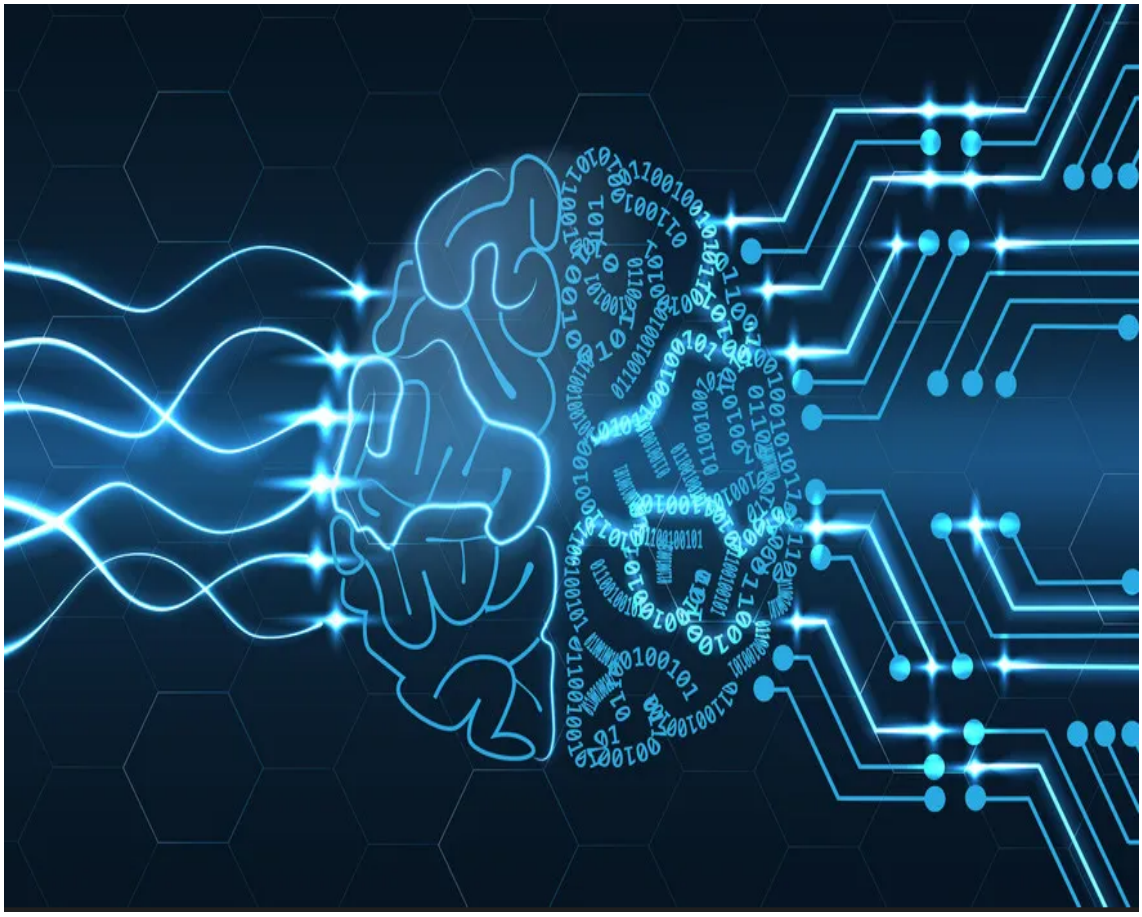
Υπολογιστική Νοημοσύνη - Τέταρτη Εργασία

Ονοματεπώνυμο: Πρωτοψάλτης Παναγιώτης
AEM: 9847
Email: pprotops@ece.auth.gr

March 10, 2024



Εισαγωγή

Σε αυτήν την εργασία θέλουμε να διερευνήσουμε την ικανότητα των μοντέλων TSK στην επίλυση προβλημάτων ταξινόμησης. Συγκεκριμένα επιλέγονται δύο σύνολα δεδομένων από το U CI repository με σκοπό την ταξινόμηση δειγμάτων στις εκάστοτε κλάσεις τους, με χρήση ασαφών νευρωνικών μοντέλων.

Εφαρμογή σε απλό dataset

Επιλέγουμε από το UCI repository το Haberman's Survival dataset, που περιέχει 306 δείγματα από 3 χαρακτηριστικά το καθένα.

Θα ακολουθήσουμε τα παρακάτω βήματα:

- Διαχωρισμός του dataset σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου (60-20-20)
- Η εκπαίδευση του μοντέλου έγινε σε 100 κύκλους εποχών.
- Εκπαίδευση TSK μοντέλων με διαφορετικές παραμέτρους, με χρήση της υβριδικής μεθόδου και με μεταβαλλόμενο πλήθος ασαφών κανόνων, με σκοπό να μελετηθεί η επίδραση της διαμέρισης του χώρου εισόδου σε συνάρτηση με την πολυπλοκότητα που επιφέρει στην απόδοση του ταξινομητή.
- Αξιολόγηση των μοντέλων με τους ακόλουθους δείκτες απόδοσης:
  - Error matrix, ο οποίος βοηθά στην οπτικοποίηση της απόδοσης ενός ταξινομητή
  - Overall accuracy, η οποία είναι η συνολική ακρίβεια ενός ταξινομητή
  - Producer's accuracy - User's accuracy
  - $\hat{K}$

Παρακάτω θα δοθούν τα γραφήματα για κάθε μία από τις παραπάνω περιπτώσεις. Για κάθε περίπτωση έχουμε 4 γραφήματα.

- 1 γράφημα για το "Learning-Curve"
- 1 γράφημα για τις συναρτήσεις συμμετοχής πριν την εκπαίδευση
- 1 γράφημα για τις συναρτήσεις συμμετοχής μετά την εκπαίδευση
- 1 γράφημα με το confusion matrix

