ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΒΙΟΙΑΤΡΙΚΗ

ΕΝΕΡΓΗ ΜΑΘΗΣΗ, ΕΚΠΑΙΔΕΥΣΗ ΣΕ ΜΕΡΙΚΩΣ ΕΠΙΣΗΜΕΙΩΜΕΝΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΒΙΟΪΑΤΡΙΚΩΝ ΕΙΚΟΝΩΝ

ΓΕΩΡΓΙΟΣ ΣΤΑΥΡΟΠΟΥΛΟΣ
ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ: 687

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Μιχάλης Σαβελώνας
Επίκουρος Καθηγητής

Λαμία, Οκτώβριος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΒΙΟΙΑΤΡΙΚΗ

ΕΝΕΡΓΗ ΜΑΘΗΣΗ, ΕΚΠΑΙΔΕΥΣΗ ΣΕ ΜΕΡΙΚΩΣ ΕΠΙΣΗΜΕΙΩΜΕΝΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΒΙΟΪΑΤΡΙΚΩΝ ΕΙΚΟΝΩΝ

ΓΕΩΡΓΙΟΣ ΣΤΑΥΡΟΠΟΥΛΟΣ
ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ: 687

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Μιχάλης Σαβελώνας
Επίκουρος Καθηγητής

Λαμία, Οκτώβριος 2022

UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
AND
BIOMEDICAL INFORMATICS

ACTIVE LEARNING, TRAINING ON PARTIALLY LABELED DATA AND BIOMEDICAL IMAGE CLASSIFICATION

GEORGIOS STAVROPOULOS

STUDENT REGISTRATION NUMBER: 687

FINAL THESIS
Supervisor

Michalis Savelonas

Assistant Professor

Lamia, October 2022

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

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

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

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

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

Ημερομηνία: 5 / .Οκτωβρίου / 2022

–Ο –
Δηλών.

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

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Δρ. Μιχάλη, Σαβελώνα, επίκουρο καθηγητή, για την καθοδήγηση που μου προσέφερε και το χρόνο που διέθεσε δίνοντάς μου χρήσιμες συμβουλές και οδηγίες για την ολοκλήρωση της πτυχιακής μου εργασίας. Ιδιαίτερα τον ευχαριστώ για την άμεση και ενεργή ανταπόκρισή του σε κάθε αίτημα μου.

Στο ίδιο πλαίσιο ευγνωμοσύνης, θα ήθελα να ευχαριστήσω όλους τους καθηγητές του τμήματος Πληροφορικής με εφαρμογές στην Βιοϊατρική για τη συμβολή τους στην επιστημονική και τεχνολογική μου συγκρότηση στα χρόνια της φοίτησής μου στο Τμήμα.

Τέλος, ένα μεγάλο ευχαριστώ στους γονείς μου για την οικονομική τους υποστήριξη, καθώς και τους συγγενείς και τους φίλους για την ηθική υποστήριξη σε όλο το διάστημα των σπουδών μου.

ΠΕΡΙΛΗΨΗ

Η διαπίστωση ότι τα ιατρικά test (PCR, rapid και self) δίνουν μεγάλο ποσοστό ψευδώς αρνητικών, σε συνδυασμό με την αδήριτη ανάγκη για περιορισμό της μεταδοτικότητας της COVID-19 οδήγησε την επιστημονική κοινότητα στην αναζήτηση συμπληρωματικών μεθόδων διάγνωσης και ανίχνευσης της νόσου. Η εξέτασή ακτινογραφιών θώρακα, λόγω κυρίως του χαμηλού κόστους και της ευρείας διαθεσιμότητας, είναι η κύρια υποψήφια για τον ρόλο μιας συμπληρωματικής ή εναλλακτικής μεθόδου διάγνωσης. Οι ελλείψεις σε υγειονομικό προσωπικό που δημιούργησε η ίδια η πανδημία, μαζί με την δυσκολία ανίχνευσης της νόσου από ακτινογραφίες με γυμνό οφθαλμό, καθιστά επιτακτική την ανάγκη η μέθοδος αυτή να είναι αυτοματοποιημένη. Η εποπτευόμενη μηχανική μάθηση, η οποία βασίζεται στην ύπαρξη ενός μεγάλου αριθμού επισημειωμένων δεδομένων, είναι προφανώς μία λύση. Το πρόβλημα όμως που αντιμετωπίζουμε συχνά είναι ότι δεν υπάρχουν αρκετές σε αριθμό επισημειωμένες ακτινογραφίες, είτε λόγω αντικειμενικής αδυναμίας είτε λόγω του κόστους επισημείωσης, ώστε να εκπαιδεύσουμε επαρκώς ένα μοντέλο πλήρως εποπτευόμενης μάθησης. Πολλές φορές μάλιστα, έχουμε μόνο θετικά επισημειωμένα δεδομένα. Οι λόγοι αυτοί είναι που μας οδήγησαν στο να μελετήσουμε μεθόδους μηχανικής μάθησης που μπορούν να λειτουργήσουν με μερικώς επισημειωμένα δεδομένα, να τις συγκρίνουμε μεταξύ τους, να συγκρίνουμε τους εκτιμητές (αλγορίθμους βάσης) που χρησιμοποιούν, να διερευνήσουμε τους περιορισμούς και τα πλεονεκτήματα. Εξετάσθηκαν όλοι οι γνωστοί μέθοδοι ημι-εποπτευόμενης μάθησης (self-training, label propagation, label spreading), η ενεργή μάθηση με έναν εκτιμητή και με επιτροπή εκτιμητών, με διάφορες παραλλαγές όσον αφορά τις στρατηγικές ερωτήσεων, καθώς επίσης εξετάσθηκαν και αλγόριθμοι που αφορούν σε μάθηση από θετικά και μη επισημειωμένα δεδομένα. Τα αποτελέσματα είναι ενθαρρυντικά, αφού αποδεικνύεται ότι δεν χρειάζεται μεγάλη βάση πραγματικά επισημειωμένων ακτινογραφιών για να μπορούμε να δημιουργήσουμε ένα αξιόπιστο μοντέλο. Στην περίπτωση της ημι-εποπτευόμενης μάθησης και της μάθησης από θετικά και μη επισημειωμένα δεδομένα αρκεί να έχουμε ένα μικρό αριθμό με πραγματικά επισημειωμένες εικόνες τις οποίες μπορούμε να συμπληρώσουμε με επισημειωμένες με αυτόματο τρόπο, ώστε να έχουμε ένα αξιόπιστο μοντέλο προβλέψεων. Στην δε ενεργή μάθηση με την χρήση καταλλήλων στρατηγικών μπορούμε να ζητήσουμε να επισημειωθεί μόνο ένας μικρός αλλά κατάλληλα επιλεγμένος αριθμός ακτινογραφιών ώστε να κτίσουμε ένα αξιόπιστο μοντέλο πρόβλεψης με πολύ λιγότερα επισημειωμένα δεδομένα από ότι θα χρειαζόταν ένα μοντέλο πλήρως εποπτευόμενης μάθησης.

ABSTRACT

The realization that medical tests (PCR, rabid, and self) give a high false negative rate, combined with the urgent need to limit the transmissibility of COVID-19, led the scientific community to search for additional methods of diagnosis and detection of the disease. Chest X-ray examination, mainly due to its low cost and wide availability, is the primary candidate for the role of a complementary or alternative diagnostic method. The shortages of health personnel created by the pandemic itself, together with the difficulty of detecting the disease from naked-eye X-rays, make it imperative that this method be automated. Supervised machine learning, which relies on the existence of a large amount of labeled data, is obviously one solution. However, the problem we often face is that there are not enough labeled radiographs, either due to objective impossibility or labeling costs, to train a fully supervised learning model adequately. In fact, many times, we have only positively labeled data. These are the reasons that led us to study machine learning methods that can work with partially labeled data, compare them, compare the estimators (base algorithms) they use, and investigate their limitations and advantages. However, the problem we often face is that there are not enough labeled radiographs, either due to objective impossibility or labeling costs, to train a fully supervised learning model adequately. In fact, many times, we have only positively labeled data. These reasons led us to study machine learning methods that can cope with partially labeled data, compare them, compare the estimators (base algorithms) they use, and investigate their limitations and advantages. All known semi-supervised learning methods (self-training, label propagation, label spreading), active learning with one estimator, and with a committee of estimators with variations in terms of questioning strategies were considered, as well as algorithms related to learning algorithms from positive and unlabeled data were examined. The results are encouraging, as it turns out that we do not need a large base of truly labeled radiographs to build a reliable model. In the case of semi-supervised learning and learning from positive and unlabeled data, it is sufficient to have a small number of truly labeled images which we can supplement with automatically labeled images to obtain a reliable prediction model. In active learning, using appropriate strategies, we can have labeled only a small but appropriately selected number of radiographs to build a reliable prediction model with much less labeled data than a fully supervised learning model would need.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΚΕΦΑΛΑΙΟ 1 ΕΙΣΑΓΩΓΗ

(Υποκεφάλαιο 1.1) Ιστορική αναδρομή της εξέλιξης της τεχνητής νοημοσύνης

Η τεχνητή νοημοσύνη, δηλαδή η κατασκευή μιας μηχανής που σκέφτεται, πάντα ασκούσε γοητεία στους ανθρώπους. Μετά την κατασκευή των πρώτων ηλεκτρονικών υπολογιστών (ENIAC και MANIAC) κατά την διάρκεια του ΙΙ Παγκοσμίου πολέμου, ορισμένοι ερευνητές άρχισαν να σκέφτονται σοβαρά την ιδέα μιας μηχανής που σκέφτεται, χωρίς κανένας τους όμως να καθορίζει με σαφήνεια τι σημαίνει αυτό . Το 1950 ο Άγγλος μαθηματικός Turing, μετά από πειράματα που έκανε κατέληξε στο συμπέρασμα ότι για να θεωρηθεί ότι μια μηχανή σκέφτεται, απαιτεί περισσότερα πράγματα από γνώσεις και συλλογιστική ικανότητα. «Η τεχνητή νοημοσύνη είναι η ικανότητα των μηχανών να εκτελούν ορισμένες εργασίες, οι οποίες χρειάζονται τη νοημοσύνη που επιδεικνύουν άνθρωποι. Αυτός ο ορισμός αποδίδεται στους Marvin Minsky and John McCarthy από τη δεκαετία του 1950, οι οποίοι γνωστοί ως οι πατέρες της τεχνητής νοημοσύνης ». (What is Artificial Inteligence, n.d.).

Την δεκαετία του 1980 γράφτηκαν πολλά βιβλία για την τεχνητή νοημοσύνη και κυρίως για τις προσδοκίες που υπήρχαν κυρίως σε 3 τομείς: **έμπειρα συστήματα, επεξεργασία φυσικής γλώσσας, υπολογιστική όραση**. Τα επόμενα 20 χρόνια, παρά τις προσπάθειες που έγιναν, δεν υπήρξαν σημαντικά αποτελέσματα.

(Υποκεφάλαιο 1.2) Η κατάσταση σήμερα

Σήμερα, η τεχνολογία της τεχνητής νοημοσύνης είναι ένας κρίσιμος άξονας μεγάλου μέρους του ψηφιακού μετασχηματισμού που λαμβάνει χώρα σε όλο τον κόσμο, καθώς οι οργανισμοί, οι εταίρες και διάφοροι ερευνητικοί φορείς προσπαθούν να αξιοποιήσουν τον συνεχώς αυξανόμενο όγκο δεδομένων που παράγονται και συλλέγονται.

Η αλλαγή αυτή οφείλεται:

- Στην δημιουργία τεράστιων αποθηκών ψηφιακών δεδομένων (Big Data), που αρχικά έγιναν με μοναδικό σκοπό την μηχανογράφηση των χειρόγραφων διαδικασιών και ιδιαίτερα την ψηφιοποίηση των αρχείων.
- Στην εκθετική αύξηση της επεξεργαστικής δύναμης των υπολογιστών που κατέστησε δυνατή την μαζική επεξεργασία τεράστιου όγκου δεδομένων

- Στην δημιουργία νέων «έξυπνων» αλγορίθμων που είναι αποτελεσματικότεροι από πλευράς ταχύτητας και αξιοπιστίας των αποτελεσμάτων.

Αυτό το αυξημένο ενδιαφέρον, στον ακαδημαϊκό χώρο, στη βιομηχανία και στην κοινότητα ανοιχτού κώδικα, έχει οδηγήσει σε σημαντική πρόοδο σε όλους τους τομείς. Από τα οικονομικά θέματα, τις πολεμικές επιχειρήσεις, την ανίχνευση κυβερνοεπιθέσεων, τα αυτό-οδηγούμενα αυτοκίνητα ακόμη και μέχρι την πρόβλεψη της έκβασης των νομικών υποθέσεων ή αποτελεσμάτων αθλητικών αγώνων. Ιδιαίτερα στον ιατρικό τομέα η δυνατότητα προβλέψεων μετά από εκπαίδευση διαφόρων μοντέλων με βάση δεδομένα που έχουν συλλεχθεί στο παρελθόν, έχει δώσει νέες σημαντικές δυνατότητες στην έρευνα και εφαρμογή της τεχνητής νοημοσύνης. Γενικά εφαρμόζεται ο κύκλος: Συλλογή – Προεπεξεργασία -Εκπαίδευση – Αξιολόγηση -Προβλέψεις.

(Υποκεφάλαιο 1.3) Επάρκεια και ποιότητα δεδομένων

Οι δυνατότητες που προσφέρθηκαν τα τελευταία χρόνια με την αύξηση της διαθεσιμότητας δεδομένων σε μεγάλες ποσότητες και με την τεράστια ισχύ των σύγχρονων υπολογιστών δεν είναι τελικά πανάκεια. Δεν αρκεί να έχουμε ένα τεράστιο αριθμό ακατέργαστων δεδομένων για να δημιουργήσουμε μια μηχανή που με αυτοματοποιημένο τρόπο θα προβλέπει και θα διαγιγνώσκει με βάση την εκπαίδευση που έχει υποστεί σε παλαιότερα δεδομένα. Αυτά τα εκπαιδευτικά δεδομένα πρέπει να είναι επισημειωμένα, δηλαδή να έχει διαπιστωθεί με ανθρώπινη εργασία σε ποια κλάση ανήκουν. Ένα άλλο πρόβλημα είναι ότι τα επισημειωμένα αυτά δεδομένα πρέπει να είναι διαθέσιμα σε μεγάλες ποσότητες, ώστε ένας αλγόριθμος εκπαίδευσης να μπορεί να αποκαλύψει την λανθάνουσα σχέση μεταξύ των χαρακτηριστικών των δεδομένων μας και των επισημειώσεων τους. Ορισμένες φορές οι μεγάλες ποσότητες δειγμάτων δεν επαρκούν για την εκπαίδευση των μοντέλων της τεχνητής νοημοσύνης, εάν αυτά δεν έχουν ορισμένα χαρακτηριστικά, όπως πχ να είναι τυχαία και αντιπροσωπευτικά. Η τροφοδότηση ενός συστήματος τεχνητής νοημοσύνης με τεράστιες αλλά ακατάλληλες ποσότητες δεδομένων, εκτός από τον μεγάλο χρόνο επεξεργασίας και την τεράστια ποσότητα επεξεργαστικής ισχύος που δαπανά, μπορεί να δώσει και άστοχες προβλέψεις. Το πρόβλημα αυτό μαζί με το γεγονός ότι η χειροκίνητη επισημείωση των δεδομένων κοστίζει, έχει οδηγήσει σημαντικό τμήμα της ερευνητικής κοινότητας στην αναζήτηση μεθόδων επιλογής δεδομένων προς επισημείωση κατά τρόπο ώστε

είτε η διαδικασία των επισημειώσεων να γίνεται αυτόματα είτε να σταλούν προς χειροκίνητη επισημείωση τα πλέον «κατάλληλα», ώστε η λανθάνουσα συνάρτηση που συνδέει τα δεδομένα με την κατηγορία στην οποία ανήκουν να μπορεί να αποκαλυφθεί με τις λιγότερες δυνατές επισημειώσεις.

(Υποκεφάλαιο 1.4) Μέθοδος προσέγγισης

(Ενότητα 1.4.α) Σκοπός

Ο σκοπός της παρούσας εργασίας είναι να διερευνηθούν οι δυνατότητες των αλγορίθμων που ασχολούνται με μηχανική μάθηση με μερικώς επισημειωμένα δεδομένα. Αυτό επιδιώκεται να επιτευχθεί με την πιλοτική εφαρμογή διαφόρων τέτοιων αλγορίθμων στην διάγνωση της νόσου COVID-19 από ακτινογραφίες θώρακα.

(Ενότητα 1.4.β) Γενικά.

Η μηχανική μάθηση και η βαθιά μάθηση είναι δύο περιοχές της τεχνητής νοημοσύνης που εφαρμόζονται για αυτοματοποιημένες προβλέψεις σε πάρα πολλούς τομείς, και ειδικότερα στον ιατρικό τομέα, όπου τα τελευταία χρόνια δίνουν εντυπωσιακά αποτελέσματα. Η βασική αρχή στην οποία στηρίζονται αυτά τα συστήματα είναι η εκπαίδευση των μοντέλων με βάση επισημειωμένα δεδομένα, ώστε αυτά να καταστούν ικανά να γενικεύσουν στην ταξινόμηση μη επισημειωμένων δεδομένων, τα οποία είναι διαφορετικά από αυτά με τα οποία εκπαιδεύτηκε το μοντέλο μας.

Το βασικό συστατικό, όμως των μεθόδων αυτών, είναι η ύπαρξη επισημειωμένων δεδομένων, δηλαδή δεδομένων που ήδη έχουν χαρακτηριστεί από ειδικούς και έχουν ταξινομηθεί σε μία κλάση (πχ covid, non covid). Οι λανθάνουσες πληροφορίες της κατανομής τους που μαθαίνονται από το μοντέλο, με την διαδικασία της εκπαίδευσης χρησιμοποιούνται για να βοηθήσουν τον ταξινομητή να κινηθεί προς τη σωστή απόφαση, επιτυγχάνοντας έτσι υψηλότερη γενίκευση και αξιοπιστία στις προβλέψεις.

Το πρόβλημα που αντιμετωπίζουμε συχνά, είναι η έλλειψη αρκετών μη επισημειωμένων δειγμάτων ώστε το μοντέλο μας να μπορεί να εκπαιδευτεί επαρκώς. Η έλλειψη των επισημειωμένων δεδομένων μπορεί να οφείλεται:

α) σε αντικειμενική αδυναμία (πχ δεν είναι πάντα δυνατή η ασφαλής διάγνωση κορονοϊού από μια ακτινογραφία και ιδιαίτερα για τους ασυμπτωματικούς ασθενείς).

β) υπερβολικό κόστος σε χρήμα και χρόνο των επισημειώσεων.

Τέλος συχνά υπάρχουν μερικά θετικά επισημειωμένα δείγματα και υπάρχει μια πληθώρα δεδομένων που δεν γνωρίζουμε εάν είναι θετικά η αρνητικά.

(Ενότητα 1.4.γ) Παραδοχές

Εκτιμάται ότι οι ακτινολογικές εικόνες μεταφέρουν βασικές πληροφορίες για τον COVID-19, επομένως, η αυτοματοποιημένη ανίχνευση πνευμονικών λοιμώξεων με τη βοήθεια τεχνητής νοημοσύνης (AI) μπορεί να χρησιμεύσει ως πιθανό διαγνωστικό εργαλείο.

Στο πείραμα που διεξάγεται στην παρούσα μελέτη το σύνολο των δεδομένων που χρησιμοποιείται είναι «πραγματικά» επισημειωμένα. Αναλόγως της περίπτωσης που εξετάζουμε κάθε φορά θεωρούμε ότι, είτε το σύνολο είτε το μεγαλύτερο τμήμα αυτού, είναι μη επισημειωμένα. Οι τιμές των μετρικών που εξετάζουμε λαμβάνουν ως αληθείς τιμές τις πραγματικές επισημειώσεις. Σε μια πραγματική κατάσταση δεν θα ήταν δυνατό να έχουμε διαθέσιμες τις πραγματικές τιμές, στην περίπτωση όμως αυτή που αξιολογούμε τα διάφορα μοντέλα, για μελλοντική χρήση είναι λογικό να τα αξιολογήσουμε συγκρίνοντας τις προβλέψεις με τις πραγματικές επισημειώσεις. Επίσης, για την διερεύνηση της περίπτωσης της ενεργούς μάθησης (active learning), όπου θεωρητικά απαιτείται η χειροκίνητη επισημείωση, τα στοιχεία επισημειώσεων λαμβάνονται από τις πραγματικές τιμές, που έχουμε διαθέσιμες με πρόγραμμα υπολογιστή που προσομοιώνει τον άνθρωπο επισημειωτή.

(Ενότητα 1.4.δ) Εξαγωγή χαρακτηριστικών

Παραδοσιακά έχουν αναπτυχθεί πολλές μέθοδοι για την εξαγωγή χαρακτηριστικών από εικόνες. Τα τελευταία χρόνια χρησιμοποιούνται τα προεκπαιδευμένα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks-CNN) όπως τα DenseNets, τα οποία επιδεικνύουν μεγάλη αξιοπιστία. Ειδικότερα για την εργασία μας έχει επιλεγεί το DenseNet169 που χρησιμοποιήθηκε και από τους (Hamid Nasiri & Alavi, 2022). Από τα εξαχθέντα χαρακτηριστικά επελέγησαν τα 20 «σχετικότερα» με βάση την μέθοδο ANOVA.

(Ενότητα 1.4.ε) Εκπαίδευση

Εκπαίδευση είναι η διαδικασία με την οποία ένας αλγόριθμος λαμβάνοντας υπόψιν διαθέσιμα δείγματα δεδομένων με τις επισημειώσεις τους καιπαράγει ένα μοντέλο το οποίο μπορεί να κάνει προβλέψεις για την κλάση στην οποία ανήκουν παρεμφερή αλλά άγνωστα δείγματα.

Ημι-επόπτευόμενη μάθηση (semi-supervised)

Σε αυτήν την περίπτωση έχουμε λίγα επισημειωμένα δείγματα, και πολλά για τα οποία δεν γνωρίζουμε σε ποια κλάση ανήκουν. Ο αριθμός των επισημειωμένων δειγμάτων δεν επαρκεί ώστε να παραχθεί ένα αξιόπιστο μοντέλο εποπτευόμενης μάθησης. Η προσπάθειά σε αυτήν την περίπτωση έγκειται στο να συμπληρώσουμε τα ήδη επισημειωμένα, επισημειώνοντας με αυτόματο τρόπο ορισμένα ή όλα τα μη επισημειωμένα. Εξετάσθηκαν οι αλγόριθμοι self-training, label-propagation και label-spreading.

Ενεργή μάθηση (active learning)

Και στην περίπτωση αυτήν δεν έχουμε αρκετά επισημειωμένα δεδομένα για να δημιουργήσουμε ένα αξιόπιστο μοντέλο πρόβλεψης. Η επισημείωση των δεδομένων που λείπουν όμως, γίνεται χειροκίνητα, δηλαδή έχει κόστος. Τα δεδομένα που στέλνονται προς επισημείωση επιλέγονται βάση μιας στρατηγικής ερωτήσεων ούτως ώστε να επιλέγονται κάθε φορά αυτά που μπορούν να εισφέρουν περισσότερο στην κατασκευή του μοντέλου.

Μάθηση από θετικά και μη επισημειωμένα δεδομένα

Αυτή είναι μια ιδιαίτερη περίπτωση που έχουμε μόνο θετικά επισημειωμένα δεδομένα και ταυτόχρονα έχουμε πολλά μη επισημειωμένα τα οποία μπορεί να είναι είτε θετικά είτ αρνητικά. Όπως και στην ημι-εποπτευόμενη μάθηση και εδώ με την χρήση των κατάλληλων αλγορίθμων επισημειώσουμε αυτόματα τα μη επισημειωμένα δεδομένα. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι οι Elkanoto, WeightedEkanoto και Bagging.

(Ενότητα 1.4.στ) Μέθοδος πιστοποίησης

Η πιστοποίηση έγινε με σύγκριση των επισημειώσεων που προέβλεψε το μοντέλο μας με τις πραγματικές επισημειώσεις των δεδομένων μας.

Έχουν χρησιμοποιηθεί διάφοροι μέθοδοι πιστοποίησης αναλόγως της περίπτωσης:

1) Εκπαίδευση με το σύνολο των διαθεσίμων επισημειωμένων και μη επισημειωμένων δεδομένων από τους αλγορίθμους ημι-εποπτευόμενης μάθησης ή από τους αλγορίθμους θετικών και μη επισημειωμένων και πιστοποίηση με το σύνολο δεδομένων που έχουμε στην διάθεσή μας (βάση δεδομένων που προήλθε από την εξαγωγή των χαρακτηριστικών). Στην ενεργή μάθηση η εκπαίδευση γίνεται με βάση τα ήδη επισημειωθέντα.

2) Διαχωρισμός δεδομένων σε εκπαιδευτικά και δοκιμών σε αναλογία 80 -20. Εδώ η εκπαίδευση γίνεται με τμήμα των δεδομένων και η δοκιμή με τα υπόλοιπα.

3) Εφαρμογή της διασταυρούμενης επικύρωσης (cross validation) με αυτήν έχουμε την δυνατότητα α) να πάρουμε τον μέσο όρο πολλών προβλέψεων β) να υπολογίσουμε την τυπική απόκλιση (standard deviation) και γ) να υπολογίσουμε την αξιολόγηση της εκπαίδευσης επιπλέον από την αξιολόγηση των δοκιμών. .

(Ενότητα 1.4.ζ) Μέθοδος παρουσίασης εργασίας

Παρουσιάζονται στην αρχή βασικές έννοιες, από την αναγνώριση προτύπων, τη μηχανική μάθηση και τη βαθιά μάθηση, γνώσεις που χρειάζονται για την εκπόνηση της εργασίας.

Οι μέθοδοι μάθησης με μερικώς επισημειωμένα δεδομένα χρησιμοποιούν διαφόρους αλγορίθμους πλήρως εποπτευόμενης μάθησης ως βάση καθώς και διάφορες στρατηγικές για να υπολογίσουν τις τιμές που χρειάζονται είτε για να ψευδοεπισημειώσουν είτε να επιλέξουν τα καταλληλότερα δείγματα προς επισημείωση. Έχουν δοκιμαστεί τα αντιπροσωπευτικότερα από αυτά και η παρουσίαση των αποτελεσμάτων γίνεται υπό μορφή διαγραμμάτων, ιστογραμμάτων, πινάκων και confusion matrix ώστε να μπορούν να γίνουν άμεσα οι συγκρίσεις και η εξαγωγή συμπερασμάτων.

ΚΕΦΑΛΑΙΟ 2 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΙΣΗ

(Υποκεφάλαιο 2.1) Εργασίες που έχουν γίνει στον τομέα της διάγνωσης του κορονοϊού με πλήρως εποπτευόμενη μάθηση

Παρακάτω παρουσιάζονται αρκετές από τις δημοσιεύσεις που έγιναν και αφορούν εργασίες στον τομέα της διάγνωσης του COVID-19 με μεθόδους τεχνητής νοημοσύνης και αφορούν στην εφαρμογή πλήρως εποπτευόμενης μάθησης. Οι περισσότερες από τις παρακάτω έχουν αντληθεί από την αναζήτηση που έχουν κάνει οι (Hasani & Nasiri, 2021), που συμπληρώθηκαν από άλλες με αναζήτηση στο διαδίκτυο και με δικές μου σημειώσεις.

COVIDNet-CT: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images

(Gunraj, Wang, & Wong, 2020).

Ανέπτυξαν ένα μοντέλο στηριγμένο σε βαθιά μάθηση (Deep Learning) για την ανίχνευση του COVID-19 και την κατηγοριοποίηση σε φυσιολογικής, μη-COVID-19 πνευμονίας και COVID-19, με ακρίβεια 92,4 τοις εκατό.

COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks

(Apostolopoulos & Mpesiana, 2020).

Εφάρμοσαν τη μάθηση με μεταφοράς (transfer learning) και χρησιμοποίησαν εικόνες ακτινών X σε τρεις κατηγορίες: Ασθενών με COVID-19, υγιών και ασθενών με πνευμονία για να αναπτύξουν το μοντέλο τους. Εκτελέσθηκαν δοκιμές με διάφορου τύπου νευρωνικά δίκτυα και η μεγαλύτερη ακρίβεια (98.75 %) επιτεύχθηκε με VGG16.

Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost

(Hasani & Nasiri, 2021).

Χρησιμοποίησαν το προεκπαιδευμένο στο ImageNet[1] DenseNet169 για να εξάγουν χαρακτηριστικά από εικόνες ακτινών X και χρησιμοποίησαν το XGBoost για ταξινόμηση. Σε δυο (covid, no findings) και τρεις κατηγορίες

[1] Το σύνολο δεδομένων ImageNet περιέχει 14.197.122 επισημειωμένες εικόνες.

(covid, pneumonia, no findings) πέτυχαν 98,24% σε δυαδική και 89,70% σε 3-κλάση ταξινόμηση, αντίστοιχα.

Hybrid deep-learning and machine-learning models for predicting COVID-19

(Talal S. Qaid, και συν., 2021).

Εφάρμοσαν τεχνικές βαθιάς μάθησης και μάθησης με μεταφορά για την εκπαίδευση μοντέλων για τον εντοπισμό του COVID-19. Ενίσχυσαν τα δεδομένα για να δημιουργήσουν μετασχηματισμένες εκδόσεις εικόνων ακτίνων Χ με COVID-19 (όπως αναστροφή, ελαφριά περιστροφή και προσθήκη μικρής παραμόρφωσης) για να αυξήσουμε τον αριθμό των δειγμάτων. Χρησιμοποιούν προεκπαιδευμένα μοντέλα για εξαγωγή χαρακτηριστικών και ακολούθως τροφοδοτούν ένα μοντέλο μη εποπτευόμενης μάθησης. Για την αρχική εξαγωγή χαρακτηριστικών χρησιμοποίησαν CNN ή VGG16[2]. Η ακρίβεια που επιτεύχθηκε σε δυαδικές ταξινομήσεις ήταν μεγαλύτερη του 98%, ενώ η ταξινόμηση πολλαπλών κλάσεων περίπου 93%.

Realizing an effective COVID-19 diagnosis system based on machine learning and IoT in smart hospital environment

(Abdulkareem, και συν., 2021)

Προτείνουν ένα μοντέλο για τον εντοπισμό περιπτώσεων COVID-19 σε έξυπνα νοσοκομεία που χρησιμοποιούν μηχανική μάθηση και το διαδίκτυο των πραγμάτων. Δοκιμάστηκαν τρία μοντέλα μηχανικής μάθησης, συγκεκριμένα, το απλό Bayes (NB), το Random Forest (RF) και το Support Vector Machine (SVM), με εργαστηριακά σύνολα δεδομένων. Παρουσιάστηκαν διαγνώσεις που βασίζονται σε πρωτότυπα και κανονικοποιημένα σύνολα δεδομένων και άλλες που βασίζονται σε επιλογή χαρακτηριστικών. Το μοντέλο SVM υπερείχε στις περισσότερες περιπτώσεις των άλλων μοντέλων που δοκιμάστηκαν (έως και 95% ακρίβεια).

Realizing an effective COVID-19 diagnosis system based on machine learning and IoT in smart hospital environment

(Chen1 & Rezaei, 2021)

Οι Chen και Rezaei προτείνουν μια μέθοδο για την εξαγωγή 18 διαφορετικών χαρακτηριστικών από εικόνες ακτίνων Χ. Τα ελάχιστα χαρακτηριστικά επιλέγονται χρησιμοποιώντας έναν μεταευρετικό αλγόριθμο που ονομάζεται

[2] Είδος CNN(convolutional neural Network)

«βελτιστοποίηση του Αρχιμήδη» για τη μείωση της πολυπλοκότητας της προσέγγισης. Επιτευχθείσα ακρίβεια 86% συγκρίθηκαν.

Optimal diagnosis of COVID-19 based on convolutional neural network and red Fox optimization algorithm

(Khorami, Babaei, & Azadeh, 2021).

Πρότειναν μια μέθοδο για την εξαγωγή συνδυασμού χαρακτηριστικών σε επίπεδο γκρι (GLCM[3]) και του διακριτού μετασχηματισμού κυματιδίων (DWT[4]) από εικόνες ακτινογραφιών X, ακολουθούμενη από ταξινόμηση χρησιμοποιώντας ένα βελτιωμένο μοντέλο CNN[5], με βάση τον αλγόριθμος βελτιστοποίησης Red Fox. Επιτευχθείσα ακρίβεια 84,56%.

CovidGAN: data augmentation using auxiliary classifier GAN for improved covid-19 detection

(Waheed, και συν., 2020)

Ανέπτυξαν ένα μοντέλο που το ονόμασαν CovidGAN και βασίζεται σε Auxiliary Classifier Generative Adversarial Network (ACGAN[6]) για τη δημιουργία συνθετικών εικόνων ακτίνων X και τη βελτίωση της ακρίβειας της ταξινόμησης COVID-19. Η ταξινόμηση με χρήση μόνο του Συνελικτικού Νευρωνικού Δικτύου (Convolutional Neural Network-CNN) απέδωσε 85% ακρίβεια. Με την προσθήκη συνθετικών εικόνων που παράγονται από το CovidGAN, η ακρίβεια αυξήθηκε στο 95% .

[3] Οι συναρτήσεις GLCM χαρακτηρίζουν την υφή μιας εικόνας υπολογίζοντας πόσο συχνά εμφανίζονται ζεύγη pixel με συγκεκριμένες τιμές και σε μια καθορισμένη χωρική σχέση σε μια εικόνα.

[4] Είναι ένας μετασχηματισμός που αποσυνθέτει ένα δεδομένο σήμα σε έναν αριθμό συνόλων, όπου κάθε σύνολο είναι μια χρονική σειρά συντελεστών που περιγράφουν τη χρονική εξέλιξη του σήματος στην αντίστοιχη ζώνη συχνοτήτων.

[5] Το συνελικτικό νευρωνικό δίκτυο (CNN) είναι ένας τύπος τεχνητού νευρωνικού δικτύου που χρησιμοποιείται κυρίως για την αναγνώριση και επεξεργασία εικόνας, λόγω της ικανότητάς του να αναγνωρίζει μοτίβα στην εικόνα

[6] Τα Generative Adversarial Networks (GAN) είναι μια κατηγορία τεχνικών μηχανικής μάθησης, όπου εκπαιδεύονται ταυτόχρονα ένα μη εποπτευόμενο μοντέλο που δημιουργεί νέα δεδομένα και ένας εποπτευόμενος ταξινομητής

Detection of coronavirus disease (COVID-19) based on deep features and support vector machine

(Kumar, Kumari, Kumar, & Biswas, 2020).

Επινόησαν μια μέθοδο που βασίζεται σε Support Vector Machines (SVM[7]) για την ανίχνευση ατόμων που έχουν μολυνθεί από τον κορονοϊό χρησιμοποιώντας εικόνες ακτίνων X. Το SVM εξετάζεται για αναγνώριση COVID-19 χρησιμοποιώντας τα χαρακτηριστικά που εξάχθηκαν από 13 διαφορετικών μοντέλων CNN. Η υψηλότερη ακρίβεια επιτεύχθηκε από τα χαρακτηριστικά που εξάχθηκαν με την βοήθεια του ResNet50 είναι 98,66.

Deep ensemble model for classification of novel coronavirus in chest X-ray images

(Ahmad, Farooq, & Ghani, 2020).

Εφάρμοσαν βαθιά μοντέλα συνελικτικών δικτύων (Convolutional Neural Networks- CNN) CNN, συγκεκριμένα τα MobileNet [8], ResNet50 [9] και InceptionV3[10], με διαφορετικές παραλλαγές, συμπεριλαμβανομένης της εκπαίδευσης του μοντέλου από την αρχή, μαζί με την προσαρμογή των μαθησιακών βαρών όλων των επιπέδων. Από αυτά, δύο μοντέλα με τις καλύτερες επιδόσεις (MobileNet και InceptionV3) επιλέχθηκαν και παρήγαγαν ακρίβεια και F1-Score 95,18% και 90,34% και 95,75% και 91,47%, αντίστοιχα. Το προτεινόμενο μοντέλο υβριδικού συνόλου που δημιουργήθηκε με τη συγχώνευση αυτών των μοντέλων παρήγαγε ακρίβεια ταξινόμησης και F1-Score 96,49% και 92,97%.

Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network

(Abbas, Abdelsamea, & Gaber, 2020)

Επαλήθευσαν ένα συνελικτικό νευρωνικό δίκτυο (CNN) που ονομάζεται Decompose transfer and Compose (DeTraC) για την ταξινόμηση εικόνων ακτίνων X θώρακα. Χρησιμοποιήθηκε μάθηση με μεταφορά ως εξής: Δοκιμάστηκαν προεκπαιδευμένα μοντέλα για την εξαγωγή χαρακτηριστικών πριν το

[7] Μια μηχανή διανυσμάτων υποστήριξης (SVM) είναι ένας τύπος αλγόριθμου βαθιάς μάθησης που εκτελεί εποπτευόμενη μάθηση για ταξινόμηση ή παλινδρόμηση ομάδων δεδομένων.

[8] Είναι μια αρχιτεκτονική CNN που είναι πολύ πιο γρήγορη καθώς και ένα μικρότερο μοντέλο που χρησιμοποιεί ένα νέο είδος συνελικτικού επιπέδου, γνωστό ως Depthwise Separable convolution.

[9] Συνελικτικό Νευρωνικό Δίκτυο με 50 επίπεδα.

[10] Είναι η τρίτη έκδοση του Inception Convolutional Neural Network της Google.

τελευταίο επίπεδο που ακολούθως τροφοδοτούν το τελευταίο επιπέδου του CNN για την ταξινόμηση. Διαπιστώθηκε ακρίβεια 93.1%

COVID-19 patients' detection in chest X-ray images via MCFF-net

(Wang, Li, Wang, Li, & Zhang, 2021).

Παρουσίασαν το Parallel Channel Attention Feature Fusion Module (PCAF) και ένα νέο συνελικτικό νευρωνικό δίκτυο το MCFF-Net που βασίζεται στο PCAF. Το δίκτυο χρησιμοποιεί τρεις ταξινομητές για να ενισχύσει την αποτελεσματικότητα της αναγνώρισης: Πλήρως συνδεδεμένο (Fully Connected - FC), συνολική μέσης συγκέντρωσης πλήρως συνδεδεμένη (global average pooling fully connected (GAP-FC)), και συνελικτική συνολική μέσης συγκέντρωσης convolution global average pooling (C-GAP)). Η ακρίβεια που επιτεύχθηκε είναι 96.79% για μια ταξινόμηση τεσσάρων τάξεων.

CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images

(Khan, Shah, Mudasir, & Bhat, 2020).

Έκαναν ταξινόμηση των εικόνων ακτινογραφίας θώρακος στις κλάσεις: Φυσιολογικές, Βακτηριακές, Πνευμονίας (ιικές) και COVID-19. Παρουσίασαν ένα μοντέλο βαθέως συνελικτικού νευρωνικού δικτύου βασισμένο στην αρχιτεκτονική Xception[11]. Χρησιμοποίησαν μια μέθοδο μάθησης με μεταφορά από προεκπαιδευμένα μοντέλα. Εκπαίδευσαν το μοντέλο σε ένα σύνολο δεδομένων που προετοιμάστηκε με τη συλλογή εικόνων ακτίνων Χ από βάσεις δεδομένων που είναι διαθέσιμες στο κοινό. Επιτεύχθηκε συνολική ακρίβεια 89,6.

A deep learning system to screen novel coronavirus disease 2019 pneumonia

(Xu, και συν., 2020)

Εφάρμοσαν τεχνικές βαθιάς μάθησης για τη δημιουργία ενός μοντέλου πρώιμου προσυμπτωματικού ελέγχου για τη διάκριση του COVID-19 από την ιογενή πνευμονία της γρίπης και υγιών περιπτώσεων χρησιμοποιώντας αξονικές τομογραφίες θώρακα.

[11] Xception *είναι μια βαθιά συνεκτική αρχιτεκτονική νευρωνικών δικτύων.*

COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images

(Hemdan, Shouman, & Karar, 2020).

Χρησιμοποίησαν 50 επισημειωμένες εικόνες ακτινογραφίας θώρακα και 25 επιβεβαιωμένα θετικά κρούσματα COVID-19 και ανέπτυξαν το COVIDX-Net, το οποίο ενσωματώνει επτά διαφορετικές αρχιτεκτονικές μοντέλων CNN. Κάθε μοντέλο είναι σε θέση να αναλύσει τις κανονικοποιημένες εντάσεις της εικόνας ακτίνων Χ για να ταξινομήσει την κατάσταση του ασθενούς σε είτε αρνητική είτε θετική περίπτωση COVID-19. Τα πειράματα και η αξιολόγηση του COVIDX-Net έχουν γίνει με επιτυχία με βάση την αναλογία 80-20% των εικόνων ακτίνων Χ για τις φάσεις εκπαίδευσης και δοκιμής του μοντέλου, αντίστοιχα. Τα μοντέλα VGG19 και DenseNet [12] έδειξαν καλή και παρόμοια απόδοση της αυτοματοποιημένης ταξινόμησης COVID-19 με σκορ f1 0,89 και 0,91.

Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning

(Minaee, Kafieh, Sonka, Yazdani, & Soufi, 2020).

Χρησιμοποίησαν δημόσια διαθέσιμα σύνολα δεδομένων για να δημιουργήσουν ένα σύνολο δεδομένων από 5000 ακτινογραφίες θώρακα. Τέσσερα μοντέλα CNN εκπαιδεύτηκαν για την ανίχνευση της νόσου COVID-19 χρησιμοποιώντας τη μάθηση με μεταφορά (transfer learning).

(Υποκεφάλαιο 2.2) Εργασίες που έχουν γίνει στον τομέα της διάγνωσης του κορονοϊού με ενεργή μάθηση.

AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data

(Santosh, 2020).

Γίνεται μια θεωρητική προσέγγιση της αντιμετώπισης του covid-19 από πολλαπλές πηγές και με ενεργή μάθηση.

Προτείνεται:

i) Αντί να χρησιμοποιούνται διαφορετικά μοντέλα μηχανικής μάθησης για κάθε τύπο δεδομένων (πχ άλλο για ακτινογραφίες και άλλο για τομογρα-

[12] VGG19 και DenseNet είναι εξελιγμένα συνελικτικών νευρωνικών δικτύων (CNN - Convolutional Neural Network)

φίες) και μετά να αναζητούμε τεχνικές συνόλου για να συνδυαστούν τα αποτελέσματα, είναι συνετό να χρησιμοποιούμε ταυτόχρονα δεδομένα διαφορετικών τύπων.

ii) Αντί να περιμένουμε να συμπληρωθεί ο αριθμός των δειγμάτων που απαιτούνται θα πρέπει να χρησιμοποιούμε τεχνικές ενεργούς μάθησης για να επιταχύνουμε την διαδικασία.

Semi-Supervised Active Learning for COVID-19 Lung Ultrasound Multi-symptom Classification

(Liu, και συν., 2020)

Έχουν δημιουργήσει ειδικά για την εργασία τους μια μεγάλη επισημειωμένη βάση δεδομένων εικόνων υπερήχων. Χρησιμοποιούν ένα συνελικτικό δίκτυο ως εργαλείο ταξινόμησης πολλαπλών επισημειώσεων και προτείνουν μια μέθοδο εποπτευόμενης ενεργούς μάθησης (active learning) δύο ρευμάτων (TSAL) η οποία μετά από δοκιμές απέδειξαν ότι απαιτεί κατά 20% λιγότερα πραγματικά επισημειωμένα δεδομένα από αυτά που απαιτεί μια πλήρως εποπτευόμενη μάθηση. Η μέθοδος αυτή λειτουργεί επαναληπτικά με επιλογή δείγματος, επικύρωση με ψευδοεπισημείωση (αυτόματα), αλληλεπίδραση ανθρώπου-μηχανής και ενημέρωση παραμέτρων του CCN. Χρησιμοποιεί μια μέθοδο αυτό-εκπαίδευσης για την απόδοση επισημειώσεων και χρησιμοποιεί ανθρώπινη παρέμβαση (ενεργή μάθηση) για την ενημέρωση του εκτιμητή. Εξάγει τα χαρακτηριστικά από την βάση δεδομένων που έχει δημιουργήσει και την επισημειώσει με μικτή μέθοδο (αυτοματοποιημένη και χειροκίνητη).

A Weakly Supervised Region-Based Active Learning Method for COVID-19 Segmentation in CT Images

(Rodriguez, και συν., 2020)

Προτείνουν ένα σύστημα ενεργητικής μάθησης για την επισημείωση αξονικών τομογραφιών. Το σύστημά επικεντρώνεται στην μείωση του χρόνου επισημείωσης. Παρουσιάζει στον «επισημειωτή» περιοχές χωρίς επισημείωση που υπόσχονται υψηλό περιεχόμενο πληροφοριών και χαμηλό κόστος επισημείωσης.

Τα πειράματα με σύνολα δεδομένων COVID-19 δείχνουν ότι η χρήση μιας μεθόδου που βασίζεται στην εντροπία για την κατάταξη περιοχών χωρίς επισημείωση αποδίδει σημαντικά καλύτερα αποτελέσματα από την τυχαία επισημείωση αυτών των περιοχών. Επίσης, αποδεικνύουν ότι η επισημείωση μι-

κρών περιοχών εικόνων είναι πιο αποτελεσματική από την επισημείωση ολόκληρων εικόνων. Τέλος, αποδεικνύουν ότι μόνο με το 7 % της προσπάθειας επισημείωσης που απαιτείται για την επισημείωση ολόκληρου του εκπαιδευτικού σετ μας δίνει περίπου το 90 % της απόδοσης που επιτυγχάνεται με την εκπαίδευση του μοντέλου στο πλήρως επισημειωμένο σύνολο.

Highly Efficient Representation and Active Learning Framework and Its Application to Imbalanced Medical Image Classification

(Hao, Moon, Didari, & Jae Oh Woo, 2022)

Προτείνουν ένα ενεργό πλαίσιο μάθησης για ταξινόμηση ακτινογραφιών. Το πλαίσιο αυτό συνδυάζει: μάθηση χωρίς επίβλεψη με την βοήθεια ενός Συνελικτικού Νευρωνικού Δικτύου (ResNet 50[13]) και τη μέθοδο Gaussian Process (GP)[14], για την επιλογή «υψηλής απόδοσης δεδομένων» για την ταξινόμηση.

Επιπλέον, και τα δύο στοιχεία είναι λιγότερο ευαίσθητα στο ζήτημα της ανισορροπίας της τάξης, χάρη στο χαρακτηριστικό του να μαθαίνει χωρίς επισημειώσεις και στην Bayesian φύση του GP. Οι εκτιμήσεις αβεβαιότητας που παρέχονται από τον GP επιτρέπουν την ενεργή μάθηση να επισημειώσει επιλεκτικά δείγματα που δείχνουν υψηλότερη αβεβαιότητα. Αποδεικνύουν ότι απαιτείται το 10% των δεδομένων με επισημείωση για να επιτευχθεί η ίδια ακρίβεια με την εκπαίδευση όλων των διαθέσιμων.

Active Learning Strategy for COVID-19 Annotated Dataset

(Nazir & Fajri, 2021)

Σε αυτό το έγγραφο, προτείνεται αλγόριθμος μάθησης (DS3) για να επιτρέψει ταχύτερη και πιο αποτελεσματική επισημείωση δεδομένων. Το πλαίσιο που έχει σχεδιαστεί ειδικά για να ταιριάζει στο φαινόμενο ανισορροπίας δεδομένων που είναι χαρακτηριστικό των δεδομένων COVID-19. Εκτεταμένα πειράματα σε τέσσερα δημόσια σύνολα δεδομένων COVID-19 πραγματικού κόσμου από διάφορες χώρες δείχνουν ότι το προτεινόμενο πλαίσιο επιτυγχάνει μια μέση 10% βελτίωση.

---

[13] Το ResNet-50 *είναι ένα συνελικτικό νευρωνικό δίκτυο με βάθος 50 στρωμάτων.*

[14] *είναι μια γενική εποπτευόμενη μέθοδος μάθησης που έχει σχεδιαστεί για την επίλυση προβλημάτων παλινδρόμησης και πιθανολογικής ταξινόμησης.*

Ο Random Forest επιλέγεται ως ο κύριος αλγόριθμος εκμάθησης και η εντροπία ως μετρητής αβεβαιότητας. Επίσης έχουν γίνει διάφορα πειράματα με διαφορετικό αρχικό αριθμό πραγματικά επισημειωμένων.

Υποκεφάλαιο 2.3) Εργασίες που έχουν γίνει στον τομέα της διάγνωσης του κορονοϊού με ημι-εποπτευόμενη μάθηση

Sample-efficient deep learning for COVID-19 diagnosis based on CT scans

(He, και συν., 2020)

Δημιούργησαν μια μέθοδο βαθιάς μάθησης για την ταξινόμηση του COVID-19. Προτείνεται μια τεχνική, η οποία συνδυάζει την αυτό-εκπαιδευόμενη (self-training) μάθηση με τη μεταφορά μάθησης (transfer learning) από την εξαγωγή χαρακτηριστικών, αποφεύγοντας ταυτόχρονα την υπερβολική προσαρμογή (overfitting). Έχουν συγκεντρώσει τις προσπάθειές τους στην δοκιμή συνδυασμών μη εποπτευόμενης διαδικασίας και αυτό-εποπτευόμενης μάθησης που είναι ανθεκτικές στην υπερπροσαρμογή. Για να λύσει το πρόβλημα της διαφορετικότητας μεταξύ των προεκπαιδευμένων δεδομένων και των δεδομένων στόχος, προτείνουν μια εποπτευόμενη μέθοδο. Χρησιμοποιούν την μέθοδο εξαγωγής χαρακτηριστικών από τυχαία αρχικοποιημένα μοντέλα και από προεκπαιδευμένα μοντέλα, επικεντρώνεται στην σύγκριση των διαφόρων μεθόδων εξαγωγής χαρακτηριστικών σε συνδυασμό με τον αλγόριθμο self-train που προτείνουν με αποτέλεσμα ποσοστό ακρίβειας 94%

Semi-supervised COVID-19 CT image segmentation using deep generative models

(Zammit, Fung, Liu, Leung, & Hu, 2022)

Προτείνεται ένα μοντέλο που χρησιμοποιεί τις ικανότητες τμηματοποίησης εικόνας των δικτύων βαθιάς συνέλιξης και τις ημι-εποπτευόμενες μάθησης των μοντέλων με αξονικές τομογραφίες θώρακα ασθενών με COVID-19.

Δημιουργούν ένα βαθύ συνελικτικό δίκτυο κατάλληλο για τομογραφίες πνεύμονα. Το τελικό στρώμα αναγκάζει το μοντέλο να ανακατασκευάσει την εικόνα εισόδου χρησιμοποιώντας μια τμηματοποίηση.

Η απόδοση του προτεινόμενου μοντέλου είναι συγκρίσιμη με εκείνη του πλήρως εποπτευόμενου μοντέλου U-Net[15]. Το μοντέλο που προτείνουν, μαθαίνει ή από μόνο τα επισημειωμένα δεδομένα ή μόνο από τα μη επισημειωμένα.

[15] Το U-Net *είναι ένα συνελικτικό νευρωνικό δίκτυο που αναπτύχθηκε*

(Alizadehsani, και συν., 2021)

Προτείνουν μια ημι-εποπτευόμενη ταξινόμηση με χρήση περιορισμένων επισημειωμένων δεδομένων (SLLLD) που βασίζεται στον εντοπισμό ακμών με φίλτρο Sobel[16] και στα Generative Adversarial Networks (GAN)[17] για την αυτοματοποίηση της διάγνωσης του COVID-19. Η έξοδος του GAN είναι μια πιθανολογική τιμή που χρησιμοποιείται για ταξινόμηση. Το προτεινόμενο σύστημα εκπαιδεύεται χρησιμοποιώντας 10.000 αξονικές τομογραφίες που συλλέγονται από νοσοκομείο, ενώ ένα δημόσιο σύνολο δεδομένων χρησιμοποιείται για την επικύρωση του συστήματος. Η προτεινόμενη μέθοδος συγκρίνεται με άλλες εποπτευόμενες μεθόδους. Το σύστημά είναι ικανό να μάθει από ένα μείγμα περιορισμένων δεδομένων με επισημείωση και χωρίς επισημείωση, όπου οι εποπτευόμενοι αλγόριθμοι αποτυγχάνουν λόγω έλλειψης επαρκούς όγκου επισημειωμένων δεδομένων. Έτσι, η ημι-εποπτευόμενη μέθοδος εκπαίδευσης ξεπερνά σημαντικά την εποπτευόμενη εκπαίδευση Συνελικτικού Νευρωνικού Δικτύου (CNN) όταν τα δεδομένα εκπαίδευσης με επισημείωση είναι σπάνια. Τα διαστήματα εμπιστοσύνης 95% για τη μέθοδό μας όσον αφορά την ακρίβεια, την ευαισθησία και την ειδικότητα είναι 99,56 +- 0,20%, 99,88 +- 0,24% και 99,40 +- 0,18%, αντίστοιχα, ενώ τα διαστήματα για το CNN (εκπαιδευμένο εποπτευόμενο) είναι 68. +- 4,11%. Γίνεται προσέγγιση σε δύο φάσεις, σε πρώτη φάση ο ταξινομητής εκπαιδεύεται να ανιχνεύει έγκυρες εικόνες αξονικής σάρωσης. Στη δεύτερη φάση, χρησιμοποιώντας την τεχνογνωσία που αποκτήθηκε από την πρώτη φάση, ο εκπαιδευμένος υπεύθυνος διάκρισης μπορεί να μάθει γρηγορότερα εικόνες CT από ασθενείς και υγιείς. Η μαθησιακή ώθηση οφείλεται στο γεγονός ότι ανεξάρτητα από το αν είναι COVID ή υγιής, κάθε εικόνα εκπαίδευσης/δοκιμής είναι μια έγκυρη αξονική τομογραφία την οποία έχει κατακτήσει ο υπεύθυνος διάκρισης χρησιμοποιώντας δεδομένα χωρίς επισημείωση.

για την τμηματοποίηση βιοϊατρικής εικόνας στο Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου του Φράιμπουργκ.

[16] Το *φίλτρο* Sobel, *χρησιμοποιείται στην επεξεργασία εικόνας και στην όραση υπολογιστή, ιδιαίτερα στους αλγόριθμους ανίχνευσης άκρων όπου δημιουργεί μια εικόνα που δίνει έμφαση στις άκρες.*

[17] Το Generative Adversarial Network (GAN) *είναι μια κατηγορία πλαισίων μηχανικής μάθησης στα οποία τα νευρωνικά δίκτυα ανταγωνίζονται μεταξύ τους όπου το κέρδος του ενός είναι η απώλεια του άλλου.*

(Han, Kim, & 3, 2021)

Προτείνουν ένα ημι-εποπτευόμενο βαθύ νευρωνικό δίκτυο για βελτιωμένη ανίχνευση του COVID-19. Η προτεινόμενη μέθοδος χρησιμοποιεί αξονικές τομογραφίες με εποπτευόμενο και χωρίς επίβλεψη τρόπο για να βελτιώσει την ακρίβεια και την ευρωστία της διάγνωσης του COVID-19. Χρησιμοποιούνται τόσο επισημασμένες όσο και μη επισημασμένες εικόνες. Για τη συστηματική αξιολόγηση της προτεινόμενης μεθόδου, χρησιμοποιούνται δύο σύνολα δεδομένων CT COVID-19 και τρία δημόσια σύνολα δεδομένων CT χωρίς επισημείωση. Κατά τη διάκριση των τομογραφιών σε COVID-19 από μη-COVID-19, η προτεινόμενη μέθοδος επιτυγχάνει συνολική ακρίβεια 99,83%, ευαισθησία 0,9286, ειδικότητα 0,9832 και θετική προγνωστική αξία (PPV) 0,9192. Για τη διάκριση μεταξύ των τομογραφιών COVID-19 και κοινής πνευμονίας, η προτεινόμενη μέθοδος λαμβάνει 97,32% ακρίβεια, 0,9971 ευαισθησία, 0,9598 ειδικότητα και 0,9326 PPV. Επιπλέον, τα συγκριτικά πειράματα σε σχέση με τις στρατηγικές εποπτευόμενης μάθησης δείχνουν ότι η προτεινόμενη μέθοδος είναι σε θέση να βελτιώσει τη διαγνωστική ακρίβεια και ευρωστία χωρίς εξαντλητική επισημείωση. Η προτεινόμενη ημι-εποπτευόμενη μέθοδος, εκμεταλλεύεται τόσο την εποπτευόμενη όσο και την μη εποπτευόμενη μάθηση.

(Υποκεφάλαιο 2.4) Εργασίες που έχουν γίνει στον τομέα της διάγνωσης του κορονοϊού με θετικά και μη επισημειωμένα δεδομένα

Semi-Supervised Screening of COVID-19 from Positive and Unlabeled Data with Constraint Non-Negative Risk Estimator

(Han, et al., 2021)

Προτείνουν μια νέα μέθοδο μάθησης από θετικά και μη επισημειωμένα δεδομένα. Εισηγείται την χρήση ενός εκτιμητή μη αρνητικού κινδύνου. Προτείνει τον εκτιμητή περιορισμού μη αρνητικού κινδύνου, ο οποίος είναι πιο ισχυρός έναντι της υπερπροσαρμογής από προηγούμενες μεθόδους μάθησης από θετικά και μη επισημειωμένα δεδομένα (PU learning) όταν έχουμε περιορισμένα θετικά δεδομένα. Ενσωματώνει επίσης έναν νέο και αποτελεσματικό αλγόριθμο βελτιστοποίησης που μπορεί να κάνει το μοντέλο να μάθει καλά σε θετικά δεδομένα και να αποφύγει την υπερβολική προσαρμογή σε

δεδομένα χωρίς επισημείωση. Από όσο γνωρίζουν, αυτή είναι η πρώτη εργασία που υλοποιεί την μάθηση PU ώστε να γίνουν διαγνωστικές προβλέψεις για τον COVID-19. Μια σειρά εμπειρικών μελετών δείχνει ότι ο αλγόριθμός τους ξεπερνά αξιοσημείωτα την τελευταία λέξη της τεχνολογίας σε πραγματικά σύνολα δεδομένων σε ακτινογραφίας και της αξονικές τομογραφίες τομογραφίας.

(Υποκεφάλαιο 2.5) Συμπεράσματα

Έχει γίνει αξιοσημείωτη προσπάθεια στον τομέα της αναζήτησης μεθόδων μηχανικής και βαθιάς μάθησης για την δημιουργία μοντέλων που αξιόπιστα θα μπορούσαν να συμπληρώσουν τις υπάρχουσες διαγνωστικές μεθόδους. Οι περισσότερες ακολουθούν την μέθοδο της μάθησης με μεταφορά με κάποιες παραλλαγές:

1. Χρησιμοποιούν κάποιο προεκπαιδευμένο συνελικτικό νευρωνικό δίκτυο για εξαγωγή των χαρακτηριστικών από τις εικόνες. Εμπλουτίζουν τεχνητά τις εικόνες για να αναγνωρίζονται τα «αντικείμενα» από διαφορετικές γωνίες και σε διαφορετικές θέσεις άλλα κυρίως ώστε να αποφεύγεται η υπερπροσαρμογή. Για τον ίδιο λόγο αλλά και για την μείωση της απαιτούμενης «υπολογιστικής ενέργειας»[18] προβαίνουν σε μείωση των χαρακτηριστικών. Τα προεκπαιδευμένα DesNet φαίνεται ότι αποδίδουν καλύτεραρα στην εξαγωγή των χαρακτηριστικών.

2. Τροφοδοτούν ένα αλγόριθμο πλήρως εποπτευόμενης μάθησης η τα τελευταία επίπεδα του συνελικτικού δικτύου, συνήθως το τελευταίο (πλήρως συνδεδεμένο) με τα χαρακτηριστικά και τις επισημειώσεις. Στην περίπτωση που τα επισημειωμένα δεδομένα δεν επαρκούν τα συμπληρώνουν είτε με μια μέθοδο ημι-εποπτευόμενης μάθησης είτε με μια μέθοδο είτε με ενεργή μάθηση, δηλαδή με ανθρώπινη παρέμβαση.

3. Εκπαιδεύουν με τον αλγόριθμο που επιλέξανε ή κατασκεύασαν ένα μοντέλο κσι το δοκιμάζουν.

4. Αποτιμούν την αξιοπιστία του με την εφαρμογή μετρικών.

Οι περισσότερες περιπτώσεις αφορούν πλήρως εποπτευόμενη μάθηση, λιγότερες αφορούν ημι-εποπτευόμενη μάθηση, ακόμη λιγότερες ενεργή μάθηση και ελάχιστες μάθηση από θετικά και μη επισημειωμένα δεδομένα.

[18] Υπολογιστή ισχύς και χρόνος

Στην ενεργή και ημι-εποπτευόμενη μάθηση στις περισσότερες των περιπτώσεων χρησιμοποιούνται αξονικές τομογραφίες οι οποίες είναι κοστοβόρες και όχι διαθέσιμες σε μεγάλες ποσότητες. Είναι προφανές ότι η διάγνωση από ακτινογραφίες είναι πιο χρήσιμη ειδικά στην περίπτωση που γίνεται σε προσυμτωματικούς ή ασυμτωματικούς ασθενείς. Άρα υπάρχει μια έλλειψη μελετών στις μεθόδους με μερικώς επισημειωμένα δεδομένα στην έρευνα με την βοήθεια απλών ακτινογραφιών. Σοβαρότατη έλλειψη υπάρχει στην ερευνά σχετικά με χρήση μεθόδων μάθησης από θετικά και μη επισημειωμένα δεδομένα.

Στις περισσότερες των περιπτώσεων δοκιμάζεται ένας αλγόριθμος ταξινόμησης, ενώ ορισμένες φορές δοκιμάζονται μερικοί αλγόριθμοι στην πρώτη φάση της εκπαίδευσης με μεταφορά, δηλαδή 2-3 διαφορετικά συνελικτικά δίκτυα, για εξαγωγή χαρακτηριστικών. Δεν υπάρχουν μελέτες που να συγκρίνουν μεταξύ τους εξαντλητικά τους αλγορίθμους εκτιμητές η ταξινομητές που εμπλέκονται στην επιλογή των προς επισημείωση ή ψευδοεπισημείωση δεδομένων.

Επίσης στις περισσότερες των περιπτώσεων αναφέρονται οι τιμές των μετρικών που επιτυγχάνονται στο πείραμα και τις συγκρίνουν με ορισμένες άλλες.

Με βάση τα παραπάνω δεδομένα αποφασίσαμε να χρησιμοποιήσουμε ως δεδομένα εισόδου κοινές ακτινογραφίες και να μελετήσουμε:

Την απόδοση των μοντέλων που παρήχθησαν από το σύνολο των γνωστών αλγορίθμων ημι-εποπτευόμενης μάθησης, της ενεργούς μάθησης και της μάθησης με θετικά και μη επισημειωμένα δεδομένα.

Να βρούμε κατά περίπτωση τον ελάχιστο αριθμό των πραγματικά επισημειωμένων δεδομένων που απαιτούνται για να πετύχουμε την επιθυμητή απόδοση (εν προκειμένω 0,90 στην τιμή f1-score).

Το βασικό ερωτήματα στα οποία προσπαθεί να απαντήσει η μελέτη μας είναι:

Πόσο αποτελεσματική μπορεί να είναι μια μέθοδος με χρήση μερικώς επισημειωμένων δεδομένων στην διάγνωση του κορονοϊού; Ποια είναι η βάση των «πραγματικά επισημειωμένων δεδομένων» που πρέπει να έχουμε ώστε να είναι αποτελεσματική η μάθηση με μερικώς επισημειωμένα δεδομένα. Ποιος αλγόριθμος είναι καλύτερος, δηλαδή επιτυγχάνει αξιόπιστα μοντέλα με τα λιγότερα πραγματικά επισημειωμένα δεδομενα.

ΚΕΦΑΛΑΙΟ 3 ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ – ΑΝΑΓΝΩΡΙΣΗ ΕΙΚΟΝΩΝ

(ΥΠΟκεφάλαιο 3.1) Αναγνώριση προτύπων - εικόνων

(Ενότητα 3.1.α) Γενικά

Η αναγνώριση προτύπων χρησιμοποιεί αλγόριθμους τεχνητής νοημοσύνης για να αναγνωρίζει αυτόματα μοτίβα και σχέσεις στα δεδομένα μας. Αυτά τα δεδομένα μπορεί να είναι οτιδήποτε, από κείμενο και εικόνες έως ήχους ή άλλες ιδιότητες που μπορούν να προσδιοριστούν. Η αναγνώριση εικόνων περιορίζεται και εξειδικεύεται στην αναγνώριση και ταξινόμηση εικόνων.

Ένα καλό σύστημα αναγνώρισης προτύπων θα πρέπει να αναγνωρίζει γνωστά «μοτίβα» γρήγορα και με ακρίβεια. Επίσης να αναγνωρίζει και ταξινομεί άγνωστα αντικείμενα, σχήματα και αντικείμενα από διαφορετικές οπτικές γωνίες.

Βασική προϋπόθεση για να γίνει αυτό είναι να υπάρχει ένα σύνολο παρατηρήσεων ή αλλιώς διάνυσμα χαρακτηριστικών (μονοδιάστατος πίνακας). Ένα χαρακτηριστικό είναι ένα διακριτικό χαρακτηριστικό ενός αγαθού ή μιας υπηρεσίας ή αντικειμένου που το ξεχωρίζει από παρόμοια είδη. Το διάνυσμα χαρακτηριστικών είναι το σύνολο των χαρακτηριστικών που αντιστοιχούν σε ένα είδος. Τα διαφορετικά είδη μπορεί να έχουν διαφορετικές τιμές χαρακτηριστικών, αλλά ένα είδος έχει πάντα τις ίδιες τιμές χαρακτηριστικών, δηλαδή το ίδιο διάνυσμα χαρακτηριστικών.

Η αναγνώριση εικόνων είναι μια πιο συγκεκριμένη περίπτωση της αναγνώρισης προτύπων (η πιο συνηθισμένη) και χρειάζεται τους εξής μηχανισμούς:

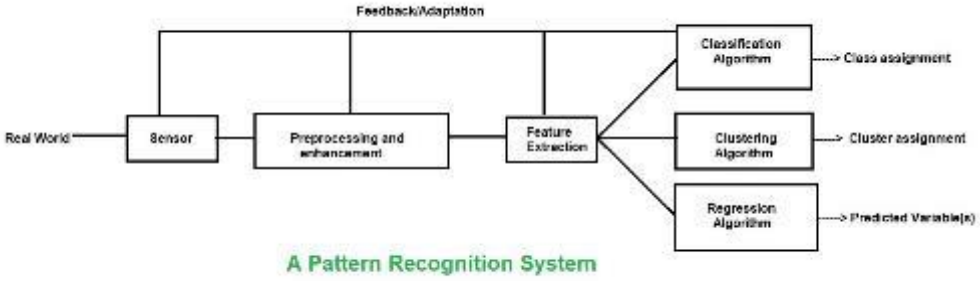
Αισθητήρας (sensor): Ο αισθητήρας είναι μια συσκευή που χρησιμοποιείται για τη μέτρηση μιας ιδιότητας, όπως η πίεση, η θέση, η θερμοκρασία, ή για την καταγραφή κάποιου φαινομένου, εικόνας, ήχου, όπως οι φωτογραφικές μηχανές, τα ακτινολογικά μηχανήματα κλπ.

Μηχανισμός εξαγωγής χαρακτηριστικών: Η εξαγωγή χαρακτηριστικών ξεκινά από ένα αρχικό σύνολο μετρούμενων δεδομένων και δημιουργεί παράγωγες τιμές (χαρακτηριστικά) διευκολύνοντας τα επόμενα βήματα της μάθησης και γενίκευσης οδηγώντας σε καλύτερες ανθρώπινες ερμηνείες.

Μηχανισμός προεπεξεργασίας: Καθαρισμός δεδομένων και προετοιμασία τους για χρήση ώστε να αποφευχθούν προβλήματα στην διαδικασία εκπαίδευσης που θα έχουν σαν συνέπεια λανθασμένα αποτελέσματα.

Μηχανισμός εκπαίδευσης: Περιλαμβάνει ένα αλγόριθμο αναγνώρισης προτύπων που στοχεύει στο να αποκαλύψει την σχέση που συνδέει τα χαρακτηριστικά με την κλάση στην οποία ανήκουν.

Σύνολο εκπαίδευσης: Τα δεδομένα εκπαίδευσης είναι ένα ορισμένο ποσοστό ενός συνόλου δεδομένων μαζί με το σύνολο δοκιμών. Κατά κανόνα, όσο καλύτερα είναι τα δεδομένα εκπαίδευσης, τόσο καλύτερα αποδίδει ο αλγόριθμος.



Εικόνα 2: Σύστημα αναγνώρισης προτύπων
(geeksforgeeks, n.d.)

Υπάρχουν διάφορες εργασίες που πρέπει να εκτελεστούν για την υλοποίηση των συστημάτων αναγνώρισης προτύπων. Οι δραστηριότητες αυτές είναι οι εξής:

- Συλλογή δεδομένων
- Εξαγωγή - επιλογή χαρακτηριστικών
- Επιλογή μοντέλου (αλγορίθμου εκπαίδευσης)
- Εκπαίδευση
- Εκτίμηση αξιοπιστίας μοντέλου

(Ενότητα 3.1.β) Συλλογή – εμπλουτισμός δεδομένων

Η απόκτηση ακατέργαστων δεδομένων όσον αφορά την ιατρική μπορεί να γίνει με μία από τις υπάρχουσες απεικονιστικές μεθόδους.

- Υπέρηχοι (Ultrasound)
- Ακτινογραφίες (X-Ray Imaging)
- Αξονικές Τομογραφίες (Computer Tomography (CT))
- Μαγνητικές (Magnetic Resonance Imaging (MRI))

- Τομογραφία εκπομπής ποζιτρονίων (Positron Emission Tomography (PET)).

«Ωστόσο, ανεξάρτητα από τον τύπο της μεθόδου απεικόνισης, η διαδικασία απόκτησης δεδομένων μπορεί να υποδιαιρεθεί σε ανίχνευση φυσικού μεγέθους που περιλαμβάνει τη μετατροπή του σε ηλεκτρικό σήμα, προετοιμασία του σήματος και ψηφιοποίηση του». (Patyuchenko, n.d.)

Η συλλογή «καλών» δεδομένων είναι συχνά το πιο δύσκολο μέρος της έρευνας. Στην ιδανική περίπτωση, θα θέλαμε μία δειγματοληψία είτε καθολική είτε τυχαία και αντιπροσωπευτική για τον πληθυσμό στόχο. Αυτή θα πρέπει να περιέχει και άλλα στοιχεία που πιθανόν να επηρεάζουν την απεικόνιση πχ ηλικία, φύλλο κλπ.

Ο (Aylward, χ.χ.) στο site https://www.aylward.org/notes/open-access-medical-image-repositories παρουσιάζει μία λίστα με πήγες ανοικτής πρόσβασης. Στην παρούσα μελέτη δεν θα χρησιμοποιήσουμε δεδομένα που παραθέτουν οι ερευνητές (Hamid Nasiri & Alavi, 2022) στην δική τους μελέτη που αφορούν πραγματικά δείγματα. Τα συγκεκριμένα δεδομένα επελέγησαν λόγω της σημαντικής ακρίβειας που παρουσιάζει η μέθοδός τους.

Πολλές φορές ειδικότερα στις εφαρμογές αναγνώρισης – ταξινόμησης εικόνας θα χρειαστεί να εμπλουτίσουμε τις εικόνες με τροποποιήσεις. Δηλαδή να τις περιστρέψουμε, να τις αλλάξουμε κλίση, να τις αλλάξουμε ε μέγεθος και να τις μετακινήσουμε ώστε το μοντέλο που θα εκπαιδεύσουμε θα σαν είσοδο τις ίδιες εικόνες σε πολλές εκδοχές.

(Ενότητα 3.1.γ) Εξαγωγή χαρακτηριστικών

«Η εξαγωγή χαρακτηριστικών (feature extraction) αναφέρεται στη διαδικασία μετατροπής ακατέργαστων δεδομένων σε αριθμητικά χαρακτηριστικά τα οποία μπορούν να υποστούν επεξεργασία διατηρώντας παράλληλα τις πληροφορίες από το αρχικό σύνολο δεδομένων. Η εξαγωγή χαρακτηριστικών αντιπροσωπεύει τα ενδιαφέροντα μέρη μιας εικόνας ως ένα διάνυσμα χαρακτηριστικών. Η εξαγωγή χαρακτηριστικών μπορεί να πραγματοποιηθεί είτε χειροκίνητα είτε αυτόματα.

Η μη αυτόματη εξαγωγή χαρακτηριστικών απαιτεί τον εντοπισμό και την περιγραφή των χαρακτηριστικών που σχετίζονται με ένα δεδομένο πρόβλημα και την εφαρμογή μιας μεθόδου εξαγωγής αυτών των χαρακτηριστικών. Λεπτομέρειες της εικόνας, όπως ο χρωματισμός, η υφή το φόντου μπορεί να

βοηθήσει στη λήψη τεκμηριωμένων αποφάσεων σχετικά με το ποια χαρακτηριστικά θα μπορούσαν να είναι χρήσιμα.

Η αυτόματη εξαγωγή χαρακτηριστικών χρησιμοποιεί εξειδικευμένους αλγόριθμους ή νευρωνικά δίκτυα για την εξαγωγή χαρακτηριστικών από σήματα ή εικόνες χωρίς την ανάγκη ανθρώπινης παρέμβασης.» (Feature Extruction, n.d.)

Υπάρχουν πολλοί αλγόριθμοι που ασχολούνται στην εξαγωγή χαρακτηριστικών εικόνων όπως:

- Ιστόγραμμα προσανατολισμένων κλίσεων (HOG)
- Scale-Invariant Feature Transform (SIRF)
- Επιταχυνόμενα στιβαρά χαρακτηριστικά (SURF)
- Χαρακτηριστικά τοπικού δυαδικού μοτίβου (LBP).
- ORB (Oriented FAST and Rotated BRIEF)
- Color Gradient Histogram
- Gabor filter

Ανεξάρτητα από την προσέγγιση που ακολουθούμε, οι εφαρμογές υπολογιστικής όρασης όπως η καταχώριση εικόνων, η ανίχνευση και ταξινόμηση αντικειμένων εικόνων και η ανάκτηση εικόνων βάσει περιεχομένου, απαιτούν αποτελεσματική αναπαράσταση των χαρακτηριστικών της εικόνας. Γενικά οι αλγόριθμοι που αναφέρθηκαν παραπάνω τείνουν να παραγκωνισθούν και να αντικατασταθούν από προεκπαιδευμένα συνελικτικά νευρωνικά δίκτυα, τα οποία εκτός από το ότι μας δίνουν την δυνατότητα να εξάγουμε χαρακτηριστικά χωρίς να γνωρίζουμε επακριβώς την διαδικασία (black box), έχουν αποδειχθεί αποτελεσματικά. Ιδιαίτερα η κατηγορία των προ-εκπαιδευμένων DenseNet τα οποία έχουν την ικανότητα να διαβλέψουν και να προτείνουν χαρακτηριστικά που είναι κατάλληλα για κάθε μοντέλο μηχανικής μάθησης.

(Ενότητα 3.1.δ) Επιλογή χαρακτηριστικών

Η επιλογή χαρακτηριστικών (feature selection) και η προβολή χαρακτηριστικών (feature projection) είναι μία διαδικασία η οποία αφορά στην απαλοιφή των χαρακτηριστικών που δεν συνεισφέρουν στην ακρίβεια του μοντέλου ή στην συγχώνευση πολλών χαρακτηριστικών σε ένα.

