Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρονικής και Υπολογιστών

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Εργασία 4η : Classification με χρήση TSK μοντέλων

Εαρινό εξάμηνο 2023/24

Δεϊρμεντζόγλου Ιωάννης
Τομέας Ηλεκτρονικής και Υπολογιστών
A.E.M.: 10015
Email: deirmentz@ece.auth.gr

Περιεχόμενα

ΕΙΣΑΓΩΓΗ

Η παρούσα εργασία έχει ως στόχο την επίλυση προβλήματος ταξινόμησης μέσω της εφαρμογής και αξιολόγησης ασαφών νευρωνικών μοντέλων TSK. Πιο συγκεκριμένα, μέσω της χρήσης ενός dataset 306 στοιχείων, γίνεται εκπαίδευση του Fuzzy Intelligence System (FIS), με την μέθοδο του Subtractive Clustering, ώστε να προκύψουν τα ασαφή σύνολα των μεταβλητών εισόδου. Συνολικά εκπαιδεύονται 4 TSK μοντέλα.

Στο πρώτο μέρος της εργασίας, εξετάζεται ένα απλό σύνολο δεδομένων, το **Haberman's Survival**, το οποίο αποτελείται από 306 δείγματα και τρία χαρακτηριστικά. Εδώ, τα TSK μοντέλα εκπαιδεύονται με διάφορες παραμέτρους, και γίνεται σύγκριση της απόδοσής τους σε υποσύνολα εκπαίδευσης, επικύρωσης και ελέγχου, με Subtractive Clustering. Στόχος είναι να μελετηθεί η επίδραση του αριθμού των ασαφών κανόνων και της πολυπλοκότητας του μοντέλου στην ακρίβεια ταξινόμησης.

Στο δεύτερο μέρος, η ανάλυση προχωρά σε πιο πολύπλοκο σύνολο δεδομένων, το **Epileptic Seizure Recognition**, το οποίο περιέχει 11.500 δείγματα και 179 χαρακτηριστικά. Λόγω της υψηλής διαστασιμότητας, εφαρμόζονται τεχνικές μείωσης χαρακτηριστικών και clustering, προκειμένου να περιοριστεί η έκρηξη κανόνων (rule explosion) και να επιτευχθεί καλύτερη απόδοση. Επιπλέον, χρησιμοποιείται η μέθοδος της διασταυρωμένης επικύρωσης (cross validation) για τη βελτιστοποίηση των παραμέτρων του μοντέλου, ενώ η διαδικασία αξιολογείται μέσω δεικτών απόδοσης όπως η συνολική ακρίβεια και η ακρίβεια ταξινόμησης ανά κλάση. Το βέλτιστο μοντέλο με το ελάχιστο μέσο σφάλμα, αξιολογείται με ορισμένες παραμέτρους ως προς την απόδοση του training.

Η εργασία ολοκληρώνεται με την ανάλυση των αποτελεσμάτων και τη σύγκριση των αποδόσεων των μοντέλων, παρέχοντας σημαντικά συμπεράσματα για τη βελτίωση της ακρίβειας και της αποδοτικότητας των TSK μοντέλων στην ταξινόμηση.

1. Εφαρμογή σε απλό dataset

Εισαγωγή Δεδομένων

Αρχικά, εισάγονται τα δεδομένα από το αρχείο “haberman.dat”. Το αρχείο αυτό περιέχει τα δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση και την αξιολόγηση των μοντέλων. Το σύνολο αυτό dataset περιέχει 306 δείγματα με 3 features το καθένα. Το dataset περιέχει μία ακόμα στήλη με τον αριθμό της κλάσης που ανήκει το κάθε στοιχείο (class 1 ή class 2).

Διαχωρισμός Δεδομένων

Η γραμμή διαχωρίζει το σύνολο δεδομένων σε τρία μέρη: δεδομένα εκπαίδευσης (60%), δεδομένα επικύρωσης (20%) και δεδομένα δοκιμών (20%). Το σύνολο εκπαίδευσης (training set) χρησιμοποιείται για την εκπαίδευση των μοντέλων, ενώ το σύνολο επικύρωσης (validation set) για τη ρύθμιση των υπερπαραμέτρων και για αποφυγή του δεδομένου της υπερεκπαίδευσης και το σύνολο δοκιμών (test set) για την τελική αξιολόγηση και τον υπολογισμό των μετρικών έπειτα από το training. Ο διαχωρισμός των δεδομένων είναι 60%, 20% και 20% αντίστοιχα.

Η διαφορά σε αυτή την περίπτωση είναι ότι για να επιτευχθεί καλή απόδοση θα πρέπει η συχνότητα εμφάνισης δειγμάτων που ανήκουν σε μία συγκεκριμένη κλάση, σε κάθε υποσύνολο της εκπαίδευσης, να είναι όσο το δυνατόν πιο όμοια με την αρχική συχνότητα εμφάνισης στο συνολικό dataset. Για να επιτευχθεί αυτό μια μέθοδος είναι η ταξινόμηση των δεδομένων ως προς την κλάση που ανήκουν (τελευταία στήλη του dataset). Έπειτα, λόγω της μεθόδου διαχωρισμού που εφαρμόζεται με την χρήση της split_scale συνάρτησης, θα υπάρχει μία αρκετά όμοια συχνότητα εμφάνισης της κάθε κλάσης σε κάθε υποσύνολο δεδομένων.

Για τον παραπάνω λόγο υλοποιείται και η συναρτηση calculate_frequencies() που έχει ως παραμέτρους: trnData (δεδομένα εκπαίδευσης), chkData (δεδομένα ελέγχου) tstData (δεδομένα δοκιμής) . Για κάθε υποσύνολο δεδομένων (εκπαίδευσης, ελέγχου, δοκιμής), η συνάρτηση υπολογίζει πόσες φορές εμφανίζεται η κλάση με ID=1 και πόσες φορές εμφανίζεται η κλάση με ID=2 και τα ποσοστά (%) της κάθε κλάσης.

Η συναρτηση βρίσκεται στο αρχείο **calculate_frequencies.m**

Σε κάθε ένα από αυτά τα σύνολα δεδομένων, η 4η στήλη περιέχει το αναγνωριστικό (ID) της κλάσης στην οποία ανήκει το κάθε δείγμα.

Για το shuffling και τον διαχωρισμό των δεδομένων χρησιμοποιήθηκε η συνάρτηση *split_scale,* που βρίσκεται στο αρχείο split_scale.m από τα μοντέλα TSK που είναι αναρτημένα στην ιστοσελίδα του μαθήματος.

Η συναρτηση **split_scale** έχει ως εισόδους τις μεταβλητές : `data`(τα δεδομένα που θα διαχωριστούν και θα προεπεξεργαστούν) και `preproc`(παράμετρος που καθορίζει το είδος προεπεξεργασίας) . Έχει ως εξόδους τα σύνολα `trnData` (δεδομένα

εκπαίδευσης) , `valData` (δεδομένα επικύρωσης) και `tstData` (δεδομένα δοκιμών). Αρχικά, τα δεδομένα ανακατεύονται τυχαία μέσω της συνάρτησης `randperm`, η οποία επιστρέφει έναν τυχαίο πίνακα με δείκτες. Στη συνέχεια, τα δεδομένα χωρίζονται και ανάλογα με την τιμή της παραμέτρου `preproc`, εφαρμόζονται δύο είδη προεπεξεργασίας:

- Περίπτωση 1: Κανονικοποίηση (Normalization) στο unit hypercube.
- Περίπτωση 2: Τυποποίηση (Standardization) σε μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση.

TSK Μοντέλα Εκπαίδευσης

Η ακτίνα ενός cluster καθορίζει το μέγεθος της περιοχής στον χώρο των δεδομένων στην οποία κάθε cluster θα ασκήσει επιρροή. Όσο μικρότερη είναι η ακτίνα, τόσο περισσότερα clusters δημιουργούνται, επειδή μικρές περιοχές του χώρου των δεδομένων καλύπτονται από κάθε cluster. Όσο μεγαλύτερη είναι η ακτίνα, τόσο λιγότερα clusters δημιουργούνται, καθώς κάθε cluster καλύπτει μια μεγαλύτερη περιοχή.

Τα clusters στον χώρο εισόδου δημιουργούν τις βάσεις για τους IF-THEN κανόνες σε ένα TSK μοντέλο. Κάθε cluster αντιστοιχεί ουσιαστικά σε έναν κανόνα IF-THEN. Όταν η ακτίνα είναι μικρή, δημιουργούνται περισσότερα clusters, άρα προκύπτουν περισσότεροι κανόνες. Αυτό σημαίνει αυξημένη πολυπλοκότητα του μοντέλου, καθώς το σύστημα χρησιμοποιεί περισσότερους κανόνες για να ταξινομήσει τα δεδομένα.

Από την εκφώνηση ζητείται να εκπαιδευθούν 4 TSK μοντέλα για το παραπάνω dataset, οπου τα χαρακτηριστικά τους συνοψίζονται στον παρακάτω πίνακα:

Μοντέλο	Διαμέριση Εισόδου	Τύπος Εξόδου	Cluster Radius
TSK_Model_1	Class Independent	Singleton	0.1 (Small)
TSK_Model_2	Class Independent	Singleton	0.9 (Large)
TSK_Model_3	Class Dependent	Singleton	0.1 (Small)
TSK_Model_4	Class Dependent	Singleton	0.9 (Large)

Η κύρια διαφορά των μοντέλων είναι ο τρόπος διαμέρισης του χώρου εισόδου, class dependent όπου διαμερίζεται για το σύνολο των δεδομένων και class independent, όπου η διαμέριση γίνεται ξεχωριστά για κάθε κλάση δεδομένων. Για κάθε τρόπο διαχωρισμού του χώρου εισόδου δημιουργούνται 2 μοντέλα.

Τα Μοντέλα 1 και 2 χρησιμοποιούν τη μεθοδολογία class independent, δηλαδή το clustering εκτελείται για όλα τα δεδομένα ανεξαρτήτως κλάσης. Το Μοντέλο 1 έχει λιγότερους κανόνες IF-THEN, ενώ το Μοντέλο 2 περισσότερους καθώς έχει έχει μικρή ακτίνα cluster (0.1).

Τα Μοντέλα 3 και 4 χρησιμοποιούν τη μεθοδολογία class dependent, όπου το clustering γίνεται ξεχωριστά για κάθε κλάση. Εδώ το Μοντέλο 3 έχει λιγότερους κανόνες και το Μοντέλο 4 περισσότερους.

Η εκπαίδευση όλων των μοντέλων γίνεται με συνδυασμό Backpropagation για τις παραμέτρους των συναρτήσεων συμμετοχής και Least Squares για τη συνάρτηση εξόδου.

Παράμετροι αξιολόγησης

Οι παράμετροι αξιολόγησης που χρησιμοποιούνται στην εργασία για την αποτίμηση της απόδοσης των ασαφών μοντέλων TSK είναι οι εξής:

1. **Error Matrix (Πίνακας Σφαλμάτων)**:
   - Πρόκειται για έναν πίνακα διαστάσεων k×k, όπου k είναι ο αριθμός των κλάσεων ταξινόμησης. Κάθε στοιχείο του πίνακα δείχνει πόσα δείγματα ταξινομήθηκαν σωστά ή λανθασμένα σε μια κλάση.
   - Τα στοιχεία της κύριας διαγώνιου περιέχουν τον αριθμό των σωστά ταξινομημένων δειγμάτων για κάθε κλάση.
   - Τα στοιχεία εκτός διαγώνιου δείχνουν τα δείγματα που ταξινομήθηκαν λάθος (δηλαδή ανήκουν σε μια κλάση αλλά το μοντέλο τα ταξινόμησε σε μια άλλη).
2. **Overall Accuracy (Συνολική Ακρίβεια)**:
   - Εκφράζει το ποσοστό των σωστά ταξινομημένων δειγμάτων σε σχέση με το συνολικό πλήθος δειγμάτων. Είναι μια συνολική εκτίμηση της απόδοσης του ταξινομητή.
   - Υπολογίζεται ως:

$$OA = \frac{1}{N} \sum_{i=1}^{k} x_{ii}$$

Όπου N είναι το συνολικό πλήθος των δειγμάτων και x_{ii} το πλήθος των δειγμάτων που ταξινομήθηκαν σωστά στην κλάση C_i.

Η αντίστοιχη συναρτηση βρίσκεται στο αρχείο OvAcc.m .

