ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΣΧΟΛΗ ΚΟΙΝΩΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΓΕΩΓΡΑΦΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΓΕΩΓΡΑΦΙΑ ΚΑΙ ΕΦΑΡΜΟΣΜΕΝΗ ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

# *Θέμα: «ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΤΑΞΙΝΟΜΗΣΗ ΔΟΡΥΦΟΡΙΚΩΝ ΕΙΚΟΝΩΝ»*

Επιβλέπων Καθηγητής: Καβρουδάκης Δημήτρης
Επιτροπή : Βαΐτης Μιχάλης, Τοπουζέλης Κωνσταντίνος

Παπαχρόνης Ιωάννης
Μυτιλήνη, Ιούλιος 2019

# Ευχαριστίες

Στα πλαίσια εκπόνησης της παρούσας διπλωματικής εργασίας για το ΠΜΣ Γεωγραφίας και Εφαρμοσμένης Γεωπληροφορικής θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που με στήριξαν για να πετύχω το στόχο μου.

Αρχικά θα ήθελα να ευχαριστήσω τον Επιβλέποντα Καθηγητή, Δ. Καβρουδάκη για τη συνεργασίας μας, για την εμπιστοσύνη που έδειξε στο πρόσωπο μου και για τη βοήθεια που προσέφερε με τις εξειδικευμένες γνώσεις του.

Επίσης θα ήθελα να ευχαριστήσω την οικογένεια μου, για τη διαχρονική και διαρκεί στήριξη που μου προσφέρει όλα αυτά τα χρόνια, με αποκορύφωμα την ολοκλήρωση των μεταπτυχιακών μου σπουδών.

Τέλος, τίποτα δε θα είχε επιτευχθεί, χωρίς την αμέριστη, ανιδιοτελή, συνεισφορά και στήριξη των φίλων μου, τόσο πριν, όσο και κατά τη διάρκεια πραγματοποίησης της εργασίας. Θα ήθελα να ευχαριστήσω ιδιαιτέρως, τον Άρη, την Κάσια, τη Χριστίνα, το Μάριο, το Γιώργο, τον Παύλο, το Λουκά, τον Αλέξη, το Βασίλη και τόσους ακόμα φίλους, που ο καθένας με τον τρόπο του μου έδωσε κουράγιο και υπομονή για την ολοκλήρωση των σπουδών μου.

# Περίληψη

Στα πλαίσια της παρούσας διπλωματικής εργασίας με θέμα «Σύγκριση αλγορίθμων εξόρυξης δεδομένων στην ταξινόμηση δορυφορικών εικόνων» έγινε μια προσπάθεια να παρουσιαστεί μια διαφορετική μεθοδολογική προσέγγιση για την ταξινόμηση μιας δορυφορικής εικόνας. Η εξόρυξη δεδομένων είναι μια μέθοδος που εφαρμόζεται σε ένα ευρύ φάσμα επιστημονικών πεδίων και αποτελεί μια λύση στο πρόβλημα της διαχείρισης μεγάλων όγκων δεδομένων.

Για την εκπόνηση της εργασίας έγινε μια προσπάθεια αναπαραγωγής των τεχνικών εξόρυξης δεδομένων στο πεδίο της τηλεπισκόπισης, χρησιμοποιώντας δορυφορικά δεδομένα για την ταξινόμηση των Χρήσεων Γης για τη Νήσο Λέσβο. Σε όλα τα στάδια υλοποίησης της εργασίας για την εκπλήρωση του στόχου χρησιμοποιήθηκαν τα λογισμικά που διέπονται από την αρχή του ανοικτού κώδικα Qgis και Rstudio και ελεύθερα δεδομένα.

Τα κεφάλαια της εργασίας αποτελούνται από την Εισαγωγή, τη Θεωρητική Ανάλυση και τους Αλγορίθμους, τη Μεθοδολογία, τα Αποτελέσματα, τα Συμπεράσματα, τις Βελτιώσεις και τέλος τον Κώδικα που χρησιμοποιήθηκε για τη διαδικασία της εξόρυξης.

Πιο συγκεκριμένα στο πρώτο κεφάλαιο αναφέρονται οι λόγοι, ο στόχος και ο σκοπός της εργασίας, στο δεύτερο κεφάλαιο αναφέρεται η θεωρία και η διαδικασία της εξόρυξης δεδομένων, της ταξινόμησης και οι αλγόριθμοι που χρησιμοποιήθηκαν στην εργασία.

Στο τρίτο και τέταρτο κεφάλαιο παρουσιάζεται η μεθοδολογία και τα αποτελέσματα της εργασίας, όπου, αρχικά γίνεται αναφορά στην περιοχή μελέτης, στη συλλογή και επεξεργασία των δεδομένων, στις διαδικασίες που ακολούθησαν στο Qgis και στο Rstudio και στη συνέχεια παρουσιάζονται τα αποτελέσματά, με τη μορφή γραφημάτων, πινάκων και γίνονται οι απαραίτητες συγκρίσεις μεταξύ των αλγορίθμων.

Τέλος στα επόμενα κεφάλαια, αναφέρονται τα συμπεράσματα της εργασίας, οι βελτιώσεις που θα μπορούσαν να γίνουν σε μελλοντικές εργασίες και ακολουθεί ο κώδικας που παράχθηκε για την πραγματοποίηση της παρούσας εργασίας.

# Abstract

In the context of this thesis on "Comparison of data mining algorithms in classification of remote sensing data", there is an attempt to present a different methodological approach for the classification of remote sensing data. Data mining is a method applied to a wide range of scientific fields and is a solution to the problem of managing large amounts of data.

For the preparation of this thesis there was an attempt to reproduce the data mining techniques in the field of remote sensing, using remote sensing data for the classification of land uses for the island of Lesvos. In all the stages of this project, the software that is used applies by the principle of open source Qgis and Rstudio and free data were used.

The chapters of the thesis consist of the introduction, the theoretical analysis and the algorithms, the methodology, the results, the conclusions, the improvements and finally the code used for the extraction process.

More specifically, the first chapter mentions the reasons, the objective and the purpose of the thesis, in the second chapter refers to the theory and the process of data mining, classification and the algorithms used in the thesis.

The third and fourth chapters show the methodology and results of the work, where, initially, a reference is made to the study area, the collection and processing of data, the procedures that followed in Qgis and Rstudio and then follow the results, presented in the form of graphs, tables and made the necessary comparisons between the algorithms.

Finally, in the following chapters, are mentioned the conclusions of the thereis, the improvements that could be made in future work and they are followed by the code produced for the implementation of this thesis.

# Περιεχόμενα


**1 ΕΙΣΑΓΩΓΗ ..... 1**

**2 ΘΕΩΡΗΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ..... 3**

2.1 Χρήσεις Γης ..... 3

2.2 Εξόρυξη Δεδομένων ..... 3

2.3 Νεες Εφαρμογες – Νεες προκλήσεις ..... 5

*2.3.1 Δεδομένα επιχειρήσεων και ηλεκτρονικού εμπορίου ..... 5*

*2.3.2 Δεδομένα Δικτύου (Web Data) ..... 5*

*2.3.3 Εμπορικές συναλλαγές( Business & E-commerce Data) ..... 6*

*2.3.4 Ηλεκτρονικό εμπόριο (Electronic commerce) ..... 6*

*2.3.5 Γονιδιωματικά δεδομένα.(Genomic data) ..... 6*

*2.3.6 Δορυφορικά Δεδομένα( Remote Sensing Data) ..... 6*

*2.3.7 Δεδομένα προσομοίωσης (Simulation Data) ..... 7*

*2.3.8 Στοιχεία υγειονομικής περίθαλψης (Health care Data) ..... 7*

2.4 Ταξινόμηση ..... 7

2.5 Διαδικάσια Data Mining ..... 9

2.6 Bagging Boosting Stacking ..... 11

*2.6.1 Bagging ..... 11*

*2.6.2 Boosting ..... 13*

*2.6.3 Stacking ..... 14*

2.7 Αλγόριθμοι που Χρησιμοποιήθηκαν ..... 14

*2.7.1 Stochastic Gradient Boosting (GBM) ..... 15*

*2.7.2 K-nearest neighbor (KNN) ..... 15*

*2.7.3 Classification and Reggresion Trees(CART) ..... 16*

*2.7.4 Random Forest (RF) ..... 17*

*2.7.5 Bagged Cart ..... 19*

*2.7.6 Linear Discriminant Analysis (LDA) ..... 20*

*2.7.7 Support Vector Machine( SVM) ..... 21*

*2.7.8 C5.0 ..... 22*

2.8 Αξιολόγηση Αλγορίθμων: Μετρικές Accuracy και Kappa ..... 23

**3 ΜΕΘΟΔΟΛΟΓΙΑ ..... 24**

3.1 Περιοχή Μελέτης ..... 24

3.2 Συλογή Δεδομένων ..... 25

3.3 Επεξεργασια Δεδομένων ..... 26

3.4 Η Γλωσσα R και το RStudio ..... 26

3.5 Διαδικασια Μέσω Rstudio ..... 27

3.6 Διάγραμμα Ροής ..... 28

**4 ΑΠΟΤΕΛΕΣΜΑΤΑ ..... 29**

4.1 Συγκρισεις - Χρονική Απόκριση ..... 31

4.2 Συγκρίσεις – Εκτιμηση Ακρίβειας ..... 33

4.3 Συγκρίσεις – Οπτικοποίηση Αποτελεσμάτων ..... 35

**5 ΣΥΜΠΕΡΑΣΜΑΤΑ-ΣΥΖΗΤΗΣΗ ..... 49**

**6 ΒΕΛΤΙΩΣΕΙΣ ..... 50**

**7 ΒΙΒΛΙΟΓΡΑΦΙΑ.....52**
7.1 Ξενόγλωσση.....52
7.2 Ελληνική.....54
7.3 Ηλεκτρονικές Πηγές.....54
7.4 Διατριβές.....56
**8 ΚΩΔΙΚΑΣ.....57**

# 1 ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια παρατηρείται μια ταχύτατη αύξηση του όγκου των αποθηκευμένων δεδομένων. Η ανάλυση όλων αυτών των δεδομένων και η ανάκτηση χρήσιμης πληροφορίας χωρίς τη χρήση εξειδικευμένων τεχνικών είναι αδύνατη. Η Εξόρυξη Δεδομένων , αντλώντας μεθοδολογίες από τη Μηχανική Μάθηση, τις Βάσεις Δεδομένων, τη Στατιστική και άλλους κλάδους, έχει στόχο την ανακάλυψη γνώσης μέσα από μεγάλους όγκους δεδομένων. (Γ. Κύρκος, 2015). Η τηλεπισκόπηση αποτελεί ένα ακόμα επιστημονικό πεδιο που χρησιμοποιεί αυτές τις μεθόδους.

Σκόπος της παρούσας διπλωματικής εργασίας αποτελεί η σύγκριση αλγορίθμων που χρησιμοποιούνται στην εξόρυξη δεδομένων για την ταξινόμηση και για την οποία παράγονται μοντέλα πρόβλεψης.

Οι στόχοι της εργασίας είναι, αρχικά να παρουσιάσω μια διαφορετική μεθοδολογική προσέγγιση για την ταξινόμηση μια δορυφορικής εικόνας, να μιλήσω για μια, νέα χρονολογικά μέθοδο, με πολλές εφαρμογές που συμβαδίζει με τις ανάγκες της εποχής και να μας δώσει μια νέα δυνατότητα, συνδυασμού επιστημονικών πεδίων , συνδυασμού μεθόδων και διαδικασιών.

Στην παρούσα εργασία έγινε μια προσπάθεια αναπαραγωγής των τεχνικών εξόρυξης δεδομένων στο πεδίο της τηλεπισκόπισης, χρησιμοποιώντας δορυφορικά δεδομένα για την ταξινόμηση των Χρήσεων Γής για τη Νήσο Λέσβο. Σε όλα τα στάδια υλοποίησης της εργασίας για την εκπλήρωση του στόχου χρησιμοποιήθηκαν τα λογισμικά που διέπονται απο την αρχή του ανοικτού κώδικα Qgis και Rstudio και ελεύθερα δεδομένα. Κατα τη διαδικασία υλοποίησης της εργασίας, επιλέχθηκαν οι Χρήσεις Γής της Νήσου Λέσβου ως μελέτη περίπτωσης και προσεγγίστηκαν 8 αλγορίθμοι ταξινόμησης.

Στην πρώτη ενότητα παρουσιάζεται για το θεωρητικό πλαίσιο της παρούσας εργασίας, όπως για την εξόρυξη δεδομένων, τις νέες εφαρμογές, την ταξινόμηση, τις μεθόδους Bagging, Boosting και Stacking, τους αλγόριθμους που χρημοποιήθηκαν στην εργασία και τέλος την περιγραφή των εκτιμητων ακρίβειας Accuracy και Kappa.

Στη συνέχεια η ενότητα που ακολουθεί αφορά τα στάσια μεθοδολογία που πραγματοποιήθηκαν για την ταξινόμηση. Ειδικότερα, θα μιλήσουμε για την περιοχή

μελέτης, τη συλλογή των δεδομένων, την επεξεργασία τους και τέλος παρουσιάζεται ένα διάγραμμα ροής της μεθοδολογίας.

Στην επόμενη ενότητα γίνεται αφορά στα αποτελέσματα στα οποία αναφέρονται, τα σημεία και οι κλάσεις που υπήρξε εκπροσώση, αποτυπώνονται τα γραφήματα με τη διάρκεια απόκρισης των μοντέλων, τα αποτελέσματα των αλγορίυμων καθώς και σχολιασμός των αποτελεσμάτων.

Ακολουθούν τα συμπεράσματα των αποτελεσμάτων της ταξινόμησης, της μεθοδολογίας και τα προβλήματα που παρουσιάστηκαν κατα τη διάρκεια της εργασίας.

Και τέλος παρατίθεται ο κώδικας που χρησιμοποιήθηκε στο Rstudio και η βιβλιογραφία για την εκπόνηση της παρούσας διπλωματικής εργασίας.

# 2 ΘΕΩΡΗΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ

## 2.1 Χρήσεις Γης

Η χρήση και η κάλυψη της γης είναι αποτέλεσμα της ανθρώπινης χρήσης της γης και των αλληλεπιδράσεων που ακολουθούν. Συνεπώς, η τακτική παρακολούθησης και αξιολόγησης της αλλαγής της χρήσης γης και της κάλυψης της γης είναι κρίσιμη για την κατανόηση της έκτασης και των επιπτώσεων τέτοιων ανθρωπογενών και φυσικών αλλαγών στη Γη σε τοπική, περιφερειακή ή παγκόσμια κλίμακα (Potapov et al., 2008). Τα δεδομένα τηλεπισκόπησης έχουν χρησιμοποιηθεί ευρέως για να ταξινομήσουν την κάλυψη της γης και να παρέχουν εκτιμήσεις για την αντίστοιχη περιοχή. Η δύναμη της τηλεπισκόπησης έγκειται στην ικανότητά της να παρέχει χωρικά σαφή πληροφόρηση για την κάλυψη μεγάλων περιοχών, ιδιαίτερα απομακρυσμένων, και με δύσκολια προσβασιμότητας (Lillesand και Kiefer, 1994).

## 2.2 Εξόρυξη Δεδομένων

Τα τελευταία χρόνια, η Εξόρυξη Δεδομένων έχει γίνει όλο και πιο δημοφιλής. Μαζί με την εποχή της πληροφορίας, η ψηφιακή επανάσταση κατέστησε αναγκαία τη χρήση ορισμένων μεθόδων για να είναι σε θέση να αναλύσει το μεγάλο όγκο των δεδομένων που είναι διαθέσιμα. Η εξορυξη δεδομένων είναι η διαδικασία εξαγωγής έμμεσων, παλαιότερα γνωστών και δυνητικά χρήσιμων πληροφοριών από τα δεδομένα της βάσης δεδομένων(Witten & Eibe, 2005). Επιπρόσθετα, είναι διαδικασία εξερεύνησης και ανάλυσης μεγάλων όγκων δεδομένων, με αυτόματων ή ημιαυτόματων μέσων, με στόχο να ανακαλυφθούν σημαντικά πρότυπα και κανόνες.(Berry, 1999) Η εξόρυξη δεδομένων αφορά όλες τις τεχνικές και διαδικασίες, εύρεσης νέας και πιθανόν χρήσιμης γνώσης από δεδομένα, που στην πράξη είναι μεγάλες Βάσεις Δεδομένων (U. Fayyad, et al, 1996).

Ένας άλλος ορισμός της εξόρυξης δεδομένων αναφέρεται ότι αποτελλεί μια διαδικασία που συνίστα την εφαρμογή αλγορίθμων ανάλυσης δεδομένων και που, κάτω από αποδεκτούς περιορισμούς υπολογιστικής απόδοσης, παράγουν μια συγκεκριμένη απαρίθμηση προτύπων (ή μοντέλων) πάνω στα δεδομένα (Fayyad,1997).

Οι τεχνικές εξόρυξης δεδομένων έχουν εφαρμοστεί επιτυχώς σε πολλούς τομείς από τις επιχείρησεις έως την επιστήμη και τον αθλητισμό. Η εξόρυξη δεδομένων έχει χρησιμοποιηθεί στο μάρκετινγκ χρησιμοποιώντας βάσεις δεδομένων, ανάλυση δεδομένων λιανικής, χρηματιστίριο κλπ. Οι τεχνολογίες εξόρυξης δεδομένων έχουν χρησιμοποιηθεί στην αστρονομία, τη μοριακή βιολογία, την ιατρική, τη γεωλογία και πολλούς άλλους τομείς. Έχει επίσης χρησιμοποιηθεί στη διαχείριση της υγειονομικής περίθαλψης, την ανίχνευση φορολογικών απάτων, την παρακολούθηση της νομιμοποίησης εσόδων από παράνομες δραστηριότητες κ.α.

Πιο συγκεκριμένα οι τομείς που έχουν εφαρμοσθεί οι τεχνικές εξόρυξης δεδομένων, έχουν ως εξής:

- Διαχείριση της Αγοράς (Market management)

Σε αυτήν την εφαρμογή στόχος είναι το μάρκετινγκ, η διαχείριση πελατειακών σχέσεων, η ανάλυση καλαθιού αγοράς, οι διασταυρούμενες πωλήσεις και η κατάτμηση της αγοράς. Όπως επίσης και στην διαχείριση κινδύνου της Αγοράς (Risk management) όπως για παράδειγμα στις προβλέψεις, στη διατήρηση πελατών, τη βελτιωμένη αναδοχή, τον έλεγχος ποιότητας, και την ανάλυση του ανταγωνισμού.

- Στην Ανίχνευση και Διαχείρηση Απάτης (Fraud management& Fraud detection).

-Σε βιομηχανικές εφαρμογές, σε τραπεζικές, χρηματοοικονομικές και κινητές αξίες όπως: η ανάλυση της κερδοφορίας (για κάθε υποκατάστημα, προϊόν, ομάδα προϊόντων, παρακολούθηση προγραμμάτων και καναλιών παρακολούθησης, ανάλυση δεδομένων πελάτη).

-Στις τηλεπικοινωνίες και μέσα ενημέρωσης.(Telecommunications and media)

Στη Βαθμολογία απόκρισης, στη διαχείριση εκστρατειών μάρκετιγκ, στην ανάλυση κερδοφορίας και κατακερματισμός πελατών.

-Στην Υγεία (Health care)

Στο Σύστημα Διαχείρισης Απάτης και Κατάχρησης που βοηθά οργανισμούς ασφάλισης υγείας που ασχολούνται με την απάτη και την κακοδιαχείρηση: ανίχνευση, διερεύνηση, διευθέτηση, πρόληψη υποτροπής.

## 2.3 Νέες Εφαρμογές – Νέες προκλήσεις

Δίνοντας λύσεις στα παραπάνω πεδία η εξόρυξη δεδομένων βρήκε κι άλλες εφαρμογές σε συνάρτηση με την εξέλιξης της τεχνολογίας. Έπίσης δημιουργείται η δυνατότητα να δοκιμαστεί σε νέα ερευνητικά και επαγγελματικά πεδία που όπου θα αναφέρω στη συνέχεια.

### *2.3.1 Δεδομένα επιχειρήσεων και ηλεκτρονικού εμπορίου.*

Το back office, το front office και οι εφαρμογές δικτύου παράγουν μεγάλα ποσά δεδομένων για επιχειρηματικές διαδικασίες. Η χρήση αυτών των δεδομένων για την αποτελεσματική λήψη αποφάσεων παραμένει βασική πρόκληση.

Επιστημονικά, Μηχανικά και Υγειονομικά Δεδομένα( Business & E-commerce Data).

