Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Εργασία Αναγνώρισης Προτύπων

Κινούς Βασίλειος Αλέξανδρος
8834
Ομάδα 27

5 Ιανουαρίου 2024

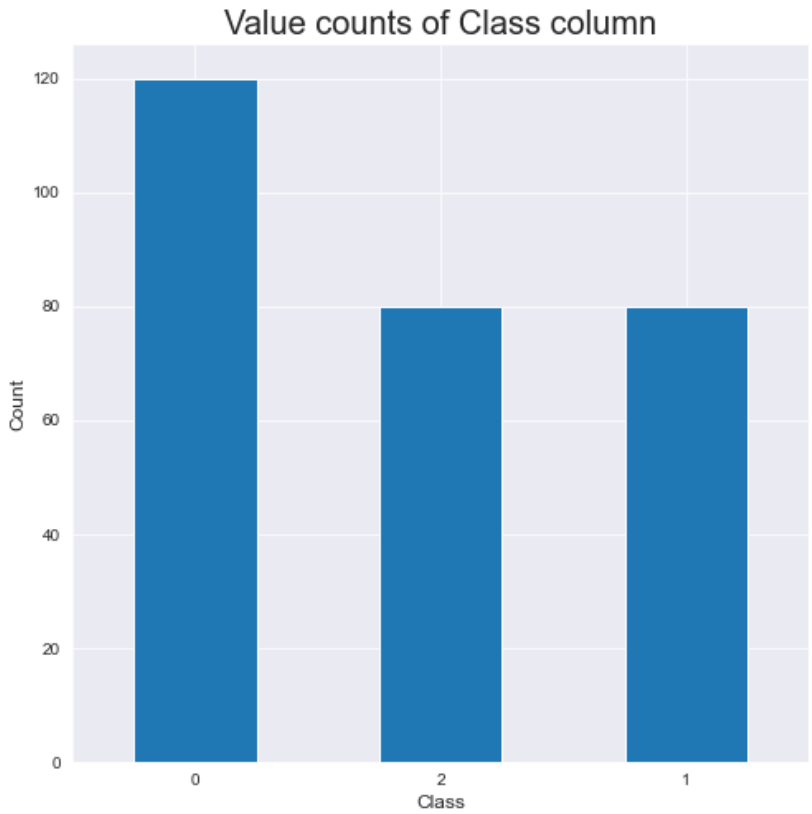
Περιεχόμενα

Κεφάλαιο 1

Άσκηση 1

1.1 Οπτικοποίηση δεδομένων

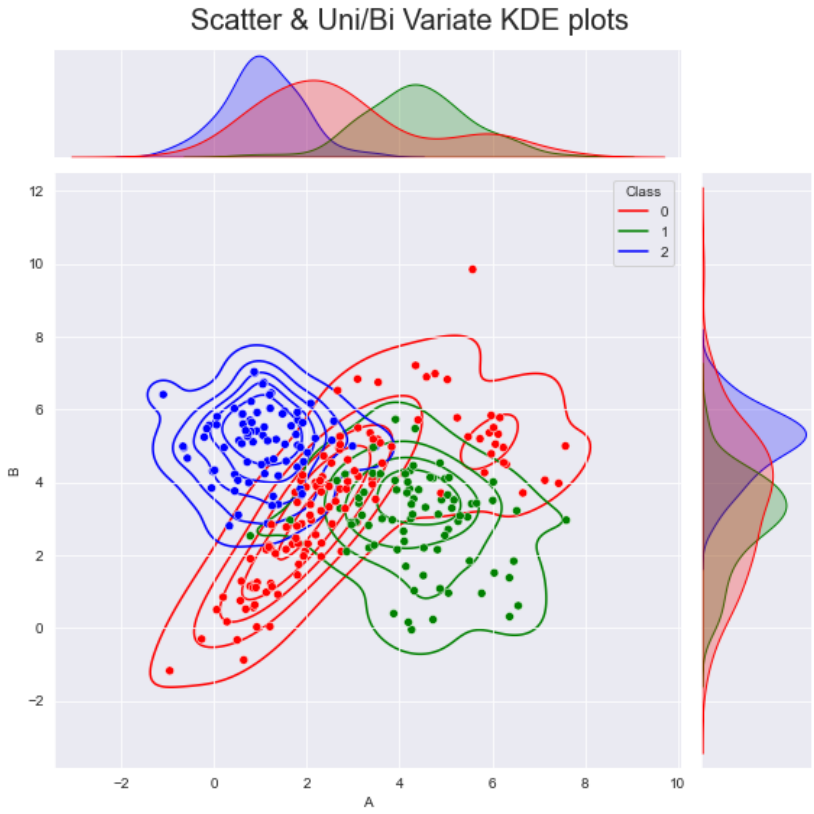
Μελετώντας το αρχειο excel των δεδομενων παρατηρειτε πως εμπεριεχονται 2 διανυσματα χαρακτηριστικων ενω οι διαφορετικες κλασεις ταξινομησης ειναι τρεις.Παρατηρειτε επισης μια μικρη ανισοροπια στις κλασεις καθως η κλαση 0 εμπεριεχει περισσοτερα δειγματα.



Σχήμα 1.1: Συνολο δειγματων που ανηκουν στην εκαστοτε κλαση.

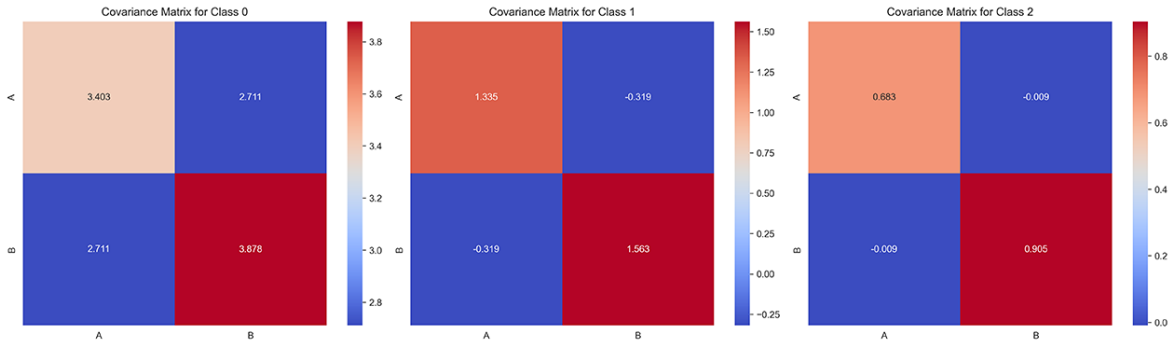
Η kernel density estimation (KDE) είναι μια μέθοδος απεικόνισης της κατανομής των παρατηρήσεων σε ένα σύνολο δεδομένων, ανάλογη με το ιστόγραμμα. Η KDE αναπα-

ριστά τα δεδομένα χρησιμοποιώντας μια συνεχή καμπύλη πυκνότητας πιθανότητας σε μία ή περισσότερες διαστάσεις.



Σχήμα 1.2: Διαγραμμα διασπορας με διμεταβλητη εκτιμηση πυκνοτητας.Παραλληλα και εγκαρσια του διαγραμματος διασπορας οι εκτιμησεις πυρηνα του καθε χαρακτηριστικου της εκαστοτε κλασης.

Διαπιστωνεται οτι τοσο η εκτιμηση πυκνοτητας μιας μεταβλητης οσο και δυο μεταβλητων δεν προσεγγιζει ιδιαιτερως κανονικη κατανομη οποτε μπορουμε να υποθεσουμε οτι μεθοδοι που δεν υποθετουν κανονικη κατανομη δεδομενων θα εχουν αυξημενη ακριβεια προβλεψεων στην εργασια της ταξινομησης δειγματων.
Στη συνεχεια θα υπολογιστουν οι ανα κλαση συνδιασπορες των χαρακτηριστικων και θα παρουσιαστουν σε heatmaps με σκοπο την ευκολη οπτικη αξιολογηση.



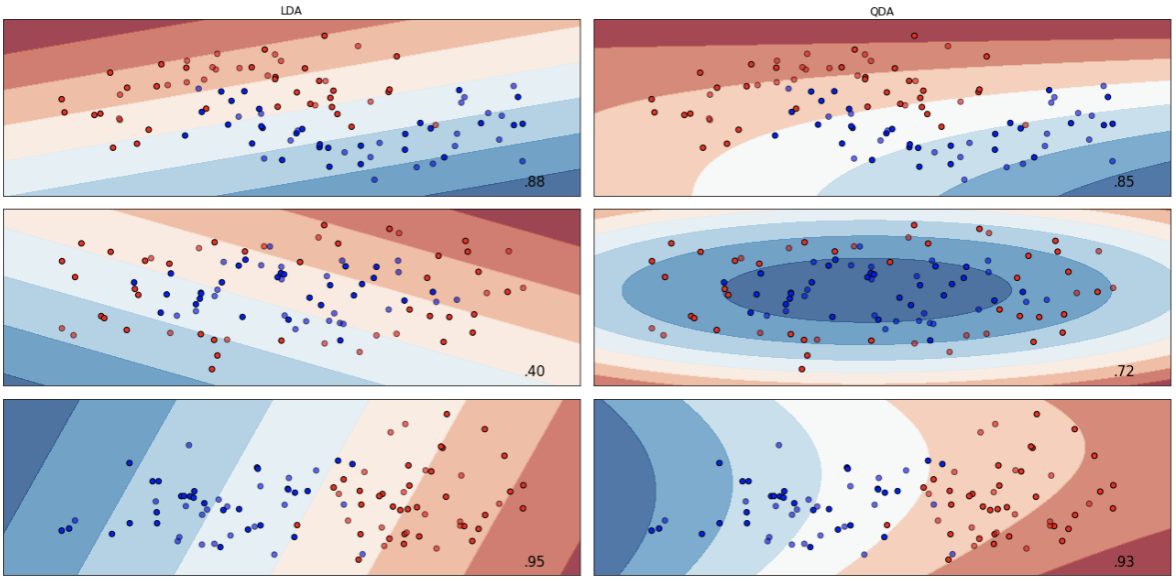
Σχήμα 1.3: Διαγραμματα συνδιασπορας

Διαπιστωνεται οτι υπαρχουν διαφορες μεταξυ των κλασεων οσων αφορα τις συνδιασπορες των χαρακτηριστικων οποτε μπορουμε και σε αυτην την περιπτωση εικαζουμε οτι

μια μεθοδος που δεν υποθετει ιδιο πινακα συνδιασπορας για ολες τις κλασεις θα παρουσιασει μεγαλυτερη ακριβεια προβλεψεων απο μια μεθοδο που υποθετει.

