ΑΣΑΦΗ ΣΥΣΤΗΜΑΤΑ
ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ
2020 - 2021

Εργασία 3 - Regression

Ονοματεπώνυμο: Παπαδάμ Στέφανος

ΑΕΜ: 8885

email: sgpapadam@ece.auth.gr

Ομάδα Εργασίας: -

Περιγραφή του προβλήματος

Η εργασία αυτή έχει σκοπό την διερεύνηση της ικανότητας των TSK μοντέλων στην μοντελοποίηση μη γραμμικών συναρτήσεων και στην επίλυση προβλημάτων παλινδρόμησης. Για αυτό το λόγο επιλέγονται δύο σύνολα δεδομένων από τα οποία το πρώτο θα χρησιμοποιηθεί για μια απλή μοντελοποίηση της απεικόνισης της συνάρτησης εισόδου στην έξοδο του συστήματος μέσω των χαρακτηριστικών που διαθέτει, ενώ το δεύτερο το οποίο περιέχει περισσότερα χαρακτηριστικά και δείγματα θα χρησιμοποιηθεί για τον ίδιο σκοπό αλλά υλοποιώντας πιο σύνθετα βήματα όπως επιλογή χαρακτηριστικών και μεθόδους βελτιστοποίησης των μοντέλων μέσω διασταυρωμένης επικύρωσης.

Μέρος 1 - Airfol Self-Noise dataset

Περιγραφή dataset

Το πρώτο dataset το οποίο χρησιμοποιείται είναι το airfol self-noise το οποίο περιλαμβάνει 1503 δείγματα και 5 χαρακτηριστικά ενώ διαθέτει μία έξοδο. Τα χαρακτηριστικά με σειρά από τη στήλη 1 έως την 5 είναι τα εξής:

- συχνότητα σε Hz.
- γωνία επίθεσης σε μοίρες.
- μήκος χορδής σε μέτρα.
- ταχύτητα ελεύθερης ροής σε m/s.
- πάχος μετατόπισης πλευρικής αναρρόφησης σε μέτρα.

Η μοναδική έξοδος του dataset είναι το:

- Κλιμακούμενο επίπεδο ηχητικής πίεσης σε decibels.

Διαχωρισμός dataset

Αρχικά θα πρέπει να γίνει διαχωρισμός του dataset σε τρία σύνολα. Τα σύνολα αυτά είναι τα training, validation και check τα οποία χρησιμοποιούνται για εκπαίδευση, επικύρωση και έλεγχο αντίστοιχα. Τα μεγέθη τους είναι 60%, 20%, 20% του αρχικού συνόλου δεδομένων αντίστοιχα. Ο διαχωρισμός του dataset πραγματοποιήθηκε με την έτοιμη συνάρτηση split_scale που μας παρέχετε η οποία χωρίζει το σύνολο στα παραπάνω μεγέθη και επιπλέον πραγματοποιεί και κανονικοποίηση. Στην αρχή προσπαθήσαμε να διαχωρίσουμε manually τα δεδομένα αλλά τα αποτελέσματα της συνάρτησης ήταν καλύτερα για αυτό και την επιλέξαμε.

Εκπαίδευση TSK μοντέλων

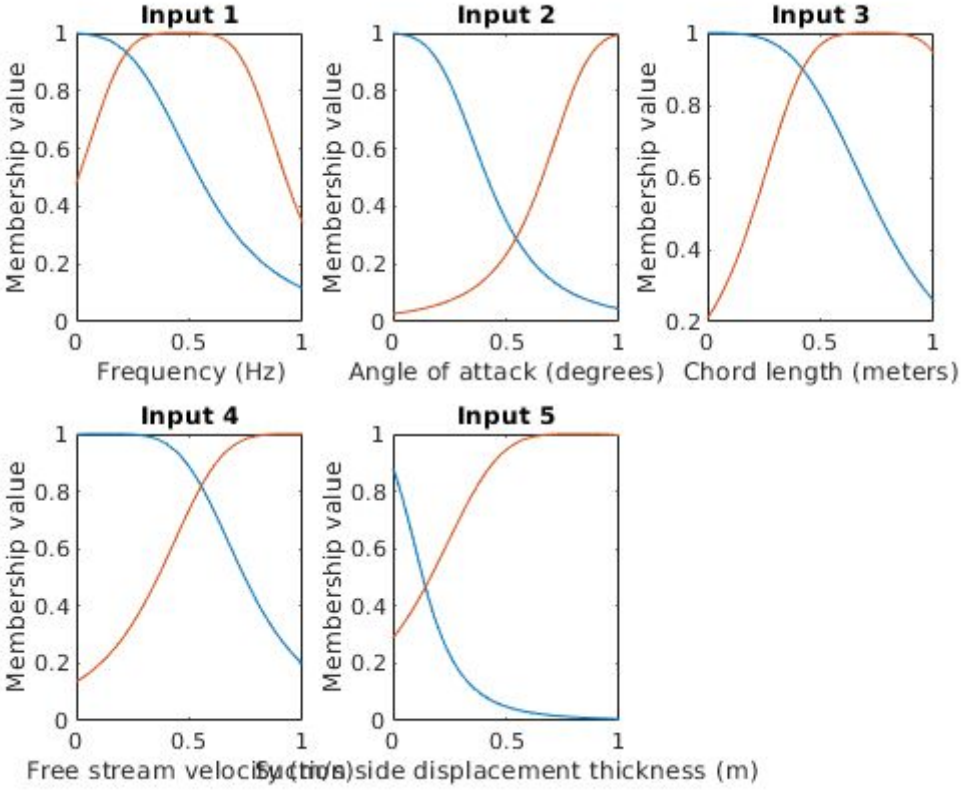
Αφού πραγματοποιήθηκε ο διαχωρισμός στη συνέχεια υλοποιούμε την εκπαίδευση των τεσσάρων διαφορετικών μοντέλων που περιγράφονται στην εκφώνηση. Τα μοντέλα διαφέρουν ως προς τη μορφή της εξόδου και τον αριθμό των συναρτήσεων συμμετοχής και

δημιουργούνται με την εντολή genfis1. Στο πρόγραμμα μας αποθηκεύουμε κάθε μοντέλο στον πίνακα ‘m’ έτσι ώστε να μπορούμε να εκτελέσουμε την ίδια διαδικασία εκπαίδευσης τέσσερις φορές. Οι εποχές που χρησιμοποιήθηκαν ήταν 100 και για την εκπαίδευση χρησιμοποιήθηκε η εντολή anfis ρυθμίζοντας τις παραμέτρους με την εντολή anfisOptions. Για την πρόβλεψη του αποτελέσματος χρησιμοποιείται η εντολή evalfis. Παρακάτω παρουσιάζονται τα αποτελέσματα των τεσσάρων μοντέλων. Όλα τα μοντέλα έχουν bell-shaped συνάρτηση συμμετοχής.