Τα επιστημονικά δεδομένα και τα μεταδεδομένα τείνουν να είναι πιο σύνθετα στη δομή από τα δεδομένα των επιχειρήσεων. Επιπλέον, οι επιστήμονες και οι μηχανικοί χρησιμοποιούν όλο και περισσότερο την προσομοίωση και τα συστήματα με γνώση στον τομέα εφαρμογής.

### *2.3.2 Δεδομένα Δικτύου (Web Data).*

Τα δεδομένα στο διαδίκτυο αυξάνονται όχι μόνο στην ένταση, αλλά επίσης

σε πολυπλοκότητα. Τα δεδομένα Ιστού περιλαμβάνουν τώρα όχι μόνο κείμενο και εικόνα, αλλά και δεδομένα και αριθμητικά δεδομένα.

Σε αυτή την ενότητα περιγράφουμε πολλές από αυτές τις εφαρμογές από κάθε κατηγορία.

### 2.3.3 Εμπορικές συναλλαγές( Business & E-commerce Data)

Σήμερα, οι επιχειρήσεις εδραιώνονται αναπτύσοντας περισσότερες επιχειρήματικές δραστηριότητες έχοντας εκατομμύρια πελάτες και δισεκατομμύρια από τις συναλλαγές τους.Θα πρέπει να κατανοήσουν τους κινδύνους της Αγοράς.

### 2.3.4 Ηλεκτρονικό εμπόριο (Electronic commerce)

Όχι μόνο το ηλεκτρονικό εμπόριο παράγει μεγάλα σύνολα δεδομένων στα οποία η ανάλυση των προτύπων μάρκετινγκ και των προτύπων κινδύνων είναι κρίσιμη,

αλλά σε αντίθεση με κάποιες από τις παραπάνω εφαρμογές, είναι επίσης σημαντικό να γίνει αυτό σε πραγματικό ή σχεδόν πραγματικό χρόνο, προκειμένου να ικανοποιηθεί η ζήτηση των συναλλαγών on-line.

### 2.3.5 Γονιδιωματικά δεδομένα.(Genomic data)

Η γονιδιωματική αλληλουχία και οι προσπάθειες χαρτογράφησης έχουν δημιουργήσει μια σειρά βάσεων δεδομένων, τα οποία είναι προσβάσιμα μέσω του Διαδικτύου. Επιπλέον, υπάρχει επίσης μια μεγάλη ποικιλία άλλων ηλεκτρονικών βάσεων δεδομένων, συμπεριλαμβανομένων εκείνων που περιέχουν πληροφορίες σχετικά με ασθένειες, κυτταρική λειτουργία και φάρμακα. Η εύρεση σχέσης μεταξύ αυτών των πηγών δεδομένων, οι οποίες είναι σε μεγάλο βαθμό ανεξερεύνητες, είναι μια άλλη βασική πρόκληση εξόρυξης δεδομένων. Πρόσφατα αναπτύχθηκαν κλιμακούμενες τεχνικές για τη σύγκριση ολόκληρων γονιδιωμάτων.

### 2.3.6 Δορυφορικά Δεδομένα( Remote Sensing Data).

Δορυφόροι, και διάφοροι άλλοι αισθητήρες παράγουν τεράστιες ποσότητες δεδομένων σχετικά με την ατμόσφαιρα της γης, τους ωκεανούς και τα εδάφη. Μια βασική πρόκληση είναι να κατανοήσουμε τις σχέσεις, συμπεριλαμβανομένων των περιστασιακών σχέσεων μεταξύ αυτών των δεδομένων. Υπάρχουν επίσης μεγάλα σύνολα δεδομένων terabyte σε petabyte που παράγονται από αισθητήρες και όργανα σε άλλους κλάδους, όπως η αστρονομία,

τη φυσική υψηλής ενέργειας και την πυρηνική φυσική.

### 2.3.7 Δεδομένα προσομοίωσης (Simulation Data)

. Η προσομοίωση είναι τώρα αποδεκτή ως ένας τρίτος τρόπος επιστήμης, συμπληρώνοντας τη θεωρία και το πείραμα. Σήμερα, όχι μόνο τα πειράματα παράγουν σύνολα δεδομένων, αλλά και οι προσομοιώσεις. Η εξόρυξη δεδομένων και γενικότερα ο υπολογισμός έντασης δεδομένων αποδεικνύεται ότι είναι ένας κρίσιμος σύνδεσμος μεταξύ θεωρίας, προσομοίωσης και πειράματος.

### 2.3.8 Στοιχεία υγειονομικής περίθαλψης (Health care Data)

Η υγειονομική περίθαλψη ήταν ο ταχύτερα αναπτυσσόμενος τομέας του ακαθάριστου εγχώριου προϊόντος των κρατών (ΑΕΠ) εδώ και κάποιο καιρό. Τα νοσοκομεία, οι οργανισμοί υγειονομικής περίθαλψης, οι ασφαλιστικές εταιρείες και η ομοσπονδιακή κυβέρνηση έχουν μεγάλη συλλογή δεδομένων για τους ασθενείς, τα προβλήματα υγείας τους, τις κλινικές διαδικασίες που χρησιμοποιήθηκαν, το κόστος τους και τα αποτελέσματα. Η κατανόηση των σχέσεων σε αυτά τα δεδομένα είναι κρίσιμη για μια ευρεία ποικιλία προβλημάτων.(Suthami, 2006)

## 2.4 Ταξινόμηση

Η ταξινόμηση της εικόνας αποτελεί ένα σημαντικό εργαλείο για την ανάλυση των ψηφιακών εικόνων διότι βοηθάει στην εξαγωγή σημαντικής πληροφορίας από μία εικόνα (Παρχαρίδης, 2015).

Στην εργασία πραγματοποιήθηκε η επιβλεπόμενη ταξινόμηση που έχει ως σκοπό τη χρήση κατάλληλων αλγορίθμων ώστε να κατατάξει τα εικονοστοιχεία μίας απεικόνισης σε συγκεκριμένες θεματικές κατηγορίες, με χρήση των δεδομένων εκπαίδευσης. Έχοντας διαθέσιμα τα δεδομένα εκπαίδευσης που χαρακτηρίζουν την κάθε τάξη και από τα οποία θα εκτιμηθούν οι φασματικές υπογραφές πριν την εκτέλεση του αλγόριθμου ταξινόμησης, ο χρήστης κατά κάποιο τρόπο ‘‘εκπαιδεύει’’ τον αλγόριθμο να αναγνωρίζει τα φασματικά χαρακτηριστικά της κάθε κατηγορίας. Έτσι έχει επικρατήσει ο

όρος της επιβλεπόμενης ταξινόμησης. Μετά το πέρας της φάσης της εκπαίδευσης ο αλγόριθμος ταξινόμησης αποδίδει το κάθε εικονοστοιχείο στην κατάλληλη κατηγορία βάση των χαρακτηριστικών της κάθε τάξης (Λάσπιας, 2012).

Σε άλλη βιβλιογραφία, αναφέρεται και ως "ελεγχόμενη ταξινόμηση" και ακολούθως περιγράφεται ως μια διαδικασία που χρησιμοποιεί δείγματα γνωστής ταυτότητας με σκοπό την ταξινόμηση των εικονοστοιχείων, των οποίων δεν έχει προσδιοριστεί η ταυτότητα. Τα δείγματα λαμβάνονται από περιοχές δειγματοληψίας που καθορίζει ο αναλυτής και συνήθως οριοθετούνται με ψηφιοποίηση επάνω στην εικόνα. Οι περιοχές αυτές πρέπει να έχουν γνωστή ταυτότητα και να εμπεριέχουν μονάχα ένα χαρακτηριστικό. Τα εικονοστοιχεία που βρίσκονται μέσα σε αυτές τις περιοχές και τα οποία χρησιμοποιούνται για την ελεγχόμενη ταξινόμηση είναι οι οδηγοί που θα χρησιμοποιηθούν από τον αλγόριθμο ταξινόμησης (Παρχαρίδης, 2015).

Η ταξινόμηση είναι η διαδικασία εύρεσης ενός μοντέλου (ή συνάρτησης) που περιγράφει και διακρίνει κατηγορίες δεδομένων ή έννοιες. Το μοντέλο προκύπτει με βάση την ανάλυση ενός συνόλου δεδομένων εκπαίδευσης.

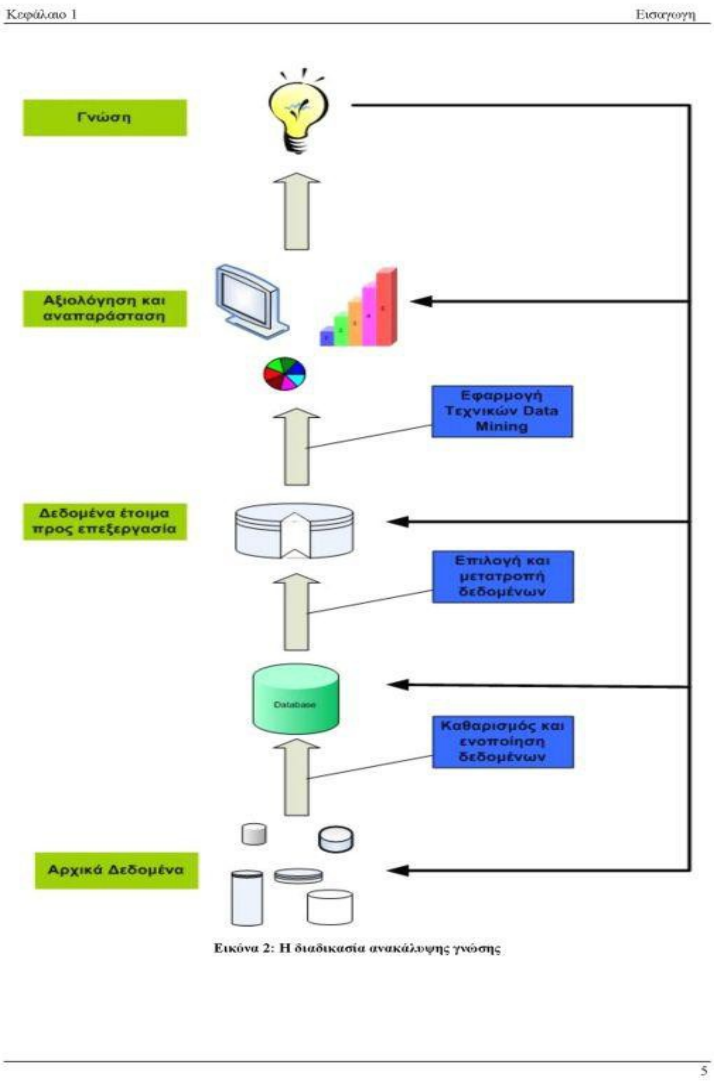
Το μοντέλο χρησιμοποιείται για την πρόβλεψη της ετικέτας κλάσης αντικειμένων, για τα οποία η ετικέτα κλάσης είναι άγνωστη. Το προερχόμενο μοντέλο μπορεί να εκπροσωπείται σε διάφορες μορφές, όπως κανόνες ταξινόμησης , δέντρα αποφάσεων, μαθηματικοί τύποι ή νευρωνικά δίκτυα.

Ένα δέντρο απόφασης (decision trees) είναι μια δομή δέντρου ροής, όπου κάθε κόμβος δηλώνει μια δοκιμή σε μια τιμή χαρακτηριστικού, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα της δοκιμής και τα φύλλα δέντρων αντιπροσωπεύουν κατηγορίες ή διανομές τάξεων. Τα δέντρα αποφάσεων μπορούν εύκολα να μετατραπούν σε κανόνες ταξινόμησης. Ένα νευρωνικό δίκτυο, όταν χρησιμοποιείται για ταξινόμηση, είναι συνήθως μια συλλογή μονάδων επεξεργασίας με σταθμισμένες συνδέσεις μεταξύ των μονάδων. Υπάρχουν πολλές άλλες μέθοδοι για την κατασκευή μοντέλων ταξινόμησης, με κυριότερες να αποτελλούν οι Naive Bayesian , Support Vector Machines, and K-Nearest-Neighbor (Jiamei et.al, 2012).

## 2.5 Διαδικάσία Data Mining

Η διαδικασία του Data Mining (Εξόρυξης Δεδομένων) αποτελείται από μια επαναληπτική ακολουθία των παρακάτω βημάτων:

1. Καθαρισμός των δεδομένων (Data Cleaning), όπου ουσιαστικά απομακρύνουμε τον θόρυβο και τα ακατάλληλα δεδομένα.
2. Ενοποίηση των δεδομένων (Data Integration), όπου πιθανώς να έχουμε πολλαπλές πηγές δεδομένων οι οποίες θα πρέπει να συνδυαστούν.
3. Επιλογή δεδομένων (Data Selection), όπου δεδομένα σχετικά με τη διαδικασία της ανάλυσης μας θα πρέπει να επιλεχθούν και να ανακτηθούν από τη βάση μας.
4. Μετατροπή των δεδομένων (Data Transformation), όπου τα δεδομένα μας θα πρέπει να μετατραπούν σε μία ενιαία μορφή κατάλληλη προς επεξεργασία.
5. Εξόρυξη δεδομένων (Data Mining), μια διαδικασία που εφαρμόζονται ευφυείς μέθοδοι προκειμένου να εξάγουμε μοτίβα-πρότυπα από τα δεδομένα μας.
6. Αξιολόγηση μοτίβων (Pattern Evaluation), η διαδικασία κατά την οποία αναγνωρίζουμε και ξεχωρίζουμε τα πραγματικά ενδιαφέροντα μοτίβα με χρήση μετρικών ενδιαφέροντος.
7. Αναπαράσταση γνώσης (Knowledge Presentation), όπου εφαρμόζουμε τεχνικές οπτικοποίησης και αναπαράστασης γνώσης προκειμένου να παρουσιάσουμε καλύτερα την εξαγόμενη γνώση στους χρήστες. (Κουρής, 2006).




***Εικόνα 1***. *Διαδικασία Εξόρυξης Δεδομένων*

## 2.6 Bagging Boosting Stacking

Οι παραδοσιακές μέθοδοι μοντελοποίησης μέσω στατιστικής και εκμάθησης μηχανών, συμπεριλαμβανομένης της γραμμικής παλινδρόμησης, της λογιστικής παλινδρόμησης, της διακριτικής ανάλυσης και των μοντέλων Bayes, είναι συχνά τα πρώτα εργαλεία που χρησιμοποιούνται για τον μοντελοποίηση δεδομένων πολλαπλών μεταβλητών. Τα νεότερα προγνωστικά μοντέλα, συμπεριλαμβανομένης της παλινδρόμησης , των δέντρων αποφάσεων, των νευρωνικών δικτύων, support vector machines και των δικτύων Bayesian , έχουν προσελκύσει την έρευνα και τις εφαρμογές εξόρυξης δεδομένων, καθώς η σύγχρονη υπολογιστική ισχύς επέτρεψε στους ερευνητές να εξερευνήσουν πιο πολύπλοκα μοντέλα.

Σε αρκετές περιπτώσεις οι ατομικοί ταξινομητές παρουσιάζουν μια καλή ακρίβεια ωστόσο . υπάρχουν και περιπτώσεις όπου η ακρίβεια δεν κρίνεται ικανοποιητική. Έτσι δημιουργήθηκε η ανάγκη για την ανάπτυξη μεθόδων που συνδυάζουν τους ταξινομητές για να παράξουν καλύτερα αποτελέσματα. . Οι πιο γνωστές μέθοδοι είναι οι “ensemble methods” ή μέθοδοι συνόλου ή “από κοινού” μέθοδοι ταξινομητών. Συνδυάζουν ταξινομητές εκπαιδευόμενους στο ίδιο σύνολο δεδομένων (Rokach,2009)

Πρόσφατες έρευνες έχουν δείξει ότι ο συνδυασμός διαφορετικών μοντέλων μπορεί να είναι πιο αποτελεσματικός από ένα ενιαίο μοντέλο που ταιριάζει σε ένα ενιαίο σύνολο δεδομένων. Μια ποικιλία επιλογών εμφανίζεται με το συνδυασμό μοντέλων τα οποία αναφέρονται ως *Bagging, boosting και stacking* και συνδυάζουν μια ισχυρή μέθοδο υπολογιστικής αναζήτησης. για την κατασκευή πολυπαραγοντικών προγνωστικών μοντέλων (Suthami, 2006).

### *2.6.1 Bagging*

Το όνομα Bagging προήλθε από τη συντομογραφία του Bootstrap AGGregatING (Breiman, 1996). Όπως υποδηλώνει το όνομα, τα δύο βασικά συστατικά του Bagging είναι bootstrap και aggregation.

Η Aggregation Bootstrap (ή Bagging), είναι μια απλή και πολύ ισχυρή μέθοδος ανάλυσης.

Είναι μια τεχνική που συνδυάζει τις προβλέψεις από πολλούς αλγορίθμους μηχανικής μάθησης μαζί για να κάνει πιο ακριβείς προβλέψεις από οποιοδήποτε μεμονωμένο μοντέλο.

Η Bootstrap Aggregation είναι μια γενική διαδικασία που μπορεί να χρησιμοποιηθεί για τη μείωση της διακύμανσης για τους αλγόριθμους που χαρακτηριζονται απο μεγάλη διακύμανση. Ένας αλγόριθμος που έχει μεγάλη διακύμανση είναι τα δέντρα απόφασης, όπως τα δέντρα ταξινόμησης και παλινδρόμησης (CART), όπου θα μιλήσουμε και στη συνέχεια.

Η μέθοδος αυτή ανήκει στην κατηγορία των τεχνικών όπου βασίζονται σε ένα αλγόριθμο εξόρυξης γνώσης και χρησιμοποιούν επαναληπτική δειγματοληψία (με επανατοποθέτηση) στα δεδομένα εκπαίδευσης για να παράγουν μια ομάδα ταξινομητών. Η κατηγορία αυτή εκμεταλλεύεται την αστάθεια που παρουσιάζουν ορισμένοι αλγόριθμοι στις μικρές αλλαγές στα δεδομένα εκπαίδευσης. ΄Σύμφωνα με τη μέθοδο του bagging η διαδικασία που ακολοθείτε είναι η εξής:

Αρχικά χωρίζουμε το σύνολο δεδομένων σε t ίσα σύνολα με τυχαία επιλογή και επανατοποθέτηση. Τα σύνολα αυτά μπορεί να είναι ίδια, παρόμοια ή και τελείως διαφορετικά. Στη συνέχεια σε κάθε υποσύνολο που δημιουργήθηκε εφαρμόζουμε τον αλγόριθμο και έτσι έχουμε ένα σύνολο ταξινομητών. Συνδυάζοντας τα αποτελέσματα θα έχουμε την τελική εκτίμηση. Ο τρόπος συνδυασμού εξαρτάται από τις εξόδους των ταξινομητών. Αν οι έξοδοι είναι συνεχείς τότε παίρνουμε ως αποτέλεσμα τον μέσο όρο των εξόδων. Σε περίπτωση που έχουμε επιγραφές κλάσεων το τελικό αποτέλεσμα συμπίπτει με την απόφαση της πλειοψηφίας (Zhi-Hua Zhou, 2012)

**Input:** Data set $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
Base learning algorithm $\mathfrak{L}$;
Number of base learners $T$.

**Process:**
1. **for** $t = 1, \ldots, T$:
2. $h_t = \mathfrak{L}(D, \mathcal{D}_{bs})$ % $\mathcal{D}_{bs}$ is the bootstrap distribution
3. **end**

**Output:** $H(\boldsymbol{x}) = \underset{y \in \mathcal{Y}}{\arg\max} \sum_{t=1}^{T} \mathbb{I}(h_t(\boldsymbol{x}) = y)$

***Εικόνα 2****.Γενική διαδικασία bagging*

### 2.6.2 Boosting

Ο όρος Boosting (ενίσχυση) αναφέρεται σε μια οικογένεια αλγορίθμων που είναι ικανοί να μετατρέψουν αδύναμους εκπαιδευτές σε δυνατούς.Ο αδύνατος εκπαιδευτής πρέπει να είναι ελαφρώς καλύτερος από την τυχαία επιλογή. Η δημιουργία των αλγορίθμων ενίσχυσης προέκυψε απο την απαντηση των Kearns και Valiant(1989) σε ένα θεωρητικό ερώτημα. Αυτό ήταν ένας αδυναμος εκπαιδευτής και ένας δυνατός εκπαιδευτής μπορούν να είναι ίσοι. Αυτό το ερώτημα είναι θεμελιώδους σημασίας, αφού εάν η απάντηση είναι θετική, κάθε αδύναμος εκπαιδευτής μπορεί δυνητικά να ενισχυθεί σε έναν ισχυρό εκπαιδευτή. Ο Schapire [1990] απέδειξε ότι η απάντηση είναι θετική και ότι η απόδειξη είναι η μέθοδος Boosting δηλαδή η ενίσχυση. Επι της ουσίας, η ιδέα του boosting είναι να αυξήσει τη δύναμη ενός αδύναμου αλγορίθμου εκμάθησης. Το boosting, εκπαιδεύει έναν αδύναμο εκπαιδευτή αρκετές φορές, χρησιμοποιώντας μια ανανεωμένη έκδοση του αρχικού συνόλου εκπαίδευσης. Ετσι, εκπαιδεύει τον πρώτο αδύναμο εκπαιδευτή με το ίδιο βάρος σε όλα τα σημεία δεδομένων του συνόλου εκπαίδευσης, στη συνέχεια εκπαιδεύει όλους τους άλλους αδύναμους εκπαιδευτές με βάση τα νέα βάρη που θα αποδώσει. Τα λάθος ταξινομημένα δεδομένα, απο τους αδύναμους εκπαιδευτές, παίρνουν μεγαλύτερο βάρος, και τα σωστά ταξινομημένα σημεία δεδομένων παίρνουν ελαφρύτερο. Με αυτόν τον τρόπο, ο επόμενος εκπαιδευτής θα προσπαθήσει να διορθώσει τα σφάλματα που κάνει ο προηγούμενος εκπαιδευτής.

