Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Εργασία στα Ασαφή Συστήματα

Επίλυση προβλήματος ταξινόμησης με χρήση μοντέλων TSK – Group4-Ser07

Μανουσαρίδης Ιωάννης (8855)

Θεσσαλονίκη, Δεκέμβριος 2019

Περιεχόμενα

Περιγραφή του Προβλήματος

Στόχος της εργασίας αυτής είναι να διερευνηθεί η ικανότητα των μοντέλων TSK στην επίλυση προβλημάτων ταξινόμησης (classification). Συγκεκριμένα, επιλέγονται δύο σύνολα δεδομένων από το UCI repository με σκοπό την ταξινόμηση, από τα διαθέσιμα δεδομένα, δειγμάτων στις εκάστοτε κλάσεις τους, με χρήση ασαφών νευρωνικών μοντέλων. Η εργασία αποτελείται από δύο μέρη, το πρώτο από τα οποία προορίζεται για μια απλή διερεύνηση της διαδικασίας εκπαίδευσης και αξιολόγησης των TSK μοντέλων, ενώ το δεύτερο περιλαμβάνει μια πιο συστηματική προσέγγιση στο πρόβλημα της εκμάθησης από δεδομένα, σε συνδυασμό με προεπεξεργαστικά βήματα όπως επιλογή χαρακτηρηστικών (feature selection) και μεθόδους βελτιστοποίησης των μοντέλων μέσω της διασταυρωμένης επικύρωσης (cross validation).

Εφαρμογή σε απλό Σετ Δεδομένων

Προετοιμασία του Σετ Δεδομένων

Το σύνολο δεδομένων Avila έχει εξαχθεί από 800 εικόνες της «Βίβλου Avila», ενός γιγάντου λατινικού αντιγράφου της Βίβλου του 12ου αιώνα. Η εργασία πρόβλεψης συνίσταται στη συσχέτιση κάθε σχεδίου με έναν αντιγραφέα.

Περισσότερες πληροφορίες: https://archive.ics.uci.edu/ml/datasets/Avila#

Αρχικά, το σετ δεδομένων ταξινομείται κατά αύξουσα σειρά με βάση τις τιμές των κλάσεων (11η στήλη – value range = [1,12]) και μετριέται η συχνότητα εμφάνισης της καθεμίας. Στη συνέχεια, πραγματοποιείται διαχωρισμός του σετ δεδομένων σε τρία μη επικαλυπτόμενα υποσύνολα με τρόπο τέτοιο ώστε οι παραπάνω συχνότητες εμφάνισης να διατηρούνται περίπου σταθερές και τα τρία σετ ανακατεύονται. Ο διαχωρισμός γίνεται ως εξής:

- 60% : σύνολο εκπαίδευσης – training data
- 20% : σύνολο επικύρωσης – validation data
- 20% : σύνολο ελέγχου – check data

Διαδικασία Εκπαίδευσης

Για τα 5 μοντέλα TSK, η εκπαίδευση γίνεται με την υβριδική μέθοδο, δηλαδή οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται με τη μέθοδο

οπισθοδιάδοσης (Backpropagation) και οι παράμετροι της πολυωνυμικής συνάρτησης εξόδου βελτιστοποιούνται με τη μέθοδο Least Squares.

Η διαμέριση του εισόδου πραγματοποιείται με τη μέθοδο του Subtractive Clustering (SC) με χρήση της συνάρτησης genfis() (χρησιμοποιείται η genfis αντί για την genfis2). Κατόπιν, τα μοντέλα εκπαιδεύονται με τη χρήση της συνάρτησης anfis(). Η αξιολόγηση των τελικών μοντέλων γίνεται με βάση τον πίνακα σφαλμάτων (error matrix), την συνολική ακρίβεια (Overall accuracy), την ακρίβεια παραγωγού και την ακρίβεια χρήστη (Producer's accuracy – User's accuracy) καθώς με ένα άλλο στατιστικό μέγεθος, την εκτίμηση της πραγματικής στατιστικής παραμέτρου $\breve{K}$.

<u>Σημείωση για την έξοδο των μοντέλων</u>:

΄Ενα σημείο το οποίο μπορεί να αποτελέσει πηγή σύγχυσης, είναι το γεγονός ότι η υλοποίηση των TSK ασαφών μοντέλων στο MATLAB είναι τέτοια ώστε η έξοδός τους να είναι πραγματική, κάτι το οποίο οδηγεί σε δυσκολίες σε προβλήματα ταξινόμησης. Για την επίλυση του προβλήματος, η έξοδος του μοντέλου στρογγυλοποιείται για κάθε στοιχείο στον πλησιέστερο ακέραιο.

Σύμφωνα με τη μέθοδο Subtractive Clustering, η τιμή ακτίνας επηρεάζει το πλήθος των IF/THEN κανόνων. Στη διαδικασία αυτή μεταβλήθηκε και μία ακόμα παράμετρος, το squash factor, των genfisOptions. Τα πέντε μοντέλα TSK προς εκπαίδευση καθώς και οι ακτίνες του και τα αντίστοιχα squash factor διακρίνονται με βάση τον πίνακα 1.

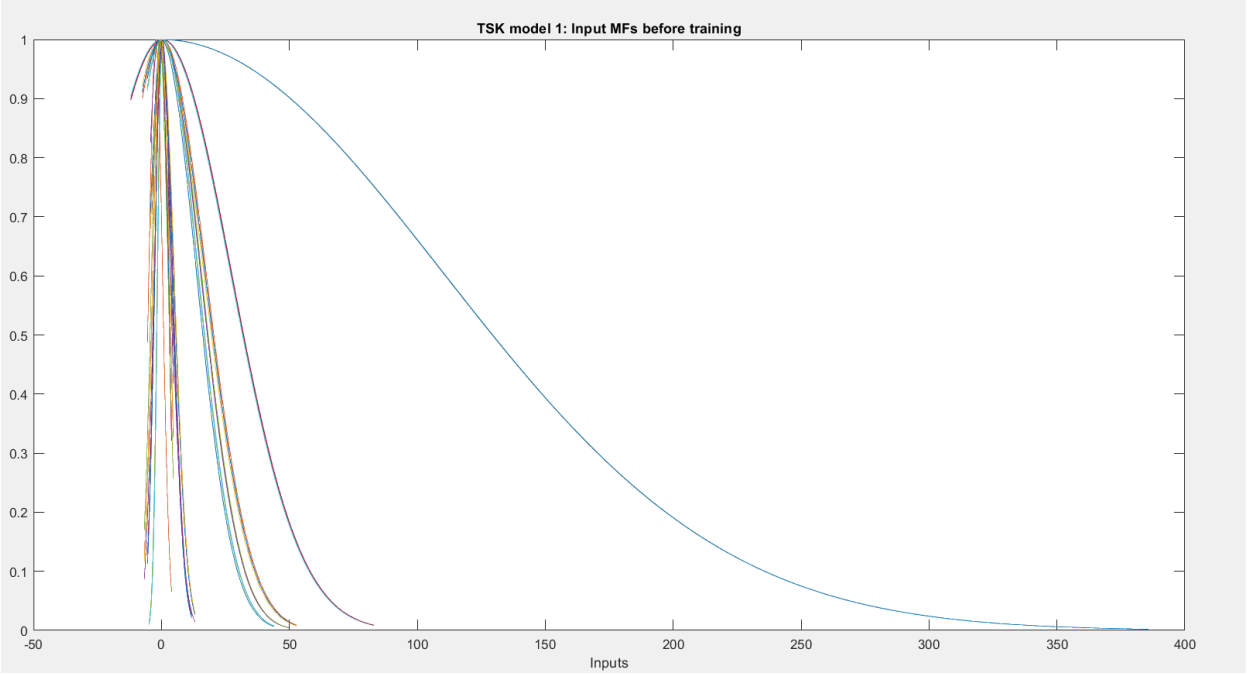
Πίνακας 1.

TSK Model	Radius	Squash Factor	Rules	Output Format
Model 1	**0.8**	**0.5**	**4**	**Singleton**
Model 2	**0.8**	**0.5**	**8**	**Singleton**
Model 3	**0.3**	**0.5**	**12**	**Singleton**
Model 4	**0.6**	**0.5**	**16**	**Singleton**
Model 5	**0.4**	**0.5**	**20**	**Singleton**

Αποτελέσματα TSK Μοντέλων και Μετρικές Σφάλματος

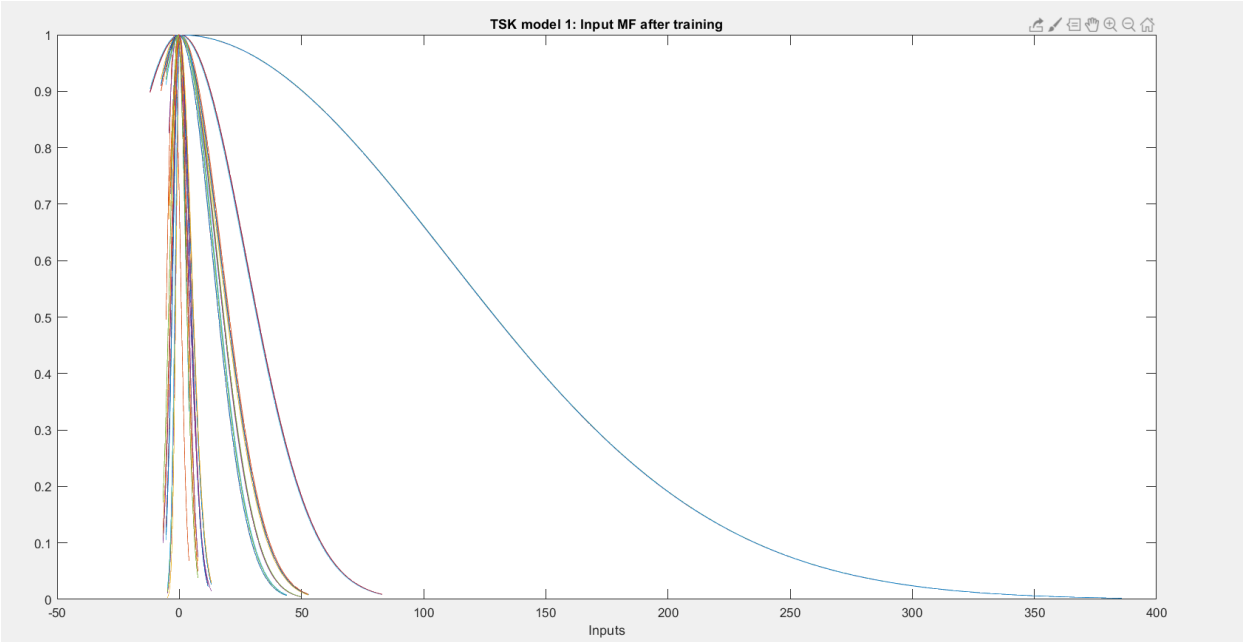
TSK Μοντέλο 1

Στο πρώτο μοντέλο TSK οι συναρτήσεις συμμετοχής πριν τη διαδικασία εκπαίδευσης φαίνονται στο Σχήμα 1.



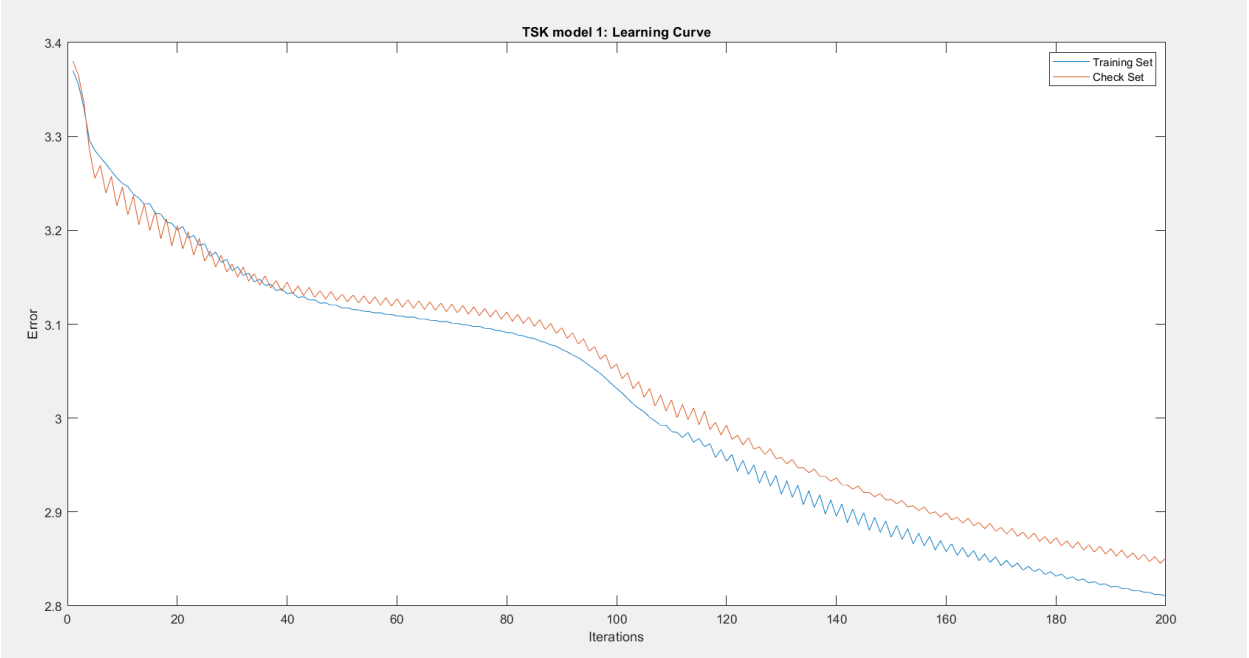
Σχήμα 1: Αρχικές Συναρτήσεις Συμμετοχής για το TSK Μοντέλο 1.