Μοντέλο 1

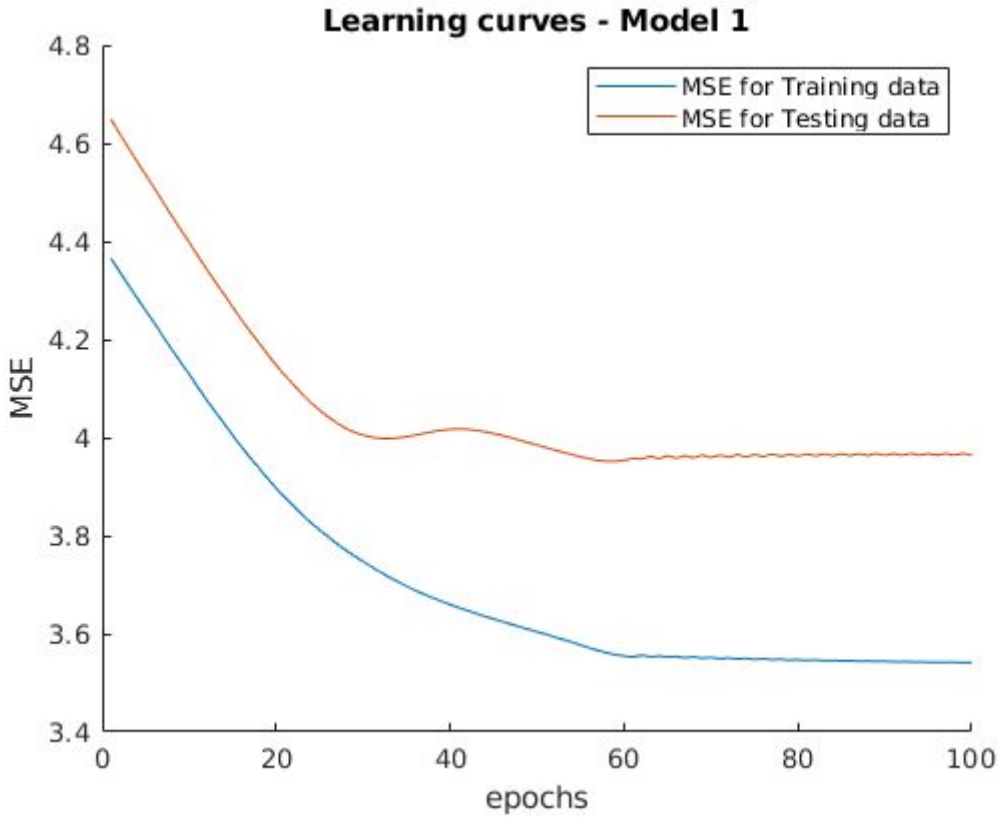
- Αριθμός εισόδων: 2
- Μορφή εξόδου: Singleton

Συναρτήσεις Συμμετοχής



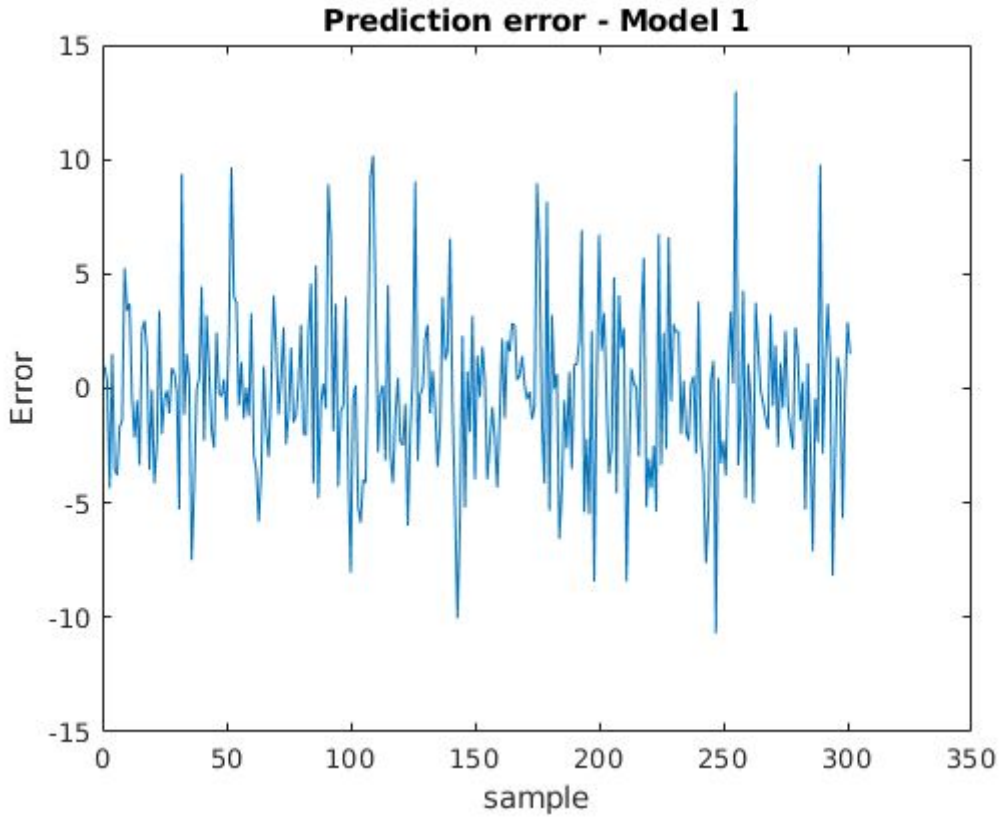
Εικόνα 1: Συναρτήσεις συμμετοχής των 5 εισόδων.

Καμπύλες μάθησης



Εικόνα 2: Καμπύλες μάθησης μοντέλου 1.

Σφάλματα πρόβλεψης

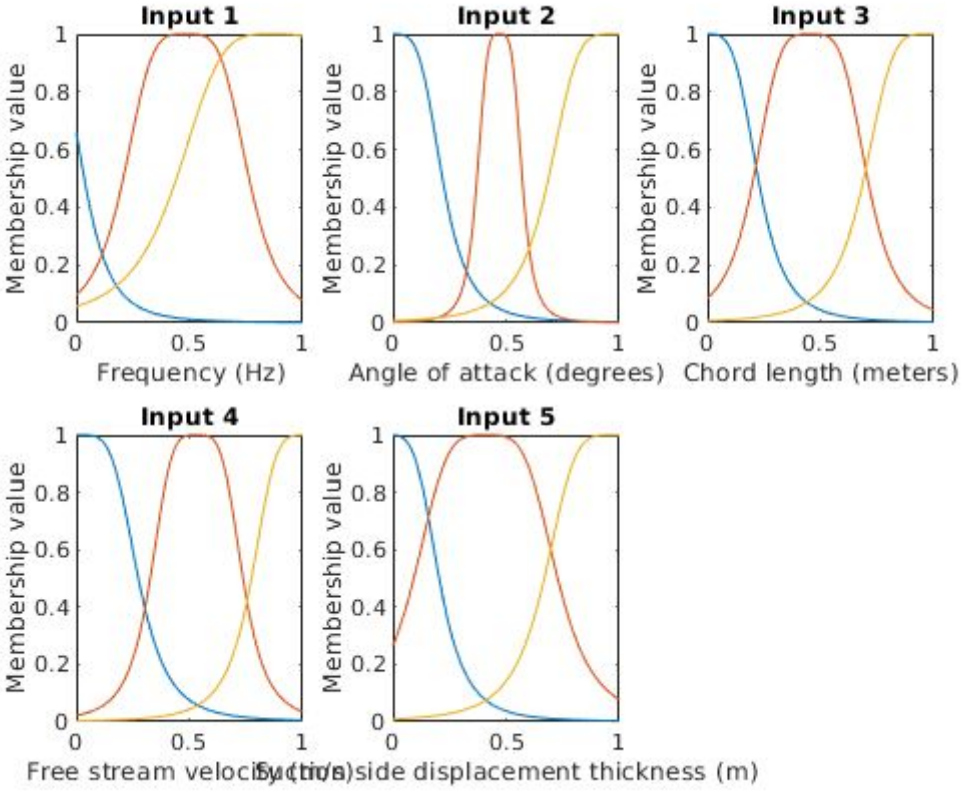


Εικόνα 3: Σφάλμα πρόβλεψης μοντέλου 1.