1.2 Linear and Quadratic Discriminant Analysis (LDA/QDA)

Η LDA είναι μια στατιστική μέθοδος που χρησιμοποιείται συνήθως για τη μείωση της διαστατικότητας και την ταξινόμηση. Στόχος της LDA είναι η εύρεση των γραμμικών συνδυασμών χαρακτηριστικών που διαχωρίζουν καλύτερα δύο ή περισσότερες κλάσεις σε ένα σύνολο δεδομένων. Επικεντρώνεται στη μεγιστοποίηση της απόστασης μεταξύ των μέσων όρων διαφορετικών κλάσεων, ενώ ελαχιστοποιεί τη διασπορά ή τη διακύμανση εντός κάθε κλάσης.
Η LDA υποθέτει ότι τα δεδομένα για κάθε κλάση κατανέμονται κανονικά και ότι ο πίνακας συνδιακύμανσης είναι ο ίδιος για όλες τις κλάσεις. Αυτό συνεπάγεται ότι οι κλάσεις έχουν παρόμοια σχήματα και προσανατολισμούς.Υποθέτει επίσης ότι τα δεδομένα είναι γραμμικά διαχωρίσιμα, πράγμα που σημαίνει ότι ένα γραμμικό όριο απόφασης μπορεί να ταξινομήσει με ακρίβεια τις διάφορες κλάσεις.
Η μεθοδος QDA οπως και η LDA υποθετει κανονικη κατανομη δεδομενων με την ειδοποιο διαφορα οτι επιτρεπει διαφορετικο πινακα συνδιακημανσης για καθε κλαση και κατα συνεπεια τετραγωνικα ορια αποφασης.Το QDA παρέχει ευελιξία στον χειρισμό πιο περίπλοκων ορίων αποφάσεων σε σύγκριση με το LDA, καθιστώντας το κατάλληλο για σενάρια όπου οι διακυμάνσεις της εκαστοτε κλασης είναι σημαντικές.



Σχήμα 1.4: Παραδειγμα οριων αποφασης LDA,QDA

1.3 Αποτελεσματα και σχολιασμος

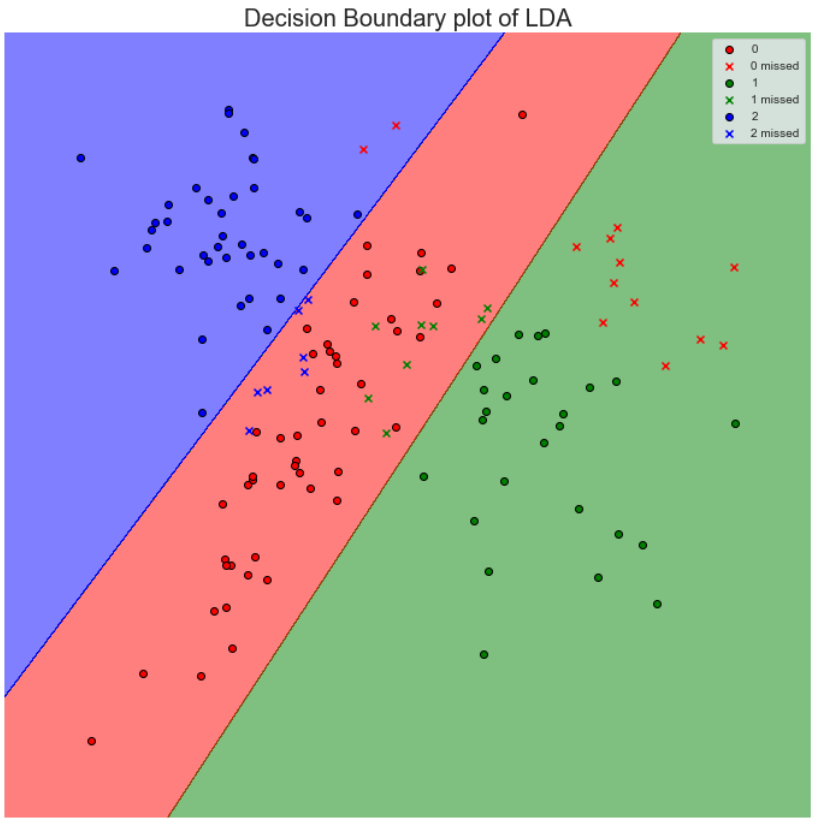
Αφου χωριστουν τα δεδομενα τυχαια κατα το ημισυ σε τραιν και τεστ θα εκπαιδευτουν μεσω της βιβλιοθηκης sklearn ενας lda και ενας qda ταξινομητης. Τα εκπαιδευμενα μοντελα θα χρησιμοποιηθουν για προβλεψεις επανω στο test set και θα υπολογιστει η ακριβεια προβλεψεων.
Παρακατω θα αναλυθουν οι μετρικες που θα χρησιμοποιηθουν για τα πειραματα.

1. Accuracy
    - Μετρά τη συνολική ορθότητα του ταξινομητή.
    - Ισούται με την αναλογία των σωστά προβλεφθέντων περιπτώσεων προς το σύνολο των περιπτώσεων.
    - Αγνοεί τις ανισορροπίες των κλάσεων και αντιμετωπίζει όλες τις κλάσεις ισότιμα.
2. Precision
    - Μετρά την ακρίβεια των θετικών προβλέψεων για κάθε κλάση.
    - Η σταθμισμένη ακρίβεια λαμβάνει υπόψη τις ανισορροπίες των κλάσεων λαμβάνοντας υπόψη τον αριθμό των αληθώς θετικών προβλέψεων για κάθε κλάση.
3. Recall
    - Μετρά την ικανότητα του ταξινομητή να συλλαμβάνει όλες τις θετικές περιπτώσεις για κάθε κλάση.
    - Η σταθμισμένη ανακληση λαμβάνει υπόψη τις ανισορροπίες των κλάσεων και ισούται με την ακρίβεια.
4. F1-Score
    - Αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης.
    - Η σταθμισμένη βαθμολογία F1 εξισορροπεί την ακρίβεια και την ανάκληση για κάθε κλάση, λαμβάνοντας υπόψη τις ανισορροπίες των κλάσεων.

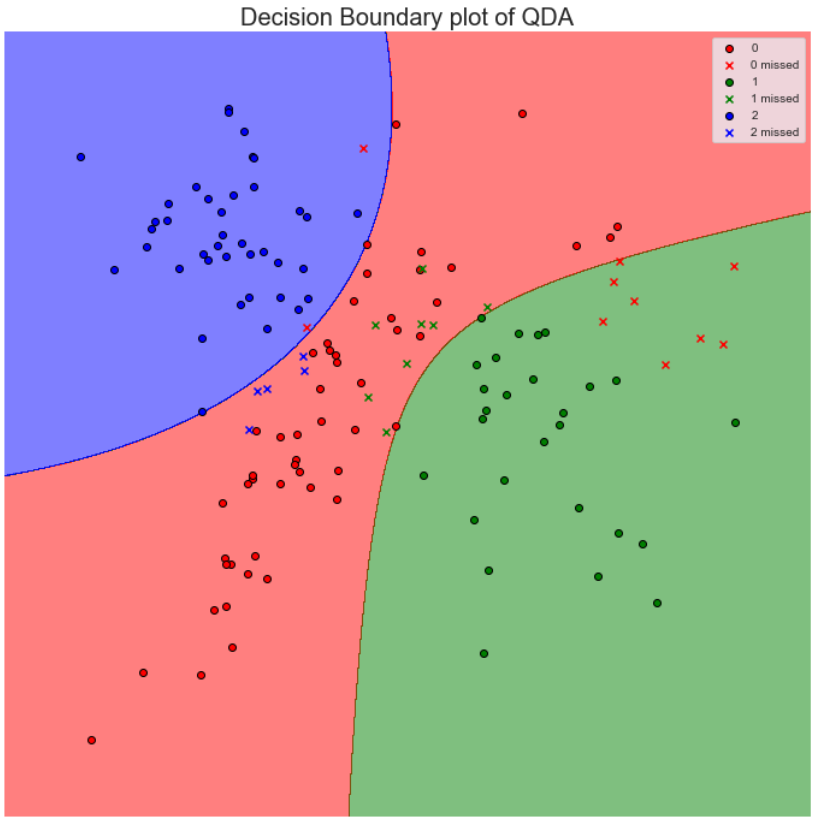
Πίνακας 1.1: Μετρικες στο συνολο δοκιμων για τις μεθοδους LDA,QDA

Metric \ Method	LDA	QDA
• Accuracy	79.28%	83.57%
• Precision	80.06%	83.90%
• Recall	79.28%	83.57%
• F1-score macro	79.53%	83.68%

Οπως βλεπουμε σε καθε διαφοτερικη μετρικη η μεθοδος qda υπερτερει της lda . Αυτο αποδιδεται στις χαλαρες υποθεσεις της qda οσων αφορα τους πινακες συνδιακημανσης και στα ευλικτα τετραγωνικα ορια αποφασεων.