. Οι τεχνικές επιλογής χαρακτηριστικών χρησιμοποιούνται για διάφορους λόγους, όπως:

- Απλοποίηση των μοντέλων
- Μικρότεροι χρόνοι εκπαίδευσης των μοντέλων.
- Μικρότερη απαιτούμενη επεξεργαστική ισχύς
- Μείωση των διαστάσεων των δεδομένων
- Βελτίωση της συμβατότητας των δεδομένων με ένα μοντέλου μηχανικής μάθησης
- Κωδικοποιούν εγγενείς συμμετρίες που υπάρχουν στον χώρο εισόδου.
- Μείωση της πιθανότητας για υπερπροσαρμογή.

Η κεντρική ιδέα κατά τη χρήση μιας τεχνικής επιλογής χαρακτηριστικών είναι ότι τα δεδομένα περιέχουν ορισμένα χαρακτηριστικά που είναι είτε περιττά είτε άσχετα, και επομένως μπορούν να αφαιρεθούν χωρίς να υπάρξει μεγάλη απώλεια πληροφοριών. Περιττό και άσχετο είναι δύο διακριτές έννοιες, καθώς ένα σχετικό χαρακτηριστικό μπορεί να είναι περιττό παρουσία ενός άλλου σχετικού χαρακτηριστικού με το οποίο συσχετίζεται ισχυρά.

Οι τεχνικές επιλογής χαρακτηριστικών χρησιμοποιούνται σε γενικές γραμμές είναι:

- Επιλέγουμε ένα υποσύνολο χαρακτηριστικών εισόδου από το σύνολο δεδομένων:
  - Χωρίς επίβλεψη: Χωρίς την χρήση τη μεταβλητής στόχου (π.χ. κατάργηση περιττών μεταβλητών).
  - Υπό επίβλεψη: Με την χρήση της μεταβλητής προορισμού (π.χ. κατάργηση μη σχετικών μεταβλητών).
- Wrapper (περιτυλίγματος): Αναζήτηση για υποσύνολα χαρακτηριστικών με καλή απόδοση.
- Εγγενείς: Αλγόριθμοι που εκτελούν αυτόματη επιλογή χαρακτηριστικών κατά τη διάρκεια της εκπαίδευσης.
- Φίλτρου: Επιλέγουμε υποσύνολα χαρακτηριστικών με βάση τη σχέση τους με τον στόχο.
- Στατιστικές μέθοδοι
- Μέθοδοι σημαντικότητας χαρακτηριστικών (feature importance), επιλέγουμε τα χαρακτηριστικά που είναι πιο σημαντικά στην σχέση μεταξύ των χαρακτηριστικών και των επισημειώσεων τους.

- Δέντρα απόφασης - κέρδος πληροφορίας (information gain[19]): Μετράμε πόση πληροφορία συνεισφέρει ένα χαρακτηριστικό. Οι μετρήσεις του κέρδους πληροφορίας γίνονται με τους δείκτες μη καθαρότητας (impurity) που είναι είτε η εντροπία είτε ο δείκτης Gini.

Με την βοήθεια της βιβλιοθήκης sklearn μπορούμε να κάνουμε τα παρακάτω:

Αφαίρεση λειτουργιών με χαμηλή διακύμανση: Είναι μια απλή προσέγγιση για την επιλογή χαρακτηριστικών. Καταργεί όλα τα χαρακτηριστικά των οποίων η διακύμανση δεν πληρεί κάποιο όριο. Από προεπιλογή, καταργεί όλα τα χαρακτηριστικά μηδενικής διακύμανσης, δηλαδή τα χαρακτηριστικά που έχουν την ίδια τιμή σε όλα τα δείγματα.

Επιλογή χαρακτηριστικών με βάση δοκιμές: Αφαιρεί όλα τα χαρακτηριστικά εκτός από αυτά που έχουν τις υψηλότερες βαθμολογίες. υπολογίζοντας τον βαθμό γραμμικής εξάρτησης μεταξύ δύο τυχαίων μεταβλητών.

Αναδρομική εξάλειψη χαρακτηριστικών: Δεδομένου ενός εξωτερικού εκτιμητή που εκχωρεί βαρύτητα σε χαρακτηριστικά (π.χ. τους συντελεστές ενός γραμμικού μοντέλου), ο στόχος της αναδρομικής εξάλειψης χαρακτηριστικών (Recursive Feature Elimination) είναι η επιλογή χαρακτηριστικών εξετάζοντας αναδρομικά όλο και μικρότερα σύνολα χαρακτηριστικών. Ο εκτιμητής εκπαιδεύεται στο αρχικό σύνολο χαρακτηριστικών και εξάγεται η σημαντικότητα κάθε χαρακτηριστικού μέσω κάποιου μετρητή όπως το feature_importance[20]. Στη συνέχεια, τα λιγότερο σημαντικά χαρακτηριστικά απαλείφονται από το σύνολο χαρακτηριστικών. Αυτή η διαδικασία επαναλαμβάνεται στο εναπομείναν σύνολο έως ότου επιτευχθεί τελικά ο επιθυμητός αριθμός χαρακτηριστικών.

Επιλογή χαρακτηριστικών με βάση δεντρικές δομές: Οι εκτιμητές που βασίζονται σε δέντρα μπορούν να χρησιμοποιηθούν για τον υπολογισμό της σημαντικότητας χαρακτηριστικών που βασίζονται στην μη καθαρότητα

[19] Gain = (Entropy of the parent node) - (average entropy of the child nodes)

[20] *αναφέρεται στον υπολογισμό μιας βαθμολογίας για τα χαρακτηριστικά εισόδου για ένα μοντέλο που αντιπροσωπεύει τη «σημασία» ενός χαρακτηριστικού όσον αφορά την επίδρασή του στον υπολογισμό των τιμών «στόχου»*

(impurity) (Εντροπία[21], Geni[22]) των χαρακτηριστικών, οι οποίες μπορούν να χρησιμοποιηθούν για την απόρριψη άσχετων χαρακτηριστικών.

Σειριακή επιλογή χαρακτηριστικών: Η σειριακή επιλογή χαρακτηριστικών (Sequential Feature Selection (SFS) μπορεί να είναι είτε προς τα εμπρός είτε προς τα πίσω. Το Forward-SFS είναι μια άπληστη[23] διαδικασία που βρίσκει επαναληπτικά το καλύτερο νέο χαρακτηριστικό για προσθήκη στο σύνολο των επιλεγμένων χαρακτηριστικών. Συγκεκριμένα, αρχίζουμε με μηδενικά χαρακτηριστικά και βρίσκουμε το ένα χαρακτηριστικό που μεγιστοποιεί μια διασταυρούμενη επικυρωμένη βαθμολογία (cross-validation) όταν ένας εκτιμητής εκπαιδεύεται σε αυτό το μεμονωμένο χαρακτηριστικό. Μόλις επιλεγεί αυτό το πρώτο χαρακτηριστικό, επαναλαμβάνουμε τη διαδικασία προσθέτοντας ένα νέο χαρακτηριστικό στο σύνολο των επιλεγμένων χαρακτηριστικών. Η διαδικασία σταματά όταν επιτευχθεί ο επιθυμητός αριθμός επιλεγμένων χαρακτηριστικών. Το Backward-SFS ακολουθεί την ίδια ιδέα, αλλά λειτουργεί προς την αντίθετη κατεύθυνση: αντί να ξεκινάμε χωρίς χαρακτηριστικά και να προσθέτουμε άπληστα χαρακτηριστικά, ξεκινάμε με όλα τα χαρακτηριστικά και αφαιρούμε άπληστα χαρακτηριστικά από το σύνολο.

(Feature selection, χ.χ.)

Η προβολή χαρακτηριστικών (ονομάζεται επίσης εξαγωγή χαρακτηριστικών ή μείωση διαστάσεων) μετατρέπει τα δεδομένα από τον χώρο υψηλών διαστάσεων σε χώρο λιγότερων διαστάσεων.

Ο (Brownlee, How to Choose a Feature Selection Method For Machine Learning, 2019) παρουσιάζει στην ιστοσελίδα του μια ολοκληρωμένη και συστηματική μεθοδολογία για την επιλογή μεθόδου επιλογής χαρακτηριστικών, η οποία παρουσιάζεται συνοπτικά παρακάτω:

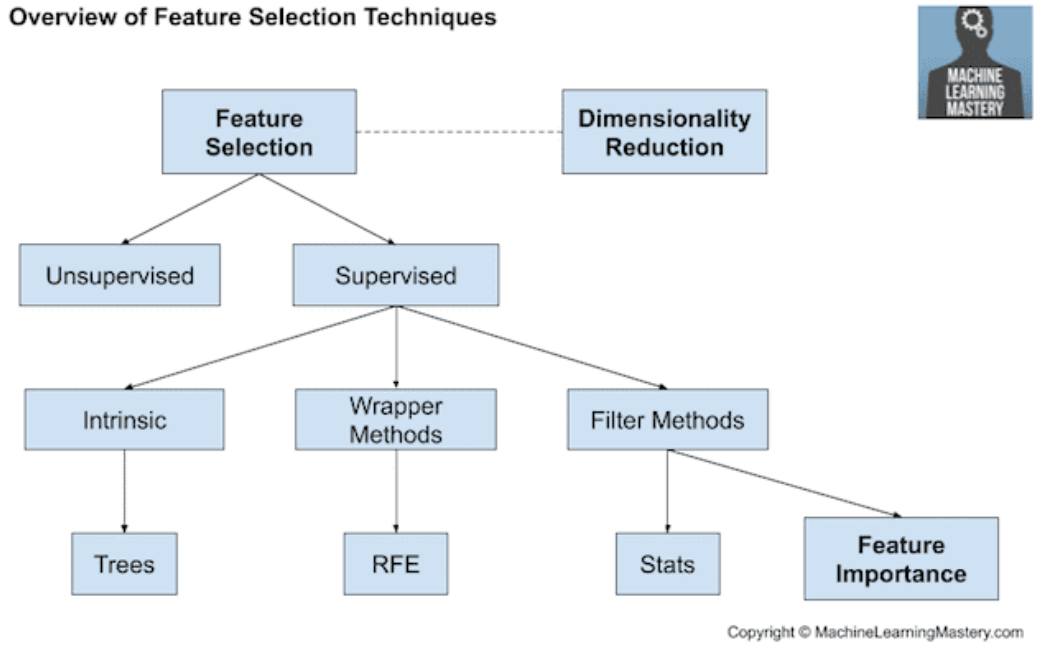
---

[21] Η *εντροπία είναι ένα μέτρο μη καθαρότητας δεδομένων που υποδεικνύει την μη κανονική θέση των χαρακτηριστικών με τον στόχο.*

[22] Ο *δείκτης* Gini *είναι μέτρο μη καθαρότητας που μετρά τη συχνότητα με την οποία οποιοδήποτε στοιχείο του συνόλου δεδομένων θα επισημειώνεται λανθασμένα όταν επισημαίνεται τυχαία.*

[23] Ένας *άπληστος* (greedy) *αλγόριθμος είναι μια προσέγγιση για την επίλυση ενός προβλήματος επιλέγοντας την καλύτερη διαθέσιμη επιλογή αυτή τη στιγμή. Δεν ανησυχεί αν το τρέχον καλύτερο αποτέλεσμα θα φέρει το συνολικό βέλτιστο αποτέλεσμα.* Ο *αλγόριθμος δεν αντιστρέφει ποτέ την προηγούμενη απόφαση ακόμα κι αν η επιλογή είναι λάθος.*

«



Εικόνα 3: Επιλογή μεθόδου επιλογής δεδομένων

Για την επιλογή της καταλληλότερης μεθόδου επιλογής χαρακτηριστικών σημαντικός παράγοντας είναι το είδος των τιμών των μεταβλητών (χαρακτηριστικών και εξόδων) λαμβάνουμε υπόψιν τα παρακάτω:.

Αριθμητική είσοδο – Αριθμητική έξοδο:
- Pearson's correlation coefficient (γραμμική).
- Spearman's rank coefficient (μη γραμμική)

Αριθμητική είσοδο – Κατηγορική έξοδο[24]:
- ANOVA correlation coefficient (γραμμική).
- Kendall's rank coefficient (μη γραμμική). Υποθέτει ότι η κατηγορική μεταβλητή εκφράζει σειρά απαρίθμησης.

Κατηγορική είσοδο – Αριθμητική έξοδο (σπάνια περίπτωση):
- ANOVA correlation coefficient (γραμμική).
- Kendall's rank coefficient (μη γρραμμική). Υποθέτει ότι η κατηγορική μεταβλητή εκφράζει σειρά απαρίθμησης.

Κατηγορική είσοδο – κατηγορική έξοδο):
- X^2 (Χι Chi-Squared test (contingency tables)).
- Αμοιβαία πληροφορία25 (Mutual Information).

»

[24] Εκφράζουν κατηγορία (πχ covid, no-covid0

[25] Από την θεωρία πληροφορίας με χρήση των δεικτών gini ή εντροπίας για προσδιορισμό του κέρδους πληροφορίας

(Brownlee, How to Choose a Feature Selection Method For Machine Learning, 2019)

(Ενότητα 3.1.ε) Καθαρισμός - μετασχηματισμός χαρακτηριστικών

Καθαρισμός δεδομένων ή χαρακτηριστικών είναι η διαδικασία διόρθωσης ή αφαίρεσης λανθασμένων, κατεστραμμένων, μη σημαντικών, διπλών ή ελλιπών χαρακτηριστικών. Δεν υπάρχει κανένας απόλυτος τρόπος για να ορίσουμε τα ακριβή βήματα στη διαδικασία καθαρισμού χαρακτηριστικών. Οι διαδικασίες θα διαφέρουν από σύνολο δεδομένων σε σύνολο δεδομένων. Ο όρος καθαρισμός δεδομένων αναφέρεται στην αφαίρεση χαρακτηριστικών και ο όρος μετασχηματισμός δεδομένων που είναι η μετατροπή των από μια μορφή ή δομή σε άλλη.

«Ο καθαρισμός δεδομένων περιέχει τα παρακάτω βήματα:

1. Χειρισμός μηδενικών τιμών (είτε αφαίρεση όλου του διανύσματος χαρακτηριστικών είτε τοποθέτηση μιας άλλης τιμής).
2. Κατάργηση διπλών ή άσχετων παρατηρήσεων
3. Διόρθωση δομικών σφαλμάτων
4. Αφαίρεση των ανεπιθύμητων υπερβολικών τιμών (outliers) εφόσον αυτό κριθεί σκόπιμο.
5. Χειρισμός των δεδομένων που λείπουν. Μπορούμε είτε να αφαιρέσουμε τις παρατηρήσεις που έχουν τιμές που λείπουν, με κίνδυνο να χαθούν πληροφορίες, είτε μπορούμε να εισάγουμε τιμές που λείπουν με κίνδυνο να χάσουμε την ακεραιότητα των δεδομένων.
6. Επικύρωση και διασφάλιση ποιότητας. »

(TABLEAU SOFTWARE, LLC, n.d.)

(Ενότητα 3.1.στ) Κωδικοποίηση Κατηγορικών Χαρακτηριστικών

Κατηγορικά δεδομένα είναι δεδομένα που έχουν ορισμένες κατηγορίες όπως, πχ (covid, noncovd), (ελέφαντας, λιοντάρι, τίγρη), (ναι, όχι) κλπ.

Δεδομένου ότι το μοντέλο μηχανικής μάθησης λειτουργεί πλήρως με μαθηματικά και αριθμούς, εάν το σύνολο χαρακτηριστικών έχει μια κατηγορική μεταβλητή, τότε μπορεί να δημιουργηθεί πρόβλημα κατά την κατασκευή του μοντέλου. Επομένως, είναι απαραίτητο να κωδικοποιήσουμε αυτές τις κατηγορικές μεταβλητές σε αριθμούς.

Παράδειγμα:

Αποτελέσματα Αγώνων	
Αποτέλεσμα Αγώνα	**Κατηγορία**
Νίκη	3
Ισοπαλία	2
Ήττα	1

Έτσι λοιπόν μετατρέπουμε τα κατηγορικά στοιχεία σε αριθμούς. Η ήττα απέχει 1 από την ισοπαλία και 2 από την Νίκη. Η κωδικοποίηση αυτή είναι λογική.

Αλλά:

Ζώα	
Αποτέλεσμα Αγώνα	**Κατηγορία**
Ελέφαντας	2
Ποντίκι	1
Τίγρη	0

Εδώ υπονοείται ότι υπάρχει κάποια σχέση πχ ο ελέφαντας είναι μεγαλύτερος από το ποντίκι και το ποντίκι μεγαλύτερο από την τίγρη.

Είναι πιο σωστό να έχουμε μετασχηματίσει τον πίνακα των χαρακτηριστικών ώστε να περιέχει 3 στήλες.

Πίνακας 1: One Hot Encoding

One Hot Encoding: Ζώα			
	Ελέφαντας	**Ποντίκι**	**Τίγρη**
Ελέφαντας	1	0	0
Ποντίκι	0	1	0
Τίγρη	0	0	1

Στην περίπτωση που αναφερόμαστε στις «επισημειώσεις» («ετικέτες» ή «τιμές output» ή «τιμές στόχος») οι περισσότεροι αλγόριθμοι μπορούν να χειριστούν από μονοί τους τις κατηγορικές τιμές, στην περίπτωση που δεν το κάνουν βάζουμε μια αριθμητική τιμή (0,1,2,3 κλπ.) για κάθε κατηγορία.

(Ενότητα 3.1.ζ) Προσαρμογή - κλιμάκωση χαρακτηριστικών

Οι όροι που χρησιμοποιούνται είναι κλιμάκωση (scaling), κανονικοποίηση (normalization) και τυποποίηση (standardization).

Η **κανονικοποίηση** (normalization) είναι μια τεχνική προσαρμογής των δεδομένων[26] στην οποία οι τιμές μετατοπίζονται και κλιμακώνονται έτσι ώστε να καταλήγουν να κυμαίνονται μεταξύ 0 και 1. Είναι επίσης γνωστή ως Min-Max scaling.

Ο τύπος κανονικοποίησης:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Το Xmax και το Xmin είναι η μέγιστη και η ελάχιστη τιμή.

Όταν η τιμή του X είναι η ελάχιστη τιμή στη στήλη, ο αριθμητής θα είναι 0 και συνεπώς το X' είναι 0. Από την άλλη πλευρά, όταν η τιμή του X είναι η μέγιστη τιμή στη στήλη, ο αριθμητής είναι ίσος με τον παρονομαστή και έτσι η τιμή του X' είναι 1. Εάν η τιμή του X είναι μεταξύ της ελάχιστης και της μέγιστης τιμής, τότε η τιμή του X 'είναι μεταξύ 0 και 1.

Η **τυποποίηση** (standardization) είναι μια άλλη τεχνική κλιμάκωσης όπου οι τιμές επικεντρώνονται γύρω από το μέσο όρο με τυπική απόκλιση μονάδας (τυπική κανονική κατανομή). Αυτό σημαίνει ότι ο μέσος όρος του χαρακτηριστικού γίνεται μηδέν και η προκείπτουσα κατανομή έχει τυπική απόκλιση μονάδας.

Εξίσωση τυποποίησης:

$$X' = \frac{X - \mu}{\sigma}$$

μ είναι η μέση τιμή των χαρακτηριστικών, σ είναι η τυπική απόκλιση των τιμών των χαρακτηριστικών.

[26] Πολλοί συγγραφείς το ονομάζουνε scaling

Ορισμένοι αλγόριθμοι μηχανικής μάθησης είναι ευαίσθητοι στην προσαρμογή των χαρακτηριστικών, ενώ άλλοι δεν επηρεάζονται.

Η κανονικοποίηση είναι προτιμότερο να χρησιμοποιείται σε αλγόριθμους που δεν υποθέτουν καμμιά κατανομή δεδομένων, όπως Κ πλησιέστεροι γείτονες (K-Nearest Neighbors) και νευρωνικά δίκτυα (Neural Networks) ενώ η τυποποίηση είναι χρήσιμη σε περιπτώσεις όπου τα δεδομένα ακολουθούν μια κανονική (Gaussian) κατανομή. Σε αντίθεση με την κανονικοποίηση, η τυποποίηση δεν έχει όριο. Έτσι, ακόμη και αν έχουμε υπερβολικά υψηλά ή χαμηλά (outliers) τα δεδομένα μας, δεν θα επηρεαστούν από την τυποποίηση.

Οι **αλγόριθμοι μηχανικής και βαθιάς μάθησης**, όπως η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση, το νευρωνικό δίκτυο κ.λπ. που χρησιμοποιούν την gradient descent[27] ως τεχνική βελτιστοποίησης απαιτούν κλιμάκωση δεδομένων.

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Η παρουσία της τιμής χαρακτηριστικού x στον τύπο, θα επηρεάσει το μέγεθος βημάτων της gradient descent. Η διαφορά στο εύρος των χαρακτηριστικών θα προκαλέσει διαφορετικά μεγέθη βημάτων για κάθε συνάρτηση. Για να διασφαλίσουμε ότι η gradient descent κινείται ομαλά προς το ελάχιστο και ότι τα βήματα ενημερώνονται με τον ίδιο ρυθμό για όλα τα χαρακτηριστικά, προσαρμόζουμε τα δεδομένα πριν τροφοδοτήσουμε στο μοντέλο. Έτσι η σύγκλιση θα επιτευχθεί ταχύτερα.

Οι **αλγόριθμοι απόστασης** όπως το (K-Nearest Neighbors) KNN, το K-means και τα Support Vector Machines (SVM) επηρεάζονται από το εύρος των χαρακτηριστικών. Αυτό συμβαίνει επειδή χρησιμοποιούν αποστάσεις μεταξύ των σημείων δεδομένων για να προσδιορίσουν την ομοιότητά τους.

---

[27] [27] **Gradient descent** *είναι ένας γενικός αλγόριθμος βελτιστοποίησης ικανός να βρει βέλτιστες λύσεις σε ένα ευρύ φάσμα προβλημάτων. Η γενική ιδέα είναι να τροποποιήσει τις παραμέτρους επαναληπτικά προκειμένου να ελαχιστοποιηθεί μια συνάρτηση κόστους. Μετρά την τοπική κλίση της συνάρτησης απώλειας για ένα σύνολο παραμέτρων και κάνει βήματα προς την κατεύθυνση της φθίνουσας κλίσης. Μόλις η κλίση είναι μηδέν, φτάσαμε στο ελάχιστο. Μια σημαντική παράμετρος είναι το μέγεθος των βημάτων που καθορίζεται από τον ρυθμό εκμάθησης* **(learning ratio).**

Εάν τα χαρακτηριστικά έχουν διαφορετικές κλίμακες, υπάρχει πιθανότητα να δοθεί υψηλότερο βάρος σε χαρακτηριστικά με μεγαλύτερο μέγεθος.

Οι **αλγόριθμοι που βασίζονται σε δέντρα**, από την άλλη πλευρά, δεν επηρεάζονται από την κλίμακα των χαρακτηριστικών. Ένα δέντρο αποφάσεων χωρίζει μόνο έναν κόμβο που βασίζεται σε ένα μόνο χαρακτηριστικό που αυξάνει την ομοιογένεια του κόμβου. Αυτή η διάσπαση σε ένα χαρακτηριστικό δεν επηρεάζεται από άλλα χαρακτηριστικά. Έτσι, δεν υπάρχει καμία επίδραση των υπόλοιπων χαρακτηριστικών στη διάσπαση.

(Ενότητα 3.1.η) Εκπαίδευση - μάθηση

Μετά τις παραπάνω διαδικασίες που αφορούν στην λήψη και προετοιμασία των δεδομένων ακολουθεί η διαδικασία επιλογής μοντέλου που θα ανακαλύψει τις κρυφές σχέσεις μεταξύ των χαρακτηριστικών και των τιμών στόχου.

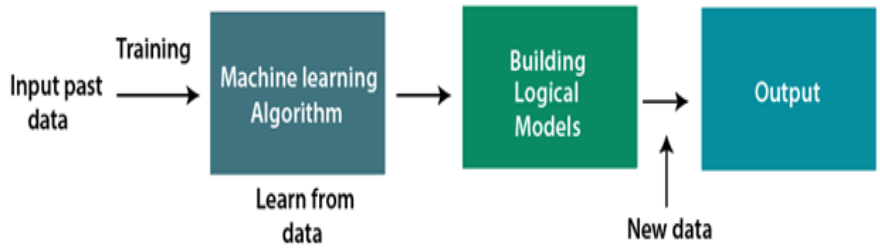
Το σημαντικότερο μέρος ενός μοντέλου που έχει κατασκευαστεί για να κάνει προβλέψεις είναι η εκπαίδευσή του, με σκοπό να «μάθει» από τα γνωστά δεδομένα ώστε να μπορεί να κάνει προβλέψεις. Η μάθηση είναι ένα φαινόμενο μέσω του οποίου ένα σύστημα εκπαιδεύεται και προσαρμόζεται ώστε να δίνει αποτελέσματα με ακριβή τρόπο. Η μάθηση είναι η πιο σημαντική φάση ως προς το πόσο καλά αποδίδει το σύστημα.

Τα είδη μάθησης – εκπαίδευσης μπορούν να χωριστούν σε δύο μεγάλες κατηγορίες την **μηχανική μάθηση**. και στην **βαθιά μάθηση** τα οποία εξετάζουμε στα επόμενα κεφάλαια.

ΚΕΦΑΛΑΙΟ 4 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

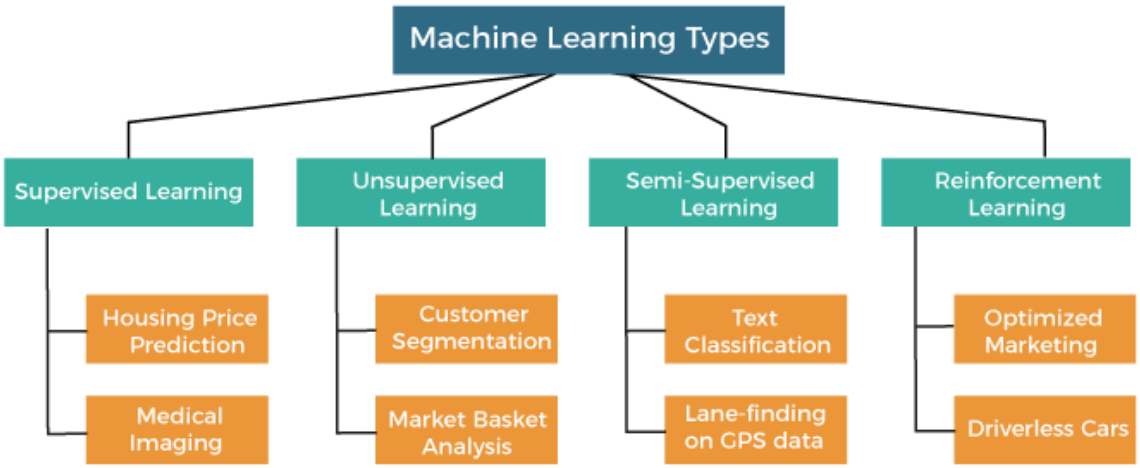
(Υποκεφάλαιο 4.1) Γενικά

«Η μηχανική μάθηση (machine learning) είναι υποσύνολο της τεχνητής νοημοσύνης και αναφέρεται στην ιδέα, ότι τα προγράμματα υπολογιστών μπορούν αυτόματα να μάθουν και να προσαρμοστούν σε νέα δεδομένα χωρίς να βοηθηθούν από τον άνθρωπο» (Gordon, 2021).



Εικόνα 4: Διαδικασία μηχανικής μάθησης

(Υποκεφάλαιο 4.2) Κατηγορίες μηχανικής μάθησης



Εικόνα 5: Κατηγορίες μηχανικής μάθησης με τους κυριότερους αλγορίθμους (javatpoint, n.d.)

(javatpoint, n.d.)

Οι κατηγορίες μηχανικής μάθησης είναι η εποπτευόμενη (supervised) και μη εποπτευόμενη μηχανική μάθηση (unsupervised), ενισχυτική ημι-εποπτευόμενη (semi-supervised) και η ενισχυτική (reinforcement) μάθηση.

(Υποκεφάλαιο 4.3) Μη εποπτευόμενη μάθηση

Ως είσοδος χρησιμοποιείται ένα σύνολο δεδομένων που δεν έχουν επισημανθεί, ταξινομηθεί ή κατηγοριοποιηθεί και ο αλγόριθμος πρέπει να ενεργήσει σε αυτά τα δεδομένα χωρίς καμία επίβλεψη. Ο σκοπός είναι η αναδιάρθρωση των δεδομένων εισόδου (πχ ο διαχωρισμός εικόνων σε ομάδες αντικειμένων με παρόμοια μοτίβα). Στην μάθηση χωρίς επίβλεψη, δεν έχουμε προκαθορισμένες κλάσεις. Ο αλγόριθμος προσπαθεί να βρει χρήσιμες πληροφορίες από τον τεράστιο όγκο δεδομένων. Μπορεί περαιτέρω να ταξινομηθεί σε δύο κατηγορίες:

Συσταδοποίηση ή ομαδοποίηση (clustering): Σε αυτόν τον τύπο έχουμε άγνωστο αριθμό κλάσεων και έχουμε διαθέσιμα αντικείμενα για τα οποία δεν είναι γνωστή οποιαδήποτε πληροφορία σχετική με την κλάση (ομάδα) στην οποία ανήκουν. Η καταχώρηση δειγμάτων στην ίδια ομάδα μεταφράζεται ως ομοιότητα των αντικειμένων αυτών και αντίστροφα.

Συσχέτιση (Association): Ελέγχει την εξάρτηση ενός στοιχείου δεδομένων σε ένα άλλο στοιχείο δεδομένων και προσπαθεί να ανακαλύψει κάποιες ενδιαφέρουσες σχέσεις ή συσχετισμούς μεταξύ των μεταβλητών του συνόλου δεδομένων.

(Υποκεφάλαιο 4.4) Εποπτευόμενη μάθηση

(Ενότητα 4.4.α) Γενικά

Η εποπτευόμενη μάθηση είναι ένας τύπος μηχανικής μάθησης κατά τον οποίο παρέχουμε στο σύστημα μηχανικής μάθησης επισημειωμένα δείγματα δεδομένων εκμάθησης προκειμένου να το εκπαιδεύσουμε και σε αυτή τη βάση, να μπορεί να προβλέπει στην έξοδο την κλάση στην οποία ανήκει κάποιο άλλο άγνωστο δείγμα. Στην εποπτευόμενη μάθηση έχουμε γνωστό αριθμό κλάσεων και διαθέσιμα αντικείμενα για τα οποία είναι γνωστή η κλάση στην οποία ανήκουν (επισημειωμένα), δηλαδή γνωρίζουμε τα δεδομένα εισόδου και τις

πιθανές τιμές των εξόδων του αλγορίθμου. Χωρίζετε σε δύο κατηγορίες: Την ταξινόμηση (classification) και την παλινδρόμηση (regression).

Η ταξινόμηση αφορά την πρόβλεψη μιας κατηγορίας (ομάδας) ενώ η παλινδρόμηση αφορά την πρόβλεψη μιας ποσότητας. Η ταξινόμηση έχει έξοδο διακριτές τιμές ενώ η παλινδρόμηση συνεχείς.

(Ενότητα 4.4.β) Παλινδρόμηση

Οι αλγόριθμοι παλινδρόμησης προβλέπουν μία ποσότητα (ύψος, βάρος, θερμοκρασία κλπ.), άρα έχουν έξοδο συνεχείς τιμές, μπορούν όμως να έχουν είσοδο είτε συνεχείς είτε διακριτές τιμές. Ένα πρόβλημα με πολλαπλές μεταβλητές εισόδου ονομάζεται πρόβλημα πολλαπλής παλινδρόμησης (multivariate regression). Ένα πρόβλημα παλινδρόμησης, όπου οι μεταβλητές εισόδου ταξινομούνται στο χρόνο ονομάζεται πρόβλημα πρόβλεψης χρονοσειρών (time series forecasting problem).

(Ενότητα 4.4.γ) Ταξινόμηση

«Στην ταξινόμηση (classification) ένας αλγόριθμος εκπαιδεύεται σε ένα σύνολο δεδομένων εκπαίδευσης και, κατηγοριοποιεί τα δεδομένα σε προκαθορισμένες κλάσεις. Ένα πρόβλημα με δύο κολάσεις (πχ «ναι», «όχι» ονομάζεται πρόβλημα δυαδικής (binary) ταξινόμησης. Ένα πρόβλημα με περισσότερες από δύο κλάσεις ονομάζεται πρόβλημα ταξινόμησης πολλαπλών κλάσεων (multi-class). Ένα πρόβλημα όπου ένα δείγμα έχει εκχωρηθεί σε πολλές κλάσεις ονομάζεται πρόβλημα πολλαπλής ταξινόμησης ή πολλαπλών επισημειώσεων (multilabel classification).

Η μη ισορροπημένη (unbalanced) ταξινόμηση αναφέρεται στις περιπτώσεις όπου ο αριθμός των δειγμάτων σε κάθε κλάση είναι άνισος.

Οι κυριότερες κατηγορίες αλγορίθμων:[28]

Naive Bayes: είναι μια οικογένεια απλών γραμμικών «πιθανολογικών ταξινομητών» που βασίζονται στην εφαρμογή του θεωρήματος του Bayes.

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM: είναι γραμμικά εποπτευόμενα μοντέλα μάθησης με αλγόριθμους που μπορούν να κάνουν και ταξινόμηση και παλινδρόμηση. Διαφέρουν από άλλους αλγόριθμους ταξινόμησης λόγω του τρόπου με τον οποίο επιλέγουν το

[28] https://scikit-learn.org/stable/supervised_learning.html υπάρχει λίστα με όλους τους αλγορίθμους

όριο απόφασης που μεγιστοποιεί την απόσταση από τα πλησιέστερα σημεία δεδομένων όλων των κλάσεων.

Λογιστική παλινδρόμηση (Logistic Regression): είναι μια διαδικασία μοντελοποίησης της πιθανότητας ενός διακριτού αποτελέσματος δεδομένης μιας μεταβλητής εισόδου. Η απλή λογιστική παλινδρόμηση διαμορφώνει ένα δυαδικό αποτέλεσμα. Η πολυωνυμική λογιστική παλινδρόμηση μπορεί να μοντελοποιήσει σενάρια όπου υπάρχουν περισσότερα από δύο πιθανά διακριτά αποτελέσματα.

k-πλησιέστεροι γείτονες (k-NN): είναι ένας μη παραμετρικός εποπτευόμενος ταξινομητής, ο οποίος χρησιμοποιεί την εγγύτητα για να κάνει ταξινομήσεις ή προβλέψεις σχετικά με την κλάση ενός μεμονωμένου σημείου δεδομένων.

Δένδρο απόφασης (Decision Tree): είναι ένας ταξινομητής σε μορφή δέντρου, όπου οι εσωτερικοί κόμβοι αντιπροσωπεύουν τα χαρακτηριστικά ενός συνόλου δεδομένων, οι κλάδοι αντιπροσωπεύουν τους κανόνες απόφασης και τα φύλλα αντιπροσωπεύουν το αποτέλεσμα. Κατά την εκπαίδευση ενός δέντρου απόφασης, το κύριο ζήτημα που προκύπτει είναι ότι πώς να επιλέξουμε το καλύτερο χαρακτηριστικό για τον ριζικό κόμβο και για τους υποκόμβους. Για την επιλογή χρησιμοποιούνται τεχνικές όπως το κέρδος πληροφοριών που βασίζεται στις τιμές των δεικτών εντροπίας και Gini και σύμφωνα με τον αλγόριθμο Classification and Regression Tree Algorithm(CART)[29].

Μέθοδοι συνόλων (ensemble): είναι τεχνικές που δημιουργούν πολλαπλά μοντέλα που στη συνέχεια συνδυάζονται για να παράγουν βελτιωμένα αποτελέσματα. Οι μέθοδοι συνόλων παράγουν συνήθως πιο ακριβείς λύσεις από ό, τι ένα μόνο μοντέλο. Το πιο γνωστό είναι το «**τυχαίο δάσος» (Random Forest)** που λειτουργεί κατασκευάζοντας κατά το χρόνο εκπαίδευσης ένα πλήθος δέντρων αποφάσεων και εξάγοντας την κλάση που είναι ο διάμεσος των κλάσεων (για ταξινόμηση) ή μέση πρόβλεψη (για παλινδρόμηση).

Μέθοδος προσαρμοστικής ενίσχυσης (Adaptive Boosting) – μέθοδος ενίσχυσης κλίσης (Gradient Boosting): Η βασική αρχή του boosting

[29] Λεπτομέρειες στο https://towardsdatascience.com/gini-index-vs-information-entropy-7a7e4fed3fcb

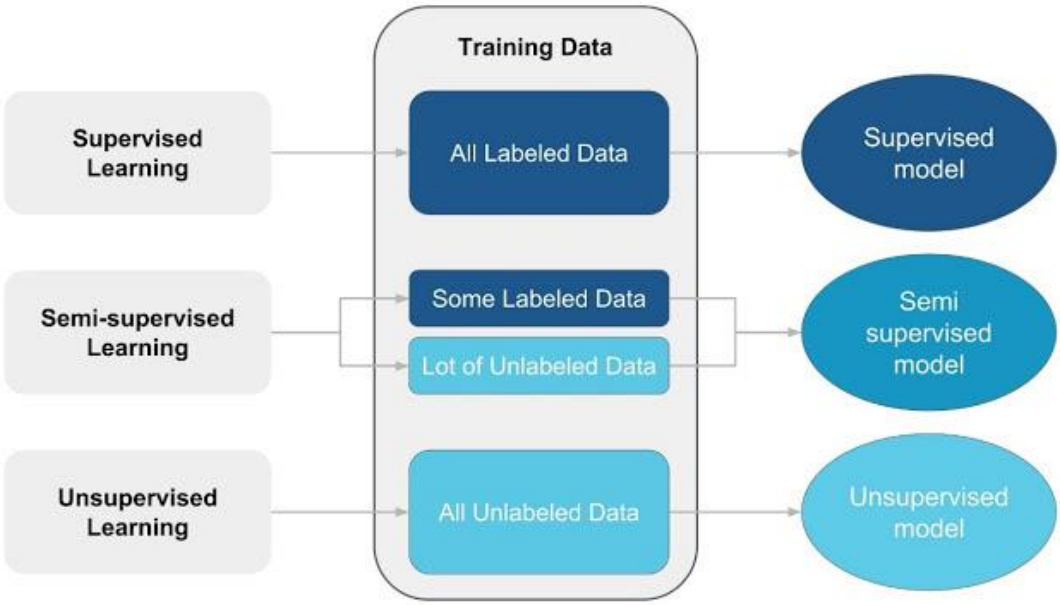
είναι η προσαρμογή μιας σειράς αδύναμων μαθητών[30] σε επανειλημμένα τροποποιημένες εκδόσεις των δεδομένων μετατρέποντας σταδιακά το μοντέλο σε ισχυρό μαθητή. «Η τεχνική του boosting χρησιμοποιεί διάφορες λειτουργίες απώλειας. Στην περίπτωση της Adaptive Boosting ή του AdaBoost, ελαχιστοποιεί τη συνάρτηση «εκθετικής απώλειας» που μπορεί να κάνει τον αλγόριθμο ευαίσθητο στις ακραίες τιμές. Με το Gradient Boosting, μπορεί να χρησιμοποιηθεί οποιαδήποτε διαφορίσημη συνάρτηση απώλειας. Ο αλγόριθμος Gradient Boosting είναι πιο ισχυρός σε ακραίες τιμές από το AdaBoost». (Choudhury, 2021).

[30] Μοντέλα που είναι ελαφρώς καλύτερα από την τυχαία εικασία, όπως μικρά δέντρα αποφάσεων.

ΚΕΦΑΛΑΙΟ 5 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΜΕ ΜΕΡΙΚΩΣ ΕΠΙΣΗΜΕΙΩΜΕΝΑ ΔΕΔΟΜΕΝΑ

(Υποκεφάλαιο 5.1) Γενικά - παραδοχές

Η ουσιαστική διάκριση μεταξύ της εποπτευόμενης και μη εποπτευόμενης μάθησης είναι ότι όλα τα δείγματα δεδομένων που χρησιμοποιούνται στην εποπτευόμενη μάθηση είναι επισημειωμένα, ενώ στην μη εποπτευόμενη δεν έχουμε κανένα επισημειωμένο δείγμα.



Εικόνα 6: Εποπτευόμενη - μη εποπτευόμενη – ημι-εποπτευόμενη μάθηση.
(Ibañez, 2019)

Οι εποπτευόμενοι αλγόριθμοι μηχανικής μάθησης απαιτούν το σύνολο των δεδομένων να έχει επισημειωθεί προσωπικά από έναν ειδικό. Αυτή η διαδικασία έχει υπερβολικό κόστος, ιδιαίτερα κατά τη διαχείριση τεράστιων συνόλων δεδομένων. Εκτός όμως από το κόστος, ορισμένες φορές είναι αδύνατον να πραγματοποιηθεί η επισημείωση.

Από την άλλη πλευρά το εύρος εφαρμογής της μη εποπτευόμενης μάθησης είναι περιορισμένο, ιδιαίτερα στον ιατρικό-διαγνωστικό τομέα. Δηλαδή, σπάνια θα ζητηθεί να χωρίσουμε τους ασθενείς σε άγνωστες ομάδες. Αυτό που θα ζητηθεί είναι, εάν κάποιος πάσχει από συγκεκριμένη νόσο ή όχι, η παρόμοια ερωτήματα πχ “COVID-19, πνευμονία, χωρίς ευρήματα κλπ.».

Επιπλέον τα περισσότερα προβλήματα που αφορούν την ιατρική διάγνωση αφορούν ταξινόμηση και όχι παλινδρόμηση.

Για την αντιμετώπιση αυτών των θεμάτων προτάθηκε η ανάπτυξη αλγορίθμων ταξινόμησης που να μπορούν να έχουν αξιόπιστα αποτελέσματα με λίγα επισημειωμένα δεδομένα. Αυτούς τους αλγόριθμους μπορούμε να τους ταξινομήσουμε σε 3 κατηγορίες:

Ημι-εποτευόμενη μάθηση: Επισημείωση των μη επισημειωμένων δεδομένων με αυτοματοποιημένο τρόπο.

Ενεργή μάθηση (Active learning): Επιλογή των καταλληλότερων δεδομένων για χειροκίνητη επισημείωση.

Μάθηση από θετικά και μη επισημειωμένα δεδομένα (positive unlabeled-PU learning).. Σε αυτήν την περίπτωση έχουμε μία ομάδα με θετικά και μια πολύ μεγαλύτερη ομάδα με μη επισημειωμένα δεδομένα. Και αυτή όπως και η ημι-εποπτευόμενη μάθηση προσπαθεί να επισημειώσει τα δεδομένα με αυτοματοποιημένο τρόπο.

Σκοπός και των τριών είναι μετά την εκπαίδευση να πέτυχουν μεγαλύτερο βαθμό αξιοπιστίας με όσον το δυνατόν μικρότερο αριθμό πραγματικά επισημειωμένων δεδομένων. Στον πρώτο και τρίτο τρόπο, που η επισημείωση γίνεται αυτοματοποιημένα, επιλέγουμε σε κάθε επανάληψη του αλγορίθμου να επισημειώσουμε αυτά που έχουν μεγαλύτερη πιθανότητα να ανήκουν σε κάποια κλάση, ενώ στην δεύτερη περίπτωση που η επισημείωση γίνεται χειροκίνητα, επιλέγουμε προς επισημείωση αυτά που έχουν τον μεγαλύτερο βαθμό αβεβαιότητας ως προς την επισημείωση που πρέπει να φέρουν.

«Για την μάθηση από μη επισημειωμένα δεδομένα, πρέπει να είναι γνωστή κάποια σχέση που προσδιορίζει την υποκείμενη κατανομή των δεδομένων. Δηλαδή να υπάρχει κάποιος τρόπος εξαγωγής της επισημείωσης των μη επισημειωμένων δεδομένων με βάση τα δεδομένα που είναι ήδη επισημειωμένα.

Οι αλγόριθμοι αυτοί χρησιμοποιούν τουλάχιστον μία από τις ακόλουθες παραδοχές:

«Παραδοχή συνέχειας / ομαλότητας: Τα σημεία που είναι κοντά το ένα στο άλλο είναι πιο πιθανό να έχουν την ίδια επισημείωση. Αυτό ισχύει γενικά και στην εποπτευόμενη μάθηση όπου υπάρχει μια προτίμηση στα απλά γεωμετρικά όρια απόφασης. Στην περίπτωση των αλγορίθμων μάθησης με

μερικώς επισημειωμένα δεδομένα, η υπόθεση ομαλότητας δείχνει μια προτίμηση για όρια απόφασης σε περιοχές χαμηλής πυκνότητας. Έτσι όταν λίγα σημεία βρίσκονται κοντά το ένα στο άλλο, αποδίδονται σε διαφορετικές κλάσεις.

Παραδοχή συστάδας: Τα δεδομένα τείνουν να σχηματίζουν διακριτά συμπλέγματα, τότε τα σημεία στο ίδιο σύμπλεγμα είναι πιο πιθανό να έχουν την ίδια επισημείωση. Αυτή είναι μια ειδική περίπτωση της υπόθεσης της ομαλότητας και οδηγεί στη μάθηση χαρακτηριστικών με αλγόριθμους ομαδοποίησης (clustering).

Πολύπτυχη (manifold) παραδοχή: Τα δεδομένα βρίσκονται περίπου σε ένα πολύπτυχο πολύ μικρότερης διάστασης από τον χώρο των δεδομένων εισόδου. Δηλαδή ενώ αρχικά φαίνεται ότι τα δεδομένα απαιτούν πολλές μεταβλητές για να περιγραφούν, τελικά χρειάζονται πολύ λιγότερες. Η πολύπτυχη υπόθεση εφαρμόζεται όταν από κάποια διαδικασία παράγονται δεδομένα υψηλών διαστάσεων με λίγους βαθμούς ελευθερίας ,τα οποία μπορεί να είναι δύσκολο να μοντελοποιηθούν άμεσα. Για παράδειγμα, η ανθρώπινη φωνή ελέγχεται από μερικές φωνητικές χορδές και οι εικόνες διαφόρων εκφράσεων του προσώπου ελέγχονται από λίγους μύες. Σε αυτές τις περιπτώσεις, είναι προτιμότερο να λαμβάνονται υπόψη οι αποστάσεις και η ομαλότητα στο φυσικό χώρο του προβλήματος παρά στο χώρο όλων των πιθανών ακουστικών κυμάτων ή εικόνων, αντίστοιχα.» (Semi-supervised learning, 2022).

(Υποκεφάλαιο 5.2) Ημι-εποπτευόμενη μάθηση

(Ενότητα 5.2.α) Γενικά

Σε αυτήν την περίπτωση έχουμε λίγα επισημειωμένα δεδομένα, και πολλά δείγματα τα οποία δεν γνωρίζουμε σε ποια κατηγορία υπάγονται. Η προσπάθειά μας έγκειται στο να ταξινομήσουμε αυτόματα όσα χρειαζόμαστε από τα μη ταξινομημένα δείγματα. Δηλαδή με την βοήθεια των δεδομένων που έχουμε να εντάξουμε τα πιο «σίγουρα» στην μία ή την άλλη κλάση. Όταν λέμε πιο «σίγουρα» εννοούμε αυτά που έχουν μεγαλύτερη πιθανότητα να ανήκουν στην μία κλάση από την άλλη.

Τα μη επισημειωμένα δεδομένα, όταν χρησιμοποιούνται σε συνδυασμό με μια μικρή ποσότητα επισημειωμένων, μπορούν να προκαλέσουν σημαντική βελτίωση στην αποτελεσματικότητα της εκπαίδευσης. Η ημι-εποπτευόμενη

μάθηση συνδυάζει αυτές τις πληροφορίες για να ξεπεράσει την απόδοση της ταξινόμησης που μπορεί να επιτευχθεί απορρίπτοντας τα μη επισημειωμένα δεδομένα και πραγματοποιώντας εποπτευόμενη μάθηση με μόνο τα επισημειωμένα.

Η ημι-εποπτευόμενη μάθηση μπορεί να αναφέρεται, είτε σε μεταγωγική μάθηση (transudative) είτε σε επαγωγική μάθηση (inductive). Ο στόχος της μεταγωγικής μάθησης είναι να συμπεράνει απευθείας τις σωστές επισημειώσεις για τα δεδομένα στα οποία αυτές δεν υπάρχουν. Ο στόχος της επαγωγικής μάθησης είναι να συμπεράνουμε τη σχεση μεταξύ των δεδομένων και των κλάσεων και μέσω αυτής να γίνει η επισημείωση.

Η διαδικασία της ημι-εποτευόμενης μάθησης γίνεται επαναληπτικά μέχρις ότου επιτευχθούν οι συνθήκες τερματισμού. Αυτές μπορεί να είναι: ολοκλήρωση του μέγιστου αριθμού επαναλήψεων που έχουν τεθεί, ολοκλήρωση των επισημειώσεων των δεδομένων που περνάνε το κατώφλι πιθανότητας να ανήκουν σε κάποια κλάση, έχει επιτευχθεί η τιμή της μετρικής που επιθυμούμε, ή έχουν επισημειωθεί όλα.

(Ενότητα 5.2.β) Κατηγορίες ημι-εποπτευόμενης Μάθησης

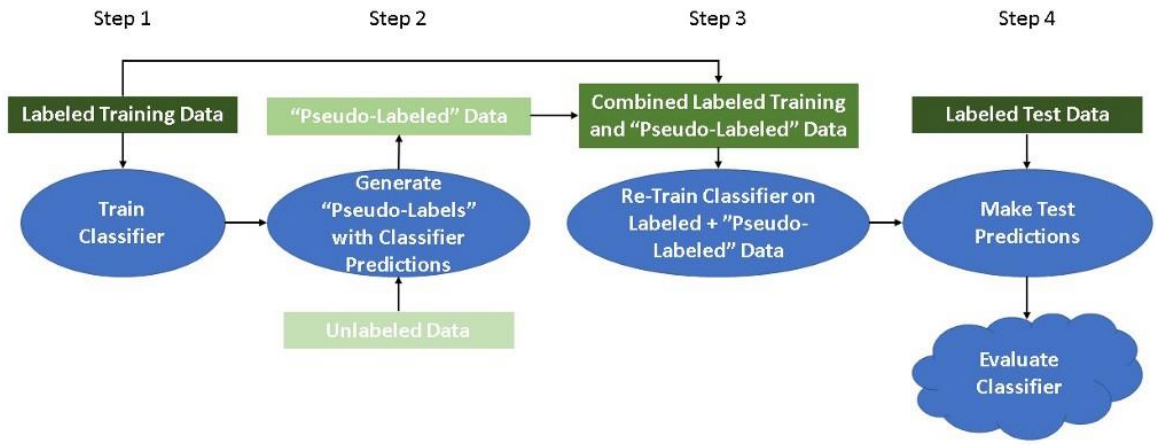
Αυτό-εκπαίδευση

Είναι μια μέθοδος επαναδειγματοληψίας που κατασκευάζει επισημειώσεις σε μη επισημειωμένα δείγματα. Θεωρείται επαγωγική (inductive), διότι τις παρατηρούμενες περιπτώσεις εκπαίδευσης εξάγονται γενικοί κανόνες, οι οποίοι στη συνέχεια εφαρμόζονται στα μη επισημειωμένα δεδομένα.

Η διαδικασία που ακολουθείται είναι:

1. Εκπαίδευση μοντέλου με το σύνολο των επισημειωμένων δεδομένων με βάση κάποιον αλγόριθμο εποπτευόμενης μάθησης.
2. Επισημείωση των μη επισημειωμένων δεδομένων με βάση της προβλέψεις που κάνουμε με το εκπαιδευμένο μοντέλο. Τα επισημειωμένα με αυτόν τον τρόπο ονομάζονται ψευδοεπισημειωμένα.
3. Προσθήκη ορισμένων από τα ψευδοεπισημειωμένα, με βάση την βεβαιότητα περί της ορθότητας της ψευδοεπισημείωσης, στο σύνολο των επισημειωμένων.
4. Επανεκπαίδευση του μοντέλου με βάση το νέο σύνολο των επισημειωμένων δεδομένων.

5. Επαναλαμβάνεται η διαδικασία από το βήμα 3 και κάτω ορισμένες φορές (συνήθως 10) ή μέχρι να επιτευχθούν οι συνθήκες τερματισμού που αναφέρθηκαν παραπάνω.



Εικόνα 7 Λογικό διάγραμμα self-training
(Steen, 2020)

Βασισμένη σε γράφους

Βασισμένη σε γράφους ή διάδοση επισημειώσεων (Graph-based ή label propagation ή Trunsductive semi - supervised machine learning): Αποδεχόμαστε ότι τα δεδομένα (τόσο με επισημείωση όσο και χωρίς επισημείωση) εισάγονται σε ένα σύμπλεγμα χαμηλής διάστασης που μπορεί να μεταδοθεί λογικά από ένα γράφημα. Όλα τα δείγματα δεδομένων αντιπροσωπεύονται από μια κορυφή σε ένα σταθμισμένο γράφημα, με τα βάρη να δίνουν την αναλογία εγγύτητας μεταξύ των κορυφών. Η μέθοδος θεωρείται μεταγωγική (transductive), καθώς προσπαθούμε να προβλέψουμε απευθείας επισημειώσεις από δεδομένα που μας έχουν δοθεί. Η υιοθέτηση μιας στρατηγικής βασισμένης σε γραφήματα περιλαμβάνει τα ακόλουθα βήματα:

i. Ανάπτυξη γραφήματος (αν δεν υπάρχει γράφημα πληροφοριών),
ii. Σημείωση των κόμβων που είναι επισημειωμένοι,
iii. Συμπερασματικές επισημειώσεις στους κόμβους χωρίς επισημείωση στο γράφημα.

Πολλές συλλογές δεδομένων συνήθως αντιπροσωπεύονται εξ αρχής από ένα γράφο, πχ τα κοινωνικά δίκτυα. Υπάρχουν όμως και περιπτώσεις που δεν είναι προφανής η αντιστοίχιση με γράφο. Τότε μπορεί να κατασκευαστεί ένας γράφος χρησιμοποιώντας γνώσεις του αντικειμένου, ομοιότητα δειγμάτων δεδομένων κλπ. Χρησιμοποιούνται κυρίως δύο (2) μέθοδοι, η σύνδεση κάθε σημείου δεδομένων με τους k πλησιέστερους γείτονές του ή με την απόσταση ε. από επισημειωμένους κόμβους. Το βάρος κάθε ακμής του γράφου υπολογίζεται με τον τύπο:

$$e^{\frac{-\|x_i - x_j\|^2}{\epsilon}}$$

Εάν όλα τα βάρη είναι ίσα με την μονάδα, η θεωρούνται ισοδύναμα τότε θεωρούμε ότι όλες οι διαδρομές (ακμές) έχουν το ίδιο βάρος (1).

Ακολούθως εξετάζουμε το γράφο που έχουμε δημιουργήσει πχ όπως αυτό στο παρακάτω σχήμα, όπου έχουμε δύο κατηγορίες επισημειώσεων (κόκκινο και πράσινο) και 4τέσσερεις χρωματισμένους κόμβους (δύο για κάθε κατηγορία). Έστω ότι θέλουμε να προβλέψουμε την επισημείωση του κόμβου 4.

Μπορούμε να περπατήσουμε τυχαία στο γράφο της παρακάτω εικόνας, ξεκινώντας από τον κόμβο 4 μέχρι να συναντήσουμε οποιονδήποτε κόμβο με επισημείωση. Όταν συναντάμε έναν κόμβο με επισημείωση, σταματάμε . Ας εξετάσουμε όλες τις πιθανές διαδρομές (walks)[31] από τον κόμβο 4 που καταλήγουν σε έναν επισημειωμένο κόμβο. Έτσι έχουμε διαδρομές που καταλήγουν:

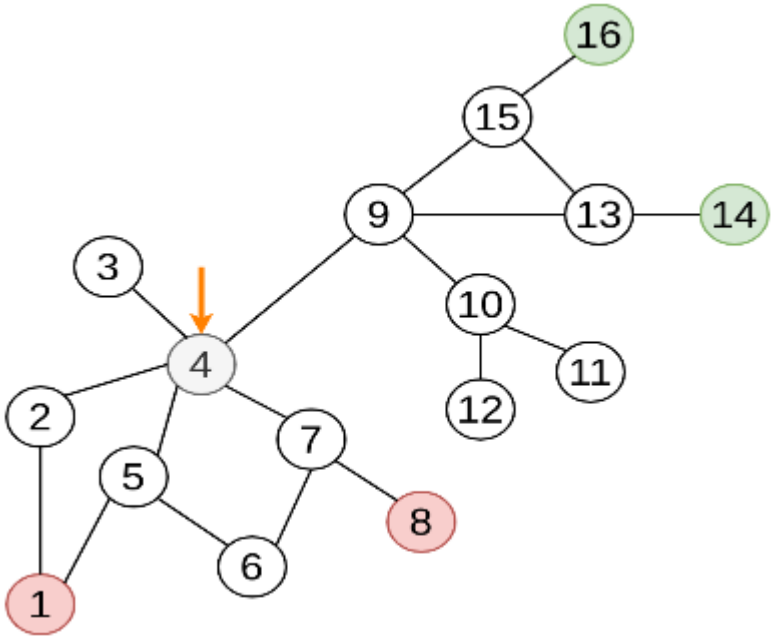
πράσινο κόμβο:
1. 4 → 9 → 15 → 16
2. 4 → 9 → 13 → 14
3. 4 → 9 → 13 → 15 → 16
4. 4 → 9 → 15 → 13 → 14

ενώ σε κόκκινο κομβο:
1. 4 → 7 → 8
2. 4 → 7 → 6 → 5 → 1
3. 4 → 5 → 1
4. 4 → 5 → 6 → 7 → 8

[31] Η *διαδρομή σε ένα γράφημα είναι μια ακολουθία ακμών και κορυφών. Όταν έχουμε ένα γράφημα και το διασχίσουμε, τότε αυτή η τραβέρσα θα είναι γνωστή ως περίπατος. Σε ένα περίπατο, μπορεί να υπάρχουν επαναλαμβανόμενες ακμές και κορυφές. Ο αριθμός των άκμων που καλύπτονται λέγεται το μήκος της διαδρομής.*

5. 4 → 2 → 1

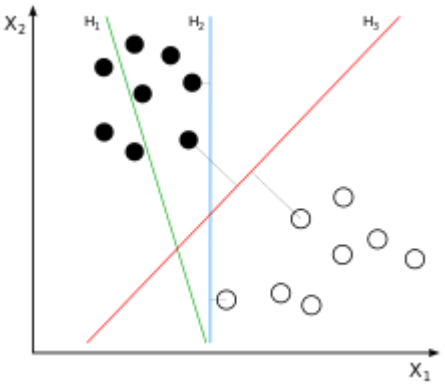
Με βάση όλους τους πιθανούς περιπάτους ξεκινώντας από τον κόμβο 4, μπορούμε να δούμε ότι η πλειονότητα των περιπάτων καταλήγει σε έναν κόκκινο κόμβο. Έτσι, μπορούμε να χρωματίσουμε τον κόμβο 4 με κόκκινο.



Εικόνα 8 Γράφος ημι-εποπτευόμενης μάθησης
(Mallawaarachchi, Label Propagation Demystified, 2020)

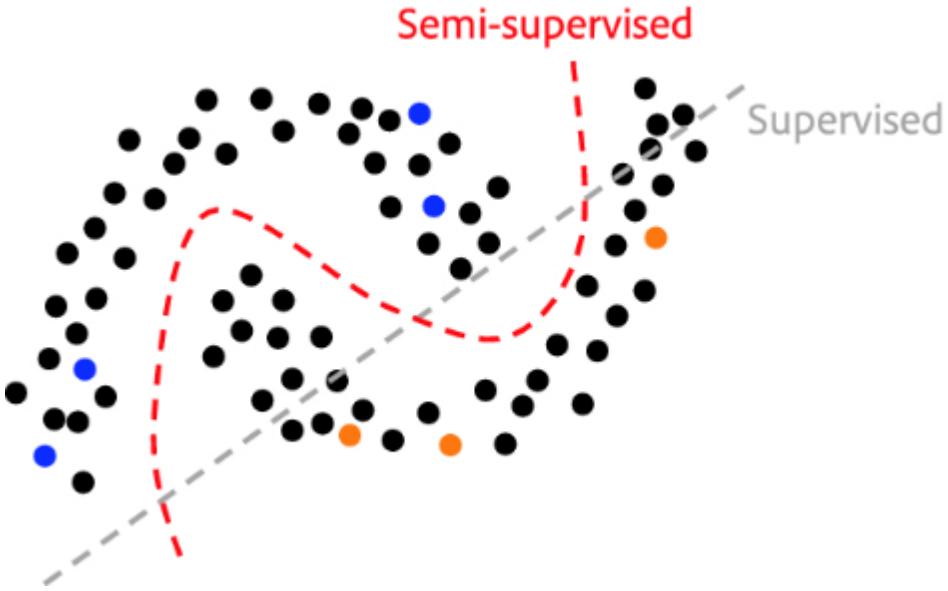
Διαχωρισμός χαμηλής πυκνότητας

Μια άλλη μεγάλη κατηγορία μεθόδων επιχειρεί να τοποθετήσει όρια σε περιοχές με λίγα σημεία δεδομένων (επισημειωμένα ή μη επισημειωμένα). Ένας από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους είναι η «μηχανή διανυσμάτων υποστήριξης μεταγωγής» ή TSVM (Transductive Support Vector Machne, (η οποία, παρά το όνομά της, μπορεί να χρησιμοποιηθεί και για επαγωγική (inductive) μάθηση). Ενώ οι μηχανές υποστήριξης διανυσμάτων (SVM) στην εποπτευόμενη μάθηση αναζητούν ένα όριο απόφασης με μέγιστο περιθώριο πάνω από τα επισημειωμένα δεδομένα, ο στόχος του TSVM είναι η επισημείωση των μη επισημειωμένων δεδομένων έτσι ώστε το όριο απόφασης να έχει μέγιστο περιθώριο για όλα τα δεδομένα (επισημειωμένα και μη).



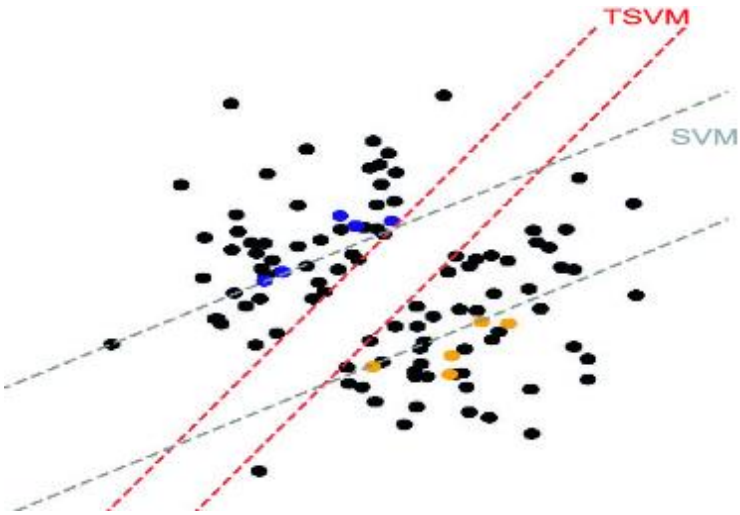
Εικόνα 9: Διαχωρισμός SVM

Στο παραπάνω σχήμα SVM, το H1 δεν χωρίζει τις τάξεις. Το H2 το κάνει, αλλά μόνο με ένα μικρό περιθώριο. Το H3 τα χωρίζει με το μέγιστο περιθώριο.



Εικόνα 10 Μέθοδος χαμηλής πυκνότητας

Η γκρίζα γραμμή αντιστοιχεί σε ένα όριο απόφασης που λαμβάνεται από έναν εποπτευόμενο ταξινομητή ο οποίος ενσωματώνει πληροφορίες μόνο από τα επισημειωμένα μπλε και πορτοκαλί σημεί). Η κόκκινη γραμμή αντιστοιχεί σε ένα όριο από μια ημι-εποπτευόμενη μέθοδο που αναζητά ένα όριο απόφασης χαμηλής πυκνότητας. Είναι μια ζώνη που περνάει από εκεί όπου τα δεδομένα είναι «αραιά»



Εικόνα 11 Ομαδοποίηση TVSM

Οι γκρίζες γραμμές αντιστοιχούν στον μέγιστο διαχωρισμό περιθωρίων για επισημειωμένα δεδομένα χρησιμοποιώντας ένα τυπικό SVM ενώ οι κόκκινες γραμμές αντιστοιχούν σε διαχωρισμό με TSVM.

(Burkhart & Shan, 2020)

Με την βοήθεια νευρωνικών δικτύων

Τόσο η βασισμένη σε γράφο όσο και η προσέγγιση διαχωρισμού χαμηλής πυκνότητας, βασίζονται στη γεωμετρία του χώρου των χαρακτηριστικών παρέχοντας μια λογική προσέγγιση στα πραγματικά χαρακτηριστικά των αντικειμένων. Καθώς τα σύνολα δεδομένων γίνονται όλο και πιο περίπλοκα και υψηλών διαστάσεων, η Ευκλείδεια απόσταση μεταξύ των διανυσμάτων των χαρακτηριστικών δεν είναι ο καλύτερος διακομιστής της σχέσης μεταξύ των δεδομένων και των επισημειώσεων τους. Για τον λόγο αυτό οι εφαρμογές που βασίζονται σε νευρωνικά δίκτυα έχουν γίνει πολύ δημοφιλείς τα τελευταία χρόνια. Ωστόσο, η βελτιστοποίηση των υπερπαραμέτρων των νευρωσικών δικτυών απαιτεί χειρισμό από πολύ ειδικούς ή/και έχει απαγορευτικό υπολογιστικό κόστος.

(Burkhart & Shan, 2020)

(Υποκεφάλαιο 5.3) Ενεργή μάθηση

«Η ενεργή μάθηση (active learning) είναι μια ειδική περίπτωση μηχανικής μάθησης στην οποία ένας αλγόριθμος εκμάθησης μπορεί να ρωτήσει αλληλεπιδραστικά έναν χρήστη (ή κάποια άλλη πηγή πληροφοριών) για να επισημάνει νέα σημεία δεδομένων με τις επιθυμητές εξόδους. Στη βιβλιογραφία

της στατιστικής, μερικές φορές ονομάζεται επίσης βέλτιστος πειραματικός σχεδιασμός. Η πηγή πληροφοριών ονομάζεται δάσκαλος ή μάντης» (teacher or oracle).

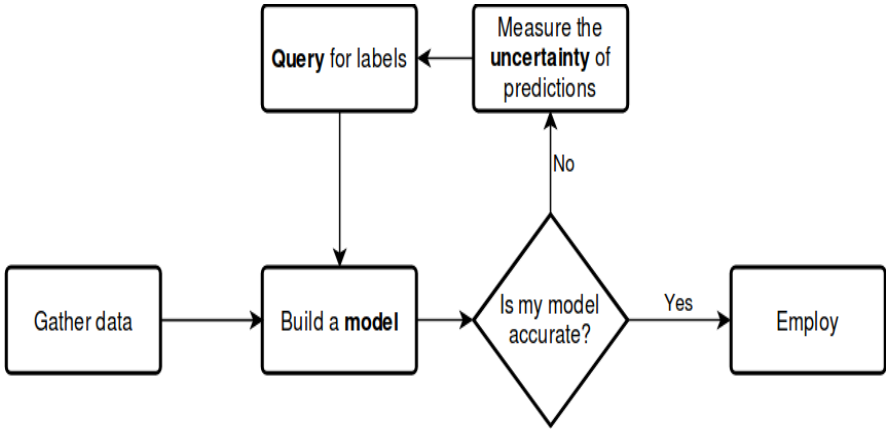
Υπάρχουν περιπτώσεις στις οποίες τα δεδομένα χωρίς επισημείωση είναι άφθονα, αλλά η χειροκίνητη επισημείωση κοστίζει. Σε ένα τέτοιο σενάριο, οι αλγόριθμοι εκμάθησης μπορούν να ρωτήσουν ενεργά τον χρήστη/δάσκαλο για επισημειώσεις. Αυτός ο τύπος επαναληπτικής εποπτευόμενης μάθησης ονομάζεται ενεργή μάθηση. Εφόσον ο εκπαιδευόμενος (αλγόριθμος-μοντέλο) επιλέγει τα δεδομένα που θα χρησιμοποιηθούν, ο αριθμός των απαιτούμενων δειγμάτων για την εκπαίδευση μπορεί να είναι πολύ μικρότερος από τον αριθμό που απαιτείται στην κανονική εποπτευόμενη μάθηση όπου η επιλογή των δειγμάτων για επισημείωση έχει γίνει τυχαία. Η ενεργή μάθηση είναι ένα πλαίσιο που μας επιτρέπει να αυξήσουμε την απόδοση του μοντέλου ταξινόμησης ζητώντας με «έξυπνα ερωτήματα» να επισημάνουμε τις πιο κατάλληλες περιπτώσεις. Οι πιο κατάλληλες περιπτώσεις, στην περίπτωση αυτή, είναι τα δεδομένα που έχουν την μεγαλύτερη αβεβαιότητα ως προς την σωστή ταξινόμηση.

Τα βασικά στοιχεία οποιασδήποτε ροής εργασιών ενεργούς μάθησης είναι: το μοντέλο που επιλέγουμε για να κάνουμε την ταξινόμηση, το μέτρο που χρησιμοποιούμε για την επιλογή των δεδομένων που θα σταλούν προς επισημείωση και η στρατηγική ερωτημάτων που εφαρμόζουμε για να επιλέξουμε τα δεδομένα προς επισημείωση.

Ας υποθέσουμε ότι μπορούμε να στείλουμε προς επισημείωση ένα δείγμα αλλά αυτό κοστίζει πολύ. Για παράδειγμα έχουμε ένα μικρό σύνολο επισημειωμένων ακτινογραφιών και ένα μεγάλο αριθμό μη επισημειωμένων. Το μικρό σύνολο δεν επαρκεί για να εκπαιδεύσει ένα μοντέλο που θα δίνει αξιόπιστες προβλέψεις. Αυτό μας οδηγεί στο να εμπλουτίσουμε τα δεδομένα μας με επισημείωση και άλλων δεδομένων. Στην ταξινόμηση που θα επιχειρήσουμε να εφαρμόσουμε υπάρχουν μέτρα που μπορούν να προσδιορίσουν ποια δεδομένα έχουν ταξινομηθεί με σχετική βεβαιότητα και ποια όχι. Η αβεβαιότητα στις επισημειώσεις συνήθως υπολογίζεται συνήθως από την εντροπία ή την διακύμανση των επισημειώσεων.

Εάν επιλέξουμε να επισημειώσουμε δεδομένα που βρίσκονται στην αβέβαιη περιοχή, σίγουρα λαμβάνουμε περισσότερες πληροφορίες από το να επι-

λέξουμε τυχαία ή να προσθέσουμε δεδομένα που έχουν ταξινομηθεί με μεγάλη βεβαιότητα. Η ενεργός μάθηση μας παρέχει ένα σύνολο εργαλείων για να χειριστούμε προβλήματα όπως αυτό. Γενικά, μια ενεργός ροή εργασιών μάθησης μοιάζει με την ακόλουθη.



Εικόνα 12 Διάγραμμα ροής ενεργούς μάθησης

(Danka, Modular Active Learning framework for Python3, 2021)

« Όσον αφορά τα δεδομένα τα χωρίζουμε σε τρεις (3) κατηγορίες:

- Αυτά που είναι επισημειωμένα
- Αυτά τα οποία δεν είναι επισημειωμένα
- Αυτά που θα ζητήσουμε να επισημειωθούν

Υπάρχουν 3 σενάρια για την επιλογή των δειγμάτων που θα ζητήσουμε να επισημειωθούν:

1. Σύνθεση ερωτήματος μέλους: Εδώ το μοντέλο που εκπαιδεύεται δημιουργεί το δικό του δείγμα από μια υποκείμενη φυσική κατανομή. Για παράδειγμα, εάν το σύνολο δεδομένων είναι εικόνες ανθρώπων και ζώων, ο εκπαιδευόμενος θα μπορούσε να στείλει μια αποκομμένη εικόνα ενός ποδιού στον δάσκαλο και να ρωτήσει εάν αυτό ανήκει σε ζώο ή άνθρωπο.

2. Δειγματοληψία από τα δεδομένα που έχουμε: Σε αυτό το σενάριο, τα δεδομένα προς επισημείωση αντλούνται από ολόκληρη τη δεξαμενή δεδομένων και τους αποδίδεται μια βαθμολογία εμπιστοσύνης, μια μέτρηση του πόσο καλά το εκπαιδευόμενο μοντέλο «καταλαβαίνει» τα δεδομένα. Στη συνέχεια, το σύστημα επιλέγει τις περιπτώσεις για τις οποίες έχει τη λιγότερη βεβαιότητα για την κλάση που ανήκουν και ζητά από τον «δάσκαλο» να τις επισημειώσει.

3. Επιλεκτική δειγματοληψία βάσει ροής: Εδώ, κάθε σημείο δεδομένων χωρίς επισημείωση εξετάζεται ένα κάθε φορά, με το μοντέλο να αξιολογεί την

πληροφόρηση κάθε στοιχείου σε σχέση με τις παραμέτρους του ερωτήματός του. Το μοντέλο αποφασίζει μόνο του εάν θα εκχωρήσει μια επισημείωση ή θα ρωτήσει τον δάσκαλο για κάθε σημείο δεδομένων. Αυτή είναι μια μικτή διαδικασία.

Οι αλγόριθμοι για τον προσδιορισμό των δειγμάτων δεδομένων που πρέπει να επισημανθούν μπορούν να κατηγοριοποιηθούν με βάση τον σκοπό τους:

1. Εξερεύνηση και εκμετάλλευση ισορροπίας (balance exploration and exploitation): «Η επιλογή των δειγμάτων προς επισημείωση θεωρείται ως δίλημμα μεταξύ της εξερεύνησης και της εκμετάλλευσης σε σχέση με την αναπαράσταση του χώρου δεδομένων. Παράδειγμα, ο αλγόριθμος με το όνομα Active Tomson Sampling (ATS), ο οποίος, σε κάθε γύρο, εκχωρεί μια κατανομή δειγματοληψίας στη δεξαμενή των δεδομένων, λαμβάνει δείγματα από αυτήν την κατανομή και θέτει ερωτήματα στον «δάσκαλο» για την επισημείωση.». (Bouneffouf, Laroche, Urvoy, Féraud, & Allesiardo, 2014).

2. Αναμενόμενη αλλαγή μοντέλου: Επιλέγουμε να επισημειώσουμε τα σημεία εκείνα που θα άλλαζαν περισσότερο το τρέχον μοντέλο.

3. Αναμενόμενη μείωση σφαλμάτων: Επιλέγουμε να επισημειώσουμε τα σημεία εκείνα που θα μείωναν περισσότερο το σφάλμα γενίκευσης του μοντέλου. Δηλαδή την/τις μετρικές που δείχνουν πόσο καλά το μοντέλο αποδίδει σε άγνωστα δεδομένα δοκιμών.

