



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ – ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

«ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΓΕΩΡΓΙΑ
ΑΚΡΙΒΕΙΑΣ»

Κωνσταντίνα Γ. Γκάνια

Επιβλέπων Καθηγητής: Θ. Έξαρχος

Κέρκυρα 2020-21

Περιεχόμενα

Περίληψη	3
Abstract	4
Εισαγωγή.....	5
ΚΕΦΑΛΑΙΟ 1	7
1.1. Γεωργία Ακριβείας	7
1.2. Αντίκτυπο της τεχνητής νοημοσύνης στη γεωργία ακριβείας.....	10
1.3. Ψηφιακή γεωργία με machine learning.....	11
ΚΕΦΑΛΑΙΟ 2	13
2.1. Machine learning σε agricultural δεδομένα.....	13
2.2. Ανάλυση ερευνών	19
2.3. Συμπεράσματα	24
ΚΕΦΑΛΑΙΟ 3	25
3.1. Μεθοδολογία	25
3.2. Ανάλυση μεθοδολογίας	26
3.2.1. Διαδικασία και μέτρα αξιολόγησης	27
3.2.2. Το περιβάλλον WEKA	27
ΚΕΦΑΛΑΙΟ 4	27
4.1. Περιγραφή αλγορίθμων.....	27
4.1.1. Δέντρα απόφασης (C4.5).....	29
4.1.2. Random Forest	30
4.1.3. kNN	30
4.1.4. Naive Bayes	30
4.2. Μέτρα αξιολόγησης	31
4.2.1. Πίνακας σύγχυσης (Confusion Matrix).....	31
4.2.2. Ακρίβεια (Accuracy).....	31
4.2.3. Rating.....	32
4.2.4. Ορθότητα (Precision).....	32
4.2.5. Εμβαδόν κάτω από την καμπύλη ROC	32
4.2.6. Ανάκληση (Recall).....	33
4.2.7. 10-fold Cross Validation	34
ΚΕΦΑΛΑΙΟ 5	35
5.1. Eucalyptus Dataset	35
5.1.1. Περιγραφή συνόλου δεδομένων	35
5.1.2. Πληροφορίες παραδειγμάτων και χαρακτηριστικώ	35

5.1.3. Διαδικασία ανάλυσης δεδομένων	38
5.1.4. Αποτελέσματα	39
5.1.5. Συμπεράσματα	43
5.2. Grub-damage Dataset	44
5.2.1. Περιγραφή συνόλου δεδομένων	44
5.2.2. Πληροφορίες παραδειγμάτων και χαρακτηριστικών	44
5.2.3. Διαδικασία ανάλυσης δεδομένων	46
5.2.4. Αποτελέσματα	47
5.2.5. Συμπεράσματα	50
5.3. Pasture Dataset.....	51
5.3.1. Περιγραφή συνόλου δεδομένων	51
5.3.2. Πληροφορίες παραδειγμάτων και χαρακτηριστικών	51
5.3.3. Διαδικασία ανάλυσης δεδομένων	53
5.3.4. Αποτελέσματα	54
5.3.5. Συμπεράσματα	57
ΚΕΦΑΛΑΙΟ 6.....	58
6.1. Σύνοψη και συμπεράσματα	58
6.2. Μελλοντικές κατευθύνσεις	59
ΒΙΒΛΙΟΓΡΑΦΙΑ	61

Περίληψη

Η γεωργία ακριβείας παίζει ζωτικό ρόλο στην οικονομική ανάπτυξη κάθε χώρας. Με την αύξηση του πληθυσμού, τις συχνές αλλαγές στις κλιματολογικές συνθήκες και τους περιορισμένους πόρους καθίσταται ένα δύσκολο έργο για την εκπλήρωση του φαγητού που απαιτείται σήμερα για τον πληθυσμό. Η γεωργία ακριβείας γνωστή και ως έξυπνη γεωργία εμφανίστηκε ως καινοτόμο εργαλείο για την αντιμετώπιση των προκλήσεων στη γεωργική βιωσιμότητα. Η τεχνολογία αιχμής που χρησιμοποιείται είναι η μηχανική μάθηση, η οποία δίνει στη μηχανή τη δυνατότητα να μαθαίνει χωρίς να είναι ρητά προγραμματισμένη. Η μηχανική μάθηση μαζί με τα αγροτικά μηχανήματα με δυνατότητα IoT (Internet of Things) [1] είναι βασικά συστατικά του επόμενου επαναστάτη στη γεωργία. Σε αυτή την εργασία παρουσιάζεται μια συστηματική ανασκόπηση εφαρμογών ML στον τομέα της γεωργίας. Οι περιοχές που επικεντρώνονται είναι η πρόβλεψη παραμέτρων του εδάφους, η πρόβλεψη απόδοσης καλλιέργειας, η ανίχνευση ασθενειών και ζιζανίων στις καλλιέργειες και η ανίχνευση ειδών. Η μηχανική μάθηση με την μηχανική όραση αναθεωρείται για την ταξινόμηση ενός διαφορετικού συνόλου εικόνων και καλλιεργειών προκειμένου να παρακολουθείται η ποιότητα της καλλιέργειας και η εκτίμηση απόδοσης. Τέλος, περιγράφεται πως μπορεί η γεωργία που βασίζεται στη γνώση, να βελτιώσει την βιώσιμη παραγωγικότητα και την ποιότητα του προϊόντος.

Abstract

Precision agriculture plays a vital role in the economic development of any country. With population growth, frequent changes in climatic conditions and limited resources, it becomes a difficult task to meet the food demand of the population today. Precision agriculture, also known as smart farming, has emerged as an innovative tool for tackling the challenges to agricultural sustainability. The cutting-edge technology used is machine learning, which enables the machine to learn without being explicitly programmed. Machine learning along with IoT (Internet of Things) [1] agricultural machinery are key components of the next revolution in agriculture. This paper presents a systematic review of ML applications in agriculture. Areas of focus are soil parameter prediction, crop disease, weed and species detection. Machine learning with computer vision is revised to classify a different set of images and crops in order to monitor crop quality and yield assessment. Finally, it is described how knowledge-based agriculture can improve sustainable productivity and product quality.

Εισαγωγή

Πριν από την εκμηχάνιση της γεωργίας και την αύξηση του μεγέθους των αγροκτημάτων, ο γεωργός γνώριζε το χωράφι του, καθώς το περπατούσε διαρκώς στη διάρκεια της βλαστικής περιόδου εκτελώντας τις καλλιεργητικές εργασίες και είχε τη δυνατότητα να διαχειρίζεται τα διάφορα μέρη του σύμφωνα με τις ανάγκες που θεωρούσε ότι είχαν. Για παράδειγμα, μπορεί να έριχνε περισσότερο σπόρο σε εκείνα τα σημεία του χωραφιού που δεν φύτρωναν, έριχνε περισσότερο λίπασμα εκεί που τα φυτά φαινόταν αδύναμα και κατέστρεφε τα ζιζάνια όπου τα έβρισκε. [2]

Με την εκμηχάνιση της γεωργίας και καθώς τα αγροκτημάτα μεγαλώνουν, ο γεωργός χάνει αυτή την άμεση αίσθηση του χωραφιού του. Μέχρι σήμερα η διαχείριση των αγροκτημάτων γίνεται με βάση τους μέσους όρους της παραγωγής, των ιδιοτήτων του εδάφους και των χαρακτηριστικών της καλλιέργειας. Η βασική υπόθεση είναι ότι οι αγροί είναι ομοιόμορφοι. Παρόλο που οι αγρότες γνώριζαν ότι υπήρχε ανομοιομορφία των αγρών τους, δεν είχαν αρκετές δυνατότητες να διαφοροποιήσουν τις καλλιεργητικές φροντίδες σύμφωνα με τις πραγματικές ανάγκες της καλλιέργειας σε κάθε σημείο του αγρού. Η τυπική διαχείριση στηρίζεται στη δειγματοληψία του εδάφους σε τυχαία σημεία του αγρού και καθορισμό της ποσότητας του λιπάσματος με βάση τον μέσο όρο των εδαφικών ιδιοτήτων. Έτσι, το λίπασμα διανέμεται ομοιόμορφα σε όλο τον αγρό, χωρίς να λαμβάνεται υπόψη η χωρική παραλλακτικότητα του εδάφους και της παραγωγής. Με τον τρόπο αυτό, σε μερικά σημεία του αγρού εφαρμόζεται μεγαλύτερη ποσότητα λιπάσματος από την απαιτούμενη με αποτέλεσμα την σπατάλη του λιπάσματος και μείωση της ποιότητας του παραγομένου προϊόντος και σε άλλες περιοχές μικρότερη ποσότητα από την απαιτούμενη, με αποτέλεσμα την μείωση της παραγωγής. Το ίδιο συνέβαινε και με τις υπόλοιπες εισροές, όπως τα φυτοφάρμακα και το νερό άρδευσης, τα οποία εφαρμόζονταν ομοιόμορφα στον αγρό.

Με την ενσωμάτωση και την εφαρμογή των νέων τεχνολογιών στη γεωργία έγινε επιτρεπτή η μέτρηση της χωρικής και χρονικής παραλλακτικότητας των παραμέτρων της παραγωγής και του εδάφους και έδωσαν τη δυνατότητα ανάπτυξης συστημάτων Γεωργίας Ακριβείας. Με τον όρο Γεωργία Ακριβείας ορίζουμε τη διαχείριση της χωρικής και χρονικής παραλλακτικότητας των αγρών, προκειμένου να βελτιωθεί η αποδοτικότητα των αγροκτημάτων και/ή να επιτευχθεί μείωση των αρνητικών επιπτώσεων στο περιβάλλον από την μη ορθολογική χρήση των εισροών. Η Γεωργία Ακριβείας είναι ένα σύστημα διαχείρισης αγροκτημάτων το οποίο χρησιμοποιώντας την πληροφορική, εφαρμοσμένη στη γεωργία, βοηθά τον γεωργό στη λήψη αποφάσεων για την καλύτερη διαχείριση του αγροκτήματος. Με τον όρο καλύτερη διαχείριση μπορεί να σημαίνει ότι θα γίνει βελτίωση της οικονομικής απόδοσης του αγροκτήματος, είτε με αύξηση της

παραγωγής, είτε με μείωση των εισροών, είτε με συνδυασμό και των δύο. Επιπλέον, μπορεί να επιτευχθεί βελτίωση των τυχόν αρνητικών επιπτώσεων της γεωργίας στο περιβάλλον, εφόσον εφαρμόζεται η αναγκαία ποσότητα εισροών σε κάθε σημείο του αγρού. Το χαρακτηριστικό είναι ότι, αντί να γίνονται οι καλλιεργητικές φροντίδες με βάση τις μέσες τιμές παραγωγής και γονιμότητας του εδάφους, είναι δυνατή η εφαρμογή διαφορετικών δόσεων εισροών και άλλων καλλιεργητικών φροντίδων για κάθε τμήμα του αγρού ανάλογα με τις πραγματικές του ανάγκες.

Ο πληθυσμός του κόσμου θα αυξηθεί στα 9.1 δισεκατομμύρια περίπου 34% από σήμερα μέχρι τα τέλη του 2050. Οι απαιτήσεις σε τρόφιμα θα αυξηθούν κατά 70% και λόγω της ταχείας αστικοποίησης η διαθέσιμη γη για την γεωργία θα μειωθεί δραστικά τα επόμενα χρόνια. [3]

Ο κύριος λόγος για τη μειωμένη παραγωγή φαγητού είναι η απουσία προγραμματισμού, οι απρόβλεπτες καιρικές συνθήκες, οι ακατάλληλες τεχνικές συγκομιδής και άρδευσης και η κακοδιαχείριση των ζώων. Η γεωργία ακριβείας είναι μια από τις λύσεις να διασφαλίσει κάποιος φαγητό για ολόκληρο τον κόσμο. Επίσης, η γεωργία ακριβείας συντομογραφείται ως ψηφιακή γεωργία και είναι ουσιαστικά η υιοθέτηση των σύγχρονων τεχνολογιών πληροφορικής, εργαλείων λογισμικού και έξυπνων ενσωματωμένων συσκευών για υποστήριξη αποφάσεων.

Στόχος του συγκεκριμένου συνόλου δεδομένων είναι να καθοριστεί ποια σπορόφυτα σε ένα είδος είναι καλύτερα για τη διατήρηση του εδάφους σε εποχιακά ξηρή χώρα. Ο προσδιορισμός βρίσκεται με μέτρηση του ύψους, της διαμέτρου κατά ύψος, της επιβίωσης και άλλων συντελεστών.

Οι πειραματικές διαδικασίες καταγραφής των δεδομένων που απαιτούν ιδιαίτερη προσοχή περιλαμβάνουν περιπτώσεις, που δεν καταγράφηκαν δεδομένα, λόγω πειραματικών διαδικασιών που δείχνουν ότι η απουσία ενός είδους από ένα αντίγραφο σε μια μια τοποθεσία αντιμετωπίστηκε ως τιμή που λείπει. Αν απουσιάζει από δύο ή περισσότερα αντίγραφα σε μια τοποθεσία τότε το είδος εξαιρείται από τις αναλύσεις της τοποθεσίας.

Τα δεδομένα που λείπουν (missing values) αφορούν στην επιβίωση, το σθένος, την αντοχή στα έντομα, τη μορφή στελέχους, τη μορφή στεφάνης και τη χρησιμότητα, ειδικά για τα δεδομένα που καταγράφηκαν στον σταθμό Morea. Αυτό θα μπορούσε να δείξει τον θάνατο των ειδών σε αυτές τις περιοχές ή την έλλειψη συλλογής δεδομένων.[4]

Στην παρούσα εργασία αναφέρονται και εξετάζονται διαφορετικές εφαρμογές μηχανικής μάθησης με αποτέλεσμα να προκύψουν κάποια συμπεράσματα για το πως οι τεχνολογίες θα ωφελήσουν την γεωργία. Επιπλέον, γίνεται αναφορά στην γνώμη της ερευνητικής κοινότητας σχετικά με την υιοθέτηση των ψηφιακών πρακτικών στο

σύστημα διαχείρισης της γεωργίας. Αναμένεται ότι οι κυβερνητικές υπηρεσίες θα πλαισιώσουν πολιτικές για την προώθηση της γεωργίας ακριβείας σε όλο τον κόσμο.

ΚΕΦΑΛΑΙΟ 1

1.1. Γεωργία Ακριβείας

Η Γεωργία Ακριβείας (Precision Agriculture) είναι μία νέα μέθοδος που διαχειρίζεται τους αγρούς σύμφωνα με την οποία οι εισροές¹ και οι καλλιεργητικές πρακτικές εφαρμόζονται ανάλογα με τις ανάγκες που έχει το κάθε έδαφος και οι καλλιέργειες, καθώς αυτές διαφοροποιούνται στον χώρο και τον χρόνο. Οι κύριοι στόχοι της γεωργίας ακριβείας είναι η αύξηση απόδοσης των καλλιεργειών, η βελτίωση της ποιότητας των παραγόμενων προϊόντων, η αποδοτική χρήση των αγροχημικών, η εξοικονόμηση ενέργειας και η προστασία του εδάφους και των νερών από τη ρύπανση.

Η διαχείριση των τμημάτων του αγρού σαν ξεχωριστή μονάδα δεν είναι καινούργια ιδέα. Η μικρότερη περιοχή που μπορεί η Γεωργία Ακριβείας να διαχειριστεί μπορεί να είναι κάθε φυτό μεμονωμένα και το έδαφος που επηρεάζει την ανάπτυξή του. Αυτή την τεχνική χρησιμοποιούσαν οι γεωργοί τα παλαιότερα χρόνια όταν έσπερναν με το χέρι κάθε φυτό. Σήμερα επειδή έχουν αυξηθεί οι καλλιεργούμενες εκτάσεις λόγω των δυνατοτήτων που έδωσε η εκμηχάνιση, για να γίνει διαχείριση σε επίπεδο φυτού χρειάζεται αναπτυγμένη τεχνολογία.

Η χωρική παραλλακτικότητα [5] είναι σημαντική προϋπόθεση για την εφαρμογή της Γεωργίας Ακριβείας και κατ' επέκταση την εφαρμογή των εισροών με μεταβλητές δόσεις²[6]. Ως χωρική παραλλακτικότητα ορίζεται η παραλλακτικότητα σε μετρούμενα χαρακτηριστικά της καλλιέργειας και του εδάφους στο χώρο. Η παραλλακτικότητα εντοπίζεται στους αγρούς και παρατηρείται στο πόσο γόνιμο και υγρό είναι το έδαφος, στη μηχανική σύστασή του, στην τοπογραφία, στην ανάπτυξη των φυτών και τους πληθυσμούς εχθρών και ασθενειών. Η παραλλακτικότητα μπορεί να είναι χωρική ή χρονική. Για παράδειγμα, μερικές εδαφικές ιδιότητες παραμένουν σταθερές με το χρόνο ή μεταβάλλονται ελάχιστα από χρόνο σε χρόνο, όπως η οργανική ουσία και η μηχανική σύσταση του εδάφους. Υπάρχουν και άλλες ιδιότητες, όπως τα επίπεδα των νιτρικών και η υγρασία του εδάφους, που μπορεί να αλλάζουν πολύ με την πάροδο του χρόνου. Η κατάσταση της καλλιέργειας μπορεί και αυτή να μεταβληθεί μέσα σε μικρό χρονικό διάστημα.

Η οικονομικότητα είναι ένας από τους πιο σημαντικούς λόγους που εξετάζονται για την μετάβαση από τον παραδοσιακό τρόπο διαχείρισης στη Γεωργία Ακριβείας.

¹ Φυτοφάρμακα, λιπάσματα, σπόροι, νερό άρδευσης

² Variable Rate Application: αναφέρεται στην εφαρμογή ενός υλικού, έτσι ώστε ο ρυθμός εφαρμογής να βασίζεται στην ακριβή τοποθεσία της εφαρμογής στη οποία εφαρμόζεται το υλικό.

Θετικά μπορεί να επηρεαστεί το κόστος παραγωγής και η πρόσοδος από την καλλιέργεια με την Γεωργία Ακριβείας. Για αυτό τον λόγο, υπάρχει δυνατότητα για μεγαλύτερες αποδόσεις χρησιμοποιώντας τις ίδιες εισροές, για ίδιες αποδόσεις με μειωμένες εισροές ή για μεγαλύτερες αποδόσεις με μειωμένες εισροές. Ο παραγωγός πρέπει να αποφασίσει για την καταλληλότερη μέθοδο διαχείρισης. Εκτός από την απόδοση, πρέπει να επιδιώξει καλύτερη ποιότητα προϊόντων χρησιμοποιώντας τις εισροές ανάλογα με τις ανάγκες των καλλιεργειών. Είναι γνωστό ότι τόσο η έλλειψη θρεπτικών στοιχείων όσο και η περίσσεια θρεπτικών συστατικών μπορεί να μειώσει την ανάπτυξη των φυτών και να χειροτερέψει την ποιότητα των προϊόντων καθώς και να οδηγήσει σε κακή ποιότητα καρπών αλλά και προβλήματα στη φυτεία αντίστοιχα.

Οι εισροές που εφαρμόζονται σήμερα με μεταβλητές δόσεις είναι τα λιπάσματα, τα φυτοφάρμακα, το νερό άρδευσης και ο σπόρος. Ένας από τους στόχους της εφαρμογής των εισροών με μεταβλητές δόσεις είναι η προστασία του περιβάλλοντος. Οι τεχνολογίες που χρησιμοποιεί η Γεωργία Ακριβείας έχουν σχέση με όλα τα στάδια παραγωγής από τη σπορά μέχρι τη συγκομιδή.

