Γιώργος Παναγιωτάτος

SVRegression

Datasets: Χρησιμοποιήθηκαν 3 διαφορετικά Datasets τα οποία προήλθαν από το kaggle. Η ιδέα ήταν να δούμε πως συμπεριφέρεται το SVM σε 3 διαφορετικές περιπτώσεις οι οποίες διαφέρουν σε δυσκολία. Ένα εύκολο Dataset για πρόβλεψη του Rating σε εφαρμογές Android. Ένα μέτριας δυσκολίας Dataset το οποίο είναι η πρόβλεψη αποτελέσματος σε εξέταση μαθηματικών. Και ένα δύσκολο Dataset το οποίο είναι η πρόβλεψη τιμής που έχουν σπίτια στην Νέα Υόρκη στην δημοφιλή πλατφόρμα Airbnb.

Preprocessing: Στο preprocessing η λογική που ακολουθήθηκε είναι η εξής: όσα δεδομένα είναι αριθμοί κανονικοποίησε με MinMaxScaler και όσα είναι κατηγορικά εφάρμοσε OneHotEncoding. Εδώ να πούμε πως δεδομένα όπως ονόματα ανθρώπων δεν χρησιμοποιήθηκαν για την δημιουργία του input καθώς και δεν προσφέρουν κάτι στο μοντέλο μας(λογικά) αλλά και για να μην γίνει πολύ μεγάλο το input.

Για το Rating σε εφαρμογές Android χρησιμοποιήθηκαν τα εξής: (διακριτά δεδομένα) "Reviews", "Price" , "Category","Installs", "Size", "Genres".

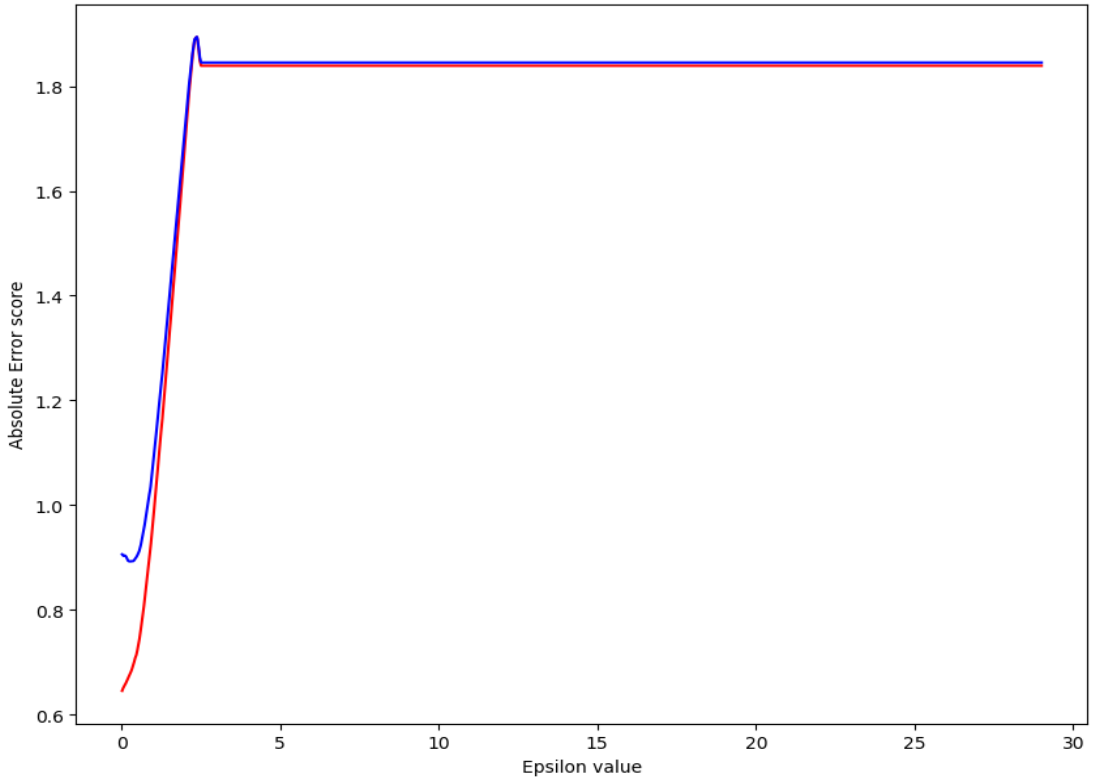
Για την πρόβλεψη του αποτελέσματος σε εξέταση μαθηματικών χρησιμοποιήθηκαν τα εξής: (συνεχή δεδομένα) "reading score","writing score" και (διακριτά δεδομένα) "gender", "race/ethnicity" , "parental level of education","lunch", "test preparation course".

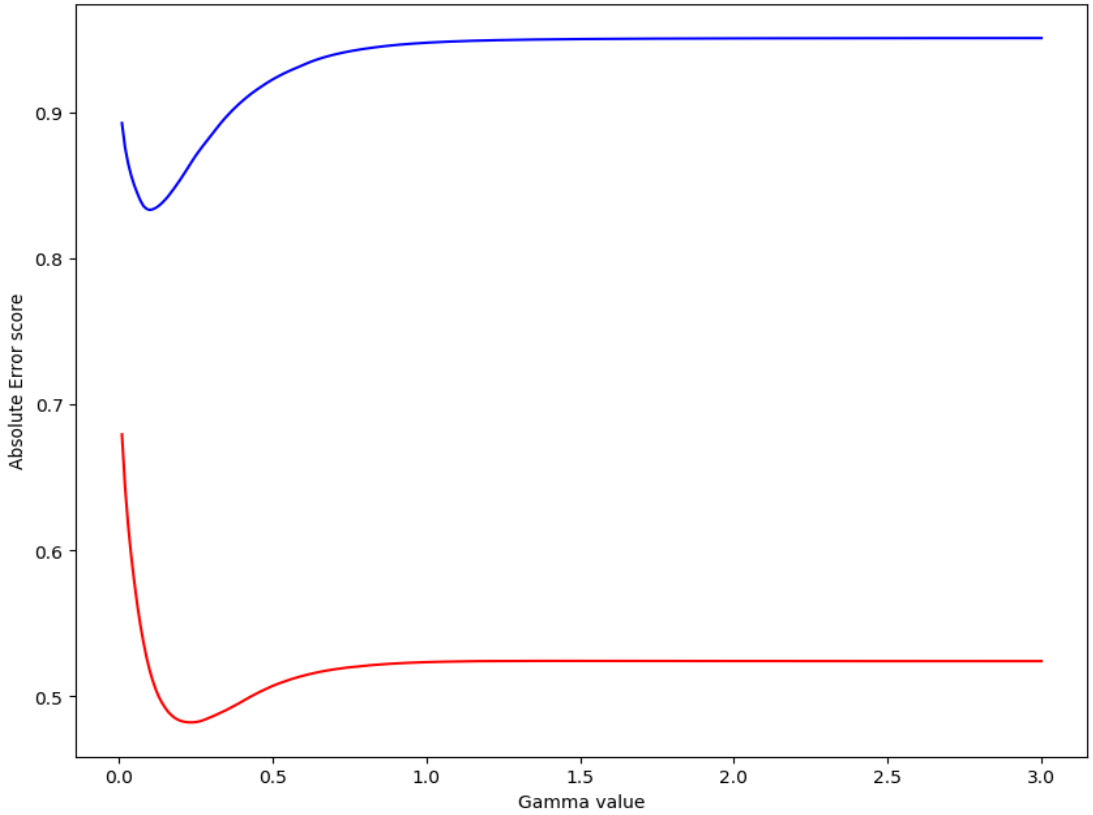
Για την πρόβλεψη τιμής σε σπίτι στην Νέα Υόρκη στο Airbnb χρησιμοποιήθηκαν τα εξής: (συνεχή δεδομένα) "latitude","longitude" , "minimum_nights", "number_of_reviews" , "reviews_per_month", "calculated_host_listings_count", "availability_365" και (διακριτά δεδομένα) "neighbourhood_group","neighbourhood", "room_type".