Οι συναρτήσεις συμμετοχής μετά την εκπαίδευση παίρνουν την εξής μορφή:



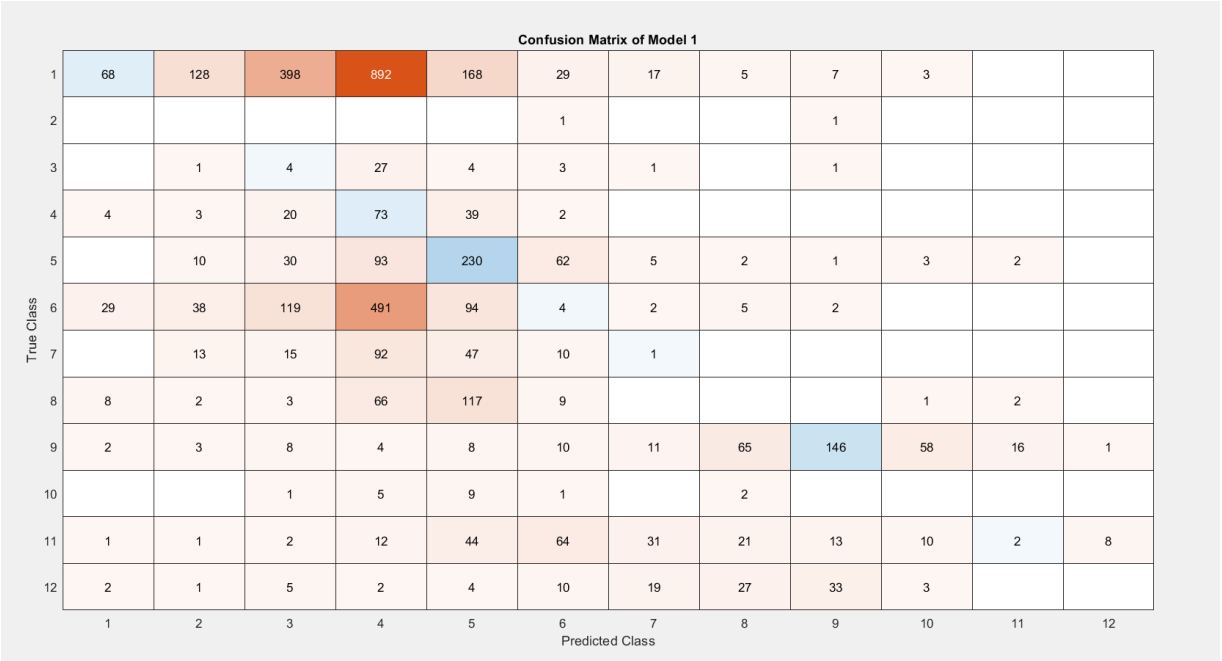
Σχήμα 2: Συναρτήσεις συμμετοχής μετά την εκπαίδευση για το TSK μοντέλο 1.

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

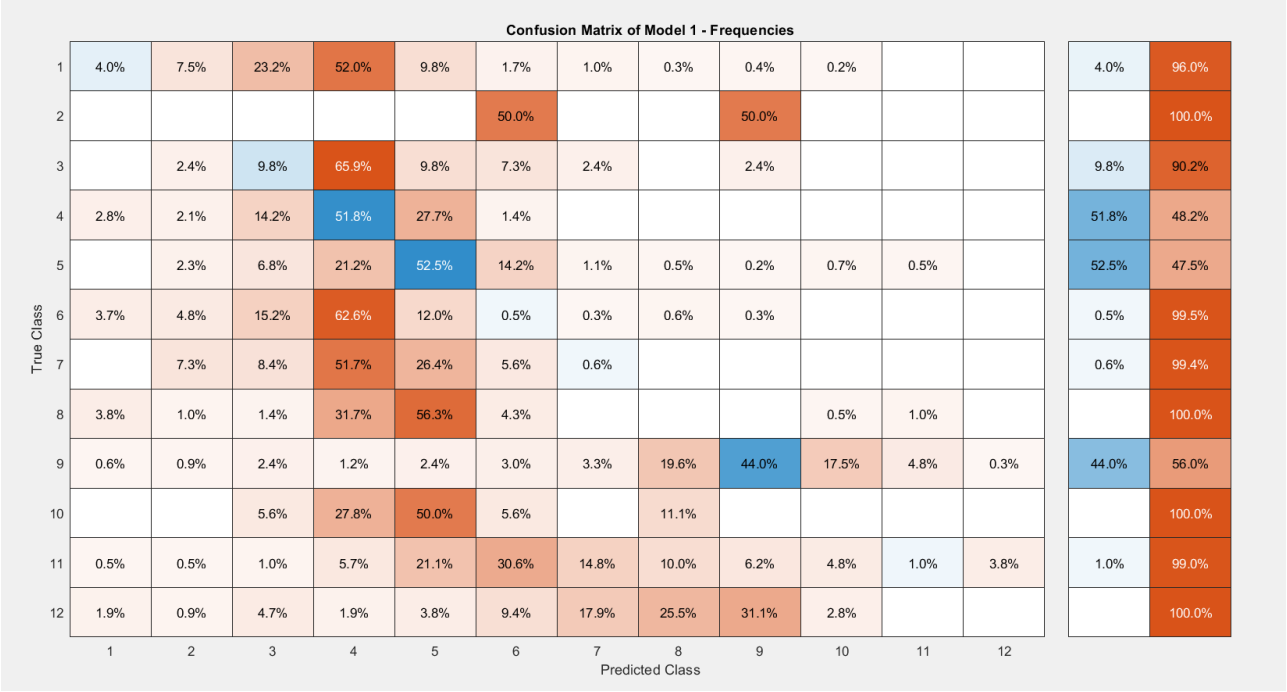


Σχήμα 3: Καμπύλες Εκμάθησης για το TSK Μοντέλο 1.

Τα σφάλματα πρόβλεψης της ταξινόμησης με τη μορφή απολύτων τιμών και συχνοτήτων φαίνονται παρακάτω.



Σχήμα 4: Σφάλματα Πρόβλεψης Ταξινόμησης για το TSK Μοντέλο 1.



Σχήμα 5: Σφάλματα Πρόβλεψης Ταξινόμησης (Συχνότητες) για το TSK Μοντέλο 1.

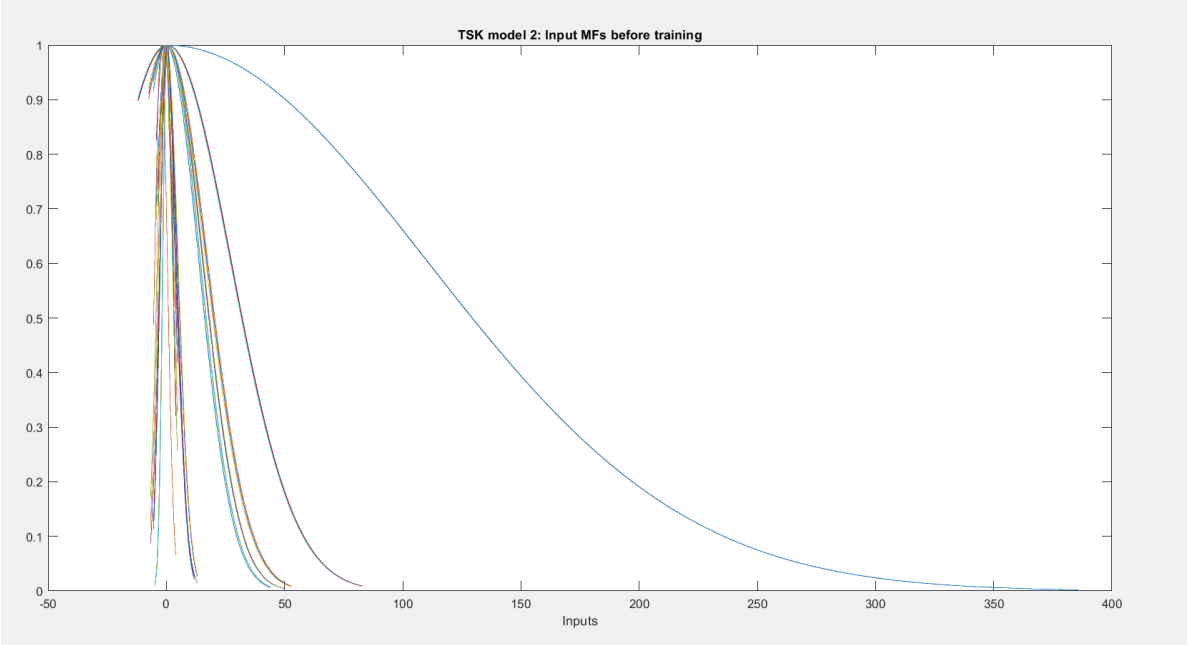
Παρακάτω φαίνεται ο πίνακας με τους ζητούμενους δείκτες απόδοσης.

Πίνακας 2.

Class Number	1	2	3	4	5	6	7	8	9	10	11	12
Producers Accuracy	0.5965	0	0.0066	0.0415	0.301	0.0195	0.0114	0	0.7157	0	0.0909	0
Users Accuracy	0.0396	0	0.0975	0.5177	0.5251	0.0051	0.0056	0	0.4397	0	0.0095	0
Overall Accuracy	0.12656					$\hat{k}$		0.068754				

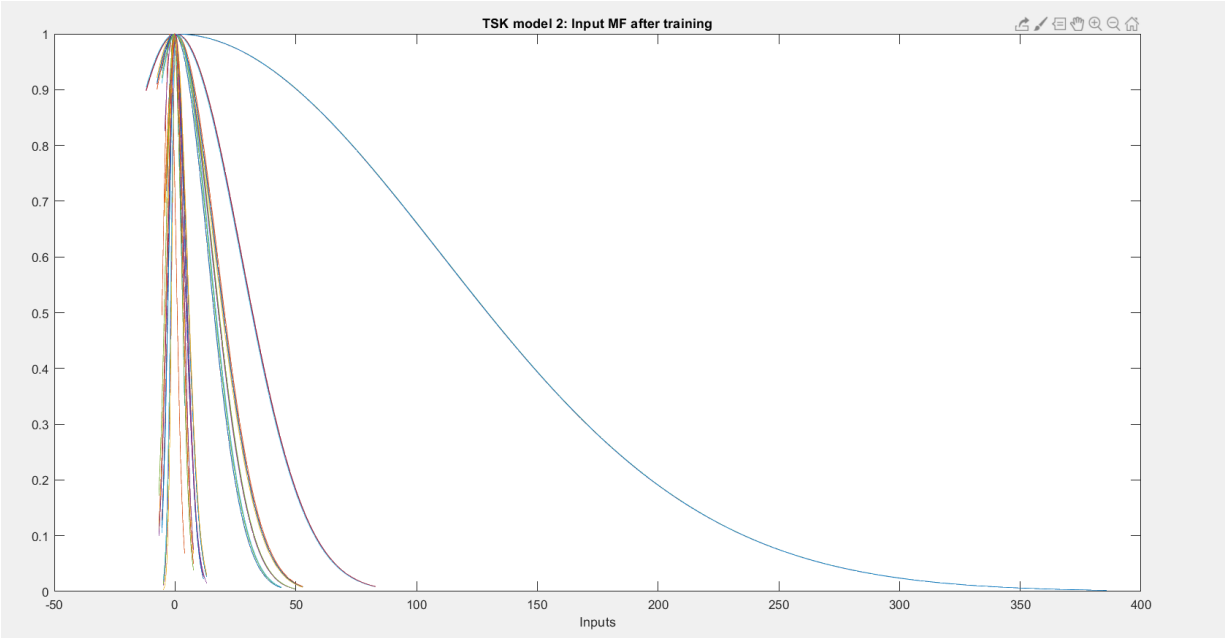
TSK Μοντέλο 2

Στο δεύτερο μοντέλο TSK οι συναρτήσεις συμμετοχής φαίνονται στο Σχήμα 6.