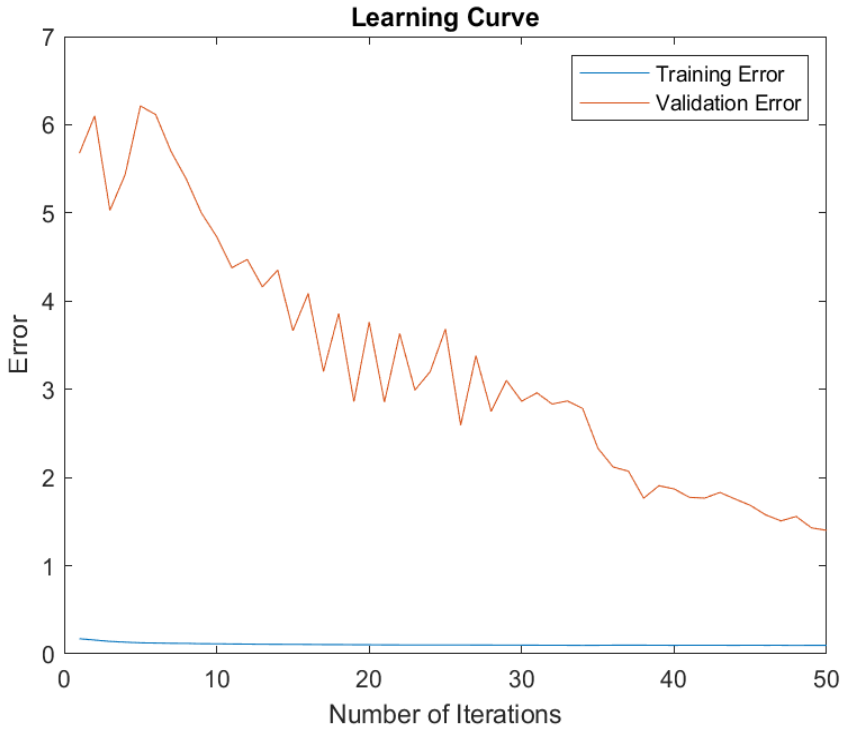


Figure 1: Καμπύλη εκμάθησής του 1ου μοντέλου

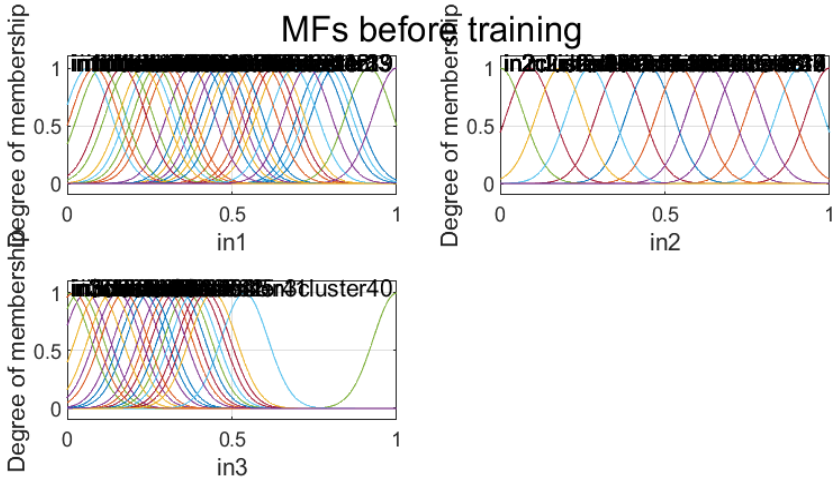


Figure 2: Συναρτήσεις συμμετοχής πριν την εκπαίδευση του 1ου μοντέλου

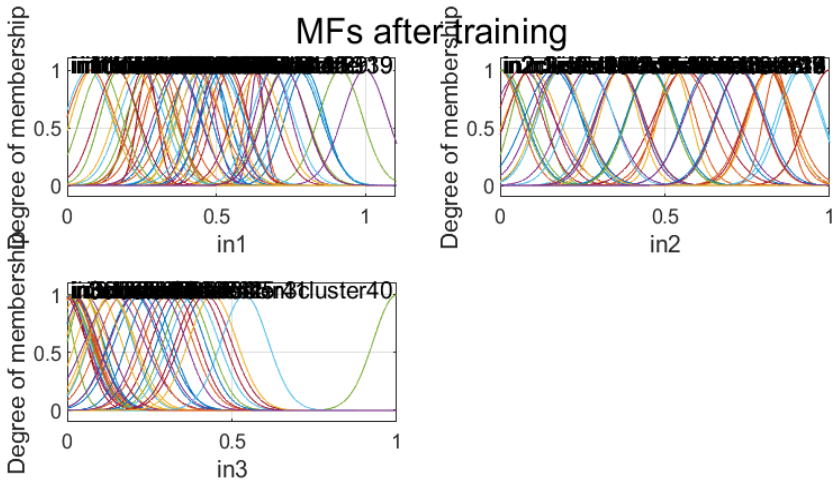


Figure 3: Συναρτήσεις συμμετοχής μετά την εκπαίδευση του 1ου μοντέλου

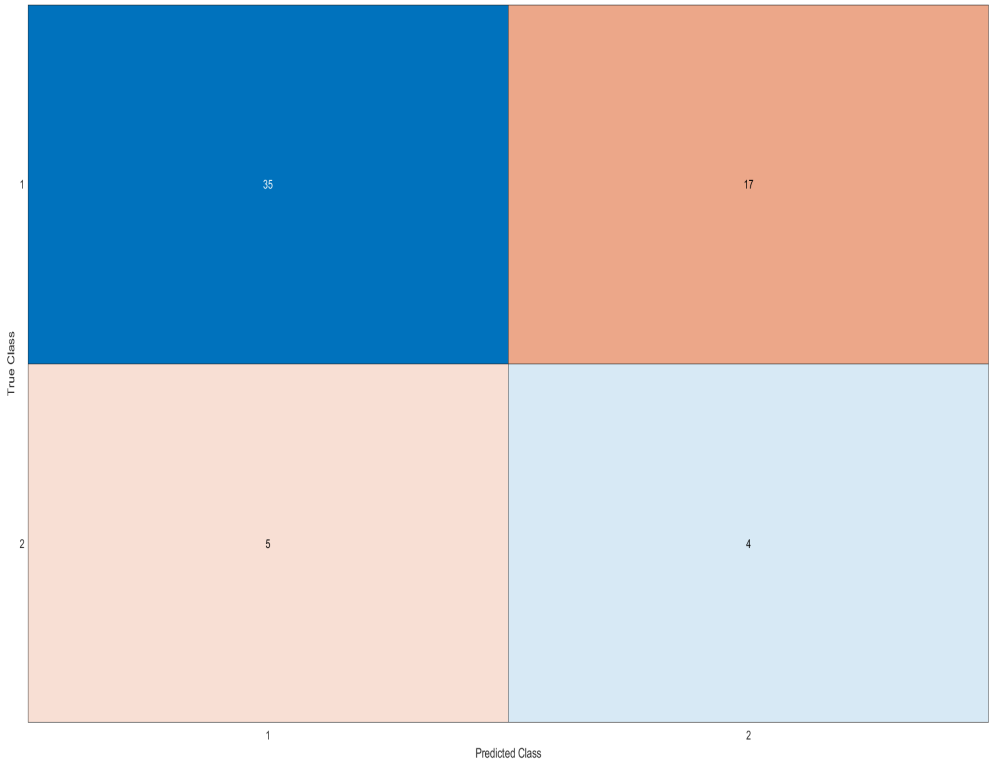


Figure 4: Confusion Matrix 1ου μοντέλου

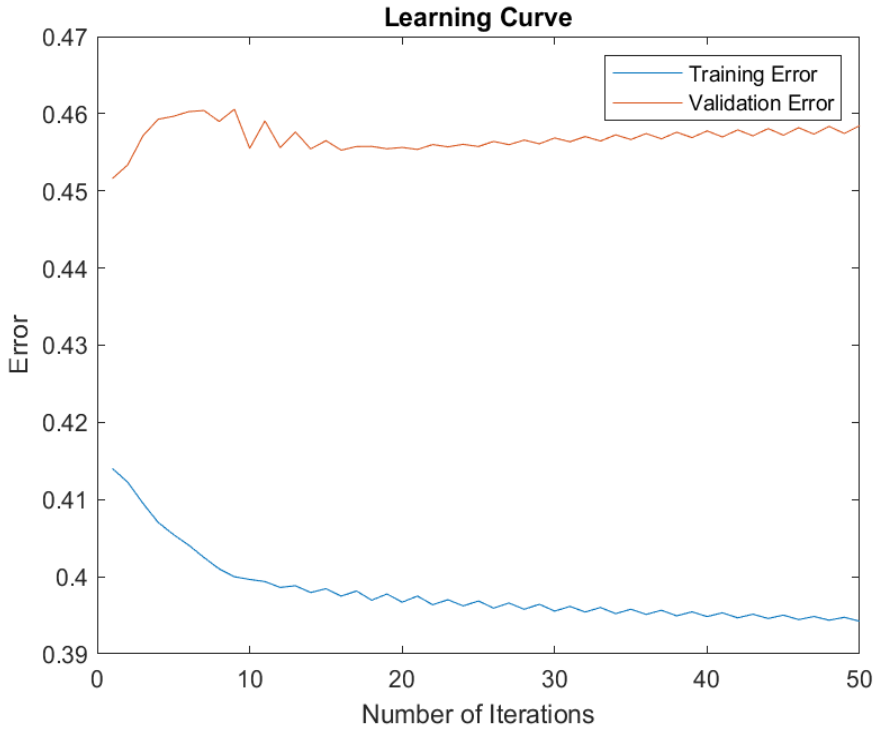


Figure 5: Καμπύλη εκμάθησής του 2ου μοντέλου

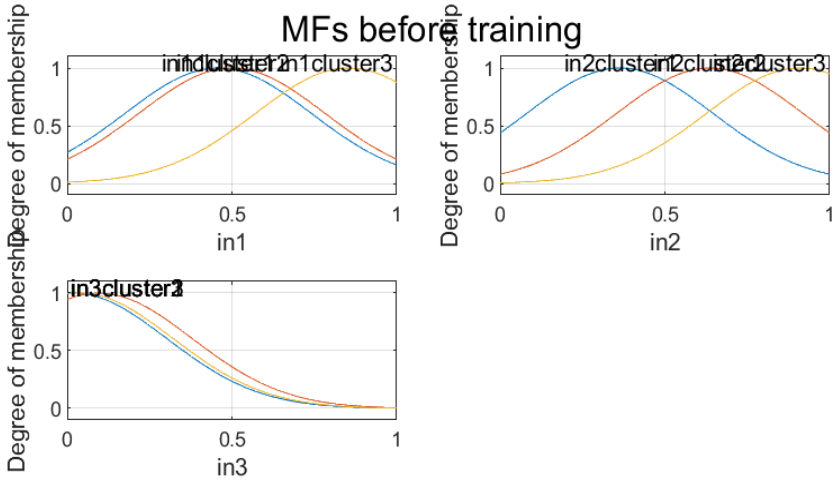


Figure 6: Συναρτήσεις συμμετοχής πριν την εκπαίδευση του 2ου μοντέλου

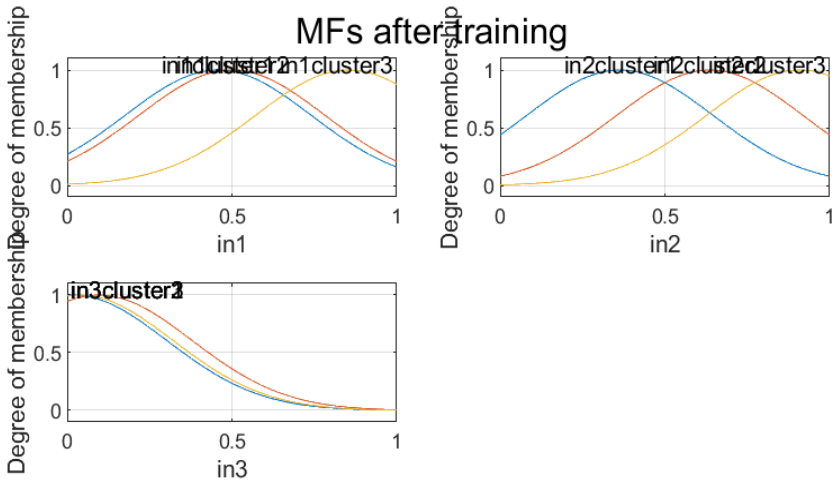


Figure 7: Συναρτήσεις συμμετοχής μετά την εκπαίδευση του 2ου μοντέλου

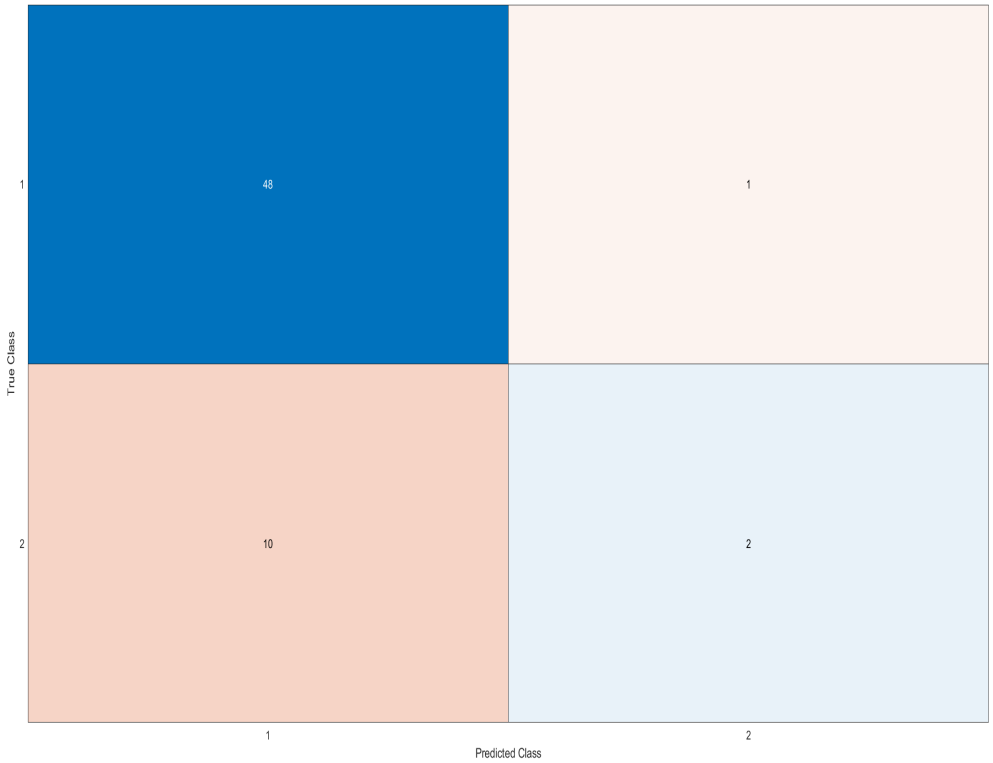


Figure 8: Confusion Matrix 2ου μοντέλου

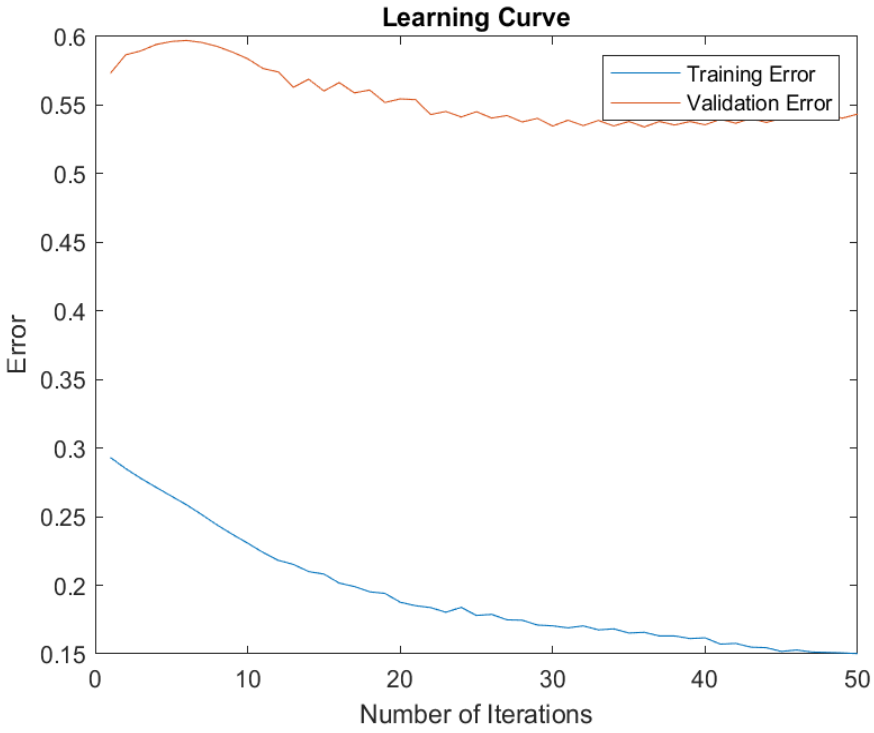


Figure 9: Καμπύλη εκμάθησής του 3ου μοντέλου

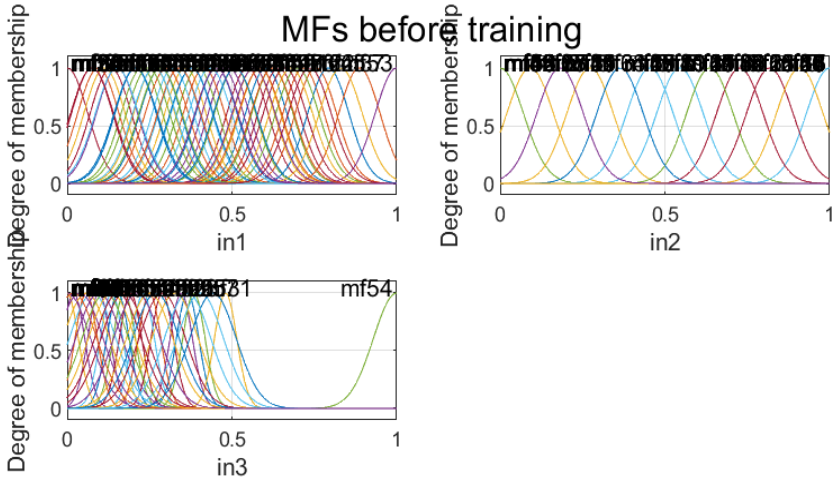


Figure 10: Συναρτήσεις συμμετοχής πριν την εκπαίδευση του 3ου μοντέλου

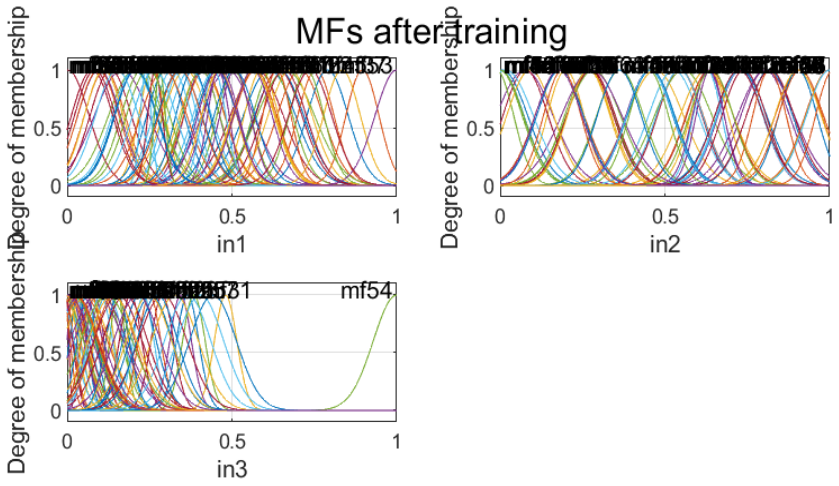


Figure 11: Συναρτήσεις συμμετοχής μετά την εκπαίδευση του 3ου μοντέλου

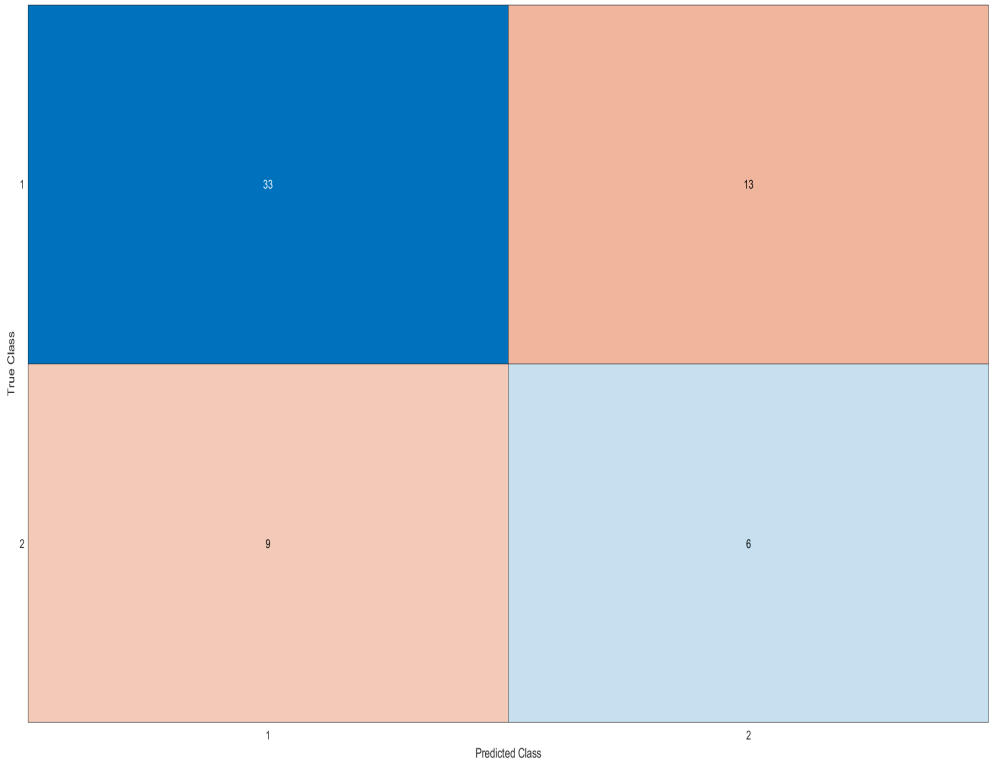


Figure 12: Confusion Matrix 3ου μοντέλου

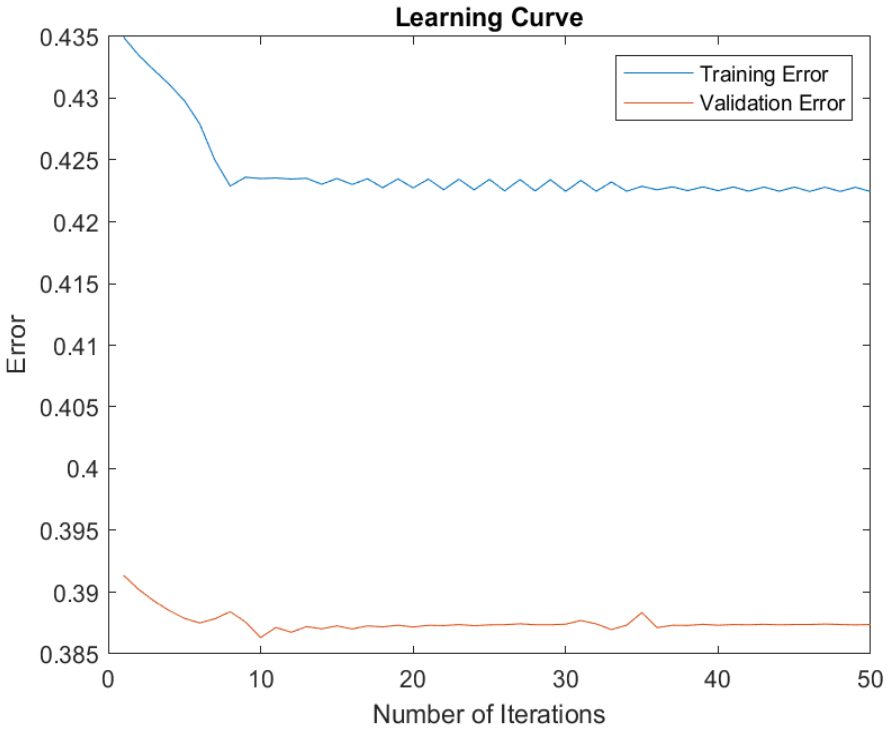


Figure 13: Καμπύλη εκμάθησής του 4ου μοντέλου

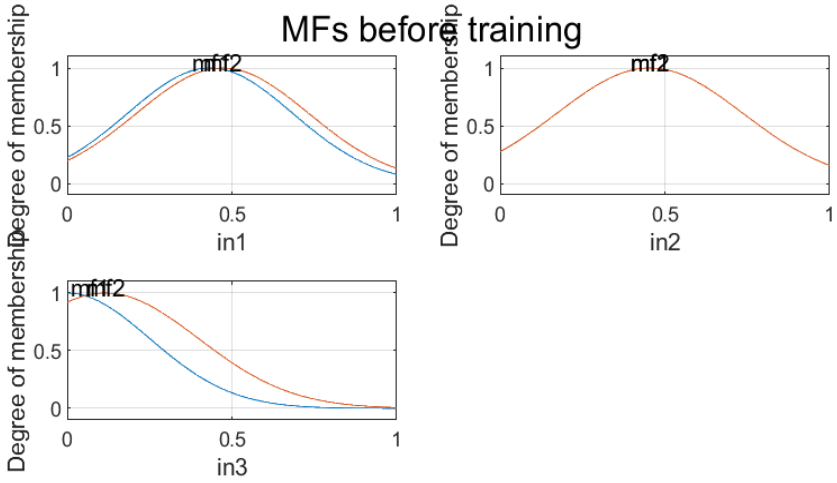


