Πανεπιστήμιο Πειραιώς
Τμήμα Ψηφιακών Συστημάτων
Π.Μ.Σ "Πληροφοριακά Συστήματα και Υπηρεσίες"
Ειδίκευση: Μεγάλα Δεδομένα και Αναλυτική

«Ανάλυση Δεδομένων COVID-19 με Συστήματα Γεωγραφικών Πληροφοριών»

ΧΑΤΖΗΔΙΑΚΟΥ ΕΛΕΝΗ

Αθήνα, Μάρτιος 2022

Πανεπιστήμιο Πειραιώς
Τμήμα Ψηφιακών Συστημάτων
Π.Μ.Σ "Πληροφοριακά Συστήματα και Υπηρεσίες"
Ειδίκευση: Μεγάλα Δεδομένα και Αναλυτική

Επιβλέπων Καθηγητής: Φιλιππάκης Μιχαήλ

Τριμελής Εξεταστική Επιτροπή

Κυριαζής Δημοσθένης

Φιλιππάκης Μιχαήλ

Χαλκίδη Μαρία

«Ανάλυση Δεδομένων COVID-19 με Συστήματα Γεωγραφικών Πληροφοριών»

ΧΑΤΖΗΔΙΑΚΟΥ ΕΛΕΝΗ

Περιεχόμενα

Περίληψη

Σε αυτή την εργασία θα μελετήσουμε τη γεωγραφική εξάπλωση του κορονοϊού. Ως πηγή δεδομένων για τον κορονοϊό είναι οι περιπτώσεις μολύνσεων και θάνατοι ανά ημέρα. Έχουν χρησιμοποιηθεί δεδομένα που περιέχουν αυτή τη στιγμή περίπου 60000 εγγραφές, οπότε είναι κατάλληλα για εφαρμογή τεχνικών ανάλυσης μεγάλων δεδομένων. Ακόμα μελετήθηκαν χωρικά δεδομένα όπως για παράδειγμα κάποια βασικά γεωγραφικά και οικονομικά δεδομένα ανά χώρα (τύπου πληθυσμός, πυκνότητα, ΑΕΠ κλπ) και δημογραφικά δεδομένα (πληθυσμός, πυκνότητα, ηλικία κλπ) για το 2020.

Επιπλέον, αναλύονται χωρικά δεδομένα διαφορετικού τύπου, όπως για παράδειγμα σχετικά με τον καιρό (π.χ. θερμοκρασία, βροχή). Τα μετεωρολογικά δεδομένα είναι για το 2020. Τα παραπάνω δεδομένα θα συνδυαστούν και θα αποτελέσουν χαρακτηριστικά προς ταξινόμηση. Η μεταβλητή στόχος θα είναι η ημερήσια αύξηση του κορονοϊού. Θα χωρίσουμε την αύξηση σε διακριτές κατηγορίες.

Στη συνέχεια θα εκπαιδεύσουμε διάφορους αλγορίθμους, για παράδειγμα K-nearest-neighbors, SVM (support vector machines), δέντρα αποφάσεων (Decision Tree) και Τυχαίο Δάσος (Random Forest). Θα γίνει 10-fold cross-validation, ώστε να χωρίσουμε τα δεδομένα σε υποσύνολα εκπαίδευσης και δοκιμής. Επίσης, θα δοκιμάσουμε διάφορες παραμέτρους, ώστε να βελτιστοποιήσουμε τα αποτελέσματα του κάθε αλγορίθμου. Με βάση τα δεδομένα δοκιμής θα αξιολογήσουμε τους ταξινομητές με διάφορες μετρικές, όπως η ακρίβεια, η ευαισθησία και η εξειδίκευση για να δούμε ποιος αλγόριθμος δίνει το καλύτερο αποτέλεσμα.

Τα γεωγραφικά δεδομένα για τον κορονοϊό που συγκεντρώθηκαν προηγουμένως θα γίνει απεικόνισή τους σε χάρτη. Επίσης, μπορεί να γίνει απεικόνιση σε χάρτη της συσχέτισης των οικονομικών, δημογραφικών και μετεωρολογικών δεδομένων με την εξάπλωση του ιού. Λαμβάνοντας υπόψη τη φύση των δεδομένων, η ανάλυση και η απεικόνιση θα γίνει ανά χώρα.

Abstract

In this research we will study the geographical spread of the coronavirus. The coronavirus source of data are the cases of infections and deaths per day. Data that currently contains about 60,000 records have been used, making them suitable for the application of big data analysis techniques. Spatial data were also studied, such as some basic geographical and economic data by country (type of population, density, GDP, etc.) and demographic data (population, density, age, etc.) for 2020.

In addition, spatial data of different types are analyzed, such as the weather (e.g., temperature, rain). The meteorological data is for 2020. The data will be combined and the characteristics will be classified. The variable goal will be the daily increase of the coronavirus. We will divide the increase into distinct categories.

Next, we are going to train various algorithms, for example K-nearest-neighbors, SVM (support vector machines), Decision Tree and Random Forest. There will be 10-fold cross-validation, so that we can divide the data into training and test subsets. Furthermore, we will test various parameters to optimize the results of each algorithm. Based on the test data we will evaluate the classifiers with various metrics such as accuracy, sensitivity and expertise to see which algorithm gives the best result.

The geographical data that was collected before for the coronavirus, will be displayed on a map. The correlation of economic, demographic and meteorological data with the spread of the virus can also be mapped. Taking into account the nature of the data, the analysis and display will be done by country.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου κ. Φιλιππάκη Μιχαήλ για την ανεκτίμητη βοήθεια του και για τον πολύτιμο χρόνο που αφιέρωσε τόσο στην επίβλεψη όσο και στην καθοδήγηση για την υλοποίηση της διπλωματικής.

Επίσης, θα ήθελα να ευχαριστήσω τη ΔΡ . Μαρία Ελένη Πούλου για την πολύτιμη βοήθεια της στην επίβλεψη της διπλωματικής.

Επιπρόσθετα, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένεια μου για όλες τις θυσίες , τους κόπους και την συνεχή συμπαράσταση και καθοδήγηση τους όλα τα χρόνια της σταδιοδρομίας μου.

1. Εισαγωγή

Ένα Γεωγραφικό Πληροφοριακό Σύστημα (Geographic Information System) είναι ένα υπολογιστικό σύστημα το οποίο επιτρέπει τη συλλογή, διαχείριση, ανάλυση και παρουσίαση χωρικών δεδομένων (spatial data) (Worboys & Duckham, 2004; Burrough et al., 2015).

Στην καρδιά ενός τέτοιου συστήματος βρίσκεται μια βάση δεδομένων, η οποία περιέχει τα χωρικά και άλλα σχετιζόμενα δεδομένα και δίνει τη δυνατότητα πρόσβασης των επιθυμητών δεδομένων στο χρήστη. Φυσικά, όπως σε κάθε σύστημα, υπάρχει το υλικό (hardware) του συστήματος και κάποια δικτυακή υποδομή για πρόσβαση των χρηστών και το λογισμικό (software), το οποίο περιλαμβάνει τη βάση δεδομένων και άλλες εφαρμογές για την ανάλυση και την απεικόνιση.

Τα χωρικά δεδομένα που προαναφέρθηκαν έχουν ιδιότητες τοποθεσίας, για παράδειγμα γεωγραφικές συντεταγμένες καθώς και μη γεωγραφικές ιδιότητες. Τα δεδομένα για να αξιοποιηθούν πρέπει να οριστεί κάποιο μοντέλο και μπορεί να χρειαστούν διάφορους μετασχηματισμούς για να λάβουμε την τελική πληροφορία. Εκτός από τις διάφορες ιδιότητες, υπάρχει και η σχέση μεταξύ διαφορετικών δεδομένων, η οποία ονομάζεται τοπολογία των δεδομένων.

2. Θεωρία Κορονοϊού

Οι κοροναϊοί είναι μια ομάδα μεγάλων, περιτυλιγμένων, μονόκλωνων ιών RNA με θετική αίσθηση που ανήκουν στην τάξη Nidovirales, οικογένεια Coronaviridae, υποοικογένεια Coronavirinae. Εικοσι εξι είναι γνωστά διάφορα είδη (Cleri, D.J. et al (2010)) και έχουν χωριστεί σε τέσσερα γένη (άλφα, βήτα, γάμα και δέλτα) που χαρακτηρίζονται από διαφορετική αντιγονική διασταυρούμενη αντιδραστικότητα και γενετική. Μόνο τα γένη άλφα και betacoronavirus περιλαμβάνουν στελέχη παθογόνα για τον άνθρωπο (Paules, C.I. et al (2020)).

Ο πρώτος γνωστός κοροναϊός, ο ιός της μολυσματικής βρογχίτιδας των πτηνών, απομονώθηκε το 1937 και ήταν την αιτία καταστροφικών λοιμώξεων στο κοτόπουλο. Ο πρώτος ανθρώπινος κορονοϊός απομονώθηκε από το ρινική κοιλότητα και πολλαπλασιάζεται σε ανθρώπινα ακτινωτά κύτταρα εμβρυϊκής τραχείας in vitro (σε δοκιμαστικό σωλήνα) από τον Tyrrell και Bynoe το 1965. Ωστόσο, οι κορονοϊοί υπήρχαν στον άνθρωπο για τουλάχιστον 500-800 χρόνια, και όλα προέρχονται από νυχτερίδες (Chan, P.K. et al (2013); Berry, M. et al (2015))

Οι κορονοϊοί έχουν αναγνωριστεί εδώ και καιρό ως σημαντικοί κτηνιατρικοί παθογόνοι παράγοντες, που προκαλούν αναπνευστικές και εντερικές ασθένειες σε θηλαστικά καθώς και σε πτηνά. Από τον γνωστό κορονοϊό είδη, μόνο έξι είναι γνωστό ότι προκαλούν ασθένειες στους ανθρώπους: HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1, σοβαρό κορονοϊό οξέος αναπνευστικού συνδρόμου (SARS-CoV) και Κοροναϊός του αναπνευστικού ιού της Μέσης Ανατολής (MERS-CoV) (Arabi, Y.M. et al (2017) · Skariyachan, S. et al (2019)). Τα τέσσερα πρώτα είναι ενδημικά έχουν συσχετιστεί κυρίως με ήπια, αυτοπεριοριζόμενη νόσο, ενώ οι δύο τελευταίες μπορούν να προκαλέσουν σοβαρή ασθένεια (Zumla, A. et al (2016) · Paules, C.I. et al (2020)). Οι SARS-CoV και MERS-CoV είναι betacoronaviruses (Zumla, A. et al (2015)), και είναι μεταξύ των παθογόνων που περιλαμβάνονται στον κατάλογο των απειλών υψηλής προτεραιότητας του Παγκόσμιου Οργανισμού Υγείας (Σχέδιο έρευνας και ανάπτυξης για δράση για την πρόληψη επιδημιών (Παγκόσμια Υγεία Οργάνωση, αναθεωρημένος Φεβρουάριος 2018)). Δεδομένης της μεγάλης επικράτησης και της ευρείας κατανομής των κορωνοϊών, της μεγάλης γενετικής τους ποικιλομορφίας καθώς και ο συχνός ανασυνδυασμός των γονιδιωμάτων τους και η αυξανόμενη δραστηριότητα στην επαφή των ανθρώπων, αυτοί οι ιοί αντιπροσωπεύουν μια συνεχή απειλή για την ανθρώπινη υγεία (Hui, D.S. et al. (2020) ·). Αυτό το γεγονός έγινε ξανά εμφανές στα τέλη του 2019 και στις αρχές του 2020, όταν ένας νέος κορονοϊός αποκαλύφθηκε ότι ήταν η αιτία μιας μεγάλης και ραγδαίας εξάπλωσης της αναπνευστικής νόσου,

συμπεριλαμβανομένης της πνευμονίας, στη Γουχάν της Κίνας (δήλωση του ΠΟΥ σχετικά με τη συστάδα περιπτώσεων πνευμονίας στο Γουχάν της Κίνας (Παγκόσμιος Οργανισμός Υγείας, 9 Ιανουαρίου 2020). Ο ιός, προσωρινά οριζόμενος 2019-nCoV, απομονώθηκε και προσδιορίστηκε η αλληλουχία του γονιδιώματος του ιού. Το 2019-nCoV χαρακτηρίστηκε ως a beta coronavirus, και έτσι έγινε το έβδομο ξεχωριστό είδος κορονοϊού ικανό να προκαλέσει ανθρώπινη ασθένεια.
Ο Παγκόσμιος Οργανισμός Υγείας, στις 30 Ιανουαρίου 2020, ανακοίνωσε το παγκόσμιο ξέσπασμα COVID-19 ως έκτακτη ανάγκη διεθνούς υγείας.

2.1 Ιστορική Εξέλιξη των Συστημάτων Γεωγραφικών Πληροφοριών

Η ιστορία των GIS ξεκίνησε συγχρόνως στη Βόρεια Αμερική, την Ευρώπη και την Αυστραλία. Στα μέσα της δεκαετίας του 1960, η εκροή μεγεθών απέφερε την προτροπή του πρώτου πραγματικού συστήματος GIS του Καναδικού Συστήματος Γεωγραφικών Πληροφοριών ή CGIS (Canada Geographic Information System). Στο τέλος της δεκαετίας του 1960, μια όξυνση νεωτερισμού έγινε στην Απογραφική Υπηρεσία των Η.Π.Α.(US Bureau of the Census) . Το Εργαστήριο Γραφικών Υπολογιστών και Χωρικής Ανάλυσης του Πανεπιστημίου του Χαρβαρντ εντόπισε την ομοιογένεια της τεχνολογίας αυτής με του συστήματος CGIS και ανέπτυξε ένα σύστημα γενικής χρήσης το οποίο να συμβάλει και στις δυο ανάγκες ταυτόχρονα. Έτσι κατέληξε στην υλοποίηση του ODYSSEY GIS, στο τέλος της δεκαετίας του 1970.
Οι ανάγκες για μείωση τόσο του χρόνου όσο και του κόστους είχαν αυξηθεί στο δεύτερο μισό της δεκαετίας του 1960. Έτσι, το 1968, η Ομάδα Πειραματικής Χαρτογραφίας του Ηνωμένου Βασιλείου ξεκίνησε τη χαρτογράφηση υψηλής ποιότητας με τη χρήση υπολογιστή. Σε συνεργασία με τη Βρετανική Υπηρεσία Γεωλογικής Τοπογραφίας, το 1973, δημοσιεύτηκαν χάρτες σχεδιασμένοι από υπολογιστή.
Τη δεκαετία του 1960 είχαν ενισχυθεί τα πρώτα συστήματα αυτοματοποιημένης χαρτογραφίας και μέχρι το τέλος της δεκαετίας του 1970 είχαν γίνει μηχανογραφημένες. Αυτό αποτέλεσε να γίνει το 1995 η πλήρης κάλυψη της Μεγάλης Βρετανίας με ψηφιακούς χάρτες αποθηκευμένους σε μία βάση δεδομένων. Ωστόσο και η τηλεπισκόπηση αποτελεί ουσιώδη ρόλο στην άνθηση των GIS, από τη δεκαετία του 1950, με τους πρώτους στρατιωτικούς δορυφόρους στη συλλογή πληροφοριών.

3. Συστήματα Γεωγραφικών Πληροφοριών

Ο Nick Chrisman απέδωσε των ορισμό των συστημάτων γεωγραφικών πληροφοριών ως τελικό επακόλουθο των ορισμών : Η οργανωμένη δραστηριότητα με την οποία οι άνθρωποι:

1. Μετρούν τις εκδηλώσεις των γεωγραφικών φαινομένων και διεργασιών·
2. Απεικονίζουν αυτές τις μετρήσεις, συνήθως με τη μορφή βάσεων δεδομένων, για να δώσουν έμφαση σε χωρικά θέματα, οντότητες, και σχέσεις μεταξύ τους·
3. Επεξεργάζονται αυτές τις αναπαραστάσεις για να δημιουργήσουν περισσότερες μετρήσεις και να ανακαλύψουν νέες σχέσεις ενοποιώντας νέες πηγές· και
4. Μετασχηματίζουν αυτές τις αναπαραστάσεις ώστε να προσαρμόζονται σε άλλα πλαίσια οντοτήτων και συσχετίσεων.

Αυτές οι δραστηριότητες αντανακλούν την έκταση του πλαισίου (θεσμικό και πολιτισμικό) όπου αυτοί οι άνθρωποι υλοποιούν τις εργασίες τους. Ως επακόλουθο, τα συστήματα GIS μπορεί να επηρεάσουν αυτές τις δομές. (Chrisman 2003, σελίδα 13)

Σε ένα ΣΓΠ τα στάδια επεξεργασίας

Εικόνα. Τα στάδια Επεξεργασίας σε ένα ΣΓΠ

Πηγή: Αρχές και Εφαρμογές Δορυφορικής Τηλεπισκόπησης

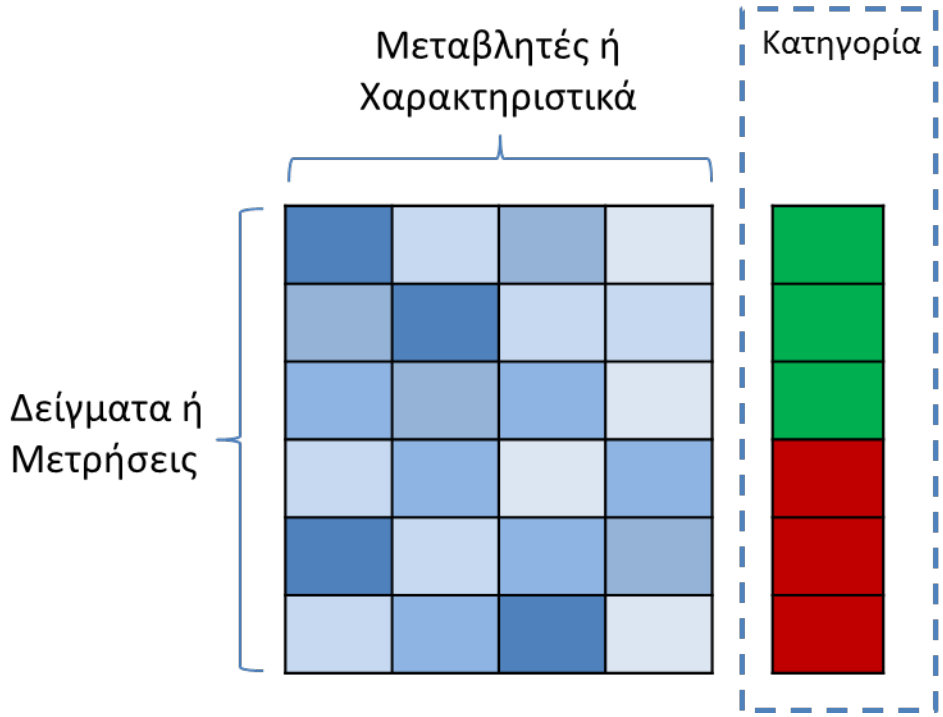
4. Ταξινόμηση δεδομένων Κορονοϊού

Σε αυτό το κεφάλαιο θα γίνει μια εισαγωγή στο πρόβλημα της ταξινόμησης και θα παρουσιάσουμε θεωρητικά τρεις μεθόδους μηχανικής μάθησης (KNN, SVM, Random Forest) για τη λύση αυτού του προβλήματος. Μετά, θα γίνει εφαρμογή στην περίπτωση της εξάπλωσης του κορονοϊού. Θα χρησιμοποιήσουμε δεδομένα για την εξάπλωσή του, καθώς και δημογραφικά δεδομένα και δεδομένα καιρού. Πρώτα θα γίνει μια γεωγραφική απεικόνιση των δεδομένων, ώστε να κατανοήσουμε καλύτερα τα δεδομένα που συλλέξαμε. Στη συνέχεια θα γίνει εκπαίδευση τριών μοντέλων με βάση τις τρεις μεθόδους που προαναφέραμε και θα γίνει σύγκριση των αποτελεσμάτων τους ως προς τη δυνατότητα πρόβλεψης του αριθμού των κρουσμάτων σε τρία επίπεδα (όχι μετάδοση, χαμηλή μετάδοση, υψηλή μετάδοση).

