Ταξινομητής Τραγουδιών βάσει του είδους της μουσικής και των μουσικών οργάνων

Φοιτητές: Ηλίας Κασιωτάκης mtn2311, Κωνσταντίνος Σταυράκης

Επιβλέπων καθηγητής: Θεόδωρος Γιαννακόπουλος

ΕΙΣΑΓΩΓΗ

Η ανάγκη για αυτόματη ταξινόμηση των τραγουδιών ανάλογα με το μουσικό είδος και το κυρίαρχο μουσικό όργανο έχει αναδειχθεί ως μία σημαντική πρόκληση στον τομέα της μηχανικής μάθησης. Η ικανότητα να αυτοματοποιήσουμε τη διαδικασία ταξινόμησης μπορεί να οδηγήσει σε βελτιωμένη εμπειρία χρήστη.

Στο παρόν έγγραφο, εξετάζουμε την εφαρμογή τεχνικών μηχανικής μάθησης για την αυτόματη ταξινόμηση τραγουδιών βάσει του μουσικού είδους και του κυρίαρχου μουσικού οργάνου. Παρουσιάζουμε μια συνοπτική επισκόπηση των μεθόδων μηχανικής μάθησης που χρησιμοποιούνται για αυτό το σκοπό και αξιολογούμε την απόδοσή τους στο IMRAS dataset.

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΕΤΡΙΚΕΣ ΑΠΟΔΟΣΗΣ

Το IMRAS Dataset περιέχει αρχεία 16 bit stereo τύπου wav, με συχνότητα δειγματοληψίας 44.1kHz διάρκειας 3 sec. Επιπλέον είναι ήδη διαχωρισμένο σε σύνολο εκπαίδευσης (train set) και δοκιμής (test set) με 6705 και 2874 αρχεία ήχου αντίστοιχα . Το dataset περιέχει μουσικά αρχεία από ποικίλες δεκαετίες, ποιότητας ήχου και από διαφορετικούς καλλιτέχνες. Επιπρόσθετα, τα μουσικά αρχεία περιλαμβάνουν μια έντονη ποικιλία μουσικών οργάνων και τρόπων ηχογράφησης.

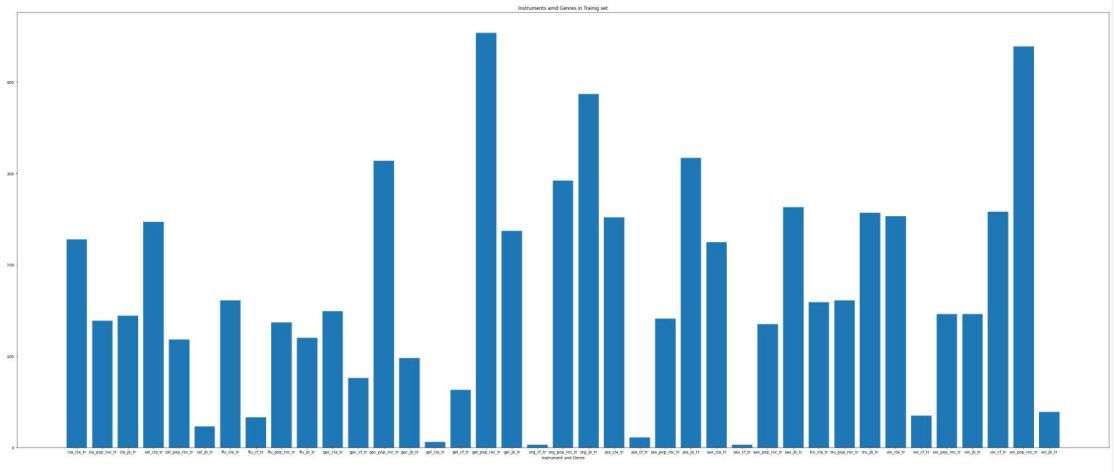
Το σύνολο εκπαίδευσης διαθέτει tags για το είδος της μουσικής και την κατηγορία του κυρίαρχου μουσικού οργάνου. Συγκεκριμένα περιέχονται 11 μουσικά όργανα και 4 είδη μουσικής. Τα μουσικά όργανα είναι τσέλο (cel), κλαρινέτο (cla), φλάουτο (flu), ακουστική (gac) και ηλεκτρική (gel) κιθάρα, εκκλησιαστικό όργανο (org), πιάνο (pia), σαξόφωνο (sax), τρομπέτα (tru), βιολί

(vio) και ανθρώπινη φωνή (voi). Τα είδη μουσικής περιορίζονται σε country-folk (cou_fol), κλασική ([cla]), pop-rock ([pop-roc]) και blue-jazz (jaz_blu). Τα tags για το είδος μουσικής (genre) και μουσικού οργάνου περιέχονταν στο όνομα του αρχείου ήχου στο train set.