Σχήμα 1.5: Ορια αποφασεων LDA στο test set



Σχήμα 1.6: Ορια αποφασεων QDA στο test set

Κεφάλαιο 2

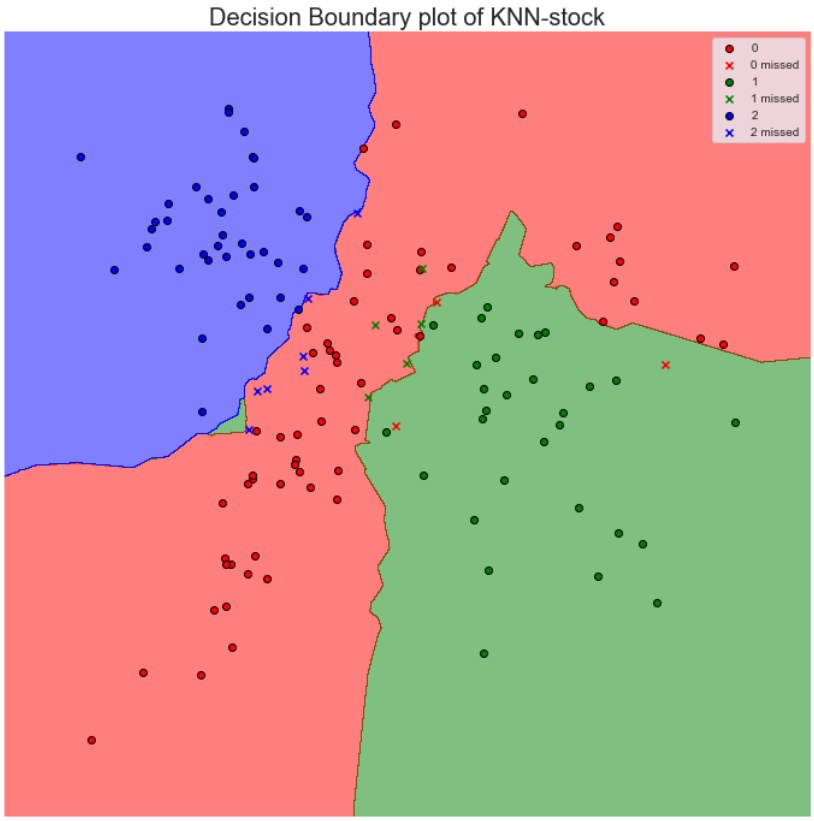
Άσκηση 2

2.1 Knn-Classifier

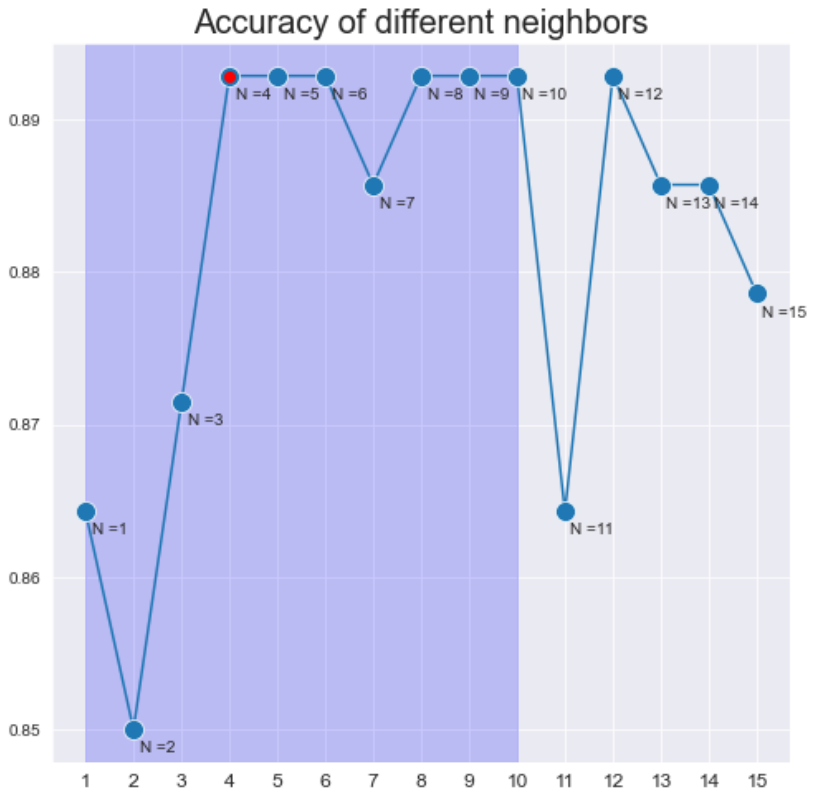
Στην ταξινόμηση πολλαπλών κατηγοριών, ο αλγόριθμος k-κοντινότεροι γείτονες (KNN) επεκτείνει την εφαρμογή του για την πρόβλεψη της κατηγορίας ενός νέου σημείου δεδομένων με βάση την πλειοψηφική κατηγορία μεταξύ των k-κοντινότερων γειτόνων του. Σε αντίθεση με τη δυαδική ταξινόμηση, όπου το αποτέλεσμα είναι μία από δύο κλάσεις, η ταξινόμηση πολλαπλών κλάσεων περιλαμβάνει την πρόβλεψη μεταξύ τριών ή περισσότερων κλάσεων.
Η KNN λειτουργεί με βάση την αρχή της εγγύτητας, υποθέτοντας ότι τα σημεία δεδομένων που ανήκουν στην ίδια κλάση βρίσκονται κοντά το ένα στο άλλο στο χώρο των χαρακτηριστικών. Όταν παρουσιάζεται ένα νέο σημείο δεδομένων, ο αλγόριθμος εντοπίζει τους k-κοντινότερους γείτονες από το σύνολο εκπαίδευσης και αποδίδει την ετικέτα κλάσης που είναι πιο διαδεδομένη μεταξύ αυτών των γειτόνων.
Η επιλογή του ¨k¨ αποτελεί κρίσιμο παράγοντα για την απόδοση του αλγορίθμου. Ένα μικρότερο ¨k¨ μπορεί να οδηγήσει σε ένα πιο ευαίσθητο μοντέλο, το οποίο ενδεχομένως να καταγράφει το θόρυβο στα δεδομένα, ενώ ένα μεγαλύτερο ¨k¨ μπορεί να εξομαλύνει τα όρια των αποφάσεων, χάνοντας ενδεχομένως λεπτά μοτίβα.
Ο αλγόριθμος χρησιμοποιεί μια μετρική απόστασης, συχνά την ευκλείδεια απόσταση, για τη μέτρηση της ομοιότητας μεταξύ σημείων δεδομένων. Ωστόσο, η αποτελεσματικότητα του KNN μπορεί να επηρεαστεί από την κλίμακα και τη συνάφεια των χαρακτηριστικών, καθιστώντας απαραίτητη την κατάλληλη προεπεξεργασία των δεδομένων.

2.2 Εκπαιδευση και αποτελεσματα

Αρχικα θα εκπαιδευτει ενα knn μοντελο με τις βασικες υπερπαραμετρους (αριθμος γειτονων = 5) .Τοσο η εκπαιδευση οσο και οι προβλεψεις θα υλοποιηθουν στα ιδια δεδομενα με προηγουμενως. Η ακριβεια προκυπτει 89.29%
Στην συνεχεια θα εκπαιδευτουν διαφορετικοι ταξινομητες με αριθμο γειτονων απο 1-15 (με εμφαση απο 1-10 που ζητειται) και θα παρουσιαστει η ακριβεια και τα ορια αποφασεων τους.

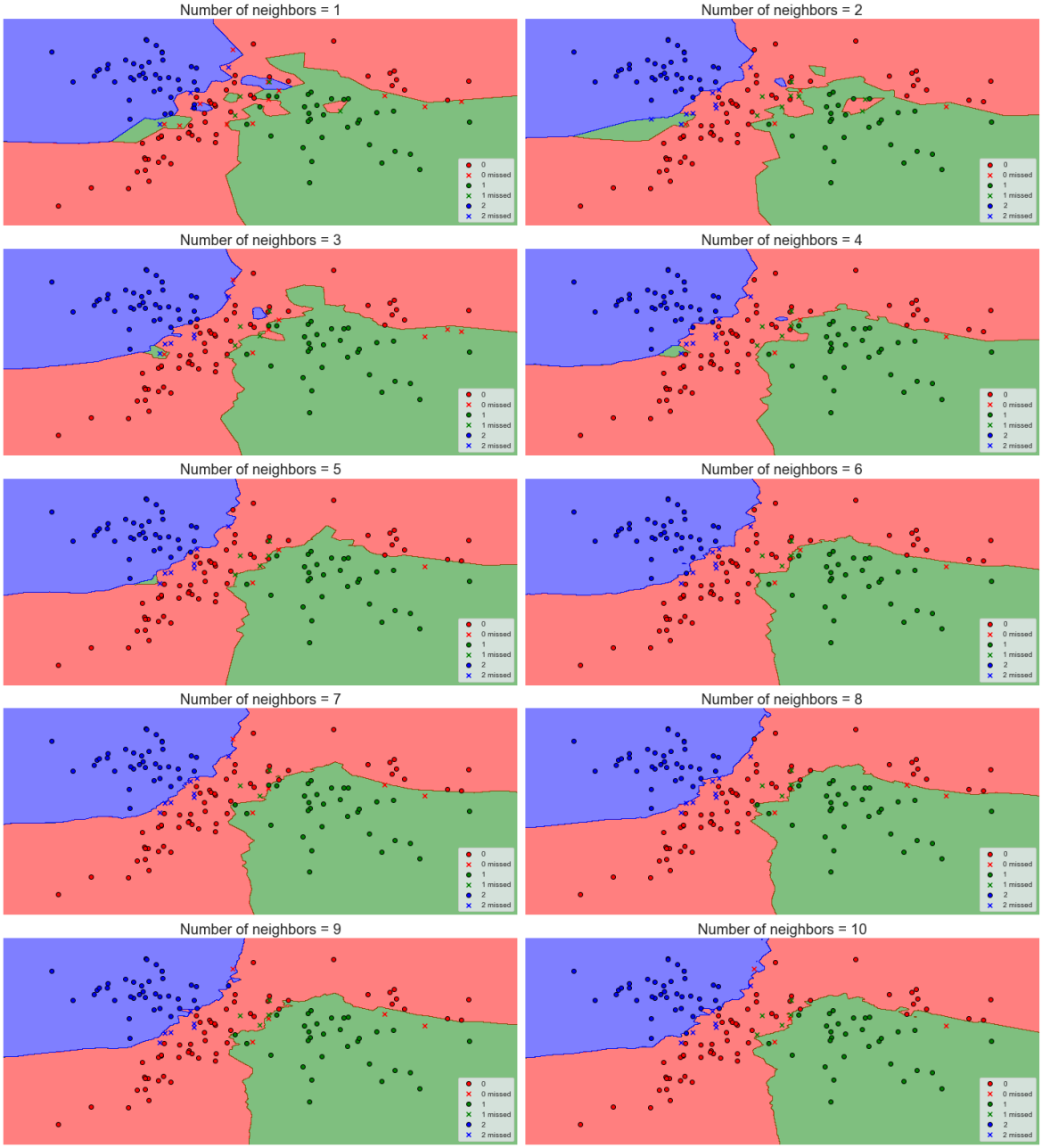


Σχήμα 2.1: Ορια αποφασεων ταξινομητη knn με αριθμο γειτονων 5.



Σχήμα 2.2: Ακριβεια διαφορετικων γειτονων.