4. «Εκθετική εξερεύνηση κλίσης για ενεργό μάθηση (exponentiated gradient (EG)): Επικεντρώνεται στην βελτίωση του αλγορίθμου και όχι στο ξεκαθάρισμα των ορίων μεταξύ των κατηγοριών. Ένας διαδοχικός αλγόριθμος που μπορεί να βελτιώσει οποιονδήποτε ενεργό αλγόριθμο μάθησης με μια βέλτιστη τυχαία εξερεύνηση». (Bouneffouf, Exponentiated Gradient Exploration for Active Learning, 2016).

5. Δειγματοληψία αβεβαιότητας: Αποστέλλουμε προς επισημείωση εκείνα τα δεδομένα για τα οποία το τρέχον μοντέλο είναι λιγότερο σίγουρο για το ποια θα πρέπει να είναι η σωστή επισημείωση.

6. Ερώτημα από επιτροπή: Μια ποικιλία μοντέλων εκπαιδεύεται στα τρέχοντα επισημειωμένα δεδομένα και ψηφίζει για την πιθανή επισημείωση για τα δεδομένα. Αποστέλλουμε προς επισημείωση τα δείγματα για τα οποία η «επιτροπή» διαφωνεί περισσότερο.

7. Ερώτηση από διαφορετικούς υποχώρους ή διαμερίσματα: Όταν το υποκείμενο μοντέλο είναι ένα δάσος από δέντρα, οι κόμβοι των φύλλων ενδέχεται

να αντιπροσωπεύουν επικαλυπτόμενες κατατμήσεις του αρχικού χώρου χαρακτηριστικών. Αυτό προσφέρει τη δυνατότητα επιλογής δειγμάτων από μη επικαλυπτόμενες ή ελάχιστα επικαλυπτόμενες κατατμήσεις για επισημείωση.

8. Μείωση διακύμανσης (variance): Αποστέλλουμε προς επισημείωση τα σημεία εκείνα που θα ελαχιστοποιούσαν τη διακύμανση εξόδου, η οποία είναι μία από τις συνιστώσες του σφάλματος.

9. Σύμμορφοι προγνώστες (conformal predictors): Αυτή η μέθοδος προβλέπει ότι ένα νέο δείγμα δεδομένων θα έχει μια επισημείωση παρόμοια με τα παλιά δείγματα δεδομένων σύμφωνα με τον βαθμό ομοιότητας που έχουν με τα παλιά δείγματα και υπολογίζεται με κάποιους τρόπους (πχ οι κ πλησιέστεροι γείτονες). Αυτός ο υπολογισμός οδηγεί στην εκτίμηση της εμπιστοσύνης στην πρόβλεψη. Αποστέλλονται για επισημείωση αυτά για τα οποία υπάρχει μικρότερη εμπιστοσύνη για την πρόβλεψη που έγινε.

10. Αναντιστοιχία πρώτης πιο απομακρυσμένης διάβασης (Mismatch-first farthest-traversal) : Το κύριο κριτήριο επιλογής είναι η αναντιστοιχία πρόβλεψης μεταξύ του τρέχοντος μοντέλου και της πρόβλεψης του πλησιέστερου γείτονα. Στοχεύει σε λανθασμένες προβλέψεις. Το δεύτερο κριτήριο επιλογής είναι η απόσταση από τα επισημειωμένα δεδομένα. Στοχεύει στη βελτιστοποίηση της ποικιλομορφίας των επιλεγμένων δεδομένων.

11. Στρατηγικές επισημείωσης με επίκεντρο τον χρήστη: Η μάθηση επιτυγχάνεται με την εφαρμογή μείωσης διαστάσεων σε γραφήματα και σχήματα όπως διαγράμματα διασποράς. Στη συνέχεια, ο χρήστης καλείται να επισημειώσει τα δεδομένα που συμμορφώνονται σε διάφορες καταστάσεις (κατηγορικά, αριθμητικά, βαθμολογίες συνάφειας, σχέση μεταξύ δύο περιπτώσεων).» (Towards AI, 2021)

Η διαφορά της ενεργούς μάθησης με τις άλλες μεθόδους είναι ότι η επισημείωσή θα γίνει από τον άνθρωπο, δηλαδή έχει κόστος. Σκοπός μας είναι να χρειαστούμε να επισημάνουμε όσον το δυνατόν λιγότερα δεδομένα. Ο αριθμός των δεδομένων που θα απαιτηθεί εξαρτάται από το μοντέλο πλήρως εποπτευόμενης μάθησης που θα εκτιμήσει την στρατηγική των ερωτήσεων (μετρητή αβεβαιότητας) που θα χρησιμοποιήσουμε.

Συνήθως στην πλήρως εποπτευόμενη μάθηση που θα ακολουθήσει, θα επιτύχουμε καλύτερα αποτελέσματα εάν από την αρχή είχαμε τον ίδιο αριθμό επισημειωμένων δειγμάτων που επιλέχθηκαν για επισημείωση τυχαία.. Αυτό

οφείλεται στο ότι τα προς επισημείωση δεδομένα έχουν «επιλεγεί» με βάση τον βαθμό αβεβαιότης προς ταξινόμηση που παρουσιάζουν σε διάφορους αλγορίθμους πλήρους εποπτευόμενης μάθησης, με αποτέλεσμα να αποδίδουν καλύτερα την σχέση που συνδέει τα δεδομένα με τις επισημειώσεις από ότι μια τυχαία επιλογή.

(Υποκεφάλαιο 5.4) Μάθηση από θετικά και μη επισημειωμένα δεδομένα

(Ενότητα 5.4.α) Γενικά – τεχνικές - αξιολόγηση

Η μάθηση από θετικά και μη επισημειωμένα δεδομένα (Learning from Positive and Unlabeled Data (PU Learning)) είναι ένα είδος δυαδικής ταξινόμησης όπου τα επισημειωμένα δεδομένα είναι μόνο θετικά και για τα υπόλοιπα δεν γνωρίζουμε σε ποια κλάση ανήκουν. Είναι μια ιδιαίτερη περίπτωση που κατά τεκμήριο είναι η ποιο συνηθισμένη στον χώρο των διαθέσιμων ακτινογραφιών για COVID-19. Είναι λογικό να έχουμε ορισμένες ακτινογραφίες από θετικά δείγματα και έναν τεράστιο αριθμό ακτινογραφιών που δεν γνωρίζουμε εάν ανήκουν σε θετικά ή αρνητικά άτομα. Παρόλη την εμφανέστατη χρησιμότητα της περίπτωσης αυτής φαίνεται να μην υπάρχει αρκετή προσπάθεια στην ερευνητική κοινότητα σε αυτήν την κατεύθυνση

Θα πρέπει να ξεχωρίσουμε την PU (θετικό και χωρίς επισημείωση) ταξινόμηση από δύο παρόμοια «προβλήματα ταξινόμησης»:

Ο πρώτος και πιο συνηθισμένος τύπος προβλήματος επισημείωσης είναι το πρόβλημα μικρού σετ εκπαίδευσης. Δηλαδή, παρόλο που έχουμε έναν αξιοπρεπή όγκο δεδομένων, μόνο ένα μικρό μέρος τους φέρει στην πραγματικότητα επισημείωση. Αυτό το πρόβλημα έχει ήδη αναφερθεί στις περιπτώσεις ημι-εποπτευόμενης και ενεργής μάθησης και έχει συγκεκριμένες μεθοδολογίες προσέγγισης.

Ένα άλλο πρόβλημα επισημείωσης (που συχνά συγχέεται με προβλήματα PU) περιλαμβάνει περιπτώσεις στις οποίες το σύνολο δεδομένων εκπαίδευσης είναι πλήρως επισημειωμένο, αλλά αποτελείται από μία μόνο κατηγορία. Ας υποθέσουμε, για παράδειγμα, ότι το μόνο που έχουμε είναι ένα σύνολο δεδομένων από «non covid» δεδομένα και θα πρέπει να χρησιμοποιήσουμε αυτό το σύνολο δεδομένων για να εκπαιδεύσουμε ένα μοντέλο για τη διάκριση μεταξύ «covid και non-covid». Αυτό είναι ένα πρόβλημα που αντιμετωπίζεται συνήθως ως πρόβλημα ανίχνευσης ακραίων στοιχείων (outliers) χωρίς επίβλεψη,.

Υπάρχουν αρκετά εργαλεία ευρέως διαθέσιμα στην μηχανική μάθηση που έχουν σχεδιαστεί ειδικά για να χειρίζονται αυτήν την περίπτωση (το OneClassSVM είναι το πιο διάσημο).

Ένα πρόβλημα ταξινόμησης PU είναι μια περίπτωση δυαδικής ταξινόμησης που περιλαμβάνει ένα σύνολο εκπαίδευσης στο οποίο μόνο μέρος των δεδομένων είναι επισημειωμένο ως θετικό, ενώ τα υπόλοιπα είναι μη επισημειωμένα και θα μπορούσαν να είναι είτε θετικά είτε αρνητικά. Για παράδειγμα έχουμε πολλές ακτινογραφίες αλλά μόνο ορισμένες από αυτές έχουν επισημειωθεί και ανήκουν μόνο σε θετικούς στον COVID-19 ενώ οι υπόλοιπες μπορεί να ανήκουν είτε σε θετικούς είτε σε αρνητικούς. Αυτό είναι λογικό γιατί οι ασθενείς με COVID-19 συνήθως υποβάλλονται σε ακτινογραφίες, ενώ οι ασυμπτωματικοί ή αρνητικοί θα έχουν υποβληθεί σε ακτινολογικές εξετάσεις για άλλους λόγους.

Οι (Bekker & Davis, Learning from positive and unlabeled data, 2020) στην έκθεσή τους σχετικά με τις μεθόδους μάθησης από «θετικά και μη επισημειωμένα δεδομένα» κάνουν μία συστηματική προσέγγιση όλων των μεθόδων που θα μπορούσαν να χρησιμοποιηθούν. Προτείνουν επτά βασικά ερωτήματα έρευνας που προκύπτουν συνήθως σε αυτόν τον τομέα και παρέχουν μια ευρεία επισκόπηση του θέματος.

Η βιβλιογραφία εκμάθησης PU υποθέτει ένα από τα δύο ακόλουθα σενάρια μάθησης:

Το πρώτο σενάριο προϋποθέτει ένα σύνολο δεδομένων που προέρχεται από κατανομή τυχαίων δειγμάτων. Ένα υποσύνολο των θετικών δειγμάτων του συνόλου δεδομένων είναι επισημειωμένα ενώ τα υπόλοιπα δείγματα είναι χωρίς επισημείωση.

Το δεύτερο σενάριο προϋποθέτει δύο ανεξάρτητα σχεδιασμένα σύνολα δεδομένων: ένα χωρίς επισημείωση και ένα με θετικές επισημειώσεις. Τα επισημειωμένα δεδομένα επιλέγονται από το θετικό υποσύνολο.

Είναι αυτονόητο ότι σε αυτού του είδους τα προβλήματα ότι πρέπει να γίνουν επιπλέον παραδοχές από αυτές που έχουν αναφερθεί για τις μεθόδους μάθησης από μερικώς επισημειωμένα δεδομένα γενικά. Η πιο συνηθισμένη παραδοχή είναι ότι τα δείγματα που έχουν επισημειωθεί έχουν επιλεγεί εντελώς τυχαία. Πρόσφατα, έχει προταθεί η παραδοχή που υποθέτει ότι ο μη-

χανισμός επισημείωσης να εξαρτάται από τα χαρακτηριστικά. Δηλαδή τα χαρακτηριστικά που εμφανίζονται περισσότερες φορές στα θετικά δείγματα, θα απουσιάζουν από τα αρνητικά.

Κάνοντας παραδοχές σχετικά με τα δεδομένα και τον μηχανισμό επισημείωσης είναι δυνατό να εκτιμηθεί η συχνότητα μιας κατηγορίας επισημείωσης. Με βάση αυτήν την διαπίστωση έχουν σχεδιαστεί πολλοί αλγόριθμοι για εκπαίδευση μοντέλων από θετικά και μη επισημειωμένα δεδομένα. Αυτό γίνεται με την εκτίμηση του αναμενόμενου αριθμού θετικών και αρνητικών δειγμάτων των δεδομένων, η οποία μπορεί να επιτευχθεί είτε σταθμίζοντας τα δεδομένα και στη συνέχεια εφαρμόζοντας τυπικούς αλγόριθμους είτε τροποποιώντας απευθείας τους αλγόριθμους ώστε να δέχονται σταθμισμένες εισόδους.

Οι περισσότερες μέθοδοι μάθησης από θετικά και μη επισημειωμένα δεδομένα ανήκουν σε μία από τις τρεις κατηγορίες: τεχνική δύο βημάτων, μεροληπτική μάθηση και μέθοδοι προηγούμενης ενσωμάτωσης κλάσης.

Οι τεχνικές δύο βημάτων ξεκινούν με τον εντοπισμό αξιόπιστων αρνητικών (και μερικές φορές θετικών) δειγμάτων και στη συνέχεια χρησιμοποιούμε τα επισημειωμένα αξιόπιστα δείγματα για την εκπαίδευση ενός ταξινομητή. Οι μεροληπτικές μέθοδοι αντιμετωπίζουν τα μη επισημειωμένα δείγματα σαν να ανήκουν στην αρνητική κλάση. Τέλος οι μέθοδοι προηγούμενης ενσωμάτωσης κλάσης σταθμίζουν τα δεδομένα χωρίς επισημείωση ή τροποποιούν αλγόριθμους μηχανικής εκμάθησης για να υπολογίσουν τον αναμενόμενο αριθμό θετικών και αρνητικών δειγμάτων στα δεδομένα χωρίς επισημείωση.

Ο (Li) στην έκθεσή του «A Survey on Postive and Unlabelled Learning» κάνει μια ανασκόπηση των υπαρχόντων μεθόδων και ιδιαίτερα της στρατηγικής των δύο βημάτων. Αναπτύσσει μεθόδους που μπορούν να χρησιμοποιηθούν σαν πρώτο βήμα (Rocchio, Naive Bayesian Classifier, Spy και 1-DNF). Επίσης αναφέρεται σε μεθόδους που βασίζονται σε σταθμισμένα θετικά και μη επισημειωμένα δείγματα («weighted positive and unlabeled») όπως «σταθμισμένη λογιστική παλινδρόμηση» (weighted logistic regression), SVM βασισμένο σε σταθμισμένα θετικά και μη επισημειωμένα δείγματα. (SVM with weighted positive and unlabeled data) και σε «θορυβώδη αρνητικά δεδομένα» («noisy negative data»).

Η αξιολόγηση του αλγορίθμου δεν είναι ίδια με τις άλλες περιπτώσεις αφού υπάρχουν μόνο πραγματικές θετικές μετρήσεις. «Αυτό μπορεί να προσεγγιστεί με τους εξής τρόπους: α) με τις παραδοσιακές μετρήσεις αξιολόγησης. Σε αυτήν την περίπτωση θα ήταν προτιμότερο να επισημειώσουμε ορισμένα αρνητικά δεδομένα, οπότε να έχουμε την πραγματική απόδοση του αλγορίθμου. β) Να σχεδιάσουμε μετρήσεις που μπορούν να υπολογιστούν με βάση μόνο τα θετικά δείγματα.» (Bekker & Davis, Learning From Positive and Unlabeled Data, 2018).

(Ενότητα 5.4.β) Αλγόριθμοι μάθησης από θετικά και μη επισημειωμένα Δεδομένα.

Οι κυριότεροι αλγόριθμοι που υπάρχουν είναι οι παρακάτω:

Κλασσικός ταξινομητής Charles Elkan and Keith Noto.

Οι (Elkan & Noto) αποδεικνύουν ότι σε ένα σύνολο δεδομένων στο οποίο έχουμε θετικά και μη επισημειωμένα δεδομένα, η πιθανότητα ένα συγκεκριμένο δείγμα να είναι θετικό ($P(y=1|x)$) ισούται με την πιθανότητα το δείγμα να είναι επισημειωμένο ($P(s=1|x)$) διαιρούμενο με την πιθανότητα που έχει ένα θετικό δείγμα, στο σύνολο δεδομένων μας να είναι επισημειωμένο ($P(s=1|y=1)$). Αυτό συμβαίνει επειδή παρόλο που δεν έχουμε αρκετά επισημειωμένα δεδομένα σε ένα σενάριο PU, ώστε να εκπαιδεύσουμε έναν ταξινομητή για να αποφασίσει εάν το δείγμα είναι θετικό ή αρνητικό, έχουμε αρκετά επισημειωμένα δεδομένα για να βρούμε την πιθανότητα που έχει ένα θετικό δείγμα να επισημειωθεί και, σύμφωνα με την μελέτη των E&N, αυτό αρκεί για να εκτιμηθεί πόσο πιθανό είναι να είναι θετικό. Δηλαδή μπορούμε να εκτιμήσουμε την πιθανότητα ένα μη επισημειωμένο δείγμα να είναι θετικό. Με βάση την πιθανότητα που έχει ένα δείγμα να είναι θετικό ($P(s=1|x)$ / $P(s=1|y=1)$), μπορούμε να χρησιμοποιήσουμε σχεδόν οποιονδήποτε ταξινομητή για να τον εκπαιδεύσουμε σύμφωνα με τα ακόλουθα βήματα:

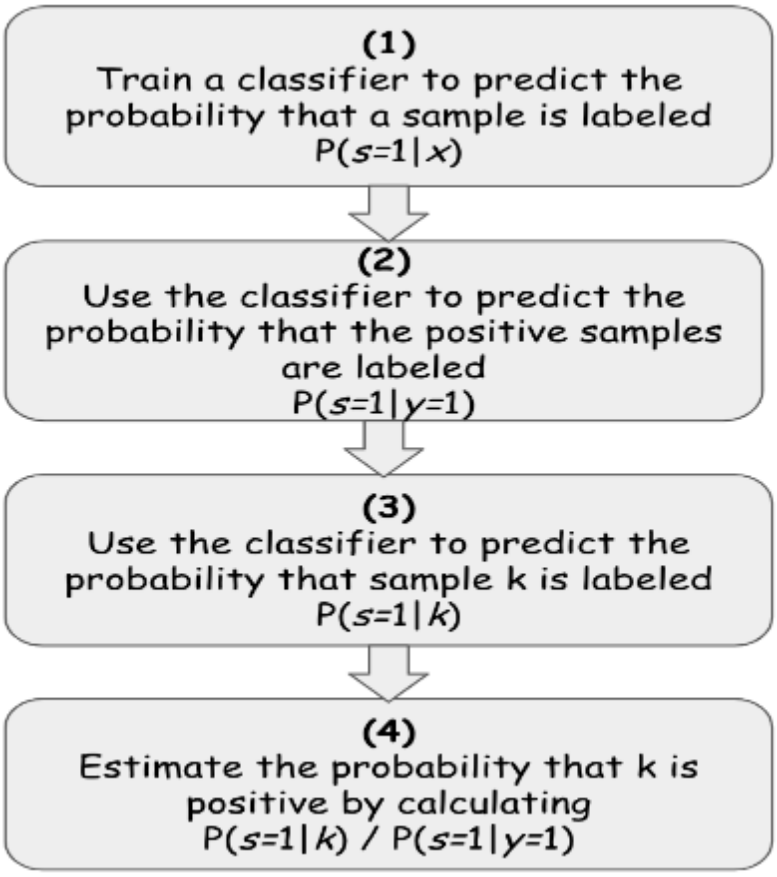
«(1) Εκπαιδεύουμε έναν ταξινομητή σε ένα σύνολο δεδομένων που περιέχει δεδομένα με επισημείωση και χωρίς επισημείωση ώστε να προβλέπει εάν ένα δείγμα έχει επισημείωση. Έτσι η πιθανότητα ένα δεδομένο δείγμα x να φέρει επισημείωση είναι $P(s=1|x)$.

(2) Χρησιμοποιούμε τον ταξινομητή για να προβλέψουμε την πιθανότητα επισημείωσης των γνωστών θετικών δειγμάτων στο σύνολο των δεδομένων

μας, έτσι ώστε τα προβλεπόμενα αποτελέσματα να αντιπροσωπεύουν την πιθανότητα να έχει επισημειωθεί ένα θετικό δείγμα (P(s=1|y=1|x)). Υπολογίζουμε τον μέσο όρο αυτών των πιθανοτήτων P(s=1|y=1).

(3) Έχοντας υπολογίσει το P(s=1|y=1), το μόνο που χρειάζεται να κάνουμε για να προβλέψουμε την πιθανότητα ένα σημείο δεδομένων k να είναι θετικό και η οποία σύμφωνα με τους E&N είναι να εκτιμήσουμε το P(s=1|k) ή την πιθανότητα να είναι επισημειωμένο που είναι ακριβώς αυτό που έχει εκπαιδευτεί να κάνει ο ταξινομητής. Χρησιμοποιούμε τον ταξινομητή στον οποίο εκπαιδεύσαμε στο βήμα (1) για να υπολογίσουμε την πιθανότητα ότι το k έχει επισημείωση ή P(s=1|k).

(4) Μόλις υπολογίζουμε το P(s=1|k), μπορούμε πραγματικά να ταξινομήσουμε το k διαιρώντας το με το P(s=1|y=1), το οποίο έχει υπολογιστεί στο βήμα (2), και να πάρουμε την πραγματική πιθανότητα να ανήκει σε οποιαδήποτε κατηγορία.» (Agmon, 2020)



Εικόνα 13: Λογικό διάγραμμα αλγορίθμου ElkaNoto
(Agmon, 2020)

Μέθοδος με στάθμιση των μη επισημειωμένων δειγμάτων Charles Elkan and Keith Noto.

Σε αυτή τη μέθοδο, οι (Elkan & Noto) προσπαθούν να εκτιμήσουν την πιθανότητα P(y = 1/x) όπου το x είναι ένα δείγμα στο σύνολο δεδομένων και το y είναι η πραγματική επισημείωσή του. Υποθέτουν ότι τα P(Positive) και U (Unlabeled) αντλούνται τυχαία από μία κατανομή που ορίζεται από μια συνάρτηση «βάρους» που υπολογίζεται από έναν ταξινομητή εκπαιδευμένο σε P και U. Ο ταξινομητής μπορεί να εκπαιδευτεί από οποιονδήποτε εποπτευόμενο αλγόριθμο μάθησης. Κάθε δείγμα στο U μπορεί να θεωρηθεί ως σταθμισμένο θετικό δείγμα και ταυτόχρονα ως σταθμισμένο αρνητικό δείγμα. Χρησιμοποιούμε SVM (ή άλλο παραδοσιακού ταξινομητή) για την εκπαίδευση ενός ταξινομητή στο P και στο σταθμισμένο U ο οποίος μοντελοποιεί το P(y = 1/x). Κάθε δείγμα στο U χρησιμοποιείται δύο φορές στην εκπαίδευση μία ως σταθμισμένο θετικό και μία ως σταθμισμένο αρνητικό.

Αυτή η μέθοδος απαιτεί από τον αλγόριθμο ταξινόμησης που εξάγει απευθείας την πιθανότητα, να έχουμε ως αποτέλεσμα «θετικό» ή «αρνητικό». Η έξοδος όμως, των περισσότερων ταξινομητών, όπως του SVM, δεν είναι η πιθανότητα της επισημείωσης αλλά η επισημείωση η ίδια, οπότε απαιτείται κάποια επεξεργασία για την μετατροπή των εξόδων σε πιθανότητες. Οι (Elkan & Noto) χρησιμοποιούν την κλιμάκωση Platt (platt Scaling)[32]. Μία άλλη λύση θα ήταν να μετατρέψουμε την ταξινόμηση σε παλινδρόμηση. Η τιμή των διάφορων μετρικών της παλινδρόμησης μας δίνει το πόσο μακριά ή κοντά βρισκόμαστε σε μία δυαδική πρόβλεψη, άρα και πόσο πιθανόν είναι να βρισκόμαστε κοντά σε μία πρόβλεψη.

Μέθοδος bagging

Το πρόβλημα μετατρέπεται σε σειρές εποπτευόμενων δυαδικών προβλημάτων ταξινόμησης που διακρίνουν τα γνωστά θετικά δείγματα από τυχαία δείγματα χωρίς επισημείωση [33].

[32] Η κλιμάκωση Platt είναι ένας τρόπος μετατροπής της εξόδου μιας ταξινόμησης σε κατανομή πιθανότητας.

[33] SVM με bagging (bootstrap aggregating). Κάθε μεμονωμένο SVM εκπαιδεύεται ανεξάρτητα χρησιμοποιώντας τυχαία επιλεγμένα δείγματα εκπαίδευσης. Τα αποτελέσματα συναξιολογούνται με διάφορους τρόπους (όπως πχ η πλειοψηφία)

Στους αλγόριθμους συνόλου (ensemble), οι μέθοδοι bagging σχηματίζουν μια κατηγορία αλγορίθμων που δημιουργούν πολλά στιγμιότυπά ενός εκτιμητή σε τυχαία υποσύνολα του αρχικού συνόλου εκπαίδευσης και στη συνέχεια συγκεντρώνουν τις μεμονωμένες προβλέψεις τους για να σχηματίσουν μια τελική πρόβλεψη. Αυτές οι μέθοδοι χρησιμοποιούνται ως τρόπος μείωσης της διακύμανσης ενός εκτιμητή βάσης (π.χ., ενός δέντρου αποφάσεων), εισάγοντας την τυχαιοποίηση στη διαδικασία κατασκευής του και στη συνέχεια δημιουργώντας ένα σύνολο από αυτό. Σε πολλές περιπτώσεις, οι μέθοδοι bagging αποτελούν έναν τρόπο βελτίωσης σε σχέση με ένα μεμονωμένο μοντέλο, καθώς παρέχουν έναν τρόπο για τη μείωση της υπερπροσαρμογής (overfitting). Οι μέθοδοι bagging λειτουργούν καλύτερα με πολύπλοκα μοντέλα (πχ πλήρως ανεπτυγμένα δέντρα αποφάσεων), σε αντίθεση με τις μεθόδους boosting που συνήθως λειτουργούν καλύτερα με αδύναμα μοντέλα (π.χ. ρηχά δέντρα απόφασης). Τα τυχαία δείγματα μπορεί να είναι τυχαία υποσύνολα των δεδομένων μας (γραμμών) η τυχαία υποσύνολα των χαρακτηριστικών (στηλών).

Οι (Mordelet & Vert, 2010) προτείνουν ένα γενικό και απλό σχήμα για την επαγωγική μάθηση PU, παρόμοιο με το bagging σε μια εποπτευόμενη δυαδική ταξινόμηση. Η μέθοδός τους, την οποία ονομάζουμε bagging SVM, συνίσταται στη συγκέντρωση ταξινομητών που είναι εκπαιδευμένοι να διακρίνουν το P από ένα μικρό τυχαίο δείγμα του U. Ο αλγόριθμος εκπαιδεύει επαναληπτικά πολλούς δυαδικούς ταξινομητές για να διακρίνουν τα γνωστά θετικά δείγματα από τυχαία δείγματα του συνόλου χωρίς επισημείωση και υπολογίζει τον μέσο όρο των προβλέψεών του. Οι (Mordelet & Vert, 2010) δείχνουν θεωρητικά και πειραματικά ότι η μέθοδος μπορεί να ξεπεράσει σε απόδοση τις άλλες μεθόδους PU, ιδιαίτερα όταν ο αριθμός των θετικών δειγμάτων είναι μικρός και το κλάσμα των αρνητικών μεταξύ των δειγμάτων χωρίς επισημείωση είναι μικρό. Η μέθοδος αυτή μπορεί επίσης να λειτουργήσει πολύ καλά όταν το σύνολο των δειγμάτων χωρίς επισημείωση είναι μεγάλο.» (Mordelet & Vert, 2010)

ΚΕΦΑΛΑΙΟ 6 ΒΑΘΙΑ ΜΑΘΗΣΗ

(Υποκεφάλαιο 6.1) Γενικά

«Η βαθιά μάθηση (deep learning) είναι μέρος μιας ευρύτερης οικογένειας μεθόδων μηχανικής μάθησης που βασίζονται σε τεχνητά νευρωνικά δίκτυα. Η μάθηση μπορεί να είναι εποπτευόμενη, ημι-εποπτευόμενη ή μη εποπτευόμενη.

Αρχιτεκτονικές βαθιάς μάθησης όπως βαθιά νευρωνικά δίκτυα, δίκτυα βαθιάς πεποίθησης, μάθηση βαθιάς ενίσχυσης, επαναλαμβανόμενα νευρωνικά δίκτυα, συνελικτικά νευρωνικά δίκτυα έχουν εφαρμοστεί σε πεδία όπως η υπολογιστική όραση, η επεξεργασία φυσικής γλώσσας, η μηχανική μετάφραση, η βιοπληροφορική, ο σχεδιασμός φαρμάκων, η ανάλυση ιατρικής εικόνας, όπου παρήγαγαν αξιοθαύμαστα αποτελέσματα.

Το επίθετο «deep» στη βαθιά μάθηση αναφέρεται στη χρήση πολλαπλών επιπέδων στο δίκτυο. Η βαθιά μάθηση είναι μια σύγχρονη παραλλαγή των νευρωνικών δικτύων που ασχολείται με έναν μεγάλο αριθμό επιπέδων, που επιτρέπει την πρακτική εφαρμογή, διατηρώντας παράλληλα τη θεωρητική καθολικότητα.

Τα βαθιά δίκτυα έχουν την ιδιότητα να γίνονται πιο ακριβή όσο περισσότερα κρυφά επίπεδα έχουν και όσο περισσότερα δεδομένα έχουν σαν είσοδο. Αυτό απαιτεί υψηλή επεξεργαστική ισχύ, η οποία κατέστη εφικτή τα τελευταία χρόνια. Επίσης ένα άλλο χαρακτηριστικό τους εκτός από τις ακριβείς προβλέψεις που μπορούν να κάνουν είναι να λειτουργήσουν και ως εξαγωγείς χαρακτηριστικών.

Η βαθιά μάθηση έχει αποδειχθεί στην πράξη ιδιαίτερα αποτελεσματική σε ιατρικές εφαρμογές. Τα σύγχρονα εργαλεία βαθιάς μάθησης επιδεικνύουν υψηλή ακρίβεια στον εντοπισμό διαφόρων ασθενειών και είναι ιδιαίτερα χρήσιμα για τη βελτίωση της αποτελεσματικότητας της διάγνωσης» .

(Deep Learning, n.d.)

(Υποκεφάλαιο 6.2) Πλήρως συνδεδεμένα νευρωνικά δίκτυα

(Ενότητα 6.2.α) Γενικά - δομικά στοιχεία νευρωνικών Δικτύων.

Το νευρωνικό δίκτυο είναι ένα δίκτυο από απλούς υπολογιστικούς κόμβους (νευρώνες), διασυνδεδεμένους μεταξύ τους. Είναι εμπνευσμένο από το

κεντρικό νευρικό σύστημα του ανθρώπου, το οποίο προσπαθεί να προσομοιώσει.

Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε νευρώνας δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), εκτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με ένα ποσό που λέγεται συναπτικό βάρος (η τιμή του αντιστοιχεί με την δύναμη της σύνδεσης) και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτεί ως όρισμα μια συνάρτηση που λέγεται συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε νευρώνας. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.

Εάν x_{ki} είναι η i-στη είσοδος του κ νευρώνα, ω_{ki} το συναπτικό βάρος του κ νευρώνα και φ η συνάρτηση ενεργοποίησης του νευρωνικού δικτύου, τότε η έξοδος y_{ki} είναι:

$$y_k = \phi\left(\sum_{i=0}^{N} x_{ki} w_{ki}\right)$$

Στον κ-οστό νευρώνα υπάρχει ένα συναπτικό βάρος ω_{k0} με ιδιαίτερη σημασία, το οποίο καλείται πόλωση ή κατώφλι (bias, threshold). Η τιμή της εισόδου του είναι πάντα η μονάδα. Εάν το συνολικό άθροισμα από τις υπόλοιπες εισόδους του νευρώνα είναι μεγαλύτερο από την τιμή αυτή, τότε ο νευρώνας ενεργοποιείται. Εάν είναι μικρότερο, τότε ο νευρώνας παραμένει ανενεργός.

Όπως είναι φανερό, κάθε στοιχείο του διανύσματος χαρακτηριστικών του προς επίλυση προβλήματος τροφοδοτεί κατά τη λειτουργία του δικτύου έναν νευρώνα εισόδου. Οι αριθμοί οι οποίοι συναποτελούν το διάνυσμα εξόδου

κάθε στοιχείο του οποίου εμφανίζεται, μετά το πέρας του ολικού υπολογισμού, σε έναν νευρώνα εξόδου, αντιστοιχούν στην πρόβλεψη.

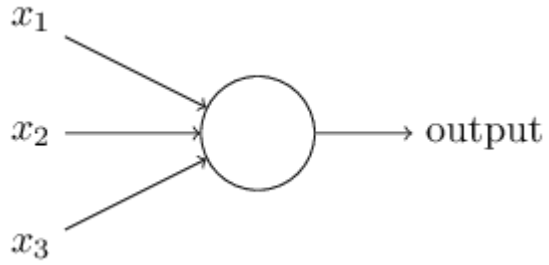
Εάν το πρόβλημα είναι πρόβλημα ταξινόμησης (classification) τότε υπάρχουν τόσοι νευρώνες εξόδου όσες οι κατηγορίες. Η πρόβλεψη θα είναι ίση με τον νευρώνα που έχει την μεγαλύτερη τιμή (συνήθως 1). Εάν το πρόβλημα είναι παλινδρόμησης (regression) τότε έχουμε σαν έξοδο ένα νευρώνα, η τιμή του οποίου μας δίνει το αποτέλεσμα.

Αυτό που μας ενδιαφέρει πρωτίστως είναι το δίκτυο να μετατρέπει με ορθό τρόπο τα διανύσματα εισόδου σε κατάλληλα διανύσματα εξόδου, το πρόβλημα δηλαδή είναι η υλοποίηση μίας συνάρτησης πολλαπλών μεταβλητών, που κατά κανόνα έχουν περίπλοκο και άγνωστο ακριβή τύπο.

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η ικανότητα μάθησης. Ως μάθηση μπορεί να οριστεί η σταδιακή βελτίωση της ικανότητας του δικτύου να επιλύει κάποιο πρόβλημα. Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης, μίας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου (συνήθως των βαρών και της πόλωσής του) σε τιμές κατάλληλες ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα. Αφού ένα δίκτυο εκπαιδευτεί, οι παράμετροί του «παγώνουν» στις κατάλληλες τιμές και από εκεί κι έπειτα είναι σε λειτουργική κατάσταση. Το ζητούμενο είναι το νευρωνικό δίκτυο να χαρακτηρίζεται από μία ικανότητα γενίκευσης. Αυτό σημαίνει πως δίνει ορθές εξόδους για εισόδους διαφορετικές από αυτές με τις οποίες εκπαιδεύτηκε, όπως ακριβώς και στα άλλα μοντέλα μηχανικής μάθησης.

(Ενότητα 6.2.β) Perceptrons – αρχιτεκτονική νευρωνικών δικτύων

Ένα perceptron είναι το δομικό στοιχείο (νευρώνας του δικτύου) που λαμβάνει αρκετές δυαδικές εισόδους, x1, x2,... και παράγει μία μόνο δυαδική έξοδο.

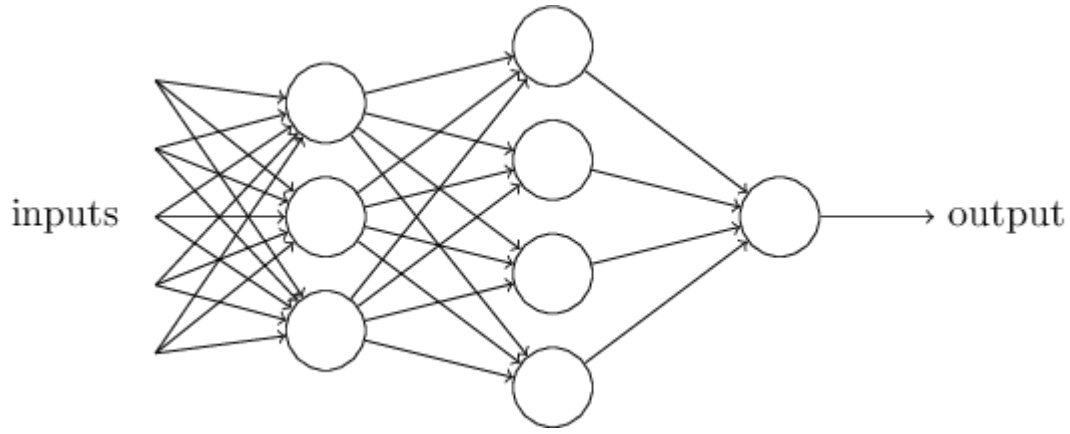


Εικόνα 14 Perceptron

Στο παραπάνω παράδειγμα το perceptron έχει τρεις εισόδους, x1, x2, x3. Σε γενικές γραμμές θα μπορούσε να έχει περισσότερες ή λιγότερες εισόδους, αλλά μία μόνο έξοδο. Υπάρχει ένας απλός κανόνας για τον υπολογίσουμε την έξοδο (output). Τα βάρη, w1, w2,..., είναι πραγματικοί αριθμοί που εκφράζουν τη σημασία των αντίστοιχων εισόδων στην έξοδο. Η έξοδος του νευρώνα, 0 ή 1, καθορίζεται από το εάν το σταθμισμένο άθροισμα $\sum W_j X_j$ είναι μικρότερο ή μεγαλύτερο από κάποια τιμή κατωφλίου (threshold). Ακριβώς όπως τα βάρη, το κατώφλι είναι ένας πραγματικός αριθμός που είναι μια παράμετρος του νευρώνα. Για να το θέσουμε με πιο ακριβείς αλγεβρικούς όρους:

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases}$$

Μπορούμε να πούμε ότι το perceptron είναι μια διάταξη που λαμβάνει αποφάσεις σταθμίζοντας στοιχεία.



Εικόνα 15 Δίκτυο από perceptrons

Ένα νευρωνικό δίκτυο αποτελείται από πολλά perceptrons, τα οποία τα χωρίζουμε σε επίπεδα (layers). Στο δίκτυο της εικόνας η πρώτη στήλη των perceptrons - αυτό που ονομάζουμε το πρώτο επίπεδο των perceptrons - λαμβάνει απλές αποφάσεις, σταθμίζοντας τα στοιχεία εισόδου. Καθένα από τα perceptrons του 2ου επιπέδου λαμβάνει μια απόφαση σταθμίζοντας τα αποτελέσματα από το πρώτο επίπεδο λήψης αποφάσεων. Με αυτόν τον τρόπο, ένα perceptron στο δεύτερο επίπεδο μπορεί να πάρει μια απόφαση σε πιο περίπλοκο και πιο αφηρημένο επίπεδο από τα perceptron στο πρώτο επίπεδο. Και ακόμη πιο περίπλοκες αποφάσεις μπορούν να ληφθούν από το perceptron στο τρίτο επίπεδο κοκ. Με αυτόν τον τρόπο, ένα δίκτυο πολλαπλών επιπέδων μπορεί να συμμετάσχει σε εξελιγμένη λήψη αποφάσεων. Όπως έχουμε πει ένα

perceptron έχει μόνο μία έξοδο. Στο δίκτυο στην παραπάνω εικόνα φαίνεται ότι τα perceptrons έχουν πολλαπλές εξόδους. Στην πραγματικότητα, είναι ενιαία έξοδος. Τα πολλαπλά βέλη εξόδου είναι απλώς ένας τρόπος ένδειξης ότι η έξοδος από ένα perceptron χρησιμοποιείται ως είσοδος σε πολλά άλλα perceptron.

Ο παραπάνω συμβολισμός, ένα percepton χωρίς είσοδο και με μία έξοδο συμβολίζει την είσοδο και όχι πραγματικό percepton.

Η συνθήκη $\Sigma_j w_i x_i >$ κατωφλίου είναι δύσχρηστη και μπορούμε να κάνουμε δύο αλλαγές για να την απλοποιήσουμε. Η πρώτη αλλαγή είναι να γράψουμε $\Sigma_j w_i x_i$ ως εσωτερικό γινόμενο, $w.x = \Sigma_j w_i x_i$ όπου w και x είναι διανύσματα των οποίων τα συστατικά είναι τα βάρη και οι είσοδοι, αντίστοιχα. Η δεύτερη αλλαγή είναι να μετακινήσουμε το κατώφλι στην άλλη πλευρά της ανισότητας και να το αντικαταστήσουμε με αυτό που είναι γνωστό ως πόλωση (bias) του perceptron (b≡threshold). Χρησιμοποιώντας την πόλωση αντί για το κατώφλι, ο κανόνας perceptron μπορεί να ξαναγραφεί:

$$\text{output} = \begin{cases} 0 & \text{if } w \cdot x + b \leq 0 \\ 1 & \text{if } w \cdot x + b > 0 \end{cases}$$

Μπορούμε να θεωρήσουμε το b σαν ένδειξη του πόσο εύκολα παίρνει η έξοδος την τιμή 1.

Μπορούμε να επινοήσουμε εκπαιδευτικούς αλγόριθμους που μπορούν να συντονίσουν αυτόματα τα βάρη και τις πολώσεις ενός δικτύου τεχνητών νευρώνων. Αυτός ο συντονισμός συμβαίνει ως απόκριση σε εξωτερικά ερεθίσματα, χωρίς άμεση παρέμβαση προγραμματιστή. Τα νευρωνικά δίκτυα δηλαδή μπορούν να μάθουν να επιλύουν προβλήματα από μονά τους.

Ας υποθέσουμε ότι έχουμε ένα δίκτυο perceptrons που θα θέλαμε να χρησιμοποιήσουμε για να μάθουμε να επιλύουμε κάποιο πρόβλημα, θα θέλαμε το δίκτυο να μάθει βάρη και πολώσεις, έτσι ώστε η έξοδος από το δίκτυο να ταξινομεί σωστά. Ας υποθέσουμε ότι κάνουμε μια μικρή αλλαγή σε κάποιο βάρος (ή πόλωση) στο δίκτυο. Αυτό που θα θέλαμε είναι αυτή η μικρή αλλαγή βάρους να προκαλεί μόνο μια μικρή αντίστοιχη αλλαγή στην έξοδο του δικτύου.

Για παράδειγμα, ας υποθέσουμε ότι το δίκτυο ταξινόμησε κατά λάθος ένα αποτέλεσμα ως «COVID» όταν θα έπρεπε να είναι "NON-COVID". Θα μπορούσαμε να καταλάβουμε πώς αν κάνουμε μια μικρή αλλαγή στα βάρη και τις πολώσεις, ώστε το δίκτυο να πλησιάσει λίγο στην ταξινόμηση της ως "NON-COVID» και μετά να το επαναλάβαμε, αλλάζοντας τα βάρη και τις πολώσεις ξανά και ξανά για να παράγουμε καλύτερη και καλύτερη απόδοση. Έτσι το δίκτυο εκπαιδεύεται.

Αυτό δεν συμβαίνει όταν το δίκτυό μας περιέχει απλά perceptrons. Στην πραγματικότητα, μια μικρή αλλαγή στα βάρη ή πολώσεις οποιουδήποτε μεμονωμένου perceptron στο δίκτυο μπορεί μερικές φορές να αλλάξει την έξοδο αυτού του perceptron εντελώς, από 0 έως 1 και αντίστροφα. Αυτό καθιστά δύσκολο να τροποποιήσουμε σταδιακά τα βάρη και τις πολώσεις, έτσι ώστε το δίκτυο να πλησιάσει την επιθυμητή συμπεριφορά. Μπορούμε να ξεπεράσουμε αυτό το πρόβλημα με έναν άλλο τύπο τεχνητού νευρώνα που θα χρησιμοποιεί μία άλλη συνάρτηση (συνάρτηση ενεργοποίησης) για να παράγει την έξοδο. Η έξοδος δεν θα είναι 0 ή 1 αλλά μπορεί να είναι οποιοσδήποτε πραγματικός αριθμός σ (w.x + b), όπου σ η συνάρτηση ενεργοποίησης. Έτσι επιτυγχάνουμε οι μικρές αλλαγές στα βάρη και την πόλωση να προκαλούν μόνο μια μικρή αλλαγή στην έξοδό τους. Αυτό είναι και το κρίσιμο γεγονός που θα επιτρέψει ένα δίκτυο νευρώνων να «μάθει».

Τα κλασικά perceptrons παίρνουν εισόδους 0 και 1 ενώ αυτά τα τροποποιημένα perceptrons οι είσοδοι μπορούν να πάρουν οποιεσδήποτε τιμές μεταξύ 0 και 1. Έτσι, για παράδειγμα, 0,638... είναι μια έγκυρη είσοδος για αυτά. Επίσης, ακριβώς όπως ένα κλασικό perceptron, ο νευρώνας αυτός έχει βάρη για κάθε είσοδο, w1, w2,... και μια συνολική πόλωση, b.

(Ενότητα 6.2.γ) Συναρτήσεις ενεργοποίησης

Κάθε νευρώνας σε ένα τεχνητό νευρωνικό δίκτυο έχει μια συνάρτηση ενεργοποίησης που εκτελεί μια μαθηματική πράξη στο άθροισμα των σταθμισμένων εισόδων και της πόλωσης του και παράγει μια έξοδο

Η συνάρτηση ενεργοποίησης μπορεί να είναι **βηματική** (step transfer function), **γραμμική** (linear transfer function**), μη γραμμική** (non linear transfer function), στοχαστική (stochastic transfer function).

Η **βηματική** συνάρτηση ενεργοποίησης μπορεί να είναι της μορφής:

$$\phi(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

ή οποιαδήποτε άλλη βηματική συνάρτηση.

Η βηματική συνάρτηση **δεν θεωρείται χρήσιμη** ως συνάρτηση ενεργοποίησης στα τεχνητά νευρωνικά δίκτυα, καθώς σύμφωνα με τον απειροστικό λογισμό έχει το βασικό μειονέκτημα ότι η παράγωγός της απειρίζεται. Έτσι προέκυψε η ανάγκη συναρτήσεων ενεργοποίησης που η γραφική παράσταση τους να μοιάζει με τη βηματική, αλλά ταυτόχρονα να είναι συνεχείς και παραγωγίσιμες σε όλο το πεδίο ορισμού τους.

Η γραμμική συνάρτηση ενεργοποίησης μπορεί να είναι της μορφής:

$$\phi(x) = x$$

ή οποιαδήποτε άλλη γραμμική συνάρτηση.

Η μη γραμμική συνάρτηση ενεργοποίησης που χρησιμοποιείται συνήθως στα νευρωνικά δίκτυα καλείται σιγμοειδής συνάρτηση. Οι τυπικές σιγμοειδείς είναι δύο:

Λογιστική σιγμοειδής, τιμές από 0 έως 1:

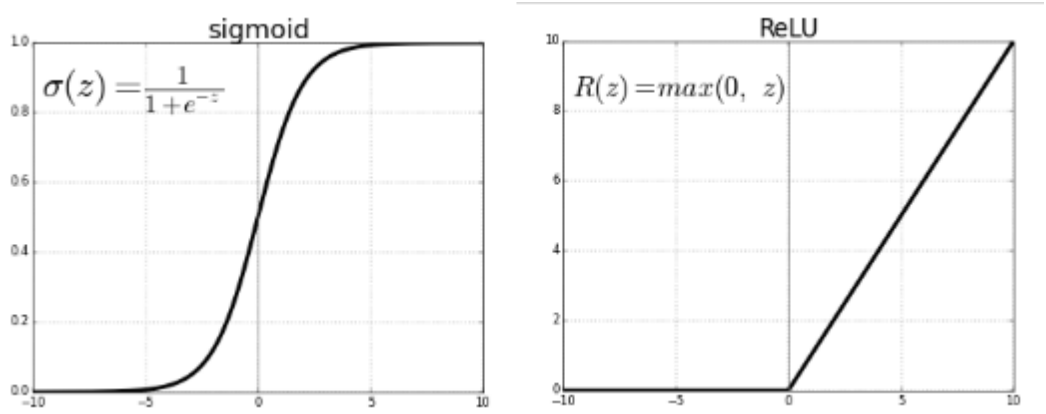
$$\phi(z) = \frac{1}{1 + e^{-z}}$$

Υπερβολική εφαπτομένη, τιμές από -1 έως 1:

$$\phi(x) = \tanh x$$

(Wikipedia, 2021)

Η σιγμοειδής συνάρτηση είναι συνεχής και παραγωγίσιμη με αποτέλεσμα να μπορούμε να βρούμε εύκολα την κλίση μεταξύ οποιονδήποτε δύο σημείων της. Η παράγωγός της δεν είναι συνεχής που σημαίνει ότι μπορεί να κωλύσει σε τοπικά ελάχιστα κατά την διάρκεια της εκπαίδευσης (βλ. gradient descent).



Εικόνα 16: Συναρτήσεις ενεργοποίησης (Σιγμοειδής και RelU)
(Sharma, Activation Functions in Neural Networks, 2017)

Η συνάρτηση ενεργοποίησης που χρησιμοποιείται στα βαθιά νευρωνικά δίκτυα είναι συνήθως η Rectified Linear Unit (RelU)[34] (P(χ) = max(0,x)) επειδή έχει δείξει πλεονεκτήματα έναντι της σιγμοειδούς. Η συνάρτηση αλλά και η παραγωγός της είναι συνεχής και παραγωγίσιμη. Κύριο αποτέλεσμα οι εύκολοι υπολογισμοί και της συνάρτησης και της παραγώγου της και το ότι δεν «κολλάει» σε τοπικά ελάχιστα.

Τελικά, η έξοδος ενός σιγμοειδούς νευρώνα με εισόδους x1, x2,..., βάρη w1, w2,... και η πόλωση b είναι:

$$\frac{1}{1 + \exp(-\sum_j w_j x_j - b)}.$$

Η λογιστική σιγμοειδής παίρνει τιμές από 0 έως 1 γι' αυτό είναι κατάλληλη για πρόβλεψη πιθανότητας, ενώ η ReLu παίρνει τιμές από 0 έως άπειρο και είναι καταλληλότερη για τα βαθιά νευρωνικά δίκτυα.

«Η συνάρτηση softmax είναι μια πιο γενικευμένη συνάρτηση ενεργοποίησης λογιστικής που χρησιμοποιείται για ταξινόμηση πολλαπλών κλάσεων.»
(Sharma, Activation Functions in Neural Networks, 2017) [35]

Χρησιμοποιούμε τη συνάρτηση ενεργοποίησης sigmoid/softmax στο τελικό επίπεδο εξόδου όταν προσπαθούμε να λύσουμε τα προβλήματα ταξινόμησης όπου οι επισημειώσεις είναι τιμές κλάσης. Χρησιμοποιούμε ReLu στο

[34] Ο νευρώνας που την υλοποιεί λέγεται ReLU, ενώ η συνάντηση Rel.
[35] Λεπτομερής Περιγραφή των συναρτήσεων ενεργοποίησης στο https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

τελευταίο επίπεδο όταν θέλουμε να προβλέψουμε πραγματικές θετικές τιμές πχ ύψος ανθρώπου.

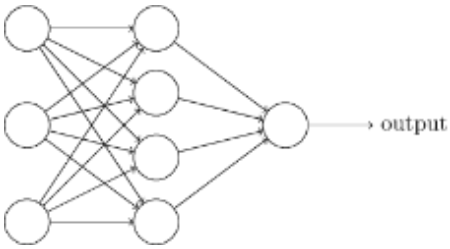
Η ομαλότητα του σ σημαίνει ότι μικρές αλλαγές Δw_iστα βάρη και Δb στη πόλωση θα παράγουν μια μικρή αλλαγή Δ στην έξοδο. Στην πραγματικότητα, το Δoutput προσεγγίζεται καλά από:

$$\Delta \text{output} \approx \sum_j \frac{\partial \,\text{output}}{\partial w_j}\Delta w_j + \frac{\partial \,\text{output}}{\partial b}\Delta b$$

όπου το άθροισμα υπολογίζεται σε όλα τα βάρη, w_j και ∂output / ∂w_j και ∂output / ∂b δηλώνουν μερικές παραγώγους της εξόδου σε σχέση με τα w_i και b, αντίστοιχα.

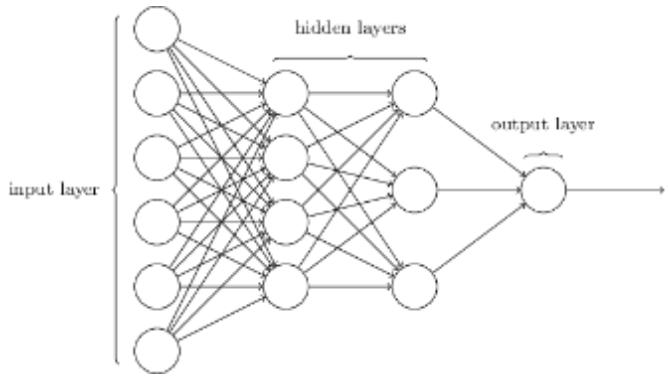
(Ενότητα 6.2.δ) Σχεδιασμός νευρωνικού δικτύου

Όπως αναφέρθηκε προηγουμένως, το αριστερότερο επίπεδο ονομάζεται επίπεδο εισόδου και οι νευρώνες εντός του 1ου αυτού επιπέδου ονομάζονται νευρώνες εισόδου. Το δεξιότερο ή το επίπεδο εξόδου περιέχει τους νευρώνες εξόδου ή, όπως στην περίπτωση του παρακάτω σχήματος, το οποίο έχει έναν μόνο νευρώνα εξόδου. Το μεσαίο στρώμα ονομάζεται κρυφό επίπεδο (στρώμα) καθώς οι νευρώνες σε αυτό το επιπεδο δεν είναι ούτε είσοδοι ούτε έξοδοι.



Εικόνα 17 Δίκτυο 3 επιπέδων (εισόδου - ενός κρυφού – εξόδου)

Το παραπάνω δίκτυο έχει μόνο ένα κρυφό επίπεδο, αλλά ορισμένα δίκτυα έχουν πολλά κρυφά επίπεδα. Για παράδειγμα, το ακόλουθο δίκτυο τεσσάρων επιπέδων έχει δύο κρυφά επίπεδα:



Εικόνα 18 Δίκτυο 3 επιπέδων (εισόδου – δύο κρυφών - εξόδου

Ο σχεδιασμός των επιπέδων εισόδου και εξόδου σε ένα δίκτυο είναι συνήθως απλός. Το επίπεδο εισόδου για παράδειγμα αντιστοιχεί στον αριθμό των χαρακτηριστικών, δηλαδή των στηλών του dataset, ενώ το επίπεδο output εάν ήταν πρόβλημα ταξινόμησης θα ήταν όσες είναι οι ομάδες (στην προκείμενη 2 (COVID-19, NON-COVID)), εάν πρόκειται για παλινδρόμηση τότε αρκεί για έξοδος ένας(1) νευρώνας που θα δείχνει την τιμή που προβλέφθηκε.

Ενώ ο σχεδιασμός των επιπέδων εισόδου και εξόδου ενός νευρωνικού δικτύου είναι γενικά απλός, μπορεί να χρειαστεί αρκετή εμπειρία και ικανότητα στο σχεδιασμό των κρυφών επιπέδων. Συνήθως αυτά βρίσκονται με διάφορες ευρετικές σχεδίασης .

(Ενότητα 6.2.ε) Σχεδιασμός και λειτουργία νευρωνικού δικτύου

Τα νευρωνικά δίκτυα είναι ένας αλγόριθμος και όπως είναι προφανές δεν μπορούμε να φτιάξουμε κανένα μοντέλο χωρίς να έχουμε κάποια μαθηματική εξίσωση. Η εξίσωση είναι ο γραμμικός συνδυασμός των εισόδων και των αντίστοιχων βαρών τους και ένας όρος πόλωσης (bias), ο οποίος είναι ο συντελεστής διόρθωσης που έχει το βάρος. Η πόλωση χρειάζεται επειδή η έξοδος του μοντέλου δεν μπορεί να είναι μηδενική όταν δεν υπάρχουν είσοδοι (X). Έτσι, η εξίσωση του νευρωνικού δικτύου είναι: $Z = b0 + W1X1 + W2X2 + \ldots + WnXn$. Το Z είναι έξοδος, W, είναι τα βάρη και η πόλωση (bias) το b0.

Σε κάθε νευρώνα εισόδου (νευρώνα $1^{ου}$ επιπέδου) αποδίδεται ένα βάρος. Αρχικά, τα βάρη κατανέμονται τυχαία. Αυτά τα βάρη πολλαπλασιάζονται με κάθε τιμή εισόδου και στη συνέχεια προστίθενται μαζί, όπως προβλέπει η εξίσωση του νευρωνικού δικτύου($Z = b0 + W1X1 + W2X2 + \ldots + WnXn$)

Η παραπάνω εξίσωση περνάει από έναν μετασχηματισμό (ενεργοποίηση). Η συνάρτηση ενεργοποίησης προσθέτει μη γραμμικότητα στη εξίσωση του

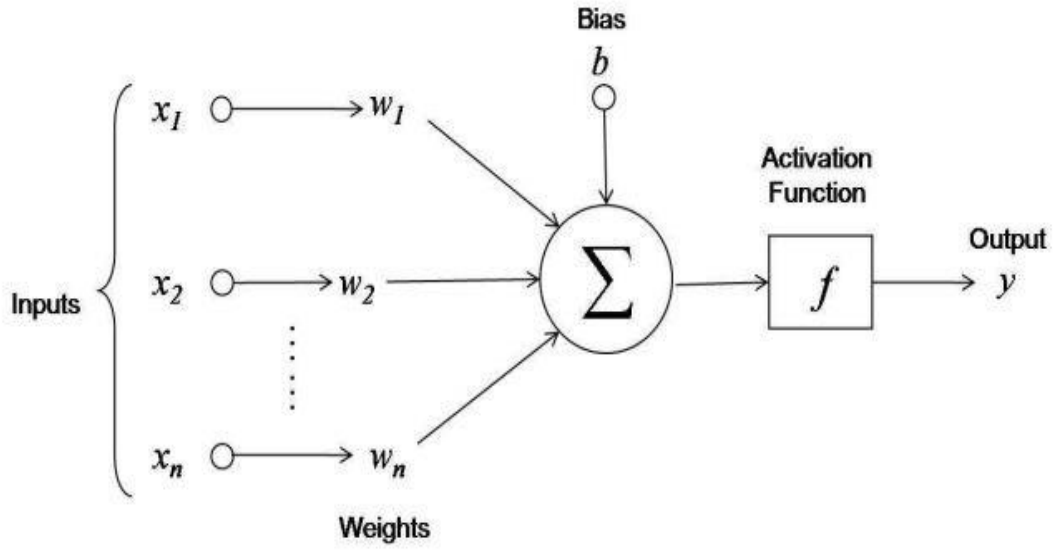
νευρωνικού δικτύου. Έτσι, για να λάβουμε υπόψη τη μη γραμμικότητα, εφαρμόζουμε κάποιο μαθηματικό μετασχηματισμό στην εξίσωση του νευρωνικού δικτύου πριν δημιουργηθεί η έξοδος.

Η συνάρτηση ενεργοποίησης εξαρτάται από τον τύπο δεδομένων και το πρόβλημα. Ως εκ τούτου, είναι μια παράμετρος για το νευρωνικό δίκτυο. Ας πούμε, για τον νευρώνα 1, N1 = Φ(Z) όπου

$$\phi(z) = \frac{1}{1 + e^{-z}}$$

μετά την εφαρμογή της σιγμοειδούς συνάρτησης ενεργοποίησης η έξοδος πλησιάζει το 0 όταν το Z γίνεται πολύ αρνητικό, το 1 όταν το Z είναι πολύ θετικό, και το 0,5 όταν Z = 0. Ωστόσο, είναι αργή και αντικαθίσταται από τη συνάρτηση RelU.

Στο επίπεδο εξόδου, η συνάρτηση ενεργοποίησης βασίζεται και στον τύπο του προβλήματος που αντιμετωπίζουμε.



Εικόνα 19 Λειτουργία νευρώνα
(Arnx, 2019)

Τα βήματα που περιλαμβάνονται για την εκπαίδευση ενός νευρωνικού δίκτυο είναι:

Τοποθετούμε τα χαρακτηριστικά στην είσοδο του νευρωνικού δικτύου (στους νευρώνες πρώτου επιπέδου).

Παίρνουμε την εξίσωση εισόδου: Z = b0 + W1X1 + W2X2 +... + WnXn και υπολογίζουμε την έξοδο εφαρμόζοντας την συνάρτηση ενεργοποίησης. Έτσι υπολογίζεται η έξοδος του νευρώνα.

Την έξοδο κάθε νευρώνα την χρησιμοποιούν σαν είσοδο οι νευρώνες του επομένου επιπέδου κοκ. Η κάθε έξοδος μεταβιβάζεται σε όλους τους νευρώνες του επομένου επιπέδου μέχρι το τελευταίο επίπεδο.

Η έξοδος των νευρώνων του τελευταίου επιπέδου μας δίνουν την πρόβλεψη. Η συνάρτηση ενεργοποίησης είναι συνήθως διαφορετική από τα άλλα επίπεδα αναλόγως του σκοπού του νευρωνικού δικτύου (σιγμοειδής για δυαδική ταξινόμηση, softmax για ταξινόμηση πολλαπλών κλάσεων και ReLU για παλινδρόμηση).

Στις περιπτώσεις ταξινόμησης στο τελευταίο επίπεδο έχουμε τόσους νευρώνες όσες οι κλάσεις στις οποίες μπορεί να ανήκει ένα αντικείμενο για το οποίο θέλουμε να κάνουμε προβλέψεις. Κάθε νευρώνας του τελευταίου επιπέδου αντιπροσωπεύει μία κλάση. Ο νευρώνας που παίρνει την τιμή 1 (ή αυτός με την μεγαλύτερη τιμή) αντιπροσωπεύει την πρόβλεψη. Στην παλινδρόμηση το τελευταίο επίπεδο έχει ένα μόνο νευρώνα που μας δίνει σαν έξοδο μια τιμή. Η διαδικασία μετάβασης από αριστερά προς τα δεξιά, δηλαδή από το επίπεδο εισόδου στο επίπεδο εξόδου για την προσαρμογή είναι γνωστή ως «εμπρόσθια διάδοση» (forward propagation).

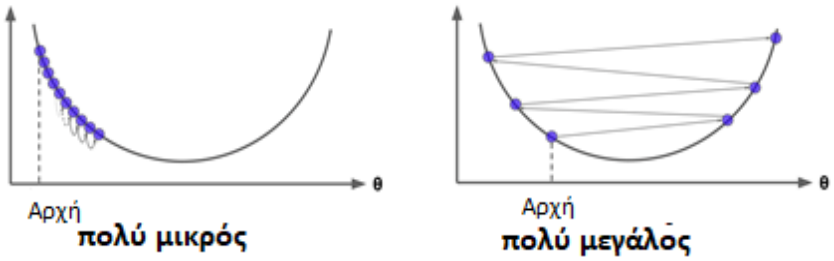
Κατόπιν υπολογίζουμε το σφάλμα (απώλεια) το οποίο μας λέει πόσο το μοντέλο αποκλίνει από τις πραγματικές τιμές. Υπολογίζεται ως η διαφορά μεταξύ των προβεφθεισών και πραγματικών τιμών.

Μέχρι τώρα, λάβαμε την είσοδο, βρήκαμε την έξοδο και υπολογίσαμε το σφάλμα. Ο τελικός στόχος είναι να ελαχιστοποιήσουμε το σφάλμα πηγαίνοντας την υπολογισμένη τιμή απώλειας πίσω σε κάθε επίπεδο, έτσι ώστε να ενημερώνει τα βάρη και τις πολώσεις με τέτοιο τρόπο ώστε να ελαχιστοποιείται η απώλεια. Αυτός ο τρόπος πρώτης προσαρμογής των βαρών και των πολώσεων είναι γνωστός ως «οπίσθιά διάδοση» (backward propagation).

(Ενότητα 6.2.στ) Κάθοδος βασισμένη στην κλίση

Η κάθοδος βασισμένη στην κλίση (gradient descent) είναι ένας πολύ γενικός αλγόριθμος βελτιστοποίησης ικανός να βρει βέλτιστες λύσεις σε ένα ευρύ φάσμα προβλημάτων. Η γενική ιδέα του gradient descent είναι να τροποποιήσει τις παραμέτρους επαναληπτικά προκειμένου να ελαχιστοποιηθεί μια συνάρτηση κόστους. Το gradient descent μετρά την τοπική κλίση (παράγωγο) της συνάρτησης απώλειας (κόστους) για ένα δεδομένο σύνολο παραμέτρων και κάνει βήματα προς την κατεύθυνση της φθίνουσας κλίσης. Μόλις η κλίση είναι μηδέν, φτάσαμε στο ελάχιστο.

Για να υπολογιστεί το gradient descent, οποιαδήποτε συνάρτησης απώλειας αυτή πρέπει να είναι διαφοροποιήσιμη. Αυτό επιτρέπει στα μοντέλα να βελτιστοποιούν διάφορες συναρτήσεις απώλειας. Μια σημαντική παράμετρος στο gradient descent είναι το μέγεθος των βημάτων που καθορίζεται από την υπερπαράμετρο «ρυθμό εκμάθησης (learning ratio)». Εάν ο ρυθμός εκμάθησης είναι πολύ μικρός, τότε ο αλγόριθμος θα χρειαστεί πολλές επαναλήψεις για να βρει το ελάχιστο. Από την άλλη πλευρά, εάν ο ρυθμός εκμάθησης είναι πολύ υψηλός, μπορεί να υπερβούμε το ελάχιστο και να καταλήξουμε πιο μακριά από ό, τι όταν ξεκινήσαμε.



Εικόνα 20 Ρυθμός εκμάθησης (βήμα)

Επιπλέον, δεν είναι όλες οι συναρτήσεις κόστους κυρτές (σε σχήμα μπολ). Μπορεί να υπάρχουν τοπικά ελάχιστα, οροπέδια και άλλα ακανόνιστες γραφικές παραστάσεις της συνάρτησης απώλειας που καθιστά δύσκολη την εύρεση του ολικού ελάχιστου.

Ο ρυθμός εκμάθησης είναι μια υπερ-παράμετρος που ελέγχει πόσο ρυθμίζουμε τα βάρη του δικτύου μας σε σχέση με την βαθμό απώλειας. Όσο χαμηλότερη είναι η τιμή, τόσο πιο αργά πηγαίνουμε. Αυτό μπορεί να είναι μια καλή ιδέα (χρησιμοποιώντας χαμηλό ρυθμό εκμάθησης), όσον αφορά τη διασφάλιση ότι δεν θα χάσουμε το ελάχιστο, θα μπορούσε επίσης να σημαίνει ότι θα χρειαστεί πολύς χρόνος για σύγκλιση - ειδικά εάν κολλήσουμε σε ένα τοπικό ακρότατο.

(Ενότητα 6.2.ζ) Εποπτευόμενα – μη εποπτευόμενα νευρωνικά δίκτυα

Ένα νευρωνικό δίχτυο λέγεται ότι μαθαίνει εποπτευόμενο, εάν η επιθυμητή έξοδος είναι ήδη γνωστή. Κατά την εκπαίδευση μία γραμμή από τα δεδομένα δίνεται στο επίπεδο εισόδου. Αυτό διαδίδεται μέσω του διχτύου (ανεξάρτητα από τη δομή του) στο επίπεδο εξόδου. Το επίπεδο εξόδου δημιουργεί την έξοδο, η οποία στη συνέχεια συγκρίνεται με τον στόχο (επισημείωση). Ανάλογα με τη διαφορά μεταξύ εξόδου και στόχου, υπολογίζεται μια τιμή σφάλματος. Αυτό το σφάλμα με την μέθοδο της οπίσθιας διάδοσης διορθώνει

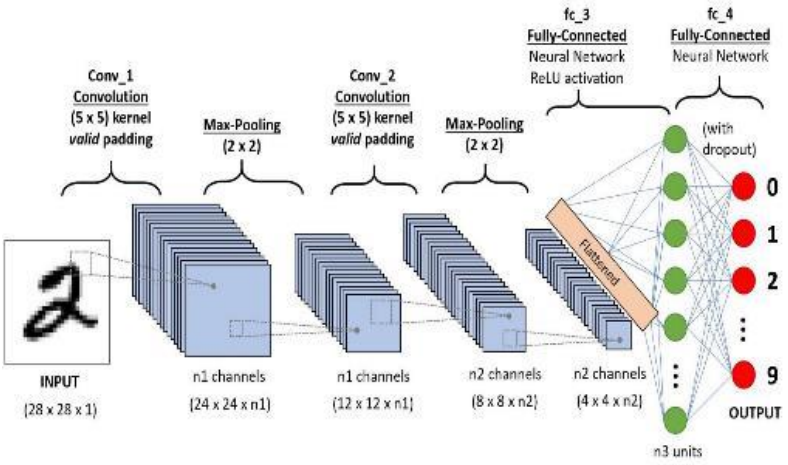
τα βάρη και τις πολώσεις σταδιακά με την επαναληπτική διαδικασία εμπρόσθιας – οπίσθιας διάδοσης μέχρις ότου μηδενιστεί το λάθος ή επιτευχθεί ο μέγιστος αριθμός επαναλήψεων (epoch). Εκπαίδευση λοιπόν είναι η εκμάθηση των βαρών και των πολώσεων που μηδενίζουν το λάθος της πρόβλεψης. Στην περίπτωση που δεν μηδενιστούν τα λάθη μετά την ολοκλήρωση του μεγίστου αριθμού των επαναλήψεων η μέγιστη τιμή των νευρώνων του τελευταίου επιπέδου δίνει την πρόβλεψη της κλάσης.

Τα νευρωνικά δίκτυα τα οποία μαθαίνουν χωρίς επίβλεψη δεν έχουν τέτοιου είδους στόχους. Δεν μπορεί να καθοριστεί ποιο θα είναι το αποτέλεσμα της μαθησιακής διαδικασίας. Κατά την διάρκεια της μαθησιακής διαδικασίας, οι μονάδες (τιμές βάρους) ενός τέτοιου νευρικού δικτύου είναι «διευθετημένες» μέσα σε ένα συγκεκριμένο εύρος, ανάλογα με τις δεδομένες τιμές εισόδου. Ο στόχος είναι η ομαδοποίηση παρόμοιων μονάδων κοντά σε ορισμένες περιοχές του εύρους τιμών.

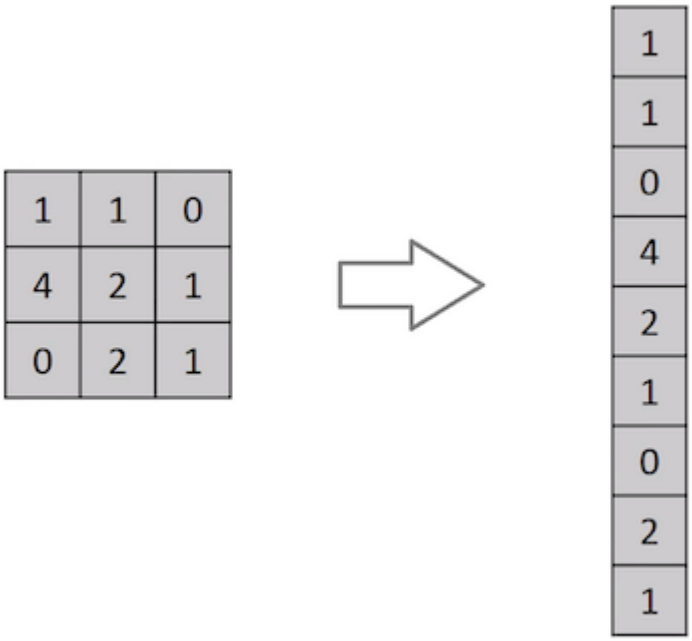
Υποκεφάλαιο 6.3) Συνελικτικά νευρωνικά δίκτυα

(Ενότητα 6.3.α) Αρχιτεκτονική -λειτουργία

Ένα συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network– CNN) είναι ένας αλγόριθμος βαθιάς μάθησης, που μπορεί να λάβει μια εικόνα εισόδου, να αποδώσει μαθησιακά βάρη και πολώσεις (bias) σε διάφορες πτυχές/αντικείμενα της εικόνας ώστε να μπορεί να ξεχωρίσει το ένα από το άλλο. Η προεπεξεργασία που απαιτείται σε ένα CNN είναι πολύ χαμηλότερη σε σύγκριση με άλλους αλγόριθμους ταξινόμησης. .

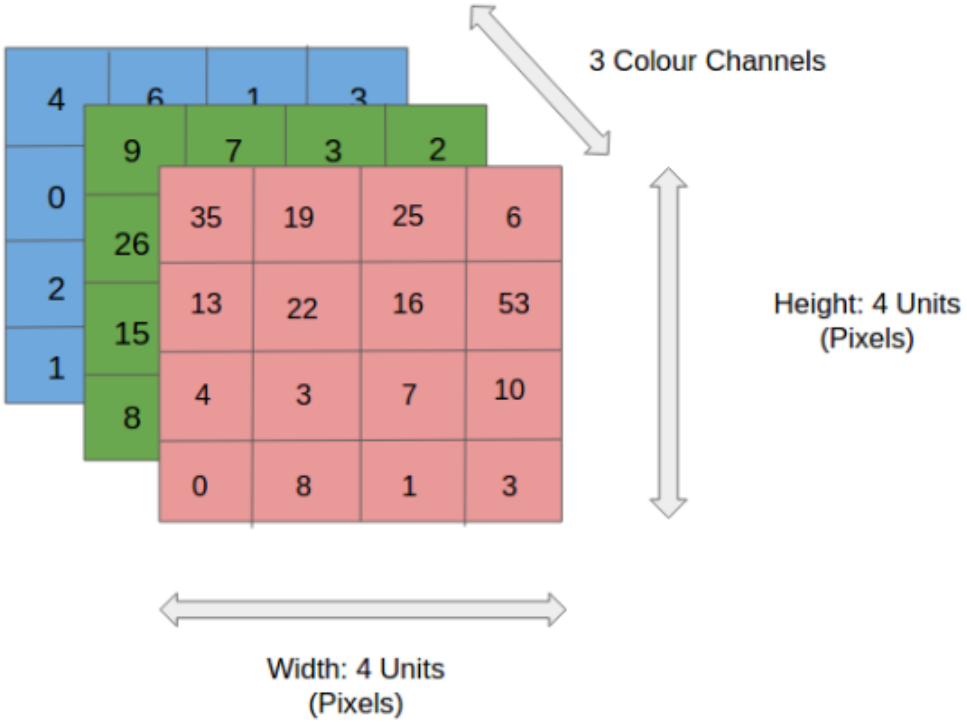


Εικόνα 21 Αρχιτεκτονική ενός συνελικτικού δικτύου με 2 συνελικτικά επίπεδα, 2 max pooling.
(Saha, A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, 2018)



Εικόνα 22 Μετατροπή μιας 3χ3 εικόνας σε μονοδιάστατο διάνυσμα 9 θέσεων
(Saha, A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, 2018)

Η μετατροπή μιας εικόνας 3x3 σε μονοδιάστατο διάνυσμα 9 θέσεων όπως γίνεται σε ένα κλασσικό νευρωνικό δίκτυο δεν μπορεί να αποδώσει σωστά τις χωρικές και χρονικές εξαρτήσεις των pixel. Ένα CNN αντιθέτως είναι σε θέση να αποτυπώσει με επιτυχία τις χωρικές και χρονικές εξαρτήσεις σε μια εικόνα μέσω εφαρμογής σχετικών φίλτρων. Η αρχιτεκτονική προσαρμόζεται καλύτερα στο σύνολο δεδομένων εικόνας λόγω της μείωσης του αριθμού των παραμέτρων που εμπλέκονται και της επαναχρησιμοποίησης των βαρών. Με άλλα λόγια, το δίκτυο μπορεί να εκπαιδευτεί ώστε να κατανοεί καλύτερα την πολυπλοκότητα της εικόνας.



