Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

# ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

## ΤΕΤΑΡΤΗ ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ

## ΕΠΙΛΥΣΗ ΠΡΟΒΛΗΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕ ΧΡΗΣΗ ΜΟΝΤΕΛΩΝ TSK

Ημερομηνία: 11/10/2024

Οικονόμου Χρήστος

Α.Ε.Μ.: 10268

Email: cnoikonom@ece.auth.gr

# ΠΕΡΙΕΧΟΜΕΝΑ

# Εισαγωγή

## Μοντέλα TSK

Τα μοντέλα TSK (Takagi – Sugeno – Kang) αποτελούν έναν τύπο συστήματος , το οποίο παράγει ασαφείς (“fuzzy”) κανόνες με βάση ένα σετ δεδομένων που δίνεται ως είσοδος. Σχεδιάζονται για να προσεγγίζουν σύνθετα, μη γραμμικά συστήματα, χρησιμοποιώντας ένα σύνολο κανόνων “αν - τότε” (“if - then” rules), ενώ συνδυάζουν την ασαφή λογική με μαθηματικές συναρτήσεις για να δημιουργήσουν μια έξοδο, καθιστώντας τα δημοφιλή για συστήματα ελέγχου, πρόβλεψης και λήψης αποφάσεων.

**Κύρια χαρακτηριστικά των μοντέλων TSK:**

1. **Ασαφείς Κανόνες**: Όπως και άλλα συστήματα ασαφούς λογικής, τα μοντέλα TSK χρησιμοποιούν κανόνες "αν-τότε". Το μέρος "αν" ορίζει ασαφή σύνολα για τις εισόδους, και το μέρος "τότε" ορίζει μια μαθηματική συνάρτηση (συνήθως γραμμική ή πολυωνυμική) που καθορίζει την έξοδο.
2. **Ακριβείς Έξοδοι**: Σε αντίθεση με τα παραδοσιακά συστήματα ασαφούς λογικής που παράγουν ασαφείς τιμές, τα μοντέλα TSK συνήθως παράγουν ακριβείς, πραγματικές τιμές. Αυτό επιτυγχάνεται με τον συνδυασμό των αποτελεσμάτων από όλους τους κανόνες.
3. **Δομή Κανόνων**: Η δομή των κανόνων ενός μοντέλου TSK είναι:
    - **Αν** (προηγούμενο): ασαφείς συνθήκες για τις μεταβλητές εισόδου.
    - **Τότε** (επόμενο): μια γραμμική ή πολυωνυμική συνάρτηση που εφαρμόζεται στις μεταβλητές εισόδου.
4. **Σταθμισμένος Μέσος για την Έξοδο**: Η συνολική έξοδος του συστήματος είναι ένας σταθμισμένος μέσος όρος των εξόδων όλων των κανόνων, με τα βάρη να καθορίζονται από τον βαθμό στον οποίο οι τιμές εισόδου ικανοποιούν τις ασαφείς συνθήκες στα προηγούμενα.

## Στόχος της Εργασίας

Η παρούσα εργασία πραγματεύεται τη δυνατότητα επίλυσης προβλημάτων ταξινόμησης (classification) με την χρήση TSK μοντέλων. Για την εξαγωγή των συμπερασμάτων, γίνεται εφαρμογή των μοντέλων σε δύο datasets από το UCI Repository, με σκοπό την εκτίμηση της μεταβλητής στόχου από τα διαθέσιμα δεδομένα, με χρήση ασαφών νευρωνικών μοντέλων. Αναλυτικότερη εξήγηση της διαδικασίας παρατίθεται στις ενότητες που ακολουθούν.

# Πρώτο Μέρος - Εφαρμογή σε Απλό Dataset

Στο πρώτο μέρος της εργασίας, εξετάζεται η εφαρμογή τεσσάρων μοντέλων TSK σε ένα σχετικά απλό σύνολο δεδομένων και γίνεται διερεύνηση του τρόπου εκπαίδευσης και αξιολόγησης των μοντέλων αυτών. Συγκεκριμένα, χρησιμοποιείται το σύνολο δεδομένων Haberman's Survival από το UCI Repository, το οποίο περιλαμβάνει 306 δείγματα (instances) και 3 χαρακτηριστικά (features). Ο κώδικας MATLAB που αποτελεί την υλοποίηση του μέρους αυτού της εργασίας είναι καταγεγραμμένος στο αρχείο "classification_a_10268.m".

Αρχικά γίνεται ο κατάλληλος διαχωρισμός του dataset, με σκοπό τον σχηματισμό τριών υποσυνόλων δεδομένων: του υποσυνόλου **εκπαίδευσης**, του υποσυνόλου **επικύρωσης** και του υποσυνόλου **ελέγχου απόδοσης** του τελικού μοντέλου ($D_{chk}$). Ο διαχωρισμός έγινε μέσω της συνάρτησης split_scale (αρχείο "split_scale.m"), ως εξής:

- Υποσύνολο Εκπαίδευσης ($D_{trn}$) → 60% του αρχικού dataset → 184 instances
- Υποσύνολο Επικύρωσης ($D_{val}$) → 20% του αρχικού dataset → 61 instances
- Υποσύνολο Ελέγχου ($D_{chk}$) → 20% του αρχικού dataset → 61 instances

Επόμενο βήμα αποτέλεσε η **εκπαίδευση** των τεσσάρων TSK μοντέλων. Όπως αναφέρθηκε και προηγουμένως, συνολικά εκπαιδεύονται τέσσερα μοντέλα. Η κύρια διαφορά των μοντέλων είναι ο τρόπος διαμέρισης του χώρου εισόδου:

- **Class Dependent**, όπου διαμερίζεται για το σύνολο των δεδομένων
- **Class Independent**, όπου η διαμέριση γίνεται ξεχωριστά για κάθε κλάση δεδομένων

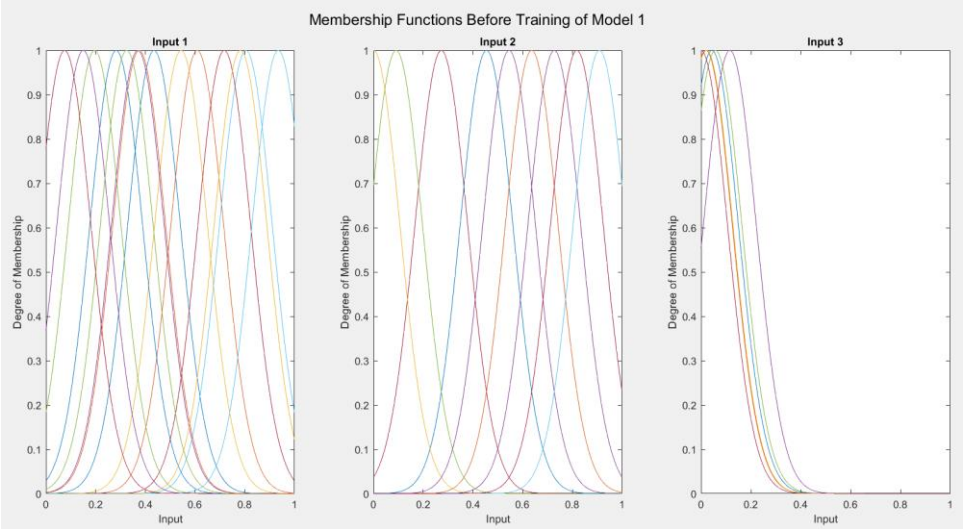
Για κάθε τρόπο διαχωρισμού του χώρου εισόδου δημιουργούνται δύο μοντέλα: ένα μικρής ακτίνας cluster (0.3) και ένα μεγαλύτερης ακτίνας cluster (0.8). Όσον αφορά την έξοδο, επιλέγεται σε όλες τις περιπτώσεις να είναι singleton, όπως υποδεικνύεται από την εκφώνηση. Τα χαρακτηριστικά των μοντέλων εκπαίδευσης συνοψίζονται στον Πίνακα 1.

| Μοντέλο | Τρόπος Διαμέρισης Εισόδου | Μορφή Εξόδου | Ακτίνα Cluster |
|---|---|---|---|
| TSK_model_1 | Class Independent | Singleton | 0.3 |
| TSK_model_2 | Class Independent | Singleton | 0.8 |
| TSK_model_3 | Class Dependent | Singleton | 0.3 |
| TSK_model_4 | Class Dependent | Singleton | 0.8 |

