ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Πρόγραμμα Σπουδών Τμήματος

Πληροφορικής και Τηλεπικοινωνιών Λαμίας

Εφαρμογή αλγορίθμων εξόρυξης δεδομένων σε δεδομένα γονιδιώματος υδρόβιων οργανισμών

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Κωνσταντινίδης Νικόλαος (ΑΜ: 2115090)

Επιβλέπων καθηγητής: Δρ. Σταμούλης Γεώργιος

Συνεπιβλέπων καθηγητής: Δρ. Κόκκινος Κωνσταντίνος

ΛΑΜΙΑ 2020

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις (1), που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1.Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί χωρίς να τα περικλείω σε εισαγωγικά και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.

2.Δέχομαι ότι η αυτολεξεί παράθεση χωρίς εισαγωγικά, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.

3.Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κ.λπ.), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια.

4.Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

Υπογραφή

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Τόπος: .....

Ημερομηνία: .....

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. .....
2. .....
3. .....

ΕΥΧΑΡΙΣΤΙΕΣ

Η πτυχιακή εργασία με τίτλο «Αναγνώριση ειδών θαλάσσιων οργανισμών με χρήση τεχνικών εξόρυξης δεδομένων σε δεδομένα γενετικών ακολουθιών ψαριών» εκτελέστηκε ως μέρος των απαιτήσεων για την απόκτηση του πτυχίου, του τμήματος «Πληροφορικής και Τηλεπικοινωνιών» της σχολής Θετικών Επιστημών του Πανεπιστημίου Θεσσαλίας, το έτος 2020.

Θεωρώ πρέπον σε αυτό το σημείο, να ευχαριστήσω όσους με βοήθησαν, άμεσα ή έμμεσα στην ολοκλήρωση αυτής της εργασίας. Χωρίς σειρά προτεραιότητας θα ήθελα να ευχαριστήσω, τους γονείς μου, που με άφησαν να επιλέξω μια σχολή μακριά από την πόλη καταγωγής μου και με στήριξαν, ώστε να ολοκληρώσω τις σπουδές μου εκεί. Έπειτα, θα ήθελα να ευχαριστήσω, τον καθηγητή και επιβλέποντα για αυτή την εργασία, κύριο Κόκκινο Κωνσταντίνο, τόσο για την εμπιστοσύνη, προς το πρόσωπό μου, όσο και για την αμεσότητα, που μου παρείχε στην επικοινωνία μας, όταν άρχισα να έχω απορίες και προβληματισμούς, για την εργασία. Τέλος, θα ήθελα να ευχαριστήσω και τους φίλους μου στη Θεσσαλονίκη, οι οποίοι πίστευαν και πιστεύουν στις δυνατότητές μου.

Κωνσταντινίδης Νικόλαος

Ημερομηνία

ΣΤΟΧΟΣ

Στόχος της παρούσας εργασίας είναι η διερεύνηση της αποτελεσματικότητας κλασσικών τεχνικών δημιουργίας μοντέλων κατηγοριοποίησης, στη διάκριση των ειδών της οικογένειας των λαγοκέφαλων. Αναλυτικότερα, αυτή η μελέτη στοχεύει στην ανάδειξη της καλύτερης τεχνικής κατηγοριοποίησης, η οποία και θα προτιμηθεί σε άλλες μελέτες, αλλά και στην ανάδειξη των αδύναμων μεθόδων, ώστε να αποφευχθούν σε μελλοντικές μελέτες.

Κρίνεται σημαντικό να σημειωθεί, ότι πρόκειται για έρευνα, που συντάχθηκε με περιορισμένο πλήθος δεδομένων και πως έγινε σε περιορισμένο μέρος των διάφορων παραμετροποιήσεων, που μπορούν να γίνουν στον κώδικα.

ΠΕΡΙΛΗΨΗ

Στην συγκεκριμένη μελέτη αξιοποιήθηκαν γενετικές ακολουθίες, του γένους του λαγοκέφαλου, που βρέθηκε στη σελίδα «www.boldsystems.org». Έπειτα έγινε καθαρισμός των δεδομένων για να εξασφαλιστεί, ότι κάθε ακολουθία «dna», υπάρχει μόνο μια φορά μέσα στο «dataset». Επίσης, έγινε απομάκρυνση όσων ακολουθιών, δεν συνοδεύονταν από όνομα είδους.

Οι αλγόριθμοι που χρησιμοποιήθηκαν και αξιολογήθηκαν αποτελούν μέλη της ομάδας των αλγορίθμων μηχανικής μάθησης(Machine Learning). Αυτοί είναι: Δέντρο Απόφασης(Decision Tree), Μηχανή Διανύσματος Υποστήριξης(Support Vector Machine), Τυχαίου Δάσους(Random Forest) και «Κ» Πλησιέστερων Γειτόνων(K Nearest Neighbors).

Έγιναν δοκιμές για να βρεθεί το βέλτιστο πλήθος γειτόνων, που θα χρησιμοποιεί το αλγόριθμος «KNN», ενώ έγιναν δοκιμές και για το βέλτιστο πλήθος δέντρων, που θα χρησιμοποιεί ο αλγόριθμος «Random Forest». Έπειτα, αξιολογήθηκαν τα αποτελέσματα όλων των αλγορίθμων για τα διάφορα μεγέθη χαρακτηριστικού, από τέσσερα έως και οκτώ, ενώ έγινε και μια εξαγωγή αποτελεσμάτων για ένα ακραίο μέγεθος χαρακτηριστικού, το είκοσι.

Τέλος, γίνεται παρουσίαση και σύγκριση μεταξύ των αποτελεσμάτων των αλγορίθμων, οι οποίοι έχουν εκτελεστεί με βάση τα χαρακτηριστικά, στα οποία έδειξαν τα καλύτερα αποτελέσματα στις προηγούμενες δοκιμές.

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

Είναι κοινώς γνωστό, πως η διώρυγα του Σουέζ, αποτελεί ένα έργο υψίστης σημασίας για τις εμπορικές δραστηριότητες, της ανθρωπότητας. Έχει επιτρέψει την θαλάσσια επικοινωνία των κρατών της Ευρώπης, αλλά και των κρατών της βόρειας Αμερικής, με κράτη της ανατολικής Αφρικής και της Ασίας, δίνοντας τη δυνατότητα αποφυγής της πλεύσης νοτίου της αφρικανικής ηπείρου. Εκτός, όμως από τα κράτη, η διώρυγα, έχει φέρει σε επαφή και εντελώς ξένα, μέχρι τη στιγμή ολοκλήρωσης του έργου, θαλάσσια οικοσυστήματα. Συγκεκριμένα, το οικοσύστημα της Μεσογείου, ήταν εντελώς διαφορετικό, από αυτό του Ινδικού ή του Ειρηνικού ωκεανού, καθώς αποτελούνταν από διαφορετικούς οργανισμούς.

Ένας οργανισμός, που υπήρχε στον Ινδικό και τον Ειρηνικό ωκεανό, έγινε γνωστός στην Ελλάδα, ως λαγοκέφαλος, με επιστημονική ονομασία «lagocephalus sceleratus». Το ψάρι αυτό ανήκει στην οικογένεια των τετραοδοντίδων και στο γένος του λαγοκέφαλου. Εμφανίστηκε στη Μεσόγειο λίγο μετά το 2003 και κατ' επέκταση ήρθε στο κοντά στην Ελλάδα, στο Αιγαίο. Η εμφάνισή του, οφείλεται στο ταξίδι, που κάνουν κάποια ψάρια μέσω της διώρυγας και το οποίο ονομάζεται λεσσεψιανή μετανάστευση, παίρνοντας το όνομά της από τον αρχιτέκτονα της διώρυγας του Σουέζ, Φερντινάρντ ντε Λεσσέψ.

Στην Ερυθρά Θάλασσα, απ' όπου και προέρχεται, ο λαγοκέφαλος ζει σε βραχώδεις πυθμένες από ρηχά παράκτια νερά μέχρι βάθους μέτρων. Ο λαγοκέφαλος μοιάζει πολύ με το ψάρι φούσκα αλλά είναι πιο επιμήκης και με συμμετρική ουρά. Η πλάτη του είναι γκρι ή καφέ με πιο σκούρα σημεία και έχει λευκή κοιλιά. Μια χαρακτηριστική ασημένια ζώνη βρίσκεται κατά μήκος των πλευρών του ψαριού. Μπορεί να έχει μήκος έως και 40 εκατοστά και θηρεύει βενθικά ασπόνδυλα, δηλαδή ασπόνδηλα ζώα, που ζουν στους βυθούς θαλασσών και ωκεανών.

Ο λόγος, που ο λαγοκέφαλος είναι άξιος μελέτης, είναι η επικινδυνότητά του, τόσο για τα άλλα ψάρια, όσο και για τον άνθρωπο. Πρόκειται για ψάρι, που παράγει την δηλητηριώδη τοξίνη, τετραδοτοξίνη, η οποία μπορεί να προκαλέσει μυϊκή παράλυση, ανεπάρκεια κυκλοφορικού συστήματος, αναπνευστικές διαταραχές και θάνατο. Αυτή ήταν η αιτία, να αποφασίσω να ασχοληθώ με το ψάρι. Αναλυτικότερα, επέλεξα να ασχοληθώ με την αναγνώριση του είδους κάθε ψαριού του γένους, του λαγοκέφαλου, μέσω της γενετικής του ακολουθίας.

Ο στόχος αυτός, μπορεί να επιτευχθεί, μέσω της εξόρυξης γνώσης από δεδομένα ή της εξόρυξης δεδομένων, όπως είναι γνωστή. Ως εξόρυξη δεδομένων νοείται η διαδικασία εξαγωγής χρήσιμων πληροφοριών από μια τεράστια ποσότητα δεδομένων. Χρησιμοποιείται για την ανακάλυψη νέων και χρήσιμων, με ακρίβεια, μοτίβων στα δεδομένα, αναζητώντας νόημα και σχετικές πληροφορίες για αυτόν, που τις χρειάζεται. Είναι εργαλείο, που χρησιμοποιείται από τους ανθρώπους.

Χρησιμοποίησα 4 αλγορίθμους εξόρυξης δεδομένων και τους σύγκρινα, ως προς την απόδοσή τους για την αναγνώριση του είδους του ψαριού, βάση της γενετικής του ακολουθίας. Οι τεχνικές, που επέλεξα, είναι : Δέντρο Απόφασης(Decision Tree),

Μηχανή Διανύσματος Υποστήριξης(Support Vector Machine), Τυχαίου Δάσους(Random Forest) και «Κ» Πλησιέστερων Γειτόνων(K Nearest Neighbors).

Η ανάπτυξη του κώδικα, για την εκπαίδευση των κατηγοριοποιητών και τον υπολογισμό των μετρικών αξιολόγησής τους, έγιναν με χρήση τής γλώσσας προγραμματισμού «python». Στα πλαίσια αυτής της εργασίας, επιλέχθηκε η μελέτη της επιρροής συγκεκριμένων παραμέτρων των συναρτήσεων, που αξιοποιούν τους αλγορίθμους. Οι υπόλοιπες παράμετροι παρέμειναν στις προτεινόμενες(default) τιμές τους, τις οποίες έχουν επιλέξει οι δημιουργοί του πακέτου «scikit-learn».

Ξεκινώντας, θα δούμε γενικές πληροφορίες για το ψάρι, λαγοκέφαλο και τον τρόπο εμφάνισής του στη Μεσόγειο. Έπειτα θα αναφερθούμε στην οικογένεια , όπου ανήκουν τα είδη τού λαγοκέφαλου, αλλά και στην επικινδυνότητά τους. Θα ακολουθήσει μία ανάλυση της έννοιας της εξόρυξης δεδομένων. Αφού γίνει αναφορά στην κατηγοριοποίηση, θα παρουσιαστούν τα διάφορα είδη κατηγοριοποιητών, που υπάρχουν, ώστε στη συνέχεια να αναλυθούν λίγο περισσότερο τα είδη, που χρησιμοποιήθηκαν στη μελέτη. Στη συνέχεια, υπάρχει παρουσίαση των δεδομένων, που χρησιμοποιήθηκαν.

Έτσι, εισαγόμαστε στο κύριο μέρος της έρευνας, όπου αποκαλύπτονται τα σημεία αξιολόγησης των τεχνικών και των αλγορίθμων. Έπονται τα αποτελέσματα, όπου για το κάθε πείραμα, εμφανίζονται τα διαγράμματα, ο πίνακας μέσων όρων των τιμών των μετρικών αξιολόγησης και όποιες παρατηρήσεις μπορούν να γίνουν βάσει αυτών. Φυσικά, στο τέλος υπάρχει σύνοψη της συνολικής εικόνας τής έρευνας και κάποια συμπεράσματα.

2 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΡΕΥΝΑ

2.1 Λαγοκέφαλος

Ο αριθμός των ειδών στη Μεσόγειο Θάλασσα που φθάνουν μέσω της διώρυγας του Σουέζ (ονομάζεται επίσης είδος «Lessepsian») συνεχίζει να αυξάνεται (Nader et al., 2012, Golani, 2010). Πρόσφατες μελέτες εκτιμούν ότι περισσότερο από το 5% των θαλάσσιων ειδών είναι μη ιθαγενή και 13,5% είναι εισβολείς, συμπεριλαμβανομένων ψαριών, ασπόνδυλων και μακροφυτών (Galil, 2009, Zenetos, 2010, Fricke et al., 2015, Zenetos et al., 2015, Golani, 2010). Αυτά τα «χωροκατακτητικά είδη» (Shine et al., 2000) εγκαθίστανται στο νέο βιότοπο, αυξάνονται σε αριθμό και εξαπλώνονται στην περιοχή, δυνητικά απειλώντας τη φυσική βιολογική ποικιλότητα (Galil et al., 2015, Coll et al., 2010) και οικονομία (Galil, 2008). Επομένως, απαιτείται ιδιαίτερη προσπάθεια από εποπτεύοντες οργανισμούς, προκειμένου να παρακολουθούν και να προβλέπουν την εξάπλωσή τους.

Μεταξύ αυτών των ειδών, ο φρύνος με το ασημένιο μάγουλο «Lagocephalus sceleratus» (Gmelin, 1789) είναι ιδιαίτερα ανησυχητικό. Τα πρώτα αξιόπιστα αρχεία στη Μεσόγειο Θάλασσα χρονολογούνται από το 2003, αλλά ο αριθμός των παρατηρήσεων αυξήθηκε ραγδαία, ώστε να θεωρείται ένα από τα ταχύτερα επεκτεινόμενα επεμβατικά είδη στη λεκάνη (Akyol et al., 2005, Peristeraki et al., 2006). Οφείλει την επιτυχία του στον υψηλό ρυθμό ανάπτυξης και αναπαραγωγής, στην έλλειψη φυσικών αρπακτικών, στην ικανότητα εκμετάλλευσης των τροφίμων, και στην ικανότητα να ανέχεται ένα ευρύ φάσμα περιβαλλοντικών συνθηκών (Yaglioglu et al., 2011). Έχει δέρμα χωρίς φολίδες, με σκούρα σημεία στην κορυφή και πλευρικές ασημένιες γραμμές. Αυτό το είδος είναι συνηθισμένο στην Ερυθρά Θάλασσα, ανήκει στην οικογένεια «Tetraodontidae», είναι εξαιρετικά δηλητηριώδες και μπορεί να είναι θανατηφόρο για τον άνθρωπο εάν καταναλωθεί, λόγω του υψηλού επιπέδου νευροτοξίνης «Tetrodotoxin» (TTX), που υπάρχει σε πολλά όργανα τού ψαριού και εκκρίνεται από το δέρμα του, ως απωθητικό μετά το πρήξιμο (Yaglioglu et al., 2011, Nader et al., 2012). Προτιμά συνήθως τα ρηχά νερά και τη μέση-υψηλή θερμοκρασία του νερού, η οποία συσχετίζεται με την ταχύτερη πρόσληψη «TTX». Έτσι, η κλιματική αλλαγή θα μπορούσε να είναι ευνοϊκή για αυτό το είδος, ιδιαίτερα στη Μεσόγειο Θάλασσα (Nader et al., 2012).

Επιστημονικές μελέτες έχουν εκτιμήσει τον πιθανό αντίκτυπο του «Lagocephalus Sceleratus» στην οικονομική και ανθρώπινη υγεία στην Ανατολική Μεσόγειο Θάλασσα (Ünal et al., 2015, Ünal et al., 2017). Σε αυτήν την περιοχή, αυτό είναι τώρα ένα από τα πιο σημαντικά είδη (σε βιομάζα) στα λιβάδια «Posidonia oceanica», που αποτελεί μείζον πρόβλημα για τη βιοτεχνική αλιεία, δεδομένου ότι βλάπτει τα αλιευτικά εργαλεία (π.χ. δίχτυα και γραμμές) και προηγείται σε μεγάλο βαθμό σε τοπικά αποθέματα καλαμαριών και χταποδιών (Kalogirou et al., 2010). Ωστόσο, αυτές οι μελέτες δεν αναφέρουν οριστικές οικολογικές και οικονομικές μελλοντικές εκτιμήσεις επιπτώσεων και συνήθως περιλαμβάνουν πιο ποιοτικές από τις ποσοτικές προβλέψεις. Συνολικά, δείχνουν ότι τα ψάρια αντιπροσωπεύουν επί του παρόντος το 4% του βάρους των συνολικών αλιευτικών προϊόντων (Nader et al., 2012) και έχει ήδη επηρεάσει αρνητικά την οικονομία ορισμένων χωρών της Μεσογείου

(Ünal et al., 2017). Επίσης, από το 2003 έχουν καταγραφεί αρκετά επεισόδια θανάτου και σοβαρής ασθένειας μετά την κατανάλωση ψαριών, καθώς οι ψαράδες και άλλοι άνθρωποι συνήθως δεν μπορούν να αναγνωρίσουν αυτό το σχετικά νέο είδος (Bentur et al., 2008, Kheifets et al., 2012).

Αυτό το σενάριο απαιτεί δράσεις προτεραιότητας για την πρόληψη, τον εντοπισμό και πιθανώς την εξάλειψη του «Lagocephalus Sceleratus» (Zenetos et al., 2016), ειδικά λαμβάνοντας υπόψη ότι η χωρητικότητα του καναλιού του Σουέζ διευρύνεται (Searight, 2016) και η κλιματική αλλαγή διευκολύνει την εισβολή (Galil et. κ.λπ., 2014, ICES, 2007, FAO, 2007). Μια προσέγγιση θα μπορούσε να είναι η χρήση επιλεκτικής αλιείας, ιδίως σε μεγάλα άτομα και σε τοπικές προληπτικές ενέργειες σε εκείνες τις περιοχές όπου το pufferfish πιθανόν να μετακινηθεί και να εγκατασταθεί τα επόμενα χρόνια (alnal et al., 2017). Επομένως, ένας χάρτης του συνεχιζόμενου μοτίβου εισβολής θα μπορούσε να καθοδηγήσει την ανάπτυξη προληπτικών και διορθωτικών ενεργειών (Zenetos et al., 2015, Zenetos et al., 2016) και θα μπορούσε επίσης να βοηθήσει στην κάλυψη ενός χάσματος μεταξύ έρευνας και διαχείρισης σχετικά με αυτό το ψάρι (Ünal et. κ.λπ., 2015).

Την τελευταία δεκαετία, υπήρξε ένα αυξανόμενο ενδιαφέρον για την εφαρμογή οικολογικών εξειδικευμένων μοντέλων (ENMs) για την πρόβλεψη της κατανομής των χωροκατακτητικών ειδών (Guisan et al., 2014). Έχουν χρησιμοποιηθεί διαφορετικές προσεγγίσεις με βάση την αξιολόγηση των εξειδικευμένων διαφορών μεταξύ της γηγενής περιοχής ενός είδους και της εισβαλλόμενης περιοχής (Peterson, 2003, Barbosa et al., 2012, Leidenberger et al., 2015). Σε ορισμένες περιπτώσεις, αυτές οι προσεγγίσεις λαμβάνουν επίσης υπόψη τον τρόπο με τον οποίο η κλιματική αλλαγή διευκολύνει την εξάπλωση του είδους στην εισβαλλόμενη περιοχή (Sax et al., 2007, Thuiller et al., 2005). Οι προσεγγίσεις που βασίζονται σε ENM για τη διείσδυση ειδών χρησιμοποιούν ποικίλα μοντέλα, συμπεριλαμβανομένων των φακέλων (Sutherst, 2000, Jeschke and Strayer, 2008), στατιστικά (Ficetola et al., 2007, Bidegain et al., 2015) και μηχανή μοντέλα μάθησης (Peterson and Robins, 2003). Τα περισσότερα από αυτά τα μοντέλα εκτιμούν τη σχέση μεταξύ της παρουσίας ενός είδους και ορισμένων περιβαλλοντικών παραμέτρων και παράγουν μια κατανομή πιθανότητας. Στη συνέχεια, προβάλλεται σε μια συγκεκριμένη περιοχή (με την πάροδο του χρόνου) για να αποκτήσει μια δυναμική απεικόνιση της εισβολής (Mellin et al., 2016, Carlos-Júnior et al., 2015). Το πιο χρησιμοποιούμενο ENM σε αυτό το πλαίσιο είναι ο «Γενετικός Αλγόριθμος για Παραγωγή Κανόνων», GARP (Stockwell, 1999), που χρησιμοποιεί μια προσέγγιση μηχανικής μάθησης (Peterson and Vieglais, 2001, Ganeshaiah et al., 2003, Sanchez-Flores et. al., 2008, Underwood et al., 2004). Ένα άλλο ευρέως χρησιμοποιούμενο μοντέλο είναι το μοντέλο της παρουσίας Maximum Entropy (Ficetola et al., 2007, West et al., 2016), ενώ τα μοντέλα παρουσίας-απουσίας, π.χ. Τα τεχνητά νευρικά δίκτυα (Kulhanek et al., 2011) και οι μηχανές διανυσμάτων υποστήριξης (Pouteau et al., 2011, Sadeghi et al., 2012), είναι λιγότερο συχνές λόγω της έλλειψης αξιόπιστων δεδομένων απουσίας. Συνήθως, οι εναλλακτικές προσεγγίσεις που βασίζονται σε ENM έχουν συμπληρωματικά χαρακτηριστικά που συλλαμβάνουν διαφορετικά χαρακτηριστικά εισβολής ενός είδους (Elith and Graham, 2009). Έτσι, είναι σύνηθες να συγκρίνουμε ή να συγχωνεύουμε την έξοδο

διαφορετικών μοντέλων για να παράγουμε μια τελική εκτίμηση spread (Castelar et al., 2015, Farashi and Najafabadi, 2015, Padalia et al., 2014, Sobek-Swant et al., 2012).

2.2 Εξόρυξη Δεδομένων

Η χειροκίνητη εξαγωγή μοτίβων και μοντέλων από δεδομένα, εμφανίστηκε πριν από αιώνες. Οι πρώτες μέθοδοι αναγνώρισης μοτίβων στα δεδομένα περιλαμβάνουν το θεώρημα του «Bayes», περί τα 1700 και την ανάλυση παλινδρόμησης, περί τα 1800. Ο πολλαπλασιασμός, η πανταχού παρούσα και αυξανόμενη ισχύς των υπολογιστών έχουν αυξήσει δραματικά τη δυνατότητα συλλογής, αποθήκευσης και χειρισμού δεδομένων. Καθώς τα σύνολα δεδομένων(data sets) έχουν αυξηθεί σε μέγεθος και πολυπλοκότητα, η άμεση "πρακτική" ανάλυση δεδομένων επαυξάνεται όλο και περισσότερο με έμμεση, αυτοματοποιημένη επεξεργασία δεδομένων, με τη βοήθεια άλλων ανακαλύψεων στην επιστήμη των υπολογιστών, ειδικά στον τομέα της μηχανικής μάθησης, όπως τα νευρωνικά δίκτυα, ανάλυση συστάδων, γενετικοί αλγόριθμοι, δέντρα αποφάσεων, κανόνες απόφασης και μηχανές διανυσμάτων υποστήριξης. Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων με σκοπό την αποκάλυψη κρυφών μοτίβων σε μεγάλα σύνολα δεδομένων. Γεφυρώνει το χάσμα από τα εφαρμοσμένα στατιστικά στοιχεία και την τεχνητή νοημοσύνη (που συνήθως παρέχουν το μαθηματικό υπόβαθρο) στη διαχείριση βάσεων δεδομένων, αξιοποιώντας τον τρόπο αποθήκευσης και ευρετηρίασης των δεδομένων σε βάσεις δεδομένων για την αποτελεσματικότερη εκτέλεση των αλγορίθμων μάθησης και ανακάλυψης, επιτρέποντας την εφαρμογή τέτοιων μεθόδων σε ολοένα και μεγαλύτερα σύνολα δεδομένων.

Η εξόρυξη δεδομένων είναι μια διαδικασία ανακάλυψης μοτίβων και μοντέλων σε μεγάλα σύνολα δεδομένων που περιλαμβάνουν μεθόδους στη διασταύρωση της μηχανικής μάθησης, στατιστικών και συστημάτων βάσεων δεδομένων. Είναι ένα διεπιστημονικό πεδίο της επιστήμης των υπολογιστών και των στατιστικών με συνολικό στόχο την εξαγωγή πληροφοριών από ένα σύνολο δεδομένων και τη μετατροπή των πληροφοριών σε κατανοητή δομή για περαιτέρω χρήση. Η εξόρυξη δεδομένων είναι το βήμα ανάλυσης της διαδικασίας "ανακάλυψη γνώσεων σε βάσεις δεδομένων"(knowledge discovery in databases) ή «KDD». Εκτός από το αρχικό βήμα ανάλυσης, περιλαμβάνει επίσης πτυχές διαχείρισης δεδομένων και βάσεων δεδομένων, προεπεξεργασία δεδομένων, θέματα μοντέλου και συμπερασμάτων, μετρήσεις ενδιαφέροντος, θέματα πολυπλοκότητας, μετα-επεξεργασία ανακαλυφθέντων δομών, οπτικοποίηση και ενημέρωση στο διαδίκτυο. Θα μπορούσε κανείς να πει, ότι η φράση «εξόρυξη δεδομένων» είναι εσφαλμένη, καθώς ο στόχος είναι η εξαγωγή μοτίβων και γνώσεων από μεγάλες ποσότητες δεδομένων και όχι η εξαγωγή (εξόρυξη) των ίδιων των δεδομένων. Είναι επίσης μια λέξη, που χρησιμοποιείται καταχρηστικά και εφαρμόζεται συχνά σε οποιαδήποτε μορφή μεγάλης κλίμακας επεξεργασίας δεδομένων ή πληροφοριών (συλλογή, εξαγωγή, αποθήκευση, ανάλυση και στατιστικά στοιχεία) καθώς και σε οποιαδήποτε εφαρμογή συστήματος υποστήριξης αποφάσεων υπολογιστών, συμπεριλαμβανομένης της τεχνητής νοημοσύνης και επιχειρηματική ευφυΐα.