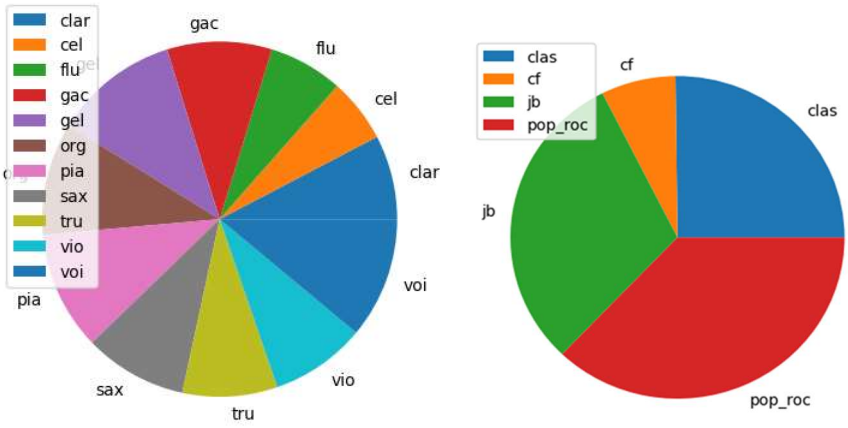
Στο test set τα αρχεία ήχου είναι ονοματισμένα με έναν συνδυασμό του ονόματος του καλλιτέχνη και του τίτλου του τραγουδιού. Το κυρίαρχο μουσικό όργανο είναι αποθηκευμένο σε ένα txt αρχείο με ίδιο όνομα με αυτό του αρχείου ήχου. Το πρόβλημα που παρουσιάζεται είναι ότι το test set δεν περιείχε καμία πληροφορία για το είδος της μουσικής. Αυτό είχε ως αποτέλεσμα την ανάγκη annotation του είδους της μουσικής σε κάθε αρχείο. Για να είναι έγκυρο το annotation χρησιμοποιήθηκε η ακόλουθη μέθοδος:

- Αναζήτηση του τίτλου του τραγουδιού μέσω μηχανής αναζήτησης (browser)
- Εύρεση τραγουδιού και καταγραφή του είδους μουσικής
- Διασταύρωση ότι βρέθηκε το σωστό τραγούδι ακούγοντάς το.

Ο συνδυασμός των μουσικών ειδών και των μουσικών οργάνων οδήγησαν συνολικά σε 39 διαφορετικές κλάσεις. Οι κλάσεις δεν αντιπροσωπεύονται ομοιόμορφα, δηλαδή το dataset είναι αρκετά imbalanced.



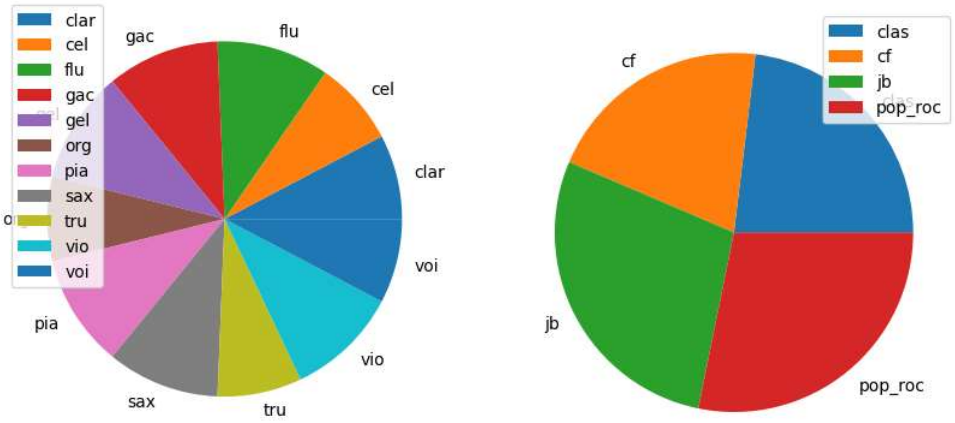
Εικόνα 1: Το πλήθος αρχείων ανά κλάση σε μορφή ραβδογράμματος στο train set



