ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

# *Ασαφή Συστήματα*

## *Εργασία #3 ~ Regression*

Δημήτρης Παππάς

ΑΕΜ: 8391

e-mail: dspappas@ece.auth.gr

# Περιεχόμενα

# 3. Εργασία #3: Regression

## 3.1 Part 1 - Εφαρμογή σε απλό dataset

Σε αυτή την εργασία μελετάμε την ικανότητα των TSK μοντέλων (Takagi Sugeno Kang) στην μοντελοποίηση μοντέλων με πολλές μεταβλητές.

Δημιουργούμε 4 μοντέλα:

| | Πλήθος συναρτήσεων συμμετοχής | Μορφή εξόδου |
|---|---|---|
| TSK_model_1 | 2 | Singleton |
| TSK_model_2 | 3 | Singleton |
| TSK_model_3 | 2 | Polynomial |
| TSK_model_4 | 3 | Polynomial |

Σχήμα 64: TSK μοντέλα

Το dataset που θα χρησιμοποιήσουμε είναι το “Airfoil Self-Noise” από το UCI repository, το οποίο περιέχει 1503 δεδομένα και αποτελείται από 5 features και 1 output.

Για κάθε ένα από τα μοντέλα υπολογίζουμε τα σφάλματα RMSE, NMSE, NDEI, R2, με τη συνάρτηση “calculate.m”, ώστε να αποφασίσουμε ποιό είναι το βέλτιστο.

Χωρίζουμε το dataset σε τρία υποσύνολα, 60% training data, 20% validation data, 20% checking data.

Για την εκπαίδευση του μοντέλου χρησιμοπούμε τη συνάρτηση anfis και επιλέγουμε EpochNumber = 200.

Επίσης, δημιουργούμε διαγράμματα για τις τελικές εκβάσεις των μοντέλων και των Συναστήσεων Συμμετοχής, για την καμπύλη εκμάθησης και για το σφάλμα πρόβλεψης.

Οι είσοδοι των μοντέλων είναι:

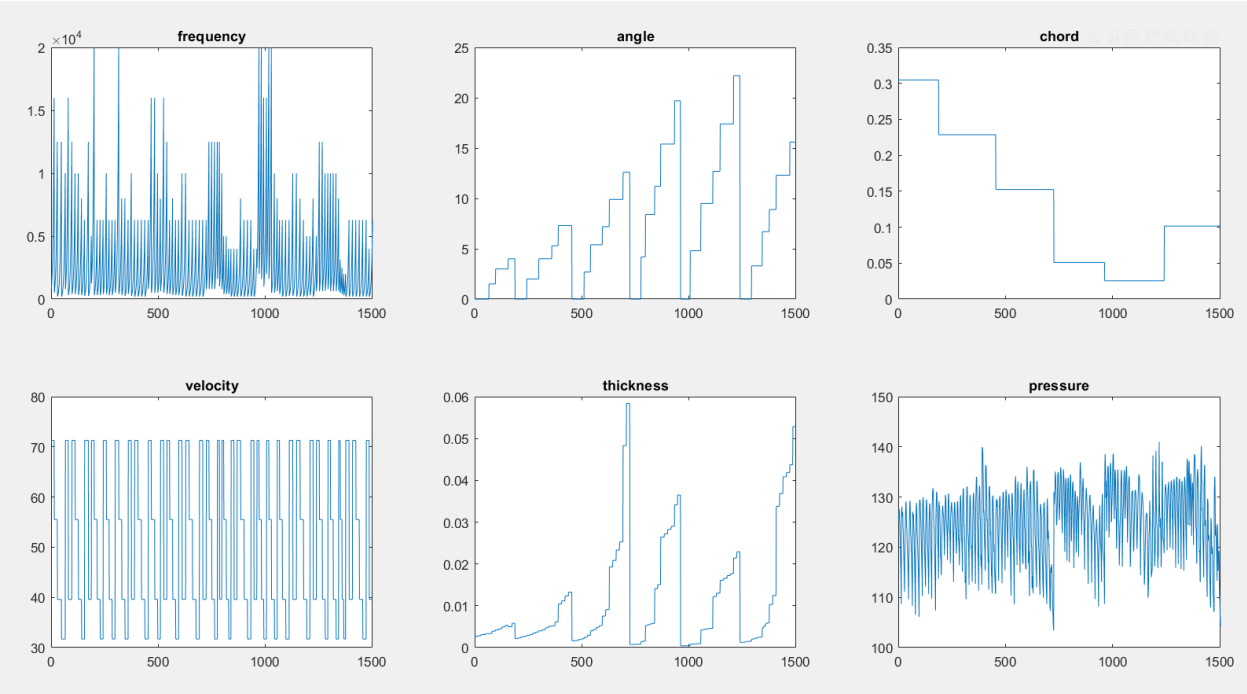
- frequency (Hz)
- angle of attack (degrees)
- chord length (m)
- free-stream velocity (m/s)

- suction side displacement thickness (m)

ενώ η έξοδος είναι :

- scaled sound pressure level (dB)

Παρακάτω φαίνονται οι τιμές των εισόδων και της εξόδου.



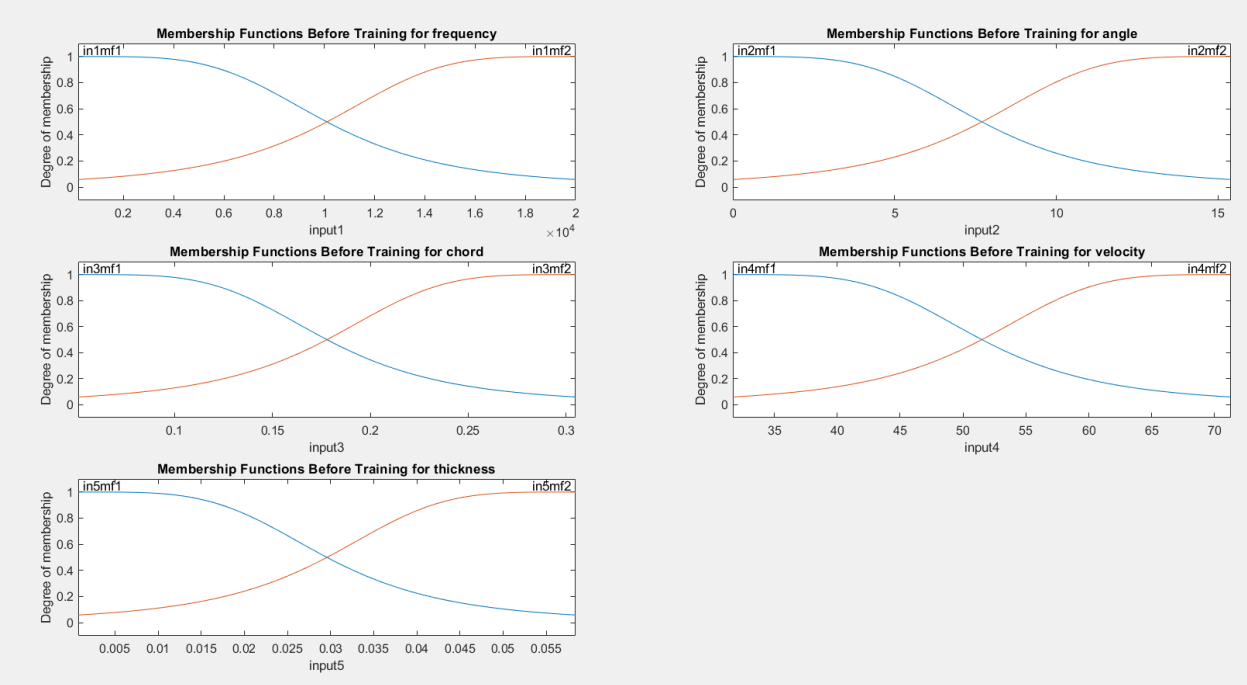

Σχήμα 65: Input & Output Values

### 3.1.1 TSK model 1

Το μοντέλο 1 υλοποιείται με το script “Regression_TSK_model_1.m”.

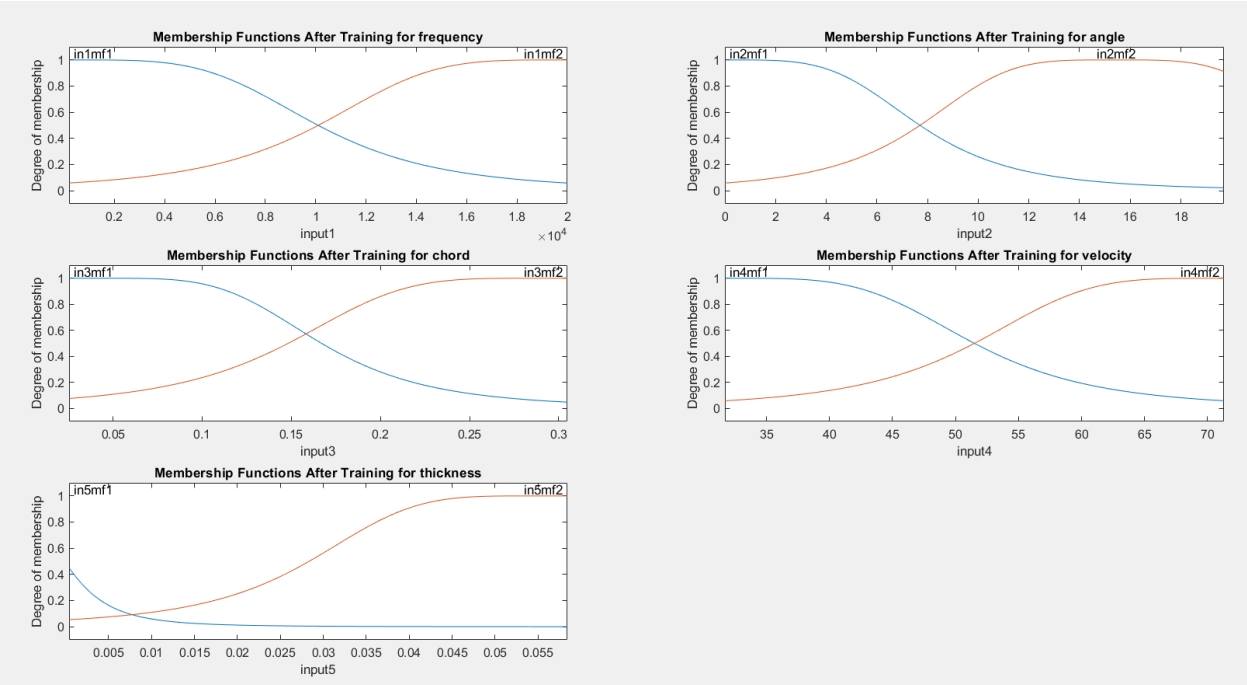
Χρησιμοποιεί μέθοδο ομαδοποίησης grid partition, 2 συναρτήσεις συμμετοχής “gbellmf” και “Singleton” output (constant).