Η πραγματική εργασία εξόρυξης δεδομένων είναι η ημιαυτόματη ή αυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένων για εξαγωγή προηγουμένως άγνωστων, ενδιαφερόντων μοτίβων, όπως ομάδες εγγραφών δεδομένων (ανάλυση συμπλέγματος), ασυνήθιστες εγγραφές (ανίχνευση ανωμαλιών) και εξαρτήσεις (εξόρυξη κανόνων συσχέτισης, διαδοχική εξόρυξη προτύπων). Αυτό συνήθως περιλαμβάνει τη χρήση τεχνικών βάσεων δεδομένων, όπως χωρικών δεικτών. Τα μοτίβα μπορούν στη συνέχεια να θεωρηθούν ως ένα είδος σύνοψης των δεδομένων εισαγωγής και είναι δυνατό να χρησιμοποιηθούν σε περαιτέρω ανάλυση ή για παράδειγμα, στη μηχανική μάθηση και σε προγνωστικές αναλύσεις. Για παράδειγμα, το βήμα εξόρυξης δεδομένων μπορεί να προσδιορίσει πολλές ομάδες στα δεδομένα, οι οποίες μπορούν στη συνέχεια να χρησιμοποιηθούν για τη λήψη ακριβέστερων αποτελεσμάτων πρόβλεψης από ένα σύστημα υποστήριξης αποφάσεων. Η συλλογή δεδομένων, η προετοιμασία δεδομένων, η ερμηνεία τους και η αναφορά αποτελεσμάτων δεν αποτελούν μέρος της εξόρυξης δεδομένων, αλλά ανήκουν στη συνολική διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων ως πρόσθετα βήματα.

Η διαφορά μεταξύ ανάλυσης δεδομένων και εξόρυξης δεδομένων είναι, ότι η ανάλυση δεδομένων χρησιμοποιείται για τη δοκιμή μοντέλων και υποθέσεων στο σύνολο δεδομένων, για παράδειγμα ανάλυση της αποτελεσματικότητας μιας καμπάνιας μάρκετινγκ, ανεξάρτητα από την ποσότητα των δεδομένων. Αντίθετα, η εξόρυξη δεδομένων χρησιμοποιεί μηχανική εκμάθηση και στατιστικά μοντέλα για να αποκαλύψει κρυφά μοτίβα σε μεγάλο όγκο δεδομένων.

Στο «data mining» υπάρχουν πολλές ομάδες αλγορίθμων, όπως είναι η κατηγοριοποίηση(classification), η ανίχνευση ανωμαλιών(anomaly detection), η συσταδοποίηση(clustering), οι κανόνες συσχέτισης(association rule learning), η παλινδρόμηση(regression) και η συνάθροιση(summarization). Από αυτές τις ομάδες, μας απασχολεί μόνο η κατηγοριοποίηση και τα διάφορα είδη κατηγοριοποιητών, που ανήκουν σε αυτή. Έχουμε κατηγοριοποιητές κανόνων(rule based classifiers), «Bayesian» δίκτυα(Bayesian networks), δέντρα απόφασης(decision trees), πλησιέστερους γείτονες(nearest neighbors), τεχνητά νευρωνικά δίκτυα(artificial neural networks), μηχανές διανυσμάτων υποστήριξης(support vector machines), τραχιά σύνολα(rough sets), ασαφής λογική(fuzzy logic), γενετικοί αλγόριθμοι(genetic algorithms) και σύνολα μεθόδων(ensemble methods).

Οι κατηγοριοποιητές κανόνων λειτουργούν με ανακάλυψη υψηλού επιπέδου και εύκολων στην ερμηνεία κανόνων ταξινόμησης, με μορφή «αν-τότε»(if-then). Οι κανόνες αποτελούνται από δύο μέρη, το προγενέστερο μέρος και το επακόλουθο μέρος κανόνα. Στο προγενέστερο τμήμα ανήκει το «αν», το οποίο καθορίζει ένα σύνολο συνθηκών τιμών χαρακτηριστικών πρόβλεψης. Στο επακόλουθο τμήμα ανήκει το «τότε», το οποίο καθορίζει την κλάση που προβλέπεται από το προηγηθέν μέρος του κανόνα. Αυτοί οι κανόνες μπορούν να δημιουργηθούν από διάφορους αλγορίθμους κατηγοριοποίησης. Οι πιο διαδεδομένοι, είναι οι αλγόριθμοι επαγωγικών δέντρων απόφασης.

Ένα «Bayesian» δίκτυο αποτελείται από έναν κατευθυνόμενο, ακυκλικό γράφο και κατανομή πιθανότητας για κάθε κόμβο σε αυτό το γράφημα δεδομένων των άμεσων προκατόχων του. Ένας ταξινομητής δικτύου «Bayes» βασίζεται σε ένα

«Bayesian» δίκτυο, που αντιπροσωπεύει μια συνδυαστική κατανομή πιθανότητας σε ένα σύνολο κατηγορικών γνωρισμάτων. Αποτελείται από δύο μέρη, τον κατευθυνόμενο ακυκλικό γράφο, που με τη σειρά του αποτελείται από κόμβους και τόξα, και τους πίνακες πιθανότητας υπό όρους. Οι κόμβοι αντιπροσωπεύουν χαρακτηριστικά, ενώ τα τόξα υποδηλώνουν άμεσες εξαρτήσεις. Η πυκνότητα των τόξων σε ένα δίκτυο «Bayes» είναι ένα δείγμα της πολυπλοκότητάς του. Αραιά δίκτυα μπορούν να αντιπροσωπεύουν απλά πιθανολογικά μοντέλα, όπως τα αφελή μοντέλα «Bayes» και τα κρυφά μοντέλα «Markov», ενώ τα πυκνά «Bayesian networks» μπορούν να συλλάβουν πολύ περίπλοκα μοντέλα. Έτσι, αυτά τα δίκτυα παρέχουν μια ευέλικτη μέθοδο για πιθανολογική μοντελοποίηση.

Ένας ταξινομητής δέντρου απόφασης συγκροτείται από ένα δέντρο απόφασης, που δημιουργείται με βάση τις περιπτώσεις. Το δέντρο αποφάσεων έχει δύο τύπους κόμβων: α) τη ρίζα και τους εσωτερικούς κόμβους, β) τα φύλλα. Η ρίζα και οι εσωτερικοί κόμβοι σχετίζονται με χαρακτηριστικά, οι κόμβοι των φύλλων σχετίζονται με κλάσεις. Συγκεκριμένα, κάθε κόμβος, που δεν είναι φύλλο, έχει έναν εξερχόμενο κλάδο για κάθε πιθανή τιμή του χαρακτηριστικού που σχετίζεται με τον κόμβο. Για να προσδιοριστεί την κλάση για μια νέα εγγραφή, χρησιμοποιώντας ένα δέντρο αποφάσεων, ξεκινάμε από τη ρίζα και επισκεπτόμαστε τους διαδοχικούς εσωτερικούς κόμβους, έως ότου φτάσουμε σε κόμβο-φύλλο. Στη ρίζα και σε κάθε εσωτερικό κόμβο, γίνεται μία αξιολόγηση, το αποτέλεσμα τής οποίας, καθορίζει ποιος θα είναι ο επόμενος κόμβος, που θα επισκεφθούμε. Η κλάση τής εγγραφής είναι η κλάση του τελευταίου κόμβου, του φύλλου.

Ο κατηγοροποιητής πλησιέστερου γείτονα υποθέτει πως όλες οι εγγραφές, που χρησιμοποιήθηκαν για να δημιουργηθεί, αντιστοιχούν σε σημεία σε χώρο «ν» διαστάσεων. Κατά της ώρα της εκπαίδευσης, το μοντέλο ταξινόμησης μαθαίνει τις εγγραφές, που συναντάει και τις θυμάται ακόμη και μετά το τέλος της. Όταν ταξινομείται μία νέα εγγραφή, αναπαρίσταται και αυτή με σημείο, τότε βρίσκονται τα «κ» κοντινότερα σε αυτό σημεία και χρησιμοποιούνται με συντελεστή βαρύτητας για το καθένα, ώστε να βγει συμπέρασμα για το νέο σημείο. Για να επιτευχθεί μεγαλύτερη ακρίβεια, συνηθίζεται να δίνονται μεγαλύτερα βάρη στα κοντινότερα σημεία.

Ένα τεχνητό νευρωνικό δίκτυο, που συχνά ονομάζεται σκέτο νευρωνικό δίκτυο είναι ένα μαθηματικό μοντέλο ή ένα υπολογιστικό μοντέλο βασισμένο σε βιολογικό νευρικό δίκτυο. Με άλλα λόγια, είναι μια προσομοίωση βιολογικού νευρικού συστήματος. Στις περισσότερες περιπτώσεις ένα τεχνητό νευρωνικό δίκτυο είναι ένα προσαρμοστικό σύστημα, που αλλάζει τη δομή του ανάλογα με εξωτερικές ή εσωτερικές πληροφορίες, που ρέουν μέσα στο δίκτυο κατά τη φάση εκπαίδευσής του. Ένας ταξινομητής νευρωνικού δικτύου βασίζεται σε νευρωνικά δίκτυα, που αποτελούνται από διασυνδεδεμένους νευρώνες. Πιο απλά, ένας νευρώνας λαμβάνει θετικά και αρνητικά ερεθίσματα(αριθμητικές τιμές) από άλλους νευρώνες και όταν το σταθμισμένο άθροισμα των ερεθισμάτων είναι μεγαλύτερο από ένα ορισμένο κατώφλι, αυτό ενεργοποιείται. Η τιμή εξόδου του νευρώνα είναι συνήθως ένας μη γραμμικός μετασχηματισμός του αθροίσματος των τιμών εισόδου. Σε πιο προηγμένα μοντέλα, ο μη γραμμικός μετασχηματισμός προσαρμόζεται από κάποια συνεχή συνάρτηση.

Η μηχανή διανύσματος υποστήριξης, ουσιαστικά είναι δυαδικός αλγόριθμος ταξινόμησης. Πρόκειται για σύστημα κατηγοριοποίησης, που προέρχεται από τη θεωρία της στατιστικής μάθησης. Έχει χρησιμοποιηθεί με επιτυχία σε εφαρμογές, όπως η κατηγοριοποίηση κειμένου, η αναγνώριση χειρόγραφων χαρακτήρων, ταξινόμηση εικόνων και στην ανάλυση βιολογικών ακολουθιών. Το «SVM» διαχωρίζει τις τάξεις με μια επιφάνεια απόφασης που μεγιστοποιεί το περιθώριο μεταξύ των τάξεων. Η επιφάνεια ονομάζεται βέλτιστο υπερπλάνο(optimal hyperplane) και τα σημεία δεδομένων που βρίσκονται πλησιέστερα στο υπερπλάνο καλούνται διανύσματα υποστήριξης(support vectors). Τα διανύσματα υποστήριξης είναι τα κρίσιμα στοιχεία του συνόλου δεδομένων εκπαίδευσης. Ο μηχανισμός, που καθορίζει τη διαδικασία χαρτογράφησης, ονομάζεται συνάρτηση πυρήνα(kernel function). Το «support vector machine» μπορεί να προσαρμοστεί ώστε να γίνει μη γραμμικός ταξινομητής μέσω της χρήσης μη γραμμικών πυρήνων. Επίσης, μπορεί να λειτουργήσει ως ένας κατηγοριοποιητής πολλαπλών κλάσεων, συνδυάζοντας διάφορους δυαδικούς κατηγοριοποιητές. Η έξοδος της ταξινόμησης είναι οι τιμές απόφασης κάθε εικονοστοιχείο για κάθε κατηγορία, οι οποίες χρησιμοποιούνται για εκτιμήσεις πιθανότητας. Οι τιμές πιθανότητας αντιπροσωπεύουν την "αληθινή" πιθανότητα, με την έννοια ότι κάθε πιθανότητα πέφτει στο εύρος μηδέν έως ένα και το άθροισμα αυτών των τιμών για κάθε εικονοστοιχείο ισούται με ένα. Στη συνέχεια, η ταξινόμηση πραγματοποιείται επιλέγοντας την υψηλότερη πιθανότητα. Η μηχανή διανύσματος υποστήριξης περιλαμβάνει μια παράμετρο ποινής, που επιτρέπει έναν ορισμένο βαθμό εσφαλμένης ταξινόμησης, η οποία είναι ιδιαίτερα σημαντική για μη διαχωρίσιμα σύνολα δεδομένων εκπαίδευσης. Η παράμετρος ποινής ελέγχει την αντιστάθμιση μεταξύ της αποδοχής σφαλμάτων εκπαίδευσης και της επιβολής αυστηρών περιθωρίων. Δημιουργεί ένα χαλαρό περιθώριο που επιτρέπει ορισμένες εσφαλμένες ταξινομήσεις, όπως επιτρέπει ορισμένα σημεία προπόνησης στη λάθος πλευρά του υπερπλάνου. Η αύξηση της τιμής της παραμέτρου ποινής αυξάνει το κόστος της εσφαλμένης ταξινόμησης εγγραφών και αναγκάζει τη δημιουργία ενός πιο ακριβούς μοντέλου που μπορεί να μην γενικευτεί καλά.

Κάθε σύνολο όλων των μη διακριτών αντικειμένων ονομάζεται στοιχειώδες σύνολο. Οποιαδήποτε ένωση ορισμένων στοιχειωδών συνόλων αναφέρεται ως τραγανό ή ακριβές σύνολο, αλλιώς το σύνολο είναι τραχύ(ανακριβές, ασαφές). Κάθε τραχύ σύνολο έχει περιπτώσεις οριακής γραμμής, δηλαδή αντικείμενα που δεν μπορούν να ταξινομηθούν με βεβαιότητα, χρησιμοποιώντας τις διαθέσιμες γνώσεις, ως μέλη του συνόλου ή του συμπληρώματός του. Προφανώς τα τραχιά σύνολα, σε αντίθεση με τα ακριβή σύνολα, δεν μπορούν να χαρακτηριστούν όσον αφορά τις πληροφορίες σχετικά με τα στοιχεία τους. Η χαμηλότερη προσέγγιση αποτελείται από όλα τα αντικείμενα που ανήκουν σίγουρα στο σύνολο και η ανώτερη προσέγγιση περιέχει όλα τα αντικείμενα που ενδεχομένως ανήκουν στο σύνολο. Η διαφορά μεταξύ της ανώτερης και της κατώτερης προσέγγισης αποτελεί την οριακή περιοχή του τραχιού συνόλου. Η προσέγγιση της ανάλυσης δεδομένων με τραχιά σύνολα έχει πολλά πλεονεκτήματα, όπως ότι προσφέρει αποδοτικούς αλγορίθμους για την εύρεση κρυφών μοτίβων στα δεδομένα, αναγνωρίζει συσχετίσεις, που θα εντοπίζονταν με χρήση στατιστικών μεθόδων, επιτρέπει ταυτόχρονα ποσοτικά και ποιοτικά δεδομένα, βρίσκει ελάχιστα σύνολα δεδομένων(μείωση δεδομένων), αξιολογεί τη σημαντικότητα των δεδομένων και είναι απλά στην κατανόηση.

Η ασαφής λογική είναι μία πολύτιμη λογική διαφορετική από την «τραγανή λογική», όπου δυαδικά σύνολα έχουν λογική δύο τιμών. Οι μεταβλητές ασαφούς λογικής έχουν τιμή αλήθειας στο εύρος μεταξύ μηδέν και ένα. Πρόκειται για ένα υπερσύνολο της κλασσικής λογικής «Boolean», που επεκτάθηκε για να διαχειριστεί την έννοια της μερικής αλήθειας. Η συνάρτηση ιδιότητας μέλους (membership function) είναι μια καμπύλη που καθορίζει τον τρόπο αντιστοίχισης κάθε σημείου στο χώρο εισόδου σε τιμή ιδιότητας μέλους (ή βαθμό συμμετοχής) μεταξύ μηδέν και ένα. Η ασαφής λογική χωρίζεται σε δύο τύπους, την ασαφή λογική τύπου ένα και την τύπου δύο. Ο τύπος ένα περιέχει τις σταθερές τιμές. Ο τύπος δύο είναι επέκταση του τύπου ένα στην οποία τα ασαφή σύνολα προέρχονται από τον υπάρχοντα τύπο ένα ασάφειας. Ένα ασαφές σύνολο τύπου δύο περιέχει τους βαθμούς συμμετοχής, που είναι οι ίδιοι ασαφείς. Ο βαθμός συμμετοχής τύπου δύο μπορεί να είναι οποιοδήποτε υποσύνολο της κύριας ιδιότητας μέλους. Για κάθε βασική ιδιότητα μέλους υπάρχει μια δευτερεύουσα ιδιότητα μέλους, που καθορίζει τις δυνατότητες για την κύρια ιδιότητα μέλους. Η ασαφής λογική τύπου ένα είναι αδύνατο να καλύψει τις αβεβαιότητες των κανόνων. Αντίθετα η τύπου δύο μπορεί να τις χειριστεί αποτελεσματικά και αποδοτικά. Τα ασαφή σύνολα τύπου δύο χαρακτηρίζονται από κανόνες αν-τότε. Επίσης, ονομάζονται «fuzzy fuzzy», όπου ο ασαφής βαθμός συμμετοχής είναι ο ίδιος ασαφής, που προκύπτει από τον τύπο ένα.

Οι γενετικοί αλγόριθμοι είναι αλγόριθμοι αναζήτησης βασισμένοι στη φυσική γενετική, που παρέχει ισχυρές δυνατότητες αναζήτησης σε πολύπλοκους χώρους, προσφέροντας μία έγκυρη προσέγγιση σε προβλήματα, που απαιτούν αποδοτικές και αποτελεσματικές διαδικασίες αναζήτησης. Ένας γενετικός αλγόριθμος είναι μια επαναληπτική διαδικασία, που λειτουργεί σε έναν πληθυσμό, για παράδειγμα σε πλήθος υποψήφιων λύσεων. Κάθε λύση λαμβάνεται μέσω ενός μηχανισμού κωδικοποίησης / αποκωδικοποίησης, ο οποίος μας επιτρέπει να αντιπροσωπεύουμε τη λύση ως χρωμόσωμα και αντίστροφα. Αρχικά, ο πληθυσμός δημιουργείται τυχαία. Κάθε άτομο στον πληθυσμό αποδίδεται, μέσω μιας λειτουργίας γυμναστικής, μιας αξίας φυσικής κατάστασης που αντικατοπτρίζει την ποιότητά του σε σχέση με την επίλυση του συγκεκριμένου προβλήματος. Ένα χρωμόσωμα αξιολογείται από μια συνάρτηση φυσικής κατάστασης για να προσδιοριστεί η ποιότητα της λύσης, δηλαδή πόσο αποτελεσματικό είναι στην επίλυση του προβλήματος. Η είσοδος της συνάρτησης φυσικής κατάστασης είναι το χρωμόσωμα και η έξοδος είναι η τιμή φυσικής κατάστασης αυτού του χρωμοσώματος. Σε κάθε κύκλο, καθορίζεται η καταλληλότητα κάθε υποψήφιας λύσης. Το επόμενο στάδιο είναι η επιλογή, όπου δημιουργείται ένας προσωρινός πληθυσμός στον οποίο τα πιο κατάλληλα άτομα είναι πιθανό να έχουν μεγαλύτερο αριθμό πιθανότητες από λιγότερο κατάλληλα άτομα, να χρησιμοποιηθούν ως γονείς για την επόμενη γενιά. Οι αναπαραγωγικοί τελεστές όπως η διασταύρωση και η μετάλλαξη εφαρμόζονται στα άτομα σε αυτόν τον πληθυσμό αποδίδοντας έναν νέο πληθυσμό.

Μια τεχνική συνόλου(ensemble method) ορίζεται ως μια τεχνική, που συνδυάζει ένα σύνολο μεμονωμένων τεχνικών μέσω ενός κανόνα συνάθροισης προκειμένου να επιλυθεί ένα δεδομένο πρόβλημα. Εννοιολογικά, οι μεμονωμένες τεχνικές που αποτελούν το σύνολο εκπαιδεύονται ξεχωριστά για την επίλυση του ίδιου έργου. Η τελική έξοδος του συνόλου είναι μια συγκέντρωση των διαφορετικών εξόδων που

δημιουργούνται από τη μοναδική τεχνική. Οι μεμονωμένες τεχνικές μπορούν να είναι είτε τεχνικές ταξινόμησης είτε παλινδρόμησης. Οι τεχνικές συνόλων μπορεί να είναι ομοιογενείς ή ετερογενείς. Το ομοιογενές σύνολο χρησιμοποιείται για να αναφέρεται: 1) ένα σύνολο που συνδυάζει μια βασική μέθοδο με τουλάχιστον δύο διαφορετικές διαμορφώσεις ή διαφορετικές παραλλαγές, 2) ένα σύνολο που συνδυάζει μια μέθοδο βάσης με ένα μετα-σύνολο, όπως «bagging», «boosting», ή «random subspace». Ο όρος ετερογενής, χρησιμοποιείται ως αναφορά σε ένα σύνολο που συνδυάζει τουλάχιστον δύο διαφορετικές βασικές μεθόδους. Η κύρια ιδέα πίσω από το σχεδιασμό αυτής της προσέγγισης ήταν να επιτευχθεί ένα υψηλό επίπεδο ακρίβειας που τουλάχιστον υπερβαίνει την ακρίβεια απόδοσης της μεμονωμένης τεχνικής από την οποία συντίθεται το σύνολο. Στην πραγματικότητα, κάθε μεμονωμένη τεχνική έχει τόσο πλεονεκτήματα όσο και μειονεκτήματα και ο συνδυασμός πολλών τεχνικών μπορεί επομένως να βελτιώσει την ακρίβεια της πρόβλεψης. Για το σκοπό αυτό, οι μεμονωμένες τεχνικές που σχηματίζουν το σύνολο πρέπει να είναι ακριβείς και ποικιλόμορφες. Οι τεχνικές μπορούν να ονομαστούν «ακριβείς» όταν δημιουργούν καλύτερη ακρίβεια από την τυχαία εικασία και είναι «ποικιλόμορφες» αν παράγουν διαφορετικά σφάλματα στην ίδια παρουσία δεδομένων. Ο συνδυασμός τέτοιων τεχνικών θα οδηγήσει συνεπώς σε πιο ακριβή αποτελέσματα, καθώς κάθε τεχνική χειρίζεται το σφάλμα που κάνουν οι άλλοι. Η επιτυχία ενός συνόλου, έγκειται στην ικανότητά του να επιτυγχάνει μια ευνοϊκή ανταλλαγή μεταξύ αυτών των δύο ιδιοτήτων.

3 ΑΝΑΛΥΣΗ ΤΕΧΝΙΚΩΝ

Είναι πολύ σημαντικό να γνωρίζουμε τους αλγορίθμους και τις τεχνικές με τις οποίες θα κάνουμε την κατηγοριοποίηση, καθώς έτσι μπορούμε να κατανοήσουμε, τον λόγο, που κάποιος παρουσιάζει καλύτερα αποτελέσματα, στις μετρικές αξιολόγησής του, από κάποιον άλλο.

3.1 Μέθοδος δέντρου απόφασης

Ένα δέντρο απόφασης είναι κατά βάση ένα δέντρο, άρα κάθε κόμβος μπορεί να έχει από κάτω του, δηλαδή να συνδέονται μαζί του και τον ακολουθούν, οποιοσδήποτε αριθμός κόμβων-παιδιών. Οι κόμβοι, που δεν έχουν δικούς τους κόμβους-παιδιά, ονομάζονται φύλλα. Ο κόμβος, που βρίσκεται στην κορυφή του δέντρου και είναι απαραίτητα μοναδικός, ονομάζεται ρίζα. Όλοι οι υπόλοιποι, λέγονται εσωτερικοί κόμβοι. Σε κάθε δέντρο απόφασης, μόνο στους κόμβους-φύλλα, δίνονται ετικέτες κατηγορίας. Οι υπόλοιποι λαμβάνουν τις συνθήκες ελέγχου χαρακτηριστικών, ώστε να διαχωρίζουν τις εγγραφές.

Υπάρχουν δύο κύριες λειτουργίες κατά την κατασκευή δέντρων: (1) αξιολόγηση διαχωρισμών για κάθε χαρακτηριστικό και επιλογή του καλύτερου διαχωρισμού και (2) δημιουργία κατατμήσεων χρησιμοποιώντας τον καλύτερο διαχωρισμό. Ένας δείκτης διαχωρισμού χρησιμοποιείται για την αξιολόγηση της ποιότητας των εναλλακτικών διαχωρισμών για ένα χαρακτηριστικό. Έχουν προταθεί διάφορα σχήματα διαχωρισμού στο παρελθόν (Rastogi & Shim 2000). Θεωρούμε δύο κοινά σχήματα: την εντροπία(entropy) και τον δείκτη «gini». Αν ένα σύνολο δεδομένων «S» περιέχει παραδείγματα από «m» τάξεις, τότε η εντροπία (S) και το Gini (S) ορίζονται ως εξής:

$$Entropy(S) = -\sum_{j=1}^{m} P_j \log P_j$$

$$Gini(S) = 1 - \sum_{j=1}^{m} P_j^2$$

Όπου «P_j» είναι η σχετική συχνότητα της κλάσης «j» στο σύνολο «S».

Ο τρόπος με τον οποίο θα δομηθεί το δέντρο, δηλαδή το πλήθος των κόμβων και των παιδιών τους, σε κάθε επίπεδο, το πλήθος των επιπέδων ή αλλιώς, βάθος του δέντρου και η απόφαση των συνθηκών ελέγχου, που θα περιέχει κάθε κόμβος, καθορίζονται κατά τη διάρκεια της εκπαίδευσης του κατηγοριοποιητή. Φυσικά, δεν δίνεται άμεση πρόσβαση, ώστε να επηρεάσουμε αυτή τη δομή, ωστόσο το πακέτο «scikit-learn», μέσω των συναρτήσεων, που παρέχει για τη δημιουργία των δέντρων, δίνει τη δυνατότητα, έμμεσης επιρροής, της δομής. Για παράδειγμα, με κατάλληλη μεταβλητή, μπορούμε να επιλέξουμε το κριτήριο βάση του οποίου θα αξιολογεί ο αλγόριθμος την ποιότητα κάθε διάσπασης σε κάθε κόμβο και με μία άλλη μεταβλητή να επιλέξουμε τη στρατηγική του διαχωρισμού στους κόμβους-παιδιά. Έπειτα, της

δημιουργίας του δέντρου, μπορεί να εφαρμοστεί μια διαδικασία, που ονομάζεται κλάδεμα(prunning), με στόχο τη μείωση του μεγέθους του δέντρου. Αυτό επιλέγεται, γιατί ένα μεγάλο σε μέγεθος δέντρο, είναι ένα δέντρο, επιρρεπές στη υπερπροσαρμογή στα δεδομένα εκπαίδευσης, κάτι που θα το απαγόρευε να γίνει χρήσιμος κατηγοριοποιητής σε δεδομένα, εκτός των δεδομένων εκπαίδευσης.

Η λειτουργία του δέντρου απόφασης μπορεί να περιγραφεί με τον εξής απλό τρόπο. Δοσμένης μιας εγγραφής και ενός δέντρου απόφασης, ελέγχεται η συνθήκη ελέγχου στη ρίζα του δέντρου. Ανάλογα το αποτέλεσμα, ακολουθείται η αντίστοιχη διακλάδωση προς τον κατάλληλο κόμβο-παιδί. Αν αυτός ο κόμβος είναι εσωτερικός, τότε γίνεται έλεγχος της εγγραφής και με τη συνθήκη ελέγχου του κόμβου, για να επιλεχθεί η ορθή διακλάδωση. Το ίδιο ακριβώς επαναληπτικό σενάριο εκτελείται, ώσπου το δέντρο, να οδηγηθεί σε φύλλο, οπότε και δίνει στην εγγραφή, την κατηγορία, που αντιστοιχεί στο φύλλο.