3. **Producer's Accuracy (Ακρίβεια Παραγωγού)**:
   - Αξιολογεί την ακρίβεια της ταξινόμησης από την οπτική της κλάσης παραγωγής, δηλαδή πόσο σωστά το μοντέλο ταξινόμησε δείγματα που ανήκουν σε μια κλάση.
   - Ορίζεται ως:

$$PA(j) = \frac{x_{jj}}{x_{jc}}$$

Οπου, x_{jj} είναι το πλήθος των δειγμάτων που ανήκουν στην κλάση C_j και ταξινομήθηκαν σωστά, και x_{jc} είναι το πλήθος όλων των δειγμάτων που ανήκουν στην κλάση C_j .

4. **User's Accuracy (Ακρίβεια Χρήστη)**:

   - Αξιολογεί την ακρίβεια της ταξινόμησης από την οπτική του χρήστη του ταξινομητή, δηλαδή πόσα από τα δείγματα που ταξινομήθηκαν σε μια συγκεκριμένη κλάση ανήκουν πραγματικά σε αυτή.

   - Ορίζεται ως:

$$UA(i) = \frac{x_{ii}}{x_{ir}}$$

   Όπου x_{ii} είναι το πλήθος των σωστά ταξινομημένων δειγμάτων στην κλάση C_i και x_{ir} είναι το πλήθος των δειγμάτων που ταξινομήθηκαν στην κλάση C_i, ανεξαρτήτως αν ανήκουν πραγματικά σε αυτή ή όχι.

   Η αντίστοιχη συναρτηση βρίσκεται στο αρχείο UsAcc.m .

5. **$\hat{K}$ (Kappa Coefficient)**:

   - Πρόκειται για έναν στατιστικό δείκτη που μετρά τη συμφωνία μεταξύ του μοντέλου και των πραγματικών ταξινομήσεων, λαμβάνοντας υπόψη και την τυχαία ταξινόμηση. Δηλώνει πόσο καλά λειτουργεί το μοντέλο σε σχέση με το αν ταξινομούσε τυχαία τα δείγματα.

   - Υπολογίζεται ως:

$$\kappa = \frac{N \sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{ir} x_{ic}}{N^2 - \sum_{i=1}^{k} x_{ir} x_{ic}}$$

   Όπου x_{ir} είναι το πλήθος των δειγμάτων που ταξινομήθηκαν στην κλάση C_i , και x_{ic} είναι το πλήθος των δειγμάτων που πραγματικά ανήκουν στην κλάση C_i .

   Η αντίστοιχη συναρτηση βρίσκεται στο αρχείο KCoef.m .

Αποτελέσματα

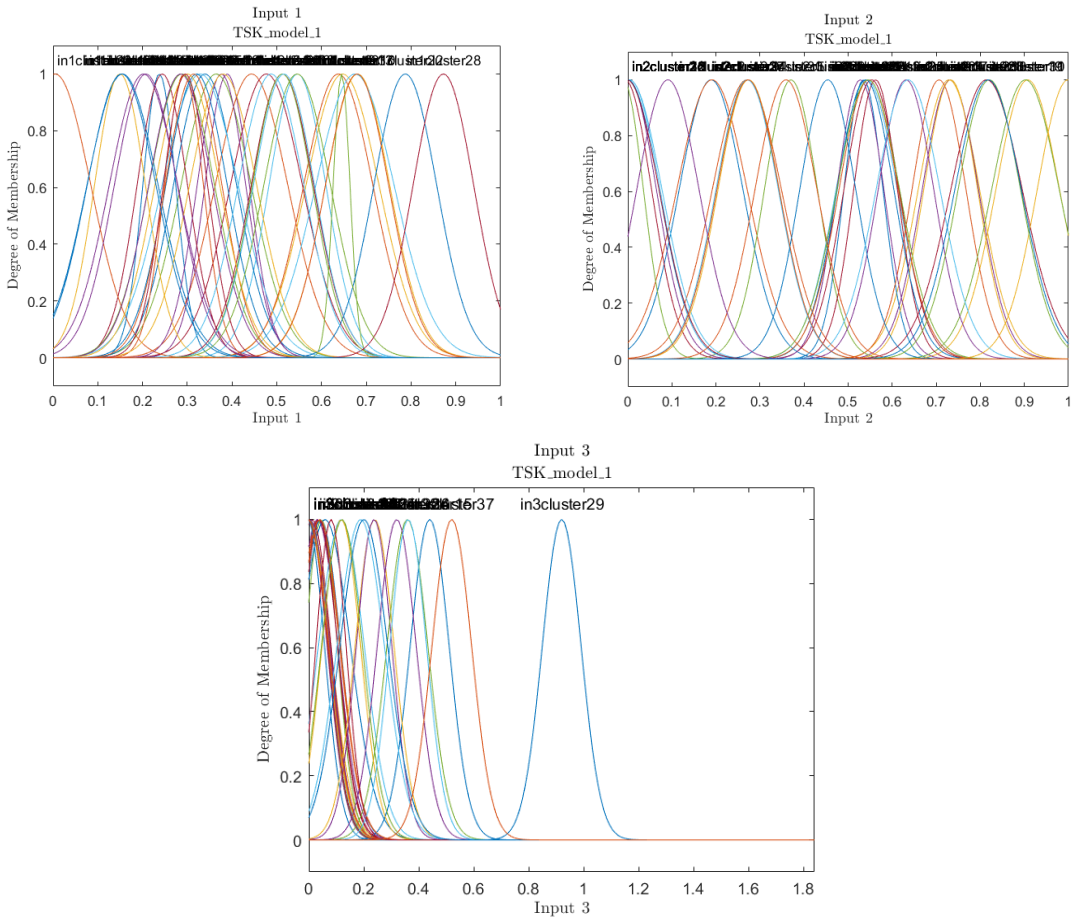
Αρχικά, χρησιμοποιείται η συναρτηση calculate_frequencies() για να υπολογιστούν οι συχνότητες εμφάνισης κάθε κλάσης σε κάθε υποσύνολο. Τα αποτελέσματα συνοψίζονται παρακάτω :

```
Percentage of training data in class 1: 73.37%
Percentage of training data in class 2: 26.63%
Percentage of check data in class 1: 73.77%
Percentage of check data in class 2: 26.23%
Percentage of test data in class 1: 73.77%
Percentage of test data in class 2: 26.23%
```

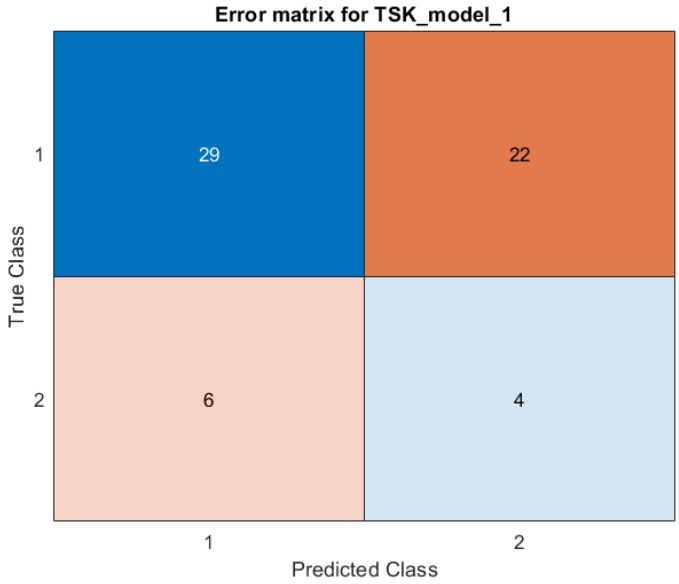
Παρατηρείται ότι καλό θα ήταν τα subsets που θα προκύψουν από το splitting έχουν παρόμοιες συχνότητες εμφάνισης για κάθε κλάση του dataset (δηλ. να πάνε τόσα σημεία από κάθε κλάση όσος και ο λόγος των μεγεθών). Στην συγκεκριμενη περιπτωση φαίνεται σχετικά δίκαιο το splitting.

Ακολουθουν οι γραφικές παραστάσεις των membership functions και οι υπολογισμοί των παραμέτρων αξιολόγησης, για κάθε μοντέλο ξεχωριστά.

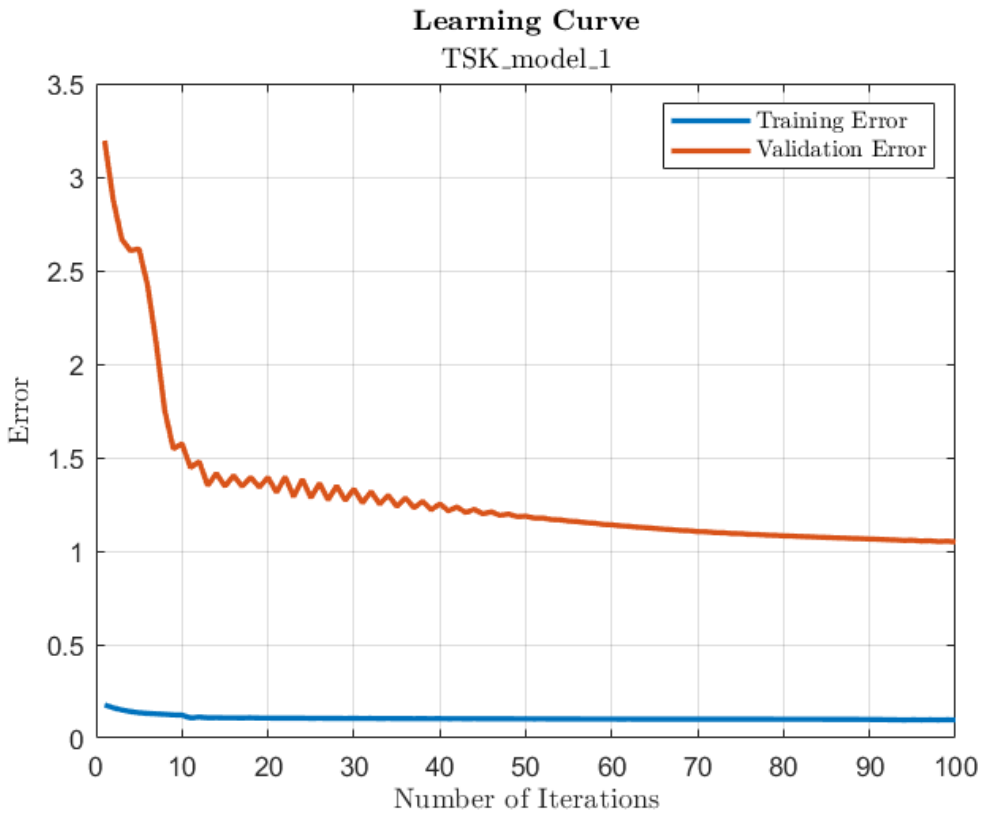
TSK Model 1 : Independent Model with Small Radius