Μοντέλο 2

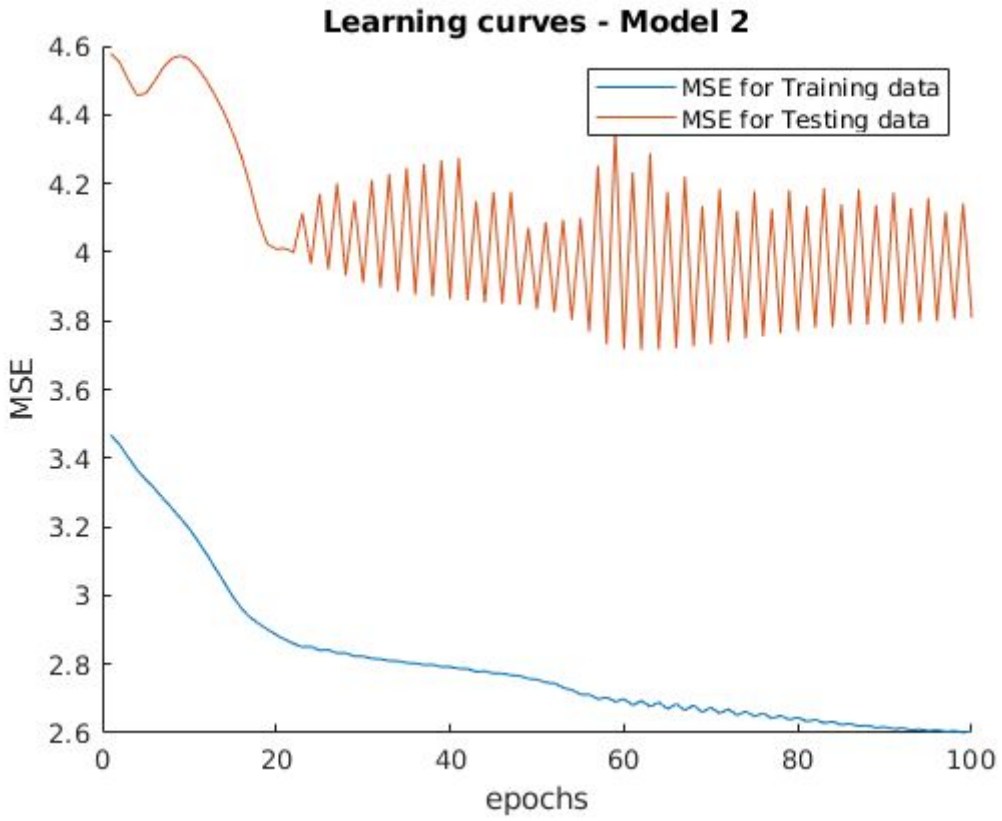
- Αριθμός εισόδων: 3
- Μορφή εξόδου: Singleton

Συναρτήσεις Συμμετοχής



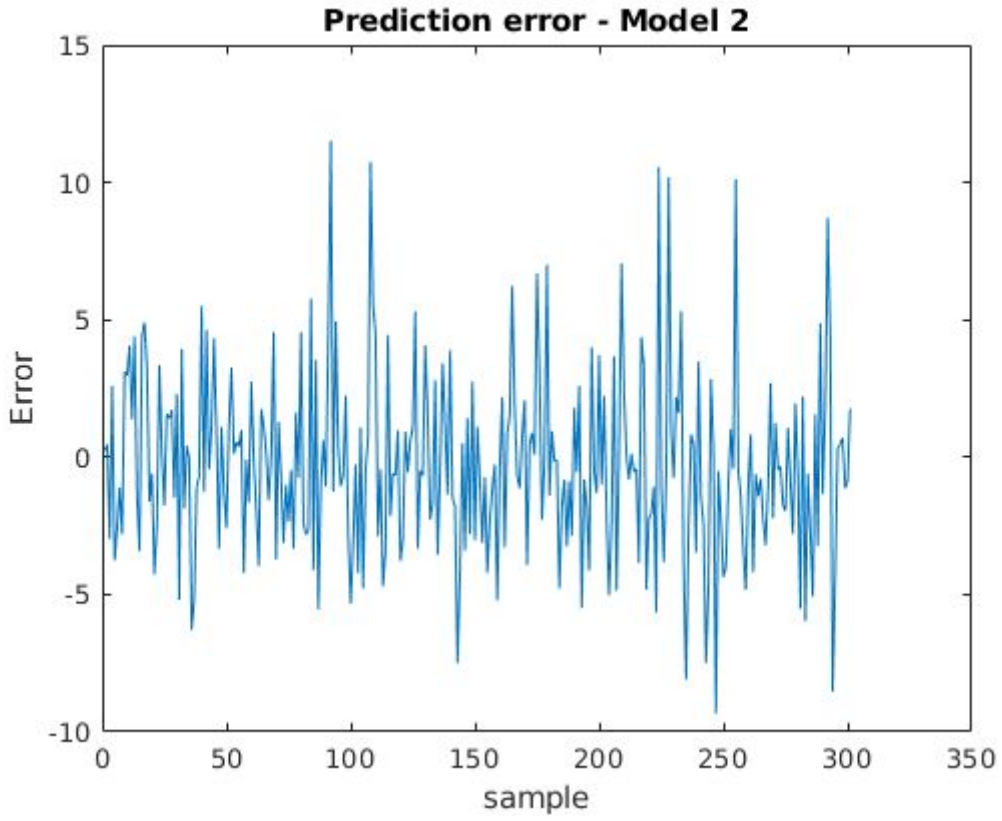
Εικόνα 4: Συναρτήσεις συμμετοχής των 5 εισόδων.

Καμπύλες μάθησης



Εικόνα 5: Καμπύλες μάθησης μοντέλου 2.

Σφάλματα πρόβλεψης

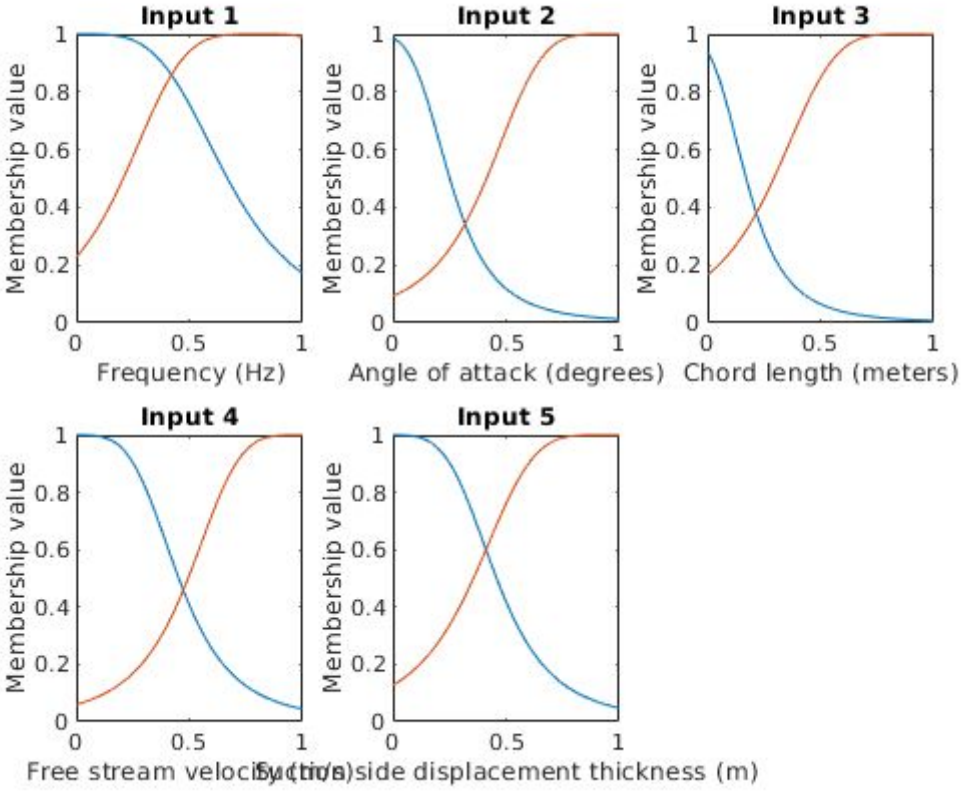


Εικόνα 6: Σφάλμα πρόβλεψης μοντέλου 2.