Ο αλγόριθμος, που χρησιμοποίησα για το δέντρο απόφασης ήταν ο «CART». Συγκεκριμένα το πακέτο «scikit-learn» αξιοποιεί μια βελτιωμένη έκδοση του «CART», ωστόσο δεν αναλύει τη λειτουργία του αλγορίθμου. «CART» σημαίνει δέντρα ταξινόμησης και παλινδρόμησης(classification and regression trees). Χαρακτηρίζεται από το γεγονός ότι κατασκευάζει δυαδικά δέντρα, δηλαδή κάθε εσωτερικός κόμβος έχει ακριβώς δύο εξερχόμενες ακμές. Οι διαχωρισμοί επιλέγονται χρησιμοποιώντας δυαδικά κριτήρια και το δέντρο που λαμβάνεται κλαδεύεται από κλάδεμα κόστους-πολυπλοκότητας. Αν ζητηθεί, ο αλγόριθμος μπορεί να εξετάσει το κόστος εσφαλμένης ταξινόμησης στην επαγωγή δέντρων. Επίσης επιτρέπει στους χρήστες να παρέχουν προηγούμενη κατανομή πιθανότητας. Ένα σημαντικό χαρακτηριστικό του CART είναι η ικανότητά του να δημιουργεί δέντρα παλινδρόμησης. Τα δέντρα παλινδρόμησης είναι δέντρα όπου τα φύλλα τους προβλέπουν πραγματικό αριθμό και όχι τάξη. Σε περίπτωση παλινδρόμησης, ο «CART» αναζητά διαχωρισμούς που ελαχιστοποιούν το τετράγωνο σφάλμα πρόβλεψης (η ελάχιστη-τετράγωνη απόκλιση). Η πρόβλεψη σε κάθε φύλλο βασίζεται στον σταθμισμένο μέσο όρο για τον κόμβο. Έχει τα ακόλουθα πλεονεκτήματα και μειονεκτήματα:

Πλεονεκτήματα

- Το CART μπορεί να χειριστεί εύκολα αριθμητικά και κατηγορικές μεταβλητές.
- Ο αλγόριθμος CART ο ίδιος θα αναγνωρίσει περισσότερο σημαντικές μεταβλητές και εξάλειψη μη σημαντικών.
- Το CART μπορεί εύκολα να χειριστεί τα έκτοπα(outliers).

Μειονεκτήματα

- Το CART μπορεί να έχει ασταθές δέντρο αποφάσεων. Ασήμαντη τροποποίηση του δείγματος εκπαίδευσης, όπως η εξάλειψη πολλών παρατηρήσεις και προκαλούν αλλαγές στο δέντρο απόφασης: αύξηση ή μείωση της πολυπλοκότητας των δέντρων, αλλαγές στο διαχωρισμό μεταβλητών και τιμών.
- Το CART κάνει διαχωρισμούς μόνο από μία μεταβλητή.

3.2 Μέθοδος «κ» πλησιέστερων γειτόνων

Ο κατηγοριοποιητής πλησιέστερων γειτόνων, ακολουθεί μια φράση η οποία είναι ιδιαίτερα γνωστή στο χώρο της μηχανικής μάθησης. Συγκεκριμένα, η φράση λέει το εξής: «Αν περπατάει σαν πάπια, φωνάζει σαν πάπια και μοιάζει οπτικά με πάπια, τότε πιθανόν να είναι πάπια». Η μέθοδος δημιουργίας του κατηγοριοποιητή, έχει ως βασική αρχή την εύρεση των δειγμάτων-εγγραφών εκπαίδευσης, των οποίων τα χαρακτηριστικά μοιάζουν με τα χαρακτηριστικά του δείγματος-εγγραφής ελέγχου. Οι εγγραφές εκπαίδευσης είναι αυτές, που αποκαλούνται «πλησιέστεροι γείτονες». Οι ίδιες εγγραφές, είναι αυτές, που χρησιμοποιούνται για την πρόβλεψη της κατηγορίας, στην οποία θα τοποθετηθεί η εγγραφή ελέγχου.

Στην περίπτωση της μεθόδου του πλησιέστερου γείτονα, οι εγγραφές εκπαίδευσης αναπαρίστανται από σημεία, σε χώρο τόσων διαστάσεων, όσα τα χαρακτηριστικά των εγγραφών. Δοσμένου ενός δείγματος-εγγραφής ελέγχου, υπολογίζεται η απόστασή του από τα σημεία του συνόλου δείγματος εκπαίδευσης. Οπότε οι «κ» γείτονες, με τη μικρότερη απόσταση, από ένα δοσμένο δείγμα «ν» αναφέρονται, στα «κ» σημεία, που είναι πλησιέστερα στο «ν». Ισχύει ο γενικός κανόνας, ότι η κατηγορία στην οποία ανήκουν, τα περισσότερα, από τα «κ» σημεία, είναι και η κατηγορία του «ν». Μπορεί, όμως να παρουσιαστεί το φαινόμενο, τα σημεία εκπαίδευσης, που ορίζουν την κατηγορία του σημείου ελέγχου, να οδηγήσουν σε ισοψηφία, ανάμεσα στις υποψήφιες κατηγορίες, για το «ν». Αν συμβεί αυτό, τότε η κατηγορία, επιλέγεται τυχαία, ανάμεσα από τις υποψήφιες και δίνεται στο «ν».

Όπως γίνεται φανερό, είναι πολύ σημαντικό να καταφέρουμε να ορίσουμε το σωστό πλήθος «κ» γειτόνων, ώστε να πάρουμε το καλύτερο δυνατό αποτέλεσμα κατά την κατηγοριοποίηση. Αν δεν καταφέρουμε να επιλέξουμε καλή τιμή, για το «κ», δύο σενάρια μπορούν να συμβούν. Το πρώτο είναι η τιμή, που θα δώσουμε στο «κ», να είναι μικρή, οπότε η τεχνική του πλησιέστερου γείτονα, μπορεί να δημιουργήσει μοντέλο με υπερπροσαρμογή στα δεδομένα εκπαίδευσης, που οφείλεται στον θόρυβο, που περιέχουν. Το δεύτερο σενάριο, είναι να επιλέξουμε μεγάλο «κ». Σε αυτή την περίπτωση, ο κατηγοριοποιητής, διατρέχει τον κίνδυνο, να κατηγοριοποιεί λανθασμένα, το δεδομένο ελέγχου, επειδή λαμβάνονται ως πλησιέστεροι γείτονες και επηρεάζουν το αποτέλεσμα, σημεία δεδομένων που βρίσκονται πολύ μακριά από την πραγματική γειτονιά του.

Στα δέντρα, που υλοποιούνται με τη μέθοδο των πλησιέστερων γειτόνων, δεν όρισα αυστηρή τιμή, στη μεταβλητή ελέγχου του αλγορίθμου, καθώς το εγχειρίδιο του «scikit-learn», αναφέρει, πως μπορεί να αποφασίσει το ίδιο για τον πιο καλό αλγόριθμο, βάσει των δεδομένων. Αφού το σύνολο δεδομένων είναι σταθερό, θα χρησιμοποιείται πάντα ο ίδιος αλγόριθμος. Συνεπώς, θα αναφερθώ και στους τρεις αλγορίθμους, που υπάρχουν, αλλά με συντομία.

Πρώτος, ο αλγόριθμος «Ball Tree», που μπορεί να θεωρηθεί μετρικό δέντρο. Τα μετρικά δέντρα οργανώνουν και δομούν σημεία δεδομένων λαμβάνοντας υπόψη τον μετρικό χώρο στον οποίο βρίσκονται τα σημεία. Χρησιμοποιώντας μετρήσεις, τα σημεία δεν χρειάζεται να είναι πεπερασμένων διαστάσεων ή σε διανύσματα (Kumar, Zhang & Nayar, 2008). Ο αλγόριθμος χωρίζει τα σημεία δεδομένων σε δύο ομάδες. Κάθε σύμπλεγμα περικλείεται από έναν κύκλο (2D) ή μια σφαίρα (3D). Η σφαίρα

ονομάζεται συχνά υπερσφαίρα. Από τη μορφή σφαίρας του συμπλέγματος, προέρχεται η ονομασία του αλγορίθμου. Κάθε σύμπλεγμα αντιπροσωπεύει έναν κόμβο του δέντρου. Τα παιδιά επιλέγονται να έχουν τη μέγιστη απόσταση μεταξύ τους, συνήθως χρησιμοποιώντας την ακόλουθη κατασκευή σε κάθε επίπεδο του δέντρου. Κατ 'αρχάς, έχει ρυθμιστεί το κεντρικό σημείο ολόκληρου του cloud των σημείων δεδομένων. Το σημείο με τη μέγιστη απόσταση από το κέντρο επιλέγεται ως το κέντρο του πρώτου συμπλέγματος και του θυγατρικού κόμβου. Το σημείο που βρίσκεται πιο μακριά από το κέντρο του πρώτου συμπλέγματος επιλέγεται ως το κεντρικό σημείο του δεύτερου συμπλέγματος. Στη συνέχεια, όλα τα άλλα σημεία δεδομένων αντιστοιχίζονται στον κόμβο και το σύμπλεγμα στο πλησιέστερο κέντρο, είτε είναι το σύμπλεγμα 1 είτε το σύμπλεγμα 2. Οποιοδήποτε σημείο μπορεί να είναι μέλος μόνο ενός συμπλέγματος. Οι γραμμές σφαίρας μπορούν να τέμνονται μεταξύ τους, αλλά τα σημεία πρέπει να αντιστοιχίζονται σαφώς σε ένα σύμπλεγμα. Εάν ένα σημείο βρίσκεται ακριβώς στη μέση και των δύο κέντρων και έχει συνεπώς την ίδια απόσταση και από τις δύο πλευρές, πρέπει να αντιστοιχιστεί σε ένα σύμπλεγμα. Οι συστάδες μπορεί να μην είναι ισορροπημένες. Η διαδικασία διαίρεσης των σημείων δεδομένων σε δύο συστάδες / σφαίρες επαναλαμβάνεται σε κάθε σύμπλεγμα έως ότου επιτευχθεί ένα καθορισμένο βάθος. Αυτό οδηγεί σε ένθετο σύμπλεγμα που περιέχει όλο και περισσότερους κύκλους.

Δεύτερος, ο αλγόριθμος «KD Tree» είναι ένας από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους πλησιέστερων γειτονικών. Τα σημεία δεδομένων χωρίζονται σε κάθε κόμβο σε δύο σύνολα. Όπως και ο προηγούμενος αλγόριθμος, ο «KD Tree» είναι ένας αλγόριθμος δυαδικού δέντρου, που καταλήγει πάντα σε δύο κόμβους το πολύ. Το κριτήριο διαχωρισμού που επιλέγεται είναι συνήθως η μέση τιμή. Ο αλγόριθμος «KD-Tree» χρησιμοποιεί πρώτα τη διάμεσο του πρώτου άξονα και στη συνέχεια, στο δεύτερο επίπεδο, τη διάμεσο του δεύτερου άξονα.

Τελευταίος, ο αλγόριθμος ωμής βίας(brute force), όπου αφού φορτωθούν τα δεδομένα και αποφασίζεται και αρχικοποιείται το πλήθος των πλησιέστερων γειτόνων. Για κάθε εγγραφή του συνόλου δεδομένων ελέγχου, υπολογίζεται η απόστασή τους από τα δεδομένα εκπαίδευσης, χρησιμοποιώντας μια συγκεκριμένη μεθοδολογία εύρεσης απόστασης, όπως η Ευκλείδεια απόσταση. Έπειτα, τα σημεία ταξινομούνται με αύξουσα σειρά απόστασης. Επιλέγονται τα «κ» πρώτα σημεία τής ταξινομημένης λίστας. Βρίσκεται η κλάση στην οποία ανήκουν τα περισσότερα και αυτή δίνεται ως προβλεπόμενη κλάση για την εγγραφή η οποία θέλαμε να κατηγοριοποιηθεί.

3.3 Μέθοδος τυχαίου δάσους

Το τυχαίο δάσος, αποτελεί μια κατηγορία μεθόδων ομάδας(ensemble methods). Η βασική λειτουργία του έγκειται στο ότι συνδυάζει τις προβλέψεις πολλών δέντρων απόφασης. Κάθε δέντρο έχει προέλθει από τις τιμές, ανεξάρτητων τυχαίων διανυσμάτων. Μια σταθερή κατανομή πιθανότητας είναι υπεύθυνη για την παραγωγή αυτών των διανυσμάτων. Με άλλα λόγια, ο αλγόριθμος δημιουργίας του μοντέλου τυχαίου δάσους, επιλέγει διαφορετικά τυχαία χαρακτηριστικά μέσα από το «dataset», σύμφωνα με τη σταθερή κατανομή πιθανότητας, ώστε βάσει κάποιου αλγορίθμου

δημιουργίας δέντρων απόφασης, συναρμολογεί τα δέντρα, που θα αποτελέσουν το δάσος. Στην περίπτωση μου, η υλοποίηση του τυχαίου δάσους στο πακέτο «scikit-learn» χρησιμοποιεί τον αλγόριθμο του δέντρου απόφασης, του ίδιου πακέτου, οπότε κάθε δέντρο έχει δημιουργηθεί από την παραλλαγή του αλγορίθμου «CART». Επιπλέον, η υλοποίηση του «random forest» στο πακέτο «scikit-learn», αντί να επιτρέπει σε κάθε δέντρο να ψηφίζει για να αποφασιστεί η κλάση τής εγγραφής ελέγχου, συνδυάζει τους κατηγοριοποιητές βγάζοντας το μέσο όρο της πιθανολογικής τους πρόβλεψης. Εφόσον έχουμε μία μέθοδο δημιουργίας μοντέλου κατηγοριοποίησης, που αποτελείται από πλήθος απλούστερων κατηγοριοποιητών, η σωστή επιλογή του πλήθους των δέντρων απόφασης, παίζει σημαντικό ρόλο στην επιτυχία του μοντέλου.

Έχει αποδειχτεί σε θεωρητικό επίπεδο, πως το σφάλμα γενίκευσης των τυχαίων δασών, έχει άνω όριο, το οποίο συγκλίνει σύμφωνα με τον παρακάτω τύπο υπολογισμού, αρκεί το πλήθος των δέντρων απόφασης να είναι αρκετά μεγάλο.

$Generror \leq \frac{\bar{P}(1-s^2)}{s^2}$ Όπου «Generror» είναι το σφάλμα γενίκευσης (Generalization Error), «$\bar{P}$» είναι η μέση συσχέτιση μεταξύ των δέντρων και «s» μια ποσότητα, η οποία μετράει την «ισχύ» των δέντρων κατηγοριοποίησης.

Η λειτουργία του τύπου του σφάλματος γενίκευσης, περιλαμβάνεται στο ότι όσο αυξάνεται η συσχέτιση μεταξύ των δέντρων ή η «ισχύς» τής ομάδας μειώνεται, τόσο το όριο του σφάλματος γενίκευσης αυξάνεται. Η τυχαιότητα των διανυσμάτων βοηθάει στην μείωση της συσχέτισης μεταξύ των δέντρων και έτσι το σφάλμα γενίκευσης μειώνεται-βελτιώνεται.

3.4 Μέθοδος μηχανής διανύσματος υποστήριξης

Η τεχνική της μηχανής διανύσματος υποστήριξης, βασίζεται στη θεωρία της στατιστικής εκπαίδευσης. Το θετικό με αυτή τη μέθοδο, είναι ότι έχει δείξει ελπιδοφόρα εμπειρικά αποτελέσματα σε πρακτικές εφαρμογές, όπως η κατηγοριοποίηση κειμένου και η αναγνώριση χειρόγραφων ψηφίων. Επίσης, έχει παρατηρηθεί ότι η συγκεκριμένη μέθοδος φέρνει καλά αποτελέσματα σε δεδομένα πολλών διαστάσεων. Ένα άλλο μοναδικό χαρακτηριστικό, της μεθόδου, είναι ότι αντιπροσωπεύει το όριο τής απόφασης χρησιμοποιώντας, ένα υποσύνολο των εγγραφών εκπαίδευσης, οι οποίες λέγονται διανύσματα υποστήριξης(support vectors).

Τα χαρακτηριστικά της μεθόδου, που την καθιστούν συχνά τη μέθοδο επιλογής των ερευνητών για κατηγοριοποίηση, είναι τέσσερα. Πρώτον, το πρόβλημα της εκπαίδευσης μιας μηχανής διανυσμάτων υποστήριξης, μπορεί να διατυπωθεί, ως ένα κυρτό πρόβλημα βελτιστοποίησης, για το οποίο υπάρχουν διαθέσιμοι αποτελεσματικοί αλγόριθμοι εύρεσης του ολικού ελάχιστου, της αντικειμενικής συνάρτησης. Δεύτερον, η μέθοδος είναι φτιαγμένη έτσι, ώστε να ελέγχει τη χωρητικότητα, μεγιστοποιώντας το περιθώριο του ορίου απόφασης. Τρίτον, η τεχνική αυτή είναι δυνατό να εφαρμοστεί σε κατηγορικά χαρακτηριστικά, εισάγοντας ψεύτικες μεταβλητές για κάθε χαρακτηριστικό, που βρίσκεται στα δεδομένα. Τέλος, οι μηχανές διανυσμάτων υποστήριξης αφορούν δυαδική κατηγοριοποίηση, δηλαδή κατηγοριοποίηση σε δύο

ομάδες. Ωστόσο, υπάρχουν τρόποι, για αυτή την μέθοδο κατηγοριοποίησης, ώστε να την επεκτείνουν, να λειτουργήσει για περισσότερες από δύο ομάδες.

Η λειτουργία της μεθόδου μπορεί να περιγραφεί ως εξής. Δίνεται ένα δείγμα εγγραφών εκπαίδευσης και ένα δείγμα εγγραφών ελέγχου. Η μέθοδος προσπαθεί να φτιάξει τον κατηγοριοποιητή επιλέγοντας τα κατάλληλα υπερεπίπεδα μέγιστου περιθωρίου(maximal margin hyperplane), όπως λέγονται, ώστε οι κατηγορίες των εγγραφών να βρίσκονται σε διαχωρισμένα μέρη-υπερεπίπεδα. Έτσι κάθε υπερεπίπεδο, αφορά μία κατηγορία. Τα διάφορα μέρη, είναι παράλληλα μεταξύ τους και διαχωρίζονται με γραμμές, που τοποθετεί ο αλγόριθμος, στον κατηγοριοποιητή, έτσι ώστε να μπορεί το μοντέλο να κατηγοριοποιεί, τις εγγραφές ελέγχου. Αυτές οι γραμμές ονομάζονται, όρια απόφασης. Το «περιθώριο» αφορά την απόσταση, που απέχουν οι γραμμές από τις πιο κοντινές τους εγγραφές, πάνω στο επίπεδο. Αν το περιθώριο είναι μικρό, τότε εγγραφές, με πολύ όμοια χαρακτηριστικά, μπορεί να βρεθούν σε διαφορετικό υπερεπίπεδο, άρα και κατηγορία, λόγω μικροδιαφορών, κάτι που μπορεί να οδηγήσει σε λάθος κατηγοριοποίηση. Αντίθετα, αν τα περιθώρια είναι μεγάλα, τότε το σύστημα, τείνει να κατηγοριοποιεί με μικρότερο σφάλμα γενίκευσης, σε δεδομένα, που δεν έχει ξαναδεί.

Βασικό στοιχείο της τεχνικής «support vector machine», είναι η προσπάθεια να μεγιστοποιήσουμε το περιθώριο μεταξύ των σημείων δεδομένων και του υπερπλάνου. Η συνάρτηση απώλειας που βοηθά στη μεγιστοποίηση του περιθωρίου είναι η απώλεια αρθρώσεων.

$$c\big(x, y, f(x)\big) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & else \end{cases} \quad c\big(x, y, f(x)\big) = \big(1 - y * f(x)\big)_+$$

Το κόστος είναι 0 εάν η προβλεπόμενη τιμή και η πραγματική τιμή είναι του ίδιου σημείου. Εάν δεν είναι, τότε υπολογίζουμε την τιμή απώλειας. Προσθέτουμε επίσης μια παράμετρο κανονικοποίησης στη συνάρτηση κόστους. Ο στόχος της παραμέτρου κανονικοποίησης είναι η εξισορρόπηση της μεγιστοποίησης του περιθωρίου και της απώλειας. Μετά την προσθήκη της παραμέτρου κανονικοποίησης, οι λειτουργίες κόστους φαίνονται όπως παρακάτω.

$$\min_w \lambda \, \|w\|^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

4 ΤΑ ΔΕΔΟΜΕΝΑ

Τα δεδομένα, οι πληροφορίες, η γνώση και η σοφία είναι στενά συνδεδεμένες έννοιες, αλλά ο καθένας έχει τον δικό του ρόλο σε σχέση με τον άλλο και κάθε όρος έχει τη δική του σημασία. Σύμφωνα με μια κοινή άποψη, τα δεδομένα συλλέγονται και αναλύονται. Τα δεδομένα καθίστανται κατάλληλες πληροφορίες για τη λήψη αποφάσεων μόλις αναλυθούν με κάποιο τρόπο. Μπορεί κανείς να πει ότι ο βαθμός στον οποίο ένα σύνολο δεδομένων είναι ενημερωτικός για κάποιον εξαρτάται από το βαθμό στον οποίο είναι απρόσμενο από αυτό το άτομο. Η ποσότητα των πληροφοριών που περιέχονται σε μια ροή δεδομένων μπορεί να χαρακτηρίζεται από την εντροπία «Shannon» της.

Η γνώση είναι η κατανόηση που βασίζεται σε εκτεταμένη εμπειρία σχετικά με πληροφορίες σχετικά με ένα θέμα. Για παράδειγμα, το ύψος του όρους Έβερεστ χαρακτηρίζεται ως ένα δεδομένο. Το ύψος μπορεί να μετρηθεί ακριβώς με ένα αλτίμετρο και να εισαχθεί σε μια βάση δεδομένων. Αυτά τα δεδομένα μπορούν να συμπεριληφθούν σε ένα βιβλίο μαζί με άλλα δεδομένα για το Όρος Έβερεστ για να περιγράψουν το βουνό με τρόπο χρήσιμο για όσους επιθυμούν να πάρουν μια απόφαση σχετικά με την καλύτερη μέθοδο για να το ανέβουν. Η κατανόηση που βασίζεται στην εμπειρία αναρρίχησης βουνών που θα μπορούσε να συμβουλεύει άτομα στο δρόμο να φτάσουν στην κορυφή του Έβερεστ μπορεί να θεωρηθεί ως «γνώση». Η πρακτική αναρρίχηση της κορυφής του όρους Έβερεστ βάσει αυτής της γνώσης μπορεί να θεωρηθεί ως «σοφία». Με άλλα λόγια, η σοφία αναφέρεται στην πρακτική εφαρμογή της γνώσης ενός ατόμου σε περιπτώσεις όπου μπορεί να προκύψει καλό. Έτσι, η σοφία συμπληρώνει και συμπληρώνει τη σειρά "δεδομένα", "πληροφορίες" και "γνώση" ολοένα και πιο αφηρημένων εννοιών.

Τα δεδομένα συχνά θεωρείται ότι είναι η λιγότερο αφηρημένη έννοια, η πληροφόρηση είναι επόμενη και η πιο αφηρημένη η γνώση. Σε αυτήν την περίπτωση, τα δεδομένα γίνονται πληροφορίες διερμηνείας. Για παράδειγμα, το ύψος του όρους Έβερεστ θεωρείται γενικά «δεδομένα», ένα βιβλίο σχετικά με τα γεωλογικά χαρακτηριστικά του Έβερεστ μπορεί να θεωρηθεί «πληροφορία» και ένας οδηγός ορειβάτη που περιέχει πρακτικές πληροφορίες σχετικά με τον καλύτερο τρόπο για να φτάσετε στην κορυφή του Έβερεστ μπορεί να θεωρηθεί «γνώση» . Οι «πληροφορίες» φέρουν μια ποικιλία εννοιών που κυμαίνονται από την καθημερινή χρήση έως την τεχνική χρήση. Αυτή περίπτωση, ωστόσο, υποστηρίχθηκε επίσης ότι αντιστρέφει τον τρόπο με τον οποίο προκύπτουν τα δεδομένα από πληροφορίες και πληροφορίες από τη γνώση. Σε γενικές γραμμές, η έννοια της πληροφορίας σχετίζεται στενά με τις έννοιες του περιορισμού, της επικοινωνίας, του ελέγχου, των δεδομένων, της μορφής, της διδασκαλίας, της γνώσης, της έννοιας, του διανοητικού ερεθίσματος, του μοτίβου, της αντίληψης και της αναπαράστασης.

Για την εκπαίδευση των μοντέλων κατηγοριοποίησης, χρησιμοποιήθηκαν δεδομένα «dna» της οικογένειας του λαγοκέφαλου. Τα δεδομένα τα κατέβασα, από την σελίδα «http://www.boldsystems.org/». Πρόκειται για το σύστημα με όνομα «Barcode Of Life Data» με το ακρωνύμιο «BOLD», το οποίο είναι ένα σύστημα αποθήκευσης δεδομένων στο νέφος(cloud) και ανάλυσής τους. Αναπτύσσεται από το «Centre for Biodiversity Genomics», στον Καναδά. Από εκεί κατέβασα ένα αρχείο τύπου «fasta»

και ξεκίνησα την επεξεργασία, ώστε να εξάγω όσα χρήσιμα δεδομένα μπορούσα να βρω, για την εκπαίδευση.

Το περιεχόμενο του αρχείου, αποτελείται από εγγραφές δύο γραμμών. Η πρώτη περιέχει το όνομα του είδους του οργανισμού και κάποια αναγνωριστικά στοιχεία, σχετικά με τη δεύτερη γραμμή. Η δεύτερη γραμμή αποτελεί το δεδομένο μεγάλου ενδιαφέροντος για εμένα, καθώς πρόκειται για την γενετική ακολουθία (dna), του οργανισμού.

Το πρώτο μου βήμα, ήταν να ανοίξω και να ελέγξω, οπτικά, το αρχείο για τυχόν εμφανή προβλήματα. Έπειτα βρήκα ένα κομμάτι κώδικα στη σελίδα «https://biopython.org/wiki/Sequence_Cleaner», του πακέτου «biopython», το οποίο έχει τη δυνατότητα να δημιουργήσει ένα νέο αρχείο τύπου «fasta», βάσει του πρωτότυπου, από το οποίο θα λείπουν εγγραφές γενετικών ακολουθιών, που εμφανίζονται πολλαπλές φορές. Επίσης ο κώδικας, μπορεί να αφαιρέσει πολύ μικρές ακολουθίες και ακολουθίες, με πολλά σύμβολα μη γνωστού νουκλεοτιδίου, δηλαδή σύμβολα «N». Ο χρήστης έχει τη δυνατότητα να επιλέξει το ελάχιστο μέγεθος των «dna» ακολουθιών, που θα διατηρηθούν και το αποδεκτό ποσοστό «N» σε κάθε ακολουθία. Επέλεξα να κρατήσω, όλες τις μοναδικές ακολουθίες, ανεξάρτητα από το μήκος τους και το ποσοστό άγνωστων νουκλεοτιδίων.

Το επόμενό μου βήμα, ήταν να διαβάσω πόσα σε πλήθος είναι τα διαφορετικά είδη, που υπάρχουν, στο νέο αρχείο-«dataset» και να το διασταυρώσω με την πληροφορία από το σύστημα «BOLD». Τα είδη ήταν δεκατρία, όπως δηλώνει το πληροφοριακό σύστημα, συν μερικές εγγραφές, στις οποίες δεν υπήρχε όνομα είδους λαγοκέφαλου, αλλά μόνο η ονομασία της οικογένειας των λαγοκέφαλων. Αφαίρεσα και αυτές τις εγγραφές από το αρχείο δεδομένων, που θα χρησιμοποιούσα. Συνεπώς, το αρχείο-«dataset», ήταν έτοιμο για χρήση στην εκπαίδευση και αξιολόγηση των μοντέλων κατηγοριοποίησης των ψαριών, στο είδος τους, βάσει της γενετικής τους ακολουθίας.