Figure 14: Συναρτήσεις συμμετοχής πριν την εκπαίδευση του 4ου μοντέλου

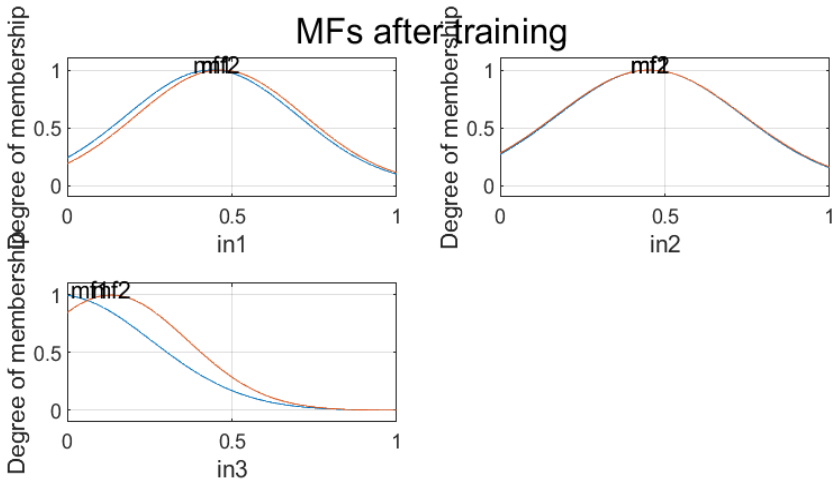


Figure 15: Συναρτήσεις συμμετοχής μετά την εκπαίδευση του 4ου μοντέλου

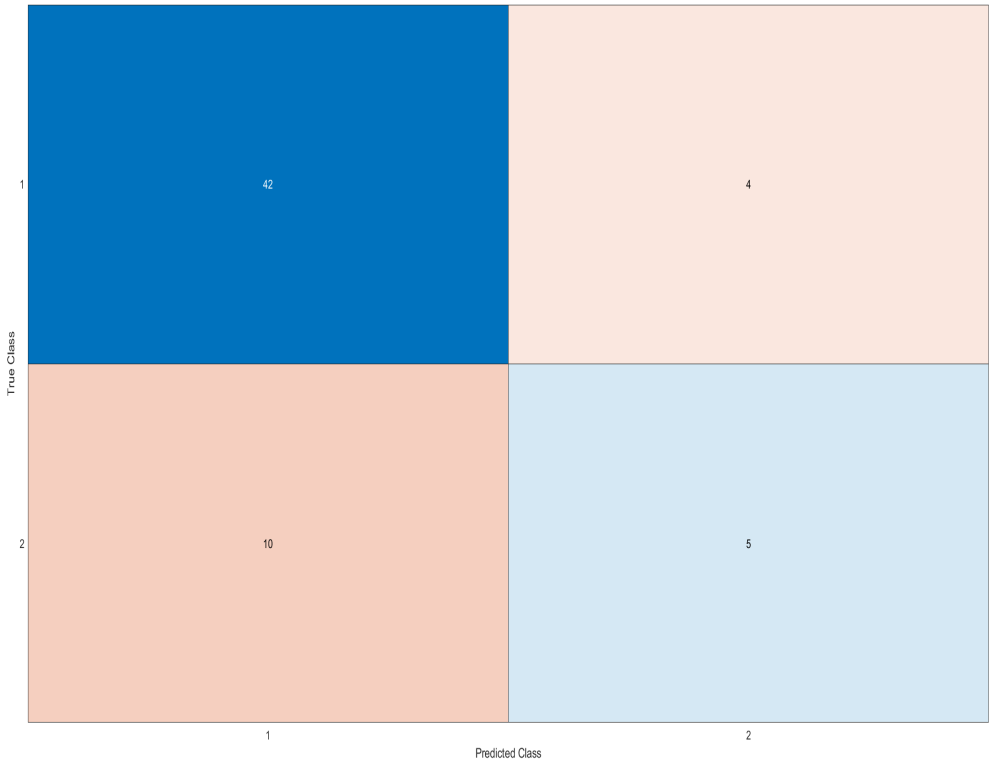


Figure 16: Confusion Matrix 4ου μοντέλου

Γενικές Παρατηρήσεις

Όπως παρατηρούμε τα validation error είναι πιο μεγάλα από τα training που είναι λογικό καθώς τα training εκπαιδεύονται σε περισσότερα δεδομένα. Όσο προχωράνε οι επαναλήψεις, το error και των δύο γραφημάτων μειώνονται και προσεγγίζουν καλύτερες τιμές. Παρατηρούμε ότι το 3ο μοντέλο πετυχαίνει το μικρότερο σφάλμα, μετά ακολουθεί το 4ο, μετά το 2ο και τέλος το 1ο. Τα μοντέλα έκαναν 8, 3, 13 και 3 δευτερόλεπτα για να τρέξουν, τρέχουν δηλαδή σε μικρό χρονικό διάστημα.

Ειδικές Παρατηρήσεις

Μοντέλο 1

Στο figure 1 βλέπουμε ότι το σφάλμα ξεκινάει από το 6 για το validation και από το 0.2 για το training και καταλήγει στο 1.2 και 0.1 αντίστοιχα. Στο figure 2 και 3 βλέπουμε τις συναρτήσεις συμμετοχής και στο figure 4 βλέπουμε το confusion matrix, το οποίο μας δείχνει τι κατανεμήθηκε σε ποια κλάση. Όπως παρατηρούμε ταξινομήθηκαν σωστά 39 δείγματα και 22 λάθος που μας κάνει ένα ποσοστό επιτυχίας 64%, το οποίο δεν είναι αρκετά καλό για έναν ταξινομητή. Παρακάτω δίνονται οι μετρικές για το εν λόγω μοντέλο.

Overall Accuracy	$\hat{K}$	Producers Accuracy	Users Accuracy