Μοντέλο 3

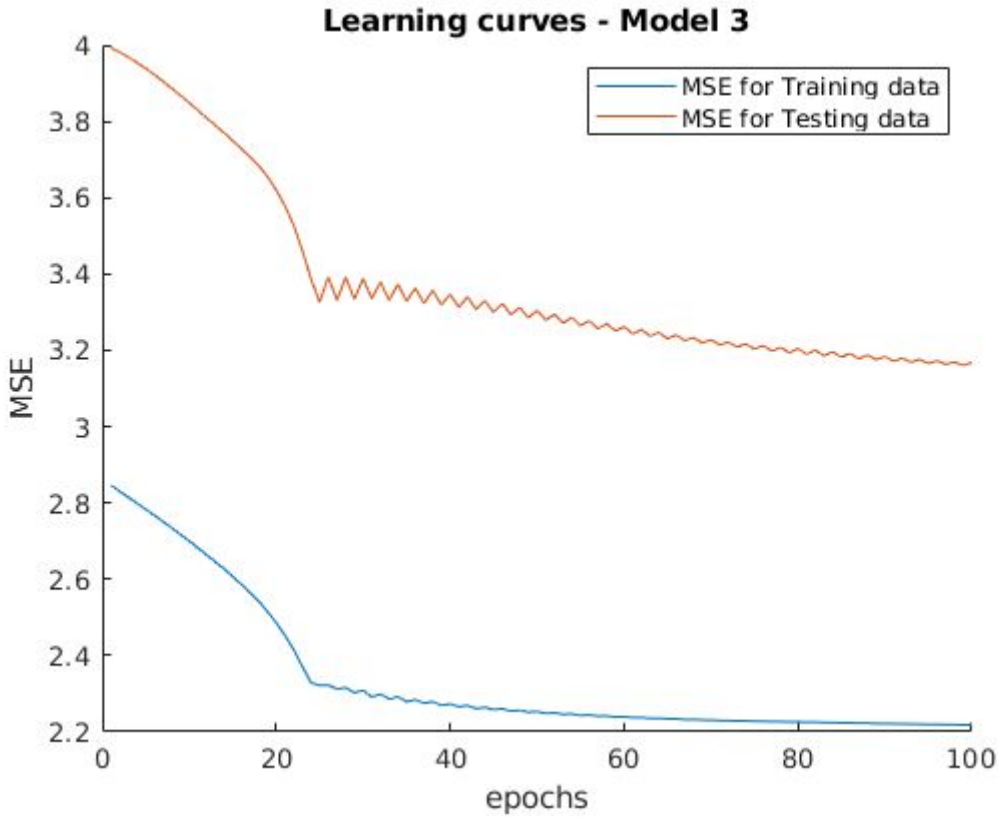
- Αριθμός εισόδων: 2
- Μορφή εξόδου: Polynomial

Συναρτήσεις Συμμετοχής



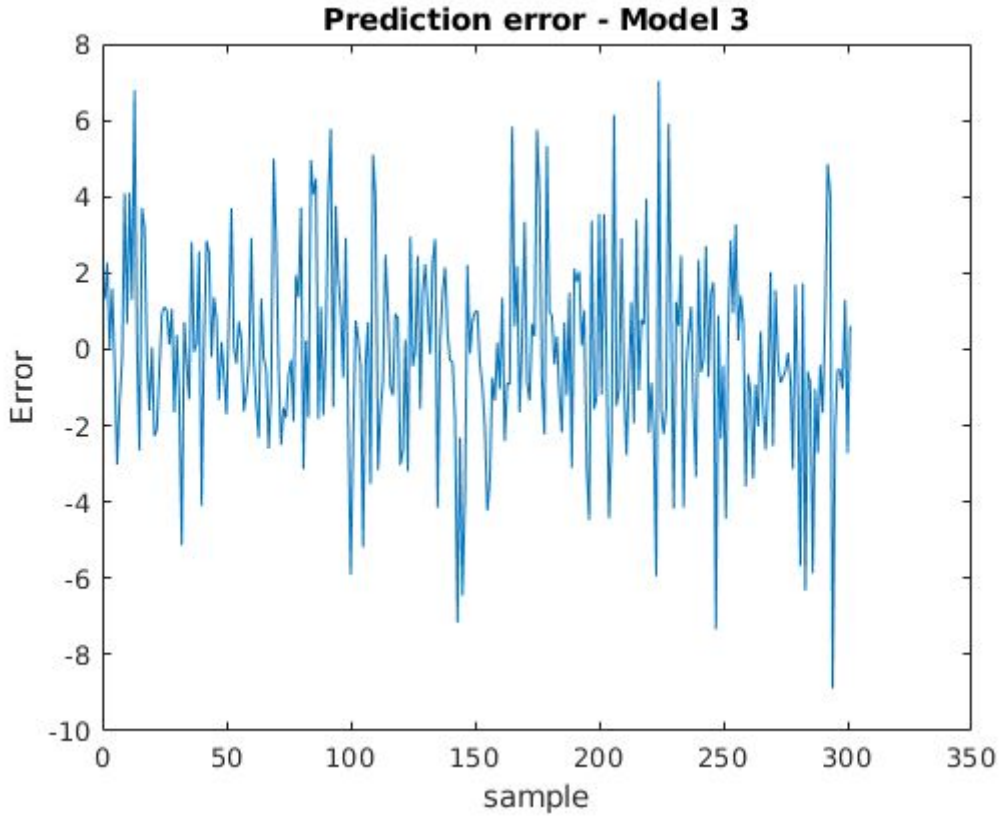
Εικόνα 7: Συναρτήσεις συμμετοχής των 5 εισόδων.

Καμπύλες μάθησης



Εικόνα 8: Καμπύλες μάθησης μοντέλου 3.

Σφάλματα πρόβλεψης

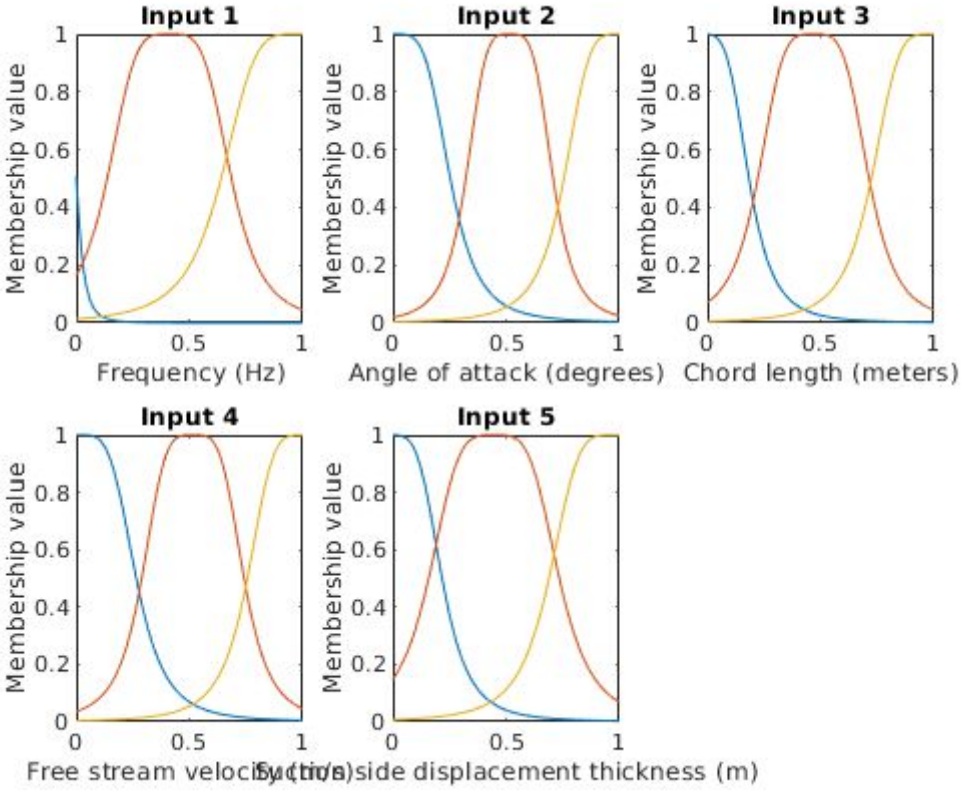


Εικόνα 9: Σφάλμα πρόβλεψης μοντέλου 3.