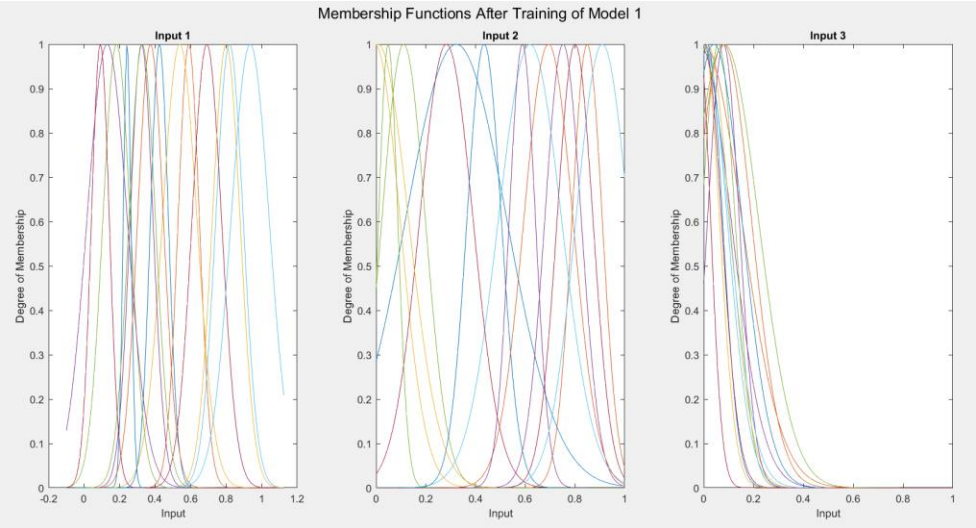
***Πίνακας 1**: Χαρακτηριστικά Μοντέλων Εκπαίδευσης*

Στη συνέχεια, παρατίθενται τα διαγράμματα που απεικονίζουν τις τελικές μορφές των ασαφών συνόλων που προέκυψαν μέσω της διαδικασίας εκπαίδευσης, οι καμπύλες εκμάθησης, καθώς και οι πίνακες σφαλμάτων για κάθε ένα από τα μοντέλα TSK που εκπαιδεύτηκαν

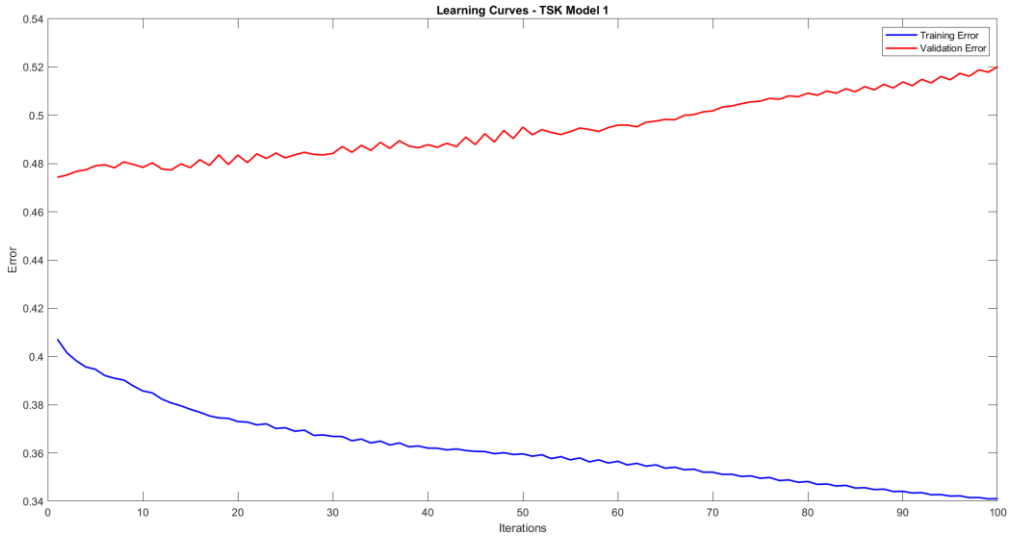
## Μοντέλο 1 (TSK_model_1)




***Διάγραμμα 1**: Membership Functions Πριν την Εκπαίδευση του Μοντέλου 1*




***Διάγραμμα 2**: Membership Functions Μετά την Εκπαίδευση του Μοντέλου 1*

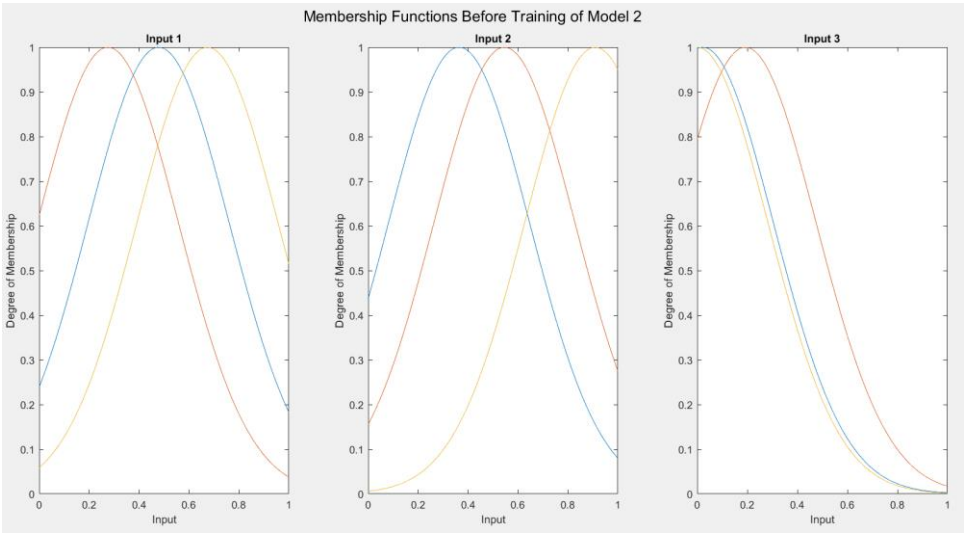



***Διάγραμμα 3**: Καμπύλες Εκμάθησης του Μοντέλου 1*

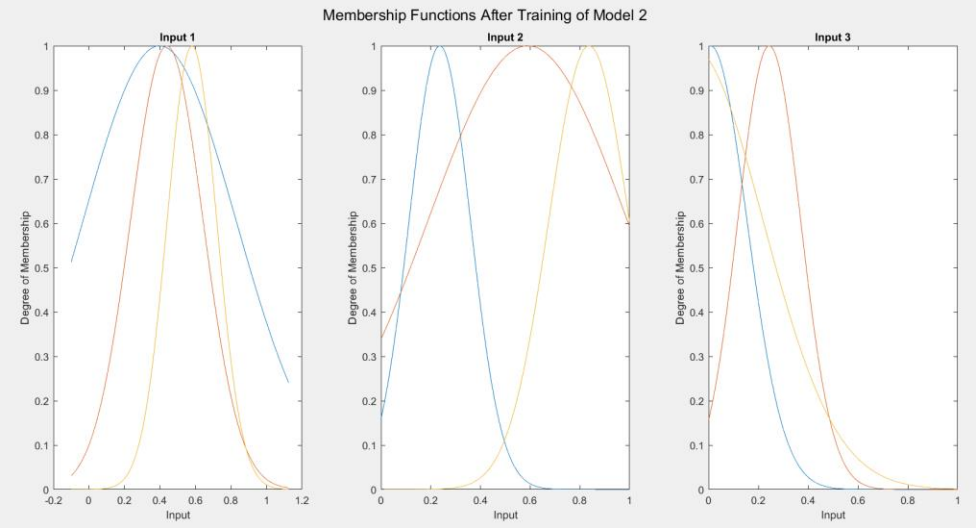
| | |
|---|---|
| 42 | 14 |
| 3 | 2 |

***Πίνακας 2**: Πίνακας Σφαλμάτων (Error Matrix) του Μοντέλου 1*

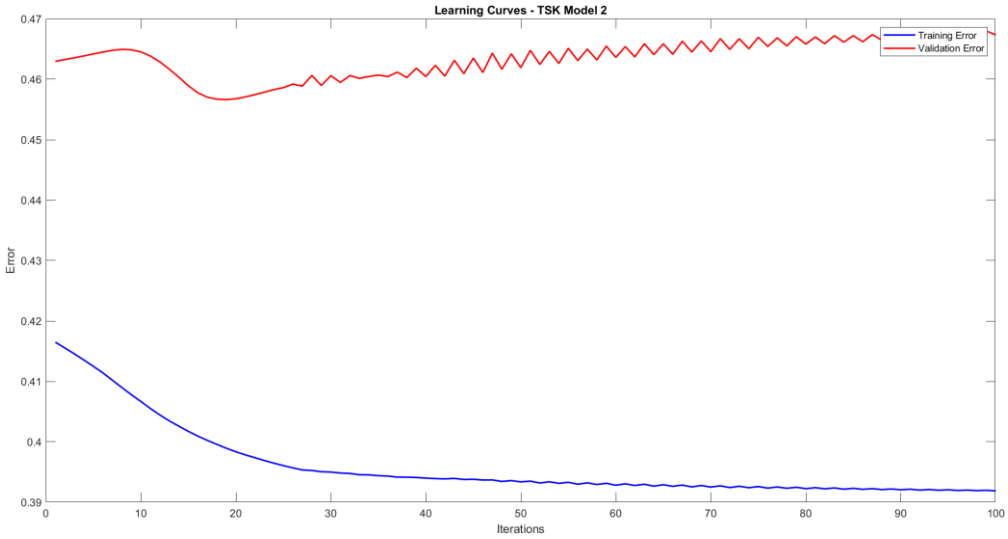
## Μοντέλο 2 (TSK_model_2)