0.639344	0.075758	[0.875000, 0.190476]	[0.673077, 0.444444]

Παρατηρούμε λοιπόν πως υπάρχει χαμηλό ποσοστό σωστής ταξινόμησης και πως αρκετά από τα δεδομένα που ανήκουν στην κλάση 1 προβλέφθηκαν ότι ανήκουν στην κλάση 2.

Μοντέλο 2

Στο figure 5 βλέπουμε ότι το σφάλμα ξεκινάει από το 0.45 για το validation και από το 0.415 για το training και καταλήγει στο 0.46 και 0.405 αντίστοιχα. Στο figure 6 και 7 βλέπουμε τις συναρτήσεις συμμετοχής και στο figure 8 βλέπουμε το confusion matrix, το οποίο μας δείχνει τι κατανεμήθηκε σε ποια κλάση. Όπως παρατηρούμε ταξινομήθηκαν σωστά 50 δείγματα και 11 λάθος που μας κάνει ένα ποσοστό επιτυχίας 82%, το οποίο δεν είναι και το καλύτερο ποσοστό που θα μπορούσε να επιτευχθεί, αλλά είναι αρκετά καλύτερο από τον πρώτο ταξινομητή. Παρακάτω δίνονται οι μετρικές για το εν λόγω μοντέλο.

Overall Accuracy	$\hat{K}$	Producers Accuracy	Users Accuracy
0.819672	0.204033	[0.827586, 0.666667]	[0.979592, 0.166667]

Παρατηρούμε λοιπόν πως υπάρχει ένα ικανοποιητικό ποσοστό σωστής ταξινόμησης, αλλά πως αρκετά από τα δεδομένα που ανήκουν στην κλάση 2 προβλέφθηκαν ότι ανήκουν στην κλάση 1.

Μοντέλο 3

Στο figure 9 βλέπουμε ότι το σφάλμα ξεκινάει από το 0.58 για το validation και από το 0.3 για το training και καταλήγει στο 0.55 και 0.15 αντίστοιχα. Στο figure 10 και 11 βλέπουμε τις συναρτήσεις συμμετοχής και στο figure 12 βλέπουμε το confusion matrix, το