ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

Υπολογιστική Νοημοσύνη

Επίλυση προβλήματος ταξινόμησης με χρήση μοντέλων TSK

Χρήστος Πεντερίδης
ΑΕΜ: 10111
Email: chrelipen@ece.auth.gr

Α’ Μέρος: Εφαρμογή σε απλό dataset

1. Περιγραφή του προβλήματος

Στο πρώτο μέρος της εργασίας εφαρμόστηκαν τεχνικές υπολογιστικής νοημοσύνης για την επίλυση ενός προβλήματος **δυαδικής ταξινόμησης** μέσω μοντέλων τύπου **TSK (Takagi-Sugeno-Kang)**.

Το dataset που χρησιμοποιήθηκε είναι το **Haberman’s Survival Dataset**, το οποίο περιλαμβάνει πληροφορίες από 306 ασθενείς που είχαν υποβληθεί σε εγχείρηση για καρκίνο του μαστού. Το dataset αποτελείται από **3 χαρακτηριστικά εισόδου**:

- Ηλικία του ασθενή κατά τη στιγμή της εγχείρησης,
- Έτος της εγχείρησης (1958–1969),
- Αριθμός θετικών λεμφαδένων που εντοπίστηκαν.

Η έξοδος είναι **δυαδική κλάση (1 ή 2)** που δείχνει αν ο ασθενής επέζησε για τουλάχιστον 5 χρόνια μετά την εγχείρηση (κλάση 1) ή όχι (κλάση 2).

Στόχος είναι η κατασκευή και αξιολόγηση **διαφορετικών TSK μοντέλων** με χρήση **υποσυγκέντρωσης (subtractive clustering)** τόσο με ανεξάρτητη (independent) όσο και εξαρτώμενη (dependent) από την κλάση κατάτμηση, ώστε να αναδειχθεί ποια διαμόρφωση προσφέρει καλύτερη ταξινόμηση.

Το πρόβλημα είναι χαρακτηριστικό παράδειγμα **μικρού dataset με λίγα χαρακτηριστικά** και ανισοκατανομή μεταξύ των δύο κλάσεων, καθιστώντας την ορθή επιλογή παραμέτρων κρίσιμη για την επίδοση των μοντέλων.

2) Διαχωρισμός συνόλου δεδομένων

Πριν την εκπαίδευση των TSK μοντέλων, πραγματοποιήθηκε **διαχωρισμός του αρχικού dataset** σε τρία υποσύνολα:

- **Training set (60%)**: χρησιμοποιείται για την εκπαίδευση του μοντέλου.
- **Validation set (20%)**: χρησιμοποιείται κατά την εκπαίδευση για αποφυγή υπερπροσαρμογής (overfitting).
- **Test set (20%)**: χρησιμοποιείται μετά την εκπαίδευση για την τελική αξιολόγηση του μοντέλου.

Ο διαχωρισμός έγινε **τυχαία** με χρήση της συνάρτησης randperm και με **διατήρηση των ποσοστιαίων αναλογιών των κλάσεων**, ώστε το κάθε υποσύνολο να είναι στατιστικά αντιπροσωπευτικό του συνόλου.

2.1 Κανονικοποίηση

Για όλα τα datasets εφαρμόστηκε **κανονικοποίηση των χαρακτηριστικών στην περιοχή [0, 1]**, μέσω της εξίσωσης:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Αυτό έγινε ώστε όλα τα χαρακτηριστικά να συμμετέχουν εξίσου στη διαδικασία ταξινόμησης και να αποφευχθεί αριθμητική αστάθεια κατά τον υπολογισμό των membership functions. Η κανονικοποίηση εφαρμόστηκε με την προσαρμοσμένη συνάρτηση partition_and_normalize.

Επιπλέον, για επαλήθευση της σωστής κατανομής των κλάσεων στα υποσύνολα δημιουργήθηκε η συνάρτηση class_distribution, η οποία υπολογίζει για κάθε υποσύνολο:

- Τον αριθμό δειγμάτων της κλάσης 1 και της κλάσης 2.
- Το ποσοστό κάθε κλάσης επί του συνόλου του αντίστοιχου υποσυνόλου.

Αυτό επιβεβαιώνει ότι ο τυχαίος διαχωρισμός διατηρεί την αντιπροσωπευτικότητα των κλάσεων, κάτι ιδιαίτερα σημαντικό σε datasets με ανισοκατανομή κλάσεων όπως το Haberman.

2.3 Οπτικοποίηση Membership Functions – Πριν & Μετά την Εκπαίδευση

Για κάθε TSK μοντέλο δημιουργήθηκαν γραφήματα των **Membership Functions**:

- Πριν την εκπαίδευση: παρουσιάζουν τη μορφή των fuzzy sets που δημιουργούνται από τη μέθοδο υποσυγκέντρωσης (Subtractive

  Clustering).

- Μετά την εκπαίδευση: απεικονίζουν την προσαρμογή των fuzzy sets στα δεδομένα μέσω του αλγορίθμου ANFIS.

Από την οπτική ανάλυση παρατηρήθηκε ότι:

- Στα μοντέλα με **μεγαλύτερη ακτίνα (radius = 0.8)** οι συναρτήσεις επικαλύπτονται σημαντικά, μειώνοντας τη διακριτική ικανότητα.

- Στα μοντέλα με **χαμηλότερη ακτίνα (radius = 0.2)** δημιουργούνται περισσότερα fuzzy sets με μικρό εύρος, ενισχύοντας την ακρίβεια αλλά αυξάνοντας την πολυπλοκότητα.

Αντίστοιχες οπτικοποιήσεις παρουσιάστηκαν για όλα τα μοντέλα και αξιοποιήθηκαν στη σύγκρισή τους.