Εικόνα 2: Το ποσοστό των μουσικών οργάνων και των μουσικών ειδών στο train set

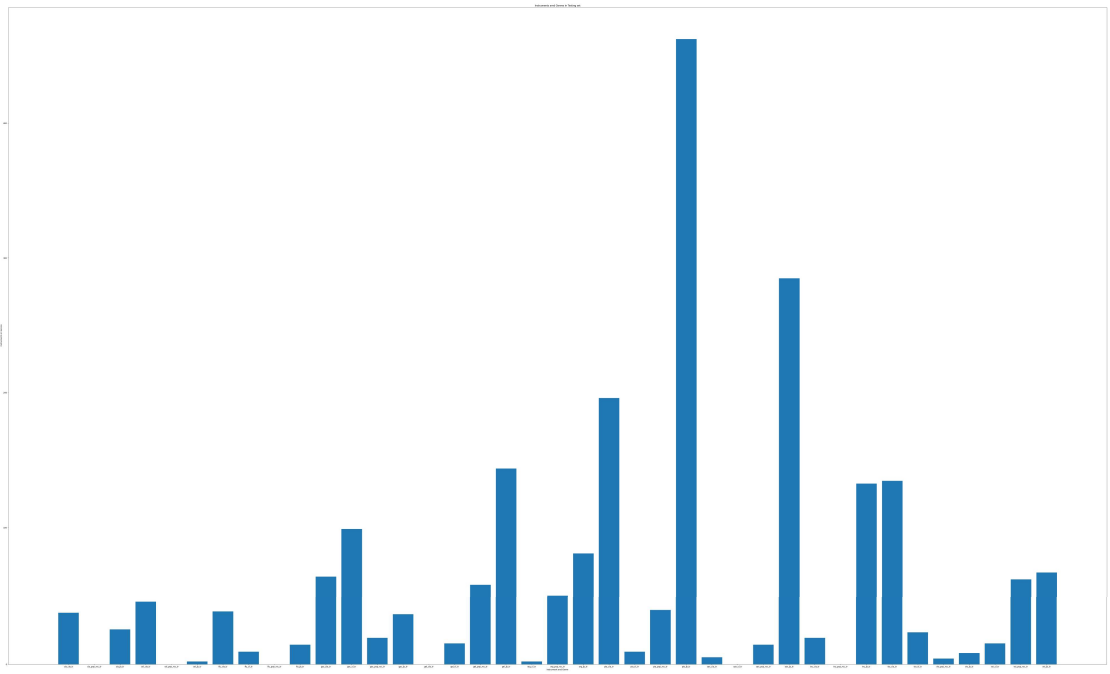
Αυτό οδήγησε στην ανάγκη να δημιουργήσουμε περισσότερα δεδομένα για τις κλάσεις που δεν αντιπροσωπεύονται αρκετά. Για αυτό τον λόγο χρησιμοποιήθηκε αρχικά η μέθοδος data augmentation. Χρησιμοποιήθηκε η προσθήκη λευκού θορύβου, η μεταβολή της έντασης του τραγουδιού μέσω time stretching , η μεταβολή του τόνου του τραγουδιού μέσω pitch shifting και η μετατόπιση στο πεδίο του χρόνου (time shifting). Επειδή οι τεχνικές αυτές δεν οδήγησαν στην παραγωγή του απαραίτητου αριθμού δεδομένων, ο αναγκαίος αριθμός συμπληρώθηκε μέσω της μεθόδου oversampling.

Εικόνα 3: Το πλήθος αρχείων ανά κλάση σε μορφή ραβδογράμματος μετά το balancing του dataset στο train set



Εικόνα 4: Το νέο ποσοστό των μουσικών οργάνων και των μουσικών ειδών στο train set

Το test set είναι αρκετά imbalanced όπως και το train set. Επομένως, η ακρίβεια δεν μπορεί να χρησιμοποιηθεί ως μετρική απόδοσης του ταξινομητή. Για αυτό τον λόγο χρησιμοποιείται η μετρική F1 weighted για τη σωστή αξιολόγηση της απόδοσης σε συνδυασμό με τον confusion matrix.