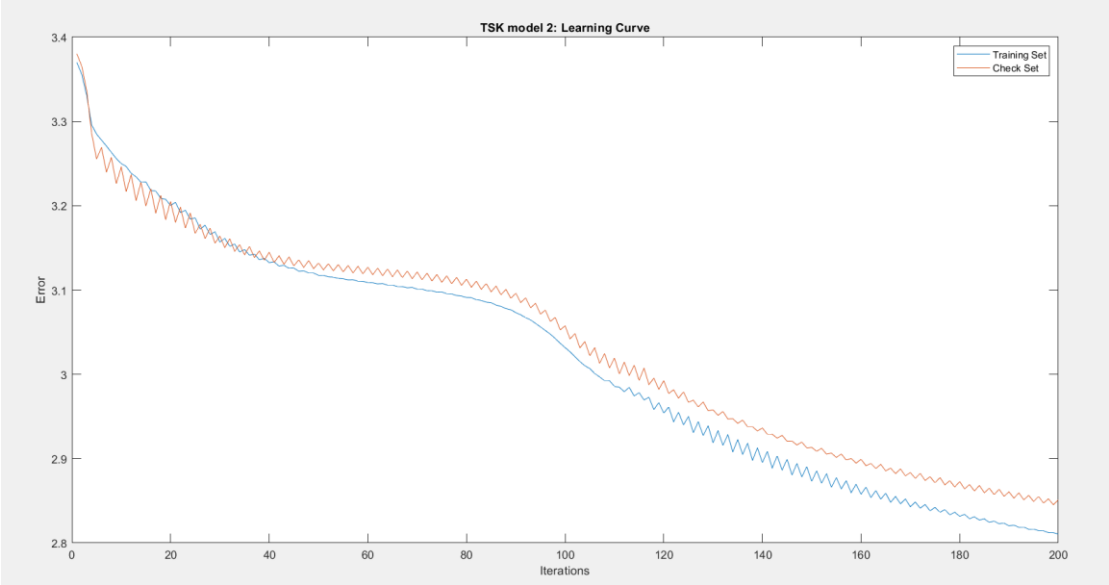
Σχήμα 6: Αρχικές Συναρτήσεις Συμμετοχής για το TSK Μοντέλο 2.

Οι συναρτήσεις συμμετοχής μετά την εκπαίδευση παίρνουν την εξής μορφή:



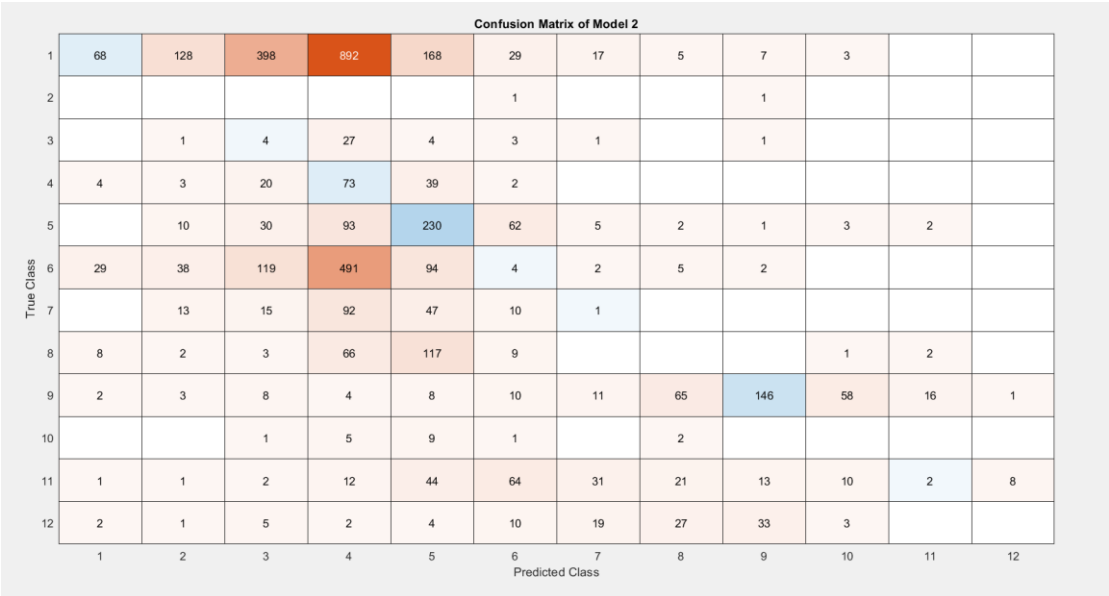
Σχήμα 7: Συναρτήσεις συμμετοχής μετά την εκπαίδευση για το TSK μοντέλο 2.

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

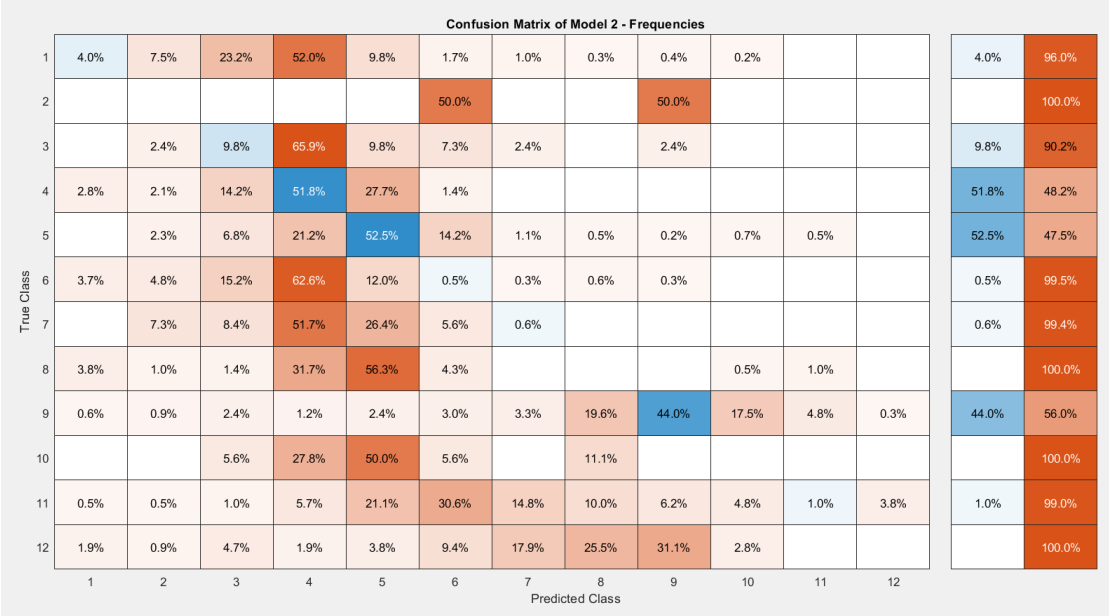


Σχήμα 8: Καμπύλες Εκμάθησης για το TSK Μοντέλο 2.

Τα σφάλματα πρόβλεψης της ταξινόμησης με τη μορφή απολύτων τιμών και συχνοτήτων φαίνονται παρακάτω.



Σχήμα 9: Σφάλματα Πρόβλεψης Ταξινόμησης για το TSK Μοντέλο 2.



Σχήμα 10: Σφάλματα Πρόβλεψης Ταξινόμησης (Συχνότητες) για το TSK Μοντέλο 2.

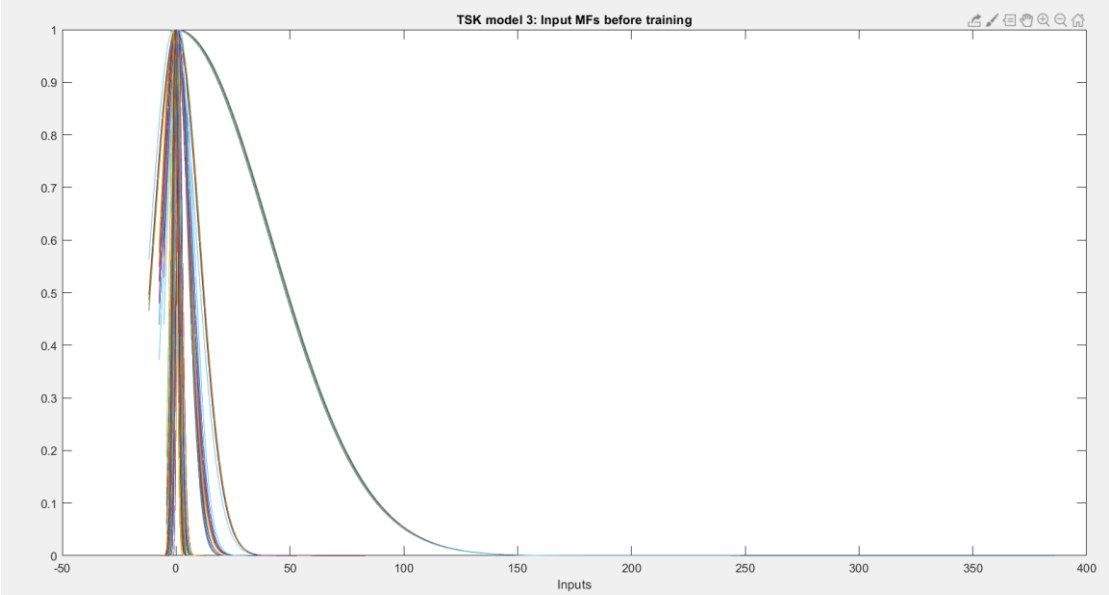
Παρακάτω φαίνεται ο πίνακας με τους ζητούμενους δείκτες απόδοσης.

Πίνακας 3.

Class Number	1	2	3	4	5	6	7	8	9	10	11	12
Producers Accuracy	0.0396	0	0.0976	0.5177	0.5251	0.0051	0.0056	0	0.4398	0	0.0096	0
Users Accuracy	0.0397	0	0.0976	0.5177	0.5251	0.0051	0.0056	0	0.4397	0	0.0096	0
Overall Accuracy	0.1266					$\hat{k}$		0.06875				

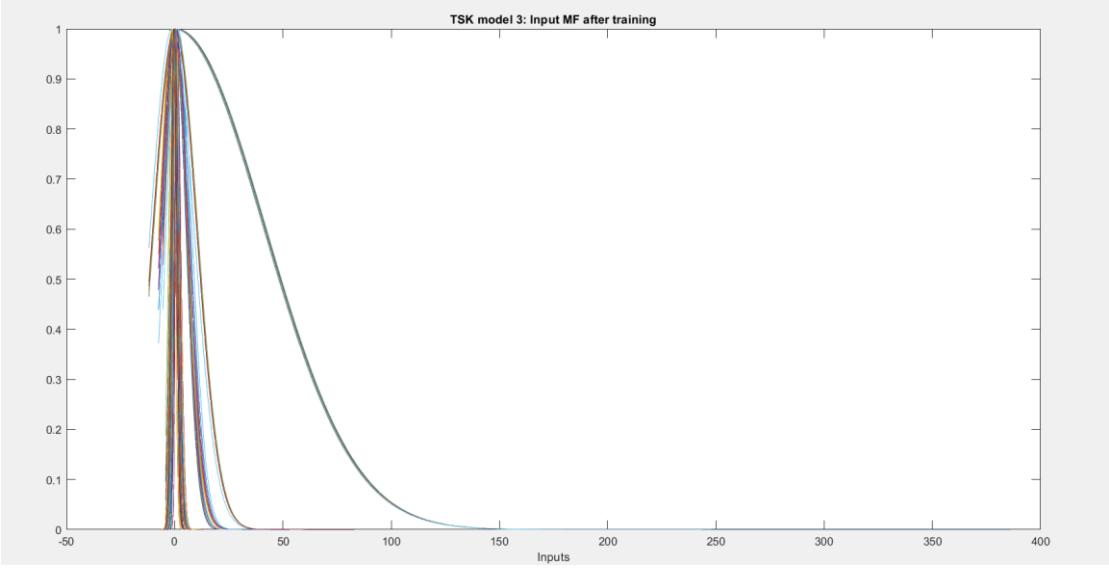
TSK Μοντέλο 3

Στο τρίτο μοντέλο TSK οι συναρτήσεις συμμετοχής φαίνονται στο Σχήμα 11.



