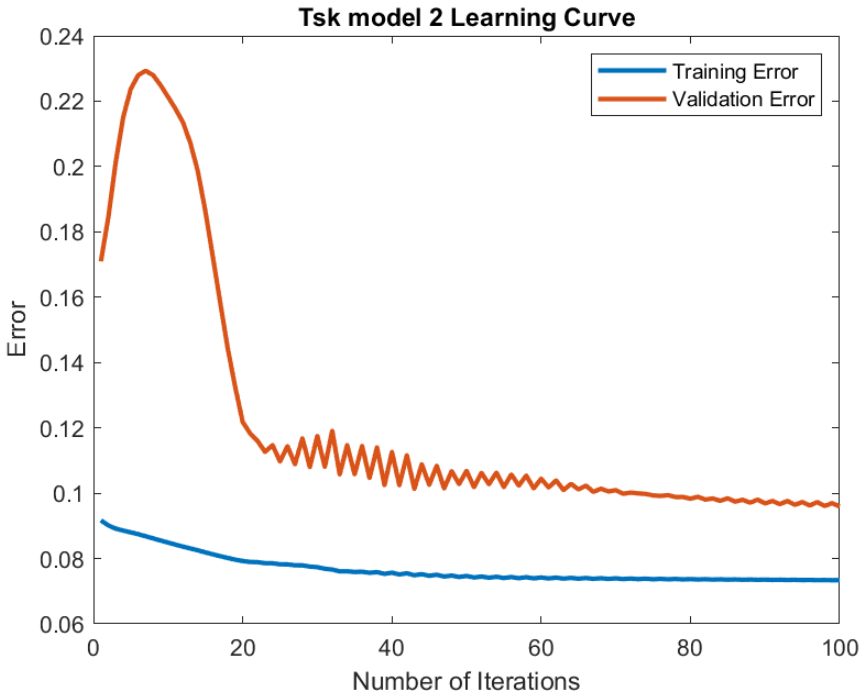
TSK_model_2



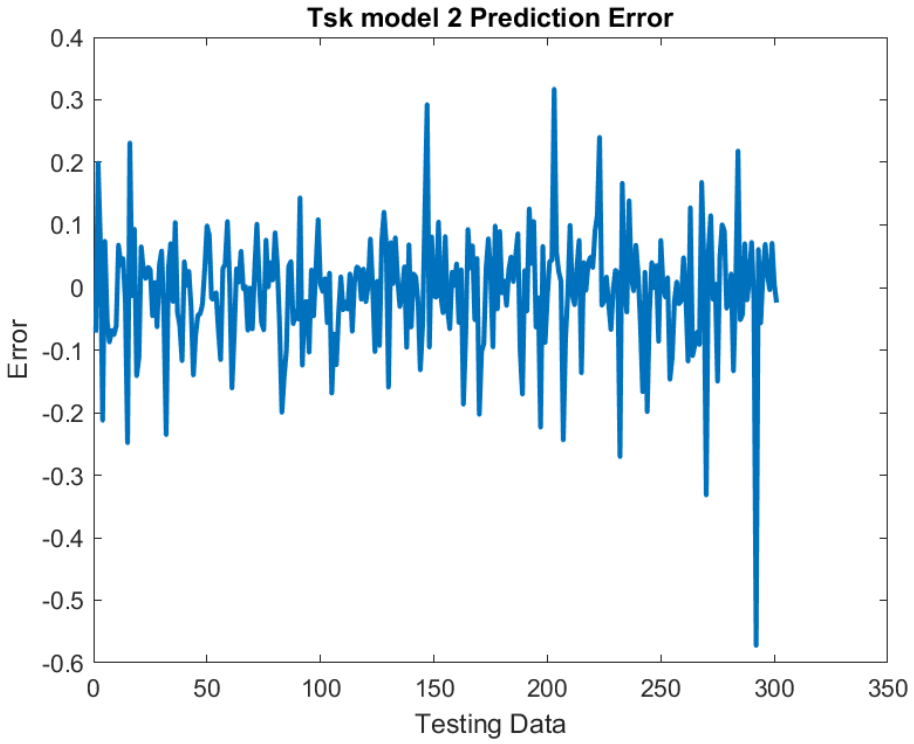
5. TSK_model2 MFs before Training(Αρχικές συναρτησεις συμμετοχης πριν την εκπαίδευση)