## Συναρτήσεις Συμμετοχής Πριν την Εκπαίδευση



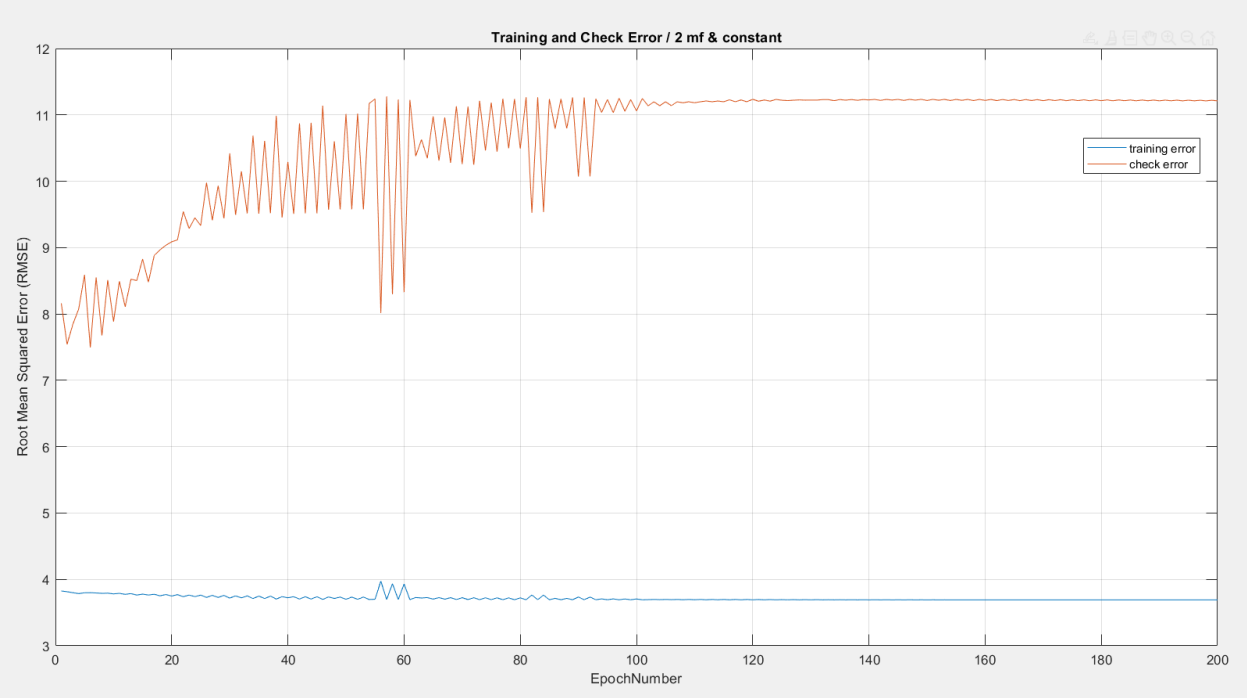

Σχήμα 66: mf Before Training

## Συναρτήσεις Συμμετοχής Μετά την Εκπαίδευση



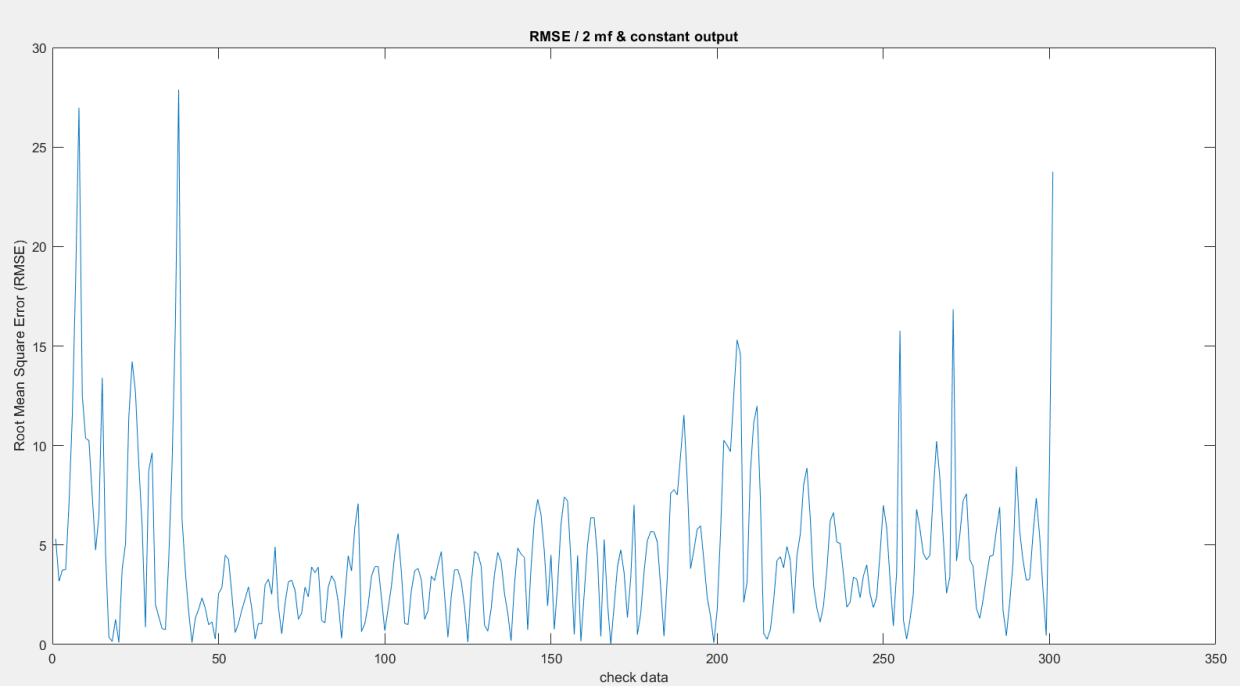

Σχήμα 67: mf After Training

## Καμπύλη Εκμάθησης



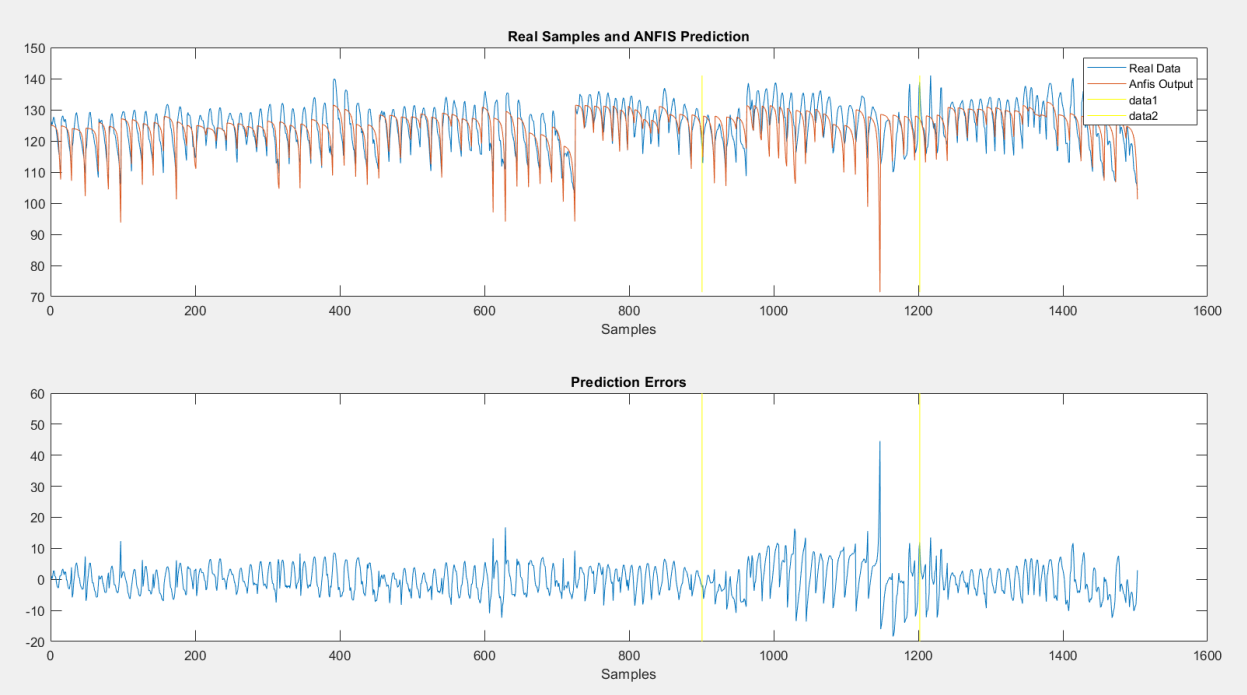

Σχήμα 68: Learning Curve

## RMSE




Σχήμα 69: RMSE

Πραγματικές Τιμές VS Τιμές Πρόβλεψης & Σφάλμα Πρόβλεψης




Σχήμα 70: Real VS Prediction Values & Prediction Error

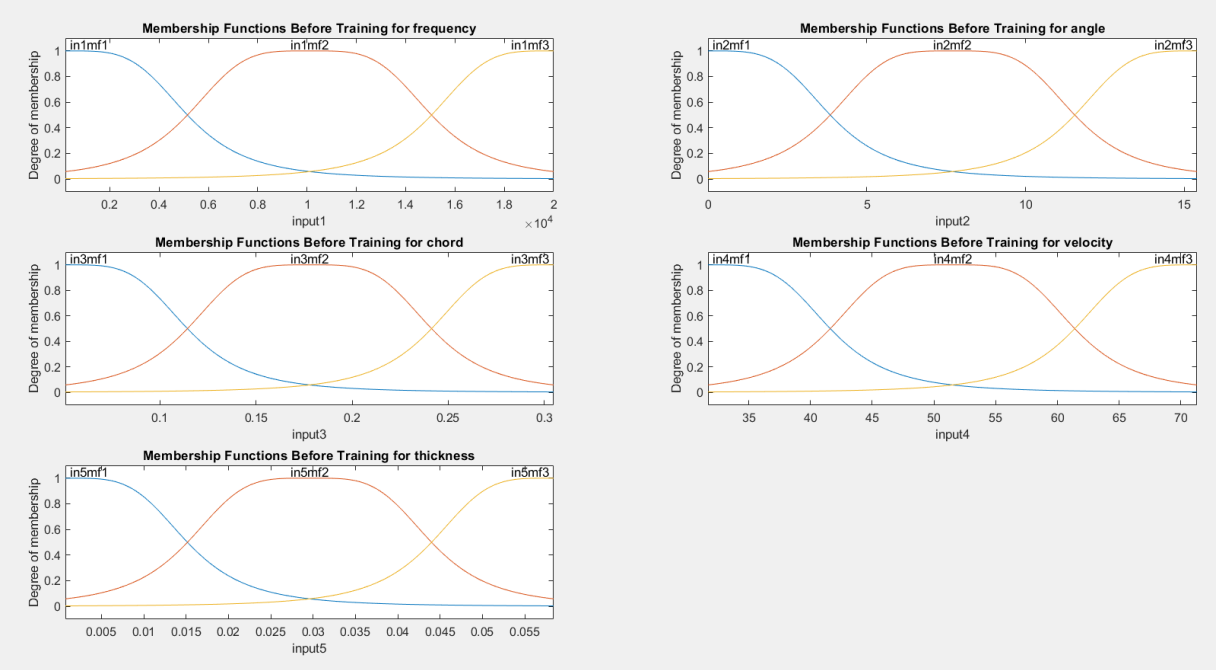
| Error | RMSE | NMSE | NDEI | R2 |
|---|---|---|---|---|
| TSK model 1 | 5.9544 | 0.6277 | 0.7923 | 0.3723 |

### 3.1.2 TSK model 2

Το μοντέλο 2 υλοποιείται με το script “Regression_TSK_model_2.m”.

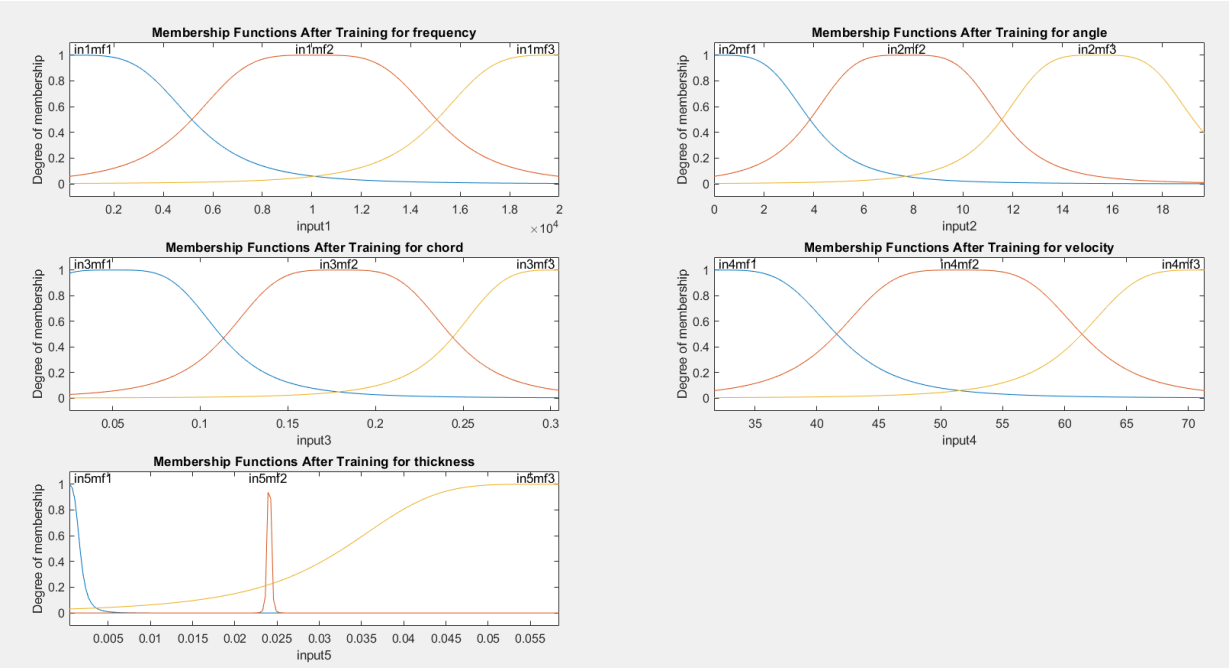
Χρησιμοποιεί μέθοδο ομαδοποίησης grid partition, 3 συναρτήσεις συμμετοχής “gbellmf” και “Singleton” output (constant).