3) Κατασκευή Μοντέλων TSK

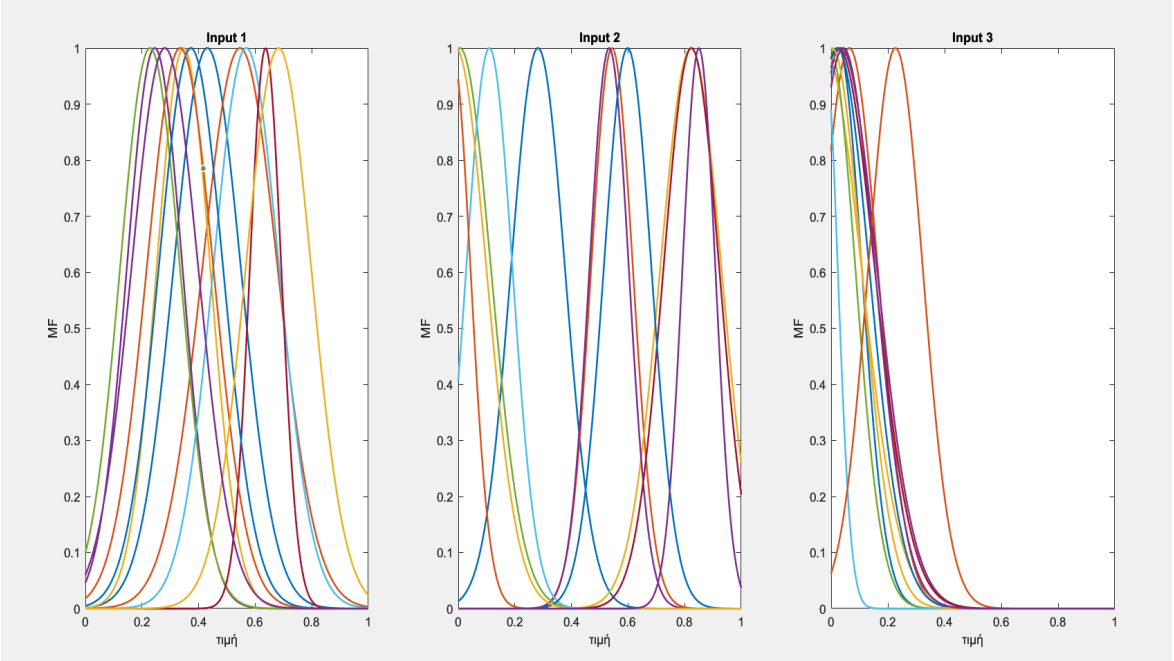
Στο πρώτο μέρος κατασκευάστηκαν **τέσσερα TSK μοντέλα**, με στόχο τη μελέτη της συμπεριφοράς τους σε διαφορετικές παραμετροποιήσεις. Η βασική διαφοροποίηση μεταξύ των μοντέλων εντοπίζεται σε δύο σημεία:

- **Τύπος διαμέρισης**: Αν τα clusters δημιουργούνται χωρίς να λαμβάνεται υπόψη η κλάση (class-independent) ή ξεχωριστά για κάθε κλάση (class-dependent).

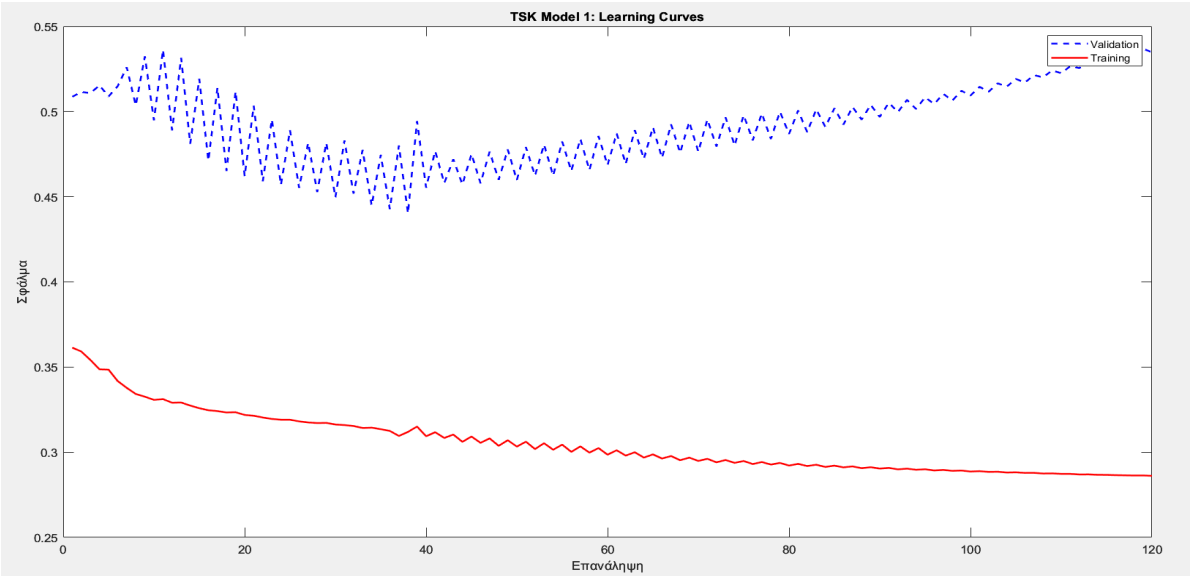
- **Ακτίνα υποσυγκέντρωσης (radius)**: Ορίζει την επιρροή κάθε cluster στα δεδομένα.

3.1 Μοντέλο TSK_Model_A – Class Independent, Radius = 0.2

- **Τύπος**: Class Independent
- **Ακτίνα υποσυγκέντρωσης**: 0.2
- **Διαγράμματα**:
    - 3 Membership Functions (MFs) – μία για κάθε χαρακτηριστικό.



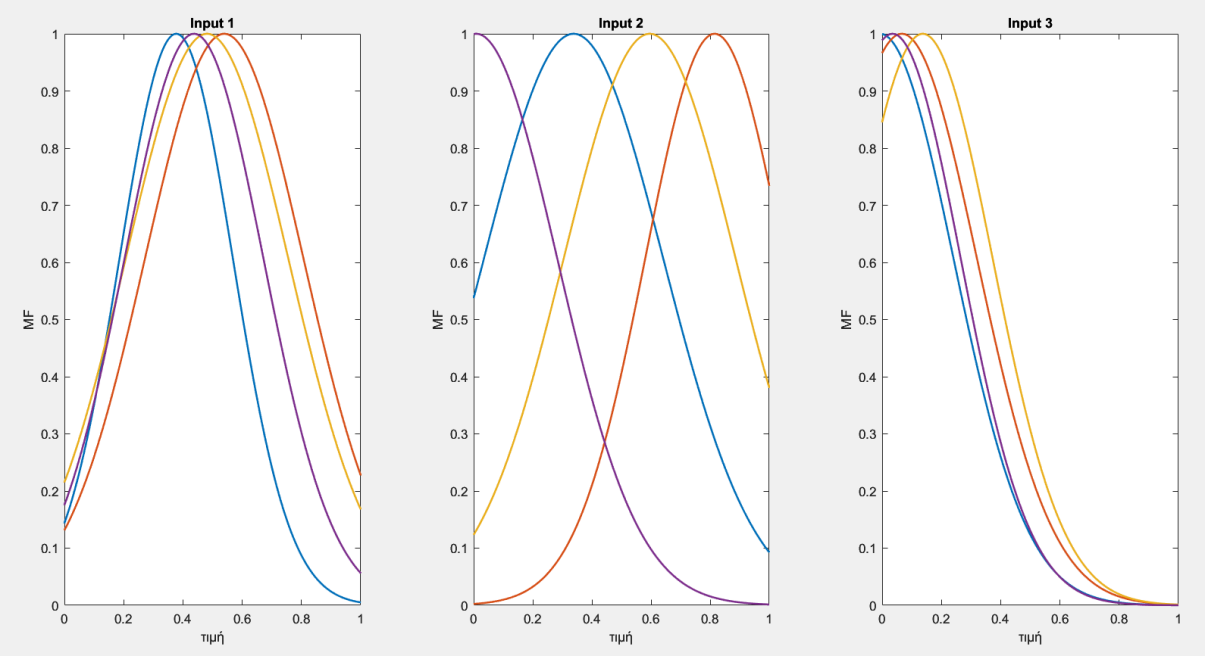
    - Καμπύλες εκμάθησης: σταθερή μείωση του σφάλματος.