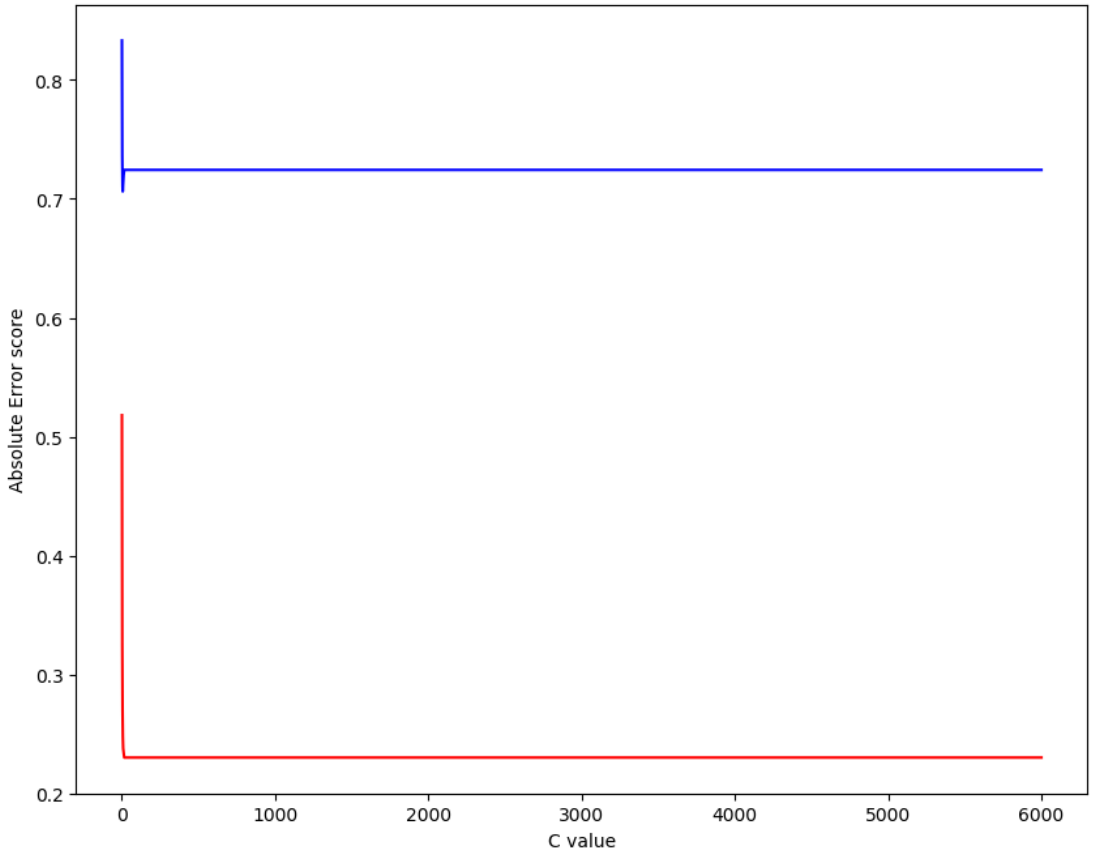
Πειράματα: Για τα πειράματα έφτιαξα μία συνάρτηση η οποία κάνει greed search χρησιμοποιώντας τις παραμέτρους. Το range των παραμέτρων ήταν τέτοιο ώστε να φανεί η επίδραση της κάθε παραμέτρου στο Dataset (τα αρνητικά και τα θετικά της). Μέσα από τα διαγράμματα θα καταλάβουμε τι συμβαίνει στα μοντέλα μας.

Αποτελέσματα Rbf kernel: Με κόκκινο στα διαγράμματα βλέπουμε το Training set και με μπλε το Test set

Πρώτο Dataset:

