# ΕΡΓΑΣΙΑ ΣΤΑ ΑΣΑΦΗ ΣΥΣΤΗΜΑΤΑ

ΟΝΟΜΑ: ΚΩΝΣΤΑΝΤΙΝΟΣ
ΕΠΙΘΕΤΟ: ΛΕΤΡΟΣ
ΣΧΟΛΗ: ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ: ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧ. ΚΑΙ ΜΗΧ. ΥΠΟΛΟΓΙΣΤΩΝ
ΑΕΜ: 8851
ΕΞΑΜΗΝΟ: 8ο
ΕΤΟΣ: 2019

# Επίλυση προβλημάτων παλινδρόμησης με χρήση μοντέλων TSK

# Ομάδα 4 – S08

## Περιεχόμενα

# Περιγραφή του Προβλήματος

Στόχος αυτής της εργασίας είναι η διερεύνηση της ικανότητας των TSK μοντέλων στην επίλυση προβλημάτων ταξινόμησης (classification) με χρήση ασαφών νευρωνικών μοντέλων. Η εργασία διακρίνεται σε δύο τμήματα στα οποία θα χρησιμοποιηθούν δύο διαφορετικά σετ δεδομένων. Σκοπός του πρώτου τμήματος είναι η εκπαίδευση και αξιολόγηση ενός πλήθους μοντέλων (με διαφορετικό πλήθος IF THEN κανόνων) τα οποία να μπορούν να ταξινομούν με επιτυχία το κάθε δείγμα στην κλάση που ανήκει. Αντίθετα στο δεύτερο μέρος, το μεγάλο πλήθος χαρακτηριστικών του δεύτερου σετ δεδομένων της εργασίας καθιστά τη διαδικασία εκμάθησης πολύ δυσκολότερη. Για την επίλυση του παραπάνω προβλήματος γίνεται μια επιλογή των πιο σημαντικών χαρακτηριστικών του σετ δεδομένων με τη βοήθεια του αλγορίθμου Relief και έπειτα αναζητείται από μια σειρά μοντέλων με διαφορετικές παραμέτρους, το μοντέλο με το ελάχιστο σφάλμα. Τέλος, γίνεται εκπαίδευση του μοντέλου αυτού, το οποίο χαρακτηρίζεται ως το βέλτιστο μοντέλο.

# Εφαρμογή στο Σετ Δεδομένων Avila

## Προετοιμασία του Σετ Δεδομένων

Το Avila Dataset της UCI περιλαμβάνει 10 χαρακτηριστικά και 20876 δείγματα ενώ έχει εξαχθεί από 800 εικόνες της «Βίβλου Avila», ενός γιγάντιου λατινικού αντιγράφου της Βίβλου του 12ου αιώνα. Ο στόχος ταξινόμησης συνίσταται στη συσχέτιση κάθε σχεδίου με ένα αντιγραφέα.

Αρχικά ταξινομούμε κατά αύξουσα σειρά το σετ δεδομένων με βάση τη στήλη που περιέχει τις διάφορες τιμές εξόδων (κλάσεις) 1–12 και προχωρούμε στην καταμέτρηση της συχνότητας εμφάνισης κάθε διαφορετικής τιμής εξόδου (με χρήση της εντολής tabulate).

Στη συνέχεια, πραγματοποιούμε διαχωρισμό του σετ δεδομένων σε τρία μη επικαλυπτόμενα υποσύνολα ως εξής:

1. 60% : Σετ Εκπαίδευσης – training set
2. 20% : Σετ Επικύρωσης – validation set
3. 20% : Σετ Ελέγχου – check set

με τρόπο τέτοιο ώστε οι παραπάνω συχνότητες εμφάνισης να διατηρούνται περίπου σταθερές και ανακατεύουμε το κάθε σετ ξεχωριστά.

## Περιγραφή της Διαδικασίας Εκπαίδευσης

Η εκπαίδευση γίνεται με την υβριδική μέθοδο, δηλαδή οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται με τη μέθοδο Backpropagation.

Δημιουργούμε, με τη συνάρτηση genfis() του MATLAB, το προς εκπαίδευση μοντέλο με βάση τα χαρακτηριστικά του πίνακα στο Σχήμα 1 για κάθε μοντέλο, κάνοντας χρήση της μεθόδου Subtractive Clustering και δίνοντας ως είσοδο τα δεδομένα εκπαίδευσης. Επίσης, οι συναρτήσεις

εξόδου μεταβάλλονται από την προκαθορισμένη μορφή Polynomial (Linear) σε Singleton (Constant) .

Στη συνέχεια εκπαιδεύουμε το μοντέλο με χρήση της συνάρτησης anfis() του MATLAB για 250 εποχές, προχωρούμε στην αξιολόγησή του και τέλος υπολογίζουμε τις ζητούμενα μεγέθη - δείκτες απόδοσης Error Matrix (Πίνακας Σφαλμάτων Ταξινόμησης), Overall Accuracy, Producer's Accuracy, User's Accuracy, $\hat{k}$ . Σημειώνουμε επίσης ότι πριν την αξιολόγηση του μοντέλου με το σετ δεδομένων ελέγχου, στρογγυλοποιούμε τις τιμές εξόδου στον κοντινότερο ακέραιο, καθώς το παρών πρόβλημα είναι πρόβλημα ταξινόμησης οπότε ο εκτιμώμενος αριθμός της κλάσης θα πρέπει να ανήκει στο σύνολο τιμών της εξόδου (δηλαδή ακέραιος από 1 έως 12).

Συνολικά εξετάζονται πέντε μοντέλα με διαφορετικές τιμές ακτίνας, δηλαδή της παραμέτρου της μεθόδου Subtractive Clustering που προσδιορίζει το εύρος επιρροής του κέντρου κάθε κλάσης, άρα και το πλήθος των IF THEN κανόνων. Τα πέντε μοντέλα λαμβάνουν τιμές ακτίνας από το σύνολο τιμών όπως φαίνεται στο Σχήμα 1.

| *TSK Model* | *Radius* | *Squash Factor* | *Number of Rules* | *Output Format* |
|---|---|---|---|---|
| *Model 1* | 0.8 | 0.5 | 4 | Singleton |
| *Model 2* | 0.8 | 0.45 | 8 | Singleton |
| *Model 3* | 0.3 | 0.475 | 12 | Singleton |
| *Model 4* | 0.7 | 0.4 | 16 | Singleton |
| *Model 5* | 0.5 | 0.432 | 20 | Singleton |

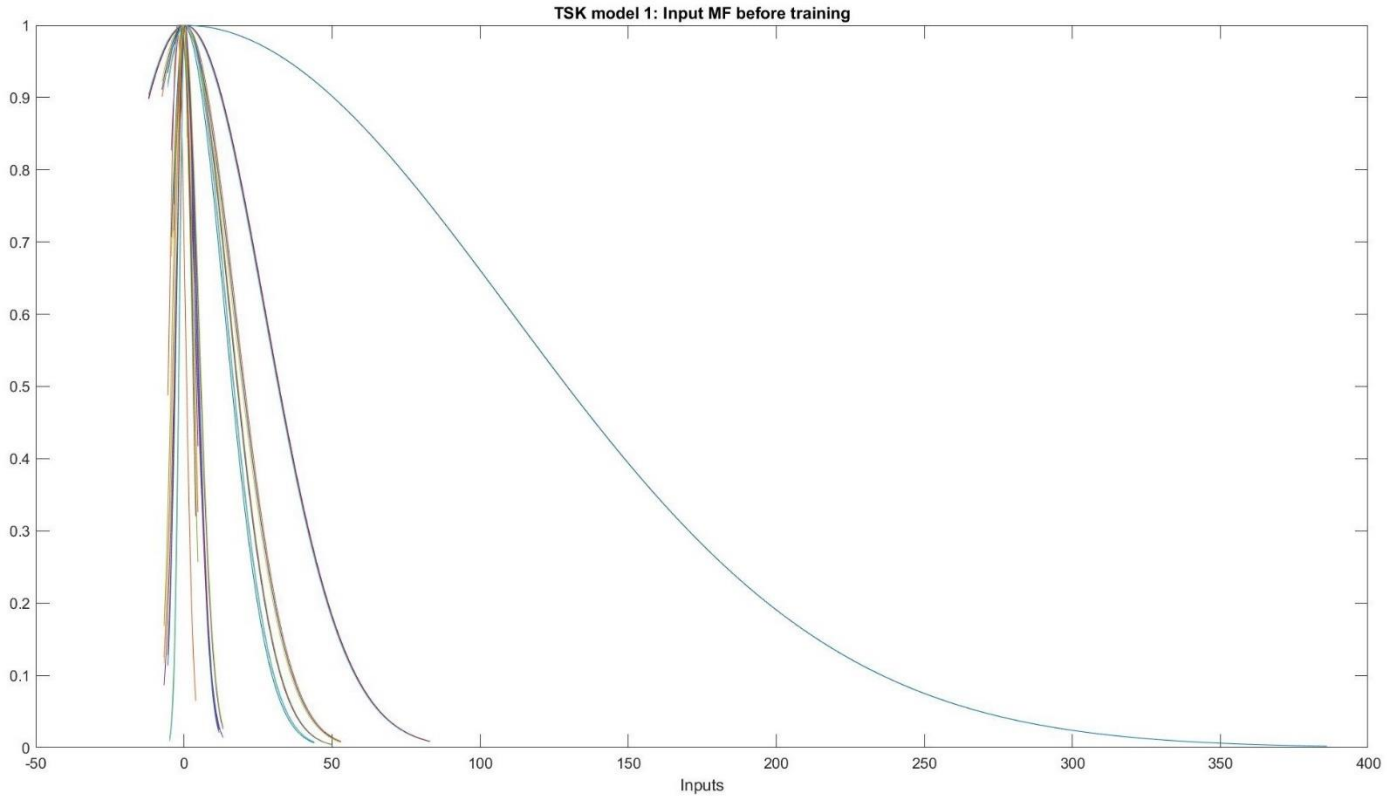
*Σχήμα 1: TSK Μοντέλα για διάφορες τιμές ακτίνας*

Στην παραπάνω εικόνα εμφανίζεται και μια ακόμα παράμετρος, Squash Factor. Για τις διάφορες τιμές ακτίνας ο αλγόριθμος παρήγαγε πολύ μικρό πλήθος κανόνων, κάτι που καθιστούσε ανέφικτο να πετύχουμε τη ζητούμενη προδιαγραφή για το πλήθος των κανόνων. Για το λόγο αυτό μεταβάλλουμε την παράμετρο Squash Factor των genfisOptions, από την προκαθορισμένη τιμή 1.25, σε αυτές που φαίνονται παραπάνω, σε κάθε περίπτωση.

## Αποτελέσματα TSK Μοντέλων και Δείκτες Απόδοσης

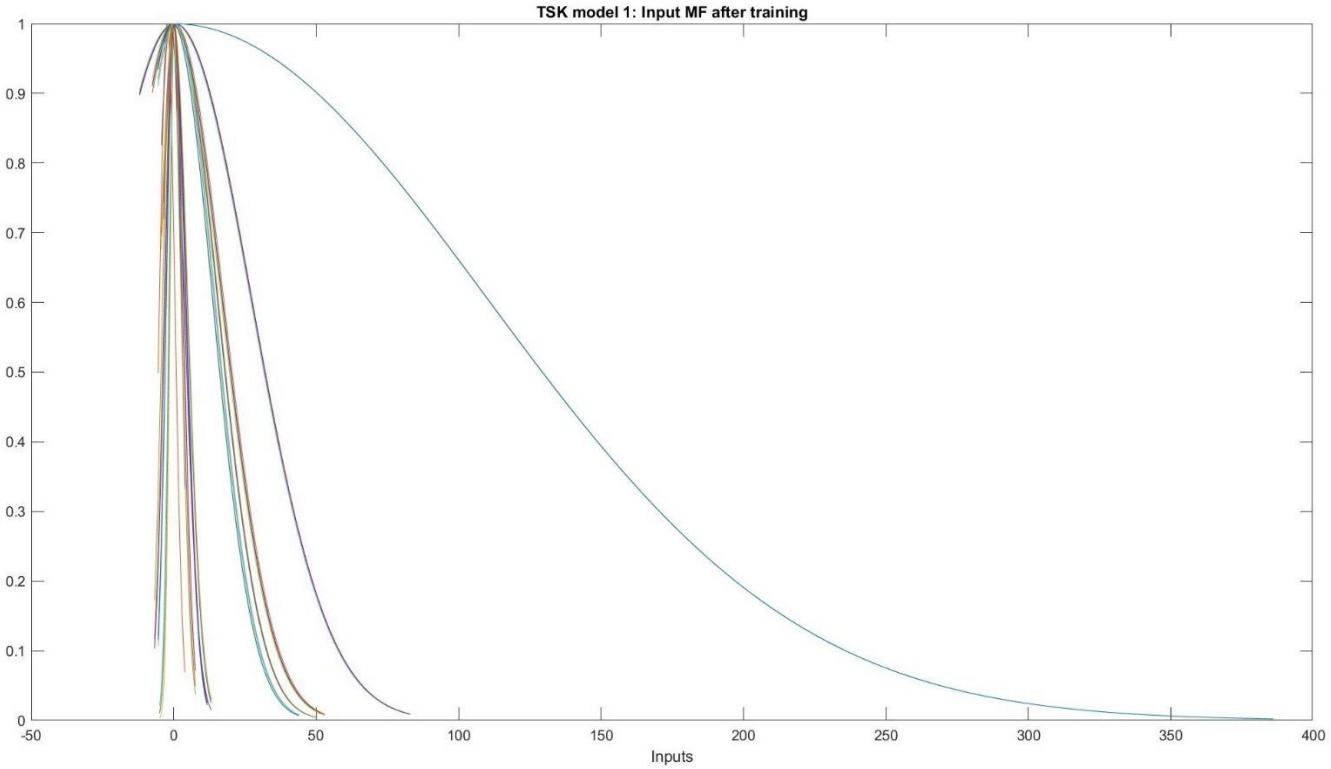
### TSK Μοντέλο 1 (4 Rules)

Οι συναρτήσεις συμμετοχής για το πρώτο μοντέλο πριν από τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 2.



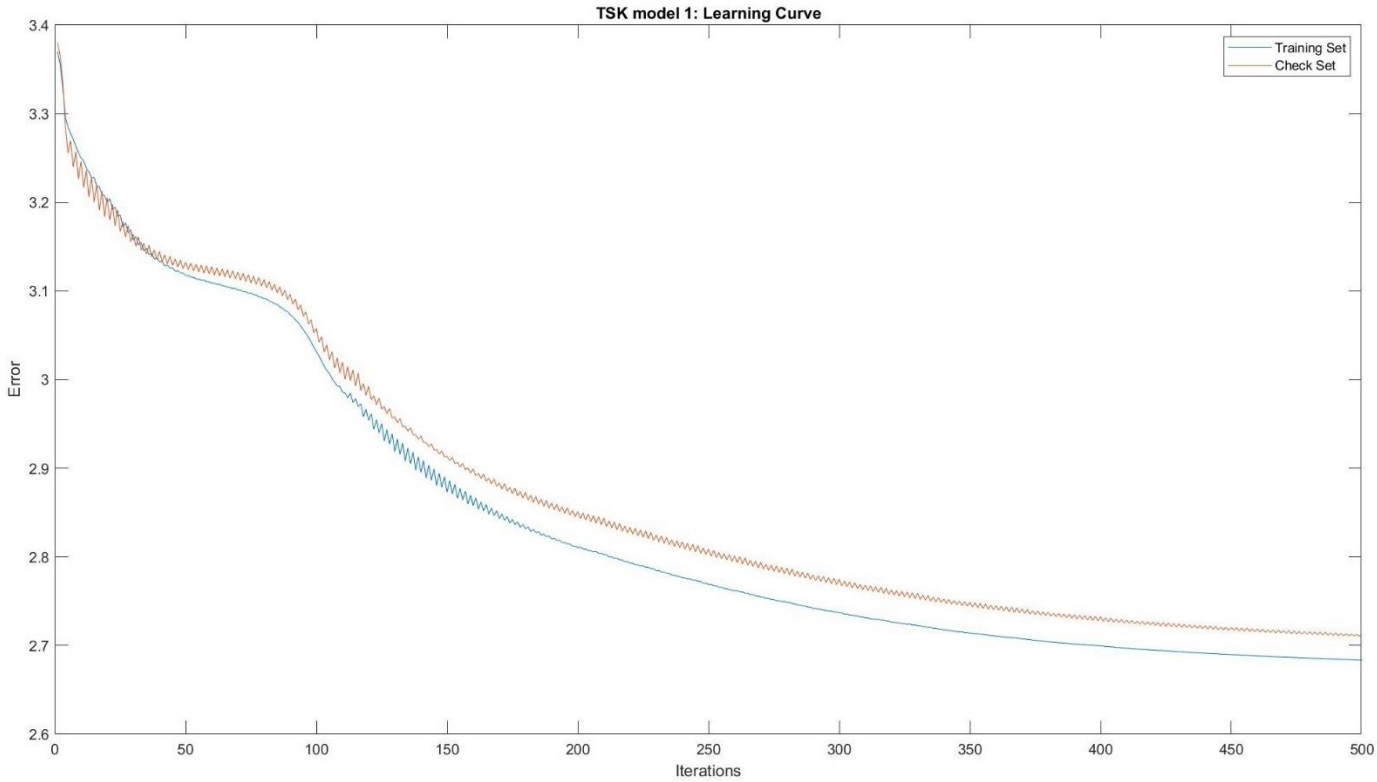

*Σχήμα 2: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 1*

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



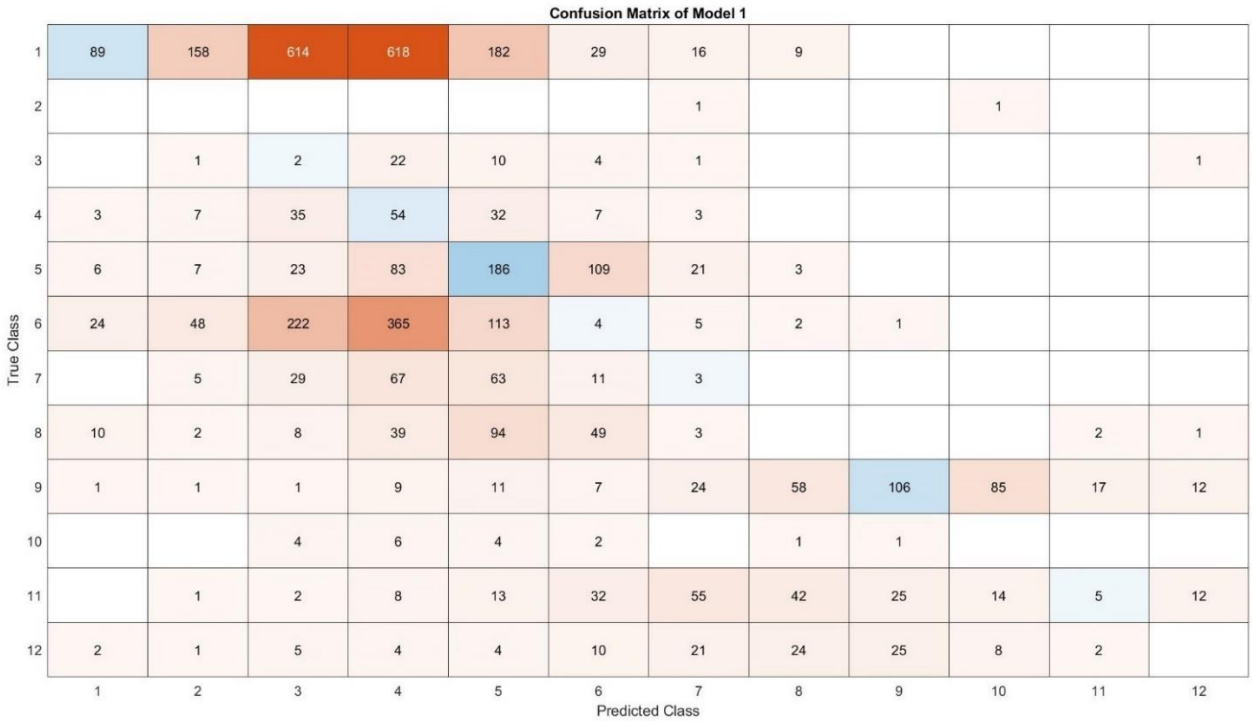

*Σχήμα 3: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 1*

Ακολουθούν οι καμπύλες εκμάθησης με βάση το RMSE στο πέρας των εποχών.