Εικόνα 1.1 : Τελικές μορφές ασαφών συνόλων του 1ου TSK μοντέλου

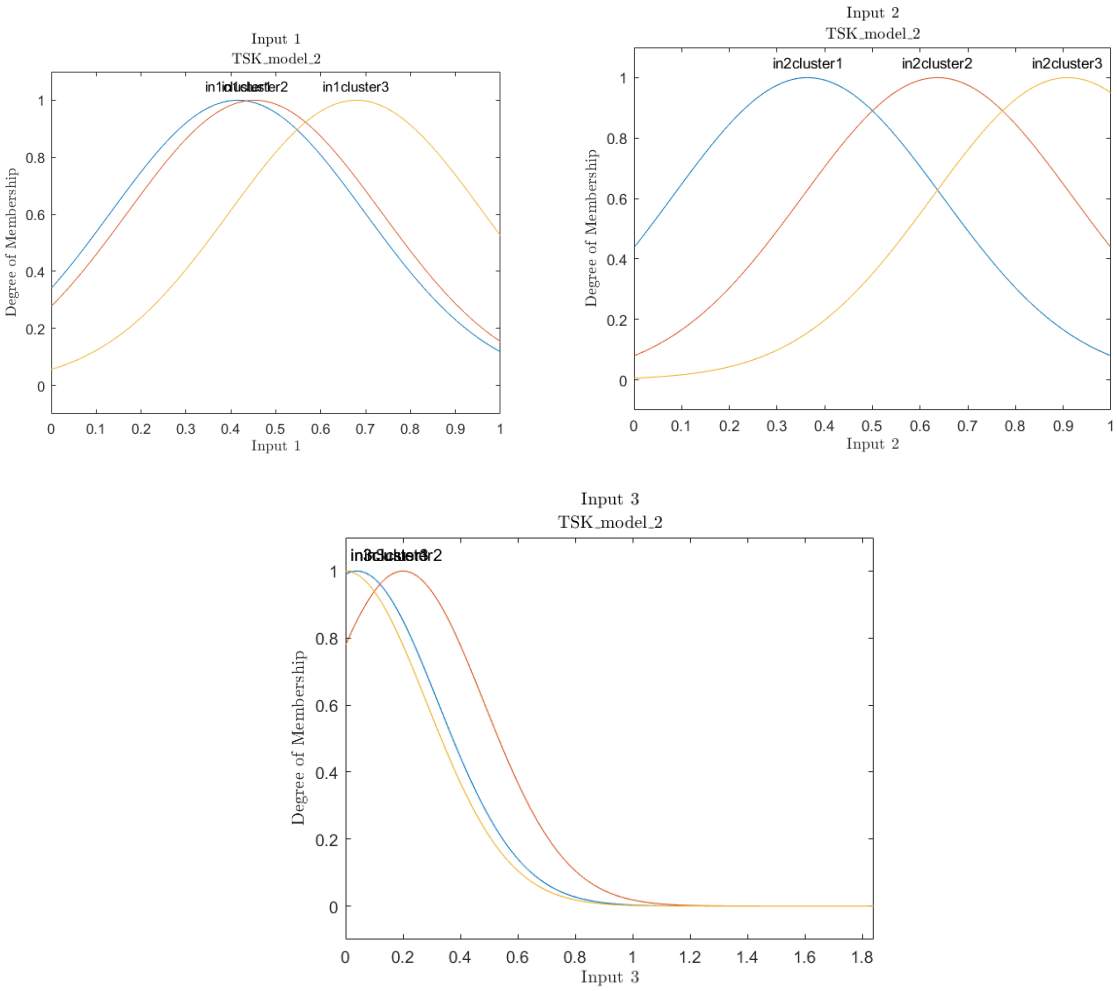


Εικόνα 1.2 : Error Matrix του 1ου TSK μοντέλου

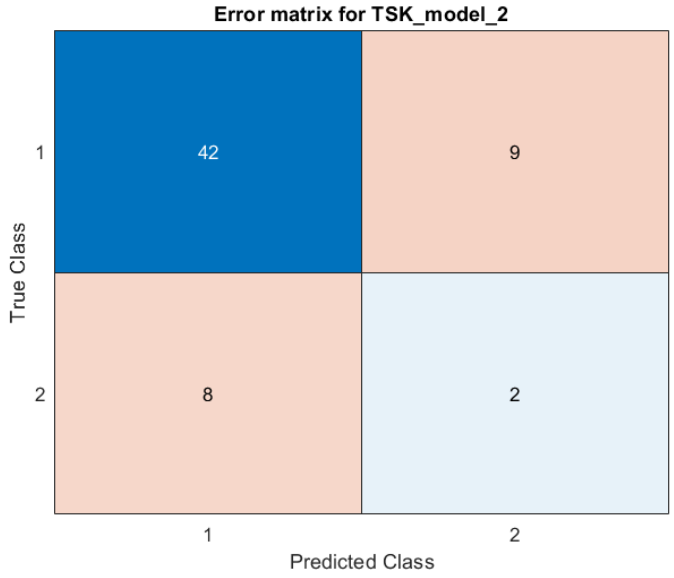


Εικόνα 1.3 : Learning Curve του 1ου TSK μοντέλου

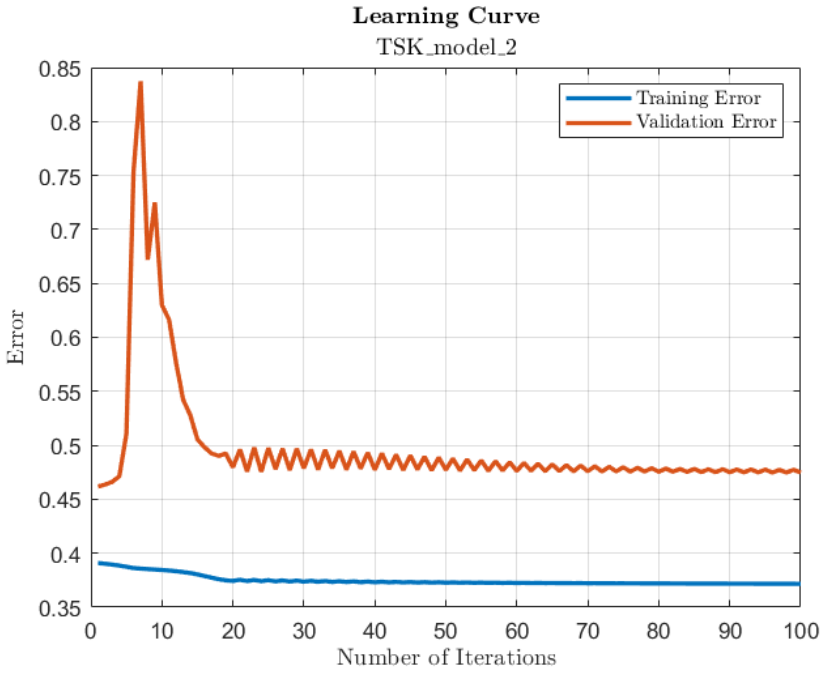
TSK Model 2 : Independent Model with Large Radius



Εικόνα 1.4 : Τελικές μορφές ασαφών συνόλων του 2ου TSK μοντέλου

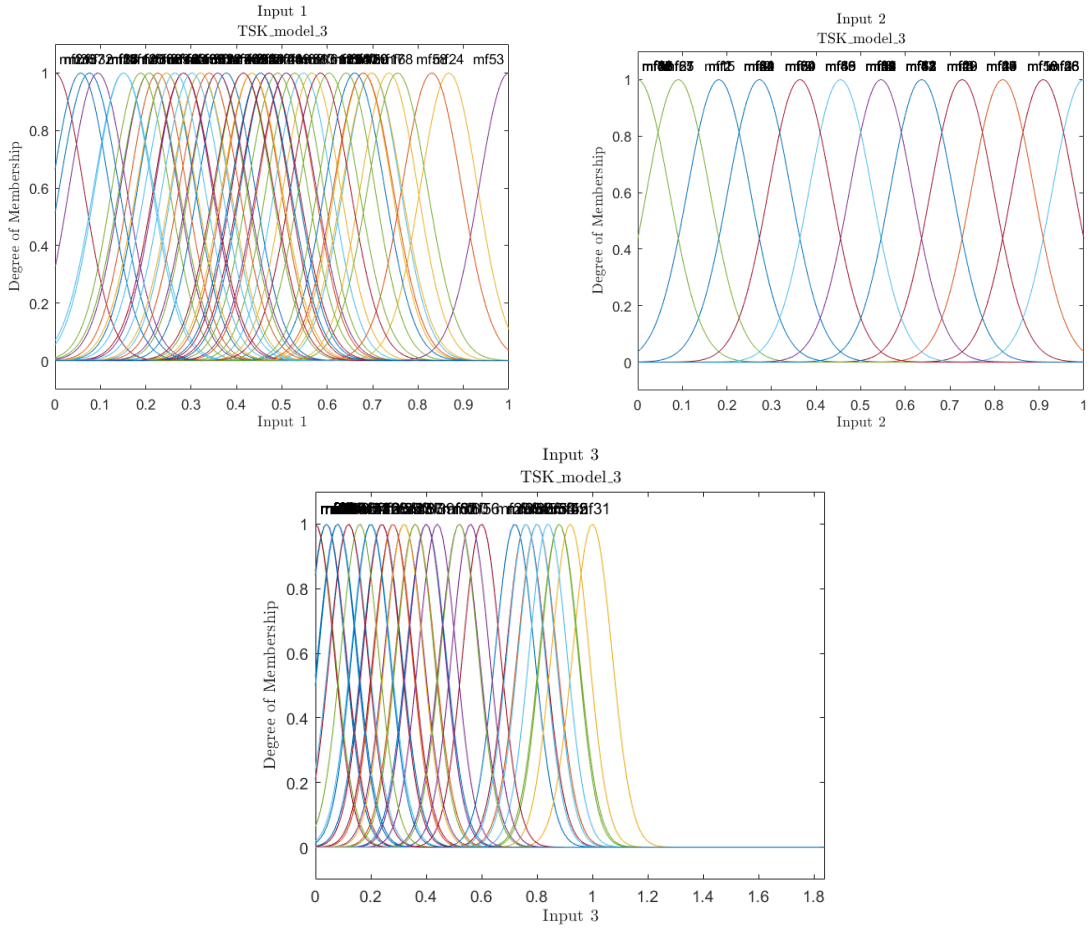


Εικόνα 1.5 : Error Matrix του 2ου TSK μοντέλου



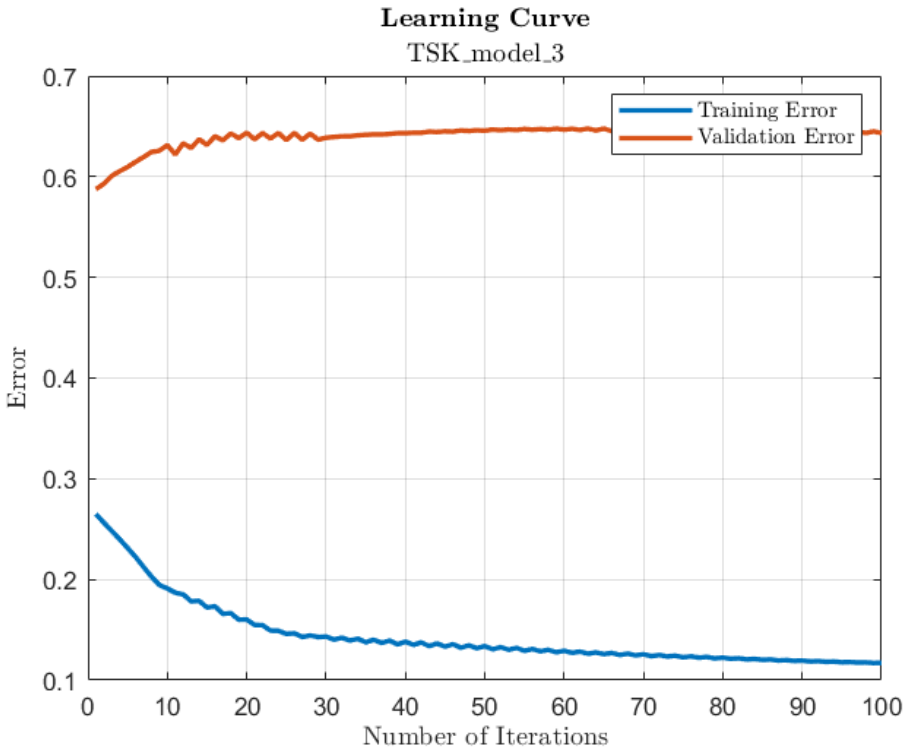
Εικόνα 1.6 : Learning Curve του 2ου TSK μοντέλου

TSK Model 3 : Dependent Model with Small Radius



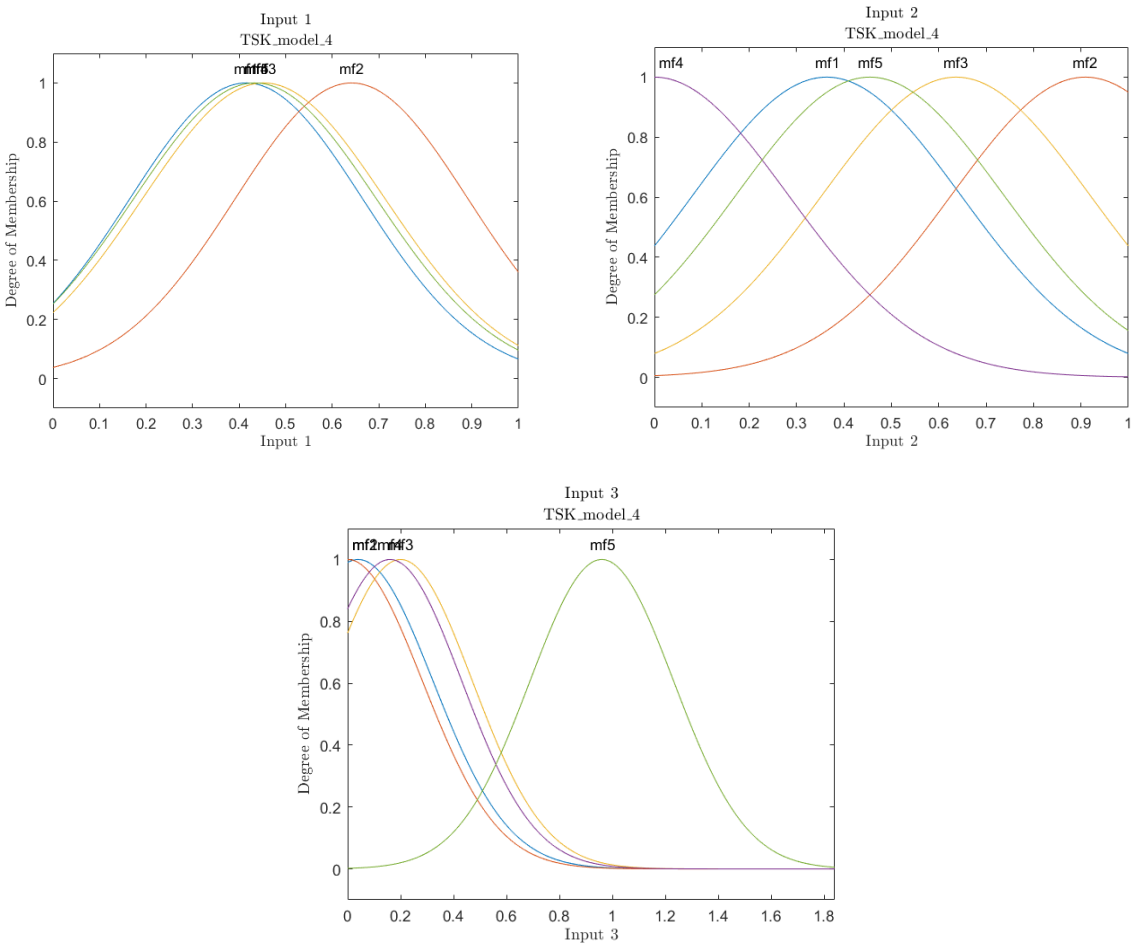
Εικόνα 1.7 : Τελικές μορφές ασαφών συνόλων του 3ου TSK μοντέλου