Παρατηρουμε οτι η μεγιστη τιμη της ακριβειας εμφανιζεται στους 4 γειτονες και επειτα ειτε μενει σταθερη ειτε μειωνεται. Μικροτερος αριθμος γειτονων δημιουργει εναν ταξινομητη ευαισθητο στο θορυβο και στις ακραιες τιμες ενω πολυ μεγαλος αριθμος παραγει ομαλοτερες επιφανειες αποφασεων χανοντας ετσι τις λεπτομερειες.



Σχήμα 2.3: Ορια αποφασεων διαφορετικων γειτονων.

Πίνακας 2.1: Σύγκριση KNN με τις μεθοδους LDA,QDA

Metric \ Method	LDA	QDA	KNN-5Neigh
• Accuracy	79.28%	83.57%	89.28%
• Precision	80.06%	83.90%	90.27%
• Recall	79.28%	83.57%	89.28%
• F1-score	79.53%	83.68%	89.35%

Η διαφορα στην ακριβεια μεταξυ knn και lda-qda μπορει να αποδοθει στο γεγονος οτι η Β με καταλληλο αριθμο γειτονων για το εκαστοτε προβλημα δεν υποθετει τιποτα για την κατανομη των δεδομενων σε αντιθεση με τις lda-qda.Οπως ειδαμε και στην οπτικοποιηση δεδομενων η υποθεση της κανονικης κατανομης δεν ηταν βελτιστη για τα δεδομενα του προβληματος.Επισης η knn μπορεί να αποτυπώσει πολύπλοκα όρια αποφάσεων, τα οποία είναι καταλληλότερα για το σύνολο δεδομένων του προβληματος.

Κεφάλαιο 3

Άσκηση 3

3.1 Linear-SVM

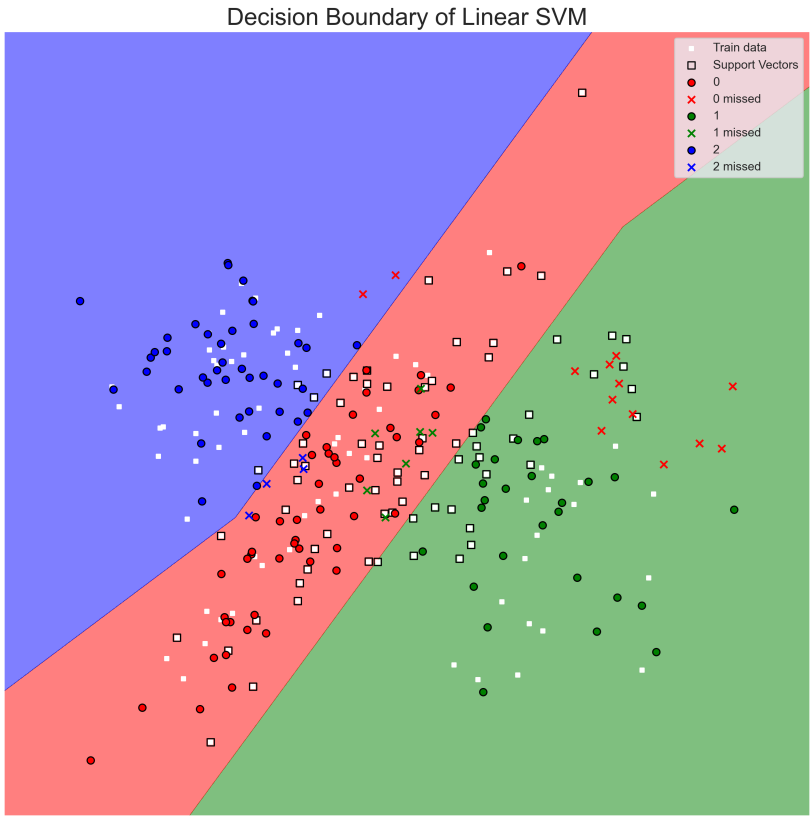
Η μηχανή διανυσμάτων υποστήριξης (SVM) είναι ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη που χρησιμοποιείται για εργασίες ταξινόμησης και παλινδρόμησης. Στο πλαίσιο της ταξινόμησης, η Γραμμική SVM είναι μια παραλλαγή της SVM που επικεντρώνεται στην εύρεση ενός υπερεπιπέδου στο χώρο χαρακτηριστικών που διαχωρίζει καλύτερα τις κλάσεις.
Ο πρωταρχικός στόχος του Γραμμικού SVM είναι ο προσδιορισμός ενός ορίου απόφασης (υπερεπίπεδο) που διαχωρίζει στο μέγιστο βαθμό τα σημεία δεδομένων διαφορετικών κλάσεων. Αυτό το υπερεπίπεδο επιλέγεται με τέτοιο τρόπο ώστε να μεγιστοποιείται το περιθώριο, το οποίο είναι η απόσταση μεταξύ του υπερεπιπέδου και του πλησιέστερου σημείου δεδομένων οποιασδήποτε κλάσης.
Τα διανύσματα υποστήριξης είναι τα σημεία δεδομένων που βρίσκονται πλησιέστερα στο όριο απόφασης και συμβάλλουν στον καθορισμό του περιθωρίου.Επιτελούν σημαντικό ρόλο επειδή έχουν τη δυνατότητα να επηρεάσουν τη θέση και τον προσανατολισμό του ορίου απόφασης.
Η γραμμική φύση του SVM συνεπάγεται ότι το όριο απόφασης είναι μια ευθεία γραμμή σε έναν δισδιάστατο χώρο ή ένα υπερεπίπεδο σε έναν χώρο υψηλότερων διαστάσεων.

3.1.1 Αποτελέσματα και σχολιασμός

Οπως και προηγουμενως θα εκπαιδευτει και θα εξεταστει ενας γραμμικος SVM στα ιδια δεδομενα με την υπερπαραμετρο C να ισουται με 4. Η ακριβεια προκυπτει 82.86%

Πίνακας 3.1: Σύγκριση KNN με τις μεθοδους LDA,QDA

Metric \ Method	LDA	QDA	KNN-5Neigh	Linear-SVM
• Accuracy	79.28%	83.57%	89.28%	82.85%
• Precision	80.06%	83.90%	90.27%	83.32%
• Recall	79.28%	83.57%	89.28%	82.85%
• F1-score	79.53%	83.68%	89.35%	83.01%



Σχήμα 3.1: Ορια αποφασεων ταξινομητη λινεαρ σμ.

Οπως ηταν αναμενομενο μεσω των προηγουμενων ασκησεων οι γραμμικες μεθοδοι δεν ειναι καταλληλες για τα δεδομενα του προβληματος και ολες οι μετρικες ειναι αισθητα μικροτερες απ τις αντιστοιχες του knn.

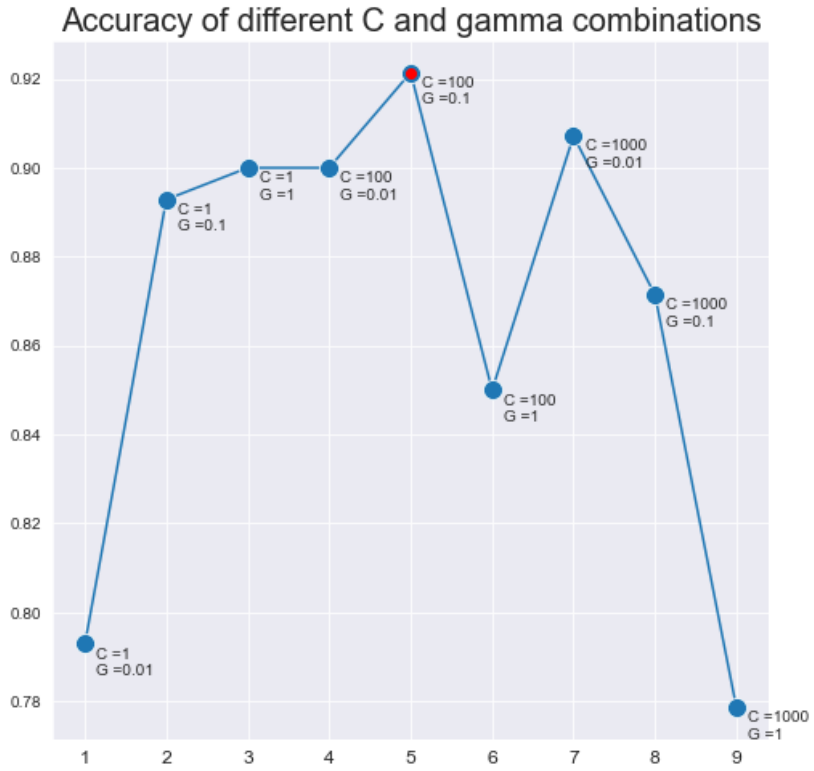
3.2 RBF-Kernel-SVM

Ο πυρήνας RBF είναι ένας τύπος συνάρτησης πυρήνα που χρησιμοποιείται στο SVM για τον χειρισμό μη γραμμικών σχέσεων μεταξύ των χαρακτηριστικών. Ο πυρήνας RBF επιτρέπει στο SVM να προβάλλει τα δεδομένα σε έναν χώρο υψηλότερων διαστάσεων, καθιστώντας δυνατή την εύρεση ενός μη γραμμικού ορίου απόφασης. Αυτό είναι ιδιαίτερα χρήσιμο όταν η υποκείμενη σχέση μεταξύ των χαρακτηριστικών είναι πολύπλοκη και δεν μπορεί να αποτυπωθεί αποτελεσματικά από ένα γραμμικό όριο απόφασης.
Στο SVM με πυρήνα RBF, η συνάρτηση απόφασης ορίζεται με βάση ένα μέτρο ομοιότητας, συγκεκριμένα τη συνάρτηση ακτινικής βάσης. Η συνάρτηση απόφασης γίνεται ένα σταθμισμένο άθροισμα συναρτήσεων ακτινικής βάσης που εφαρμόζεται στις αποστάσεις ανά ζεύγη μεταξύ των σημείων δεδομένων και ενός συνόλου παραμέτρων.
Τα διανύσματα υποστήριξης είναι οι περιπτώσεις που συμβάλλουν στην υποστήριξη της επιφάνειας απόφασης. Είναι τα σημεία δεδομένων των οποίων οι αποστάσεις ανά ζεύγη με άλλα σημεία επηρεάζουν τον υπολογισμό της συνάρτησης πυρήνα RBF. Αυτές οι περιπτώσεις είναι ζωτικής σημασίας για την ικανότητα του μοντέλου να γενικεύει καλά σε νέα δεδομένα και να χειρίζεται πολύπλοκες, μη γραμμικές σχέσεις.
Η **παράμετρος gamma** καθορίζει το πλάτος της συνάρτησης πυρήνα RBF. Μια μικρή τιμή γάμμα σημαίνει έναν ευρύ πυρήνα και τα σημεία που απέχουν περισσότερο μεταξύ

τους θεωρούνται παρόμοια. Αντίθετα, μια μεγάλη τιμή gamma οδηγεί σε στενό πυρήνα και τα σημεία πρέπει να είναι πολύ κοντά για να θεωρηθούν παρόμοια. Η ρύθμιση του gamma μπορεί να επηρεάσει σημαντικά την ομαλότητα και την πολυπλοκότητα του ορίου απόφασης.
Η **παράμετρος C** είναι η παράμετρος κανονικοποίησης που ελέγχει τον συμβιβασμό μεταξύ της επίτευξης ενός ομαλού ορίου απόφασης και της ορθής ταξινόμησης των σημείων εκπαίδευσης. Ένα μικρότερο C δίνει έμφαση σε ένα πιο ομαλό όριο απόφασης, επιτρέποντας ενδεχομένως κάποιες λανθασμένες ταξινομήσεις, ενώ ένα μεγαλύτερο C τιμωρεί τις λανθασμένες ταξινομήσεις πιο έντονα, με αποτέλεσμα ένα πιο πολύπλοκο όριο απόφασης.