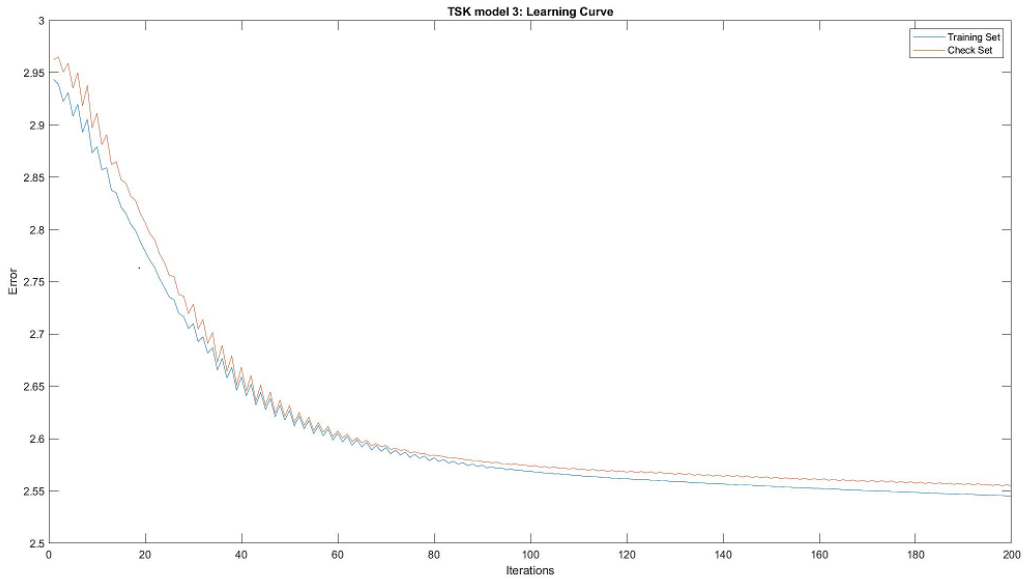
Σχήμα 11: Αρχικές Συναρτήσεις Συμμετοχής για το TSK Μοντέλο 3.

Οι συναρτήσεις συμμετοχής μετά την εκπαίδευση παίρνουν την εξής μορφή:



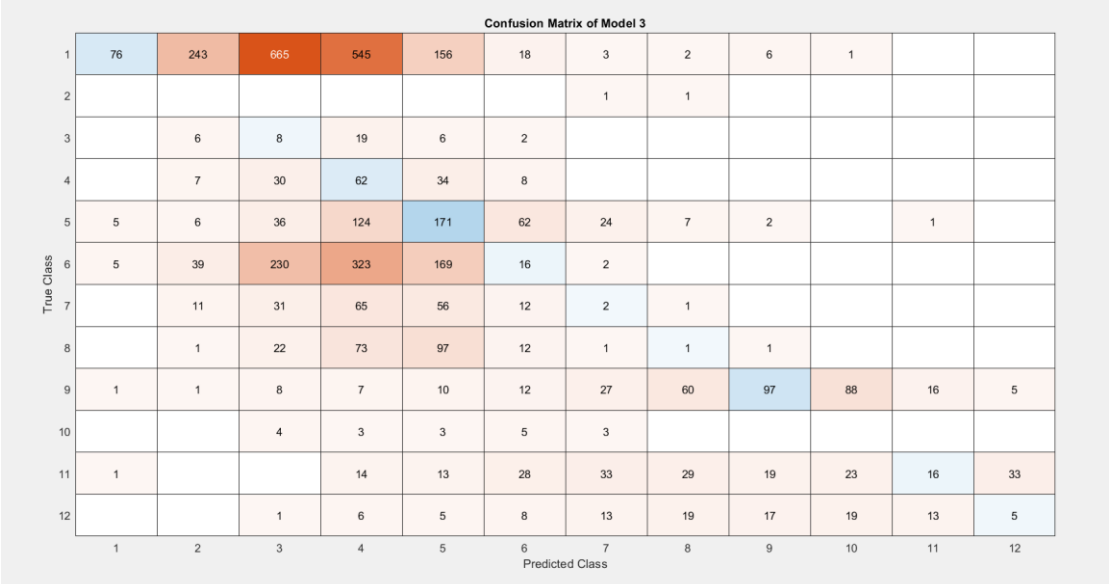
Σχήμα 12: Συναρτήσεις συμμετοχής μετά την εκπαίδευση για το TSK μοντέλο 3.

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

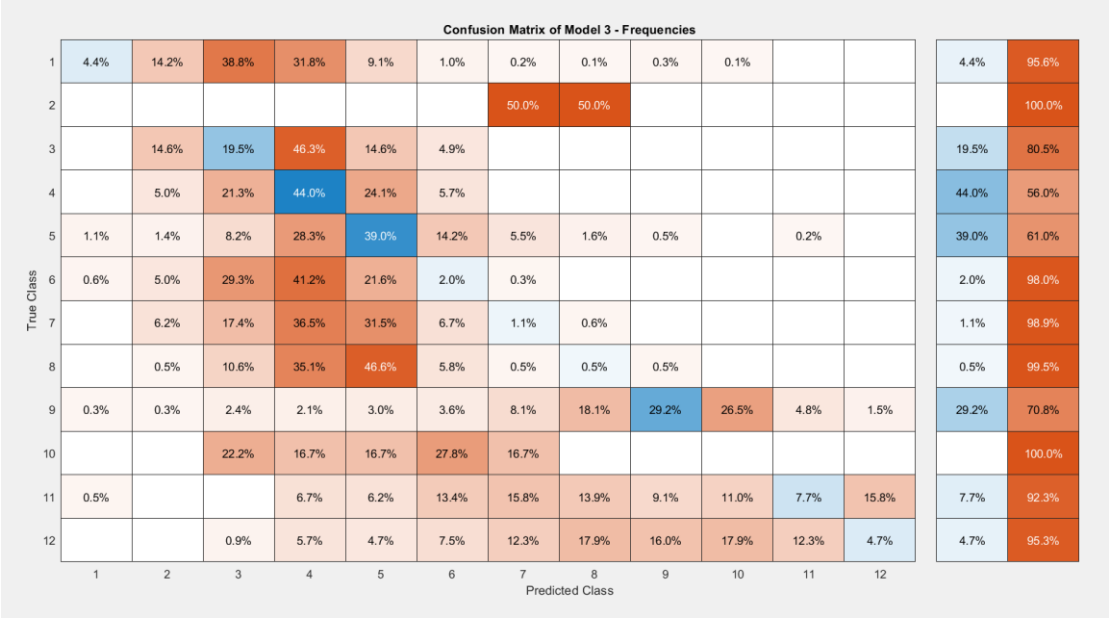


Σχήμα 13: Καμπύλες Εκμάθησης για το TSK Μοντέλο 3.

Τα σφάλματα πρόβλεψης της ταξινόμησης με τη μορφή απολύτων τιμών και συχνοτήτων φαίνονται παρακάτω.



Σχήμα 14: Σφάλματα Πρόβλεψης Ταξινόμησης για το TSK Μοντέλο 3.



Σχήμα 15: Σφάλματα Πρόβλεψης Ταξινόμησης (Συχνότητες) για το Μοντέλο 3.

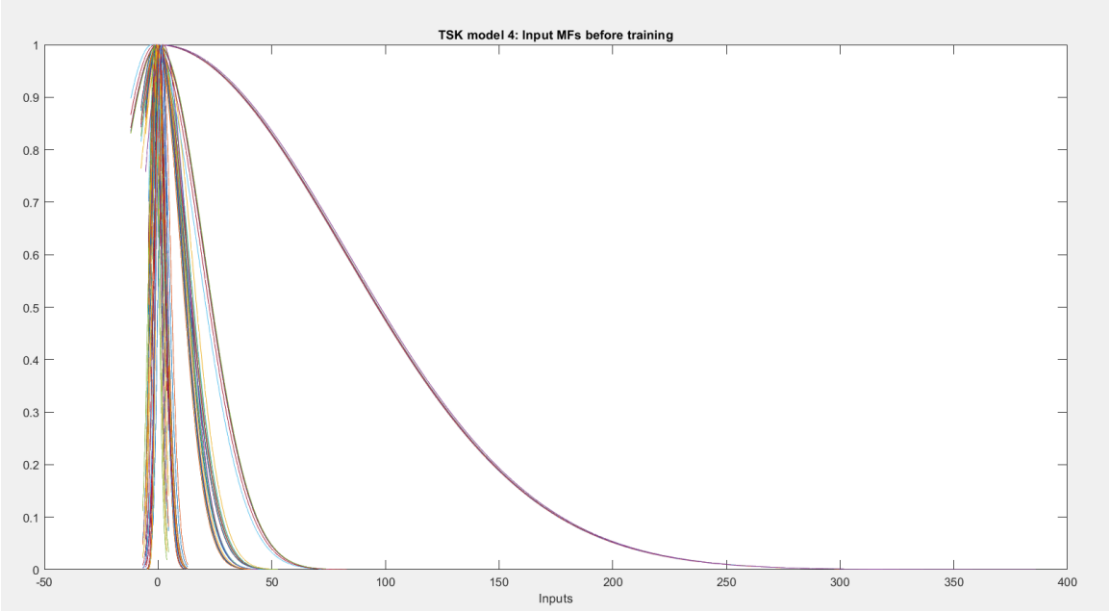
Παρακάτω φαίνεται ο πίνακας με τους ζητούμενους δείκτες απόδοσης.

Πίνακας 4.

Class Number	1	2	3	4	5	6	7	8	9	10	11	12
Producers Accuracy	0.8636	0	0.0077	0.05	0.2375	0.0874	0.0183	0.0083	0.6831	0	0.3478	0.1163
Users Accuracy	0.0443	0	0.1951	0.4397	0.3904	0.0204	0.0112	0.0048	0.2922	0	0.0766	0.0472
Overall Accuracy	0.1088					$\hat{k}$		0.05818				

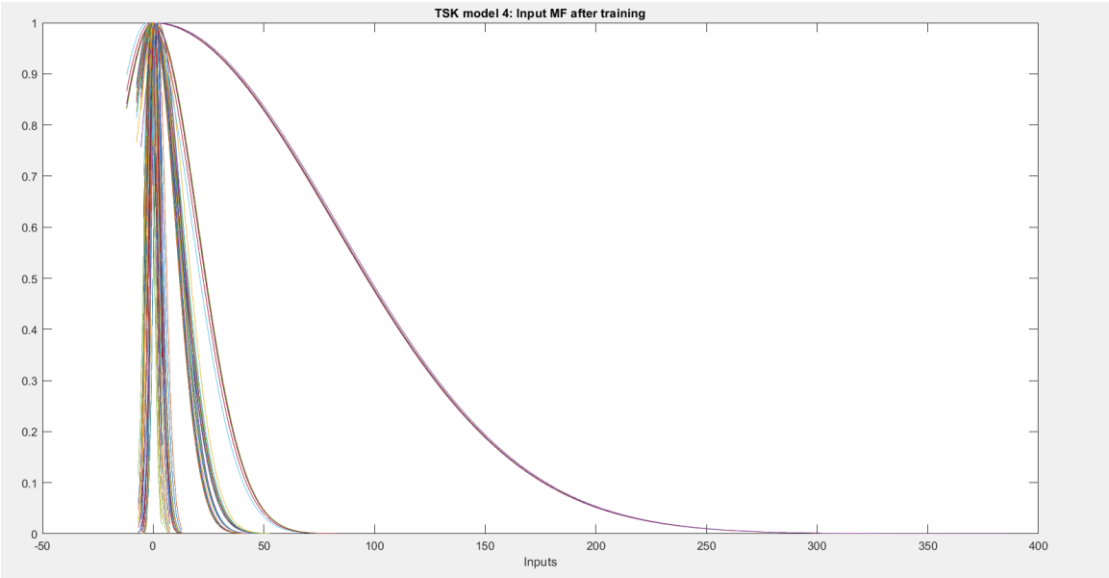
TSK Μοντέλο 4

Στο τέταρτο μοντέλο TSK οι συναρτήσεις συμμετοχής φαίνονται στο Σχήμα 16.