οποίο μας δείχνει τι κατανεμήθηκε σε ποια κλάση. Όπως παρατηρούμε ταξινομήθηκαν σωστά 39 δείγματα και 22 λάθος που μας κάνει ένα ποσοστό επιτυχίας 64%, το οποίο δεν είναι αρκετά καλό για έναν ταξινομητή και είναι ακριβώς ολόιδιο με αυτό του 1ου ταξινομητή, η διαφορά τους είναι ότι αυτός ο ταξινομητής δεν ταξινόμησε σωστά τα δείγματα ομοιόμορφα δηλαδή είχε περίπου τα ίδια false positive και false negative, καθώς επίσης ότι έχει μεγαλύτερο K που σημαίνει ότι είναι καλύτερος από το 1ο. Παρακάτω δίνονται οι μετρικές για το εν λόγω μοντέλο.

Overall Accuracy	$\hat{K}$	Producers Accuracy	Users Accuracy
0.639344	0.107713	[0.785714, 0.315789]	[0.717391, 0.400000]

Παρατηρούμε λοιπόν πως υπάρχει χαμηλό ποσοστό σωστής ταξινόμησης και πως αρκετά από τα δεδομένα που ανήκουν στην κλάση 1 και κλάση 2 προβλέφθηκαν ότι ανήκουν στην άλλη κλάση.

Μοντέλο 4

Στο figure 13 βλέπουμε ότι το σφάλμα ξεκινάει από το 0.435 για το validation και από το 0.39 για το training και καταλήγει στο 0.425 και 0.385 αντίστοιχα. Στο figure 14 και 15 βλέπουμε τις συναρτήσεις συμμετοχής και στο figure 16 βλέπουμε το confusion matrix, το οποίο μας δείχνει τι κατανεμήθηκε σε ποια κλάση. Όπως παρατηρούμε ταξινομήθηκαν σωστά 47 δείγματα και 14 λάθος που μας κάνει ένα ποσοστό επιτυχίας 77%, το οποίο είναι μέτριο για έναν ταξινομητή. Παρακάτω δίνονται οι μετρικές για το εν λόγω μοντέλο.