**Input:** Sample distribution $\mathcal{D}$;
Base learning algorithm $\mathfrak{L}$;
Number of learning rounds $T$.
**Process:**
1. $\mathcal{D}_1 = \mathcal{D}$. % Initialize distribution
2. **for** $t = 1, \ldots, T$:
3. $h_t = \mathfrak{L}(\mathcal{D}_t)$; % Train a weak learner from distribution $\mathcal{D}_t$
4. $\epsilon_t = P_{\boldsymbol{x} \sim \mathcal{D}_t}(h_t(\boldsymbol{x}) \neq f(\boldsymbol{x}))$; % Evaluate the error of $h_t$
5. $\mathcal{D}_{t+1} = Adjust\_Distribution(\mathcal{D}_t, \epsilon_t)$
6. **end**

**Output:** $H(\boldsymbol{x}) = Combine\_Outputs(\{h_1(\boldsymbol{x}), \ldots, h_t(\boldsymbol{x})\})$

***Εικόνα 3****.Γενική διαδικασία Boosting*

### 2.6.3 Stacking

Τέλος η μέθοδος που θα δούμε ειναι η μέθοδος stacking ή η μέθοδος της συσσωρευμένης γενίκευσης (stacked generalization). Αποτελεί μια νεα σχετικά προσέγγιση για τον συνδυασμό των ταξινομητών και η οποία προτάθηκε από τον David Wolpert. Παρότι χρησιμοποιείται τα τελευταία χρόνια δεν είναι το ίδιο γνωστή μέθοδος όσο αποτελλούν οι παραπάνω μέθοδοι boosting και bagging καθώς δεν υπάρχει ένας κοινά αποδεκτός καλύτερος τρόπος να την χρησιμοποιήσουμε και επίσης είναι δύσκολο να αναλυθεί θεωρητικά. Ανήκει στην κατηγορία των μεθόδων που συνδυάζουν περισσότερους από έναν αλγόριθμους ταξινόμησης για την δημιουργία του μοντέλου.(Witten & Eibe,2005)

,Η μέθοδος stacking είναι μια γενική μέθοδος χρήσης ενός μοντέλου υψηλού επιπέδου για το συνδυασμό μοντέλων χαμηλότερου επιπέδου για την επίτευξη μεγαλύτερης προγνωστικής ακρίβειας(Raschka, 2019)

**Input:** Data set $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
First-level learning algorithms $\mathfrak{L}_1, \ldots, \mathfrak{L}_T$;
Second-level learning algorithm $\mathfrak{L}$.
**Process:**
**for** $t = 1, \ldots, T$: % Train a first-level learner by applying the
$h_t = \mathfrak{L}_t(D)$; % first-level learning algorithm $\mathfrak{L}_t$
**end**
$D' = \emptyset$; % Generate a new data set
**for** $i = 1, \ldots, m$:
**for** $t = 1, \ldots, T$:
$z_{it} = h_t(\boldsymbol{x}_i)$;
**end**
$D' = D' \cup ((z_{i1}, \ldots, z_{iT}), y_i)$;
**end**
$h' = \mathfrak{L}(D')$; % Train the second-level learner $h'$ by applying
% the second-level learning algorithm $\mathfrak{L}$ to the
% new data set $D'$.
**Output:** $H(\boldsymbol{x}) = h'(h_1(\boldsymbol{x}), \ldots, h_T(\boldsymbol{x}))$

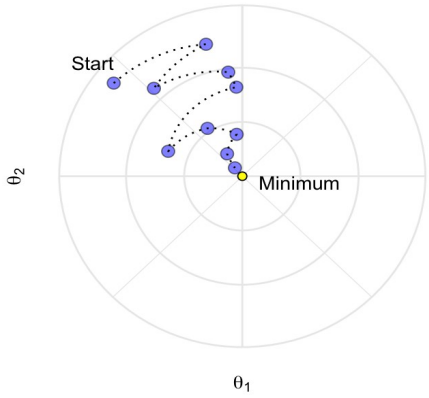
***Εικόνα 4****.Γενική διαδικασία Stacking*

## 2.7 Αλγόριθμοι που Χρησιμοποιήθηκαν

Σε αυτή την υποενότητα θα ακολουθήσει μια μικρή επεξήγηση των αλγόριθμων που χρησιμοποιήθηκαν κατα τη διαδικασία της ταξινόμησης, όπως επίσης και στη διαδικασία εκτίμησης της ακρίβεια των μοντέλων που αναπαράχθηκαν.

### 2.7.1 Stochastic Gradient Boosting (GBM)

Με τη διαδικασία του Boosting, ο Breiman(1996) εισήγαγε τον παράγοντα της τυχαιότητας στις διαδικασίες εκτίμησης ώστε να βελτιώση την απόδοση των αποτελεσμάτων. Επίσης με μια σημαντική διορατική σκέψη που έκανε ο Breiman (1996,2001) στην ανάπτυξη του bagging και του αλγόριθμου Random forest (που θα μιλήσουμε στη συνέχεια) ήταν ότι η εκπαίδευση του αλγορίθμου σε ένα τυχαίο υπο-δείγμα του συνόλου δεδομένων εκπαίδευσης προσέφερε πρόσθετη μείωση της συσχέτισης των δέντρων και ως εκ τούτου, βελτίωση της ακρίβειας της πρόβλεψης. Ο Freidman (2002) χρησιμοποίησε την ίδια λογική και ενημέρωσε τον αλγόριθμο Boosting με την ίδια λογική. Αυτή η διαδικασία είναι γνωστή ως Stochastic Gradient Boosting και, όπως απεικονίζεται στο σχήμα παρακάτω, βοηθά στη μείωση των πιθανοτήτων να κολλήσει σε τοπικό ελάχιστο επίπεδο (όπως οροπέδια και οποιοδήποτε άλλο ακανόνιστο έδαφος) έτσι ώστε να μπορούμε να φτάσουμε στο βέλτιστο αποτέλεσμα.
(*Boehmke & Greenwell,2019* )




***Εικόνα 5****.Stochastic Gradient Boosting πηγη:*
*https://bradleyboehmke.github.io/HOML/gbm.html#stochastic-gbms*

### 2.7.2 K-nearest neighbor (KNN)

Ο Κ-κοντινότερος γείτονας (KNN) είναι ένας πολύ απλός αλγόριθμος στον οποίο κάθε παρατήρηση προβλέπεται βάσει της «ομοιότητας» της με άλλες παρατηρήσεις. Το KNN είναι ένας αλγόριθμος που βασίζεται στη μνήμη και δεν είναι δυνατό να συνοψιστεί από ένα μοντέλο κλειστής φόρμας. Αυτό σημαίνει ότι τα δείγματα εκπαίδευσης πραγματοποιούνται κατά το χρόνο εκτέλεσης και οι προβλέψεις γίνονται

απευθείας από τις σχέσεις του δείγματος. Κατά συνέπεια, οι KNNs είναι επίσης γνωστοί ως τεμπέληδες μαθητές (Cunningham και Delany 2007) και μπορεί να είναι υπολογιστικώς αναποτελεσματικοί.

Τα KNNs είναι ένας πολύ απλοϊκός, και διαισθητικός αλγόριθμος, ο οποίος μπορεί να παρέχει μέσο όρο στην αξιοπρεπή προγνωστική ισχύ, ειδικά όταν η απόκριση εξαρτάται από την τοπική δομή των χαρακτηριστικών. Ωστόσο, ένα σημαντικό μειονέκτημα των KNNs είναι ο χρόνος υπολογισμού τους, ο οποίος αυξάνεται με $n \times p$ για κάθε παρατήρηση.

Οι και τα KNNs σπάνια παρέχουν την καλύτερη προγνωστική απόδοση, έχουν πολλά οφέλη. για παράδειγμα, στη μηχανική χαρακτηριστικών και στον καθαρισμό και την προεπεξεργασία δεδομένων.

$$\sqrt{\sum_{j=1}^{P}(x_{aj} - x_{bj})^2}$$

$$\sum_{j=1}^{P}|x_{aj} - x_{bj}|$$

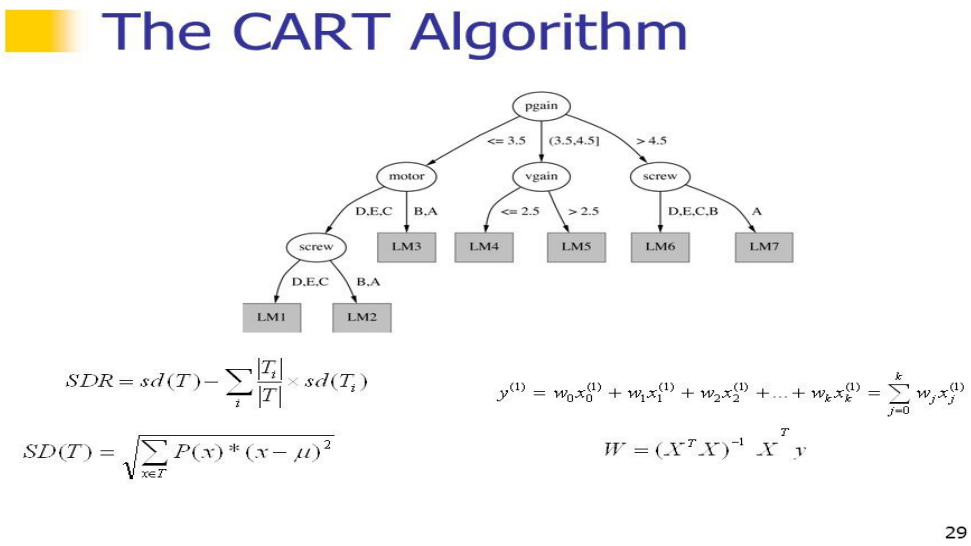
***Εικόνα 6.*** *Υπολογισμός Ευκλείδιας απόσταση κ Μανχάταν*

### *2.7.3 Classification and Reggresion Trees(CART)*

Υπάρχουν πολλές μεθοδολογίες για την κατασκευή δέντρων αποφάσεων, αλλά η πιο γνωστή είναι η ταξινόμηση και το δέντρο παλινδρόμησης (CART) αλγόριθμος, που προτείνεται απο τον Breiman (1984). Ένα τυπικό δέντρο απόφασης χωρίζει τα δεδομένα εκπαίδευσης σε ομοιογενείς υπο-ομάδες και στη συνέχεια ταιριάζει μια απλή σταθερά σε κάθε υπο-ομάδα . Οι υπο-ομάδες σχηματίζονται χρησιμοποιώντας δυαδικά χωρίσματα που σχηματίζονται ρωτώντας απλές ερωτήσεις ναι ή όχι για κάθε χαρακτηριστικό. Αυτό γίνεται αρκετές φορές μέχρι να ικανοποιηθεί μια κατάλληλη συνθήκη. Μετά από όλο το διαμοιρασμό που έχει γίνει, το μοντέλο προβλέπει την παραγωγή χρησιμοποιώντας τις μέσες τιμές απόκρισης για όλες τις παρατηρήσεις που εμπίπτουν σε αυτή την υποομάδα (πρόβλημα παλινδρόμησης), ή η κλάση που έχει

πλειοψηφία (πρόβλημα ταξινόμησης). Για την ταξινόμηση, οι προβλεπόμενες πιθανότητες μπορούν να ληφθούν με τη χρήση του ποσοστού κάθε κλάσης εντός της υπο-ομάδας.

Τα πλεονεκτήματα των Δέντρων Απόφασης απαιτούν, ελάχιστη προ-επεξεργασία και λειτουργουν με ευκολία τη διαδικασία της ταξινόμησης. Επίσης οι ακραίες τιμές δεν επηρεάζουν συνήθως το αποτέλεσμα καθώς ο δυαδικός διαχωρισμός απλά αναζητά μια μεμονωμένη θέση για να κάνει μια διαίρεση εντός της κατανομής κάθε χαρακτηριστικού (Boehmke & Greenwell, 2019).




***Εικόνα 7****. Ο αλγόριθμος Classification and Regression Trees (CART)*

*Πηγή:https://slideplayer.com/slide/5121894/, Hong Kong University of Science and Technology*

### *2.7.4 Random Forest (RF)*

Η Random Forest (Τυχαίο Δάσος), είναι μια συνδυαστικη μέθοδος ταξινόμησης που προτάθηκε από την Breiman το 2001. Χρησιμοποιώντας τη μέθοδο bagging, η Random Forest θα αντλήσει πολλαπλά σύνολα δειγμάτων εκπαίδευσης που είναι διαφορετικά μεταξύ τους.

Κάθε σύνολο δειγμάτων δημιουργεί ένα δέντρο αποφάσεων με τυχαία επιλεγμένα χαρακτηριστικά.

.

Το τυχαίο δάσος χρησιμοποιεί αλγόριθμο CART για την κατασκευή δέντρων. Λαμβάνοντας υπόψη τον μεγάλο αριθμό δομημένων δένδρων, η Random Forest χαρακτηρίζεται από καλή ικανότητα να αντισταθεί στον θόρυβο και παρουσιάζει εξαιρετικές επιδόσεις στην ταξινόμηση.

Η Random forest ορίζεται ως ένα σύνολο δέντρων αποφάσεων, {h (x, θk), k = 1, ...}, όπου το h (x, θk) είναι μετα-ταξινομητής, δηλαδή ένα μη επεξεργασμένο δέντρο αποφάσεων που δημιουργήθηκε χρησιμοποιώντας αλγόριθμο CART . Το x χρησιμεύει ως διάνυσμα εισόδου, ενώ {θk} είναι ανεξάρτητος και κατανεμημένος πανομοιότυπα τυχαίος φορέας.

Επίσης, καθορίζουν τη διαδικασία ανάπτυξης κάθε δέντρου αποφάσεων.

Σύμφωνα με τα δεδομένα εισόδου, κάθε δέντρο απόφασης θα δώσει ένα αποτέλεσμα. η ενσωμάτωση πολλαπλών αποτελεσμάτων θα δώσει την τελική παραγωγή ενός τυχαίου δάσους. Στο τυχαίο δάσος, η διαδικασία ανάπτυξης ενός ενιαίου δέντρου αποφάσεων έχει ως εξής:

1. Για τα αρχικά σύνολα εκπαίδευσης, η μέθοδος bagging

χρησιμοποιείται για την επιλογή τυχαίων δεδομένων με αντικατάσταση

2. Τα χαρακτηριστικά επιλέγονται με δειγματοληψία. Αν υποτεθεί ότι ένα σύνολο δεδομένων έχει χαρακτηριστικά N, τότε τα χαρακτηριστικά M θα δειγματοληφθούν από το N, όπου M << N.

Για κάθε εξαγόμενο σύνολο εκπαίδευσης, μόνο τα τυχαία επιλεγμένα χαρακτηριστικά M και όχι όλα τα χαρακτηριστικά N θα χρησιμοποιηθούν για τον διαχωρισμό των κόμβων, στην κατασκευή δένδρων.

3. Όλα τα δέντρα απόφασης που έχουν δημιουργηθεί θα αναπτυχθούν ελεύθερα χωρίς

κλάδεμα. Το τελικό αποτέλεσμα μπορεί να ενσωματωθεί χρησιμοποιώντας τη μεθοδος της πλειοψηφίας (για προβλήματα ταξινόμησης) ή με μαθηματικές μεθόδους μεταξύ των αποτελεσμάτων των δέντρων αποφάσεων. (Miner et al,2009)

Πλεονεκτήματα

Ο Breiman πρότεινε έναν νέο και πολλά υποσχόμενο ταξινομητή που ονομάζεται τυχαίο δάσος (Random Forest), το οποίο παρουσιάζει πολλά πλεονεκτήματα για την εφαρμογή του στην τηλεπισκόπηση:

•Λειτουργεί αποτελεσματικά σε μεγάλες βάσεις δεδομένων.

•Μπορεί να χειριστεί χιλιάδες μεταβλητές εισόδου χωρίς μεταβλητή διαγραφή.

•Παρέχει εκτιμήσεις για ποιες μεταβλητές είναι σημαντικές στην ταξινόμηση.

•Δημιουργεί μια εσωτερική αμερόληπτη εκτίμηση του σφάλματος γενίκευσης (σφάλμα oob).

•Υπολογίζει τις εγγύτητες μεταξύ ζευγών περιπτώσεων που μπορούν να χρησιμοποιηθούν για τον εντοπισμό των ακραίων τιμών.

•Είναι σχετικά ανθεκτικό στις υπερβολικές τιμές..

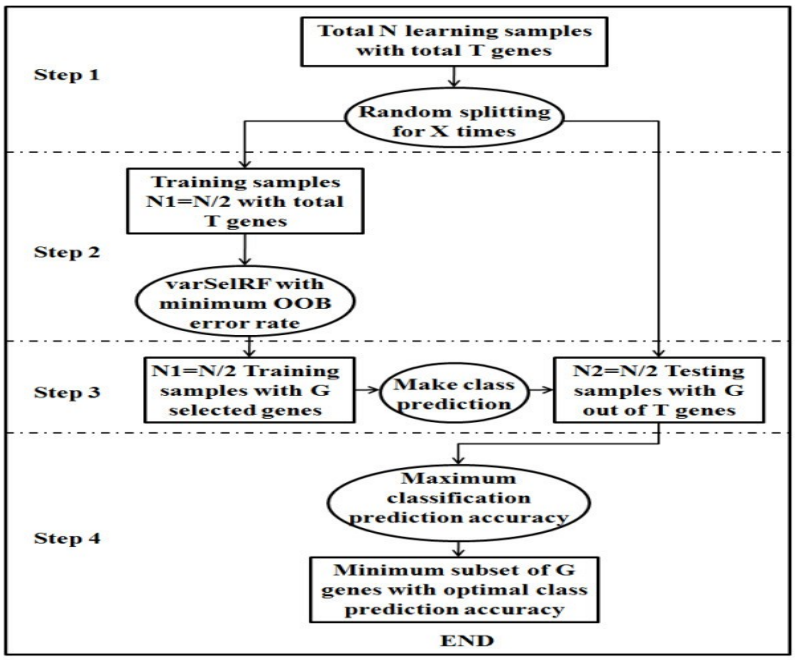
•Είναι υπολογιστικά ελαφρύτερο από άλλες μεθόδους συνόλων δέντρων. (Yingchun, 2014)

### *2.7.5 Bagged Cart*

Είναι ο συνδυασμός της μεθόδου Bagging, για τον οποίο έγινε ανφορά στην προηγούμενη υπο-ενότητα, με τον αλγόριθμο CART που επίσης αναφέραμε.

Το Bagging είναι μια μέθοδος που μπορεί χρησιμοποιηθεί για τη μείωση της διακύμανσης ενός αλγορίθμου μεγάλης διακύμανσης. Τα δέντρα αποφάσεων, όπως ο αλγόριθμος CART, αποτελεί μία τέτοια περίπτωση. Τα δέντρα αποφάσεων είναι ευαίσθητα στα δεδομένα για τα οποία εκπαιδεύονται. Εάν τα δεδομένα εκπαίδευσης

αλλάξουν, το δέντρο απόφασης που προκύπτει μπορεί να είναι αρκετά διαφορετικό και με τη σειρά του και οι προβλέψεις. Τα δέντρα απόφασης, τείνουν να αναπτύσονται περισσότερο, χωρίς κλάδεμα, για να αποφύγουν την υπερ-προσαρμογή (Overfitting) στο μοντέλο. Έτσι τα δέντρα απόφασης θα έχουν μεγάλη διακύμανση και χαμηλή προ-κατάληψη (Brownlee, 2016).