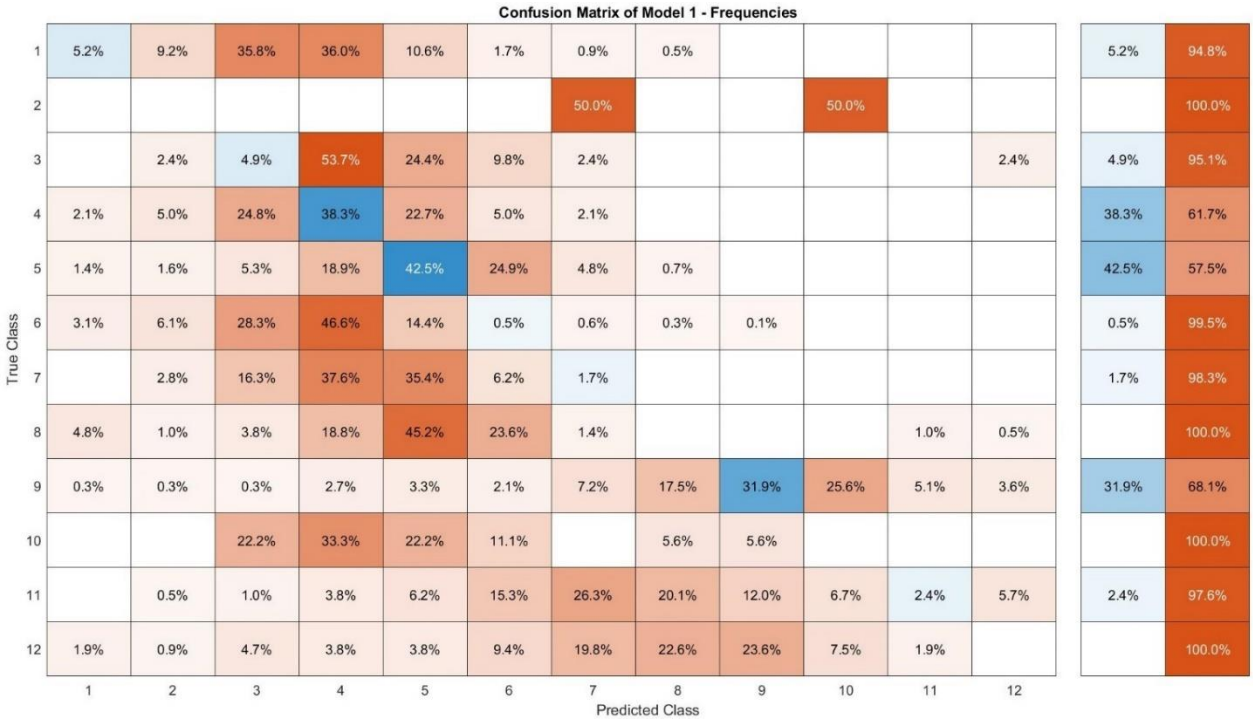



*Σχήμα 4: Καμπύλες Εκμάθησης - TSK Μοντέλο 1*

Ακόμα, βλέπουμε τους Πίνακες Σφαλμάτων Ταξινόμησης με τη μορφή απολύτων τιμών αλλά και ποσοστιαία καθώς και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.




*Σχήμα 5: Πίνακας Σφαλμάτων Ταξινόμησης - TSK Μοντέλο 1*



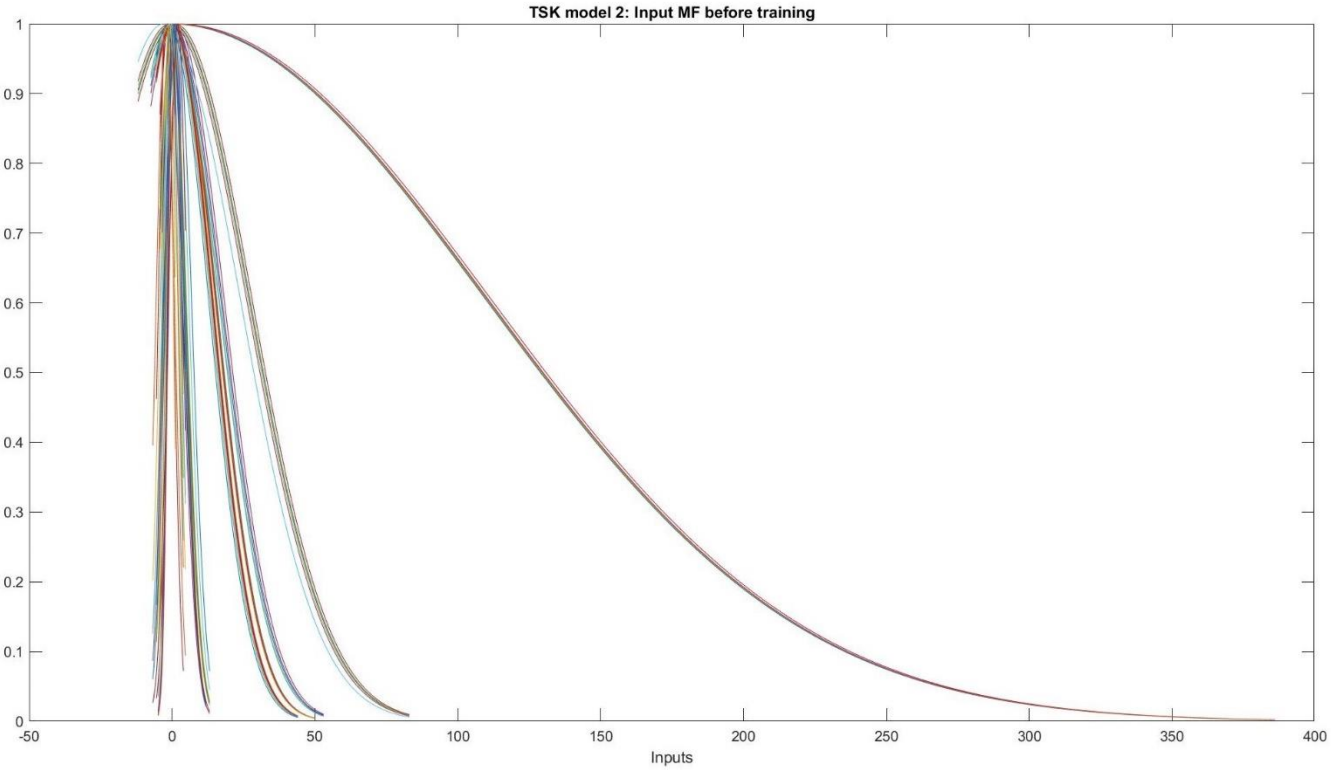

*Σχήμα 6: Πίνακας Σφαλμάτων Ταξινόμησης (Συχνότητες) - TSK Μοντέλο 1*

Τέλος, οι ζητούμενοι δείκτες απόδοσης για το μοντέλο αυτό φαίνονται παρακάτω.

| *Class Number* | Producers Accuracy | Users Accuracy | *Class Number* | Producers Accuracy | Users Accuracy | *Overall* Accuracy |
|---|---|---|---|---|---|---|
| 1 | 0.6593 | 0.0519 | 7 | 0.0196 | 0.0169 | 0.1076 |
| 2 | 0 | 0 | 8 | 0 | 0 | |
| 3 | 0.0021 | 0.0488 | 9 | 0.6709 | 0.3193 | $\hat{k}$ |
| 4 | 0.0424 | 0.3830 | 10 | 0 | 0 | |
| 5 | 0.2612 | 0.4247 | 11 | 0.1923 | 0.0239 | 0.0481 |
| 6 | 0.0152 | 0.0051 | 12 | 0 | 0 | |

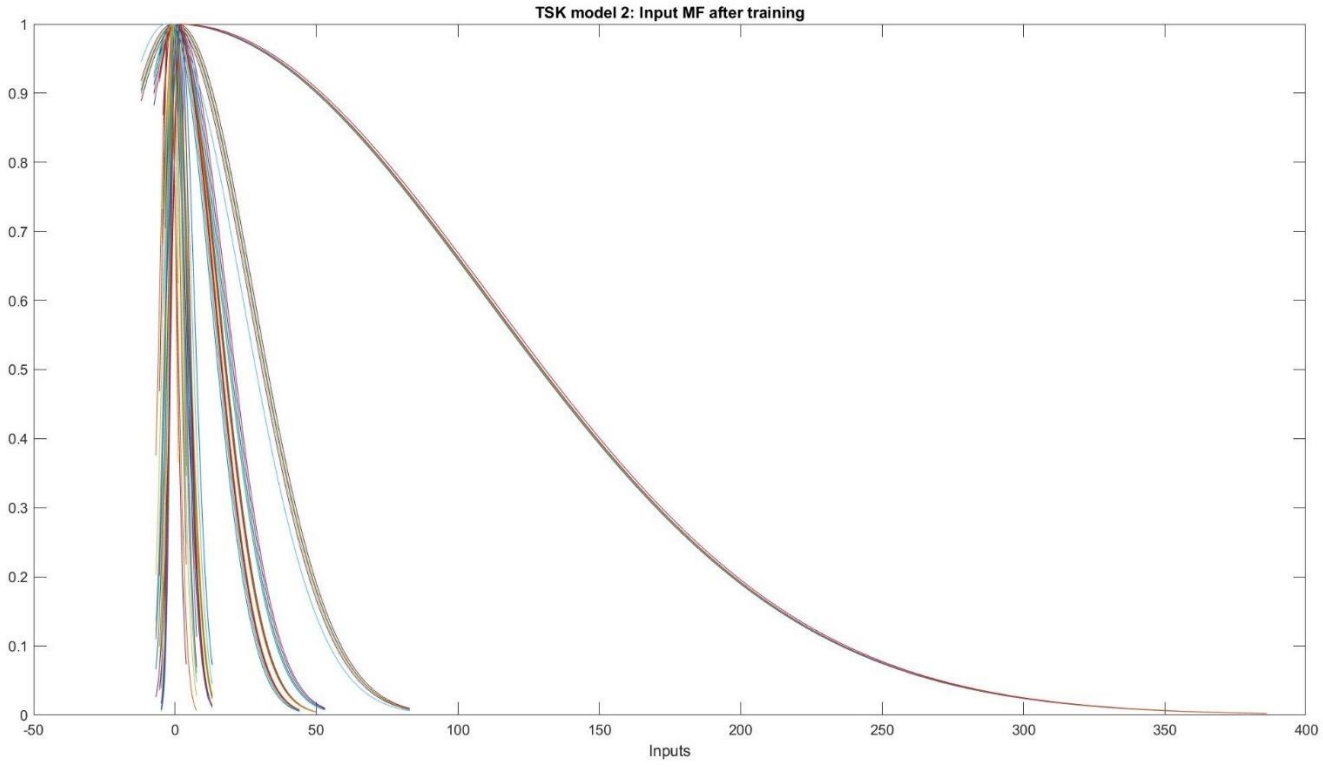
## TSK Μοντέλο 2 (8 Rules)

Οι συναρτήσεις συμμετοχής για το δεύτερο μοντέλο πριν από τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 7.



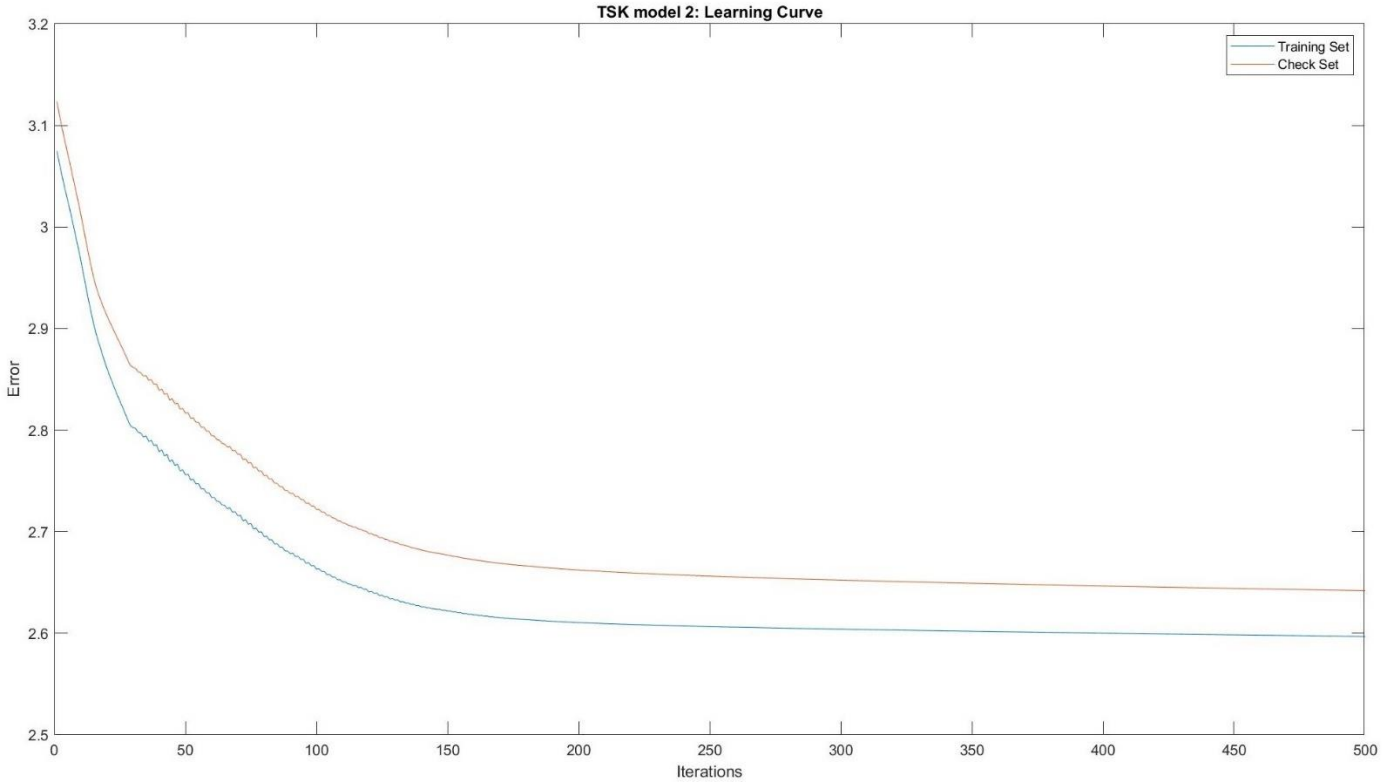

*Σχήμα 7: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 2*

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



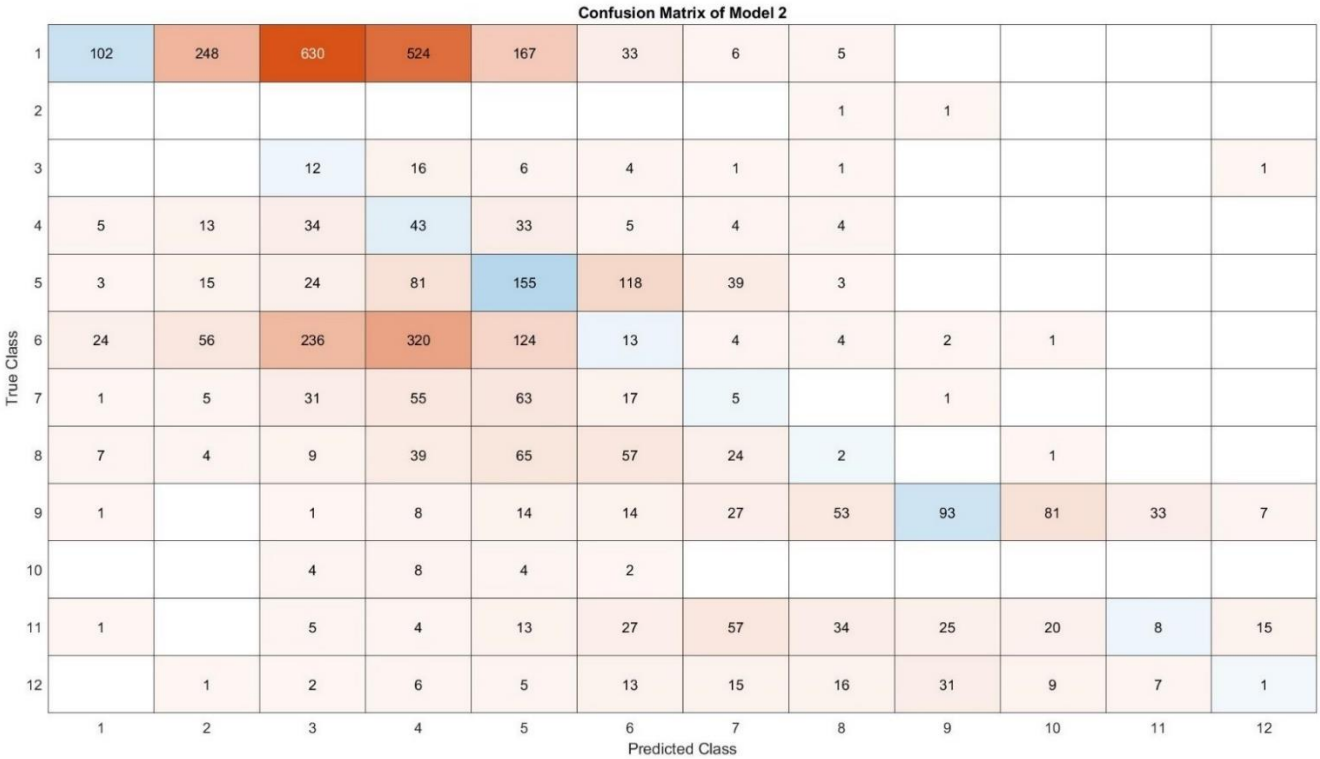

*Σχήμα 8: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 2*

Ακολουθούν οι καμπύλες εκμάθησης με βάση το RMSE στο πέρας των εποχών.