Μοντέλο 4

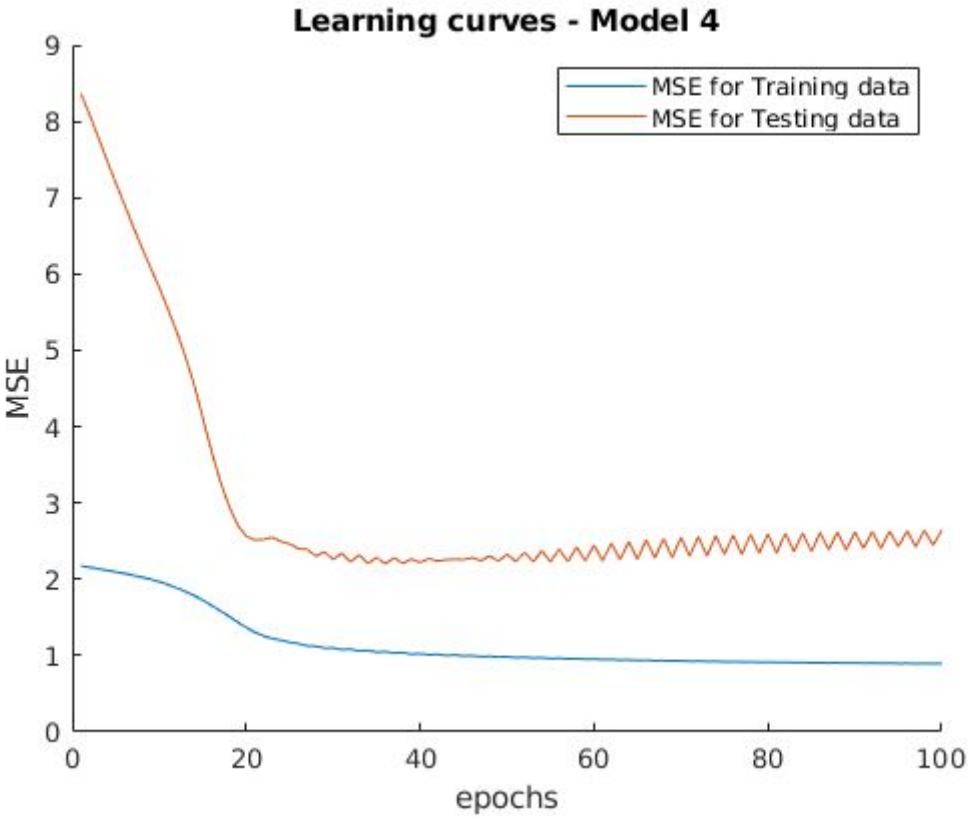
- Αριθμός εισόδων: 3
- Μορφή εξόδου: Polynomial

Συναρτήσεις Συμμετοχής



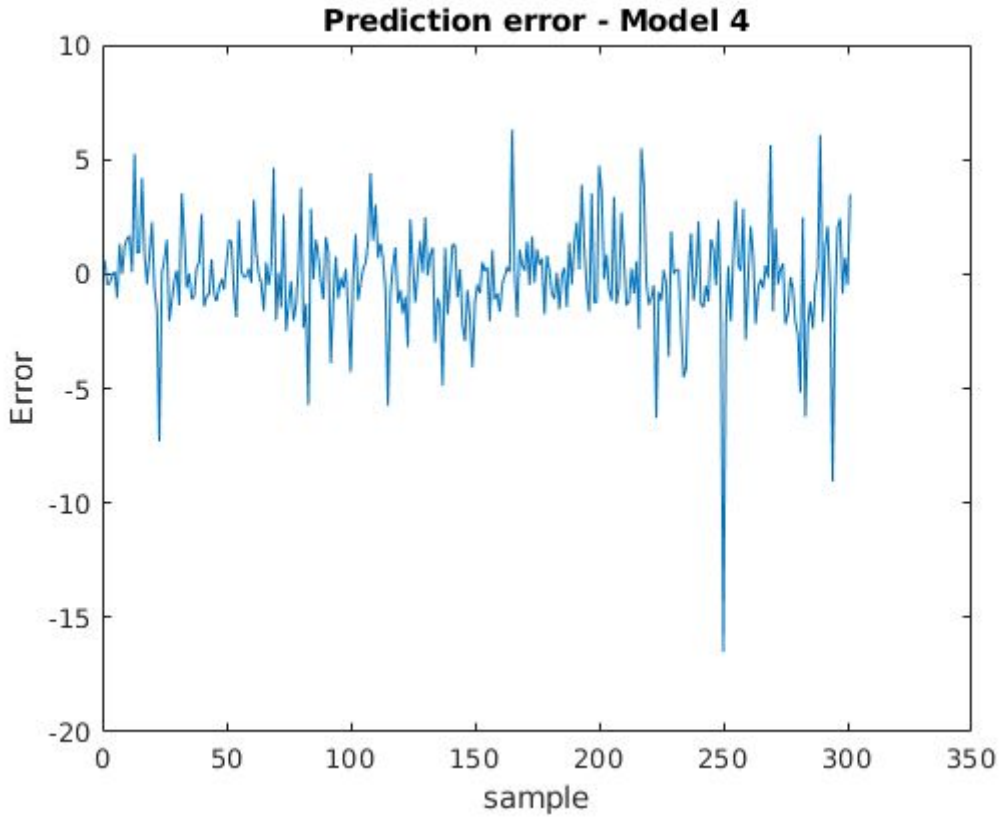
Εικόνα 10: Συναρτήσεις συμμετοχής των 5 εισόδων.

Καμπύλες μάθησης



Εικόνα 11: Καμπύλες μάθησης μοντέλου 4.

Σφάλματα πρόβλεψης



Εικόνα 12: Σφάλμα πρόβλεψης μοντέλου 4.

Μετρικές

Μοντέλο	Συναρτήσεις Συμμετοχής	Έξοδος	MSE	RMSE	R^2	NMSE	NDEI
1	2	Singleton	13.1825	3.6308	0.7006	0.2994	0.5472
2	3	Singleton	10.8953	3.3008	0.7526	0.2474	0.4974
3	2	Polynomial	6.3221	2.5367	0.8539	0.1461	0.3823
4	3	Polynomial	5.0512	2.2475	0.8853	0.1147	0.3387

Πίνακας 1: Μετρικές αξιολόγησης.

Σχολιασμός

Στην εκφώνηση μας ζητείται να σχολιάσουμε τα αποτελέσματα που πήραμε με βάση τη μορφή της εξόδου και τη διαμέριση του χώρου. Παρατηρούμε από τον πίνακα με τις μετρικές αξιολόγησης ότι με βάση την μορφή της εξόδου, τα δύο μοντέλα που έχουν πολυωνυμική έξοδο υπερτερούν σε ακρίβεια από τα μοντέλα που έχουν έξοδο singleton καθώς όλες οι μετρικές σφάλματος (MSE, RMSE, NMSE και NDEI) έχουν μικρότερη τιμή στα μοντέλα με πολυωνυμική έξοδο από ότι στα μοντέλα με έξοδο singleton. Σε ότι αφορά τον αριθμό των εισόδων παρατηρούμε ότι μεταξύ των μοντέλων με έξοδο Singleton αλλά και των μοντέλων με έξοδο πολυωνυμική υπερτερούν τα μοντέλα με περισσότερες εισόδους. Συνολικά, μπορεί να βγεί το συμπέρασμα ότι τον κύριο ρόλο τον έχει η μορφή της εξόδου καθώς τα μοντέλα με πολυωνυμικές εξόδους είναι καλύτερα από τα μοντέλα με έξοδο Singleton ανεξάρτητα από τον αριθμό των εισόδων. Σε δεύτερη φάση, ο μεγαλύτερος αριθμός εισόδων έχει καλύτερα αποτελέσματα ανάμεσα στα μοντέλα ίδιας εξόδου.
Σε ότι αφορά την υπερεκπαίδευση παρατηρούμε ότι στα μοντέλα 1,2,3 το σφάλμα του συνόλου ελέγχου μειώνεται συνεχώς μέχρι περίπου τις 20 εποχές όπου και συγκλίνει σε μία σταθερή τιμή και μειώνεται με πολύ αργό ρυθμό. Στο διάγραμμα του μοντέλου 2 μπορούμε να παρατηρήσουμε ότι το σφάλμα παρουσιάζει μία έντονη ταλάντωση ωστόσο δεν μπορεί να χαρακτηριστεί υπερεκπαιδευση καθώς δεν αποκλίνει. Η μοναδική περίπτωση που μπορεί να θεωρηθεί ότι παρουσιάζει υπερεκπαίδευση είναι το μοντέλο 4 στο οποίο η καμπύλη μάθησης του σφάλματος πρόβλεψης (πορτοκαλί καμπύλη) παρουσιάζει μια μικρή αυξητική πορεία δηλώνοντας ότι το σφάλμα αυξάνεται και κατ επέκταση ότι το μοντέλο δεν μπορεί να γενικεύσει τα αποτελέσματα στο σετ ελέγχου.