***Εικόνα 8***. *Ο αλγόριθμος του Τυχαίου Δάσους (Random Forest)*
*Πηγή: https://www.researchgate.net/figure/Flowchart-of-the-Splitting-Random-Forest-SRF-Algorithm_fig1_225055544*

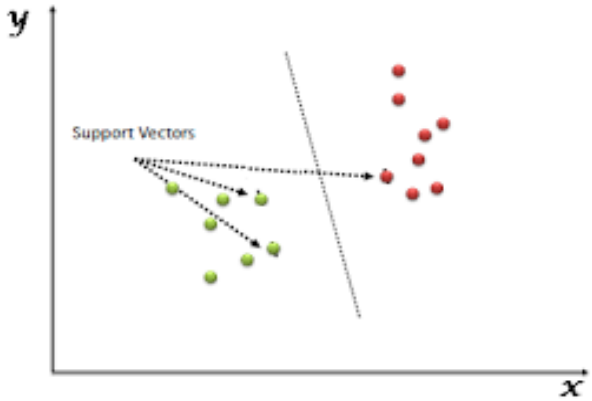
### *2.7.6 Linear Discriminant Analysis (LDA)*

Η Γραμμική Διακριτική Ανάλυση είναι, μια μέθοδος που χρησιμοποιείται στην στατιστική, στην αναγνώριση μοτίβων και την εκμάθηση μηχανών για να βρει έναν γραμμικό συνδυασμό χαρακτηριστικών που χαρακτηρίζουν ή διαχωρίζουν δύο ή περισσότερες κατηγορίες αντικειμένων. Αυτή η μέθοδος, προσφέρει ένα σύνολο δεδομένων σε μικρότερη διάσταση με έναν καλό διαχωρισμό, ώστε να αποφεύγεται η υπερτοποθέτηση και να μειώνονται τα υπολογιστικά σφάλματα. Ο συνδυασμός που προκύπτει μπορεί να χρησιμοποιηθεί ως μια γραμμική ταξινόμηση.

Με λίγα λόγια, η Γραμμική Διακριτική Ανάλυση (LDA) αποτελείται απο 3 στάδια Στο πρώτο στάδιο είναι ο υπολογισμός του διαχωρισμού μεταξύ διαφορετικών κλάσεων που ονομάζεται επίσης και διακύμανση μεταξύ κατηγοριών, στο δεύτερο στάδιο, είναι να υπολογιστεί η απόσταση μεταξύ του μέσου όρου και του δείγματος κάθε κλάσης, η οποία ονομάζεται διακύμανση εντός της κατηγορίας και τέλος στο τρίτο στάδιο είναι η κατασκευή του χώρου με τις μικρότερες διαστάσεις που μεγιστοποιεί τη διακύμανση μεταξύ κατηγοριών και ελαχιστοποιεί την διακύμανση εντός της κλάσης (Sawla, 2018)

### 2.7.7 Support Vector Machine( SVM)

Οι μηχανές διανυσμάτων υποστήριξης (SVMs) είναι μία ομάδα αλγορίθμων επιτηρούμενης μάθησης που αρχικά χρησιμοποιήθηκαν για την κατηγοριοποίηση ενώ αργότερα εφαρμόστηκαν και σε προβλήματα παλινδρόμησης. Η κατηγοριοποίηση των δεδομένων στηρίζεται στην εύρεση ενός βέλτιστου υπερεπιπέδου που διαχωρίζει τα δεδομένα δημιουργώντας το μέγιστο περιθώριο. Στην περίπτωση που ο γραμμικός διαχωρισμός είναι αδύνατος, γίνεται χρήση κατάλληλων απεικονίσεων που μεταφέρουν το σύνολο των δεδομένων σε μεγαλύτερη διάσταση ώστε να επιτευχθεί τελικά ο διαχωρισμός τους. Η ικανότητα γενίκευσης της χρήσης των SVM σε μη γραμμικά δεδομένα στηρίζεται στο τέχνασμα του πυρήνα (kernel trick). Κάθε μηχανή διανυσμάτων υποστήριξης είναι ένας δυαδικός ταξινομητής, έχει δηλαδή τη δυνατότητα κατηγοριοποίησης σε δύο κλάσεις. Εάν οι κλάσεις είναι περισσότερες, τότε κρίνεται απαραίτητη η χρήση περισσότερων μηχανών διανυσμάτων υποστήριξης και η εφαρμογή διάφορων τεχνικών που θα αναλυθούν(Παπαποστόλου, 2017).




***Εικόνα 9.****Ο αλγόριθμος Supprt Vector Machine (SVM).*
*Πηγή https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/*

### 2.7.8 C5.0

Ο αλγόριθμος C5.0 έδραιωθεί ως ένας βασικός αλγόριθμος για την παραγωγή δέντρων αποφάσεων, επειδή μπορεί να διαχειριστεί πολυ καλά διαφορετικούς τύπους προβλημάτων λειτουργώντας με πολυ διαφορετική προσέγγιση. Σε σύγκριση με άλλα προηγμένα μοντέλα μηχανικής μάθησης, τα δέντρα απόφασης που κατασκευάστηκαν από τον αλγόριθμο C 5.0 γενικά εκτελούν σχεδόν εξίσου καλά και είναι πιο εύκολα στην κατανόηση και την ανάπτυξή τους.

Υπάρχουν διάφοροι τρόποι με τους οποίους μπορεί να επιτευχθεί ένας διαχωρισμός στα δέντρα αποφάσεων. Ο αλγόριθμος C5.0 χρησιμοποιεί την εντροπία με στόχο τη μείωση της ώστε να αυξήση την ομοιογένεια εντός των συνόλων.

Ο μαθηματικός τύπος της εντροπίας έχει ώς εξής :

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \; log_2(p_i)$$

***Εικόνα 10.*** *Μαθηματικός τύπος, Εντροπία*

Τα πλεονεκτήμα του αλγόριθμου ειναι η οτι, μπορεί να λειτουργήσει σαν ταξινομητής πολλών διαφορετικών προβλημάτων, διαθέτει υψηλές δυνατότες αυτόματης μάθησης η οποία μπορεί να χειριστεί ονομαστικά, αριθμητικά καθώς και δεδομένα που λείπουν, εξαιρεί ασήμαντα χαρακτηριστικά, μπορεί να χρησιμοποιηθεί για μεγάλα ή μικρά σύνολα δεδομένων, μπορεί να εξάγει αποτελέσματα χωρίς μαθηματικό υπόβαθρο και τέλος μπορεί να είναι πιο αποτελεσματικό σε σχέση με πιο σύνθετα μοντέλα.

Στο αντίποδα, μειονεκτήματα αποτελλούν, ότι τα δέντρα αποφάσεων που αναπαράγονται απο τον αλγόριθμο, ενδέχεται να μεροληπτούν σε χαρακτηριστικά που αντιστοιχούν σε μεγάλο αριθμό επιπέδων, υπερπροσαρμόζεται( Overfitting) ή υποπροσαρμόζεται (Underfitting) εύκολα, μπορεί να έχει πρόβλημα στη μοντελοποίηση κάποιων σχέσεων λόγω της σχέσης των παράλληλων διαιρέσεων με τον άξονα, οι μικρές αλλαγές στα δεδομένα εκπαίδευσης μπορούν να οδηγήσουν σε μεγάλες αλλαγές στη λογική

της απόφασης και τέλος τα μεγάλα δέντρα απόφαση που αναπαράγονται μπορεί να έχουν μια δυσκολία στην ερμηνεία και να είναι δύσχτηστα (Packt editorial staff, 2019).

## 2.8 Αξιολόγηση Αλγορίθμων: Μετρικές Accuracy και Kappa

Η εκτίμηση της ακρίβειας (Accuracy Assessment) είναι μια διαδικασία για να προσδιοριστεί ποσοτικά η εγγυρότητα ενός ταξινομητή, δηλαδή πόσο ακριβής είναι η ταξινόμηση. Η εκτίμηση της ακρίβειας είναι ένα σημαντικό μέρος της ταξινόμησης και πραγματοποιείται συνήθως με τη σύγκριση του αποτελέσματος ταξινόμησης με κάποια δεδομένα αναφοράς που θεωρείται ότι αντικατοπτρίζουν την πραγματική κάλυψη της γης με ακρίβεια (Brownlee, 2016).

$$Accuracy = \frac{Correct\ Classified\ Pixels}{Total\ of\ Pixels}$$

Οι τιμές Kappa, είνα ένα στατιστικό μέτρο σύγκρισης μεταξύ των προβλέψεων και των πραγματικών . Μπορεί επίσης να ερμηνευτεί ως σύγκριση της συνολικής ακρίβειας με την αναμενόμενη ακρίβεια τυχαίας πιθανότητας. Όσο υψηλότερη είναι η μέτρηση Kappa, τόσο καλύτερα ο ταξινομητής σας συγκρίνεται με έναν τυχαίο τυχαίο δείκτη. Η διαίσθηση πίσω από την στατιστική Kappa είναι η ίδια με τις μετρήσεις τυχαίων εικασιών που μόλις συζητήσαμε. Ωστόσο, η αναμενόμενη ακρίβεια που χρησιμοποιείται για τον υπολογισμό του Kappa βασίζεται τόσο στις πραγματικές όσο και στις προβλεπόμενες κατανομές (Brownlee, 2016).

$$Kappa = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy}$$

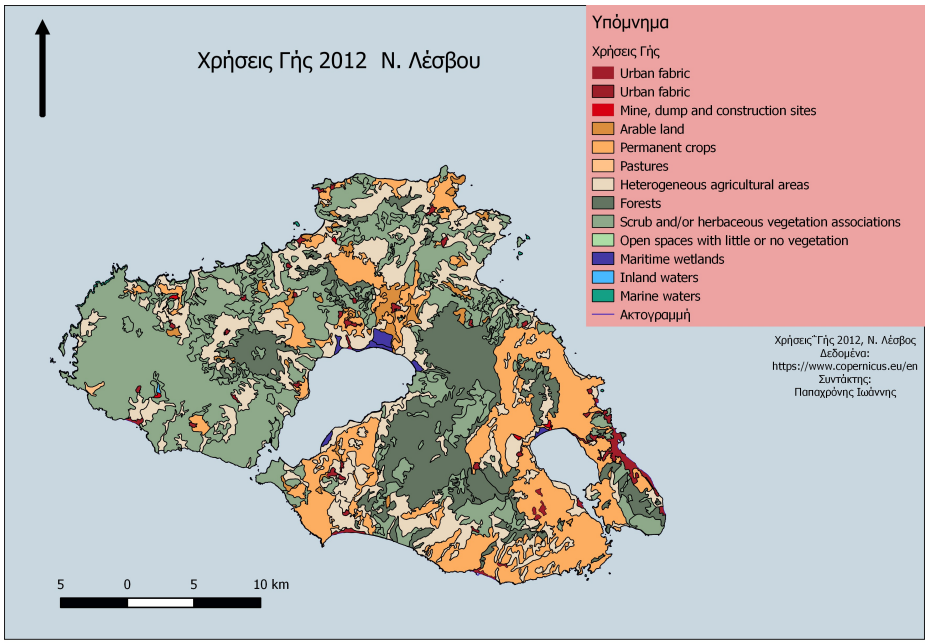
# 3 ΜΕΘΟΔΟΛΟΓΙΑ

Σε αυτήν την ενότητα αναφέρεται η περιοχή μελέτης και στα πλαίσια της εργασιας προτείνεται μεθοδολογική προσέγγιση που συνοψίζεται σε 3 βασικά βήματα Τα βήματα αυτά, αφορούν, τη συλλογή δεδομένων, την επεξεργασία τους στο πρόγραμμα QGIS 2.18 και τέλος στη χρήση αλγορίθμων για την ταξινόμηση της δορυφορικής εικόνας μέσω του στατιστικού πακέτου Rstudio

## 3.1 Περιοχή Μελέτης

Η Ν. Λέσβος ανήκει στο βορειοανατολικό τμήμα του Αιγαίου, στο νομό Λέσβου μαζί με τη Λήμνο και τον Άι Στράτη. Είναι το τρίτο σε μέγεθος ελληνικό νησί με έκταση 1.632.500 στρέμματα και ανάπτυγμα ακτών 320χλμ. Το ανάγλυφο του νησιού παρουσιάζει απότομες και ισχυρές εδαφικές κλίσεις με ποσοστό εδαφών με κλίση >16% που ανέρχεται στο 50,3% του συνόλου της έκτασης με συνέπεια την ανάπτυξη πυκνού υδρογραφικού δικτύου με μεγάλο όγκο απορροής νερού (Τσαγκαλίδης,2013).

Με με τη χρήση του Συστήματος Γεωγραφικών Πληροφοριών (ΣΓΠ) Qgis συντάχθηκε χάρτης με τις Χρήσεις Γής της Ν. Λέσβου χρησιμοποιώντας το Corine Land Cover (CLC) 2012. Στο χάρτη απεικονίζονται οι κλάσεις που αφορούν το δεύτερο επιπεδο πληροφορίας που χρειάστηκε για την ταξινόμηση.



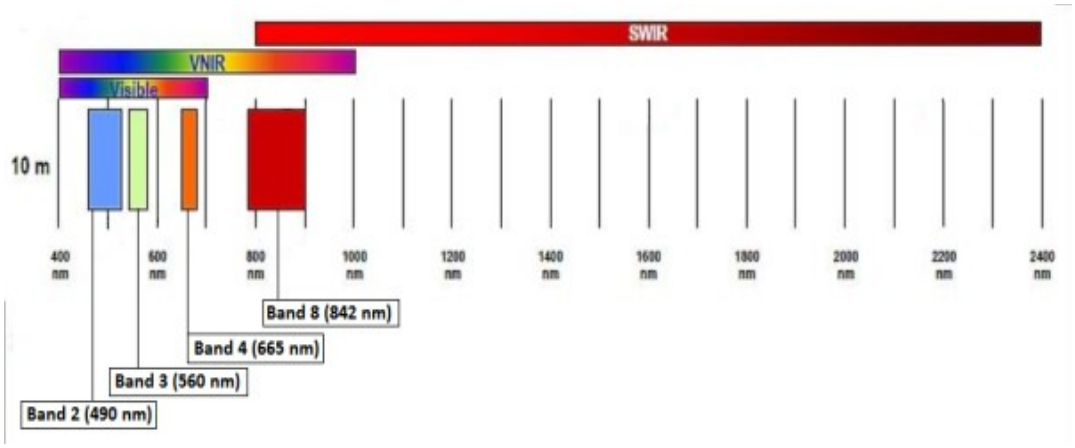

. ***Χάρτης 1.*** *Χρήσης Γής 2012 για τη Ν.Λέσβο, Qgis*

## 3.2 Συλογή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στη συγκεκριμένη εργασία αποτελλούνται απο:

- δορυφορική εικόνα Sentinel 2-A της Ν. Λέσβου
- το Corine 2012 (CLC 2012)
- Υπόμνημα Κατηγοριών Χρήσεων Γης (CLC legend)
- την ακτογραμμή της περιοχής μελέτης

Η δορυφορική εικόνα Sentinel 2-A συλλέχθηκε απο τη διαδικτυακή πλατφόρμα Earth Explorer (https://earthexplorer.usgs.gov) με ημερομηνία καταγραφής 27/08/2016. Ο δορυφόρος Sentinel 2-A διαθέτει 13 κανάλια με χωρική διακριτική ικανότητα 20m έως 60m και θεωρείται κατάλληλος για έρευνες που αφορούν την ανίχνευση αλλαγών και τις χρήσεις Γής, την Κληματική Αλλαγή και στην αντιμετώπιση καταστροφών. Τα δεδομένα είναι γεωαναφερμένα και το προβολικό τους είναι σε GGRS87 (Greek Grid).




***Εικόνα 11.****Τα δεκατρία φασματικά κανάλια του δοευφόρουSentinel 2-A*
*Πηγή: https://earth.esa.int*

Οι Χρήσεις Γης που χρησιμοποιήθηκαν αφορούν της Χρήσεις Γής 2012( CLC2012) σε μορφή γεωχωρικών δεδομένων (Shapefile) και περιλμβάνει τρία επίπεδα πληροφορίας χωρισμένα σε πέντε θεματικές κατηγορίες. Στην εργασία χρησιμοποιήθηκε το δεύτερο επιπεδο πληροφορίας για την ταξινόμηση. Τα δεδομένα

συλλέχθηκαν απο την μέσα από το πλατφόρμα Κοπέρνικος (https://www.copernicus.eu/en), δωρεάν και με ελεύθερη πρόσβαση όπως και πληροφορίες κατηγοριοποίησης. Τέλος, χρησιμοιποιήθηκε η ακτογραμμή της Λέσβου σε μορφή Shapefile απο το Εργαστήριο Χαρτογραφίας του Πανεπιστημίου Αιγαίου, του Τμήματος Γεωγραφίας.

## 3.3 Επεξεργασία Δεδομένων

Για την Επεξεργασίας των δεδομένων χρησιμοιποιήθηκε το Quantum GIS ή QGIS (v. 2.18.17) (http://www.qgis.org) είναι ένα Σύστημα Γεωγραφικών Πληροφοριών.Το QGIS διαθέτει φιλικό γραφικό περιβάλλον επικοινωνίας με το χρήστη και ενσωματώνει εργαλεία και λειτουργίες όπως η σχεδίαση και η ταυτόχρονη σχεδίαση διανυσματικών και κανονικοποιημένων γεωχωρικών δεδομένων, η διαχείριση και ο μετασχηματισμός του Συστήματος Αναφοράς Συντεταγμένων (ΣΑΣ), η διερεύνηση των δεδομένων και η χαρτοσύνθεση, η συλλογή, η επεξεργασία, η διαχείριση και η εξαγωγή, η χωρική ανάλυση και η γεωεπεξεργασία των δεδομένων, η δημοσιοποίηση στο διαδίκτυο κ.ά. Υποστηρίζει πολλαπλούς μορφότυπους διανυσματικών και κανονικοποιημένων δεδομένων και 2 επικοινωνεί με χωρικές βάσεις δεδομένων.

Μέσω του Qgis έγινε η προετοιμασία των δεδομένων, πιο συγκεκριμένα, με τη διαδικασία clip, περιορίστηκε ο όγκος των δεδομένων απο τις χρήσεις γής στα όρια της περιοχής μελέτης χρησιμοποιώντας την ακτογραμμή. Στη συνέχεια έγινε ο διαχωρισμός των καναλιών της δορυφορικής εικόνας σε 3 σετ με βαση τη διακριτική τους ικανόητα, των 10μ, 20μ και 60μ.

Τέλος με τη διαδικασία του clip, περιορίσαμε τον όγκο πληροφορίας των εικόνων στα όρια της περιοχής μελέτης.

## 3.4 Η Γλώσσα R και το RStudio

Η R είναι μια γλώσσα προγραμματισμού για στατιστικούς υπολογισμούς και γραφικά. Παρέχει μια μεγάλη ποικιλία γραμμικών και μη γραμμικών μοντέλων,

κλασικών στατιστικών εξετάσεων, ανάλυσης χρονικής σειράς, ταξινόμησης, συσταδοποίησης ,τεχνικών γραφικών και έχει μεγάλη επέκταση. Ένα από τα πλεονεκτήματα της R είναι η ευκολία με την οποία μπορεί να οπτικοποιήσει τα δεδομένα και τα αποτελέσματα με μεγάλη ποιοτική ακρίβεια συμπεριλαμβανομένων μαθηματικών συμβόλων και τύπων όπου χρειάζεται. Η γλώσσα R είναι διαθέσιμη ως ελεύθερο λογισμικό και διέπεται απο την αρχή του ανοιχτού κώδικα.

Το RStudio είναι ένα ελεύθερο και ανοιχτού κώδικα ολοκληρωμένο περιβάλλον ανάπτυξης για την R, μια γλώσσα προγραμματισμού για στατιστικούς υπολογισμούς και γραφικά. Περιλαμβάνει μια κονσόλα, συντάκτης επισημάνσεως σύνταξης που υποστηρίζει την άμεση εκτέλεση κώδικα, καθώς και εργαλεία σχεδίασης, ιστορικού, εντοπισμού σφαλμάτων και διαχείρισης χώρου εργασίας (https://www.r-project.org/about.html).

## 3.5 Διαδικασία Μέσω Rstudio

Μέσω του Rstudio. πραγματοποιήθηκε ή διαδικασία της ταξινόμησης της δορυφορικής εικόνας με τη χρήση αλγορίθμων.