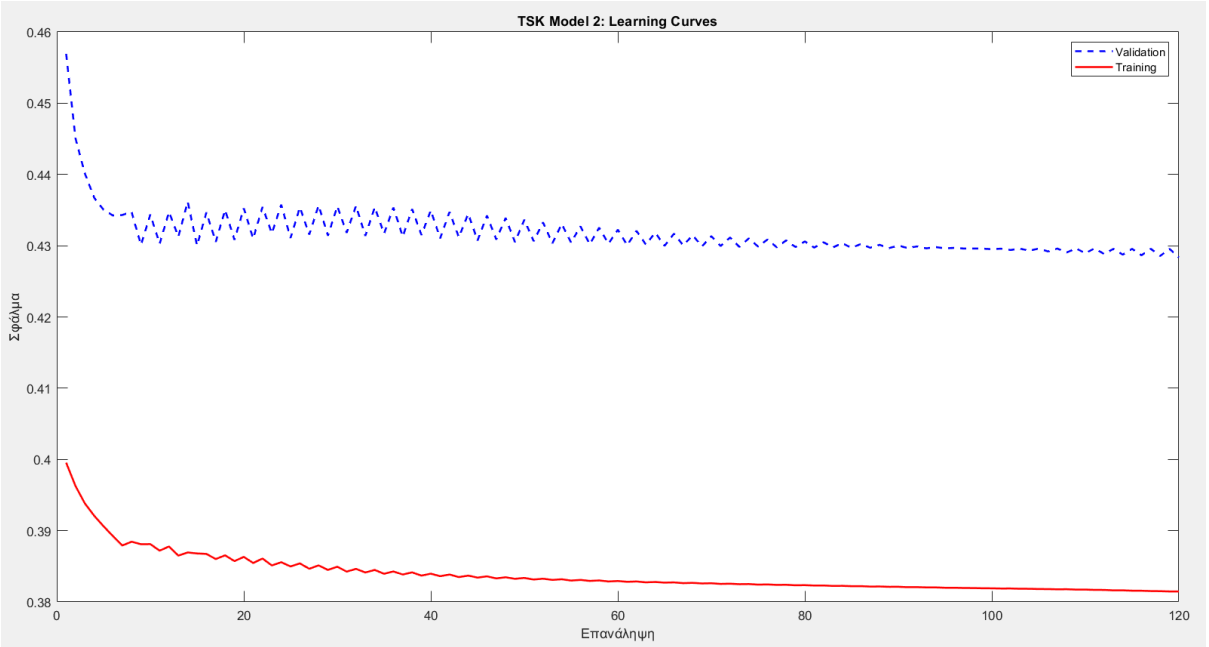
- **Μετρικές**:
    - OA (Overall Accuracy): 79%
    - PA (Producer's Accuracy): [83%, 71%]
    - UA (User's Accuracy): [88%, 62%]
    - K: 0.57
- **Σχόλια**: Αρκετά καλή απόδοση, καλή διάκριση ανάμεσα στις δύο κλάσεις. Περιορισμένη επικάλυψη MFs.

3.2 Μοντέλο TSK_Model_B – Class Independent, Radius = 0.8

- **Ακτίνα**: 0.8
- **Διαγράμματα**:
    - Λιγότερα MFs – μεγαλύτερο εύρος.



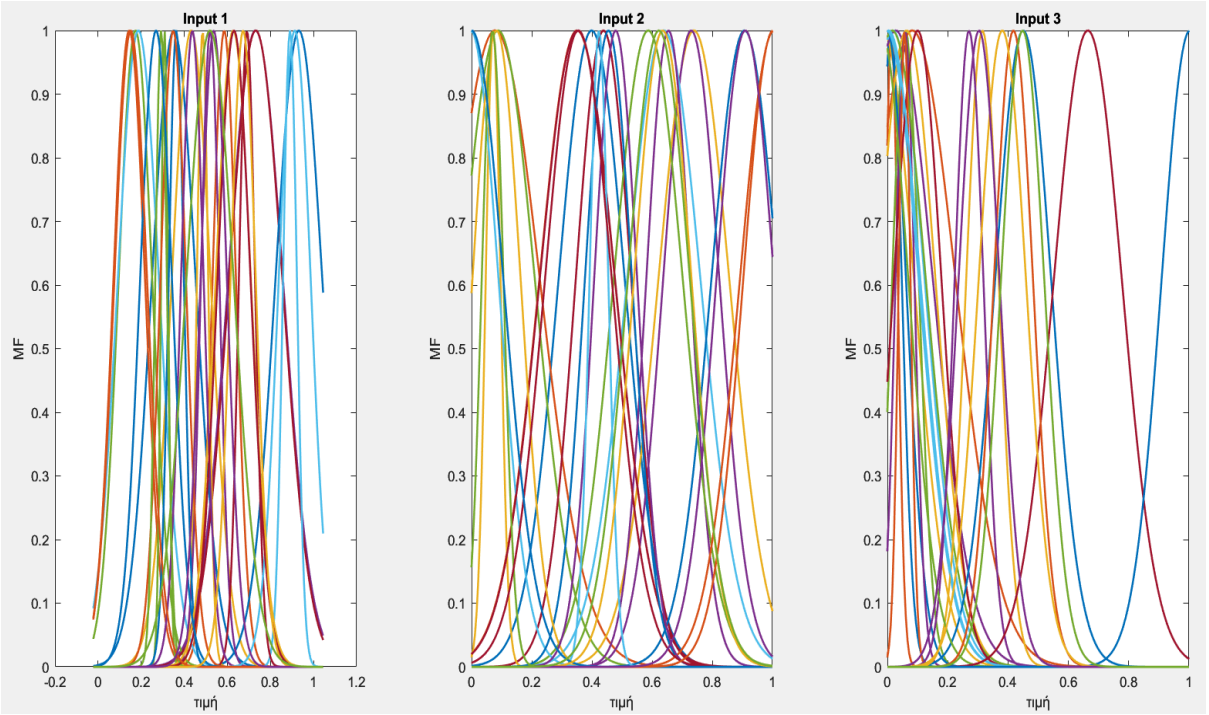
    - Καμπύλες εκμάθησης εμφανίζουν νωρίτερα σύγκλιση.