6. TSK_model2 MFs after Training(Τελικές συναρτήσεις συμμετοχής μετά την εκπαίδευση)

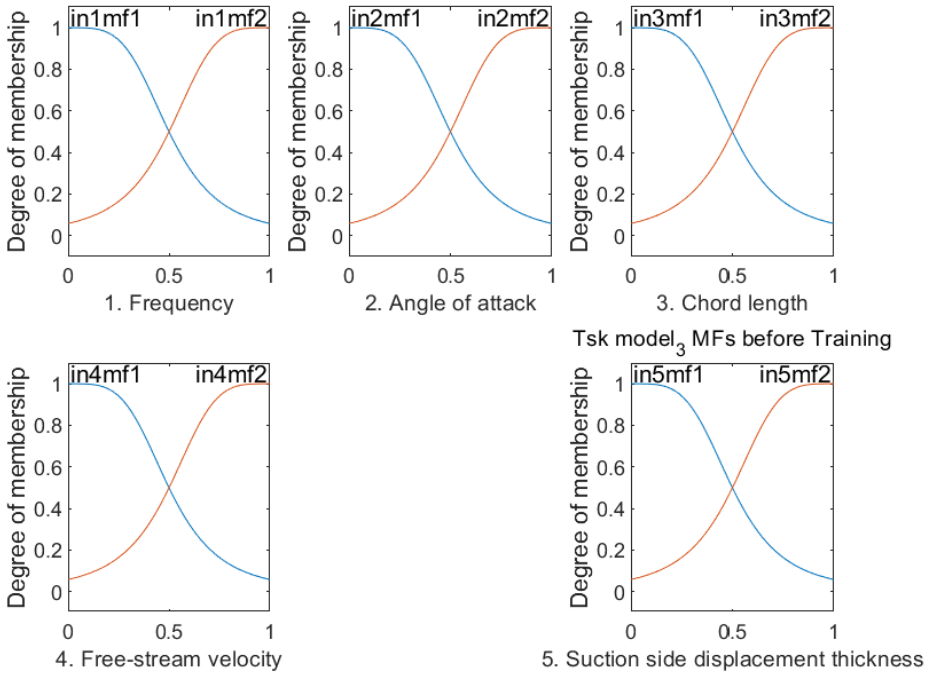


7. TSK_model_2 Learning Curve(Καμπύλες εκπαίδευσης)

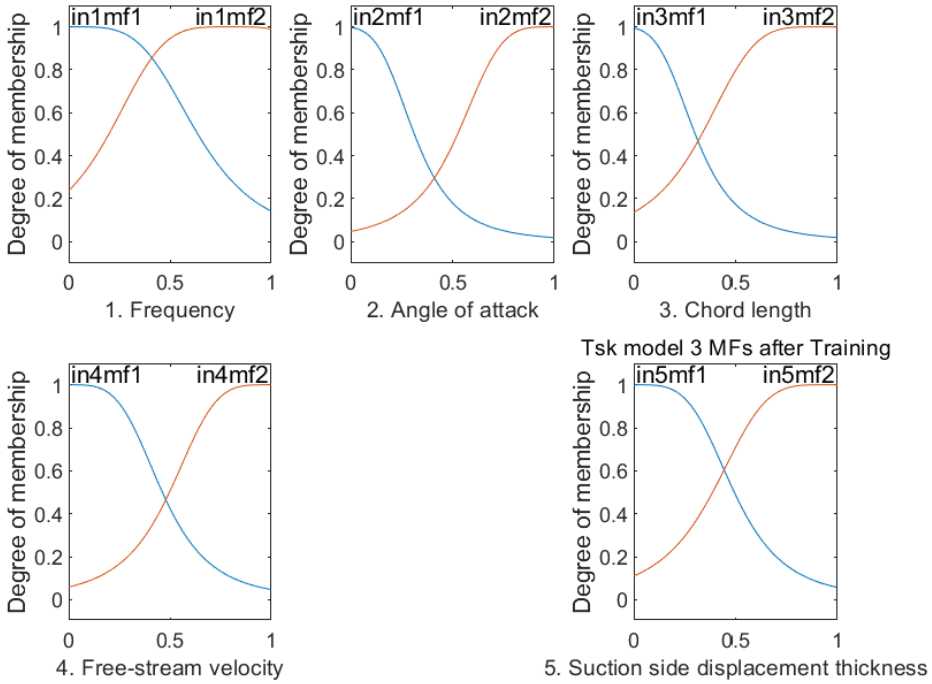