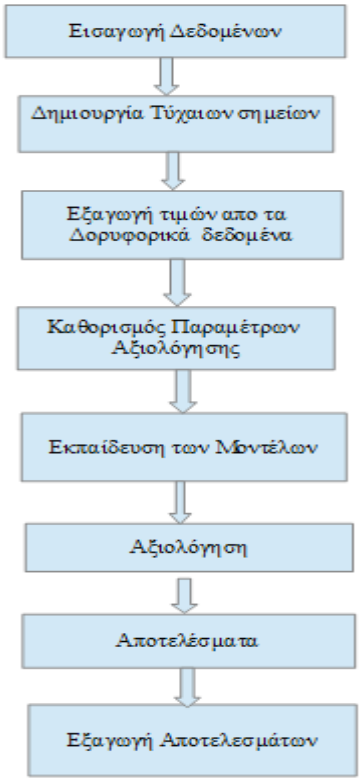
Τα βήματα που ακολουθήθηκαν για να ολοκληρωθεί η διαδικασία ταξινόμησης έχουν ως εξής:

- Αρχικά έγινε ο καθορισμός των δειγμάτων εκπαίδευσης των αλγορίθμων της ταξινόμησης
- Στη συνέχεια έγινε η χρήση της τεχνικής της τυχαίας δειγματοληψίας με την οποία τοποθετήθηκαν 1000 τυχαία σημεία (Random Samping).
  Στην εφαρμογής της τυχαίας δειγματοληπτικής μεθόδου, ως μοναδικό κριτήριο ορίστηκε το εμβαδό που καταλαμβάνει η κάθε χρήσει γής με βάση το CLC2012
- Για κάθε σετ εικόνων έγινε η εξαγωγή των αντίστοιχων τιμών στα σημεία ώστε να τροφοδοτήσουν την εκπαίδευση των μοντέλων
- Έπειτα γίνεται η εκπαίδευση των μοντέλων με την οποία πρόκειται να γίνει η επιβλεπόμενη ταξινόμηση

- Γίνεται ο ορισμό των κατηγοριών, (formula(class ~ b2 + b3 + b4 + b8)
- Η προεπεξεργασία των δεδομένων , preProc=c(“center”, “scale”)
- Ο έλεγχος και η επικύρωση των αποτελεσμάτων, control trainControl(method="repeatedcv", number=10, repeats=3)
- Η εκτίμηση ακρίβειας των αποτελεσμάτων metric ← c("Accuracy", “Kappa”)
- Δημιουργείται μία λίστα για την αποθήκευση των τριών μοντέλων για κάθε αλγόριθμο (fit...<-(list)
- Γίνεται η εκτίμηση και πρόβλεψη των αποτελεσμάτων (“evaluation”, “predict”)
- Εξαγωγή γραφημάτων Accuracy & Kappa
- Τέλος έχουμε την εξαγωγή των αποτελεσμάτων των μοντέλων.

## 3.6 Διάγραμμα Ροής

Στο Διάγραμμα ροής που ακολουθεί, αποτυπώνεται η διαδικασία της μεθοδολογίας που ακολουθήθηκε για την εργασία αυτή.




***Εικόνα 12***. *Διάγραμμα Ροής*

# 4 ΑΠΟΤΕΛΕΣΜΑΤΑ

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα των ταξινομήσεων όπως αυτά προέκυψαν έπειτα από την εφαρμογή 8 διαφορετικών αλγορίθμων ταξινόμησης για τη νήσο Λέσβο. Ως αποτέλεσμα του σταδίου της προ-επεξεργασίας είναι άξιο αναφοράς το ότι έγινε διαχωρισμός της εικόνας σε 3 σετ εικόνων με βάση τη διακριτική ικανότητα της κάθε μπάντας: 4 μπάντες με 10m (2 – Red, 3 – Green, 4 – Blue, 8 – Infrared), 6 μπάντες με 20m (5 – Vegetation red edge, 6 – Vegetation red edge, 7 – Vegetation red edge, 8A – Narrow NIR, 11 – SWIR 1, 12 – SWIR 2) 3 μπάντες με 60m (1 – Coastal aerosol, 9 – Water vapour, 10 Swir – Cirrus).

Αρχικά, απαραίτητος κρίνεται ο καθορισμός των δειγμάτων εκπαίδευσης των αλγορίθμων ταξινόμησης. Η επιλογή της στρατηγικής για την εύρεση των βέλτιστων δειγμάτων εκπαίδευσης βασίζεται στην παραδοχή της ορθότητας του Corine Land Cover (CLC) 2012 και στη χρήση τεχνικών τυχαίας δειγματοληψίας (Random Sampling ). Η εφαρμογή της τυχαίας δειγματοληπτικής μεθόδου κατά τη διαδικασία εύρεσης αντιπροσωπευτικών δεδομένων εκπαίδευσης έχει μοναδικό κριτήριο το εμβαδόν που καταλαμβάνει η κάθε χρήση του CLC. Τα αποτελέσματά της μεθόδου για 1000 σημεία παρουσιάζονται στον Πίνακα 1.

| **α/α** | **Ονομασία χρήσης (κωδικός)** | **N = 1000** |
|---|---|---|
| 1 | Urban fabric (11) | 11 |
| 2 | Industrial, commercial and transport units (12) | 1 |
| 3 | Mine, dump and construction sites (13) | 0 |
| 4 | Arable land (21) | 20 |
| 5 | Permanent crops (22) | 212 |
| 6 | Pastures (23) | 1 |
| 7 | Heterogeneous agricultural areas (24) | 225 |
| 8 | Forests (31) | 155 |
| 9 | Scrub and/or herbaceous vegetation associations (32) | 361 |

| | | |
|---|---|---|
| 10 | Open spaces with little or no vegetation (33) | 0 |
| 11 | Maritime wetlands (42) | 7 |
| 12 | Inland waters (51) | 0 |
| 13 | Marine waters (52) | 7 |
| **Σύνολο** | | **1000** |

**Πίνακας 1.** Δημιουργία τυχαίων σημείων

Για κάθε σετ εικόνων (stack) γίνεται εξαγωγή των αντίστοιχων τιμών στα σημεία ώστε σε επόμενο βήμα να αποτελέσουν τροφοδοσία στη διαδικασία εκπαίδευσης των μοντέλων. Από τον Πίνακα 1 παρατηρείται ότι 3 εκ των χρήσεων παρουσιάζουν μηδενική εκπροσώπηση τιμών, 5 κατηγορίες υπό-εκπροσώπούνται και οι υπόλοιπες υπερ-εκπροσωπούνται. Το φαινόμενο αυτό πρόκειται να επιδράσει στην εκπαίδευση των μοντέλων με αρνητικό τρόπο και άρα ενδέχεται να επηρεάσει την αξιοπιστία και εγκυρότητα των αποτελεσμάτων.

Σε επόμενο στάδιο λαμβάνει χώρα η εκπαίδευση των μοντέλων με βάση τα οποία πρόκειται να γίνει η επιβλεπόμενη ταξινόμηση. Σημαντικές παράμετροι που αφορούν την εκπαίδευση είναι α) ο ορισμός των κατηγοριών ταξινόμησης και των αντίστοιχων τιμών που αντιπροσωπεύουν κάθε κατηγορία για κάθε σετ εικόνων β) προεπεξεργασία των δεδομένων η οποία λειτουργεί ως βοηθητική στην εφαρμογή ορισμένων αλγορίθμων, γ) ο τρόπος ελέγχου/επικύρωσης των αποτελεσμάτων και δ) η εκτίμηση ακρίβειας των αποτελεσμάτων.

α) *formula(class ~ b2 + b3 + b4 + b8)*

β) *preProc=c("center", "scale")*

γ) *control ← trainControl(method="repeatedcv", number=10, repeats=3)*

δ) *metric ← c("Accuracy", "Kappa")*

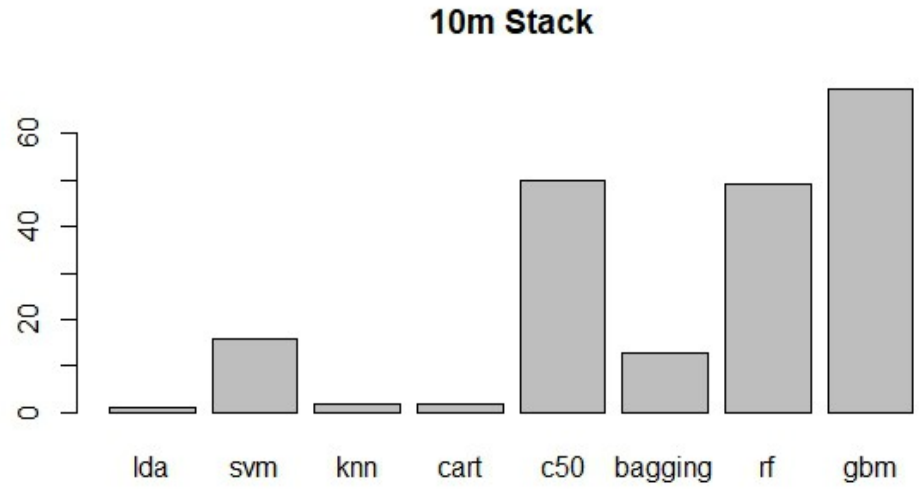
*train(formula, train data, method, metric, pre-processing, control)*

Αυξημένης σημαντικότητας είναι οι παράγοντες control και metric. Ο έλεγχος επικύρωσης των δεδομένων (control) χρησιμοποιεί την μέθοδο επαναλαμβανόμενης διασταυρωμένης επικύρωσης (Repeated Cross Validation) κατά την οποία κατακερματίζεται το δείγμα σε 10 ίσα τμήματα (number = 10) χρησιμοποιεί τα 9/10 ή 90% για εκπαιδευτικούς σκοπούς και το 10% για επικύρωση. Η διαδικασία επικύρωσης επαναλαμβάνεtai 3 φορές (repeats = 3) για ισχυρότερες/καλύτερες εκτιμήσεις.

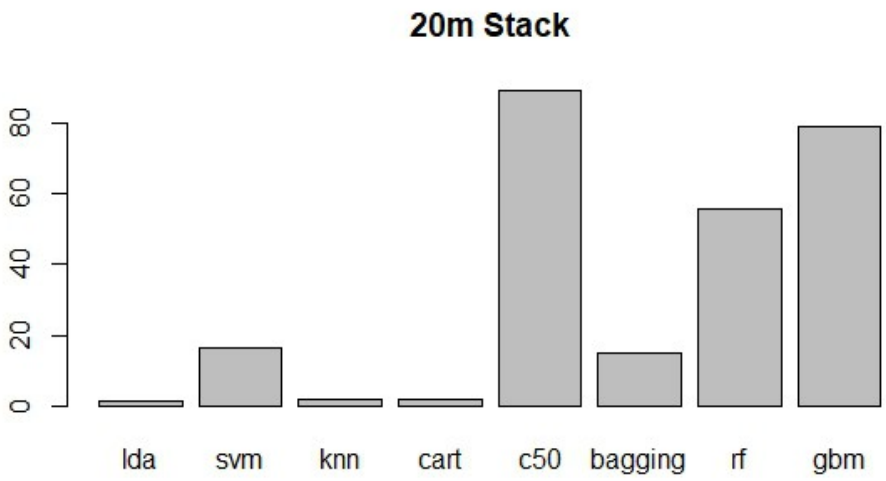
Η σύγκριση/αξιολόγηση των μοντέλων είναι τώρα εφικτή καθώς η εκπαιδευτική διαδικασία των μοντέλων έχει τελειώσει επιτυχώς. Στα πλαίσια των μοντέλων εντάσσεται ο α) χρόνος επεξεργασίας ο οποίος εξαρτάται και από τις δυνατότητες υπολογισμού του Η/Υ, του λογισμικού και λειτουργικό σύστημα και β) οι τιμές accuracy και kappa.

Η υπολογιστική μηχανή που έλαβαν χώρα  οι υπολογισμοί αποτελείται από 4 πυρήνες 3 γενιάς (Intel i3) επεξεργαστικής ισχύος 2.5GHz, μνήμη ταχείας προσπέλασης 4GB ταχύτητας 1600Mhz, σκληρό δίσκο SSD 400MB writing 450 MB reading. Οι χρόνοι απόκρισης των αλγορίθμων για τα τρία σετ εικόνων (ανάλυση 10m, 20m και 60m) παρουσιάζονται στα Γραφήματα 1, 2 και 3.

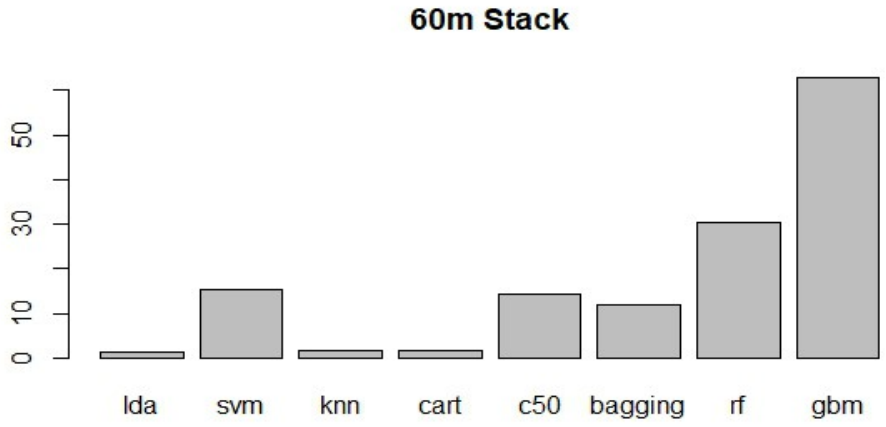
## 4.1 Συγκρίσεις - Χρονική Απόκριση




***Γράφημα 1.*** *Διάρκεια απόκρισης μοντέλων (σε δευτερόλεπτα) για το σετ εικόνων με 10m ανάλυση*




*Γράφημα 2. Διάρκεια απόκρισης μοντέλων (σε δευτερόλεπτα) για το σετ εικόνων με 20m ανάλυση*




***Γράφημα 3.*** *Διάρκεια απόκρισης μοντέλων (σε δευτερόλεπτα) για το σετ εικόνων με 60m ανάλυση*

Οι αλγόριθμοι lda, knn και cart παρουσιάζουν εξαιρετικά χαμηλούς χρόνους (έως 3 δευτερόλεπτα), ενδιάμεσους χρόνους έχουν και στα 3 σετ εικόνων οι svm και bagging ενώ oI rf και gbm και στις 3 περιπτώσεις απαιτούν τον περισσότερο. Ο αλγόριθμος c50 από τη μία στα δύο πρώτα σετ εικόνων συγκαταλέγεται στους αλγορίθμους με σχετικά υψηλότερο χρόνο υπολογισμού στο τρίτο σετ εικόνων συγκαταλέγεται στους αλγορίθμους με ενδιάμεσο χρόνο.

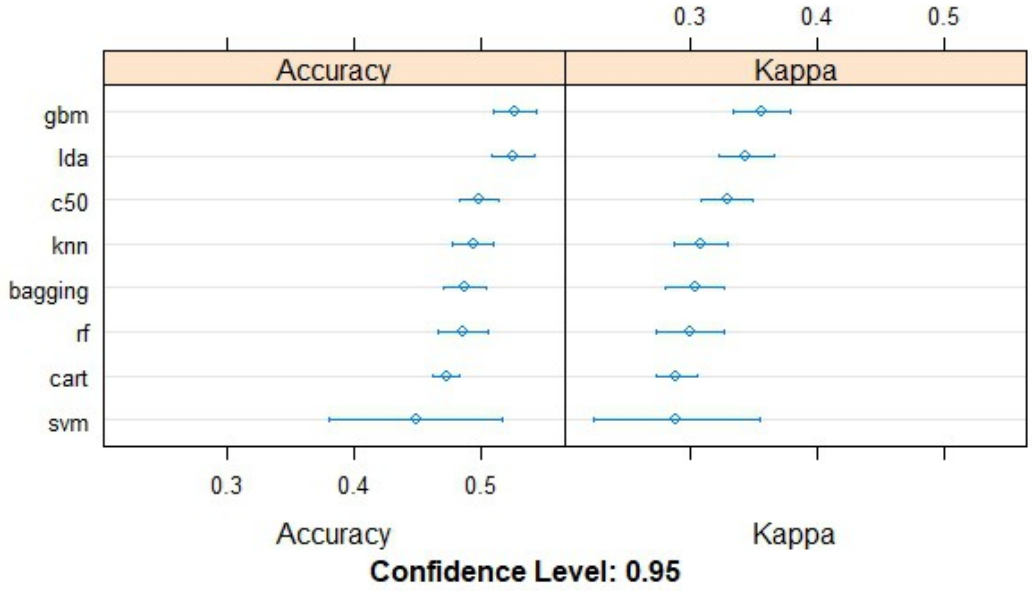
## 4.2 Συγκρίσεις – Εκτίμηση Ακρίβειας

Την αξιολόγηση των αλγορίθμων με βάση το χρόνο ακολουθεί η αξιολόγηση των αλγορίθμων με βάση την ακρίβεια των αποτελεσμάτων. Για το λόγο αυτό υπολογίζονται οι τιμές Accuracy και Kappa για κάθε αλγόριθμο όπως απεικονίζεται στα γραφήματα 4 έως 6. Οι τιμές Accuracy των αλγορίθμων αφορούν στον λόγο των σωστά ταξινομήμένων εικονοστοιχείων ως προς το σύνολο των εικονοστοιχείων (1) και λαμβάνουν τιμές από 0 έως 1.

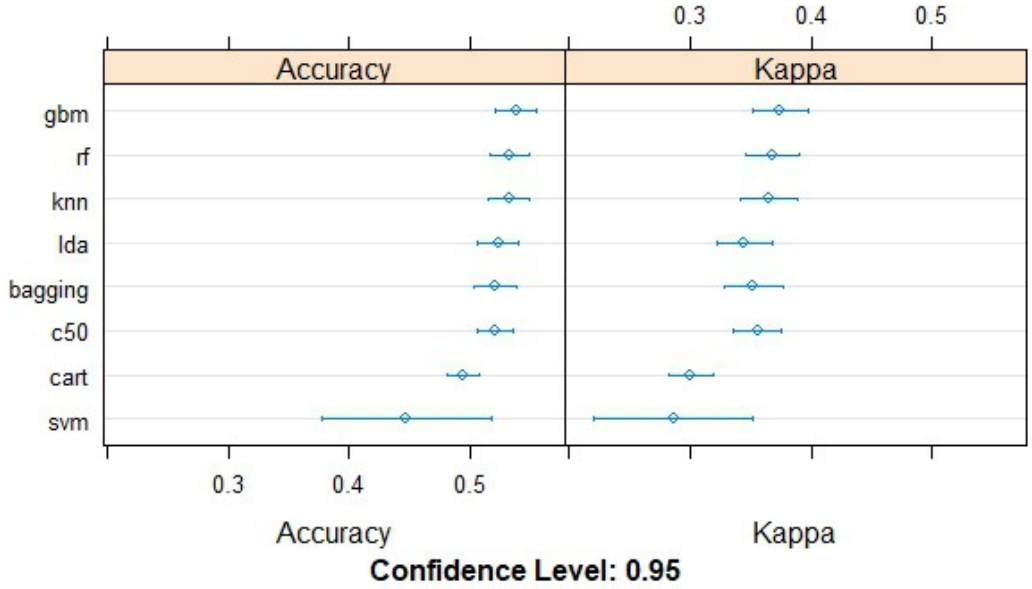
$$Accuracy = \frac{Correct\ Classified\ Pixels}{Total\ of\ Pixels} \tag{1}$$

Αντίστοιχα οι τιμές Kappa αποτελούν στατιστικό μέτρο σύγκρισης των της αναμενόμενης ακρίβειας των αποτελεσμάτων ταξινόμησης σε σχέσει με την παρατηρούμενη ακρίβεια των αποτελεσμάτων (2). Ο δείκτης Kappa, όπως και ο Accuracy, παίρνει τιμές στην κλίμακα από 0 έως 1.