Μερικές από τις τεχνολογίες που χρησιμοποιούνται είναι οι εξής:

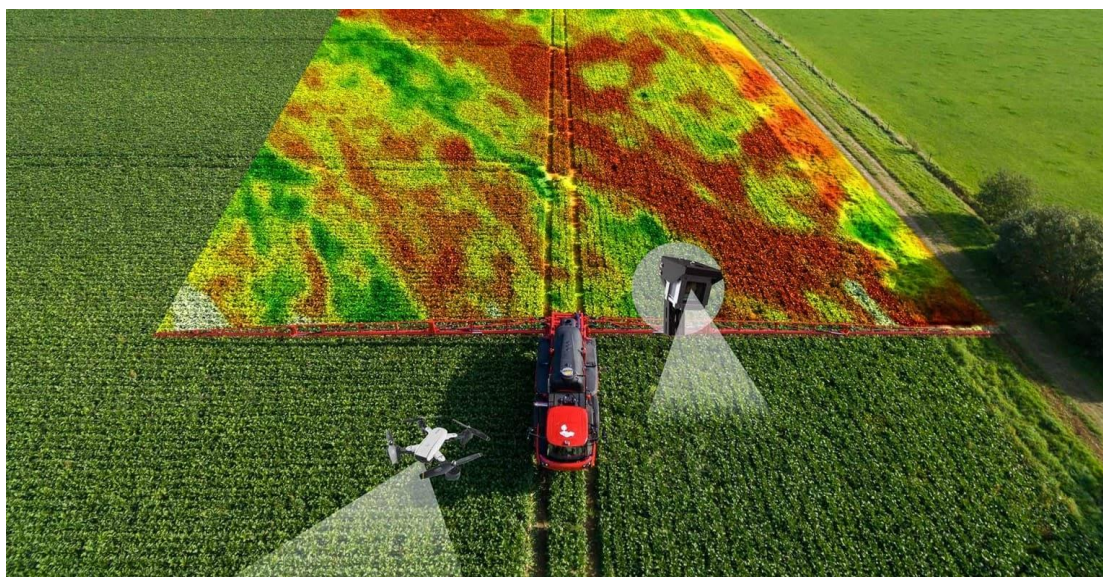
- GPS και GIS³ είναι συστήματα που επιτρέπουν την ακριβή χαρτογράφηση των αγρών και την ερμηνεία της παραλλακτικότητας των αγρών.
- Χαρτογράφηση παραγωγής, με την οποία γίνεται καταγραφή και συλλογή δεδομένων της παραγωγής από συγκεκριμένες θέσεις στον αγρό.
- Χαρτογράφηση εδαφικών ιδιοτήτων με την οποία γίνεται καταγραφή της γονιμότητας των αγρών.
- Μία από τις απλούστερες και οικονομικότερες μετρήσεις παραμέτρων του εδάφους που χρησιμοποιείται στη γεωργία ακριβείας αποτελεί η φαινομενική ηλεκτρική αγωγιμότητα. Η φαινομενική ηλεκτρική αγωγιμότητα του εδάφους ενοποιεί ένα ευρύτερο σύνολο παραγόντων που επηρεάζουν την παραγωγή μιας καλλιέργειας. Η χαρτογράφηση της ηλεκτρικής αγωγιμότητας μπορεί να πραγματοποιηθεί συνδέοντας μία συσκευή μέτρησης αγωγιμότητας σε παρελκόμενο όχημα και προσαρμόζοντας ένα δέκτη εντοπισμού γεωγραφικής θέσης (GPS).
- Τηλεπισκόπηση. Με την τηλεπισκόπηση συλλέγονται πληροφορίες για ένα αντικείμενο, χωρίς να υπάρχει επαφή. Οι δύο πιο συνηθισμένες μέθοδοι τηλεπισκόπησης είναι οι αεροφωτογραφίες και οι δορυφορικές εικόνες. Η ηλεκτρομαγνητική ακτινοβολία είναι το βασικό στοιχείο της τηλεπισκόπησης.

³ GPS/Global Positioning System: Παγκόσμιο Σύστημα Στιγματοθέτησης, GIS/Geographic Information System: Σύστημα Γεωγραφικών Πληροφοριών

Όταν η ηλεκτρομαγνητική ακτινοβολία έρχεται σε επαφή με ένα αντικείμενο, μπορεί να ανακλαστεί, να απορροφηθεί ή να διέλθει. Ανάλογα με το αντικείμενο στο οποίο προσπίπτει η ηλεκτρομαγνητική ακτινοβολία, διαφορετικά μήκη κύματος της ακτινοβολίας αντιδρούν με διαφορετικό τρόπο. Μετρώντας την ανακλώμενη ακτινοβολία από τα φυτά μπορεί να γίνει συλλογή πληροφοριών για την περιεκτικότητα των φυτικών ιστών σε νερό, για τη θρεπτική κατάσταση των φυτών και για άλλα χαρακτηριστικά των φυτών. Για αυτό το λόγο χρησιμοποιούνται οι δείκτες βλάστησης που είναι μαθηματικοί συνδυασμοί καναλιών. Με τον NDVI⁴ γίνεται μια αντικειμενική παρουσίαση της ζωντανής βλάστησης, ώστε να είναι εφικτή η παρακολούθηση της εξέλιξης της κατά τη διάρκεια του βιολογικού κύκλου των φυτών. Ακόμη, είναι δυνατά η αναγνώριση και ο διαχωρισμός των διαφόρων καλλιεργειών καθώς και η εισαγωγή δεδομένων τηλεπισκόπησης σε αγρο-κλιματικά μοντέλα.

- Variable Rate Application (VRA)[7]. Με την τεχνολογία αυτή οι εισροές εφαρμόζονται στον αγρό σε διαφορετικές δόσεις στις διάφορες περιοχές του αγρού ανάλογα με τις ανάγκες της κάθε περιοχής. Υπάρχουν δύο μέθοδοι VRA, οι οποίες είναι η βασισμένη σε χάρτες και η βασισμένη σε αισθητήρες. Η βασισμένη σε χάρτες απαιτεί έναν χάρτη εφαρμογής (prescription map) και ένα GPS που καθορίζει τη θέση στον αγρό. Καθώς το μηχάνημα που εφαρμόζει τις εισροές προχωρά στον αγρό, αλλάζει τη δόση με βάση τον χάρτη εφαρμογής (τις συντεταγμένες των ζωνών διαχείρισης). Η μέθοδος που βασίζεται σε αισθητήρες δεν απαιτεί ούτε χάρτη ούτε GPS. Αισθητήρες είναι τοποθετημένοι στη μηχανή εφαρμογής και μετρούν χαρακτηριστικά του εδάφους ή της καλλιέργειας, καθώς κινείται στον αγρό. Η πληροφορία μεταδίδεται σε ένα πρόγραμμα που υπολογίζει τις ανάγκες του εδάφους ή των φυτών και μεταφέρει την πληροφορία σε μια διάταξη εφαρμογής που διανέμει τις εισροές.

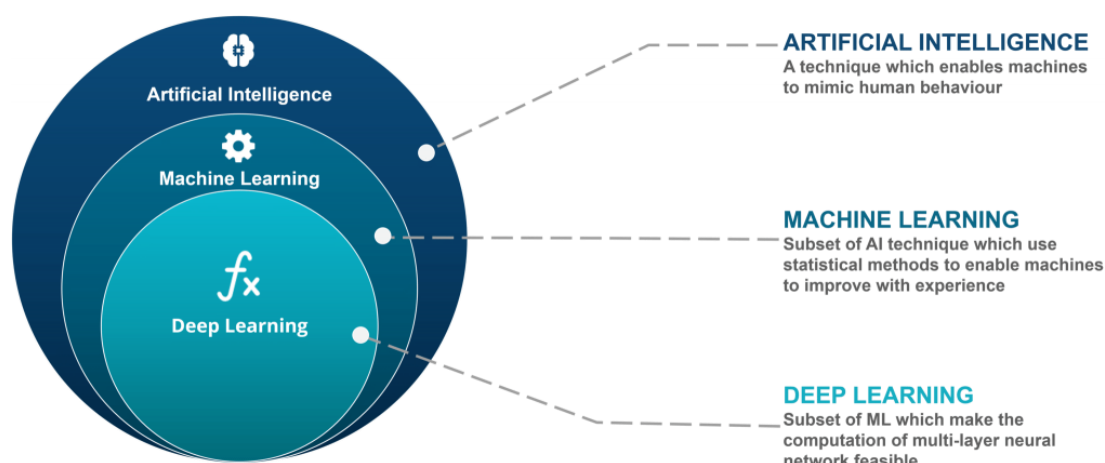
⁴ Στη Γεωργία Ακριβείας τα δεδομένα NDVI παρέχουν μια μέτρηση υγείας των καλλιεργειών.



Εικόνα 1: Ψηφιακή Γεωργία Ακριβείας

1.2. Αντίκτυπο της τεχνητής νοημοσύνης στη γεωργία ακριβείας

Η τεχνολογία ΑΙ παρέχει υπολογιστική νοημοσύνη στις μηχανές έτσι ώστε να μπορούν να μάθουν, να κατανοήσουν και να αντιδράσουν ανάλογα με την κατάσταση. Η μηχανική μάθηση και άλλες παρόμοιες τεχνικές είναι τα υποπεδία της τεχνητής νοημοσύνης.



Εικόνα 2: Τεχνικές της τεχνητής νοημοσύνης

Αυτό το πεδίο βρίσκει πάρα πολλές εφαρμογές σε διαφορετικούς τομείς της ανθρώπινης ζωής. Τώρα η γεωργική βιομηχανία υιοθετεί τις έξυπνες τεχνολογίες για

την αποτελεσματική καλλιέργεια βιολογικών προϊόντων σε περιορισμένες περιοχές, καθώς και για την υπέρβαση των παραδοσιακών προκλήσεων των αγροτών.

Το έξυπνο σύστημα γεωργίας που βασίζεται στο IoT έχει δημιουργηθεί για την παρακολούθηση των θρεπτικών συστατικών του εδάφους και την υγρασία του εδάφους χρησιμοποιώντας αισθητήρες. Οι αλγόριθμοι της μηχανικής μάθησης διερευνώνται για τον προσδιορισμό της βέλτιστης ποσότητας λιπασμάτων που απαιτούνται για τα εδάφη πριν από την σπορά των καλλιεργειών.[8]

1.3. Ψηφιακή γεωργία με machine learning

Η ψηφιακή γεωργία, με την ενσωμάτωση τεχνολογιών που βασίζονται στο Internet of Things (IoT), παρουσιάζει την ικανότητα αξιολόγησης ενός συστήματος σε πολλαπλά επίπεδα (ατομικό, τοπικό, περιφερειακό και παγκόσμιο) και δημιουργούν εργαλεία που επιτρέπουν τη βελτίωση της λήψης αποφάσεων σε κάθε περίπτωση. Πρόσφατες εξελίξεις στο υλικό IoT, όπως τα δίκτυα των ετερογενών ενσωματωμένων συσκευών και λογισμικών, όπως τους αλγορίθμους υπολογιστικής όρασης και οι βελτιστοποιημένες λύσεις του cloud, γίνεται δυνατή η συλλογή δεδομένων και η αποτελεσματική επεξεργασία δεδομένων από διάφορες πηγές σε συνδεδεμένη (έξυπνη) φάρμα.

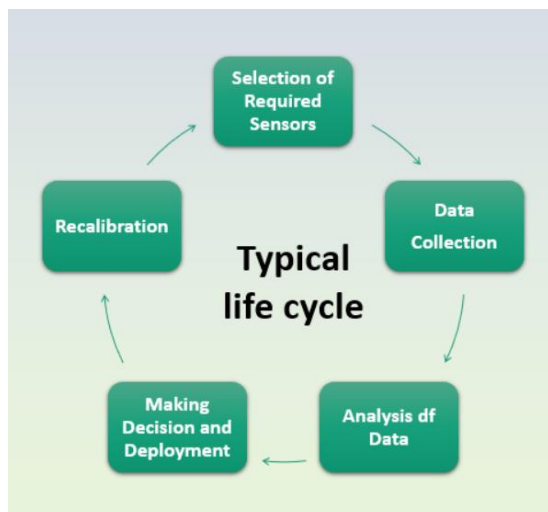
Συνδέοντας τις IoT συσκευές, συχνά σε μεγάλες εκτάσεις αγροτικής γης, είναι δυνατόν να συλλεχτούν δεδομένα από πολλαπλές γεωργικές εργασίες και σε διαφορετικό χρονικά πλαίσια, συμπεριλαμβανομένων σχεδόν σε πραγματικό χρόνο, δηλαδή, καθυστερήσεις μερικών δεκάδων δευτεριλέπτων.

Αυτά τα δεδομένα μπορούν στη συνέχεια να χρησιμοποιηθούν για χρήσιμες πληροφορίες, όπως, ακριβείς εφαρμογές συμπληρωμάτων εδάφους, μειώνοντας έτσι τα απόβλητα και τη ρύπανση του περιβάλλοντος. Τέτοιες γνώσεις θα είναι πιο ισχυρές κατά την αξιοποίηση των τεχνικών μάθησης για τη χρήση δεδομένων από πολλαπλά αγροκτήματα.

Συσκευές IoT που μπορούν να ανιχνεύσουν διαφορετικά είδη πληροφοριών μπορούν να συνδεθούν μεταξύ τους σε έναν ασύρματο αισθητήρα δικτύου (WSN) [9]. Για παράδειγμα, σε συστήματα καλλιέργειας σειράς, μεγάλα ποσά δεδομένων δημιουργούνται από μια ποικιλία επιχειρηματικών πεδίων και αποτελούν μια πλούσια πηγή για τη βελτιστοποίηση της συλλογής δεδομένων IoT και λειτουργίες ανάλυσης.

Αυτές οι λειτουργίες περιλαμβάνουν δειγματοληψία εδάφους, εφαρμογές λιπάσματος, φύτευση, ανίχνευση, ψεκασμό, συγκομιδή, διανομή και επεξεργασία. Γίνεται αξιοποίηση των διαφορετικών και συχνά ροών συνόλων δεδομένων που παράγονται από αγροκτήματα εξοπλισμένα με IoT και ακολουθούνται οι τύποι συνόλων δεδομένων σε τυπικές λειτουργίες πεδίου από τον κύκλο ζωής για τα δεδομένα, την

αποθήκευση και τις γρήγορες λύσεις ανάκτησης πληροφοριών και τέλος για τον τρόπο εκδημοκρατισμού της μηχανής μάθησης και του υπολογιστικού νέφους για τους γεωπόνους με την πρόβλεψη και την εφαρμογή πλαισίων επεξεργασίας δεδομένων ανοιχτού κώδικα.



Εικόνα 3: IoT in agriculture

Η ψηφιακή γεωργία [10] είναι η ενσωμάτωση της ψηφιακής τεχνολογίας σε όλα τα στάδια της γεωργικής παραγωγής μέχρι το τελικό προϊόν να φτάσει στον καταναλωτή. Αυτές οι τεχνολογίες μπορούν να παρέχουν στη γεωργική βιομηχανία εργαλεία και πληροφορίες για τη λήψη αποφάσεων και τη βελτίωση της παραγωγικότητας. Είναι μια μέθοδος διαχείρισης των αγρών χρησιμοποιώντας τεχνολογία της πληροφορίας, δορυφορικά δεδομένα και απομακρυσμένους αισθητήρες. Οι τεχνολογίες αυτές έχουν ως στόχο την βελτιστοποίηση των αποδόσεων των εισροών ενώ ταυτόχρονα μειώνουν τις περιβαλλοντικές επιπτώσεις. Η υφιστάμενη κατάσταση της σύγχρονης επιστήμης που αφορά στην Ψηφιακή Γεωργία σε αρόσιμη γη, μόνιμες καλλιέργειες και γαλακτοκομικές παραγωγές εξετάζονται κυρίως στο ευρωπαϊκό πλαίσιο μαζί με ορισμένες πτυχές της ψηφιακής γεωργίας.[11]

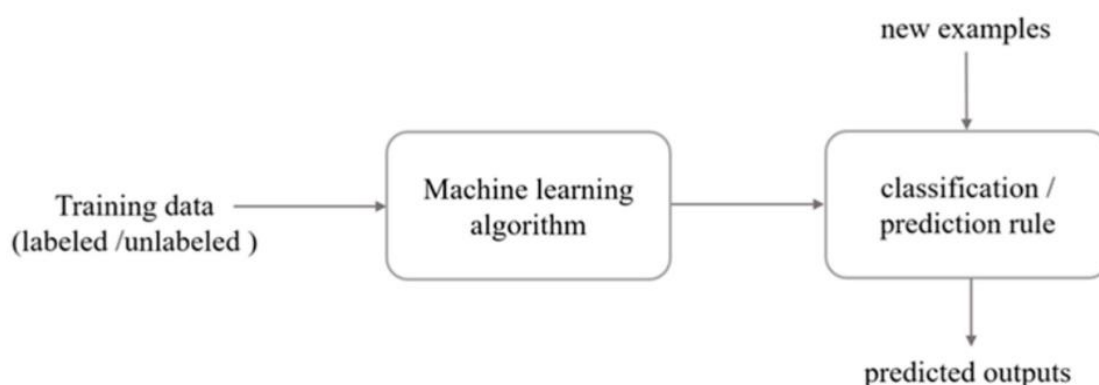
Η γεωργική τεχνολογία και η γεωργία ακριβείας, που ονομάζονται ψηφιακή γεωργία, έχουν προκύψει ως νέα επιστημονικά πεδία που χρησιμοποιούν έντονες προσεγγίσεις δεδομένων για την αύξηση της γεωργικής παραγωγικότητας ελαχιστοποιώντας παράλληλα τις περιβαλλοντικές επιπτώσεις της. Τα δεδομένα που παράγονται στις σύγχρονες γεωργικές εργασίες παρέχονται από μια ποικιλία διαφορετικών αισθητήρων που επιτρέπουν την καλύτερη κατανόηση του επιχειρησιακού περιβάλλοντος και της ίδιας της λειτουργίας, ώστε οι λήψεις αποφάσεων να είναι πιο ακριβείς και ταχύτερες.

Το μοντέλο της ψηφιακής γεωργίας είναι αποτέλεσμα των ραγδαίων εξελίξεων του IoT και του υπολογιστικού νέφους, τα οποία διαθέτουν επίγνωση του περιεχομένου

και γεγονότα σε πραγματικό χρόνο και παρουσιάζουν έρευνες σχετικά με βιομηχανίες έξυπνων αγροτικών εκμεταλλεύσεων καθώς και πολυεπιστημονικά μοντέλα.

Ρόλος της μηχανικής μάθησης (Machine Learning) είναι η δημιουργία νέων ευκαιριών που αφορούν στην αποκάλυψη, τον ποσοτικό προσδιορισμό και την κατανόηση διαδικασιών σε γεωργικά επιχειρησιακά περιβάλλοντα. Ως μηχανική μάθηση εννοείται το επιστημονικό πεδίο που δίνει στις μηχανές να μάθουν χωρίς να είναι αυστηρά προγραμματισμένες.

Συνήθως οι μεθοδολογίες του Machine Learning περιλαμβάνουν μια διαδικασία εκμάθησης με στόχο να μάθουν από τα εκπαιδευτικά δεδομένα για την εκτέλεση μιας εργασίας. Αυτά τα δεδομένα αποτελούνται από ένα σύνολο παραδειγμάτων που τις περισσότερες φορές περιγράφονται από ένα σύνολο χαρακτηριστικών ή μεταβλητών. Η απόδοση του μοντέλου σε μια συγκεκριμένη εργασία υπολογίζεται από μια μέτρηση απόδοσης που βελτιώνεται με το πέρασμα του χρόνου.



Εικόνα 4: Μια τυπική προσέγγιση του Machine Learning

ΚΕΦΑΛΑΙΟ 2

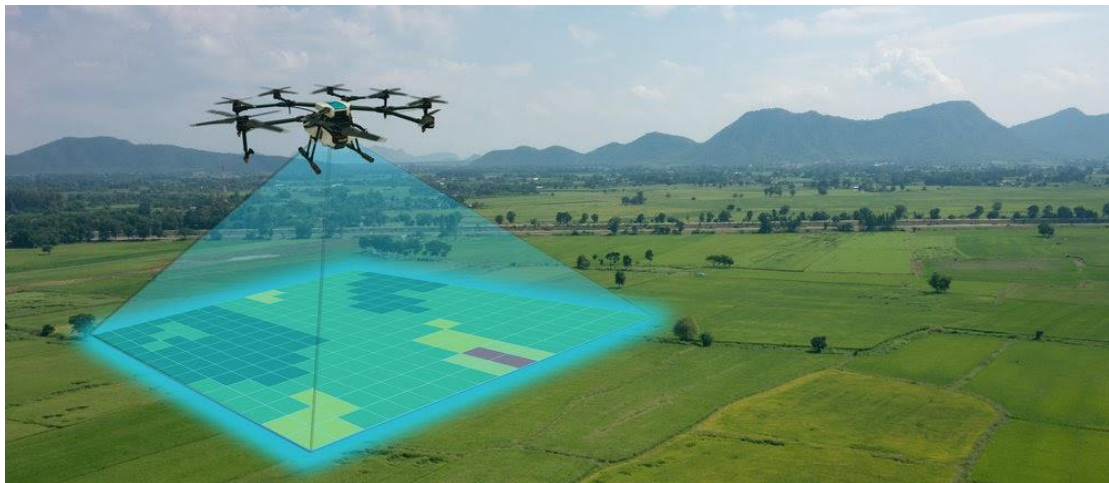
2.1. Machine learning σε agricultural δεδομένα

Η μηχανική μάθηση [12] είναι μια αναδυόμενη τεχνολογία που μπορεί να βοηθήσει στην ανακάλυψη κανόνων και μοτίβων σε ένα σύνολο δεδομένων. Σύμφωνα με παρατηρήσεις διαπιστώνεται ότι ο όγκος των καταγεγραμμένων δεδομένων αυξάνεται με εκπληκτικό ρυθμό που ξεπερνά την ικανότητα του ανθρώπου να βγάζει νόημα και η φράση “εξόρυξη βάσης δεδομένων” χρησιμοποιείται ώστε να περιγράψει τις προσπάθειες για αυτόματη ανάλυση συνόλων δεδομένων. Εφαρμογές αυτών των τεχνικών χρησιμοποιούνται και στον τομέα της γεωργίας. Όσον αφορά στο κομμάτι της μηχανικής μάθησης, το πιο σημαντικό χαρακτηριστικό ενός προβληματικού τομέα είναι η μορφή που παίρνουν τα δεδομένα. Οι περισσότερες τεχνικές που έχουν όντως εφαρμοστεί υποθέτουν ότι τα δεδομένα παρουσιάζονται σε μια απλή μορφή

χαρακτηριστικού-τιμής στην οποία μια εγγραφή έχει έναν σταθερό αριθμό πεδίων ή ιδιοτήτων σταθερής αξίας.[13]

Σκοπός αυτής της εργασίας είναι να δείξει θεωρητικά και πειραματικά αποτελέσματα, με στόχο την βελτίωση στην ανάλυση και την διαχείριση δεδομένων σε μικρές βιομηχανικές επιχειρήσεις και σε πιθανά εδαφικά πλαίσια που είναι ανθεκτικά στην καινοτομία. Πριν αναπτυχθεί η τεχνολογία, στο παρελθόν μικρές ποσότητες των καλά δομημένων δεδομένων αντιμετωπίστηκαν κερδοσκοπικά χρησιμοποιώντας λίγα προσαρμοστικά μαθηματικά μοντέλα που έρχονται από στατιστικές και αριθμητικές θεωρίες. Για τις πηγές δεδομένων υπάρχουν τρεις κύριες διαδικασίες για την συλλογή και τη δημιουργία τους.[14]

- 1) Machine Generated (MG): δεδομένα που προέρχονται από αισθητήρες και έξυπνα μηχανήματα όπως Drones, εναέρια οχήματα(UAV)⁵ [15] και GPS. Αυτά αντιπροσωπεύουν το πρότυπο IoT και την δομή τους που κυμαίνονται από απλά έως πολύπλοκα, αλλά γενικά καλοσχηματισμένες αριθμητικές εγγραφές. Αυτά τα δεδομένα αυξάνονται κρίσιμα σε όγκους και ταχύτητα και οι παραδοσιακές προσεγγίσεις σήμερα δεν επαρκούν για την θεραπεία τους.