## Συναρτήσεις Συμμετοχής Πριν την Εκπαίδευση



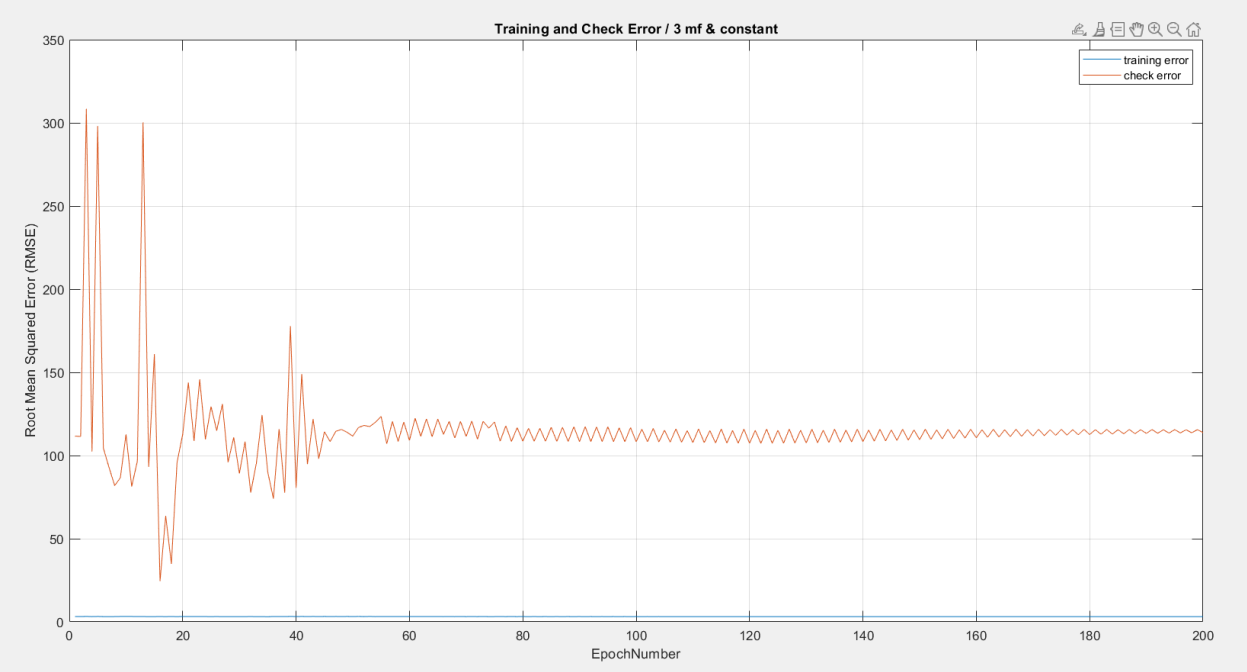

Σχήμα 71: mf Before Training

## Συναρτήσεις Συμμετοχής Μετά την Εκπαίδευση



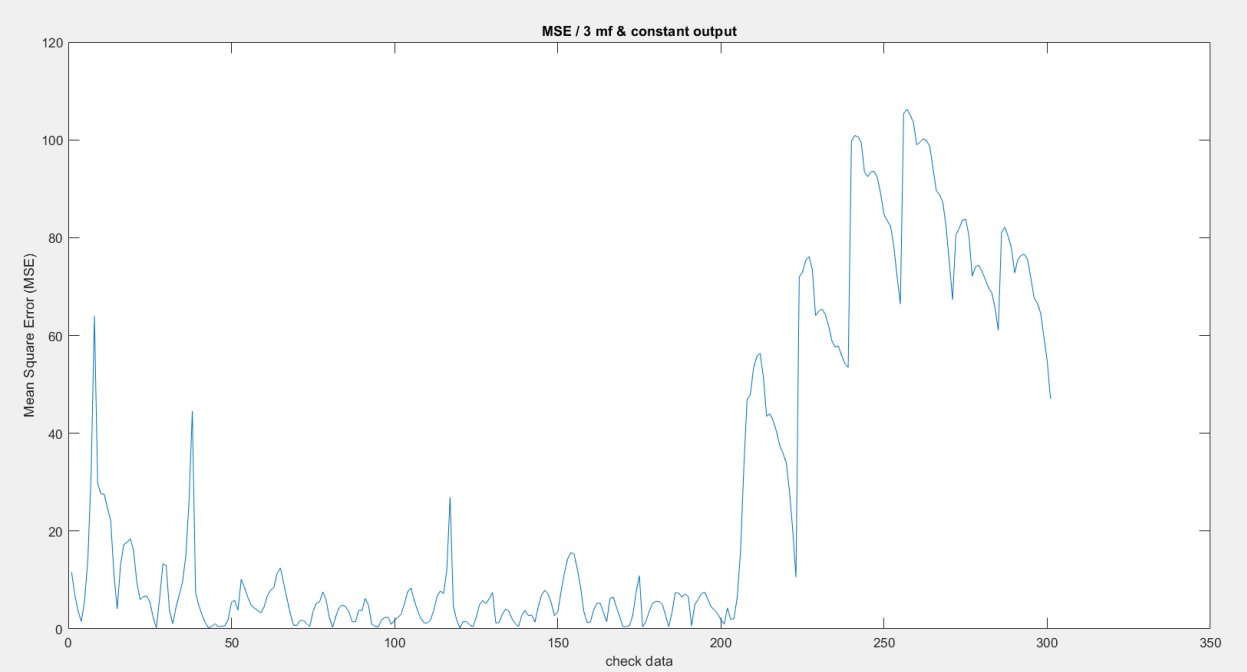

Σχήμα 72: mf After Training

## Καμπύλη Εκμάθησης



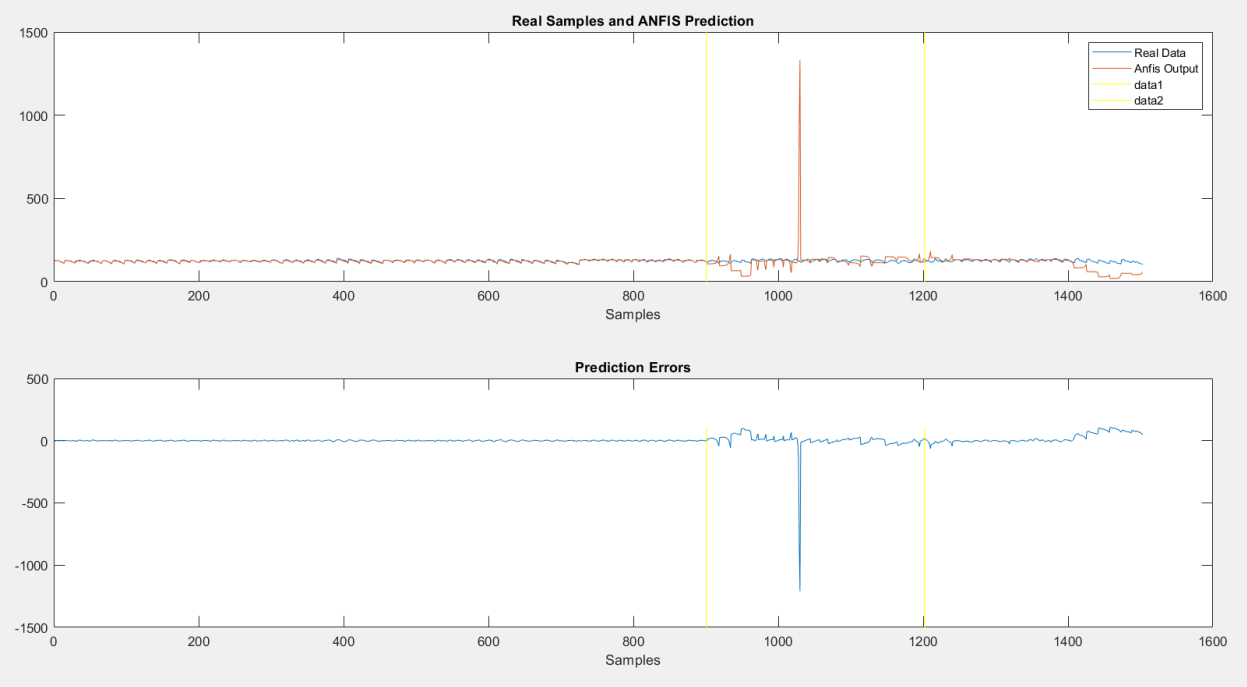

Σχήμα 73: Learning Curve

## RMSE




Σχήμα 74: RMSE

Πραγματικές Τιμές VS Τιμές Πρόβλεψης & Σφάλμα Πρόβλεψης




Σχήμα 75: Real VS Prediction Values & Prediction Error

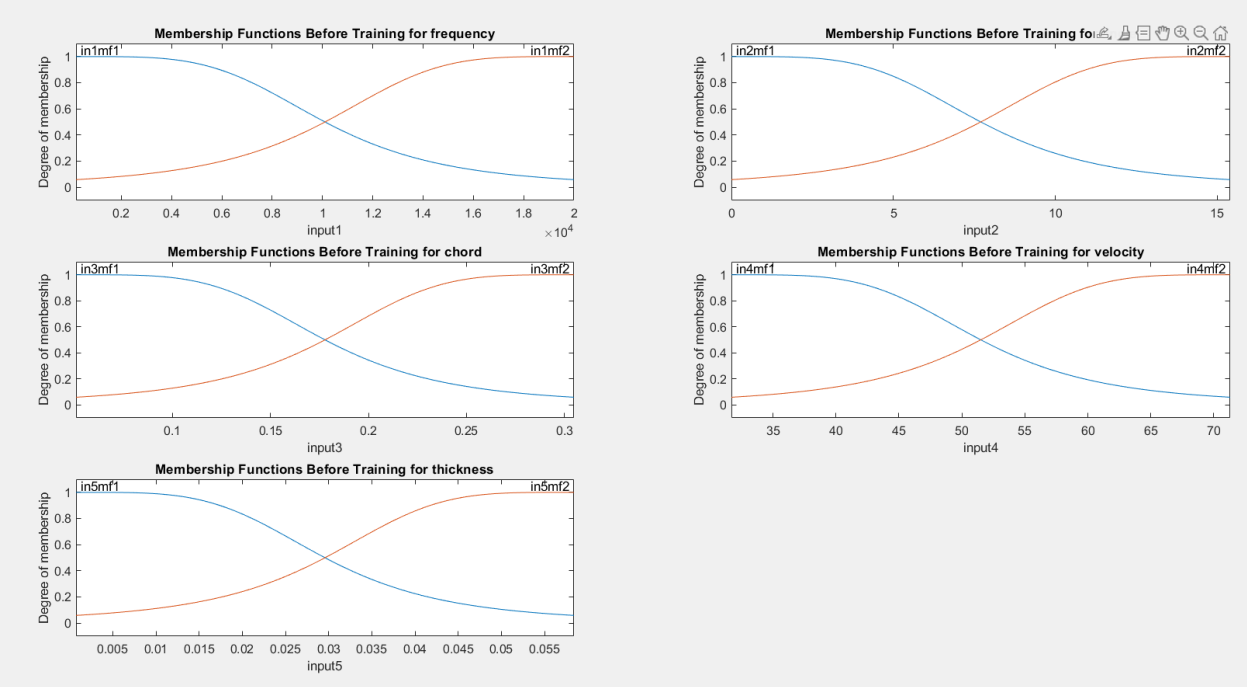
| Error | RMSE | NMSE | NDEI | R2 |
|---|---|---|---|---|
| TSK model 2 | 42.6229 | 32.1637 | 5.6713 | -31.1637 |

### 3.1.3 TSK model 3

Το μοντέλο 3 υλοποιείται με το script “Regression_TSK_model_3.m”.