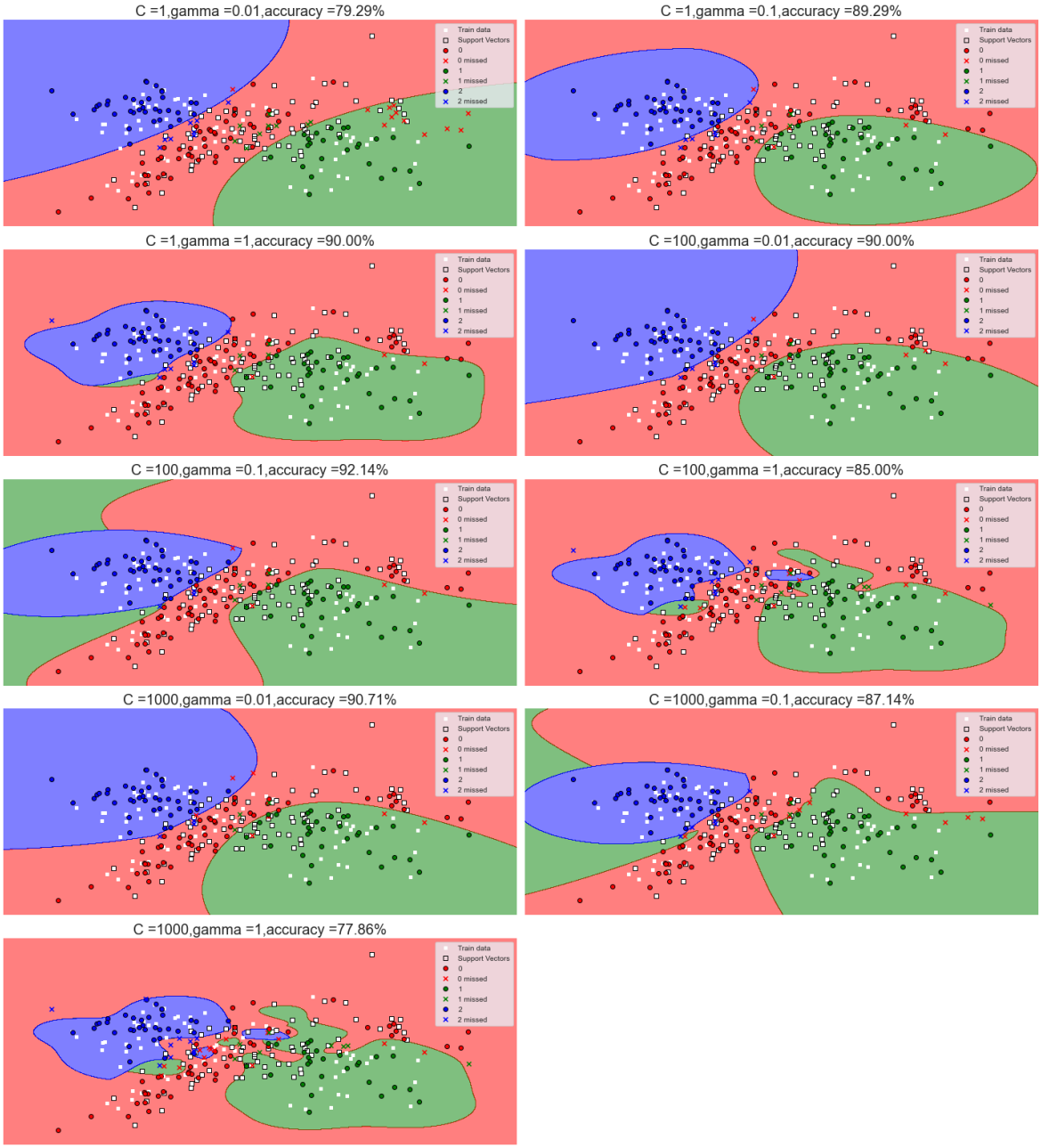
3.2.1 Αποτελέσματα και σχολιασμός

Θα υλοποιηθει ενας ελεγχος διαφορετικων υπερπαραμετρων C και γαμμα για το svm στα ιδια δεδομενα με τα προηγουμενα ερωτηματα και θα παρουσιαστουν τοσο οι ακριβεια οσο και τα ορια αποφασεων των διαφορετικων υπερπαραμετρων.



Σχήμα 3.2: Ακριβεια ταξινομητη rbf kernel svm για διαφορετικες υπερπαραμετρους.

Παρατηρουμε οτι η μεγιστη ακριβεια του ταξινομητη προκυπτει απ τις υπερπαραμετρους C =100 , gamma = 0.1 οποτε θα γινει μια τελικη επανεκπαιδευση με αυτες τις υπερπαραμετρους για εξαγωγη ολων των μετρικων οι οποιες θα παρουσιαστουν παρακατω.



Σχήμα 3.3: Ορια αποφασεων ταξινομητη rbf kernel svm για διαφορετικες υπερπαραμετρους.

Πίνακας 3.2: Σύγκριση RBF-SVM με τις μεθοδους LDA,QDA,KNN,LinearSVM

Metric \ Method	LDA	QDA	KNN-5Neigh	Linear-SVM	RBF-SVM
• Accuracy	79.28%	83.57%	89.28%	82.85%	92.14%
• Precision	80.06%	83.90%	90.27%	83.32%	92.24%
• Recall	79.28%	83.57%	89.28%	82.85%	92.14%
• F1-score	79.53%	83.68%	89.35%	83.01%	92.16%

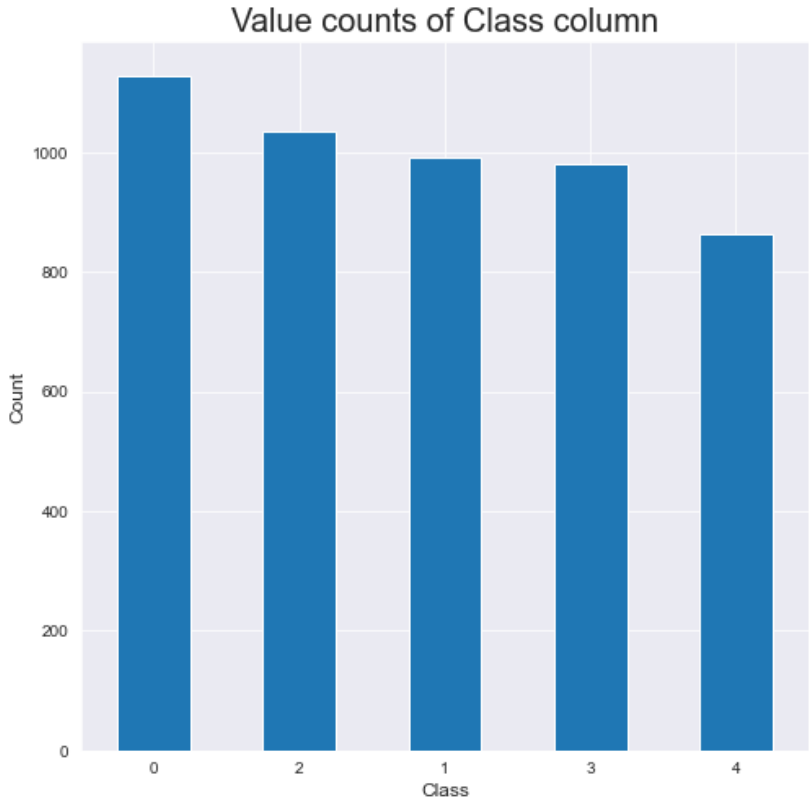
Ο πυρήνας RBF SVM είναι ιδιαίτερα αποτελεσματικός όταν αντιμετωπίζει μη γραμμικές σχέσεις στα δεδομένα. Εάν το όριο απόφασης είναι πολύπλοκο και μη γραμμικό, ο πυρήνας RBF μπορεί να αντιστοιχίσει τα χαρακτηριστικά εισόδου σε έναν χώρο υψηλότερων διαστάσεων, επιτρέποντας ένα πιο ευέλικτο όριο απόφασης.Επίσης οι SVM είναι γενικά ανθεκτικοί στις ακραίες τιμές και στον θόρυβο.Γενικότερα ο RBF SVM ηταν ο καταλληλότερος ταξινομητής για το πρόβλημα καθως ούτε επιτελεί εσωτερικές υποθέσεις για την κατανομή των δεδομένων ενω η μη γραμμικότητα και η ανθεκτικότητα του ,του επιτρέπουν να διακρίνει ακριβέστερα τις σχέσεις μεταξύ των δεδομένων.