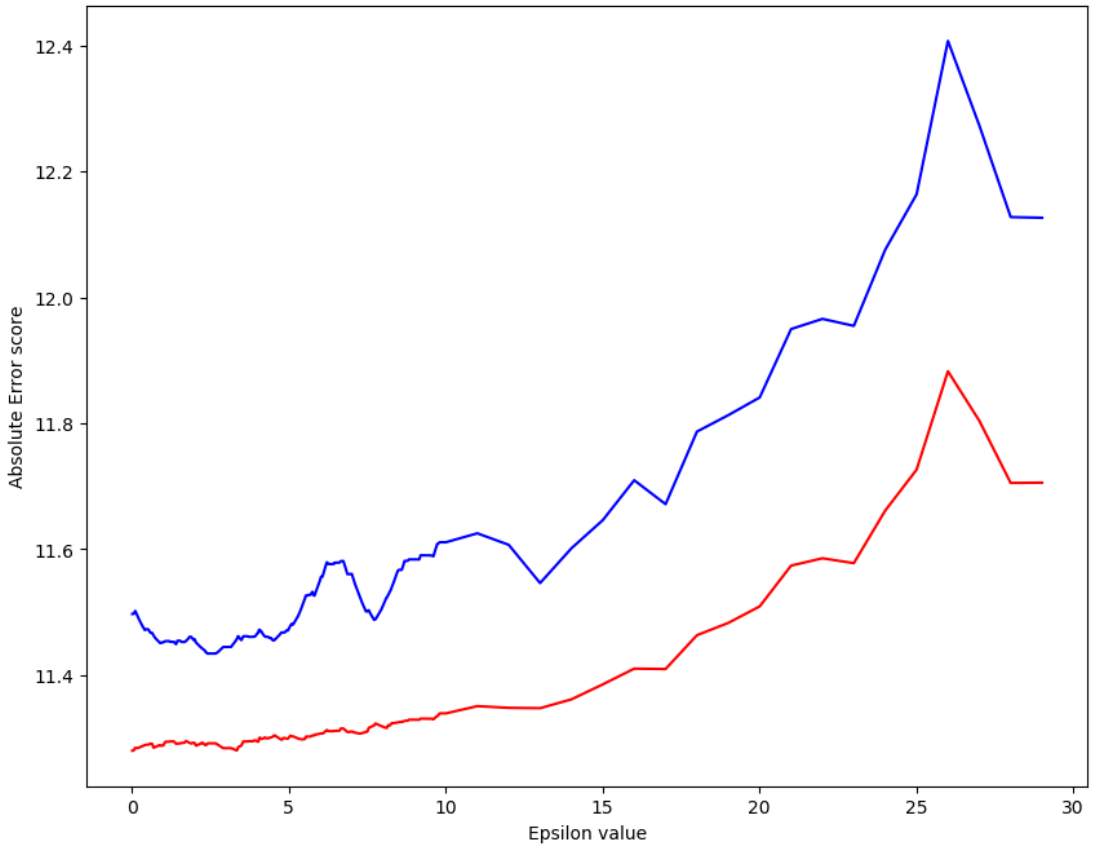


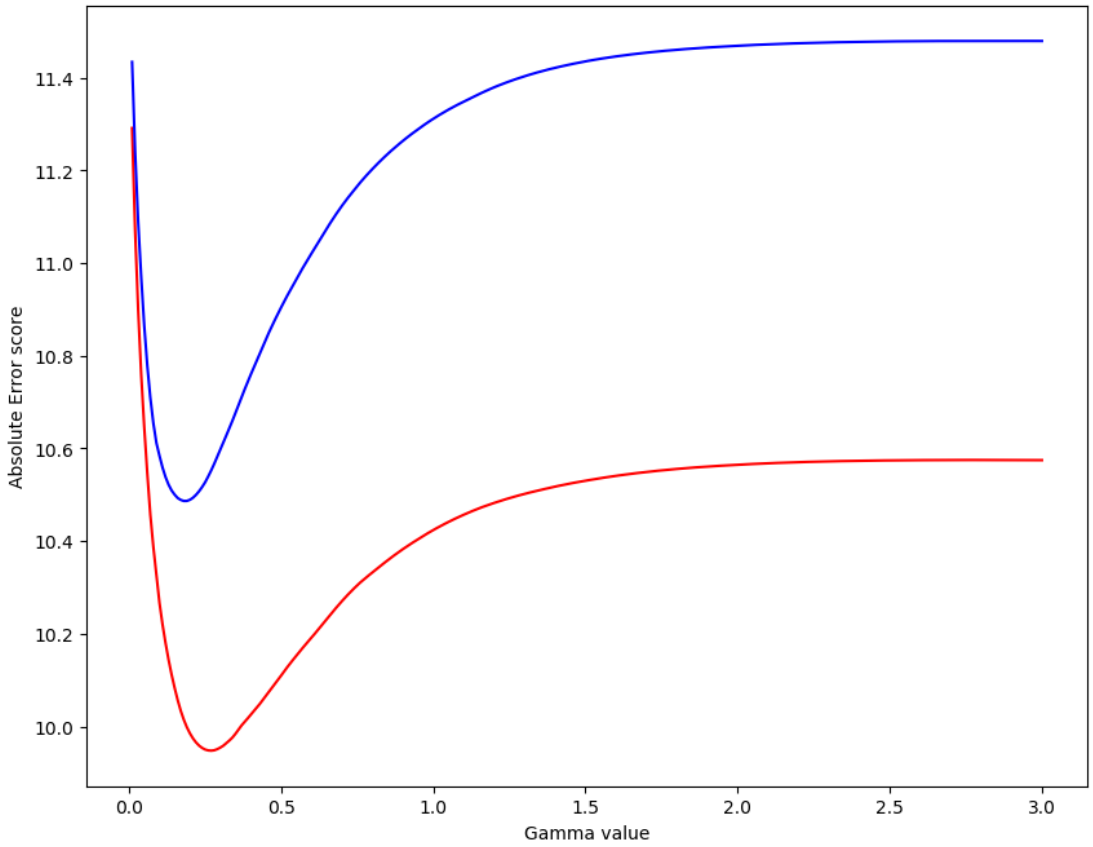


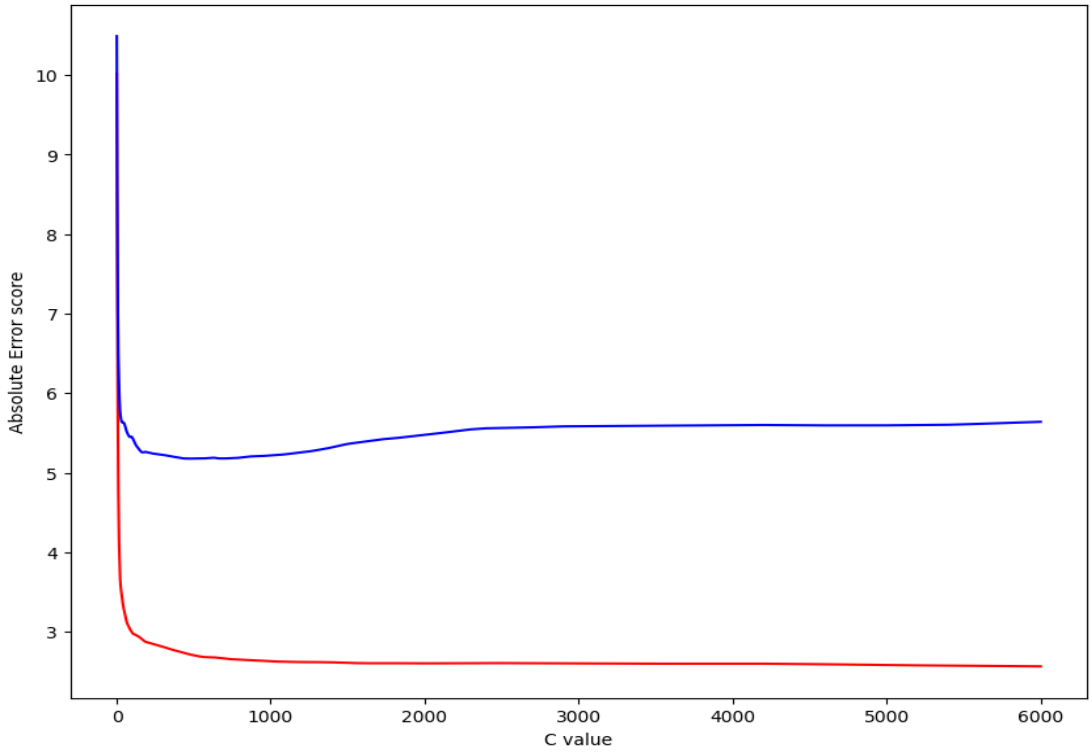


Παρατηρήσεις: Σε όλα τα διαγράμματα παρατηρούμε πως από ένα σημείο και μετά το σφάλμα παραμένει σταθερό τόσο Train_set αλλά και στο Test_set. Αυτό μας οδηγεί στο συμπέρασμα πως μετά από εκείνο το σημείο όσο και αν αλλάξω τις παραμέτρους δεν θα αλλάξει κάτι για το μοντέλο μου. Στην epsilon παράμετρο αυτό που γίνεται μετά από το σημείο που συζητάμε είναι να μην χρησιμοποιεί πλέον κανένα σημείο ως SV και χρησιμοποιεί μια μέση τιμή του train για να προβλέψει την έξοδο. Αυτό έχει ως αποτέλεσμα σταθερό σφάλμα και στο train και στο test. Για την gamma παράμετρο τώρα, από μία τιμή και μετά χρησιμοποιεί σχεδόν όλο το Dataset για τον υπολογισμό του υπερεπιπέδου που περικλείει τα δεδομένα μας. Κατά αυτόν τον τρόπο, μετά όσο και να το αυξήσεις απλά δεν έχει άλλο σημείο να λάβει υπόψιν. Στην C παράμετρο αυτό που συμβαίνει είναι το εξής: από μία τιμή και μετά σε όλα τα σημεία που είναι SV τοποθετεί μια γκαουσιανή και προβλέπει με 0 σφάλμα αυτά και στα υπόλοιπα το σφάλμα είναι σταθερό (είτε ανήκουν στο Train είτε ανήκουν στο Test).

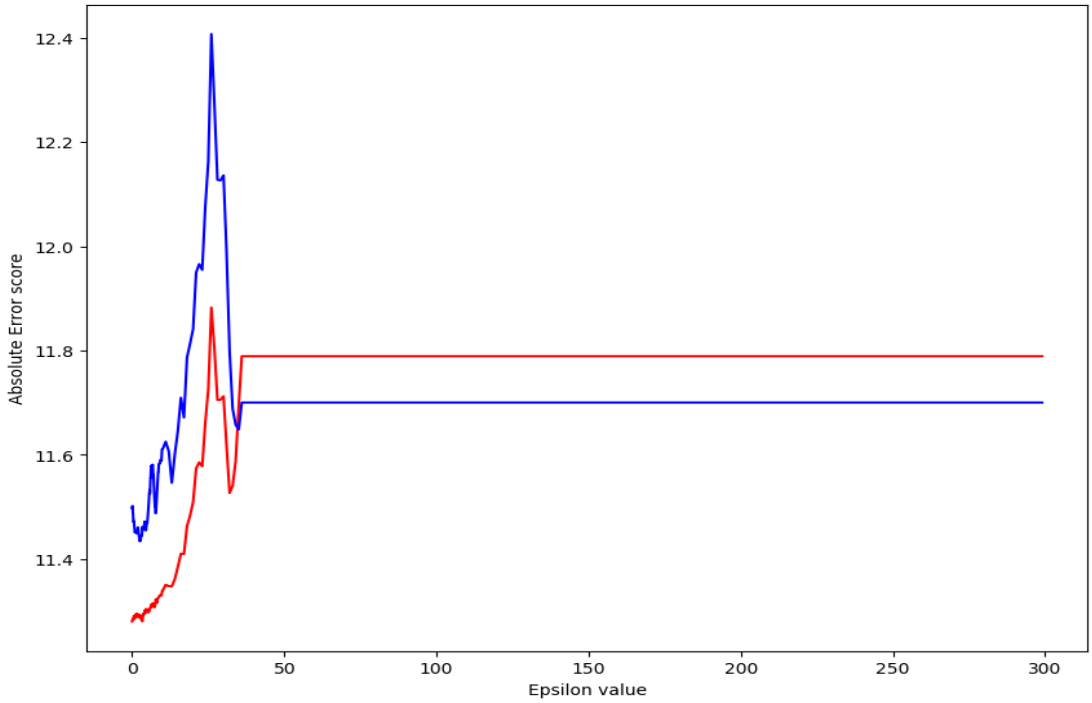
Δεύτερο Dataset:



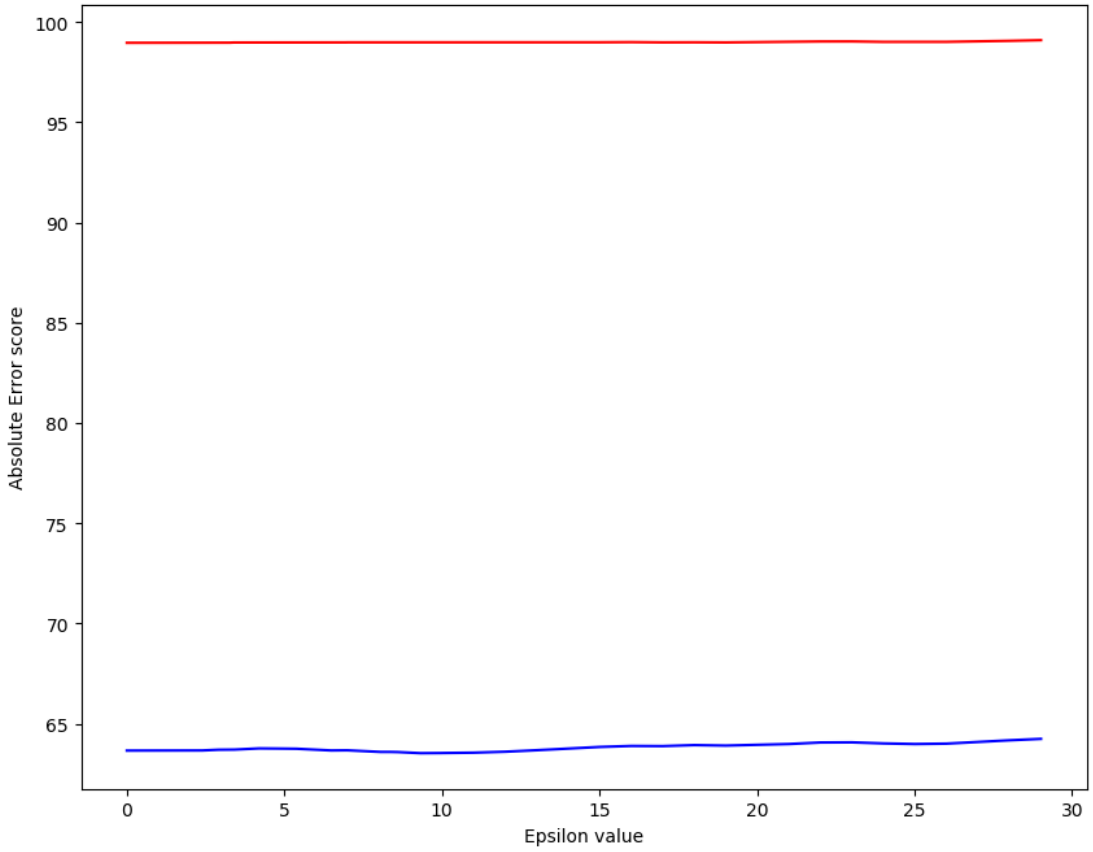


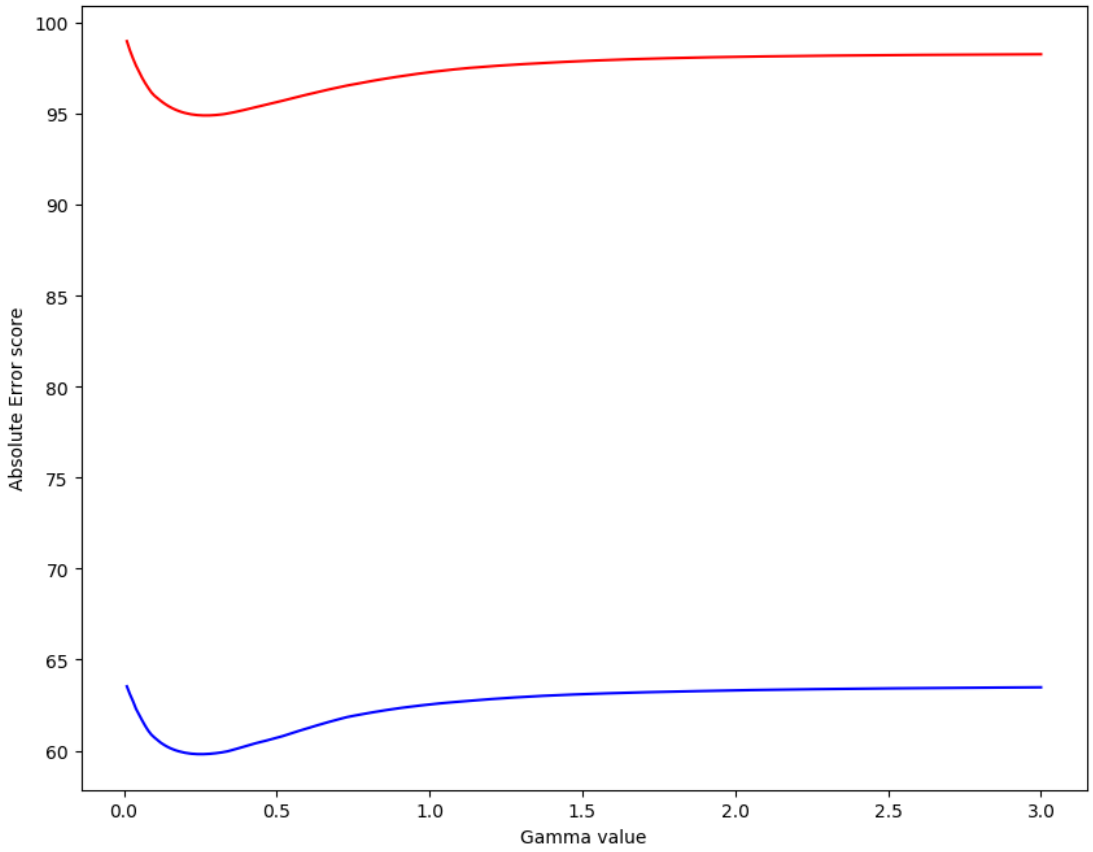


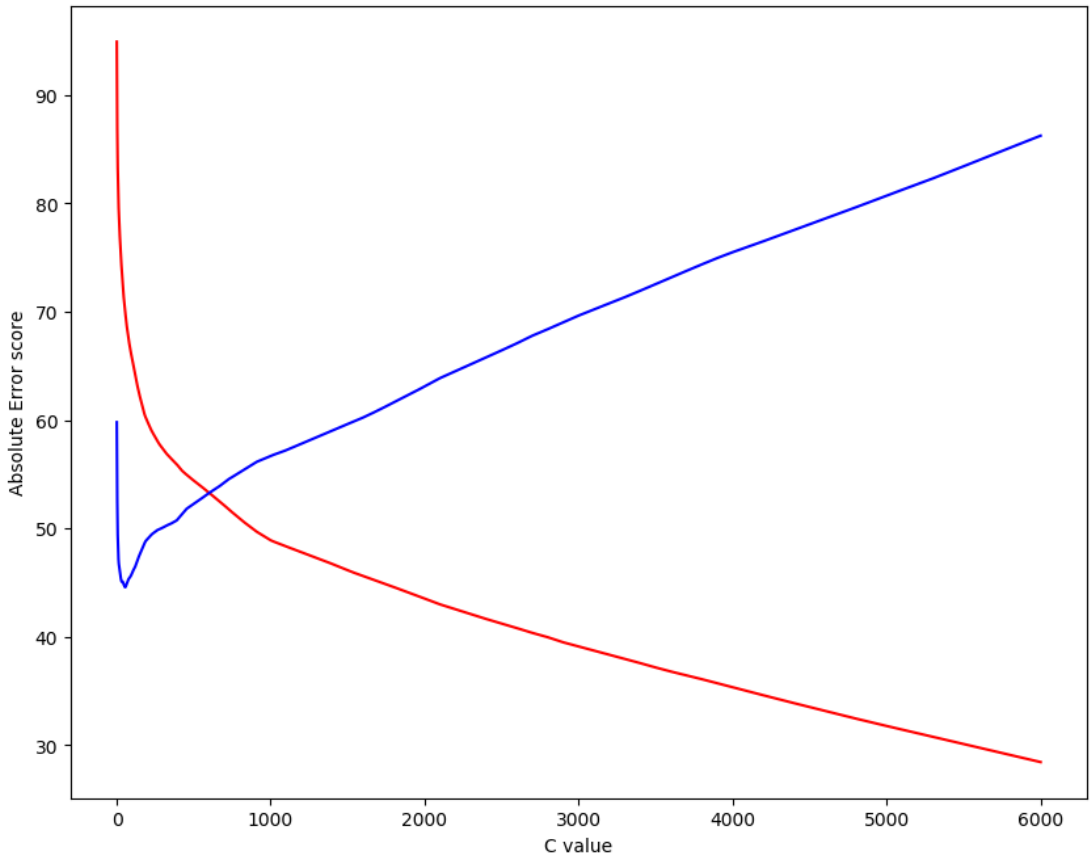
Παρατηρήσεις: Στο δεύτερο Dataset για την παράμετρο gamma και C συμβαίνει ακριβώς το ίδιο με το πρώτο Dataset με την μόνη ίσως διαφορά ότι συγκλίνει λίγο πιο μετά από το πρώτο (λογικό καθώς είναι ελαφρώς πιο δύσκολο). Για την τιμή του epsilon τώρα αυτό που συμβαίνει φαίνεται να είναι διαφορετικό από το πρώτο Dataset. Όμως είναι ακριβώς το ίδιο απλά βρίσκεται σε διαφορετική φάση το ένα από το άλλο. Αν μεγαλώσουμε το epsilon θα παρατηρήσουμε το ίδιο ακριβώς μοτίβο.



Τρίτο Dataset:

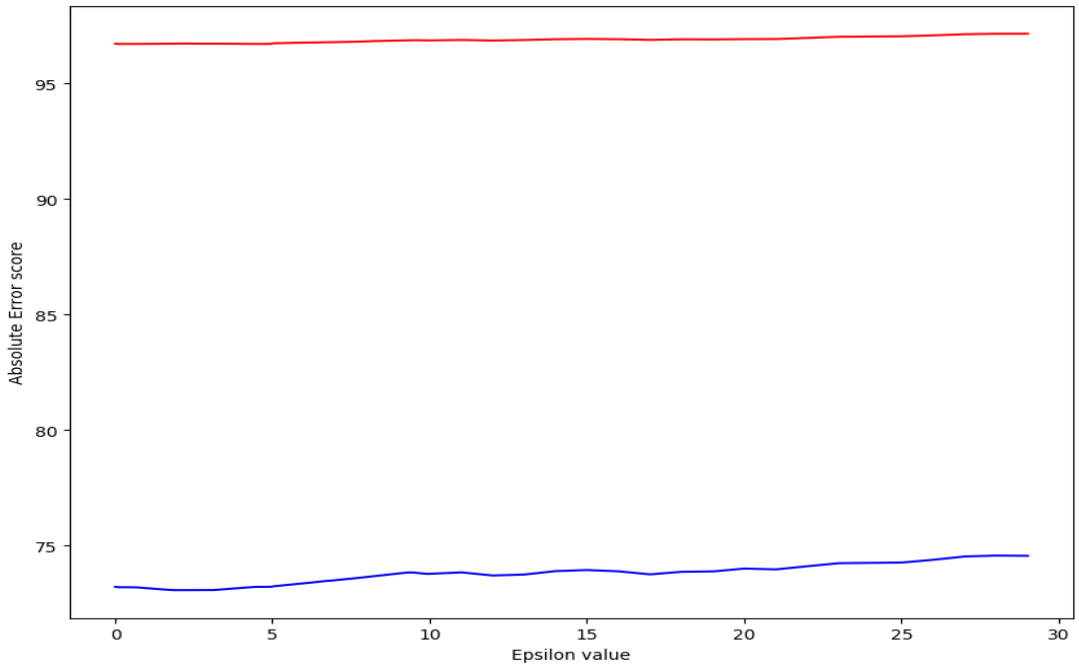


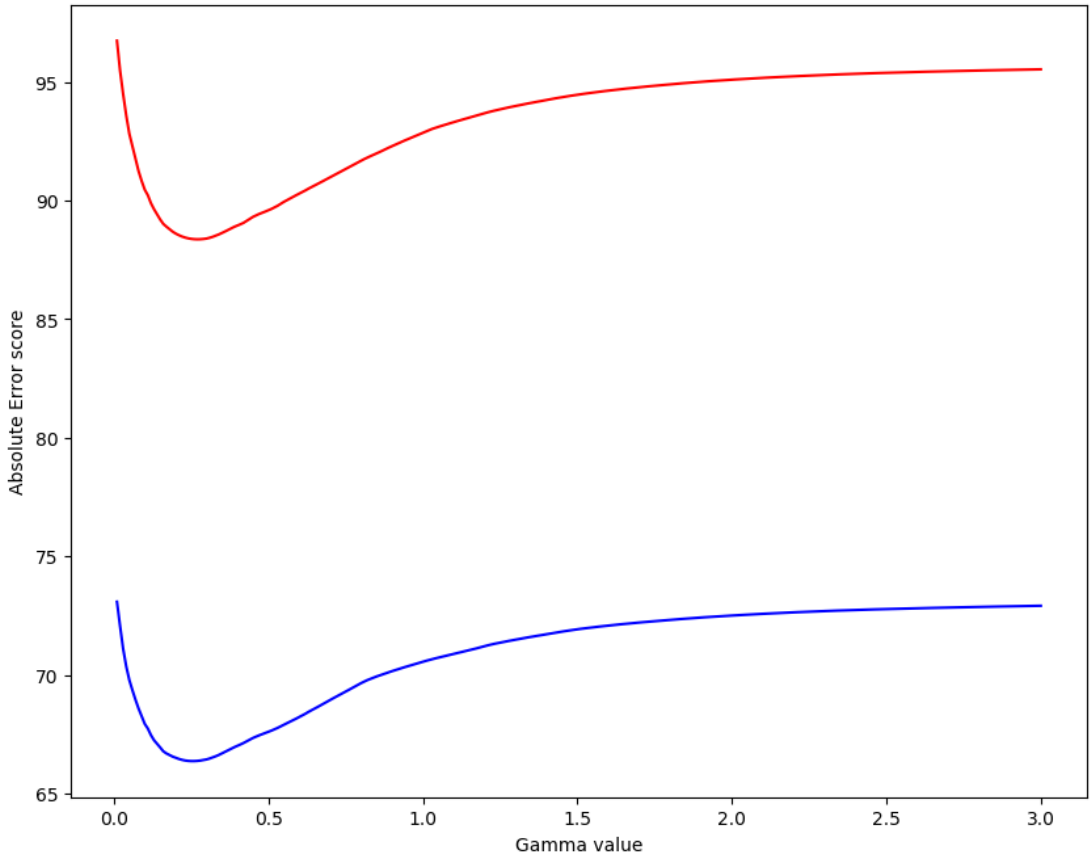


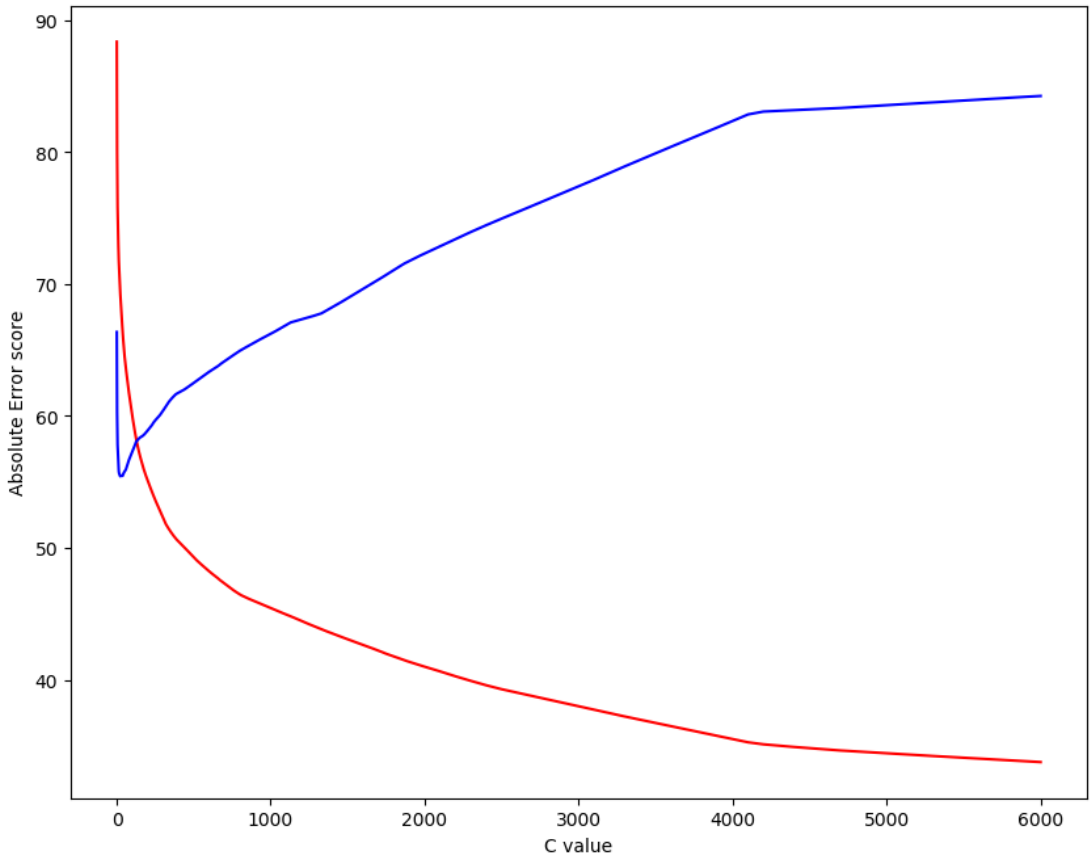


Παρατηρήσεις: Στο τρίτο και πιο δύσκολο Dataset παρατηρούμε παρόμοια συμπεριφορά μόνο που αυτήν την φορά το σφάλμα αργεί περισσότερο αν συγκλίνει.

Αποτελέσματα Polynomial Kernel 10^{ου} βαθμού:

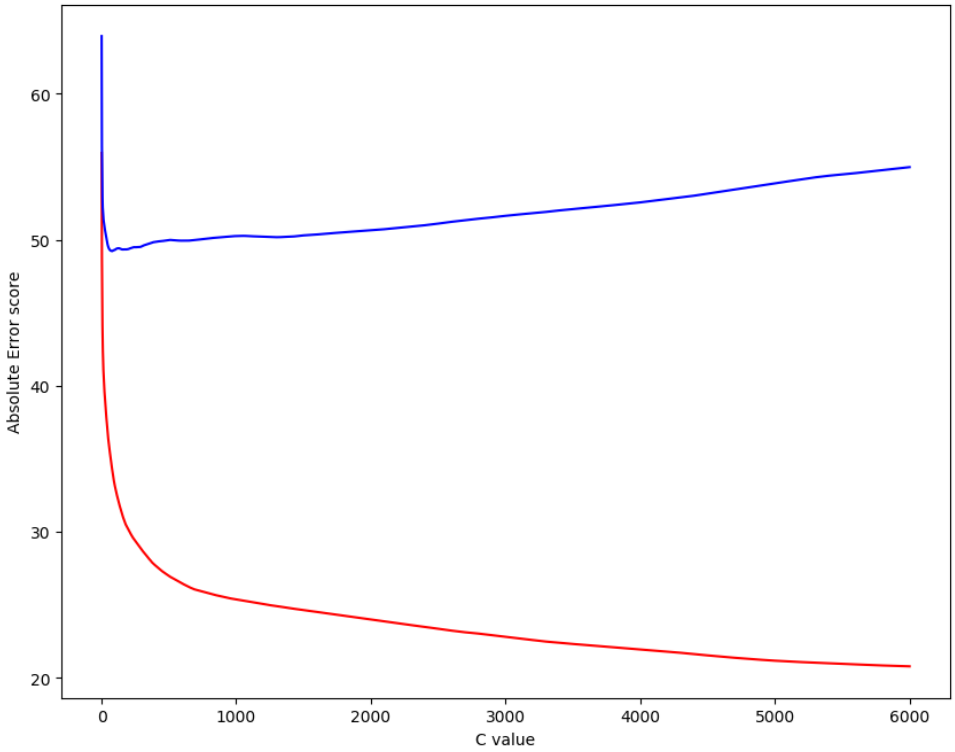



Absolute Error score
