Covid-19 / Symptoms Prediction - Big Data Mining

Φοίβος Τζάβελλος και Αθανάσιος Παπανικολάου

Εξόρυξη Δεδομένων 2020-21
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας, Βόλος
{ftzavellos,atpapanikolaou}@e-ce.uth.gr

Περίληψη Όσο προχωράει η επιστήμη των δεδομένων και πιο συγκεκριμένα η εξόρυξη τους τόσο προχωράνε και οι προβλέψεις πάνω σε κάθε επιστημονικό τομέα. Ένας από αυτούς τους τομείς που επωφελείται πάρα πολύ από την επιστήμη των προβλέψεων είναι η ιατρική. Τα ατελείωτου όγκου δεδομένα που παράγονται από τον κάθε ασθενή παγκόσμια μετατρέπονται, με γοργούς ρυθμούς και σταθερά βήματα, στις πιο εξελιγμένες προβλέψεις που μπορούν να καθοδηγήσουν κάθε ιατρό και ιατρικό προσωπικό στην σωστή πρόγνωση ακόμα και όταν τα εμφανή συμπτώματα του ασθενή δεν είναι αρκετά "για το γυμνό μάτι". Αυτό το τεράστιο ιστορικό παρελθοντικών ασθενών/ασθενειών είναι ικανό στο μέλλον να τροφοδοτεί τα μοντέλα προβλέψεων με τόσα πολλά δεδομένα ώστε να προβλέπεται το πρόβλημα υγείας του ασθενή με μεγάλη ακρίβεια χρησιμοποιώντας σαν είσοδο στο μοντέλο μόνο τα συμπτώματα του και πιθανώς μερικά προσωπικά χαρακτηριστικά.

Λέξεις Κλειδιά: covid19, συμπτώματα, προβλέψεις

1 Εισαγωγή

Κάποια μέρα στο μέλλον θα φτάσουμε στο σημείο να προβλέπουμε ασθένειες και πολλά άλλα δεδομένα με τέλειους αλγορίθμους και ελάχιστα λάθη. Δυστυχώς δεν έχουμε φτάσει ακόμα σε αυτό το επίπεδο αλλά με αυτό το στόχο προσπαθήσαμε σε αυτήν την εργασία να δείξουμε με ποιον τρόπο μπορούμε να προβλέψουμε την ασθένεια ενός ανθρώπου χρησημοποιώντας τα εμφανή συμπτώματα του και στο δεύτερο μέρος της εργασίας να παρουσιάσουμε δεδομένα για τον ιό covid-19 και να προβλεψουμε με παρόμοιο τρόπο αν θα ζήσει ο ασθενής με covid-19 χρησιμοποιώντας μερικά προσωπικά του στοιχεία.

2 Σχετική Βιβλιογραφία

Στην εργασία χρησιμοποιήθηκαν δωρεάν datasets από το https://www.kaggle.com/ [2] καθώς και από την ιστοδελίδα της Μεξικανικής Κυβέρνησης [1]. Το διαδίκτυο, και συγκεκριμένες ιστοσελίδες όπως το https://towardsdatascience.com/ παρέχουν πληθώρα υλικού από το οποίο αντλήσαμε ιδέες για το θέμα της εργασίας. Επίσης χρησιμοποιήθηκαν γνώσεις και ιδέες από τα μαθήματα του κυρίου Ηλία Χούστη, πιο συγκεκριμένα από τα μαθήματα "Περιβάλλοντα Επίλυσης Προβλημάτων για Εφαρμογές στην Επιστήμη Δεδομένων", "Machine Learning" και "Στατιστική".

3 Εργαλεία που χρησιμοποιήθηκαν

Και στα δύο κομμάτια της εργασίας χρησιμοποιήθηκε το Jupyter Notebook και η Python 3.0. Η Python σαν γλώσσα είναι αρκετά βολική για εμάς, ευκολόχρηστη αλλά και με την κατάλληλη πολυπλοκότητα ώστε να είμαστε σίγουροι για την εγκυρότητα των αποτελεσμάτων. Οι βιβλιοθήκες που παρέχει μας λύνουν τα χέρια . Μέσα στα παραδοτέα αρχεία με κώδικα αναφέρεται και ποιες βιβλιοθήκες χρησιμοποιήθηκαν συγκεκριμένα (sklearn, ggplot, seaborn, pandas, numpy και άλλες). Το Jupyter Notebook (επισυνάπτεται οδηγός χρήσης του για να τρέξει ο παραδοτέος κώδικας) είναι το μέσο που μας επιτρέπει να βγάλουμε αποτελέσματα με αποδοτικό τρόπο αλλα και να τα κάνουμε ευπαρουσίαστα.