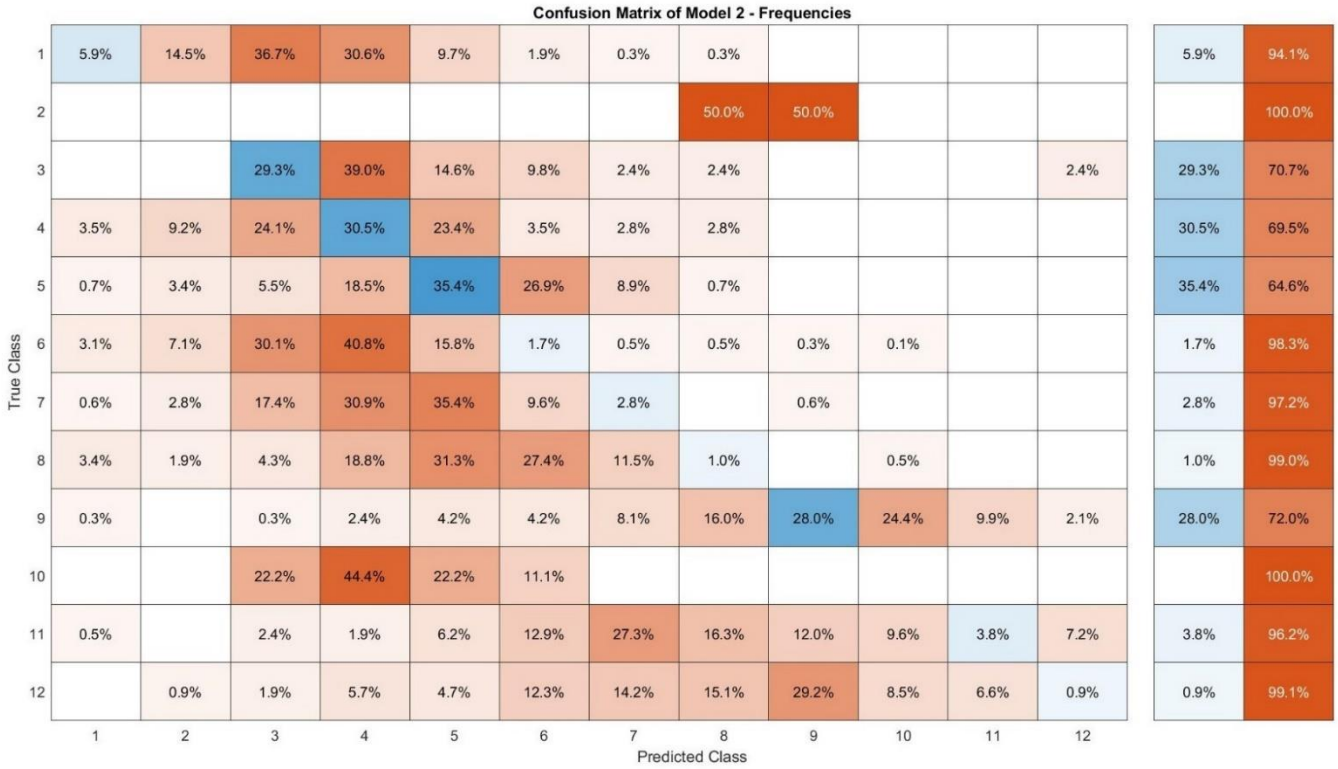



*Σχήμα 9: Καμπύλες Εκμάθησης - TSK Μοντέλο 2*

Ακόμα, βλέπουμε τους Πίνακες Σφαλμάτων Ταξινόμησης με τη μορφή απολύτων τιμών αλλά και ποσοστιαία καθώς και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.




*Σχήμα 10: Πίνακες Σφαλμάτων Ταξινόμησης - TSK Μοντέλο 2*



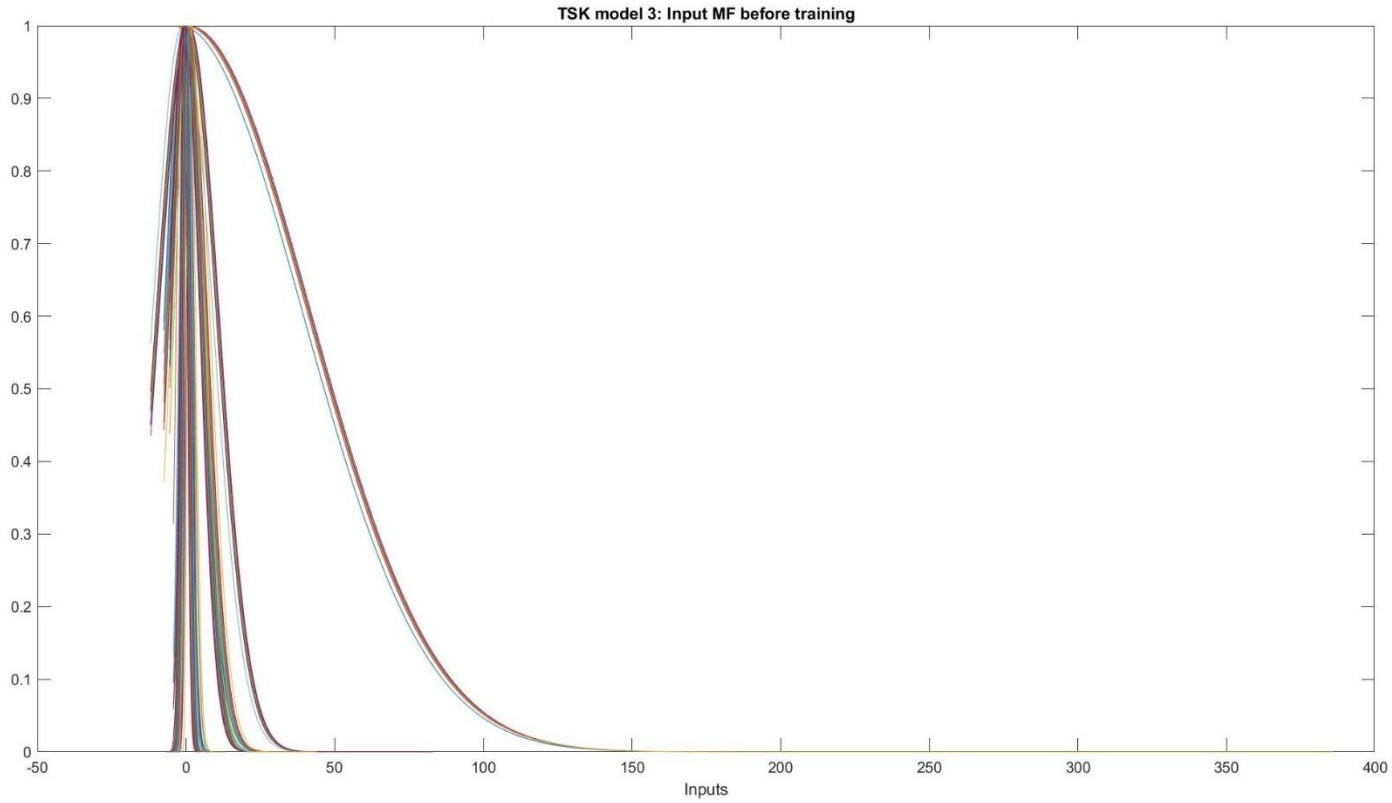

*Σχήμα 11: Πίνακες Σφαλμάτων Ταξινόμησης (Συχνότητες) - TSK Μοντέλο 2*

Τέλος, οι ζητούμενοι δείκτες απόδοσης για το μοντέλο αυτό φαίνονται παρακάτω.

| *Class Number* | Producers Accuracy | Users Accuracy | *Class Number* | Producers Accuracy | Users Accuracy | *Overall* Accuracy |
|---|---|---|---|---|---|---|
| 1 | 0.7083 | 0.0595 | 7 | 0.0275 | 0.0281 | 0.1040 |
| 2 | 0 | 0 | 8 | 0.0163 | 0.0096 | |
| 3 | 0.0121 | 0.2927 | 9 | 0.6078 | 0.2801 | $\hat{k}$ |
| 4 | 0.0389 | 0.3050 | 10 | 0 | 0 | |
| 5 | 0.2388 | 0.3539 | 11 | 0.1667 | 0.0383 | 0.0442 |
| 6 | 0.0429 | 0.0166 | 12 | 0.0417 | 0.0094 | |

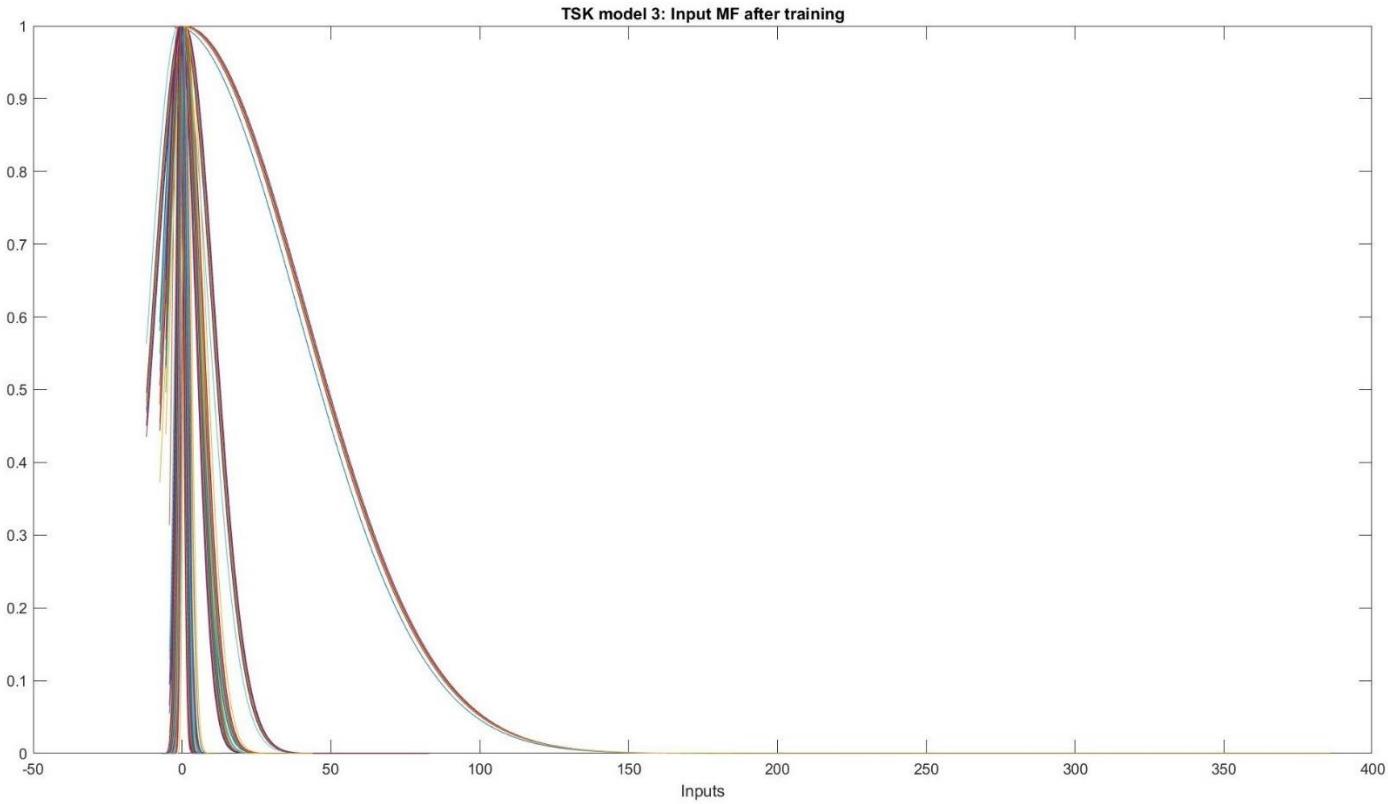
## TSK Μοντέλο 3 (12 Rules)

Οι συναρτήσεις συμμετοχής για το τρίτο μοντέλο πριν από τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 12.



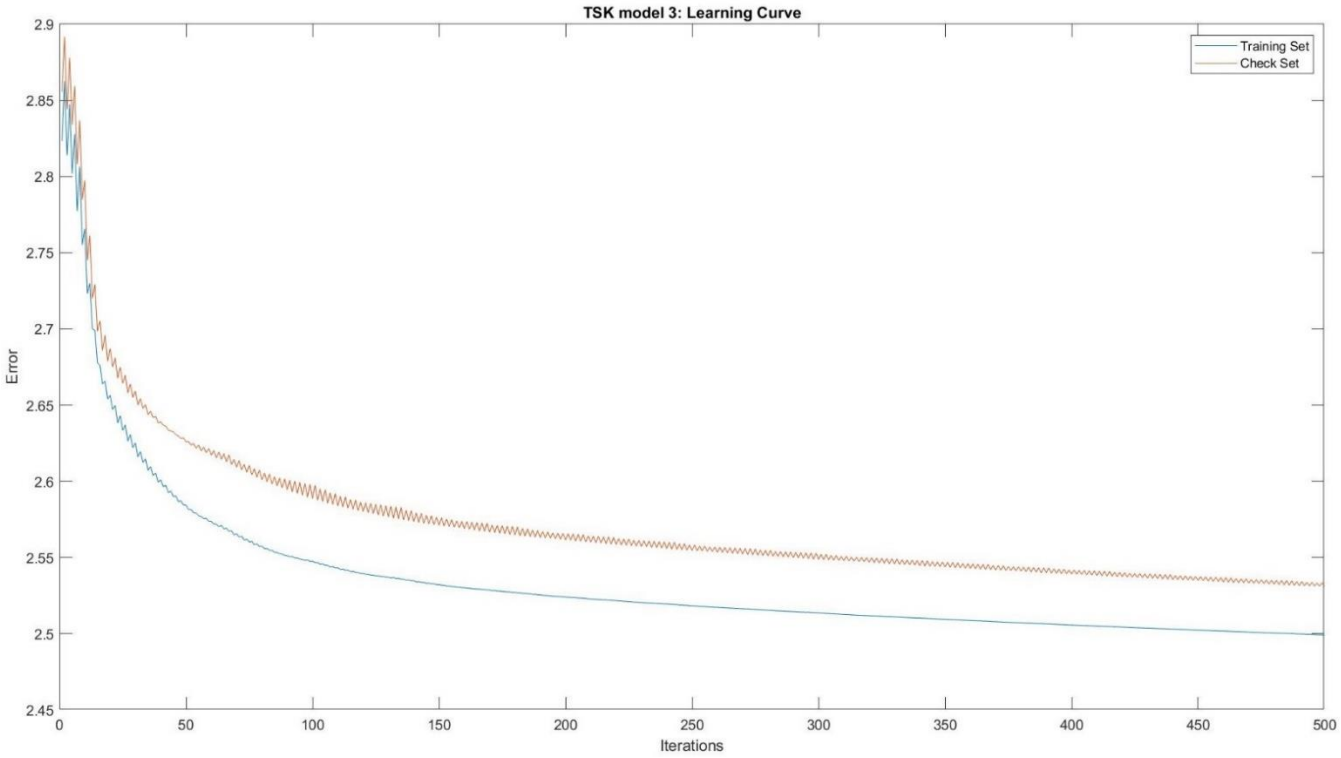

*Σχήμα 12: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 3*

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



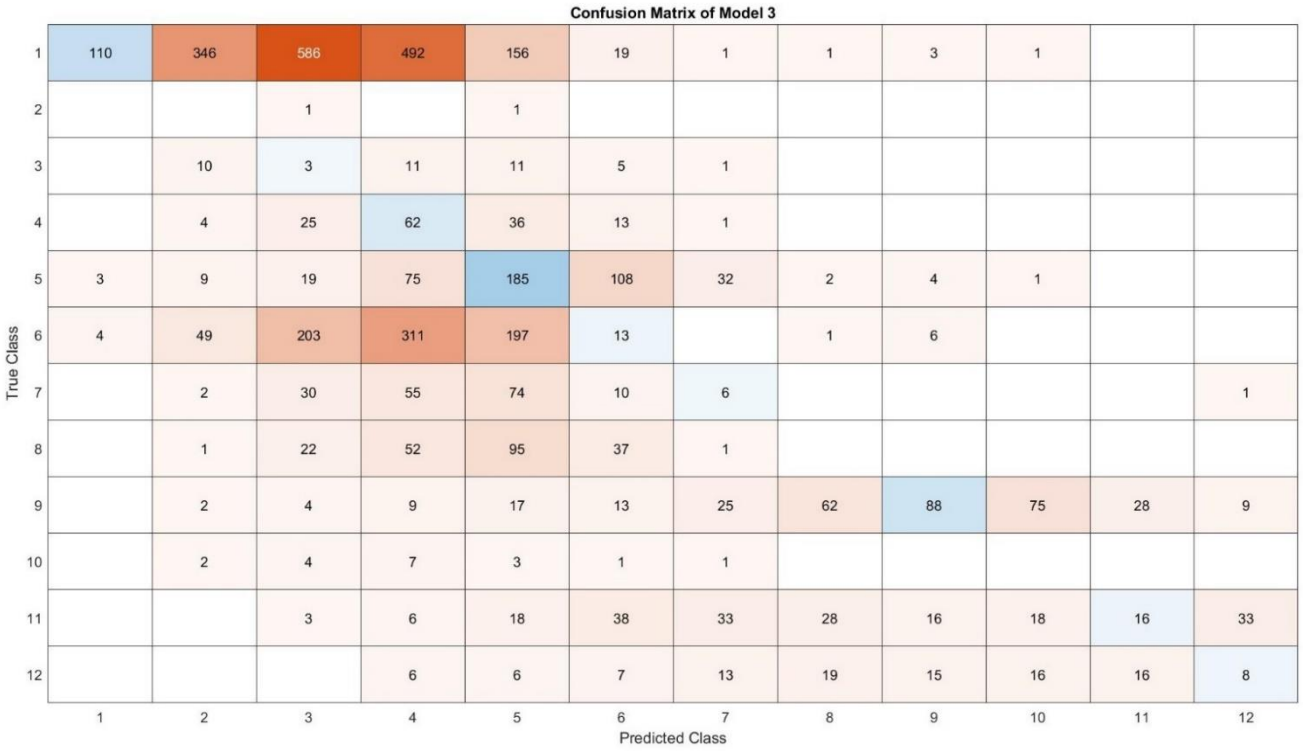

*Σχήμα 13: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 3*

Ακολουθούν οι καμπύλες εκμάθησης με βάση το RMSE στο πέρας των εποχών.

