Regression

1. Εφαρμογή σε απλό dataset

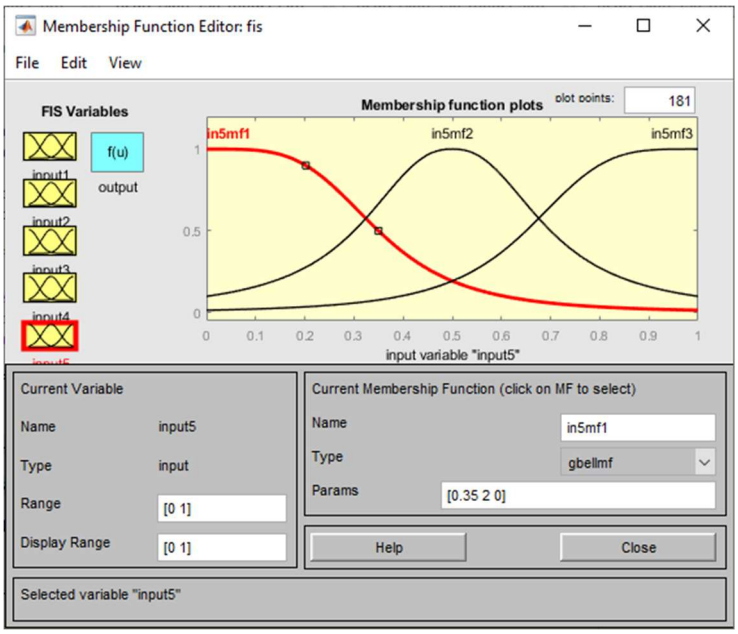
Το τμήμα της εργασίας που αφορά σε αυτό το dataset έγινε με το *MATLAB **R2019a***.

Στα δεδομένα εφαρμόστηκε κανονικοποίηση στο διάστημα [0,1].

Για τα μοντέλα με 2 συναρτήσεις συμμετοχής, θεωρήθηκε πως η default αρχικοποίηση των συναρτήσεων συμμετοχής είναι επαρκής ώστε να θεωρηθεί ότι τα διαδοχικά ασαφή σύνολα της εισόδου παρουσιάζουν βαθμό επικάλυψης περίπου 0.5.

Για τα μοντέλα με 3 συναρτήσεις συμμετοχής, προκειμένου να θεωρηθεί ότι τα διαδοχικά ασαφή σύνολα της εισόδου παρουσιάζουν βαθμό επικάλυψης περίπου 0.5, οι παράμετροι των 3 συναρτήσεων συμμετοχής, αρχικοποιήθηκαν ως εξής:

Membership function 1: gbellmf με παραμέτρους [0.35 2 0]
Membership function 2: gbellmf με παραμέτρους [0.2 1.2 0.5]
Membership function 3: gbellmf με παραμέτρους [0.35 2 1]



Και τα 4 μοντέλα εκπαιδεύτηκαν για 500 εποχές, με error goal =0, step size = 0,01, decrease rate = 0,9, increase rate = 1,1.

Σε κάθε περίπτωση ως τελικό μοντέλο επιλέχθηκε εκείνο που αντιστοιχεί στο μικρότερο σφάλμα στο σύνολο επικύρωσης.

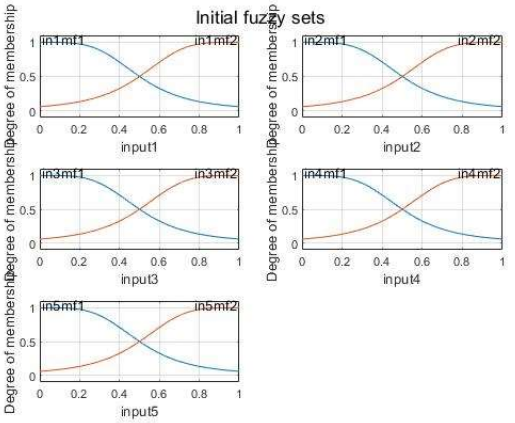
Να σημειωθεί ότι κατά την εκτέλεση της εργασίας, χάριν ευκολίας, τα scripts `Regression_TSK_model_2.m` και `Regression_TSK_model_4.m` εκτελέστηκαν τοποθετώντας

ένα log αμέσως μετά την εντολή `mfedit(fis) - line 45`, και πατώντας *continue* μετά την τροποποίηση των παραμέτρων των εισόδων στο GUI, για την εκτέλεση του υπολοίπου.

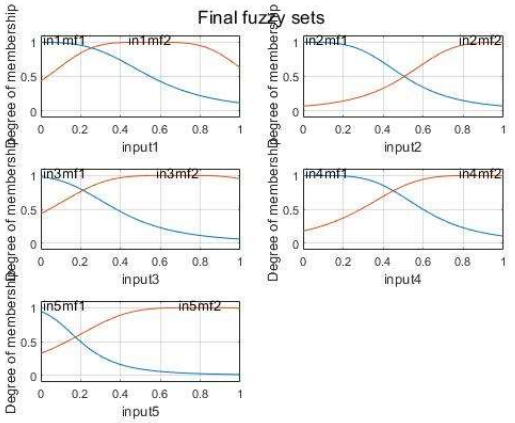
TSK_model_1: πλήθος συναρτήσεων συμμετοχής: 2, μορφή εξόδου: Singleton

5 είσοδοι, 1 έξοδος, 32 κανόνες

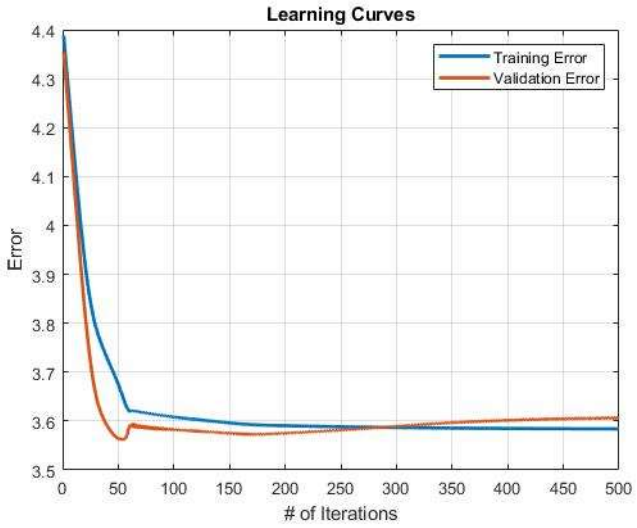
Εικόνα 1: Αρχικές μορφές των ασαφών συνόλων