8. TSK_model_1 Prediction Error(Σφάλμα πρόβλεψης)

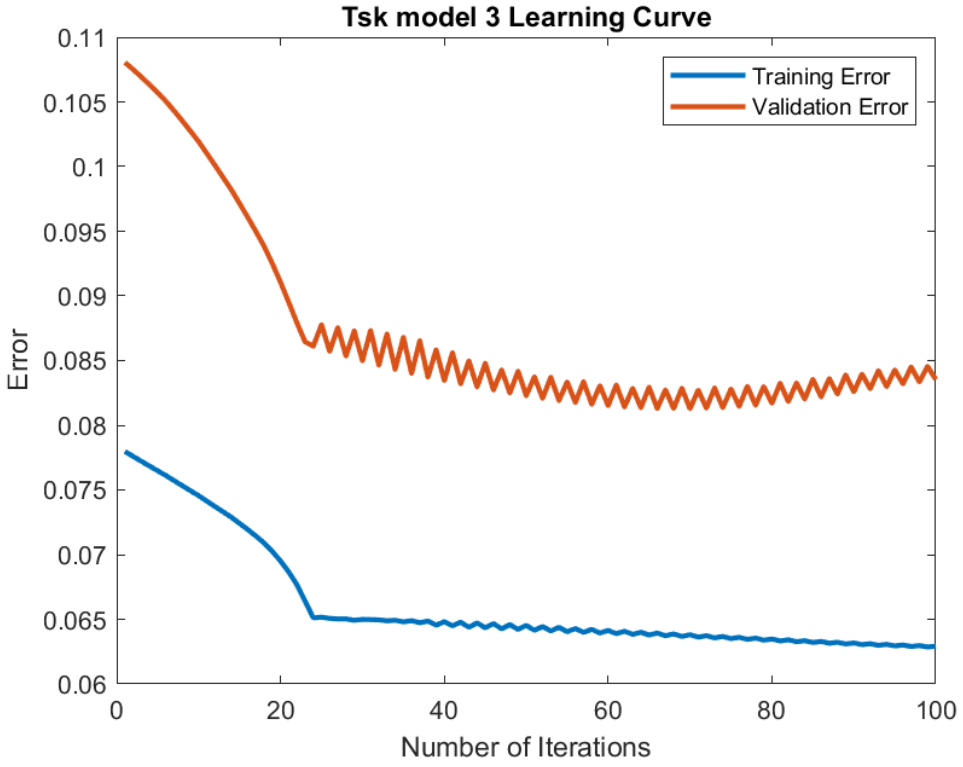
TSK_model_3



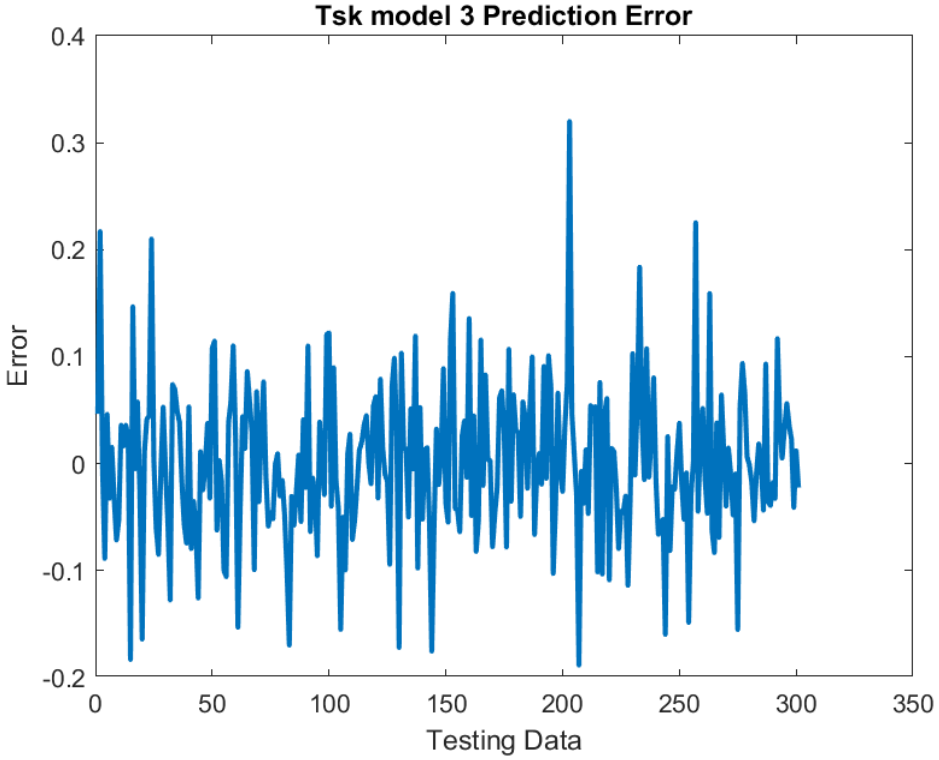
9. TSK_mode3 MFs before Training(Αρχικές συναρτησεις συμμετοχης πριν την εκπαίδευση)