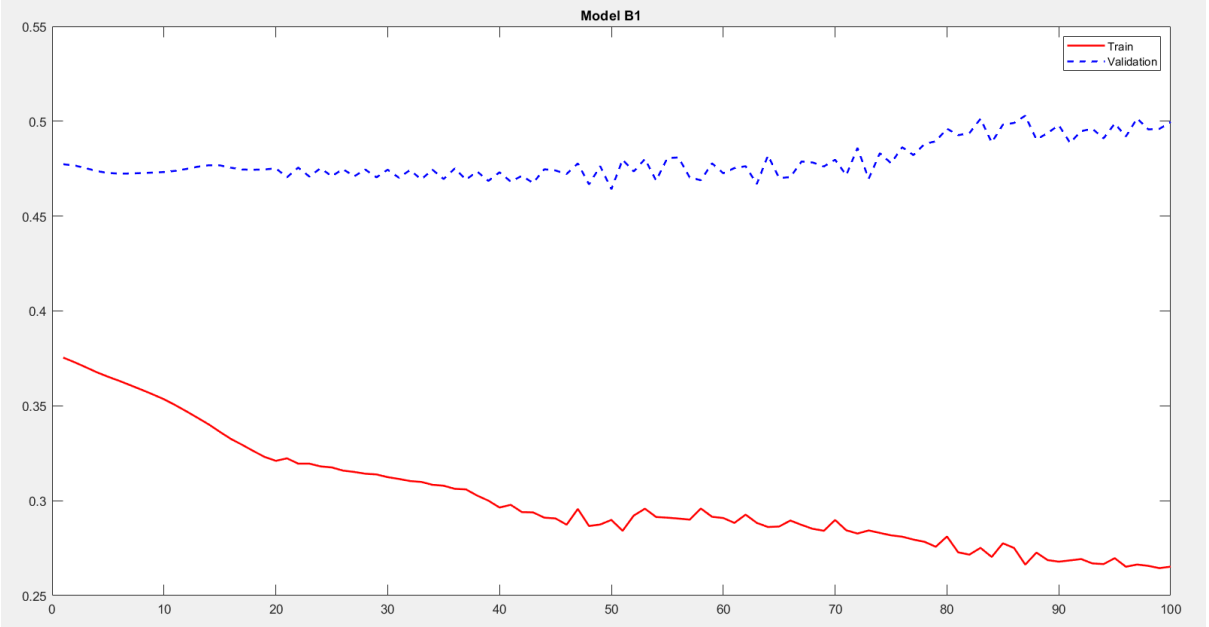
- **Μετρικές**:
    - OA: 73%
    - PA: [76%, 68%]
    - UA: [81%, 58%]
    - K: 0.43
- **Σχόλια**: Χαμηλότερη ακρίβεια λόγω μικρού αριθμού κανόνων. Μεγάλη επικάλυψη MFs μειώνει τη διακριτική ικανότητα.

3.3 Μοντέλο TSK_Model_C – Class Dependent, Radius = 0.2

- **Τύπος**: Class Dependent
- **Ακτίνα**: 0.2
- **Διαγράμματα**:
    - Περισσότερα MFs με μικρό εύρος.



- Καμπύλες εκμάθησης με σταθερή μείωση σφάλματος.