95
90
85
80
75
70
65
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Gamma value

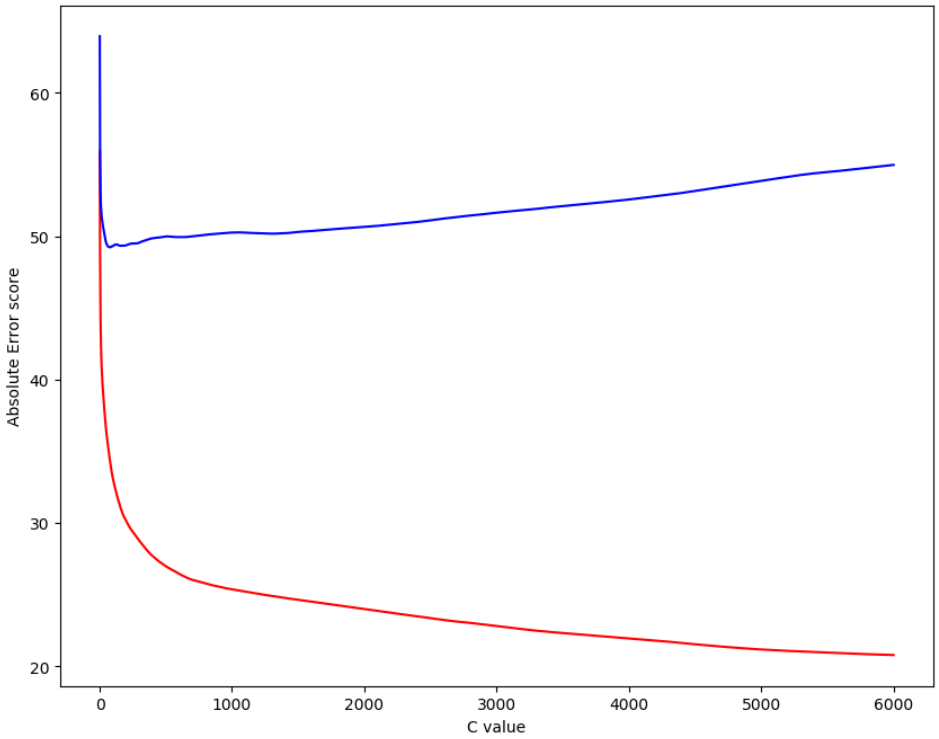

Absolute Error score
90
80
70
60
50
40
0
1000
2000
3000
4000
5000
6000
C value

<u>Παρατηρήσεις:</u> Παρατηρούμε ότι μοιάζουν πάρα πολύ οι καμπύλες των παραμέτρων με τις καμπύλες με RBF Kernel. Τα αποτελέσματα είναι μόνο στο τρίτο και πιο δύσκολο Dataset.