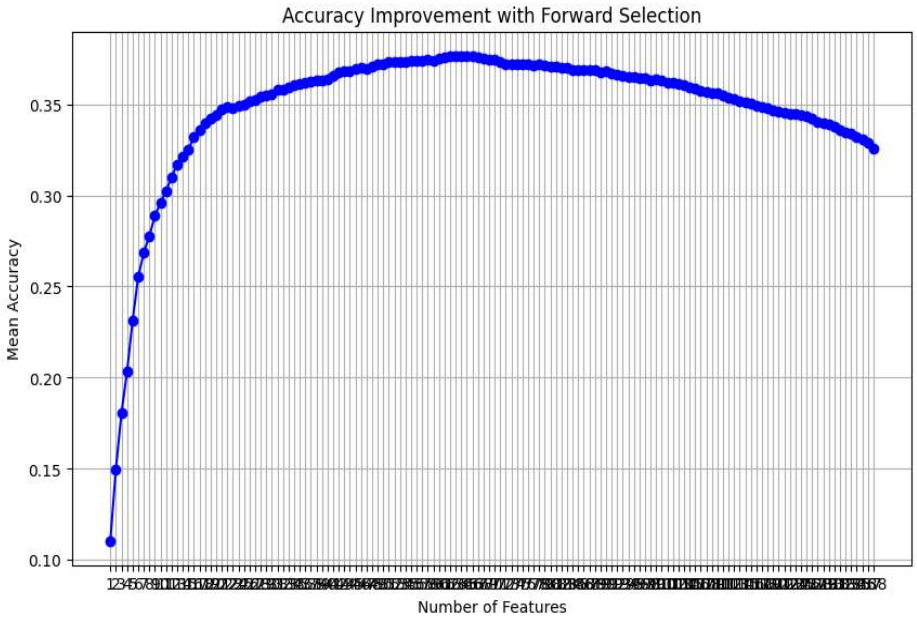
Εικόνα 5: Το πλήθος αρχείων ανά κλάση σε μορφή ραβδογράμματος στο test set

ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE EXTRACTION)

Εφόσον το dataset περιείχε μόνο αρχεία ήχου ήταν ανάγκη η εξαγωγή των απαραίτητων χαρακτηριστικών ώστε να εκπαιδευτούν οι ταξινομητές. Αυτό επιτυγχάνεται μέσω του framing. Κάθε αρχείο ήχου χωρίζεται σε τμήματα (frames) μίας συγκεκριμένης διάρκειας και για κάθε frame εξάγεται ένα σύνολο χαρακτηριστικών . Στη συνέχεια, υπολογίζεται η μέση τιμή και η τυπική απόκλιση κάθε χαρακτηριστικού και χρησιμοποιούνται για να αναπαραστήσουν την πληροφορία για ολόκληρο το αρχείο. Η συγκεκριμένη διαδικασία πραγματοποιήθηκε μέσω της βιβλιοθήκης pyAudioAnalysis. Για κάθε αρχείο έγινε εξαγωγή 138 χαρακτηριστικών και επιλέχθηκε ως μήκος των frames τα 50 msec. Στη συνέχεια τα δεδομένα υπέστησαν scaling μέσω της MinMax μεθόδου.

Αφελής Ταξινομητής Μπέυζ

Στην αρχή χρησιμοποιήθηκε ο Αφελής Μπευζιανός Ταξινομητής για την επίλυση του προβλήματος. Προκειμένου να μειωθεί η πολυπλοκότητα χρησιμοποιήθηκε η μέθοδος forward selection έτσι ώστε ο ταξινομητής να εκπαιδευτεί με ένα μικρότερο αριθμό δεδομένων. Η επιλογή γίνεται εισάγοντας σταδιακά κάθε χαρακτηριστικό στον ταξινομητή. Εάν η προσθήκη του χαρακτηριστικού οδηγήσει σε αύξηση της απόδοσης κατά τη φάση εκπαίδευσης τότε χρησιμοποιείται για την εκπαίδευση του μοντέλου, διαφορετικά απορρίπτεται.