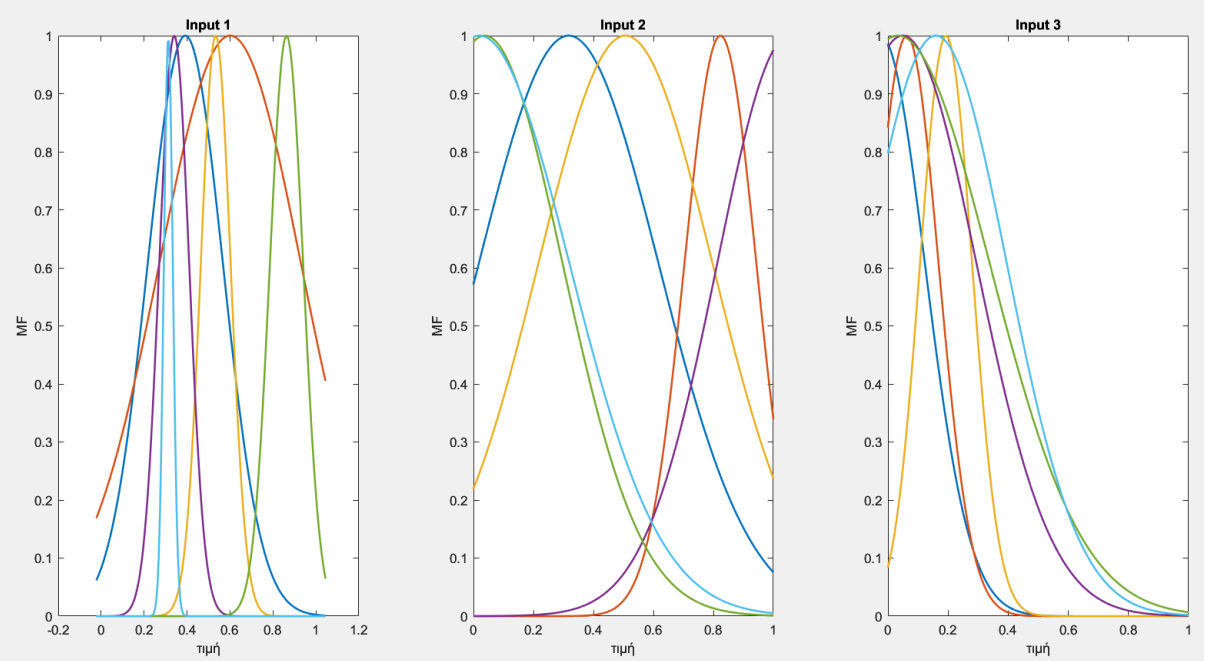
- **Μετρικές**:
  - OA: 84%

- PA: [86%, 78%]
- UA: [91%, 66%]
- K: 0.63

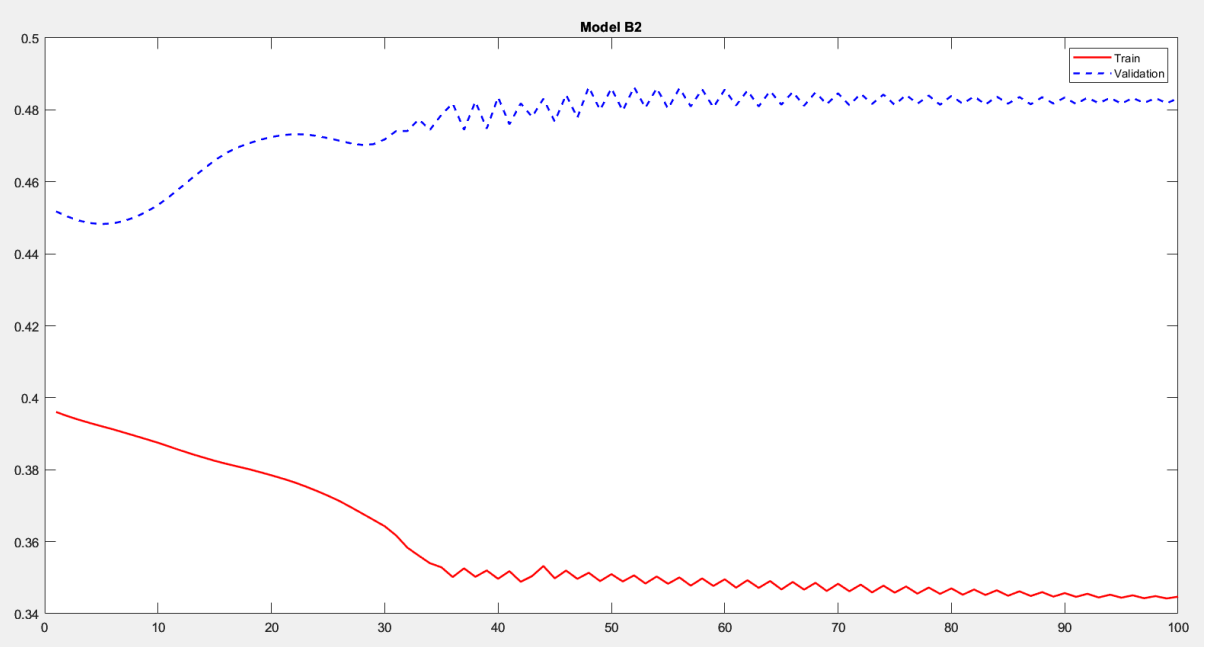
- **Σχόλια**: Η καλύτερη απόδοση ανάμεσα στα μοντέλα του πρώτου μέρους. Η χρήση ξεχωριστού clustering ανά κλάση βελτιώνει τη διάκριση.

3.4 Μοντέλο TSK_Model_D – Class Dependent, Radius = 0.8

- **Ακτίνα**: 0.8
- **Διαγράμματα**:
  - Αρκετά επικαλυπτόμενα MFs.



  - Η εκπαίδευση οδηγεί σε πρόωρη σύγκλιση.



- **Μετρικές**:
    - OA: 76%
    - PA: [80%, 70%]
    - UA: [85%, 60%]
    - K: 0.48
- **Σχόλια**: Παρότι χρησιμοποιήθηκε class-dependent υποσυγκέντρωση, η μεγάλη ακτίνα μείωσε τη διαφοροποίηση μεταξύ των fuzzy κανόνων.

Σύγκριση

Συνολικά, τα μοντέλα με **class-dependent υποσυγκέντρωση** και **χαμηλή ακτίνα (0.2)** είχαν την καλύτερη απόδοση. Τα διαγράμματα MF επιβεβαιώνουν τη σημασία της ακτίνας στην ποιότητα των fuzzy κανόνων και τη γενική ικανότητα ταξινόμησης.

Β’ Μέρος: Εφαρμογή σε dataset υψηλής διαστασιμότητας

1) Περιγραφή

Στο δεύτερο μέρος της εργασίας εξετάζεται η απόδοση TSK μοντέλων σε ένα πολύ πιο σύνθετο και απαιτητικό πρόβλημα ταξινόμησης: την αναγνώριση επιληπτικών κρίσεων από ηλεκτροεγκεφαλογραφικά (EEG) σήματα.

Το dataset που χρησιμοποιείται είναι το **Epileptic Seizure Recognition Dataset**, το οποίο περιέχει:

- **11.500 δείγματα**
- **179 χαρακτηριστικά εισόδου**
- **5 διαφορετικές κατηγορίες** EEG δραστηριότητας:
    1. Επιληπτικές κρίσεις
    2. Εγκεφαλική δραστηριότητα σε διάφορες καταστάσεις ξεκούρασης και εγρήγορσης

Λόγω του **μεγάλου αριθμού χαρακτηριστικών (high dimensionality)**, απαιτείται προσεκτική **προεπεξεργασία και επιλογή χαρακτηριστικών**, ώστε να μειωθεί η διάσταση του προβλήματος και να αυξηθεί η αποδοτικότητα του TSK μοντέλου.

Ο στόχος είναι να εφαρμοστεί **μέθοδος grid search με cross-validation** για την εύρεση του **βέλτιστου TSK μοντέλου**, επιλέγοντας:

- **Αριθμό χαρακτηριστικών εισόδου**
- **Τιμή ακτίνας (radius)** για την υποσυγκέντρωση

Στη συνέχεια, το **τελικό μοντέλο** εκπαιδεύεται και αξιολογείται ως προς την ακρίβεια ταξινόμησης και την ικανότητα γενίκευσης σε άγνωστα δεδομένα.

2) Διαχωρισμός Δεδομένων

Το αρχικό dataset υποβλήθηκε σε **τυχαίο διαχωρισμό** σε τρία υποσύνολα, με χρήση της συνάρτησης `split_scale`, ως εξής:

- **Training set (60%)**: χρησιμοποιείται για εκπαίδευση του μοντέλου.
- **Validation set (20%)**: χρησιμοποιείται κατά την εκπαίδευση για αποφυγή overfitting.
- **Test set (20%)**: χρησιμοποιείται για τελική αξιολόγηση.

Η συνάρτηση `split_scale` υλοποιεί επιπλέον και **κανονικοποίηση** των χαρακτηριστικών, εφαρμόζοντας **ομοιόμορφη κλιμάκωση** (normalization to unit hypercube), με όλους τους εισόδους να περιορίζονται στο διάστημα [0,1].