***Διάγραμμα 4**: Membership Functions Πριν την Εκπαίδευση του Μοντέλου 2*




***Διάγραμμα 5**: Membership Functions Μετά την Εκπαίδευση του Μοντέλου 2*

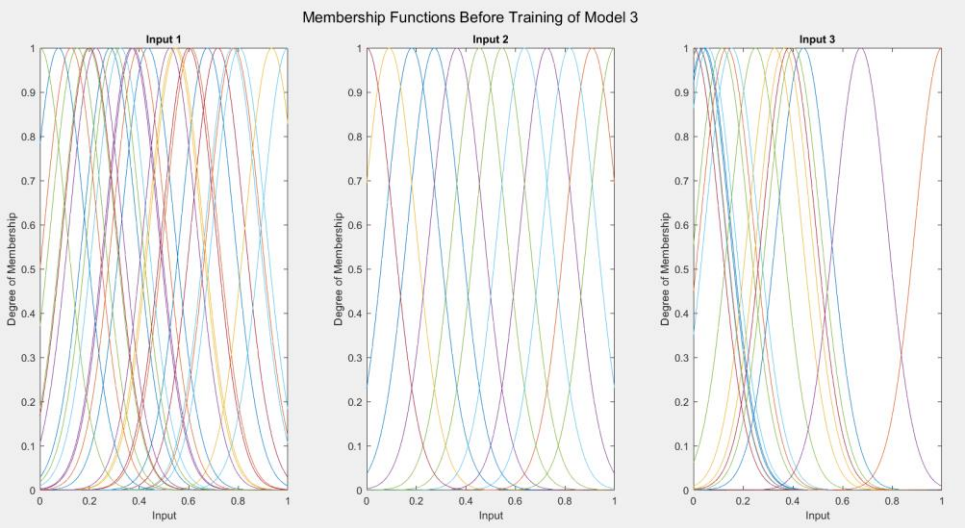



***Διάγραμμα 6**: Καμπύλες Εκμάθησης του Μοντέλου 2*

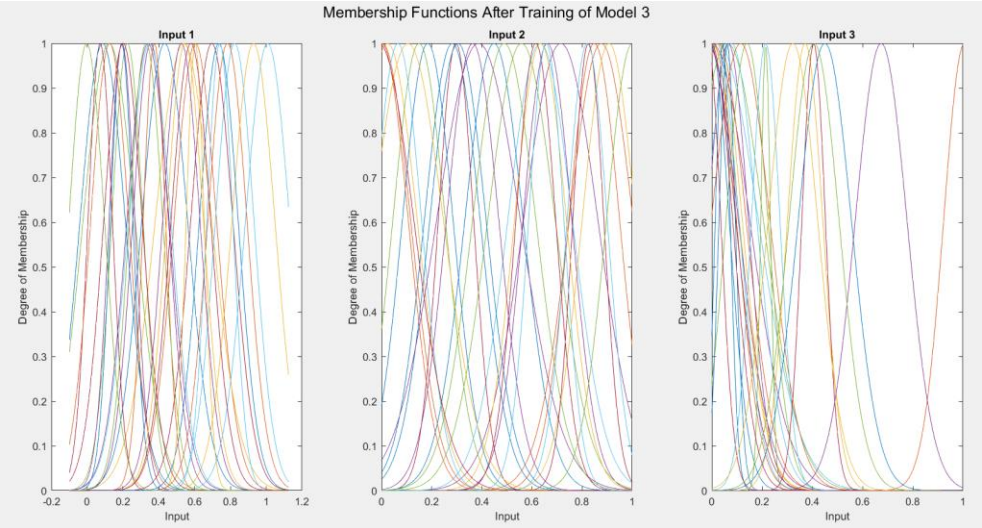
| | |
|---|---|
| 40 | 13 |
| 5 | 3 |

***Πίνακας 3**: Πίνακας Σφαλμάτων (Error Matrix) του Μοντέλου 2*

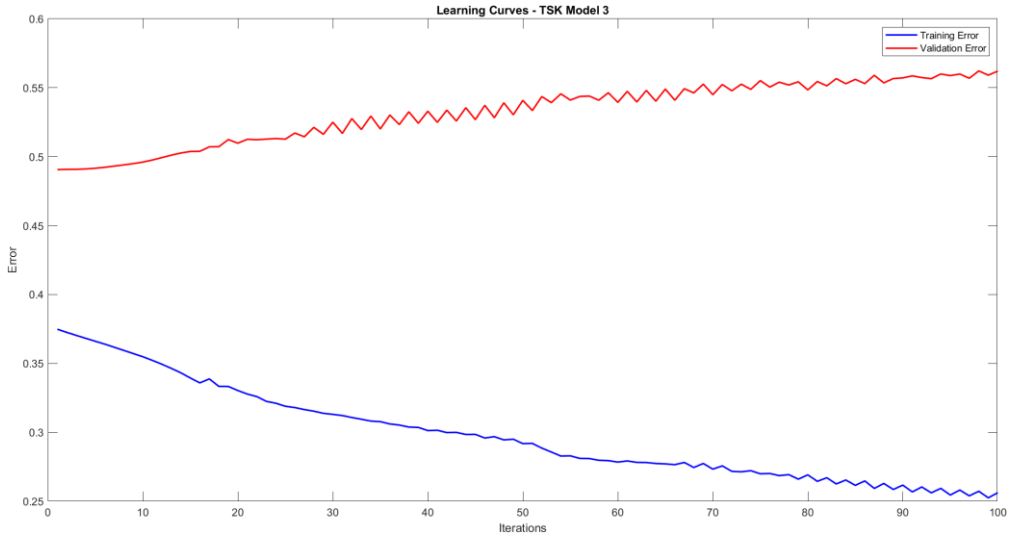
## Μοντέλο 3 (TSK_model_3)