Εισαγωγή στο πρόβλημα της ταξινόμησης

Μια βασική κατηγορία προβλημάτων της μηχανικής μάθησης είναι η ταξινόμηση. Τα δεδομένα συνήθως έχουν μορφή πίνακα, όπου οι γραμμές αντιστοιχούν σε ένα δείγμα ή μέτρηση, ενώ οι στήλες είναι οι μεταβλητές ή τα χαρακτηριστικά που παίρνουμε σε κάθε μέτρηση. Οι μεταβλητές χωρίζονται σε ποιοτικές και ποσοτικές μεταβλητές. Οι ποιοτικές μεταβλητές διακρίνονται σε διατάξιμες και ονομαστικές. Οι ποσοτικές μεταβλητές διακρίνονται σε μεταβλητές διαστηματικής κλίμακας, κλίμακας λόγου καθώς και σε συνεχείς και διακριτές. Οι συνεχείς μεταβλητές περιλαμβάνουν οποιαδήποτε τιμή πραγματικών αριθμών ενώ οι διακριτές μεταβλητές περιέχουν μόνο ακέραιες τιμές οι οποίες δε σχετίζονται μεταξύ τους κατά καθορισμένες ποσότητες.

Στην περίπτωση της ταξινόμησης, κάθε δείγμα ανήκει σε μια διακριτή κατηγορία ή κλάση (Εικόνα 1). Αν έχουμε δύο κατηγορίες, η ταξινόμηση χαρακτηρίζεται ως δυαδική (binary), αλλιώς για παραπάνω κατηγορίες πολλαπλών κλάσεων (multi-class). Οπότε ο σκοπός είναι να δημιουργήσουμε ένα μοντέλο μηχανικής μάθησης, το οποίο θα μπορεί από τις μεταβλητές να μάθει τη σχέση τους με την κατηγορία. Έτσι, για νέα δείγματα άγνωστης κατηγορίας θα μπορεί να πραγματοποιήσει πρόβλεψη της κατηγορίας. Επειδή στην ταξινόμηση χρειαζόμαστε ένα σύνολο δεδομένων με γνωστές κατηγορίες για την δημιουργία ενός μοντέλου, το πρόβλημα της ταξινόμησης χαρακτηρίζεται ως επιβλεπόμενη μάθηση (supervised learning). Οπότε, έχουμε δύο φάσεις, τη φάση εκπαίδευσης (train), όπου δημιουργείται ένα μοντέλο και τη φάση δοκιμής (test), όπου πραγματοποιείται η πρόβλεψη.



Εικόνα 1. Δεδομένα για το πρόβλημα της ταξινόμησης.

Πηγή: Ιδίας Επεξεργασίας

Επειδή συνήθως ένα σύνολο δεδομένων, για να αξιολογήσουμε την απόδοση ενός αλγορίθμου ταξινόμησης, πραγματοποιούμε εκπαίδευση και πρόβλεψη με 10-fold cross-validation. Δηλαδή, θα διαιρέσουμε τα δεδομένα σε 10 τμήματα και σε 10 επαναλήψεις τα 9 τμήματα (90% των δεδομένων) θα χρησιμοποιηθούν για εκπαίδευση και το άλλο τμήμα (10% των δεδομένων) για δοκιμή. Με τα δεδομένα εκπαίδευσης δημιουργείται ένα μοντέλο και χρησιμοποιείται πάνω στα δεδομένα δοκιμής για πρόβλεψη της κατηγορίας. Μετά, συγκρίνουμε την πρόβλεψη με τις πραγματικές τιμές και υπολογίζουμε την ακρίβεια. Η διαδικασία επαναλαμβάνεται 10 φορές, ώστε όλα τα δεδομένα να χρησιμοποιηθούν ως δεδομένα δοκιμής και παίρνουμε στο τέλος τη μέση ακρίβεια των 10 δοκιμών. Αυτό φαίνεται σχηματικά στην εικόνα 2.

Εικόνα 2. Η διαδικασία 10-fold cross-validation.

Πηγή: [1].

5. Αλγόριθμοι Μηχανικής Μάθησης για ταξινόμηση

5.1 KNN

Ο αλγόριθμος K-πλησιέστερων γειτόνων (KNN) είναι ένας τύπος προβλεπόμενων αλγορίθμων εκμάθησης μηχανών, ο οποίος παρουσιάστηκε για πρώτη φορά το 1951 από τους Fix και Hodges [1]. Ο KNN στη βασική του μορφή εφαρμόζεται με κατανοητό τρόπο διότι δεν έχει μια εξειδικευμένη φάση κατάρτισης αλλά ταυτόχρονα εκτελεί πολύπλοκες εργασίες ταξινόμησης. Απεναντίας, χειρίζεται όλα τα δεδομένα για εκπαίδευση κατά την ταξινόμηση ενός νέου σημείου δεδομένων ή στιγμιότυπου.

Ο KNN είναι ένας αλγόριθμος εκμάθησης χωρίς παραμέτρους, που υποδηλώνει ότι δεν κάνει υποθέσεις για τα υποκείμενα δεδομένα. Αυτό είναι ένα εξαιρετικά ωφέλιμο γνώρισμα, δεδομένου ότι στη πλειοψηφία των δεδομένων του πραγματικού κόσμου δεν ακολουθούν πραγματικά καμία θεωρητική υπόθεση, για παράδειγμα γραμμική διαχωρισιμότητα, ομοιόμορφη κατανομή και άλλα.

Πιο συγκεκριμένα, για κάθε άγνωστο δείγμα, ο KNN βρίσκει τους K πλησιέστερους γείτονες με χρήση κάποιας μετρικής απόστασης, όπως η Ευκλείδεια. Στη συνέχεια, βρίσκει ποια κλάση έχει η πλειοψηφία των γειτόνων και την αναθέτει στο άγνωστο δείγμα.

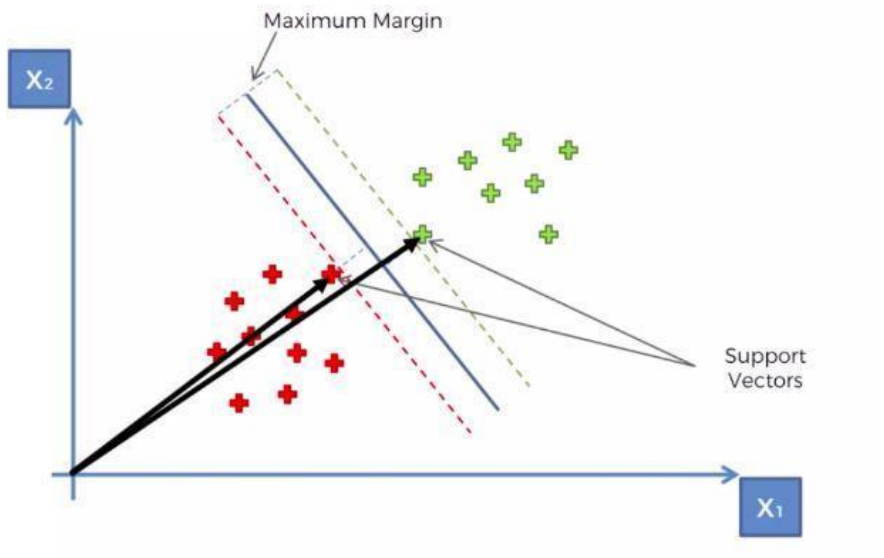
5.2.Support Vector Machine (SVM)

Ο Support Vector Machine (SVM) είναι ένας αλγόριθμος μηχανικής μάθησης, ο οποίος έχει την ικανότητα να εφαρμόζει τεχνικές ταξινόμησης, παλινδρόμησης ακόμη και ανίχνευση εξωστρέφειας [2]. Ο γραμμικός ταξινομητής SVM (Linear SVM - LSVM) λειτουργεί απεικονίζοντας μια ευθεία γραμμή μεταξύ δύο κατηγοριών (Εικόνα 3). Διακρίνεται σε δύο κλάσεις, με τη πρώτη κλάση να εμπεριέχει όλα τα σημεία δεδομένων που εμπίπτουν στη μία πλευρά της γραμμής και τη δεύτερη στην οποία επισημαίνονται όλα τα σημεία που εμπίπτουν στην άλλη πλευρά. Η επιλογή των δύο κλάσεων δεν είναι τόσο απλή διότι υφίστανται ένας άπειρος αριθμός γραμμών. Η βέλτιστη γραμμή της ταξινόμησης των δεδομένων αποδίδεται με το LSVM αλγόριθμο. Ο αλγόριθμος LSVM θα επιλέξει μια γραμμή που όχι μόνο χωρίζει τις δύο κλάσεις αλλά παραμένει όσο πιο μακριά γίνεται από τα κοντινά δείγματα. Ο αλγόριθμος αυτός είναι κατάλληλος για δυαδική ταξινόμηση μόνο, αλλά μπορεί να επεκταθεί σε ταξινόμηση πολλαπλών κλάσεων. Υπάρχουν δύο στρατηγικές, είτε θα δημιουργήσουμε ένα μοντέλο για κάθε κατηγορία, που θα ξεχωρίζει την κάθε κατηγορία από τις υπόλοιπες (one-versus-all) είτε θα δημιουργήσουμε

ένα μοντέλο για κάθε ζεύγος κατηγοριών (one-versus-one). Άλλες παραλλαγές του SVM χρησιμοποιούν τον πυρήνα radial basis function (RBF) για να διαχωρίσουν τις κλάσεις αντί για απλή γραμμή. Η εξίσωση του πυρήνα RBF είναι:

$$K(x, x') = e^{-\gamma\|x-x'\|}$$

όπου x και x' είναι τα διανύσματα δύο δειγμάτων.

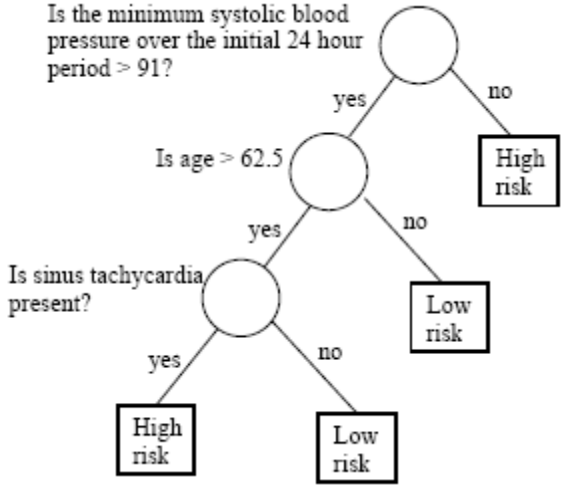


Εικόνα 3. Ο αλγόριθμος SVM
Πηγή:towardsdatascience.com

5.3. Δέντρα Απόφασης και Random Forest

Ένα δένδρο απόφασης δημιουργείται ως εξής: σε κάθε κόμβο δοκιμάζουμε για κάθε μεταβλητή ποια τιμή της διαχωρίζει καλύτερα τα δεδομένα. Οπότε διαλέγουμε την μεταβλητή και την τιμή με το καλύτερο αποτέλεσμα. Τα δεδομένα με τιμή της μεταβλητής μικρότερη από το όριο πάνε δεξιά, ενώ με μεγαλύτερη αριστερά και η διαδικασία επαναλαμβάνεται σε κάθε κόμβο παιδί.
Για να γίνει πιο κατανοητός ο αλγόριθμος, παρουσιάζουμε ένα απλό παράδειγμα δένδρου αποφάσεων στην εικόνα 4. Ο σκοπός είναι να αποφασίσουμε τη βαρύτητα της κατάστασης ενός ασθενή με βάση τις μεταβλητές συστολική πίεση, ηλικία, ύπαρξη ταχυκαρδίας όπως καταγράφηκαν το πρώτο 24ώρο νοσηλείας. Πρώτα εξετάζουμε την συστολική πίεση και αν είναι κάτω από 91, είναι υψηλού κινδύνου. Αλλιώς εξετάζουμε την ηλικία και αν είναι κάτω των 62,5

ετών, είναι χαμηλού κινδύνου. Αν είναι άνω των 62,5 ετών, τότε εξετάζουμε αν έχει ταχυκαρδία. Αν ναι, ανήκει στην κατηγορία υψηλού κινδύνου, αλλιώς χαμηλού κινδύνου.



Εικόνα 4. Παράδειγμα δέντρου αποφάσεων
Πηγή: https://online.stat.psu.edu/stat508/lesson/11

Επειδή τα δέντρα απόφασης είναι σχετικά αδύναμος αλγόριθμος, αναπτύχθηκε ο αλγόριθμος Random Forest [4], ο οποίος χρησιμοποιεί πολλά δέντρα για να εξάγει μια καλύτερη πρόβλεψη. Ο λόγος που Δέντρα Απόφασης λειτουργούν τόσο καλά είναι στην ύπαρξη μεγάλου αριθμού σχετικά μη σχετιζόμενων μοντέλων (δένδρων) που λειτουργούν ως επιτροπή για το πως θα ξεπεράσει κάθε ένα από τα επιμέρους συστατικά μοντέλα.

Έτσι, για τα Δέντρα Απόφασης οι προϋποθέσεις που αποδίδουν καλά είναι:

1. Πρέπει να υπάρχει κάποιο πραγματικό σήμα στις δυνατότητές μας, ώστε τα μοντέλα που έχουν κατασκευαστεί με αυτά τα χαρακτηριστικά να είναι καλύτερα από τυχαία εικασία.
2. Οι προβλέψεις (και κατά συνέπεια τα σφάλματα) που γίνονται από τα μεμονωμένα δέντρα πρέπει να έχουν χαμηλές συσχετίσεις μεταξύ τους.

Πιο συγκεκριμένα, αυτός ο αλγόριθμος εισάγει τυχαιότητα, ώστε τα δέντρα να είναι διαφορετικά μεταξύ τους και να μπορούν να αποφύγουν να «παγιδευτούν» σε κάποιο τοπικό ελάχιστο (σε κάποια μη βέλτιστη απόφαση). Συγκεκριμένα, σε κάθε δέντρο δίνεται τυχαία ένα υποσύνολο των δειγμάτων (περίπου το 67%) ενώ σε κάθε κόμβο εξετάζουμε τυχαία μόνο ένα υποσύνολο των μεταβλητών (καθορίζεται από μια παράμετρο με το όνομα mtry). Κάθε μεμονωμένο δέντρο στο

τυχαίο δάσος δίνει μια πρόβλεψη κλάσης και η τάξη με τις περισσότερες ψήφους γίνεται πρόβλεψη του μοντέλου.

6. Επιδημιολογική και Χωρική Μοντελοποίηση με Χρήση Τεχνητής Νοημοσύνης

6.1. Κοινωνικο-δημογραφικοί παράγοντες που σχετίζονται με λοιμώδη νοσήματα

Όπως έχει διαπιστωθεί, μία λοιμώδης νόσος δεν κάνει διακρίσεις και μπορεί να επηρεάσει οποιονδήποτε. Ωστόσο, κοινωνικο-δημογραφικοί παράγοντες, όπως η φυλή, το φύλο, η ηλικία και η φτώχεια μπορούν να οδηγήσουν σε δυσανάλογη νοσηρότητα ορισμένων ομάδων. Η αναγνώριση των παραγόντων κινδύνου εκείνων που μπορούν να οδηγήσουν σε αυξημένη θνησιμότητα είναι σημαντική τόσο για την προστασία των ατόμων υψηλού κινδύνου όσο και για την προστασία της ευρύτερης κοινότητας

Καθώς ο COVID-19 εξαπλώθηκε από την Κίνα διαφορετικές εθνολογικά περιοχές παγκοσμίως, όπως για παράδειγμα το Ηνωμένο Βασίλειο, όπου το 13% του πληθυσμού προέρχεται από εθνικές μειονότητες, έγινε πιο προφανές ότι αυτές οι ομάδες επηρεάζονται περισσότερο (Pareek et al., 2020). Για παράδειγμα, 11 γιατροί που πέθαναν από COVID-19 ήταν από μαύρες, ασιατικές και μειονοτικές εθνικές κοινότητες (black, Asian and minority ethnic-BAME) (Kirby, 2020). Επιπλέον, ένας πρόσφατος Εθνικός Έλεγχος Εντατικής Θεραπείας ανέφερε ότι το ένα τρίτο των ασθενών σε μονάδες εντατικής θεραπείας προέρχονται από εθνικές μειονότητες. Περαιτέρω μελέτες έδειξαν ότι έως τις 22 Απριλίου 2020, το 63% των θανάτων που σχετίζονται με τον COVID-19 σε εργαζόμενους στον τομέα της υγείας ήταν από ομάδες BAME. Προφανώς, τα δεδομένα δείχνουν ότι τα άτομα BAME διατρέχουν αυξημένο κίνδυνο θανάτου από COVID-19.