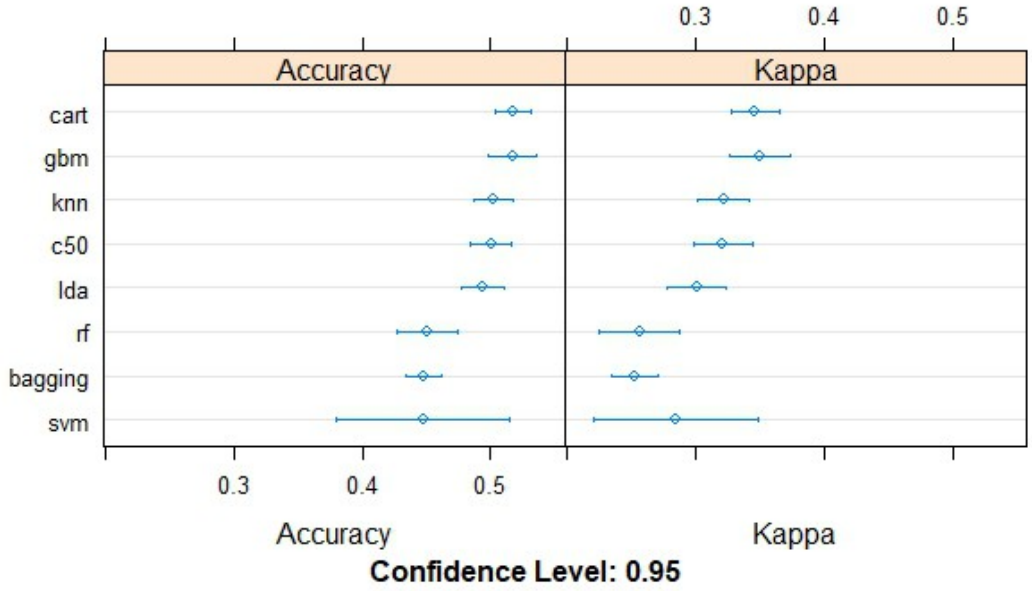
$$Kappa = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy} \tag{2}$$




***Γράφημα 4.*** *Τιμές accuracy και kappa για τους 8 αλγορίθμους που αφορούν το σετ εικόνων με ανάλυση 10m*




***Γράφημα 5.*** *Τιμές accuracy και kappa για τους 8 αλγορίθμους που αφορούν το σετ εικόνων με ανάλυση 20m*

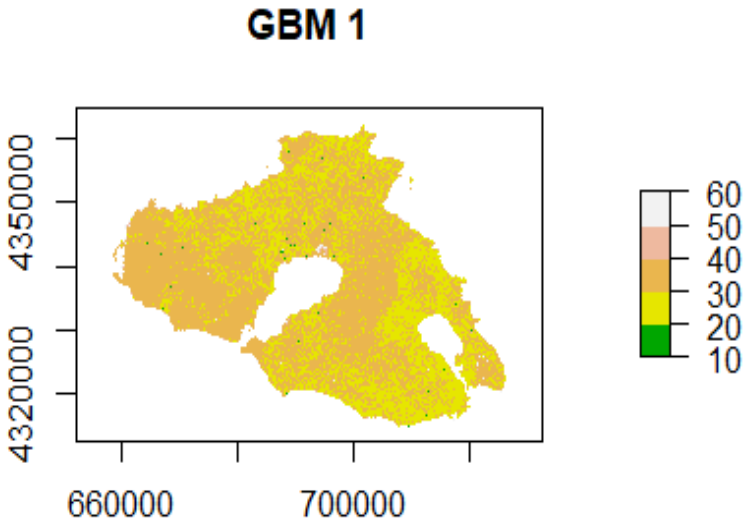



***Γράφημα 6.*** *Τιμές accuracy και kappa για τους 8 αλγορίθμους που αφορούν το σετ εικόνων με ανάλυση 60m*

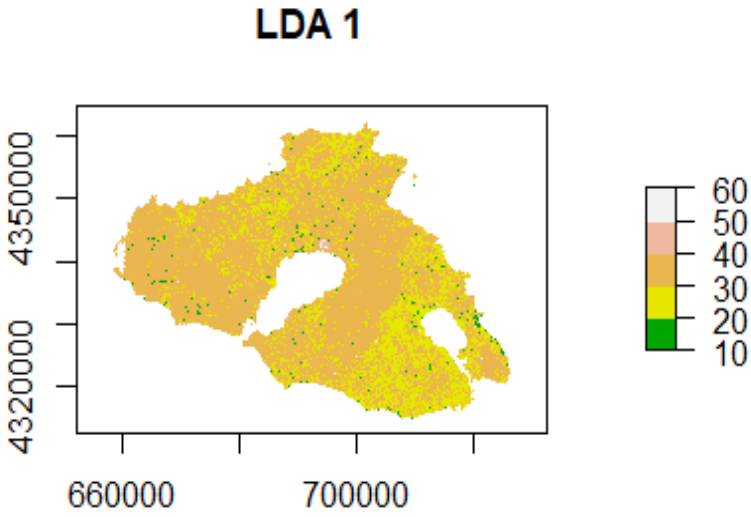
Η εκτίμηση ακρίβειας των μοντέλων έγινε σε επίπεδο εμπιστοσύνης 95%. Η βασική παρατήρηση που ταυτόχρονα αποτελεί και αρνητικό στοιχείο και αφορά τα παραπάνω 3 γραφήματα είναι ότι οι τιμές των στατιστικών μέτρων εκτίμησης ακρίβειας είναι εξαιρετικά χαμηλές. Ωστόσο στα γραφήματα 4 έως 6 ο αλγόριθμος gbm φαίνεται να έχει από τις καλύτερες επιδόσεις και στα δύο στατιστικά μέτρα ενώ και στα 3 σετ εικόνων ο αλγόριθμος SVM παρουσιάζει αντίστοιχα τα χειρότερα αποτελέσματα. Στο Γράφημα 4 ακόμα μια πολύ καλή επίδοση φαίνεται είναι αυτή του αλγορίθμου lda. Στο Γράφημα 5 μαζί με τον gbm αντίστοιχα υψηλά αποτελέσματα παρουσιάζουν οι αλγόριμοι knn και rf. Τέλος, στο Γράφημα 6 ο αλγόριμος cart κατατάσσεται πρώτος στις τιμές Accuracy και δεύτερος στις τιμές Kappa.

## 4.3 Συγκρίσεις – Οπτικοποίηση Αποτελεσμάτων

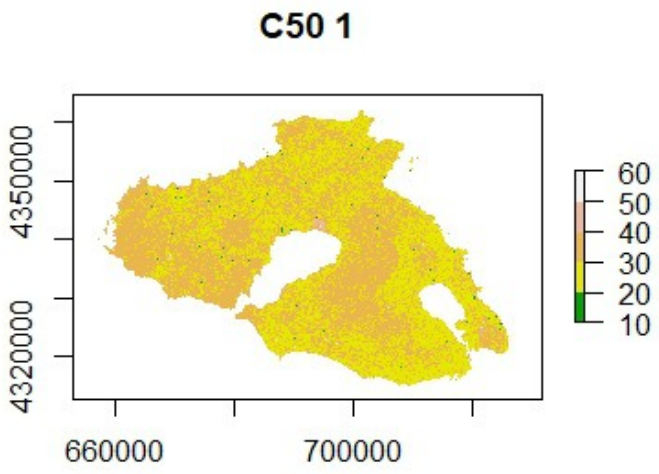
Στα γραφήματα από από 7 έως 30 παρουσιάζονται οι οπτικοποιήσεις των ταξινομήσεων χρησιμοποιώντας τεχνικές παράλληλου προγραμματισμού που αφορούν το σετ εικόνων με διακριτική ικανότητα 10m, 20m και 30m.