Εικόνα 23: 4x4x3 RGB Image

(Saha, A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, 2018)

Στην παραπάνω εικόνα, έχουμε μια εικόνα RGB που αποτελείται από τρία χρωματικά της επίπεδα - Κόκκινο, Πράσινο και Μπλε. Υπάρχουν πολλοί τέτοιοι χρωματικοί χώροι όπως κλίμακα του γκρι, GBR, HSV, CMYK, κ.λπ.

Ο ρόλος ενός CNN είναι να μειώσει την εικόνα ώστε να είναι εύκολη στην επεξεργασία, χωρίς να χάσει χαρακτηριστικά που είναι κρίσιμα για την επίτευξη καλής πρόβλεψης.

Τα κλασσικά νευρωνικά δίκτυα χρειαζόμαστε πολλούς νευρώνες οι οποίοι διασυνδέονται με όλους τους νευρώνες του επομένου επιπέδου. Αυτή η πλήρης συνδεσιμότητα και ο τεράστιος αριθμός παραμέτρων εκτός της δυσκολίας στην επεξεργασία θα οδηγούσε γρήγορα σε υπερβολική προσαρμογή (overfitting).

Τα συνελικτικά νευρωνικά δίκτυα εκμεταλλεύονται το γεγονός ότι η είσοδος αποτελείται μόνο από εικόνες και περιορίζουν την αρχιτεκτονική. Συγκεκριμένα, σε αντίθεση με ένα κλασικό νευρωνικό δίκτυο, τα επίπεδα ενός CNN έχουν νευρώνες διατεταγμένους σε 3 διαστάσεις: πλάτος, ύψος, βάθος. (η λέξη βάθος εδώ αναφέρεται στην τρίτη διάσταση των χρωμάτων). Οι νευρώνες συνδέονται μόνο με μια μικρή περιοχή του προηγούμενου επιπέδου, αντί με όλους τους νευρώνες όπως συμβαίνει σε ένα κλασσικό νευρωνικό δί-

κτυο. Το τελικό επίπεδο εξόδου θα έχει διαστάσεις 1x1χαριθμό κλάσεων, επειδή μέχρι το τέλος της αρχιτεκτονικής CNN θα μειώσουμε την πλήρη εικόνα σε ένα ενιαίο διάνυσμα βαθμολογιών κλάσης.

Χρησιμοποιούμε τρεις κύριους τύπους επιπέδων για τη δημιουργία αρχιτεκτονικών CNN: είσοδος (Input-IL) συνελικτικό (convolutiona-CLl), συγκέντρωσης (Pooling-PL) , ReLU (RL) και πλήρως συνδεδεμένο (Fully-Connected -FC) επίπεδο.

Το επίπεδο FC (δηλαδή πλήρως συνδεδεμένο, τελευταίο επίπεδο) θα υπολογίσει τις βαθμολογίες της κλάσης, καταλήγοντας σε μέγεθος [1x1x2] για δύο κλάσεις.

(Convolutional Neural Networks (CNNs / ConvNets), χ.χ.)

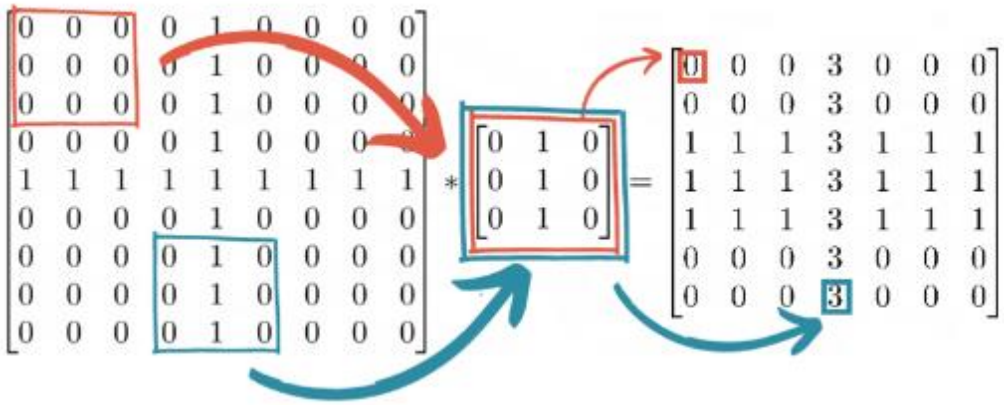
Ένα CNN αποτελείται από 1 input layer, πολλά σετ (convolutional-ReLU-Pooling) και 1 fully connected.

Συνελικτικό επίπεδο (convolutional layer): Σε ένα συνελικτικό νευρωνικό δίκτυο, ένα συνελικτικό επίπεδο είναι υπεύθυνο για τη συστηματική εφαρμογή ενός ή περισσότερων φίλτρων (kernels-πυρήνες) σε μια είσοδο. Τα φίλτρα αυτά είναι τετράγωνα τα οποία γλιστρούν πάνω από την εικόνα και αναζητούν μοτίβα. Ο πολλαπλασιασμός του φίλτρου με την εικόνα εισόδου οδηγεί **σε μία μόνο έξοδο**. Η είσοδος είναι τρισδιάστατες εικόνες (π.χ. σειρές, στήλες και κανάλια) και τα φίλτρα είναι επίσης τρισδιάστατα με τον ίδιο αριθμό καναλιών και λιγότερες σειρές και στήλες από την εικόνα εισόδου. Το φίλτρο εφαρμόζεται επανειλημμένα σε κάθε μέρος της εικόνας εισόδου, με αποτέλεσμα **έναν δισδιάστατο χάρτη εξόδου** των ενεργοποιήσεων, που ονομάζεται χάρτης χαρακτηριστικών. Όπου αυτό το τμήμα της εικόνας ταιριάζει με το μοτίβο του πυρήνα, ο πυρήνας επιστρέφει μια μεγάλη θετική τιμή και όταν δεν υπάρχει αντιστοίχιση, ο πυρήνας επιστρέφει μηδέν ή μικρότερη τιμή.

Έστω ότι θέλουμε να δοκιμάσουμε το φίλτρο 3χ3 ανιχνευτή κάθετης γραμμής στην εικόνα του σταυρού. Για να εκτελέσουμε τη συνέλιξη, σύρουμε τον πυρήνα συνέλιξης πάνω από την εικόνα. Σε κάθε θέση, πολλαπλασιάζουμε κάθε στοιχείο του πυρήνα συνέλιξης με το στοιχείο της εικόνας που καλύπτει και αθροίζουμε τα αποτελέσματα.

Δεδομένου ότι ο πυρήνας έχει πλάτος 3, μπορεί να τοποθετηθεί μόνο σε 7 διαφορετικές θέσεις οριζόντια σε μια εικόνα πλάτους 9. Έτσι το τελικό

αποτέλεσμα της λειτουργίας συνέλιξης σε μια εικόνα μεγέθους 9x9 με έναν πυρήνα συνέλιξης 3x3 είναι μια νέα εικόνα μεγέθους 7x7 .



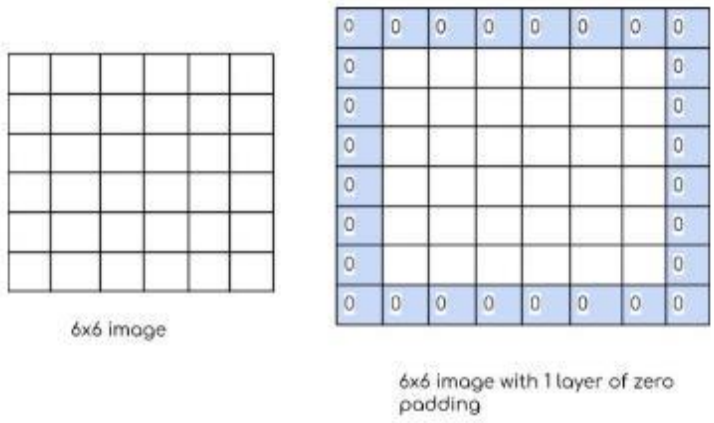
Εικόνα 24 Λειτουργία συνέλιξης με φίλτρο ανιχνευτή κάθετης γραμμής (Wood, n.d.).

Παρατηρούμε όπου υπάρχουν κάθετες γραμμές έχουμε την τιμή 3, όπου έχουμε οριζόντιες την τιμή 1 και όπου δεν υπάρχουν καθόλου γραμμές έχουμε 0.

Στην περίπτωση εικόνων με πολλά κανάλια (π.χ. RGB), ο πυρήνας έχει το ίδιο βάθος με αυτό της εικόνας εισόδου δηλαδή 3. Ο πολλαπλασιασμός μήτρας εκτελείται μεταξύ του κάθε πυρήνα και του κάθε καναλιού και όλα τα αποτελέσματα αθροίζονται με την πόλωση (bias) για να μας δώσουν ένα κανάλι βάθους 1 (convoluted features output).

Ο στόχος της λειτουργίας συνέλιξης είναι να εξαγάγει τα χαρακτηριστικά υψηλού επιπέδου, από την εικόνα εισόδου. Τα CNN συνήθως έχουν πολλά συνελικτικά επίπεδα. Το πρώτο επίπεδο είναι υπεύθυνο για την καταγραφή των χαρακτηριστικών χαμηλού επιπέδου, όπως άκρες, χρώμα, προσανατολισμό κλίσης, κ.λπ. Με πρόσθετα επίπεδα, η αρχιτεκτονική προσαρμόζεται και στα χαρακτηριστικά υψηλού επιπέδου, δίνοντάς μας ένα δίκτυο που έχει την πλήρη κατανόηση εικόνων στο σύνολο δεδομένων.

Υπάρχουν δύο τύποι αποτελεσμάτων α) το συνελιγμένο χαρακτηριστικό το οποίο μειώνεται σε διαστάσεις σε σύγκριση με την είσοδο, με την εφαρμογή «valid padding» χωρίς να προσθέσουμε περίγραμμα με μηδενικά και β) η διάσταση είτε αυξάνεται είτε παραμένει η ίδια, σε αυτήν την περίπτωση προσθέτουμε στην εικόνα περίγραμμα με 0).



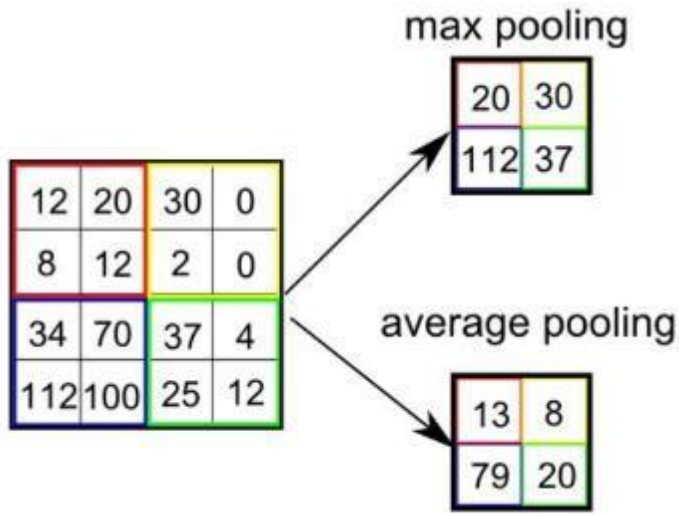
Εικόνα 25 Εικόνα με Padding και χωρίς Padding
(Singirikonda, 2020)

Στην έξοδο που ακολουθεί κάθε επίπεδο συνέλιξης εφαρμόζεται η διορθωμένη συνάρτηση γραμμικής ενεργοποίησης (ReLU). Η ReLU θα εξάγει την τιμή απευθείας την είσοδο εάν είναι θετική, διαφορετικά, θα βγάζει μηδέν.

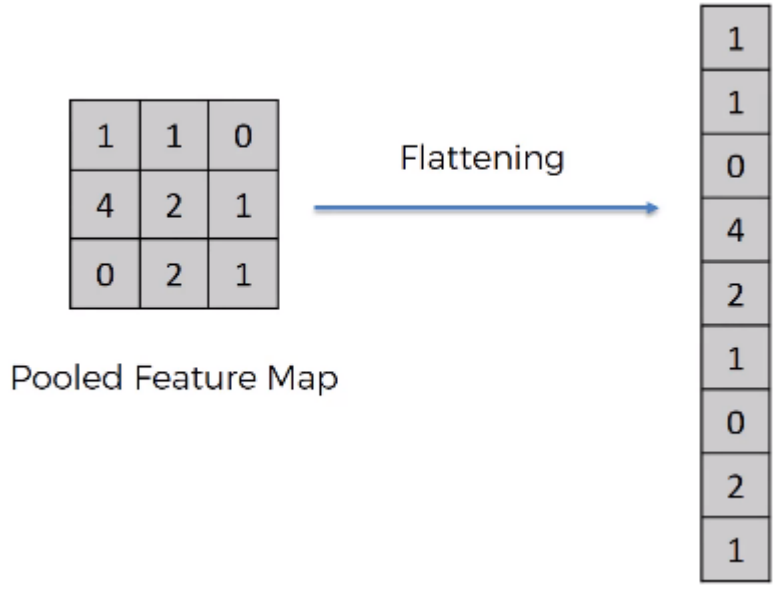
$$f(x) = \max(0, x)$$

Επίπεδο συγκέντρωσης (pooling layer): Όπως το συνελικτικό επίπεδο (convolutional) το επίπεδο συγκέντρωσης (pooling layer) είναι υπεύθυνο για τη μείωση του χωρικού μεγέθους του συνελιγμένων χαρακτηριστικών (convolved features). Αυτό γίνεται για να μειωθεί η υπολογιστική ισχύς που απαιτείται για την επεξεργασία των δεδομένων και επιπλέον, είναι χρήσιμο για την εξαγωγή κυρίαρχων χαρακτηριστικών που είναι αμετάβλητα λόγω περιστροφής και θέσης, διατηρώντας έτσι τη διαδικασία αποτελεσματικής εκπαίδευσης του μοντέλου. Το επίπεδο συγκέντρωσης προστίθεται μετά σε κάθε συνελικτικό επίπεδο, και συγκεκριμένα, μετά την εφαρμογή της συνάρτησης ενεργοποίησης (συνήθως ReLU).

Υπάρχουν δύο τύποι συγκέντρωσης (pooling) : Μέγιστης Τιμής (max pooling) και Μέσης Τιμής (average pooling). Το «max pooling» επιστρέφει τη μέγιστη τιμή από το τμήμα της εικόνας που καλύπτεται από τον πυρήνα και το «average pooling» επιστρέφει τον μέσο όρο όλων των τιμών από το τμήμα της εικόνας που καλύπτεται από τον πυρήνα.

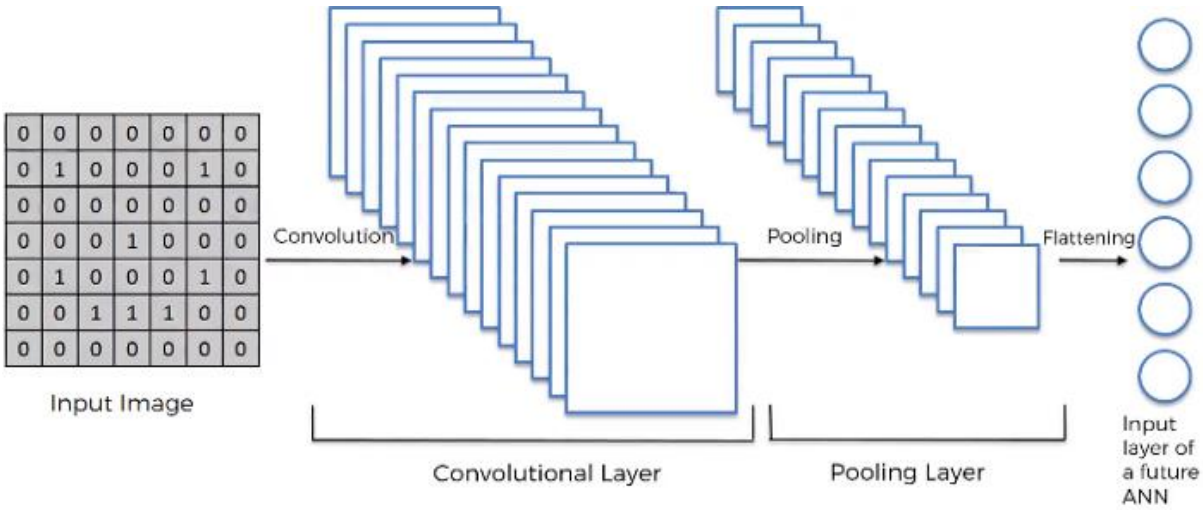


Εικόνα 26 Τύποι pooling

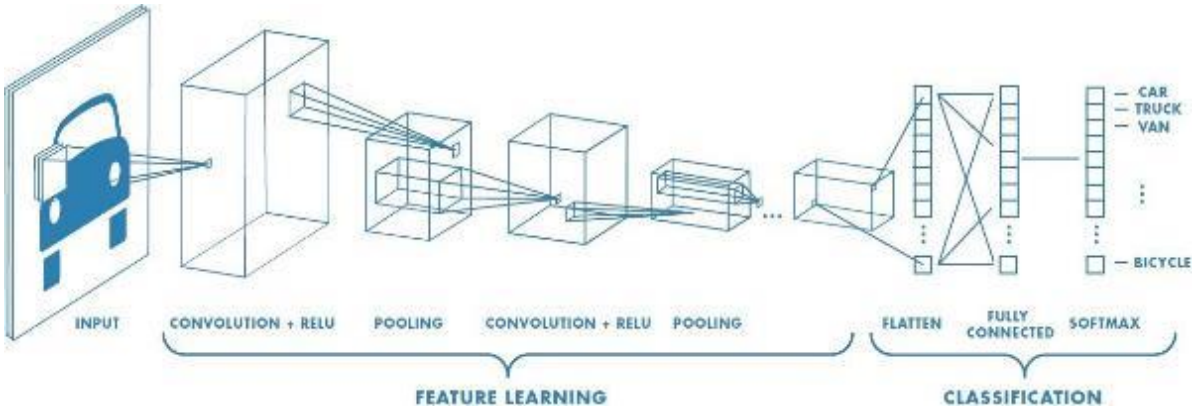


Εικόνα 27 Επιπεδοποίηση pooled feature map
(Girgin, 2019)

Αφού περάσαμε από την διαδικασία είσοδος-(convolution, RelU, Pooling) έχουμε επιτυχώς ενεργοποιήσει το μοντέλο να κατανοήσει τα χαρακτηριστικά. Ακολούθως, θα επιπεδοποιήσουμε την τελική έξοδο και θα την τροφοδοτήσουμε σε ένα κανονικό νευρωνικό δίκτυο για λόγους ταξινόμησης ή θα μεταβιβάσουμε την έξοδο του τελευταίου επιπέδου συγκέντρωσης στο «πλήρως συνδεδεμένο επίπεδο» (fully connected layer).



Εικόνα 28 Είσοδος - convolutional layer -pooling – flattening



Εικόνα 29 Αναπαράσταση πλήρους συνελικτικού νευρωνικού δικτύου.
(Saha, A Comprehensive Guide to Convolutional Neural Networks , 2018)

Η προσθήκη ενός πλήρως συνδεδεμένου επιπέδου είναι ένας τρόπος εκμάθησης μη γραμμικών συνδυασμών των χαρακτηριστικών υψηλού επιπέδου όπως αντιπροσωπεύονται από την έξοδο του συνελικτικού επιπέδου. Το επίπεδο Fully Connected μαθαίνει μια συνάρτηση σε αυτόν τον χώρο και μετατρέπει την εικόνα σε διάνυσμα στήλης. Η επιπεδοποιημένη έξοδος τροφοδοτείται σε ένα νευρωνικό δίκτυο «τροφοδοσίας προς τα εμπρός» (feed forward) και εφαρμόζεται αντίστροφη διάδοση (backward propagation) σε κάθε επανάληψη της εκπαίδευσης. Μετά από μια σειρά επαναλήψεων, το μοντέλο είναι σε θέση να διακρίνει μεταξύ κυρίαρχων χαρακτηριστικών και χαρακτηριστικών χαμηλού επιπέδου σε εικόνες και να τις ταξινομήσει.

Στο τελευταίο επίπεδο εφαρμόζεται η softmax για ταξινόμηση πολλαπλών κλάσεων και η σιγμοειδής για δυαδική ταξινόμηση.

Υπάρχουν διάφορες αρχιτεκτονικές των διαθέσιμων CNN που έχουν παίξει καθοριστικό ρόλο στη δημιουργία αλγορίθμων. Μερικές από αυτές είναι είναι: LeNet,, AlexNet,, VGGNet, Googlenet, Resnet, ZFnet.

(Ενότητα 6.3.β) Επαναληπτική δομή ενός συνελικτικού νευρωνικού δικτύου

Ένα βασικό συνελικτικό νευρωνικό δίκτυο μπορεί να θεωρηθεί ως μια σειρά συνελικτικών επιπέδων, ακολουθούμενη από μια συνάρτηση ενεργοποίησης, ακολουθούμενη από ένα επίπεδο συγκέντρωσης που επαναλαμβάνεται πολλές φορές. Με τον επαναλαμβανόμενο συνδυασμό αυτών των λειτουργιών, το πρώτο επίπεδο ανιχνεύει απλά χαρακτηριστικά, όπως άκρες σε μια εικόνα, και το δεύτερο επίπεδο αρχίζει να ανιχνεύει χαρακτηριστικά υψηλότερου επιπέδου κοκ. Μέχρι το δέκατο επίπεδο, ένα συνελικτικό νευρωνικό δίκτυο είναι σε θέση να ανιχνεύσει πολύπλοκα σχήματα. Μέχρι το εικοστό επίπεδο μπορεί να ξεχωρίσει ανθρώπινα πρόσωπα..

(Υποκεφάλαιο 6.4) Πλήρως συνδεδεμένα έναντι συνελικτικών νευρωνικών δικτύων

Ένα συνελικτικό νευρωνικό δίκτυο είναι ένα ειδικό είδος νευρωνικού δικτύου με λιγότερες συνδέσεις από ένα πλήρως συνδεδεμένο δίκτυο. Σε ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο εμπρόσθιας προώθησης, κάθε νευρώνας συνδέεται με όλους τους του επομένου επιπέδου και δεν υπάρχει πυρήνας συνέλιξης ο οποίος μειώνει τον αριθμό των χαρακτηριστικών.

Είναι σαφές ότι ένα συνελικτικό νευρωνικό δίκτυο χρησιμοποιεί πολύ λιγότερες παραμέτρους (βάρη) από το ισοδύναμο πλήρως συνδεδεμένο νευρωνικό δίκτυο με τα ίδια επίπεδα. Το τελικό επίπεδο ενός συνελικτικού νευρωνικού δικτύου είναι συνήθως πλήρως συνδεδεμένο.

Η διαδικασία εκπαίδευσης ενός συνελικτικού νευρωνικού δικτύου είναι ουσιαστικά η ίδια με την εκπαίδευση οποιουδήποτε άλλου νευρωνικού δικτύου. Αρχικά, το δίκτυο δημιουργείται με τυχαίες τιμές σε όλα τα βάρη και τις πολώσεις. Ακολούθως και στα δύο ακολουθείται επαναληπτικά το feed-forward – backpropagation για ένα αριθμό φορών (epoch).

(Υποκεφάλαιο 6.5) Πυκνά συνδεδεμένα συνελικτικά Δίκτυα

Η συνολική αρχιτεκτονική ενός CNN αποτελείται από δύο βασικά μέρη: έναν εξαγωγέα χαρακτηριστικών και έναν ταξινομητή. Τα επίπεδα συνέλιξης (convolution) και συγκέντρωσης (pooling) είναι τα δύο βασικά επίπεδα της

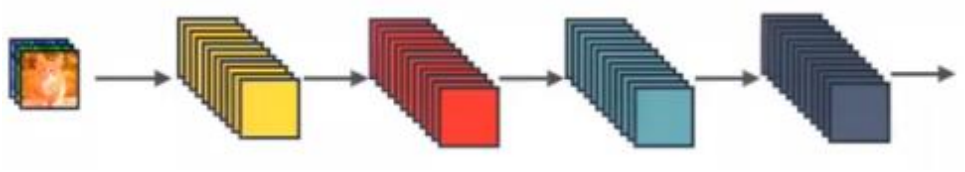
αρχιτεκτονικής του CNN. Κάθε νευρώνας στο επίπεδο συνέλιξης εξάγει χαρακτηριστικά από τις εικόνες εισόδου εκτελώντας μια λειτουργία συνέλιξης. Το επίπεδο max-pooling αφαιρεί τα χαρακτηριστικά υπολογίζοντας τον μέσο όρο ή υπολογίζοντας τη μέγιστη τιμή των κόμβων εισόδου.

Το πρόβλημα που προκύπτει για τα CNN όταν αποκτήσουν πολλά επίπεδα, η διαδρομή για τις πληροφορίες από το επίπεδο εισόδου μέχρι το επίπεδο εξόδου (και για την κλίση προς την αντίθετη κατεύθυνση) γίνεται τόσο μεγάλη, που μπορούν να εξαφανιστούν πριν φτάσουν στην άλλη πλευρά. Αυτό έχει ως αποτέλεσμα η κλίση να γίνει 0 ή πολύ μεγάλη. Έτσι, όταν αυξάνουμε τον αριθμό των επιπέδων, το ποσοστό σφαλμάτων εκπαίδευσης και δοκιμής αυξάνεται επίσης.

Για να αντιμετωπισθεί αυτό το πρόβλημα της εξαφάνισης κλίσης οι

(He, Xiangyu Zhang, Shaoqing Ren, & Sun, 2015) πρότειναν το ResNet, συντομογραφία του Residual Network. Σε αυτό το δίκτυο, χρησιμοποιούμε μια τεχνική που ονομάζεται παράκαμψη συνδέσεων. Δηλαδή συνδέουμε τις ενεργοποιήσεις ενός επιπέδου με άλλα επιπεδα παρακάμπτοντας ορισμένα ενδιάμεσα επίπεδα.

Το πλεονέκτημα της προσθήκης αυτού του τύπου «σύνδεσης παράκαμψης» είναι ότι εάν οποιοδήποτε επίπεδο βλάψει την απόδοση της αρχιτεκτονικής, μπορεί να παραλειφθεί. Αυτό έχει ως αποτέλεσμα την εκπαίδευση ενός πολύ βαθέως νευρωνικού δικτύου χωρίς τα προβλήματα που προκαλούνται από την εξαφάνιση της κλίσης.

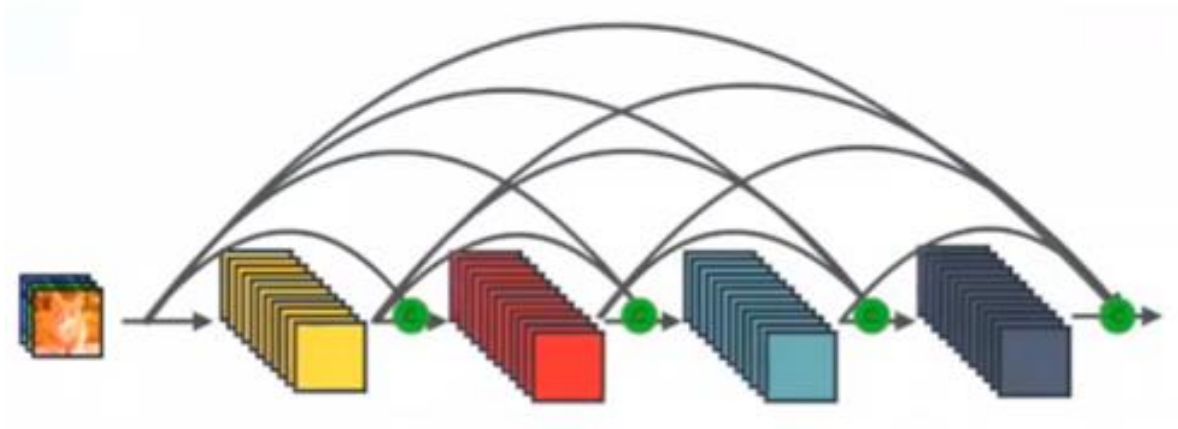


Εικόνα 30: Η γενική ιδέα ενός κλασικού CNN
(Tsang, Review: DenseNet — Dense Convolutional Network (Image Classification), 2018)

Οι (Gao Huang, Liu, Laurens van der Maaten, & Weinberger, 2018) πρότειναν τα Densely Connected Convolutional Networks που αντιμετωπίζουν καλύτερα το πρόβλημα. Τα «Πυκνά Συνελικτικά Δίκτυα ή DenseNets είναι μια αρχιτεκτονική CNN όπου τα επίπεδα του δικτύου χωρίζονται σε blocks. Μέσα σε ένα block κάθε επίπεδο συνδέεται με όλα τα προηγούμενα. Στο Dense

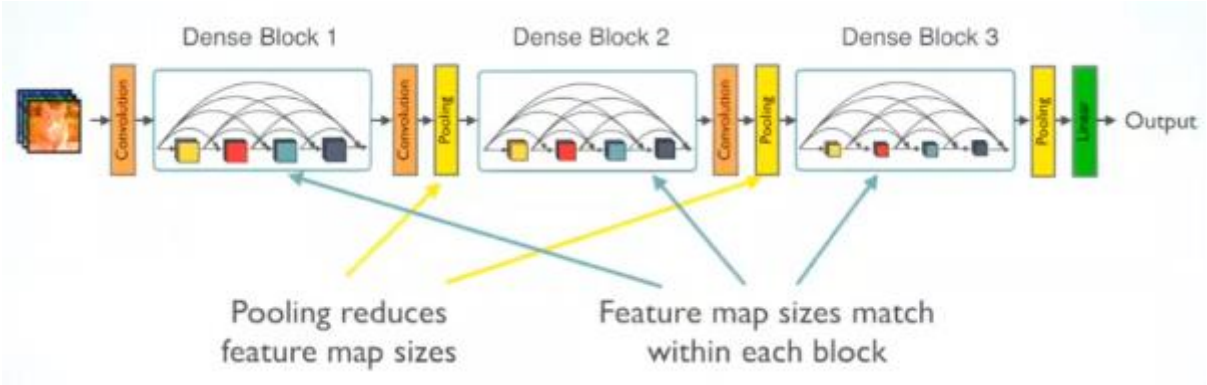
Block, το κάθε επίπεδο παίρνει ως είσοδο όλους τους χάρτες χαρακτηριστικών $x_1, x_2, x_3 \ldots . x_l$ από τα προηγούμενα επιπεδα, το οποίο περιγράφεται από : $x_l = H_1([x_1, x_2, x_3, \ldots . x_l])$.[36] Μέσα σε κάθε μπλοκ το μέγεθος των χαρτών χαρακτηριστικών παραμένει το ίδιο ώστε να μπορούν να ενωθούν μεταξύ τους.

Το DenseNet είναι ένα δίκτυο υψηλής εποπτείας που περιέχει Dense-Blocks που το καθένα έχει ένα αριθμό επιπέδων. Η έξοδος κάθε επιπέδου σε ένα «πυκνό μπλοκ» (dense block) περιλαμβάνει την έξοδο όλων των προηγούμενων επιπέδων, ενσωματώνοντας χαρακτηριστικά χαμηλού και υψηλού επιπέδου της εικόνας εισόδου.



Εικόνα 31 Dense Block
(Tsang, Review: DenseNet — Dense Convolutional Network (Image Classification), 2018)

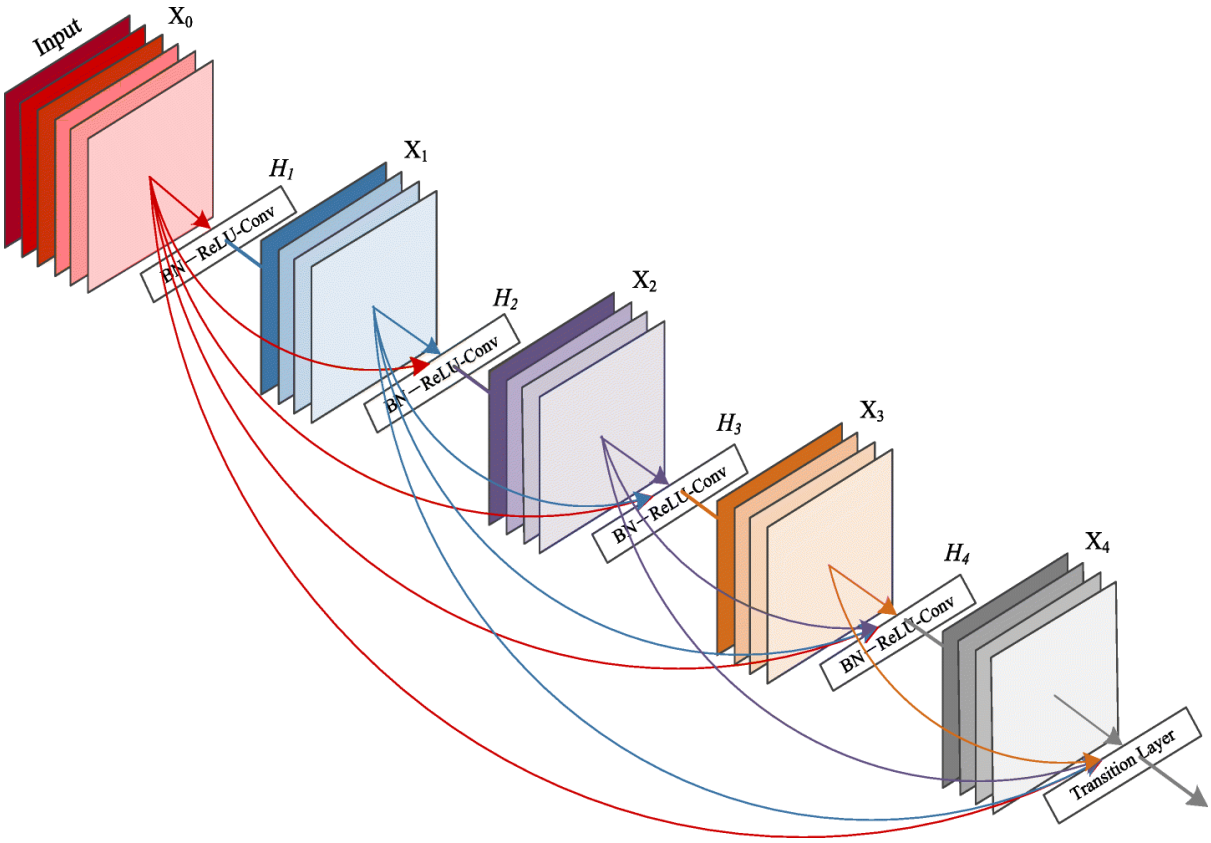
.



Εικόνα 32 Διαδικασία μεταφοράς χαρακτηριστικών σε ένα DenseNet
(Tsang, Review: DenseNet — Dense Convolutional Network (Image Classification), 2018)

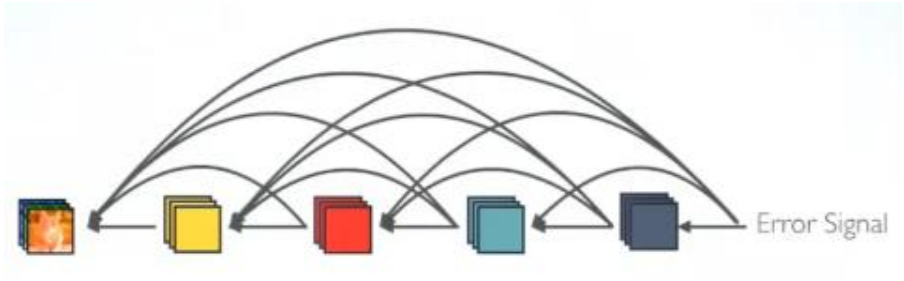
[36] Συνένωση χαρτών χαρακτηριστικών (concatenation of feature maps) από όλα τα προηγούμενα επίπεδα.

Ενδιάμεσα από δύο Dense blocks υπάρχουν ένα convolutional επίπεδο 1x1 ακολουθούμενο από ένα 2×2 επίπεδο «μέσου συνόλου» (average pooling).



Εικόνα 33: Ένα DenseBlock με 4 set BN [37], RelU, Convolutional
(Ruiz, Understanding and visualizing DenseNets, 2018)

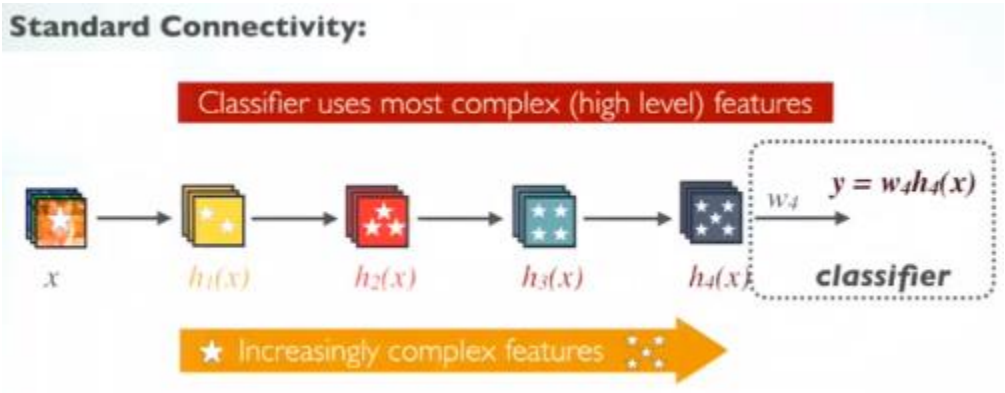
Ως μεταβατικά στρώματα μεταξύ δύο συνεχόμενων πυκνών μπλοκ χρησιμοποιούνται 1×1 μετατροπείς (Convolutional Πυρήνες) ακολουθούμενοι από μέση συγκέντρωση μέσου (averaged pooling) 2×2.



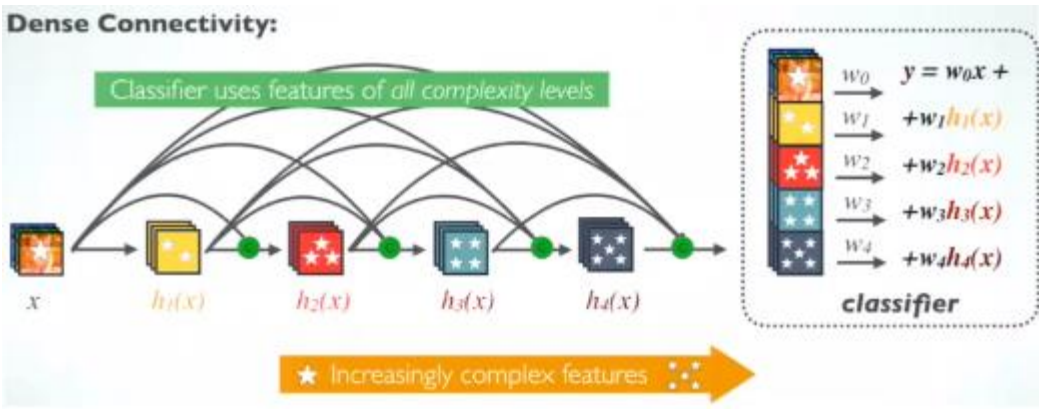
Εικόνα 34: Μετάδοση σήματος λάθους στο DenseNet

[37] BN= Bach Normalization: *είναι μια τεχνική για την τυποποίηση των εισόδων σε ένα δίκτυο, που εφαρμόζεται στις ενεργοποιήσεις ενός προηγούμενου επιπέδου ή των εισόδων απευθείας* (βλ. (Ενότητα 3.ζ.) Προσαρμογή – Κλιμάκωση Χαρακτηριστικών)

Το σήμα σφάλματος μπορεί εύκολα να μεταδοθεί σε προηγούμενα στρώματα πιο άμεσα.



Εικόνα 35 Κλασσικό CCN



Εικόνα 36: Το DenseNet χρησιμοποιεί χαρακτηριστικά όλων των επίπεδων.
(Tsang, Review: DenseNet — Dense Convolutional Network (Image Classification), 2018)

Το DenseNet αποδίδει καλά όταν τα δεδομένα εκπαίδευσης είναι ανεπαρκή, καθώς το DenseNet χρησιμοποιεί λειτουργίες όλων των επιπέδων πολυπλοκότητας.

Τα DenseNets-B είναι απλά τα κανονικά DenseNets που εκμεταλλεύονται τη συνέλιξη 1x1 για να μειώσουν το μέγεθος των χαρτών χαρακτηριστικών πριν από τη συνέλιξη 3x3.

«Το DenseNet χρησιμοποιεί συνέλιξη 1x1 και avg pooling 2X2 ως μεταβατικά επίπεδα (transition layers) μεταξύ γειτονικών «πυκνών μπλοκ» (Dense Blocks). Πραγματοποιείται μια συνολική μέση συγκέντρωση (avg pooling) στο τέλος του τελευταίου πυκνού μπλοκ και στη συνέχεια συνδέεται ένας ταξινομητής Softmax.» (M. Adnan, 2018)

(Υποκεφάλαιο 6.6) Μάθηση με μεταφορά – προεκπαιδευμένα μοντέλα.

(Ενότητα 6.6.α) Γενικά

Στη μάθηση με μεταφορά (transfer learning), η γνώση ενός ήδη εκπαιδευμένου μοντέλου μηχανικής μάθησης εφαρμόζεται σε ένα διαφορετικό αλλά σχετικό πρόβλημα. Για παράδειγμα, εάν εκπαιδεύσουμε έναν απλό ταξινομητή για να προβλέψουμε εάν μια εικόνα περιέχει ένα κατοικίδιο, θα μπορούσαμε να χρησιμοποιήσουμε τη γνώση που απέκτησε το μοντέλο κατά την εκπαίδευσή του για να αναγνωρίσει άλλα αντικείμενα όπως πχ άγρια ζώα.

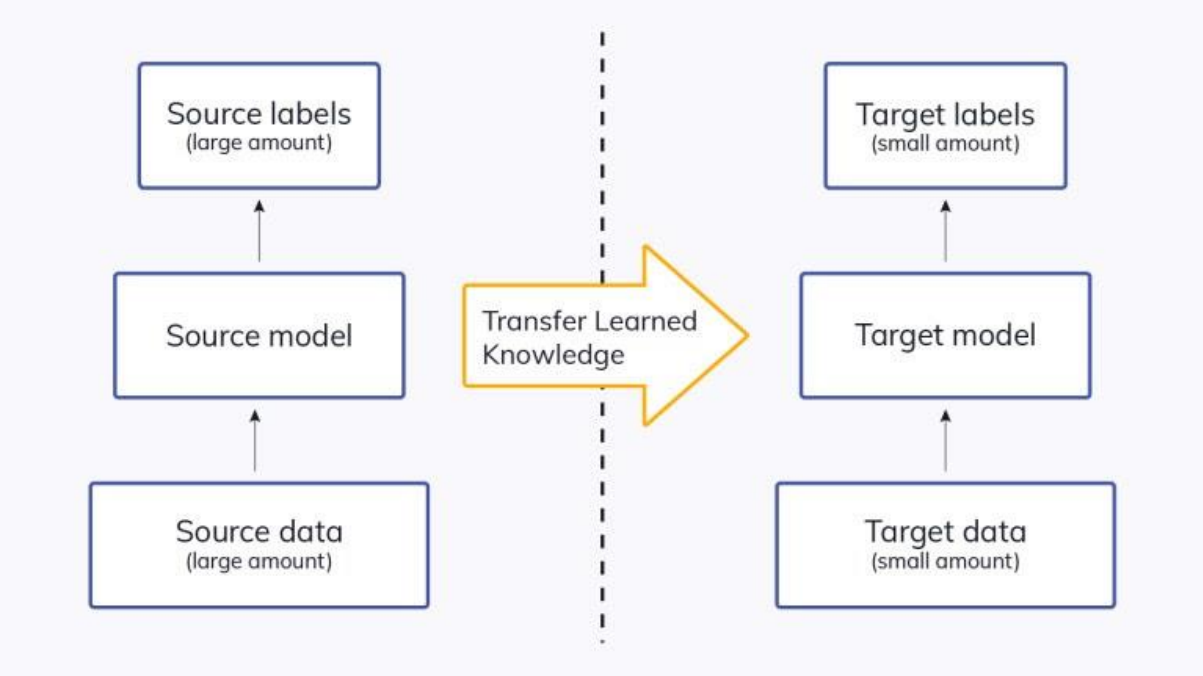
Με τη μάθηση με μεταφορά, προσπαθούμε να εκμεταλλευτούμε όσα έχουμε «μάθει» σε μια εργασία για να βελτιώσουμε τη δυνατότητα πρόβλεψης σε μια άλλη. Μεταφέρουμε τα βάρη που έχει μάθει ένα δίκτυο στην «εργασία Α» σε μια νέα «εργασία Β».

Αυτός ο τύπος μάθησης είναι πολύ δημοφιλής στη «βαθιά μάθηση», επειδή μπορεί να εκπαιδεύσει βαθιά νευρωνικά δίκτυα με σχετικά λίγα δεδομένα..

Στην υπολογιστική όραση (Computer Vision), για παράδειγμα, τα νευρωνικά δίκτυα συνήθως προσπαθούν να ανιχνεύσουν ακμές στα πρώτα επίπεδα, σχήματα στο μεσαίο επίπεδο και ορισμένα χαρακτηριστικά ειδικά για την εργασία που εκτελούν στα τελευταία επίπεδα. Στη μάθηση με μεταφορά, χρησιμοποιείται η έξοδος των πρώτων και μεσαίων επίπεδων και επανεκπαιδεύσουμε μόνο τα τελευταία.

Τα κύρια πλεονεκτήματα είναι η εξοικονόμηση χρόνου εκπαίδευσης, η καλύτερη απόδοση των νευρωνικών δικτύων και η μη ανάγκη πολλών δεδομένων για να εκπαιδεύσουμε το δίκτυο για την εργασία που θέλουμε, αφού ήδη μεγάλο ποσοστό της εργασίας έγινε από άλλο μοντέλο.

Συνήθως, απαιτούνται πολλά δεδομένα για την εκπαίδευση ενός νευρωνικού δικτύου από την αρχή, αλλά η πρόσβαση σε αυτά τα δεδομένα δεν είναι πάντα διαθέσιμη και γι' αυτό είναι χρήσιμη η μάθηση με μεταφορά. Με την μάθηση με μεταφορά μπορεί να κατασκευαστεί ένα συμπαγές μοντέλο μηχανικής μάθησης με σχετικά λίγα δεδομένα εκπαίδευσης, επειδή το μοντέλο είναι ήδη **προεκπαιδευμένο**.



Εικόνα 37 Μάθηση με μεταφορά
(Baheti, 2022)

Η μάθηση με μεταφορά ενδείκνυται να χρησιμοποιείται όταν:

- δεν υπάρχουν αρκετά επισημειωμένα δεδομένα για να εκπαιδεύσουμε το δίκτυό μας επαρκώς.
- υπάρχει ήδη ένα δίκτυο που είναι προεκπαιδευμένο σε παρόμοια εργασία, το οποίο συνήθως έχει εκπαιδευτεί σε τεράστιες ποσότητες δεδομένων.
- Όταν δύο εργασίες έχουν την ίδια είσοδο.

Η μάθηση με μεταφορά λειτουργεί μόνο εάν τα χαρακτηριστικά που μαθαίνονται από την πρώτη εργασία είναι γενικά, που σημαίνει ότι μπορούν να είναι χρήσιμα και για άλλη σχετική εργασία. Επίσης, η είσοδος του μοντέλου πρέπει να έχει το ίδιο μέγεθος με το οποίο εκπαιδεύτηκε αρχικά. Εάν δεν συμβαίνει αυτό, πρέπει, προσθέσουμε ένα βήμα προεπεξεργασίας για να αλλάξουμε το μέγεθος της εισόδου στο απαιτούμενο μέγεθος (πχ το μέγεθος μιας εικόνας).

(Ενότητα 6.6.β) Χρήση προεκπαιδευμένων μοντέλων

Μπορούμε να χρησιμοποιήσουμε τα προεκπαιδευμένα μοντέλα με διάφορους τρόπους:

-Εάν θέλουμε να εκτελέσουμε την εργασία Α, αλλά δεν έχουμε αρκετά δεδομένα για να εκπαιδεύσουμε ένα βαθύ νευρωνικό δίκτυο, ένας τρόπος είναι να βρούμε μια σχετική εργασία Β με άφθονα δεδομένα, να εκπαιδεύσουμε

ένα βαθύ νευρωνικό δίκτυο στην εργασία Β και να χρησιμοποιήσουμε το μοντέλο Β ως σημείο εκκίνησης για την επίλυση της εργασίας Α.

-Να αγνοήσουμε την εκπαίδευσή τους και να χρησιμοποιήσουμε μόνο την αρχιτεκτονική τους, όπως σε όλους τους αλγορίθμους, δηλαδή να τα αρχικοποιήσουμε και να τα εκπαιδεύσουμε το μοντέλο με τα δεδομένα που έχουμε και να δημιουργήσουμε ένα νέο μοντέλο που μπορεί να κάνει προβλέψεις για το πρόβλημά μας.

-Μπορούμε να τροποποιήσουμε την έξοδο ή την είσοδο αναλόγως με το πρόβλημά μας. Μπορούμε να τροποποιήσουμε πχ την έξοδο ενός μοντέλου που έχει προεκπαιδευτεί στο ImageNet, ώστε να προβλέπει μόνο δύο κατηγορίες (πχ άλογο ή σκύλος)

-Μπορούμε να κρατήσουμε («παγώσουμε») την εκπαίδευση ορισμένων (αρχικών επίπεδων) και να επανεκπαιδεύσουμε τα υπόλοιπα. Πόσα επίπεδα θα επαναχρησιμοποιηθούν και πόσα θα επανεκπαιδευτούν εξαρτάται από το πρόβλημα. Μπορούμε με δοκιμές να αποφασίσουμε πόσα επίπεδα πρέπει να «παγώσουν» και πόσα να επανεκπαιδευτούν.

-Μπορούμε να χρησιμοποιήσουμε ένα προεκπαιδευμένο μοντέλο ως μηχανισμό εξαγωγής χαρακτηριστικών. Συνήθως αφαιρέσουμε το επίπεδο εξόδου (αυτό που δίνει τις κλάσεις για τις οποίες έχει προεκπαιδευτεί), «παγώνουμε» τα υπόλοιπα επίπεδα και στη συνέχεια τα χρησιμοποιήσουμε ως εξαγωγέα χαρακτηριστικών για το νέο σύνολο δεδομένων δηλαδή για να κάνει προβλέψεις για τα νέα δεδομένα.

Γενικά το μέγεθος και η ομοιότητα μεταξύ του παλιού και νέου μοντέλου μας δείχνει πως ενδείκνυται να χρησιμοποιούνται τα προεκπαιδευμένα μοντέλα.

«

Σενάριο 1: Το μέγεθος του συνόλου δεδομένων είναι μικρό ενώ η ομοιότητα των δεδομένων είναι πολύ υψηλή. Σε αυτήν την περίπτωση, καθώς η ομοιότητα δεδομένων είναι πολύ υψηλή, δεν χρειάζεται να επανεκπαιδεύσουμε το μοντέλο. Το μόνο που χρειάζεται είναι να προσαρμόσουμε και να τροποποιήσουμε τα επίπεδα εξόδου σύμφωνα με τη δήλωση του προβλήματος μας. Ας υποθέσουμε ότι αποφασίσαμε να χρησιμοποιήσουμε μοντέλα τα οποία έχουν εκπαιδευτεί στο ImageNet[38] για να προσδιορίσουμε εάν το νέο σύνολο

[38] *Το σύνολο δεδομένων* ImageNet *έχει χρησιμοποιηθεί ευρέως για την κατασκευή διαφόρων αρχιτεκτονικών, καθώς είναι αρκετά μεγάλο*

εικόνων που έχει γάτες ή σκύλους. Εδώ, οι εικόνες που πρέπει να αναγνωρίσουμε θα είναι παρόμοιες με το imageNet, ωστόσο χρειαζόμαστε μόνο δύο κατηγορίες ως έξοδο - γάτες ή σκύλους. Σε αυτήν την περίπτωση το μόνο που κάνουμε είναι απλώς να τροποποιήσουμε το τελικό επίπεδο softmax ώστε να βγουν 2 κατηγορίες.

Σενάριο 2: Το μέγεθος των δεδομένων είναι μικρό και η ομοιότητα δεδομένων πολύ χαμηλή. Σε αυτήν την περίπτωση μπορούμε να παγώσουμε τα αρχικά επίπεδα του προεκπαιδευμένου μοντέλου και να εκπαιδεύσουμε ξανά μόνο τα υπόλοιπα επίπεδα. Στη συνέχεια, τα ανώτερα επίπεδα θα προσαρμοστούν στο νέο σύνολο δεδομένων. Δεδομένου ότι το νέο σύνολο δεδομένων έχει χαμηλή ομοιότητα, είναι σημαντικό να επανεκπαιδεύσουμε και να προσαρμόσουμε τα υψηλότερα επίπεδα σύμφωνα με το νέο σύνολο δεδομένων. Το μικρό μέγεθος του συνόλου δεδομένων αντισταθμίζεται από το γεγονός ότι τα αρχικά στρώματα διατηρούνται προεκπαιδευμένα (έχουν εκπαιδευτεί προηγουμένως σε ένα μεγάλο σύνολο δεδομένων).

Σενάριο 3: Το μέγεθος του συνόλου δεδομένων είναι μεγάλο, και η ομοιότητα δεδομένων είναι πολύ χαμηλή. Καθώς τα δεδομένα που έχουμε είναι πολύ διαφορετικά σε σύγκριση με τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των προεκπαιδευμένων μοντέλων μας, οι προβλέψεις που γίνονται χρησιμοποιώντας αυτά τα προεκπαιδευμένα μοντέλα δεν θα ήταν αποτελεσματικές. Γι' αυτό, είναι καλύτερο να εκπαιδεύσουμε το δίκτυο από την αρχή σύμφωνα με τα δεδομένα μας.

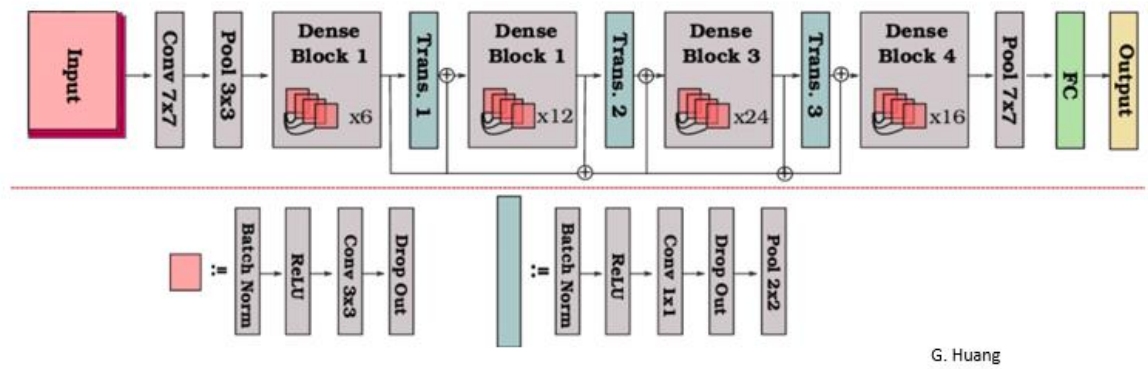
Σενάριο 4: Το μέγεθος των δεδομένων είναι μεγάλο και η ομοιότητα δεδομένων είναι υψηλή: Αυτή είναι η ιδανική κατάσταση καθώς μπορούμε να χρησιμοποιήσουμε το μοντέλο ως έχει, διατηρώντας και την αρχιτεκτονική και τα αρχικά βάρη και να το χρησιμοποιήσουμε απευθείας για να κάνουμε προβλέψεις»

(Gupta D. , 2021)

(1,2M εικόνες) για να δημιουργήσει ένα γενικευμένο μοντέλο. Τα προεκπαιδευμένα μοντέλα με το ImageNet μπορούν να ταξινομήσουν σωστά τις εικόνες σε 20000 ξεχωριστές κατηγορίες αντικειμένων. Αυτές οι 20.000 κατηγορίες εικόνων αντιπροσωπεύουν κατηγορίες αντικειμένων που συναντάμε στην καθημερινή μας ζωή, όπως είδη σκύλων, γατών, διάφορα οικιακά αντικείμενα, τύπους οχημάτων κ.λπ.

(Ενότητα 6.6.γ) Προεκπαιδευμένα DenseNets

Μία σημαντική περίπτωση προεκπαιδευμένων μοντέλων με την βάση δεδομένων ImageNet η οποία περιέχει πάνω από 14 εκατομμύρια εικόνες επισημειωμένες με το χέρι σε πάνω από 20.000 κατηγορίες, είναι τα DenseNets. Αυτά είναι τα: DenseNet121, DenseNet169, DenseNet201 και DenseNet 264. Ο αριθμός που τα ακολουθεί αναφέρεται στον συνολικό αριθμό των επιπέδων. Στην παρούσα εργασία έχει χρησιμοποιηθεί το DenseNet169 ως εξαγωγέας χαρακτηριστικών.



Εικόνα 38: Αρχιτεκτονική DenseNet 121
(Introduction to DenseNet with TensorFlow, 2020) από
(Tsang, Review: DenseNet — Dense Convolutional Network (Image Classification), 2018)

Στο παραπάνω σχήμα:

- Είσοδος έγχρωμη (GBR) 224 x 224.
- Ακολουθεί συνελικτικού επίπεδο 7X7, με βήμα (stride[39]) = 2
- Ακολουθεί επίπεδο συγκέντρωσης (pooling) 3x3 με βήμα 2.
- Ακολουθούν 4 Dense Blocks
  - 1ο Block (Conv 1x1, Conv 3x3) x6
  - 2ο Block (Conv 1x1, Conv 3x3) x12
  - 3ο Block (Conv 1x1, Conv 3x3) x24
  - 4ο Block (Conv 1x1, Conv 3x3) x16
- 3 Μεταβατικά επίπεδα (Transition layers)
  - Conv 1x1
  - Avg Pooling 2x2
- 1 7x7 Convolution
- 58 3x3 Convolution
- 61 1x1 Convolution
- 4 Avg Pooling
- 1 Fully Connected

[39] Το Stride είναι μια παράμετρος του φίλτρου νευρωνικού δικτύου που τροποποιεί το μέγεθος της κίνησης πάνω από την εικόνα. Για παράδειγμα, εάν το stride=1 οριστεί στο 1, το φίλτρο θα

Σύνολο 121 επίπεδα.

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv; 3 × 3 conv] × 24	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48	[1 × 1 conv; 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv; 3 × 3 conv] × 16	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Εικόνα 39 Το μέγεθος των εξόδων και συνελικτικών πυρήνων των διαφόρων προεκπαιδευμένων DenseNet στο ImageNET.

(Tsang, Review: DenseNet — Dense Convolutional Network (Image Classification), 2018)

ΚΕΦΑΛΑΙΟ 7 ΔΟΚΙΜΕΣ ΚΑΙ ΕΠΙΚΥΡΩΣΗ

(Υποκεφάλαιο 7.1) Συλλογή – επιλογή δεδομένων

Για να εκπαιδεύσουμε κάποιο μοντέλο με μια από τις εποπτευόμενες μεθόδους μηχανικής μάθησης ή βαθιάς μάθησης πρέπει να έχουμε **επαρκή ποσότητα** δεδομένων ώστε να εκπαιδευτεί το μοντέλο μας δηλαδή να «συλλάβει» την σχέση που συνδέει τα δεδομένα μας με την κατηγορία που ανήκουν.

Αλλά και **η ποιότητα** των δεδομένων μας παίζει σημαντικό ρόλο. Η συλλογή των δεδομένων πρέπει να είναι τυχαία, αντιπροσωπευτική και να έχει γίνει από τον πραγματικό κόσμο, η δε επισημείωσή τους να έχει γίνει από κατάλληλα άτομα. Εάν η ποσότητα των επισημειωμένων δεδομένων δεν επαρκεί μπορούμε να χρησιμοποιήσουμε μία μέθοδο ημι-εποπτευόμενης μάθησης για να επισημειώσουμε αυτόματα ορισμένα από αυτά. Επίσης μπορούμε να χρησιμοποιήσουμε μία από τις μεθόδους ενεργούς μάθησης ώστε να επιλέξουμε προς επισημείωση τα πλέον κατάλληλα μη επισημειωμένα προς επισημείωση. Με Λιγότερα αλλά ποιοτικότερα δεδομένα μπορούμε να πέτυχουμε καλύτερα αποτελέσματα.

Μέσα στην ποιότητα των δεδομένων μας συμπεριλαμβάνουμε και την ποικιλία, δηλαδή τα δεδομένα μας πρέπει προέρχονται από όλο των φάσμα του τομέα που εξετάζουμε. Πχ εάν πρόκειται για ακτινογραφίες πρέπει να έχουμε δείγματα από διάφορα διαγνωστικά κέντρα και μηχανήματα αλλιώς κινδυνεύουμε να έχουμε αυτό που λέμε πόλωση (bias).

Αφού κατασκευάσουμε και εκπαιδεύσουμε το μοντέλο μας η επόμενη φάση περιλαμβάνει την διενέργεια δοκιμών (tests) ώστε να μετρήσουμε την αξιοπιστία των μοντέλου μας να κάνει προβλέψεις σε γνωστά (σε αυτά που έχει εκπαιδευτεί) αλλά κυρίως σε άγνωστα δεδομένα.

(Υποκεφάλαιο 7.2) Διαχωρισμός δεδομένων

Είθισται να χωρίζουμε τα δεδομένα σε 2-3 κατηγορίες για την διεξαγωγή των δοκιμών και της επικύρωσης. Αυτό το κάνουμε ώστε να μπορούμε να αξιολογήσουμε και επικυρώσουμε το μοντέλο με διαφορετικά δεδομένα από αυτά της εκπαίδευσης. Κάθε σετ δεδομένων αποτελείται από ένα σύνολο διανυσμάτων χαρακτηριστικών με τις αντίστοιχες επισημειώσεις.

Σετ εκπαίδευσης (training set): Το σετ εκπαίδευσης χρησιμοποιείται για την κατασκευή ενός μοντέλου με την διαδικασία της εκπαίδευσης. Αποτελείται από ένα σύνολο χαρακτηριστικών που χρησιμοποιούνται από τους διαφόρους αλγορίθμους μηχανικής ή βαθιάς μάθησης για την εκπαίδευση του μοντέλου. Οι αλγόριθμοι εκπαίδευσης έχουν ως σκοπό να συσχετίσουν τα δεδομένα εισόδου με τις αποφάσεις εξόδου. Το σύστημα εκπαιδεύεται με την εφαρμογή αυτών των αλγορίθμων στο σετ των δεδομένων εκπαίδευσης. Γενικά, το 75-80% των δεδομένων του συνόλου δεδομένων λαμβάνονται ως δεδομένα εκπαίδευσης.

Σετ δοκιμών (test set): Τα δεδομένα δοκιμών χρησιμοποιούνται για τον έλεγχο της αποτελεσματικότητας του μοντέλου που έχουμε εκπαιδεύσει. Είναι το σύνολο των δεδομένων που χρησιμοποιείται για να επαληθεύσουν εάν το σύστημα παράγει τη σωστή έξοδο μετά την εκπαίδευση. Γενικά, το 20-25% του συνόλου δεδομένων χρησιμοποιείται για δοκιμές.

Σετ επικύρωσης (validation set): Αυτό είναι ένα διαφορετικό σετ από το σετ δοκιμών. Το σετ δοκιμών χρησιμοποιείται για την αξιολόγηση του μοντέλου μετά την εκπαίδευση, ενώ το σετ επικύρωσης χρησιμοποιείται για την βελτιστοποίηση των «υπερπαραμέτρων» του μοντέλου.

(Υποκεφάλαιο 7.4) Πόλωση – διακύμανση, υποπροσαρμογή - υπερπροσαρμογή

Πόλωση (Bias): «Η διαφορά μεταξύ της μέσης πρόβλεψης του μοντέλου μας και της σωστής τιμής που προσπαθούμε να προβλέψουμε. Ένα μοντέλο με υψηλή πόλωση δίνει πολύ λίγη προσοχή στα δεδομένα εκπαίδευσης και υπεραπλουστεύει το μοντέλο. Οδηγεί πάντα σε υψηλό σφάλμα στα δεδομένα εκπαίδευσης και δοκιμών» (Singh, 2018). Δηλαδή, εάν τα δεδομένα εκπαίδευσης δεν έχουν την ποικιλομορφία ή την ποσότητα που απαιτείται τότε το μοντέλο μας μπορεί να αποδίδει καλύτερα σε ορισμένα δεδομένα από άλλα.

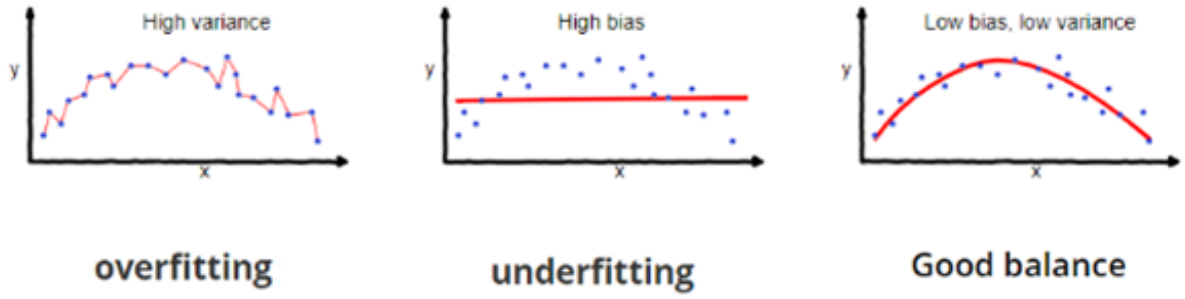
Διακύμανση (Variance): «Το ποσό κατά το οποίο θα αλλάξει η εκτίμηση της συνάρτησης στόχου εάν χρησιμοποιηθούν διαφορετικά δεδομένα από αυτά της εκπαίδευσης.» (Brownlee, 2019). «Το μοντέλο με υψηλή διακύμανση δίνει μεγάλη προσοχή στα δεδομένα εκπαίδευσης και «δεν γενικεύει» (δεν μπορεί να κάνει προβλέψεις) για δεδομένα που δεν έχει δει πριν. Ως αποτέλεσμα, τέτοια μοντέλα αποδίδουν πολύ καλά στα δεδομένα εκπαίδευσης, αλλά έχουν υψηλά ποσοστά σφάλματος στα δεδομένα δοκιμής» (Singh, 2018).

«Ο στόχος οποιουδήποτε αλγόριθμου εποπτευόμενης μηχανικής μάθησης είναι η επίτευξη χαμηλών bias και variance. Η παραμετροποίηση των αλγορίθμων μηχανικής μάθησης είναι μια μάχη για την εξισορρόπηση των bias και variance». (Brownlee, 2019)

«**Υπερπροσαρμογή (overfitting),** συμβαίνει όταν ένα μοντέλο ταιριάζει ακριβώς με τα εκπαιδευτικά δεδομένα. Όταν συμβαίνει αυτό, ο αλγόριθμος δεν μπορεί να αποδώσει με ακρίβεια σε δεδομένα που δεν έχουν παρατηρηθεί. Χαμηλό bias και υψηλό variance είναι ένδειξη για overfitting» (Singh, 2018) Πχ πολύ μεγάλη ακρίβεια του μοντέλου με τα δεδομένα που εκπαιδεύτηκε σε αντίθεση με την μικρή ακρίβεια στα δεδομένα δοκιμών. Μπορούμε να ελαττώσουμε τις πιθανότητες για υπερπροσαρμογή:

1. Μεγεθύνοντας το σύνολο δεδομένων και χρησιμοποιώντας τεχνικές αύξησης όπως αναστροφή, περιστροφή, μεγέθυνση κ.λπ.
2. Χρησιμοποιώντας τεχνικές αποβολής (drop out), αδρανοποιώντας μονάδες του μοντέλου κατά την διάρκεια της εκπαίδευσης πχ θέτοντας 0 σαν είσοδο σε ορισμένους νευρώνες.
3. Μειώνοντας τα χαρακτηριστικά.
4. Μία από τις καλές τεχνικές είναι να κάνουμε πρόωρη διακοπή. Σε όποια επανάληψη βλέπουμε ότι το μοντέλο πάει για υπερπροσαρμογή την διακόπτουμε.
5. Χρήση διασταυρούμενης επικύρωσης για εκπαίδευση/δοκιμή του μοντέλου και παρατήρηση της διαφοράς μεταξύ των μετρικών εκπαίδευσης και δοκιμών μας δίνει μία καλή ένδειξη για υπερπροσαρμογή.

Υποπροσαρμογή (underfitting), συμβαίνει όταν ένα μοντέλο δεν είναι σε θέση να συλλάβει με ακρίβεια τη σχέση μεταξύ των μεταβλητών εισόδου και εξόδου, δημιουργώντας υψηλό ποσοστό σφάλματος. Υψηλό bias και χαμηλό variance είναι ένδειξη για underfitting» (Singh, 2018). Μικρή ακρίβεια του μοντέλου στις δοκιμές των δεδομένων εκπαίδευσης είναι ένδειξη για underfitting. Επίσης εάν κάνουμε την διαδικασία της διασταυρούμενης επικύρωσης και μετρήσουμε την τυπικά απόκλιση μεταξύ των μετρικών που μας δίνει κάθε fold και διαπιστώσουμε ότι υπάρχει μεγάλη διαφορά τότε έχουμε υποπροσαρμογή.



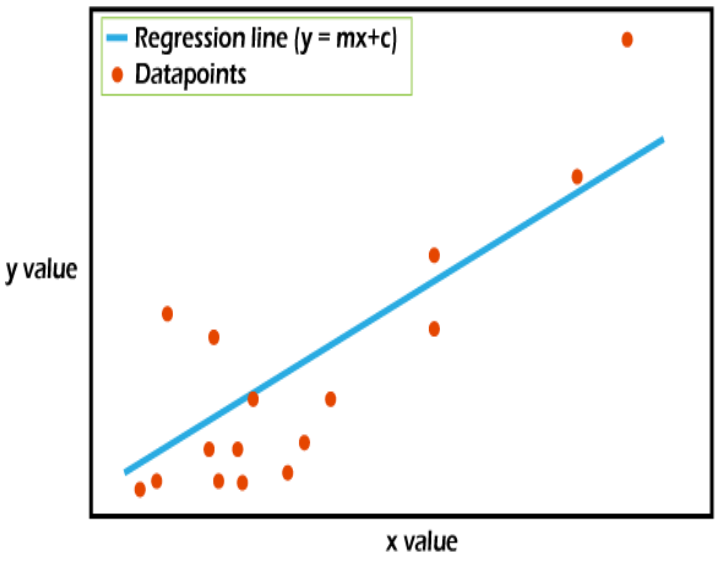
Εικόνα 40 Υπερπροσαρμογή -υποπροσαρμογή – ισορροπημένη εφαρμογή
(Singh, 2018).

(Υποκεφάλαιο 7.5) Ρύθμιση υπερπαραμέτρων

Στην μηχανική και βαθιά μάθηση υπάρχει διαφορά στους όρους παράμετροι και υπερπαράμετροι.

Οι παράμετροι ενός μοντέλου είναι μεταβλητές διαμόρφωσης που είναι εσωτερικές στο μοντέλο και ένα μοντέλο τις μαθαίνει μόνο του. Για παράδειγμα, τα βάρη ή οι συντελεστές ανεξάρτητων μεταβλητών στο μοντέλο γραμμικής παλινδρόμησης. Βάρη ή συντελεστές ανεξάρτητων μεταβλητών στο SVM, βάρος και πολώσεις ενός νευρωνικού δικτύου κλπ.

Μπορούμε να κατανοήσουμε τις παραμέτρους του μοντέλου χρησιμοποιώντας την παρακάτω εικόνα:



Εικόνα 41: Παράμετροι και υπερπαράμετροι
(Difference between Model Parameter and Hyperparameter, n.d.)

Το παραπάνω διάγραμμα δείχνει την αναπαράσταση του μοντέλου της απλής γραμμικής παλινδρόμησης. Εδώ, το x είναι μια ανεξάρτητη μεταβλητή,

το y είναι η εξαρτημένη μεταβλητή και ο στόχος είναι να προσαρμόσουμε την καλύτερη γραμμή παλινδρόμησης για τα δεδομένα για να ορίσουμε μια σχέση μεταξύ x και y. Η γραμμή παλινδρόμησης μπορεί να δοθεί από την εξίσωση: y= mx+c. Οι m και c είναι παράμετροι.

Μερικά βασικά σημεία για τις παραμέτρους του μοντέλου είναι τα εξής:

- Το μοντέλο τις χρησιμοποιεί για να κάνει προβλέψεις.
- Οι τιμές τους γίνονται γνωστές μετά την εκπαίδευση και αποτελούν μέρος του μοντέλου που προήλθε από την εκπαίδευση.
- Δεν μπορούμε να τις ρυθμίσουμε χειροκίνητα.

Υπερπαράμετροι από την άλλη πλευρά είναι εκείνες οι παράμετροι που ορίζονται ρητά από τον χρήστη για τον έλεγχο της εκπαιδευτικής διαδικασίας. Οι υπερπαράμετροι χρησιμοποιούνται από τον αλγόριθμο μάθησης κατά την διάρκεια της εκπαίδευσης, αλλά δεν αποτελούν μέρος του προκύπτοντος μοντέλου. Αυτές εξαρτώνται από το είδος του αλγορίθμου που χρησιμοποιούμε. Μπορεί να είναι πχ για ένα νευρωνικό δίκτυο: ο αριθμός των επαναλήψεων (epoch), ο ρυθμός εκμάθησης (learning rate), ο αριθμός των κρυφών επιπέδων και κόμβων κλπ. Για ένα αλγόριθμο boosting μπορεί να είναι: το μέγιστο βάθος των δένδρων, ο αριθμός των δένδρων, ο ρυθμός εκμάθησης κλπ. Η ρύθμιση των υπερπαραμετρων είναι σημαντικό μέρος και απαιτεί αρκετή εργασία η οποία καλό είναι να γίνει συστηματικά. Μπορεί να γίνει χειροκίνητα και αυτοματοποιημένα. Όταν συντονίζουμε τις υπερπαραμέτρους χειροκίνητα χρησιμοποιούμε επαναληπτικά την τεχνική δοκιμή και λάθος .«Οι αυτοματοποιημένες μέθοδοι συντονισμού υπερπαραμέτρων χρησιμοποιούν έναν αλγόριθμο για την αναζήτηση των βέλτιστων τιμών. Μερικές από τις πιο δημοφιλείς αυτοματοποιημένες μεθόδους σήμερα είναι η αναζήτηση πλέγματος, η τυχαία αναζήτηση και η Bayesian βελτιστοποίηση». (Navas, 2022)

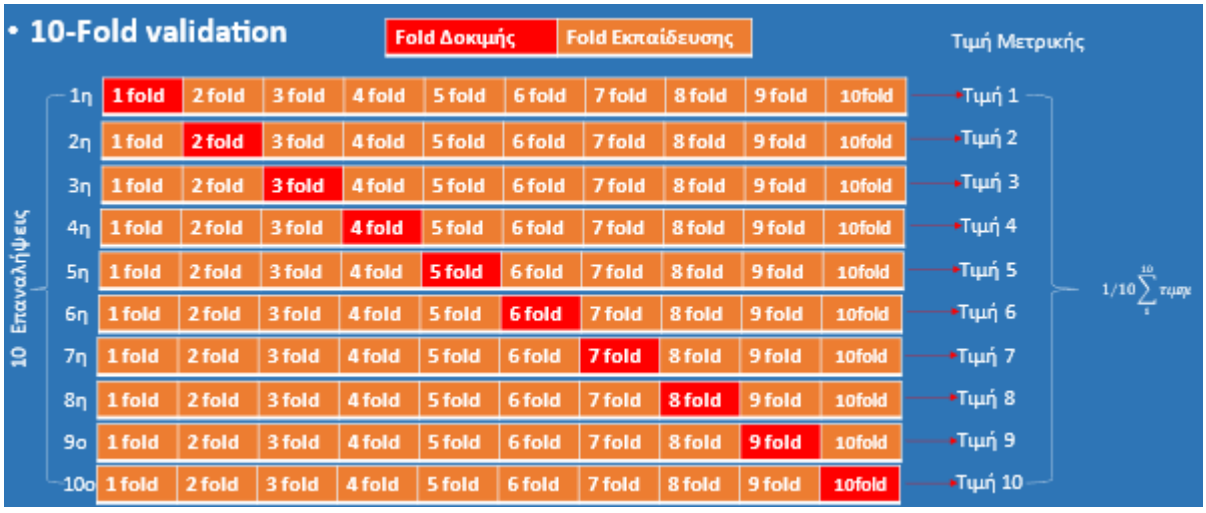
(Υποκεφάλαιο 7.6) Εκτέλεση δοκιμών

Αφού περάσουμε το στάδιο της διακρίβωσης της ποσότητας, ποιότητας και ποικιλομορφίας των δεδομένων και της ρύθμισης των υπερπαραμετρων περνάμε στο στάδιο των δοκιμών. Σε αυτό το στάδιο γίνεται ο διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης (συνήθως το 75-80%) των δεδομένων μας και σε δεδομένα δοκιμών (τα υπόλοιπα). Ακολούθως εκτελούμε την εκπαίδευση με τα δεδομένα εκπαίδευσης και κάνουμε προβλέψεις με τα δεδο-

μένα δοκιμών. Συγκρίνουμε τις πραγματικές κλάσεις με αυτές που προέβλεψε το εκπαιδευμένο μοντέλο και υπολογίζουμε τις μετρικές (βλ. παρακάτω), για να καθορίσουμε την αξιοπιστία του μοντέλου μας.