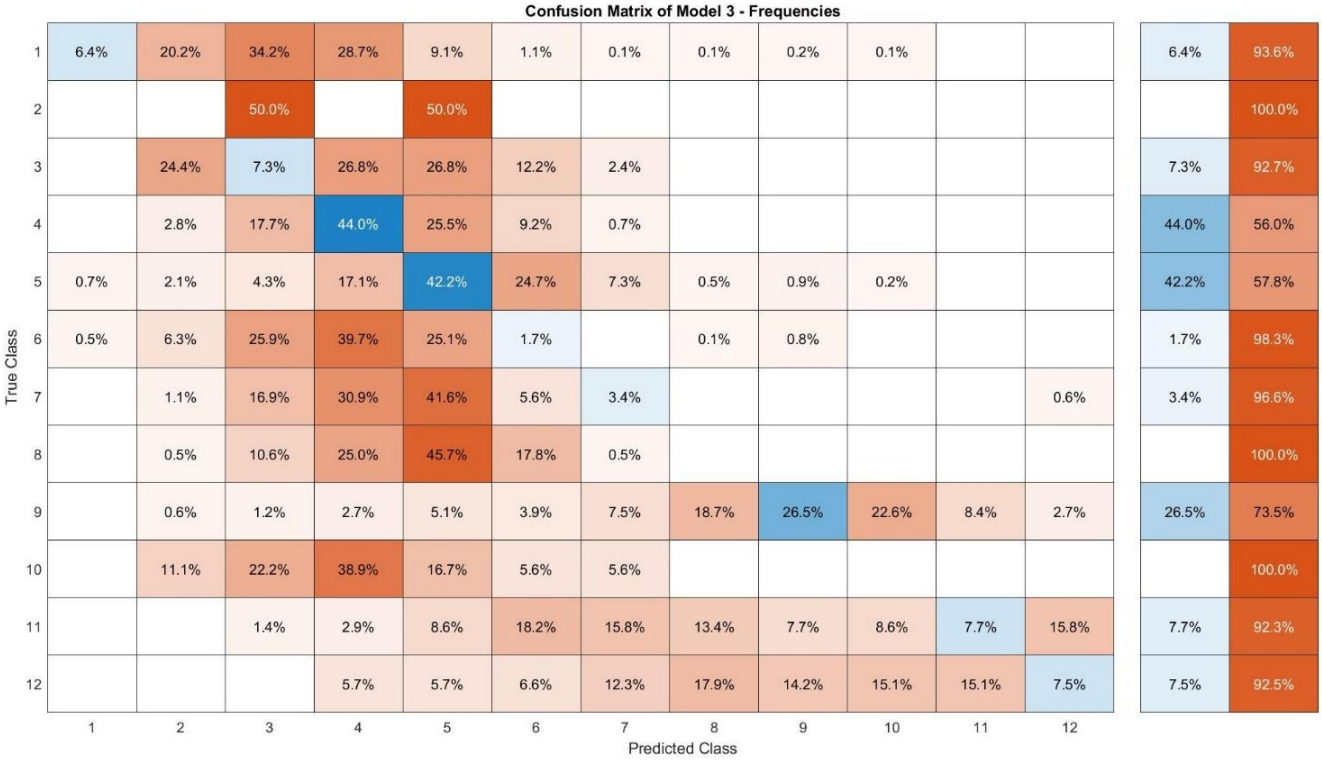



*Σχήμα 14: Καμπύλες Εκμάθησης - TSK Μοντέλο 3*

Ακόμα, βλέπουμε τους Πίνακες Σφαλμάτων Ταξινόμησης με τη μορφή απολύτων τιμών αλλά και ποσοστιαία καθώς και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.




*Σχήμα 15: Πίνακες Σφαλμάτων Ταξινόμησης - TSK Μοντέλο 3*



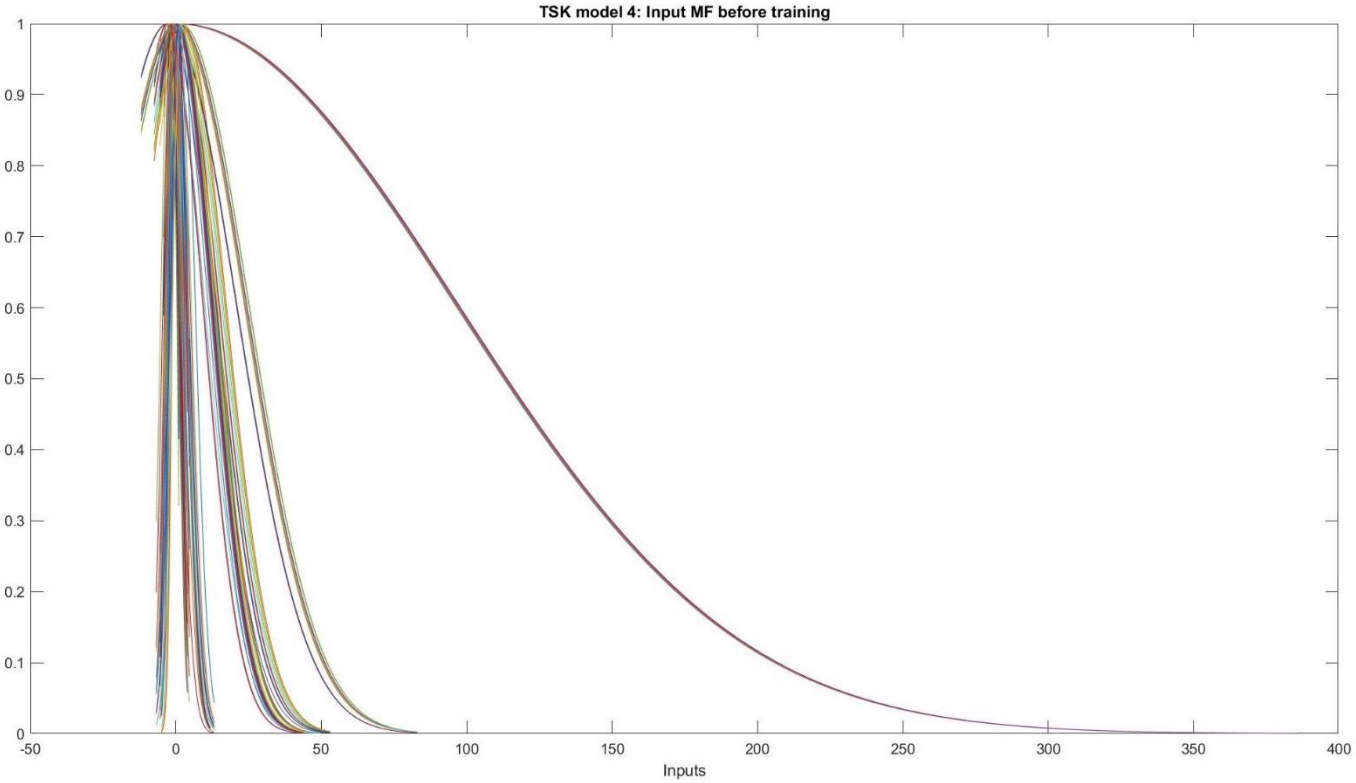

*Σχήμα 16: Πίνακες Σφαλμάτων Ταξινόμησης (Συχνότητες) - TSK Μοντέλο 3*

Τέλος, οι ζητούμενοι δείκτες απόδοσης για το μοντέλο αυτό φαίνονται παρακάτω.

| *Class Number* | Producers Accuracy | Users Accuracy | *Class Number* | Producers Accuracy | Users Accuracy | *Overall* Accuracy |
|---|---|---|---|---|---|---|
| 1 | 0.9402 | 0.0641 | 7 | 0.0526 | 0.0337 | 0.1177 |
| 2 | 0 | 0 | 8 | 0 | 0 | |
| 3 | 0.0033 | 0.0732 | 9 | 0.6667 | 0.2651 | $\hat{k}$ |
| 4 | 0.0571 | 0.4397 | 10 | 0 | 0 | |
| 5 | 0.2315 | 0.4224 | 11 | 0.2667 | 0.0766 | 0.0607 |
| 6 | 0.0492 | 0.0166 | 12 | 0.1569 | 0.0755 | |

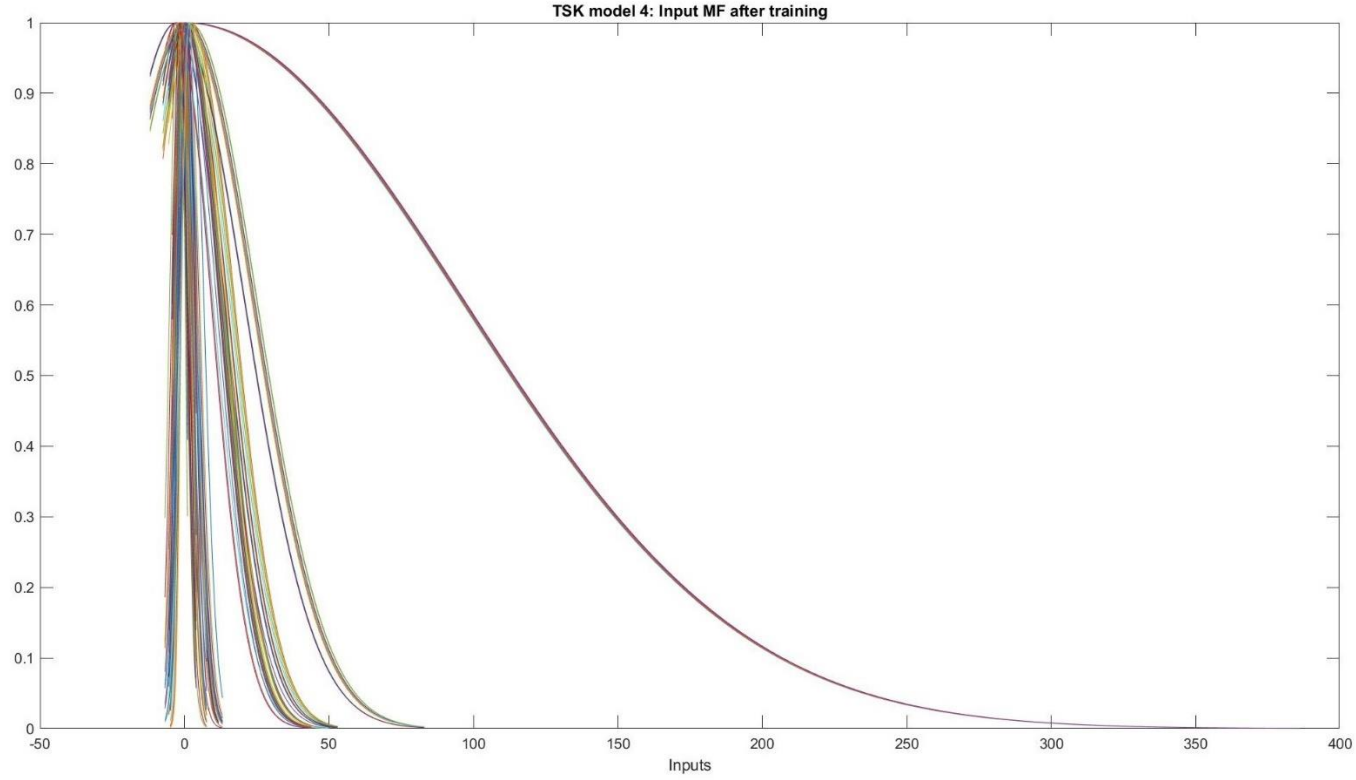
## TSK Μοντέλο 4 (16 Rules)

Οι συναρτήσεις συμμετοχής για το τέταρτο μοντέλο πριν από τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 17.



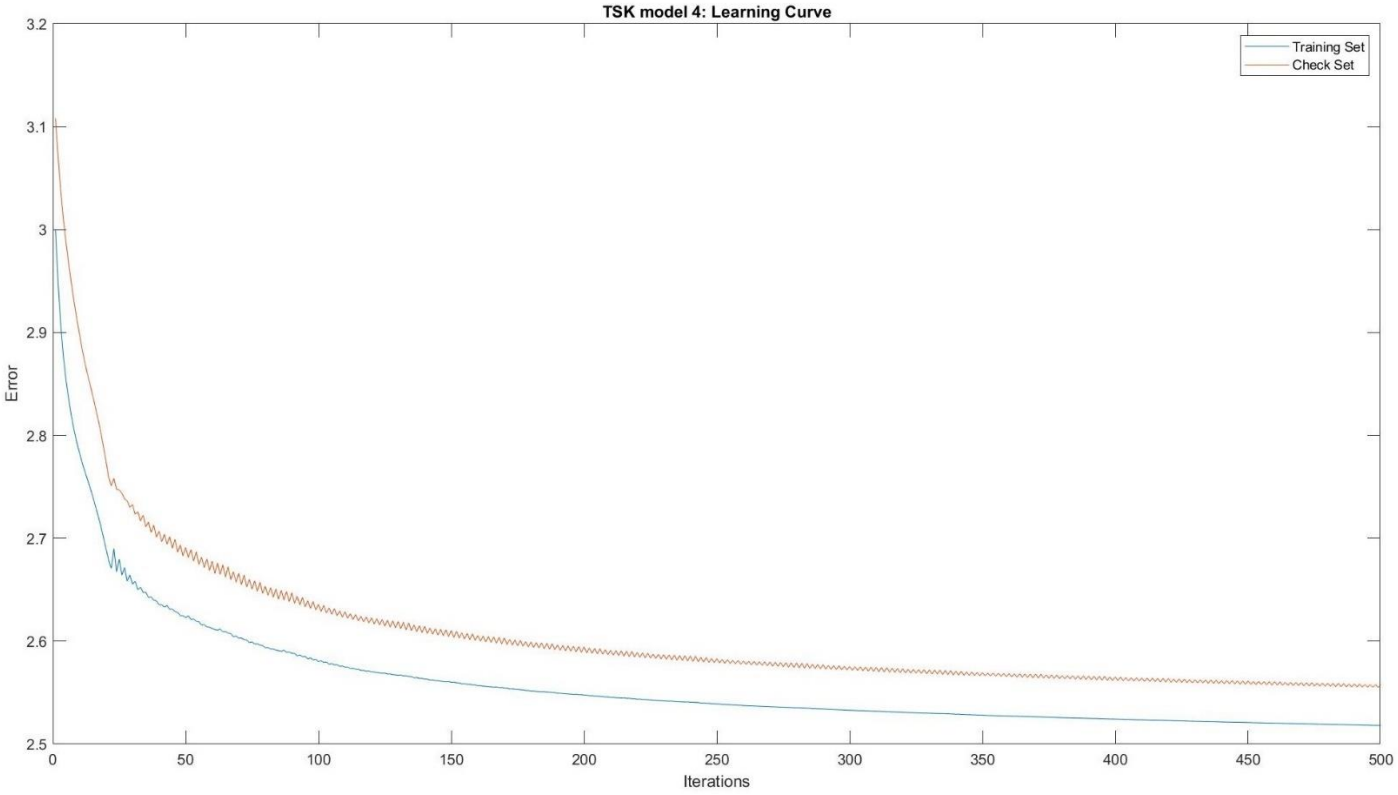

*Σχήμα 17: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 4*

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



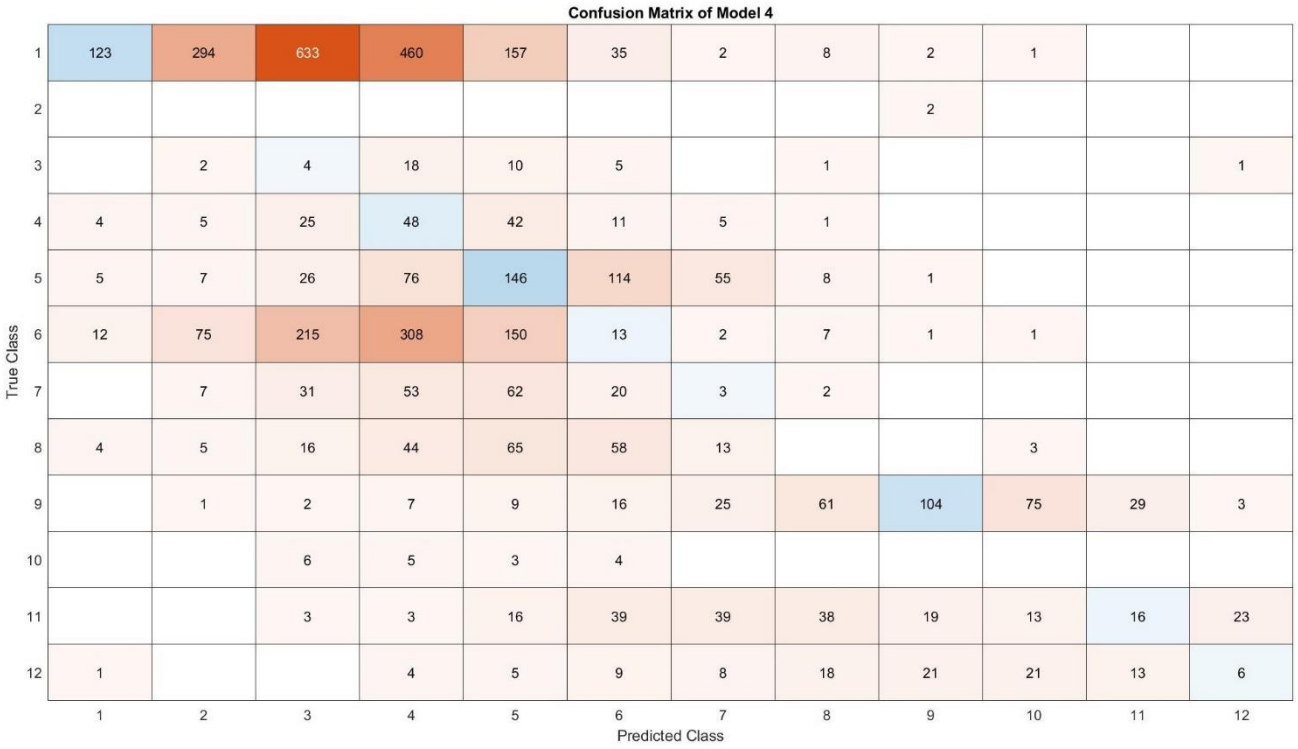

*Σχήμα 18: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 4*

Ακολουθούν οι καμπύλες εκμάθησης με βάση το RMSE στο πέρας των εποχών.