4 Πρόβλεψη Ασθενειών

Σε αυτό το κομμάτι της εργασίας στόχος είναι η πρόβλεψη του προβλήματος υγείας του ασθενή χρησιμοποιώντας μόνο τα συμπτώματα του και για να επιτευχθεί αυτό εκπαιδεύουμε μοντέλα πολλών αλγορίθμων με παρελθοντικούς ασθενείς/ασθένειες με στόχο της παλινδρόμηση αλλά και την κατηγοριοποίηση.

4.1 Δεδομένα

Χρησιμοποιείται ένα μοναδικό dataset σε όλη αυτήν την διαδικασία το οποίο αποτελείται απο 4920 γραμμές και 133 στήλες εκ των οποίων η μία είναι η ασθένεια και οι άλλες 132 είναι τα πιθανά συμπτώματα. Κάθε στήλη συμπτώματος είναι γεμάτη με άσσους και μηδενικά συμβολίζοντας την ύπαρξη και την απουσία αντίστοιχα του συγκεκριμένου συμπτώματος στην αντίστοιχη περίπτωση ασθένειας, δηλαδή κάθε περίπτωση ασθένειας, δηλαδή κάθε γραμμή αποτελείται από την ασθένεια σαν όνομα και από 132 μηδενικά και άσσους.

4.2 Διαδικασία

Μετά από κατάλληλη επεξεργασία των δεδομένων χρησιμοποιούμε συνολικά 7 αλγόριθμους και 11 μοντέλα για την τελική πρόβλεψη. Πιο συγκεκριμένα χρησιμοποιούνται οι Linear Regression, Logistic Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Machine Regressor και K-Nearest-Neighbors Regressor από την βιβλιοθήκη sklearn με σκοπό την παλινδρόμηση και όσων αφορά την κατηγοριοποίηση χρησιμοποιούνται οι Naive Byes Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine Classifier και K-Nearest-Neighbors Classifier επίσης από την βιβλιοθήκη sklearn.

Linear Regression Η γραμμική παλινδρόμηση είναι ένας βασικός και κοινώς χρησιμοποιούμενος τύπος προγνωστικής ανάλυσης. Η συνολική ιδέα της παλινδρόμησης είναι να εξετάσουμε δύο πράγματα: (1) κάνει ένα σύνολο μεταβλητών πρόβλεψης καλή δουλειά στην πρόβλεψη μιας μεταβλητής έκβασης (εξαρτώμενη); και (2) ποιες μεταβλητές

ειδικότερα είναι σημαντικοί προγνωστικοί παράγοντες της μεταβλητής αποτελεσμάτων και με ποιον τρόπο επηρεάζουν τη μεταβλητή αποτελέσματος; Αυτές οι εκτιμήσεις παλινδρόμησης χρησιμοποιούνται για να εξηγήσουν τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών.

Logistic Regression Η λογιστική παλινδρόμηση είναι ένα στατιστικό μοντέλο που στη βασική του μορφή χρησιμοποιεί μια λογιστική συνάρτηση για τη μοντελοποίηση μιας δυαδικής εξαρτώμενης μεταβλητής, αν και υπάρχουν πολλές πιο περίπλοκες επεκτάσεις. Στην ανάλυση παλινδρόμησης, η λογιστική παλινδρόμηση εκτιμά τις παραμέτρους ενός λογιστικού μοντέλου (μια μορφή δυαδικής παλινδρόμησης). Μαθηματικά, ένα δυαδικό λογιστικό μοντέλο έχει μια εξαρτημένη μεταβλητή με δύο πιθανές τιμές, όπως pass / fail που αντιπροσωπεύεται από μια μεταβλητή δείκτη, όπου οι δύο τιμές φέρουν την ένδειξη ”0” και ”1”.