Η πρώτη και μεγαλύτερη δια τομεακή ανάλυση πραγματοποιήθηκε από τους (de Lusignan et al., 2020).Στην εργασία τους, χρησιμοποιώντας λογισμικό R, ανέλυσαν δεδομένα με πολυμεταβλητά μοντέλα λογιστικής παλινδρόμησης πολλαπλών υπολογισμών, προκειμένου να εντοπίσουν μεταβλητές που σχετίζονται με τον COVID-19. Διαπιστώθηκε ότι το 15,5% που βρέθηκαν θετικοί ήταν λευκοί και το 62,1% ήταν έγχρωμοι. Συνολικά, οι μαύροι επηρεάστηκαν δυσανάλογα, ακόμη και όταν έγιναν προσαρμογές για συγχυτικές μεταβλητές όπως η υπέρταση και ο διαβήτης. Άλλοι κλινικοί και δημογραφικοί παράγοντες κινδύνου του COVID-19 ήταν η χρόνια νεφρική νόσο, η παχυσαρκία, η ηλικιακή ομάδα μεταξύ 40 και 64 ετών και η ζωή σε υποβαθμισμένες περιοχές. Η μεγάλη πρόκληση σε αυτήν τη μελέτη, όπως φαίνεται σε πολλές άλλες μελέτες που χρησιμοποιούν δεδομένα του NHS, είναι η απουσία μεγάλου ποσοστού δεδομένων. Για παράδειγμα, για του 1014 από τους 3802 ασθενείς, δεν υπήρχαν δεδομένα σχετικά με την εθνότητά τους αφού τα πιστοποιητικά θανάτου δεν περιλαμβάνουν αυτές τις πληροφορίες. Το πρόβλημα αυτό ξεπεράστηκε με την τυχαία αντιστοίχιση μιας εθνοτικής ομάδας σύμφωνα με την αναλογία

εθνότητας στην περιοχή, όπως προέκυπτε από τα δεδομένα της απογραφής. Άλλα ημιτελή δεδομένα αντιμετωπίστηκαν μέσω πολυμεταβλητών καταλογισμών με αλυσιδωτές εξισώσεις (multivariate Imputation by chained equations-MICE) και ανάλυση ευαισθησίας χρησιμοποιώντας πλήρη περιπτωσιολογική ανάλυση (Van Buuren & Groothuis-Oudshoorn, 2011).

Πρόσφατα δεδομένα δημοσιεύθηκαν από την Εθνική Στατιστική Υπηρεσία του Ηνωμένου Βασιλείου τα οποία εξέταζαν τους θανάτους που σχετίζονται με τον COVID-19 ανά εθνοτική ομάδα, κατά την περίοδο 2 Μαρτίου έως 10 Απριλίου 2020[1]. Χρησιμοποιώντας ένα μοντέλο λογιστικής παλινδρόμησης, διαπιστώθηκε ότι οι ομάδες BAME παρουσίαζαν υψηλότερο κίνδυνο θανάτου από COVID-19. Προκειμένου να ποσοτικοποιηθεί ο εθνοτικός κίνδυνος χωρίς συνθετικές μεταβλητές, κατέστη αναγκαία η θέσπιση προσαρμογών για κοινωνικο-δημογραφικούς παράγοντες. Το επίπεδο στέρησης, η αγροτική αστική ταξινόμηση, η ηλικία, η σύνθεση της οικογένειας και η κοινωνικο-οικονομική κατάσταση προσαρμόστηκαν όλα και το καθένα ξεχωριστά και παρουσιάστηκαν ως περιττές αναλογίες. Μία πρώτη εξήγηση για αυτά τα ευρήματα ήταν ότι οι ομάδες BAME παρουσιάζουν υψηλότερη συχνότητα εμφάνισης προϋπαρχόντων χρόνιων καταστάσεων όπως είναι η υψηλή αρτηριακή πίεση, ο διαβήτης, το άσθμα και η παχυσαρκία, γεγονός που αυξάνει τον κίνδυνο για COVID-19 (Kirby, 2020). Παρόλα αυτά, οι φυλετικές ανισότητες σε αυτές τις προϋπάρχουσες συνθήκες δεν αντικατοπτρίζουν τις τεράστιες ανισότητες θανάτου από COVID-19. Στην πραγματικότητα, η εθνότητα επηρεάζει την κουλτούρα και τη συμπεριφορά οι οποίες με τη σειρά τους είναι σε θέση να αυξήσουν την ευαισθησία σε ασθένειες. Οι ομάδες BAME για παράδειγμα, τείνουν να εργάζονται σε χαμηλότερα αμειβόμενες θέσεις εργασίας, να ζουν σε πιο πολυπληθείς οικογένειες και σε πιο πυκνοκατοικημένες περιοχές, γεγονός που αυξάνει τον κίνδυνο έκθεσής τους στο ιό[2]. Πολλές από τις μελέτες σε αυτήν την βιβλιογραφική ανασκόπηση δεν έλαβαν υπόψη κοινωνικο-οικονομικούς παράγοντες όπως η απασχόληση σε θέσεις υψηλού κινδύνου, το εισόδημα, τα ποσοστά αναλφαβητισμού, την πρόσβαση στην υγειονομική περίθαλψη και την εκπαίδευση. Η συγκέντρωση ομάδων πληθυσμού με κριτήριο τη θνησιμότητα αποτελεί μια υπεραπλούστευση και δεν μπορεί να ερμηνεύσει τον μεγάλο αριθμό θανάτων στις ομάδες BAME. Ο διαχωρισμός

[1] Coronavirus-related deaths by ethnic group, England and Wales methodology - Office for National Statistics. Ons.gov.uk; 2020. Available:https://www.ons.gov.uk/peoplepo pulationandcommunity/birthsdeathsandmar riages/deaths/methodologies/coronavirusre lateddeathsbyethnicgroupenglandandwale smethodolog

[2] COVID-19: Review of disparities in risks and outcomes. GOV.UK; 2020. Available from: Available:https://www.gov.uk/government/publications/covid-19-review-of-disparitiesin-risks-and outcomes

αυτών των ομάδων λαμβάνοντας υπόψη χαρακτηριστικά όπως είναι ο τόπος κατοικίας, οδηγεί στην καλύτερη κατανόηση των παραγόντων που προκαλούν αυτές τις δυσαναλογίες. Ωστόσο, λόγω της πολύπλοκης φύσης της εθνότητας που χαρακτηρίζεται από γενετικούς, συμπεριφορικούς και κοινωνικούς παράγοντες οι οποίοι αλληλοεπιδρούν μεταξύ τους, απαιτείται περαιτέρω διερεύνηση αυτών των παρατηρήσεων με σχετική ισχυρή ανάλυση. Η περιγραφή των ευρημάτων αυτών καθίσταται σημαντική για την κατανόηση περισσότερων στοιχείων σχετικά με αυτήν την περίπλοκη ασθένεια. Πρόσφατη αλληλογραφία που δημοσιεύτηκε από το Ινστιτούτο Παγκόσμιας Υγείας, University College του Λονδίνου στο Lancet χαρακτήρισε την έρευνα για την εθνότητα ως «επείγουσα ερευνητική προτεραιότητα για τη δημόσια υγεία» (Pareek et al., 2020). Ωστόσο, μια από τις μελλοντικές προκλήσεις για τον εντοπισμό εθνοτικών ομάδων υψηλού κινδύνου αποτελεί το γεγονός ότι για την αναφορά θνησιμότητας στο Ηνωμένο Βασίλειο δεν απαιτείται νομικά να περιλαμβάνεται η εθνότητα. Επιπλέον, η συγκέντρωση των δημοσιευμένων δεδομένων σχετικά με την εθνότητα καθώς και οι συσχετισμοί που πιθανόν προκύψουν από αυτά μπορεί να οδηγήσουν στον σχηματισμό ανακριβών συμπερασμάτων. Ωστόσο, η εύρεση της οριστικής αιτίας για τυχόν εθνοτικές ανισότητες είναι σημαντική. Η τεχνητή νοημοσύνη (AI) και η τεχνολογία μηχανικής μάθησης σε αυτόν τον τομέα θα μπορούσαν να είναι χρήσιμες για την κατανόηση αυτής της ανισότητας και, τελικά, την προστασία αυτών που διατρέχουν υψηλότερο κίνδυνο.

6.2. Τεχνητή Νοημοσύνη και Μηχανική Μάθηση

Η τεχνητή νοημοσύνη αποτελεί κλάδο της επιστήμης των υπολογιστών κατά τον οποίο οι υπολογιστές είναι σε θέση να μιμούνται την ανθρώπινη νοημοσύνη προκειμένου να εκτελούν σύνθετες εργασίες (Kersting, 2018). Ένας τύπος τεχνητής νοημοσύνης είναι η μηχανική εκμάθηση που προσδιορίζει μοτίβα και αποχρώσεις από δεδομένα, ενώ το σύστημα μαθαίνει και προσαρμόζεται μέσα από την εμπειρία προκειμένου να βελτιώνει τις επιδόσεις του.

6.2.1. Τεχνητή Νοημοσύνη και Μηχανική Μάθηση στην Ιατρική

Η λοιμώδης νόσος αποτελεί ένα σύνθετο πρόβλημα το οποίο απαιτεί καινοτόμες λύσεις. Η τεχνητή νοημοσύνη θα μπορούσε να αποτελέσει τη σύγχρονη τεχνολογία που παρέχει αυτές τις λύσεις. Καθώς οι δυνατότητες αποθήκευσης δεδομένων έχουν βελτιωθεί αρκετά στον τομέα της υγειονομικής περίθαλψης και πολύτιμα δεδομένα έχουν πια συγκεντρωθεί, εργαλεία μηχανικής μάθησης (ML) έχουν εφαρμοστεί για τον εντοπισμό προτύπων δεδομένων και την πρόβλεψη

μελλοντικών αποτελεσμάτων (Agrebi & Larbi, 2020). Η τεχνητή νοημοσύνη έχει ήδη αφήσει το στίγμα της στην ιατρική παίζοντας σημαντικό ρόλο στην πρόβλεψη φαινοτύπων από γονότυπους, στην ακτινολογία και στην παθολογική διάγνωση (Chandiok & Chaturvedi, 2015). Η πρόοδος στην έρευνα της τεχνητής νοημοσύνης έγινε ιδιαιτέρως εμφανής στην επιτυχία που σημειώθηκε στον τομέα της ακτινολογίας. Οι αλγόριθμοι ML που χρησιμοποιούνται για την ερμηνεία πολύπλοκων μοτίβων δεδομένων απεικόνισης σε εργασίες που βασίζονται σε εικόνες έχουν ξεπεράσει τον άνθρωπο σε συγκεκριμένες εφαρμογές. Η ικανότητα μιας μεθόδου ML να μαθαίνει και να βελτιώνεται καθώς εξάγει μοτίβα και χαρακτηριστικά από τον μεγάλο όγκο συνόλων δεδομένων εισόδου που βρίσκονται στα αρχεία υγειονομικής περίθαλψης βελτιώνει τα προγνωστικά χαρακτηριστικά και κατ' επέκτασιν την ικανότητα λήψης αποφάσεων.

6.2.2. Τεχνητή Νοημοσύνη και Μηχανική Μάθηση σε Μολυσματικές Ασθένειες

Η τεχνητή νοημοσύνη υπερβαίνει τα κανονικά μοντέλα που βασίζονται σε εξισώσεις. Οι αλγόριθμοι τεχνητής νοημοσύνης μαθαίνουν από τα δεδομένα, επιτρέποντας σε άλλους αλγόριθμους να λαμβάνουν αποφάσεις χρησιμοποιώντας την εμπειρία η οποία είναι αποθηκευμένη στη βάση γνώσης (Wong et al., 2019). Μια μέθοδος λήψης αποφάσεων που ονομάζεται παλινδρόμηση της διαδικασίας Gauss χρησιμοποιήθηκε από ομάδα του Πανεπιστημίου της Οξφόρδης με σκοπό να βρει τη βέλτιστη πολιτική για την ελονοσία που θα έπρεπε να χρησιμοποιήσει η κυβέρνηση με βάση την εξάπλωση της νόσου σε περιοχές του πληθυσμού. Η εφαρμογή των σωστών πολιτικών την κατάλληλη στιγμή κατέστη σημαντικότατη καθώς οι περικοπές χρηματοδότησης για την πρόληψη ασθενειών απαιτούσαν αλλαγές στον προϋπολογισμό.

Χρησιμοποιώντας εφαρμογές που βασίζονται σε ML, γίνονται προσπάθειες για την πρόβλεψη μεμονωμένων παραγόντων κινδύνου όσον αφορά την υγεία οι οποίοι συμβάλλουν στον συνολικό κίνδυνο ασθένειας. Οι (Wiens et al., 2016) στην εργασία τους ανέπτυξαν μια εφαρμογή που μαθαίνει να χαρτογραφεί δεδομένα όπως το ιστορικό του ασθενούς, τα εργαστηριακά αποτελέσματα και τα δημογραφικά στοιχεία για την πρόβλεψη του κινδύνου του ασθενούς από τη νοσοκομειακή Clostridium difficile, μια κοινή νοσοκομειακή επίκτητη βακτηριακή λοίμωξη. Το μοντέλο διαστρωμάτωσης κινδύνου βάσει δεδομένων μπορεί επίσης να προσδιορίσει περιοχές hotspot ενός νοσοκομείου που ενδέχεται να απαιτούν πιο ενδελεχή απολύμανση. Ένας περιορισμός που έθεσε το μοντέλο ήταν ότι αναλάμβανε έναν σταθερό κίνδυνο μόλυνσης από

Clostridium difficile κατά τη διάρκεια της νοσηλείας ενός ασθενούς, ενώ στην πραγματικότητα το παραπάνω αποτελεί μια δυναμική μεταβλητή. Οι ημερήσιες εκτιμήσεις κινδύνου παρήχθησαν από μια μηχανή πολλαπλών εργασιών, επιτρέποντας την εξέταση μεταβλητών που εξαρτώνται από το χρόνο. Αυτές οι εκτιμήσεις επιτρέπουν στο προσωπικό υγειονομικής περίθαλψης να απομονώνει γρήγορα από τους υπόλοιπους ασθενείς τους ασθενείς υψηλού κινδύνου που εντοπίστηκαν, μειώνοντας έτσι την εξάπλωση της νόσου.

Μετά την ανασκόπηση πολλών εφαρμογών τεχνητής νοημοσύνης στον ιατρικό τομέα, η κύρια πρόκληση που τίθεται είναι ο τεράστιος όγκος των ιατρικών δεδομένων που αποθηκεύονται στα ηλεκτρονικά αρχεία υγείας. Ενώ αυτός ο μεγάλος όγκος δεδομένων είναι ζωτικής σημασίας για μοντέλα μηχανικής εκμάθησης που απαιτούν δεδομένα, πολλά από τα δεδομένα περιγράφονται ως "θορυβώδη" (Sarkar & Chakrabarti, 2020). Τα θορυβώδη δεδομένα περιέχουν μεγάλο ποσοστό δεδομένων που λείπουν ή είναι άσχετα. Δυστυχώς, αυτά τα πλήρως αυτοματοποιημένα συστήματα ML αρχίζουν να μαθαίνουν νέες έννοιες από τα θορυβώδη δεδομένα με αποτέλεσμα ανακριβείς προβλέψεις, έννοια γνωστή ως «overfitting». Συνηθέστερα, το ζήτημα αυτό μπορεί να ξεπεραστεί με τη χρήση μεθόδων τακτοποίησης της μάθησης. Δυστυχώς, η τεχνητή νοημοσύνη αντιμετωπίζει περισσότερες προκλήσεις με την ανάλυση μολυσματικών ασθενειών, αφού κάθε ασθένεια έχει τα δικά της μοναδικά φυσικά χαρακτηριστικά, όπως η περίοδος επώασης και η μεταδοτικότητα (Bent et al., 2018). Τα χαρακτηριστικά αυτά είναι δύσκολο να προβλεφθούν μέχρι να εμφανιστεί η ασθένεια, γεγονός που καθιστά την έγκαιρη πρόβλεψη δύσκολη.

6.2.3. Τεχνητή Νοημοσύνη και Μηχανική Μάθηση για τον COVID-19

Καθώς η συχνότητα εμφάνισης του COVID-19 αυξήθηκε εκθετικά, τα συστήματα υγειονομικής περίθαλψης έχουν κατακλυστεί και οι πόροι έχουν καταπονηθεί. Με σκοπό τη μείωση του φόρτου αυτού, έχουν εφαρμοστεί προτεινόμενα μοντέλα πρόβλεψης από μοντέλα βασισμένα σε κανόνες έως πολύ προηγμένα μοντέλα ML. ΜΕ τον τρόπο αυτό επιτυγχάνεται ο εντοπισμός ασθενών υψηλού κινδύνου, η διάγνωση και η πρόβλεψη των εκβάσεων και η πρόγνωσης της νόσου (Wynants et al., 2020).

Ένα μοντέλο AI που προτάθηκε από τους (Pourhomayoun & Shakibi, 2020) υπολόγισε τον κίνδυνο θνησιμότητας για ασθενείς θετικούς στον COVID-19, προκειμένου να βοηθήσει στην αξιολόγηση και την ιεράρχηση των ασθενών με κίνδυνο πιο αποτελεσματικά και με ακρίβεια στο τρέχον καταπονημένο σύστημα υγειονομικής περίθαλψης (Yan et al., 2020). Οι αλγόριθμοι ML χρησιμοποίησαν 117.000 ασθενείς θετικούς στον COVID-19 προκειμένου εκπαιδεύσουν ένα

μοντέλο που θα προβλέπει τους κινδύνους θνησιμότητας με βάση 42 χαρακτηριστικά, τα οποία αποτελούνταν από φυσιολογικές καταστάσεις και δημογραφικά χαρακτηριστικά που εξήχθησαν από ιατρικά αρχεία. Η Μηχανή Φορέα Υποστήριξης (SVM), τα νευρωνικά δίκτυα και το τυχαίο δάσος ήταν οι αλγόριθμοι που χρησιμοποιήθηκαν για τη δημιουργία του προγνωστικού μοντέλου. Οι μετρήσεις απόδοσης πραγματοποιήθηκαν με τη χρήση καμπυλών AUC - ROC και η ακρίβεια των προβλέψεων υπολογίστηκε χρησιμοποιώντας 10πλάσια διασταυρούμενη επικύρωση. Τα αποτελέσματα έδειξαν ότι οι αλγόριθμοι νευρωνικών δικτύων ήταν οι πιο ακριβείς σε ποσοστό που έφθανε το 93,75%.

Μια συστηματική ανασκόπηση που αξιολογούσε κριτικά 66 μοντέλα από 51 μελέτες διαπίστωσε ότι η πλειονότητα των μελετών που χρησιμοποίησαν μοντέλα ML χρησιμοποιήθηκαν για ανάλυση εικόνας αξονικών τομογράφων και ακτινογραφιών για ανάλυση του COVID-19. Τα υπόλοιπα μοντέλα ML χρησιμοποιήθηκαν για την πρόβλεψη θνησιμότητας χρησιμοποιώντας δημογραφικά στοιχεία όπως η ηλικία και οι βιοδείκτες από τις εξετάσεις αίματος (Sarkar & Chakrabarti, 2020). Μια ομάδα στην Κίνα χρησιμοποίησε το μοντέλο XGBoost το οποίο είναι ένας αλγόριθμος ML υψηλής απόδοσης (Yan et al., 2020). Χρησιμοποιώντας 909 δείγματα αίματος από 485 ασθενείς, αναπτύχθηκε ένα λειτουργικό δέντρο αποφάσεων από τρεις επιλεγμένους βιοδείκτες αίματος (LDL, hsCRP, λεμφοκύτταρα). Σε όλη τη βιβλιογραφία, αυτοί οι τρεις βιοδείκτες χρησιμοποιούνται συχνά για την πρόγνωση του COVID-19. Το προγνωστικό μοντέλο μπορεί να ποσοτικοποιήσει τον κίνδυνο θανάτου στο 90%. Το σημαντικότερο όμως είναι ότι αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για να δώσει προτεραιότητα σε ασθενείς που χρειάζονται εξειδικευμένη φροντίδα, η οποία είναι απίστευτα χρήσιμη όπου οι πόροι είναι περιορισμένοι. Η ακρίβεια του μοντέλου θα μπορούσε να βελτιωθεί περαιτέρω με περισσότερα σύνολα δεδομένων. Παρά το γεγονός ότι χρησιμοποιούνται μόνο τρεις βιοδείκτες για την αποφυγή υπερβολικής προσαρμογής, η πρόοδος στην κατανόησή για την παθοφυσιολογία του COVID-19 θα επιτρέψει τον εντοπισμό των βιοδεικτών εκείνων που επηρεάζουν σημαντικά τη σοβαρότητα της νόσου, όπως είναι η Ιντερλευκίνη-6 (Messner et al., 2020). Η εφαρμογή τους σε ένα μοντέλο πρόβλεψης θα μπορούσε να βελτιώσει περαιτέρω την απόδοση του μοντέλου.