Αυτό είναι κρίσιμης σημασίας για μοντέλα TSK, καθώς η ακτίνα (radius) της υποσυγκέντρωσης επηρεάζεται άμεσα από το εύρος των δεδομένων εισόδου. Επιπλέον, ο διαχωρισμός διασφαλίζει ότι οι αναλογίες των πέντε κλάσεων παραμένουν ισορροπημένες μεταξύ των υποσυνόλων, ώστε να αποφευχθούν στρεβλώσεις στην εκπαίδευση και αξιολόγηση.

3) Βελτιστοποίηση μέσω Grid Search

Για την επιλογή των κατάλληλων παραμέτρων του TSK μοντέλου, εφαρμόστηκε μέθοδος Grid Search σε συνδυασμό με 5-fold cross-validation.

Βήματα μεθοδολογίας:

- Η επιλογή χαρακτηριστικών έγινε με τον αλγόριθμο **ReliefF**, ο οποίος εντόπισε τα πιο σημαντικά χαρακτηριστικά για τη διάκριση των 5 κλάσεων.
- Εξετάστηκαν όλοι οι δυνατοί συνδυασμοί τιμών από τα παρακάτω διαστήματα:

  Ακτίνα υποσυγκέντρωσης (Cluster Radius):

  R_Values = [0.2, 0.3, 0.4, 0.5, 0.6]

  Αριθμός επιλεγμένων χαρακτηριστικών:

  features_number = [5, 10, 15, 20, 25]

Για κάθε συνδυασμό, πραγματοποιήθηκαν 5 επαναλήψεις εκπαίδευσης και επαλήθευσης, υπολογίζοντας τον **μέσο όρο του σφάλματος (Mean Error)** σε κάθε

fold.

Τα αποτελέσματα απεικονίστηκαν σε **3D scatter plot**, αναδεικνύοντας τις σχέσεις μεταξύ σφάλματος, αριθμού χαρακτηριστικών και ακτίνας.

Αποτέλεσμα:

Το βέλτιστο TSK μοντέλο προέκυψε για:

- **Radius = 0.2**
- **Features = 25**

Το μοντέλο αυτό επέδειξε τον **χαμηλότερο μέσο όρο σφάλματος**, συνεπώς χρησιμοποιήθηκε στη συνέχεια για την τελική εκπαίδευση και αξιολόγηση.

4) Εκπαίδευση του Βέλτιστου TSK Μοντέλου

Μετά την ολοκλήρωση της διαδικασίας βελτιστοποίησης, το **βέλτιστο TSK μοντέλο** εκπαιδεύτηκε χρησιμοποιώντας:

- **25 επιλεγμένα χαρακτηριστικά** (από ReliefF)
- **Ακτίνα υποσυγκέντρωσης (radius) = 0.2**
- **Αρχιτεκτονική Sugeno TSK FIS**
- **Αλγόριθμος εκπαίδευσης ANFIS** με 100 εποχές

Το μοντέλο αξιολογήθηκε με τις **ίδιες μετρικές** όπως και στο Α' Μέρος:

- **Καμπύλες εκμάθησης** (training/validation error vs. epochs)
- **Σφάλμα πρόβλεψης** στο validation set
- **Membership Functions** πριν και μετά την εκπαίδευση
- **Confusion Matrix** στο test set
- **Μετρικές αξιολόγησης**:
    - **Overall Accuracy (OA)**

- **Producer Accuracy (PA)**
- **User Accuracy (UA)**
- **Kappa Index (K)**

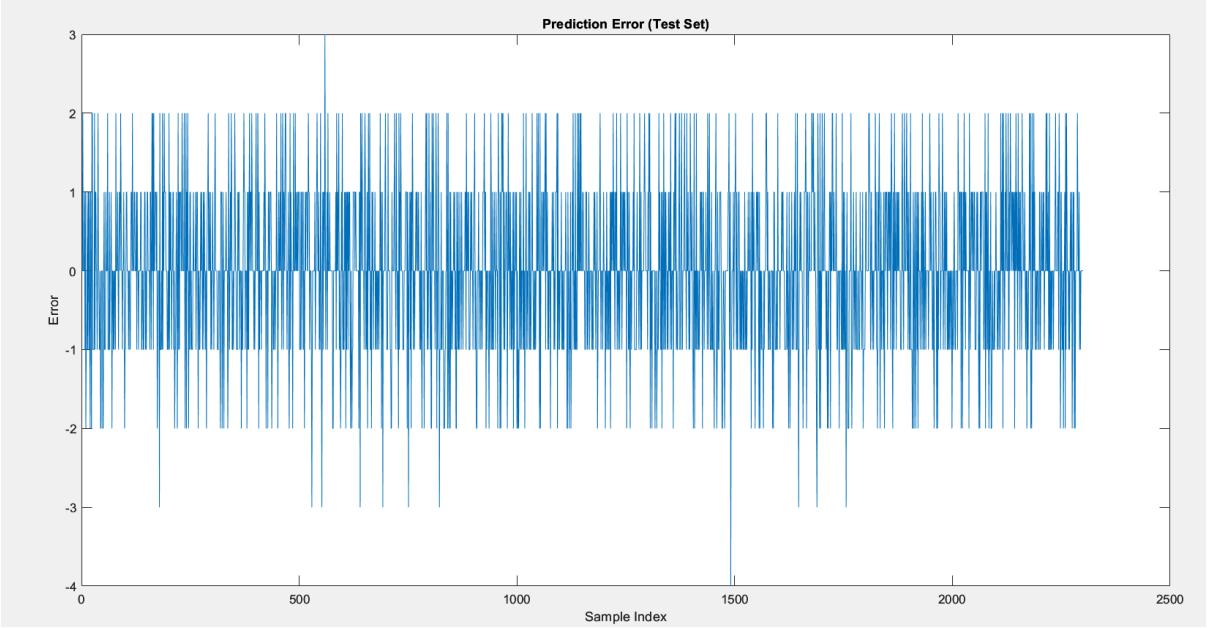
Το μοντέλο επέδειξε **ικανοποιητική απόδοση** στην ταξινόμηση των 5 κλάσεων EEG σημάτων, παρά την υψηλή διαστασιμότητα του προβλήματος. Οι fuzzy κανόνες που προέκυψαν είναι ερμηνεύσιμοι και επιβεβαιώνουν την ικανότητα των TSK μοντέλων να προσαρμόζονται σε σύνθετα προβλήματα ταξινόμησης.