Εικόνα 5: Ανάλυση της υγείας των καλλιεργειών με drones

- 2) Process Mediated (PM): παραδοσιακά εμπορικά δεδομένα που προέρχονται από αναφορές επιχειρηματικών διαδικασιών σε εταιρικές εκδηλώσεις όπως αγορές και παραγγελίες. Είναι αρκετά δομημένα με ποικιλία τύπων δεδομένων και συνήθως αποθηκεύονται σε σχεσιακές βάσεις δεδομένων.

⁵ Drones/Unmanned Aerial System(UAV): ιπτάμενα οχήματα που δεν έχουν χειριστή στην άτρακτο τους, αλλά πραγματοποιούν πτήσεις είτε αυτόνομα είτε μέσω τηλεκατεύθυνσης.

- 3) Με την έννοια Human Sourced (HS) εννοείται η επιβεβαίωση ανθρώπινων εμπειριών καταγεγραμμένων σε βιβλία, φωτογραφίες, ήχο και βίντεο. Οι ανθρώπινες εμπειρίες είναι πια ψηφιοποιημένες σε ψηφιακές συσκευές και κοινωνικά δίκτυα που δεν έχουν επικυρωθεί επίσημα. Η διαχείριση, ανάλυση και αποθήκευση αυτών των δεδομένων είναι προβληματική και ανοιχτή στην έρευνα [16].

Η αγροτική παραγωγή είναι μια πολύπλοκη αλυσίδα εφοδιασμού. Η μηχανική μάθηση αλλάζει τον τρόπο παραγωγής, διανομής και κατανάλωσης των αγροτικών προϊόντων. Οι ερευνητές χρησιμοποιούν τεχνολογίες που βασίζονται στην τεχνητή νοημοσύνη και την μηχανική μάθηση ώστε να παρέχουν γνώσεις και καθοδήγηση σχετικά με τον προγραμματισμό της εναλλαγής των καλλιεργειών, τους χρόνους φύτευσης, τη διαχείριση νερού και θρεπτικών συστατικών, τη διαχείριση επιβλαβών οργανισμών, τον έλεγχο των ασθενειών και τη βέλτιστη συγκομιδή.

Τα γεωργικά δεδομένα και η γεωργία ακριβείας σε ευρύτερο βαθμό για να επωφεληθούν από την μηχανική μάθηση περνάνε από κάποια στάδια τα οποία θα εξηγηθούν παρακάτω.

Species Management/Διαχείριση ειδών

Species breeding/Αναπαραγωγή ειδών: Αυτή η εφαρμογή είναι τόσο λογική αλλά τόσο απροσδόκητη γιατί προβλέπει κυρίως την συγκομιδή ή την διαχείριση των συνθηκών του περιβάλλοντος σε μεταγενέστερα στάδια. Η επιλογή ειδών είναι μια κουραστική διαδικασία αναζήτησης συγκεκριμένων γονιδίων που καθορίζουν την αποτελεσματικότητα της χρήσης νερού και θρεπτικών συστατικών, την προσαρμογή στην κλιματική αλλαγή, την αντοχή στις ασθένειες καθώς και το περιεχόμενο σε θρεπτικά συστατικά ή μια καλύτερη γεύση. Η μηχανική μάθηση, ειδικότερα οι αλγόριθμοι της βαθιάς μάθησης (Deep Learning) [17], χρειάζονται άπειρα δεδομένα πεδίου για να αναλυθεί η απόδοση των καλλιεργειών σε διάφορα κλίματα και στα νέα χαρακτηριστικά που αναπτύχθηκαν στην διαδικασία. Με βάση αυτά τα δεδομένα μπορούν να δημιουργήσουν ένα μοντέλο πιθανότητας που θα προβλέπει ποια γονίδια θα συμβάλλουν πιο πολύ σε ένα ευεργετικό χαρακτηριστικό του φυτού.

Species recognition/Αναγνώριση ειδών: Όσο η παραδοσιακή ανθρώπινη προσέγγιση για την ταξινόμηση των φυτών θα ήταν η σύγκριση του χρώματος και του σχήματος των φύλλων, τόσο η μηχανική μάθηση μπορεί να προσφέρει ακριβέστερα και γρηγορότερα αποτελέσματα αναλύοντας την μορφολογία της φλέβας των φύλλων που δίνει περισσότερες πληροφορίες για τις ιδιότητες των φύλλων.



Εικόνα 6: Species recognition

Field Conditions Management/Διαχείριση συνθηκών πεδίου

Soil management/Διαχείριση εδάφους: Για αυτούς που ασχολούνται με τη γεωργία, το έδαφος είναι ένας ετερογενής φυσικός πόρος με πολύπλοκες διαδικασίες και ασαφείς μηχανισμούς. Η θερμοκρασία του και μόνο αυτή και μπορεί να δώσει πληροφορίες για τις επιπτώσεις της κλιματικής αλλαγής στην απόδοση της περιοχής. Οι αλγόριθμοι μηχανικής μάθησης μελετούν τις διαδικασίες εξάτμισης, την υγρασία και τη θερμοκρασία τους εδάφους για να κατανοήσουν τη δυναμική των οικοσυστημάτων και την πρόσκρουση στη γεωργία.

Water management/Διαχείριση νερού: Η διαχείριση των υδάτων στην γεωργία επηρεάζει την υδρολογική, κλιματική και αγρονομική ισορροπία. Μέχρι στιγμής οι πιο ανεπτυγμένες εφαρμογές που βασίζονται σε μηχανική μάθηση συνδέονται με την εκτίμηση της ημερήσιας, εβδομαδιαίας ή μηνιαίας εξάτμισης επιτρέποντας μια πιο αποτελεσματική χρήση των συστημάτων άρδευσης και πρόβλεψη της ημερήσιας θερμοκρασίας η οποία βοηθά στον εντοπισμό των αναμενόμενων καιρικών φαινομένων και στην εκτίμηση της εξάτμισης.

Crop Management/Διαχείριση καλλιεργειών

Yield prediction/Πρόβλεψη απόδοσης: Η πρόβλεψη απόδοσης είναι ένα από τα πιο σημαντικά και δημοφιλή θέματα στη γεωργία ακριβείας, καθώς καθορίζει τη χαρτογράφηση και την εκτίμηση της απόδοσης, την αντιστοίχιση της προσφοράς των

καλλιιεργειών με τη ζήτηση και τη διαχείριση των καλλιιεργειών. Οι υπερσύγχρονες προσεγγίσεις έχουν ξεπεράσει κατά πολύ την απλή πρόβλεψη που βασίζεται στα ιστορικά δεδομένα, αλλά ενσωματώνουν τεχνολογίες μηχανικής οράσεως (computer vision)⁶ για να παρέχουν δεδομένα εν κινήσει και ολοκληρωμένη πολυδιάστατη ανάλυση των καλλιιεργειών, των καιρικών συνθηκών και των οικονομικών συνθηκών για να αξιοποιήσουν στο έπακρο την απόδοση για τους αγρότες και τον πληθυσμό γενικότερα.

Crop quality/Ποιότητα καλλιέργειας: Η ακριβής ανίχνευση και ταξινόμηση των χαρακτηριστικών ποιότητας της καλλιέργειας μπορεί να αυξήσει την τιμή του προϊόντος και να μειώσει τα απόβλητα. Σε σύγκριση με τους ειδικούς ανθρώπους, οι μηχανές μπορούν να χρησιμοποιήσουν φαινομενικά ανούσια δεδομένα και διασυνδέσεις για να αποκαλύψουν νέες ιδιότητες που παίζουν ρόλο στη συνολική ποιότητα των καλλιιεργειών και να τις εντοπίσουν.



Εικόνα 7: Crop quality

⁶ Computer Vision: μηχανική όραση, υπολογιστική όραση ή τεχνητή όραση είναι ένα επιστημονικό πεδίο της τεχνητής νοημοσύνης το οποίο επιχειρεί να αναπαράγει αλγοριθμικά της αίσθηση της όρασης συνήθως σε ηλεκτρονικό υπολογιστή ή ρομπότ.

Disease detection/Ανίχνευση ασθενειών: Τόσο σε υπαίθριες όσο και σε θερμοκηπιακές συνθήκες, η πιο ευρέως χρησιμοποιούμενη πρακτική στον έλεγχο παρασίτων και ασθενειών είναι ο ομοιόμορφος ψεκασμός φυτοφαρμάκων στην περιοχή της καλλιέργειας. Για να είναι αποτελεσματική αυτή η προσέγγιση απαιτεί σημαντικές ποσότητες φυτοφαρμάκων που οδηγούν σε υψηλό οικονομικό και σημαντικό περιβαλλοντικό κόστος. Η μηχανική μάθηση χρησιμοποιείται ως μέρος της γενικής διαχείρισης γεωργίας ακριβείας όπου οι εισροές αγροχημικών στοχεύουν ως προς τον χρόνο, τον τόπο και τα επηρεαζόμενα φυτά.



Εικόνα 8: Ανίχνευση ασθένειας

Weed detection/Ανίχνευση ζιζανίων: Εκτός από ασθένειες τα ζιζάνια είναι οι σημαντικότερες απειλές για την καλλιέργεια. Λόγω του ότι είναι δύσκολο να εντοπιστούν και να διακρίνουν τις καλλιέργειες, τα ζιζάνια αποτελούν μεγάλο πρόβλημα στην καταπολέμησή τους. Οι αλγόριθμοι της μηχανικής οράσεως και της μηχανικής μάθησης έχουν τη δυνατότητα βελτίωσης της ανίχνευσης και της διάκρισης των ζιζανίων με χαμηλό κόστος και χωρίς περιβαλλοντικά ζητήματα και παρενέργειες. Στο μέλλον αυτές οι τεχνολογίες μπορούν να οδηγήσουν τα ρομπότ στην καταστροφή των ζιζανίων ελαχιστοποιώντας την ανάγκη για ζιζανιοκτόνα.

Όσον αφορά στην παρακολούθηση των καλλιεργειών πρέπει να ειπωθεί ότι οι συμβατικές μέθοδοι παρακολούθησης της υγείας των καλλιεργειών είναι εντατικές και χρονοβόρες. Η χρήση της μηχανικής μάθησης είναι ένας αποτελεσματικός τρόπος

για την παρακολούθηση και τον εντοπισμό πιθανών θεμάτων υγείας των καλλιεργειών ή ελλείψεων σε θρεπτικά συστατικά στο έδαφος. Με αυτόν τον τρόπο αναπτύσσονται εφαρμογές για την ανάλυση των προτύπων υγείας των φυτών στη γεωργία. Τέτοιες εφαρμογές με τη δυνατότητα και της τεχνητής νοημοσύνης είναι καθοριστικές για την καλύτερη κατανόηση, την υγεία του εδάφους, τα παράσιτα των φυτών και τις ασθένειες των φυτών.

Τα αγροκτήματα παράγουν πολλά δεδομένα στο έδαφος καθημερινά. Με τη βοήθεια της μηχανικής μάθησης οι αγρότες μπορούν να αναλύσουν μια ποικιλία από δεδομένα σε πραγματικό χρόνο, όπως τις καιρικές συνθήκες, τη θερμοκρασία, την χρήση νερού και τις συνθήκες του εδάφους που συλλέγονται από το αγρόκτημά τους προκειμένου να έχουν μια καλύτερη εικόνα για να πάρουν τις αποφάσεις τους. Οι τεχνολογίες αυτές επιτρέπουν στους αγρότες να εκμεταλλευτούν τα δεδομένα που έχουν στη διάθεσή τους για να καλλιεργήσουν υγιείς καλλιέργειες ενώ χρησιμοποιούν λιγότερους φυσικούς πόρους.

Για παράδειγμα, με τη χρήση ενός συστήματος προτάσεων τεχνητής νοημοσύνης με μηχανική μάθηση για γεωργική έρευνα μπορεί να μεγιστοποιηθεί η χρήση δεδομένων που σχετίζονται με την επίλυση αγροτικών προβλημάτων και να βελτιωθεί η αποδοτικότητα του επιστημονικού δυναμικού, βελτιώνοντας παράλληλα την ακρίβεια των εκτιμήσεων του ποσού των παραγόμενων τροφίμων. Με λίγα λόγια αυτό το σύστημα παρέχει μια ισχυρή προσέγγιση για την χρήση μεγάλων ποσοτήτων δεδομένων και επιστημονικής εμπειρογνωμοσύνης σε μια γεωργική επιχείρηση για την πρόβλεψη της δυναμικής του αγροοικοσυστήματος υπό μεταβαλλόμενες περιβαλλοντικές συνθήκες.

2.2. Ανάλυση ερευνών

Η αυξανόμενη διαθεσιμότητα δεδομένων του εδάφους που μπορούν να αποκτηθούν αποτελεσματικά από απόσταση και με εγγυρότητα και οι αλγόριθμοι ανοιχτού κώδικα που είναι διαθέσιμοι έχουν οδηγήσει σε επιταχυνόμενη υιοθέτηση τεχνικών machine learning για την ανάλυση δεδομένων του εδάφους. Σύμφωνα με αυτήν την έρευνα [18] διαπιστώθηκε ότι υπάρχει αυξανόμενη χρήση μεθόδων ML κυρίως που αφορούν στο έδαφος και άλλες έρευνες που εξετάστηκαν μπορούν να ομαδοποιηθούν σε 12 θέματα, τα οποία είναι τηλεπισκόπηση, οργανικός άνθρακας του εδάφους, μόλυνση του νερού, διάβρωση, μέθοδοι με νευρωνικά δίκτυα, support vector machines, καλλιέργειες και μοντελοποίηση. Με βάση αυτά, βρέθηκαν κάποια ερευνητικά κενά σχετικά με τις προφυλάξεις που πρέπει να ληφθούν για να αποφευχθεί η υπερβολική προσαρμογή και ότι η ερμηνεία των μοντέλων ML είναι μια σημαντική πτυχή στην εφαρμογή των προηγμένων μεθόδων της, ώστε να βελτιωθεί η κατανόηση και η κατανομή του εδάφους.

Η διαδικασία μοντελοποίησης και επικύρωσης μπορεί να χρησιμοποιηθεί για την διαμόρφωση ενός μοντέλου που να εξηγεί την κατανομή του εδάφους [19]. Η μοντελοποίηση μέσω ML μπορεί να χρησιμοποιηθεί ακόμα και για τις αιτίες της

μεταβολής του εδάφους. Τα αποτελέσματα από τα μοντέλα μπορούν να ενημερώσουν για το ποιες περιβαλλοντικές μεταβλητές ελέγχουν την κατανομή του εδάφους.

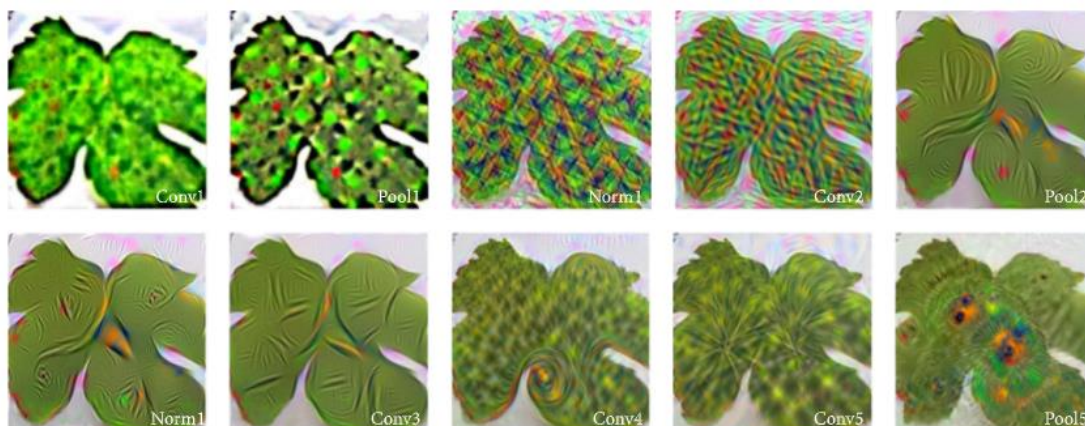
Αυτή η έρευνα χρησιμοποιεί μια συγκεκριμένη τεχνική μοντελοποίησης για να αναλύσει την τάση στην εφαρμογή της μηχανικής μάθησης στην επιστήμη του εδάφους και μέσα από αυτήν προκύπτουν ορισμένα ερωτήματα τα οποία πρέπει να απαντηθούν. Αρχικά, ποιος χρησιμοποιεί την μηχανική μάθηση και αν είναι τόσο δημοφιλής η εφαρμογή της όσο πιστεύεται. Ακόμη, ποιες μέθοδοι ML χρησιμοποιούνται συνήθως και πόσο συχνά έχουν χρησιμοποιηθεί. Σε ποιους τομείς των εδαφολογικών επιστημών χρησιμοποιείται η μηχανική μάθηση και πως ομαδοποιείται και συσχετίζεται σε αυτές. Μπορούν οι μέθοδοι ML να προσομοιώσουν τις διαδικασίες του εδάφους στον χρόνο και στον χώρο. Τέλος, μπορούν να χρησιμοποιηθούν αυτοί οι μέθοδοι για να βελτιωθούν οι γνώσεις και η κατανόηση του εδάφους.

Εδώ τα μοντέλα αναφέρονται ως απλά ή σύνθετα/προχωρημένα. Για να απεικονιστεί αυτή η σχέση μεταξύ σύνθετου και απλού οι ερευνητές θεώρησαν ότι ένα γραμμικό μοντέλο με 2 μεταβλητές είναι απλό σε σύγκριση με ένα που έχει 100. Ένα δέντρο ταξινόμησης και παλινδρόμησης με 2 κλαδιά είναι απλό σε σύγκριση με ένα δέντρο με 100 κλαδιά. Τέλος, ένα δέντρο με 2 κλαδιά είναι απλό σε σύγκριση με ένα γραμμικό μοντέλο με 100 μεταβλητές. Επίσης, ένα μπντέλο με βαθύ νευρωνικό δίκτυο έχει πολλές παραμέτρους και είναι πιο περίπλοκο με ένα δέντρο ταξινόμησης.

Για να καθοριστεί σε ποιες περιοχές των εδαφολογικών επιστημών χρησιμοποιείται η μηχανική μάθηση έγινε επιλογή ενός αλγορίθμου που χρησιμοποιείται στην μοντελοποίηση που αναφέρθηκε και ονομάζεται λανθανουσα κατανομή[20].

Από την μεγάλη ποικιλία των μοντέλων μηχανικής μάθησης που είναι διαθέσιμα βρέθηκαν πάνω από 100 διαφορετικές παραλλαγές που έχουν εφαρμοστεί στις επιστήμες του εδάφους. Υπάρχει μεγάλη αύξηση στη χρήση όλων των μοντέλων αλλά σε κάποια από αυτά έχει παρατηρηθεί μείωση, για παράδειγμα στα support vector machines, στην καμπύλη παλινδρόμησης και στα δέντρα ταξινόμησης, με αποτέλεσμα να βρίσκονται οι λύσεις στον random forest. Ο συγκεκριμένος αλγόριθμος έχει αναπτυχθεί αρκετά και έχει χρησιμοποιηθεί σε πολλά θέματα. Τέλος, η εμφάνιση του deep learning είναι πλέον αρκετά αισθητή.