Ας πάρουμε μια εικόνα από τη δομή των δεδομένων, μέσα στο αρχείο(dataset), που προέκυψε μετά την προεπεξεργασία των δεδομένων.

```
>ABFJ191-07|Lagocephalus spadiceus|COI-5P|JF952772
CCTCTATCTAGTATTTGGTGCCTGAGCCGGAATAGTGGGAACGGCCCTGAGCCTCCTTATTCGGGCAGAGCT
>ANGBF24862-19|Lagocephalus gloveri|COI-5P|KP641407
TCCTCCTTATTCGGGCAGAGCTAATCCAGCCGGGTGCTCTCCTAGGTGACGATCAGATTTATAACGTAATCC
>ANGBF24869-19|Lagocephalus guentheri|COI-5P|KU508429
TTATTCGGGCAGAGCTAAGCCAACCAGGTGCTCTCCTGGGGGACGACCAGATTTATAATGTAATCGTCACGC
>ANGBF24871-19|Lagocephalus guentheri|COI-5P|KX758092
CTATCTAGTATTTGGTGCCTGAGCCGGAATAGTGGGAACGGCCCTGAGCCTCCTTATTCGGGCAGAGCTAAC
>ANGBF24873-19|Lagocephalus guentheri|COI-5P|MF588654
AGCCTCCTTATTCGGGCAGAGCTAAGCCAACCAGGTGCTCTCCTGGGGGACGACCAGATTTATAATGTAATC
```

Ο τύπος αρχείου στον οποίο ανήκει το σύνολο δεδομένων, ονομάζεται «fasta». Στη βιοπληροφορική και τη βιοχημεία, η μορφή «fasta» είναι μια μορφή αρχείου βασισμένη σε κείμενο, για την αναπαράσταση αλληλουχιών νουκλεοτιδίων ή αλληλουχιών αμινοξέων (πρωτεΐνης), όπου τα νουκλεοτίδια ή τα αμινοξέα αντιπροσωπεύονται χρησιμοποιώντας κωδικούς ενός γράμματος. Η μορφή επιτρέπει επίσης τα ονόματα ακολουθιών και τα σχόλια να προηγούνται των ακολουθιών. Η μορφή προέρχεται από το πακέτο λογισμικού «fasta», αλλά έχει πλέον γίνει σχεδόν

καθολικό πρότυπο στον τομέα της βιοπληροφορικής. Η απλότητα της μορφής «fasta» καθιστά εύκολο τον χειρισμό και την ανάλυση των ακολουθιών χρησιμοποιώντας εργαλεία επεξεργασίας κειμένου και «scripting» γλώσσες προγραμματισμού, όπως οι «R», «Python», «Ruby» και «Perl».

Όπως παρουσιάζεται το παραπάνω κομμάτι του πακέτου δεδομένων, κάθε εγγραφή απλώνεται σε δύο γραμμές του αρχείου, οπότε έχουμε πέντε εγγραφές. Το σύμβολο «>» υποδηλώνει έναρξη νέα εγγραφής. Σε κάθε εγγραφή, η πάνω γραμμή, ονομάζεται κεφαλίδα(header) και περιέχει διάφορα αναγνωριστικά, αλλά και το όνομα του είδους στο οποίο ανήκει ο οργανισμός από τον οποίο προέρχεται το «dna», που ακολουθεί στην δεύτερη γραμμή. Στο δικό μου αρχείο, η χρήσιμη πληροφορία της πρώτης γραμμής είναι μόνο το είδος στο οποίο ανήκει το ψάρι, άρα η δεύτερη κατά σειρά πληροφορία, μετά το πρώτο «|». Η δεύτερη γραμμή κάθε εγγραφής περιέχει τη γενετική ακολουθία, στην περίπτωσή μου, τμήμα «dna» από τον οργανισμό. Φυσικά, στην εικόνα, δεν φαίνεται το πραγματικό μέγεθος των δεύτερων γραμμών των πέντε εγγραφών. Από όλη τη δομή του «dna», μας απασχολούν μόνο οι αζωτούχες βάσεις, οι οποίες είναι και αυτές, που φαίνονται στην εικόνα. Συγκεκριμένα, οι αζωτούχες βάσεις είναι η αδενίνη(A),η θυμίνη(T),η γουανίνη(G) και η κυτοσίνη(C). Αυτά τα στοιχεία σχηματίζουν δεσμούς μεταξύ τους, ανά δύο και δομούν την διπλή έλικα, το «dna». Οι δεσμοί είναι αυστηρά καθορισμένοι και επιτρέπουν μόνο σύνδεση αδενίνης με θυμίνη και γουανίνης με κυτοσίνη. Βάσει αυτού του χαρακτηριστικού, το «dna» χαρακτηρίζεται ως δίκλωνο, δηλαδή έχοντας τη μία από τις δύο έλικες, μπορούμε να υπολογίσουμε την άλλη, τον άλλο κλώνο. Έτσι γίνεται φανερό, γιατί στο αρχείο υπάρχει μόνο μία γραμμή για το «dna», το οποίο κανονικά αποτελείται από δύο. Σημαντικό στοιχείο των αρχείων «fasta» είναι ότι μπορεί να εμφανίζονται και άλλα σύμβολα, εκτός από τις τέσσερις αζωτούχες βάσεις, που το καθένα κάτι υποδηλώνει. Το μόνο από τα επιπλέον σύμβολα, που εμφανίζεται στο «dataset» είναι η παύλα(-), η οποία υποδηλώνει κενό μη υπολογιζόμενου πλήθους αζωτούχων βάσεων.

5 Η ΕΡΕΥΝΑ

Στόχος της έρευνας ήταν οι τέσσερις τεχνικές, μαζί με τους αλγόριθμους που τις υλοποιούν, να έχουν την καλύτερη δυνατή ευκαιρία στην κατηγοριοποίηση, βάσει των παραμέτρων, που επέλεξα να μελετήσω. Αυτό σημαίνει, πως πρώτα για τους δύο αλγόριθμους τυχαίου δάσους και πλησιέστερων γειτόνων, έπρεπε να προσδιορίσω το βέλτιστο πλήθος δέντρων και πλήθος γειτόνων, αντίστοιχα. Είχα επιλέξει τρία διαφορετικά πλήθη δέντρων και γειτόνων, για να ελέγξω τα αποτελέσματά τους. Οι δύο άλλες τεχνικές, δέντρου απόφασης και μηχανή διανύσματος υποστήριξης, δεν έχουν παράμετρο πλήθους, για να εξεταστεί. Πριν την εφαρμογή των αλγορίθμων, έπρεπε να δημιουργήσω τα χαρακτηριστικά, μέσα από τα «dna», οπότε το επόμενο βήμα, ήταν να βρω το κατάλληλο μέγεθος χαρακτηριστικών για κάθε αλγόριθμο.

Οι τεχνικές κατηγοριοποίησης, που έχω επιλέξει, έχουν τη δυνατότητα να λειτουργήσουν καλύτερα, αν αντί για ολόκληρες τις ακολουθίες, τους εισάγω χαρακτηριστικά για κάθε ακολουθία. Αυτό το πετυχαίνω με σπάσιμο των ακολουθιών σε ίσα τμήματα «k» μεγέθους, «k-mers» ή και «k-grams», όπως λέγονται. Τα τμήματα, που θα δημιουργηθούν από το σπάσιμο κάθε «dna», θα τοποθετηθούν στο διάνυσμα χαρακτηριστικών της αντίστοιχης ακολουθίας. Έπειτα, τα διανύσματα που προκύπτουν, θα αξιοποιήσουν οι αλγόριθμοι, ως είσοδο, για να εκπαιδεύσουν τα μοντέλα κατηγοριοποίησης. Τα «k-mers», που επέλεξα να μελετήσω, είναι μεγέθους τεσσάρων, πέντε, έξι, εφτά, οκτώ και ως ακραία περίπτωση είκοσι χαρακτήρων. Φυσικά, όπως πριν για κάθε τεχνική θα επιλεγεί το μέγεθος χαρακτηριστικών, που θα παρουσιάσει τα καλύτερα αποτελέσματα στις δοκιμές. Η σημαντική διαφορά, που πρέπει να αναφερθεί εδώ, είναι ότι για τις δοκιμές του μεγέθους των χαρακτηριστικών, θα αξιοποιηθούν τα αποτελέσματα του πρώτου ελέγχου για το πλήθος δέντρων και γειτόνων, στους ανάλογους αλγορίθμους, όπου ανήκουν.

Πριν αναφερθώ στις μετρικές αξιολόγησης, θεωρώ λογικό να δείξω μερικά τμήματα του κώδικα. Ιδιαίτερα κάποια, που επαναλαμβάνονται στα διάφορα πειράματα και ήταν κρίσιμης σημασίας. Εφόσον έχω επιλέξει να αντιμετωπίσω τις γενετικές ακολουθίες ως κείμενο, τότε μπορώ να ακολουθήσω την παρακάτω προσέγγιση.

```
def getKmers(sequence, size):
    return [sequence[x:x+size].lower() for x in range(len(sequence) - size + 1)]
```

Χρησιμοποιώντας αυτή τη συνάρτηση χωρίζω κάθε γενετική ακολουθία σε μικρότερα κομμάτια, τα οποία μπορούμε να χαρακτηρίσουμε ως λέξεις. Ειδικό χαρακτηριστικό αυτής της συνάρτησης, που προκύπτει από τον κώδικα, που περιέχει εντός της δομής «for», είναι ότι κάθε επόμενη λέξη ξεκινάει με το δεύτερο γράμμα της προηγούμενης. Αυτό σημαίνει, πως ανάλογα το μέγεθος των λέξεων, θα έχουμε επικάλυψη μεταξύ τους και μάλιστα τη μέγιστη δυνατή, καθώς μετακινούμαστε μόνο ένα σύμβολο δεξιά για να πάρουμε την επόμενη λέξη.

```
sentences = []
for k in range(len(clear_records)):
    sentence = ' '.join(getKmers(str(clear_records[k].seq), size=6))
    sentences.insert(k,sentence)
```

Στη φυσική γλώσσα, χρησιμοποιώντας λέξεις σχηματίζουμε προτάσεις για να επικοινωνήσουμε. Το ίδιο έκανα και για την έρευνά μου, με τη δομή επανάληψης «for», της παραπάνω εικόνας. Πρόκειται για βήμα, που δεν μπορώ να παραλείψω, καθώς με αυτόν τον τρόπο, επιλύω το πρόβλημα της διαφοράς μεταξύ των «dna» ακολουθιών σε μέγεθος.

```
# Splitting the dataset into the Training set and Test set
X_train, X_test, y_train, y_test=train_test_split(X,y,test_size = 0.2)
```

Σε αυτό το σημείο του κώδικα, όπως λέει και το σχόλιο της πράσινης γραμμής, γίνεται ο διαχωρισμός του συνόλου δεδομένων σε δεδομένα εκπαίδευσης(training set) και δεδομένα ελέγχου(test set). Τα σύμβολα «x» και «y» συμβολίζουν την εξαρτημένη μεταβλητή του προβλήματος(y) και την ανεξάρτητη(x). Επίσης, η τιμή «0,2» στη μεταβλητή που εκχωρείται δηλώνει τι ποσοστό των συνολικών δεδομένων θα μπουν στα δεδομένα ελέγχου.

```
#fitting classifier to the training set
classifier = DecisionTreeClassifier(criterion='entropy')
classifier.fit(X_train,y_train)
```

```
# Fitting classifier to the Training set
classifier = KNeighborsClassifier(n_neighbors = 5,weights='uniform',algorithm='auto',p=2,metric='minkowski')
classifier.fit(X_train,y_train)
```

```
#fitting classifier to the training set
classifier = RandomForestClassifier(n_estimators = 20, criterion = 'entropy')
classifier.fit(X_train,y_train)
```

```
# Fitting classifier to the Training set
classifier = SVC(kernel = 'rbf')
classifier.fit(X_train,y_train)
```

Οι τέσσερις αυτές εικόνες παρουσιάζουν τις απαραίτητες εντολές, ώστε να ξεκινήσει η εκπαίδευση του κατηγοριοποιητή, βάσει των δεδομένων εκπαίδευσης. Φαίνονται οι απαραίτητες εντολές για κάθε είδος κατηγοριοποιητή. Σε κάθε περίπτωση, στην πρώτη γραμμή γίνεται η κλήση της κατάλληλης συνάρτησης «python», για να προετοιμαστεί το αντικείμενο «classifier» και στη δεύτερη γραμμή αρχίζει η εκπαίδευση.

Αφού αναφέρθηκαν λίγες πληροφορίες για την έρευνα, είναι καλό σημείο για να δοθούν λεπτομέρειες για τα σημεία αξιολόγησης των τεχνικών δημιουργίας μοντέλων κατηγοριοποίησης. Τέσσερις είναι οι μετρικές, που χρησιμοποιήθηκαν και τα ονόματά τους είναι, «accuracy», «precision», «recall» και «f1-score».

Πρώτη, επιστράτευσα την πιο απλή μετρική, το ποσοστό επιτυχίας (accuracy). Μέσω αυτού θα φανεί ξεκάθαρα, πόσο καλά ο κατηγοριοποιητής κατηγοριοποιεί τις εγγραφές, που δεν έχει συναντήσει στην εκπαίδευσή του. Ωστόσο η μετρική αυτή ενέχει έναν μεγάλο κίνδυνο. Αν τα δεδομένα εκπαίδευσης, προέρχονται από σετ με μη ισορροπημένες κλάσεις, τότε υπάρχει η πιθανότητα να βλέπουμε υψηλές τιμές ακρίβειας, λόγω του ότι το μοντέλο προβλέπει για παράδειγμα μία κλάση συνέχεια, και επειδή τα δεδομένα της είναι περισσότερα, κατηγοριοποιεί ορθά. Ο γενικός τύπος του ποσοστού επιτυχίας φαίνεται στην επόμενη εικόνα.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Δεύτερη στη σειρά, επιλέχθηκε η ακρίβεια(precision), δηλαδή το ποσοστό των ορθών κατηγοριοποιήσεων από το μοντέλο, για την κλάση, που προβλέπει. Αυτή η μετρική, έχει μεγάλη σημασία σε δύο κυρίως σενάρια. Πρώτον, όταν η πιθανότητα μιας εγγραφής να μπει εσφαλμένα στην κατηγορία που μελετάμε, έχει μεγάλο κόστος. Για παράδειγμα, σε ιατρικές εφαρμογές, μας ενδιαφέρει να αποφύγουμε την πιθανότητα το σύστημα να αναγνωρίσει έναν υγιή άνθρωπο ως ασθενή για κάποια ασθένεια, καθώς οι λάθος αναγνωρίσεις θα οδηγήσουν σε ανθρώπους να κάνουν περιττές εξετάσεις, πιθανώς με μεγάλο οικονομικό κόστος και να φοβούνται για τη ζωή τους, ενώ θα ήταν απόλυτα υγιείς. Δεύτερον, όταν το πακέτο δεδομένων περιέχει κλάσεις με μεγάλη διαφορά στο πλήθος εγγραφών, όπου και η πρώτη μετρική (accuracy), χάνει αξία. Ο γενικός τύπος της ακρίβειας είναι ο παρακάτω.

$$Precision = \frac{TP}{TP + FP}$$

Τρίτη μετρική κατά σειρά, είναι η «recall». Εξάγοντας αυτήν, μπορούμε και έχουμε την πληροφορία για το ποσοστό των εγγραφών που κατηγοριοποιήθηκαν σε κάθε κλάση, συγκριτικά με το πόσες εγγραφές ανήκουν πραγματικά σε αυτήν την κλάση. Ο γενικός τύπος υπολογισμού του «recall» παρουσιάζεται στην επόμενη εικόνα.

$$Recall = \frac{TP}{TP + FN}$$

Το τέταρτο σημείο αξιολόγησης των τεχνικών κατηγοριοποίησης ονομάζεται «f1-score». Για να υπολογιστεί το «f1-score», υπολογίζεται ο αρμονικός μέσος της «precision» και της «recall». Είναι σημαντικό να σημειωθεί, ότι το «f1-score» μάς ενδιαφέρει στις περιπτώσεις κατηγοριοποιητών, όπου δεν μπορούμε να δώσουμε μεγαλύτερη βαρύτητα και αξία στην ακρίβεια ή στην «recall», οπότε το «f1» μάς δίνει τη δυνατότητα να τις συνδυάζουμε σε ένα σημείο σύγκρισης. Ο γενικός τύπος του «f1-score» δίνεται στην ακόλουθη εικόνα.

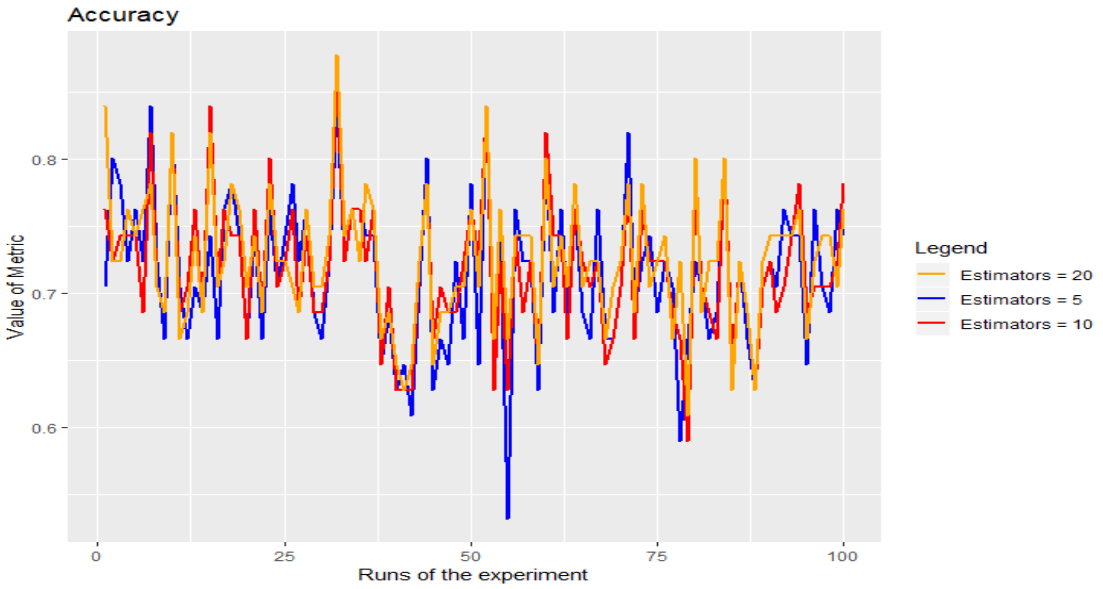
$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Οι τέσσερις αυτές μετρικές δηλώνουν ένα καλύτερο μοντέλο κατηγοριοποίησης, όσο πλησιάζουν στην τιμή ένα, ενώ μαρτυρούν ένα υποδεέστερο μοντέλο, όσο προσεγγίζουν το μηδέν.

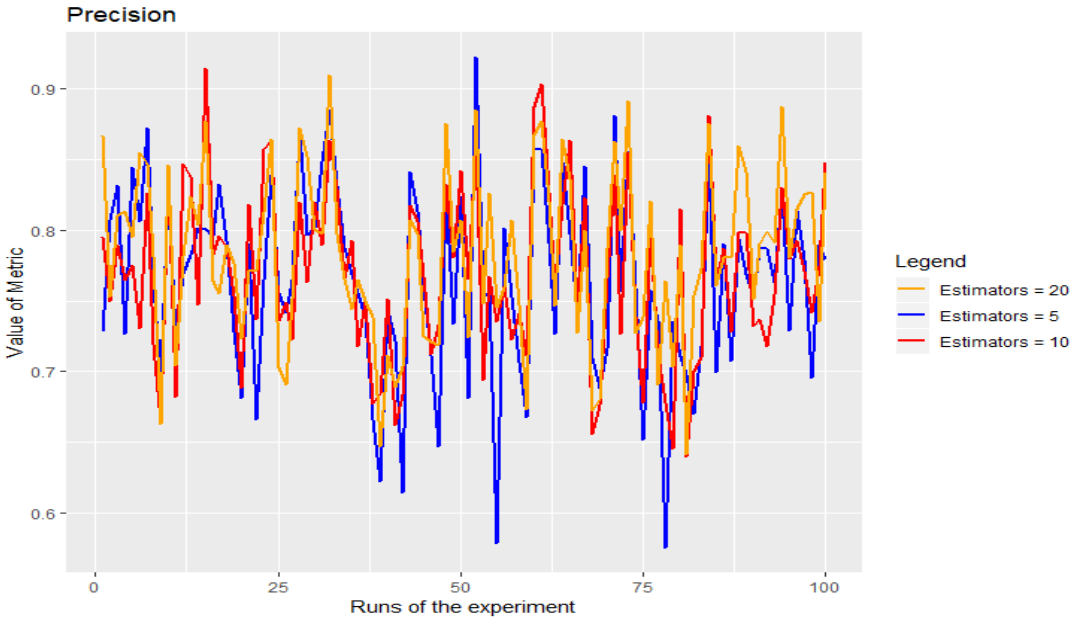
6 ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

6.1 Μελέτη πλήθους δέντρων στη μέθοδο τυχαίου δάσους

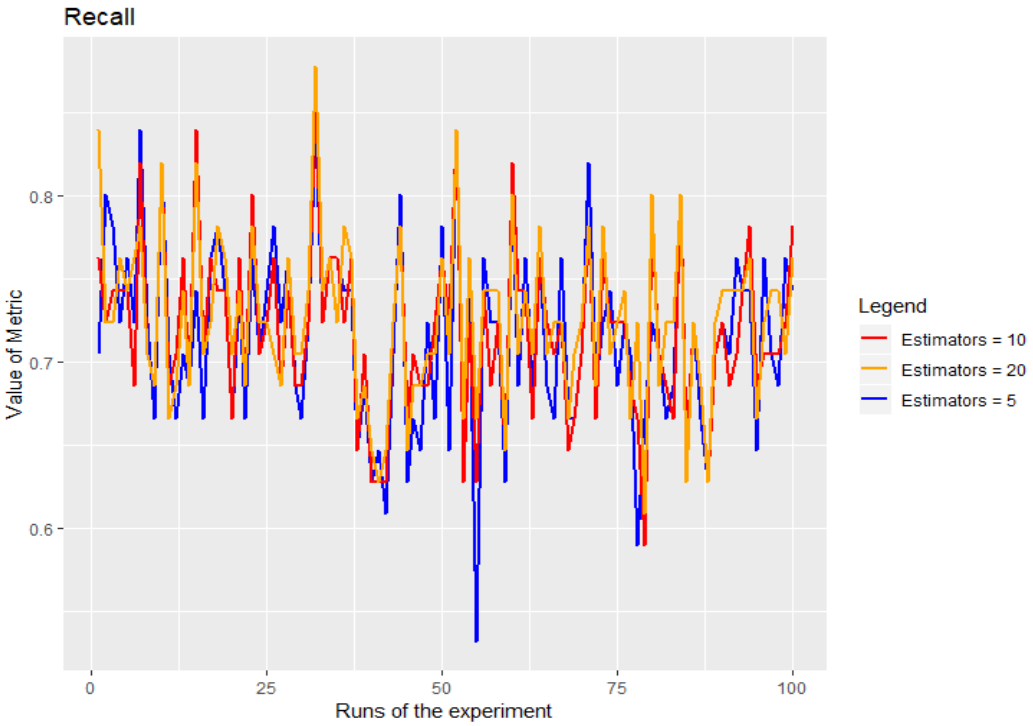
Για τη μέθοδο του τυχαίου δάσους(random forest), έπρεπε να επιλεγεί το πλήθος των δέντρων, που θα αποτελούν τον κατηγοριοποιητή. Έγιναν δοκιμές του αλγορίθμου με πέντε, δέκα και είκοσι εκτιμητές(estimators), ώστε να συγκριθούν τα αποτελέσματα των σημείων αξιολόγησής τους. Φυσικά, ο αριθμός που θα δώσει τα καλύτερα αποτελέσματα, είναι και αυτός, που θα εισαχθεί ως τιμή της παραμέτρου «n_estimators» στη συνάρτηση παραγωγής του μοντέλου. Ακολουθώντας αυτή τη μέθοδο, δίνουμε στον κατηγοριοποιητή, τη δυνατότητα να πετύχει τα βέλτιστα αποτελέσματα στις αξιολογήσεις, που θα επέλθουν. Δημιουργήθηκαν εκατό κατηγοριοποιητές για κάθε διαφορετικό πλήθος δέντρων, οι οποίοι ωστόσο είχαν σταθερή τιμή στην παράμετρο «random_state», για να πετύχουμε όσο το δυνατόν ελεγχόμενο περιβάλλον στο πείραμα. Φρόντισα, ώστε οι τεχνικές να χρησιμοποιούν σε κάθε επανάληψη εκτέλεσής τους, τον ίδιο ακριβώς διαχωρισμό των εγγραφών του πακέτου δεδομένων, ώστε οι μετρικές αξιολόγησης, να είναι συγκρίσιμες μεταξύ τους. Τα διαγράμματα, που ακολουθούν απεικονίζουν τις τιμές των μετρικών αξιολόγησης, που προέκυψαν.