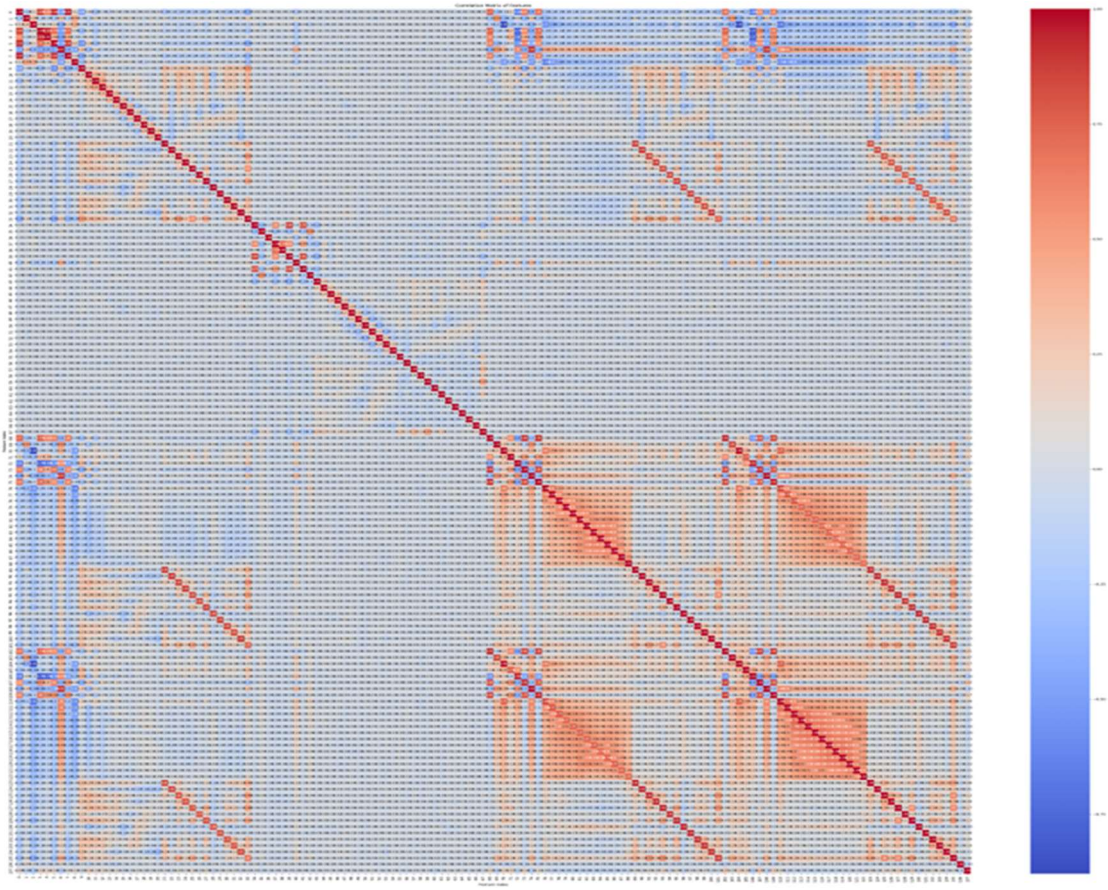


Εικόνα 6: Η απόδοση του ταξινομητή με την προσθήκη κάθε χαρακτηριστικού

Παρόλα αυτά ο Αφελής Μπευζιανός ταξινομητής απέτυχε στο να προβλέψει σωστά τα test δεδομένα. Συγκεκριμένα η ακρίβειά του ήταν 8% και η F1 weighted 5%. Η αποτυχία του συγκεκριμένου αλγορίθμου φαίνεται και από την ακρίβεια που πετυχαίνει ο αλγόριθμος και στα train δεδομένα. Η παράμετρος F1 κατά το training είναι 39%. Επομένως, ο αλγόριθμος δεν μπόρεσε να μάθει από τα δεδομένα. Αυτό μπορεί να οφείλεται στην κατάρα της διαστατικότητας καθώς ο αριθμός των συνολικών features συνέχιζε να είναι αρκετά μεγάλος και στο γεγονός ότι δεν ισχύει η υπόθεση ότι τα features είναι ανεξάρτητα μεταξύ τους.

Random Forest και RBF SVM

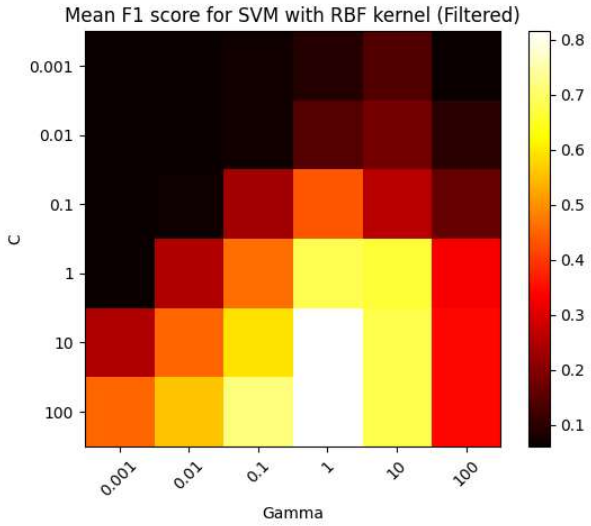
Στη συνέχεια χρησιμοποιήθηκαν οι αλγόριθμοι Random Forest και RBF SVM για την εκπαίδευση του μοντέλου. Λόγω του υψηλού υπολογιστικού κόστους της forward selection και των ίδιων των αλγορίθμων, για την επιλογή των απαραίτητων χαρακτηριστικών χρησιμοποιήθηκε η μέθοδος correlation matrix. Ο correlation matrix είναι ένας τετραγωνικός πίνακας που περιέχει τις συσχετίσεις μεταξύ των χαρακτηριστικών, δηλαδή μετρά το βαθμό με τον οποίο οι μεταβολές σε ένα χαρακτηριστικό συσχετίζονται με τις μεταβολές σε ένα άλλο. Απορρίφθηκαν τα χαρακτηριστικά που παρουσίαζαν συσχέτιση 80% και πάνω.