Οι περιορισμοί σε όλα τα μοντέλα που εξετάστηκαν περιλάμβαναν τον υψηλό κίνδυνο μεροληψίας λόγω χαμηλού μεγέθους δείγματος και φτωχής αναφοράς (Wynants et al., 2002). Επιπλέον, τα μοντέλα έδειξαν μεταβλητά προβλεπόμενα αποτελέσματα, τα οποία μπορεί να οδηγήσουν σε εσφαλμένη βαθμονόμηση. Ένα ακόμη χαρακτηριστικό είναι το γεγονός ότι δεν

υπήρχε επαρκής προσδιορισμός των πληθυσμών-στόχων, στοιχείο σημαντικό για μια ενοποιημένη προσέγγιση στην αξιολόγηση του μοντέλου. Επί του παρόντος, κανένα από αυτά τα μοντέλα δεν είναι σε θέση να εφαρμοστεί στην πράξη και εξαιτίας των σοβαρών χρονικών περιορισμών για τη δημοσίευση των ευρημάτων ενώ οι εργασίες στην πλειονότητά τους δεν έχουν αξιολογηθεί. Ωστόσο, αυτά τα μοντέλα παρέχουν μια πλατφόρμα για περαιτέρω ανάπτυξη ώστε τελικά να παραχθεί ένα ισχυρό μοντέλο πρόβλεψης που να είναι σε θέση να χρησιμοποιηθεί στην πράξη. Καθώς συγκεντρώνονται περισσότερα κλινικά σύνολα δεδομένων, καθίσταται δυνατή η ανάπτυξη ισχυρότερων μοντέλων πρόβλεψης. Μετά την επικύρωση από έναν ανεξάρτητο εξωτερικό φορέα, οι κλινικοί γιατροί θα είναι σε θέση να μπορούν να εφαρμόσουν τα μοντέλα αυτά στην πράξη. Έχουν συνταχθεί περιορισμένες εργασίες που έχουν χρησιμοποιήσει μοντέλα ML για τον εντοπισμό κοινωνικο-δημογραφικών παραγόντων που σχετίζονται με τον COVID-19. Πρόκειται για έναν τομέα μελλοντικής έρευνας, όπου αυτά τα δυναμικά μοντέλα ML μπορούν να εντοπίσουν άτομα με παράγοντες κινδύνου που σχετίζονται με τον COVID-19 στον πληθυσμό. Η απομόνωση αυτών των ατόμων και η πρόληψη της έκθεσης στη νόσο θα μπορούσε να μειώσει τον αριθμό των ασθενών που εισάγονται στο νοσοκομείο με σοβαρά συμπτώματα COVID-19 και για τα οποία συμπτώματα απαιτείται μεταφορά σε Μονάδες Εντατικής Θεραπείας, γεγονός καθοριστικής σημασίας σε μια περίοδο όπου η ικανότητα εντατικής θεραπείας είναι περιορισμένη.

6.3.Χωρική Μοντελοποίηση και Επιδημιολογία

Η χωρική μοντελοποίηση στην επιδημιολογία ασθενειών περιλαμβάνει την εισαγωγή χωρικών δεδομένων με συστήματα γεωγραφικών πληροφοριών σε μοντέλα προκειμένου να εντοπιστεί και να απεικονιστεί η γεωγραφική κατανομή της νόσου με βάση δημογραφικούς, περιβαλλοντικούς και κοινωνικο-δημογραφικούς παράγοντες (Elliott, & Wartenberg, 2004). Οι περισσότερες, αν όχι όλες οι μολυσματικές ασθένειες επηρεάζονται σε μεγάλο βαθμό από περιβαλλοντικούς παράγοντες. Ως εκ τούτου, η ενσωμάτωση γεω-χωρικών δεδομένων σε πολύπλοκα χωρικά μοντέλα προκειμένου να γίνει πιο κατανοητή η δυναμική της μετάδοσης ήταν απίστευτα χρήσιμη στην επιδημιολογία ασθενειών. Το GIS και η χωρική μοντελοποίηση υπερβαίνουν την απλή οπτικοποίηση δεδομένων. Τα χωρικά μοντέλα μπορεί να είναι ζωτικής σημασίας για τον εντοπισμό περιοχών υποκλίμακας με αναδυόμενες περιπτώσεις COVID-19 (O'Sullivan et al., 2020). Καθώς αρχίζουν να αίρονται τα εθνικά μέτρα περιορισμού (lockdown), η εφαρμογή μικρών χωρικών κλιμάκων με χρήση χωρικών μοντέλων επιτρέπει τον εντοπισμό μικρότερης κλίμακας

περιοχών όπως πόλεις ή γειτονιές που διατρέχουν υψηλότερο κίνδυνο, κάτι που θα βοηθήσει να αποτραπεί σημαντικά η εξάπλωση του COVID-19 στην ευρύτερη κοινότητα και έτσι να αποτραπεί περαιτέρω επέκταση του ιού.

Ομάδα μελετητών στο Ιράν πραγματοποίησε χωρική μοντελοποίηση και χαρτογράφηση κινδύνου με βάση την τεχνική της τυχαίας μηχανικής εκμάθησης δασών (RF-MLT) χρησιμοποιώντας το "τυχαίο πακέτο δασών" σε λογισμικό R (Pourghasemi et al., 2020). Το τυχαίο Δάσος είναι ουσιαστικά ένας τύπος εποπτευόμενου αλγόριθμου μηχανικής μάθησης που συνδυάζει πολλαπλά δέντρα αποφάσεων μέσω συνάθροισης bootstrap. Το πλεονέκτημα του συνδυασμού αυτών των δέντρων απόφασης είναι η μείωση τόσο της μεροληψίας όσο και της πιθανότητας υπερπροσαρμογής, κάτι που συνεπάγεται την υψηλή ακρίβεια της πρόβλεψης. Στο μοντέλο αυτό δοκιμάστηκαν 16 μεταβλητές, συμπεριλαμβανομένων αυτών που αφορούν κλιματικούς παράγοντες, όπως είναι η θερμοκρασία, οι οποίοι ελήφθησαν από σύνολα δεδομένων WorldClim και γεωγραφικών παραγόντων όπως η απόσταση από τους δρόμους και η πυκνότητα των πόλεων. Αυτές οι μεταβλητές χαρτογραφήθηκαν με σκοπό την πρόβλεψη της μετάδοσης της νόσου. Οι κοινωνικά πυκνοί χώροι, όπως οι σταθμοί λεωφορείων και οι χώροι λατρείας, αναλύθηκαν με τη χρήση χωρικών εργαλείων ArcGIS. Δημιουργήθηκαν χάρτες θερμότητας που έδειξαν την κατανομή υψηλότερων συγκεντρώσεων μόλυνσης σε διάφορες περιοχές του Ιράν. Η χωρική ανάλυση έδειξε συσχέτιση υψηλού ποσοστού μόλυνσης μεταξύ δύο περιοχών. Για παράδειγμα, οι πόλεις Alborz και Qazvin παρουσίασαν και οι δύο υψηλά επίπεδα μόλυνσης ως αποτέλεσμα των πολυσύχναστων διασυνδέσεων με αυτοκινητόδρομους. Κι ενώ η επικύρωση των χαρτών κινδύνου μέσω ROC-AUC έδειξε βαθμολογία 0,886 που χαρακτηρίζεται ως "πολύ καλή", οι μέθοδοι που βασίζονται σε GIS αποτελούν την πιο πιο συχνά χρησιμοποιούμενη μέθοδο.

6.3.1. Μοντελοποίηση GIS

Η γεωπληροφορική έχει παίξει σημαντικό ρόλο στην πρόβλεψη μελλοντικών εστιών ασθενειών και στην παρακολούθηση της εξάπλωσης της νόσου. Η χρήση τεχνολογίας επιδημιολογικής χαρτογράφησης με συστήματα προειδοποίησης βάσει τοποθεσίας επιτρέπει την οπτικοποίηση των τάσεων και των προτύπων που αποτελούν χρήσιμη πηγή για τους υπεύθυνους λήψης αποφάσεων στον τομέα της δημόσιας υγείας (Franch-Pardo et al., 2002). Για περισσότερα από 20 χρόνια το μεγαλύτερο εργαλείο του Παγκόσμιου Οργανισμού Υγείας (ΠΟΥ) στη χαρτογράφηση ασθενειών αποτέλεσε η γεωπληροφορική, και συγκεκριμένα τα γεωγραφικά συστήματα πληροφοριών (geographical information systems-GIS). Ο ΠΟΥ χρησιμοποιούσε GIS για να παρακολουθεί

εστίες ασθενειών που μεταδίδονται με φορείς όπως ο δάγκειος πυρετός και να αναλύει τις ευάλωτες περιοχές όπου απαιτούνταν επείγοντα προγράμματα εμβολιασμού (Khan et al., 2010). Η χαρτογράφηση του κλίματος και της πυκνότητας του πληθυσμού για τον ιό του δάγκειου πυρετού έδειξε ότι η πυκνότητα του πληθυσμού είναι ο πιο σημαντικός παράγοντας στην εξάπλωση της συγκεκριμένης ασθένειας (Butt et al., 2019).

Κατά τη διάρκεια της πρώιμης πανδημίας, το Πανεπιστήμιο Johns Hopkins δημιούργησε έναν ευρέως δημοφιλή διαδραστικό διαδικτυακό πίνακα ελέγχου χρησιμοποιώντας δεδομένα σε πραγματικό χρόνο από το CDC των ΗΠΑ (Κέντρα Ελέγχου και Πρόληψης Νοσημάτων) και ECDC (Ευρωπαϊκό Κέντρο Πρόληψης και Ελέγχου Νοσημάτων) (Dong et al., 2020). Ο πίνακας ενημερωνόταν τακτικά μέσω της ομάδας ArcGIS Living Atlas της Esri, μεταδίδοντας δεδομένα από την πραγματική ζωή, τα οποία έδιναν πλήρεις λεπτομέρειες και οπτικοποιούσαν την εξέλιξη του ιού καθώς και τα επιβεβαιωμένα κρούσματα παγκοσμίως. Ωστόσο, η έλλειψη δεδομένων σε πραγματικό χρόνο απέτρεψε την ανάλυση των προτύπων κινητικότητας, στοιχείο καθοριστικής σημασίας για τη χαρτογράφηση της εξάπλωσης της νόσου. Το GIS είναι ιδιαίτερα χρήσιμο καθώς επιτρέπει τον εντοπισμό περιοχών πληθυσμού υψηλού κινδύνου ώστε να δίνει τη δυνατότητα στις τοπικές αρχές τόσο στο να παρέχουν έγκαιρες πληροφορίες στο κοινό όσο και στο να διοχετεύουν πόρους υγειονομικής περίθαλψης για έγκαιρη παρέμβαση αυτής. Για παράδειγμα, στον Καναδά χρησιμοποιώντας δεδομένα από την Canada's Community Health Survey, οι περιοχές με υψηλή πυκνότητα πληθυσμού με υποκείμενα νοσήματα και γηρασμένου προσδιορίζονται μέσω πίνακα, βοηθώντας έτσι την τοπική κυβέρνηση να ανακατευθύνει τους πόρους για να προετοιμαστεί για αύξηση των εισαγωγών στα νοσοκομεία.[3]

6.3.2. Προσδιορισμός Κοινωνικο-Δημογραφικών του COVID-19 με GIS

Η πλειοψηφία των GIS έχει χρησιμοποιηθεί για την πρόβλεψη νέων εστιών, την παρακολούθηση της εξάπλωσης ασθενειών και την ανάλυση των επιπτώσεων της κοινωνικής απόστασης. Ωστόσο, περιορισμένες δημοσιεύσεις έχουν χρησιμοποιήσει GIS για τον εντοπισμό των παραγόντων κινδύνου που σχετίζονται με τον COVID-19 σε σχέση με διάφορες γεωγραφικές περιοχές. Ένα χρήσιμο εργαλείο στο λογισμικό ArcGIS είναι η προσθήκη επιπέδων δεδομένων δημογραφικών

[3] COVID-19 Response: GIS best practices in Local Government. Data-Smart City Solutions; 2021. Available:https://datasmart.ash.harvard.ed u/news/article/covid-19-response-gis-bestpractices- local-government

χαρακτηριστικών από τον «Ζωντανό άτλαντα του κόσμου» στον πίνακα προκειμένου να εντοπιστούν κοινωνικοδημογραφικά χαρακτηριστικά που σχετίζονται με τα hotspot του COVID-19. Ο εντοπισμός αυτών των παραγόντων κινδύνου που επηρεάζουν τη σοβαρότητα και την πρόγνωση της νόσου είναι ζωτικής σημασίας για την κατανόηση της εξάπλωσης της νόσου, τη διευκόλυνση του σχεδιασμού της δημόσιας υγείας και την προστασία των ευάλωτων ατόμων.

7. Βιβλιογραφική Ανασκόπηση Μεθοδολογιών GIS για την Ανάλυση της Δυναμικής του COVID-19

Από τις αρχές του 2020 οπότε και εμφανίστηκε, η πανδημία COVID-19 αποτελεί σημαντική απειλή για τη δημόσια υγεία παγκοσμίως. Ο COVID-19 προκαλείται από το σοβαρό οξύ αναπνευστικό σύνδρομο κορονοϊού 2 (Shi et al., 2020). Η πανδημία θεωρείται ότι ξεκίνησε σε αγορά θαλασσινών στη Wuhan, στην επαρχία Hubei, στην Κίνα. Ο ιός εξαπλώθηκε γρήγορα και σε άλλες χώρες της ανατολικής Ασίας, στην Ευρώπη και τελικώς σε όλο τον κόσμο (Spagnuolo et al., 2020). Η εξάπλωση της πανδημίας έχει μελετηθεί ευρέως και οι στρατηγικές και τα εργαλεία που εφαρμόστηκαν για την έρευνα των χωρικών και χρονικών αλλαγών στη μετάδοση του COVID-19 ποικίλουν ενώ χαρακτηρίσθηκαν από γρήγορη εξέλιξη με σκοπό να προσαρμοστούν στα διαθέσιμα δεδομένα και στη γνώση όσον αφορά τη νόσο. Από την έναρξη της πανδημίας, πάνω από 163 εκατομμύρια άτομα έχουν μολυνθεί και πάνω από 3,3 εκατομμύρια έχουν πεθάνει. Η διάρκεια της πανδημίας, μαζί με τις τεράστιες επιπτώσεις στις κοινωνίες, τις οικονομίες, την πολιτική και τη δημόσια υγεία, έχουν επηρεάσει σε μεγάλο βαθμό τον ρόλο της χωρικής ανάλυσης στην κατανόηση του COVID-19.

Τα χωρικά εργαλεία και οι τεχνικές που χρησιμοποιήθηκαν για την κατανόηση του COVID-19 παρουσίασαν μεγάλη ποικιλία ενώ η αποκτηθείσα γνώση μπορεί να βοηθήσει τους ερευνητές στην ανάπτυξη και την βελτίωση των μεθοδολογιών που υιοθετήθηκαν για να επηρεάσουν τη λήψη αποφάσεων, τις κατευθυντήριες γραμμές για τη δημόσια υγεία και την κατανομή πόρων όπως είναι τα εμβόλια και τέστ (Bernasconi & Grandi, 2021). Η ανασκόπηση των μεθοδολογιών που χρησιμοποιούνται μπορεί να συμβάλει στη βελτίωση της διαχείρισης του COVID, της διαχείρισης της μετά τον COVID κατάστασης και του σχεδιασμού προκειμένου για μελλοντικές πανδημίες.

Το πρώτο εξάμηνο του 2020, υπήρχαν πολύ λιγότερες πληροφορίες για τον COVID-19, εξαιτίας της εφοδιαστικής, της καινοτομίας της νόσου και της έλλειψης έρευνας και δεδομένων. Καθώς προχωρούσε ο καιρός περνούσε, διαπιστώθηκε μια πολύ σημαντική αύξηση στον αριθμό των μελετών των σχετικών με την πανδημία, ειδικά εκείνων που χρησιμοποίησαν χωρική ανάλυση. Από αυτή την άποψη, με βάση όσα δημοσιεύτηκαν το πρώτο εξάμηνο του 2020 (Ιανουάριος – Μάιος 2020), οι (Franch-Pardo et al., 2020) προσδιόρισαν 63 έργα που εφάρμοσαν συστήματα γεωγραφικών πληροφοριών (GIS) και χωρική επιστήμη για την ανάλυση του COVID-19, ομαδοποιώντας τις μελέτες σε πέντε κύριες κατηγορίες: χώρο-χρονική ανάλυση, υγεία και κοινωνική γεωγραφία, περιβαλλοντικές μεταβλητές, εξόρυξη δεδομένων και χαρτογράφηση μέσω

ιστού. Εμπνευσμένη από προηγούμενες εργασίες, παρέχεται πια μια ενημερωμένη συστηματική ανασκόπηση των μελετών που δημοσιεύθηκαν το δεύτερο εξάμηνο του 2020, όπου παρουσιάζεται η θεματική εξέλιξη του GIS και της χωρικής ανάλυσης που χρησιμοποιήθηκαν στις μελέτες COVID-19 σε σύγκριση με το πρώτο εξάμηνο του 2020. Κατά τους πρώτους μήνες της πανδημίας δημοσιεύθηκαν αξιοσημείωτες έρευνες σχετικές με τη χρησιμότητα αυτών των τεχνολογιών καθώς και προσεγγίσεις για την πανδημική έρευνα (Casti, 2020; Devasia et al., 2020). Σκοπός είναι η παροχή μιας πολύτιμης πηγής για ερευνητές που εφαρμόζουν GIS και χωρική ανάλυση για τη μελέτη του COVID-19.

7.1. Χωρικές στατιστικές και COVID-19

Οι χωρικές στατιστικές χρησιμοποιήθηκαν κυρίως για την ανάλυση των κοινωνικό-οικονομικών και δημογραφικών παραγόντων κινδύνου του COVID-19 (Iyanda et al., 2020; Urban & Nakada, 2021). Η ποιότητα του αέρα, οι υποδομές υγείας και η κινητικότητα έχουν επίσης εξεταστεί και συνδέονται με τον κίνδυνο και τη μετάδοση που χαρακτηρίζουν τον COVID-19. Επιπλέον έχει διαπιστωθεί μια εισροή τόσο χωρικών όσο και χώρο-χρονικών στατιστικών μεθόδων που εφαρμόζονται στην περίπτωση της μελέτης του COVID-19, με συνέπεια πολλές από τις μελέτες να χρησιμοποιούν τα αποτελέσματα για την παραγωγή χαρτών κινδύνου και κοινωνικής ευπάθειας.