Για να επιβεβαιώσουμε την αξία των προβλέψεων μας αυτές εκτελούνται πολλές φορές με διαφορετικό σετ εκπαίδευσης. Επιπρόσθετα κάνουμε έλεγχο για υπερπροσαρμογή ή υποπροσαρμογή (βλ. παραπάνω).

Ένας πιο ολοκληρωμένος τρόπος να γίνουν αυτά είναι η διασταυρούμενη επικύρωση (k-fold, cross-validation) που είναι μια διαδικασία δειγματοληψίας που χρησιμοποιείται για την αξιολόγηση μοντέλων μηχανικής μάθησης. Το k αναφέρεται στον αριθμό ομάδων στις οποίες πρέπει να διαιρεθεί ένα δείγμα δεδομένων. Μία από αυτές τις ομάδες την καθορίζουμε ως δεδομένα δοκιμής και τις υπόλοιπες (κ-1) ως δεδομένα εκπαίδευσης. Εκπαιδεύουμε σύμφωνα με κάποιο μοντέλο και το αξιολογούμε με την ομάδα δεδομένων δοκιμής (μπορούμε ταυτόχρονα και εκπαίδευσης). Αφού επαναλάβουμε την διαδικασία, ούτως ώστε κάθε ομάδα να γίνει ομάδα δοκιμής, συνοψίζουμε τα αποτελέσματα των αξιολογήσεων (μέση τιμή ή την τυπική απόκλιση).
Η τιμή k πρέπει να επιλέγεται έτσι ώστε κάθε ομάδα δειγμάτων δοκιμής να είναι αρκετά μεγάλη ώστε να είναι στατιστικά αντιπροσωπευτική του ευρύτερου συνόλου δεδομένων.



Εικόνα 42 10-fold cross validation

Επίσης καλό είναι να βλέπουμε και τα αποτελέσματα των μετρικών με προβλέψεις που γίνονται με τα δεδομένα εκπαίδευσης. Η σύγκριση των αποτελεσμάτων με προβλέψεις που γίνονται με τα δεδομένα δοκιμών και αυτών που γίνονται με τα δεδομένα εκπαίδευσης σε συνδυασμό με την τυπική

απόκλιση μπορούν να μας δώσουν χρήσιμες πληροφορίες για υπερπροσαρμογή ή υποπροσαρμογή.

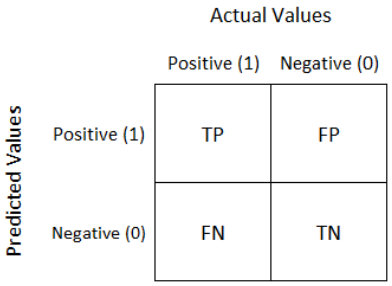
(Υποκεφάλαιο 7.6) Μετρικές

Μετρικές είναι οι μετρήσεις που γίνονται για να αξιολογηθεί η απόδοση κάποιου μοντέλου. Οι ονομασίες των διαφόρων μετρικών θα γίνει στα αγγλικά είτε επειδή η ακριβής μετάφραση σε ορισμένες περιπτώσεις μπορεί να δημιουργήσει σύγχυση π.χ. accuracy και precision

(Ενότητα 7.6.α) Μετρικές ταξινόμησης

Confusion matrix

Confusion matrix είναι ένας πίνακας με 4 διαφορετικούς συνδυασμούς προβλεπόμενων και πραγματικών τιμών.



Εικόνα 43 Confusion matrix

(Narkhede, 2018)

Αληθινά θετικό (True Positive-TP): Προβλέψαμε θετικό και είναι αλήθεια.

Αληθινά αρνητικό (True Negative-TN) : Προβλέψατε αρνητικό και είναι αλήθεια.

Ψευδώς θετικό (False Positive - FP): Προβλέψαμε θετικά και είναι λάθος (Σφάλμα τύπου Ι)-

Ψευδώς αρνητικό (False Negative - FN): Προβλέψαμε αρνητικό και είναι λάθος. (Σφάλμα τύπου ΙΙ):

Στον «confusion matrix» μπορούν να τοποθετηθούν είτε ακέραιες τιμές, είτε πραγματικές που αντιστοιχούν σε ποσοστά, είτε και τα δύο.

Precision.

Το μέτρο των σωστά αναγνωρισμένων θετικών περιπτώσεων από όλες τις προβλεπόμενες θετικές περιπτώσεις.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

Recall ή sensitivity

Το μέτρο των σωστά αναγνωρισμένων θετικών περιπτώσεων από όλες τις πραγματικές θετικές περιπτώσεις.

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$$

Accuracy

Το μέτρο όλων των σωστά προσδιορισμένων περιπτώσεων.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative})}$$

F1-Score

Ο αρμονικός μέσος όρος της Precision.

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Specificity

Η ειδικότητα είναι η μέτρηση που αξιολογεί την ικανότητα ενός μοντέλου να προβλέπει τα TN.

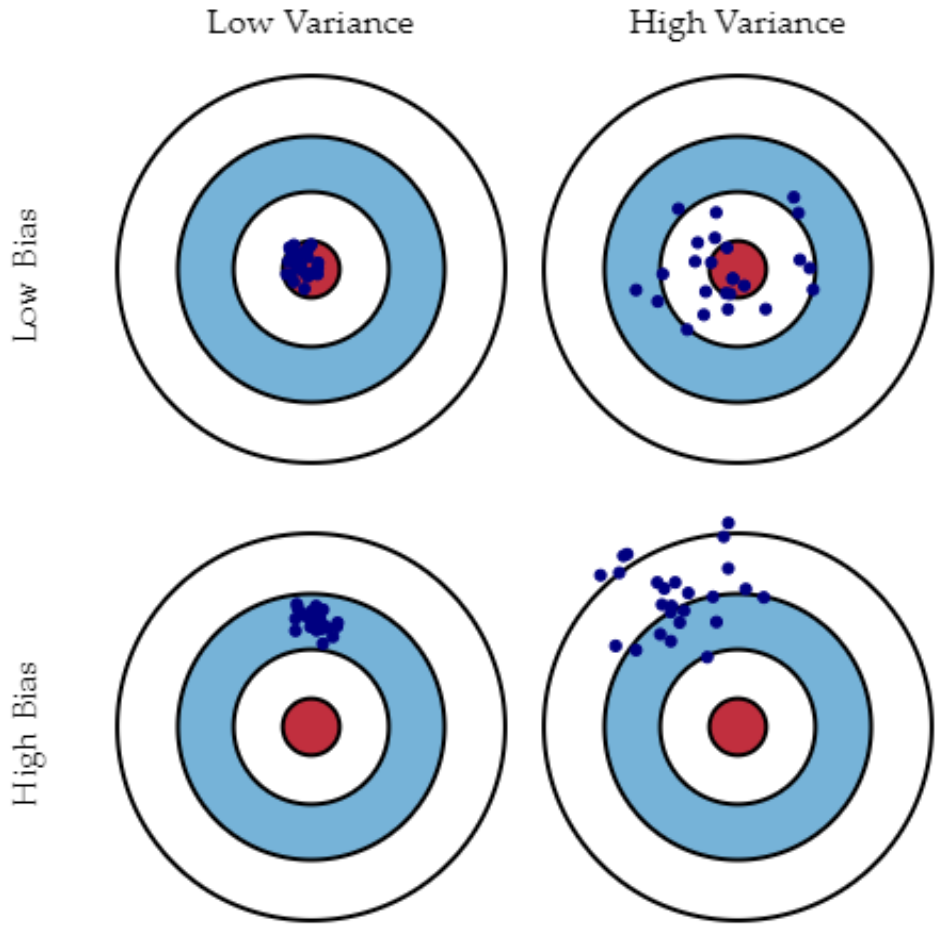
$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

(Huilgol, 2019)

Εάν υπάρχουν πάνω από δύο κλάσεις υπάρχουν διάφορες προσεγγίσεις οι κυριότερες είναι: α) Macro Averaged: Υπολογίζουμε την μετρική για όλες τις κλάσεις και μετά παίρνουμε την μέση τιμή. β) Micro averaged: Υπολογίζουμε τα true positive, true negative, false positive, false negative για κάθε κλάση και μετά χρησιμοποιούμε αυτά για να υπολογίσουμε τις μετρικές (Gupta A. , n.d.)

(Ενότητα 7.6.β) Σημασία αποτελεσμάτων δοκιμών

Γενικά πρέπει να έχουμε υπόψιν μας ότι οι μετρικές που έχουν τιμή πολύ μεγάλη στα δεδομένα εκπαίδευσης και μικρή στα δεδομένα δοκιμών μας δίνουν μία ένδειξη ύπαρξης υπερπροσαρμογής (overfitting). Όταν υπάρχει μεγάλη τυπική απόκλιση μεταξύ των μετρικών δοκιμών τότε έχουμε ένδειξη υποπροσαρμογής (underfitting).



Εικόνα 44: Υπερπροσαρμογή-υποπροσαρμογή σε σχέση την πόλωση και διακύμανση
(Fortmann-Roe, 2012)

Μία χαμηλή βαθμολογία σε οποιαδήποτε μετρική μπορεί να σημαίνει, κακή επιλογή ταξινομητή, λάθος ρύθμιση υπερπαραμετρων ή και λάθος στην διαδικασία συλλογής και προετοιμασίας των δεδομένων.

Πρέπει να συναξιολογούμε το σύνολο των μετρικών για να διαπιστώσουμε την ικανότητα ενός μοντέλου. Μία μετρική από μόνη της είναι ανεπαρκής να αξιολογήσει το μοντέλο.

Όταν προσπαθούμε να βελτιώσουμε το precision χειροτερεύουμε το sensitivity και αντιστρόφως. Πχ Εάν προσπαθήσουμε να μειώσουμε τις περιπτώσεις «μη ασθενών» που επισημειώνονται ως «ασθενείς» (sensitivity: FN/Type-II error), δεν σημαίνει ότι θα υπάρξει βελτίωση στην μετρική που μετράει τους ασθενείς που χαρακτηρίζονται ως μη υγιείς (precision: FP/Type-I error). Τις περισσότερες φορές μάλιστα όταν αυξάνεται το ένα μειώνεται το άλλο και αντιστρόφως.

Precision

Η μέτρηση precision απαντάει στο ερώτημα «Ποιο ποσοστό των θετικών ταυτοποιήσεων ήταν πραγματικά σωστό;». Εστιάζει στο σφάλμα τύπου I (FP), δηλαδή όταν χαρακτηρίζουμε εσφαλμένα ανθρώπους που είναι ασθενείς ως υγιείς. Μια βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινό θετικό (TP). Αυτό που δεν μπορεί να μετρήσει είναι η ύπαρξη σφάλματος Τύπου II δηλαδή ψευδώς αρνητικών, δηλαδή οι περιπτώσεις που ένας υγιής αναγνωρίζεται ως ασθενής.

Sensitivity (recall)

Η ευαισθησία (sensitivity) απαντά στο ερώτημα «Ποιο ποσοστό των πραγματικών θετικών προσδιορίστηκε σωστά; Η ευαισθησία ενός μοντέλου είναι η αναλογία των ατόμων που αναγνωρίστηκαν θετικοί για τη νόσο μεταξύ εκείνων που πράγματι έχουν τη νόσο.

Μία βαθμολογία προς το 100% θα σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινό θετικό. Αυτό που δεν μπορεί να μετρήσει είναι οι περιπτώσεις που κάποιος προσδιορίζεται υγιής ενώ είναι ασθενής.

Η ευαισθησία ενός μοντέλου είναι η ικανότητα του τεστ να εντοπίζει σωστά άτομα με τη νόσο (πραγματικό θετικό ποσοστό).

Specificity

Η ειδικότητα απαντά στο ερώτημα ποιο ποσοστό των πραγματικά αρνητικών προσδιορίστηκε σωστά. Η ειδικότητα είναι η αναλογία των ατόμων που αναγνωρίστηκαν αρνητικοί για τη νόσο μεταξύ εκείνων που δεν έχουν τη νόσο.

Μία βαθμολογία προς το 100% θα σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά αρνητικό. Αυτό που δεν μπορεί να μετρήσει είναι οι περιπτώσεις που κάποιος προσδιορίζεται ασθενής ενώ είναι υγιής.

Η ειδικότητα του τεστ είναι η ικανότητα του τεστ να εντοπίζει σωστά άτομα χωρίς τη νόσο (πραγματικό αρνητικό ποσοστό).

F1-score

«Η μετρική F1-score είναι ένας συνδυασμός precision και sensitivity (αρμονικός μέσος). Πολύ υψηλή βαθμολογία σημαίνει ότι και οι δύο μετρικές που συνδυάζονται έχουν καλές τιμές, πολύ χαμηλή βαθμολογία σημαίνει ότι precision και sensitivity δεν έχουν καλές ενδείξεις. Μέτρια βαθμολογία σημαίνει ότι μια από τις precision και sensitivity δεν έχει καλή τιμή ή και οι δύο έχουν μέτριες».

(Bajaj, 2019)

Είναι δείκτης κατάλληλος να μετρήσει την αξιοπιστία μοντέλων που δεν έχουν ισορροπία στα δεδομένα (πολλά δεδομένα μιας κλάσης και πολύ λιγότερα άλλης).

Accuracy

Είναι το πηλίκο όλων των σωστών προβλέψεων προς το σύνολο των προβλέψεων. Η accuracy ειδικά όταν υπάρχει ανισορροπία (μεγάλη διαφορά μεταξύ του αριθμού των θετικών και αρνητικών) δεν μας λέει «την αλήθεια». Πχ έχουμε accuracy 90%, που σημαίνει 90 σωστές προβλέψεις από 100 δείγματα. Εκ πρώτης όψεως φαίνεται ότι ο ταξινομητής μας κάνει πολύ καλή δουλειά. Αν όμως από τα 100 δείγματα ασθενών τα 90 είναι υγιή άτομα (89 TN και 1 FP) και τα 10 είναι ασθενείς (1TP και 9FN). Από τα 90 άτομα το μοντέλο προσδιορίζει σωστά ως υγιείς τους 89, που είναι καλό αλλά από τους 10 ασθενείς το μοντέλο προσδιορίζει σωστά μόνο 1. Το μοντέλο μας δηλαδή δεν έχει προγνωστική ακρίβεια στην διάκριση των ασθενών από τους μη ασθενείς. Γιατρό στις περιπτώσεις αυτές που έχουμε μη ισορροπημένα δεδομένα, καλύτερη μετρική για την είναι η f1-score.

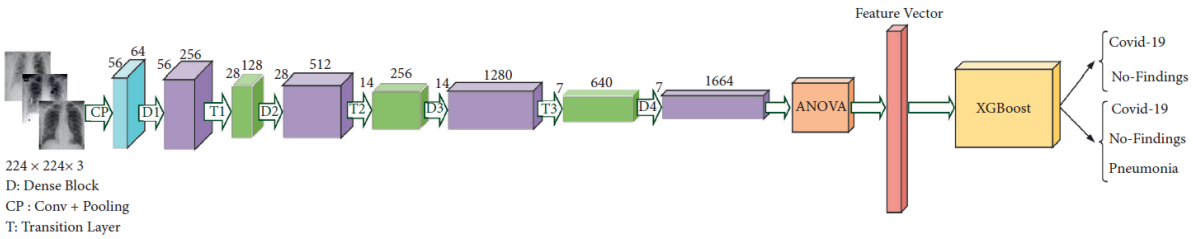
(Machine Leraning, n.d.)

ΚΕΦΑΛΑΙΟ 8 ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΜΕ ΜΕΡΙΚΩΣ ΕΠΙΣΗΜΕΙΩΜΕΝΑ ΔΕΔΟΜΕΝΑ

(Υποκεφάλαιο 8.1) Βάση εκκίνησης

Σαν βάση εκκίνησης για την εργασία αυτή έχει επιλεγεί μια μελέτη πλήρως εποπτευόμενης μάθησης, η οποία δίνει υψηλά ποσοστά αξιοπιστίας στις προβλέψεις «COVID, NON COVID». Αυτή είναι η εργασία «A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images» των (Hasani & Nasiri, 2021)

Στην εργασία αυτή τα χαρακτηριστικά που εξήχθησαν με την βοήθεια του DenseNet169 αφού μειώθηκαν σε αριθμό με την μέθοδο ANOVA τροφοδότησαν για εκπαίδευση τον αλγόριθμο Extreme Gradient Boosting (XGBoost) και παρήχθη ένα εκπαιδευμένο μοντέλο το οποίο αξιολογήθηκε με την μέθοδο 5-fold cross-validation.



Εικόνα 45: Αρχιτεκτονική μοντέλου πλήρως εποπτευόμενης μάθησης

(Hasani & Nasiri, 2021)

Η αρχιτεκτονική του προτεινόμενου μοντέλου φαίνεται στο παραπάνω σχήμα.

Το 5-fold cross validation είχε μέση ακρίβεια 98,72%. Η αρχιτεκτονική αυτή εντόπισε σωστά τα COVID-19 και No-Findings με ακρίβεια 100% και 98,43%, αντίστοιχα. Η precision, recall, and specificity ήταν 99.21%, 93.33%, and 100%, αντίστοιχα.»

(Υποκεφάλαιο 8.2) Προσέγγιση προβλήματος

(Ενότητα 8.2.α) Επιλογή μεθοδολογίας

Η ακρίβεια της μεθόδου που χρησιμοποίησαν οι (Hasani & Nasiri, 2021) είναι εντυπωσιακή, 100% στην ανίχνευση των θετικών. Για τον λόγο αυτό δεν κρίναμε ότι υπάρχει λόγος να αλλάξουμε την αρχιτεκτονική μέχρι την και δημιουργία του τελικού πίνακα χαρακτηριστικών.

Χρησιμοποιήθηκε η ίδια μεθοδολογία και τα ίδια αρχεία ακτινογραφιών, ώστε να έχουμε την ίδια βάση εκκίνησης. Δηλαδή χρησιμοποιήθηκε η τεχνική της μάθησης με μεταφορά από ένα προεκπαιδευμένο δίκτυο συγκεκριμένα το DenseNet169 (βλ. Ενότητα που αφορά την μάθηση με μεταφορά).Οι (Hasani & Nasiri, 2021) δηλώνουν ότι δοκιμάσανε διάφορα μοντέλα και ότι αυτό είχε τα καλυτέρα αποτελέσματα στην πλήρως εποπτευόμενη μάθηση με βάση τον XGBoost που ακολούθησε.

Το DenseNet169 έχει προεκπαιδευτεί στο ImageNet. Το ImageNet είναι μια μεγάλη οπτική βάση δεδομένων που έχει σχεδιαστεί για χρήση στην έ-ρευνα λογισμικού αναγνώρισης οπτικών αντικειμένων. Περισσότερες από 14 εκατομμύρια εικόνες έχουν επισημειωθεί με το χέρι για να υποδείξουν ποια αντικείμενα απεικονίζονται. Το InageNet περιέχει περισσότερες από 20.000 κατηγορίες οι οποίες επισημειώθηκαν χειροκίνητα.

Χρησιμοποιούμε τις πληροφορίες (προεκπαιδευμένα βάρη) που έμαθε από την προεκπαίδευση το DenseNet169 για να εξάγουμε χαρακτηριστικά από νέα δείγματα που αποκτήσαμε δηλαδή από τις ακτινογραφίες που έχουμε.

(Transfer learning and fine-tuning, 2022-06-08)

Τέλος θα εξετάσουμε πολλούς αλγόριθμους μάθησης με μερικώς επισημειωμένα δεδομένα (self-trainıng, label-propagation,label spreading), την ενεργή μάθηση (active learning) με πολλές διαφορετικές στρατηγικές ερωτήσεων και τους γνωστούς αλγόριθμους μάθησης από θετικά και μη επισημειωμένα δεόμενα (PUlearning), δηλ. τους Elkanoto, Weighted Elkanoto και Bagging, οι οποίοι θα εκπαιδευτούν με τα χαρακτηριστικά που εξάχθηκαν με βάση το προεκπαιδευμένο μοντέλο.

Οι αλγόριθμοι αυτοί, πλην του Label-Propagation και Label-Spreading, οι οποίοι χρησιμοποιούν κάποιον πύρινα (*'knn' ή 'rbf'*) , χρησιμοποιούν έναν αλγόριθμο πλήρους εποπτευόμενης μάθησης ως αλγόριθμο βάσης, δηλαδή ως εκτιμητή που τους βοηθά στην ταξινόμηση και τις εκτιμήσεις που κάνουν.

Στην παρούσα μελέτη χρησιμοποιούμε διάφορους αλγόριθμους εποπτευόμενης μάθησης ως αλγόριθμους βάσης για να κάνουμε τα πειράματά μας με επιλογή να έχουμε ει δυνατόν ένα από κάθε μια κατηγορία (γραμμικό, σιγμοειδές, δεντρικό, συνόλου, ενισχυτικό και νευρωνικό δίκτυο).

Ακολουθήθηκε η παρακάτω γενική ροή εργασιών:

1. Εξέταση και κατανόηση των δεδομένων,
2. Προ-επεξεργασία των δεδομένων,
3. Φόρτωση του προεκπαιδευμένου μοντέλου (προεκπαιδευμένα βάρη),
4. Χρήση του προεκπαιδευμένο μοντέλου για εξαγωγή χαρακτηριστικών από τις εικόνες μας.
5. Εκπαιδεύουμε ένα μοντέλο με τους αλγορίθμους για μερικώς επισημειωμένα δεδομένα. Οι αλγόριθμοι αυτοί χρησιμοποιούν ως αλγορίθμους βάσης (εκτιμητές) τους αλγορίθμους πλήρως εποπτευόμενης μάθησης διαδοχικά.
6. Αξιολογούμε τα μοντέλα και τους εκτιμητές.

(Ενότητα 8.2.β) Επιλογή αρχείων-εξέταση και κατανόηση δεδομένων

Χρησιμοποιούμε τις ίδιες ακτινογραφίες που χρησιμοποίησαν οι Nassiri και Allani. Αυτές αποτελούν ένα σύνολο έγχρωμων RGB ακτινογραφιών που έχουν επισημειωθεί ως «covid-19», «no-findings¨» και «pneumonia». Οι ακτινογραφίες που έχουν επισημειωθεί ως πνευμονία δεν έχουν χρησιμοποιηθεί.

Τα δεδομένα μας αφορούν έγχρωμες ακτινογραφίες τύπου RGB διαφόρων μεγεθών.

(Ενότητα 8.2.γ) Προεπεξεργασία Δεδομένων

Φορτώνουμε τις εικόνες και τις επισημειώσεις τους στις κατάλληλες δομές πινάκων. Δηλαδή προβαίνουμε στις παρακάτω ενέργειες:

- Ενσωματώνουμε τις απαραίτητες βιβλιοθήκες
- Διαβάζουμε τα αρχεία των εικόνων που έχουμε μεταφέρει στον υπολογιστή μας και τα μεταφέρουμε σε ανάλογο πίνακα
- Τις μετατρέπουμε σε μέγεθος 224 X 224 (μέγεθος εικόνων στο οποίο έχει προεκπαιδευτεί το DenseNet 169).
- Τις μετατρέπουμε από χρωματικό χώρο RGB σε BGR σύμφωνα επίσης με την προεκπαίδευση.

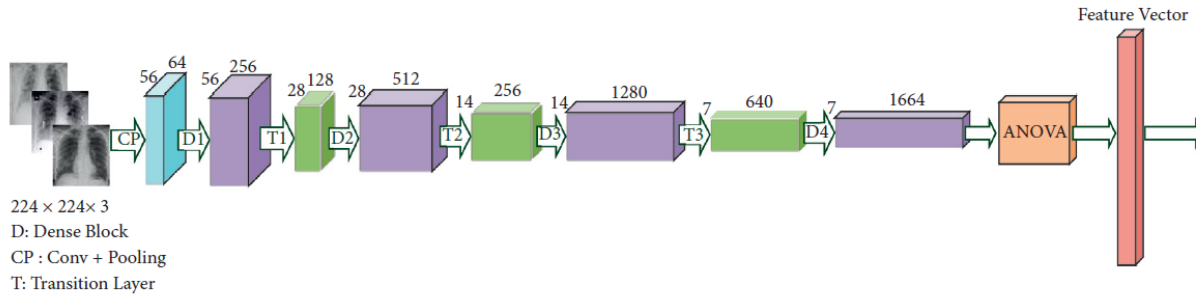
- Κρατάμε τις πραγματικές τιμές των επισημειώσεων ώστε να μπορούμε να κάνουμε την αξιολόγηση του μοντέλου.
- Μετατρέπουμε όλους τους πίνακες και λίστες σε πίνακες τύπου numpy.
- Κανονικοποιούμε δηλαδή αντικαθιστούμε όλες τις τιμές των εικονοστοιχείων, με τιμές από 0-1. Αυτό στην περίπτωση των εικόνων που η τιμή είναι για κάθε επίπεδο από 0-255, επιτυγχάνεται με διαίρεση με το 255 (images = images/255.0). Επίσης το κάθε μοντέλο, δέχεται συγκεκριμένα input εν προκείμενο το νευρωνικό δίκτυο τύπου densenet169, δέχεται τιμές; από 0-1.

Ο αριθμός των εικόνων που χρησιμοποιήθηκαν:

```
Σύνολο εικόνων:  625
Πραγματικά θετικές:  125
Πραγματικά αρνητικές:  500
```

Το τελικό επίπεδο του δικτύου DenseNet169, το οποίο χρησιμοποιήθηκε για την πρόβλεψη δεδομένων ImageNet, καταργείται. Ένα «global avg pooling» , προστέθηκε στο τελικό επίπεδο του δικτύου. Ένα από τα οφέλη «global avg pooling» είναι ότι δεν υπάρχουν παράμετροι για προσαρμογή σε αυτό το επίπεδο. επομένως, δεν απαιτείται εκπαίδευση..

```
# Δημιουργία μοντέλου για εξαγωγή χαρακτηριστικών
```



Εικόνα 46 Αρχιτεκτονική εξαγωγής χαρακτηριστικών

(Hasani & Nasiri, 2021)

```
from keras.applications import DenseNet169
DenseNet169 = DenseNet169(input_shape = SIZE +  [3], weights='imagenet',
              include_top=False, pooling='avg')

#input_shape = SIZE +  [3]:Το SIZE ήδη έχει δηλωθεί ως 224 X 224,
#3 = διαστάσεις – χρώματα.
#weights='imagenet': προ-εκπαιδευμένα βάρη από το imagenet
#include_top=False: Αφαίρεση του τελευταίου επιπέδου
```

```
    for layer in DenseNet169.layers: #παγωμα επιπεδων
        layer.trainable = False
```

 # Μορφή δικτύου.

```
    DenseNet169.summary()
```

Model: "densenet169"

```
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
=
 input_2 (InputLayer)                [(None, 224, 224, 3  0           []    )]
zero_padding2d_2 (ZeroPadding2       (None, 230, 230, 3)  0           ['input_2[0][0']] D)
 conv1/conv (Conv2D)                 (None, 112, 112, 64  9408        ['zero_padding2d_2[0][0']] )
 conv1/bn (BatchNormalization)       (None, 112, 112, 64  256         ['conv1/conv[0][0']]  )
 conv1/relu (Activation)             (None, 112, 112, 64  0           ['conv1/bn[0][0']]   )
 zero_padding2d_3 (ZeroPadding2)     (None, 114, 114, 64  0           ['conv1/relu[0][0']]  D)   )
 pool1 (MaxPooling2D)                (None, 56, 56, 64)   0           ['zero_padding2d_3[0][0']]
 conv2_block1_0_bn (BatchNormalization)  (None, 56, 56, 64)  256      ['pool1[0][0']]
 conv2_block1_0_relu (Activation)    (None, 56, 56, 64)   0           ['conv2_block1_0_bn[0][0']]
 conv2_block1_1_conv (Conv2D)        (None, 56, 56, 128)  8192        ['conv2_block1_0_relu[0][0']]
 conv2_block1_1_bn (BatchNormalization)  (None, 56, 56, 128)  512    ['conv2_block1_1_conv[0][0']]
 conv2_block1_1_relu (Activation)    (None, 56, 56, 128)  0           ['conv2_block1_1_bn[0][0']]
 conv2_block1_2_conv (Conv2D)        (None, 56, 56, 32)   36864       ['conv2_block1_1_relu[0][0']]
 conv2_block1_concat (Concatenation  (None, 56, 56, 96)   0           ['pool1[0][0']
 'conv2_block1_2_conv[0][0']]
```

……

```
 conv5_block31_2_conv (Conv2D)  (None, 7, 7, 32)    36864                    ['conv5_block31_1_relu[0][0']]
conv5_block31_concat (Concatenate)          (None, 7, 7, 1632)  0           ['conv5_block30_concat[0][0']',
                                                                             conv5_block31_2_conv[0][0']]
 conv5_block32_0_bn (BatchNormalization     (None, 7, 7, 1632)  6528        ['conv5_block31_concat[0][0']]
 conv5_block32_0_relu (Activatiion          (None, 7, 7, 1632)  0           ['conv5_block32_0_bn[0][0']]
 conv5_block32_1_conv (Conv2D)              (None, 7, 7, 128)   208896      ['conv5_block32_0_relu[0][0']]
 conv5_block32_1_bn (BatchNormalization)    (None, 7, 7, 128)   512         ['conv5_block32_1_conv[0][0']]
 conv5_block32_1_relu (Activatiion )         (None, 7, 7, 128)   0
['conv5_block32_1_bn[0][0']]
 conv5_block32_2_conv (Conv2D)              (None, 7, 7, 32)    36864       ['conv5_block32_1_relu[0][0']]
 conv5_block32_concat (Concatenate)          (None, 7, 7, 1664)  0           ['conv5_block31_concat[0][0']',
                                                                              'conv5_block32_2_conv[0][0']]
 bn (BatchNormalization)                    (None, 7, 7, 1664)  6656        ['conv5_block32_concat[0][0']]
 relu (Activation)                          (None, 7, 7, 1664)  0           ['bn[0][0']]
 avg_pool (GlobalAveragePooling 2D)             (None, 1664)       0       ['relu[0][0']]

 ==================================================================================================
=
 Total params: 12,642,880
 Trainable params: 0
 Non-trainable params: 12,642,880
 _________________________________________________________________________________________________
```

Ακολούθως εξάγουμε τα χαρακτηριστικά

 # Εξαγωγή χαρακτηριστικών (Feature extraction)

```
    features = DenseNet169.predict(images)
    print(features.shape
    print(features.shape)
    (625, 1664)
```

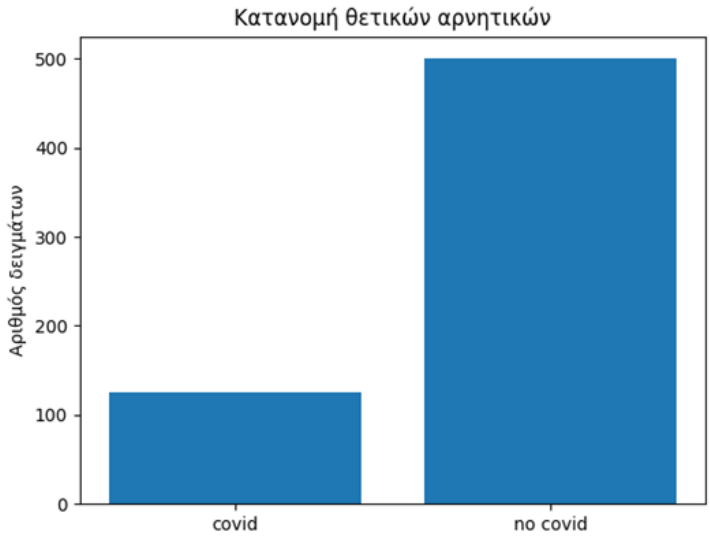
Δηλαδή από 625 εικόνες έχουμε εξάγει 1664 χαρακτηριστικά.

 # Επιλογή χαρακτηριστικών (Feature selection)

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
features =
 SelectKBest(score_func=f_classif, k=20).fit_transform(features,Actual_labels)
print (features.shape)
(20, 1664)
```

Όταν σε ένα μοντέλο μάθησης δίνονται πολλά χαρακτηριστικά και λίγα δείγματα, είναι πιθανό να υπερπροσαρμόζεται (overfitting), προκαλώντας υποβάθμιση της απόδοσής του. Προκειμένου να μειωθεί ο χρόνος ταξινόμησης να αυξηθεί η απόδοση του ταξινομητή, και να αποφευχθεί η υπερπροσαρμογή χρησιμοποιήθηκε η μέθοδος επιλογής χαρακτηριστικών ANOVA για τη μείωση του αριθμού των χαρακτηριστικών σε 20. Αυτός ο αριθμός βρέθηκε με δοκιμές. Τελικά έχουμε πίνακα 625 X 20 για τα χαρακτηριστικά και μονοδιάστατο 625 θέσεων για τις επισημειώσεις.

 # Κατασκευή Διαγράμματος Κατανομής Θετικών - Αρνητικών



Εικόνα 47 Κατανομή θετικών - αρνητικών

Υπάρχουν περισσότερες από διπλάσιες περιπτώσεις της κλάσης «no covid» από τις περιπτώσεις της covid . Με μια μη ισορροπημένη κατάσταση κλάσεων όπως αυτή, η accuracy όπως ήδη έχει αναφερθεί δεν είναι η καλύτερη επιλογή. **Το f1-score** μας δίνει μια ισορροπημένη ένδειξη και είναι η καλύτερη επιλογή για να μετρήσουμε την αξιοπιστία ενός μοντέλου σε περιπτώσεις ανισορροπίας κλάσεων.

(Υποκεφάλαιο 8.3) Εφαρμογή αλγορίθμων μάθησης με μερικώς επισημειωμένα δεδομένα.

(Ενότητα 8.3.α) Γενικά

Η μάθηση με μερικώς επισημειωμένα δεδομένα είναι μια κατάσταση κατά την οποία τα περισσότερα από τα δεδομένα εκπαίδευσης δεν φέρουν επισημείωση. Οι αλγόριθμοι που εφαρμόζουν μάθηση με μερικώς επισημειωμένα δεδομένα χρησιμοποιήσουν αυτά τα πρόσθετα δεδομένα χωρίς επισημείωση για να αποτυπώσουν καλύτερα το σχήμα της υποκείμενης κατανομής δεδομένων και να επιχειρήσουν να γενικεύσουν με επισημειώσεις που κατ' αρχάς φαίνεται ότι είναι ανεπαρκείς.

Οι αλγόριθμοι αυτοί χωρίζονται σε τρεις γενικές κατηγορίες ημι-εποπτευόμενη μάθηση που περιλαμβάνει την αυτοεκπαίδευση (self-learning) και την βασισμένη σε γράφους (label-propagation και label-spreading), την ενεργή μάθηση (active learning) και την μάθηση από θετικά και μη επισημειωμένα δεδομένα (Elkanoto, Weighted Elkanoto, Bagging). Όλοι τους χρησιμοποιούν έναν αλγόριθμο βάσης (εκτιμητή) ο οποίος είναι ένας αλγόριθμος πλήρους εποπτευόμενης μάθησης. Στην περίπτωση των label propagation και label spreading αντί αλγορίθμου βάσης χρησιμοποιούνται οι πυρήνες «rbf» και «knn»

Οι self-training, label-propagation, label-spreading και οι Elkanoto, Weighted Elkanoto, αι Bagging ανήκουν στην ημι-εποπτευόμενη μάθηση δηλαδή τα μη επισημειωμένα δεδομένα τα επισημειώνουν αυτόματα (ψευδοεπισημειώνουν).

Για να συλλέξουμε τα στοιχεία που επιθυμούμε, σχεδιάσαμε και υλοποιήσαμε τα παρακάτω πειράματα. Υπενθυμίζεται ότι μετρά την εξαγωγή και την μείωση των χαρακτηριστικών, έχουμε μια βάση δεδομένων με πραγματικά επισημειωμένα δεδομένα.

Πείραμα Semi Supervised και PU learning

Ανεξάρτητες μεταβλητές:

1. Αλγόριθμος μερικώς επισημειωμένων δεδομένων,
2. Αλγόριθμοι βάσης (Logistic Regression, SVC, Multi-Layer Perceptron, Decision Tree, Random Forest, και LGMBMC) Στην περίπτωση των

Label-Propagation και Label-Spreading χρησιμοποιήθηκαν οι πυρήνες «rb»f και «knn»),

3. Αριθμός πραγματικά επισημειωμένων.

Εξαρτημένη μεταβλητή: Η τιμή f1-score του αλγορίθμου μερικώς επισημειωμένων.

Διεξαγωγή πειράματος:

1. Θεωρούμε αρχικά ως πραγματικά επισημειωμένα δεδομένα 10 τυχαία δείγματα από την βάση δεδομένων. Στα υπόλοιπα δείγματα τοποθετούμε ως επισημείωση το -1. Δηλαδή τα θεωρούμε ως μη επισημειωμένα.

2. Διεξάγουμε την εκπαίδευση με τα επισημειωμένα και μη επισημειωμένα δεδομένα που έχουν τιμές 0,1 και -1 (0=αρνητικά (δεν υπάρχουν στο PU), 1 θετικά, -1 μη επισημειωμένα).

3. Εκτελούμε πρόβλεψη με το εκπαιδευμένο μοντέλο για το σύνολο των δεδομένων μας με επισημειώσεις τις 0,1, -1. Οι προβλέψεις θα περιέχουν μόνο 0 και 1 (θετικά ή αρνητικά). Συγκρίνουμε τις προβλέψεις με τις πραγματικές επισημειώσεις.

4. Καταγράφουμε τις τιμές της f1-score

5. Αυξάνουμε τα θεωρούμενα επισημειωμένα δεδομένα κατά 10 επιλέγοντας πάλι τυχαία από την βάση δεδομένων και πάμε στο βήμα 2.

6. Ο κύκλος αυτός τερματίζεται όταν το f1-score γίνει μεγαλύτερο του 90%.

Η παραπάνω διαδικασία επαναλαμβάνεται για κάθε αλγόριθμο βάσης.

Πείραμα Active learning

Ανεξάρτητες μεταβλητές:

1. Αλγόριθμος active learning.

2. Αλγόριθμοι βάσης (Logistic Regression, Random Forest, LGMBMC και **committee**).

3. Στρατηγικές ερωτήσεων (entropy sampling, margin sampling, uncertainty sampling)

4. Αριθμός επισημειωθέντων.

Εξαρτημένη μεταβλητή: Η τιμή f1-score του αλγορίθμου.

Διεξαγωγή πειράματος:

1. Θεωρούμε ως επισημειωμένα δεδομένα τυχαία 10 δείγματα τα οποία αφαιρούμε από την βάση δεδομένων των επισημειωμένων δειγμάτων. Τα υπόλοιπα τα κρατάμε στην βάση δεδομένων μας.
2. Διεξάγουμε την εκπαίδευση με τα δεδομένα που είναι επισημειωμένα.
3. Εκτελούμε πρόβλεψη με το εκπαιδευμένο μοντέλο με το σύνολο των δεδομένων.
4. Συγκρίνουμε τις προβλέψεις με τις πραγματικές επισημειώσεις.
5. Καταγράφουμε τις τιμές της f1-score
6. Αυξάνουμε τα επισημειωμένα δεδομένα κατά 10 αφαιρώντας τα από την βάση δεδομένων με τα επισημειωμένα δεδομένα. Η επιλογή από την βάση γίνεται σύμφωνα με την στρατηγική ερωτήσεων που εφαρμόζουμε κάθε φορά. Πάμε στο βήμα 2.
7. Ο κύκλος αυτός τερματίζεται όταν το f1-score γίνει μεγαλύτερο του 90%.

Η παραπάνω διαδικασία επαναλαμβάνεται για κάθε αλγόριθμο βάσης, για την επιτροπή των αλγορίθμων βάσης καθώς και για κάθε στρατηγική ερωτήσεων.

Σκοπός είναι να βρεθεί ο αλγόριθμος που επιτυγχάνει μια υψηλή απόδοση f1-score με τα λιγότερα επισημειωμένα δεδομένα.

Τα αποτελέσματα τα οπτικοποιούμε στις παρακάτω μορφές:

1. Confusion matrix
2. Διαγράμματα εξέλιξης τιμών
3. Ιστογράμματα απόδοσης f1-score
4. Αναλυτικοί πίνακες αποτελεσμάτων διασταυρούμενης επικύρωσης.

Εκτός από τις τιμές των μετρικών εκπαίδευσης και δοκιμών, τις οποίες υπολογίζουμε στην διασταυρούμενη επικύρωση, υπολογίζουμε και την διαφορά μεταξύ μετρικών εκπαιδευτικών δεδομένων και δοκιμών, καθώς και την τυπική απόκλιση των τιμών των μετρικών. Έτσι μπορούμε να βγάλουμε συμπεράσματα για υποπροσαρμογή ή υπερπροσαρμογή. Εάν οι μετρήσεις των εκπαιδευτικών δεδομένων είναι καλές αλλά των δοκιμών όχι, τότε σημαίνει ότι έχουμε υπερπροσαρμογή, την οποία πρέπει να αντιμετωπίσουμε. Εάν η

τυπική απόκλιση είναι μεγάλη, τότε σημαίνει ότι έχουμε ένδειξη για υποπροσαρμογή.

Για την υλοποίηση των δοκιμών έχε επιλεγεί η βιβλιοθήκη sklearn (scikit-learn.org, n.d.). η βιβλιοθήκη ModAL (Danka, A modular active learning framework for Python3, 2018) και η βιβλιοθήκη pulearn (Elkan, Noto, Drouin, AditraAS, & Wright., n.d.).

(Ενότητα 8.3.β) Αυτοεκπαίδευση

Η κλάση SelfTrainingClassifier

Η αυτοεκπαίδευση (self training) του sklearn βασίζεται στον αλγόριθμο του Yarowsky (Yarowsky, 1995). Χρησιμοποιώντας αυτόν τον αλγόριθμο, ένας αλγόριθμος πλήρως εποπτευόμενης μάθησης (αλγόριθμος βάσης) μπορεί να λειτουργήσει ως εκτιμητής για να βοηθήσει στην εκτίμηση της επισημείωσης των μη επισημειωμένων δεδομένων.

lass `sklearn.semi_supervised.`**`SelfTrainingClassifier`**(*base_estimator, threshold=0.75, criterion='threshold', k_best=10, max_iter=10, verbose=False*)

Ο SelfTrainingClassifier μπορεί να χρησιμοποιήσει ως εκτιμητή οποιονδήποτε ταξινομητή υλοποιεί την συνάρτηση predict_proba,την οποία εμπεριέχουν όλοι οι αλγόριθμοι εποπτευόμενης μάθησης του Sklearn. Η συνάρτηση αυτή προβλέπει την πιθανότητα ένα μη επισημειωμένο δείγμα, να πάρει τιμή 0 ή 1.

Ο αλγόριθμος λειτουργεί επαναληπτικά. Σε κάθε επανάληψη, ο βασικός ταξινομητής προβλέπει την κατηγορία στην οποία ανήκουν τα δείγματα που δεν είναι επισημειωμένα και προσθέτει ένα υποσύνολο αυτών στο επισημειωμένο σύνολο δεδομένων. Η επιλογή αυτού του υποσυνόλου καθορίζεται από το κριτήριο επιλογής. Εάν το κριτήριο επιλογής είναι k_best=k, τότε σε κάθε επανάληψη του αλγορίθμου, επιλέγονται τα κ «καλύτερα», δηλαδή τα κ που έχουν την καλύτερη πιθανότητα να ανήκουν στην μια ή την άλλη κατηγορία. Η επιλογή μπορεί επίσης να γίνει χρησιμοποιώντας ένα όριο (κατώφλι-threshold) στις πιθανότητες να ανήκει σε κάποια κλάση το υποψήφιο δείγμα προς ψευδοεπισημείωση. Δηλαδή επιλέγει ένα όριο πιθανότητας πχ 0,90, κάτω από το οποίο δεν θα ψευδοεπισημειώνονται δεδομένα. Εδώ θέλει προσοχή, εάν βάλουμε χαμηλό κατώφλι πχ κάτω από 0,5 για κάθε κλάση, ο ταξι-

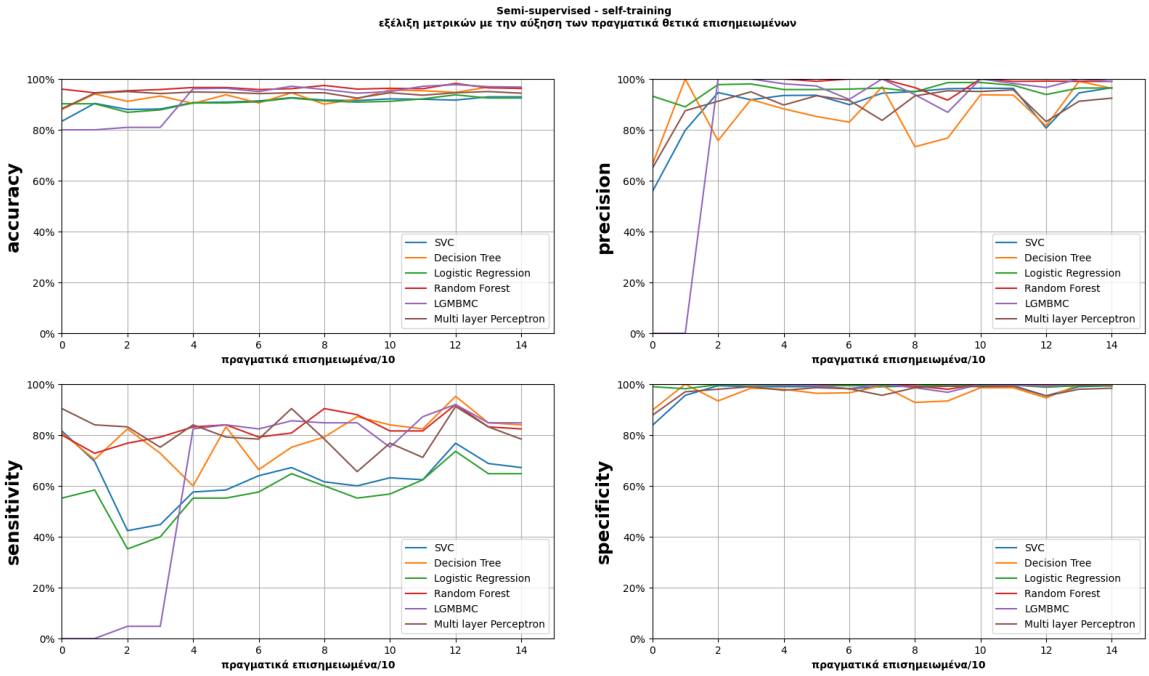
νομητής μαθαίνει από δείγματα που επισημειώθηκαν με χαμηλή εμπιστοσύνη. Αυτά τα δείγματα χαμηλής εμπιστοσύνης είναι πιθανό να δώσουν εσφαλμένες προβλέψεις, δηλαδή να έχουμε υποπροσαρμογή. Στην περίπτωση αυτή ο ταξινομητής επισημειώνει όλα τα δείγματα με ελάχιστες επαναλήψεις. Για πολύ υψηλά κατώφλια ο ταξινομητής μετά από κάποιο σημείο δεν επαυξάνει το σύνολο δεδομένων του (έχει σταματήσει να ψευδοεπισημειώνει).

Οι συνθήκες τερματισμού είναι μια από τις παρακάτω:

- Φθάσαμε τον μεγαλύτερο αριθμό επαναλήψεων που έχουμε θέσει (max_iter)
- Ο αλγόριθμος δε παράγει καινούργιες ψευδοεπισημειώσεις,
- Όλα τα μη επισημειωμένα δείγματα έχουν επισημειωθεί πριν φθάσουμε το max_iter.

Από εδώ και πέρα ακολουθούμε αυτά που περιεγράφηκαν στη 8.3.α.

Αποτελέσματα self-training



Εικόνα 48: Semi-supervised, self-training - εξέλιξη μετρικών με την αύξηση των πραγματικά επισημειωμένων

Από τον παραπάνω διάγραμμα βγάζουμε τα εξής συμπεράσματα:

Σαν γενική παρατήρηση οι γραμμές όλων των μετρικών από ένα σημείο και πέρα (40 πραγματικά επισημειωμένα) είναι σχεδόν παράλληλες προς τον

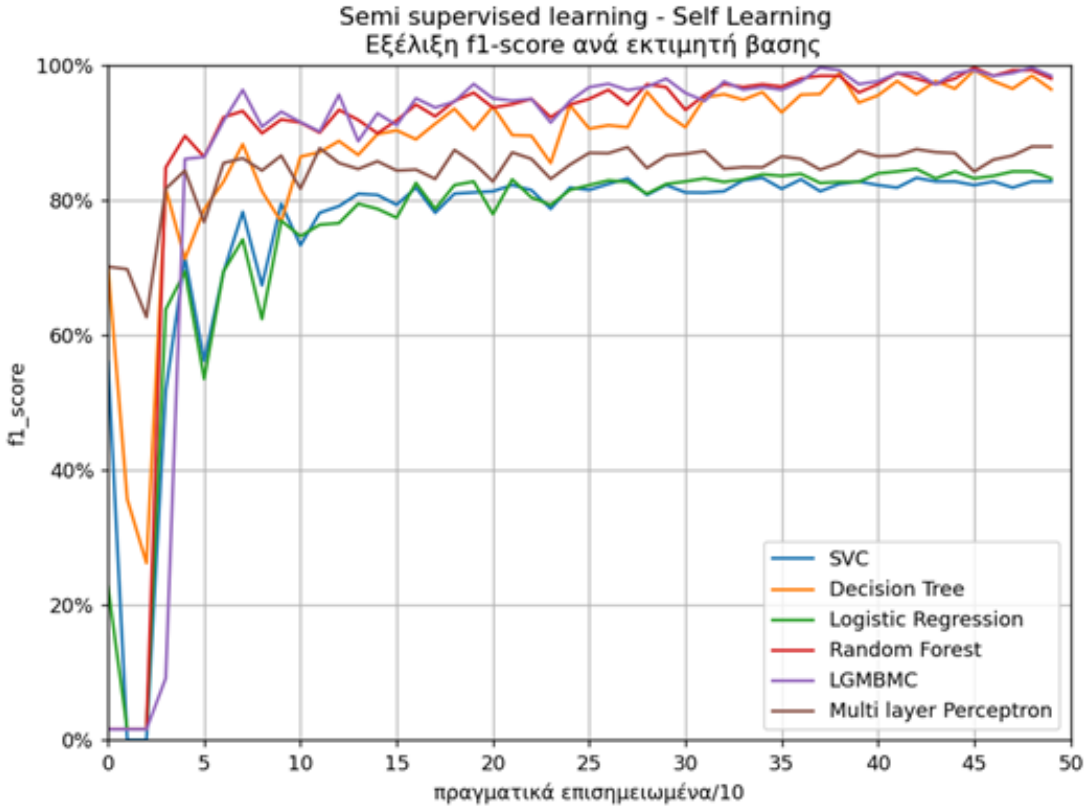
άξονα των χ, που σημαίνει ότι δεν επηρεάζονται ιδιαίτερα από την αύξηση των πραγματικά επισημειωμένων.

-**Η accuracy** είναι σταθερά πάνω από 80%, για όλους τους αλγορίθμους βάσης. Ενώ όλοι δίνουν τιμές accuracy πάνω από 90% όταν έχουμε πάνω από 40 επισημειωμένα. Λόγω του ότι τα δεδομένα μας είναι μη ισορροπημένα (έχουμε πολύ περισσότερα αρνητικά από θετικά), η accuracy δεν είναι η πλέον κατάλληλη μετρική, για την μέτρηση της αποτελεσματικότητας του μοντέλου.

-**Η precision** απαντά στο ερώτημα «ποιο ποσοστό των θετικών ταυτοποιήσεων ήταν πραγματικά σωστό;». Μια βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινό θετικό (TP). Όμως το 100% μπορεί να οφείλεται και στον τρόπο που υπολογίζεται το precision. Όταν ο παρονομαστής (FP+TP) είναι 0, τότε το αποτέλεσμα συμβατικά είναι 1, παρότι δεν είχαμε ούτε ένα (1) True Positive. Το ίδιο συμβαίνει και όταν μόνο του το FP είναι ίσο με 0. Το Decision Tree έχει ορισμένες διακυμάνσεις, οι υπόλοιποι αλγόριθμοι βάσης έχουν τιμές σταθερά πάνω από 80%.

-**Η sensitivity**, η οποία είναι η ικανότητα του τεστ να εντοπίζει σωστά άτομα τα οποία νοσούν (πραγματικό θετικό ποσοστό). Μία βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά θετικό. Η εξέλιξη της sensitivity είναι σημαντική καθόσον το να αναγνωρίσουμε με ακρίβεια τους θετικούς μπορεί να συντελέσει στον περιορισμό της εξάπλωσης της νόσου. Οι Multi-Layer Perceptron, Random Forest και Decision Tree υπερέχουν έναντι των Logistic Regression, LGMBMC και SVC.

-**Η specificity**, δηλαδή η ικανότητα να ανακαλύπτει τους πραγματικά αρνητικούς, είναι σε υψηλά επίπεδα από μικρό αριθμό θετικά επισημειωμένων. Το γεγονός αυτό εν μέρει δικαιολογείται από τον μεγάλο αριθμό των υπαρχόντων αρνητικών. Μία τιμή 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά αρνητικό.

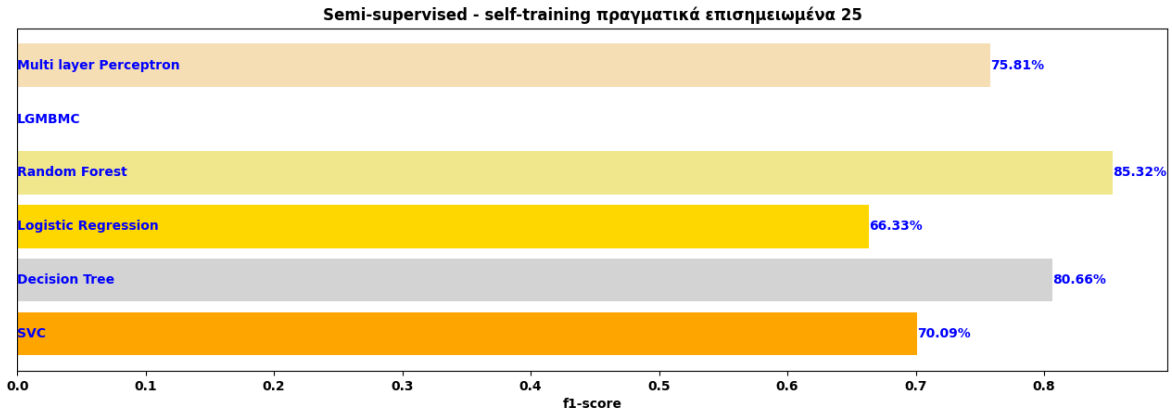


Εικόνα 49: Self-training – εξέλιξη f1-score με την αύξηση των πραγματικά επισημειωμένων δεδομένων.

Στο παραπάνω διάγραμμα φαίνεται η εξέλιξη, της μετρικής f1-score με την αύξηση των πραγματικά επισημειωμένων δειγμάτων. Το συμπέρασμα είναι ότι όλοι εκτιμητές δημιουργούν σταθερά υψηλή απόδοση όταν υπάρχουν πάνω από 200 πραγματικά επισημειωμένα δείγματα. Μερικοί όμως από αυτούς, και συγκεκριμένα οι Random Forest,, Decision Tree, LGMBMC δημιουργούν αποδώσεις πάνω από 90% όταν έχουν στην διάθεσή τους πάνω από 50 επισημειωμένα. Οι υπόλοιποι παραμένουν στα επίπεδα του 80%.

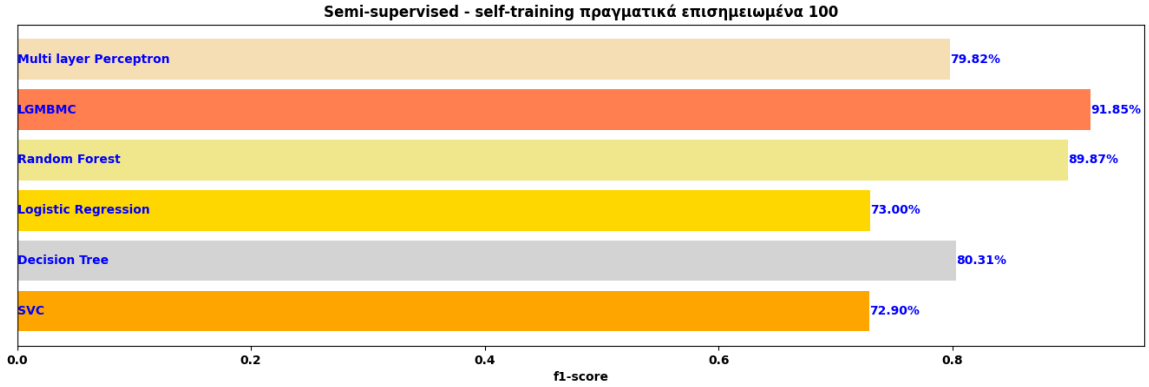
Ένα άλλο συμπέρασμα που βγαίνει είναι ότι οι εκτιμητές βάσης είναι περίπου ισοδύναμοι εάν θεωρήσουμε το 80% ως το αποδεκτό όριο για το f1_score, αλλιώς εάν θεωρηθεί το 90% ως όριο τότε μόνο οι Decision Tree, Random Forest και LGMBMC μπορούν να το πέτυχουν. Ένα άλλο συμπέρασμα είναι ότι δεν έχει νόημα να έχουμε πάνω από 100 πραγματικά επισημειωμένα δείγματα γιατί δεν θα βελτιωθεί σημαντικά η f1-score.

Σημεία προς περαιτέρω μελέτη τα 50 (εκεί που φαίνεται ότι ορισμένοι αλγόριθμοι περνούν το 80%, όσον αφορά στην τιμή της f1-score), το 100 (εκεί που ορισμένοι αλγόριθμοι περνούν το 90% στο f1-score, και 200 όπου φαίνεται ότι από εκεί και πέρα επιβραδύνεται σοβαρά η βελτίωση.

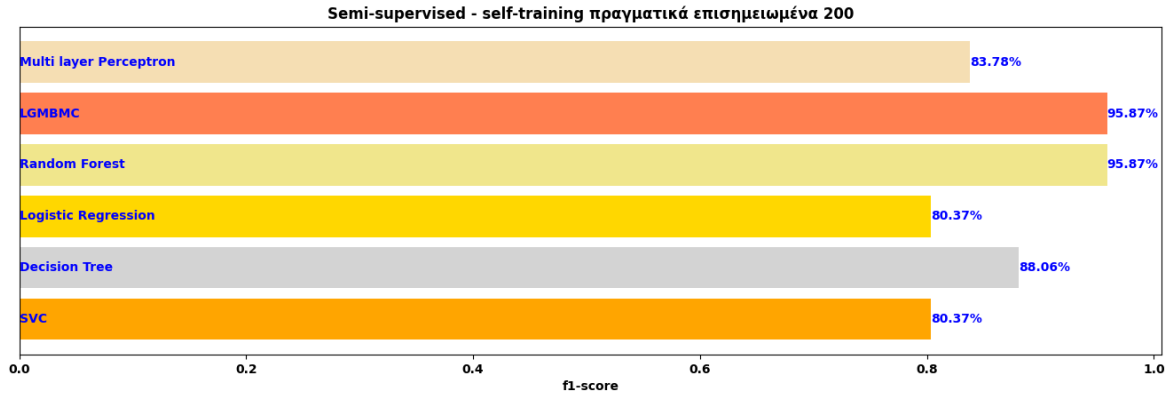


Εικόνα 50: Self-training - διάγραμμα απόδοσης των εκτιμητών βάσης για 25 πραγματικά επισημειωμένα.

Στο επίπεδο των 25 επισημειωμένων έχουμε τους Random Forest, Decision Tree και Multi-Layer Perceptron που έχουν f1-score πάνω από 80%.



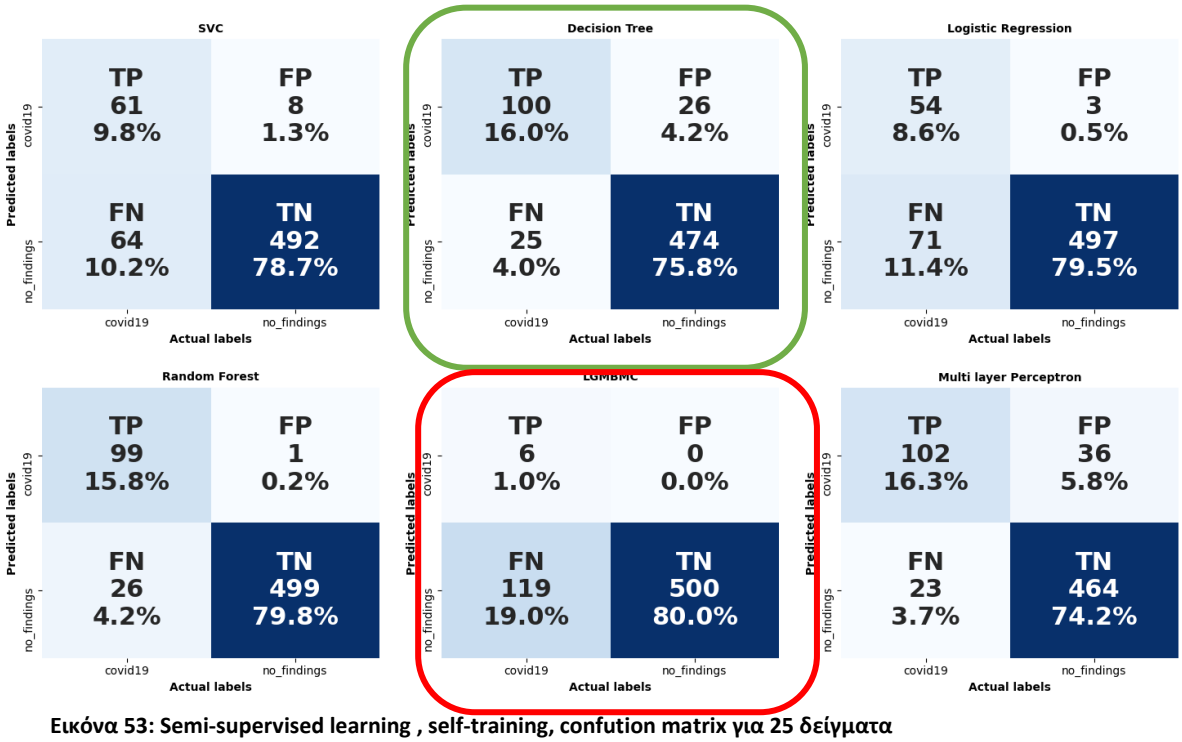
Εικόνα 51: Self-training - διάγραμμα απόδοσης των εκτιμητών βάσης για 100 πραγματικά επισημειωμένα.



Εικόνα 52: Self-training - διάγραμμα απόδοσης των εκτιμητών βάσης για 200 πραγματικά επισημειωμένα.

Παρατηρούμε πράγματι μετά τα 100 επισημειωμένα δεν υπάρχει σημαντική πρόοδος.

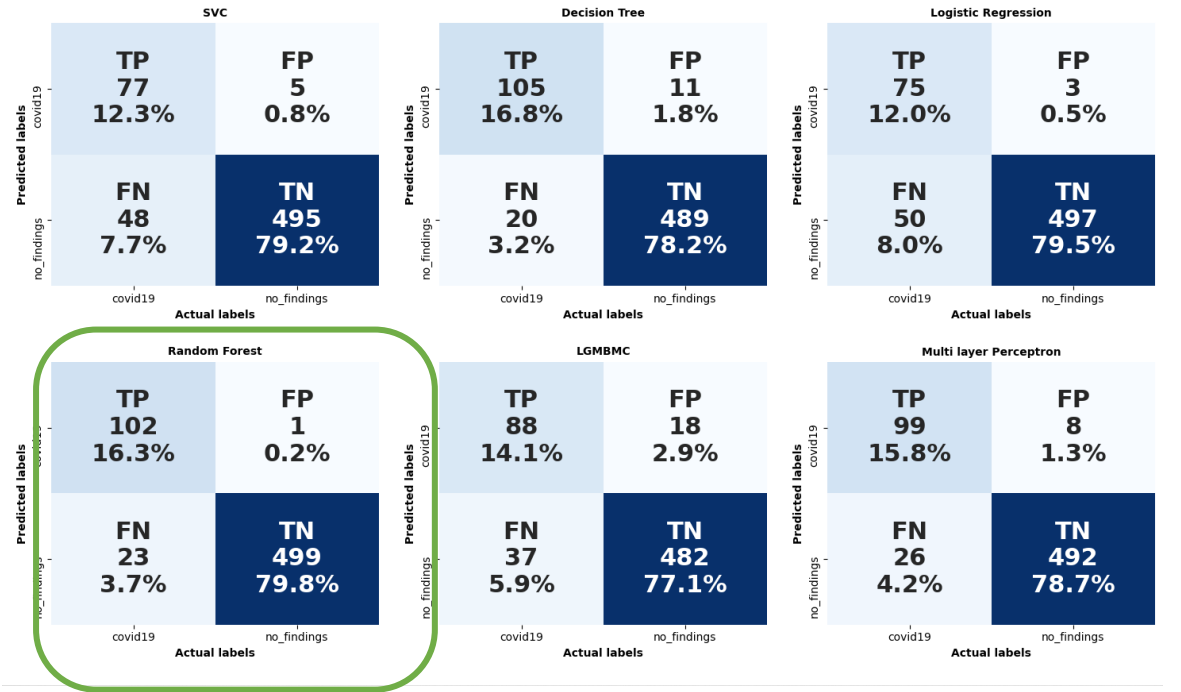
Self training
confusion matrix με 25 στά 625(4.0%) πραγματικά επισημειωμένα δείγματα



Εικόνα 53: Semi-supervised learning , self-training, confution matrix για 25 δείγματα

Στα 25 επισημειωμένα, ο LGMBMC αναγνωρίζει όλα τα αρνητικά, αλλά πολύ λίγα θετικά, ο Decision αναγνωρίζει 474 από τα 500 αρνητικά αλλά και 100 από τα 125 θετικά.

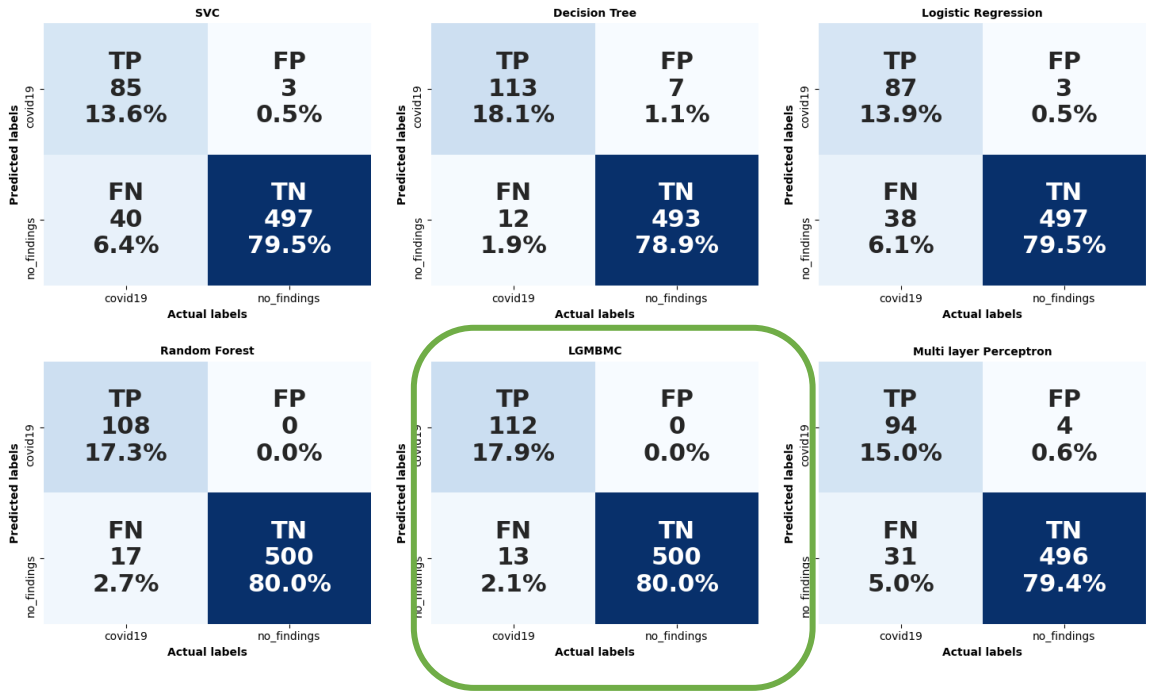
Self training
confusion matrix με 100 στά 625(16.0%) πραγματικά επισημειωμένα δείγματα



Εικόνα 54: Semi-supervised learning , self-training, confution matrix για 100 δείγματα.

Στα 100 επισημειωμένα προηγείται ο Random Forest που αναγνωρίζει σχεδόν όλα τα αρνητικά, και 102 από τα 125 θετικά.

Self training
confusion matrix με 200 στά 625(32.0%) πραγματικά επισημειωμένα δείγματα



Εικόνα 55: Semi-supervised learning , self-training, confution matrix για 200 δείγματα.

Στα 200 επισημειωμένα προηγείται ο LMGBMC, ο οποίος αναγνωρίζει όλα τα αρνητικά και 122 από τα 125 θετικά.

Πίνακας 2: Self-training - μετρικές με 25, 100, 200 πραγματικά επισημειωμένα δεδομένα.

Πραγματικά επισημειωθέντα: 25						
estimator	set	accuracy	precision	sensitivity	specificity	f1-score
SVC	test	0.8670	0.9430	0.3723	0.9920	0.4781
	train	0.8702	0.9242	0.3877	0.9889	0.5175
	Delta	0.0032	0.0188	0.0154	0.0031	0.0394
	STD	0.0480	0.0405	0.2374	0.0056	0.2792
Decision Tree	test	0.8606	0.7130	0.5380	0.9422	0.6063
	train	0.8702	0.7207	0.5809	0.9428	0.6422
	Delta	0.0096	0.0077	0.0429	0.0006	0.0359
	STD	0.0079	0.0975	0.0463	0.0265	0.0115
Logistic Regression	test	0.8317	0.6667	0.1749	1.0000	0.2681
	train	0.8277	1.0000	0.1347	1.0000	0.2204
	Delta	0.0040	0.3333	0.0402	0.0000	0.0477
	STD	0.0471	0.4714	0.1558	0.0000	0.2243
Random Forest	test	0.8750	0.6667	0.3522	1.0000	0.4603
	train	0.8638	1.0000	0.3229	1.0000	0.4382
	Delta	0.0112	0.3333	0.0293	0.0000	0.0221
	STD	0.0379	0.4714	0.2512	0.0000	0.3267
LGMBMC	test	0.8013	0.0000	0.0000	1.0000	0.0000
	train	0.8061	1.0000	0.0242	1.0000	0.0473
	Delta	0.0048	1.0000	0.0242	0.0000	0.0473
	STD	0.0099	0.0000	0.0000	0.0000	0.0000
Multi layer Perceptron	test	0.9279	0.9685	0.6625	0.9940	0.7844
	train	0.9287	0.9646	0.6732	0.9930	0.7872

	Delta	0.0008	0.0039	0.0106	0.0009	0.0028
	STD	0.0068	0.0242	0.0586	0.0050	0.0342
Πραγματικά επισημειωθέντα: 100						
estimator	**set**	**accuracy**	**precision**	**sensitivity**	**specificity**	**f1-score**
SVC	test	0.9151	0.9421	0.6203	0.9901	0.7440
	train	0.9183	0.9402	0.6314	0.9900	0.7534
	Delta	0.0032	0.0018	0.0111	0.0001	0.0095
	STD	0.0163	0.0442	0.0757	0.0075	0.0493
Decision Tree	test	0.9119	0.7844	0.7843	0.9465	0.7786
	train	0.9367	0.8350	0.8556	0.9568	0.8448
	Delta	0.0248	0.0506	0.0712	0.0102	0.0662
	STD	0.0082	0.0786	0.0566	0.0192	0.0211
Logistic Regression	test	0.9135	0.9722	0.5857	0.9961	0.7305
	train	0.9151	0.9683	0.5933	0.9950	0.7333
	Delta	0.0016	0.0039	0.0076	0.0011	0.0028
	STD	0.0104	0.0196	0.0260	0.0028	0.0190
Random Forest	test	0.9471	0.9453	0.7852	0.9880	0.8549
	train	0.9583	0.9639	0.8220	0.9919	0.8869
	Delta	0.0112	0.0186	0.0368	0.0040	0.0320
	STD	0.0104	0.0323	0.0657	0.0084	0.0240
LGMBMC	test	0.9615	0.9709	0.8294	0.9941	0.8933
	train	0.9623	0.9425	0.8675	0.9859	0.9027
	Delta	0.0008	0.0284	0.0381	0.0081	0.0094
	STD	0.0104	0.0223	0.0607	0.0049	0.0383
Multi layer Perceptron	test	0.9471	0.9334	0.7859	0.9859	0.8520
	train	0.9423	0.9097	0.7966	0.9790	0.8458
	Delta	0.0048	0.0237	0.0107	0.0069	0.0062
	STD	0.0079	0.0099	0.0648	0.0031	0.0418
Πραγματικά επισημειωθέντα: 200						
estimator	**set**	**accuracy**	**precision**	**sensitivity**	**specificity**	**f1-score**
SVC	test	0.9215	0.9432	0.6479	0.9899	0.7681
	train	0.9255	0.9407	0.6738	0.9889	0.7850
	Delta	0.0040	0.0025	0.0259	0.0010	0.0168
	STD	0.0113	0.0095	0.0032	0.0031	0.0037
Decision Tree	test	0.9343	0.8365	0.8470	0.9583	0.8352
	train	0.9583	0.9135	0.8743	0.9789	0.8928
	Delta	0.0240	0.0770	0.0273	0.0206	0.0575
	STD	0.0138	0.0612	0.1083	0.0169	0.0404
Logistic Regression	test	0.9151	0.9460	0.6046	0.9921	0.7367
	train	0.9247	0.9544	0.6523	0.9919	0.7739
	Delta	0.0096	0.0084	0.0478	0.0001	0.0372
	STD	0.0099	0.0266	0.0459	0.0025	0.0368
Random Forest	test	0.9631	0.9915	0.8241	0.9980	0.8997
	train	0.9720	0.9952	0.8639	0.9990	0.9249
	Delta	0.0088	0.0037	0.0398	0.0011	0.0252
	STD	0.0060	0.0121	0.0272	0.0029	0.0145
LGMBMC	test	0.9599	0.9459	0.8536	0.9884	0.8944
	train	0.9768	0.9823	0.8997	0.9960	0.9392
	Delta	0.0168	0.0364	0.0462	0.0076	0.0448
	STD	0.0113	0.0764	0.0381	0.0163	0.0297
Multi layer Perceptron	test	0.9407	0.9423	0.7541	0.9878	0.8370
	train	0.9407	0.9503	0.7430	0.9899	0.8336
	Delta	0.0000	0.0080	0.0111	0.0021	0.0034
	STD	0.0138	0.0171	0.0388	0.0051	0.0227

Συμπεράσματα self-training:

1. 200 πραγματικά επισημειωθέντα είναι αρκετά για να ψευδοεπισημειώσουμε τα υπόλοιπα 415 και για να δημιουργήσουμε ένα πολύ αξιόπιστο ταξινομητή (f1-score > 75%) με οποιονδήποτε αλγόριθμο βάσης.
2. Πιο λίγα, δηλαδή 100 επισημειωμένα δείγματα χρειάζονται οι LGBMC, Random Forest και Decision Tree για να αναπτυχθεί ένα μοντέλο με f1-score >80%
3. Εάν θέλουμε να έχουμε σίγουρα f1-score 90% τότε πρέπει να έχουμε 200 δείγματα και έναν από τους αλγόριθμους LGBMC ή Random Forest ή Decition Tree.
4. Η specificity είναι πολύ υψηλή από τα λίγα επισημειωμένα, που σημαίνει ότι τα αρνητικά ανά γνωρίζονται πιο εύκολα.
5. Δεν υπάρχει νόημα για κανένα εκτιμητή βάσης να αυξήσουμε τα πραγματικά επισημειωμένα, πάνω από 200 για κανένα εκτιμητή βάσης καθόσον από εκεί και πάνω αυξάνει η f1-score αυξάνει με πολύ μικρό ρυθμό σε σχέση με την αύξηση του αριθμού των επισημειωμένων.. Ήδη όμως πάνω από 100 ο ρυθμός αύξησης του f1-score είναι πολύ μικρός σε σχέση με την αύξηση των επισημειωμένων. Η απόδοση των δύο κορυφαίων επιλογών όταν έχουμε 200 πραγματικά επισημειωμένα

estimator	set	accuracy	precision	sensitivity	specificity	f1-score
Random Forest	test	0.9631	0.9915	0.8241	0.9980	0.8997
	train	0.9720	0.9952	0.8639	0.9990	0.9249
	Delta	0.0088	0.0037	0.0398	0.0011	0.0252
	STD	0.0060	0.0121	0.0272	0.0029	0.0145
LGMBMC	test	0.9599	0.9459	0.8536	0.9884	0.8944
	train	0.9768	0.9823	0.8997	0.9960	0.9392
	Delta	0.0168	0.0364	0.0462	0.0076	0.0448
	STD	0.0113	0.0764	0.0381	0.0163	0.0297

.

(Ενότητα 8.3.γ) Διάδοση επισημειώσεων

Οι κλάσεις LabelPropagation και LabelSpreading

Εδώ έχουμε δύο περιπτώσεις:

class sklearn.semi_supervised.**LabelPropagation**(*kernel='rbf', *, gamma=20, n_neighbors=7, max_iter=1000, tol=0.001, n_jobs=None*) ¶

class sklearn.semi_supervised.**LabelSpreading**(*kernel='rbf', *, gamma=20, n_neighbors=7, alpha=0.2, max_iter=30, tol=0.001, n_jobs=None*)

Ο Label-Propagation χρησιμοποιεί τον πίνακα ομοιότητας (similarity graph) ή γειτνίασης που κατασκευάστηκε από τα δεδομένα χωρίς τροποποιήσεις. Ο Label-Spreading χρησιμοποιεί μια τροποποιημένη έκδοση του αρχικού γράφου (κανονικοποιεί τα βάρη των ακμών υπολογίζοντας τον κανονικοποιημένο πίνακα Laplace[40]).

Έχουμε δυνατότητα να χρησιμοποιήσουμε δύο πυρήνες ως αλγόριθμο (εκτιμητή βάσης). Τον «rbf» που ρυθμίζεται από την παράμετρο «gamma» και την «knn» που ρυθμίζεται από την n_neighbors.

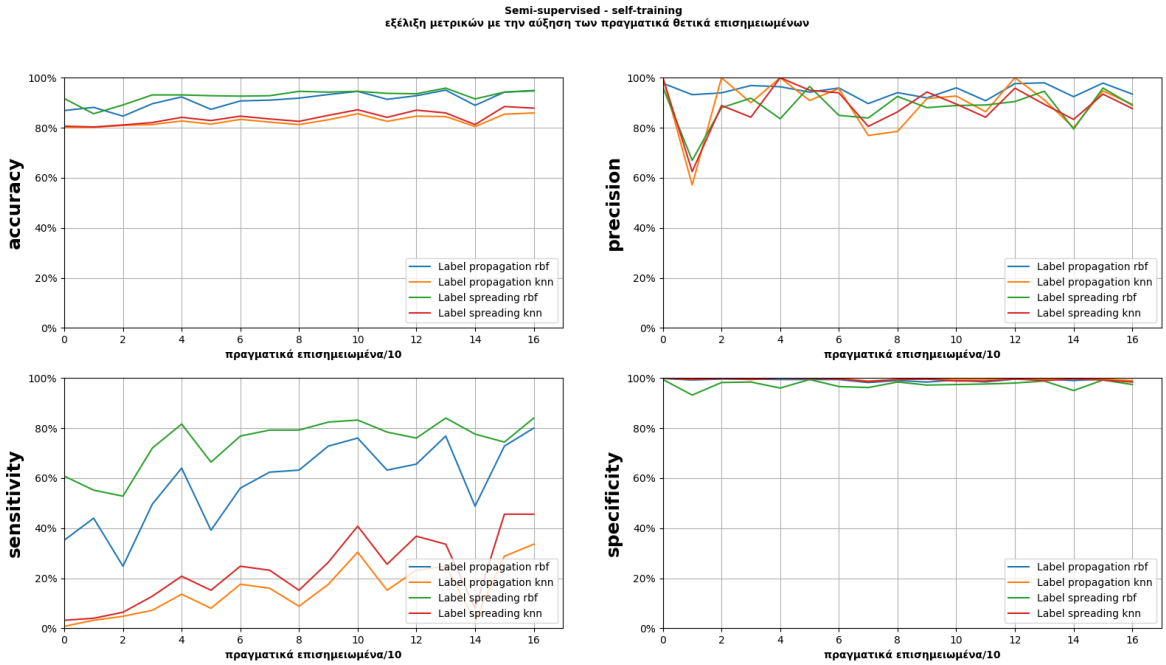
Ο πυρήνας «rbf» θα παράγει ένα πλήρως συνδεδεμένο γράφημα το οποίο αναπαρίσταται στη μνήμη από έναν πυκνό πίνακα. Αυτός ο πίνακας μπορεί να είναι πολύ μεγάλος και σε συνδυασμό με το κόστος εκτέλεσης ενός πλήρους πολλαπλασιασμού μήτρας, για κάθε επανάληψη του αλγορίθμου μπορεί να οδηγήσει σε απαγορευτικά μεγάλους χρόνους εκτέλεσης. Από την άλλη πλευρά, ο πυρήνας του «knn»παράγει μια πολύ πιο φιλική προς τη μνήμη αραιή μήτρα που μπορεί να μειώσει δραστικά τους χρόνους εκτέλεσης.

Όπως ο self-training οι αλγόριθμοι Label Propagation και Label-Spreading εκτελούνται επαναληπτικά μέχρις ότου, είτε φτάσουν τον μέγιστο αριθμό επαναλήψεων είτε δεν επιστρέφονται νέες ψευδοεπισημειώσεις.

Σκοπός είναι να βρούμε α) ποιος είναι ο ελάχιστος αριθμός πραγματικά επισημειωμένων δειγμάτων τον οποίο μπορούμε να χρησιμοποιήσουμε ως βάση και αφού τα συμπληρώσουμε με ψευδοεπισημειωμένα (αυτόματα επισημειωμένα) δεδομένα, μπορούμε να καταλήξουμε σε ένα μοντέλο που μπορεί να κάνει αξιοπρεπείς προβλέψεις, β) πιο μοντέλο βάσης μπορεί να έχει καλύτερα αποτελέσματα (υψηλότερες τιμές μετρικών με λιγότερα πραγματικά επισημειωμένα).

Η διαδικασία που ακολουθείται για την εκτέλεση των δοκιμών περιγράφεται στην ενότητα 8.3.α.

[40] Πίνακας Laplace είναι (πίνακας βαθμών – πίνακας ομοιότητας ενός γράφου).



Εικόνα 56: Label propagation-spreading - confusion matrix με 5 πραγματικά επισημειωμένα δεδομένα.

Από τον παραπάνω διάγραμμα βγάζουμε τα εξής συμπεράσματα:

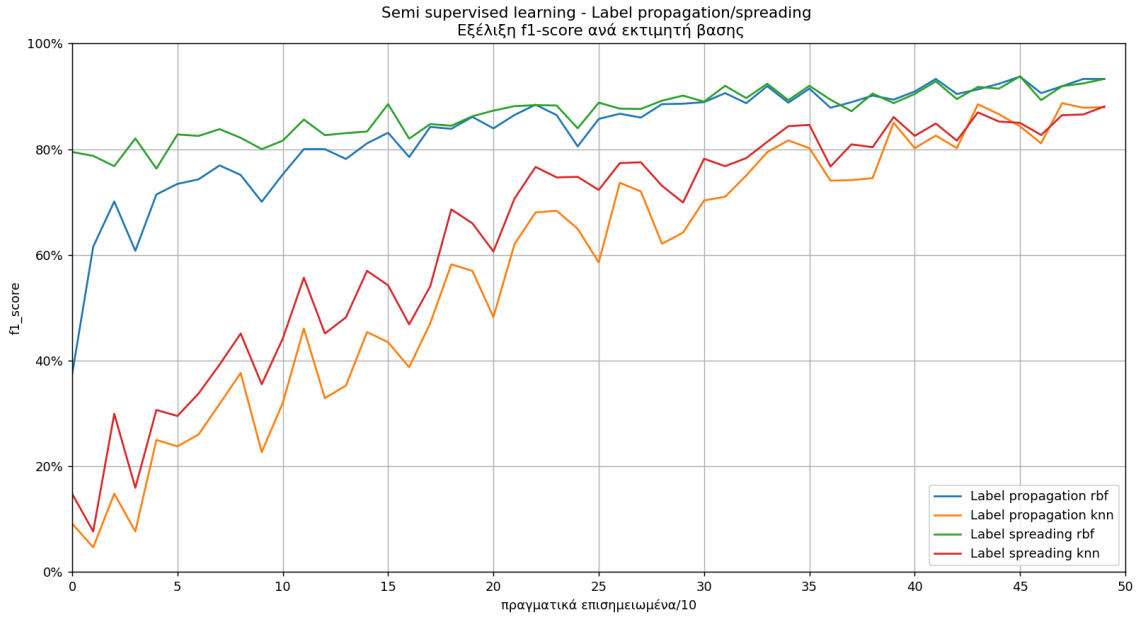
Σαν γενική παρατήρηση οι γραμμές όλων των μετρικών (πλην sensitivity) από ένα σημείο και πέρα (30 πραγματικά επισημειωμένα) είναι σχεδόν παράλληλες προς τον άξονα των χ, που σημαίνει ότι δεν επηρεάζονται ιδιαίτερα από την αύξηση των πραγματικά επισημειωμένων.

-**Η accuracy** είναι σταθερά πάνω από 80%, για τον πυρήνα «knn" και πάνω από 90% για τον πυρήνα «rbf». Λόγω όμως του ότι τα δεδομένα μας είναι μη ισορροπημένα (έχουμε πολύ περισσότερα αρνητικά) από θετικά, η accuracy δεν είναι η πλέον κατάλληλη μετρική, για την μέτρηση της αποτελεσματικότητας του μοντέλου.

-**Η precision** απαντά στο ερώτημα «ποιο ποσοστό των θετικών ταυτοποιήσεων ήταν πραγματικά σωστό;». Μια βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινό θετικό (TP). Όμως το 100% μπορεί να οφείλεται και στον τρόπο που υπολογίζεται το precision. Όταν ο παρονομαστής (FP+TP) είναι 0, τότε το αποτέλεσμα συμβατικά είναι 1, παρότι δεν είχαμε ούτε ένα (1) True Positive. Το ίδιο συμβαίνει και όταν μόνο του το FP είναι ίσο με 0. Σε όλους τους αλγορίθμους έχουμε μια τιμή πάνω από 80% με μία μικρή υπεροχή αυτών που χρησιμοποιούν πυρήνα «rbf».

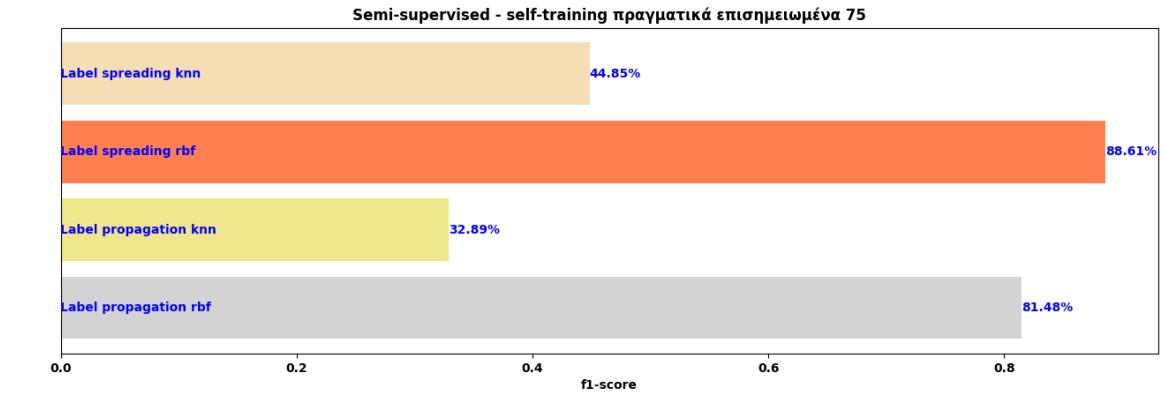
-**Η sensitivity**, είναι η ικανότητα του τεστ να εντοπίζει σωστά άτομα που νοσούν (πραγματικό θετικό ποσοστό). Μία βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά θετικό. Η εξέλιξη της sensitivity είναι σημαντική καθόσον το να αναγνωρίσουμε με ακρίβεια τους θετικούς μπορεί να συντελέσει στον περιορισμό της εξάπλωσης της νόσου. Εδώ έχουμε ένα σαφές προβάδισμα του πύρινα «rbf». Στα 160 πραγματικά επισημειωμένα η sensitivity με πυρήνα «rbf» ξεπερνά το 80%.

-**Η specificity**, δηλαδή η ικανότητα να ανακαλύπτει τους πραγματικά αρνητικούς από μικρό αριθμό θετικά επισημειωμένων είναι σε υψηλά επίπεδα. Το γεγονός αυτό, εν μέρει δικαιολογείται από τον μεγάλο αριθμό των υπαρχόντων αρνητικών. Μία τιμή 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά αρνητικό. Όμως εμείς προτιμούμε να μην χάσει κανένα θετικό.

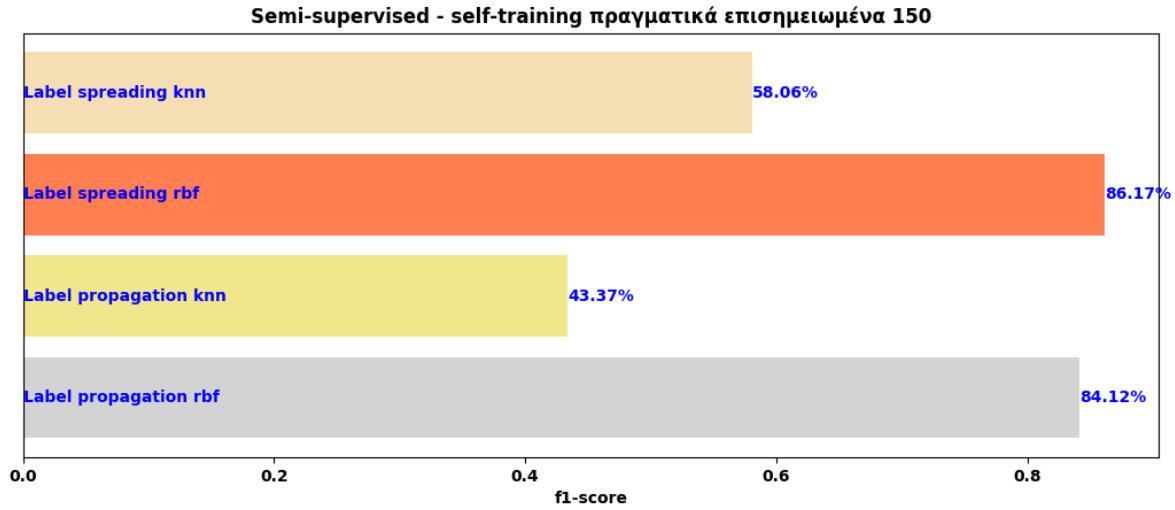


Εικόνα 57: Semi-supervised learning, label propagation/label spreading, εξέλιξη f1-score ανά αλγόριθμο.

Στο παραπάνω διάγραμμα φαίνεται η εξέλιξη, της μετρικής f1-score, με την αύξηση των πραγματικά επισημειωμένων δειγμάτων ανά αλγόριθμο. Εδώ βλέπουμε ότι ο αλγόριθμος που χρησιμοποιείται (label propagation ή label spreading) δεν έχει τόση μεγάλη σημασία όσο πυρήνας, δηλαδή ο αλγόριθμος βάσης. Ο «rbf» έχει καλύτερη πορεία από τον «knn». Βλέπουμε ότι **ο πυρήνας rbf στα 150 επισημειωμένα δείγματα έχει σταθεροποιηθεί σε υψηλά επίπεδα (80%), ενώ ο πυρήνας knn θα το πετύχει στα 450**. Αυτό αποδεικνύεται και από τα παρακάτω ιστογράμματα.



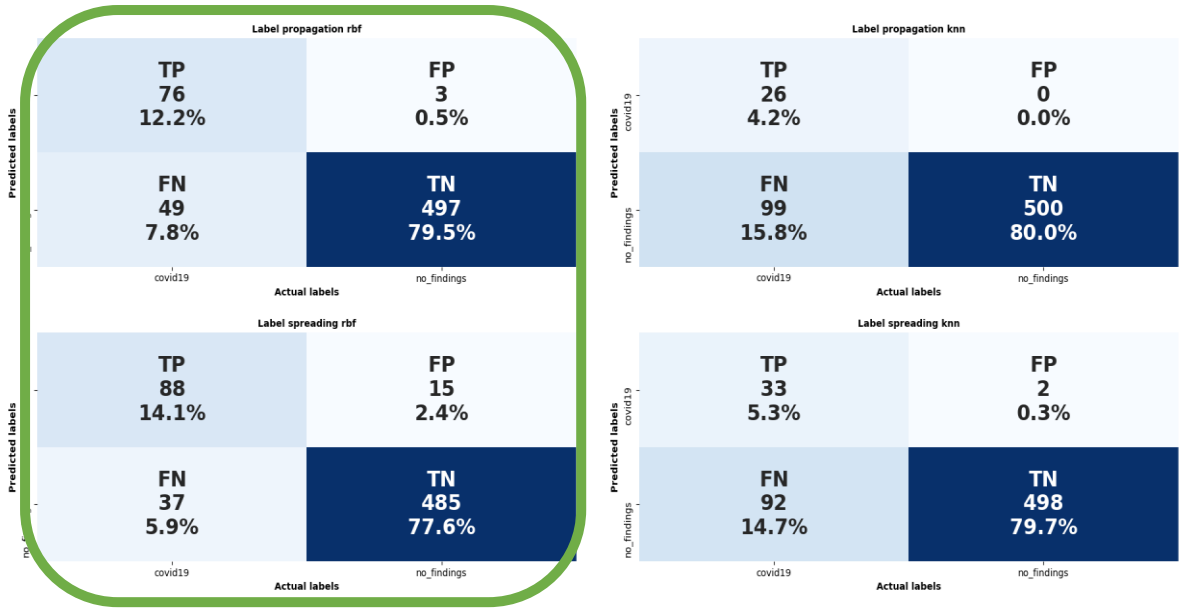
Εικόνα 58: Label propagation-spreading - ιστόγραμμα απόδοσης των εκτιμητών βάσης για 75 πραγματικά επισημειωμένα.



Εικόνα 59: Label propagation-spreading – ιστόγραμμα απόδοσης των εκτιμητών βάσης για 120 πραγματικά επισημειωμένα.

Εικόνα 60: Label propagation-spreading - ιστόγραμμα απόδοσης των εκτιμητών βάσης για 450 πραγματικά επισημειωμένα.

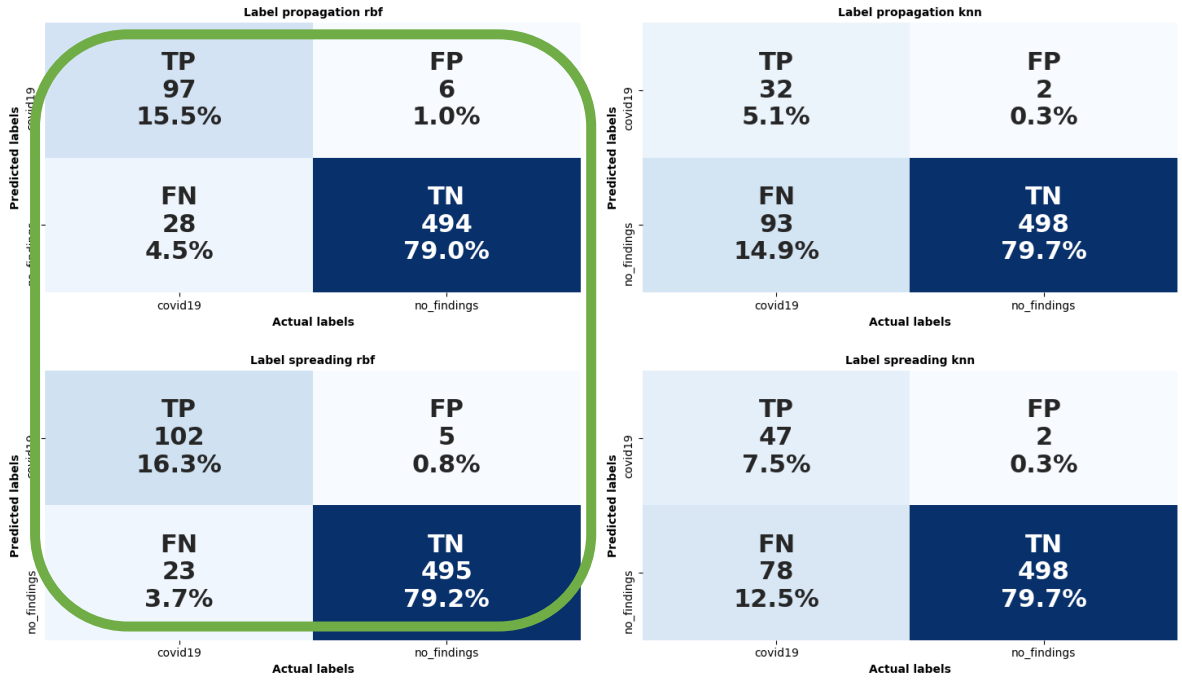
Semi supervised learning - Label propagation/spreading
confusion matrix με 75 στά 625(12.0%) πραγματικά επισημειωμένα δείγματα



Εικόνα 61: Semi-supervised confusion matrix για 75 πραγματικά επισημειωμένα.

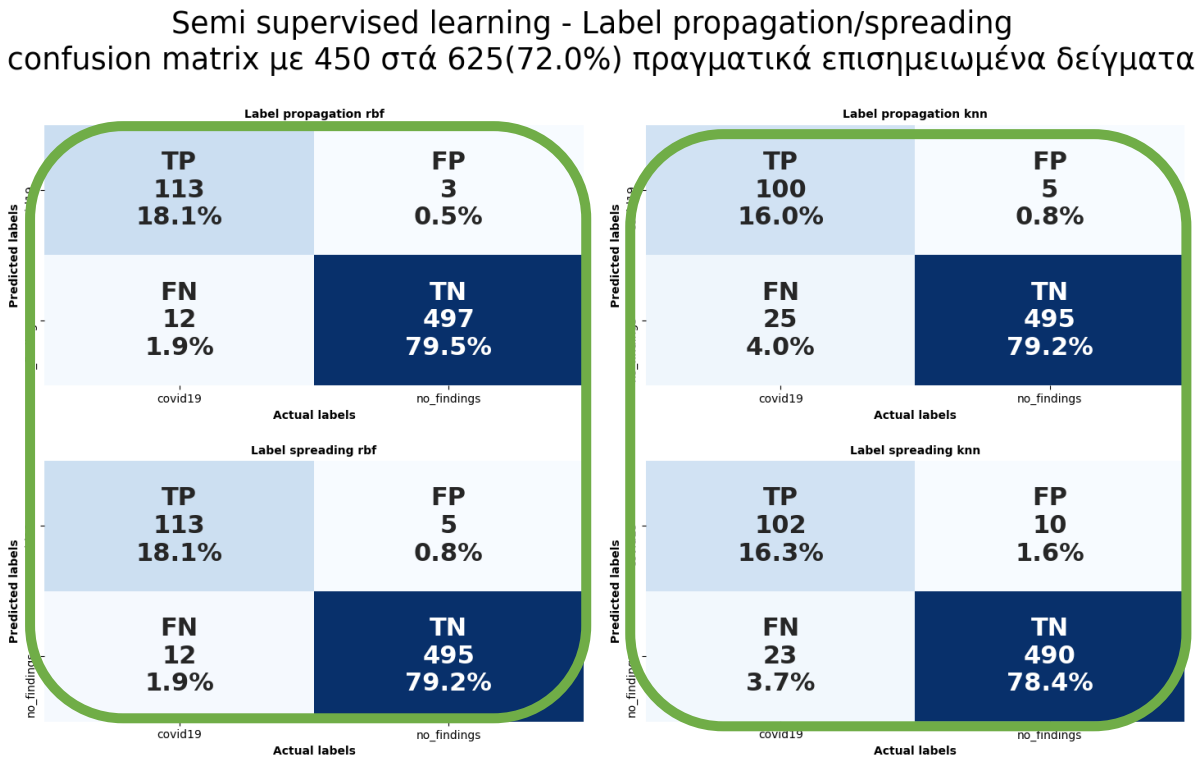
Παρατηρούμε ότι οι πυρήνες «rbf» υπερτερούν των knn στα 75 πραγματικά επισημειωμένα.

Semi supervised learning - Label propagation/spreading
confusion matrix με 150 στά 625(24.0%) πραγματικά επισημειωμένα δείγματα



Εικόνα 62: Semi-supervised confusion matrix για 150 πραγματικά επισημειωμένα.

Παρατηρούμε ότι οι πυρήνες «rbf» υπερτερούν των knn και στα 150πραγματικά επισημειωμένα. Οι «rbf" ήδη μπορούν να βρουν σωστά περί τους 100 θετικούς από τους 125.



Εικόνα 63: Semi-supervised confusion matrix για 450 πραγματικά επισημειωμένα.

Παρατηρούμε ότι οι πυρήνες «rbf» υπερτερούν των knn και στα 450 πραγματικά επισημειωμένα. Οι «knn» βρίσκονται στην κατάσταση που βρισκόταν οι «rbf" στα 150 επισημειωμένα.

Πίνακας 3: Label – propagation – spreading - μετρικές με 75-150-450 πραγματικά επισημειωμένα δεδομένα.

Πραγματικά επισημειωθέντα: 75						
estimator	set	accuracy	precision	sensitivity	specificity	f1-score
Label propagation rbf	test	0.9054	0.9315	0.5495	0.9899	0.6808
	train	0.9103	0.9404	0.5904	0.9910	0.7216
	Delta	0.0048	0.0089	0.0408	0.0011	0.0408
	STD	0.0149	0.0083	0.1379	0.0030	0.1170
Label propagation knn	test	0.8077	0.4444	0.0584	0.9941	0.1029
	train	0.8325	0.8603	0.1958	0.9919	0.3184
	Delta	0.0248	0.4159	0.1374	0.0022	0.2155
	STD	0.0180	0.4157	0.0630	0.0047	0.1095
Label spreading rbf	test	0.9119	0.8067	0.7568	0.9541	0.7706
	train	0.9263	0.8440	0.7871	0.9608	0.8121
	Delta	0.0144	0.0373	0.0303	0.0067	0.0415
	STD	0.0267	0.0332	0.1446	0.0144	0.0738
Label spreading knn	test	0.8494	0.7564	0.3192	0.9782	0.4424
	train	0.8486	0.8810	0.2852	0.9900	0.4299
	Delta	0.0008	0.1246	0.0340	0.0118	0.0125
	STD	0.0060	0.2043	0.1233	0.0158	0.1490

Πραγματικά επισημειωθέντα: 150						
estimator	set	accuracy	precision	sensitivity	specificity	f1-score
Label propagation rbf	test	0.9151	0.9622	0.5974	0.9940	0.7338
	train	0.9279	0.9744	0.6528	0.9960	0.7791
	Delta	0.0128	0.0122	0.0554	0.0020	0.0453
	STD	0.0159	0.0328	0.0792	0.0049	0.0589
Label propagation knn	test	0.8349	0.8571	0.1939	0.9942	0.3140
	train	0.8558	0.9615	0.2934	0.9969	0.4480
	Delta	0.0208	0.1044	0.0995	0.0028	0.1340
	STD	0.0082	0.2020	0.0731	0.0082	0.1071
Label spreading rbf	test	0.9054	0.9103	0.6065	0.9845	0.7165
	train	0.9343	0.9485	0.6989	0.9899	0.7960
	Delta	0.0288	0.0382	0.0924	0.0054	0.0795
	STD	0.0334	0.1026	0.1161	0.0180	0.0865
Label spreading knn	test	0.8782	0.8460	0.4880	0.9736	0.6081
	train	0.8790	0.9151	0.4407	0.9891	0.5930
	Delta	0.0008	0.0691	0.0473	0.0154	0.0151
	STD	0.0023	0.0600	0.0982	0.0154	0.0699
Πραγματικά επισημειωθέντα: 450						
estimator	set	accuracy	precision	sensitivity	specificity	f1-score
Label propagation rbf	test	0.9343	0.9174	0.7344	0.9840	0.8154
	train	0.9655	0.9770	0.8479	0.9950	0.9078
	Delta	0.0312	0.0595	0.1135	0.0110	0.0924
	STD	0.0149	0.0359	0.0598	0.0055	0.0513
Label propagation knn	test	0.9151	0.9213	0.6186	0.9881	0.7381
	train	0.9415	0.9744	0.7232	0.9949	0.8278
	Delta	0.0264	0.0531	0.1047	0.0068	0.0897
	STD	0.0149	0.0398	0.0778	0.0045	0.0648
Label spreading rbf	test	0.9327	0.9348	0.7264	0.9863	0.8115
	train	0.9688	0.9732	0.8677	0.9940	0.9174
	Delta	0.0361	0.0384	0.1413	0.0077	0.1059
	STD	0.0104	0.0699	0.0725	0.0154	0.0228
Label spreading knn	test	0.9215	0.8778	0.7170	0.9740	0.7827
	train	0.9415	0.9196	0.7761	0.9830	0.8415
	Delta	0.0200	0.0418	0.0591	0.0090	0.0589
	STD	0.0194	0.0314	0.1090	0.0112	0.0548

Πολύ μεγάλη τυπική απόκλιση και διαφορά μεταξύ test και train στο Label propagation «knn» των 75 επισημασμένων, όπου έχουμε άριστο precision εκπαίδευσής και κακό δοκιμών. Εδώ υπάρχουν σαφώς ενδείξεις υπερπροσαρμογής.

Συμπεράσματα Label propagation και spreading

1. Ο πυρήνας «rbf» σαφώς υπερέχει του «knn". Οι αλγόριθμοι Label-Spreading και Label-Propagation δεν διαφέρουν ως προς την αποτελεσματικότητά τους.
2. Ο πυρήνας «rbf» απαιτεί περίπου 150 πραγματικά επισημειωμένα για να έχει accuracy 0.9151, precision 0.9622, sensitivity 0.5974, specificity 0.9940, f1-score 0.7338 ενώ ο «knn» θέλει περίπου 450 για να φτάσει στα ίδια επίπεδα.
3. Πάνω από τα 450 πραγματικά επισημειωμένα δεν υπάρχει εμφανής βελτίωση της f1-score με την αύξηση των πραγματεία επισημειωμένων., για κανένα πυρήνα και κανένα αλγόριθμο.
4. Ο πυρήνας «rbf» μαζί με τον αλγόριθμο Label-Propagation όταν έκανε χρήση 450 πραγματικών επισημειωμένων πέτυχε accuracy 0.9343,

precision 0.9174, sensitivity 0.7344, specificity 0.9840 και f1-score 0.8154.

(Ενότητα 8.3.δ) Ενεργή μάθηση.

Χρησιμοποιήθηκε η βιβλιοθήκη modal η οποία είναι ένα πλαίσιο για την ενεργή μάθηση χτισμένο πάνω στο scikit-learn.

Η κλάση ActiveLearner

Για να δημιουργήσουμε ένα αντικείμενο της κλάσης ActiveLearner, πρέπει να παρέχουμε δύο πράγματα: α) Ένα αλγόριθμο βάσης εποπτευόμενης μαθησης της scikit-learn και β) μια συνάρτηση στρατηγικής ερωτήματος.

Εάν έχουμε διαθέσιμα αρχικά δεδομένα εκπαίδευσης, μπορούμε να εκπαιδεύσουμε τον εκτιμητή περνώντας τον μέσω των ορισμάτων X_training και y_training.

Πχ