Εικόνα 7: Ο correlation matrix των χαρακτηριστικών

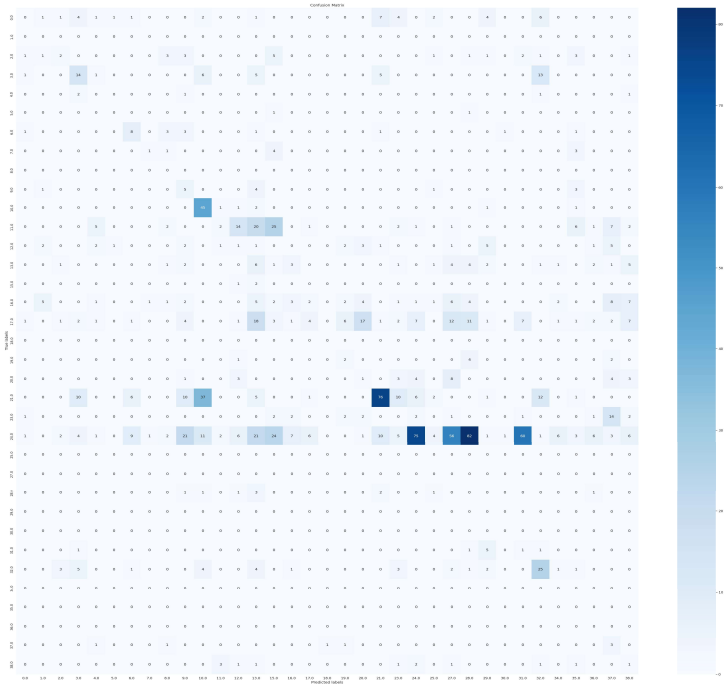
Ο αλγόριθμος SVM (Support Vector Machine) με RBF (Radial Basis Function) πυρήνα (kernel) αναζητεί το βέλτιστο υπερεπίπεδο που διαχωρίζει τα δεδομένα των κλάσεων. Για να γίνει πιο εύκολη η γραμμική διαχωρισιμότητα, ο RBF πυρήνας μετασχηματίζει τα δεδομένα εισόδου σε χώρο υψηλής διάστασης.

Για την σωστή εκπαίδευση του μοντέλου είναι απαραίτητη η σωστή επιλογή των υπερπαραμέτρων C και γάμμα. Η υπερπαράμετρος C σχετίζεται με την ικανότητα γενικότητας του SVM μοντέλου, ενώ η παράμετρος γάμμα σχετίζεται με τον RBF kernel και κρίνει την επιρροή των data points στη δημιουργία του συνόρου απόφασης. Λόγω της απουσίας validation set, χρησιμοποιήθηκε η μέθοδος 5 fold cross validation. Αρχικά επιλέχθηκε να είναι 13 αντί για 5 ώστε το μέγεθος του μπλοκ που χρησιμοποιείται κάθε φορά στον ρόλο του validation set να αναπαριστά το 10% των συνολικών δεδομένων. Παρόλα αυτά, το υπολογιστικό κόστος ήταν αρκετά μεγάλο και χρειάστηκαν ώρες εκπαίδευσης. Επομένως, επιλέχθηκε ο αριθμός 5 για εξοικονόμηση χρόνου. Έτσι, επιλέχθηκε ο συνδυασμός των κατάλληλων υπερπαραμέτρων όπου C=10 και γ=1.