7.2. Χωρικές Παλινδρομήσεις

Σε μελέτες χωρικής μοντελοποίησης, είναι σύνηθες να ξεκινάει κανείς με τυπική παλινδρόμηση ελαχίστων τετραγώνων (OLS) προκειμένου να εντοπιστούν σημαντικές σχέσεις μεταξύ των εξαρτημένων και των ανεξάρτητων μεταβλητών. Εάν τα υπολείμματα ενός μοντέλου OLS είναι χωρικά αυτοσυσχετισμένα, τότε είναι σκόπιμο να χρησιμοποιηθούν μέθοδοι που βασίζονται σε χωρική παλινδρόμηση (Delmelle et al., 2016). Για παράδειγμα, ένα μοντέλο χωρικής υστέρησης (spatial lag model-SLM) μπορεί να χρησιμοποιηθεί για να εξετάσει το πώς τα γεγονότα σε μια τοποθεσία επηρεάζουν παρόμοια γεγονότα στις γύρω τοποθεσίες (δηλαδή, χωρική αλληλεπίδραση) και το πώς ένα μοντέλο χωρικού σφάλματος (spatial error model-SEM) μπορεί να εφαρμοστεί για να ληφθεί υπόψη η αυτοσυσχέτιση των υπολειπόμενων (Iyanda et al., 2020; Urban & Nakada, 2021). Για τον COVID-19, χωρικά συνδυασμένα αυτοπαλινδρομικά μοντέλα (spatially combined autoregressive-SAC) έχουν επίσης χρησιμοποιηθεί ως συνδυασμός των

προηγούμενων μοντέλων για την ταυτόχρονη εξέταση της χωρικής καθυστέρησης και των παραμέτρων χωρικού σφάλματος (Sun et al., 2020).

Μια άλλη κοινή μέθοδος είναι η γεωγραφικά σταθμισμένη παλινδρόμηση (geographically weighted regression-GWR), η οποία χρησιμοποιεί τις μεταβλητές που περιλαμβάνονταν προηγουμένως στην παλινδρόμηση OLS (Alkhaldy, 2020; Urban & Nakada, 2021). Η παλινδρόμηση GWR δημιουργεί ένα τοπικό μοντέλο και υπολογίζει τις παραμέτρους για όλα τα σημεία του δείγματος λαμβάνοντας υπόψη τη χωρική διακύμανση στις σχέσεις (Maiti et al., 2021). Αυτό το είδος παλινδρόμησης μπορεί να εξετάσει μη στάσιμες μεταβλητές (όπως είναι το κλίμα και οι δημογραφικοί και περιβαλλοντικοί παράγοντες) και μοντελοποιεί τις τοπικές σχέσεις μεταξύ αυτών των προγνωστικών παραγόντων και των υπό μελέτη προτύπων. Διευκολύνει την ανάλυση της χωρικής διακύμανσης ενός φαινομένου σε ένα δεδομένο μέρος, σύμφωνα με τον πρώτο νόμο της γεωγραφίας του Tobler - «όλα σχετίζονται με όλα τα άλλα, αλλά τα κοντινά πράγματα σχετίζονται περισσότερο από τα μακρινά. Όσον αφορά τον COVID-19, η παλινδρόμηση GWR έχει χρησιμοποιηθεί για να εξετάσει τις σχέσεις μεταξύ της νόσου και της ποιότητας του αέρα καθώς και μιας ποικιλίας κοινωνικό-οικονομικών μεταβλητών. Για παράδειγμα, οι (Fan et al., 2020) εφάρμοσαν την GWR καθώς και άλλα μοντέλα για να αξιολογήσουν την εξέλιξη της ατμοσφαιρικής ρύπανσης κατά το 2020 σε αστικά περιβάλλοντα στην Κίνα. Οι (Murgante et al., 2020) εξέτασαν τους γεωγραφικούς παραλλήλους μεταξύ των πληγεισών περιοχών στην κοιλάδα του Πάδου, στην Ιταλία και στη Γουχάν, στην Κίνα και διαπιστώθηκε ότι η ρύπανση και η χρήση γης διαδραματίζουν σημαντικό ρόλο στον τρόπο κατανομής του COVID-19 και στις δύο περιοχές. Οι (Mansour et al., 2020) χρησιμοποίησαν την GWR για να προσδιορίσουν τις σχέσεις μεταξύ κοινωνικό-δημογραφικών μεταβλητών (πυκνότητα πληθυσμού, ηλικιακές ομάδες, διαβητικοί) και του COVID-19 στο Ομάν.

Η GWR ωστόσο υπόκειται στον περιορισμό να υποθέτει ότι όλες οι διαδικασίες που μοντελοποιούνται λειτουργούν στην ίδια χωρική κλίμακα, γεγονός που οδήγησε στην ανάπτυξη της γεωγραφικά σταθμισμένης παλινδρόμησης (multiscale geographically weighted regression-MGWR) όπου η μοντελοποίηση μπορεί να εφαρμοστεί σε διαφορετικές χωρικές κλίμακες (Maiti et al., 2020). Οι (Iyanda et al. 2020) χρησιμοποίησαν την MGWR για να εντοπίσουν σημαντικούς κοινωνικό-δημογραφικούς και οικονομικούς παράγοντες του COVID-19 σε επίπεδο χώρας. Οι (Sannigrahi et al., 2020) υπολόγισαν τους τοπικούς χωρικούς συντελεστές συσχέτισης μεταξύ κοινωνικό-δημογραφικών μεταβλητών και δεδομένων COVID-19 στην Ευρώπη. Οι (Mollalo et

al., 2020), από την άλλη πλευρά, χρησιμοποίησαν την MGWR για μια εξέταση σε τοπικό επίπεδο της χωρικής μη σταθερότητας μεταξύ 35 περιβαλλοντικών, κοινωνικό-οικονομικών, τοπογραφικών και δημογραφικών μεταβλητών στις Ηνωμένες Πολιτείες.

Η γεωγραφικά σταθμισμένη ανάλυση κύριου συστατικού (Geographically weighted principal component analysis-GWPCA) είναι μια επέκταση της κλασικής ανάλυσης κύριου συστατικού που προσαρμόζει την προσέγγιση για χρήση με γεωγραφικά δεδομένα λαμβάνοντας υπόψη τη χωρική. Για τον COVID-19, οι (Das et al., 2021) χρησιμοποίησαν την GWPCA για την ανάπτυξη του βελτιωμένου δείκτη πολλαπλών στερήσεων σε περιοχές της Καλκούτα της Ινδίας, στηριζόμενοι στις συνθήκες στέγασης, τις οικιακές ανέσεις, το νερό, την αποχέτευση και την υγιεινή, την κατοχή περιουσιακών στοιχείων και την ανισότητα μεταξύ των φύλων.

Οι (Nomura et al., 2020), στον νομό Φουκουόκα της Ιαπωνίας, χρησιμοποίησαν μια τροποποιημένη δοκιμή συντελεστών χωρικής συσχέτισης προκειμένου να συσχετίσουν τον αριθμό των επιβεβαιωμένων με PCR COVID-19 κρουσμάτων σε μια εφαρμογή κοινωνικού δικτύου που παρέχει παρακολούθηση σε πραγματικό χρόνο των αυτό αναφερόμενων συμπτωμάτων του COVID-19. Η εξαιρετικά σημαντική συσχέτιση που αναφέρουν υποδεικνύει τη χρησιμότητα των δεδομένων που προέρχονται από πλήθος στη χωρική ανάλυση, η οποία μπορεί να βοηθήσει σε αξιολογήσεις πολιτικής, όπως οι δηλώσεις έκτακτης ανάγκης.

7.3.Hotspots και Ομαδοποίηση

Μια άλλη κοινή μέθοδος που χρησιμοποιείται για τη μελέτη του COVID-19 είναι η ανάλυση hotspot, η οποία μπορεί να διευκολύνει στοχευμένες παρεμβάσεις από τοπικούς, πολιτειακούς και ομοσπονδιακούς φορείς. Δύο από τα πιο κοινά εργαλεία για τη μέτρηση της τοπικής ομαδοποίησης είναι η στατιστική Getis-Ord Gi και η πυκνότητα πυρήνα. Η στατιστική Gi χρησιμοποιεί έναν συνολικό δείκτη για να μετρήσει το επίπεδο χωρικής αυτοσυσχέτισης, δηλαδή τον βαθμό στον οποίο τα αντικείμενα ή οι δραστηριότητες σε μια γεωγραφική ενότητα είναι παρόμοια με άλλα αντικείμενα ή δραστηριότητες σε κοντινές γεωγραφικές μονάδες. Έχει χρησιμοποιηθεί σε πολυάριθμες μελέτες για την κατανόηση της κατανομής των περιπτώσεων COVID-19, της χωρικής τους εξέλιξης με την πάροδο του χρόνου καθώς και για χάρτες ευπάθειας και κινδύνου.

Η εκτίμηση πυκνότητας πυρήνα (Kernel density estimation-KDE) χρησιμοποιείται για την εκτίμηση των πυκνοτήτων δεδομένων που δεν έχουν παραμετρικές στατιστικές συμπεριφορές, δηλαδή δεν ακολουθούν κανονικές, διωνυμικές ή εκθετικές κατανομές. Οι (Rex et al., 2020)

εφάρμοσαν το KDE για να εντοπίσουν περιοχές με υψηλή πυκνότητα κρουσμάτων COVID-19 ενώ οι (Nian et al. (2020) εξέτασαν δεδομένα κινητικότητας μέσων μαζικής μεταφοράς.

Μετά την ανασκόπηση των μεθόδων χωρικής αυτοσυσχέτισης, η συνολική μονομεταβλητή μέθοδος Moran's I είναι η πιο ευρέως χρησιμοποιούμενη, αν και έχει χρησιμοποιηθεί κυρίως με κοινωνικό-οικονομικά δεδομένα και δεδομένα COVID-19. Η μέθοδος αυτή μπορεί να αξιολογήσει το εάν τα δεδομένα τείνουν να είναι ομαδοποιημένα, διασκορπισμένα ή χωρικά τυχαία. Έχει χρησιμοποιηθεί για τον εντοπισμό ομαδοποίησης του COVID-19 (Wu et al., 2020) και διευκολύνει την παραγωγή χαρτών ευπάθειας και κινδύνου.

Ως παγκόσμιος δείκτης, το Moran's I παραμελεί την αστάθεια των τοπικών χωρικών διαδικασιών, η οποία οδήγησε στην ανάπτυξη της τοπικής έκδοσης του Moran's I που προσδιορίζει τόσο τη χωρική ομαδοποίηση οντοτήτων με παρόμοιες τιμές όσο και την εμφάνιση αποκλίνων αξιών. Αυτή η τελευταία έκδοση είναι επίσης γνωστή ως τοπικός δείκτης του χωρικού συσχετισμού (local indicator of spatial association -LISA). Οι (Xie et al. (2020) στην Κίνα, οι (Murgante et al., 2020) στις Ηνωμένες Πολιτείες και οι Santana et al., 2020) στο Μεξικό παρήγαγαν χάρτες συστάδων LISA για να αναλύσουν τα χαρακτηριστικά του COVID-19 σε διάφορα χωρικά επίπεδα συγκέντρωσης.

Άλλες πολλά υποσχόμενες τεχνικές ομαδοποίησης είναι:

• Οι αυτό οργανωτικοί χάρτες, γνωστοί και ως δίκτυα Kohonen. Πρόκειται για έναν συγκεκριμένο τύπο μη εποπτευόμενου νευρωνικού δικτύου που εκτελεί χωρικές ομαδοποιήσεις δεδομένων βασισμένο σε παρόμοιες συμπεριφορές. Οι (Melin et al., 2020) χρησιμοποίησαν έναν αυτό οργανωμένο χάρτη για να εξετάσουν περιπτώσεις COVID-19 σε διάφορες χώρες.

- Τα στατιστικά στοιχεία χωρικής σάρωσης για τον εντοπισμό σημαντικών συστάδων αθροιστικών περιπτώσεων COVID-19.
- Το GeoMEDD. Πρόκειται για μια νέα μεθοδολογία ανίχνευσης συστάδων σε πραγματικό χρόνο που παρέχει δείκτες για τη χωρική εξέλιξη της νόσου, με βάση την πρόσβαση σε διάφορες δημόσιες πηγές που λαμβάνουν υπόψη την τοποθεσία και το χρόνο των κρουσμάτων.

7.4.Παρεμβολή και γεωστατιστική

Οι μέθοδοι παρεμβολής που χρησιμοποιούνται σε έργα που αφορούν τα χωρικά και χωροχρονικά πρότυπα της πανδημίας, σχετίζονται με ατμοσφαιρικά (ρύπανση και κλίμα) καθώς και κοινωνικοοικονομικά θέματα.

Πολλές εργασίες δίνουν ιδιαίτερη προσοχή στα αυστηρά lockdown που έχουν επιβληθεί σε διάφορες χώρες, κυρίως στην Ευρώπη και την Ασία.

• Παρεμβολή στάθμισης αντίστροφης απόστασης (Inverse distance weighting interpolation-IDW): Οι (Saha et al., 2020) ανέλυσαν τον αντίκτυπο του lockdown του COVID-19 στην κινητικότητα της κοινότητας σε διάφορες πολιτείες της Ινδίας και χρησιμοποίησαν την IDW για να δείξουν τις τάσεις κίνησης πριν και μετά το lockdown.

• Voronoi: Οι (Bherwani et al., 2020) χρησιμοποίησαν πιθανοτική μοντελοποίηση Bayes για να κατανοήσουν τη σχέση μεταξύ των περιπτώσεων COVID-19 και της πυκνότητας του πληθυσμού σε μια δεδομένη περιοχή μαζί με διαγράμματα Voronoi που βασίζονται σε GIS για τον εντοπισμό περιοχών υψηλού κινδύνου. Τα πολύγωνα Thiessen οριοθετούν τα όρια της ζώνης κινδύνου.

• Η γεωστατιστική εκτιμά τις τιμές των φαινομένων σε περιοχές όπου οι τιμές είναι αβέβαιες ή δεν υπάρχουν με βάση τη συνδιακύμανση (το πώς αλλάζουν δύο τυχαίες μεταβλητές) ή το βαριόγραμμα (χωρική εξάρτηση μεταξύ στοχαστικών (τυχαίων) διεργασιών). Οι παρεμβολές Spline είναι μια άλλη τεχνική που χρησιμοποιεί διακριτά σημεία δεδομένων για τη μοντελοποίηση μιας συνεχούς μεταβλητής. Οι (Sui et al., 2020) χρησιμοποίησαν μια τεχνική κυβικού spline για να υπολογίσουν την ταχύτητα των λεωφορείων και των ταξί μέσω συσκευών δεδομένων GPS οχημάτων και αναζήτησαν την πιθανή αλλαγή στις εκπομπές στην μετά -COVID περίοδο.

• Η Kriging είναι ίσως η πιο συχνά χρησιμοποιούμενη γεωστατιστική μέθοδος και έχει χρησιμοποιηθεί στη βιβλιογραφία του COVID-19 όσον αφορά την πρόβλεψη κλιματικών μεταβλητών και κοινών ατμοσφαιρικών ρύπων. Στην τελευταία περίπτωση, η έρευνα χρησιμοποίησε το kriging για να εντοπίσει συσχετίσεις μεταξύ της ατμοσφαιρικής ρύπανσης και του COVID-19. Μια σχετική τεχνική είναι το cokriging. Οι (Kerimray et al., 2020) χρησιμοποίησαν το cokriging για να χαρτογραφήσουν τις κατανομές των PM 2,5 και του βενζολίου στο Αλμάτι του Καζακστάν, για τις περιόδους 2018–2019 και 2020, αντίστοιχα. Συγκεκριμένα, εξέτασαν την επίδραση των κρατικών περιορισμών στις συγκεντρώσεις των ρύπων.

7.5.Αχωρικά μοντέλα σε GIS

Πολλές μελέτες χρησιμοποιούν μη χωρικά μοντέλα και μεθόδους που συνδυάζονται με το GIS, στοιχεία επίσης σημαντικό να αναφερθεί λόγω της γεωγραφικής τους συμβολής στη μελέτη της πανδημίας.

• Η παλινδρόμηση Poisson έχει εφαρμοστεί για τον COVID-19 σε κοινωνικοοικονομικές μελέτες με υποδομές υπηρεσιών υγείας, ειδικά σε αστικά περιβάλλοντα.

• Η συσχέτιση Pearson έχει χρησιμοποιηθεί με όλα τα είδη μεταβλητών, αλλά ειδικά με κοινωνικό-οικονομικά δεδομένα, για χώρο-χρονική ανάλυση, χάρτες κινδύνου, προσβασιμότητα στην υγεία και περιβαλλοντικές επιπτώσεις λόγω της πανδημίας.

• Τα τεστ Spearman και Kendall έχουν χρησιμοποιηθεί σε επιβεβαιωμένα κρούσματα COVID-19 και κοινωνικό-οικονομικές μεταβλητές, καθώς και στην ποιότητα του κλίματος και του αέρα, για την ανάλυση της χώρο-χρονικής εξέλιξης της πανδημίας, κυρίως σε αστικά περιβάλλοντα.

• Η παλινδρόμηση Cox έχει χρησιμοποιηθεί για την παραγωγή χαρτών ευπάθειας και κινδύνου σε αστικά περιβάλλοντα.

• Το K-means είναι ένας αλγόριθμος ομαδοποίησης χωρίς επίβλεψη που χωρίζει τα δεδομένα με βάση τον πλησιέστερο μέσο όρο. Στην ανασκόπηση των (Franch-Pardo et al., 2020) προσδιορίζονται εκείνα τα έργα που χρησιμοποίησαν τον εν λόγω αλγόριθμο για την εκτέλεση ομαδοποίησης ως μέρος της διαδικασίας μιας χωρικής ανάλυσης. Οι (Lai et al., 2020) χρησιμοποίησαν τον αλγόριθμο αυτό με σκοπό να ομαδοποιήσουν τις κομητείες των ΗΠΑ με βάση τα κοινωνικό-δημογραφικά χαρακτηριστικά και τα δεδομένα COVID-19 ενώ οι (Abdallah et al., 2020) για δεδομένα τοποθεσίας GPS.

• Τα μοντέλα SIR μπορούν να προσθέσουν σαφείς γεωγραφικές μεταβλητές για τη μελέτη της δυναμικής της επιδημίας.