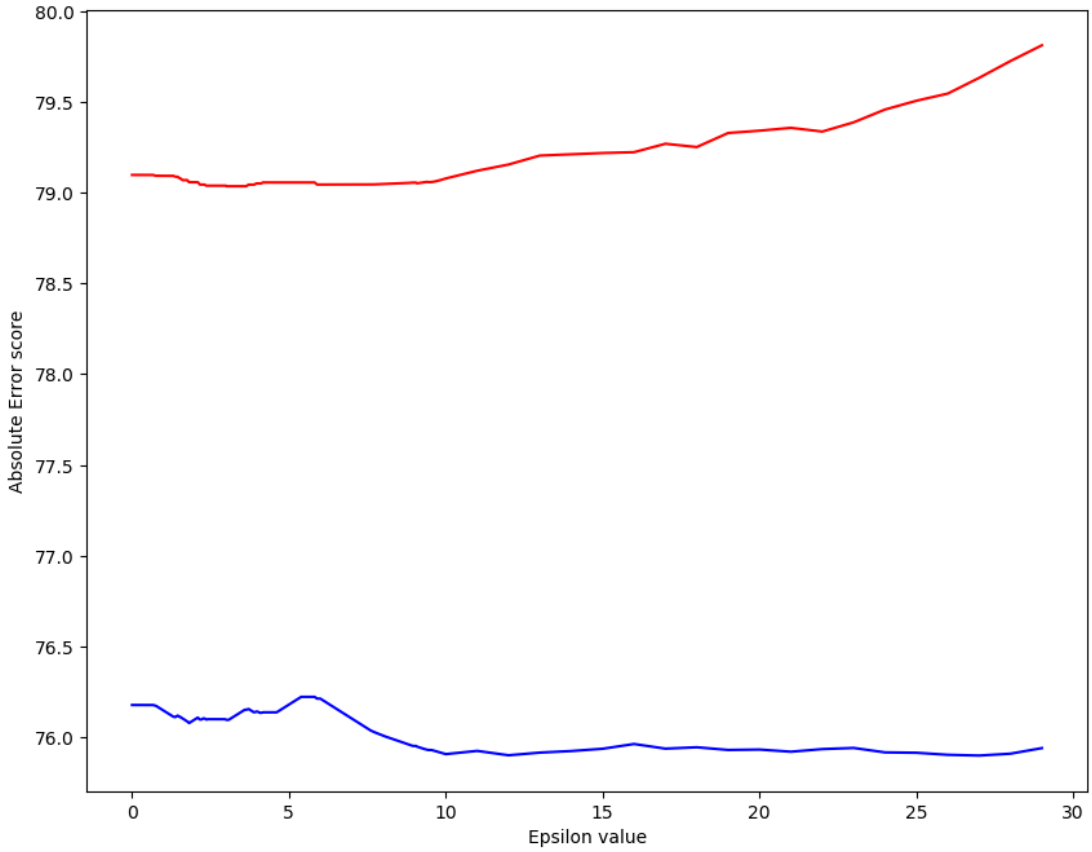
Ανεβάζοντας λίγο τον βαθμό του πολυωνύμου σε 20^{ου} παρατηρούμε βελτίωση στα αποτελέσματα και επιδόσεις καλύτερες από αυτές του RBF:

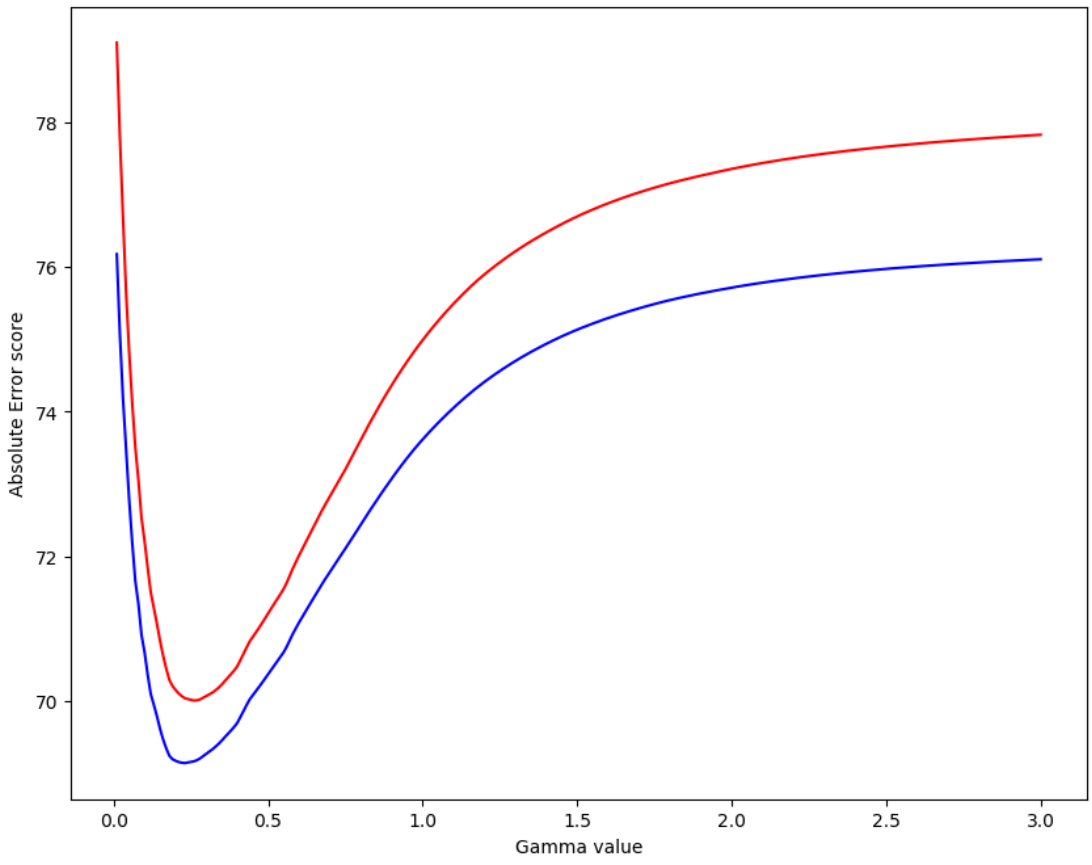


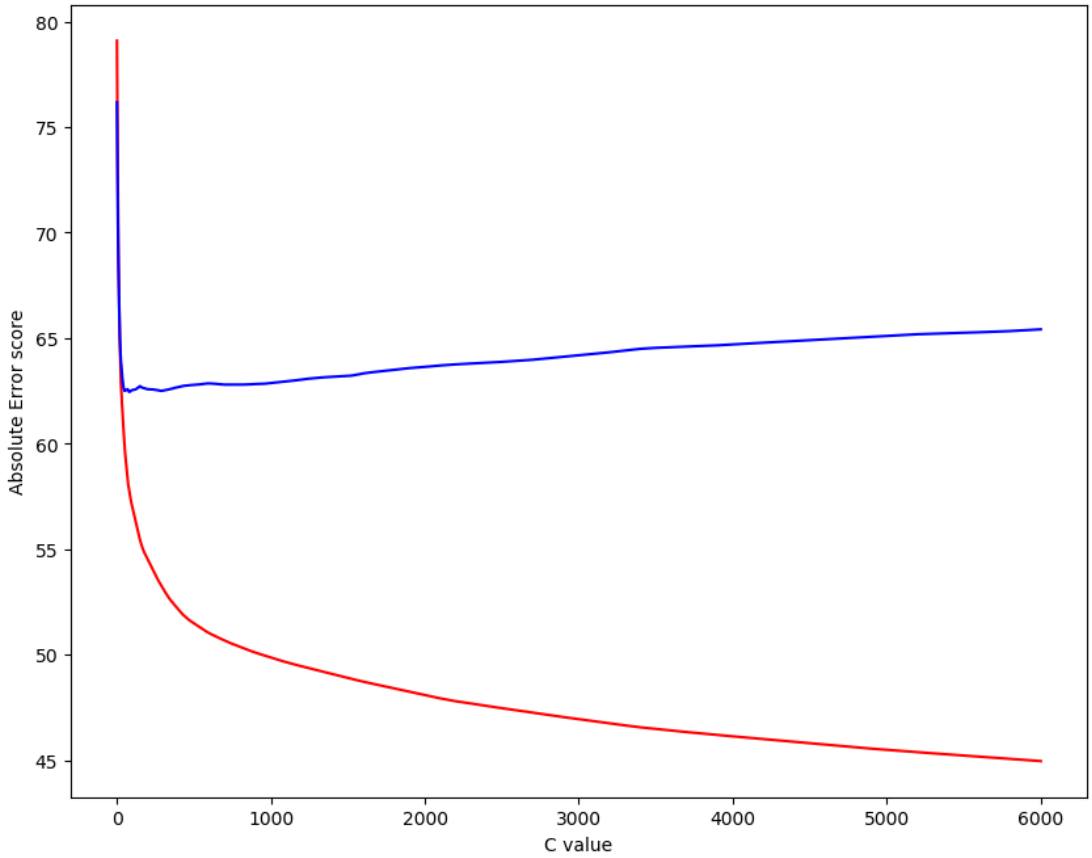
Για 30^{ου} βαθμού πολυώνυμο δεν παρατηρούμε περαιτέρω βελτίωση:



Για Linear Kernel:







<u>Παρατηρήσεις:</u> Παρατηρούμε ότι πάλι οι καμπύλες των παραμέτρων μοιάζουν πολύ στις καμπύλες για RBF Kernel. Επίσης τα αποτελέσματα ακόμα και στο δύσκολο Dataset είναι ικανοποιητικά(φυσικά όχι όσο καλά όσο των Polynomial και RBF).

KNN-regressor: Παρατηρούμε ότι παρότι πιάνει καλά αποτελέσματα ο KNN-regressor, δεν φτάνει σε επιδόσεις τα SVM.

Πρώτο Dataset:

```
With 1 Neighbors the absolute error at test set is: 0.885
With 2 Neighbors the absolute error at test set is: 0.8392499999999999
With 3 Neighbors the absolute error at test set is: 0.8641666666666665
With 4 Neighbors the absolute error at test set is: 0.882125
With 5 Neighbors the absolute error at test set is: 0.8821000000000001
With 6 Neighbors the absolute error at test set is: 0.9108333333333334
With 7 Neighbors the absolute error at test set is: 0.9070714285714284
With 8 Neighbors the absolute error at test set is: 0.8960625
With 9 Neighbors the absolute error at test set is: 0.9098333333333334
With 10 Neighbors the absolute error at test set is: 0.9034
With 11 Neighbors the absolute error at test set is: 0.9065454545454545
With 12 Neighbors the absolute error at test set is: 0.9214583333333334
With 13 Neighbors the absolute error at test set is: 0.9204230769230768
With 14 Neighbors the absolute error at test set is: 0.9157857142857142
With 15 Neighbors the absolute error at test set is: 0.9067333333333333
With 16 Neighbors the absolute error at test set is: 0.9049375000000001
With 17 Neighbors the absolute error at test set is: 0.8884117647058823
With 18 Neighbors the absolute error at test set is: 0.8841111111111111
With 19 Neighbors the absolute error at test set is: 0.8780526315789473
With 20 Neighbors the absolute error at test set is: 0.8750999999999999
With 21 Neighbors the absolute error at test set is: 0.8667857142857142
With 22 Neighbors the absolute error at test set is: 0.8564545454545454
With 23 Neighbors the absolute error at test set is: 0.8528913043478259
With 24 Neighbors the absolute error at test set is: 0.8461874999999999
With 25 Neighbors the absolute error at test set is: 0.84836
With 26 Neighbors the absolute error at test set is: 0.8421923076923076
With 27 Neighbors the absolute error at test set is: 0.8371851851851853
With 28 Neighbors the absolute error at test set is: 0.8335714285714286
With 29 Neighbors the absolute error at test set is: 0.8337931034482758
With 30 Neighbors the absolute error at test set is: 0.8370833333333333
```

Δεύτερο Dataset:

```
With 1 Neighbors the absolute error at test set is: 10.485
With 2 Neighbors the absolute error at test set is: 9.39625
With 3 Neighbors the absolute error at test set is: 9.38
With 4 Neighbors the absolute error at test set is: 9.186875
With 5 Neighbors the absolute error at test set is: 8.8785
With 6 Neighbors the absolute error at test set is: 8.625416666666668
With 7 Neighbors the absolute error at test set is: 8.532142857142858
With 8 Neighbors the absolute error at test set is: 8.2578125
With 9 Neighbors the absolute error at test set is: 8.04111111111111
With 10 Neighbors the absolute error at test set is: 7.970750000000001
With 11 Neighbors the absolute error at test set is: 7.861363636363635
With 12 Neighbors the absolute error at test set is: 7.748124999999999
With 13 Neighbors the absolute error at test set is: 7.72846153846154
With 14 Neighbors the absolute error at test set is: 7.651964285714286
With 15 Neighbors the absolute error at test set is: 7.629833333333334
With 16 Neighbors the absolute error at test set is: 7.6359375
With 17 Neighbors the absolute error at test set is: 7.671617647058825
With 18 Neighbors the absolute error at test set is: 7.69875
With 19 Neighbors the absolute error at test set is: 7.794868421052631
With 20 Neighbors the absolute error at test set is: 7.863875
With 21 Neighbors the absolute error at test set is: 7.906547619047619
With 22 Neighbors the absolute error at test set is: 7.962272727272728
With 23 Neighbors the absolute error at test set is: 8.047608695652174
With 24 Neighbors the absolute error at test set is: 8.131979166666667
With 25 Neighbors the absolute error at test set is: 8.2152
With 26 Neighbors the absolute error at test set is: 8.251057692307693
With 27 Neighbors the absolute error at test set is: 8.330277777777777
With 28 Neighbors the absolute error at test set is: 8.41
With 29 Neighbors the absolute error at test set is: 8.476120689655172
With 30 Neighbors the absolute error at test set is: 8.507583333333335
```

Τρίτο Dataset:

```
With 1 Neighbors the absolute error at test set is: 72.04
With 2 Neighbors the absolute error at test set is: 65.71
With 3 Neighbors the absolute error at test set is: 63.85666666666666
With 4 Neighbors the absolute error at test set is: 63.24375
With 5 Neighbors the absolute error at test set is: 62.005
With 6 Neighbors the absolute error at test set is: 61.98083333333334
With 7 Neighbors the absolute error at test set is: 59.65642857142857
With 8 Neighbors the absolute error at test set is: 59.12125
With 9 Neighbors the absolute error at test set is: 58.22444444444444
With 10 Neighbors the absolute error at test set is: 58.630500000000005
With 11 Neighbors the absolute error at test set is: 59.85727272727273
With 12 Neighbors the absolute error at test set is: 59.78333333333333
With 13 Neighbors the absolute error at test set is: 59.96269230769231
With 14 Neighbors the absolute error at test set is: 61.13964285714286
With 15 Neighbors the absolute error at test set is: 60.559
With 16 Neighbors the absolute error at test set is: 60.063125
With 17 Neighbors the absolute error at test set is: 60.91029411764706
With 18 Neighbors the absolute error at test set is: 61.51166666666667
With 19 Neighbors the absolute error at test set is: 61.820263157894736
With 20 Neighbors the absolute error at test set is: 62.872749999999996
With 21 Neighbors the absolute error at test set is: 63.047380952380955
With 22 Neighbors the absolute error at test set is: 63.82818181818181
With 23 Neighbors the absolute error at test set is: 63.83521739130434
With 24 Neighbors the absolute error at test set is: 64.36604166666667
With 25 Neighbors the absolute error at test set is: 64.527
With 26 Neighbors the absolute error at test set is: 64.20249999999999
With 27 Neighbors the absolute error at test set is: 63.89222222222222
With 28 Neighbors the absolute error at test set is: 63.32178571428572
With 29 Neighbors the absolute error at test set is: 63.61241379310344
With 30 Neighbors the absolute error at test set is: 63.7575
```

***Το μέγεθος των δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων ήταν 500.**

Datasets Links:

https://www.kaggle.com/lava18/google-play-store-apps

https://www.kaggle.com/spscientist/students-performance-in-exams

https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data