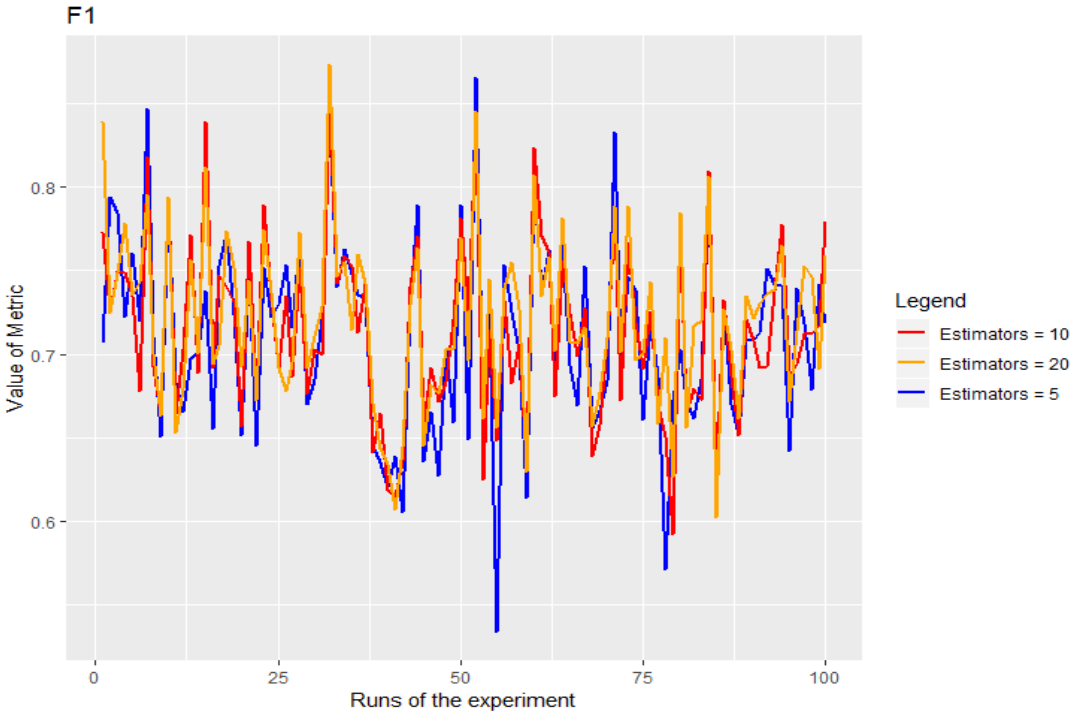
Διάγραμμα 6.1: Ποσοστό επιτυχίας κατηγοριοποίησης «Random Forest»



Διάγραμμα 6.2: Ακρίβεια κατηγοριοποίησης «Random Forest»



Διάγραμμα 6.3: «Recall» κατηγοριοποίησης «Random Forest»



Διάγραμμα 6.4: «F1-score» κατηγοριοποίησης «Random Forest»

Πίνακας 6.1: Μέσοι όροι για την τεχνική «RF», ανάλογα το πλήθος των δέντρων

RF	Estimators=5	Estimators =10	Estimators =20
Accuracy	0,714231	0,717885	0,725192
Precision	0,763645	0,768869	0,782785
Recall	0,714231	0,717885	0,725192
F1	0,70952	0,714128	0,720599

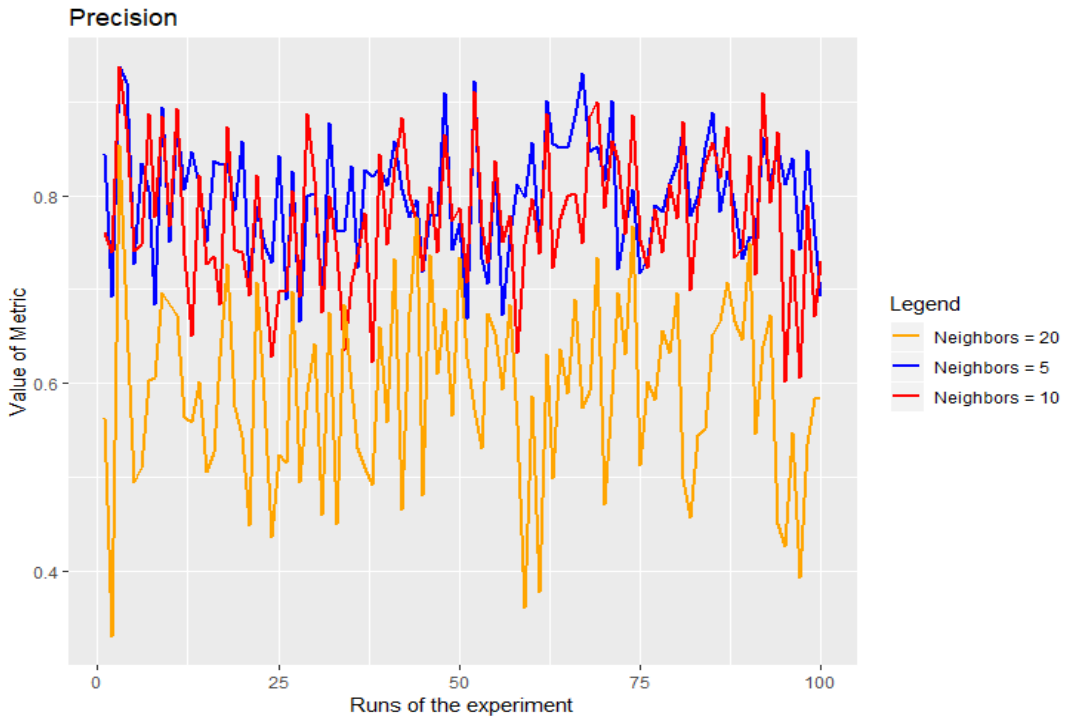
Από τα διαγράμματα δεν είναι ξεκάθαρο ποιο πλήθος έφερε τα βέλτιστα αποτελέσματα κατά τη σύγκριση, οπότε υπολόγισα τον μέσο όρο κάθε μετρικής. Το πείραμα, έδειξε πως τις υψηλότερες τιμές μετρικών, πέτυχε ο αλγόριθμος, όταν δημιούργησε είκοσι δέντρα για το δάσος κατηγοριοποίησης.

6.2 Μελέτη πλήθους γειτόνων στη μέθοδο πλησιέστερων γειτόνων

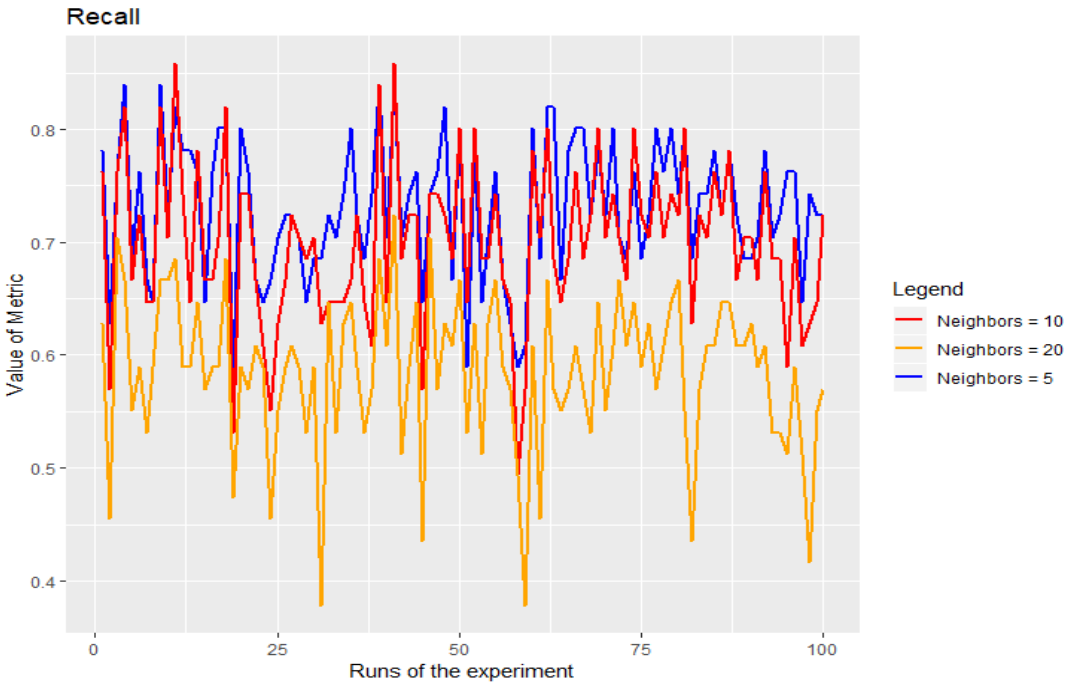
Για τη μέθοδο των πλησιέστερων γειτόνων έπρεπε να αποφασιστεί, το πλήθος των γειτόνων, που θα λαμβάνει υπόψιν το σύστημα για να επιλέγει την κατηγορία, που θα τοποθετεί τις εγγραφές ελέγχου. Όμοια με τη μέθοδο τυχαίου δάσους, επιλέχθηκαν να συγκριθούν οι περιπτώσεις με πέντε, δέκα και είκοσι γείτονες να λαμβάνονται υπόψιν από τον κατηγοριοποιητή. Ο αριθμός, που θα παρουσιάσει τα θετικότερα αποτελέσματα στις μετρικές, θα τοποθετηθεί ως τιμή, της μεταβλητής «n_neighbors» στη συνάρτηση, που δημιουργεί το μοντέλο κατηγοριοποίησης. Οι υπόλοιπες παράμετροι διατηρούν την ίδια τιμή και κάθε κλήση της συνάρτησης δημιουργίας

μοντέλου αξιοποιεί τις ίδιες εγγραφές εκπαίδευσης από το σύνολο δεδομένων. Τα αποτελέσματα φαίνονται στα διαγράμματα, που ακολουθούν:

Διάγραμμα 6.5: Ποσοστό επιτυχίας κατηγοριοποίησης «K Nearest Neighbors»



Διάγραμμα 6.6: Ακρίβεια κατηγοριοποίησης «K Nearest Neighbors»



Διάγραμμα 6.7: «Recall» κατηγοριοποίησης «K Nearest Neighbors»

Διάγραμμα 6.8: «F1-score» κατηγοριοποίησης «K Nearest Neighbors»

Πίνακας 6.2: Μέσοι όροι για την τεχνική «KNN», ανάλογα το πλήθος των γειτόνων

KNN	Neighbors=5	Neighbors=10	Neighbors=20
Accuracy	0,731346	0,700192	0,585192
Precision	0,80005	0,77574	0,591342
Recall	0,731346	0,700192	0,585192
F1	0,722857	0,688123	0,549496

Από τα διαγράμματα, γίνεται ξεκάθαρο, ότι τα μοντέλα με τους είκοσι γείτονες, αποφασίζουν τις κατηγορίες των εγγραφών ελέγχου με μεγαλύτερο βαθμό αποτυχίας, από τα μοντέλα με πέντε και δέκα γείτονες. Φαίνεται, πως δεν υπάρχει ούτε ένα σημείο αξιολόγησης, στο οποίο να υπερτερεί η περίπτωση, που είκοσι γειτονικές εγγραφές συναποφασίζουν την κατηγορία της εγγραφής ελέγχου. Το γεγονός αυτό, μας επιτρέπει να συμπεράνουμε, ότι όταν έχουμε πολλούς γείτονες να «συναποφασίζουν», το μοντέλο υποπέφτει στο σφάλμα, να συνυπολογίζει εγγραφές άλλων γειτονιών, στις αποφάσεις του.

Ανάμεσα στις τιμές, πέντε και δέκα γειτόνων, καλύτερα αποτελέσματα μοιάζει να έχει η περίπτωση των πέντε γειτόνων. Αυτό επιβεβαιώνεται από τους μέσους όρους των μετρικών, καθώς όταν το μοντέλο έχει εκπαιδευτεί με τις πέντε πλησιέστερες εγγραφές, έχει μεγαλύτερες τιμές σε όλο το φάσμα των μετρικών από το μοντέλο με τις δέκα εγγραφές.

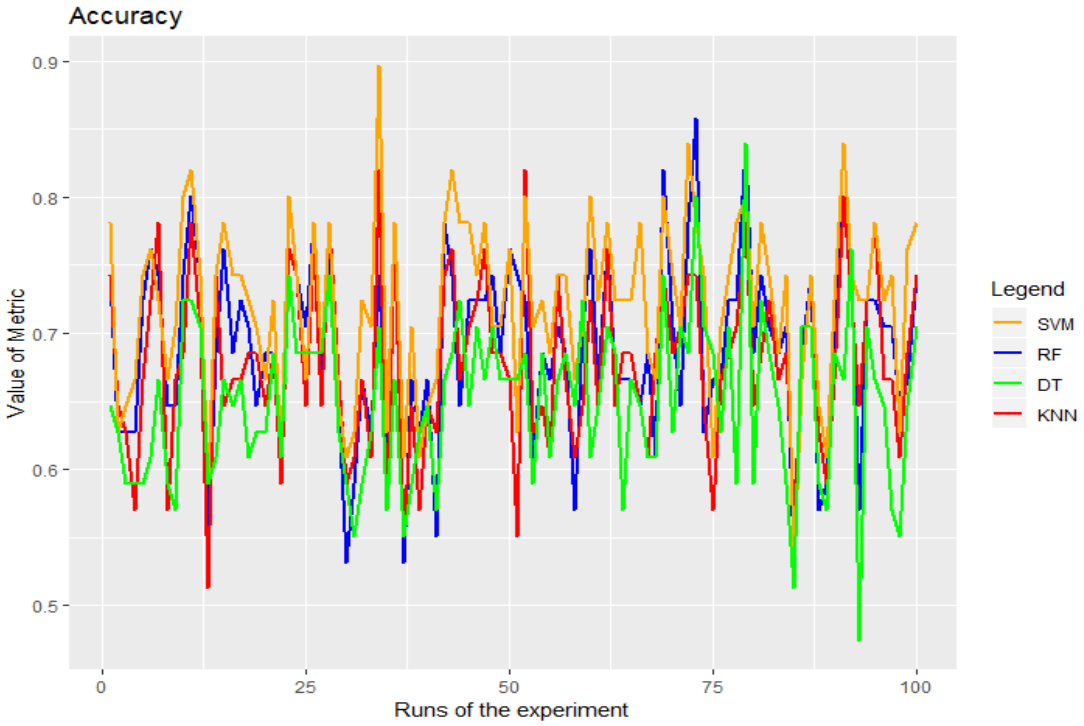
6.3 Μελέτη μεγέθους χαρακτηριστικών

Θεωρώντας τα αποτελέσματα των δύο προηγούμενων ως βάση για τα επόμενα, εξασφαλίζω, ότι το μοντέλο του τυχαίου δάσους, θα περιέχει είκοσι δέντρα και το μοντέλο των πλησιέστερων γειτόνων, θα έχει πέντε γείτονες να συναποφασίζουν κάθε κατηγοριοποίηση. Μένει στα ακόλουθα πειράματα να εντοπίσω το βέλτιστο μέγεθος χαρακτηριστικών(features), δηλαδή το μέγεθος των τμημάτων στα οποία θα χωρίσω τις γενετικές ακολουθίες, ώστε οι μέθοδοι να έχουν ξανά τη μέγιστη δυνατή επιτυχία στις κατηγοριοποιήσεις τους.

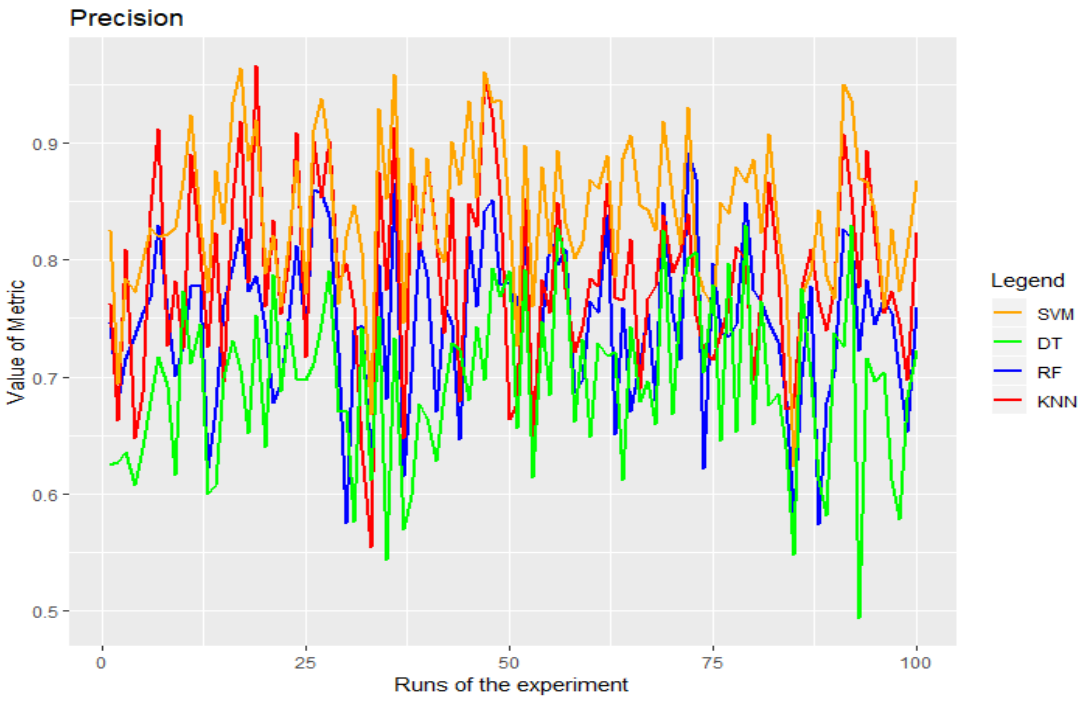
Υπολογίστηκαν οι μετρικές «accuracy», «precision», «recall» και «f1», για κάθε τεχνική με μεγέθη «features» από τέσσερα, έως οκτώ, συν μία ακραία περίπτωση το μέγεθος είκοσι. Τα διάφορα «features», που θα υπολογιστούν για τη μελέτη, αποτελούν υποακολουθίες, μεγαλύτερων ακολουθιών. Οι υποακολουθίες στον χώρο της βιοπληροφορικής, είναι γνωστές με το όνομα «k-mers», όπου το «k» συμβολίζει το πλήθος συμβόλων, που περιέχει κάθε υποακολουθία.

Σύμφωνα, με τα παραπάνω εγώ ασχολήθηκα με τα «4-mers», «5-mers», «6-mers», «7-mers», «8-mers» και «20-mers».

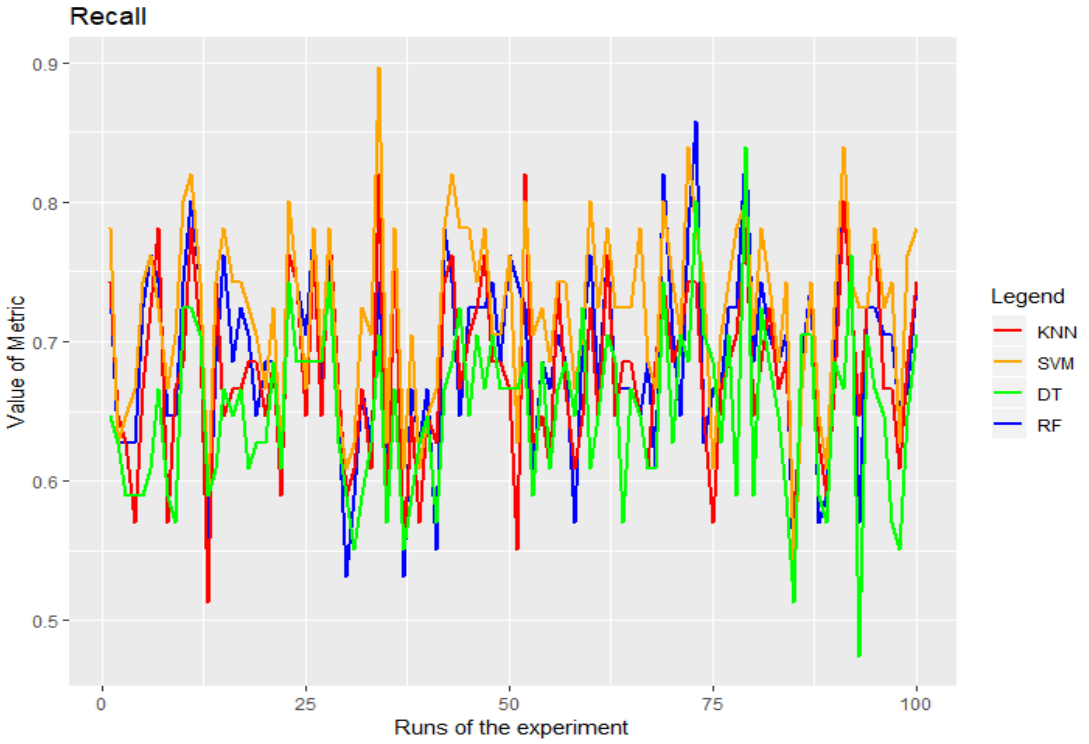
6.3.1 4-mers



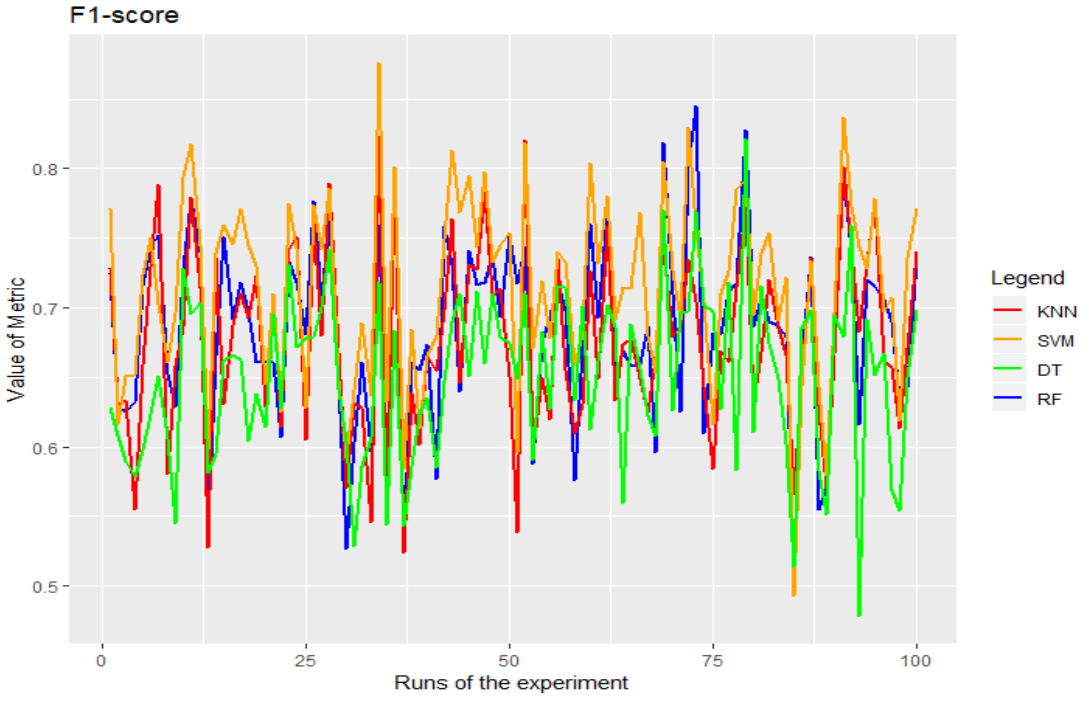
Διάγραμμα 6.9: Ποσοστά επιτυχίας κατηγοριοποίησης για μέγεθος 4 συμβόλων



Διάγραμμα 6.10: Ακρίβεια κατηγοριοποίησης για μέγεθος 4 συμβόλων



Διάγραμμα 6.11: «Recall» κατηγοριοποίησης για μέγεθος 4 συμβόλων



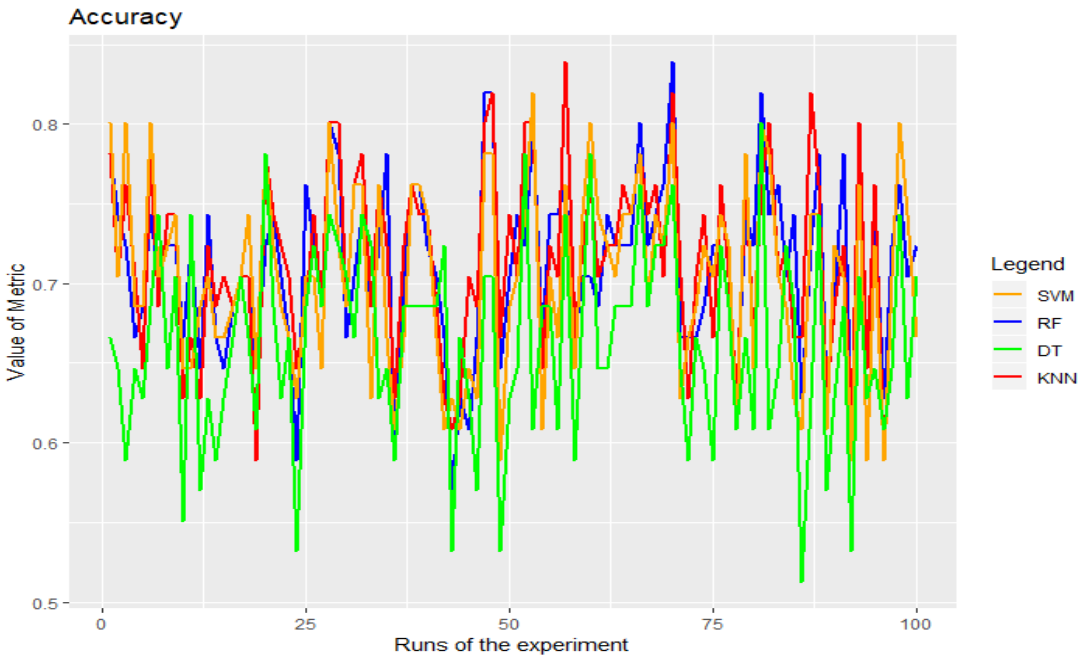
Διάγραμμα 6.12: «F1-score» κατηγοριοποίησης για μέγεθος 4 συμβόλων

Πίνακας 6.3: Μέσοι όροι με 4 σύμβολα ως μέγεθος χαρακτηριστικών, για κάθε τεχνική

4-mers	RF	KNN	SVM	DT
Accuracy	0,69	0,677692	0,721731	0,650385
Precision	0,746551	0,787066	0,840981	0,69495
Recall	0,69	0,677692	0,721731	0,650385
F1	0,685526	0,676324	0,71493	0,650251

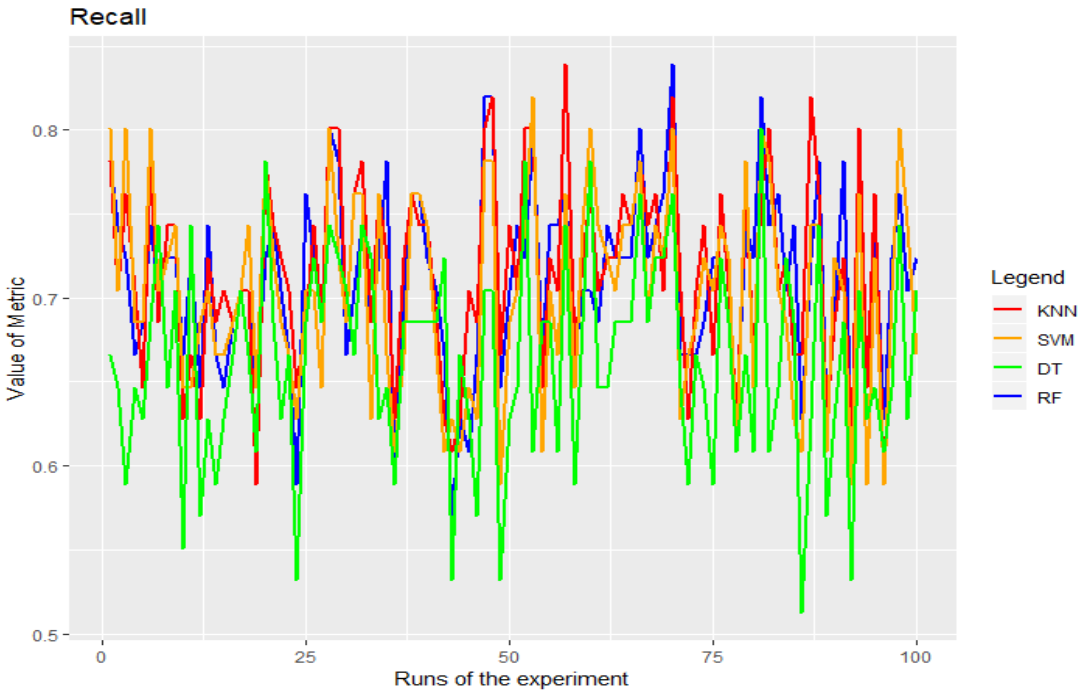
Στην περίπτωση αυτή των «4-mers», όπου είναι και το μικρότερο μέγεθος χαρακτηριστικού που χρησιμοποίησα, φαίνεται από τα διαγράμματα, ότι ο αλγόριθμος των μηχανών διανυσμάτων υποστήριξης ευνοείται από το μικρό μέγεθος χαρακτηριστικού. Η παρατήρηση αυτή, επιβεβαιώνεται από τους μέσους όρους των μετρικών, καθώς ο αλγόριθμος αυτός παρουσιάζει υψηλότερες τιμές σε όλες τις μετρικές, έναντι των υπολοίπων αλγορίθμων. Ακολουθεί το τυχαίο δάσος, οι πλησιέστεροι γείτονες και τελευταίο έρχεται το δέντρο απόφασης. Είναι σημαντικό να τονίσουμε, πως στην μετρική τής ακρίβειας, οι πλησιέστεροι γείτονες, έχουν καλύτερα αποτελέσματα από το τυχαίο δάσος, οπότε ανάλογα τη μετρική, που μας ενδιαφέρει περισσότερο, μπορούμε να υποστηρίξουμε, ότι η τεχνική των πλησιέστερων γειτόνων είναι καλύτερη με χαρακτηριστικά μεγέθους τεσσάρων χαρακτήρων, από την τεχνική του τυχαίου δάσους. Εγώ θεωρώ, το αντίθετο, καθώς τρεις από τις τέσσερεις μετρικές έχουν υψηλότερη τιμή για το τυχαίο δάσος, σε σχέση με τους πλησιέστερους γείτονες.