7.6.Πολυκριτηριακή ανάλυση

Η διεργασία αναλυτικής ιεραρχίας (analytic hierarchy process-AHP) αποτελεί ένα πολυκριτηριακό μοντέλο λήψης αποφάσεων. Πρόκειται για μία προσθετική και αντισταθμιστική τεχνική σύγκρισης βάσει ζευγαριών, που βασίζεται σε τρεις αρχές: την αποσύνθεση, τη συγκριτική αξιολόγηση και τον καθορισμό προτεραιοτήτων. Είναι ουσιαστικά μια διαδικασία για τον εντοπισμό, την κατανόηση και την αξιολόγηση των αλληλεπιδράσεων ενός συστήματος με ολιστικό τρόπο παρέχοντας μια κλίμακα για τη μέτρηση άυλων παραγόντων και μια μέθοδο

καθορισμού προτεραιοτήτων. Όσον αφορά τον COVID-19, έχει χρησιμοποιηθεί ευρέως σε θέματα που σχετίζονται με το περιβάλλον και τη γεωγραφία της υγείας μέσω της ανάπτυξης χαρτών κοινωνικής ευπάθειας και κινδύνων καθώς και της προσβασιμότητας στην υγεία. Οι (Requia et al., 2020) ανέπτυξαν ένα ιεραρχικό δίκτυο για ζητήματα χρήσης γης, κοινωνικό-οικονομικά, πληθυσμιακά, συνθήκες υγείας και για το σύστημα υγειονομικής περίθαλψης στη Βραζιλία. Οι (Mishra et al., 2020) χρησιμοποίησαν το AHP για να δημιουργήσουν έναν δείκτη ευπάθειας COVID-19 για αστικά περιβάλλοντα στην Ινδία και οι (Fang et al., 2020) πραγματοποίησαν παρόμοια ανάλυση για το νησί Xiamen, στην Κίνα.

Με τα εργαλεία GIS, έχει χρησιμοποιηθεί επίσης πολυκριτηριακή ανάλυση απόφασης (multi-criteria decision analysis-MCDA). Πρόκειται για μια μεθοδολογία που αφορά την αξιολόγηση εναλλακτικών λύσεων σε συγκεκριμένα θέματα, συχνά αντικρουόμενα, και τον συνδυασμό των λύσεων αυτών σε μια γενική αξιολόγηση. Χρησιμοποιώντας αυτή τη διαδικασία, οι (Sánchez-Sánchez et al., 2020) δημιούργησαν ένα μοντέλο ευπάθειας πληθυσμού όπου αξιολογείται και χαρτογραφείται η ευαισθησία στον κίνδυνο μετάδοσης του COVID-19 στο Chetumal του Μεξικού. Αντίστοιχες μελέτες πραγματοποιήθηκαν για την Καμπάλα της Ουγκάντα και το Νεπάλ.

7.7.GPS και δίκτυα

Οι μελέτες που χρησιμοποιούν εξόρυξη δεδομένων για τη μελέτη της ανθρώπινης κινητικότητας τείνουν να επικεντρώνονται σε περιοχές στις οποίες έχουν επιβληθεί lockdown.

Πραγματοποιείται διαφοροποίηση των μελετών αυτών βάση των εισροών τους:

• Τοποθεσία κινητού τηλεφώνου

• Διαδίκτυο και κοινωνικά δίκτυα: Google, Facebook, tweets με γεωγραφική ετικέτα στο Twitter, Weibo, Quingbo.

• Αεροπορικές και θαλάσσιες μεταφορές.

• Παρακολούθηση των δημόσιων συγκοινωνιών: λεωφορεία, ταξί και αστικά ποδήλατα (Nian et al., 2020)

• Δεδομένα με γεωγραφικές ετικέτες για εντοπισμένες λοιμώξεις που συγκεντρώθηκαν με ταξιδιωτικά ιστορικά.

Σχετικά με τη χρήση επιπέδων στα δίκτυα και τη ροή μεταφορών, δόθηκε έμφαση στα θέματα των οδικών δικτύων για την προσβασιμότητα στις υπηρεσίες υγείας και στην ανάλυση της αλλαγής στη ροή των οχημάτων λόγω του lockdown (Bessa & da Luz, 2020)

Περίοπτη θέση καταλαμβάνουν και οι πολεοδομικές μελέτες. Ως αποτέλεσμα, αντιμετωπίζεται το ζήτημα της υποδομής υγείας και της προσβασιμότητας, ανοίγοντας μια πολιτική συζήτηση για τους δημόσιους χώρους. Θα πρέπει να επισημανθεί η χρήση των εργαλείων OSMnx για τη λήψη χωρικών δεδομένων από το OpenStreetMap και τη μοντελοποίηση, την προβολή, την οπτικοποίηση και την ανάλυση δικτύων δρόμων.

8. Εφαρμογή στα δεδομένα κορονοϊού

Η υλοποίηση έγινε σε γλώσσα προγραμματισμού Python 3.8.3. Κατεβάσαμε το πακέτο Anaconda, το οποίο έχει προ εγκατεστημένες πολλές χρήσιμες βιβλιοθήκες για διαχείριση δεδομένων (pandas, numpy) και μηχανική μάθηση (sklearn). Ο κώδικας δίνεται στο παράρτημα, ο οποίος έχει κατάλληλα ονόματα μεταβλητών και σχόλια όπου χρειάζεται για να είναι κατανοητές οι διαδικασίες που υλοποιούνται.

Ως πηγή δεδομένων για τον κορονοϊό χρησιμοποιήθηκε η ιστοσελίδα [5]:
https://ourworldindata.org/coronavirus-data.

Η προηγούμενη πηγή συμπεριλαμβάνει δημογραφικά δεδομένα κάθε χώρας (πληθυσμός, πυκνότητα, κατά κεφαλήν ΑΕΠ κλπ) για το 2020.

Τα μετεωρολογικά δεδομένα για το 2020 λήφθηκαν από τον παρακάτω σύνδεσμο:
https://www.kaggle.com/vishalvjoseph/weather-dataset-for-covid19-predictions.

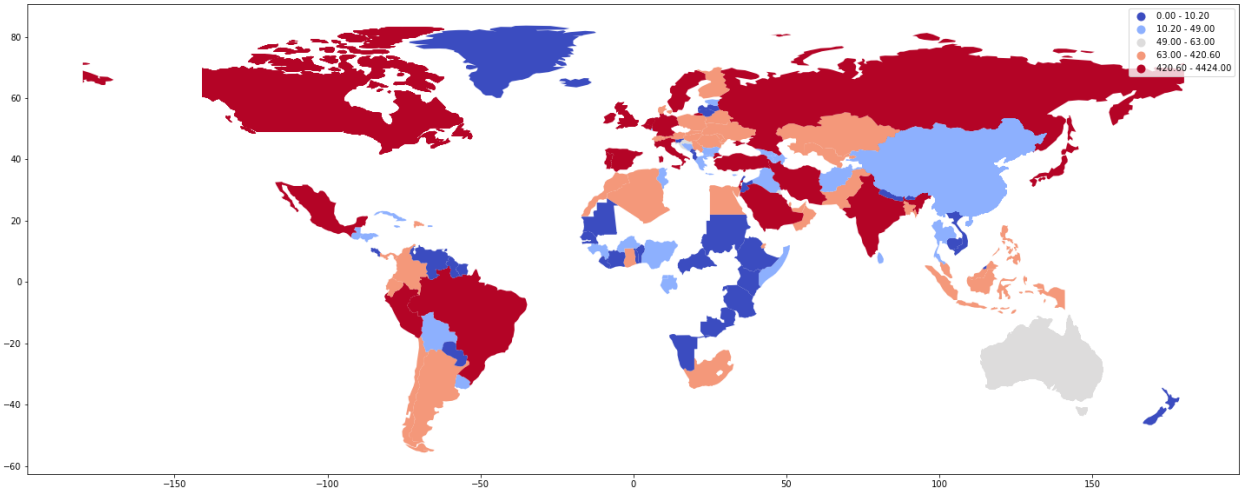
Ως μεταβλητές που θα χρησιμοποιηθούν για την μελέτη της εξάπλωσης του κορωνοϊού συγκεντρώνουμε δημογραφικά δεδομένα (πληθυσμός, πυκνότητα πληθυσμού, διάμεση ηλικία), οικονομικά δεδομένα (κατά κεφαλήν ΑΕΠ) και δεδομένα καιρού, όπως μια περιγραφή του καιρού (π.χ. ηλιοφάνεια, βροχή, χιόνι κλπ), την πιθανότητα βροχής, τη μέγιστη θερμοκρασία και την υγρασία. Επίσης, για έχουμε ένα πλήρες σύνολο δεδομένων, αφαιρούμε τις ελλιπείς τιμές (NaN). Αρχικά φορτώνουμε τα δεδομένα που συλλέξαμε σε csv αρχεία για κορονοϊό και καιρό στην Python. Συνδυάζουμε τα δεδομένα ανά χώρα και ημέρα σε ένα pandas dataframe.

8.1. Γεωγραφική Απεικόνιση

Επιλέξαμε το πακέτο geopandas για την απεικόνιση δεδομένων σε χάρτη, στο οποίο δίνουμε τα δεδομένα με το όνομα της χώρας και τα απεικονίζει σε χάρτη. Επίσης, διαλέξαμε δύο χρονικά σημεία των δεδομένων, την 15/02/2020 και την 15/04/2020 που είναι οι περίοδοι έναρξης και κορύφωσης του πρώτου κύματος κορονοϊού. Αρχικά απεικονίζουμε σε παγκόσμιο χάρτη τον αριθμό κρουσμάτων στις δύο ημερομηνίες στις εικόνες 5 και 6. Όπως παρατηρούμε, αρχικά έχουμε μεγάλο αριθμό κρουσμάτων μόνο στην Κίνα, ενώ μετά ο ιός έχει διαδοθεί στον υπόλοιπο κόσμο και κυρίως στο βόρειο ημισφαίριο.

Εικόνα 5. Κατανομή κρουσμάτων στις 15/02/2020.

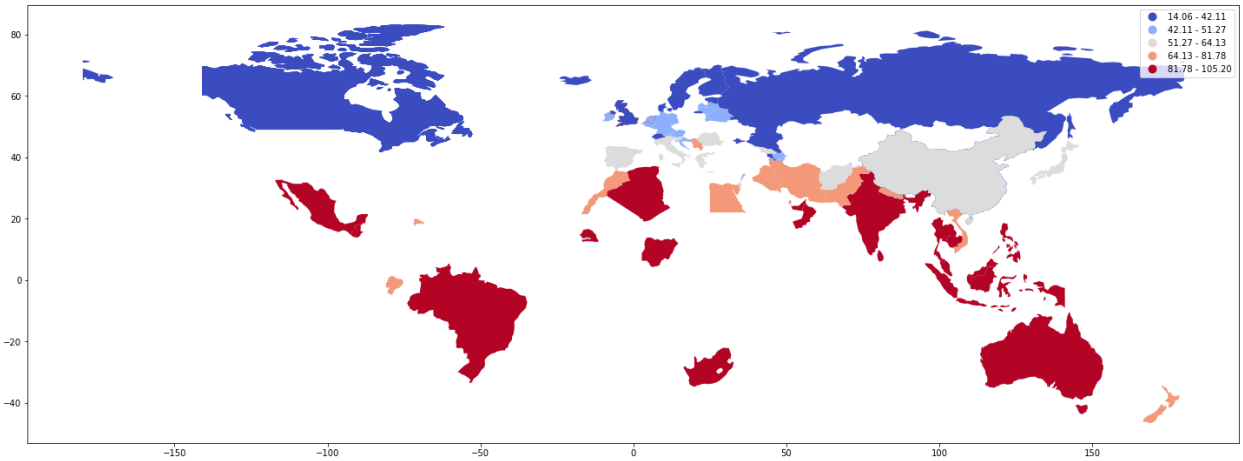
Πηγή: Ιδίας Επεξεργασίας



Εικόνα 6. Κατανομή κρουσμάτων στις 15/04/2020.

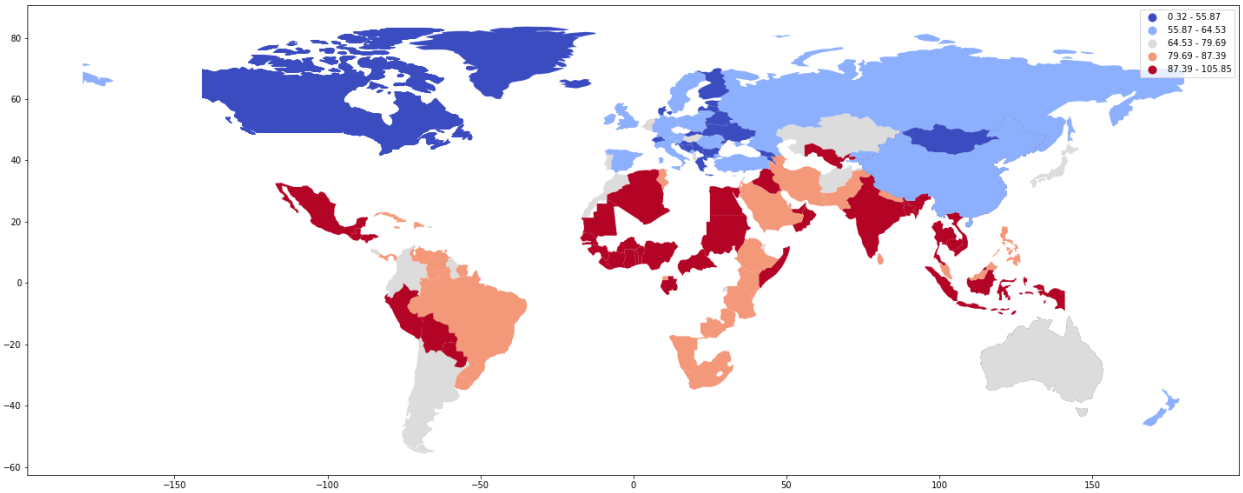
Πηγή: Ιδίας Επεξεργασίας

Μετά, στις εικόνες 7 και 8 απεικονίζεται η μέγιστη θερμοκρασία της ημέρας ανά χώρα για τις δύο ημερομηνίες. Όπως αναμένεται με βάση τις ημερομηνίες, έχουμε υψηλότερες θερμοκρασίες στη δεύτερη ημερομηνία για το βόρειο ημισφαίριο, αλλά χαμηλότερες για το νότιο.



Εικόνα 7. Μέγιστη θερμοκρασία στις 15/02/2020.

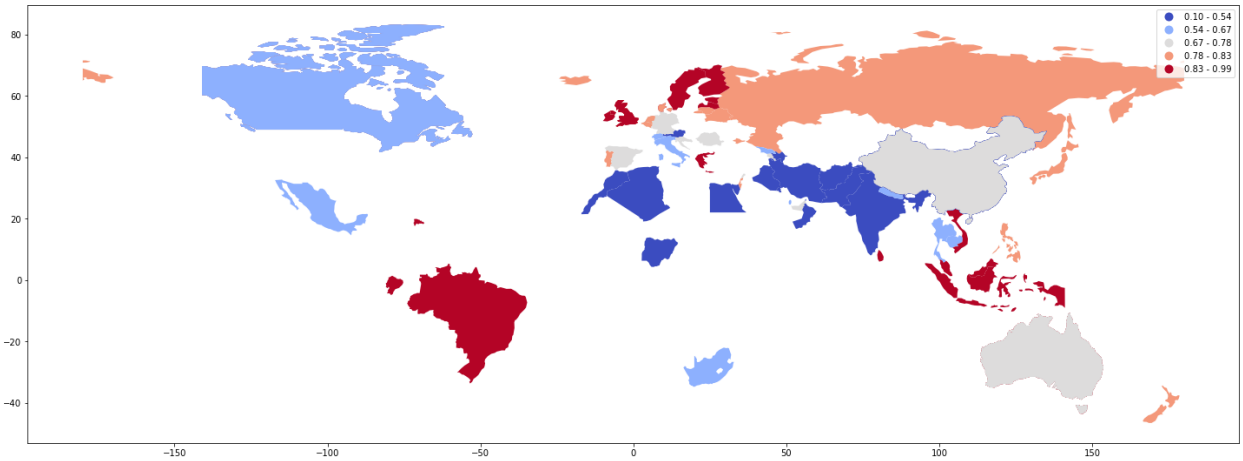
Πηγή: Ιδίας Επεξεργασίας



Εικόνα 8. Μέγιστη θερμοκρασία στις 15/04/2020.

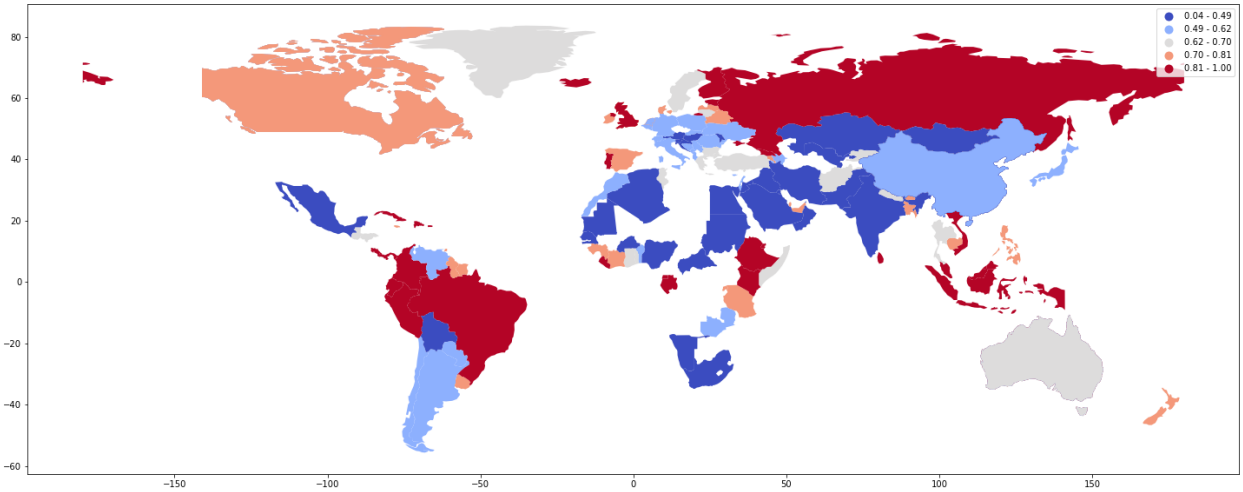
Πηγή: Ιδίας Επεξεργασίας

Ύστερα, απεικονίζουμε την υγρασία ανά χώρα για τις δύο ημερομηνίες στις εικόνες 9 και 10. Ανάλογα με την περιοχή, έχουμε διαφορές στην υγρασία, για παράδειγμα η Β. Ευρώπη έχει μικρότερη υγρασία, ενώ οι τροπικές περιοχές μεγαλύτερη.



Εικόνα 9. Υγρασία στις 15/02/2020.