Decision Tree Το δέντρο απόφασης είναι το πιο ισχυρό και δημοφιλές εργαλείο ταξινόμησης και πρόβλεψης. Ένα δέντρο απόφασης είναι ένα διάγραμμα ροής σαν δομή δέντρου, όπου κάθε εσωτερικός κόμβος δηλώνει μια δοκιμή σε ένα χαρακτηριστικό, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα της δοκιμής και κάθε κόμβος φύλλων (τερματικός κόμβος) κρατά μια ετικέτα κλάσης.

Random Forest Το τυχαίο δάσος είναι ένας εποπτευόμενος αλγόριθμος μάθησης που χρησιμοποιείται τόσο για την ταξινόμηση όσο και για την παλινδρόμηση αλλά χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης. Ο αλγόριθμος δημιουργεί δέντρα αποφάσεων σε δείγματα δεδομένων και στη συνέχεια λαμβάνει την πρόφαση από καθένα από αυτά και τελικά επιλέγει την καλύτερη λύση μέσω ψηφοφορίας. Είναι μια μέθοδος σύνοψης που είναι καλύτερη από ένα δέντρο αποφάσεων, επειδή μειώνει την προσαρμογή κατά μέσο όρο του αποτελέσματος.

Support Vector Machine Οι μηχανές υποστήριξης-φορέα (SVM) είναι εποπτευόμενα μοντέλα μάθησης με συναφείς αλγόριθμους μάθησης που αναλύουν τα δεδομένα που χρησιμοποιούνται για την ταξινόμηση και την ανάλυση παλινδρόμησης, αλλά κυρίως για στόχους ταξινόμησης. Ο στόχος αυτού του αλγορίθμου είναι η εύρεση ενός υπερπλάνου σε έναν N-διαστατικό χώρο (N - ο αριθμός των χαρακτηριστικών) που ταξινομεί διακριτά τα σημεία δεδομένων. Σε αυτόν τον αλγόριθμο, προσπαθούμε να μεγιστοποιήσουμε το περιθώριο μεταξύ των σημείων δεδομένων και του υπερπλάνου.

K-Nearest-Neighbors Το K-Nearest Neighbours είναι μια τεχνική και αλγόριθμος μηχανικής εκμάθησης που μπορούν να χρησιμοποιηθούν τόσο για εργασίες παλινδρόμησης όσο και για ταξινόμηση. Εξετάζει τις ετικέτες ενός επιλεγμένου αριθμού σημείων δεδομένων που περιβάλλουν ένα σημείο δεδομένων προορισμού, προκειμένου να κάνει μια πρόβλεψη για την κλάση στην οποία εμπίπτει το σημείο δεδομένων. Είναι ένας εννοιολογικά απλός αλλά πολύ ισχυρός αλγόριθμος και για αυτούς τους λόγους, είναι ένας από τους πιο δημοφιλείς αλγόριθμους μηχανικής μάθησης.

Naive-Bayes Ο Naïve Bayes είναι ένας πιθανολογικός αλγόριθμος μηχανικής εκμάθησης που βασίζεται στο Θεώρημα Bayes, που χρησιμοποιείται σε μια μεγάλη ποικιλία εργασιών ταξινόμησης. Οι απλούστερες λύσεις είναι συνήθως οι πιο ισχυρές και ο Naïve Bayes είναι ένα καλό παράδειγμα. Παρά τις εξελίξεις στη Μηχανική Εκμάθηση τα τελευταία χρόνια, έχει αποδειχθεί όχι μόνο απλός αλλά και γρήγορος, ακριβής και αξιόπιστος. Έχει χρησιμοποιηθεί με επιτυχία για πολλούς σκοπούς, αλλά λειτουργεί ιδιαίτερα καλά με προβλήματα επεξεργασίας φυσικής γλώσσας (NLP). Το Naïve Bayes είναι ένας πιθανός αλγόριθμος μηχανικής εκμάθησης που βασίζεται στο Θεώρημα Bayes, που χρησιμοποιείται σε μια μεγάλη ποικιλία εργασιών ταξινόμησης.