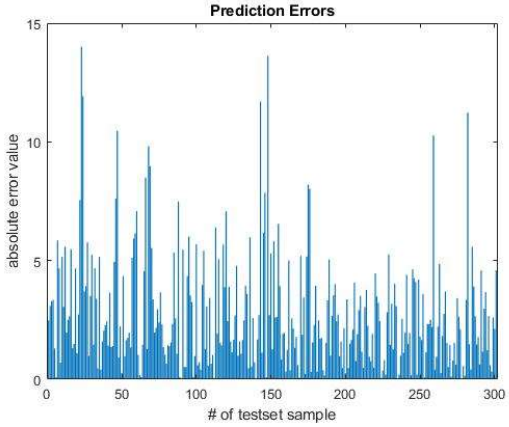
Εικόνα 2: Τελικές μορφές των ασαφών συνόλων



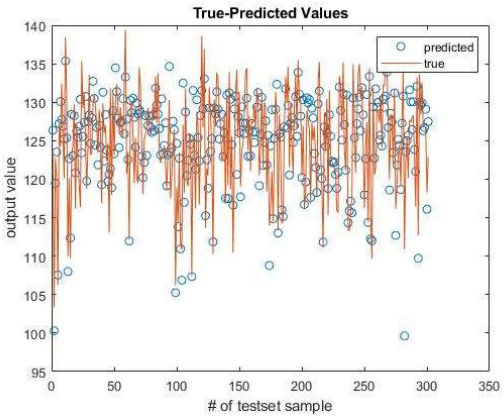
Εικόνα 3: Διάγραμμα μάθησης



Εικόνα 4: Σφάλματα πρόβλεψης κατ' απόλυτη τιμή



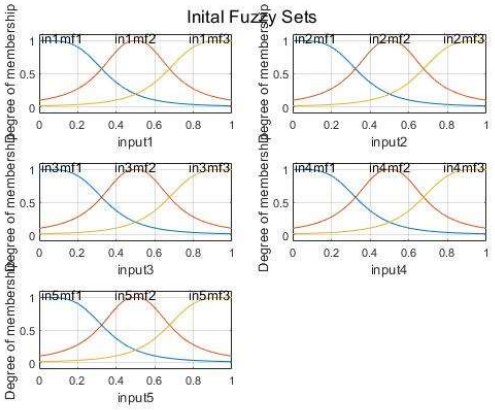
Εικόνα 5: Πραγματικές τιμές & τιμές πρόβλεψης



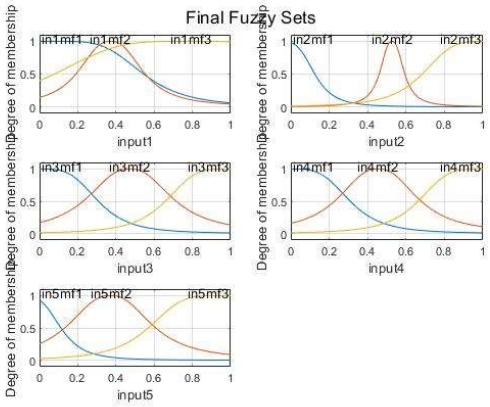
TSK_model_2: πλήθος συναρτήσεων συμμετοχής: 3, μορφή εξόδου: Singleton

5 είσοδοι, 1 έξοδος, 243 κανόνες

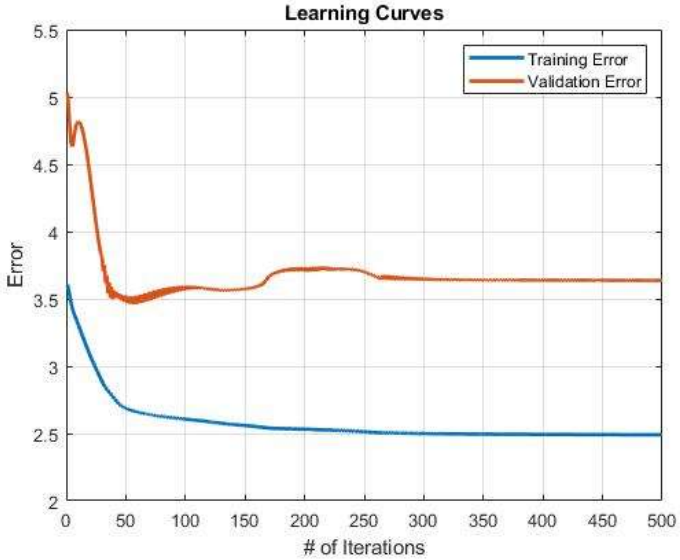
Εικόνα 2: Αρχικές μορφές των ασαφών συνόλων



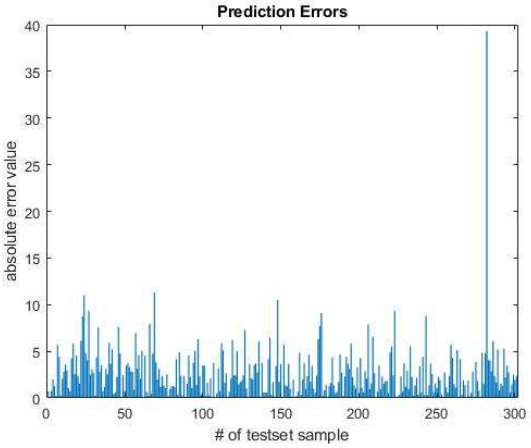
Εικόνα 2: Τελικές μορφές των ασαφών συνόλων