Κεφάλαιο 4

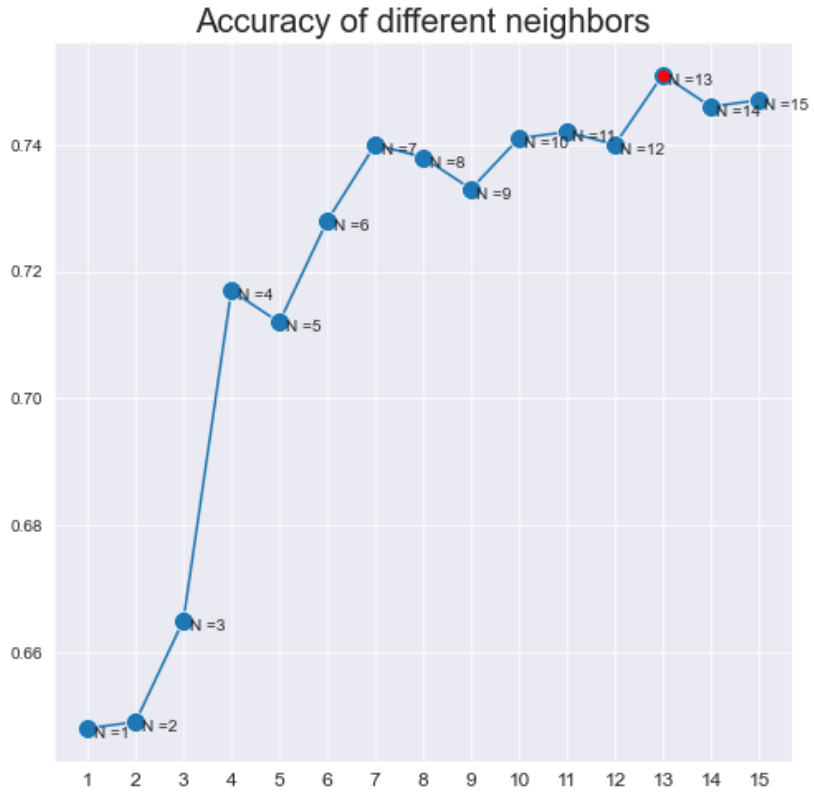
Άσκηση 4

4.1 Dataset C

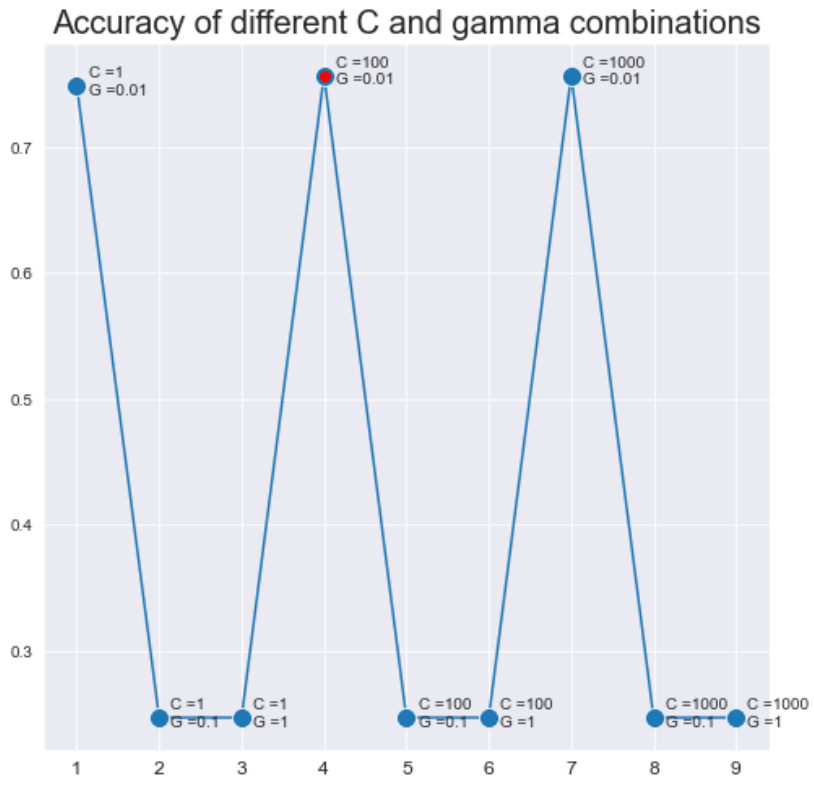
Κοιτώντας το αρχειο excel της τεταρτης ασκησης βλεπουμε οτι αυτη τη φορα προκειται για προβλημα ταξινομησης 5 κλασεων ,τα δειγματα ειναι 5000 και τα χαρακτηριστικα 400. Σαν μια πρωτη κινηση θα δοκιμασουμε να εκπαιδευσουμε τα καλυτερα μοντελα των προηγουμενων ερωτηματων (knn,rbf-svm) στο προβλημα για να δουμε τι ακριβεια επιτυγχανεται χωρις καποια προεπεξεργασια δεδομενων. Τα δεδομενα θα χωριστουν και το test set θα ειναι το 20% του train (1000 δειγματα) δηλαδη οσο και το εξωτερικο αρχειο υνλαβελεδ προβλεψεων που μας ενδιαφερει.



Σχήμα 4.1: Ανισορροπία των κλάσεων στο dataset C.



Σχήμα 4.2: Ακριβεια knn στο dataset C.



Σχήμα 4.3: Ακριβεια rbf svm στο dataset C.

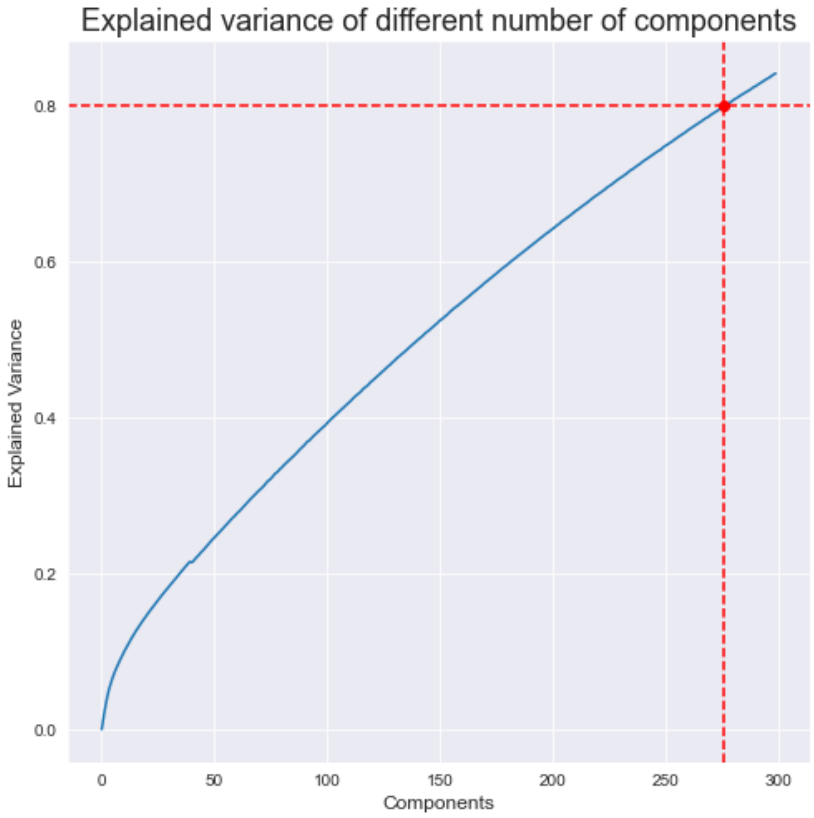
Πίνακας 4.1: Σύγκριση knn με rbf svm στο Dataset C

Metric \ Method	KNN	RBF-SVM
• Accuracy	75.10%	75.70%
• Precision	78.80%	79.23%
• Recall	75.10%	75.70%
• F1-score	74.23%	72.00%

Παρολο που ο rbf-svm εχει μεγαλυτερη ακριβεια και ανακληση το f1-score του ειναι χαμηλοτερο απο το knn.Το σταθμισμένο σκορ F1 θεωρεί τον σταθμισμένο μέσο όρο της ακρίβειας και της ανάκλησης, όπου τα βάρη καθορίζονται από τον αριθμό των περιπτώσεων σε κάθε κατηγορία. Εάν ο knn έχει μεγαλύτερη ακρίβεια και ανάκληση σε κλάσεις με περισσότερες περιπτώσεις, θα μπορούσε να συνεισφέρει περισσότερο στη σταθμισμένη βαθμολογία F1, ακόμη και αν η ακρίβεια και η ανάκληση είναι μεμονωμένα χαμηλότερες σε ορισμένες κλάσεις.
Σε καθε περιπτωση τα αποτελεσματα δεν ειναι ικανοποιητικα και μια πρωτη σκεψη ειναι η μειωση της διαστατικοτητας των δεδομενων μεσω της μεθοδου principant component analysis (PCA).

4.1.1 PCA

Ο πρωταρχικός στόχος της PCA είναι ο μετασχηματισμός ενός συνόλου δεδομένων υψηλής διάστασης σε μια αναπαράσταση χαμηλότερης διάστασης, διατηρώντας παράλληλα όσο το δυνατόν περισσότερες από τις αρχικές πληροφορίες.Αυτό επιτυγχάνετε με τον εντοπισμό και τον υπολογισμό των κύριων συνιστωσών των δεδομένων. Οι κύριες συνιστώσες είναι γραμμικοί συνδυασμοί των αρχικών χαρακτηριστικών και είναι ορθογώνιες μεταξύ τους. Η πρώτη κύρια συνιστώσα συλλαμβάνει τη μέγιστη διακύμανση των δεδομένων και κάθε επόμενη συνιστώσα εξηγεί την υπόλοιπη διακύμανση με φθίνουσα σειρά.
Η εξηγούμενη διακύμανση είναι μια κρίσιμη έννοια στην PCA. Αντιπροσωπεύει το ποσοστό της συνολικής διακύμανσης του συνόλου δεδομένων που εξηγείται από κάθε κύρια συνιστώσα. Η αθροιστική εξηγούμενη διακύμανση παρέχει εικόνα για το πόση από την αρχική πληροφορία διατηρείται καθώς περισσότερες συνιστώσες περιλαμβάνονται στον χώρο μειωμένων διαστάσεων.
Ετσι θα παρουσιαστει ενα διαγραμμα της εξηγουμενης διακυμανσης ανα προστιθεμενη συνιστωσα με σκοπο να μειωσουμε μεν τις διαστασεις αλλα να διατηρουμε το 80% της αρχικης πληροφοριας.



Σχήμα 4.4: Αθροισμα εξηγουμενης διακυμανσης ανα πρωστιθεμενη συνιστωσα.

Βλεπουμε οτι για να διατηρηθει ενα 80% της αρχικης πληροφοριας πρεπει να φτασουμε σχεδον 300 συνιστωσες οποτε η μειωση διαστασεων δεν θα υλοποιηθει και ετσι θα επιλεχθει μια μεθοδος ταξινομησης ευρωστη σε μεγαλες διαστασεις και σε ακραιες τιμες.