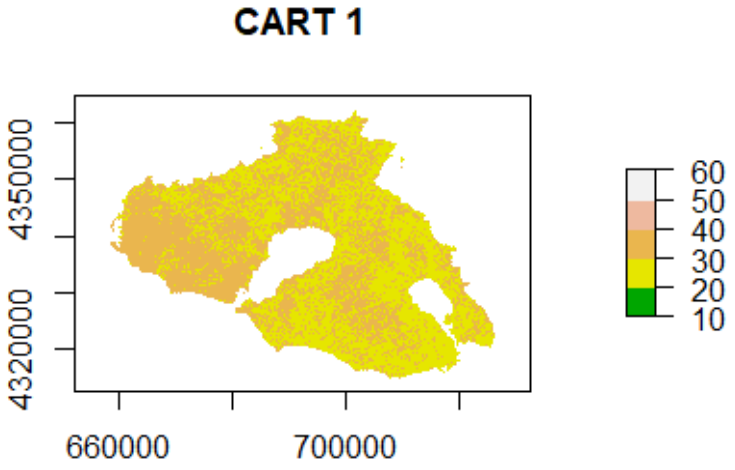

***Γράφημα 7.*** *Αποτέλεσμα αλγορίθμου gbm για το πρώτο σετ εικόνων*



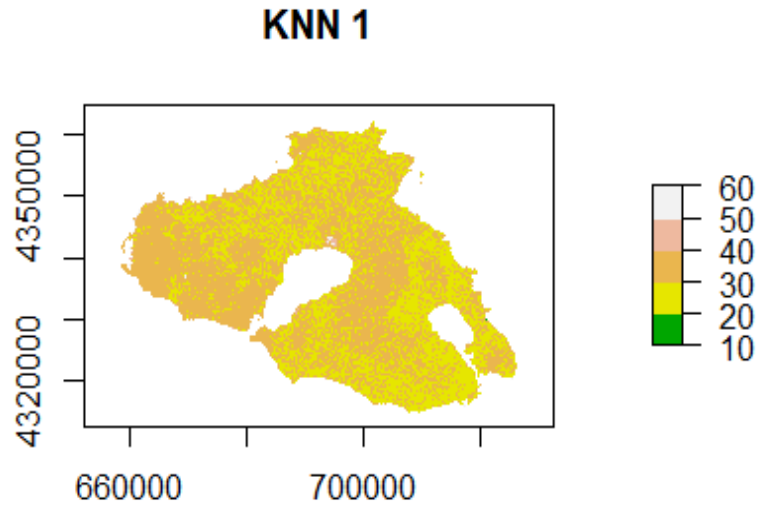

***Γράφημα 8.*** *Αποτέλεσμα αλγορίθμου lda για το πρώτο σετ εικόνων*



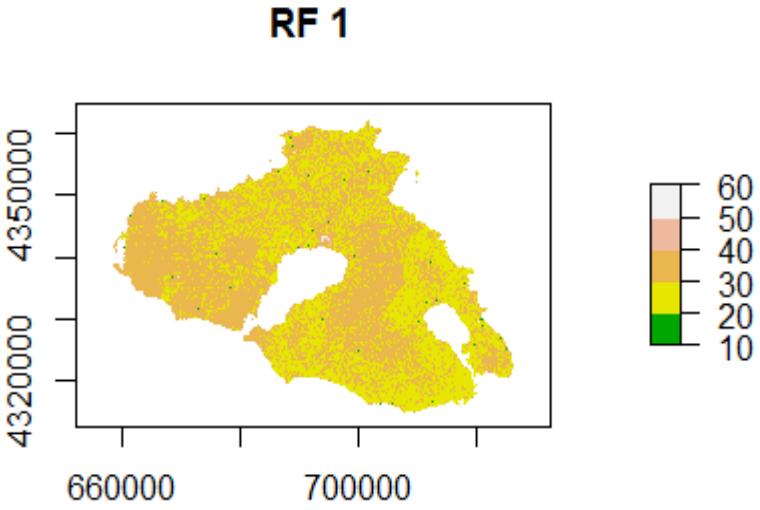

***Γράφημα 9.*** *Αποτέλεσμα αλγορίθμου c50 για το πρώτο σετ εικόνων*



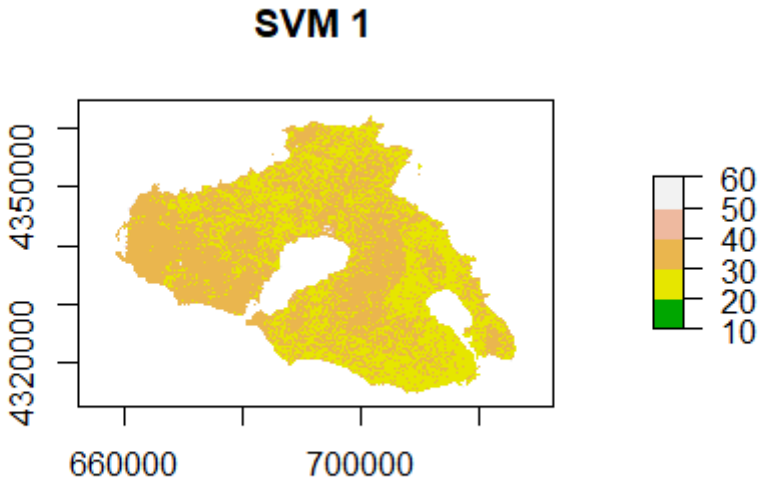

***Γράφημα 10.*** *Αποτέλεσμα αλγορίθμου cart για το πρώτο σετ εικόνων*



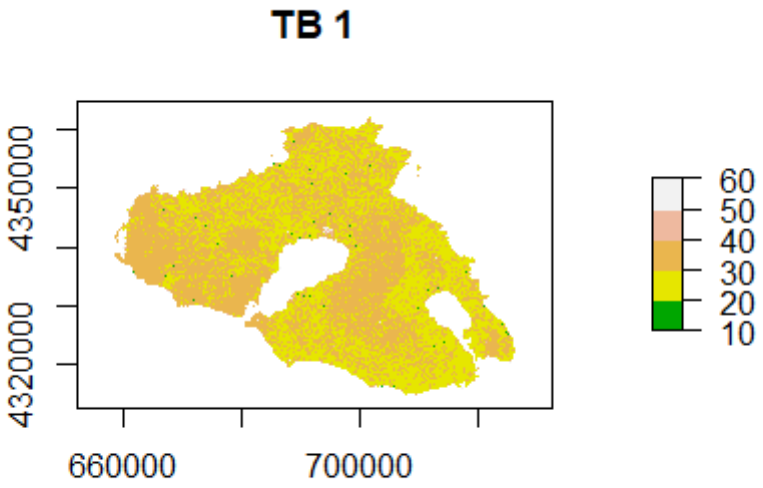

***Γράφημα 11.*** *Αποτέλεσμα αλγορίθμου knn για το πρώτο σετ εικόνων*



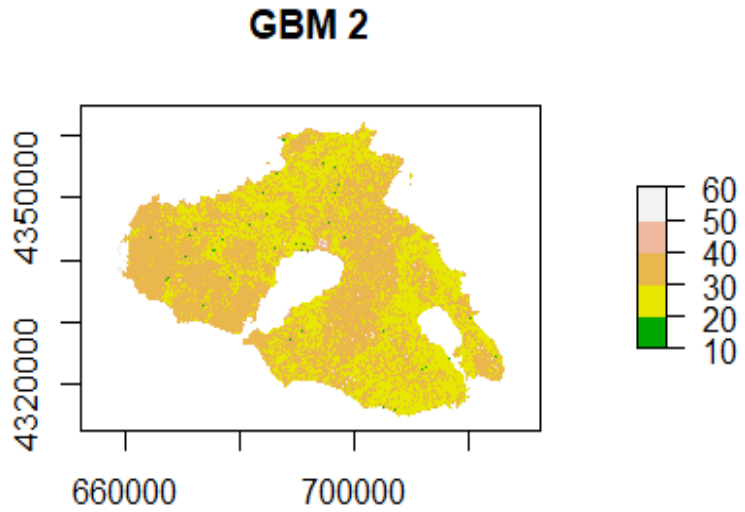

***Γράφημα 12.*** *Αποτέλεσμα αλγορίθμου Rf για το πρώτο σετ εικόνων*



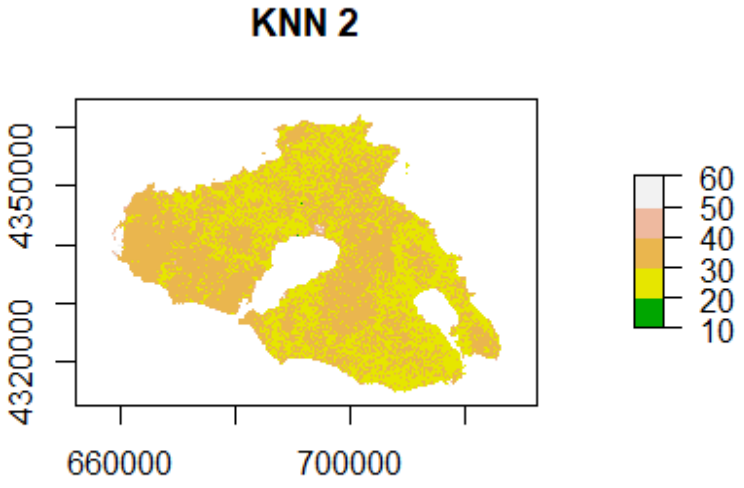

***Γράφημα 13.*** *Αποτέλεσμα αλγορίθμου svm για το πρώτο σετ εικόνων*



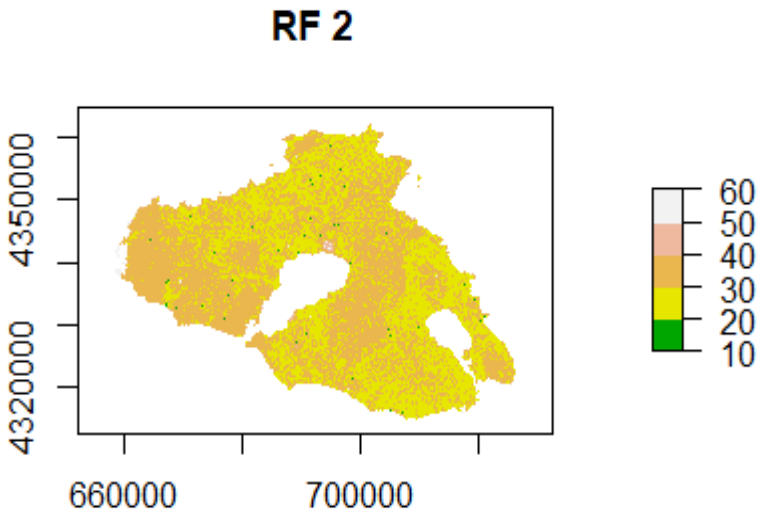

***Γράφημα 14.*** *Αποτέλεσμα αλγορίθμου TB(bagging) για το πρώτο σετ εικόνων*



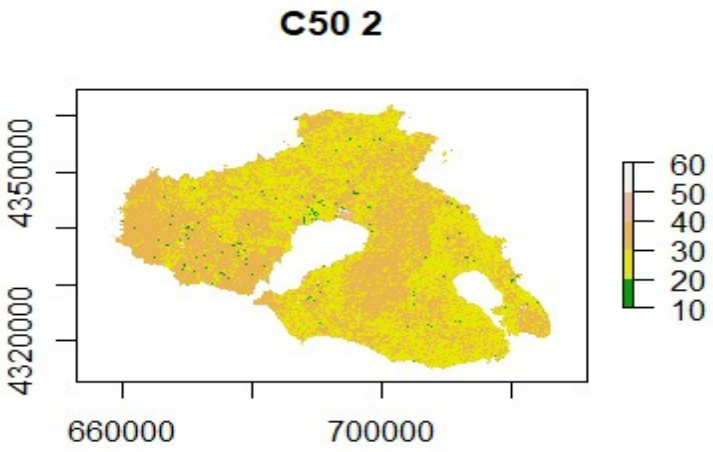

***Γράφημα 15.*** *Αποτέλεσμα αλγορίθμου gbm για το δεύτερο σετ εικόνων*



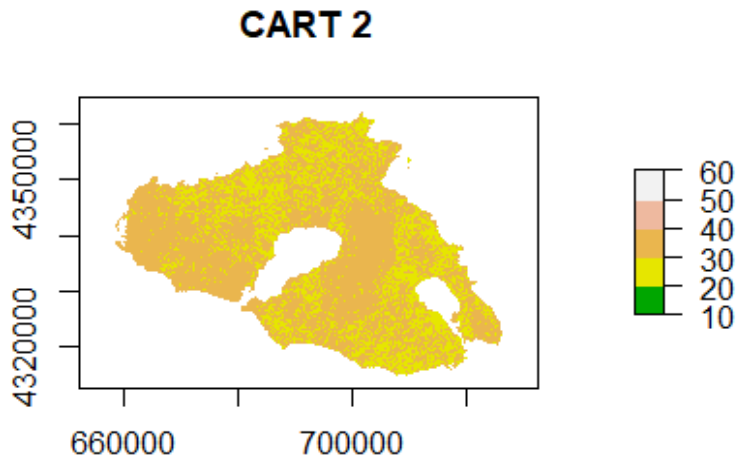

***Γράφημα 16.*** *Αποτέλεσμα αλγορίθμου knn για το δεύτερο σετ εικόνων*



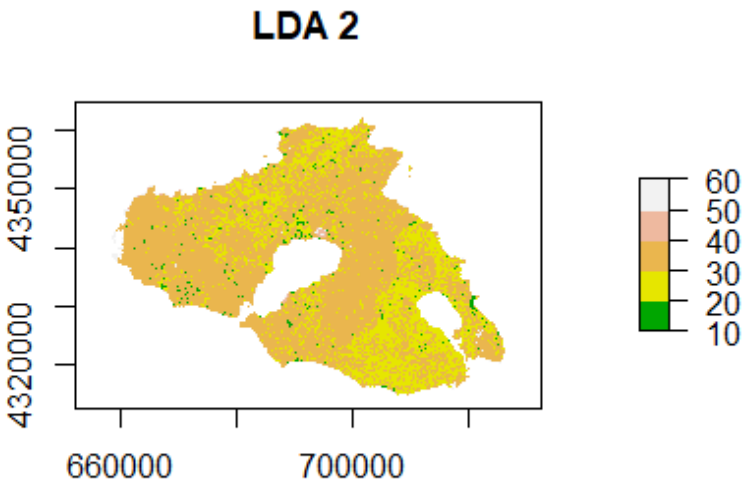

***Γράφημα 17.*** *Αποτέλεσμα αλγορίθμου rf για το δεύτερο σετ εικόνων*



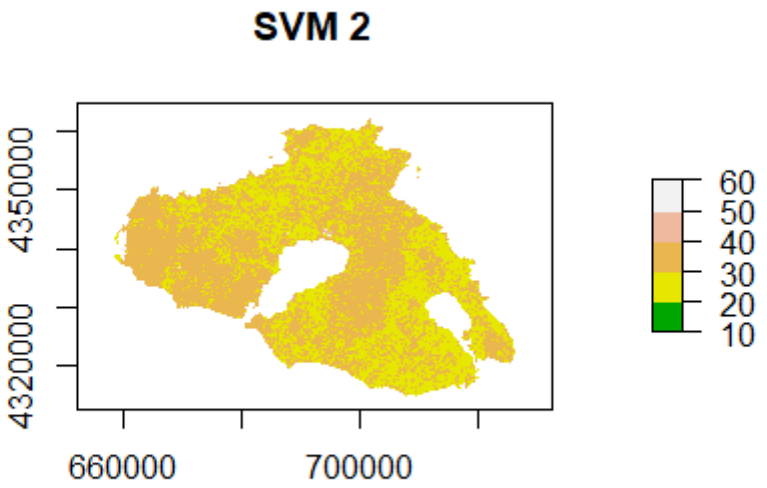

***Γράφημα 18.*** *Αποτέλεσμα αλγορίθμου c50 για το δεύτερο σετ εικόνων*



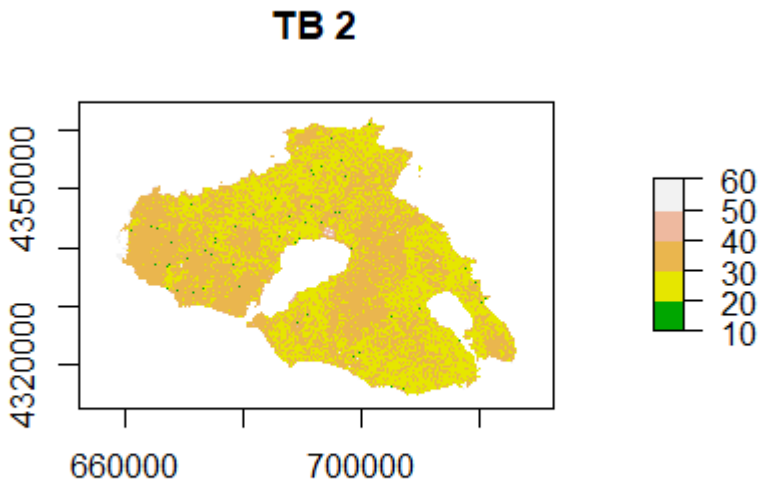

***Γράφημα 19.*** *Αποτέλεσμα αλγορίθμου cart για το δεύτερο σετ εικόνων*



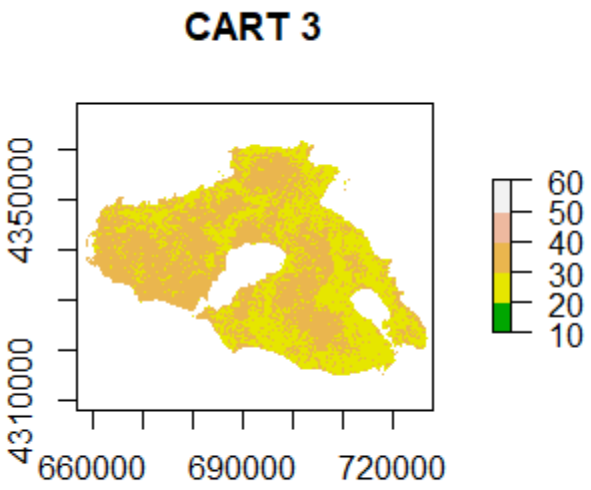

***Γράφημα 20.*** *Αποτέλεσμα αλγορίθμου lda για το δεύτερο σετ εικόνων*



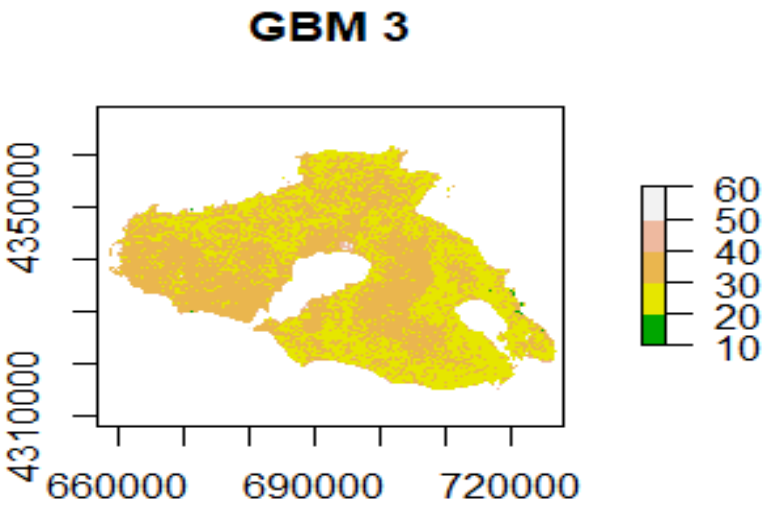

***Γράφημα 21.*** *Αποτέλεσμα αλγορίθμου SVM για το δεύτερο σετ εικόνων*



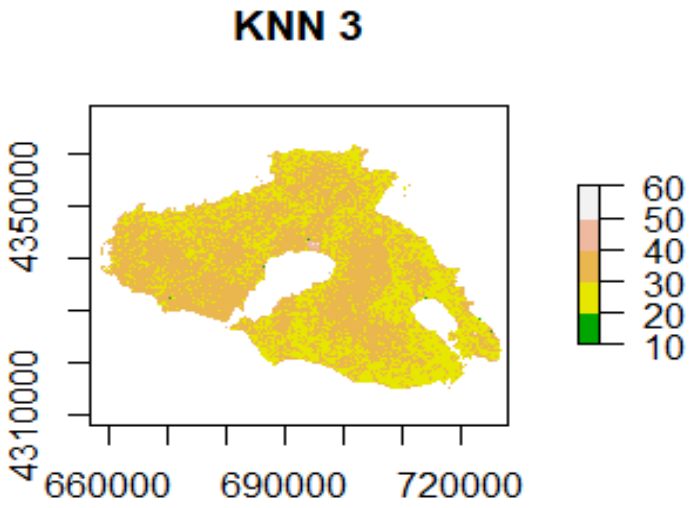

***Γράφημα 22.*** *Αποτέλεσμα αλγορίθμου tb για το δεύτερο σετ εικόνων*



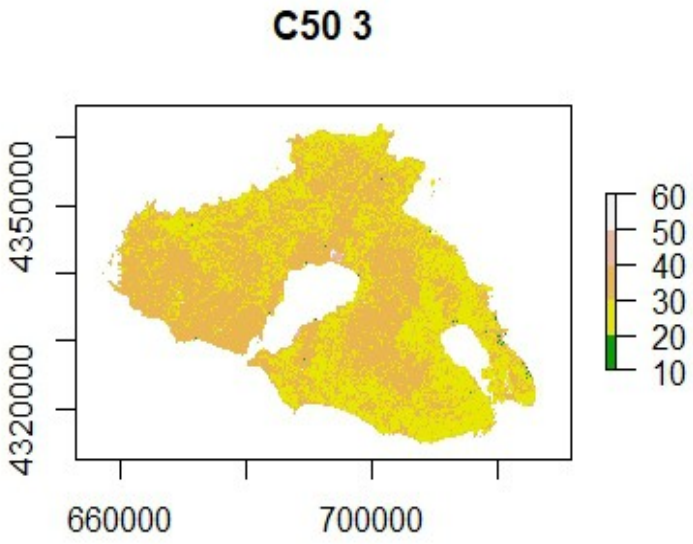

***Γράφημα 23.*** *Αποτέλεσμα αλγορίθμου cart για το τρίτο σετ εικόνων*



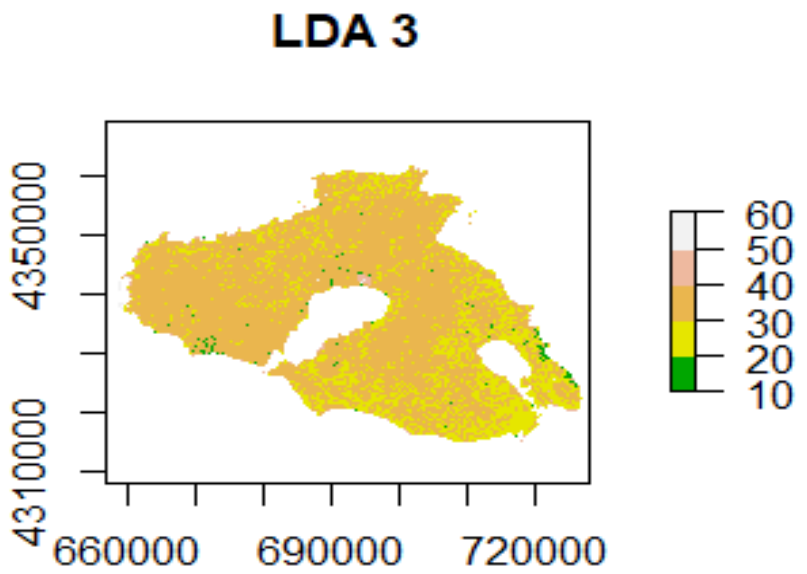

***Γράφημα 24.*** *Αποτέλεσμα αλγορίθμου gbm για το τρίτο σετ εικόνων*



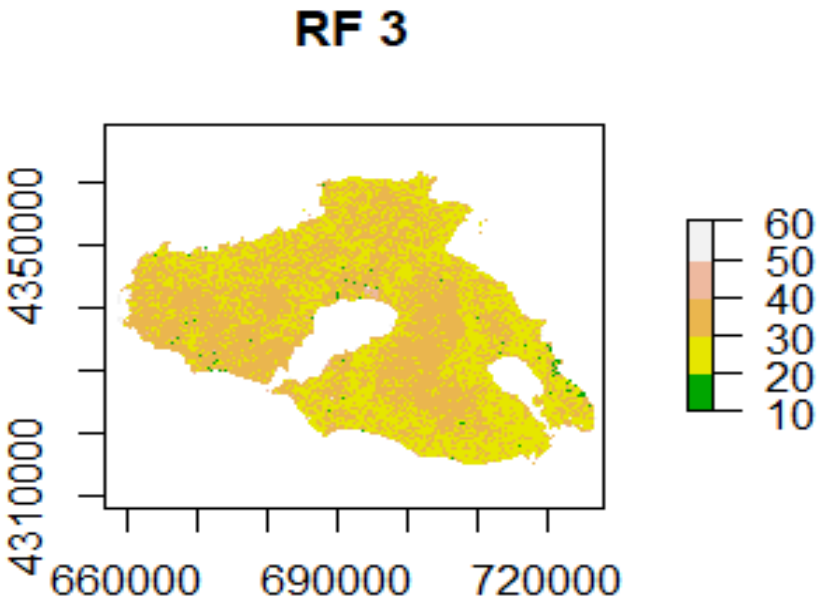

***Γράφημα 25.*** *Αποτέλεσμα αλγορίθμου knn για το τρίτο σετ εικόνων*



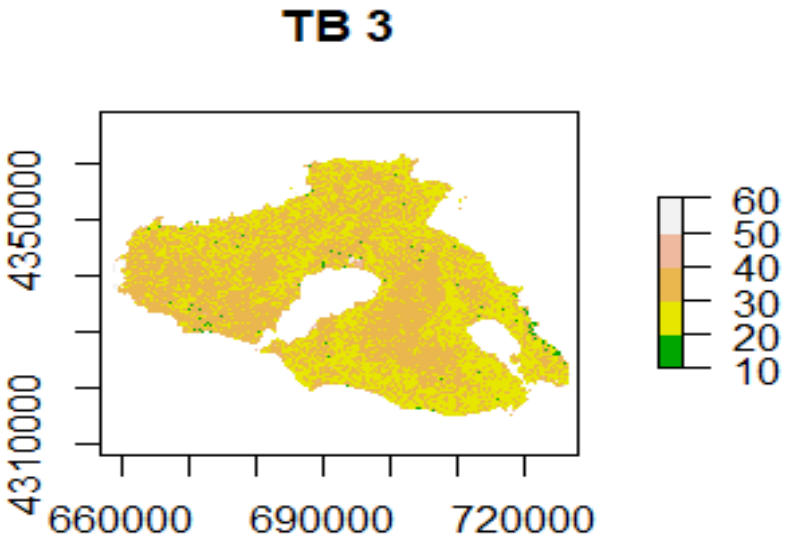

***Γράφημα 26.*** *Αποτέλεσμα αλγορίθμου c50 για το τρίτο σετ εικόνων*



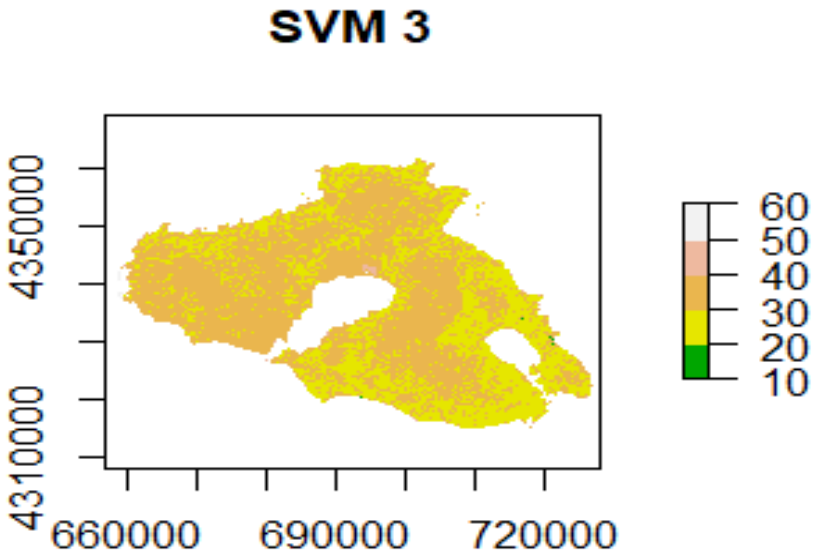

***Γράφημα 27.*** *Αποτέλεσμα αλγορίθμου lda για το τρίτο σετ εικόνων*




***Γράφημα 28.*** *Αποτέλεσμα αλγορίθμου rf για το τρίτο σετ εικόνων*




***Γράφημα 29.*** *Αποτέλεσμα αλγορίθμου tb για το τρίτο σετ εικόνων*




***Γράφημα 30.*** *Αποτέλεσμα αλγορίθμου svm για το τρίτο σετ εικόνων*

Από τα παραπάνω γραφήματα παρουσίασης των αποτελεσμάτων ταξινόμησης παρατηρείται γενικευμένη προβληματική απεικόνισης στο σύνολο των κατηγοριών. Πιο συγκεκριμένα, σε όλα τα σετ εικόνων παρουσιάζονται σημαντικές διαφοροποιήσεις στην κατανομή των εικονοστοιχείων μεταξύ των κατηγοριών πράγμα που οπτικά εντοπίζεται συγκριτικά με τα όρια του χάρτη χρήσεων γης. Ακόμα και οι αλγόριθμοι που αξιολογήθηκαν ως καλύτεροι σε κάθε περίπτωση υστερούν σημαντικά σε όλες τις κατηγορίες με ιδιαίτερη έμφαση στις τεχνητές επιφάνειες, στου υγροτόπους και στις υδατικές επιφάνειες όπου συγκριτικά με τον χάρτη χρήσεων γης παρουσιάζονται πολύ σημαντικές διαφορές.

Ο αλγόριθμος lda χρησιμοποιώντας γραμμική προσέγγιση τείνει να αποτυπώνει με καλύτερο τρόπο την κατανομή των εικονοστοιχείων της κατηγορίας τεχνητών επιφανειών ακολουθούν οι αλγόριθμοι συλλογικής μάθησης ενώ φανερά υστερούν οι αλγόριθμοι που χρησιμοποιούν μη γραμμικές ταξινομήσεις (svm, knn). Η ταξινόμηση στις υπόλοιπες κλάσεις (γεωργικές περιοχές και δαση ημι-φυσικές περιοχές) λόγω αυξημένης εκπροώπησης των τιμών στα δείγματα εκπαίδευσης φαίνεται να είναι αντιπροσωπευτική με σημαντικές ωστόσο αποκλίσεις από την πραγματικότητα (χρήσεις γης).

# 5 ΣΥΜΠΕΡΑΣΜΑΤΑ-ΣΥΖΗΤΗΣΗ

Στην τελευταία ενότητα θα συζητήσουμε τα συμπεράσματα που εξήχθησαν σύμφωνα με τα αποτελέσματά, στην προηγούμενη ενότητα. Στην εργασία πραγματοποιήθηκε η ταξινόμησης των Χρήσεων Γης της δορυφορικής εικόνας, με τη χρήση 8 αλγορίθμων ταξινόμησης για τη Νήσο Λέσβου, με δεδομένα του δορυφόρου Sentinel 2-A, Χρήσεων Γής (Corine 2012) και με τη χρήση του στατιστικού πακέτου Rstudio. Η στρατηγικής για την εύρεση των βέλτιστων δειγμάτων εκπαίδευσης βασίζεται στην παραδοχή της ορθότητας του Corine Land Cover (CLC) 2012 και στη χρήση τεχνικών τυχαίας δειγματοληψίας (Random Sampling ). Η εφαρμογή της τυχαίας δειγματοληπτικής μεθόδου πραγματοποιήθηκε για 1000 σημεια και κατά τη διαδικασία εύρεσης αντιπροσωπευτικών δεδομένων εκπαίδευσης έχει μοναδικό κριτήριο το εμβαδόν που καταλαμβάνει η κάθε χρήση του CLC .

Ωστόσο, για την ταξινόμηση της δορυφορικής εικόνας για τη Νήσος Λέσβο, σε δεκατρείς κατηγορίες με την τοποθέτηση 1000 τυχαίων σημείων εμφάνισε προβλήματα, όπως της μηδενικής εκπροσώπησης κατηγοριών, υπο-εκπροσώπησης αλλά και υπερ-εκπροσώπησης σε άλλες υποκατηγορίες. Το φαινόμενο αυτό επηρέασε αρνητικά την αξιοπιστία και την εγγυρότητα των αποτελεσμάτων. Έτσι μεγάλες εκτάσεις, όπως ετερογενείς γεωργικές εκτάσεις, δάση, ποώδης βλάστηση και καλλιέργιες συγκέντρωσαν το μεγαλύτερο μέρος των σημείων και αποτυπώθηκαν με σχετικά αντιπροσωπευτική ακρίβεια, σε αντίθεση με τις υπόλοιπες, που παρουσιάζουν μεγάλες αποκλίσεις. Τα παραπάνω επιβεβαιώνονται απο τις τιμές των στατιστικών μέτρων εκτίμησης ακρίβειας που είναι εξαιρετικά χαμηλές. Πιό συγκεκριμένα οι τιμές της στατιστικής Kappa είχαν οριακά αποδεκτές τιμές που κυμαίνονται περίπου από 0.2-04. Όπως αναφέρεται, Ο Cohen πρότεινε το αποτέλεσμα Kappa να ερμηνευτεί ως εξής: οι τιμές ≤ 0 δεν δείχνουν συμφωνία και 0,01-0,20 ως μη ελαφρές, 0,21-0,40 ως δίκαιες, 0,41- 0,60 ως μέτριες, 0,61-0,80 ως ουσιαστικές και 0,81-1,00 ως σχεδόν τέλεια συμφωνία (McHugh, 2012) Έτσι σε όλα τα σετ εικόνων παρουσιάζονται σημαντικές διαφοροποιήσεις στην κατανομή των εικονοστοιχείων μεταξύ των κατηγοριών πράγμα που οπτικά εντοπίζεται συγκριτικά με τα όρια του χάρτη χρήσεων γης.

# 6 ΒΕΛΤΙΩΣΕΙΣ

Ύστερα από τα αποτελέσματα και τα συμπεράσματα που εξήχθησαν στα πλαίσια εκπόνησης αυτής της διπλωματικής εργασίας μπορούν να αναφερθούν ορισμένες βελτιώσεις για την αύξηση της ακρίβειας των αποτελεσμάτων αυτών. Σύμφωνα με την ταξινόμηση που πραγματοποιήθηκε, κατά την οποία ορισμένες υποκατηγορίες των χρήσεων γης δεν εκπροσωπήθηκαν, ενδεχομένως, η τοποθέτηση περισσότερων τυχαίων σημείων, θα μπορούσε να προσφέρει καλύτερα αποτελέσματά.