6.3.2 5-mers

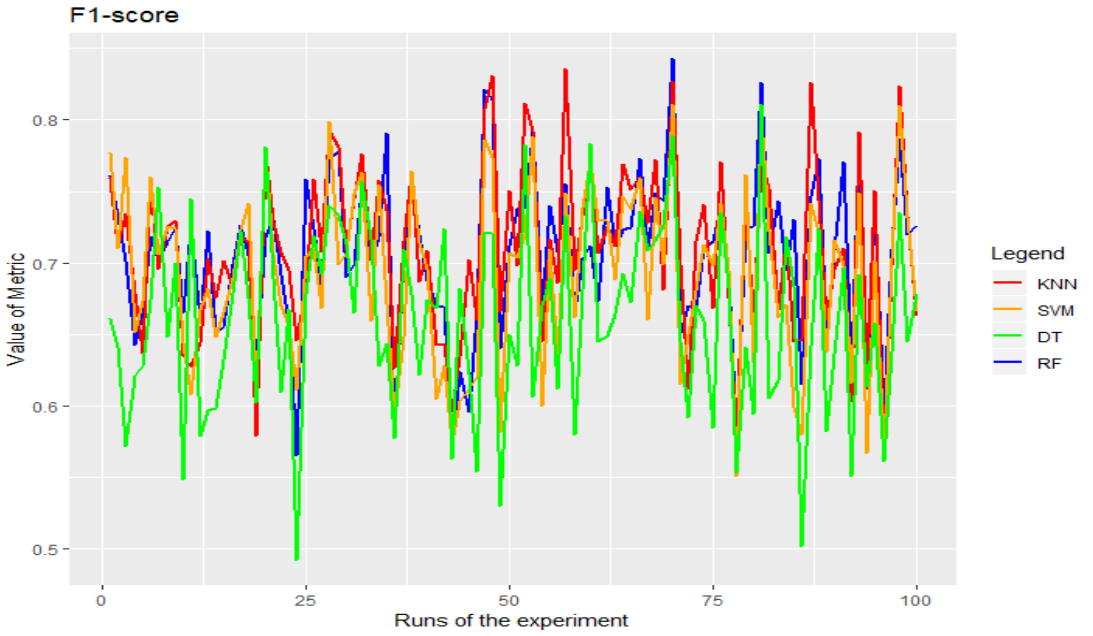


Διάγραμμα 6.13: Ποσοστά επιτυχίας κατηγοριοποίησης για μέγεθος 5 συμβόλων

Διάγραμμα 6.14: Ακρίβεια κατηγοριοποίησης για μέγεθος 5 συμβόλων



Διάγραμμα 6.15: «Recall» κατηγοριοποίησης για μέγεθος 5 συμβόλων



Διάγραμμα 6.16: «F1-score» κατηγοριοποίησης για μέγεθος 5 συμβόλων

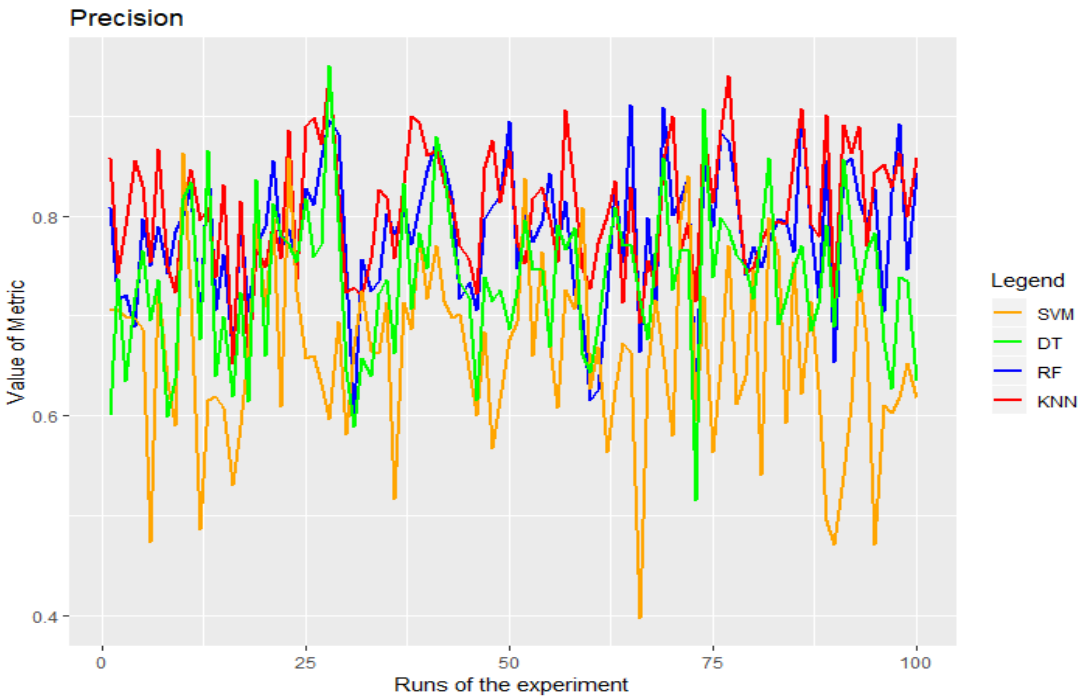
Πίνακας 6.4: Μέσοι όροι με 5 σύμβολα ως μέγεθος χαρακτηριστικών, για κάθε τεχνική

5-mers	RF	KNN	SVM	DT
Accuracy	0,711731	0,716346	0,703462	0,663077
Precision	0,766741	0,811113	0,808115	0,707777
Recall	0,711731	0,716346	0,703462	0,663077
F1	0,706287	0,710229	0,691741	0,660084

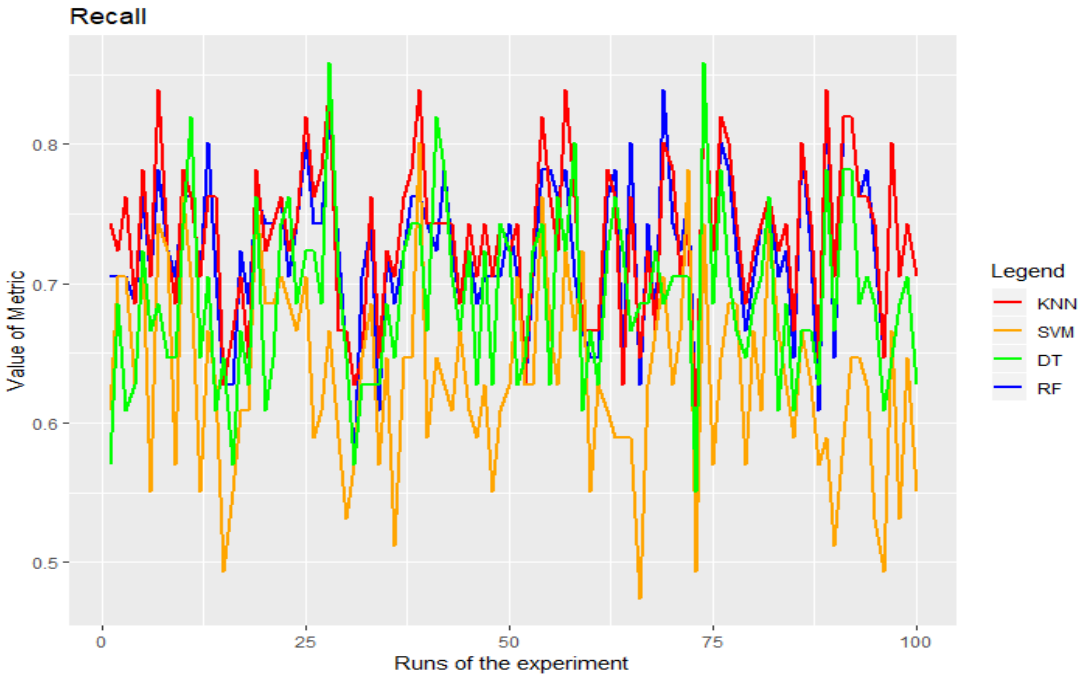
Στην περίπτωση των χαρακτηριστικών μήκους 5 χαρακτήρων, δεν είναι ξεκάθαρο από τα διαγράμματα, ποιος αλγόριθμος παρουσιάζει τα βέλτιστα αποτελέσματα στην κατηγοριοποίηση των εγγραφών ελέγχου. Καταφεύγοντας στον πίνακα με τον μέσο όρο των μετρικών για κάθε αλγόριθμο, έπειτα από εκατό εκτελέσεις του κώδικα, αντιλαμβανόμαστε, γιατί τα διαγράμματα δεν είναι ξεκάθαρα. Αν εξαιρέσουμε το δέντρο απόφασης, ως την τεχνική με την χειρότερη αξιολόγηση, τότε οι αλγόριθμοι, που απομένουν δεν παρουσιάζουν σημαντικές διαφορές. Οι διαφορές στην αξιολόγησή τους, ωστόσο αν και μη σημαντικές, είναι μετρήσιμες, για αυτό και παρουσιάζονται στον πίνακα των μέσων όρων. Ως συμπέρασμα, μπορούμε να βγάλουμε, ότι ο αλγόριθμος πλησιέστερων γειτόνων και μάλιστα με πλήθος γειτόνων, που προέκυψε από τα προηγηθέντα πειράματα(5), είναι ο βέλτιστος για μέγεθος πέντε συμβόλων στα χαρακτηριστικά.

6.3.3 6-mers

Διάγραμμα 6.17: Ποσοστό επιτυχίας κατηγοριοποίησης για μέγεθος 6 συμβόλων



Διάγραμμα 6.18: Ακρίβεια κατηγοριοποίησης για μέγεθος 6 συμβόλων



Διάγραμμα 6.19: «Recall» κατηγοριοποίησης για μέγεθος 6 συμβόλων

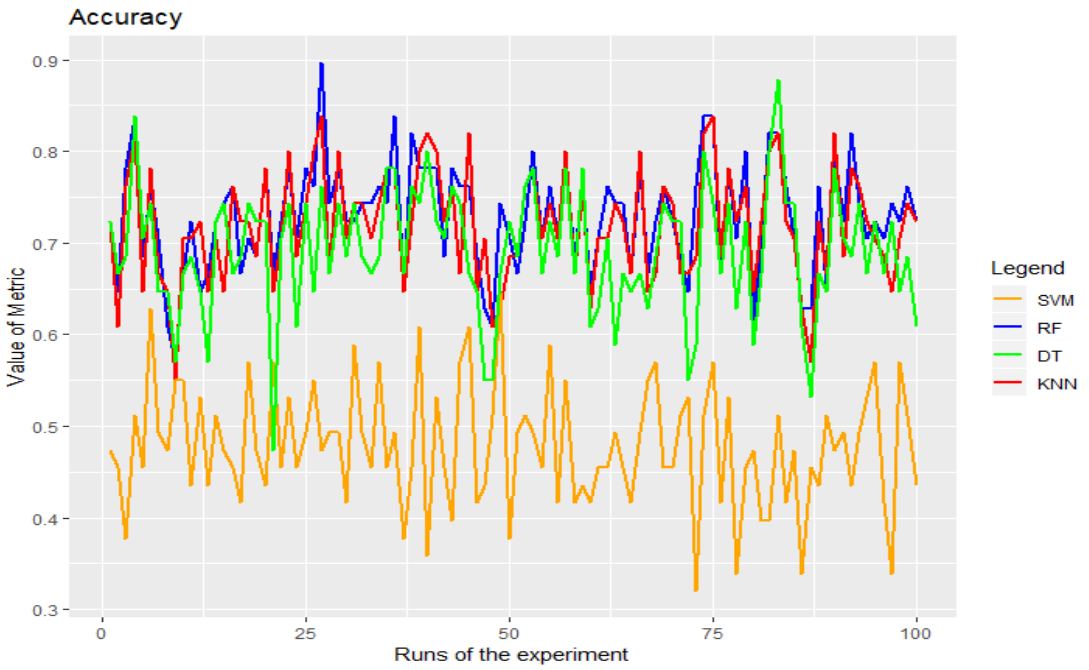
Διάγραμμα 6.20: «F1-score» κατηγοριοποίησης για μέγεθος 6 συμβόλων

Πίνακας 6.5: Μέσοι όροι με 6 σύμβολα ως μέγεθος χαρακτηριστικών, για κάθε τεχνική

6-mers	RF	KNN	SVM	DT
Accuracy	0,724615	0,734038	0,633077	0,688462
Precision	0,782607	0,807035	0,661939	0,736108
Recall	0,724615	0,734038	0,633077	0,688462
F1	0,721736	0,727344	0,591268	0,689033

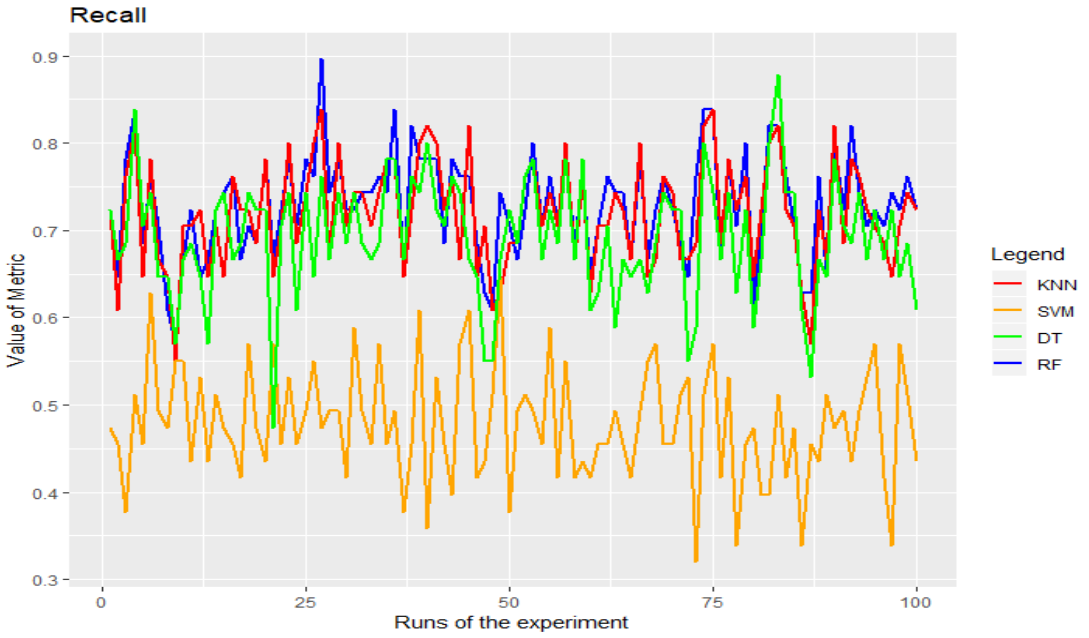
Αυτό το μέγεθος «feature», όπως φαίνεται τόσο από τα διαγράμματα, όσο και από τον πίνακα μέσων όρων, είναι αρκετά μεγάλο, ώστε η τεχνική «SVM», να προσφέρει μειωμένη αποτελεσματικότητα στην κατηγοριοποίηση, σε σχέση με τις άλλες τεχνικές. Αυτό το φαινόμενο θα επαναληφθεί και θα ενταθεί όσο ελέγχουμε μεγαλύτερα μεγέθη χαρακτηριστικών. Οι άλλες τεχνικές δείχνουν και αυτές μια πτώση στις τιμές των μετρικών αξιολόγησής τους. Το σημαντικό γεγονός εδώ, είναι ότι από τον πίνακα φαίνεται να ξεχωρίζει η τεχνική του «knn» και να βγάζει τα θετικότερα αποτελέσματα. Τέλος, το δέντρο απόφασης παρουσιάζει βελτίωση στην αξιολόγησή του, όσο αυξάνεται το μέγεθος των «k-mers», ωστόσο η διαφορά του από τις κυρίαρχες τεχνικές είναι μετρήσιμη και παραμένει υπαρκτή σε κάθε επιλογή «k-mers» μέχρι τώρα.

6.3.4 7-mers

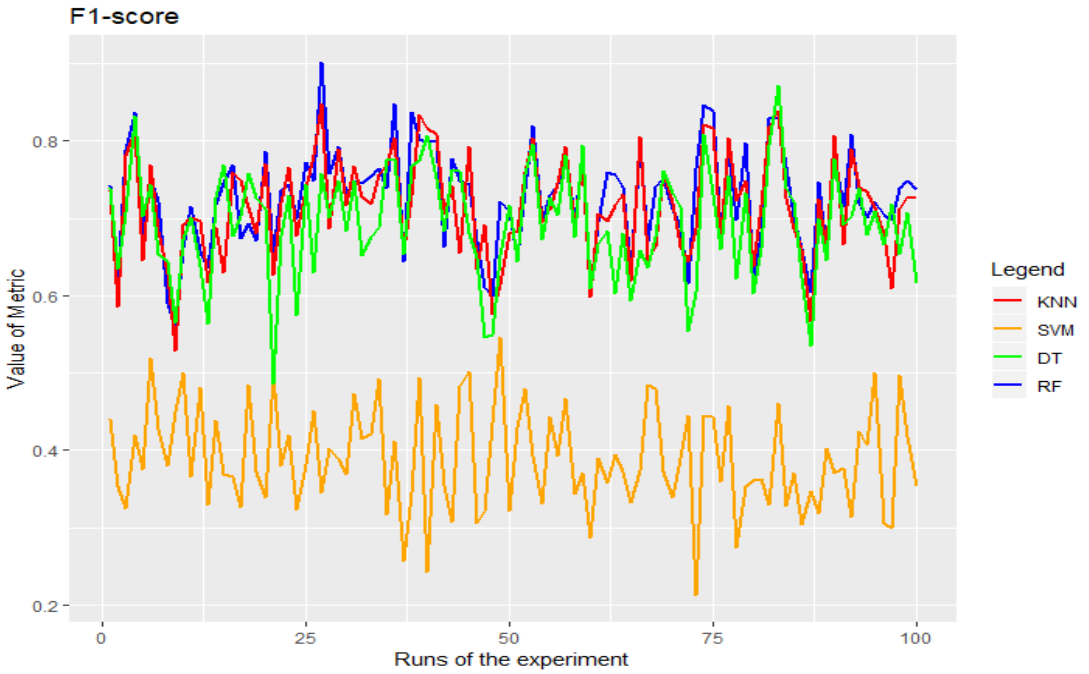


Διάγραμμα 6.21: Ποσοστό επιτυχίας κατηγοριοποίησης για μέγεθος 7 συμβόλων

Διάγραμμα 6.22: Ακρίβεια κατηγοριοποίησης για μέγεθος 7 συμβόλων



Διάγραμμα 6.23: «Recall» κατηγοριοποίησης για μέγεθος 7 συμβόλων



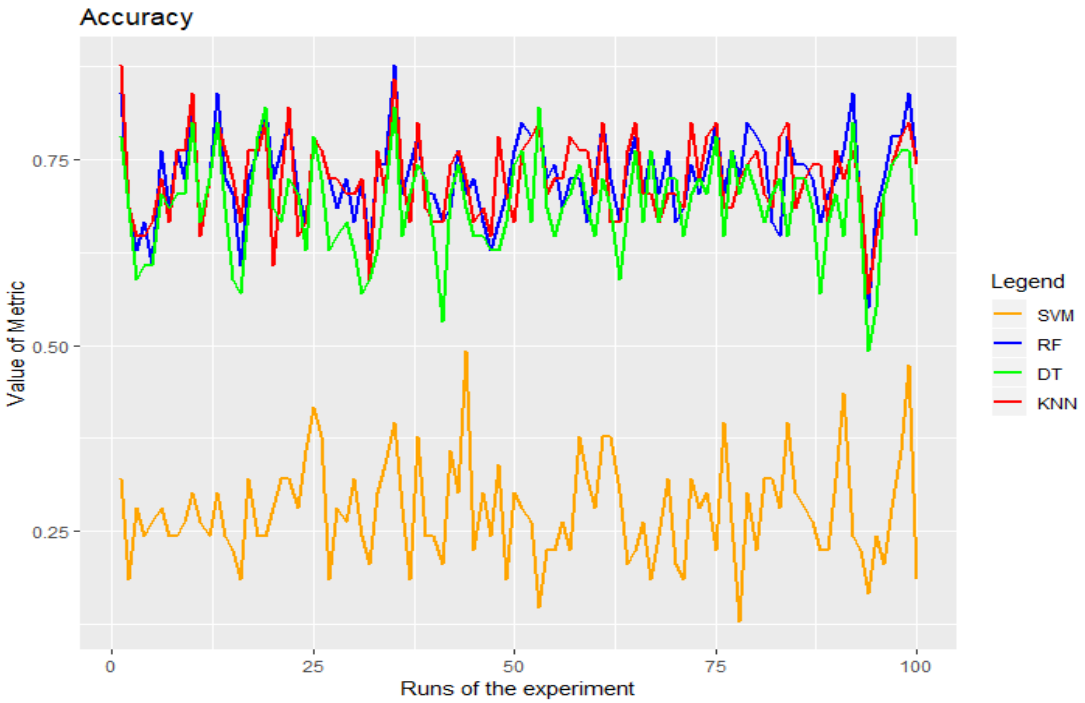
Διάγραμμα 6.24: «F1-score» κατηγοριοποίησης για μέγεθος 7 συμβόλων

Πίνακας 6.6: Μέσοι όροι με 7 σύμβολα ως μέγεθος χαρακτηριστικών, για κάθε τεχνική

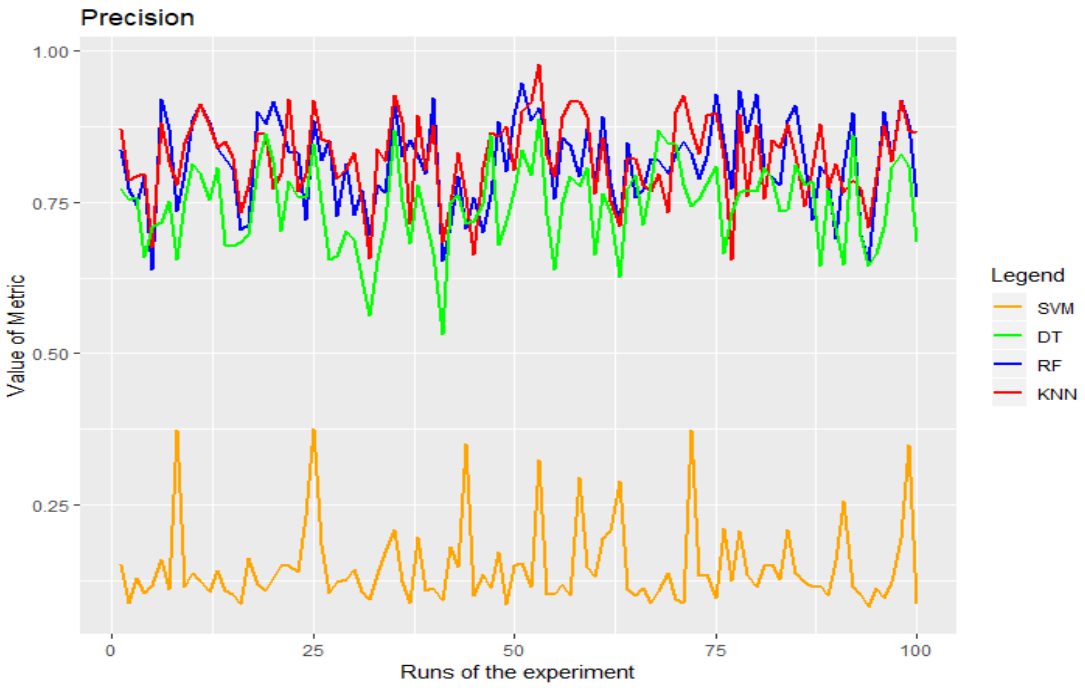
7-mers	RF	KNN	SVM	DT
Accuracy	0,728846	0,719615	0,478846	0,691346
Precision	0,794376	0,792167	0,374094	0,744084
Recall	0,728846	0,719615	0,478846	0,691346
F1	0,726386	0,715492	0,389629	0,692729

Σε χαρακτηριστικά μεγέθους επτά συμβόλων, όπως προείπαμε, ο «Support Vector Machine» αλγόριθμος δείχνει να υστερεί σε σχέση με τους τρεις άλλους. Η αξιολόγηση του «Decision Tree» συνεχίζει να βελτιώνεται σε κάθε ένα από τα σημεία αξιολόγησης, αν και η βελτίωση που παρατηρείται είναι μηδαμινή. Σε καμία περίπτωση δεν είναι αρκετή, ώστε να καλύψει τη διαφορά με τους «Random Forest» και «K Nearest Neighbors» και να παρουσιάζει μεγαλύτερη αποτελεσματικότητα στην κατηγοριοποίηση. Όμοια συμπεριφορά παρουσιάζει και ο «Random Forest», ο οποίος δείχνει να ευνοείται από την αύξηση του μεγέθους των «features». Αντίθετη απόκριση σε αυτή την αλλαγή φαίνεται να έχει ο «K Nearest Neighbors», ο οποίος χάνει μικρό μέρος της απόδοσής του σε κάθε μία από τις μετρικές, ωστόσο η διαφορά θα μπορούσε να χαρακτηριστεί αμελητέα.