***Διάγραμμα 7**: Membership Functions Πριν την Εκπαίδευση του Μοντέλου 3*




***Διάγραμμα 8**: Membership Functions Μετά την Εκπαίδευση του Μοντέλου 3*

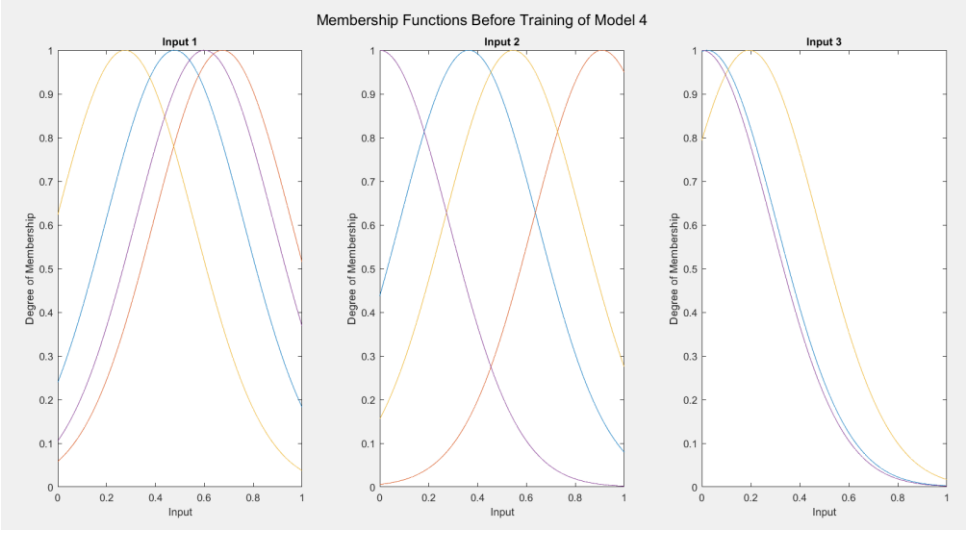



***Διάγραμμα 9**: Καμπύλες Εκμάθησης του Μοντέλου 3*

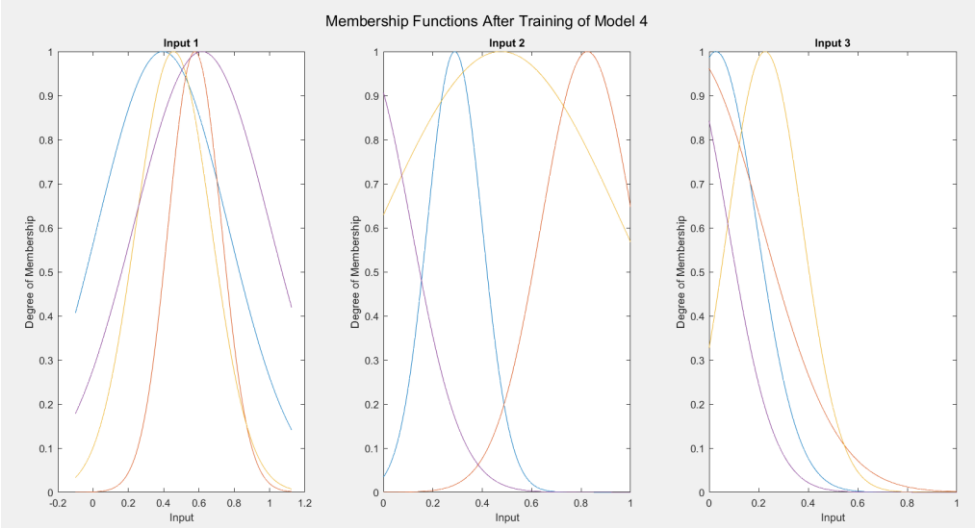
| | |
|---|---|
| 39 | 10 |
| 6 | 6 |

***Πίνακας 4**: Πίνακας Σφαλμάτων (Error Matrix) του Μοντέλου 3*

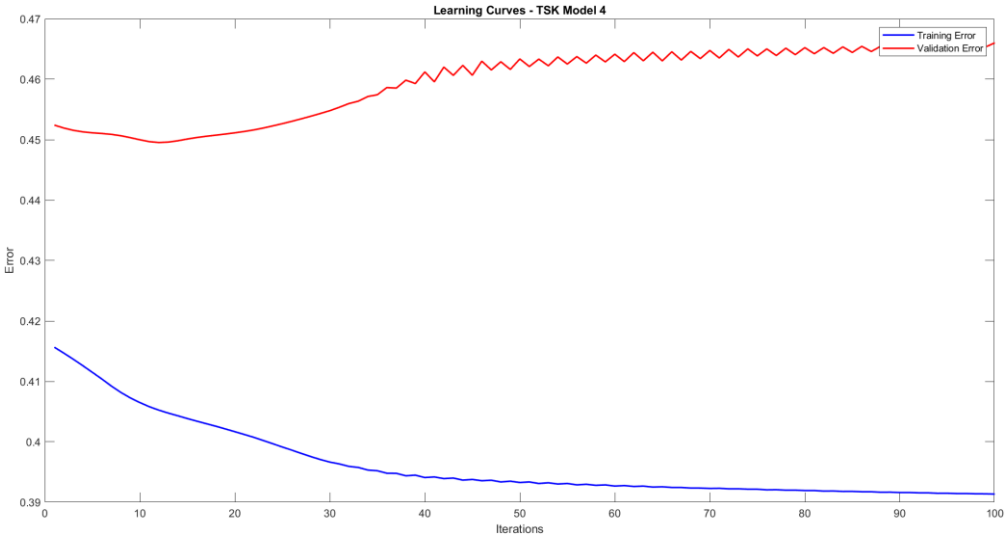
## Μοντέλο 4 (TSK_model_4)




***Διάγραμμα 10**: Membership Functions Πριν την Εκπαίδευση του Μοντέλου 4*




***Διάγραμμα 11**: Membership Functions Μετά την Εκπαίδευση του Μοντέλου 4*




***Διάγραμμα 12**: Καμπύλες Εκμάθησης του Μοντέλου 4*

| | |
|---|---|
| 43 | 14 |
| 2 | 2 |

***Πίνακας 5**: Πίνακας Σφαλμάτων (Error Matrix) του Μοντέλου 4*

Τέλος, παρατίθεται ο πίνακας με τις τιμές των δεικτών απόδοσης - αξιολόγησης των τεσσάρων μοντέλων TSK.

| Μοντέλο | OA | PA _1 | PA_2 | UA _1 | UA_2 | K |
|---|---|---|---|---|---|---|
| TSK_model_1 | 0.7377 | 0.9333 | 0.1875 | 0.7636 | 0.5000 | 0.7377 |
| TSK_model_2 | 0.7049 | 0.9111 | 0.1250 | 0.7455 | 0.3333 | 0.7049 |
| TSK_model_3 | 0.7213 | 0.8667 | 0.3125 | 0.7800 | 0.4545 | 0.7213 |
| TSK_model_4 | 0.7337 | 0.9333 | 0.1875 | 0.7636 | 0.5000 | 0.7337 |

***Πίνακας 6**: Τιμές Δεικτών Απόδοσης*

**Σημείωση:**

Οι δείκτες PA_1 (UA_1) και PA_2 (UA_2) αναφέρονται στο Producer's Accuracy (User's Accuracy) των κλάσεων 1 και 2 που απεικονίζονται στους πίνακες σφαλμάτων, αντίστοιχα.

## Συμπεράσματα

Παρατηρώντας τόσο τα παραπάνω διαγράμματα, όσο και τις τιμές που αναγράφονται στον Πίνακα 6, μπορούν να γίνουν ορισμένες παρατηρήσεις και να εξαχθούν τα εξής συμπεράσματα:

- Τα Μοντέλα 3 και 4 είναι τα πιο ισορροπημένα μοντέλα με βάση τη συνολική ακρίβεια και τη μετρική K. Το Μοντέλο 3 φαίνεται να είναι πιο ισορροπημένο μεταξύ των false positives και false negatives, ενώ το Μοντέλο 4 έχει την καλύτερη απόδοση σε true positives αλλά αντιμετωπίζει μεγαλύτερα false positives.

- Το Μοντέλο 1 αποδίδει σχετικά καλά, αλλά έχει περισσότερα false negatives και χαμηλότερη ακρίβεια για την κλάση 2.

- Το Μοντέλο 2 εμφανίζει τη χαμηλότερη απόδοση στις περισσότερες μετρικές και είναι το λιγότερο αποτελεσματικό από τα 4 μοντέλα.

- Τα class dependent μοντέλα (3 και 4), είναι περισσότερο αποδοτικά στην διαδικασία ταξινόμησης σε σχέση με τα class independent μοντέλα (1 και 2). Στα class independent μοντέλα αποδοτικότερη μέθοδος είναι η χρήση μίας μεγάλης ακτίνας cluster, ενώ στα class dependent μοντέλα αποδοτικότερο είναι αυτό με μικρή ακτίνα clusters. Οπότε, δεν υπάρχει κάποια συσχέτιση όσο αφορά την ακτίνα.

- Στα μοντέλα με μικρή ακτίνα clusters, υπάρχει επικάλυψη μεταξύ ασαφών συνόλων, αλλά γενικά τα σύνολα είναι απλωμένα στο πεδίο ορισμού τους. Στα σύνολα με μεγάλη ακτίνα clusters, τα ασαφή σύνολα είναι λιγότερα αλλά υπάρχει μεγάλη επικάλυψη μεταξύ τους. Επομένως, υπάρχει επικάλυψη και στην ενεργοποίηση των κανόνων, κάτι που συνεπάγεται μειωμένη απόδοση του συστήματος ταξινόμησης.

Με βάση αυτά τα δεδομένα, θα μπορούσε να προτιμηθεί το **Μοντέλο 3** για μια ισορροπημένη προσέγγιση ή το **Μοντέλο 4**, εάν ο στόχος είναι η μέγιστη ακρίβεια στην κλάση 1.

# Δεύτερο Μέρος - Εφαρμογή σε Dataset με Υψηλή Διαστασιμότητα

Το δεύτερο σκέλος της εργασίας αφορά τη μελέτη ενός dataset με υψηλή διαστασιμότητα και, συνεπώς, αρκετά μεγαλύτερο όγκο δεδομένων από αυτό που εξετάστηκε κατά το πρώτο μέρος. Το γεγονός ότι ο αριθμός κανόνων αυξάνεται εκθετικά σε σχέση με το πλήθος των εισόδων για την περίπτωση του grid partitioning, οδηγεί στην αναζήτηση διαφορετικών μεθόδων για την μοντελοποίηση αυτού του προβλήματος. Προτεραιότητα, πλέον, αποτελεί η **μείωση της διαστασιμότητας**, κάτι που επιτυγχάνεται μέσω της **επιλογής χαρακτηριστικών**, καθώς και της **χρήσης τεχνικών grid searching**.