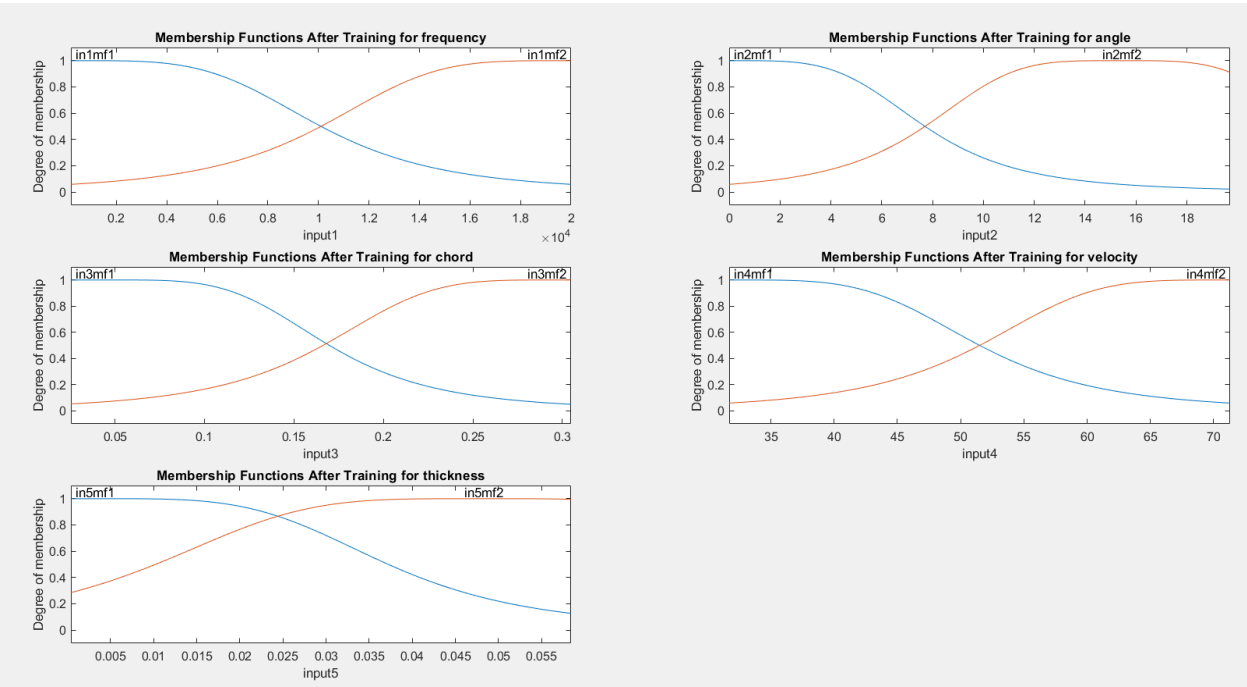
Χρησιμοποιεί μέθοδο ομαδοποίησης grid partition, 2 συναρτήσεις συμμετοχής “gbellmf” και “Polynomial” output (linear).

## Συναρτήσεις Συμμετοχής Πριν την Εκπαίδευση



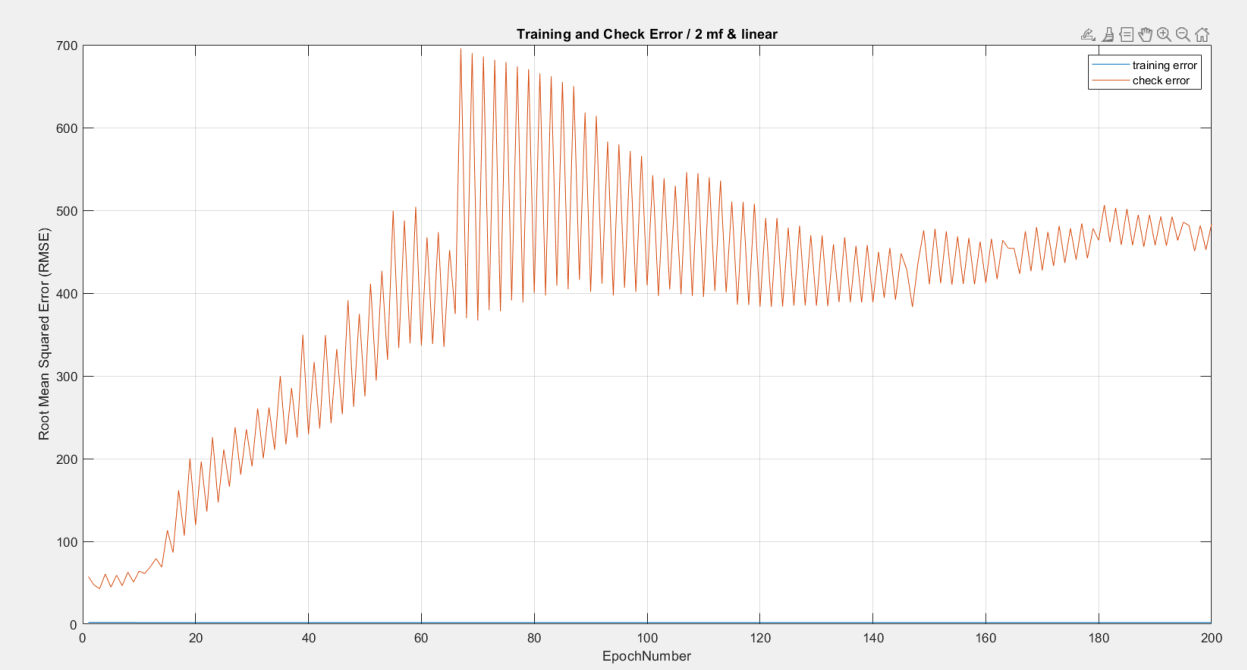

Σχήμα 76: mf Before Training

## Συναρτήσεις Συμμετοχής Μετά την Εκπαίδευση



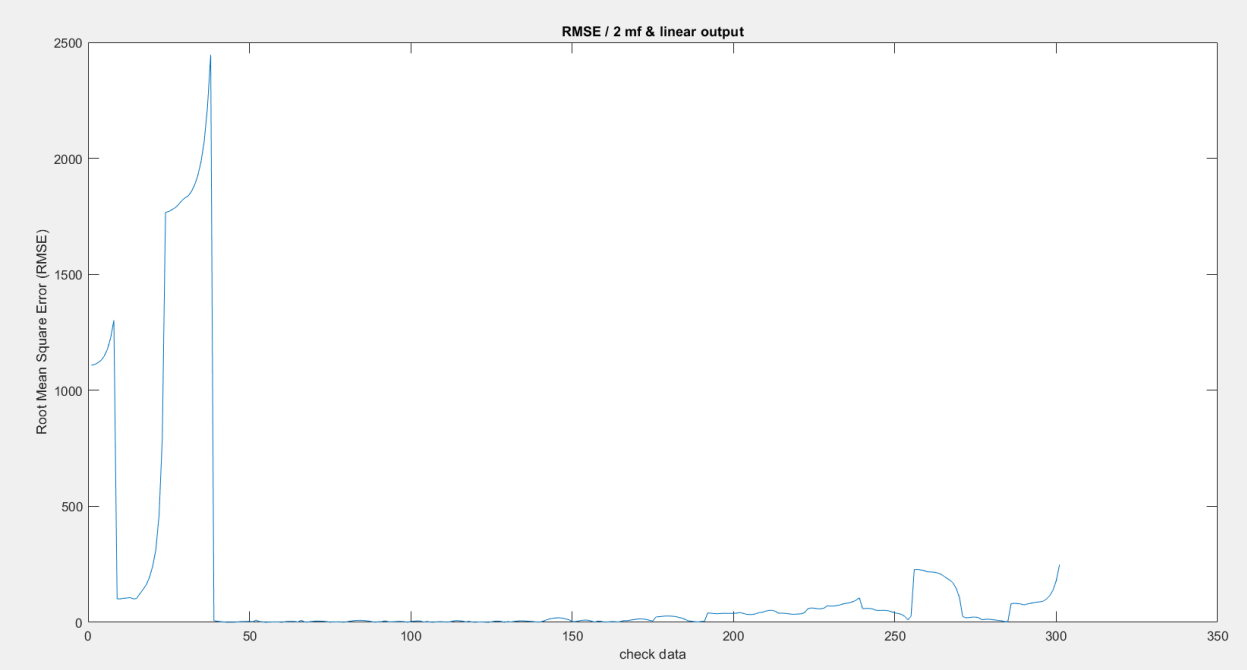

Σχήμα 77: mf After Training

## Καμπύλη Εκμάθησης



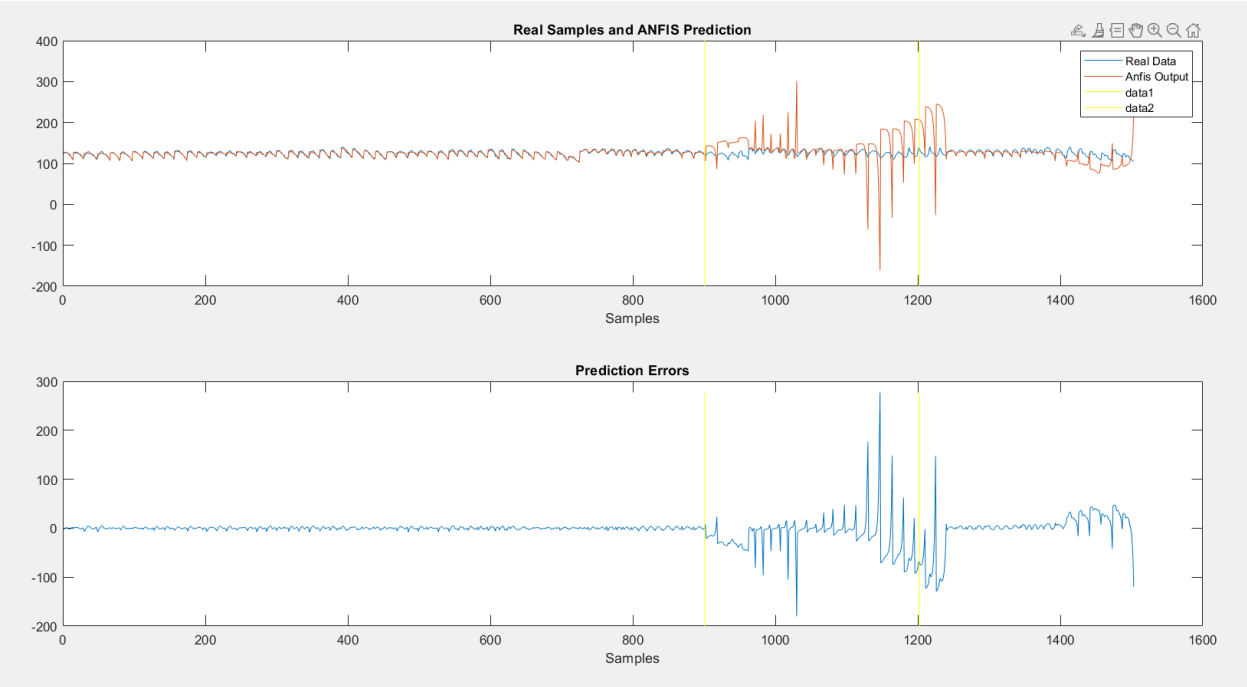

Σχήμα 78: Learning Curve

## RMSE




Σχήμα 79: RMSE

Πραγματικές Τιμές VS Τιμές Πρόβλεψης & Σφάλμα Πρόβλεψης




Σχήμα 80: Real VS Prediction Values & Prediction Error

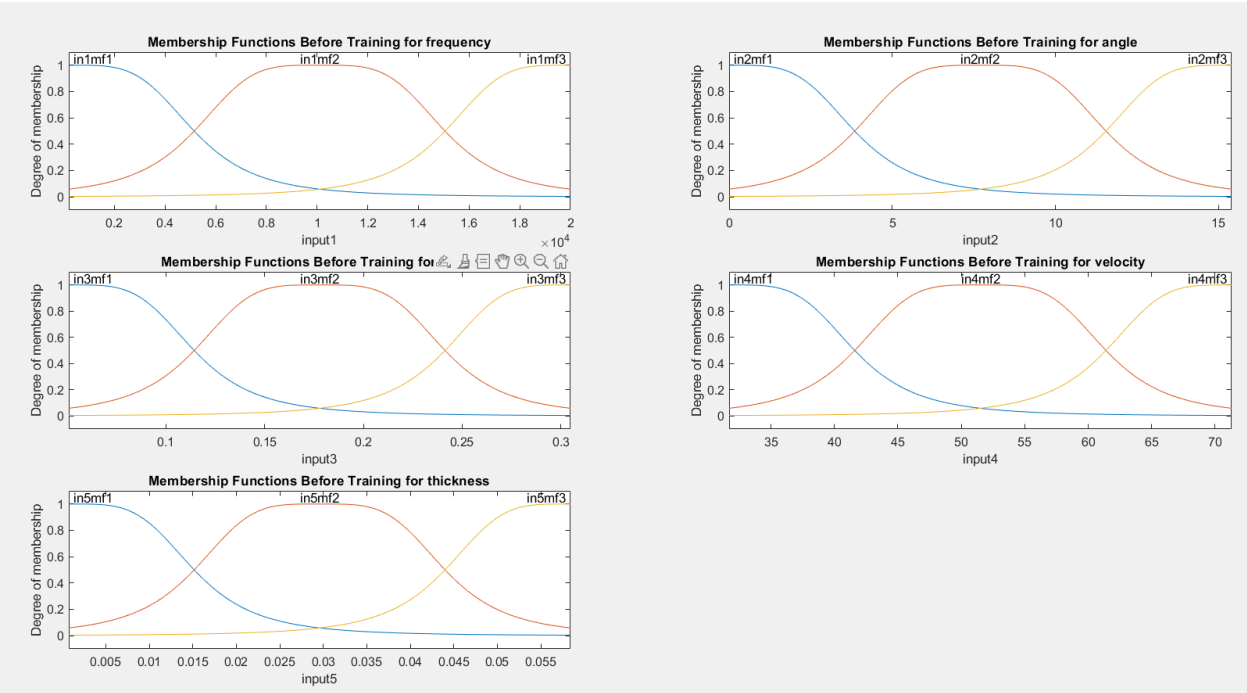
| Error | RMSE | NMSE | NDEI | R2 |
|---|---|---|---|---|
| TSK model 3 | 478.1159 | 4047.1 | 63.6170 | -4046.1 |

### 3.1.4 TSK model 4

Το μοντέλο 4 υλοποιείται με το script “Regression_TSK_model_3.m”.