Η επεξεργασία εικόνας και η υπολογιστική μηχανική έχουν αυξηθεί αρκετά στη γεωργία λόγω του μειωμένου εξοπλισμού της αυξημένης υπολογιστικής ισχύος και του υψηλού ενδιαφέροντος για να μειωθούν οι καταστροφικές μεθόδους εκτίμησης τροφίμων [21]. Η χρήση αυτών των τεχνικών παρουσιάζουν αρκετά πλεονεκτήματα συγκριτικά με τις παραδοσιακές μεθόδους που βασίζονται στη χειρωνακτική εργασία. Η χρήση της μηχανικής όρασης έχει αυξηθεί τα τελευταία χρόνια για να καλύψει την μεγάλη ζήτηση για γρήγορες και ακριβείς μεθόδους παρακολούθησης της παραγωγής των σιτιρών.



Εικόνα 9: Output layers images [22]

Όσον αφορά στις χειρωνακτικές μεθόδους πρέπει να ειπωθεί ότι υπάρχουν αρκετές δυσκολίες τόσο στην εκπαίδευση των ατομών που ασχολούνται με αυτές όσο και στον χρόνο. Ο χρόνος που απαιτείται για την υλοποίηση τους εμποδίζει την γρήγορη λήψη αποφάσεων και την αξιολόγηση μεγάλης κλίμακας. Με την χρήση της μηχανικής όρασης που χρησιμοποιεί υπέρυθρη ακτινοβολία, μαγνητική φασματοσκοπία⁷ [23] , ακτίνες X και υπερφασματικές εικόνες μπορεί να ξεπεραστούν αυτοί οι περιορισμοί των χειρωνακτικών μεθόδων.

Η συγκεκριμένη έρευνα έχει ως στόχο να βρει τις έρευνες που σχετίζονται με την χρήση της μηχανικής όρασης και της τεχνητής νοημοσύνης για τις πέντε μεγαλύτερες καλλιέργειες σιτιρών λαμβάνοντας υπόψη τους συνολικούς τόνους που συγκομίστηκαν για αυτές τις καλλιέργειες. Για την ανάλυση της βιβλιογραφίας χρησιμοποιήθηκαν πέντε βάσεις δεδομένων. Το χρονικό διάστημα που χρησιμοποιήθηκε για την έρευνα ήταν από το 2013 μέχρι το 2018.

Οι μέθοδοι που αναλύθηκαν χρησιμοποιούν το ίδιο μοτίβο δηλαδή απόκτηση εικόνας, επεξεργασία και ταξινόμηση. Ωστόσο κάθε κατάσταση απαιτεί την επιλογή των κατάλληλων αλγορίθμων υπολογίζοντας τα μεμονωμένα χαρακτηριστικά του προβλήματος που θα αντιμετωπιστεί. Ο κατακερματισμός των εικόνων παίζει σημαντικό ρόλο στο να επιτυγχάνονται τα επιθυμητά αποτελέσματα. Σε αυτήν την περίπτωση οι εικόνες λαμβάνονται κατευθείαν από την καλλιέργεια.

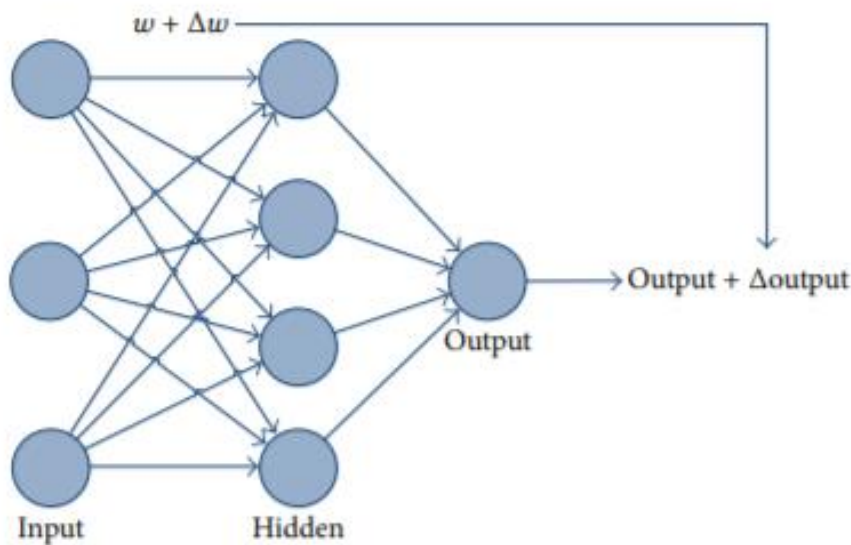
Οι αλγόριθμοι μηχανικής μάθησης είναι ευαίσθητοι στον όγκο και την ποιότητα των πληροφοριών που χρησιμοποιούνται στην εκπαίδευση τους. Κάποιες φορές όταν ο αλγόριθμος μαθαίνει για τον θόρυβο και τις υπερβολικές λεπτομέρειες του συνόλου εισόδου τότε είναι εύκολο να επηρεαστεί αρνητικά από την γενικευμένη χωρητικότητα του ενσωματωμένου συνόλου. Ωστόσο όταν τα δεδομένα δεν

⁷ Φασματοσκοπία είναι η μελέτη του ηλεκτρομαγνητικού φάσματος μιας πηγής ηλεκτρομαγνητικής ακτινοβολίας σε ένα εύρος συχνοτήτων.

επαρκούν, το μοντέλο μπορεί να μην είναι σε θέση να γενικευεί όταν χρησιμοποιείται με επικυρωμένα δεδομένα. Οι έρευνες που αξιολογήθηκαν έδειξαν διαφορετικές προσεγγίσεις για να αντιμετωπιστεί αυτή η κατάσταση και έγινε επιλογή σωστών χαρακτηριστικών προκειμένου να υπάρξει καλύτερη απόδοση. Ακόμη, για να υπάρξει εγγύηση είναι απαραίτητο να μειωθεί η διάσταση της εισόδου των διανυσμάτων για τους αλγορίθμους ταξινόμησης, ώστε να βελτιστοποιηθεί η επεξεργασία των αλγορίθμων εκμάθησης. Ο ταξινομητής support vector machines (SVM) που χρησιμοποιείται στο 56% των ερευνών που αναλύονται μπορεί να θεωρηθεί ως κυρίαρχη τεχνική έχοντας καλά αποτελέσματα ακρίβειας και απόδοσης. Παρόλα αυτά επειδή είναι ικανός να ταξινομήσει δεδομένα σε δύο ομάδες, οι προσεγγίσεις που μελετήθηκαν καταλήγουν να χρησιμοποιούν μια ιεραρχική δομή για να αυξήσουν το μεγαλύτερο μέρος των κλάσεων.

Το πρόβλημα της αποτελεσματικής προστασίας των φυτικών ασθενειών σχετίζεται συχνά με τα προβλήματα της γεωργίας και της κλιματικής αλλαγής. Τα αποτελέσματα της έρευνας δείχνουν ότι η κλιματική αλλαγή μπορεί να αλλοιώσει τα στάδια και τους ρυθμούς ανάπτυξης των παθογόνων μικροοργανισμών, μπορεί επίσης να τροποποιήσει την αντίσταση του ξενιστή, η οποία οδηγεί πολλές φορές σε αλλαγές των αλληλεπιδράσεων ξενιστή και παθογόνου.

Οι εξελίξεις στην μηχανική όραση παρέχουν μια ευκαιρία να επεκτείνουν και να ενισχύσουν την πρακτική της ακριβέστερης προστασίας των φυτών και να επεκτείνει την αγορά εφαρμογών μηχανικής όρασης στην γεωργία ακριβείας. Αξιοποιώντας κοινές τεχνικές ψηφιακής επεξεργασίας εικόνας όπως για παράδειγμα η ανάλυση χρώματος χρησιμοποιήθηκαν, προκειμένου να ανιχνευθούν και να ταξινομηθούν οι ασθένειες των φυτών [24]. Διάφορες τεχνικές προσέγγισης χρησιμοποιούνται και οι πιο συχνές είναι τα τεχνητά νευρωνικά δίκτυα (ANNs) και τα SVM. Συνδυάζονται με διαφορετικές μεθόδους επεξεργασίας εικόνας για την καλύτερη εξαγωγή χαρακτηριστικών.

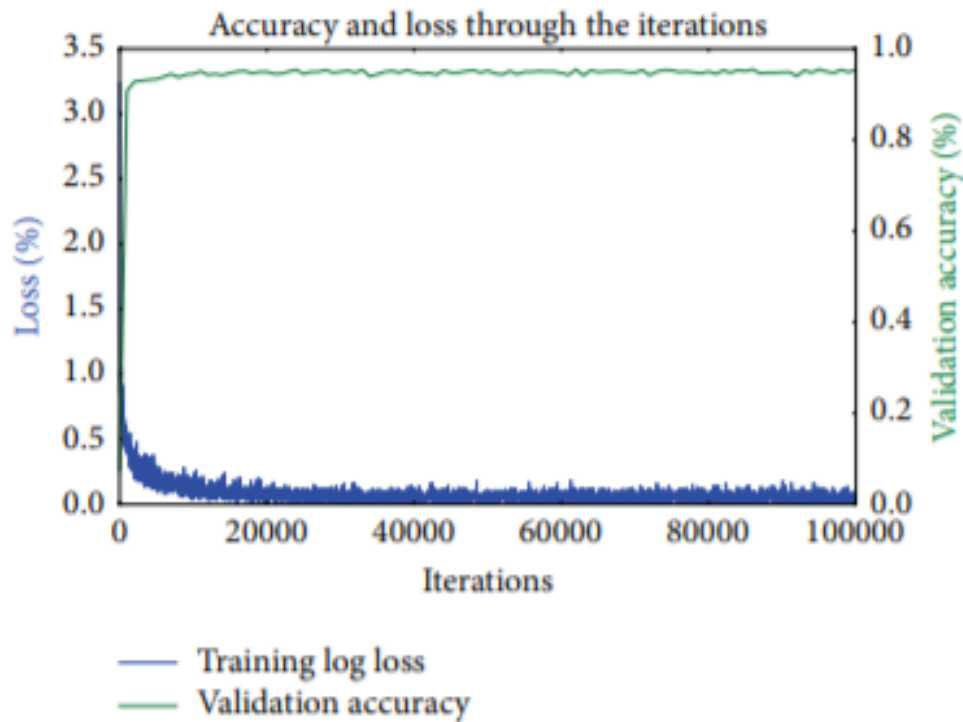


Εικόνα 10: Ένα απλό μοντέλο ANN

Η μέθοδος που περιγράφεται σε αυτή την έρευνα είναι μια καινούργια προσέγγιση στην ανίχνευση ασθενειών των φυτών χρησιμοποιώντας νευρωνικά δίκτυα ώστε να ταιριάζει με ακρίβεια στην βάση δεδομένων των φύλλων ενός φυτού. Τα υγιή φύλλα και οι εικόνες είναι σύμφωνες με άλλες κλάσεις επιτρέποντας το μοντέλο να διακρίνει τα φύλλα που έχουν ασθένεια με αυτά που δεν έχουν χρησιμοποιώντας ANN.

Τα αποτελέσματα που παρουσιάζονται εδώ είναι σχετικά με την εκπαίδευση με ολόκληρη τη βάση δεδομένων που περιέχει τόσο πρωτότυπες όσο και επαυξημένες εικόνες. Αυτά τα δίκτυα μπορούν να μάθουν χαρακτηριστικά όταν εκπαιδεύονται σε μεγαλύτερα σύνολα δεδομένων και τα αποτελέσματα επιτυγχάνονται όταν εκπαιδεύονται μόνο με πρωτότυπες εικόνες που δεν διερευνώνται.

Μετά τη ρύθμιση των παραμέτρων του δικτύου με πολλές λεπτομέρειες, επιτεύχθηκε συνολική ακρίβεια 96,3%, μετά την 100η επανάληψη εκπαίδευσης. Ακόμα και μετά τις 30 επαναλήψεις εκπαίδευσης επιτεύχθηκαν αποτελέσματα υψηλής ακρίβειας με εξαιρετικά μειωμένη απώλεια, αλλά μετά την 60η επανάληψη πάλι η ισορροπία στην ακρίβεια και την απώλεια πραγματοποιήθηκε με μεγάλη ακρίβεια. Η πράσινη γραμμή στην παρακάτω εικόνα δείχνει την επιτυχία του δικτύου στο σετ δοκιμών επικύρωσης, μέσω επαναλήψεων εκπαίδευσης. Μετά από κάθε 10 χιλιάδες επαναλήψεις εκπαίδευσης, το στιγμιότυπο του μοντέλου λήφθηκε. Η μπλε γραμμή στην εικόνα αντιπροσωπεύει την απώλεια κατά το στάδιο της εκπαίδευσης. Μέσα από τις επαναλήψεις εκπαίδευσης, η απώλεια μειώθηκε γρήγορα.



Εικόνα 11: Ακρίβεια των λεπτομεριών του CNN

2.3. Συμπεράσματα

Στο κεφάλαιο 2 γίνεται μια σύντομη περιγραφή διάφορων ερευνών που έχουν πραγματοποιηθεί στον κλάδο της γεωργίας ακριβείας κυρίως με τεχνικές machine learning, artificial intelligence και deep learning. Τα δεδομένα που χρησιμοποιήθηκαν αφορούν στο κομμάτι του εδάφους και για το πως γίνεται διάγνωση μιας ασθένειας των φυτών. Για την ανάλυση, τη διάγνωση και την επεξεργασία αυτών χρησιμοποιούνται πολλές αρχιτεκτονικές μοντέλων, κάποιες πιο παλιές και σίγουρες και κάποιες καινούργιες των οποίων τα αποτελέσματα φαίνεται να είναι πολύ θετικά ειδικά όσο αναπτύσσεται αυτό το κομμάτι της επιστήμης.

Οι αλγόριθμοι που χρησιμοποιούνται κυρίως είναι τα δέντρα αποφάσεων, τα νευρωνικά δίκτυα και οι support vector machines. Οι καινούργιες τεχνολογίες είναι η ανάλυση εικόνων και η μηχανική όραση [25]. Ακόμη είναι στο στάδιο της ανάπτυξης αλλά τα αποτελέσματα δείχνουν να είναι πολύ καλύτερα και με υψηλή ακρίβεια.

ΚΕΦΑΛΑΙΟ 3

3.1. Μεθοδολογία

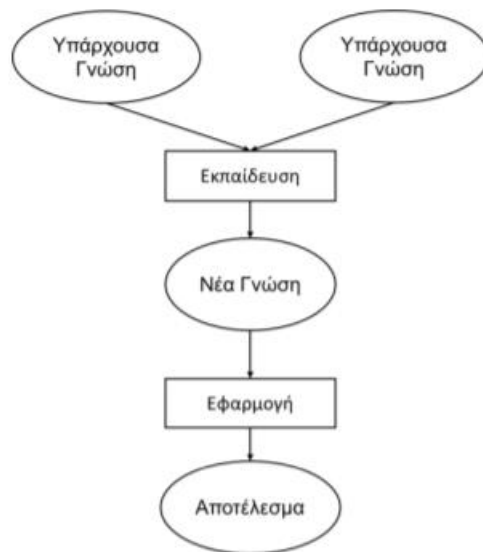
Στο προηγούμενο κεφάλαιο έγινε αναφορά και ανάλυση ερευνών εξηγώντας πως τα μοντέλα μηχανικής μάθησης λειτουργούν για την προστασία του εδάφους από διάφορους παράγοντες και την ανίχνευση ασθενειών.

Στην συγκεκριμένη εργασία γίνεται ανάλυση των συνόλων δεδομένων που επιλέχθηκαν και αφορούν κυρίως στο πως κάποια φυτά μπορούν να καλλιεργηθούν και να αντέξουν σε διαφορετικά εδάφη. Ακόμη, φαίνεται με πρακτικά και πειραματικά αποτελέσματα πως μπορεί να βελτιωθεί η διαχείριση και η ανάλυση δεδομένων σε μικρές βιομηχανικές επιχειρήσεις και σε ενδεχόμενα εδαφικά πλαίσια, συχνά ανθεκτικά στην καινοτομία. Σκοπός της είναι η ανάπτυξη αλγορίθμων και τεχνικών για την εκπαίδευση μοντέλων υπολογιστικής νοημοσύνης.

Ένα από τα βασικά εργαλεία μηχανικής μάθησης είναι τα δέντρα απόφασης. Η εκμάθηση με δέντρο απόφασης χρησιμοποιεί ένα δέντρο απόφασης ως προγνωστικό μοντέλο, το οποίο αντιστοιχίζει τις παρατηρήσεις με ένα στοιχείο συμπεράσμα σχετικά με την τιμή στόχο του αντικειμένου στο δέντρο απόφασης και έχουν ως βασική ιδιότητα την ικανότητα «εκπαίδευσής» τους μέσω της διαδικασίας της εποπτευόμενης μάθησης (supervised learning). Ακόμη, ένα δίκτυο Bayes ή ένα δίκτυο εμπιστοσύνης είναι ένα πιθανό - θεωρητικό γραφικό μοντέλο που απεικονίζει ένα σύνολο τυχαίων μεταβλητών και την μεταξύ τους υποθετική ανεξαρτησία. Για παράδειγμα, ένα δίκτυο Bayes μπορεί να αναπαραστήσει την πιθανό - θεωρητική σχέση μεταξύ ασθενειών και συμπτωμάτων ενός φυτού. Το συγκεκριμένο δίκτυο μπορεί να χρησιμοποιηθεί για να υπολογίσει τις πιθανότητες διαφόρων ασθενειών.

Στη συγκεκριμένη εργασία η μέθοδος που χρησιμοποιείται είναι κυρίως η μέθοδος των επαγωγικών δένδρων απόφασης. Οι λόγοι που έγινε η συγκεκριμένη επιλογή είναι γιατί τα χρησιμοποιούμενα δεδομένα, τα οποία παρουσιάζονται αναλυτικά σε επόμενο κεφάλαιο, παρουσιάζουν ανομοιομορφία ως προς τον τύπο τους (Αριθμητικά/Ονομαστικά). Έχει αποδειχθεί από παρόμοιες προσεγγίσεις διαφόρων προβλημάτων του πραγματικού κόσμου πως η συγκεκριμένη μέθοδος είναι αποδοτικότερη κατά την προσαρμογή της σε παρόμοια δεδομένα σε σχέση με άλλες νοήμονες μεθόδους όπως τα νευρωνικά δίκτυα. Θεωρείται η πιο κατανοητή μέθοδος σε σχέση με άλλες μεθόδους.

Τέλος, στην εργασία η διαδικασία εκπαίδευσης γίνεται με επιβλεπόμενη μάθηση ή εποπτευόμενη μάθηση. Η συγκεκριμένη προσέγγιση, μαθαίνει από παραδείγματα. Πιο συγκεκριμένα το υπολογιστικό μοντέλο εφαρμόζει επιτυχώς κάθε καταχώρηση στο σύνολο εκπαίδευσης εφόσον του έχει δοθεί το σύνολο δεδομένων εκπαίδευσης σε συνδυασμό με τις σωστές απαντήσεις.



Εικόνα 12: Γενικός τρόπος λειτουργίας των αλγορίθμων μάθησης

3.2. Ανάλυση μεθοδολογίας

Ο στόχος της εργασίας είναι η ανάλυση των γεωργικών δεδομένων χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Οι αλγόριθμοι εξόρυξης δεδομένων λειτουργούν με διαφορετικές αρχές, έχοντας τη δυνατότητα να επηρεάζονται από διαφορετικά είδη συσχετίσεων στα δεδομένα. Για την εξασφάλιση δικαιότερων συνθηκών στην αξιολόγηση, αυτή η εργασία βρίσκει τη βέλτιστη μέθοδο ομαδοποίησης για την ανάλυση δεδομένων γεωργίας. Η προτεινόμενη εργασία υιοθετεί τις εξωτερικές μετρήσεις ποιότητας όπως Precision, Recall and F μέτρα για σύγκριση των μεθόδων ομαδοποίησης.

Η ακρίβεια υπολογίζεται ως το κλάσμα αυτών που τοποθετούνται σωστά στο ίδιο σύμπλεγμα. Η ανάκληση αντιπροσωπεύει το κλάσμα των πραγματικών ζευγαριών που αναγνωρίστηκαν. Το μέτρο F υποδεικνύει τον αρμονικό μέσο ακριβείας και ανάκλησης. Η τιμή μετρήσεων υψηλότερης ποιότητας αντιπροσωπεύει την καλύτερη ποιότητα συμπλέγματος.