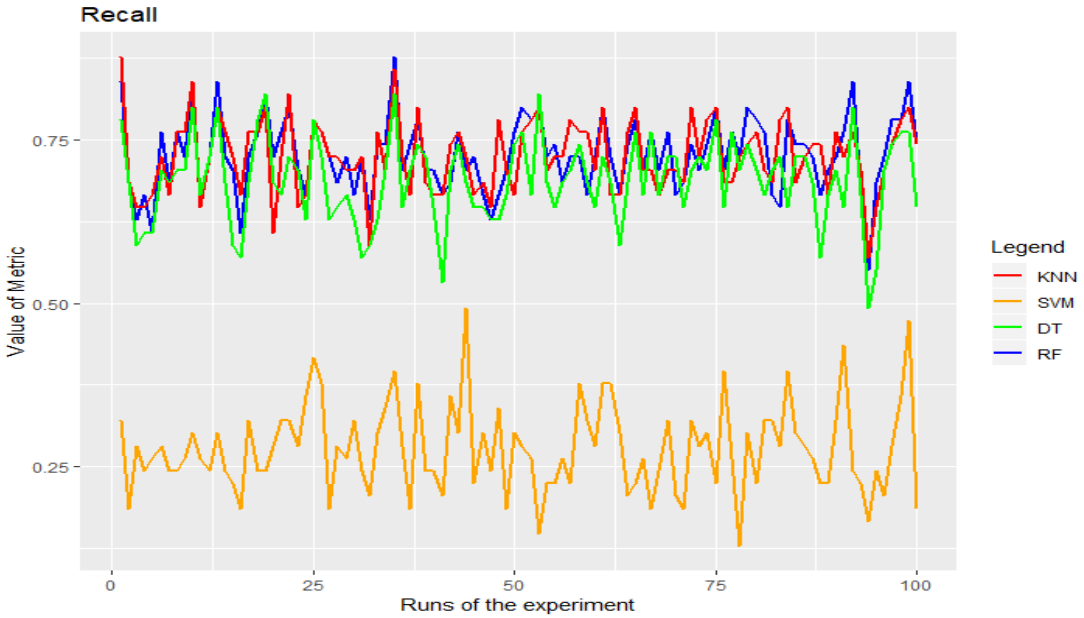
6.3.5 8-mers



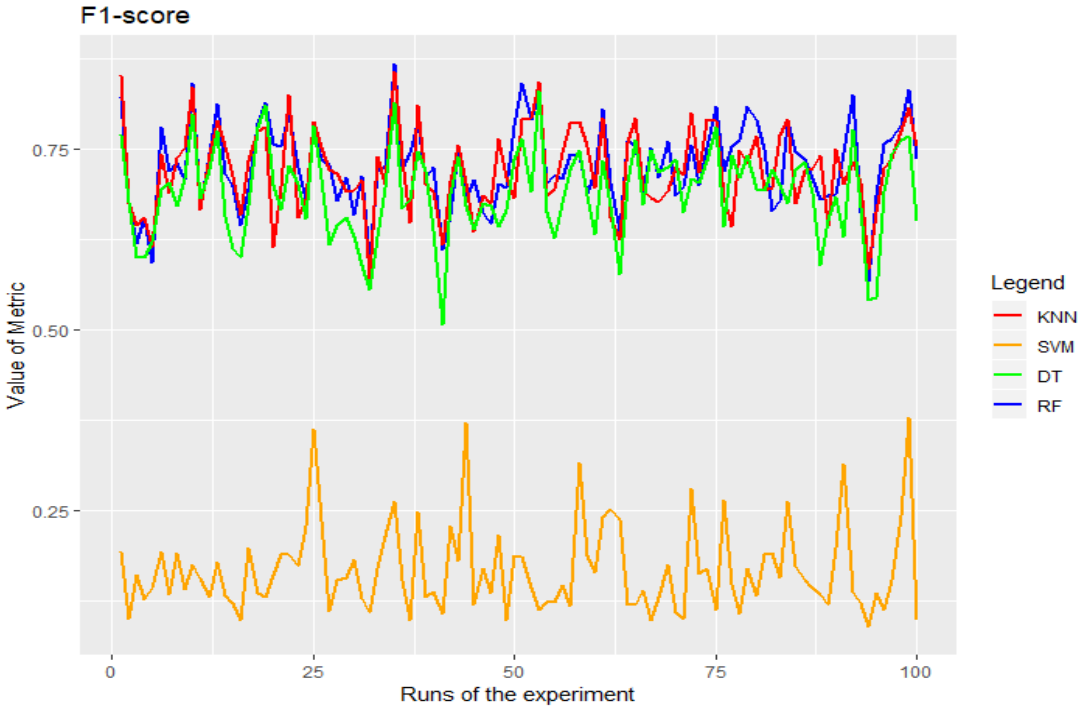
Διάγραμμα 6.25: Ποσοστό επιτυχίας κατηγοριοποίησης για μέγεθος 8 συμβόλων



Διάγραμμα 6.26: Ακρίβεια κατηγοριοποίησης για μέγεθος 8 συμβόλων



Διάγραμμα 6.27: «Recall» κατηγοριοποίησης για μέγεθος 8 συμβόλων



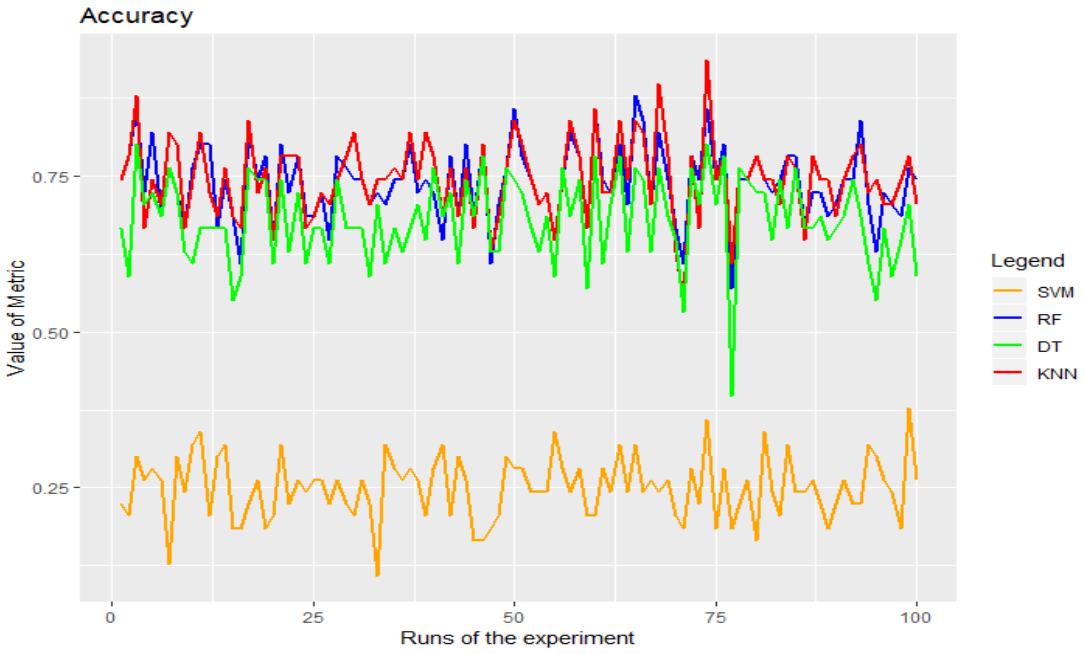
Διάγραμμα 6.28: «F1-score» κατηγοριοποίησης για μέγεθος 8 συμβόλων

Πίνακας 6.7: Μέσοι όροι με 8 σύμβολα ως μέγεθος χαρακτηριστικών, για κάθε τεχνική

8-mers	RF	KNN	SVM	DT
Accuracy	0,7275	0,725769	0,277308	0,688269
Precision	0,814846	0,823417	0,147573	0,744314
Recall	0,7275	0,725769	0,277308	0,688269
F1	0,727393	0,722182	0,168717	0,688891

Σε αυτό το πείραμα οι μετρικές αξιολόγησης του «SVM» μειώνονται κατά πολύ από το προηγούμενο. Δε δίνει ιδιαίτερο θάρρος για να υπάρξουν και επόμενες δοκιμές με άλλα μεγέθη «features», ωστόσο θα έχει ενδιαφέρον να δούμε πως θα ανταποκριθεί στην περίπτωση των «20-mers», που απέχουν πολύ σε πλήθος συμβόλων από τα «8-mers» αυτού του πειράματος. Οι τεχνικές «RF» και «DT» μειώνουν τις τιμές των μετρικών τους, όχι όμως τόσο δραματικά όσο ο «SVM». Αντίθετα ο «KNN» αυξάνει τις τιμές των σημείων αξιολόγησής του. Όπως γίνεται εύκολα αντιληπτό είναι και ο μόνος, που σημειώνει βελτίωση, αν και δεν αγγίζει τις μέγιστες τιμές του. Αυτό το έχει πετύχει πιο πριν, με τα «6-mers».

6.3.6 20-mers



Διάγραμμα 6.29: Ποσοστό επιτυχίας κατηγοριοποίησης για μέγεθος 20 συμβόλων

Διάγραμμα 6.30: Ακρίβεια κατηγοριοποίησης για μέγεθος 20 συμβόλων

Διάγραμμα 6.31: «Recall» κατηγοριοποίησης για μέγεθος 20 συμβόλων

Διάγραμμα 6.32: «F1-score» κατηγοριοποίησης για μέγεθος 20 συμβόλων

Πίνακας 6.8: Μέσοι όροι με 20 σύμβολα ως μέγεθος χαρακτηριστικών, για κάθε τεχνική

20-mers	RF	KNN	SVM	DT
Accuracy	0,738654	0,745385	0,25	0,679423
Precision	0,859477	0,854295	0,104904	0,811595
Recall	0,738654	0,745385	0,25	0,679423
F1	0,745145	0,746311	0,127983	0,693435

Ως τελευταία περίπτωση «k-mers», αποφασίστηκε να ελέγξουμε, πως θα αντιδράσουν οι αλγόριθμοι με ένα πολύ μεγαλύτερο μέγεθος «features», σε σχέση με όσα είχαν ήδη εξεταστεί. Έτσι παρατηρείται ακόμη μία μείωση στην αξιολόγηση του «SVM». Ο «DT» παρουσιάζει αξιοσημείωτες αλλαγές, καθώς το ποσοστό επιτυχίας του(accuracy) και η «recall» σημειώνουν μικρή πτώση, ενώ η ακρίβεια(precision) και το «f1-score» σημειώνουν σημαντική και μικρή αύξηση, αντίστοιχα. Οι «RF» και «KNN», παρουσίασαν καλύτερες τιμές σε όλες τις μετρικές αξιολόγησης. Σε αυτό το πείραμα δημιουργήθηκαν κατηγοριοποιητές με τις θετικότερες κατηροποιήσεις, έως τώρα. Είναι σημαντικό, να αναφέρουμε, πως αυτή η περίπτωση εξετάστηκε καθαρά για ακαδημαϊκούς λόγους και για αυτό, ανεξάρτητα από τα αποτελέσματά της, δεν πρόκειται να επιλεγεί μέγεθος χαρακτηριστικών ίσο με είκοσι, για καμία από τις τεχνικές κατηγοριοποίησης.

6.4 ΠΑΡΑΤΗΡΗΣΕΙΣ ΑΠΟ ΤΑ ΠΕΙΡΑΜΑΤΑ ΓΙΑ ΤΑ ΒΕΛΤΙΣΤΑ K-MERS

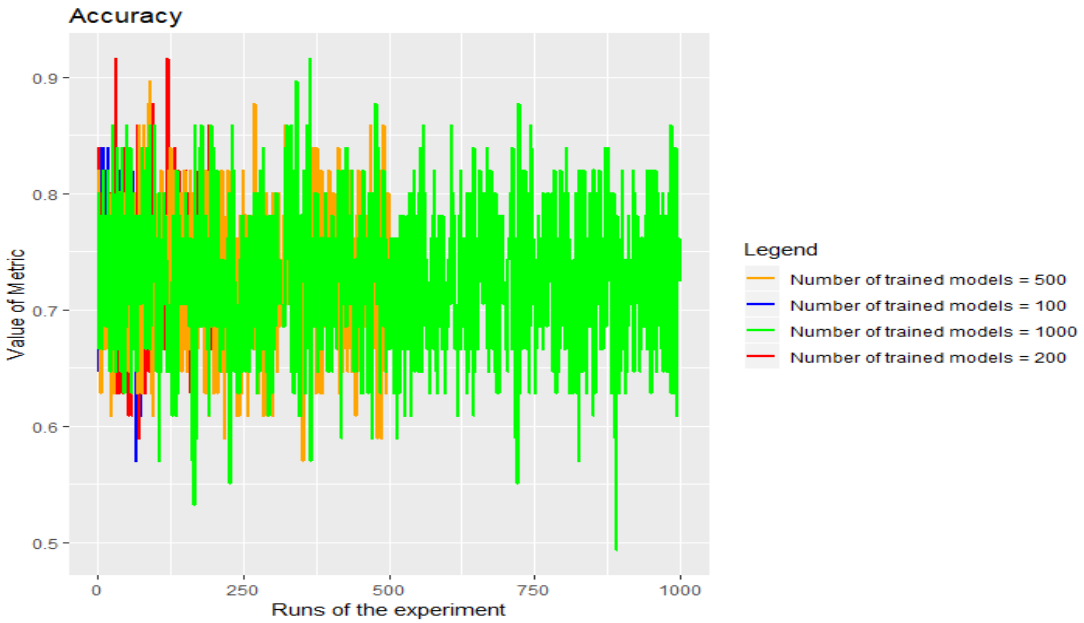
Αυτό, που καταφέραμε με τα προηγηθέντα πειράματα, είναι να επιλέξουμε τις βέλτιστες τιμές, ανάμεσα σε όσες δοκιμάσαμε, για τις παραμέτρους, που μελετήσαμε. Έτσι θα επιτύχουμε ο κάθε αλγόριθμος να δώσει τα θετικότερα αποτελέσματα, που μπορεί στα τελικά πειράματα, που θα ακολουθήσουν. Για το πλήθος των δέντρων στη μέθοδο τυχαίου δάσους, καταλήξαμε στον αριθμό είκοσι και για το πλήθος των γειτόνων στη μέθοδο πλησιέστερων γειτόνων, καταλήξαμε στον αριθμό πέντε. Για το μέγεθος των «k-mers» καταλήξαμε στη μέθοδο τυχαίου δάσους στα «7-mers», στη μέθοδο πλησιέστερου γείτονα στα «6-mers», στη μέθοδο μηχανών διανυσμάτων υποστήριξης στα «4-mers» και στη μέθοδο δέντρου απόφασης στα «7-mers». Πρέπει να αναφερθεί, πως για τη μέθοδο τυχαίου -δάσους, τα «7-mers» και τα «8-mers» είχαν σε δύο μετρικές το καθένα τη μεγαλύτερη τιμή, οπότε επέλεξα το μικρότερο μέγεθος «feature», γιατί θα δώσει πιο πολλά σε πλήθος «features» κάτι, που πιθανό να παίξει ρόλο στις πολλαπλές εκτελέσεις των μοντέλων στις τελευταίες δοκιμές. Επίσης, όπως δηλώθηκε παραπάνω, τα «20-mers» εξετάστηκαν καθαρά για ακαδημαϊκούς λόγους και δεν θα χρησιμοποιηθούν παρακάτω παρά τα ιδιαίτερα θετικά αποτελέσματα τους για τις δύο από τις τέσσερις τεχνικές.

Μία σημαντική παρατήρηση, που μπορούμε να κάνουμε είναι, ότι οι μετρικές των αλγορίθμων «Random Forest», «K Nearest Neighbors» και «Decision Tree» φαίνεται να παίρνουν τιμές, που ταλαντεύονται γύρω από κάποιες τιμές-«κέντρα».

6.5 ΜΕΛΕΤΗ ΤΩΝ ΤΕΛΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

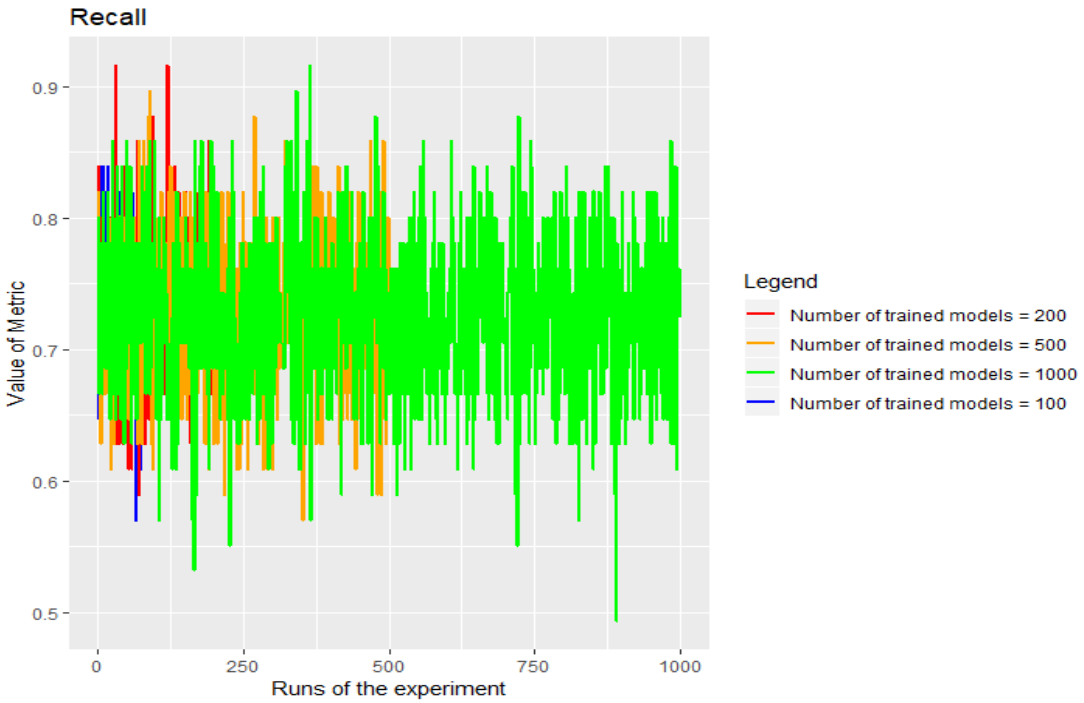
Σε αυτό το σημείο θα παρουσιαστούν τα τελικά αποτελέσματα κάθε αλγορίθμου για τα ακόλουθα πλήθη εκτελέσεων κατηγοριοποίησης: εκατό, διακόσια, πεντακόσια και χίλια. Αυτό θα έχει ενδιαφέρον για να δούμε, αν οι μέσοι όροι των μετρικών αξιολόγησης των αλγορίθμων επωφελούνται από τις πολλαπλές εκτελέσεις ή αν έχουν τόσο σταθερές επιδόσεις, ώστε να μη προκύπτει βελτίωση έπειτα από κάποιο πλήθος εκτελέσεων. Σε κάθε ένα από τα διαγράμματα παρουσιάζονται οι αξίες της κάθε μετρικής για το πλήθος των εκτελέσεων δημιουργίας των μοντέλων κατηγοριοποίησης και της αξιολόγησής τους.

6.5.1 Random Forest

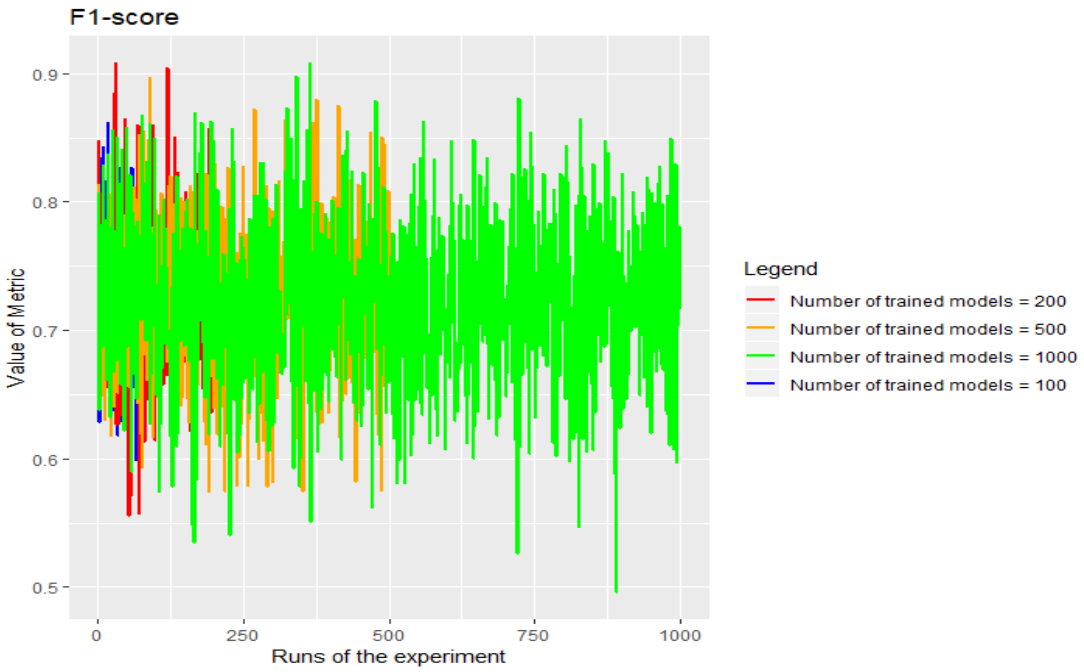


Διάγραμμα 6.33: Ποσοστό επιτυχίας κατηγοριοποίησης «Random Forest»

Διάγραμμα 6.34: Ακρίβεια κατηγοριοποίησης «Random Forest»



Διάγραμμα 6.35: «Recall» κατηγοριοποίησης «Random Forest»



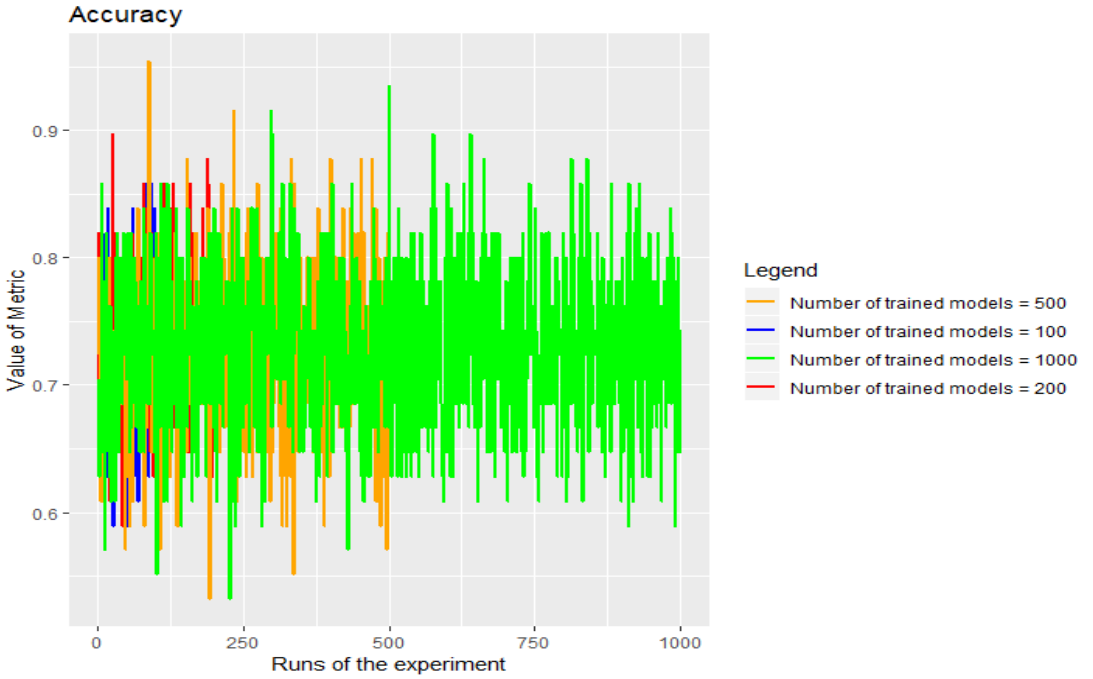
Διάγραμμα 6.36: «F1-score» κατηγοριοποίησης «Random Forest»

Πίνακας 6.9: Μέσοι όροι της τεχνικής «RF», για διαφορετικά πλήθη κατηγοριοποιητών

RF	100 Runs	200 Runs	500 Runs	1000 Runs
Accuracy	0,736731	0,735385	0,726077	0,725269
Precision	0,802564	0,799108	0,789389	0,791815
Recall	0,736731	0,735385	0,726077	0,725269
F1	0,735756	0,731133	0,723684	0,721651

Στην τεχνική του τυχαίου δάσους, παρατηρούμε ότι και στις τρεις μετρικές, δεν είναι εμφανή τα διαγράμματα των πειραμάτων με τα λιγότερα των χιλίων τρεξίματα του κώδικα. Αυτό μας δίνει τη δυνατότητα να συμπεράνουμε, ότι μέχρι και τις πρώτες πεντακόσιες εκτελέσεις, παρατηρείται ομοιότητα στην αξιολόγηση του κατηγοριοποιητή. Αυτό το φαινόμενο, είναι θετικό και επιθυμητό χαρακτηριστικό στα μοντέλα, καθώς μας επιτρέπει να προβλέψουμε την ποιότητά τους, χωρίς να χρειαστεί να τα παράγουμε περισσότερες φορές, για να δούμε τη συμπεριφορά τους. Επίσης, στον πίνακα μέσων όρων φαίνεται, ότι οι μέσοι όροι των τεσσάρων πειραμάτων είναι σταθεροί, καθώς απέχουν ελάχιστα μεταξύ τους, κάτι που δηλώνει σταθερότητα των μοντέλων, όπως και η επικάλυψη των διαγραμμάτων. Φυσικά, όπως σε κάθε πείραμα, υπάρχουν περιπτώσεις τιμών μετρήσεων, που απέχουν πολύ από τον μέσο όρο και ειδικά στην περίπτωση των χιλίων εκτελέσεων παρατηρούμε στα διαγράμματα, τέτοιες περιπτώσεις. Καθώς, αυτές οι περιπτώσεις είναι λίγες, μπορούμε να τις θεωρήσουμε αμελητέες.