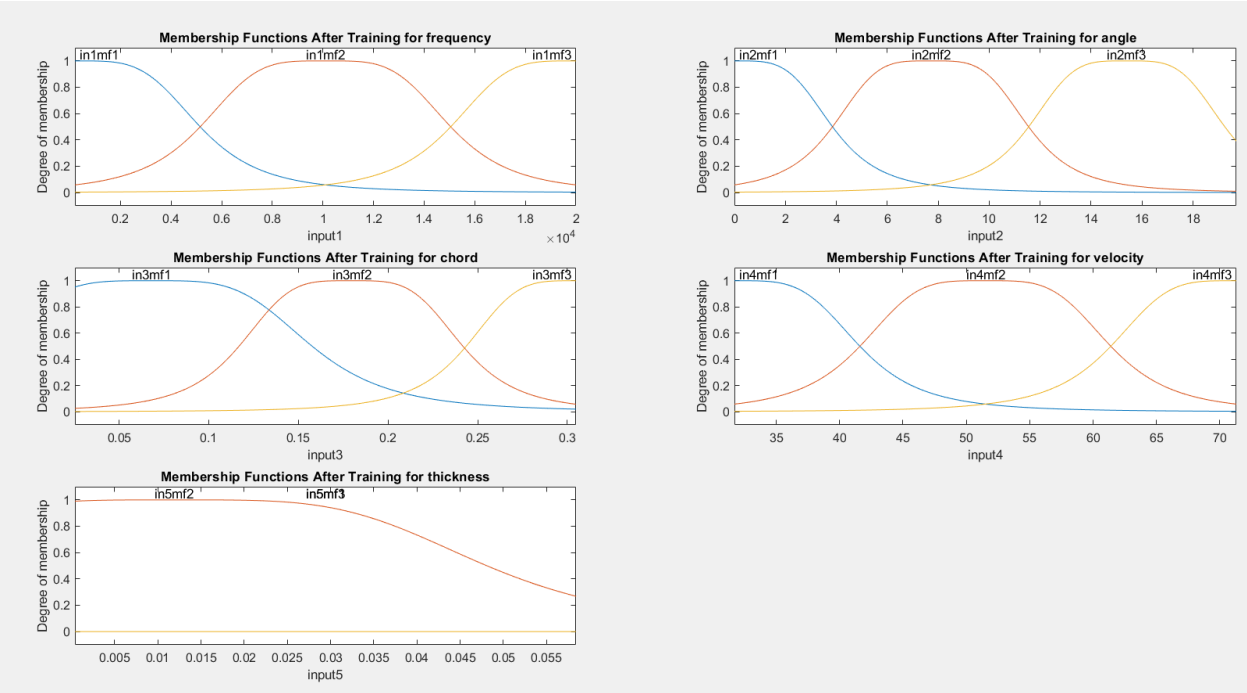
Χρησιμοποιεί μέθοδο ομαδοποίησης grid partition, 3 συναρτήσεις συμμετοχής “gbellmf” και “Polynomial” output (linear).

## Συναρτήσεις Συμμετοχής Πριν την Εκπαίδευση



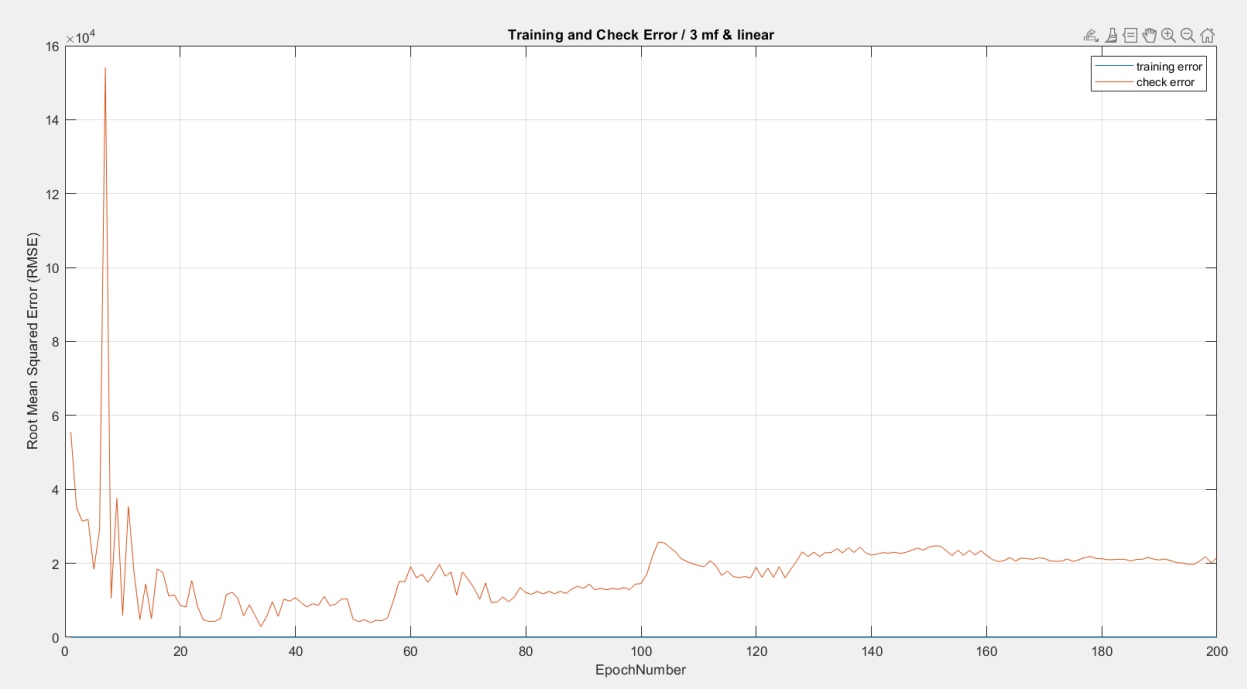

Σχήμα 81: mf Before Training

## Συναρτήσεις Συμμετοχής Μετά την Εκπαίδευση



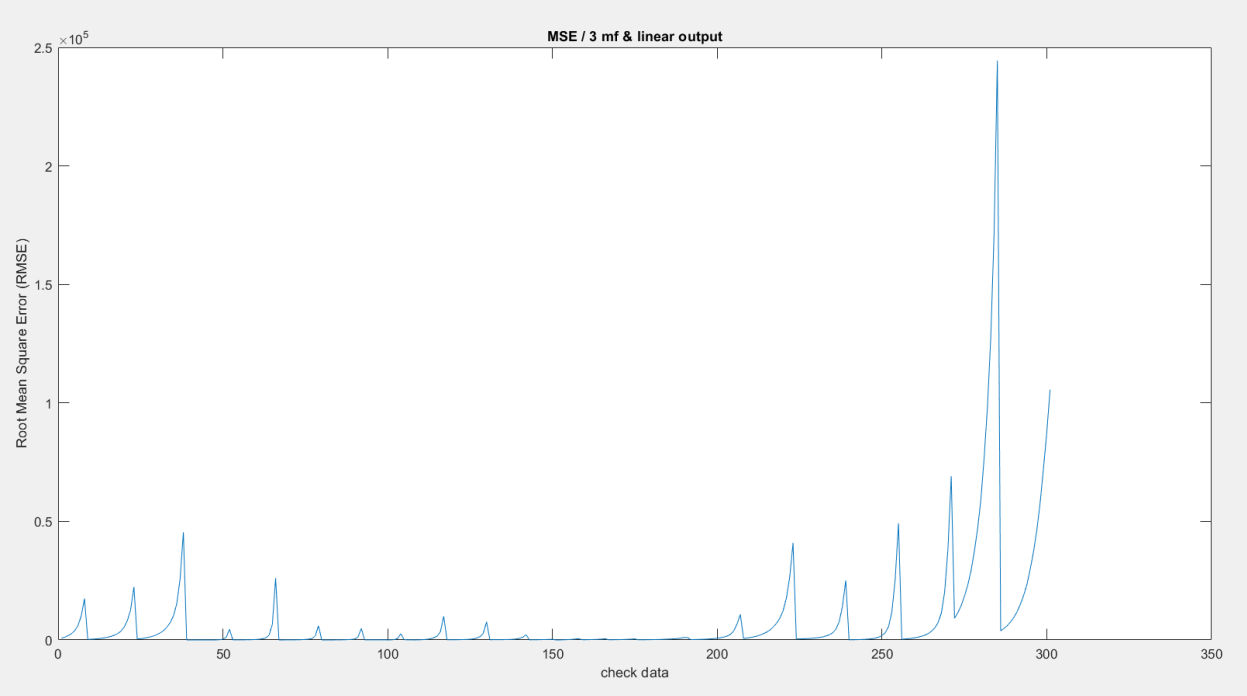

Σχήμα 82: mf After Training

## Καμπύλη Εκμάθησης



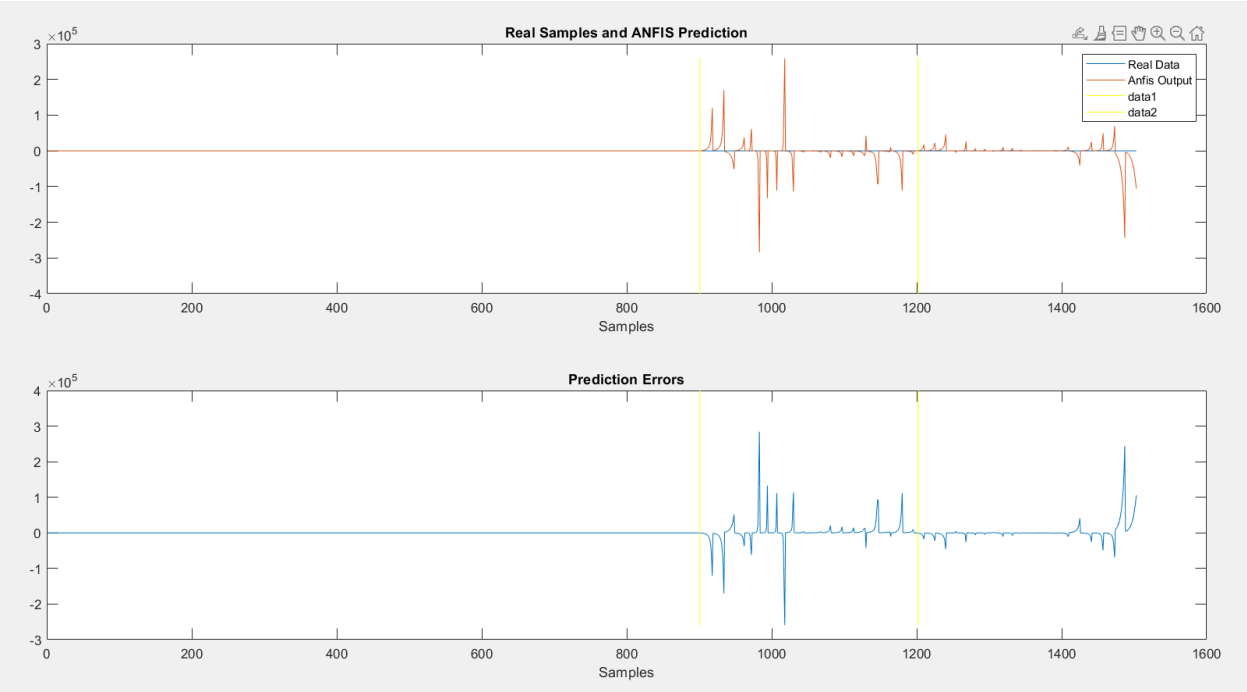

Σχήμα 83: Learning Curve

## RMSE




Σχήμα 84: RMSE

Πραγματικές Τιμές VS Τιμές Πρόβλεψης & Σφάλμα Πρόβλεψης




Σχήμα 85: Real VS Prediction Values & Prediction Error

| Error | RMSE | NMSE | NDEI | R2 |
|---|---|---|---|---|
| TSK model 4 | 24685 | 10787850.9 | 3284.5 | -10787849.9 |

### 3.1.5 Συμπεράσματα και Επιλογή μοντέλου

Σφάλματα και των τεσσάρων μοντέλων.

| Μοντέλο\Σφάλμα | RMSE | NMSE | NDEI | R2 |
|---|---|---|---|---|
| TSK model 1 | 5.9544 | 0.6277 | 0.7923 | 0.3723 |
| TSK model 2 | 42.6229 | 32.1637 | 5.6713 | -31.1637 |
| TSK model 3 | 478.1159 | 4047.1 | 63.6170 | -4046.1 |
| TSK model 4 | 24685 | 10787850.9 | 3284.5 | -10787849.9 |

Το καλύτερο μοντέλο είναι το μοντέλο 1.