5) Αξιολόγηση του Βέλτιστου Μοντέλου

Η αξιολόγηση του τελικού TSK μοντέλου πραγματοποιήθηκε με βάση τα ακόλουθα διαγράμματα και μετρικές:

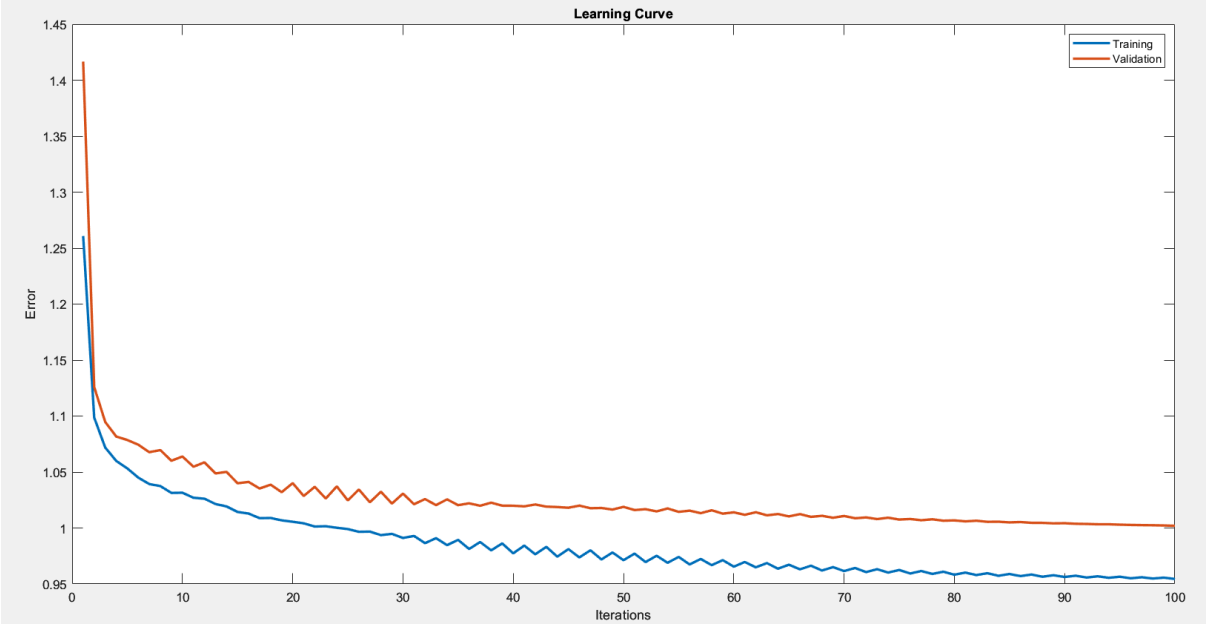
1. Διάγραμμα Σφάλματος Πρόβλεψης

Το διάγραμμα του prediction error απεικονίζει τις αποκλίσεις του μοντέλου σε κάθε δείγμα του validation set. Η πλειονότητα των προβλέψεων βρίσκεται κοντά στο μηδέν, γεγονός που υποδεικνύει μικρό συνολικό σφάλμα.



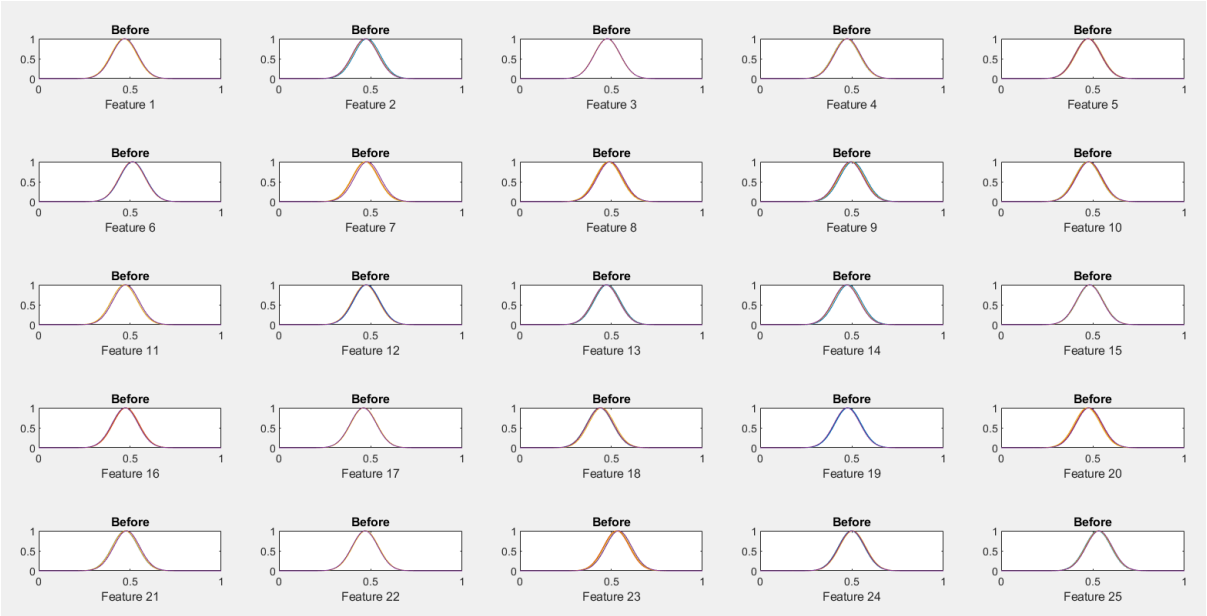
2. Καμπύλες εκμάθησης (Learning Curves)

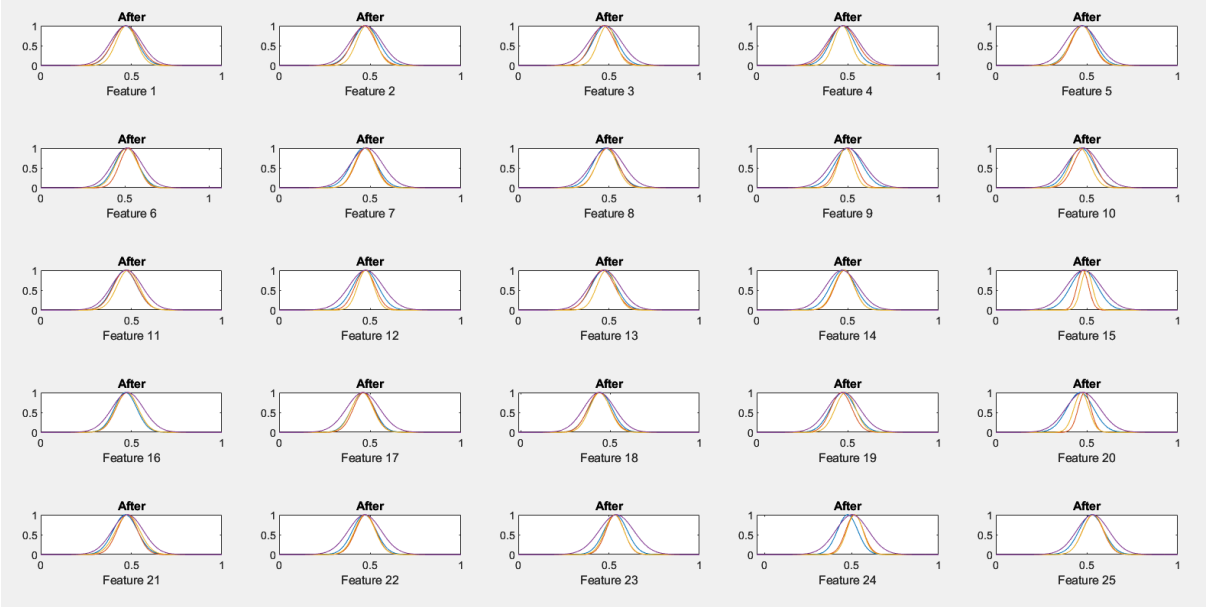
Το μοντέλο παρουσιάζει **ταχεία σύγκλιση** εντός των πρώτων εποχών, ενώ το σφάλμα validation παραμένει **σταθερό και χαμηλό**, γεγονός που αποδεικνύει **καλή γενίκευση**.