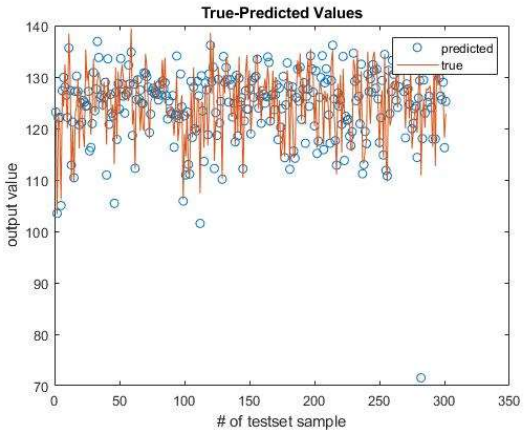
Εικόνα 3: Διάγραμμα μάθησης



Εικόνα 4: Σφάλματα πρόβλεψης κατ' απόλυτη τιμή



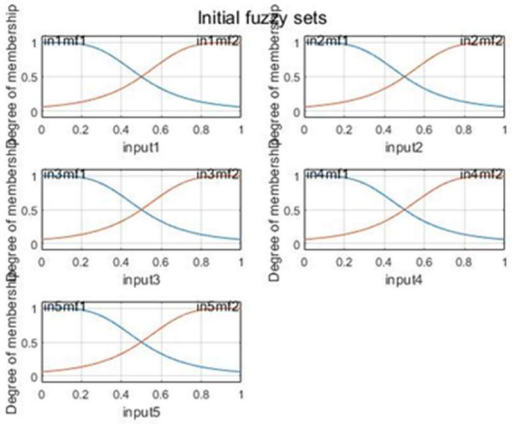
Εικόνα 5: Πραγματικές τιμές & τιμές πρόβλεψης



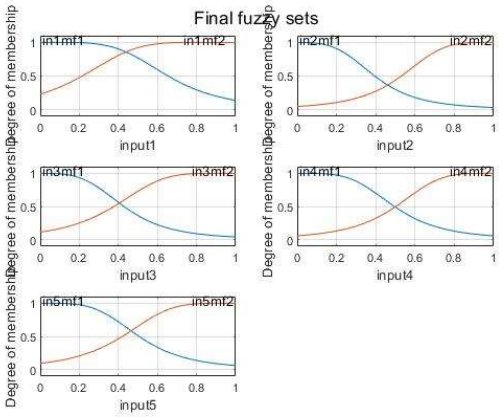
TSK_model_3: πλήθος συναρτήσεων συμμετοχής: 2, μορφή εξόδου: Polynomial

5 είσοδοι, 1 έξοδος, 32 κανόνες

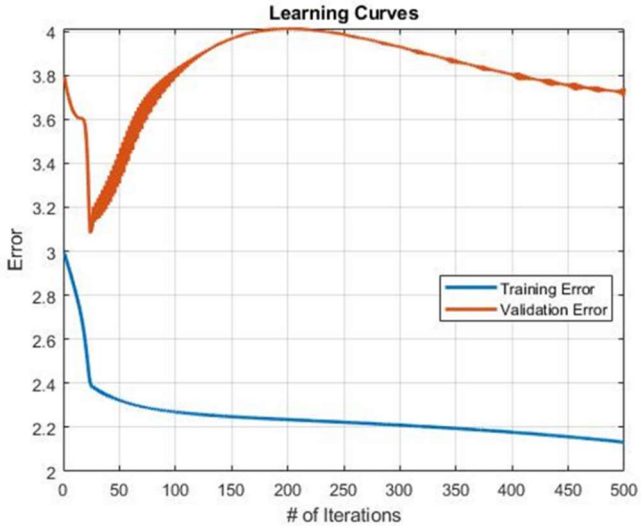
Εικόνα 3: Αρχικές μορφές των ασαφών συνόλων



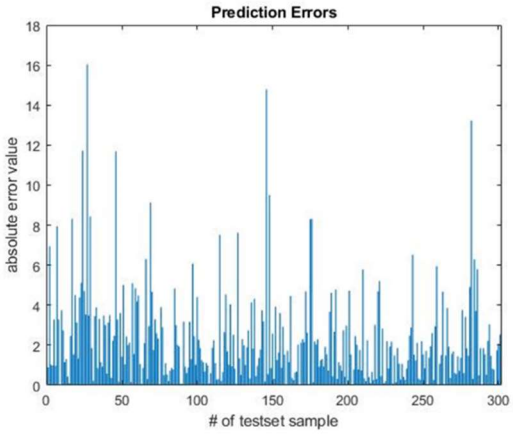
Εικόνα 2: Τελικές μορφές των ασαφών συνόλων



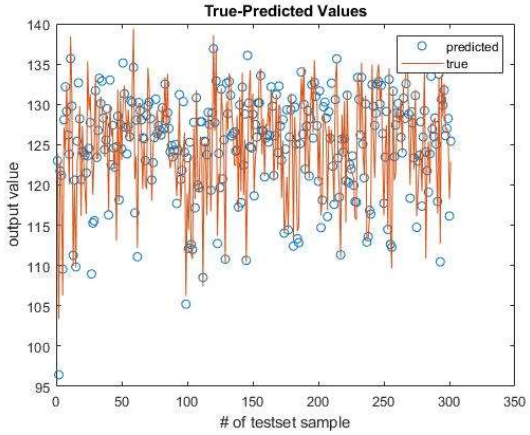
Εικόνα 3: Διάγραμμα μάθησης



Εικόνα 4: Σφάλματα πρόβλεψης κατ' απόλυτη τιμή



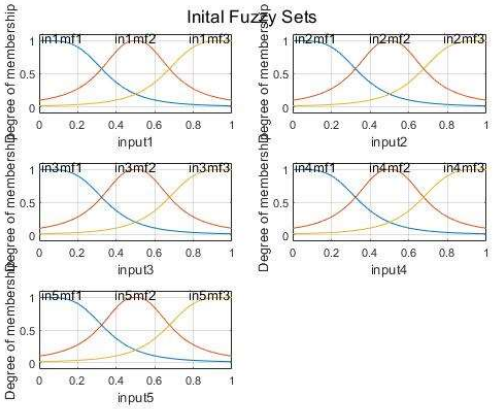
Εικόνα 5: Πραγματικές τιμές & τιμές πρόβλεψης