10. TSK_model3 MFs after Training(Τελικές συναρτήσεις συμμετοχής μετά την εκπαίδευση)

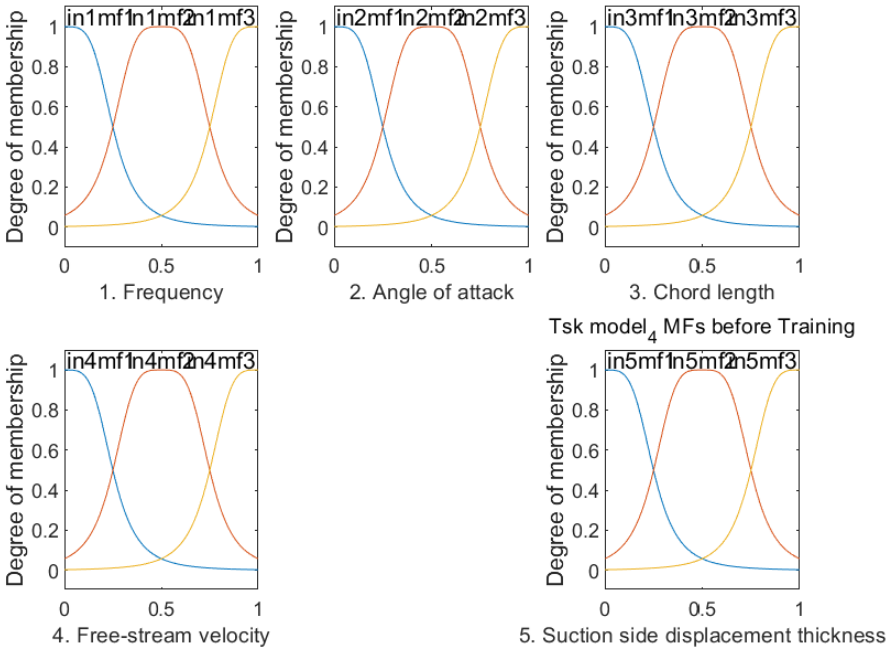


11. TSK_model_3 Learning Curve(Καμπύλες εκπαίδευσης)

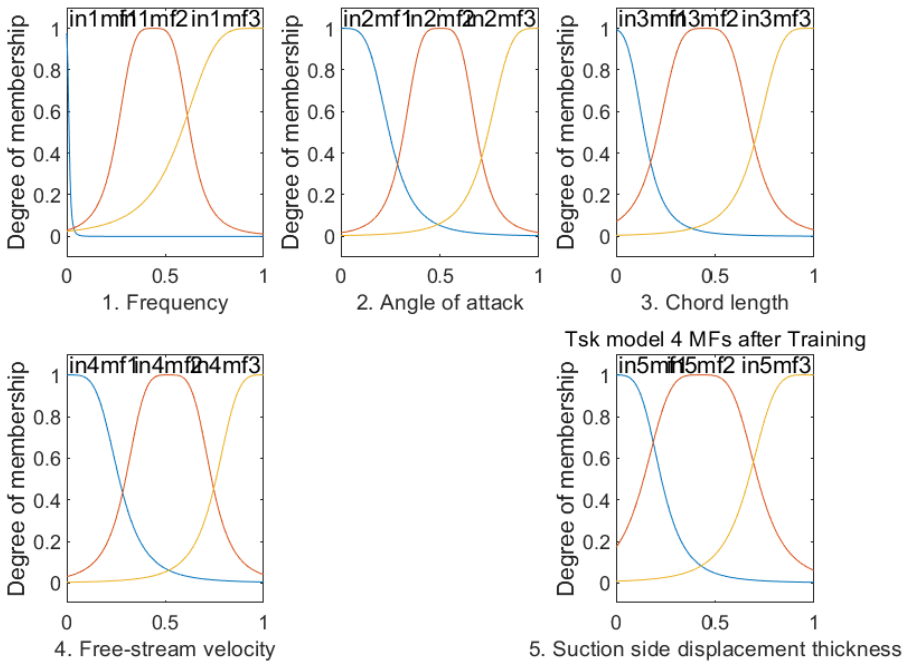


12. TSK_model_3 Prediction Error(Σφάλμα πρόβλεψης)

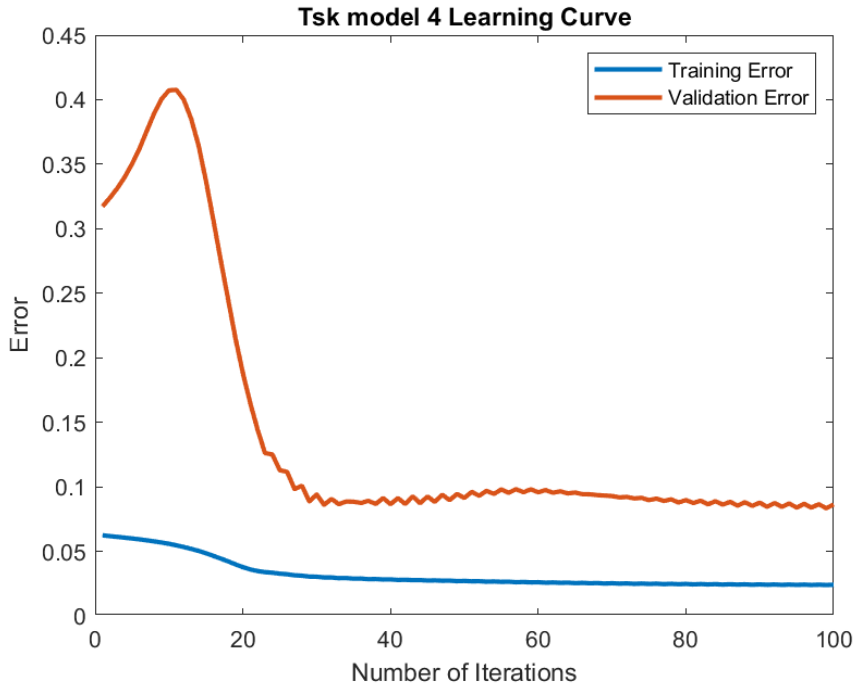
TSK_model_3



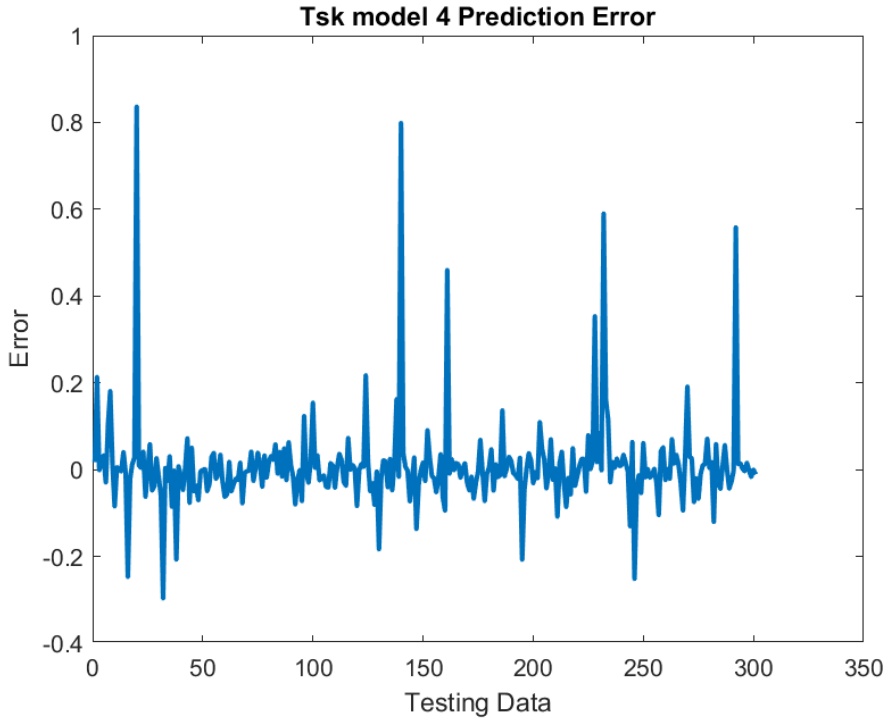
13. TSK_mode4 MFs before Training(Αρχικές συναρτησεις συμμετοχης πριν την εκπαίδευση)



14. TSK_model4 MFs after Training(Τελικές συναρτήσεις συμμετοχής μετά την εκπαίδευση)



15. TSK_model_4 Learning Curve(Καμπύλες εκπαίδευσης)



16. TSK_model_4 Prediction Error(Σφάλμα πρόβλεψης)

Δείκτες απόδοσης μοντέλων

TSK_model_1
RMSE = 0.114496
NMSE = 0.389447
NDEI = 0.624057
R^2 = 0.610553

TSK_model_2
RMSE = 0.170372
NMSE = 0.862306
NDEI = 0.928605
R^2 = 0.137694

TSK_model_3
RMSE = 0.082873
NMSE = 0.204030
NDEI = 0.451696
R^2 = 0.795970

TSK_model_4
RMSE = 0.095544
NMSE = 0.271192
NDEI = 0.520761
R^2 = 0.728808

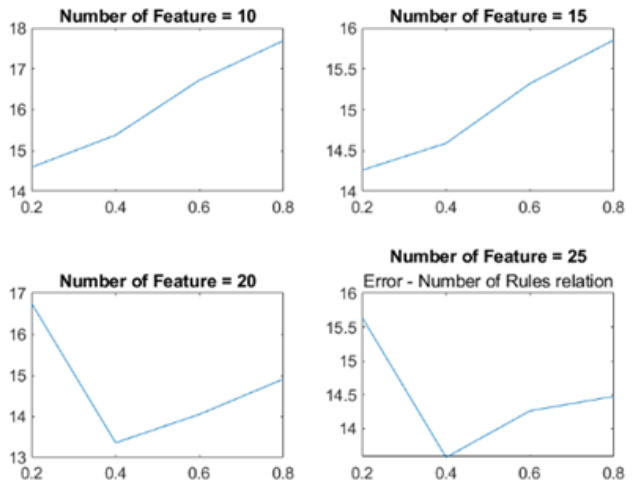
Συμπεράσματα:

Απο της γραφικές παραστάσεις των καμπυλών εκπαίδευσης μπορούμε απευθείας να δούμε ότι το μοντέλο 3 έχει υπερεκπαιδευτεί γιατί τα validation data αποκλείουν , αλλά έχει και τα καλύτερα αποτελέσματα. Συγκριτικά τα μοντέλα που έχουν μόνο 2(1 και 3) συναρτήσεις συμμετοχής έχουν καλύτερα αποτελέσματα από αυτά με 3(2 και 4) , δηλαδή περισσότερες συναρτήσεις συμμετοχής δεν σημαίνει απαραίτητα καλύτερο μοντέλο και οδηγούμαστε πιο εύκολα σε υπερεκπαίδευση. Επίσης φαίνεται ότι κάποιες συνάρτησης συμμετοχής εμφανίζουν μεγαλύτερο βαθμό συμμετοχής από κάποιες άλλες άρα και τα μοντέλα μας προσαρμόζονται περισσότερο σε κάποιες από αυτές. Βλέπουμε πως τα μοντέλα μας έχουν καλά αποτελέσματα οσον αφορα την έξοδο , δηλαδή την προσεγγίζει ικανοποιητικά αλλά τα μοντέλα με πολυωνυμική έξοδο προβλέπουν καλύτερα. Είναι κατανοητό ότι όσο περισσότερες συναρτήσεις συμμετοχής έχουμε τόσο μειώνεται το σφάλμα εξόδου χωρίς να συμένη πως αυτό είναι απαραίτητα καλό αφού μπορεί πιο εύκολα να υπάρξει υπερεκπαίδευση . Γενικά το καλύτερο μοντέλο απο τα ποιό πάνω είναι το 3 καθώς έχει τα καλύτερα αποτελέσματα αλλά βλέπουμε πως ίσως να έχει υποστεί υπερεκπαίδευση.