Overall Accuracy	$\hat{K}$	Producers Accuracy	Users Accuracy
0.770492	0.284757	[0.807692, 0.555556]	[0.913043, 0.333333]

Παρατηρούμε λοιπόν πως υπάρχει χαμηλό ποσοστό σωστής ταξινόμησης και πως αρκετά από τα δεδομένα που ανήκουν στην κλάση 2 προβλέφθηκαν ότι ανήκουν στην κλάση 1.

Τελικό συμπέρασμα

Τέλος από τα γραφήματα και από τα πινακάκια που δόθηκαν καταλαβαίνουμε πως το 2ο μοντέλο είναι το πιο αξιόπιστο για ταξινόμηση, μετά είναι το 4ο και τέλος σε ισοβαθμία ποσοστού ταξινόμησης βρίσκονται το 1ο και το 3ο, αλλά το είναι πιο "επικίνδυνο" καθώς προβλέπει περισσότερα false negative από το 3ο, το οποίο είναι κακό να συμβαίνει σε ότι αφορά την υγεία, καθώς άμα διαγνωστεί κάποιο false positive, θα βγούνε αρνητικές οι επόμενες εξετάσεις του ασθενή για την εν λόγω ασθένεια, αλλά με ένα false negative θα μπορούσε ο ασθενής να μην κάνει άλλη εξέταση καθησυχασμένος και να το συνειδητοποιήσει αργότερα όταν θα είναι πολύ αργά. Επίσης το 3ο έχει μεγαλύτερη σταθερα τυχαιότητας και συνεπώς είναι καλύτερο μοντέλο από το 1ο.

Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Στη δεύτερη φάση της εργασίας θα ασχοληθούμε με μία πιο συστηματική προσέγγιση στο πρόβλημα της χρήσης των ασαφών νευρωνικών μοντέλων σε προβλήματα ταξινόμησης.

Το dataset που θα επιλεχθεί είναι το Epileptic Seizure Recognition, το οποίο έχει υψηλότερο βαθμό διαστασιμότητας από το 1ο, με 11500 δείγματα και με 179 χαρακτηριστικά το καθένα. Είναι εύκολο να δούμε ότι σε αυτό το dataset είναι δύσκολη η εφαρμογή ενός TSK μοντέλου, όπως ήταν τα μοντέλα του 1ου μέρους. Καλούμαστε λοιπόν, να επιλέξουμε τα χαρακτηριστικά που εισάγει το πρόβλημα και να χρησιμοποιήσουμε την ασαφή ομαδοποίηση, ώστε να μειώσουμε την πολυπλοκότητα, αλλά θα εισαχθούν στο πρόβλημα ο αριθμός των χαρακτηριστικών προς επιλογή και ο αριθμός των ομάδων που θα δημιουργηθούν.

Σε αυτή την εργασία χρησιμοποιώντας το "grid-search" θα βρούμε τις βέλτιστες τιμές για τις δύο παραπάνω παραμέτρους. Επιλέχθηκαν οι τιμές 0.2, 0.4, 0.6 και 1 για την ακτίνα και οι τιμές 5, 10, 15, 20 για τον αριθμό των χαρακτηριστικών. Παρακάτω θα δοθούν κάποια διαγράμματα για τις τιμές του σφάλματος, της ακρίβειας και τον αριθμό των κανόνων για κάθε συνδυασμό αυτών των επιλογών (δηλαδή 0.2-5, 0.2-10, 0.2-15, 0.2-20, 0.4-5, 0.4-10 κτλ). Αρχικά για να βρούμε τα επικρατέστερα 5,10,15 ή 20 χαρακτηριστικά του dataset, χρησιμοποιήθηκε η συνάρτηση relieff του matlab, που μας επιστρέφει την σειρά από τα επικρατέστερα αυτά χαρακτηριστικά. Έπειτα για κάθε ένα μοντέλο έγινε το "cross-validation", με το οποίο καταφέραμε να βρούμε το σφάλμα, την ακρίβεια και τον αριθμό των κανόνων, τα οποία μας βοήθησαν στην καλύτερη επιλογή συνδυασμού που θα δούμε παρακάτω.