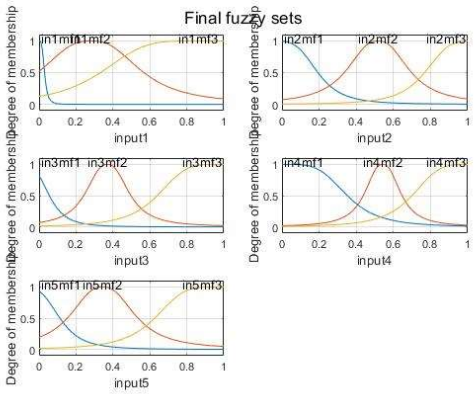
TSK_model_4: πλήθος συναρτήσεων συμμετοχής: 3, μορφή εξόδου: Polynomial

5 είσοδοι, 1 έξοδος, 243 κανόνες

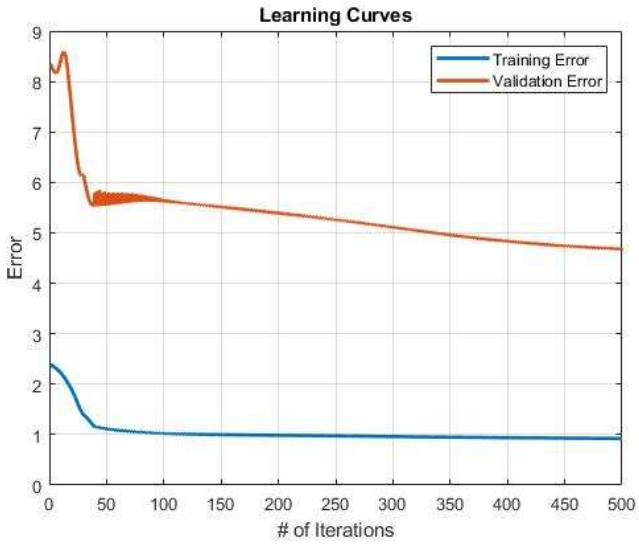
Εικόνα 4: Αρχικές μορφές των ασαφών συνόλων



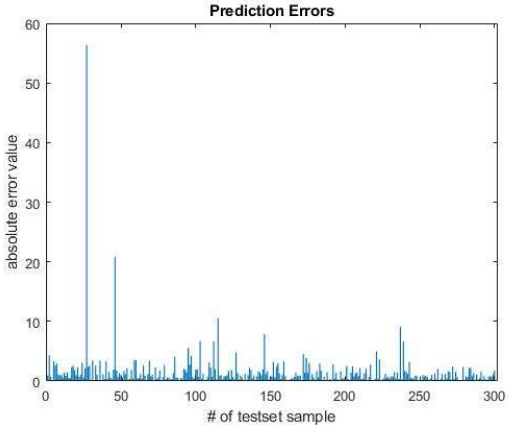
Εικόνα 2: Τελικές μορφές των ασαφών συνόλων



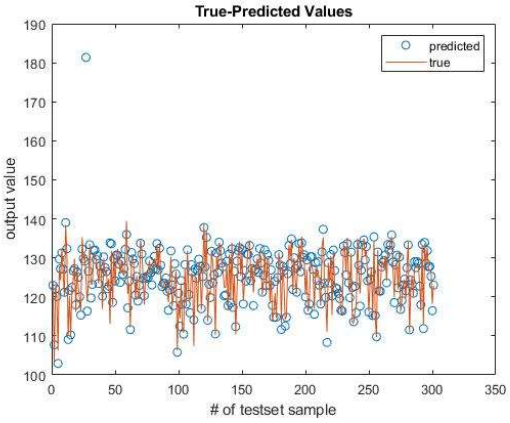
Εικόνα 3: Διάγραμμα μάθησης



Εικόνα 4: Σφάλματα πρόβλεψης κατ' απόλυτη τιμή



Εικόνα 5: Πραγματικές τιμές & τιμές πρόβλεψης



	TSK_model_1 **2, singleton**	*TSK_model_2* **3, singleton**	*TSK_model_3* **2, polynomial**	*TSK_model_4* **3, polynomial**
	# of rules = 32	# of rules = 243	# of rules = 32	# of rules = 243
RMSE	3.6578	4.0784	3.2868	3.9553
NMSE	0.30601	0.38042	0.24707	0.35781
NDEI	0.55318	0.61678	0.49706	0.59817
R^2	0.69399	0.61958	0.75293	0.64219

Το μοντέλο *TSK_model_*3, με διαμέριση κάθε εισόδου σε 2 ασαφή σύνολα, και συνεπώς 2^5 = 32 κανόνες, παρουσιάζει το μικρότερο σφάλμα (σύμφωνα και με τους τρεις δείκτες RMSE, NMSE και NDEI) και είναι το μόνο στο όποιο ο δείκτης R^2 (συντελεστής προσδιορισμού προσαρμογής) είναι μεγαλύτερος του 70%, συγκεκριμένα 75,3%.

Ως προς τον αριθμό των συναρτήσεων εισόδου : Τα μοντέλα *TSK_model_2* και *TSK_model_4*, στα οποία ο χώρος εισόδου διαμερίζεται σε 3 ασαφή σύνολα, παρουσιάζουν χειρότερη απόδοση σε όλους τους δείκτες από τα *TSK_model_1* και *TSK_model_3*, αντίστοιχα. Πράγματι λοιπόν μπορεί να παρατηρηθεί το φαινόμενο της υπερεκπαίδευσης, οι 32 κανόνες δομούν καλύτερο μοντέλο από τους 243.

Ως προς τη μορφή της εξόδου : Τα μοντέλα *TSK_model_3* και *TSK_model_4*, με εξόδους που μοντελοποιούνται από πολυωνυμική (*linear*) συνάρτηση παρουσιάζουν καλύτερη απόδοση σε όλους τους δείκτες από τα *TSK_model_1* και *TSK_model_2*, αντίστοιχα. Η έξοδος πολυωνυμικής μορφής επιτρέπει συνεπώς κατασκευή καλύτερου μοντέλου στην περίπτωσή μας, επιτρέπει μεγαλύτερη ακρίβεια σε αντίθεση με έξοδο της μορφής ενός σταθερού αριθμού που εδώ λειτουργεί περιοριστικά.