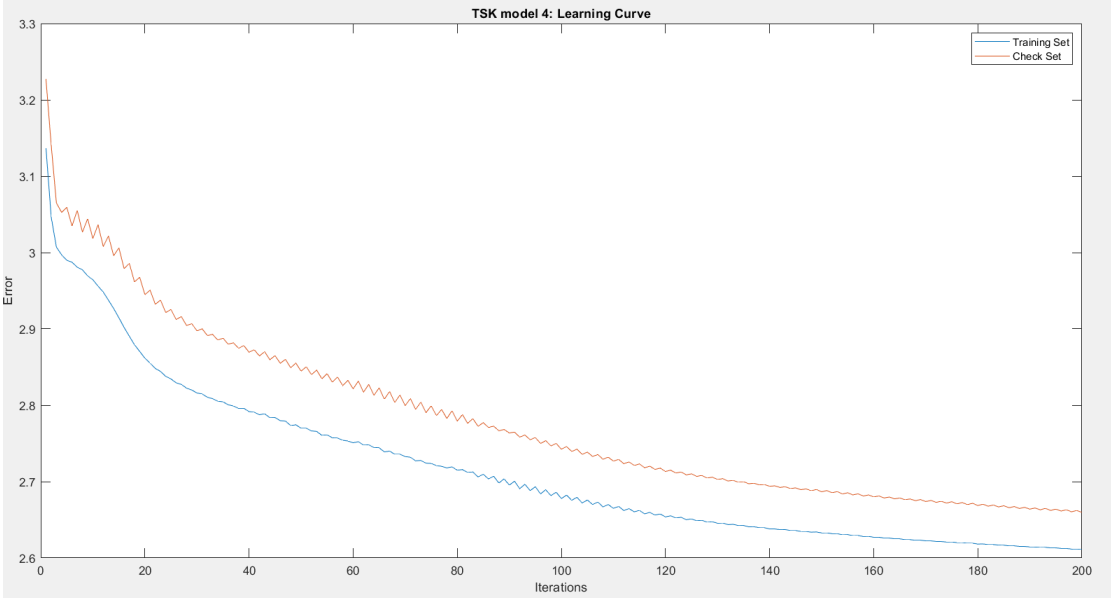
Σχήμα 16: Αρχικές Συναρτήσεις Συμμετοχής για το TSK Μοντέλο 4.

Οι συναρτήσεις συμμετοχής μετά την εκπαίδευση παίρνουν την εξής μορφή:



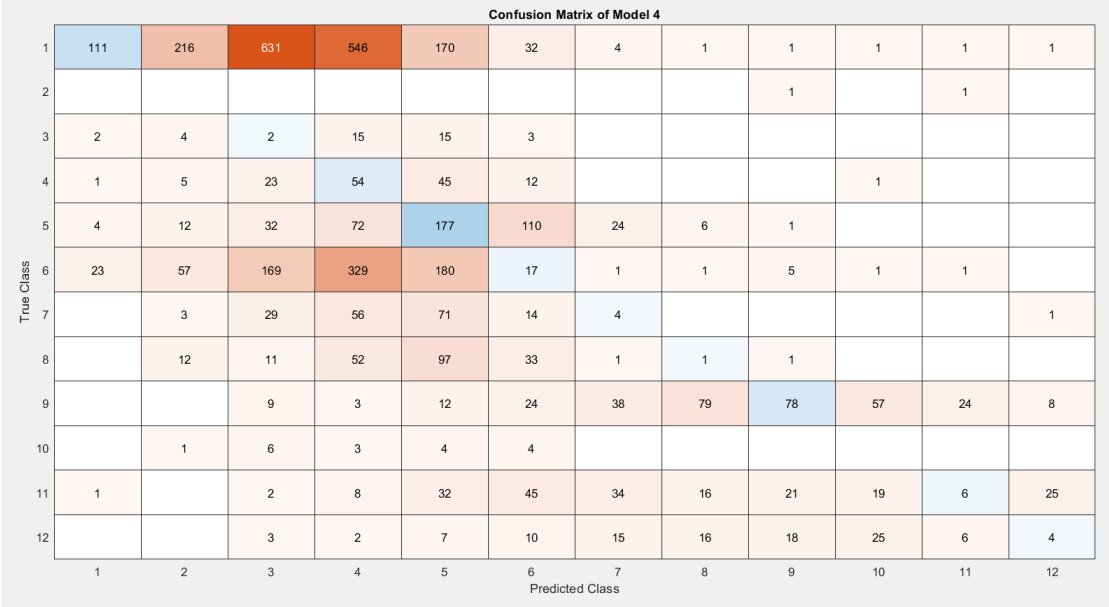
Σχήμα 17: Συναρτήσεις συμμετοχής μετά την εκπαίδευση για το TSK μοντέλο 4.

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

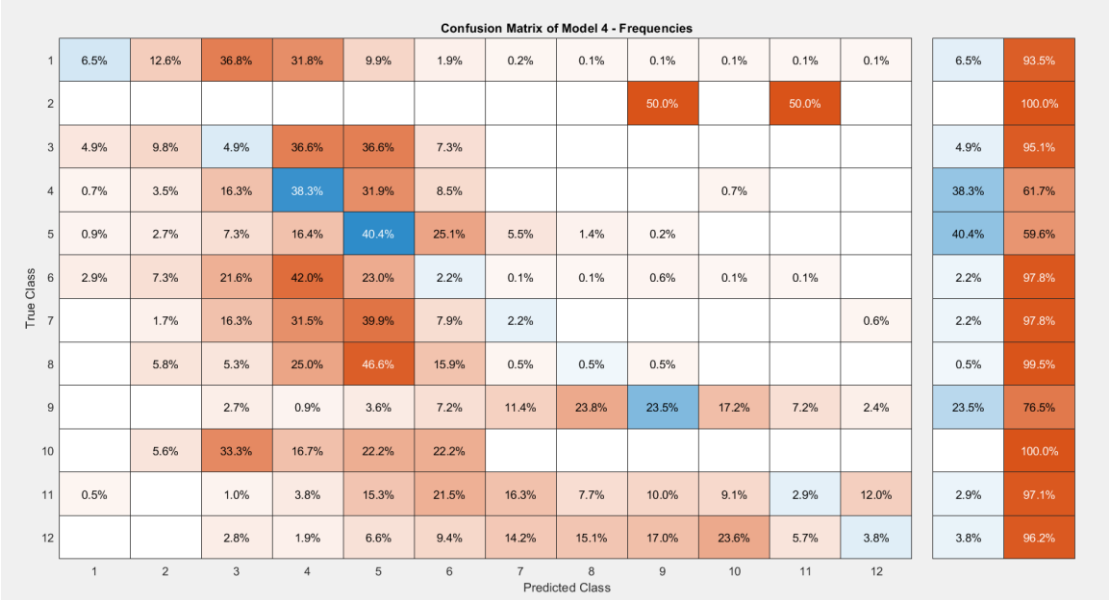


Σχήμα 18: Καμπύλες Εκμάθησης για το TSK Μοντέλο 4.

Τα σφάλματα πρόβλεψης της ταξινόμησης με τη μορφή απολύτων τιμών και συχνοτήτων φαίνονται παρακάτω.



Σχήμα 19: Σφάλματα Πρόβλεψης Ταξινόμησης για το TSK Μοντέλο 4.

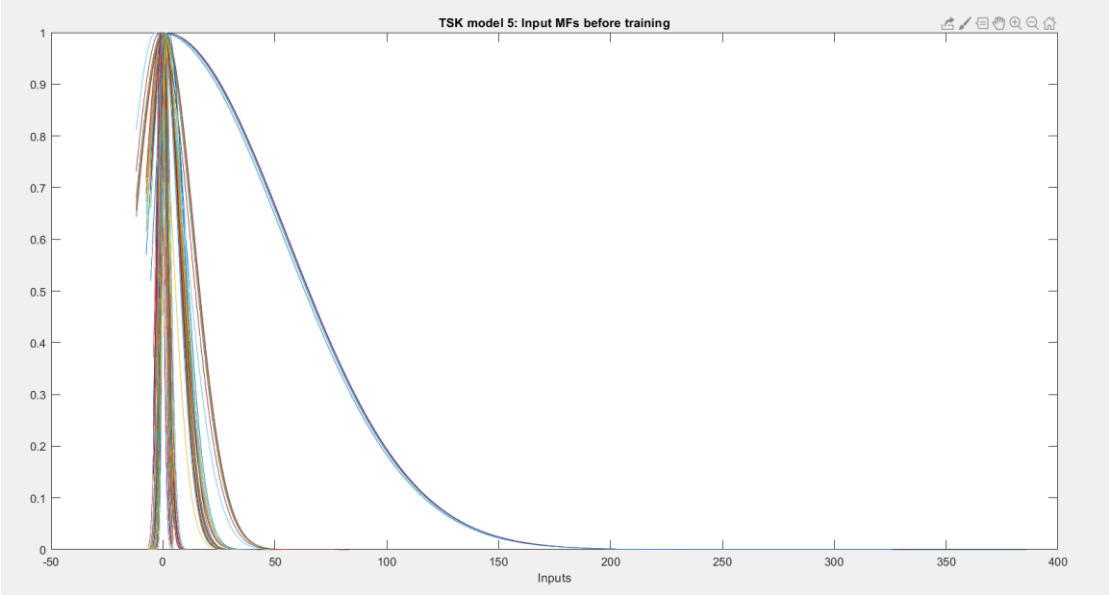


Σχήμα 20: Σφάλματα Πρόβλεψης Ταξινόμησης (Συχνότητες) για το TSK Μοντέλο 4.

Class Number	1	2	3	4	5	6	7	8	9	10	11	12
Producers Accuracy	0.7817	0	0.0022	0.0474	0.2185	0.0559	0.0331	0.0083	0.6190	0	0.1538	0.1026
Users Accuracy	0.0647	0	0.0488	0.3830	0.4041	0.0217	0.0225	0.0048	0.2349	0	0.0766	0.0472
Overall Accuracy	0.1088					$\hat{k}$		0.04647				

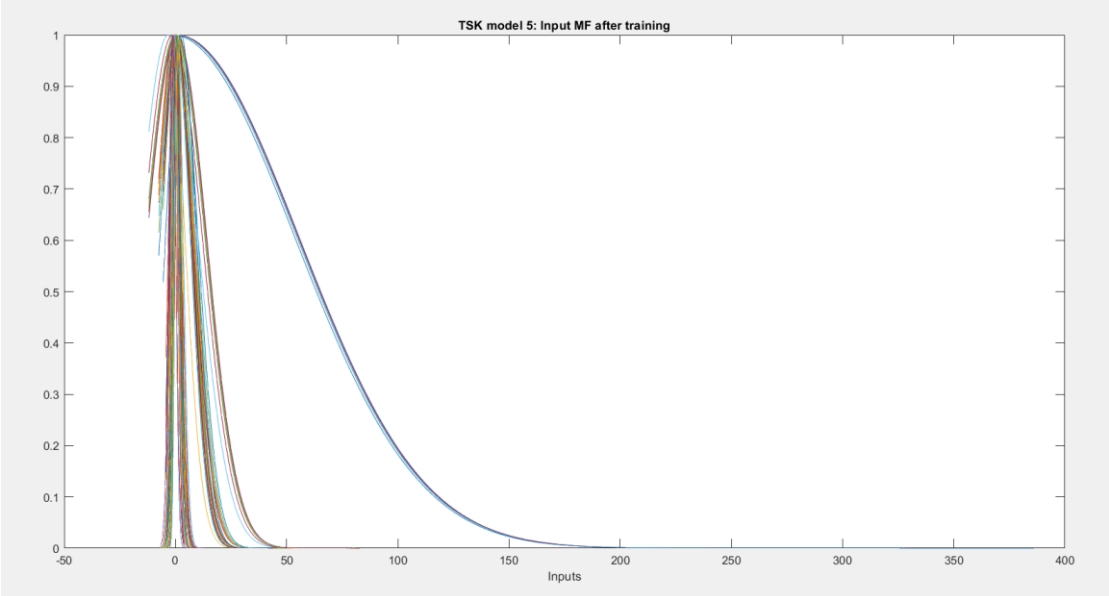
TSK Μοντέλο 5

Στο τέταρτο μοντέλο TSK οι συναρτήσεις συμμετοχής φαίνονται στο Σχήμα 21.