4.3 Πειράματα και Αποτελέσματα

Δυστυχώς τα αποτελέσματα των μοντέλων δεν είναι πολύ χρήσιμα. Το dataset είναι γενικά μικρό και τα δεδομένα του είναι έτσι φτιαγμένα ώστε κάθε φορά να γίνεται η τέλεια πρόβλεψη με 100% ακρίβεια, με εξαίρεση την Linear Regression που είχε μικρή επιτυχία και το SVM Regressor το οποίο είχε πολύ καλή αλλά όχι και τέλεια ακρίβεια στην πρόβλεψη συγκριτικά με τα υπόλοιπα.

Στο τέλος έγινε και μία αναπαράσταση τύπου ”chat bot” στο οποίο ο ασθενής ή ο ιατρός μπορεί να γράψει μερικά συμπτώματα απαντώντας σε μερικές ερωτήσεις και να εμφανιστούν οι προβλέψεις και από τα 11 μοντέλα με σκοπό την καθοδήγηση προς την πραγματικά σωστή διάγνωση της ασθένειας. Δυστυχώς υπήρχαν αρκετές αποκλίσεις ανάμεσα στις 11 απαντήσεις, κάτι που είναι περίεργο όταν σχεδόν όλα τα μοντέλα έχουν 100% ακρίβεια στις προβλέψεις τους. Η αλήθεια όμως πίσω από αυτές τις διαφορετικές προβλέψεις κρύβεται στο ότι η τυχαία συμπλήρωση συμπτωμάτων στο chat bot απέχει πολύ από τα πολύ ειδικά δεδομένα του dataset τα οποία αντιστοιχίζονται με ειδικό τρόπο με τις ασθένειες.

4.4 Συμπεράσματα και πιθανά μελλοντικά σχέδια

Συμπερασματικά δεν είδαμε κάποιο ιδιαίτερο αποτέλεσμα λόγω των τέλειων προβλέψεων, αλλά έχει τεθεί η βάση για κάποιο μελλοντικό dataset, μεγαλύτερο και πιο ”απελευθερομένο” στην σχέση μεταξύ των συμπτωμάτων μεταξύ τους. Με ένα τέτοιο dataset θα μπορούσαμε να βλέπαμε πραγματικά αποτελέσματα στις προβλέψεις και το ”chat bot” να έδινε απαντήσεις παρόμοιες αν όχι ίδιες μεταξύ τους. Η εξέλιξη αυτής της ιδέας και προφανώς των datasets μία μέρα στο μέλλον θα γίνει ένα χρήσιμο και αξιόπιστο εργαλείο στα χέρια της ιατρικής.

Κλείνοντας, η χρήση νευρωνικών δικτύων πιστεύουμε ότι θα ήταν επίσης πολύ αποτελεσματική και χρήσιμη και θα παρήγαγε ενδιαφέροντα αποτελέσματα.

5 Πρόβλεψη COVID

Σε αυτό το κομμάτι της εργασίας ο στόχος είναι να χρησιμοποιήσουμε τα δεδομένα από ασθενείς (ηλικία, φύλο, ασθένειες) για να εξάγουμε κάποια συμπεράσματα τα οποία

παρουσιάζουμε σε γραφήματα, για να αναδείξουμε την σημαντικότητα συγκεκριμένων συμπτωμάτων στην αύξηση του ποσοστού θνησιμότητας καθώς και να προβλέψουμε το αν ο ασθενής θα πεθάνει με βάση τα χαρακτηριστικά στοιχεία του.

5.1 Δεδομένα

Όπως προαναφέραμε, χρησιμοποιούμε τα δεδομένα που συγκέντρωσε η Μεξικανική κυβέρνηση [1] σχετικά με τον κορονοϊό. Περιέχει 566.602 εισαγωγές (άνθρωποι) και 23 χαρακτηριστικά (φύλο, ηλικία, αν έχει έρθει σε επαφή με επιβεβαιωμένο κρούσμα, αν ειναι θετικός στον κορονοϊό, αν έχει κάποια ασθένεια όπως υπέρταση ή άσθμα και πολλά άλλα τα οποία αναλύονται αναλυτικά στον κώδικα).

5.2 Διαγράμματα

Μέσα από την κατάλληλη προεπεξεργασία των δεδομένων προκύπτουν συγκεκριμένα γραφήματα (ιστογράμματα, countplots, catplots) και εξάγονται κάποια ενδιαφέροντα συμπεράσματα, για παράδειγμα ότι η κατηγορία με την μεγαλύτερη θνησιμότητα είναι οι ηλικίες 75-90 και αμέσως μετά οι ασθενείς με χρόνια νεφρική νόσο.