4.2 Xgboost

Το δέντρο αποφάσεων είναι ένας θεμελιώδης αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Μοντελοποιεί μια διαδικασία λήψης αποφάσεων μέσω μιας δομής που μοιάζει με δέντρο, όπου κάθε κόμβος αντιπροσωπεύει μια απόφαση βάσει ενός χαρακτηριστικού και κάθε κόμβος φύλλου αντιπροσωπεύει το προβλεπόμενο αποτέλεσμα. Το XGBoost χρησιμοποιεί μια τεχνική ενίσχυσης, που σημαίνει ότι δημιουργεί ένα σύνολο δέντρων απόφασης διαδοχικά. Κάθε δέντρο διορθώνει τα σφάλματα που έγιναν από τα προηγούμενα, εστιάζοντας σε περιπτώσεις που είχαν ταξινομηθεί εσφαλμένα. Η διαδικασία αυτή συνεχίζεται μέχρι να επιτευχθεί ένας συγκεκριμένος αριθμός δέντρων ή μέχρι να μην παρατηρηθεί περαιτέρω βελτίωση. Το XGBoost ενσωματώνει όρους κανονικοποίησης στην αντικειμενική του συνάρτηση, τιμωρώντας τα υπερβολικά πολύπλοκα μοντέλα. Αυτό συμβάλλει στην αποτροπή της υπερβολικής προσαρμογής.. Με υψηλότερες διαστάσεις, τα μοντέλα μπορεί να γίνουν υπερβολικά πολύπλοκα, προσαρμόζοντας το θόρυβο αντί για το υποκείμενο πρότυπο. η χρήση ρηχών δέντρων στο σύνολο, που ενθαρρύνεται από την κανονικοποίηση, είναι ιδιαίτερα επωφελής. Τα ρηχά δέντρα συλλαμβάνουν απλούστερα μοτίβα στα δεδομένα και είναι λιγότερο πιθανό να προσαρμοστούν στο θόρυβο.

ο χαρακτήρας του συνόλου του XGBoost, όπου συνδυάζονται πολλαπλά ρηχά δέντρα, βοηθά στη δημιουργία ενός ισχυρού μοντέλου.

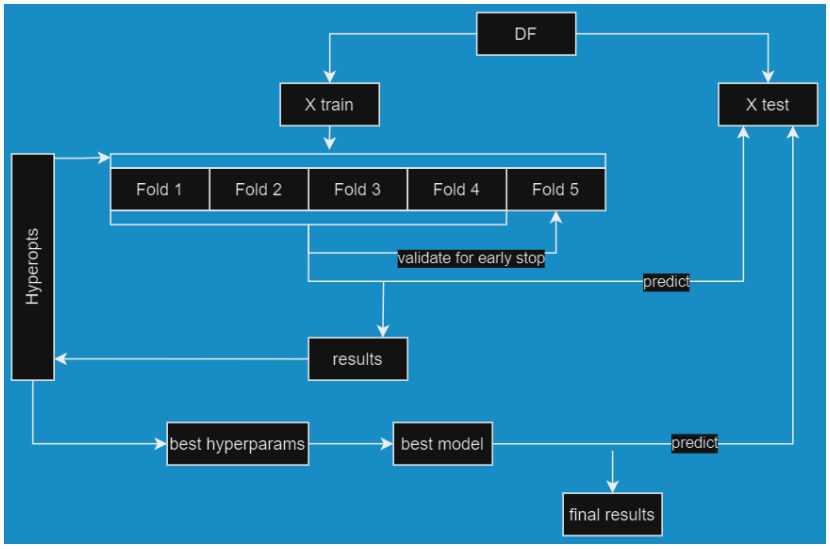
4.2.1 Υπερπαράμετροι Εκπαίδευσης και εκπαίδευση

- **Max Depth**:Ελέγχει την πολυπλοκότητα των μεμονωμένων δέντρων περιορίζοντας το βάθος τους. Ένα βαθύτερο δέντρο μπορεί να συλλάβει περίπλοκα μοτίβα στα δεδομένα, αλλά μπορεί να οδηγήσει σε υπερβολική προσαρμογή, ειδικά όταν το σύνολο δεδομένων είναι θορυβώδες ή έχει ακραίες τιμές. Ο καθορισμός ενός κατάλληλου μαξ δεπτη βοηθά στην εύρεση της σωστής ισορροπίας μεταξύ της σύλληψης των σχετικών μοτίβων και της αποτροπής της προσαρμογής του μοντέλου στο θόρυβο.
- **Min Child Weight**:Ρυθμίζει τον ελάχιστο αριθμό δειγμάτων που απαιτούνται σε κάθε κόμβο-παιδί κατά την κατασκευή του δέντρου. Αυτή η υπερπαράμετρος είναι ζωτικής σημασίας για την αποτροπή του αλγορίθμου από τη δημιουργία τμημάτων με πολύ λίγες περιπτώσεις, οι οποίες ενδέχεται να καταγράφουν θόρυβο ή ακραίες τιμές.
- **Subsample**:Εισάγει την τυχαιότητα καθορίζοντας το κλάσμα του συνόλου δεδομένων εκπαίδευσης που επιλέγεται τυχαία για την ανάπτυξη κάθε δέντρου. Αυτό συμβάλλει στην αποφυγή της υπερπροσαρμογής και βελτιώνει τη γενίκευση με την εκπαίδευση σε διαφορετικά υποσύνολα των δεδομένων.
- **ColByTree**:Ελέγχει το κλάσμα των χαρακτηριστικών (στήλες) που επιλέγονται τυχαία για την κατασκευή κάθε δέντρου. Παρόμοια με το συβσαμπλε, το σολσαμπλε βψτρεε εισάγει τυχαιότητα κατά την εκπαίδευση. Βοηθά στην αποφυγή της υπερβολικής προσαρμογής με την εκπαίδευση σε διαφορετικά σύνολα χαρακτηριστικών για κάθε δέντρο βελτιώνοντας την ικανότητα γενίκευσης του μοντέλου.
- **LR**:Κλιμακώνει τη συνεισφορά κάθε δέντρου στο σύνολο. Ένας χαμηλότερος ρυθμός μάθησης απαιτεί περισσότερα δέντρα για να επιτευχθεί το ίδιο επίπεδο προσαρμογής. Βοηθάει στον έλεγχο της υπερπροσαρμογής κάνοντας τη διαδικασία μάθησης πιο σταδιακή. Ο σωστός συντονισμός του ρυθμού μάθησης είναι απαραίτητος για την εύρεση του σωστού συμβιβασμού μεταξύ πολυπλοκότητας του μοντέλου και χρόνου εκπαίδευσης.
- **Gamma**:Ελέγχει την ελάχιστη μείωση των απωλειών που απαιτείται για να επιτραπεί η διάσπαση ενός κόμβου φύλλου. Υψηλότερες τιμές του γαμμα οδηγούν σε λιγότερες διασπάσεις, αποτρέποντας τον αλγόριθμο από τη δημιουργία κόμβων που δεν συμβάλλουν σημαντικά στη μείωση των απωλειών.

Εκπαίδευση

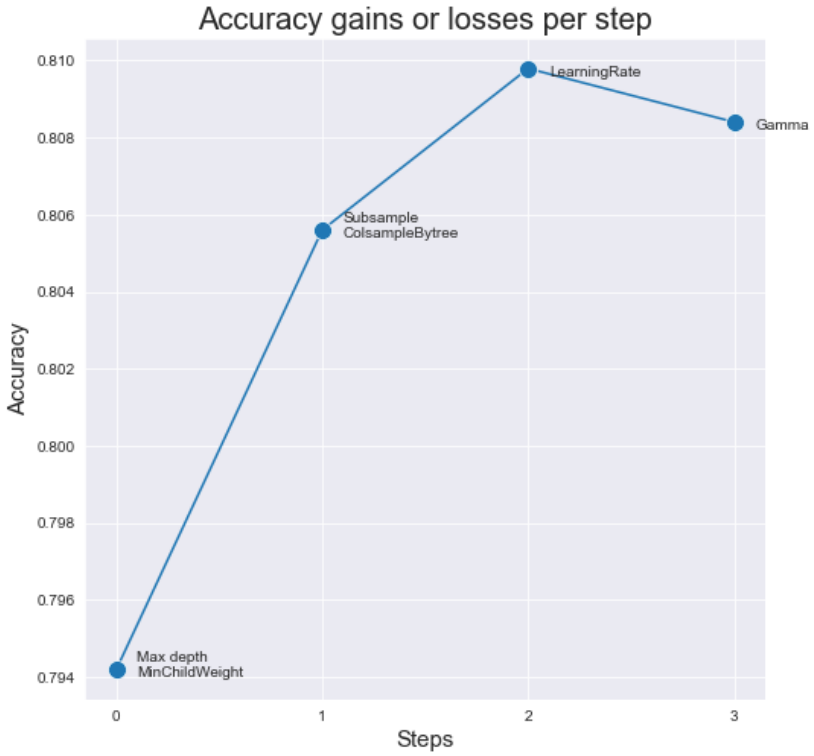
Στη ρύθμιση των υπερπαραμέτρων, ο στόχος είναι η μεγιστοποίηση ή η ελαχιστοποίηση μιας δεδομένης μετρικής, όπως η ακρίβεια ή το σκορ F1.
Το Hyperopt εξερευνά έξυπνα τον χώρο υπερπαραμέτρων, προσαρμόζοντας τη στρατηγική αναζήτησης με βάση τις προηγούμενες αξιολογήσεις, συγκλίνοντας τελικά προς το βέλτιστο σύνολο υπερπαραμέτρων.Το Hyperopt χρησιμοποιεί την Μπεϋζιανή βελτιστοποίηση για την αποτελεσματική αναζήτηση σε έναν χώρο υπερπαραμέτρων, με στόχο την εύρεση του συνόλου υπερπαραμέτρων που αποδίδει την καλύτερη απόδοση του

μοντέλου.
Η μέθοδος cross validation είναι μια ισχυρή τεχνική που χρησιμοποιείται για την αξιολόγηση της απόδοσης και της ικανότητας γενίκευσης ενός μοντέλου μηχανικής μάθησης. Περιλαμβάνει την κατάτμηση του συνόλου δεδομένων σε πολλαπλά υποσύνολα, γνωστά ως folds, και την επαναληπτική εκπαίδευση και αξιολόγηση του μοντέλου σε διαφορετικούς συνδυασμούς αυτών των folds.
Θα χρησιμοποιηθει τμηματικη ευρεση των καλυτερων υπερπαραμετρων του xgboost. Η εκπαιδευση θα ξεκινησει ψαχνωντας τις καλυτερες τιμες των max depth,minimum child weight επειτα subsample,colsample bytree ,επειτα learning rate και τελος gamma.Η ρυθμιση θα γινει με cross validation εντος του πλαισιου hyperopt (nested cross validation).Πρακτικα θα βρεθουν οι υπερπαραμετροι που μεγιστοποιουν την μεση ακριβεια των προβλεψεων σε καθε fold.