Μέρος 2 - Superconductivity dataset

Περιγραφή - Διαχωρισμός dataset

Το dataset που επιλέγεται στο δεύτερο μέρος της εργασίας έχει υψηλότερο βαθμό διαστασιμότητας και αποτελείται από 21263 δείγματα και 81 χαρακτηριστικά. Επειδή ο αριθμός δειγμάτων και χαρακτηριστικών είναι πολύ μεγάλος δεν είναι δυνατόν να χρησιμοποιηθεί η μέθοδος του πρώτου μέρους γιατί ο αριθμός κανόνων θα αυξάνονταν σε πολύ μεγάλο βαθμό. Για να αποφευχθεί αυτό θα χρησιμοποιήσουμε τη μεθοδο της επιλογής χαρακτηριστικών και της διαμέρισης του χώρου οι οποίες βέβαια εισάγουν δύο ελεύθερες μεταβλητές στο πρόβλημα (αριθμός χαρακτηριστικών προς επιλογή και αριθμός ομάδων που δημιουργούνται) οι οποίες θα προσεγγιστούν με τη μέθοδο αναζήτησης πλέγματος. Ο διαχωρισμός του dataset πραγματοποιείται και σε αυτή την περίπτωση με την έτοιμη συνάρτηση split_scale που μας παρέχετε. Το dataset χωρίζεται κατά 60%, 20%, 20% για τα training, validation και testing σετ αντίστοιχα.

Περιγραφή διαδικασίας

Αρχικά επιλέγονται οι τιμές για τον αριθμό των χαρακτηριστικών (features number) και οι τιμές για την ακτίνα των clusters (cluster radius). Οι δύο πίνακες με τις τιμές τους είναι οι εξής:

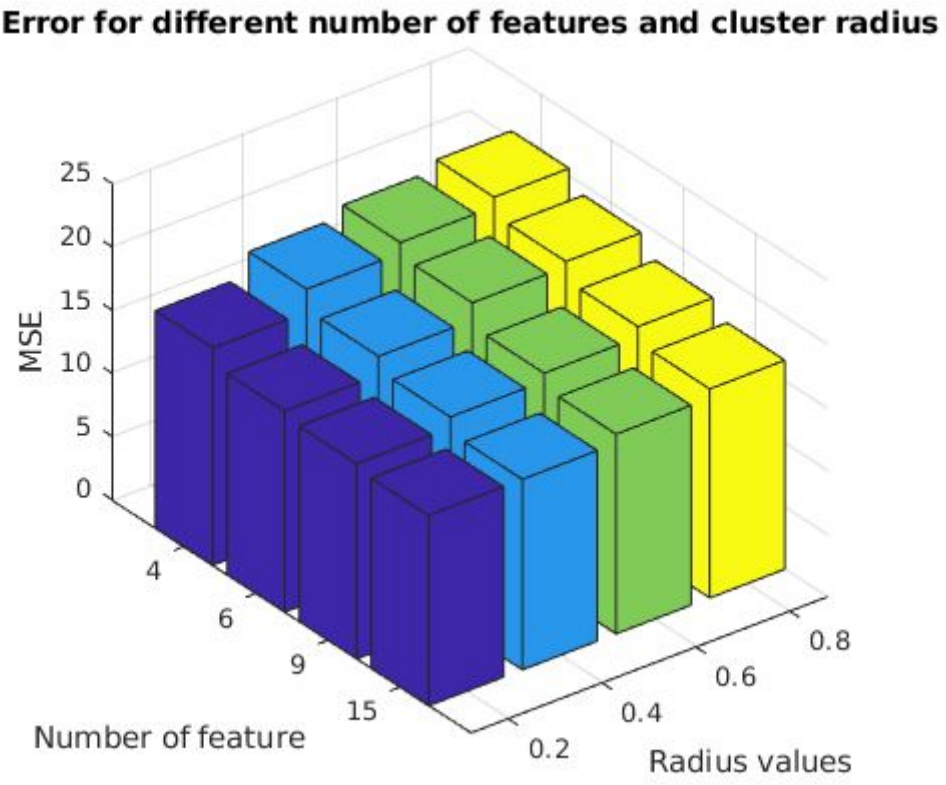
- features_number = [4, 6, 9, 15]
- cluster_radius = [0.2, 0.4, 0.6, 0.8]

Η αξιολόγηση μέσω διασταυρωμένης επικύρωσης γίνεται με την συνάρτηση cvpartition του MATLAB επιλέγοντας 5 folds. Για την αξιολόγηση τον χαρακτηριστικών χρησιμοποιείται ο αλγόριθμος relief χρησιμοποιώντας την έτοιμη συνάρτηση του MATLAB η οποία επιστρέφει στον πίνακα idx τα πιο σημαντικά χαρακτηριστικά και στον πίνακα weights τα αντίστοιχα βάρη. Έπειτα για τη βέλτιστη επιλογή παραμέτρων ξεκινάει μια σειρά από τρία loops. Η πρώτη επανάληψη τρέχει για κάθε αριθμό από features, η δεύτερη τρέχει για κάθε ακτίνα των clusters και η τρίτη για κάθε fold. Μέσα στην τριπλή loop εκπαιδεύουμε ένα fis model για κάθε fold του αλγορίθμου. Η εκπαίδευση πραγματοποιείται για κάθε fold του αλγορίθμου και τρέχει για 50 εποχές. Στη συνέχεια αποθηκεύουμε για κάθε feature και radius το μέσο σφάλμα. Όταν τελειώσει η τριπλή επανάληψη επιλέγουμε το βέλτιστο μοντέλο με βάση το ελάχιστο σφάλμα. Στο τέλος εκπαιδεύουμε το βέλτιστο μοντέλο και υπολογίζουμε τις αντίστοιχες μετρικές. Για την απεικόνιση των διαφόρων διαγραμμάτων χρησιμοποιούμε το script “plot_script.m”.

Διαγράμματα

Παρακάτω στην Εικόνα 13 παρουσιάζεται το διάγραμμα μπάρας σε συνάρτηση με τον αριθμό χαρακτηριστικών και των διάφορων ακτινών. Επίσης, στην εικόνα 14 παρουσιάζεται ο αριθμός των κανόνων σε συνάρτηση με τις ίδιες ποσότητες.

Διάγραμμα σφαλμάτων



Εικόνα 13: Διάγραμμα μπάρας για το μέσο τετραγωνικό σφάλμα σε συνάρτηση με τον αριθμό χαρακτηριστικών και τις ακτίνες.

Στον πίνακα 2 παρουσιάζονται τα αποτελέσματα των μέσων τετραγωνικών σφαλμάτων για τους διάφορους συνδυασμούς χαρακτηριστικών και ακτινών.