Βάσει των παραπάνω φαίνεται πως είναι λογικό το *TSK_model_3* να παρουσιάζει την καλύτερη απόδοση από τα 4 μοντέλα.

2. Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Το τμήμα της εργασίας που αφορά σε αυτό το dataset έγινε με το *MATLAB **R2018a***.

Το Superconductivity dataset περιλαμβάνει 21263 δείγματα με 81 μεταβλητές/χαρακτηριστικά.

Όπως αναφέρεται στην εκφώνηση, με 81 μεταβλητές αν διαμερίζαμε το χώρο εισόδου κάθε μεταβλητής με δύο ασαφή σύνολα, θα προέκυπταν 2^{81} κανόνες.

Υλοποίηση

Προεπεξεργασία δεδομένων: Ελέγχθηκε ότι το dataset Superconductivity δεν περιέχει NaN τιμές και στα δεδομένα εφαρμόστηκε κανονικοποίηση στο διάστημα [0,1].

Μέσω του αλγορίθμου *relieff* του Matlab, οι 81 μεταβλητές ταξινομήθηκαν με βάση το βαθμό σημαντικότητάς τους, και αποθηκεύτηκαν στη μεταβλητή *ranksReg*.

Το αρχικό dataset χωρίστηκε σε train set και test set, με αναλογία 80/20.

Αρχικοποιήθηκαν οι εξής μεταβλητές:

- Πλήθος εποχών εκπαίδευσης, epochs = 500
- Εύρος πλήθους χαρακτηριστικών που επιλεχθούν για την εκπαίδευση, G1 = [3,10]
- Εύρος τιμών της ακτίνας που θα χρησιμοποιηθεί για την ομαδοποίηση (clustering), G2 = [0.3, 0.4, 0.5, 0.6, 0.7, 0.8]

Η αναζήτηση πλέγματος (grid search) υλοποιήθηκε στο G1 x G2.

Για κάθε ζεύγος του G1xG2, έτρεξε 5-πτυχη διασταυρωμένη επικύρωση (5-fold cross validation).

Σε κάθε γύρο του cross validation,

1. το train set, χωρίζεται σε **trn set** και **val set** με αναλογία 80/20,

2. αρχικοποιείται μοντέλο FIS (*fis* του Matlab), με εισόδους το **trn set** (X & Y) και παραμέτρους 'SubtractiveClustering' και την ακτίνα που αντιστοιχεί στο τρέχον ζεύγος του G1xG2,
3. το μοντέλο εκπαιδεύεται με εισόδους τα **trn set** και **val set** για 500 εποχές, με step size = 0.01, decrease rate = 0.9, increase rate = 1.1
4. για το τρέχον μοντέλο αποθηκεύεται το ελάχιστο σφάλμα επικύρωσης, όπως αυτό επιστρέφεται από τη συνάρτηση *anfis*. Το σφάλμα αυτό αποθηκεύεται μαζί με το αντίστοιχο ζεύγος του G1xG2.

Αφού εκτελεστεί το grid search για όλα τα ζεύγη του G1xG2, επιλέγεται ως βέλτιστο το ζεύγος που αντιστοιχεί στο ελάχιστο σφάλμα.
Από το σύνολο train set (το 80% του αρχικού συνόλου), επιλέγονται τόσα χαρακτηριστικά όσα υποδεικνύει η τιμή του G1 από το ζεύγος που επιλέχθηκε ως το βέλτιστο. Τα χαρακτηριστικά αυτά επιλέγονται με βάση το πως αξιολογήθηκαν από τον αλγόριθμο *relieff* (να σημειωθεί ότι το ίδιο γίνεται και στο βήμα του cross-validation).

Ακολουθεί εκπαίδευση μοντέλου με είσοδο το παραπάνω σύνολο με τα βέλτιστα G1 και G2, και με ακριβώς τις ίδιες παραμέτρους όπως και στο βήμα του cross-validation (500 εποχές, 0.01 step size, κλπ).

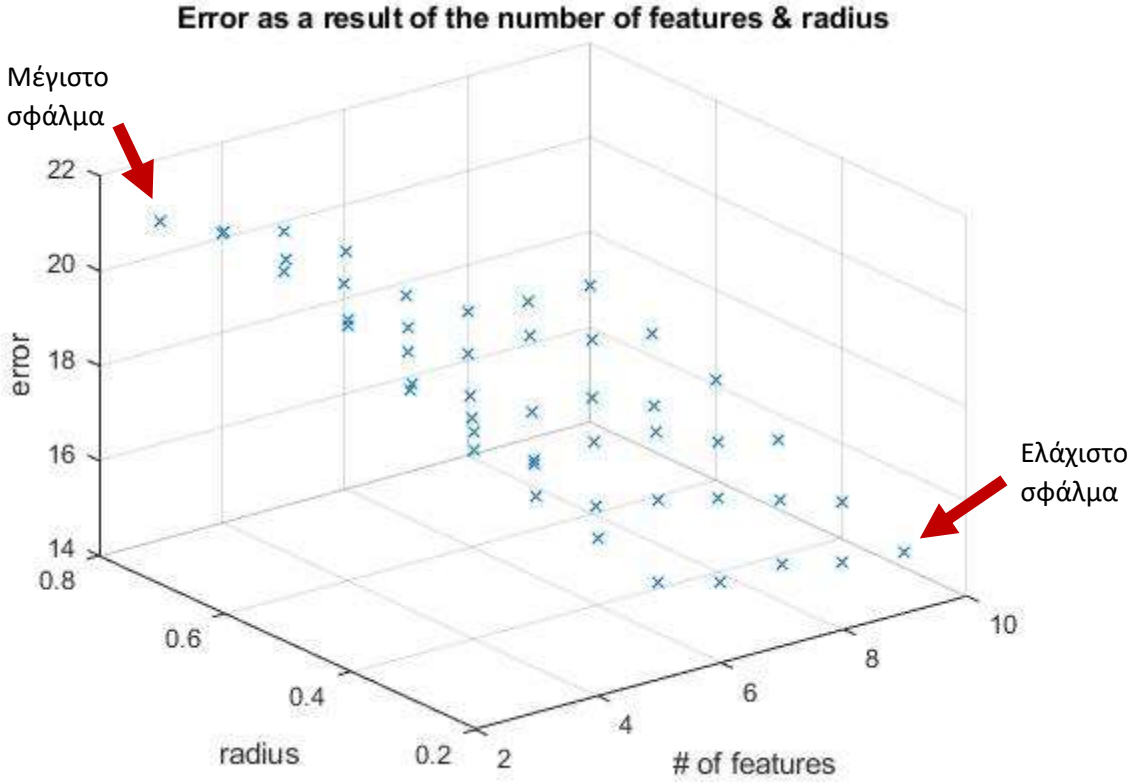
Τέλος με βάση το εκπαιδευμένο αυτό μοντέλο γίνεται πρόβλεψη στο test set (20% του αρχικού συνόλου).