Το σύνολο δεδομένων που θα μελετηθεί είναι το Epileptic Seizure Recognition dataset από το UCI Repository, το οποίο περιλαμβάνει 11500 δείγματα (instances) και 179 χαρακτηριστικά (features). Ο κώδικας MATLAB που αποτελεί την υλοποίηση του μέρους αυτού της εργασίας είναι καταγεγραμμένος στο αρχείο “classification_b_10268.m”.

Αρχικά, όπως και στο πρώτο μέρος της εργασίας, γίνεται ο κατάλληλος διαχωρισμός του dataset, με σκοπό τον σχηματισμό τριών υποσυνόλων δεδομένων: του υποσυνόλου **εκπαίδευσης**, του υποσυνόλου **επικύρωσης** και του υποσυνόλου **ελέγχου απόδοσης** του τελικού μοντέλου ($D_{chk}$). Ο διαχωρισμός έγινε μέσω της συνάρτησης split_scale (αρχείο “split_scale.m”), ως εξής:

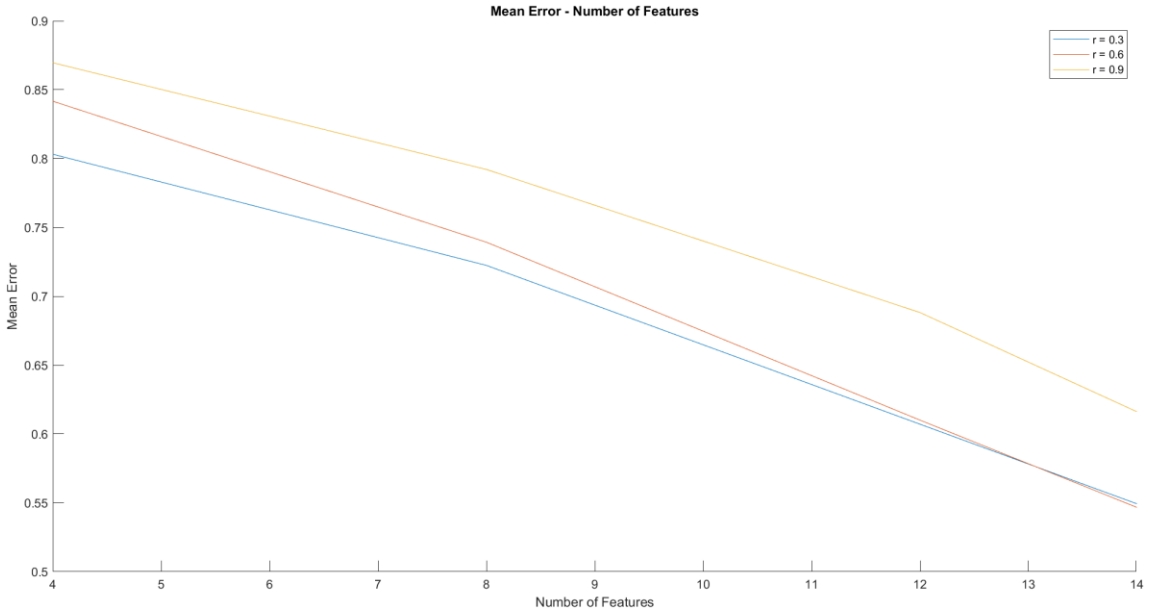
- Υποσύνολο Εκπαίδευσης ($D_{trn}$) → 60% του αρχικού dataset → 6900 instances
- Υποσύνολο Επικύρωσης ($D_{val}$) → 20% του αρχικού dataset → 2300 instances
- Υποσύνολο Ελέγχου ($D_{chk}$) → 20% του αρχικού dataset → 2300 instances

Όπως αναφέρεται και στην εκφώνηση, για τους σκοπούς της εργασίας έχουν οριστεί δύο βασικές παράμετροι: το **πλήθος των χαρακτηριστικών** που θα χρησιμοποιηθούν στην εκπαίδευση των μοντέλων και η **ακτίνα επιρροής** των clusters, η οποία επηρεάζει και το πλήθος των κανόνων. Για τις παραμέτρους αυτές, επιλέγονται οι τιμές που φαίνονται στον Πίνακα 7.

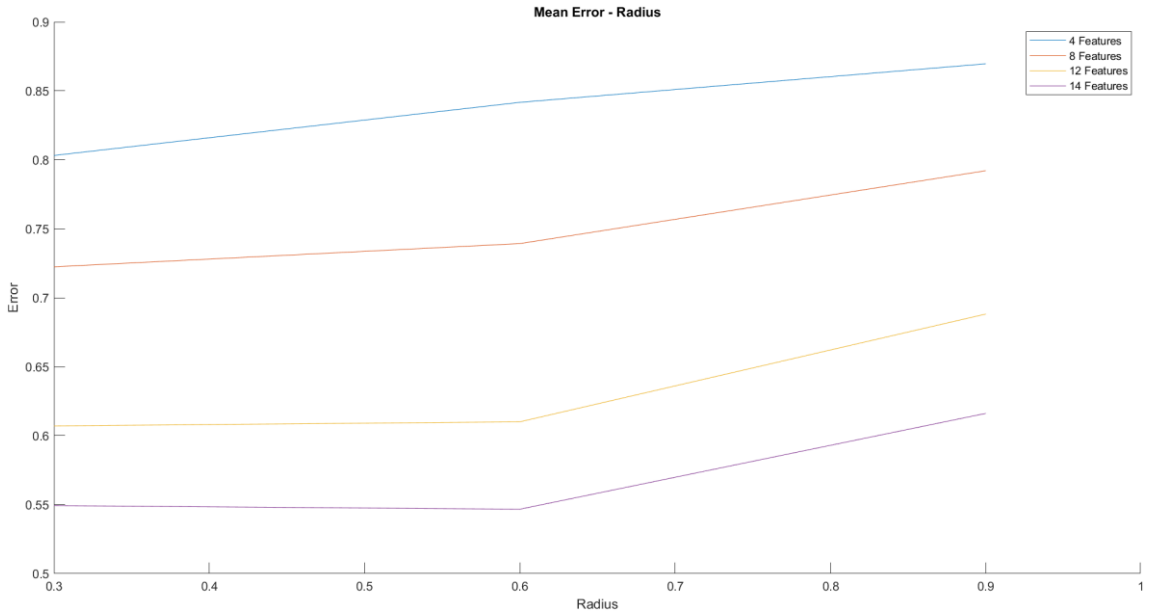
| Τιμές Παραμέτρων | | | | |
|---|---|---|---|---|
| Πλήθος Χαρακτηριστικών | 4 | 8 | 12 | 14 |
| Ακτίνα Επιρροής Clusters | 0.3 | 0.6 | 0.9 | |

***Πίνακας 7**: Επιλογή Τιμών Παραμέτρων*

Συνεπώς, θα εφαρμοστεί **5-fold-cross-validation** για κάθε μοντέλο που προκύπτει από τους συνδυασμούς των τιμών του Πίνακα 3, και τα μοντέλα αυτά θα αξιολογηθούν με βάση το μέσο σφάλμα.




***Διάγραμμα 13**: Μέσο Σφάλμα Συναρτήσει του Αριθμού Χαρακτηριστικών για τις Επιλεγμένες Τιμές της Ακτίνας Επιρροής*




***Διάγραμμα 14**: Μέσο Σφάλμα Συναρτήσει της Ακτίνας Επιρροής για τις Επιλεγμένες Τιμές του Πλήθους Χαρακτηριστικών*

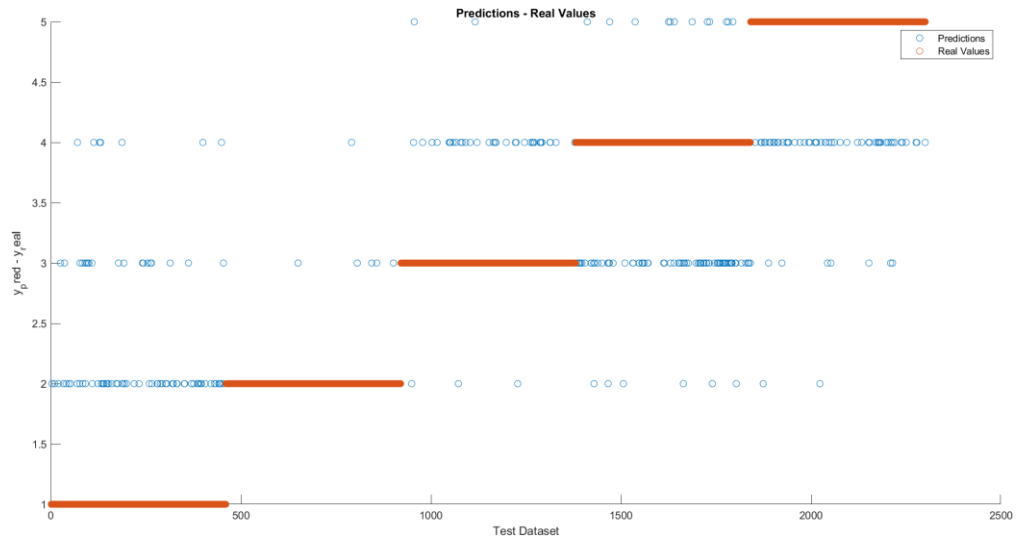
Όσον αφορά την ακτίνα επιρροής, γίνεται αντιληπτό πως, από ένα σημείο και μετά, όσο μεγαλύτερη γίνεται, μπορεί να οδηγήσει σε χειρότερες προβλέψεις (μεγαλύτερο σφάλμα). Για τον λόγο αυτό, θα πρέπει να επιλεχθεί μία μεσαία τιμή, ώστε να επιτευχθεί η καλύτερη δυνατή πρόβλεψη, αυτή, δηλαδή, με το ελάχιστο δυνατό σφάλμα.