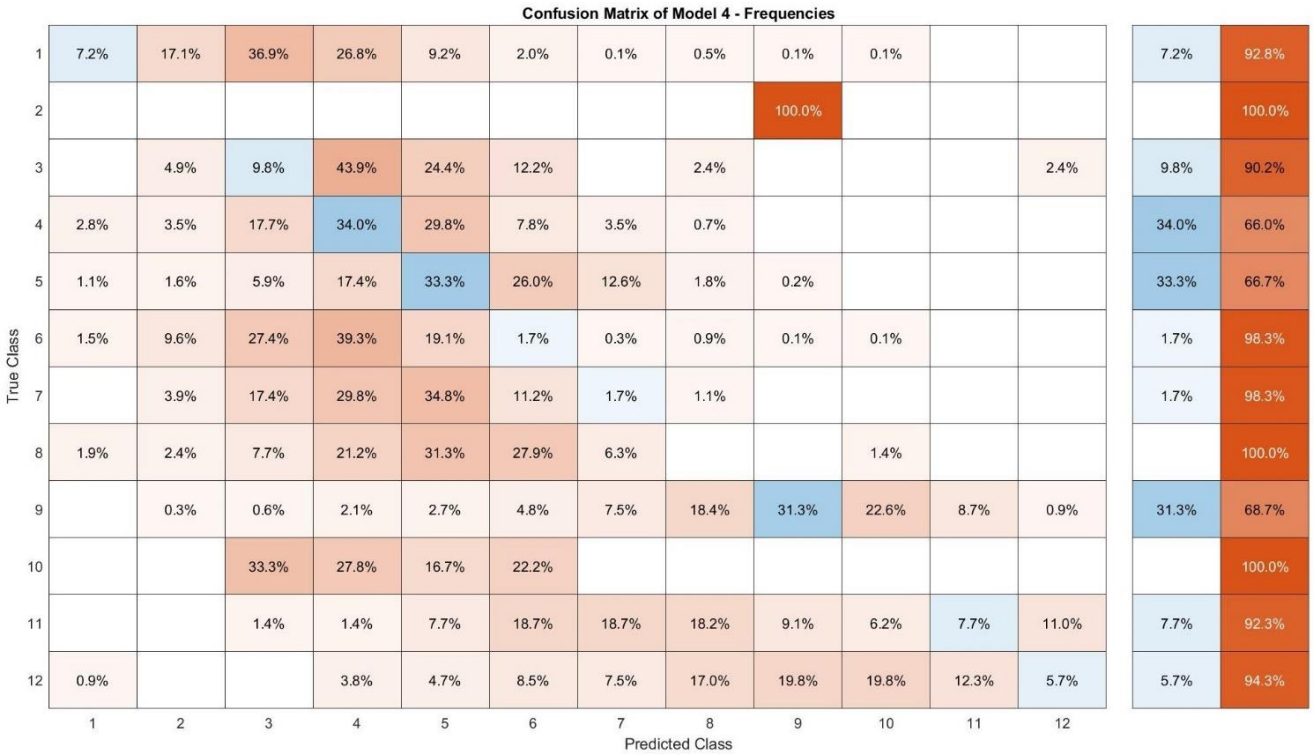



*Σχήμα 19: Καμπύλες Εκμάθησης - TSK Μοντέλο 4*

Ακόμα, βλέπουμε τους Πίνακες Σφαλμάτων Ταξινόμησης με τη μορφή απολύτων τιμών αλλά και ποσοστιαία καθώς και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.




*Σχήμα 20: Πίνακες Σφαλμάτων Ταξινόμησης - TSK Μοντέλο 4*



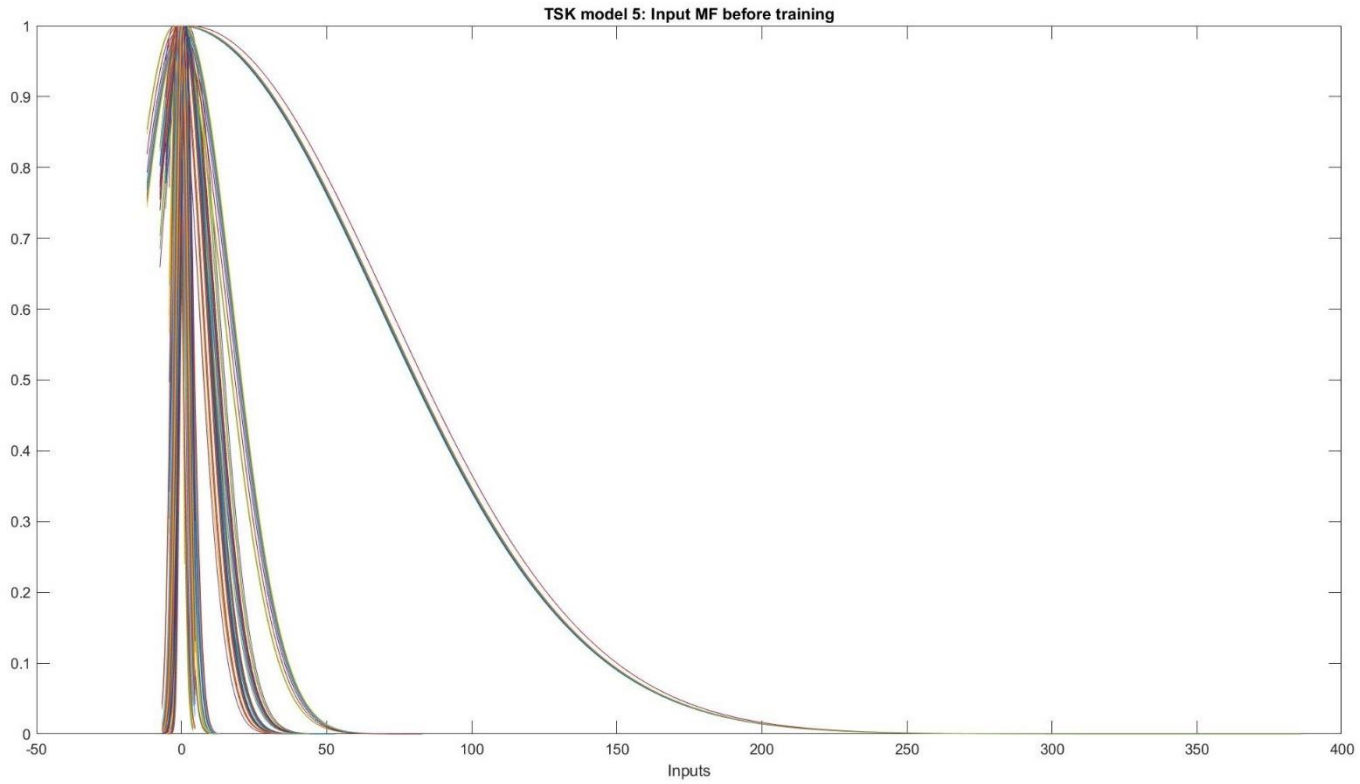

*Σχήμα 21: Πίνακες Σφαλμάτων Ταξινόμησης (Συχνότητες) - TSK Μοντέλο 4*

Τέλος, οι ζητούμενοι δείκτες απόδοσης για το μοντέλο αυτό φαίνονται παρακάτω.

| *Class Number* | Producers Accuracy | Users Accuracy | *Class Number* | Producers Accuracy | Users Accuracy | *Overall* Accuracy |
|---|---|---|---|---|---|---|
| 1 | 0.8255 | 0.0717 | 7 | 0.0197 | 0.0169 | 0.1110 |
| 2 | 0 | 0 | 8 | 0 | 0 | |
| 3 | 0.0042 | 0.0976 | 9 | 0.6933 | 0.3133 | $\hat{k}$ |
| 4 | 0.0468 | 0.3404 | 10 | 0 | 0 | |
| 5 | 0.2195 | 0.3333 | 11 | 0.2759 | 0.0766 | 0.0504 |
| 6 | 0.0401 | 0.0166 | 12 | 0.1818 | 0.0566 | |

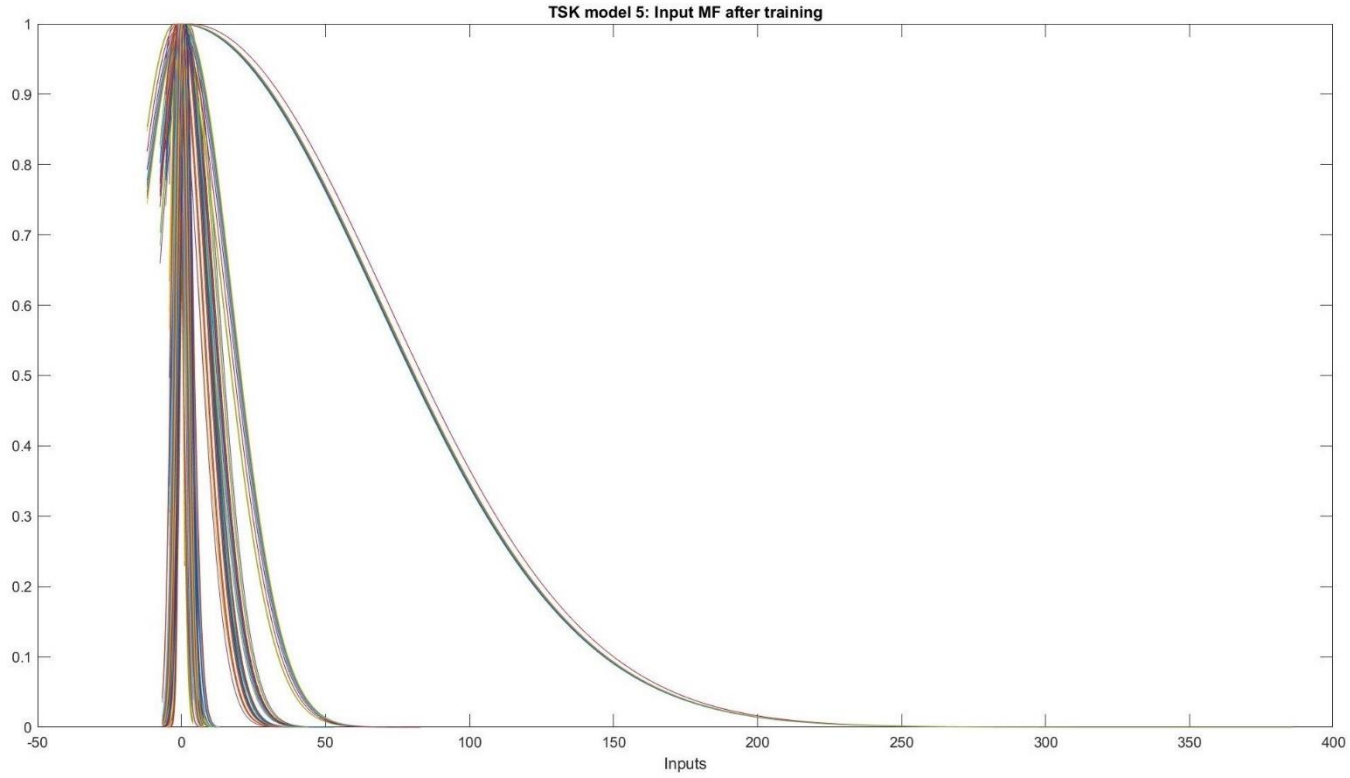
## TSK Μοντέλο 5 (20 Rules)

Οι συναρτήσεις συμμετοχής για το πέμπτο μοντέλο πριν από τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 22.



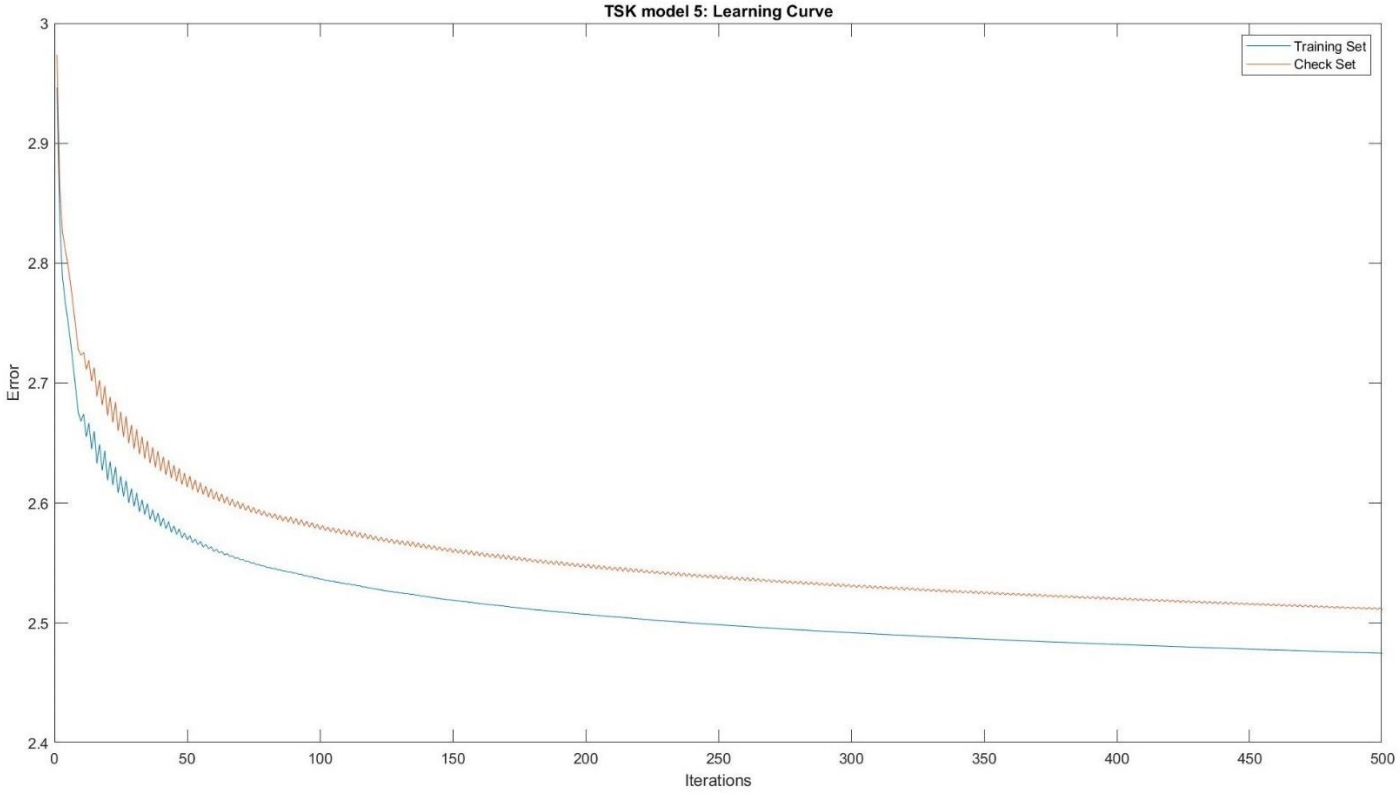

*Σχήμα 22: Αρχικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 5*

Τα αποτελέσματα της παραπάνω διαδικασίας φαίνονται στη συνέχεια. Αρχικά βλέπουμε τη μορφή των συναρτήσεων συμμετοχής του μοντέλου μετά την εκπαίδευση.