Τα επαγωγικά δένδρα [26] είναι δένδρο-διαμορφωμένες δομές που αντιπροσωπεύουν τα σύνολα αποφάσεων. Αυτές οι αποφάσεις παράγουν τους κανόνες για την ταξινόμηση ενός συνόλου δεδομένων. Πρόκειται για μια απλή δομή όπου οι μη τερματικοί κόμβοι αντιπροσωπεύουν τα αποτελέσματα των αποφάσεων. Τα δένδρα αποφάσεων έχουν διάφορα πλεονεκτήματα, όπως το ότι είναι ευκολότερα στην κατανόηση, μπορούν να μετασχηματιστούν σε κανόνες και πειραματικά έχει αποδειχθεί ότι λειτουργούν πολύ καλά. Τα δένδρα απόφασης κατασκευάζονται χρησιμοποιώντας μόνο εκείνα τα γνωρίσματα που είναι σε θέση να διακρίνουν τις έννοιες προς εκμάθηση. Για να χτιστεί ένα δέντρο απόφασης, πρέπει αρχικά να γίνει

επιλογή ενός υποσυνόλου περιπτώσεων από το σύνολο των δεδομένων που θα χρησιμοποιηθούν στην εκπαίδευση (υποσύνολο δεδομένων εκπαίδευσης - training set). Αυτό το υποσύνολο (δεδομένα ελέγχου - test set) χρησιμοποιείται έπειτα από τον αλγόριθμο για να κατασκευάσει το δέντρο απόφασης. Τα υπόλοιπα δεδομένα, τα δεδομένα εκπαίδευσης (training set), χρησιμοποιούνται στην εξέταση της ακρίβειας του κατασκευασμένου δέντρου.

3.2.1. Διαδικασία και μέτρα αξιολόγησης

Η διαδικασία αξιολόγησης που ακολουθήθηκε για την εκτίμηση των επιδόσεων των παραγόμενων μοντέλων είναι η Διασταυρούμενη Επικύρωση (Cross Validation). Η τεχνική διασταύρωσης και αξιολόγησης των αποτελεσμάτων, γνωστή ως Cross Validation είναι μια μέθοδος που χρησιμοποιεί επαναλαμβανόμενες δοκιμές σε άγνωστα δεδομένα. Συγκεκριμένα πραγματοποιεί διαίρεση των δεδομένων σε K τυχαία υποσύνολα του ίδιου μεγέθους. Το ένα υποσύνολο χρησιμοποιείται ως δεδομένα δοκιμής και τα υπόλοιπα ως δεδομένα εκπαίδευσης. Στα πειράματα που έγιναν το K πήρε την τιμή $K=10$. Τέλος, τα μέτρα αξιολόγησης των προαναφερόμενων μοντέλων ήταν η ακρίβεια ταξινόμησης και το μέγεθος του δένδρου του κάθε μοντέλου

3.2.2. Το περιβάλλον WEKA

Το περιβάλλον, όπου αναπτύχθηκαν τα μοντέλα της συγκεκριμένης εργασίας, ήταν το WEKA. Το WEKA αναπτύχθηκε στο πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και αποτελεί μια βιβλιοθήκη μεθόδων ανάλυσης δεδομένων. Στη βιβλιοθήκη του WEKA περιλαμβάνεται τόσο συμβατικές μεθόδους ανάλυσης δεδομένων όσο και νοήμονες μεθόδους όπως τα επαγωγικά δένδρα απόφασης. Θεωρείται ένα από τα πιο εύχρηστα και λειτουργικά λογισμικά καθώς έχει φιλικό προς το χρήστη περιβάλλον εργασίας και δεν απαιτεί υψηλές γνώσεις προγραμματισμού. Η γλώσσα προγραμματισμού, με την οποία έχει σχεδιαστεί και αναπτυχθεί το WEKA είναι η Java και τα αρχεία δεδομένων που αναγνωρίζει είναι της μορφής .arff.

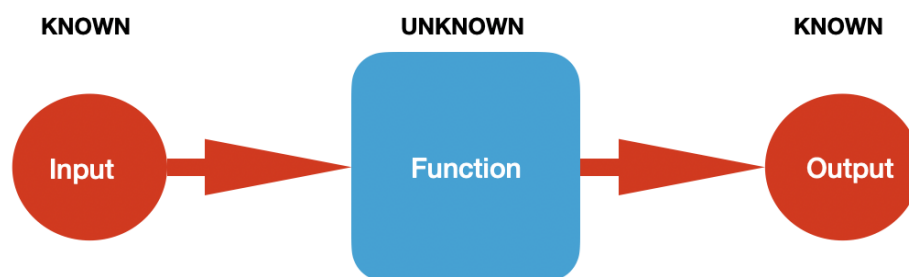
ΚΕΦΑΛΑΙΟ 4

4.1. Περιγραφή αλγορίθμων

Σήμερα λόγω της διαθεσιμότητας καινοτόμων αλγορίθμων και των μεγάλων συνόλων δεδομένων μέσω βιομηχανικών πόρων στο διαδίκτυο και οι ερευνητικές κοινότητες χρησιμοποιούν ευρέως τους αλγορίθμους μηχανικής μάθησης για την επίλυση ποικίλων συνόλων προβλημάτων. Αυτοί οι συγκεκριμένοι αλγόριθμοι κατηγοριοποιούνται ως επιβλεπόμενη μάθηση (supervised learning), μη επιβλεπόμενη μάθηση (unsupervised learning) και υποβοηθούμενη μάθηση.

Η επιβλεπόμενη μάθηση όπως και αποκαλείται είναι η μάθηση με κάποιον επόπτη ή δάσκαλο. Σε αυτό το σύνολο αλγορίθμων στην κάθε είσοδο του σετ δεδομένων αντιστοιχεί και μια έξοδος. Ο αλγόριθμος δημιουργεί μια σχέση εισόδου-εξόδου με αυτό το σύνολο δεδομένων και στη συνέχεια γίνεται γενίκευση ή πρόβλεψη των εξόδων για αόρατες εισόδους. Συνήθως οι αλγόριθμοι που χρησιμοποιούνται για την πρόβλεψη της κατηγορικής τιμής είναι οι αλγόριθμοι ταξινόμησης και οι αλγόριθμοι που χρησιμοποιούνται για την πρόβλεψη της αριθμητικής τιμής είναι γνωστοί και ως αλγόριθμοι παλινδρόμησης.

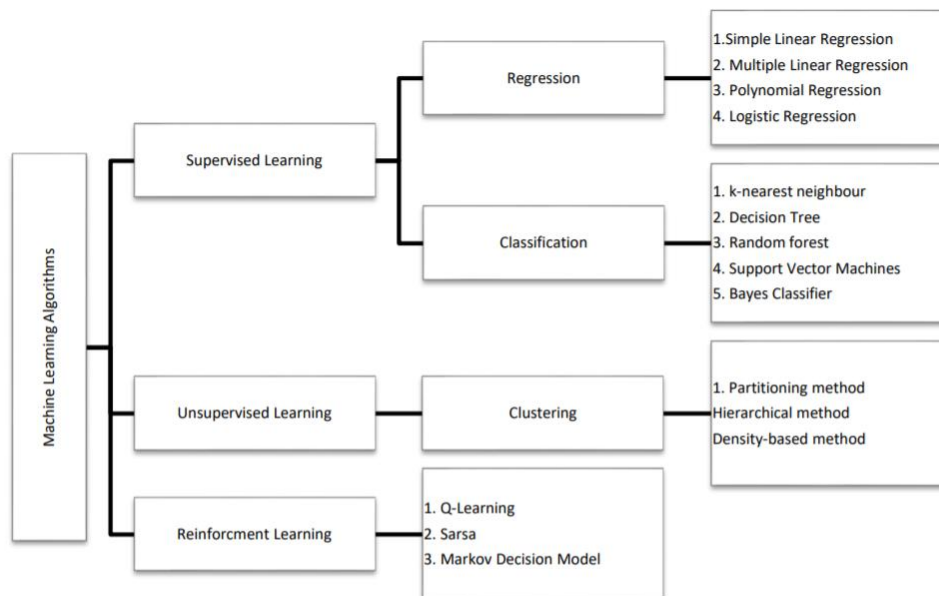
Οι αλγόριθμοι της μη επιβλεπόμενης μάθησης λειτουργούν με δεδομένα χωρίς ετικέτα και ανακαλύπτουν άγνωστα αντικείμενα ομαδοποιώντας παρόμοια αντικείμενα. Στόχος της είναι οι αλγόριθμοι να εξάγουν κρυφές γνώσεις από το σύνολο δεδομένων που εκπαιδεύεται και για αυτόν τον λόγο αυτή η προσέγγιση είναι δύσκολο να υλοποιηθεί από επιβλεπόμενους αλγορίθμους μάθησης.



Εικόνα 13: Supervised Learning

Η υποβοηθούμενη μάθηση είναι μια άλλη προσέγγιση που μαθαίνει από το περιβάλλον μέσω της ανταμοιβής και της τιμωρίας.

Σε αυτή την εργασία οι αλγόριθμοι που χρησιμοποιούνται και αναλύονται είναι αυτοί της επιβλεπόμενης μάθησης με στόχο την πρόβλεψη του εδάφους από τα σύνολα δεδομένα που χρησιμοποιούνται. Παρακάτω γίνεται περιγραφή των αλγορίθμων και αυτοί είναι οι C4.5, Random Forest, kNN και Naive Bayes.



Εικόνα 14: Κατηγοριοποίηση αλγορίθμων μηχανικής μάθησης

4.1.1. Δέντρα απόφασης (C4.5)

Ένα Δέντρο Απόφασης (Decision Tree) είναι ένας γράφος με την κλασική δενδρική δομή. Δηλαδή, μεταξύ των κόμβων διακρίνεται α) ένας ειδικός κόμβος, η ρίζα, β) οι εσωτερικοί κόμβοι και γ) οι εξωτερικοί κόμβοι, τα φύλλα. Η προσέγγιση του δέντρου απόφασης είναι πολύ χρήσιμη στα προβλήματα κατηγοριοποίησης. Με αυτή την τεχνική κατασκευάζεται ένα δέντρο για να μοντελοποιήσει την διαδικασία της κατηγοριοποίησης. Μόλις χτιστεί ένα δέντρο, εφαρμόζεται σε κάθε μία πλειάδα στη βάση δεδομένων και καταλήγει σε μια κατηγοριοποίηση για κάθε πλειάδα. Υπάρχουν δύο βασικά βήματα στην τεχνική αυτή. Το πρώτο είναι η κατασκευή του δέντρου και το δεύτερο είναι εφαρμογή στη βάση δεδομένων. Δημιουργεί δέντρα απόφασης από ένα σύνολο δεδομένων εκπαίδευσης χρησιμοποιώντας την έννοια της εντροπίας πληροφοριών. Χρησιμοποιεί το γεγονός ότι κάθε χαρακτηριστικό των δεδομένων μπορεί να χρησιμοποιηθεί για να λάβει μια απόφαση, η οποία χωρίζει τα δεδομένα σε μικρότερα υποσύνολα. Εξετάζει το ομαλοποιημένο κέρδος πληροφοριών (information gain – διαφορά στην εντροπία) που προκύπτει από την επιλογή ενός χαρακτηριστικού για το διαχωρισμό των δεδομένων. Το χαρακτηριστικό με το υψηλότερο ομαλοποιημένο κέρδος είναι αυτό που χρησιμοποιείται για να ληφθεί μια απόφαση.

Υπάρχουν αρκετά πλεονεκτήματα από τη χρήση των δέντρων απόφασης για την κατηγοριοποίηση. Τα δέντρα απόφασης είναι εύκολα στην χρήση και αποτελεσματικά. Μπορούν να δημιουργηθούν οι κανόνες οι οποίοι κατανοούνται και

ερμηνεύονται εύκολα. Ακόμη, αποδίδουν καλά για μεγάλες βάσεις δεδομένων επειδή το μέγεθος του δέντρου είναι ανεξάρτητο από το μέγεθος της βάσης δεδομένων.

4.1.2. Random Forest

Ο Random Forest είναι το μοντέλο ταξινόμησης συνόλου που συνδυάζει μια σειρά ταξινομητών των δέντρων αποφάσεων. Η τελική κλάση ενός νέου αντικειμένου διαπιστώνεται με βάση την πλειοψηφία που προβλέπεται από διαφορετικούς ταξινομητές των δέντρων αποφάσεων.

4.1.3. kNN

Ο kNN είναι ένας απλός επιβλεπόμενος αλγόριθμος ταξινόμησης. Η συγκεκριμένη τεχνική προϋποθέτει ότι το σύνολο εκπαίδευσης δεν περιλαμβάνει μόνο τα δεδομένα αλλά επίσης και την επιθυμητή κατηγοριοποίηση για κάθε στοιχείο. Αυτό έχει ως αποτέλεσμα τα δεδομένα εκπαίδευσης να αποτελούν το μοντέλο κατηγοριοποίησης. Όταν πρόκειται να γίνει μια κατηγοριοποίηση για ένα καινούργιο στοιχείο πρέπει να καθοριστεί η απόστασή του από κάθε στοιχείο του συνόλου εκπαίδευσης και μόνο οι K πλησιέστερες εκχωρήσεις στο σύνολο δεδομένων λαμβάνονται υπόψη. Το νέο στοιχείο τοποθετείται στην κατηγορία που περιέχει τα περισσότερα στοιχεία από το σύνολο των κοντινότερων στοιχείων.

Υπάρχουν ορισμένα βήματα που πρέπει να πραγματοποιηθούν[27]:

- Λαμβάνει μια ομάδα διανυσμάτων χαρακτηριστικών ως σύνολο δεδομένων που πρέπει να ομαδοποιούνται.
- Επιλέγει τυχαία το ποσοστό κατά k για να είναι το κέντρο του συμπλέγματος.
- Υπολογίζει τον μέσο όρο κάθε σημείου δεδομένων στα συμπλέγματα για να μάθει το καινούργιο και στη συνέχεια κάθε σημείο εκχωρείται στο πλησιέστερο κέντρο συμπλέγματος.
- Επαναλαμβάνει τα δύο προηγούμενα βήματα μέχρι να πληρείται κάποιο κριτήριο σύγκλισης ή η εκχώρηση να μην μπορεί να αλλάξει.

4.1.4. Naive Bayes

Ο Naive Bayes ταξινομητής προσφέρει μια απλή πιθανοθεωρητική προσέγγιση στα προβλήματα μάθησης με επίβλεψη, όπου στόχος είναι η πρόβλεψη επακριβώς στην κατηγορία-κλάση των στιγμιότυπων δοκιμής χρησιμοποιώντας ταξινομημένα στιγμιότυπα εκπαίδευσης που περιλαμβάνουν την πληροφορία της κλάσης που ανήκουν. Βασίζεται σε δύο σημαντικές υποθέσεις. Συγκεκριμένα, υποθέτει ότι κάθε χαρακτηριστικό των στιγμιότυπων είναι στοχαστικά ανεξάρτητα των υπολοίπων, δεδομένης της κλάσης και ότι δεν υπάρχουν άλλα κρυφά χαρακτηριστικά που να επηρεάζουν την διαδικασία της πρόβλεψης.

4.2. Μέτρα αξιολόγησης

4.2.1. Πίνακας σύγχυσης (Confusion Matrix)

	Positive Class	Negative Class
Classified as Positive	TP	FP
Classified as Negative	FN	TN

Πίνακας 1: Πίνακας σύγχυσης ενός προβλήματος ταξινόμησης δύο κλάσεων

Στον παραπάνω πίνακα [28] (Πίνακας 1) εμφανίζεται ο πίνακας σύγχυσης (confusion matrix) που αναπαριστά τις τέσσερις δυνατές εκβάσεις μιας προσπάθειας δυαδικής ταξινόμησης. Στην κύρια διαγώνιο του πίνακα οι συμβολισμοί TP (True Positive) και TN (True Negative) αντιστοιχούν στις περιπτώσεις ορθής ταξινόμησης ενός θετικού παραδείγματος στην κλάση των θετικών και ενός αρνητικού στην κλάση των αρνητικών. Οι άλλες δύο περιπτώσεις αναφέρονται σε εσφαλμένες ταξινομήσεις που συμμετέχουν στον υπολογισμό του συνολικού σφάλματος του ταξινομητή. Η περίπτωση FN (False Negative) που αντιστοιχεί στην εσφαλμένη ταξινόμηση ενός θετικού παραδείγματος ως αρνητικό, αναφέρεται στη βιβλιογραφία ως Σφάλμα Τύπου I (Error Type I). Η αντίθετη περίπτωση FP (False Positive) αποτελεί το Σφάλμα Τύπου II (Error Type II). Ας σημειωθεί επίσης ότι το άθροισμα TP + FN ισούται με το σύνολο των θετικών παραδειγμάτων που χρησιμοποιήθηκαν κατά τον έλεγχο του ταξινομητή και κατ' επέκταση το FP + TN με το πλήθος των αρνητικών που ή ταξινομήθηκαν σωστά ή εσφαλμένα. [29]

4.2.2. Ακρίβεια (Accuracy)

Ως ακρίβεια [30] (accuracy) ορίζεται η αναλογία όλων των προβλέψεων που ήταν σωστές:

$$accuracy = \frac{\text{αριθμός σωστών προβλέψεων}}{\text{αριθμός προβλέψεων}}$$

Για ένα πρόβλημα ταξινόμησης δύο κλάσεων, η ποσότητα εκφράζεται με την βοήθεια του πλήθους των σωστών και εσφαλμένων ταξινομήσεων ενός συστήματος ως εξής:

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

4.2.3. Rating

Στη μηχανική μάθηση, το πραγματικό θετικό ποσοστό, που αναφέρεται επίσης στην ευαισθησία ή την ανάκληση, χρησιμοποιείται για τη μέτρηση του ποσοστού των πραγματικών θετικών που έχουν προσδιοριστεί σωστά. [31]

4.2.4. Ορθότητα (Precision)

Ως ορθότητα (precision – degree of soundness) ορίζουμε τη δεσμευμένη πιθανότητα αν ταυτίζεται η κλάση που προβλέπει ένας ταξινομητής για ένα στιγμιότυπο με την πραγματική του κλάση:

$$precision\ c = \frac{\text{αριθμός σωστών προβλέψεων κλάσης } c}{\text{αριθμός προβλέψεων κλάσης } c}$$

Για ένα πρόβλημα ταξινόμησης δύο κλάσεων έχουμε:

$$precision\ p = \frac{TP}{TP+FP}, \text{ precision } n = \frac{TN}{TN+FN}$$

4.2.5. Εμβαδόν κάτω από την καμπύλη ROC

Η καμπύλη ROC (Receiver Operating Characteristic) είναι η γραφική παράσταση του ρυθμού των σωστών ταξινομήσεων των θετικών παραδειγμάτων (true positive rate ή recall), $\Pr [f(x) > \theta | Y = 1]$, ως προς το ρυθμό των ταξινομήσεων με σφάλμα των θετικών παραδειγμάτων (false positive rate), $\Pr [f(x) > \theta | Y = -1]$, καθώς γίνεται μεταβολή στο κατώφλι απόφασης θ . Η καμπύλη ROC αποτελεί ένα δισδιάστατο μέτρο αξιολόγησης ενός ταξινομητή, ενώ ένα άλλο μέτρο αξιολόγησης που χρησιμοποιείται πιο συχνά είναι το εμβαδόν κάτω από την καμπύλη (area under curve – AUC). Όταν το εμβαδόν είναι ίσο με 1 τότε ο ταξινομητής επιτυγχάνει τέλεια

ακρίβεια, αν το κατώφλι απόφασης επιλεγθεί σωστά, ενώ ένας ταξινομητής που ταξινομεί τα δεδομένα τυχαία έχει AUC ίσο με 0,5.

4.2.6. Ανάκληση (Recall)

Ως ανάκληση (recall – degree of completeness) ορίζεται η δεσμευμένη πιθανότητα αν ένα στιγμιότυπο ανήκει σε μια κλάση, έστω c , αυτή να αναγνωρισθεί σωστά από τον ταξινομητή:

$$recall\ p = \frac{TP}{TP+FN}$$

$$recall\ c = \frac{\text{αριθμός σωστών προβλέψεων κλάσης } c}{\text{αριθμός δεδομένων κλάσης } c}$$

Για ένα πρόβλημα ταξινόμησης δύο κλάσεων είναι:

$$recall\ p = \frac{TP}{TP + FN}$$

$$recall\ n = \frac{TN}{TN + FP}$$

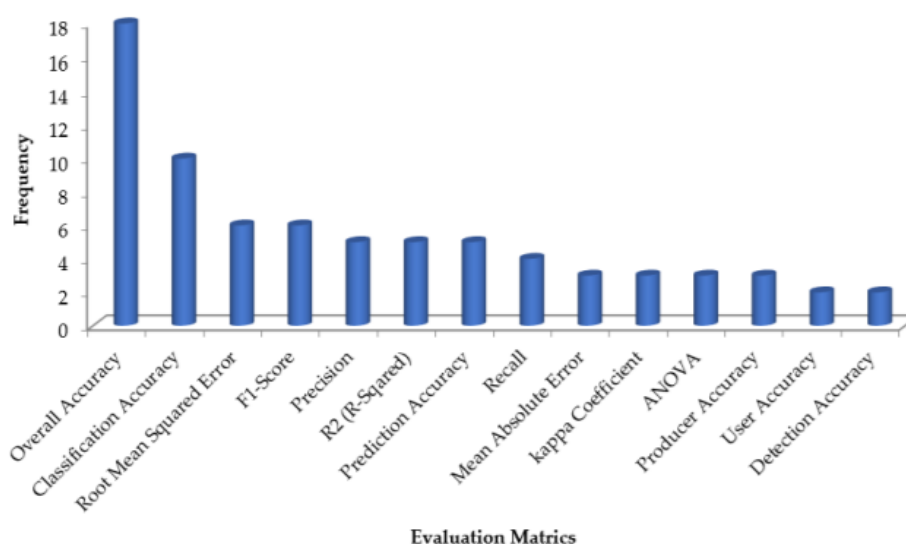
Όταν ένα σύνολο δεδομένων επεξεργάζεται μέσω του WEKA, στόχος είναι συνήθως να γίνεται διαχωρισμός των περιπτώσεων σε κλάσεις ενός από τα χαρακτηριστικά του. Για παράδειγμα, όπως θα αναλυθεί και παρακάτω, στο σύνολο δεδομένων της παραγωγής των βοσκοτόπων (pasture production) το χαρακτηριστικό της κλάσης ήταν η παραγωγή των βοσκοτόπων, το οποίο θα μπορούσε να είναι μια από τις τρεις απαριθμημένες τιμές, συγκεκριμένα lo, med και hi.