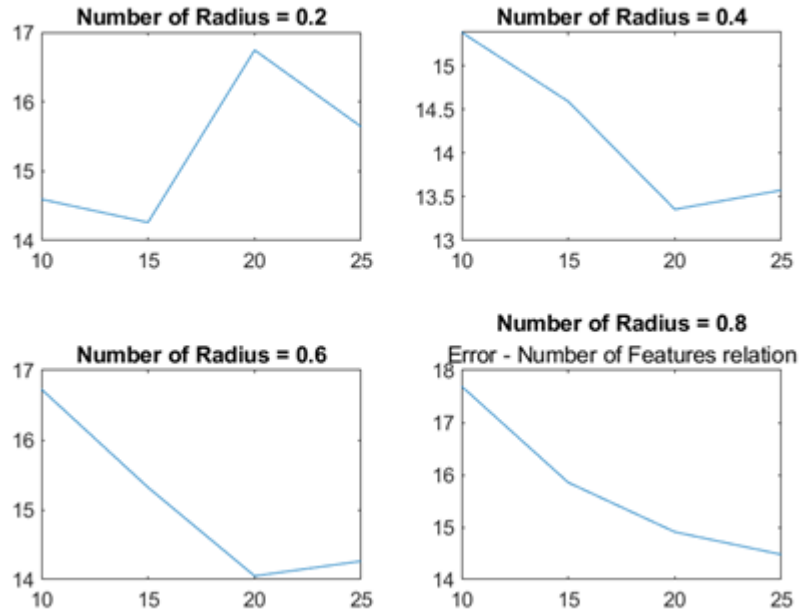
Δεύτερο ζητούμενο , εφαρμογή σε απλό dataset υψηλών διαστάσεων

Στην δεύτερη φάση της εργασίας, επιλέγεται από το UCI repository το Superconductivity data set, το οποίο περιλαμβάνει 121263 δείγματα, όπου το κάθε ένα περιγράφεται από 81 μεταβλητές/χαρακτηριστικά. Είναι φανερό ότι το μέγεθος του data set καθιστά απαγορευτική μια απλή εφαρμογή ενός TSK μοντέλου. Αφού πήραμε τα δεδομένα μας , σε πρώτη φάση τα χωρίσαμε σε δύο μη επικαλυπτόμενα υποσύνολα Dtrn, Dval, Dchk απο τα οποία το πρώτο το χρησιμοποιούμε για να εκπαιδεύσουμε το μοντέλο , το δεύτερο για επικύρωση και αποφυγή υπερεκπαιδευσης και το τρίτο για έλεγχο του μοντέλου. Τα δεδομένα τα χωρίσαμε σε 60% εκπαίδευσης , 20% επικύρωσης και 20% επαλήθευσης. Μετά χρησιμοποιούμε grid-search (αναζήτηση πλέγματος) και για αξιολόγηση χρησιμοποιούμε 5-fold cross validation(5-πτυχης διασταυρωμένης επικύρωσης) για βελτιστοποίηση παραμέτρων. Κάθε φορά απο τις 5 επαναλήψεις χρησιμοποιούμε το 80% για εκπαίδευση και 20% για επικύρωση και αποθηκεύεται το μέσο σφάλμα. Για την δημιουργία των κανόνων χρησιμοποιείται ο αλγόριθμος Subtractive Clustering (SC) και για την επιλογή χαρακτηριστικών ο αλγόριθμος Relief.