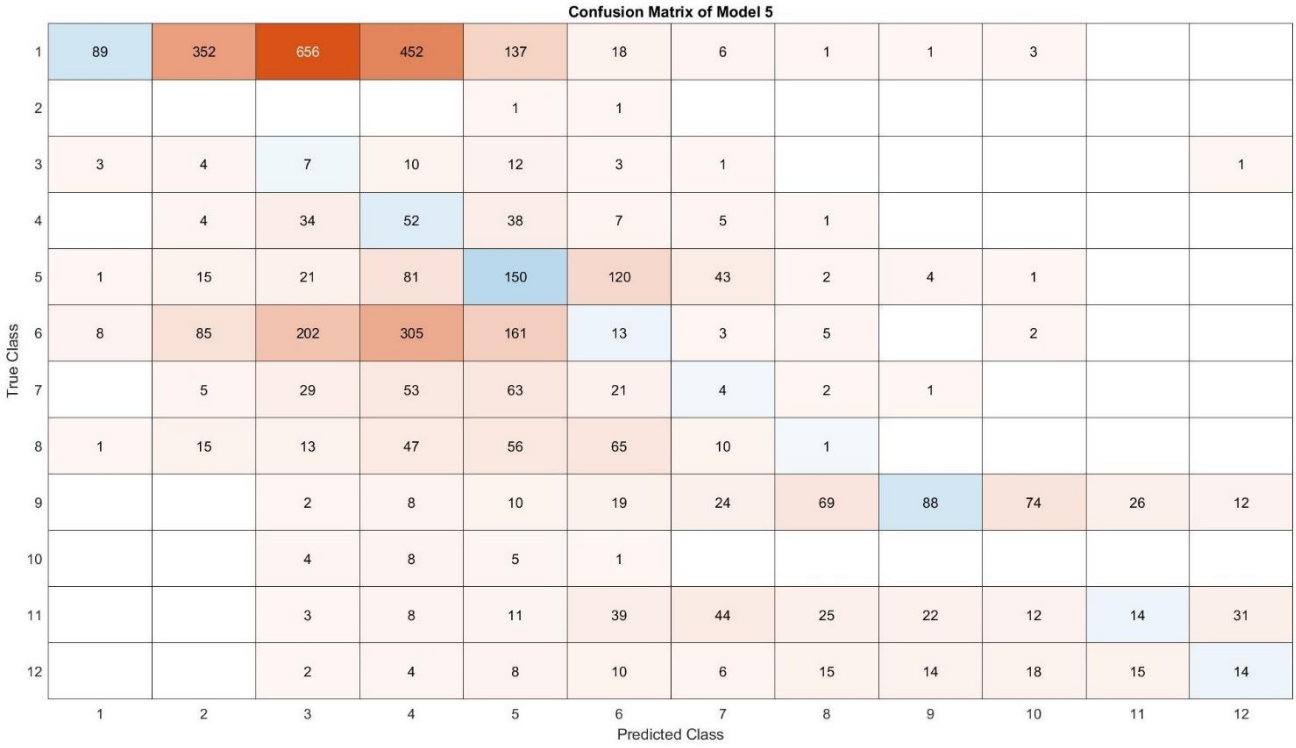

*Σχήμα 23: Τελικές Συναρτήσεις Συμμετοχής - TSK Μοντέλο 5*

Ακολουθούν οι καμπύλες εκμάθησης με βάση το RMSE στο πέρας των εποχών.

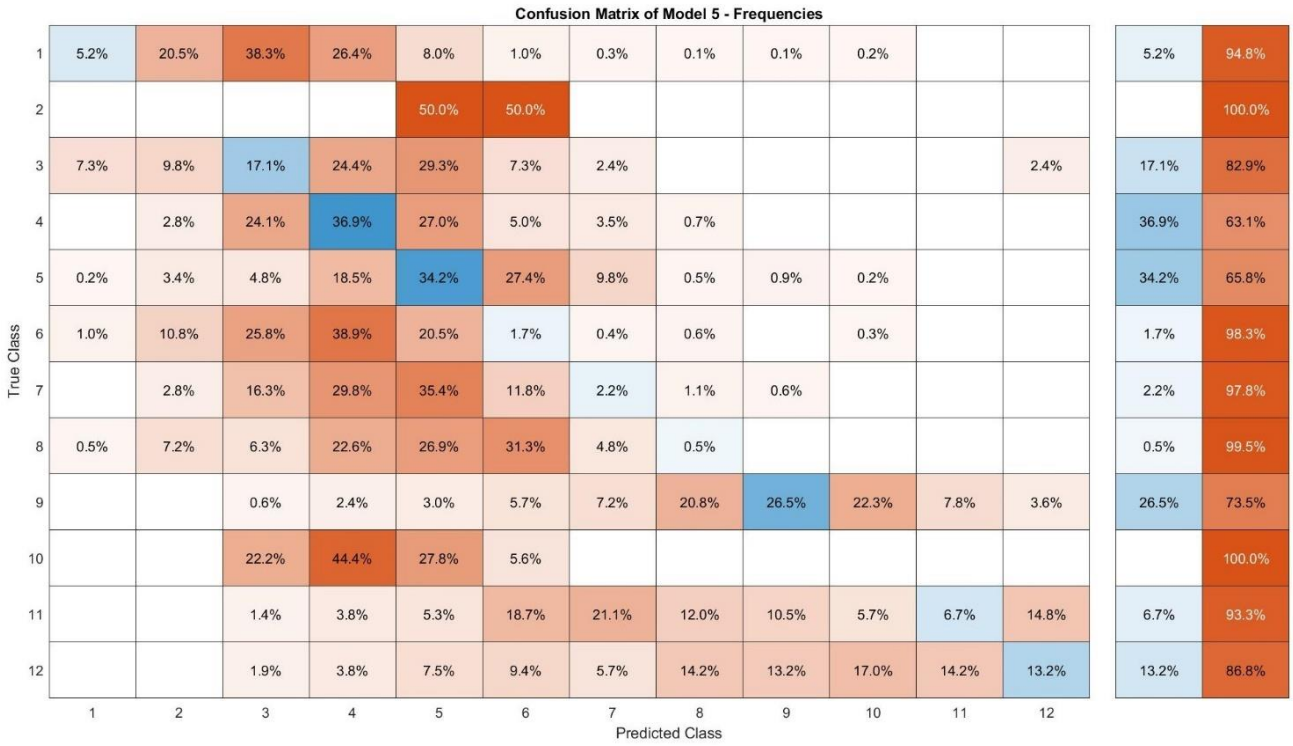



*Σχήμα 24: Καμπύλες Εκμάθησης - TSK Μοντέλο 5*

Ακόμα, βλέπουμε τους Πίνακες Σφαλμάτων Ταξινόμησης με τη μορφή απολύτων τιμών αλλά και ποσοστιαία καθώς και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.




*Σχήμα 25: Πίνακες Σφαλμάτων Ταξινόμησης - TSK Μοντέλο 5*




*Σχήμα 26: Πίνακες Σφαλμάτων Ταξινόμησης (Συχνότητες) - TSK Μοντέλο 5*

Τέλος, οι ζητούμενοι δείκτες απόδοσης για το μοντέλο αυτό φαίνονται παρακάτω.

| *Class Number* | Producers Accuracy | Users Accuracy | *Class Number* | Producers Accuracy | Users Accuracy | *Overall* Accuracy |
|---|---|---|---|---|---|---|
| 1 | 0.8725 | 0.0519 | 7 | 0.0274 | 0.0225 | 0.1035 |
| 2 | 0 | 0 | 8 | 0.0083 | 0.0048 | |
| 3 | 0.0072 | 0.1707 | 9 | 0.6769 | 0.2651 | $\hat{k}$ |
| 4 | 0.0506 | 0.3688 | 10 | 0 | 0 | |
| 5 | 0.2301 | 0.3425 | 11 | 0.2545 | 0.0670 | 0.0484 |
| 6 | 0.0410 | 0.0166 | 12 | 0.2414 | 0.1321 | |

## Σχολιασμός Αποτελεσμάτων και Συμπεράσματα

Από τα παραπάνω βλέπουμε ότι το μοντέλο με τη μεγαλύτερη ακρίβεια είναι αυτό με τους 12 κανόνες με Overall Accuracy ίσο με 11.77% και δείκτη $\hat{k}$ ίσο με 0.0607. Είναι προφανές ότι τα αποτελέσματα από την εκτέλεση του αλγορίθμου δεν είναι ιδιαίτερα ικανοποιητικά κάτι που μπορεί να οφείλεται σε διάφορους παράγοντες.

Ο βασικότερος παράγοντας αφορά τα δεδομένα με τα οποία εκπαιδεύουμε τα ζητούμενα μοντέλα. Συγκεκριμένα, το Dataset που έχουμε στη διάθεσή μας παρουσιάζει το φαινόμενο «Class

Imbalance» που σημαίνει ότι τα πλήθος των δεδομένων να μην είναι «δίκαια» μοιρασμένο σε κάθε κλάση όπως φαίνεται ξεκάθαρα από την παρακάτω εικόνα. Το πρόβλημα αυτό θα μπορούσε πιθανώς να επιλυθεί με χρήση της τεχνικής Oversampling (ή Undersampling σε άλλη περίσταση) ωστόσο δεν χρησιμοποιήθηκε στη συγκεκριμένη εκπαίδευση καθώς ο χρόνος εκτέλεσης του αλγορίθμου θα αυξανόταν σημαντικά.

Χρησιμοποιώντας την εντολή tabulate του MATLAB παρατηρούμε ότι πάνω από το 41% των δεδομένων που έχουμε στη διάθεσή μας ανήκουν στην πρώτη κλάση ενώ ταυτόχρονα λιγότερο από το 0.05% των δεδομένων ανήκει στη δεύτερη.

| Output_Values | Avila_Set | Training_Set | Validation_Set | Check_Set |
|---|---|---|---|---|
| 1 | 41.0792% | 41.0783% | 41.0539% | 41.1074% |
| 2 | 0.0479226% | 0.0479233% | 0.0479042% | 0.0479386% |
| 3 | 0.987205% | 0.990415% | 0.982036% | 0.982742% |
| 4 | 3.37854% | 3.37859% | 3.37725% | 3.37967% |
| 5 | 10.495% | 10.4952% | 10.491% | 10.4986% |
| 6 | 18.8% | 18.8019% | 18.8024% | 18.7919% |
| 7 | 4.27948% | 4.28115% | 4.28743% | 4.26654% |
| 8 | 4.97915% | 4.97604% | 4.98204% | 4.98562% |
| 9 | 7.96952% | 7.97125% | 7.97605% | 7.95781% |
| 10 | 0.426511% | 0.423323% | 0.431138% | 0.431448% |
| 11 | 5.00311% | 5% | 5.00599% | 5.00959% |
| 12 | 2.55427% | 2.55591% | 2.56287% | 2.54075% |

*Σχήμα 27: Διαμοιρασμός Δεδομένων στα διάφορα Σετ*

Επίσης, πέρα από την ανισορροπία που αναφέρθηκε το μεγάλο πλήθος των κλάσεων και μικρό απόλυτο πλήθος δεδομένων για ορισμένες κλάσεις καθιστά την εκπαίδευση ακόμα δυσκολότερη. Για παράδειγμα, το dataset περιέχει συνολικά μόνο δέκα καταχωρήσεις που αφορούν τη δεύτερη κλάση (σε αντίθεση με την πρώτη κλάση που περιέχει 8572). Είναι επομένως αναμενόμενο να παρουσιάζονται δυσκολίες κατά την εκπαίδευση των μοντέλων και τη ρύθμιση των βαρών, ώστε να προσαρμοστούν κατάλληλα και να μπορούν να ταξινομούν σωστά τα δεδομένα που ανήκουν στη δεύτερη κλάση.

Ένας ακόμα παράγοντας είναι η διαθέσιμη υπολογιστική ισχύς και ο περιορισμός, στο πλαίσιο της εργασίας, στη χρήση του αλγορίθμου Anfis. Με διαφορετική επιλογή αλγορίθμου εκπαίδευσης και με τη διάθεση περισσότερων υπολογιστικών πόρων-ισχύος είναι πιθανό να πετύχουμε αρκετά καλύτερα αποτελέσματα, κάτι που ξεφεύγει, ωστόσο, από το σκοπό της παρούσας εργασίας.

Τέλος, αξίζει να σημειωθεί ότι για την εκπαίδευση των πέντε παραπάνω μοντέλων χρειάστηκαν συνολικά 1603.12 δευτερόλεπτα, δηλαδή σχεδόν 27 λεπτά.

# Εφαρμογή στο Σετ Δεδομένων Isolet

## Αντιμετώπιση σετ δεδομένων υψηλής διαστασιμότητας

Το Isolet Dataset πρόκειται για ένα πολύ μεγαλύτερο σετ δεδομένων σε σχέση με το Avila, καθώς περιέχει 617 διαφορετικά χαρακτηριστικά και 7797 δεδομένα. Στόχος του τμήματος της εργασίας αυτού είναι η ορθή ταξινόμηση των δειγμάτων με βάση τα χαρακτηριστικά αυτά. Ο μεγάλος όγκος των χαρακτηριστικών και δεδομένων καθιστά την εκπαίδευση του ζητούμενου μοντέλου πρακτικά ανέφικτη, καθώς ο χρόνος που απαιτείται για αυτήν είναι υπερβολικά μεγάλος. Για το λόγο αυτό, θα χρειαστεί να επιλέξουμε ένα αρκετά πιο περιορισμένο πλήθος χαρακτηριστικών, και συγκεκριμένα τα πιο αντιπροσωπευτικά του δείγματος, η επιλογή των οποίων γίνεται με χρήση του αλγορίθμου Relief.

## Εύρεση Πλήθους Χαρακτηριστικών και Κανόνων για βέλτιστη Μοντελοποίηση

Αρχικά ταξινομούμε κατά αύξουσα σειρά το σετ δεδομένων με βάση τη στήλη που περιέχει τις διάφορες τιμές εξόδων (κλάσεις) και προχωρούμε στην καταμέτρηση της συχνότητας εμφάνισης κάθε διαφορετικής τιμής εξόδου.

Στη συνέχεια, πραγματοποιούμε διαχωρισμό του σετ δεδομένων σε τρία μη επικαλυπτόμενα υποσύνολα ως εξής:

1. 60% : Σετ Εκπαίδευσης – training set
2. 20% : Σετ Επικύρωσης – validation set
3. 20% : Σετ Ελέγχου – check set

με τρόπο τέτοιο ώστε οι παραπάνω συχνότητες εμφάνισης να διατηρούνται περίπου σταθερές.

Στο σημείο αυτό είναι καλό να εφαρμόσουμε μια προεπεξεργασία στα δεδομένα μας και συγκεκριμένα να ελέγξουμε ότι δεν υπάρχουν κενές τιμές και διπλότυπα δείγματα. Με τον τρόπο αυτό, θα είναι αποτελεσματικότερη, αλλά και ταχύτερη, η διαδικασία εκπαίδευσης. Τέλος, αφού διαπιστώσουμε ότι δεν υπάρχουν NaN τιμές στο Dataset, ολοκληρώνεται η διαδικασία προεπεξεργασίας του Dataset.

Στη συνέχεια εφαρμόζουμε τον αλγόριθμο Relief επιλέγοντας ως αριθμό γειτόνων το 50 ώστε να γίνει εκτίμηση των σημαντικότερων χαρακτηριστικών κατά φθίνουσα σειρά όπως εμφανίζονται στον πίνακα ranks.

Έπειτα χρησιμοποιούμε το συνδυασμό των μεθόδων Grid Search και 5-Fold Cross Validation ώστε να βρούμε το μοντέλο που εκτιμάει καλύτερα την επιθυμητή έξοδο, μέσω της αξιολόγησης μιας σειράς διαφορετικών μοντέλων. Συγκεκριμένα η μέθοδος k-Fold Cross Validation, με τιμή k=5, αποτελείται από τα εξής βήματα:

1. Αρχικά, διαχωρίζουμε το set δεδομένων εκπαίδευσης σε δύο νέα τμήματα, ένα νέο set δεδομένων εκπαίδευσης (80% του αρχικού set εκπαίδευσης) και ένα νέο set δεδομένων επικύρωσης (20% του αρχικού set εκπαίδευσης). Για το κάνουμε αυτό αναδιατάσσουμε τα 5 folds δεδομένων κάθε φορά ως 4/5 folds για set εκπαίδευσης και 1/5 folds για set ελέγχου με όλους τους δυνατούς τρόπους δημιουργώντας τελικά πέντε νέα δευτερεύοντα μοντέλα. Στο σημείο αυτό φροντίζουμε ότι κάθε set (training και validation) σε κάθε Fold αποτελείται από δεδομένα με όλες τις πιθανές εξόδους, ίσα σε ποσοστό με αυτό κάθε εξόδου του αρχικού set για σωστή εκπαίδευση.

2. Εκπαιδεύουμε καθένα από αυτά τα δευτερεύοντα μοντέλα (Folds) και στη συνέχεια υπολογίζουμε το σφάλμα του καθενός ως το μέσο τετραγωνικό σφάλμα MSE.

3. Τέλος, υπολογίζουμε τη μέση τιμή των προηγουμένως υπολογισμένων σφαλμάτων για κάθε κύριο μοντέλο (δηλαδή η μέση τιμή MSE των πέντε folds κάθε κύριου μοντέλου), η οποία αποτελεί αντιπροσωπευτικό δείγμα του πραγματικού σφάλματος για το συνολικό κύριο μοντέλο.

Η παραπάνω διαδικασία συνδυάζεται με τη μέθοδο Grid Search, δηλαδή εκτελείται μια επαναληπτική διαδικασία στην οποία εφαρμόζεται συνεχώς η μέθοδος 5-Fold Cross Validation για διάφορα κύρια μοντέλα μεταβάλλοντας κάθε φορά τόσο το πλήθος των IF THEN κανόνων όσο και το πλήθος χαρακτηριστικών που λαμβάνονται υπόψιν. Έπειτα συγκεντρώνονται όλα τα μέσα σφάλματα, που υπολογίζονται όπως αναφέρθηκε προηγουμένως για κάθε κύριο μοντέλο, και επιλέγεται το βέλτιστο μοντέλο ως αυτό που παρουσιάζει το ελάχιστο μέσο σφάλμα.