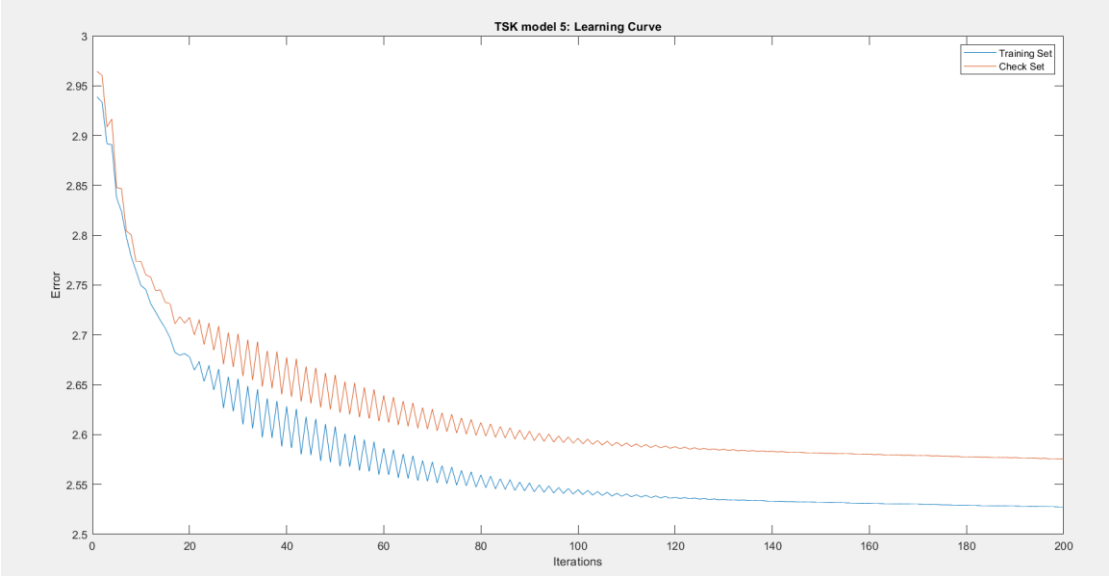
Σχήμα 21: Αρχικές Συναρτήσεις Συμμετοχής για το TSK Μοντέλο 5.

Οι συναρτήσεις συμμετοχής μετά την εκπαίδευση παίρνουν την εξής μορφή:



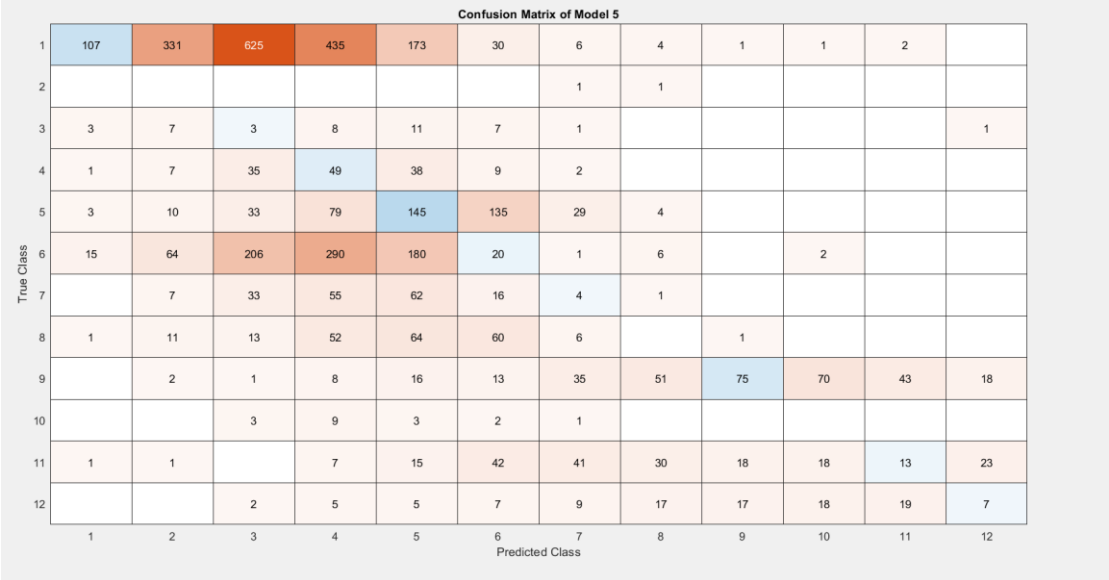
Σχήμα 22: Συναρτήσεις συμμετοχής μετά την εκπαίδευση για το TSK μοντέλο 5.

Ακολουθούν οι καμπύλες εκμάθησης στο πέρας των εποχών.

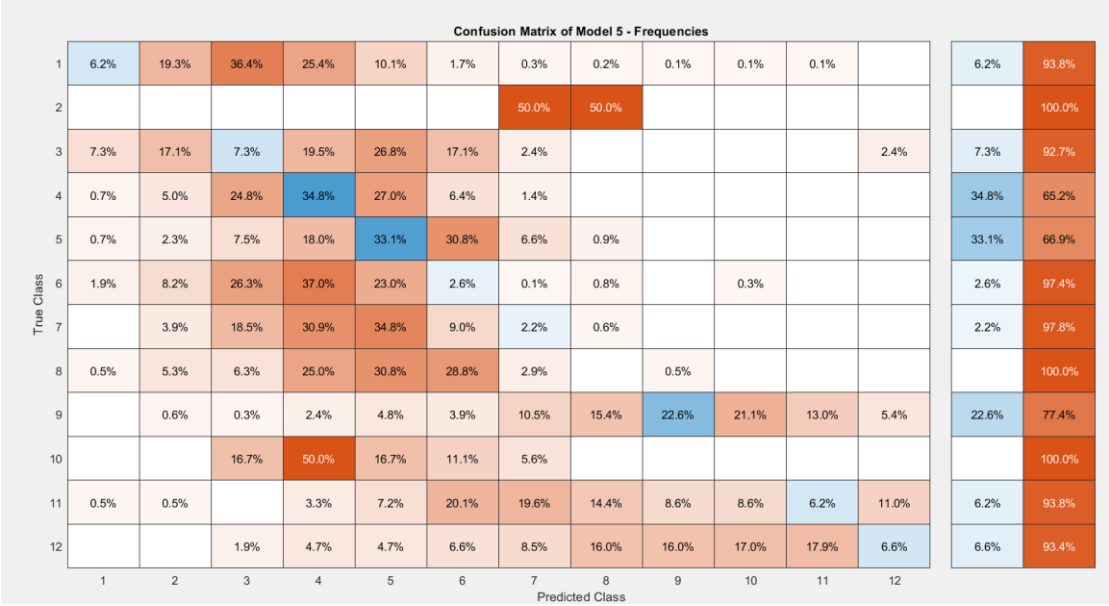


Σχήμα 23: Καμπύλες Εκμάθησης για TSK Μοντέλο 5.

Τα σφάλματα πρόβλεψης της ταξινόμησης με τη μορφή απολύτων τιμών και συχνοτήτων φαίνονται παρακάτω.



Σχήμα 24: Σφάλματα Πρόβλεψης Ταξινόμησης για το TSK Μοντέλο 5.



Σχήμα 25: Σφάλματα Πρόβλεψης Ταξινόμησης (Συχνότητες) για τοTSK Μοντέλο 5.

Παρακάτω φαίνεται ο πίνακας με τους ζητούμενους δείκτες απόδοσης.

Πίνακας 6.

Class Number	1	2	3	4	5	6	7	8	9	10	11	12
Producers Accuracy	0.8168	0	0.0031	0.0491	0.2037	0.0587	0.0294	0	0.6696	0	0.1688	0.1429
Users Accuracy	0.0624	0	0.0732	0.3475	0.3311	0.0255	0.0225	0	0.2259	0	0.0622	0.0660
Overall Accuracy	0.10139					$\hat{k}$		0.041192				

Σύνοψη και Συμπεράσματα μοντέλων

Τα πρώτα δύο μοντέλα με 4 και 8 κανόνες παρουσιάζουν τη μεγαλύτερη συνολική ακρίβεια με 12.66 % και $\hat{k}$ 6.88%. Τα αποτελέσματα δεν είναι ικανοποιητικά και λόγος οφείλεται στην ανισορροπία των δεδομένων (Class Imbalance), δηλαδή στο μη δίκαιο μοίρασμα των δεδομένων σε κάθε κλάση. Σύμφωνα με το σχήμα 21, στη κλάση 1 ανήκει κάτι λιγότερο από το μισό του συνόλου των δεδομένων, ενώ άλλες κλάσεις όπως η 2 έχουν ένα ποσοστό κάτω του 1%. Το σύνολο των δεδομένων 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. Τέλος, σημειώνεται ότι ο συνολικός χρόνος εκπαίδευσης ήταν 554 δευτερόλεπτα ή αλλιώς 9 λεπτά και 13 δευτερόλεπτα.

classes_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 26: Τα διάφορα σετ δεδομένων που δημιουργήθηκαν.

Εφαρμογή σε σύνολο δεδομένων με υψηλή διαστασιμότητα

Isolet Dataset

Το dasaset στο δεύτερο μέρος της εργασίας είναι το Isolet Dataset από το UCI repository. Το συγκεκριμένο dataset, περιλαμβάνει 7797 δείγματα, καθένα από τα οποία περιγράφεται από 618 μεταβλητές/χαρακτηριστικά. Το datset περιλαμβάνει 617 χαρακτηριστικά και μία μεταβλητή κλάσης από το 1 μέχρι το 26. Στόχος του συγκεκριμένου dataset είναι η πρόβλεψη του γράμματος που ειπώθηκε μέσω ταξινόμησης.

Περισσότερες πληροφορίες: https://archive.ics.uci.edu/ml/datasets/isolet

Είναι φανερό ότι το μέγεθος του dataset καθιστά απαγορευτική μια απλή εφαρμογή ενός TSK μοντέλου, σαν αυτή του προηγούμενου μέρους της εργασίας. Ο μεγάλος αριθμός μεταβλητών καθιστά αναγκαία τη χρήση μεθόδων μείωσης της διαστασιμότητας καθώς και του αριθμού των IF-THEN κανόνων. Ο στόχος αυτός θα επιτευχθεί μέσω της επιλογής χαρακτηριστικών, με χρήση του αλγορίθμου Relief, και της χρήσης ασαφούς ομαδοποίησης.

Προετοιμασία του Σετ Δεδομένων

Αρχικά, το σετ δεδομένων ταξινομείται κατά αύξουσα σειρά με βάση τις τιμές των κλάσεων και μετριέται η συχνότητα εμφάνισης της καθεμίας. Στη συνέχεια, πραγματοποιείται διαχωρισμός του σετ δεδομένων σε τρία μη επικαλυπτόμενα υποσύνολα με τρόπο τέτοιο ώστε οι παραπάνω συχνότητες εμφάνισης να διατηρούνται περίπου σταθερές και τα τρία σετ ανακατεύονται. Ο διαχωρισμός γίνεται ως εξής:

- 60% : σύνολο εκπαίδευσης – training data
- 20% : σύνολο επικύρωσης – validation data
- 20% : σύνολο ελέγχου – check data

Στην συνέχεια, πραγματοποιείται ένας έλεγχος των δεδομένων για διπλότυπα δείγματα και κενές τιμές. Μόλις ολοκληρωθεί ο έλεγχος εφαρμόζεται ο αλγόριθμος Relief επιλέγοντας ως αριθμό γειτόνων το 50.

Εύρεση Βέλτιστου Μοντέλου

Για την εύρεση του βέλτιστου μοντέλου (εύρεση αριθμού χαρακτηριστικών και κανόνων) χρησιμοποιείται συνδυαστικά η μέθοδος του Grid Search και του 5-Fold Cross Validation .

Το 5-Fold Cross Validation (5-πτυχη διασταυρωμένη επικύρωση) πραγματοποιείται με τη χρήση της cvpartition του MATLAB. Πιο συγκεκριμένα, αρχικά το training set δεδομένων χωρίζεται σε δύο νέα τμήματα, ένα νέο training set δεδομένων, 80% του αρχικού training set εκπαίδευσης, και ένα νέο validation set δεδομένων, 20% του αρχικού training set. Συνολικά η διαδικασία επαναλαμβάνεται πέντε φορές με διαφορετική

αναδιάταξη των δεδομένων κάθε φορά δημιουργώντας τελικά πέντε νέα δευτερεύοντα μοντέλα. Στη συνέχεια, κάθε ένα από τα δευτερεύοντα αυτά μοντέλα εκπαιδεύεται και έπειτα υπολογίζεται το μέσο τετραγωνικό σφάλμα MSE. Τέλος, υπολογίζεται η μέση τιμή των προηγουμένως υπολογισμένων σφαλμάτων, η οποία αποτελεί αντιπροσωπευτικό δείγμα του πραγματικού σφάλματος για το συνολικό κύριο μοντέλο.