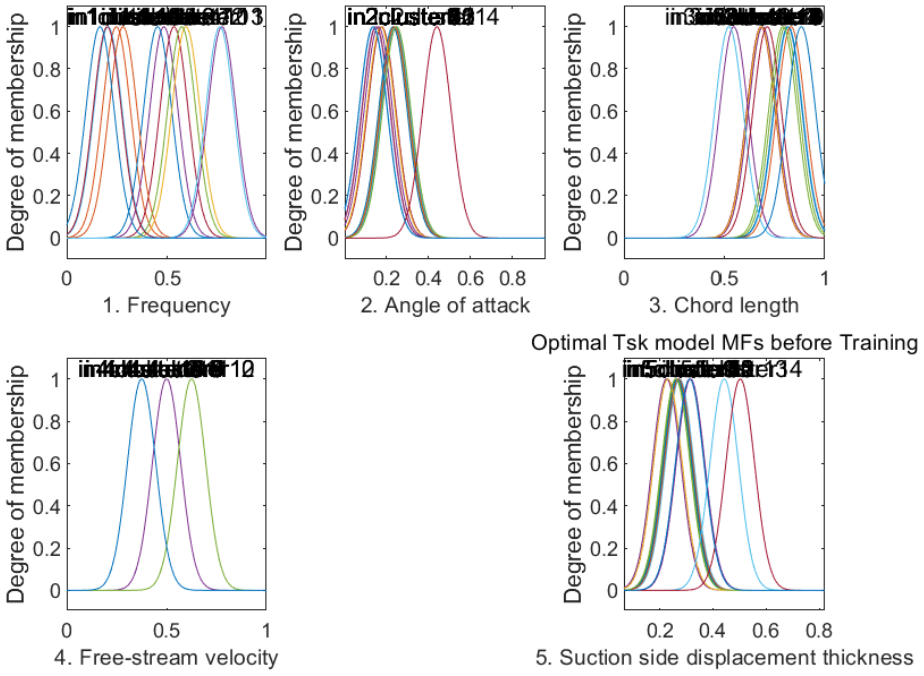
Αποτελέσματα Δεύτερου ζητούμενου



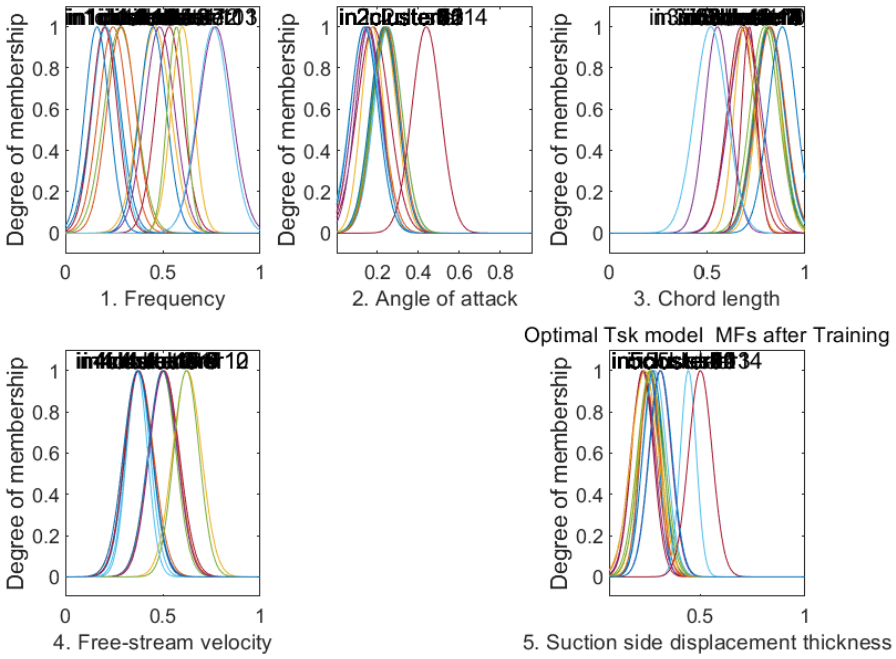
17. Σφάλμα για διαφορετικές τιμές χαρακτηριστικών