Εικόνα 1.8 : Error Matrix του 3ου TSK μοντέλου



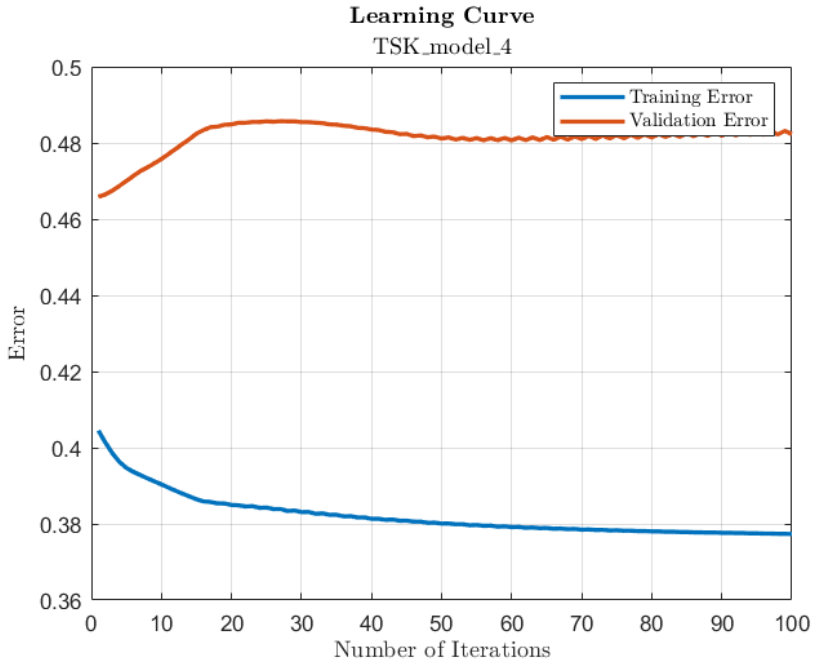
Εικόνα 1.9 : Learning Curve του 3ου TSK μοντέλου

TSK Model 4 : Dependent Model with Large Radius



Εικόνα 1.10 : Τελικές μορφές ασαφών συνόλων του 4ου TSK μοντέλου

Εικόνα 1.11 : Error Matrix του 4ου TSK μοντέλου



Εικόνα 1.12 : Learning Curve του 4ου TSK μοντέλου

Τέλος, οι μετρικές των τεσσάρων μοντέλων εμφανίζονται σε μορφή πίνακα, συμπεριλαμβανομένων των OA, PA, UA, K_hat, καθώς και ο αριθμός των κανόνων (rules) που παράγει κάθε μοντέλο.

Models/Metrics	**OA**	**PA**	**UA**	**K_hat**
Model 1	0.5410	0.8286/0.1538	0.5686/0.4	0.191
Model 2	0.7213	0.84/0.1818	0.8235/0.2	0.226
Model 3	0.6885	0.82/0.0909	0.8039/0.1	0.924
Model 4	0.7705	0.8627/0.3	0.8627/0.3	1.627

Η πιο σημαντική παράμετρος για την αξιολόγηση των μοντέλων στην επίλυση του προβλήματος ταξινόμησης είναι το Overall Accuracy.

2. Εφαρμογή σε high-dimensional dataset

Το δεύτερο μέρος της εργασίας εστιάζει στην εφαρμογή των ασαφών μοντέλων TSK σε προβλήματα ταξινόμησης με υψηλότερη διαστασιμότητα. Σε αυτή τη φάση, επιλέγεται το σύνολο δεδομένων **Epileptic Seizure Recognition** από το UCI repository, το οποίο περιλαμβάνει 11.500 δείγματα και 179 χαρακτηριστικά. Αν το πρόβλημα λυνόταν με την κλασσική μέθοδο του grid searching, θα δημιουργούνταν πάρα πολλοί κανόνες (2^{179}), άρα θα υπήρχε ένα δύσκολο μοντέλο για επίλυση. Λόγω του μεγάλου αριθμού μεταβλητών, προκύπτει η ανάγκη για εφαρμογή τεχνικών μείωσης διαστασιμότητας, καθώς και για περιορισμό του αριθμού των IF-THEN κανόνων μέσω Subtractive Clustering.

Η επιλογή της ακτίνας των cluster αλλά και του πλήθους των χαρακτηριστικών γίνεται με την μέθοδο του grid searching. Το TSK μοντέλο που θα έχει το μικρότερο μέσο σφάλμα θα θεωρηθεί το βέλτιστο και θα εκπαιδευτεί ώστε να γίνει η αξιολόγηση με βάση τις παραμέτρους που χρησιμοποιήθηκαν και στο Α μέρος.

Dataset

Το **Epileptic Seizure Recognition dataset** είναι ένα σύνολο δεδομένων που χρησιμοποιείται για την ανάλυση και αναγνώριση επιληπτικών κρίσεων με βάση ηλεκτροεγκεφαλικά (EEG) σήματα. Το συγκεκριμένο σύνολο δεδομένων περιλαμβάνεται στο UCI Machine Learning Repository και είναι ευρέως χρησιμοποιούμενο για προβλήματα ταξινόμησης, ιδίως στην ιατρική έρευνα και ανάλυση.

Διαχωρισμός Δεδομένων

Όπως και σε όλα τα training που χρησιμοποιήθηκαν TSK μοντέλα, το dataset χωρίζεται σε 3 μη επικαλυπτόμενα subsets. Το training (60%) που χρησιμοποιείται για την εκπαίδευση, το validation (20%) που χρησιμοποιείται κατά την εκπαίδευση για την αποφυγή του φαινομένου της υπερεκπαίδευσης και το test (20%) που χρησιμοποιείται για την αξιολόγηση του training. Για τον διαχωρισμό και σε αυτή την περίπτωση χρησιμοποιείται η συνάρτηση split_scale.

Μείωση Διαστασιμότητας Dataset/ Επιλογή Χαρακτηριστικών (Feature Selection)

Επειδή το δοσμένο dataset έχει high-dimensionality, ο αριθμός των απαιτούμενων κανόνων με βάση το grid partitioning στο χώρο των εισόδων –features του μοντέλου θα είναι τεράστιος. Για το λόγο αυτό, όπως αναφέρεται και στην εκφώνηση του $2^{ου}$ μέρους της παρούσας εργασίας, θα χρησιμοποιήσουμε τεχνικές μείωσης της διαστασιμότητας και τεχνικές ομαδοποίησης για το διαχωρισμό του χώρου των εισόδων με σκοπό τη περαιτέρω μείωση του απαιτούμενου αριθμού ασαφών κανόνων (καθώς αυτοί θα λάβουν σαν είσοδο όχι τα features αλλά τα ομαδοποιημένα features). Για την επιλογή των πιο σημαντικών χαρακτηριστικών από το dataset χρησιμοποιείται ο αλγόριθμος ReliefF. Ο αλγόριθμος αυτός είναι μια μέθοδος για την

επιλογή χαρακτηριστικών που βασίζεται στη μέτρηση της συνεισφοράς κάθε χαρακτηριστικού στην απόσταση μεταξύ των δειγμάτων διαφορετικών κλάσεων. Κατά την διαδικασία του ReliefF:

1. Ο αλγόριθμος συγκρίνει κάθε δείγμα του dataset με τους κοντινότερους γείτονές του (nearest neighbors), τόσο εντός της ίδιας κλάσης όσο και με δείγματα από διαφορετικές κλάσεις.
2. Αν ένα χαρακτηριστικό συμβάλλει στη διαφοροποίηση μεταξύ των δειγμάτων διαφορετικών κλάσεων, η συνεισφορά του αυξάνεται (δηλαδή, το βάρος του χαρακτηριστικού αυξάνεται).
3. Αν ένα χαρακτηριστικό δεν συμβάλλει στη διαφοροποίηση, το βάρος του παραμένει μικρό.
4. Στο τέλος, τα χαρακτηριστικά κατατάσσονται με βάση το βάρος τους, και τα χαρακτηριστικά με τα υψηλότερα βάρη θεωρούνται τα πιο σημαντικά για την ταξινόμηση.

Στη συνέχεια, αφού επιλεγούν τα σημαντικότερα χαρακτηριστικά με βάση την κατάταξη του αλγορίθμου ReliefF, ο αλγόριθμος εκτελεί διαδικασία clustering (ομαδοποίηση), χρησιμοποιώντας την τεχνική Subtractive Clustering. Πρόκειται για έναν αλγόριθμο clustering που χρησιμοποιείται για τον εντοπισμό των κέντρων των clusters σε ένα dataset. Τα κέντρα αυτά θα αποτελέσουν τους πυρήνες των ασαφών κανόνων IF-THEN. Η ακτίνα (radius) των clusters καθορίζει το πόσο ευρεία θα είναι η περιοχή επιρροής κάθε cluster. Μικρότερη ακτίνα οδηγεί σε περισσότερα clusters, άρα και σε περισσότερους κανόνες IF-THEN, ενώ μεγαλύτερη ακτίνα μειώνει τον αριθμό των clusters και των κανόνων.

Grid Searching και Βελτιστοποίηση Παραμέτρων

Το **grid searching** είναι μια μέθοδος που χρησιμοποιείται για την εύρεση των βέλτιστων τιμών παραμέτρων στο σύστημα Takagi-Sugeno-Kang (TSK), συγκεκριμένα για την ακτίνα των clusters (**RValues**) και τον αριθμό των χαρακτηριστικών (**numFeatures**). Ο στόχος είναι να απλοποιηθεί το σύστημα **FIS** ώστε να περιγράφεται επαρκώς με όσο το δυνατόν μικρότερο αριθμό κανόνων, χωρίς να μειώνεται η απόδοσή του.

Επιλογή Παραμέτρων

Για τον προσδιορισμό των βέλτιστων τιμών των παραμέτρων, επιλέγονται τα ακόλουθα σύνολα τιμών με βάση την εμπειρία και δοκιμές:

- **Ακτίνα των clusters (RValues):** [0.2, 0.4, 0.6, 0.8, 1]
- **Αριθμός χαρακτηριστικών (numFeatures):** [5, 8, 10, 12, 15]

Η διαδικασία grid search δοκιμάζει όλους τους πιθανούς συνδυασμούς αυτών των τιμών. Για κάθε ζεύγος παραμέτρων, εκπαιδεύεται ένα νέο μοντέλο ασαφούς συμπερασμού (**FIS**) και υπολογίζεται το σφάλμα εκπαίδευσης (training error).

Εκπαίδευση μέσω Cross-Validation

Για να αυξηθεί η ακρίβεια κατά την επιλογή του βέλτιστου TSK μοντέλου, χρησιμοποιείται η μέθοδος της **5-fold cross-validation**. Αυτή η μέθοδος βελτιώνει την αξιοπιστία της διαδικασίας εκπαίδευσης με τον εξής τρόπο:

- Το dataset χωρίζεται σε πέντε υποσύνολα (folds). Σε κάθε επανάληψη, τέσσερα από τα πέντε υποσύνολα χρησιμοποιούνται για εκπαίδευση και το πέμπτο για επικύρωση.
- Η διαδικασία επαναλαμβάνεται πέντε φορές, εναλλάσσοντας τα υποσύνολα, και στο τέλος υπολογίζεται ο μέσος όρος του σφάλματος (mean error) για κάθε συνδυασμό τιμών παραμέτρων.