Σχήμα 4.5: Διάγραμμα διαδικασίας εκπαίδευσης

4.3 Αποτελέσματα και σχολιασμός



Σχήμα 4.6: Αύξηση μείωση της ακριβειας ανα ρυθμιση σετ παραμετρων

Παρατηρουμε οτι η χρηση της καλυτερη τιμης της παραμετρου γαμμα μειωνει την ακριβεια οποτε θα διατηρηθουν οι καλυτερες τιμες των υπολοιπων παραμετρων για την τελικη επανεκπαιδευση οι οποιες ειναι :

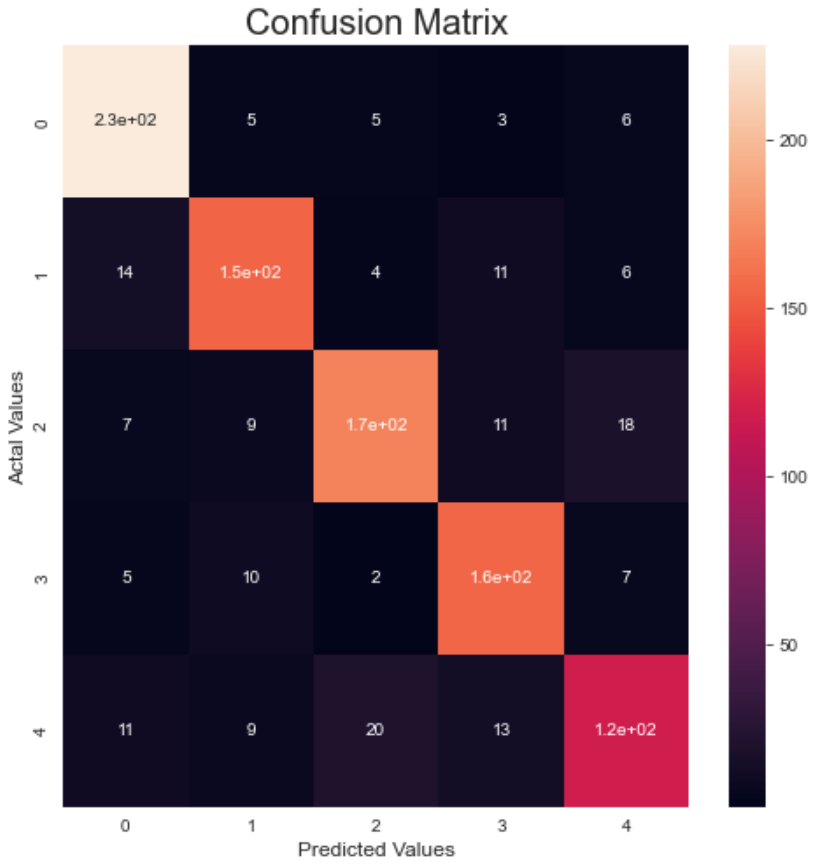
- Max Depth = 6
- Min Child Weight = 7
- Subsample = 0.49423
- ColByTree = 0.38780
- LR = 0.02905

Επειτα γινεται επανεκπαιδευση σε ολο το train set με τις καλυτερες υπερπαραμετρους μια φορα με το αυτουσιο train set και μια φορα μεσω δειγματοληψιας για εξισορροπηση των κλασεων με την μεθοδο Smote-Tomek.
Η SMOTE-Tomek (Synthetic Minority Over-sampling Technique with Tomek links) είναι μια υβριδική τεχνική δειγματοληψίας που χρησιμοποιείται για την αντιμετώπιση μη ισορροπημένων συνόλων δεδομένων στη μηχανική μάθηση.Για κάθε κλάση μειονότητας στο σύνολο δεδομένων πολλαπλών κατηγοριών, το SMOTE παράγει ανεξάρτητα συνθετικά δείγματα με παρεμβολή μεταξύ υφιστάμενων περιπτώσεων. Λαμβάνει υπόψη τους πλησιέστερους γείτονες κάθε περίπτωσης στην κατηγορία μειονότητας και δημιουργεί συνθετικά δείγματα κατά μήκος των τμημάτων γραμμής που τα συνδέουν.Ένας σύνδεσμος Tomek περιλαμβάνει δύο περιπτώσεις, κάθε μία από μια διαφορετική κλάση, που είναι οι πλησιέστεροι γείτονες της άλλης. Σε σενάρια πολλαπλών κλάσεων, η ύπαρξη

ενός συνδέσμου Tomek μπορεί να περιλαμβάνει περιπτώσεις από διαφορετικά ζεύγη κλάσεων.Οι περιπτώσεις που εμπλέκονται σε συνδέσμους Tomek αφαιρούνται από το σύνολο δεδομένων. Αυτό το βήμα γίνεται για να εξαλειφθούν ζεύγη περιπτώσεων που μπορεί να συμβάλουν στη σύγχυση και το θόρυβο στη διαδικασία εκπαίδευσης.

Πίνακας 4.2: Σύγκριση RBF-SVM,KNN,Xgboost,Xgboost-smote-tomek

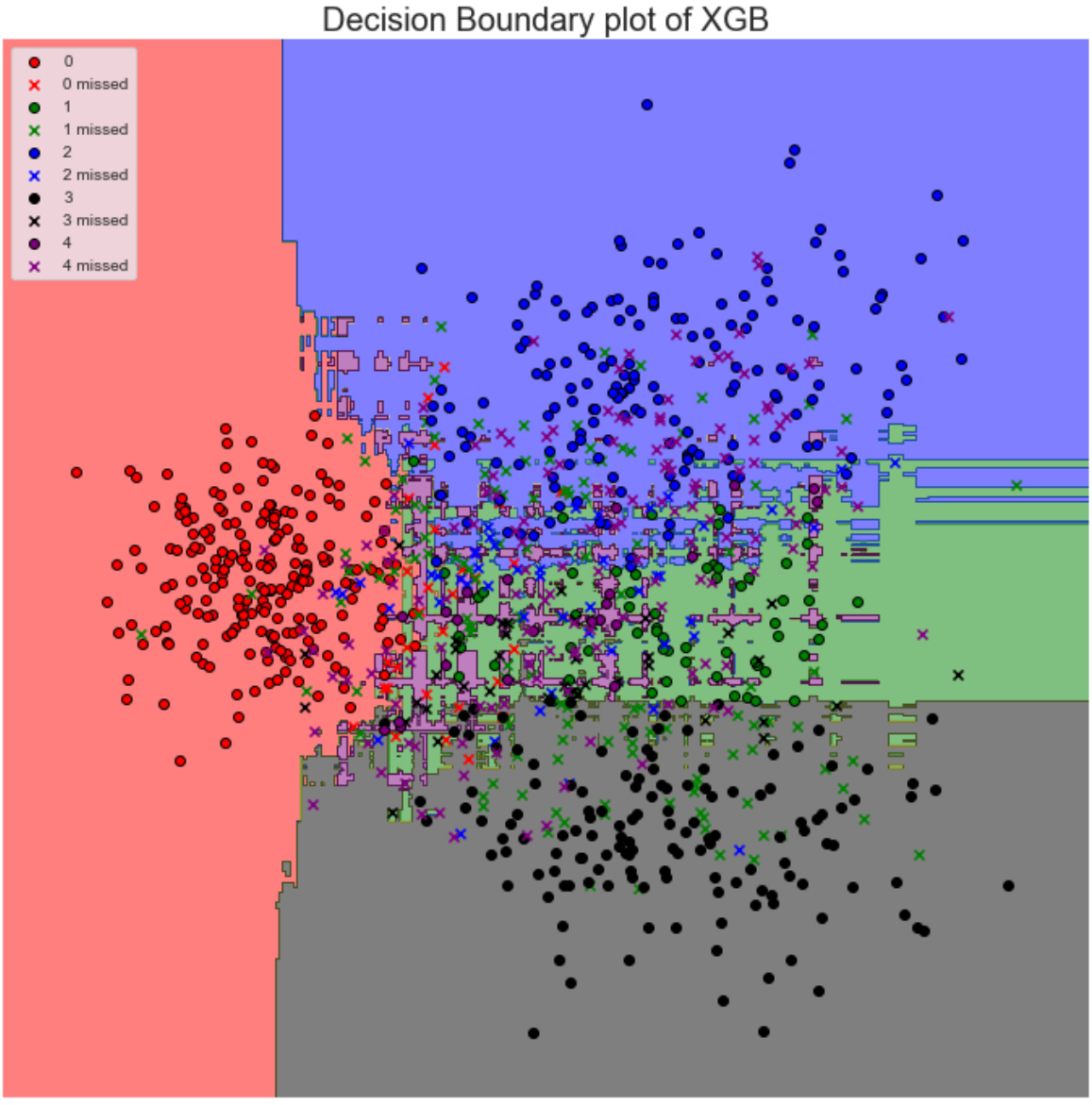
Method / Metric	RBF-SVM	KNN-13N	XGB	XGB-Smote
• Accuracy	75.70%	75.10%	82.10%	82.40%
• Precision	79.23%	78.80%	81.99%	82.29%
• Recall	75.70%	75.10%	82.10%	82.40%
• F1-score	74.23%	72.00%	81.95%	82.24%



Σχήμα 4.7: Πίνακας σύγχησης του Xgb-Smote

Επειτα θα επανεκπαιδευθει ολο το συνολο δεδομενων στις καλυτερες υπερπαραμετρους xgboost-smote και τα αποτελεσματα θα αποθηκευτουν σε καταλληλο αρχειο .npy.

Τέλος θα γινει μια προσπαθεια απεικονισης των οριων αποφασης του μοντελου xgboost μειωνοντας τις διαστασεις σε δυο μεσω της τεχνικης pca.



Σχήμα 4.8: Ορια αποφασης Xgboost