Τα αποτελέσματα του grid-search έχουν αποθηκευτεί στο αρχείο resultsTable.xls, το οποίο επισυνάπτεται στα αρχεία της εργασίας.

Ως βέλτιστες παράμετροι από το δοσμένο εύρος επιλέχθηκαν:

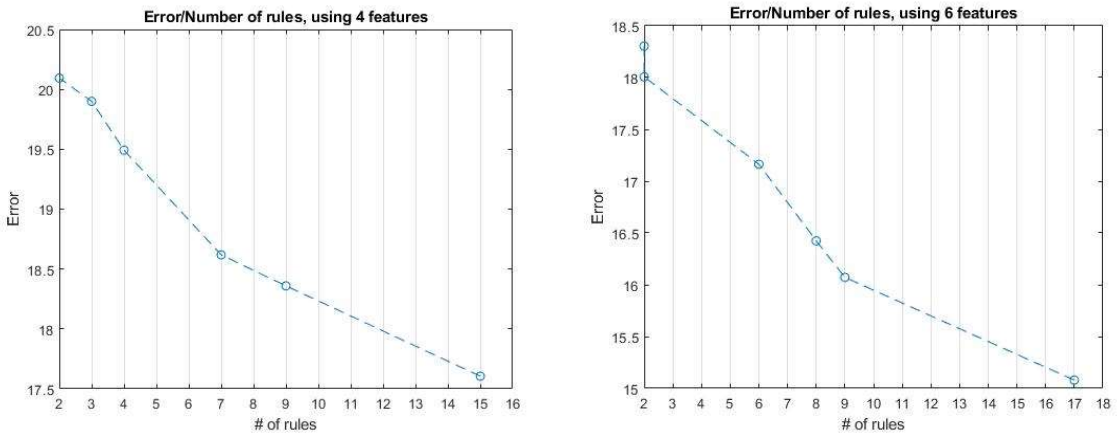
- 10 μεταβλητές/χαρακτηριστικά,
- ακτίνα ομαδοποίησης ίση με 0,3
- ελάχιστο σφάλμα 14,324

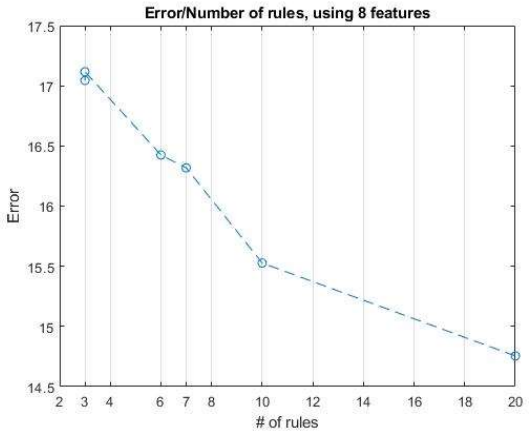
Εικόνα 1: Σφάλμα συναρτήσει πλήθους χαρακτηριστικών και ακτίνας

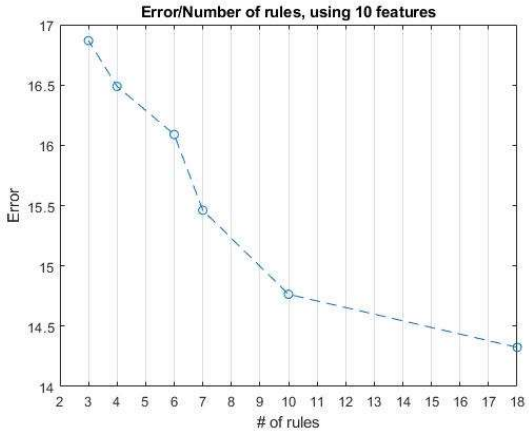


Στο εύρος στο οποίο εκτελέστηκε το grid-search, το σφάλμα αυξάνεται καθώς αυξάνεται η ακτίνα που χρησιμοποιείται για την ομαδοποίηση και μειώνεται καθώς αυξάνεται το πλήθος των χαρακτηριστικών που χρησιμοποιούνται. Το αποτέλεσμα είναι λογικό αφού όσο μικρότερη είναι ακτίνα τόσο μικρότερες ομάδες έχουμε και συνεπώς περισσότερους κανόνες. Ομοίως όσο περισσότερα χαρακτηριστικά χρησιμοποιούνται τόσο περισσότεροι κανόνες προκύπτουν. Όσο περισσότεροι κανόνες, τόσο αυξάνεται η δυνατότητα για περιγραφή από το μοντέλο, μέχρι βεβαίως το σημείο που ξεκινάει η υπερεκπαίδευση.

Εικόνα 2: Σφάλμα σε σχέση με τον αριθμό των κανόνων, ενδεικτικά διαγράμματα για 4,6,8 και 10 χαρακτηριστικά







Από τα διαγράμματα φαίνεται ότι το σφάλμα μειώνεται καθώς αυξάνεται ο αριθμός των κανόνων που χρησιμοποιεί το μοντέλο για δεδομένο πλήθος χαρακτηριστικών.