Η επαναληπτική διαδικασία επιτρέπει την εκπαίδευση με διαφορετικό διαμοιρασμό δεδομένων κάθε φορά (χρησιμοποιώντας τη συνάρτηση cvpartition στο MATLAB), εξασφαλίζοντας μεγαλύτερη ακρίβεια και σταθερότητα στα αποτελέσματα, καθώς μειώνεται η πιθανότητα υπερεκπαίδευσης σε συγκεκριμένα δεδομένα.

Επιλογή Χαρακτηριστικών με τον Αλγόριθμο ReliefF

Για κάθε συνδυασμό παραμέτρων, ο αλγόριθμος ReliefF χρησιμοποιείται για την επιλογή των πιο σημαντικών χαρακτηριστικών από το dataset, το οποίο περιέχει 176 μεταβλητές. Ο αλγόριθμος αυτός κατατάσσει τα χαρακτηριστικά με βάση τη συμβολή τους στη διάκριση μεταξύ των κλάσεων.

Τα σημαντικότερα χαρακτηριστικά καταγράφονται στον πίνακα Index, ο οποίος περιέχει τις θέσεις των χαρακτηριστικών που θα χρησιμοποιηθούν κατά την εκπαίδευση του μοντέλου. Ο αριθμός των χαρακτηριστικών που επιλέγονται εξαρτάται από την τιμή του numFeatures που εξετάζεται κάθε φορά.

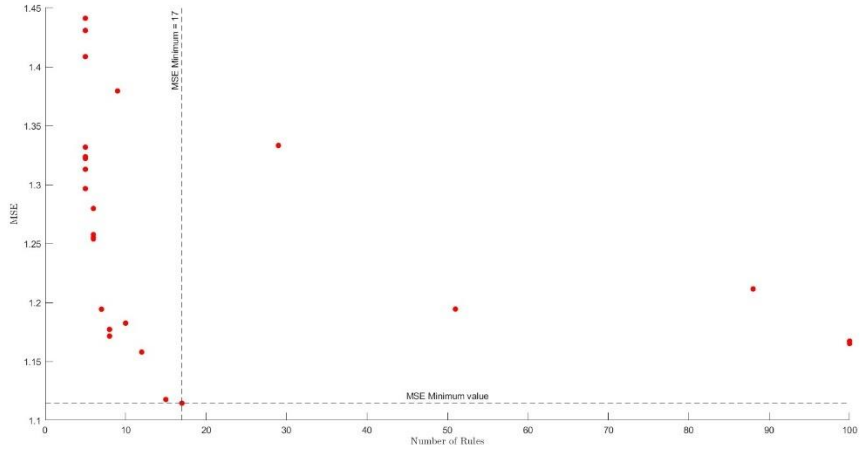
Το σύνολο του παραπάνω κώδικα βρίσκεται στο αρχείο Classification_Part2.m

Αποτελέσματα Grid Searching

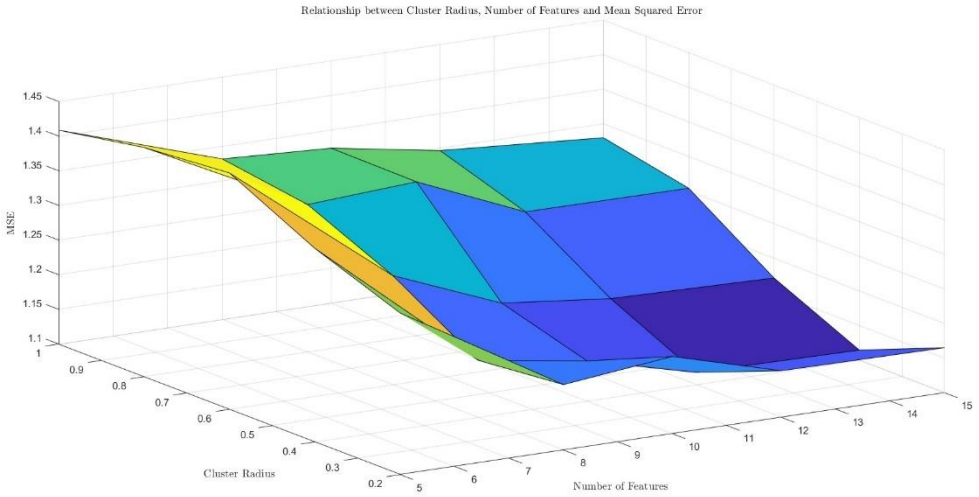
Από το grid search προέκυψε ότι :

```
Optimal TSK Model:
--------------------------
Minimum Error (MSE): 1.1145
Optimal Cluster Radius: 0.40
Optimal Number of Features: 15
```

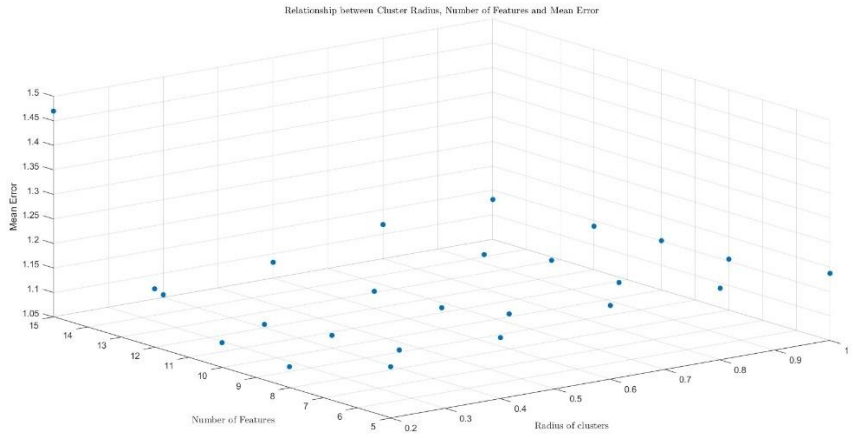
Στην συνέχεια, παρατίθενται κάποια ενδεικτικά γραφήματα μέσου σφάλματος σε συνδυασμό με τον αριθμό των χαρακτηριστικών και της ακτίνας των clusters διαγράμματα. Απεικονίζουν το MSE, RMSE συναρτήσει των numFeatures και της ακτίνας clusterRadius με βάση τα αποτελέσματα που προέκυψαν :



Εικόνα 2.1:Διάγραμμα MSE συναρτήσει του αριθμού των κανόνων



Εικόνα 2.2:Διάγραμμα μέσου τετραγωνικού σφάλματος συναρτήσει του αριθμού χαρακτηριστικών και της ακτίνας των clusters



Εικόνα 2.3:Διάγραμμα μέσου σφάλματος συναρτήσει του αριθμού χαρακτηριστικών και της ακτίνας των clusters

Από τα αποτελέσματα του grid search, φαίνεται ότι η μείωση του αριθμού των χαρακτηριστικών (features) οδηγεί σε αυξημένα validation errors, κάτι αναμενόμενο λόγω της υψηλής διαστασιμότητας του dataset. Συνεπώς, είναι λογικό ο βέλτιστος συνδυασμός να περιλαμβάνει έναν από τους σχετικά μεγάλους αριθμούς διαθέσιμων χαρακτηριστικών, δηλαδή 15, όπως και παρατηρείται.

Αναφορικά με την ακτίνα cluster, παρατηρούμε ότι για έναν δεδομένο αριθμό χαρακτηριστικών, μικρότερες τιμές ακτίνας οδηγούν σε χαμηλότερα validation errors. Στην περίπτωση των 15 χαρακτηριστικών, η βέλτιστη ακτίνα είναι 0.4, ενώ κοντά στο ελάχιστο σφάλμα βρίσκονται και άλλες χαμηλές τιμές ακτίνας, όπως η 0.2. Αυτή η τάση υποδεικνύει ότι μικρότερη ακτίνα βελτιώνει την ακρίβεια του μοντέλου.

Επίσης, παρατίθενται και τα αποτελέσματα των μετρικών της εκφώνησης για κάθε συνδυασμό του grid search:

OA

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.3567	0.3332	0.3133	0.3256	0.3289
8	0.3967	0.3836	0.3707	0.3403	0.3413
10	0.3943	0.3890	0.3882	0.3366	0.3372
12	0.3062	0.4046	0.3932	0.3671	0.3533
15	0.3377	0.4046	0.3818	0.3596	0.354

UA

Class 1

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.6802	0.659	0.6679	0.6679	0.6693
8	0.775	0.7476	0.7538	0.7524	0.759
10	0.7505	0.7613	0.7703	0.7618	0.767
12	0.4132	0.7741	0.7594	0.775	0.7731
15	0.8269	0.7783	0.7613	0.7486	0.7731

Class 2

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.0516	0.0516	0.0472	0.048	0.0427
8	0.0797	0.0545	0.0606	0.0561	0.0508
10	0.113	0.0622	0.0516	0.0589	0.0533
12	0.1431	0.065	0.0541	0.048	0.0398
15	0.348	0.0919	0.0528	0.0618	0.0524

Class 3

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.6705	0.6308	0.643	0.6802	0.7342
8	0.6076	0.646	0.5726	0.4675	0.4789
10	0.5911	0.6409	0.611	0.4354	0.4321
12	0.5595	0.6726	0.6278	0.5473	0.4979
15	0.3831	0.6291	0.6473	0.5388	0.4911

Class 4

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.4136	0.367	0.2475	0.271	0.2348
8	0.5208	0.5032	0.5226	0.4851	0.4783
10	0.5403	0.491	0.5561	0.4882	0.4964
12	0.4005	0.4805	0.5624	0.5208	0.519
15	0.0765	0.4846	0.4719	0.5023	0.5131

Class 5

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.0128	0.0009	0	0	0
8	0.0564	0.0214	0.0021	0	0
10	0.0303	0.044	0.012	0	0
12	0.0355	0.0838	0.0205	0.0056	0
15	0.0846	0.0919	0.0299	0.0038	0.0021

PA

Class 1

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.8935	0.9037	0.8951	0.9048	0.8941
8	0.8891	0.9057	0.9107	0.9104	0.9060
10	0.8859	0.9007	0.9008	0.9023	0.8989
12	0.8242	0.8982	0.9025	0.8998	0.9036
15	0.6238	0.8943	0.9145	0.9144	0.8997

Class 2

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.1837	0.1863	0.1843	0.1889	0.1948
8	0.2694	0.2036	0.2382	0.2212	0.2243
10	0.2774	0.2223	0.2177	0.2335	0.2363
12	0.1726	0.2253	0.1932	0.1992	0.1817
15	0.2402	0.2737	0.1984	0.2150	0.2064

Class 3

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.2719	0.2564	0.2328	0.2431	0.2450
8	0.3029	0.3054	0.2838	0.2580	0.2560
10	0.3158	0.3067	0.2977	0.2491	0.2477
12	0.2873	0.3148	0.3042	0.2919	0.2695
15	0.2522	0.3203	0.2967	0.2845	0.2690

Class 4

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.2791	0.2366	0.1989	0.2242	0.2284
8	0.2967	0.2829	0.2681	0.2220	0.2238
10	0.2964	0.2832	0.2967	0.2184	0.2195
12	0.2444	0.3030	0.3098	0.2503	0.2409
15	0.1740	0.2885	0.2783	0.2450	0.2411

Class 5

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.3919	NaN	NaN	NaN	NaN
8	0.4556	0.3632	NaN	NaN	NaN
10	0.3121	0.4361	0.3052	NaN	NaN
12	0.5165	0.4975	0.3894	NaN	NaN
15	0.3597	0.4661	0.4204	0.4446	NaN

K_hat

numFeatures/ clusterRadius	0.2	0.4	0.6	0.8	1
5	0.1963	0.1671	0.1414	0.1567	0.1604
8	0.2474	0.2310	0.2154	0.1780	0.1794
10	0.2444	0.2377	0.2372	0.1737	0.1746
12	0.1316	0.2569	0.2431	0.2114	0.1945
15	0.1702	0.2570	0.2286	0.2016	0.1951

Overall Accuracy (OA):

Παρατηρείται ότι η συνολική ακρίβεια (OA) μειώνεται καθώς η ακτίνα cluster αυξάνεται. Για παράδειγμα, με 5 χαρακτηριστικά, η OA μειώνεται από 0.3567 σε 0.3289 όταν η ακτίνα cluster αυξάνεται από 0.2 σε 1. Οι καλύτερες τιμές OA επιτυγχάνονται για μεγαλο αριθμό features και μικρή ακτίνα cluster (0.2 και 0.4).