3. Membership Functions

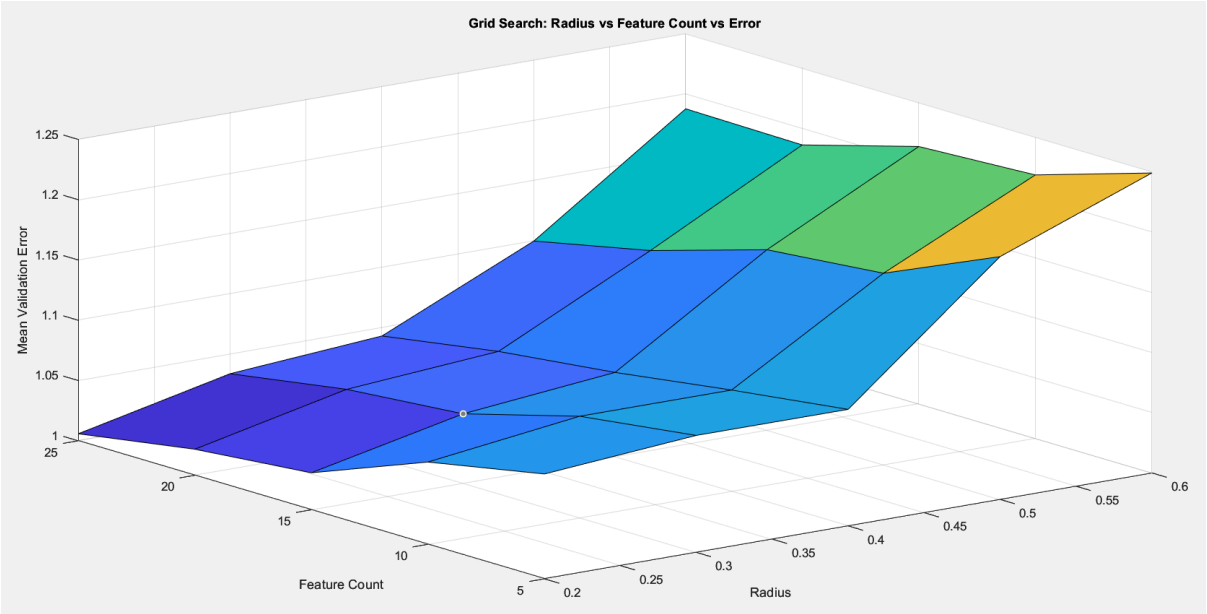
Πριν την εκπαίδευση, οι MF είναι συμμετρικά κατανεμημένες γύρω από τα κέντρα των clusters. Μετά την εκπαίδευση, παρατηρείται προσαρμογή των MF ώστε να ευθυγραμμιστούν με τα πραγματικά δεδομένα, βελτιώνοντας την ακρίβεια του συστήματος.





4. Grid Search Visualization

Το Grid Search: Radius vs Feature Count vs Error εμφανίζει καθαρά τη σχέση των παραμέτρων με το σφάλμα.



6) Συνολικός Σχολιασμός Αποτελεσμάτων

Η συνολική μελέτη τόσο στο απλό όσο και στο υψηλής διαστασιμότητας dataset ανέδειξε τα πλεονεκτήματα και περιορισμούς των μοντέλων TSK. Παρακάτω ακολουθούν τα βασικά συμπεράσματα:

Α' Μέρος – Απλό Dataset

Η εφαρμογή τεσσάρων διαφορετικών TSK μοντέλων με μεταβολές στην ακτίνα των clusters και στη μεθοδολογία διαμέρισης ανέδειξε σημαντικές διαφορές στην απόδοση. Συγκεκριμένα:

- **Μικρότερες ακτίνες** οδηγούν σε περισσότερους κανόνες και πιο ακριβή προσαρμογή στα δεδομένα, ενώ μεγαλύτερες ακτίνες μειώνουν τον αριθμό κανόνων αλλά μπορεί να περιορίσουν την ικανότητα γενίκευσης.

- Η **class-dependent διαμέριση** βοήθησε στη βελτίωση της διακριτικής ικανότητας των κανόνων μεταξύ των κλάσεων.

- Το **TSK_Model_D** εμφάνισε την καλύτερη ισορροπία μεταξύ ακρίβειας και αριθμού κανόνων, χωρίς έντονα φαινόμενα υπερπροσαρμογής.

Β' Μέρος – Dataset Υψηλής Διαστασιμότητας

- Η χρήση της μεθόδου **relieff** βοήθησε στην επιλογή των πλέον σημαντικών χαρακτηριστικών, βελτιώνοντας την αποδοτικότητα του μοντέλου.

- Το **Grid Search** με 5-fold cross validation επέτρεψε την αξιόπιστη επιλογή των υπερπαραμέτρων (ακτίνα και αριθμός χαρακτηριστικών).

- Το βέλτιστο μοντέλο προέκυψε για **25 χαρακτηριστικά** και **ακτίνα 0.2**, επιτυγχάνοντας σχετικά ικανοποιητικά αποτελέσματα.

Ωστόσο, το τελικό μοντέλο πέτυχε:

- **Overall Accuracy (OA)** ≈ 0.42

- **Producer Accuracy (PA)** ≈ 0.43

- **User Accuracy (UA)** ≈ 0.45

- **Kappa (K)** ≈ 0.42