Ωστόσο, αν και προέκυψαν μοντέλα με παραπάνω κανόνες από 18, από τον παρακάτω πίνακα μπορεί να παρατηρηθεί ότι το σφάλμα σε αυτές τις περιπτώσεις δεν ήταν μικρότερο. Για τις περιπτώσεις των 19, 20, 21 κανόνων με χρήση 7,8 και 9 χαρακτηριστικών αντίστοιχα, το σφάλμα ήταν πολύ κοντινό στο ελάχιστο. Φαίνεται να επηρεάζει το πλήθος των χαρακτηριστικών που χρησιμοποιείται.

# χαρακτηριστικών	# κανόνων	Σφάλμα	Ακτίνα
5	18	16,372	0,3
7	19	14,753	0,3
8	20	14,755	0,3
9	21	14,458	0,3
10	18	14,324	0,3

Δεδομένου ότι, λόγω μεγάλων απαιτήσεων σε υπολογιστική ισχύ, η διερεύνησή μας (το grid search) έγινε σε περιορισμένο εύρος, κατά βάση όσον αφορά στο πλήθος των μεταβλητών (καθώς το εύρος G2 είναι ικανοποιητικό για τις τιμές που μπορεί να πάρει η ακτίνα). Στη γενική περίπτωση για καλύτερο αποτέλεσμα θα ήταν ορθότερο να γίνει grid-search στο σύνολο [10-81] ως προς τη διάσταση που αφορά το πλήθος των χαρακτηριστικών.

- ***Το τελικό μοντέλο περιέχει 10 εισόδους, 1 έξοδο και 18 κανόνες.***

Με 10 εισόδους, grid partitioning και 2 ασαφή σύνολα ανά είσοδο, θα είχαμε 2^{10} = 1024 κανόνες, με 3 ασαφή σύνολα ανά είσοδο θα είχαμε 3^{10} = 59049 κανόνες.

Οι 18 κανόνες είναι σημαντικά καλύτερο αποτέλεσμα ακόμα και από τους 1024.