18.Σφάλμα για διαφορετικές τιμές ακτίνας

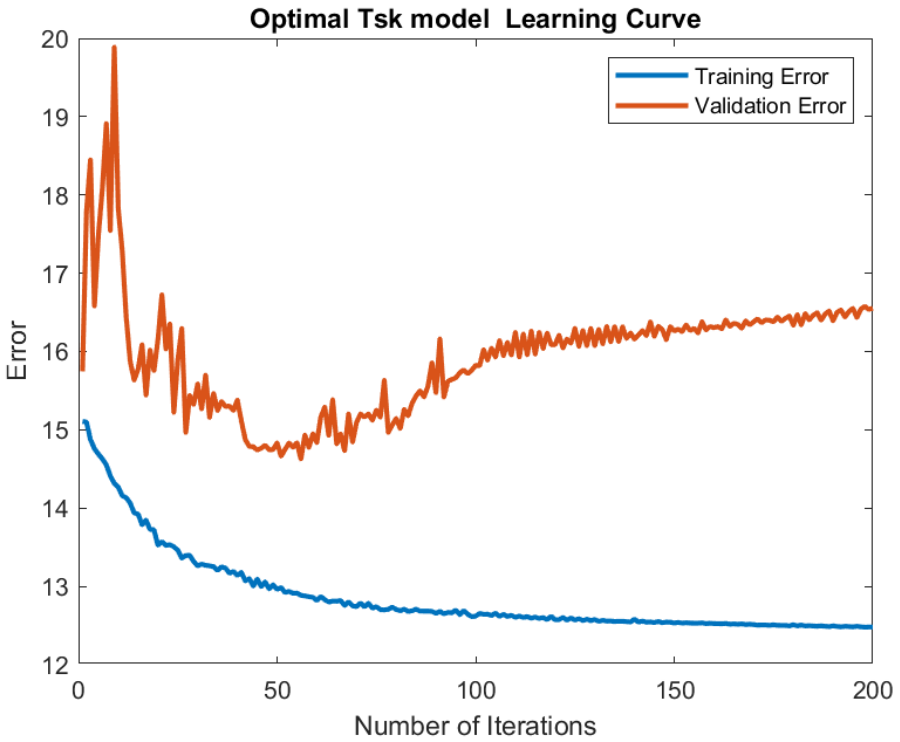


19. Optimal TSK_model MFs before Training (MF's πριν την εκπαίδευση)

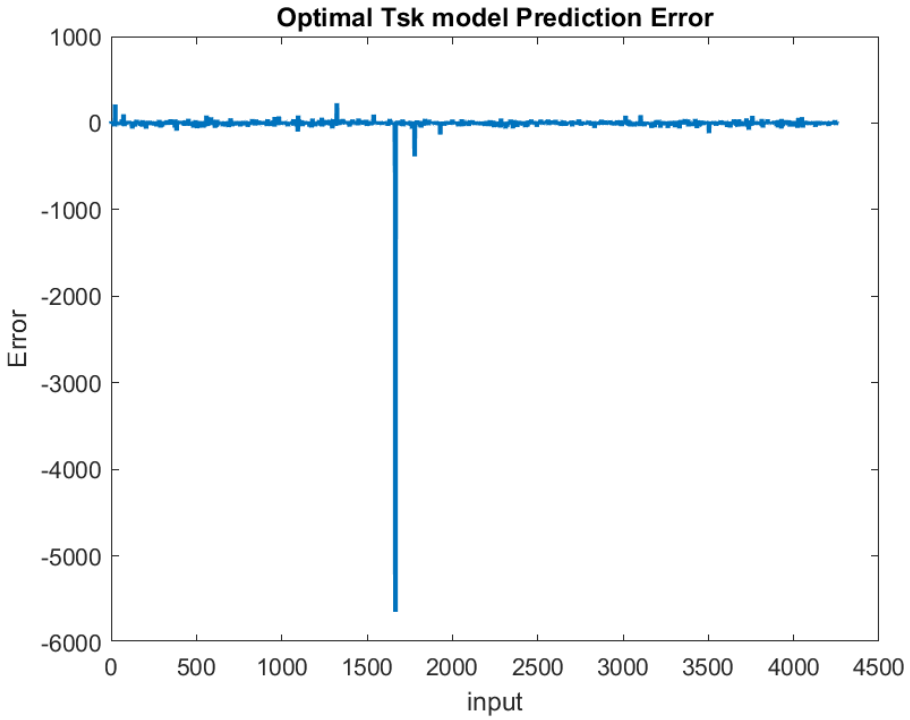


20.Optimal TSK_model MFs after Training(MF's μετά την εκπαίδευση)

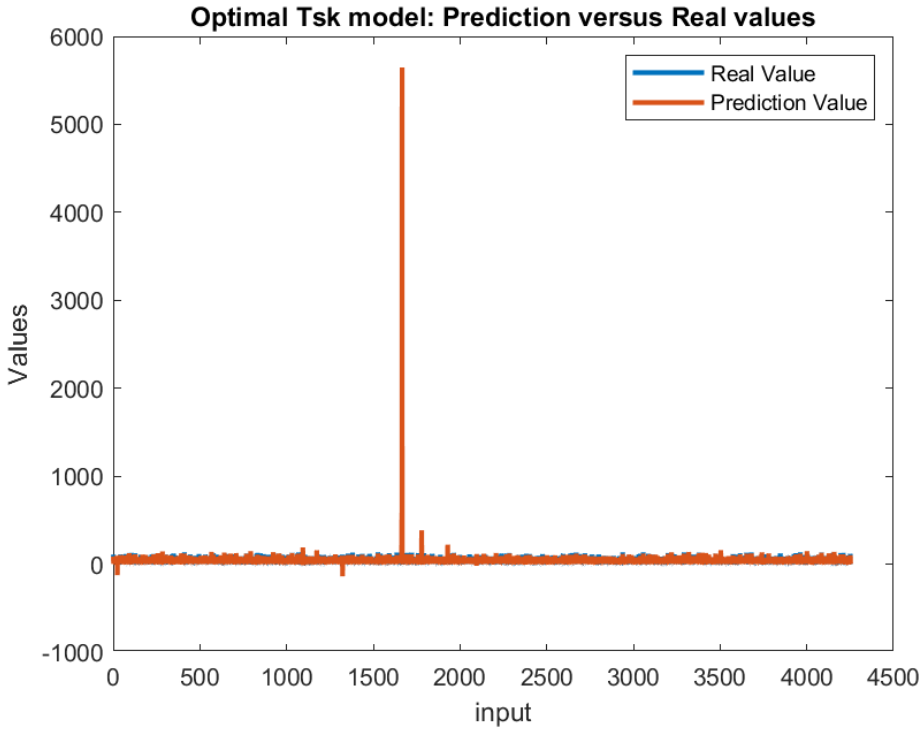
Για το βέλτιστο μοντέλο έχουμε τις εξείς καμπύλες



21.Optimal TSK_model Learning Curve(Λάθος ανα αριθμό επαναλήψεων)



22. Optimal TSK_model Prediction Error (σφάλμα πρόβλεψης βέλτιστου μοντέλου)



23. Optimal TSK_model Model Prediction(Τιμές που προέβλεψε το μοντέλω σε σχέση με τις πραγματικές)

Τα αποτελέσματα για το TSK_model_optimal είναι RMSE = 16.180522, NMSE = 0.225092, NDEI = 0.474439 και = 0.774908.

Συμπεράσματα:

Σύμφωνα με τα ποιό πάνω αποτελέσματα , έχουμε ικανοποιητικό R^2 αρκετά κοντά στο 1 άρα οι εκτιμώμενες τιμές σε σχέση με τις πραγματικές είναι πολύ κοντά όπως βέβαι φένεται και απο τα χαμηλά NMSE και NDEI