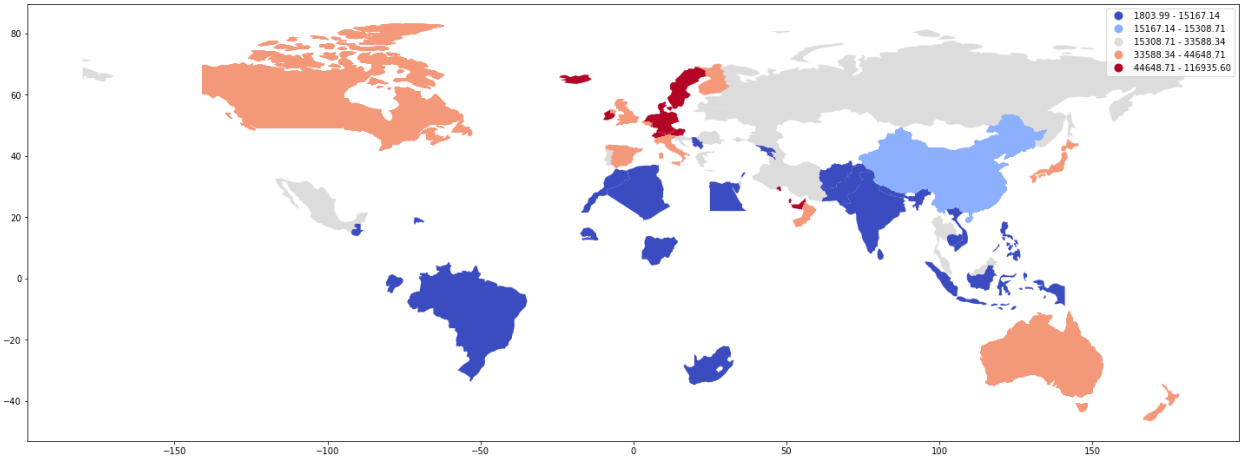
Πηγή: Ιδίας Επεξεργασίας



Εικόνα 10. Υγρασία στις 15/04/2020.

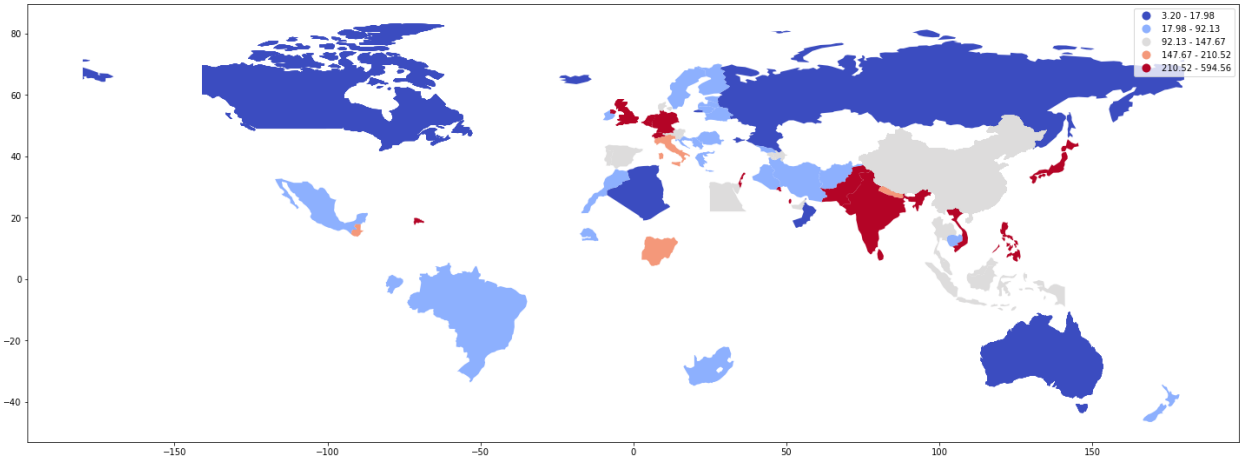
Πηγή: Ιδίας Επεξεργασίας

Τέλος, απεικονίζουμε τα δημογραφικά δεδομένα κατά κεφαλήν ΑΕΠ (εικόνα 11) και πυκνότητα πληθυσμού (εικόνα 12) για το 2020 για όλες τις χώρες.



Εικόνα 11. Κατά κεφαλήν ΑΕΠ.

Πηγή: Ιδίας Επεξεργασίας



Εικόνα 12. Πυκνότητα πληθυσμού

Πηγή: Ιδίας Επεξεργασίας

8.2. Αποτελέσματα Ταξινόμησης

Εδώ θέλουμε να κάνουμε πρόβλεψη για τη μεταβλητή y = "νέα κρούσματα ανά εκατομμύριο κατοίκων». Αρχικά, κάνουμε διακριτοποίηση της μεταβλητής σε τρία επίπεδα ανά εκατομμύριο:

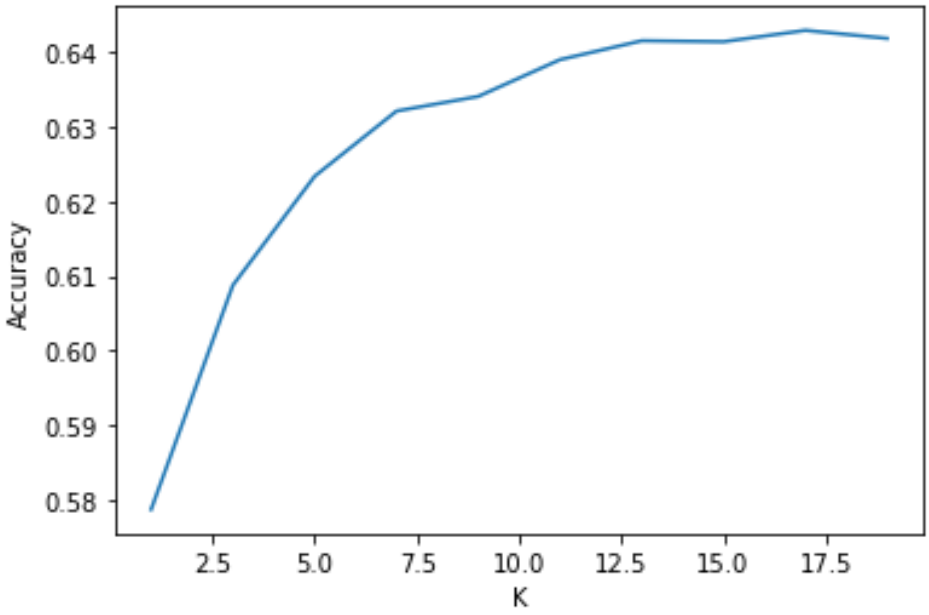
- Κατηγορία 0:
  y = 0 (καθόλου κρούσματα),
- Κατηγορία 1:
  $0 < y \leq 1$ (μικρός ρυθμός εξάπλωσης)
- Κατηγορία 2:

`1 < y ≤ 10 (μεγάλος ρυθμός εξάπλωσης)`

Από τα δεδομένα παρατηρούμε ότι δεν υπάρχει y > 10.

Στη συνέχεια, εφαρμόζουμε διάφορους αλγόριθμους μηχανικής μάθησης στα δεδομένα για ταξινόμηση. Συγκεκριμένα, χρησιμοποιούμε 10-fold cross-validation για να χωρίσουμε τυχαία τα δεδομένα σε 90% δεδομένα εκπαίδευσης και 10% δοκιμής.

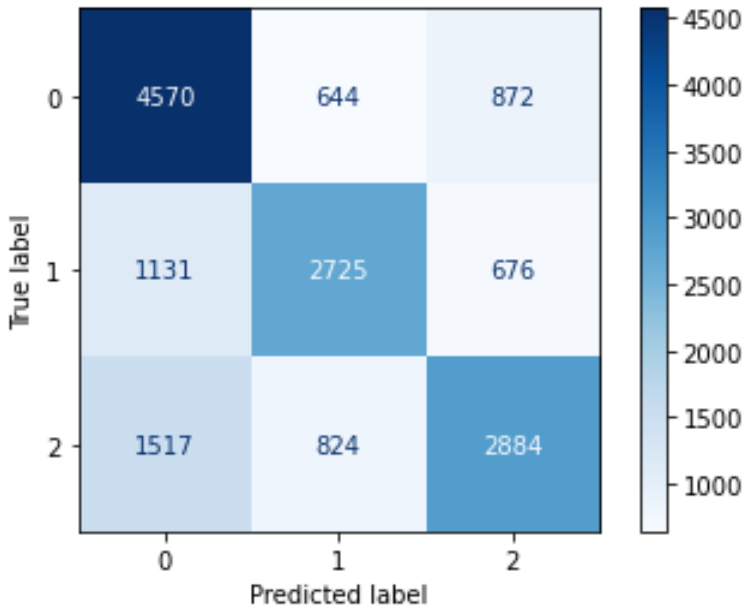
Πρώτα δοκιμάζουμε τον αλγόριθμο KNN. Η ακρίβεια πρόβλεψης φαίνεται στην εικόνα 13 για διάφορες τιμές της παραμέτρου K (αριθμός γειτόνων). Η μέγιστη ακρίβεια είναι 64.28% για K = 17.



Εικόνα 13. Ακρίβεια του αλγορίθμου KNN για διάφορες τιμές του K.

Πηγή: Ιδίας Επεξεργασίας

Οπότε εκτελούμε ξανά τον αλγόριθμο για K = 17 και παίρνουμε τον πίνακα σύγχυσης που δείχνει τις πραγματικές και τις προβλεφθείσες κατηγορίες. Ιδανικά, η διαγώνιος πρέπει να περιέχει μη μηδενικές τιμές και τα υπόλοιπα κελιά μηδενικές τιμές. Ο πίνακας σύγχυσης φαίνεται παρακάτω στην εικόνα 14.

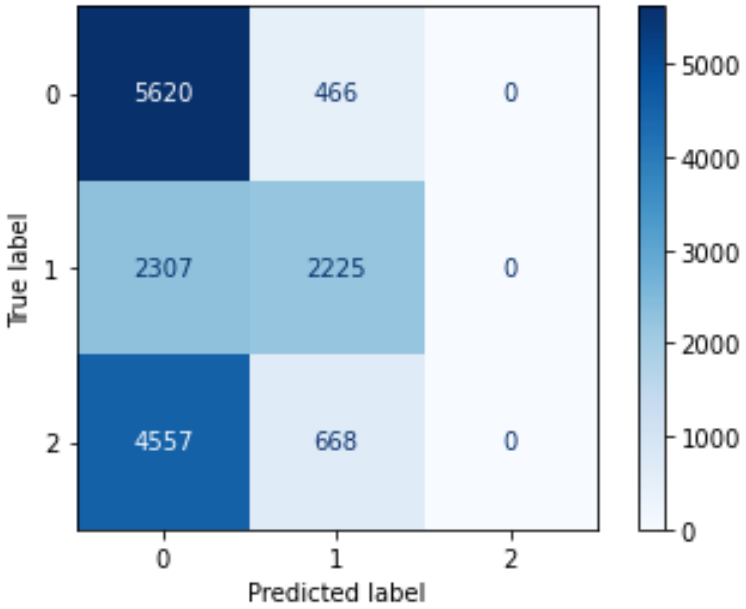


Εικόνα 14. Πίνακας σύγχυσης για τον αλγόριθμο KNN.

Πηγή: Ιδίας Επεξεργασίας

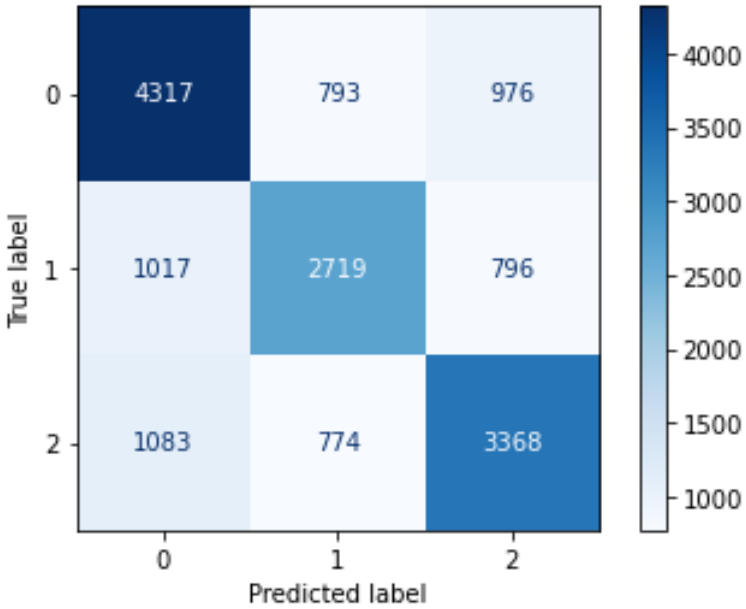
Στη συνέχεια εκτελέσαμε τον αλγόριθμο SVM με πυρήνα RBF και αυτόματη εύρεση της παραμέτρου γ. Επειδή έχουμε τρεις κατηγορίες, η προεπιλογή της βιβλιοθήκης scikit-learn είναι η δημιουργία one-versus-one μοντέλων. Η ακρίβεια είναι 49.52%. Από τον πίνακα σύγχυσης της εικόνας 15 παρατηρούμε ότι ο αλγόριθμος αποτυγχάνει να ανιχνεύει την τρίτη κατηγορία, για αυτό έχει χαμηλή απόδοση.

Τέλος, εφαρμόζουμε τον αλγόριθμο Random Forest με 100 δέντρα, ενώ η παράμετρος mtry τέθηκε στην προεπιλεγμένη τιμή της τετραγωνικής ρίζας των μεταβλητών, εδώ 3. Η ακρίβεια είναι 65.66. Παρακάτω στην εικόνα 16 φαίνεται ο πίνακας σύγχυσης.



Εικόνα 15.Πίνακας σύγχυσης για τον αλγόριθμο SVM.

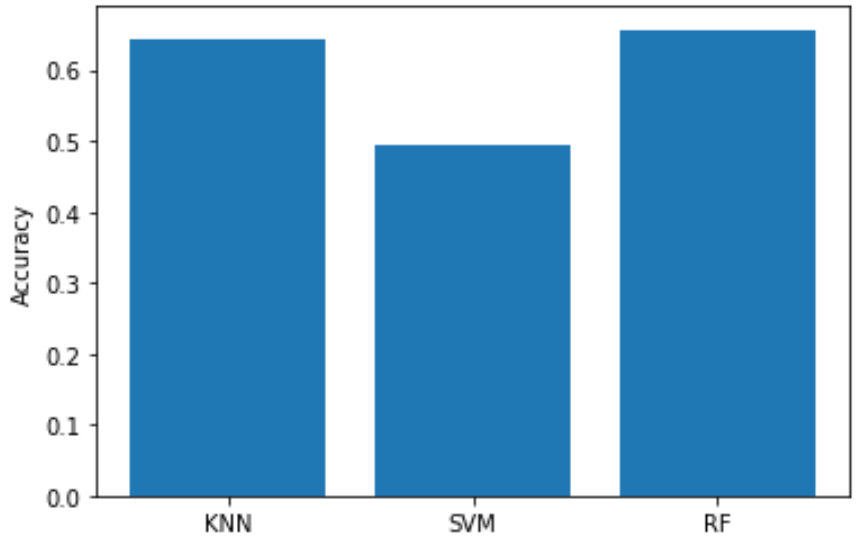
Πηγή: Ιδίας Επεξεργασίας



Εικόνα 16. Πίνακας σύγχυσης για τον αλγόριθμο Random Forest.

Πηγή: Ιδίας Επεξεργασίας

Η ακρίβεια των τριών αλγορίθμων φαίνεται συγκεντρωτικά στην εικόνα 17. Η καλύτερη ακρίβεια βρέθηκε με τον αλγόριθμο Random Forest, αλλά είναι λίγο μεγαλύτερη από τον KNN, ενώ ο SMV δεν μπορεί να δώσει σωστά αποτελέσματα σε αυτά τα δεδομένα.



Εικόνα 17. Σύγκριση της ακρίβειας των τριών αλγορίθμων

Πηγή: Ιδίας Επεξεργασίας

Επίσης, από τον αλγόριθμο Random Forest μπορούμε να πάρουμε μια μετρική, την Variable Importance, η οποία δείχνει πόσο σημαντική είναι η κάθε μεταβλητή για την πρόβλεψη. Το αποτέλεσμα φαίνεται στον Πίνακα 1. Παρατηρούμε ότι τα δημογραφικά και οικονομικά στοιχεία δεν παίζουν τόσο σημαντικό ρόλο, όσο ο καιρός. Η πιο σημαντική μεταβλητή ήταν η μέγιστη θερμοκρασία, μετά η υγρασία και τέλος η πιθανότητα βροχής. Οπότε, τα δεδομένα καιρού φαίνεται να συσχετίζονται περισσότερο με την μετάδοση του κορονοϊού. Επίσης, πρέπει να παρατηρήσουμε ότι τα δημογραφικά δεδομένα είναι σταθερά για κάθε χώρα, ενώ για τον καιρό έχουμε καθημερινές παρατηρήσεις, όπως και για τον αριθμό κρουσμάτων. Σημειώνεται ότι η ακρίβεια για τρεις κλάσεις με τυχαία πρόβλεψη είναι 33%, οπότε οι αλγόριθμοι KNN και Random Forest πετυχαίνουν αρκετά ικανοποιητική ακρίβεια.

Πίνακας 1.Σημαντικότητα Μεταβλητών με βάση το Random Forest.

Μεταβλητή	**Σημαντικότητα**
population	0.09695
population_density	0.04880
median_age	0.05824
gdp_per_capita	0.06671
weather	0.04234
precipProbability	0.18632
temperatureHigh	0.30983
humidity	0.19078

Πηγή: Ιδίας Επεξεργασίας

9. Συμπεράσματα

Σε αυτή την εργασία εξετάστηκε το πρόβλημα της πρόβλεψης της εξάπλωσης του κορονοϊού με μεθόδους μηχανικής μάθησης. Αρχικά παρουσιάστηκαν θεωρητικά το πεδίο της μηχανικής μάθησης και μερικές δημοφιλείς μέθοδοι μηχανικής μάθησης για ταξινόμηση: KNN, SVM, Random Forest. Πιο συγκεκριμένα, ο σκοπός είναι να προβλέψουμε τον ημερήσιο αριθμό των νέων κρουσμάτων ανά εκατομμύριο πληθυσμού. Τα δεδομένα που χρησιμοποιήθηκαν ως μεταβλητές είναι δημογραφικά, οικονομικά και ο καιρός. Πρώτα έγιναν γεωγραφικές απεικονίσεις των δεδομένων για να κατανοήσουμε πιο εύκολα την κατανομή τους ανά χώρα. Μετά προχωρήσαμε στην εκπαίδευση τριών μοντέλων με βάση τις μεθόδους που προαναφέραμε. Η καλύτερη απόδοση σε ακρίβεια βρέθηκε από τον αλγόριθμο Random Forest και ήταν 65.66%. Η απόδοση του αλγορίθμου KNN ήταν παραπλήσια (64%), ενώ ο αλγόριθμος SVM είχε χειρότερη απόδοση σε αυτό το πρόβλημα (50%). Λαμβάνοντας υπόψη ότι ο αλγόριθμος Random Forest είναι πιο πολύπλοκος, είναι αναμενόμενο να έχει καλύτερες επιδόσεις. Επίσης, από τον αλγόριθμο Random Forest μπορούμε να εξάγουμε τη σημαντικότητα των μεταβλητών και να κατανοήσουμε ποιες μεταβλητές παίζουν σημαντικό ρόλο για την πρόβλεψη της κατηγορίας. Διαπιστώσαμε ότι πιο σημαντικές είναι οι μεταβλητές που σχετίζονται με τον καιρό, δηλαδή η θερμοκρασία, η υγρασία και η πιθανότητα βροχής. Όπως είναι γνωστό, οι περισσότεροι ιοί μεταδίδονται περισσότερο το χειμώνα, όποτε και παρατηρείται έξαρση των κρουσμάτων, ενώ η εξάπλωση περιορίζεται όταν βελτιώνεται ο καιρός. Σημειώνεται ότι αυτό προκύπτει μόνο από τα δεδομένα (data-driven), χωρίς να θεωρήσουμε εμείς εκ των προτέρων κάποια μεταβλητή πιο σημαντική από κάποια άλλη. Ως μελλοντική εργασία, μπορούν να δοκιμαστούν διαφορετικοί αλγόριθμοι ταξινόμησης, που ενδεχομένως δώσουν καλύτερη ακρίβεια ταξινόμησης. Επίσης, μπορεί κάποιος να δοκιμάσει να χρησιμοποιήσει νέες μεταβλητές, οι οποίες ενδεχομένως να αναδείξουν επιπλέον λόγους εξάπλωσης του κορονοϊού.