Το χειρότερο μοντέλο είναι το μοντέλο 4. Αν και το σφάλμα είναι ακραία μεγάλο, οπότε υπάρχει η πιθανότητα λάθους στον υπολογισμό.

Τα μοντέλα με Polynomial έξοδο παρουσιάζουμε μεμονωμένα μεγάλες τιμές Prediction Error, πράγμα που χαλάει την ακρίβεια του μοντέλου και μεγαλώνει το μέσο όρο του σφάλματος.

Τα μοντέλα φαίνεται από το Learning Curve να κάνουν υπερεκπαίδευση, αλλά στη συνέχεια σταθεροποιείται η καμπύλη. Τα μοντέλα με 3 mf είναι πιο επιρρεπή σε αυτή τη συμπεριφορά.

# 3.2 Part 2 - Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Στο δεύτερο μέρος της εργασίας θα χρησιμοποιήσουμε το dataset “Superconductinity” από το UCI repository, το οποίο είναι dataset με υψηλό βαθμό διαστασιμότητας. Περιέχει 21263 δεδομένα και αποτελείται από 81 features και 1 output.

Είναι απαραίτητο να επιλέξουμε ένα μικρότερο σύνολο features με τα οποία θα εργαστούμε, διότι τα 81 είναι απαγορευτικά μεγάλο νούμερο, λογω της έκρηξης των κανόνων If-Then.

Το μοντέλο υλοποιείται στο script “Regression_part2_v2.m”.

Σύμφωνα με το script δοκιμάζουμε διαφορετικές τιμές για τον αριθμό των σημαντικότερων features που θα κρατήσουμε για να τροφοδοτήσουμε το μοντέλο, καθώς και διάφορες τιμές για την ακτίνα Ra των clusters. Χρησιμοποιούμε την τεχνική Grid Search και διαχωρίζουμε τα δεδομένα με την μέθοδο CV Partition k-fold με k=5.

Η μέθοδος ομαδοποίησης για την δημιουργία των κανόνων είναι η “Subtractive Clustering”, που υλοποιείται με την εντολή genfis2.

Κάθε μοντέλο εκπαιδεύεται με την μέθοδο anfis και χρησιμοποιούμε 100 EpochNumber.

Η επιλογή των σημαντικότερων features έγινε με την εντολή Relieff.

### 3.2.1 Αποτελέσματα Σφαλμάτων του Grid Search

- Πίνακας **RMSE** (Root Mean Square Error)

| **Features\Ra** | **0.2** | **0.4** | **0.6** | **0.8** | **1** |
|---|---|---|---|---|---|
| **5** | 16.4775 | 17.6595 | 18.7122 | 20.3321 | 20.4394 |
| **8** | 16.9144 | 16.8142 | 17.5359 | 20.1406 | 20.2999 |
| **11** | 36.3985 | 15.6254 | 16.6085 | 17.9002 | 18.1107 |
| **14** | 20.4471 | 15.2011 | 15.8688 | 17.2750 | 17.8301 |

- Πίνακας **NMSE** (Normalized Mean Square Error)

| Features\Ra | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
|---|---|---|---|---|---|
| **5** | 0.2314 | 0.2658 | 0.2985 | 0.3523 | 0.3560 |
| **8** | 0.2459 | 0.2411 | 0.2621 | 0.3457 | 0.3512 |
| **11** | 2.6399 | 0.2081 | 0.2351 | 0.2732 | 0.2796 |
| **14** | 0.4280 | 0.1969 | 0.2146 | 0.2544 | 0.2710 |

- Πίνακας **NDEI** (Root of NMSE)

| Features\Ra | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
|---|---|---|---|---|---|
| **5** | 0.4810 | 0.5155 | 0.5462 | 0.5935 | 0.5967 |
| **8** | 0.4938 | 0.4908 | 0.5119 | 0.5879 | 0.5926 |
| **11** | 1.0618 | 0.4561 | 0.4848 | 0.5225 | 0.5287 |
| **14** | 0.5970 | 0.4437 | 0.4632 | 0.5043 | 0.5205 |

- Πίνακας **R2** (Coefficient of Determination)

| Features\Ra | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
|---|---|---|---|---|---|
| **5** | 0.7685 | 0.7341 | 0.7014 | 0.6476 | 0.6439 |
| **8** | 0.7540 | 0.7588 | 0.7378 | 0.6542 | 0.6487 |
| **11** | -1.6399 | 0.7918 | 0.7648 | 0.7267 | 0.7203 |
| **14** | 0.5719 | 0.8030 | 0.7853 | 0.7455 | 0.7289 |

Σύμφωνα με τα παραπάνω αποτελέσματα πετυχαίνουμε το ελάχιστο σφάλμα

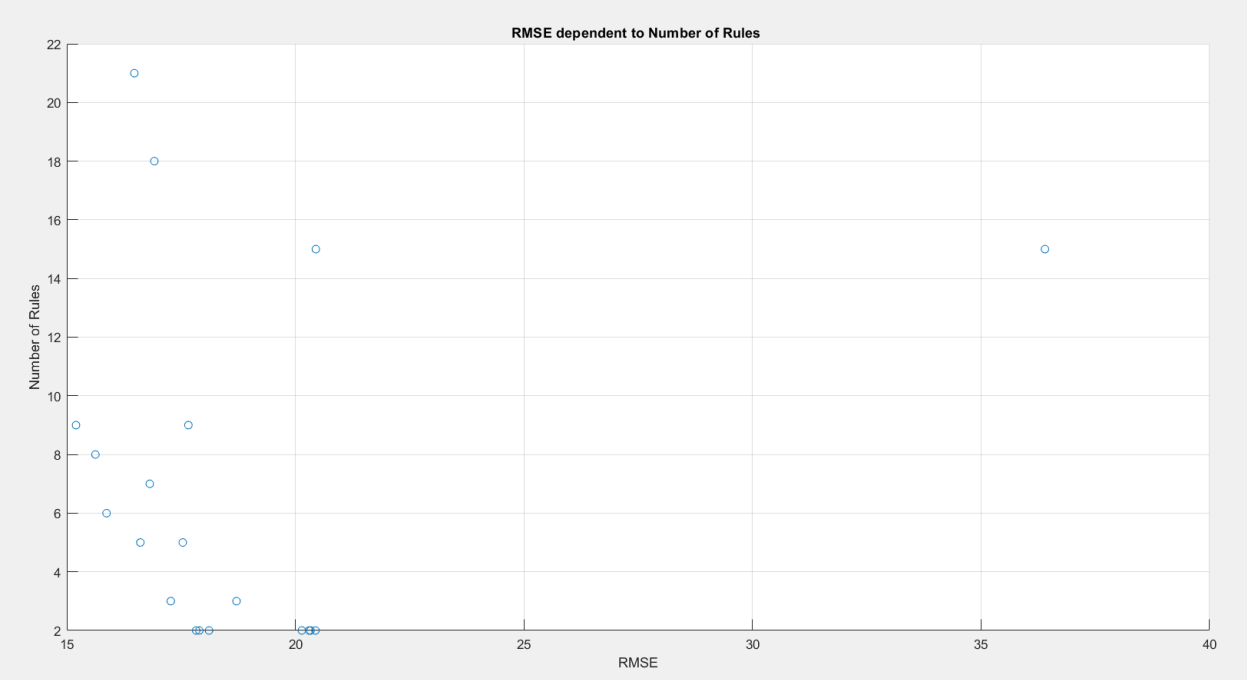
**RMSE = 15.2011** με παραμέτρους:

- features = 14
- rules = 9
- radius = 0.4

### 3.2.2 Διαγράμματα Grid Search

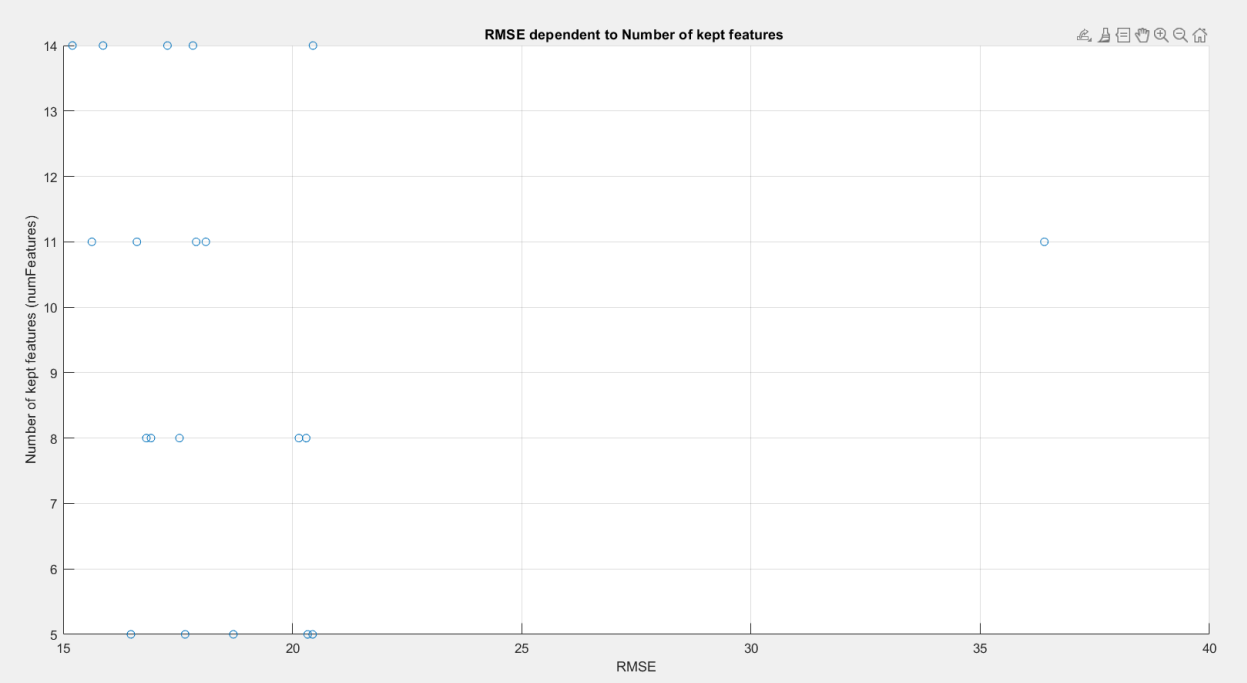
Παρακάτω φαίνονται τα διαγράμματα scatters για το σφάλμα RMSE συναρτήσει του αριθμού των σημαντικότερων features, του αριθμού των κανόνων και της ακτίνας Ra.