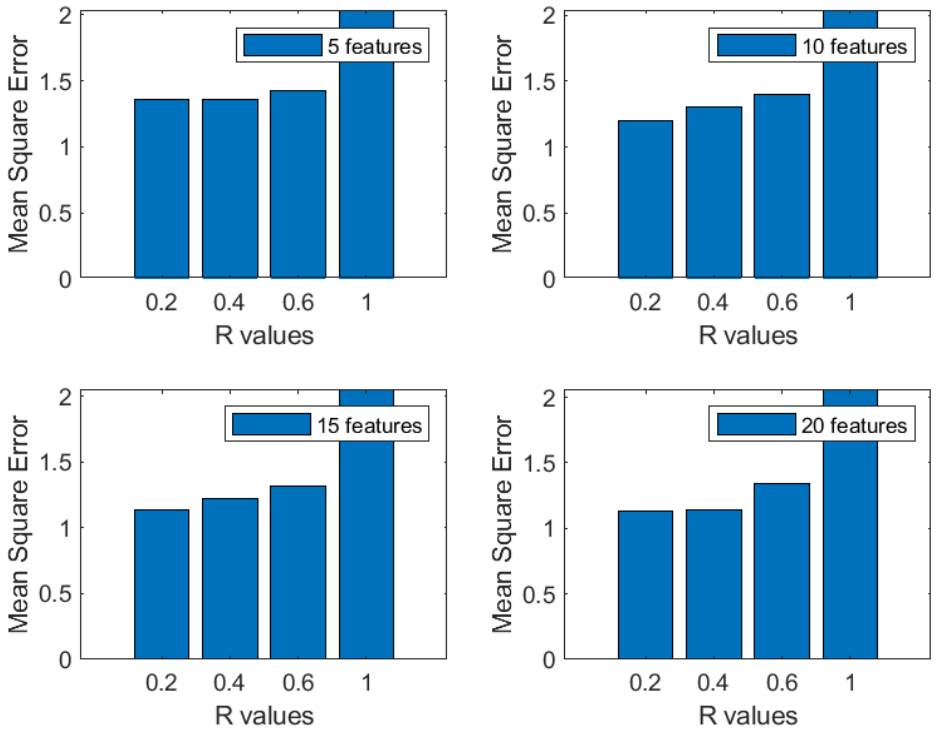


Figure 17: Σφάλμα πρόβλεψης για κάθε συνδυασμού σε 2D

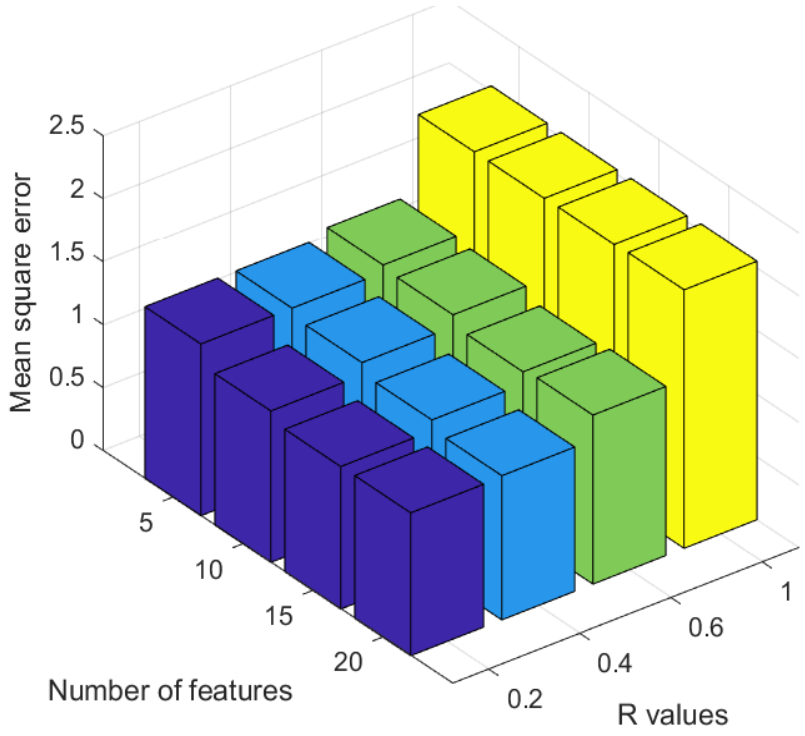


Figure 18: Μέσο τετραγωνικό σφάλμα σε 3D

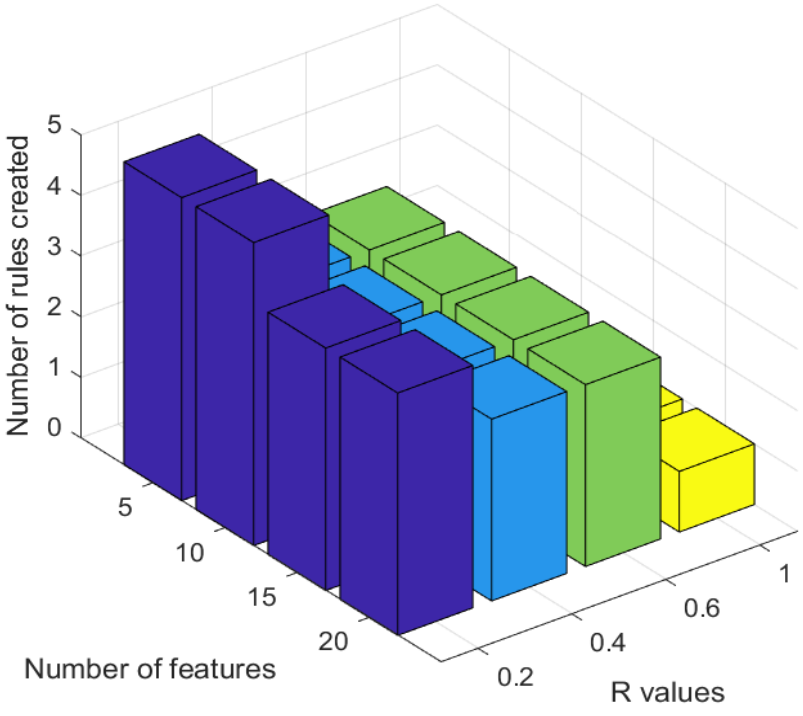


Figure 19: Ο αριθμός των κανόνων σε 3D

Παρατηρήσεις

Ο κώδικας για το παραπάνω κομμάτι έτρεξε σε **20 λεπτά**, λόγω της μεγάλης επιλογής των χαρακτηριστικών. Στα παραπάνω διαγράμματα παρατηρούμε ότι ο συνδυασμός τιμών που έχει το μικρότερο σφάλμα είναι αυτός του **0.2 ακτίνα και 15 αριθμός χαρακτηριστικών**. Ο συνδυασμός αυτός ήταν κοντά στους συνδυασμούς 0.2-20 και 0.4-20. Παρόλα αυτά κρίνουμε πως θα ήταν καλύτερο να επιλεγεί αυτός επειδή θέλουμε να φτιάξουμε ένα μοντέλο με το μικρότερο δυνατό σφάλμα και έτσι επιλέχθηκε ο συνδυασμός 0.2-15.

Τελικό TSK μοντέλο με βέλτιστες τιμές παραμέτρων

Με τον συνδυασμό που επιλέχθηκε, δηλαδή 0.2-15, θα κάνουμε ένα μοντέλο παρόμοιας λογικής με το πρώτο μέρος της εργασίας. Θα χωρίσουμε το dataset σε 60-20-20 για training, validation και testing αντίστοιχα και θα το εκπαιδεύσουμε για 150 κύκλους επαναλήψεων. Παρακάτω δίνονται τα διαγράμματα για την εκπαίδευση αυτού του μοντέλου.