```
learner = ActiveLearner(
    estimator=RandomForestClassifier(),
    query_strategy=uncertainty_sampling
    X_training=X_training, y_training=y_training
)
```

Αφού δημιουργήσουμε το αντικείμενο ActiveLearner, αυτό είναι έτοιμο να ρωτήσει και να συμπληρώσει τις γνώσεις του. Το μοντέλο που δημιουργήσαμε παρακολουθεί τα δεδομένα εκπαίδευσης που έχει δει κατά τη διάρκεια της ζωής του.

Για να εκπαιδεύσουμε αρχικά ένα μοντέλο ActiveLearner (πχ AL), το κάνουμε όπως ακριβώς εκπαιδεύουμε ένα μοντέλο πλήρως εποπτευόμενης μάθησης (Al.fit (X,y)), όπου X τα χαρακτηριστικά και y οι επισημειώσεις τους. Για να διδάξουμε το μοντέλο που δημιουργήσαμε με το ActiveLearner με νέες γνώσεις (X_new, y_new) που αποκτήθηκαν πρόσφατα, θα πρέπει να χρησιμοποιήσουμε τη μέθοδο .teach(X_new, y_new). Αυτό αυξάνει τα διαθέσιμα δεδομένα εκπαίδευσης με τα νέα δείγματα (X_new) και τις επισημειώσεις τους (y_new). Δηλαδή συμπληρώνει την εκπαίδευση και στη συνέχεια, επανατοποθετεί τον εκτιμητή σε αυτό το επαυξημένο σύνολο δεδομένων εκπαίδευσης. Η εντολή είναι:

```
learner.teach(X_new, y_new)
```

Εάν θέλουμε να ξεκινήσουμε την εκπαίδευση από το μηδέν, χρησιμοποιήσουμε τη μέθοδο .fit(X, y) . Τότε το μοντέλο «ξεχνάει όλα όσα έχει μάθει» και ξεκινάει ένα εντελώς νέο μοντέλο.

Για εκπαίδευση μόνο στα δεδομένα που αποκτήθηκαν πρόσφατα, θα πρέπει να περάσουμε στη μέθοδο .teach(), την υπερπαράμετρο only_new=True

Τα μοντέλα active learner ονομάζονται active επειδή εάν τους παρέχουμε δείγματα χωρίς επισημείωση, μπορούν να επιλέξουν τα καλύτερα από αυτά για να επισημειωθούν από κάποιο φυσικό πρόσωπο. Στο modAL, μπορούμε να το επιτύχουμε καλώντας τη μέθοδο .query(X), όπου X το σύνολο των μη επισημειωμένων δεδομένων που έχουμε.

```
query_idx, query_sample = Learner.query(X)

# ...απόκτηση νέων επισημειωμένων δεδομένων

Learner.teach(query_sample, query_label)
```

Η μέθοδος .query(X) καλεί τη συνάρτηση στρατηγικής ερωτήματος που καθορίσαμε κατά την αρχικοποίηση του ActiveLearner.

Οι διαθέσιμες ενσωματωμένες στρατηγικές ερωτημάτων είναι η δειγματοληψία μέγιστης αβεβαιότητας, η δειγματοληψία μέγιστου περιθωρίου και η δειγματοληψία εντροπίας (*max uncertainty sampling*, *max margin sampling* and *entropy sampling*.).

Για να χρησιμοποιήσουμε το ActiveLearner για προβλέψεις καλούμε την predict (X,y), όσον αφορά τις δοκιμές και την πιστοποίηση ακριβώς το ίδιο όπως σε μία πλήρως εποπτευόμενη μάθηση

Η κλάση Committee

Μία από τις δημοφιλείς στρατηγικές ενεργούς μάθησης είναι το «Query by Committee», που χρησιμοποιεί πολλούς αλγορίθμους βάσης και οι απαντήσεις στα ερωτήματα (query strategies) βασίζονται στην γνώμη της «επιτροπής». Δηλαδή, εκτελείται εσωτερικά μία διαδικασία συναξιολόγησης. Στο modAL, αυτό το μοντέλο εφαρμόζεται με την κλάση Committee.

Για να δημιουργήσουμε ένα αντικείμενο Committee, πρέπει να παρέχουμε δύο πράγματα: μια λίστα αντικειμένων ActiveLearner και μια συνάρ-

τηση στρατηγικής ερωτήματος. Μια λίστα με εκτιμητές scikit-learn δεν αρκεί, επειδή κάθε εκπαιδευόμενο μοντέλο πρέπει να παρακολουθεί τα δείγματα εκπαίδευσης που έχει δει.

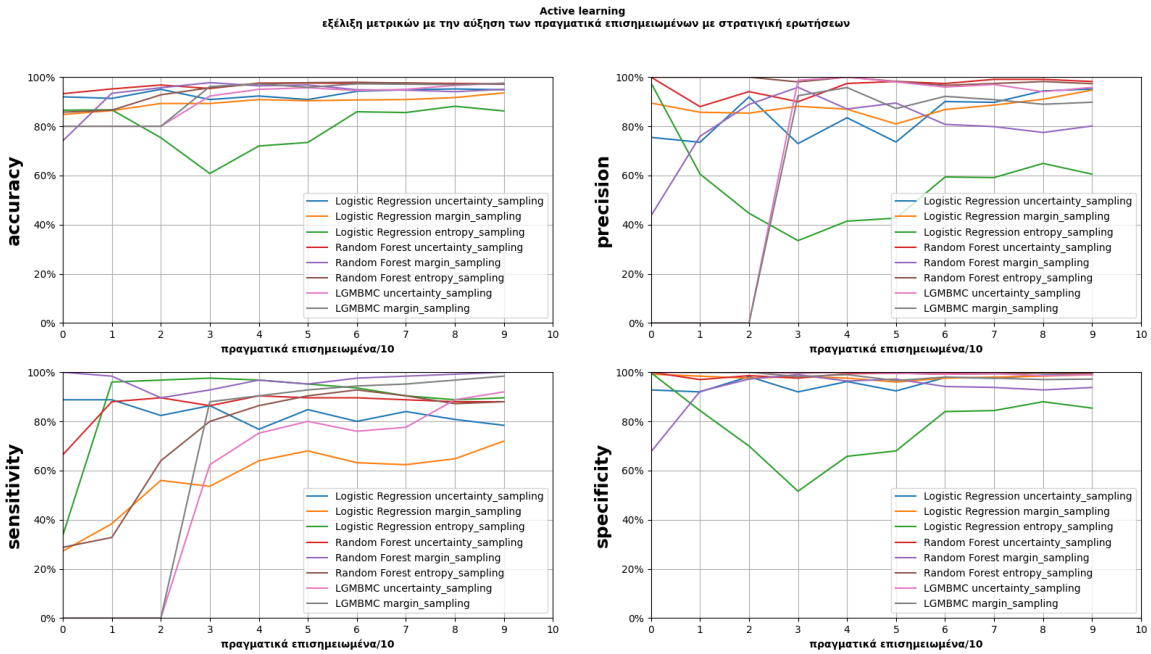
Η εκπαίδευση και η αναζήτηση δεδομένων προς επισημείωση λειτουργούν ακριβώς όπως και για το ActiveLearner. Για να διδάξουμε νέα δείγματα με την «επιτροπή», χρησιμοποιήσουμε επίσης τη μέθοδο .teach(X_new, y_new), όπου το X_new περιέχει τα νέα δείγματα εκπαίδευσης και το y_new τις αντίστοιχες επισημειώσεις τους.

Για να επιλέξουμε τις καλύτερες παρουσίες για επισημείωση, χρησιμοποιούμε τη μέθοδο .query(X), όπως και στο ActiveLearner, η οποία καλεί την συνάρτηση που έχουμε ορίσει ως στρατηγική ερωτήσεων. Επί του παρόντος, υπάρχουν τρεις ενσωματωμένες στρατηγικές ερωτήματος από επιτροπή στο modAL: μέγιστη εντροπία ψήφου, μέγιστη εντροπία αβεβαιότητας και μέγιστηδιαφωνία (*max_vote_entropy*, *max_vote_uncertainty_entropy* και

max_vote_ disagreement).

Η διαδικασία είναι αυτή που περιεγράφηκε στο 8.3.α . Η στρατηγική ερωτήσεων θα περιλαμβάνει τις τρείς (3) στρατηγικές ερωτήσεων οι οποίες προβλέπονται στο ModAl. Γενικά θα εξετασθούν όλοι οι συνδυασμοί «Active learner-αλγόριθμος βάσης-στρατηγική ερωτήσεων, αριθμός επισημειωμένων» και θα καταγραφεί το αποτέλεσμα που φέρουν όσον αφορά την τιμή της f1-score.

Ο σκοπός είναι να βρεθεί ο μικρότερος αριθμός των απαιτούμενων δεδομένων προς επισημείωση, ώστε με βάση την εκπαίδευση σε αυτά, να κατασκευάσουμε ένα αποδοτικό μοντέλο πλήρως εποπτευόμενης μάθησης. Ζητώντας επαναληπτικά περισσότερες ακτινογραφίες για επισημείωση μέχρι να φτάσουμε στην επιθυμητή τιμή f1-score. Εφαρμόσαμε, 3 διαφορετικούς αλγορίθμους βάσης και από τρεις διαφορετικές στρατηγικές ερωτήσεων για τον καθένα. Η επιτροπή συγκροτήθηκε από όλους του αλγόριθμους βάσης που δοκιμάσαμε, δηλαδή όλα τα μέλη της είναι διαφορετικού τύπου. Η κάθε ερώτηση του αλγορίθμου προς τον «μάντη ή δάσκαλο» που είναι το φυσικό πρόσωπο που επισημειώνει τις ακτινογραφίες και προσομοιώνεται από το πρόγραμμά μας αφορά ομάδα των δέκα (10).



Εικόνα 64: Active learning-εξέλιξη μετρικών με την αύξηση των επισημειωμένων

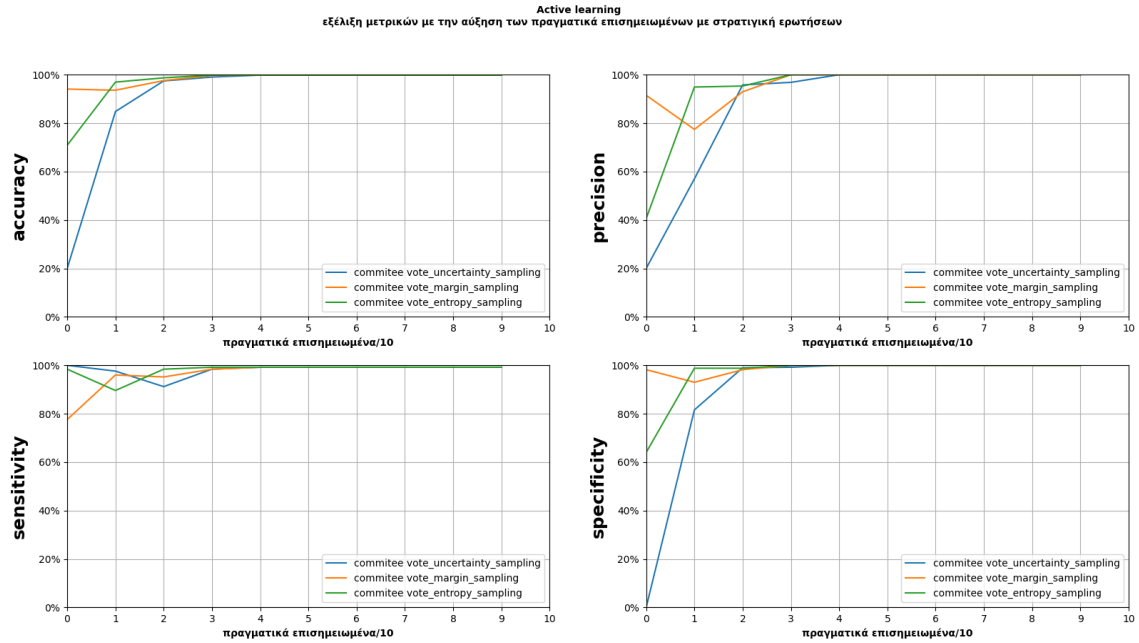
Από τα παραπάνω διαγράμματα βγάζουμε τα εξής συμπεράσματα:

-**Η accuracy** είναι σταθερά πάνω από 80%, πλην των Logistic Regression Margin Sampling, και Random Forest Margin Sampling, οι οποίοι φτάνουν σε αυτό το ύψος όταν επισημειωθούν περισσότερα από 60 δείγματα. Φαίνεται ότι οι «Random Forest Uncertainty Sampling» και «LGMBMC Margin Sampling» μετά τα 40 επισημειωθέντα υπερβαίνουν την τιμή 95% στην accuracy. Λόγω της ανισορροπίας δεδομένων η accuracy δεν είναι η καταλληλότερη μετρική για την αξιολόγηση του μοντέλου.

-**Η precision** απαντά στο ερώτημα «ποιο ποσοστό των θετικών ταυτοποιήσεων ήταν πραγματικά σωστό;». Μια βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινό θετικό (TP). Το 100% οφείλεται και στον τρόπο που υπολογίζεται το precision. Όταν ο παρονομαστής (FP+TP) είναι 0, τότε το αποτέλεσμα συμβατικά είναι 1, παρότι δεν είχαμε ούτε ένα (1) True Positive. Το ίδιο συμβαίνει και όταν μόνο του το FP είναι με 0. Η Logistic Regression Uncertainty Sampling υπερβαίνει διαρκώς το 95%.

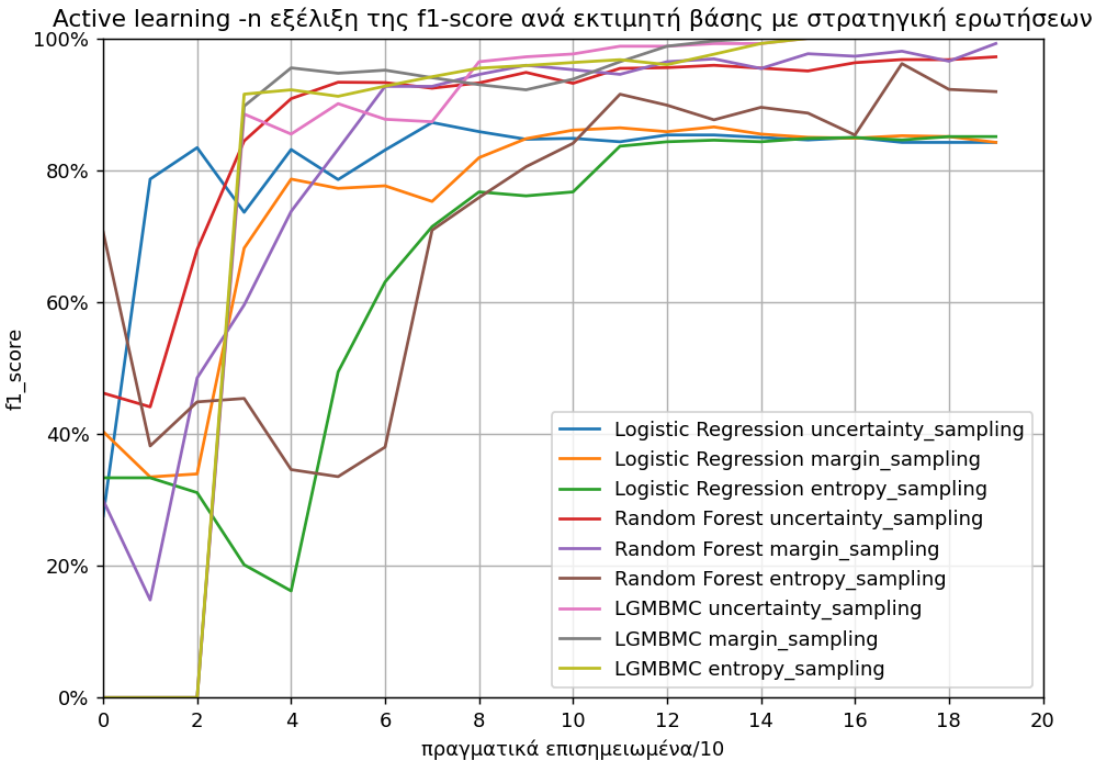
-**Η sensitivity**, η οποία είναι η ικανότητα του τεστ να εντοπίζει σωστά άτομα με τη νόσο (τιμή 100% σημαίνει ότι δεν έχασε κανένα πραγματικά θετικό) **εξελίσσεται διαφορετικά σε κάθε αλγόριθμο και στρατηγική ερωτήσεων.** Μέχρι τα 100 επισημειωθέντα υπερέχουν οι «Random Forest Margin Sampling» και «Logistic Regression Margin Sampling».

-**H specificity**, δηλαδή η ικανότητα να ανακαλύπτει τους πραγματικά αρνητικούς. Μία τιμή 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά αρνητικό. Εδώ υπερέχουν οι «Logistic Regression Uncertainty Sampling», «Logistic Regression Entropy Sampling» και «LGMBMC Margin Sampling».



Εικόνα 65: Active learning - εξέλιξη μετρικών με χρήση επιτροπής και στρατηγικών ερωτήσεων.

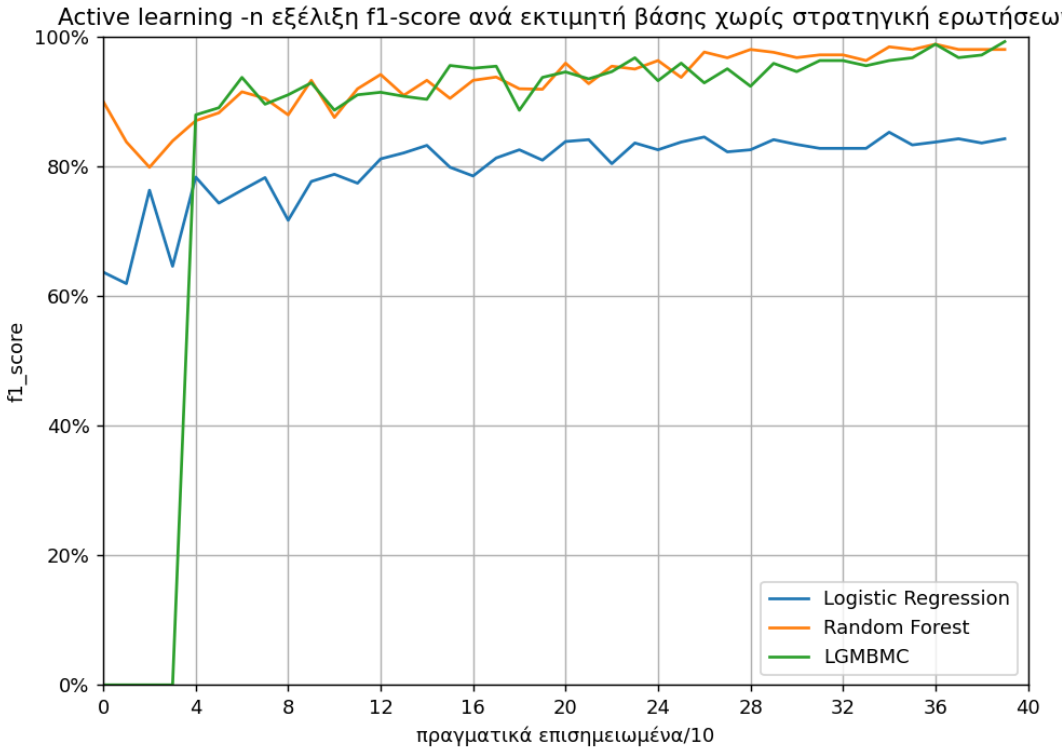
Από τα παραπάνω διαγράμματα προκύπτει αβίαστα ότι η επιτροπή υπερτερεί σε όλες τις μετρικές



Εικόνα 66 Active learning εξέλιξη f1 με στρατηγική ερωτήσεων.

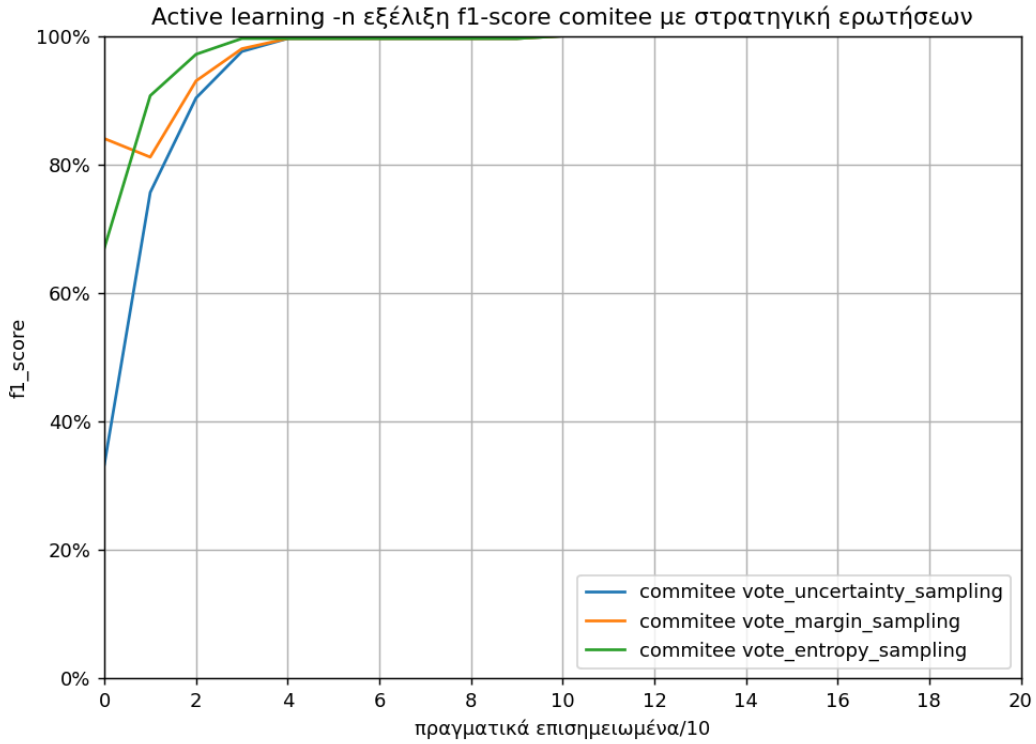
Παρατηρούμε ότι το σύνολο των αλγορίθμων βάσης, πλην του Logistic Regression περνάει το **80% f1-score στα 50 επισημειωμένα. Οι Random Forest και LGMBMC πλησιάζουν το 100%, όταν υπάρχουν 150 πραγματικά επισημειωμένα. Στα 100 επισημειωμένα όλα πλην του Logistic Regression, ξεπερνούν το 95%.**

Στο παρακάτω διάγραμμα παρατηρούμε ότι η συμπλήρωση των επισημειωμένων με τυχαία επιλογή, υστερεί αυτής με στρατηγική ερωτήσεων. Μάλιστα η διαφορά γίνεται εμφανέστερη όσο αυξάνεται ο αριθμός των επισημειωμένων. Αυτό δικαιολογείται επειδή, όσες περισσότερα επιλεγμένα επισημειωμένα προστίθενται, τόσο η βάση μας γίνεται καλύτερη. Στα 100 επισημειωμένα με στρατηγική ερωτήσεων, ο Random Forest και ο LGMBMC θα φτάσουν στα 140 επισημειωμένα το 95%, ενώ στην τυχαία επιλογή αυτό θα συμβεί στα 400. Γενικά φαίνεται και στις δύο περιπτώσεις ότι ο Logistic Regression υστερεί έναντι των άλλων.

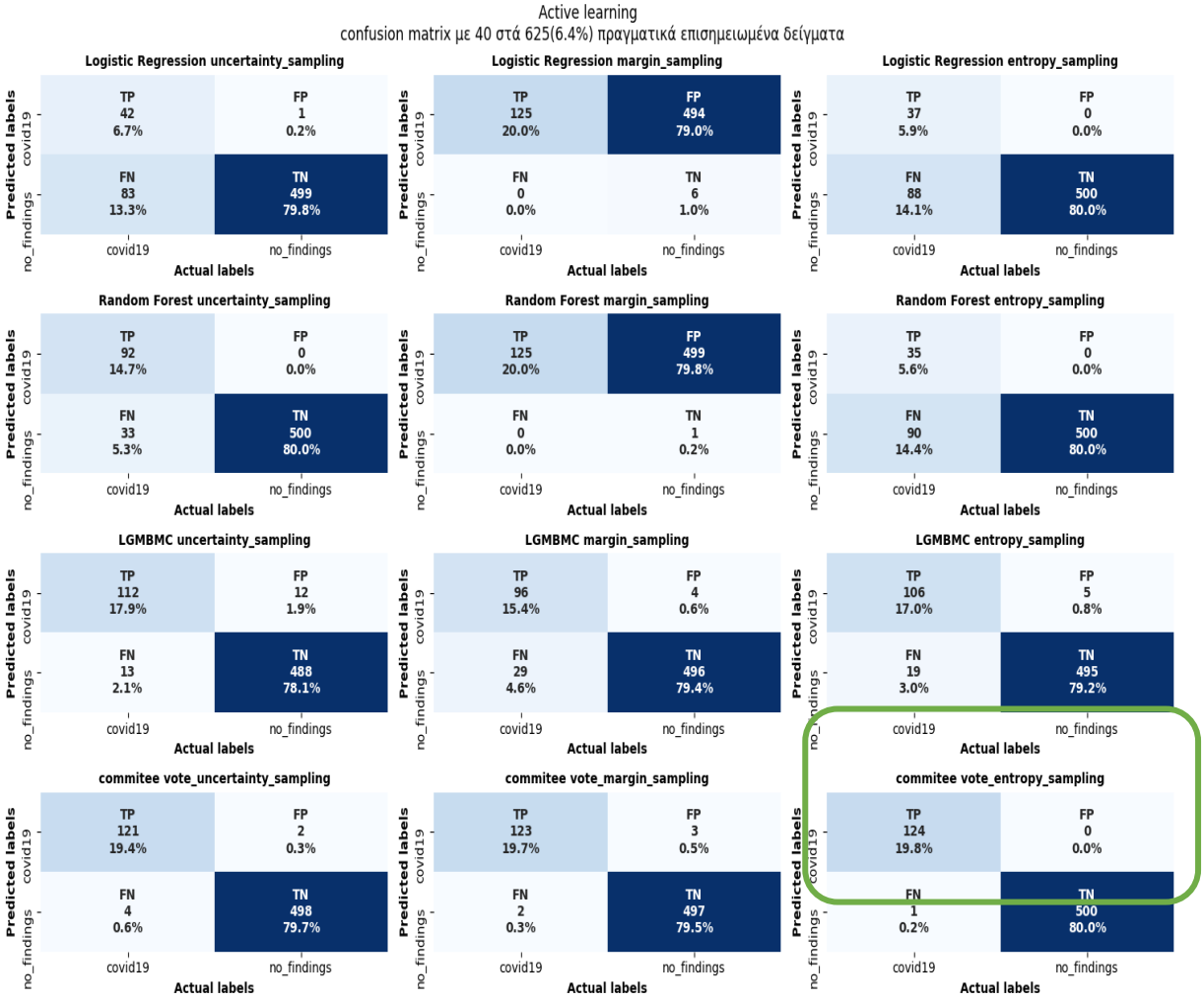


Εικόνα 67: Active Learning , εξέλιξη μετρικών χωρίς στρατηγική ερωτήσεων.

Στο παρακάτω διάγραμμα βλέπουμε ότι η επιτροπή των αλγορίθμων υ-περτερεί κατά πολύ αφού καταφέρνει να φτάσει το 100% στα 40 επισημειω-μένα δείγματα.

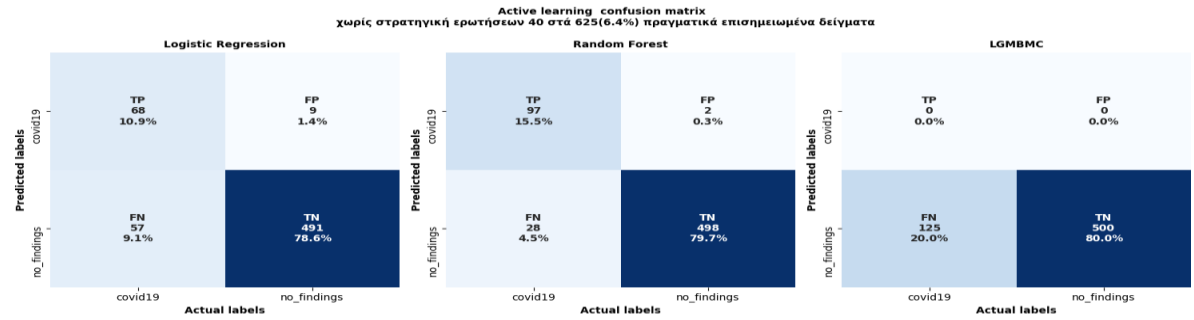


Εικόνα 68: Active learning, επιτροπή με στρατηγική ερωτήσεων.

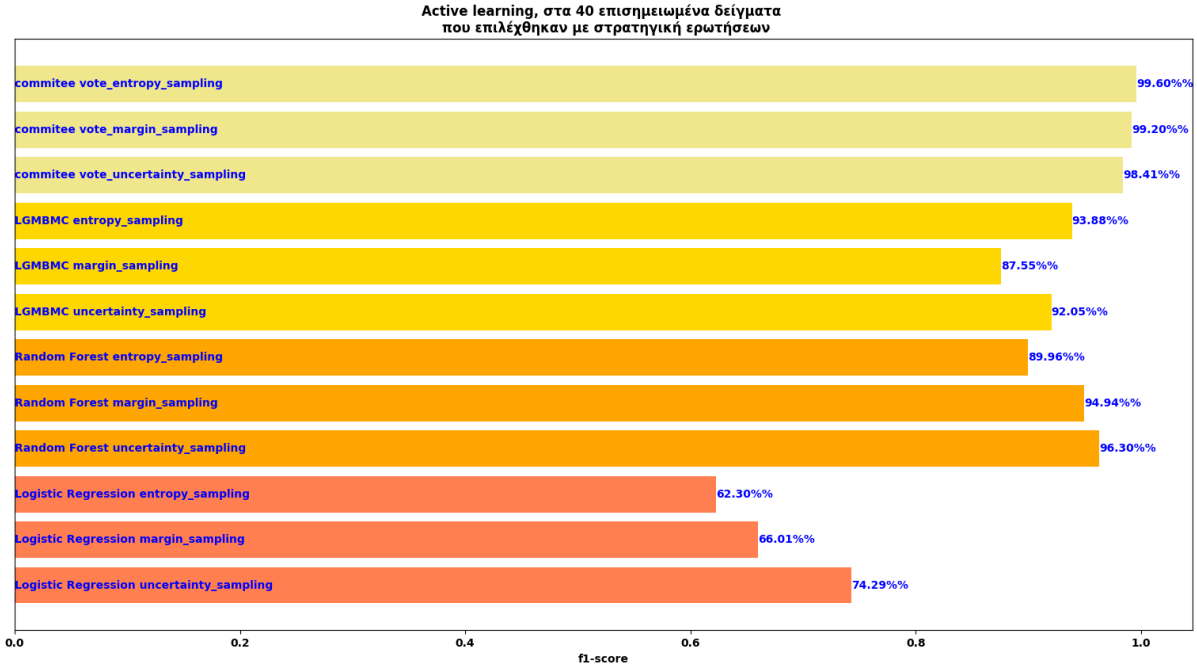


Εικόνα 69 Active learning - confusion matrix στα 40 επισημειωμένα με στρατηγική ερωτήσεων.

Όπως βλέπουμε στην παραπάνω εικόνα, στα 40 επισημειωμένα οι εκτιμήσεις των επιτροπών με οποιαδήποτε στρατηγική ειδικά δε με την vote_entropy_sampling έχει χάσει μόνο 1 στα 125 θετικά και κανένα από τα 500 αρνητικά. Ενώ στην παρακάτω εικόνα **με τυχαία επιλογή** των δειγμάτων που επισημειώνονται κάθε φορά, βλέπουμε ότι ο LGMBMC δεν βρήκε κανένα θετικό, ο Random Forest βρήκε 97 θετικά στα 125 και ο Logistic Regression 68 θετικά στα 125. Στην αποκάλυψη των αρνητικών τα πήγαν όλοι πολλοί καλά.

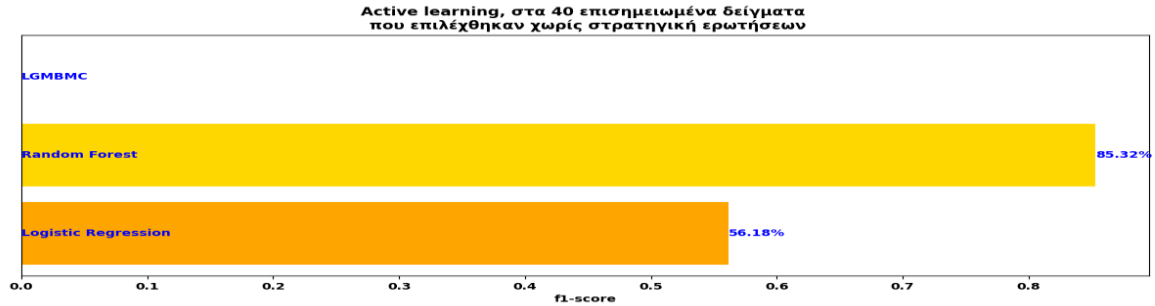


Εικόνα 70: Active Learner- confusion matrix στα 40 πραγματικά επισημειωμένα χωρίς στρατηγική ερωτήσεων.



Εικόνα 71: Active learning, ιστόγραμμα f1_score στα 40 δείγματα επιλεγμένα με την αντίστοιχη στρατηγική

Όπως φαίνεται από το ιστόγραμμα, η επιτροπή με committee-vote-entropy-sampling πετυχαίνει με 40 μόνο επισημειωμένα f1-score 99.6%, Χειρότερος είναι ο Logistic Regression με entropy-sampling f1-score 62.30%. Μία άλλη διαπίστωση που μπορούμε να κάνουμε από το παραπάνω ιστόγραμμα είναι, ότι μεγαλύτερη σημασία έχει το μοντέλο παρά η στρατηγική των ερωτήσεων στην διαμόρφωση των τιμών της f1-score.



Εικόνα 72: Active learning, ιστόγραμμα f1_score στα 40 επισημειωμένα που επελέγησαν χωρίς στρατηγική ερωτήσεων

Στο παραπάνω ιστόγραμμα, φαίνεται η εξέλιξη μετρικών με τυχαία επιλογή των προς επισημείωση, το f1_score απαιτεί αρκετή περισσότερη προσπάθεια για να φτάσει σε υψηλά επίπεδα. Στα 40 επισημειωμένα που επελέγησαν στην τύχη, έχουμε τιμές μετρικών από 0 έως 82,24%. Οι τιμές αυτές όταν έχουμε επισημειώσει μικρό αριθμό, τυχαία επιλεγμένων δειγμάτων ενέχουν και την τυχαιότητα.

Πίνακας 4: Μετρικές Active learning για 40 επισημειωμένα δείγματα που επελέγησαν χωρίς στρατηγική ερωτήσεων.

estimator	set	accuracy	precision	sensitivity	specificity	f1-score
Logistic Regression	test	0.9088	0.9722	0.5600	0.9960	0.7107
	train	0.9250	1.0000	0.6250	1.0000	0.7692
	Delta	0.0162	0.0278	0.0650	0.0040	0.0586
Random Forest	test	0.9536	1.0000	0.7680	1.0000	0.8688
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0464	0.0000	0.2320	0.0000	0.1312
LGMBMC	test	0.8000	0.0000	0.0000	1.0000	0.0000
	train	0.6750	0.0000	0.0000	1.0000	0.0000
	Delta	0.1250	0.0000	0.0000	0.0000	0.0000

Παρατηρούμε στον παραπάνω πίνακα, ότι με 40 επισημειωμένα δείγματα που επελέγησαν τυχαία δεν υπάρχει κανένας αλγόριθμος που να μας δίνει μοντέλο με αξιοπιστία πάνω από 75%.

Πίνακας 5: Μετρικές Active learning για 40 επισημειωμένα δείγματα που επελέγησαν με στρατηγική ερωτήσεων.

estimator	set	accuracy	precision	sensitivity	specificity	f1-score
Log Regr uncertainty_sampling	test	0.9360	0.8761	0.7920	0.9720	0.8319
	train	0.6000	0.8000	0.5714	0.6667	0.6667
	Delta	0.3360	0.0761	0.2206	0.3053	0.1653
Log Regr margin_sampling	test	0.8464	0.5714	0.9280	0.8260	0.7073
	train	0.7000	0.0000	0.0000	0.7778	0.0000
	Delta	0.1464	0.5714	0.9280	0.0482	0.7073
Log Regr entropy_sampling	test	0.9280	0.9545	0.6720	0.9920	0.7887
	train	0.6000	0.0000	0.0000	1.0000	0.0000
	Delta	0.3280	0.9545	0.6720	0.0080	0.7887
R. Forest uncertainty_sampling	test	0.9808	0.9669	0.9360	0.9920	0.9512
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0192	0.0331	0.0640	0.0080	0.0488
Random Forest margin_sampling	test	0.9504	0.8092	0.9840	0.9420	0.8881
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0496	0.1908	0.0160	0.0580	0.1119
Random Forest entropy_sampling	test	0.9040	1.0000	0.5200	1.0000	0.6842
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0960	0.0000	0.4800	0.0000	0.3158
LGMBMC uncertainty_sampl	test	0.9552	0.9709	0.8000	0.9940	0.8772
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0448	0.0291	0.2000	0.0060	0.1228
LGMBMC margin_sampling	test	0.9712	0.9908	0.8640	0.9980	0.9231
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0288	0.0092	0.1360	0.0020	0.0769
LGMBMC entropy_sampling	test	0.9632	0.9322	0.8800	0.9840	0.9053
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0368	0.0678	0.1200	0.0160	0.0947
commitee vote_uncert_sampl	test	0.9904	0.9760	0.9760	0.9940	0.9760
	train	1.0000	1.0000	1.0000	0.0000	1.0000
	Delta	0.0096	0.0240	0.0240	0.9940	0.0240
commitee vote_margin_sampl	test	0.9968	1.0000	0.9840	1.0000	0.9919
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0032	0.0000	0.0160	0.0000	0.0081
commitee vote_entropy_sampling	test	0.9984	1.0000	0.9920	1.0000	0.9960
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0016	0.0000	0.0080	0.0000	0.0040

Από τον παραπάνω πίνακα διαπιστώνουμε την υπεροχή της επιτροπής έναντι όλων των υπολοίπων μεθόδων. Χρειαζόμαστε πολύ πιο λίγες επισημειώσεις όταν η επιλογή των προς επισημείωση γίνεται από επιτροπή αλγορίθμων. Ο παραπάνω πίνακας προήλθε από δοκιμή σε όλα τα διαθέσιμα δεδομένα (625 δείγματα, ενώ η εκπαίδευση έγινε μόνο σε 40), που σημαίνει ότι τα αποτελέσματα αντιπροσωπεύουν επαρκή γενίκευση. Το commitee vote_entropy_sampling πέτυχε: accuracy=0.9984, precision=1.0000, sensitivity= 0.9920, specificity=1.0000, f1-score= 0.9960

Επιπρόσθετα παρατηρούμε ότι στο Logistic Regression, με στρατηγική margin-sampling έχουμε υποεφαρμογή, καθόσον στις περισσότερες μετρικές εκπαίδευσης έχουμε μηδενικές τιμές.

Συμπεράσματα Active learning:

1. Χρειαζόμαστε πολύ πιο λίγες επισημειώσεις όταν η επιλογή των προς επισημείωση γίνεται σύμφωνα με μία στρατηγική ερωτήσεων και όχι τυχαία.

2. Από ότι φαίνεται, μεγαλύτερη σημασία έχει το αλγόριθμος βάσης του μοντέλου και μικρότερη η στρατηγική των ερωτήσεων, με την προϋπόθεση βέβαια ότι ακολουθείται κάποια στρατηγική.

3. Οι επιτροπές με συμμετοχή διαφόρων τύπων μοντέλων είναι αυτές που μπορούν να φέρουν με λίγα επισημειωμένα δείγματα μεγάλες αποδόσεις (υψηλή τιμή f1-score.

4. Ο πρωταθλητής είναι ο committee vote_entropy_sampling, αλλά και οι υπολοίπου αλγόριθμοι επιτροπών από τα 40 δείγματα το f1_score πλησιάζει το 100%. Ο commitee vote_entropy_sampling πέτυχε: accuracy=0.9984, precision=1.0000, sensitivity= 0.9920, specificity=1.0000, f1-score= 0.9960.

(Ενότητα 8.3.ε) Μάθηση από θετικά και μη επισημειωμένα.

Στην εργασία μας χρησιμοποιήθηκαν οι αλγόριθμοι των Charles Elkan και Keith Noto (κλασικός και σταθμισμένος) καθώς και ο αλγόριθμος Bagging των Mordelet and Vert. Οι αλγόριθμοι των Elkan και Noto αντλήθηκαν από την βιβλιοθήκη pulearn όπως έχουν αναπτυχθεί (Drouin, AditraAS, Palachy, & Wright, n.d.) ενώ η bagging αντλήθηκε από Roy Wright (roywright on GitHub).

Classic Elkanoto

```
   from pulearn import ElkanotoPuClassifier
    from sklearn.svm import SVC
    svc = SVC(C=10, kernel='rbf', gamma=0.4, probability=True)
    pu_estimator = ElkanotoPuClassifier(estimator=svc, hold_out_ra-
tio=0.2)
       pu_estimator.fit(X, y)
```

Χρησιμοποιήθηκε ο παραπάνω ταξινομητής από την βιβλιοθήκη pulern που εφαρμόζει τον κλασικό αλγόριθμο Elkanoto, με διαφορετικούς αλγόριθμους βάσης, οι οποίοι και αξιολογήθηκαν.

Η μέθοδος που χρησιμοποιήθηκε περιγράφεται στην ενότητα 8.3.α.

Weighted Elkanoto

Χρησιμοποιήθηκε ο παρακάτω ταξινομητής από την βιβλιοθήκη της python ο οποίος υλοποιεί τον αλγόριθμο Weighted Elkanoto με τον ίδιο τρόπο που χρησιμοποιήθηκε ο Elkanoto.