Επειδή το πρόγραμμα της μηχανικής μάθησης προσπαθεί να δημιουργήσει ένα σύνολο κανόνων καλύτερης προσαρμογής, οι κανόνες αυτοί δεν επιτρέπουν απαραίτητα για όλες τις περιπτώσεις να είναι σωστά τοποθετημένες στην κατάλληλη κατηγορία τους όπως καθορίζεται στο χαρακτηριστικό για την συγκεκριμένη περίπτωση. Το πρόγραμμα μηχανικής μάθησης υπολογίζει συνήθως αυτόματα ένα

μέτρο ακριβείας, δηλαδή το ποσοστό των περιπτώσεων που έχουν τοποθετηθεί σωστά στην κατάλληλη τάξη για κάθε επεξεργασία που εκτελείται δια μέσω αυτού.

Υποθετικά θα μπορούσε κανείς να δημιουργήσει ένα σύνολο κανόνων, ο αριθμός των οποίων θα ισούται με τον αριθμό περιπτώσεων για το σύνολο δεδομένων δοκιμής, το οποίο θα ταξινομούσε κάθε περίπτωση με 100% ακρίβεια. Έτσι για ένα σύνολο δεδομένων με 1000 περιπτώσεις θα χρειαζόταν ένα δέντρο αποφάσεων με 1000 κανόνες.

Τα συστήματα μηχανικής μάθησης επιτρέπουν τον καθορισμό παραμέτρων που θα καθορίσουν την ισορροπία μεταξύ των δέντρων με μεγάλο πλήθος κανόνων και υψηλή ακρίβεια σε πολύ μικρότερα και απλούστερα σύνολα κανόνων με ίσως ελαφρώς μειωμένη ακρίβεια.



Εικόνα 15: Οι συχνότεροι παράμετροι

4.2.7. 10-fold Cross Validation

Στη μέθοδο επικύρωσης 10 τμημάτων (10 fold cross validation) το σύνολο δεδομένων διαιρείται σε 10 υποσύνολα. Κάθε υποσύνολο περιέχει διαφορετικές παρατηρήσεις. Η επιλογή των υποσυνόλων είναι τυχαία. Ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο επικύρωσης και τα υπόλοιπα εννέα συνενώνονται και δημιουργούν το σύνολο εκπαίδευσης. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης και δοκιμάζεται έναντι του συνόλου επικύρωσης. Η διαδικασία επαναλαμβάνεται δέκα φορές, κάθε φορά χρησιμοποιώντας ένα διαφορετικό σύνολο

ως σύνολο επικύρωσης και τα υπόλοιπα εννέα ως σύνολο εκπαίδευσης. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου. Η μέθοδος μπορεί να διαφοροποιηθεί ως προς το πλήθος των τμημάτων. Γενικότερα ονομάζεται μέθοδος επικύρωσης k τμημάτων, όπου k συμβολίζει τον αριθμό των υποσυνόλων που δημιουργούνται και των επαναλήψεων [32].

ΚΕΦΑΛΑΙΟ 5

5.1. Eucalyptus Dataset

5.1.1. Περιγραφή συνόλου δεδομένων

Με βάση τα δεδομένα που συλλέχθηκαν τα μεγαλύτερα είδη καλλιέργειας φαίνονται πιο κατάλληλα σε περιοχές με ξηρό λόφο. Ο προσδιορισμός εντοπίζεται με μέτρηση του ύψους, της διαμέτρου κατά ύψος, της επιβίωσης και άλλων συντελεστών. Είναι σημαντικό να σημειωθεί ότι οι μέθοδοι δοκιμής ευκαλύπτου άλλαξαν με την πάροδο του χρόνου. Το πεδίο γνώσης επέτρεψε στους ερευνητές να καθορίσουν ότι τα μεγαλύτερα δέντρα θα εξυπηρετούσαν καλύτερα καθεστώτα διατήρησης/παραγωγής/αγροδοκιμασίας και τα μικρότερα σε ευρύτερα διαστήματα διατήρησης/ποιμαντισμού καθεστώτα. Μια σημαντική σημείωση της έρευνας είναι ότι οι συγκρίσεις προέλευσης σε είδη υποστηρίζουν τα ευρήματα πιο εκτεταμένων δοκιμών γενετικής βελτίωσης επικεντρώθηκαν στους ευκαλύπτους με τις περισσότερες δυνατότητες παραγωγής ξυλείας.

5.1.2. Πληροφορίες παραδειγμάτων και χαρακτηριστικώ

Τα σύνολα δεδομένων μπορεί να έχουν τιμές που λείπουν (missing values) και αυτό μπορεί να προκαλέσει προβλήματα σε πολλούς αλγόριθμους. Είναι καλή πρακτική να γίνεται προσδιορισμός και αντικατάσταση των τιμών που λείπουν για κάθε στήλη στα δεδομένα εισόδου πριν από τη μοντελοποίηση.

- Ορισμένες περιπτώσεις περιείχαν missing values[3]. Τα missing values συνήθως αντιπροσωπεύουν ένα φυτό που πεθαίνει κατά τη διάρκεια του πειράματος.
- Παραδείγματα: 736
- Χαρακτηριστικά: 20

Τα αρχικά δεδομένα περιείχαν δύο σύνολα δεδομένων. Το πρώτο περιγράφει τον τόπο και τις λεπτομέρειες της τοποθεσίας και το δεύτερο περιγράφει την ανάπτυξη του ευκαλύπτου, σύμφωνα με τις διαφορετικές περιοχές.

Είναι σημαντικό να σημειωθεί ότι οι μέθοδοι δοκιμής ευκαλύπτου άλλαξαν με την πάροδο του χρόνου. Οι προηγούμενες δοκιμές περιελάμβαναν κυρίως σπορόφυτο με

ύψος 15 - 30 εκατοστά που καλλιεργήθηκε σε τυρφώνες και οι μεταγενέστερες δοκιμές περιελάμβαναν τρεις αναπαραγωγές οκτώ δένδρων. Αυτή η αλλαγή μπορεί να συμβάλλει σε λιγότερο σημαντικά αποτελέσματα.

Οι πειραματικές διαδικασίες καταγραφής δεδομένων που είναι σημαντικές περιλαμβάνουν:

- περιπτώσεις χωρίς εγγραφή δεδομένων για τον λόγο ότι οι πειραματικές διαδικασίες απαιτούν την απουσία ενός είδους από ένα αντίγραφο σε μια τοποθεσία για να θεωρηθεί missing value, αλλά αν απουσιάζει από δύο ή περισσότερες επαναλήψεις σε μια τοποθεσία το είδος θα αποκλειστεί από τις αναλύσεις του τύπου.
- δεδομένα που λείπουν για επιβίωση, σφρίγος, αντοχή στα έντομα μορφή στελέχους και χρησιμότητα. Αυτό θα μπορούσε να οδηγήσει στον θάνατο πολλών ειδών σε αυτές τις περιοχές ή στην έλλειψη συλλογής δεδομένων.

Πληροφορίες χαρακτηριστικών

	Attribute	Type	Values	Missing
1	Sp (species)	enumerated	co, fr, ..., te	0
2	PMCNo (seedlot number)	integer	1.0 - 3.275.0	7
3	DBH (best diameter base height, cm)	real	0.58 - 42.085.0	1
4	Ht (m)	real	1.12 - 21.79	0
5	Surv (survival)	integer	1.0 - 100.0	94
6	Vig (vigour)	real	0.5 - 5.0	69
7	Ins_res (insect resistance)	real	0.0 - 4.5	69

8	Stem Fm (stem form)	real	0.0 - 0.5	0
9	Crown Fm (crown form)	real	0.0 - 5.0	0
10	Brnch Fm (branch form)	real	0.0 - 5.0	69
11	Utility (utility rating)	real	0.0 - 5.0	94
12	Name (name of site)	enumerated (15)	Clydebank, Craggy Range Road,.....,Wimbledon Species	0
13	Abbreviation (site abbreviation)	enumerated (15)	Cly, Cra,.....,WSp	0
14	Locality (where the site is in the North Island)	enumerated	Central_Hawkes_Bay, Northern_Hawkes_Bay,...., South_Wairarapa	0
15	Map Ref (map location in the North Island, NZMS1)	enumerated	N135_382/137, N116_848/985,.....,N151_922/226	0
16	Latitude approximation (South - deg.,min)	enumerated	29_38, 39_00,.....,82_32	0
17	Altitude (m)	integer	70 - 300	0
18	Rainfall (mm pa)	integer	850 - 1,750	0
19	Frosts (deg.C)	integer	-3.0 - -2.0	0
20	Year (year of planting)	integer	1980 - 1986	0

Πληροφορίες κλάσης

Υπάρχουν δύο τιμές κλάσης που δημιουργήθηκαν για αυτό το σύνολο δεδομένων από τα χαρακτηριστικά που ήδη υπάρχουν.

Utility

1. None	180
2. Low	107
3. Average	130
4. Good	214
5. Best	105

5.1.3. Διαδικασία ανάλυσης δεδομένων

Τα δύο σύνολα δεδομένων συνδυάστηκαν για να δημιουργηθεί ένα σύνολο δεδομένων με χαρακτηριστικά και από τα δύο.

Η αρχική μέτρηση χρησιμότητας κατηγοριοποιήθηκε σε ένα καινούργιο utility rating:

- none < 0.1
- low < 1.48
- average < 2.65
- good < 3.83
- best < 5.00

- Το απλοποιημένο δέντρο έχει μέγεθος: 62 κόμβους με 47.7% σφάλματα και το κυρίαρχο χαρακτηριστικό είναι το 'Species', το οποίο διαίρεσαι τα δεδομένα περισσότερο από τον αριθμό της παρτίδα των σπόρων αφού πολλά είδη ήταν όλα σε μια κατηγορία.

- Το απλοποιημένο δέντρο έχει μέγεθος: 234 κόμβους με 14.7% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι τα 'Vigor', 'Species' και 'PMCNo'.

- Το απλοποιημένο δέντρο έχει μέγεθος: 227 κόμβους με 7.7% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι τα ‘Year’, διαιρώντας τα δεδομένα στο 1983, ‘Vigour’, ‘Survival’ και ‘Height’.
- Το απλοποιημένο δέντρο έχει μέγεθος: 151 κόμβους με 26,5% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι τα ‘Year’, διαιρώντας τα δεδομένα στο 1983, ‘Vigour’ και ‘Species’.

Τα τρία τελευταία δέντρα αποφάσεων χρησιμοποίησαν το ‘Vigour’ ως ένα από τους πρώτους διαιρέτες των δεδομένων. Υπάρχουν missing values για τα survival, vigour, insect resistance, stem form, crown form, branch form και utility. Τέλος δεν καταγράφηκαν στοιχεία για το 1984 και το 1985 που εξηγεί την πρώτη διαίρεση σε δύο από αυτά τα δέντρα αποφάσεων δίνοντας ‘none’ ως utility rating.

5.1.4. Αποτελέσματα

C 4.5

Παρακάτω γίνεται ανάλυση του συγκεκριμένου συνόλου δεδομένων με τον αλγόριθμο C4.5 με 10-fold cross validation όπως και οι υπόλοιποι.

Το δέντρο που δημιουργείται έχει την ακόλουθη δομή:

- Αριθμός φύλλων : 81
- Μέγεθος του δέντρου : 111

Τα αποτελέσματα κατάταξης είναι τα ακόλουθα:

Algorithm	TP Rate	Precision	ROC Area	Accuracy
C4.5	0.639	0.651	0.842	63.8587%

Πίνακας 3

Confusion Matrix

a	b	c	d	e	Classified as
151	28	1	0	0	a = none
46	64	18	9	0	b = low
0	21	55	52	2	c = average
0	3	36	160	15	d = good
0	0	6	59	40	e = best

Από τον πίνακα σύγχυσης προκύπτει TP Rate = 0.639, Precision = 0.651, ROC Area = 0.842 και Accuracy = 63.8587%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου είναι πολύ καλή.

Random Forest

Algorithm	TP Rate	Precision	ROC Area	Accuracy
Random Forest	0.575	0.575	0.861	57.4748%

Πίνακας 4

Confusion Matrix

a	b	c	d	e	Classified as
139	30	5	5	1	a = none
31	43	24	9	0	b = low
4	21	54	48	3	c = average
2	4	39	135	34	d = good

0	1	5	47	52	e = best
---	---	---	----	----	-----------------

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.575, Precision = 0.575, ROC Area = 0.861 και Accuracy = 57.4748%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου είναι καλή.

KNN

Algorithm	TP Rate	Precision	ROC Area	Accuracy
KNN/IBk	0.516	0.524	0.701	51.6304%

Πίνακας 5

Confusion Matrix

a	b	c	d	e	Classified as
128	34	9	6	3	a = none
26	39	25	15	2	b = low
7	26	47	45	5	c = average
3	15	43	109	44	d = good
0	1	8	39	57	e = best

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.516, Precision = 0.651, ROC Area = 0.701 και Accuracy = 51.6304%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου δεν είναι πολύ καλή και ο αλγόριθμος IBk δεν μπορεί να θεωρηθεί κατάλληλος ταξινομητής.

Naive Bayes

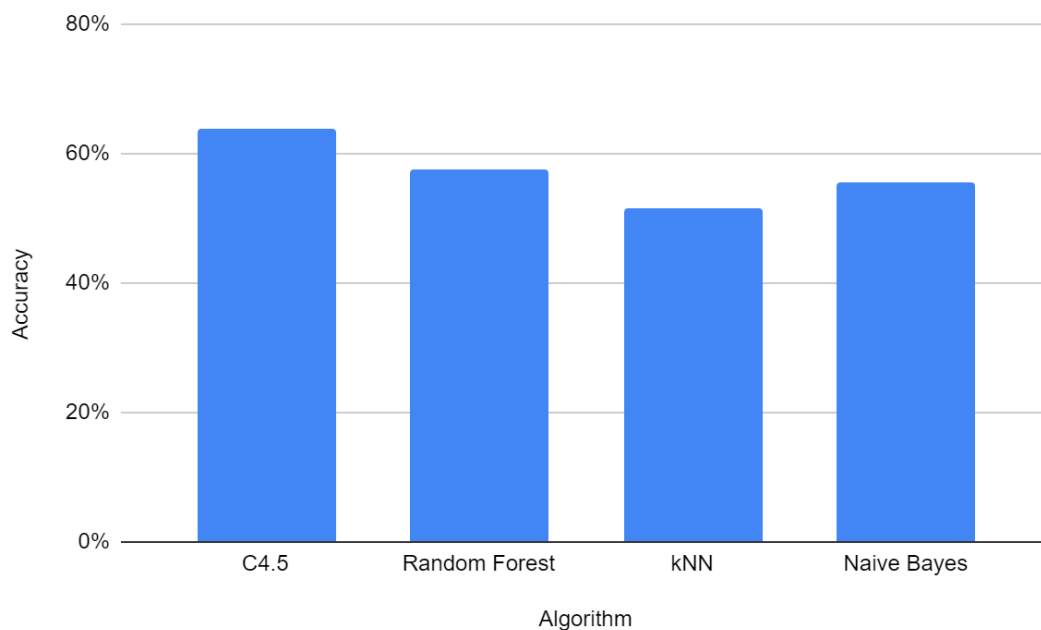
Algorithm	TP Rate	Precision	ROC Area	Accuracy
Naive Bayes	0.556	0.628	0.863	55.5707%

Πίνακας 6

Confusion Matrix

a	b	c	d	e	Classified as
104	55	15	6	0	a = none
5	66	24	12	0	b = low
2	31	59	38	0	c = average
0	7	46	152	9	d = good
0	2	18	57	28	e = best

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.556, Precision = 0.628, ROC Area = 0.863 και Accuracy = 55.5707%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου είναι καλή.



Πίνακας 7

Στο παραπάνω γράφημα φαίνεται ότι αλγόριθμος C4.5 έχει μεγαλύτερη ακρίβεια από τους υπόλοιπους και έτσι θεωρείται ο καλύτερος ταξινομητής για αυτό το σύνολο δεδομένων. Το ποσοστό ακρίβειας του kNN είναι αρκετά χαμηλό οπότε δεν είναι κατάλληλος.

5.1.5. Συμπεράσματα

Λόγω της αρχικής ανάλυσης που πραγματοποιήθηκε με το «PMcNo» ως ακέραιο τα πρώτα τέσσερα αποτελέσματα απορρίφθηκαν. Ο σωστός τύπος από το «PMcNo» είναι της απαρίθμησης.

Τα δέντρα και οι κανόνες που βρέθηκαν από την ανάλυση του αλγορίθμου C4.5 πρέπει να είναι μικρότεροι και ακριβέστεροι, για παράδειγμα λιγότεροι από 20 κόμβους και σφάλματα 10%. Τα δέντρα με τις καλύτερες αποφάσεις έχουν ποσοστά σφάλματος πάνω από 10%.

Ιδανικά, τα δέντρα και οι κανόνες που προκύπτουν από την ανάλυση του αλγορίθμου C4.5 θα πρέπει να είναι μικρότερα και ακριβέστερα, για παράδειγμα λιγότεροι από 20 κόμβοι και 10% σφάλματα. Τα καλύτερα δέντρα αποφάσεων έχουν ποσοστά σφάλματος πάνω από 10%. Αυτά τα ποσοστά σφάλματος μπορεί να οφείλονται στο γεγονός ότι έλειπαν δεδομένα από μια σειρά χαρακτηριστικών. Ορισμένες από αυτές τις τιμές που λείπουν μπορεί να είναι άγνωστες ή μη μετρημένες τιμές.

Η ανάλυση μηχανικής μάθησης έχει οδηγήσει στην ανάγκη για εντοπισμό σχέσεων ή μοτίβων στα δεδομένα. Κάποια δεδομένα μπορεί να διαφέρουν από ένα καθιερωμένο μοτίβο. Αν και τα δεδομένα ενδεχομένως να είναι σωστά, η διαφοροποίηση μπορεί να μην αναγνωρίζεται από τον αλγόριθμο, που χρησιμοποιείται, και για αυτό αποδίδεται υψηλότερος αριθμός σφαλμάτων. Σε αυτή την περίπτωση το γνωστικό πεδίο και η έρευνα είναι σημαντικά για τον καθορισμό ή μη του αριθμού σφαλμάτων. Ο μεγάλος αριθμός κόμβων στα περισσότερα δέντρα μπορεί να αποδοθεί από τον μεγάλο αριθμό χαρακτηριστικών και των τιμών σε καθένα από αυτά τα χαρακτηριστικά.

Τα αποτελέσματα που προκύπτουν από την μηχανική μάθηση είναι ενδιαφέροντα. Λόγω του μεγάλου αριθμού σφαλμάτων, το γνωστικό πεδίο απαιτείται για περαιτέρω ανάλυση των αποτελεσμάτων. Τα δέντρα δείχνουν ποιες ιδιότητες και ποιες τιμές συμβάλλουν στην καλύτερη διατήρηση του εδάφους που υπάρχουν τα δέντρα ευκαλύπτου.

5.2. Grub-damage Dataset

5.2.1. Περιγραφή συνόλου δεδομένων

Η κάμπια στο γρασίδι είναι ένα από τα κύρια έντομα στους βοσκοτόπους και μπορούν να προκαλέσουν σοβαρές ζημιές στα λιβάδια και πολλές οικονομικές απώλειες. Η ποιμαντική βλάβη μπορεί να συμβαίνει περιοδικά σε ευρείες περιοχές. Οι πληθυσμοί των συγκεκριμένων εντόμων επηρεάζονται από βιοτικούς παράγοντες, όπως ασθένειες και γεωργικές πρακτικές και άρδευση. Ο σκοπός αυτού του συνόλου δεδομένων πρόκειται να υποβάλλει αναφορά για τον πληθυσμό και τις ζημιές των εντόμων στο γρασίδι, ώστε να παρέχει αντικειμενικές εκτιμήσεις των ετήσιων απωλειών που προκαλούνται από τις εκβολές αυτών.