Στην παρούσα διπλωματική χρησιμοποιήθηκε το δεύτερο επίπεδο πληροφορίας των χρήσεων Γης με αποτέλεσμα να εμφανιστούν δεκατρείς υποκατηγορίες ταξινόμησης. Επίσης, παρατηρείται πως για την ταξινόμηση των χρήσεων γης έχουν πραγματοποιηθεί μελέτες οι οποίες πραγματοποιούν ταξινόμηση για μία μοναδική κλάση χρησιμοποιώντας την μέθοδος One Class Classification (OCC) καθώς όπως αναφέρεται είναι πολύ συχνό το φαινόμενο της ταξινόμησης μόνο του αστικού τοπίου ή μόνο των αγροτικών περιοχών (Deng et al 2018). Επίσης όπως υποστηρίζεται παρόλο που μια επιβλεπόμενη ταξινόμηση παρουσιάζει πολύ ελπιδοφόρες επιδόσεις όσον αφορά την ακρίβεια της ταξινόμησης, επικεντρώνεται κυρίως στην ταξινόμηση πολλών κατηγοριών. Οι ταξινομητές πολλαπλών τάξεων απαιτούν εξαντλητική επισήμανση όλων των κατηγοριών που υπάρχουν σε μια περιοχή μελέτης. Επιπλέον, ο στόχος σε πολλές περιπτώσεις είναι να βελτιστοποιηθεί η ακρίβεια ταξινόμησης για όλες τις κατηγορίες κάλυψης γης και όχι για μια συγκεκριμένη κλάση ή λίγες κατηγορίες ενδιαφέροντος. Ωστόσο, σε πολλές εφαρμογές, είναι δύσκολο να συλλεχθούν δεδομένα αναφοράς για όλες τις κατηγορίες κάλυψης γης στην περιοχή μελέτης. Είναι επίσης πολύ συνηθισμένο το γεγονός ότι σε πολλές εφαρμογές δεν επικεντρώνεται σε όλες τις κατηγορίες κάλυψης γης. Αντίθετα, μόνο μία συγκεκριμένη τάξη ή λίγες τάξεις έχουν πραγματικό ενδιαφέρον (Song, B. et al., 2016). Ενδεχομένως, μια ξεχωριστή ταξινόμηση ανά κατηγορία των χρήσεων γης θα μπορούσε να προσφέρει συνολικά μεγαλύτερη ακρίβεια στην μελέτης περίπτωσης που πραγματοποιήθηκε στην παρούσα εργασία.

Τέλος μια επίσης ενδιαφέρουσα προσέγγιση θα ήταν η εφαρμογή μια μεθοδολογίας που αποδεικνύει πως η ενσωμάτωση των κλιματολογικών και τοπογραφικών συνθηκών μπορεί να βοηθήσει στην οριοθέτηση της αλληλεπικαλυπτόμενη πληροφορίας, όπως παρουσιάζεται στη μελέτη των Saadat, H. et al. (2011) όπου με τη χρήση τριών

δορυφορικών εικόνων Landsat, την άνοιξη, το καλοκαίρι και στο τέλος του καλοκαιριού, πραγματοποιήθηκε ταξινόμηση των χρήσεων γης και επισημαίνεται πως θα μπορούσε η ίδια ανάλυση να πραγματοποιηθεί για διαφορετικές χρονικές περιόδους με τη χρήση πολλών δορυφορικών εικόνων και για διαφορετικές κλιματικές ζώνες.

# 7 ΒΙΒΛΙΟΓΡΑΦΙΑ

## 7.1 Ξενόγλωσση

Potapov, P. et al. (2008) Mapping the World's Intact Forest Landscapes by Remote Sensing, Ecology and Society. doi: 51.

Lillesand, T. M. and Kiefer, R. W. (1994) Remote sensing and image interpretation -- 3rd ed., John Wiley and Sons, Inc., New York.

Witten, I. H., Frank, E. and Hall, M. a (2005) Data Mining:Practical Machine Learning Tools and Techniques second edition, Complementary literature None. doi: 0120884070, 9780120884070.

Sumathi, S. and Sivanandam, S. N. (2006) Introduction to Data Mining and its Applications. Doi: 10.1007/978-3-540-34351-6.

Berry, M. A. (2000) 'Mastering Data Mining: The Art and Science of Customer Relationship Management', Industrial Management & Data Systems. doi: 10.1108/imds.2000.100.5.245.2.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery: an overview', in Advances in knowledge discovery and data mining.

Fayyad, U. (1997) 'Knowledge discovery in databases: An overview', in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Doi: 10.1007/3540635149_30.

Jiawei Han, M. K. and Pei, J. (2012) 'Data Mining: Concepts and Techniques, Third Edition - Books24x7', Morgan Kaufmann Publishers. doi: 10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Rokach, L. (2009) 'Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography', Computational Statistics and Data Analysis. doi: 10.1016/j.csda.2009.07.017.

Zhou, Z. H. (2012) Ensemble methods: Foundations and algorithms, Ensemble Methods: Foundations and Algorithms. doi: 10.1201/b12207.

Breiman, L. (1996) 'Bagging predictors - Springer', Machine learning. doi: 10.1007/BF00058655.

Breiman, L. (2001) 'Randomforest2001', Machine Learning. doi: 10.1017/CBO9781107415324.004.

Cunningham, P. and Delany, S. J. (2007) 'k-Nearest Neighbour Classifiers', Multiple Classifier Systems. doi: 10.1016/S0031-3203(00)00099-6.

Liu, Y. (2014) 'Random forest algorithm in big data environment', Computer Modelling & New Technologies, 18, pp. 147–151

Friedman, J. H. (2002) 'Stochastic gradient boosting', Computational Statistics and Data Analysis. doi: 10.1016/S0167-9473(01)00065-2.

Miner, G., Nisbet, R. and Elder, J. (2009) Handbook of Statistical Analysis and Data Mining Applications, Handbook of Statistical Analysis and Data Mining Applications. doi: 10.1016/B978-0-12-374765-5.X0001-0.

McHugh, M. L. (2012) 'Interrater reliability: the kappa statistic.', *Biochemia medica*.

Deng, X. et al. (2018) 'One-class remote sensing classification: One-class vs. Binary classifiers', International Journal of Remote Sensing. Taylor & Francis, 39(6), pp. 1890–1910. doi: 10.1080/01431161.2017.1416697.

Song, B. et al. (2016) 'One-Class Classification of Remote Sensing Images Using Kernel Sparse Representation', IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(4), pp. 1613–1623. doi: 10.1109/JSTARS.2015.2508285.

Saadat, H. et al. (2011) 'Land use and land cover classification over a large area in Iran based on single date analysis of satellite imagery', ISPRS Journal of Photogrammetry and Remote Sensing. doi: 10.1016/j.isprsjprs.2011.04.001.

## 7.2 Ελληνική

Παρχαρίδης, Ι. (2015) Αρχές Δορυφορικής Τηλεπισκόπησης Θεωρία Και Εφαρμογές. Available at: www.kallipos.gr.

Κύρκος, Ε. Γ. (2015) Επιχειρηματική Ευφυΐα & Εξόρυξη Δεδομένων.

## 7.3 Ηλεκτρονικές Πηγές

K. M. Ting, I. H. Witten(1999), Issues in Stacked Generalization access : https://jair.org/index.php/jair/article/view/10228 [accessed Jun, 2019]

Qiang Yang "Classification with Decision Trees II"— Presentation transcript, access: https://slideplayer.com/slide/5121894, [accessed Jun, 2019]

Xiaowei Guan (2012), Splitting random forest (SRF) for determining compact sets of genes that distinguish between cancer subtypes, Flowchart of the Splitting Random Forest (SRF) Algorithm. https://www.researchgate.net/figure/Flowchart-of-the-Splitting-Random-Forest-SRF-Algorithm_fig1_225055544 [accessed Jun, 2019]

Packt edditorial Staff (2016), Support Vector Machines as a Classification Engine, https://hub.packtpub.com/support-vector-machines-classification-engine/ [accessed Jun, 2019]

Sebastian, Raschka (2014-2019), StackingClassifier, An ensemble-learning meta-classifier for stacking. https://rasbt.github.io/mlxtend/user_guide/classifier/StackingClassifier/ [accessed Jun, 2019]

Srishti Sawla (2018), Linear Discriminant Analysis, https://medium.com/@srishtisawla/linear-discriminant-analysis-d38decf48105 [accessed Jun, 2019]

Packt edditorial Staff (2019), Brett Lantz on implementing a decision tree using C5.0 algorithm in R, https://hub.packtpub.com/brett-lantz-on-implementing-a-decision-tree-using-c5-0-algorithm-in-r/ [accessed Jun, 2019]

Jason Brownlee (2016), How to Build an Ensemble Of Machine Learning Algorithms in R https://machinelearningmastery.com/machine-learning-ensembles-with-r/ [accessed May, 2019]

Jason Brownlee (2016) Machine Learning Evaluation Metrics in R, https://machinelearningmastery.com/machine-learning-evaluation-metrics-in-r/ [accessed May, 2019]

Brad Boehmke & Brandon Greenwell (2019), Hands-on Machine Learning with R, https://bradleyboehmke.github.io/HOML/DT.html#structure [accessed May, 2019]

Brad Boehmke & Brandon Greenwell (2019), Hands-on Machine Learning with R, https://bradleyboehmke.github.io/HOML/gbm.html [accessed May, 2019]

Sunil Ray (2017), Understanding Support Vector Machine algorithm from examples (along with code), https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/, [accessed Jun, 2019]

https://earthexplorer.usgs.gov
https://earth.esa.int
https://www.copernicus.eu/en
https://www.r-project.org/
https://www.qgis.org

## 7.4 Διατριβές

Λάσπιας Ευστάθιος (2012) Επιβλεπόμενη και μη επιβλεπόμενη ταξινόμηση πολυφασματικών εικόνων τηλεπισκόπησης και θεματικές εφαρμογές στο χώρο τους : ανάπτυξη σε περιβάλλον wiki, ΕΜΠ, Σχολή Αγρονόμων και Τοπογράγων Μηχανικών, Εργαστήριο Τηλεπισκόπισης

Παπαποστόλου Μαρία (2017) Μεταπτυχιακή διπλωματική εργασία: Κατηγοριοποίηση με μηχανές διανυσμάτων υποστήριξης, ΑΠΘ, Σχολή Θετικών Επιστημών, Τμήμα Μαθηματικών, ΠΜΣ Στατιστική και μοντελοποίηση

Τσαγκαλίδης ΑΘ. (2013). Μεταπτυχιακή διπλωματική διατριβή : Στρατηγική μελέτη περιβαλλοντικών επιπτώσεων σχεδίου διαχείρισης υδατικών πόρων Ν. Λέσβου, ΑΠΘ, Τμήμα Αγρονόμων και Τοπογράφων Μηχανικών, ΠΜΣ Γεωπληροφορική Υδατικών Πόρων

Κουρής Ι. (2006) Διδακτορική Διατριβή : Εφαρμογή Τεχνικών Data Mining σε Συστήματα Ηλεκτρονικού Εμπορίου, Πανεπιστήμιο Πατρών, Πολυτεχνική Σχολή, Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής.

# 8 ΚΩΔΙΚΑΣ

```
#
#  Initialize
#

# Set Working Directory
setwd(".\comp2")

# load Libraries
library(rgdal)
library(sp)
library(raster)
library(caret)
library(mlbench)

# import/load Remote Sence Data .tif
rst <- stack("./Composite_2_3_4_8_10m_2100_clip.tif")
rst_20m <- stack("./Composite_5_6_7_8A_11_12_20m_2100_clip.tif")
rst_60m <- stack("./Composite_1_9_10_60m_2100_clip.tif")
rs <- list(rst, rst_20m, rst_60m)

# load .shp file
clc <- readOGR("./les_clip.shp")
```

```
#
#  Calculate how many Random Points Each area will get
#

# set random Seed (so we have random data that we can reproduce for all tesings)
set.seed(1821)

# calculate how many random points each area will get
clc$pc <- clc$area/sum(clc$area)
n<-1000
clc$pts <- clc$pc * 1000
sum(clc$pts)
clc$pts_integer <- round(clc$pts)
sum(clc$pts_integer)

# becaouse by rounding the numbers [round(clc$pts)] we loose some points
# we add them manually where they shuld be
d <- n - sum(clc$pts_integer)
clc$diaf <- clc$pts-clc$pts_integer
diaf <- sort(clc$diaf, decreasing = T)
for (i in 1:d) {
  clc$pts_integer[clc$diaf==diaf[i]] <- clc$pts_integer[clc$diaf==diaf[i]] + 1
}

# check that we have 1000 points
sum(clc$pts_integer)
```

```
#
#  Add the Train Data
#

# locate diferent/unique land use
un <- as.numeric(as.character(unique(clc$level_12)))
ls <- list()
rp <- list()
df <- data.frame()
vl_10m <- 0
vl_20m <- 0
vl_60m <- 0
m<-0

# Place Random Data
train_10m <- data.frame()
train_20m <- data.frame()
train_60m <- data.frame()
for (i in 1:length(un)) {
  nm <- un[i]
  ls[[i]] <- subset(clc, clc$level_12==nm)
  n <- sum(as.numeric(ls[[i]]$pts_integer))
  m <- m +n
  print(n)
  print(nm)
  df <- rbind(df, nm, n)
  if (n>0) {
    rp[[i]] <- spsample(ls[[i]], n=n, type = "random", iter=100)
    plot(rp[[i]])
    vl_10m <- extract(rst,rp[[i]])
```

```
      vl_10m <-  cbind(vl_10m,class=rep(nm))
      vl_20m <- extract(rst_20m,rp[[i]])
      vl_20m <-  cbind(vl_20m,class=rep(nm))
      vl_60m <- extract(rst_60m,rp[[i]])
      vl_60m <-  cbind(vl_60m,class=rep(nm))
      train_10m <- rbind(train_10m, vl_10m)
      train_20m <- rbind(train_20m, vl_20m)
      train_60m <- rbind(train_60m, vl_60m)
    }
  }

  # clear NA (Missing Values) Data
  train_10m[is.na(train_10m)] <- 0
  train_20m[is.na(train_20m)] <- 0
  train_60m[is.na(train_60m)] <- 0

  # Formula
  frml_10m <- formula(as.factor(class) ~ Composite_2_3_4_8_10m_2100_clip.1 +
Composite_2_3_4_8_10m_2100_clip.2 + Composite_2_3_4_8_10m_2100_clip.3 +
Composite_2_3_4_8_10m_2100_clip.4)
  frml_20m <- formula(as.factor(class) ~
Composite_5_6_7_8A_11_12_20m_2100_clip.1 +
Composite_5_6_7_8A_11_12_20m_2100_clip.2 +
Composite_5_6_7_8A_11_12_20m_2100_clip.3 +
Composite_5_6_7_8A_11_12_20m_2100_clip.4 +
Composite_5_6_7_8A_11_12_20m_2100_clip.5 +
Composite_5_6_7_8A_11_12_20m_2100_clip.6)
  frml_60m <- formula(as.factor(class) ~ Composite_1_9_10_60m_2100_clip.1 +
Composite_1_9_10_60m_2100_clip.2 + Composite_1_9_10_60m_2100_clip.3)
  frml <- list(frml_10m, frml_20m, frml_60m)
  # train data
  train <- list(train_10m, train_20m, train_60m)
```

```
# Print the Points
print(m)



# run a test to check the Model
set.seed(1821)
control <- trainControl(method="repeatedcv", number=10, repeats=3)
metric <- c("Accuracy")




#
#  Train the Models
#

fit.lda <- list()
fit.svmRadial <- list()
fit.knn <- list()
fit.cart <- list()
fit.treebag <- list()
fit.rf <- list()
fit.gbm <- list()
fit.c50 <- list()
results <- list()



# loop throw the different formula and train
for (i in 1:3) {
```

```
# train the models
 fit.lda[[i]] <- train(frml[[i]], data=train[[i]], method="lda", metric=metric, preProc=c("center", "scale"), trControl=control) # lda - Linear Discriminant Analysis
  fit.svmRadial[[i]] <- train(frml[[i]], data=train[[i]], method="svmRadial", metric=metric, preProc=c("center", "scale"), trControl=control, fit=FALSE) # SVM Radial
 fit.knn[[i]] <- train(frml[[i]], data=train[[i]], method="knn",metric=metric, preProc=c("center", "scale"), trControl=control) # kNN
 fit.cart[[i]] <- train(frml[[i]], data=train[[i]], method="rpart",metric=metric, preProc=c("center", "scale"), trControl=control) # CART
fit.treebag[[i]] <- train(frml[[i]], data=train[[i]], method="treebag", metric=metric, preProc=c("center", "scale"), trControl=control) # Bagged CART
  fit.rf[[i]] <- train(frml[[i]], data=train[[i]], method="rf", metric=metric, preProc=c("center", "scale"), trControl=control) # Random Forest
 fit.gbm[[i]] <- train(frml[[i]], data=train[[i]], method="gbm", metric=metric, preProc=c("center", "scale"), trControl=control,verbose=FALSE) # Stochastic Gradient Boosting (Generalized Boosted Modeling)
 fit.c50[[i]] <- train(frml[[i]], data=train[[i]], method="C5.0", metric=metric, preProc=c("center", "scale"), trControl=control,verbose=FALSE) # C5.0



# var results will be used to Summarise and Evaluate the Results
results[[i]] <- resamples(list(lda=fit.lda[[i]],
                               svm=fit.svmRadial[[i]],
                               knn=fit.knn[[i]],
                               cart=fit.cart[[i]],
                               bagging=fit.treebag[[i]],
                               rf=fit.rf[[i]],
                               gbm=fit.gbm[[i]],
                               c50=fit.c50[[i]]))
```

```
}

# Results for 10m bands
rslt_1 <- summary(results[[1]])
bwplot(results[[1]])
dotplot(results[[1]])
modelCor(results[[1]])
splom(results[[1]])

# Results for 20m bands
rslt_2 <- summary(results[[2]])
bwplot(results[[2]])
dotplot(results[[2]])
modelCor(results[[2]])
splom(results[[2]])

# Results for 60m bands
rslt_3 <- summary(results[[3]])
bwplot(results[[3]])
dotplot(results[[3]])
modelCor(results[[3]])
splom(results[[3]])

#
# Evaluate The Results
#

beginCluster()
lda <- list()
svm <- list()
knn <- list()
cart <- list()
```

```
tb <- list()
rf <- list()
gbm <- list()
c50<- list()

# Evaluation Each Model
for (i in 1:3) {
  lda[[i]] <- clusterR(rs[[i]], raster::predict, args = list(model = fit.lda[[i]]))
  svm[[i]] <- clusterR(rs[[i]], raster::predict, args = list(model = fit.svmRadial[[i]]))
  knn[[i]] <- clusterR(rs[[i]], raster::predict, args = list(model = fit.knn[[i]]))
  cart[[i]] <- clusterR(rs[[i]], raster::predict, args = list(model = fit.cart[[i]]))
  tb[[i]] <- clusterR(rs[[i]], raster::predict, args = list(model = fit.treebag[[i]]))
  rf[[i]] <- clusterR(rs[[i]], raster::predict, args = list(model = fit.rf[[i]]))
  gbm[[i]] <- clusterR(rs[[i]], raster::predict, args = list(model = fit.gbm[[i]]))
  c50[[i]] <- clusterR(rs[[i]], raster::predict, args = list(model = fit.c50[[i]]))
}
endCluster()

# Show Results and Evaluation of each model
for (i in 1:3) {
   plot(lda[[i]], main=paste0("LDA ",i), breaks = c(10, 20, 30, 40, 50, 60), col=ter-
rain.colors(5))
   plot(svm[[i]],main=paste0("SVM ", i), breaks = c(10, 20, 30, 40, 50, 60), col=ter-
rain.colors(5))
   plot(knn[[i]],main=paste0("KNN ", i), breaks = c(10, 20, 30, 40, 50, 60), col=ter-
rain.colors(5))
   plot(cart[[i]],main=paste0("CART ", i), breaks = c(10, 20, 30, 40, 50, 60), col=ter-
rain.colors(5))
   plot(tb[[i]],main=paste0("TB ", i), breaks = c(10, 20, 30, 40, 50, 60), col=terrain.-
colors(5))
   plot(rf[[i]],main=paste0("RF ", i), breaks = c(10, 20, 30, 40, 50, 60), col=terrain.-
colors(5))
```

```
    plot(gbm[[i]],main=paste0("GBM ", i), breaks = c(10, 20, 30, 40, 50, 60), col=ter-
rain.colors(5))
    plot(c50[[i]],main=paste0("GBM ", i), breaks = c(10, 20, 30, 40, 50, 60), col=ter-
rain.colors(5))
  }
```