```
from pulearn import WeightedElkanotoPuClassifier
    from sklearn.svm import SVC
    svc = SVC(C=10, kernel='rbf', gamma=0.4, probability=True)
    pu_estimator = WeightedElkanotoPuClassifier(
        estimator=svc, labeled=10, unlabeled=20, hold_out_ratio=0.2)
       pu_estimator.fit(X, y)
```

Bagging PU Classifier

Χρησιμοποιήθηκε ο κώδικας που ανέπτυξε ο (Wright., 2017) στην βάση της δημοσίευσης των (Mordelet & Vert, 2010) με τίτλο A bagging SVM to learn from positive and unlabeled examples και αυτός με τον ίδιο τρόπο που χρησιμοποιήθηκαν και οι προηγούμενοι.

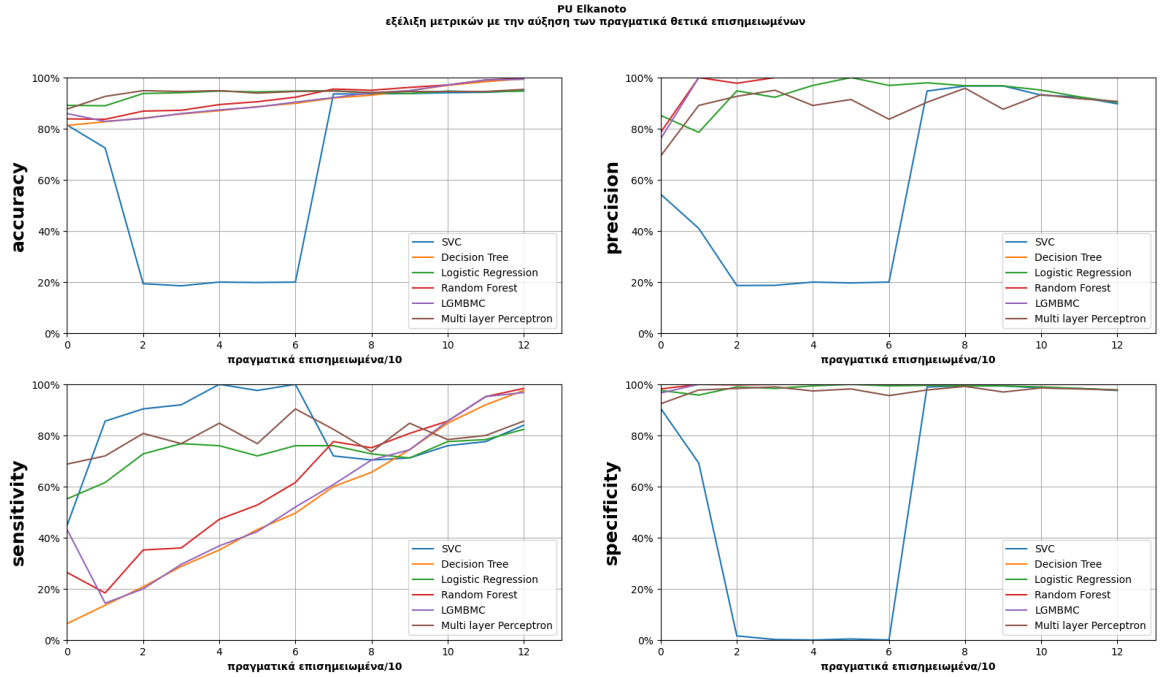
```
    from pulearn import BaggingPuClassifier
    from sklearn.svm import SVC
    svc = SVC(C=10, kernel='rbf', gamma=0.4, probability=True)
    pu_estimator = BaggingPuClassifier(
        base_estimator=svc, n_estimators=15)
       pu_estimator.fit(X, y)
```

Ο αλγόριθμος βασίζεται στον sklearn.ensemble.BaggingClassifiep ο οποίος τροποποιήθηκε ώστε να αντιμετωπίζει το πρόβλημα της «μάθησης με θετικά και μη επισημειωμένα δεδομένα.

Η εκπαίδευση γίνεται με βάση τους δύο αλγορίθμους Elkanoto (κλασικός και σταθμισμένος) καθόσον και την μέθοδο bagging. Οι αλγόριθμοι αυτοί χρησιμοποιούν ως εκτιμητές βάσης τους: SVC, Logistic Regression, Dcision Tree, Multi layer Perceptron, Random Forest και LGMBMC, οι οποίοι επελέγησαν λόγω του ότι ανήκουν σε διαφορετικές κατηγορίες (γραμμικός, σιγμοειδής, δένδρο αποφάσεων, τυχαίου δάσους (ensemble), gradient boosting και νευρωνικό δίκτυο.

Η μέθοδος που χρησιμοποιήθηκε περιγράφεται στην ενότητα 8.3.α.

Αποτελέσματα Elkanoto:



Εικόνα 73 PU learning Elkanoto, εξέλιξη μετρικών με την αύξηση των πραγματικά θετικά επισημειωμένων

Από τα παραπάνω διαγράμματα βγάζουμε τα εξής συμπεράσματα:

-**Η accuracy** είναι σε όλους τους αλγορίθμους βάσης, είναι σταθερά πάνω από 80%, όταν έχουμε περισσότερα από 20 θετικά επισημειωμένα στα 625 συνολικά δείγματα, εκτός από τον SVC. Όμως παρατηρούμε ότι όταν έχουμε διαθέσιμα κάτω από 40 θετικά επισημειωμένα, οι αλγόριθμοι Logistic Regression και Multi-Layer Perceptron, υπερτερούν των άλλων. Επιπλέον, φαίνεται ότι η accuracy, με αυτούς τους αλγορίθμους είναι ανεξάρτητη του λόγου θετικών επισημειωμένων προς σύνολο δεδομένων, καθώς αυτή δεν αυξάνεται σαφώς με την αύξηση των θετικά επισημειωμένων.

Φαίνεται ότι, οι Random Forest, LGMBMC, και Decision Tree υπερέχουν ελαφρώς στην τιμή της accuracy, έναντι των υπολοίπων όταν πλησιάζουμε να φτάσουμε τα 125 θετικά επισημειωμένα, τα οποία είναι το σύνολο των διαθεσίμων θετικών.

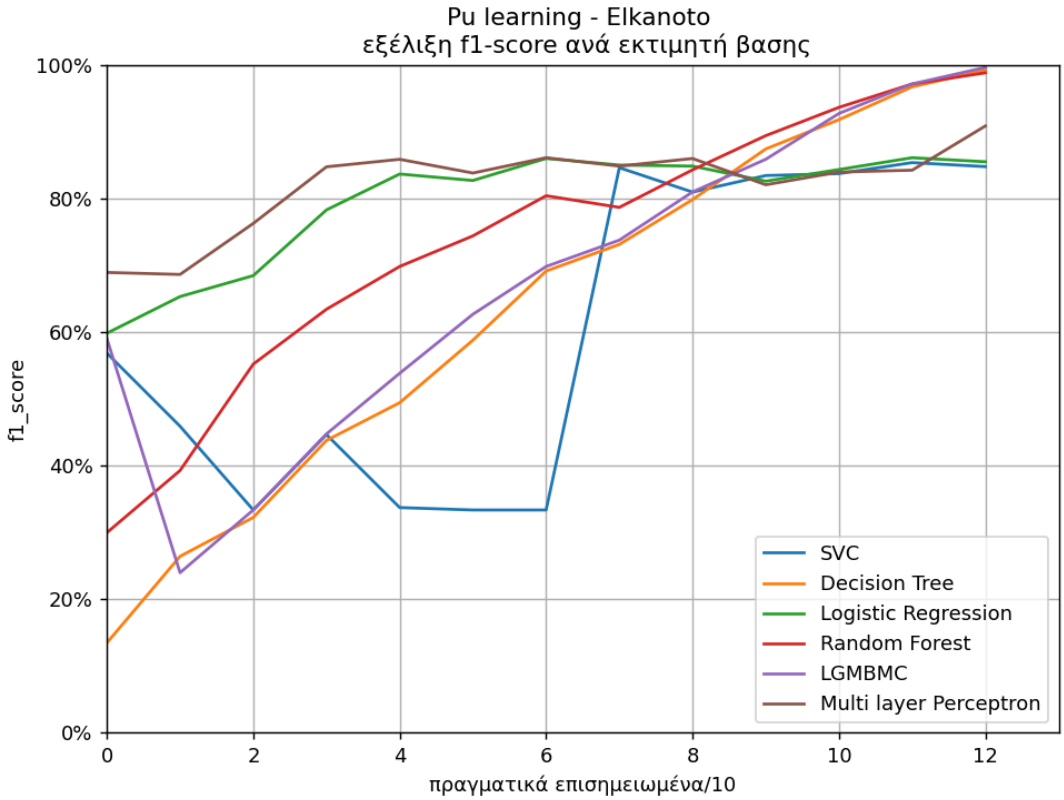
Ο SVC μέχρι τα 60 θετικά επισημειωμένα δεν έχει «καλή» accuracy, το οποίο οφείλεται, όπως θα δούμε παρακάτω, στην αδυναμία του να εντοπίσει τα θετικά, όταν τα θετικά επισημειωμένα είναι κάτω από 60.

Πρέπει να θυμόμαστε, ότι έχουμε ανισορροπία δεδομένων που σημαίνει ότι η accuracy δεν είναι η πλέον κατάλληλη μετρική.

-**Η precision** απαντά στο ερώτημα, «ποιο ποσοστό των θετικών ταυτοποιήσεων είναι πραγματικά σωστό;». Μια βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινό θετικό (TP). Όταν ο παρονομαστής (FP+TP) είναι 0, τότε το αποτέλεσμα συμβατικά είναι 1, παρότι δεν είχαμε ούτε ένα (1) True Positive. Το ίδιο συμβαίνει και όταν μόνο του το FP είναι ίσο με 0.Στην παρούσα περίπτωση οι Decision Tree, LGMBMC, και Random Forest εμφανίζονται με 100% precision που οφείλεται στο ότι το FP, όπως θα δούμε παρακάτω, είναι 0, που σημαίνει ότι οι αλγόριθμοι αυτοί βρίσκουν όλα τα αρνητικά χωρίς κανένα λάθος. Αυτό οφείλεται εν μέρει και στην ανισορροπία των αρνητικών και θετικών στην βάση με τα πραγματικά δεδομένα.

-**Η sensitivity**, είναι η ικανότητα του μοντέλου να εντοπίζει σωστά άτομα με τη νόσο (πραγματικό θετικό ποσοστό). Μία βαθμολογία στο 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινό θετικό. Στους Random Forest, LGMBMC, και Decision Tree, η sensitivity **είναι ανάλογη με την αύξηση των διαθεσίμων θετικά επισημειωμένων**. Οι SVC, Logistic Regression και Multi–Layer Perceptron επιδεικνύουν μια σταθερότητα στην τιμή της sensitivity περίπου στο 80%, όταν έχουν στην διάθεσή τους πάνω από 60 θετικά επισημειωμένα. Αυτό σημαίνει ότι από εκεί και πάνω, υπάρχει μία ανεξαρτησία της τιμής της sensitivity από την αναλογία του αριθμού των θετικών επισημειωμένων προς το σύνολο των δεδομένων. Η τιμή της sensitivity είναι σημαντική, καθόσον αναγνωρίζοντας με ακρίβεια τους θετικούς, μπορεί να συντελέσει στον περιορισμό της εξάπλωσης της νόσου.

-**Η specificity**, δηλαδή η ικανότητα να ανακαλύπτει τους πραγματικά αρνητικούς (το ποσοστό των πραγματικά αρνητικών που προσδιορίστηκε σωστά), φαίνεται ότι, από μικρό αριθμό θετικά επισημειωμένων, έχει υψηλές τιμές, πράγμα που εν μέρει δικαιολογείται από τον μεγάλο αριθμό των υπαρχόντων αρνητικών. Μία βαθμολογία προς το 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά αρνητικό. Από το διάγραμμα φαίνεται, ότι όλοι οι αλγόριθμοι βάσης πετυχαίνουν ποσοστό κοντά στο 100% (ο SVC υστερεί λίγο). Αυτό οφείλεται στο ότι ο αριθμός των FP είναι μηδέν.



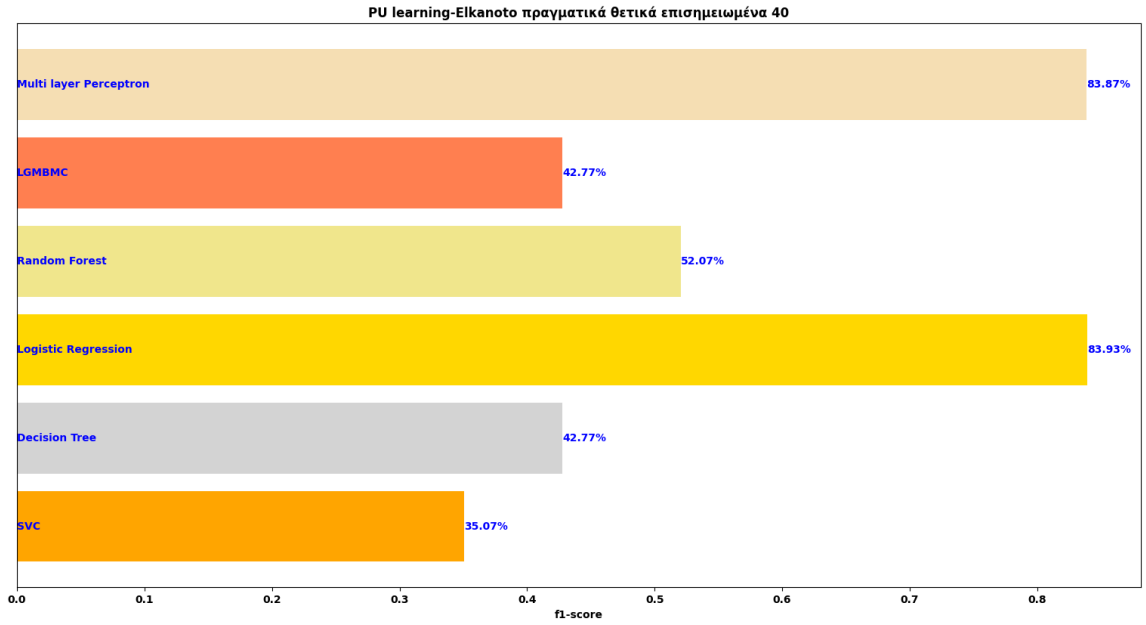
Εικόνα 74 Σύγκριση εξέλιξης f1 score ανά εκτιμητή βάσης.

Από την παραπάνω εικόνα, διαπιστώνουμε ότι η f1-score έχει την πιο σταθερή εξέλιξη, όταν εκτιμητής βάσης είναι ένα από τους Decision Tree, Random Forest και LGMBC, δηλαδή η αύξηση της τιμής της είναι ανάλογη με την αύξηση των θετικά επισημειωμένων. Έτσι μπορούμε να συμπεράνουμε, ότι παρότι τελικά που όλοι οι αλγόριθμοι βάσης οδηγούν σε ψηλά επίπεδα f1-score, όταν χρησιμοποιηθούν και τα 125 από τα θετικά επισημειωμένα δεδομένα, αυτοί οι τρείς (3) αλγόριθμοι είναι: α) πιο προβλέψιμοι σε σχέση με την αύξηση των θετικών επισημειωμένων, και β) έστω κατά λίγο υπερέχουν στην τελική τιμή της f1-score. Εάν όμως δεν έχουμε τόσο πολλά δεδομένα στην διάθεσή μας, πρέπει να προτιμήσουμε τον Multi-Layer Perceptron ή τον Logistic Regression, επειδή έχουν f1-score πάνω από 80% και στα λίγα θετικά επισημειωμένα, φαίνεται όμως απίθανο να μας δώσουν τιμή f1-score πάνω από 90%, με οποιονδήποτε αριθμό θετικά επισημειωμένων. Οι μόνοι εκτιμητές (αλγόριθμοι βάσης) που μπορούν να το κάνουν αυτό είναι: οι Decision Tree, LGMBC και Random Forest. Η σταθερή πορεία του αλγορίθμου Multi-Layer Perceptron και Logistic Regression, δείχνει ότι υπάρχει μία

ανεξαρτησία της τιμής της f1-score σε αυτούς τους αλγορίθμους από την αύξηση των θετικών επισημειωμένων, η οποία μπορεί να είναι χρήσιμη για εκτιμήσεις και σε άλλα σύνολα δεδομένων.

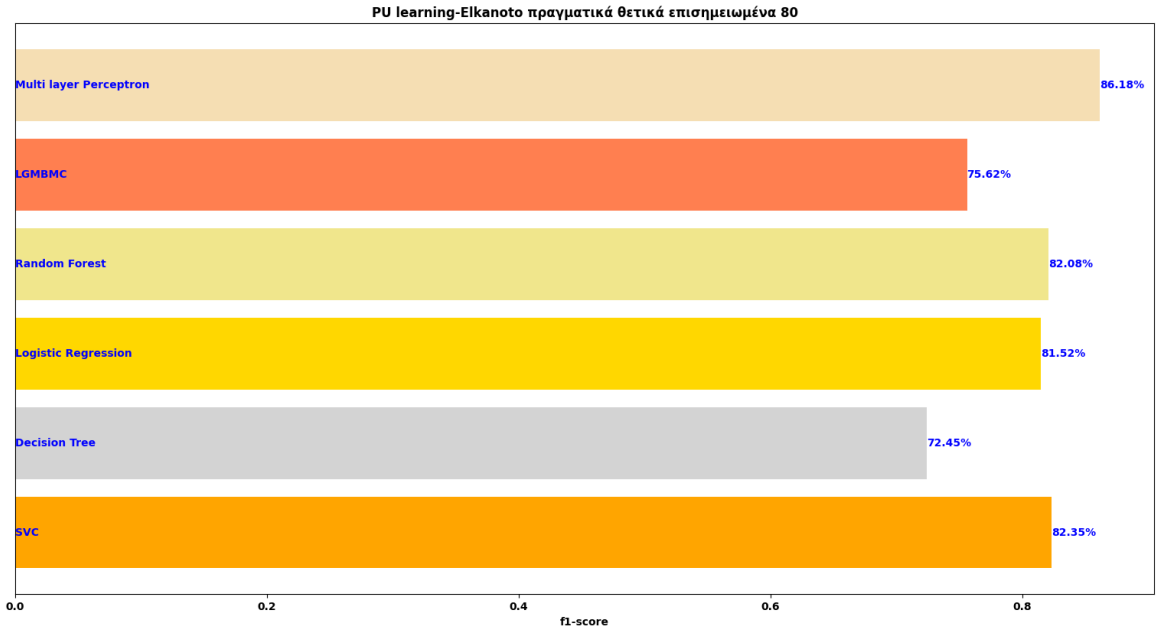
Το SVC κάτω από τα 80 θετικά επισημειωμένα δεν αναπτύσσεται αναλογικά με την αύξηση των θετικά επισημειωμένων. Φαίνεται σαφώς, ότι σε κάθε περίπτωση, πρέπει να κινηθούμε πάνω από τα 70 διαθέσιμα θετικά επισημειωμένα για να έχουμε μία απόδοση πάνω από 80%.

Τα σημεία που πρέπει να εξετασθούν με περισσότερη λεπτομέρεια είναι τα 40, τα 85 και τα 125 θετικά. Τα 40, επειδή οι αλγόριθμοι Multi-Layer Perceptron και Logistic Regression, ξεπερνούν το 80% στην τιμή της f1-score και την διατηρούν από εκεί και πέρα, τα 80 επειδή είναι σημείο όπου όλοι οι αλγόριθμοι έχουν περίπου την ίδια τιμή f1-score και τέλος τα 125 θετικά επισημειωμένα που είναι το σύνολο των θετικά επισημειωμένων που έχουμε.



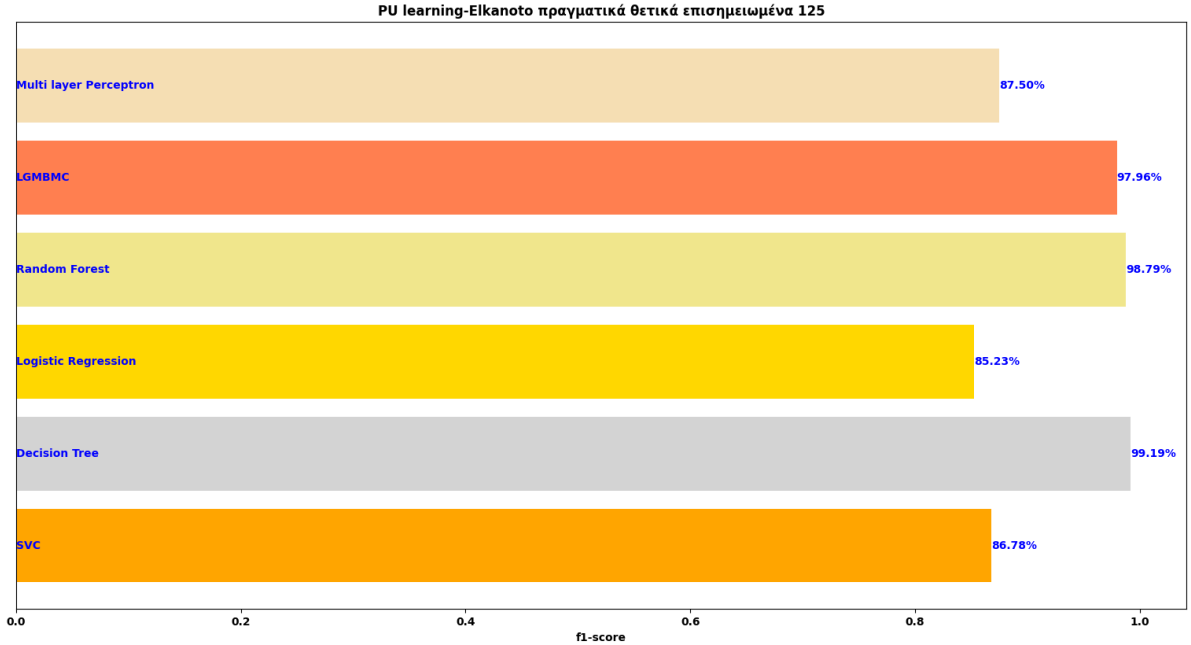
Εικόνα 75: PU learning Elkanoto – ιστόγραμμα f1-score ανά εκτιμητή βάσης στα 40 θετικά επισημειωμένα δείγματα.

Στη παραπάνω εικόνα φαίνεται, ότι ο εκτιμητής βάσης Multi-Layer Perceptron δίνει f1-score 83.87% και ο Logistic Regression 83.93%



Εικόνα 76: PU learning Elkanoto – ιστόγραμμα f1-score ανά εκτιμητή βάσης στα 80 θετικά επισημειωμένα δείγματα.

Στα 80 θετικά επισημειωμένα, η f1-score του Multi-Layer Perceptron είναι πάνω από 85%. Ενώ οι Random Forest, Logistic Regression και SVC είναι πάνω από 80%.

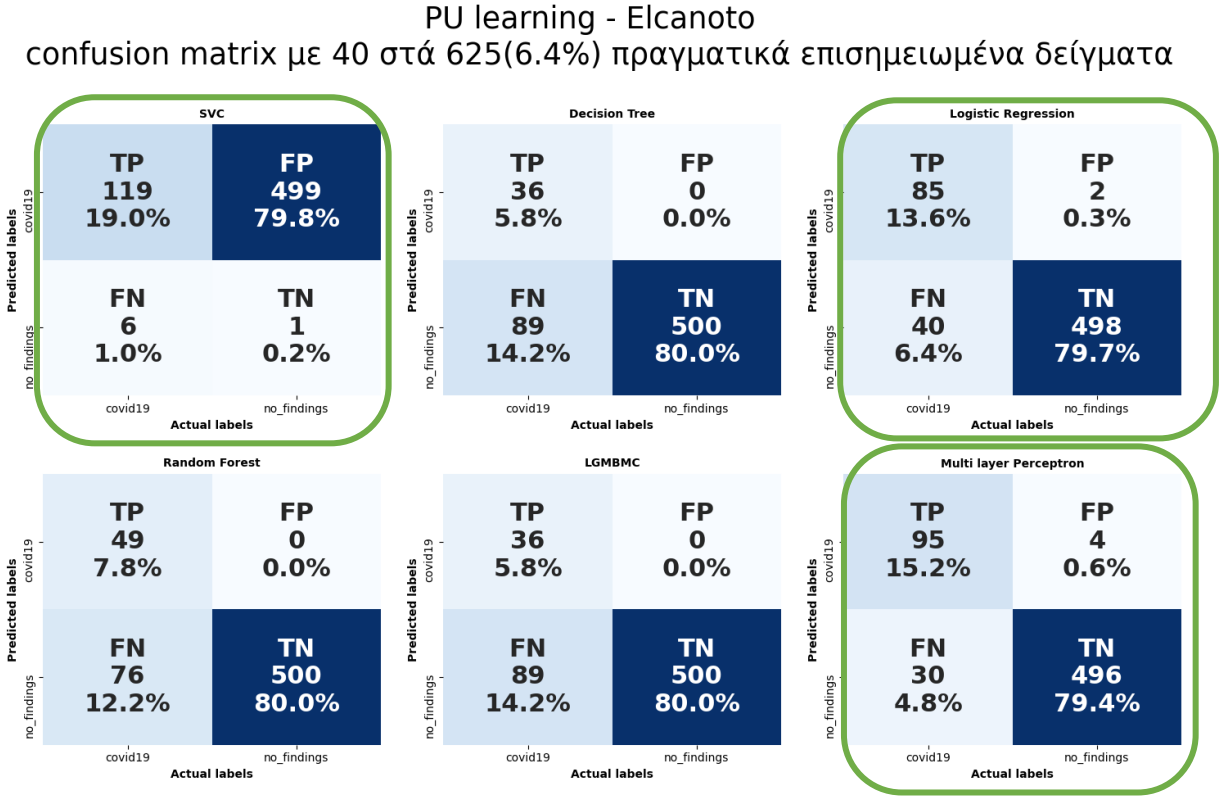


Εικόνα 77: PU learning Elkanoto – f1-score ανά εκτιμητή βάσης στα 125 θετικά επισημειωμένα δείγματα.

Στα 125 θετικά επισημειωμένα έχουμε τιμή f1-score πάνω από 95%, για τους LGMBMC, Random Forest, και Decision Tree. Παρατηρούμε δηλαδή, σε αυτές τις περιπτώσεις, όταν ο αλγόριθμός μας εξαντλήσει όλα τα θετικά

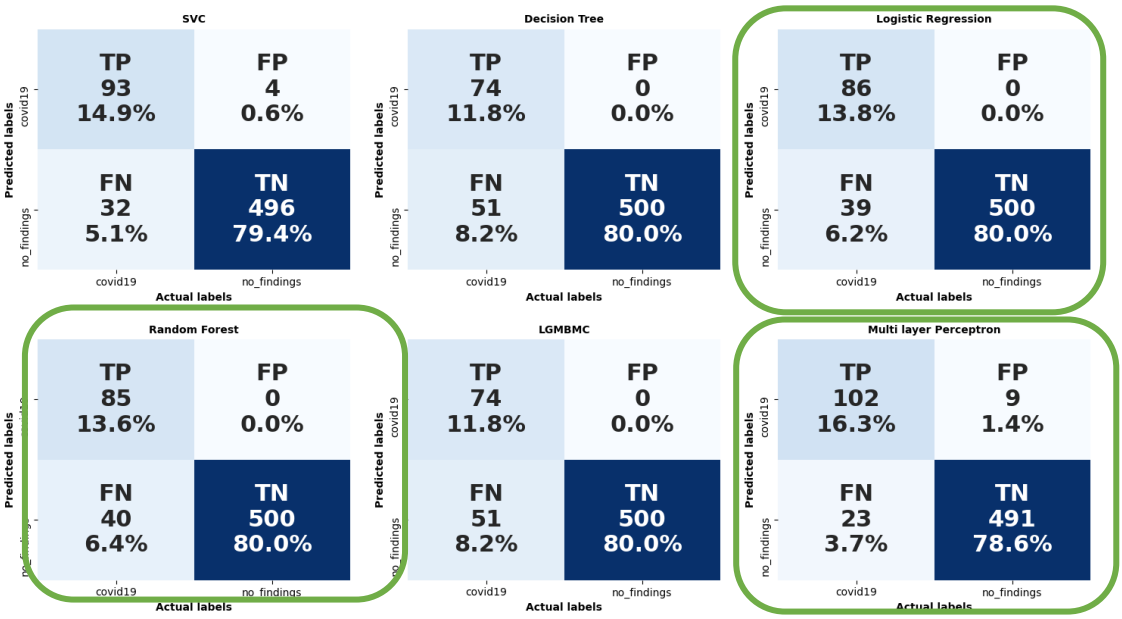
(125), αυτοί μπορούν να βρουν αποτελεσματικότατα ότι όλα τα υπόλοιπα είναι αρνητικά. Στην περίπτωση που θέλουμε να βγάλουμε γενικό συμπέρασμα ανεξάρτητο του μεγέθους του συνόλου των δεδομένων μας, πρέπει να αναζητήσουμε έναν αλγόριθμο που δείχνει σταθερότητα στην εξέλιξη της f1-score και να αποφύγουμε αυτούς που έχουνε στενή σχέση με την αύξηση των θετικών δειγμάτων. Τέτοιοι αλγόριθμοι είναι οι Logistic Regression και Multi-layer perceptron. Αυτοί οι δύο αλγόριθμοι ανεβάζουν γρήγορα την f1-score, στα 35-40 θετικά, σε τιμές πάνω από 80%, οι οποίες, από εκεί και πέρα διατηρούνται, ανεξάρτητα από την αύξηση των θετικών επισημειωμένων. Αυτοί οι αλγόριθμοι, έχουν μεγαλύτερη πιθανότητα να έχουν την ίδια συμπεριφορά σε άλλα σύνολα δεδομένων. Για τον λόγο αυτό το σημείο των 40 θετικά επισημειωμένων θα το μελετήσουμε περισσότερο.

Στα παρακάτω confusion matrix, ραρατηρούμε ότι, με 40 μόνο θετικά επισημειωμένα από τα 625, οι αλγόριθμοι, Multi-layer Perceptron, και Logistic Regression, αναγνωρίζουν σχεδόν όλα τα αρνητικά και 85-95 θετικά από τα 125. Ο SVC υπερτερεί στην αναγνώριση των θετικών αλλά υστερεί χαρακτηριστικά στην αναγνώριση των αρνητικών, και γι' αυτό απορρίπτεται (θεωρεί όλα σχεδόν τα αρνητικά ως θετικά).



Εικόνα 78 PU learning – Elkanoto, confusion matrix με 40 διαθέσιμα θετικά.

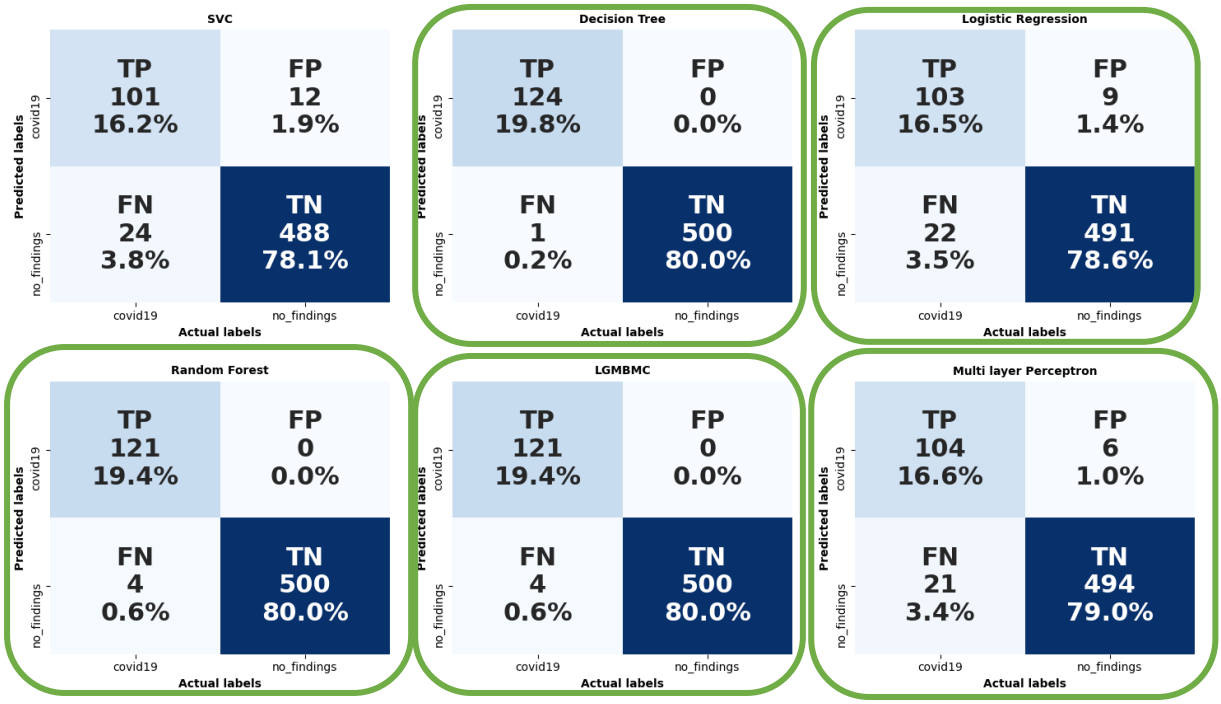
PU learning - Elcanoto
confusion matrix με 80 στά 625(12.8%) πραγματικά επισημειωμένα δείγματα



Εικόνα 79 PU learning – Elkanoto, confusion matrix με 80 διαθέσιμα θετικά.

Στα 80 θετικά επισημειωμένα βλέπουμε ότι όλοι οι αλγόριθμοι έχουν πλησιάσει μεταξύ τους. Οι Logistic Regression, Multi-Layer Perceptron, και Random Forest όμως υπερτερούν.

PU learning - Elcanoto
confusion matrix με 125 στά 625(20.0%) πραγματικά επισημειωμένα δείγματα



Εικόνα 80 PU learning – Elkanoto confusion matrix με 80 διαθέσιμα θετικά.

Οι αλγόριθμοι LGMBMC, Decision Tree, Random Forest ξεπερνούν σε απόδοση τους Logistic Regression και Multi-Layer Perceptron, όμως αυτοί όπως ήδη είπαμε, διατηρούν σταθερή απόδοση, άρα υπάρχουν πολλές πιθανότητες να έχουν την ίδια συμπεριφορά σε άλλα σετ δεδομένων.

Πίνακας 6: Μετρικές PU Learning – ElkaNoto θεωρώντας ότι τα 40, 90, 125 θετικά έχουν ληφθεί υπόψη.

Πραγματικά επισημειωθέντα: 40						
estimator	**set**	**accuracy**	**precision**	**sensitivity**	**specificity**	**f1-score**
SVC	test	0.2821	0.2018	0.8691	0.1357	0.3249
	train	0.3189	0.2307	0.9520	0.1622	0.3672
	Delta	0.0369	0.0288	0.0829	0.0264	0.0424
	STD	0.1243	0.0157	0.1503	0.1919	0.0193
Decision Tree	test	0.8462	0.9167	0.2693	0.9922	0.4094
	train	0.8534	1.0000	0.2679	1.0000	0.4220
	Delta	0.0072	0.0833	0.0014	0.0078	0.0126
	STD	0.0208	0.1179	0.0465	0.0110	0.0393
Logistic Regression	test	0.9263	0.9683	0.6695	0.9918	0.7686
	train	0.9303	0.9747	0.6730	0.9950	0.7929
	Delta	0.0040	0.0065	0.0035	0.0032	0.0242
	STD	0.0401	0.0449	0.2021	0.0116	0.1494
Random Forest	test	0.9279	0.9596	0.6761	0.9922	0.7889
	train	0.8774	1.0000	0.3862	1.0000	0.5552
	Delta	0.0505	0.0404	0.2898	0.0078	0.2337
	STD	0.0079	0.0571	0.0621	0.0111	0.0208
LGMBMC	test	0.8878	0.9872	0.4435	0.9980	0.5937
	train	0.8598	1.0000	0.3002	1.0000	0.4618
	Delta	0.0280	0.0128	0.1433	0.0020	0.1320
	STD	0.0267	0.0181	0.1545	0.0029	0.1601
Multi layer Perceptron	test	0.9455	0.9061	0.8092	0.9777	0.8521
	train	0.9431	0.9159	0.7886	0.9821	0.8461
	Delta	0.0024	0.0099	0.0206	0.0043	0.0059
	STD	0.0163	0.0451	0.0912	0.0128	0.0571
Πραγματικά επισημειωθέντα: 90						
estimator	**set**	**accuracy**	**precision**	**sensitivity**	**specificity**	**f1-score**
SVC	test	0.9263	0.9366	0.6800	0.9880	0.7876
	train	0.9279	0.9437	0.6797	0.9900	0.7895
	Delta	0.0016	0.0071	0.0003	0.0020	0.0020
	STD	0.0099	0.0496	0.0074	0.0099	0.0222
Decision Tree	test	0.9135	0.9543	0.6041	0.9918	0.7392
	train	0.9375	1.0000	0.6887	1.0000	0.8156
	Delta	0.0240	0.0457	0.0845	0.0082	0.0763
	STD	0.0219	0.0383	0.0304	0.0080	0.0264
Logistic Regression	test	0.9391	0.9610	0.7328	0.9918	0.8312
	train	0.9415	0.9731	0.7302	0.9950	0.8335
	Delta	0.0024	0.0121	0.0027	0.0032	0.0022
	STD	0.0216	0.0354	0.0474	0.0078	0.0425
Random Forest	test	0.9696	0.9737	0.8696	0.9940	0.9172
	train	0.9471	1.0000	0.7350	1.0000	0.8470
	Delta	0.0224	0.0263	0.1346	0.0060	0.0702
	STD	0.0082	0.0195	0.0609	0.0049	0.0291
LGMBMC	test	0.9391	1.0000	0.6971	1.0000	0.8192
	train	0.9351	1.0000	0.6763	1.0000	0.8068
	Delta	0.0040	0.0000	0.0208	0.0000	0.0124
	STD	0.0163	0.0000	0.0764	0.0000	0.0516
Multi layer Perceptron	test	0.9471	0.9405	0.7960	0.9861	0.8532
	train	0.9375	0.9274	0.7514	0.9839	0.8278
	Delta	0.0096	0.0132	0.0446	0.0021	0.0254
	STD	0.0196	0.0383	0.1344	0.0112	0.0616
Πραγματικά επισημειωθέντα: 125						
estimator	set	accuracy	precision	sensitivity	specificity	f1-score
SVC	test	0.9391	0.8796	0.8384	0.9711	0.8450
	train	0.9423	0.8844	0.8191	0.9728	0.8504
	Delta	0.0032	0.0048	0.0193	0.0017	0.0055
	STD	0.0177	0.1050	0.1141	0.0261	0.0363
Decision Tree	test	0.9599	0.9112	0.8867	0.9779	0.8976
	train	0.9968	1.0000	0.9842	1.0000	0.9920
	Delta	0.0369	0.0888	0.0975	0.0221	0.0943
	STD	0.0023	0.0246	0.0416	0.0077	0.0124

Logistic Regression	test	0.9407	0.8988	0.8008	0.9781	0.8431
	train	0.9487	0.9053	0.8317	0.9779	0.8668
	Delta	0.0080	0.0064	0.0309	0.0002	0.0237
	STD	0.0255	0.0298	0.1057	0.0052	0.0632
Random Forest	test	0.9728	0.9722	0.8884	0.9942	0.9274
	train	0.9952	1.0000	0.9757	1.0000	0.9877
	Delta	0.0224	0.0278	0.0874	0.0058	0.0603
	STD	0.0023	0.0393	0.0244	0.0082	0.0110
LGMBMC	test	0.9744	0.9803	0.8845	0.9961	0.9298
	train	0.9952	1.0000	0.9755	1.0000	0.9875
	Delta	0.0208	0.0197	0.0909	0.0039	0.0577
	STD	0.0060	0.0140	0.0341	0.0028	0.0242
Multi layer Perceptron	test	0.9471	0.9091	0.8273	0.9775	0.8661
	train	0.9543	0.9144	0.8532	0.9799	0.8827
	Delta	0.0072	0.0053	0.0260	0.0024	0.0166
	STD	0.0208	0.0460	0.0260	0.0151	0.0346

Συμπεράσματα Elkanoto:

1. Με ένα αριθμό 90 πραγματικά θετικά επισημειωμένων δειγμάτων στα συνολικά 625 δείγματα, σχεδόν όλα τα μοντέλα που στηρίζονται στον Elkanoto δίνουν μετρικές πάνω από 80%.

2. Οι τιμές της f1-score, διαμορφώνονται αναλογικά με τα διαθέσιμα θετικά επισημειωμένα για τους Decision Tree, Random Forest και LGMBC. Όταν ο αριθμός των θετικά επισημειωμένων είναι 125, ο LGMBMC έχει τιμές: accuracy 0.9744, precision 0.9803, sensitivity 0.8845, specificity 0.9961, f1-score 0.9298. Γι' αυτό, στην περίπτωση που έχουμε μεγάλο αριθμό θετικά επισημειωμένων προτιμούμε τον LGMBMC.

3. Στο επίπεδο των 40 θετικών επισημειωμένων στον Random Forest έχουμε μεγάλη διαφορά στην μετρική sensitivity και f1-score μεταξύ εκπαίδευσης και δοκιμών, υπέρ των δοκιμών. Αυτό είναι ένδειξη για υποπροσαρμογή σε αυτό το επίπεδο.

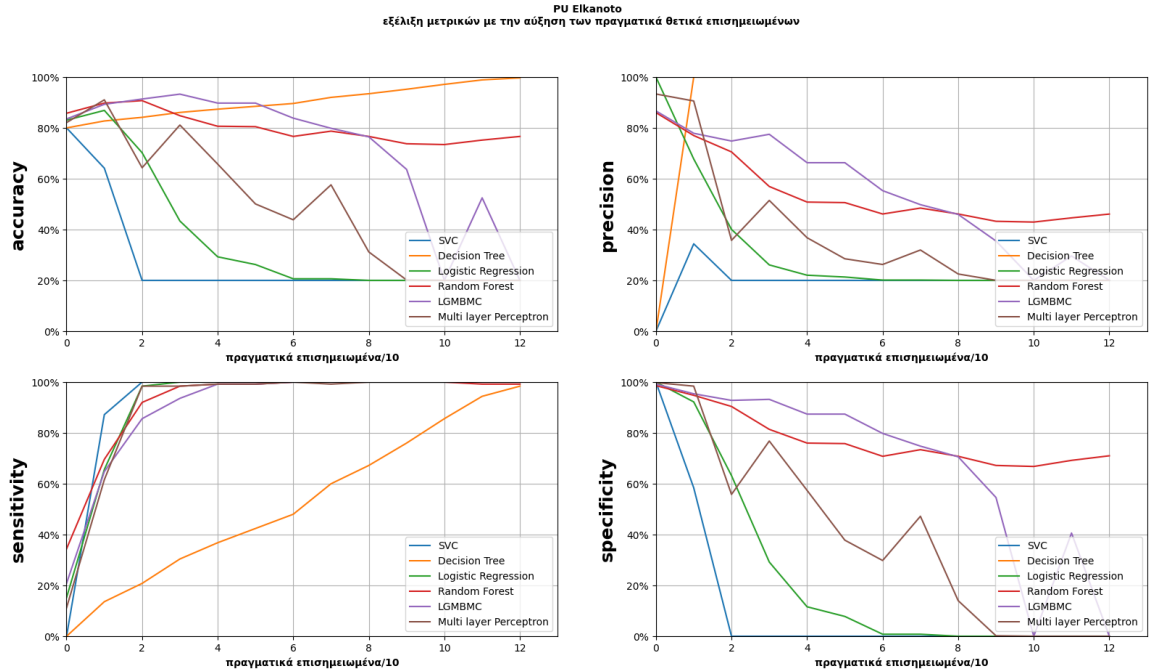
4. Όταν τα θετικά επισημειωμένα, τα οποία έχουμε στην διάθεσή μας, είναι κάτω από 20, κανένας αλγόριθμος δεν δίνει f1-score πάνω από 60%, άρα δεν υπάρχει δυνατότητα να εκπαιδεύσουμε σε αυτά τα επίπεδα αξιόπιστο μοντέλο.

5. Ο SVC έχει υψηλή sensitivity μετά τα 40 θετικά επισημειωμένα, πράγμα το οποίο σημαίνει, ότι αναγνωρίζει με μεγάλη ακρίβεια τα θετικά, σημαντικό στοιχείο για τον περιορισμό της διάδοσης του COVID-19. Όμως πρέπει να έχουμε τουλάχιστον 90 θετικά επισημειωμένα για να έχουμε ταυτόχρονα μια ικανοποιητική τιμή για την specificity.

6. Γενικά, η specificity έχει καλύτερες τιμές από την sensitivity, με όλους τους αλγόριθμους βάσης, το οποίο σημαίνει ότι τα αρνητικά αναγνωρίζονται με μεγαλύτερη ακρίβεια.

7. Τέλος οι Logistic Regression και Multilayer Percepton, φαίνεται ότι όταν διαθέτουν πάνω από 40 θετικά δείγματα, δεν επηρεάζονται από τον αριθμό των θετικά επισημειωμένων, στοιχείο που δείχνει ότι ο αριθμός 40 πιθανόν να έχει εφαρμογή και σε άλλα σύνολα δεδομένων.

Αποτελέσματα weighted Elkanoto



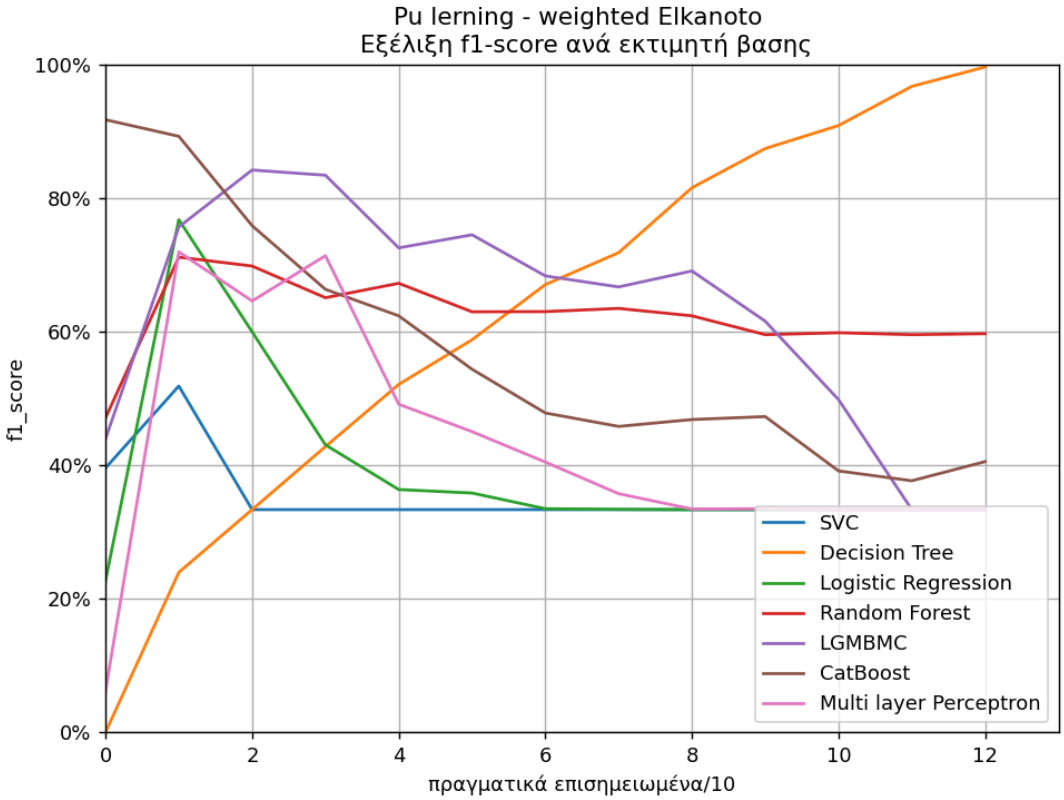
Εικόνα 81 : Pu learning – εξέλιξη μετρικών με την αύξηση των θετικά επισημειωμένων.

Από τα παραπάνω διαγράμματα διαπιστώνεται ότι, εκτός από την sensitivity, είτε δεν υπάρχει καλή συνεργασία μεταξύ του αλγορίθμου weighted elkanoto και των αλγορίθμων βάσης που επελέγησαν, είτε τα δεδομένα που χρησιμοποιήθηκαν είναι ανεπαρκή για να επιτρέψουν να καταγραφεί η λανθάνουσα σχέση, η οποία τα συνδέει με τις επισημειώσεις τους. Όσον αφορά τους αλγορίθμους βάσης, ο Decision Tree ακολουθεί σε όλες τις μετρικές, πλην specificity, μία πορεία ανάλογη του αριθμού των διαθεσίμων θετικά επισημειωμένων.

Όσον αφορά την sensitivity, η οποία αντιπροσωπεύει την ικανότητα αναγνώρισης των πραγματικά θετικών, και γι' αυτό τον λόγο, η εξέλιξη της είναι σημαντική καθώς το να αναγνωρίζουμε με ακρίβεια τους θετικούς, μπορεί να συντελέσει στον περιορισμό της εξάπλωσης της νόσου, παρατηρούμε ότι: α) όλοι οι αλγόριθμοι βάσης, όταν έχουν στην διάθεσή τους πάνω από τα 20 θετικά επισημειωμένα, έχουν τιμή sensitivity πάνω από 80%. β) μετά τα 40 πραγματικά θετικά επισημειωμένα η τιμή της sensitivity είναι περίπου

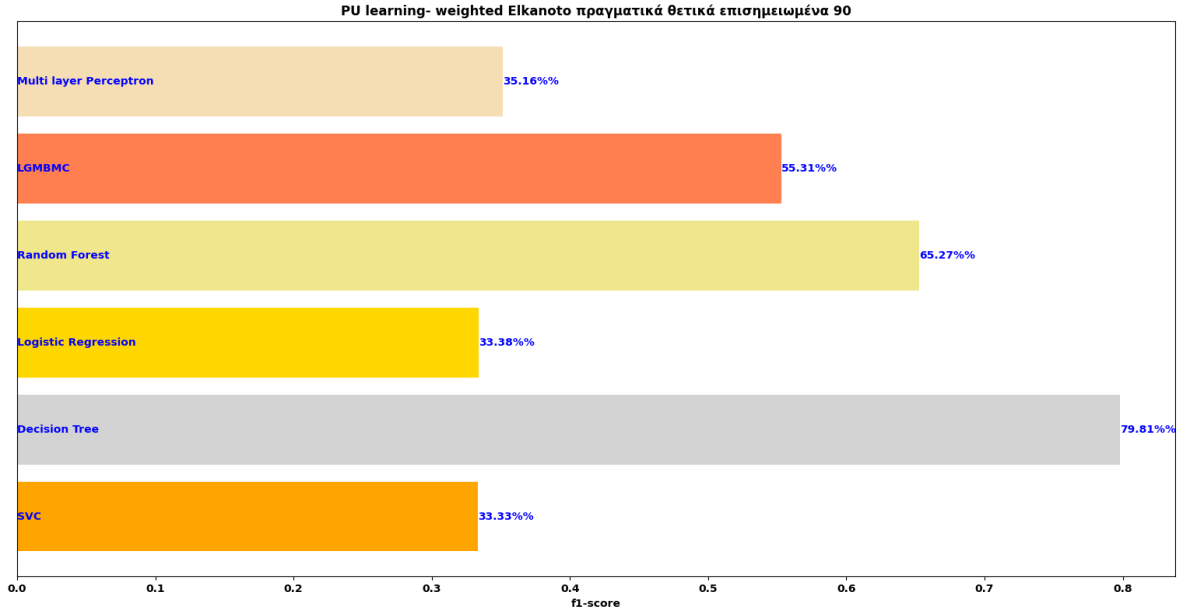
100%, το οποίο σημαίνει, ότι αναγνωρίζει απόλυτα όλους τους θετικούς. Επιπλέον η γραφική αναπαράσταση δείχνει ότι, από τα 40 θετικά επισημειωμένα και πάνω, η sensitivity είναι ανεξάρτητη της αναλογίας θετικών προς σύνολο δειγμάτων. Αυτό σημαίνει ο αριθμός αυτός (40) πιθανόν να ισχύει και για άλλα σύνολα δεδομένων.

Από το παραπάνω διάγραμμα διαπιστώνουμε, ότι κανένας ουσιαστικά από τους αλγορίθμους που επιλέξαμε δεν μπορεί να μας δώσει γενική λύση πλην του Decision Tree. Αυτός μας δίνει πολύ καλά αποτελέσματα, όταν έχει στην διάθεσή του, πάνω από 100 θετικά επισημειωμένα. Στα 125 θετικά επισημειωμένα, δηλαδή εάν τον έχουμε τροφοδοτήσει με το 100% των πραγματικών θετικών, τότε φαίνεται ότι ο αλγόριθμος Decision Tree, θα αντιληφθεί ότι όλα τα υπόλοιπα είναι αρνητικά. Με λιγότερα από 60 θετικά επισημειωμένα, η f1-score όλων, πλην Decision Tree, δεν έχει πάρει μία συγκεκριμένη κατεύθυνση. Αυτό σημαίνει ότι τελικά, είτε το σύνολο που έχουμε στην διάθεσή μας, το οποίο αποτελείται από 125 θετικά και 500 αρνητικά δεν επαρκεί για να βγάλουμε συμπεράσματα, είτε οι διάφοροι αλγόριθμοι βάσης που επιλέξαμε πλην του Decision Tree δεν συνεργάζονται καλά με τον weighted Elksnoto, τουλάχιστον σε αυτά τα επίπεδα αριθμού δεδομένων. Το συμπέρασμα αυτό έχει διαπιστωθεί και προηγουμένως όταν εξετάσθηκαν οι υπόλοιπες μετρικές.

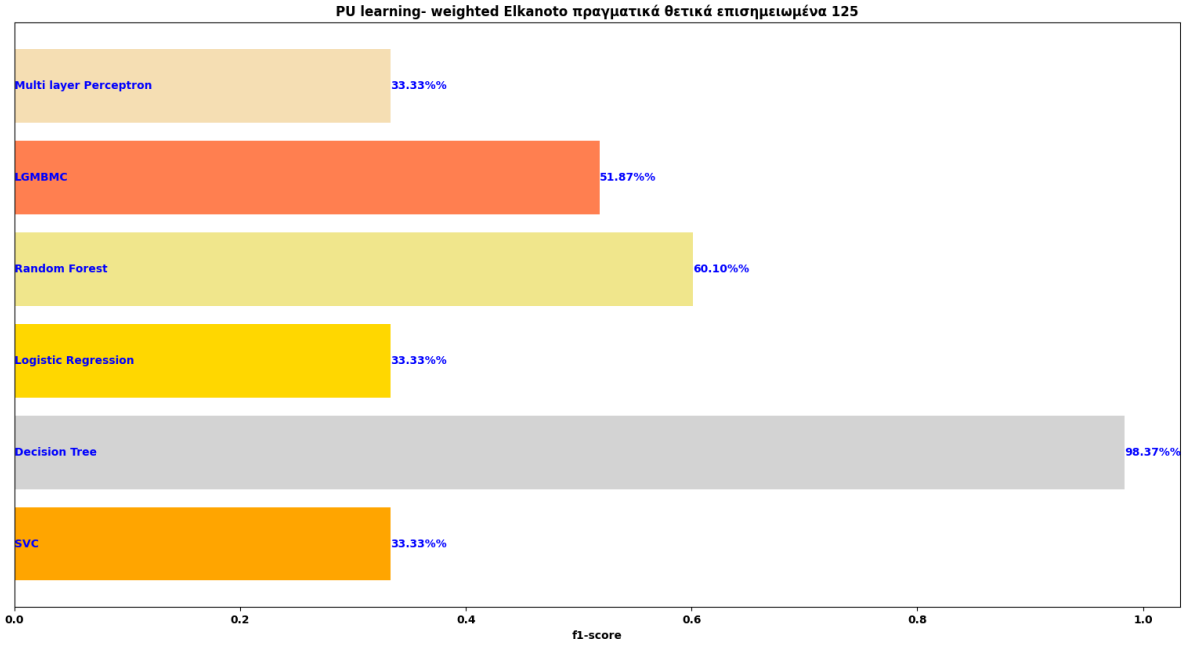


Εικόνα 82 :Pu learning – weighted Elkanoto, εξέλιξη f1-score ανά εκτιμητή βάσης.

Όσον αφορά τα σημεία ενδιαφέροντος είναι τα 90 και τα 125. Στα 90 θετικά επισημειωμένα, έχει οριζοντιοποιηθεί η γραμμή του Random Forest και το 125 είναι το σύνολο των θετικά επισημειωμένων που διαθέτουμε.

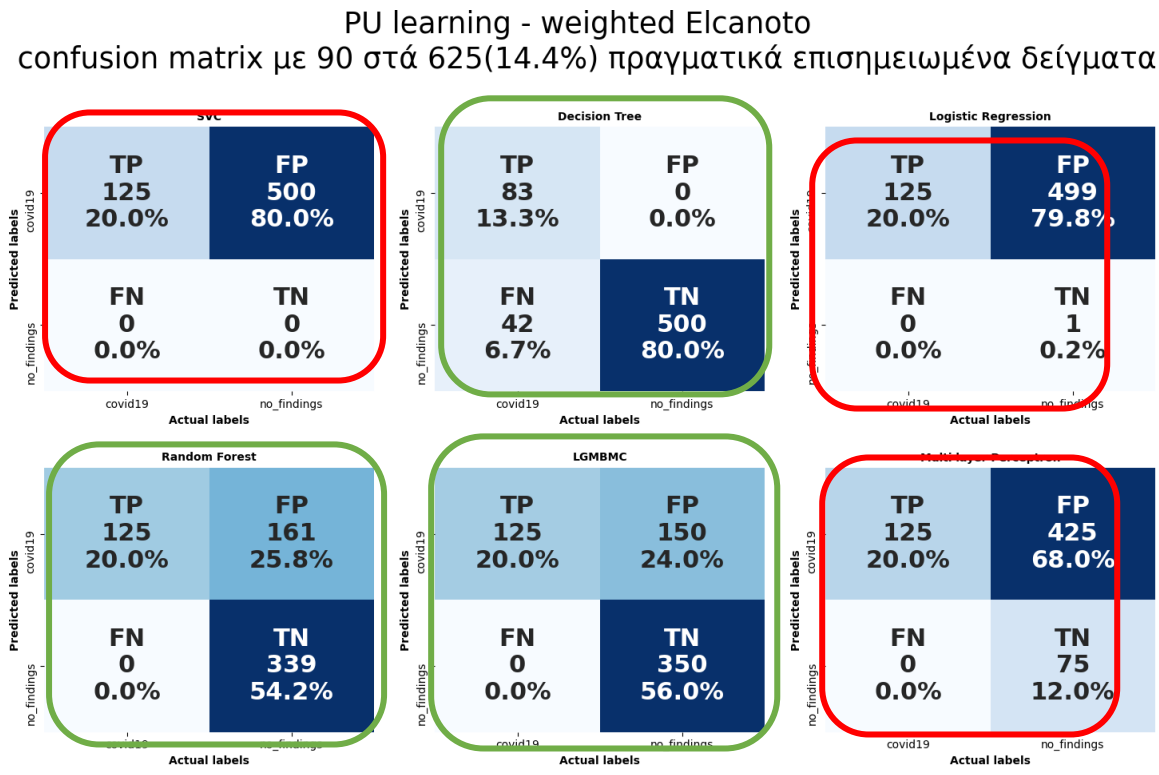


Εικόνα 83: PU learning weighted Elkanoto,, ιστόγραμμα απόδοσης αλγορίθμων βάσης στα 90 θετικά επισημειωμένα.



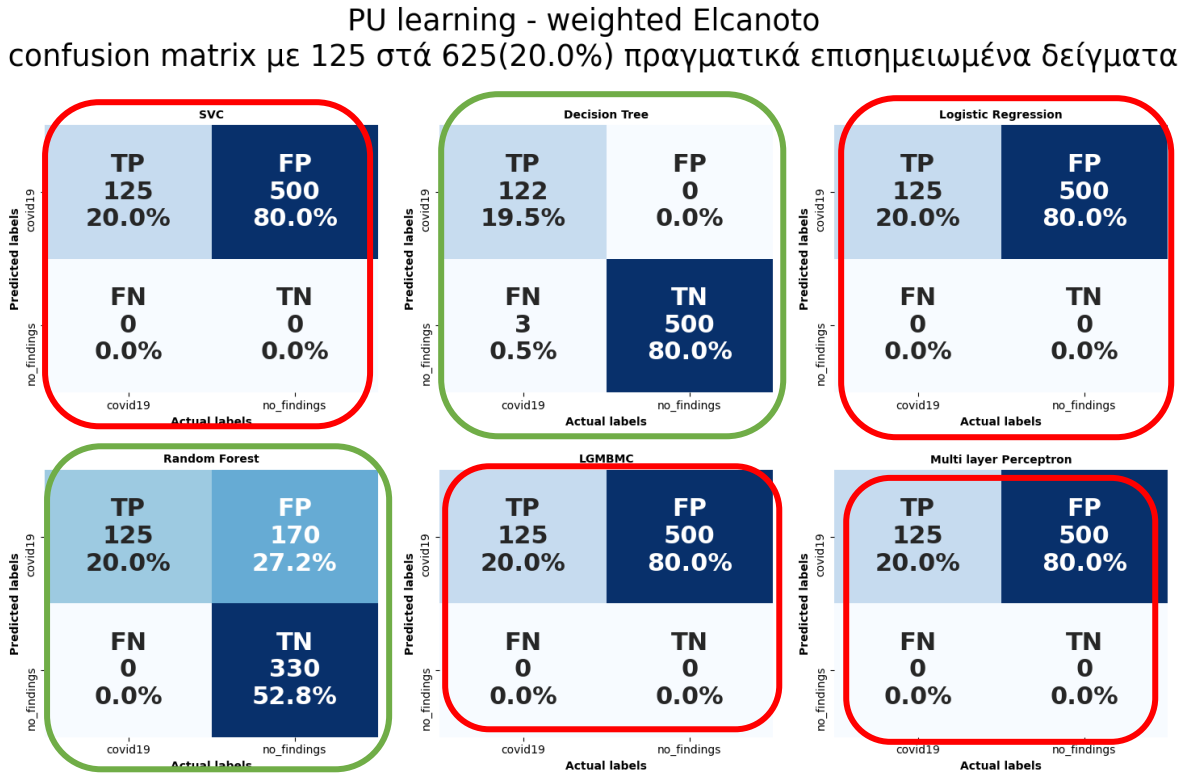
Εικόνα 84: PU learning weighted Elkanoto, ιστόγραμμα απόδοσης αλγορίθμων βάσης στα 125 θετικά επισημειωμένα.

Στα παραπάνω ιστογράμματα, εκτός από την αυξανόμενη αποτελεσματικότητα του Decision Tree, με την αύξηση των θετικών επισημειωμένων, παρατηρούμε την σχετική σταθερότητα του Random Forest στα δύο ιστογράμματα. Αυτό μας δίνει μία ένδειξη ότι πάνω από τα 90 πραγματικά επισημειωμένα θα έχουμε μια σταθερή απόδοση στον Random Forest, πάνω από 60% για f1-score, που πιθανόν να ισχύει και για μεγαλύτερα σύνολα δεδομένων.



Εικόνα 85: PU learning weighted Elkanoto, confusion matrix 90 θετικά επισημειωμένα.

Εδώ βλέπουμε ότι, ο Random Forest βρήκε στα 90 θετικά επισημειωμένα, όλα τα θετικά και 339 από τα 500 αρνητικά. **Το Decision Tree, παρότι έχει καλύτερο f1-score βρίσκει μόνο 83 από τα 125 θετικά**, βρίσκει όμως όλα τα αρνητικά. Αυτό σημαίνει ότι η επιλογή του Random Forest θα συμβάλλει περισσότερο στην μείωση της εξάπλωσης της νόσου. Ο SVC από την πλευρά του, βρίσκει και αυτός, σε αυτό το σημείο, όλα τα θετικά, αλλά κανένα αρνητικό, δηλαδή τα θεωρεί όλα θετικά.



Εικόνα 86: PU learning weighted Elkanoto, confusion matrix 90 θετικά επισημειωμένα.

Στα 125 (όλα) θετικά επισημειωμένα, ο Random Forest **βρίσκει όλα τα θετικά** και 330 από τα 500 αρνητικά και έχει f1-score 60.10%. Ο Decision Tree βρίσκει όλα τα αρνητικά και 122 από τα θετικά με f1-score 98.37 %. Το Random Forest παρότι έχει πολύ μικρότερη f1-score, μπορούμε να το θωρήσουμε ως καλύτερο, αφού εάν διαθέτουμε 90 θετικά επισημειωμένα και πάνω, μπορεί να βρει όλα τα υπόλοιπα θετικά. Οι SVC, Logistic Regression, Multi–Layer Perceptron και LGMBMC, βρίσκουν όλα τα θετικά αλλά σχεδόν κανένα αρνητικό, άρα απορρίπτονται σε αυτά τα επίπεδα.

Το σημαντικότερο στοιχείο του Random Forest είναι ότι διατηρεί την απόδοσή του μέχρι εξαντλήσεως των θετικών, τα οποία έχουμε στην διάθεσή μας. Δηλαδή, από τον αριθμό 90 και πάνω, η τιμή f1-score του Random Forest δεν ακολουθεί με την αύξηση του αριθμού των θετικά επισημειωμένων.

.

Πίνακας 7: Μετρικές PU Learning – Weighted Elkanoto, με συμμετοχή 50, 90, 125 θετικών.

Πραγματικά επισημειωμένα: 50						
estimator	set	accuracy	precision	sensitivity	specificity	f1-score
SVC	test	0.2212	0.2018	0.9773	0.0321	0.3344
	train	0.2019	0.2006	1.0000	0.0020	0.3342
	Delta	0.0192	0.0012	0.0227	0.0301	0.0002
	STD	0.0257	0.0071	0.0321	0.0298	0.0096
Decision Tree	test	0.8397	0.6042	0.2148	0.9941	0.3158
	train	0.8494	0.6667	0.2445	1.0000	0.3575
	Delta	0.0096	0.0625	0.0297	0.0059	0.0417
	STD	0.0230	0.4340	0.1522	0.0084	0.2233
Logistic Regression	test	0.7099	0.4120	0.9919	0.6386	0.5812
	train	0.3854	0.2484	1.0000	0.2319	0.3972
	Delta	0.3245	0.1636	0.0081	0.4067	0.1840
Random Forest	test	0.8926	0.6639	0.9767	0.8722	0.7857
	train	0.8269	0.5497	0.9839	0.7875	0.7013
	Delta	0.0657	0.1141	0.0072	0.0847	0.0844
	STD	0.0276	0.0852	0.0329	0.0406	0.0511
LGMBMC	test	0.9247	0.8891	0.7443	0.9746	0.7941
	train	0.8934	0.6589	0.9759	0.8725	0.7865
	Delta	0.0312	0.2302	0.2316	0.1021	0.0076
	STD	0.0181	0.0904	0.1387	0.0231	0.0524
Multi layer Perceptron	test	0.8125	0.5658	0.9690	0.7759	0.6955
	train	0.5785	0.3522	0.9958	0.4724	0.5108
	Delta	0.2340	0.2137	0.0268	0.3035	0.1847
	STD	0.1093	0.1799	0.0274	0.1364	0.1391
Πραγματικά επισημειωθέντα: 90						
estimator	**set**	**accuracy**	**precision**	**sensitivity**	**specificity**	**f1-score**
SVC	test	0.2019	0.2007	1.0000	0.0021	0.3337
	train	0.2003	0.2003	1.0000	0.0000	0.3336
	Delta	0.0016	0.0004	0.0000	0.0021	0.0001
	STD	0.0245	0.0228	0.0000	0.0029	0.0314
Decision Tree	test	0.8942	0.8908	0.5532	0.9770	0.6769
	train	0.9351	1.0000	0.6762	1.0000	0.8065
	Delta	0.0409	0.1092	0.1230	0.0230	0.1296
	STD	0.0283	0.0783	0.0715	0.0180	0.0433
Logistic Regression	test	0.2804	0.2253	1.0000	0.1063	0.3612
	train	0.2067	0.2016	1.0000	0.0080	0.3348
	Delta	0.0737	0.0237	0.0000	0.0984	0.0264
	STD	0.1402	0.0781	0.0000	0.1350	0.1031
Random Forest	test	0.7532	0.4495	1.0000	0.6913	0.6189
	train	0.7596	0.4545	1.0000	0.6991	0.6249
	Delta	0.0064	0.0049	0.0000	0.0078	0.0061
	STD	0.0365	0.0462	0.0000	0.0445	0.0441
LGMBMC	test	0.9022	0.6803	0.9596	0.8890	0.7927
	train	0.8117	0.5176	1.0000	0.7654	0.6799
	Delta	0.0905	0.1627	0.0404	0.1235	0.1129
	STD	0.0252	0.0925	0.0113	0.0297	0.0616
Multi layer Perceptron	test	0.3942	0.2536	0.9922	0.2476	0.4014
	train	0.2115	0.2013	1.0000	0.0160	0.3352
	Delta	0.1827	0.0522	0.0078	0.2316	0.0662
	STD	0.1319	0.0474	0.0110	0.1590	0.0595
Πραγματικά επισημειωθέντα: 125						
estimator	**set**	**accuracy**	**precision**	**sensitivity**	**specificity**	**f1-score**
SVC	test	0.2003	0.2003	1.0000	0.0000	0.3333
	train	0.2003	0.2003	1.0000	0.0000	0.3337
	Delta	0.0000	0.0000	0.0000	0.0000	0.0003
	STD	0.0194	0.0194	0.0000	0.0000	0.0272
Decision Tree	test	0.9647	0.9176	0.9048	0.9801	0.9099
	train	0.9936	1.0000	0.9681	1.0000	0.9838
	Delta	0.0288	0.0824	0.0633	0.0199	0.0739
	STD	0.0023	0.0451	0.0280	0.0099	0.0150
Logistic Regression	test	0.2019	0.2006	1.0000	0.0019	0.3336
	train	0.2003	0.2003	1.0000	0.0000	0.3336
	Delta	0.0016	0.0003	0.0000	0.0019	0.0000
	STD	0.0196	0.0213	0.0000	0.0027	0.0296
Random Forest	test	0.5833	0.3245	1.0000	0.4772	0.4898
	train	0.7003	0.4028	1.0000	0.6258	0.5729
	Delta	0.1170	0.0783	0.0000	0.1486	0.0831
	STD	0.0441	0.0138	0.0000	0.0682	0.0156
LGMBMC	test	0.2484	0.2141	1.0000	0.0637	0.3498
	train	0.2396	0.2082	1.0000	0.0478	0.3447

	Delta	0.0088	0.0059	0.0000	0.0159	0.0051
	STD	0.0998	0.0514	0.0000	0.0901	0.0678
Multi-layer Perceptron	test	0.1987	0.1987	1.0000	0.0000	0.3313
	train	0.1995	0.1989	1.0000	0.0010	0.3317
	Delta	0.0008	0.0002	0.0000	0.0010	0.0004
	STD	0.0159	0.0159	0.0000	0.0000	0.0219

Συμπεράσματα Weighted Elkanoto:

1. Η τιμή της f1-score του Decision Tree είναι ανάλογη του αριθμού των θετικών επισημειωμένων. Όταν φτάσουμε τα 125, δηλαδή εξαντλήσουμε τον αριθμό των θετικών, τότε έχουμε πετύχει accuracy 0.9647, precision 0.9176, sensitivity 0.9048, specificity 0.9801 και f1-score 0.9099.

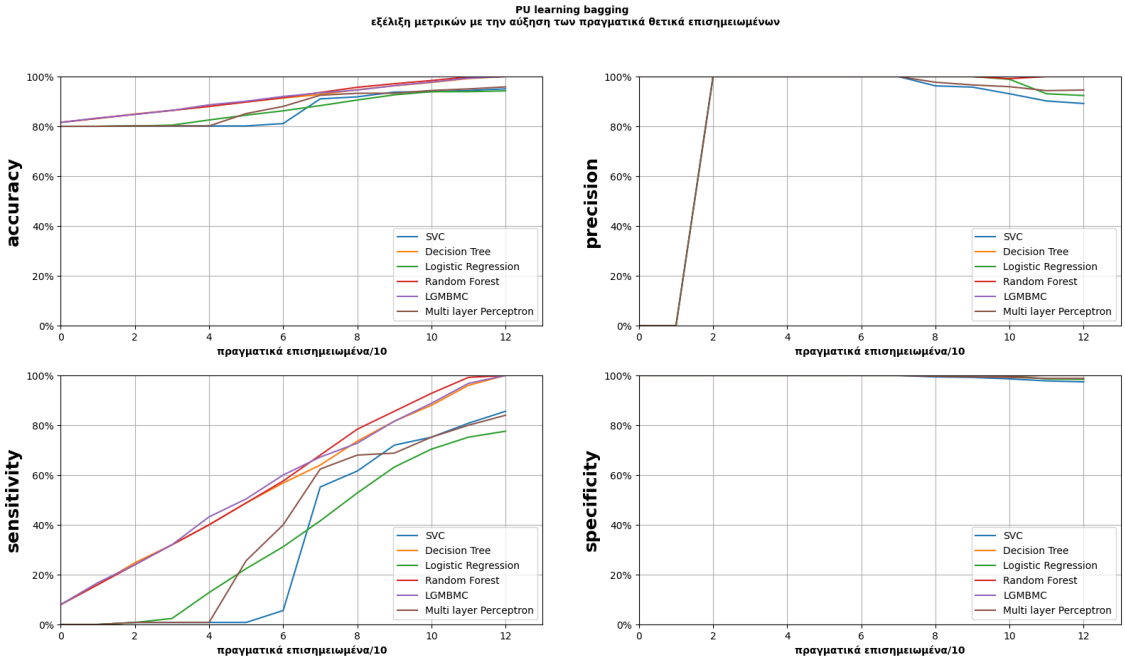
2. Ο Random Forest μετά τα 90 θετικά, παρότι η f1-score είναι κοντά στο 50%, μας δίνει sensitivity 100%, που σημαίνει ότι αναγνωρίζει όλα τα θετικά. Αυτό είναι σημαντικό επειδή α) για την καταπολέμηση της εξάπλωσης της νόσου σημαντικό είναι να μπορούμε να αναγνωρίζουμε τα θετικά. β) πάνω από τον αριθμό 90 επισημειωμένων η απόδοση του Random Forest είναι σταθερή και ανεξάρτητη του αριθμού των επισημειωμένων θετικών.

3. Ο SVC έχει sensitivity 100% και specificity 0%, σε όλες τις περιπτώσεις δηλαδή προβλέπει ότι όλα είναι θετικά και ως εκ τούτου απορρίπτεται. Για τον ίδιο λόγο απορρίπτονται, οι Logistic Regression, Multi–Layer Perceptron και LGMBMC, βρίσκουν όλα τα θετικά αλλά σχεδόν κανένα αρνητικό. Δηλαδή καταλήγουμε στο Decision Tree με πάνω από 100 επισημειωμένα (τα 90 είναι οριακά ανεπαρκή).

4. Γενικά το sensitivity υπερτερεί του specificity, το οποίο σημαίνει ότι τα θετικά αποκαλύπτονται με μεγαλύτερη ακρίβεια.

5. Όλοι οι αλγόριθμοι που δοκιμάστηκαν πλην του Decision Tree είτε δεν συνεργάζονται ικανοποιητικά με τον weighted Elkanoto είτε τα δεδομένα μας είναι ποσοτικά ανεπαρκή για να δημιουργηθεί ένα αποτελεσματικό μοντέλο προβλέψεων.

Αποτελέσματα bagging



Εικόνα 87: PU learning – bagging εξέλιξη μετρικών

Από τα παραπάνω διαγράμματα βγάζουμε τα εξής συμπεράσματα:

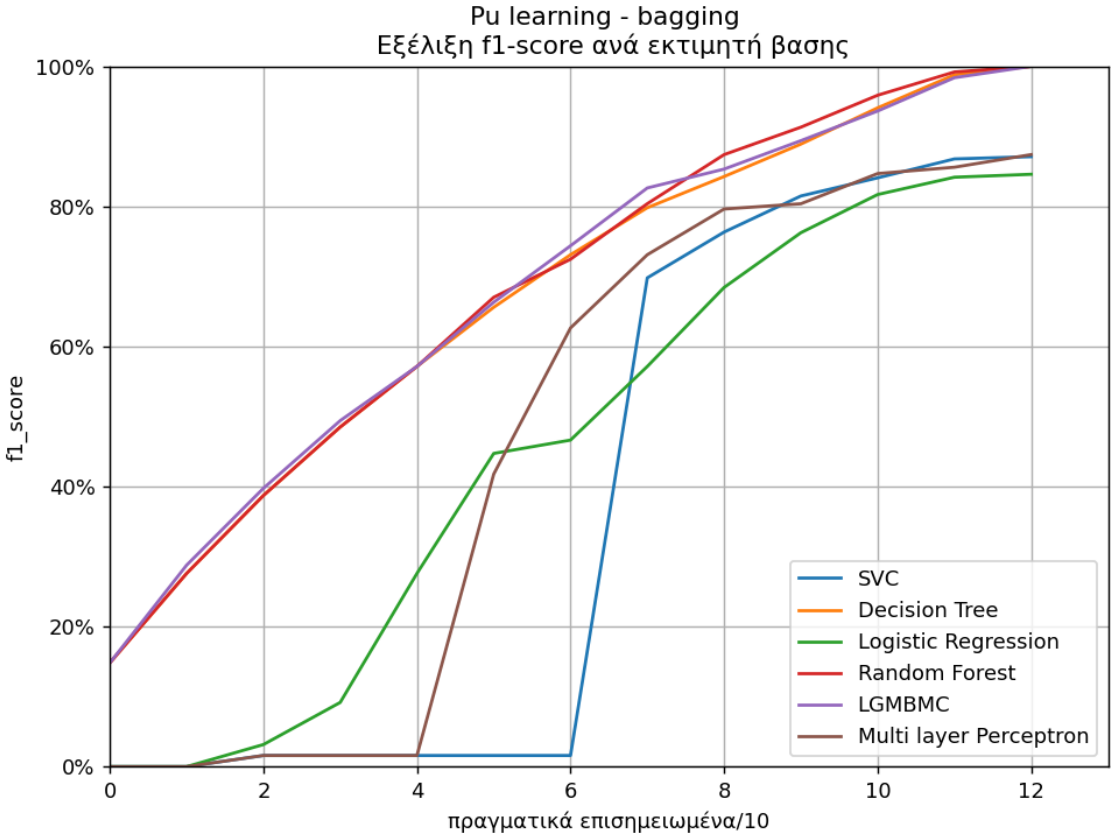
-**Η accuracy** είναι σταθερά πάνω από 80%, όταν έχουμε περισσότερα από 20 θετικά επισημειωμένα μέσα στα 625 συνολικά δείγματα. Φαίνεται ότι οι Random Forest, LGMBMC, και Decision Tree υπερέχουν ελαφρώς έναντι των Logistic Regression, Multi- Layer Perceptron και SVC σε όλα τα επίπεδα.

-**Η precision** απαντά στο ερώτημα «ποιο ποσοστό των θετικών ταυτοποιήσεων ήταν πραγματικά σωστό;». Μια βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινό θετικό (TP). Όμως στην παρούσα περίπτωση είναι εν μέρει αληθές. Το 100% οφείλεται και στον τρόπο που υπολογίζεται το precision. Όταν ο παρονομαστής (FP+TP) είναι 0, τότε το αποτέλεσμα συμβατικά είναι 1, παρότι δεν είχαμε ούτε ένα (1) True Positive. Το ίδιο συμβαίνει και όταν μόνο του το FP είναι με 0.

-**Η sensitivity**, η οποία είναι η ικανότητα του τεστ να εντοπίζει σωστά άτομα με τη νόσο (πραγματικό θετικό ποσοστό) **είναι ανάλογη με την αύξηση αυτών που θεωρούμε θετικά επισημειωμένα**. Μία βαθμολογία 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά θετικό. Η εξέλιξη της sensitivity είναι σημαντική καθόσον το να αναγνωρίσουμε με ακρίβεια τους θετικούς μπορεί να συντελέσει στον περιορισμό της εξάπλωσης της νόσου. Οι Random Forest, LGMBMC, και Decision Tree υπερέχουν έναντι

των Logistic Regression, Multi- Layer Perceptron και SVC σε όλα τα επίπεδα κυρίως όμως στα κάτω από 60 θετικά επισημειωμένα.

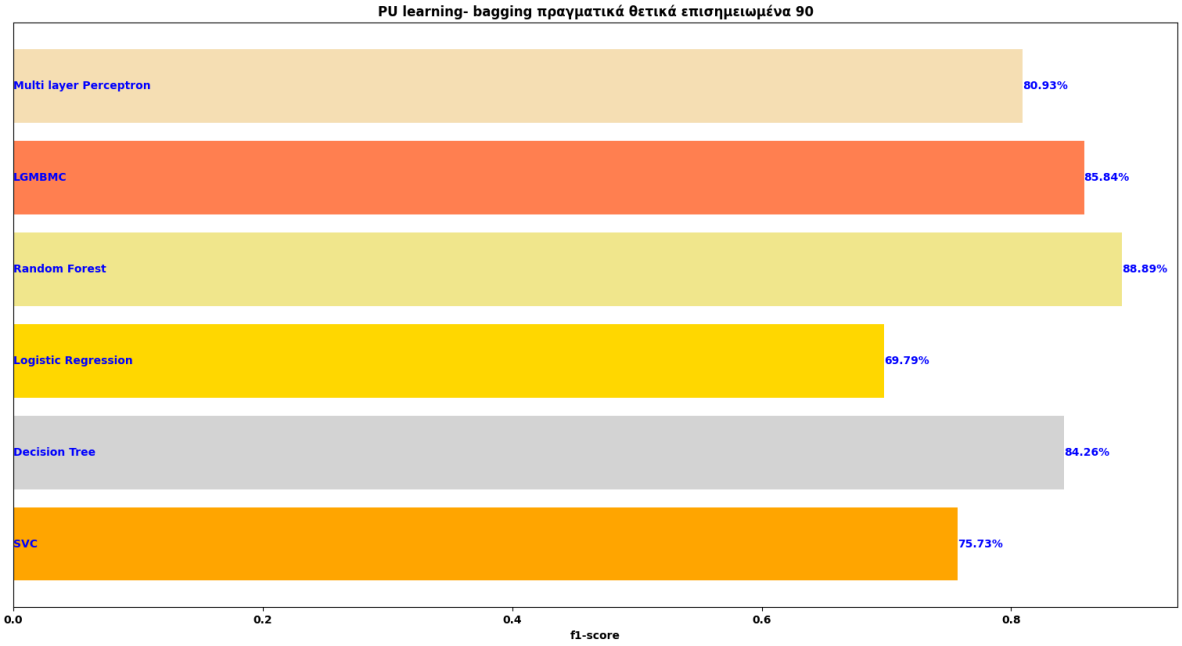
-**Η specificity**, δηλαδή η ικανότητα να ανακαλύπτει τους πραγματικά αρνητικούς (το ποσοστό των πραγματικά αρνητικών που προσδιορίστηκε σωστά) από μικρό αριθμό θετικά επισημειωμένων είναι σε υψηλά επίπεδα, πράγμα που εν μέρει δικαιολογείται από τον μεγάλο αριθμό των υπαρχόντων αρνητικών. Μία τιμή 100% σημαίνει ότι το μοντέλο μας δεν έχασε κανένα αληθινά αρνητικό.



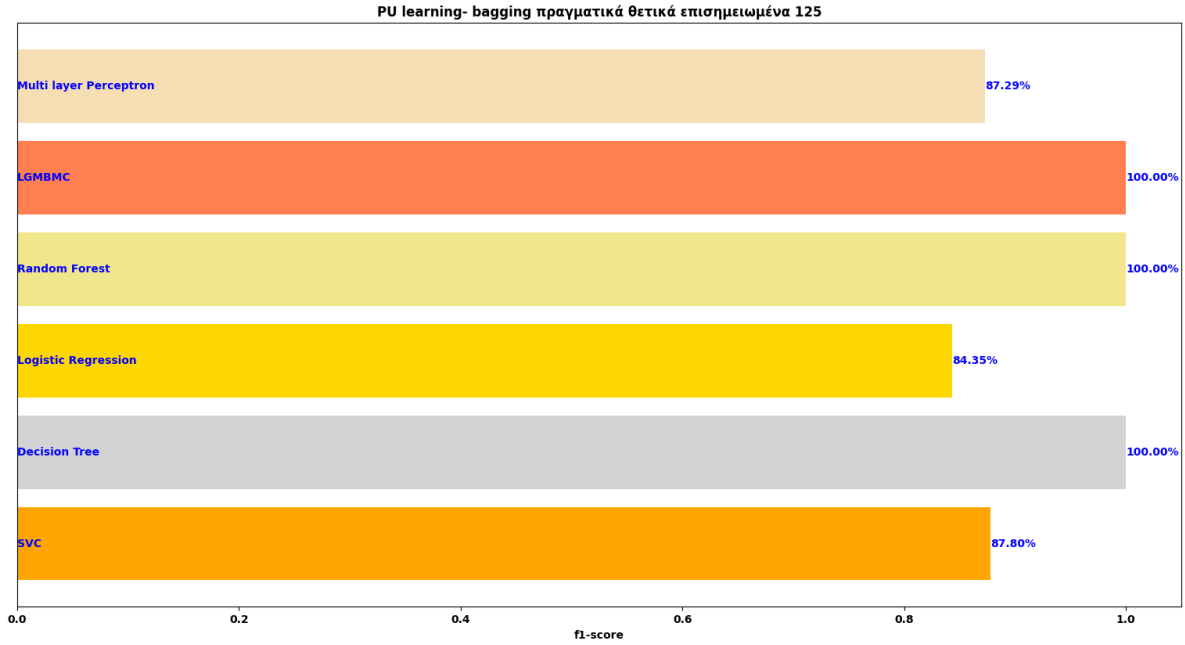
Εικόνα 88:Pu learning – bagging εξέλιξη f1-score ανά εκτιμητή βάσης.

Βλέπουμε ότι η f1-score των Decision Tree, Random Forest και LGMBMC είναι ανάλογη, με την αύξηση των θετικών δειγμάτων. Σε γενικές γραμμές το ίδιο ισχύει και για τους υπόλοιπους αλγόριθμους βάσης που εξετάζουμε. Σαν σημεία ενδιαφέροντος για περαιτέρω μελέτη είναι τα σημεία 80, όπου οι Multi-layer Perceptron, Logistic Regression, και SVC πλησιάζουν το 80% και 125 όπου εξαντλούνται τα διαθέσιμα θετικά.

Από το παραπάνω διάγραμμα επίσης προκύπτει ότι όλοι οι αλγόριθμοι όταν έχουν στην διάθεσή τους πάνω από 100 θετικά επισημειωμένα, τότε η f1-score τους περνάει το 90%.



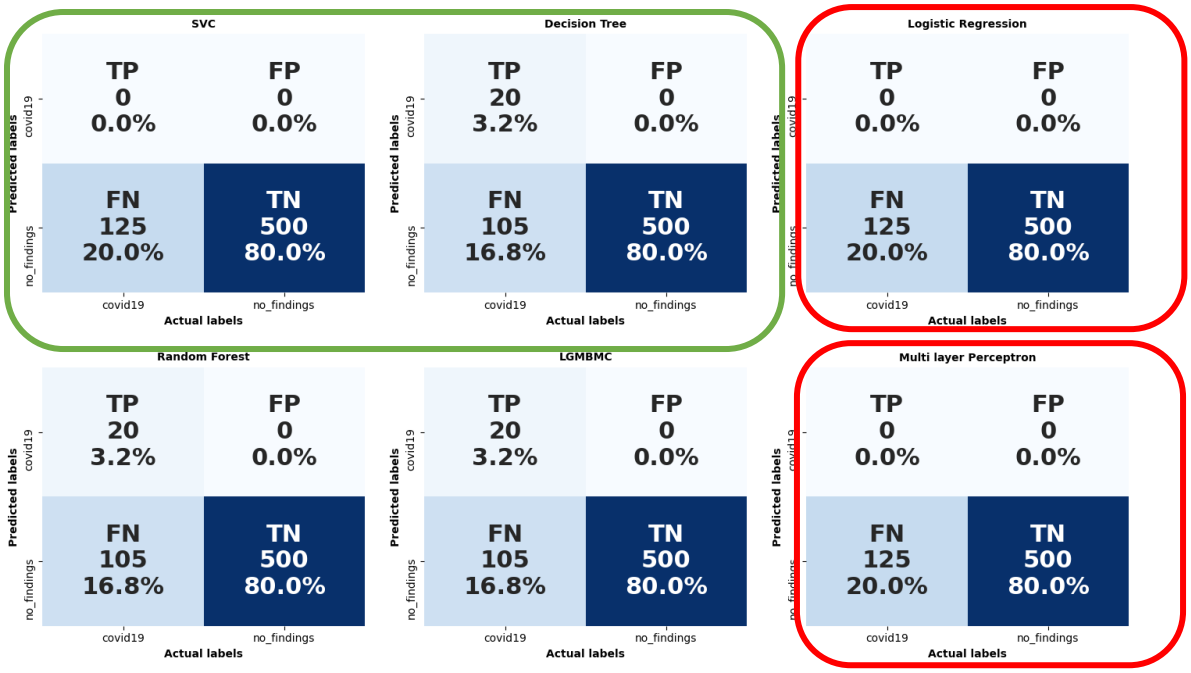
Εικόνα 89: PU learning - bagging ιστόγραμμα απόδοσης αλγορίθμων βάσης στα 80 θετικά.



Εικόνα 90: PU learning - bagging ιστόγραμμα απόδοσης αλγορίθμων βάσης στα 125 θετικά.

Βλέπουμε πράγματι με την ολοκλήρωση των θετικών, οι LGMBMC, Decision Tree και Random forest φτάνουν το 100%. Παρακάτω βλέπουμε τα αντίστοιχα confusion matrix.

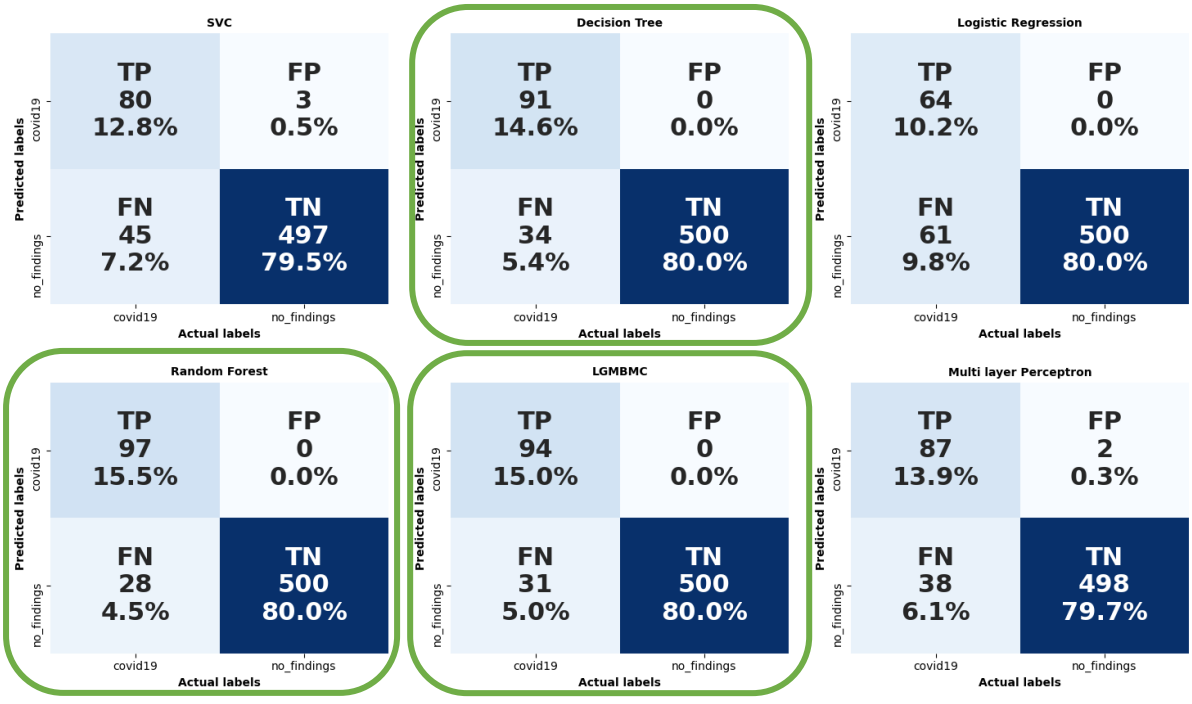
PU learning bagging
confusion matrix με 20 στά 625(3.2%) πραγματικά επισημειωμένα δείγματα



Εικόνα 91: PU learning - bagging confusion matrix απόδοσης αλγορίθμων βάσης στα 20 θετικά.

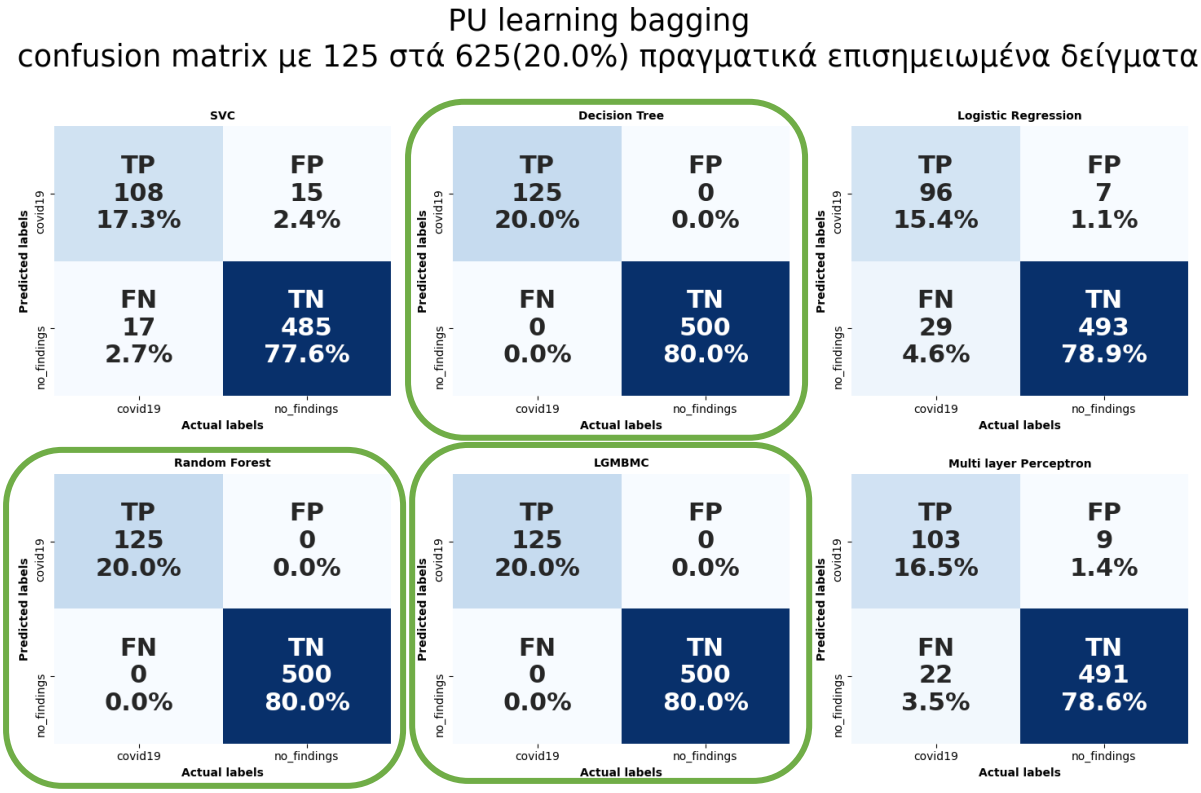
Στον παραπάνω πίνακα βλέπουμε ότι όλα τα FP + TP είναι 0 και γι' αυτό η precision είναι ίση με 1. Αλλά και μόνο όταν η FP = 0, η precision εμφανίζεται = 1.

PU learning bagging
confusion matrix με 90 στά 625(14.4%) πραγματικά επισημειωμένα δείγματα



Εικόνα 92: PU learning - bagging confusion matrix απόδοσης αλγορίθμων βάσης στα 90 θετικά.

Από την παραπάνω εικόνα διαπιστώνουμε ότι, ενώ όλοι οι αλγόριθμοι βάσης βρίσκουν όλα τα αρνητικά, έχουν μια δυσκολία στην αναγνώριση των θετικών. Καλύτερος είναι ο Random Forest, ο οποίος αναγνωρίζει 97 από τα 125 θετικά, ακολουθεί ο LGMBMC με 94 και ο Decision Tree με 91. Τα False Positives που είναι 0 δίνουν precision 1 και specificity 1.



Εικόνα 93: PU learning - bagging confusion matrix απόδοσης αλγορίθμων βάσης στα 125 θετικά.

Με την εξάντληση και των 125 θετικών που υπάρχουν στο σύνολο των δεδομένων μας, βλέπουμε ότι οι LGMBMC, Random Forest, και Decision Tree έχουν προβλέψει σωστά όλα τα θετικά και όλα τα αρνητικά. Παρακάτω οι τιμές μετά από 3-fold cross validation.

Πίνακας 8: Μετρικές PU Learning – bagging, με συμμετοχή 90, 125 θετικών.

Πραγματικά επισημειωθέντα: 90						
estimator	set	accuracy	precision	sensitivity	specificity	f1-score
SVC	test	0.9231	0.9786	0.6224	0.9959	0.7587
	train	0.9199	0.9683	0.6209	0.9950	0.7563
	Delta	0.0032	0.0103	0.0016	0.0009	0.0023
	STD	0.0039	0.0154	0.0607	0.0029	0.0423
Decision Tree	test	0.9423	0.9222	0.7720	0.9839	0.8401
	train	0.9447	1.0000	0.7257	1.0000	0.8409
	Delta	0.0024	0.0778	0.0463	0.0161	0.0008
	STD	0.0039	0.0157	0.0289	0.0032	0.0185
Logistic Regression	test	0.8974	1.0000	0.4876	1.0000	0.6548
	train	0.8974	1.0000	0.4880	1.0000	0.6559
	Delta	0.0000	0.0000	0.0004	0.0000	0.0011
	STD	0.0060	0.0000	0.0344	0.0000	0.0316
Random Forest	test	0.9615	0.9917	0.8140	0.9980	0.8933

	train	0.9543	1.0000	0.7719	1.0000	0.8713
	Delta	0.0072	0.0083	0.0420	0.0020	0.0221
	STD	0.0039	0.0118	0.0373	0.0029	0.0172
LGMBMC	test	0.9551	1.0000	0.7803	1.0000	0.8718
	train	0.9471	1.0000	0.7377	1.0000	0.8487
	Delta	0.0080	0.0000	0.0426	0.0000	0.0231
	STD	0.0278	0.0000	0.1166	0.0000	0.0732
Multi-layer Perceptron	test	0.9215	1.0000	0.6066	1.0000	0.7549
	train	0.9231	0.9939	0.6196	0.9990	0.7616
	Delta	0.0016	0.0061	0.0130	0.0010	0.0067
	STD	0.0060	0.0000	0.0221	0.0000	0.0171
Πραγματικά επισημειωθέντα: 125						
estimator	**set**	**accuracy**	**precision**	**sensitivity**	**specificity**	**f1-score**
SVC	test	0.9375	0.8613	0.8258	0.9634	0.8404
	train	0.9423	0.8648	0.8440	0.9670	0.8538
	Delta	0.0048	0.0034	0.0182	0.0036	0.0135
	STD	0.0204	0.0544	0.0773	0.0184	0.0446
Decision Tree	test	0.9631	0.9260	0.8892	0.9820	0.9057
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0369	0.0740	0.1108	0.0180	0.0943
	STD	0.0099	0.0320	0.0532	0.0083	0.0265
Logistic Regression	test	0.9375	0.9159	0.7627	0.9822	0.8297
	train	0.9479	0.9510	0.7792	0.9900	0.8564
	Delta	0.0104	0.0351	0.0165	0.0078	0.0267
	STD	0.0079	0.0681	0.0353	0.0143	0.0200
Random Forest	test	0.9744	0.9820	0.8891	0.9961	0.9329
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0256	0.0180	0.1109	0.0039	0.0671
	STD	0.0045	0.0255	0.0100	0.0056	0.0096
LGMBMC	test	0.9631	0.9216	0.8927	0.9823	0.9052
	train	1.0000	1.0000	1.0000	1.0000	1.0000
	Delta	0.0369	0.0784	0.1073	0.0177	0.0948
	STD	0.0045	0.0474	0.0334	0.0091	0.0127
Multi-layer Perceptron	test	0.9503	0.9116	0.8337	0.9800	0.8704
	train	0.9471	0.9145	0.8115	0.9809	0.8597
	Delta	0.0032	0.0028	0.0222	0.0009	0.0107
	STD	0.0060	0.0162	0.0288	0.0024	0.0120

Στην γενική εικόνα η specificity υπερτερεί της sensitivity που σημαίνει ότι τα αρνητικά αποκαλύπτονται με μεγαλύτερη αξιοπιστία από ότι τα θετικά.

Συμπεράσματα bagging:

1. Οι Multi-Layer Perceptron και Logistic Regression μέχρι τα 40 θετικά επισημειωμένα έχουν f1-score μηδέν, ενώ για το SVC η τιμή της f1-score είναι μηδενική μέχρι τα 80 θετικά επισημειωμένα. Από εκεί και πέρα, όταν έχουμε διαθέσιμα 100 θετικά επισημειωμένα σε ένα σύνολο 625, τότε το f1-score περνάει το 90%.

2. Όταν έχουμε 120 θετικά επισημειωμένα, οι επιδώσεις των Decision Tree, Random Forest και LGBMC εξακολουθούν να ανεβαίνουν, ενώ οι υπόλοιπες διατηρούνται στο ίδιο επίπεδο, το οποίο όμως ελάχιστα υστερεί. Οι επιδώσεις της LGMBMC στα 125 θετικά επισημειωμένα είναι accuracy 0.9631, precision 0.9216, sensitivity 0.8927 0.9823 0.9052.

3. Γενικά η specificity είναι μεγαλύτερη της sensitivity, το οποίο σημαίνει ότι τα αρνητικά αποκαλύπτονται με πιο μεγάλη ακρίβεια από τα θετικά.

4. Πάνω από 90 θετικά, η διαφορά μεταξύ train και τεστ, όπως και η τυπική απόκλιση μεταξύ των αποτελεσμάτων της διασταυρούμενης επικύρωσης δεν μας δίνουν καμία ένδειξη για υπερπροσαρμογή, ή υπερπροσαρμογή.

5. Όλοι οι αλγόριθμοι βάσης όταν έχουν στην διάθεσή τους πάνω από 100 θετικά επισημειωμένα, τότε η f1-score τους περνάει το 90%.

ΚΕΦΑΛΑΙΟ 9 ΑΝΑΚΕΦΑΛΑΙΩΣΗ

Έχουν γίνει αρκετές προσπάθειες μέχρι σήμερα από στην επιστημονική κοινότητα για αναζήτηση συμπληρωματικών διαγνωστικών μεθόδων με σκοπό την έγκαιρη διάγνωση και πρόληψη της μετάδοσης της νόσου COVID-19 με χρήση ιατρικών εικόνων σε συνδυασμό με την μηχανική μάθηση. Οι προσπάθειες αυτές αφορούν κυρίως πλήρως εποπτευόμενη μάθηση.

Στον τομέα της ημι-εποπτευόμενης μάθησης και της ενεργούς μάθησης οι προσπάθειες που έγιναν ήταν σχετικά λίγες και αφορούσαν κυρίως μάθηση από αξονικές τομογραφίες, οι οποίες ασφαλώς έχουν την χρησιμότητα τους, αλλά οι απλές ακτινογραφίες λόγω της τιμής τους αλλά και της συχνότητας με την οποία επιβάλλονται στους ασθενείς από τους γιατρούς έπρεπε κατά την γνώμη μας να τύχουν μεγαλύτερης προσοχής. Στον τομέα δε της μάθησης από θετικά και μη επισημειωμένα δεδομένα, υπάρχουν ελάχιστες δημοσιευμένες προσπάθειες. Όπως είναι φανερό αυτή θα είναι η πιο συνηθισμένη περίπτωση, να έχουμε δηλαδή πολλές ακτινογραφίες διαθέσιμες, αλλά μόνο λίγες από αυτές θα είναι επισημειωμένες και μάλιστα μόνον θετικά.

Οι παραπάνω λόγοι μας οδήγησαν να επιλέξουμε τις ακτινογραφίες ως πηγή δεδομένων και δόθηκε λίγο περισσότερη προσοχή στην διερεύνηση των αλγορίθμων της κατηγορίας PU learning. Σκοπός της εργασίας δεν είναι να συμπληρώσει τις υπάρχουσες μεθόδους με αλγορίθμους αλλά απλώς να τις μελετήσει και συγκρίνει μεταξύ τους και να διαπιστώσει διάφορα κενά ή ελλείψεις. Δεν εξετάσθηκε η βελτιστοποίηση των υπερπαραμετρων των διάφορων μοντέλων, που μπορεί να αλλάξει άρδην τα αποτελέσματα της παρούσας εργασίας.