Στόχος ήταν να βρεθεί μια σχέση μεταξύ του αριθμού των grass grubs, της άρδευσης και της κατάταξης ζημιών μεταξύ ενός συγκεκριμένου χρονικού διαστήματος

Πραγματοποιήθηκαν 2 μελέτες εκ των οποίων η μία υπολόγισε τον μέσο όρο της πυκνότητας του πληθυσμού των grass grubs τον μήνα του Αυγούστου και κυμάνθηκε από 32-109/m². Η δεύτερη μελέτη ανέφερε ότι ο μέσος όρος της πυκνότητας του πληθυσμού τον μήνα του Σεπτεμβρίου κυμάνθηκε από 37-114/m². Ο μέσος όρος της πυκνότητας του πληθυσμού μέσα σε αυτούς τους δύο μήνες συγκρίθηκε με τις καιρικές συνθήκες και με μια σημαντική συσχέτιση που βρέθηκε με βροχοπτώσεις τον προηγούμενο Νοέμβριο. Ο μέσος όρος των ζημιών των βοσκοτόπων εκτιμήθηκε στο 8%. Σημαντικές σχέσεις βρέθηκαν μεταξύ της ζημιάς και της πυκνότητας των grass grubs και μεταξύ της ζημιάς και των φθινοπωρινών βροχοπτώσεων.

5.2.2. Πληροφορίες παραδειγμάτων και χαρακτηριστικών

- Παραδείγματα: 155
- Χαρακτηριστικά: 9

Πληροφορίες χαρακτηριστικών

	Attribute	Type	Values	Missing
1	Year_Zone	enumerated (20 values)	0c, 0f, 0m,.....9c, 9f, 9m	0

2	Year	enumerated	1986 - 1992	0
3	strip	integer	0.0 - 5.0	0
4	pk (paddock)	integer	0.0 - 5.0	0
5	damage_rankRJT	integer	0.0 - 5.0	17
6	damage_rankALL	enumerated	0.0 - 5.0	5
7	GG_per_m2 (grass grubs per metre square)	real	0.0 - 708.0	12
8	dry_or_irr	enumerated	D (dryland), O (irrigated overhead), B (irrigated border dyke)	6
9	zone	enumerated	F (foothills), M (midplain), C (coastal)	0

Πίνακας 8: Πίνακας χαρακτηριστικών

Πληροφορίες κλάσης

Για αυτό το σύνολο δεδομένων έχει δημιουργηθεί μια μόνο κλάση.

GG_new

- | | |
|--------------|----|
| 1. low | 49 |
| 2. average | 41 |
| 3. high | 46 |
| 4. very high | 19 |

Η νέα κλάση GG_new καθορίζεται από το χαρακτηριστικό GG_per_m2 υπολογίζοντας:

- low < 60
- average < 130
- high < 280
- very high >280

Σε όλες τις ακόλουθες περιπτώσεις τα χαρακτηριστικά year, lane και Mantra δεν χρησιμοποιήθηκαν. Τα συγκεκριμένα δεν θα μπορούσαν να προβλέψουν κανένα άλλο παράδειγμα εκείνων που βρίσκονται μέσα στο training set.

5.2.3. Διαδικασία ανάλυσης δεδομένων

Οι δοκιμές στο training set ήταν:

- ο C4.5 χρησιμοποιήθηκε με την κλάση GG_new και τα χαρακτηριστικά damage_rankRGT, damage+rankALL, dry_or_irr και zone. Τα αποτελέσματα του δέντρου έχουν μέγεθος 15 κόμβους και 43.2% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι damage_rankRJ και damage_rankALL.
- ο C4.5 χρησιμοποιήθηκε με την κλάση GG_new και τα χαρακτηριστικά dry_or_irr και zone. Τα αποτελέσματα του δέντρου έχουν μέγεθος 7 κόμβους και 54.8% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι dry_or_irr και zone. Τα δεδομένα χωρίστηκαν έτσι ώστε τα αρδευόμενα (Ο ή Β) να είναι ταξινομήμενα κατά μέσο όρο. Τα ξηρά οικόπεδα, οι πρόποδες έχουν low grass grubs και τα παράκτια έχουν high grass grubs.
- ο C4.5 χρησιμοποιήθηκε ως κλάση την damage_rankRJT και τα χαρακτηριστικά damage_rankALL, GG_new, dry_or_irr και zone. Τα αποτελέσματα του δέντρου έχουν μέγεθος 21 κόμβους και 39.4% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι damage_rankALL, GG_new και dry_ir_irr.
- ο C4.5 χρησιμοποιήθηκε ως κλάση την damage_rankRJT και τα χαρακτηριστικά GG_per_m2, GG_new, dry_or_irr και zone. Τα αποτελέσματα του δέντρου έχουν μέγεθος 60 κόμβους και 43.2% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι GG_per_m2, GG_new, dry_or_irr και zone.
- ο C4.5 χρησιμοποιήθηκε ως κλάση την damage_rankALL και τα χαρακτηριστικά damage_rankRJT, GG_new, dry_or_irr και zone. Τα αποτελέσματα του δέντρου έχουν μέγεθος 10 κόμβους και 40.6% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι damage_rankRJT και GG_new.

- ο C4.5 χρησιμοποιήθηκε ως κλάση την damage_rankALL και τα χαρακτηριστικά GG_per_m2, dry_or_irr και zone. Τα αποτελέσματα του δέντρου έχουν μέγεθος 59 κόμβους και 43.2% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι GG_per_m2.

5.2.4. Αποτελέσματα

C 4.5

Παρακάτω γίνεται ανάλυση του συγκεκριμένου συνόλου δεδομένων με τον αλγόριθμο C4.5 με 10-fold cross validation όπως και οι υπόλοιποι.

Το δέντρο που δημιουργείται έχει την ακόλουθη δομή:

- Αριθμός φύλλων : 56

- Μέγεθος του δέντρου : 69

Τα αποτελέσματα κατάταξης είναι τα ακόλουθα:

Algorithm	TP Rate	Precision	ROC Area	Accuracy
C4.5	0.335	0.322	0.592	33.5484%

Πίνακας 9

Confusion Matrix

a	b	c	d	Classified as
28	10	8	3	a = low
13	10	15	3	b = average
13	18	11	4	c = high
4	5	7	3	d = very high

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.335, Precision = 0.332, ROC Area = 0.592 και Accuracy = 33.5484%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου δεν είναι καθόλου καλή.

Random Forest

Algorithm	TP Rate	Precision	ROC Area	Accuracy
Random Forest	0.348	0.343	0.649	34.8387%

Πίνακας 10

Confusion Matrix

a	b	c	d	Classified as
24	15	8	2	a = low
15	11	11	4	b = average
12	14	16	4	c = high
5	4	7	3	d = very high

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.348, Precision = 0.343, ROC Area = 0.649 και Accuracy = 34.8387%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου δεν είναι καθόλου καλή.

KNN

Algorithm	TP Rate	Precision	ROC Area	Accuracy
KNN/IBk	0.394	0.387	0.587	39.3548%

Πίνακας 11

Confusion Matrix

a	b	c	d	Classified as
27	7	10	5	a = low
11	14	12	4	b = average
11	15	16	4	c = high
7	4	4	4	d = very high

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.394, Precision = 0.387, ROC Area = 0.587 και Accuracy = 39.3548%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου δεν είναι πολύ καλή και ο αλγόριθμος IBk δίνει μεγαλύτερο ποσοστό ακρίβειας σε σχέση με τους άλλους.

Naive Bayes

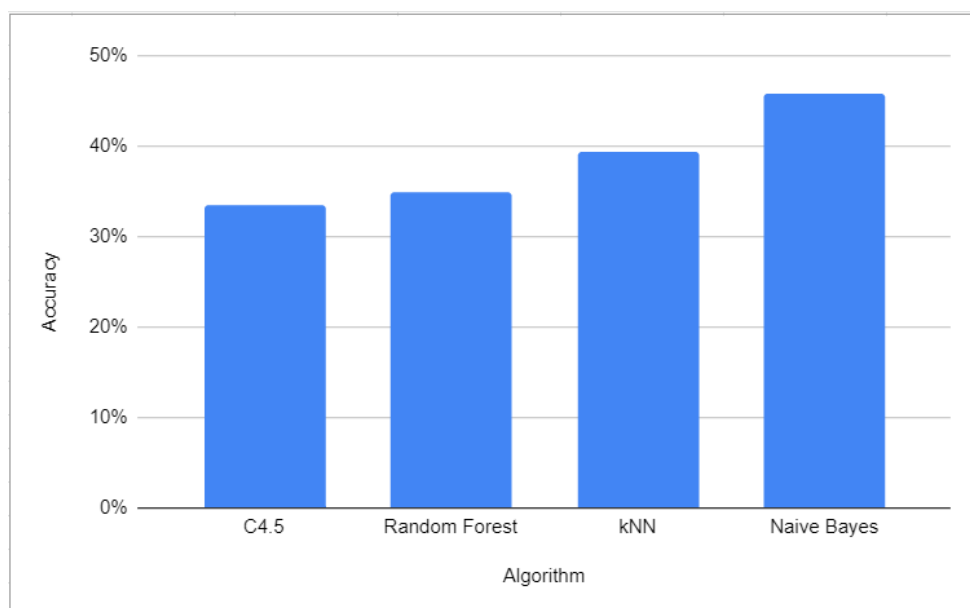
Algorithm	TP Rate	Precision	ROC Area	Accuracy
Naive Bayes	0.458	0.457	0.722	45.8065%

Πίνακας 12

Confusion Matrix

a	b	c	d	Classified as
32	12	2	3	a = low
9	10	17	5	b = average
9	15	19	3	c = high
2	5	2	10	d = very high

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.458, Precision = 0.457, ROC Area = 0.722 και Accuracy = 45.8065%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου είναι η καλύτερη μέχρι στιγμής με το υψηλότερο ποσοστό ακρίβειας.



Πίνακας 13

Στο παραπάνω γράφημα φαίνεται ότι αλγόριθμος Naive Bayes έχει μεγαλύτερη ακρίβεια από τους υπόλοιπους και έτσι θεωρείται ο καλύτερος ταξινομητής για αυτό το σύνολο δεδομένων. Το ποσοστό ακρίβειας όλων των υπόλοιπων αλγορίθμων είναι αρκετά χαμηλό.

5.2.5. Συμπεράσματα

Η κατάταξη των ζημιών τόσο του RJT όσο και του ALL σχετίζονται με τον αριθμό των grass grubs. Οι ζώνες και οι μέθοδοι άρδευσης δεν φαίνεται να παίζουν καθοριστικό ρόλο στον προσδιορισμό του αριθμού των grass grubs όταν χρησιμοποιούνται με χαρακτηριστικά damage rank. Ωστόσο, όταν το GG_new χρησιμοποιήθηκε σαν κλάση με τα zone και dry_or_irr μπόρεσε να γίνει διαχωρισμός των irrigated paddocks από τα dry paddocks. Όταν ένα από τα χαρακτηριστικά της κατάταξης ζημιάς χρησιμοποιήθηκε ως κλάση, το άλλο χαρακτηριστικό ήταν αυτό που συνέβαλε στο δέντρο. Αυτό δείχνει ότι οι βαθμολογίες είναι παρόμοιες. Όταν έγινε χρήση ενός από τα χαρακτηριστικά της κατάταξης ζημιάς ως κλάση χωρίς τα άλλα χαρακτηριστικά, το δέντρο είναι πολύ μεγάλο με τον αριθμό των grass grubs να παίζουν σημαντικότερο ρόλο στον καθορισμό της ζημιάς. Σε όλα τα δέντρα αποφάσεων υπάρχει μεγάλος αριθμός σφαλμάτων με αποτέλεσμα να αμφισβητείται η εγκυρότητα των δέντρων.

Στην αρχή αυτό το σύνολο δεδομένων φάνηκε να είναι αρκετά ευνοϊκό για την μηχανική μάθηση καθώς έχει χαρακτηριστικά που μπορούν εύκολα να χρησιμοποιηθούν ως κλάση. Παρόλα αυτά τα δέντρα αποφάσεων που δημιουργήθηκαν ήταν αρκετά μικρά και είχαν υψηλό ποσοστό σφαλμάτων. Το υψηλό πλήθος σφαλμάτων είναι πιθανό να οφείλεται στον αριθμό των τιμών που λείπουν για τα χαρακτηριστικά `damage_rankRJT` και `damage_rankALL`. Μικρότερο ποσοστό σφαλμάτων θα μπορούσε να βρεθεί αν όλοι οι τύποι είχαν απαριθμηθεί σε ομάδες μεταξύ πέντε και επτά τιμών.

5.3. Pasture Dataset

5.3.1. Περιγραφή συνόλου δεδομένων

Στόχος του συγκεκριμένου συνόλου δεδομένων ήταν να γίνει πρόβλεψη της παραγωγής του βοσκοτόπου από μια ποικιλία βιοφυσικών παραγόντων. Οι μεταβλητές της βλάστησης και του εδάφους από διάφορα σημεία ενός λόφου (εφαρμογή λιπάσματος/ποσοστό αποθέματος) μετρήθηκαν και υποδιαιρέθηκαν σε 36 αγρούς. Δεκαεννέα τιμές βλάστησης επιλέχθηκαν ως χρήσιμοι βιοφυσικοί δείκτες.

Το αποτέλεσμα για αυτές τις μεταβλητές είναι ότι βοτανική σύνθεση των βοσκοτόπων, κυρίως αυτοί που περιέχουν χόρτα υψηλής απόκρισης στη γονιμότητα και σε ποώδη είδη ζιζανίων, και η μάζα του γαιοσκώληκα ανά μονάδα της κάθε περιοχής εξήγησαν το μεγαλύτερο μέρος της διακύμανσης των παραγόμενων δεδομένων. Αυτές οι μεταβλητές συσχετίστηκαν πολύ και με την παραγωγή των βοτάνων, έναν δείκτη πιθανών οικονομικών επιδόσεων.

Το αρχικό σύνολο περιλαμβάνει ακόμα μια νέα μεταβλητή, την συγκέντρωση φωσφόρου στο φύλλο (ug/g).

5.3.2. Πληροφορίες παραδειγμάτων και χαρακτηριστικών

- Παραδείγματα: 36
- Χαρακτηριστικά: 23

Περιγραφή χαρακτηριστικών

	Attribute	Type	Values	Missing
1	Fertiliser	enumerated	LL, LN, HN, HH	0

2	Slope	integer	7.0 - 33.0	0
3	Aspect (deviation from north-west)	integer	17.0 - 172.0	0
4	OlsenP	integer	5.0 - 42.0	0
5	MinN	integer	162.0 - 409.0	0
6	TS	integer	175.0 - 655.0	0
7	Ca:Mg (Calcium Magnesium ratio)	real	2.22 - 6.36	0
8	LOM (g/100g)	real	1.3 - 3.7	0
9	NFIX mean	real	0.012 - 0.229	0
10	Eworms (main 3 spp - g/m ²)	real	0.0 - 285.7	0
11	Eworms (number of spp)	integer	1.0 - 5.0	0
12	KUnSat (mm/hr)	real	4.4 - 57.3	0
13	OM	real	5.5 - 11.7	0
14	Air-Perm	real	2.63e-13 - 1.03e-11	0
15	Porosity	real	0.126 - 0.212	0
16	HFRG% (mean)	real	1.31 - 57.26	0
17	legume-yield (kgDM/ha)	real	26.4 - 642.7	0

18	OSPP% (mean)	real	1.19 - 19.07	0
19	Jan-Mar-mean-TDR	real	18.0 - 39.3	0
20	Annual-Mean-Runoff (mm)	real	578.9 - 893.5	0
21	root-surface-area (m2 /m3)	real	56.2 - 520.6	0
22	Leaf-P (ppm)	real	1162.0 - 3993.0	0
23	pasture-prod-class (class variable)	enumerated	LO, MED, HI	0

Πίνακας 14: Πίνακας χαρακτηριστικών

Πληροφορίες κλάσης

Για αυτό το σύνολο δεδομένων έχει δημιουργηθεί μια μόνο κλάση.

Pasture-prod-class

1. LO (12 lowest values)
2. MED (12 intermediate values)
3. HI (12 highest values)

5.3.3. Διαδικασία ανάλυσης δεδομένων

Οι δοκιμές στο training set ήταν:

- ο C4.5 χρησιμοποιήθηκε με την κλάση pasture-prod-class και τα χαρακτηριστικά paddock_id, Air-Perm και Leaf-P. Τα αποτελέσματα του δέντρου έχουν μέγεθος 9 κόμβους και 2.8% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι HFRG%mean, Leaf-P, legume-yield και Eworms-main-3.

- ο C4.5 χρησιμοποιήθηκε με την κλάση pasture-prod-class και τα χαρακτηριστικά paddock_id, Air-Perm και Leaf-P. Τα αποτελέσματα του δέντρου έχουν μέγεθος 9 κόμβους και 2.8% σφάλματα και τα κυρίαρχα χαρακτηριστικά είναι HFRG%mean, NFIX-mean, legume-yield και Eworms-main-3.

5.3.4. Αποτελέσματα

C 4.5

Παρακάτω γίνεται ανάλυση του συγκεκριμένου συνόλου δεδομένων με τον αλγόριθμο C4.5 με 10-fold cross validation όπως και οι υπόλοιποι.

Το δέντρο που δημιουργείται έχει την ακόλουθη δομή:

- Αριθμός φύλλων : 6
- Μέγεθος του δέντρου : 9

Τα αποτελέσματα κατάταξης είναι τα ακόλουθα:

Algorithm	TP Rate	Precision	ROC Area	Accuracy
C4.5	0.778	0.772	0.841	77.7778%

Πίνακας 15

Confusion Matrix

a	b	c	Classified as
9	3	0	a = LO
4	7	1	b = MED
0	0	12	c = HI

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.778, Precision = 0.772, ROC Area = 0.841 και Accuracy = 77.7778%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου είναι πάρα πολύ καλή.

Random Forest

Algorithm	TP Rate	Precision	ROC Area	Accuracy
Random Forest	0.861	0.868	0.935	86.1111%

Πίνακας 16

Confusion Matrix

a	b	c	Classified as
10	2	0	a = LO
2	10	0	b = MED
0	1	11	c = HI

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.861, Precision = 0.868, ROC Area = 0.935 και Accuracy = 86.1111%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου είναι πολύ καλή και το ποσοστό ακρίβειας πολύ υψηλό.

KNN

Algorithm	TP Rate	Precision	ROC Area	Accuracy
KNN/IBk	0.694	0.719	0.826	69.4444%

Πίνακας 17

Confusion Matrix

a	b	c	Classified as
7	4	1	a = LO
2	9	1	b = MED
0	3	9	c = HI

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.694, Precision = 0.719, ROC Area = 0.826 και Accuracy = 69.4444%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου δεν είναι τόσο καλή συγκριτικά με τους άλλους αλγορίθμους.

Naive Bayes

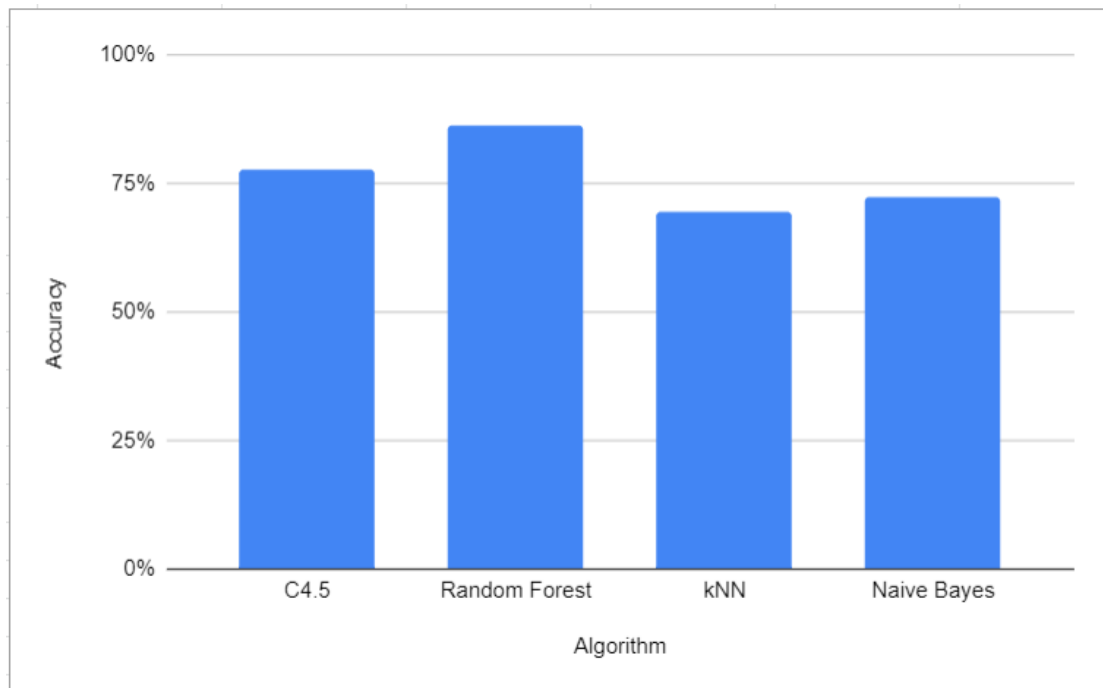
Algorithm	TP Rate	Precision	ROC Area	Accuracy
Naive Bayes	0.722	0.724	0.837	72.2222%

Πίνακας 18

Confusion Matrix

a	b	c	Classified as
10	2	0	a = LO
3	7	2	b = MED
0	3	9	c = HI

Από τον πίνακα σύγκρισης προκύπτει TP Rate = 0.722, Precision = 0.724, ROC Area = 0.837 και Accuracy = 72.2222%. Η ταξινόμηση με τη χρήση αυτή του αλγορίθμου είναι αρκετά καλή με υψηλό ποσοστό ακρίβειας.