Users Accuracy (UA)

Κλάση 1: Η ακρίβεια χρήστη για την Κλάση 1 είναι σταθερά υψηλή, ανεξαρτήτως πλήθους χαρακτηριστικών και ακτίνας cluster. Τα καλύτερα αποτελέσματα παρατηρούνται με 15 χαρακτηριστικά και ακτίνα cluster 0.2, με UA = 0.8269.Η επίδοση της Κλάσης 1 δεν επηρεάζεται πολύ από τις αλλαγές στην ακτίνα cluster, κάτι που δείχνει ότι είναι μια εύκολα διακριτή κλάση.

Κλάση 2: Η Κλάση 2 εμφανίζει χαμηλή UA σε όλες τις παραμέτρους. Με 15 χαρακτηριστικά και ακτίνα cluster 0.2, η UA φτάνει τη μέγιστη τιμή 0.348. Παρατηρείται ότι η απόδοση πέφτει δραματικά με αύξηση της ακτίνας cluster, γεγονός που δείχνει δυσκολία στο να διακριθεί σωστά αυτή η κλάση.

Κλάση 3: Η UA για την Κλάση 3 μειώνεται καθώς η ακτίνα cluster αυξάνεται, ειδικά για 8 και 10 χαρακτηριστικά. Η καλύτερη επίδοση παρατηρείται με 5 χαρακτηριστικά και ακτίνα cluster 1 (UA = 0.7342), αλλά γενικά, όσο αυξάνονται οι παράμετροι, η απόδοση πέφτει.

Κλάση 4: Για την Κλάση 4, η ακρίβεια χρήστη (UA) εμφανίζει σχετική σταθερότητα με μικρές ακτίνες cluster, με την καλύτερη τιμή να εμφανίζεται με 10 χαρακτηριστικά και ακτίνα cluster 0.6 (UA = 0.5561).

Κλάση 5: Η UA για την Κλάση 5 είναι εξαιρετικά χαμηλή, με σχεδόν μηδενικές τιμές σε πολλές παραμέτρους. Η υψηλότερη UA (0.3919) εμφανίζεται με 5 χαρακτηριστικά και ακτίνα cluster 0.2. Η Κλάση 5 είναι σαφώς πιο δύσκολη στην ταξινόμηση.

Producers Accuracy (PA)

Κλάση 1:Η PA για την Κλάση 1 παραμένει εξαιρετικά υψηλή, με τις καλύτερες τιμές να εμφανίζονται για 15 χαρακτηριστικά και ακτίνα cluster 0.2, με PA = 0.9145. Η PA δεν επηρεάζεται έντονα από την αλλαγή της ακτίνας cluster, παρόλο που υπάρχουν μικρές διακυμάνσεις.

Κλάση 2:Η PA για την Κλάση 2 είναι σχετικά χαμηλή. Η υψηλότερη τιμή εμφανίζεται με 8 χαρακτηριστικά και ακτίνα cluster 0.2 (PA = 0.2694). Αυτή η κλάση φαίνεται να αντιμετωπίζει προβλήματα αναγνώρισης, όπως και στην UA.

Κλάση 3: Για την Κλάση 3, η PA εμφανίζει μια μικρή βελτίωση με μικρές ακτίνες cluster. Η καλύτερη τιμή (0.3203) παρατηρείται με 15 χαρακτηριστικά και ακτίνα cluster 0.4.

Κλάση 4: Η PA για την Κλάση 4 εμφανίζει μια σχετικά σταθερή συμπεριφορά, με υψηλότερη τιμή (PA = 0.3098) για 12 χαρακτηριστικά και ακτίνα cluster 0.6.

Κλάση 5: Η PA για την Κλάση 5 είναι και πάλι πολύ χαμηλή, με τιμές που φτάνουν το πολύ έως 0.5165 με 12 χαρακτηριστικά και ακτίνα cluster 0.2. Σημαντική πτώση παρατηρείται με αύξηση της ακτίνας cluster.

Εκπαίδευση βέλτιστου TSK Μοντέλου με βάση το grid searching

Η εκπαίδευση του βέλτιστου TSK μοντέλου πραγματοποιείται με βάση τα δεδομένα που προέκυψαν από την 3-D απεικόνιση. Από την απεικόνιση αυτή, παρατηρείται ότι το ελάχιστο σφάλμα επιτυγχάνεται όταν το μοντέλο χρησιμοποιεί 15 χαρακτηριστικά (features) και ακτίνα cluster ίση με 0.4. Έχοντας καθορίσει αυτές τις βέλτιστες παραμέτρους, η διαδικασία εκπαίδευσης προχωρά με τη χρήση της μεθόδου subtractive clustering, για να διαμορφωθεί το τελικό μοντέλο.

Μετά την εκπαίδευση, το μοντέλο αξιολογείται, ώστε να επιβεβαιωθεί η απόδοσή του και να ελεγχθεί η ακρίβεια των προβλέψεών του με βάση τα επιλεγμένα χαρακτηριστικά και την ακτίνα cluster. Ο κώδικας που υλοποιήθηκε για αυτόν τον σκοπό περιγράφεται παρακάτω με τα αποτελέσματα και βρίσκεται στο αρχείο classification_Part2_OptimalModel.m

Αρχικά, με βάση τα αποτελέσματα από την 3D απεικόνιση που παρουσιάζει το MSE (Mean Squared Error), επιλέγονται οι βέλτιστες τιμές:

- **Features : 15**
- **Ακτίνα cluster : 0.4**

Στη συνέχεια, φορτώνονται τα δεδομένα από μία προηγούμενη διαδικασία αναζήτησης πλέγματος (grid search). Τα δεδομένα εκπαίδευσης, επικύρωσης και δοκιμών διαμορφώνονται, ώστε να περιέχουν τα 10 πιο σημαντικά χαρακτηριστικά, βάσει των επιλεγμένων βέλτιστων τιμών. Τα δεδομένα εκπαίδευσης διαχωρίζονται σε κατηγορίες και εκτελείται η διαδικασία "subtractive clustering" για κάθε κατηγορία ξεχωριστά. Αυτό οδηγεί στον υπολογισμό των κεντρικών σημείων και των τυπικών αποκλίσεων των clusters για κάθε κατηγορία.

Αφού υπολογιστούν τα clusters, δημιουργείται ένα αρχικό FIS (Sugeno Fuzzy Inference System). Σε αυτό το μοντέλο προστίθενται τα απαραίτητα εισαγόμενα δεδομένα (inputs) με συναρτήσεις συμμετοχής Gaussian (gaussmf), οι οποίες διαμορφώνονται σύμφωνα με τα clusters κάθε κατηγορίας. Η έξοδος του μοντέλου επίσης καθορίζεται και περιλαμβάνει συναρτήσεις συμμετοχής με σταθερές τιμές (singleton) που αντιστοιχούν στις κατηγορίες των δεδομένων.

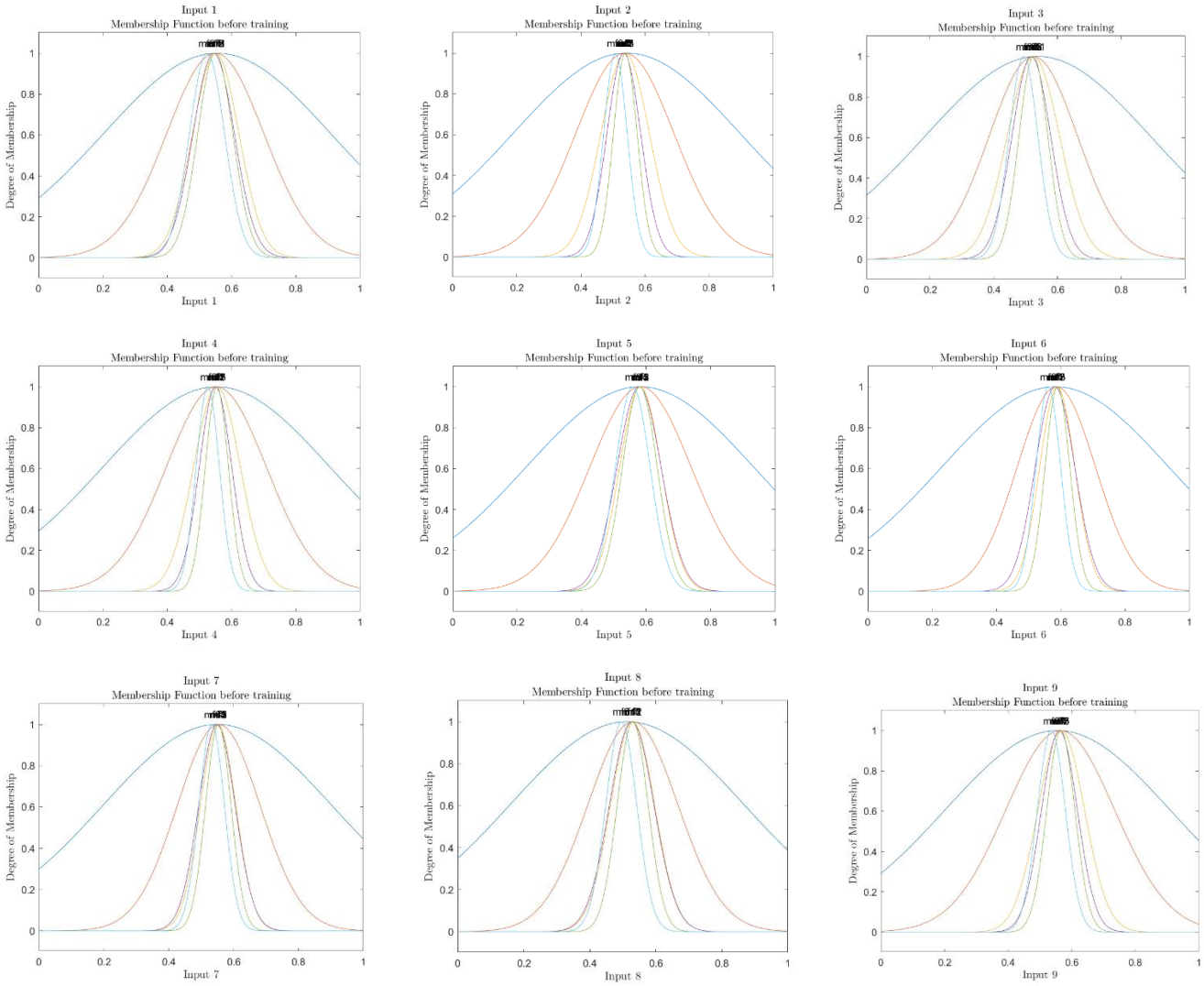
Η βάση κανόνων (rule base) διαμορφώνεται με τη σύνδεση των inputs με την έξοδο, και οι κανόνες παράγονται από τα clusters που έχουν υπολογιστεί για κάθε κατηγορία. Αφού διαμορφωθεί η αρχική μορφή του μοντέλου, πραγματοποιείται απεικόνιση των συναρτήσεων συμμετοχής των εισόδων, πριν ξεκινήσει η διαδικασία εκπαίδευσης. Η εκπαίδευση του μοντέλου πραγματοποιείται με τη χρήση του ANFIS (Adaptive Neuro-Fuzzy Inference System). Ορίζονται οι κατάλληλες παράμετροι εκπαίδευσης και το μοντέλο εκπαιδεύεται χρησιμοποιώντας τα δεδομένα

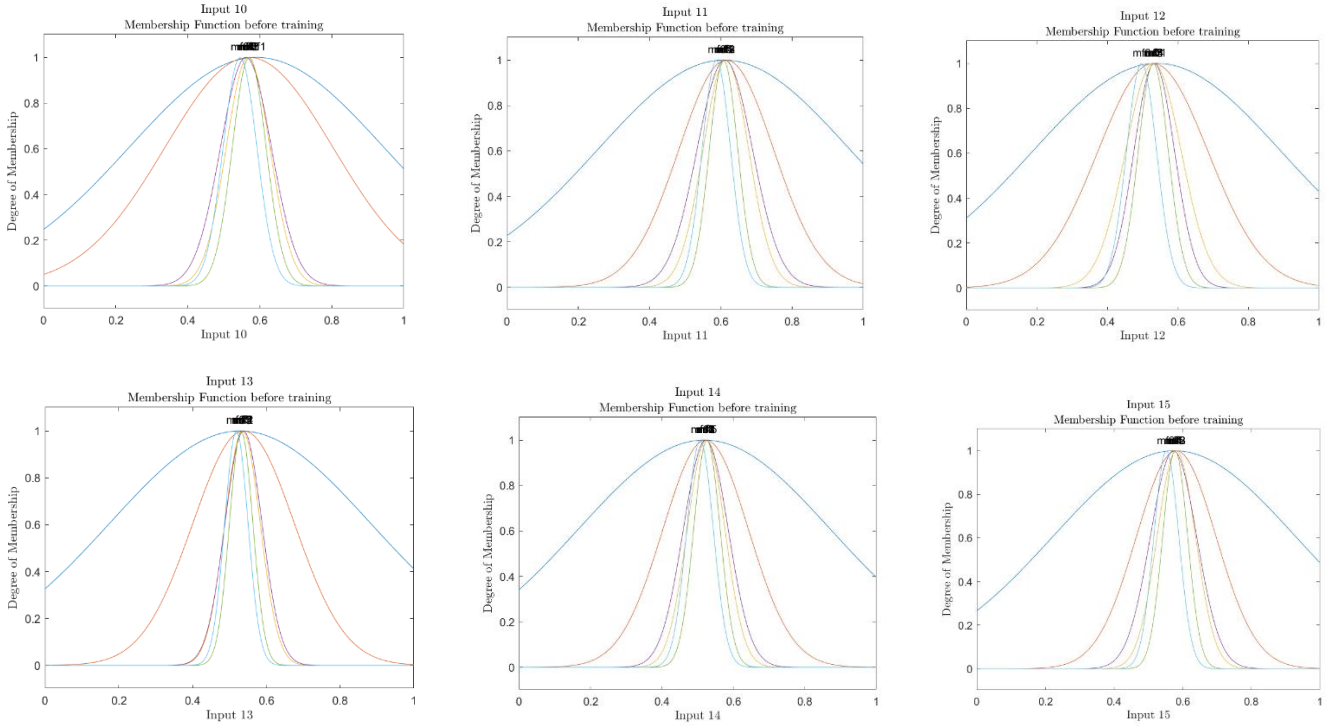
εκπαίδευσης, ενώ αξιολογείται με τα δεδομένα επικύρωσης. Μετά την εκπαίδευση, πραγματοποιείται νέα απεικόνιση των συναρτήσεων συμμετοχής, ώστε να συγκριθούν οι αλλαγές που προέκυψαν από την εκπαίδευση.