Όλοι αυτοί οι αλγόριθμοι που ασχολούνται με μάθηση από μερικώς επισημειωμένα δεδομένα, χρησιμοποιούν αλγορίθμους πλήρως εποπτευόμενης μάθησης ως αλγορίθμους βάσης για να κάνουν τις εκτιμήσεις τους. Σκοπός της παρούσας αφορά την μελέτη της αποτελεσματικότητας αυτών των αλγορίθμων βάσης και την σύγκριση μεταξύ τους.

Όσον αφορά την διαδικασία κατασκευής του μοντέλου, ακολουθήθηκαν τα κλασικά βήματα:

Συλλογή δεδομένων. Επελέγη η κατηγορία που θέλουμε να ασχοληθούμε (ακτινογραφίες), έγινε αναζήτηση στο διαδίκτυο και επελέγη ένα μικρό σύνολο (625) επισημειωμένων ακτινογραφιών το οποίο έχει φέρει αξιόλογα αποτελέσματα στην κατασκευή ενός μοντέλου πλήρως εποπτευόμενης μάθησης.

Εξαγωγή χαρακτηριστικών με τελευταίας τεχνολογίας (state of the art) μέθοδο μεταφοράς μάθησης από προεκπαιδευμένο συνελικτικό δίκτυο και συγκεκριμένα από το DenseNet169.

Προεπεξεργασία με κανονικοποίηση και απομείωση των χαρακτηριστικών σε 20, για αποφυγή υπερπροσαρμογών και ελαχιστοποίηση της απαιτούμενης επεξεργαστικής ισχύος. Δεν κρίθηκε σκόπιμο να γίνει ενίσχυση των δεδομένων.

Ακολούθησε το σημαντικότερο κομμάτι που είναι η υλοποίηση μοντέλων μηχανικής μάθησης και η δοκιμή τους με διάφορες συνθήκες. Η βασική υπόθεση της εργασίας είναι ότι ο αριθμός των πραγματικά (χειροκίνητα) επισημειωμένων δειγμάτων που έχουμε στην διάθεσή μας επηρεάζει την αποτελεσματικότητα του μοντέλου που εκπαιδεύουμε.

Όλοι οι αλγόριθμοι που αφορούν σε μηχανική μάθηση με μερικώς επισημειωμένα δεδομένα χρησιμοποιούν έναν αλγόριθμο βάσης, ο οποίος είναι ένας αλγόριθμος εποπτευόμενης μάθησης για να κάνουν τις εκτιμήσεις τους.

Η βασική διαδικασία του πειράματος που σχεδιάσαμε είναι να αυξάνουμε προοδευτικά τον αριθμό των θεωρούμενων πραγματικά επισημειωμένων, μέχρις ότου πέτυχουμε τιμή f1-score για το μοντέλο μας > 90%. Κατά την εφαρμογή της διαδικασίας εκτελούμε μετρήσεις, καθώς αυξάνουμε τα θεωρούμενα ως επισημειωμένα ανά 10 σε κάθε επανάληψη.

Αξιολογούμε τα μοντέλα που δημιουργούν οι διάφοροι αλγόριθμοι βάσης μεταξύ τους με 3-fold cross-validation.

Παρουσιάζουμε τα αποτελέσματα υπό μορφή διαγραμμάτων, confusion matrix, ιστογραμμάτων και πινάκων.

Οι αλγόριθμοι που εξετάσθηκαν είναι:

Η active learning, μία μέθοδος πλήρως εποπτευόμενης μάθησης η οποία επιλέγει τα προς επισημείωση δεδομένα με βάση μια στρατηγική ερωτήσεων.

Η δεύτερη μεγάλη κατηγορία που εξετάσθηκε είναι η ημι-εποπτευόμενη μάθηση. Σε αυτήν επισημειώνονται αυτόματα ορισμένα ή όλα τα μη επισημειωμένα δεδομένα. Αξιολογήθηκαν οι τρεις κυριότεροι αλγόριθμοι: Η αυτοεκπαίδευση (self-training), διάδοση και η εξάπλωση των επισημειώσεων (Label Propagation – Label Spreading).

Τέλος δοκιμάστηκαν αλγόριθμοι μάθησης με θετικά και μη επισημειωμένα δεδομένα. Συγκεκριμένα οι λεγόμενοι Elkanoto, Weighted Elkanoto και Bagging. Η λειτουργία τους είναι παρόμοια με αυτήν της ημι-εποπτευόμενης μάθησης.

Τελικά δεν βρέθηκε ένας αλγόριθμος βάσης που να υπερτερεί ε όλες τις περιπτώσεις αλλά έγιναν διαπιστώσεις κατά περίπτωση. Η αδιαμφισβήτητη διαπίστωση είναι ότι ο καλύτερος αλγόριθμος βάσης για την ενεργό μάθηση είναι η επιτροπή αλγορίθμων.

ΚΕΦΑΛΑΙΟ 10 ΣΥΜΠΕΡΑΣΜΑΤΑ

(Υποκεφάλαιο 10.1) Συμπεράσματα με βάση τις δοκιμές

(Ενότητα 10.1.α) Γενικές διαπιστώσεις

Η μηχανική μάθηση με μερικώς επισημειωμένα δεδομένα είναι εφικτή και μπορεί άριστα να αντικαταστήσει την πλήρως εποπτευόμενη μάθηση όπου υπάρχουν οι προϋποθέσεις και απαιτείται.

Σε όλους τους αλγορίθμους που αντιμετωπίζουν το πρόβλημα της μάθησης με μερικώς επισημειωμένα δεδομένα υπάρχει ένας εκτιμητής ή αλγόριθμος βάσης, ο οποίος μπορεί να είναι οποιασδήποτε αλγόριθμος εποπτευόμενης μάθησης. Η δουλειά του εκτιμητή είναι να συμβουλεύσει τον αλγόριθμο της «μάθησης από μερικώς επισημειωμένα δεδομένα» σε ποια κλάση να τοποθετήσει κάθε δείγμα και με πια πιθανότητα έκανε αυτήν την εκτίμηση. Από τις δοκιμές που έγιναν με όλους τους συνδυασμούς «αλγόριθμος μάθησης με μερικώς επισημειωμένα δεδομένα -αλγόριθμος βάσης - αριθμός υπαρχόντων πραγματικά επισημειωμένων», προέκυψαν τα παρακάτω:

1. Σε κάθε επανάληψη των δοκιμών ο κάθε τριπλός συνδυασμός έδινε μονότονα το ίδιο αποτέλεσμα f1-score, με πολύ μικρές διαφορές, που σημαίνει ότι ο κάθε συνδυασμός «αλγόριθμος μάθησης με μερικώς επισημειωμένα δεδομένα – αλγόριθμος βάσης – αριθμός επισημειωμένων δεδομένων» δίνει πάντα το ίδιο αποτέλεσμα.

2. Η αύξηση του αριθμού των αρχικά πραγματικά επισημειωμένων δειγμάτων πετυχαίνει γενικά υψηλότερες τιμές για το f1-score. Όμως διαπιστώθηκαν και παρεκκλίσεις από τον κανόνα αυτό, οι οποίες εντοπίσθηκαν και καταγράφηκαν.

3. Στην περίπτωση της ανισορροπίας δεδομένων όπως είναι η δική μας περίπτωση, η accuracy δεν είναι κατάλληλη για την μέτρηση της αποτελεσματικότητας, η καταλληλότερη μετρική στις περιπτώσεις ταξινόμησης που η μία κατηγορία υστερεί η υπερτερεί αριθμητικά από την άλλη είναι η f1-score.

(Ενότητα 10.1.β) Συμπεράσματα επί των αλγορίθμων μερικώς επισημειωμένων δεδομένων και των αλγορίθμων βάσης που χρησιμοποιούν

Συμπεράσματα Semi-supervised – self-training:

Τα 200 πραγματικά επισημειωθέντα δείγματα είναι αρκετά για να ψευδοεπισημειώσουμε τα υπόλοιπα 415 ώστε να δημιουργήσουμε ένα ταξινομητή με f1-score > 75% μάθησης με μερικώς επισημειωμένα δεδομένα με οποιονδήποτε ταξινομητή βάσης. Οι LGMBMC και Random Forest χρειάζονται μόλις 100 πραγματικά επισημειωμένα για να φτάσουμε τιμές f1-score > 80%.

Ο ρυθμός ανόδου του f1_score πάνω από τα 100 πραγματικά επισημειωμένα δείγματα επιβραδύνεται σημαντικά, για όλους τους αλγορίθμους βάσης ενώ πάνω από τα 200 παραμένει σχεδόν σταθερός.

Εάν στόχος μας είναι ένα μοντέλο με απόδοση f1-score > 90% τότε πρέπει να έχουμε 200 πραγματικά επισημειωμένα δείγματα και να χρησιμοποιήσουμε έναν από τους LGMBMC, Random Forest, Decision Tree ως αλγόριθμο βάσης.

Στα 200 πραγματικά επισημειωμένα δείγματα ο Random Forest επιτυγχάνει: accuracy 0.9631, precision 0.9915, sensitivity 0.8241, specificity 0.9980 και f1-score 0.8997, στα ίδια επίπεδα περίπου κυμαίνεται και ο LGMBMC.

Η specificity είναι πολύ υψηλότερη σε σχέση με την sensitivity που σημαίνει ότι τα αρνητικά αναγνωρίζονται πιο εύκολα.

Συμπεράσματα semi supervised – label propagation και spreading

Οι αλγόριθμοι label-propagation και label-spreading δεν έχουν διαφορά όσον αφορά τις τιμές των μετρικών, δηλαδή την αξιοπιστία των μοντέλων που παράγουν.

Αυτό που κάνει την διαφορά είναι οι αλγόριθμοι (πυρήνες βάσης). Ο πυρήνας «rbf» στον Label-Spreading απαιτεί περίπου 150 πραγματικά επισημειωμένα για να έχει accuracy 0.9151, precision 0.9622, sensitivity 0.5974, specificity 0.9940 και f1-score 0.7338 ενώ ο «knn» θέλει περίπου 450 για να φτάσει στα ίδια επίπεδα. Δηλαδή ο «rbf» υπερέχει σαφώς.

Πάνω από τα 450 πραγματικά επισημειωμένα δεν υπάρχει εμφανής βελτίωση της f1-score με την αύξησή των πραγματικά επισημειωμένων για κανένα πυρήνα και κανένα αλγόριθμο.

Ο πυρήνας «rbf» μαζί με τον αλγόριθμο label-propagation όταν έκανε χρήση **450** πραγματικών επισημειωμένων πέτυχε: accuracy 0.9343, precision 0.9174, sensitivity 0.7344, specificity 0.9840 και f1-score 0.8154.

Συμπεράσματα active learning

Χρειαζόμαστε πολύ πιο λίγες επισημειώσεις όταν η επιλογή των προς επισημείωση γίνεται σύμφωνα με μία στρατηγική ερωτήσεων και όχι τυχαία.

Μεγαλύτερη σημασία έχει το αλγόριθμος βάσης του μοντέλου και μικρότερη η στρατηγική των ερωτήσεων, με την προϋπόθεση βέβαια ότι ακολουθείται κάποια στρατηγική.

Οι επιτροπές με συμμετοχή διαφόρων τύπων μοντέλων είναι αυτές που μπορούν να φέρουν με λίγα δείγματα μεγάλες αποδόσεις όσον αφορά την τιμή της f1-score.

Ο committee vote_entropy_sampling, αλλά και οι υπολοίπου αλγόριθμοι επιτροπών από τα 40 δείγματα το f1_score πλησιάζει το 100%. **Ο commitee vote_entropy_sampling πέτυχε: accuracy=0.9984, precision=1.0000, sensitivity= 0.9920, specificity=1.0000, f1-score= 0.9960 στα 40 επισημειωμένα δείγματα.**

Συμπεράσματα αλγορίθμου PU Elkanoto:

Όταν τα θετικά επισημειωμένα, τα οποία έχουμε στην διάθεσή μας, είναι κάτω από 20, κανένας αλγόριθμος δεν δίνει f1-score πάνω από 60%, άρα δεν υπάρχει δυνατότητα να εκπαιδεύσουμε σε αυτά τα επίπεδα αξιόπιστο μοντέλο.

Με ένα αριθμό 90 πραγματικά θετικά επισημειωμένων δειγμάτων στα συνολικά 625 δείγματα, σχεδόν όλα τα μοντέλα που στηρίζονται στον Elkanoto δίνουν μετρικές πάνω από 80%.

Οι τιμές της f1-score, διαμορφώνονται αναλογικά με τα διαθέσιμα θετικά επισημειωμένα για τους Decision Tree, Random Forest και LGMBC. **Όταν ο αριθμός των θετικά επισημειωμένων είναι 125, ο LGMBMC έχει τιμές: accuracy 0.9744, precision 0.9803, sensitivity 0.8845, specificity 0.9961, f1-score 0.9298**. Γι' αυτό, στην περίπτωση που έχουμε μεγάλο αριθμό θετικά επισημειωμένων προτιμούμε τον LGMBMC.

Γενικά, η specificity έχει καλύτερες τιμές από την sensitivity με όλους τους αλγόριθμους βάσης, το οποίο σημαίνει ότι τα αρνητικά αναγνωρίζονται με μεγαλύτερη ευκολία.

Τέλος οι Logistic Regression και Multilayer Percepton, όταν διαθέτουν πάνω από 40 θετικά δείγματα, δεν επηρεάζονται από τον αριθμό των θετικά επισημειωμένων, στοιχείο που δείχνει ότι ο αριθμός 40 πιθανόν να έχει εφαρμογή και σε άλλα σύνολα δεδομένων.

Συμπεράσματα Weighted Elkanoto:

Η τιμή της f1-score του Decision Tree είναι ανάλογη του αριθμού των θετικών επισημειωμένων. Όταν φτάσουμε τα 125, δηλαδή εξαντλήσουμε τον αριθμό των θετικών, τότε έχουμε πετύχει accuracy 0.9647, precision 0.9176, sensitivity 0.9048, specificity 0.9801 και f1-score 0.9099.

Ο Random Forest μετά τα 90 θετικά, παρότι η f1-score είναι κοντά στο 50%, μας δίνει sensitivity 100%, το οποίο σημαίνει ότι αναγνωρίζει όλα τα θετικά. Αυτό είναι σημαντικό επειδή α) για την καταπολέμηση της εξάπλωσης της νόσου σημαντικό είναι να μπορούμε να αναγνωρίζουμε τα θετικά. β) πάνω από τον αριθμό 90 επισημειωμένων η απόδοση του Random Forest είναι σταθερή και ανεξάρτητη του αριθμού των επισημειωμένων θετικών.

Ο SVC έχει sensitivity 100% και specificity 0%, σε όλες τις περιπτώσεις. Δηλαδή προβλέπει ότι όλα είναι θετικά και ως εκ τούτου απορρίπτεται. Για τον ίδιο λόγο απορρίπτονται, οι Logistic Regression, Multi–Layer Perceptron και LGMBMC, βρίσκουν όλα τα θετικά αλλά σχεδόν κανένα αρνητικό Τελικά καταλήγουμε στο Decision Tree με πάνω από 100 επισημειωμένα (τα 90 είναι οριακά ανεπαρκή).

Γενικά το sensitivity υπερτερεί του specificity, το οποίο σημαίνει ότι τα θετικά αποκαλύπτονται με μεγαλύτερη ευκολία, το οποίο συνεπάγεται ότι γνωρίζοντας τους θετικούς με ακρίβεια μπορεί να συντελέσει στον περιορισμό της μετάδοσης της νόσου.

Όλοι οι αλγόριθμοι που δοκιμάστηκαν πλην του Decision Tree είτε δεν συνεργάζονται ικανοποιητικά με τον weighted Elkanoto είτε τα δεδομένα μας είναι ποσοτικά ανεπαρκή για να δημιουργηθεί ένα αποτελεσματικό μοντέλο προβλέψεων.

Συμπεράσματα Bagging:

Οι Multi-Layer Perceptron και Logistic Regression μέχρι τα 40 θετικά επισημειωμένα έχουν f1-score μηδέν, ενώ για το SVC η τιμή της f1-score είναι μηδενική μέχρι τα 80 θετικά επισημειωμένα. Από εκεί και πέρα, όταν

έχουμε διαθέσιμα 100 θετικά επισημειωμένα σε ένα σύνολο 625, τότε το f1-score περνάει το 90%.

Όταν έχουμε 120 θετικά επισημειωμένα, οι επιδώσεις των Decision Tree, Random Forest και LGBMC εξακολουθούν να ανεβαίνουν, ενώ οι υπόλοιπες διατηρούνται στο ίδιο επίπεδο, το οποίο όμως ελάχιστα υστερεί.

Οι επιδώσεις της LGMBMC στα 125 θετικά επισημειωμένα (‘ολα τα πραγματικά διαθέσιμα) είναι accuracy 0.9631, precision 0.9216, sensitivity 0.8927 specificity 0.9823 και f1-score 0.9052.

Γενικά η specificity είναι μεγαλύτερη της sensitivity, το οποίο σημαίνει ότι τα αρνητικά αποκαλύπτονται με πιο μεγάλη ευκολία από τα θετικά.

(Υποκεφάλαιο 10.2) Συμπεράσματα με βάση την μελέτη των αλγορίθμων

(Ενότητα 10.2.α) Προβλήματα μάθησης από θετικά και μη επισημειωμένα.

Το ερώτημα που δημιουργήθηκε κατά την εξέταση των μεθόδων μάθησης με θετικά και μη επισημειωμένα δεδομένα είναι: όταν στον πραγματικό κόσμο, ο σκοπός μας θα είναι να φτιάξουμε ένα μοντέλο από θετικά και μη επισημειωμένα δεδομένα, πως μπορούμε να αξιολογήσουμε το μοντέλο αυτό; Είναι προφανές ότι αφού έχουμε στην διάθεσή μας ένα σύνολο δεδομένων, στο οποίο οι επισημειώσεις είναι μόνο θετικές, δεν μπορούμε να υπολογίσουμε τις μετρικές στις οποίες εμπλέκονται τα αρνητικά, δηλαδή τα TN και FP, αφού δεν γνωρίζουμε ποια είναι αυτά στην πραγματικότητα, γνωρίζουμε μόνο ορισμένα από τα θετικά. Η λύση είναι να μετράμε μόνο την sensitivity = TP/(TP+FN) ή να κατασκευάσουμε δίκες μας μετρικές πχ FN/TP. Είναι επόμενο, ότι δεν θα γνωρίζουμε την αξιοπιστία των προβλέψεών μας όσον αφορά την αναγνώριση αρνητικών.

(Ενότητα 10.2.β) Σύγκριση – Χρήση αλγορίθμων μάθησης με μερικώς επισημειωμένα δεδομένα.

Οι αλγόριθμοι ενεργούς μάθησης, ημι-εποπτευόμενης μάθησης με θετικά και μη επισημειωμένα δεν είναι ανταγωνιστικοί μεταξύ τους, καθόσον ασχολούνται με διαφορετικού τύπου προβλήματα. Ο αλγόριθμος μάθησης με θετικά και μη επισημειωμένα δεδομένα εμπλέκεται μόνο σε προβλήματα που υπάρχουν θετικά και μη επισημειωμένα, και τα οποία δεν αφορούν ούτε την ενεργή μάθηση, ούτε την ημι-εποπτευόμενη μάθηση. Όμως η ενεργή μάθηση και η ημι-εποπτευόμενη μπορούν να συνεργαστούν ως εξής:

Πχ υπάρχουν ορισμένα επισημειωμένα δεδομένα όμως που δεν επαρκούν για πλήρως εποπτευόμενη μάθηση, τότε ακολουθούμε την παρακάτω διαδικασία:

Δοκιμάζουμε την ημι-εποπτευόμενη μάθηση ψευδοεπισημειώνοντας (με αυτόματο τρόπο) έναν αριθμό μη επισημειωμένων (πχ. πάνω από ένα συγκεκριμένο επίπεδο εμπιστοσύνης ή και όλα τα διαθέσιμα μη επισημειωμένα), εάν και πάλι δεν μπορούμε να φτιάξουμε ένα αξιόπιστο μοντέλο, καλείται η ενεργή μάθηση να μας φέρει ορισμένο αριθμό επισημειωμένων, με την διαδικασία της επιτροπής και μία στρατηγική ερωτήσεων. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου έχουμε το αποτέλεσμα που επιθυμούμε.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aarohi. (22, Ιουνίου 2022). *DenseNet | Densely Connected Convolutional Networks*. Ανάκτηση από Youtube: https://www.youtube.com/watch?v=hCg9bolMeJM

Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2020, Σεπτεμβρίου 5). Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Applied Intelligence*, σσ. vol. 51, no. 2, pp. 854–864.

Abdulkareem, K. H., Mohammed, M. A., Salim, A., Arif, M., Geman, O., Gupta, D., & Khanna, A. (2021, Νοεμβρίου 01). Realizing an effective COVID-19 diagnosis system based on machine learning and IoT in smart hospital environment. *IEEE Internet of Things Journal*, σσ. vol. 8, no. 21, pp. 15919–15928.

Agmon, A. (2020, Μαρτίου 6). *Semi-Supervised Classification of Unlabeled Data (PU Learning)*. Ανάκτηση από Towards Data Science: https://towardsdatascience.com/semi-supervised-classification-of-unlabeled-data-pu-learning-81f96e96f7cb

Ahmad, F., Farooq, A., & Ghani, M. U. (2020, Ιανουαρίου 12). Deep ensemble model for classification of novel coronavirus in chest X-ray images. *Computational Intelligence and Neuroscience*, σσ. vol. 2021, Article ID 8890226, 17 pages.

Alizadehsani, R., Sharifrazi, D., Izadi3, N. H., Javad Hassannataj Joloudari, Shoeibi5, A., Gorriz, J. M., . . . Mohammed, S. (2021, Δεκεμβρίου 25). *Electrical Engineering and Systems Science > Image and Video Processing*. Ανάκτηση από Cornell University-arxiv: https://arxiv.org/abs/2102.06388

Apostolopoulos, I. D., & Mpesiana, T. A. (2020, Απριλίου 3). COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, σσ. vol. 43, no. 2, pp. 635–640.

Arnx, A. (2019, Ιανουαρίου 13). *First neural network for beginners explained*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/first-neural-network-for-beginners-explained-with-code-4cfd37e06eaf

Aylward, S. R. (χ.χ.). *Open-Access Medical Image Repositories*. Ανάκτηση από www.aylward.org: https://www.aylward.org/notes/open-access-medical-image-repositories

Baheti, P. (2022, Ιουλίου 19). *A Newbie-Friendly Guide to Transfer Learning*. Ανάκτηση από v7labs: https://www.v7labs.com/blog/transfer-learning-guide

Bajaj, A. (2019, Δεκεμβρίου 25). *What does your classification metric tell about your data*. Ανάκτηση από Towards Data Science: https://towardsdatascience.com/what-does-your-classification-metric-tell-about-your-data-4a8f35408a8b

Bekker, J., & Davis, J. (2018, Νοεμβρίου 12). *Learning From Positive and Unlabeled Data*. Ανάκτηση από Research Gate: https://www.researchgate.net/

Bekker, J., & Davis, J. (2020, Απριλίου 02). *Learning from positive and unlabeled data*. Ανάκτηση από springer: https://link.springer.com/article/10.1007/s10994-020-05877-5

Boehmke, B., & Greenwell, B. (2020, Φεβρουαρίου 1). *Gradient Boosting*. Ανάκτηση από Hands-On Machine Learning with R: https://bradleyboehmke.github.io/HOML/gbm.html

Bouneffouf, D. (2016, Ιανουαρίου 8). *Exponentiated Gradient Exploration for Active Learning*. Ανάκτηση από MDPI: https://www.mdpi.com/2073-431X/5/1/1

Bouneffouf, D., Laroche, R., Urvoy, T., Féraud, R., & Allesiardo, R. (2014, Σεπτεμβρίου 29). *Contextual Bandit for Active Learning*. Ανάκτηση από HAL: https://hal.archives-ouvertes.fr/hal-01069802

Brownlee, J. (2019, Οκτοβρίου 25). Ανάκτηση από https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

Brownlee, J. (2019, Οκτωβρίου 25). *Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning*. Ανάκτηση από Machine Learning Mastering: https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

Brownlee, J. (2019, Νοεμβρίου 27). *How to Choose a Feature Selection Method For Machine Learning*. Ανάκτηση από Machine Learning Mastery: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Brownlee, J. (2019, Νοεμβρίου 27). *How to Choose a Feature Selection Method For Machine Learning*. Ανάκτηση από Machine Learning Mastery: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Brownlee, J. (2021, 16 Φεβρουαρίου). *Regression Metrics for Machine Learning*. Ανάκτηση από https://machinelearningmastery.com: https://machinelearningmastery.com/regression-metrics-for-machine-learning/

Burkhart, M. C., & Shan, K. (2020, Ιουνίου 15). *Deep Low-Density Separation for Semi-supervised Classification*. Ανάκτηση από springer.: https://link.springer.com/chapter/10.1007/978-3-030-50420-5_22

Burkhart, M. C., & Shan, K. (2020, Ιουνίου 15). *Deep Low-Density Separation for Semi-supervised Classification*. Ανάκτηση από Springer Link: https://link.springer.com/chapter/10.1007/978-3-030-50420-5_22

Chen1, L., & Rezaei, T. (2021, Αυγούστου 24). A new optimal diagnosis system for coronavirus (COVID-19) diagnosis. *Computational Intelligence and Neuroscience*, σσ. vol. 2021, Article ID 7788491, 9 pages, 2021.

Choudhury, A. (2021, Ιανουαρίου 18). *AdaBoost Vs Gradient Boosting: A Comparison Of Leading Boosting Algorithms*. Ανάκτηση από Analytics India Magazine: JANUARY 18, 2021

Convolutional Neural Networks (CNNs / ConvNets). (χ.χ.). Ανάκτηση από cs231n.github.io: https://cs231n.github.io/convolutional-networks/

Danka, T. (2018). *A modular active learning framework for Python3*. Ανάκτηση από Modal: https://modal-python.readthedocs.io/en/latest/

Danka, T. (2021, Ιανουαρίου 7). *Modular Active Learning framework for Python3*. Ανάκτηση από Git Hub: https://github.com/modAL-python/modAL

Deep Learning. (χ.χ.). Ανάκτηση από Wikipedia: https://en.wikipedia.org/wiki/Deep_learning

Difference between Model Parameter and Hyperparameter. (χ.χ.). Ανάκτηση από Java T point: https://www.javatpoint.com/model-parameter-vs-hyperparameter

Drouin, A., AditraAS, Palachy, S., & Wright, R. (χ.χ.). *Module PUlearn*. Ανάκτηση από PUlern: https://pulearn.github.io/pulearn/doc/pulearn/#credits

Elkan, C., & Noto, K. (χ.χ.). *Learning Classifiers from only Positive Only Positive and Unlabeled Data.*

Elkan, Noto, Drouin, A., AditraAS, & Wright., R. (χ.χ.). *Module pulearn*. Ανάκτηση από pulearn.github: https://pulearn.github.io/pulearn/doc/pulearn/#credits

Feature Extruction. (χ.χ.). Ανάκτηση από MathWorks: https://www.mathworks.com/discovery/feature-extraction.html

Feature selection. (χ.χ.). Ανάκτηση από scikit-learn.org: https://scikit-learn.org/stable/modules/feature_selection.html

FERNANDO, J. (2021, Μαρτίου 9). *R-Squared Definition*. Ανάκτηση από https://www.investopedia.com: https://www.investopedia.com/terms/r/r-squared.asp

Fortmann-Roe, S. (2012, Ιούνιος). *Understanding the Bias-Variance Tradeoff*. Ανάκτηση από Scott Fortmann-Roe: https://scott.fortmann-roe.com/docs/BiasVariance.html

Gao Huang, Liu, Z., Laurens van der Maaten, & Weinberger, K. Q. (2018, Ιανουαρίου 28). *Densely Connected Convolutional Networks*. Ανάκτηση από Cornel University: 28 Jan 2018

geeksforgeeks. (χ.χ.). *geeksforgeeks*. Ανάκτηση από https://www.geeksforgeeks.org/pattern-recognition-basics-and-design-principles/

Girgin, S. (2019, Σεπτεμβρίου 20). *Day-41 Deep Learning-6 (CNN-2)*. Ανάκτηση από PursuitData: https://medium.com/pursuitnotes/day-41-deep-learning-6-cnn-2-d38b355bd0ba

Gordon, S. (2021, Μαρτίου 8). *Artificial Intelligence*. Ανάκτηση από investopedia: https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp

Gunraj, H., Wang, L., & Wong, A. (2020, Δεκεμβρίου 23). COVIDNet-CT: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images. *Frontiers of Medicine, vol. 7, pp. 1–12, 2020.*, σσ. Vol 7, pp1-22. Ανάκτηση από H. Gunraj, L. Wang, and A. Wong, “COVIDNet-CT: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images,” Frontiers of Medicine, vol. 7, pp. 1–12, 2020.

Gupta, A. (χ.χ.). *Evaluation Metrics for Multi-Class Classification*. Ανάκτηση από kaggle.com: https://www.kaggle.com/nkitgupta/evaluation-metrics-for-multi-class-classification

Gupta, D. (2021, Ιανουαριος 4). *Transfer learning and the art of using Pre-trained Models in Deep Learning*. Ανάκτηση από Analytics Vidhya.: https://www.analyticsvidhya.com/blog/2017/06/transfer-learning-the-art-of-fine-tuning-a-pre-trained-model/

Hamid Nasiri, & Alavi, S. A. (2022, Ιανουαρίου 7). A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images. *Computational Intelligence and neuroscience*, σσ. Volume 2022 , Article ID 4694567. Ανάκτηση από https://www.hindawi.com/journals/cin/2022/4694567/

Han, C. H., Kim, M., & 3, J. T. (2021, Απριλίου 1). *Semi-supervised learning for an improved diagnosis of COVID-19 in CT images*. Ανάκτηση από National Center of Biotechnology Information, Korea: https://pubmed.ncbi.nlm.nih.gov/33793650/

Han, Z., He, R., Li, T., Wei, B., Wang, J., & Yin, Y. (2021). Semi-Supervised Screening of COVID-19 from Positive and Unlabeled Data with Constraint Non-Negative Risk Estimator. Στο A. Feragen, S. Sommer, J. Schnabel', & M. Nielsen, *Information Processing in Medical Imaging* (σσ. 611–623). Cham: Springer Nature Switzerland AG . Ανάκτηση από https://link.springer.com/chapter/10.1007/978-3-030-78191-0_47#editor-information

Hao, H., Moon, H., Didari, S., & Jae Oh Woo, P. B. (2022, Ιουνίου 22). *Highly Efficient Representation and Active Learning Framework and Its Application to Imbalanced Medical Image Classification*. Ανάκτηση από Cornell Univesity: https://arxiv.org/abs/2103.05109

Hasani, H., & Nasiri, S. (2021, Σεπτεμβρίου 3). *Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost*. Ανάκτηση από Cornel University: https://arxiv.org/abs/2109.02428

He, K., Xiangyu Zhang, Shaoqing Ren, & Sun, J. (2015, ΔΕκεμβρίου 10). *Deep Residual Learning for Image Recognition*. Ανάκτηση από Cornel University: https://arxiv.org/abs/1512.03385

He, X., Yang, X., Zhang, S., Zhao, J., Zhang, Y., Xing, E., & Xie, P. (2020, Απριλίου 13). Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. *IEEE Transactions on Medical Imaging*, σσ. vol. XX, no. Xx.

Hemdan, E. E.-D., Shouman, M. A., & Karar, M. E. (2020, Μαρτίου 24). *COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images*. Ανάκτηση από Cornell University: https://arxiv.org/abs/2003.11055

https://www.business-standard.com/about/what-is-artificial-intelligence#collapse. (χ.χ.). *What is Artificial Inteligence*. Ανάκτηση από Bussiness Standard: https://www.business-standard.com/about/what-is-artificial-intelligence#collapse

Huilgol, P. (2019, Αυγούστου 24). *Accuracy vs. F1-Score*. Ανάκτηση από analytics-vidhya: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

Ibañez, A. (2019, Μαίου 30). *Semi-Supervised Learning… the great unknown*. Ανάκτηση από Telefonica: https://business.blogthinkbig.com/semi-supervised-learning-the-great-unknown/

IBM. (2021, Μαρτίου 3). *IBM Cloud Eduvations*. Ανάκτηση από IBM Cloud Eduvations: https://www.ibm.com/cloud/learn/overfitting και https://www.ibm.com/cloud/learn/underfitting

Introduction to DenseNet with TensorFlow. (2020, Μαίου 06). Ανάκτηση από pluralsigh: https://www.pluralsight.com/guides/introduction-to-densenet-with-tensorflow

Javapoint. (χ.χ.). *Regression vs. Classification in Machine Learning*. Ανάκτηση από Javapoint: https://www.javatpoint.com/regression-vs-classification-in-machine-learning

javatpoint. (χ.χ.). *machine-learning*. Ανάκτηση από javatpoint.: https://www.javatpoint.com/machine-learning

javatpoint. (χ.χ.). *types-of-machine-learning*. Ανάκτηση από javatpoint.com: https://www.javatpoint.com/types-of-machine-learning

Jetchev, N. (χ.χ.). *Graph-and-similarity-matrix-connection*. Ανάκτηση από researchgate.net/: https://www.researchgate.net/figure/Graph-and-similarity-matrix-connection_fig4_242508535

Khan, A. I., Shah, J. L., Mudasir, M., & Bhat. (2020, Ιουνίου 10). CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Computer Methods and Programs in Biomedicine*, σσ. vol. 196, Article ID 105581, 2020.

Khorami, E., Babaei, F. M., & Azadeh, A. (2021, Αυγούστου 20). Optimal diagnosis of COVID-19 based on convolutional neural network and red Fox optimization algorithm. *Computational Intelligence and Neuroscience*, σσ. vol. 2021, Article ID 4454507, 11 pages.

Kumar, S. P., Kumari, B. S., Kumar, R. P., & Biswas, P. (2020, Απριλίου 15). Detection of coronavirus disease (COVID-19) based on deep features and support vector machine.

International Journal of Mathematical, Engineering and Management Sciences, σσ. vol. 5, no. 4, pp. 643–651.

Li, G. (χ.χ.). *A Survey on Postive and Unlabelled Learning.* Newark: University of Delaware.

Liu, L., Lei, W., Wan, X., Liu, L., Luo, Y., & Feng, C. (2020, Δεκεμβρίου 24). Semi-Supervised Active Learning for COVID-19 Lung Ultrasound Multi-symptom Classification. *IEEE 32nd International Conference on Tools with Artificial Intelligence*, σσ. pp. 1268-1273.

M. Adnan, F. R. (2018). Handwritten bangla character recognition using inception convolutional neural network,". *International Journal of Computer Application,*, σσ. v ol. 181, no. 17, pp 48–59.

Machine Leraning. (χ.χ.). Ανάκτηση από Developers Google: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Mallawaarachchi, V. (2020, Μαρτίου 6). *Label Propagation Demystified*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/label-propagation-demystified-cd5390f27472

Mallawaarachchi, V. (2020, Μαρτίου 6). *Label Propagation Demystified*. Ανάκτηση από towards datascience: https://towardsdatascience.com/label-propagation-demystified-cd5390f27472

Mazur, M. (χ.χ.). *A Step by Step Backpropagation Example*. Ανάκτηση από https://mattmazur.com: https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/

Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., & Soufi, G. J. (2020, Αυγούστου 8). Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning. *Medical Image Analysis*, σσ. ,vol. 65, Article ID 101794,.

Mordelet, F., & Vert, J.-P. (2010, Οκτοβρίου 6). *A bagging SVM to learn from positive and unlabeled examples.* Ανάκτηση από Cornell University: https://arxiv.org/abs/1010.0772

Narin, A., Kaya, C., & Ziynet, P. (2020, Οκτωμβριος 5). *Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks*. Ανάκτηση από Cornell University: https://arxiv.org/abs/2003.10849

Narkhede, S. (2018, Μαίου 2018). *Understanding Confusion Matrix*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Nasiri, H., & Alavi, S. A. (2022, Ιανουαρίου 7). A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images. *Computational Intelligence and Neuroscience,*, σσ. vol. 2022, Article ID 4694567, 11 pages. Ανάκτηση από https://www.hindawi.com/journals/cin/2022/4694567/

Nasiri, H., & Alavi, S. A. (2022, Ιανουαρίου 7). A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images. *Computational Intelligence and Neuroscience*, σσ. Article ID 4694567, 11 pages, 202.

Navas, J. (2022, Φεβρουαρίου 8). *What is hyperparameter tuning?* Ανάκτηση από www.anyscale.com: https://www.anyscale.com/blog/what-is-hyperparameter-tuning

Nazir, A., & Fajri, R. M. (2021, Νοεμβρίου 23). *Active Learning Strategy for COVID-19 Annotated Dataset*. Ανάκτηση από IEEE: https://ieeexplore.ieee.org/document/9625938

Ozturka, T., Talob, M., Yildirim, E. A., Baloglud, U. B., Yildirime, O., & Acharya, U. (2020, Ιούνιος). Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, σσ. vol. 121, Article ID 103792.

Patyuchenko, A. (χ.χ.). *medical-image-processing-from-formation-to-interpretation*. Ανάκτηση από Analog Services: https://www.analog.com/ru/technical-articles/medical-image-processing-from-formation-to-interpretation.html

Probability calibration. (χ.χ.). Ανάκτηση από https://scikit-learn.org: https://scikit-learn.org/stable/modules/calibration.html#calibration

Reinforcement Learning Tutorial. (χ.χ.). Ανάκτηση από Java T Point: https://www.javatpoint.com/reinforcement-learning

Rodriguez, I. L., Branchaud-Charron, Frederic, Keegan, L., Atighehchian, P., Parker, W., . . . Nowrouzezahrai, D. (2020, Ιουλίου 7). *A Weakly Supervised Region-Based Active Learning Method for COVID-19 Segmentation in CT Images*. Ανάκτηση από Cornell University: https://arxiv.org/abs/2007.07012

Ruiz, P. (2018, Οκτωβρίου 10). *Understanding and visualizing DenseNets*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

Ruiz, P. (2018, Οκτωβρίου 10). *Understanding and visualizing DenseNets*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

Saha, S. (2018, 15 Δεκεμβρίου). *A Comprehensive Guide to Convolutional Neural Networks* . Ανάκτηση από Towards Data Science: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Saha, S. (2018, Δεκεμβρίου 15). *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Santosh, K. C. (2020, Μαρτίου 18). AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. *Journal of Medical Systems (2020)*, σ. 44:93.

scikit-learn.org. (χ.χ.). *scikit-learn.org*. Ανάκτηση από https://scikit-learn.org/

Semi-supervised learning. (2022, Ιουλίου 10). Ανάκτηση από Wikipedeia: https://en.wikipedia.org/wiki/Semi-supervised_learning

Sethneha. (2021, Απριλίου 1). *Is Gradient Descent sufficient for Neural Network*. Ανάκτηση από https://www.analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2021/04/is-gradient-descent-sufficient-for-neural-network/

Sharma, S. (2017, Σεπτεμβρίου 06). *Activation Functions in Neural Networks*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

Sharma, S. (2017, Σεπτεμβρίου 7). *Activation Functions in Neural Networks*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

simplilearn. (2022, Ιουλίου 7). *The Best Introduction to Data Science*. Ανάκτηση από https://www.simplilearn.com/tutorials/data-science-tutorial/introduction-to-data-science

Simplilearn. (20222, Σεπτεμβριου 4). *What is Data Science: Lifecycle, Applications, Prerequisites and Tools*. Ανάκτηση από www.simplilearn.com: https://www.simplilearn.com/tutorials/data-science-tutorial/what-is-data-science

Singh, S. (2018, Μαίου 21). *Understanding the Bias-Variance Tradeoff*. Ανάκτηση από towords data science: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Singirikonda, M. (2020, Δεκεμβρίου 22). *How Padding helps in CNN*. Ανάκτηση από https://mahithas.medium.com/: https://mahithas.medium.com/how-padding-helps-in-cnn-2b87957e1b

Steen, D. (2020, Αυγόυστου 30). *A Gentle Introduction to Self-Training and Semi-Supervised Learning*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/a-gentle-introduction-to-self-training-and-semi-supervised-learning-ceee73178b38

TABLEAU SOFTWARE, LLC. (χ.χ.). *components ofquality data*. Ανάκτηση από https://www.tableau.com: https://www.tableau.com/learn/articles/what-is-data-cleaning

Talal S. Qaid, Mazaar, H., H, M. Y., Al-Shamri, Alqahtani, M. S., Raweh, A. A., & Alakwaa, W. (2021, Αυγούστου 6). Hybrid deep-learning and machine-learning models for predicting COVID-19. *Computational Intelligence and Neuroscience,*, σσ. vol. 2021, Article ID 9996737, 11 pages.

Towards AI. (2021, Δεκεμβριου 23). *towardsai*. Ανάκτηση από Mismatch-first Farthest-search in Active Learning: https://towardsai.net/p/l/mismatch-first-farthest-search-in-active-learning

Transfer learning and fine-tuning. (2022-06-08, Ιουνίου 08). Ανάκτηση από TensorFlow: https://www.tensorflow.org/tutorials/images/transfer_learning

Tsang, S.-H. (2018, Νοεμβρίου 25). *DenseNet — Dense Convolutional Network (Image Classification)*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803

Tsang, S.-H. (2018, Νοεμβρίου 25). *Review: DenseNet — Dense Convolutional Network (Image Classification)*. Ανάκτηση από towardsdatascience: https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803

UC Business Analytics. (χ.χ.). *Gradient Boosting Machines*. Ανάκτηση από UC Business Analytics R Programming Guide: http://uc-r.github.io/gbm_regression

Ucara, F., & Korkmazb, D. (2020, Ιούλιος). Medical Hypotheses. *COVID diagnosis-Net: deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images*, σσ. vol. 140, Article ID 109761.

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., & Pinheiro, P. R. (2020, Μαίου 14). CovidGAN: data augmentation using auxiliary classifier GAN for improved covid-19 detection. *IEEE Access*, σσ. vol. 8, pp. 91916–91923.

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., . . . Xu, B. (2021, Φεβρουαρίου 23). A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *European Radiology*, σσ. vol. 31, pp. 1–9.

Wang, W., Li, Y., Wang, X., Li, J., & Zhang, P. (2021, Μαίου 25). COVID-19 patients detection in chest X-ray images via MCFF-net. *13th International Conference on Advanced Computational Intelligence (ICACI), Wanzhou, China*, σσ. pp. 318–322.

What is Artificial Inteligence. (χ.χ.). Ανάκτηση από Business Standard: https://www.business-standard.com/about/what-is-artificial-intelligence#collapse

Wikipedia. (2021, Αυγούστου 23). *Νευρωνικό δίκτυο*. Ανάκτηση από Wikipedia: https://el.wikipedia.org/wiki/%CE%9D%CE%B5%CF%85%CF%81%CF%89%CE%BD%CE%B9%CE%BA%CF%8C_%CE%B4%CE%AF%CE%BA%CF%84%CF%85%CE%BF

Wolff, R. (2020, Αυγούστου 26). *5 Types of Classification Algorithms in Machine Learning*. Ανάκτηση από Monkey Learn: https://monkeylearn.com/blog/classification-algorithms/

Wood, T. (χ.χ.). *Convolutional Neural Network*. Ανάκτηση από Deep AI: https://deepai.org/machine-learning-glossary-and-terms/convolutional-neural-network

Wright., R. (2017, Νοεμβρίου 14 Nov 2017). *pu_learning/blob/master/baggingPU.py*. Ανάκτηση από github: https://github.com/roywright/pu_learning/blob/master/baggingPU.py

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., . . . Chen, Y. (2020, Δεκεμβρίου 6). A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, σσ. vol. 6, no. 10, pp. 1122–1129.

Yarowsky, D. (1995, Ιουνίου 26). Unsupervised word sense disambiguation rivaling supervised methods. *ACL '95: Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, σσ. 189–196. Ανάκτηση από ACL '95: Proceedings of the 33rd annual meeting on Association for Computational Linguistics.

Zammit, J., Fung, D. L., Liu, Q., Leung, C. K.-S., & Hu, P. (2022, Αυγούστου 7). *Semi-supervised COVID-19 CT image segmentation using deep generative models. BMC Bioinformatics*. Ανάκτηση από https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04878-6