Για την ομαδοποίηση και τη δημιουργία των IF THEN κανόνων χρησιμοποιείται η μέθοδος Fuzzy C-Means (FCM) ενώ οι διάφορες περιπτώσεις των μοντέλων που διερευνώνται αποτελούνται από τους συνδυασμούς πλήθους χαρακτηριστικών και IF THEN κανόνων όπως προκύπτουν από το καρτεσιανό γινόμενο των συνόλων αντίστοιχα,

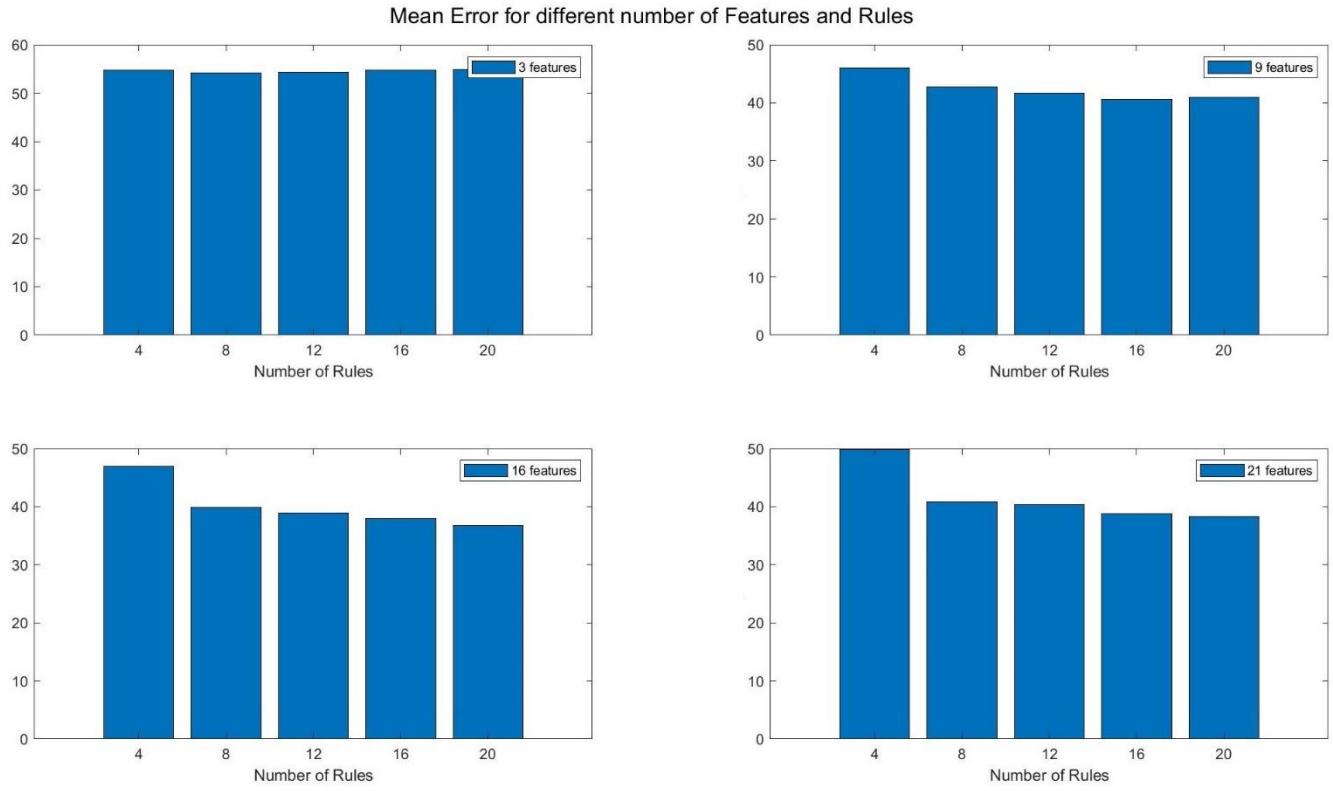
$$NF \times NR = \{3, 9, 16, 21\} \times \{4, 8, 12, 16, 20\}$$

Συνεπώς εξετάζεται η απόδοση 20 διαφορετικών κύριων μοντέλων του αρχικού σετ εκπαίδευσης με βάση τα 5 δευτερεύοντα μοντέλα στα οποία διακρίνεται το καθένα από αυτά (Μέθοδος 5-Fold Validation). Για κάθε κύριο μοντέλο πραγματοποιείται εκπαίδευση, καθενός από τα πέντε δευτερεύοντα μοντέλα του (συνολικά 100 μοντέλα) για 150 εποχές το καθένα, και υπολογίζεται το σφάλμα καθενός από αυτά. Τέλος, υπολογίζεται ο μέσος όρος των 5 σφαλμάτων ο οποίος αποτελεί το κριτήριο για την εύρεση του βέλτιστου από τα κύρια μοντέλα, όπως αναφέρθηκε προηγουμένως.

Στον παρακάτω πίνακα παρουσιάζεται το μέσο MSE για τα 20 διαφορετικά κύρια μοντέλα.

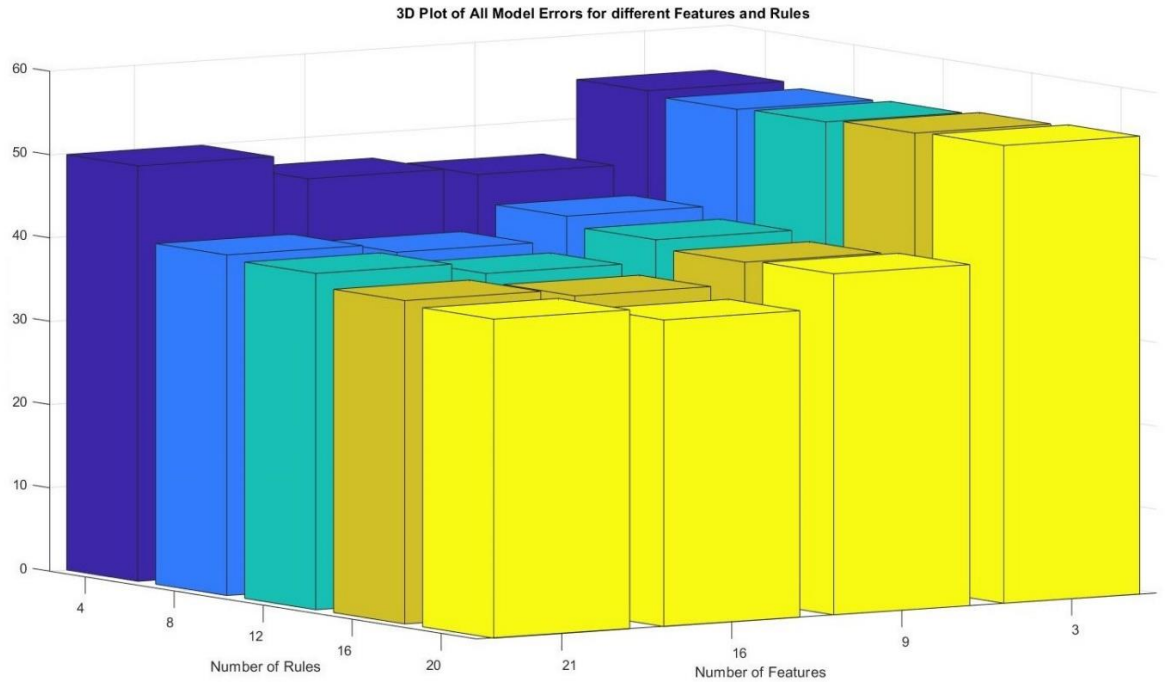
| *Rules* / *Features* | 4 | 8 | 12 | 16 | 20 |
|---|---|---|---|---|---|
| 3 | 54.6965 | 54.1664 | 54.3628 | 54.7276 | 54.9521 |
| 9 | 46.0294 | 42.7510 | 41.6201 | 40.6237 | 40.9394 |
| 16 | 46.9267 | 39.8126 | 38.9600 | 37.9859 | 36.7977 |
| 21 | 49.8663 | 40.8838 | 40.3626 | 38.8072 | 38.2950 |

Στα παρακάτω διαγράμματα φαίνονται γραφικά οι τιμές του μέσου σφάλματος για τις διάφορες τιμές χαρακτηριστικών και κανόνων.




*Σχήμα 28: Μέσο σφάλμα μοντέλων για τις διάφορες τιμές πλήθους χαρακτηριστικών και κανόνων*

Τέλος, τα παραπάνω σφάλματα παρουσιάζονται και σε ένα κοινό διάγραμμα τριών διαστάσεων.




*Σχήμα 29: Κοινό 3D Διάγραμμα Μέσου Σφάλματος των διάφορων μοντέλων*

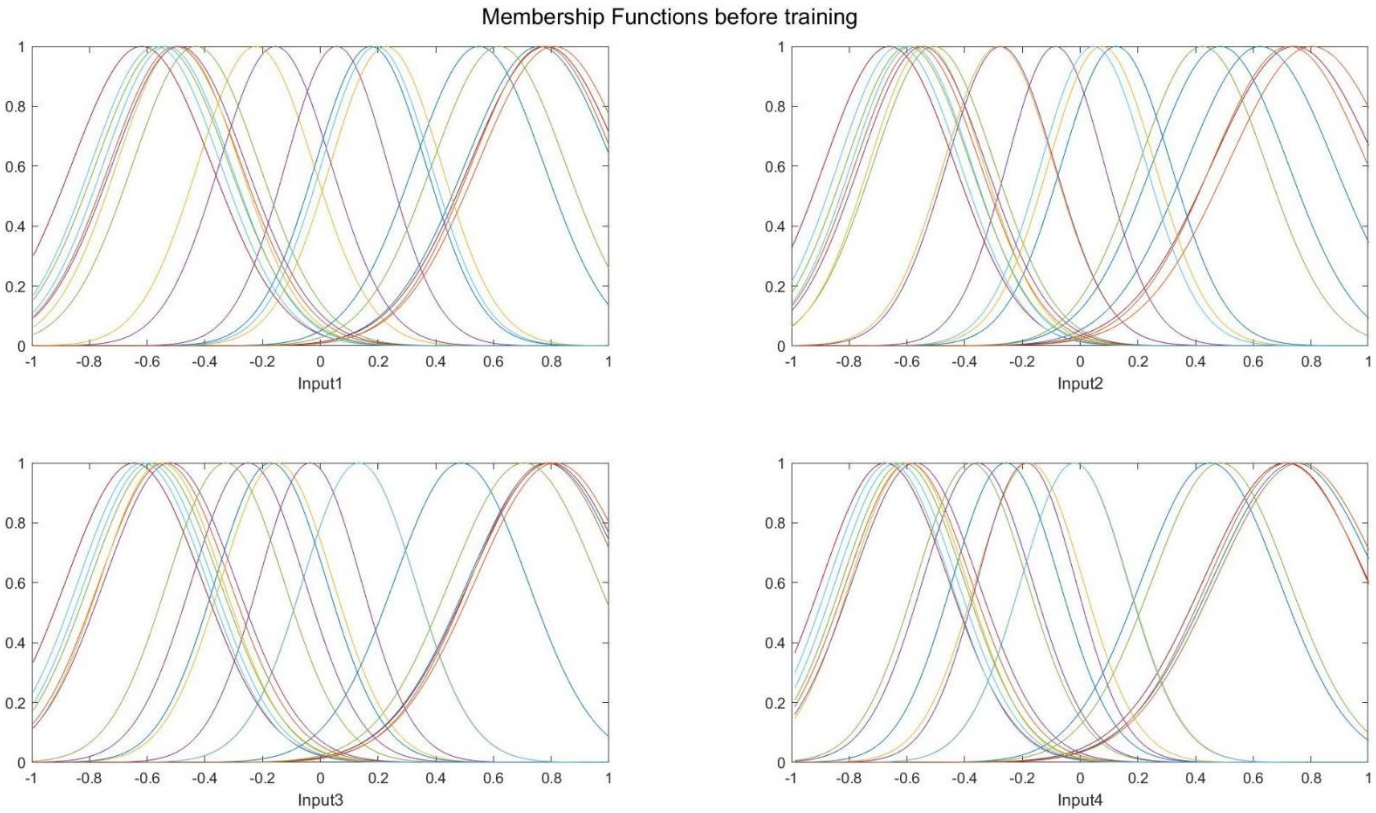
Από τα παραπάνω είναι εμφανές ότι το βέλτιστο, από τα εξεταστέα μοντέλα, είναι αυτό με τα 16 χαρακτηριστικά και τους 20 κανόνες. Παρατηρούμε επίσης ότι όσο αυξάνεται η πολυπλοκότητα του μοντέλου (πλήθος χαρακτηριστικών και κανόνων) τόσο αυξάνεται και ο χρόνος εκτέλεσης του αλγορίθμου, ωστόσο δεν βελτιώνεται απαραίτητα η ικανότητα εκτίμησης - ταξινόμησης του μοντέλου. Τέλος, επισημαίνουμε ότι ο διαμοιρασμός στα διάφορα set γίνεται με τέτοιο τρόπο, ώστε να περιέχουν ίσο σε ποσοστό αριθμό δεδομένων από κάθε κλάση, όπως φαίνεται παρακάτω.

| Output_Values | Isolet_Set | Training_Set | Validation_Set | Check_Set |
|---|---|---|---|---|
| 1 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 2 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 3 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 4 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 5 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 6 | 3.822% | 3.8264% | 3.8462% | 3.7845% |
| 7 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 8 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 9 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 10 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 11 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 12 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 13 | 3.8348% | 3.8264% | 3.8462% | 3.8486% |
| 14 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 15 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 16 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 17 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 18 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 19 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 20 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 21 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 22 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 23 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 24 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 25 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 26 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |

*Σχήμα 30: Διαμοιρασμός Δεδομένων στα διάφορα Σετ*

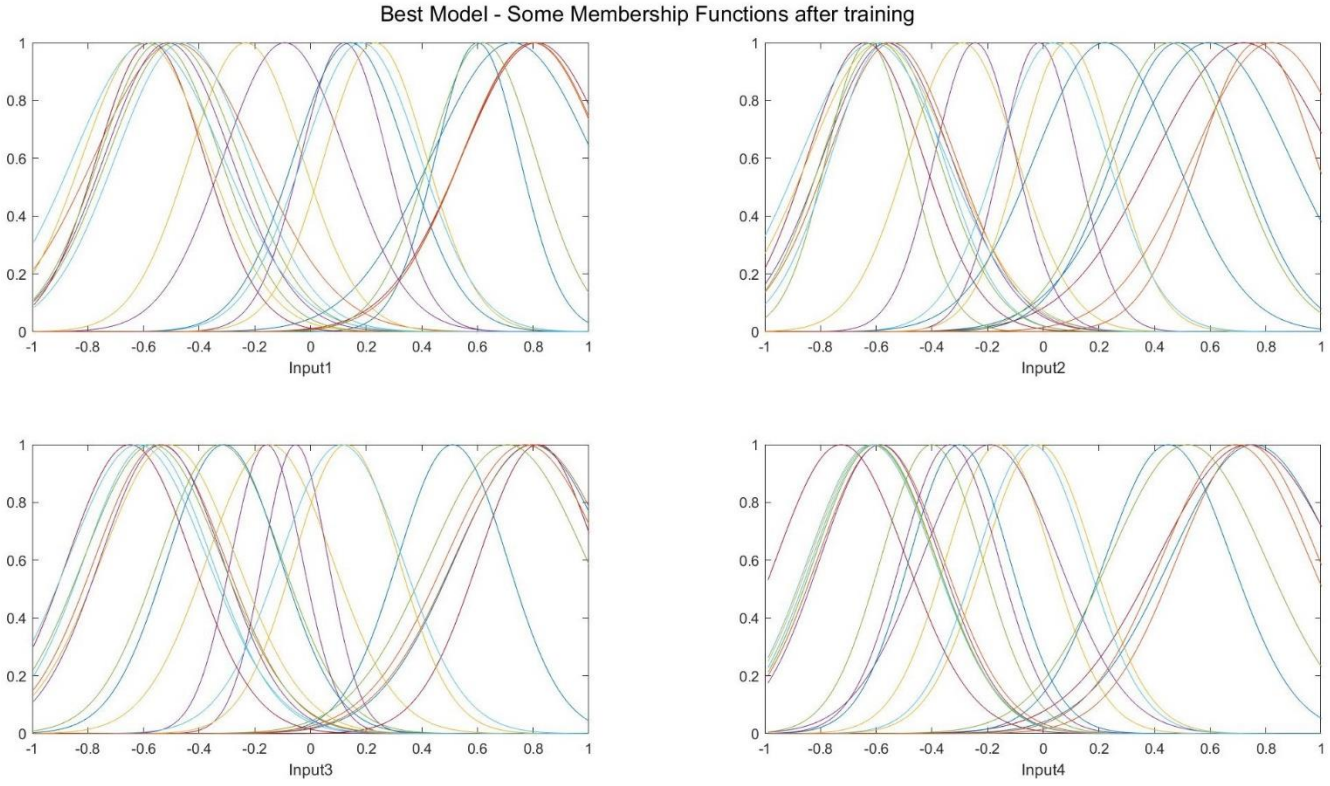
## Εκπαίδευση Βέλτιστου TSK Μοντέλου (20 Features - 16 Rules)

Αρχικά παρουσιάζουμε ορισμένες από τις συναρτήσεις συμμετοχής του βέλτιστου μοντέλου πριν την εκπαίδευσή του.