Radius Features	0.2	0.4	0.6	0.8
4	17.18377	18.99633	19.83665	20.58821
6	15.90612	17.39158	18.78277	19.19771
9	15.44827	16.29878	16.91477	17.75959
15	15.05695	15.06847	15.79888	16.53030

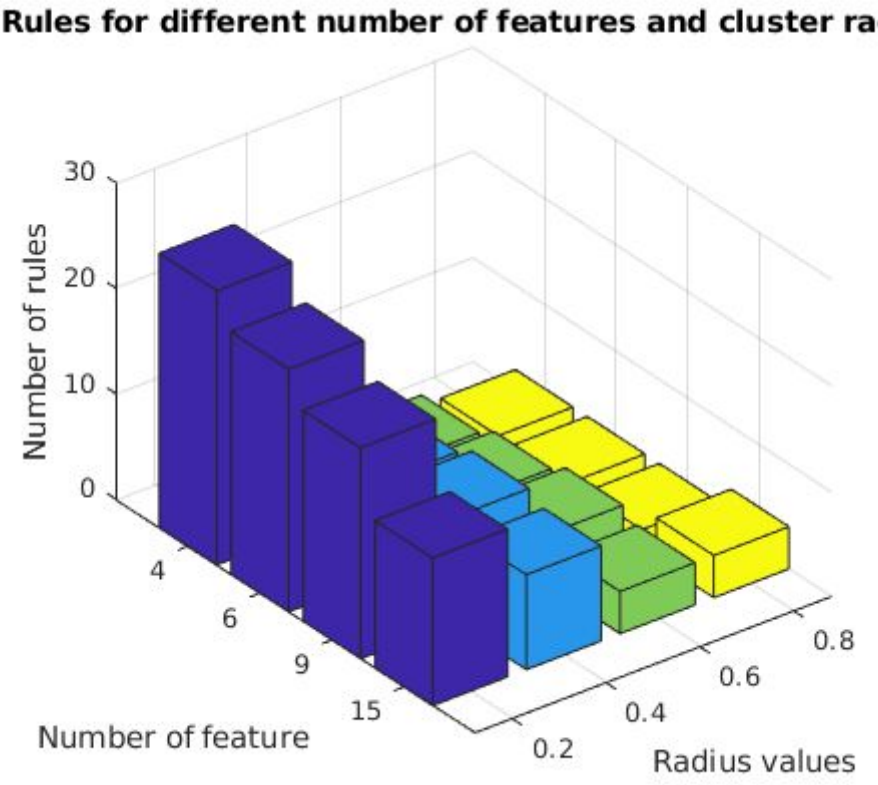
Πίνακας 2: Αποτελέσματα των μέσων τετραγωνικών σφαλμάτων για τους διάφορους συνδυασμούς χαρακτηριστικών και ακτινών.

Σχολιασμός Αποτελεσμάτων Πίνακα

Από τον πίνακα των σφαλμάτων γίνεται αντιληπτό ότι ότι για μεγαλύτερο αριθμό χαρακτηριστικών έχουμε μικρότερο σφάλμα όταν η ακτίνα είναι σταθερή. Από την άλλη πλευρά παρατηρούμε ότι όσο κρατάμε σταθερό τον αριθμό των χαρακτηριστικών έχουμε μικρότερο σφάλμα για μικρότερες τιμές ακτινών. Η βέλτιστη τιμή εντοπίζεται στο πορτοκαλι

κελί για ακτίνα 0.2 και αριθμό χαρακτηριστικών 15. Επίσης, αρκετά καλά αποτελέσματα κοντά στη βέλτιστη τιμή παρουσιάζουν και οι εξής συνδυασμοί χαρακτηριστικών - ακτίνας οι οποίοι φαίνονται με μπλε χρώμα: (15, 0.4) (9, 0.2).

Διάγραμμα κανόνων

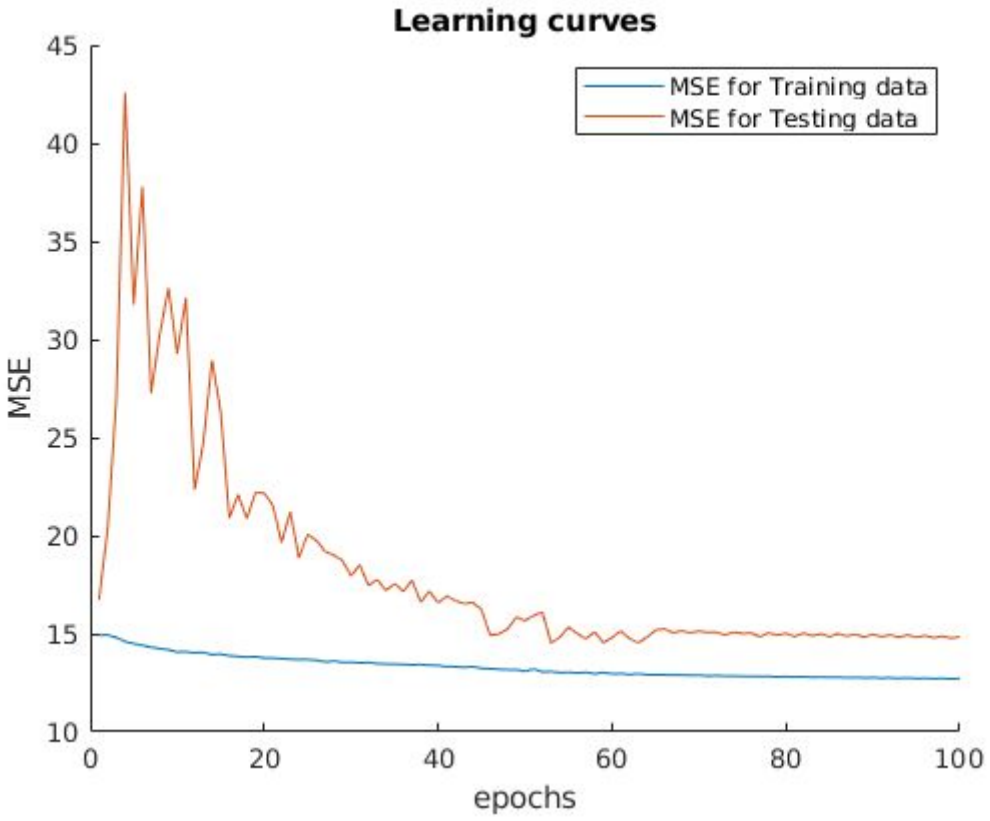


Εικόνα 14: Διάγραμμα μπάρας για τον αριθμό κανόνων σε συνάρτηση με τον αριθμό χαρακτηριστικών και τις ακτίνες.

Σχολιασμός διαγραμμάτων κανόνων

Από το διάγραμμα κανόνων φαίνεται ότι για μικρή τιμή της ακτίνας και μικρό αριθμό χαρακτηριστικών έχουμε μεγάλο αριθμό κανόνων γεγονός που ενδεχομένως να μην μας επιτρέπει την εκτέλεση του αλγορίθμου λόγω μεγάλου χρόνου εκτέλεσης. Δεδομένου του αριθμού των χαρακτηριστικών όσο η τιμή της ακτίνας αυξάνεται ο αριθμός των κανόνων μειώνεται. Δεδομένης της ακτίνας, εκτός από την περίπτωση της ακτίνας ίσης με 0.2, στις υπόλοιπες περιπτώσεις ο αριθμός των κανόνων που δημιουργείται φαίνεται περίπου ίδιος για όλες τις τιμές των χαρακτηριστικών. Στην περίπτωση που η ακτίνα είναι 0.2 τότε αυξάνοντας τον αριθμό των χαρακτηριστικών ο αριθμός των κανόνων μειώνεται.

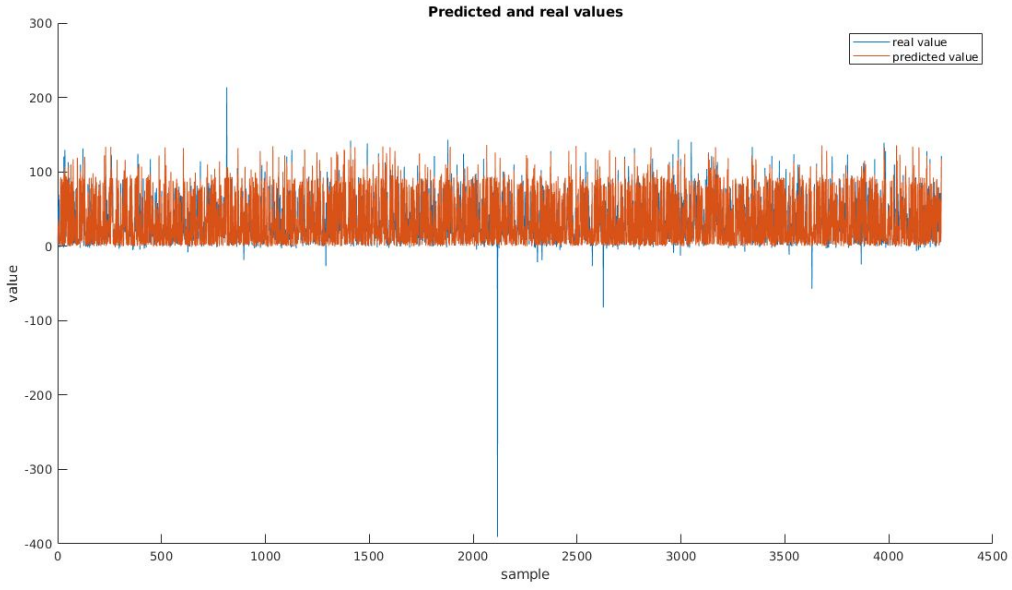
Καμπύλες μάθησης



Εικόνα 15: Καμπύλη μάθησης

Στο διάγραμμα των καμπυλών μάθησης παρατηρείται ότι το σφάλμα μειώνεται συνεχώς συγκλίνοντας περίπου στις 50 εποχές όπου αρχίζει και μειώνεται με πολύ μικρό ρυθμό.