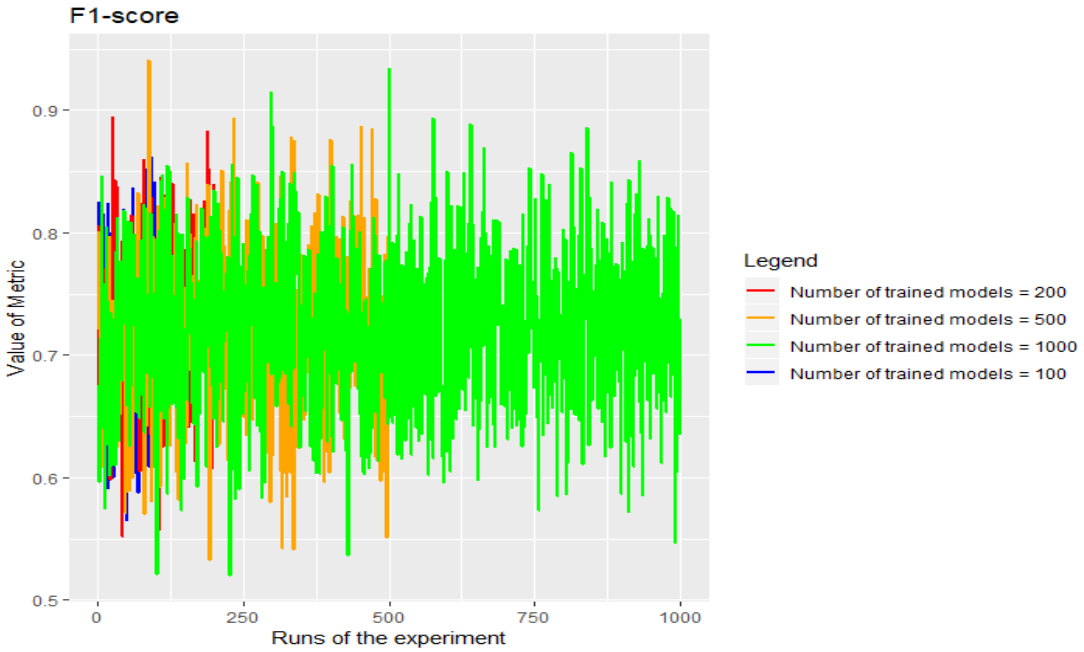
6.5.2 K Nearest Neighbors



Διάγραμμα 6.37: Ποσοστό επιτυχίας κατηγοριοποίησης «K Nearest Neighbors»

Διάγραμμα 6.38: Ακρίβεια κατηγοριοποίησης «K Nearest Neighbors»

Διάγραμμα 6.39: «Recall» κατηγοριοποίησης «K Nearest Neighbors»



Διάγραμμα 6.40: «F1-score» κατηγοριοποίησης «K Nearest Neighbors»

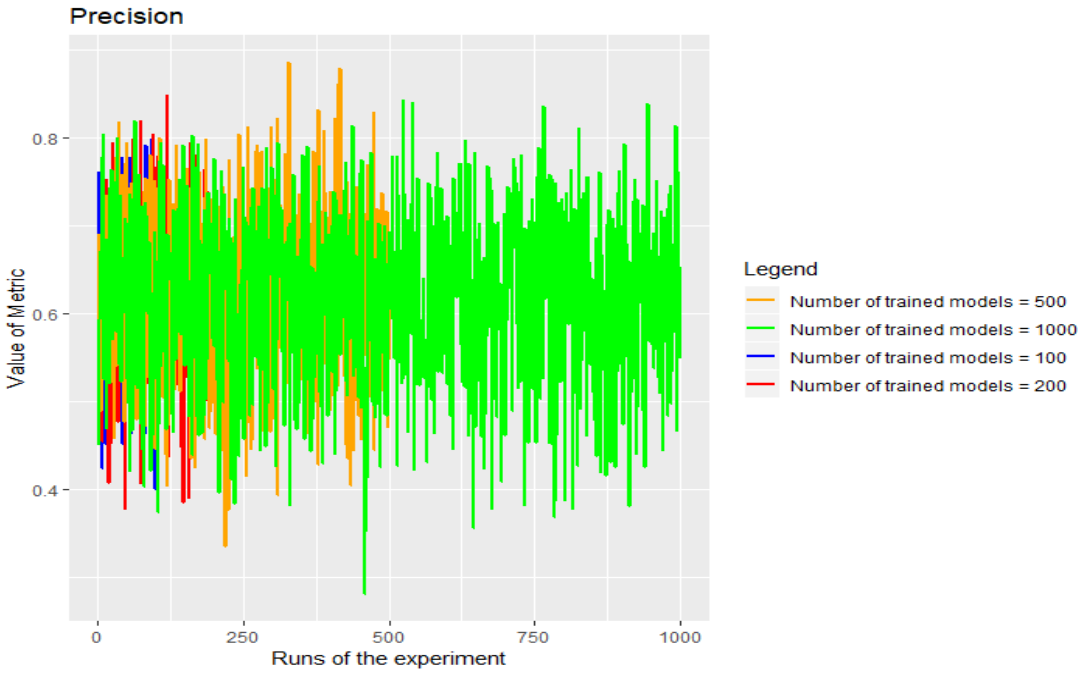
Πίνακας 6.10: Μέσοι όροι της τεχνικής «KNN», για διαφορετικά πλήθη κατηγοριοποιητών

KNN	100 Runs	200 Runs	500 Runs	1000 Runs
Accuracy	0,722115	0,730865	0,729154	0,730346
Precision	0,802256	0,805392	0,803709	0,805494
Recall	0,722115	0,730865	0,729154	0,730346
F1	0,713356	0,723785	0,721244	0,722978

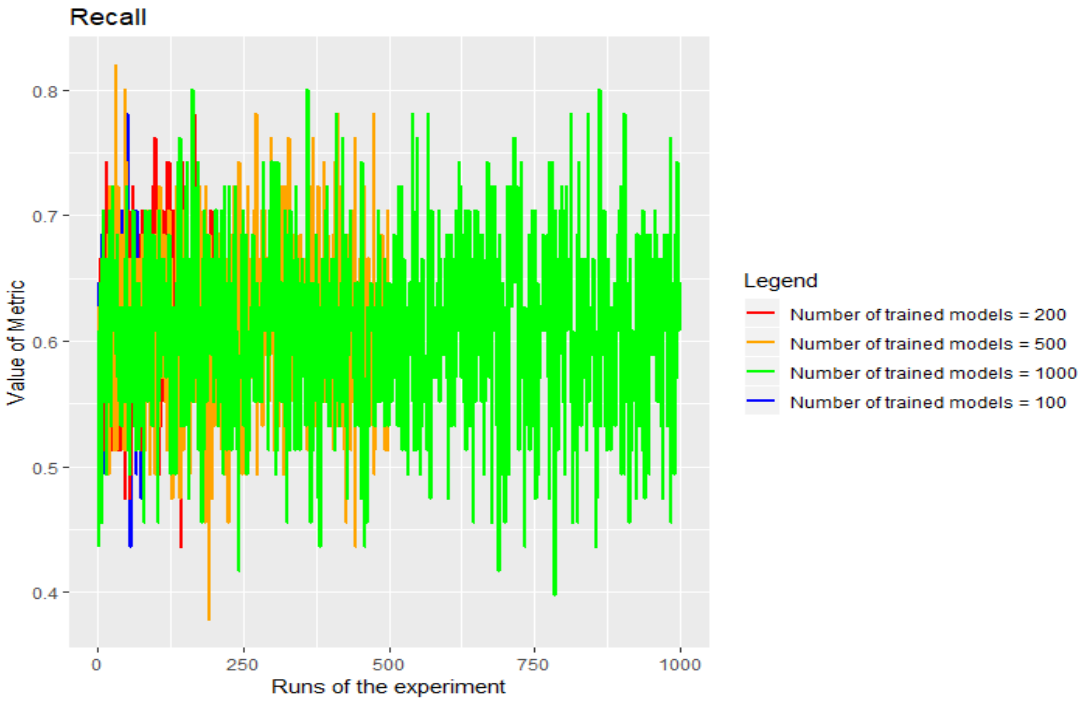
Στην τεχνική των πλησιέστερων γειτόνων, εδώ των πέντε πλησιέστερων γειτόνων, σε όλα τα διαγράμματα παρατηρείται μεγάλη συγκέντρωση των τιμών γύρω από τους μέσους όρους των μετρικών. Με άλλα λόγια οι περισσότερες τιμές των μετρικών απέχουν λίγο από τον αντίστοιχο μέσο όρο τους και άρα αυτό κάνει την τεχνική να θεωρείται σταθερή στην αποτελεσματικότητά της στην κατηγοριοποίηση, σε αυτή τη μελέτη. Απόρροια αυτού, είναι να μειώνεται η απομάκρυνση, που απαιτείται να έχουν οι τιμές των μετρικών, από τον μέσο όρο, για να θεωρηθούν έκτοπες(outliers). Τέτοιες τιμές εμφανίζονται, κυρίως στα διαγράμματα του ποσοστού επιτυχίας(accuracy) και στην ακρίβεια(precision). Πάλι έχουμε ένα πείραμα, όπου τα διαγράμματα των μετρικών των επιμέρους υποπειραμάτων παρουσιάζουν σημαντικά μεγάλη επικάλυψη και οι μέσοι όροι στον πίνακα είναι πολύ κοντά μεταξύ τους, οπότε έχουμε άλλη μία περίπτωση σταθερότητας στην ποιότητα των μοντέλων, ανεξάρτητα από το πόσες εκπαιδεύσεις μοντέλου δοκιμάζουμε. Το γεγονός αυτό επαληθεύεται από τη σύγκριση των μέσων όρων κάθε τεχνικής για το πλήθος εκπαιδεύσεων, στον πίνακα μέσων όρων.

6.5.3 Support Vector Machines

Διάγραμμα 6.41: Ποσοστό επιτυχίας κατηγοριοποίησης «Support Vector Machines»



Διάγραμμα 6.42: Ακρίβεια κατηγοριοποίησης «Support Vector Machines»



Διάγραμμα 6.43: «Recall» κατηγοριοποίησης «Support Vector Machines»

Διάγραμμα 6.44: «F1-score» κατηγοριοποίησης «Support Vector Machines»

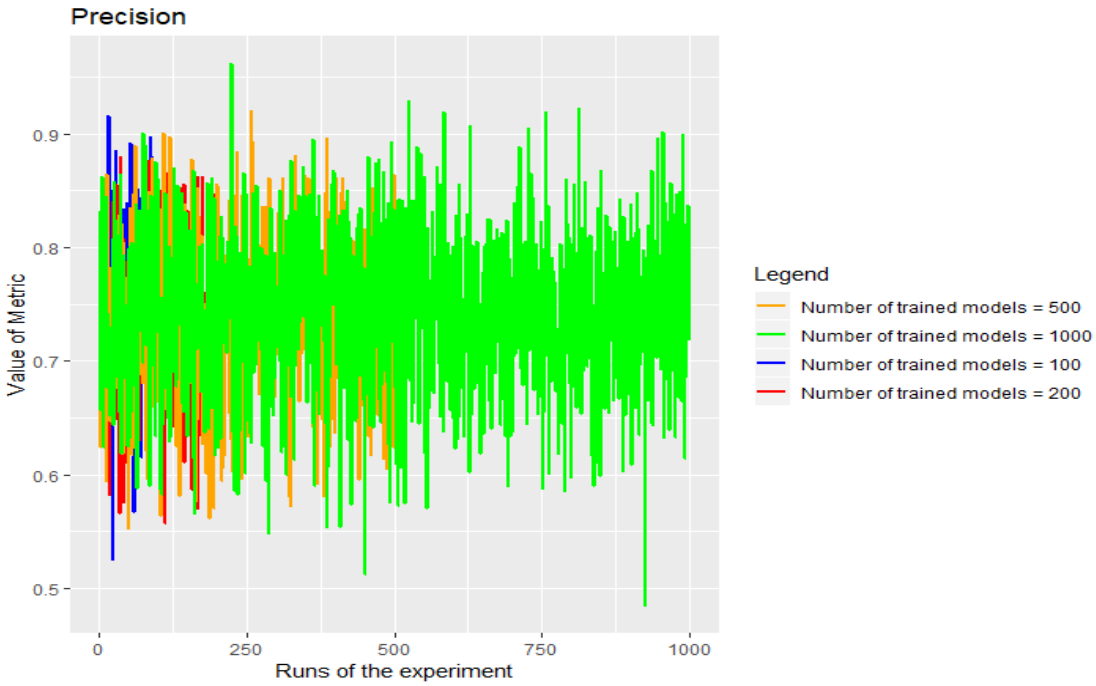
Πίνακας 6.11: Μέσοι όροι της τεχνικής «SVM», για διαφορετικά πλήθη κατηγοριοποιητών

SVM	100 Runs	200 Runs	500 Runs	1000 Runs
Accuracy	0,601923	0,608558	0,612577	0,606827
Precision	0,615006	0,621584	0,626806	0,617465
Recall	0,601923	0,608558	0,612577	0,606827
F1	0,554113	0,563021	0,56739	0,559529

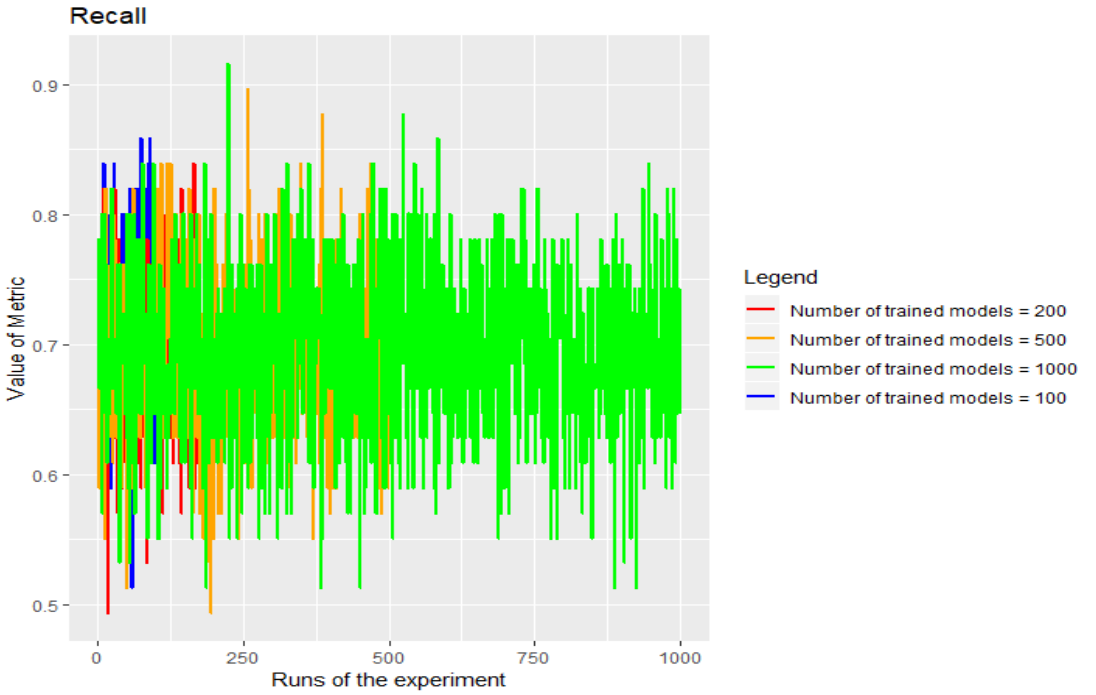
Οι μηχανές διανυσμάτων υποστήριξης, από τα διαγράμματα των σημείων αξιολόγησης και μόνο, δεν δείχνουν να υπόσχονται πολλά για την έρευνά μου. Ανατρέχοντας στον πίνακα των μέσων όρων γίνεται ξεκάθαρη η υποψία μου. Η «SVM», έχει καταφέρει ιδιαίτερα χαμηλά αποτελέσματα στις μετρικές, που χρησιμοποιώ στα πειράματα, με τις παραμέτρους, που επέλεξα για να την φτιάξω, στην τελική αξιολόγηση. Φυσικά, τα διαγράμματά της είναι μετατοπισμένα προς τα κάτω, σε σχέση με τα διαγράμματα των δύο προηγηθέντων τεχνικών, ωστόσο υπάρχει κάτι αξιοσημείωτο. Όλες οι μετρικές αυτής της τεχνικής έχουν διάσπαρτες τις τιμές τους, πάνω στον άξονα «y» των διαγραμμάτων. Με άλλα λόγια, οι τιμές τους, παρουσιάζουν μικρότερη συγκέντρωση γύρω από τους αντίστοιχους μέσους όρους. Βλέπουμε, πως ακόμη αυτή η απόσταση από τους μέσους όρους ποικίλει σημαντικά, με μεγάλες αποκλίσεις τόσο προς τους μεγάλους, όσο και προς τους μικρούς αριθμούς. Συνεπώς, σημαίνει ότι η αποτελεσματικότητα της τεχνικής αμφισβητήσιμη για τη δική μας έρευνα.

6.5.4 Decision Tree

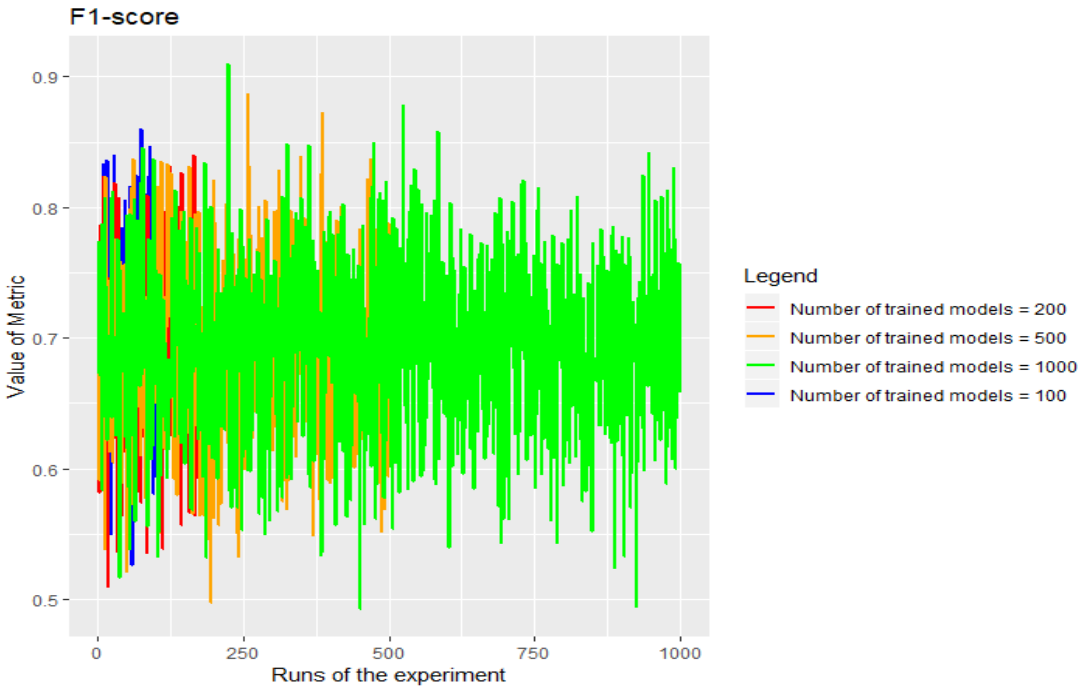
Διάγραμμα 6.45: Ποσοστό επιτυχίας κατηγοριοποίησης «Decision Tree»



Διάγραμμα 6.46: Ακρίβεια κατηγοριοποίησης «Decision Tree»



Διάγραμμα 6.47: «Recall» κατηγοριοποίησης «Decision Tree»



Διάγραμμα 6.48: «F1-score» κατηγοριοποίησης «Decision Tree»

Πίνακας 6.12: Μέσοι όροι της τεχνικής «DT», για διαφορετικά πλήθη κατηγοριοποιητών

DT	100 Runs	200 Runs	500 Runs	1000 Runs
Accuracy	0,710577	0,6925	0,695962	0,694365
Precision	0,755282	0,738003	0,740923	0,741898
Recall	0,710577	0,6925	0,695962	0,694365
F1	0,70997	0,689423	0,694693	0,692863

Τελευταία τεχνική, στη σειρά, είναι το δέντρο απόφασης, το οποίο δεν εμφανίζει τους χειρότερους μέσους όρους, από τις τεχνικές. Το «decision tree» έχει πετύχει να βρίσκεται κοντά στο «random forest» και στο «k nearest neighbors», αλλά τα ποσοστά του, το τοποθετούν ως λιγότερο αποτελεσματικό. Ενδιαφέρον, αποτελεί και το γεγονός, ότι και εδώ, όπως και στο «support vector machine», οι μετρικές ανά εκπαίδευση, έλαβαν τιμές, που απέχουν πολύ από τον τελικό μέσο όρο κάθε μίας. Έτσι έχουμε μεγάλες διαφορές στην ποιότητα των δέντρων απόφασης, που παράγει κάθε εκπαίδευση. Σε κάθε περίπτωση, τα ποσοστά της τεχνικής αυτής για τη δημιουργία κατηγοριοποιητών, είναι χαμηλότερα από των δύο πρώτων, κατά σειρά αναφοράς.

7 ΣΥΜΠΕΡΑΣΜΑΤΑ

Όπως έγινε φανερό, για να ολοκληρωθεί αυτή τη σύγκριση των τεχνικών και των αλγορίθμων τους, χρησιμοποιήθηκε η προσέγγιση διεξοδικής ή εξαντλητικής αναζήτησης, γνωστή στο χώρο της πληροφορικής ως «exhaustive search». Είναι σημαντικό, πριν ξεκινήσουμε να βγάζουμε συμπεράσματα, πως όλα όσα θα ακολουθήσουν, αφορούν την έρευνα με αυτές τις παραμέτρους και αυτά τα δεδομένα, που είχα στη διάθεσή μου. Υπάρχει σοβαρή πιθανότητα με περισσότερα δεδομένα ή και διαφορετικές τιμές στις παραμέτρους των συναρτήσεων δημιουργίας των κατηγοριοποιητών, τα αποτελέσματα να αλλάξουν σημαντικά.

Ας ξεκινήσουμε με τα συμπεράσματα από την τελική φάση αξιολόγησης, όπου κάθε τεχνική και αλγόριθμος είχε την καλύτερη ευκαιρία να παρουσιάσει θετικά αποτελέσματα, έχοντας ως δεδομένες τις τιμές στις παραμέτρους, που προέκυψαν από τα προηγούμενα πειράματα. Ο αλγόριθμος των «support vector machines» παρουσίασε τις μικρότερες τιμές σε όλες τις μετρικές αξιολόγησης, σε σχέση με τους άλλους αλγορίθμους. Με μέσο όρο ποσοστού επιτυχίας προβλέψεων σταθερά στο 60%, ανεξάρτητα από πόσες δοκιμές θα γίνουν, δεν είναι αξιοποιήσιμος για εφαρμογές της καθημερινής ζωής. Επόμενος, είναι ο «decision tree», ο οποίος με ποσοστό επιτυχίας πρόβλεψης στο 70% κατά μέσο όρο, είναι επίσης πολύ αδύναμος για να εφαρμοστεί. Ωστόσο, αξίζει να σημειωθεί, πως παρουσιάζεται μεγάλη διαφορά στους μέσους όρους των μετρικών αξιολόγησης, των δύο υποδεέστερων τεχνικών της έρευνας, δηλαδή το δέντρο απόφασης παρουσιάζει 70% ποσοστό επιτυχίας, 74% ακρίβεια, 71% «recall», 71% «f1-score», ενώ η μηχανή διανύσματος υποστήριξης δίνει 60%, 61%, 60% και 55%, αντίστοιχα. Οι δύο άλλοι αλγόριθμοι, «k nearest neighbors» και «random forest», εμφανίζουν πολύ όμοια μεταξύ τους αποτελέσματα. Ο «knn» βγάζει 73% ποσοστό επιτυχίας ταξινόμησης, με 80% ακρίβεια, 73% «recall» και 72% «f1-score». Όμοια ο «rf» δίνει 73% ποσοστό επιτυχίας κατηγοριοποίησης, με 79% ακρίβεια, 73% «recall» και 73% «f1-score». Τα ποσοστά αυτά είναι προσεγγίσεις των μέσων όρων, στα τελικά πειράματα. Έτσι ο «knn» δείχνει πως αξίζει περισσότερη προσοχή από όλους τους αλγορίθμους, ως αυτός με τη μεγαλύτερη δυνατότητα να γίνει αξιοποιήσιμος, αν μελετηθεί η συμπεριφορά του και βρεθούν καλύτερες τιμές στις μεταβλητές του, κατά το σχηματισμό του μοντέλου κατηγοριοποίησης.

Στη συνέχεια, αξίζει να προσέξουμε, πως οι δύο επικρατούσες τεχνικές είναι πολύ κοντά μεταξύ τους και με την παραμικρή αλλαγή, είτε στο σύνολο των δεδομένων εκπαίδευσης, είτε στην τιμή κάποιας από τις παραμέτρους, που κράτησα στην προεπιλεγμένη της τιμή, θα μπορούσε να αλλάξει αυτή τη σειρά προτεραιότητας.

Είναι ενδιαφέρον, πως ο κατηγοριοποιητής, που αποτελείται από ένα δέντρο απόφασης, αποδίδει σημαντικά λιγότερο από τον κατηγοριοποιητή, που αποτελείται από πολλά μικρότερα δέντρα απόφασης. Το αξιοσημείωτο είναι ότι στο πακέτο «scikit-learn» τόσο το μονό δέντρο απόφασης, όσο και τα δέντρα απόφασης, του τυχαίου δάσους, προκύπτουν από τον ίδιο βελτιωμένο «cart» αλγόριθμο. Είναι ασφαλές λοιπόν, να εξάγουμε το συμπέρασμα, πως η διαφορά τους προκύπτει από τη «δύναμη των πολλών» και από τη συνάρτηση συνένωσης των αποτελεσμάτων των δέντρων για να βγει το αποτέλεσμα του κατηγοριοποιητή.

Παρατήρηση, όμοιου ενδιαφέροντος, αποτελεί ότι οι δύο αλγόριθμοι, που απεικονίζουν τις εγγραφές των δεδομένων εκπαίδευσης στον χώρο, είναι και αυτές με τη μεγαλύτερη απόκλιση στην αποτελεσματικότητά τους. Από το γεγονός αυτό, μπορούμε να καταλήξουμε σε δύο πιθανά συμπεράσματα. Το πρώτο είναι, ότι δεν υπάρχουν γραμμές κατάλληλες, ώστε να χωρίσουν σωστά τις εγγραφές στα υπερεπίπεδα, που ανήκουν και κάθε υπερεπίπεδο να περιέχει μίας κατηγορίας εγγραφές. Το δεύτερο είναι, ότι κάποια από τις παραμέτρους, που αφήσαμε στην προεπιλεγμένη της τιμή, κατά την εκπαίδευση του ταξινομητή, έκανε μεγάλη ζημιά στην απόδοση του μοντέλου, που προέκυψε.

Σε κάθε περίπτωση, αυτό που πρέπει να κρατήσουμε είναι ότι καμία από τις τέσσερις τεχνικές, με τα υπάρχοντα δεδομένα δεν έδωσε μοντέλο κατηγοριοποίησης, έτοιμο για αξιοποίηση. Ωστόσο, οι τεχνικές του τυχαίου δάσους και των πλησιέστερων γειτόνων, είναι πολλά υποσχόμενες και αξίζει να μελετηθούν περισσότερο. Κρίνω απαραίτητο και με δεδομένο, ότι τα δεδομένα ακολουθιών «dna» δεν είναι εύκολο και φθηνό να συλλεχθούν, να γίνει έρευνα με τις υπόλοιπες παραμέτρους των συναρτήσεων δημιουργίας των κατηγοριοποιητών. Παράλληλα, θα πρότεινα να διερευνηθούν περισσότεροι αλγόριθμοι ταξινόμησης και η συλλογή περισσότερων δεδομένων να ξεκινήσει αν δεν υπάρξει κάποια πρόοδος με τις προαναφερθείσες προσεγγίσεις του προβλήματος.

ΠΗΓΕΣ

1. Βικιπαίδεια στα ελληνικά-Λαγοκέφαλος https://el.wikipedia.org/wiki/%CE%9B%CE%B1%CE%B3%CE%BF%CE%BA%CE%AD%CF%86%CE%B1%CE%BB%CE%BF%CF%82
2. Βικιπαίδεια στα αγγλικά- Lessepsian_migration https://en.wikipedia.org/wiki/Lessepsian_migration
3. Βικιπαίδεια στα αγγλικά-FASTA format https://en.wikipedia.org/wiki/FASTA_format
4. Βικιπαίδεια στα αγγλικά-Data mining https://en.wikipedia.org/wiki/Data_mining
5. [Chris Nicholson] Evaluation Metrics for Machine Learning-Accuracy, Precision, Recall, and F1 https://wiki.pathmind.com/accuracy-precision-recall-f1
6. [Shervin Minaee Oct 28, 2019]20 Popular Machine Learning Metrics Part 1. Classification & Regression Evaluation Metrics https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce
7. [Boaz Shmueli Jul 2, 2019] Multi-Class Metrics Made Simple, Part 1: Precision and Recall https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2
8. [Boaz Shmueli Jul 3, 2019] Multi-Class Metrics Made Simple, Part 2: the F1-score https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1
9. [Max Miller Feb 10, 2020] The basics: Evaluating classifiers https://towardsdatascience.com/the-basics-evaluating-classifiers-b0078a097732
10. [Βασίλειος Σ. Βερύκιος, Σταύρος Σουραβλάς, 2017] Εισαγωγή στην εξόρυξη δεδομένων. Μετάφραση του «Introduction to Data Mining», Pang-ning Tan, Michael Steinbach, Vipin Kumar
11. [Beniwal, Sunita & Arora, Jitender. 2012]. Classification and feature selection techniques in data mining. International Journal of Engineering Research and Technology. 1. https://www.researchgate.net/publication/263662705_Classification_and_feature_selection_techniques_in_data_mining
12. [Wenliang Du, Zhijun Zhan 2002]. Building Decision Tree Classifier on Private Data https://surface.syr.edu/cgi/viewcontent.cgi?article=1007&context=eecs
13. [Hucker Marius Jun 15,2020]. Tree algorithms explained: Ball Tree Algorithm vs KD Tree vs Brute Force https://towardsdatascience.com/tree-algorithms-explained-ball-tree-algorithm-vs-kd-tree-vs-brute-force-9746debcd940
14. Zhengzheng Xing, Jian Pei, Eamonn Keogh. A Brief Survey on Sequence Classification

https://www.cs.sfu.ca/~jpei/publications/Sequence%20Classification.pdf?fbclid=IwAR2ltfIE4w6iMqGv1nK1q6-51buuxiLJ9eQTZrNLNgoIlftDZswBNpzED7Q

15. [Gianpaolo Coro, Luis Gonzalez Vilas, Chiara Magliozzi, Anton Ellenbroek, Paolo Scarponi, Pasquale Pagano, March 2018]. Forecasting the ongoing invasion of Lagocephalus sceleratus in the Mediterranean Sea. Ecological Modelling. Volume 371. http://www.sciencedirect.com/science/article/pii/S0304380018300164