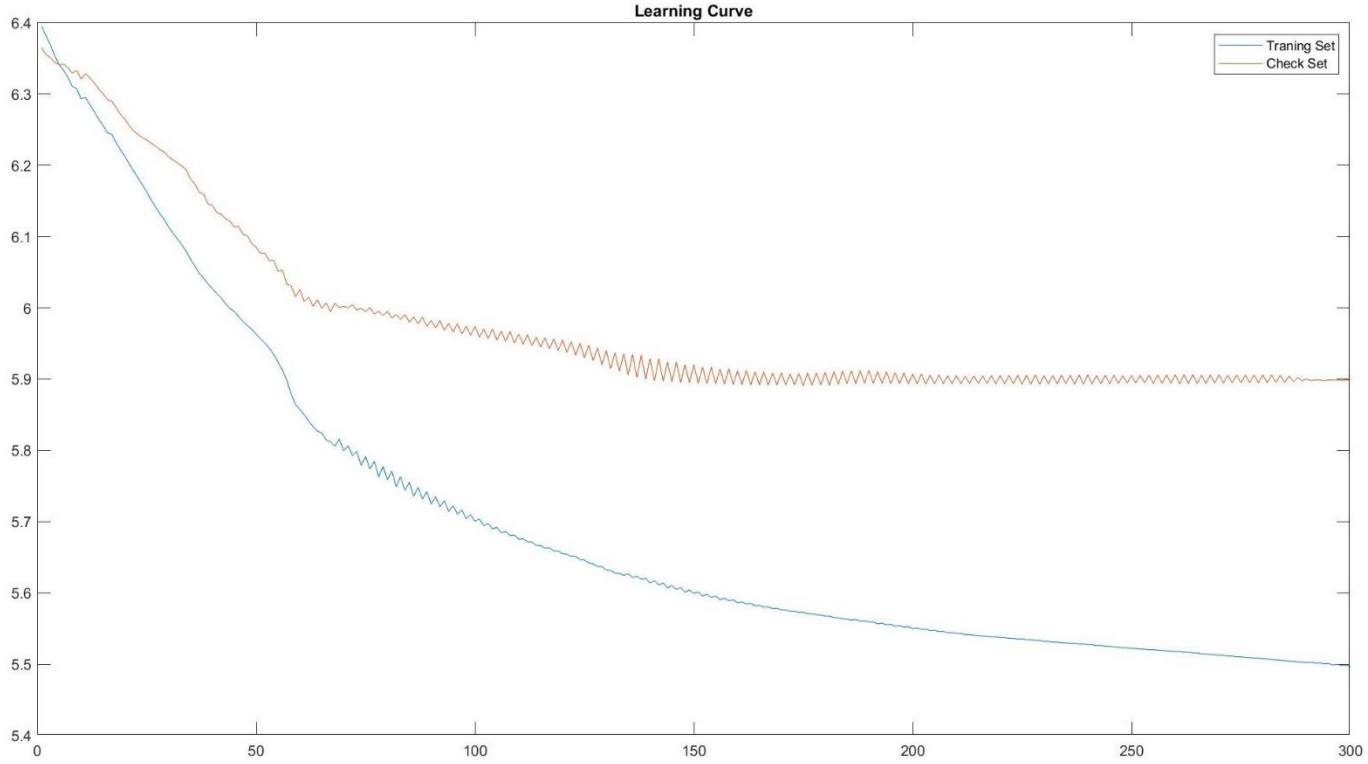

*Σχήμα 31: Συναρτήσεις Συμμετοχής πριν την εκπαίδευση του Βέλτιστου Μοντέλου*

Μετά από εκπαίδευση σε 300 εποχές οι παραπάνω συναρτήσεις συμμετοχής λαμβάνουν την παρακάτω μορφή



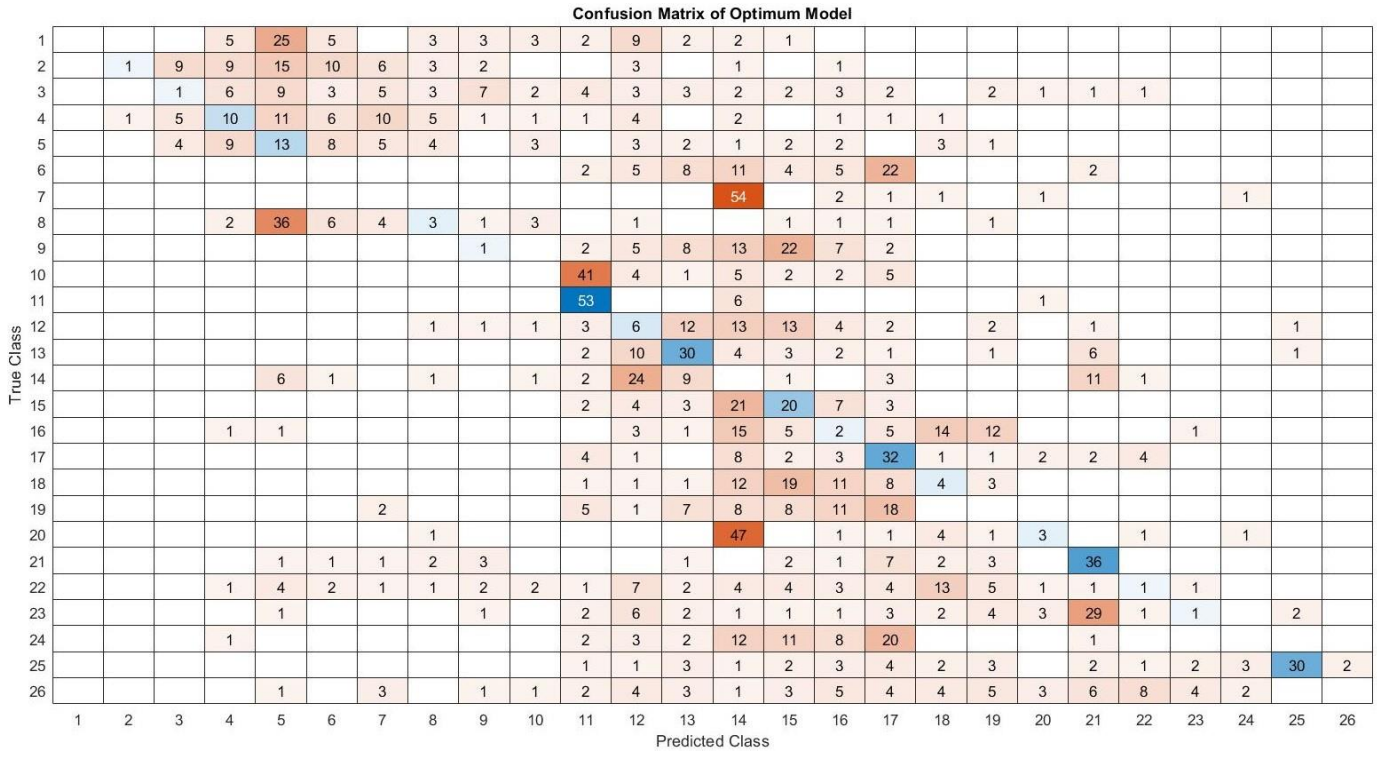

*Σχήμα 32: Συναρτήσεις Συμμετοχής μετά την εκπαίδευση του Βέλτιστου Μοντέλου*

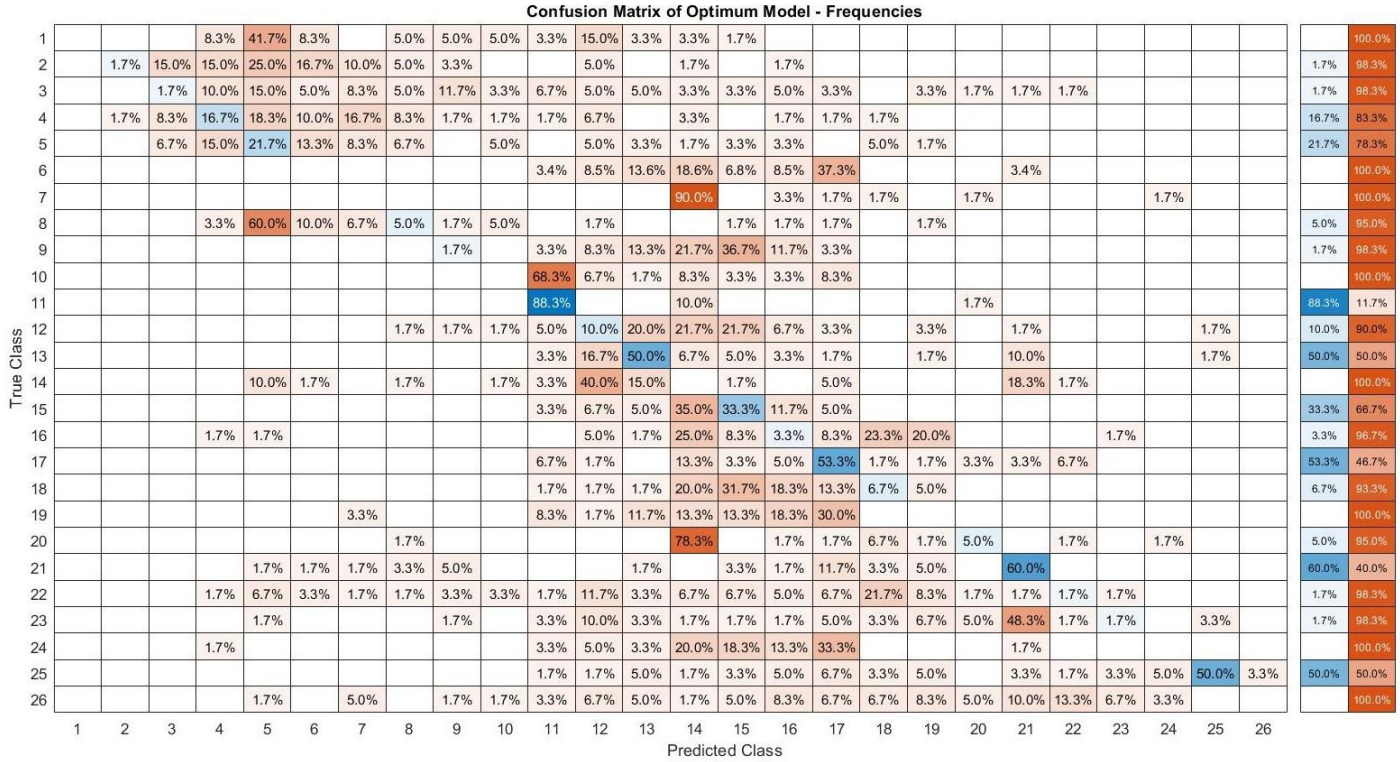
Ακολουθούν οι καμπύλες εκμάθησης με βάση το RMSE στο πέρας των εποχών.



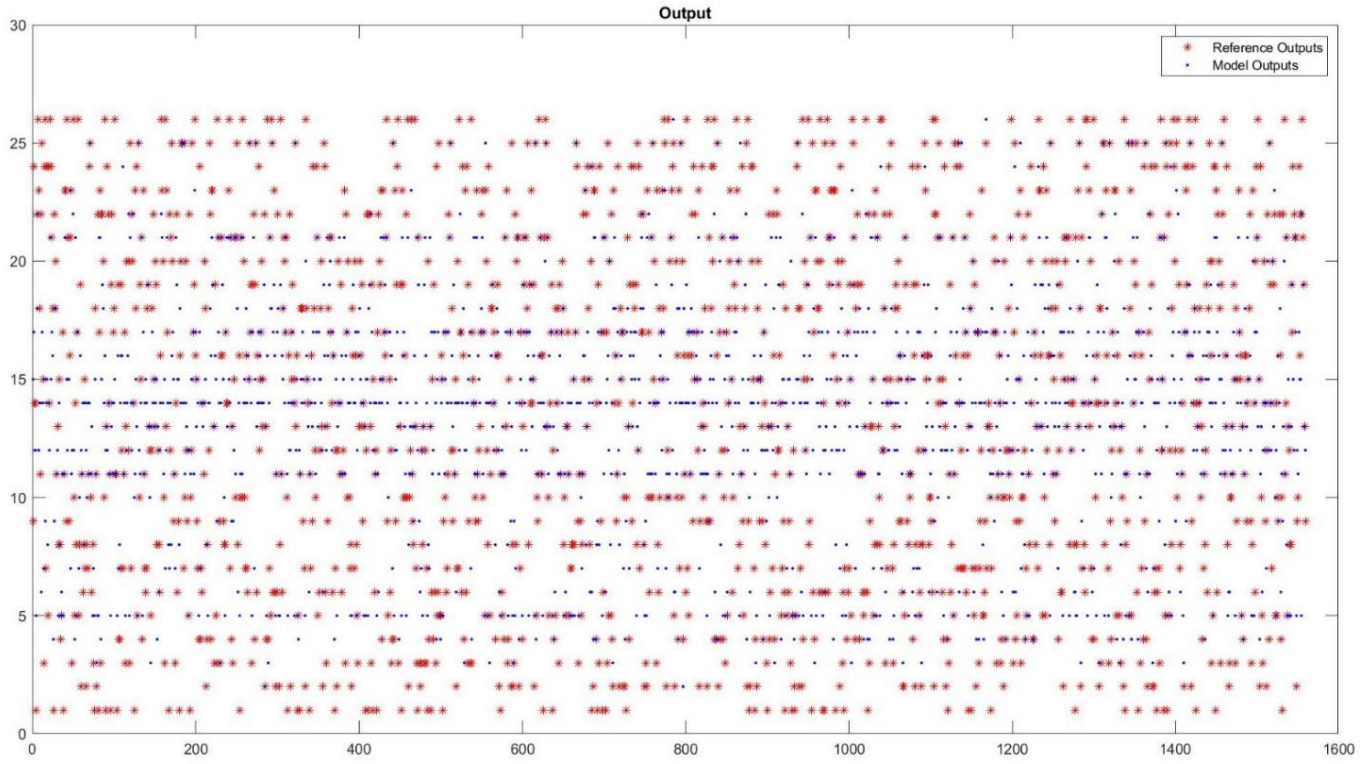

*Σχήμα 33: Καμπύλες Εκμάθησης - TSK Βέλτιστο Μοντέλο*

Ακόμα, βλέπουμε τους Πίνακες Σφαλμάτων Ταξινόμησης με τη μορφή απολύτων τιμών αλλά και ποσοστιαία καθώς και τις τιμές πραγματικής και εκτιμήτριας εξόδου για το σύνολο των δεδομένων ελέγχου.






*Σχήμα 34: Πίνακες Σφαλμάτων Ταξινόμησης – TSK Βέλτιστο Μοντέλο*




*Σχήμα 35: Πραγματική και Εκτιμήτρια Έξοδος – TSK Βέλτιστο Μοντέλο*

## Δείκτες Απόδοσης και Χρόνος Εκτέλεσης

Στον παρακάτω πίνακα βλέπουμε τους δείκτες απόδοσης και το χρόνο εκτέλεσης για την εκπαίδευση και αξιολόγηση του βέλτιστου μοντέλου.

| *Accuracy Metrics* | *Class Number* | Producers Accuracy | Users Accuracy | *Class Number* | Producers Accuracy | Users Accuracy |
|---|---|---|---|---|---|---|
| | 1 | NaN * | 0 | 14 | 0 | 0 |
| *Overall Accuracy* | 2 | 0.5000 | 0.0167 | 15 | 0.1563 | 0.3333 |
| | 3 | 0.0526 | 0.0167 | 16 | 0.0233 | 0.0333 |
| 0.1584 | 4 | 0.2273 | 0.1667 | 17 | 0.2148 | 0.5333 |
| | 5 | 0.1057 | 0.2167 | 18 | 0.0784 | 0.0667 |
| $\hat{k}$ | 6 | 0 | 0 | 19 | 0 | 0 |
| | 7 | 0 | 0 | 20 | 0.2000 | 0.0500 |
| 0.1248 | 8 | 0.1111 | 0.0500 | 21 | 0.3673 | 0.6000 |
| | 9 | 0.0435 | 0.0167 | 22 | 0.0556 | 0.0167 |
| *Elapsed Time* | 10 | 0 | 0 | 23 | 0.1111 | 0.0167 |
| | 11 | 0.4015 | 0.8833 | 24 | 0 | 0 |
| 229.366 sec | 12 | 0.0556 | 0.1000 | 25 | 0.8824 | 0.5000 |
| | 13 | 0.3000 | 0.5000 | 26 | 0 | 0 |

*NaN: Not a Number, εμφανίζεται καθώς γίνεται απόπειρα εκτέλεσης της πράξης 0/0

Από την παραπάνω εικόνα βλέπουμε ότι το βέλτιστο μοντέλο με τους 16 κανόνες και τα 20 χαρακτηριστικά πετυχαίνει, μετά από 300 εποχές εκπαίδευσης, Overall Accuracy ίσο με 15.84% και δείκτη $\hat{k}$ ίσο με 0.1248. Είναι προφανές ότι τα αποτελέσματα από την εκτέλεση του αλγορίθμου δεν είναι ιδιαίτερα ικανοποιητικά κάτι που οφείλεται κυρίως στο μεγάλο πλήθος των κλάσεων συγκριτικά με το μικρό πλήθος δεδομένων που είναι διαθέσιμο.

Επίσης, υπεισέρχεται και πάλι ο παράγοντας της διαθέσιμης υπολογιστικής ισχύος και ο περιορισμός στη χρήση του αλγορίθμου Anfis. Με διαφορετική επιλογή αλγορίθμου εκπαίδευσης και με τη διάθεση περισσότερων υπολογιστικών πόρων-ισχύος είναι πιθανό να πετύχουμε αρκετά καλύτερα αποτελέσματα, ξεφεύγοντας, ωστόσο, από το σκοπό της παρούσας εργασίας.

Όμοια με τα προηγούμενα τμήματα της εργασίας επισημαίνουμε ότι ο διαμοιρασμός στα διάφορα set γίνεται με τέτοιο τρόπο, ώστε να περιέχουν ίσο σε ποσοστό αριθμό δεδομένων από κάθε κλάση, όπως φαίνεται παρακάτω.

| Output_Values | Isolet_Set | Training_Set | Validation_Set | Check_Set |
|---|---|---|---|---|
| 1 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 2 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 3 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 4 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 5 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 6 | 3.822% | 3.8264% | 3.8462% | 3.7845% |
| 7 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 8 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 9 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 10 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 11 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 12 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 13 | 3.8348% | 3.8264% | 3.8462% | 3.8486% |
| 14 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 15 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 16 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 17 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 18 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 19 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 20 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 21 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 22 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 23 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 24 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 25 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |
| 26 | 3.8476% | 3.8478% | 3.8462% | 3.8486% |

*Σχήμα 37: Διαμοιρασμός Δεδομένων στα διάφορα Σετ - Βέλτιστο Μοντέλο*

## Αρχεία MATLAB

1. avilaModel.m : MATLAB Script – Υλοποίηση πρώτου τμήματος της εργασίας (Avila Dataset).

2. gridSearch.m : MATLAB Script – Υλοποίηση δεύτερου τμήματος της εργασίας (Isolet Dataset). Ο χρήστης ενημερώνεται σε ζωντανό χρόνο για την πρόοδο της διαδικασίας εκπαίδευσης των 100 μοντέλων και τις παραμέτρους (πλήθος χαρακτηριστικών, πλήθος κανόνων, αριθμός πτυχής) του μοντέλου που εκπαιδεύεται κάθε φορά. Τέλος, δημιουργείται και ένα αρχείο με όνομα optimum_model.mat, το οποίο περιλαμβάνει τον αριθμό των χαρακτηριστικών και κανόνων του βέλτιστου μοντέλου καθώς και το απαραίτητο τμήμα του του πίνακα ranks, που καθορίζει με φθίνουσα σειρά σημασίας ποιες από τις στήλες των χαρακτηριστικών χρησιμοποιήθηκαν.

3. optimumModel.m : MATLAB Script – Εκπαίδευση του βέλτιστου TSK μοντέλου και υπολογισμός των απαραίτητων δεικτών απόδοσης.