RMSE – Number of Rules



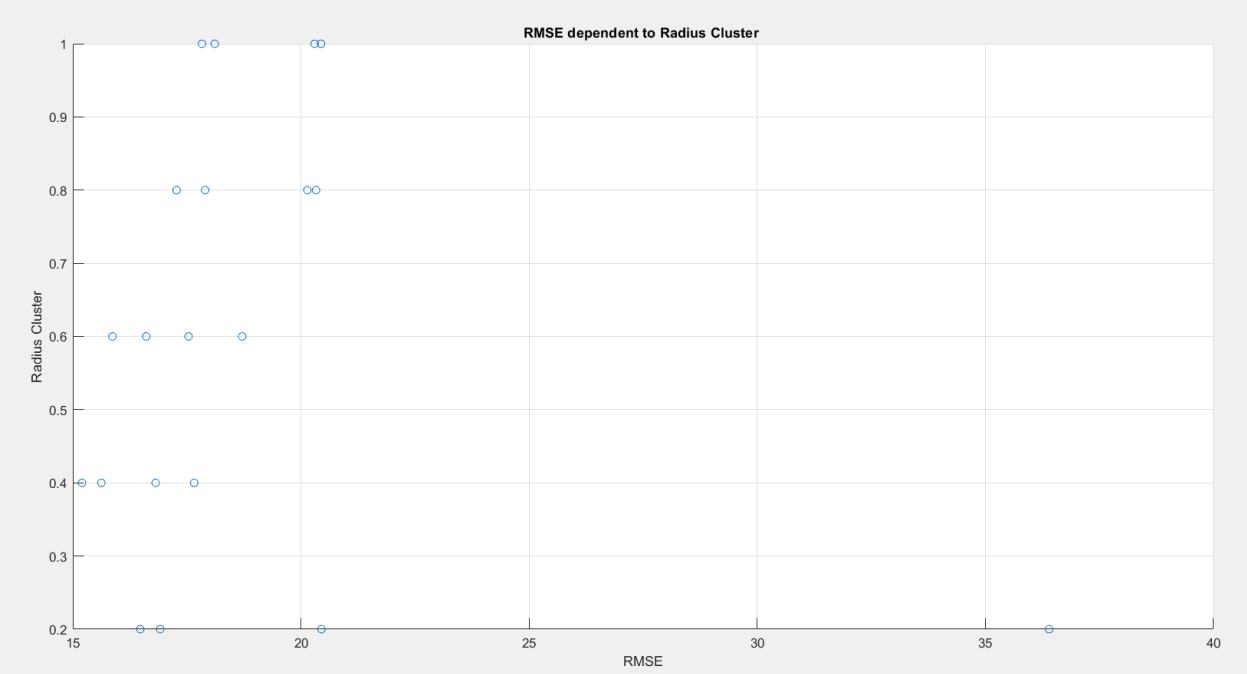

Σχήμα 86: RMSE συναρτήσει του Αριθμού των Κανόνων

RMSE – Number of Features




Σχήμα 87: RMSE συναρτήσει του Αριθμού των Features

RMSE – Radius Ra




Σχήμα 88: RMSE συναρτήσει της Ακτίνας Ra

Συμπεραίνουμε ότι το σφάλμα RMSE μεγαλώνει, όσο αυξάνεται η ακτίνα Ra, θεωρόντας σταθερό τον αριθμό των features. Με μικρή ακτίνα το μοντέλο εκπαιδεύεται με καλύτερη ακρίβεια και κάνει καλύτερες προβλέψεις. Αυτό όμως ισχύει μέχρι ένα σημείο, διότι με πολύ μικρή ακτίνα το μοντέλο κάνει overfitting και δεν μπορεί να εφαρμώσει αυτά που έμαθε και να τα γενικεύσει στο checking dataset. Γι αυτό το λόγο η βέλτιστη ακτίνα είναι η 0.4 και όχι η 0.2, όπως μας επιβεβαιώνει και το script “Regression_part2_v2.m”.

Επίσης, παρατηρούμε ότι το σφάλμα RMSE μειώνεται όσο αυξάνουμε τον αριθμό των features που χρησιμοπούμε. Αυτό είναι λογικό, διότι τροφοδοτούμε το μοντέλο με περισσότερη πληροφορία. Όμως, δεν μπορούμε να το αυξήσουμε παραπάνω, γιατί ο χρόνος εκπαίδευσης αυξάνει εκθετικά και το μοντέλο θα γίνει μη λειτουργικό.

### 3.3.3 Βέλτιστο Μοντέλο

Το βέλτιστο μοντέλο επιλέχθηκε από το script “Regression_part2_v2.m” και υλοποιείται στο script “Optimal_TSK_model.m”.

Αποτελείται από 8 κανόνες, 14 features και 0.4 ακτίνα.

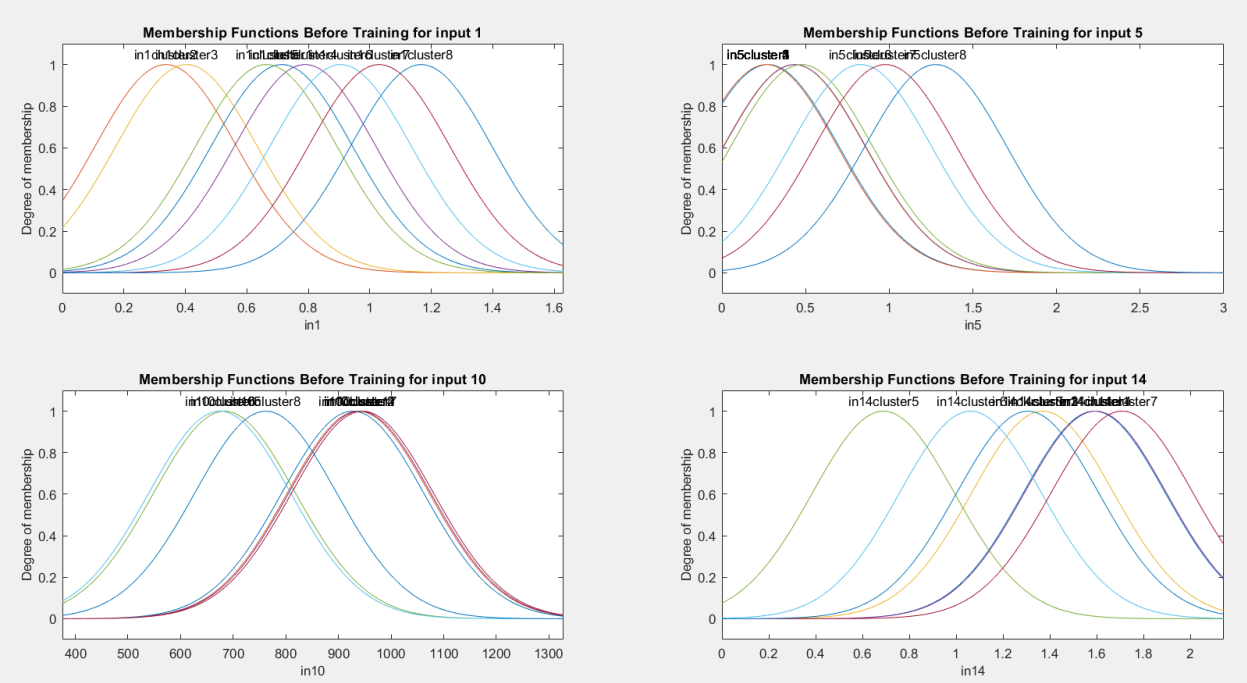
- Πίνακας Σφαλμάτων Βέλτιστου Μοντέλου:

| Error | RMSE | NMSE | NDEI | R2 |
|---|---|---|---|---|
| Optimal Model | 15.1346 | 0.1952 | 0.4418 | 0.8047 |

Παρατηρούμε ότι πετυχαίνουμε πολύ ικανοποιητικά αποτελέσματα.

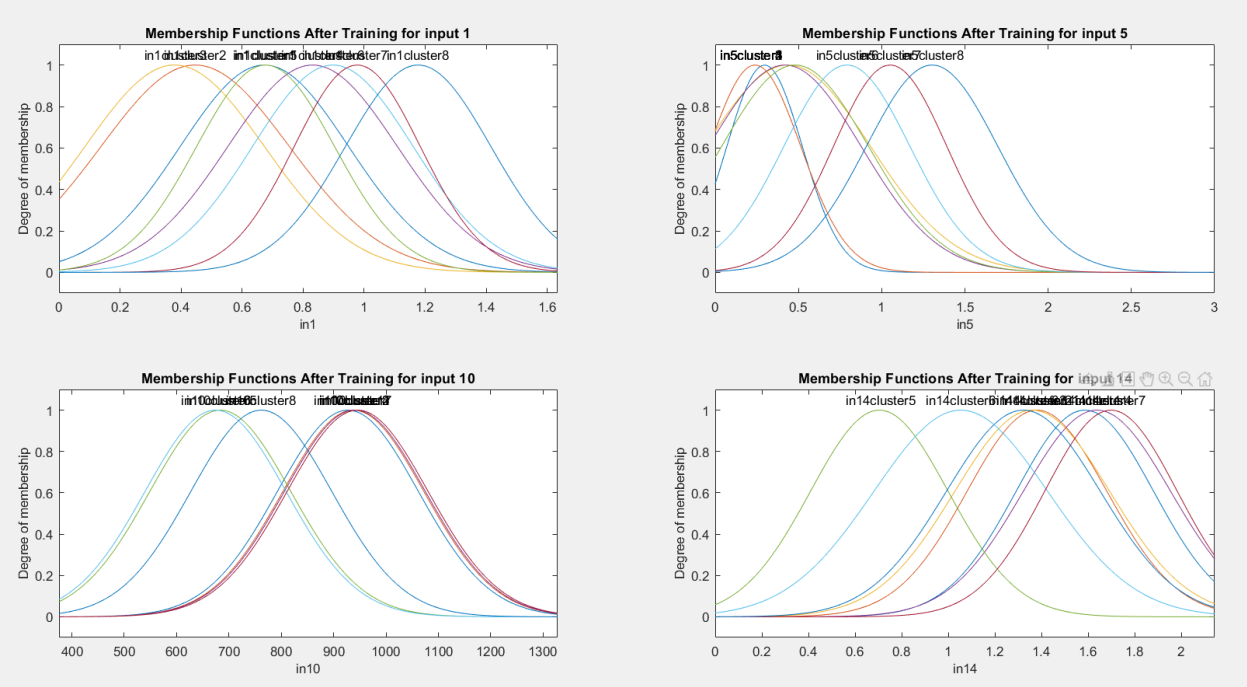
- Διαγράμματα Βέλτιστου Μοντέλου:

Συναρτήσεις Συμμετοχής πριν την εκπαίδευση



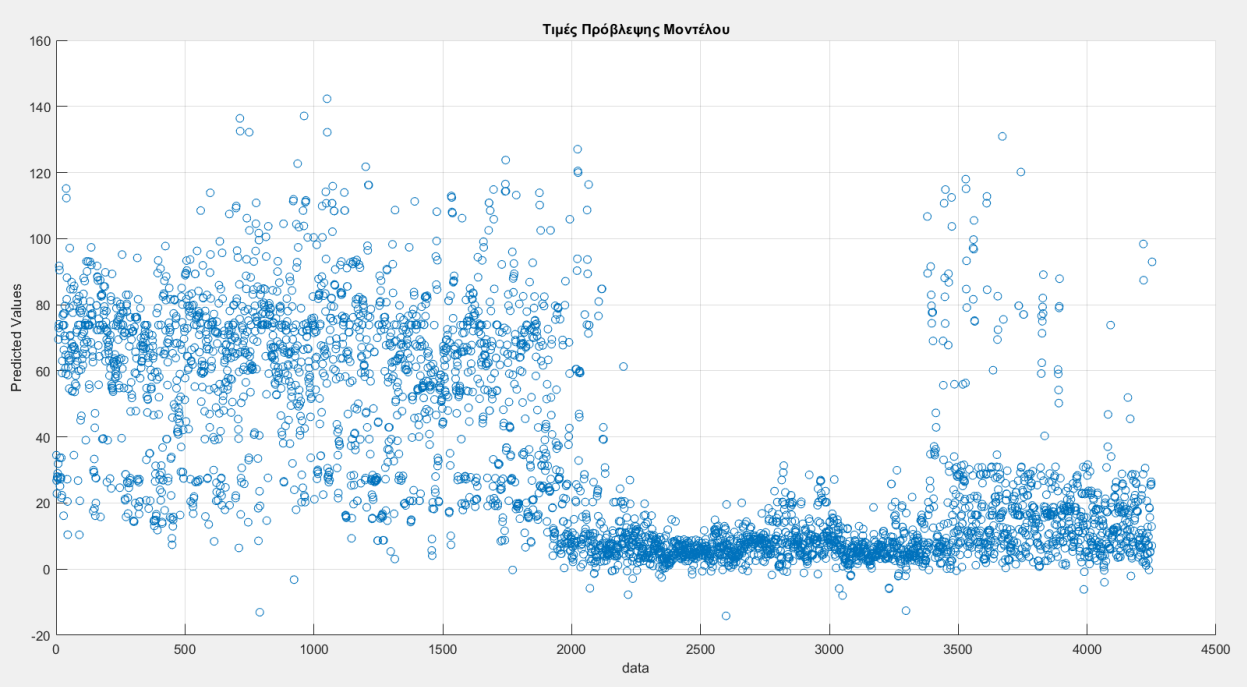

Σχήμα 89: mf Before Training

## Συναρτήσεις Συμμετοχής μετά την εκπαίδευση



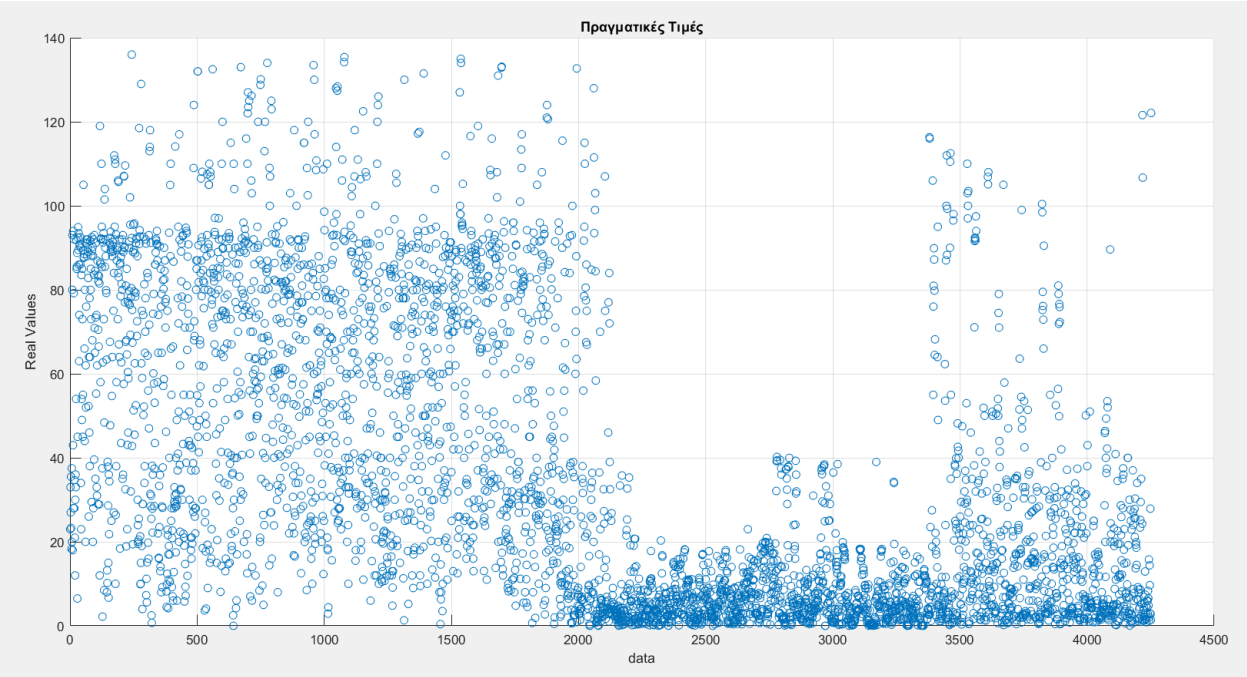

Σχήμα 90: mf After Training

## Τιμές Πρόβλεψης



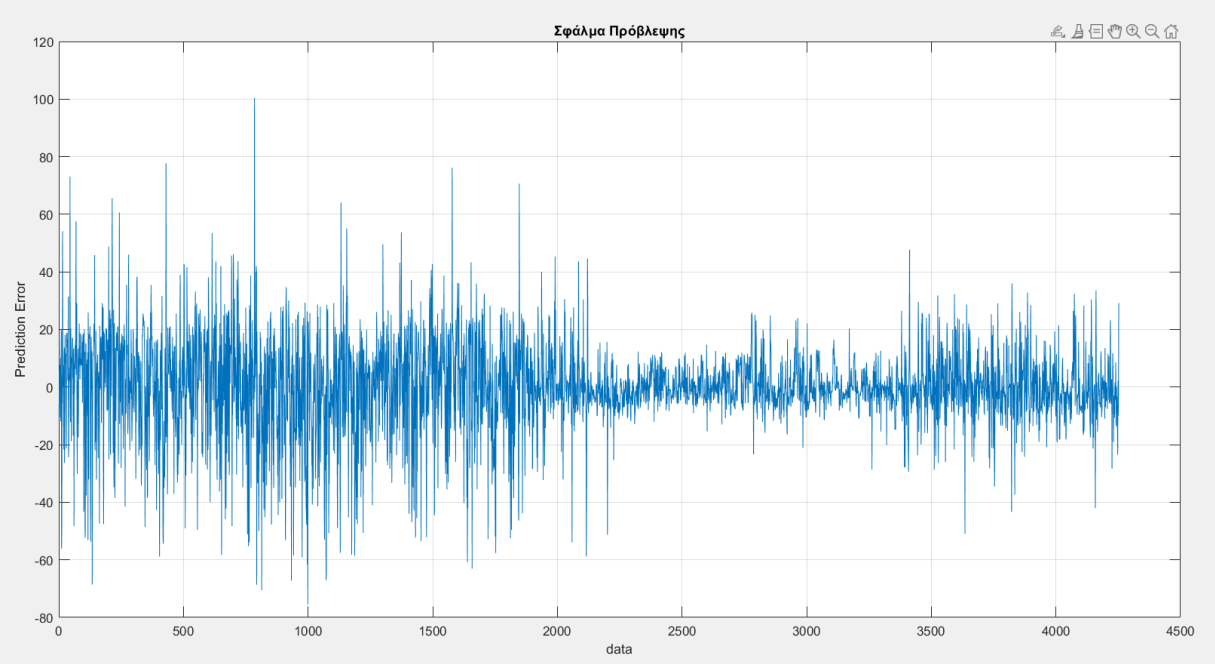

Σχήμα 91: Prediction Values

Πραγματικές Τιμές



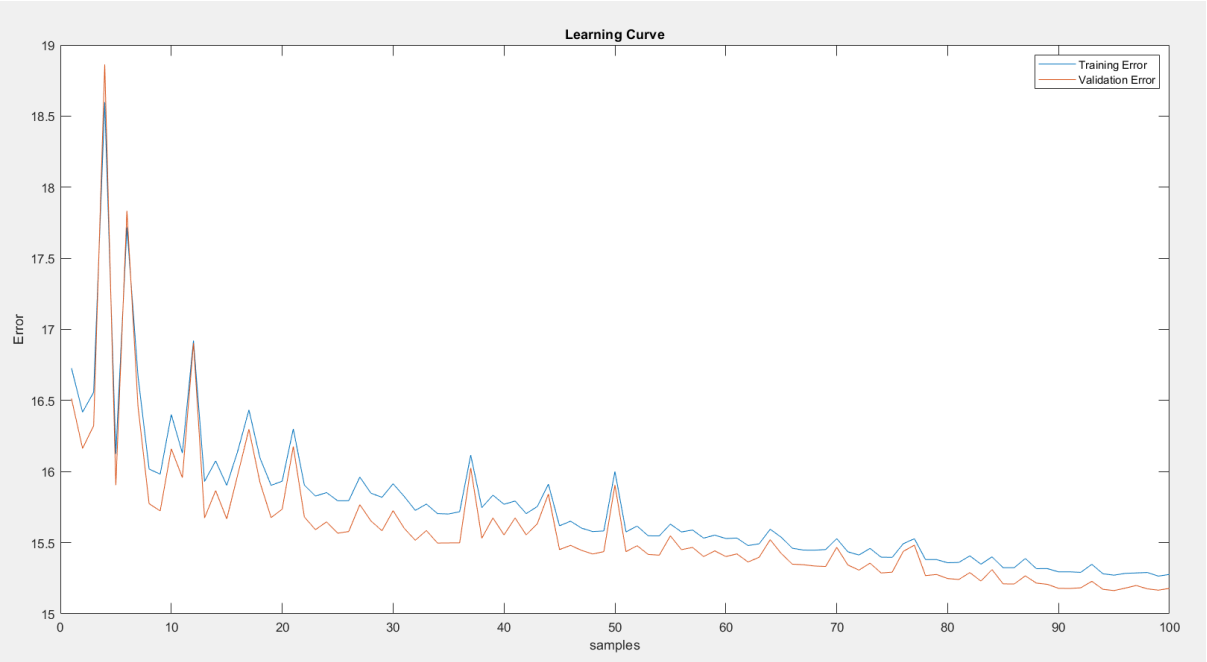

Σχήμα 92: Real Values

## Σφάλμα Πρόβλεψης




Σχήμα 93: Prediction Error

## Καμπύλη Εκμάθησης




Σχήμα 94: Learning Curve

- Οι κανόνες εκπαίδευσης είναι 8 και φαίνονται από την εντολή trainedFis.Rules:

1 "in1==in1cluster1 & in2==in2cluster1 & in3==in3cluster1 & in4==in4cluster1 & in5==in5cluster1 & in6==in6cluster1 & in7==in7cluster1 & in8==in8cluster1 & in9==in9cluster1 & in10==in10cluster1 & in11==in11cluster1 & in12==in12cluster1 & in13==in13cluster1 & in14==in14cluster1 => out1=out1cluster1 (1)"

2 "in1==in1cluster2 & in2==in2cluster2 & in3==in3cluster2 & in4==in4cluster2 & in5==in5cluster2 & in6==in6cluster2 & in7==in7cluster2 & in8==in8cluster2 & in9==in9cluster2 & in10==in10cluster2 & in11==in11cluster2 & in12==in12cluster2 & in13==in13cluster2 & in14==in14cluster2 => out1=out1cluster2 (1)"

3 "in1==in1cluster3 & in2==in2cluster3 & in3==in3cluster3 & in4==in4cluster3 & in5==in5cluster3 & in6==in6cluster3 & in7==in7cluster3 & in8==in8cluster3 & in9==in9cluster3 & in10==in10cluster3 & in11==in11cluster3 & in12==in12cluster3 & in13==in13cluster3 & in14==in14cluster3 => out1=out1cluster3 (1)"

4 "in1==in1cluster4 & in2==in2cluster4 & in3==in3cluster4 & in4==in4cluster4 & in5==in5cluster4 & in6==in6cluster4 & in7==in7cluster4 & in8==in8cluster4 & in9==in9cluster4 & in10==in10cluster4 & in11==in11cluster4 & in12==in12cluster4 & in13==in13cluster4 & in14==in14cluster4 => out1=out1cluster4 (1)"

5 "in1==in1cluster5 & in2==in2cluster5 & in3==in3cluster5 & in4==in4cluster5 & in5==in5cluster5 & in6==in6cluster5 & in7==in7cluster5 & in8==in8cluster5 & in9==in9cluster5 & in10==in10cluster5 & in11==in11cluster5 & in12==in12cluster5 & in13==in13cluster5 & in14==in14cluster5 => out1=out1cluster5 (1)"

6 "in1==in1cluster6 & in2==in2cluster6 & in3==in3cluster6 & in4==in4cluster6 & in5==in5cluster6 & in6==in6cluster6 & in7==in7cluster6 & in8==in8cluster6 & in9==in9cluster6 & in10==in10cluster6 & in11==in11cluster6 & in12==in12cluster6 & in13==in13cluster6 & in14==in14cluster6 => out1=out1cluster6 (1)"

7 "in1==in1cluster7 & in2==in2cluster7 & in3==in3cluster7 & in4==in4cluster7 & in5==in5cluster7 & in6==in6cluster7 & in7==in7cluster7 & in8==in8cluster7 & in9==in9cluster7 & in10==in10cluster7 & in11==in11cluster7 & in12==in12cluster7 & in13==in13cluster7 & in14==in14cluster7 => out1=out1cluster7 (1)"

8 "in1==in1cluster8 & in2==in2cluster8 & in3==in3cluster8 & in4==in4cluster8 & in5==in5cluster8 & in6==in6cluster8 & in7==in7cluster8 & in8==in8cluster8 & in9==in9cluster8 & in10==in10cluster8 & in11==in11cluster8 & in12==in12cluster8 & in13==in13cluster8 & in14==in14cluster8 => out1=out1cluster8 (1)"

- Συμπεράσματα

Από τα διαγράμματα παρατηρούμε ότι το μοντέλο προβλέπει σε ικανοποιητικό βαθμό τις πραγματικές τιμές.

Αν χρησιμοποιούσαμε Grid Partition αντί για Subtractive Clustering και υποθέσουμε ότι για κάθε είσοδο είχαμε 3 Ασαφή Σύνολα, αυτό σημαίνει ότι το μοντέλο θα χρειαζόταν $3^{14}$ κανόνες, έναντι των 8 που χρησιμοποιούμε σε αυτή την εργασία.