Η διαδικασία αυτή εκτελείται συνδυαστικά με τη Grid Search μέθοδο. Η συνδυαστική επαναληπτική αυτή διαδικασία χρησιμοποιώντας το 5-Fold Cross Validation, εξετάζει διαφορετικά μοντέλα ως προς τον αριθμό των κανόνων (IF-THEN) και το πλήθος των χαρακτηριστικών. Το βέλτιστο μοντέλο είναι εκείνο με 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:

Δοκιμή :

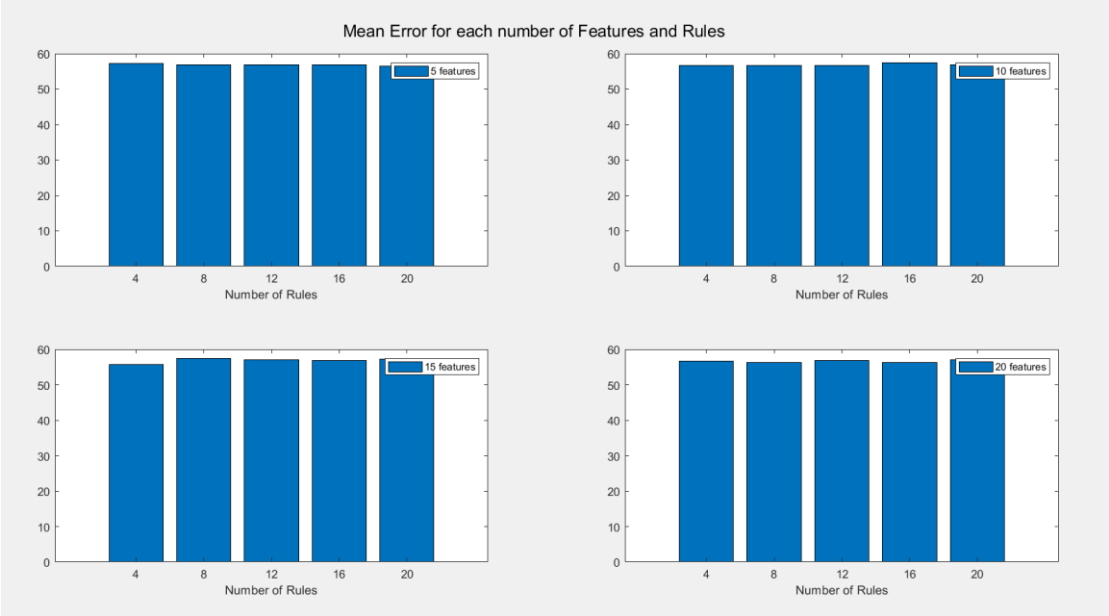
NF x NR = {5, 10, 15, 20} x {4, 8, 12, 16, 20}

Η δοκιμή αποτελείται από 20 κύρια μοντέλα. Η εκπαίδευση των κύριων μοντέλων έγινε με βάση 5 δευτερεύοντα (5 Fold Validation). Όλα τα κύρια μοντέλα (σύνολο είκοσι) , το καθένα ξεχωριστά δέχεται εκπαίδευση, με βάση τα 5 δευτερεύοντα μοντέλα (σύνολο εκατό) για 400 εποχές το καθένα, και υπολογίζεται το σφάλμα αυτών. Τέλος, υπολογίζεται ο μέσος όρος αυτών των 5 σφαλμάτων που αποτελεί και τη μετρική σύγκρισης για την εύρεση του βέλτιστου μοντέλου.

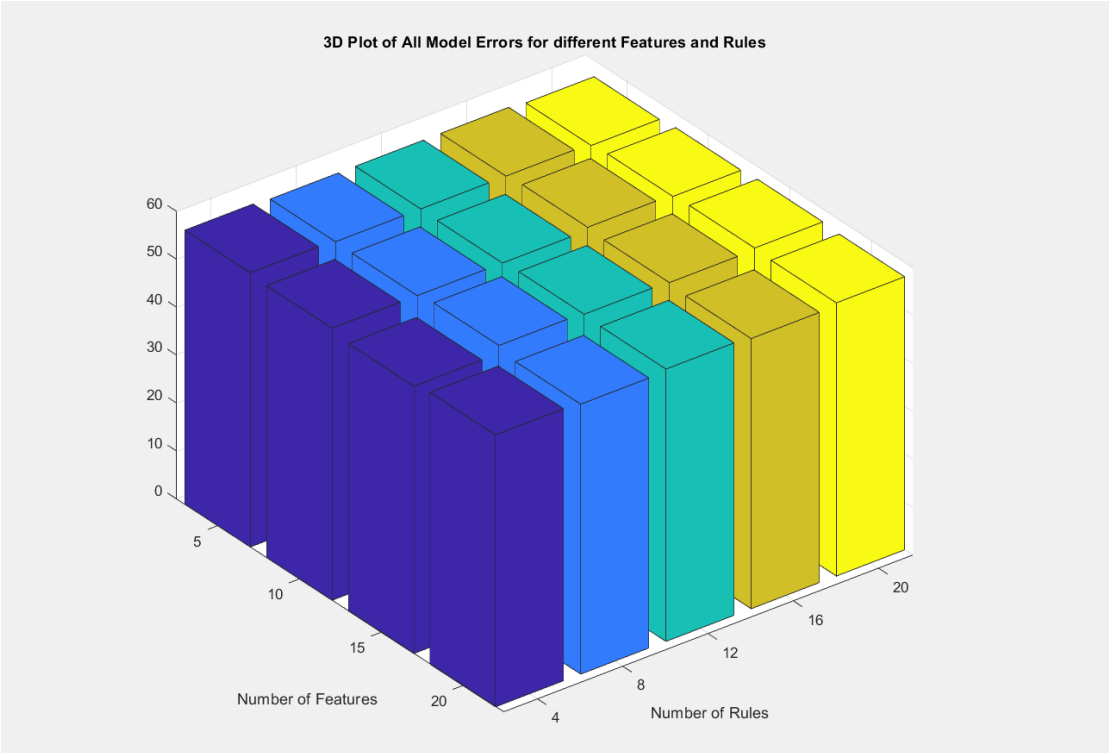
Η δοκιμή διήρκησε συνολικά 1 ώρα, 35 λεπτά και 24 δευτερόλεπτα (5723.55 seconds). Στον παρακάτω πίνακα παρουσιάζονται τα μέσα σφάλματα των 5 δευτερευόντων μοντέλων για κάθε ένα από τα 20 κύρια μοντέλα, ενώ στα σχήματα 22 και 23 φαίνονται οι τιμές του μέσου σφάλματος για το διαφορετικό αριθμό κανόνων και χαρακτηριστικών.

Πίνακας 7.

Number of Rules / Number of Features	4 Rules	8 Rules	12 Rules	16 Rules	20 Rules
5 Features	57.2041	56.8619	56.8815	56.9003	56.4623
10 Features	56.7571	56.7242	56.7118	57.3510	56.9521
15 Features	55.7736	57.4558	57.1386	56.9226	57.3127
20 Features	56.7381	56.3214	56.8210	56.3617	56.9858



Σχήμα 27: Μέσο σφάλμα μοντέλων για τις διάφορες τιμές χαρακτηριστικών και κανόνων.



Σχήμα 28: Κοινό 3D Διάγραμμα Μέσου Σφάλματος των διάφορων μοντέλων.

Με βάση τα σφάλματα το βέλτιστο μοντέλο είναι το μοντέλο 11, αυτό με τα 15 χαρακτηριστικά και τους 4 κανόνες. Για τα μοντέλα που δοκιμάστηκαν ισχύει ότι όσο αυξάνεται η πολυπλοκότητα των μοντέλων, ο χρόνος εκτέλεσης αυξάνεται κατά πολύ, ενώ το σφάλμα δεν μειώνεται αντίστοιχα σε τόσο μεγάλο βαθμό, ενώ μερικές φορές αυξάνεται κιόλας. Τέλος, επισημαίνεται ότι ο διαμοιρασμός στα διάφορα set γίνεται με

τέτοιο τρόπο, ώστε να περιέχουν ίσο σε ποσοστό αριθμό δεδομένων από κάθε κλάση, όπως φαίνεται παρακάτω στο σχήμα 24.

classes_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 29: Τα διάφορα σετ δεδομένων που δημιουργήθηκαν.

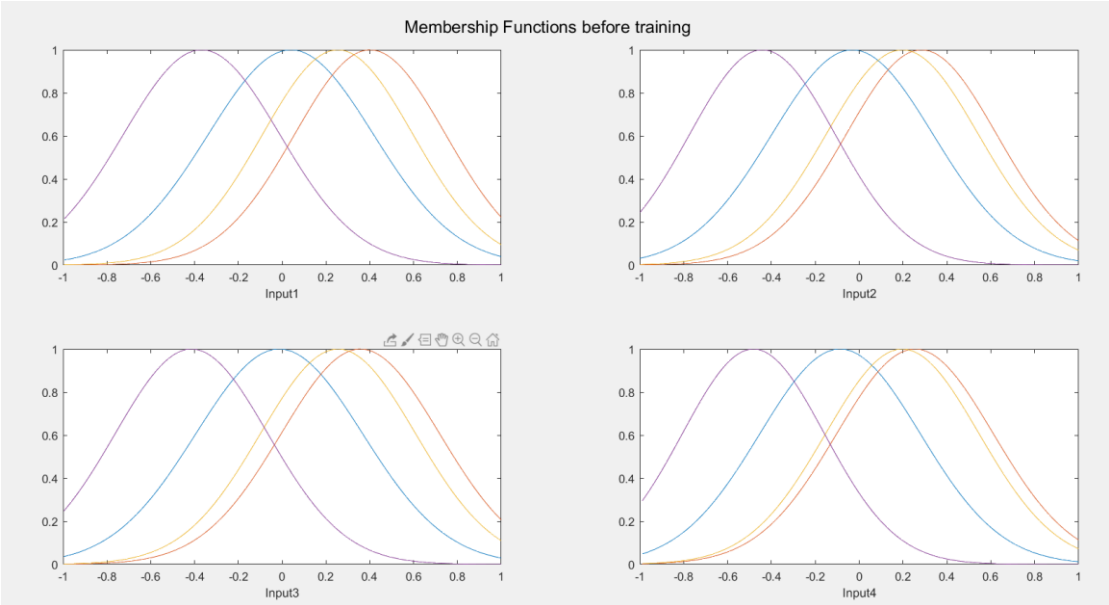
<u>Σημείωση</u>:

Οι προσομοιώσεις πραγματοποιήθηκαν σε laptop με

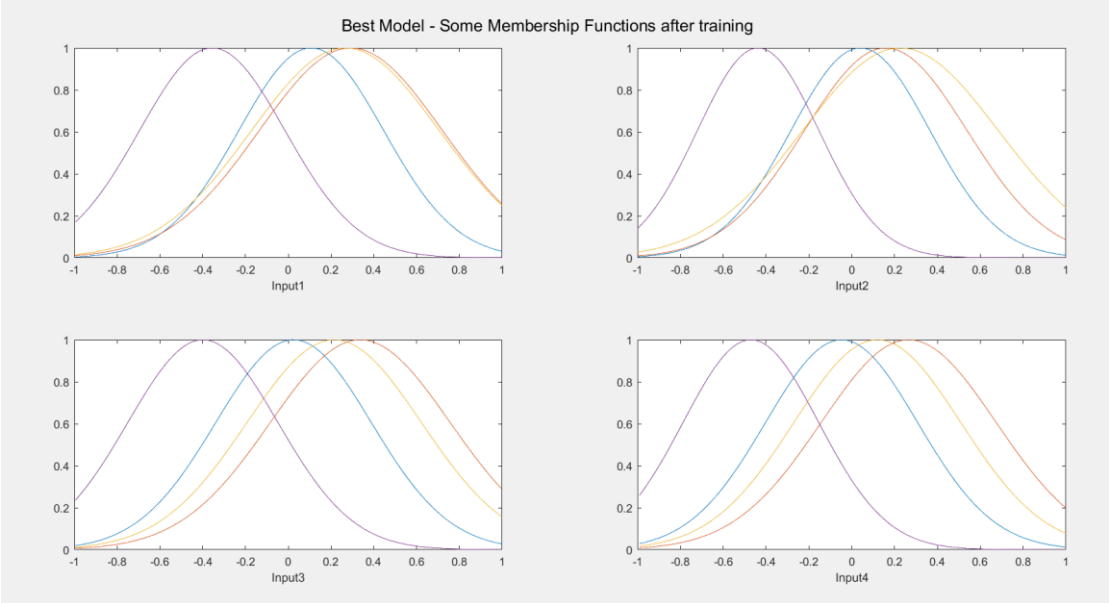
- Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz
- Radeon (TM) R5 M420 & Intel(R) HD Graphics 620
- 4.0 GB RAM