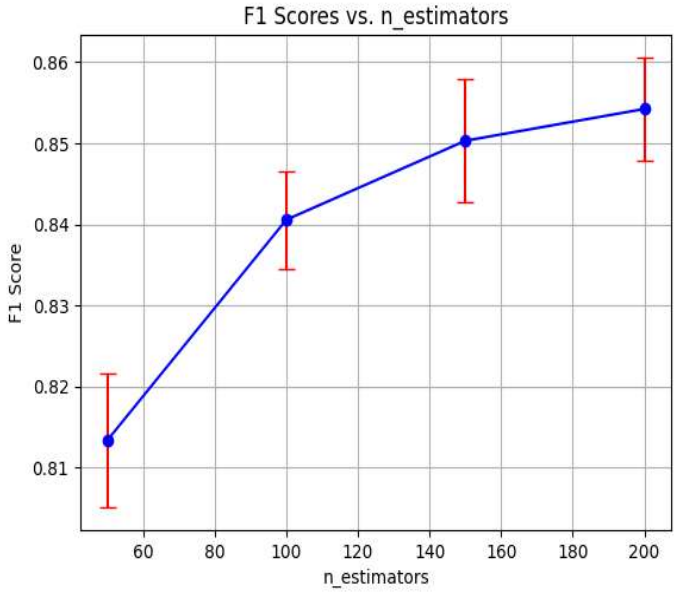
Εικόνα 8: Ο τρόπος με τον οποίο ο συνδυασμός των παραμέτρων C και γ επηρεάζει την τιμή της F1 weighted.

Παρόλα αυτά η ακρίβεια είναι πολύ χαμηλή (21%) το ίδιο και η F1 weighted (24%). Η απόδοση είναι χειρότερη από αυτή ενός random classifier. Επομένως ο RBF SVM αποτυγχάνει στην αναγνώριση των κλάσεων.



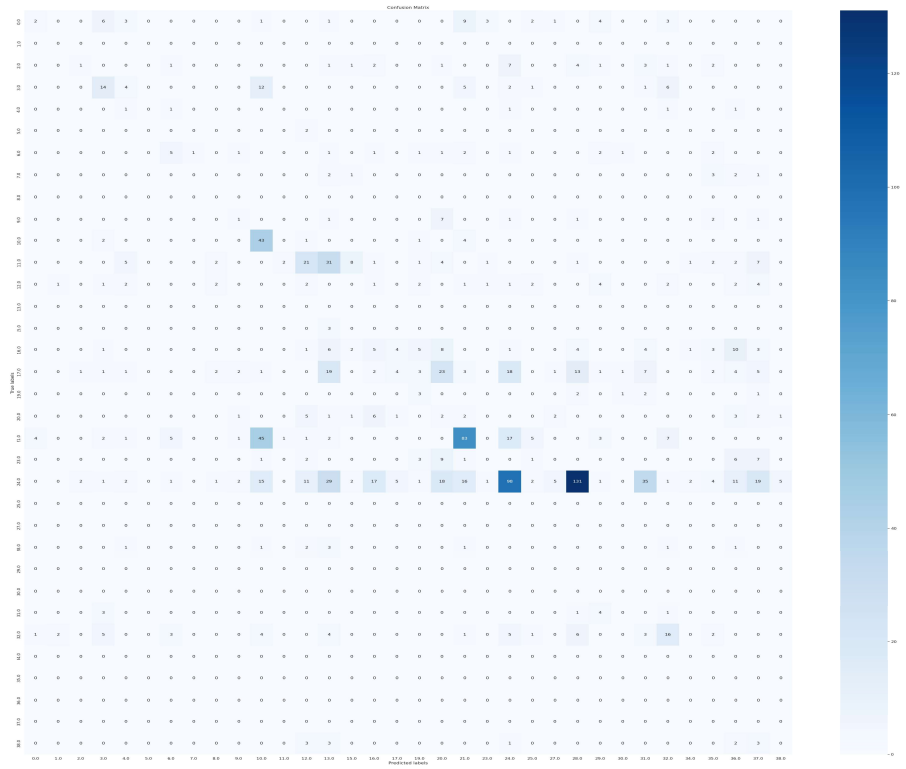
Εικόνα 9: Ο confusion matrix του RBF SVM

Ο Random Forest συνδυάζει τις προβλέψεις από πολλούς απλούς εκτιμητές (estimators) για να βελτιώσει την ακρίβεια των προβλέψεων του. Η βασική ιδέα πίσω από το Random Forest είναι η δημιουργία ενός "δάσους" (forest) από δέντρα αποφάσεων (decision trees). Κάθε δέντρο λαμβάνει τυχαία υποσύνολα των δεδομένων εκπαίδευσης και βάσει των χαρακτηριστικών τους επιλέγουν τις αντίστοιχες κλάσεις. Στη συνέχεια, συγκεντρώνονται οι προβλέψεις από όλα τα δέντρα και επιλέγεται η πιο συχνή κλάση. Ο αριθμός των estimators επιλέγεται ίσος με 200 μέσω 13 fold cross validation.



Εικόνα 10: Πώς ο αριθμός των estimators επηρεάζει την τιμή της F1 weighted.