5.3 Gradient Boosting

Το Gradient Boosting είναι μια διαδικασία της μηχανικής μάθησης που παράγει ένα μοντέλο προβλέψεων στην μορφή μιας συλλογής από αδύναμα μοντέλα πρόβλεψης, συνήθως δέντρα αποφάσεων, τα οποία συνδυάζει σε έναν δυνατό μοντέλο. Στην περιπτωσή μας το χρησιμοποιούμε για να αναδείξουμε την κρισιμότητα των συμπτωμάτων, με το ίδιο dataset απλά διαφορετική προεπεξεργασία του. Προκύπτει ότι η ηλικία είναι το χαρακτηριστικό που επηρεάζει το περισσότερο την κρισιμότητα θανάτου, και η πνευμονία την κρισιμότητα εισαγωγής σε ΜΕΘ και την κρισιμότητα διασωλήνωσης.

5.4 Πρόβλεψη Θανάτου

Με αυτή την κάπως νοσηρή επικεφαλίδα αναφερόμαστε στην πρόβλεψη θνησιμότητας ενός ασθενή με βάση τα χαρακτηριστικά του. Χρησιμοποιούμε πάλι το ίδιο dataset με διαφορετική προεπεξεργασία. Οι αλγόριθμοι που χρησιμοποιούνται είναι οι K Κοντινότεροι Γείτονες, τα Δέντρα Αποφάσεων, η Λογιστική Παλινδρόμηση, οι Μηχανές Υποστήριξης Φορέα, το Random Forest και το Gradient Boosting.

5.5 Πειράματα και Αποτελέσματα

Με βάση τα αποτελέσματα, την μεγαλύτερη ακρίβεια την πετυχαίνουν τα δεντρα αποφάσεων, και αμέσως μετά η λογιστική παλινδρόμηση και το Gradient Boosting, με όλα να κυμάινονται κοντά στο 80% οσων αφορά την απόδοση. Έχουμε φτίαξει ένα είδους chatbot όπως και στο προηγούμενο κομμάτι, στο οποίο ο χρήστης εισάγει σαν είσοδο τα χαρακτηριστικά του ασθενή και παίρνει σαν έξοδο 6 προβλέψεις (με βάση τα μοντέλα/αλγορίθμους που προαναφέρθηκαν) σχετικά με το αν ο ασθενής θα πεθάνει ή όχι. Σε κάποιες περιπτώσεις, ανάλογα και με την είσοδο που θα βάλει ο χρήστης, υπάρχει ασυμφωνία στα αποτελέσματα μεταξύ των αλγορίθμων, κάτι το οποίο ίσως να οφείλεται στον διαφορετικό τρόπο με τον οποίο ”μαθαίνει” το κάθε μοντέλο.

5.6 Συμπεράσματα και πιθανά μελλοντικά σχέδια

Σαν συμπέρασμα θα πούμε ότι ίσως το θέμα της πρόβλεψης του αν ένας θετικός στον ιό θα ζήσει είναι τόσο πολυσύνθετο και σημαντικό που δεν θα ήμασταν καλυμμένοι απλά και μόνο από τα αποτελέσματα που προκύπτουν μέσα από τα συγκεκριμένα μοντέλα μηχανικής μάθησης. Δεν λαμβάνονται υπόψιν παράγοντες όπως η πληρότητα των ΜΕΘ στην περιοχή ή η φροντίδα που θα του παρέχει το ιατρικό προσωπικό. Ένα πιο ανεπτυγμένο μοντέλο με κάποιες συγκεκριμένες βελτιστοποιήσεις ίσως να είχε και καλύτερη απόδοση.

Αναφορές

1. Datos Abiertos - Dirección General de Epidemiología — Secretaría de Salud — Gobierno — gob.mx, (n.d.).
   https://www.gob.mx/salud/documentos/datos-abiertos-152127
2. https://www.kaggle.com/neelima98/disease-prediction-using-machine-learning?fbclid=IwAR0dBjUNRlGtVQtNAP4x3cNSH3lDw$_V EtgT2QF84TvJ8jI24mtEzmCsO8z4$