Εκπαίδευση τελικού TSK μοντέλου

Όσον αφορά το τελικό μοντέλο, αρχικά παρουσιάζονται οι συναρτήσεις συμμετοχής πριν την εκπαίδευση, οι οποίες είναι οι εξής:

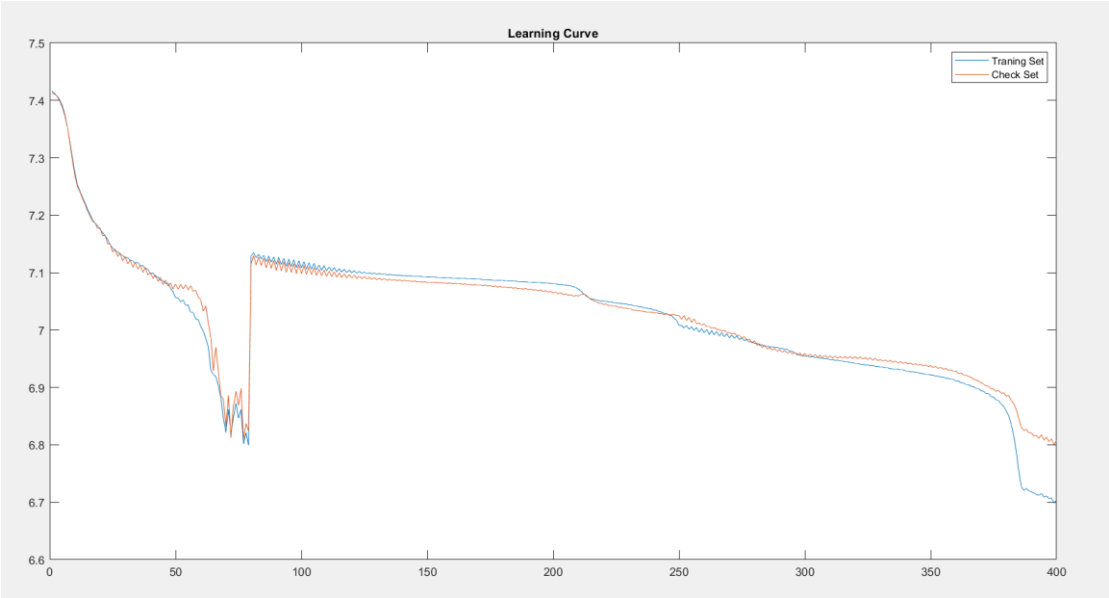


Σχήμα 30: Τελικό TSK model - Συναρτήσεις συμμετοχής πριν την εκπαίδευση.



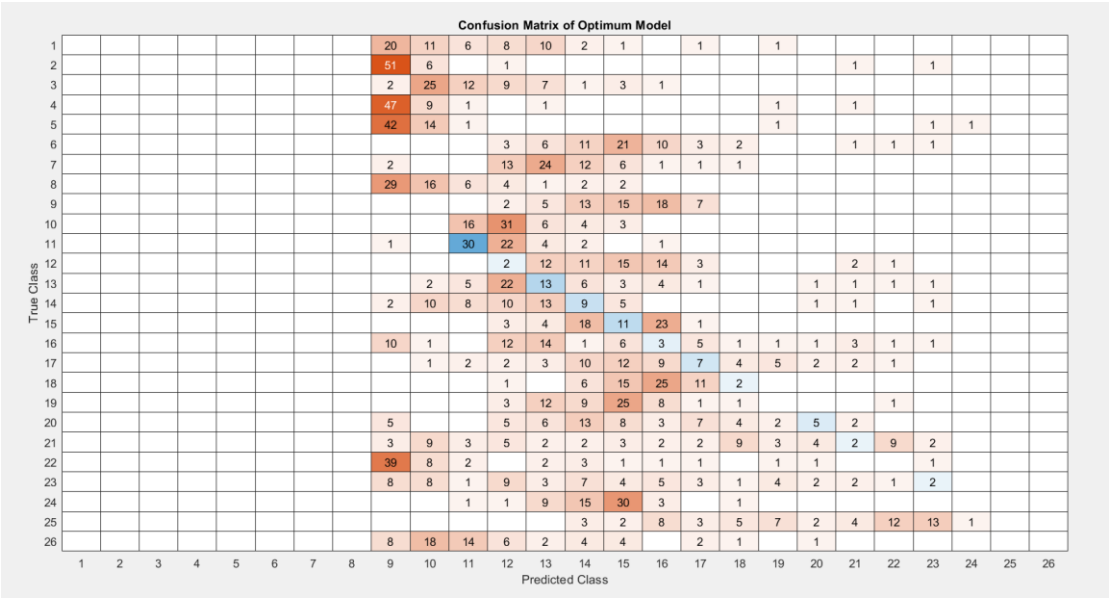
Σχήμα 31: Τελικό TSK model - Συναρτήσεις συμμετοχής μετά από εκπαίδευση 400 εποχών.

Η καμπύλη εκμάθησης και τα σφάλματα πρόβλεψης φαίνονται στα παρακάτω διαγράμματα.

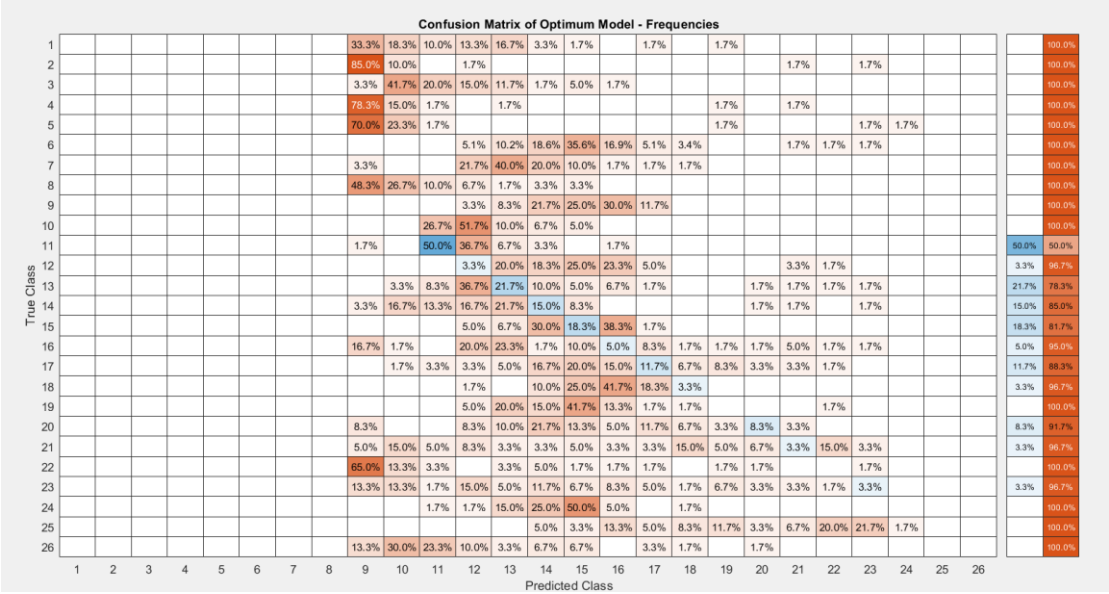


Σχήμα 32: Τελικό TSK model - Καμπύλη εκμάθησης

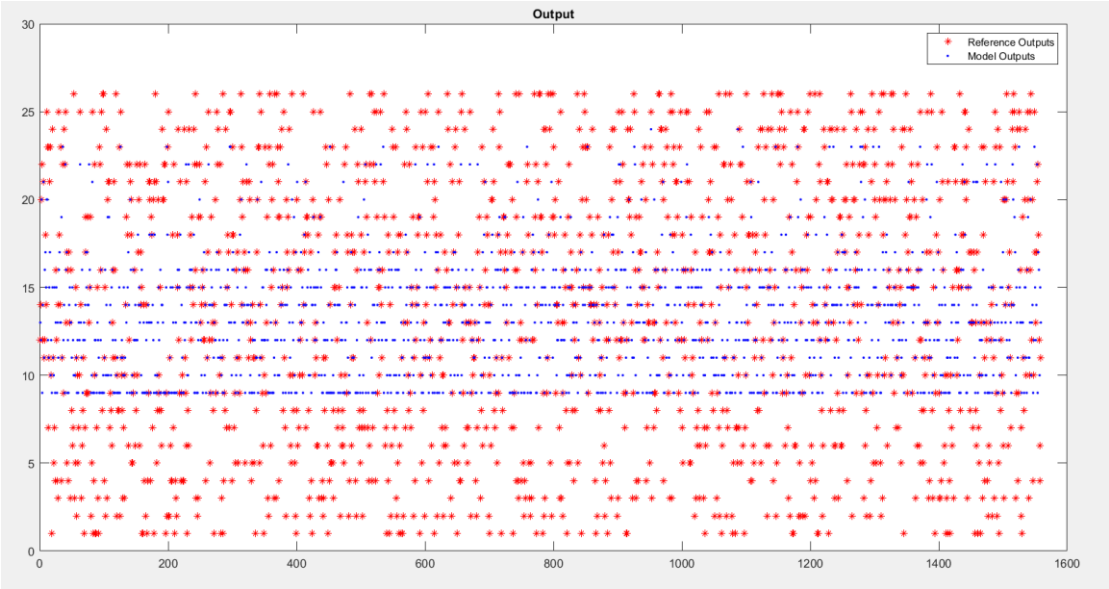
Τα σφάλματα πρόβλεψης της ταξινόμησης με τη μορφή απολύτων τιμών και συχνοτήτων φαίνονται παρακάτω.



Σχήμα 33: Σφάλματα Πρόβλεψης Ταξινόμησης για το βέλτιστο TSK.



Σχήμα 34: Σφάλματα Πρόβλεψης Ταξινόμησης (Συχνότητες) για το βέλτιστο TSK.



Σχήμα 35: Πραγματική και Εκτιμήτρια Έξοδος για το βέλτιστο TSK.

Στον παρακάτω πίνακα φαίνονται οι δείκτες απόδοσης και ο χρόνος εκτέλεσης για την εκπαίδευση και αξιολόγηση του βέλτιστου μοντέλου.

Πίνακας 8.

Class Number	1	2	3	4	5	6	7	8	9	10	11	12	13
Producers Accuracy	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0	0	0.5000	0.0333	0.2167
Users Accuracy	0	0	0	0	0	0	0	0	0	0	0.5000	0.0333	0.2167
Class Number	14	15	16	17	18	19	20	21	22	23	24	25	26
Producers Accuracy	0.150	0.1833	0.0500	0.1167	0.0333	0	0.0833	0.0333	0	0.0333	0	0	0
Users Accuracy	0.1500	0.1833	0.0500	0.1167	0.0333	0	0.0833	0.0333	0	0.0333	0	0	0
Overall Accuracy	0.0552			$\hat{k}$		0.0173		Elapsed Time for Algorithm Execution			72.977976 seconds		

Από τον πίνακα 8 φαίνεται ότι το βέλτιστο μοντέλο παρουσιάζει μη ικανοποιητικά αποτελέσματα. Μετά το πέρας των 100 εποχών τα σφάλματα μεγαλώνουν απότομα και στην συνέχεια μέχρι το πέρας των 400 εποχών μειώνονται. Το Overall Acurracy είναι ίσο με μόλις 15.84% και ο δείκτης $\hat{k}$ ίσος με 0.0173.

Το πρόβλημα των μη ικανοποιητικών αποτελεσμάτων οφείλεται κατά βάση στο γεγονός ότι το πλήθος το δεδομένων είναι μικρό συγκριτικά με τον μεγάλο αριθμό του πλήθους των κλάσεων. Ενδεχομένως, τα αποτελέσματα να βελτιωνόντουσαν αν το μοντέλο υποβάλλονταν σε διαφορετική μέθοδο εκπαίδευσης, το οποίο όμως δεν είναι στο πλαίσιο της παρούσας εργασίας.

Παρακάτω παρουσιάζονται τα διάφορα σετ που δημιουργήθηκαν.

classes_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 36: Τα διάφορα σετ δεδομένων που δημιουργήθηκαν.

Επεξήγηση παραδοτέων αρχείων MATLAB

- **task1.m** : MATLAB Script - Εκπαίδευση και αξιολόγηση των πέντε TSK models.
- **grid_search.m:** MATLAB Script - Επιλογή των βέλτιστων παραμέτρων
- **final_tsk_model.m:** MATLAB Script - Εκπαίδευση και αξιολόγηση του τελικού μοντέλου