Για να αξιολογηθεί το εκπαιδευμένο μοντέλο, χρησιμοποιούνται τα δεδομένα δοκιμών. Το μοντέλο προβλέπει τις κατηγορίες των δεδομένων και αυτές οι προβλέψεις συγκρίνονται με τις πραγματικές τιμές. Δημιουργείται ένας πίνακας σύγχυσης (confusion matrix) για την ανάλυση της ακρίβειας του μοντέλου και για να υπολογιστούν διάφορα στατιστικά όπως η συνολική ακρίβεια (Overall Accuracy - OA), η ακρίβεια ανά κατηγορία (PA - Precision Accuracy), και ο συντελεστής Kappa (K-hat). Στη συνέχεια, γίνεται οπτικοποίηση των πραγματικών και προβλεπόμενων δεδομένων μέσω γραφημάτων, όπως και η καμπύλη μάθησης, η οποία παρουσιάζει την εξέλιξη του σφάλματος κατά την εκπαίδευση και επικύρωση.

Αποτελέσματα βέλτιστου TSK Μοντέλου

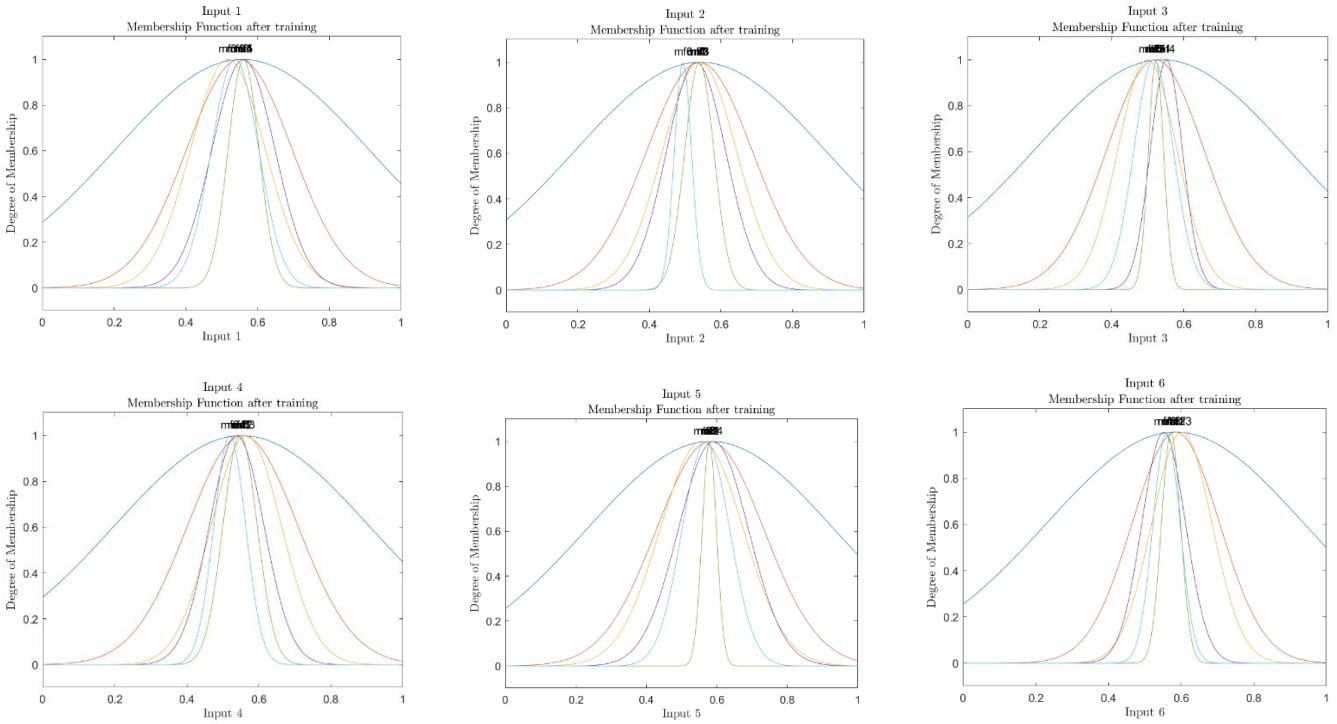
Παρακάτω, δίνονται οι αρχικές συναρτήσεις συμμετοχής (MFs)των λεκτικών τιμών - clusters για τις εισόδους-features του τελικού TSK μοντέλου:

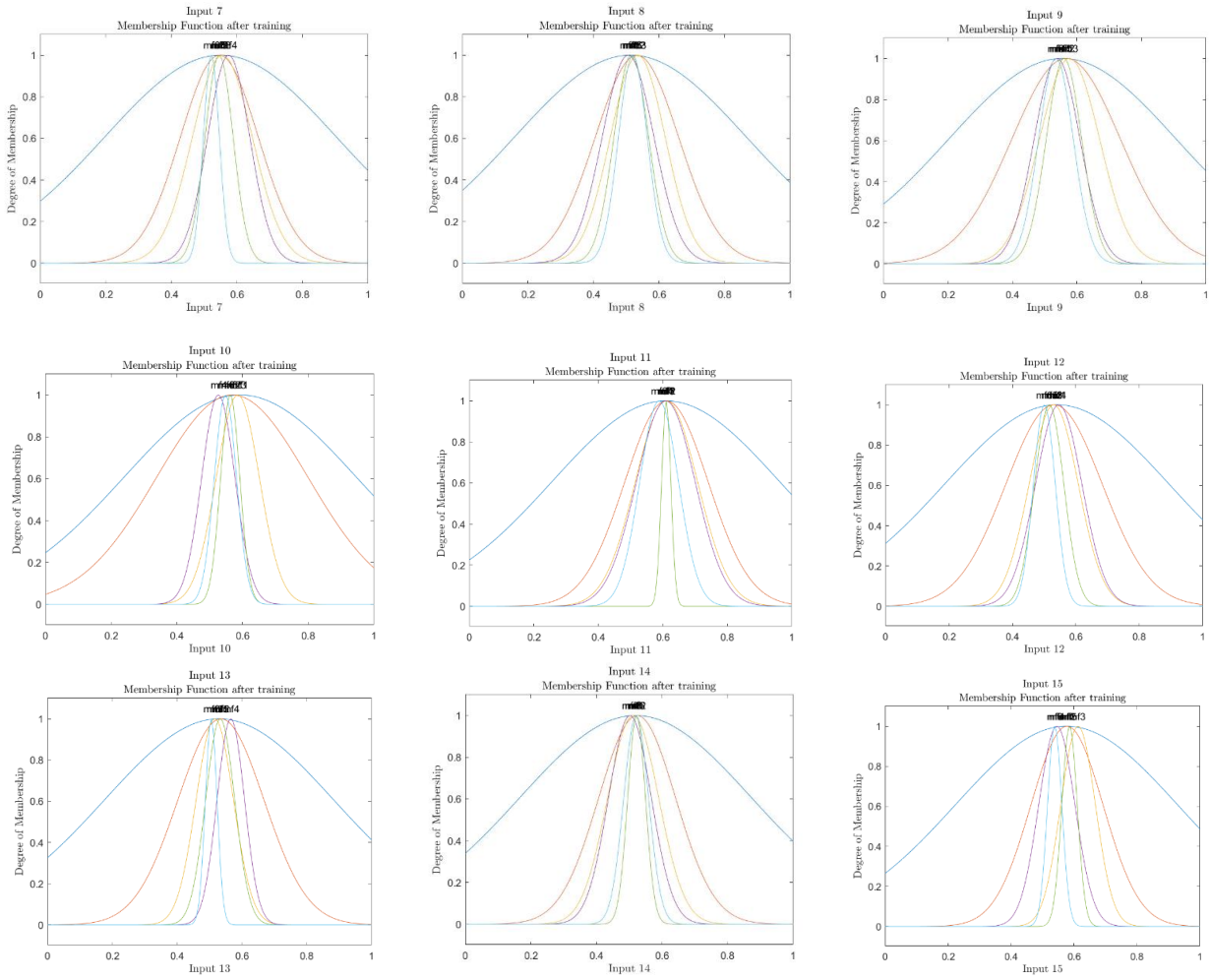




Εικόνα 2.4: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου αρχικού μοντέλου

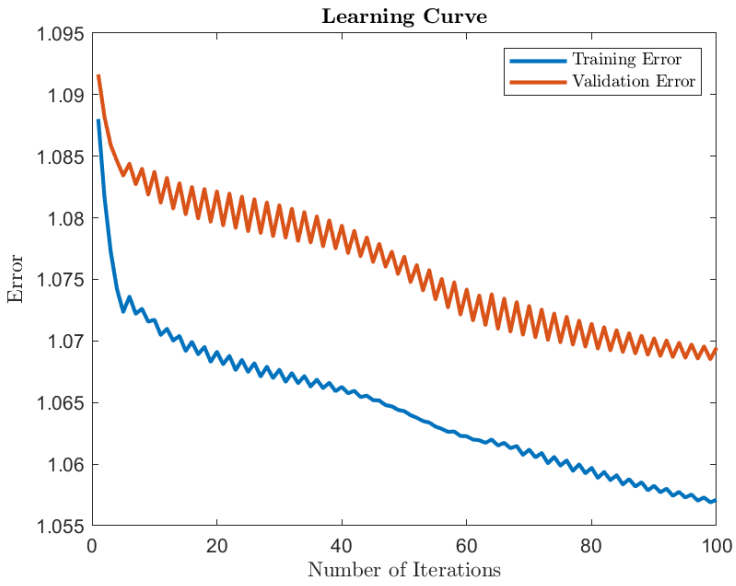
ενώ ακολούθως δίνονται οι τελικές μορφές των παραπάνω MFs:





Εικόνα 2.5: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου τελικού μοντέλου

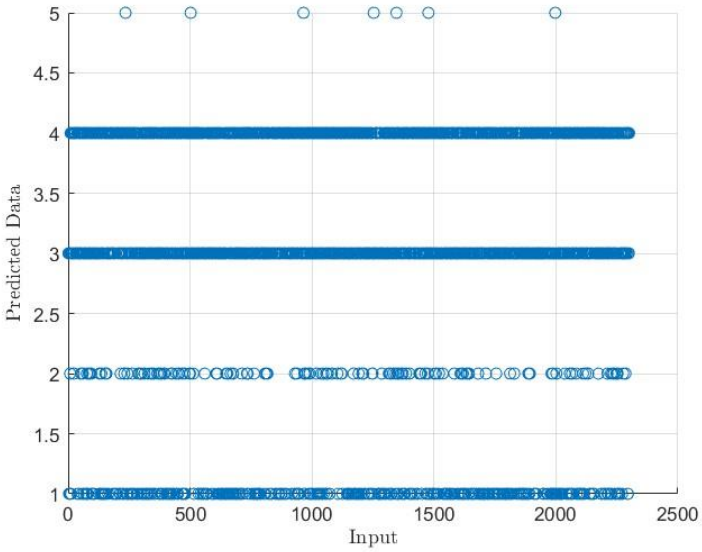
Παρακάτω, δίνονται οι καμπύλες μάθησης (learning curves) του τελικού TSK μοντέλου με τον βέλτιστο συνδυασμό χαρακτηριστικών –κανόνων:



Εικόνα 16: Διάγραμμα καμπυλών εκμάθησης του βέλτιστου TSK μοντέλου

Ακολούθως, σε ότι αφορά την απόδοση του τελικού μοντέλου, δίνεται το διάγραμμα με τα σφάλματα πρόβλεψης του μοντέλου ως προς τις πραγματικές τιμές του test set:

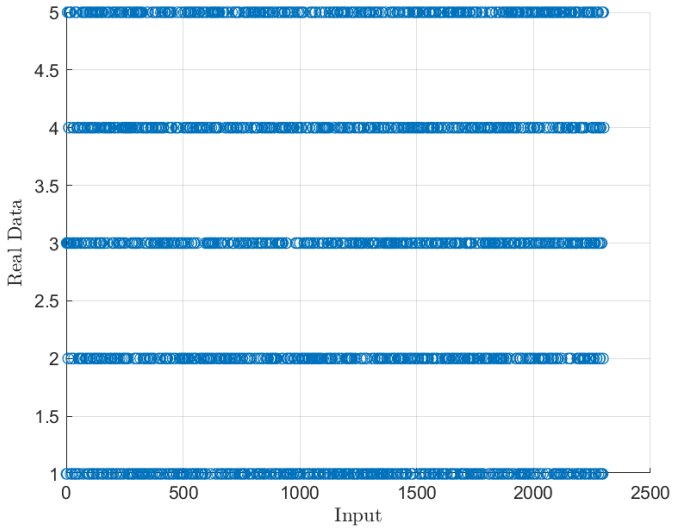
Προβλέψεις τελικού μοντέλου



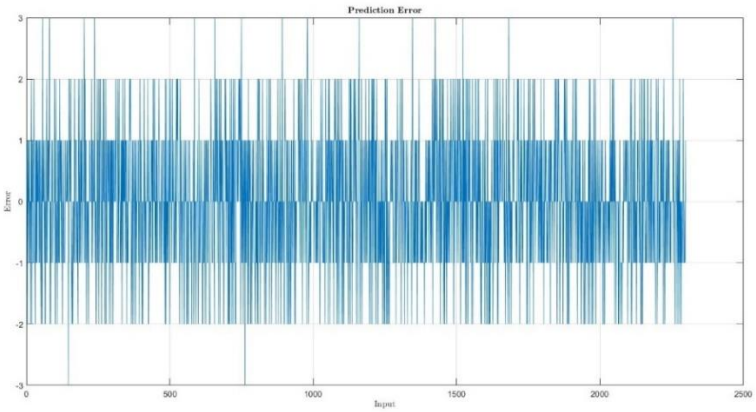
***Εικόνα 2.7 :* Προβλέψεις τελικού μοντέλου**

Το διάγραμμα δείχνει τις προβλέψεις του μοντέλου σε σχέση με τα πραγματικά δεδομένα. Φαίνεται ότι το μοντέλο έχει δυσκολία να ξεχωρίσει σαφώς τις διαφορετικές κλάσεις, ιδιαίτερα την κλάση 5, καθώς οι προβλέψεις της κλάσης 5 βρίσκονται διασκορπισμένες σε όλο το εύρος των τιμών.

Πραγματικές τιμές target του dataset

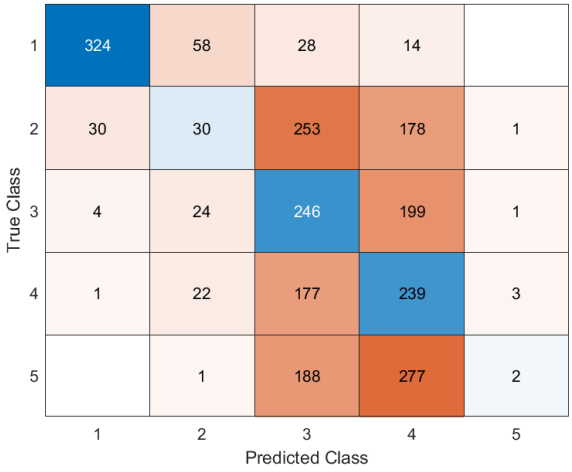


Εικόνα 2.8 : Πραγματικές τιμές target του dataset



Εικόνα 2.9: Σφάλματα πρόβλεψης κατά την εφαρμογή του τελικού TSK μοντέλου στο test set

Τέλος δίνονται ο confusion matrix και οι μετρικές απόδοσης του τελικού μοντέλου στο classification task στο testing set:



Optimal values model

OA	0.36565

Optimal values model

PA(1)	0.90251
PA(2)	0.22222
PA(3)	0.27578
PA(4)	0.26351
PA(5)	0.28571

Optimal values model

UA(1)	0.76415
UA(2)	0.060976
UA(3)	0.51899
UA(4)	0.54072
UA(5)	0.0042735

Optimal values model

K_hat	0.20938

Εικόνα 2.10: Confusion Matrix και μετρικές απόδοσης τελικού TSK μοντέλου

Σχολιασμός Αποτελεσμάτων

Από τον Confusion Matrix παρατηρούνται τα εξής :

- **Κλάση 1**: Το μοντέλο ταξινομεί σωστά 324 δείγματα, αλλά 58 δείγματα ταξινομούνται λανθασμένα ως κλάση 2 και 28 ως κλάση 3. Μικρότερες λανθασμένες ταξινομήσεις υπάρχουν στις κλάσεις 4 και 5.
- **Κλάση 2**: Μόνο 30 δείγματα ταξινομούνται σωστά στην κλάση 2, ενώ πολλά (253) ταξινομούνται ως κλάση 3 και 178 ως κλάση 4.
- **Κλάση 3**: 246 δείγματα ταξινομούνται σωστά, αλλά υπάρχει σύγχυση κυρίως με την κλάση 4, όπου 199 δείγματα καταλήγουν λανθασμένα.
- **Κλάση 4**: Το μοντέλο ταξινομεί σωστά 239 δείγματα, αλλά 177 ταξινομούνται λανθασμένα στην κλάση 3.
- **Κλάση 5**: Μόνο 2 δείγματα ταξινομούνται σωστά, με τα περισσότερα (277) να ταξινομούνται λανθασμένα ως κλάση 4 και 188 ως κλάση 3.

Όσον αφορά τα αποτελέσματα του τελικού βέλτιστου μοντέλου ταξινόμησης :

Overall Accuracy - OA:

Η συνολική ακρίβεια του μοντέλου είναι **36.57%**, κάτι που δείχνει ότι το μοντέλο δυσκολεύεται να ταξινομήσει με ακρίβεια τα δεδομένα σε σχέση με τις πραγματικές τιμές. Παρά το γεγονός ότι η OA είναι χαμηλή, παρατηρούνται διακυμάνσεις στις επιμέρους ακρίβειες ανά κατηγορία.

Producer’s Accuracy - PA:

- Για την **κλάση 1**, η ακρίβεια του παραγωγού είναι **90.25%**, κάτι που δείχνει ότι το μοντέλο ταξινομεί σωστά τα περισσότερα δείγματα αυτής της κλάσης.
- Για τις **κλάσεις 2, 3, 4 και 5**, οι τιμές PA είναι αρκετά χαμηλές, με την **κλάση 2** να έχει **22.22%**, την **κλάση 3 27.58%**, την **κλάση 4 26.35%**, και την **κλάση 5 28.57%**. Αυτά τα ποσοστά υποδεικνύουν ότι το μοντέλο αποτυγχάνει να ταξινομήσει με ακρίβεια τις περισσότερες παρατηρήσεις αυτών των κλάσεων.

User’s Accuracy - UA:

- Η **UA για την κλάση 1** είναι **76.42%**, υποδεικνύοντας ότι από τα δείγματα που έχουν ταξινομηθεί ως κλάση 1, περίπου το 76.42% είναι σωστές ταξινομήσεις.
- Αντίθετα, η **κλάση 2** έχει εξαιρετικά χαμηλή αξιοπιστία χρήστη, με μόλις **6.10%**, κάτι που σημαίνει ότι μόνο ένα μικρό ποσοστό των δειγμάτων που έχουν ταξινομηθεί ως κλάση 2 είναι σωστά.
- Η **UA για την κλάση 3** είναι **51.90%**, ενώ για την **κλάση 4** είναι **54.07%**, δείχνοντας ότι οι ταξινομήσεις για αυτές τις κατηγορίες είναι μέτριας ακρίβειας.
- Τέλος, η **UA για την κλάση 5** είναι εξαιρετικά χαμηλή, μόλις **0.43%**, κάτι που υποδηλώνει ότι σχεδόν όλες οι ταξινομήσεις στην κλάση 5 είναι λανθασμένες.

Συντελεστής Kappa (K-hat): Ο συντελεστής Kappa είναι **0.209**, δείχνοντας μια μέτρια αλλά περιορισμένη συμφωνία του μοντέλου με τις πραγματικές κατηγορίες. Αυτό το αποτέλεσμα υποδεικνύει ότι η απόδοση του μοντέλου είναι μόλις ελαφρώς καλύτερη από τυχαία ταξινόμηση.

Γενικά συμπεράσματα

Αύξηση των χαρακτηριστικών: Παρατηρείται ότι με την αύξηση των χαρακτηριστικών, το μοντέλο δυσκολεύεται να διαχωρίσει σωστά τις κλάσεις, κάτι που αντανακλάται στον confusion matrix. Ειδικότερα, η σύγχυση μεταξύ των κλάσεων 3, 4 και 5 είναι έντονη. Για παράδειγμα, πολλά δείγματα της κλάσης 5 ταξινομούνται λανθασμένα στις κλάσεις 3 και 4. Αυτή η επικάλυψη δείχνει ότι τα χαρακτηριστικά δεν είναι αρκετά διακριτά, γεγονός που οδηγεί σε αυξημένο σφάλμα πρόβλεψης.

Απόδοση του βέλτιστου μοντέλου: Ακόμα και με το βέλτιστο μοντέλο, που επιλέχθηκε μέσω grid search, παρατηρείται ότι το σφάλμα ταξινόμησης παραμένει σημαντικό. Ειδικά για την κλάση 5, το ποσοστό σωστής ταξινόμησης είναι εξαιρετικά χαμηλό, με τη συντριπτική πλειονότητα των δειγμάτων να ταξινομούνται λανθασμένα ως κλάση 3 ή 4. Αυτό το αποτέλεσμα υποδεικνύει ότι το μοντέλο έχει σοβαρές δυσκολίες να διαχειριστεί το αυξημένο εύρος χαρακτηριστικών και την πολυπλοκότητα των δεδομένων.

Επιρροή του αριθμού χαρακτηριστικών: Είναι φανερό ότι η αύξηση των χαρακτηριστικών, αν και μπορεί να αυξάνει την πληροφορία που δέχεται το μοντέλο, αυξάνει και την πολυπλοκότητα, με αποτέλεσμα να γίνεται πιο δύσκολη η σωστή εκμάθηση των κανόνων από το fuzzy inference system (FIS). Η απλοποίηση του μοντέλου με λιγότερα χαρακτηριστικά, παρόλο που θα μείωνε την ακρίβεια, ίσως να βοηθούσε στη μείωση των λανθασμένων ταξινομήσεων και στην καλύτερη γενίκευση του μοντέλου.

Διαφορά μεταξύ εκπαίδευσης και γενίκευσης: Η απόδοση του τελικού μοντέλου στο testing set δείχνει ότι το σφάλμα εκπαίδευσης μπορεί να παραμένει μικρό, αλλά όταν εφαρμόζεται στο test set παρατηρείται μείωση στην απόδοση. Αυτό οφείλεται εν μέρει στην επικάλυψη των ασαφών συνόλων. Οι ασαφείς κανόνες του μοντέλου αποτυγχάνουν να αποδώσουν καλά όταν τα όρια μεταξύ των κλάσεων δεν είναι ξεκάθαρα.

Συνολική απόδοση ταξινομητή: Το μοντέλο καταφέρνει να ταξινομήσει με σχετική ακρίβεια κάποιες κλάσεις, όπως την κλάση 1, αλλά η απόδοσή του μειώνεται δραματικά σε άλλες, κυρίως στην κλάση 5. Το γεγονός ότι η κλάση 5 σχεδόν εξ ολοκλήρου κατανέμεται στις κλάσεις 3 και 4 καθιστά το μοντέλο ανεπαρκές. Παρά το grid search, το ποσοστό ακρίβειας υποδηλώνει ότι το πρόβλημα ταξινόμησης παραμένει δύσκολο για τα TSK μοντέλα, ειδικά σε datasets με υψηλή διαστασιμότητα.