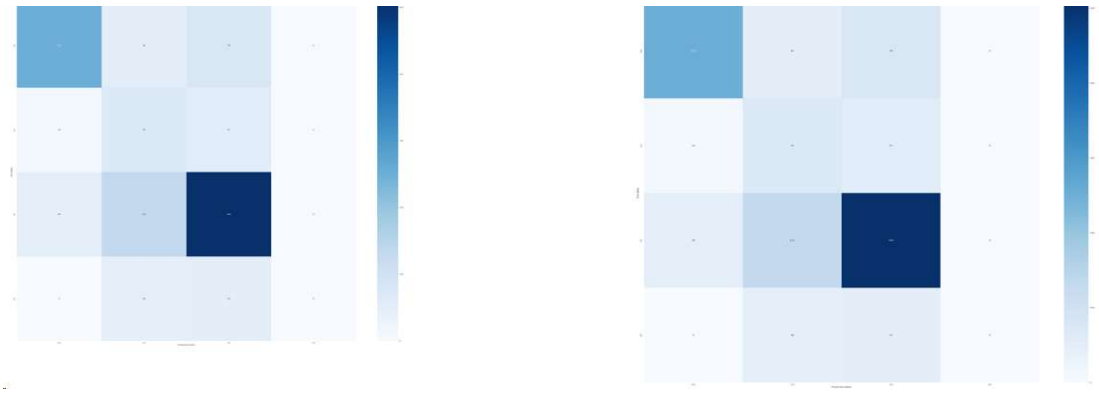
Παρόλα αυτά και ο Random Forest παρουσιάζει εξίσου κακή απόδοση με τον RBF SVM με μόλις 22% ακρίβεια και 26% F1 weighted.



Εικόνα 11: Ο confusion matrix του Random Forest

ΑΠΟΔΟΣΗ ΤΩΝ ΤΑΞΙΝΟΜΗΤΩΝ ΩΣ ΠΡΟΣ ΤΟ ΕΙΔΟΣ ΤΗΣ ΜΟΥΣΙΚΗΣ ΚΑΙ ΤΩΝ ΜΟΥΣΙΚΩΝ ΟΡΓΑΝΩΝ

Η πολύ κακή απόδοση των RBF SVM και Random Forest ταξινομητών οδήγησαν στην υπόθεση ότι το annotation των ειδών μουσικής στο test set ήταν λανθασμένο. Παρόλα αυτά η υπόθεση απορρίφθηκε καθώς οι ταξινομητές μπορούν να ταξινομήσουν το είδος της μουσικής με πολύ καλή ακρίβεια. Συγκεκριμένα, ο RBF SVM ταξινομητής σημείωσε ακρίβεια 66% και F1 weighted 68%, συγκριτικά καλύτερος από τον Random Forest με ακρίβεια και F1 weighted 63%. Παρατηρείται από τον confusion matrix έντονο misclassification στην κλάση 1 (pop rock). Αυτό οφείλεται σε λανθασμένο annotation καθώς και στο γεγονός ότι τραγούδια που κρίθηκαν ως ποπ ροκ περιείχαν στοιχεία και από άλλα μουσικά είδη, όπως η country.



Εικόνα 12: Αριστερά: Ο confusion matrix του Random Forest. Δεξιά: Ο confusion matrix του SVM RBF. Οι κλάσεις αποτελούν μόνο τα είδη μουσικής

Αντίθετα, οι κακές προβλέψεις εντοπίζονται ως προς την κατηγοριοποίηση των μουσικών οργάνων με τους ταξινομητές να παρουσιάζουν και οι δύο ακρίβεια 30%. Ο SVM παρουσιάζει F1 weighted 34% ενώ ο Random Forest 33%.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Ο αφελής ταξινομητής Μπέυζ αδυνατεί να εκπαιδευτεί από τα δεδομένα λόγω εξάρτησης μεταξύ των features και του curse of dimensionality. Πιο κατάλληλοι για την ταξινόμηση είναι οι αλγόριθμοι Random Forest και RBF SVM. Ο RBF SVM κατηγοριοποιεί καλύτερα το είδος της μουσικής από τον Random Forest ταξινομητή. Ωστόσο, παρουσιάζεται έντονο misclassification της pop rock το οποίο θα μπορούσε να αντιμετωπιστεί αν είχαμε στη διάθεσή μας περισσότερα είδη της ροκ μουσικής. Λόγω της λανθασμένης ταξινόμησης των κυρίαρχων μουσικών οργάνων πρέπει να ελεγχθεί η επιρροή των augmented και oversampled δεδομένων στην ακρίβεια των ταξινομητών καθώς θα μπορούσαν να οδήγησαν σε overfitting και εισαγωγή θορύβου.