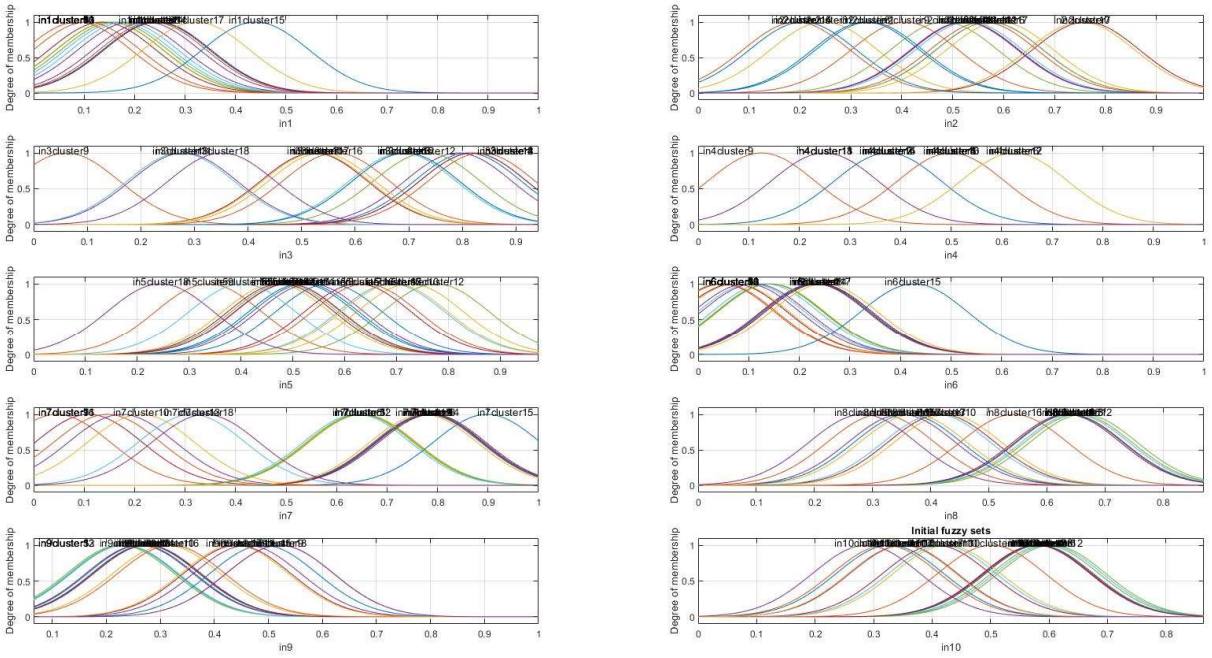
Πίνακας Δεικτών Απόδοσης

RMSE	14.6934
NMSE	0.1881
NDEI	0.4338

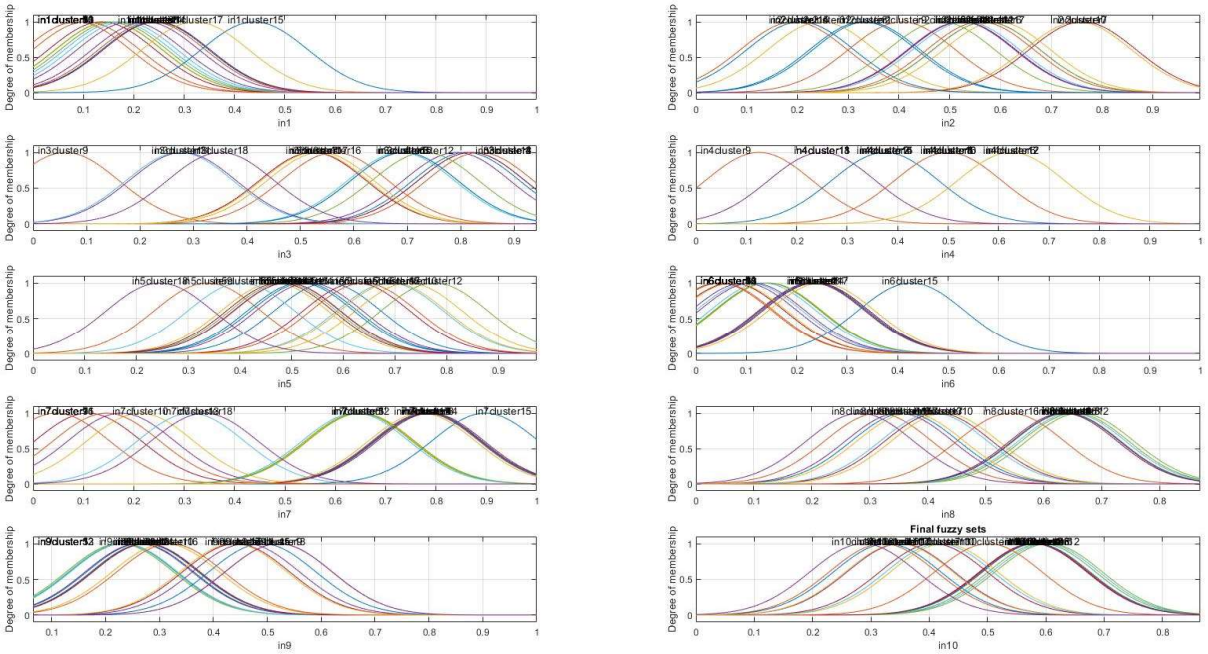
R2	0.8119

Τα υπόλοιπα αποτελέσματα παρουσιάζονται παρακάτω.

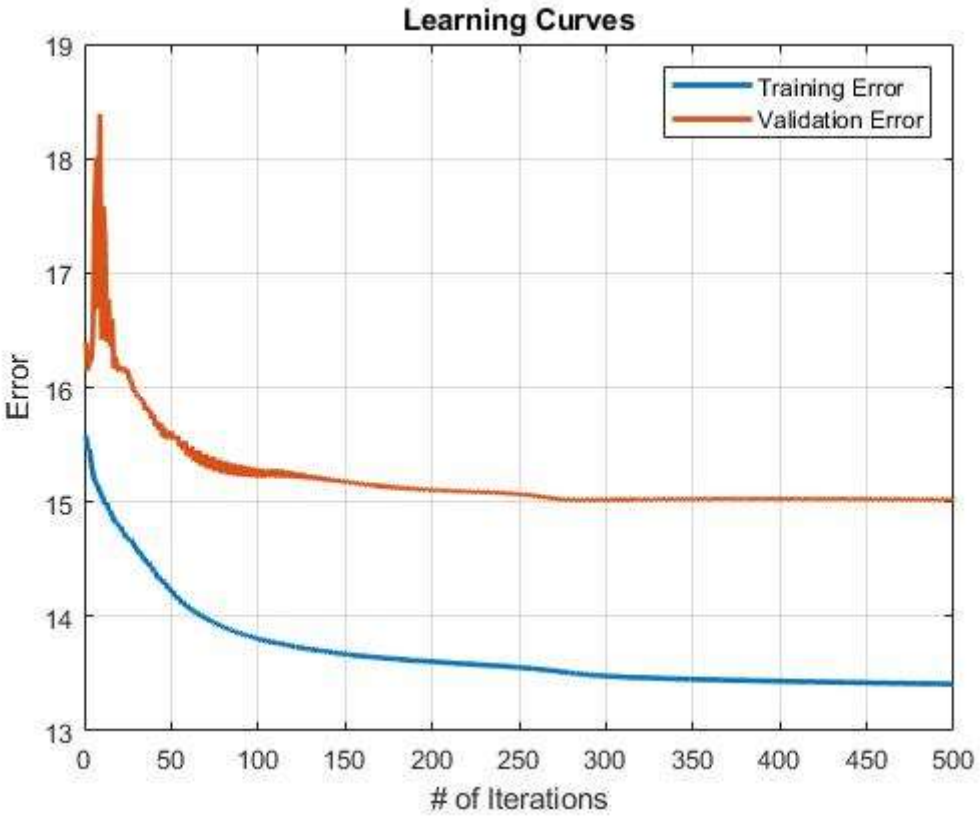
Εικόνα 3: Αρχικές μορφές των ασαφών συνόλων



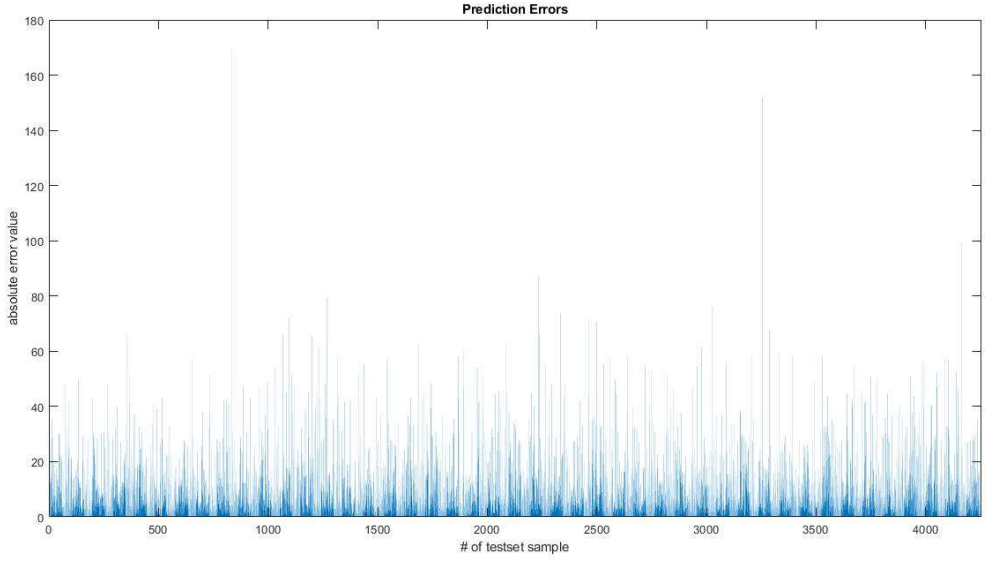
Εικόνα 4: Τελικές μορφές των ασαφών συνόλων



Εικόνα 5: Διάγραμμα μάθησης



Εικόνα 6: Σφάλματα πρόβλεψης, κατ' απόλυτη τιμή



Εικόνα 7: Πραγματικές τιμές & τιμές πρόβλεψης