Σχετικά με το πλήθος των χαρακτηριστικών, γενικότερα παρατηρείται η τάση να παρουσιάζεται μικρότερο σφάλμα όσο αυξάνεται το πλήθος τους, οπότε πρέπει να επιλεχθεί ένα πλήθος χαρακτηριστικών, το οποίο να μας οδηγεί σε ένα ικανοποιητικό μοντέλο, το οποίο ταυτόχρονα να μην απαιτεί πολύ μεγάλη υπολογιστική ισχύ.

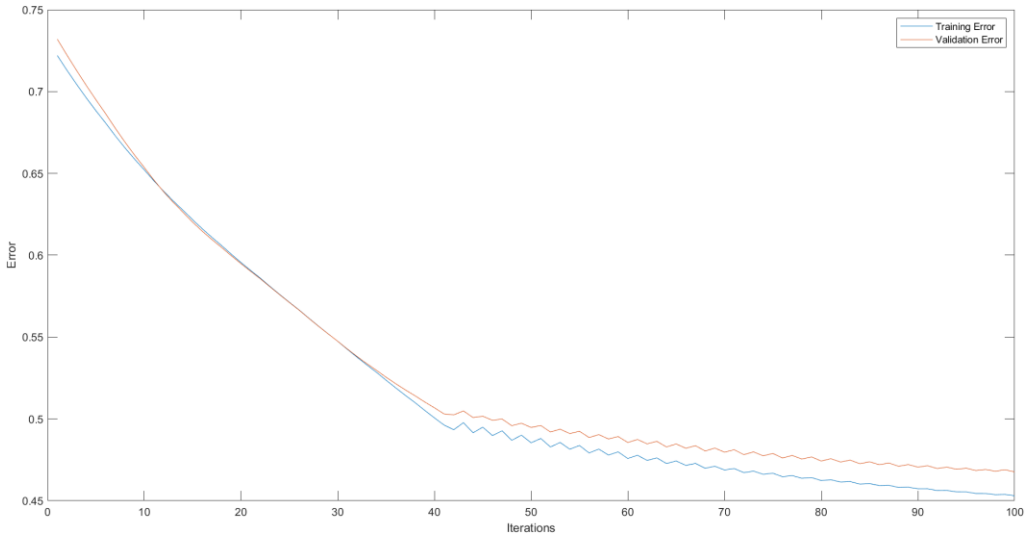
Σύμφωνα με τις παραπάνω παρατηρήσεις, οι οποίες έγιναν με βάση τα σχετικά διαγράμματα, προκύπτει πως ο βέλτιστος συνδυασμός τιμών για τις δύο κύριες παραμέτρους, δηλαδή αυτός με τα μικρότερα σφάλματα, είναι:

- Πλήθος Χαρακτηριστικών = **14**
- Ακτίνα Επιρροής Clusters = **0.6**

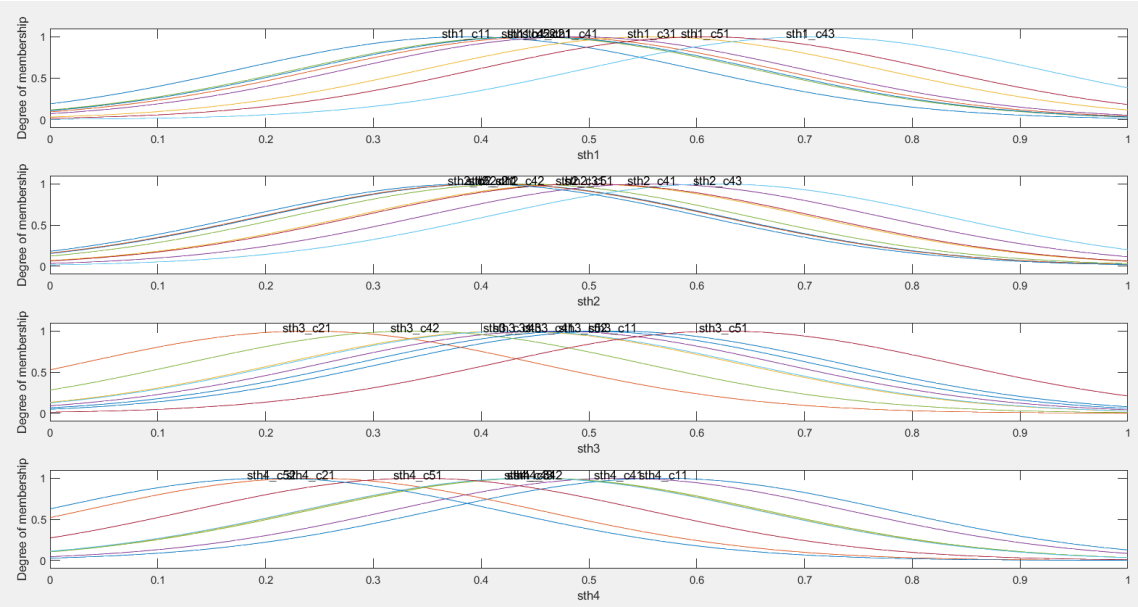
Για το βέλτιστο μοντέλο, παρατίθενται τα εξής διαγράμματα:



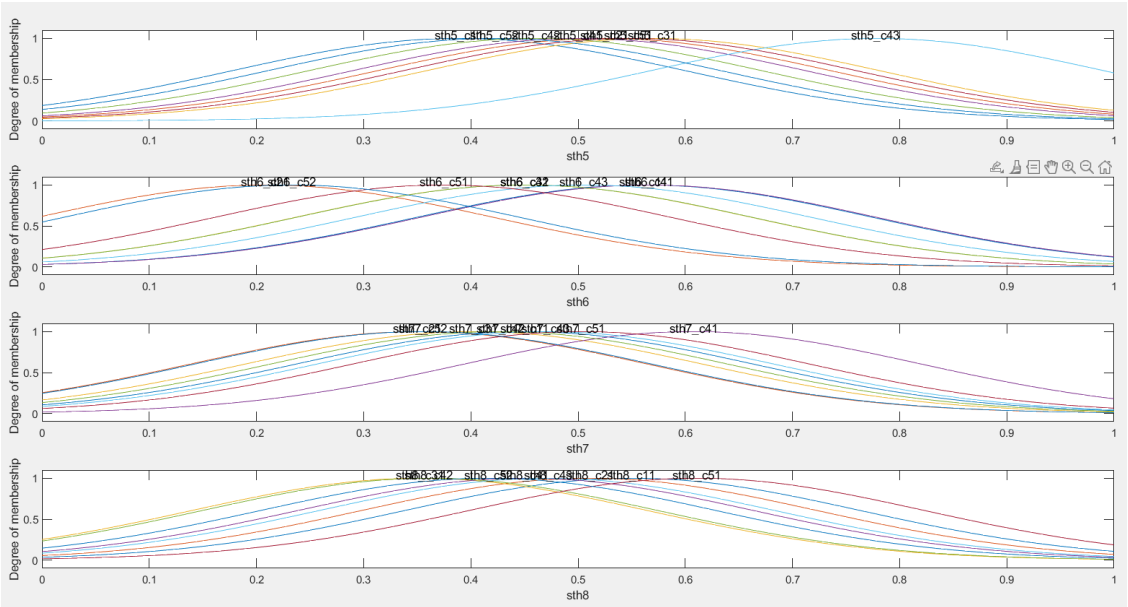

***Διάγραμμα 15**: Προβλέψεις Τελικού Μοντέλου και Πραγματικές Τιμές*



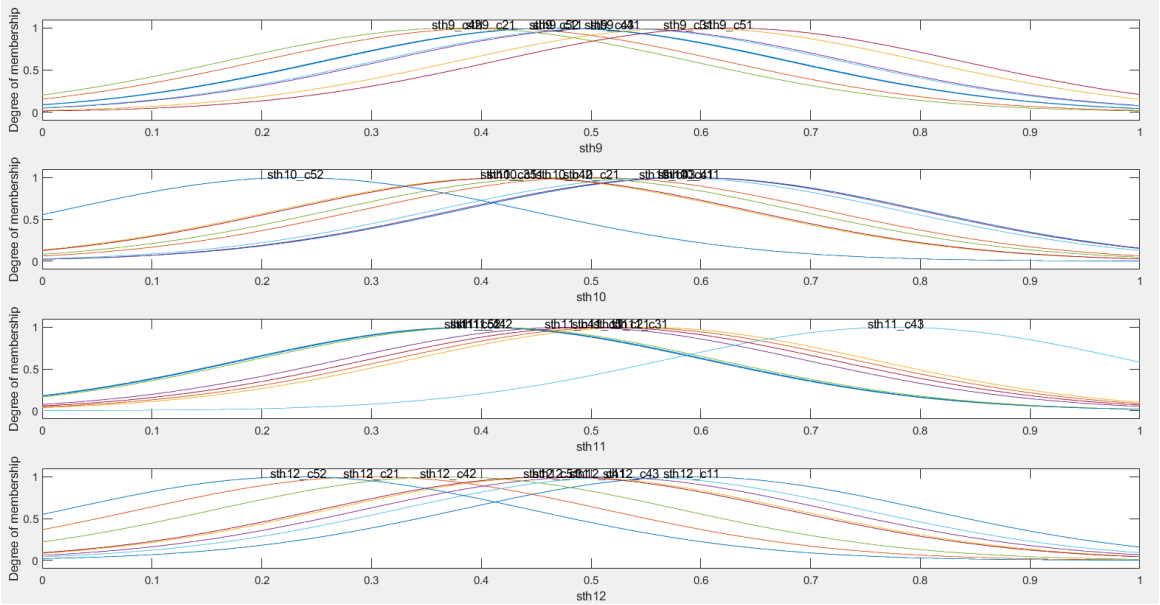

***Διάγραμμα 16**: Καμπύλες Εκμάθησης Τελικού Μοντέλου*



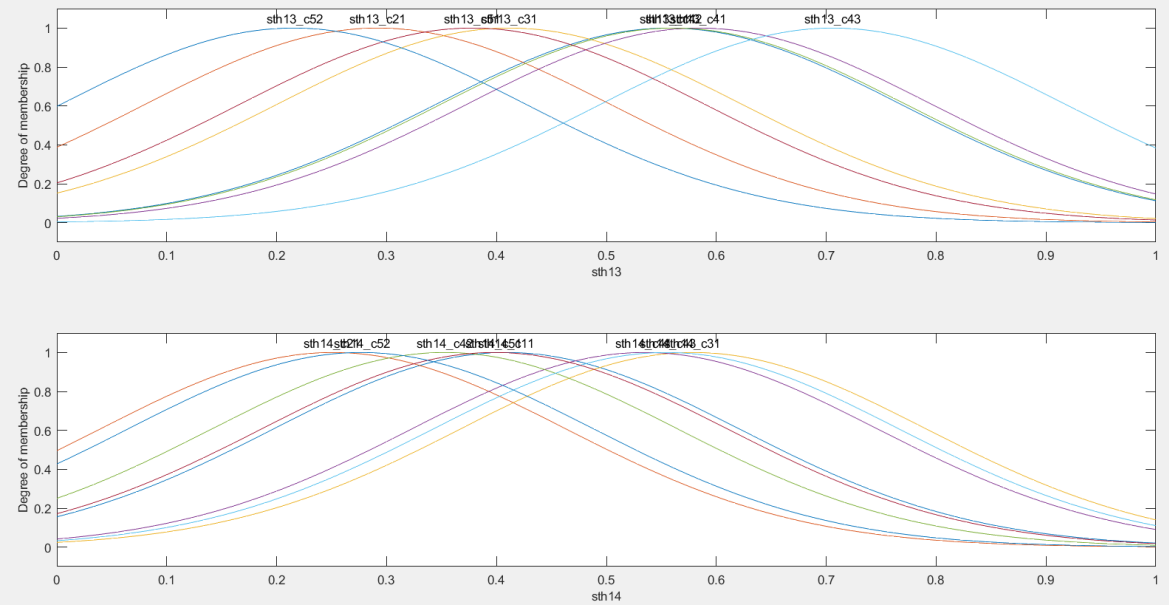

***Διάγραμμα 17**: Membership Functions Πριν την Εκπαίδευση του Βέλτιστου Μοντέλου (Inputs 1 - 4)*



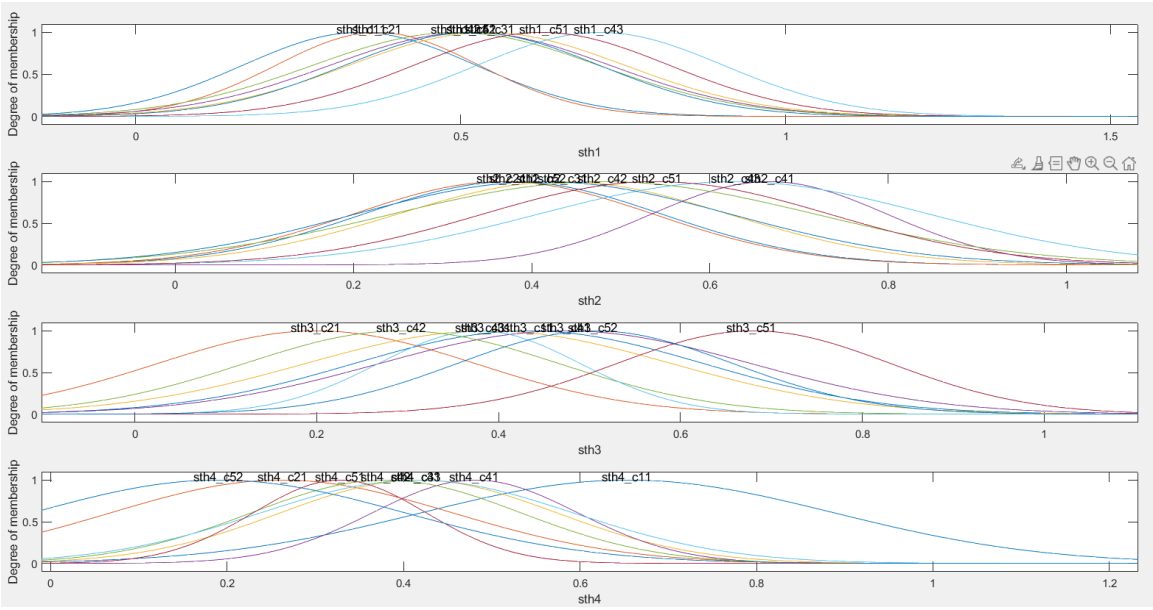

***Διάγραμμα 18**: Membership Functions Πριν την Εκπαίδευση του Βέλτιστου Μοντέλου (Inputs 5 - 8)*



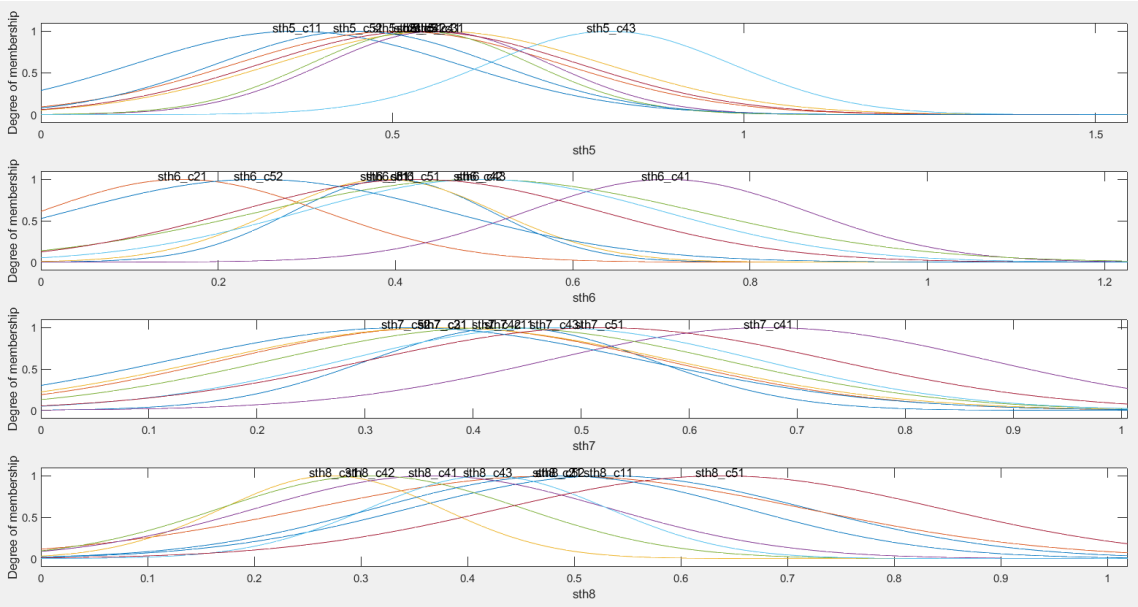

***Διάγραμμα 19**: Membership Functions Πριν την Εκπαίδευση του Βέλτιστου Μοντέλου (Inputs 9 - 12)*



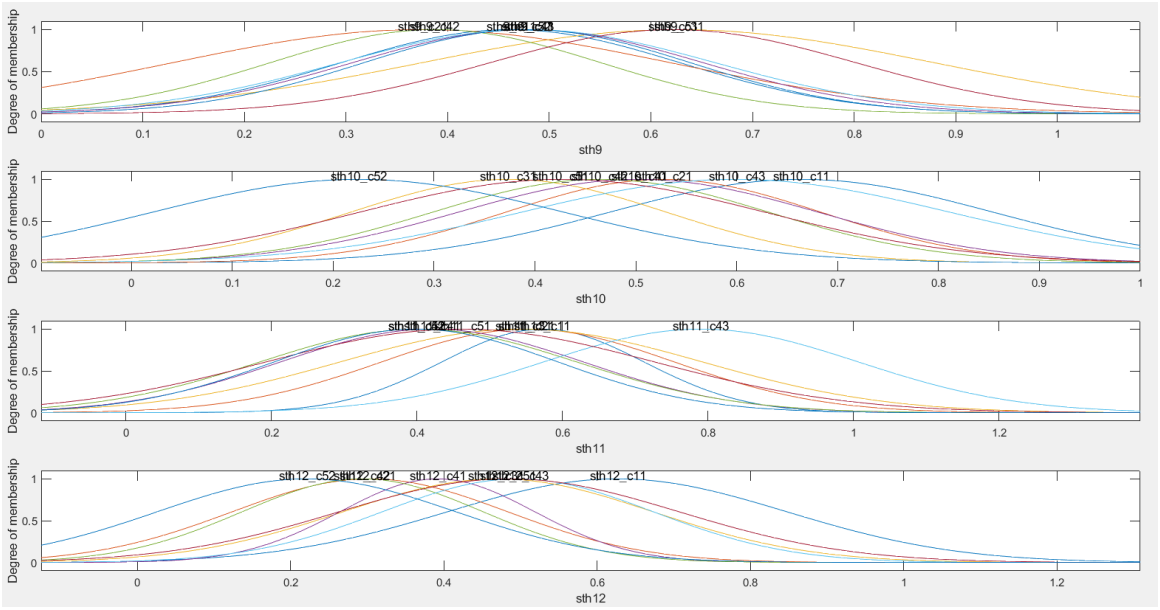

***Διάγραμμα 20**: Membership Functions Πριν την Εκπαίδευση του Βέλτιστου Μοντέλου (Inputs 13 - 14)*



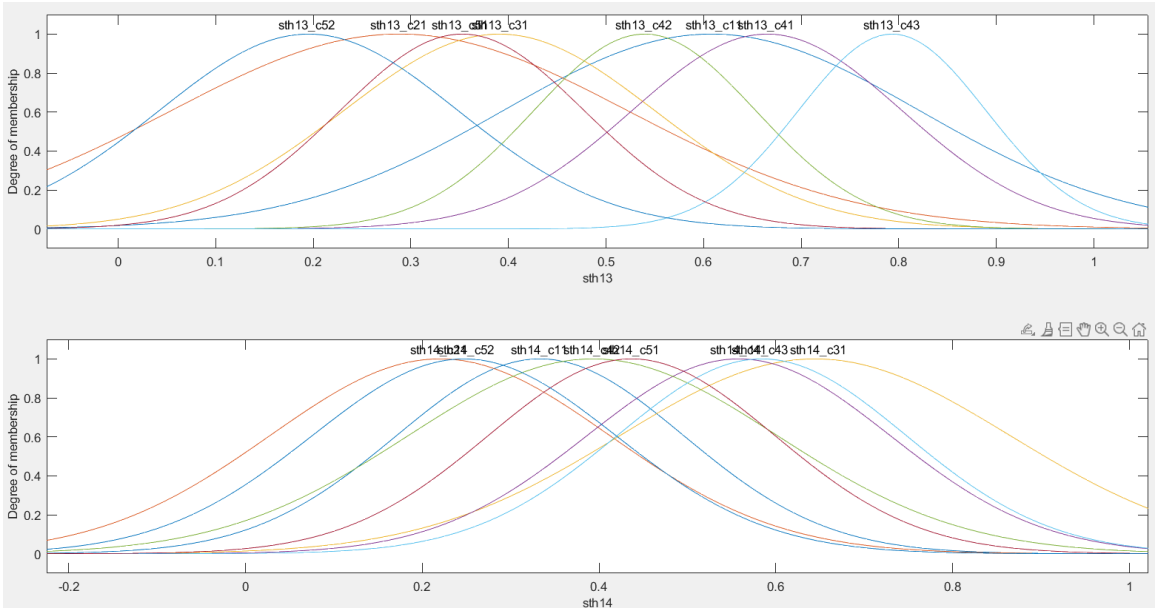

***Διάγραμμα 21**: Membership Functions Μετά την Εκπαίδευση του Βέλτιστου Μοντέλου (Inputs 1 - 4)*




***Διάγραμμα 22**: Membership Functions Μετά την Εκπαίδευση του Βέλτιστου Μοντέλου (Inputs 5 - 8)*




***Διάγραμμα 23**: Membership Functions Μετά την Εκπαίδευση του Βέλτιστου Μοντέλου (Inputs 9 - 12)*




***Διάγραμμα 23**: Membership Functions Μετά την Εκπαίδευση του Βέλτιστου Μοντέλου (Inputs 13 - 14)*

| | | | | |
|---|---|---|---|---|
| 376 | 0 | 0 | 0 | 0 |