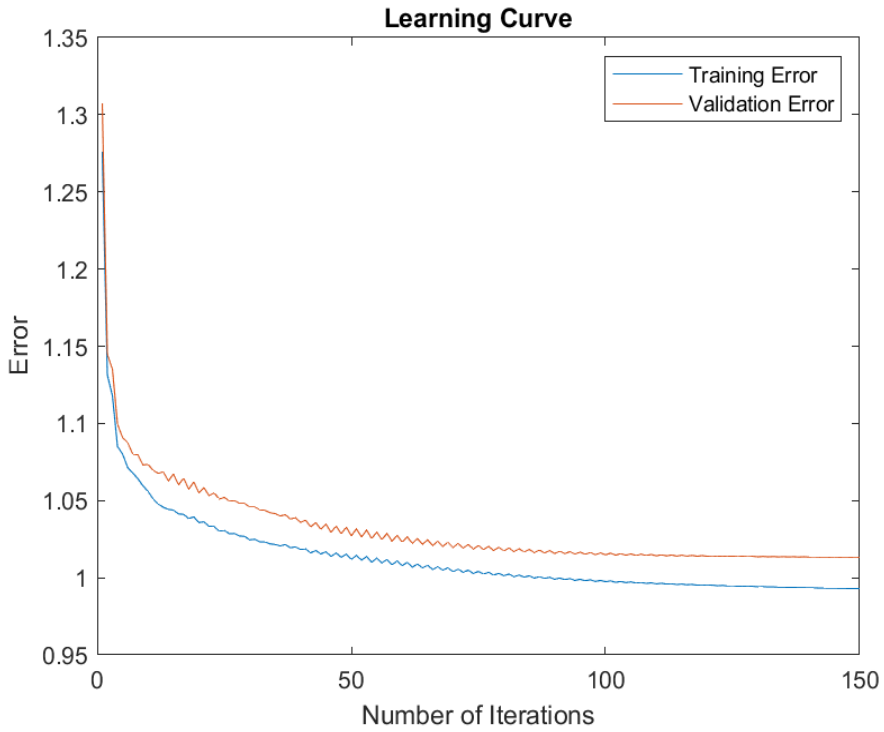


Figure 20: Καμπύλη εκμάθησής του τελικού μοντέλου

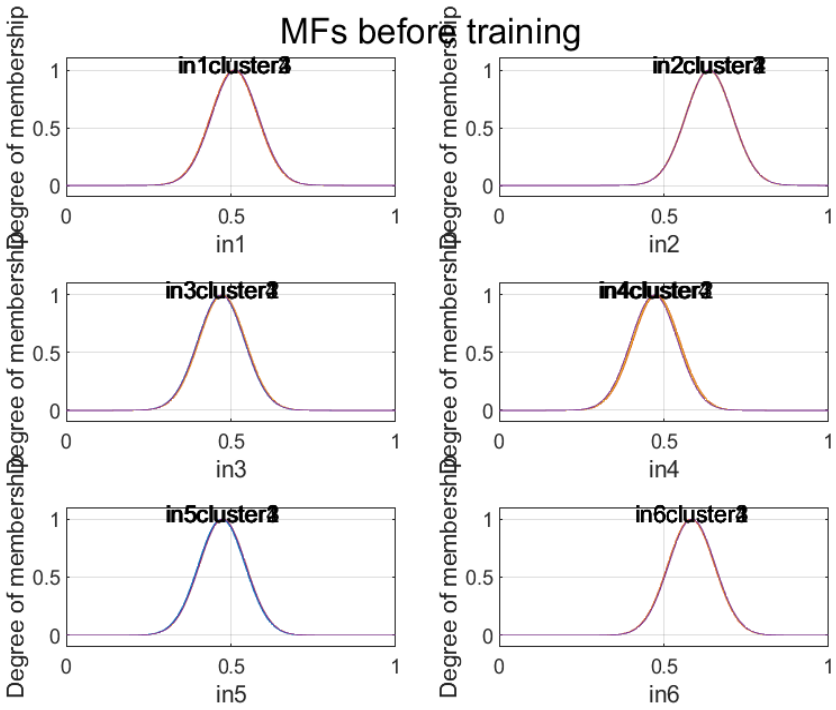


Figure 21: Συναρτήσεις συμμετοχής πριν την εκπαίδευση του τελικού μοντέλου

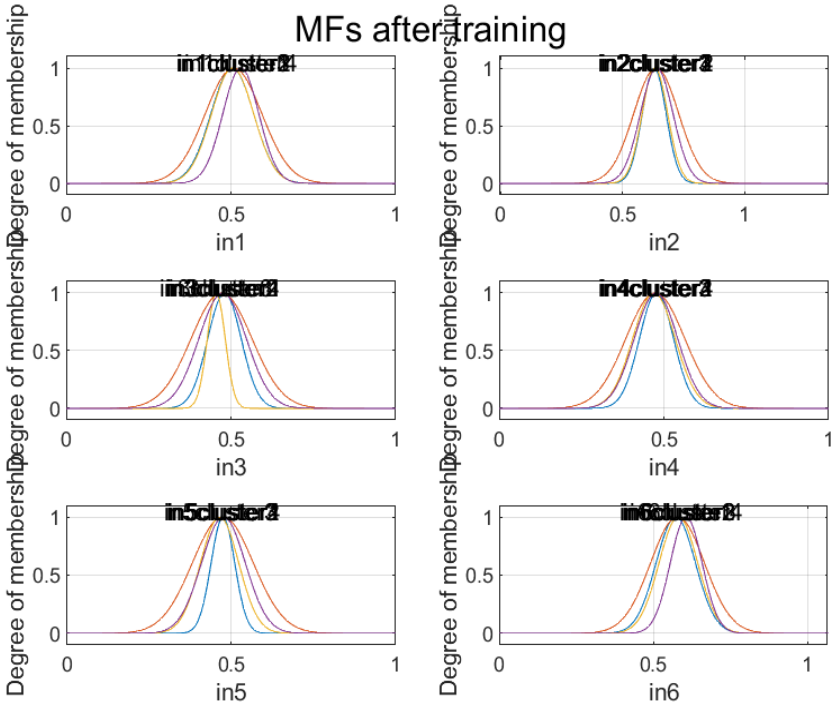


Figure 22: Συναρτήσεις συμμετοχής μετά την εκπαίδευση του τελικού μοντέλου

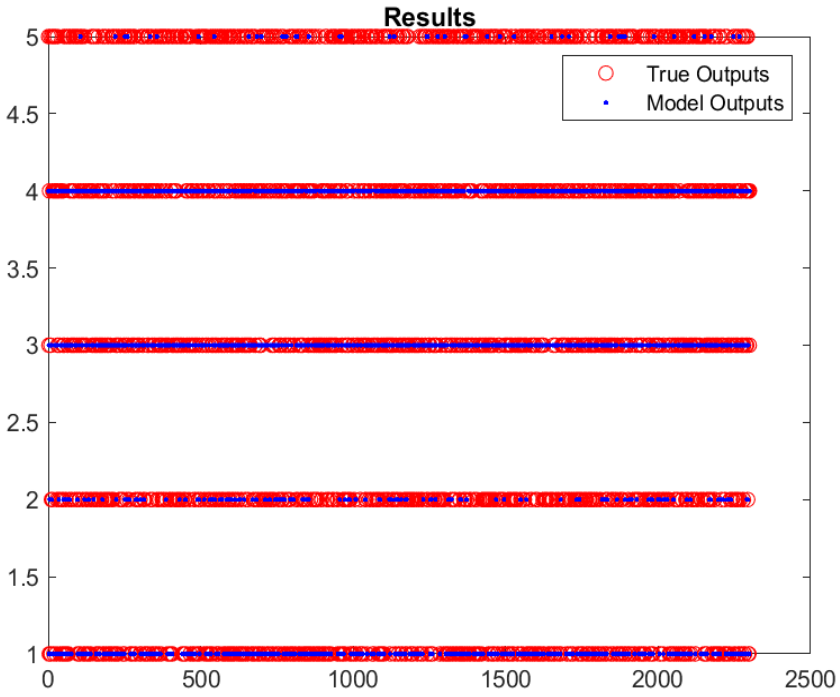


Figure 23: Διαφορά πραγματικής και εκτιμώμενης εξόδου

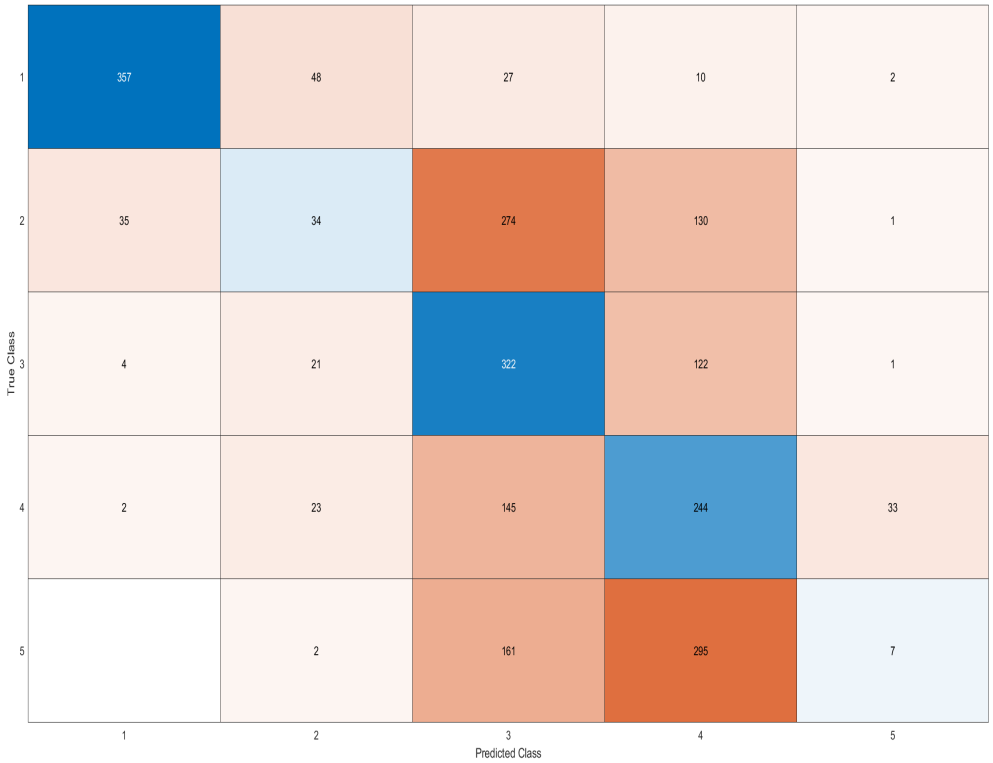


Figure 24: Confusion Matrix του μοντέλου

Παρατηρήσεις

Ο πίνακας για τις τιμές των μετρικών του μοντέλου είναι ο παρακάτω.

Overall Accuracy	$\hat{K}$	**Producers Accuracy**	**Users Accuracy**
0.419130	0.274852	[0.896985, 0.265625, 0.346609, 0.304619, 0.159091]	[0.804054, 0.071730, 0.685106, 0.545861, 0.015054]

Παρατηρούμε, ότι το μοντέλο για τις βέλτιστες τιμές των παραμέτρων που επιλέχθηκαν δεν είναι καθόλου ακριβές, καθώς έχει ακρίβεια 0.42. Παρόλα αυτά το K είναι μεγαλύτερο από τα μοντέλα του πρώτου μέρους της εργασίας. Στο figure 20 βλέπουμε ότι το validation error ξεκινάει από το 1.3 και το training από το 1.25 και καταλήγουν στο 1.05 και 1 αντίστοιχα, χωρίς την ύπαρξη ταλαντώσεων ή υπερεκπαίδευσης. Στο figure 21 και 22 βλέπουμε τις συναρτήσεις συμμετοχής πριν και μετά την εκπαίδευση του μοντέλου και στο figure 23 βλέπουμε την πραγματική έξοδο με το κόκκινο χρώμα και την εκτιμώμενη έξοδο από το μοντέλο μας.

Συμπεράσματα

Οι κανόνες που χρησιμοποιήθηκαν όπως βλέπουμε και στο figure 19 είναι ελάχιστοι. Σε περίπτωση που είχαμε επιλέξει **grid-partitioning** αντί για grid-search, τότε θα είχαμε $2^1 70$ ή $3^1 70$ κανόνες που είναι απαγορευτικό για τον χρόνο που θα έκανε να τρέξει ο κώδικας μας, οπότε η επιλογή του grid-search ήταν καλύτερη του grid-partitioning.

Όπως παρατηρούμε στο confusion matrix του μοντέλου μας, στην διαγώνιο βρίσκονται τα δείγματα που ταξινομήθηκαν σωστά. Παρατηρούμε ότι υπάρχει μία δυσκολία από το μοντέλο να αναγνωρίσει την 2η και την 5η κλάση, καθώς τις ταξινομεί κατά βάσει ως 3η και 4η. Παρ' όλα αυτά υπάρχει ένα καλό ποσοστό επιτυχίας στην κλάση 1 και 3.