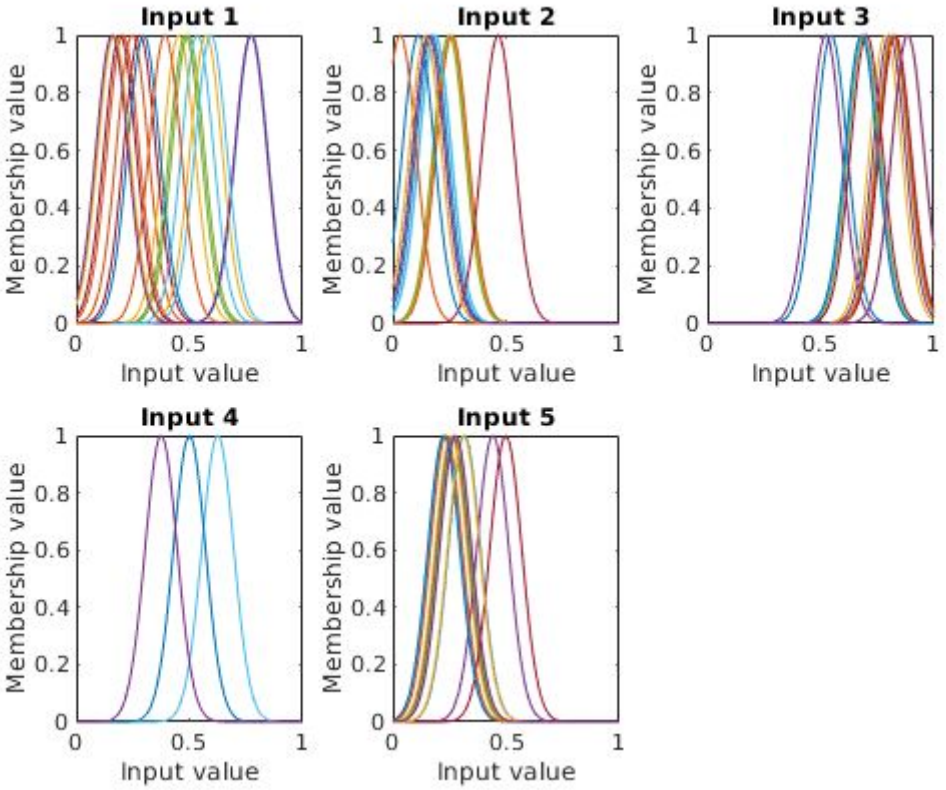
Διάγραμμα πραγματικών και προβλέψιμων τιμών



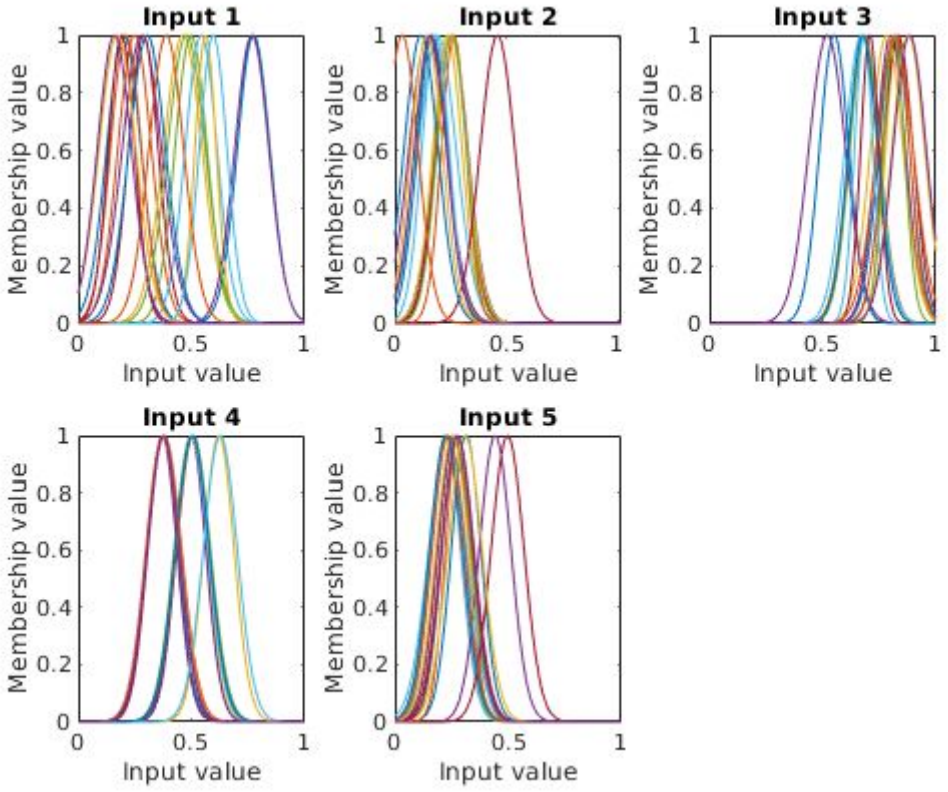
Εικόνα 16: Διάγραμμα πραγματικών και προβλέψιμων τιμών.

Παρακάτω παρουσιάζονται μερικά από τα αρχικά και τελικά σύνολα που προκύπτουν μετά τη βελτιστοποίηση έτσι ώστε να συγκριθούν οι μορφές τους.

Αρχικά Σύνολα



Τελικά Σύνολα



Μετρικές

RMSE	NMSE	NDEI	R^2
15.5296	0.2037	0.4513	0.7963

Παρατηρούμε στον πίνακα με τις μετρικές αξιολόγησης ότι πέρα από το δείκτη RMSE ο οποίος είναι σχετικά μεγάλος οι υπόλοιπες μετρικές είναι μικρές και πολύ ικανοποιητικές. Σε σχέση με την υλοποίηση του πρώτου μέρους παρατηρούμε ότι τα αποτελέσματα είναι χειρότερα καθώς έχουμε μεγαλύτερες τιμές σφαλμάτων. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι στο πρώτο μέρος χρησιμοποιούνται όλοι οι είσοδοι που δίνονται στο dataset ενώ εδώ μόνο ένα μέρος αυτών καθώς έχουμε πολύ περισσότερα χαρακτηριστικά. Σχετικά με την απόδοση του συστήματος παρατηρούμε ότι είναι αρκετά ικανοποιητική καθώς χρησιμοποιούμε σχετικά μικρό αριθμό χαρακτηριστικών και κανόνων. Αν είχαμε επιλέξει τη μέθοδο του grid partitioning με 81 χαρακτηριστικά θα έπρεπε να χρησιμοποιηθούν 2^{81} και 3^{81} κανόνες για 2 και 3 εισόδους αντίστοιχα σε αντίθεση με τη δική μας περίπτωση που χρησιμοποιούνται μόνο 14 κανόνες. Σίγουρα, η ακρίβεια θα ήταν μεγαλύτερη αλλά ο χρόνος εκτέλεσης θα ήταν απαγορευτικός. Συμπεραίνουμε λοιπόν ότι δεδομένου του μικρού

αριθμού κανόνων που χρησιμοποιούμε τα αποτελέσματα είναι αρκετά ικανοποιητικά ως προς την ακρίβεια και τον χρόνο εκτέλεσης.