| 58 | 454 | 3 | 6 | 2 |
| 19 | 5 | 421 | 72 | 7 |
| 7 | 1 | 34 | 370 | 51 |
| 0 | 0 | 2 | 12 | 400 |

***Πίνακας 8**: Πίνακας Σφαλμάτων Ταξινόμησης*

| OA | PA_1 | PA_2 | PA_3 | PA_4 | PA_5 |
|---|---|---|---|---|---|
| 0.8787 | 0.8174 | 0.9870 | 0.9152 | 0.8043 | 0.8696 |
| UA_1 | UA_2 | UA_3 | UA_4 | UA_5 | K |
| 1.0000 | 0.8681 | 0.8034 | 0.7991 | 0.9662 | 0.8787 |

***Πίνακας 9**: Τιμές Δεικτών Απόδοσης Τελικού Μοντέλου*

### Συμπεράσματα

Κρίνοντας από τα διαγράμματα που παρατέθηκαν παραπάνω, αλλά και από τις τιμές των δύο τελευταίων πινάκων, η απόδοση του μοντέλου είναι αρκετά ικανοποιητική. Όπως φαίνεται, η αύξηση των κανόνων (δηλαδή η μείωση της ακτίνας) δεν βελτιώνει πάντα το μοντέλο, κάτι που οφείλεται στην ασάφεια που προκαλείται από την επικάλυψη που παρατηρείται στις συναρτήσεις συμμετοχής. Όσον αφορά τους κανόνες, διαπιστώνουμε ότι είναι πολύ λίγοι – μόλις 8 στο τελικό μοντέλο. Στην περίπτωση όπου εξετάζονταν 14 χαρακτηριστικά με grid partitioning, θα υπήρχαν πολύ περισσότεροι κανόνες ($3^{12}$ ή $2^{12}$). Επιπλέον, παρατηρείται ότι, μερικές φορές, η καμπύλη σφάλματος εκπαίδευσης βρίσκεται κάτω από αυτή της δοκιμής, γεγονός που είναι τυχαίο.

## Βιβλιογραφικές Πηγές

1. Γιάννης Μπουτάλης, Γεώργιος Συρακούσης, "*Υπολογιστική Νοημοσύνη & Εφαρμογές*", Ξάνθη 2019