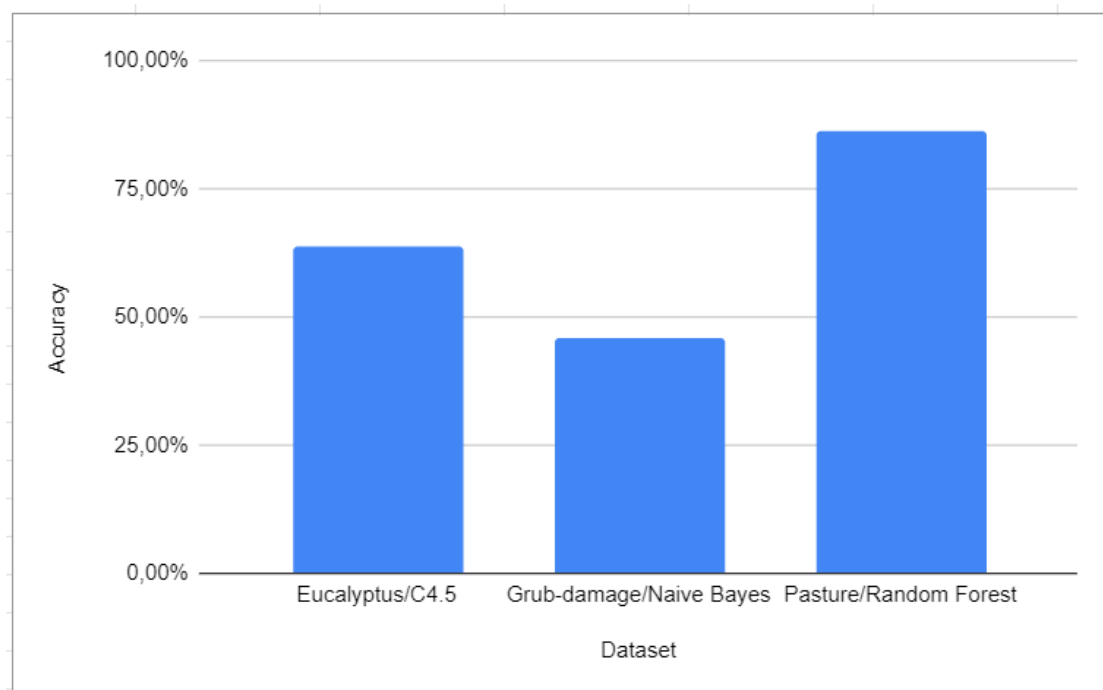
Πίνακας 19

Στο παραπάνω γράφημα φαίνεται ότι αλγόριθμος Random Forest έχει μεγαλύτερη ακρίβεια από τους υπόλοιπους και έτσι θεωρείται ο καλύτερος ταξινομητής για αυτό το σύνολο δεδομένων. Το ποσοστό ακρίβειας όλων των υπόλοιπων αλγορίθμων είναι αρκετά υψηλό.

5.3.5. Συμπεράσματα

Παράχθηκαν μικρά δέντρα αποφάσεων που περιγράφουν πιθανούς τρόπους ταξινόμησης της παραγωγής των βοσκοτόπων. Είναι ενδιαφέρον ότι όταν αφαιρεθεί το χαρακτηριστικό Leaf-P το δέντρο που προκύπτει είναι αρκετά παρόμοιο με τα δέντρα των χαρακτηριστικών Eworms-main-3 και legume-yield.

Το συγκεκριμένο σύνολο δεδομένων είναι κατάλληλο για να αναλυθεί με μηχανική μάθηση αλλά το μέγεθος του είναι αρκετά μικρό ώστε τα αποτελέσματα να είναι σίγουρα ακριβή. Τα αποτελέσματα που βρέθηκαν θα μπορούσαν να είναι πολύ σαφή. Η ανάλυση με περισσότερα παραδείγματα θα μπορούσαν να δημιουργήσουν ένα διαφορετικό αποτέλεσμα.



Πίνακας 20

Στο παραπάνω διάγραμμα φαίνεται η υψηλότερη ακρίβεια κάθε συνόλου δεδομένων με την χρήση διαφορετικών αλγορίθμων. Υπάρχει μεγάλη διαφορά μεταξύ των τιμών τους γιατί ο αριθμός των χαρακτηριστικών και των παραδειγμάτων που περιλαμβάνει το κάθε σύνολο δεδομένων είναι διαφορετικός. Τα αποτελέσματα των πειραμάτων που έγιναν στην έρευνα [3] σε συνδυασμό με τις πειραματικές διαδικασίες σε αυτήν εδώ την εργασία συμπίπτουν. Χρησιμοποιήθηκαν κοινά μοντέλα ταξινόμησης αλλά και διαφορετικά, όπως Random Forest, Naïve Bayes και αποδείχτηκε ότι είναι εξίσου αποδοτικά.

ΚΕΦΑΛΑΙΟ 6

6.1. Σύνοψη και συμπεράσματα

Στα κεφάλαια 1 και 2 γίνεται περιγραφή της Γεωργίας Ακριβείας, της Ψηφιακής Γεωργίας και πως η Μηχανική Μάθηση επιδρά σε αυτές. Ακόμη, αναλύονται κάποιες έρευνες και γίνεται αναφορά στα αποτελέσματα. Στα κεφάλαια 3 και 4 γίνεται ανάλυση της μεθοδολογίας που επιλέχτηκε για αυτή την εργασία και περιγραφή των αλγορίθμων που χρησιμοποιούνται για την ανάλυση των τριών συνόλων δεδομένων αντίστοιχα. Στο κεφάλαιο 5 γίνονται οι πειραματικές διαδικασίες, μέσω του λογισμικού WEKA [33], των μοντέλων για κάθε σύνολο δεδομένων με τα ανάλογα αποτελέσματα που προκύπτουν. Τέλος, γίνεται μια σύνοψη της συνολικής εργασίας.

6.2. Μελλοντικές κατευθύνσεις

Αν και είναι πάντα συναρπαστικό να διαβάζει κάποιος για το μέλλον, το πιο σημαντικό μέρος είναι η τεχνολογία που ανοίγει τον δρόμο για αυτό. Η γεωργική μηχανική εκμάθηση δεν είναι ένα περιέργο τέχνασμα αλλά ένα σύνολο από καλά καθορισμένα μοντέλα που συλλέγουν συγκεκριμένα δεδομένα και εφαρμόζουν συγκεκριμένους αλγορίθμους για να πετύχουν τα αναμενόμενα αποτελέσματα.

Μέχρι στιγμής η κατανομή της μηχανικής μάθησης είναι άνιση σε όλη την γεωργία. Κυρίως οι τεχνικές μηχανικής μάθησης χρησιμοποιούνται σε διαδικασίες διαχείρισης καλλιεργειών ακολουθώντας την διαχείριση των συνθηκών εκτροφής και την διαχείριση των ζώων.

Η βιβλιογραφική ανασκόπηση δείχνει ότι τα πιο δημοφιλή μοντέλα στη γεωργία ακριβείας είναι τα τεχνητά και βαθιά νευρωνικά δίκτυα (Artificial and Deep Neural Networks/ANNs and DL) και οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines/SVMs).

Τα ANNs είναι εμπνευσμένα από τη λειτουργικότητα του ανθρώπινου εγκεφάλου και αντιπροσωπεύουν ένα απλοποιημένο μοντέλο της δομής του βιολογικού νευρωνικού δικτύου που μιμείται πολύπλοκες λειτουργίες όπως η δημιουργία προτύπων, η γνώση, η μάθηση και η λήψη αποφάσεων. Τέτοια μοντέλα χρησιμοποιούνται συνήθως για εργασίες παλινδρόμησης και ταξινόμησης που αποδεικνύουν τη χρησιμότητα τους στη διαχείριση των καλλιεργειών και τον εντοπισμό των ζιζανίων, ασθενειών ή συγκεκριμένων χαρακτηριστικών. Η πρόσφατη εξέλιξη των ANN στο deep learning που επεκτείνεται στο πεδίο της εφαρμογής αυτών σε όλους τους τομείς περιλαμβάνει και την γεωργία ακριβείας.

Τα SVM είναι δυαδικοί ταξινομητές που κατασκευάζουν ένα γραμμικό διαχωριστικό υπερπλάνο για να ταξινομήσουν δεδομένα. Στη γεωργία χρησιμοποιούνται για να προβλέψουν την απόδοση και την ποιότητα των καλλιεργειών καθώς και την κτηνοτροφία.

Αν και ακόμη στην αρχή της πορείας του, τα αγροκτήματα που διαχειρίζονται από μηχανική μάθηση εξελίσσονται ήδη σε συστήματα τεχνητής νοημοσύνης. Προς το παρόν οι λύσεις μηχανικής μάθησης αντιμετωπίζουν μεμονωμένα προβλήματα αλλά με ευρύτερη ενσωμάτωση της αυτόματης καταγραφής και ανάλυσης δεδομένων, μηχανικής μάθησης και λήψης αποφάσεων σε ένα σύστημα που συνδέονται όλα αυτά. Οι γεωργικές πρακτικές θα αλλάξουν με την λεγόμενη γεωργία που βασίζεται στη γνώση και θα είναι σε θέση να αυξήσει τα επίπεδα παραγωγής και την ποιότητα των προϊόντων.

Αξιοσημείωτα είναι όμως και τα κανούργια μοντέλα που έχουν αναπτυχθεί και εξελίσσονται ακόμα. Τα συστήματα μηχανικής όρασης χρησιμοποιούνται ήδη ευρέως σε διαφορετικά τμήματα της γεωργικής παραγωγής και της βιομηχανικής παραγωγής τροφίμων. Η χρήση του είναι δικαιολογημένη λόγω τα οφέλη που αποκτήθηκαν. Η χρήση τέτοιων συστημάτων παρέχει μια απλή και αντικειμενική ανάλυση των δειγμάτων, παράγοντας ακριβή περιγραφικά δεδομένα. Μέσω αυτών των συστημάτων είναι δυνατό να αυτοματοποιηθούν επίπονες έρευνες παράγοντας επαρκή δεδομένα για μελλοντική ανάλυση. Διαπιστώθηκε ότι υπάρχουν κενά που πρέπει να καλυφθούν με την ανάπτυξη έξυπνων συσκευών που χρησιμοποιούν μηχανική όραση και τεχνητή νοημοσύνη για αυτοματοποίηση εργασιών στον τομέα, καθώς και την ενσωμάτωσή τους με γεωργικές μηχανές και μη επανδρωμένα αεροσκάφη. Για την μελλοντική εργασία ως εναλλακτικές λύσεις προτείνονται προηγμένες τεχνικές τεχνητής νοημοσύνης, καθώς υπόσχονται πολλά. Διάφορες καλλιέργειες θα μπορούσαν να επωφεληθούν από ένα σύστημα υπολογιστών που στοχεύει στη μείωση της πολυπλοκότητας και το κόστος στην ταξινόμηση των χαρακτηριστικών τους. Η μελέτη των ερευνών που γίνονται συνέχεια και τα προβλήματα που πρέπει να λυθούν μαζί με τις νέες εξελίξεις στην μηχανική όραση και την τεχνητή νοημοσύνη μπορούν να οδηγήσουν σε νέες λύσεις για τη γεωργία φέρνοντας κέρδη στην παραγωγή, την ποιότητα και την ασφάλεια των τροφίμων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] F. Wortmann and K. Flüchter, “Internet of Things: Technology and Value Added,” *Bus. Inf. Syst. Eng.*, vol. 57, no. 3, pp. 221–224, 2015, doi: 10.1007/s12599-015-0383-3.
- [2] Σ. Φουντάς and Θ. Γέμτος, *Γεωργία Ακριβείας*. 2015.
- [3] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, “Machine Learning Applications for Precision Agriculture: A Comprehensive Review,” *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.
- [4] K. Thomson and R. J. Mcqueen, “Machine Learning Applied to Fourteen Agricultural Datasets Table of Contents,” *Mach. Learn.*, 1996.
- [5] P. Toscano, A. Castrignanò, S. F. Di Gennaro, A. V. Vonella, D. Ventrella, and A. Matese, “A precision agriculture approach for durum wheat yield assessment using remote sensing data and yield mapping,” *Agronomy*, vol. 9, no. 8, 2019, doi: 10.3390/agronomy9080437.
- [6] N. J. Miller, T. W. Griffin, I. A. Ciampitti, and A. Sharda, “Farm adoption of embodied knowledge and information intensive precision agriculture technology bundles,” *Precis. Agric.*, vol. 20, no. 2, pp. 348–361, 2019, doi: 10.1007/s11119-018-9611-4.
- [7] “Department of Agrosystems and Bioclimatology, Mendel University in Brno, Czech Republic Agriculture Research Ltd., Troubsko, Czech Republic,” p. 5593.
- [8] Y. Mekonnen, S. Namuduri, L. Burton, A. Sarwat, and S. Bhansali, “Review—Machine Learning Techniques in Wireless Sensor Network Based Precision Agriculture,” *J. Electrochem. Soc.*, vol. 167, no. 3, p. 037522, 2020, doi: 10.1149/2.0222003jes.
- [9] K. Haseeb, N. Islam, A. Almogren, and I. Ud Din, “Intrusion prevention framework for secure routing in WSN-Based mobile internet of things,” *IEEE Access*, vol. 7, pp. 185496–185505, 2019, doi: 10.1109/ACCESS.2019.2960633.
- [10] I. Workshop, F. P. O. Project, and D. Hutchison, *and Open-Source Solutions*. 2014.
- [11] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, “Machine learning in agriculture: A review,” *Sensors (Switzerland)*, vol. 18, no. 8, pp. 1–29, 2018, doi: 10.3390/s18082674.
- [12] Bharadwaj, K. B. Prakash, and G. R. Kanagachidambaresan, *Pattern Recognition and Machine Learning*. 2021.

- [13] F. Balducci, D. Impedovo, and G. Pirlo, "Machine learning applications on agricultural datasets for smart farm enhancement," *Machines*, vol. 6, no. 3, 2018, doi: 10.3390/machines6030038.
- [14] W. Tae Kim and Y. J. Park, "Batalin-Tyutin quantization of the (2+1) dimensional nonabelian Chern-Simons field theory," *Phys. Lett. B*, vol. 336, no. 3–4, pp. 376–380, 1994, doi: 10.1016/0370-2693(94)90548-7.
- [15] H. Yao, R. Qin, and X. Chen, "Unmanned aerial vehicle for remote sensing applications - A review," *Remote Sens.*, vol. 11, no. 12, pp. 1–22, 2019, doi: 10.3390/rs11121443.
- [16] S. Y. Liu, "Artificial Intelligence (AI) in Agriculture," *IT Prof.*, vol. 22, no. 3, pp. 14–15, 2020, doi: 10.1109/MITP.2020.2986121.
- [17] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019, doi: 10.1109/ACCESS.2019.2912200.
- [18] J. Padarian, B. Minasny, and A. B. McBratney, "Machine learning and soil sciences: a review aided by machine learning tools," *SOIL*, vol. 6, no. 1, pp. 35–52, Feb. 2020, doi: 10.5194/soil-6-35-2020.
- [19] C. W. Brungard, J. L. Boettinger, M. C. Duniway, S. A. Wills, and T. C. Edwards, "Machine learning for predicting soil classes in three semi-arid landscapes," *Geoderma*, vol. 239, pp. 68–83, 2015, doi: 10.1016/j.geoderma.2014.09.019.
- [20] A. U. Asuncion, P. Smyth, and M. Welling, "Asynchronous distributed estimation of topic models for document analysis," *Stat. Methodol.*, vol. 8, no. 1, pp. 3–17, 2011, doi: 10.1016/j.stamet.2010.03.002.
- [21] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Comput. Electron. Agric.*, vol. 153, no. April, pp. 69–81, 2018, doi: 10.1016/j.compag.2018.08.001.
- [22] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification," *Comput. Intell. Neurosci.*, vol. 2016, 2016, doi: 10.1155/2016/3289801.
- [23] F. Nicolson, M. F. Kircher, N. Stone, and P. Matousek, "Spatially offset Raman spectroscopy for biomedical applications," *Chem. Soc. Rev.*, vol. 50, no. 1, pp. 556–568, 2021, doi: 10.1039/d0cs00855a.
- [24] I. A. and O. L. N., "Digital Image Processing for Detecting and Classifying Plant Diseases," *Circ. Comput. Sci.*, vol. 2, no. 11, pp. 1–7, 2017, doi: 10.22632/ccs-2017-252-66.
- [25] M. K. Tripathi and D. D. Maktedar, "A role of computer vision in fruits and vegetables among various horticulture products of agriculture fields: A

- survey,” *Inf. Process. Agric.*, vol. 7, no. 2, pp. 183–203, 2020, doi: 10.1016/j.inpa.2019.07.003.
- [26] Y. Yang, I. G. Morillo, and T. M. Hospedales, “Deep Neural Decision Trees,” no. Whi, 2018, [Online]. Available: <http://arxiv.org/abs/1806.06988>.
- [27] A. Bansal, M. Sharma, and S. Goel, “Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining,” *Int. J. Comput. Appl.*, vol. 157, no. 6, pp. 35–40, 2017, doi: 10.5120/ijca2017912719.
- [28] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.
- [29] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, “An improved method to construct basic probability assignment based on the confusion matrix for classification problem,” *Inf. Sci. (Ny)*, vol. 340–341, pp. 250–261, 2016, doi: 10.1016/j.ins.2016.01.033.
- [30] J. G. Richens, C. M. Lee, and S. Johri, “Improving the accuracy of medical diagnosis with causal machine learning,” *Nat. Commun.*, vol. 11, no. 1, pp. 1–9, 2020, doi: 10.1038/s41467-020-17419-7.
- [31] S. Figini and P. Giudici, “Statistical merging of rating models,” *J. Oper. Res. Soc.*, vol. 62, no. 6, pp. 1067–1074, 2011, doi: 10.1057/jors.2010.41.
- [32] R. O. Characteristics, “10 Εναλλακτικές Μέθοδοι Και Ειδικά Θέματα Κατηγοριοποίησης,” 2011.
- [33] W. Explorer, W. Explorer, and M. Perceptron, “13 Οδηγός WEKA.”

Ηλεκτρονικές πηγές

- Jason Brownlee (May, 2020), “Statistical Imputation for Missing Values in Machine Learning”, *Machine Learning Mastery, Making Developers Awesome At Machine Learning*, <https://machinelearningmastery.com/statistical-imputation-for-missing-values-in-machine-learning/>
- Christopher M. Bishop Pattern Recognition and Machine Learning Bishop-
Pattern-Recognition-And-Machine-Learning-Springer-2006.pdf
- Analytics Vidhya, Weka – GUI way to learn Machine Learning, <https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/>

- De Gruyter (2019), “Calibration of near infrared spectroscopy (NIRS) data of three Eucalyptus species with extractive contents determines by ASE extraction for rapid identification of species and high extractive contents”, *Holzforschung Journal*, <https://www.degruyter.com/document/doi/10.1515/hf-2018-0166/html>
- José Padarian, Budiman Minasny, and Alex B. McBratney, “Machine Learning and soil sciences: a review aided by machine learning tools”, European Geosciences Union, *Articles Volume 6, issue 1, SOIL*, 6, 35–52, 2020, <https://soil.copernicus.org/articles/6/35/2020/>

Πηγές εικόνων

- <https://www.analyticsvidhya.com/blog/2020/11/artificial-intelligence-in-agriculture-using-modern-day-ai-to-solve-traditional-farming-problems/>
- <https://www.xyonix.com/crop-monitoring>
- <https://www.agrifac.com/ca/sustainable-farming/spot-spraying/>
- <https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning>
- <https://www.educba.com/classification-algorithms/>
- <https://i.ibb.co/7gHWhyC/0-Yw-Jn-KL-s-QUIVy-IPW.png>
- <https://i.ibb.co/4t20nPQ/0-y-Ob-Hm6-MUDw-ZTz-Lqu.png>
- <https://ospreydata.com/wp-content/uploads/AI-vs-ML-vs-Deep-Learning-1024x431.png>
- <https://ospreydata.com/ai-ml-models-101-what-is-a-model/>
- <http://breakthrough.unglobalcompact.org/disruptive-technologies/digital-agriculture/>
- <https://www.instructables.com/How-to-Detect-Plant-Diseases-Using-Machine-Learnin/>