Βιβλιογραφία

Abdallah, H. S., Khafagy, M. H., & Omara, F. A. (2020). Case Study: Spark GPU-Enabled Framework to Control COVID-19 Spread Using Cell-Phone Spatio-Temporal Data. *CMC-COMPUTERS MATERIALS & CONTINUA*, *65*(2), 1303-1320.

Agrebi, S., & Larbi, A. (2020). Use of artificial intelligence in infectious diseases. In *Artificial intelligence in precision health* (pp. 415-438). Academic Press.

Alkhaldy, I. A. GIS APPLICATION FOR MODELING COVID-19 RISK IN THE MAKKAH REGION, SAUDI ARABIA, BASED ON POPULATION AND POPULATION DENSITY.

Ashfaque, J. M., & Iqbal, A. (2019). Introduction to Support Vector Machines and Kernel Methods. publication at https://www. researchgate. net/publication/332370436.

Bent, O., Remy, S., Roberts, S., & Walcott-Bryant, A. (2018, April). Novel exploration techniques (NETs) for malaria policy interventions. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Bernasconi, A., & Grandi, S. (2021). A Conceptual Model for Geo-Online Exploratory Data Visualization: The Case of the COVID-19 Pandemic. *Information*, *12*(2), 69.

Bessa, K., & da Luz, R. A. (2020). A pandemia de Covid-19 e as particularidades regionais da sua difusão no segmento de rede urbana no estado do Tocantins, Brasil. *Ateliê Geográfico*, *14*(2), 6-28.

Bherwani, H., Anjum, S., Kumar, S., Gautam, S., Gupta, A., Kumbhare, H., ... & Kumar, R. (2021). Understanding COVID-19 transmission through Bayesian probabilistic modeling and GIS-based Voronoi approach: a policy perspective. *Environment, Development and Sustainability*, *23*(4), 5846-5864.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Butt, M. A., Khalid, A., Ali, A., Mahmood, S. A., Sami, J., Qureshi, J., & Waheed, K. (2019). Towards a Web GIS-based approach for mapping a dengue outbreak. *Applied Geomatics*, 1-11.

Casti, E. (2020). GEOGRAFIA A "VELE SPIEGATE". ANALISI TERRITORIALE E MAPPING RIFLESSIVO SUL COVID-19 IN ITALIA. *documenti geografici*, (1), 61-83.

Chandiok, A., & Chaturvedi, D. K. (2015, December). Machine learning techniques for cognitive decision making. In *2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI)* (pp. 1-6). IEEE.

Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297.

Das, A., Ghosh, S., Das, K., Basu, T., Dutta, I., & Das, M. (2021). Living environment matters: Unravelling the spatial clustering of COVID-19 hotspots in Kolkata megacity, India. *Sustainable Cities and Society*, *65*, 102577.

de Lusignan, S., Dorward, J., Correa, A., Jones, N., Akinyemi, O., Amirthalingam, G., ... & Hobbs, F. R. (2020). Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. *The Lancet Infectious Diseases*, *20*(9), 1034-1042.

Delmelle, E., Hagenlocher, M., Kienberger, S., & Casas, I. (2016). A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia. *Acta tropica*, *164*, 169-176.

Devasia, J. T., Lakshminarayanan, S., & Kar, S. S. (2020). How Modern Geographical Information Systems Based Mapping and Tracking Can Help to Combat Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Pandemic around the World and India. *International Journal of Health Systems and Implementation Research*, *4*(1), 30-54.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, *20*(5), 533-534.

Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environmental health perspectives*, *112*(9), 998-1006.

Fan, Z., Zhan, Q., Yang, C., Liu, H., & Zhan, M. (2020). How did distribution patterns of particulate matter air pollution (PM2. 5 and PM10) change in China during the COVID-19 outbreak: A spatiotemporal investigation at Chinese city-level. *International Journal of Environmental Research and Public Health*, *17*(17), 6274.

Fang, L., Huang, J., Zhang, Z., & Nitivattananon, V. (2020). Data-driven framework for delineating urban population dynamic patterns: Case study on Xiamen Island, China. *Sustainable Cities and Society*, *62*, 102365.

Fix, Evelyn; Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties *(Report). USAF School of Aviation Medicin*e, Randolph Field, Texas.

Franch-Pardo, B. N. Spatial analysis and GIS in the study of COVID-19. A review Ivan Franch-Pardo, Brian M. Napoletano b, Fernando Rosete-Verges, Lawal Billa c.

Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. *Science of The Total Environment*, *739*, 140033.

Hannah Ritchie, Esteban Ortiz-Ospina, Diana Beltekian, Edouard Mathieu, Joe Hasell, Bobbie Macdonald, Charlie Giattino, Cameron Appel, Lucas Rodés-Guirao and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource].

Iyanda, A. E., Adeleke, R., Lu, Y., Osayomi, T., Adaralegbe, A., Lasode, M., ... & Osundina, A. M. (2020). A retrospective cross-national examination of COVID-19 outbreak in 175 countries: a multiscale geographically weighted regression analysis (January 11-June 28, 2020). *Journal of infection and public health*, *13*(10), 1438-1445.

Kerimray, A., Baimatova, N., Ibragimova, O. P., Bukenov, B., Kenessov, B., Plotitsyn, P., & Karaca, F. (2020). Assessing air quality changes in large cities during COVID-19 lockdowns: The impacts of traffic-free urban conditions in Almaty, Kazakhstan. *Science of the Total Environment*, *730*, 139179.

Kersting, K. (2018). Machine learning and artificial intelligence: two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in big Data*, *1*, 6.

Khan, O. A., Davenhall, W., Ali, M., Castillo-Salgado, C., Vazquez-Prokopec, G., Kitron, U., ... & Clements, A. C. A. (2010). Geographical information systems and tropical medicine. *Annals of Tropical Medicine & Parasitology*, *104*(4), 303-318.

Kirby, T. (2020). Evidence mounts on the disproportionate effect of COVID-19 on ethnic minorities. *The Lancet Respiratory Medicine*, *8*(6), 547-548.

Lai, Y., Charpignon, M. L., Ebner, D. K., & Celi, L. A. (2020). Unsupervised learning for county-level typological classification for COVID-19 research. *Intelligence-based medicine*, *1*, 100002.

Maiti, A., Zhang, Q., Sannigrahi, S., Pramanik, S., Chakraborti, S., Cerda, A., & Pilla, F. (2021). Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the contiguous United States. *Sustainable cities and society*, *68*, 102784.

Mansour, S., Al Kindi, A., Al-Said, A., Al-Said, A., & Atkinson, P. (2021). Sociodemographic determinants of COVID-19 incidence rates in Oman: Geospatial modelling using multiscale geographically weighted regression (MGWR). *Sustainable cities and society*, *65*, 102627.

Melin, P., Monica, J. C., Sanchez, D., & Castillo, O. (2020). Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps. *Chaos, Solitons & Fractals*, *138*, 109917.

Messner, C. B., Demichev, V., Wendisch, D., Michalick, L., White, M., Freiwald, A., ... & Ralser, M. (2020). Clinical classifiers of COVID-19 infection from novel ultra-high-throughput proteomics. *MedRxiv*.

Mishra, S. V., Gayen, A., & Haque, S. M. (2020). COVID-19 and urban vulnerability in India. *Habitat international*, *103*, 102230.

Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the total environment*, *728*, 138884.

O'Sullivan, D., Gahegan, M., Exeter, D. J., & Adams, B. (2020). Spatially explicit models for exploring COVID-19 lockdown strategies. *Transactions in GIS*, *24*(4), 967-1000.

Pareek, M., Bangash, M. N., Pareek, N., Pan, D., Sze, S., Minhas, J. S., ... & Khunti, K. (2020). Ethnicity and COVID-19: an urgent public health research priority. *The Lancet*, *395*(10234), 1421-1422.

Pourghasemi, H. R., Pouyan, S., Heidari, B., Farajzadeh, Z., Shamsi, S. R. F., Babaei, S., ... & Sadeghian, F. (2020). Spatial modeling, risk mapping, change detection, and outbreak trend analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020). *International Journal of Infectious Diseases*, *98*, 90-108.

Pourhomayoun, M., & Shakibi, M. (2020). Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. *MedRxiv*.

Sarkar, J., & Chakrabarti, P. (2020). A machine learning model reveals older age and delayed hospitalization as predictors of mortality in patients with COVID-19. medRxiv. *Preprint posted online March*, *30*.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1-67.

Wiens, J., Guttag, J., & Horvitz, E. (2016). Patient risk stratification with time-varying parameters: a multitask learning approach. *The Journal of Machine Learning Research*, *17*(1), 2797-2819.

Wong, Z. S., Zhou, J., & Zhang, Q. (2019). Artificial intelligence for infectious disease big data analytics. *Infection, disease & health*, *24*(1), 44-48.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., ... & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, *369*.

Yan, L., Zhang, H., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... & Yuan, Y. (2020). A machine learning-based model for survival prediction in patients with severe COVID-19 infection. medRxiv. *Preprint posted online on March*, *17*.

Παράρτημα – Κώδικας

Αρχείο data_prepare.py

```
# -*- coding: utf-8 -*-

import pandas as pd

#get weather data
df_w = pd.read_csv('daily_weather_2020.csv')
#keep important columns
df_w = df_w[['Country/Region', 'Province/State', 'time',
'icon','precipIntensity', 'precipProbability', 'precipType',
'temperatureHigh', 'temperatureLow', 'humidity',
'pressure','windSpeed','cloudCover','ozone' ]]
#rename columns
df_w.rename(columns={"Country/Region": "country"}, inplace=True)
df_w.rename(columns={"time": "date"}, inplace=True)

#get covid data
df_c = pd.read_csv('covid.csv')
#keep important columns
df_c =
df_c[['iso_code','continent','location','date','total_cases','new_cases','tot
al_deaths','new_deaths','total_cases_per_million','new_cases_per_million','to
tal_deaths_per_million','new_deaths_per_million','population','population_den
sity','median_age','aged_65_older','aged_70_older','gdp_per_capita','life_exp
ectancy','extreme_poverty','cardiovasc_death_rate',
'diabetes_prevalence','female_smokers','male_smokers','hospital_beds_per_thou
sand','human_development_index']]
#rename columns
df_c.rename(columns={"location": "country"}, inplace=True)


#merge covid and weather data based on country, date
df = df_c.merge(df_w, on=['country', 'date'], how='inner')

#get weather categories as numbers
weather = pd.factorize(df['icon'])
df['weather'] = weather[0]
print(weather[1])

#save to python pickle file
df.to_pickle("df_covid_data.pkl")
```

Αρχείο country_plots.py

```
# -*- coding: utf-8 -*-

#first time run:
#conda install geopandas
#conda install mapclassify
#conda install descartes

import pandas as pd
import geopandas as gpd

df = pd.read_pickle("df_covid_data.pkl")

world = gpd.read_file(gpd.datasets.get_path("naturalearth_lowres"))

world.columns=['pop_est', 'continent', 'name', 'iso_code', 'gdp_md_est',
'geometry']

#first date
date_sel = '2020-02-15'
df_date = df[df['date']==date_sel]

merge=pd.merge(world,df_date,on='iso_code')

merge.plot(column='new_cases', scheme="quantiles",
           figsize=(25, 20),
           legend=True,cmap='coolwarm')

merge.plot(column='temperatureHigh', scheme="quantiles",
           figsize=(25, 20),
           legend=True,cmap='coolwarm')

merge.plot(column='humidity', scheme="quantiles",
           figsize=(25, 20),
           legend=True,cmap='coolwarm')

merge.plot(column='gdp_per_capita', scheme="quantiles",
           figsize=(25, 20),
           legend=True,cmap='coolwarm')

merge.plot(column='population_density', scheme="quantiles",
           figsize=(25, 20),
           legend=True,cmap='coolwarm')



#second date
date_sel = '2020-03-15'
df_date = df[df['date']==date_sel]

merge2=pd.merge(world,df_date,on='iso_code')

merge2.plot(column='new_cases', scheme="quantiles",
           figsize=(25, 20),
           legend=True,cmap='coolwarm')
```

```
merge2.plot(column='temperatureHigh', scheme="quantiles",
            figsize=(25, 20),
            legend=True,cmap='coolwarm')

merge2.plot(column='humidity', scheme="quantiles",
            figsize=(25, 20),
            legend=True,cmap='coolwarm')
```

Αρχείο classify.py

```
# -*- coding: utf-8 -*-

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import KFold
from sklearn.neighbors import KNeighborsClassifier
from sklearn import svm
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay

#%% classify - prepare data

#load data from pickle file
df = pd.read_pickle("df_covid_data.pkl")

#keep important columns
data =
df[['population','population_density','median_age','gdp_per_capita','weather'
,'precipProbability','temperatureHigh','humidity', 'new_cases_per_million']]
#remove rows with na
data = data.dropna()

#create labels for classification
labels = data['new_cases_per_million']
#remove labels from data
data.drop(["new_cases_per_million"], axis=1, inplace=True)

#convert to table
data = np.array(data)
#keep values>=0
x = labels>=0
labels = labels[x]
data = data[x,:]
#create categories
labels[labels==0] = 0
labels[(labels>0) & (labels<=1)] = 1
labels[(labels>1) & (labels<=10)] = 2
labels[(labels>10) ] = 3 #no category 3
labels = np.array(labels, dtype=np.int32) #convert to int

N = data.shape #data size
#10 fold crossvalidation
kf = KFold(n_splits=10, shuffle=True,  random_state=101)


#%% KNN

#try K = 1 to 21 with step = 2

pred_knn = np.zeros([N[0]])
acc_knn_all = np.zeros(20)
for nn in range(1,21,2):
    i = 0
    for tr, ts in kf.split(data,labels):
```

```
        print(i)
        x_train = data[tr,:]
        y_train = labels[tr]
        x_test = data[ts,:]
        y_test = labels[tr]

        knn = KNeighborsClassifier(n_neighbors=nn)
        knn.fit(x_train, y_train)
        pred_knn[ts] = knn.predict(x_test)

        i = i + 1
    #end_for

    pred_knn = pred_knn.astype('int64')
    acc_knn_all[nn] = np.sum(pred_knn==labels)/len(labels)
    print("Acc_knn = ", acc_knn_all[nn])

#create plot
x = np.arange(1,21,2)
plt.plot(x,acc_knn_all[x])
plt.xlabel('K')
plt.ylabel('Accuracy')
#plt.axis([0, 20, 0, 0.8])
plt.show()

print(np.max(acc_knn_all))
print(np.argmax(acc_knn_all))

#%% run for best k
nn = np.argmax(acc_knn_all)
i = 0
for tr, ts in kf.split(data,labels):
    print(i)
    x_train = data[tr,:]
    y_train = labels[tr]
    y_train2 = labels[tr]
    x_test = data[ts,:]


    knn = KNeighborsClassifier(n_neighbors=nn)
    knn.fit(x_train, y_train2)
    pred_knn[ts] = knn.predict(x_test)

    i = i + 1

#end_for

pred_knn = pred_knn.astype('int64')
acc_knn = np.sum(pred_knn == labels) / len(labels)
conf_knn = confusion_matrix(labels, pred_knn)
print("KNN")
print("Accuracy    ", acc_knn)
print("Confusion Matrix")
print(conf_knn)
h = ConfusionMatrixDisplay(conf_knn)
h.plot(cmap='Blues')
```

```
#%% SVM rbf
pred_svm = np.zeros([N[0]])

i = 0
for tr, ts in kf.split(data,labels):
    print(i)
    x_train = data[tr,:]
    y_train = labels[tr]
    x_test = data[ts,:]

    svm_model = svm.SVC() #rbf kernel, auto-search for best gamma
    svm_model.fit(x_train, y_train)
    pred_svm[ts] = svm_model.predict(x_test)

    i = i + 1

#end_for


pred_svm = pred_svm.astype('int64')
acc_svm = np.sum(pred_svm==labels)/len(labels)
conf_svm = confusion_matrix(labels, pred_svm)
print("SVM")
print("Accuracy    ", acc_svm)
print("Confusion Matrix")
print(conf_svm)
h = ConfusionMatrixDisplay(conf_svm)
h.plot(cmap='Blues')

#%% RF
np.random.seed(1) #set random seed, so we always get same random numbers

#mtry = np.round(0.5*np.sqrt(N[1])).astype('int64')
mtry = np.round(np.sqrt(N[1])).astype('int64')
#mtry = np.round(2*np.sqrt(N[1])).astype('int64')

pred_rf = np.zeros([N[0]])
imp_rf = np.zeros((10, N[1]))
i = 0
for tr, ts in kf.split(data,labels):
    print(i)
    x_train = data[tr,:]
    y_train = labels[tr]
    x_test = data[ts,:]

    rf = RandomForestClassifier(n_estimators=100, max_features = mtry)
    rf.fit(x_train, y_train)
    pred_rf[ts] = rf.predict(x_test)
    imp_rf[i,:] = rf.feature_importances_

    i = i + 1

#end_for


pred_rf = pred_rf.astype('int64')
acc_rf = np.sum(pred_rf==labels)/len(labels)
```

```
conf_rf = confusion_matrix(labels, pred_rf)
print("RF")
print("Accuracy    ", acc_rf)
print("Confusion Matrix")
print(conf_rf)
h = ConfusionMatrixDisplay(conf_rf)
h.plot(cmap='Blues')
#get feature importance (mean of 10 folds)
imp = np.mean(imp_rf,axis=0)
print(imp)

#%% plots - compare algorithms
accs = [acc_knn, acc_svm, acc_rf]

plt.bar([1,2,3],accs,tick_label=['KNN','SVM','RF'])
plt.ylabel('Accuracy')
plt.show()
```